Cognition in Flux

Proceedings of the 32nd Annual Meeting of the Cognitive Science Society

Portland, Oregon, August 11-14, 2010

Edited by

Stellan Ohlsson
University of Illinois at Chicago

Richard Catrambone
Georgia Institute of Technology



Austin, TX: Cognitive Science Society

How to cite a paper in these Proceedings:

APA formatted citation for a 6-page paper:
Author, A. & Author, B. (2010). This is the title of the paper. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. NUMBERS). Austin, TX: Cognitive Science Society.

APA formatted citation for a published abstract:
Author, A. & Author, B. (2010). This is the title of the abstract [Abstract]. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (p. NUMBER). Austin, TX: Cognitive Science Society.

APA formatted citation for a talk (or poster) presentation:
Author, A. & Author, B. (2010, August). This is the title of the talk or poster. Paper (or Poster) presented at the 32nd Annual Conference of the Cognitive Science Society. Portland, OR.

Table of Contents

Table of Contents i

Introduction 1

Organizing Committee 3

Program Committee 4

Awards 5

Sponsors 7

Invited Plenary Presentations 8

Symposia 9

Tutorials and Workshops 13

Foundations: Steps to Unification

Wild Systems Theory: Overcoming the Computational-Ecological Divide 15
Jerome Scott Jordan

Framing and Resource Activation: Bridging the Cognitive-Situative Divide Using a Dynamic Unit of Cognitive Analysis 19
Luke Conlin, Ayush Gupta, David Hammer

Modeling Personality and Individual Differences: The Approach-Avoid-Conflict Triad 25
Karl Fua, William Revelle, Andrew Ortony

Understanding the Brain as an Endogenously Active Mechanism 31
William Bechtel, Adele Abrahamsen

Language: Novel Approaches

Likability-Based Genres: Analysis and Evaluation of the Netflix Dataset 37
Andrew Olney

Constructing Typing-Time Corpora: A New Way to Answer Old Questions 43
Uriel Cohen Priva

Large-Scale Acquisition of Feature-Based Conceptual Representations from Textual Corpora 49
Barry Devereux, Nicholas Pilkington, Thierry Poibeau, Anna Korhonen

Are Random Representations Accurate Approximations of Lexical Semantics? 55
Brendan Johns, Michael Jones

Induction: Basic Processes

A Neural Model of Rule Generation in Inductive Reasoning 61
Daniel Rasmussen, Chris Eliasmith

The Effects of Domain and Type of Knowledge on Category-Based Inductive Reasoning 67
Aimée Kay Crisp-Bright, Aidan Feeney

Learning hypothesis spaces and dimensions through concept learning 73
Joseph Austerweil, Thomas Griffiths

The Role of "Explaining Away" in Human Abstract Rule Induction 79
Colin Dawson, LouAnn Gerken

Cognitive Architecture: Central Processing

An Analysis of the Working Memory Capacity Paradox 85
Eddy J. Davelaar

Working Memory Load Affects Device-Specific but Not Task-Specific Error Rates 91
Maartje Ament, Anna Cox, Ann Blandford, Duncan Brumby

An examination of the ERP correlates of recognition memory using state-trace analysis. 97
Emily Freeman, Simon Dennis, John Dunn

Modeling Change in Recognition Bias with the Progression of Alzheimer's 103
James Pooley, Michael Lee, William Shankle

Social Cognition: Alignment in Dialogue

Context, Syntactic Priming, and Referential Form in an Interactive Dialogue Task: Implications for Models of Alignment 109
Kathleen Carbary, Ellen Frohning, Michael Tanenhaus

Converging Hands or Converging Minds? 115
Lisette Mol, Emiel Krahmer, Alfons Maes, Marc Swerts

Vocal Interaction Dynamics of Children With and Without Autism 121
Anne S. Warlaumont, D. Kimbrough Oller, Rick Dale, Jeffrey A. Richards, Jill Gilkerson, Dongxin Xu

Virtually accommodating: Speech rate accommodation to a virtual interlocutor. 127
Laura Staum Casasanto, Kyle Jasmin, Daniel Casasanto

Education: Knowledge and Strategies

Connecting the Visible to the Invisible: Helping Middle School Students Understand Complex Ecosystem Processes. 133
Sameer Honwad, Cindy Hmelo-Silver, Rebecca Jordan, Catherine Eberbach, Steven Gray, Suparna Sinha, Ashok Goel, Swaroop Vattam, Spencer Rugaber, David Joyner

Response Times and Misconception-like Responses to Science Questions 139
Andrew Heckler, Thomas Scaife, Eleanor Sayre

Effects of Problem Context on Strategy Use within Functional Thinking 145
Katherine McEldoon, Caroline Cochrane-Braswell , Bethany Rittle-Johnson

The Application of the Less is More Hypothesis in Foreign Language Learning 150
Simone Chin, Alan Kersten

Decision Making: Sequential Choices

Are People Successful at Learning Sequential Decisions on a Perceptual Matching Task? 156
Reiko Yakushijin, Robert Jacobs

The Impact of Perceptual Aliasing on Exploration and Learning in a Dynamic Decision Making Task 162
Lisa Zaval, Todd Gureckis

Human foraging behavior: A virtual reality investigation on area restricted search in humans 168
Christopher Kalff, Thomas Hills, Jan Malte Wiener

Learning in Multiple-cue Judgment Tasks 174
Bettina von Helversen, Joerg Rieskamp

Symposium: Success in Theory of Mind

Success in Theory of Mind 180
Rose Scott, Adam Petrashek, Noah Goodman, Rebecca Saxe, Renee Baillargeon, Ori Friedman

Foundations: Representation

Disentangling Representation from Conceptualisation 182
Nancy Salay

Thought, Language and Mental Representation 188
Jonathan Trigg, Michael Kalish

What Is Domain Specificity (and Why Does It Matter)? 194
Muhammad Ali Khalidi

Cognition for action: an architectural account for "grounded interaction" 200
Anthony Harrison, J. Gregory Trafton

Language: Role of Syntax

The avoidance of ambiguity during conversation: More than mere priming or mimicry? 206
Jennifer M. Roche, Rick Dale, Roger J. Kreuz

Syntax Drives Phonological Choice - Even Independently of Word Choice 212
Marie Nilsenova, Marije van Amelsvoort

Corpus Evidence for Age Effects on Priming in Child Language 218
Jeffrey Gerard, Frank Keller, Themis Palpanas

Understanding acceptability judgments: Additivity and working memory effects 224
Laura Staum Casasanto, Philip Hofmeister, Ivan Sag

Induction: Facilitating Factors

The Role of Linguistic Labels in Categorization 230
Sophia Deng, Vladimir Sloutsky

The Price is Right: A High Information Access Cost Facilitates Category Learning 236
Michael Wood, Michael Fry, Mark Blair

Effects of generative and discriminative learning on use of category variability 242
Anne Hsu, Thomas Griffiths

Category Learning Through Active Sampling 248
Doug Markant, Todd Gureckis

Cognitive Architecture: Encoding & Retrieval

SARKAE - Modeling the Co-Evolution of Event Memory and Knowledge 254
Angela Nelson, Richard Shiffrin

Looking at Nothing Indicates Memory Search in Multiattribute Decision Making 260
Frank Renkewitz, Georg Jahn

Motor Simulation in a Memory Task: Evidence from Rock Climbing 266
Giovanni Pezzulo, Laura Barca, Alessandro Lamberti Bocconi, Anna M. Borghi

Social Cognition: Joint Attention

Enactive Social Cognition 272
Tobias Schlicht

Building a Model of Infant Social Interaction 278
Joshua Lewis, Gedeon Deák, Hector Jasso, Jochen Triesch

The Emergence of Referential Gaze and Perspective-taking in Infants 284
R. Joanne Jao, Marybel Robledo, Gedeon O. Deák

Joint perception: gaze and beliefs about social context 290
Daniel Richardson, Chris Street, Joanne Tan

Education: Meta-Cognitive Aspects

Teaching Students Self-Assessment and Task-Selection Skills with Video-Based Modeling Examples 296
Tamara van Gog, Danny Kostons, Fred Paas

Confidence without Competence in the Evaluation of Scientific Claims 302
Andrew Shtulman

Individual Differences as Predictors of Learning and Engagement 308
Sidney D'Mello, Claire Williams, Patrick Hays, Andrew Olney

Do Tutors' Content Knowledge and Beliefs About Learning Influence Their Assessment of Tutees' Understanding? 314
Stephanie Herppich, Jörg Wittwer, Matthias Nückles, Alexander Renkl

Decision Making: Risk

Risk attitude in decision making: A clash of three approaches 320
Eldad Yechiam, Eyal Ert

Making Assessments While Taking Sequential Risks 326
Avishai Wershbale, Timothy Pleskac

Testing Two Explanations for the Disjunction Effect in Prisoner's Dilemma Games: Complexity and Quasi-Magical Thinking 332
Evgenia Hristova, Maurice Grinberg

Mentalizing in games: A subtractive behavioral study of Prisoner's Dilemma 338
Antonio Napoli, Danilo Fum

Symposium: Prospective Perception

Prospective Perception 344
Jerome Scott Jordan, Jessica Witt, Michael Riley

Conceptual Spaces: Examples and Methods

Constructing Spatial Concepts from Universal Primitives 346
Yang Xu, Charles Kemp

Replicating Color Term Universals through Human Iterated Learning 352
Jing Xu, Thomas Griffiths, Mike Dowman

Similarity judgments reflect both language and cross-language tendencies: Evidence from two semantic domains 358
Naveen Khetarpal, Asifa Majid, Barbara Malt, Steven Sloman, Terry Regier

Language: Speech Perception

Adults' self-directed learning of an artificial lexicon: The dynamics of neighborhood reorganization 364
Neil Bardhan, Richard Aslin, Michael Tanenhaus

Perceptual Advantage from Generalized Linguistic Knowledge 369
Bozena Pajak

A rational account of perceptual compensation for coarticulation 375
Morgan Sonderegger, Alan Yu

Effects of Pragmatic Inference on Phoneme Identification 381
Hannah Rohde, Marc Ettlinger

Induction: Differential Factors

Individual Differences in Attention During Category Learning 387
Michael Lee, Ruud Wetzels

Categorisation, Deference and Cognitive Style 393
Nick Braisby, Sharon Hanlon

When Comparison Helps: The Role of Language, Prior Knowledge and Similarity in Categorizing Novel Objects 399
Clare Sims, Eliana Colunga

Category Learning and Adaptive Benefits of Aging 405
Angela Merritt, Linnea Karlsson, Edward Cokely

Cognitive Architecture: Higher-Order Structure

Encoding higher-order structure in visual working memory: A probabilistic model 411
Timothy Brady, Joshua Tenenbaum

Expertise in Visual Art is Associated with Altered Perceptual Strategies Within and Across Domains: Evidence from Eye Tracking 417
Kuba Glazek, Robert Weisberg

Melody Recognition: Effects of Articulation Format 423
Stephen Wee Hun Lim, Winston D. Goh

On the Relationship Between Entropy and Meaning in Music: An Exploration with Recurrent Neural Networks 429
Greg Cox

Social Cognition: Human-Robot Interaction

Simulating Cognitive Coping Strategies for Intelligent Support Agents 435
Azizi Ab Aziz, Jan Treur, Michel Klein

An Adaptive Integrative Ambient Agent Model to Intervene in the Dynamics of Beliefs and Emotions 441
Zulfiqar Ali Memon, Jan Treur

Proposing Artificial Subtle Expressions as an Intuitive Notification Methodology for Artificial Agents' Internal States 447
Takanori Komatsu, Seiji Yamada, Kazuki Kobayashi, Kotaro Funakoshi, Mikio Nakano

Experiments for Assessing Floating Reinstatement in Argument-based Reasoning 453
Iyad Rahwan, Jean-Francois Bonnefon, Mohammed Iqbal Madakkatel, Ruqiyabi Naz Awan, Sherief Abdallah

Education: Transfer

Optimizing Learning Environments: An Individual Difference Approach to Learning and Transfer 459
Daniel Belenky, Timothy Nokes

The Effects of Similarity and Individual Differences on Comparison and Transfer 465
Samuel Day, Robert L. Goldstone, Thomas Hills

Seeing Language Learning Inside the Math: Cognitive Analysis Yields Transfer 471
Ken Koedinger, Elizabeth McLaughlin

Initial Evidence of the Effects of Linguistic Framing on Transfer 477
Randi A. Engle, Adam Mendelson, Phi D. Nguyen

Decision Making: Non-Optimality

Impatience as Intertemporal Egoism 483
Daniel Bartels, Oleg Urminsky

WIN!! vs. win: Impact of "Outcome" Salience on Illusion of Control 489
Stefania Mereu, Alejandro Lleras

Signs of Non-Linearity in Base-Rate Neglect 495
Christopher Erb, Heidi Kloos

Mathematically Modeling Anchoring Effects 501
Jessica Choplin, Mark Tawney

Symposium: The Philosophy of Affective Neuroscience

The Philosophy of Affective Neuroscience 507
Rami Gabriel, Jaak Panksepp, Stephen Asma, Glennon Curran

Poster Session 1

Response Choice When Telling Lies 509
Emma Williams, Lewis Bott, Michael Lewis, John Patrick

Does Practice Narrow the Radius of Spatial Interference in Mental Images? 510
Don Lyon

A Sense of Order: Numerical ordering ability predicts complex mental arithmetic performance .. 511
Ian Lyons, Sian Beilock

Imaginary affordances shape children's preference judgments 512
Tania Henetz, Daniel Casasanto

Testing fMRI predictions of a Cognitive Model of the Problem State Multitasking Bottleneck 513
Jelmer Borst, Niels Taatgen, Hedderik Van Rijn, Andrea Stocco

Grounded Congruency Effects Between Vertical Meaning and Vertical Responding: Not Replicated 514
Lauren McDonough, Christine Wilson-Mendenhall, Lawrence Barsalou

MIReR: Media Integration Reflection Resource 515
Andreea Danielescu, Ellen Campana

Social Influences on the, um, Use of, uh, Fillers 516
Esther Walker, Evan Risko, Alan Kingstone

Probability estimation by mice in an interval timing task 517
Aaron Kheifets, C. Randy Gallistel

The Specificity of Non-Arbitrary Sound-to-Meaning Correspondences in Spoken Language 518
Christina Y. Tzeng, Lynne C. Nygaard, Laura L. Namy

Pattern Recognition Principle Theoretical Model of Mind-Brain Functioning 519
Gilberto de Paiva

Individual Differences in Explaining Noisy Data 520
Daniel R. Little, Richard M. Shiffrin

Making a good impression (formation model): a more complete account of processing 521
Tei Laine, Swati Gupta, Brian M. Monroe

The representation of idiom words in the mental lexicon 522
Simone Sprenger, Hedderik van Rijn

Assessing the Effectiveness of Wayfinding Directions 523
Alycia Hund, Amanda Padgitt

How Agent Placement Can Influence Perceived Boss/Co-worker Agreement in a Simulated Work Environment 524
Justin L. Matthews, Teenie Matlock

The structure of event representations: behavioral, imaging, and computational investigations ... 525
Anna Schapiro, Timothy Rogers, Matthew Botvinick

Pair Analysis and Joint Action Theory: A Research Protocol to Study Cognition and Interaction in Visual Analytics ... 526
Richard Arias Hernández, Linda T. Kaastra, Brian D. Fisher

The Importance of Visual Modeling in Children's Understanding of Physical Science ... 527
Nancy L Stein, Marc W. Hernandez

Reversing the side-effect effect: the 'Rational Scientist' explanation ... 528
Kevin Uttich, Tania Lombrozo

MHP/RT: Model Human Processor with Real Time Constraints ... 529
Makoto Toyota, Muneo Kitajima

CCE: Cognitive Chrono-Ethnography ... 530
Muneo Kitajima, Makoto Toyota

On the diversity of folk morality: Measuring classical positions in moral philosophy ... 531
Stephanie Müller, Bernd-Christian Otto, Edward Cokely

Linguistic Control in Monolingual and Bilingual Language Learners ... 532
James Bartolotti, Viorica Marian

Intent discerning agent for more intuitive visualizations ... 533
Tera Marie Green, Steve DiPaola

A Difference in Working Memory Capacity among Chinese Speakers Using Different Computer Word Typing Methods ... 534
Jenn-Yeu Chen, Cheng-Yi Li

Cognitive Arithmetic revisited: Effects of equation presentation format ... 535
Michael C. W. Yip

The Capacity to Discover: Working Memory and the Ability to Use Self-Explanation to Discover Early Algebra Concepts ... 536
Marci DeCaro, Bethany Rittle-Johnson

Number, Language, and Object Individuation ... 537
Lisa Cantrell, Linda B. Smith

The Cognitive and Motor Performance of Children with Functional Articulation Disorders ... 538
Rong-Ju Cherng, Hung-Yi Chen, Jenn-Yeu Chen, Yung-Jung Chen

I let the music speak: a model of music perception that predicts speech segmentation ... 539
Geraint Wiggins

Peer Reviewing in Undergraduate Psychology Students ... 540
Joanna Salapska-Gelleri

Training University Students on the Balance Scale Problem ... 541
Thomas Scaife, Andrew Heckler

Verb tense and aspect in scene descriptions in a humanoid robot ... 542
Carol J. Madden, Stéphane Lallée, Peter Ford Dominey

The importance of being present: The effect of a real or videotaped person on visual attention ... 543
Kaitlin Laidlaw, Tom Foulsham, Gustav Kuhn, Alan Kingstone

Exploring Phonological Levenshtein Distance Effects in Auditory Lexical Decision 544
Lidia Suárez, Seok Hui Tan, Melvin J. Yap, Winston D. Goh

How is children's exploratory play influenced by evidence conflicting with their theory? 545
Tessa J. P. van Schijndel, Maartje E. J. Raijmakers

Causal reasoning in decision making: A test of causal model theory of choice 546
Motoyuki Saito, Tsuneo Shimazaki

Gaze movement and language production when talking about events in live-recorded video clips .. 547
Monique Flecken, Christiane von Stutterheim, Mary Carroll

Fluency and cognitive control in judgment: Influences of memory and elaborative encoding 548
Paula Parpart, Edward T. Cokely

Bridging the Implementation Gap: From Sensorimotor Experience to Abstract Conceptual Knowledge 549
Anna Koop, Leah Hackman, Rich Sutton

An examination of learner control during web-based instruction 550
Jessica Federman, Ryan Morris, Lisa Dragoni

The Linguistic Distribution of Relational Categories 551
Micah Goldwater, Jon Willits

Multisensory stimuli improve numerical matching abilities of preschool children 552
Kerry Jordan, Joseph Baker, K.S. Rodzon

Wayfinding Tasks and Heuristics 553
Simon J. Buechner, Christoph Hölscher

Auditory distraction during semantic processing: Data and a model 554
Philip Beaman, John Marsh, Dylan Jones

How to Foster the Integration of Text and Diagrams: An Eye Tracking Study on the Use of Signals in Multimedia Learning 555
Katharina Scheiter, Alexander Eitel

Cognitive Modeling Repository 556
Jay Myung, Mark Pitt

The hindsight bias in temporal predictions of animated automobile accidents 557
Dustin Calvillo, Dayna Gomes

Strategies for multitasking: An fMRI study of individual differences in multitasking ability 558
Winston Jones, Jarrod Moss, Stephanie Doane

Facilitation in Second Language Word Meaning Evaluation from Masked Primes 559
Robert Zheng, Fernando Rubio, Dan Woltz

Physical design tools support and hinder innovative engineering design 560
Jooyoung Jang, Christian Schunn

Seductive Images and Metacomprehension of Science Texts 561
Allison Jaeger, Jennifer Wiley

Emotion and association-memory 562
Christopher Madan, Christine Lau, Jeremy Caplan, Esther Fujiwara

The Effects of Alcohol Use on Creative Problem Solving 563
Andrew Jarosz, Gregory Colflesh, Jennifer Wiley

Tapping into Student Knowledge about Science Systems 564
Jodi Davenport, Edys Quellmalz, Mike Timms

Preschoolers writing of multidigit numbers: From an additive to multiplicative representational system? 565
Sandra Street, Richard Prather, Cody Stitzel, Linda Smith, Kelly Mix

Knowledge about the role of illustrations on motivation for reading 566
Hideaki Shimada

Effects of Self-Explanation and Prompts Depend on the Students' Need for Cognition 567
Kyung Soo Do, Hyo-hee Lee, Hanna Kim

Engineering Models of Human Behavior 568
Spyridon Revithis

Attention for Action: Attentional Modulation by the Hands 569
Holger Schultheis, Laura Carlson, Richard Abrams

The effect of conventionality and aptness on suppression of metaphor-irrelevant meaning 570
Tomohiro Taira, Takashi Kusumi

A Categorization of Face Recognition Deficits in Congenital Prosopagnosia 571
Rainer Stollhoff, Jürgen Jost, Ingo Kennerknecht

Focusing on the Intermediate Event Makes the Chain Structure More Learnable 572
Kyung Soo Do, JaeHyuk Choi

Effects of Physical Structure and Creators' Intentions on Judgments of Function 573
Kyung Soo Do, Kyuhee Kim

A Computation Model synthesizing the Rule based and Experience based Cognitive Processes of Chinese Characters 574
Sau-chin Chen, Jon-Fan Hu, Ping Li

Virtual Brainstorming: Avatar Visibility and Group Size 575
Thomas Ward, Matthew Guerdat, Beverly Roskos-Ewoldsen

Semantic richness modulates early word processing within left-lateralized visual brain areas and enhances repetition priming 576
Milena Rabovsky, Werner Sommer, Rasha Abdel Rahman

The role of stimulus familiarity in non-linguistic sequence learning 577
Jennifer A. Sturm, Kenny Smith

The Development of Numeracy: Fingers Count! 578
Marcie Penner-Wilger, Lisa Fast, Jo-Anne LeFevre, Brenda L. Smith-Chant, Sheri-Lynn Skwarchuk, Deepthi Kamawar, Jeffrey Bisanz

Location, Location, Location: Environmental constraints on interpreting spatial terms 579
Kevin Mickey, Laura Carlson, Scott Freundschuh

Goals and the Perception of Distance and Time in Virtual Spaces 580
Angie Johnson, Kenny Coventry, Emine Mine Thompson

Artificial Cognitive Systems for Human-like Situation Awareness Ability 581
Soo-Young Lee

Comprehension and a Complex Task: A construction-integration study of individual performance in a non-routine task situation 582
Paul Ladny, Jordan McGuire, Randy J. Brou, Stephanie M. Doane

Competitive Routes to Belief and Their Impact on Future Learning 583
Carlos R. Salas, Thomas D. Griffin

Unique and Additive Effects of Self-Explaining and Contrasting Cases on Learning Fraction Division 584
Shanta Hattikudur, Pooja G. Sidney, Martha W. Alibali

When dog is more wolf than bone: Computational and electrophysiological evidence for featural organization of semantic memory 585
Sarah Laszlo, Blair Armstrong, Joseph MacInnes, David Plaut, Kara Federmeier

Hindsight bias in judgments of others' performance on inattentional blindness tasks 586
Alan Penaloza, Dustin P. Calvillo, Richard Brooks , Dayna M. Gomes

A decision science blind to decision procedures would be "unfair": The effect of decision process on decision-outcome satisfaction and subsequent choice in a performance environment 587
Daniel DeCaro, Joseph Johnson

A Functional, Hormonal, and Computational Study of Sex Differences in Working Memory 588
Brandon Abbs, Jill Goldstein

Now You See It, Now You Dont: Social Attention in a Magic Trick, Live and On Video 589
Robert Teszka, Evan Risko, Gustav Kuhn, Alan Kingstone

Declarative and procedural memory abilities as predictors of successful adult language learning .. 590
Katherine Brill, Mandy Faretta, Francis Wong, Patrick Wong, Kara Morgan-Short

A Bird's-Eye View of Numerical Discrimination in the Wild 591
Alexis Garland, Jason Low, K.C. Burns

Distinguishing first-line defaults and second-line conceptualization in reasoning about humans, robots, and computers 592
Daniel Levin, Megan Saylor, Simon Lynn

Cross-Modality Strategy Transfer: A behavioral study of strategic discrimination skill acquisition and transfer across auditory and visual modalities 593
Hao Bai, Paul Ladny, J. Gregory Trafton, Randy J. Brou, Stephanie M. Doane

"That's what she said": The effect of emotional prosody on the interpretation of intent. 594
Jennifer M. Roche, Rick Dale

Turn that frown upside down and to the left: Memory for faces is affected by their gravitational orientation 595
Nicolas Davidenko, Stephen Flusberg

Individual Differences in Successful Second Language Learning: The Roles of Working Memory and Intelligence 596
Brendan McCarthy, Mandy Faretta, Francis Wong, Patrick Wong, Ph.D., Kara Morgan-Short, Ph.D.

Individual differences in anticipatory eye-movements: Vocabulary size is associated with speed of noun-verb integration 597
Arielle Borovsky, Jeffrey Elman

Reasoning through Mindful Actions: Effect of Instruction and Spatial Ability on Understanding Dynamic Systems 598
Margaret Chan

Are children irrational category learners? Evidence from a process model 599
Gavin Jenkins, Jodi Smith, John Spencer, Larissa Samuelson

Brain Response Over Time to Structured and Unstructured Musical Sequences 600
Kat Agres, Hia Datta, Jason Zevin

Seeing the world through a visual language: Visual world paradigm in British Sign Language 601
Robin L. Thompson, David P. Vinson, Neil Fox, Gabriella Vigliocco

Insight into dynamics of speech perception in English and Japanese native speakers using a mouse-tracking paradigm 602
Hia Datta, Ran Liu, Jason Zevin

Concrete Models as Aids to Representational Translation of Molecular Diagrams 603
Andrew Stull, Hegarty Mary, Stieff Mike, Dixon Bonnie

I spy with your eye: On the perception of others' gaze 604
Nicola Anderson, Craig Anderson, Evan Risko, Alan Kingstone

Beyond binary: One small step across the artificial-naturalistic divide in understanding human category learning 605
Kimery Levering, Kenneth Kurtz

Dimension Word Knowledge and Flexible Attention Shifting 606
Rima Hanania, Thea Ionescu, Linda B. Smith

Lay Theories and Linguistic Framing in Teaching Children about Nutrition 607
Sarah Gripshover, Ellen Markman

The hindsight bias with dynamic stimuli and the propensity effect with static stimuli 608
Dayna Gomes, Dustin Calvillo

Category Learning in Second Life: Effects of Learning Context on Mechanisms of Categorization 609
Joshua Sturm, Peter Nachbaur, Alex Goldberg, Julianne Herts, Jan Andrews, Ken Livingston

Deciding Whether or Not to Guess the Answer Predicts Subsequent Learning 610
Sean Kang, Michael Mozer, Harold Pashler

Am I a Robot? How Verb Agency and Agent Description Influence Perspective-Taking in Visual Scenes 611
Michelle D. Greenwood, Teenie Matlock, Michael J. Spivey, Justin L. Matthews

The role of conventional number knowledge in young children's nonverbal number matching: Is "two" special? 612
Mee-Kyoung Kwon, Yoonkyung Jeong, Susan Levine

A sequence analysis of actions in complex system comprehension 613
Patrick Jeuniaux, Sebastien Tremblay, Jean-François Gagnon, Daniel Lafond, François Bernier

A valid separation of location memory based on allocentric and egocentric reference frames 614
Jonna Nilsson, Kenny Coventry, Nicol Ferrier

The relationship between similarities computed by LSA and several types of association 615
Keisuke Inohara, Takashi Kusumi

Reanalysis of Linda Problem 616
SangSuk Yoon, MinGyung Choi, HyunJung Shin

A computational model for the acquisition of referring subjects in discourse 617
Jacolien van Rij, Hedderik van Rijn, Petra Hendriks

Word-Form Typicality and Its Influence on Grammatical Category Assignment 618
Thomas Farmer, Padraic Monaghan, Jennifer Misyak, Morten Christiansen

Did you say "gross snails" or "gross nails?" The problem of segmenting co-occurrences of the same segment 619
Dahee Kim, Colin Widmer, Christine Szostak, Mark Pitt

Expectations of common ground with a computer dialog agent 620
Donna Byron, Joy Hanna, William Hartmann

Interaction of bottom-up and top-down attentional influences on the processing of contingency information 621
Kelly Goedert, Brianna Eiter

Influence on memory of the temporal schedule of repetitions over multiple days and its modulation by the retention interval 622
Emilie GERBIER, Olivier KOENIG

Matching Exact Posterior Probabilities in the Multinomial Interactive Activation Model 623
Pranav Khaitan, James L. McClelland

Threat and anxiety interactively impair task switching ability 624
Wolfgang Rauch, Marie Lauer-Schmaltz

Grammars for Funk Drumming: Symbolic and Motor-Spatial Aspects 625
Richard Ashley

About the Validity of Computer Models in Cognitive Science 626
Ricardo Sanz, Carlos Hernández, Jaime Gómez, Guadalupe Sánchez, Adolfo Hernando

The roles of working memory capacity and spatial ability in first-time solution of the Tower of Hanoi 627
Patrick Cushen, Jennifer Wiley

Causal Learning in Joint Activity: Comparing Collaborative, Active, and Passive Contexts 628
Andrew G. Young, Martha W. Alibali, Charles W. Kalish

What makes for inspirational examples in design? The effects of example modality, distance, and familiarity. 629
Joel Chan, Katherine Fu, Christian Schunn, Kristin Wood, Jonathan Cagan, Kenneth Kotovsky

False Recognition in the DRM-Paradigm reflects False Encoding 630
Tamella M. Pettitt, Eddy J. Davelaar

Monday is before Tuesday in speech, but left of Tuesday in gesture. 631
Daniel Casasanto, Kyle Jasmin

Enhanced visuo-spatial learning and memory effects in time-space synesthesia 632
Ursina Teuscher, David Brang, Vilayanur S. Ramachandran, Seana Coulson

Analyzing Discourse Functions in Student Research Reports to Assess Gains Due To Research Experiences 633
Roman Taraban, Brianna Bennett, Xiaofang Zeng

Embodying attentional states: The role of posture in task performance 634
Joseph Chisholm, Evan Risko, Alan Kingstone

Similarity avoidance in processing consonants: apparent exceptions from Polynesian languages .. 635
John Alderete

Cross-Situational Word Learning in Bilinguals 636
Viridiana Benitez, Linda B. Smith

Metacognition and Writing: How an Academically Gifted Adolescent Organizes and Controls the Writing Process 637
Delayne Connor

Immediate Introduction to Multiple Procedures Supports Procedural Flexibility in Equation Solving 638
Kelley Durkin, Bethany Rittle-Johnson, Jon Star

Promoting Cross-Disciplinary Communication in Nanotechnology 639
Sarah Kriz, Karen Cheng, Marco Rolandi, Yeechi Chen

Nouns are more stable than Verbs: Patterns of semantic change in 19th century English 640
Eyal Sagi

Context in distributed situated cognition 641
Hedda Rahel Schmidtke, Michael Beigl

Exploring Active Learning in a Bayesian Framework 642
Stephen Denton, John Kruschke

Linguistic Mediation of Visual Search: The effects of relative timing of speech and display 643
Eric Chiu, Michael Spivey

Belief bias in judgments of sample-size adequacy 644
Richard Anderson, Leisha Colyn, Beth Hartzler

Must analysis of meaning follow analysis of form? A time course analysis 645
Laurie Beth Feldman, Fermín Moscoso del Prado Martín, Patrick A O'Connor

Learning Cross-Modal Contingencies through Attentional Cues 646
Daniel Yurovsky, Rachel Wu, Natasha Kirkham, Chen Yu

The Effects of Alcohol on Working Memory and Change Detection 647
Gregory Colflesh, Andrew Jarosz, Jennifer Wiley

Effects of the Exploration Perspective on Pointing Accuracy 648
Julia Frankenstein, Manuel Vidal, Michael Rouillé, Stéphane Donikian, Mohamed Zaoui, Alain Berthoz

The Effect of Processing Type on Re-Categorization 649
David G. Cosejo, Stellan Ohlsson

Fractioning Factors that Influence Phonological Word Form Learning 650
Libo Zhao, Prahlad Gupta

Large differences in the distribution of instances of common object-based categories in early childhood 651
Alfredo F. Pereira, Karin H. James, Susan S. SJones, Linda B. Smith

Phonetic symbolism for size and shape 652
Patrick Thompson, Zachary Estes

Using Embodied Cognition in the Instruction of Abstract Programming Concepts 653
Cameron L. Fadjo, John B. Black, JeeHye Hong, Chun-Hao Chang

Decision-Making in Older Adults: Sometimes Older is Wiser 654
Darrell Worthy, W. Todd Maddox

Coordination dynamics in speech and lexical semantics 655
Christopher Kello, Theo Rhodes, Geoff Hollis, Bryan Kerster

Knowing who knows what best: Preschoolers selectively use others' past accuracy in causal learning 656
Chris Vredenburgh, Lauren Schneider, Andy Hsia , Tamar Kushnir

The effects of perspective on understanding of projective spatial terms 657
Takatsugu Kojima

Inferring Object Structure from Human Action at 9 Months 658
Stephen Killingsworth, John Jacobson, Megan Saylor

Social Indexing: How the People Around Us Aid Cognition 659
Chris N.H. Street, Daniel. C Richardson

Impact of Diverse Abilities on Learning to Write through Peer-Review 660
Melissa Patchan, Christian Schunn

Look who's talking (and follow the leader)! Eye movements in a social interaction reveal effects of speaking and social status 661
Tom Foulsham, Joey Cheng, Jessica Tracy, Joseph Henrich, Alan Kingstone

Using analogical learning in science curricula to improve conceptual understanding 662
J. Elizabeth Richey, Alicia Chang, Timothy J. Nokes, Christian D. Schunn

Modeling age of exposure in L2 learning of vowel categories 663
Meghan Clayards, Joseph Toscano

Mental models of virology in experts and novices 664
Benjamin Jee, David Uttal, Caroline Crouch, Amy Spiegel, Judy Diamond

Facilitating Educator Evaluation of Online Instructional Materials: Does Conceptual Browsing Impact Cognitive Processing? 665
Kirsten Butcher, Robert Zheng, Anne Cook, Lisa Ferrara, Sarah Davies, Ashley Crockett Mazal, Aaron Dewald

Are Hindu-Arabic Numerals Concrete or Abstract Symbols? 666
Percival Matthews

Order Effects in Categorization: Identifying "the Nuts" in Poker 667
Brian D. Gane, Richard Catrambone

Semantics in the wild: Context-sensitive inferences about mammals 668
Jeremy Glick, James McClelland

Judgements of relative order: Mechanisms underlying subspan versus supraspan lists 669
Yang Liu, Michelle Chan, Jeremy Caplan

Verb-body part associations across users of English, Telugu, and Hindi 670
Raju Bapi, Jigar Patel, Viswanath Naidu, Sireesha Jala, Vasanta Duggirala, Suvarna Alladi

Reasoning in pedagogical versus deceptive situations 671
Russell Warner, Todd Stoess, Patrick Shafto

Interactions Between the Fast and Slow Mental Processes 672
William Kennedy, Magdalena Bugajska

Word Length Effects and the Serial vs. Parallel Debate in Connectionist Models of Reading Aloud 673
Alan H. Kawamoto

Differential effects of dopamine dysfunction on context usage in people with autism and schizophrenia: A computational exploration 674
Trent Kriete, David C. Noelle

Can Statistics Change our Minds? The Role of Causal Explanation in Accommodation of Base Rate Statistics 675
Edward Munnich, Saera Khan, Melissa Latham, Michelle Brewer, Valesia Ho, Sierra Walton

Handedness and Hand Used Differentially Affect Object Facing 676
Jyotsna Vaid, Hsin-Chin Chen, Rebecca Rhodes, Sumeyra Tosun

Experience word learning predicts children's ability to generalize novel labels 677
Emily Thom, Catherine Sandhofer

A Biologically Plausible Account of the Computational Utility of Consciousness 678
William B. St. Clair, David C. Noelle

Perceiving the Other during Joint Action 679
Jerome Scott Jordan, Andrew Kenning, James Clinton, Justin Durtschi, J. Cooper Cutting

A Model of Cognitive Rehabilitation: Recovering with Constraints 680
Shin-ichi Asakawa, Yoshihiro Itaguchi

Restructuring representations in analogy making by children: the role of cognitive flexibility 681
Jean-Pierre Thibaut, Robert French, Yannick Gerard

White- and Grey-Matter Damage Differentially Impair Learning and Generalization in a Computational Model of the Raven Matrices Task 682
Vincent G. Berthiaume, Thomas R. Shultz, Olaf Dammann

Neural networks for word recognition: Is a hidden layer necessary? 688
Frederic Dandurand, Thomas Hannagan, Jonathan Grainger

The dimensionality of episodic images 694
Vishnu Sreekumar, Yuwen Zhuang, Simon Dennis, Mikhail Belkin

Distributional Analyses in Visual Lexical Decision: Orthographic Neighborhood Density and Word Frequency Effects 700
Stephen Wee Hun Lim, Melvin J. Yap

Linguistic and Non-Linguistic Influences on Learning Biases for Vowel Harmony 706
Sara Finley, William Badecker

Structural Constraints and Real-World Plausibility in Analogical Inference 712
Linsey Smith, Dedre Gentner

Cross-Modal Influence on Binocular Rivalry 718
Joshua Lewis, Adam Fouse, Virginia de Sa

The Necessity of Ordinary Experience 724
Robin Flanagan

A Cross-linguistic Model of the Acquisition of Inflectional Morphology in English and Modern Greek 730
Themis Karaminis, Michael Thomas

Handling what the other sees: the effects of seeing and being seen on gesture production 736
Lisette Mol, Emiel Krahmer

On Attractiveness of Surprising Ideas: How Memory for Counterintuitive Ideas Drives Cultural Dynamics 742
M. Afzal Upal

Consciousness is Data Compression 748
Phil Maguire, Rebecca Maguire

Feature repetition effects on object familiarity: Evidence from an old/new recognition task 754
Selda Eren, Annette Hohenberger

Beyond Transitional Probabilities: Human Learners Impose a Parsimony Bias in Statistical Word Segmentation 760
Michael Frank, Harry Tily, Inbal Arnon, Sharon Goldwater

Effects of Goal Specificity on a Search in a Hypothesis Space and an Instance Space 766
Miki Matsumuro, Kazuhisa Miwa

Speaker's choice of frame based on rarity information 772
Hidehito Honda, Toshihiko Matsuka

Discovering Structure by Learning Sparse Graphs 778
Brenden Lake, Joshua Tenenbaum

Using the Social of Tagging: The Interplay of Social Tags and the Strength of Association in Navigation and Learning Processes 784
Christoph Held, Ulrike Cress

Parallel Processing During Spoken Language Comprehension in Bimodal Bilinguals 790
Anthony Shook, Viorica Marian

Conditions of Directed Attention Inhibit Recognition Performance for Target-Aligned Stimuli 796
Andrew Dewald, Scott Sinnett, Leonidas Domas

Information Selection in the Blogosphere: The Effect of Expertise, Community Rating, and Age . 802
Stephan Winter, Nicole C. Krämer, Jana Appel, Kathrin Schielke

The Impact of Syntax on the Interpretation and Graphical Depiction of Underspecified Propositions 808
Aaron Kalb, Dave Barker-Plummer, Richard Cox, Robert Dale, Deonne Castaneda, Christopher Potts

Development of Prototype Abstraction and Exemplar Memorization 814
Irina Beatu, Thomas Shultz

Person, place, and past influence eye movements during visual search 820
Barbara Hidalgo-Sotelo, Aude Oliva

Perspectivizing Space in Bānlā Discourse 826
Samir Karmakar, Rajesh Kasturirangan

Egocentric and allocentric spatial references in children with Cerebral Palsy 831
Laura Barca, Giovanni Pezzulo , Enrico Castelli

Vocabulary Spurt; Are Infants full of Zipf? 836
Julien Mayor, Kim Plunkett

Context and Category Information in Children and Adults 842
Adam Osth, Simon Dennis, Vladimir Sloutsky

Visual Similarity is ObViS 848
Michel Brudzinski, Chris Sims, Wayne Gray, Michael Schoelles

Learning Structured Preferences 853
Leon Bergen, Owain Evans, Joshua Tenenbaum

Conceptual Spaces: Processes

Beyond Boolean logic: exploring representation languages for learning complex concepts 859
Steven Piantadosi, Joshua Tenenbaum, Noah Goodman

Encoding Sequential Information in Vector Space Models of Semantics: Comparing Holographic Reduced Representation and Random Permutation 865
Gabriel Recchia, Michael Jones, Magnus Sahlgren, Pentti Kanerva

Learning and Generalization of Abstract Semantic Relations: Preliminary Investigation of Bayesian Approaches 871
Dawn Chen, Hongjing Lu, Keith Holyoak

You Can't Wear a Coat Rack: A Binding Framework to Avoid Illusory Feature Migrations in Perceptually Grounded Semantic Models 877
Michael Jones, Gabriel Recchia

Language: Impact on Imagery

The evocative power of words: Activation of visual information by verbal and nonverbal means .. 883
Gary Lupyan, Sharon Thompson-Schill

Perceptual Simulations of Temporal Uses of In and On in First and Second Language Processing 889
Luca Onnis, Daniel Jackson, Michael Spivey

A Motion Aftereffect from Literal and Metaphorical Motion Language: Individual Differences ... 895
Alexia Toskos Dils, Lera Boroditsky

Language-Driven Motor Simulation is Sensitive to Social Context 901
Heeyeon Y. Dennison, Benjamin K. Bergen

Causal Inference: Statistical Assumptions

Connecting Causal Events: Learning Causal Structures Through Repeated Interventions Over Time 907
Benjamin Rottman, Frank Keil

The induction of hidden causes: Causal mediation and violations of independent causal influence 913
Christopher Carroll, Patricia Cheng

Edge replacement and nonindependence in causation 919
David Buchanan, Joshua Tenenbaum, David Sobel

Agents and Causes: A Bayesian Error Attribution Model of Causal Reasoning 925
Ralf Mayrhofer, York Hagmayer, Michael Waldmann

Cognitive Architecture: Cost & Control

Modeling strategies in Stroop with a general architecture of executive control 931
Tomasz Smole", Adam Chuderski

Modelling the Correlation Between Two Putative Inhibition Tasks: An Analytic Approach 937
Eddy Davelaar, Richard Cooper

Estimation of Trade-off between Costs of Preprocessing and Primary Processing 943
Akihiro Maehigashi, Kazuhisa Miwa

Increasing Information Access Cost to Protect Against Interruption Effects during Problem Solving 949
Phillip Morgan, John Patrick, Tanya Patrick

Social Cognition: Networks

An Agent-based Simulation of the Effectiveness of Creative Leadership 955
Stefan Leijnen, Liane Gabora

Linguistic cues predict fraudulent events in a corporate social network 961
Max Louwerse, David Lin, Amanda Drescher, Gun Semin

The Cognitive Cost of Ethnocentrism 967
Artem Kaznatcheev

Self-esteem and the Matching Effect in Mate Selection 972
Artem Kaznatcheev, Kyler Brown, Thomas Shultz

Communication: Complex Inferences

Determining the Internal Consistency of Attitude Attributions 978
Kyle Jennings

Cohesion, Coherence, and Expert Evaluations of Writing Proficiency 984
Scott Crossley, Danielle McNamara

An acquired taste: How reading literature affects sensitivity to word distributions when judging literary texts 990
Justine Kao, Robert Ryan, Melody Dye, Michael Ramscar

What You Did and Didn't Mean: Noise, Context, and Human Skill 996
Tiziana Ligorio, Susan L. Epstein, Rebecca J. Passonneau, Joshua B. Gordon

Thinking: Representations and Processes

Spatial Reasoning as Verbal Reasoning 1002
Antje Krumnack, Leandra Bucher , Jelica Nejasmic, Markus Knauff

Arbitrating Between Theory-Theory and Simulation Theory: Evidence from a Think-Aloud Study of Counterfactual Reasoning 1008
Meredith Wilkinson, Linden Ball, Rachel Cooper

Less-is-more effects in knowledge-based heuristic inference 1014
Philip Beaman, Philip Smith, Rachel McCloy

What Makes a Good Reasoner?: Brain Potentials and Heuristic Bias Susceptibility 1020
Wim De Neys, Nikolay Novitskiy , Jennifer Ramautar, Johan Wagemans

Symposium: The Mechanics of Embodiment

The Mechanics of Embodiment 1026
McRae Ken, Fischer Martin

Semantics: Intended Meaning

On the Notion of Intended Meaning 1028
Marco Cruciani

Predicative Metaphor Comprehension as Indirect Categorization 1034
Akira Utsumi, Maki Sakamoto

Abstract and Belief-Based Language Differentiate Joking, Pretending, and Literal Toddler-Directed Speech 1040
Elena Hoicka, Ruth Campbell

Wrongness and Representational Thought 1046
Elena Hoicka, Merideth Gattis

Language: Impact on Vision

Spatial Position in Language and Visual Memory: A Cross-Linguistic Comparison 1052
Solveig Bosse, Anna Papafragou

Generalized Event Knowledge Activation During Online Language Comprehension 1058
Ross Metusalem, Marta Kutas, Mary Hare, Ken McRae, Jeffrey L. Elman

Framed: Factors influencing reference frame choice in tabletop space 1064
Laurie Robinette, Michele Feist, Michael Kalish

Sentence Production in Naturalistic Scenes with Referential Ambiguity 1070
Moreno I. Coco, Frank Keller

Causal Inference: Context Effects

Uncertainty in causal inference: The case of retrospective revaluation 1076
Christopher Carroll, Patricia Cheng, Hongjing Lu

The Role of Causal Schemas in Inductive Reasoning 1082
Ralf Mayrhofer, Jonas Nagel, Michael Waldmann

Causal Conditional Reasoning and Conditional Likelihood 1088
Philip Fernbach, Adam Darlow

Causal Models Interact with Structure Mapping to Guide Analogical Inference 1094
Hee Seung Lee, Keith Holyoak

Cognitive Architecture: Brain-Like Computation

Symbolic Reasoning in Spiking Neurons: A Model of the Cortex/Basal Ganglia/Thalamus Loop 1100
Terrence Stewart, Xuan Choo, Chris Eliasmith

A Hubel Weisel model for hierarchical representation of concepts in textual documents 1106
Kiruthika Ramanathan, Luping Shi, Tow Chong Chong

Automatic and Controlled Processes in Semantic Priming: an Attractor Neural Network Model with Latching Dynamics 1112
Itamar Lerner, Shlomo Bentin, Oren Shriki

Social Cognition: Wisdom of Crowds

Cognitive Models and the Wisdom of Crowds: A Case Study Using the Bandit Problem 1118
Shunan Zhang, Michael Lee

The Accuracy of Small-Group Estimation and the Wisdom of Crowds 1124
Michael Lee, Jenny Shi

The Wisdom of Crowds with Informative Priors 1130
Pernille Hemmer, Mark Steyvers, Brent Miller

Reading: Words and Eyes

The Emergence of Adaptive Eye Movements in Reading 1136
Yanping Liu, Erik Reichle

Rational eye movements in reading combining uncertainty about previous words with contextual probability 1142
Klinton Bicknell, Roger Levy

A New Perspective on Visual Word Processing Efficiency 1148
Joseph Houpt, James Townsend

The online processing of modal verbs: Parallel activation of competing mental models 1154
Stephanie Huette, Teenie Matlock, Michael Spivey

Thinking: Sources of Error

INFLUENCE OF GRAMMATICAL GENDER ON DEDUCTIVE REASONING ABOUT SEX-SPECIFIC PROPERTIES OF ANIMALS 1160
Mutsumi Imai, Lennart Schalk, Henrik Saalbach, Hiroyuki Okada

A Comparison of the Belief-Adjustment Model and the Quantum Inference Model as Explanations of Order Effects in Human Inference 1166
Jennifer Trueblood, Jerome Busemeyer

Information Relevance in Pseudodiagnostic Reasoning 1172
Frederic Vallee-Tourangeau, Gaelle Villejoubert

Accessing the Unsaid: The Role of Scalar Alternatives in Children's Pragmatic Inference 1178
Neon Brooks, Alan Bale, David Barner

Symposium: Developmental and Computational Perspectives on Infant Social Cognition

Developmental and computational perspectives on infant social cognition 1184
Noah Goodman, Chris Baker, Joshua Tenenbaum, Chris Lucas, Kiley Hamlin, Tamar Kushnir, Tomer Ullman, Elizabeth Spelke

Semantics: Reference - 1

Linking meaning to language: linguistic universals and variation 1186
Joshua Hartshorne, Tim O'Donnell, Yasutada Sudo, Miki Uruwashi, Jesse Snedeker

Comprehending Negated Sentences With Binary States and Locations 1192
Sarah Anderson, Stephanie Huette, Teenie Matlock, Michael Spivey

On-line Interactions of Context and Grammatical Aspect 1198
Sarah Anderson, Teenie Matlock, Michael Spivey

Anaphors and Local Coherences 1204
Lars Konieczny, Helmut Weldle, Sascha Wolfer, Daniel Müller, Peter Baumann

Vision: Linguistic and Social Factors

Self-directed speech alters visual processing 1210
Gary Lupyan, Daniel Swingley

Is categorical perception really verbally mediated perception? 1216
Andrew Hendrickson, George Kachergis, Todd Gureckis, Robert Goldstone

Visual Similarity Effects in Categorical Search 1222
Robert Alexander, Wei Zhang, Gregory Zelinsky

Social Cues Support Learning about Objects from Statistics in Infancy 1228
Rachel Wu, Alison Gopnik, Daniel Richardson, Natasha Kirkham

Mental Representation of Number

Electrophysiological Evidence for Multiple Representations of Number in the Human Brain 1234
Frank Kanayet, John Opfer, William Cunningham

The perception of number from long-term memory 1240
Jiaying Zhao, Nicholas Turk-Browne

Mapping number words to approximate magnitudes: associative learning or structure mapping? 1246
Jess Sullivan, David Barner

Analogue Magnitudes and Knower-Levels: Re-Visiting the Variability Argument 1252
James Negen, Barbara Sarnecka

Subsymbolic Learning: Reinfoecement

Integrating Reinforcement Learning with Models of Representation Learning 1258
Matt Jones, Fabian Cañas

Attention and Reinforcement Learning: Constructing Representations from Indirect Feedback .. 1264
Fabián Cañas, Matt Jones

Learning to selectively attend 1270
Samuel Gershman, Jonathan Cohen, Yael Niv

Pavlovian conditioning from a foraging perspective 1276
James Anderson, Chloe Bracis, Andrew Goodwin

Social Cognition: Cultural Factors

A cognitive model of punishment 1282
Francesca Giardini, Giulia Andrighetto, Rosaria Conte

Culturally-Guided Beliefs about Opposing Categories and Their Consequences for Action: The Case of Cooperation and Competition 1289
Josh Keller, Jeffrey Loewenstein, Jin Yan

Theories of God: Explanatory Coherence in a Non-Scientific Domain 1295
Andrew Shtulman

The Cultural Transmission of Explanations: Evidence that Teleological Explanations are Preferentially Remembered 1301
Nicholas Gwynne, Tania Lombrozo

Reading: Words and Brains

Semantic integration of novel word meanings after a single exposure in context 1307
Arielle Borovsky, Jeffrey Elman, Marta Kutas

Fixation durations in first-pass reading reflect uncertainty about word identity 1313
Nathaniel Smith, Roger Levy

An fMRI Study of Strategic Reading Comprehension 1319
Jarrod Moss, Christian Schunn, Walter Schneider, Danielle McNamara, Kurt VanLehn

Gricean Brainwaves: Brain Responses to Pragmatic Violations in Dialogues 1325
John Hoeks, Petra Hendriks, Gisela Redeker, Laurie Stowe

Thinking: Impact of Language

Can grammar influence voting? 1330
Caitlin Fausey, Teenie Matlock

How much for a transitive? Subtle linguistic cues influence blame and punishment 1336
Caitlin Fausey, Lera Boroditsky

Can mirror-reading reverse the flow of time? 1342
Daniel Casasanto, Roberto Bottini

Implicit spatial length modulates time estimates, but not vice versa. 1348
Roberto Bottini, Daniel Casasanto

Symposium: Emerging Insights from Eye-Movement Research on Category Leaning

Emerging Insights from Eye-Movement Research on Category Learning 1354
Bob Rehder, Mark Blair, Aaron Hoffman, Marcus Watson

Semantics: Reference - 2

Comparing Apples to Fruit: Parent's Comparisons of Labels are Related to First and Second Label Learning 1356
Chandra Brojde, Eliana Colunga

Thinking With Your Body: Modelling Spatial Biases in Categorization Using a Real Humanoid Robot 1362
Anthony Morse, Tony Belpaeme, Angelo Cangelosi, Linda Smith

Effects of simultaneously presented visual information on adults' and infants' auditory statistical learning 1368
Erik Thiessen

Gesture in language: How sound symbolic words are processed in the brain 1374
Mamiko Arata, Mutsumi Imai, Jiro Okuda, Hiroyuki Okada, Tetsuya Matsuda

Word Order, Case Forms and Structural Priming in Czech Children's Comprehension 1380
Filip Smolík, Jií Lukavský

Subsymbolic Learning: Complex Patterns

Modeling Implicit and Explicit Processes in Recursive Sequence Structure Learning 1381
Jamie Alexandre

The Impact of Starting Small on the Learnability of Recursion 1387
Jun Lai, Fenna Poletiek

Dissociating Sources of Knowledge in Artificial Grammar Learning 1393
MIchelle Hendricks, Christopher Conway, Ronald Kellogg

Why Streaks Are Special: The Time of Patterns 1399
Yanlong Sun, Hongbin Wang

Social Cognition: Context Facilitation

Social Learning and Cumulative Mutual Improvement in a Networked Group 1405
Thomas Wisdom, Robert Goldstone

Social Context Effects on the Impact of Category Labels 1411
Rachel Stephens, Amy Perfors, Daniel Navarro

Pedagogical Cues Influence Children's Inductive Inference and Exploratory Play 1417
Lucas Butler, Ellen Markman

The Facilitative Effect of Context on Second-Order Social Reasoning 1423
Ben Meijering, Leendert Van Maanen, Hedderik Van Rijn, Rineke Verbrugge

Reading: Phonological Coding

Phonological instability in young adult poor readers 1429
James Magnuson, Anuenue Kukona, David Braze, Clint Johns, Julie Van Dyke, Whiteny Tabor, Kenneth Pugh, Einar Mencl

Phonological Encoding in Word Naming and Word Typing 1435
Jenn-Yeu Chen, Cheng-Yi Li

Visual and Task characteristics may explain hemispheric asymmetry in visual word recognition 1441
Kloser Chee Fung Cheung, Janet Hui-wen Hsiao

Effects of Near and Distant Phonological Neighbors on Picture Naming 1447
Daniel Mirman, Audrey K. Kittredge, Gary S. Dell

Thinking: Impact of Emotion

How Does Anxiety Influence Analogical Mapping? 1453
Veselina Feldman, Penka Hristova, Boicho Kokinov

Impact of mood induction on temporal processing 1459
Katrina Rodzon, Kerry Jordan

Hot Cognitions in Coherence-Based Reasoning and Decision-Making 1465
Stephen Read, Dan Simon, Douglas Stenstrom

Comparison-Induced Sequence Effects on Hedonic Evaluations 1471
Jessica Choplin, Megan Lombardi

Poster Session 2

The Effect of Cognitive Load and Meaning on Selective Attention 1477
Rebecca Weast, Nicole Neiman

Is perceptual acuity asymmetric in isolated word recognition? Evidence from an ideal-observer reverse-engineering approach 1483
Nathaniel Smith, Wen-Hsuan Chan, Roger Levy

Are grunters cheaters? The effects of grunting when judging the direction of a tennis shot 1489
Scott Sinnett, Alan Kingstone

Children's Inductive Inference with Synonymous Labels 1493
Bryan Matlen, Anna Fisher, Karrie Godwin

Development of Relational Reasoning with Semantically Similar Labels 1499
Sheela Ramesh, Bryan Matlen, Anna Fisher

Negative Transfer in Matchstick Arithmetic Insight Problems 1505
Trina Kershaw, Jason Braasch, Christopher Flynn

A critique of multi-voxel pattern analysis 1511
Michael Anderson, Tim Oates

The Origins of Collective Overvaluation: Irrational exuberance emerges from simple, honest and rational individual behavior 1517
Michael Anderson, C. Athena Aktipis

A category theory explanation for systematicity 1523
Steven Phillips, William Wilson

The Effect of Word-internal Properties on Syntactic Categorization: A Computational Modeling Approach 1529
Fatemeh Torabi Asr, Afsaneh Fazly, Zohreh Azimifar

Multiple-choice testing can improve the retention of nontested related information 1535
Jeri Little, Elizabeth Ligon Bjork

Holographic stimulus representation and judgement of grammaticality in an exemplar model: Combining item and serial-order information 1541
Randall K. Jamieson, D. J. K. Mewhort

A Cross-Cultural Study of Change Blindness in Turkish and American Students 1547
Treysi Terziyan, Joan Gilkey

Analogical Mapping Through Visual Abstraction 1553
Jim Davies, Patrick Yaner

A Bottom-Up Parsing Model of Local Coherence Effects 1559
Emily Morgan, Frank Keller, Mark Steedman

Heuristics for Choosing Features to Represent Stimuli 1565
Mathew Zeigenfuse, Michael Lee

When Two Plus Two Does Not Equal Four: Event-Related Potential Responses to Semantically Incongrous Arithmetic Word Problems 1571
Kristie Fisher, Miriam Bassok, Lee Osterhout

Simplifying the Mapping from Referring Expression to Referent in a Conceptual Semantics of Reference 1577
Jerry Ball

Toward a Functional Model of Human Language Processing 1583
Jerry Ball, Mary Freiman, Stuart Rodgers, Christopher Myers

Linking Learning to Looking: Habituation and Association in Infant Statistical Language Learning 1589
Daniel Yurovsky, Shohei Hidaka, Chen Yu, Linda Smith

Simultaneous Cross-situational Learning of Category and Object Names 1595
Tarun Gangwani, George Kachergis, Chen Yu

Extending Beyond Space 1601
Brooke Breaux, Michele Feist

A computational model of cognitive interference without neural inhibitory mechanisms 1607
Serge Thill, Robert Lowe

Phonetic training makes word learning easier 1613
Amy Perfors, David Dunbar

The Influence of Within-Category Structure on Stimulus Similarity and Stimulus Generalization 1619
James Close, Ulrike Hahn, Robert Honey

Conservatism in Belief Revision and Participant Skepticism 1625
Adam Corner, Adam Harris, Ulrike Hahn

The Game Lies in the Eye of the Beholder: The Influence of Expertise on Watching Soccer 1631
Michael Smuc, Eva Mayr, Florian Windhager

When Robot Gaze Helps Human Listeners: Attentional versus Intentional Account 1637
Maria Staudte, Matthew W. Crocker

How Action Understanding can be Rational, Bayesian and Tractable 1643
Mark Blokpoel, Johan Kwisthout, Theo P. van der Weide, Iris van Rooij

Facilitating Low-Achieving Students' Diagram Use in Algebraic Story Problems 1649
Julie Booth, Kenneth Koedinger

Why do four- year- olds show poor cross-modal transfer between haptic and vision? 1655
Hilary Kalagher, Susan Jones

Collaborative Sensemaking in the Blogosphere 1661
Richard Alterman, Johann Larusson

Inflectional Suffix Priming in Czech Verbs and Nouns 1667
Filip Smolík

Restructuring Causal Concepts 1673
Eric Taylor

Analysis of the Variability of Three-Dimensional Spatial Relations in Visual Short-Term Memory 1679
Carsten Winkelholz, Michael Kleiber, Christopher Marc Schlick

Spatial Factors in Social and Asocial Learning 1685
Alex Metz, Thomas Shultz

Taking a Look (Literally!) at the Raven's Intelligence Test: Two Visual Solution Strategies 1691
Maithilee Kunda, Keith McGreggor, Ashok Goel

The Dice are Cast: The Role of Intended versus Actual Contributions in Responsibility Attribution 1697
Tobias Gerstenberg, David A. Lagnado, Yaakov Kareev

Learning perceptual aspects of diagnosis in medicine via eye movement modeling examples on patient video cases 1703
Halszka Jarodzka, Thomas Balslev, Kenneth Holmqvist, Marcus Nyström, Katharina Scheiter, Peter Gerjets, Berit Eika

Structure Awareness in Action-Outcome Learning Eradicates the Detrimental Effect of Reinforcement Delays 1709
William Greville, Adam Cassar, Mark Johansen, Marc Buehner

More than One Kind of Probability Matching: Evidence from a Dual-Task Paradigm 1715
A. Ross Otto, Arthur Markman, Eric Taylor

Probabilistic language acquisition: Theoretical, computational, and experimental analysis 1720
Anne Hsu, NIck Chater

Learning concepts from sketches via analogical generalization and near-misses 1726
Matthew D. McLure, Scott E. Friedman, Kenneth D. Forbus

Inferring Multitasking Breakpoints from Single-Task Data 1732
Peter Bogunovich, Dario Salvucci

Comparing Human-Human to Human-Computer Tutorial 1738
Natalie Steinhauser, Gwendolyn Campbell, Katherine Harrison, Leanne Taylor, Myroslava Dzikovska, Johanna Moore

Perceptually Grounded Word Meaning Acquisition: A Computational Model 1744
Claudius Gläser, Frank Joublin

A Motivationally Based Computational Interpretation of Social Anxiety Induced Stereotype Bias 1750
Nicholas Wilson, Ron Sun, Robert Mathews

Cultural Network Analysis: Mapping Cultural Theories of Mind 1756
Winston Sieck, Louise Rasmussen

Some Attention Learning "Biases" in Adaptive Network Models of Categorization 1762
Toshihiko Matsuka, James E. Corter, Arther B. Markman

Investigating Insight Using Compound Remote Associate Problems 1768
Edward Cranford, Jarrod Moss

Writing: The Process of Discovery 1774
Veerle Baaijen, David Galbraith, Kees de Glopper

Number Representations and their Development: A Connectionist Model of Number Comparison 1780
Mark Lewis, Sashank Varma

Causal stream location effects in preschoolers 1786
David Buchanan, David Sobel

In What Sense is P(A—B) P(B) = P(A,B)? The Relationship between Distributional Format and Subjective Probability Estimates 1792
Belinda Bruza, Matthew Welsh, Daniel Navarro, Stephen Begg

Of parrots and parsimony: reconsidering Morgan's canon 1798
Matthew Welsh

Socially induced motor plasticity affects language comprehension 1804
David Havas, Julia Jenvey, Hayley Shilling, Mitchell Nathan

Considering the Source: Preschoolers (and Adults) Use Talker Acoustics Predictively and Flexibly in On-Line Sentence Processing 1810
Sarah Creel

Hebbian learning for deciding optimally among many alternatives (almost) 1816
Patrick Simen, Tyler McMillen, Sam Behseta

Contributions of Prosodic and Distributional Features of Caregivers' Speech in Early Word Learning 1822
Soroush Vosoughi, Brandon Roy, Michael Frank, Deb Roy

Concreteness and Relational Matching in Preschoolers 1828
Jennifer Kaminski, Vladimir Sloutsky

How does the presence of a label affect attention to other features? 1834
Amy Perfors, Daniel Navarro

Wisdom of the Crowds in Minimum Spanning Tree Problems 1840
Sheng Kung Yi, Mark Steyvers, Michael Lee, Matthew Dry

The Effect of Labels on Visual Attention: An Eye Tracking Study 1846
Catherine Best, Christopher Robinson, Vladimir Sloutsky

Mental Representations of Diagrams, Views about Diagrams, and Problem Solving 1852
Emmanuel Manalo, Yuri Uesaka, Yoshio Yano

Metacognitive Judgments of Improvement are Uncorrelated with Learning Rate. 1858
Corinne Townsend, Evan Heit

Learning to Explore the World through its Statistics: Infants' Visual Search in the A-not-B task 1863
Hanna Popick, Michael Ramscar, Natasha Kirkham

The Impact of Collective Opinion on Online Judgment 1869
Yasuaki Sakamoto

Enhancing Acquisition of Intuition versus Planning in Problem Solving 1875
Dawn Chen, Keith Holyoak

Infants expect others to help one another achieve a goal 1881
Woo-yeol Lee, Eun Young Kim, Jeong-ae Won, Yoonha Lee, Yoon Kim

Priming Effects on Event Types Classication: Effects of Word and Picture Stimuli 1886
Alessandra Zarcone, Alessandro Lenci

Motor Effects in Rating Lines' Length Using a Dichotomous Scale 1892
Lyuben Laskin

The Role of Event Knowledge in Comprehending Synesthetic Metaphors 1898
Tetsuaki Nakamura, Maki Sakamoto, Akira Utsumi

Simple Pointing to Objects may Facilitate Remembering 1904
Georgi Petkov, Prolet Nikolova

A Bayesian Antidote Against Strategy Sprawl 1910
Benjamin Scheibehenne, Jörg Rieskamp

Meaning Representation in Natural Language Categorization 1916
Trevor Fountain, Mirella Lapata

Temporal Chunk Signal Reflecting Five Hierarchical Levels in Writing Sentences 1922
Erlijn van Genuchten, P. C-H. Cheng

Emotion in Good Luck and Bad Luck: Predictions from Simplicity Theory 1928
Jean-Louis Dessalles

Scan Patterns on Visual Scenes predict Sentence Production 1934
Moreno I. Coco, Frank Keller

Modulation of motor-meaning congruity effects for valenced words 1940
Geoffrey Brookshire, Richard Ivry, Daniel Casasanto

Does micro-variability make models more complex? A comparison between diffusive and linear evidence accumulation 1946
Christopher Donkin, Richard Shiffrin, Scott Brown, Andrew Heathcote

SELF-ORGANIZATION, EMBODIED COGNITION AND THE BOUNDED RATIONALITY CONCEPT 1952
MARIA LUISA BIZZOTTO

The Role of Inhibition in Theory of Mind Performance: Evidence for a Non-Modular View of Theory of Mind 1957
Lindsey Frederixon Byom, Margarita Kaushanskaya

An ACT-R List Learning Representation for Training Prediction 1963
Michael Matessa

Eye Movements During Mental Imagery are Not Reenactments of Perception 1968
Roger Johansson, Jana Holsanova, Kenneth Holmqvist

The Role of Vagueness in the Numerical Translation of Verbal Probabilities: A Fuzzy Approach 1974
Franziska Bocklisch, Steffen F. Bocklisch, Martin R. K. Baumann, Agnes Scholz, Josef F. Krems

Selective attention and development of categorization: An eye tracking study 1980
Xin Yao, Vladimir Sloutsky

Eye-Movements of Dyslexic Children Reading in Regular Orthography: Exploring Word Frequency and Length Effects 1986
Evgenia Hristova, Alexander Gerganov, Ekaterina Todorova, Severina Georgieva

Mathematical reasoning with higher-order anti-unifcation 1992
Markus Guhe, Alison Pease, Alan Smaill, Martin Schmidt, Helmar Gust, Kai-Uwe Kühnberger, Ulf Krumnack

Head and Hand Movements in the Orchestration of Dialogue 1998
Stuart Battersby, Patrick Healey

Tracking Lexical and Syntactic Alignment in Conversation 2004
Christine Howes, Patrick G T Healey, Matthew Purver

Investigating phonotactics using xenolinguistics: A novel word-picture matching paradigm 2010
Vsevolod Kapatsinski, Lamia Johnston

Computer-based Learning of Neuroanatomy: A Longitudinal Study of Learning, Transfer, and Retention 2016
Julia Chariker, Farah Naaz, John Pani

Introspection and Mindreading as Mental Simulation 2022
Paul Bello, Marcello Guarini

Illusions of consistency in quantified assertions 2028
Niklas Kunze, Sangeet Khemlani, Max Lotstein, Phil Johnson-Laird

Can similarity-based models of induction handle negative evidence? 2033
Daniel Heussen, Wouter Voorspoels, Gert Storms

The Role of Dynamic Visualizations and Spatial Layout of Static Visualizations for Learning How to Classify Locomotion Patterns 2039
Birgit Imhof, Katharina Scheiter, Peter Gerjets, Jörg Edelmann

Working Memory Constraints on Multiple Center-Embedding 2045
Fred Karlsson

Coordination of understanding in face-to-face narrative dialogue 2051
Kathleen Eberhard, Hannele Nicholson

The Interpretation of Null and Overt Pronouns in Japanese: Grammatical and Pragmatic Factors 2057
Mieko Ueno, Andrew Kehler

Novel Words in Novel Contexts: The Role of Distributional Information in Form-class Category Learning 2063
Patricia Reeder, Elissa Newport, Richard Aslin

Optimal Language Learning: The Importance of Starting Representative 2069
Anna Rafferty, Thomas Griffiths

A Cognitive Model of Positive and Negative Congruency Effects in Unmasked Priming: The Role of Attentional Limit and Conflict 2075
Ahmad Sohrabi, Robert West

The role of action in perceiving and comparing functional relations 2081
Ivan Vankov, Boicho Kokinov

How Causal Reasoning Can Bias Empirical Evidence 2087
Momme von Sydow, York Hagmayer, Björn Meder, Michael R. Waldmann

Group Stratification and Coordination Failure in a Continuous N-Player Stag Hunt 2093
Seth Frey, Robert L Goldstone

Learning from Failures for Cognitive Flexibility 2099
Dongkyu Choi, Stellan Ohlsson

Motor Affordances in Object Perception 2105
Stephen Flusberg, Alexia Toskos Dils, Lera Boroditsky

Order Effects in Moral Judgment 2111
Alex Wiegmann, Yas Okan, Jonas Nagel, Stefan Mangold

Preferences in Cardinal Direction 2117
Marco Ragni, Benedikt Becker

Expanding Retrieval Promotes Long Term Retention by Preventing Rapid Rates of Forgetting .. 2123
Aimee Callender

Functional and Causal Abstractions of Complex Systems 2128
Ashok Goel, Swaroop Vattam, Spencer Rugaber, David Joyner, Cindy Hmelo-Silver, Rebecca Jordan, Sameer Honwad, Steven Gray, Suparna Sinha

The Effects of Work Shift and Strategy on an Orientation Task 2134
Tim Halverson, Glenn Gunzelmann, L. Richard Moore Jr., Hans P.A. Van Dongen

A Distributional Account of Covariance Effects and Talker Adaptation in Infant and Adult Phonetic Category Recognition 2140
Bevan Jones

Getting at the Cognitive Complexity of Linguistic Metadata Annotation: A Pilot Study Using Eye-Tracking 2146
Steffen Lohmann, Katrin Tomanek, Jürgen Ziegler, Udo Hahn

Subject-Object Asymmetries in Korean Sentence Comprehension 2152
Jiwon Yun, John Whitman, John Hale

The Benefit of Imitating Particular Individuals 2158
Yasuaki Sakamoto, Hongyuan Shi

Arithmetic Notation...now in 3D! 2164
David Landy, Sally Linkenauger

Computational Semantic Detection of Information Overlap in Text 2170
Julia Taylor

Do Baseball Fans Experience the Fan Effect? 2176
Travis Ricks, Jennifer Wiley

Prior expectations in pedagogical situations 2182
Patrick Shafto, Noah Goodman, Ben Gerstle, Francy Ladusaw

A Spiking Neuron Model of Serial-Order Recall 2188
Feng-Xuan Choo, Chris Eliasmith

Nonverbal Semantic Processing Disrupts Visual Word Recognition in Healthy Adults 2194
Lang Chen, Timothy T. Rogers

Effects of Anticipatory Coarticulation on Lexical Access 2200
Stephen Tobin, Pyeong Whan Cho, Patrick Jennet, James Magnuson

Cross Cultural Differences in Implicit Learning 2206
Sachiko Kiyokawa, Zoltán Dienes, Daisuke Tanaka, Ayumi Yamada

Interaction between lexical and syntactic structures in transcoding from verbal to Arabic numerals 2212
Rafael Hurtado, Mariela Orozco-Hormaza, Diego F. Guerrero

Language specific preferences in anaphor resolution: Exposure or Gricean maxims? 2218
Barbara Hemforth, Lars Konieczny, Christoph Scheepers, Savéria Colonna, Sarah Schimke, Peter Baumann, Joël Pynte

Designing state-trace experiments to assess the number of latent psychological variables underlying binary choices 2224
Guy Hawkins, Melissa Prince, Scott Brown, Andrew Heathcote

Simulating individual differences in language ability and genetic differences in FOXP2 using a neural network model of the SRT task 2230
Joseph Toscano, Kathryn Mueller, Bob McMurray, Bruce Tomblin

Time Course of Visual Attention in Statistical Learning of Words and Categories 2236
Chi-hsin Chen, Chen Yu, Damian Fricker, Thomas Smith, Lisa Gershkoff-Stowe

Development in the Estimation of Degree Measure: Integrating Analog and Discrete Representations 2242
Jonathan Vitale, John Black, Carson Eric, Chun-Hao Chang

The Disproportionate Face Inversion Effect in Recognition Memory 2248
Melissa Prince, Andrew Heathcote

Decomposing Externally Cued Task Switching Costs 2254
Christina Wasylyshyn

Deconfounding Hypothesis Generation and Evaluation in Bayesian Models 2260
Elizabeth Baraff Bonawitz, Thomas L. Griffiths

Finding a Bigger Fish Bowl: Higher Difficulty Helps Transitive Inferences 2266
Sarah Schwind, Heidi Kloos

Preschoolers sample from probability distributions 2272
Stephanie Denison, Elizabeth Baraff Bonawitz, Alison Gopnik, Thomas L. Griffiths

Topical Relevance and Information Quality in Cognitive Models of Web Search Behavior: Introducing Epistemic Scent into Information Foraging Theory 2278
Peter Gerjets, Yvonne Kammerer

Simulating the Elimination and Enhancement of the Plosivity Effect in Reading Aloud 2284
Qiang Liu, Alan Kawamoto

Ideal representations in a similarity space 2290
Wouter Voorspoels, Wolf Vanpaemel, Gert Storms

Learning Structured Generative Concepts 2296
Andreas Stuhlmüller, Joshua B. Tenenbaum, Noah D. Goodman

Short-Term Word Priming Across Eye Movements 2302
Stephen Denton, Richard M. Shiffrin

Toward a Large-Scale Characterization of the Learning Chain Reaction 2308
Alexei Samsonovich

The Effects of Communication Medium Upon Collaborative Orientation Task Performance 2314
Laura D'Andrea, Wai-Tat Fu

Semi-supervised learning is observed in a speeded but not an unspeeded 2D categorization task .. 2320
Timothy Rogers, Charles Kalish, Bryan Gibson, Joseph Harrison, Xiaojin Zhu

Embodiment and Virtual Reality in Learning to Visualize Anatomy 2326
Susan Jang, John Black, Robert Jyung

Predicting Students' Retention of Facts from Feedback during Study 2332
Robert Lindsey, Owen Lewis, Harold Pashler, Michael Mozer

Differences in the Development of Analogy Across Cultures: A Computational Account 2338
Leonidas Doumas, Robert Morrison, Lindsey Richland

Discerning Affect in Student Discussions 2344
Jihie Kim, Erin Shaw, Saul Wyner, Taehwan Kim, Jia Li

Recognizability of Individual Creative Style Within and Across Domains: Preliminary Studies .. 2350
Liane Gabora

Computational Modeling of Emotional Content in Music 2356
Kristine Monteith, Tony Martinez, Dan Ventura

Cross-Situational Statistical Learning: Implicit or Intentional? 2362
George Kachergis, Chen Yu, Richard Shiffrin

An Interactive Environment for Explanatory Biological Modeling 2368
Pat Langley

On the limits of dynamic imagination: A mental extrapolation task 2374
Florent Levillain, Luca Bonatti

(Category) Learning by Doing: How Goal Directed Tasks Constrain Conceptual Acquisition 2381
Seth Chin-Parker

Different Kinds of Pragmatic Factors Explain Failures of Default-to-Stereotype Inferences 2387
Matthias Unterhuber, Gerhard Schurz

The Effect of Graph Design Type on Word Preferences In the Description of Trend and Cyclic Events 2388
Ozge Alacam, Annette Hohenberger, Kursat Cagiltay

When to Switch? Understanding How Performance Tradeoffs Shape Dual-Task Strategy 2389
Duncan Brumby, Nina del Rosario , Christian Janssen

When More Load Leads to Less Distraction in Multimedia Learning: An Event-Related Potential Approach 2390
Krista DeLeeuw, Richard Mayer, Barry Giesbrecht

Information Search in Decisions from Experience: Do Our Patterns of Sampling Influence Our Decisions? 2391
Thomas Hills, Ralph Hertwig

The sensory nature of knowledge 2392
Lionel Brunel, Guillaume Vallet , Benoit Riou, Mathieu Lesourd, Elodie Labeye, Rémy Versace

Wrong prediction by experts provide more support than that by novices 2393
Kuninori Nakamura

Change in Encoding Facilitates Principle Acquisition 2394
Richard Prather

Complementary processing systems: A PDP model of the simultaneous perception of multiple objects 2395
Cynthia Henderson, James McClelland

Ascribing Causality and Intention to 2D Animations 2396
David Pautler, Bryan Koenig, Boon-Kiat Quek, Andrew Ortony

Limits in Monitoring and Recall with Constant and Changing Memory Sets 2397
Daniel Cassenti, Richard Carlson, Troy Kelley

Hick's Law in the Random-Dot Motion Task 2398
Leendert van Maanen, Eric-Jan Wagenmakers

Are Place Names Merely Referencing Expressions? 2399
Paula Engelbrecht, Michael Tull

Sociocultural history of the machine metaphor 2400
Vladimir Glebkin

Agile Software Development Process: A Case Of Collaborative Cognition In Flux 2401
Nik Nailah Binti Abdullah, Robert G.M Hausmann, Shinichi Honiden, Helen Sharp

The Influence of Prior Knowledge on Recall for Height 2402
Pernille Hemmer, Jenny Shi, Mark Steyvers

Harry Potter and the Sorcerer's Scope: Scope Biases in Explanatory Reasoning 2403
Sangeet Khemlani, Abigail Sussman, Daniel Oppenheimer

Cognitive Load Measurement through Multimodal Behaviour Patterns 2404
Natalie Ruiz, Ronnie Taib, Fang Chen

Conceptual understanding of the relationship between consciousness, memory, and attention ... 2405
Eunsook Kim, Hyunjung Shin

To Be Subtle or To Be Clear?: Comparing Strategies for Changing People's Attitudes Towards Social Groups 2406
M. Afzal Upal

Category directedness 2407
Steven Verheyen, Gert Storms

'Meryem (reportedly) missed her flight': Cognitive Implications of the Turkish Evidential 2408
Sumeyra Tosun, Jyotsna Vaid, Lisa Geraci

Abstract Perceptual Learning of Hidden Patterns 2409
Everett Mettler, Hongjing Lu, Philip Kellman

Metathesis in English and Hebrew: A Computational Account of Usaged-Based Phonology 2410
Paul De Palma, George Luger

Explaining the Minimal Counterintuitiveness Effect Without Assuming A Strongly Modular Mind 2411
M. Afzal Upal

Interpretate Novel Conceptual Combinations: Age-Related Impact of Memory Availability 2412
Sandra Jhean-Larose, Fabiola Martinez

Interactive Representation in the Motor Control 2413
Daniel Hsi-wen Liu

The network properties of episodic graphs 2414
Yuwen Zhuang, Vishnu Sreekumar, Mikhail Belkin, Simon Dennis

Weakly Supervised Learning: What Could It Do and What Could Not? 2415
Jinhui Yuan, Bo Zhang

Models of Information Integration in Perceptual Decision Making 2416
Jared Hotaling, Andrew Cohen, Jerome Busemeyer, Richard Shiffrin

The Influence of Integration and Counterintuitiveness on Memory for Text 2417
M. Afzal Upal, Mary Harmon-Vukic

An Attention Based Theory to Explore Affordances of Textual and Diagrammatic Proofs 2418
Peter Coppin, James Burton

Socially Facilitated Alignment and Novelty in Separate Channels of Communication 2419
Monica Riordan, Rick Dale, Roger Kreuz

The Influence of Causal Information on Memory for Misinformation 2420
Jessecae Marsh, Sarah Kulkofsky

The Interaction of Age and Skill for Recognition of Chess Positions 2421
Jerad H. Moxley, K. Anders Ericsson, Tyler J. Towne, Ryan Best

Thinking Ahead: How Children Reason About the Future 2422
Janani Prabhakar, Judith Hudson

When Distance Relationship Contradicts Similarity in SUSTAIN 2423
Chung-Yu Wang, Lee-Xieng Yang

Adaptive Information Indexing in Re-finding Information 2424
J. Michelle Moon, Wai-Tat Fu

Variability Helps Children Balance a Beam 2425
David Pfeiffer, Daniel Bullard, Heidi Kloos

Mutual Alignment Analogy Facilitates Abstraction and Transfer of a Complex Scientific Principle 2426
Judy Orton, Florencia Anggoro, Benjamin Jee

A Lexical Gap in the Humor Domain of Japanese and Its Possible Implications for Theories of Conceptual Language 2427
Whitney Vandiver

Reactive Task-Set Switching Ability, Not Working Memory Capacity, Predicts Change Blindness Sensitivity 2428
Robert Youmans

Perception of Visual Similarity: Modeling Feature-Based Effects 2429
Michael Romano, Michael Spivey

Preschooler's Performance in Three Visual Perspective Taking Tasks 2430
Yue Yu, Yanjie Su, Raymond Chan

Implicit Learning of Spatial Context by School-Age Children 2431
Hanako Yoshida, Kevin Darby, Joseph Burling

Studies of the Effects on Creativity of Having Very Different Parents 2432
Brian O'Connor, Liane Gabora

A Dynamic Memory Model 2433
Eva Cadez, Evan Heit, Vladimir Cadez

Effects of the Target Distribution on Numerical Prediction 2434
Jason Jones

Cognitive leaps and multiple epistemological resources: an agent-based modeling approach 2435
Paulo Blikstein

Representing Conceptual Knowledge: A Network Analysis 2436
Takashi Yamauchi, Ricardo Gutierrez-Osuna, James Caverlee

Development of the Semantic Network: From a random to a complex network 2437
Shohei Hidaka

Predicting Coreference: the role of alternative constructions 2438
Peter Baumann, Lars Konieczny, Barbara Hemforth

The Effects of Transcranial Magnetic Stimulation over Premotor Cortex on the Perception of Biological Motion 2439
Bianca van Kemenade, Neil Muggleton, Vincent Walsh, Ayse Pinar Saygin

Questioning the Free Will Comprehension Question 2440
Edward Cokely, Adam Feltz

Verb-action versus role relations congruence effects: Evidence from ERPs in picture-sentence verification 2446
Pia Knoeferle, Thomas P. Urbach, Marta Kuas

Vocabulary: Cross-Situational Learning

Integrating Syntactic Knowledge into a Model of Cross-situational Word Learning 2452
Afra Alishahi, Afsaneh Fazly

Sentence Processing Mechanisms Influence Cross-Situational Word Learning 2458
Judith Koehne, Matthew W. Crocker

Adaptive Constraints and Inference in Cross-Situational Word Learning 2464
George Kachergis, Chen Yu, Richard Shiffrin

Desirable Difficulties in Cross-Situational Word Learning 2470
Haley Vlach, Catherine Sandhofer

Vision: Attention in Childhood

The Goldilocks Effect: Infants' preference for stimuli that are neither too predictable nor too surprising 2476
Celeste Kidd, Steven T. Piantadosi, Richard N. Aslin

Infants' Visual Processing of Faces and Objects: Age-Related Changes in Interest, and Stability of Individual Differences 2482
Marybel Robledo, Gedeon Deák, Thorsten Kolling

Mechanisms of Sustained Selective Attention in 3- to 5-year-old Children: Evidence from a New Object Tracking Task 2488
Anna Fisher

Spatial Cognition: Routes, Surveys and Frames

The Influence of Route Planning and its Execution on Spatial Learning 2494
Kayoko Ohtsu, Yoshihiro Ouchi

The Direction Bias and the Incremental Construction of Survey Knowledge 2500
Tobias Meilinger, Heinrich H. Bülthoff

Alignment of Spatial Perspective 2506
Elena Andonova

Spatial Representations with Conflicting Intrinsic Frames of Reference 2512
Franklin Tamborello, Yanlong Sun, Hongbin Wang

Subsymbolic Learning: Environmental Regularities

Learning from Environmental Regularities is Grounded in Specific Objects not Abstract Categories 2518
Lauren Emberson, Dani Rubinstein

Adult language learners under cognitive load do not over-regularize like children 2524
Amy Perfors, Nicholas Burns

Descriptive Assessment of Jeffrey's Rule 2530
Jiaying Zhao, Daniel Osherson

Effects of Varied Priority Training on Complex Perceptual-Motor Learning 2536
Yi Wang, Michelle Moon, Wai-Tat Fu, Walter Boot, Kirk Erickson, Arthur Kramer

Action: Intention and Movement

Non-verbal responses to being ignored: Evidence of cognitive deconstruction? 2542
Emiel Krahmer, Juliette Schaafsma, Marc Swerts, Martijn Balsters, Ad Vingerhoets

The Bodily Movements of Liars 2548
Natasha Eapen, Sam Baron, Chris Street, Daniel Richardson

Exploring the mental space of autonomous intentional agents 2554
Peter Pantelis, Jacob Feldman

Actor-Observer Differences in Intentional Action Intuitions 2560
Adam Feltz, Maegan Harris, Ashley Perez

Symbolic Learning: Error Correction

Learning from Errors by Counterfactual Reasoning in a Unified Cognitive Architecture 2566
Andreea Danielescu, David J. Stracuzzi, Nan Li, Pat Langley

Optimal Inference and Feedback for Representational Change 2572
Yun Tang, Christopher J. Young, Jay I. Myung, Mark A. Pitt, John E. Opfer

Learning from Errors in Game-Based versus Formal Mathematics Contexts 2578
Lori Petersen, Jennifer Heil, Nicole McNeil, Gerald Haeffel

Collaborative Facilitation through Error-Detection: A Classroom Experiment 2583
Soniya Gadgil, Timothy Nokes

Thinking: Moral Concepts

A Double Causal Contrast Theory of Moral Intuitions in Trolley Dilemmas 2589
Michael R. Waldmann, Alex Wiegmann

Deconfounding Distance Effects in Moral Reasoning 2595
Jonas Nagel, Michael Waldmann

Developing notions of free will: Preschoolers' understanding of how intangible constraints bind their freedom of choice 2601
Nadia Chernyak, Tamar Kushnir, Henry Wellman

Symposium: Dynamic Decision Making

Dynamic Decision Making 2607
Todd Gureckis, Jared Hotaling, Michael Lee, Bradley Love, Dylan Simon

Vocabulary: Facilitating Factors

The Active Role of Partial Knowledge in Cross-Situational Word Learning 2609
Daniel Yurovsky, Damian Fricker, Chen Yu, Linda Smith

Cross-situational Learning of Low Frequency Words: The Role of Context Familiarity and Age of Exposure 2615
Afsaneh Fazly, Fatemeh Ahmadi-Fakhr, Afra Alishahi, Suzanne Stevenson

Object and Word Familiarization Differentially Boost Retention in Fast-Mapping 2621
Sarah C. Kucker, Larissa K. Samuelson

Attentional Control and Early Word Learning 2627
Hanako Yoshida, Duc Tran, Viridiana Benitez, Megumi Kuwabara

Perception: Cross-Modal Processing

A Bayesian Nonparametric Approach to Multisensory Perception 2633
Ilker Yildirim, Robert Jacobs

Attention and cross-modal processing: Evidence from heart rate analyses 2639
Chris Robinson, Vladimir Sloutsky

Evidence for auditory dominance in a passive oddball task 2644
Chris Robinson, Nayef Ahmar, Vladimir Sloutsky

Spatial Cognition: Functions of Diagrams

Effects of Problem Difficulty and Student Expertise on the Utility of Provided Diagrams in Probability Problem Solving 2650
Eliza J. Bobek, James E. Corter

Interactive Effects of Diagrammatic Format and Teleological Beliefs on Tree Thinking 2656
Brenda Phillips, Laura Novick, Kefyn Catley, Daniel Funk

Thinking with Networks 2662
Jeffrey V. Nickerson, Barbara Tversky, James E. Corter, Lixiu Yu, Yun Jin Rho, David Mason

Constructing internal diagrammatic proofs from external logic diagrams 2668
Yuri SATO, Koji MINESHIMA, Ryo TAKEMURA

Subsymbolic Learning: Language

Learning verb alternations in a usage-based Bayesian model 2674
Christopher Parisien, Suzanne Stevenson

Frequent Frames as Cues to Part-of-Speech in Dutch: Why Filler Frequency Matters 2680
Richard Leibbrandt, David Powers

When 'More' in Statistical Learning Means 'Less' in Language: Individual Differences in Predictive Processing of Adjacent Dependencies 2686
Jennifer B. Misyak, Morten H. Christiansen

Statistical Learning of Complex Questions 2692
Hartmut Fitz

Action: Perception and Simulation

Effector-specific Motor Interference in Action Simulation 2698
Peggy Tausche, Anne Springer, Wolfgang Prinz

Simulation from Schematics: Dorsal Stream Processing and the Perception of Implied Motion .. 2704
Kevin J. Holmes, Phillip Wolff

Assessing Behavioral and Computational Approaches to Naturalistic Action Segmentation 2710
Meredith Meyer, Philip DeCamp, Bridgette Hard, Dare Baldwin, Deb Roy

The Perception of Humans and Robots: Uncanny Hills in Parietal Cortex 2716
Ayse Pinar Saygin, Thierry Chaminade, Hiroshi Ishiguro

Symbolic Learning: Role of Disequilibrium

Modeling Cognitive-Affective Dynamics with Hidden Markov Models 2721
Sidney D'Mello, Art Graesser

Productive Failure in Learning the Concept of Variance 2727
Manu Kapur

Finding the Sweet Spot: Is There a Fixed Template for Culturally Successful Counterintuitive Narratives? 2733
M. Afzal Upal

Fortune Favors the Bold (and the Italicized): Effects of Disfluency on Educational Outcomes .. 2739
Daniel Oppenheimer, Connor Diemand-Yauman, Erikka Vaughan

Thinking: Relational Structures

Relational Versus Attributional Mode of Problem Solving? 2743
Svetoslav Bliznashki, Boicho Kokinov

Executive Control in Analogical Mapping: Two Facets 2749
Anna Chuderska

Selective Attention by Structural Alignment: An Eyetracking Study 2755
Aaron Hoffman, Bradley Love, Arthur Markman

A Structure-Mapping Model of Raven's Progressive Matrices 2761
Andrew Lovett, Kenneth Forbus, Jeffrey Usher

Symposium: Bridging the Gap from Cognitive Anthropology to Cognitive Science

Bridging the Gap: From Cognitive Anthropology to Cognitive Science 2767
Andrea Bender, Sieghard Beller, Giovanni Bennardo, James S. Boster, Asifa Majid, Douglas L. Medin

Vocabulary: Growth and Attrition

Semantic network connectivity is related to vocabulary growth rate in children 2769
Nicole Beckage, Linda Smith, Thomas Hills

Effects of Maternal Input on Language in the Absence of Genetic Confounds: Vocabulary Development in Internationally Adopted Children 2775
Carissa L. Shafto, Joy Geren, Jesse Snedeker

Who's afraid of similarity? Effects of phonological and semantic similarity on lexical acquisition. 2781
Sarah Sahni

A SOM Model of First Language Lexical Attrition 2787
Benjamin Zinszer, Ping Li

Perception: Stimulus Dimensions

The Encoding of Spatial Information During Small-Set Enumeration 2793
Harry Haladjian, Manish Singh, Zenon Pylyshyn, Randy Gallistel

Multiple visual cues enhance quantitative perception in infancy 2799
Joseph Baker, Jessica Feigleson, Kerry Jordan

Multidimensional Scaling Methods for Absolute Identification Data 2804
Pennie Dodds, Chris Donkin, Scott Brown, Andrew Heathcote

Mind Reading by Machine Learning: A Doubly Bayesian Method for Inferring Mental Representations 2810
Ferenc Huszar, Uta Noppeney, Mate Lengyel

Spatial Cognition: Acquisition

The Development and Assessment of Cross-Sectioning Ability in Young Children 2816
Kristin Ratliff, Charlotte McGinnis, Susan Levine

What can Information Extraction from Scenes and Causal Systems Tell us about Learning from Text and Pictures? 2822
Alexander Eitel, Katharina Scheiter, Anne Schüler

Does Spatial Verbal Information Interfere with Picture Processing in Working Memory? The Role of the Visuo-spatial Sketchpad in Multimedia Learning 2828
Anne Schueler, Katharina Scheiter, Peter Gerjets

Impact of placing icons next to hyperlinks on information-retrieval tasks on the web 2834
Saraschandra Karanam, Janhavi Viswanathan, Anand Theertha, Bipin Indurkhya, Herre van Oostendorp

Subsymbolic Learning: Childhood

Theory Learning as Stochastic Search 2840
Tomer Ullman, Noah Goodman, Joshua Tenenbaum

Is it me or the world? 16-month-olds distinguish competing hypotheses about the cause of failed interventions 2846
Hyowon Gweon, Laura Schulz

Developmental differences in learning the forms of causal relationships 2852
Christopher Lucas, Alison Gopnik, Thomas Griffiths

Children's Imitation of Action Sequences is Influenced by Statistical Evidence and Inferred Causal Structure 2858
Daphna Buchsbaum, Alison Gopnik, Tom Griffiths

Action: Thinking with the Body

Thinking with the Body 2864
David Kirsh

Symbolic Learning: Tutoring

The More the Merrier? Examining Three Interaction Hypothese 2870
Min Chi, Kurt VanLehn, Diane Litman

Comparing Worked Examples and Tutored Problem Solving: Pure vs. Mixed Approaches 2876
Rob Weitz, Ron Salden, Ryung Kim, Neil Heffernan

Interleaving Worked Examples and Cognitive Tutor Support for Algebraic Modeling of Problem Situations 2882
Albert Corbett, Stephen Reed, Bob Hoffman, Ben MacLaren, Angela Wagner

Learning during Intelligent Tutoring: When Do Integrated Visual-Verbal Representations Improve Student Outcomes? 2888
Kirsten Butcher, Vincent Aleven

Thinking: Explanation

Explanatory Reasoning for Inductive Confidence 2894
David Landy, John Hummel

Explanations make inconsistencies harder to detect 2900
Sangeet Khemlani, Phil Johnson-Laird

Why Does Explaining Help Learning? Insight From an Explanation Impairment Effect. 2906
Joseph Jay Williams, Tania Lombrozo, Bob Rehder

Explanation Constrains Learning, and Prior Knowledge Constrains Explanation 2912
Joseph Jay Williams, Tania Lombrozo

Symposium: Flux: Fundamental or Frivolous?

Flux: Fundamental or Frivolous? 2918
lera boroditsky, helen neville, christina karns, arthur markman, michael spivey

Reviewers List 2920

Author Index 2927

Cognition in Flux

Stellan Ohlsson and Richard Catrambone
Conference Co-Chairs

The slogan "Cognition in Flux" was chosen to emphasize two points. First, cognitive research is to a large extent research into how cognition changes over time. Second, the field of cognitive research – our shared theoretical vocabulary and our repertoire of research practices -- is itself continually changing. Both points are well illustrated by the content of program for the 2010 meeting of the Cognitive Science Society.

Explaining change is a central enterprise in the cognitive sciences. We received approximately 850 submissions, covering all areas of cognitive research. The majority of those submissions focused on cognitive change in some form, whether it be through models of subsymbolic learning, developmental studies of word learning, experimental evaluation of training procedures, classroom studies of instructional techniques, or through models of brain-like computations.

The fact that cognitive science itself is changing is equally evident in the submissions. Perhaps the most stunning trend over the past three decades is the dissemination of cognitive science concepts, techniques and results to all areas of study that pertain to humans, social entities and computational entities. From its core concerns with memory representations, processing limits and problem solving strategies 30 years ago, the perspective of cognitive science has rolled outwards in an ever widening circle to reach areas of inquiry that were once thought to lie outside its reach. The 2010 program included papers that presented cognitive perspectives on blame and punishment, the explanatory coherence of religious thought and the detection of fraud in corporate email networks, to mention only a few of the topics that would have raised eyebrows three decades earlier.

The labels for sessions and tracks that emerged out of the pool of accepted papers reflected this widening ring of influence. We had, for the first time, a track on Social Cognition; with deliberate provocation, we included the session on Human-Robot Interaction in that track. Another interesting trend is that work on perception and action, the input-output devices relegated to the periphery of the cognitive system in the first decades of cognitive science, has migrated towards the center of focus. This trend was represented by multiple sessions on perception or action, and by several of the symposia. Another remarkable feature of the 2010 program is the extraordinary attention paid to issues regarding language. The entire track B and all but the first two sessions of track A were devoted to research on various aspects of language, and even so the topic of language spilled over into sessions in other tracks.

As the stock of research topics grows broader, the repertoire of research techniques grows also. This is necessarily so. A viable scientific enterprise does not define itself by its methods but by its questions. Methods are tools, and as the research questions morph, so do the tools for answering them. There was much evidence for the evolution of tools in the 2010 program. The Internet provides new ways of collecting data, and ways of collecting new types of data. New computational techniques are applied to reverse

engineer the mental representations of experimental subjects. New modes of analyses reveal novel phenomena. If we had any bias in organizing the program, it was a bias against basing sessions and tracks on the similarity of methods in favor of basing them on the similarity of research questions. Hence, there were no tracks on Bayesian models, studies of children or the use of eye movement recordings. Instead, papers that included Bayesian models, observations on children or eye movements were sorted with other papers that addressed the same or some conceptually related research question.

Of the approximately 850 submissions, 270 or 30 % were included in the program as talks. The two poster sessions on Thursday and Friday afternoon included more than 200 posters each. Papers and abstracts are included in full in these proceedings. The program also featured invited talks, tutorials, workshops and symposia. The latter are listed. We hope that the Proceedings will serve as a useful collection of cutting-edge papers.

CogSci 2010 Organizing Committee

CogSci 2010 Program Co-Chairs:

Richard Catrambone and Stellan Ohlsson

Awards Co-Chairs:
Laura Carlson and Thomas Shipley

Member Abstracts Chair:
Christoph Hoelscher

Publicity Chair:
Mitchell J. Nathan

Scheduling Chair:
Simon Dennis

Sponsors Chair:
Timothy T. Rogers

Student Volunteers Chairs:
Glenn Gunzelmann and
Christopher Myers

Symposia Chair:
Leendert van Maanen

Tutorials and Workshops Chair:
Duncan Brumby

Web Page Chair:
Dongkyu Choi

CS Conference Officer:
Kevin Gluck

PCS Manager:
James Stewart

Local Arrangements Chairs:
Deborah Gruber and Nicole Dillon

CogSci2010 Program Committee

Altmann, Erik
Atkinson, Robert
Barley, Mike
Bassok, Miriam
Bello, Paul
Best, Brad
Billman, Dorrit
Blessing, Steve
Bonnefon, Jean-Francois
Bosse, Tibor
Brook, Andrew
Brumby, Duncan
Burns, Bruce
Busemeyer, Jerry
Byrne, Ruth
Carlson, Richard
Cassimatis, Nicholas
Castelfranchi, Cristiano
Chipman, Susan
Clancey, Bill
Clement, Cathy
Cox, Anna
Dennis, Simon
Doane, Stephanie
Epstein, Susan L.
Forbus, Ken
Frank, Mike
French, Bob
Fu, Wai-Tat
Fum, Danilo
Gentner, Dedre
Goel, Ashok K.
Goldstone, Robert
Griffiths, Tom
Gunzelmann, Glenn
Hegarty, Mary
Helie, Sebastien
Hoelscher, Christoph
Holyoak, Keith
Howes, Andrew
Hummel, John
Hutchins, Edwin
Kamawar, Deepthi
Kashak, Mike
Keane, Mark
Kemp, Charles
Klenk, Matthew
Koedinger, Ken
Kokinov, Boicho
Landy, David
Langley, Pat
Larreamendy, Joerns
Lee, Michael D.
Lombrozo, Tania
Luger, George
Magnani, Lorenzo
Maloney, Larry
Markman, Art
McDaniel, Mark
McNeil, Nicole
Miller, Craig
Minda, John Paul
Mitrovic, Tanja
Navarro, Dan
Nokes, Timothy
Opfer, John E.
Ormerod, Tom
Pani, John
Peebles, David
Petrov, Alex
Pleskac, Tim
Rapp, David
Reed, Steve
Rehder, Robert
Scheutz, Matthias
Schmalhofer, Franz
Schoelles, Mike
Schunn, Christian
Shafto, Mike
Shipley, Thomas
Shultz, Thomas
Sloman, Steve
Sloutsky, Vladimir
Stracuzzi, David
Sun, Ron
Taatgen, Niels
Tenenbaum, Josh
Thagard, Paul
Trafton, Greg
Treur, Jan
Uttal, David
van Elst, Ludger
van Maanen, Leendert
van Rijn, Hedderik
Youmans, Robert
Young, Richard
Zuidema, Willem

CogSci 2010 Awards

Marr Prize

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

The winner of the 2010 Marr Prize for Best Student Paper is:

Hyowon Gweon and Laura Schulz

Is it me or the world? 16-month-olds distinguish competing hypotheses about the cause of failed interventions

Saturday, 1:00 p.m., Track D

Computational Modeling Prizes

Four prizes worth $1,000 each are awarded for the best full paper submissions to CogSci 2010 that involve computation cognitive modeling. The four prizes represent the best modeling work in the areas of perception/action, language, higher-level cognition, and applied cognition. These prizes are all sponsored by The Cognitive Science Society.

The winners of the 2010 Computational Modeling Prizes are:

Applied Cognition

Eldad Yechiam and Eyal Ert
Risk attitude in decision making: A clash of three approaches

Thursday, 12:00 noon, Track G

Perception/Action

Celeste Kidd, Steven T. Piantadosi, and Richard N. Aslin
The Goldilocks Effect: Infants' preference for stimuli that are neither too predictable nor too surprising

Saturday, 10:00 a.m., Track B

Language

Yang Xu and Charles Kemp
Constructing spatial concepts from universal primitives

Thursday, 1:30 p.m., Track A

Higher-Level Cognition

Daniel Rasmussen and Chris Eliasmith
A neural model of rule generation in inductive reasoning

Thursday, 10:30 a.m., Track C

Awards, Cont'd

Cognition and Student Learning (CaSL) Prize

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

The winner of the 2010 Cognition and Student Learning Prize is:

Daniel Oppenheimer, Connor Diemand-Yauman, and Erikka Vaughan
Fortune Favors the Bold (and the Italicized): Effects of Disfluency on Educational Outcomes

Saturday, 11:30 a.m., Track F

Student Travel Awards

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

The 2010 Travel Awards went to:

Jamie Alexandre	Kyle Jennings	Ben Rottman
Daniel Belenky	Brendan Johns	Solveig Bosse
Christopher Carroll	Artem Kaznatcheev	Peggy Tausche
Colin Dawson	Celeste Kidd	Haley Vlach
Heeyeon Y.Dennison	Itamar Lerner	Jing Xu
Hyowon Gweon	Khetarpal Naveen	Yang Xu
Nicholas Gwynne	Peter Pantelis	Benjamin Zinszer
Harry Haladjian	Daniel Rasmussen	

Awards Committee

Laura A. Carlson (co-chair), Thomas F. Shipley (co-chair), Erik M. Altmann, Felice Bedford, Gary Dell, Frank H. Durgin, Cynthia Fisher, Art Graesser, Scott Johnson, Barbara Landau, Jeffrey Lidz, Mark McDaniel, Ennio Mingolla, Willis F. Overton, Thomas J. Palmeri, Hal Pashler, Terry Regier, Alexander Renkl, Michael K. Tanenhaus, John Trueswell.

CogSci 2010 Sponsors

We sincerely thank the sponsors of the 32nd Annual Meeting of the Cognitive Science Society for their support of the conference awards and the tutorials, and for supporting student participation through reduced registration fees and coverage of travel costs.

Air Force Office of Scientific Research (AFOSR)

National Science Foundation (NSF)

Air Force Research Laboratory (AFRL)

Office of Naval Research (ONR)

Institute of Education Sciences (IES)

The Robert J. Glushko and Pamela Samuelson Foundation

Invited Plenary Presentations

Rumelhart Prize Lecture:
Emergence of Semantic Structure from Experience
James McClelland, Stanford University.
Thursday, August 12, 3:00 – 4:00 p.m.

A Hierarchical Competing Systems Model of the Emergence and Early Development of Executive Function
Philip Zelazo, University of Minnesota.
Thursday, August 12, 9:15 – 10:15 a.m.

Accelerated Learning through Adaptive, Data-Driven Instructional Design
Marsha Lovett, Carnegie-Mellon University.
Friday, August 13, 8:45 – 9:45 a.m.

Bridge Over Troubled Water: From Cognitive Science to Designing Digital Instruction
Peter Gerjets, University of Tübingen.
Saturday, August 14, 8:45 – 9:45 a.m.

Symposia

Rumelhart Prize Symposium:
Graded, Distributed, and Interactive: How Parallel Distributed Processing Has Influenced Cognitive Science.
Friday, August 13, 10:00 – 11:30 p.m.

This symposium honors the career of James L. McClelland, winner of the 2010 Rumelhart Prize. When Jay entered the field, cognitive science looked very different: representations and processes were thought to be modular, independent, symbolic, and discrete, and the discipline proceeded independently from neuroscience. Through computational modeling, elegant empirical studies, and close attention to neuroscience, Jay's work contributed to an alternative view: that cognitive representations are graded and not discrete; that representations and processes are highly distributed and non-modular; that cognitive processes are highly interactive and not functionally independent; and that information about the brain importantly constrains theories of cognitive functioning. This symposium illustrates how these themes have shaped current thinking in influential research on the graded nature of sublexical representations, the interactive nature of language comprehension and perception, the distributed nature of neural representations revealed by multi-voxel pattern analysis, and the functional and neuroanatomical organization of the semantic system.

Timothy T. Rogers, Bruce Hayes, Michael Spivey and Nikolaus Kreigeskorte

Invited Symposium 1:
Abductive Reasoning: Inferring Explanations.
Friday, August 13, 10:00 – 11:30 p.m.

The American philosopher Charles Peirce used the term abduction to describe inference to explanatory hypotheses, including both the initial generation of hypotheses and their evaluation, which is now commonly called inference to the best explanation. This symposium will discuss current interdisciplinary work on abduction, including research in artificial intelligence and linguistics (Hobbs), psychology (Lombrozo), philosophy (Magnani), and computational neuroscience (Thagard).

Paul Thagard, Jerry Hobbs, Tania Lombrozo, and Lorenzo Magnani.

Invited Symposium 2:
Balancing Internal and External Cognition: A Learning Process.
Friday, August 13, 10:00 – 11:30 p.m.

Human cognition is able to accomplish amazing feats, even though it has to deal with limited cognitive resources. A solution to dealing with cognitive limitations is to outsource as much as possible to the outside world: instead of relying on representations in the head we rely on representations in the world. In this symposium we will examine the relationship between information in the head and in the world, and how to optimize it. This can be in terms of finding the optimal division of labor between internal and external, to examine learning processes that converge to such an optimal solution, to study how external representations can lead to optimal teaching solutions, and how good performance in multitasking can be related to internal and external control.

Niels Taatgen, Wai-Tat Fu, Ken Koedinger, and Andrew Howes.

The following symposia are listed in the order in which they were presented:

Success in the Theory of Mind

Thursday, August 12, 10:30 a.m. – 12: 00 noon.

Peter wants to get the beer he left in the refrigerator. Predicting Peter's behavior correctly is usually an easy matter, but understanding how people correctly predict his behavior with ease is a much more difficult task. Thirty years of research on theory of mind has focused on the interesting few cases in which fail to reason about mental states correctly. However, it is perhaps more interesting to explore the common, reliable cases of successful theory of mind reasoning. This symposium presents cutting-edge research using several different experimental approaches to studying the processes involved in successful instances of theory of reasoning, as well as the processes involved in developing the ability to succeed consistently across the life span. In this symposium, research employing a variety of measures – with toddlers, preschoolers, school-age children, and adults – takes aim at current debates central to the field and delivers weighty results.

Rose Scott, Adam Petrashek, Noah Goodman, Adam Cohen, Rebecca Saxe, Renee Baillargeon, Ori Friedman and Tamsin German

Prospective Perception

Thursday, August 12, 12:00 noon – 1:30 p.m.

Recent data indicate that perception is inherently prospective (i.e., anticipatory). The purpose of this symposium is to examine the research of three scholars who approach prospective perception from three different theoretical perspectives: the Theory of Event Coding, the Economy of Action theory, and the Ecological Theory. The panelists will examine differences among these theories and address the extent to which prospective perception research affords a means of potentially integrate these three theories.

Jerome S Jordan, Jessica Witt and Michael Riley

The Philosophy of Affective Neuroscience

Thursday, August 12, 1:30 – 3:00 p.m.

This panel showcases the interdisciplinary cutting edge innovations of the cognitive sciences. It is the unique meeting of the founder of Affective Neuroscience with an interdisciplinary set of scholars who follow the implications of this work through the philosophy of psychology, the philosophy of Self, and neuroscience and law.

Rami Gabriel, Jaak Panksepp, Stephen Asma and Glennon Curran

The Mechanics of Embodiment
Friday, August 13, 12:30 – 2:00 p.m.

There currently exist a large number of interesting and intriguing empirical effects regarding embodied cognition. A critical next step in the development of embodied theories is to flesh out ideas in terms of implemented computational models. This symposium features speakers who currently are working toward that goal. These researchers describe and discuss challenges for embodied models. The major focus is on presenting current efforts to model human cognition in a physical agent with sensory and motor capabilities, implementing the perceptual symbols systems framework, and modeling the dynamic on-line influences of integrated sensori-motor processes.

Giovanni Pezzulo, Angelo Cangelosi, Michael Spivey, Lawrence Barsalou, Martin Fischer and Ken McRae

Developmental and Computational Perspectives on Infant Social Cognition
Friday, August 13, 2:00 – 3:30 p.m.

Adults effortlessly and automatically infer complex patterns of goals, beliefs and other mental states as the causes of others' actions. Yet before the last decade little was known about the developmental origins of these abilities in early infancy. Our understanding of infant social cognition has now improved dramatically. Even preverbal infants appear to perceive goals, preferences, and even beliefs from sparse observations of intentional agents' behavior. Furthermore, they use these inferences to predict others' behavior in novel contexts and to make social evaluations.

Noah Goodman, Chris Baker, Joshua Tenenbaum, Chris Lucas, Kiley Hamlin, Tamar Kushnir, Tomer Ullman and Elizbeth Spelke

Emerging Insights from Eye-Movement Research on Category Learning
Friday, August 13, 3:30 – 5:00 p.m.

This symposium brings together four talks on eye-tracking and categorization. Each talk focuses on a different aspect of categorization and demonstrates how eye-tracking can extend our knowledge. One recent trend in category learning is the use of alternative training procedures. The inference learning task is the most popular of these procedures and in the first talk Aaron Hoffman presents eye-tracking data illuminating the differences between inference learning and categorization. Bob Rehder then presents his recent work on understanding the learning difficulties associated with Parkinson's disease. Marcus Watson discusses work using eye-tracking to inform our understanding of the basic issue in category learning: error. Finally, Mark Blair discusses the relationship between working memory, attention and performance in a category learning task.

Bob Rehder, Mark Blair, Aaron Hoffman and Marcus Watson

Dynamic Decision Making
Saturday, August 14, 10:00 – 11:30 a.m.

The experimental study of decision-making has historically focused on simple single-trial judgment or reasoning tasks. However, real world behavior often necessitates on-line decision making, planning and sequentially organized behavior. The goal of the proposed symposium is to bring together researchers who are working to understand the cognitive processes underlying dynamic decision-making, defined as tasks or contexts that are structured as a sequence of interdependent decision. A symposium on this topic is particularly timely since research in this area is having a tremendous impact on the field of psychology. The key topics covered are: how people plan sequences of actions to accomplish goals; the neurobiology of sequential decision-making and planning; how cognitive representations of the task environment support planning and decision-making; and how people balance exploration and exploitation to arrive at effective decision strategies in unknown environments.

Todd Gureckis Jared Hotaling, Michael Lee, Bradley Love and Dylan Simon

Bridging the Gap: From Cognitive Anthropology to Cognitive Science
Saturday, August 14, 11:30 a.m. – 1:00 p.m.

Although cognitive anthropology once was a pioneer in the cognitive revolution and founding member of the cognitive sciences, over the years its participation and influence have diminished – to the detriment of both cognitive anthropology and cognitive science. Meanwhile, though, interactions between culture and cognition are increasingly recognized as being of prime interest for cognitive science. Among the most important issues that call for anthropological expertise is the question of cognitive and/or linguistic universals. Anthropology, with its expertise in culture and language, thus becomes an invaluable partner for cognitive science research. But only recently have initiatives been launched to re-calibrate the relationships among the subfields of cognitive science. This seminar will review such initiatives.

Andrea Bender, Sieghard Beller, Giovanni Bennardo, James S Boster, Asifa Majid and Douglas Medin

Flux: Fundamental or Frivolous?
Saturday, August 14, 1:00 – 2:30 p.m.

A broad range of findings across the cognitive sciences has emerged, revealing surprising flexibility and dynamic flux in a large range of cognitive domains. These include exciting new discoveries of neuroplasticity well into adulthood, of great cognitive variability as a function of the statistical properties of one's environment (from the patterns in natural language to those in embodied experience), and discoveries of the surprisingly dynamic microstructure of cognition. Do such findings demonstrate that many fundamental aspects of cognition are indeed quite flexible? Or does a finding that some aspect of cognition is flexible mean that it is therefore not fundamental? Or is flux the only truly fundamental thing about cognition in the first place? The talks in this symposium will speak to these questions from a variety of perspectives (incorporating ideas from development, neuroscience, computational studies, and cross-cultural approaches) and they aim to help us clarify our thinking about what such findings mean.

Lera Boroditsky, Helen Neville, Christina Karns, Arthur Markman and Michael Spivey

Workshops and Tutorials

All workshops and tutorials took place on Wednesday, August 11.

All Day Workshops

The Invited Workshop on Cognitive Social Sciences – Grounding the Social Sciences in Cognition?
Organizer: Ron Sun
9:00 a.m. – 17:00 p.m., Room A105

Compositional Connectionism in Cognitive Science II: The Localist/Distributed Dimension
Organizers: Ross W. Gayler and Simon D. Levy
9:00 a.m. – 17:00 p.m., Room A106

Semantic Development: An Interdisciplinary Approach
Organizers: David Barner and Susan Carey
9:00 a.m. – 17:00 p.m., Oregon Ballroom 204

Workshop on Understanding, Predicting and Mitigating Error in Routine Procedural Tasks
Organizers: Anna L. Cox, Duncan P. Brumby, and Jonathan Back
9:00 a.m. – 17:00 p.m., Room B110-111

Half Day Workshop:

Staying in the Academic Pipeline: Growing Professionally in an Economic Drought
Organizers: Janet van Hell, Laurie Feldman, Judith Kroll, and Suparna Rajaram
13:30 – 17:00 p.m., Room B115-116

All Day Tutorials

An Introduction to Agent-Based Computer Modelling for Cognitive Research
Organizer: Paulo Blikstein
9:00 a.m. – 17:00 p.m., Room 107-108

Bayesian Models of Inductive Learning
Organizers: Thomas L. Griffiths, Charles Kemp, and Joshua B. Tenenbaum
9:00 a.m. – 17:00 p.m., Oregon Ballroom 201

Doing Bayesian Data Analysis with R and BUGS
Organizer: John K. Kruschke
9:00 a.m. – 17:00 p.m., Room B113-114

Dynamic Field Theory: Conceptual Foundations and Applications in the Cognitive and Developmental Sciences
Organizers: Gregor Schoner, Anne Schutte, and Sammy Perone
9:00 a.m. – 17:00 p.m., Room B112

All Day Tutorial

Nengo and the Neural Engineering Framework: Connecting Cognitive Theory to Neuroscience
Organizers: Chris Eliasmith and Terrence C. Stewart
9:00 a.m. – 17:00 p.m., Room A109

Half Day Tutorials:

Tutorial on Model Comparison Methods
Organizers: Jay I. Myung and Mark A. Pitt
9:00 a.m. – 12:30 p.m., Room B115-116

Building Models of Learning and Expertise with CHREST
Organizers: Peter C. R. Lane and Fernand Gobet
9:00 a.m. – 12:30 p.m., Room A104

The CLARION Cognitive Architecture
Organizers: Nicholas Wilson, Michael Lynch, and Ron Sun
13:30 p.m. – 17:00 p.m., Room A104

Tutorials and Workshops Committee

Duncan P. Brumby (Chair), Andrew Brook, Matthew W. Crocker, Thomas L. Griffiths, Ulrike Hahn, Tim Halverson, Gary Jones, Jelena Mirkovic, Padraic Monaghan, Frank Ritter, Terrence C. Stewart, Michael Thomas, and Richard M. Young

Wild Systems Theory:
Overcoming the Computational-Ecological Divide via Self-Sustaining Systems

J. Scott Jordan (jsjorda@ilstu.edu)
Department of Psychology, Illinois State University, Campus Box 4620
Normal, IL 61790-4620

Abstract

For years, there has been a tension between computationalist cognitive scientists who utilize the notion of representation and efficient-cause in their accounts of mind, and dynamical-systems oriented ecological psychologists who eschew representationalism and efficient-cause in favor of multi-scale, contingent interactions and embodiment. The present paper presents a recently-developed theory of embodiment, Wild Systems Theory (WST), that was developed to overcome this riff. WST conceptualizes organisms as multi-scale self-sustaining embodiments of the phylogenetic, cultural, social, and developmental contexts in which they emerged and in which they sustain themselves. Such self-sustaining embodiments of context are *naturally* and *necessarily* about the multi-scale contexts they embody. As a result, meaning (i.e., content) is constitutive of what they are. This approach to content overcomes the computationalist need for representation while simultaneously satisfying the ecological penchant for multi-scale contingent interactions.

Keywords: representation; phenomenology; embodiment; philosophy.

Wild Systems Theory

For years, there has been a tension between computationalist cognitive scientists who utilize the notion of representation and efficient-cause in their accounts of mind, and dynamical-systems oriented ecological psychologists who eschew representationalism and efficient-cause in favor of multi-scale, contingent interactions and embodiment. Wild Systems Theory (WST) is a new theory of embodiment that was developed to overcome this riff. My central thesis is that organisms (i.e., bodies) *are* meaning (and ultimately mind), precisely because they constitute embodiments of the external constraints (i.e., contexts) they have had to phylogenetically, as well as ontogenetically internalize in order to sustain themselves (Jordan, 1998). Within this framework, fins constitute an embodiment of the hydrodynamic properties of water, bones, an embodiment of the constraints that need to be overcome in order to propel a body through a gravity field, and teeth, an embodiment of the make-up of plants and what it takes to release the chemical energy they contain. In every case, these embodiments are *naturally* and *necessarily* "about" the environmental constraints they evolved to address. It is this necessary "aboutness" that I want to define as meaning and, ultimately *mind*.

But does this notion of internalized constraints really naturalize meaning and, ultimately mind? One could argue that the body of a submarine, the body of a car and the body of certain farm tools also constitute embodiments of water, gravity and plants, respectively. Do they really constitute *meaning*? Of course, I have to say yes, but I would also add that this is not the type of meaning that ultimately evolved into mind. To be sure, the designers of submarines, cars and farm tools constructed such bodies so that their internal structure reflected the external constraints within which they have to function. The difference between these bodies however, and biological bodies is the means by which they sustain themselves. Biological bodies do so by continuously taking in, transforming, and dissipating energy. Non-biological bodies do not. It is my position that it is this wild, interactive-internalization of local context (i.e., energy transformation) that afforded, and continues to afford biological systems the means by which their embodied meaning was and is capable of evolving into mind. This is because the work (i.e., energy transformation) that constitutes biological bodies is *self-sustaining*. That is, it produces products that feed back into and sustain the work. Kaufmann (1995) recognized this principle at the chemical level and refereed to it as *autocatalysis*. Specifically, an autocatalytic (i.e., self-sustaining) chemical system is one in which the work (energy transformations) taking place among molecules, produces its own catalyst. By producing its own catalyst, the work sustains itself, as well as the system as a whole. Kauffman conceptualizes such work as a self-sustaining metabolism and argues that the emergence of such systems constituted the emergence of living systems.

According to Wild Systems theory, such self-sustaining "work" constituted a type of meaning—what Jordan and Ghin (2006) refer to as *content*—that proved capable of evolving into mind. It constituted *meaning* because the work, as well as the global whole it sustained, was naturally and necessarily "about" the external constraints the system had to embody in order to sustain itself. It constituted *content* because it gave rise to (i.e., was *for*) the global whole (i.e., the body) it sustained, while the body (i.e., the sustained global whole) synergistically provided a sustained context in which the internal work could be *for* something (cf., Bickhard, 2001; Jordan & Ghin, 2006). And it proved capable of affording the evolution of mind because it constituted a potential fuel source (i.e. encapsulated energy). That is, the energy entailed in such a system could be captured by another system. But to be capable of doing so, the latter had to internalize (i.e., embody) all the constraints that needed to be addressed in order to capture the energy encapsulated in the former. Said another way, once plant energy was widely available, it provided a context in which a system could emerge that sustained itself on plant energy. From this perspective, herbivores can be seen as

embodiments of the constraints that need to be internalized in order for a system to sustain itself on the energy encapsulated in plants, and carnivores, the constraints to be addressed to sustain a system on the energy encapsulated in herbivores. What we have here then, is a continuing recursion on a simple theme; specifically, *the fuel source dictates the consumer*. From this perspective, the world of nature is conceptualized as a self-organizing energy transformation hierarchy (Odum,1988; Vandervert, 1995) in which any newly emerging systems constitute embodiments of the constraints they have to address to sustain themselves within this transformation hierarchy.

According to WST, within the context of such a self-sustaining hierarchy, *mind* emerges when systems emerge that are capable of embodying (i.e., internalizing) *virtual* content. By virtual, I simply mean content that is "about" events that are non-existent in the present context. Take, for example, a lion chasing a gazelle. Lotka (1945) recognized that in order to capture the energy entailed in the gazelle, the lion must propel itself as a whole on an *anticipatory* pursuit curve. What makes the pursuit curve anticipatory is the fact the lion runs toward a location the gazelle does not yet occupy. In short, it propels itself toward the gazelle's *future*. The reason it can do so is because it has embodied (i.e., internalized) the constraint of having to capture a moving energy source. Specifically, certain structures in the lion's cerebellum have access to both the movement commands leaving motor cortex, and the immediate sensory consequences of the resultant movements. These cerebellar structures project back up to motor cortex and influence its activity. This is important, for it affords the lion the ability to embody (i.e., internalize), in the weights of its cerebral-cerebellar circuitry, patterns between motor commands and their resultant sensory effects. Thus, as the lion garners experience controlling its body in relation to moving prey, successful command-feedback patterns become embodied in the cerebral-cerebellar circuits. And given these cerebral-cerebellar loops influence motor cortex and function at a time scale of 10-20 milliseconds, versus the 120 millisecond time-scale between motor commands and sensory feedback, the system can basically control is propulsion on *virtual* feedback (Clark, 1997; Grush, 2004) and, as a result, propel itself toward internalized (i.e., embodied) *virtual* prey locations (i.e., where the prey will be in the next 200 or so milliseconds).

There are five important points to be made about such *virtual* content. First, it is not virtual in the sense it does not exist. To the contrary, it does exist. It is virtual in the sense it is about *future* body-prey states. Second, it is possible for the lion to embed (i.e., embody) such content within its brain because neural networks function according to the principle of self-sustaining work. Hebb (1949) recognized this aspect of neural work and refereed to it as the cell-assembly; the notion that neurons sustain themselves by becoming part of a neural network. Edelman (1989) also noted this principle in the developing brain, and referred to it as *Neuronal Darwinism*. In short, the work of being a neuron (i.e., producing action potentials and forming synapses with other neurons) sustains the neuron. Thus, patterns of neural activity sustain themselves, and factors that cause neural patterns to repeat (i.e., command-feedback patterns in cerebral-cerebellar loops and their relationship to prey patterns) become embedded (i.e., embodied) within these self-sustaining neural patterns.

Third, all of this embodied work is naturally and necessarily about the external (as well as internal) contexts (patterns) that have to addressed in order for the work to sustain itself; from the single neuron, to the neural circuit, to the neuro-muscular system, to the organism as a whole. Thus, there is no epistemic divide between internal and external states (including virtual states)—organisms are reciprocally nested eco-systems of self-sustaining work. They are a representation, at every level, of the phylogenetic, as well as ontogenetic constraints their species has had to overcome in order to sustain itself.

This leads to the fourth point. Virtual content emerged in self-sustaining systems precisely because of their need to capture energy that was on the move. The virtual content therefore, is necessarily about the *other*. That is, it is not just about the command-feedback patterns in the lion's brain, but rather, the relationship between command-feedback patterns and their relationship to prey patterns. The point I'm after here is that the virtual content is inherently *other-relative*. If we assume that the ability to chase gazelles phylogenetically emerged prior to the ability to have self-consciousness about chasing gazelles, it seems to be the case that *others* were in the brain before the *self* was. In short, the brain has never been alone. This claim is supported by the discovery of areas in the brain (i.e., mirror neurons) that are active both when one plans a goal related action, as well as when one observes another execute such an action (Rizzolatti, Fadiga Fogassi, & Gallese, 2002). This means that as others produce goal-directed actions, they simultaneously put my brain in a planning state for the same goal-related action. The discovery of such mechanisms indicates that resonance (i.e., doing what others are doing) constitutes the default value in human interaction. Kinsbourne (2002) agrees with this position and argues that infant imitation is actually uninhibited perception "on the fly". Only as the cortex develops inhibitory circuits, he argues, are we able to "not" resonate to the actions of others. He cites echopraxia as further evidence of this claim. Rizzolatti et al. agree with this notion of resonance, and distinguish between low- and high-level resonance. While the former refers to the ability of an organism's body movements to entrain similar movements in conspecifics (e.g., a school of fish moving together, or a flock of birds flying together), the latter refers to resonance at the level of goal related actions (e.g., a chimp watching another eat a peanut, or a person watching another dance). Collectively, these findings indicate that the *othe*r was embodied in the structure of the brain very early on, and has been there ever since.

And finally, the fifth point about virtual content is that it sets the stage for the emergence of phenomenal self-experience (Ghin, 2005; Metzinger, 2003). For since neural networks emerge and function according to the principle of self-sustaining work, the virtual content embedded in a brain is always available for "capture" by newly-emerging neural networks (Grush, 2004). The content of these new circuits will necessarily constitute an abstraction from the content embedded and sustained in the network it is tapping into.

As systems emerged that were capable of externalizing and sharing virtual content (i.e., communicate), the ability to "capture" such content required the system be able to distinguish its own, internally-generated virtual content from that entering the system from the outside. These are the constraints that I believe forced the emergence of "self" and "other" (Jordan, 2003c; Jordan & Knoblich, 2003; Knoblich & Jordan, 2003). In short, the self emerges as foreground amidst a background of virtual others, and it does so in order to sustain itself with those others in virtual contexts (i.e., within a world of ideas). The phenomenal self then garners its content (i.e., phenomenal properties) as do all self-sustaining systems; from the fact it is naturally and necessarily "about" the context (i.e., the externalized virtual content of others) it must embody in order to sustain itself.

The idea that the other has always been there, embodied within us, seems to render communication more an act of self-sustaining resonance among embodied others than an act of information exchange between lone cognizers. It does so because self-sustaining systems do not need to "perceive" their environment in order to be "about" it. Rather, they are naturally and necessarily about the contexts they have embodied, including the context of others. Environments therefore, including the world of others, modulate (versus 'cause') what self-sustaining systems are "about". Communication therefore, at least among self-sustaining embodiments, is an act of reciprocal modulation (i.e., resonance). And in order for such resonance to sustain itself, participants must generate work (e.g., eye-contact, gestures and head nods) to sustain the joint modulation. In short, communication itself is a self-sustaining process. Instead of constituting work among chemical systems embedded in a pre-biotic soup however, it constitutes work among embodied others embedded in a sea of virtual meaning.

Overcoming the Divide

Given its ability to satisfy the concerns of both computationalists and ecological psychologists without violating the assumptions of either, WST might be in a position to integrate the two theories. As regards computationalism, WST address the notion of representation by arguing that all aspects of an organism constitute representations, in that, all aspects of the organism constitute embodiments of context. In short, an organism represents all the scales of context that have had to be addressed for it to phylogenetically emerge and sustain itself. Representation, therefore, is not a property that distinguishes brains for other aspects of an organism. What distinguishes brains however, is the time-scale at which embodiment takes place. The emergence of a particular memory emerges and sustains itself at a much faster set of time scales than the time-scales by which individual neurons, neural nets, and entire brains emerge and sustain themselves. Regardless of this difference however, representation is there at every time-scale of self-sustaining work.

In addition to addressing representation in an ecologically-friendly way, WST also addresses computationalism's reliance on efficient cause as an explanation of content manipulation. Computationalism is led to efficient-cause by its assumption there exist specific levels in a cognitive architecture that are sufficiently isolated from other levels to enable them to 'bear' content. This assertion is proving increasingly difficult to defend as neuroscience provides more and more data indicating the immensely recursive, interconnected nature of neural organization. WST address this issue by conceptualizing neural dynamics in terms of multi-scale, contingent interactions. Given such embodiments are naturally and necessarily about the contexts they embody, WST encounters no need to pose sufficiently isolated 'vehicles' of content. Content is constitutive of what self-sustaining embodiments are. And conscious and cognition are not so much computational processes that take place in specific levels of a cognitive architecture, as they are emergent levels of self-sustaining work whose 'aboutness' cannot be reduced to any one level of work. Consciousness and cognition are irreducibly 'about' all such levels of work.

References

Bickhard, M. H. (2001). The emergence of contentful experience. In T. Kitamura (Ed.), *What should be computed to understand and model brain function?* Singapore: World Scientific.

Clark, A. (1997). *Being there: Putting brain, body, and world together again*. London: MIT Press.

Edelman, G. M. (1989). *Neural Darwinism: The theory of group neuronal selection*. Oxford: Oxford University Press.

Ghin, M. (2005, June). What a self could be. *Psyche*, *11(5)*. Online: http://psyche.cs.monash.edu.au/symposia/metzinger/Marcello.pdf

Grush, R. (2004). The emulation theory of representation: motor control, imagery, and perception. *Behavioral and Brain Sciences, 27,* 377-442.

Hebb, D. O. (1949). *The organization of behavior: A neuropsychological theory*. New York: Wiley.

Jordan, J.S. (1998). Recasting Dewey's critique of the reflex-arc concept via a theory of anticipatory consciousness: Implications for theories of perception. *New Ideas in Psychology*, *16(3),* 165-187.

Jordan, J. S. (2003a). The embodiment of intentionality. In

W. Tschacher & J. Dauwalder (Eds.), *Dynamical systems approaches to embodied cognition*. *Berlin:* Springer Verlag.

Jordan, J.S. (2003b). Consciousness on the edge: The intentional nature of experience. *Science and Consciousness Review* (December, No.1). Online serial, URL:http://www.scicon.org/news/articles/20040101.html

Jordan, J. S. (2003c). Emergence of self and other in perception and action. *Consciousness and Cognition*, *12*, 633-646.

Jordan, J. S., & Ghin, M. (2006). (Proto-) consciousness as a contextually emergent property of self-sustaining systems. *Mind & Matter, 4(1),* 45-68.

Jordan, J. S., & Knoblich, G. (2004). Spatial perception and control. *Psychonomic Bulletin and Review, 11(1)*, 54-59.

Kauffman, S. (1995). *At home in the universe*. New York: Oxford University Press.

Kinsbourne, M. (2002). The role of imitation in body ownership and mental growth. In A. Meltzoff and W. Prinz (Eds.), The imitative mind. New York, Oxford: Oxford University Press.

Knoblich, G., & Jordan, J. S. (2003). Action coordination in groups and individuals: Learning anticipatory control. Journal of Experimental Psychology: Learning, Memory, and Cognition, 29(5), 1006-1016.

Lotka, A. J. (1945). The law of evolution as a maximal principle. *Human Biology, 17*, 167-194.

Metzinger, T. (2003). *Being no one. The self-model theory of subjectivity*. The MIT Press.

Odum, H. T. (1988). Self-organization, transformity, and information. *Science*, *242*, 132-1139.

Rizzolatti G., Fadiga L., Fogassi L., Gallese V. (2002). *From mirror neurons to imitation: facts and speculations.* In A. M: The Imitative Mind Development, Evolution and Brain Bases. A. Meltzoff & W. Prinz (Eds.), Cambridge: Cambridge University Press.

Vandervert, L. (1995). Chaos theory and the evolution of consciousness and mind: A thermodynamic-holographic resolution to the mind-body problem. *New Ideas in Psychology, 13(2),* 107-127.

Framing and Resource Activation: Bridging the Cognitive-Situative Divide Using a Dynamic Unit of Cognitive Analysis

Luke David Conlin (luke.conlin@gmail.com)
Department of Curriculum & Instruction, 2226 Benjamin Building
College Park, MD 20742 USA

Ayush Gupta (ayush@umd.edu)
Department of Physics, Toll Building
College Park, MD 20742 USA

David Hammer (davidham@umd.edu)
Departments of Physics and Curriculum & Instruction, Toll Building
College Park, MD 20742 USA

Abstract

Theory in cognitive science often splits into those who treat cognition as occurring in individual minds those who treat it as situated or distributed, as irreducibly a matter of an individual-in-a-setting or of multiple individuals and artifacts. Prominent accounts have treated this split as between incommensurable paradigms (Sfard, 1998), competing theories (Greeno, 1997), and as complementary perspectives (Cobb, 1994). In the present paper, however, we argue that the accounts can be seen as theoretically continuous, differing in the scale of dynamics, such that a "society of mind" (Minsky, 1988) model of individual cognition is theoretically continuous with a "mind of society" model of social cognition. We sketch our framework and show how it leads to this continuity. We also argue that the relevant scale in any instance should be guided by the evidence, rather than based on purely *a priori* commitments.

Keywords: Modeling cognition; situated cognition; distributed cognition; resources; framing, education, collaborative learning

Theoretical Backdrop

Cognitive science has undergone dramatic advances that have forced us to question our basic assumptions of the nature of mind and its relation to the world. This progress has followed a path analogous to the conceptual changes in astronomy over the centuries. As astronomers have extended their gaze outward into the cosmos, they have revolutionized our view of the world and our place in it. These revolutions have been patently decentralizing—the Copernican revolution displaced the Earth from the center of the universe, and Einstein's cosmology went so far as to remove the very concept of 'center' from the universe.

A similarly decentralizing pattern of revolutions has also been the fruit of our gaze inward, using the tools and trade of cognitive science. While ancient views of consciousness assumed a central role for the heart, neuroscience has followed Hippocrates in focusing on the brain as lexis of mental life (Finger, 2001). Descartes in particular placed the "center of consciousness" squarely between the ears by postulating that the connection between spirit and body occurs in the pineal gland near the center of the brain. Modern cognitive science has shown that Descartes was wrong not only about the function of the pineal gland, but that the very concept of a 'center' can apply to consciousness and cognition—there is apparently no single place or time in the brain where it all 'comes together' (Dennett & Weiner, 1993). Vision provides a case in point: we have moved away from the assumption that the visual cortex functions something like a neural correspondence of our visual field, finding instead that vision is hierarchically distributed over various parts of the brain (Felleman & Van Essen, 1991)[1].

This decentralized view of mind has been highlighted by researchers working within the traditions of situated and distributed cognition. Situativity theorists claim that cognition cannot be defined apart from the situation in which it takes place and so take the appropriate unit of analysis the individual-in-a-setting (Greeno, 1997; Greeno & Moore, 1993; Lave, Murtaugh, & de la Rocha, 1984). In a commonly cited example, Lave et al. (1984) argued that whether or not a person knows how to find 3/4 of 2/3 a cup of cottage cheese depends critically on how the person takes up the affordances of the situation at hand; whereas the person may be unable to solve the problem via manipulation of symbolic fractions, they may still get the correct result by manipulating the physical objects. Theorists of distributed cognition have decentralized the mind even further by considering how information processing can be distributed across multiple individuals as well as artifacts. Hutchins (1995) has detailed a paradigmatic example by arguing that it is the cockpit—not any individual pilot—that remembers the safe landing speed of an airplane.

[1] Even if one of the area of cortical 'projection' is damaged, so that a blindsight patient reports seeing nothing at all, their 'visual location' capabilities can be quite intact, as evidenced by their ability to 'guess' well above the level of chance where an object is in their field of 'vision'.

Meanwhile, researchers in the 'cognitivist' tradition have resisted extending the border of cognition past the most intuitively obvious one—the brain. Where we draw the line around cognition has important consequences for how educational research is carried out, the conclusions we can draw from such research, and the recommendations we can then provide to practitioners. Anderson, Reder, and Simon (1996), for example, argued that the educational implications of situated theories of learning are often misguided. They advocated for the importance of training by abstraction, in contrast to training purely through concrete examples as situated theories would seem to favor. In his counter, Greeno (1997) took issue with this characterization but did point out a specific instructional consequence of situated cognition: teaching algorithmic skills is insufficient for achieving one of the main goals of education, namely getting students to "reason successfully in their everyday activity outside of school" (p. 7).

The cognitivist, situated, and distributed perspectives appear to have drastically different ontologies of mind. After all, there seems to be a vast ontological divide between claiming that it is a person who is remembering, rather than a cockpit. Such conceptual differences have contributed to the miscommunication between these camps, as several researchers have noted (e.g., Greeno, 1997; Sfard, 1998).

In this paper, we sketch a framework for cognitive analysis that has the potential to bridge these major ontological rifts in cognitive science. This is afforded, in part, by the dynamic unit of cognitive analysis we adopt in our model. We suggest heuristics for basing the unit of analysis on the data, rather than prescribing the cognitive unit based purely on theoretical commitments. Our account thus has the potential to unify or coordinate these perspectives.

Our Theoretical Framework

We work from a view of mind as a complex, dynamic system involving manifold cognitive resources, a generalization in line with schema theory (Bartlett, 1932, Rumelhart, 1980), Minsky's (1988) "society of mind" in which cognition is distributed within the mind across manifold "agents," and diSessa's "knowledge in pieces" (1993). "Resources" is a generic term for cognitive elements at various grain sizes that may be in different states of activation at any given moment (Hammer, et. al. 2005). For example, a student might explain the motion of a ball tossed into the air by saying it slows down as the force from your hand 'dies away,' but a moment later claim that it stops at the top of the trajectory because gravity has exactly balanced by the force from your toss[2]. Rather than assume the student is utterly confused, we find it productive to explain the dynamics of reasoning in terms of activation of fine-grained cognitive elements – "dying away" in one instance and "balancing" in the other (diSessa, 1993) – and the contextual features that cue these different resources. On this view, the phenomenology of reasoning is understood in terms of the activations of resources, of which there must be many kinds, including conceptual resources such as for understanding causal mechanisms (diSessa, 1993) or mathematical expressions (Sherin, 2001), as well as epistemological resources (Hammer & Elby, 2002), which will be of more central concern here. Resources often activate in stable patterns, and in what follows we will be concerned with the dynamics and patterns of resource activations, in particular with what the evidence suggests is involved in their formation and stabilities.

We refer to these patterns as "frames," (Hammer, et. al. 2005), building from accounts in the literature of frames as structures of expectation (Bateson, 1955; Minsky, 1988; Tannen, 1993) that undergird our sense of "what is it that is going on here" (Goffman & Berger, 1974). In the analyses below, we focus on the dynamics of how students, as individuals or as groups, frame what they are doing primarily with respect to knowledge, which we refer to as *epistemological framing* (Redish, 2004).

Phenomenological and ontological views of framing Describing a frame as a sense of 'what is going on' may be called a phenomenological view of framing. Most accounts in the literature on framing are phenomenological, focused on evidence of how individuals or groups understand what is taking place, as well as how individuals send "metamessages" (Bateson, 1955; Redish, 2004) to signal how they are framing the situation, in order to help each other interpret the accompanying message. For instance, a student who uses a rising intonation while offering an idea may convey more uncertainty than if they had delivered the idea with a falling intonation (Ward & Hirschberg, 1985).

Our account also incorporates an ontological view of framing by describing frames in terms of coherent activation patterns of resources. For instance, Rosenberg, Hammer, & Phelan (2006) found that when students framed their discussion of the rock cycle as "storytelling" they stably activated a set of epistemological resources including 'knowledge as fabricated stuff', 'knowledge as mental imagery', and 'knowledge as connectable through causal relations.'

Dynamics of framing. The phenomenological accounts in the literature cited above emphasize the dynamic nature of framing—Tannen (1993) prefers the gerund to emphasize the dynamic process, citing Bartlett's account of schemas as "active organized settings." The ontological view suggests models of framings as emergent patterns in a complex system. We may ask, then, what contributes to the dynamics of the system?

We suggest that the stability of a framing, as a pattern of activations, may just as easily involve manifold resources within an individual mind as across minds or across minds

[2] Phenomenological primitives (DiSessa, 1993) are examples of resources, but this by no means exhausts the set nor scale of resources.

and materials. That is, given an ontology of mind as comprised of manifold resources—a society of agents or a complex system of conceptual primitives—it is natural to expect dynamics that involve particular resources of one mind interacting with particular resources of others. To put this succinctly, a "society of mind" view of individuals (Minsky, 1988) should be consistent and continuous with a "mind of society" view of social cognition. It is a question of the scale of the relevant system (or subsystem) that is involved in the particular phenomena under study.

Thus we look for evidence of what contributes to the dynamics, and we expect that the relevant unit of analysis may vary from the individual (or perhaps even smaller) to much larger groups. Here, we limit ourselves to groups of four. We look for evidence, as we elaborate below, in the data for the scale of the dynamics involved for any particular instance.

Dynamic Unit of Analysis

Since both resources and frames exist at many different grain sizes, and may be activated on many levels at once, it makes little sense to limit our empirical studies to one level of analysis. Roth (2001) has also argued for the need to dynamically focus on multiple 'zoom' levels while analyzing cognition, and has provided some of the epistemological justification for doing so. Mandelblit & Zachar (1998) have laid out ontological considerations that allow for a dynamic unit of analysis, and have discussed how such a tack may be useful in bridging disparate traditions in cognitive science.

Epistemological considerations One good reason to seek out a dynamic unit of analysis is to avoid the temptation of doing *a priori* science. By rigidly adhering to only one cognitive unit, we may be effectively telling the world how it ought to be. If the individual is the unit of cognition, this is something that should be empirically supported, not just theoretically presumed.

Perhaps the gravest risk of such myopia is that of missing salient data. We all know that our perceptions are contingent on our attention. So if we focus our attention merely on the individual as the cognitive unit, we risk missing critical data relevant not only to the behavior of that individual, but also the group or situation of which she is a part (e.g. a jury in deliberation, a romantic couple in an argument, or a group of students working on a problem). As Roth (2001) puts it, "[b]y changing focus and by zooming, phenomena pertaining to different fields of attention become visible and are of different grain sizes and time scales" (p. 55).

Ontological considerations In motivating the concept of a dynamic unit of analysis, Mandelblit & Zachar (1998) describe several varieties of fundamental unity. Each of these various forms of unity "is formed under *different environmental restrictions* and is characterized by *different patterns of correlation*" (p. 234, emphasis in original). Physics provides many illustrative examples: The electron, for instance, is considered a spatially integrated unit in some circumstances (e.g. a point charge, or a small sphere of charge), but becomes an inseparable part of a dynamically integrated unit called a "Cooper pair" within a superconductor. Although such an ontological commitment violates some of our intuitions about what an "object" is, it is underwritten by the explanatory and predictive success of the BCS theory of superconductivity.

A dynamic unit of analysis also has explanatory and predictive power in the social sciences. It is often noted that people can form groups that are more (or less) than the sum of their parts, and although this may sound like mere rhetoric, it becomes a matter of practical significance when considering the differences between how individuals and groups act and make decisions. That crowds behave coherently as a unit and in ways that differ substantially from how the individuals that comprise them might otherwise act has long been noted (see McPhail & Wohlstein, 1983 for a review), and has important consequences in many areas including, for example, fire safety (Cocking & Drury, 2008). Research on small groups has found important differences between how individuals and groups make decisions, something that has important consequences for some of our most influential decision makers, such as juries. Studies of simulated juries suggest that juries are, as Moscovici & Doise (1994) have put it, "something other than a dozen jurors" (p. 110) since they polarize towards the majority opinion regardless of what that opinion is (e.g., Myers & Kaplan, 1976). Although such research is far removed from our own work, it does highlight the need for a way of incorporating multiple units of analysis into a theoretical framework of decision-making, behaviors, and cognition.

Empirical considerations Our empirical work has led us to posit a set of heuristics for identifying the cognitive unit, which is to say the scale at which the evidence suggests the dynamics of framing occur: *clustering*, *persistence*, *resistance*, and *transitions*. Each of these guides us in making a reliable identification of the unit of cognitive analysis at various grain sizes and time scales. We describe these heuristics in greater detail elsewhere (Conlin, Gupta, & Hammer, forthcoming).

Scherr & Hammer (2009) provides an illustrative example of the work that motivated these heuristics. They found that in small student groups working on physics tutorials, various behaviors tended to cluster together both within and across the students. They identified four distinct clusters, which were sufficient to account for most of the time spent in tutorial. These clusters can be stable for several minutes on the level of the student group. Scherr & Hammer also provided instances in which a cluster was resilient to bids from students to change clusters. The groups, when they did transition, tended to do so abruptly and synchronously. These clusters and the timing of the transitions were coded

with over 90% inter rater reliability, within 5 seconds accuracy.

These four behavioral clusters indicated four distinct epistemological frames (Scherr & Hammer, 2009). One frame corresponded with disproportionate quality of evidence for a measure of scientific reasoning (Conlin, Gupta, Scherr, & Hammer, 2009,). We will now offer two brief analyses of video data from these tutorials in order to illustrate the utility of having a dynamic unit of cognitive analysis.

Data & Discussion

The data comes from an algebra-based introductory physics course in which the students participated in worksheet-guided inquiry discussions (i.e., 'tutorials'). The students were mostly pre-med majors, and the worksheets focused on conceptual and epistemological issues in physics.

The students get many conflicting metamessages from the tutorials—messages about how to interpret what sort of activity they are engaged in and how to act accordingly. For example, students are given a worksheet, and this document can be framed in many contrasting ways. For instance, they may see the worksheet as "something to be completed," an interpretation they have long associated with worksheets in their school experience. On the other hand, they may see them as "something to guide them through their discussion," which was explicitly encouraged in several ways. One metamessage meant to encourage such a framing is the seating arrangement: there are four stools placed around a table so that the students faced inward, which is a common way of setting up a classroom for a discussion.

The tension between these alternate interpretations is typically never resolved once-and-for-all by the students. Rather, what we have found is that their behaviors indicate that their framing of the tutorial changes over multiple time scales—over the course of a few minutes, or over the whole hour of tutorial, or over the course of the semester. We have focused primarily on the minute-to-minute dynamics in framing.

Clustering heuristic applied to the individual

Throughout the course of the tutorial, the students exhibit a range of behaviors. It has been observed that a small set of behaviors tend to cluster together for each individual student in the tutorial. For instance, a student's gaze angle, hand position, and posture do not vary independently from each other but rather consistently cluster together in a few distinct sets. Two such sets are depicted in Figure 1. A downward gaze tends to cluster with hands on the table (often writing or resting) and a hunched-over posture (Fig 1a), while a horizontal gaze angle clusters with hands off the table (often gesturing) and an upright posture.

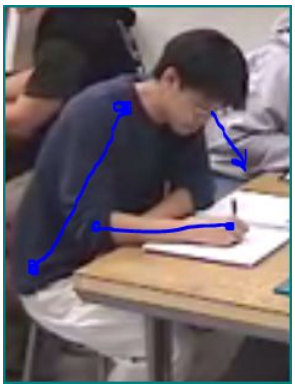
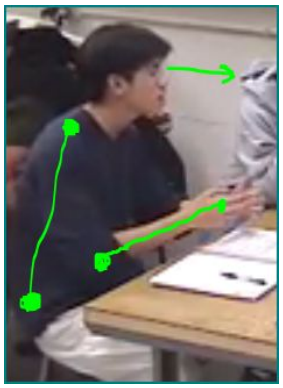
(a)  (b) 

Figure 1: (a) and (b) Two different behavior clusters for an individual

Clustering heuristic applied to the group

The same clusters of behaviors that are found on the individual level also are found at the group level. In fact, it was at the group level the behavioral clusters first drew Scherr's attention via abrupt and synchronous transitions by the group from one cluster to another. The clusters persist across individuals from tens to hundreds of seconds and just four distinct group-level behavioral clusters were enough to account for about 86% of time spent in a single tutorial session.

The tutorial groups' behavioral clusters serve as a robust and reliable indicator of the group's framing of the activity. There is a high degree of inter rater reliability (95% on the cluster code, 90% on the timing of the transitions). The coding is done without a transcript[3] and the analysis of discourse confirms the nature of the frame. The fact that the group spends most of the tutorial transitioning back and forth synchronously between the same set of activities indicates that it is appropriate to take the group (as well as the individual students) as the unit of analysis.

In what follows, we present two cases from our corpus of data and analyze them in light of our empirical heuristics. The first case supports taking the group to be the unit of cognitive analysis, while the second does not.

Case of group level cognitive analysis

This case comes from a tutorial on Newton's third law, during which the students are to find the speed a car gains when hit by a truck of twice the mass that loses 5m/s. In the first part of clip, the students are all looking down, so there is clustering of gaze angle across students. They are also hunched over, speaking softly, with their hands on their desks and their eyes on their worksheet. This is what Scherr and Hammer (2009) called the *blue* behavioral cluster (Fig 2a).

There is a sharp transition in behavior, in which the students all sit up, make eye contact, use animated voices, and gesture prolifically. This is what Scherr & Hammer (2009) called the *green* behavioral cluster (Fig 2b).

[3] The coding can be reliably done without even listening to the content of the speech.

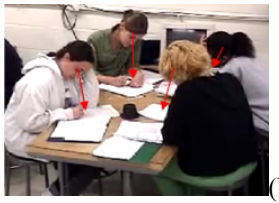
(a)  (b) 

Figure 2: The *blue* (a) and *green* (b) behavioral clusters.

Analysis of the group's discourse also falls in line with this transition. While in the blue cluster, the students are making intuitive guesses of the answer to a tutorial question (e.g., "Car speeds up by five"), with little or no justification provided. Along with the transition to the green behavioral cluster comes a corresponding transition in the substance of their discourse. They begin to describe the mechanism at work in the physical situation described in the worksheet question, as evidenced by metaphorical gestures of the collision as well as an analysis of the group's mechanistic reasoning (Russ, Scherr, Hammer, & Mikeska, 2008). When taking the behavior and discourse in conjunction, it becomes apparent that the group as a whole is changing activities from what might be called *completing the worksheet* to one of *having a discussion*. This transition also comprises a shift in activated conceptual and epistemological resources that are distributed across individuals, such that the activities of *completing the worksheet* and *having a discussion* are frames definable at the group level.

Case of individual level cognitive analysis

A contrasting example comes from a different group, working on a shadows and light tutorial, in which they are asked whether the light made by a bulb shining through a through an aperture onto a screen will move up or down when the bulb is lifted, and why (Lising & Elby, 2005).

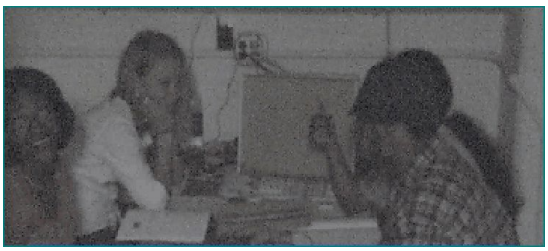


Figure 3: Lack of clustering of behaviors across students.

In this clip, there is no cohesive clustering of behaviors across students, and there is a lack of cohesion in their speech. Although their discourse centers on the same conceptual content , they are at this moment engaged in very different epistemological activities.

One student, Veronica, provides an intuitive explanation for why the light would go down as the bulb goes up, using gestures and colloquial speech. Another student, Jan, provides an 'explanation' that amounts to a gerrymandered list of physics vocabulary. When Veronica objects, "you're making it too complicated," Jan explains that she is "just trying to make it more physics oriented." Veronica retorts, "It *is* physics oriented. It's just how it is." Even though they both report taking part in a 'physics oriented' activity, through their activities and speech they express very different notions of what 'physics oriented' entails. For Veronica it apparently means explaining 'how it is,' while Jan thinks using words like "vectors" and "polarized" make it more 'physics oriented.' Their individual behaviors cluster with individualized epistemological and conceptual stances, and thus do not warrant a group level of analysis (for this interaction).

A Common Basis for Cognition in Action

There has been disagreement over the nature of the distinction between cognitivist, situated, and distributed accounts of cognition. This disagreement has fueled debate over how the debate can be settled, whether it can be settled, and even whether it *should* be resolved. While Anderson, Reder, and Simon (1996) have suggested the debate largely concerns the use of language, Greeno (1997) has contended that the issue can be settled as it becomes clear which tradition is better equipped for doing productive empirical work.

Others have argued that the distinction between cognitive and situated accounts of cognition lie with their preferred metaphors for learning. According to Sfard (1998) cognitivists follow a long tradition of viewing learning as an *acquisition* of knowledge, while situativity theorist view learning as an evolution of *participation* within a community of knowing. Rather than resolve their apparently incompatible ontological claims, she argues that they should be considered incommensurable and complementary. She thereby advocates for the peaceful coexistence of the paradigms, since "empirical evidence is unlikely to serve as an effective weapon in paradigm wars" (1998, p. 12).

We argue that our alternative account affords an ontological continuity between the cognitivist and situated/distributed traditions. Thus, in our account we can avoid the metaphorical paradigm war by distilling the choice of metaphor to an empirically informed decision about the unit of analysis. We therefore avoid surrendering to incommensurability, which if taken seriously leads to formidable methodological problems (and if taken *too* seriously descends into naïve relativism). Cobb and Bowers (1999) have also noted the need for a common basis for communication between these paradigms in order to avoid methodological problems. We hope that our account will provide such a basis, since it is founded upon established theories of cognition and is compatible with the connectionist principles that undergird both sides of the cognitivist/situativist divide.

Conclusion

We have described an account of cognition, in terms of resources & framing (Hammer, et. al., 2005), that provides

an ontological and epistemological basis for connecting these traditions within cognitive science. This connection is made possible by adopting a dynamic unit of analysis that can be grounded in the data, rather than based on entrenched theoretical commitments. We have provided empirical heuristics for assessing the unit of analysis. Finally, we have shown two contrasting empirical analyses to demonstrate the empirical nature of the unit of analysis as afforded by the resources & framing account.

One of the most remarkable aspects of cognition that science has uncovered is its decentralized nature—we have learned that there is no one place where our perception, thought, and conscious experience all 'come together.' Given the decentralized, distributed, and contextually sensitive functioning of the brain during cognition, it is not such a stretch to extend the distributed nature of cognition past the skull and into the surrounding environment. Although this may seem counterintuitive, the empirical and theoretical gains made by doing so may warrant the refinement of that persistent intuition that our minds reside in—and are confined to—our heads.

References

Anderson, J. R., Reder, L. M., & Simon, H. A. (1996). Situated learning and education. *Educational researcher*, *25*(4), 5.

Bateson, G. (1955). A theory of play and fantasy. *Psychiatric research reports*, *2*(39), 39–51.

Cobb, P. (1994). Where is the mind? Constructivist and sociocultural perspectives on mathematical development. *Educational Researcher, 23*(7), 13-20.

Cobb, P., & Bowers, J. (1999). Cognitive and situated learning perspectives in theory and practice. *Educational researcher*, *28*(2), 4.

Cocking, C. , & Drury, J. (2008). The mass psychology of disasters and emergency evacuations: A research report and implications for practice. *Fire Safety, Technology and Management, 10,* 13-19.

Conlin, L. D., Gupta, A., & Hammer, D. (forthcoming). Where to find the mind: Identifying the scale of cognitive dynamics. *To appear in the Proceedings of the 9th International Conference of the Learning Sciences (ICLS), Chicago, IL.*

Conlin, L. D., Gupta, A., Scherr, R. E., & Hammer, D. (2009). The dynamics of students' behaviors and reasoning during collaborative physics tutorial sessions. *Red*, *24*(15), 4.

Dennett, D. C., & Weiner, P. (1993). *Consciousness explained*. London: Penguin Press.

diSessa, A. A. (1993). Toward an epistemology of physics. *Cognition and Instruction*, *10*(2), 105–225.

Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral cortex*, *1*(1), 1.

Finger, S. (2001). *Origins of neuroscience: a history of explorations into brain function*. New York: Oxford University Press, USA.

Goffman, E., & Berger, B. M. (1974). Frame analysis: An essay on the organization of experience. Cambridge, MA: Harvard University Press.

Greeno, J. G. (1997). On claims that answer the wrong questions. *Educational researcher*, *26*(1), 5.

Greeno, J. G., & Moore, J. L. (1993). Situativity and symbols: Response to Vera and Simon. *Cognitive Science: A Multidisciplinary Journal*, *17*(1), 49–59.

Hutchins, E. (1995). How a cockpit remembers its speeds. *Cognitive Science: A Multidisciplinary Journal*, *19*(3), 265–288.

Lave, J., Murtaugh, M., & de la Rocha, O. (1984). The dialectic of arithmetic in grocery shopping. *Everyday cognition: Its development in social context*, 67–94.

Lising, L., & Elby, A. (2005). The impact of epistemology on learning: a case study from introductory physics. *American Journal of Physics, 73*(4), 372-382.

Mandelblit, N., & Zachar, O. (1998). The notion of dynamic unit: Conceptual developments in cognitive science. *Cognitive Science: A Multidisciplinary Journal*, *22*(2), 229–268.

McPhail, C., & Wohlstein, R. T. (1983). Individual and collective behaviors within gatherings, demonstrations, and riots. *Annual Review of Sociology*, *9*(1), 579–600.

Minsky, M. (1988). *The society of mind*. New York: Simon and Schuster.

Moscovici, S., Doise, W., & Halls, W. D. (1994). *Conflict and consensus*. London: Sage Publications.

Myers, D. G., & Kaplan, M. F. (1976). Group-induced polarization in simulated juries. *Personality and Social Psychology Bulletin*, *2*(1), 63.

Rosenberg, Hammer, & Phelan, (2006). Multiple epistemological coherences in an eighth-grade discussion of the rock cycle. *The Journal of the Learning Sciences, 15*(2), 261-292.

Roth, W. M. (2001). Situating cognition. *Journal of the Learning Sciences*, *10*(1), 27–61.

Russ, R. S., Scherr, R. E., Hammer, D., & Mikeska, J. (2008). Recognizing mechanistic reasoning in student scientific inquiry: A framework for discourse analysis developed from philosophy of science. *Science Education*, *92*(3), 499.

Scherr, R. E., & Hammer, D. (2009). Student behavior and epistemological framing: Examples from collaborative active-learning activities in physics. *Cognition and Instruction*, *27*(2), 147–174.

Sfard, A. (1998). On two metaphors for learning and the dangers of choosing just one. *Educational researcher*, *27*(2), 4.

Tannen, D. (1993). *Framing in discourse*. New York: Oxford University Press, USA.

Ward, G., & Hirschberg, J., (1985). Implicating uncertainty: the pragmatics of fall-rise intonation. *Language, 61*(4), 747-776.

Modeling Personality and Individual Differences: The Approach-Avoid-Conflict Triad

Karl Fua[1,2]**, William Revelle**[1] **and Andrew Ortony**[1,2]
(karl.fua@gmail.com, {revelle,ortony}@northwestern.edu)
[1]Northwestern University, Evanston, IL 60208 USA
[2]Computational Cognition for Social Systems, Institute of High Performance Computing,
Agency for Science, Technology and Research, Singapore

Abstract

Personality is the unique patterning of affect, behavior, cognition and desires in individuals across time and situations. This patterning can occur on different information processing levels, specifically, the Reactive, Routine, and Reflective levels (Ortony et al., 2005), across these four domains. Reinforcement Sensitivity Theory (RST; Gray & McNaughton, 2000) provides a biological account of the functional subdivision of the reactive level into the approach, avoidance, and conflict systems. These systems differ in their sensitivities to different classes of cues, giving rise to personality differences. But, individuals also differ at the routine and reflective levels in terms of how they respond (cognitively, affectively, behaviorally and motivationally) to approach situations, avoidance situations, and internal conflicts. In this paper, we discuss how the approach-avoidance-conflict (AAC) triad can be used as a broad framework for incorporating personality and individual differences into theories of human cognitive architectures. We also present work in progress on a computer implementation of the AAC triad at the reactive level.

Keywords: Reinforcement Sensitivity Theory, Personality, Behavior, Affect, Motivation.

Introduction

A trolley is hurtling along a track toward five children, all of whom are tied to the track. Should you flip a switch to divert the trolley onto another track on which only one child is tied so that only one life is sacrificed instead of five?

The moral dilemma posed by this well-known "trolley problem" illustrates the presence of high level motivational conflicts that arise when incompatible goals and values are coactivated. In this case, one would want to save the five lives but at the same time avoid taking the life of another. The recognition that the conflict exists results in rumination and reasoning as an individual seeks to resolve the conflict. Internal conflicts of this kind often occur in social situations, as for example, when a person wants to approach a potential date while also wanting to avoid rejection. Such conflicts lead to indecisive behavior such as dithering between approach and shyly looking away, and paying greater attention to hints that might inform the individual if approach (or shying away) would be a more suitable course of action.

Important in the present context is the fact that individuals differ in how they perceive and weigh alternatives (taking a life versus not saving a life), and how they handle different (approach or avoidance) goals and conflicts. Reward oriented individuals are prone to engage in riskier behavior, such as brazenly approaching a potential date. In the same situation, punishment oriented (averse) individuals will be more likely to shy away for fear of rejection. Yet others, who are prone to indecisiveness, are likely to spend time ruminating about the pros and cons of approaching and avoiding. We believe that this patterning of affect, behavior, cognition and motivation occurs across all three information processing levels proposed by Ortony et al. (2005) in their (ONR) model, from the reactive (lowest) to the reflective (highest). For instance, chronically goal/reward oriented individuals tend to exhibit more pro-social behavior (e.g., attending lively parties, dating more often; Paunonen, 2003), and are biased toward speed (maximizing hits) rather than accuracy (avoiding misses) when completing simple reactive tasks (Higgins & Spiegel, 2004).

Many models of the human cognitive architecture have been proposed, for example in cognitive psychology (Anderson & Lebiere, 1998; Ortony et al., 2005; Broadbent, 1971), personality psychology (Carver et al., 2009; Revelle, 1993), and artificial intelligence (Sloman & Chrisley, 2005; Newell, 1990). Many of these architectures, such as H-CogAff, ACT-R, and Soar, are highly elaborated and have been studied in great detail. However, although personality is a key moderator of individuals' affect, behavior, cognition and motivations, there has been little effort to include an account of personality and individual differences in these architectures. On the other hand, computational models of personality focus on describing specific aspects (e.g., the motivational aspect; Read et al., 2010) of personality but not the systematic integration of personality in the broader framework of cognitive, affective, motivational, and behavioral processes.

In this paper, we argue that there are three main classes of sensitivities (sensitivity to cues for reward, cues for punishment, and internal conflict), related to approach, avoidance, and conflict resolution (AAC) respectively, and that the AAC framework has implications for how different cognitive processes interact with each other and how the mechanisms driving personality and individual differences can be modeled systematically. Behavioral/motivational processes are commonly studied in terms of approach and avoidance (e.g., Carver et al., 2009; Elliot & Thrash, 2002). An individual's sensitivities to cues for reward and for punishment refer to how that individual's approach and avoidance systems react to and learn from such cues. Inspired by RST, we propose that conflict resolution is a third component that should be considered in conjunction with the approach-avoidance pair, and that it is associated with the sensitivity to conflicts. We define sensitivity to conflicts as the threshold beyond

which incompatible behavioral tendencies activate the conflict system–a system that we take to be distinct from the approach and avoidance systems and that is responsible for triggering conflict resolution processes (e.g., information gathering and rumination). Differences in the three kinds of sensitivities underlie broad personality dimensions such as the Giant 3 (Extraversion-Neuroticism-Psychoticism; Eysenck et al., 1985) or the Big 5 (Openness, Neuroticism, Conscientiousness, Extraversion, Agreeableness; Costa & McCrae, 1992; Goldberg, 1990). The AAC triad allows us to study how personality arises in individuals and how it influences cognitive processes like strategizing or resolving dilemmas, enabling us to address questions such as why, in the same situation, the plans an extravert makes differ from those of an introvert?

We believe that a theory of human cognitive architecture should be capable of accommodating differences in the ways in which different individuals feel, want, think, and act. To this end, our current work examines the structure of the reactive, routine and reflective levels (in the ONR model) in terms of the AAC triad, and looks at how these structures influence the organization of different parameters in systems. Our goal here is to propose a general framework that augments existing architectures to help in thinking about personality, and to elucidate how high and low level processes interact with each other.

The Approach-Avoidance-Conflict (AAC) Triad: From the Reactive to the Reflective Level

The Reactive Level

According to the ONR model, at the lowest level, behavioral responses to environmental cues are immediate and reactive. Automatic responses like the instinctive flight at the sight of a predator belong to this level. Reinforcement Sensitivity Theory (RST; Gray & McNaughton, 2000) was originally developed as an animal model of fear and anxiety, and has also been extensively studied in personality psychology (Corr, 2008; Smillie et al., 2010). RST proposes three functionally distinct subsystems–the Behavioral Approach System (BAS), the Fight-Flight-Freeze System (FFFS) and the Behavioral Inhibition System (BIS), each responding to different classes of cues with different sensitivities. RST offers neurobiological evidence that low-level, rapid behavioral responses, which we think of as the reactive level, have, at least functionally, the approach-avoid-conflict triadic structure.

The Approach-Avoid-Conflict Triad in RST The BAS, FFFS and BIS handle approach, avoidance, and conflict respectively. The approach system (BAS) is associated with the dopamine system and reacts to cues for reward, and is implicated in the learning of reinforcing signals of reward. The reactivity of BAS is highly correlated with trait extraversion and an individual who has an overactive BAS is prone to exhibit impulsive approach behaviors toward hedonic rewards.

Similarly, the avoidance system (FFFS) handles cues for punishment. The FFFS is primarily associated with fear, panic and avoidance behaviors, resulting from the activation of the periaqueductal grey, medial hypothalamus and amygdala regions of the neural system. The avoidance system is specifically modulated by panicolytic (suppression) and panicogenic (stimulating) drugs; individuals with a high sensitivity to cues for punishment are susceptible to phobias and panic attacks.

A major part of RST focuses on the functions of the conflict system (BIS). This system is associated with the septo-hippocampal system (SHS) and its major role is to detect conflicts and trigger appropriate conflict resolution behavior. The BIS handles two forms of conflicts: conflicts in motivations, and conflicts in expectations. Motivation conflicts occur between or within the approach and avoidance systems. An example of an approach-avoid conflict is the desire to escape from a burning building conflicting with the desire to save a trapped loved one. Expectation conflicts occur either when a stimulus is detected but not expected (novelty) or when an expected stimulus is absent. Examples would be suddenly seeing a furtive shadow in your house at night (novelty), or turning on the lights expecting to see a burglar but seeing an empty room instead (absent expected stimulus). The BIS detects such expectation violations with a comparator (the CA3-comparator in the SHS) that compares the signal (presence or absence of an expectation) from the entorhinal cortex stream and the signal (presence or absence of an actual stimulus) from the medial septum. When conflicts are detected, the BIS sends inhibiting signals to the conflicting systems to inhibit prepotent responses, and triggers behaviors such as information gathering. Importantly, the BIS does not actually resolve conflicts but rather triggers potentially appropriate higher-level cognitive processes and behaviors to do the resolution. Unlike the FFFS, the BIS has been shown to be insensitive to panicolytics/panicogenics but instead responds to anxiolytic/anxiogenic drugs. The BIS is therefore a separate system that is specifically associated with anxiety (as opposed to fear that is associated with the FFFS) and is implicated in Generalized Anxiety Disorders. The BIS is also highly correlated with trait anxiety and neuroticism. At least from a functional standpoint, different individuals must possess different thresholds (sensitivities to conflict) for the activation of the conflict system, and these thresholds are independent of and exist in parallel with the sensitivities to cues for reward and punishment that reside in the approach and avoidance systems respectively.

The Routine Level

The routine level resides between the reactive and reflective levels and deals with habitual, routinized behaviors. It deals with expectations over a longer time span than the moment-to-moment activities in the reactive level. While the reactive level is concerned with cues and their immediate implications, such as hearing a gunshot close by and instinctively taking cover, the routine level deals with more sophisticated

expectations and implications of cues. For example, the series of actions one executes after making the decision to drive home–getting into one's car, putting the key into the ignition, and turning the key with a foot on the brake.

As in the reactive level, individual differences can be analyzed at the routine level in terms of approach, avoidance and conflicts. Consider individuals at a party. In this case, a conceivable routinized behavior is the act of approaching and talking to a stranger. Extraverts, having a high sensitivity to cues for reward, may have learned that a stimulating conversation is rewarding, and so tend to engage in such approach behaviors. On the other hand, individuals who are highly sensitive to punishing cues tend to be afraid of approaching others at parties (Costa & McCrae, 1980) and so tend to engage in routine avoidance behaviors such as staying away from large groups. Individuals who are sensitive to conflicts (who can also, independently, differ in their sensitivities to cues for reward and punishment) easily feel frustration or annoyance if a conversation turns out to be less stimulating than expected, or feel anxious and unsure when the conversation partner shows signs of boredom, prompting the individual to try even harder to make the conversation work. This latter case should be differentiated from ones in which an individual is very sensitive to cues for punishment, in which case the individual will likely back off and try to avoid conversation altogether.

The Reflective Level

Often known also as the deliberative level, the reflective level is the home of high-level cognitive processes such as planning and conscious reasoning. The reflective level functions as the overall executive control that 'oversees' the operation of the lower levels. However, we want to suggest that the reflective level also embodies the same triadic AAC structure with individual differences in the corresponding sensitivities.

Of course, appeal to the approach-avoidance dyad in studies of motivation is anything but a new idea. It can be found in numerous theories (see Elliot 1999 for review), and it is widely recognized that individuals differ in their sensitivities to reward and punishment on the reflective level (Carver & White, 1994; Torrubia et al., 2001). Although most such studies have been designed to assess aspects of RST (which is reactive) in humans, the items in instruments used to do this in fact tap into routine and mainly reflective level processes. For example, the Sensitivity to Punishment, Sensitivity to Reward questionnaire (SPSR; Torrubia et al., 2001), includes items such as

- Does the good prospect of obtaining money motivate you strongly to do some things?
- Do you often renounce your rights when you know you can avoid a quarrel with a person or an organization?
- Do you think a lot before complaining in a restaurant if your meal is not well prepared?

The behaviors that correspond to such items obviously involve very reflective processes. Similarly, Regulatory Focus theory (Higgins, 1997) shows that there are chronic individual differences in motivation. A promotion-focused individual is concerned with nurturance-related motivations and is therefore sensitive to cues for reward. In contrast, a prevention-focused individual focuses more on security related needs, resulting in a bias toward cues for punishment. Regulatory Focus theory has been studied in a wide variety of contexts including goal pursuit and moral judgments, indicating that at least the approach-avoidance structure and individual differences in sensitivities to different classes of cues do exist on the reflective level.

However, the conflict system and individual differences in sensitivity to conflicts have received much less attention than the approach-avoidance pair, even though it is just as important an aspect of the motivation system. As already mentioned, conflicts arise within and between the approach and avoidance systems, but an individual's sensitivity to conflicts is independent of the sensitivities to cues for reward and punishment. It might be tempting to equate sensitivity to conflict with sensitivity to cues for punishment, but as the sample item from the SPSR questionnaire about complaining in a restaurant indicates, there is a difference between experiencing the conflicting desires of seeking redress and avoiding potential unpleasantness with the restaurant staff, on the one hand, and simply having a high desire to avoid unpleasantness, on the other. The experience of conflict triggers rumination about the choice that the individual faces, whereas the desire to avoid the unpleasantness of a confrontation could have been avoided by the person just leaving when the service was bad. Functionally, sensitivity to conflicts is the threshold that determines *when* a conflict is experienced and produces a separate class of affective states and behaviors from those that result from simple approach or avoidance. The rumination, anxiety, and heart-wrenching despair that arise when one is forced to make choices are the products of internal conflicts and are not mere amalgams of behavior or affect produced in the approach-avoid systems. If one thinks of approach and avoidance tendencies as having different activation levels, then the sensitivity to conflicts is the threshold above which activation levels of incompatible motivations are experienced as internal conflict, while sensitivities to cues for reward and punishment influence how fast the respective activation levels change.

Connecting the Levels

The structure described in the previous sections provides a general framework for organizing personality parameters and for suggesting how these parameters might influence processes on the reactive, routine, and reflective levels. In appealing to the AAC structure we, of course, do not mean that, for example, a reflective level module should be split into three; we are certainly not proposing an approach-planning module, an avoidance-planning mod-

ule and a conflict-planning module, each associated with a distinct brain region. Rather, we are suggesting that there exist at least three broad classes of parameters which influence an individual's selection of and access to different classes of strategies, or memories, or knowledge. We argue that broad personality dimensions arise from systematic differences in the set points of these parameters and should be modeled as such in cognitive architectures.

The patterning of a person's sensitivities is consistent across the different levels of processing. So, for example, a person who is highly sensitive to cues for reward reacts and learns faster to cues for reward, engages more readily in habitual behaviors that he or she associates with reward, and values high level achievement goals. The consistency also implies that the relative relationships between the different sensitivities are preserved across levels. Therefore, a person who has a relatively higher sensitivity to cues for punishment than rewards will exhibit this difference in sensitivities across the three processing levels. The absolute magnitude of the sensitivities on each level can differ, but the relationships should remain consistent. Inconsistencies in the biases could explain behaviors that might be viewed as uncharacteristic of a person, as for example, when a person who typically values safety and security indulges in a spur-of-the-moment risky behavior such as reckless gambling.

Another consequence of the AAC structure is that similar systems on the three levels might be more tightly coupled than they are to others. That is, other things being equal, an activated approach system on the reactive level is more likely to cause responses in the approach systems on the routine and reflective levels, acting as a mechanism for the endogenous generation of related higher level goals and actions (e.g., Cacioppo et al., 1993). Consider the case of being the target of a scathing remark. The immediate reactive response might be to lash out and perhaps even to retaliate physically. However, because of the fear of reprisal and possible physical harm to oneself, the immediate response is suppressed. The reaction to the stimulus (insult) can trigger higher level processes, for example, to devise an elaborate plan for exacting revenge that acts to suppress the reactive level urge. The reverse also holds, where higher level goals and values, being more persistent, bias the perception of and sensitivity to different cues at the reactive and routine levels. An example is the cognitive bias that results from different task framing which influences actual task performance. Anxious students who want to do well, but are afraid of being seen as incompetent, perform better when the task is reframed to emphasize its difficulty (Born et al., 2002; Weiner & Schneider, 1971). In the case of the reframed task, the system that deals with wanting to avoid appearing incompetent is less activated because the task is perceived as being highly difficult anyway, and therefore reduces conflict with the approach system, resulting in lower state anxiety and allowing approach system behaviors (e.g., persisting in performing the task) to manifest themselves.

Implementing the AAC triad on the Reactive Level

Our prototype implementation of the AAC structure on the reactive level is inspired by RST and is combined with the Cues-Tendency-Action (CTA) re-parameterization of the Dynamics of Action model (Revelle, 1986; Atkinson & Birch, 1970). CTA models the dynamic interaction between cues and tendencies within and between the approach, avoidance and conflict systems. In particular, it models the interaction between conflicting tendencies and actions. It also includes the feedback mechanism for the interaction of consummatory actions with the behavioral tendencies. The hybrid RST-CTA architecture is shown in Figure 1.

The model is implemented on a set of virtual characters using the Twig animation system (Horswill, 2009). Screenshots from the simulation are shown in Figure 2. The focus of our simulation is the yellow child, who interacts with the red child, the ball, and the yellow adult–his parent. The yellow child perceives the other "objects" (i.e., the red child, ball, and adult) in the environment as stimuli. An input stimulus perceived by the agent (yellow child) is a tuple comprising the object, the object's action and the object's distance from the agent, in the form I = (name,action,distance). For example, I = (red child,play,2.3) indicates that the yellow child sees the red child playing 2.3 distance units away.

The expectation module (Figure 1a) uses the input I to form an expectation about what type of cue the stimulus is along four dimensions–reward (R+), non-reward (R-), punishment (P+) and non-punishment (P-). For instance, the red child is a cue for both reward (R+, playmate) and punishment (P+, aggressive child), the degree of which is scaled by his action (R+ is higher if the red child is being friendly) and distance (an aggressive child is less threatening if he is further away). The agent also uses the current stimulus (I) and information he has about its current action (A) to generate an expectation of what he should expect at the next moment. For example, a hostile red child is expected to approach the agent aggressively after issuing a threat (see Figure 2b) and the expected action will be flagged as highly punishing (P+), and if the agent runs back to the adult, he will expect to attain a certain amount of safety when he is close to the adult (P-).

The behavioral tendencies (Figure 1b) react to the input stimulus based on how the stimulus is evaluated along the four dimensions (R+, R-, P+, P-) with different sensitivities. A stimulus can cause changes to more than one behavioral tendency. In the case of the red child who is both a R+ and P+ to the yellow child, the presence of the red child activates the yellow child's tendencies to both approach and avoid. The module also responds to consummatory actions taken by the agent at a rate defined by the sensitivity S_{cons} (Figure 1h). For example, the act of playing with a ball is a consummatory action that reduces the tendency to continue playing.

The BIS module (Figure 1c) detects conflicts in behavioral tendencies and expectations, and responds by activating information-gathering behavior and inhibiting the conflicting

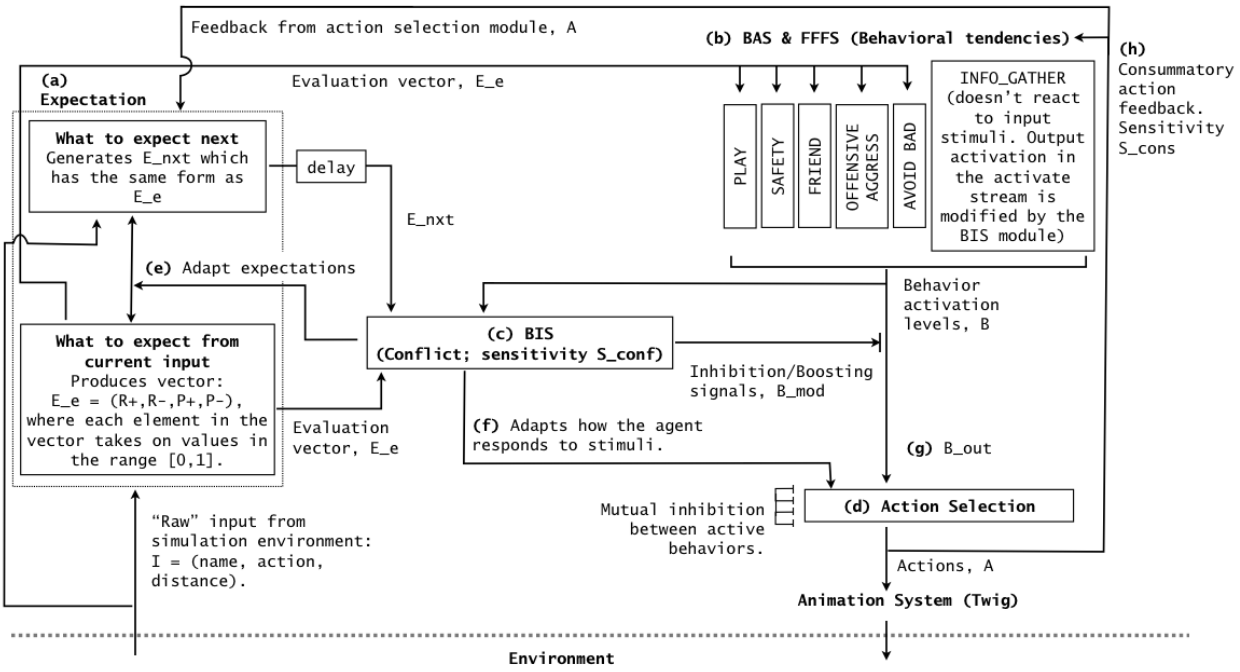


Figure 1: The RST-CTA model on the reactive level

tendencies (Figure 1g). In Figure 2a, the yellow child wants to approach the red child but is also afraid that the red child might hurt him, causing the yellow child to inhibit his actions and look nervously at the red child. A child with high sensitivity to conflict (trait anxiety) will pause more often. On the other hand, if the yellow child has low sensitivity to conflict and high sensitivity to reward (extraversion), he will engage in "riskier" behavior and approach the red child with less hesitation. The BIS module also sends learning signals to adapt expectations and actions (Figure 1e and 1f). For example, if the red child threatens the yellow child, the yellow child will expect physical aggression from the red child to follow. An expectation violation occurs when the red child does not follow up his threat with physical violence. This causes the yellow child to eventually learn that a vocal threat from the red child is less of a cue for punishment. If the red child is friendly instead of aggressive, a yellow child who is highly sensitive to reward will react more strongly to the R+ component of the stimulus and learn more quickly that the red child is much more of a cue for reward (Figure 2c).

The action selection module (Figure 1d) receives the activation levels of the behavioral tendencies and activates the actions corresponding to the behavioral tendency with the maximum activation level. The appropriate commands associated with the selected action, *A*, is passed to the Twig animation system.

Conclusion

Personality and individual differences are coherent patterns of thoughts, feelings, desires and behaviors within individuals. Inspired by RST, we have proposed that personality can be organized and investigated in terms of approach, avoidance, and, importantly, conflict on the reactive, routine, and reflective levels, where an individual is consistent in his/her sensitivity to different types of cues on all three levels. We believe that this organization is useful for examining the influence of personality on other cognitive processes such as planning and moral deliberation. We also hope that the proposed structure might be informative for applications that require integrating personality into AI systems, such as in interactive games and drama, and various kinds of simulations of, for example, interpersonal interactions.

References

Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought*. Mahwah, NJ: Erlbaum.

Atkinson, J. W., & Birch, D. (1970). *The dynamics of action*. New York: John Wiley.

Born, W. K., Revelle, W., & Pinto, L. H. (2002). Improving biology performance with workshop groups. *Journal of Science Education and Technology*, *11*, 347-365.

Broadbent, D. E. (1971). *Decision and stress*. London: Academic Press.

Cacioppo, J. T., Priester, J. R., & Berntson, G. G. (1993). Rudimentary determinants of attitudes. II: Arm flexion and extension have differential effects on attitudes. *Journal of Personality and Social Psychology*, *65*, 5-17.

Carver, C. S., Johnson, S. L., & Joorman, J. (2009). Two-mode models of self-regulation as a tool for conceptualizing effects of the serotonergic system in normal behavior



(a) A highly anxious yellow child (trait anxiety).

(b) The yellow child running away from the aggressive red child (fearfulness, panic).

(c) The reward sensitive yellow child interacting with the friendly red child (trait extraversion).

Figure 2: Screenshots from Twig showing some of the behaviors the yellow child can exhibit with different biases in his sensitivities to different environmental cues.

and diverse disorders. *Current Directions in Psychological Science*, *18*, 195-199.

Carver, C. S., & White, T. L. (1994). Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: The BIS/BAS scales. *Journal of Personality and Social Psychology*, *67*, 319-333.

Corr, P. J. (2008). *The reinforcement sensitivity theory of personality*. New York: Cambridge University Press.

Costa, P. T., & McCrae, R. R. (1980). Influence of extraversion and neuroticism on subjective well-being: Happy and unhappy people. *Journal of Personality and Social Psychology*, *38*(4), 668-678.

Costa, P. T., & McCrae, R. R. (1992). Four ways five factors are basic. *Personality and Individual Differences*, *13*(6), 653-665.

Elliot, A. J. (1999). Approach and avoidance motivation and achievement goals. *Educational Psychologist*, *34*, 149-169.

Elliot, A. J., & Thrash, T. M. (2002). Approach-Avoidance motivation in personality: approach avoidance temperaments and goals. *Journal of Personality and Social Pyschology*, *82*, 804-818.

Eysenck, S. B. G., Eysenck, H. J., & Barrett, P. (1985). A revised version of the pyschoticism scale. *Personality and Individual Differences*, *6*(1), 21-29.

Goldberg, L. R. (1990). An alternative "description of personality": The big-five factor structure. *Journal of Personality and Social Psychology*, *59*(6), 1216-1229.

Gray, J., & McNaughton, N. (2000). *The neuropsychology of anxiety*. New York: Oxford University Press.

Higgins, E. T. (1997). Beyond pleasure and pain. *American Psychologist*, *52*(12), 1280-1300.

Higgins, E. T., & Spiegel, S. (2004). Promotion and prevention strategies for self regulation. In R. F. Baumeister & K. D. Vohs (Eds.), *Handbook of self-regulation* (p. 171-187). New York: The Guilford Press.

Horswill, I. (2009). Lightweight Procedural Animation With Believable Physical Interactions. *IEEE Transactions on Computational Intelligence and AI in Games*, *1*(1), 39-49.

Newell, A. (1990). *Unified theories of cognition*. Cambridge, MA: Harvard University Press.

Ortony, A., Norman, D. A., & Revelle, W. (2005). Affect and proto-affect in effective functioning. In J. M. Fellous & M. A. Arbib (Eds.), *Who Needs Emotions? The Brain Meets the Machine* (p. 173-202). New York: Oxford University Press.

Paunonen, S. V. (2003). Big five factors of personality and replicated predictions of behavior. *Journal of Personality and Social Psychology*, *84*(2), 411-422.

Read, S., Monroe, B., Brownstein, A., Yang, Y., Chopra, G., & Miller, L. (2010). A neural network model of the structure and dynamics of human personality. *Psychological Review*, *117*(1), 61-92.

Revelle, W. (1986). Motivation and efficiency of cognitive performance. In D. R. Brown & J. Veroff (Eds.), *Frontiers of Motivational Psychology: Essays in honor of J. W. Atkinson* (p. 107-131). Berlin: Springer.

Revelle, W. (1993). Individual differences in personality and motivation: Non-cognitive determinants of cognitive performance. In A. Baddeley & L. Weiskrantz (Eds.), *Attention: Selection, awareness and control: A tribute to Donald Broadbent* (p. 346-373). Oxford: Oxford University Press.

Sloman, A., & Chrisley, R. L. (2005). More things than are dreamt of in your biology: Information-processing in biologically-inspired robots. *Cognitive Systems Research*, *6*(2), 145-174.

Smillie, L. D., Loxton, N. J., & Avery, R. E. (2010). Reinforcement Sensitivity Theory, Research, Applications and Future. In T. Chamorro-Premuzic, A. F. Furnham, & S. von Stumm (Eds.), *Handbook of Individual Differences.* London,UK: Wiley-Blackwell.

Torrubia, R., Ávila, C., Moltó, J., & Caseras, X. (2001). The sensitivity to punishment and sensitivity to reward questionnaire (SPSRQ) as a measure of Grays anxiety and impulsivity dimensions. *Personality and Individual Differences*, *31*, 837-862.

Weiner, B., & Schneider, K. (1971). Drive versus Cognitive Theory: A Reply to Boor and Harmon. *Journal of Personality and Social Psychology*, *18*(2), 258-262.

Understanding the Brain as an Endogenously Active Mechanism

William Bechtel (bill@mechanism.ucsd.edu)
Department of Philosophy, University of California, San Diego
La Jolla, CA 92093-0119 USA

Adele Abrahamsen (aabrahamsen@ucsd.edu)
Center for Research in Language, University of California, San Diego
La Jolla, CA 92093 USA

Abstract

Although a reactive framework has long been dominant in cognitive science and neuroscience, an alternative framework emphasizing dynamics and endogenous activity has recently gained prominence. We review some of the evidence for endogenous activity and consider the implications not only for understanding cognition but also for accounts of explanation offered by philosophers of science. Our recent characterization of *dynamic mechanistic explanation* emphasizes the coordination of accounts of mechanisms that identify parts and operations with computational models of their activity. These can, and should, be extended to incorporate attention to mechanisms that are not only active, but endogenously active.

Keywords: philosophy of science; mechanistic explanation; dynamics; endogenous brain activity, resting state fMRI, brain default network

Introduction

Observe a living organism, from a bacterium to a fellow human being, and you see an endogenously active system. Introspect and you will observe, as did William James, a continual flow of thoughts. If pressed, most cognitive scientists will acknowledge that neural systems—from individual neurons to the brain as a whole—exhibit endogenous activity. That is, some of the activity is internally (Greek *endo*) produced (German *gennan*); the causes and control of this activity is inside the system rather than reactive to inputs from outside the system. But cognitive scientists tend to disregard this when designing studies. Those in psychology present discrete stimuli in structured tasks designed to permit statistical analysis of the behavioral effects of independent variables. Those in neuroscience, following the tradition of Charles Scott Sherrington (1923), commonly treat the brain as a reactive system in which sensory inputs initiate neural processing that results ultimately in motor responses. They may stimulate specific neurons or provide sensory inputs with specific properties so that recorded neural activity can be analyzed in terms of responses to inputs. In both fields, variations in activity that cannot be associated with an input are treated as random fluctuations (noise). There is no doubt that this reactive framework in psychology and neuroscience has been enormously productive in identifying the parts, operations, and organization of the mechanisms responsible for cognition. It soon reaches its limits, though, in seeking accounts of the orchestrated functioning of those components: their dynamics and coordination in real time.

The investigation of endogenous activity, though less influential, has historical roots nearly as deep as those of the reactive approach. It was promoted by Thomas Graham Brown (1914), for example, who studied decerebrate and deafferented cats in Sherrington's laboratory at Liverpool from 1910 to 1913. He found that the isolated spinal cord, even when not receiving inputs, generates patterns of activity comparable to those exhibited during motor behavior elicited by stimuli. Brown's emphasis on endogenous activity initially was largely ignored (for discussion, see Stuart & Hultborn, 2008) but was revived several decades later when biologists recognized a class of neural circuits—*central pattern generators*—whose self-sustaining patterns of activity generated rhythmic motor behavior even in the absence of sensory input. After Wilson and Wyman (1965) pioneered this construct in their account of locust flight, others identified central pattern generators in the brain stem and spinal cord for walking, swimming, respiration, circulation, and other behaviors for which oscillatory control was crucial (Grillner, 2003). Endogenous activity has received far less attention from those studying sensory processing and central cognition rather than motor control, despite indications of endogenous oscillatory activity in cerebral cortex using techniques ranging from single cell recording to EEG and fMRI. In the next section we describe highlights from this research and in the subsequent section briefly explore the implications for reconstruing how we understand cognitive activity. Most important, if the conception of the brain as endogenously active is taken seriously, it profoundly challenges the reactive perspective that has dominated much of cognitive science as well as neuroscience: stimuli or tasks must be regarded not as initiating activity in an inactive system, but rather as perturbing endogenous dynamic behavior.

The slow pace at which these fields are achieving a change of perspective is unsurprising considering the history of other sciences. Although Max Planck was exaggerating when he said "A new scientific truth does not triumph by convincing its opponents . . . but rather because its opponents eventually die . . .," the considerable costs and uncertain benefits of change make it a tough sell. Uneven acceptance of Einstein's revolutionary proposals is a familiar example. Less remarked upon is the delayed impact of changes in the sciences on *philosophy* of science. For example, this young field (which did not even have a journal until 1934) did not exhibit acute concern with the epistemological foundations of science until it was confronted with Ein-

stein's proposals and their aftermath—a response that necessarily involved at least a short delay. However, delays in uptake have been far greater for developments in sciences other than physics, notably the biological and cognitive sciences. Philosophers of science did not even recognize the dominant mode of explanation in these sciences—mechanistic explanation—until the 1990s and especially after 2000. More recently, we have argued that such developments as computational modeling of the dynamics of cognitive and neural mechanisms require philosophers of science to extend their notion of mechanism to include *dynamic mechanistic explanation*. In the last section of this paper we will briefly characterize these two explanatory frameworks and consider how the philosophical understanding of dynamic mechanistic explanation can incorporate the implications of scientific work on endogenous activity.

Evidence that the Brain is Endogenously Active

Although lesion and stimulation techniques have been important in identifying brain regions involved in different cognitive activities, since the mid-20th century the greatest insights have come from techniques in which researchers record brain activity of individual neurons (single or multi-cell recording) or brain regions (EEG and fMRI). Most commonly these techniques have been employed within the reactive framework in which stimuli are presented or tasks are assigned, responses within the brain recorded, and these responses pooled for analysis to remove variability not associated with the intervention.

Each of these techniques, though, also has been employed in ways that reveal endogenous brain activity. Notably, Rodolfo Llinás employed intracellular recordings to identify systematic variations in the conductance of calcium ions across neural membranes. He showed how the manner in which these conductances varied through time enabled neurons in the inferior olive, a brainstem nucleus, to function as single-cell oscillators "capable of self-sustained rhythmic firing independent of synaptic input" (Llinás, 1988, p. 1659). (For a review of evidence and models showing how these intrinsic oscillations when combined with synaptic processes can generate synchronous thalamocortical oscillations, see Destexhe & Sejnowski, 2003.)

A second line of evidence for endogenous brain activity, consistent with that of single-cell recording, emerged from earlier studies by Hans Berger (1929) pioneering the identification of distinctive waveforms in electroencephalograph (EEG) recordings of brain activity. When he presented no stimuli or task demands but simply had subjects sit awake with their eyes closed, he obtained high-amplitude oscillations between 8 and 12 Hz that he dubbed *alpha waves*. When subjects instead viewed a stimulus or solved a problem, alpha waves were supplanted by lower-amplitude, higher-frequency *beta waves* (12-30 Hz). Soon thereafter it was determined that the EGG signal captured, not action potentials, but rather synchronized sub-threshold electrical potentials across a population of neurons. In the 1960s, the development of digital EEG and of powerful statistical techniques for decomposing complex EEG signals into component waveforms brought further discoveries; notably, very high-frequency (25-100 Hz) *gamma waves* were prominent in addition to beta waves when people performed various cognitive tasks. Moreover, synchronized oscillations at all of these frequencies were found in both active and passive conditions, but at different amplitudes.

Thus, both single-cell recording and EEG studies have provided evidence for endogenous brain activity. In this paper we will focus on yet another line of evidence offered by recent work on *resting-state fMRI*. The BOLD (blood oxygen level dependent) signal employed in fMRI research registers the oxygen concentrations in the brain within areas that can be as small as 2 mm. Until recently fMRI research focused nearly exclusively on finding higher values in the BOLD signal when a task condition is compared to a control or resting state condition.[1] For example, semantic processing of words (task condition) would be contrasted to reading words aloud (control condition) or to lying still in the scanner with eyes closed (resting condition). The interest in neuroimaging during a resting state, rather than during task performance, developed from researchers' occasional observations that a number of brain areas routinely exhibited less activity in task situations than in the resting state. To explore further these intriguing observations, Shulman et al. (1997) conducted a meta-analysis of studies in which a task condition was compared to a non-task condition in which the same stimulus was present. They found that the areas commonly less active in task situations included posterior cingulate cortex (PCC), precuneus, inferior parietal cortex (IPC), left dorsal lateral prefrontal cortex (left DLPFC), and a medial frontal strip that continued through the inferior anterior cingulate cortex (ACC), left inferior frontal cortex, and left inferior frontal gyrus to the right amygdala. Turning the focus from the fact that these areas are less active during tasks to the fact that they are more active in the absence of task requirements, Raichle and his collaborators (Raichle et al., 2001) suggested that together these areas constitute a *default network*.

A major advance in understanding the default network resulted from analyzing the temporal dynamics of the BOLD signal. A pioneering dynamical analysis of fMRI data was provided by Biswal, Yetkin, Haughton, and Hyde (1995), who obtained BOLD signal values every 250 msec after a hand movement and identified spontaneous low frequency

[1] In referring to resting states, the assumption is not that the subject's brain is resting, but that he or she is not engaged in a specific task or responding to a specific stimulus. Often the subject is asked to fixate on a cross-hair or lie still in the scanner with eyes closed but not asleep. Fluctuations in activity that can be linked to physiological activity (cardiac or respiratory activity) are eliminated from the data through linear regression. In a critique of this research, Morcom and Fletcher (2007) focused on the privileging of the resting state. The insights into the default network on which we focus, however, do not rely on the resting state being privileged but simply as revealing ongoing activity in brain networks not employed in cognitive tasks.

(less than 0.1 Hz) fluctuations in sensorimotor cortex. These fluctuations were synchronized across the left and right hemispheres and with those in other motor areas, which was interpreted as evidence of functional connectivity among all these areas. Accordingly, the approach is referred to as *functional connectivity MRI* (fcMRI).

Employing fcMRI, Greicius, Krasnow, Reiss, and Menon (2003) demonstrated that if they used the PCC as a seed for statistical analysis, they could identify synchronized fluctuations in a large cluster of areas: medial prefrontal cortex (including inferior ACC and orbitofrontal cortex), left DLPFC, inferior parietal cortex bilaterally, left inferolateral temporal cortex, and left parahippocampal gyrus. Taking instead the inferior ACC as the seed area, they found correlated fluctuations in the PCC, medial prefrontal cortex/orbital frontal cortex, the nucleus accumbens, and the hypothalamus/midbrain. Since these regions were virtually the same as those showing activity in Shulman's resting state data, Greicius et al. construed this as evidence for "a cohesive, tonically active, default mode network" (p. 256) with two subnetworks.

While the default network exhibits greater activity in the resting state than in task conditions, the areas showing greater activity in task conditions still generate a BOLD signal in the resting state and one can find correlations in the dynamics across these areas (synchronized oscillations). These synchronized oscillations are, however, out of phase with those in the default network. Comparing the default network with one that exhibited greater activation in an attention-demanding task (intraparietal sulcus, frontal eye field, middle temporal region, supplementary motor areas, and the insula), Fox et al. (2005) described oscillations in the two networks as anticorrelated, whereas oscillations for different areas within each network were positively correlated. This shows that both the default network and the network involved in attention-demanding tasks are coordinating their activities within themselves in the absence of external stimulation or task demands.

Researchers subsequently identified additional networks using this strategy. That is, a set of areas with correlated dynamics (synchronized oscillations) under resting state conditions were posited to constitute a network, further evidenced by negative correlations with other networks (e.g., Mantini, Perrucci, Del Gratta, Romani, & Corbetta, 2007, differentiate six anticorrelated networks). Fox and Raichle (2007) concluded: "A consistent finding is that regions with similar functionality—that is, regions that are similarly modulated by various task paradigms—tend to be correlated in their spontaneous BOLD activity."

Although the oscillations revealed in fMRI are of a much lower frequency (< 0.1 Hz) than those usually reported in EEG (1-80 Hz), researchers have found ways to relate them. Mantini et al., for example, found that "Each brain network was associated with a specific combination of EEG rhythms, a neurophysiological signature that constitutes a baseline for evaluating changes in oscillatory signals during active behavior" (p. 13170). For example, the default network showed positive correlations with amplitude in alpha and beta band oscillations while the attention network exhibited negative correlations in these frequency bands. These correlations may reflect systemic coherence in brain functioning. In the cortex of mammals, the amplitude (power density) of EEG oscillations has been found to be inversely proportional to their frequency ($1/f$). Even more interesting, the phase of lower-frequency oscillations seems to modulate the amplitude of those at higher frequencies, which results in a nesting relation between the frequency bands. (Lakatos et al., 2005, refer to this as "oscillatory hierarchy hypothesis") In addition, oscillations at lower frequencies tend to synchronize over more widely distributed areas of the brain than those at higher frequencies (Buzsáki & Draguhn, 2004). Such coupling can be particularly important when the brain is perturbed by a stimulus, since a modulation in low-frequency oscillations can, through phase-locking with higher-frequency oscillations, yield rapid changes at those frequencies.

The Significance of Endogenous Brain Activity for Understanding Cognition

One might acknowledge endogenous activity in various brain networks, but deny that it is of any cognitive significance. Perhaps it merely reflects basic metabolic activity and bears no implications for cognition. However, the fact that each network oscillates at a characteristic frequency, ,rather than fluctuating randomly, suggests that endogenous activity has implications for understanding brain activity generally—including activity during cognitive functioning. We briefly explore different ways in which endogenous activity may be important for understanding the brain as a system for cognition.

First, if a mechanism responds to a stimulus by increasing its activity, and that activity already is oscillating, response to the stimulus will vary depending on the phase of the oscillation when the stimulus arrives. This is true of individual neurons. If the membrane voltage of a neuron oscillates endogenously in a range below zero mV, as the evidence developed by Llinás and others indicates, then it will require stronger input to exceed the threshold for generating an action potential when it happens to be at its most negative phase. The same principle applies to populations of neurons whose oscillations are synchronized. In a variety of tasks in which a stimulus evokes a behavioral response, it is known that the response correlates with the magnitude of the BOLD signal. Fox, Snyder, Zacks, and Raichle (2005) therefore investigated whether these effects could be explained by synchronized spontaneous fluctuations in neuronal activity detectable with fMRI. Subjects were instructed to press a button with the right hand when a stimulus was detected, resulting in evoked activity in the left somatosensory cortex. The researchers hypothesized that the ongoing spontaneous fluctuations in the right somatosensory cortex provided an accurate measure of the spontaneous contribution to activity in the left somatosensory area at each timestep and succeeded in showing that these sponta-

neous fluctuations contributed significantly to the amplitude of blood flow in the left somatosensory areas after each stimulus. In fact, the task-related increased blood flow could be analyzed as a linear addition to the current amplitude of the spontaneous fluctuation. From this they inferred that the underlying spontaneous fluctuations affected perception and behavior. They supported this conclusion more directly in a subsequent study, in which they determined that spontaneous fluctuations accounted for variability in the force with which subjects pressed the button (Fox, Snyder, Vincent, & Raichle, 2007). When subjects were instructed as to how forcefully they should press the button, the pattern of neuronal activity was very different than that which arose when they were not instructed, allowing the investigators to discount the possibility that what they took to be spontaneous variability was in fact an evoked response. Thus, their study can be taken as initial evidence that the variability in endogenous brain activity is one source of the variability in measures of cognitive activity.

Second, endogenous activity in the brain's default network is the most obvious candidate for the neural underpinnings of *mindwandering* (Antrobus, Singer, Goldstein, & Fortgang, 1970). In one of the early fMRI studies using the resting state, Andreasen et al. (1995) queried subjects about what they were doing and elicited reports of being engaged in "a mixture of freely wandering past recollection, future plans, and other personal thoughts and experiences." Since these activities involve episodic memory, and episodic memory tasks are among those which do not lead to lower activity in the default network, Andreasen et al. and subsequent researchers (e.g., Buckner & Carroll, 2007) have suggested that the default network is involved in recalling personal experiences and anticipating future ones. Intriguingly, Li, Yan, Bergquist, and Sinha (2007) correlated trials on which subjects failed to detect stop signals in behavioral tasks with increased activity in the default network, as one would expect if that network were involved in a person thinking distracting thoughts about past and future experiences. One factor that renders problematic such a characterization of the activity of the default network is that the oscillatory behavior of the default network is maintained as well in sleep (Fukunaga et al., 2006) and under anesthesia (Vincent et al., 2007), when presumably spontaneous thoughts are not occurring.

Third, endogenous brain activity might be crucial for building and maintaining certain types of organization in the nervous system required for cognitive activity. There is growing evidence that the brain exhibits *small-world* organization (Watts & Strogratz, 1998) in which most connections link neighboring neurons, creating clusters that can collaborate in processing specific information, but a few long range connections enable overall coordination (Sporns & Zwi, 2004). There also is evidence that while most brain areas have connections to only a few other areas, some have a large number of connections, thereby constituting hubs. Such an architecture provides a highly efficient organization for information processing, and it is notable that the default network itself exhibits a small-world architecture with hubs. An important question is how such organization might arise. Rubinov, Sporns, van Leeuwen, and Breakspear (2009) advanced the intriguing possibility that oscillatory neurons, developing connections when synchronized, might self organize into a small world network with hubs. In support of this proposal they described a model by Gong and van Leeuwen (2004) that employs a logistic map activation function for individual units that endogenously exhibit chaotic behavior. This enables the emergence of temporary patterns of synchronized oscillations even in the absence of external stimulation. A Hebbian learning procedure establishes new connections between pairs of units whose activity is synchronized and prunes those between unsynchronized units. Even when these networks begin with random connectivity, they develop clusters linked to each other through hubs. However, in real brains the initial state already involves local regions with interconnections and experience further shapes the emerging organization such that the outcome is a highly correlated brain capable of maintaining multiple anticorrelated networks. That is, the architecture of the information processing system may be shaped by both endogenous and exogenous activity.

In this section we have considered three suggestions as to how endogenous activity in the brain may contribute to its functioning as a cognitive system. Although it is too early to judge which will prove most fruitful, clearly the time for dismissing endogenous activity as mere noise has passed.

Endogenously Activity and Mechanistic Explanation

The evidence for endogenous activity in brains presents challenges not only to the ways in which cognitive scientists understand cognitive activity but also to philosophers' construal of the explanatory frameworks used in science. We mentioned above that these construals lag behind the sciences, often far more than necessary. Until recently, philosophical accounts of explanation focused primarily on laws and construed explanation as the subsumption of phenomena to be explained under these laws. While such an approach might work in physics, where there are many well established laws, it does not characterize explanations in the life sciences, where there are few laws but an abundance of phenomena to be explained (Cummins, 2000). What form of explanation is appropriate? In the past 20 years a number of philosophers of science have finally paid attention to biologists and, following their lead, construed explanation as the characterization of the mechanism responsible for a phenomenon of interest (Bechtel & Richardson, 1993; Bechtel & Abrahamsen, 2005; Machamer, Darden, & Craver, 2000; Thagard, 2006).

Although there are minor differences among these various accounts of mechanistic explanation, they concur in construing a mechanism as consisting of component parts, each of which performs one or more operations. Each operation produces change in another part that triggers or affects the operation of that part, and so forth. Cognitive psychologists,

traditionally have posited operations that transform, copy, or move representations without localizing them in parts of the brain. Cognitive neuroscientists (and growing numbers of cognitive psychologists) emphasize localization and choose operations at the appropriate grain for their brain recording technology (Bechtel, 2008).

Given the focus on specifying a mechanism to explain a given phenomenon, it is natural to conceive of the mechanism as having a specific beginning condition and continuing its operations until its task is completed. This sequential conception of mechanism is most clearly captured in the definition offered by Machamer, Darden, and Craver (2000): "Mechanisms are entities [parts] and activities [operations] organized such that they are productive of regular changes from start or set-up to finish or termination conditions." If the start or set up conditions involve a stimulus or task originating from outside the mechanism, we arrive at the construal of a mechanism not only as sequential but also as reactive.

This reactive conception of a mechanism accords well with the accounts offered in many areas of biology and cognitive science, but it is not adequate to characterize endogenously active systems as discussed in the previous sections. A sequentially organized mechanism will not exhibit endogenous activity. A minimal first step towards a mechanism capable of endogenous activity retains the general sequential conception of the overall functioning of the mechanism but allows operations that are viewed as later in the sequential order to feed back, either negatively or positively, on operations thought of as earlier. With even a single negative feedback loop it is possible to generate oscillatory behavior. It has long been known (Goodwin, 1965) that if the operations are appropriately non-linear *and* the system is open to sources of energy, these oscillations may be self-sustained and not dampen to a steady state over time. The same is true of mechanisms employing positive feedback or cyclic organization (see Bechtel & Abrahamsen, 2010).

Accommodating these organizational principles requires dropping the sequential characterization of a mechanism and instead coordinating accounts of parts and operations with accounts of their dynamics. The conception of mechanism hence becomes more dynamic: "A mechanism is a structure performing a function in virtue of its component parts, component operations, and their organization. The orchestrated functioning of the mechanism, **manifested in patterns of change over time in properties of its parts and operations,** is responsible for one or more phenomena" (Bechtel & Abrahamsen, in press). Accounts that utilize this conception exemplify what we have recently called *dynamic mechanistic explanation*. Often such accounts incorporate computational modeling of the real-time dynamics produced by feedback loops and other forms of cyclic organization. Moreover, a dynamic conception of mechanism and mechanistic explanation is compatible with the non-sequential organization, non-linear interactions, and openness to energy required for endogenous operation.

A self-sustaining oscillatory mechanism can account for the endogenous activity found in the brain, but now new explanatory tasks arise. First, the phenomenon of interest is typically not generated by a single oscillatory mechanism but by the coordinated behavior of multiple oscillators. Since Huygens we have known that if a signal can be passed between oscillators, they can synchronize their oscillations. However, depending on the particular ways in which oscillators are organized into a system, a population of oscillators can come to exhibit extremely complex behavior. Second, even a single oscillator can be perturbed by external inputs and the resulting change in its functioning can be complex. Complexity is even greater when a population of oscillators already exhibiting complex behavior is perturbed. These are the sorts of challenges faced in understanding how the brain, viewed as an endogenously active system, is presented with stimuli or tasks. Philosophical accounts of explanation must also reflect these challenges confronted in neuroscience and cognitive science.

References

Andreasen, N. C., O'Leary, D. S., Cizadlo, T., Arndt, S., Rezai, K., Watkins, G. L., et al. (1995). Remembering the past: two facets of episodic memory explored with positron emission tomography. *American Journal of Psychiatry, 152* (11), 1576-1585.

Antrobus, J. S., Singer, J. L., Goldstein, S., & Fortgang, M. (1970). Mindwandering and cognitive structure. *Transactions of the New York Academy of Sciences, 32*, 242-252.

Bechtel, W. (2008). *Mental mechanisms*. London: Routledge.

Bechtel, W., & Abrahamsen, A. (2005). Explanation: A mechanist alternative. *Studies in History and Philosophy of Biological and Biomedical Sciences, 36*, 421-441.

Bechtel, W., & Abrahamsen, A. (2010). Complex biological mechanisms: Cyclic, oscillatory, and autonomous. In C. A. Hooker (Ed.), *Philosophy of complex systems. Handbook of the philosophy of science, Volume 10*. New York: Elsevier.

Bechtel, W., & Abrahamsen, A. (in press). Dynamic mechanistic explanation: Computational modeling of circadian rhythms as an exemplar for cognitive science. *Studies in History and Philosophy of Science Part A.*

Bechtel, W., & Richardson, R. C. (1993). *Discovering complexity: Decomposition and localization as strategies in scientific research*. Princeton, NJ: Princeton University Press.

Berger, H. (1929). Über daas Elektroenkephalogramm des Menschen. *Archiv für Psychiatrie und Nervenkrankheiten, 87* (527-570).

Biswal, B., Yetkin, F. Z., Haughton, V. M., & Hyde, J. S. (1995). Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. *Magnetic Resonance in Medicine, 34* (4), 537-541.

Brown, T. G. (1914). On the nature of the fundamental activity of the nervous centres; together with an analysis of the conditioning of rhythmic activity in progression, and a

theory of the evolution of function in the nervous system. *The Journal of Physiology, 48* (1), 18-46.

Buckner, R. L., & Carroll, D. C. (2007). Self-projection and the brain. *Trends in Cognitive Sciences, 11* (2), 49-57.

Buzsáki, G., & Draguhn, A. (2004). Neuronal oscillations in cortical networks. *Science, 304* (5679), 1926-1929.

Cummins, R. (2000). "How does it work?" versus "what are the laws?": Two conceptions of psychological explanation. In F. Keil & R. Wilson (Eds.), *Explanation and cognition* (pp. 117-144). Cambridge, MA: MIT Press.

Destexhe, A., & Sejnowski, T. J. (2003). Interactions between membrane conductances underlying thalamocortical slow-wave oscillations. *Physiological Reviews, 83* (4), 1401-1453.

Fox, M. D., & Raichle, M. E. (2007). Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. *Nature Reviews Neuroscience, 8* (9), 700-711.

Fox, M. D., Snyder, A. Z., Vincent, J. L., Corbetta, M., Van Essen, D. C., & Raichle, M. E. (2005). The human brain is intrinsically organized into dynamic, anticorrelated functional networks. *Proceedings of the National Academy of Sciences of the United States of America, 102* (27), 9673-9678.

Fox, M. D., Snyder, A. Z., Vincent, J. L., & Raichle, M. E. (2007). Intrinsic fluctuations within cortical systems account for intertrial variability in human behavior. *Neuron, 56* (1), 171-184.

Fox, M. D., Snyder, A. Z., Zacks, J. M., & Raichle, M. E. (2005). Coherent spontaneous activity accounts for trial-to-trial variability in human evoked brain responses. *Nature Neuroscience, 9*, 23-25.

Fukunaga, M., Horovitz, S. G., van Gelderen, P., de Zwart, J. A., Jansma, J. M., Ikonomidou, V. N., et al. (2006). Large-amplitude, spatially correlated fluctuations in BOLD fMRI signals during extended rest and early sleep stages. *Magnetic Resonance Imaging, 24* (8), 979-992.

Gong, P., & van Leeuwen, C. (2004). Evolution to a small-world network with chaotic units. *Europhysics Letters, 67*, 328-333.

Goodwin, B. C. (1965). Oscillatory behavior in enzymatic control processes. *Advances in Enzyme Regulation, 3*, 425-428.

Greicius, M. D., Krasnow, B., Reiss, A. L., & Menon, V. (2003). Functional connectivity in the resting brain: A network analysis of the default mode hypothesis. *Proceedings of the National Academy of Sciences of the United States of America, 100* (1), 253-258.

Grillner, S. (2003). The motor infrastructure: from ion channels to neuronal networks. *Nature Reviews Neuroscience, 4* (7), 573-586.

Lakatos, P., Shah, A. S., Knuth, K. H., Ulbert, I., Karmos, G., & Schroeder, C. E. (2005). An oscillatory hierarchy controlling neuronal excitability and stimulus processing in the auditory cortex. *Journal of Neurophysiology, 94* (3), 1904-1911.

Li, C.-S. R., Yan, P., Bergquist, K. L., & Sinha, R. (2007). Greater activation of the "default" brain regions predicts stop signal errors. *Neuroimage, 38* (3), 640-648.

Llinás, R. R. (1988). The intrinsic electrophysiological properties of mammalian neurons: Insights into central nervous system function. *Science, 242* (4886), 1654-1664.

Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science, 67*, 1-25.

Mantini, D., Perrucci, M. G., Del Gratta, C., Romani, G. L., & Corbetta, M. (2007). Electrophysiological signatures of resting state networks in the human brain. *Proceedings of the National Academy of Sciences, 104* (32), 13170-13175.

Morcom, A. M., & Fletcher, P. C. (2007). Does the brain have a baseline? Why we should be resisting a rest. *Neuroimage, 37* (4), 1073-1082.

Raichle, M. E., MacLeod, A. M., Snyder, A. Z., Powers, W. J., Gusnard, D. A., & Shulman, G. L. (2001). A default mode of brain function. *Proceedings of the National Academy of Sciences of the United States of America, 98* (2), 676-682.

Rubinov, M., Sporns, O., van Leeuwen, C., & Breakspear, M. (2009). Symbiotic relationship between brain structure and dynamics. *BMC Neuroscience, 10* (1), 55.

Sherrington, C. S. (1923). *The integrative action of the nervous system*. New Haven: Yale University Press.

Shulman, G. L., Fiez, J. A., Corbetta, M., Buckner, R. L., Miezin, F. M., Raichle, M. E., et al. (1997). Common blood flow changes across visual tasks. II. Decreases in cerebral cortex. *Journal of Cognitive Neuroscience, 9* (5), 648-663.

Sporns, O., & Zwi, J. D. (2004). The small world of the cerebral cortex. *Neuroinformatics, 2* (2), 145-162.

Stuart, D. G., & Hultborn, H. (2008). Thomas Graham Brown (1882-1965), Anders Lundberg (1920-), and the neural control of stepping. *Brain Research Reviews, 59* (1), 74-95.

Thagard, P. (2006). *Hot thought: Mechanisms and applications of emotional cognition*. Cambridge, MA: MIT Press.

Vincent, J. L., Patel, G. H., Fox, M. D., Snyder, A. Z., Baker, J. T., Van Essen, D. C., et al. (2007). Intrinsic functional architecture in the anaesthetized monkey brain. *Nature, 447* (7140), 83-86.

Watts, D., & Strogratz, S. (1998). Collective dynamics of small worlds. *Nature, 393* (440-442).

Wilson, D. M., & Wyman, R. J. (1965). Motor output patterns during random and rhythmic stimulation of locust thoracic ganglia. *Biophysical Journal, 5* (2), 121-143.

Likability-Based Genres: Analysis and Evaluation of the Netflix Dataset

Andrew M. Olney (aolney@memphis.edu)
Institute for Intelligent Systems, 365 Innovation Drive
Memphis, TN 38152 USA

Abstract

This paper describes a new approach to defining genre. A model is presented that defines genre based on likability ratings rather than features of the content itself. By collecting hundreds of thousands of likability ratings, and incorporating these into a topic model, one can create genre categories that are interesting and intuitively plausible. Moreover, we give evidence that likability-based features can be used to predict human annotated genre labels more successfully than content-based features for the same data. Implications for outstanding questions in genre theory are discussed.

Keywords: Genre; topic model; Netflix; likability;

Introduction

Many web sites, e.g. Amazon, allow users to rate items along several dimensions, the most common being likability or overall satisfaction. These ratings allow other users to roughly estimate their own probable satisfaction with the item, leading to better item selection and better satisfaction with the web site itself. Moreover, the same rating information can be exploited by a website to make personalized recommendations for the user producing the ratings. In theory, highly accurate recommendations might influence the user to purchase additional products, again leading to greater profitability for the web site in question.

This process of tracking ratings and using ratings to make personal recommendations often falls under the classification of "recommender system" or "collaborative filtering," and is a widely studied problem in the data mining/machine learning field (Resnick & Varian, 1997). To assist the development of new and better algorithms, some companies like Netflix have released large datasets containing hundreds of thousands of ratings by hundreds of thousands of users (*The Netflix Prize Rules*, 2010). These datasets can be analyzed in multiple ways, and an interesting perspective is to view them as a kind of graph or social network. By viewing users as nodes and items as edges, we can study how users are related to each other through item connectivity. Conversely, we can study how items are related to each other through users who have rated them. Another way of looking at this second scenario is as "mass criticism" wherein each user is afforded the same status as a critic, and the mass action of all critics determines not only the overall value of the item (through ratings) but also the association of an item with other items (through connectivity).

In film theory, criticism and genre theory are likewise intertwined (Stam, 2000), creating relationships between the value of film and its taxonomic place. Intuitively, a film might be called, "a good comedy" or "a poor horror," in the sense that the genre defines a kind of rubric or context by which the film is evaluated. Genre theorists often attempt to go beyond such normative characterizations to consider genre in terms of sociocultural effects between film, audience, and author. However, even in a more elaborated perspective, there are a number of outstanding issues in genre theory, which can loosely be divided into problems of definition and problems of analysis.

Problems of definition in genre theory include circularity and the monolithic assumption (Stam, 2000). The problem of circularity arises when one tries to define a genre in terms of features like those given in Table 1.

Table 1: Genre Features (Adapted from Chandler (1997)

Feature	Example
Time	Films of the 1930s
Author	Stephan King
Age of audience	Kid movie
Technology	Animated
Star	Sylvester Stallone
Director	Quentin Tarantino
Structure	Narrative
Ideology	Christian
Culture of origin	Bollywood
Subject matter	Disaster movie
Location	Western

A feature based analysis requires first assembling all the films representative of that genre and then analyzing their features. However, gathering the films requires knowing their genre in the first place, otherwise how would one know which films to assemble? A second problem of definition is the monolithic assumption, in which a film is assumed to belong to one and only one genre. While the monolithic assumption in some ways makes the task of genre definition simpler, it nevertheless ignores genres that are part of our public discourse, e.g. "romantic comedy."

Genre theory is also plagued by problems of analysis. Some questions with regard to genre analysis of film are as follows (Stam, 2000). First, are genres real or imagined? In other words, are they merely analytic constructs, or do they have some status in the world. Second, are the number of genre categories finite or infinite? Third, are genres timeless or are they culture-driven and therefore trendy? Finally, are genres universal, or are they culturebound? As questions about genre, these four questions are inherently tied back to the definition of what genre is. Therefore to answer them, we must first define genre.

In this paper, we analyze the information implicit in user

ratings to build a model of genre. Our study focuses on the ratings from the Netflix dataset, which we incorporate into a probabilistic topic model (Griffiths, Steyvers, & Tenenbaum, 2007). Moreover, we show how the extracted genres can be used to predict human annotated genres with better performance than typical features used by genre critics. That a content-free analysis, based purely on likability ratings, can predict genres is surprising and provocative. We argue that the ability of a likability-based analysis to predict genre with more success than a traditional feature-based approach suggests that likability ratings not only represent a new way of considering genre, but they also represent a significant force in shaping genre categories, a force that is possibly more significant than the content itself.

Study 1: Modeling

Method

The data used in this study consisted of the Netflix dataset, which is freely available online (*The Netflix Prize Rules*, 2010). The dataset has a collection of information applicable to both training a model as well as evaluating the model using the Netflix API (*The Netflix Prize Rules*, 2010). In this study and succeeding studies, only the training data was used. The training data consists of two logical components. The first is a master file which lists for each movie a unique id, along with the title and release year for the movie. The second component is a folder which contains, for each movie id, the set of ratings given to that id by various users. Each rating is a triple consisting of user id, rating, and date of rating. Each rating is an integral number from 1 to 5. There are 17,770 movies in the dataset, 480,189 users, and 100,480,507 ratings. The dataset is sparse, meaning that not every user has rated every movie.

Topic models (Griffiths & Steyvers, 2002; Griffiths et al., 2007), also known in other communities as Latent Dirichlet Allocation (Blei, Ng, & Jordan, 2003), are a class of generative statistical models typically applied to text. Topic models use "bag of words" assumption, making them somewhat similar to methods such as latent semantic analysis (Landauer, Foltz, & Laham, 1998; Landauer, McNamara, Dennis, & Kintsch, 2007), however there are significant differences. Rather than reduce the dimensionality of the data according an optimal least-squares approximation, topic models use a probabilistic model that assumes the data was generated by an underlying process involving hidden variables. Thus while LSA expresses the data along latent dimensions, i.e. singular vectors, which have no clear semantic interpretation, topic models express the data according to the topics that generated the data, and these topics are expressed as a collection of semantically related words, i.e. the words that are most probable given a topic.

More specifically, the standard topic model makes the following assumptions. For each document, there is an associated distribution of topics. Each of these topics has an associated distribution of words. Thus to generate a document, one first probabilistically samples a from the distribution of topics, yielding a particular topic. One then probabilistically samples from the distribution of words associated with that particular topic, yielding a word. This process can be repeated to generate more words and more documents. Thus a topic model specifies how to generate the observed data; however a model may be fitted to existing data using probabilistic inference. Briefly, this is accomplished by randomly initializing the model and then using Gibbs sampling to reestimate the model's parameters, iteratively, until the model converges. For more details see Griffiths, Kemp, and Tenenbaum (2008).

Though topic models have primarily been applied to text in the cognitive science community, the model itself is agnostic to the underlying data it represents, so long as that data has a form consistent with the assumptions of the model. One generalization of these assumptions would be as follows: data consists of a set of samples, each sample has a distribution of topics, and each item in the sample is generated from one of these topics. It doesn't matter whether the samples are documents or whether the items are words. Using this intuition, it is fairly straightforward to map the Netflix dataset into a form consistent with the topic model. Indeed there are alternate mappings (Rubin & Steyvers, 2009), but in what follows we will only consider one.

Our mapping is as follows. Each customer is a mixture of genres, and each genre is a distribution over movies. To transform the existing Netflix dataset using this mapping, we collect all of the movies seen by a customer. The number of stars given that movie is represented by the number of times that movies label appears. For example, if a customer had only rated the movie "Whale Rider" and gave it three stars, then the customer would be represented as (Whale Rider, Whale Rider, Whale Rider), analogous to a document containing the same word three times. Under the assumptions of this mapping and the underlying topic model, each star in a customer's rating can generated by a different genre. For example two stars of "Whale Rider" might be generated by the drama genre, and one star might be generated by the foreign film genre.

The inference algorithm to fit our model to the Netflix data is identical to that used in typical topic models. However, given the large size of the dataset and the widespread availability of multi-core processors, we have created and make publicly available our code for fast parallel topic models in the C# language [1]. Inference parameters were as follows. The number of topics was 50, the prior for topics appearing in a document (α) was 1, and the prior for words appearing in a topic (β) was 0.01. The α and β smoothing parameters are typical (Steyvers & Griffiths, 2007). The model was run for 200 iterations.

Results

An initial inspection of the genres found by the model reveals intuitive categories, as displayed in Table 2. The intuitive

[1] http://andrewmolney.name

Table 2: Selected Genres.

Genre 1	Genre 2	Genre 3	Genre 4
Bowling for Columbine	The Mummy Returns	Spirit: Stallion of the Cimarron	My Big Fat Greek Wedding
Fahrenheit 9/11	Bad Boys II	Brother Bear	Sweet Home Alabama
Whale Rider	Face/Off	Treasure Planet	How to Lose a Guy in 10 Days
Super Size Me	Behind Enemy Lines	The Lion King 1 1/2	Pretty Woman
Hotel Rwanda	Tomb Raider	Stuart Little 2	Legally Blonde
Maria Full of Grace	The Fast and the Furious	Garfield: The Movie	Two Weeks Notice
City of God	Rush Hour 2	Spy Kids 2	When Harry Met Sally
The Motorcycle Diaries	Gone in 60 Seconds	Home on the Range	Bridget Jones's Diary
Spellbound	XXX: Special Edition	Scooby-Doo 2	13 Going on 30
Rabbit-Proof Fence	The Mummy	SpongeBob SquarePants	The Wedding Planner

appeal of these genres is consistent with word-based topics presented in the topic model literature (Steyvers & Griffiths, 2007). Each genre list is rank ordered by probabilistic membership. Therefore the first ranked film in each genre is the most probable film given that genre, and so on. This ranking is derived from the ϕ matrix of the topic model (Steyvers & Griffiths, 2007).

Consistencies in Table 2 are evident. For example, Genre 1 could be considered documentaries or biographically inspired independent films. Genre 2 consists of action films that veer towards the fantastic. Genre 3 is made up of animated films directed at children. And Genre 4 lists romantic comedies. However, inconsistencies are also apparent. For example is "Bad Boys II" really as fantastic as a film about mummies? Or are Michael Moore films really that much like "Whale Rider"? Under this critical view, what can be gleaned from Table 2 is somewhat mixed. On the one hand, it is clear that some sense of genre can be driven by likability ratings alone. On the other, it is unclear to what extent these ratings-driven genres correspond to typical film genres. Without a correspondence-based evaluation, it is unclear whether the genres in Table 2 represent strong coherent categories or an observer bias towards any category that might make them coherent.

Study 2: Correspondence-based Evaluation

Method

To carry out a correspondence-based evaluation of our model, it is necessary to find a large existing dataset with human annotated genres for each movie. Fortunately such a dataset exists and is freely available: the Internet Movie Database (IMDB). IMDB contains an enormous amount of information for a given film, ranging from the director and year of release to less commonly known information such as the art department. Including amongst the hundreds of pieces of information associated with each movie is a set of 28 genres, listed in Table 3.

Each film in IMDB is associated with one or more of the genres in Table 3. For example, the biopic, "Ray," based on the story of Ray Charles, is labeled with Biography, Drama, and Music. How these genre labels were generated for IMDB is not clear, and interrater reliability for these genres is not available. The task of correspondence is then to match up every film in the Netflix dataset (which contains all the likability ratings) with the genres in the IMDB dataset. Unfortunately, this is less straightforward than it might first appear. The Netflix dataset is intentionally sparse, including only title, year, and ratings for each film.

Table 3: IMDB Genres.

Documentary	Animation	Family	Sport
Crime	Drama	Mystery	Action
Sci-Fi	Comedy	Short	Game-Show
Romance	Fantasy	Adventure	Music
Thriller	Biography	History	Musical
Horror	Adult	War	Film-Noir
Reality-TV	Western	Talk-Show	News

IMDbPy is the Python-based software library used for manipulating the IMDB data (*IMDbPy*, 2010). IMDbPy provides a search capability for querying a particular title. This search capability purposely returns more than single title in order to accommodate alternate title forms. Using IMDbPy, a correspondence requiring an exact match of both year and title yields only 8,283 exact matches out of a possible 17,770. Relaxing the exact match requirement so that years match and titles match up to the colon yields an additional 1,082 matches.

Inspection of the data reveals that failures to match have a variety of reasons. First, typographic conventions differ between datasets, such that a foreign film may have its original title spelling in one dataset and an Anglicized title in another, e.g. "Character" and "Charackter." In addition, year information may be off by one between the two databases. Sequels and series are a particular problem, such that one database may precede the name of an episode with the name of the series, whereas the other does not. Some errors also exist in the matched films. It is possible, though rare, for two films to be released in the same year with the same name. For

example, "Ray," the biopic of Ray Charles, appeared in the same year as a genre short of the same name. Finally, because to the inconsistencies with series naming conventions and the partial match strategy described above, some within-genre mismatches can occur, e.g. "Star Trek: Insurrection" and "Star Trek: First Contact." However, the distribution of genres is very similar in both the matched set and the original set, as shown in Table 4. Additionally, the correlation between the proportional distributions for original and matched sets is .978.

Table 4: Proportion of Genres.

Genre	Matched	Original
Action	0.14	0.12
Adult	0	0.02
Adventure	0.04	0.04
Animation	0.04	0.05
Biography	0.03	0.02
Comedy	0.24	0.2
Crime	0.06	0.05
Documentary	0.08	0.1
Drama	0.21	0.19
Family	0.02	0.02
Fantasy	0.01	0.01
Film-Noir	0	0
Game-Show	0	0
History	0	0
Horror	0.05	0.04
Music	0.02	0.02
Musical	0.01	0.01
Mystery	0.01	0.01
News	0	0
None (missing)	0	0.05
Reality-TV	0	0
Romance	0.01	0.01
Sci-Fi	0.01	0.01
Short	0.01	0.03
Sport	0	0
Talk-Show	0	0
Thriller	0.02	0.01
War	0	0
Western	0.01	0.01

Once the 9,249 films were paired, the WEKA toolkit (Hall et al., 2009) was used to build two sets of predictive models. The first set uses as features only the distribution of topics associated with each movie, a row vector. For example, position 1 would be the probability that a movie belongs in genre 1, position 2 to probability a movie belongs in genre 2, and so on for all 50 genres. The second set of models uses as features a collection of information from IMDB, chosen to best match the features sometimes used by film critics to determine the genre of a film, as described in Table 1. These features are listed in Table 5.

Table 5: IMDB Features.

Feature	Type
Plot	NUMERIC
Title	NUMERIC
Actor1	NOMINAL
Actor2	NOMINAL
Director	NOMINAL
Year	NUMERIC
MPAA	NOMINAL
Genre	NOMINAL

A few features of Table 5 warrant brief remarks. Plot is a plot synopsis of the film. The two actor features are the first and second named actors on the billing, i.e. the stars of the film. MPAA is the rating of the film, e.g. PG-13. The other features are self-explanatory.

Some of these features are nominal, such as actor and director names, meaning that they are associated with a fixed set of labels as is genre in Table 3. However, the IMDB plot synopsis is an arbitrary string of considerable length, e.g. 500 words, and the title is a shorter but equally arbitrary string. In order to be usable features that two films could have in common, both plot and title were transformed using term frequency/inverse document frequency such that each word in the string became its own feature. This large set of features was considerably pruned using stop words and stemming, so that only 1,420 features remained. The WEKA command line used to convert plot and title to these numeric features was "StringToWordVector -R1,2 -W100 -prune-rate-1.0 -C -T -I -N0 -L -S -SnowballStemmer -M1 -WordTokenizer".

In both the first and second sets, the genre class to be predicted is the first genre listed by IMDB. This restriction is due to WEKA's inability to perform multi-class classifications, and implies that overall performance of the models is significantly lower than would be the case if any genre label associated with a movie was permitted as a correct answer.

The two differing data formats is what separates the first and second sets of models. Within each set, the same machine learning algorithms were used to predict genre. These include the following five models. First, ZeroR, which predicts the most prevalent class, e.g. Comedy. Secondly, NaiveBayes, which assumes features are independent and uses Bayes Rule to construct a classifier. Thirdly, AdaBoostM1 uses an ensemble of weak learners, in this case a decision stump, using the boosting approach (Schapire, 2003). Fourthly, J48 is a decision tree whose internal branching on attribute values is constructed to maximally discriminate amongst the training data. And finally, Ibk is an instance/prototype based classifier, i.e. k nearest neighbors where k has been set to 10 neighbors. These five algorithms were selected because they represent a cross section of the most widespread and effective machine learning techniques (Wu et al., 2007).

Each model was trained using 10 fold cross validation in

which the dataset is divided into ten bins, and the model trained 10 times, using a different bin as test data each time. Significant differences were measured using a paired samples t-test, p = .05, corrected for the variability introduced by cross validation (Nadeau & Bengio, 2003).

Results

The results of the predictive models are displayed in Table 6. Numbers shown indicate percent correct, aggregated across all genre categories. All significant differences are relative to the ZeroR model for each set.

Table 6: Results in Percent Correct.

Model	Likability Based	Content Based
rules.ZeroR	23.51	23.51
bayes.NaiveBayes	9.94	27.12
meta.AdaBoostM1	23.96	23.51
trees.J48	37.30	29.21
lazy.IBk	41.22	27.50

Interestingly there is a fair distribution of performance across all models for the first set (likability-based genres). The worst performer is NaiveBayes, worse than the ZeroR model, while the best performer is IBk-10, at 41%. All differences in this first set are significant.

Performance on the second set of models is worse than the performance on the first set. There is very little deviation away from ZeroR. All differences are significant, except AdaBoostM1, which is not significantly different from ZeroR. The best model of the second set, J48, has only 29% accuracy compared to 41% for IBk in the first set. This performance is particularly poor considering the base rate (ZeroR) is 23%.

Two important points are clear from this data. The first is that the likability-based genres are indeed strong and coherent, predicting the correct human annotated label in 41% of cases. The second is that the likability-based features are more successful at predicting the human annotated label than are the content-based features.

Discussion

Perhaps the most significant finding of both studies is that genres can be extracted from just ratings. Although the percent accuracy using just ratings is 41%, that is still a large figure given two observations. The first is that the 41% performance is based on a single genre classification, when IMDB allows multiple classifications. So 41% performance represents the lowest, most conservative figure. The second observation is that the likability-based performance is considerably higher than the content-based performance at 29%. This difference suggests that likability-based genre classification is a more accurate model of how humans classify film genres than is content-based classification.

The topic model we use makes very few assumptions, and yet the assumptions it does make are quite strong. The basic premise of the model is that people are a mixture of genres. These genres, in turn, generate the ratings observed. To claim that people are a mixture of genres, when genres are typically considered to be a property of artifacts, is a strong and radical claim. The results of the two studies presented above not only support this claim but also suggest that it should be taken seriously as a new approach to genre.

Suppose that likability-based genres are taken seriously. Are they useful, particularly in regard to existing genre studies? The current focus on film suggests that they are. Recall the complementary problems of genre definition and analysis discussed in the introduction. Using likability-based genres as a framework, these can be addressed straightforwardly.

As before, the problems of definition include circularity and the monolithic assumption (Stam, 2000). The basic problem of circularity lies in a supervised approach in which a critic tries to align film features with a given genre category. A likability-based model, as an unsupervised model, avoids this problem entirely because their is no initial assumption of genre used to define the features of genre. Instead, genre emerges from genre-agnostic likability ratings. The second problem of definition, the monolithic assumption, is addressed by the structure of the topic model. Under this model, every movie has some probability of membership in every genre. Study 2 above illustrates that it is not necessary to pigeonhole a movie into a genre in order to create meaningful genres: even using a probabilistic definition of genre, one can still approximate the monolithic assumption to 41% accuracy. Pluralistic genres, like "romantic comedy," are not a special case but are represented in the same way as any other genre.

Using the likability-based definition of genre, we can also clarify problems of analysis that have been raised (Stam, 2000). First, are genres real or imagined? According to our approach, genres are only manifested through people's preferences. Therefore they do not have any status in the world except as a consensus of preferences across large groups of people. On whether the number of genre categories finite or infinite, the structure of the topic model suggests that the number of genres is completely arbitrary, and is controllable using the parameter T, the number of topics. This suggests that likability-based genres are potentially infinite. Third, on whether genres are timeless or are trendy, the likability-based model suggests that they are trendy. Any new ratings that are assimilated into the model can change the resulting genres. As long as the people making the new ratings represent a new mixture of genres, the genres will shift towards the trendy. Finally, as to whether the genres are universal or culturebound, one can speculate that they are culturebound to the extent that one culture may rate movies consistently differently from another culture. This is intuitively plausible, e.g. Bollywood movies rated in India vs. the United States, and may be accounted for in the same way as the timeless or trendy prob-

lem.

Likability-based genres also extend beyond the traditional conceptualization of genre and correspond to the notion of intertextuality. In film, intertextuality has be described as having several properties (Stam, 2000). The first overarching property is that every film is necessarily related to every other film. Second, intertextuality is an active process, so rather than "belonging" to a genre, a film dynamically relates to other films. Finally, intertextuality involves not only all other films, but potentially other arts and media. Clearly the likability-based model corresponds to each of these three properties, by being based on the connectivity amongst all movies via ratings, using an active data-driven model, and using the abstract notion of rating, which can be applied to heterogeneous items like film, music, and books simultaneously. Thus the likability-based model can apply to modern intertextual theories of media in addition to traditional notions of genre.

In summary, likability-based genres offer a novel and useful way of considering genre: people are a mixture of genres. Likability-based genres can predict a significant percentage of genres in the Netflix dataset. Moreover, likability-based genres can also be used to address fundamental problems of definition and analysis in film theory. Likability-based genres can also be extended to broader frameworks than genre, such as intertextuality. However, likability-based genres as described in this paper do not represent a complete theory. In order to understand this phenomenon fully, it is necessary to understand how the ratings themselves are generated as well as how likability-based genres manifest in other contexts.

Acknowledgments

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education,through Grant R305A080594 and by the National Science Foundation, through Grant BCS-0826825, to the University of Memphis. The opinions expressed are those of the authors and do not represent views of the Institute, the U.S. Department of Education, or the National Science Foundation.

References

Blei, D., Ng, A., & Jordan, M. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, *3*, 993–1022.

Chandler, D. (1997). *An introduction to genre theory.* Available from http://www.aber.ac.uk/media/Documents/intgenre /chandler_genre_theory.pdf

Griffiths, T. L., Kemp, C., & Tenenbaum, J. B. (2008). Bayesian models of cognition. In R. Sun (Ed.), *The Cambridge handbook of computational psychology* (pp. 59–100). New York: Cambridge University Press.

Griffiths, T. L., & Steyvers, M. (2002). A probabilistic approach to semantic representation. In W. D. Gray & C. D. Schunn (Eds.), *Proceedings of the 24th annual conference of the cognitive science society* (pp. 381–386). Lawrence Erlbaum Associates.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological review*, *114*(2), 211–244.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The weka data mining software: an update. *SIGKDD Explor. Newsl.*, *11*(1), 10–18.

Imdbpy. (2010, February). Available from http://imdbpy.sourceforge.net/index.php?page=main

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes*, *25*(2&3), 259-284.

Landauer, T. K., McNamara, D. S., Dennis, S., & Kintsch, W. (2007). *Handbook of latent semantic analysis*. Lawrence Erlbaum.

Nadeau, C., & Bengio, Y. (2003). Inference for the generalization error. *Mach. Learn.*, *52*(3), 239–281.

The netflix prize rules. (2010, February). Available from http://www.netflixprize.com/rules

Resnick, P., & Varian, H. R. (1997). Recommender systems. *Commun. ACM*, *40*(3), 56–58.

Rubin, T., & Steyvers, M. (2009). A topic model for movie choices and ratings. In *Proceedings of the ninth international conference on cognitive modeling.* Manchester, UK.

Schapire, R. E. (2003). The boosting approach to machine learning: An overview. In D. D. Denison, M. H. Hansen, C. C. Holmes, B. Mallick, & B. Yu (Eds.), *Nonlinear estimation and classification* (Vol. 171, pp. 149–172). New York: Springer Verlag.

Stam, R. (2000). *Film theory: an introduction*. Malden, Mass.: Wiley-Blackwell.

Steyvers, M., & Griffiths, T. L. (2007). Probabilistic topic models. In T. Landauer, D. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis* (pp. 424–440). Lawrence Erlbaum.

Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., et al. (2007). Top 10 algorithms in data mining. *Knowl. Inf. Syst.*, *14*(1), 1–37.

Constructing Typing-Time Corpora: A New Way to Answer Old Questions

Uriel Cohen Priva (urielc@stanford.edu)
Department of Linguistics, 450 Serra Mall
Stanford, CA 94305 USA

Abstract

Many current studies in linguistics and psycholinguistics require the use of phonetically labeled speech data. Collecting and annotating such data is expensive and slow. An alternative approach makes use of pre-labeled speech corpora, but these are available for very few languages, might not contain the desired linguistic environment, and the construction of new ones is still expensive and time-consuming. We present a fast and cost-efficient method for constructing a new type of corpus which retains many of the advantages of phonetically labeled speech, *typing-time corpora*. In this paper we show that an English typing-time corpus collected over the web is sufficient to replicate word frequency and neighborhood density effects. We then demonstrate the transferability of this method to less studied languages and to different orthographies. We show that a smaller Hebrew typing corpus collected over the web can be used to find lengthening effects in infrequent Hebrew words.

Keywords: Typing-time; Corpora; Frequency; Neighborhood density; Amazon Mechanical Turk

Introduction

Many studies in linguistics and psycholinguistic require either the precise annotation of durations and latencies for speech data gathered in carefully controlled experiments, or the availability of phonetically labeled corpora. For example, evidence for *neighborhood density* in production depends on measuring speech production latencies (Vitevitch, 2002). The study of *production difficulties* relies on measuring the lengthening of words in a difficult context (Fox Tree & Clark, 1997). Studies of *frequency* and predictability-dependent phonetic reduction (Van Son & Van Santen, 2005; Pluymaekers, Ernestus, & Baayen, 2005; Aylett & Turk, 2006; Bell, Brenier, Gregory, Girand, & Jurafsky, 2009) make use of corpora containing exact word and phone durations, such as the Switchboard Corpus (Godfrey & Holliman, 1997) and Buckeye Corpus of Conversational Speech (Pitt et al., 2007) for English, the Spoken Dutch Corpus (Oostdijk, 2000) and the Kiel Corpus of Spontaneous Speech (Kohler, Pätzold, & Simpson, 1995) for German.

However, neither the experimental approach nor the corpus-based one may be a feasible option when trying to address the problem of data availability in less studied languages. Subjects may not always be available on the one hand, and Switchboard-like corpora do not exist for most languages on the other hand. In addition, even when a corpus is accessible, it might not contain the relevant linguistic environments for addressing the questions at hand. The creation of even a small-scale corpus is an expensive and time-consuming project, and therefore, there is much gain in finding a simpler alternative. In this paper, we propose a solution to this problem. We show that by wedding two methodological advancements, tracking typing speed and collection of data over the web, we can create an alternative both to experiments which require phonetic labeling and to phonetically-labeled corpora, *typing-time corpora* — corpora of typed data in which each letter and word is annotated with the time it took to type.

A number of studies (Weingarten, Nottbusch, & Will, 2004; Zesiger, Orliaguet, Boë, & Mounoud, 1994 among others) demonstrate that typing is sensitive to language-based effects. Weingarten et al. (2004) show that typing is sensitive to phonological and morphological properties of the words being typed. Zesiger et al. (1994) show that actual words are typed faster than pseudo-words and that frequent words are typed faster than infrequent words. These effects demonstrate that even though a typing task is different from spoken speech production, it does exhibit linguistic effects that are normally associated with speech.

Not only does typing time provides a window to linguistic performance, but it also holds a big advantage, as it allows the automatic gathering of large amounts of data through the web. A simple way to utilize this is by using Amazon Mechanical Turk (AMT), a virtual work marketplace created by Amazon.com. On AMT, requesters can upload work requests in the form of HTML pages, which workers can access online. Several researchers in the natural language processing community (Callison-Burch, 2009; Colowick & Pool, 2007; Snow, O'Connor, Jurafsky, & Ng, 2008 among others) make use of AMT to construct corpora for which human-labeled data is not available, or to annotate new data sets. In this paper, we demonstrate that extending the use of AMT to the construction of typing-time corpora provides an easy and cost-efficient alternative to laboratory experiments and extant corpora. We show evidence that supports the applicability of this methodology by constructing a typing time corpus for English, and using it to replicate two well known effects on language production: word frequency (Bell et al., 2009) and neighborhood density (Coltheart, Davelaar, Jonasson, & Besner, 1977; Vitevitch, 2002; Adelman & Brown, 2007). We then extend these results to a less studied language. We

show the effect of word frequency on typing time in a smaller Hebrew corpus, exemplifying that the paradigm holds even for relatively small typing-time corpora, and for different languages with varying orthographies.[1]

Previous production studies

Frequency effects

Much current work in linguistics stresses the importance of word-frequency in the minute modulations in the duration of words, morphemes, syllables and phones in various contexts. These durations are taken from corpora of spontaneous or read speech in which phone durations were hand-labeled by linguists. Bell et al. (2009) show that frequent English words tend to reduce more than infrequent words. Pluymaekers et al. (2005) show the reduction of Dutch morphemes in predictable contexts. Aylett and Turk (2006) show reduction in predictable English syllables. Van Son and Van Santen (2005) show that some contextually predictable consonants are more likely to reduce.

Neighborhood effects

A wide range of studies has shown language users to be sensitive to the effects of neighborhood-density (Coltheart et al., 1977; Vitevitch, 2002; Adelman & Brown, 2007; Peereman & Content, 1997). Coltheart et al. (1977) defines the neighborhood density of a given word as "the number of words that can be produced by changing just one of the letters in the string to another letter, preserving letter positions." Two different definitions of neighborhood density follow naturally from this one: Coltheart's original spelling-based definition, in which the substitutions are of single orthographic characters, and a phonological definition, in which the substitution is based on phonemes. Peereman and Content (1997) argue that the best approximation of neighborhood density is phonographic, that is, the cases in which the spelling neighbor is also the phonological neighbor. Furthermore, neighborhood density has been shown to have different consequences in production and comprehension. For English, Vitevitch (2002) shows that a dense neighborhood facilitates spoken word production, whereas Vitevitch and Luce (1998) show that a dense neighborhood inhibits word comprehension.

Motivating typing-time corpora

The studies cited above demonstrate the benefit of investigating slight modulations of durations and latencies in spoken language production. However, many of them presuppose rather ideal experimental settings: a laboratory with accurate recording equipments, access to relevant human subjects in the proximity of that laboratory, sufficient time to label large amounts of data, and ample funds. The often easier alternative of using a pre-labeled speech corpus is not available when the linguistic environment being studied is not present in the corpus, or when no such corpus is present, as is in fact the case with most languages. Therefore, there would be much to benefit from a new methodology for investigating language production.

The following two sections describe the components of the solution we propose for this problem: an experimental approach to the collection of typing-time data, and the collection of large amounts of data over the web. By combining these methodologies we can create *typing-time corpora*, which provide an answer to the problem we presented above; they do not require any special equipment, subjects from remote locations can provide experimental data over the web, and no further labeling is required.

Online data collection

Even basic web technology allows the collection of data through the web. Every search request on the web involves sending data to some webserver, which can collect the data it receives. However, utilizing web technology for data collection requires finding enough workers to perform the specific task. Amazon Mechanical Turk (AMT) provides a simple platform to do so. AMT is a virtual marketplace in which requests and workers can interact. The requester uploads tasks in the form of HTML pages to the website and proposes to pay a given price for the completion of each task. Workers can choose among available tasks, perform them, and submit the results through AMT. The requester can then review and approve the results, which leads to the transfer of the proposed sum of money from his account to the workers' accounts. AMT handles the overhead involving all other aspects of the interaction: the exchange of money and the collection of the results.

Several recent studies have already made use of AMT (Callison-Burch, 2009; Colowick & Pool, 2007). Colowick and Pool (2007) use AMT to find preferences for semantic scope ambiguity, and Callison-Burch (2009) uses AMT to evaluate the quality of automatic translations.

One possible concern with data collected this way is whether it can be as accurate as data collected under controlled conditions. However, Snow et al. (2008) compare the performance of AMT annotators with that of professional annotators, and they find that by increasing the number of annotators, untrained annotators over the web can match the performance of expert annotators. Increasing the number of data points per observation type is a key concept in handling noisy data collected

[1] The typing time approach is, of course, limited to languages that have a letter-based written standard (unlike, e.g., Chinese). While not all languages have such a written form, or any kind of written form at all, the proposed methodology would still allow access to a large number of currently less studied languages.

over the web. Since the gathering of data over the web is fast and inexpensive, enough data points can be collected to ensure that noisy data would be as sensitive as a smaller amount of data collected under ideal conditions.

Typing time experiments

Several studies have demonstrated that typing speed is affected by linguistic factors. Gentner (1982) shows that a sequence of keystrokes is more predictive of the time it would take to strike one key if the sequence does not span word boundaries. Gentner, Larochelle, and Grudin (1988) show that the same four-key sequence is typed faster in frequent words than in infrequent words of comparable length. Weingarten et al. (2004) show that typing is sensitive to morphological-syllabic boundaries, by comparing the lag between typing two specific keys, held constant across conditions, and varying between syllable and morpheme boundaries.

Since typing requires moving the hands and fingers to different locations on the keyboard, the baseline lag between the typing of a given key and the preceding key varies dramatically based on the preceding and possibly the following keystrokes. Gentner (1982) shows that almost 50% of the variability is controlled for if we control for the immediately preceding keystroke. He also shows that adding up to one more key to the preceding context of the target key, and up to one following key, can account for most of the location-based variability.

While typing time studies clearly show the potential of using typing time as a segue to assessing linguistic performance, the factorial methods used in Gentner et al. (1988) and Weingarten et al. (2004) are not always replicable in further languages. Weingarten et al. (2004), who investigate lexical access effects in German, keep the same two-key sequences while varying the morphological and syllabic environment. However, many languages would not necessarily allow the same two-key sequence to appear in every condition, making a factorial design impossible. It would be beneficial to see such effects even if only some of the conditions exist for each two-key environment. Gentner et al. (1988) uses identical four-key sequences embedded in words of varying frequencies, but in orthographic systems in which vowels are not assigned a separate letter (e.g. Arabic or Hebrew), words that contain identical four-key sequences would usually belong to the same stem or the same neighborhood. These issues can be remedied by the proposed methodology of constructing a typing-time corpus.

Building typing-time corpora

We construct the typing-time corpus in the following manner. AMT workers (or other web users) are presented with an HTML form in which they first fill in some basic details. We request our subjects to say whether they are left- or right-handed, and whether they look at the keyboard while typing. They are also requested to type in the keyboard keys below the digits 1–6 in order to identify the keyboard layout, and to fill in the first two languages they speak, following an example in which the first language is not the language we want to investigate. In order to reduce the variance, submissions from anyone who is left-handed, looks at the keyboard, is not using the most common keyboard layout (QWERTY in the case of English) or did not fill in the language we want to investigate were not included in the analysis (but were still accepted and paid).

After the basic details are collected, the subjects move to ten open text fields. After they choose the field, text appears to the right (and in right-to-left languages, to the left) of the open text field, and the subjects are instructed to copy it. Once they move to the next field, the field they leave is locked, and they are no longer able to change it. While they type, a javascript program running in the web page collects the exact time of each key press.

The output of the collected data is then parsed and assigned additional attributes. Each keystroke is associated with the word it belongs to and the key typed in that word. Corrected text is recorded as *corrected*, and words that contain it are marked as *corrected*. Words that do not match the target text are recorded as *wrong*. Keystrokes that took more than 500ms to type are considered a *break*, and words that contain breaks past the first characters are considered *interrupted*. When a word is marked as *interrupted*, *corrected* or *wrong*, all the keystrokes that comprise it are marked as having a *interrupted*, *corrected* or *wrong* attribute, respectively.

Several tests can be performed on the collected corpus. It is possible to check in which contexts we find typing errors, which environment cause significant lags in typing time, etc. This paper concentrates on the modulation of inter-key duration, which we will call *lags*. The distribution of lags is not normal, leading us to use percentiles and medians rather than means and standard deviations. We first exclude all data from AMT workers that submitted more than five tasks, all keys that originate in *interrupted*, *corrected* or *wrong* words, all word-initial keys, and the top and bottom five percentiles of remaining lags. Following Gentner (1982) we build on the fact that the variance of keystrokes is reduced when preceding and following keys are taken into account as context. Like Weingarten et al. (2004) we use the preceding key for small corpora, but we also include the following key if at least ninety percent of every three-key sequences appear in the corpus at least five times. The median of each set of keys sequences is used as the *expected* lag of the the target key in that context. Our predicted value is the ratio between the actual lag and the expected lag, rather than the actual size of each lag. In this way, we can compare lag modulation across different words, and not limit ourselves to a specific key sequences. For ex-

ample, if the lag for the letter 'e' in the context 'r*e*a' has an expected baseline of 220ms (based on all occurrences of the 'rea' in the corpus), but in a specific instance of the word 'great' we measure it to be 140ms long, we would like to explain why that particular 'e' is shorter, the value to predict being 140:220 (figure 1). Since the predicted value is the ratio, we can compare the ratios of different keystrokes, in different contexts.

key	g	⟷	r	⟷	e	⟷	a	⟷	t
actual lag		100		140		30		90	
expected lag		210		220		100		150	
ratio		0.48		0.64		0.30		0.61	

Figure 1: sample actual:expected ratios

Study 1: Lexical frequency and neighborhood density in English

In the first study, we construct a typing time corpus for English, and use it to investigate the effects of neighborhood density and frequency on the typing-time. We predict a facilitatory effect of word frequency on its typing time. Additionally, we expect to find an effect of neighborhood density.

Constructing an English typing-time corpus

The English typing-time corpus was built using AMT, using the procedure described above. Each AMT task was unique, but workers could participate in the study up to five times.

In order to choose the stimuli words to be typed, each word in the CMU Pronunciation Dictionary (Weide, 1998) was matched with its frequency and its most common letter case in the New York Times section of English Gigaword Third Edition (Graff, Kong, Chen, & Maeda, 2007): Gigaword-NYT. The corpus has two sections, which correspond to data collected using two different kinds of stimuli. Both tasks were used in order to calculate the expected lag of each keystroke in the context of one preceding and one following keystrokes.

In the first data collection task, AMT workers were requested to type in four randomly chosen words in each item. The words were independent from one another. Each word was in one of the top ten thousand lowercase words in Gigaword-NYT. A total of 475 AMT tasks were collected, and each took about two minutes to perform. No worker had to type the same word twice within the same hour.

The second data collection task required AMT workers to type in five words that form a coherent sentence, which was sampled from Gigaword-NYT. All sentences were exclusively in lowercase in the original corpus except for the first character, which was also changed into lowercase for the construction of the stimuli. The sentences were comprised only of words that are in the top five thousand most frequent words in Gigaword-NYT. No sentence had conjunctions or WH-words. Pronouns, if they appeared at all, occurred only before the verb. Each sentence had a verb and a noun following it. A total of 190 AMT tasks were collected, and each took about two minutes to perform. No worker had to type the same sentence twice within the same hour.

Methods and materials

We investigate the effects of neighborhood density and frequency on the modulation of inter-key typing lag. A linear regression was used to estimate the predicted value, which was defined as the log ratio between a lag and its expected value. Only lags from the first section of the English typing-time corpus (words in isolation) were estimated. The key's position in the word, the predicted lag, AMT workers' typing rate across all items, their typing rate in the corresponding item and the logged predicted lag time were used as controls. The word frequencies used were the corresponding word counts in the NYT section of English Gigaword Third Edition (Graff et al., 2007): Gigaword-NYT. Two frequency measurement were tested. The first was the negative log unigram probability of that word: $-\log \Pr(word)$. The second word frequency measurement was based on the word lemmas: $-\log \Pr(lemma)$, calculated using WordNet (Miller, 1995).[2]

Neighborhood density was calculated using the CMU dictionary. We tested three variants of neighborhood density: the number of spelling neighbors (substitution of one letter), the number of phonological neighbors (substitution of one segment) and the number phonographic neighbors (substitution of one letter and one segment).

The linear regression model was selected using R's (R Development Core Team, 2010) `step()` function which uses AIC (Akaike, 1974) for model selection. The model was also re-evaluated using a mixed-effect model with worker and word as random effects. No significant changes to the significance and direction of the reported coefficients were found.

Results and discussion

Both word and lemma frequency alone have a significant facilitatory effect on typing speed (words which are frequent or whose lemma is frequent are typed faster). However, in the final model only the frequency of lemma remained significant, as it masks the effect of the frequency of the word. The lemma unigram frequency has a significant ($p < 10^{-7}$) facilitatory effect and is significantly superior to word probability ($p < 0.02$).

All three neighborhood density measurements have a significant facilitatory effect on typing speed (words with a dense neighborhood are typed faster). However, in the

[2]If a word was ambiguous between two parts of speech, the shorter lemma was associated with the word.

final model only phonological density remains significant ($p < 0.001$). Phonological neighborhood density is not significantly better than spelling neighborhood density ($p = 0.097$) or phonographic neighborhood density ($p = 0.13$). The adjusted R^2 is 0.237

These results show that typing-time corpora are indeed sensitive to the well known effects of word frequency and neighborhood density. The fact that it is the frequency of lemmas rather than words suggests that lexicon access is active during typing, as shown in Weingarten et al. (2004). The fact that neighborhood density has a *facilitatory* effect is of particular importance, since it has been shown that in English a dense neighborhood facilitates productions whereas it inhibits comprehension (Vitevitch, 2002; Vitevitch & Luce, 1998). Therefore, although the typing task arguably involves both production and comprehension, the results suggest that this method is indeed tapping into the effects of production.

Study 2: Lexical frequency in Hebrew

In the second study, we construct a typing time corpus for Hebrew. We use it to demonstrate that this paradigm is extensible to other languages, and can be collected outside AMT. We show that Hebrew demonstrates word frequency effects on typing-time.

Constructing a Hebrew typing-time corpus

Hebrew orthography is different from that of English in several crucial aspects. It is written from right to left, it has no uppercase-lowercase distinction, and most importantly it does not incorporate most vowels. Furthermore, norms regarding the use of space are different — several very frequently occurring clitics (such as *ve* 'and') are glommed to the following word.

The Hebrew typing-time corpus was built using an online form, the results of which were collected by a web server.[3] One hundred unique tasks were generated, and each task was performed no more than three times, by different subjects. Subjects could participate in the study up to five times.

In order to choose the stimuli words to be typed, we collected 1300 articles from the *Haaretz*, a Hebrew news website. We calculated the frequency of each word and used Hspell (Har'El & Kenigsberg, 2006) to stem it from possible adjoining clitics. Word-frequencies were estimated using the same data from *Haaretz*. We calculated the expected lag of each keystroke in the context of one preceding keystroke.[4]

The data collection task was similar to the isolated word section of the English corpus described in study 1. Subjects were asked to type in five randomly chosen words in each item. The words were independent from one another. Each word was in one of the top five thousand in *Haaretz*. A total of 72 web tasks were collected, and each took about two minutes to perform. No worker had to type the same word twice.

Methods and materials

We investigate the effects of frequency on the modulation of inter-key typing lag. As in study 1, a linear regression was used to estimate the predicted value, which was defined as the log ratio between a lag and its expected value. Once again, the key's position in the word, the predicted lag, the subjects' typing rate across all items, their typing rate in the corresponding item and the logged predicted lag time were used as controls. We limited ourselves to words that had no clitics. The word frequencies used were the corresponding word counts *Haaretz*. Two frequency measurement were tested. The first was the negative log unigram probability of that word including its clitics when they occur, $-\log \Pr(clitics + word)$. The second word frequency measurement was based on the stemmed words (which still include morphological inflections, but not adjoining clitics) $-\log \Pr(word)$.

The linear regression model was evaluated as in Study 1. The linear regression model was selected using R's (R Development Core Team, 2010) `step()` function (Hastie & Pregibon, 1992) which uses AIC (Akaike, 1974) for model selection. The model was also re-evaluated using a mixed-effect model with worker and word as random effects. No significant changes to the significance and direction of the reported coefficients were found.

Results and Discussion

Of the two word-frequency measurements, only the frequency of the word form that included its clitics came up significant $p < 0.05$. The frequency of the bare form did not come up significant even when we excluded the frequency of the cliticized form. The adjusted R^2 is 0.1462.

These results show that even with a much smaller typing-time corpus, frequency effects can be seen.

General Discussion

The experimental results shown in both studies provide strong support of our proposal that typing time corpora can provide a simple method to investigate linguistic performance. Further investigation is required to assess the many different ways in which production is similar or different across the typed and spoken modalities.

Study 1 shows that the reduction of frequent words, an effect shown by both laboratory experiments and phonetically tagged corpora, has a corollary in typing time which can be replicated using our corpus. It also shows that a facilitatory effect of neighborhood density can be observed using our corpus, which shows that it patterns with production rather than comprehension.

[3] We did not use AMT because there are currently not enough native speakers of Hebrew in AMT

[4] There was not enough data to use the following key as well.

Study 2 demonstrates that this methodology can be easily extended to other, less studied languages. The results show a shortening of typing lag in more frequent words in Hebrew, as was shown for English in Study 1. This demonstrates that the method is applicable to new languages, even those with non-Roman orthographies.

Acknowledgments

This research was partially supported by the NSF via award IIS-0624345. Special thanks to Dan Jurafsky, Roey Gafter, Chigusa Kurumada, Victor Kuperman, Matthew Adams and Meghan Sumner.

References

Adelman, J. S., & Brown, G. D. A. (2007). Phonographic neighbors, not orthographic neighbors, determine word naming latencies. *Psychonomic Bulletin & Review*, *14*(3), 455–459.

Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, *19*(6), 716–723.

Aylett, M., & Turk, A. (2006). Language redundancy predicts syllabic duration and the spectral characteristics of vocalic syllable nuclei. *The Journal of the Acoustical Society of America*, *119*, 3048–3058.

Bell, A., Brenier, J. M., Gregory, M., Girand, C., & Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational english. *Journal of Memory and Language*, *60*(1), 92–111.

Callison-Burch, C. (2009). Fast, cheap, and creative: Evaluating translation quality using Amazon's Mechanical Turk. In *Proceedings of EMNLP 2009.*

Colowick, S. M., & Pool, J. (2007). Disambiguating for the web: a test of two methods. In *Proceedings of K-CAP '07* (pp. 173–174). ACM.

Coltheart, M., Davelaar, E., Jonasson, J. T., & Besner, D. (1977). Access to the internal lexicon. In S. Dornic (Ed.), *Attention and performance VI* (pp. 535–555). Hillsdale, NJ: Erlbaum.

Fox Tree, J. E., & Clark, H. H. (1997). Pronouncing *the* as *thee* to signal problems in speaking. *Cognition*, *62*, 151–167.

Gentner, D. R. (1982). Evidence against a central control model of timing in typing. *Journal of Experimental Psychology: Human Perception and Performance*, *8*(6), 793–810.

Gentner, D. R., Larochelle, S., & Grudin, J. (1988). Lexical, sublexical, and peripheral effects in skilled typewriting. *Cognitive Psychology*, *20*(4), 524–548.

Godfrey, J. J., & Holliman, E. (1997). *Switchboard-1, Release 2.* Linguistic Data Consortium, Philadelphia.

Graff, D., Kong, J., Chen, K., & Maeda, K. (2007). *English gigaword third edition.* Linguistic Data Consortium, Philadelphia.

Har'El, N., & Kenigsberg, D. (2006). *Hspell.* `http://hspell.ivrix.org.il/`.

Hastie, T. J., & Pregibon, D. (1992). Generalized linear models. In J. M. Chambers & T. J. Hastie (Eds.), *Statistical models in S* (chap. 6). Pacific Grove, CA: Wadsworth and Brooks / Cole.

Kohler, K., Pätzold, M., & Simpson, A. (1995). *From scenario to segment: the controlled elicitation, transcription, segmentation and labelling of spontaneous speech. AIPUK 29.* Kiel: IPDS.

Miller, G. A. (1995). Wordnet: a lexical database for english. *Commun. ACM*, *38*(11), 39–41.

Oostdijk, N. (2000). The spoken dutch corpus project. *ELRA newsletter*(5), 4–8.

Peereman, R., & Content, A. (1997). Orthographic and phonological neighborhoods in naming: Not all neighbors are equally influential in orthographic space. *Journal of Memory and Language*, *37*(3), 382–410.

Pitt, M. A., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E., et al. (2007). *Buckeye corpus of conversational speech, 2nd release.* Department of Psychology, Ohio State University.

Pluymaekers, M., Ernestus, M., & Baayen, R. H. (2005). Articulatory planning is continuous and sensitive to informational redundancy. *Phonetica*, *62*, 146–159.

R Development Core Team. (2010). R: A language and environment for statistical computing [Computer software manual]. Available from `http://www.R-project.org`

Snow, R., O'Connor, B., Jurafsky, D., & Ng, A. Y. (2008). Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of EMNLP 2008* (pp. 254–263).

Van Son, R. J. J. H., & Van Santen, J. P. (2005). Duration and spectral balance of intervocalic consonants: A case for efficient communication. *Speech Communication*, *47*, 100–123.

Vitevitch, M. S. (2002). The influence of phonological similarity neighborhoods on speech production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(4), 735–747.

Vitevitch, M. S., & Luce, P. A. (1998). When words compete: Levels of processing in perception of spoken words. *Psychological Science*, *9*(4), 325.

Weide, R. (1998). *The CMU pronunciation dictionary, release 0.6.* (Carnegie Mellon University)

Weingarten, R., Nottbusch, G., & Will, U. (2004). Morphemes, syllables and graphemes in written word production. In T. Pechmann & C. Habel (Eds.), *Multidisciplinary approaches to language production* (Vol. 157, pp. 529–572). Berlin: Mouton de Gruyter.

Zesiger, P., Orliaguet, J., Boë, L., & Mounoud, P. (1994). The influence of syllabic structure in handwriting and typing production. In C. Faure, P. Keuss, G. Lorette, & A. Vinter (Eds.), *Advances in handwriting and drawing: a multidisplinary approach* (pp. 389–401). Paris: Europia.

Large-Scale Acquisition of Feature-Based Conceptual Representations from Textual Corpora

Barry Devereux (barry@csl.psychol.cam.ac.uk)[1], Nicholas Pilkington (ncvp2@cam.ac.uk)[2], Thierry Poibeau (thierry.poibeau@ens.fr)[3], Anna Korhonen (alk23@cam.ac.uk)[2]

[1] Centre for Speech, Language and the Brain, Department of Experimental Psychology, University of Cambridge

[2] Computer Laboratory, University of Cambridge

[3] Laboratoire LaTTiCe-CNRS, Paris

Abstract

Methods for estimating people's conceptual knowledge have the potential to be very useful to theoretical research on conceptual semantics. Traditionally, feature-based conceptual representations have been estimated using property norm data; however, computational techniques have the potential to build such representations automatically. The automatic acquisition of feature-based conceptual representations from corpora is a challenging task, given the unconstrained nature of what can constitute a semantic feature. Existing computational methods typically do not target the full range of concept-relation-feature triples occurring in human generated norms (e.g. *tiger have stripes*) but rather focus on concept-feature tuples (e.g. *tiger – stripes*) or triples involving specific relations only. We investigate the large-scale extraction of concept-relation-feature triples and the usefulness of encyclopedic, syntactic and semantic information in guiding the extraction process. Our method extracts candidate triples (e.g. *tiger have stripes*, *flute produce sound*) from parsed corpus data and ranks them on the basis of semantic information. Our investigation shows the usefulness of external knowledge in guiding feature extraction and highlights issues of methodology and evaluation which need to be addressed in developing models for this task.

Keywords: distributed conceptual representations; semantic features; corpus-based acquisition

Introduction

Concrete concepts like TIGER, APPLE and CHISEL constitute a fundamental part of people's coherent mental representations of the world around them. A key question in cognitive science is how these semantic representations are organised and accessed. Most theories of conceptual representation assume a distributed, feature-based model of conceptual knowledge (e.g. Cree, McNorgan, & McRae, 2006; Randall, Moss, Rodd, Greer, & Tyler, 2004; Tyler, Moss, Durrant-Peatfield, & Levy, 2000). According to such theories, conceptual knowledge is distributed across a network of interconnected feature units (such as *has_eyes*, *has_ears*, *has_stripes*) with concepts' meanings being represented as patterns of activation across these units. The relative prominence of this distributed, feature-based account of conceptual representation in the literature reflects the many perceived strengths of such a framework.

A key issue for all studies which aim to test distributed theories of concepts is the accurate estimation of the knowledge that people are likely to represent in such a system. Recent connectionist, behavioural and neuropsychological studies (e.g. Cree et al., 2006; Grondin, Lupker, & McRae, 2009; Randall et al., 2004; Tyler et al., 2000; Taylor, Salamoura, Randall, Moss, & Tyler, 2008) have relied on data derived from property norming studies. Currently, the largest set of norms available is that collected by Ken McRae and colleagues which contains features for 541 concrete concepts (McRae, Cree, Seidenberg, & McNorgan, 2005). Participants listed features for each concept word and McRae et al. normalised them by mapping different feature descriptions with the same meaning to the same feature label.

Feature-based representations of concepts based on property-norming studies have played an important role in testing theories of conceptual knowledge. However, property norms come with several important caveats (see e.g. Murphy, 2002, for a discussion). One issue is that participants tend to under-report features which are present in many of the concepts in a category (McRae et al., 2005; Murphy, 2002, p. 32); for TIGER for example, participants list salient features like *has_teeth* but not less salient features like *has_eyes*. Thus *has_eyes* is not listed for TIGER although presumably all McRae et al.'s participants knew that tigers have eyes. Another concern is the size of the currently available property norms. Although the largest collection of norms lists features for over 500 concepts, larger sets of norms would be useful given the number of confounding variables (word length, familiarity, etc) that need to be controlled for in studies of concepts and word meaning. Unfortunately, large scale property norming studies are costly and time consuming.

In recent years, researchers have begun to develop methods which can automatically extract feature norm-like representations using corpus-based computational techniques (e.g. Almuhareb & Poesio, 2005; Barbu, 2008; Baroni, Murphy, Barbu, & Poesio, 2009). These approaches – and the approach we present in this paper – have their antecedents in early methods for extracting and organizing the semantic feature information implicit in dictionary definitions (e.g. Chodorow, Byrd, & Heidorn, 1985). The automatic approach is cost-effective and can gather large-scale frequency data from text corpora. As corpora contain words denoting concepts and their features in natural language, they provide ideal material for feature generation. However, current methods target concept-feature tuples only or are restricted to specific relations between concepts and their features. For example, Almuhareb and Poesio (2005) targeted *is-a* and *part-of* relations, whilst Barbu (2008) combined linguistic patterns with a co-occurrence based method to extract six types of features: *superordinate*, *part*, *stuff*, *location*, *quality* and *action*.

The Strudel model (Baroni et al., 2009) also uses linguis-

tic patterns, but more generally. Strudel uses "connector patterns" consisting of sequences of part-of-speech tags to look for candidate feature terms near a target concept. Properties are scored based on the number of distinct patterns connecting them to a concept, rather than on the overall number of corpus co-occurrences. When evaluated against the ESSLLI dataset that includes 44 concepts from the McRae norms (Baroni, Evert, & Lenci, 2008), Strudel yields the precision of 23.9% – which is the best state of the art result for unconstrained acquisition of concept feature tuples.

Due to the difficulty of the task, we believe that additional linguistic and world knowledge will be required to extract more accurate representations. Moreover, Strudel has the limitation that it produces concept-feature tuples – not concept-relation-feature triples similar to those in human generated norms (although the distribution of the connector patterns for a tuple does cue information about the broad class of semantic relation that holds between concept and feature).

In this paper, we investigate the challenges that need to be met in both methodology and evaluation when aiming to move towards unconstrained, large-scale extraction of concept-relation-feature triples in corpus data. The extraction of such realistic, human-like feature norms is extremely challenging and we do not predict a high level of accuracy in these first experiments. We investigate the usefulness of three types of external knowledge in guiding feature extraction: encyclopedic, syntactic and semantic knowledge. We first compile large automatically parsed corpora from Wikipedia which contains encyclopedic information. We then introduce a novel method which extracts concept-relation-feature triples from grammatical dependences produced by a parser. We use probabilistic information about semantic classes of features and concepts to guide the acquisition process. Our investigation shows that external knowledge can be useful in guiding the extraction of human-like norms.

Extraction Method

Corpora

We chose Wikipedia as our corpus as it is a freely available and comprehensive encyclopedia that includes basic information on many everyday topics. Almost all concepts in the norms have their own Wikipedia articles, and the articles often include facts similar to those elicited in norming studies (e.g. the article *Elephant* describes how elephants are large, are mammals, and live in Africa). By using Wikipedia, we investigate the usefulness of a smaller amount of more focused (encyclopedic) corpus data for the task.

The XML dump of Wikipedia was filtered to remove non-encyclopedic articles (e.g. talk pages), article sections that are unlikely to contain parsable text (e.g. bibliography sections), and inline references (e.g. book citations). The remaining content was preprocessed with Wikiprep (Gabrilovich & Markovitch, 2007), removing tables, unparsable elements (e.g. Wikipedia infoboxes) and the WikiMedia mark-up, yielding a plaintext version of each article. Two subcorpora were created from the resultant set of 1.84 million articles. The first of these (Wiki500) includes the Wikipedia articles that correspond to each of the McRae concepts. It contains *c*. 500 articles (1.1 million words). The second subcorpus consists of those articles which contain one of the McRae concept words in the title and the title is less than five words long.[1] This Wiki110K corpus includes 109,648 plaintext articles (36.5 million words).

Recoding the McRae features

We recoded a British English version of the McRae norms to a uniform representation that is more appropriate for our computational work. Each concept-feature pair in the norms (e.g. TIGER *has_stripes*) was automatically recoded to a triple of the form *concept relation feature-head* where *concept* was the singular of the concept noun (e.g. 'tiger'), *relation* was the root form of a verb (e.g. 'have') and *feature-head* was always a singular noun or an adjective (e.g. 'stripe'). Feature-heads containing more complex information than could be captured with a single noun or adjective were split into two or more triples (for example, the norm feature *is a_musical instrument* for ACCORDION was recoded to the two triples *accordion be instrument* and *accordion be musical*). Where "beh" and "inbeh" appeared in features in the norms (indicating behaviour features of animate and inanimate concepts; e.g. DOG *beh_bark*) this was replaced with the verb "do". Prepositions and determiners were also removed when constructing the triples. Although this recoding involves a loss of information to some extent, it also enables us to clearly distinguish between the relation and feature-head parts in each feature norm. It is triples of this form that we aim to extract with our computational method.

Candidate feature extraction

Our method for extracting concept-relation-feature triples consists of two stages: we first extract large sets of candidate feature triples for each target concept from the corpus, and then re-rank and filter the triples with the aim of retaining only those triples which are most likely to be true semantic features.

For the first stage, the corpora are parsed using the Robust Accurate Statistical Parsing (RASP) system (Briscoe, Carroll, & Watson, 2006). For each sentence in the corpora, this yields the set of grammatical relations (GRs) for the most probable analysis returned by the parser. The GR sets for each sentence containing the target concept noun are then retrieved from the corpus. We construct an undirected acyclyic graph of the GRs that spans the sentence and which has the target concept word as its root node. The nodes are labelled by the words occurring in the sentence and an edge is present when a GR links those two words in the sentence. Edges can thus be labelled by the GR types. For example, the graph

[1] The subset was limited to articles with titles less than five words long in order to avoid articles on very specific topics which are unlikely to contain basic information about the target concept (e.g. *Coptic Orthodox Church of Alexandria* for CHURCH.)

constructed for the sentence *Tabby tigers can often have pale stripes* contains a path connecting *tiger*, *have* and *stripe*.

Our method considers the set of paths through the tree between the target concept root node and the other nodes which are either an adjective or a noun; these adjectives and nouns are the potential feature heads in the concept-relation-feature triples. If there is a verb in the path between the target concept and the feature head, we extract the candidate triple *concept verb feature-head*. The first stage of our method extracts all possible candidate triples from the set of paths. As this method is maximally greedy, the second stage evaluates the quality of these extracted candidates using semantic information, with the aim of filtering out the poor quality features.

Re-ranking based on semantic information

The more often a triple is extracted for a concept, the more likely it is that the triple corresponds to a feature related to the concept. However, production frequency alone is an inadequate measure of the quality of the feature term because concept terms and candidate feature terms can co-occur for all sorts of reasons. For example, one of the extracted triples for TIGER is *tiger have squadron* (because of the RAF squadron called the Tigers).

The probability of a feature being part of a concept's representation is dependent on the semantic category that the concept belongs to (*used for_cutting* should have low probability for animals, for example). We conducted an analysis of the norms to quantify this type of semantic information. Our aim was to identify higher-order structure in the distribution of semantic classes for features and concepts, with the goal of investigating whether this information is useful in feature extraction. More formally, we assume that there is a 2-dimensional probability distribution over concept and feature classes, $P(C,F)$, where C is a concept class (e.g. *Animal*) and F is a feature class (e.g. *Body-Part*). Knowing this distribution gives a way of evaluating how likely it is that a candidate feature f is true for a concept c, assuming that we know that $c \in C$ and $f \in F$. We can regard the McRae norms as being a sample drawn from this distribution, provided the concept and feature terms appearing in the norms can be assigned to suitable concept and feature classes. Clustering was used to identify such classes.

Clustering Our cluster analysis used Lin's (1998) similarity metric, which uses the WordNet ontology as the basis for calculating similarity. Such a measure is appropriate for our purposes as we are interested in generating suitable superordinate classes for which we can calculate the distributional statistics. The concepts and feature-head terms appearing in the recoded norms were each clustered independently into 50 clusters using hierarchical clustering. Table 1 presents three concept clusters and three feature clusters with five representative members of each cluster (we have given intuitive labels to the clusters for explanatory purposes). In general, semantically similar concepts and features clustered together.

We calculated the conditional probability $P(F|C)$ of a feature cluster given a concept cluster using the data in the McRae norms. Table 2 gives the conditional probability for each of the three feature clusters given each of the three concept clusters that were presented in Table 1. For example, $P(\textit{Body Parts}|\textit{Reptiles})$ is higher than $P(\textit{Body Parts}|\textit{Vehicles})$: given a concept in the *Reptiles* cluster the probability of a *Body_Part* feature is relatively high whereas given a concept in the *Vehicle* cluster the probability of a *Body_Part* feature is low. The cluster analysis therefore supports our hypothesis that the likelihood of a particular feature for a particular concept is not independent of the semantic categories that the concept and feature belong to.

Clusters	Example Members
Concept clusters	
Reptiles	alligator, crocodile, iguana, rattlesnake
Fruit/Veg	cucumber, honeydew, mushroom, plum
Vehicles	ambulance, helicopter, car, rocket, jet
Feature clusters	
Body Parts	ear, foot, fuzz, nose, tongue
Plant Parts	bark, berry, blade, grape, prune
Activities	cluck, drip, emergency, flow, funeral

Table 1: Example members of concept and feature clusters

	Reptiles	Fruit/Veg	Vehicles
Body Parts	0.164	0.031	0.023
Plant Parts	0.009	0.130	0.014
Activities	0.100	0.060	0.140

Table 2: $P(F|C)$ for $C \in \{\text{Reptiles, Fruit/Veg, Vehicles}\}$ and $F \in \{\text{Body Parts, Plant Parts, Activities}\}$

Reranking We used this distributional semantic information to improve the quality of the *concept relation feature* candidate triples, by using the conditional probabilities of the appropriate feature cluster given the concept cluster as a weighting factor. To get the probabilities for a triple, we first find the clusters that the concept and the feature-head words belong to. When the feature-head word of the extracted triple appears in the norms, its cluster membership is looked up directly; when it is not in the norms we assign the feature-head to the feature cluster with which it has the highest average similarity. Given the concept and feature clusters determined for the concept and feature in the triple, we reweight the triple's frequency by multiplying it by the conditional probability. This helps downgrade incorrect triples that occur frequently in the data and boost the evidence for correct triples.

Baseline model For the purposes of evaluation, we also implemented a co-occurrence-based model based on the "SVD" (Singular Value Decomposition) model described by Baroni et al. (2009). A word-by-word co-occurrence matrix was constructed for both our corpora, storing how often each target word co-occurred in the same sentence as each context word. Context words were defined to be the 5,000 most frequent content words in the corpora. Target words were the concept names in the recoded norms, supplemented with the 10,000

most frequent content words in the corpora (with the exception of the 10 most frequent words). The dimensionality of the co-occurrence matrix was reduced to 150 columns by singular value decomposition. Cosine similarity between pairs of target words was calculated and, for each concept word, we chose the 200 most similar target words to be the feature-head terms extracted by the model.

Experimental Evaluation

Methods of Evaluation

We considered several methods for evaluating the quality of the extracted feature triples. One method is to calculate precision and recall for the extracted triples with respect to the McRae norms "gold standard". However, direct comparison with the recoded norms is problematic since an extracted feature which is semantically equivalent to a triple in the norms may have a different lexical form. For example, *avocado have stone* appears in the recoded norms whilst *avocado contain pit* is extracted by our method; direct comparison of these two triples results in *avocado contain pit* being incorrectly counted as an error. To deal with the fact that semantically identical features can be lexically different, we followed the approach taken in the ESSLLI 2008 Workshop on semantic models (Baroni et al., 2008). The gold standard for the ESSLLI task was the top 10 features for 44 of the McRae concepts: for each feature an expansion set was given, listing words that were synonyms of the feature term that appeared in the norms. For example, the feature *lives on water* was expanded to the set {*aquatic*, *lake*, *ocean*, *river*, *sea*, *water*}.

We expect to find correct features in corpus data which are not in the "gold standard" (e.g. *breathes_air* is listed for WHALE but for no other animal). We therefore aim for high recall in the evaluation against the ESSLLI set (since all features in the norms should ideally be extracted) but not necessarily high precision (since extracted features that are not in the norms may still be correct; e.g. *breathes_air* for TIGER). To evaluate the ability of our model to generate such novel features, we also conducted a manual evaluation of the highest ranked extracted features which did not appear in the norms. Finally, we introduce a novel evaluation method which makes no direct use of McRae norms. This is based on analysis of the extracted feature-based semantic reprentations in terms of conceptual structure properties. Conceptual structure statistics such as feature distinctiveness, sharedness and correlation strength have an important role to play in testing distributed theories of conceptual knowledge (e.g. see Randall et al., 2004; Taylor et al., 2008). Therefore, we were interested in the accuracy of the conceptual structure statistics that can be calculated from the extracted features. If the conceptual structure statistics calculated for the extracted features resemble those obtained from human-generated norms, it provides evidence that the extracted features capture important aspects of the semantics of concrete concepts.

Extraction set	Corpus	Prec.	Recall
SVD Baseline	Wiki500	0.0235	0.4712
	Wiki110K	0.0140	0.2798
Method - unfiltered	Wiki500	0.0239	0.5081
	Wiki110K	0.0068	0.8083
Method - top 25% unweighted	Wiki500	0.0470	0.2735
	Wiki110K	0.0179	0.6260
Method - top 25% weighted	Wiki500	0.0814	0.4167
	Wiki110K	0.0230	0.6851

Table 3: Results for the baseline model and the extraction method, when matching on features but not relations.

Precision and Recall

The *recall score* for a concept is defined as the number of extracted features for the concept that appear in the recoded norms divided by the total number of features for that concept in the norms. High recall indicates that a high proportion of the McRae features are being extracted. The *precision score* for a concept is defined as the number of extracted features for that concept that appear in the norms divided by the total number of features extracted for the concept.[2] As discussed above, we aim to maximize recall.

Table 3 presents the results when we evaluate using the feature-head term alone (i.e. in calculating precision and recall we disregard the relation verb and require only a match between the feature-head terms in the extracted triples and the recoded norms). Evaluating tuples (rather than triples) is how large-scale models of feature extraction have typically been evaluated in the past (e.g. Baroni et al., 2009).

Results for four sets of extractions are presented. The first set is the set of features extracted by the SVD baseline. The second set of extracted triples are the full set of triples extracted by our method, prior to the reweighting stage. "Top 25% unweighted" gives the results when all but the top 25% most frequently extracted triples for each concept are filtered out. Note that the filtering criteria here is raw extraction frequency, without reweighting by conditional probabilities. "Top 25% weighted" are the corresponding results when the features are weighted by the conditional probability factors prior to filtering; that is, using the top 25% reranked features. The effectiveness of using the semantic class-based analysis data in our method can thus be assessed by comparing the filtered results with and without feature weighting.

For the baseline implementation, the results are better using the smaller Wiki500 corpus than the larger Wiki110K corpus. This is not surprising, since the smaller corpus contains only the articles corresponding to the concepts in the norms. This smaller corpus thus minimizes sources of noise such as word polysemy that are more apparent in the larger corpus (e.g. "tiger" almost always refers to the animal in the Wiki500 corpus, but can have other meanings in larger or general cor-

[2]Since we define precision over the whole set of extracted features, our precision score is not comparable to Baroni et al. (2009), where the top 10 extracted features are used.

pora (the RAF squadron called the Tigers, etc)).

The results for the baseline model and the unfiltered experimental method are quite similar for the Wiki500 corpus. As our extraction method is deliberately greedy, extracting many candidate features per sentence, it is not surprising that its performance is comparable to a purely co-occurrence-based method. The innovation of our method is that it uses information about the GR-graph of the sentence to also extract the verb which appears in the path linking the concept and feature terms in the sentence, which is not possible in a purely co-occurrence-based model.

The results for the unfiltered model using the Wiki110K corpus give the maximum recall achieved by our method; 81% of the features are extracted. Precision is low (because of the large number of features being extracted) although, as discussed above, we are less interested in precision, particularly for the unfiltered model. For the results of the filtered feature sets, where all but the top 25% of features were discarded, we see the benefit of reranking, with the reranked frequencies yielding higher precision and recall scores than the method using the unweighted extracted frequencies.

We also evaluated the extracted triples using the full relation + feature-head pair (i.e. both the feature and the relation verb have to be correct). Previous researchers have typically only compared extracted features to the feature-head term; to our knowledge our work is the first to try and compare extracted features to the full relation + feature norm. Unsurprisingly, this reduces recall and precision compared to the case where only the feature-head terms need match. For example, for the Wiki110K corpus recall falls from 69% to 35% for the filtered re-ranked model. However, given that we impose no constraints on what the relation verb can be and that we do not have expanded synonym sets for verbs it is actually impressive that the verb agrees with what is in the recoded norms about 50% of the time.

Manual Evaluation Analysis

Inspection of the extracted triples reveals that some of them are correct although they do not appear in the gold standard norms. One motivation for developing NLP technology for feature extraction is the need to enrich existing models of conceptual representation with novel features. To evaluate the method's ability to learn this type of novel data, 10 concepts were selected at random from among the McRae concepts and the top 20 extracted triples not present in the norms were selected. Two judges evaluated whether these were genuine errors or valid data missing from the norms. The judges rated each "erroneous" triple as correct, plausible, wrong, or wrong but related. The judges worked first independently and then discussed the results to reach consensus. Across the 10 concepts, 23% and 26% of the relation+feature pairs were considered correct and plausible respectively, indicating roughly half of the errors were not true errors but potentially valid triples missing from the norms. This demonstrates the potential of NLP methods in enriching existing models of conceptual representation.

Measure	Correl	p
Number of features	0.203	< 0.001
Number of distinctive features	0.168	< 0.001
Number of shared features	0.113	0.983
Mean distinctiveness	0.167	< 0.001
Proportion of shared features	0.155	< 0.001
Mean correlational strength	-0.118	0.014

Table 4: Evaluation in terms of CSA variables

Evaluation in terms of conceptual structure

Of particular interest to distributed, feature-based theories of conceptual knowledge is how relationships which exist between the features of concepts influence conceptual processing. Statistics capturing such relationships have proven useful in testing theories of distributed semantic representation, including the conceptual structure account (Randall et al., 2004; Tyler et al., 2000). Researchers have calculated several variables from norm data which capture various aspects of the structural organization of the semantic space (e.g. McRae et al., 2005; Randall et al., 2004). Here, we propose a novel method for evaluating feature extraction methods which is based on testing whether conceptual structure statistics calculated from the extracted features exhibit similar qualities to those calculated on the McRae norms.

Various kinds of conceptual structure variables can be calculated. The simplest is the *number of features* in the concept (i.e. the number of features with non-zero production frequency). Features can also be distinguished by whether they are shared or distinctive. Highly shared features occur in many concepts (e.g. *has_legs*); highly distinctive features occur in few concepts (e.g. *has_an udder*). The reciprocal of the number of concepts that a feature occurs in is a measure of the feature's *distinctiveness* (so a feature occurring in two concepts has distinctiveness of 0.5). In particular, a feature is defined to be *distinguishing* if it occurs in one or two concepts and *shared* if it occurs in more than two concepts. For each concept, we can then define the mean distinctiveness of its features, the number of shared and distinguishing features it has, and the proportion of shared features. We can also define a measure of the strength of interconnection between a pair of features. For example, *has_eyes* and *has_ears* co-occur together in concepts more often than do the features *is_gray* and *has_teeth*. The correlation strength for a pair of features is calculated as the Pearson correlation of their production frequencies across concepts. We can then calculate the *mean correlational strength* of a concept's constituent features (using only the shared features; see Cree et al., 2006; Taylor et al., 2008).We therefore define a total of six conceptual structure variables, summarized in Table 4.

The results show a significant correlation between the norms and the extracted triples for five of the six conceptual structure variables. This is important as it indicates that the semantic representations generated from the extracted features are capturing some aspects of the conceptual structure that is present in the norms. However, the correlations are

quite weak, and we do not see expected differences between living and non-living domains that are observed in the McRae norms. What we wish to highlight here is the potential usefulness of conceptual structure statistics as a means for evaluating models: improvements to the extraction method should yield better quality conceptual structure statistics.

Discussion

The feature acquisition method that we have presented above aims to extract semantically unconstrained concept-relation-feature triples from corpus data. High accuracy extraction of such general representations from corpora is unrealistic given the state of the art. The main goal of our experiment was to investigate issues in both methodology and evaluation which need to be addressed when aiming towards higher accuracy feature extraction in the future. In particular, we examined the usefulness of three types of knowledge for guiding feature extraction: encyclopedic, syntactic, and lexical-semantic. We have also compared different approaches to evaluation: direct evaluation against existing norms, qualitative analysis, and evaluation against conceptual structure variables.

Our extraction method performs better than the co-occurrence-based baseline, demonstrating the benefits of using syntactic information for feature extraction. Using GRs also allows us to extract a relation verb for each concept-feature pair, which is not possible using a purely co-occurrence-based approach like the SVD baseline. Performance was improved further by using semantic constraints calculated from the concept and feature clusters: the re-weighting of features based on distributional data increased the rank of higher-quality features.

Our paper highlights the difficulties inherent in evaluating the quality of extracted features. Evaluation that tests against existing property norms is problematic, since participants in property norming studies list features in unsystematic ways. Furthermore, as property norms are created by normalizing participants' responses to a set of feature labels, direct lexical comparison with property norms is not necessarily meaningful. Although the ESSLLI sub-set of the norms which expands the set of features in the norms with their synonyms goes some way towards addressing the latter issue, the former issue remains: norms are not complete in the sense that there are true features which are not included in the norms.

We therefore considered other forms of evaluation. Our qualitative analysis shows that about 50% of the errors against the recoded norms are in fact correct or plausible features. Our novel evaluation in terms of the conceptual structure variables acts as a valuable task-based evaluation that avoids direct comparison with the norms, and instead compares higher-level structural properties of concepts. Future work can aim for larger-scale qualitative evaluation using multiple judges as well as investigate other task-based evaluations.

Acknowledgments

This research was supported by EPSRC grant EP/F030061/1. We thank McRae et al. for making their norms available.

References

Almuhareb, A., & Poesio, M. (2005). Concept learning and categorization from the web. In *Proceedings of the 27th Annual Meeting of the Cognitive Science Society* (pp. 103–108).

Barbu, E. (2008). Combining methods to learn feature-norm-like concept descriptions. In *Proceedings of the ESSLLI Workshop on Distributional Lexical Semantics* (pp. 9–16).

Baroni, M., Evert, S., & Lenci, A. (Eds.). (2008). *ESSLLI 2008 Workshop on Distributional Lexical Semantics.*

Baroni, M., Murphy, B., Barbu, E., & Poesio, M. (2009). Strudel: A corpus-based semantic model based on properties and types. *Cognitive Science*, 1–33.

Briscoe, E., Carroll, J., & Watson, R. (2006). The second release of the RASP system. In *Proceedings of the Interactive Demo Session of COLING/ACL-06* (pp. 77–80).

Chodorow, M. S., Byrd, R. J., & Heidorn, G. E. (1985). Extracting semantic hierarchies from a large on-line dictionary. In *Proceedings of the 23rd Annual Meeting of the Association for Computational Linguistics* (pp. 299–304).

Cree, G. S., McNorgan, C., & McRae, K. (2006). Distinctive features hold a privileged status in the computation of word meaning: Implications for theories of semantic memory. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *32*(4), 643–58.

Gabrilovich, E., & Markovitch, S. (2007). Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *Proceedings of IJCAI'07* (pp. 1606–1611).

Grondin, R., Lupker, S. J., & McRae, K. (2009). Shared features dominate semantic richness effects for concrete concepts. *Journal of Memory and Language*, *60*(1), 1–19.

Lin, D. (1998). An information-theoretic definition of similarity. In *Proceedings of ICML'98* (p. 296-304).

McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, *37*, 547–559.

Murphy, G. (2002). *The big book of concepts*. Cambridge, MA: The MIT Press.

Randall, B., Moss, H. E., Rodd, J. M., Greer, M., & Tyler, L. K. (2004). Distinctiveness and correlation in conceptual structure: Behavioral and computational studies. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *30*(2), 393–406.

Taylor, K. I., Salamoura, A., Randall, B., Moss, H., & Tyler, L. K. (2008). Clarifying the nature of the distinctiveness by domain interaction in conceptual structure: comment on cree, McNorgan, and McRae (2006). *Journal of Experimental Psychology: Learning, Memory & Cognition*, *34*(3), 719–725.

Tyler, L. K., Moss, H. E., Durrant-Peatfield, M. R., & Levy, J. P. (2000). Conceptual structure and the structure of concepts: A distributed account of category-specific deficits. *Brain and Language*, *75*(2), 195–231.

Are Random Representations Accurate Approximations of Lexical Semantics?

Brendan T. Johns (johns4@indiana.edu)
Michael N. Jones (jonesmn@indiana.edu)
Department of Psychological and Brain Sciences, Indiana University
1101 E. Tenth St., Bloomington, In 47405 USA

Abstract

A common assumption made by cognitive models is that lexical semantics can be approximated using randomly generated representations to stand in for word meaning. However, the use of random representations contains the hidden assumption that semantic similarity across randomly selected words is normally distributed. We evaluated this assumption by computing similarity distributions for randomly selected words from a number of well-know semantic measures and comparing them with the distributions from random representations commonly used in memory models.

Keywords: Memory models; semantics; episodic recognition

Introduction

A model of a cognitive phenomenon typically requires an account of both representation and process, and how the two interact (Estes, 1975). These two aspects of a model are interdependent, with the process requiring a representation on which to operate, and the representation requiring a process to simulate behavior. A common practice in cognitive modeling is to use randomly generated representations if the theorist wishes to evaluate a process mechanism, but is unsure of the correct psychological structure or features to use as a representation. This practice makes it unlikely that the representation is biased towards supporting the process model, and the process account can be later refined when further research reveals the correct representation. Over the history of computational modeling, emphasis has been placed on processing over representation.

If insufficient research exists to point towards the correct representation, random representations often provide a useful alternative or simulation of the process would be impossible. An excellent example is Hintzman's (1986) use of random representations to simulate schema abstraction using Posner and Keele's (1968) stimuli. Briefly, stimuli were random dot patterns, and exemplars of the same category were random perturbations of a prototype pattern. Without needing to account for how the human visual system represents dot patterns, Hintzman was able to create equivalent structure in his simulation by generating random prototypes and exemplars.

Random representations have been commonly used in models of episodic memory, for example, recognition, recall, and paired-associate learning. In global matching models of recognition memory (e.g., Hintzman, 1986; Murdock, 1982; Shiffrin & Steyvers, 1997) decisions are made by assessing the similarity of the probe word to the (usually noisy) study items with a particular processing and decision mechanism. The use of random representations in these models produces a hidden assumption that the distribution of similarity across randomly selected words is symmetric and approximately Gaussian.

The distributional assumption comes from the design of a typical memory experiment in which random words are used. In these experiments, random words are selected from a word pool (e.g., Friendly, et al., 1982). Because words are randomly selected, they are assumed to have only random similarity on dimensions extraneous to the experimental manipulation (e.g., orthography, phonology, semantics, etc.); however, this assumption is unlikely to be true. Hence, it is common to explicitly control extraneous factors such as frequency. In this examination, we focus on semantics—a factor often ignored because it is difficult to quantify and control. In assuming that two randomly selected words have only a random expected semantic similarity, random representations seem appropriate.

However, the use of these representations assumes that semantic similarity is randomly distributed across all sampled words. We demonstrate in the following analysis that this is unlikely to be the case with real words, and may produce consequences for conclusions drawn from process models that have used random representations.

Analysis

To evaluate the assumption of random similarity, comparison distributions are needed. Our analysis will utilize three types of semantic similarity measures to create distributions—similarity measures computed from: 1) free association data, 2) a hand-coded lexical ontology (WordNet), and 3) corpus-based co-occurrence models.

Semantic Measures

1. Word Association Space (WAS). Steyvers, Shiffrin, and Nelson (2004) developed a method for inferring semantic representations from free association data. Steyvers et al. represented the free association data for the 5000 cue words from Nelson, McEvoy, and Schreiber's (1999) norms in a word-by-word matrix, where each entry was the probability of a cue word (the row) eliciting the response (the column). This matrix was then reduced in dimensionality using singular value decomposition so that each word was represented by an abstracted 400-dimensional vector. Steyvers et al. demonstrated that the resulting vectors are a good predictor of similarity effects in recognition, recall, and other behaviors.

2. **WordNet Similarity**. WordNet (Miller, 1990) is a hand-coded lexical database encoded as a network in which nodes contain one or more synonymous words. These nodes are then linked together via different types of lexical relationships (e.g. hypernymy and holonymy) and based on these relationships it is possible to build a measure of semantic similarity between two given words using network statistics. A variety of methods that have proposed to do compute similarity, but the measure that seems to best map onto human similarity ratings is the Jiang-Conrath distance measure (JCN; Maki, McKinely, & Thompson, 2004). JCN is a network distance measure that basically counts the number of nodes and edges between two concepts in the database.

3. Latent Semantic Analysis (LSA). This method (and those that follow) differs from the WAS of Steyvers, et al. (2004) in that it does not use human behavioral data to create a semantic representation but, rather, uses statistical regularities computed from a large text corpus. In LSA (Landauer & Dumais, 1997), a word-by-document matrix is created by tabulating the frequency that each word occurs in a given document, inversely weighted by the word's marginal frequency and entropy over documents. The dimensionality of this matrix is then reduced using singular value decomposition so that each word is represented by a vector containing the 300-400 dimensions with the largest eigenvalues. Words that frequently co-occur in similar documents will be represented by similar vectors.

4. BEAGLE. In the BEAGLE model of Jones and Mewhort (2007), a distributed holographic representation of a word is built through experience with a text corpus. Words are initially represented by random Gaussian vectors, and a word's semantic representation is created by summing and convolving (cf. Murdock, 1982) other words that occur in sentences with a target word. The use of convolution allows order information to be included (the sentential position of the word relative to other words), as well as the basic co-occurrence information in LSA. This associative mechanism affords inclusion of rudimentary syntactic knowledge in the vector representation of the word.

5. The COALS model. Unlike the two previous models, COALS (Rohde, Gonnerman, & Plaut, submitted) is not designed to explain human learning, but rather to create a co-occurrence metric that yields the best predictions on a variety of semantic tasks. The model creates a word-by-word matrix, with modifications to how values within the matrix are computed (i.e. correlations are used instead of pure co-occurrence count). This large, sparse matrix is subsequently reduced in dimensionality with SVD in the same way LSA reduces a co-occurrence matrix.

6. Pointwise Mutual Information (PMI). PMI uses a pure co-occurrence count across a large text corpus to create a measure of similarity between two words (e.g., Recchia & Jones, 2009). As with COALS, PMI is not meant to be a model of human learning or representation, but rather a scalar measure of similarity between two words. PMI is essentially computed by taking the probability of observing word x and word y together and dividing by the probability of observing x and y independently. Recchia & Jones computed PMI values over a very large corpus of Wikipedia articles (approximately 400,000 articles), and found that PMI produced a significantly better fit to human rating data than LSA or other semantic similarity metrics.

Random Representations

To compare to the distributions created by the semantic measures, we explored five common types of random vectors that have been used to represent semantics in influential models of memory.

1. Random Gaussian Vectors. A word's representation is created by randomly sampling vector elements from a Gaussian distribution with a certain mean (typically zero) and variance (usually 1/N, where N is vector dimensionality). This type of representation has been used in a variety of models of recognition (e.g. Murdock, 1982), and recall, among others. In the following analysis, vectors were created as in Murdock (1982), with a vector size of 250, a mean of 0 and an SD of ($\sqrt{1/250}$).

2. Gamma Vectors. A word vector is created by sampling integers from a gamma distribution:

$$P[V = j] = (1-g)^{j-1} g, j = 1,...,\infty \quad (1)$$

Where g is a parameter between 0 and 1 that defines the environmental base rates for the different feature values. This type of representation has been used in the highly successful REM model of recognition memory (Shiffrin & Steyvers, 1997), and related models. We constructed these vectors as specified in Shiffrin & Steyvers (1997), with a length of 20, and a g = 0.45 (the parameter used to create high frequency words).

3. MINERVA vectors. In the influential MINERVA 2 model of memory (Hintzman, 1986), vector elements are assumed to be randomly selected from the set of {-1, 0, 1}. A value of 1 is intended to represent a positive link between the word and that feature, a -1 represents an inhibitory link, while a 0 is defined as either irrelevant or unknown for that particular word and feature. Vectors were constructed with a length of 20. Similarity for these vectors was calculated with the following equation:

$$s_i = \sum_{j=1}^{D} \frac{P_i \cdot T_{i,j}}{n} \quad (2)$$

Where D is the size of the vectors, P is the probe word, T is a studied memory trace and n is the number of non-zero items in P. The value is then transformed by cubing it.

4. Sparse Binary Vectors. In this type of distributed representation, the majority of entries are zero, with some entries having the value of 1 at random locations. For instance, in Plaut (1995) items in a word's semantic representation had a 10% probability of being non-zero. Sparse binary vectors have been used to model lexical priming (Plaut) and recognition memory (Dennis & Humphreys, 2001), among other domains. Similar to Plaut's

simulations we generated vectors with a length of 100 and each item having a 10% probability of being non-zero. In addition, binomial distributions (with a sparsity of 50%) will also be tested to examine the effect of sparseness on the similarity distributions.

5. Dichotomous Vectors. Another common type of representation used in connectionist modeling is a random vector composed equally of 1 or -1. These are similar to MINERVA vectors, but without any zero-valued elements. Dichotomous vectors have been used in variety of models, such as connectionist models of semantic priming (e.g., Masson, 1995). We use vectors with a length of 100 in the following simulations.

Method

To calculate similarity distributions using the semantic measures, 1000 words were selected from the Toronto word pool (Friendly, et al., 1982), and the similarity between each word in the pool was computed. Next, 50,000 of these semantic comparison values were randomly sampled to examine the distribution of similarity values. In the WAS, LSA, and BEAGLE models the similarity metric used was a vector cosine (a normalized dot-product), while in COALS Pearson's correlation was used.

For the randomly generated representations, we created a distribution of 100,000 similarity comparisons for each representation type. The distribution was constructed by randomly generating two vectors from the given representation type and computing the similarity between them. Similarity was vector cosine for all representations.

Toevaluate distribution shape, two different methods of assessing normality were employed: 1) skewness, and 2) normal quantile-quantile (Q-Q) plots. Skewness is the third moment about the mean, and signals asymmetry in a distribution. Q-Q plots are used to assess the difference between an observed distribution and a theoretical (in this case Gaussian) distribution. The standardized values of the comparison distribution are plotted against the respective values for the Gaussian, and any discrepancy signals a deviation from the theoretical Gaussian distribution.

Results

The skewness values for the similarity distributions of both the semantic spaces and random representations are displayed in Figure 1. As the figure shows, all the semantic spaces create positively skewed similarity distributions. That is, there tends to be a greater number of low similarity scores and a small number of high similarity scores in a given distribution of randomly selected words. Co-occurrence models (LSA, BEAGLE, and COALS) have the lowest skew (from 1.06 for BEAGLE to 2.01 for COALS). The PMI distribution produced the largest skew, likely due to the fact that this method does not abstract across documents, but is instead a pure co-occurrence count. Even with this shortcoming, PMI has been shown to be very effective in fitting human semantic similarity ratings

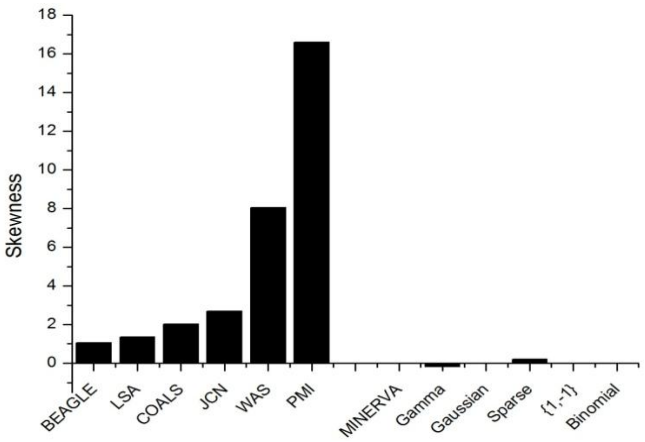


Figure 1. Levels of skewness for the different distributions.

(Recchia & Jones, 2009). In the middle was the JCN measure with a skewness of 2.61 and the WAS of Steyvers, et al. (2004) with a skewness of 8.04, which signals a highly skewed distribution.

In contrast, all of the random representations produced skewness values of essentially zero (this is expected by their construction). The only distribution that is mildly positively skewed is the sparse binomial distribution with a skewness of 0.21, while the Gamma distribution is actually mildly negatively skewed with a value of -0.17.

The Q-Q plots are displayed in Figure 2 for the semantic space distributions (left panel) and the distributions computed from the random representations (right panel). Due to space limitations, only 4 graphs were included, but these are diagnostic of the remaining distributions. Again, the semantic space distributions show significant deviation from the expected Gaussian distribution. Specifically, the semantic space distributions are skewed to the right, with all of the models having lower than expected number of large similarity values. They also tend to have greater than expected low similarity values. Again, the random representation distributions produce very different results—there is little deviation from normality.

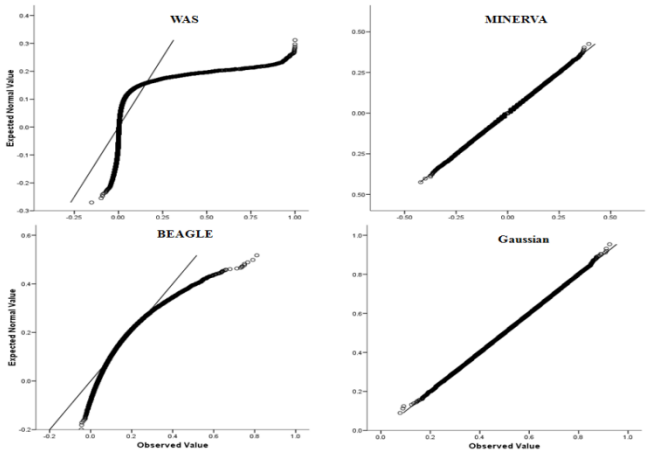


Figure 2. Q-Q plots for semantic and random vectors.

This simple analysis demonstrates that the similarity distributions created by semantic space models and randomly generated representations are considerably different. Two randomly selected words are likely to be less similar (relative to the other values in the distribution) for semantic models, than for random representations.

Demonstrations

In order to show the potential impact that the use of random representations may have, two simple demonstrations were conducted using data from recognition memory tasks.

Demonstration #1: Signal Detection Theory

The purpose of this demonstration is to show what effect skewed similarity distributions will have on a signal detection theory (SDT) based process, which is the dominant decision making process within recognition memory (Shiffrin & Steyvers, 1997; Dennis & Humphreys, 2001). In order to accomplish this, a recognition process with SDT is simulated by sampling from both skewed (semantic) similarity distributions as well as normal (random) similarity distributions. Recognition is then simulated by fitting an optimal criterion to separate old and new items, and the resulting d-prime values for the different distributions will be compared to behavioral results.

In order to compare the different similarity distributions, a normalization procedure was necessary. This was done by taking the distributions from each of the semantic metrics and random representations and normalizing them to have a range of 0 and 0.5 and a mean of 0.25. This procedure allows us to evaluate the shape of the distribution while centering the distributions on the same mean.

Evidence distributions for new and old items were simulated for lists of 20 words. The evidence for a probe was the similarity of the probe to the 20 items on the list. For 'new' probes, this evidence was simply the mean of 20 randomly sampled similarity values (as new probes are randomly similar to the contents of memory). For 'old' probes, this evidence was the average of the similarity of the item to itself and the other items on the list (simulated as the mean of 19 randomly sampled similarities and the value of 1, representing the similarity of the word to itself). This process was repeated 50,000 times for each similarity distribution.

To compare the resulting evidence values, the discriminability (measured with d-prime) was calculated for each simulation—d-prime is a measure of how distinct studied items are from non-studied items. Figure 3 displays the d-prime values for the different similarity distributions compared with the d-prime from a simple recognition experiment which used a list length of 20 (Dennis, Lee, & Kinnel, 2008). As the figure illustrates, all of the semantic distributions have higher d-prime than do the random distributions. In addition, the d-prime values for the random representations are much closer to the behavioral data from Dennis, et al. The difference in magnitude demonstrated for d-prime values for semantic and random similarity was statistically reliable, $t(11) = 4.75$, $p < 0.001$. To evaluate the effect of skew in the similarity distributions on the resulting d-prime values, we computed the partial correlation between d-prime and skewness (controlling for kurtosis and variance) for the distributions, which resulted in a robust $r = 0.913$, $p < 0.001$.

The skewness of the similarity distribution has a large effect on the calculation of evidence distributions because the probability of sampling lower similarity values is much greater than in a symmetric distribution. Hence, with 'true' semantic representations an old item tends to be more distinct from other random items on the list, producing a greater difference between old and new evidence distributions. This demonstration is certainly not meant as a refutation of signal detection theory, but instead demonstrates that using realistic representations of semantics will impose significant constraint on a processing model's ability to simulate data.

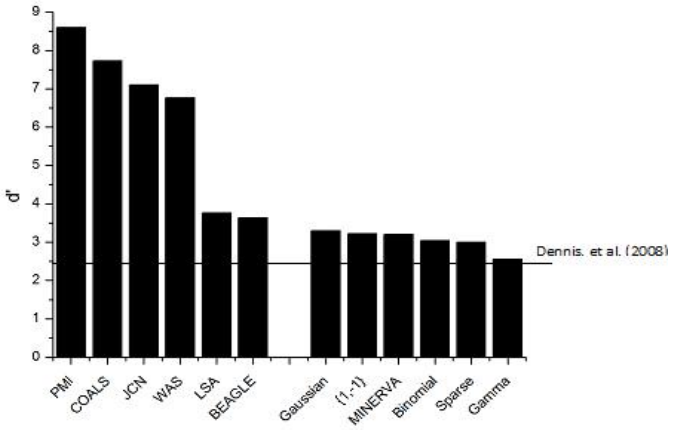


Figure 3. Levels of discriminability (d-prime) for SDT simulations; behavioral data from Dennis, et al. (2008).

Demonstration #2: MINERVA 2 and False Recognition

This demonstration was conducted in order to show that random representations provide an increase amount of freedom to fit data. The MINERVA 2 model of Hintzman (1986) has been used to successfully account for a variety of categorical false recognition effects (Arndt & Hirshman, 1998). Here, we simulate associative false recognition with the model, using both random and structured representations of semantics. Robinson and Roediger (1997) found that as the number of studied items that are related to a critical lure is increased, so is the probability of falsely recognizing that critical lure. The purpose of this demonstration is to compare the ease with which a simple process model like MINERVA is able to model this effect when using random representations versus when it is using representations that contain knowledge about the similarity structure of the actual words.

To construct MINERVA vectors that contain plausible semantic structure, we transformed the WAS representations from Steyvers et al. (2003). Typical applications of MINERVA use ternary vectors with a fairly low dimensionality. Hence, WAS vectors were collapsed from

400 to 20 dimensions by summing every 20 quadrants in the WAS vector into a single element in the reduced vector. This reduced vector was then transformed into a ternary vector with values of the set {-1, 0, 1}; the magnitude of the summed WAS values were recoded so that the highest third were assigned +1 (representing a high weighting on that feature), the middle third 0, and the lowest third -1. To ensure that the MINERA transformed vectors still reflected the semantic structure in the original WAS vectors, we computed the word-by-word cosines between vectors in both representations, and correlated the two matrices: The original vectors and their ternary transformed versions were highly correlated, $r = .67$, $p < .001$, indicating that the transformed vectors contain an arrangement of elements that reflects the semantic structure in the original WAS vectors. Using the false recognition lists from Stadtler, Roediger and McDermott, (1998) and Gallo and Roediger (2002), there was a high average similarity of the critical word's representation to the representations of the list items across the 52 word lists, $r = 0.35$, $p < .001$.

Random representations for critical words and their corresponding lists were created as in Arndt and Hirshman (1998), by using prototype and exemplar vectors. A prototype vector (representing the critical word) is first generated by randomly sampling elements from the set {1, 0, -1} with equal probability. Each item in the word list is then created by randomly perturbing elements in the prototype vector. This process requires a distortion parameter, which determines the probability of switching elements from the prototype vector when creating a list item vector. The distortion parameter determines how similar the list items are to the critical word. The important point is that both the semantic and random representations contain the exact same elements (same number of -1, 0, and 1s). The difference is that the elements are arranged independently for the random representations, whereas they are arranged to respect the inter-word similarity structure from WAS in the semantic version.

For MINERVA with a semantic representation, the results of Robinson and Roediger (1997) were modeled by randomly selecting 3 word lists, and adding 3, 6, or 9 items from one of the lists into a study list. Because the word lists in Robinson and Roediger were longer (they also used 12 and 15 associates), 27 words selected randomly from the Toronto word pool were added into the study list. To simulate this with MINERVA using random representation, 3, 6, or 9 exemplars were created for 3 random prototypes and added into the study list. Additionally, 27 random vectors were added into the study list to make the two simulations equivalent. Decisions are based on activation levels of a probe to the studied items (echo intensity: Hintzman, 1986), calculated by summing the similarity across all items in the study list.

For the MINERVA with semantic representations, there are two free parameters: 1) a criterion to make a new-old decision based on activation levels, and 2) a forgetting parameter which determines the probability of a non-zero element switching to zero during study. The simulation with random representations includes an additional distortion parameter (described above) to create the semantic structure. These parameters were fit to the data from Robinson & Roediger (1997) data using a Nelder-Mead simplex algorithm. The results of the simulation are displayed in Figure 4: the MINERVA model that utilizes random representations was able to reproduce the overall trend in the data. However, this was not the case with the MINERVA model that used semantic representations—this model tended to falsely recognize critical items over studied items, which is not the case with the human data. The random representation version of the model produced an excellent account of the data, $R^2 = 0.98$, $p < .001$. However, the version based on the true semantic similarity of the words used fit no better than chance, $R^2 = 0.05$, $p = .45$.

This simulation provides a simple demonstration of how a process model that has false representation assumptions may be incorrectly accepted as a plausible model. The only difference between the two models is in their representation structure—the process is identical. While the semantic version contains the "true" semantic structure for the exact words used in the experiment, the random version uses the distortion parameter to create the semantic structure that is most likely if this process account is correct. It is exclusively the incorrect inferred semantic structure that allows the process account to fit these data. If the correct representational structure were used, the process account would be rejected. The point is that random representations allow unnecessary freedom for the model to fit the data.

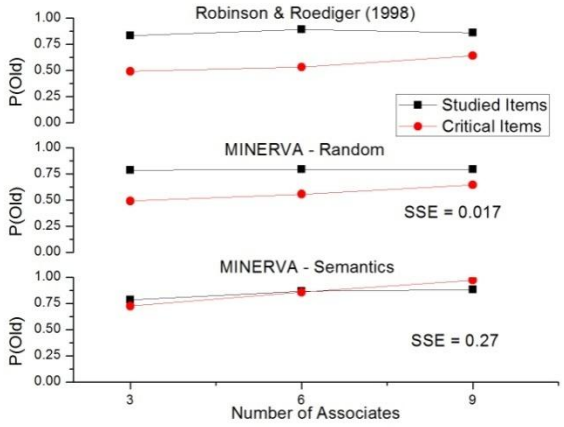


Figure 4. Results of false recognition simulation**.**

General Discussion

The use of randomly generated representations contains the assumption that semantic similarity is normally distributed over randomly selected pairs of words. This assumption was shown to be false across many different semantic metrics that have demonstrated success at accounting for human data. In experiments using words, two randomly selected words are likely to be relatively less similar (compared to the distribution of all possible pairs) than would be implied

using randomly generated representations for lexical semantics. Because similarity plays a central role in the processing mechanisms used by many memory models, the use of random representations may have consequences for conclusions drawn from simulations using these models.

As McClelland (2009) has noted, "…simplification is essential, but it comes at a cost, and real understanding depends in part on understanding the effects of simplification." (p. 18). The use of random representations in the development of cognitive models has been a necessary simplification for our understanding of cognitive processes. In doing so, researchers have made use of representations whose assumptions may not be entirely accurate, but through the use of this simplification modelers have made fundamental discoveries about how memory processes work. However without this assumption these results would not have been possible. It has only been within the last decade that researchers have had access to realistic representations of lexical semantics. The task for the future is to integrate semantic representations with processing models of memory for a fuller understanding of how they work together to produce observable behavior.

In accordance, recent models have begun to conduct this type of integration. For example, Monaco, Abbott, & Kahana (2007) have created a neural network model of the mirror effect of frequency, utilizing lexical semantic representations taken from the WAS of Steyvers, et al. (2004). Ideally, future models will combine a learning process that builds a representation through exposure to environmental information, which can then feed into a processing mechanism. For example, Johns and Jones (2009) have utilized representations built through a co-occurrence learning process to drive a processing model of both false recognition and false recall. These models suggest that it is no longer necessary to assume random representations for lexical semantics when modeling cognitive phenomena, but that item-specific semantic representations are now freely available and offer additional modeling constraints about the structure of semantic similarity that a process mechanism must operate on to produce behavior in a given task.

References

Arndt, J., & Hirshman, E. (1998). True and false recognition in MINERVA2: Explanations for a global matching perspective. *Journal of Memory and Language*, 39, 371-391.

Dennis, S., & Humphreys, M. S. (2001). A context noise model of episodic word recognition.*Psychological Review*, *108*, 452-478.

Dennis, S., Lee, M. D., & Kinnel, A. (2008). Bayesian analysis of recognition memory: The case of the list-length effect. *Journal of Memory and Language*, *59*, 361-376.

Friendly, M., Franklin, P. E., Hoffman, D., & Rubin, D. C. (1982). Norms for Toronto word pool: Norms for imagery, concreteness, orthographic variables, and grammatical usage for 1,080 words. *Behavior Research Methods and Instrumentation, 14*, 375-399.

Johns, B. T., & Jones, M. N. (2009). Simulating false recall as an integration of semantic search and recognition. *Proceedings of the 31st Annual Cognitive Science Society.*

Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review, 114*, 1-37.

Gallo, D.A., & Roediger, H.L. (2002). Variability among word lists in eliciting memory illusions: evidence for associative activation and monitoring. *Journal of Memory and Language, 47*, 469-497.

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review, 93*, 411-428.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent Semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review,104*, 211-240.

Masson, M. E. J. (1995). A distributed memory model of semantic priming. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 21*, 3-23.

McClelland, J. L. (2009). The place of modeling in cognitive science. *Topics in Cognitive Science, 1*, 11-38.

Monaco, J. D., Abbott, L. F., & Kahana, M. J. (2007). Lexico-semantic structure and the word-frequency effect in recognition memory. *Learning and Memory, 14*, 204–13.

Miller, G. A. (Ed.) (1990). WordNet: An on-line lexical database. *International Journal of Lexicography, 3*.

Murdock, B. B. (1982). A theory for the storage and retrieval of item and associative information. *Psychological Review, 89*, 609-626.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). The University of South Florida word association, rhyme, and word fragment norms. http://www.usf.edu/FreeAssociation/.

Plaut, D. C. (1995). Semantic and associative priming in a distributed attractor network. *Proceedings of the 17th Annual Conference of the Cognitive Science Society.*

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology, 77*, 353-363.

Recchia, G. L., & Jones, M. N. (2009). More data trumps smarter algorithms: Comparing pointwise mutual information to latent semantic analysis. *Behavior Research Methods,* 41, 657-663.

Robinson, K., & Roediger, H. L. (1997). Associative processes in false recall and false recognition. *Psychological Science*, 8, 389-393.

Rohde, D. L. T., Gonnerman, L., and Plaut, D. C. (submitted). An improved model of semantic similarity based on lexical co-occurrence. *Cognitive Science.*

Shiffrin, R.M. & Steyvers, M. (1997). A model for recognition memory: REM: Retrieving Effectively from Memory. *Psychonomic Bulletin & Review, 4*, 145-166.

Stadler, M.A., Roediger, H.L., & McDermott, K.B. (1999). Norms for word lists that create memories. *Memory & Cognition*, 29, 424-432.

Steyvers, M., Shiffrin, R. M., & Nelson, D. L. (2004).Word Association Spaces for Predicting Semantic Similarity Effects in Episodic Memory. In A. Healy (Ed.), *Experimental Cognitive Psychology and its Applications.*

A Neural Model of Rule Generation in Inductive Reasoning

Daniel Rasmussen (drasmuss@uwaterloo.ca)
Chris Eliasmith (celiasmith@uwaterloo.ca)
Centre for Theoretical Neuroscience, University of Waterloo
Waterloo, ON, Canada, N2J 3G1

Abstract

Inductive reasoning is a fundamental and complex aspect of human intelligence. In particular, how do subjects, given a set of particular examples, generate general descriptions of the rules governing that set? We present a biologically plausible method of accomplishing this task, and implement it in a spiking neuron model. We demonstrate the success of this model by applying it to the problem domain of Raven's Progressive Matrices, a widely used tool in the field of intelligence testing. The model is able to generate the rules necessary to correctly solve Raven's items, as well as recreate many of the experimental effects observed in human subjects.

Keywords: inductive reasoning; neural engineering framework; fluid intelligence; Raven's Progressive Matrices; vector symbolic architectures; cognitive modeling

Introduction

Inductive reasoning is the process of using a set of examples to infer a general rule which both describes the relationships shared by those examples and allows us to predict future items in the set. For example, if a person were watching objects in a river or lake and saw a stick, a wooden rowboat, and a telephone pole float past, they might induce the rule that "wooden things float". This rule both describes the relationship which linked those items (being wooden) and allows the person to predict future items which would also float (a wooden bookcase). Given even more examples—some non-wooden floating objects—they might infer the general rule that objects float when they displace a volume of water equal to their weight.

This type of reasoning is fundamental to our ability to make sense of the world, and represents a key facet of human intelligence. It determines our ability to be presented with a novel situation or problem and extract meaning from it. As such, it is a process which has been made central to many tests of general intelligence. One of the most widely used and well respected tools in this field is the Raven's Progressive Matrices (RPM) test (Raven, 1962). In the RPM, subjects are presented with a 3x3 matrix, in which each cell in the matrix contains various geometrical figures with the exception of the final cell which is blank (Figure 1). The subject's task is to determine which one of eight possible answers belongs in the blank cell. They accomplish this by examining the other rows and columns and inducing rules which govern the features in those cells. They can then apply those rules to the last row/column to determine which answer belongs in the blank cell.

Although there has been much experimental and theoretical effort put into understanding the mental processes involved in performing RPM-like tasks, to our knowledge there have been no models of the inductive process of rule generation. In this paper we present a method of rule generation, and implement it in a neural model using simulated spiking neurons. This model can induce the rules necessary to solve Raven's matrices, and also displays many of the most interesting cognitive effects observed in humans: improved accuracy in rule generation over multiple trials, variable performance in repeated trials, and both quantitative and qualitative changes in individual performance.

Background

Raven's Progressive Matrices

There are several variations of the RPM; the Standard and Coloured versions are generally used to test children or adults with cognitive deficits, while the Advanced is used to differentiate average/above-average adults. In our work we focus on the Advanced version.

Figure 1 depicts an example of a simple Raven's-style matrix.[1] The matrix is shown at the top with one blank cell, and the 8 possible candidates for that blank cell are along the bottom. In order to solve this matrix the subject needs to generate three rules: 1) the number of instances of each shape increases by one across the row, 2) the orientation of the shapes within a cell is constant across the row, 3) each cell in a row contains one shape type from the set {square, triangle, circle}. Subjects can then determine which elements belong in the blank cell by applying the rules to the third row (i.e. there should be $2+1=3$ shapes, they should be arranged in the same orientation (vertically), and they should be triangles, since circle and square are already taken). Once they have

[1] For copyright reasons we have created a modified matrix to present here, the model works with the true Raven's matrices.

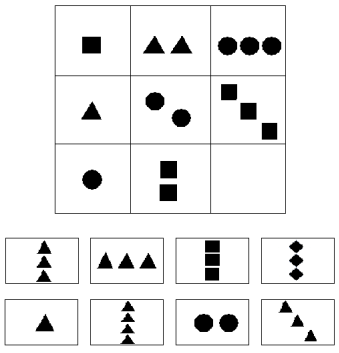


Figure 1: A simple Raven's-style matrix

generated their hypothesis as to what the blank cell should look like, they can check for a match among the 8 possible answers. Not all subjects will explicitly generate these exact rules, and their route to the answer may be more roundabout, but they do need to extract equivalent information if they are to correctly solve the problem.

Despite the test's broad use, the only other computational model for the RPM is that of Carpenter et al. (1990). Their model accurately recreates high-level human data, but does not reflect the flexibility and variability of individual human performance nor take into account neurological data. In addition, Carpenter et al.'s model has no ability to generate new rules; all the rules are pre-programmed. This limitation of their model reflects a general lack of explanation in the literature as to how this inductive process is performed.

The two default assumptions regarding the origin of the rules are that people are either 1) born with, or 2) learn earlier in life, a library of rules. During the RPM, these pre-existing rules are then applied to the current inductive problem. Hunt described this theory as early as 1973, and also pointed out the necessary conclusion of this explanation: if RPM performance is dependent on a library of known rules, then the RPM is testing our crystallized intelligence (our ability to acquire and use knowledge or experience) rather than fluid intelligence (our novel problem solving ability). In other words, the RPM would be a similar task to acquiring a large vocabulary and using it to communicate well. However, this is in direct contradiction to the experimental evidence, which shows the RPM strongly and consistently correlating with other measures of fluid intelligence (Marshalek et al., 1983), and psychometric/neuroimaging practice, which uses the RPM as an index of subjects' fluid reasoning ability (Perfetti et al., 2009; Prabhakaran et al., 1997; Gray et al., 2003). A large amount of work has been informed by the assumption that the RPM measures fluid intelligence, yet the problem raised by Hunt has been largely ignored. Consequently, there is a need for a better explanation of rule induction; by providing a technique to dynamically generate rules, we remove the dependence on a past library, and thereby resolve the problem.

In contrast to the paucity of theoretical results, there has been an abundance of experimental work on the RPM. This has brought to light a number of important aspects of human performance on the test that need to be accounted for by any potential model. First, there are a number of learning effects: subjects improve with practice if given the RPM multiple times (Bors & Vigneau, 2003), and also show learning within the span of a single test (Verguts & De Boeck, 2002). Second, there are both qualitative and quantitative differences in individuals' ability; they exhibit the expected variability in "processing power" (variously attributed to working memory, attention, learning ability, or executive functions), but also consistent differences in high-level problem-solving strategy between low-scoring and high-scoring individuals (Vigneau et al., 2006). Third, a given subject's performance is far from deterministic; given the same test multiple times, subjects will get previously correct answers wrong and vice versa (Bors & Vigneau, 2003). In the Results section we demonstrate how each of these observations is accounted for by our model.

Vector encoding

In order to represent a Raven's matrix in neurons and work on it computationally, we need to translate the visual information into a symbolic form. Vector Symbolic Architectures (VSAs; Gayler, 2003) are one set of proposals for how to construct such representations. VSAs represent information as vectors, and implement mathematical operations to combine those vectors in meaningful ways.

To implement a VSA it is essential to define a binding operation (which ties two vectors together) and a superposition operation (which combines vectors into a set). We use circular convolution for binding, and vector addition for superposition (Plate, 2003). Circular convolution is defined as

$$\begin{aligned} C &= A \otimes B \\ \text{where} & \\ c_j &= \sum_{k=0}^{n-1} a_k b_{j-k \bmod n} \end{aligned} \tag{1}$$

Along with this we employ the idea of a transformation vector T between two vectors A and B, defined as

$$\begin{aligned} A \otimes T &= B \\ \text{or} & \\ T &= A' \otimes B \end{aligned} \tag{2}$$

where A' denotes the approximate inverse of A.

With these elements we can create a vector representation of the information in any Raven's matrix. For example, suppose we wanted to encode the information contained in the third cell of Figure 1. The first step is to define a vocabulary, the elemental vectors which will be used as building blocks. These vectors are randomly generated, and the number of vectors that can be held in a vocabulary and still be distinguishable as unique "words" is determined by the dimensionality of those vectors (the more words in the vocabulary, the higher the dimension of the vectors needed to represent them).

Once the vocabulary has been generated it is possible to encode the structural information in the third cell. A simple method to do this is by using a set of $attribute \otimes value$ pairs: $shape \otimes circle + number \otimes three + colour \otimes black + orientation \otimes horizontal + shading \otimes solid$ and so on, allowing us to encode arbitrary amounts of information. As descriptions become more detailed it is necessary to use more complex encoding; however, ultimately it does not matter to the inductive system how the VSA descriptions are implemented, as long as they encode the necessary information. Thus these descriptions can be made as simple or as complex as desired without impacting the underlying model.

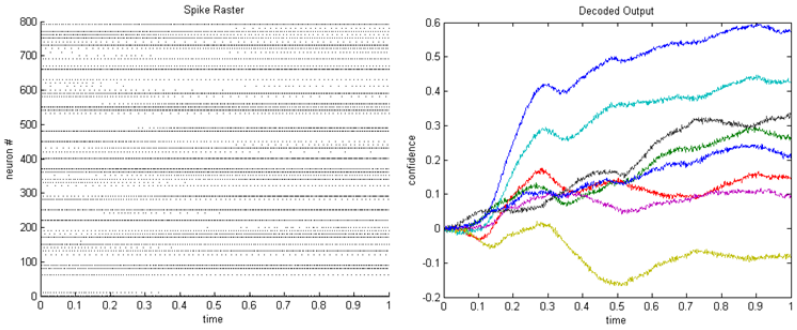


Figure 2: Recordings from the output population of the model, which expresses the similarity between the predicted answer and each of the 8 possible choices. On the left is the spike raster, and on the right is the decoded information from those spikes. The model correctly picks answer number one (the top line).

VSAs have a number of other advantages: vectors are easier to represent in populations of neurons than complex visual information, they are easier to manipulate mathematically, and perhaps most importantly the logical operation of the inductive system is not dependent on the details of the visual system. All that our neural model requires is that the Raven's matrices are represented in some structured vector form; the visual processing which accomplishes this, though a very difficult and interesting problem in itself (see Meo et al. 2007 for an example of the complexities involved), is beyond the scope of the current model. This helps preserve the generality of the inductive system: the techniques presented here will apply to any problem that can be represented in VSAs, not only problems sharing the visual structure of the RPM.

Neural encoding

Having described a method to represent the high-level problem in structured vectors, we now define how to represent those vectors and carry out the VSA operations in networks of simulated spiking neurons. There are several important reasons to consider a neural model. First, by tying the model to the biology we are better able to relate the results of the model to the experimental human data, both at the low level (eg. fMRI or PET) and at the high level (eg. non-deterministic performance and individual differences). Second, our goal is to model human inductive processes, so it is essential to determine whether or not a proposed solution can be realized in a neural implementation. Neuroscience has provided us with an abundance of data from the neural level that we can use to provide constraints on the system. This ensures that the end result is indeed a model of the human inductive system, not a theoretical construct with infinite capacity or power.

We use the techniques of the Neural Engineering Framework (Eliasmith & Anderson, 2003) to represent vectors and carry out the necessary mathematical operations in spiking neurons. To encode a vector $x(t)$ into the spike train of neuron a_i we define

$$a_i(x(t)) = G_i\left[\alpha_i\tilde{\phi}_i x(t) + J_i^{bias}\right] \tag{3}$$

G_i is a function representing the nonlinear neuron characteristics—essentially, how will the neuron spike given the input described within the brackets. In our model we use Leaky Integrate and Fire neurons, but the advantage of this formulation is that any neuron model can be substituted for G_i without changing the overall framework. α_i is a gain on the input, determined by the characteristics of this particular neuron. J_i^{bias} is the background current, modelling the activity in the network which is not a direct input to this neuron. $\tilde{\phi}_i$ represents the neuron's preferred stimulus, that is, which inputs will make it fire more strongly. Broadly speaking, the activity of neuron a_i is a result of its unique response (determined by its preferred stimulus) to the input $x(t)$, passed through a nonlinear neuron model in order to generate spikes.

We can then define the decoding from spike train to vector as

$$\hat{x}(t) = \sum_i h(t) * a_i(x(t))\phi_i \tag{4}$$

where $h(t)$ is a model of the post-synaptic current generated by one spike, $a_i(x(t))$ are the spikes generated by Equation 3, and ϕ_i are the optimal linear decoders. The optimal linear decoders are calculated analytically so as to provide the best linear representation of the original input $x(t)$; they are essentially a weight on the post-synaptic current generated by each neuron (the result of summing the current generated by each spike).

We have defined how to transform a vector into neural activity and how to turn that neural activity back into a vector, but we also need to be able to carry out the VSA operations (binding and superposition) on those representations. One of the primary advantages of the NEF is that we can calculate the synaptic weights for arbitrary transformations analytically, rather than learning them. If we want to calculate a transformation of the form $z = C_1 x + C_2 y$ (C_1 and C_2 are any matrix), and x and y are represented in the a and b neural populations (we can add or remove these terms as necessary to perform operations on different numbers of variables), respectively, then we describe the activity in the output popula-

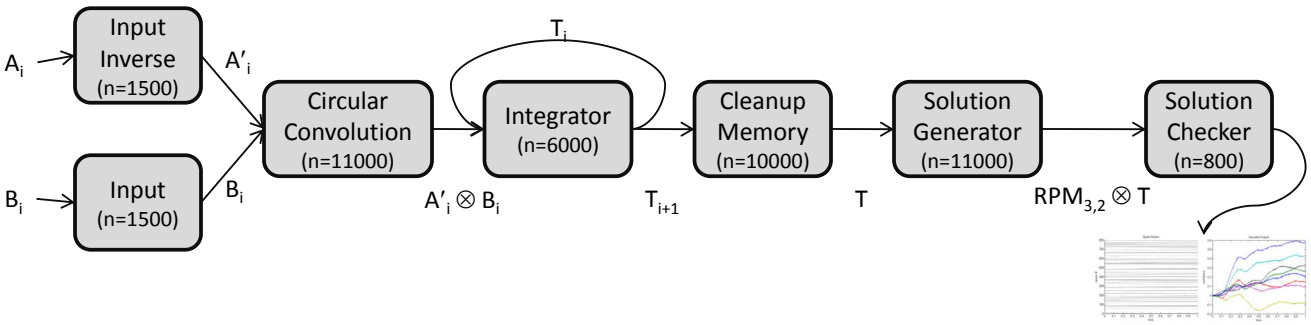


Figure 3: Schematic diagram of the rule generation section with cleanup memory, displaying the approximate number of neurons used in each submodule. The inputs (A_i and B_i) represent two adjacent cells in the matrix. The "Input Inverse" module calculates A_i', while "Input" simply leaves B_i unchanged. The "Circular Convolution" module calculates $A_i' \otimes B_i$ (the rule for that particular pair of cells). "Integrator" is storing the calculated rule so far (based on previous pairs of adjacent cells), which we combine with the current calculation. The output of "Integrator" is the overall rule, which we pass through a cleanup memory, potentially giving us a less noisy version of that rule. Finally, "Solution Generator" generates a prediction of what should be in the blank cell by convolving the second-last cell with our calculated rule, and then "Solution Checker" calculates the similarity between that hypothesis and each of the eight possible answers given in the problem.

tion as

$$c_k(C_1x+C_2y) = G_k\left[\sum_i \omega_{ki}a_i(x) + \sum_j \omega_{kj}b_j(y) + J_k^{bias}\right]$$

where c_k, a_i, and b_j describe the activity of the kth, ith, and jth neuron in their respective populations. The ω are our synaptic weights: $\omega_{ki} = \alpha_k\langle\tilde{\phi}_k C_1 \phi_i^x\rangle_m$ and $\omega_{kj} = \alpha_k\langle\tilde{\phi}_k C_2 \phi_j^y\rangle_m$. Referring back to our descriptions of the variables in Equations 3 and 4, this means that the connection weight between neuron a_i and c_k is determined by the preferred stimulus of c_k, multiplied by the desired transformation and the decoders for a_i. To calculate different transformations all we need to do is modify the C matrices in the weight calculations, allowing us to carry out all the linear computations necessary in this model. For a more detailed description of this process, and a demonstration of implementing the nonlinear circular convolution (Equation 1), see Eliasmith (2005).

The Model and Results

Rule generation

The key to our model is the idea of the transformation vector (Equation 2). Since we have our Raven's matrix items encoded as vectors, we can represent rules as transformations on those vectors. For example, if A is the vector representation of one square, and B is the vector representation of two squares, then the transformation vector $T = A' \otimes B$ will be analogous to the rule "number of squares increases by one". However, we do not just want to calculate individual transformations, we want general rules for the whole matrix. To accomplish this we treat all adjacent pairs of cells as a set of A and B vectors, and extract a general transformation from that set of examples. Neumann (2001) has shown that we can accomplish this by calculating

$$T = \frac{1}{n}\sum_{i=0}^{n} A_i' \otimes B_i$$

In order to perform this operation in neurons (where we cannnot instantly sum over a set of examples) we translate it into the equivalent learning rule, where each pair of A and B vectors is presented sequentially:

$$T_{i+1} = T_i - w_i(T_i - A_i' \otimes B_i)$$

We implement this by combining a neural integrator (to maintain the overall value of T) with a network which calculates the T_i for the current pair of examples. We present the examples in a top-down row-wise fashion, as that is the general scanning strategy employed by humans as revealed by eye-tracking studies (Carpenter et al., 1990; Vigneau et al., 2006). Let us again take Figure 1 as an example, and examine how the model induces one of the rules necessary to solve the matrix: "number of objects increases by one". A_0 is the vector representation of one square, and B_0 is the vector representation of two triangles (we will omit orientation in this example to keep things simple, but it is treated in exactly the same way). The network calculates $T_1 = A_0' \otimes B_0$, which is something like the rule "number of objects increases by one and squares become triangles", and that value is stored in the neural integrator. In the next step A_1 is two triangles and B_1 is three circles, and T_2 is "number of objects increases by one and triangles become circles". However, when T_2 is added to the neural integrator, "number of objects increases by one" is reinforced (since it was already present) while the other information is not. This process continues with the next two rows. Thus we begin with a very noisy rule, but over time relations which are particular to individual A and B pairs are

drowned out by the relation which all the pairs have in common: "number of objects increases by one".[2]

Once this process is complete we have the overall T vector, representing a general rule for the problem. Thus we have accomplished our primary goal, to provide an explanation as to how subjects can inductively generate descriptions of the rules governing a set of examples. We use these rules by applying them to the second-last cell of the Raven's matrix $A \otimes T$ giving us B, a vector representing what our rules tell us should be in the blank cell. We then compare this hypothesis to the eight possible answers and take the most similar (determined by the dot product between the two vectors) as our final answer (see Figures 2 and 3).

Cleanup memory

In addition to being able to generate the rules to solve a matrix, the model should improve at this process given practice. We accomplish this by adding a cleanup memory, a system which stores certain values and, when given a noisy version of those values as input, outputs the clean version stored in memory. A cleanup memory can be implemented in neurons by creating a network which contains neural populations tuned to respond only to certain inputs and output the clean version of those values (Stewart et al., 2009). We implement a cleanup memory in this model by storing the past rules the system has induced. The current rule generated by the network, which will be perturbed by neural noise and the details of the particular Raven's matrix, is passed through this cleanup memory, and if the cleanup memory contains a similar rule then that clean version of the rule is output.

The cleanup memory is improved over time by two mechanisms. First, if the cleanup memory receives an input that it does not recognize, it adds that input to its memory so that it will be recognized in the future. Second, if the cleanup memory receives an input that it does recognize, it uses that input to refine the value stored in memory, so that the stored value becomes increasingly accurate. Thus as the system encounters rules it has calculated before it will be able to draw on its past efforts to provide a more accurate output. See Figure 4 for a demonstration of how this improvement in cleanup memory can lead to improved inductive performance.

The cleanup memory not only helps account for observed learning effects, it also bridges the gap between this model of inductive rule generation and theories of a "library" of known rules. In short, we are improving on current theories by explaining where that past knowledge comes from, and why its use is a dynamic, fluid process.

Higher level processes

In addition to the inductive process of rule generation, there are high-level problem solving effects (what we might call the subject's "strategy") which will have a significant impact on performance. For example, how does the subject decide when and where to apply the rule generation system? When there are multiple rules to be found, how does the subject differentiate them, and how do they decide they have found all the rules? How does the subject decide whether their hypothesis is good enough to settle on as a final answer? These are important questions, but they are dependent on the particular problem the subject is solving.

[2]This same process will help eliminate the noise added at the neural level.

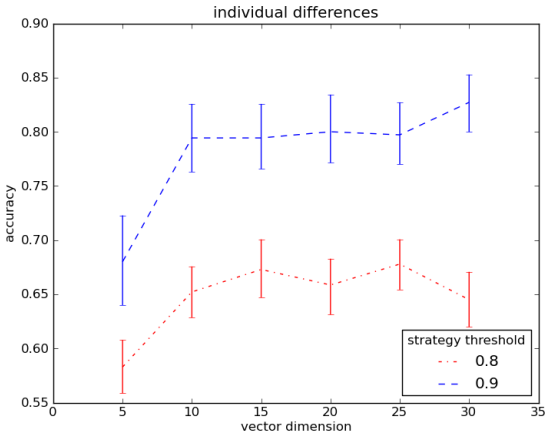


Figure 5: A demonstration of both low-level (vector dimension) and high-level (strategy) influences on accuracy (displaying 95% confidence intervals).

We have implemented such a strategy system for the RPM (although not at the neural level) in order to collect aggregate test results and explore individual differences. Figure 5 shows an example of these results, demonstrating the model's ability to recreate differences caused by both low-level neural processing power and high-level strategy. The low-level variable is the dimensionality of the vectors, higher dimension vectors requiring more neurons to represent. The high-level variable is how willing the model is to decide it has found a correct rule: the lower line represents a subject who has less stringent standards, and is willing to accept rules that may not be completely correct, whereas the top line represents a subject employing a more conservative strategy. These results demonstrate that both low and high level variables have a significant impact on accuracy, and reflect the quantitative and qualitative individual differences observed in human performance. Figure 5 also reveals that although the overall performance trends are clear, there is significant variability (average $\sigma = 0.13$) in any given trial, another parallel of human subjects. There are many such interesting avenues of exploration, however we will not go into the details of the strategy system here; the primary contribution of this research is the general rule-induction system described above, which is not dependent on the higher level framework within which it is used.

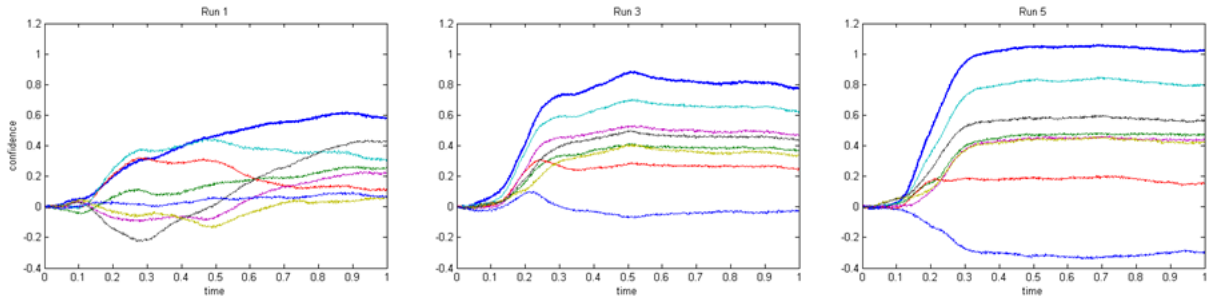


Figure 4: An example of the model's ability to learn over time. The model was presented with a series of matrices that appeared different but required the same underlying rules to solve; as we can see, the model is able to more quickly and definitively pick out the correct answer on later matrices.

Conclusion

We have presented a novel, neurally-based model of inductive rule generation, and we have applied this system to the particular problem of Raven's Progressive Matrices. The success of the system is demonstrated in its ability to correctly find general rules that enable it to solve these matrices, as well as in the model's ability to recreate the interesting effects observed in human subjects, such as learning over time, non-deterministic performance, and both quantitative and qualitative variability of individual differences. These results demonstrate the potential for gaining a deeper understanding of human induction by adopting a neurally plausible approach to modeling cognitive systems.

Acknowledgments

This work was supported by the Natural Sciences and Engineering Research Council of Canada, CFI/OIT, Canada Research Chairs, and the Ontario Ministry of Training, Colleges, and Universities.

References

Bors, D., & Vigneau, F. (2003). The effect of practice on Raven's Advanced Progressive Matrices. *Learning and Individual Differences*, *13*, 291-312.

Carpenter, P., Just, M., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven's Progressive Matrices test. *Psychological Review*, *97*, 404-431.

Eliasmith, C. (2005). Cognition with neurons: A large-scale, biologically realistic model of the Wason task. In B. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the 27th annual conference of the Cognitive Science Society.* Stresa: Cognitive Science Society.

Eliasmith, C., & Anderson, C. (2003). *Neural engineering: Computation, representation, and dynamics in neurobiological systems*. Cambridge: MIT Press.

Gayler, R. (2003). Vector Symbolic Architectures answer Jackendoff's challenges for cognitive neuroscience. In P. Slezak (Ed.), *ICCS/ASCS international conference on cognitive science* (p. 133-138).

Gray, J., Chabris, C., & Braver, T. (2003). Neural mechanisms of general fluid intelligence. *Nature Neuroscience*, *6*, 316-322.

Hunt, E. (1973). Quote the Raven? Nevermore! In L. Gregg (Ed.), *Knowledge and cognition* (p. 129-157). Potomac: Lawrence Erlbaum Associates.

Marshalek, B., Lohman, D., & Snow, R. (1983). The complexity continuum in the radex and hierarchical models of intelligence. *Intelligence*, *7*, 107-127.

Meo, M., Roberts, M., & Marucci, F. (2007). Element salience as a predictor of item difficulty for Raven's Progressive Matrices. *Intelligence*, *35*, 359-368.

Neumann, J. (2001). *Holistic processing of hierarchical structures in connectionist networks*. Unpublished doctoral dissertation, University of Edinburgh.

Perfetti, B., Saggino, A., Ferretti, A., Caulo, M., Romani, G., & Onofrj, M. (2009). Differential patterns of cortical activation as a function of fluid reasoning complexity. *Human Brain Mapping*, *30*, 497-510.

Plate, T. (2003). *Holographic reduced representations*. Stanford: CLSI Publications.

Prabhakaran, V., Smith, J., Desmond, J., Glover, G., & Gabrieli, J. (1997). Neural substrates of fluid reasoning: An fMRI study of neocortical activation during performance of the Raven's Progressive Matrices test. *Cognitive Psychology*, *33*, 43-63.

Raven, J. (1962). *Advanced progressive matrices (sets I and II)*. London: Lewis.

Stewart, T., Tang, Y., & Eliasmith, C. (2009). A biologically realistic cleanup memory: Autoassociation in spiking neurons. In *9th International Conference on Cognitive Modelling.*

Verguts, T., & De Boeck, P. (2002). The induction of solution rules in Raven's Progressive Matrices test. *European Journal of Cognitive Psychology*, *14*, 521-547.

Vigneau, F., Caissie, A., & Bors, D. (2006). Eye-movement analysis demonstrates strategic influences on intelligence. *Intelligence*, *34*, 261-272.

The Effects of Domain and Type of Knowledge on Category-Based Inductive Reasoning

Aimée Kay Crisp-Bright (a.k.crisp@dur.ac.uk)
Department of Psychology, Science Site, South Road
Durham, DH1 3LE, UK

Aidan Feeney (a.feeney@qub.ac.uk)
School of Psychology, Queen's University Belfast
Belfast, BT7 1NN, UK

Abstract

Accounts of category-based inductive reasoning can be distinguished by the emphasis they place on structured versus unstructured knowledge. In addition, it has been claimed that certain domains of structured knowledge are more available than others. Using a speeded task paradigm, participants rated the strength of inductive arguments in which the categories were either strongly or weakly associated and shared a taxonomic or causal relation.. Strongly associated categories received higher inductive strength ratings than weakly associated category pairs, regardless of the domain by which the categories were related. Strength of association was highly predictive of inductive strength ratings, but more additional variance was accounted for by beliefs about taxonomic and causal relations when people were not under time pressure. This suggests that, regardless of knowledge domain, maximizing inductive potency relies on the use of both structured and unstructured knowledge, depending on available mental resources.

Keywords: Category-Based Induction; Knowledge; Categorical Inferences; Reasoning.

Knowledge and Category-Based Induction

Category-based generalizations cover a class of inferences in which an object's category membership supports people's inferences about properties shared with other category members. For example, classifying an animal as a rabbit allows us to infer that it probably lives in a burrow. Furthermore, if we observe that the animal we have classified as a rabbit eats carrots, we are likely to infer that other rabbits and, perhaps hares, also eat carrots.

In order to understand what determines the likelihood that a property will be generalized from a known to a novel instance, we need to identify which aspects of our background knowledge are central to the induction process. Whereas some approaches view category-based induction as driven solely by associative or unstructured knowledge, such as featural overlap (Sloman, 1993), perceptual similarity (Sloutsky & Fisher, 2004) or semantic associations (Rogers & McClelland, 2004), apparently contradictory approaches place theory-based or structured knowledge at the centre of the inductive process, such as knowledge about stable category-hierarchies (Osherson, et al., 1990) and causal relations between categories (Kemp &Tenenbaum, 2009). These contrasting types of knowledge in turn possess unique processing characteristics which differentially affect the reasoning output.

Unstructured Knowledge and Induction

Unstructured knowledge cannot be described by a higher order structure, abstract interrelationships or theories. It can include relations between entities based on contiguity, co-occurrence, similarity or associations. Several studies suggest that early category formation and induction is driven by the statistical properties inherent in the environment, such as co-occurrence and statistical distribution of perceptual features. For example, Sloutsky and Fisher's (2004) model of Similarity, Induction and Categorization (SINC) assumes that children perform categorization and inductive reasoning on the basis of perceptual similarity, in which the category label is simply treated as another feature contributing to increased similarity between different instances. These researchers also claim that there is only a gradual and developmentally late transition from exclusive reliance on similarity to the use of category membership as a basis for induction. This transition is largely seen as the product of explicit instruction and learning about general characteristics of categories (Fisher & Sloutsky, 2005).

Some proponents of associative approaches to category-based induction advocate that adult categorization and induction is also heavily influenced by similarity (Sloman, 1993) and associations in semantic memory (Rogers & McClelland, 2004). For example, Sloman's (1993) feature-based model explains generalizations purely in associative terms as the degree to which the presentation of the premise instances activates overlapping features of the conclusion instance. Arguments in which premise and conclusion categories share more features are stronger than arguments with little featural overlap between premise and conclusion. Consequently, there is no need to assume a stable category hierarchy. Sloman (1998) does not preclude the possibility that assessment of similarity can at times reflect a more effortful process which draws on knowledge about stable category hierarchies. However, he does suggest that the default mode of category-based induction reflects a predominantly intuitive thought process, requiring no processing effort or reference to class inclusion relations,

especially when people lack relevant knowledge, are under time pressure or have not been explicitly instructed to carefully consider their responses.

Structured Knowledge in Induction

An opposing approach to explaining inductive reasoning focuses on the influence of structured knowledge. The justification for assuming that structured knowledge can play an important role in category-based induction arises from several reasoning phenomena that cannot be explained exclusively by the use of unstructured or associative knowledge.

Osherson et al's (1990) Similarity-Coverage Model posits knowledge about stable taxonomic structure as an important source of information that people rely on when evaluating categorical arguments. Inductive evaluations reflect the weighted sum of two primary parameters, similarity and coverage. Similarity refers to the maximum average similarity between the premise and conclusion categories. Coverage refers to the degree to which the premise categories cover the featural space of the inclusive superordinate category and thus, calculation of coverage requires structured knowledge in the form of a stable hierarchy of categories. The coverage component of the model gives rise to the diversity effect, whereby dissimilar premise categories act as stronger evidence than similar premise categories. Although this phenomenon can be explained by Sloman's model, the developmental trajectory of the diversity effect (Lopez, Gelman, Gutheil & Smith, 1992) is more compatible with the assumption that people draw on structured knowledge about stable category hierarchies. Similarly, if sensitivity to diversity was based exclusively on unstructured associative knowledge, it would not be related to general cognitive ability (Feeney, 2007).

Approaches emphasizing the importance of unstructured knowledge also have no means of explaining effects that arise from considering underlying higher-order interrelationships between categories. Tenenbaum and Kemp (2009) and Shafto et al. (2008) have demonstrated that inductive reasoning about causal transmission can be dissociated from inductive inferences about physiological properties. Such dissociations suggest that the context or property people are reasoning about prompts them to draw on different and most relevant sources of structured knowledge. Making use of this kind of structured knowledge also gives rise to phenomena such as the causal asymmetry effect, whereby inferences about the transmission of diseases are deemed stronger from prey to predator than from predator to prey (Medin, Coley, Storms & Hayes, 2003; Shafto, et al., 2008). Again, it is hard to see how approaches relying exclusively on nondirectional unstructured knowledge might cogently explain such effects.

Processing Differences

On the surface it appears that approaches placing divergent emphasis on different types of knowledge are incompatible. However, recent evidence suggests that both structured and unstructured types of knowledge play an important role in inductive reasoning, and that they may be a source of individual differences. One of the major distinguishing features appears to be the nature of the mental processes that mediate the use of these contrasting types of knowledge. For example, Rehder (2009) explicitly suggests that the use of structured knowledge relies on an elaborate, analytical thought processes, whereas associative knowledge influences inductive reasoning fairly automatically and without much cognitive effort. Rehder (2009) taught participants about the causal links between category features of artificial categories. In line with the assumption that people draw on extensive causal knowledge, he demonstrated various phenomena, such as a causal asymmetry effect. However, he also found that there was a substantial minority of people whose patterns of inductions did not adhere to those predicted by his causal-based generalization model. Instead, they seemed to rely more on nondirectional associations between the category features.

This suggests that selective inductive reasoning can either be driven by structured knowledge based on theoretical conceptions about relations between categories within a domain, or on unstructured knowledge based on temporal contiguity or degree of association between the categories.

Testing for Effects of Knowledge Type

To test our hypothesis that category-based induction might be driven by different types of knowledge we used a paradigm developed by Shafto, Coley & Baldwin (2007) who were interested in the effects of knowledge domain on induction. Shafto et al (2007) presented participants with arguments consisting of taxonomically or ecologically related categories and manipulated time to respond. To test our hypothesis about differential effects of knowledge type, we also included a manipulation of between-category association. As access to structured knowledge seems to require slower and more elaborate reasoning, we expected people to rely more on unstructured knowledge when under time pressure.

Our design also allowed us to attempt to replicate Shafto et al's finding that whereas people's inferences about taxonomically related categories were unaffected when under time pressure, they gave lower inductive strength ratings to ecologically related categories when they had to respond rapidly. Because Shafto et al. did not control for level of association between their category pairs, it will be of interest to examine whether processing differences between knowledge domains still emerge when degree of association is equated between domains.

Methods

Participants

40 participants took part in the study. They were volunteers from Durham University, who received course credit for their participation. Their mean age was 24.2 years (SD= 7.8 years).

Design

The experiment had a 2 (timing: speeded versus delayed) by 2 (property: cells or disease) by 2 (relation: taxonomic or causal) by 2 (level of association: high versus low) mixed design, with timing as the between-subjects variable.

Materials and Procedure

There were 20 reasoning items consisting of a base category, a causally related target category and a taxonomically related target category. Causally related pairs were always from different superordinate categories, for example, plants and animals, or mammals and reptiles. In contrast, taxonomically related pairs were always from the same superordinate taxonomic category

For each item, there was a causal problem and a matching taxonomic induction problem, resulting in a total of 40 problems.

In order to control for level of association between the base category and its two target categories, 18 Durham University students were asked to rate how strongly pairs of words were associated on a scale from 1 (unrelated) to 9 (very strong association). Whilst no specific examples were given, when generating each rating participants were instructed to consider all kinds of possible relations, such as causal, functional, taxonomic etc, and were asked to give the first answer that came to mind. We selected only those 20 items with a similar level of association between the base and its alternative causal and taxonomic target categories. We then also derived a more objective measure of co-occurrence against which to verify our notion of association. We calculated the frequency with which the two categories co-occurred within six words on the World Wide Web by using a Google proximity search and used a formula suggested by Heylighen (2001) to calculate the conditional probability of co-occurrence:

$$Aw_{1\&}w_2 = P(w_1 \mid w_2) = \frac{P(w_1 \& w_2)}{P(w_1)} = \frac{N(w_1 \& w_2)}{N(w_1)}$$

In this equation, $P(w_1 \& w_2)$ represents the probability that a text contains both words w_1 and w_2, $P(w_1)$ represents the probability that it contains w_1 on its own. To calculate the conditional probability, one can simply count the number of times w_1 and w_2 co-occur and divide this by the number of times w_1 occurs by chance in the same text sample. We then took the mean of these two conditional probabilities and correlated this with our association strength ratings. These two measures were significantly correlated (Spearman's rho= .56, $p< .01$) supporting our contention that we are indeed measuring a construct of associative strength in which the activation of one leads to activation of the other, irrespective of the nature of relation between the two categories.

To explore the role that level of association plays in the availability of knowledge from different domains, a median split based on level of association was carried out on the selected items. Thus, for 10 items the association between the base and its target categories was classed as strong and for the remaining 10 items this association was classed as weak. For half the strongly and weakly associated items participants generalized diseases. For the other half, people evaluated inductive conclusions about cells, so whilst property was manipulated within-subjects, content was counterbalanced across participants in a Latin-square design.

Participants learnt that the base category had either a blank disease, such as disease 9T4, or blank cells, such as cells Lo8. They then rated the likelihood that the target category shared the disease or cells on a 9-point scale. For example, participants might be presented with the following induction problems:

Carrots have disease 3dfT.
How likely is it that Rabbits have disease 3dfT?
(causal/disease)

Carrots have disease ww3T.
How likely is it that Radishes have disease ww3T?
(taxonomic/disease)

Acorns have cells T4H.
How likely is it that Squirrels have cells T4H?
(causal/cells)

Acorns have cells eR2.
How likely is it that Walnuts have cells eR2?
(taxonomic/cells)

The induction problems were presented on a laptop. The premise and conclusions were presented simultaneously and appeared in a red font. Participants could only enter their response once the font changed to green. In the speeded condition, the font changed from green to red after one second and participants were instructed to read the problem and respond as fast as possible without sacrificing accuracy. In the delayed condition, the font only changed colour after 10 seconds and participants were instructed to carefully consider their responses. They entered their response on the key board by giving a rating between 1 and 9.

Post-Test

The post-test assessed people's beliefs about taxonomic and causal relatedness. For each of the 40 category pairs, participants were asked two questions, resulting in a total of 80 questions. One question asked them whether they believed that the two categories were from the same biological class and the other asked whether the two

categories were part of the same food chain. Participants could respond with YES, NO or DON'T KNOW, but were instructed to use the third option sparingly, as the emphasis was on their intuitions and beliefs rather than on factual correctness. The mean proportion of positive responses to the two post-test questions about biological group membership and food chain relations across the two timing conditions did not correlate with our web-based measure of co-occurrence (Spearman rho correlation coefficients ranged from -.18 to .16, all *p*'s > .27), nor did it correlate with our subjective measure of associative strength (Spearman rho correlation coefficients ranged from .1 to .2, all *p*'s > .18) suggesting that these measures did not reflect associative strength but represents beliefs based on more structured knowledge.

Results

To facilitate an initial factorial analysis of the data, mean inductive strength scores were calculated for the 5 problems representing the unique property by association by relation combination, resulting in 8 means for each participant. These were subjected to a 2 (property: disease or cell) by 2 (relation: causal or taxonomic) by association (high versus low) by 2 (timing: delayed or speeded) mixed-design ANOVA, with timing as the between-subject variable. We predicted effects of degree of association in our results. However, if association does not play an important role, we would expect to observe an interaction between timing and relation, with timing affecting causal but not taxonomic inferences, thus replicating Shafto et al's (2007) findings.

Although the effects of relation, $F_{(1, 38)} = 3.39$, $p = .073$, effect size d = .66, and timing, $F_{(1, 38)} = 3.18$, $p = .082$, effect size d = .6, were approaching significance, timing did not interact with any of the other variables. Thus, when we control for degree of association we do not replicate Shafto et al's finding.

The only large and reliable significant main effect was strength of association, $F_{(1, 38)} = 28.82$, $p < .0001$, effect size d = 2.0. As expected, inferences about closely associated categories ($M = 4.52$, $SE = .14$) were rated stronger than inferences about weakly associated categories ($M = 3.98$, $SE = .14$).

The only significant two-way interaction was between property and relation, $F_{(1, 38)} = 25.68$, $p < .0001$, effect size d = 1.7, suggesting that people showed some context-sensitive reasoning. Bonferroni posthoc tests showed that when reasoning about cells, people rated taxonomic inferences ($M = 5.01$, $SE = .2$) significantly stronger than causal inferences ($M = 3.79$, $SE = .22$, $p < .0001$, effect size d = .9). When reasoning about diseases, people rated causal inferences slightly higher ($M = 4.32$, $SE = .26$) than taxonomic inferences ($M = 3.89$, $SE = .17$) although this difference was not significant ($p = .16$, effect size d = .3). This might suggest that whereas physiological inferences are predominantly supported by taxonomic relations between categories, inferences about diseases can be made on the basis of external mechanisms, in this case causal transmission, but also on the basis of more internal mechanisms, in this case taxonomic links and thus genetic relatedness.

None of the other higher-order interactions were significant (all *p*'s > .08)

Regression Analyses

To explore how structured and unstructured types of knowledge influence category-based inductions under different conditions, we calculated mean inductive strength ratings for each item separately for the two types of property and timing conditions, resulting in 4 inductive strength scores for each item. Similarly, for each item we calculated the mean proportion of positive responses to the two post-test questions about biological group membership and food chain relations across the two timing conditions.

Multiple regression analyses were carried out on the mean inductive strength scores. We make the theoretical assumption that people will be influenced by strength of association regardless of timing manipulations. Hence, we entered this variable in block 1. In a second block, we added proportion of positive responses to the biological group question and food chain question as the independent predictor variables. This enabled us to evaluate the degree to which adding variables reflecting structured knowledge accounted for additional variance above and beyond strength of association.

All four regression analyses were significant, but different relevant knowledge influenced inductive strength under different conditions. Overall, larger multiple correlation coefficients were observed in the delayed condition, suggesting that people used different types of knowledge to inform their inferences when they had time to do so, whereas under time pressure, the ability to recruit relevant knowledge seemed to be attenuated.

Inferences about Diseases

As Figure 1 shows, speeded inductive reasoning about diseases (R = .59) was significantly predicted by strength of association (beta = .45, t = 3.13, $p = .003$). In the second block, knowledge about relevant causal food chain relations was also a significant predictor (beta = .35, t = 2.04, $p = .05$), whereas taxonomic knowledge was not a significant predictor (beta = .08, t = .44, $p = .67$). Together, adding these two structured knowledge variables accounted for a nonsignificant amount of additional variance (R^2 Change: 9.6%, $F_{(2, 36)} = 2.64$, $p = .09$).

In contrast, reasoning about diseases under delayed conditions (R = .68) was no longer significantly predicted by association (beta = .24, t = 1.8, $p = .08$). However, inductive strength was strongly predicted by relevant knowledge about food chain relations (beta = .61, t = 4.34, $p < .001$), but also by beliefs about biological relatedness (beta = .34, t = 2.33, $p = .03$). Adding the structured knowledge predictors in a second block did account for significantly more variance in inductive strength ratings

than strength of association on its own (R^2 Change: 25.8%, $F_{(2, 36)} = 9.46$, $p < .001$).

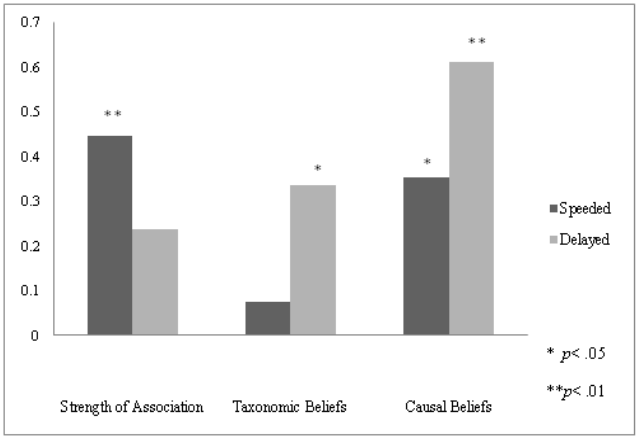


Figure 1: Standardized Regression Coefficients for Predictive Relations between Taxonomic and Causal Beliefs, Strength of Association and Inductive Strength Ratings for Diseases

Inferences about Cells

Reasoning about cells showed a different pattern as shown in Figure 2. Under delayed conditions, strength of association was not a significant predictor of inductive strength (beta = .19, t = .15, p = .14). Inductive inferences were however predicted by beliefs about biological relatedness (R = .72) (beta = .48, t = 3.48, p = .001), and were negatively predicted by beliefs about causal relatedness (beta = -.31, t = -2.29, p = .03). Given that we had selected causal targets that were always from different superordinate categories, it is not surprising that causal beliefs were a negative predictor of inferences about cells.

As when reasoning about diseases, adding the structured knowledge predictors in a second block accounted for significantly more variance in inductive strength ratings than strength of association on its own when people were not under time pressure (R^2 Change: 44.2%, $F_{(2, 36)} = 16.33$, $p < .001$)

Speeded inductions about cells (R = .64) were predicted by strength of association (beta = .51, t = 3.76, p = .001) and were negatively predicted by beliefs about causal relatedness (beta = -.34, t = -2.5, p = .05). Taxonomic beliefs were not a significant predictor of speeded inductive strength ratings (beta = .11, t = .65, p = .52). However, adding the structured knowledge coefficients did explain some additional variance above strength of association on its own (R^2 Change: 16.7%, %, $F_{(2, 36)} = 5.09$, p = .01).

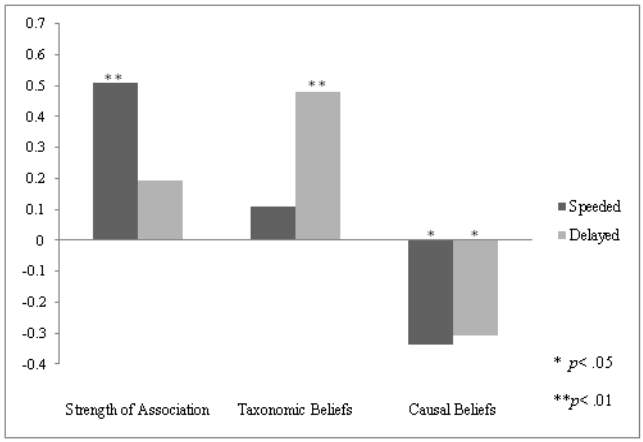


Figure 2: Standardized Regression Coefficients for Predictive Relations between Strength of Association, Taxonomic and Causal Beliefs and Inductive Strength Ratings for Cells

Discussion

Our main proposal was that knowledge effects in category-based induction can be distinguished with regards to two contrasting types of knowledge: effortlessly computable, unstructured knowledge such as strength of association (Rogers & McClelland, 2004) or similarity (Sloman 1993; Sloutsky & Fisher, 2004) on the one hand, and structured knowledge (Kemp & Tenenbaum, 2009, Shafto et al, 2008, Rehder, 2009), which requires more time and processing effort, on the other. Overall, our results strongly support this distinction between different types of knowledge that differ in their processing characteristics. The response timing paradigm used in the current experiment showed that strength of association was a stronger predictor of inductive strength ratings when people had to respond quickly. In contrast, structured causal and taxonomic knowledge became more important when people were forced to delay their response and hence had time to consider the nature of the relationship between the categories.

A secondary goal of this experiment was to explore whether differences in the accessibility of knowledge from different domains arises when level of association is controlled for. The results showed that once level of association was equated across causally and taxonomically related category pairs, the previously observed advantage for taxonomic knowledge (e.g. Shafto et al., 2007) was no longer observed. This suggests that no domain of knowledge is more privileged than any other.

With regards to our main proposal, there are several benefits of being able to draw on two types of knowledge that differ in their processing characteristics. The potency of inductive inferences can be maximized by recruiting structured knowledge, making inferences more sensitive to

contextual factors and relational constraints. It is difficult to see how connectionist models, whose hallmark processes are instantiated by nondirectional and automatic spreading activation, could explain how additional sources of knowledge, such as causal and taxonomic knowledge, selectively influence people's inferences about diseases when people have time but not when they have to respond rapidly.

However, people may not always have the time and available mental resources to try and draw on elaborate background knowledge. Thus, unstructured knowledge acquired through associations, temporal contiguity, or co-occurrence provides a rich source of ecologically valid information at little or no processing cost (Evans, 2008; Smith & DeCoster, 2000). As demonstrated by Rogers & McClelland's (2004) PDP model, it is conceivable that frequently co-occurring categories would lead to a gradual adjustment of their semantic representations in memory, so that activation of one would either 'prime' or partially activate the representation of strongly associated categories.

Conclusion

We provide support for the claim that category-based inductive reasoning is influenced by two types of knowledge, structured and unstructured knowledge, which are mediated by two contrasting mental processes (Rehder, 2009). Use of unstructured knowledge, such as nondirectional associative strength (Sloman, 1993; Rogers & McClelland; 2004) seems to reflect a relatively effortless process, in which inductions are proportional to the degree to which activation of the premise and conclusion category representations in semantic memory overlap. However, this can be supplemented by the use of more elaborate structured knowledge (Kemp & Tenenbaum, 2009; Shafto et al., 2008). Structured knowledge encodes intuitive theories about the structural relationships between categories, such as knowledge about taxonomic connections or causal interactions. Use of this type of knowledge is constrained by cognitive resources but can maximize inductive potency of inferences beyond mere associative strength between categories.

Acknowledgments

This research was funded by an ESRC postgraduate research studentship awarded to A. K. Crisp-Bright.

References

Evans, J. S. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology, 59*, 255-278.

Feeney, A. (2007). How many processes underline category-based induction? Effect of conclusion specificity and cognitive ability. *Memory & Cognition, 35*, 1830-1839.

Fisher, A. V., & Sloutsky, V. M. (2005). When induction meets memory: Evidence for gradual transition from similarity-based to category-based induction. *Child Development, 76*(3), 583-597.

Heylighen, F. (2001). *Mining Associative Meanings from the Web: from Word Disambiguation to the Global Brain*. Standaard Publishers.

Kemp, C., & Tenenbaum, J. B. (2009). Structured Statistical Models of Inductive Reasoning. *Psychological Review, 116*(1), 20-58.

Lopez, A., Gelman, S. A., Gutheil, G., & Smith, E. E. (1992). The Development of Category-Based Induction. *Child Development, 63*(5), 1070-1090.

Medin, D. L., Coley, J. D., Storms, G., & Hayes, B. K. (2003). A relevance theory of induction. *Psychonomic Bulletin & Review, 10*(3), 517-532

Osherson, D. N., Smith, E. E., Wilkie, O., Lopez, A., & Shafir, E. (1990). Category-based induction. *Psychological Review, 97*(2), 185-200.

Rehder, B. (2009). Causal-Based Property Generalization. *Cognitive Science, 33*(3), 301-344

Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: a parallel distributed processing approach.* London: MIT.

Shafto, P., Coley, J. D., & Baldwin, D. (2007). Effects of time pressure on context-sensitive property induction. *Psychonomic Bulletin & Review, 14*(5), 890-894.

Shafto, P., Kemp, C., Bonawitz, E. B., Coley, J. D., & Tenenbaum, J. B. (2008). Inductive reasoning about causally transmitted properties. *Cognition, 109*(2), 175-192.

Sloman, S. A. (1993). Feature-Based Induction. *Cognitive Psychology, 25*(2), 231-280.

Sloman, S. A. (1998). Categorical inference is not a tree: The myth of inheritance hierarchies. *Cognitive Psychology, 35*(1), 1-33.

Sloutsky, V. M., & Fisher, A. V. (2004). Induction and categorization in young children: A similarity-based model. *Journal of Experimental Psychology-General, 133*(2), 166-188.

Smith, E. R., & DeCoster, J. (2000). Dual-process models in social and cognitive psychology: Conceptual integration and links to underlying memory systems. *Personality and Social Psychology Review, 4*(2), 108-131.

Learning hypothesis spaces and dimensions through concept learning

Joseph L. Austerweil (Joseph.Austerweil@gmail.com)
Thomas L. Griffiths (Tom_Griffiths@berkeley.edu)
Department of Psychology, University of California, Berkeley, Berkeley, CA 94720-1650 USA

Abstract

Generalizing a property from a set of objects to a new object is a fundamental problem faced by the human cognitive system, and a long-standing topic of investigation in psychology. Classic analyses suggest that the probability with which people generalize a property from one stimulus to another depends on the distance between those stimuli in psychological space. This raises the question of how people identify an appropriate metric for determining the distance between novel stimuli. In particular, how do people determine if two dimensions should be treated as separable, with distance measured along each dimension independently (as in an L_1 metric), or integral, supporting Euclidean distance (as in an L_2 metric)? We build on an existing Bayesian model of generalization to show that learning a metric can be formalized as a problem of learning a hypothesis space for generalization, and that both ideal and human learners can learn appropriate hypothesis spaces for a novel domain by learning concepts expressed in that domain.
Keywords: generalization; categorization; Bayesian modeling; similarity; integral and separable dimensions

Introduction

Almost every two objects, events, or situations (or the sensory data for the *same* object at two different moments) that we encounter are unique. Despite this fact, when people (and animals) learn that one stimulus has a property, they reliably and systematically believe certain other stimuli have that property and others do not (Shepard, 1987). For example, if you learn a dark, large circle is a *gnarble*, how likely is a dark, slightly smaller circle or a dark very small circle to be a *gnarble*? This is the problem of *generalization*, which is pervasive across cognitive science. It occurs in many forms from higher-level cognition (e.g., concept learning, Tenenbaum, 2000) to linguistics (e.g., word learning, Xu & Tenenbaum, 2007) to perception (e.g., color categorization, Kay & McDaniel, 1978). How should an ideal learner generalize a property from a group of stimuli observed to have the property to other stimuli?

One of the most celebrated theoretical results of cognitive psychology provides a deceptively simple answer to this question, indicating that we should generalize a property from one object to another object when the two objects are similar, or equivalently, close in some psychological space (Shepard, 1987). However, this establishes a new problem: How should the distance between objects be measured? More formally, the problem is one of identifying a *metric* on a space, a basic challenge that also arises when using machine learning methods that rely on computing distances, such as nearest-neighbor classification (Xing, Ng, Jordan, & Russell, 2002; Davis, Kulis, Jain, Sra, & Dhillon, 2007). Cognitive psychologists have determined that people use two different kinds of metrics when forming generalizations about multidimensional stimuli: *separable* dimensions are associated with "city-block" distance or the L_1 metric, while *integral* dimensions are associated with Euclidean distance or the L_2 metric (Garner, 1974). These different metrics also have consequences beyond generalization behavior, influencing how people categorize objects varying along different dimensions (Handel & Imai, 1972) and whether people can selectively attend to each dimension (Garner & Felfoldy, 1970).

Analyses of human generalization have tended to treat the metric as a fixed property of stimuli. However, determining the appropriate metric on a psychological space is an important step towards developing an appropriate representation for the properties of novel objects. If two dimensions are separable, then those dimensions form privileged axes for representing locations in the psychological space, and it is easier to learn categories defined by rules that align with those axes (Kruschke, 1993). This is qualitatively different from an integral representation, in which there are no natural axes for representing the space. Identifying whether dimensions should be separable or integral is thus just as basic a step towards forming a representation for a novel domain as determining the number of dimensions, or the locations of each stimulus in the resulting space.

In this paper, we consider how a learner could identify the appropriate metric for representing a novel domain, comparing an ideal Bayesian learner with human judgments. The starting point for this investigation is an existing Bayesian model of generalization, introduced by Shepard (1987) and extended by Tenenbaum and Griffiths (2001). In this model, the property of interest is possessed by all stimuli within an unknown region of psychological space, and the probability of generalizing to a new stimulus is computed by summing over all candidate regions containing the new stimulus and the previous stimuli observed to have some property, weighted by the posterior probability of that region. The difference between separable and integral dimensions emerges as the result of probabilistic inference with different hypothesis spaces of regions (Shepard, 1987, 1991; Davidenko & Tenenbaum, 2001). The hypothesis spaces that produce generalization corresponding to separable and integral dimensions consist of axis-aligned and axis-indifferent regions in the space, respectively (see Figure 1). Axis-aligned regions produce stronger generalization along the axes, while axis-indifferent regions produce generalization that depends only on the Euclidean distance between stimuli.

This analysis of separable and integral dimensions lays the groundwork for our account of how people learn an appropriate metric for a novel space. Learning a metric thus becomes a matter of inferring an appropriate hypothesis space on which to base generalization. We define a hierarchical

Bayesian model that makes this inference from a set of observed concepts. We demonstrate that this model infers a city-block or Euclidean generalization metric when given axis-aligned or axis-indifferent concepts, respectively, and that people infer a hypothesis space for generalization based on the concepts they learn in a way that is consistent with this ideal observer analysis. This extends previous results by Goldstone (1994) who changed dimensions from being integral to separable via repeated training of a single concept.

The plan of the paper is as follows. The next section provides the theoretical background for our approach, summarizing the basic generalization model, revisiting some of the literature on separable and integral dimensions, and laying out our approach to hypothesis space learning. We then present a test of the predictions of this model with human learners. Finally, we conclude the paper with a discussion of our results and possible future directions.

Theoretical Framework

Our theoretical framework builds directly on the Bayesian generalization model introduced in Shepard (1987) and Tenenbaum and Griffiths (2001), so we begin by summarizing the key ideas behind this approach. We then show how this approach produces separable and integral generalization, and how it can be extended to allow an ideal learner to infer an appropriate representation for novel stimuli.

The Bayesian Generalization Model

Let $\mathcal{X}$ be the stimulus space and $\mathcal{H}$ be the hypothesis space, where $h \in \mathcal{H}$ is a hypothesis as to which objects have and do not have the property of interest (i.e., a hypothesis is a set of $x \in \mathcal{X}$). After observing that a set of stimuli $X = \{x_1, \ldots, x_n\}, x_i \in \mathcal{X}$, stimuli have some property, how should you update your belief in: (1) which property it is and (2) which other stimuli have that property? Assuming that stimuli are generated uniformly and independently under the true hypothesis at random for the property ($p(X|h) = \prod_i p(x_i|h) = |h|^{-n}$ for a hypothesis containing all stimuli in the given set; $p(X|h) = 0$ otherwise) and taking some prior over hypotheses $p(h)$, the posterior probability that a hypothesis h is the property that n given stimuli share is

$$p(h|X) = \frac{p(h)\prod_{i=1}^{n} p(x_i|h)}{\sum_{h' \in \mathcal{H}} p(h')\prod_{i=1}^{n} p(x_i|h')} \quad (1)$$

which is simply Bayes' rule. Using Equation 1, we can derive the probability of generalizing from X to some other stimulus y as the sum over the posterior probability of hypotheses containing y

$$p(y|X) = \sum_{h:y \in h} P(h|X) \quad (2)$$

which constitutes a form of *hypothesis averaging* (Robert, 2007). The predictions of the model depends intimately on the nature of the hypotheses under consideration, with different hypothesis spaces leading to different generalization patterns.

Separable and Integral Dimensions

Psychological explorations of human similarity metrics of multidimensional stimuli discovered two different ways in which people use these dimensions: separable and integral (Shepard, 1987). Separable dimensions can be interpreted independently and form natural axes for representing a space, while integral dimensions are difficult to perceive independently. The dimensional structure of stimuli affects many aspects of human information processing, including the ease of categorizing objects into groups and perceived distance between objects (Garner, 1974). For example, Garner and Felfoldy (1970) found that categorization time was facilitated for objects with integral dimensions (e.g., saturation and lightness of a color) into groups where the values of the dimensions of the objects in each group are correlated (light and desaturated vs. dark and saturated). However, there was interference for objects categorized into groups of objects where the values of the dimensions are orthogonal (light and satured vs. dark and desaturated). Conversely, there were no major differences in categorization time for these types of categorization structures when the dimensions were separable.

Dimensional structure also affects the perceived distances between objects (Shepard, 1991). The perceived distance metric for objects with separable dimensions is the "city-block" distance, also known as the L_1 metric, with the distance between two stimuli x_i and x_j being $d(x_i, x_j) = \sum_k |x_{ik} - x_{jk}|$, where k ranges over dimensions and x_k is the value of stimulus x on dimension k. The perceived distance metric for objects with integral dimensions is the Euclidean distance, or L_2 metric, with $d(x_i, x_j) = \sqrt{\sum_k (x_{ik} - x_{jk})^2}$. The use of these different distance metrics is consistent with the different properties of separable and integral dimensions: city-block distance sums the distance along each axis separately for all points in the space, while Euclidean distance is insensitive to whether a point is located along an axis, and is thus invariant to changes in the axes used to represent the space. Recent extensions of classic multidimensional scaling techniques bear out these results, and provide a way to identify whether people seem to use separable or integral dimensions in their representation of a set of stimuli (Lee, 2008).

In the Bayesian generalization model introduced in the previous section, the difference between integral and separable dimensions emerges from using two different hypothesis spaces (Shepard, 1987). Using a hypothesis space in which regions are aligned with the axes results in behavior consistent with separable dimensions, while a hypothesis space in which regions are indifferent to the axes results in behavior consistent with integral dimensions. Figure 1 shows a schematic of two such hypothesis spaces, restricted to rectangular regions in two dimensions, together with the generalization gradient for a single exemplar concept in each space.[1]

[1] We calculated the generalization gradients by sampling from the prior distribution over hypotheses for the axis-aligned and axis-indifferent hypothesis spaces, then weighting each hypothesis by the likelihood given the single exemplar $E5$. The gradients were evaluated on a discretized 9×9 grid.

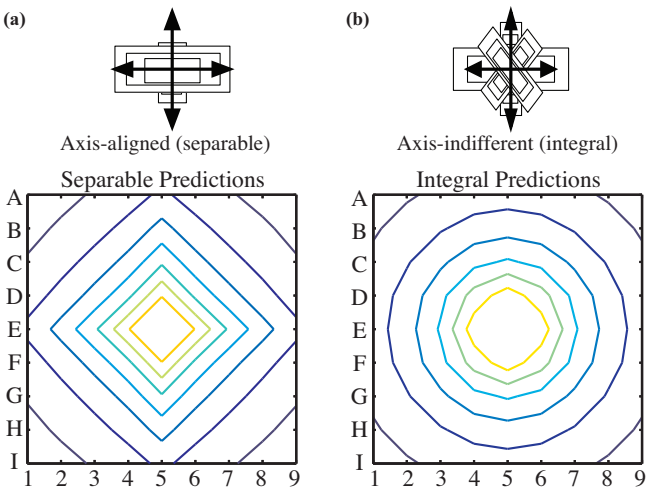


Figure 1: Hypothesis spaces and generalization gradients. (a) Axis-aligned (separable) and axis-indifferent (integral) hypothesis spaces. (b) Resulting generalization gradients for each hypothesis space given a single exemplar of a concept.

The generalization gradient resulting from the axis-aligned hypothesis space given a single exemplar of a concept decreases with distance under a city-block metric, while the gradient resulting from the axis-indifferent hypothesis space decreases with Euclidean distance. Models using the appropriate hypothesis spaces capture generalization judgments well for concept learning tasks using separable and integral dimensions for both single and multiple exemplars (Davidenko & Tenenbaum, 2001; Tenenbaum, 1999).

Learning a Hypothesis Space

The Bayesian generalization framework naturally extends to learning an appropriate hypothesis space by introducing the hypothesis space itself as a higher-level random variable in a hierarchical Bayesian model. Given an enumerable set of hypothesis spaces $\mathcal{M} = \{\mathcal{H}_1, \ldots, \mathcal{H}_M\}$, the probability that an ideal observer generalizes to a new stimulus y given a set of stimuli X have a property and a set of previously observed concepts $\mathcal{C}$ (where each concept itself is a set of stimuli) is

$$P(y|X,\mathcal{C}) = \sum_{m=1}^{M} P(y|\mathcal{H}_m, X) P(\mathcal{H}_m|\mathcal{C}, X) \qquad (3)$$

where the first term is the probability of generalizing from X to y under hypothesis space $\mathcal{H}_m$ (as specified by Equation 2), and the second term is the posterior probability of hypothesis space $\mathcal{H}_m$ given the previous concepts $\mathcal{C}$ and the observed stimuli of the current concept of interest. This posterior probability can be computed by applying Bayes' rule

$$P(\mathcal{H}_m|\mathcal{C}, X) = \frac{P(\mathcal{C}, X|\mathcal{H}_m) P(\mathcal{H}_m)}{\sum_{m=1}^{M} P(\mathcal{C}, X|\mathcal{H}_m) P(\mathcal{H}_m)} \qquad (4)$$

where $P(\mathcal{C}, X|\mathcal{H}_m)$ is the probability of observing a set of concepts $\mathcal{C}$ and the currently observed stimuli under hypothesis space $\mathcal{H}_m$ and $P(\mathcal{H}_m)$ is the prior probability of hypothesis space $\mathcal{H}_m$. The probability of concepts $\mathcal{C}$ and current stimuli X under hypothesis space $\mathcal{H}_m$ is

$$P(\mathcal{C}, X|\mathcal{H}_m) = \prod_{C \in (\mathcal{C} \cup X)} \sum_{h \in \mathcal{H}_m} P(h|\mathcal{H}_m) \prod_{x \in C} P(x|h) \qquad (5)$$

where C plays the same role as X, but for the previously observed concepts.

Intuitively, the model can be thought as being composed of m Bayesian generalization "submodels" (each with their own hypothesis space). The model's generalization judgments are made by averaging over the generalizations made by the individual submodels (given the current stimulus X) weighted by how well the submodel explains the previously and currently observed stimuli. Thus, the model "learns" to use hypothesis spaces that explain the observed concepts well.

Human Learning of Hypothesis Spaces

The model presented in the previous section predicts that a learner should be able to infer whether dimensions are integral or separable for a novel domain after seeing some examples of concepts expressed in that domain. Preliminary support for this idea is provided by the results of Goldstone (1994), who showed that teaching people a novel axis-aligned concept could affect generalization along that axis in both integral and separable spaces. However, shifting a representation all the way towards integral or separable dimensions will require learning more than one concept. To test whether human learners behaved in this way, we conducted an experiment in which we examined how the generalization judgments that people produce depend on the concepts they have learned. We used rectangles varying in width and height as our set of stimuli, and participants learned 20 concepts that were either aligned with or orthogonal to these dimensions (rectangles with the same aspect ratio or area). The key prediction was that participants observing axis-aligned concepts should show a generalization gradient consistent with a city-block metric, whereas participants observing concepts indifferent to these axes should show a generalization gradient consistent with a Euclidean metric. This prediction results from the different hypothesis spaces the two groups of participants should infer are appropriate for these domains.

Stimuli and Methods

The stimuli for this experiment were rectangles where the two manipulated dimensions were the width and height (ranging from 13 to 115 pixels in increments of approximately 25 pixels). The stimulus set is shown in Figure 2. We chose rectangles because it is easy to think of concepts on our two manipulated dimensions (same width or height) and the diagonals of the dimensions (same aspect ratio or area). Previously, Krantz and Tversky (1975) found people weakly favor using area and aspect ratio as separable dimensions (the diagonals of separable dimension space). However, people can use any of the four potential dimensions for generalization depending on the context rectangles are in. This natural flexibility

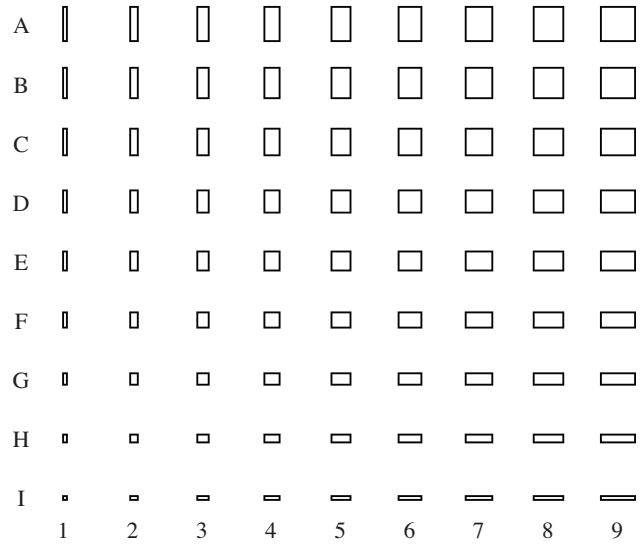


Figure 2: Stimuli used in our experiment (not to scale).

makes rectangles an ideal candidate for training participants to represent rectangles using different dimensional structures.

There were two phases to the experiment: training and test. For the training phase, there were two between-subjects conditions: the *separable* condition ($n = 15$), in which people observed axis-aligned concepts, and the *integral* condition ($n = 18$)[2] in which people observed axis-indifferent concepts. The test phase was the same for all participants. The cover story for the experiment was:

> On a small island in the Pacific Ocean, scientists found the ancient ruins of a small civilization. While excavating the ruins, they discovered objects on the doors of particular houses. They believe that the objects carry information about the people in the houses. Some of the objects the scientists found had names written under them.

Stimuli were then presented as objects with names, and people guessed what other objects would share the same name.

The 20 concepts shown to the training groups are shown in Figure 3 (each concept is a straight line picking out several points, corresponding to stimuli). The concepts for the two conditions were chosen such that each condition saw each object an equal number of times, there were two to four objects in each concept, and the concepts spanned the space of objects. The 20 concepts were presented to participants in a random order as examples of objects that were called different nonsense names randomly chosen from a standardized list. While the objects in each concept were on the screen, participants were asked whether or not they thought every object in $\{A,C,E,G,I\} \times \{1,3,5,7,9\}$ shown individually below the objects in the concept could be called that name.

The test phase of the experiment was identical to the first phase except participants' generalizations were tested for

[2]The different number of participants in each group was due to the computer crashing mid-experiment.

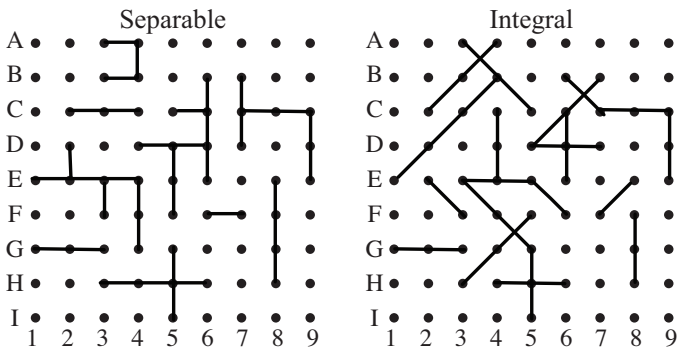


Figure 3: The 20 concepts for each training condition. Each concept is the collection of objects on a straight line on the grid. The separable concepts are axis aligned and the integral concepts are indifferent to axes.

concepts consisting of single objects ($\{B2, B8, E5, H2, H8\}$ were tested) over the total 9×9 set of objects.

Results

Figure 4 shows averaged results for single exemplar generalization for the test phase in the two conditions. The single exemplar concept results were re-aligned to $\{E,5\}$ and then averaged over the five concepts per participant and over participants. We then took the difference between the generalization gradients for the two conditions, and compared them with the difference between the generalization gradients produced by the Bayesian model. The integral group generalizes more on the diagonals and less on the axes than the separable group as predicted if the integral and separable groups used Euclidean and city-block distance metrics respectively.

To test quantitatively that the two groups learn integral and separable dimensions, we found that the integral training group generalized significantly more often on diagonals than axes (averaging over $\{C,D,F,G\} \times \{3,4,6,7\}$ vs. $C5, D5, F5, G5$, $t(32) = 3.23, p < 0.005$). Within the separable group, the generalization judgments on the axes were significantly greater than the diagonals ($t(34) = 2.66, p < 0.05$); however, the integral group did not differentiate between changes on the axes and the diagonals ($t(30) = 0.43, p = 0.43$). Interestingly, both groups of participants treated the positive diagonal ($F3, F4, G3, G4, C6, C7, D6, D7$) differently than the negative diagonal ($C3, C4, D3, D4, F6, F7, G6, G7$) ($t(34) = 2.58, p < 0.05$ for separable and $t(30) = 2.63, p < 0.05$ for integral). This replicates Krantz and Tversky (1975)'s finding that people tend to generalize rectangles based on constant aspect ratio. This is not surprising as constant aspect ratio is an important invariance of an object's projection on the retina as it changes in depth (keeping the viewpoint orientation constant) due to perspective projection (Palmer, 1999).

Finally, we calculated a mixed effect 2×2 ANOVA that corrobrates the conclusions of our other statistical tests. It identified a main effect of generalizing on the diagonal vs. the

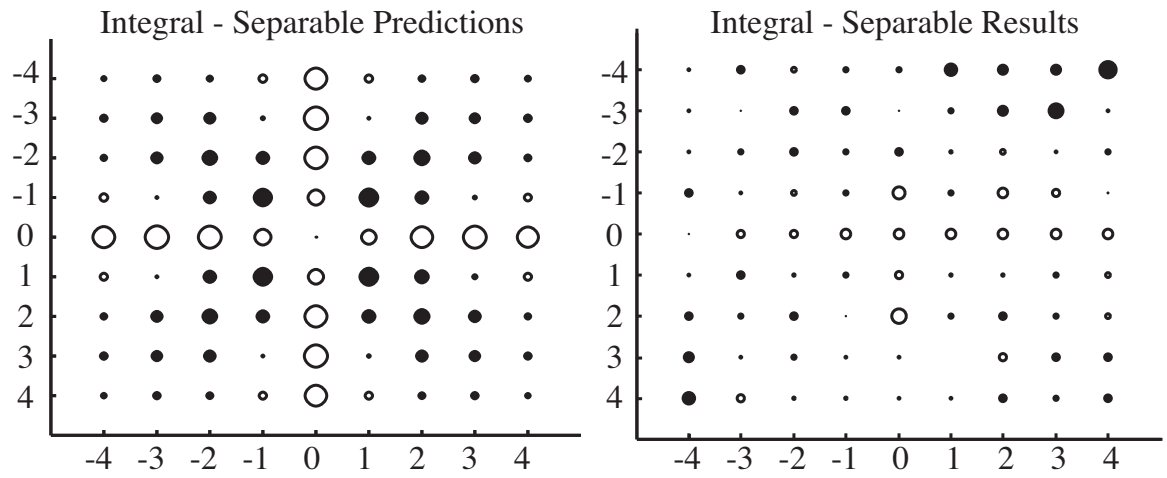


Figure 4: Predictions of the difference between the two Bayesian models formed by model averaging given the separable and integral concepts, and difference between the human generalization results from the two conditions. The results are presented as bubble plots where the size of the bubble represents the degree of generalization. Solid and open bubbles represent positive and negative values respectively. Each single exemplar concept results were re-aligned to $E5$ and then averaged over the five concepts per participant and over participants. Notice how the differences on the axes aligned with the given stimulus ($E5$) are negative and the differences on the diagonals are positive.

axes ($F(1,32) = 44.258, p < 0.001$) and an interaction between generalizing on the diagonal vs. the axes and the training group ($F(1,32) = 10.453, p < 0.005$). This suggests that in the future we should include a hypothesis space into our hierarchy that includes regions varying on the axes and the positive diagonal (but not the negative diagonal).

Discussion

Generalization is an essential problem that basically every cognitive system needs to solve in virtually every domain. Previous analyses of the generalization problem (Shepard, 1987; Tenenbaum & Griffiths, 2001) indicated how an ideal learner should act assuming that an appropriate representation of the stimuli and hypothesis space for generalizations is known. However, how people arrive at a representation and hypothesis space has been left as an open question. As it seems unlikely that people would be born with the appropriate representation and hypothesis space for all possible domains, people need to be able to infer this information from their observations of the properties of stimuli. Using the problem of learning a metric as an example, our analysis shows how an ideal learner would go about inferring such hypothesis spaces, and our experimental results suggest that people do so in a way that is consistent with this model.

To our knowledge, our results provide the first behavioral evidence that people can learn whether stimuli should be represented with separable or integral dimensions. Our results also provide compelling support for the idea that the difference between separable and integral dimensions can be thought of as the result of different hypothesis spaces for generalization, building on (Shepard, 1987, 1991; Davidenko & Tenenbaum, 2001). In future work, it would be interesting to further test this account of separable and integral dimensions by exploring if after training participants show other consequences of having separable or integral dimensions, such as classification and attentional effects. Additionally, this would address a potential confound that the training affects the attention participants pay to each dimension. Fortunately, our larger conclusion that people use the concepts they are given to learn the appropriate hypothesis space for a domain holds regardless of the potential confound (as this conclusion is agnostic to the exact mechanism affecting generalization).

One attractive aspect of this analysis (over using a different solution, like model selection) is that it provides a way to explain why the empirical literature suggests that integrality has been found to be a fuzzy rather than a binary distinction (Garner, 1974). Such fuzzy boundaries emerge as a consequence of Bayesian inference when there is uncertainty to which hypothesis space is appropriate for generalization. We would predict that the "integrality" of natural dimensions are a consequence of how real world objects are categorized along those dimensions. For example, the reason why the saturation and brightness of a color are integral is because in our environment we do not make distinctions between colors at different saturations and brightnesses. "Light" green is a typical color word; however, "saturated" green is an esoteric word, reserved only for artists, designers, and perceptual psychologists. In fact, Goldstone (1994) and Burns and Shepp (1988) found that these dimensions are separable in people who regularly distinguish between the two (color experts and participants trained to distinguish between the two), which implies that they have concepts aligned with the axes of brightness and saturation.

Another important implication of our results is that humans learn the metric appropriate for generalization in a particular domain from the concepts they observe. It would be interesting to compare how metric learning algorithms developed in machine learning (e.g., Xing et al., 2002; Davis et al., 2007) compare to human metric learning on this task, and after learning other types of concepts. This could pave the way towards new machine learning algorithms that automatically infer dimensions intuitive to people from a given set of concepts. Dimensionality reduction techniques like multi-dimensional scaling and principal component analysis are some of the most widely used tools for scientific data analysis, but only produce the equivalent of integral dimensions. An algorithm that determines whether a space is better represented by separable or integral dimensions, and produces interpretable separable dimensions, would be a valuable addition to any data analysis toolkit.

Though Bayesian models have become very popular and successful at explaining different cognitive phenomena (Chater, Tenenbaum, & Yuille, 2006), the hypothesis spaces used in the models are handpicked by the modeler and usually specific to the particular investigated phenomenon. This leaves open the question of how people choose the hypotheses for a set of observed stimuli. Our framework presents an answer to this problem – a hypothesis space is used for a set of observed stimuli depending on how well it explains the observed stimuli and its prior probability. We provide behavioral evidence for our framework in the case study of learning whether or not two dimensions should be separable or integral. Futhermore, this introduces an interesting equivalence between learning the structure of dimensions used to represent stimuli and the set of candidate hypotheses for generalization, which we plan to investigate in future research.

Acknowledgments. We thank Rob Goldstone, Stephen Palmer, Karen Schloss, Tania Lombrozo, and the Berkeley Computational Cognitive Science Lab for insightful discussions, and Shubin Li, Brian Tang, and David Belford for help with experiment construction, running participants, and data analysis and four anonymous reviewers and Nick Chater for their comments on a previous draft of the paper. This work was supported by grant FA9550-07-1-0351 from the Air Force Office of Scientific Research.

References

Burns, B., & Shepp, B. E. (1988). Dimensional interactions and the structure of psychological space: The representation of hue, saturation, and brightness. *Perception and Psychophysics*, *43*, 494-507.

Chater, N., Tenenbaum, J. B., & Yuille, A. (2006). Special issue on "Probabilistic models of cognition". *Trends in Cognitive Sciences*, *10*(7), 287-344.

Davidenko, N., & Tenenbaum, J. B. (2001). Concept generalization in separable and integral stimulus spaces. In *Proceedings of the 23rd Annual Conference of the Cognitive Science Society.* Mahwah, NJ.

Davis, J. V., Kulis, B., Jain, P., Sra, S., & Dhillon, I. S. (2007). Information-theoretic metric learning. In *Proceedings of the 24th International Conference on Machine Learning.* Corvallis, OR.

Garner, W. R. (1974). *The Processing of Information and Structure*. Maryland: Erlbaum.

Garner, W. R., & Felfoldy, G. L. (1970). Integrality of stimulus dimensions in various types of information processing. *Cognitive Psychology*, *1*, 225-241.

Goldstone, R. (1994). Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology: General*, *2*(123), 178-200.

Handel, S., & Imai, S. (1972). The free classification of analyzable and unanalyzable stimuli. *Perception & Psychophysics*, *12*, 108-116.

Kay, P., & McDaniel, C. K. (1978). The linguistic significance of the meanings of basic color terms. *Language*, *54*, 610-646.

Krantz, D. H., & Tversky, A. (1975). Similarity of rectangles: An analysis of subjective dimensions. *Journal of Mathematical Psychology*, *12*, 4-34.

Kruschke, J. K. (1993). Human category learning: Implications for backpropagation models. *Connection Science*, *5*, 3-36.

Lee, M. D. (2008). Three case studies in the Bayesian analysis of cognitive models. *Psychonomic Bulletin and Review*, *15*(1), 1-15.

Palmer, S. E. (1999). *Vision Science*. Cambridge, MA: MIT Press.

Robert, C. P. (2007). *The Bayesian choice: A Decision-theoretic Motivation*. New York: Springer.

Shepard, R. N. (1987). Towards a universal law of generalization for psychological science. *Science*, *237*, 1317-1323.

Shepard, R. N. (1991). Integrality versus separability of stimulus dimensions: from an early convergence of evidence to a proposed theoretical basis. In *The Perception of Structure: Essays in Honor of Wendell R. Garner* (p. 53-71). Washington, DC: American Psychological Association.

Tenenbaum, J. B. (1999). Bayesian modeling of human concept learning. In M. S. Kearns, S. A. Solla, & D. A. Cohn (Eds.), *Advances in Neural Information Processing Systems 11* (p. 59-65). Cambridge, MA: MIT Press.

Tenenbaum, J. B. (2000). Rules and similarity in concept learning. In S. A. Solla, T. K. Leen, & K.-R. Muller (Eds.), *Advances in Neural Information Processing Systems 12.* Cambridge, MA: MIT Press.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, *24*, 629-641.

Xing, E. P., Ng, A. Y., Jordan, M. I., & Russell, S. (2002). Distance metric learning, with application to clustering with side-information. In *Advances in Neural Information Processing Systems* (Vol. 12). Cambridge, MA: MIT Press.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, *114*, 245-272.

The Role of "Explaining Away" in Human Abstract Rule Induction

Colin Reimer Dawson, M.A. (CDawson@Email.Arizona.Edu)
Department of Psychology, 1503 E. University Blvd.
Tucson, AZ 85721 USA

LouAnn Gerken, Ph.D. (Gerken@Email.Arizona.Edu)
Department of Psychology, 1503 E. University Blvd.
Tucson, AZ 85721 USA

Abstract

Of great interest to cognitive science is how human learning is constrained to avoid spurious generalizations. While many constraints must be relatively experience-independent, past experience provides a rich source of guidance for subsequent learning. If a learner discovers some structure in part of the environment, this can inform her future hypotheses about that domain. If a general structure parsimoniously accounts for particular sub-patterns, a rational learner should not stipulate separate explanations for each detail without additional evidence, as the general structure has "explained away" the original evidence. In a grammar-learning experiment using tone sequences, manipulating learners' prior exposure to a tone environment affects their sensitivity to the grammar-defining feature, in this case consecutive repeated tones. Grammar-learning performance is worse if context melodies are "smooth", that is, if small intervals occur more often than large ones, as this smoothness is a general property that accounts for a high rate of repetition.

Keywords: statistical learning; artificial grammar; Bayesian inference; language acquisition; music cognition

Introduction

In traditional theories of learning, the relationship between knowledge[1] and learning is fairly static. Some initial knowledge is provided by experience-invariant biology to constrain learning. Within these *a priori* constraints, learning builds one's body of knowledge. An area that has been explored relatively little is the dynamic interplay between learning and knowledge: namely, (how) can the results of learning actually change how subsequent learning proceeds? If this feedback loop is ignored, observed constraints on learning may be incorrectly attributed to initial biologically provided knowledge instead of to learning that has already taken place.

Untangling the relative contributions of experience-independent biology and prior learning has been particularly important in the study of infant cognition, not least infant language acquisition. If an adult can learn one pattern and not another in the absence of *a priori* differences in difficulty, there are often ready explanations in terms of her years of experience in the world. In contrast, if a young infant exhibits the same discrepancy, it is tempting to attribute it to biology. This conclusion would be premature without further examination, however.

Indeed, previous research suggests that infants reorganize their domain knowledge within the first year of life, and even within the laboratory. In language, infants reorganize their phonetic categories (Werker & Tees, 1984; Bosch and Sebastian-Galles, 2003; Maye, Werker, & Gerken, 2002) and even exhibit shifts in what features a stress rule can reference (Gerken and Bollt, 2008). In music, attention shifts from absolute pitches to relative intervals (Saffran and Griepentrog, 2001; Saffran, 2003), and infants' tonal and rhythmic categories change as a function of cultural context (Hannon and Trehub, 2005; Lynch and Eilers, 1992).

Marcus, et al. (1999) and Marcus, Fernandes and Johnson (2007) found that 7-month-old infants can learn an AAB or ABB pattern in three-element sequences, provided the elements are syllables. Infants at the same age failed at the same task when the elements were non-linguistic events such as musical tones or animal noises. It was suggested that the child's initial endowment may tell her that speech can be structured by abstract, relational properties, but other auditory stimuli cannot.

While this is possible, subsequent research has revealed AAB-style learning in infants with other stimuli, such as pictures of dogs (Saffran, et al., 2007), and simple shapes (Johnson, et al., in press). Murphy, Mondragon and Murphy (2008) found that even rats can learn such generalizations from both speech and tones. These results cast doubt on the notion that language is privileged for abstract pattern-learning.

"Explaining Away" Details With Generalities

Dawson and Gerken (2009) found that while 7-month-olds fail at learning AAB and ABA patterns with tones, 4-month-olds succeed given the same input. They suggested that 7-month-olds' failure may be due to their having learned certain general properties about music. In particular, if they have learned that (a) melodies tend to move in small intervals from pitch to pitch, and (b) individual melodies tend to use only a restricted set of pitches (Temperley, 2008), the presence of a large number of repetitions would become much less surprising, and hence less informative about the abstract structure in the AAB-style task. This change in informativeness is an example of a phenomenon known as "explaining away", central to several cognitive models in a variety of areas

[1] Here, "knowledge" is meant in a broad sense – roughly, "information about the environment". This can be anything that affects behavior, or, critically, the interpretation of experiences.

including visual inference (Kersten, Mamassian and Yuille, 2004), linguistic processing (Ciaramita and Johnson. 2000), and infant causal reasoning (Xu and Garcia, 2008; Gergely and Csibra, 2003).

The basic idea is as follows. When an observed pattern could arise from multiple hidden causes, the causes "compete" with each other over the evidence contained in the data, even when the underlying hypotheses do not conflict with each other a priori. For example, suppose during a card game you peek at the dealer's hand, and you notice that on one hand, she has three aces, and on the next she has the nine through king of hearts. If you assume the game is poker, this unusually lucky sequence might raise suspicion that the dealer has stacked the deck to give herself a favorable hand. However, if you later learn that the players are engaged in a friendly pinochle match, in which only the cards nine through Ace are used, the dealer's hands are less surprising given a fair deal. Although the dealer may still be stacking the deck, the evidence for this hypothesis must be discounted, or "explained away".

In a musical context, repetition is an ambiguous event. On the one hand, it constitutes a "sameness" relation between two tones. At the same time, it is also an interval of magnitude zero between successive pitches. If one assumes that melodies are random, and that any tone is equally likely at any point (i.e., the tone distribution is uniform), hearing every melody begin with two repeated notes would be quite surprising, and evidence for a "sameness" interpretation would be strong. If, however, one knows that tones nearby in time also tend to be nearby in pitch (i.e., melodies are usually "smooth"), repetition becomes a more common event (*qua* interval of distance zero), and it should take more evidence to conclude that repetition is special. Similarly, as the set of tones shrinks, the probability of chance repetitions increases (as with the three aces in the Pinochle hand), and the evidentiary bar for learning a repetition grammar should be raised.

The present experiment provides a test of the first of these two predictions with human adults. Participants are first placed in one of three melodic environments: one where every tone is equally likely at any point (the Uniform condition); one in which small intervals are more common than large intervals (the Smooth condition); and one in which repetition alone is more frequent than other intervals (the Repetition condition). Following this exposure, participants are given a grammar-induction task where the "grammatical" melodies have either an AABCD or DCBAA structure. If learners model the interval distribution in the larger environment, the Smooth context should lead them to represent repetition as the result of a general constraint on melodies, and not as a specific grammatical feature. Hence, learners should exhibit decreased sensitivity to positional repetition, as well as decreased grammar-learning performance.

In contrast, in the Repetition environment, the only way to explain the high rate of repeated tones is to represent it explicitly. This unexplained repetition may even *increase* learners' attention to that feature, improving their performance relative to the Uniform group.

Methods

Participants

One hundred and twenty University of Arizona undergraduates participated in the study for course credit. An additional eighteen participated but were excluded from analysis due to their failure to score above chance on a melodic-discrimination screening task.

Materials and Procedures

The experiment consists of a "context" phase and a grammar-learning phase. The latter contains four blocks, each with a training component and a test component. All "sentences" consist of five tones generated using the FM Synthesizer in the MIDI Toolbox for MATLAB (Eerola & Toiviainen, 2004), which produces a horn-like sound. The first four notes are 250 msec each, with 50 msec gaps after each one. The last note is 500 msec. In music terms, the melodies contain four eighth notes followed by a quarter note, played at 200 beats per minute.

Procedures: Context Phase The context phase consists of two blocks of 100 sentences, in random order. Ten are "probe" sentences, after which either the same sentence is repeated or one of the other ten probe sentences is played. On the probe trials, participants have 3 seconds to press the "1" or "0" key on the keyboard to register "same" or "different" sentence pairs. The absence of a response is coded as incorrect. Each block lasts about five minutes. Data from participants who did not perform above chance on this discrimination task (15 or more out of 20 correct) was discarded, as these participants presumably either could not distinguish differences among melodies, or were not attempting to succeed.

During context exposure, all participants see a group of eight cartoon "aliens" (Folstein, Van Petten and Rose, 2007). Half are "star-chested" and half are "brick-chested".

Materials: Context Phase Participants are assigned to one of three context conditions: Uniform (n = 24), Smooth (n = 48) or Repetition (n = 48). The Smooth and Repetition conditions are further divided into High Variance (HV) and Low Variance (LV) sub-conditions. In all cases, context melodies are drawn from a "vocabulary" of six tones: A3, A#3, C#4, E4, G4 and G#4 (MIDI values 57, 58, 61, 64, 67 and 68).

In the Uniform condition, each tone is equally likely and independent of the last. As such, the probability of a repetition at any given point is 1/6 (in the 200 generated melodies, the empirical rate was 18.1%). The resulting distribution of intervals is shown in Fig. 1a.

In the Smooth condition, melodies are generated as follows. The first tone is chosen from a uniform distribution over the six tones. For each subsequent tone, a

sample is generated from a normal distribution, truncated between 0.5 and 6.5. The mean of the distribution is an integer corresponding to the previous tone (the lowest tone is 1; the highest tone 6). The standard deviation is 2 in the HV condition and 1.2 in the LV condition. The sampled value is rounded to the nearest integer to generate the tone. The resulting distribution reflects the bias toward small intervals in typical folk music (Dawson, 2007). The rate of repetition across the 200 melodies is 39.3% of all intervals in the LV condition (Fig. 1b), and 26.3% in the HV condition (Fig. 1c).

The Repetition conditions control for the actual rate of repetition, while removing the overall "smoothness" constraint. Here, the HV and LV conditions (Fig. 1d-e) are matched to their Smooth counterparts for the number of repetitions, but unlike in the Smooth cases, the remaining notes are equiprobable. Here, the high rate of repetition cannot be explained by a general bias for small intervals; instead, a learner modeling the tone distribution must encode repetitions separately to achieve a good fit.

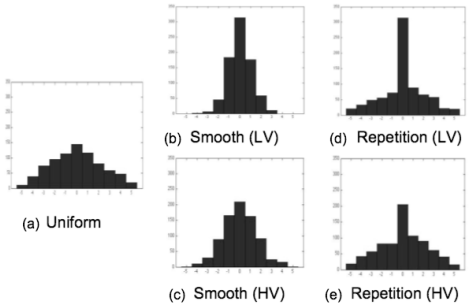


Figure 1(a-e): Interval Counts in the Context Phase[2]

Procedures: Grammar-Learning Phase After the context phase, participants move on to the grammar-learning phase. They are asked to detect "spies" attempting to infiltrate the "Qixian" colony, and are told that they can distinguish Qixians from spies by the grammaticality or ungrammaticality of their speech.

In each training block, participants hear thirty "grammatical" sentences in random order while an image of four star-chested aliens is displayed.

After each training block, participants hear twenty-four test sentences, half grammatical. After each sentence, participants make a continuous grammaticality judgment by clicking on a line (Fig. 2), where the left pole represents "definitely grammatical", the right pole represents "definitely ungrammatical", and every gradient response in between is possible. There is no time limit. The computer records a binary response, based on whether the participant clicks left or right of center, and a continuous "discrimination score" calculated by subtracting from 100 the percentage of the line lying between the response and the correct pole. Participants experience four training-test cycles on the same grammar.

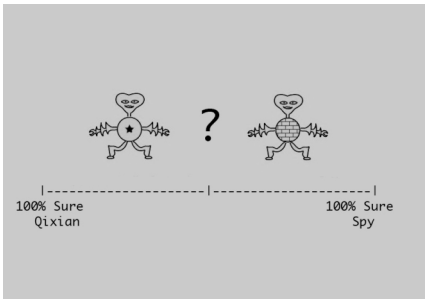


Figure 2: Test Prompt

Materials: Grammar-Learning Phase The "Qixian" and "spy" sentences are again five tones in length. Each participant is trained using one of two five-tone vocabularies. The first (V1) contains the tones A3, C4, D#4, F#4 and G4 (MIDI 57, 60, 63, 66 and 67); the second (V2) contains the tones A#3, B3, D4, F4 and G#4 (MIDI 58, 59, 62, 65 and 68). Each set shares two tones with the context vocabulary.

For half of participants, the "grammatical" sentences follow an AABCD pattern (with a repetition at the beginning and nowhere else), while the "ungrammatical" sentences have a DCBAA pattern. For the other half of participants, the labels are reversed.

Of the 120 sentences possible in each grammar, 60 are used as training items, and 24 as test items. The chosen items were balanced for pitch contour: ¼ in each section had a rising segment followed by a falling segment (in addition to the repetition), ¼ had the reverse; ¼ had a rise-fall-rise pattern and ¼ a fall-rise-fall pattern.

Thirty training sentences are used in the first two learning blocks; the other thirty in the last two blocks. On odd-numbered test blocks, participants are tested with items from the training vocabulary; on even blocks they hear items from the opposite vocabulary. Both vocabularies were used to test whether the context manipulation has an effect on the level of abstraction at which participants learn the grammar. The training vocabulary always comes first, as the vocabulary switch could provide a clue to the nature of the grammar (i.e., that it was vocabulary-independent), and if the new vocabulary came first, participants could not demonstrate mastery independent of this "hint".

Results

Of primary interest is whether prior exposure to the Smooth distribution will impair participants' detection of the repetition pattern. If so, this will suggest that learners are establishing a higher baseline for repetition, which (partially) explains away the training pattern. The key comparison is between the Smooth and Repetition conditions, as these are matched for number of repetitions,

[2] Here, 0 is a repetition, +1 is a step to the next-highest note, etc. The distributions are combined across all pitches. The truncation of the pitch range results in more small intervals across all conditions. If the interval distributions were separated by preceding pitch, those for the Uniform and Repetition conditions would each be flat, except for the peak at 0 in the Repetition case.

differing only in the presence or absence of a larger-scale regularity that accounts for that frequency.

A secondary question is whether the presence of an *inexplicably* high rate of repetitions will encourage learners to encode discrete "same" and "different" relations at the expense of the continuous relations among frequencies, thereby increasing the proportion of attention allocated to repetition and hence increasing performance in grammar-learning. If so, the Repetition group should outperform the Uniform group.

Pilot data revealed that many participants performed near ceiling at discriminating grammatical and ungrammatical sentences, while another large set performed at chance overall. For many of these, presumably only a fairly strong manipulation would observably shift performance. As such, the particular values of the scores received by these participants are mostly uninformative, and contribute noise that could obscure effects of the manipulations.

To address this issue, participants were separated into quartiles within each context condition based on their combined number of correct responses throughout the four test blocks, and two sets of analyses were conducted. The first used all of the data; the second discarded the highest- and lowest-performing quartiles in each condition, thereby greatly reducing the proportion of participants performing either at floor or ceiling. When "floor" is defined as producing fewer than 57 correct binary responses out of 96 (the one-tailed $p < 0.05$ cutoff under coin-flip guessing), and "ceiling" is defined as 88 or more correct (i.e., the same distance from 100% as floor is from 50%), then of the 60 participants in the trimmed sample, only 9 were still at floor, and 7 at ceiling. Of the 30 participants excluded for low performance, all but 2 were at floor, and of the 30 excluded for high performance, all but 4 were at ceiling.

Full Sample Analysis

Both the binary and continuous responses were analyzed, yielding qualitatively similar results. In the interest of concision, we report only the latter here. Mean scores were computed for each participant at each block and entered into an ANOVA with between-subjects factor Context Condition (five levels: Uniform, Smooth (High Variance), Smooth (Low Variance), Repetition (High Variance) and Repetition (Low Variance)), and within-subjects factor Block (1 through 4). Four planned contrasts were used for the Context factor: the first three concerned the Repetition and Smooth groups, corresponding to main effects of (1) distribution type and (2) variance, and (3) the interaction between distribution and variance; the last comparison contrasted the two Repetition groups with the Uniform group.

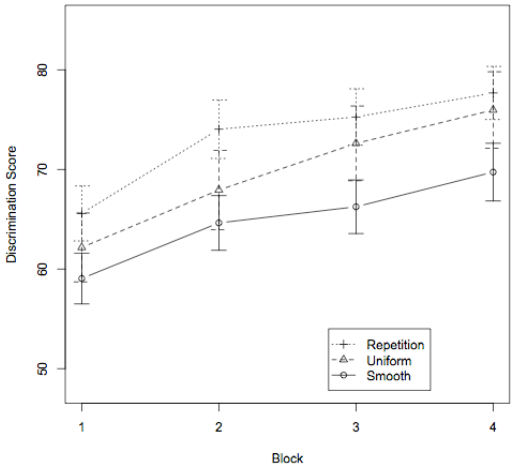


Figure 3: Mean Discrimination Scores by Context Condition and Block, Full Sample

The main effect of Block was significant ($F(3, 345) = 30.59$, $p < 10^{-15}$) but the Block X Context interaction was nonsignificant ($F(12, 345) = 0.55$, n.s.). Of the contrasts among context conditions, only the contrast between the Repetition and Smooth groups reached significance ($F(1, 115) = 5.63$, $p < 0.02$). The contrast between the Repetition group and the Uniform group was nonsignificant ($F(1, 115) = 0.03$, *n.s.*), as were the contrast between the High and Low Variance groups ($F(1, 115) = 1.08$, *n.s.*) and the Distribution X Variance interaction ($F(1, 115) = 0.12$, *n.s.*). Means and standard errors for each block and each group (collapsing the High and Low Variance groups) are displayed in Fig. 3.

Trimmed-Sample Analysis

The above analysis was repeated using only those participants in the second and third quartiles within each context group, as determined by total number correct collapsed across blocks. The effect of Block was significant ($F(3, 165) = 26.89$, $p < 10^{-13}$), but the Block X Context interaction was not ($F(12, 165) = 1.12$, *n.s.*). Of the contrasts among context conditions, only the contrast between the Repetition group and the Smooth group reached significance ($F(1, 55) = 16.16$, $p < 0.001$). The contrast between the Repetition group and the Uniform group was nonsignificant ($F(1, 55) = 0.41$, *n.s.*), as were the contrast between the High and Low Variance groups ($F(1, 55) = 1.58$, *n.s.*) and the Distribution X Variance interaction ($F(1, 55) = 0.004$, *n.s.*). Means and standard errors for this trimmed sample are displayed in Fig. 4.

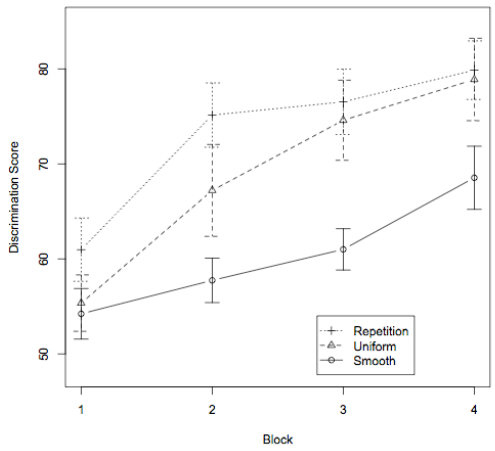


Figure 4: Mean Discrimination Scores, 2nd and 3rd Quartiles Only

Discussion

The present experiment set out to investigate the contribution of rational, generative "explaining away" to induction of an abstract repetition rule over a set of tone sequences. In the context of sequences of musical tones, repetition has a dual nature, first as an identity relation between two consecutive events, and second as an interval of magnitude zero between two tones on a continuum. Hence when a repetition occurs, it is ambiguous which of these two descriptions should be attached to it. The central finding was that adult humans appear to take into account a global "smoothness constraint" on melodies, which have a statistical tendency to move in small intervals, to set a baseline expectation for the rate of repetitions. This reduces the informational value of a repeated tone as a cue to an abstract rule.

The secondary prediction was that participants in the Repetition condition (in which repetitions are uniquely frequent and cannot be explained except by representing them explicitly) would be inclined to downplay the ordinal relations among tones in their representation of the environment, focusing instead on abstract, "discrete" relations like "same" and "different". Since these are precisely the relations needed to learn the repetition grammar, participants in this group were expected to learn the rule more easily than those in the Uniform group. Support for this prediction is tenuous at best: although scores in the Repetition group were numerically higher than those in the Uniform group, this difference failed to reach significance. Exploratory analyses suggested that an advantage for the Repetition group may be present in the early blocks, disappearing later, but the difference was only marginally significant, and in any case the analysis was *post hoc*. It may be that a larger sample is needed to detect a difference if one indeed exists.

At first glance, the presence of differences in participants' use of information resembles situations in which learners come to focus on features that are predictive of a relevant task outcome, filtering out redundant information (Haider and French, 1996; Pellegrino, Doane, Fischer and Alderton, 1991; Doane, Sohn and Schrieber, 1999). Although many of the same mechanisms may be involved here, the nature of the learning is somewhat different. Whereas in the preceding experiments participants were engaged in a specific task all along, in the present experiment the key manipulation occurs before participants become aware of what it is they will be asked to learn. As such, it is not simply a matter of repetition being predictive of a particular response (or even of other stimulus features); rather this result suggests that learners in this experiment are creating an *explanatory* model of the alien environment, and forming hypotheses about how their input is being generated.

Although ultimately the value of explanation may be connected to the future ability to make predictions, the absence of explicit behavioral demands frees learners to pursue a general goal of understanding the underlying nature of the environment. Here, in the Smooth environment repetitions do not appear to be an essential component of the environment at all, whereas in the Repetition environment it is necessary to represent them in order to understand the distribution of intervals. This concept of the learning process as rational hypothesis testing fits nicely into the wealth of recent literature using Bayesian models to capture aspects of cognitive functioning (see, e.g., Tenenbaum, Griffiths and Kemp (2006), for a review).

The present set of findings is of great relevance to the rule-learning literature initiated by Marcus, et al. (1999), and is particularly supportive of the conjecture by Dawson and Gerken (2009) that 7.5-month-olds may have "learned to fail" at learning AAB rules due to the acquisition of knowledge about tonality and the smoothness of natural melodies. We are currently carrying out a version of the present experiment adapted to infants to determine whether the explaining away process observed here in adults comes into play in infancy as well. If so, it will add a new explanatory tool to be applied to the puzzle of why formally analogous rules are easier to learn in some contexts than others. More generally, the sort of "metalearning" observed here may play an important role in the formation of apparently domain-specific biases and constraints. In general, when a potential role for differential experience exists, caution should be exercised before proposing innate biases.

Finally, in order to explain away, learners must be explaining in the first place. The present findings add to a growing body of evidence (Gopnik, 1998; Schulz and Bonawitz, 2007; Xu and Garcia, 2008; Gerken, 2010) that learning is a lot like science: in addition to making specific predictions, an important role of cognition is to build explanatory models of the environment, and to construct and test hypotheses about why the world works the way it does.

Acknowledgments

This research was supported by an NSF Graduate Research Fellowship to Colin Dawson, as well as NIH grant R01 HD042170 to LouAnn Gerken. The authors also wish to thank Brianna McMillan and Kailey Tucker of the Tweety Language Development Laboratory at the University of Arizona for their assistance with data collection.

References

Agresti, A. (2002). *Categorical Data Analysis*. Hoboken, NJ: John Wiley & Sons, Inc.

Bosch, L., & Sebastián-Gallés, N. (2003). Simultaneous bilingualism and the perception of a language-specific vowel contrast in the first year of life. Language and Speech, 46, 217-243.

Ciaramita, M., & Johnson, M. (2000). Explaining away ambiguity: Learning verb selectional preference with Bayesian networks. In Proceedings of the 18th International Conference on Computational Linguistics, 187-193, Saarbrucken, Germany.

Dawson, C. (2007). Infants Learn to Attend to Different Relations When Forming Generalizations in Different Domains. Unpublished Master's thesis.

Dawson, C., & Gerken, L. (2009). From domain-generality to domain-sensitivity: 4-Month-olds learn an abstract repetition rule in music that 7-month-olds do not. *Cognition, 111(3)*, 378-382.

Doane, S., Sohn, Y. W., & Schrieber, B. (1999). The role of processing strategies in the acquisition and transfer of a cognitive skill. *Journal of Experimental Psychology: Human Perception and Performance, 25(5)*, 1390-1410.

Eerola, T. & Toiviainen, P. (2004). MIDI Toolbox: MATLAB Tools for Music Research. University of Jyväskylä: Kopijyvä, Jyväskylä, Finland.

Folstein, J. R., Van Petten, C., Rose, S. A. (2007) Novelty and conflict in the Categorization of complex stimuli. *Psychophysiology*, 45, 467-479.

Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: the naïve theory of rational action. *Trends in Cognitive Sciences, 7*, 287-292.

Gerken, L. A. (2010). Infants use rational decision criteria for choosing among models of their input. Cognition, 115(2), 362-6.

Gerken, L., & Bollt, A. (2008). Three exemplars allow at least some linguistic generalizations: Implications for generalization mechanisms. *Language Learning and Development, 4(3)*, 228-248.

Gopnik, A. (1998) Explanation as orgasm. Minds and Machines, 8(1), 101-118.

Hannon, E. E., & Trehub, S. E. (2005). Tuning into musical rhythms: Infants learn more readily than adults. *Proceedings of the National Academy of Sciences*, 102(35), 12639-12643.

Haider, H., & Frensch, P. A. (1996). The role of information reduction in skill acquisition. *Cognitive Psychology, 30(3)*, 304-337.

Johnson, S. P., Fernandes, K. J., Frank, M. C., Kirkham, N. Z., Marcus, G. F., Rabagliati, H., & Slemmer, J. A. (in press). Abstract rule learning for visual sequences in 8- and 11-month-olds. *Infancy*.

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as Bayesian inference. *Annual Review Psychology*, 55, 271-304.

Lynch, M. P., & Eilers, R. E. (1992). A study of perceptual development for musical tuning. *Perception and Psychophysics*, 52(6), 599-608.

Marcus, G., Fernandes, K., & Johnson, S. (2007). Infant Rule-Learning Facilitated by Speech. *Psychological Science*, 18, 387-391.

Marcus, G. F., Vijayan, S., Rao, S. B., & Vishton, P. M. (1999). Rule learning by seven-month-old infants. *Science*, 283, 77–80.

Maye, J., Werker, J., & Gerken, L. A. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*, 101-111.

Murphy, R. A., Mondragon, E., & Murphy, V. A. (2008). Rule learning by rats. *Science*, *319*, 1849-1851.

Pellegrino, J. W., Doane, S. M., Fischer, S. C., & Alderton, D. (1991). Stimulus complexity effects in visual comparisons: The effects of practice and learning context. *Journal of Experimental Psychology: Human Perception and Performance, 17(3)*, 781-791.

Saffran, J. (2003). Statistical language learning: Mechanisms and constraints. *Current Directions in Psychological Science, 12*, 110-114.

Saffran, J. R., & Griepentrog, G. J. (2001). Absolute pitch in infant auditory learning: Evidence for developmental reorganization. *Developmental Psychology*, *37*, 74-85.

Saffran, J. R., Pollack, S. D., Seibel, R. L., & Shkolnik, A. (2007). Dog is a dog is a dog: Infant rule learning is not specific to language. *Cognition, 105*, 669-680.

Schulz, L., & Bonawitz, E.B. (2007) Serious fun: Preschoolers play more when evidence is confounded. Developmental Psychology, 43(4), 145-150.

Temperley, D. (2008). A probabilistic model of melody perception. *Cognitive Science: A Multidisciplinary Journal*, 32(2), 418–444.

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences, 10(7)*, 309-318.

Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, *7*, 49–63.

Xu, F., & Garcia, V. (2008). Intuitive statistics by 8-month-old infants. *Proceedings of the National Academy of Sciences of the United States of America, 105(13)*, 5012-5015.

An Analysis of the Working Memory Capacity Paradox

Eddy J. Davelaar (e.davelaar@bbk.ac.uk)
Department of Psychological Sciences, Birkbeck College
London, WC1E 7HX United Kingdom

Abstract

In the literature on working memory (WM), a paradox exists according to which very similar memory tasks provide support for very different estimates of working memory capacity. The current paper analyses the conflicting estimates of a capacity of 4+/-1 with a capacity of 1. To this end a dynamic process model of short-term recognition is used to generate data to which exponential speed-accuracy trade-off functions are fitted. The results show that even though the process model has a capacity larger than one, the exponential SAT functions indicate a one-chunk hypothesis. Further nested modeling reveals, counter to the dominant belief, that retrieval rate is insensitive to differences in WM capacity. The resolution of the WM capacity paradox lies in the choice of dependent measure.

Keywords: working memory capacity; speed-accuracy tradeoff; memory retrieval; model comparison.

Introduction

The last ten years have seen increased efforts in elucidating various aspects of working memory. Currently, there are several theories of working memory (see the chapters in Miyake & Shah, 1999) giving different explanations of behavioural data. Although many similarities exist among the theories, there are also important differences. In this paper, I will address the paradox of different estimates of working memory capacity and contrast the view that working memory can hold about 4 +/- 1 chunks (Cowan, 2001) with the view that the focus of attention is limited to 1 chunk (McElree, 2006). The paradox lies in the fact that the behavioural paradigms that provided different estimates are very similar – presentation of a sequence of words – whereas the dependent measure differs. I will use an activation-based model of working memory that has been applied to the list presentation paradigm (Davelaar, et al., 2005, 2006) and assess whether the model can reconcile the different views. Stated differently, is it possible that the estimate of 4 +/- 1 is compatible with the estimate of 1, when the paradigm-specific feature, i.e., the dependent measure, is taken into account?

The starting point is the paper by Nelson Cowan (2001) in which he reviewed a wide literature on attention and memory and concluded that the capacity limit or the focus of attention is around four chunks. Such a limit was suggested previously in a review by Donald Broadbent (1975) based on similar analyses of the literature. Furthermore, computational analyses using models such as the Search of Associative Memory (SAM; Raaijmakers & Shiffrin, 1980) supported the estimate of around four (Raaijmakers, 1982).

The commentaries based on Cowan's target article included empirical arguments supporting the view that the focus of attention is limited to one chunk (McElree & Dosher, 2001). This particular empirical argument focuses on the speed of retrieval from working memory and is central to the current paper. McElree and Dosher (2001) based their argument on data obtained using the response-signal speed-accuracy tradeoff (SAT) procedure. In this procedure, participants are presented with a sequence of words and receive a test probe after the final item. The participant has to indicate whether the test probe is one of the items in the just-presented sequence. Instead of freely responding, the participant makes a response as soon as a signal (e.g., a beep) is given. The profile of retrieval can be mapped out by employing a wide range of response signal delays. With very short delays, the participant is unlikely to have processed the test probe and performance is at chance. With a longer delay, performance rises above chance and with very long delays, performance asymptotes. The function that is traced by this procedure is called the speed-accuracy tradeoff function and can be described by or fitted with Equation 1 that involves three parameters: the intercept (T_0), the rate (s), and the asymptote (d'_{asy}).

$$d' = d'_{asy}\,(1 - e^{-s(t-T_0)}) \qquad \text{for } t>T_0,\ 0 \text{ otherwise} \quad (1)$$

The argument favouring the one-chunk hypothesis is as follows. Assume that the representation can either be in or outside the focus of attention. When it is in the focus of attention it is more readily accessible and should therefore lead to a faster rate of retrieval. This is measured by the rate parameter of the SAT function. Empirical studies consistently show (e.g., McElree, 1996; McElree & Dosher, 1989; Wickelgren, Corbett & Dosher, 1980) that the SAT function for the very last item has a faster rate than the SAT functions of the other items. In addition, the retrieval speeds for all pre-final items are equal. This suggests that the very last item is in the focus of attention, while the other items are not and thus that the capacity is limited to one item – the very last presented (or the very last processed McElree, 1998) item.

Initially, one would comment that it is possible that the most recent item is consistently in working memory, whereas the pre-final items reside in working memory with a lower probability. Therefore the estimated retrieval speeds for those items is a mixture of the fast and slow speeds, where the slow speed correspond with retrieval of presented items that are displaced from working memory (Cowan,

2001). The implied assumption underlying this view is that the probability of residing in working memory is a constant factor. Two objections to this assumption can be articulated. First, if a fixed-capacity buffer is used to encode a *sequence* of words, the probability of being in the buffer is highest for the most recent item. Thus theoretically, there is recency gradient *within* the buffer. Second, empirical observations show a recency gradient over the last four items for accuracy and reaction times (e.g., McElree & Dosher, 1989; McKone, 1995; Ratcliff, 1978), suggesting that if these items are in the buffer, a recency gradient must exist within the buffer.

To appreciate the complexities of these findings, consider that the encoding phase in the paradigms used by Raaijmakers (1982) and McElree and Dosher (1989) is identical but that the test phase differs. In addition, whereas Raaijmakers (1982) and Cowan (2001) focused on memory accuracy, McElree and Dosher (2001) focused on retrieval rate, which they argue provides direct evidence for distinct representational states. It should be said that the asymptotic accuracy of the SAT functions show a typical recency gradient. Therefore the paradox might be recast as a difference in opinion about what constitutes a proper dependent measure. This might well be the critical factor that prevents resolution of this central feature of working memory. The proposed way forward is to use a computational model with a capacity larger than one and produce the SAT functions. This requires (1) a *process* model of recognition memory that (2) implements a dynamic buffer, and (3) is capable of producing retrieval dynamics that can produce SAT functions. Several process models of recognition memory exist (Gillund & Shiffrin, 1986; Hintzman, 1984; Hockley & Murdock, 1987; McClelland & Chappell, 1998; Norman & O'Reilly, 2003; Shiffrin & Steyvers, 1997), but only a subset have been applied to SAT functions (Diller, Nobel & Shiffrin, 2001). Instead of readjusting the models to also include a dynamic buffer, the research strategy followed here is to extend a dynamic buffer model (Davelaar, et al., 2005; Haarmann & Usher, 2001) with a matching process that allows for a yes/no-recognition decision. This involves combining the dynamic buffer model with Ratcliff's (1978) diffusion model.

Model Description

The dynamic buffer model is based on the view that the content of working memory is the active part of long-term memory. More precisely, representations in consolidated memory, such as semantic long-term memory, phonological long-term memory (Baddeley, Gathercole & Papagno, 1997), and other modalities in long-term memory, are activated through sensory information. This activation is short-lived and would decay to baseline activation if there was not an active process that counteracts this decay. This process of active maintenance is a function of working memory (Baddeley, 1996) and has been called primary memory (Norman, 1968). The consequence of this process is that more than one representation can be activated simultaneously, albeit at different levels of activation. Previous work has shown that this model, which has many points of contact with Cowan's embedded processes framework (1995, 2001), is able to capture several observations in list memory paradigms. The core aspect of the model is the differential Equation 2 that governs the change of activation for every representation in long-term memory per timestep,

$$\frac{dx_i}{dt} = -x_i + \alpha F(x_i) - \beta \sum_{j \neq i}^{N} F(x_j) + I_i + \Phi(0, \sigma) \quad (2)$$

where x_i is the internal activation of representation *i*, F = 1/(1+x) is the output activation function, α captures the process of active maintenance. When α = 0, the model reduces to system with a capacity of one and is indistinguishable from theoretical models that purport to assume that only one representation can be active at any one moment (Brown, Neath & Chater, 2007; Howard & Kahana, 2002)[1]. All representations compete with each other through the inhibition parameter, β = 0.2, which governs the maximum capacity. Each representation receives activation, I_i = 0.33, from sensory processing levels. The activation dynamics is supplemented with zero-mean Gaussian noise with standard deviation, σ = 1.0. Representations that are active above a fixed threshold θ = 0.2 interact with other aspects of the cognitive system. This includes episodic memory encoding and probe matching.

The diffusion model as used by Ratcliff (1978) is in essence a dynamic signal detection model and includes the mean drift rate, ξ, which represents the amount of match between the probe and the memory item. From trial to trial the amount of match varies and this variability is captured by the standard deviation, η, of the drift rate. When applying the diffusion model to behavioural tasks, the effective drift rate for a given trial is drawn from a normal distribution with mean υ and standard deviation η. For each unit of time, zero-mean Gaussian noise with standard deviation 0.1 is added to the mean drift rate causing the total amount of evidence indicating a match or mismatch to drift towards a boundary. When a match boundary is reached, system responds with a yes-response. When a non-match boundary is reached, a no-response is emitted. The original diffusion model has many more parameters and has been applied to a wide range of reaction time paradigms. Relevant to the current discussion is that the diffusion model has been

[1] So-called single-store models include some form of relative strength calculation. When reimplementing those models in a connectionist form in order to allow direct comparison, these models require a stage where multiple representations are active to allow for the ratio-rule type of calculation. An extreme version of this is where only one representation is allowed to be active during encoding, while multiple representations are active during retrieval (Sederberg, et al., 2008).

applied to the response-signal speed-accuracy tradeoff procedure (McElree & Dosher, 1989; Ratcliff, 1978, 2006).

The diffusion model takes the value for the drift rate from the dynamic buffer model. Specifically, the drift rate on each trial is the above-threshold activation for that representation. To produce SAT functions, the following two situations need to be explicated. First, when the response-signal appears and the diffusion process has not reached any boundary, the response is based on whether the process is moving towards the yes- or no-boundary. This represents making decisions based on partial information (see for discussion, Ratcliff, 2006). Second, when a boundary has been reached before the response-signal, the corresponding decision will be given at the time of the response-signal. The resulting decision probabilities are converted into d' scores and the full SAT functions are fitted with two version of Equation 1. In version 1, all parameters are free to vary across conditions, yielding 18 free parameters. In version 2, the reduced model that is supported by the empirical literature is used. This model has a fixed T_0 for all conditions and two different rates, yielding 9 free parameters.

The process model as described above was applied to a sequence of six words. Each of six representations was activated sequentially for 1,000 iterations. Then one of the six positions was probed and a SAT function created for that serial position by using response-signals at 100, 200, 300, 400, 500, 750, 1,000, 1,500, 2,000, and 3,000 iterations. Each serial position was probed 1,000 times at each of the ten response-signal delays. The effective capacity of the model is easily assessed by counting the number of representations that are active above threshold at t = 6,000 iterations. In order to address the possibility that different parameters obtained from the exponential SAT function are sensitive to different working memory capacities, the simulations are repeated for $\alpha = 0$ (no buffer), $\alpha = 1.8$ (small capacity), and $\alpha = 2.0$ (large capacity).

Simulation Results

Figure 1 shows a noise-less simulation of a sequence (with $\alpha = 2.0$). At time = 6,000, the very last item is the most active and activation levels decrease with the temporal distance of presentation. Figure 2 shows the frequency distribution of the activations for each of the items in Figure 1 at t = 6,000 iterations. As can be seen, items that are still in the activation buffer at time of test show a step-like function, with the very last item being more active than all other active items, which in turn have similar activation levels. The reason for this is immediately apparent when taking a closer look at Equation 2. Assume that at time of test, the activation level does not change and is above threshold. The resulting $F(x_i)$ is governed by α and β, leading to convergence of the activations. Only the very last item still receives external input, leading to a higher activation.

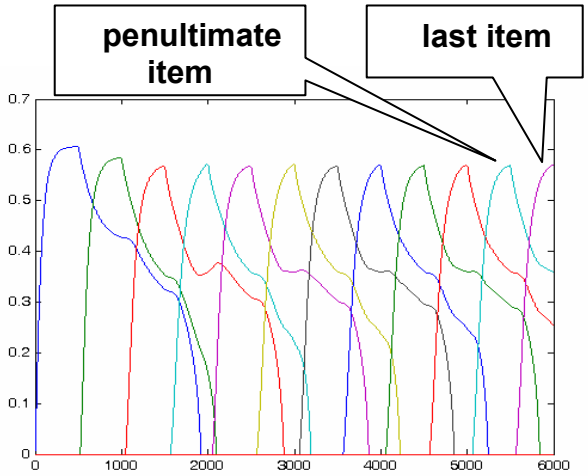


Figure 1. A noise-less simulation of 12 sequentially activated items. The x-axis indicates time in iterations. The y-axis indicates activation level, $F(x_i)$.

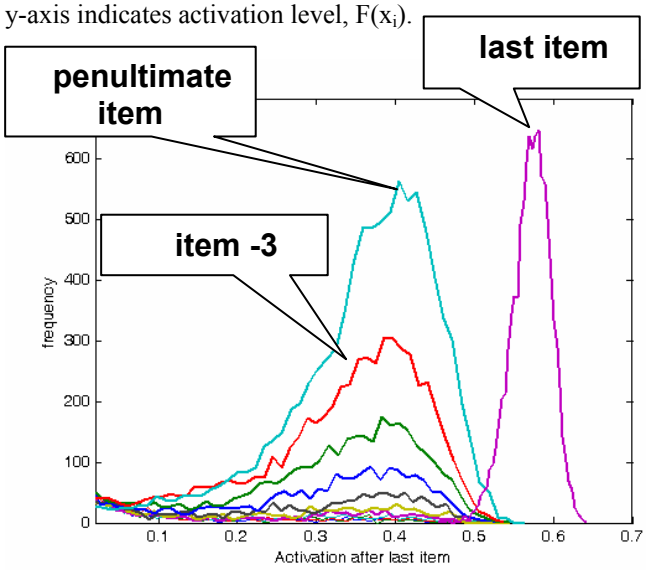


Figure 2. Frequency distributions of the activation levels of the 12 items in Figure 1 at t = 6,000 iterations.

The simulated data and corresponding best-fitting SAT functions for the simulation of $\alpha = 2.0$ are presented in Figure 3. Table 1 shows the parameter values of the best-fitting reduced model for each of the values of α. The models were fit by maximising the adjusted R^2.

Although the reduced model fits the data less well compared to the saturated model, the change in goodness of fit, ΔR^2, is negligible given the amount of variability present in real data. This supports the findings in the empirical literature that led to the one-chunk hypothesis. However, the model maintains multiple items at the time of test, as seen by the capacities. The capacity at $\alpha = 2.0$ is higher than at $\alpha = 1.8$.

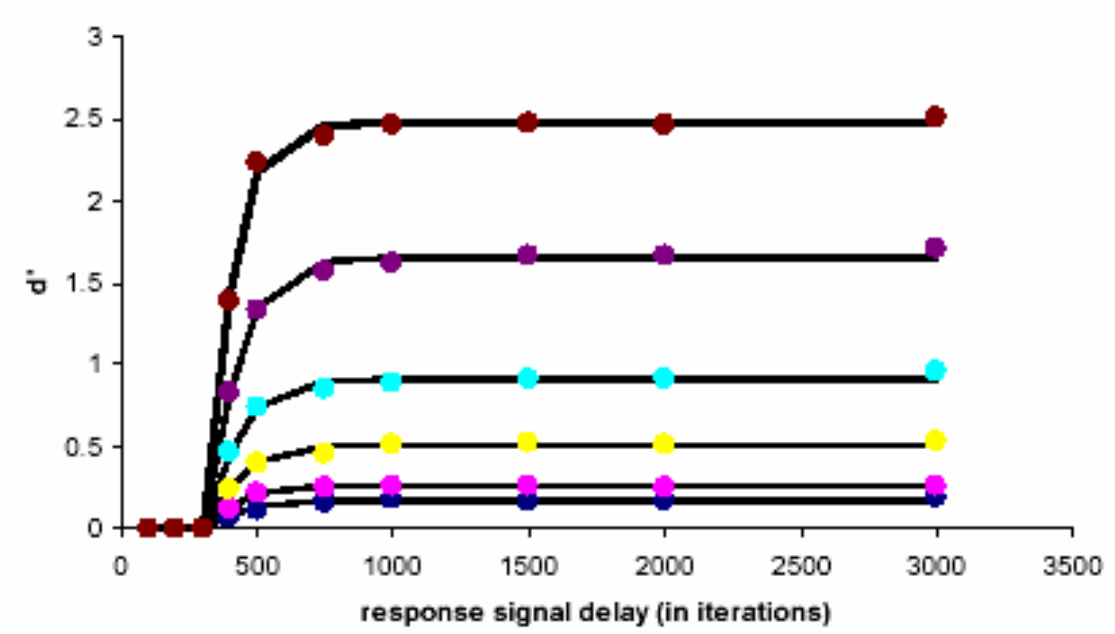


Figure 3. Simulation data and best-fitting reduced model for the simulation with α = 2.0.

Table 1: Parameter estimates for the 9-parameter exponential SAT function and the estimates of buffer capacity.

		simulation		
parameters	Serial position	α = 0	α = 1.8	α = 2.0
d'_{asy}	1	0.015	0.014	0.173
d'_{asy}	2	0.028	0.031	0.261
d'_{asy}	3	0.000	0.107	0.509
d'_{asy}	4	0.025	0.632	0.910
d'_{asy}	5	0.018	1.966	1.652
d'_{asy}	6	1.208	3.760	2.471
T_0	1-6	279.56	338.12	33.92
s	1-5	0.0005	0.0068	0.0102
s	6	0.0019	0.0088	0.0129
R^2-adjusted		.996	.999	.999
ΔR^2		0	.001	0.0002
capacity		1	2.64	3.38

Note: the capacity was estimated by counting the number of above-threshold representations at t = 6,000 iterations.

The parameter values for the d'_{asy} are well-fitted by an exponential function, allowing the 6 free parameters to be reduced to 2 free parameters. In addition, s could be fitted with a function with only 1 parameter. Therefore, the best-fitting 9-parameter model could be further reduced to a 4-parameter model. This further parameter reduction allowed an examination of model fit as a function of differences in buffer capacity. To do this the data form the simulations with α = 1.8 and α = 2.0 were compared. This resulted in a "full" model having 8 free parameters with 4 parameters for each α-level. The 8-parameter model, $[2F(d'_{asy}) - 2G(s) - 2H(T_0)]$, (F(x) has 2 parameters) and all nested models were fit to 120 datapoints by maximizing the adjusted R^2. Of special interest was the identification of parameters that reduce the fit and thus carry the difference in buffer capacity. The results are shown in Table 2 and are clear-cut. The goodness of fit is largely unaffected when G(s) or $H(T_0)$ is fixed between the two levels of α. However, a 5% decrease in fit is observed when $F(d'_{asy})$ is fixed. The interpretation of this finding is that differences in buffer capacity are only picked up in the differences in *gradient* of the d'_{asy} function. The rate parameter seems insensitive to variation in buffer capacity and is therefore only useful to assess which item or one-chunk was the most-recently processed.

Table 2: Results of nested modeling fits on the data from the two different WM capacity simulations. The number of free parameters are given between brackets after each model.

Model	Degrees of freedom	adjusted R^2
Full model (8)	112	.989
F-fixed (6)	114	.942
G-fixed (7)	113	.988
H-fixed (7)	113	.989
F/G-fixed (5)	115	.942
F/H-fixed (5)	115	.943
G/H-fixed (6)	114	.987
All fixed (4)	116	.943

Discussion

This paper focused on the paradox that different estimates of working memory capacity are estimated based on very similar tasks. Using a dynamic model of short-term recognition, data were generated and fitted by exponential SAT functions. Contrary to what was previously thought, the results show that the rate of retrieval from WM is insensitive to the WM capacity and instead is most sensitive to the recency of cognitive processing. The asymptotic accuracy is found to be the only parameter that is sensitive to WM capacity. The resolution of the WM paradox lies in the choice of dependent measure, with accuracy being the preferred measure for estimating WM capacity and retrieval rate being the preferred measure for identifying the most recently processed chunk in WM.

The process model predicts that items that are not in WM will lead to misses. Therefore for items that were presented a very long time ago, only misses should happen. This is partially correct. One would, however, expect that deactivated items require an additional process of episodic retrieval to allow for contextual matching. This is likely to result in slower retrieval dynamics and quite likely to a larger intercept. The problem is that in order to assess this possibility, trials would have to be separated into those in which the probe matches with a deactivated item and trials in which the probe matches a pre-recency active item. This is not possible experimentally and thus differences in intercept for pre-recency items are always mixtures. The same holds for the retrieval speeds. With long lists, very early items could be probed and used to check if they do have the slowest retrieval speed and the largest intercept. The difficulty here is that performance is close to chance (Wickelgren, Corbett & Dosher, 1980). Wickelgren et al. used a 16-word list and measured the SAT of the list item -12 (position 4). In some of the participants, the intercept for the item -12 was larger than all other items. Although this might suggest that the intercept is the preferred parameter to assess whether items are retrieved from WM or form long-term memory, a thorough empirical investigation waits.

What does the reinterpretation of the exponential SAT-parameters mean for the use of the exponential SAT-procedure? Several authors have commented that exponential and diffusion SAT are too similar to be distinguished (McElree & Dosher, 1989; Ratcliff, 2006). Others have argued that diffusion SAT should be used as it is based on an actual theory of memory retrieval (Ratcliff, 2006), whereas the exponential SAT is not based on a theory and therefore only of statistically-descriptive use. Despite the finding that exponential SAT can not be used to address capacity estimates, it is able to identify the last processed item (McElree, 1998). This utility depends heavily on the assumption that across many trials, participants process the stimuli in identical ways. Whether the SAT-procedure is robust against violation of the identical-processing assumption remains for future analyses. What does all this mean for WM capacity? The analyses presented here suggest that WM can hold multiple items in an active stat to varying degrees, but that the very last processes item is in a highly accessible state. The work also demonstrates more generally the importance of using explicit formal analyses to verify the interpretations based on statistical tests.

References

Baddeley, A. D. (1996). Exploring the central executive. *Quarterly Journal of Experimental Psychology, 49*, 5-28.

Baddeley, A. D., Gathercole, S., & Papagno, C. (1998). The phonological loop as a language learning device. *Psychological Review, 105*, 158–73.

Broadbent, D. E. (1975). The magic number seven after fifteen year, In: *Studies in long-term memory* (Kennedy, A. and Wilkes, A., eds), pp. 3–18, John Wiley and Sons.

Brown, G. D. A., Neath, I., & Chater, N. (2007). A temporal ratio model of memory. *Psychological Review, 114*, 539-576.

Cowan, N. (1995). *Attention and memory: an integrated framework*. Oxford: Oxford University Press.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24*, 87-185.

Davelaar, E. J., Goshen-Gottstein, Y., Ashkenazi, A., Haarmann, H. J., & Usher, M. (2005). The demise of short-term memory revisited: empirical and computational investigations of recency effects. *Psychological Review, 112*, 3-42.

Davelaar, E. J., Haarmann, H. J., Goshen-Gottstein, Y., & Usher, M. (2006). Semantic similarity dissociates short- from long-term memory: testing a neurocomputational model of list memory. *Memory & Cognition, 34*, 323-334.

Diller, D. E., Nobel, P.A., & Shiffrin, R. M. (2001). An ARC-REM model for accuracy and response time in recognition and cued recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*, 414-435.

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review, 91*, 1-67.

Haarmann, H. J., & Usher, M. (2001). Maintenance of semantic information in capacity limited item short-term memory. *Psychonomic Bulletin & Review, 8*, 568-578.

Hintzman, D. L. (1984). MINERVA 2: A simulation model of human memory. *Behavior Research Methods, Instruments, & Computers, 16*, 96-101.

Hockley, W. E., & Murdock, B. B. (1987). A decision model for accuracy and response latency in recognition memory. *Psychological Review, 94*, 341-358.

Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology, 46*, 269-299.

McClelland, J. L., & Chappell, M. (1998). Familiarity breeds differentiation: a subjective-likelihood appoach to the effects of experience in recognition memory. *Psychological Review, 105*, 724-760.

McElree, B. & Dosher, B. A. (1989). Serial position and set size in short-term memory: Time course of recognition. *Journal of Experimental Psychology: General, 18*, 346-373.

McElree, B. & Dosher, B. A. (1993). Serial retrieval processes in the recovery of order information. *Journal of Experimental Psychology: General, 122*, 291-315.

McElree, B. (1996). Accessing short-term memory with semantic and phonological information: A time-course analysis. *Memory & Cognition, 24*, 173-187.

McElree, B. (1998). Attended and non-attended states in working memory: Accessing categorized structures. *Journal of Memory & Language, 38*, 225-252.

McElree, B. (2001). Working memory and focal attention. *Journal of Experimental Psychology: Learning, Memory & Cognition, 27*, 817-835.

McElree, B. (2006). Accessing recent events. In B. H. Ross (Ed.), *The psychology of learning and motivation*, Vol. 46. San Diego: Academic Press.

McElree, B., & Dosher, B. A. (2001). The focus of attention across space and across time. *Behavioral and Brain Sciences, 24*, 129-130.

McKone, E. (1995). Short term implicit memory for words and nonwords. *Journal of Experimental Psychology: Learning, Memory and Cognition, 21*, 1108-1126

Miyake, A., & Shah, P. (Eds.) (1999). *Models of working memory: Mechanisms of active maintenance and executive control.* New York: Cambridge University Press.

Norman, D. A. (1968). Toward a theory of memory and attention. *Psychological Review, 75*, 522-536.

Norman, K. A., & O'Reilly, R. C. (2003). Modeling hippocampal and neocortical contributions to recognition memory: A complementary learning systems approach. *Psychological Review, 110*, 611-646.

Raaijmakers, J. G. W. (1982). A note on the measurement of primary memory capacity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 8*, 343-352.

Raaijmakers, J. G. W., & Shiffrin, R. M. (1981). Search of assocative memory. *Psychological Review, 88*, 93-134.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*, 59-108.

Ratcliff, R. (2006). Modeling response signal and response time data. *Cognitive Psychology*, *53*, 195-237.

Sederberg P. B., Howard M. W., & Kahana M. J. (2008) A context-based theory of recency and contiguity in free recall. *Psychological Review*, 115, 893-912.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM: Retrieving Effectively from Memory. *Psychonomic Bulletin & Review, 4*, 145-166.

Wickelgren, W. A., Corbett, A. T., & Dosher, B. A. (1980). Priming and retrieval from short-term memory: a speed accuracy trade-off analysis. *Journal of Verbal Learning and Verbal Behavior, 19*, 387-404.

Working Memory Load Affects Device-Specific but Not Task-Specific Error Rates

Maartje G. A. Ament (M.Ament@ucl.ac.uk)

Anna L. Cox (Anna.Cox@ucl.ac.uk)

Ann Blandford (A. Blandford@ucl.ac.uk)

Duncan Brumby (Brumby@cs.ucl.ac.uk)

UCL Interaction Centre
University College London
Gower Street London, WC1E 6BT

Abstract

Human error in routine procedural tasks is often attributed to momentary failures to remember what step to perform. We argue that task-specific steps, which can be defined as actions required to achieve a particular goal across a variety of different devices, are far less prone to error than device-specific steps, which can be defined as actions that are required for the operation of the device but do not directly contribute to the goal. An experiment is reported that supports this distinction, showing that device-specific steps are more error prone than task-specific steps. Moreover, we argue that these errors reflect a failure of memory because the error rate for device-specific steps was sensitive to increased working memory load, while the error rate for task-specific steps was not. The current work demonstrates that a distinction between device- and task-specific steps can be effective in explaining error patterns observed on a specific task.

Keywords: human error; device-specific error; working memory load.

Introduction

While routine procedural errors occur only occasionally, they are persistent. A growing body of empirical work has studied these errors in the laboratory. Most of them have focussed on the post-completion error (PCE) (e.g. Byrne & Bovair, 1997; Chung & Byrne, 2008; Li, Blandford, Cairns, & Young, 2008), a cognitive slip that occurs when the final step in a task is omitted after the main goal has already been completed.

The PCE is theoretically well understood. An influential account is the *memory-for-goals* model developed by Altmann and Trafton (2002). This account assumes that goals are declarative memory representations (chunks) with an associated activation level. The interference level is defined as the 'collective effect of distractor goals'. In order to direct behaviour, the relevant goal needs to be above the interference level. In order to overcome the interference level, the activation of goals must be strengthened. A goal that is retrieved more often or the most recently retrieved subgoal will have a higher activation value than others with less history. Associative links between goals allow activation to spread to other goals. The PC step is usually remembered because it receives associative activation from the step preceding it. Moreover, Byrne and Bovair (1997) have argued that upon completion of the main goal, the sources of activation for the PC subgoal are reduced, leading to lower activation on the PC subgoal, often to a point where it cannot be retrieved.

Another step that is associated with a relatively high error rate is the device-initialisation (DI) step. A device initialisation step is an action that must be executed before the main task steps can be completed (e.g. pressing a 'mode' key before setting the alarm on a digital watch). Li et al. (2008) and Hiltz, Back & Blandford (2010) found relatively high error rates on both the post-completion and the device-initialisation steps. However, this error is less well understood, and it is not clear how the *memory-for-goals* model would account for it. For this error, the main goal has not yet been completed, so should still provide activation for the device-initialisation step.

A common factor that the PC step and the DI step share is that they are both device-specific (Cox & Young, 2000). This means that they do not make a direct contribution towards the main goal, but are only required for the correct operation of the device. Task-specific steps, on the other hand, do make a direct contribution towards the main goal and are required regardless of the type of device they are carried out on. Consider the example of using a state-of-the-art induction hob. A typical task-specific step may be to increase or decrease the power output by pressing the '+' or '-' button, whereas a device-specific step may be to press the selector button to cycle through the different hobs until you have selected the one for which you want to adjust the power. While a number of previous studies have discussed concepts similar to device- and task-specific steps (e.g. Cox & Young, 2000; Kirschenbaum, Gray, Ehret, & Miller, 1996; Gray, 2000), this is a novel approach to explaining routine procedural errors.

In this paper, we propose that the distinction between task-specific and device-specific steps can explain why some steps in a procedure appear to be more error prone than others. Our account relies on the user having a task model (how to do the task) and a device model (how to do the task using a particular device), two concepts widely used

in the field of human-computer interaction research (Young, 1983). Device-specific steps are only represented in the device model, whereas task-specific steps are represented in both. Using an activation-based approach, the current work hypothesises that device-specific steps have lower activation levels, because they have only one source of activation (the device model), whereas task-specific steps receive activation from two sources (the device model and the task model). These lower activation levels make it more likely that device-specific steps fall below the interference level, resulting in a slip. Ament, Blandford & Cox (2009) describe an experiment in which device-specific error rates on the 'Spy task' were significantly higher than those on task-specific steps, as predicted.

There are two aims to this paper. First, we seek to provide empirical evidence to support the idea that error rates are higher on device-specific steps than on task-specific steps. Second, we investigate the effect that varying working memory load has on these two classes of steps. We argue there is good reason to believe that device-specific steps are more susceptible to the deleterious effects of increased working memory load than task-specific steps.

Byrne and Bovair (1997) argued that post-completion errors are memory-based failures. Therefore, they investigated how working memory load affects the PCE. They found that the frequency of the PCE increased under a high working memory load. Byrne and Bovair (1997) argued that a higher working memory load leads to the scaling back of activation on all items in memory. This means that the decay rate is higher, and items are displaced from memory faster. If the source of activation for an item is lost, such as on the post-completion step, it is more likely that that step will not reach the threshold necessary to be executed and a post-completion error will be likely.

However, this account does not explain how working memory load would affect other device-specific errors, since their source of activation is not lost like that of the PC step. In the *memory-for-goals* model (Altmann & Trafton, 2002), higher working memory load is represented by an increased interference level. While no direct predictions about the effect of this are made, it seems clear that an increased interference level makes it more likely that the activation level for a given action falls below it, leading to an error. We therefore hypothesise that device-specific errors should be particularly affected by an increase in working memory load, because a higher interference level makes it even more difficult for device-specific steps to overcome this. Conversely, task-specific steps are expected to be affected less, because their higher activation levels make them more robust to increases in the interference level.

We investigate the effect of working memory load on device-specific and task-specific error rates, by means of a secondary load task. It is expected that in low memory load conditions, participants will make fewer errors overall compared to high load conditions. Critically, it is expected that, under high load, there will be proportionally more errors on device- specific steps than on task-specific steps.

Method

Participants

Forty participants were recruited from a dedicated psychology subject database. They were aged between 18

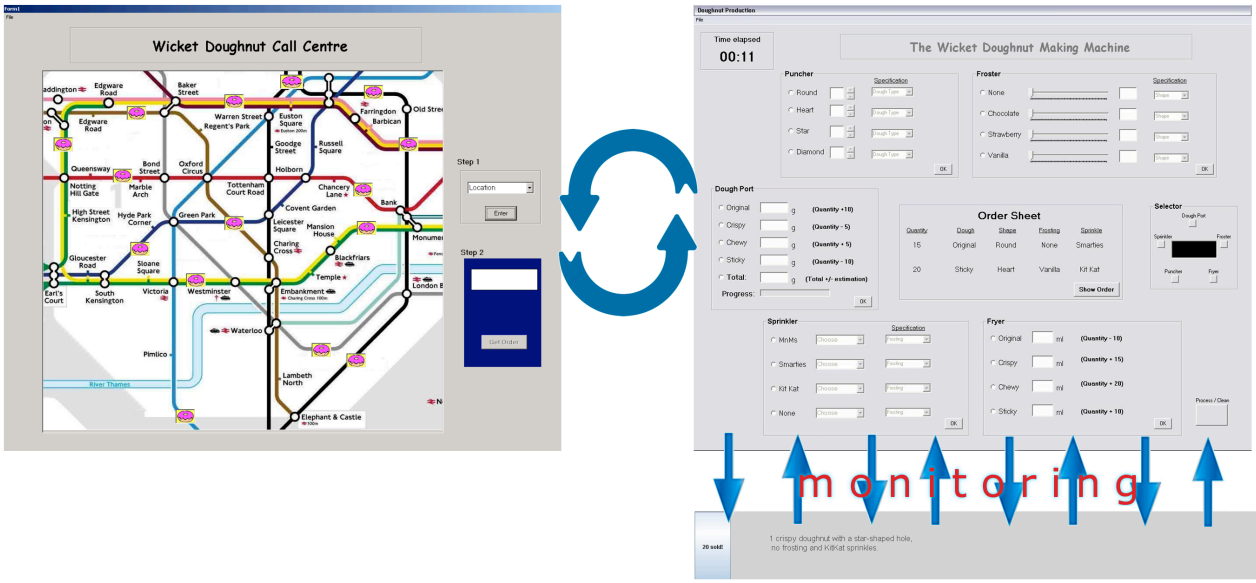


Figure 1: Diagrammatic representation of the Doughnut task. On the top right is the main Doughnut task interface. While making the doughnuts, participants monitor the Doughnut Live Feed, displayed directly underneath the main Doughnut task interface. In between doughnut making trials, participants answer a call at the Call Centre, displayed on the left.

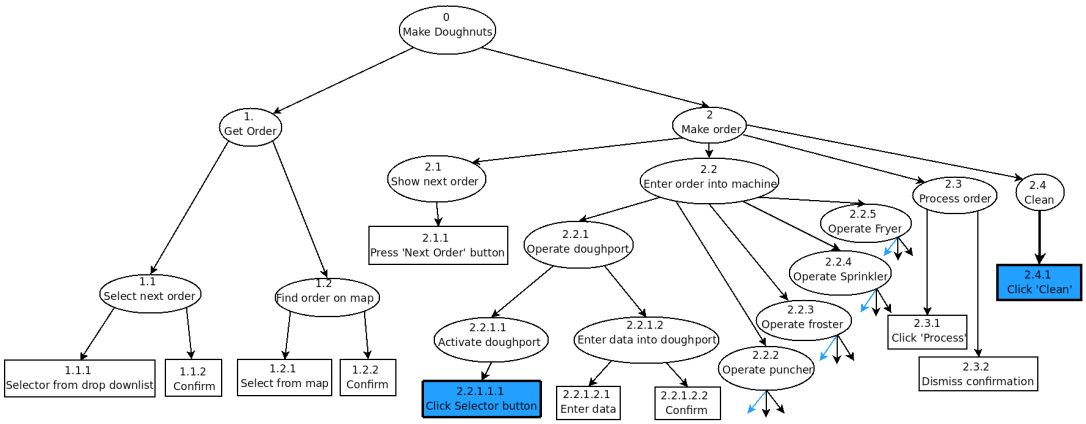


Figure 2: Hierarchical task analysis of the doughnut task. Step 2.2.1.1.1 is the device-initialisation/device-specific step, whereas step 2.4.1 is the post-completion step; both are shaded. Note that the 'Operate Puncher', 'Operate Froster', 'Operate Sprinkler' and 'Operate Fryer' subgoals are not defined further to save space; they are identical in structure to 'Operate Doughport' and as such also contain a device-specific step at the beginning.

and 33 with a mean age of 22.0, and 27 were female. The majority of participants were students, and they were paid £6 for their time.

Materials

The Wicket Doughnut task (Li, 2006), a routine procedural task in which participants have to follow a defined procedure to make virtual doughnuts, was used. Figure 1 shows the components of the doughnut task: the main doughnut interface, the call centre (both developed by Li (2006)), and the live feed (developed for the current study). Figure 2 shows a hierarchical task analysis of the doughnut and call centre tasks. The main task consists of two subtasks (represented as ovals), which are further subdivided into smaller subgoals. The square boxes represent the lowest-level goals and correspond to discrete actions. Device-specific steps are shaded. While only two are shown in the figure to save space, the task contained a total of 6 device-specific steps; the steps that are not shown are the initial selector steps on the Puncher, Froster, Sprinkler and Fryer subtasks.

A trial starts with taking a call at the call centre to get the next order, done on a separate computer terminal. It involves selecting the correct doughnut shop from a list, and finding it on a map. After confirming, the order is then 'transferred' to the Doughnut task interface on another computer terminal.

The main doughnut task consists of five compartments, or widgets, in which participants have to enter information from the order sheet. These need to be operated in the order: Dough Port → Puncher → Froster → Sprinkler → Fryer Before data can be entered, a widget needs to be activated by clicking the appropriate selector button on the selector panel on the right-hand side. Clicking the Ok button then confirms the entry for that widget. Once all widgets have been completed, the order needs to be processed by clicking the 'Process' button. A pop-up screen then indicates the completion of the trial, and the number of doughnuts made. At the end of the trial, the machine must be cleaned by clicking the 'Clean' button. While Li et al. (2008) used interruptions at certain points during the task, the current experiment did not.

To vary working memory load, a monitoring task was added in which participants had to count the number of doughnuts sold in the shops. The Doughnut Live Feed was shown at the bottom of the screen, where occasionally a description of a doughnut was shown. Participants had to attend to a specific characteristic of the doughnut (such as dough type, hole shape or frosting) and keep count of how many with that characteristic were sold. In the low working memory load condition, participants were asked to attend to and keep track of doughnuts with a specific dough type, for instance Crispy. In the high working memory load condition, participants were asked to attend to and separately keep track of doughnuts with a specific dough type and those with a specific hole shape. In both conditions, once a participant had counted 20 doughnuts of



Figure 3: the doughnut live feed. A cycle starts out completely white (a). The background then quickly fades to grey, while the item fades from white to black (b). Halfway through the cycle, the background and the item are at its darkest, and the item is clearly visible (c). At the end of the cycle, the background fades to white again while the item may either stay visible or fade as well (d).

the specified type, they had to click the button on the left of the live feed and start counting from zero again. This allowed the experimenter to assess whether a participant was successfully monitoring the live feed.

To ensure effective monitoring, new items on the live feed did not capture visual attention. This was achieved by using a background that changed from grey to white and back in continuous cycles. Each doughnut description faded in on top of that from white to black, and faded out again after a random number of cycles. Figure 3 shows the progression through one cycle. Each cycle took three seconds, and items remained visible for between 2 and 4 cycles. This randomness made it impossible for participants to predict when a new doughnut description would be shown. The monitoring task and primary tasks were carried out simultaneously.

A number of device-specific steps were present in the doughnut task. Selecting the first compartment, the dough port, was a device-initialisation step. The other selecting steps were counted as other device-specific steps. The last step in the procedure, cleaning the machine, was a post-completion step. A false completion signal was given in the form of a pop-up screen indicating that the doughnuts were ready. In addition, a flashing message notifying the participant of the next call provided a competing signal for the post-completion step. After dismissing this pop-up, the post-completion step took place.

Two separate computer terminals were used; one for the call centre and one for the doughnut making task and live feed. Both screens were operating at a resolution of 1280 x 1024 pixels.

Design

A mixed design was used, with two levels for each independent variable. The first independent variable was working memory load; this was varied between participants. This variable had two levels: low load and high load. The second independent variable was the type of step; this was varied within participant. This variable had two main levels, device-specific and task-specific.

The dependent variable was the error rate. Errors were counted systematically according to the required steps. An error is defined as any action that deviates from the required action at a certain step. To ensure only inappropriate actions are counted and not each individual inappropriate click, only one error could be made on each step.

Procedure

Participants carried out the experiment individually. During the training phase, participants were given an instruction sheet that explained in detail what their task was, and all the procedures necessary to complete the task. After reading the instruction sheet, they observed the experimenter doing the task once, after which they were allowed to practice it twice. Any errors made during the training trials were pointed out immediately using the default Windows XP notification sound and were required to be corrected before the participant was allowed to move on. After each practice trial, the experimenter asked the participant how many doughnuts they had counted on the live feed, and encouraged more accurate performance if necessary.

Participants were instructed to complete the doughnut task as quickly and as accurately as possible. A timer was displayed on the screen throughout the experiment to encourage swift performance; it was reset after each trial. After processing the doughnuts, a pop-up screen notified the participant of the number of doughnuts made. Participants were also told to count the doughnuts in the live feed as accurately as possible; this was further encouraged by the '20 doughnuts' button. Participants were not aware that errors were being studied.

During the experimental phase, the participants completed 11 trials, with the opportunity of a short break after 6 trials. Any errors were pointed out immediately and had to be corrected before the participant was allowed to carry on. The total duration of the experiment was approximately 60 minutes.

Results

Data from 12 participants was excluded from the analysis. The reasons for excluding participants varied. Three participants were excluded because they failed to follow the instructions to monitor the live feed correctly. One participant's data sheet was lost. Eight participants were excluded because they made omission errors at any step on more than 65% of trials. The reason for excluding these error-prone participants is that such high error rates likely

Type of Step	Error count (Opportunity)	Mean error rate (SD), in %
Total	292 (5852)	4.99 (2.51)
Task-specific	57 (4004)	1.42 (0.96)
Device-specific	235 (1848)	12.7 (7.44)
Device-initialisation	84 (308)	27.27 (20.55)
Post-completion	66 (308)	21.43 (21.60)
Other device-specific	85 (1232)	6.90 (6.47)

Table 1: Total error counts and mean error rates across all participants and conditions for the different types of steps.

indicate that the participant has not correctly learnt how to perform the task. We present analysis of error-rate for the remaining twenty-eight participants.

Due to the failure of so many participants to perform the task to criterion, we first examine whether error rate decreased as participants gained more experience at performing the task. There was no evidence of a learning effect over consecutive trials; that is, there was no relationship between number of errors per trial and trial number (τ = -0.26, p = 0.27). This suggests that those included in the analysis had been effectively trained before conducting the study.

We were primarily interested in error rates at device-specific and task-specific steps. Error rates were calculated for each participant for the relevant step types. Only one error was possible on each of the steps. Step 19 (dismissing the pop-up screen) was removed from further discussion, because no error was possible on this step, since the pop-up screen blocked action on the main screen. Thus, a total of 19 errors could be made on a single trial. Each participant did 11 trials, and data from 28 participants was analysed, giving a total opportunity for errors of 19 x 11 x 28 = 5852. Across all participants, a total of 292 errors were made, giving an overall error rate of 4.99%.

It was hypothesised that error rates were higher on device- than on task-specific steps. Table 1 shows the average error rates across all participants on the different types of steps. A repeated-measures ANOVA, comparing error rates on task-specific, device-initialisation, post-completion and other device-specific steps, showed a significant difference between the types of steps, F(3,81) = 19.46, p = 0.000, with Greenhouse-Geisser correction. A post-hoc comparison showed that task-specific steps had significantly lower error rates than all device-specific steps. Looking more specifically at the different types of device-specific steps, it becomes clear that the error rates on DI and PC steps are higher than on the other steps. Post-hoc tests confirm that PC and DI steps have significantly higher error rates than both task-specific steps and other device-specific steps, although there is no significant difference between PC and DI steps.

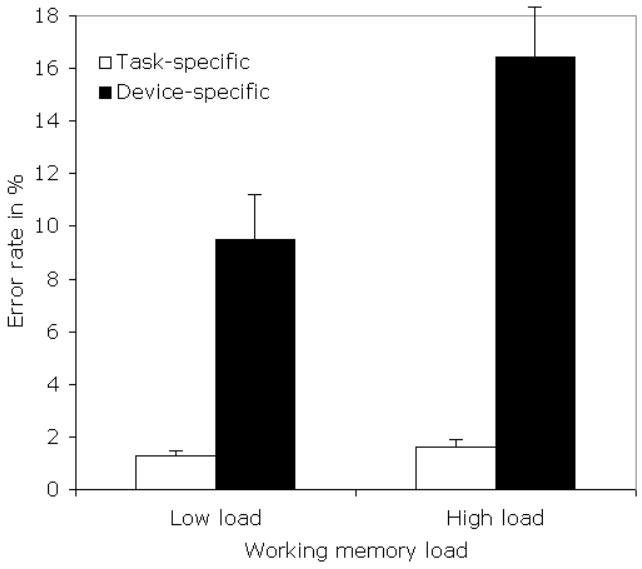


Figure 4: Error rates across working memory load and type of step conditions. Error bars represent the standard error of the mean.

Working memory load was also manipulated on two levels, low load and high load. Figure 4 shows the error rates on the different working memory load levels, for both device- and task-specific steps. Error rates on task-specific steps remained stable across all conditions, while error rates on device-specific steps increased under high working memory load. A 2 x 2 mixed-design ANOVA with type of step as the within-subjects variable and working memory load as the between-subjects variable revealed a main effect of working memory load, F(1,26) = 8.10, p = 0.009. An interaction effect was also found, F(1,26) = 6.68, p = 0.016. A main effect of type of step was also found to be significant, F(1,26) = 81.90, p = 0.000. Simple effects analysis showed that there was no simple effect of working memory load on task-specific steps, F(1,26) = 0.95, p = 0.339. There was a simple effect of working memory load on device-specific steps, F(1,26) = 7.53, p = 0.011.

Discussion

The current experiment investigated the hypothesis that error rates on device-specific steps are higher than on task-specific steps, and that working memory load has a differential influence on them. The results of this study show that the error rates observed at device-specific steps is greater than the error rates observed at task-specific steps. Also, a high working memory load resulted in higher error rates overall. In addition, an interaction effect of working memory load and type of step was found. This supports our predictions.

It can be argued that the finding that error rates are higher on device-specific than on task-specific steps is mainly due to the high error rates on device-initialisation and post-completion steps. However, it should be noted that the error rate on the 'other device-specific steps' was also found to be higher than that on task-specific steps. This indicates that device-specific steps are indeed associated with higher error rates than task-specific steps. Nevertheless, the relatively high error rates on the PC and DI steps may indicate that other factors play a role as well.

Byrne and Bovair (1997) found that only low-capacity individuals were affected by a high working memory load. Although we did not administer working memory capacity tests to participants, the fact that working memory load had a significant effect without dividing participants into low and high capacity groups suggests that this is unlikely to have adversely affected the results.

As expected, working memory load increases the overall error rates. The significant interaction indicated that this

effect is much stronger on device-specific than on task-specific steps. This confirms our predictions.

The current work has implications for theoretical models of error. We hypothesised that device-specific steps have lower activation levels, and are therefore more likely to fall below the interference level. The higher error rates on device-specific steps are in line with this explanation. In addition, the differential influence of working memory load on the two types of steps further supports our theory. It is not clear how the *memory-for-goals* model would account for the lower activation on device-specific steps, highlighting a possible limitation of the model.

Apart from higher error rates and a greater influence of working memory load, these lower activation levels make a number of further predictions. First, reaction times should be longer on device-specific steps. A lower activation level on such steps means that more time is needed for the activation level to increase above the interference level, in order to execute the associated step. Due to the nature of the steps within the doughnut task, it is not appropriate to conduct this analysis on the data from the experiment reported in this paper. Future studies should use a more suitable task to investigate the differences in reaction times on device- and task-specific steps.

Second, device-specific errors should be qualitatively different from task-specific errors. It is more difficult for device-specific steps to overcome the interference level, making it more likely that the step's activation inadvertently falls below the interference level. When this happens, it is likely that the next step has the highest activation level and directs behaviour: an omission error occurs. On the other hand, the higher activation levels on task-specific steps make it less likely that the step accidentally falls below the interference level. Instead, other errors such as incorrect sequence errors (i.e. performing a different task-specific step that is out of sequence) may be more common.

The current work also has implications for the design of interactive systems by going beyond the well-studied PCE. While PC steps are relatively rare, device-specific steps occur on many devices. The current results have demonstrated that device-specific steps are more prone to errors than their task-specific counterparts, and therefore these steps should be avoided in task design where possible.

Conclusion

The current study demonstrated that people are more likely to make errors on device-specific steps than on task-specific steps, providing support for the claim that this distinction can be effective in explaining observed error patterns. Moreover, working memory load was found to have a greater effect on device-specific error rates than on task-specific ones, providing support for our hypothesis that device-specific steps have lower activation levels. Future studies can look more closely at the mechanisms underlying device- and task-specific steps, and investigate how these can lead to different activation levels.

Acknowledgements

This work is supported by an EPSRC DTA studentship. We thank Simon Li for the use of his Doughnut Task.

References

Altmann, E. M., & Trafton, J. G. (2002). Memory for goals: An activation-based model. *Cognitive Science, 26*, 39-83.

Ament, M. G. A., Blandford, A., & Cox, A. L. (2009). Different cognitive mechanisms account for different types of procedural steps. In N. A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31th Annual Conference of the Cognitive Science Society* (p. 2170-2175).

Byrne, M. D., & Bovair, S. (1997). A working memory model of a common procedural error. *Cognitive Science, 21(1)*, 31-61.

Byrne, M. D., & Davis, (2006) Task structure and postcompletion error in the execution of a routine procedure. *Human Factors, 48,* 627-638.

Chung, P. H., & Byrne, M. D. (2004). Visual cues to reduce errors in a routine procedural task. In *Proceedings of the Twenty-Sixth Annual Conference of the Cognitive Science Society*.

Chung, P. H., & Byrne, M. D. (2008). Cue effectiveness in mitigating postcompletion errors in a routine procedural task. *International Journal of Human-Computer Studies, 66,* 217-232.

Cox, A. L., & Young, R. M. (2000). Device-oriented and task-oriented exploratory learning of interactive devices. In N. Taatgen & J. Aasman (Eds.), *Proceedings of the Third International Conference on Cognitive Modelling* (p. 70-77). Veenendaal, The Netherlands: Universal Press.

Gray, W. D. (2000). The nature and processing of errors in interactive behaviour. *Cognitive Science, 24(2)*, 205248.

Hiltz, Back & Blandford (2010) The roles of conceptual device models and user goals in avoiding device initialization errors. *Interacting with Computers*. DOI http://dx.doi.org/10.1016/j.intcom.2010.01.001.

Kirschenbaum, S. S., Gray, W. D., Ehret, B. D., & Miller, S. L. (1996). When using the tool interferes with doing the task. In *Proceedings of CHI '96*. Vancouver, Canada.

Li, S. Y.-W. (2006). *An empirical investigation of post-completion error: A cognitive perspective*. Unpublished doctoral dissertation, Department of Psychology, UCL.

Li, S. Y.-W., Blandford, A., Cairns, P., & Young, R. M. (2008). The effect of interruptions on postcompletion and other procedural errors: An account based on the activation-based goal memory model. *Journal of Experimental Psychology: Applied, 14(4)*, 314-328.

Young, R.M. (1983) Surrogates and Mappings: Two Kinds of Conceptual Models for Interactive Devices. In Gentner, D. and Stevens, A.L. (Eds.), *Mental Models*. Lawrence Erlbaum Associates Inc., pp 35-52.

An examination of the ERP correlates of recognition memory using state-trace analysis.

Emily Freeman (emily.e.freeman@gmail.com)
School of Psychology, The University of Newcastle, NSW, 2308, Australia
Simon Dennis (simon.dennis@gmail.com)
Department of Psychology, The Ohio State University, OH, 43210, USA
John Dunn (john.c.dunn@adelaide.edu.au)
School of Psychology, The University of Adelaide, SA, 5005, Australia

Abstract
There has been much debate in recent years as to whether recognition memory is best described using a single or dual process model. State-trace analysis provides an atheoretical approach to determining the number of underlying psychological variables, or processes, that mediate the effect of one or more independent variables on the measured dependent variables. Recently, state-trace analysis has shown strong support for a single process interpretation of the behavioral results from recognition memory experiments. In this paper, we demonstrate, using state-trace analysis, that both the behavioral and electrophysiological results from recognition memory experiments are also supportive of a single process interpretation.
Keywords: recognition memory; event-related potentials; single process models; dual process models

The study of recognition memory aims to determine the process(es) underlying how one recognizes something, or someone, as having been previously encountered (Mandler, 1980). In a typical recognition experiment, participants study a list of items, and at test are asked to discriminate between both studied (old) and unstudied (new) items. Two measures are obtained: the hit rate (proportion of old items correctly identified as being old) and the false alarm rate (the proportion of new items incorrectly identified as being old). The hit and false alarm rates can be combined to indicate an overall level of accuracy[1].

A number of mathematical models have been proposed attempting to describe the basis of recognition memory. These models can be grouped into two main frameworks: single and dual process models. This paper will attempt to assess the validity of these two classes of models by testing their basic assumptions using electrophysiological data from a recognition memory experiment. First these two frameworks will be described as well as some of the supporting behavioral, imaging and electrophysiological evidence. Next, an atheoretical method that can be used to test the basic assumptions of these two classes of models will be described. Following which, the results from an experiment, designed to test these underlying assumptions are presented.

[1] For example, d' is calculated by subtracting the z-transformed false alarm rate from the z-transformed hit rate.

Models of Recognition Memory

It has long been debated whether recognition memory decisions are performed on the basis of a single memory process, referred to as either strength, familiarity, or matching, or whether a recall-like component is also involved (Clark, 1999). The first dual process models were developed in the 1970s (e.g., Atkinson & Juola, 1974), but were overtaken in popularity when single process, global memory/matching models, were developed in the 1980s. Dual process models regained popularity in the early 1990s and as such the debate as to which type of model best describes memory is ongoing.

Single Process Models

Single process theories are based on the signal detection framework (Green & Swets, 1967). In its simplest form, signal detection theory considers two basic aspects of detection: the underlying representations, which are interpreted as psychological distributions, and a decision aspect, which involves the use of decision criteria to arrive at a response (DeCarlo, 2002). Signal detection theory can be applied in any task in which participants are required to discriminate between two or more classes of stimuli (Stanislaw & Todorov, 1999).

Signal detection memory models assume that when a participant is presented with a test stimulus it is directly matched to multiple memory representations in parallel and the fit of these matches is used to calculate a familiarity value (Clark, 1999). Familiarity is thought to be based on associative information and information about other items in memory, as well as on stored item-specific information about the test item. In a recognition memory experiment, stimuli presented in the study phase have familiarity values drawn from the 'old' normal distribution, while the familiarity values for new items are drawn from the 'new' normal distribution. The mean of the old distribution is assumed to be higher than the mean for the

new distribution. Each old and new condition has its own response distribution and the criterion is placed at a point chosen by the participant that determines whether an old or new response is made.

There are a number of specific single process theories that have been developed to account for findings in recognition memory. Although each of these models is considered to contain a single process, they vary quite substantially in their focus. For example, Attention-Likelihood Theory (ALT, Glanzer & Adams, 1990) is based on the idea of feature marking first proposed by Glanzer and Bowles (1976). Retrieving Effectively from Memory (REM, Shiffrin & Steyvers, 1997) is centered around item noise, while at the other end of the spectrum, the Bind Cue Decide Model of Episodic Memory (BCDMEM, Dennis & Humphreys, 2001) is focused on context noise.

Dual Process Models

Dual process models assume that recognition is based on two memory processes: familiarity and recollection, which are assumed to make independent contributions to recognition (Clark, 1999). Familiarity is assumed to be a fast process and is equivalent to the signal detection process described by single process theories. On the other hand, recollection is assumed to be a slow, deliberate, and relatively accurate search process whereby information about the study episode is retrieved (Arndt & Reder, 2002; Yonelinas, 1999). Generally, dual process theorists propose that the hit rate in a recognition experiment is driven by recollection and the false alarm rate is driven by familiarity (e.g., Joordens & Hockley, 2000).

Yonelinas (2002) presented a high-threshold dual process model of recognition memory. He proposed that recollection and familiarity are independent parallel processes that differ in the type of information they provide and the extent to which they influence a person's confidence. Familiarity reflects the assessment of quantitative memory strength information in the same manner as signal detection theory used in single process theories. The variable strength of familiarity leads to a wide range of confidence ratings. Recollection reflects a threshold retrieval process in which qualitative information about a previous event is retrieved, producing a high level of confidence.

A number of pieces of evidence have been put forward in support of the dual process models of recognition memory. The most dominant of these behavioral, imaging and electrophysiological findings are presented in the following section.

Behavioral, Imaging and Electrophysiological Evidence

The Remember-Know paradigm (Tulving, 1985) has been used to add support to the claim that recognition memory is best described using a dual process model. This procedure requires participants to indicate whether their 'old' responses in a recognition memory test are based upon familiarity alone (Know) or whether they recollect seeing the item in the study list (Remember). Some researchers (e.g., Gardiner & Java, 1990) have suggested that the mere finding that participants are able to distinguish between these two types of responses is evidence that both familiarity and recollection contribute to the recognition memory task. However, experiments finding dissociations between remember and know responses provide much more compelling arguments. For example, Gardiner (1988) reported a dissociation between remember and know responses such that deeper levels of processing at study led to more remember responses at test, but did not affect know responses. Since this early finding, numerous studies have been reported finding dissociations between remember and know responses (e.g., Gardiner & Java, 1990, 1991; Glanc & Greene, 2007; Joordens & Hockley, 2000; Park, Reder, & Dickison, 2005; Rajaram, 1993).

Although these dissociations between remember and know responses are often taken as evidence for dual process models (e.g., Jacoby, Yonelinas, & Jennings, 1997), a number of single process advocates have argued that remember and know responses are simply classifications of different levels of confidence, and as such can also be accounted for by single process models (e.g., Donaldson, 1996). Dunn (2004) put forward a compelling argument for remember and know responses representing higher and lower levels of confidence, respectively. In an analysis of 72 studies, Dunn showed that the arguments against remember-know data being described by a signal detection, single process framework could not be ruled out, and provided an equally plausible account of the data.

Since it appears that behavioral data can be well explained using single process models, researchers have recently started looking at the neurological basis of recognition memory, in order to determine if there is any biological evidence for familiarity and recollection playing a role in the decision process. Despite evidence that the remember-know procedure does not necessarily separate recollection and familiarity, it has been widely used in imaging and electrophysiological studies. Here the aim is to find either separate brain regions (in fMRI studies), or distinct event-related potentials (ERPs) related to remember and know responses, which are then interpreted as being related to recollection and familiarity, respectively.

Recently, Yonelinas, Otten, Shaw and Rugg (2005) suggested that they had found a neural signature of recollection that was distinct from familiarity. Because past researchers (e.g., Dunn, 2004) had suggested that remember responses simply reflect a subject's high level of confidence, Yonelinas et al. had their subjects respond 'remember' if they could remember something specific about the study episode, otherwise they were asked to give

a confidence rating that the item was studied using a four-point scale (sure old / sure new). Yonelinas et al. found different neural signatures for remember and high confidence familiar responses, which led them to the conclusion that recollection and familiarity are two distinct processes (but see Dunn & Dennis, submitted, for a conflicting interpretation of these results).

Curran (1999, 2004) and colleagues (e.g., Curran & Dien, 2003; Curran, DeBuse, Woroch, & Hirshman, 2006; Curran, Tepe, & Piatt, 2006; Curran, DeBuse, & Leynes, 2007) have focused on differentiating recollection and familiarity using ERPs. Two time periods of interest have been identified. The first, occurring 300-500ms after stimulus onset is commonly referred to as the FN400 as it is a frontal negative peak. The second, occurring 400-800ms after stimulus onset has received numerous names, but the most common is the LPC, or late positive component, and is more dominant in the parietal brain region. Curran et al. have argued that the FN400 is an old/new decision component related to item familiarity, while the LPC is related to the recollection process. Evidence for this distinction also comes from studies using the remember-know procedure. Studies have shown that studied items produce a more negative FN400 than unstudied items, and that 'remembered' items produce a more positive LPC than 'known' items (e.g., Rugg et al., 1998; Rugg & Curran, 2007; Rugg & Yonelinas, 2003). However, as Finnigan, Humphreys, Dennis, and Geffen (2002) have demonstrated, these findings can be easily fit by a single process model whereby the FN400 reflects an individual's old/new decision, and the LPC reflects their confidence.

Obviously there is much controversy as to how both the behavioral and neurological data should be interpreted. The following section outlines a technique that can be used to determine the number of processes that are needed to account for a given data set, without making any assumptions about single or dual process models.

State-Trace Analysis

State-trace analysis (Bamber, 1979) is based on the premise that two dependent variables will covary with each other to the extent that they are affected by the same independent variable. By producing a plot of one dependent variable as a function of another dependent variable, one can determine the number of intervening psychological variables, or processes, that mediate the effect of one or more independent variables on the measured dependent variables. If the resulting scatter plot is one dimensional, that is all the data points lie on a single monotonically increasing (or decreasing) curve, then it can be assumed that the two dependent variables are functions of the same latent variable.

Dunn (2008) performed a state-trace analysis on the data from 37 remember-know studies. When the old/new hit rate was plotted as a function of the remember (or high confidence) hit rate, a predominately one dimensional curve was found, suggesting that the remember-know task is best described by a single process model. Further, when the z-transform of the state-trace was computed, a straight line with a slope of one was obtained. This finding is also in accordance with an unequal variance, signal detection, or single process model.

Experiment

The aim of the present research is to examine the ERP correlates of recognition memory. To do this, state-trace analyses will be applied to behavioral and ERP data obtained from an experiment that manipulates two independent variables identified by Yonelinas (2002) to affect either familiarity or recollection. The behavioral state-trace will plot the low confidence hit rate (LCHR) as a function of the high confidence hit rate (HCHR) and the ERP state-trace will plot the FN400 as a function of the LPC. If the HCHR/LPC reflects recollection, the state-trace plots should show two lines, separated on the dimension outlined by Yonelinas to reflect recollection. However, if the state-trace plots show a one dimensional curve, this will be indicative of a single process underlying recognition memory, and will provide strong evidence in favor of single process models.

Specifically, in our experiment the number of study repetitions (1/2/4) and attention at study (focused/divided) were manipulated. According to Yonelinas (2002), the attention manipulation should affect recollection, but not familiarity, and the study repetition manipulation should affect both familiarity and recollection. If the dual process interpretation of recognition memory is accurate, the state-trace plots should show two monotonic functions, separated by the attention manipulation. Specifically, the LCHR/FN400 should become more positive as the number of study repetitions increases, and the focused attention condition should be shifted to the right (i.e., a more positive HCHR/LPC) compared to the divided attention condition. However, if the resulting state-trace plot is one-dimensional, this will indicate that both the number of study repetitions, and attention at study are related to the same latent variable, or memory process, indicating that a single process interpretation of the data is accurate.

Method

Participants

54 students from the Ohio State University participated in return for course credit.

Stimuli

The stimuli consisted of 240 high frequency words with a mean frequency of 155 (ratings taken from the Celex

database, Baayen, Piepenbrock, & van Rijn, 1993). Words were 4-8 letters in length (mean 4.6). Words were randomly divided into 5 lists for each participant and items within each list were randomly allocated to old/new, focused/divided attention, and repetition conditions.

Design

The experiment was a 2x3 design, with attention at study (focused/divided) and number of study repetitions (1/2/4) manipulated within-subjects.

Procedure

Participants were first briefed on the requirements of the study, signed a consent form and were fitted with the Geodesic Electrode Net.

Each study list consisted of 24 words. Half the words were presented alone on the screen (focused attention condition) while the other half were presented flanked by two numbers (divided attention condition). In the divided attention condition, the flankers appeared for 200ms and were then covered by a mask. The numbers differed in both their numerical value, and their font size. After the target word was removed from the screen, the participants were asked to report which number (left or right) was larger in either value or size by pressing the appropriate key on the keyboard. One third of the study items were presented once, one third were presented twice, and one third were presented four times during the study phase to give a total of 56 study trials. Repeated words were always repeated within the same attention condition. Words were presented for three seconds followed by a one second interstimulus interval (isi). Following the study phase, participants completed several math problems for a period of approximately three minutes.

The test lists consisted of 48 words, with an equal number of old and new items. Each word was presented for two seconds followed by a response cue, at which time the participant was required to give their response by pressing the appropriate key on the keyboard using a six-point confidence rating scale (sure old/sure new). Participants were instructed to wait for the cue before responding, to stay as still as possible, and to minimize eye blinks.

Each study/test cycle took approximately 20 minutes to complete. After each cycle, the Electrode Net was checked to ensure that impedances remained below 50kΩ. Participants completed as many study/test cycles as possible during the two hour time period, with most completing an average of four cycles.

EEG Recording

Scalp voltages were collected using a 128-channel Electrical Geodesics Sensor Net connected to a high impedance amplifier (300kΩ Net Amps™, Electrical Geodesics Inc, Eugene, OR, USA). Amplified analog voltages (0.1-100Hz bandpass, -3dB) were digitized at 500Hz. Individual sensors were adjusted until each reached an impedance of less than 50kΩ. The EEG was digitally low-pass filtered at 40Hz.

Results

Trials were discarded from the analysis if they contained eye movements (EOG over 70μV), or more than 20% of channels were bad (average amplitude over 200μV or transit amplitude over 100ms). Individual bad channels were replaced on a trial-by-trial basis with a spherical spline algorithm (Srinivasan, Nunez, Silberstein, Tucker, & Cadusch, 1996). Consistently bad channels for a given subject were replaced throughout that subject's entire dataset (bad channels per subject: median = mode = 1, range = 0 - 3). EEG was measured with respect to a vertex reference (Cz), but an average-reference transformation was used to minimize the effects of reference-site activity and accurately estimate the scalp topography of the measured electrical fields (Dien, 1998; Picton, Lins, & Scherg, 1995). Average-reference ERPs were computed for each channel as the voltage difference between that channel and the average of all channels. The average reference was corrected for the polar average reference effect (Junghofer, Elbert, Tucker, & Braun, 1999). ERPs were baseline-corrected with respect to a 100ms prestimulus recording interval.

Figure one shows the state-trace plot obtained by plotting the mean LPC against the mean FN400 for each attention by repetition condition. Both the FN400 and the LPC were found to increase (i.e., become more positive) with increasing study repetitions. Additionally, there is no significant differentiation between the focused and divided attention conditions.

The behavioral state-trace plot, also shown in Figure one, shows that both high and low confidence hit rates increase with increasing study repetitions. Further, there is no differentiation between the attention conditions.

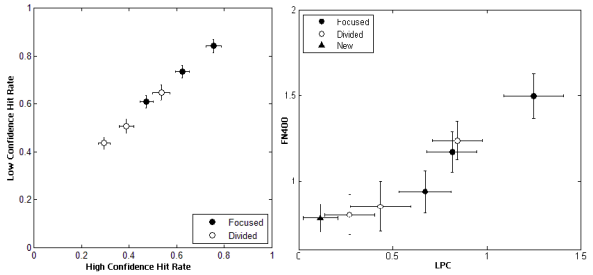


Figure 1: State-trace plots of the behavioral (left) and ERP (right) results for each of the attention by repetition conditions.

Discussion

The state-trace plots produced from the analysis of our experiment are clearly one-dimensional and thus provide very little evidence in support of the dual process interpretation of recognition memory. Rather, our findings

show strong support for a single process interpretation of recognition memory.

Numerous previous recognition memory studies looking at ERPs have assumed that the LPC is reflective of recollection (e.g. Curran, 2004; Curran, Tepe, & Piatt, 2006). These studies have rejected a single process interpretation of the FN400 and LPC because it "does not explain the double-dissociation between mid-frontal and parietal effects observed by Woodruff et al." (Rugg & Curran, 2007, p.264). In the study which Rugg and Curran (2007) refer to, participants we asked to respond using a variation of the remember-know procedure, in which they either made a graded, confidence-based familiarity judgment on a 4-point scale, or a remember/recollection response. The authors report differing ERP patterns for the FN400 and the LPC and suggest that the ordering of the waveforms (1<2<3<4=R and 1=2=3=4<R, respectively) are evidence that the FN400 represents familiarity, and the LPC represents recollection. However, as explained by Dunn and Kirsner (2003), this is actually a classic non-double-association. The error is that the authors implicitly assume that changes in volts (a physical variable) are linearly related to memory strength (a psychological variable). If on the other hand, one assumes that this relationship is at best monotonic and different for frontal and parietal, the underlying assumption of state-trace analysis, then there is no dissociation. At both sites, the underlying pattern of memory strength is 1<2<3<4<R but mapped onto frontal and parietal volts by different functions. By using state-trace analysis to interpret our research findings, we have not only avoided this common error, but have also shown strong support for a single, rather than dual, process interpretation of ERP results.

Additionally, the results from our experiment add weight to the suggestion by Finnigan et al. (2002) that the LPC is not reflective of recollection. Finnigan et al. suggested that the LPC may instead be related to confidence. Although not specifically addressed in this analysis, our research methods provided an opportunity for testing this idea in the future.

Yonelinas (2002) suggested that dividing attention at study would affect recollection at test, such that items in the focused attention condition would have higher levels of recollection than items in the divided attention condition. The analysis of our behavioral data produced a single monotonic state-trace curve (see also Dunn, Heathcote, Dennis, & deZubicary, in preparation). Our ERP findings extend these behavioral findings, supporting a single process interpretation of recognition memory.

Combined, these findings suggest that not only is the LPC not reflective of recollection as suggested by Curran and colleagues, but they also suggest that the notion of recollection itself may be flawed, further supporting the predictions of single process models of recognition memory.

Acknowledgements

The authors would like to acknowledge Nayef Ahmar for his assistance with the analysis. The project was funded by an Australian Research Council Grant DP0558407.

References

Arndt, J., & Reder, L.M. (2002). Word frequency and receiver operating characteristic curves in recognition memory: Evidence for a dual-process interpretation. *Journal of Experimental Psychology, 28(5)*, 830-842.

Atkinson, R.C., & Juola, J.F. (1974). Search and decision processes in recognition memory. In D.H. Krantz, R.C. Atkinson, R.D. Luce, & P. Suppes (Eds.), *Contemporary developments in mathematical psychology: Vol. 1. Learning, memory & thinking* (pp. 242-293). San Francisco: Freeman.

Baayen, R.H., Piepenbrock, R. & van Rijn, H. (1993). The CELEX lexical database [CD-ROM]. Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania.

Clark, S.E. (1999). Recalling to recognize and recognizing to recall. In C. Izawa (Ed.) *On Human Memory: Evolution, Progress, and reflections on the 30th Anniversary of the Atkinson-Shiffrin Model* (pp. 215-244). Mahwah, NJ: Earlbaum.

Curran, T. (1999). The electrophysiology of incidental and intentional retrieval: ERP old/new effects in lexical decision and recognition memory. *Neuropsychologia, 37*, 771-785.

Curran, T. (2004). Effects of attention and confidence on the hypothesized ERP correlates of recollection and familiarity. *Neuropsychologia, 42*, 1088-1106.

Curran, T., DeBuse, C., & Leynes, P.A. (2007). Conflict and criterion setting in recognition memory. *Journal of Experimental Psychology: Learning, Memory & Cognition, 33(1)*, 2-17.

Curran, T., DeBuse, C., Woroch, B., & Hirshman, E. (2006). Combined pharmacological and electrophysiological dissociation of familiarity and recollection. *The Journal of Neuroscience, 26(7)*, 1979-1985.

Curran, T., & Dien, J. (2003). Differentiating amodal familiarity from modality-specific memory processes: An ERP study. *Psychophysiology, 40*, 979-988.

Curran, T., Tepe, K.L., & Piatt, C. (2006). Event-related potential explorations of dual processes in recognition memory. In H.D. Zimmer, A. Mecklinger, & U. Linderberger (Eds.), *Binding in Human Memory: A Neurocognitive Approach* (pp. 467-492), Oxford: Oxford University Press.

DeCarlo, L.T. (2002). Signal detection theory with finite mixture distributions: Theoretical developments with applications to recognition memory. *Psychological Review, 109*, 710-721.

Dennis, S. & Humphreys, M.S. (2001). A context noise model of episodic word recognition. *Psychological Review, 108(2)*, 452-478.

Dien, J. (1998) Issues in the application of the average reference: Review, critiques and recommendations. *Behavior Research Methods Instruments & Computers, 30*, 34-43.

Donaldson, W. (1996). The role of decision processes in remembering and knowing. *Memory & Cognition, 24*, 523-533.

Dunn, J.C. (2004). Remember-Know: A matter of confidence. *Psychological Review, 95*, 91-101.

Dunn, J.C. (2008). The dimensionality of the remember-know task: A state-trace analysis. *Psychological Review, 115(2)*, 426-446.

Dunn, J.C., & Dennis, S. (submitted). Separating the brain regions involved in recollection and familiarity in recognition memory: A comment on Yonelinas, Otten, Shaw & Rugg (2005).

Dunn, J.C., Heathcote, A., Dennis, S., & deZubicary, G. (in preparation). Single and dual-process models of recognition memory: A state-trace analysis.

Dunn, J. C. & Kirsner, K. (2003). What can we infer from double dissociations? *Cortex, 39*, 1-7.

Finnigan, S., Humphreys, M.S., Dennis, S., & Geffen, G. (2002). ERP 'old/new' effects: Memory strength and decisional factor(s). *Neuropsychologia, 40*, 2288-2304.

Gardiner, J.M. (1988). Functional aspects of recollective experience. *Memory & Cognition, 16*, 309-313.

Gardiner, J.M., & Java, R.I. (1990). Recollective experience in word and nonword recognition. *Memory & Cognition, 18(1)*, 23-30.

Gardiner, J.M., & Java, R.I. (1991). Forgetting in recognition memory with and without recollective experience. *Memory & Cognition, 19(6)*, 617-623.

Glanc, G.A., & Greene, R.L. (2007). Orthographic neighborhood size effects in recognition memory. *Memory & Cognition, 35(2)*, 365-371.

Glanzer, M., & Adams, J.K. (1990). The mirror effect in recognition memory: Data and theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16(1)*, 5-16.

Glanzer, M., & Bowles, N. (1976). Analysis of the word-frequency effect in recognition memory. *Journal of Experimental Psychology: Human Learning and Memory, 2(1)*, 21-31.

Green, D.M., & Swets, J.A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Jacoby, L.L., Yonelinas, A.P., & Jennings, J.M. (1997). The relationship between conscious and unconscious (automatic) influences: A declaration of independence. In J. Cohen & J.W. Schooler (Eds.), *Scientific approaches to the question of consciousness* (pp. 13-47). Hillsdale, NJ: Erlbaum.

Joordens, S., & Hockley, W.E. (2000). Recollection and familiarity through the looking glass: When old does not mirror new. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26(6)*, 1534-1555.

Junghofer, M., Elbert, D., Tucker, T., & Braun, C. (1999). The polar effect of average reference: A bias in estimating the head surface integral in EEG recording. *Electroencephalography and Clinical Neurophysiology, 110*, 1149-1155.

Mandler, G. (1980). Recognizing: The judgment of previous occurrence. *Psychological Review, 87*, 252-271.

Park, H., Reder, L.M., & Dickison, D. (2005). The effects of word frequency and similarity on recognition judgments: The role of recollection. *Journal of Experimental Psychology: Learning, Memory & Cognition, 31(3)*, 567-578.

Rajaram, S. (1993). Remembering and knowing: Two means of access to the personal past. *Memory & Cognition, 21(1)*, 89-102.

Rugg, M.D., & Curran, T. (2007). Event-related potentials and recognition memory. *Trends in Cognitive Sciences, 11(6)*, 251-257.

Rugg, M.D., Mark, R.E., Walla, P., Schloerscheidt, A.M., Birch, C.S., & Allan, K. (1998). Dissociation of the neural correlates of implicit and explicit memory. *Nature, 392*, 595-598.

Rugg, M.D., & Yonelinas, A.P. (2003). Human recognition memory: A cognitive neuroscience perpective. *Trends in Cognitive Sciences, 7(7)*, 313-319.

Shiffrin, R.M. & Steyvers, M. (1997). A model for recognition memory: REM – retrieving effectively from memory. *Psychonomic Bulletin & Review, 4*, 145-166.

Snodgrass, J.G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General, 117(1)*, 34-50.

Srinivasan, R., Nunez, P.L., Silberstein, R.B., Tucker, D.M., & Cadusch, P.J. (1996). Spatial sampling and filtering of EEG with spline-Laplacians to estimate cortical potentials. *Brain Topography, 8*, 355–366.

Stanislaw, H. & Todorov, N. (1999). Calculation of Signal Detection Theory Measures. *Behaviour Research Methods, Instruments and Computers, 31*, 137-149.

Tulving, E. (1985). Memory and Consciousness. *Canadian Psychology / Psychologie Canadienne, 26(1)*, 1-12.

Yonelinas, A.P. (1999). The contribution of recollection and familiarity to recognition and source-memory judgments: A formal dual-process model and an analysis of receiver operating characteristics. *Journal of Experimental Psychology: Learning, Memory, and Language, 25(2)*, 1415-1434.

Yonelinas, A.P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language, 46*, 441-517.

Yonelinas, A.P., Otten, L.J., Shaw, K.N., & Rugg, M.D. (2005). Separating brain regions involved in recollection and familiarity in recognition memory. *The Journal of Neuroscience, 25(11)*, 3002-3008.

Modeling Change in Recognition Bias with the Progression of Alzheimer's

James P. Pooley (jpooley@uci.edu)
Michael D. Lee (mdlee@uci.edu)
Department of Cognitive Sciences, 3151 Social Science Plaza
University of California, Irvine, CA, 92697-5100

William R. Shankle (rshankle@mccare.com)
Medical Care Corporation, 19782 MacArthur Boulevard
Irvine, CA 92612

Abstract

One of the key memory tests in the clinical assessment and diagnosis of Alzheimer's Disease (AD) is the recognition memory task. Models developed in cognitive psychology have previously been applied to help understand clinical data. In particular, Signal Detection Theory (SDT) models have been used, to separate people's memory capabilities from their decision-making strategies. An important finding in this literature is that people with AD change their decision strategy in response to memory impairment, applying a more liberal criterion than people without AD. In this paper, we analyze clinical data that measures the progression of AD in a detailed way, using a theoretically motivated version of SDT, and applying hierarchical Bayesian methods to model individual differences. Our results corroborate many of the previous findings, but provide a more detailed focus on recognition performance with AD progression.

Keywords: Alzheimer's disease; Cognitive psychometrics; Hierarchical Bayesian modeling; Human recognition memory; Signal detection theory

Introduction

The clinical assessment and diagnosis of Alzhiemer's disease (AD) routinely involves the administration of memory tests that are familiar to cognitive scientists who study human memory. In particular, recognition, immediate free recall, and delayed free recall are large subcomponents of assessment tools such as the MCIS and the ADAS-Cog (e.g., Morris, Heyman, & Mohs, 1989). This link means there is an important role for theories and models of memory, as developed in the cognitive sciences (for an overview, see Norman Detre, & Polyn, 2008), in helping understand AD. In particular, memory models can provide quantitative measurement tools that allow for patient behavior to be interpreted in terms of psychologically meaningful latent parameters (e.g., Riefer, Knapp, Batchelder, Bamber, & Manifold, 2002).

A good example of the potential for applying memory models to clinical data is provided by a literature that uses equal-variance Signal Detection Theory (SDT) models (e.g., MacMillan & Creelman, 2004). SDT is widely-used as a basic model of the recognition memory task, and has the theoretical attraction of separating memory capabilities from decision processes when explaining people's behavior (e.g., Budson Wolk, Chong, & Waring, 2006; Snodgrass & Corwin, 1988). This is a very important capability, because there is considerable evidence that AD patients do have different decision-making strategies in tasks like recognition memory.

The recent review by Budson et al. (2006) notes that the application of SDT models to clinical data has repeatedly shown that patients with AD use a more liberal criterion in identifying previously studied words. This strategy is usually interpreted as a response to awareness of diminishing memory capabilities. Additionally, Budson et al. (2006) report the results of an experiment which addressed several potential confounds in the existing experiments, including unequal numbers of old and new words and semantic and/or perceptual relatedness of the old and new words. Again, AD patients were found to have abnormally liberal response biases compared to non-AD patients.

In this paper, we extend the application of SDT models to clinical recognition memory data. We do this in a number of ways. First, we use a large new clinical database, which has the advantage of measuring the progression of AD in some detail. This lets us conduct a finer-grained analysis of how recognition memory changes as AD progresses. Second, we use a simple variant of the standard SDT model that builds in an unequal-variance assumption. This is theoretically preferable, given empirical evidence that there is more variability in people's memory for studied than non-studied words. Third, we embed our SDT analyses with a hierarchical Bayesian framework for statistical inference. This lets us provide a coherent model-based account of variation, at both the level of individual patients, and the level of clinical sub-populations.

The plan of the paper is as follows. We begin by describing the clinical data, and then the unequal-variance SDT model we use. We show that the model provides a good account of the data, and show how inference about the model's parameters gives an interpretable account of changes in recognition memory with the progression of AD. We then extend the modeling to account explicitly for changes in decision bias, and conclude by discussing how our findings relate to the existing literature.

Clinical Data

Our data come from two neurology clinics where 1350 patients completed a standard old/new recognition mem-

ory test. The patient was shown a study list of 10 words to memorize, and was then tested on their ability to recognize the 10 studied *old* words from 10 unstudied *new* words. This means there are 20 test trials, on each of which the patient was shown a word and simply asked to decide whether or not the word was on the study list. Consequently, the patient's behavior on each trial naturally falls into one of the standard SDT classes of hits, misses, false alarms, and correct rejections. The words themselves were selected from the CERAD (Consortium to Establish a Registry for Alzheimer's Disease) word list (Shankle, Mangrola, Chan, & Hara, 2009).

Independent of patient performance on the recognition memory tests, a trained neurologist used the Functional Assessment Staging Test (FAST) to assess the severity of each patient's AD. The FAST (Reisberg, 1988) is a well-validated diagnostic tool used by clinicians to classify patients into one of the seven *stages* of AD, each of which corresponds to a level of functional impairment. Specifically, stage 1 corresponds to 'normal aging', stage 2 to 'possible mild cognitive impairment', stage 3 to 'mild cognitive impairment', stage 4 to 'mild dementia', stage 5 to 'moderate dementia', stage 6 to 'moderately severe dementia' and stage 7 to 'severe dementia'. We focus on only FAST Stages 1–5, because patients diagnosed into Stages 6 and 7 have very limited functional capabilities, and cannot necessarily understand and complete memory tasks. In our sample of 1350 patients, 288 were classified as Stage 1, 308 as Stage 2, 129 as Stage 3, 436 as Stage 4, and 189 at Stage 5.

Hierarchical SDT Model

In this section, we describe the hierarchical SDT model we use to analyze the clinical data. We start with a standard SDT model, and then describe how our hierarchical extensions add the capability to model individual differences and changes in bias. We then implement the model as a graphical model to allow Bayesian inference.

Signal Detection Theory

The basic SDT model shown in Figure 1 assumes that, on each trial, the presented word evokes some memory strength. The memory strengths of both old and new words are assumed to have Gaussian distributions, with the mean of the new distribution separated from the mean of the old distribution by a distance $d' > 0$. In this way, d' measures the *discriminability* of the old from the new words, and so represents the acuity of memory for the words.

Due to the assumed overlap of the old and new distributions, an individual needs a decision strategy for relating memory strength to responses in a recognition test. SDT models assume this is done using a criterion level of memory strength k below which the individual will respond studied and above which the individual will respond non-studied. The area h under the old distribution above the criterion corresponds to the hit rate, and the area f under the new distribution above the criterion corresponds to the false-alarm rate.

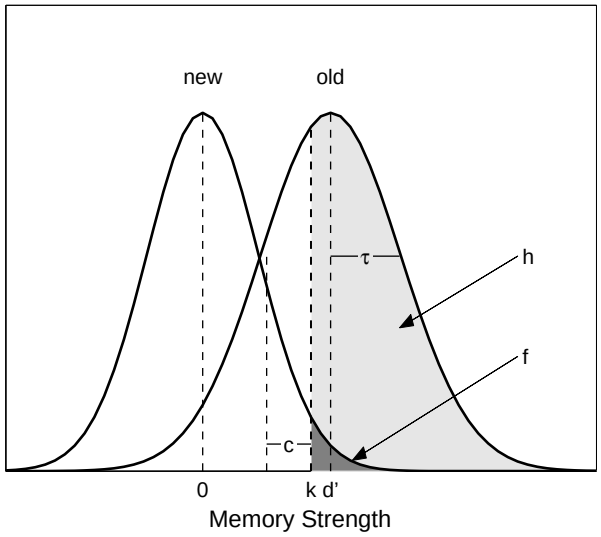


Figure 1: The unequal-variance SDT model and parameters.

The distance c between this criterion and unbiased responding is commonly used as a measure of *response bias* due to its purported independence from d' (Snodgrass & Corwin, 1988). The response bias measures the tendency of an individual to give one response rather than another.

Extension for Unequal Variance

Most SDT modeling in psychology assumes that the standard deviations of the old and new distributions are equal, with $\sigma_{\text{old}} = \sigma_{\text{new}} = 1$ for convenience. Results of recognition memory experiments (e.g., Mickes, Wixted, & Wais, 2007), however, support a version of SDT in which the standard deviation of the old distribution is 25% larger than the standard deviation of the new distribution, so that $\sigma_{\text{new}}/\sigma_{\text{old}} = 0.8$. This finding is usually interpreted as coming from variability in the encoding of studied words. Our SDT model adopts an unequal-variance assumption, using the approach developed by Dennis, Lee, & Kinnell (2008).

Extension for Individual Differences

Most previous applications of SDT models to the recognition memory data of Alzheimer's patients have also ignored the issue of individual differences. To address this shortcoming, we apply hierarchical methods to extend the standard SDT model (e.g., Dennis, Lee, & Kinnell, 2008; Rouder & Lu, 2005). The basic idea is to introduce sub-populations at a group-level that allow for different parameter values for different levels of severity in AD. An individual patient's discriminability and response bias parameters are then drawn from the appropriate group-level distribution for their level of severity. In this way, the model allows freedom for different individuals to have different parameters, but still maintain a

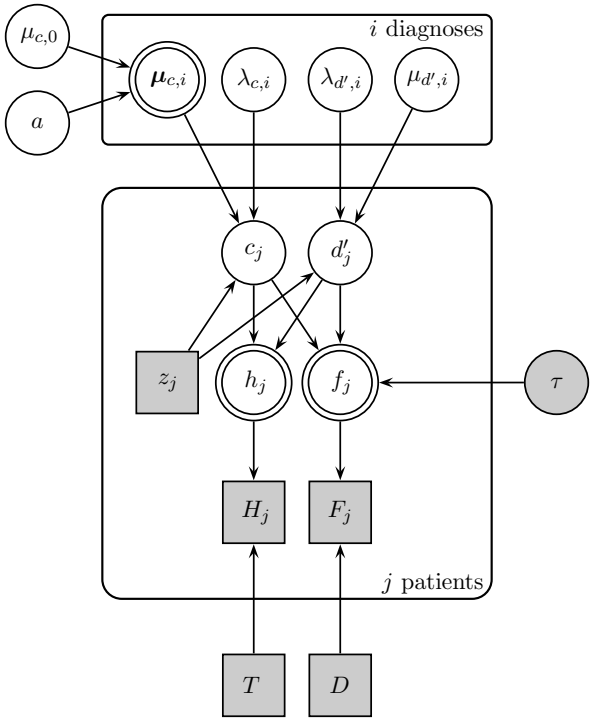


Figure 2: Graphical model implementation of the hierarchical SDT model.

similarity to other patients with a similar level of cognitive impairment.

Extension for Modeling Change

Most previous analyses focusing on changes in response bias with AD progression have taken a purely statistical approach. Typically, they have tested for significant differences in bias or criterion parameters, as inferred separately from AD and non-AD patients. We adopt a different approach based on cognitive modeling, building assumptions about how bias changes into the model itself. This is consistent with the basic idea of *generative* models, which try to provide formal accounts of how latent parameters produce and co-vary with observed behavior, and can be contrasted with the *discriminative* philosophy of post-hoc statistical tests. In the generative approach, a model of change is incorporated into the SDT model, with the goal of providing a complete and integrated account of how the criterion changes with the progression of AD.

Graphical Model Implementation

We implemented our hierarchical SDT model in the form of a Bayesian graphical model, a formalism widely used statistics and computer science (e.g., Jordan, 2004). In graphical models, nodes correspond to variables, and their interdependencies show the causal relationships between the variables. In particular, graphical models show how unobserved variables (i.e., parameters) generate observed variables (i.e., data). Details and tutorials are aimed at cognitive scientists are provided by Lee (2008) and Shiffrin, Lee, Kim, and Wagenmakers (2008). The practical advantage of graphical models is that sophisticated and relatively general-purpose Markov Chain Monte Carlo (MCMC) algorithms exist that can sample from the full joint posterior distribution of the parameters conditional on the observed data.

It is easiest to understand the graphical model in Figure 2 by starting with the d'_j and c_j nodes, which are the discriminability and bias parameters for the jth patient. These parameters can be used to generate the hit and false-alarm rates for that patient, according to the SDT model. The hit rate is $h_j = \Phi(d'_j/2 - c_j)$ and the false alarm rate is $f_j = \Phi(-(d'_j/2 + c_j)/\tau)$, where $\tau = 0.8$ gives the unequal-variance model advocated by Mickes, Wixted, and Wais (2007). Based on these hit and false alarm rates and the $O = 10$ old and $N = 10$ new words presented to all patients during the recognition tests, the jth patient produces $H_j \sim \text{Binomial}(h_j, T)$ hits and $F_j \sim \text{Binomial}(f_j, D)$ false-alarms.

The distributions of discriminability and bias for different AD diagnoses, at the group or sub-population level, are controlled by the mean μ and precision λ variables. There is a Gaussian group distribution for each group. If, for example, we use FAST stage diagnoses to define groups, and the jth patient belongs to stage z_j, then $d'_j \sim \text{Gaussian}(\mu_{d',z_j}, \lambda_{d',z_j})$ and $c_j \sim \text{Gaussian}(\mu_{c,z_j}, \lambda_{c,z_j})$.

Finally, the graphical model in Figure 2 implements a basic model of change for response bias. Following previous analyses (e.g., Snodgrass & Corwin, 1988), we just consider the change from non-AD to AD patients. The parameter $\mu_{c,0}$ measures the non-AD response bias, and a quantifies the change, so that $\mu_{c,1}, \mu_{c,2} = \mu_{c,0}$ and $\mu_{c,3}, \ldots, \mu_{c,5} = \mu_{c,0} + a$.

Modeling Results

In order to perform Bayesian inference, we implemented the graphical models in WinBUGS (Spiegelhalter, Thomas, & Best, 2004. This software uses a range of MCMC computational methods to obtain samples from the posterior distributions of the relevant parameters (e.g., Mackay, 2003). All of our analyses are based on 10,000 posterior samples collected following a burn-in of 1000 samples, using multiple chains to check convergence.

Assessing Model Fit

Posterior predictive distributions provide an intuitive and principled to assessing the descriptive adequacy of a Bayesian model (Gelman, Carlin, Stern, & Rubin, 2004, pp. 165–172). A posterior prediction corresponds to the data the model expects, based on the parameter values it has inferred, and naturally takes into account uncertainty in those parameter estimates.

Figure 3 shows a posterior predictive analysis for the

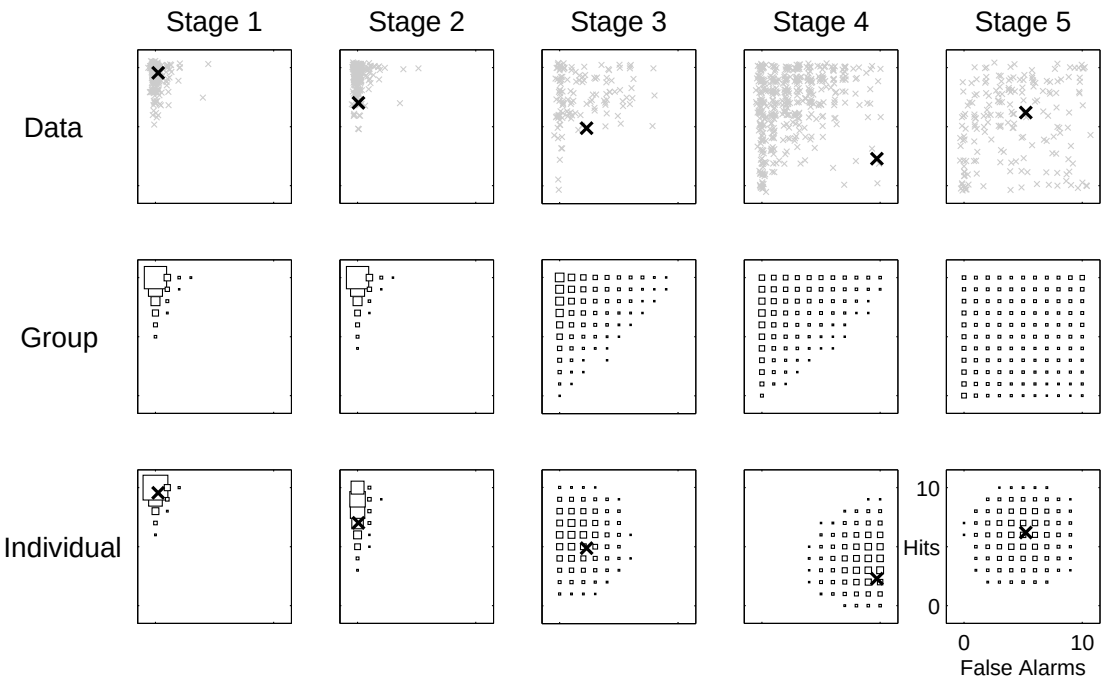


Figure 3: Posterior predictive assessment of the fit of the hierarchical SDT model. The first row shows the hit and false-alarm counts for each patient, according to their FAST stage, with the counts for a randomly selected patient shown in bold. The second and third rows show the corresponding posterior predictive distributions for hit and false alarm counts for the group data and for the individual patient data. In the posterior predictive panels, the box sizes are proportional to the mass of the posterior predictive distribution for that combination of hits and false alarms.

hierarchical SDT model. The first row corresponds to the behavioral data, the second row to the group-level inferences of the model, and the third-level to the individual-level inferences of the model. The columns correspond to the five FAST stages. Each panel shows the distribution of data or predicted data in terms of hit and false-alarm counts, as in standard Receiver Operation Characteristic (ROC) analysis (e.g., MacMillan Creelman, 2004).

The observed data for all patients are shown as gray crosses, except for one highlighted individual—selected out to test the individual-level predictions of the model—shown by a black cross. For the group level, the model's posterior predictions are shown by squares, with areas proportional to predictive mass. It is clear that the group-level predictions match the data, and show a degradation in performance, with fewer hits and more false-alarms, as the severity of AD progresses. In this sense, the model provides an accurate description of the similarities and differences between clinical sub-populations. In the individual-level model predictions, the area of the squares again correspond to predictive mass, and provide accurate fits to the observed data. We note that several of the individuals were deliberately chosen to be outliers within their clinical sub-population. The ability of the model for describe these individuals well, while simultaneously describing group-level performance, highlights the advantages of the hierarchical approach we have taken to modeling individual differences.

Assessing Discriminability and Bias

Figure 4 shows the joint and marginal posterior distributions for both discriminability and bias, at the level of the FAST stage groups. The main panel shows samples from the joint distribution for each of the five FAST stages. The side panels show the marginal distributions for both discriminability and bias.

As would be expected, discriminability decreases as AD severity progresses, starting around $d' = 4$ for non-AD patients in the first two stages, and decreasing to $d' < 1$ for patients in stage 5. The pattern change in recognition bias across the stages is more revealing. Patients in the non-AD stages start with a conservative bias, with $c > 0$, meaning they are more likely to fail to recognize studied words than to false-alarm to non-studied words. This bias changes significantly for the AD patients, and becomes much more liberal, shifting to a position almost consistent with unbiased responding at $c = 0$.

Assessing Change in Recognition Criterion

Figure 4 shows that the change in criterion is sudden and sustained. At FAST stage 3—which is the first AD

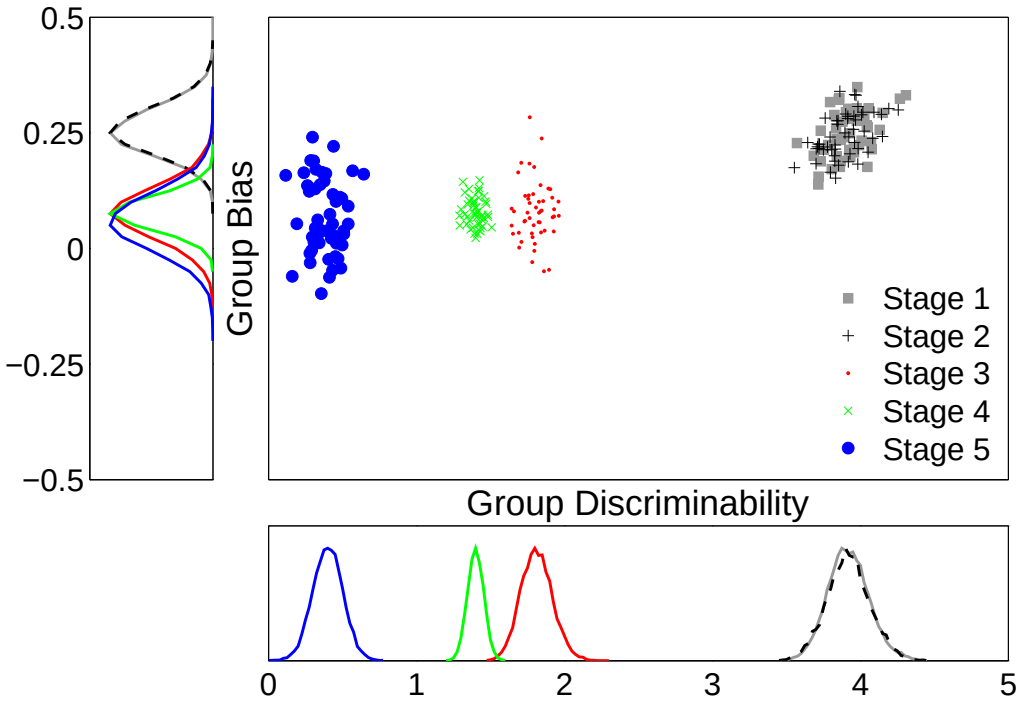


Figure 4: Joint and marginal posterior distributions for the group-level discriminability and bias parameters, for each of the five FAST stages.

stage—the distribution of individual response bias moves to a smaller value, and it sustains approximately the same distribution over subsequent progression through stages 4 and 5.

Our generative model of change allows an immediate inference about the significance of these apparent change in response bias, via the posterior distribution of the a parameter. This is the parameter that control the step-change in response bias between AD and non-AD diagnoses. Its posterior distribution is shown in Figure 5, and is clearly negative, and does not include zero, confirming the liberal change in bias at the onset of AD.

Discussion

Our results are largely consistent with previous findings, but are not identical. We have corroborated the most important existing finding, which is that the onset of AD leads to a liberalization in response bias in recognition memory tasks. Our results, however, extend the previous understanding of the change in response bias, through using a clinical data set with more FAST stage information about AD progression. Using this more detailed measure we found, perhaps surprisingly, that the change in response bias seems to involve a sudden shift at the onset of AD, rather than gradual change over its progression.

Unlike most previous studies, we found non-AD patients starting from a conservative criterion setting—being more likely to miss than to false-alarm—and so the liberalization actually leads to more unbiased decision-making in the AD patients. There are many possible reasons for this difference, which are worth further investigation. One possibility involves methodological issues, including details of the assessment tasks, such as differences in the word lists used. Another possibility relates to more fundamental theoretical and modeling differences in our analysis. We have introduced a number of innovations, any (or all) of which might lead to different findings from more standard analyses.

We think the modeling approach we have used has some clear advantages over previous work. As AD progresses, memory capabilities and decision strategies change in important and interpretable ways. But there remains variability in the characteristics of individual patients, even though they can appropriately be classified within groups like FAST stages. Our hierarchical approach naturally incorporates this interplay between clinical sub-populations and individual patients, making it suitable for both broad characterization of AD progression and for individual diagnosis.

Throughout our modeling, we used a simple extension of the standard SDT model to allow for unequal-variances between studied and non-studied words. We think this theoretically preferable, although we did not observe very different results when we repeated the current analyses with equal-variance SDT. Perhaps the most striking difference was that the posterior for the response bias parameter in Figure 5 showed a much stronger change in bias for the non-AD versus AD comparison.

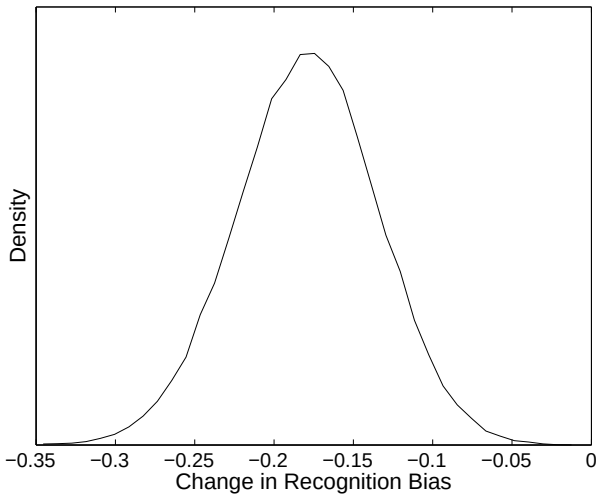


Figure 5: Posterior distribution for the a parameter, which controls the change in response bias from non-AD to AD patients.

It may be that equal-variance SDT overstates the change in decision strategies.

We believe the framework for modeling change we have introduced also has great potential, but realize we have only taken the smallest first step. The key idea is that group-level parameters like discriminability and bias can now be inter-related across diagnoses or classifications like FAST stages. We used a simple step function between non-AD and AD patients, but much more sophisticated functional relationships could be modeled, expressing a theory of how key psychological variables change throughout AD progression. Even more generally, graphical models provide a natural vehicle for modeling and evaluating changes in these variables due to external factors like treatments in clinical trials, or for expressing these variables in terms of causal or co-variate information like demographic or other properties of people. These sorts of extended possibilities highlight the potential of using cognitive models like SDT and hierarchical Bayesian analysis to understand Alzheimer's Disease.

Acknowledgments

This research was supported by award NIRG-08-90460 from the Alzheimer's Association.

References

Budson, A. E., Wolk, D. A., Chong, H., & Waring, J. D. (2006). Episodic memory deficits in Alzheimer's disease: Separating response bias from discrimination. *Neuropsychologia, 44*, 2222–2232.

Dennis, S., Lee, M., & Kinnell, A. (2008). Bayesian analysis of recognition memory: The case of the list-length effect. *Journal of Memory and Language, 59*, 361-376.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC.

Jordan, M. I. (2004). Graphical models. *Statistical Science, 19*, 140-155.

Lee, M. D. (2008). Three case studies in the Bayesian analysis of cognitive models. *Psychonomic Bulletin & Review, 15*, 1-15.

MacKay, D. J. C. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge: Cambridge University Press.

MacMillan, N., & Creelman, C. D. (2004). *Detection theory: A users guide* (2nd ed.). Hillsdale, NJ: Erlbaum.

Mickes, L., Wixted, J. T., & Wais, P. E. (2007). A direct test of the unequal-variance signal-detection model of recognition memory. *Psychonomic Bulletin & Review, 14*, 858-865.

Morris, J. C., Heyman, A., & Mohs, R. C. (1989). The Consortium to Establish a Registry for Alzheimer's Disease (CERAD) Part I. Clinical and neuropsychological assessment of Alzheimers disease. *Neurology, 39*, 1159-1165.

Norman, K. A., Detre, G. J., & Polyn, S. M. (2008). Computational models of episodic memory. In R. Sun (Ed.), *The Cambridge handbook of computational psychology* (pp. 189-224). New York: Cambridge University Press.

Reisberg, B. (1988). Functional assessment staging (FAST). *Psychopharmacology Bulletin, 24*, 653–659.

Riefer, D. M., Knapp, B. R., Batchelder, W. H., Bamber, D., & Manifold, V. (2002). Cognitive psychometrics: Assessing storage and retrieval deficits in special populations with multinomial processing tree models. *Psychological Assessment, 14*, 184-201.

Rouder, J. N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review, 12*, 573–604.

Shankle, W. R., Mangrola, T., Chan, T., & Hara, J. (2009). The CERAD wordlist memory performance index: Development and validation. *Alzheimer's & Dementia, 5*, 295-306.

Shiffrin, R. M., Lee, M. D., Kim, W., & Wagenmakers, E.-J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive Science, 32*, 1248-1284.

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General, 117*, 34–50.

Spiegelhalter, D. J., Thomas, A., & Best, N. G. (2004). *WinBUGS Version 1.4 User Manual*. Cambridge, UK: Medical Research Council Biostatistics Unit.

Context, Syntactic Priming, and Referential Form in an Interactive Dialogue Task: Implications for Models of Alignment

Kathleen M. Carbary (kcarbary@bcs.rochester.edu)
Ellen E. Frohning (efrohnin@u.rochester.edu)
Michael K. Tanenhaus (mtan@bcs.rochester.edu)
Department of Brain and Cognitive Sciences, University of Rochester
Rochester, NY 14627 USA

Abstract

Two experiments examined how context and syntactic priming interact to determine interlocutors' choice of referential form. Pairs of naïve participants took turns producing descriptions of target pictures from a set of alternatives. The first experiment established that a contrast picture in the display (e.g. a striped cat in a display where a spotted cat was the target) primarily determined whether an adjective was used. Priming with an adjective had a only small, secondary effect on adjective use. However, when an adjective was used, it was more likely to occur in the prime-congruent structure than the alternative structure. Experiment 2 compared the effects of a prime produced by the dialogue partner with the effects of a pre-recorded prime played through headphones. Syntactic priming was significant only for the dialogue prime trials, indicating that priming may be stronger in dialogue than outside of dialogue, as previous work has suggested. However, even in dialogue, the primary factor that determined referential form was the set of alternatives. Our results begin to clarify the role of syntactic priming in dialogue, suggesting that it has at most a small effect on message formulation.

Keywords: dialogue; language production; referential form; syntactic priming; alignment; message formulation.

Introduction

How does a speaker choose the content and form of a referring expression? In addition to being a classic question in the philosophy of language, it is an important problem for language generation systems. Such systems aim to approximate the types of utterances that a speaker would produce in a task-oriented dialogue, and thus provide important data for evaluating models of dialogue developed within psycholinguistics.

Work on reference production in the Gricean tradition assumes that a speaker will provide sufficient information for her addressee to identify an intended referent, taking into account the purpose of the conversation (Grice, 1975). Speakers should be specific enough for the addressee to identify the intended referent, without being overly specific by providing unnecessary information. For example, a speaker might say "the cat" when referring to a single cat among several other animals, but refer to the same animal as "the striped cat" when there are multiple cats present.

A more striking observation in language production research is that interlocutors not only converge on the same referring expressions (Clark & Wilkes-Gibbs, 1986), but also begin to use the same syntactic forms. This effect is often referred to as structural persistence, syntactic persistence, structural priming, or – as we will call it in this paper – syntactic priming (Bock, 1986; Ferreira & Bock, 2006; Pickering & Ferreira, 2008). Some research suggests that syntactic priming might be stronger in dialogue settings than in other types of experimental tasks (Branigan, Pickering, McLean, & Cleland, 2007; Branigan, Pickering, & Cleland, 2000). Pickering & Garrod account for this trend by suggesting that "a major reason why priming effects occur is to facilitate alignment, and therefore they are likely to be particularly strong during natural language interactions" (p. 174, 2004).

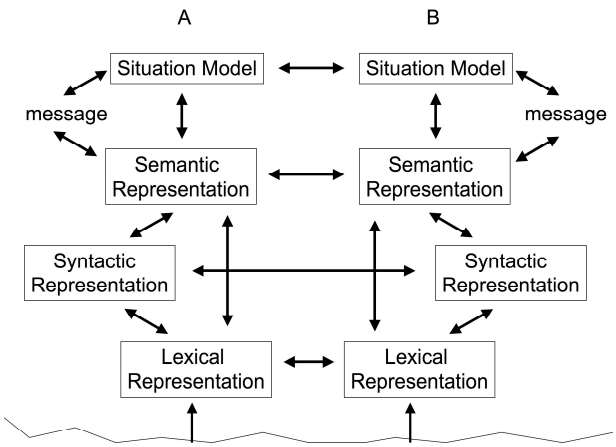


Figure 1: The Interactive Alignment Model (based on Pickering and Garrod, 2004); the mental representations of hypothetical interlocutors, A and B, are shown.

As a first step toward extending this idea and integrating it with a broader model of language production in dialogue, Pickering & Garrod (2004) proposed the Interactive Alignment Model, illustrated in Figure 1. The model assigns a central role to syntactic priming, casting it as a mechanism that aligns interlocutors' representations at multiple levels. Alignment at one level affects alignment at higher and lower levels of representation, making the model interactive.

Notably, the model assumes that the process through which interlocutors come to have the same representation of a situation is unconscious and automatic. Alignment at the syntactic level – that is, syntactic priming – is treated as an important factor that allows for communicative success, by increasing alignment at these other levels (Pickering &

Garrod, 2004, 2006). The model treats syntactic priming as a mechanism for alignment at other levels, which in turn explains how interlocutors are able to communicate successfully. In this way, syntactic priming has played a major empirical and theoretical role in the development of alignment models of dialogue. However, few, if any, investigations have attempted to examine how and when syntactic priming might affect representations and processes at various levels of representation in dialogue. We focus on two levels of representation that are typically distinguished in models of language production: message formulation and building the syntactic form of the utterances (see Bock & Levelt, 1994, for a classic language production model that makes this distinction).

According to Bock & Levelt, the message formulation level "captures features of the speaker's intended meaning and provides the raw material for the processes of grammatical encoding" (p. 946, 1994). In other words, the message formulation stage involves planning the information to be communicated at a pre-grammatical level. Bock & Levelt present this as a stage that occurs before grammatical encoding begins. However, the Interactive Alignment Model is compatible with the possibility that grammatical encoding, such as syntactic category selection, could indirectly affect the message (refer to Figure 1 for a general idea of how this would work in that model). While message selection is most naturally affected by Gricean processes, like including the appropriate amount of information in an utterance, it is possible that the availability of syntactic structures could have some influence on the message that is formed.

Optional adjective use within a noun phrase is ideal for investigating this possibility. Adjectives are often used in referential expressions, even when the context does not require it from a Gricean perspective. Speakers' are prone to over-informative adjective use when describing objects, including adjectives in their descriptions unnecessarily up to 46% of the time (Belke, 2006; Sedivy, 2003). This makes noun phrases that can optionally contain adjectives (e.g. *the [striped] cat*) ideal for an investigation of how syntactic priming and message formulation interact. Syntactic priming with an adjective-containing structure might increase the likelihood that the subsequent message will include information associated with an adjective, so that the primed syntactic structure can be used again. This structure provides a special opportunity to observe any potential effects of priming on message content, since speakers are free to use adjectives, even when the context does not require it. In addition, previous work has indicated that noun phrase structures containing adjectives are susceptible to syntactic priming effects (Branigan, McLean, & Jones, 2005). This structure should therefore allow us to observe any possible effects of syntactic priming on message formulation.

The hypothesis is that the message selection level of representation could be subtly affected by the increased availability of a primed syntactic representation. We compare two conditions: one in which syntactic priming could affect message formulation by increasing the likelihood of subsequent adjective use, and another where no increase in adjective use would be expected. For example, will a speaker be more likely to refer to a single cat, among other potential referents, as "the striped cat," if their interlocutor had used an adjective-containing noun phrase construction on the previous trial?

Our design also allows us to explore the claim that syntactic priming effects are stronger in dialogue than in non-dialogue situations. Although some previous research has suggested that syntactic priming effects are stronger in dialogue, this conclusion has been primarily based on post-hoc comparisons of priming effects between experiments that use different methods (but see Branigan, et. al., 2000, 2007 for exceptions). This creates the potential for confounds which could mimic a difference between dialogue and non-dialogue, and makes it difficult to determine whether differences are significant.

In addition, the few experiments (Branigan, et. al., 2000, 2007) that directly examine syntactic priming effects in dialogue are scripted confederate studies, in which a participant takes part in a highly controlled task with a trained assistant. In this setting, many factors that would normally affect what is said – such as referential context and lexical availability – are highly controlled by the situation, and unlikely to have a strong effect. This is a problem, since syntactic priming effects may appear to be larger when other influences on a referential expression are minimized. It is unclear whether the magnitude of priming effects in such a setting can be considered evidence that priming is a special mechanism that causes language production to occur differently in dialogue than in other experimental settings.

We report two experiments that investigate how syntactic priming affects referential form in an unscripted dialogue task. The first experiment examined the effects of syntactic priming on message formulation during dialogue. The second experiment compared priming effects in dialogue with priming effects outside the dialogue, in an otherwise identical task. Both experiments allowed us to explore how referential constraints interact with syntactic priming to determine referential form, and to address the relationship between syntactic priming and successful communication.

Experiment 1

Experiment 1 was a first step towards examining how syntactic priming in dialogue affects other levels of representation. Specifically, does priming affect alignment at the level of message formulation, as one interpretation of the Interactive Alignment Model suggests? Or, does syntactic priming exert an effect on language production only after the message to be communicated has been fully planned, as the Bock & Levelt (1994) model predicts?

Materials and Methods

Participants Fifteen pairs of friends from the University of Rochester were paid to participate. All were native English speakers and naïve to the purpose of the experiment.

Experimental Setup Individual participants sat at separate computers on either side of a large cardboard barrier, so that they could not see each other or each other's computer screens. To ensure that they could clearly hear each other, participants wore headphones and spoke into microphones. This setup facilitated audio recording of the entire session. To initiate each trial, one participant clicked a central fixation cross. The same set of four clip-art pictures then appeared on both screens. To discourage participants from using expressions like "the top left picture," picture locations were pseudo-randomized. After a 2-second delay, Participant 1 saw a circle appear around the target picture. Her task was to instruct her partner to click on that picture, using any description she chose. The trial ended when Participant 2 clicked the target picture. The overall error rate was less than 1%, and participants were given no feedback about their performance. Participants alternated between giving and responding to instructions, and found the task very easy and natural (see Figure 2).



Figure 2: The experimental setup.

The order of prime target pairs was pseudo-randomized so that different participant pairs saw the displays in different counter-balanced orders. The experiment was divided into blocks, so that participants had 5 breaks throughout the experiment.

Experimental Items There were two types of displays that occurred in pairs: *prime displays* and *response displays*. Half of the prime displays were *adjective primes,* designed to elicit descriptions that included either a pre- or post-nominal adjective; for example, "click the striped cat" or "click the cat with stripes." This was achieved using a contrast set, including the target and a picture that differed from the target in only one adjectival property (e.g. a striped cat vs. a spotted cat). This required participants to use an adjective in their description in order to uniquely identify the target.[1] The *no adjective prime* displays contained a target picture with no related pictures in the display, allowing participants to successfully describe the target without an adjective.

Each prime display was followed by a response display. The referential context of the response displays was manipulated so that the target was part of a contrast set half the time, and appeared with unrelated pictures only half the time. When there was *contrast* in the display, an adjective was required for a felicitous referential expression, and when there was *no contrast* an adjective was unnecessary. This 3 x 2 design allowed us to test the effects of prime type (no-adjective, prenominal, postnominal) and contrast (present or absent) on the referential expression produced in a response display.

Coding and Analysis The entire interaction was digitally recorded. Participants' descriptions of the pictures were later transcribed word-by-word, and coded by the second author according to the syntactic structure had been used (e.g. prenominal, postnominal, noun only, etc.). Task-irrelevant utterances were not included in the analysis. All statistical comparisons were made using mixed-effects regression models,[2] which were computed using the R data analysis software, version 2.6.1 (2007).

Results and Discussion

We wanted to answer two questions: was there a basic syntactic priming effect, and if so, did priming affect message formulation by increasing adjective use? Looking first at the subset of data where an adjective *was* used in the description of the response display, we asked whether prenominal and postnominal primes types had an impact on the syntactic structure of the description. If syntactic priming effects in dialogue are strong, then we would expect to see a strong syntactic priming effect: participants should produce more prenominal structures following prenominal primes, and more postnominal structures following postnominal primes. When the property associated with an adjective was already included in the message, we expected that the prime type would affect the structure in which the adjective appeared. Two separate mixed-effects regression models, with subject pair and item as random effects, were

[1] Norming data allowed us to classify prime displays as being likely to generate prenominal or postnominal descriptions, and the experiment included half of each display type. This prompted participants to use a prenominal adjective on 47% of prime descriptions, and a postnominal adjective on 43% of prime descriptions, even though no limitations were placed on what participants could say.

[2] Mixed-effects regression models were more appropriate for our dataset than ANOVAs, since the unscripted nature of the task led to unequal numbers and variances in each cell of the design. For a discussion of why this choice was appropriate, see Jaeger (2008).

used to test for significance of prenominal and postnominal priming.[3]

This analysis revealed that the use of a prenominal prime significantly predicted the use of a prenominal adjective in the subsequent response description ($B = 0.45$, SE = 0.20, $p = 0.05$). Similarly, the use of a postnominal prime significantly predicted the use of a postnominal response description ($B = 0.46$, SE = 0.21, $p < 0.05$). As shown by the coefficients, the magnitude of the effect was approximately equal for pre- and postnominal primes. On average, participants produced a prime-congruent response (i.e. a response that contained the same structure as the prime) 61% of the time, and an incongruent response 39% of the time. This 22% difference is similar to what has been found in classic priming studies not involving dialogue (e.g. the alternating dative priming effect shown by Bock, 1986).

Having established a syntactic priming effect when the message includes an adjective, we evaluated the extent to which priming affected message content. Figure 3 shows the rate of adjective use for response descriptions following each prime type. The pre- and postnominal prime types did not produce different effects, and so they were collapsed into one "adjective prime" type for the purposes of analysis. A mixed-effects regression model with subject pair and item as random effects tested the significance of three predictor variables: Display Type, Trial Order, and Prime Type. The coefficient and significance level for each of these factors is shown in Table 1.

Table 1: The effects of contrast, trial order, and prime type on adjective use.

Predictor	Coefficient (SE)	Significance
Contrast in Display	4.53 (0.43)	$p < 0.001$
No Adjective Prime	0.60 (0.61)	n.s.
Trial Order	-0.0004 (0.002)	n.s.
No Adjective Prime x Contrast in Display	-2.06 (0.88)	$p < 0.05$
No Adjective Prime x Trial Order	-0.0015 (0.005)	n.s.
Contrast in Display x Trial Order	-0.0011 (0.0038)	n.s.
Contrast x Trial Order x No Adjective Prime	-0.0040 (0.0074)	n.s.

There was no effect of trial order, indicating that participants did not prime each other more as the experiment unfolded. Instead, the degree of syntactic priming remained constant over the course of the experiment. This is not what would be expected if syntactic priming was associated with successful dialogue, as participants became faster and better at this communication task as the experiment unfolded.

As shown in Table 1, adjective use in response display descriptions was predicted only by a main effect of contrast and an interaction between prime type and contrast.

The primary determiner of whether the message included an adjective was the referential context. When a contrast was present in the display, there was a small additional increase in adjective use when the preceding prime had contained an adjective (see Figure 3). However, there was no main effect of adjective prime indicating a complex relationship between priming and adjective use. This was true even though the message could have been modified to include more information based on the presence of an adjective in the preceding prime without any negative consequences for communication.

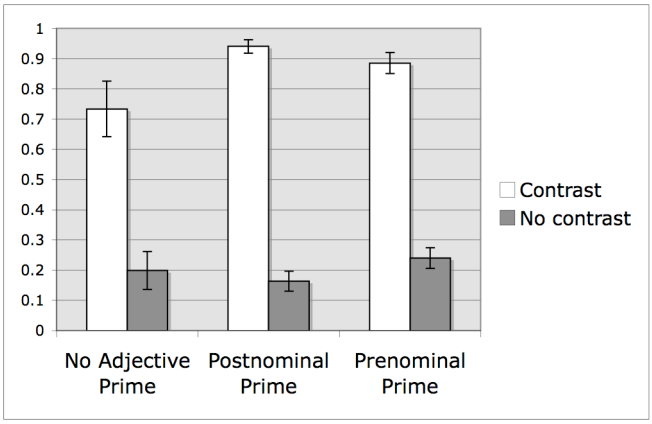


Figure 3: Mean (standard error) adjective inclusion rates in response descriptions following three prime types, for contrast and no-contrast displays.

The only suggestion of an effect of syntactic priming on message content was the slightly higher rate of adjective use following adjective primes. Priming appeared to increase the rate of adjective use only when a contrast was present. An alternative interpretation is that adjective use following no-adjective primes was artificially low. This may have occurred because some of no-adjective primes involved single words that were coded as nouns, but that could also have been considered adjectives (e.g. *wood* for a tree branch or *fluffy* for a Persian cat). This small subset of the data may have increased the likelihood that an alternative adjective-containing structure would be used again, thereby reducing the rate of pre- and postnominal adjective use following no-adjective primes. This is a viable alternative explanation for the lower rate of adjective use following no-adjective primes in this study, which will need to be carefully explored in future work.

Experiment 2

Experiment 2 was designed to extend the results of Experiment 1, by directly comparing the effects of dialogue and non-dialogue primes using a within-subjects design. If

[3] We tested for prenominal and postnominal priming separately, to determine whether one structure caused stronger priming than the other, and to rule out the possibility that the overall priming effect was driven by only one of these structures. Since priming effects were comparable for both structures, subsequent analyses treat prenominal and postnominal priming together.

syntactic priming is stronger in an unscripted dialogue setting than in a non-dialogue setting, participants should be more likely to reuse a syntactic structure generated by the conversation partner than a description that had been pre-recorded by a speaker not participating in the dialogue.

Materials and Methods

Participants Seventeen pairs of friends from the University of Rochester were paid to participate. All were native English speakers who had not taken part in the first experiment.

Experimental Setup and Items The setup was the same as for the first experiment, with a few notable changes. First, primes were now divided into two new categories, depending on dialogue status. Dialogue primes involved one participant describing a prime display to her partner; this was followed by the other participant describing a response display. For one third of trials, non-dialogue primes that had been pre-recorded by a trained female speaker were played through headphones to the participant who was the listener on that trial. The other participant, who would normally be generating the prime description, did not hear the prime, and instead completed an unrelated task (clicking a dot that appeared in an unpredictable location). This prevented the pre-recorded prime from becoming part of the participants' shared knowledge about the situation, or become introduced to the dialogue in any another way. All the response descriptions were participant-generated, regardless of prime status. In order to include enough trials in each condition to support the comparison between dialogue and non-dialogue, the no-adjective primes were eliminated. Thus, we manipulated *prime type* (prenominal or postnominal) and *prime status* (dialogue or non-dialogue) independently.

Results and Discussion

If syntactic priming in dialogue is truly stronger than outside of dialogue, then participants should be more likely to re-use the syntactic structure just used by an interlocutor than a structure just produced by a prerecorded voice. The results of our second experiment supported this prediction. When all the descriptions that included adjectives were considered together, we saw a small but significant priming effect for both prenominal and postnominal primes ($p < 0.05$), just as in experiment 1. However, when these response descriptions were examined separately by prime status, participant-generated primes had a greater impact on the subsequent descriptions than pre-recorded primes (see Figure 4).

A mixed effects regression model with participant pair and item as random effects was used to test for significance. Whether a response description was syntactically congruent or incongruent with the preceding prime was predicted from the prime status. When only trials containing adjectives were considered, incongruent responses were significantly more likely following pre-recorded primes than following dialogue primes (B =0.58, SE = 0.16, $p < 0.001$). The observation of priming effects in this paradigm depended on dialogue, since syntactic priming was not observed with the pre-recorded primes.

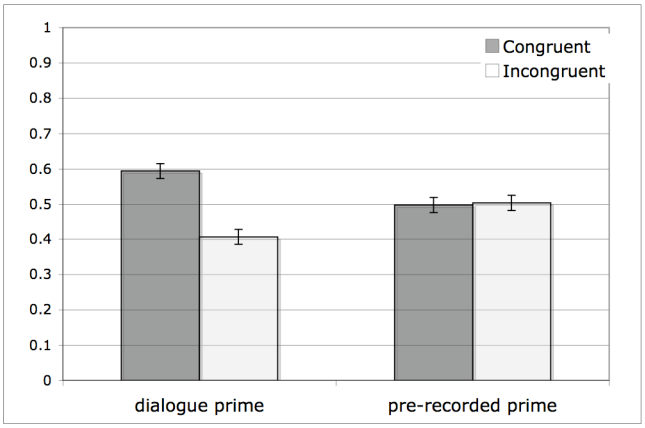


Figure 4: The proportion (standard error) of prime-congruent and prime-incongruent response descriptions following dialogue and non-dialogue primes, when only trials containing an adjective are considered.

Next, we wanted to address the hypothesis that syntactic priming effects in dialogue are instrumental in successful communication, as has been suggested by Pickering and Garrod (2006). One way to test this hypothesis is to examine priming over the course of the experiment. As the experiment unfolded, participants become better at the task, making fewer mistakes, and competing the trials more quickly. If syntactic priming promotes successful communication by increasing alignment at other levels, then we might expect that increased levels of priming should be correlated with this improvement at the task. However, this was not the case: syntactic priming did not significantly increase or decrease over the course of the experiment. Moreover, the degree of syntactic alignment, that is, the proportion of trials where participants re-used the primed structure, was not correlated with a pair completing the task more quickly (Spearman's rho = -0.197, *n.s.*). This was true both for the subset of trials where the response description included an adjective and for all of the trials.

When examined as part of a larger system of language production in dialogue, syntactic priming appeared to play only a small part in determining referential forms. There was no evidence from this experiment to support the idea that syntactic priming contributed to task success. This suggests that syntactic priming and successful communication are not necessarily related.

General Discussion and Conclusions

In Experiment 1, we examined how referential context and syntactic priming interact to affect referential form. At the level of message formulation, where a speaker makes decisions about what information to include in an utterance, content was determined primarily by referential context.

One hypothesis was that syntactic priming would increase the likelihood that a speaker would include an adjectival property in the message, in order to re-use the structure that had just been primed. When the context strongly supported including an adjective in the message, priming with structures containing an adjective had a small additional effect on adjective use. However, when the context did not support adjective inclusion, priming had no affect on message content. This rules out the possibility that syntactic priming has a strong affect on message formulation independent of other factors. Our results are compatible with a model in which context constrains message content, and syntactic priming exerts a small additional affect. However, it is also possible that syntactic priming affected the message structure, but not the content; the rate of adjective use following no-adjective primes might have been lower due to adjective-like content being incorporated into the message in other ways.

In Experiment 2, we compared syntactic priming in dialogue and non-dialogue trials during an unscripted interaction between two naive participants. We found that syntactic priming depended on a prime that was generated by the conversation partner, as the Interactive Alignment Model suggests. This is in line with the trends that have been observed in previous experiments: syntactic priming effects are greater in dialogue than in response to a non-dialogue prime. We did not, however, find a relationship between syntactic alignment and task success. These results, taken together with the findings of previous work, raise questions about whether priming facilitates communication by aligning interlocutors' mental representations. In future research it will be important to address the relationship between priming and task success more directly. This could involve using more complex tasks, where there is a greater likelihood of differences in how well participants perform in a task-oriented dialogue.

These experiments shed light on how syntactic priming affects the selection of referential forms in dialogue, suggesting that while priming occurs, it is secondary to contextual factors that more strongly constrain what is said. This represents an initial step toward more carefully evaluating if and how syntactic priming impacts other levels of representation in dialogue. It also highlights the importance of using experimental designs where potential priming can be observed in interaction with other variables affecting message formulation. Experimental situations in which speakers have a larger range of options, (e.g. Gómez Gallo, Jaeger & Smyth, 2008), will allow priming to be examined in conjunction with such variables in single utterances and pairs of utterances. Situations like these are also ideal for future investigations because they closely approximating natural dialogue settings.

Acknowledgments

We thank Dana Subik for years of assistance with data collection, and current and former members of the Tanenhaus lab for helpful comments on this project. These experiments were supported by the NIH grant HD-27026 to Michael Tanenhaus.

References

Belke, E. (2006). Visual determinants of preferred adjective order. *Visual Cognition, 14(3),* 261-294.

Bock, J.K., & Levelt, W.J.M. (1994). Language production: grammatical encoding. In M. Gernsbacher (Ed.), *Handbook of Psycholinguistics* (pp. 945-984). Orlando: Academic Press.

Bock, J.K. (1986). Syntactic persistence in language production. *Cognitive Psychology, 18,* 355-387.

Branigan, H., Pickering, M., McLean, J., & Cleland, A. (2007). Syntactic alignment and participant role in dialogue. *Cognition, 104,* 163-197.

Branigan, H., McLean, J., & Jones, M.W. (2005). A blue cat or a cat that is blue? Evidence for abstract syntax in young children's noun phrases. In *Proceedings of the 29th Annual Boston University Conference on Language Development,* (p. 109-121). Somerville, MA: Cascadilla Press.

Branigan, H., Pickering, M., & Cleland, A. (2000). Syntactic co-ordination in dialogue. *Cognition, 75*, B13-B25.

Clark, H.H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition, 22*, 1-30.

Gómez Gallo, C., Jaeger, T.F., & Smyth, R. Incremental syntactic planning across clauses. *In Proceedings of the 30th Annual Meeting of the Cognitive Science Society* (pp. 845-850). Austin, TX: Cognitive Science Society.

Ferreira, V. & Bock, K. (2006). The functions of structural priming. *Language and Cognitive Processes, 21(7-8),* 1011-1029.

Grice, H.P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and Semantics*. *Volume 3: Speech Acts* (pp. 41-58): Academic Press.

Jaeger, T.F. (2008). Categorical data analysis: away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language (Special Issue on Emerging Data Analysis), 59,* 434-446.

Pickering, M., & Ferreira, V. (2008). Structural priming: a critical review. *Psychological Bulletin, 134(3),* 427-459.

Pickering, M. & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences, 27,* 169-226.

Pickering, M. & Garrod, S. (2006). Alignment as the basis for successful communication. Research on Language and Computation, 4, 203-228.

R, version 2.6.1, The R Foundation for Statistical Computing, 2007.

Sedivy, J. C. (2003). Pragmatic versus form-based accounts of referential contrast: evidence for effects of informativity expectations. *Journal of Psycholinguistic Research,* 32(1), 3-23.

Converging Hands or Converging Minds?

Lisette Mol (L.Mol@uvt.nl)

Emiel Krahmer (E.J.Krahmer@uvt.nl)

Alfons Maes (Maes@uvt.nl)

Marc Swerts (M.G.J.Swerts@uvt.nl)
Tilburg Centre for Cognition and Communication (TiCC), School of Humanities, Tilburg University
P.O. Box 90135, NL-5000 LE Tilburg, The Netherlands

Abstract

Interlocutors sometimes repeat each other's representational hand gestures. We investigated if this is a case of direct mimicry of form, or whether perceiving a gesture gives rise to a semantic representation, which subsequently informs gesture production. For this we used an interactive route description task, in which a confederate's gestures indicated the route in either the vertical or the horizontal plane and either with one or four fingers extended as an index. We found that perceiving vertical gestures led to an increase not only in participants' production of vertical gestures, but also in their use of one finger as an index, suggesting that seeing vertical gestures caused participants to think of the route as on a map, which led them to point with one finger (as is common on a map) rather than four. Our results support the notion that repetition of meaningful gesture forms results from converging semantic representations.

Keywords: Gesture; Adaptation, Lexical Entrainment.

Introduction

It is well established that when people interact in dialogue, they tend to adapt to each other in many ways (for an overview, see Pickering & Garrod, 2004). For example, interlocutors reuse each other's (referring) expressions (e.g. Brennan & Clark, 1996) and syntactic constructions (e.g. Branigan, Pickering, & Cleland, 2000). In one study, Levelt and Kelter (1982) found that if shop keepers where asked in Dutch "(At) what time does your shop close?", their answer tended to match the question in surface form, either including or omitting 'at'. Similarly, repetitions of form across interlocutors have also been found for co-speech hand gestures (e.g. De Fornel, 1992). Such gestures are spontaneous movements of the hands and arms during speech, which can convey information, or emphasize certain parts of speech (e.g. McNeill, 1992). Elements of a gesture's physical form (*articulators*), like the shape and orientation of the hand, the direction and size of the movement, and where it is performed relative to the speaker can be repeated in subsequent gestures by the same or another speaker.

Some scholars believe that speech and gesture jointly express a speaker's ideas (McNeill, 1992), or that speech and gesture are both part of a speaker's communicative effort (Kendon, 2004). From this perspective, it seems likely that repetition of each other's gesture forms would resemble repetition of each other's (other) linguistic forms. On the other hand, repetitions in physical behavior are found in many species, and need not be tied to speech (Parrill & Kimbara, 2006). In this paper, we focus on gestures that depict some of the content a speaker is conveying, which are known as *illustrators* (Ekman & Friesen, 1969) or *representational gestures* (McNeill, 1992). We compare the repetition across speakers of certain articulators of such gestures to the repetition of meaningful units in speech, specifically *lexical entrainment,* as well as to non-linguistic forms of behavioral mimicry. We first explain a difference between direct behavioral mimicry and lexical entrainment. We then describe some empirical results on the repetition of gesture forms across speakers. This will lead to our research question: Is the repetition of meaningful gesture forms across interlocutors a consequence of converging semantic representations, or is there a more direct link between perceiving a form and producing a form?

Mimicry and Adaptation

Mimicry is defined as one person repeating the behavior of another person (Bock, 1986). Some forms of mimicry enable the transfer of important functional behaviors (Tomasello, Savage-Rumbaugh, & Kruger, 1993). It has also been found that repeating others can have social benefits. Van Baaren et al. (2003) found that a waitress received higher tips when repeating her customers' orders literally, than when signaling in some other way that she understood the order. Yet for some repetitions of behavior, the functional or social purpose is less clear (Chartrand & Bargh, 1999). For example, if one person starts yawning, oftentimes those around will start yawning as well. Chartrand and Bargh explain this type of behavior in terms of the *perception-behavior link*, meaning "the mere perception of another's behavior automatically increases the likelihood of engaging in that behavior oneself", p. 893. Notably, they state that although such mimicry may act as a kind of 'social glue', intent or conscious effort are not required for it to occur. We will subsequently use the term 'mimicry' to refer to such automated repetitions of behavior.

Pickering and Garrod (2004) propose that similar automatic priming underlies the repetition of linguistic

behaviors across interlocutors, a form of adaptation which they call *alignment*. They state that at each linguistic level, "the activation of a representation in one interlocutor leads to the activation of the matching representation in the other interlocutor directly", p. 177. These representations are thought to be used in both language production and processing (parity of representation). Thus, if a certain lexical or semantic representation has just been constructed as a result of hearing an utterance, it can subsequently be used for production. In addition to this direct source of alignment across interlocutors, alignment at one level can also enhance alignment at certain other levels within a speaker, because of bidirectional connections between the representations at different levels. Thus, if a lexical representation is connected to a semantic one, activation of that semantic representation may subsequently activate the lexical one.

Let us focus on one particular case of converging linguistic behavior: the repetition of referring expressions across speakers, known as *lexical entrainment* (Brennan & Clark, 1996). Brennan and Clark propose that interlocutors use the same words to refer to the same objects, because they use similar conceptualizations of that object. For example, suppose a particular object could be thought of as a document, a picture, or a map. When a speaker refers to it with 'the map', she conceptualizes the object for the current purpose as such. If the addressee agrees with this conceptualization and a *conceptual pact* is formed, both interlocutors can subsequently use 'the map' as a reference to both the object and the particular conceptualization of it. In this view, the repetition of references across interlocutors results from the establishment of conceptual pacts.

In both the model by Pickering and Garrod and the model by Brennan and Clark, lexical representations and semantic representations are linked. This is where lexical entrainment seems to differ from some instances of direct behavioral mimicry. In mimicry, we may repeat forms without being aware of their meaning. In other words, the perception of a form directly leads to the production of that form. In lexical entrainment on the other hand, there seems to be an intermediate stage: meaning. A form that is perceived is coupled with a certain meaning. Only when that meaning is activated again is the same form a likely candidate for repetition.

Repetition of Gesture Form

Is meaning also involved in the repetition of gesture forms across interlocutors? Kimbara (2008) observed interlocutors while they were discussing an animated cartoon. She found that their gestures looked more similar if they could see each other, compared to when they were separated by an opaque screen. Thus, it seems that adaptation occurs in gesture like it does in speech. Yet if these similarities in gesture form resulted from similarities in semantic representations, one could argue that they would also occur when interlocutors cannot see each other, since similarities in semantic representations can also be arrived at through

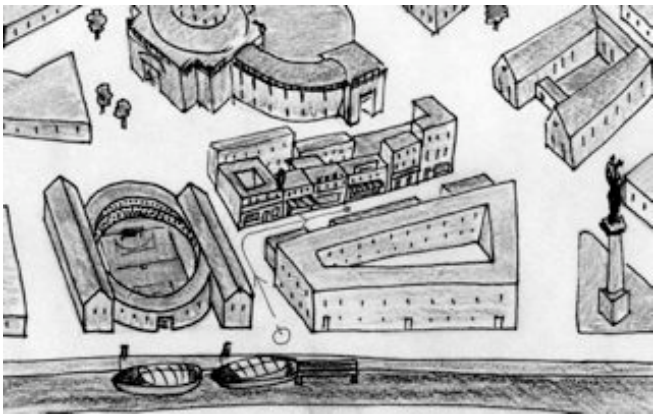


Figure 1: Part of a scene used in the experiment, note the route starting at the bottom-center.

speech. Therefore, it may be the case that the repetitions of gesture form resulted from the automatic across-speaker path of alignment (direct copying of form), rather than from connections between gesture forms and semantic representations, or the establishment of conceptual pacts.

In a previous study, we have investigated how relevant the semantic context was for certain gestures to be repeated across interlocutors (Mol, Krahmer, & Swerts, 2009). Gestures were either performed with speech matching the gesture's form in meaning (e.g. a speaker moved his arms as though running, while talking about running), or with speech that expressed a very different meaning (e.g. the same gesture performed while describing looking through binoculars). We found that repetition did not occur when a gesture was shown in a non-matching semantic context. This suggests that repetition of form in gesture may result from the coupling of a certain form to a certain meaning, rather than from direct copying of form. However, since the mismatches were designed to be very clear in this study, the non-matching gestures may not have been processed very deeply to begin with, or participants may simply be less likely to adapt to a less coherent speaker. We thus need to investigate further whether repetition of gesture forms across interlocutors results from converging meanings.

Present Study

In this study we investigate whether a perceived gesture form can influence the construction of meaning (whether it be any semantic representation or a conceptual pact), which subsequently influences gesture production (also see Cassell, McNeill, & McCullough, 1998). Suppose that certain articulators of a perceived gesture give rise to the construction of meaning. Then when this meaning is subsequently expressed in gesture, all articulators of the gesture produced will likely be consistent with this meaning. Therefore, we would expect that articulators of the perceived gesture that are inconsistent with the meaning constructed would not be repeated as frequently. On the other hand, if repetition of gesture form happens without the

semantic level being involved, any combination of perceived articulators could be repeated in gesture production. In this case, whether an articulator matches the constructed meaning will not affect how frequently it is repeated.

To test this we use a task in which a confederate and a participant give each other route directions repeatedly. We present participants with bird's view drawings of a city scene, with a short route indicated on them (see Figure 1 for an example). These scenes are neither presented vertically nor horizontally, but at an angle. Therefore, the production task can be thought of as either describing a route on a vertically oriented map, *or* as describing a route through an actual (horizontal) city.

In each condition, the confederate expresses only one of these conceptualizations in her gesturing. While speech is kept constant, she gestures either as though moving along a route in a horizontal city, or as though indicating the route on a vertical map. This is done using two articulators: the plane in which the gesture is produced (either horizontal or vertical) and hand orientation (with fingers moving along with the route, or pointing forward as though on a map).

It is interesting in itself to see whether participants adapt to the confederate's perspective in their gesturing. Yet this alone would not tell us whether this is based on direct mimicry of form, or on the convergence of semantic representations. Therefore, we manipulate a third articulator (hand shape) independently. Gestures are produced either with one finger, or four fingers extended as an index. Now if it is the case that gesture form is perceived and reproduced directly, without mediation of meaning, both the confederate's perspective and her hand shape could be repeated independently by participants in their own gesturing. There may be a difference in how frequently each aspect is repeated, but what we would not expect based on this view, is for the confederate's perspective to influence participants' hand shapes or for the confederate's hand shape to influence participants' perspective.

On the other hand, if meaning does form an intermediate stage between the perception and production of gesture forms, we would expect such cross-effects to occur. For example, it is more common to point at a map using a single finger, than it is to point at a map using four fingers at once. Therefore, if the confederate's vertical perspective would lead participants to think of the communication task as describing a route on a map, then their gestures may be produced more frequently with only one finger as an index (even if the confederate uses more than one finger). This would mean there is an effect of the confederate's gestures' perspective on the hand shape of participants' gestures. This effect may also be found in the opposite direction: the confederate's use of more fingers as an index may lead participants to think of the route as through an actual city rather than on a map, causing them to gesture horizontally more frequently.



Figure 2: Partial view of the experimental set-up. Participants were seated on the right.

Method

Participants

48 Native Dutch speakers, all students of Tilburg University took part in this study. The data of eight participants could not be used for analysis (six participants did not produce any relevant gestures). The remaining 40 participants (33 female) had a mean age of 20.5, range 18 – 25.

Procedure

The participant and the confederate came to the lab and were introduced by the experimenter. They each received a written instruction and were seated across from each other. The instruction explained the communication task, and stated that the couple with most correct responses could win a book voucher (in reality there was a random draw). To their side (right to the participant) was a table, on which there was a flip chart for each of them. In between these flip charts was a screen, such as to keep information private. The screen did not keep the interlocutors from seeing each other, see Figure 2. Both behind the confederate and the participant was a camera capturing the other interlocutor.

After reading the instruction, both 'participants' were allowed to ask questions. The confederate always asked one question, after which the experimenter quickly went over the task again. Then the experimenter turned on the cameras and left the room.

The confederate started by studying a little map and memorizing the route on it. Each route had one turn, see Figure 1 for an example. She then turned the page of her flip chart and described the route to the participant, for example: "Je begint bij de rondvaartboot, dan ga je langs het voetbalstadion en dan rechts een winkelstraat in tot ongeveer halverwege." ("You start at the tour boat, then you go along the soccer stadium and then into a shopping street on the right until about halfway.") The confederate's speech always followed the same script. Gestures were timed naturally with speech and gazed at by the confederate.

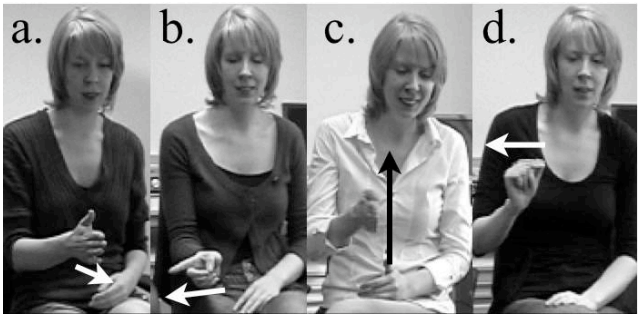


Figure 3: The confederate's path gestures. a: Hand/ Route; b: Finger/ Route; c: Hand/ Map; d: Finger/ Map.

After the confederate's description, the participant turned a page and was to choose which route had just been described, selecting from four alternatives by pronouncing the corresponding letter. No feedback was provided. Then it was the participant's turn to study a route. This route was always on the same scene that the confederate's route had been on. After turning the page (rendering a blank page) the participant described the route to the confederate, who then turned a page and selected one of the four alternatives. This ended one cycle of the experiment. In total each participant perceived and produced five route descriptions.

Afterward, both the confederate and the participant filled out a questionnaire, which included questions on the presumed purpose of the experiment and whether the participant noticed anything peculiar, as well as some questions on how they liked their interaction partner. It ended with the question whether the participant was left or right handed. When the participant was done filling out the forms, the confederate revealed herself and asked the participant's consent for the use of the data. Participants were also asked if they had suspected any deception. Two participants were excluded from our analysis, because they indicated having been suspicious about either the goal of the experiment or the role of the confederate.

Design, Coding and Analysis

We used a 2 x 2 between subjects design. The independent variables were the hand shape (one or four fingers extended) and perspective (route or map) of the confederate's path gestures. In the route perspective, gestures were performed with the index in the direction of the hand movement and movement was in the horizontal plane, as though following a virtual route (Figure 3a, 3b). In the map perspective, gestures were performed in the vertical plane and the index was always pointing forward, as though pointing on a virtual map (Figure 3c, 3d).

The confederate gestured with her right hand. The first direction of a route was always straight, which was depicted with either a forward or an upward movement. These movements were of comparable size. The gesture for the second direction (to the side) was placed relative to the first gesture; it started where the first gesture had ended.



Figure 4: Examples of participants' path gestures (published with permission of the people depicted). a: Hand/ Vertical Map; b: Finger/ Vertical Map; c: Hand/ Route; d: Finger/ Horizontal Map.

We coded all *path gestures* participants produced, that is, all gestures in which one or more fingers were extended as an index, there was hand movement along some virtual path, and the co-occurring speech mentioned a direction to take. Within the stroke phase of each path gesture, we coded hand shape and perspective. The labels for hand shape were *Finger*, when one finger was extended as an index, and *Hand*, if more than one finger was extended. The label for perspective was based on three articulators: location in the gesture space, hand orientation, and movement (direction and size). It turned out that in addition to the two perspectives the confederate had used, participants occasionally used an alternative one, as though pointing on a horizontal map. Therefore, we chose from three labels: *Route, Vertical Map*, and *Horizontal Map.* A gesture in the route perspective would typically have horizontal movement in front of and to the side of the speaker, with the fingers pointing in the direction of the hand movement (Figure 4c). Vertical Map gestures on the other hand would typically have vertical movement, with relative sizes mapping onto distances on the map, fingers pointing forward and the location in the gestures space corresponding to the location on the map (Figure 4a, 4b). Horizontal Map gestures (Figure 4d) differ from Route gestures in their hand orientation (fingers pointing down), and their relative size and location. The label that could explain most of the articulators was assigned to each gesture. Figure 4 shows some examples of participants' path gestures and our coding.

Each of the confederate's verbal descriptions contained two target landmarks, which also appeared along the participant's route. For each of these landmarks it was

determined whether the participant referred to it, and if so whether it was a *literal repetition*, an elaboration or shortening of the confederate's reference (both counted as *partial match*) or a complete *mismatch*. For example, if the confederate said "hoog gebouw" (tall building), "gebouw" (building) and "hoog grijs gebouw" (tall grey building) would be labeled as *partial match* whereas "flat" (apartment building) would be a *mismatch*.

The data of 6 participants were excluded because these participants did not produce any path gestures. This left 40 participants, 10 in each cell. Analysis was done using ANOVA, with factors perspective (levels: Route & Map) and hand shape (levels: Finger & Hand) of the confederate's gestures. The significance threshold was .05 and we used partial eta squared as a measure of effect size.

Results

Neither the confederate's hand shape nor her perspective significantly influenced the total number of path gestures participants produced (M = 5.5, SD = 3.9) and there was no interaction between these factors. We did not find a significant effect of gender, or left or right handedness on the amount or type of path gestures produced. Analysis of the answers to the questionnaire showed no significant effect of condition on how the participants perceived the confederate.

Verbal Alignment

Neither the confederate's perspective (p = .63), nor her hand shape (p = .81) had a significant effect on the number of target nouns repeated by participants (M = 6.2, SD = 1.3), or on the number of partial matches or mismatches.

Effects of the Confederate's Perspective

The confederate's perspective influenced participants' perspective. When the confederate gestured as though on a map, the mean proportion of participants' path gestures in the vertical map perspective was higher (M = .46, SD = .35) than when she gestured as though following a route (M = .11, SD = .20), $F(1, 36) = 14.88$, $p < .001$, $\eta^2 = .29$. Similarly, when the confederate gestured as though following a route, participants produced a higher proportion of gestures with the route perspective (M = .77, SD = .32) than when she gestured as though on a map (M =.52, SD = .39), $F(1, 36) = 12.35$, $p < .001$, $\eta^2 = .14$, see Table 1.

The confederate's gestures' perspective also influenced the hand shape used by participants, $F(1, 36) = 5.00$, $p < .05$, $\eta^2 = .12$. The proportion of gestures with more than one finger extended was higher when the confederate used the route perspective (M = .78, SD = .37) than when she used the map perspective (M = .52, SD = .39), whereas the proportion of gestures with one finger extended was higher when she used the map perspective (M = .48, SD = .39), compared to when she used the route perspective (M = .22, SD = .37), see Table 2.

Table 1: *Means* and *standard deviations* of the proportion of path gestures participants produced from each perspective.

Confederate's Perspective	Prop. Route	Prop. Vertical Map	Prop. Hor. Map
Route	0.77 (.32)	0.11 (.20)	0.12 (.26)
Map	0.43 (.31)	0.46 (.35)	0.11 (.25)

Table 2: *Means* and *standard deviations* of the proportion of path gestures participants produced with each hand shape.

Confederate's Perspective	Prop. Hand	Prop. Finger
Route	.78 (.37)	.22 (.37)
Map	.52 (.39)	.48 (.39)

Effects of the Confederate's Hand Shape

We did not find that the confederate's hand shape influenced participants' hand shape $F(1, 36) = .04$, p = .85, nor that her hand shape influenced the proportion of gestures in the map, $F(1, 36) = .38$, p = .54, or route perspective, $F(1, 36) = .030$, p = .86.

Discussion

We found some of the cross-effects we expected if perceiving gestures would lead to the construction of meaning, which in turn would influence gesture production. The perspective of the confederate's gestures influenced the hand shape of participants' gestures: participants more frequently pointed with one finger if the confederate gestured as though on a vertical map. This can be explained by the confederate's vertical gestures leading participants to think of the route as on a map, which caused them to point with their finger more frequently.

Gestures, like speech, seem to allow for the convergence of representations of meaning across interlocutors. This leads to the question of whether the same representations underlie adaptation in both gesture and speech, and whether these representations can also be influenced by both gesture and speech.

Adaptation in Gesture through Speech

The results of an additional study indeed point in this direction. In this study, the confederate gestured with one finger extended and in the map perspective. Thus, all articulators in gesture suggested a vertically oriented map. Yet we added a condition (N = 10) to the previous study, in which speech also matched this perspective. Rather than using horizontal terms like "rechtdoor" (straight), the confederate now used vertical terms like "naar boven" (up) instead. Note that the first direction was always straight/ up.

When comparing this condition to the Finger/ Map condition with horizontal speech, we found that the perspective of the confederate's speech had an additional effect on the perspective of participants' gestures. With

vertical terms, participants produced a lower rate of gestures with the route perspective (M = .17, SD = .25*)* than with horizontal terms (M = .54, SD = .36), $F(1, 18) = 7.21$, $p < .02$, $\eta^2 = .29$. This supports the notion that semantic representations were converging across interlocutors, rather than surface forms. In addition, it suggests that these representations may be shared between speech and gesture.

Future Work

A limitation of our studies is that the confederate always acted according to a script, and thus was not exactly like a spontaneous partner in forming a conceptual pact. Whereas this usually can be thought of as an interactive process between both interlocutors, the confederate always stuck to her own initial proposal. It would be interesting to see how spontaneous interaction is similar to or differs from this partly staged interaction.

Overall, perspective was repeated more than hand shape. This may be because perspective was expressed in two articulators, whereas hand shape is only one. Thus, the one articulator not matching the constructed meaning may have been adapted to the two matching ones. Another explanation would be that in this task, perspective carried a more important meaning than did hand shape. A vertical gesture cannot possibly depict a route one can walk (at least not in the Netherlands), whereas the distinctions between the different hand shapes are probably far subtler. Therefore, the perspective of gestures may have influenced the construction of meaning more readily than their hand shape. Apparently, in this task, hand shape was not a likely candidate for the type of direct alignment at one level between interlocutors that Pickering and Garrod (2004) proposed. However, in other settings it may very well be. It would be interesting to investigate whether adaptation in hand shape depends on to what extent hand shape carries meaning, and similarly for other articulators. Additionally, other types of gestures (especially non representational gestures) need to be looked at, since they may carry meaning in a way different from the path gestures we studied, and may or may not be linked to semantic representations.

Conclusion

In the adaptation of one interlocutor to another, some hand gestures seem to behave truly like linguistic forms. Whether they are repeated across interlocutors depends on whether their form corresponds to the semantic context (Mol et al., 2009), and the repetition of forms is mediated by mental representations of meaning, rather than being based on a direct perception-action link.

Acknowledgements

We like to thank Susan Brennan and Sotaro Kita for valuable discussions on this work. We also thank Nathalie Bastiaansen for drawing the stimuli, our co-workers at Tilburg University for assisting in the data collection, and the anonymous reviewers.

References

Bock, J. (1986). Syntactic persistence in language production. *Cognitive Psychology, 18*, 355-387.

Branigan, H. P., Pickering, M. J., & Cleland, A. A. (2000). Syntactic co-ordination in dialogue. *Cognition, 75*(2), B13-B25.

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology-Learning Memory and Cognition, 22*(6), 1482-1493.

Cassell, J., McNeill, D., & McCullough, K.-E. (1998). Speech-gesture mismatches: Evidence for one underlying representation of linguistic & nonlinguistic information. *Pragmatics & Cognition, 6*(2), 1-33.

Chartrand, T. L., & Bargh, J. A. (1999). The Chameleon effect: The perception-behavior link and social interaction. *Journal of Personality and Social Psychology, 76*(6), 893-910.

De Fornel, M. (1992). The return gesture. . In P. Auer & A. di Luzio (Eds.), *The contextualization of language*. Amsterdam: John Benjamins.

Ekman, P., & Friesen, W. V. (1969). The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotica, 1*, 49-98.

Kendon, A. (2004). *Gesture: Visible Action as Utterance*. Cambridge: Cambridge University Press.

Kimbara, I. (2008). Gesture form convergence in joint description. *Journal of Nonverbal Behavior, 32*(2), 123-131.

Levelt, W. J. M., & Kelter, S. (1982). Surface form and memory in question answering. *Cognitive Psychology, 14*(1), 78-106.

McNeill, D. (1992). *Hand and Mind: what gestures reveal about thought*. Chicago and London: The University of Chicago Press.

Mol, L., Krahmer, E., & Swerts, M. (2009). *Alignment in iconic gestures: Does it make sense?* Paper presented at the The eight international conference on auditory-visual speech processing, Norwich, United Kingdom.

Parrill, F., & Kimbara, I. (2006). Seeing and hearing double: The influence of mimicry in speech and gesture on observerse. *Journal of Nonverbal Behavior, 30*(4), 157-166.

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences, 27*(2), 169-225.

Tomasello, M., Savage-Rumbaugh, S., & Kruger, A. (1993). Imitative learning of actions on objects by children, chimpanzees and enculturated chimpanzees. *Child Development, 64*, 1688-1705.

Van Baaren, R. B., Holland, R. W., Steenaert, B., & Van Knippenberg, A. (2003). Mimicry for money: Behavioral consequences of imitation. *Journal of Experimental Social Psychology, 39*(4), 393-398.

Vocal Interaction Dynamics of Children With and Without Autism

Anne S. Warlaumont (awarlmnt@memphis.edu)
D. Kimbrough Oller (koller@memphis.edu)
Speech-Language Pathology, University of Memphis, 807 Jefferson Ave.
Memphis, TN 38105 USA

Rick Dale (radale@memphis.edu)
Department of Psychology, University of Memphis, 202 Psychology Building
Memphis, TN 38152 USA

Jeffrey A. Richards (JeffRichards@lenafoundation.org)
Jill Gilkerson (JillGilkerson@lenafoundation.org)
Dongxin Xu (DongxinXu@lenafoundation.org)
LENA Foundation, 5525 Central Ave., Suite 100
Boulder, CO 80301 USA

Abstract

This study examines the temporal and directional characteristics of child-adult vocal exchanges in day-long naturalistic recordings of autism and typical control groups. In both populations, adults responded frequently (on average about 40% of responses) within 1s or less, a time thought to be conducive for contingency learning by the child. However, the time to adult response tended to be longer for the autism population. In the autism group, children also tended to follow more and lead less relative to the control group, as measured by differences in diagonal recurrence profiles computed based on cross recurrence plots. The results inform on the dynamics of naturalistic communicative interaction in normal development and therefore on the social context in which language develops. They also illustrate how large datasets and modern interaction analyses can expand our understanding of differences in children with autism, a population with both social and language deficits.

Keywords: Social interaction; autism; temporal dynamics; cross recurrence; language development; naturalistic recording; response time; social contingency

Introduction

In this paper, we examine fundamental issues related to the fine-grained temporal organization of vocal interaction between children and their social environment, primarily, caregivers. Recent years have seen an abundance of interest in joint action and coordinative processes in both children and adults (Galantucci & Sebanz, 2009). In the current study, we make use of latency response measures as well as the technique of cross recurrence analysis to identify leading and following patterns in the vocal exchanges between children and adults. We find a distinct signature of leading in normal children and find that a distinct breakdown of this signature is identifiable in children with autism. These results show that analysis of naturalistic recordings may reveal socio-dynamic indicators of at-risk children. We conclude with a brief discussion of the relevance of our findings to models of language acquisition in normal and disordered individuals.

Interaction and Contingency in Language Development

The fact that language learning occurs in a dynamic and interactive social context is becoming increasingly appreciated. Children are not passive information processors nor do they learn language purely on the basis of contingent reinforcement. They are, rather, actively engaged in perceptual learning and responding to communicative acts produced by others as well as being engaged in behavioral and motor exploration for which they at least sometimes receive feedback in the form of communicative response by adults and other children in their environment.

For example, in a video-recording study performed in the participants' homes, Keller et al. (1999) found that mothers often respond within one second to their three-month-old infants' communicative acts. Relating this to the fact that one second had been previously shown to be about the amount of time within which a contingent response must occur in order for the infant to detect that contingency, the authors concluded that mothers' communicative responses to their infants' communicative attempts tend to occur within the necessary window of time for the infant to perceive them as contingent. In other words, their results support the notion that caregivers' responses to their children support infant communicative development by serving as contingent reinforcers for the infant's own communicative acts.

In a more recent study, Gros-Louis et al. (2006) observed naturalistic interactions in a laboratory setting and found that mothers responded contingently to their infant's vocalizations over 70% of the time and that the type of response they gave depended on the phonological characteristics of the infant's vocalizations. Furthermore, Goldstein, King, and West (2003) and Goldstein and Schwade (2008) have found experimentally that mothers' contingent responses do appear to shape the infant's speech-

related vocal development as measured through follow-up tests.

Recently, cross recurrence analysis of time series has allowed for additional quantitative measures of interactive contingency to be measured in naturalistic child-caregiver interaction. For example, patterns of leading and following by interlocuters can be examined at multiple timescales concurrently. Dale and Spivey (2006) examined diagonal cross recurrence profiles calculated on syntactic patterns (specifically, part of speech bigrams) for three well-known child-caregiver conversation corpora. They found individual differences among the three children in their tendency to lead versus follow their caregiver. Abe (Kucjaz, 1976), who had the most advanced language out of the three children also had the greatest tendency to lead rather than follow the caregiver. This work lays foundations for application of cross recurrence analysis to other vocal interaction phenomena, to larger naturalistic datasets, and, as carried out here, to the study of populations with autism.

Autism Spectrum Disorders (ASD)

Impaired social interaction and language learning are two components of the DSM autism diagnostic criteria. With regard to social interaction, children with ASD have exhibited differences in initiation, turn-taking, imitation, and joint attention behaviors.

In recent years, technology has become available to permit day-long naturalistic recording of infant's acoustic environments, including their own vocalizations and the speech and other environmental sounds in the infant's vicinity. Warren et al. (2009) evaluated social interaction in all-day recordings (5,256 hours over 438 sessions) in ASD and control groups. The authors discovered differences between typically developing and autistic children in the frequencies of both conversational turns and child vocalizations. These results, based on summary measures, encourage analysis at a more fine-grained level of temporal detail in order to address such issues as the directionality of the conversational exchanges and temporal characteristics of adult-child interactions. Both latency to response and diagonal cross recurrence profiles can be automatically calculated, making them suitable for application to large-scale naturalistic recordings.

This Study

In the present study, we first looked at response latencies in a way that was similar to Keller et al. (1999). However, we evaluated much more data and used more naturalistic recordings (collected at home, daycare, and therapy as opposed to only at home in a single post-sleep, post-feeding context with experimenters present and videotaping). Other differences are that we looked at the vocal modality only, and that we evaluated age, autism, gender, and maternal education as predictive factors. We also applied cross recurrence analysis to the data and investigate leading and following tendencies in the recordings. The application of this method with large-scale recordings of adult-child speech is unique as is its application to the autism population.

Method

Participants

The participant recruitment, recordings and the automated labeling of them were conducted as part of previous studies. Warren et al. (2009) provide more detailed information on the procedures. The present study includes data from 26 children between 16-48 months who have been diagnosed under the classic autism subtype except for two who received Pervasive Developmental Disorder-Not Otherwise Specified subtype diagnoses; documentation of ASD diagnoses by trained professionals was provided by the children's parents. No child was reported to have a diagnosis that included echolalia (pathological repetition of previously heard speech). The study also includes data from 78 typically developing (TD) children who were selected form a larger normative database such that for each child with ASD there were three TD controls of the same gender and socioeconomic status (SES), as measured by the mother's education level, and collectively across the three controls spanning the same range of ages as the ASD child.

Recording

Recordings were made using LENA digital language processor devices. These recorders fit into a pocket sewn into the front of custom-designed clothing and record a single channel of audio for up to 16 hours at a time. The device records the child's voice as well as other sounds within approximately a 6-10' radius of the child. Parents were mailed the devices and were instructed to begin recording when the child awoke in the morning and left the recorder on throughout the day. Recordings contexts included the home, preschool, and speech-language therapy. There were 438 recordings in total, each lasting at least 12 hours. The present study is thus based on over 5,256 hours of naturalistic recording.

Automated Labeling

Each recording was processed using the professional version of the LENA analysis software. The software analyzes and time segments the entire recording according to the likely source of the signal, e.g., the child wearing the recorder, another child, an adult, a television or radio, silence; every part of the recording is given a label. Within child segments it also labels some sub-segments (termed *vocalizations* by the system) as speech-like or as cry/vegetative/fixed. Reliability for the automated labeling compared to human raters on TD child recordings is approximately 82% correct for adult speaker, 76% correct for key child, 75% correct for child speech-like, and 84% for child cry/vegetative/fixed (Xu et al., 2009). The software allows for exporting these sound source and child vocalization type segmentations along with other information in XML format.

We developed a set of Perl scripts to extract the specific information of interest for this study from the XML files (exported as *.its* files by the LENA software). Specifically, we extracted the start and end times of each segment labeled with relatively high confidence as coming from the child wearing the recorder (*child near*, *CHN*, segments, labeled as such because they fell near the maximum of the Gaussian mixture model that gave the likelihood that a segment was produced by the child) and of each segment labeled as coming from an adult with relatively high confidence (*female adult near*, *FAN*, or *male adult near*, *MAN*). Note that loudness and nearness to the child increase the confidence of segment coding and therefore increase the likelihood of a sound being included in the present study. Also, there were minimum duration thresholds for each segment label type; thus, a long string of vocalization by the same speaker could only be split if there was an intervening silence, TV, or other-speaker vocalization meeting the minimum duration requirement. We also identified child segments that contained only speech-like sub-segments as well as those that contained only cry/vegetative sub-segments. All the subsequent response time and cross recurrence calculations were made using only those child segments that contained no cry/vegetative sub-segments and at least one speech-like sub-segment.

Response Time Analysis

We developed a set of programs written in Perl and R that automatically extracted and calculated response time information from the speaker labels. First, adult response times were calculated according to the following procedure. For each child segment, we determined whether an adult segment followed without any child segment intervening. Other sound source labels were permitted to intervene between the child segment and the subsequent adult segment. Then the time between the offset of the child segment and the onset of the adult segment was calculated. Child response times were calculated in exactly the same manner, except with the speaker labels reversed. Based on these response times, the median adult response time and the median child response time were calculated as well as the proportion of adult responses occurring within 1s and the proportion of child responses occurring within 1s.

Cross Recurrence Analysis

Cross Recurrence Plots Cross recurrence plots (Marwan et al., 2007; Richardson et al., 2007) are matrices that indicate correspondence or lack of correspondence for every possible combination of events or times in one event time series and the events or times in another series. In our case, the vertical dimension of the matrix corresponds to the presence and absence of child segments and the horizontal dimension of the matrix corresponds to the presence and absence of adult segments (Fig. 1). Each element in the plot matrix is assigned a value of 1 (marked in black in the figure) if there is a child segment at the row corresponding to the element's row number as well as an adult segment at the segment corresponding to the element's column number.

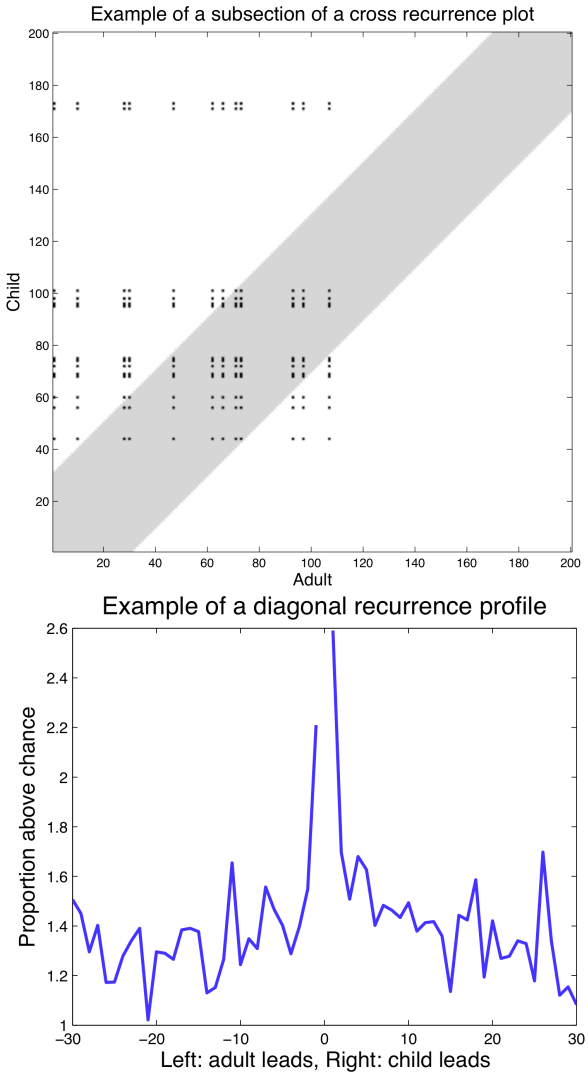


Figure 1: On the top is the cross recurrence plot for the first 200s of one of the recordings. The gray region indicates the portion from which the diagonal recurrence profile is calculated. On the bottom is the diagonal recurrence profile for the entire recording.

The time series that were used for making these charts were broken into 1s chunks. When either the child or an adult was speaking, a 1s chunk was coded as 1 in the speaker's series and as null value in the non-speaker's series. Regardless of the actual length of the segment, it was coded as lasting 1s so that long speaker segments would not be treated as having long lags to segments by the other speaker. When neither child nor adult were speaking, both time series were coded as null values for the duration of the no-speaker time, rounded down to the nearest second. The recurrence plot is square since both the vertical and the horizontal dimensions have length equal to the total number of 1s chunks in the recording.

Diagonal Cross Recurrence Profiles A number of measures, such as recurrence rate, determinism, etc. can be extracted directly from the cross recurrence plot (Marwan et al., 2007). However, in this study we focused on measures that were extracted from the plot's diagonal recurrence profile (explained below) after it had been derived from the recurrence plot. In the physical sciences, this is sometimes referred to as the recurrence probability or the recurrence spectrum (Marwan et al., 2007). Richardson and Dale (2005) and Richardson et al. (2007) have used this measure in analyses of linguistic coordination. It can be interpreted as a lag profile that reflects co-occurrence patterns between utterances at varying relative lags. We provide some further description here.

The diagonal on the recurrence plot running from the origin to the final event on both axes reflects when the child and caregiver are speaking at the same time. Sometimes this main line is referred to as the "line of synchronization," since any points on this line reflect matching on/off states for the child and the adult(s). However, since the automatic labeling procedure does not allow overlapping speaker labels, there will never be a match along this diagonal.

The next diagonal line just below-right of the primary diagonal contains the matches between the child's on/off states and those of the adult series one step into the future. In other words, the elements of this adjacent below-right diagonal are given a point on the plot when a given child segment was immediately followed by an adult segment (i.e., the adult spoke one time step later during the interaction). Conversely, the elements of the adjacent above-left diagonal line have a point when a given adult segment was immediately followed by a child segment Moving to diagonals further below-right or above-left give indication of when the adult followed the child at larger lags and when the child followed the adult at larger lags, respectively.

For each diagonal line parallel to the primary diagonal, the number of 1's can be added and divided by the total number of elements in that diagonal to give the proportion cross recurrence for the speaker order and lag amount corresponding to that diagonal. These proportions can then be plotted to create a diagonal recurrence profile (Fig. 1). By randomly shuffling the speaker labels and recalculating the diagonal recurrence profile, and by repeatedly doing this and averaging across the shuffled label profiles, one can obtain a bootstrapped estimate of the baseline diagonal recurrence profile that would be expected if were no systematic leading-following relationship between the speakers. Dividing the actual diagonal recurrence profile by the baseline estimate gives a normalized diagonal recurrence profile that represents proportion above chance leading/following tendencies.

In this study, we measured three characteristics of the normalized diagonal recurrence profile for a given recording. The first was the height of the profile at the point immediately right from center. This gives an indication of how often the adult vocalized immediately after a child vocalization. The second was the height at the point immediately left of center; it tells how often the child's vocalizations immediately followed the adult's. The third measure is the ratio of the sum of values on the right side of the profile (which is higher when the adult tended to follow the child) to the sum of values on the left side of the profile (higher when the child tended to follow the adult). This gives a measure of the general balance between leading and following across the two speakers.

Results

Response Time Results

The adult response times and child response times for the ASD and control groups are plotted as averaged histograms in Figure 2.

For each of the four response time independent measures (adult median response time, adult proportion within 1s, child median response time, and child proportion within 1 s) we ran a mixed model regression with participant ID as a random effect and ASD status, age in weeks, gender, and mother's education level (a measure of the family's socioeconomic status) as fixed effects.

Adult median response time was significantly longer for children with ASD (M = 2.32s, SD = 1.22) than for the controls (M = 1.65s, SD = 0.78), $p < 0.001$, $\beta = 0.331$, and was significantly shorter as maternal education increased, $p < 0.001$, $\beta = -0.234$. Adult proportion of responses within 1s was significantly smaller for children with ASD (M = 0.37, SD = 0.10) than for the controls (M = 0.43, SD = 0.08), $p < 0.001$, $\beta = -0.348$, and was significantly larger as maternal education increased ($p < 0.001$, $\beta = 0.284$).

Child median response time was longer for children with ASD (M = 2.70s, SD = 1.38) than for the controls (M = 2.37s, SD = 0.92) though this did not reach statistical significance, $p = 0.063$, $\beta = 0.161$, and was significantly shorter as maternal education increased, $p = 0.016$, $\beta = -0.153$. Child proportion of responses within 1s was significantly larger as maternal education increased, $p = 0.010$, $\beta = 0.174$.

Age and gender did not significantly predict any of the four independent variables.

Cross Recurrence Results

The averaged diagonal recurrence profiles for the ASD and control groups are plotted in Figure 3. As with the response time measures, each of the three dependent variables (height immediately right of center, height immediately left of center, and ratio of right side to left side) was regressed on participant ID as a random effect and ASD status, age in weeks, gender, and mother's education level as fixed effects.

The height at the point immediately right of center was only significantly predicted by age, decreasing as age increased, $p < 0.001$, $\beta = -0.228$.

The height at the point immediately left of center, which represents the frequency with which the child immediately

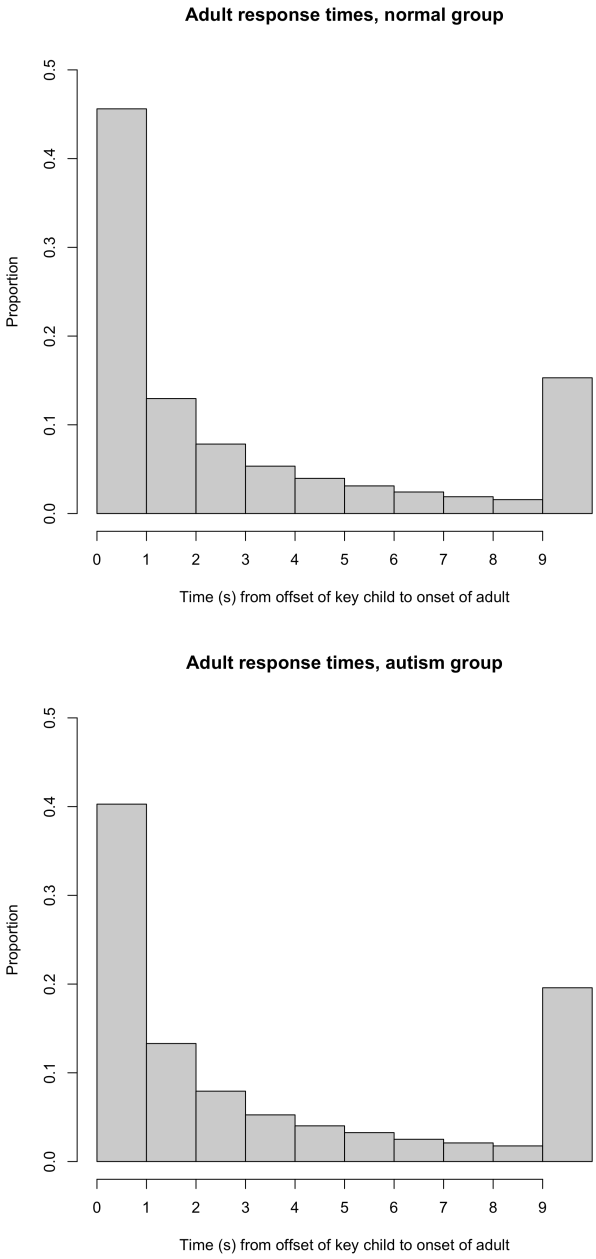


Figure 2: Histograms of adult response latencies for children with and without ASD.

followed the adult, was significantly higher for the ASD group (M = 1.73, SD = 0.91) than for the control group (M = 1.53, SD = 1.09), $p < 0.001$, $\beta = 0.207$. The height at this point was also significantly lower as maternal education increased, $p = 0.017$, $\beta = -0.183$, and was lower as age increased, $p = 0.002$, $\beta = -0.214$.

The ratio of the right side (from lag 1 through lag 10) to the left side (from lag 1 through lag 10) of the diagonal cross recurrence profile was smaller for the ASD group (M = 1.06, SD = 0.18) than for the TD group (M = 1.25, SD = 0.30), $p < 0.001$, $\beta = -0.398$, indicating that the general tendency for the child to lead and for the adult to follow was lessened in the autism group. A small but significant increase was accounted for by age, $p = 0.01$, $\beta = 0.136$.

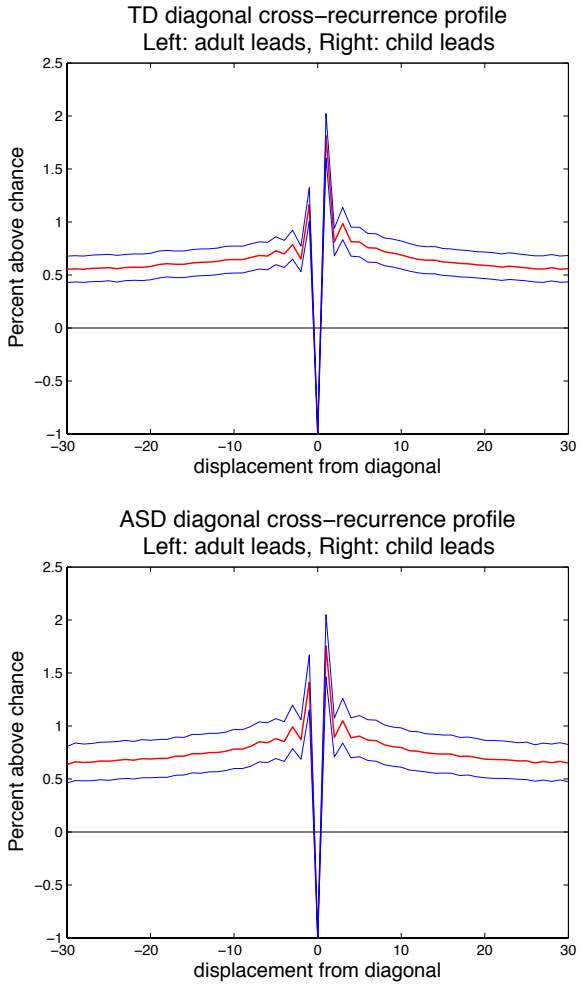


Figure 3. Diagonal recurrence profiles averaged across all recordings in the TD group (top) and all recordings in the ASD group (bottom). In each profile, the red line indicates the mean values across recordings. Blue lines indicate 95% confidence intervals. Displacement from diagonal is in seconds.

Discussion

This study provides new information from automated analysis over large naturalistic recordings in support of the idea that social interaction is impaired in ASD. Interestingly, the strongest trend concerned the adult's responses to the child rather than the child's responses to the adult. There were differences in both the dynamics and the directionality of adult-child interaction in ASD. The length of time before an adult responded to an ASD child's speech or speech-like vocalization was larger in ASD than for TD children with a smaller percentage of responses occurring within the 1s window considered ideal for contingency detection. In addition, ASD children's speech and speech-like vocalizations were more of a tendency to follow the adult vocalizations than TD children's.

The shift of the balance toward child following (and adults leading) and increased latency of adult responses to the child when they did occur could be due to less initiation of communication on the part of the ASD children and/or to

reduced communicative content or other deficiencies in the vocalizations of children with ASD. It could also be due to adults' reduced attentiveness to the vocalizations of children with the disorder. This pattern of following vs. being followed may have feedback effects on the child's language development, reducing the quality of the contingency-based input available to the child with ASD as they acquire speech, language, and other communication skills. At present, there are very few computational models that attempt to capture the interplay among cognitive agents in a realistic way (one exception may be language evolutionary models; see Cangelosi & Parisi, 2002, for examples). The dynamic interplay between cognitive agents during development, such as speech-contingency patterns, may produce feedback loops that substantially impact learning within an individual system.

The present work is relevant to theoretical, including computational, modeling of speech-language development. Language learning occurs in the context of social interactions during which the child hears what other speakers say but also receives contingent reinforcement for their own vocalizations. Understanding the typical dynamics of these interactions may help guide the development of models that take into account the dynamic interactive social context of language learning. They may also help inform models of autism. Some of the deficits present in autism may be the result of a negative feedback loop in which children with autism produce fewer or lower-quality conversation initiations, leading to adults' responding with lower frequency and more latency, which in turn leads to poorer learning of language and communication-related skills by the child.

From a practical standpoint, measures of conversational dynamics, both at short and long timescales could potentially be applied for early identification of autism or other communicative disorders. Being a disorder that involves profound social and cognitive impairments, differences in patterns of communicative interaction, such as in leading-following and elicitation of quick responses, might indicate risk for autism. For example, automatically computed interaction-based measures (such as the ones used in the present study) could supplement the acoustic measures used in an existing autism screening tool (Xu et al., 2009)

Acknowledgments

Thanks go to the LENA Foundation, particularly to Terry and Judith Paul, for funding and executing the development of the LENA recording and automated labeling system as well as the autism recordings database used in this study. Author ASW was supported by a Department of Energy Computational Science Graduate Fellowship. DKO was supported by the Plough Foundation.

References

Cangelosi, A., & Parisi, D. (2002). *Simulating the evolution of language*. London: Springer.

Dale, R. & Spivey, M. J. (2006). Unraveling the dyad: using recurrence analysis to explore patterns of syntactic coordination between children and caregivers in conversation. *Language Learning*, *56*, 391-430.

Galantucci, B., & Sebanz, N. (2009). Joint action: current perspectives. *Topics in Cognitive Science*, *1*, 255-259.

Goldstein, M. H., Kin, A. P., & West, M. J. (2003). Social interaction shapes babbling: testing parallels between birdsong and speech. *Proceedings of the National Academy of Sciences of the United States of America*, *100*, 8030-8035.

Goldstein, M. H., & Schwade, J. A. (2008). Social feedback to infants' babbling facilitates rapid phonological learning. *Psychological Science*, *19*, 515-523.

Gros-Louis, J., West, M. J., Goldstein, M. H., & King, A. P. (2006). Mothers provide differential feedback to infants' prelinguistic sounds. *International Journal of Behavioral Development*, *30*, 509-516.

Keller, H., Lohaus, A., Völker, S., Cappenberg, M., Athanasios, C. (1999). Temporal contingency as an independent component of parenting behavior. *Child Development*, *70*, 474-485.

Kuczaj, S. (1976). –ing, -s, and –ed: a study of the acquisition of certain verb inflections. Unpublished doctoral dissertatio, University of Minnesota.

Marwan, N., Romano, M., Thiel, M., & Kurths, J. (2007). Recurrence plots for the analysis of complex systems. *Physics Reports*, *438*, 237-329.

Richardson, D. C. & Dale, R. (2005). Looking to understand: The coupling between speakers' and listeners' eye movements and its relationship to discourse comprehension. *Cognitive Science*, *29*, 39-54.

Richardson, D. C., Dale, R., & Kirkham, N. (2007). The art of conversation is coordination: common ground and the coupling of eye movements during dialogue. *Psychological Science*, *18*, 407-413.

Warren, S. F., Gilkerson, J., Richards, J. A., Oller, D. K., Xu, D., Yapanel, U., & Gray, S. (2009). What automated vocal analysis reveals about the vocal production and language learning environment of young children with autism. *Journal of Autism and Developmental Disorders*. Advance online publication. doi: 10.1007/s10803-009-0902-5

Xu, D., Richards, J. A., Gilkerson, J., Yapanel, U., Gray, S., Hansen, J. (2009). Child vocalization composition as discriminant information for automatic autism detection. *Proceedings of the 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (pp. 2518-2522). Minneapolis, MN: IEEE.

Xu, D., Yapanel, U., & Gray, S. (2009). *Reliability of the LENA^{TM} Language Environment Analysis System in young children's natural home environment* (LENA Foundation Technical Report LTR-05-02). Retrieved from *http://www.lenafoundation.org/TechReport.aspx/Reliability/LTR-05-2*

Virtually accommodating: Speech rate accommodation to a virtual interlocutor.

Laura Staum Casasanto[1] (laura.casasanto@mpi.nl)
Kyle Jasmin[1] (kyle.jasmin@mpi.nl)
Daniel Casasanto[1,2] (daniel.casasanto@mpi.nl)

[1]Max Planck Institute for Psycholinguistics, Neurobiology of Language Group, Nijmegen, NL
[2]Donders Center for Brain, Cognition, and Behavior, Radboud University, Nijmegen, NL

Abstract

Why do people accommodate to each other's linguistic behavior? Studies of natural interactions (Giles, Taylor & Bourhis, 1973) suggest that speakers accommodate to achieve interactional goals, influencing what their interlocutor thinks or feels about them. But is this the only reason speakers accommodate? In real-world conversations, interactional motivations are ubiquitous, making it difficult to assess the extent to which they drive accommodation. Do speakers still accommodate even when interactional goals cannot be achieved, for instance, when their interlocutor cannot interpret their accommodation behavior? To find out, we asked participants to enter an immersive virtual reality (VR) environment and to converse with a virtual interlocutor. Participants accommodated to the speech rate of their virtual interlocutor even though he could not interpret their linguistic behavior, and thus accommodation could not possibly help them to achieve interactional goals. Results show that accommodation does not require explicit interactional goals, and suggest other social motivations for accommodation.

Keywords: Conversation; interaction; accommodation; alignment; virtual reality.

Introduction

Why do people accommodate to each other's linguistic behavior? Studies of multi-party interactions both in the laboratory and in natural conversation have suggested that two or more speakers in a conversation tend to align their linguistic behavior along several dimensions: lexical choice (Barr & Keysar, 2002; Bortfeld & Brennan, 1997; Niederhoffer & Pennebaker, 2002), phonetics (Alim, 2004; Pardo, 2006), and syntactic constructions (Gries, 2005), among others. The fact that accommodation occurs is well accepted, but the motivations for this convergence between speakers have been a matter of debate.

Studies of natural conversational interactions have identified social/interactional factors that influence how much speakers accommodate (Giles, Taylor & Bourhis, 1973). Based on these studies, it appears that a speaker accommodates towards or away from their interlocutor to achieve interactional goals: to make one's interlocutor do, think, or feel things. This can involve conveying social information, such as information about their social stances toward their interlocutor or toward a group that their interlocutor belongs to (Coupland, 1985). Accommodation could also help to coordinate joint actions being negotiated through conversation (Brennan & Clark, 1996).

But are immediate social motivations necessary to make speakers accommodate? Or might speakers accommodate even in the absence of a desire to achieve direct interactional goals? Mechanistic theories of dialogue (Pickering & Garrod, 2004) offer one possible alternative. Automatic alignment processes could account for convergence in linguistic behavior. That is, speakers might use similar linguistic forms to those used by their interlocutors because these forms are highly active and thus have an advantage over alternatives in the selection process.

On another alternative, accommodation could be a consequence of a speaker's attempt to achieve longer-term social goals: accommodation could be part of how speakers develop the linguistic styles that they use to communicate social information about themselves to others and to indicate their membership in social groups. This could occur in tandem with or independent of interaction-specific social goals.

Distinguishing these alternatives is difficult because in real-world conversations, interactional motivations are ubiquitous; in any conversation between two real people, the interlocutors may have social goals and relationships that could be influencing their linguistic behavior. Experimenters have attempted to deal with this complexity in a few ways. Experiments using pre-recorded speech in repetition paradigms have uncovered alignment between a speaker and a recording (Babel, 2009). However, because these experiments do not involve conversation, it's difficult to know whether the same mechanisms underlie speakers' production in these tasks and their accommodation in conversations.

In other experimental paradigms, the conversational setting is retained by using a confederate (Hannah & Murachver, 1999). However, no human confederate can entirely prevent his or her speech from being influenced by the naïve participant's speech. Introducing a confederate means losing experimental control over precisely those social and linguistic variables that might matter the most. This makes it difficult to assess the extent to which accommodation on the part of the participant depends upon their own interactional motivations, and to what extent it is a response to their interlocutor's behavior.

Virtual Reality (VR) provides an opportunity to engage participants in a conversational interaction with an interlocutor whose speech is not influenced by their speech, and can be varied systematically along a single dimension.

Moreover, a virtual interlocutor cannot feel or think at all, so participants cannot hope to influence the thoughts or feelings of the virtual interlocutor by accommodating to him. What happens in a conversational situation where interactional goals cannot be achieved? Do speakers still accommodate when their interlocutor cannot interpret their accommodation behavior? And if so, are they motivated by other, longer-term social goals, or is it a fully automatic process that is independent of social factors?

To find out whether speakers accommodate in a conversation with someone who cannot interpret their accommodation, we asked participants to enter an immersive VR environment and to converse with a virtual interlocutor, VIRTUO. While accommodation could theoretically occur along many dimensions at once, this experiment focused on the single dimension of speech rate, because this was easily manipulated in the virtual interlocutor. We varied VIRTUO's speech rate between participants to see whether participants would adjust their own speech rate to better match the rate at which their virtual conversational partner was speaking.

If immediate interactional goals motivate accommodation, then speakers in a conversation with VIRTUO should not accommodate to his speech rate, because they cannot hope to influence his thoughts, feelings or behavior by accommodating to him. If speakers *do* accommodate to VIRTUO by adjusting their speech rate towards his, there are two possible explanations: either accommodation occurs entirely automatically, or it can be motivated by social goals with a locus outside the current interaction (i.e., long-term social goals). To distinguish between these possibilities, we administered a post-experiment questionnaire investigating how participants judged VIRTUO on relevant social dimensions. If participants' judgments of VIRTUO correlate with their degree of accommodation to him, then this suggests that social goals beyond the level of the individual conversation influence accommodation.

Methods

Participants

Members of the Radboud University community (N=62, 30 male) participated in exchange for payment. Participants were all native speakers of Dutch between the ages of 17 and 28.

Materials

VIRTUO's speech was pre-recorded by a male native Dutch speaker reading in a conversational tone from a script of statements and questions designed to simulate a conversation about products in a grocery store. The speed of the original recordings was manipulated without changing the pitch of the recordings using the "change speed" function in the audio manipulation software package Audacity, which removes or replicates short intervals of the acoustic signal in order to extend or shorten the overall length of a sound clip. Participants in the FAST condition heard these recordings sped up by 12%, and those in the SLOW condition heard them slowed down by 12%. Both sets of recordings remained within the range of possible speaking rates of a Dutch speaker, but the two rates were noticeably different.

The virtual environment (VE) was a supermarket, which was custom-designed for this experiment using Adobe 3ds Max 4. The virtual supermarket consisted of a single long aisle with shelves on both sides, stocked with products, providing a variety of items for VIRTUO to inquire about.

The experiment was programmed and run using WorldViz's Vizard software. Participants wore an NVIS nVisor SX60 head-mounted display (HMD), which presented the VE at 1280x1024 resolution with a 60 degree monocular field of view. Mounted on the HMD was a set of 8 reflective markers linked to a passive infrared DTrack 2 motion tracking system from ART Tracking, the data from which was used to update the participant's viewpoint as he moved his head. Sounds in the VE, including the voice of the avatar, were rendered with a 24-channel WorldViz Ambisonic Auralizer System. The sound system was supplemented by 4 floor shakers mounted on a raised platform. These produced vibrations that contributed to an illusion of motion as participants were driven through the supermarket by VIRTUO in a specially modified virtual golf cart.

VIRTUO was represented by a stock avatar produced by WorldViz. The avatar's appearance suggested that he was a Caucasian male in his mid-twenties (the average age guessed by participants in debriefing was 26 years old), which matched the age of the Dutch speaker who recorded his speech.



Figure 1. VIRTUO in the virtual supermarket, from the perspective of a participant. The arrow indicates the next item that VIRTUO and the participants should discuss (here, ketchup). The steering wheel of the virtual golf cart is visible in the bottom left corner.

Procedure

Prior to entering the VE, participants were told that they would be having a conversation with VIRTUO, a virtual

agent who wanted to learn more about the human world. They entered the VE by putting on the HMD, which showed them a virtual supermarket. When participants moved their heads, the display changed, so they could explore the virtual world by looking around. Participants remained seated on a chair throughout the experiment. They traveled through the virtual supermarket in a virtual golf cart with VIRTUO in the drivers' seat, so there was no need for participants to walk in order to move down the aisle of the grocery store.

Participants were randomly assigned to the Fast or Slow speech condition automatically by the experiment program, so that the experimenter was not aware of which condition participants would be in until the experiment had begun. This minimized the possibility of experimenter expectancy effects influencing participants' speech rate before they spoke with VIRTUO. Once the experiment began, all instructions were written; therefore participants did not have any verbal interaction with the experimenter, which could have influenced their speech rate.

The experiment consisted of a Baseline block of trials followed by a Conversation block. During the Baseline trials, participants were alone in the VE, and had an opportunity to get accustomed to their surroundings. We collected a sample of speech during this time to use as a Baseline speech rate. To elicit speech, we gave participants written instructions (via the HMD) to look at 4 of the products on the shelves in front of them, one at a time, and describe each product briefly.

After the four Baseline trials, participants met VIRTUO, who introduced himself in a few sentences. VIRTUO then took participants on a tour of the grocery store, stopping at six products (bananas, ketchup, light bulbs, toothpaste, cat food, and beer) to ask them three or four questions about each one. Participants responded with information about the identity of the products, what they were made of, how they are used in the human world, etc. Participants' speech was recorded through a microphone suspended from the HMD.

VIRTUO's speech behavior created a conversational setting, but he did not have the ability to understand or flexibly respond to participants' utterances. The experimenter listened to participants' responses from a control booth, and pressed a button to advance VIRTUO to the next utterance in his script. VIRTUO's speech began after a random delay between 150 and 400 ms, so that the experimenter's button-pressing (i.e., turn-taking behavior) could not directly influence the speech rate of the participant. If the next item in VIRTUO's script did not constitute a sensible response to something a participant said, the experimenter pressed a button that caused VIRTUO to say that he did not understand, and that they should move on.

Speech rate (in words per second) was calculated by transcribing participants' speech and marking the boundaries of their utterances as intervals in the audio and video transcription and coding software ELAN, then dividing the number of words transcribed by the number of seconds in the interval. Each participant's speech rate during the Conversation block was compared to their own Baseline rate.

Results and Discussion

Participants were assigned randomly to the two speed conditions, resulting in 33 participants in the Fast condition and 29 participants in the Slow condition. Mean speech rates during Baseline and Conversation periods are shown for participants assigned to the Fast and Slow conditions in Figure 2. Results indicate that VIRTUO's speech rate influenced how fast the participants spoke during their Conversation with him. Participants in the Fast condition spoke significantly faster during their Conversation with VIRTUO than during their Baseline measurement ($t(1,32)=4.02$, $p=.0003$), and significantly faster than participants in the SLOW condition ($t(1,60)=2.24$, $p=.03$), whose Conversational speech rate did not differ from their Baseline rate ($t<1$). This resulted in the predicted interaction between Condition (Fast, Slow) and Measurement (Baseline, Conversation; $F(1,60)=4.36$, $p=.04$; Figure 2).

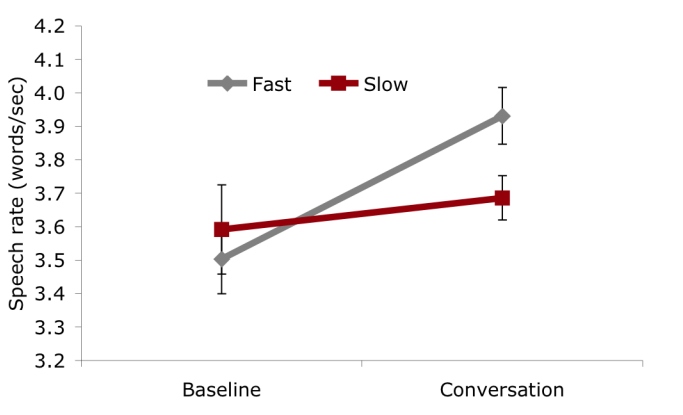


Figure 2. Experimental results. Speech rate differed between the Fast and Slow conditions during the Conversation period but not during the Baseline period.

The fact that the Baseline speech rate did not differ between conditions confirms that speaking with the experimenter prior to the experiment did not differently influence the speech of the Fast and Slow groups; rather, the differences that appeared in the Conversation period were a result of speaking to VIRTUO.

Participants in the SLOW condition did not speed up significantly, but there was a trend toward speaking faster in the Conversation condition than in the Baseline condition among these participants. This occurred despite the fact that their virtual interlocutor spoke slower than even their Baseline rate on average throughout their Conversation (VIRTUO spoke at 3.31 words per second on average in the Slow condition, and 4.20 words per second in the Fast condition). The slight increase in speed from Baseline to Conversation even among participants in the Slow condition suggests that while participants were influenced by VIRTUO's speech rate, they also may have been influenced

by other factors, such as increasing "immersion" in the virtual world (Heeter, 1992). This may have counteracted the effects of VIRTUO's speech rate on participants in the Slow condition. Importantly, the critical interaction between Condition and Measurement indicates that VIRTUO's speech rate affected participants' speech rate above and beyond any unexpected task-related effects.

To find out how early in the Conversation period accommodation could be detected, we first conducted an analysis comparing participants' speech rate in the Baseline period to their speech rate during their responses to VIRTUO's questions about the first item they discussed in their Conversation.

Results of this analysis suggest that speech rate can be adjusted quite quickly; a comparison of the Baseline speech rate with participants' speech in just the first item yields a significant interaction between Measurement and Condition ($F(1,58)=4.82$, $p=.03$), indicating that speakers in the Fast condition had already sped up more than speakers in the Slow condition over the course of the first 4 question-answer pairs. These results suggest that accommodation occurred rapidly and did not develop slowly over the course of the experiment.

This might seem surprising given the widespread assumption that accommodation is a process that occurs gradually over time. However, there are many respects in which speakers adjust to their interlocutors in the beginning of a conversation quite immediately; for example, when a Spanish-English bilingual speaker is approached by a stranger who begins to speak English to them, they are very likely to speak English in response immediately. There is no period of gradual adjustment in choice of language. Similarly, if someone begins a conversation with a friend in a sorrowful tone of voice, the friend is unlikely to respond back in a chipper tone; they will adjust to the emotional tone of the conversation immediately, without requiring a period of gradual change.

Participants' speech rate in response to VIRTUO's questions did not relate to the position in the experiment where the question appeared. This is consistent with rapid change immediately after meeting VIRTUO; perhaps speech rate accommodation does not occur gradually over time, but instead happens early in a conversation and is maintained fairly consistently throughout the interaction. However, the fact that accommodation did not increase over time in this experiment must be interpreted with caution, because the order of VIRTUO's questions was fixed rather than counterbalanced across subjects. Differences in content between the questions might have influenced participants' speech rate, which could have obscured any possible effects of the passing of time.

According to the questionnaire participants filled out after they finished the VR portion of the experiment, speakers accommodated more to VIRTUO when they judged themselves to be more similar to him ($r=.25$, $p=.05$). This correlation suggests that in addition to whatever automatic mechanisms might cause accommodation, people accommodate more to an interlocutor they identify with for longer-term social reasons (a point we will return to below).

General Discussion

The results of this experiment indicate that participants accommodated to the speech rate of their virtual interlocutor. Participants who spoke to a fast-talking VIRTUO sped up significantly from their Baseline speech rate, and spoke significantly faster than their counterparts in the Slow condition. This was true even though VIRTUO could not interpret participants' linguistic behavior, and thus accommodation could not possibly help them to influence VIRTUO's thoughts or behavior. Why, then, did participants accommodate to VIRTUO?

On one possibility, participants accommodated to VIRTUO through fully automatic mechanisms, without any social component. This would be consistent with studies showing alignment between a speaker and non-conversational speech (Babel, 2009; Goldinger, 1998), and might suggest that social motivations are unnecessary to cause accommodation in conversation.

However, results of the post-test support the idea that long-term social goals may be a factor that drives accommodation. Speakers may have social goals that extended beyond their current interaction. For example, some participants may have been motivated to accommodate to VIRTUO by a general tendency to speak similarly to other speakers, especially those that they can identify with to some extent.

Accommodation to certain interlocutors may be one of the mechanisms by which speakers develop a coherent linguistic style over a longer time scale, perhaps even playing a critical role in sound change (Niedzielski & Giles, 1996; Auer & Hinskens, 2005). Selective accommodation, to people with the right social characteristics, may help speakers speak in a way that reflects the way their in-group members speak.

The tendency to speak more like someone who one identifies with is fundamental to the organization of linguistic variation. Linguistic behavior at many levels, including phonetics, word choice, and choice of syntactic constructions, is subject to variation – there are many possible ways for a speaker to communicate approximately the same thing. Although this variation can seem random, these choices can often be predicted by a speaker's membership in various social groups. These groups can correspond to macrosociological categories (e.g. gender, age, ethnicity), or they can be locally defined (e.g. communities of practice) (Eckert & McConnell-Ginet, 1992). But how do these relationships between social group membership and linguistic behavior get established?

Individuals' linguistic behavior must be influenced by the behavior of others whom they consider to be in-group members in order for social groups to become correlated with linguistic behavior. The mechanisms underlying this process of sociolinguistic differentiation and identification are not entirely well-understood; however, they must

operate on the level of actual language use, i.e. individual conversations.

There is another way that social motivations could have influenced accommodation even in a virtual interaction. Perhaps people accommodated to VIRTUO because participants were somehow confused into thinking that he can interpret their social behaviors. VIRTUO does, after all, resemble a human interlocutor in many ways; perhaps speakers do not realize his limitations? It seems unlikely that participants were truly confused about this, given the restrictions in how VIRTUO could respond to them in this experiment. However, the principle that humans might interact with VIRTUO as though he were a real human even though he is not could still explain their accommodation behavior.

Some social behaviors seem to be so automatic that they do not disappear in human-computer interaction even when they are totally illogical in these scenarios. For example, humans have been shown to exhibit politeness and reciprocity to computers (Nass & Moon, 2004), in what Nass and colleagues have called "overlearned social behaviors." If accommodation is such an overlearned social behavior, then people might accommodate to VIRTUO not because they think that they will influence his beliefs about them or behavior toward them, but because this behavior is applied automatically regardless of its applicability in a specific situation.

If this is in fact the reason why speakers accommodate to VIRTUO, then it suggests a reinterpretation of the accommodation we see in natural conversation as well. That is, interactional motivations may underlie linguistic accommodation, but in an automatic, overlearned way. If so, then speakers may not have specific intentions about the interaction they are engaged in, and they may have very little control over their accommodation behavior. This is consistent with the idea that accommodation can be motivated by general social goals, even in the absence of short-term social motivations.

Conclusions

In real-world conversations, accommodation may often be motivated by efforts to achieve interactional goals: people accommodate to make others do, think and feel things. But the present data show that this is not the only reason that people accommodate. Since people accommodate to a virtual interlocutor, we can conclude that accommodation is not necessarily driven by immediate attempts to influence social relationships or convey social messages. Yet, social motivations at a broader level may motivate accommodation, which may be a tool by which people develop linguistic styles, over the long term. The finding that the degree to which people accommodate correlates with how much they identify with their interlocutor suggests that accommodation is not merely a reflex. However, these results do not rule out some role for alignment processes that are engaged automatically. In real conversations, social and interactional factors may combine with automatic factors to produce linguistic accommodation.

Acknowledgments

Thanks to Albert Russel, Gerd Klaas, and Jeroen Derks for technical assistance in setting up the VR lab and programming the virtual world; to Sanne Berends, Daphne van Moerkerken, Merel van Rees Vellinga, and Tomas Bergvelt for help with data collection and coding; and to Pieter Seuren and Matthias Sjerps for help creating the verbal materials.

References

Alim, H. S. (2004). *You Know My Steez: An Ethnographic and Sociolinguistic Study of Styleshifting in a Black American Speech Community.* Durham, NC: Duke Univ. Press, 303 pgs.

Auer, P. & Hinskens, F. (2005). The role of interpersonal accommodation in a theory of language change. In P. Auer, F. Hinskens & P. Kerwswill (eds). *Dialect change: The convergence and divergence of dialects in European languages.* Cambridge University Press.

Babel, M. (2009). *Phonetic and Social Selectivity in Speech Accommodation.* Doctoral dissertation, Department of Linguistics, UC Berkeley.

Barr, D. & Keysar, B. (2002). Anchoring comprehension in linguistic precedents. *Journal of Memory and Language*, 46, 391-418.

Bortfeld, H. & Brennan, S. E. (1997). Use and acquisition of idiomatic expressions in referring by native and non-native speakers. *Discourse Processes*, 23, 119-147.

Brennan, S., & Clark, H. (1996). Conceptual Pacts and Lexical Choice in Conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* 22(6), 1482–1493.

Coupland, N. (1985). 'Hark, Hark, the Lark': Social motivations for phonological style-shifting. Language & Communication, 5(3):153-171.

Eckert, P. & McConnell-Ginet, S. (1992). Think practically and look locally: Language and gender as community-based practice. *Annual Review of Anthropology* 21:461–90.

Giles, Howard, Donald M. Taylor, and Richard Bourhis. (1973). Towards a theory of inter- personal accommodation through language: some Canadian data. *Language in Society* 2:177–192.

Goldinger, Stephen D. 1998. Echoes of echoes? An episodic theory of lexical access. *Psychological Review* 105:251–279.

Hannah, A. & Murachver, T. (1999). Gender and Conversational Style as Predictors of Conversational Behavior. *Journal of Language and Social Psychology,* 18(2):153-174.

Heeter, C. (1992). Being There: The Subjective Experience of Presence. *Presence: Teleoperators and Virtual Environments*, MIT Press.

Nass, C. & Moon, Y. (2000). Machines and Mindlessness: Social Responses to Computers. *Journal of Social Issues* 56(1):86-103.

Niederhoffer, K. and Pennebaker, J. (2002). Linguistic Style Matching in Social Interaction. *Journal of Language and Social Psychology*, 21(4).

Niedzielski, N., Giles, H., 1996. Linguistic accommodation (Essay no. 39). In: Goebl, H., Nelde, P.H., Stary, Z. and Wolck, W. (Eds.), *Contact Linguistics: An International Handbook of Contemporary Research.* De Gruyter, Berlin, pp. 33-342.

Pardo, Jennifer S. 2006. On phonetic convergence during conversational interaction. *Journal of the Acoustical Society of America* 119:2382–2393.

Pickering, M. J., & Garrod, S. (2004) Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences,* 27(2), 169-226.

Connecting the Visible to the Invisible: Helping Middle School Students Understand Complex Ecosystem Processes

Sameer Honwad[1], Cindy Hmelo-Silver[1], Rebecca Jordan[2], Catherine Eberbach[1], Steven Gray[2], Suparna Sinha[1]

[1]Department of Educational Psychology, Graduate School of Education, Rutgers University, New Brunswick 08901

[2]School of Environmental and Biological Sciences, Department of Ecology, Evolution, and Natural Resources and the Program in Science Learning, Rutgers University, New Brunswick, NJ 08901.

[3]Ashok Goel, [3]Swaroop [3]Vattam, [3]Spencer Rugaber, [3]David Joyner

[3]Design & Intelligence Laboratory, School of Interactive Computing, Georgia Institute of Technology, Atlanta, GA 30332.

Abstract

Learning about ecosystems is challenging because, like any complex system, they are simultaneously multidimensional and dynamic. Often, learners engage only with the visible components of an ecosystem and draw either single or linear causal connections between components. In this study, we explored how using a Structure-Behavior-Function framework supported middle school students' conceptual and complex reasoning about the visible and invisible components of an ecosystem. Research shows that learners often engage only with the visible components of an ecosystem and draw linear/single causal connections between the components of the ecosystem. Our findings suggest that a combination of using structure, behavior, and function approach along with a set of carefully designed technology tools can push the students toward a better understanding of the ecosystem functioning. The results show that along with the visible components of the ecosystem, students have started to identify the invisible components of the ecosystem.

Keywords: Ecosystems learning, SBF, complex systems, Science education

Introduction

Given the urgent need to empower the future generation with knowledge to help them make informed decisions about their ecosystems and environment, both national and local science standards have a growing focus on ecosystems learning (e.g., National Research Council, 1996; New Jersey Department of Education, 2006). Developing ecosystems understanding is challenging, because it requires learners to understand how different aspects of an ecosystem are interconnected, and the processes that occur within such systems (Anderson, 2008; Covitt & Gunckel, 2008; Jordan et al., 2009).

Ecosystem processes are challenging for learners, because these are complex systems that transcend spatial, temporal and cognitive boundaries (Pickett, et al 1997). Similar to other complex systems, ecosystems are also characterized by multidimensional processes that connect visible and invisible components of the system to one another (Hmelo-Silver & Azevedo, 2006). These visible and invisible components within the ecosystem are interdependent. The components have their own behavior patterns and any change in the patterns, affects not only other components, but also overall functioning of the system (Jordan, et al 2009). The dynamic and multifaceted nature of an ecosystem makes it difficult for learners to grasp the associations and interactions among system components (Gallegos et al 1994).

Learners find it challenging to think beyond the linear relationships and visible components of an ecosystem (e.g., food chains: Reiner & Eilam, 2001; aquaria: Hmelo-Silver, Marathe, & Liu, 2007; systems: Hogan, 2000, food webs/nutrient cycles: Hogan & Fisherkeller, 1996, energy flow: Leach et al. 1996; water cycle: Covitt & Gunkel, 2008). When asked to draw or name components of an ecosystem, learners often focus on the visible components of the ecosystem (Gellert, 1962; Hmelo, Holton, & Kolodner, 2000). Expert-novice studies suggest that that it is hard for young learners to conceptualize the invisible components within an ecosystem such as: oxygen, nitrogen, and bacteria, (Hmelo-Silver, Marathe, & Liu, 2007). It is also challenging for students to think beyond single causality and linear connections between ecosystem components (Grotzer & Basca 2003).

In this paper, we present the results of a technology-intensive classroom intervention designed to teach middle schools students about aquatic ecosystems. The goals of our intervention are to help learners develop deep understanding of ecosystems and to use tools that make the invisible visible and the interconnections explicit.

Aquariums as Models for Learning

To help students understand complex systems, we implemented a two-week aquarium unit that was designed by a team of learning scientists, middle school classroom teachers, and ecologists. The technology consisted of a suite of tools: a function-oriented hypermedia (Liu & Hmelo-Silver, 2009), simulations of macro- and micro-level processes (Liu & Hmelo-Silver, 2008; Gray et al. 2008), and the Aquarium Construction Kit (ACT; Goel, Rugaber, & Vattam, 2009). The unit was grounded in the structure behavior and function approach.

Our approach to instruction is grounded in the structure-

behavior-function theory (Goel et al., 2009). The structure behavior function (SBF) approach is useful to explain dynamic systems with multiple components and levels (Goel et al., 2009; Liu & Hmelo-Silver, 2009). We view SBF theory as providing a conceptual representation with canonical explanations in biological systems, as well as, being consistent with expert understanding (Bechtel & Abrahamson, 2005; Hmelo-Silver et al., 2007). In addition to helping students organize their system knowledge the SBF representation also provides a scaffold for overall knowledge organization. The approach helps the learner to breakdown and distinguish individual parts of the complex system.

In a biological system, structure refers to components of an ecosystem that have form. Structures can be macro (e.g Fish, plants) or micro (e.g bacteria, fungi) in nature. Behavior represents the process of how structures achieve their functions, and, finally, functions are roles the structures play in an ecosystem.

Technology Support for Learning about Complex System

It is difficult for learners to understand many aspects of ecosystems because they have not had opportunities to engage with those processes that are dynamic and outside their perceptual understanding. In addition to helping students organize their system knowledge, the SBF representation also provides a scaffold for overall knowledge organization because it helps learners consider the relationships among form and function as well as the causal behaviors. We make SBF explicit through the use of hypermedia, organized in terms of SBF, and through the Aquarium Construction Toolkit (ACT) (Figure 1a and 1b).

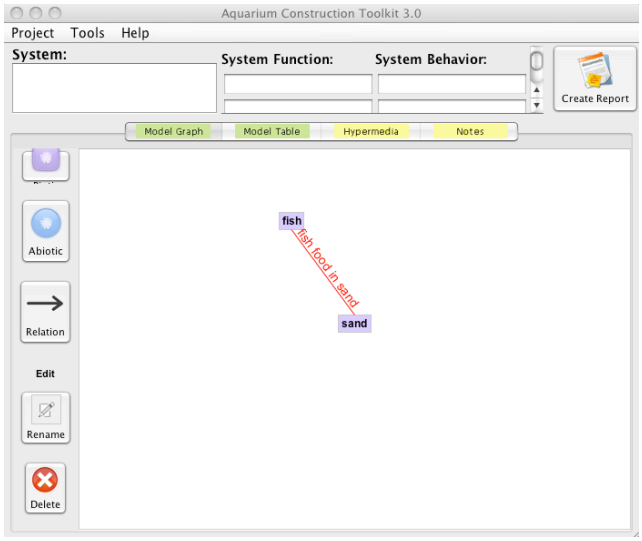


Figure 1a. ACT: A space to create models

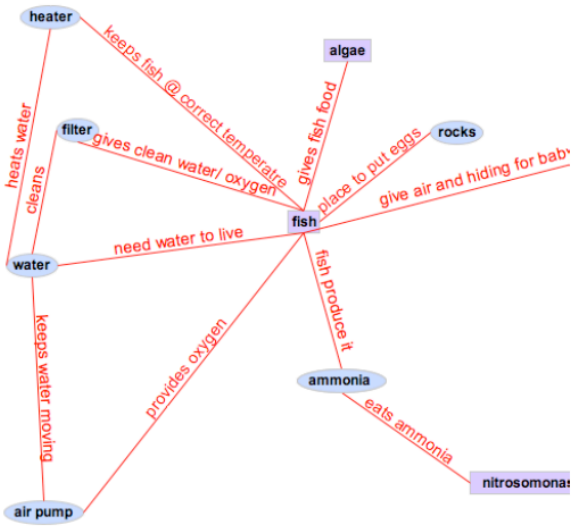


Figure 1b. ACT: Example of model created by a student.



Figure 2. SBF is used to organize the hypermedia

Along with the hypermedia and ACT tools students also used NetLogo simulations to learn about behaviors and functions within an ecosystem (Wilensky & Reisman, 2006). Using these simulations, (Figure 3) students learned about how to keep an ecosystem 'healthy.' For example, the macro fishspawn simulation allowed students to manipulate different aspects of the ecosystem such as initial population, spawning, filtrations, and amount of food. Thus if the students overfed the fish then the increasing ammonia (due to fish waste) within the water would affect water quality and the fish would die. This helps problematize water quality, which is a black box in the macro simulation. This creates the need for students to identify some of the invisible components within an ecosystem, and students also start to see the importance of these invisible components. For example, the students can observe how crucial nitrification cycle is for the overall health of an ecosystem. They also can learn that many components of the ecosystem involved in the nitrification process are invisible. These

behaviors and functions can then be observed in the micro level simulation.

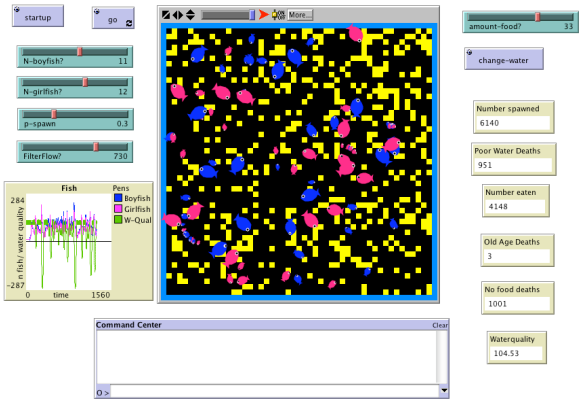


Figure 3: NetLogo simulation

Classroom Instruction

The science teacher introduced the unit by asking students to articulate their thoughts about ecosystems functions. This allowed the teacher to gauge the students' prior knowledge. The teacher then moved on to the ACT modeling tool and asked the students to represent their thoughts about ecosystems as structures behaviors and functions. The students recorded their ideas in a table within the ACT tool (Figure 4).

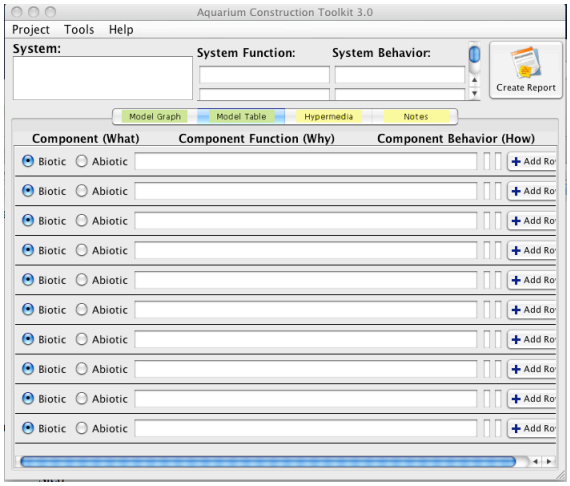


Figure 4: ACT table where students record ideas as structure, behavior, and function

The teacher also encouraged the students to use the hypermedia to sharpen their existing ideas about the ecosystems. The teacher then proceeded to the modeling activity using NetLogo simulations. In the NetLogo simulations students manipulated various ecosystem components (number of fish, amount of food, etc) in order to maintain a healthy ecosystem. The students worked in groups and were given the freedom to continuously refine their models. Finally at the end of the two-week period the students presented their models in front of the entire classroom.

Methods

Participants

Fifty-four seventh grade students from a suburban public middle school in the northeast United States participated in this study during their regular science instruction time. Two of the participants reported having an aquarium at home and most had been to a public aquarium. Many had also been on excursions to the beach or on fishing trips with adults in their families.

Data Sources

The students were given pre and post-tests before and after the intervention. In the pre and post-test students were asked to draw components of an aquatic ecosystem, and show relationships between them. They were also asked to label all of the components and relationships between those components.

Coding for pre and post tests

There were three parts to the coding. The first part of the coding scheme involved counting the number of visible and invisible ecosystem components that were drawn by the students. The second part of the coding scheme involved counting the number of relationships that the students observed between the components in their drawing. Care was taken to make sure that the relationships were scientifically plausible. We coded the connections on a three-point scale. We gave a connection one point if students made implausible connections between components of the ecosystem. A connection was assigned two points if students made plausible connections within the same level of an ecosystem (e,g. visible component to visible component; invisibile to invisible). One example of this is a connection that shows fish eat plants. Here both fish and plants are visible components of the ecosystem. A connection was assigned three points if students were able to make plausible connections between the visible and invisible components of the ecosystem. An example of this would be a connection showing that fish breathe oxygen (Figure 5). Here fish is the visible component of the ecosystem and oxygen is the invisible component of the ecosystem.

The third part of the coding scheme was designed to find out the type of connections the students made between the different components of the ecosystem. As components within an ecosystem function nonlinearly, it was important to find out whether student understanding of ecosystem functioning went beyond linear-single cause relationships. The coding scheme for the third part (types of connections) was adapted from Grotzer & Basca (2003).

This part of the coding scheme was also coded on a three-point scale. A connection was assigned a point if students made a 'simple linear' connection between the components of the ecosystem. A simple linear connection was observed as a connection that was linear, one directional and

indicating single cause and effect. For example, fish eat plants is a linear connection because it indicates that only fish benefit from the plants. A connection was given two points if the students made a 'complex linear' connection between the components of the ecosystem.

A complex linear connection was defined as a linear connection that had more than one cause and effect. For example, plants get energy from the sun, fish eat plants and thus fish get energy from the plants is a complex linear relationship because it shows one directional relationship between more than two components of the ecosystem. This code was also used when students represented symbiotic relationships/mutually beneficial relationships.

Finally a connection was given three points if the connection was observed to connect more than two components in a mutually benefiting relationship. The connection was called 'cyclic' (Figure 5). For example, fish waste produces ammonia, a form of nitrogen that is then transformed by different bacteria into new forms of nitrogen that support plant growth, which in turn benefit the fish.



Figure 5: Connecting the visible (fish) to invisible (ammonia). Also, an example of a cyclic connection.

Reliability was calculated by having three independent raters code the entire sample. The overall reliability was 98% agreement.

Results

We expected the students to start identifying invisible components within an ecosystem. Since the intervention provided them with an opportunity to learn about the system in depth the results show that the students have identified more invisible components (Table 1). However, the students did not show any significant change in identifying the visible components (Table 1).

Table 1: Components coding (N=54).

	Visible	Invisible
Pretest Mean (SD)	8.50 (3.34)	0.28 (0.49)
Posttest Mean (SD)	7.61 (3.21)	2.87 (2.17)
Sample Size	54	54
T (53)	1.90	8.86*
Effect Size	0.13	0.64

*p< 0.05

We also expected the students to make more plausible connections between the components because the instruction was designed to scaffold students' understanding of how ecosystem components are connected to each other. The results, shown in Table 2, demonstrate that students made significant progress in making plausible connections within levels (visible to visible and invisible to invisible) and between levels (visible to invisible).

Table 2: Plausible connections made between ecosystems components (N=54)

	Plausible connections made within level	Plausible connections made between levels:
Pretest Mean (SD)	3.81 (2.15)	0.17 (0.61)
Posttest Mean (SD)	4.87 (2.91)	1.43 (1.53)
Sample Size	54	54
T (53)	2.55*	7.09*
Effect Size	0.21	0.47

*p< 0.05

Finally we investigated whether the types of plausible connections students were making were demonstrating the complexity of ecosystem functions. It was not clear whether students were able to move beyond making linear connections or complex linear connections. We found that the number of students making simple linear connections increased from pre to post. However, there was no significant change in the number of students making complex linear connections. Finally there was a significant change in the number of students making cyclic connections. Although, the results clearly showed that only a small number of students made a leap to making more complex connections between the ecosystem components (Table 3).

Table 3: Types of connections made by students

	Linear Relationships	Linear Complex relationships	Cyclic
Pretest Mean (SD)	1.85 (1.87)	0.52 (0.91)	0.02 (0.14)
Posttest Mean (SD)	2.80 (2.33)	0.74 (0.96)	0.15 (0.36)
T (53)	2.76*	1.73	2.81*
Effect Size	0.21	0.11	0.09

*$p < 0.05$

Discussion and Conclusion

Our results show that students find it challenging to conceptualize the role of invisible components within an ecosystem. Consistent with other research, students initially focus on the interactions between the visible components of the ecosystem (e.g., Hmelo-Silver et al., 2007). For example, most students represented the fish eating fish (prey predator) relationship as the primary relationship within an ecosystem. However the study also shows that students are on a trajectory of conceptual change and began to consider invisible components of the ecosystem and how they connect to what is visible.

Our findings suggest that a combination of using structure, behavior, and function approach along with a set of carefully designed technology tools can push the students toward a better understanding of the ecosystem. Another study (Goel et al. 2010) that looks at how the ACT tool helps students construct SBF models of complex ecosystem processes is a part of the proceedings.

The results show that along with the visible components of the ecosystem, students have started to identify the invisible components of the ecosystem. They are still not completely making a sophisticated model that includes the visible and invisible components connected to each other, but this is the first step. Moving students to a more robust and rich understanding of complex systems requires more than a two week intervention. In our ongoing research, we are exploring how SBF thinking can provide a tool for students to understand complex biological systems that are pervasive in the world in which they live and are key components of helping students become scientifically and environmentally literate citizens.

Acknowledgements

The authors would like to thank Staci Klienbaum, Rebecca Digiglo and Courtney Farruggia for data organization and coding.

References

Anderson, C.W. (2008). *Learning Progression for Environmental Science Literacy: Overview of the Interactive Poster Symposium.* Paper presented at the annual meeting of the National Association for Research in Science Teaching, 2008.

Bechtel, W., & Abrahamson, A. (2005). Explanation: A mechanist alternative. *Studies in the History and Philosophy of Biological and Biomedical Sciences,* 36, 421-441.

Covitt, B.A., Gunckel, K.L. & Anderson C.W. (2009). Students' Developing Understanding of Water in Environmental Systems. *Journal of Environment Education*, 40(3), 37-51.

Gallegos, L., Jerezano, M.E., & Flores, F. (1994). Preconceptions and relations used by children in the construction of food chains. *Journal of Research in Science Teaching*, 22, 421-426.

Gellert, E. (1962) Children's conceptions of the content and functions of the human body. *Genetic Psychology Monographs,* 65, 293-411.

Goel, A. K., Gomez de Silva Garza, A., Grué, N., Murdock, J. W., Recker, M. M., & Govinderaj, T.(1996). Towards designing learning environments -I: Exploring

Goel, A.K., Vattam, S.S., Rugaber, S., Joyner, D., Hmelo-Silver, C., Jordan, R., Honwad, S., Gray, S. & Sinha, S. (2010). Functional and causal abstractions of complex systems. Paper to be presented at the annual conference of the Cognitive Science Society, 2010.

Grotzer, T. & Basca, B.B. (2003). How does grasping the underlying causal structures of ecosystems impact students' understanding? *Journal of Biological Education*. 38 (1)16-35.

Hmelo-Silver, C. E. & Azevedo, R. (2006). Understanding complex systems: Some core challenges. *Journal of the Learning Sciences*, 15, 53-61.

Hmelo, C. E., Holton, D. L., & Kolodner, J. L. (2000). Designing to learn about complex systems. *Journal of the Learning Sciences,* 9(3), 247– 298.

Hmelo-Silver, C. Marathe, S., & Liu, L. (2007). Fish swim, rocks sit, and lungs breathe: expert- novice understanding of complex systems. *Journal of the Learning Sciences*, 16, 307- 331.

Hogan, K. & Fisherkeller, J. (1996) Representing students' thinking about nutrient cycling in ecosystems: Bidemensional coding of a complex topic. *Journal of Research in Science Teaching*, 33, 941-970.

Hogan, K. (2000). Assessing students' systems reasoning in ecology. *Journal of Biological Education,* 35, 22-28.

Hungerford, H.R., Bluhm, W.J., Volk, T.L., & Ramsey, J.M. (*eds.)* (2001) *Essential Readings in Environmental Education; 2nd Edition*. Stipes: Champaign, IL.

Jordan, R., Gray, S., Demeter, M., Lui, L. & Hmelo-Silver, C (2009). An Assessment of Student' Understanding of Ecosystem Concepts: Conflating Ecological Systems and Cycles. *Applied Environment Education and Communication*, 8, 40-48.

Leach, J., Driver, R., Scott, P. & Wood-Robinson, C. (1996). Children's ideas about ecology 2: ideas found in children aged 5-16 about the cycling of matter. *International Journal of Science* Education, 18, 19-34.

Liu, L. & Hmelo-Silver, C. (2009). Promoting complex systems learning through the use of conceptual representations in hypermedia. *Journal of Research in Science Teaching,* 46, 1023-1040.

Liu, L., & Hmelo-Silver, C. E. (2008). Collaborative Scientific Conceptual Change in a Simulation-supported Learning Environment. In G. Kanselaar, Jonker, V., Kirschner, P., & Prins, F. (Ed.), *Proceedings of the Eight International Conference for the Learning Sciences* (Vol. 1, pp. 477-484). Utrecht, The Netherlands: International Society for the Learning Sciences.

National Research Council. (1996). National science education standards. Washington D.C.: National Academy Press.

Pickett, S.T.A., Burch, W.R., Dalton, S.E., Foresman, T.W., Grove, M.J. & Rowntree, R. (1997). A conceptual framework for the study of human ecosystems in urban areas. *Urban Ecosystems*, 1, 185-199.

Reiner, M. & Eilam, B. (2001). A systems view of learning. *International Journal of Science Education,* 23, 551-568.

Vattam, S., Goel, A. K., Rugaber, S., Hmelo-Silver, C. E., & Jordan, R. (2009). *From conceptual models to agent-based simulations: Why and How.* Paper presented at the Fourteenth International Conference on AI in Education, Brighton UK.

Wilensky, U. & Reisman, K. (2006). Thinking like a wolf, a sheep or a firefly: Learning biology through constructing and testing computational theories -- An embodied modeling approach. *Cognition and Instruction,* 24, 171-209.

Response Times and Misconception-like Responses to Science Questions

Andrew F. Heckler (heckler.6@osu.edu)
Department of Physics, Ohio State University
Columbus, OH 43210 USA

Thomas M. Scaife (scaife.1@osu.edu)
Department of Physics, Ohio State University
Columbus, OH 43210 USA

Eleanor C. Sayre (esayre@gmail.com)
Department of Physics, Wabash College
Crawfordsville, IN USA 47933

Abstract

Patterns of incorrect answering or "misconception-like" responses to scientific concept questions have been well documented. Here we investigate both response choices and response times to gain insight into the nature of misconception-like responses. In a series of experiments involving questions on graphs in which participants must compare the slopes of two points, we find that students answering with misconception-like responses, namely comparing heights rather than the slopes, do so consistently and more rapidly than those answering correctly. We also find in a speeded experiment, that all students are able to compare slopes and heights, but comparing heights requires less time than comparing slopes. Finally, by imposing a delay in responding that is long enough for the responder to process both slopes and heights, we find a reduction in misconception-like responses. Thus the misconception-like responses can be explained in terms of speed-accuracy tradeoff models in which responders place high priority on answering quickly.

Keywords: Scientific misconceptions, graphs, response time, speed-accuracy tradeoff, physics education.

Introduction

It is well documented in science education that students often respond to scientific concept questions in regular and persistent patterns of errors (Pfundt & Duit, 2000). For example, when presented with qualitative questions about the time of flight of a projectiles with various trajectories, many students will incorrectly answer that both range and height of the trajectory influence the time of flight, when in fact only the height determines the time of flight. For convenience, we will refer to such patterns of incorrect answers as *misconception-like* responses, as we do not know whether they stem from coherent and explicit "misconceptions" of the students or some other mechanism.

While past studies of student difficulties with answering science concept questions have examined the patterns of response choices, in this study we investigated the response *times* as well as the response choices in order to address two main questions. First, are there interesting patterns of response times when comparing correct and misconception-like responses? Second, does response time data shed any light on the processes involved in choosing correct or misconception-like answers?

A number of investigators have examined response times on standardized tests. These studies use both response time and response accuracy in order, for example, to eliminate the effect of guessing and thus improve the accuracy of the tests (e.g., Schnipke & Scrams, 1997; van der Linden, 2008), or to detect cheating (van der Linden & van Krimpen-Stoop, 2003). In this study we investigate questions that evoke misconception-like responses. As we will see in Experiment 1, these incorrect responses are not guesses but rather a coherent pattern of answering.

In addition to studies on standardized tests, a long history of response time studies in a wide range of tasks has revealed the well-known phenomenon of the speed-accuracy tradeoff, namely that there often exists a monotonically increasing relation between response time and response accuracy (Wickelgren, 1977). There are two classes of models used to explain the speed-accuracy tradeoff curve. The first is the *fast-guess* model which proposes that students use a mixture of guesses, which are fast, and non-guesses, which are slow. As mentioned earlier, since there is very little guessing in the responses in this study, we will not consider this class of models.

The second class of models postulates that response choices are a result of *decision criteria* applied to evidence that accumulates over time. As time increases, the amount of information increases, thus increasing accuracy, which explains the speed-accuracy tradeoff curve (e.g., Ratcliff, 1978; Smith & Vickers, 1988).

Let us consider the decision-criteria model with respect to response times on scientific concept questions that often evoke misconception-like responses. If correct answers and misconception-like answers require different solution paths, then it is possible that the response times of the two paths will be different. For example, if the time needed for the process involved in obtaining the misconception-like answer is inherently shorter than the process for obtaining the correct answer, then one would expect the misconception-like response times to be shorter.

In addition to expecting different response times for correct and misconception-like responses, in this model the

actual response time also depends on decision criteria. For example, there may be a minimum amount of information needed before a decision can be made. On the other hand, there may also be a maximum amount of time allotted for the decision. Therefore, misconception-like responses may be a result of implicit decision criteria rather than the responder's absolute ability to determine the correct answer.

This study proceeds as follows. In the first experiment, we established that our example science concept question evokes misconception-like responses, and we characterized the difference in the response times of the correct and misconception-like responses. In Experiment 2, we measured and characterized the response times needed to process the main underlying tasks required for obtaining the correct and misconception-like responses. In Experiment 3, we impose a minimum time to respond in order to determine whether this will affect the response choices.

Experiment 1

The first experiment investigates a well-known student difficulty with interpreting graphs commonly used in math and physics courses at the high school and introductory university level. Specifically, when a variable of interest corresponds to the slope of a line at a given point, students instead often attend to the value (i.e. the height) of the point on the line rather than the slope. For example, When students are presented with a position versus time graphs for an object (see Figure 1) and asked "at which point does the object have a higher speed?", many incorrectly answer according to the higher point rather than the greater slope (McDermott, Rosenquist, & van Zee, 1987).

Experiment 1 was designed to achieve three goals. The first goal was to replicate the misconception-like response pattern indicating a tendency to attend to values (heights) of points on a line rather than slopes at those points. The second was to determine whether this pattern was due to students' fundamental inability to compare the slopes of points on a line, or if it was instead a function of familiarity with the question context. Finally the third goal was to compare the response times of students answering correctly vs. those answering incorrectly to determine if there was a pattern in response times corresponding to answer choice.

Experiment 1 used a between-subjects design employing three conditions: math graphs, kinematic graphs, and electric potential graphs (see Figure 1). Each condition presented a series of graphs and participants were asked to compare two points on a curved (or straight) line on the graph. Figure 1 presents examples of the graphs in the three conditions, including the question posed for each graph.

In addition to the fact that the series of graphs in the three conditions were identical (except for the labels on the axes), the questions posed for each graph are also conceptually analogous. In particular, the math graph condition asked for a comparison of the slopes at two points (magnitude of slope = $|dx/dt|$), and the other two conditions also effectively asked for a comparison of slopes since *speed* is the slope for the position-time (kinematic) graph (speed = $|dx/dt|$), and the magnitude of *electric field* is the slope of the electric potential (V)-position (x) graph (magnitude of electric field = $|dV/dx|$).

The three graph conditions were also at differing levels of familiarity for the participants. The math graphs were the most familiar, as they are introduced in standard curricula before and throughout high school. The kinematic graphs were the next most familiar. They are typically introduced in high school physics or physical science courses, and used frequently in the university level physics course that was a prerequisite to the physics course in which the participants were enrolled at the time of the study. Finally, the electric potential graphs were the least familiar, as they are not part of standard pre-university curriculum and most participants saw them for the first time in physics course in which they were enrolled at the time of the study.

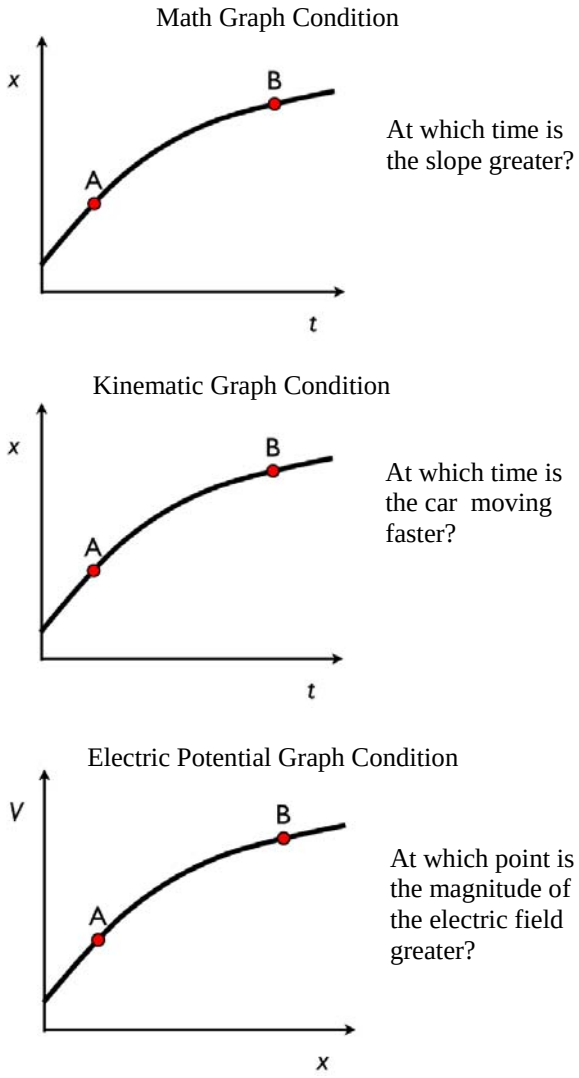


Figure 1. Examples of the graphs and questions used in the three conditions in Experiment 1. The answer choices for all three were: "A", "B", or "the same at A and B".

Method

Participants Participants were enrolled in one of two undergraduate calculus-based introductory physics courses. The first course covered the topic of classical mechanics and the topic of the second course covered electromagnetism. The courses are part of a three-course introductory physics series, and are typically populated with engineering majors. Participants received partial course credit for participation, and the participation rate for both courses was > 95% of all students enrolled in course.

Participants were randomly chosen to be placed in each condition. For the math graphs condition, 28 participants were chosen from the mechanics course and 49 were chosen from the electromagnetism course, for a total of 77 participants. For the kinematics graphs condition, 94 participants were chosen from the mechanics class. For the electric potential graphs condition, 38 students were chosen from the electromagnetism course.

Procedure, materials and design All testing was presented to individual participants on a computer screen in a quiet room. They proceeded through testing at their own pace, and their response choices and response times were electronically recorded.

In each condition students were presented with a series of graphs and asked to compare relevant values at two points on each graph. Participants were given no feedback as to the correctness of their answers. See Figure 1 for examples of graphs and specific questions asked.

Testing consisted of a comparison of two points on 14 graphs (presently serially) with various curve shapes: 8 graphs in which the higher point had a lower slope (these are the difficult "target" questions), 2 graphs in which the higher point had a higher slope, 2 graphs in which both points had the same slope, and 2 graphs in which the two points had the same height but different slopes. The graphs types were placed in a fixed random order, and this sequence was presented to all participants in all conditions. Thus the graphs were mixed such that the correct response was not always "A", and not always the lower or higher point. Our previous pilots studies did not reveal any significant effects of order of graph type on answering. Furthermore, Experiment 3 uses a design to counterbalance for order, with similar results to Experiment 1. Therefore we are confident that the results here are not an artifact of question order.

Results

Analysis of response choices We first report on the performance on the "target" questions, namely those graphs in which the higher point has a lower slope (see Figure 1 for examples). These type of questions are important for investigating graph difficulties, since the correct answer choice (the point with the greater slope, but with a lower height on the graph) is opposite of the common "misconception" that, for example, "the higher point has greater speed".

Figure 2 presents the average scores for the target questions for each condition. The averages depended strongly on the graph type, with scores of 94% for the Math graphs, 72% for the Kinematic graphs an 47% for the Electric Potential graphs (One-way ANOVA with Bonferroni adjusted post-hoc comparisons, *p*s < 0.0001). Thus the less familiar the graph context, the lower the score.

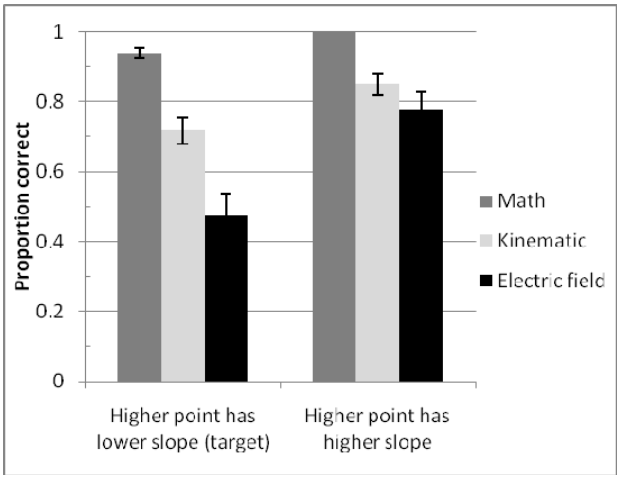


Figure 2. Experiment 1, mean scores for the Math graphs, the Kinematic graphs, and the Electric potential graphs conditions. Scores are shown for target questions in which one of the points has a higher slope but lower value, for "aligned" questions in which one of the points has a higher slope and higher value. Error bars are 1 S.E.M.

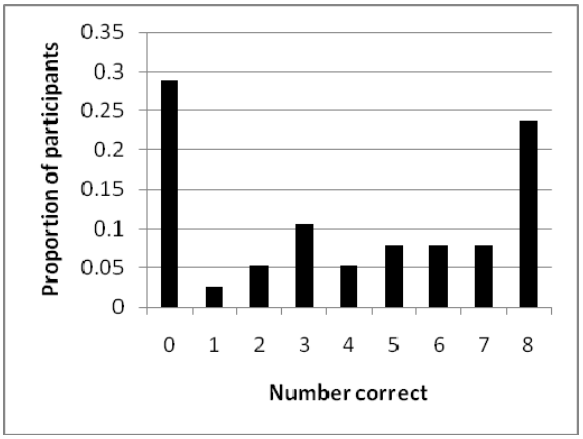


Figure 3. Distribution of scores on target questions for the Electric Potential graph questions. Rather than a random binomial distribution, this distribution is bimodal, indicating that most participants are not guessing.

The patterns of specific answer choices also revealed that answering was not random, and those choosing incorrect answers consistently choose the main misconception-like answer. There are two kinds of evidence to support this. First, Figure 3 presents the score distributions for students

answering the electric potential graph questions. If the answering were random, one would expect a binomial distribution of scores. Instead Figure 3 shows a strong bimodal distribution, with most students either answering all questions correctly or answering all questions incorrectly. Note that over 95% of the incorrect answers were the main misconception-like distracter, namely the point with the higher value; few incorrect answerers chose that the points had the same electric field.

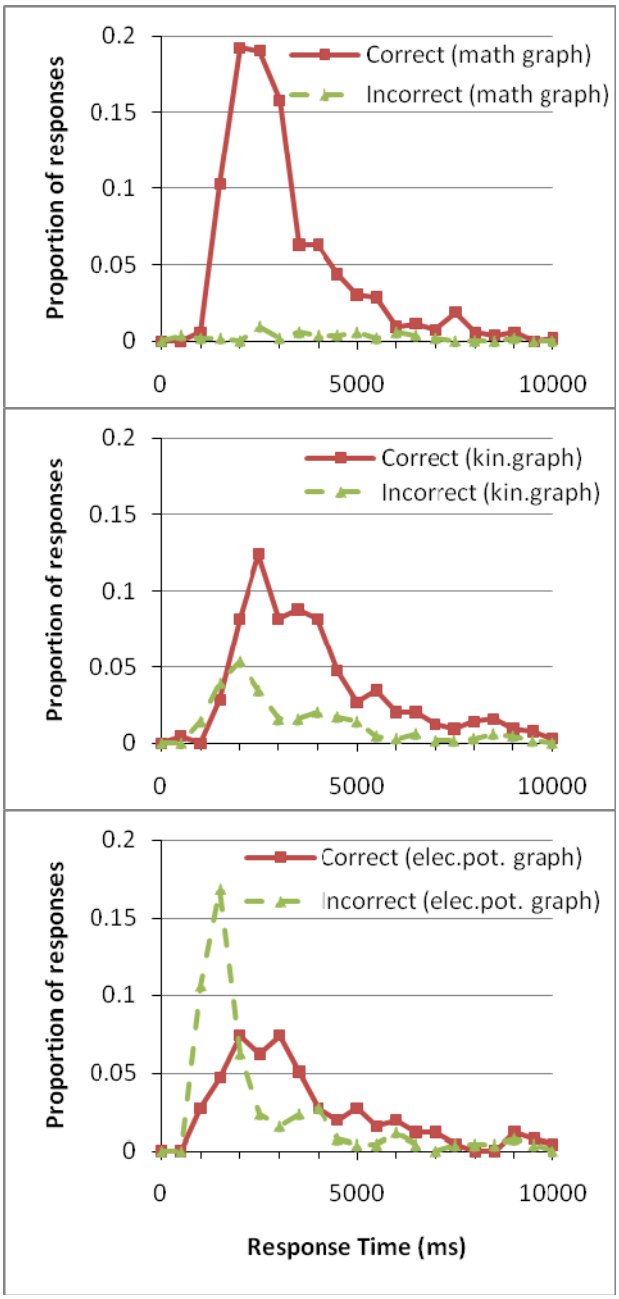


Figure 4: Experiment 1 distribution of response times on target questions in the math (top), kinematic (middle), and electric potential (bottom) graphs conditions. The area under the curves represents the total proportion of correct or incorrect (misconception-like) responses for each condition.

The second kind of evidence to support the fact that students answering incorrectly consistently chose the misconception-like distracter comes from the non-target type of questions. For example, Figure 2 shows the score for the questions for which the point with the highest values was also the point with the highest slope. In all three conditions, participants scored higher on these "aligned" questions in which the slope and values were both greater compared to the target questions in which the point with the higher slope had the lower value (paired t–tests, ps < 0.003).

Analysis of response times Figure 4 presents the distribution of response times for each question in each condition, separated out by all questions answered correctly and those answered incorrectly. Note that the response times for all students were pooled together, so this graph represents both between student and within student data mixed together.

There are two main points about the data presented in Figure 4. First, for the kinematic and electric potential graphs conditions, the response times are shorter for the incorrect answers than the correct answers (Mann-Whitney U test used because of long tails in distribution, ps < 0.0001). The peaks of the distribution for the incorrect answers are about 500 ms earlier than for the correct answers. There are so few incorrect responses for the math graphs that no reliable comparisons can be made for that case. Second, the peaks of correct answers for all three conditions are at the same place (about 2000 ms) for all three conditions.

At first glance, the fact that the response times for the incorrect responses are shorter than the correct responses may not be a surprise: the speed-accuracy tradeoff is a well known phenomenon. However, as discussed earlier, the incorrect answers are not random guesses, so one cannot conclude that the shorter response times are due to fast guessing. Rather, there is a pattern to the guessing.

This leads us to the question of whether there is an inherent difference in time required to perform the two different response modes, which in Experiment 1 translate to systematically correct vs. "incorrect" (misconception-like) responses. The underlying task to determine the correct answer is to compare the slopes at the two points and the underlying task to determine the misconception-like answer is to compare the heights of the two points. Therefore, in Experiment 2 we determine the time required to perform these two basic tasks.

Experiment 2

The goal of Experiment 2 is to compare the response times for the tasks of comparing the heights of two points vs. comparing the slopes at two points.

Method

Participants Eighteen undergraduate students participated, receiving partial credit for a calculus-based introductory physics course.

Procedure, materials and design The procedure was similar to Experiment 1. Participants were presented with examples depicting various position time curves for a car (see Figure 1 for an example). For each graph, two points on the curve were marked, indicating the position and time of the car at two different times. In a within-subject design, participants were asked to determine as quickly as they can without making a mistake either which point was higher, or at which point the slope was greater. The test was administered in blocks of 9 questions of the same type (compare height or compare slope). Question type blocks were presented in an alternating sequence, with 2 blocks for each question type, for a total of 4 blocks (36 questions).

Results

The mean score for both the compare height questions and the compare slope questions was >97%. Because the response times in the first two blocks were initially relatively high and decayed to an asymptote within 3-4 questions, and the times were near a steady asymptote in the second two blocks, we only compared the response times in the second two blocks (third and forth block). The response times in the first two blocks showed the same trend. Figure 5 presents the distributions of response times for the height and slope comparison tasks. The mean response time was significantly lower for the comparison of height questions (788 ms) verses the comparison of slope questions (1216 ms), (paired-sample $t(17)= 7.04$, $p < 0.001$, $d = 1.28$).

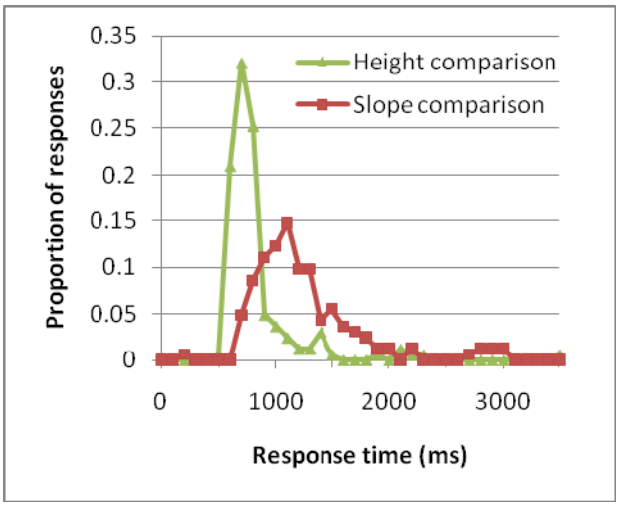


Figure 5: Distribution of response times for the height comparison and slope comparison tasks in Experiment 2.

Figure 5 is similar to the results from the electric potential graphs in Figure 4, with the participants choosing the point with the greater height answering significantly faster than those choosing the point with the greater slope. The main difference is that the peaks in Experiment 2 are earlier and the widths are narrower. One possibility for the difference is that in Experiment 2, the participants were asked to answer as quickly as possible. Therefore the time to peak represents a typical minimum time needed to perform the task.

These results suggest that the difference in response times between the correct answer (comparing slopes) and misconception-like answer (comparing heights) is due to these answers employing different procedures to complete, and these two procedures require different amounts of time.

Experiment 3

The results of Experiments 1 and 2 demonstrate that response times of misconception-like responses are shorter than those of correct responses, and the underlying task necessary for determining the misconception-like response (comparing heights) takes less time than the task necessary for determining the correct response (comparing slopes). Considering the decision-criteria model discussed earlier, one way to help explain misconception-like responses on these questions is to propose that students self-imposed a decision criterion that gave high priority to answering quickly. In this case, then students may have tended to choose the information that was processed first, namely information about the relative heights of the points, and this lead to an incorrect response. The information about the relative slopes would lead to the correct answer but took longer to process, so it was excluded from the decision.

Experiment 3 aims to test this idea by imposing a minimum time delay before responding. That is, participants are shown the question and may answer only after a short delay. If the delay is long enough to allow for the processing of both faster solution (comparing heights) and slower solution (comparing slopes), then they would have both kinds of information available. This could then result in participants with the delay answering more frequently with the response consistent with the slower process compared to participants who had no delay imposed. The delay was set to 3 seconds, since the majority of participants who answered correctly in Experiment 1 did so within this time.

Method

Participants A total of 72 undergraduates enrolled in a calculus-based introductory physics courses in electromagnetism participated, receiving partial course credit for participation. Participants were randomly assigned to one of two conditions: 37 in the delay condition and 35 in the control condition.

Procedure, materials and design The procedure was similar to Experiment 1. Participants in the control condition were presented with the same graphs as in the electric potential graph condition in Experiment 1. Participants in the delay condition were presented with the same graphs. However, before the questions began they were presented a screen with the following message: "On each slide, you will see the question with a message at the bottom of the screen. At first the message will read: 'Take a moment to carefully consider your answer.' While this message is displayed, you will not be able to answer the question. After a couple of seconds, the message will

change and prompt you for an answer. Please press the key that corresponds to your answer at that time." They were then given a simple math-fractions problem as an example of the delay, then they proceeded to the graph questions.

Therefore the students in the delay condition were required to wait 3 seconds before responding. The only other difference in Experiment 3 was to randomly assign students within each condition into one of two question-order conditions, to counterbalance for any question order effects. Note that there were no significant differences in performance between the control in Experiment 3 compared to the electric potential graphs condition in Experiment 1.

Results

As shown in Figure 6, participants in the Delay condition score significantly higher than those in the control condition (70% vs. 49%, $t(70) = 2.07$, $p = .04$, $d = .5$).

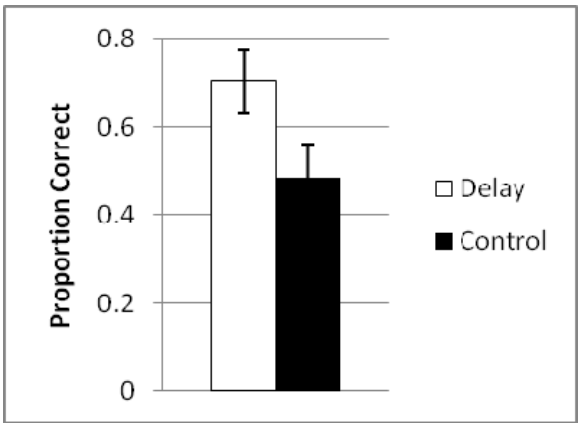


Figure 6: Results of Experiment 3. Error bars are 1 S.E.M.

Discussion and Conclusion

There are three main results of this paper. First, in Experiment 1 we found a clear difference in the pattern of response times for correct vs. misconception-like responses. This cannot be explained by correct vs. guessing responses because the misconception-like responses are not guesses, rather a consistent pattern of answering. For the particular example used in this study, we found that students will often compare heights of points on a graph, even in cases when they are supposed to compare the slopes. The participants answering with the misconception-like response tend to respond more quickly than those answering correctly.

Second we found in Experiment 2 that participants were able to compare heights and slopes with near-perfect accuracy, and it takes longer to compare slopes than heights. This response-time pattern is consistent with Experiment 1.

Third, when a delay for responding is imposed on the participants, they tend to answer correctly more frequently. This suggests that participants are able to arrive at the correct answers for these kinds of questions, but there is another factor influencing their responses.

The basic structure of the decision-criterion model may at least qualitatively provide an explanation for these results. The key feature of the model is that there exists a set of criteria for responding. Let us hypothesize two criteria that can explain the results. The first criterion is the need for information about the comparison of the two points that is *plausibly relevant*. The second criterion is the need for rapid responding. If the information on the comparison of heights is plausible enough, the responder who is free to respond at any time may tend to use *only* the height information since it is obtained quickly, and thus respond consistently and incorrectly. If, on the other hand, a time delay were imposed that was long enough to allow the responder to process both height and slope comparison information, then the response choice will be based on *both* height and slope information (and an additional decision is made on which is more relevant). This could naturally result in an increase in respondents choosing the correct answer.

Therefore, these results suggest that for the graphs questions studied here, an implicit tendency to answer rapidly coupled with the fact that an incorrect answer with sufficient plausibility is arrived at rapidly may be at least partially responsible for the misconception-like answers. The respondents are capable of answering correctly, but instead they tend to answer quickly. This prevents them from processing additional relevant information and considering alternative possibilities that may be more valid.

Acknowledgments

This research is supported by a grant from the Institute of Educational Sciences of the U.S. Department of Education (#R305H050125).

References

McDermott, L. C., Rosenquist, M. L. & van Zee, E. (1987). Student difficulties in connecting graphs and physics: Examples from kinematics. *American Journal of Physics, 55*, 503.

Pfundt, H., Duit, R. (2000). *Bibliogaphy: Students' Alternative Frameworks and Science Education* (5th Ed.). Kiel, Germany: Institute for Education Science.

Ratcliff, R., 1978. A theory of memory retrieval. *Psychol. Rev.* 85, 59–108.

Schnipke, D. L., & Scrams, D. J. (1997). Modeling item-response times with a two-state mixture model. *Journal of Educational Measurement*, 34, 213–232.

Smith, P.L., Vickers, D., 1988. The accumulator model of two-choice discrimination. J. Math. Psychol. 32, 135–168.

van der Linden, W. J. (2008). Using response times for item selection in adaptive testing. Journal of Educational and Behavioral Statistics, 33, 5-20.

van der Linden, W. J., & van Krimpen-Stoop, E. M. L. A. (2003). Using response times to detect aberrant response patterns in computerized adaptive testing. *Psychometrika*, 68, 251–265.

Wickelgren, W.A., 1977. Speed-accuracy tradeoff and information processing dynamics. *Acta Psycholog.* 41, 67–85.

Effects of Problem Context on Strategy Use within Functional Thinking

Katherine L. McEldoon (Katherine.L.McEldoon@Vanderbilt.Edu)
Caroline Cochrane-Braswell (Caroline.E.Cochrane-Braswel@Vanderbilt.Edu)
Bethany Rittle-Johnson (Bethany.Rittle-Johnson@Vanderbilt.Edu)
Department of Psychology & Human Development, Vanderbilt University, Peabody College #0552, 230 Appleton Place
Nashville, TN 37203-5701 USA

Abstract

The effect of problem format on problem solving strategy selection is investigated within the early algebra domain of functional thinking. Functional Thinking is a type of algebraic reasoning appropriate for elementary students, in which a relationship exists between two sets of values. Three function table problems were given to students in grades two through six (N=232) in three different problem contexts. Problem context affected student strategy selection. Presenting the problem with non-indexical X values elicited the most correct strategy use, whereas the format with indexical X values elicited the most naïve and incorrect strategy use. Presenting the problem in a story context did not help correct strategy selection, but it decreased incorrect strategy use. Findings highlight factors influencing strategy selection, and have implication for instructional design and problem solving.

Keywords: Problem Solving; Context; Strategy Use; Mathematics; Functional Thinking; Algebra.

Problem Format and Problem Solving

The format a problem is presented in can affect how well a student understands the underlying concepts and skills the item is tapping (Collins & Ferguson, 1993; Day, 1988; Kirshner, 1989; Zhang, 1997). One problem context is a story context. Teachers and researchers often believe that story problems are harder for students than symbolic problems (Nathan & Koedinger, 2000; Nathan, Long, Alibali, 2002). National assessment data support this notion. Elementary student performance on story problems is generally worse than symbolic problems in the US (Carpenter, Corbitt, Kepner & Reys, 1980; Koba, Carpenter & Swafford, 1989). However, linguistic difficulties seem to account for younger children's poor performance on arithmetic story problems rather than inadequate knowledge of mathematics (Briars & Larkin, 1984; Cummins et al., 1988; de Corte, Verschaffel, & de Win, 1985; Hudson, 1983; Kintsch & Greeno, 1985; Riley et al. 1983).

Once students have proficient linguistic skills, story contexts can have an advantage. In high school students, a verbal advantage of story problems has been found with in algebra and arithmetic. The advantage of story problems was not only a consequence of situated world knowledge facilitating understanding. This advantage was also due to difficulties comprehending the formal symbolic representations of the symbolic problem formats (Koedinger & Nathan, 2004). The effect of problem context was further clarified in college students. A story context was advantageous when the underlying problem was simple, but a symbolic context was best when the problem was complex, presumably because these students had expertise interpreting the symbolic notation (Koedinger, Alibali & Nathan, 2008). Based on these findings, the ideal problem context seems to be dependent on the student's relative familiarity with linguistic and mathematic symbol systems.

When introducing early algebra concepts to elementary school students, mathematics education researchers stress the importance of rich and intuitive background contexts. These are thought to ground students' understanding of the new mathematic concepts they are learning (Carraher, Martinez, & Schliemann, 2008). Story contexts were found to help third grade students solving arithmetic problems over comparable symbolic contexts (Baranes, Perry, & Stigler, 1989). These ideas are in line with learning theories that have emphasized the role of contextual knowledge in supporting the development of symbolic knowledge (e.g., Greeno, Collins, & Resnick, 1996; Vygotsky, 1978).

Problem context can be varied in ways other than adding a story. The presentation of numeric information can be changed in ways which might alter how much attention is given to surface features versus the deep structure of the problem (Bassok, 1996). Having an understanding of the deep structure of a problem is important for fully understanding and correctly solving a problem.

This study investigates the effect of problem context on problem solving strategy within functional thinking, a type of early algebraic reasoning. Functional thinking tasks are appropriate for students that range in age from early elementary, where story context has been shown to hurt, to early middle, where story context has been shown to help. Giving the same task to students in this age range will help elucidate the effect of problem context on problem solving strategy use.

Functional Thinking

Functional Thinking is a type of mathematical thinking which focuses on the relationship between two (or more) varying quantities, specifically the kinds of thinking that lead from specific relationships to generalizations of that relationship across instances (Smith, 2008). The understanding of functions is also one of the core strands of the National Council of Teachers of Mathematics expectations for mathematics curriculum. At the heart of functional thinking is a relationship between two particular quantities that can be described by a *rule of correspondence* (Blanton & Kaput, 2005). This rule of correspondence can

be used to find other sets of particular quantities that adhere to the same rule.

Functional Thinking encapsulates some of the most important core components of early algebraic reasoning, such as generalization and covariation, and provides a developmentally appropriate way to scaffold these ways of thought through elementary mathematics education.

X	1	2	3	4	5
Y	5	6	7	8	9

Figure 1: Function Table (Y = X + 4).

Difficulties in Functional Thinking

A critical aspect of functional thinking is understanding the functional relationship between XY pairs. Functional thinking problems can be represented in a *function table* (Figure 1). A function table has an X and a Y column, filled with values that are all related by a function (e.g., Y = X + 4). This is the functional relationship of the table (Carraher et al., 2008). An understanding the functional relationship requires considering the relationship across the columns; between the X and Y values. However, the table can be interpreted another way, by only looking at the relationships within one column, such as between Y_1 and Y_2. This is the recursive relationship within the table (Carraher et al., 2008). Considering this recursive relationship is often temping, particularly when the X values are arranged indexically, with regular intervals, and therefore there are also regular intervals between the Y values. When the problem is presented in this format, to find later Y values in the function table, all one would have to do is extend the pattern within the Y values. However, this relationship is only useful when the X values increase at a constant rate. Additionally, this relationship cannot be efficiently used to predict a Y value for new X value.

Children tend to begin with this recursive strategy, particularly when they are unfamiliar with problems of this type (Carrahar et al., 2008). Broadly speaking, the power of functions is in the functional relationship between X and Y values, so a focus on the recursive relationship is misguided.

Mathematics educators have suggested different problem presentation contexts to help students get out of using the recursive strategy and into using the functional XY strategy. One way is to present the function table with an X axis that has irregular intervals between values (Carraher & Earnest, 2003; Warren & Cooper, 2005), or even clearly defined visual breaks in the table structure itself (Carraher et al., 2008; Schliemann, Carraher, & Brizuela, 2001). These break up the regular pattern in the Y values, thus discouraging children's strategy of simply looking to only the pattern in the Y values to determine the missing Y values later in the table.

Another way this can be overcome is to present the function problem with a story context, so the student can have an intuitive understanding of the underlying functional relationship. A story context can help ensure that students are considering the relationship between multiple input and output values (Schliemann et al., 2001). In this way, students are less likely to utilize a shallow recursive strategy. These instructional techniques help guide students away from the initial recursive strategy and into the correct functional strategy.

There is much writing as to which problem contexts are best for learning, but no systematic investigation for elementary level functional thinking problems. This study investigates the effect of problem context on strategy use within function table problems.

X	Y
2	6
3	7
4	8
5	9
6	
14	
	25
41	

X	Y
2	6
4	8
5	9
7	11
8	
14	
	25
41	

Cost of Present	Cost of Present with Gift Wrapping
2	6
3	7
4	8
5	9
6	
14	
	25
41	

Story Context: At a gift shop, you can pay extra to have your present gift-wrapped, as shown in the table below. What is the *total cost of the present with gift-wrapping* if the *cost of the present* is $6? $14? What about $41? If the *total cost of a present with gift-wrapping* is $36, what was the *cost of the present* itself?

Figure 2: Function Table Formats. Indexical, Non-Indexical, and Story Context.

Method

An assessment on functional thinking was given to students in grades two through six. Three function table items were included, which asked the students to fill in missing Y values in a function table, and to find the rule of correspondence. The underlying functions for these items were additive (Y = X + 4), multiplicative (Y = 3X), and a combination (Y = 3X + 2). These three items were presented in three contexts: a function table with indexically increasing X values and no story context (*indexical*), a function table with non-indexically increasing X values and no story context (*non-indexical*), and a story problems with indexically increasing X values (*story*) (see Figure 2). The items in these three conditions were kept as similar as possible, with the only differences being the factors that we were manipulating. The story contexts were about the cost of having a present gift-wrapped, saving money for a bicycle, and how many people could be seating at different arrangements of dinner tables. These story contexts were adapted from instructional materials created by math education researchers (e.g. Schliemann et al., 2001). The rule was not articulated in the story context and had to be deduced from the function table values. Each individual assessment contained the additive, multiplicative, and combination function table problems in the same context condition. All problems on a given assessment were in the

same format, therefore there were three versions of the assessment; indexical, non-indexical, and story.

The assessments were randomly distributed to 232 2nd through 6th grade students in a middle class suburban elementary and middle school in the southeastern United States. The general instructions for the assessment were read aloud to the second grade students, but they read the individual problems themselves.

Coding The students' work was coded for strategy use. The student's strategy was determined from the values they wrote in the function tables. Students' strategy use was coded as *correct* if they used the correct functional strategy, *recursive* if they used a recursive strategy, and *other*. If the student gave the correct entries, regardless of a correctly written rule of correspondence, or gave an incorrect entry for one blank, but gave a correct rule of correspondence, they were coded as *correct*. Students were given a *recursive* code if they had filled in the table by looking at the pattern in one column, instead of the relationship between the two columns. Students often used *other* strategies, such as an incorrect functional strategy, a mix of a functional and recursive strategy, an indiscernible strategy, or if the student left the table blank. There were no systematic differences in other strategy use of these types between conditions, and so they were collapsed in all further analyses. See Table 1 for a breakdown of strategy use by condition. Only correct and recursive strategy use was considered in this analysis.

Strategy	Description	Sample Student Response	Frequency		
			Index	Story	Non Index
Correct	Used correct functional rule to fill in table	**Y = 3X** X: 2 3 4 5 6 12 52 Y: 6 9 12 15 **18 36 156**	39.1%	39.4%	48.3%
Recursive	Filled in table following Y pattern, instead of between X and Y	**Y2 = Y1 + 3** X: 2 3 4 5 6 12 52 Y: 6 9 12 15 **18 21 24**	16.4%	9.1%	6.25%
Other	Incorrect Functional, Mixed Functional and Recursive, Unclear, and Blank		44.5%	51.5%	45.5%
Coded as strategy even if one entry in the table was incorrect or blank					

Table 1: Strategy Use Percentages by Condition

Results

We compared the effect of problem context (indexical, non-indexical, and story) on strategy use (correct or recursive). There was an overall effect of problem format on both correct and recursive strategy use.

Correct Strategy Use

The correct strategy was utilized the most overall, with it being used 39% of the time in both the indexical and story context, and 48% of the time in the non-indexical context (See Figure 3). The effect of problem context on correct strategy use was evaluated through a series of ANCOVAs with correct strategy use as a dependent variable, condition and grade as between subjects factors and a grade by condition interaction term. Grade was treated as a continuous variable. The initial model tested for a grade by condition interaction, which was not significant, and therefore the interaction term was dropped from all further analyses. Problem context had a significant effect on correct strategy use, $F(2, 225) = 3.23$, $p = .042$, $\eta^2 = .028$. A post hoc analysis of correct strategy use revealed that differences between conditions were significant when comparing the *non-indexical* and *story problem* contexts, $F(1,151) = 5.74$, $p=.018$, $\eta^2 = .037$. The difference between the *indexical* and *non-indexical* was marginal, $F(1,149) = 2.778$, $p=.098$, $\eta^2 = .018$. There was no difference in correct strategy use in the *indexical* and *story* contexts $F(1,146) = .523$, $p = .47$, $\eta^2 = .004$. This pattern of results was the same when students were split into younger (2nd and 3rd) and older (4th through 6th) groups, showing that this effect was not dependent on grade. Average accuracy performance was similar within these groupings, and so were collapsed for summative analyses. The younger students used the correct strategy in the *non-indexical* condition the most (29.1%), and less in the *indexical* and *story* contexts (15.5% and 11.6%). The older students used it 66.7% in the *non-indexical* context, and 55.8% and 61.2% in the *indexical* and *story* contexts. Overall, the *non-indexical* context was the most conducive to the correct problem solving strategy, and there was no difference in strategy use in the *indexical* and *story* contexts.

Recursive Strategy Use

The recursive strategy was utilized less often, with it being used 16% of the time in the *indexical* context, 6% of the time in the *non-indexical* context, and 9% of the time in the *story* context. Problem context had a significant effect on recursive strategy use $F(2, 225) = 3.49$, $p=.032$, $\eta^2 = .03$. There was a significant difference in strategy use between the *indexical* and *non-indexical* contexts $F(1,149) = 6.217$, $p=.014$, $\eta^2 = .04$. There was no significant difference when directly contrasting the other conditions (*story* vs. *non-indexical*, $F(1,151) = 1.003$, $p=.318$, $\eta^2 = .007$; *story* vs. *indexical*, $F(1,146) = 2.28$, $p=.133$, $\eta^2 = .015$). Again, this pattern of results was the same when students were split into younger (2nd and 3rd) and older (4th through 6th) groups. The younger students used the recursive strategy in the *indexical* condition the most (28.2%), and less in the *non-indexical* and *story* contexts (13.6% and 13.6%). The older students used it 6.2% in the *indexical* context, and 2.3% and 5.4% in the *non-indexical* and *story* contexts. This effect was not dependent on grade. Interestingly, there was a trend towards a stronger effect of problem context on recursive strategy use when the type of underlying function (i.e., multiplicative) was difficult for the student. The *indexical* context elicited the most recursive strategy use, and there was no difference in strategy use in the *non-indexical* and *story* contexts.

Discussion

Problem context had an effect on problem solving strategy use. Particularly, the *non-indexical* context encouraged the use of the correct strategy relative to other formats, and the *indexical* context encouraged the use of the recursive strategy relative to other formats. Interestingly, the *story* context discouraged use of the recursive strategy, but it did not encourage use of the correct strategy.

These findings have direct implications for the teaching and learning of function tables. In pedagogical contexts, function table problems should be presented with non-indexical X values to facilitate student understanding.

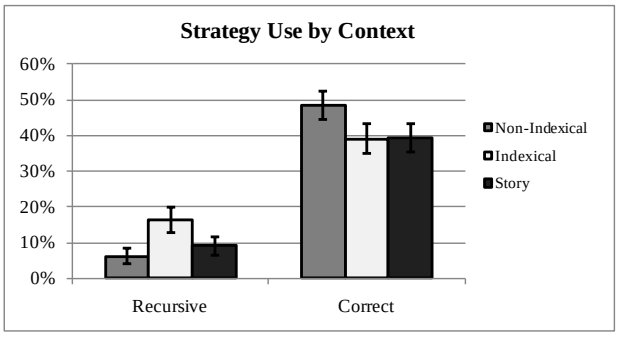


Figure 3: Problem Solving Strategy Use by Problem Context.

The indexicality of the X values had a large effect on student strategy use. Particularly, the *indexical* context encouraged the use of the naïve and incorrect recursive strategy. This could be the case because the students may have utilized the surface feature of the constant pattern in the Y values, and found it sufficient to determine the missing values. Specific aspects of content, context, and phrasing of a problem often play a crucial role in helping people determine the structure of a problem. Because of this, different structures may be abstracted from formally isomorphic problems that have different surface features (Bassok, 1996). The differences in surface features between function tables with indexical and non-indexical X values seemed to have been enough to invoke different structural interpretations in students. By arranging surface features of a problem, the learner's attention can be directed in more or less efficient manner to the underlying structure.

The story context did reduce the use of the naïve recursive strategy, but it did not support use of the correct strategy. Previous research suggests that the benefits of story contexts are dependent on the learner's relative familiarity with the linguistic and mathematic symbol systems (Koedinger, Alibali & Nathan, 2008; Rittle-Johnson & Koedinger, 2005). Our population included a range of students whose reading ability varied from novice to proficient, allowing us to address the effect of story context on students with different reading skill. The second and third grade students read the story context to themselves, yet the pattern of results between conditions was the same as those of the older students. This suggests that reading difficulties were not an issue, and that there was no verbal disadvantage for younger students. Standardized state test data was available for a subset of the 3rd through 6th grade students, and performance on our whole functional thinking assessment did not highly correlated with reading scores ($r(89) = .613$, $p < .01$). The story context seemed to reduce the tendency to focus on the Y_1Y_2 recursive relationship. This may be because the familiar and semantic information contained in the story helped form the students' understanding of the underlying problem structure. However, this story context was not enough to encourage correct strategy use, by considering the XY relationship. This effect of story context may be different from previous research findings, as the domain of functional thinking does not, at this elementary level, involve any mathematic symbolic notation. The problems only contain whole numbers, and the new concept is the focus on the XY relationship. As such, story contexts might not have as great a benefit as they do in arithmetic and algebra.

In this study, we wanted to isolate the effects of indexicality of the X values and a story context. How the two problem presentation features would interact was an open question. Given the results of this study, it is clear that future investigations should include story contexts with non-indexical X values. Perhaps the combination of both the real world context and numeric values without tempting surface patterns will be the most powerful in facilitating correct functional strategy use.

This study shows that seemingly small changes in problem context can affect the strategies a learner uses to solve a problem.

Acknowledgments

The first author is supported by a predoctorial training grant provided by the Institute of Education Sciences, U.S. Department of Education, through Vanderbilt's Experimental Education Research Training (ExpERTII) grant (David S. Cordray, Director; grant number R305B080025). The opinions expressed are those of the authors and do not represent views of the U.S. Department of Education. This work was also supported by an NSF CAREER Grant (#DRL0746565) awarded to Dr. Bethany Rittle-Johnson.

References

Baranes, R., Perry, M., & Stigler, J. W. (1989). Activation of real-world knowledge in the solution of word problems. *Cognition and Instruction, 6*, 287–318.

Bassok, M. (1996). Using Content to Interpret Structure: Effects on Analogical Transfer. *Current Diections in Psychological Science, 5*(2), 54-58.

Blanton, M., & Kaput, J.J. (2005). Characterizing a classroom practice that promotes algebraic reasoning. *Journal for Research in Mathematics Education, 36*(5), 412-446.

Briars, D. J., & Larkin, J. H. (1984). An integrated model of skill in solving elementary word problems. *Cognition and Instruction, 1,* 245–296.

Carpenter, T. P., Corbitt, M. K., Kepner, H. S., & Reys, R. E. (1980). Solving verbal problems: Results and implications from national assessment. *Arithmetic Teacher, 28*(1), 8–12.

Carraher, D., Martinez, M., & Schliemann, A. (2008). Early algebra and mathematical generalization. *ZDM Mathematics Education*, 403-422.

Carraher, D.W., & Earnest, D. (2003). Guess my rule revisited. *Proceedings of the 27th International Conference for the Psychology of Mathematics Education*, Honolulu, HI.

Collins, A., & Ferguson, W. (1993). Epistemic forms and epistemic games: Structures and strategies to guide inquiry. *Educational Psychologist, 28*, 25–42.

Cummins, D., Kintsch, W., Reusser, K., & Weimer, R. (1988). The role of understanding in solving word problems. *Cognitive Psychology, 20,* 405–438.

Day, R.S. (1988). Alternative representations. In G.H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 22). San Diego, CA: Academic Press.

de Corte, E., Verschaffel, L., & deWin, L. (1985). Influence of rewording verbal problems on children's problem representations and solutions. *Journal of Educational Psychology, 77,* 460–470.

Greeno, J.G., Collins, A.M., & Resnick, L.B. (1996). Cognition and learning. In D. Berliner & R. Calfee (Eds.), *Handbook of educational psychology*. New York: Macmillan.

Hudson, T. (1983). Correspondences and numerical differences between disjoint sets. *Child Development, 54,* 84–90.

Kirshner, D. (1989). The visual syntax of algebra. *Journal for Research in Mathematics Education, 20,* 274–287.

Koedinger, K.R., & Nathan, M.J. (2004). The Real Story Behind Story Problems: Effects of Representations on Quantitative Reasoning. *Journal of the Learning Sciences, 13*(2), 129-164.

Koedinger, K.R., Alibali, M. W., & Nathan, M.J. (2008). Trade-Offs Between Grounded and Abstract Representations: Evidence from Algebra Problem Solving. *Cognitive Science, 32*(2), 366-397.

Kouba, V.L., Carpenter, T.P., & Swafford, J.O. (1989). Number and operations. In M.M. Lindquist (Ed.), *Results from the fourth mathematics assessment of the national assessment of educational progress* (pp. 64–93). Reston, VA: National Council of Teachers of Mathematics.

Nathan, M.J., & Koedinger, K.R. (2000). Teachers' and researchers' beliefs of early algebra development. *Journal of Mathematics Education Research, 31*(2), 168–190.

Nathan, M.J., Long, S.D., & Alibali, M.W. (2002). Symbol precedence in mathematics textbooks: A corpus analysis. *Discourse Processes, 33*, 1–21.

Riley, M.S., Greeno, J.G., & Heller, J.J. (1983). Development of children's problem-solving ability in arithmetic. In H. Ginsburg (Ed.), *The development of mathematical thinking*. New York: Academic.

Rittle-Johnson, B., & Koedinger, K.R. (2005). Designing Knowledge Scaffolds to Support Mathematical Problem Solving. *Cognition and Instruction, 23*(3), 313-349.

Schliemann, A.D., Carraher, D.W., & Brizuela, B.M. (2001). When tables become function tables. *Proceedings of the XXV Conference of the International Group for the Psychology of Mathematics Education*, Utrecht, The Netherlands.

Smith, E. (2008). Representational Thinking as a Framework for Introducing Functions in the Elementary Curriculum. *Algebra in the Early Grades*. J. J. Kaput, Carraher, D. W., Blanton, M. L. New York, NY, Lawrence Erlbaum Associates.

Vygotsky, L.S. (1978). *Mind in society: The development of higher psychological processes.* Cambridge, MA: Harvard University Press.

Warren, E., & Cooper, T. (2005). Introducing Functional Thinking in Year 2: a case study of early algebra teaching. *Contemporary Issues in Early Childhood 6*, 150-162.

Zhang, J. (1997). The nature of external representations in problem solving. *Cognitive Science, 21*, 179–217.

The Application of the Less is More Hypothesis in Foreign Language Learning

Simone L. Chin (schin2@fau.edu)
Florida Atlantic University, Psychology Department
777 Glades Road, Boca Raton, FL 33431

Alan W. Kersten (akersten@fau.edu)
Florida Atlantic University, Psychology Department
777 Glades Road, Boca Raton, FL 33431

Abstract

The goal of this study was to test a foreign language teaching method inspired by Newport (1990)'s Less is More hypothesis. Computerized French language lessons were presented to 112 adults over two one-hour sessions. Learning trials were presented either in full sentences to resemble the adult learning environment, or in small phrases that incrementally increased in length to full sentences, resembling the steadily expanding processing capabilities of children. Trials were also ordered randomly or blocked such that multiple examples of the same objects and verbs were presented consecutively, in order to promote attention to individual words within those sentences. Language proficiency tests of vocabulary and grammar were administered after the lessons. The incremental and blocked conditions outperformed the randomly ordered full sentence conditions on the grammar measure. This outcome suggests that a teaching method based on Newport's Less is More hypothesis can be advantageous in learning a foreign language.

Keywords: adult language acquisition; constraints; starting small

Introduction

For many years, second language acquisition researchers and educators have been trying to sidestep the age effect problem in foreign language acquisition to help older children and adults reach a near native level of proficiency. Findings from this research demonstrate that (a) language is better learned at an earlier age, (b) despite numerous methods of explicit language instruction, older children and adult learners do not reach a native level of language proficiency, and (c) adults generally learn the word order and semantic aspects of language more quickly than children but usually never master the grammatical aspects (Newport, 1990). The demand for a solution to the age effect problem is essential to our multicultural society. A first step to solving this issue may be to investigate why young children are better language learners.

We suggest that second language educators may benefit from exploring developmental theories of language acquisition that (a) explain the robust findings of child-adult differences in language proficiency and (b) provide insight for methods of instruction to the second language teaching community. Developmental psychologists propose that the mind of a young child is more suitable for certain complex learning processes like language acquisition than is that of an older child or adult (Newport, 1990; Turkewitz & Kenny, 1982). Newport's Less is More theory explains that in the beginning stages of language learning, limited input helps children learn language (Newport, 1990). Young children's limited processing capacity and working memory only allow them to take in a small amount of the language heard around them, and as a result, they attend to limited language input such as individual words or morphemes. When learning a language, children must learn to map morphemes to specific meanings, and then combine those morphemes in original ways to create new sentences. Initial limited input may create the opportunity for children to analyze simple morphemes and create a small number of linguistic form-to-meaning mappings. When children's cognitive processes develop (working memory and processing capacity increase), they are then able to process more complex input, allowing them to learn the rules for combining morphemes in grammatical production. These cognitive processes fully develop around puberty (Newport, 1990).

The Less is More theory also explains why older children and adults do not learn language as well as young children. When older children and adults begin to learn language, they use their fully developed working memories and processing capacity to attend to complex sentences that contain multiple morphemes. From processing the complex input, adults (a) tend not to analyze individual morphemes but learn "frozen" combinations of multiple morphemes and (b) create many form-to-meaning mappings that are susceptible to noise. Out of the many possible morpheme mappings, only a few are correct, resulting in inconsistent and often incorrect language production. Therefore, adults do not learn the same morphological structure of a language as children.

Newport's research shows that late learners of ASL were more inclined to produce variable ungrammatical signs (Newport, 1990). Her theoretical explanation for these findings is the late learners immediately processed complex whole signs as units rather than analyzing the individual morphemes that make up a sign, encouraging the use of imitative unanalyzed signs. In contrast, the native and early learners of ASL used their developing cognitive abilities to process small parts of signs, enabling them to learn the individual morphemes and eventually produce original grammatically correct combinations of the morphemes.

Previous language learning studies found that initial limited exposure can be beneficial in learning the morphological structure of the language. Goldowsky &

Newport (1993) show that a computational model that has a filter restricting the amount of input when exposed to linguistic form-to-meaning mappings learns the correct mappings better than one without restrictions. Cochran, McDonald, & Parault (1999) demonstrate that restricting adults' language input by having them attend to an external working memory task or encouraging them to concentrate on small parts while being exposed to complex ASL morphology led to more consistent ASL production than did attending to the full complexity of the language. Kersten & Earles (2001) found that adults who were presented with an artificial language in small phrases that gradually increased in complexity performed better on vocabulary and morphology measures than did adults who were immediately presented with the full complexity of the language. These studies suggest that starting with limited input of language can facilitate learning.

Other research demonstrates learning benefits of a different approach to starting small. In particular, research on category learning has revealed that manipulating the order of learning trials, such that multiple examples of the same category are presented sequentially before exposing learners to the full range of variability in category exemplars, results in superior ultimate learning. For example, Sandhofer & Doumas (2008) show that manipulating the order of presentation so that children are presented with multiple examples of the same color category before introducing a new color led to better learning than did random presentation.

Additionally, Elio & Anderson (1981) introduced adults to two categories that differed on multiple attributes. The exemplars of each category were presented in either random or blocked order, in which multiple examples of the same category were presented before moving onto the next category. They found that learning the attributes of each category and generalizing new examples to the correct category were better in the blocked condition than the random condition.

These results suggest that blocking manipulations may function similarly to incremental presentation, encouraging learners to focus on the commonalities among members of an individual category and to ignore the variability associated with other, orthogonally-varying categories. Once learners acquire a basic vocabulary of individual categories, they may be in a better position to learn the more complex rules for combining those categories. Thus, teaching methods that encourage attention to simple information, and then gradually encouraging attention to more complex material, may be beneficial in learning the categories, and ultimately the structure, of language.

Present Research and Predictions

The present research tests the applicability of the Less is More hypothesis to second language learning by attempting to teach adults a foreign language using teaching approaches inspired by the theory. Adults participated in an unsupervised experiment consisting of two sessions of French language lessons in which they watched short videos and heard French descriptions. Two different methods were used to encourage participants to initially focus on individual French phrases before attempting to learn the complex grammar of entire French sentences. First, consistent with the method used by Kersten & Earles (2001), some participants were initially presented with individual phrases that gradually increased in complexity as learning progressed, whereas others were immediately presented with entire sentences. Second, consistent with the methods of Sandhofer & Doumas (2008) and Elio & Anderson (1981), some participants were presented with learning examples in a blocked order that encouraged the acquisition of individual words within the sentences that accompanied the videos, whereas others were presented with learning examples in a random order. Crossing these two factors led to four between-subjects conditions.

The incremental random (IR) condition presented a set of French lessons initially in individual phrases that gradually increased to full-length sentences. First, participants viewed a set of videos and heard only the direct object that corresponded to the video in the French language. In the second phase, participants viewed similar videos but heard entire phrases (including the verb and direct object). In the final phase, participants viewed the same videos from the previous phases but heard the full complex sentences that consisted of subject, verb and direct object. The trials within each learning phase were presented in random order. We predicted that initially presenting the descriptions in individual phrases would promote the learning of individual words and their meanings. As longer phrases were presented, participants were expected to learn additional words as well as the rules for combining those words

The incremental blocked (IB) condition presented participants with the French language in incrementally larger phrases as in the IR condition, but the order of presentation within each learning phase was blocked by similarity of object and verb. Similar to methods used in category learning, a block consisted of presenting two examples of the same object or verb sequentially before presenting a new object or verb. In the first learning phase, trials were blocked by similarity of object so that participants were presented with two trials containing the same object before moving onto a new object. In the second and third learning phases, in which participants heard verb phrases and full sentences, respectively, trials were blocked by similarity of verb meaning. The intention of blocking trials was to encourage acquisition of individual words within the speech stream. We predicted that both incremental and blocking manipulations would ultimately produce advantages in grammar acquisition.

The sentence blocked (SB) condition immediately presented the language in full sentences but in a blocked order similar to the IB condition. Even though these participants were not initially exposed to individual phrases, we expected that the blocked order of presentation would still encourage attention to individual words within the

speech stream, leading to better learning of those words and ultimately giving a grammar learning advantage to this group.

Emulating the adult learning experience, the sentence random (SR) or control condition was immediately exposed to the foreign language without restrictions. The SR condition viewed videos accompanied by full sentences in random order. Therefore, we did not expect these participants to focus on learning the individual morphemes or structure and do as well on the morphology measure as the other conditions.

Following the lessons were test trials assessing language acquisition. The measures of acquisition were vocabulary and grammar (word order and morphology) from the presented foreign sentences, as well as tests of inductive grammar, measuring the ability to extract grammatical rules and apply them to novel sentences. Although word order is a component of grammar, in this study it was measured separately from grammatical morphology. Performance on the vocabulary and word-order measures was predicted to be similar among all groups, as vocabulary and word order are generally acquired without difficulty in both adults and children. However, a disparity was expected in performance among conditions on both grammar measures that test participants' knowledge of morphology. The three experimental conditions (IR, IB, and SB) were expected to outperform the control condition (SR) in the measures of morphology.

Method

Participants

One hundred twelve native English speakers from Florida Atlantic University participated in this experiment. Only participants who reported in a language background questionnaire that they did not speak a Romance language and knew fewer than 30% of the French words on the vocabulary pre-test were included in the data analysis. The average age of the participants was 22.2 (SD = 4.6) years. Each participant was randomly assigned to one of the four conditions of the experiment.

Stimuli

French Language The stimuli included sentences of active and reflexive verb forms made up of high frequency French words. The French active sentences in the stimuli share similar structure to the English language. Each active sentence consists of a subject, action verb, and object. However, English and French languages have grammatical distinctions when conveying a reflexive action. In English, possessive pronouns are used to express a reflexive action, but in French, reflexive pronouns are used. The pronoun "se" is added and placed in front of the verb and definite articles describe the object rather than possessive pronouns. For example, a man brushing his hair is described as "L'homme se brosse les cheveux". If the "se" is omitted from the sentence, the statement changes the meaning to "The man brushes the hair," implying the hair of an object. The word order of the French language is similar to English. Lesson and test trials were in French and only the instructions and examples of the tasks were in English.

Trials All trials were programmed into Superlab Pro 4.0 and displayed on computers. Each trial consisted of a video playing on the screen and a corresponding French description presented audibly through headphones. French text was not available. Each video consisted of an actor performing a specific action on him/herself or on an object. Six different actors performed the same actions in different contexts. A female native French speaker recited French descriptions into a recorder and the recordings were linked to correspond with the videos. Each trial was approximately 3 seconds in length.

Learning Trials Learning trials were designed to teach participants the semantics and grammatical structure of sentences using 8 verbs and 16 nouns. Lessons were made up of three learning phases. Each learning phase consisted of 32 trials that presented 4 examples of each verb (2 in the active and 2 in the reflexive form) and 2 examples of each noun. In the IR condition, phase 1 comprised 32 trials of videos accompanied by only the direct object description from the corresponding sentences, phase 2 consisted of 32 trials accompanied by the verb and direct object, and phase 3 comprised 32 trials with full sentence descriptions. The IB condition presented the same trials from the IR condition in blocked order. Phase 1 consisted of videos and direct object descriptions, ordered such that the trials with the same direct objects were presented one after the other. Phases 2 and 3 involved the exact trials from the IR condition, but ordered such that trials involving the same verb were presented one after the other, first in the active form then in the reflexive form. The SB condition consisted of three phases of trials with full sentence descriptions, but presented in blocked order similar to the IB condition. Participants in the control condition (SR) were exposed to trials with full sentence descriptions in random order identical to the last phase of the IR condition for all learning phases. See Table 1 for an illustration of the presentation of two learning trials between the groups.

Test Trials To measure participants' knowledge of the French vocabulary, a judgment task of 16 trials of videos and corresponding French sentences were presented. The correct trial included the appropriate object and verb description of the video, whereas the incorrect trial contained an incorrect noun or verb. The incorrect French sentences were taken from the learning trials but linked to an incorrect video from the learning trials entailing a different object or action.

As one test of participants' knowledge of French grammar, a word order forced-choice task for the active and reflexive sentences was administered. The word order task consisted of 8 trials. The correct trials resembled the learning trials, whereas the incorrect trials included videos

Table 1: Illustration of two learning trials for each condition.

Experimental Condition	Learning Phase 1		Learning Phases 2 & 3		
	Video	**Audio**	**Video for 2 & 3**	**Audio Phase 2**	**Audio Phase 3**
Incremental Random	woman sprays car	"la voiture"	Same as phase 1	"asperge la voiture"	"La femme asperge la voiture."
	woman shakes bottle	"la bouteille"		"secoue la bouteille"	"La femme secoue la bouteille."
Incremental Blocked	woman sprays car	"la voiture"	woman sprays car	"asperge la voiture"	"La femme asperge la voiture."
	woman washes car	"la voiture"	woman sprays her face	"s'asperge le visage"	"La femme s'asperge le visage."
Sentence Blocked	woman sprays car	"La femme asperge la voiture."	woman sprays car	Same as Learning phase1	
	woman washes car	"La femme lave la voiture."	woman sprays her face	"La femme s'asperge le visage."	
Sentence Random	woman brushes her hair	"La femme se brosse les cheveux."	Same as phase 1	Same as Learning phase1	
	woman cuts the ticket	"La femme coupe le billet."			

from the learning trials linked to French sentences with incorrect word order.

The grammar forced-choice task assessed participants' understanding of the morphological structure underlying the active and reflexive sentence forms. This task comprised 16 test trials. The goal of this task was to determine if the participants could correctly identify, discriminate, and link the active and reflexive sentence forms to the appropriate video. In other words, would participants learn to understand that reflexive actions are linked to reflexive sentences with the "se" pronoun and that active sentences (without the reflexive pronoun) are used to express actions on objects rather than to self? The correct choices consisted of videos with the correct corresponding French sentences resembling the learning trials. The incorrect choices included videos with incorrect grammatical French sentences of the active or reflexive forms.

The purpose of the inductive task was to assess participants' ability to apply the learned French grammatical rules to new stimuli. First, there was a learning phase of 8 videos and sentences (4 active and 4 reflexive sentences), each presented twice. The intention of the learning phase was to introduce participants to new vocabulary. The testing phase consisted of a forced-choice task of 8 trials (4 reflexive and 4 active). The videos and French descriptions in the test trials were novel. In particular, a verb presented only in the active form during learning was presented in the reflexive form in testing. To succeed in the inductive task, participants had to use the grammar rules of verb forms extracted from the lessons to fit the video and description of the task. Since this was a forced-choice task, participants did not have to produce the verb; however, they had to decide which one of the provided sentences contained the correct verb form.

Procedure

The experiment consisted of two one-hour sessions. The first session entailed lessons and a vocabulary test. During the lessons portion, participants were instructed to view each video and listen to the French description. They were told to repeat the description after each trial. Participants viewed learning trials from each phase twice before moving on to the next phase, totaling 192 trials. Every participant was issued one-minute breaks after every 32 trials. Once the learning trials were all presented and another break was given, participants took the word-meaning test.

Two days later, the participants returned for the second session. First, participants viewed lessons identical to the third learning phase in session one, totaling 64 trials. After a break, participants completed the word meaning, grammar and word order measures. Following a final break, participants completed the inductive task, and then a short questionnaire on prior knowledge of the French language, concluding the procedure of the experiment.

Results

The results of this experiment are presented in Table 2. Each measure was scored as the percentage of the total correct acceptances and rejections out of the total number of test trials for each task. Analyses of the word meaning judgment task were split into two separate measures of verbs and nouns. A 2(noun vs. verb) X 2(1st session vs. 2nd session) X 2(incremental vs. sentence) X 2(block vs. random order) repeated measures ANOVA was conducted to investigate within and between group differences on word meaning measures. The results revealed a significant within-group difference between noun and verb learning, $F(1, 107) = 146.765$, $p < .001$, $MSE = 3.752$. Participants scored higher on the noun items than verb items of the vocabulary measures. However, there were no significant increment, blocking, or interaction effects on vocabulary (all ps $> .05$).

Table 2: Means (Standard Deviations) for Groups

Language Measures	Incremental Random n = 31	Incremental Blocked n = 25	Sentence Blocked n = 27	Sentence Random n = 29
Vocabulary				
1st session nouns	.742(.21)	.613(19)	.785(.18)	.830(.13)
2nd session nouns	.811(.13)	.825(.16)	.810(.18)	.806(.20)
1sr session verbs	.612(.19)	.567(.18)	.638(.20)	.625(.24)
2nd session verbs	.641(.16)	.610(18)	.690(.18)	.586(.21)
Word order	.879(.14)	.930(.09)	.912(.12)	.897(.12)
Grammar (based on trials from lessons)	.621(.14)	.655(.13)	.683(.18)	.586(.13)
Inductive grammar (based on trials from inductive task	.605(.20)	.640(.18)	.644(.25)	.543(.24)

A two way ANOVA was conducted on performance that tested participants' knowledge of word order for the active and reflexive French sentences. The results revealed no significant main or interaction effects on word order performance (all *p*s > .05).

Analysis of the grammar measures was conducted using three planned orthogonal contrasts. The first contrast compared the control participants (SR) to the three experimental conditions, examining whether manipulations that promote attention to low-level sentence elements promote acquisition of French grammar. The second contrast compared the IB condition to the IR and SB conditions, examining whether receiving both manipulations yields better knowledge of French grammar than receiving just one. The third contrast compared the incremental IR condition to the SB condition, examining whether one manipulation promoting attention to low-level sentence elements yielded better knowledge of French grammar than the other.

The first of these three contrasts was significant, indicating that the three experimental conditions scored significantly higher on the grammar tasks than did the control condition (p = .023). However, the other two contrasts were not significant (both *p*s > .05). These results suggest that the increment or blocking methods produced better performance than the sentence random method on the grammar tasks, but having both the increment and blocked methods did not facilitate performance any better than having one of the two methods.

Discussion

The goal of this study was to investigate the usefulness of language teaching methods that were based on the Less is More hypothesis. The results suggest that presenting the language lessons in increments, or in a blocked order does not lead to better performance on the word meaning and word order measures. However, the results of the other grammar measures suggest that these teaching methods can be advantageous in grammar acquisition.

The generally high levels of performance on the vocabulary measures suggest that participants had little difficulty learning the meanings of the words. All groups learned more nouns than verbs. This finding is consistent with the notion that concrete count nouns are easier to learn than verbs, and with findings that children learn the names of objects more quickly than verbs (Gentner & Boroditsky, 2001). It can be said that the lessons promoted noun learning because the nouns were concrete, consistent and appeared at the end of each phrase.

The intent of the word order measure was not to test the effects of incremental presentation or order of the lessons but to confirm that the participants comprehended the lessons. All groups performed at similarly high levels near ceiling on the word order measure.

The crux of the experiment was the outcome of the grammar measures. Generally, adults have difficulty mastering the grammar of foreign languages. In this experiment, French language lessons inspired by the Less is More hypothesis were presented to adults as an attempt to overcome this challenge. The results revealed that learning strategies that encouraged the learning of language in small pieces by incremental presentation or blocked order of presentation facilitated learning the rules of grammar, with the three experimental groups performing better than the control group on these measures. Further analysis revealed that groups that used either incremental or blocked presentation, performed just as well as the condition that received both of these manipulations. These results suggest that either incremental presentation or blocking was sufficient to encourage attention to lower-level sentence elements and thus to ultimately yield better grammatical acquisition. The results of the study thus support our predictions that teaching methods inspired by the Less is More hypothesis may be fruitful in facilitating the acquisition of the grammar of a second language.

Limitations

The results of the experiment are consistent with Newport's Less is More theory, but there remain several limitations of the study that must be remedied before fully endorsing this theory. First, the grammar task involved only a single, relatively simple grammatical alternation. Despite this simplicity, the participants in the study still had difficulty learning the grammatical rule, performing only slightly above floor on the grammar measures. This difficulty likely reflects the limited amount of exposure to the language, involving only two one-hour lessons, and the

high demands of the task. High demands include auditory rather than visual presentation and a lack of explicit grammar instruction. Previous studies found that grammar performance is better in adults when using visual modes of instruction and testing and when using methods that encourage explicit rule learning instead of implicit learning (Conway et al., 2003, Dekeyser & Larson-Hall, 2001). The acquisition of more complex grammatical rules may require much longer, more varied training. However, this study provides a good steppingstone for further experiments.

Second, we designed our method with the assumption that children with limited cognitive processes preferentially attend to the ends of sentences, reflecting a recency effect in working memory. For that reason, the lessons in the incremental condition were presented in small increments starting with the last word of each sentence. This design gave participants simple and consistent lessons. However, the Less is More hypothesis does not state that children always attend to the last word or part of a sentence. If children do not always attend to the last word of a sentence but rather are equally likely to attend to any part of a sentence, then the incremental condition is not fully representative of children's language learning strategies. Changing the incremental condition to involve presenting randomly-chosen pieces of sentences would make the incremental condition less consistent, and as a result, language learning would likely be more difficult for adults.

Lastly, one may argue that the lessons resemble infant-directed talk (IDT). Studies suggest that IDT plays a role in language acquisition and can facilitate adult foreign vocabulary acquisition (Baldwin & Meyer, 2007; Golinkoff & Alioto, 1995). Though the Less is More hypothesis and IDT could potentially coexist in explaining children's language learning, further experimentation would be needed to investigate separate effects of IDT and incremental presentation on adults.

Implications

To address the challenge of learning foreign language grammar in adulthood, this study shows that foreign language educators may profit from incorporating teaching methods based on developmental theories. Presenting a foreign language in increments or in blocked order that promotes the learning of small pieces during the initial stages of language learning are alternative approaches for adults to learn the vocabulary and grammatical structure of a foreign language. Widely-used language teaching methods such as immersion programs or explicit instruction that heavily emphasize focus on semantics in the initial stages of foreign language learning tend to fall short of getting adults to the native level of proficiency in grammar (Harley, 1998). Although our proposal counters these established methods of adult foreign language instruction, it would be worthwhile for the foreign language community to further explore the potential of developmental theories such as Newport's Less is More hypothesis that can offer insightful new methods for foreign language learning.

Acknowledgments

The authors would like to thank David Bjorklund and Erika Hoff for their helpful comments on this research.

References

Baldwin, D. & Meyer, M. (2007). How inherently social is language? In E. Hoff & M. Shatz (Eds.), *Blackwell handbook of language development*. Oxford, U.K.: Blackwell Publishing.

Cochran, B., McDonald, J. & Parault, S. (1999). Too smart for their own good: The disadvantage of superior processing capacity for adult language learners, *Journal of Memory and Language, 41,* 30-58.

Conway, CM, Ellefson, MR., Christiansen, MH. (2003). When less is less and when less is more: Starting small with staged input. *Proceedings of the 25th Annual Conference of the Cognitive Science Society* (pp.270-275). Boston, MA: Cognitive Sciency Society.

DeKeyser, R., & Larson-Hall, J. (2005). What does the critical period really mean? In J. F. Kroll & A. M. D. de Groot (Eds.), *Handbook of bilingualism: psycholinguistic approaches* (pp. 88-108). New York: Oxford University Press.

Elio, R. & Anderson, J. (1981). The effects of category generalizations and instance similarity on schema abstraction, *Memory & Cognition, 12,* 20-30.

Gentner, D., & Boroditsky, L. (2001). Individuation, relativity, and early word learning. In M. Bowerman & S. Levinson, (Eds.), *Language acquisition and conceptual development*. Cambridge, UK: Cambridge University Press.

Goldowsky, B. & Newport, E. (1993). Modeling effects of processing limitations on the acquisition of morphology: The less is more hypothesis. In E. V. Clark (Ed.), *The proceedings of the Twenty-Fourth Annual Child Language Research Forum*. Stanford, CA: Center for the Study of Language and Information.

Golinkoff, R. & Alioto, A. (1995). Infant-directed speech facilitates lexical learning in adults hearing Chinese: implications for language acquisition, *Child Language, 22,* 703-726.

Harley, B. (1998). The role of focus-on-form in child second language acquisition. In C. Doughty & J. Williams (Eds.), *Focus on form in classroom second language acquisition*. Cambridge: Cambridge University Press.

Kersten, A. & Earles, J. (2001). Less really is more for adults learning a miniature artificial language, *Journal of Memory and Language, 44,* 250-273.

Newport, E. (1990). Maturational constraints on language learning. *Cognitive Science, 14,* 11-28.

Sandhofer, C. & Doumas, L. (2008). Order of presentation effects in learning color categories, *Journal of Cognition and Development, 9(2),* 194-221.

Turkewitz, G. & Kenny, P. (1982). Limitations on input as a basis for neural organization and perceptual development: A preliminary theoretical statement, *Developmental Psychobiology, 15(4),* 357-368.

Are People Successful at Learning Sequential Decisions on a Perceptual Matching Task?

Reiko Yakushijin (yaku@cl.aoyama.ac.jp)
Department of Psychology, Aoyama Gakuin University, Shibuya, Tokyo, 150-8366, Japan

Robert A. Jacobs (robbie@bcs.rochester.edu)
Department of Brain & Cognitive Sciences, University of Rochester, Rochester, NY 14627, USA

Abstract

Sequential decision-making tasks are commonplace in our everyday lives. We report the results of an experiment in which human subjects were trained to perform a perceptual matching task, an instance of a sequential decision-making task. We use two benchmarks to evaluate the quality of subjects' learning. One benchmark is based on optimal performance as defined by a dynamic programming procedure. The other is based on an adaptive computational agent that uses a reinforcement learning method known as Q-learning to learn to perform the task. Our analyses suggest that subjects learned to perform the perceptual matching task in a near-optimal manner at the end of training. Subjects were able to achieve near-optimal performance because they learned, at least partially, the causal structure underlying the task. Subjects' learning curves were broadly consistent with those of model-based reinforcement-learning agents that built and used internal models of how their actions influenced the external environment. We hypothesize that, in general, people will achieve near-optimal performances on sequential decision-making tasks when they can detect the effects of their actions on the environment, and when they can represent and reason about these effects using an internal mental model.

Keywords: sequential decision making; optimal performance; dynamic programming; reinforcement learning

Introduction

Tasks requiring people to make a sequence of decisions to reach a goal are commonplace in our lives. When playing chess, a person must choose a sequence of chess moves to capture an opponent's king. When driving to work, a person must choose a sequence of left and right turns to arrive at work in a timely manner. And when pursuing financial goals, a person must choose a sequence of saving and spending options to achieve a financial target. Interest in sequential decision-making tasks among cognitive scientists has increased dramatically in recent years (e.g., Busemeyer, 2002; Chhabra & Jacobs, 2006; Fu & Anderson, 2006; Gibson, Fichman, & Plaut, 1997; Gureckis & Love, 2009; Lee, 2006; Sutton & Barto, 1998; Shanks, Tunney, & McCarthy, 2002).

Here, we are interested in whether people are successful at learning to perform sequential decision-making tasks. There are at least two ways in which the quality of learning can be evaluated. These ways differ in terms of the benchmark to which the performances of a learner are compared. One way uses a benchmark of optimal performance on a task. Analyses based on optimal performance are referred to as ideal observer analyses, ideal actor analyses, or rational analyses in the literatures on perception, motor control, and cognition, respectively. At each moment during training with a task, a learner's performance can be compared to the optimal performance for that task. If a learner achieves near-optimal performance at the end of training, then it can be claimed that the learner has been successful.

A second way of evaluating a learner is to compare the learner's performances with those of an adaptive computational agent that is trained to perform the same task. We consider here an agent that learns via "reinforcement learning" methods developed by researchers interested in artificial intelligence (Sutton & Barto, 1998). Cognitive scientists have begun to use reinforcement learning methods to develop new theories of biological learning (e.g., Busemeyer & Pleskac, 2009; Daw & Touretzky, 2002; Schultz, Dayan, & Montague, 1997; Fu & Anderson, 2006). To date, however, there are few comparisons of the learning curves of people and agents based on reinforcement learning methods. Because reinforcement learning is regarded as effective and well-understood from an engineering perspective, and as plausible from psychological and neurophysiological perspectives, the performances of agents based on this form of learning can provide useful benchmarks for evaluating a person's learning. If a person's performance during training improves at the same rate as that of a reinforcement-learning agent, then it can be argued that the person is a successful learner. If a person's performance improves at a slower rate, then the person is not learning as much from experience as he or she could learn. Experimentation is often required to identify the cognitive "bottlenecks" preventing the person from learning faster. Lastly, if a person's performance improves at a faster rate, then this suggests that the person is using information sources or information processing operations that are not available to the agent. A new, more complex agent should be considered in this case.

We report the results of an experiment in which human subjects were trained to perform a perceptual matching task. This task was designed to contain a number of desirable features. Importantly, the perceptual matching task is an instance of a sequential decision-making task. Subjects made a sequence of decisions (or, equivalently, took a sequence of actions) to modify an environmental state to a goal state. In addition, efficient performance on the perceptual matching task required knowledge of how different properties of an environment interacted with each other. In many everyday tasks, people are required to understand the interactions, or "causal relations", among multiple components (Busemeyer, 2002; Gopnik &

Shulz, 2007). For example, when reaching for a coffee mug, a person must understand that forces exerted at the shoulder also influence the positions and velocities of the elbow, wrist, and fingers. To make an efficient movement, a person must use this knowledge of the causal interactions among motor components to design an effective motor plan.

Subjects' performances on the perceptual matching task were evaluated via two benchmarks. Using an optimization technique known as dynamic programming, optimal performance on this task was calculated. In addition, computer simulations of an adaptive agent were conducted in which the agent was trained to perform the perceptual matching task using a reinforcement learning method known as Q-learning (Sutton & Barto, 1998; Watkins, 1989). Comparisons of subjects' performances during training with optimal performance and with those of the adaptive agent suggest that: (i) subjects learned to perform the perceptual matching task in a near-optimal manner at the end of training; (ii) subjects learned, at least partially, the causal structure underlying the task; (iii) subjects' learning curves were consistent with those of model-based reinforcement-learning agents; and (iv) subjects may have learned by building and using mental models of how their actions influenced the external environment. Additional details and results are reported in Yakushijin & Jacobs (2010).

Experiment

Methods: Twenty-four undergraduate students at the University of Rochester participated in the experiment. Subjects were paid $10 for their participation. All subjects had normal or corrected-to-normal vision. Subjects were randomly assigned to one of six experimental conditions. Each condition included both training and test trials. Only the results of training trials are discussed here due to space limitations.

On a training trial, subjects performed a perceptual matching task which used visual objects from a class of parameterized objects known as "supershapes" (highly realistic but unfamiliar shapes; see Gielis, 2003). The parameters were latent (hidden) variables whose values determined the shapes of the objects. On each trial, subjects viewed a target object, a comparison object, and a set of six buttons (see left panel of Figure 1). Buttons were organized into three pairs, and each pair could be used to decrease or increase the value of an action variable. By pressing the buttons, subjects could change the values of the action variables which, in turn, changed the values of the parameters underlying the comparison object's shape which, in turn, changed the shape of the comparison object. Subjects' task was to press one or more buttons (i.e., to change the values of the action variables) to modify the shape of the comparison object until it matched the shape of the target object using as few button presses as possible.

An experimental condition was characterized by a specific set of causal relations among the latent shape parameters. For example, one such set is schematically illustrated in the right panel of Figure 1. Here, the three action variables are denoted A, B, and C. These variables are observable in the sense that subjects could directly and easily control their values through the use of the buttons. The values of the action variables determined the values of the shape parameters, denoted X, Y, and Z. Note that there are causal relations among the shape parameters. According to the network in Figure 1, if the value of X is changed, then this leads to a modification of Y which, in turn, leads to a modification of Z. The shape parameters determine the shape of the comparison object, whose perceptual features are denoted f_1, f_2, f_3, f_4, f_5, and f_6. The perceptual features used by a subject to assess the similarity of target and comparison object shapes may only be implicitly known by a subject, and may differ between subjects.

Importantly, to efficiently convert the comparison object's shape to the target object's shape (i.e., with the fewest number of button presses) often requires an understanding of the causal relations among the shape parameters. For instance, if the values of parameters X, Y, and Z all need to be modified, a person who does not understand the causal relations among shape parameters may decide to change the value of action variable C (thereby changing shape parameter Z), then the value of action variable B (thereby changing Y and Z), and finally the value of action variable A (thereby changing X, Y, and Z). In many cases, this will be an inefficient strategy. A person with good knowledge of the causal relations among the shape parameters knows that he or she can change the values of X, Y, and Z with a single button press that decreases or increases the value of action variable A. Thus, a good understanding of the causal relations among the shape parameters will lead to efficient task performance, whereas a poor understanding of the causal relations will lead to many more button presses than necessary.

The six experimental conditions differed in the causal relations among the latent shape parameters X, Y, and Z. Two of the causal relations were "linear" structures (one parameter had a direct causal influence on a second parameter which, in turn, had a direct causal influence on a third parameter; e.g., $X \rightarrow Y \rightarrow Z$ or $Y \rightarrow X \rightarrow Z$), two of the relations were "common cause" structures (one parameter had direct causal influences on the two remaining parameters; e.g., $Y \leftarrow X \rightarrow Z$ or $X \leftarrow Y \rightarrow Z$), and two of the relations were "common effect" structures (two parameters had direct causal influences on a third parameter; e.g., $X \rightarrow Y \leftarrow Z$ or $Y \rightarrow X \leftarrow Z$).

An experimental session consisted of 7 blocks of trials where a block contained a set of training trials followed by a set of test trials. (Test trials evaluated subjects' one-step look-ahead knowledge; on a test trial, a subject decided if a comparison object could be converted to a target object using a single button press, and the subject did not receive feedback. Again, test trials are not discussed here.) Each set contained 26 trials, one trial for each possible perturbation of a target object shape to form an initial comparison object shape.

Results: Task Performances: As a benchmark for evaluating subjects' performances on training trials, we computed optimal performances on these trials using an optimization method known as dynamic programming (Bellman, 1957).

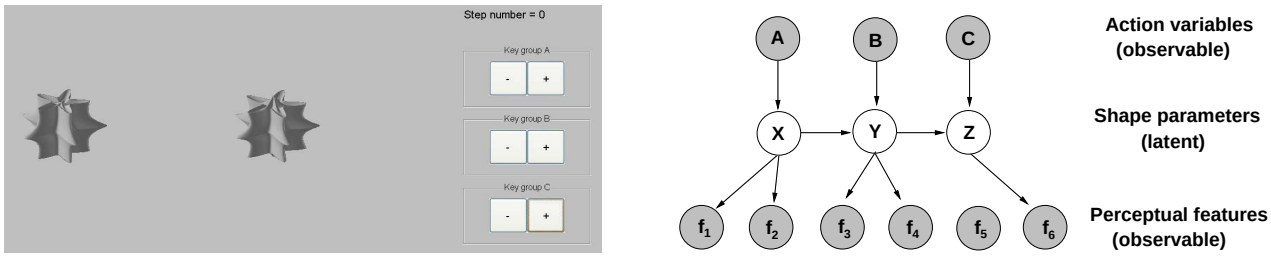


Figure 1: Left: Example of an experimental display. Right: Bayesian network representing the causal relations (in one of the experimental conditions) among the action variables, shape parameters, and perceptual features. For simplicity, the network does not represent the fact that subjects' button presses determined the values of the action variables.

In brief, dynamic programming is a technique for computing optimal solutions to multi-stage decision tasks. That is, dynamic programming finds the shortest sequences of actions that move a system from an initial state to a goal state when all states are fully observable. In the context of a training trial, the initial state corresponds to the initial values of the shape parameters X, Y, and Z for the comparison object, and the goal state corresponds to the values of the shape parameters for the target object. The dynamic programming algorithm is provided with full state information. This means that the algorithm knows the values of the comparison object's shape parameters at every time step. It also knows the state transition dynamics, meaning that it knows the causal relations among the shape parameters and, thus, knows how any button press will change the values of the shape parameters. Relative to our subjects, the dynamic programming algorithm is at an advantage. At the start of the experiment, our subjects did not know the values of the shape parameters or the causal relations among the parameters. Consequently, it would be impressive if subjects learned to perform the task as well as the dynamic programming algorithm.

We determined the optimal performances in the six experimental conditions via dynamic programming. Our analysis revealed that the range (1-5 steps or button presses) and the average length (2.54 steps) of the optimal action sequences were identical for all conditions. Thus, the conditions were well balanced in terms of their intrinsic difficulties.

Figure 2 shows subjects' learning curves on training trials in the two experimental conditions with linear causal structures among shape parameters. Due to space limitations, we do not show results for conditions with common-cause and common-effect structures, though subjects in these conditions showed very similar results to subjects in linear structure conditions (Yakushijin & Jacobs, 2010). Eight subjects participated in linear structure conditions and, thus, the figure contains eight graphs. The horizontal axis of each graph gives the block number, and the vertical axis gives the average difference between the number of steps (i.e., button presses) used by a subject during a trial and the optimal number of steps for that trial as computed by the dynamic programming procedure. These graphs show a number of interesting features. Many subjects found the task to be difficult toward the start of the experiment and, thus, their performances were highly sub-optimal during this time period. However, every subject learned during the course of the experiment. Importantly, every subject achieved near-optimal performance at the end of training: The average difference between a subject's performance and the optimal performance at the end of training is less than 1/2 of a step (mean = 0.434; standard deviation = 0.324).

Results: Causal Learning: The data from the training trials show that subjects achieved near-optimal performances. These results are consistent with the idea that subjects learned about the causal relations among the latent shape parameters. Additional analyses of training and test trials, not described here due to space limitations, confirm that subjects did indeed learn (at least partially) about these causal relations, and that this knowledge played a role in their task performances. Details can be found in Yakushijin & Jacobs (2010).

Reinforcement Learning Agents

Above, our analysis of subjects' data used a benchmark of optimal performance based on dynamic programming. Although very useful, this analysis does not allow us to evaluate the quality of subjects' rates of learning. To do so, we use a different benchmark based on an adaptive computational agent that uses a reinforcement learning method known as Q-learning to learn to perform the perceptual matching task (Sutton & Barto, 1998; Watkins, 1989). Without going into the mathematical details, the reader should note that Q-learning is an approximate dynamic programming method (Si et al., 2004). It is easy to show that, under mild conditions, the sequence of decisions found by an agent using Q-learning is guaranteed to converge to an optimal sequence found by dynamic programming (Watkins & Dayan, 1992). Hence, the benchmarks based on dynamic programming and on Q-learning are related.

In a reinforcement learning framework, it is assumed that an agent attempts to choose actions so as to receive the most reward possible. The agent explores its environment by assessing its current state and choosing an action. After executing this action, the agent will be in a new state, and will receive a reward (possibly zero) associated with this new state.

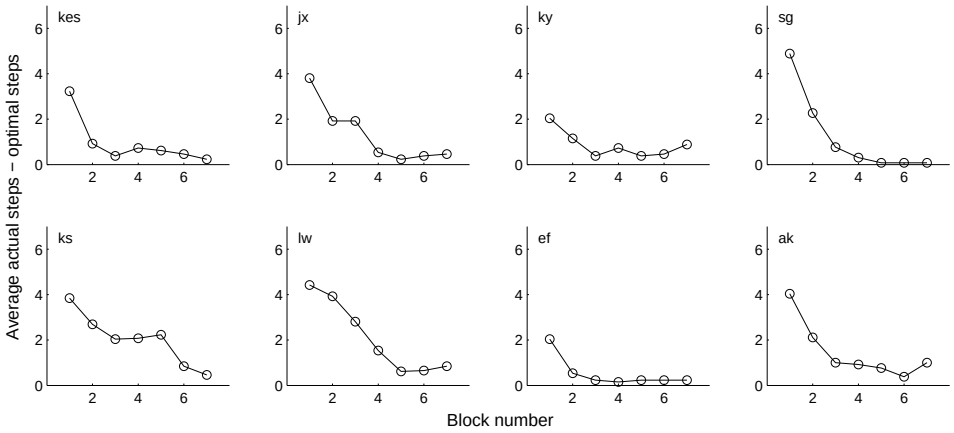


Figure 2: Subjects' learning performances on training trials in the two experimental conditions with linear causal structures among shape parameters (top row: $X \rightarrow Y \rightarrow Z$; bottom row: $Y \rightarrow X \rightarrow Z$).

The agent adapts its behavior in a trial-by-trial manner by noticing which actions tend to be followed by future rewards and which actions are not. To choose good actions, the agent needs to estimate the long-term reward values of selecting possible actions from possible states. Ideally, the value of selecting action a_t in state s_t at time t, denoted $Q(s_t, a_t)$, should equal the sum of rewards that the agent can expect to receive in the future if it takes action a_t in state s_t: $Q(s_t, a_t) = E[\sum_{k=0}^{\infty} \gamma_k \, r_{t+k+1}]$ where t is the current time step, k is an index over future time steps, r_{t+k+1} is the reward received at time $t+k+1$, and γ $(0 < \gamma \leq 1)$ is a term that serves to discount rewards that occur in the far future more than rewards that occur in the near future. An agent can learn accurate estimates of these ideal values on the basis of experience if it updates its estimates at each time step using the equation:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_{t+1} + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t)]$$

where the agent makes action a_t in state s_t and receives reward r_{t+1}, and α is a step size or learning rate parameter (Sutton & Barto, 1998; Watkins, 1989).

In our first set of simulations in which a reinforcement-learning agent was trained to perform the perceptual matching task, all "Q-values" were initialized to zero, the discount rate γ was set to 0.7, and the learning rate α was set to 0.45. In preliminary simulations, these values were found to be best in the sense that they led to performances that most closely matched human performances. At each time step, the state of the agent represented the difference in shape between the comparison and target objects. It was a three-dimensional vector whose elements were set to the values of the shape parameters for the comparison object minus the values of these parameters for the target object. Six possible actions were available to the agent corresponding to the six buttons that a subject could press to modify the action variables. The agent chose an action using an ε-greedy strategy, meaning that the agent chose the action a that maximized $Q(s_t, a)$ with probability $1 - \varepsilon$ (ties were broken at random), and chose a random action with probability ε. The value of ε was initialized to one, and then it was slowly decreased during the course of a simulation. As a result, the agent tended to "explore" a wide range of actions toward the beginning of a simulation, and tended to "exploit" its current estimates of the best action to take toward the middle and end of a simulation. If the agent chose an action that caused the comparison object to have the same shape as the target object, the agent received a reward of 100. Otherwise, it received a reward of -1. The agent performed the training trials of the experiment in the same manner as our human subjects—it performed 7 blocks of training trials with 26 trials per block. To accurately estimate the agent's performances during training, the agent was simulated 1000 times.

The results for experimental conditions using linear causal structures are shown in the left graph of Figure 3 (results for other conditions were similar). The horizontal axis plots the block number, and the vertical axis plots the average difference between the number of steps (i.e., actions or button presses) used by the agent or by human subjects during a trial and the optimal number of steps for that trial as computed by the dynamic programming procedure (as in Figure 2; the error bars in Figure 3 indicate the standard deviations). The solid line shows the data for the simulated agent, and the dotted line shows the data for our human subjects. Interestingly, the learning curves of the simulated agent and of the human subjects have similar shapes, though subjects learned faster than the agent at nearly all stages of training in all experimental conditions. Modifications of the agent by either using different values for the agent's parameters or by adding "eligibility traces" did not significantly alter this basic finding.

Why did subjects show better learning performances than the simulated agent? In the machine learning literature, a distinction is made between model-free versus model-based reinforcement learning agents. The agent described above is an instance of a model-free agent. Although model-free agents are more common in the literature, we hypothesized

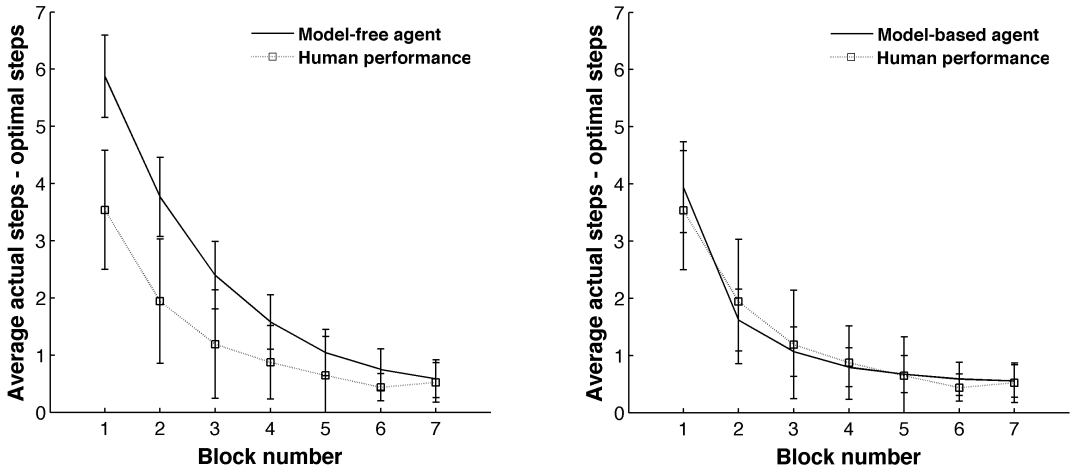


Figure 3: Left: Learning curves for the simulated agent trained via Q-learning (solid line) and for the human subjects (dotted line) in experimental conditions using linear causal structures (error bars plot standard deviations). Right: Identical to the left graph except that the simulated agent learned a model of how actions influenced the environment, and used this model to reason about good actions to take at each time step.

that a model-based reinforcement learning agent may provide a better account of our subjects' performances. Model-based agents typically learn faster than model-free agents, albeit with greater computational expense. Based on real-world experiences, a model-based agent learns an internal model of how its actions influence the environment. The agent updates its Q-values from both real-world experiences with the environment and from simulated experiences with the model (see Sutton and Barto, 1998, for details).

In our simulations, the model was an artificial neural network. Its six input units corresponded to the six possible actions or key presses (an action variable could either increase or decrease in value, and there were three action variables). Its nine output units corresponded to the nine possible influences on the comparison objects' shape parameters (a shape parameter could either increase in value, decrease in value, or maintain the same value, and there were three shape parameters). The network did not contain any hidden units.

When updating its Q-values, the model-based agent used 'prioritized sweeping' (Moore & Atkeson, 1993). This is an efficient method for focusing Q-value updates to state-action pairs associated with large changes in expected reward. Large changes occur, for example, when the current state is a non-goal state and the agent discovers a previously unfamiliar action that leads to a goal state. Large changes also occur when the current state is a non-goal state, and the agent discovers a new action that leads to a new non-goal state known to lie on a path toward a goal state.

In brief, our simulations used prioritized sweeping as follows. At each moment in time, the model-based agent maintained a queue of state-action pairs whose Q-values would change based on either real or simulated experiences. For each update based on a real experience, there were up to N updates based on simulated experiences. The items on the queue were prioritized by the absolute amount that their Q-values would be modified. For example, suppose that at some moment in time, state-action pair (s^*, a^*) had the highest priority. Then $Q(s^*, a^*)$ would be updated. If performing this update on the basis of simulated experience, the agent used the model to predict the resulting new state. In addition, the agent also used the model to examine changes to the Q-values for all state-action pairs predicted to lead to state s^*, known as predecessor state-action pairs. These predecessor state-action pairs were added to the queue, along with their corresponding priorities.

The simulations with the model-based agent were identical to those with the model-free agent. However, the model-based agent used different parameter values. Its discount rate γ was set to 0.3, its learning rate α was set to 0.05, and N, the number of Q-value updates based on simulated experiences for each update based on a real experience, was set to 5. In preliminary simulations, these values were found to be best in the sense that they led to performances that most closely matched human performances.

The combined results for the experimental conditions using linear causal structures are shown in the right graph of Figure 3 (once again, results for the other experimental conditions were similar). The learning curves of the model-based agent are more similar to those of human subjects than the curves of the model-free agent. Indeed, the curves of the model-based agent and of the human subjects are nearly identical. Our findings suggest (but do not prove) that subjects may have achieved near-optimal performances on the perceptual matching task by building internal models of how their actions influenced the external environment. By using these models to reason about possible action sequences, subjects quickly learned to perform the task.

Conclusions

Sequential decision-making tasks are commonplace in our everyday lives. Here, we studied whether people were successful at learning to perform a perceptual matching task, an instance of a sequential decision-making task. We used two benchmarks to evaluate the quality of subjects' learning. One benchmark was based on optimal performance as defined by a dynamic programming procedure. The other was based on an adaptive computational agent that used Q-learning to learn to perform the task. Overall, our analyses suggest that subjects learned to perform the perceptual matching task in a near-optimal manner. When doing so, subjects learned, at least partially, the causal structure underlying the task. In addition, subjects' learning curves were broadly consistent with those of model-based reinforcement-learning agents that built and used internal models of how their actions influenced the external environment.

The cognitive science literature now contains several studies of human performance on sequential decision-making tasks. Some studies have suggested that human performance is optimal, whereas other studies have suggested the opposite. To date, our field does not have a good understanding of the factors influencing whether people will achieve optimal performance on a task. Future research will need to focus on this critical issue. Previous articles in the literature suggested that perceptual aliasing (Stankiewicz et al., 2006) or the existence of actions leading to large rewards in the short-term but not the long-term (Neth, Sims, & Gray, 2006; Gureckis & Love, 2009) seem to be factors leading to sub-optimal performance. Here, we propose a new understanding of when people will (or will not) achieve optimal performance. We hypothesize that people will achieve near-optimal performance on sequential-decision making tasks when they can detect the effects of their actions on the environment, and when they can represent and reason about these effects using an internal mental model.

Acknowledgments

This work was supported by a Grant-in-Aid for Scientific Research (#20730480) from the Japan Society for the Promotion of Science, by a research grant from the Air Force Office of Scientific Research (FA9550-06-1-0492), and by a research grant from the National Science Foundation (DRL-0817250).

References

Bellman, R. (1957). *Dynamic Programming*. Princeton, NJ: Princeton University Press.

Busemeyer, J. R. (2002). Dynamic decision making. In N. J. Smelser & P. B. Baltes (Eds.), *International Encyclopedia of the Social and Behavioral Sciences*. Oxford, UK: Elsevier Press.

Busemeyer, J. R. & Pleskac, T. J. (2009). Theoretical tools for understanding and aiding dynamic decision making. *Journal of Mathematical Psychology, 53*, 126-138.

Chhabra, M. & Jacobs, R. A. (2006). Near-optimal human adaptive control across different noise environments. *The Journal of Neuroscience, 26*, 10883-10887.

Daw, N. D. & Touretzky, D. S. (2002). Long-term reward prediction in TD models of the dopamine system. *Neural Computation*, *14*, 2567-2583.

Fu, W.-T. & Anderson, J. R. (2006). From recurrent choice to skill learning: A reinforcement-learning model. *Journal of Experimental Psychology: General*, *135*, 184-206.

Gibson, F. P., Fichman, M., & Plaut, D. C. (1997). Learning in dynamic decision tasks: Computational model and empirical evidence. *Organizational Behavior and Human Decision Processes*, *71*, 1-35.

Gielis, J. (2003). A generic geometric transformation that unifies a wide range of natural and abstract shapes. *American Journal of Botany*, *90*, 333-338.

Gopnik, A. & Shulz, L. (2007). *Causal Learning: Psychology, Philosophy, and Computation*. New York: Oxford University Press.

Gureckis, T. M. & Love, B. C. (2009). Short-term gains, long-term pains: How cues about state aid learning in dynamic environments. *Cognition*, *113*, 293-313.

Lee, M. D. (2006). A hierarchical Bayesian model of human decision-making on an optimal stopping problem. *Cognitive Science*, *30*, 1-26.

Moore, A. & Atkeson, C. (1993). Prioritized sweeping: Reinforcement learning with less data and less real time. *Machine Learning*, *13*, 103-130.

Neth, H., Sims, C. R., & Gray, W. D. (2006). Melioration dominates maximization: Stable suboptimal performance despite global feedback. *Proceedings of the 28th Annual Meeting of the Cognitive Science Society*.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*, 1593-1598.

Shanks, D. R., Tunney, R. J., & McCarthy, J. D. (2002). A re-examination of melioration and rational choice. *Journal of Behavioral Decision Making*, *15*, 233-250.

Si, J., Barto, A. G., Powell, W. B., & Wunsch, D. (2004). *Handbook of Learning and Approximate Dynamic Programming*. Piscataway, NJ: Wiley-IEEE.

Stankiewicz, B. J., Legge, G. E., Mansfield, J. S., & Schlicht, E. J. (2006). Lost in virtual space: Studies in human and ideal spatial navigation. *Journal of Experimental Psychology: Human Perception and Performance*, *32*, 688-704.

Sutton, R. S. & Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press.

Watkins, C. J. C. H. (1989). Learning From Delayed Rewards. Unpublished doctoral dissertation. Cambridge, UK: Cambridge, University.

Watkins, C. J. C. H. & Dayan, P. (1992). Q-learning. *Machine Learning*, *8*, 279-292.

Yakushijin, R. & Jacobs, R. A. (2010). Are people successful at learning sequential decisions on a perceptual matching task? Manuscript submitted for journal publication.

The Impact of Perceptual Aliasing on Exploration and Learning in a Dynamic Decision Making Task

Lisa Zaval (lz2261@columbia.edu)
Columbia University, Department of Psychology
416 Schermerhorn, 1190 Amsterdam Ave., New York, NY 10027 USA

Todd M. Gureckis (todd.gureckis@nyu.edu)
New York University, Department of Psychology
6 Washington Place, New York, NY 10003 USA

Abstract

Perceptual aliasing arises in situations where multiple, distinct states of the world give rise to the same percept. In this study, we examine how the degree of perceptual aliasing in a task impacts the ability of human agents to learn reward-maximizing decision strategies. Previous work has shown that the presence of perceptual cues that help signal distinct states of the environment can improve the ability of learners to adopt an optimal decision strategy in sequential decision making tasks (Gureckis & Love, 2009). In our experiments, we parametrically manipulated the *degree* of perceptual aliasing afforded by certain perceptual cues in a similar task. Our empirical results and simulations show how the ability of the learner improves as relevant states in the world uniquely map to differentiated percepts. The results provide further support for the model of sequential decision making proposed by Gureckis & Love (2009) and highlight the important role that state representations may have on behavior in dynamic decision making and learning tasks. **Keywords:** perceptual aliasing, dynamic decision making, reinforcement learning

Introduction

A crucial problem facing both human and artificial learners is correctly perceiving and interpreting the current state of the environment. For instance, imagine a traveler staying in an unfamiliar hotel, with each floor and exit decorated identically. Based on perceptual cues alone, this guest may experience difficulty navigating towards his room, since each floor is effectively indistinguishable. In order for navigation to be successful, the traveler must overcome the problem of *perceptual aliasing*, in which relevant "states" or situations in the world map to a single percept (Whitehead & Ballard, 1991; McCallum, 1993). In this example, that current state is the location of the traveler in the building, and the percept is the various cues available that might indicate this location. Note that environments may be aliased along a continuum from the perspective of any individual. For example, suppose that only every other floor in the building is decorated identically. In this case, the guest will be able to differentiate at least half the floors, and his ability to navigate might be somewhat improved. This example can be extended to cases where each floor of the hotel is uniquely decorated, such that salient perceptual cues indicate the traveler's location at any moment. Across these cases, the decision-making ability of the learner is expected to improve as the potential confusion is reduced, and relevant states in the world become mapped to differentiated percepts.

In this paper, we examine how the degree of perceptual aliasing in a task environment impacts the ability of humans to learn effective decision strategies in a dynamic task environment. A growing body of work suggests that human trial-and-error learning shares a similar computational foundation with algorithms developed in the reinforcement learning (RL) literature (see Dayan & Daw, 2008 for a review). However, less work has examined how the identification and categorization of distinct task states might interact with these learning and decision-making processes to determine human performance.

Previous Work

Our work builds upon previous studies of behavior in the "Farming on Mars" task (Gureckis & Love, 2009b, 2009a; Otto, Gureckis, Love, & Markman, 2009). In this task, participants make repeated selections between two "robots" presented on a computer screen. Selection of each robot results in a certain number of "oxygen" points. Participants' goal is to maximize the total amount of oxygen generated over the entire experiment. One robot (the "Short-term" option) always returns more points than the other (the "Long-term" option). However, unknown to participants at the start of the task, the experienced reward structure (i.e., payoff for selecting either robot) continually changes in response to the recent choice history of the participant. In particular, a dynamic is set up so that when the immediately attractive alternative is selected (i.e., the Short-term option), the long-term expected value of both robots is generally lowered on the following trial (Figure 1 illustrates the payout function used in previous Farming on Mars task experiments). Conversely, selections of the immediately worse option (the Long-term option) cause the expected value of both options to increase (in particular, the payoff for each option depends on the number of selections of the Long-term option over the last nine trials). As a result, the optimal reward-harvesting strategy is to learn to choose the option that

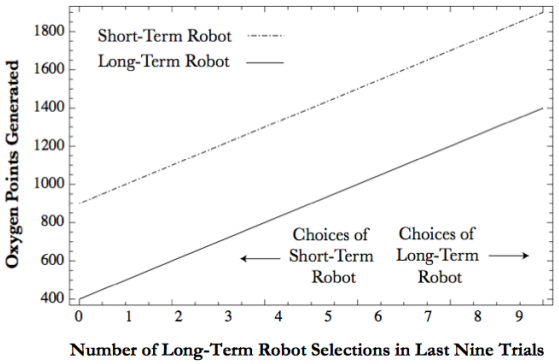


Figure 1: Illustrative payout function of the Farming on Mars Task. The horizontal axis in the figure represents the number of selections out of the last nine in which the Long-Term robot was chosen. The upper diagonal line measures the reward earned from choosing the Short-Term robot as a function of recent choice history, while the lower line illustrates the reward produced from Long-Term selections.

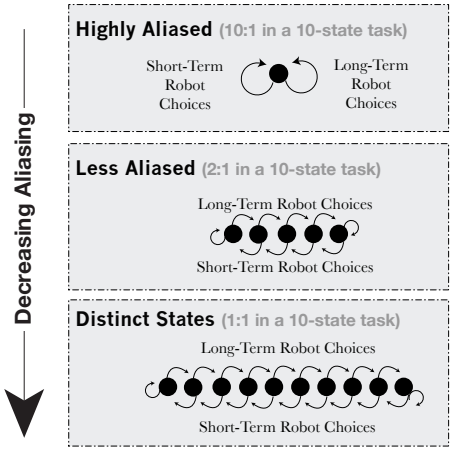


Figure 2: Degrees of perceptual aliasing. At the top is an example of a highly aliased environment where multiple distinct states maps onto a single percept (many-to-one). At the other extreme, distinct perceptual information disambiguates all states (one-to-one). Intermediate levels maps a subset of states to a single percept.

appears worse on each individual trial, since this strategy leads to the greatest cumulative reward.

Critically, performance in the task requires an appropriate balance of *exploration* (in order to discover the hidden contingencies) as well as *exploitation* of choice options known to be rewarding. In addition, a key observation about this task is that there are multiple distinct "states" of the environment (which correspond to the number of Long-term robot selections over the previous trials). When participants fail to recognize this structure, and the fact that the state of the system is changing as a function of their past response history, it becomes difficult to learn the reward-maximizing strategy. Consistent with this, Gureckis & Love (2009a,b) found that providing participants with simple perceptual cues that readily aligned with the state structure of the task improved their ability to learn the reward maximizing strategy. In their experiment, participants' display screen was augmented with a horizontal row of ten indicator lights which served as a cue indicative of the current state of the system. Participants who were given cues that correlated with the underlying task state performed better than participants attempting to learn without these cues. Further, results revealed that cues which supported generalization from one situation to the next had a more beneficial effect on performance relative to cues that effectively limited such generalization (see also Otto, et al., 2009). Gureckis & Love suggested that associating separate perceptual cues with each task "state" could reduce perceptual aliasing and facilitate more effective learning in the same way that appropriate state representations help artificial learning agents based on Q-learning (Sutton & Barto, 1998; Watkins, 1989).

The Present Studies

The present studies were designed to test a key prediction of Gureckis & Love's RL model. As anticipated by our example of the traveler in an unfamiliar hotel, the perceptual aliasing of states in the environment to distinct percepts can vary along a continuum (see Figure 2). At one extreme, every state in the world could map to the same percept (a many-to-one relationship). At the other extreme, each state in the world could map to a distinct percept (a one-to-one relationship). Intermediate cases exist where only a subset of distinct environmental states are perceptually aliased. One possibility is that any time distinct states are poorly differentiated, performance in situations such as the Farming on Mars task should suffer. Alternatively, it is possible that learners may still be able to acquire effective decision strategies when the representation of the task suggested by perceptual cues and the true structure of the task misalign, given that this misalignment takes a particular form. In other words, learners may not need to have a completely accurate representation of the task environment in order to still acquire a near-optimal reward-maximizing strategy. Indeed, this later hypothesis is what is predicted by Gureckis & Love's RL model which can still find optimal policies in some cases given misleading or inaccurate cues about the structure of the task. In the following experiments, we explore how various types of misalignment between perceptual information and task state information influences human learning. In particular, we are interested in how misalignments between perception of the world and the actual structure of contingencies influence learning and exploration behavior. Understanding the nature of this process is important since it is unlikely that human learners have completely accurate informa-

tion about the state structure of the environment at all times.

Experiment 1

In Experiment 1, each subject was randomly assigned to one of four conditions in the Farming on Mars task. Participants in each condition were given different types of perceptual cues which suggested a different interpretation of the nature of the task. Besides the type of cues displayed, each condition was identical with respect to the payoff function and task dynamics. The overall manipulation (providing different types of perceptual cues to learners in the task) parallels the approach in Gureckis & Love (2009).

In one condition (the *no-cue* condition), participants were given no additional cues as part of the display, and thus had to rely on memory and non-perceptual cues in order to uncover the optimal task strategy (c.f., Bogacz, McClure, Li, Cohen, & Montague, 2007). In the second condition (the *two-cue* condition), the interface screen was augmented with a simple cue consisting of two lights. At any point in time, only one of these lights was active, and a shift between the two cues indicated a change in the underlying task system. The position of the activated light was determined by the number of times the Long-term robot was selected over the previous nine trials of the experiment (this condition reflects a many-to-one situation with 5 states mapping to each percept). In the third condition (the *five-cue* condition), a circle of five lights (see Figure 3) was presented on the interface. The indicator lights were organized in a consistent array along the circle, such that the active light moved one position either clockwise or counterclockwise as the task state was updated. The five lights were mapped onto the underlying task system using a "modulus" rule, resulting in two distinct task states mapping to each percept. In the final condition, a display of ten lights was employed, such that each light corresponded exactly to a distinct numerical state in the underlying task system (one-to-one mapping).

Consistent with Gureckis & Love (2009a), we predicted that providing participants with light cue arrays which readily align with the underlying state of the system will limit the aliasing of functionally distinct states, and improve subjects' ability to learn the reward maximizing strategy. Thus, we predict that conditions where perceptual cues limit this aliasing (i.e., the ten-state condition) will result in better overall performance. In addition, we expect that participants' induced representation of the task will strongly influence the strategies they use to balance exploration and exploitation in the task.

Methods

Participants One hundred and ninety-two New York University undergraduates participated for course credit and a small cash bonus based on task performance. A total of 12 participants were dropped from the analysis for responding with the same button on more than 95% of the trials. The remaining participants were randomly assigned to one of four conditions: the *no-cue* condition (N = 44), the *two-cue* condition (N = 45), the *five-cue* condition (N = 45), and the *ten-cue* condition (N = 46).

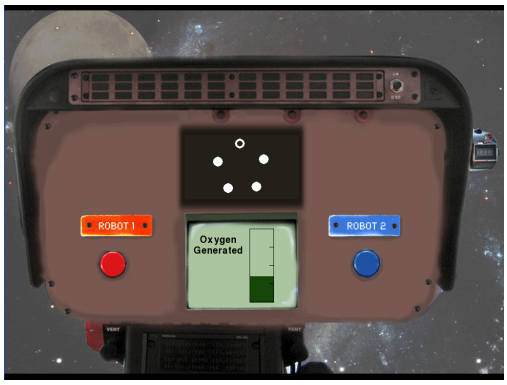


Figure 3: Example of the task interface used in the experiment. The display shows the indicator lights used in the five-cue condition. Additionally, the screen illustrates how rewards were conveyed to participants.

Materials and design The experiment was administered on standard Macintosh computers using an in-house data collection system written in Python[1]. Participants were tested individually over a single one-hour session. Extraneous display variables, such as which robot corresponded to the left or right choice option, the position of the lights, and which direction the active light moved (clockwise or counter-clockwise), were counterbalanced across participants. On each trial, the payoff for selecting the Long-term robot was $40+70*h/9$, where h is the number of times the Long-term robot was selected in the last 9 trials. In contrast, the payoff on each trial for the Short-term robot was $30 + 70 * h/9$. The final values were scaled by 110 and displayed as a percentage on the sliding oxygen meter.

Procedure Participants were tested in the basic Farming on Mars task as described above. At the beginning of the experiment, subjects were presented instructions on the screen which conveyed the basic cover story for the task. The instructions were identical for all conditions, and there was no explicit reference to the function or purpose of the indicator lights/cues. On each trial, participants were shown a display with two large response buttons. Between these buttons was a video display which presented trial-relevant feedback. After a robot selection was made, the quantity of oxygen produced for that trial was presented on the video display. The amount of oxygen points earned was presented visually with a vertical, sliding bar which filled green to

[1] http://www.pypsyexp.org

varying levels. The oxygen level display was shown for 800 ms, after which the the screen was reset to indicate the start of a new trial. No information regarding cumulative oxygen generation was presented, but instructions did emphasize that participants should try to "maximize the number of oxygen points generated over the entire experiment." In the two-light, five-light, and ten-light conditions (but not in the no-cue condition), the screen was augmented with an array of indicator lights as described above and shown in Figure 3. The experiment consisted of 500 separate trials divided into five blocks of 100 trials. In order to maintain motivation, participants were informed that they would receive a small cash bonus of $2-5 dollars based on total oxygen generated by the end of the task.

Results

The primary dependent measure in our experiment was the proportion of Long-term robot selections (i.e., reward-maximizing responses) made by the participant. Total mean proportions by condition are presented in Figure 4. Overall, the proportion of Long-term choices were significantly higher than chance in all conditions, except for the five-cue condition (all $p < .05$). Given the binary outcome choice data, we conducted a series of binomial regressions using the χ^2 distributed deviance-based test as our measure of model selection[2]. There was an overall significant effect of condition $\chi^2(3) = 15.6$, $p = .001$. In addition, the pattern of results across conditions was best predicted as a quadratic function of the number of perceptually distinct task states compared to a linear relationship ($\chi^2(1) = 11.32$, $p < .001$, the quadratic term was reliably above zero, $\beta_{cond^2} = .02$, $p < .001$). Pairwise contrasts (using an Bonferroni-adjusted $\alpha = .05/4 = .012$) between the individual conditions revealed a significantly higher proportion of maximizing responses in the ten-cue condition compared to both the five-cue condition, $\chi^2(1) = 13.46$, $p < .001$, and the two-cue condition, $\chi^2(1) = 11.62$, $p < .001$. Surprisingly, there was a relatively small difference between the ten-cue and no-cue conditions which did not reach significance, $\chi^2(1) = 3.59$, $p = .06$. Note, however, that in a similar task, Gureckis & Love (2009b) and Otto, et al. (2009) found an advantage for one-to-one percept-state representations. Also, note that when given only two state cues, performance was not significantly better than when participants are given five state cues, $\chi^2(1) = 1.04$, $p = .3$.

In order to better understand the genesis of the aliasing effect, we examined the *dynamics* of exploration in the task. In particular, even if the marginal proportion of maximizing choices is constant, it is possible that the distribution of those choices in time could vary. For example, participants in the different conditions might adopt alternative strategies for exploring the task. One way to quantify these differences is to plot the percentage of total trials participants spent in each true (latent) state in the task. Remember that "states" in this dynamic task are defined by the percent allocation of choices to the Long-term option over the last nine trials. Figure 1 plots this distribution for each of the four conditions. Interestingly, the structure of the cues in the task has a strong impact on the way participants explored the task dynamic. In particular, participants in the two-cue condition spent a much larger percentage of time in intermediate states (indicated roughly equal allocation to both choices for extended periods of time). For example, a one-way ANOVA on proportion of time spent in states 3-7 revealed an effect of condition, $F(3, 132) = 4.57$, $p < .005$. Specifically, participants in the two-cue condition spent more total time in these intermediate states than in the no-cue, $t(64) = 2.95, p < .005$, five-cue, $t(66) = 2.31, p < .02$, and ten-cue, $t(66) = 3.43, p = .001$, conditions (since these are post-hoc analyses significance should be interpreted using a conservative $\alpha = .05/3 = .016$). On the other hand, there was also a significant effect of condition on how long participants spend in the end point states (i.e., state 1 & 2 and 9 & 10), $F(3, 132) = 3.25$, $p < .025$. Post-hoc test revealed this was driven primarily by the lower percentage of total time spent in these states in the two-cue condition condition compared to the 10-cue condition, $t(66) = 3.17$, $p < .003$.

Discussion

The results of Experiment 1 show that participant's conceptualization of the state structure of the task can influence both their exploration strategies as well as their ability to identify a reward maximizing strategy. In particular, when cues about the underlying state of the states were more highly aliased (the two-cue and five-cue conditions) participant's overall task performance suffered. Closer examination of the way in which participants explored the task revealed that the alignment of the cues in the task had a dramatic effect on behavior, even when overall performance differences appeared smaller. In particular, relative to the other conditions, participants in the two-cue condition spent a considerably longer time in intermediate states, consistent with a choice strategy involving alternations between the short-term and long-term options.

Experiment 2

In Experiment 1 we found that reward-maximizing performance was worst when a circle of five indicator lights was presented on the interface, such that two different task states mapped to the same perceptual display. However, it is as yet unclear if the performance difference for highly aliased environments results from the num-

[2]We also analyzed these data through a one-way ANOVA and a series of t-tests which revealed an identical pattern of significant results.

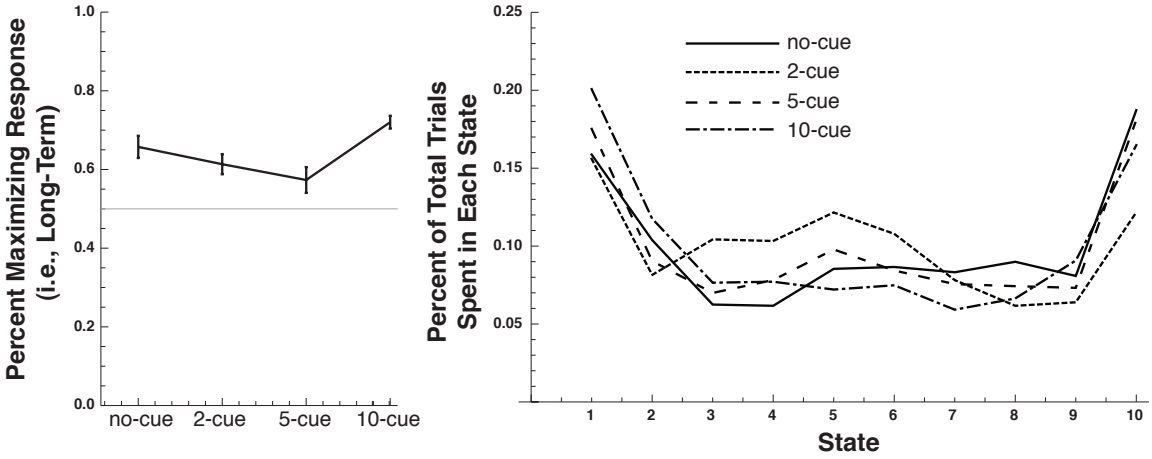


Figure 4: *Panel A*: Average proportion of Long-Term (maximizing) responses made throughout the experiment as a function of condition. The horizontal line at 0.5 shows chance performance. Error bars are standard errors of the mean. *Panel B*: Average percentage of total experiment spent in each state. State 1 corresponds to 0 of the last nine choices being to the Long-term option. State 10 corresponds to 9 of the last 9 choices being to the Long-term option.

ber of implied states (5) or how those states "blend together" by the dynamics of the focal cue (i.e., the active light). For example, in the *five-cue* condition of Experiment 1, the active cue moved one position either to the left or right as the state of the underlying system was updated. Thus, a participant who steadily progressed from states 1-10 would experience the active light looping twice around the circle of indicator lights. An alternative display which maintains the same level of perceptual aliasing (two true states for every one distinct percept) would be to have the active light remain in the same position across two consecutive state updates. In this design, a participant who steadily progressed from states 1-10 would observe the active light making a single loop around the five indicator lights, 'doubling-up' at each individual light position. In other words, if the letter A-E represent the five locations for the state cue, then the mapping from the 10 latent task states to the display would be 1,2,3,4,5,6,7,8,9,10→A,A,B,B,C,C,D,D,E,E. In Experiment 2, we compare task performance in this *single-looped* condition with performance in the *twice-looped* condition (which is identical to the 'five-cue' condition of Experiment 1).

Our prediction was that performance in the twice-looped condition would be lower than in the single-looped condition. The rationale was that participants in the single-looped condition would be better able to recognize that the "gradient" of reward was rising as the light moved in a particular direction. In contrast, the twice-looped condition would be more likely to be confused as a state that they had previously experienced to have low reward (e.g., state cue position A) might later also be associated with high reward. The prediction that the perception of a correlation between the movement of the light and the magnitude of the reward is supported by previous studies showing that participants use such information even when it is against their best interest in the task (Otto et al., 2009).

Methods

Participants Forty New York University undergraduates participated for course credit and a small cash bonus based on task performance. Participants were randomly assigned to either the *twice-looped* condition (N=21) or the *single-looped* condition (N = 19).

Materials and design All aspect of the materials and design were identical to Experiment 1, except for the changes to the five-cue display described above.

Procedure The general procedure was the same as in Experiment 1.

Results

As before, the primary dependent measure in our experiment was the proportion of Long-term robot selections (i.e., reward-maximizing responses) made by the participant. However, there was no overall effect of condition $\chi^2(1) = 0.26$, $p = .61$, M=0.52 in the twice-looped condition and M=0.54 in the single-looped condition. Closer examination of the distribution of overall performance scores indicated that the distribution was strongly bimodal in the twice-looped condition, while it was uni-model in the single-looped condition. As shown in Figure 5, this bi-modality arose from the way that participants explored the latent task states. In particular, a 2-way repeated measures ANOVA on condition

and time spent in each state found a significant effect of state, $F(9, 342) = 4.12$, $p < .001$, and a significant state by condition interaction, $F(9, 342) = 3.17$, $p < .001$. At least a subset of participants in the twice-looped condition appeared to have spent a disproportion amount of time in state 6 which is the point where the display looped back on itself suggesting that they were attempting to keep the state cue from crossing back around to the state associated with the lowest reward. In contrast, participants in the single-looped condition spent more time in the lower states (1-4) indicating that they had an overall bias towards the short-term option that a subset of participants eventually overcame.

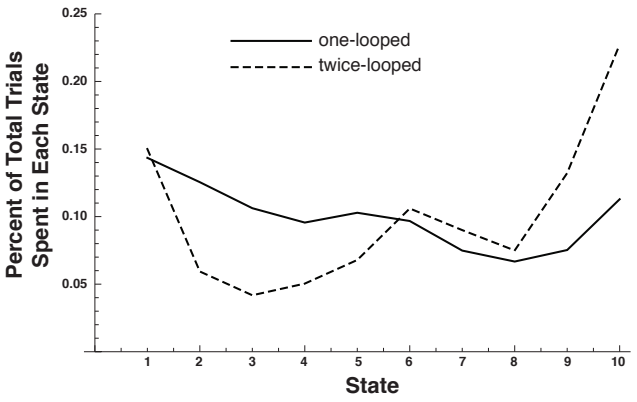


Figure 5: Average percentage of total experiment spent in each state in Experiment 2.

General Discussion

Across a set of two experiments we explored how perceptual cues concerning the underlying state structure of a dynamic decision making task influenced learning. Consistent with previous work (Gureckis & Love, 2009b, 2009a), we find that when task states are aliased, participants' ability to identify an optimal task strategy is impaired. It is important to point out that the effects we see here are unlikely to be a simple consequence of participants ignoring the primary task (to earn oxygen points) and instead exploring aspects of the display. First, participants were clearly instructed that the primary goal was to control the system to earn as many points as possible. In addition, participants were paid a small cash bonus tied to their performance in the task which increased the relevance of the primary task. Finally, our analysis of the dynamics of exploration (i.e., the percent of time spent in each state) reveal systematic differences related to the structure of the cues we provided.

One possibility is that the structure of the perceptual cues provide a kind of strategy "affordance" in the task, limiting the space of exploration/response policies that participants considered. Note that in a separate study, we recently found that motivational manipulations can also impact participant's exploration behavior in a similar task (Otto, Markman, Gureckis, & Love, in review). A theoretical analysis of these results and evaluation of their implication for the Gureckis & Love (2009) model are currently underway. However, preliminary simulations show a close correspondence between the results reported here and the behavior of the model. Future work will continue to evaluate how RL models can be used to understand the motivational and cognitive influences underlying dynamic decision-making.

Acknowledgements We thank Louis Tur and Nathaniel Blanco for programming assistance and discussion in the development of this project.

References

Bogacz, R., McClure, S., Li, J., Cohen, J., & Montague, P. (2007). Short-term memory traces for action bias in human reinforcement learning. *Brain Research*, *1153*, 111-121.

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

Gureckis, T., & Love, B. C. (2009a). Learning in noise: Dynamic decision-making in a variable environment. *Journal of Mathematical Psychology*, *53*, 180-193.

Gureckis, T., & Love, B. C. (2009b). Short term gains, long term pains: How cues about state aid learning in dynamic environments. *Cognition*, *113*(3), 293-313.

McCallum, R. (1993). Overcoming incomplete perception with utile distinction memory. In *The proceedings of the tenth international machine learning conference (ml'93)*. Amherst, MA.

Otto, A., Gureckis, T., Love, B., & Markman, A. (2009). Navigating through abstract decision spaces: Evaluating the role of state knowledge in a dynamic decision making task. *Psychonomic Bulletin and Review*, *16*(5), 957-963.

Otto, A., Markman, A., Gureckis, T., & Love, B. (in review). Regulatory fit in a dyanmic decision-making environment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Sutton, R., & Barto, A. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.

Watkins, C. (1989). *Learning from delayed rewards*. Unpublished doctoral dissertation, Cambridge University, Cambridge, England.

Whitehead, S., & Ballard, D. (1991). Learning to perceive and act by trial and error. *Machine Learning*, *7*(1), 45-83.

Human foraging behavior: A virtual reality investigation on area restricted search in humans

Christopher Kalff (kalff@cognition.uni-freiburg.de)*
Center for Cognitive Science, University of Freiburg, Friedrichstr. 50,
79098 Freiburg, Germany.

Thomas Hills (thomhills@gmail.com)*
Department of Psychology, University of Basel, Missionsstr. 60/62,
4055 Basel, Switzerland.

Jan M. Wiener (jwiener@bournemouth.ac.uk)*
Department of Psychology, Bournemouth University,
Talbot Campus, Poole, Dorset, BH12 5BB, United Kingdom.

* The authors contributed equally to this work

Abstract

The control of attention and the control of movement in space share a similar optimal control structure—mediating the trade-off between exploiting one locale and exploring others. A common spatial foraging strategy observed in many species is area-restricted search, in which animals respond to resources or their absence by moving between local and global search strategies, respectively. When resources are clustered, area-restricted search can represent an optimal foraging strategy. Surprisingly few studies have investigated whether humans display such behavior in the context of spatial navigation. Here we present two experiments in which human participants search for resources distributed over a large virtual environment. By systematically manipulating the specific distribution of the resources the first experiment investigates human's ability to perform area-restricted search. The second experiment probes for the patch-leaving rules humans apply when facing resources distributed in patches that differ in quality. Our results indicate that humans forage in space using an area-restricted search, but do so in a non-optimal way—consistent with other studies showing non-optimal search strategies in memory.

Keywords: Foraging; area-restricted search; navigation.

Introduction

Picking bananas from banana trees, searching for nebulas in the night sky, and hunting for schools of tuna in the open ocean all involve the ability to detect and respond to spatial resource distributions. Since the foundations of animal foraging behavior were laid by MacArthur and Pianka (1966) and Emlen (1966) decades of research have shown that non-human animals respond adaptively to these spatial resource distributions; moreover, their responses are often optimal with respect to long-term rate maximizing models (reviewed in Stephens & Krebs, 1987). For humans, these models have been shown to predict patterns of search in information foraging on the internet (Pirolli & Card, 1999), the foraging strategies of hunter-gatherer societies (e.g., Hawkes, Hill, & O'Connell, 1982), and the search patterns of humans in their own memory (Hills, Todd, & Jones, 2009; Hills, Todd, & Goldstone, 2008). However, surprisingly, almost nothing is known about how humans search in 3-dimensional environments like those described for the bananas, nebulas, and tuna (but cf. Smith, 1983, for an overview of anthropological research).

How *do* humans forage in space? Are they capable of detecting and localizing resources in space, with or without the help of visual cues? Moreover, are their foraging strategies adaptive, or near optimal in terms of rate maximization? In this article, we use 3-dimensional virtual representations of fields and orchards to investigate how people forage in open environments, and in particular, whether or not they show patterns consistent with area-restricted search.

Area-restricted search (ARS) is one of the most well-studied behavioral patterns in animal foraging, and has been observed in a wide variety of animals (e.g., Hills, Brockie, & Maricq, 2004; Krebs, 1973; Smith, 1974). It can also produce patterns of movement that look like Levy walks—another commonly observed foraging pattern (Benhamou, 2007). ARS involves high turning angles following resource encounters but lower turning angles elsewhere. It indicates an adaptive response to spatial distributions in clustered (or patched) environments because in clustered environments - when prior knowledge about resource locations is limited to the time since they were last encountered - ARS is optimal (Walsh, 1996; Grunbaum, 1999). ARS, like an annealing strategy, localizes animals where resources are most dense (Karieva & Odell, 1987). The success of this strategy and its minimal information requirement are consistent with the evidence that ARS had an early evolutionary origin amongst mobile animals. Moreover, the evolution of this strategy may have provided the biological building blocks for the subsequent evolution of human attention (proposed in Hills, 2006).

If humans respond to clustered resources with increased turning, but don't do so when resources are uniformly or dispersedly distributed, they are showing foraging patterns consistent with ARS. However, evidence for ARS in human

spatial foraging requires more than simply noting that humans respond to clustered resources with more turning. A number of potentially viable foraging heuristics are consistent with ARS at a gross level, but fail to meet its more strict definition of *turning mediated by decaying memories of resources*. These alternate hypotheses include the fixed-number rule and the fixed-time rule (see Stephens & Krebs, 1987).

For the *fixed-number rule*, the forager collects roughly the same amount of items in every patch regardless of the time to achieve this goal: $n_1 \approx n_2 \approx n_i$. If participants used a fixed-number-rule, they would yield insignificant differences in gathered items across patches. Additionally, re-visited (and therefore emptier) patches should receive significantly more time than during first encounter.

The *fixed-time rule* states that a foraging organism will devote roughly the same time to all patches it visits: $t_1 \approx t_2 \approx t_i$. If humans used a fixed-time-rule, there should be no significant differences in patch visit times, regardless of patch quality. Additionally, re-visited (and therefore emptier) patches should receive the same attention than during first encounter

Like the fixed-time rule, ARS uses temporal cues to determine patch departures. However, ARS adds time to the total patch residence time by incrementing the time in the patch (by turning) following each resource encounter. If a certain temporal threshold without resources is exceeded, the patch is abandoned. As Iwasa, Higashi and Yamamura (1981) mention, this heuristic—sometimes called the *incremental rule* or *Green's assessment rule* (Green, 1984) — is highly appropriate among variable patch sizes. Several studies have indicated that humans use this kind of incremental strategy when foraging in a lexical problem space (Payne, Duggan & Neth, 2007; Wilke, Todd, & Hutchinson, 2009).

In the present study we investigate human spatial foraging in a 3-dimensional environment by first asking if participants show behavior consistent with area-restricted search in clustered resource distributions (versus uniform distributions). Second, we ask if humans can detect the difference between high and low quality patches, and if so, do they respond using one of the foraging heuristics described above. That is, are their foraging patterns most consistent with an incremental rule, or are they more likely to be fixed-time or fixed-number rules?

Experiment I

Experiment 1 investigated whether human foragers are sensitive to the distribution of resources in the environment displaying a foraging pattern consistent with area-restricted search (ARS). In the experiment, participants were placed in large virtual environments that contained resource items. These were either uniformly distributed about the entire space or organized in patches. Participants could not see the items prior to encountering them; there were no visual cues to help them harvest resources. Participants had to actively navigate through the environment, searching for resource items.

Method

Environments A circular virtual environment with a radius of 110m was constructed. The environment consisted of a textured ground plane resembling a large meadow and was surrounded by a fence. Three large landmarks (mountain, city skyline, and skyscraper) surrounded the environment providing global direction cues (see Figure 1).

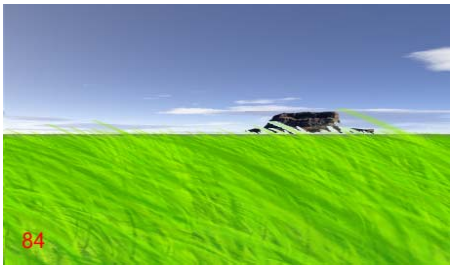


Fig. 1: Participants' perspective during the task. One of the global landmarks (mountain) is visible in the distance.

1440 individual resource items (mushrooms, modeled as 3d objects) were then either evenly distributed about the environment (*dispersed condition*) or they were arranged in 24 patches that were randomly scattered about the environment (*patched condition*; see Figure 2). Each patch had a radius of 8.65 m and contained 60 resource items. The minimal distance between any two resource items in the patched environments was 1.53m, in the distributed world it was 2.35m. For each type of resource distribution (dispersed or patched) five different environments that differed in the specific arrangement of the resource items were created. The resource items in both conditions were visible only from close proximity – i.e. from a distance smaller than 1.25m – similar to real mushrooms in long grass.

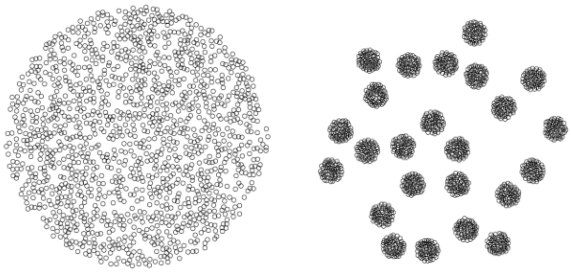


Fig. 2: The two types of resource distributions: left: one of five 'dispersed' environments; right: one out of five 'patched' environments.

Participants navigated through the environments in the first person perspective using the arrow keys of the

keyboard. Translation speed was set to 4m/s and turning velocity was set to 75°/s. The motion model allowed for either forward translations or rotations but did not allow combining translations and rotations. Thus, the resulting trajectories resembled segmented paths (see Figure 4). Participants collected resource items simply by moving closer than .75m to an item. This pick-up distance was just below half of the minimal distance between any two resource items and therefore assured that participants gathered only one item at any time. The collection of a resource item was signaled by an auditory cue. Once an item had been collected it was removed from the environment.

Participants Thirty-two participants (17 women) aged 19 to 28 (M = 22.28, SD = 2.41) took part in the experiment. They were mainly students from Freiburg University and received course credits or monetary compensation for their participation.

Procedure Participants were randomly assigned to either the dispersed or the patched condition (counterbalanced for gender) and were then briefed about the experiment: Their task was to navigate through the environment and to collect resource items. Each participant was given 5 trials. Each trial was carried out in a different environment with the same type of resource distribution (dispersed or patched). At the beginning of each trial, participants were placed in the center of the environment. A single trial was terminated either after 600 seconds or when participants collected 90 resource items. The experiment ended after participants completed all 5 trials. Participants were offered a fixed compensation, independent of the time required to do the experiment. Thus, they were motivated to finish as quickly as possible and the usual (biological) energy cost variable was transformed into a temporal equivalent.

Results

Search time A two-way mixed ANOVA (factors: trial, condition; sphericity assumed: $\chi^2(9) = 14.015$, $p = .122$) reveals a main effect for trials: $F(4, 120) = 4.703$, $p < .01$, partial-$\eta^2 = .136$ which is due to significant differences between trials one and three, and one and five (both Sidak-corrected p's < .05). Even though completion time was higher for the patched versions (M = 442.63, SE = 12.94 vs. dispersed: M = 422.95, SE = 12.94) there is no main effect of condition ($F(1, 30) = 1.158$, $p = .291$, partial- $\eta^2 = .037$), as well as no significant interaction between trials and condition: $F(4, 120) < 1$ (see Figure 3).

Search time results did not demonstrate a significant difference between experimental conditions (patched vs. dispersed condition). The reduction in search time over trials, however, indicates an adaptation of search strategy (see Figure 4) leading to a higher rate of item encounter.

An alternative explanation is that participants learned to control their movements more effectively as the experiment progressed.

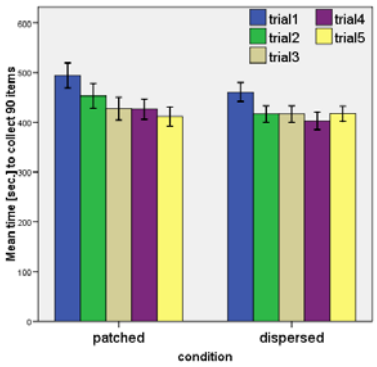


Fig. 3: Mean search time for each trial in the two conditions. Error bars depict one SE.

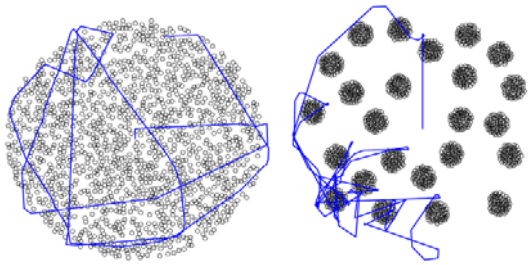


Fig. 4: Example trajectories in the dispersed (left) and the patched condition (right).

Turning rate As argued above, an increase in overall turning rate in environments with clustered resources as compared to environments with evenly distributed resources indicates an adaptive response to spatial distributions. This would be perfectly consistent with area-restricted search. Figure 5 shows the average total turning angles per second.

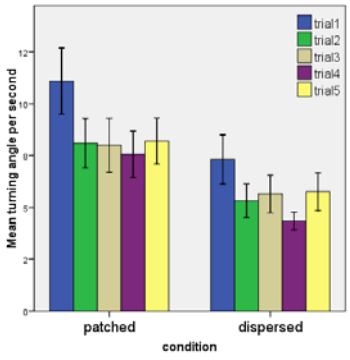


Fig. 5: Average turning angle per second for the two conditions. Error bars depict one SE.

A two-way mixed ANOVA (factors: trial, condition; due to violation of sphericity [$\chi^2(9) = 52.293$, $p < .001$, $\varepsilon = .589$] the Huynh-Feldt correction for degrees of freedom was used) demonstrates both, a main effect of trials ($F(2.356, 70.675) = 6.353$, $p < .01$, partial- $\eta^2 = .175$) as well as a main effect of condition ($F(1, 30) = 5.143$, $p < .05$, partial-$\eta^2 = .146$). Specifically, total turning angle per second in the

patched condition was higher than in the dispersed condition (patched: M = 8.59°, SE = .91; dispersed: M = 5.67°, SE = .91), demonstrating an adaptive response to the specific distribution of resources. The interaction of trial and condition did not yield a significant effect: $F(4, 120) < 1$.

Trajectories and turn rate after item encounter Visual inspection of the trajectories corroborates the latter analysis that demonstrates that participants search behavior differed in the patched and the dispersed condition (cf. Figure 4). These findings, however, do not necessarily demonstrate area restricted search, which specifically involves an increase in turning angle after resource encounter.

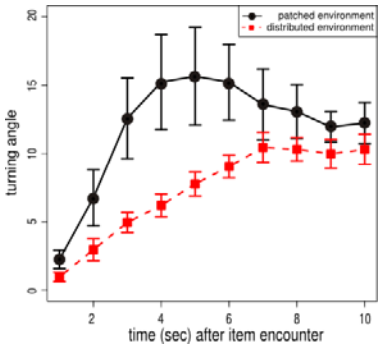


Fig. 6: Turning angle as a function of time after item encounter. Error bars represent one SE.

Figure 6 shows the effect on turning immediately after encountering a resource item. Note that in order to capture items, participants must be moving forward in a straight line. Therefore, turning angle at time of capture will always be zero. Participants in the patched condition are clearly turning more sharply following a resource encounter, as demonstrated by a mixed-model ANOVA (within-participant factor: time after item encounter; between-participant factor: condition). Due to violation of the sphericity assumption ($\chi^2(44) = 270.664$, $p < .001$, $\varepsilon = .327$) the Huynh-Feldt correction for degrees of freedom was used. The interaction of time and condition is significant ($F(2.943, 88.283) = 3.616$, $p < .05$, partial-$\eta^2 = .108$), because turning angles at time bins '3', '4', and '5' differed significantly (all Sidak-corrected p's < .05). This also resulted in a main effect of condition: $F(1, 30) = 4.403$, $p < .05$, partial-$\eta^2 = .128$.

Discussion

How *do* humans forage in space? Do they detect resources in the environment and adapt their search behavior when facing different distributions of resources? Experiment 1 demonstrated that participants increased their turning rate and turned more sharply after resource encounters in environments in which resources were patched. These results are consistent with area-restricted search and suggest that human foragers adapted their search strategy according to the specific distribution of resources in the environment. People do search differently when faced with different spatial distributions of resources; moreover, they show more evidence of area-restricted search in environments where such a search strategy is optimal.

Experiment II

Results from Experiment 1 suggest that human foragers, when searching for resources in a spatial context, are sensitive to the distribution of the resources. However, in Experiment 1 all resource patches featured the same amount of items, i.e. the quality of all patches was identical. Experiment 2 was designed to investigate how human foragers interact with resource patches that differ in quality, but that are visually identified by the presence of a tree. Can they tell a good from a bad patch? And, given that foragers have a priori knowledge about the distance to the next patch (by the distribution of trees), how do they determine when to leave a patch in order to harvest at another patch? Also, this experiment allowed to more directly test for area-restricted search in comparison with the other patch leaving rules outlined in the introduction.

Method

Environment The same circular virtual environment as in Experiment 1 was used. In addition, 19 trees, arranged on a hexagonal grid (see Figure 7), were planted in the virtual environment. Resource items were distributed under the trees in patches with a radius of 8 meters. The hexagonal arrangement of the trees (patches) ensured that for each patch the distances to all neighboring patches were identical. Each patch featured either 15 (*poor patches*) or 30 (*rich patches*) resource items.

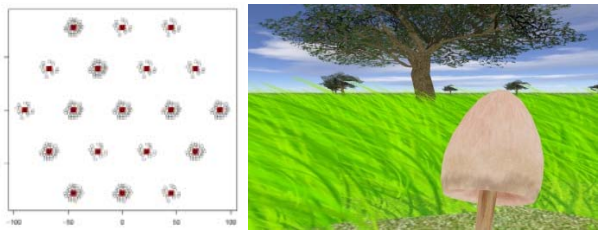


Fig. 7: Left: overview of the environment and the 19 patches (9 poor and 10 rich); right: participants' view while navigating.

As in Experiment 1, participants navigated through the environment (see Figure 7) using the arrow keys of the keyboard. They could not see resource items from the distance, but only in close proximity (viewing distance: 1m). In order to increase the costs associated with moving from one patch to another, translation speed was reduced to 2m/s. Thus, moving from one tree to a neighboring tree took 25 seconds.

Participants Thirty-two participants (16 women) aged 14 to 30 (M = 23.06, SD = 3.37) took part in the experiment. They were mainly students from Freiburg University and received course credits or monetary compensation for their

participation. None of the participants took part in Experiment 1.

Procedure Participants were first briefed about the experiment: Their task was to navigate through the environment and to collect a total of 125 resource items. Participants were also told that resource items were to be found in the vicinity of the trees: The instruction mentioned a certain type of mushroom that only grows under and in the close vicinity of trees, but never further away. Participants were unaware that the patch quality differed between patches. At the beginning of the experiment, participants were placed in the center of the environment. The experiment was terminated after participants collected the last of the 125 resource items required. As in Experiment 1, participants were motivated by being assured that they would receive a fixed compensation for their participation, independent of the time required to solve the task. For the purposes of patch leaving rule analyses, the first and the last patch participants visited were discarded from the analyses.

Results

Patches visited On average, participants visited 18.84 (SD = 10.02) patches (including re-visits). The minimum number of patches visited was seven the maximum 45 visits (see Figure 8).

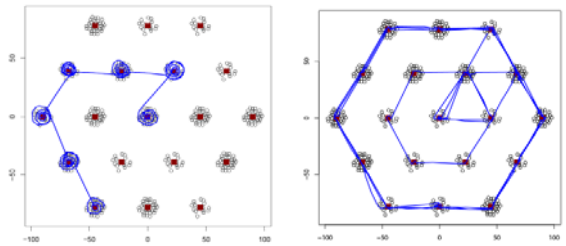


Fig. 8: Trajectories in the actual environment for the minimum (left) and maximum (right) amount of visited patches.

Time On average, participants needed 1804.87 (SD = 347.81) seconds to complete the experiment. Participants considerably differed with respect to the time needed to complete the experiment (range = 1161). Given the time required to move between patches (25 seconds), time to complete the experiment strongly correlated with the number of patches visited ($r(32) = .55$, $p < .001$).

Time spent in patches As stated above, the fixed-time rule would result in foragers devoting the same amount of time to every patch regardless of its quality. *Time in patch* is calculated as the time difference between the first and the last item encounter within each patch. In order to control for quality of the patch at time of encounter, only the first visit of each patch entered this analysis; revisits were discarded. Participants spent significantly more time – roughly twice as much – in richer patches (M = 112.25 sec, SE = 3.15) than in smaller patches (M = 64.57 sec, SE = 3.16; random-factor ANOVA: $F(1, 34.319) = 63.663$, $p < .001$, partial- $\eta^2 = .65$). This indicates that participants were not using a fixed-time rule when foraging in this environment.

Amount of collected items The fixed-number rule predicts that foragers collect an equal amount of items in every patch, regardless of the time it would take to succeed. Again, in order to control for the patch quality at the time of encounter, only the first visit of each patch entered this analysis; revisits were discarded. Participants collected more than twice as many items in rich patches (M = 13.71, SE = .27) as compared to poorer patches (M = 6.06, SE = .27): random-factor ANOVA: $F(1, 32.405) = 95.685$, $p < .001$, partial-$\eta^2 = .747$. This poses strong evidence that participants were not using a fixed-number rule.

Giving-up-densities Giving up density was lower for rich patches (M = 54.3%, SE = .012) than for poor patches (M = 59.6%, SE = .012; $F(1, 34.552) = 5.99$, $p < .05$, partial-$\eta^2 = .148$). This may indicate that participants are leaving the different patches at different inter-item retrieval times, i.e. they are more patient in rich than in poor patches.

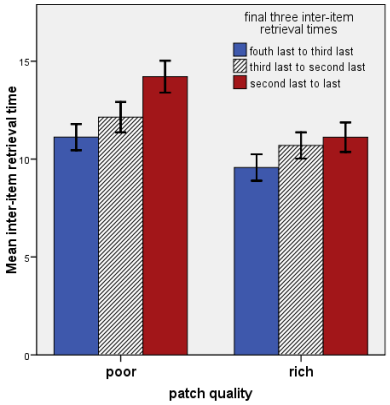


Fig. 9: Mean inter-item retrieval times for the last four item encounters over patch quality. Error bars represent one SE.

To test this, we subjected the time between encounters for the last three items within a patch to a random-factor ANOVA. Results (see Figure 9) show a significant difference between the two patch qualities[1]: $F(1, 34.28) = 6.631$, $p < .05$, partial- $\eta^2 = .162$. This is due to the last three inter-item retrieval times being higher in the poorer (M = 12.94 sec, SE = .45) than in the richer patches (M = 11.09 sec., SE = .39).

Discussion

In summary, Experiment 2 demonstrates that participants were not using fixed-time or fixed number rules, but were

[1] There is also a main effect for the development of the last three inter-item-retrieval times ($F(2, 71.085) = 3.277$, $p < .05$, partial- $\eta^2 = .144$) which is due to an increase in time towards the last time difference. There is no interaction of the two measures ($F < 1$).

instead using a strategy similar to area-restricted search (i.e., an incremental rule) - staying longer in richer patches and shorter in poorer ones. Given the nature of the patch types (one rich and one poor), this strategy is optimal. However, unlike the optimal foraging strategy predicted by the marginal value theorem (Charnov, 1976), participants do not appear to be leaving patches at equal rates of resource capture.

General Discussion

Our results provide evidence that people are using an evolutionarily old foraging strategy—area-restricted search—when foraging in patchily distributed spatial environments. The same strategy has been observed in a variety of 'internal' foraging tasks (e.g., Payne et al., 2007; Hutchinson et al., 2008). Moreover, the same neuromolecular processes facilitate area-restricted search across species as facilitate the control of human attention, suggesting a possible evolutionary origin for human attention (reviewed in Hills, 2006). This is a fascinating possibility because fluid intelligence, working memory, executive control processes, and spatial foraging may all be largely about appropriately mediating a similar kind of trade-off between exploitation and exploration of goal structures and associative relations (e.g., Kane & Engle, 2002). Optimal control of focus is a problem common to many tasks, both internal and external.

Interestingly, while our participants show evidence of utilizing ARS, they do so non-optimally—using different departure rules for different quality patches. This too has been observed in memory search (Young, 2004), and suggests that foraging tasks may provide an important paradigm for understanding the control of attention and the influence of environmental structure on that control.

Acknowledgments

We thank Gavan Wilhite and Inka Hähnlein for their help building the virtual environment, recruiting participants and collecting the data. This work was partially funded by the Volkswagen-Stiftung, and the SFB/TR8 of the German Research Foundation (DFG).

References

Benhamou, S. (2007). How many animals really do the Levy walk? *Ecology*, 88, 1962-1969.

Charnov, E. L. (1976). Optimal foraging, the marginal value theorem. *Theoretical Population Biology*, 9, 129-136.

Emlen, J. M. (1966). The role of time and energy in food preference. *American Naturalist*, 100, 611-617.

Goldstone, R. L., & Ashpole, B. C. (2004). Human foraging behavior in a virtual environment. *Psychonomic Bulletin & Review*, 11(3), 508-514.

Green, R. F. (1984). Stopping rules for optimal foragers. *American Naturalist*, 123, 30-43.

Grunbaum, D. (1999). Advection-diffusion equations for generalized tactic searching behaviors. *Journal of Mathematical Biology*, 38, 164-194.

Hawkes, K., Hill, K., & O'Connell, J. (1982). Why hunters gather: optimal foraging and the Ache of eastern Paraguay. *American Ethnologist*, 9, 379-398.

Hills, T. (2006). Animal foraging and the evolution of goal-directed cognition. *Cognition*, 30, 3-41.

Hills, T., Brockie, P., & Maricq, A.V. (2004). Dopamine and glutamate control area-restricted search behavior in Caenorhabditis elegans. *Journal of Neuroscience*, 24, 1217-1225.

Hills, T., Todd, P. M., & Jones, M. (2009). Optimal foraging in semantic memory. *Proceedings of the 31st Annual Conference of the Cognitive Science Society*.

Hills, T., Todd, P. M., & Goldstone, R. L. (2008). Search in external and internal spaces. Evidence for generalized cognitive search processes. *Psychological Science*, 19(8), 676-683.

Hutchinson, J. M. C., Wilke, A., & Todd, P. M. (2008). Patch leaving in humans: can a generalist adapt its rules to dispersal of items across patches. *Animal Behaviour*, 75(4), 1331-1349.

Iwasa, Y., Higashi, M., & Yamamura, N. (1981). Prey distribution as a factor determining the choice of optimal foraging strategy. *American Naturalist*, 117, 710-723.

Kane, M. J., & Engle, R. W. (2002). The role of the prefrontal cortes in working-memory capacity, executive attention, and general fluid intelligence: An individual-differences perspective. *Psychonomic Bulletin & Review*, 9, 637-671.

Karieva, P., & Odell, G. (1987). Swarms of predators exhibit 'prey taxis' if individuals use area-restricted search. *American Naturalist*, 130, 233-270.

Krebs, J. R. (1973). Behavioral aspects of predation. In: Bateson, P. P. G., & Klopfer, P. H. (eds.), *Perspectives in ethology*, vol. 1, New York: Plenum Press, 73-111.

MacArthur, R. H., & Pianka, E. (1966). On optimal use of a patchy environment. *American Naturalist*, 100, 603-609.

Payne, S. J, Duggan, G. B., & Neth, H. (2007). Discretionary task interleaving: heuristics for time allocation in cognitive foraging. *Journal of Experimental Psychology*, 136(3), 370-388.

Pirolli, P., & Card, S. (1999). Information foraging. *Psychological Review*, 106, 643-675.

Smith, J. N. M. (1974). The food searching behaviour of two European thrushes. II. The adaptiveness of the search patterns. *Behaviour*, 49, 1–61.

Smith, E. A. (1983). Anthropological applications of optimal foraging theory: A critical review. *Current Anthropology*, 24(5), 625-651.

Stephens, D. W., & Krebs, J. R. (1986). *Foraging theory*. Princeton, NJ: Princeton University Press.

Walsh, P. D. (1996). Area-restricted search and the scale dependence of patch quality discrimination. *Journal of Theoretical Biology*, 183, 351-361.

Young, C. J. (2004). Contributions of metaknowledge to retrieval of natural categories in semantic memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30, 909-916.

Learning in Multiple-cue Judgment Tasks

Bettina von Helversen (Bettina.vonhelversen@unibas.ch)
University of Basel, Department of Psychology, Missionsstr, 62a
4057, Basel, Switzerland

Jörg Rieskamp (joerg.rieskamp@unibas.ch)
University of Basel, Department of Psychology, Missionsstr, 62a
4057, Basel, Switzerland

Abstract

In our daily lives we often make quantitative judgments based on multiple pieces of information such as evaluating a student's paper based on form and content. Psychological research suggests that humans rely on several strategies to make multiple-cue judgments. The strategy that is used depends on the structure of the task. In contrast, recent research on learning in judgment tasks suggests that learning is relatively independent of task structure. In a simulation study we investigated how the performance of several learning models is influenced by the structure of the task and the amount of learning experience. We found that a linear additive neuronal network model performed well regardless of task structure and amount of learning. However, with little learning a heuristic model performed similarly well, and with extensive learning, associative learning models caught up with the linear additive model.

Keywords: Learning; multiple-cue judgments; Computational modeling

Multiple-cue Judgments

When judging objects on a continuous criterion such as the quality of a research paper, people often rely on multiple sources of information. For example, the clarity of the writing, the novelty of the research and the methodological precision may be used as important aspects for evaluating a paper. Several models have been developed to describe how humans solve these judgment problems. Traditionally, linear additive models have been employed to capture how humans weigh and integrate information. Social Judgment Theory (SJT; see Doherty and Kurz, 1996; Cooksey, 1996) relied on multiple-linear regression models to capture decision policies and researchers have used this approach successfully to describe judgments in many areas (see Brehmer, 1988). Similarly, Anderson (1981) suggested that humans combine information in a linear additive fashion. However, recently it has been suggested that humans may have multiple cognitive strategies available to make multiple-cue judgments. Juslin, Karlsson, and Olsson (2008) suggested that depending on the structure of the tasks, humans may switch between a rule-based cue abstraction approach and a similarity–based exemplar approach. Similarly, von Helversen and Rieskamp (2008, 2009) suggested the mapping model, a heuristic model for multiple-cue judgments, and showed that the model that was best in describing participants' behavior depended on the task structure. More specifically, they showed that the mapping model described participants' responses well in tasks that could not be solved by a linear model and where participants had knowledge about the cues' polarity; that is, the sign of the correlation between a cue and the criterion. The exemplar model performed well, in non-linear environments with no prior knowledge about cue polarity, and a linear additive model performed well if the task structure was linear.

Learning in Multiple-cue Judgment Tasks

Although many studies in multiple-cue judgment research rely on extensive learning phases, there have been relatively few attempts to understand and model the learning process. However, the learning process is crucial to understand how people come to make judgments and which cognitive processes they rely on. Particularly, if — as suggested — people rely in their judgment on multiple cognitive processes, this should also be reflected in the learning phase. Additionally, the learning phase itself could play an important role in determining how later judgments are made. Recently, Kelley and Busemeyer (2008) compared how well several models could describe the learning process in various multiple-cue judgment tasks. They compared a rule-based neuronal network model with a delta-learning rule (e.g. Gluck & Bower, 1988), which can be seen as a learning version of a linear additive model with an associative connectionist network model (ALM, Busemeyer, Byun, DeLosh, & McDaniel, 1997; Busemeyer, Myung, & McDaniel, 1993). They found that the rule-based neuronal network models described the learning process best in the majority of the tasks, suggesting that learning may be relatively independent of task structure.

These results are somewhat contrary to the research by Juslin et al. (2008) and von Helversen and Rieskamp (2009) on multiple-cue judgments, suggesting that humans rely on a variety of strategies, depending on the structure of the task (e.g. Juslin, et al., 2008; Rieskamp & Otto, 2006). This raises the question of whether learning depends on the task structure and what may be the mechanisms that lead to a switch in cognitive processing during learning. In this paper we investigate two reasons that may cause a shift in cognitive processing during learning in a multiple-cue judgment task. One reason to rely on different learning strategies may be that their learning performance differs depending on the structure of the task. Thus, we will

investigate if task structure influences how well various learning procedures perform that are imbedded in different cognitive models of multiple-cue judgments (e.g. Juslin et al., 2008; Kelley & Busemeyer, 2008; von Helversen & Rieskamp, 2008). Second, the reliance on different learning procedures could also be due to differences in how fast the procedures adapt to different judgment structures. Therefore, we additionally examined if the models differ with respect to their learning speed.

Learning Models

We tested learning versions of cognitive models suggested in the literature for multiple-cue judgments. As a learning model for the linear additive model we relied on a rule-based neuronal network model as implemented by Kelley and Busemeyer (2008). As an exemplar model we extended the ALCOVE model (Kruschke, 1992) to continuous judgments. ALCOVE has been successfully used to model exemplar-based learning in categorization. We also tested a version of the mapping model (von Helversen & Rieskamp, 2008) to allow for learning. Additionally, we included the ALM model as implemented by Kelley and Busemeyer (2008).

Linear Additive Model Much research has shown that linear additive models are good at describing human judgments (Brehmer, 1994). The linear additive model assumes that people weigh each piece of information according to its importance and then add the weighed evidence to reach a judgment. Traditionally, a multiple linear regression is used to capture how much weight people put on each piece of information (i.e. cue). Kelley and Busemeyer (2008) used a rule-based neuronal network with a linear additive structure:

$$g_t = a_1 \cdot c_1 + a_2 \cdot c_2 + ... + a_k \cdot c_k, \qquad (1)$$

where the model prediction g at time t is given by the sum of the cue values c for k cues weighted by their importance a at time t. This learning model updates the weight for each cue according to a delta rule (Gluck & Bower, 1988) with a learning parameter δ capturing the learning rate. An additional decay parameter ω controls the impact of new information.

$$a_{k,t} = a_{k,t-1} + (\delta / t^{\omega})(Y_{t-1} - g_{t-1})c_{k,t-1}, \qquad (2)$$

with Y indicating the feedback (i.e. the criterion value) and g the model prediction at time t-1.

Mapping model We extended the mapping model (von Helversen & Rieskamp, 2008) to allow for learning. The mapping model follows a simple cognitive strategy that makes judgments by first categorizing an object and then retrieving a typical estimate for the category it was put in. According to the mapping model, an object is placed into a category based on the sum of (standardized) cue values, implying that all cues are weighted equally. The judgment is then determined by the median of the criterion values of all objects in the respective cue sum category. The learning procedure we suggest describes how and how many cue sum categories are formed during learning. In the beginning it is assumed that only a single category is used. In each learning trial, the decision is then made as to whether the new object is put into a new category or into an existing category. A new category is formed if the difference between the cue sum of a new object and the cue sum of each existing category is larger than a distance parameter d. The criterion value estimated for each category is the mean of the criterion values of the objects falling into this category and is updated whenever a new object falls within a category.

ALM The ALM model is an associative connectionist network model. It assumes a layer of input nodes representing each combination of cue values (2^Number of cues, with binary cue data). The input nodes are connected to a layer of r output nodes reflecting the criterion values via a one-dimensional grid of equally spaced values. Input nodes are activated by a stimulus based on the similarity of the stimulis' cue values C to the input node's cue values I.

$$A_t = \exp(-\gamma(C_t - I_t)^2), \qquad (3)$$

with the activation A of the input nodes at time t further depending on a scaling parameter γ that determines the slope of the activation gradient. The activation of the input nodes is spread to the output nodes via connection weights. The activation of an output node O_r is given by the sum of activations of the input nodes weighted by the connection weights between the input nodes and the output node. The probability of choosing an output node is given by the ratio of the activation of the output node to the summed activation of all output nodes. The judgment is a weighted average of the output nodes, where each output node is weighted with the probability with which it is chosen. Connection weights are updated at each trial according to a delta-learning rule. For this it is assumed that the feedback criterion value produces a feedback activation of each output node F_r based on the similarity of the feedback value p_t to the output node p_r:

$$F_r(p_t) = \exp(-\gamma(p_t - p_r)^2). \qquad (4)$$

The connection weights α are updated based on the feedback activation F, the predicted activation O and the input activation A, with a learning parameter δ capturing the learning rate:

$$\alpha_t = \alpha_{t-1} + \delta\left[F_{t-1} - O_{t-1}\right]A_{t-1}. \qquad (5)$$

ALCOVE We extended ALCOVE (Kruschke, 1992) to continuous judgments. ALCOVE has a similar structure as

the ALM model; however, the input nodes of ALCOVE are restricted to the exemplars encountered during learning. As in ALM the activation of an input node is based on the similarity of the stimulus object to the input node. However in ALCOVE, similarity depends also on the attention given to each cue dimension k, which is controlled by a set of attention weights w.

$$A = \exp(-\gamma \sum_k w_k [c_k - i_k]^2), \quad (6)$$

with the activation A of an input node based on the squared distance of the stimulus value c on dimension k to the value of the input node i on cue dimension k, weighted by the attention w given to this cue dimension and a scaling parameter γ determining the slope of the activation gradient. In the original ALCOVE model, one output node is chosen as response. To allow for continuous judgments we extended ALCOVE with the ALM's estimation mechanism described above.

In ALCOVE, the connection weights are updated in the same way as in ALM, with learning parameter δ_1 capturing the learning rate (see Equations 4 and 5). Additionally, the attention weights are also updated according to a delta learning rule. The learning rate is captured in an additional free parameter δ_2. The attention weights w are updated according to the following rule:

$$w_{k,t} = w_{k,t-1} - \delta_2 \gamma \sum_r \left[(F_r - O_r) \sum_n A_n \alpha_{n,r} (c_k - i_k)^2 \right], \quad (7)$$

with r indexing the output nodes, n the input nodes and k the cue dimensions; F gives the respective feedback activation and O the predicted activation of an output node. A indicates the respective activation of an input node, α is the connection weights between the input and output node and c_k and i_k provide the stimulus value and the input node value on cue dimension k.

Method

To test how the performance of the learning models in solving judgment tasks depend on the task structure, we compared the models' performance by computer simulations in two environments: a linear environment and a multiplicative environment. Furthermore, we varied the amount of learning to examine the relationship between the models' performance and the size of the training set.

Simulation Environments We created two different environments: a linear environment and a multiplicative environment similar to the environments used by von Helversen and Rieskamp (2008; Experiment 3), which revealed a strong effect of task structure on people's judgment processes. Each environment consisted of 1000 objects described by 5 binary cues, with randomly drawn values (0 or 1). The criterion in the linear environment Y_L was generated by a linear additive function:

$$Y_L = 30 + 33c_1 + 22c_2 + 20c_3 + 15\,c_4 + 5c_5 + \varepsilon. \quad (8)$$

The error term ε was drawn from a normal distribution with a mean of zero and a standard deviation of 10. The multiplicative criterion Y_M was generated by a multiplicative function:

$$Y_M = 1.2 \cdot \exp(Y_L / 30), \quad (9)$$

resulting in criterion values with similar ranges (about 0 to 140) in both environments.

Simulation Procedure For the simulation we drew a random training sample of 250 objects 50 times and a hold-out set of 100 from each of the environments. Then we fitted the free parameters of the four models to the training data minimizing the square deviation between the model prediction and the training data. For the linear additive model we assumed that in the beginning, equal weight would be given to all cues. For the associative models we assumed that the connections weights and attention weights had equal starting values. Based on the estimated parameter values we generated model predictions for the hold-out set after seeing 20, 50, 150 and 250 objects from the training set. As a measure of prediction accuracy we calculated the root mean square deviation (*RMSD*) between the model prediction and the criterion in the hold-out set after the four points of learning and averaged across the trials of the simulation at each point of learning. Since parameters are fit on a separate data set, the performance of the models in the hold-out set can be compared without needing to further adjust for the complexity of the models.

Results

The mean best fitting parameter values for the models are reported in Table 1, indicating similar learning in both environments.

Table 1: Mean parameter values (SD)

	Environment	
Parameters	Linear	Multiplicative
Linear additive: δ	.45 (.30)	.30 (.17)
Linear additive: ω	.45(.14)	.47 (.13)
Mapping: d	0 (0)	.02 (.14)
ALCOVE: γ	.30 (.36)	.22 (.17)
ALCOVE: δ_1	.42 (.56)	.46 (1.44)
ALCOVE: δ_2	145 (50)	173 (63)
ALM: γ	2.72 (.31)	1.78 (.30)
ALM: δ	.14 (.07)	.22 (.07)

The models differed with regard to how well they learned the criterion values in the training set. In particular, the two

associative models performed less well than the mapping model and the linear additive model (see Table 2).

Table 2: Mean model performance in *RMSD* (SE) in the training set

	Environment	
Models	Linear	Multiplicative
Linear additive	11.09 (.07)	9.78 (.21)
Mapping	14.60 (.08)	9.87 (.16)
ALCOVE	15.18 (.09)	10.32 (.18)
ALM	15.05 (.12)	11.51 (.17)

The results in the hold-out set suggest that the performance differences in the training set are partly due to a slow initial learning process of the associative models. Figures 1 (linear environment) and 2 (multiplicative environment) show that the linear additive model and the mapping model learn rather quickly even with as little as 20 learning trials. However, the two associative models that performed worse with less than 50 learning trials caught up with the other two models after extensive learning of 150 trials or more.

The environment crucially influenced the performance of the models. Unsurprisingly, in the linear environment, the linear additive model performed best regardless of the amount of training. With fewer than 50 learning trials, the mapping model performed somewhat worse than the linear model, but better than the associative models. However, with more than 150 trials of learning the two associative models performed better than the mapping model and almost as good as the linear additive model.

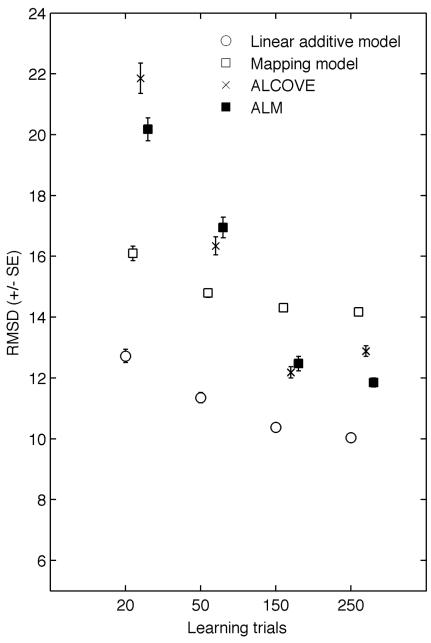


Figure 1: Model performance (*RMSD*) in the hold-out set in the linear environment after 20, 50, 150 and 250 trials of learning. Error bars denote one standard error of the mean.

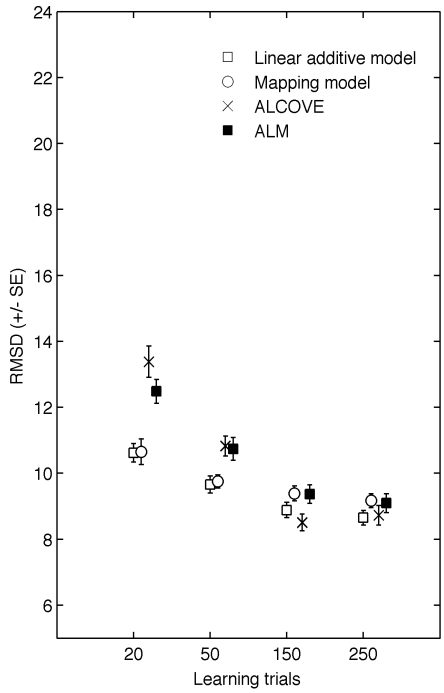


Figure 2: Model performance (*RMSD*) in the hold-out set in the multiplicative environment after 20, 50, 150 and 250 trials of learning. Error bars denote one standard error of the mean.

In the multiplicative environment, the advantage of the linear additive model was less pronounced. To begin with, it performed equally well as the mapping model, but gained a bit on the mapping model with more than 150 trials of learning. The two associative models again performed worse than the linear and the mapping models with little learning with fewer than 50 learning trials, but caught up after more than 150 trials of learning.

In summary, the linear additive model performed well in both environments and at all stages of learning. Furthermore, we found evidence that the amount of training affected which models are well suited to making accurate judgments. More specifically, the associative models only made accurate judgments after extensive training. In contrast, the mapping model performed reasonably well with little training, but failed to improve to a similar degree as the other models with further training.

Discussion

We investigated how different learning models can solve a multiple-cue judgment task depending on the amount of learning and the structure of the task. We found that a linear additive neural network model performed well in both environments and regardless of the amount of training. However, we also found differences due to task structure. In the multiplicative environment, the mapping model was

equally as good as the linear additive model, in particular with little learning experience. With extensive learning experience the two associative models, ALCOVE and ALM, performed similarly well to the linear additive model and the mapping model. The results are in line with the finding of Kelley and Busemeyer (2008) that a neural network with a linear basis was well suited to describe participants' judgments over a broad range of tasks. Our results also support research illustrating the robust performance of linear models for judgment tasks (Hogarth & Kareleia, 2007).

However, our results seem to contradict results that suggest task-dependent changes in strategy use in multiple-cue judgments (Juslin, et al., 2008; von Helversen & Rieskamp, 2008, 2009). These authors found in a task with a similar structure as in our simulation, that the model that was best in describing participants' judgments changed depending on the task structure. However, the judgment process people rely on might not only depend on the judgment performance of the learning process (e.g. Ashby, Alfonso-Reese, Waldron & Turken, 1998). Instead, the learning speed and also other factors such as the cognitive effort of relying on a specific cognitive process could also influence which judgment and learning process people follow (see also Enkvist, Newell, Juslin, & Olsson, 2006). Particularly, in the multiplicative environment the mapping model may provide an equally good but arguably cognitively simpler alternative, which could explain why a majority of participants were best described by the mapping model in the multiplicative condition of Experiment 3 by von Helversen and Rieskamp (2008). On the other hand, associative processes seem to provide a valid alternative to a linear additive model after extensive training, in particular in a multiplicative environment. If following the assumption that associative similarity-based processes may be executed without conscious awareness and be thus cognitively less demanding (e.g. Ashby & Maddox, 1994), this could still make it attractive for participants to rely on such processes, particularly after extensive training. This could explain the reliance on exemplar-based processes (Juslin, et al., 2008) and also the considerable minority of participants that were best described by the ALM model (see Kelley & Busemeyer, 2008).

Lastly, the available context information may also influence people's strategy choices. Information about cue polarity seems to trigger rule-based processes (Newell, Weston, Tunney, & Shanks, 2009; von Helversen & Rieskamp, 2009). While in the study by Juslin et al., (2008) participants had no information about cue polarity, most studies analyzed by Kelley and Busemeyer (2008) provided context information that allows drawing conclusions about cue polarity and thus could have increased the reliance on rule-based processes.

Conclusion

In sum, our results suggest that linear additive learning models are generally robust. However, the performance advantage depends on the task structure and the amount of learning opportunity. On the basis of these results future research will test whether people's judgments depend on task characteristics and learning opportunities.

Acknowledgments

This research was supported by a grant of the German Research Foundation to the first and second author (RI 1226/5).

References

Anderson, N. H. (1981). *Foundations of information integration theory*. New York: Academic Press

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review, 105*, 442-481.

Brehmer, B. (1994). The psychology of linear judgement models. *Acta Psychologica, 87*, 137-154.

Brehmer, B., & Joyce, C. R. B. (1988). *Human judgment: The SJT view*. Amsterdam: Elsevier/North Holland.

Busemeyer, J.R., Myung, I.J., & McDaniel, M.A. (1993) Cue competition effects: Theoretical implications for adaptive network learning models. *Psychological Science,* 4, 196-202

Busemeyer, J. R., Byun, E., DeLosh, E. L., & McDaniel, M. A. (1997). Learning functional relations based on experience with input-output pairs by humans and artificial neural networks. In K. Lamberts & D. Shanks (Eds.), *Knowledge concepts and categories* (pp. 405–437). Cambridge, MA: MIT Press.

Cooksey, R. W. (1996). Judgment analysis : theory, methods, and applications. San Diego: Academic Press.

Doherty, M. E., & Kurz, E. M. (1996). Social judgment theory. *Thinking and Reasoning, 2*, 109-140.

Enkvist, T., Newell, B. R., Juslin, P., & Olsson, H. (2006). On the role of causal intervention in multiple-cue judgment: Positive and negative effects on learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*, 163-179.

Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General, 117*, 227-247.

Hogarth, R. M., & Karelaia, N. (2007). Heuristic and linear models of judgment: matching rules and environments. *Psychological Review, 114*, 733.

Juslin, P., Karlsson, L., & Olsson, H. (2008). Information integration in multiple cue judgment: A division of labor hypothesis. *Cognition, 106*, 259-298.

Kelley, H., & Busemeyer, J. (2008). A comparison of models for learning how to dynamically integrate multiple cues in order to forecast continuous criteria. *Journal of Mathematical Psychology, 52*, 218-240.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review, 99*, 22-44.

Newell, B., Weston, N., Tunney, R., & Shanks, D. (2009). The effectiveness of feedback in multiple-cue probability

learning. *The Quarterly Journal of Experimental Psychology, 62*, 890-908.

Rieskamp, J., & Otto, P. E. (2006). SSL: A theory of how people learn to select strategies. *Journal of Experimental Psychology: General, 135*, 207-236.

von Helversen, B., & Rieskamp, J. (2008). The mapping model: A cognitive theory of quantitative estimation. *Journal of Experimental Psychology: General, 137*, 73-96.

von Helversen, B., & Rieskamp, J. (2009). Models of quantitative estimations: Rule-based and exemplar-based processes compared. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*, 867-889.

Success in Theory of Mind

Rose M. Scott (rmscott2@uiuc.edu)
Department of Psychology, University of Illinois at Urbana-Champaign, 603 East Daniel St., Champaign, IL 61820 USA

Adam R. Petrashek* (arpetras@uwaterloo.ca)
Department of Psychology, University of Waterloo, 200 University Ave, Waterloo, ON N2L 3G1 Canada

Noah D. Goodman (ndg@mit.edu)
Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology,
77 Massachusetts Ave, Cambridge, MA 02139 USA

Rebecca Saxe (saxe@mit.edu)**
Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology,
77 Massachusetts Ave, Cambridge, MA 02139 USA

* denotes organizer, ** denotes discussant

Keywords: Theory of mind; cognitive development; social cognition; executive function; social learning; domain-specificity; probabilistic modeling; reaction time.

Introduction

Peter wants to get the beer that he left in the refrigerator. Predicting Peter's behaviour correctly is usually an easy matter, but understanding how people correctly predict his behaviour with ease is a much more difficult task. Thirty years of research on theory of mind has focused on the interesting few cases in which people fail to reason about mental states correctly, however it is perhaps more interesting to explore the common, reliable cases of successful theory of mind reasoning. This symposium presents research exploring successful instances of theory of mind reasoning using a variety of experimental approaches, and examines the ability to succeed consistently across the lifespan, with results from toddlers, preschoolers, young children, and adults. Important conclusions are drawn from the presented research, which includes the first evidence that children as young as 2.5 years of age can succeed on explicit false belief tasks (Scott & Baillargeon), the most direct behavioral evidence to date for inhibitory processing in successful behavior prediction based on false belief and avoidance desire in preschoolers and young children (Petrashek & Friedman), and, in adults, evidence from a probabilistic modeling approach to theory of mind and social learning development with extensions to pragmatic language usage and natural pedagogy (Goodman).

Why do infants succeed in false-belief tasks when toddlers fail? Evidence for a response account

Rose M. Scott & Renée Baillargeon

Recent evidence suggests that infants in the second year of life can represent a variety of different false beliefs, as well as reason about false perceptions and deception (e.g., Baillargeon, Scott, & He, in press). If infants can represent false beliefs, then why do children fail standard tasks until age 4? Here we argue that this discrepancy reflects the use of different responses. Traditional tasks require children to answer a direct question about an agent's false belief (elicited-response tasks), whereas recent tasks measure children's spontaneous reactions to a scene (spontaneous-response tasks). Simultaneously representing a false belief and planning a response may be too difficult for young children. Since spontaneous tasks do not require a planned response, children succeed much earlier. To examine this possibility, we tested 2.5-year-olds in a novel false-belief task that closely matched the demands of standard tasks but did not require answering a question. While viewing a picture book, children heard a story about an agent who hid her apple in one of two locations; in her absence, the apple was moved to the other location. In the test trial, one picture showed the agent searching for her apple where she had originally hidden it, and one picture showed the agent searching for her apple in its current location. Children looked reliably longer at the original- than at the current-location picture, suggesting that they successfully represented the agent's false belief.

We next tested whether 2.5-year-olds could succeed in an elicited-response task if the response component were made easier for them. Specifically, we provided children with practice with the required response (pointing to one of two locations). In each trial, an experimenter either recited a line of the story (story trials) or asked a question (question trials). On story trials, one picture was shown; on question trials, two pictures were shown and the question required the children to point to one of them. In the final trial, children were asked to point to where the agent would look for her apple. Most children pointed to the correct location (e.g., where the agent falsely believed her apple was

located), suggesting that even 2.5-year-olds can succeed at an elicited-response false-belief task when the response demands are reduced.

The signature of inhibition in theory of mind

Adam R. Petrashek & Ori Friedman

Three-year-olds typically fail standard false belief tasks, whereas four-year-olds typically pass. Much has been made of this transition from failure to success, and it is now widely believed that improvements in inhibitory processing during the preschool years are at least partly responsible for improvements in theory of mind reasoning during the same period (Carlson & Moses, 2001). However, the role of inhibition remains unclear. One promising possibility is that inhibitory processing is involved in certain types of explicit mental state reasoning, such as predicting behaviour based on false belief, and directly affects *how* children perform on theory of mind tasks (Leslie, Friedman, & German, 2004).

Our research capitalizes on the lingering property of inhibition – once a response is inhibited, this inhibition lingers, making it more difficult to select than uninhibited responses. This signature of inhibition is highlighted in inhibitory accounts of negative priming and inhibition of return, which both occur in children.

In four experiments, we provide decisive evidence for the view that inhibitory processing is necessary to make explicit behavioural predictions based on avoidance desires and false beliefs. Attributing false beliefs may require inhibiting a default tendency to attribute true beliefs and, in Experiments 1 and 2, we show that inhibition lingers after 5- and 6-year olds predict an agent's behaviour based on a false belief. Attributing avoidance desires may require identifying the target to be avoided and then inhibiting it. In Experiments 3 and 4, we show that inhibition also lingers after 3-year-olds predict behaviour based on avoidance desire. In demonstrating a signature of inhibition in children's theory of mind reasoning, these four experiments clearly support the view that inhibitory processing is involved in how children successfully predict behaviour based on avoidance desires and false beliefs.

Learning what others know

Noah D. Goodman

Civilization is possible because no human needs to re-discover every fact and idea from the natural world alone. Instead, we can learn what other humans already know. What computational processes underlie this social learning, particularly early in development, before formal schooling begins? I will describe a probabilistic modeling approach to theory of mind, which addresses this problem. In this approach an understanding of other agents as goal-directed and an assumption that they are knowledgeable about the world supports social learning which is much more rapid than learning from the natural world alone. I will apply this framework to explain several experiments on social learning, and indicate how it extends to aspects of pragmatic usage of language and natural pedagogy.

Discussant

Rebecca Saxe

An Assistant Professor at the Massachusetts Institute of Technology, Dr. Saxe utilizes a multi-method, multi-directional approach to studying the cognitive neuroscience of theory of mind in both typical and atypical populations of infants, children, and adults. Saxe has received several prestigious awards and has been published extensively in top journals, including *Trends in Cognitive Sciences*, *Psychological Science*, and *Cognition*.

References

Baillargeon, R., Scott., R. M., & He, Z. (in press). False-belief understanding in infancy. *Trends in Cognitive Sciences*.

Carlson, S. M., & Moses, L. J. (2001). Individual differences in inhibitory control and children's theory of mind. *Child Development, 72,* 1032-1053.

Leslie, A.M., Friedman, O., & German, T.P. (2004). Core mechanisms in 'theory of mind'. *Trends in Cognitive Sciences,* 8, 528–533.

Disentangling Representation from Conceptualisation

Nancy Adrienne Salay (salay@queensu.ca)
Queen's University, Department of Philosophy, Watson Hall 309
Kingston, ON K7L 3N6 CANADA

Abstract

Drawing on recent work in the area of episodic memory, I suggest a novel way of dissolving the representation/anti-representation debate; if we treat representation and conceptualisation as two separate capacities, the latter being parasitic on the former, we unify the insights of both camps, but succumb to none of their failings. I provide a sketch of how we might understand de-conceptualised representation and I show that, on this new approach, many of the old problems, e.g. grounding, disappear.

Keywords: representation; anti-representation; intentionality; episodic memory; cognitivism; cognition; dynamic systems

Introduction

One of the central debates in cognitive science is the dispute over the role of representation in cognition: on one view, cognition is taken to be a kind of conceptual process; on the other, it's seen as a particular sort of physical process. But this characterisation of the two perspectives leads to a deeply unsatisfying theoretical divide: either our interactions with our world are mediated by our representations or we do a lot less modelling than we think. Accounts situated in the representational camp are plagued by the problem of intentionality, while those leaning towards the anti-representational side seem incapable of saying anything theoretically interesting about the higher-level conceptual type of cognition that purportedly develops out of the underlying non-representational physical processes. This unhelpful polarisation is due, I think, to a presupposition that drives the debate; the idea that representations are conceptual[1], that is, that they stand in for classes of things, not for particulars. I want to suggest that if we unpack this, and treat representation and conceptualisation as two separate capacities, the latter being parasitic on the former, we get a more nuanced and robust theory of cognition that unifies the insights of both camps.

In what follows, I'll take Hubert Dreyfus' arguments against cognitivism (computational/representational approaches) as paradigmatic of the anti-representational camp, since his are the most explicit and best worked out in the literature. I'll describe the competing views, uncover the underlying intuitions that motivate the respective positions, and then suggest a new paradigm for understanding mental re-presentations as symbols of particulars.

Why Cognitivism is in Trouble

The central point of contention between Dreyfus and the cognitivist is whether or not cognition is ultimately a matter of *symbol* manipulation. Cognitivism begins with the idea that a cognitive system responds intelligently to its environment *by way of* its internal symbols but it is precisely this commitment, Dreyfus thinks, that constitutes cognitivism's fatal flaw. As soon as one thinks that internal stuff can stand for external stuff, the problem of how the symbols are connected to their objects arises. What makes the symbols *symbols*? How do the inner representations get to be *about* the things they represent? What makes the inner models, *models* of the external world and, further, models of this specific part of the external world? If, as Quine so eloquently argued (1951), concepts are so deeply entangled with one another that having one entails that one must also have a host of others, this web of concepts – even if internally coherent – could never model the stuff out there; such a system could never decide, for example, that a particular real-world situation is a birthday party situation because the rule for recognising a birthday party situation would cite other concepts that would in turn need rules of application and those rules would also cite concepts that would need rules of application and so on *ad infinitum*.

Now of course one way to hold on to the representational story, but avoid the infinite regress, is to reject concept holism and suppose that there is some level of symbols for which the fixing process is not mediated by other symbols, that is direct in some non-representational way. This is the project of giving a naturalistic account of intentionality. If all holistically inter-linked symbols are 'grounded' in a level of symbols that have their contents fixed in some non-intentional way, the idea goes, the regress will end there. Recognition that representational accounts that don't have some grounding layer – we should think of these as disembodied or ungrounded cognitivist approaches – face serious problems of exactly the sort that Dreyfus raises has made the project of "solving the grounding problem" an increasingly topical one.[2] But, for Dreyfus, grounding cannot be a way out of the vicious circle since grounding requires a commitment to some form of concept atomism – the view that individual symbols can represent

[1] Indeed, Fodor (1990) goes to some length to argue that being a generalisation is a *requirement* for being a representation; symbols that stand in for particulars only are mere "labels" on his view. I think this is deeply mistaken and rests on the confusion that is the focus of this paper.

[2] See Taddeo & Floridi (2005) for a good review of the current state of work in this area.

independently of all other symbols – and he thinks that there are independent reasons for thinking that concept atomism must be false.

Dreyfus thinks that concept atomism is wrong in two ways. Not only does it not speak to the Quinean intuitions about concept holism – that concepts come in groups, not as individuals – but it assumes that our most basic relation with the world is a conceptual one. Dreyfus thinks that both the atomists and (some) holists are wrong in thinking that, at base, we come to know our world by theorising about it. He follows Heidegger in insisting that fundamentally our relation to the world is not a theoretical one, but a practical one. On this Heideggerian view, the idea that *things in the world* are meaningful in isolation from our *practices* is incoherent. Things are what they are and play the role they play partly because of their natural features – being heavy, being sharp, and so on – and partly because of the background practises of the culture within which those roles develop.

> Although practical understanding – everyday coping with things and people – involves explicit beliefs and hypotheses, these can only be meaningful in specific contexts and against a background of shared practices. And just as we can learn to swim without consciously or unconsciously acquiring a theory of swimming, we acquire these social background practices by being brought up in them, not by forming beliefs and learning rules. (Dreyfus, 1980, p.7)

Practical understanding underwrites and makes theoretical understanding possible: meaning is always situated, that is, it arises out of holistic, dynamic, inter-relations between agents and their environment. Meaning *arises*, as a whole, out of activity, and never individually, as a result of assignments:

> To say a hammer has the function of being for hammering leaves out the defining relation of hammers to nails and other equipment, to the point of building things, and to our skills – all of which Heidegger called readiness-to-hand – and so attributing functions to brute facts couldn't capture the meaningful organization of the everyday world. (Dreyfus, 2008, p.1138)

Thus, it's the complex nexus of background practices *and* (sometimes) conceptual relations that holds together hammers, nails, wood, etc., that connects *hammer* with hammers. Any view that *requires* a base-level of concept-detectors, as it appears cognitivism does if it is to avoid the regress, has been completely derailed. The fact that it was the initial assumption of representationalism that was responsible for this flight towards conceptual atomism, Dreyfus urges, should be a red flag that something is deeply wrong with that assumption.

Now some might think that Dreyfus is creating much ado about nothing, that if there is any debate here it's merely a terminological one since the internal states, or at least some subset of them, that underlie our practical understanding just *are* the low-level representations that ultimately ground a higher-level theoretical conceptual structure. The rising influence of neuroscience in the cognitive sciences coupled with the widespread acceptance of some kind of information theoretic account of representation has made this idea that conceptual cognition might be grounded in a simple capacity for object detection a natural one.

But this is no mere terminological conflict; it is a deep and confounding burden of proof debate: there seems to be no non-question-begging way of specifying what constitutes a representational system. If we are loose with our use of the term "representation" and we suppose that nomic covariance relations are sufficient to establish representation relations, then we are in danger of begging the question in one way, of assuming that *using a representation* and *acting in a way that can be interpreted as using a representation* are two sides of the same coin. This "loose" understanding is mainstream in neuroscience. When a neuron or a cluster of neurons is found to be 'sensitive' to a particular class of objects, and the underlying explanation is taken to be that a nomic causal relation between the objects of some class and the activation of a neuron or cluster of neurons has been found, that neuron or cluster is said to represent that class. There is a dispute, to be sure, over whether or not localist, single-cell, representations are possible or whether neural representations are distributed over clusters of neurons; but there is very little discussion (except among the anti-representationalists of course) about whether or not nomic covariance is sufficient to warrant representational attribution[3].

But surely, one might think, there is a difference between what I do when I consult a map in order to find the shortest route across the city and what I do when I follow a series of instructions for crossing the city. In the first case I am using the map in virtue of its content, but in the second case, while the entire sequence of steps taken together could be viewed as a model of the shortest distance across the city, I do not follow the instructions in virtue of their semantic content, I follow them in virtue of their syntactic properties – turn left at Bank Street, proceed for two blocks, and so on. If I didn't understand the semantic features of the map, e.g. if I didn't know that the black lines represented streets and the red lines stood for highways, I wouldn't be able to use the map; on the other hand, whether or not I understood that, taken as a whole, the sequence of steps represented the shortest path across the city, I could follow the instructions for taking that path. In the first case I am using the representation, but in the second I am merely acting in a way that could be interpreted as using a representation. Neuronal chain reactions, looked at from an investigator's vantage point, can certainly be interpreted as

[3] See, for example, the recent debate – rekindled by Jeffrey Bowers (2009) – concerning localist vs. distributed neural representations. Nowhere in this discussion is the question of whether or not we should be calling these regularities representational at all raised.

representational, in the same way that the sequence of steps can be seen as a model of the shortest route across the city, but unless one has some kind of story to tell about how one part of the system, or perhaps the system taken as a whole, makes use of the content of those neural states, we have no reason to think that these neuronal impulses actually play a representational role. Of course, by stating the distinction in this way, I am also begging the question, in the other way, since to get to my conclusion one has to first assume that *using a representation* and *acting in a way that can be interpreted as using a representation* are different. This is why the debate about representation seems so intractable; the competing intuitions that undergird the various positions are so polarised.

William Ramsey does a good job of making the gap between these alternative perspectives explicit when he argues that any account of representation must meet what he calls the *functional specification challenge*: "a minimal requirement for a successful functional specification of any notion of representation is that the content—or, if you like, the fact that the representation has semantic content—be an explanatorily relevant fact about that state." (2003, p. 129) In other words, one needn't go so far as to show that a system is actually using a representation in order to make the case that the system is a representational one (as in my map example) – it's unclear how one could ever give a naturalistic account of intentionality if this were the requirement – but one does need to provide a justification for treating a system as though it were using a representation; that is, the fact that the internal indicator states have some content must play *some* kind of role in one's account.

Fred Dretske's information theoretic account of representation (1988), perhaps the most robust and ambitious indicator theory of representation that has been offered thus far, looks like the best candidate for meeting this challenge. According to that account, what makes one internal state X a representation of some class of things or actions Y is the following:

1. The presence/absence of X covaries with the presence/absence of members of Y;

2. The co-variance is under-written by a nomic causal relation, that is, the presence/absence of members of Y cause or are a necessary part of the cause of the presence/absence of X; and,

3. The functional role of X, within the system within which it arises, is to carry information about the presence/absence of members of Y.

It's condition three that makes this account a candidate for meeting the functional specification challenge, since it's this requirement that makes the content of the purportedly representational state relevant to a complete description of its functional role in the system. Or so it seems. Ramsey argues that it doesn't. To support his contention, Ramsey develops a distinction between *carrying information about* – "possess[ing] states that could inform about other states of affairs" (2003, p.135) and *being an informer*— "be[ing] plugged into the right sort of system in the right sort of way, such that the relevant entailment relations are put to a very specific sort of use." (2003, p. 135) It's the latter that is required to meet the functional specification challenge, since only in such cases is the information actually playing some kind of role in the overall account. But in none of Dretske's examples, Ramsey argues, is the *informer* condition met.

I won't rehearse here Ramsey's support for this claim since ultimately it's not important that we be convinced of Ramsey's conclusion; indeed, one of the morals of this paper is that so long as some of our key presuppositions about the nature of representation remain, we will never be able to solve this burden of proof debate. Ramsey's insights are important, however, because they uncover the fact that information theoretic accounts of representation are convincing only if we assume a particular (impoverished, on one view) understanding of representation; consequently, we shouldn't be optimistic that an information theoretic grounding account could ever settle the score.

But as I've already noted, for anti-representationalists like Dreyfus, the entire grounding agenda, information theoretic or not, is misguided ultimately because it cannot accommodate our dynamic nature as systems who are continually responding to and causing changes in our environment. Any view on which it makes sense to see coping skills as decomposing into finer and finer grained skills at dealing with object types, even those that are "action-oriented" or Gibsonian, is a representational view by Dreyfus' lights and thus one that he rejects. On the representational view, our interactions with the world are mediated by categories and we see the world *as* divided up into hammers and tables and chairs. In coping, on the other hand, there is no "seeing as" at all. One copes with situation wholes, as unfolding happenings, rather than as composites of objects.

Dreyfus' anti-representationalism is thus quite radical; he rejects *any and all* representational interpretations of the internal states that underpin our coping skills. There is no mere terminological argument here.

Re-Presentation: A New Model

Dreyfus' deep and important insight into the way we think has led us to the following impasse: any disembodied or ungrounded AI founded on the principle that cognition is fundamentally a matter of concept manipulation will be caught in an infinite regress of concept consultation and, consequently, its concepts will fail to be about the things they purportedly represent. But the mainstream cognitivist response, to close the concept-world gap by grounding conceptual schemas, requires that we take concept atomism seriously. Dreyfus rejects this route and takes the fact that this looks to be the only way out of the infinite regress as a

clue that the initial representational assumption must be at fault. Others, who are more firmly rooted in the Cartesian tradition, are not as quick to reject the grounding possibility; they think that some kind of information theoretic atomism will eventually provide the answer. But, as Ramsey show us, low-level concept detectors alone can't provide us with a naturalistic account of representation, since they cannot play the required functional role of concepts or even proto-concepts unless they are already part of an intentional system, a system capable of *using* semantic content. Such grounding theories, instead of solving the problem, merely push it back a level. Dreyfus urges us towards an anti-representational AI, but I suspect the radical see-saw between the full-blooded representationalist and the strident anti-representationalist is a tug 'o war that no-one is likely to win, likely because, as Andy Clark and others (1994, 1997, 2002) have been suggesting all along, the truth lies somewhere in the middle. I suggest that we take Dreyfus' charge seriously, but explore other ways out of the impasse. His solution is to reject the foundational commitment to some form of representationalism outright, but I want to argue that even if we accept Dreyfus' arguments that cognition isn't fundamentally a conceptual process, we need not accept his more radical and less helpful conclusion that it is also not a representational process.

As evolved biological agents, there is no doubt that, as Dreyfus emphasises, we first and foremost cope with our environment in an entirely non-representational way – we avoid obstacles, recoil from harmful situations, and are drawn towards safe and pleasant ones in an unmediated way. But we're not *just* biological agents; we're cognitive ones as well. And as such we are able to respond not only to the intricacies of present situations, but to past and future ones as well. I am able to respond to the subtleties of the ebb and flow of traffic, while I'm driving my car, while at the same time, thinking about how my class went yesterday and considering ways in which I will do things differently or the same in next week's class. But while this ability does require that we have some capacity to re-present the past and imagine future situations, it need not require a conceptual ability, an ability to generalise beyond the specific cases to a class. I have in mind here the distinction, first proposed by Endel Tulving, between episodic and semantic memory. Episodic memories are re-presentations of past experiences (and imaginings of future ones), while semantic memories are conceptualisations of past experiences – they consist in the knowledge we distil from our experiences, that is, the generalisations we make on the basis of experience. Proust's *In Search of Lost Time* gives us wonderfully vivid descriptions of both. The narrator has an episodic memory of a particular moment in his childhood, when he'd tasted a bite of his aunt's lime-tea-soaked madeleine, and in this re-experiencing he tells us that "immediately the old grey house upon the street, where her [his aunt's] room was, rose up like a stage set to attach itself to the little pavilion opening on to the garden which had been built out behind it for my parents (the isolated segment which until that moment had been all that I could see); and with the house the town, from morning to night and in all weathers, the Square where I used to be sent before lunch, the streets along which I used to run errands, the country roads we took when it was fine." (p.50) This memory is portrayed in stark contrast with his more conventional semantic memories of Combray, the village of his childhood summers, where his aunt lived: "Many years had elapsed during which nothing of Combray, save what was comprised in the theatre and the drama of my going to bed there, had any existence for me." (*ibid.*) The ability to have an episodic memory, then, is the ability to re-experience some situation not present, while to have a semantic memory is to have some capacity for generalisation. Tulving, both when he first suggested the distinction and today, sees episodic memory as parasitic on semantic memory:

> Episodic memory is a recently evolved, late-developing, and early-deteriorating past-oriented memory system, more vulnerable than other memory systems to neuronal dysfunction, and probably unique to humans. It makes possible mental time travel through subjective time, from the present to the past, thus allowing one to re-experience, through autonoetic awareness, one's own previous experiences. Its operations require, but go beyond, the semantic memory system. (Tulving, p. 6)

Martin Conway, however, has recently suggested an intriguing new way of understanding episodic memories as a tripartite structure only elements of which are entwined with semantic memories. His analysis provides us with a new way of understanding the relationship between mental representations and conceptualisation, one that can serve as the foundation for the theoretical bridge between our coping skills and our conceptual abilities.

On this new picture, episodic memories can be analysed into inter-related parts:

1. Episodic elements (EE's) – these are snippets of experiences, Proustian experience snapshots;
2. Semantic episodic memories (SEM's) – these are small sets of EE's grouped by a contextualising conceptual frame, for example, one's breakfast routine; and,
3. Conceptual episodic memories (CEM's) – these are groupings of SEM's by a higher-order conceptual frame, for example, a day at work.

Of these, the most basic are EE's:

> Episodic elements are the most event-specific, most experience-near representations in long-term memory. They are often in the form of visual images (which may be the main representational format of episodic memories) and they represent moments of experience or summaries of moments of experience, particularly

> and perhaps exclusively, moments of conscious experience. (Conway, p. 2308)

Importantly, on Conway's view, and contra the received wisdom, EE's are a-conceptual and conceptually *a priori*. This means that the capacity for semantic memory is not a requirement for having episodic elements; indeed, he suggests, the relation goes the other way – episodic elements are required for semantic memory.

> An interesting question that then arises is: how can EEs be associated with conceptual frames in an infant's memory? One answer to this is that the ability to form EEs is hard-wired and functioning prior to birth. Conceptual knowledge is abstracted from EEs. (p. 2312)

Conway here is suggesting that free-floating EE's, what I'm calling re-presentations, that aren't yet grouped and framed, might be the building blocks of our concepts. At some point in a human infant's development, EE's begin to be grouped. These groupings form 'proto-SEM's', the beginnings of conceptual frames. The mechanism for grouping is of course the big ticket question since, on this view, this just is the mechanism for conceptualisation. An initial suggestion is that both the temporal contiguity of EE's and closeness with respect to sensory attributes are likely factors in how EE's are grouped in long-term memory.

Obviously, more work needs to be done in this area before we can claim anything as bold as a theory of concept development, but in speaking to the insights of both the representational and anti-representational camps, EE's as re-presentations play an important role in the cognitive story we have collectively been telling thus far. The anti-representationalist is motivated by the bottom-up observation that cognitive agents such as ourselves are, most fundamentally, dynamical physical systems and, as such, are best described in terms of the low-level mechanisms from which our higher level capacities emerge. That re-presentations are the building blocks out of which our conceptual capacities emerge supports this picture since re-presentations themselves are just responses to past experiences, neurally encoded and re-played and, in that sense, are no different from other coping responses. The representationalist, on the other hand, is motivated by the top-down observation that, unlike most physical systems, cognitive agents are able to respond to past events and possible future ones in addition to present situations. Here again, the insight is captured since a re-presentation is either a response to an experience that has already happened or is a playing out of a possible future one: "The temporal dimension in episodic memory extends then both backward and forward in time and we have recently termed this the *remembering-imaging window.*" (p. 2307)

Recall that attempts to ground conceptual schemas directly in some kind of body-world relation (this is what the detector accounts try to do) can never resolve the clash of intuitions. Either the attribution of representation to internal indicator states will be unwarranted, given a more robust notion of representation, or it will be justified only if we assume that the overall system is already intentional, which is the very thing we are trying to explain. Instead of responding to this observation by embracing the opposing intuition, as Dreyfus does, I'm suggesting a possible theoretical middle ground. We do rely on inner models of our environment in our interactions, but, as part of a low-primitive cognitive capacity, these models are wholly particular, not conceptual. There is no gap between a re-presentation and the world that needs to be bridged (as in the case of conceptual representation); the re-presentation is just another situation to be (re)experienced. Re-presentations do not threaten the underlying anti-Cartesian picture of ourselves as, ultimately, dynamic copers, since re-presentations don't mediate our interactions with the world; although they do make mediation possible, since it is out of these snapshots that concepts develop. Finally, re-presentations meet the functional specification challenge without begging the intentional question. Being a neuronal response and thus embedded within a network of connections, a re-presentation triggers other responses as well. As a first response, of course, not re-experienced, there is nothing representational about the underlying neuronal structure that encodes that response. But once that set of neural encodings is re-activated, it now has the role of carrying information about the original situation for the system to which it is being re-presented. Not just any set of neural encodings can count as representational then, no matter how detector-like they behave; only those in a system that is capable of re-presenting to itself count as representations, since only in such systems is the experience as a whole, that is to say, the content of the experience, playing a role.

Conclusions and Speculations

It sounds like we get to have our cake and eat it too. Is this too good to be true? Perhaps, but it certainly opens up some new avenues of investigation where now we seem to be stalled. The anti-representationalist seems incapable of offering a theory of cognition, distinct from a theory of, say, action, because he ignores our theoretical capacities; the representationalist, on the other hand, seems incapable of offering a theory of embodied cognition because she begins the cognitive account too high up – already at the level of concepts. If we let go of the idea that a representation must relate some particular to a class, we have a way of marrying the insights of either approach and moving forward with a new conception of intentionality: an intentional being is one that has the capacity to respond to its own response to some past experience. Intentionality, on this view, is a *prerequisite for* conceptual cognition; that is, giving an account of intentionality and giving an account of mental concepts are separate endeavours. And this is very happy news because until now we've had the proper order of investigation backwards; the thought was that once we

understood conceptual representation, the right story about intentionality would follow. But no naturalistic account of conceptual cognition is forthcoming, no surprise, since that story is parasitic on, not precedent to, an account of intentionality. This scaled back understanding of intentionality as the capacity for re-presentation, however, looks like a much more promising candidate for naturalisation. If we can manage that, we'll have, at last, a naturalistic grounding for a theory of conceptual representation.

References

Clark, A. & Toribio, J. (1994). Doing without representing? *Synthese*, *101*, 401–431.

Clark, A. (1997). *Being there: putting brain, body, and world together again.* Cambridge: MIT Press.

Clark, A. (2002). Is seeing all it seems? Action, reason, and the grand illusion. *Journal of Consciousness Studies, 9(5/6)*.

Conway, M.A. (2009). Episodic memories. N*europsychologia*, *47*, 2305–2313.

Dretske, F. (1988). *Explaining Behavior*. Cambridge: MIT Press.

Dreyfus, H. (1980). Holism and Hermeneutics. *The Review of Metaphysics*, *34 (1)*, 3-23.

Dreyfus, H. (2002). Intelligence without representation – Merleau-Ponty's critique of mental representation. *Phenomenology and the Cognitive Sciences*, *1*, 367–383.

Dreyfus, H. (2008). Why Heideggerian AI failed and how fixing it would require making it more Heideggerian. *Artificial Intelligence*, *171(18)*, 1137–1160.

Freeman, W. (2000). *How brains make up their minds*. New York: Columbia University Press.

Fodor, J. (1990). Information and representation. In P. Hanson (Ed.), *Information, language, and cognition*. Vancouver: University of British Columbia Press.

Harnad, S. (1990). The symbol grounding problem. *Physica D*, *42*, 335-346.

Keijzer, F. (1998). Doing without representations which specify what to do. *Philosophical Psychology*, *II(3)*, 269-302.

Keijzer, F. (2002). Representation in dynamical and embodied cognition. *Cognitive Systems Research, 3*, 275–288.

Kirsh, D. (1990). When is information explicitly represented? In P. Hanson (Ed.), *Information, language, and cognition*. Vancouver: University of British Columbia Press.

Prinz, J. & Barsalou, L. (2000). Steering a course for embodied representation. In E. Dietrich & A. Markman (Eds.), *Cognitive dynamics: Conceptual changes in humans and machines* (51-77). Cambridge: MIT Press.

Proust, M. (1913-27). *Remembrance of Things Past. Volume 1: Swann's Way: Within a Budding Grove*. The definitive French Pleiade edition translated by C.K. Scott Moncrieff and Terence Kilmartin. New York: Vintage.

Quine, W.V.O. (1951). Two Dogmas of Empiricism. *The Philosophical Review, 60*, 20-43.

Ramsey, W. (2003). Are receptors representations? *Journal of Experimental & Theoretical Artificial Intelligence, 15:2*, 125-141.

Taddeo, M. & Floridi, L. (2005). Solving the symbol grounding problem: a critical review of fifteen years of research. *Journal of Experimental & Theoretical Artificial Intelligence*, *17 (4)*, 419–445.

Tulving, E. (2002). Episodic memory: From mind to brain. *Annu. Rev. Psychology*, *53*, 1–25.

Van Gelder, T. (1995). What might cognition be, if not computation? *Journal of Philosophy*, *XCII(7)*, 345-381.

Varela, F., Thompson, E., and Rosch, E. (1991). *The embodied mind*. Cambridge: MIT Press.

Wheeler, M. (2008). Cognition in context: phenomenology, situated robotics and the frame problem. *International Journal of Philosophical Studies, 16 (3)*, 323-349.

Thought, language and mental representation

Jonathan Trigg (jon.trigg@louisiana.edu)
Program in Philosophy, UL Lafayette
Lafayette, LA 70504 USA

Michael Kalish (kalish@louisiana.edu)
Institute of Cognitive Science, UL Lafayette
Lafayette, LA 70504 USA

Abstract

We examine the proposal that thinking is a combinatorial operation on mental representations, and argue that it cannot be. If the argument is successful it shows that cognitive science cannot explain intelligent linguistic behavior by explaining what thinking is. We point out that this does not impugn the practice of cognitive scientists interested in human language, which, properly understood, consists in the framing and testing of hypotheses about the causally necessary enabling conditions of intelligent linguistic behavior.

Keywords: Wittgenstein; thought, language, mental representation, language of thought.

Introduction

> We regard understanding as the essential thing, and signs as something inessential. (Wittgenstein, 1974)
>
> ...the limits of possible thought are the limits of the possible expression of thought. (Bennett & Hacker, 2003)

If strong Wittgensteinian cross currents still ran into the mainstream of contemporary philosophy of mind and language, these philosophical waters would be much more turbulent than they now are. Indeed it is not even clear that they would be running in roughly their present direction. As things stand the river flows wide and slow, almost undisturbed by substantial impediments to its progress, and serious attempts to change its course are liable to seem naive or over ambitious – uncomprehending of the forces at work.

The argument we present here, which is Wittgensteinian in spirit, is meant to push quite hard against the prevailing drift. It's an argument to the conclusion that thinking is not mental representing and thoughts are not mental representations. We are not, by any means, the first to make this sort of argument. Indeed, on a plausible reading of his two great works, the very theory Wittgenstein (1921/2001) defends in the *Tractatus* is the theory he repudiates in the *Philosophical Investigations* (Wittgenstein, 1958), and at its heart is the thesis that thinking is mental representing. Our aim here is to say precisely what it is about the claim that thoughts are mental representations (and that thinking is mental representing) that runs counter to the direction in which Wittgenstein's later arguments lead. This, we believe, is not a direction cognitive scientists should be reluctant to travel in; for it takes us further and further away from Cartesian conceptions of mentality that can only hamstring research.

Overview of the argument and some preliminary points.

The argument we present here starts with the premise that to think that something is so is to perform combinatorial operations on representations, and it moves to the negative conclusion that to think that something is so cannot be to perform combinatorial operations on *mental* representations. This should not be taken to imply that thinking is a combinatorial operation on non-mental representations like words and sentences. Certain cases of thinking may be that[1], but there is nothing in the argument which entails the strong view that to think is to speak silently to oneself in a natural language. The aim is only to rule out a widely held view about what thinking is, not to defend a competing one. The view that is ruled out is that for Sam to judge that ducks run is for her to perform an operation on representations of a radically different kind than those on which she would operate were she to make a corresponding assertion. On such a view when she asserts that ducks run, operations are performed on two sorts of representation, one mental one not; and when she judges that ducks run but asserts nothing they are performed only on mental representations. On this sort of view internal mental representing need not be connected to external spoken representing in any way; so what a creature can say sets no limits on what it can think.

If the argument of this paper is successful we will have shown that it can't be because linguistic behaviour is linked with hidden events and processes that it counts as linguistic. This is liable to upset very widely held assumptions both about the nature of thinking and about the shape of cognitive scientific explanations. On these assumptions, what is important about Sam's assertion that ducks run is how it stands to various items or events in, or states of, her brain. Illuminating explanations will consist in claims about this standing. That implies a certain conception of what makes assertions into meaningful speech acts and differentiates them from grunts and squawks. On it, for Sam

[1] Bennett & Hacker (2003) contains an excellent series of reminders of how many different things we ordinarily call thinking.

to assert that ducks run is for her to make known the occurrence of a certain sort of mental act – a judgement, and for her to judge that ducks run is for certain syntactical operations to be performed (by Sam, or perhaps by a mind, a brain, or a brain-part) on certain information bearing states or structures (mental representations), which are themselves states of, or items/structures in, brains (or are realised in such states, items or structures). If that is what assertions are then linguistic behaviour must be an incidental outward accompaniment of mental representation. It is only because linguistic behaviour serves to indicate the occurrence of internal processes that it counts as *linguistic* (the parrot's squawks may sound like bits of linguistic behaviour but they aren't), but mental representations counts as mental representations whether or not they happen to be indicated by observable behaviour. On this view – which is profoundly Cartesian - the link with mental representation is an essential feature of linguistic behaviour, but the link with behaviour is an inessential feature of mental representation.

This conception of the relation between thought and linguistic behaviour could easily seem an essential feature of a cognitivist account of language.[2] The aim of cognitive science is precisely to explain intelligent behaviour by appeal to internal processes, and the distinction between speaking and thinking looks like a kind of paradigm of the distinction between intelligent behaviour and internal process (or processing). If we accept this appearance at face value, and if we identify mental representation with thinking, it will seem that a theory of mental representation will perfectly conform to the basic aim of cognitive science by providing an explanation of linguistic behaviour by appeal to internal goings on.

These appearances are deceptive. On our view, cognitive science can retain its commitment to explain behaviour by appeal to inner processes, events and states, whilst abandoning the Cartesian project of explaining linguistic behaviour by appeal to thought, and giving up the identification of thinking with mental representation. We take it that a cognitivist account of linguistic behaviour should take the form of an attempt to model and identify the states of and processes in (and around) the brain that are causally necessary conditions of linguistic behaviour. To identify such states and processes would be to provide an explanation of behaviour by appeal to obscure inner states and processes; so it would conform to the cognitivist brief. But to discover the causally necessary enabling conditions of linguistic behaviour is not to discover what thoughts are and what thinking is – for thoughts are expressed in behaviour, but a causal condition is not expressed in what it causally enables (see Trigg & Kalish 2010 (submitted).

[2] E.g., Hauser, Chomsky & Fitch (2002) say, "In the varieties of modern linguistics that concern us here, the term "language" is used quite differently to refer to an internal component of the mind/brain (sometimes called "internal language" or "I-language"). We assume that this is the primary object of interest for the study of the evolution and function of the language faculty." p1569

It is all too easy to confuse thinking, which speaking and writing is often expressive of, with the brain events and processes that are causally necessary conditions for intelligent linguistic behaviour. A Cartesian conception of the relation between thought and its linguistic expression makes such confusion almost inevitable, for in characterising thought as an inner accompaniment to observable linguistic behaviour it lumps it in with the various neural events and processes which are causally necessary for linguistic behaviour. But, Wittgenstein reminds us, intelligent speech does not consist in a series of observable movements of the face and throat on the one hand, and a series of hidden mental events and processes on the other. Rather, thought is *in* intelligent speech in roughly the way that distress is in an anguished cry and amusement is in a spontaneous peel of laughter. So, contrary to the apparently innocent Cartesian intuition, a person's thoughts are typically not hidden behind their words, but precisely revealed by them. If we remember this we will not so easily seek to identify thinking with the neural conditions of speech, for it is obvious that these conditions are not revealed, as a person's thoughts typically are, by the things they say. Of course these neural conditions are not hidden behind the linguistic behaviour they causally enable in the dramatic way Cartesian thoughts are supposed to be hidden behind some of the noises people make, for they count as hidden only because they are in the thinker's skull, not because they are in the thinker's mind.

Cognitive science, on our view, has the job of framing and testing hypotheses about the causally necessary enabling conditions of familiar psychological phenomena like thinking, imagining, remembering and willing. As long as these familiar phenomena are not conceived as private, inner accompaniments to observable behaviour and then *identified* with unfamiliar operations on mental representations, cognitive science can make a substantial contribution to our understanding of them. There is even a sense in which it might be appropriate to characterise neurological states, structures, events and processes which causally enable familiar phenomena like assertoric thinking (judgement) as mental representations, since without them it would be impossible for people to represent the world in the way that they do.

Initial Clarifications – Mental representation and the language of thought

If one holds that cognition is mental representation, and that mental representation consists in combinatorial operations performed on mental representations, then one seems compelled to accept that cognition is a quasi-linguistic operation. The basic idea here is that to combine representations to form complex ones is to perform *syntactic* operations analogous to those involved in combining words and sentences to form more complex sentences. (Fodor, 1975, 2008) In both cases, it may be supposed, the representational properties of complex representations will depend in a systematic way on the representational

properties of the simpler ones. Thus a finite stock of elemental representations together with a finite number of syntactic rules for the combination of these, will yield an unlimited number of possible complex representations. This picture, which is surely very widely accepted (see, e.g., Schneider, 2010), seems to require what Fodor famously called a 'language of thought'. The thesis that mental representation, like spoken and written representation, is linguistic, still plays an important, though poorly defined role in cognitive science. It is not hard to see why it is so popular. For one thing it seems to provide a way of explaining the unlimited number of different intelligent things intelligent creatures can do in reference to an equally unlimited number of different representational states their minds or brains can be in. For another it seems to support a computational theory of mind or cognition, since the relevant combinatorial or syntactic operations seem well suited to be conceived as computational operations of a kind that might be run by something like an organic equivalent to a computer.

Whilst the hypothesis that there is a language of mental representation consisting, not of words and sentences but of mental representations is a natural way to develop the concept of mental representation, it is no more than that. It could be discarded and the notion of mental representation retained. So it is important to appreciate that the argument we present is directed at the very idea that there are two ways to represent, for example, ducks as running, one mental, the other not. If it tells against this idea it will, as a matter of course, also tell against formulations of it on which to think that ducks run is to make an assertion in the (or a) silent language of thought.

The Mental Representation Argument

On a picture of the relation between thought and language that deserves to be called the classical picture the purpose of speech and writing is to make thought, which is essentially private and psychological, public and perceptible. The basic problem with this picture, however its details are worked out, is this: if, when Paul says 'Ducks run.' there are two types of operation he is performing (one mental, inner and private, one physical, outer and public), and two kinds of item he is operating on (one mental, inner and private, one physical, outer and public), we will have to explain what the relation is between these operations and these items. That is, we will have to explain how, on the one hand, the combining of mental representations stands to the combining of words; and on the other, how the combined mental representations stand to the combined words. It should be clear, at least with a little reflection, that the prospects of providing such an explanation look dim. If we conceive mental representations as components of a non-verbal language we will have to explain how non-verbal thoughts are to be translated into verbal utterances; if we conceive mental representations as mental images or pictures we will have to explain what it is to translate images into words. But such explanations are hopeless for at least three sorts of reason. There is no such thing as translating what is said in a mental language into something said in a public one, so no such thing as doing it correctly or incorrectly (try to compare your translation of a mental representation with the mental representation it is a translation of so as to check if you've translated it correctly). Neither is there any such thing as translating – as opposed to describing – an image. Finally, since any image can be described in indefinitely many equally faithful ways, images do not determine what does and doesn't count as faithfully describing them. So being able to say what one is thinking cannot be a matter of being able to describe a certain mental image correctly; since any description of a particular image could count as a correct description of it, talk of the (or even of a) correct description of an image is empty.[3]

If these weighty considerations do not convince thus baldly presented, we can turn to Wittgenstein's celebrated (misunderstood and neglected) private language argument to drive them home. This succeeds in showing, quite categorically, that when, e.g., Paul says 'Ducks run.' he does not perform two types of operation, one on mental representations and one on verbal representations, but only one. It shows this by showing that operating on representations must be a normative or rule-governed affair (the type of thing that can be done incorrectly or correctly) and that operations performed on items, events or states available only to the operator could not be a normative or rule-governed affair. According to our argument if one assumes that thinking is a combinatorial operation on representations one has to deny that thinking is mental representing; that is interesting, for it is precisely that assumption that has led so many cognitive scientists to conclude that it must be mental representing. What follows then, is an argument to the conclusion that thinking cannot be an operation on mental representations which exploits Wittgenstein's argument that language cannot be private:

To think, in the sense under discussion, is to think that something or other is so. To do that is to have a thought or to make a judgement, such that the thought one has or the judgement one makes will be the thought or judgement that things are thus and so. If things are that way the thought one has will be true, if not, then false.

For the purposes of argument let us assume that to think (or to judge, doubt or suppose) that something is so is to perform a certain kind of operation on a certain kind of representation. This operation must be productive of further representations that have the characteristic of being evaluable for truth; so it must be an operation on representations (that may or may not be of a truth-evaluable type) that yields representations (that are of a truth-evaluable type). So, for example, it could be an operation on words or sentences that produces sentences. Such operations must be combinatorial. What other than a combinatorial operation could produce the type of representation that is

[3] See e.g. G. McCulloch (1989). pp. 152-163.

evaluable for truth out of representations that are, very often, not evaluable for truth; and what other than a combinatorial operation could generate indefinitely many truth-evaluable representations out of a few non-truth evaluable ones?

Not every possible combination of representations will yield representations capable of truth. For example 'Green was a depressing silently were or jam' is a combination of representations, but it is nonsense, so the question of its truth cannot arise. There are many different ways of combining representations in such a way as to produce nonsensical representations not capable of truth. To combine representations in at least some of these ways that produce nonsense is to combine representations incorrectly. What can be done correctly or incorrectly must be a rule-governed activity. That is to say that for representations to be combined in the relevant way is not just for a series of orderly events to occur, but for a rule-governed activity to be engaged in. When a given process does not unfold as it usually does we can say that an irregularity has occurred in it but not that a mistake has been made in carrying it out. But many nonsensical combinations of representations are not just irregular; they are wrong. That entails that they run counter to rules of representation combination that do not merely capture actual regularities exhibited by representation combining activities but prescribe how those activities should be carried out.

Now if a given representation is knowable as the representation it is only by whatever it is that performs combinatorial operations on it, the rules determining how it may be combined with other representations must be private rules. There can't be public rules that determine how private items, events or states should be manipulated, because it would be impossible to assess putative observations of such rules for correctness. If the rule is, "Perform operation *p* when *y*-type items appear, or *x*-type events occur, or *r*-type states are actualised," and if *y*-type items, *x*-type events and *r*-type states are knowable only to the performer of *p*, then performances of *p* cannot be publically checked or assessed for correctness.[4] So they can be checked for correctness only privately.

Now it seems that *mental* representations must precisely be representations knowable as what they are only by whichever thinker or representer is operating on them. To say that a given representation is a *mental* representation is to say that the dealings thinkers have with it are not perceptual. For a mental representation to be available to a thinker need not require that a certain publically available mark be seen or sound heard, but only that thinking be going on. Thinking is conceived here precisely in contrast to observable behaviour, as an inner or psychological operation. Such an operation must be an operation on items, structures or states knowable as the representations they are only to the relevant operator.

If that is right, and if there can't be public rules for the combination of private representations, then the rules that determine what it is to combine private mental representations correctly must be private rules.

Having reached this result, it only remains to be established that there can be no private rules for the combination of representations (or anything else), and it will have been shown that thinking cannot be a combinatorial operation on mental representations. Wittgenstein showed how to establish exactly that.

There can be no private rules governing the operations performed on mental representations because there can be no difference between its seeming to a thinker or representer on a given occasion that they are following a private rule and their really following a private rule on that occasion. Say Sam's putative rule, *p,* is – "Whenever an item relevantly similar to *this* one (pointing inwardly to a relevant sample) comes before my mind, I will perform operation *r* on it (or whenever I am in *this* sort of state – pointing inwardly to an appropriate state – perform operation *r* on it)". In these sorts of case there could be no difference between its seeming to Sam at *t* that the relevant item was before her mind, or that she was in the relevant state, and that item really being before her mind, or her really being in that state. In that case, Sam cannot have invented a rule *p* governing his performance of operation *r*, because any future performance that seems to Sam to be a performance of *r* in accordance with *p*, will thereby be a performance of *r* in accordance with *p*. In situations that do not allow for a distinction between what seems justified and what is, talk of being justified or unjustified is out of place.

To conclude, if thinking is a combinatorial operation on representations it must be a rule-governed combinatorial operation on representations; but now, since private rules for the combination of representations are impossible, and rules for the combination of *mental* representations would have to be private, thinking cannot be a combinatorial operation on *mental* representations. This argument shows that there is a fundamental conflict between the idea that thinking is an activity subject to normative constraint, and the idea that thinking is a private psychological affair: if thinking can be done incorrectly it cannot consist in manipulations of private mental representations.

Here is a concise formulation of the argument just given.

1. To think – in the relevant sense – is not just to think of something but to think that something is the case.
2. To think that something is the case is to combine representations that may or may not be the sort of representations capable of truth, so as to produce representations that are the sort of representations capable of truth.
3. It is possible to combine representations so as to produce representations that are not capable of truth as well as those that are.
4. To combine representations in such a way as to produce representations not capable of truth is to combine representations incorrectly.

[4] See Wittgenstein (1958) e.g. section 258.

5. If something, *p*, can be done incorrectly there must be rules that determine what counts as doing *p* correctly; *p* must be a rule-governed activity.
6. By 2, 3, 4 and 5, thinking must be a rule-governed activity. (Not just a process exhibiting regularities).
7. Rules for the combination of representations knowable as what they are by only one representer would have to be private rules.
8. Representations that are *mental* must be knowable as what they are only by one representer.
9. So – by 7 and 8 – rules for the combination of mental representations must be private rules.
10. There can be no private rules.
11. The combination of *mental* representations – by 7, 8, 9 and 10 – cannot be a rule-governed activity.
12. Thinking – by 6 - must be a rule-governed activity.
13. Thinking – by 11 and 12 – cannot be a combinatorial operation on *mental* representations.

An Objection

Without further ado, let us consider an objection to this argument. It concerns premise eight - the claim that mental representations, *qua mental*, must be private, that is, must be knowable as the representations they are only by whatever it is that represents by operating on them. Many cognitive scientists and philosophers of mind might eagerly reject this premise on the ground that it depends upon an unacceptably Cartesian notion of what mental representations are. It will be said that if we reject this out-dated Cartesianism, and think of the mental representations in question as states of, or items in, brains, (or as realised in such items or states), we can deny that they are private, and so allow room for the idea that there could be public rules for their combination.

The problem with this objection is that it is inconsistent with the claim that thinking is *mental* representing, so can't be used to defend it. The claim that thinking is mental representing depends on a certain way of conceiving the distinction between representing done in thought and representing done in (public) language. On this conception representing done in thought (mental representing) is what makes representing done in public language what it is; it is because Sam's assertions do, but his sneezes do not, depend somehow on his thoughts, that his assertions count as meaningful utterances rather than mere noises. It turns out, we will now argue, that this conception allows for the possibility that a representer may be wrong about the *assertion* they are making, but it excludes the possibility that a representer may be wrong about the *judgement* they are making (the thought they are having). We argue that rejection of premise 8 is incompatible with the Cartesian view that a thinker cannot be wrong about which judgement they are making, and that this Cartesian view is an essential feature of the position rejection of this premise is meant to defend. Thus, whilst the anti-Cartesian feel of the objection now under discussion may be congenial in itself, it is quite inconsistent with the conception of the relation between thought, mental representation and language that it is meant to defend.

Speakers are sometimes wrong about what they are saying. This is not to say that speakers are sometimes insincere; it is to say that speakers can think they are saying one thing when they are really saying another. How shall we explain the possibility of this sort of mistake?

If we hold that to think is mentally to represent, and that thoughts are mental representations we will have to hold that to make a sincere assertion is to translate or otherwise convert a mental representation into a non-mental one. This commits defenders of this conception of thinking to the view that for a speaker to be wrong about what they are saying is for a speaker to be wrong about the relation between what they are saying and what they are thinking. Mistakes of that kind are mistakes made in translating or converting mental into non-mental representations. Now, on this conception of how it is possible for a speaker to be wrong about what they are saying, no conceptual room is left for the possibility that they could be wrong about what they are thinking. This is because, on this view, whilst a speaker has to convert or translate their thoughts into sentences in order to make an assertion, they do not have to translate or convert their thoughts into anything in order to have them. If one can be wrong about what one is saying because one can translate or convert one's thoughts into words incorrectly, and one does not have to translate or convert one's thoughts into anything in order to have them, one cannot be wrong about what thought one is having at a given time.

To see this, consider the following argument. If we explain what it is for someone to make a mistake about what they are saying by appeal to the idea that they can be wrong about the relation between what they are saying and what they are thinking, we will have to deny that they can be wrong about what they are thinking, on pain of generating an infinite regress. For if it were possible for a thinker, Sam, to be wrong at *t* about what she is thinking at *t*, that possibility would require explanation. Any such explanation would have to appeal to a difference between what Sam is thinking at *t* and what she thinks she is thinking at *t* – we will have to say that whilst she thinks she is thinking one thing she is really thinking another. But as soon as we say that, we will also have to allow the possibility that there can be a further difference between what Sam thinks she is thinking at *t* and what she thinks she thinks she is thinking at *t*, and so on. This regress is by no means benign, for it requires a thinker to have an infinite number of appropriate thoughts at *t* if they are to know what they are thinking at *t*. As soon as we open an anti-Cartesian gap between what Sam is thinking at *t* and what she thinks she is thinking at *t*, for her to know what she is thinking at *t* it will not be enough that she thinks it at *t*. Suppose she thinks that ducks run at *t*. If she is to know that she is thinking that ducks run at *t*, she has to think that she is thinking that ducks run at *t*, and if she is to know that, she has to think that she is thinking that she thinks that ducks run at *t*, and so on.

So if we are to avoid this regress we have to embrace Cartesianism; we have to say, that is, that Henry's knowledge of what he is thinking is both incorrigible and evident; incorrigible because if he thinks he is thinking that *p* he is thinking that *p*, and evident because if he thinks that *p* he thinks that he thinks that *p*. If we take thoughts to be mental representations and assertions to be translations of mental representations into perceptible signs, we commit ourselves to an explanation of how a speaker can be wrong about what they are saying which only Cartesianism will save from incoherence.

Now of course the relevant point is that this Cartesian account of the relation between a thinker and the thoughts they have is flatly incompatible with the proposal that mental representations are not private. To see this, it is important to appreciate first that to say that a given mental representation is publically available is not just to say that it is identical with certain brain states that are publically available. It is conceivable that Sam should be acquainted with a certain brain state, *p*, which is in fact identical to a certain mental representation *r*, but know neither that *p* is a mental representation nor that *p* is mental representation *r*. (If I know the butcher, and the butcher is the president, then I know the president, but I may not know that the butcher is the president). So what is required is that George and Harry can come to know that Grace is thinking that *p* by becoming acquainted with a certain state of Grace's brain.

Now that possibility is rather dramatically incompatible with the Cartesian conception of thinking to which we have just shown our opponent to be committed. It makes Grace's way of finding out what thought she is currently having into just one of many ways of finding that out. So a situation will be conceivable in which Grace tries to find out what she is thinking using her introspective method, George and Henry do the same by observing her brain, and Grace fails whilst George and Henry succeed. If that is thinkable, then not only could Grace be wrong, and George right, about what Grace is thinking at any time, but Grace could be wrong and George right about what she is thinking at all times!

The problem remember is not just that these possibilities are absurd in themselves – although the idea that something could count as a thinker whilst *always* being wrong about what it thought is pretty unsatisfying all on its own – but that they are incompatible with the Cartesian conception of the relation between thought and language to which our opponent is committed.

So the thesis that to think is to operate on mental representations cannot be defended by rejection of premise 8. If mental representations are constitutive of thoughts they must be private, and if mental representations are private combining them cannot be a rule-governed operation.

Conclusion

We take it that the mental representation argument shows that thinking cannot be mental representation, that is, that it cannot be a combinatorial operation on mental representations. While there are objections to this argument that we have not explicitly considered here, they will have to turn either on the denial that thinking can be done incorrectly or on the claim that there can be private rules, and these responses seem to head off in unpromising directions. The only plausible option open to the representationalist is to conceive thinking as a combinatorial operation on the representations constitutive of a natural language like English (Malcolm,1973). We have said nothing either for or against that position here – though it is perhaps worth noting that anyone attracted to it will have to hold that it is persons as we ordinarily conceive them and not minds or brains that think (since it is indubitably human beings and not minds or brains that know how to use the words of a natural language).

So thinking is not mental representing, and, for example, asserting is not converting or translating mental into non-mental representations. Does this result show that a cognitive scientific account of linguistic behaviour is impossible? Not at all. It shows that if we identify mental representing with thinking we cannot explain intelligent linguistic behaviour by appeal to mental representing. But if we think of mental representing not as identical to thinking but as a causally necessary condition on it (Trigg & Kalish, 2010), then the idea that mental representation underlies intelligent linguistic behaviour is in good shape. It is, of course, profoundly plausible that if certain very complex events did not take place in a person's brain at *t* they would not be able to think or speak at *t*. Nothing in the argument just presented is incompatible with this idea, and nothing suggests that the difficult business of finding out about these events is not a scientific undertaking of the greatest interest.

References

Bennett, M. & Hacker, P.M.S. (2003). *The Philosophical Foundations of Neuroscience.* Oxford, Blackwell.

Hauser, M., Chomsky, N., & Fitch, T. (2002). The faculty of language: What is it, who has it, and how did it evolve? *Science, 298,* 1569-1579.

Fodor, J. (1975). *The Language of Thought.* Harvard, Harvard University Press.

Fodor, J. (2008). *LOT2: The Language of Thought Revisited.* U.S.A: Oxford University Press.

Malcolm, N (1973). 'Thoughtless brutes'. *The Proceedings and Addresses of the North American Philosophical Society.* 46, 5-20.

Schneider, S. (2010). The language of thought. In J. Symons & P. Calvo (Eds.) *Routledge Companion to Philosophy o f Psychology*. London, Routledge.

Trigg, J. & Kalish, M. (2010). Explaining how the mind works: on the relation between cognitive science and philosophy. Manuscript submitted for publication.

Wittgenstein, L. (1921/2001). *Tractatus Logico-Philosophicus*, London, Routledge.

Wittgenstein, L. (1958). *Philosophical Investigations*, Oxford, Blackwell.

Wittgenstein, L. (1974). *Philosophical Grammar.* ed. R. Rhees, trans. A.J.P. Kenny. Blackwell: Oxford.

What Is Domain Specificity (and Why Does It Matter)?

Muhammad Ali Khalidi (khalidi@yorku.ca)
Department of Philosophy and Cognitive Science Program,
York University, 4700 Keele Street
Toronto, ON M3J 1P3 Canada

Abstract

The distinction between domain specificity and domain generality is widespread in cognitive science. Yet, the difference between the two types of cognitive capacities has rarely been made in a principled manner. Moreover, some of the examples that are put forward to illustrate it in the literature are either spurious or misleading. In this paper, I use a number of examples to determine what domain specificity is, and just as importantly, what it is not. A domain-specific cognitive system is one that is in principle generalizable, but which the cognizer does not extend to cases that the system did not originally evolve to deal with.

Keywords: domain specificity; modularity; innateness; animal cognition.

Introduction

There are many contexts in cognitive science in which it is useful to distinguish domain-specific cognitive capacities from domain-general ones. For instance, according to many evolutionary psychologists, human cognition consists largely of domain-specific cognitive capacities, and this feature of our cognitive makeup provides evidentiary support for the pervasive influence of evolutionary processes on the formation of the human mind (Carey & Spelke 1994, Cosmides & Tooby 1994). By contrast, according to other researchers, domain-general cognitive abilities are the norm in the human mind and are what distinguish human cognition from that of most other animals (Samuels 1998, Fodor 2000). Yet, the distinction between the two kinds of capacities is not easily drawn and it is often drawn erroneously by the researchers who aim to make it. After identifying some misleading examples of domain specificity, I will put forward some better examples of the phenomenon drawn from the literature. Then, I will use these examples to try and spell out a more satisfactory account of the phenomenon of domain specificity. Finally, I will draw on this new understanding to try to shed light on the significance of this concept and its importance for cognitive science.

Domain Specificity and Its Confounds

Domain specificity is a feature of cognitive capacities that is often associated with several other such features, notably: modularity, innateness, and brain localization. In this section, I will examine the connections that may or may not exist between domain specificity and these other features, in order to gain a preliminary understanding of the concept of domain specificity.

By virtue of the way in which modularity was initially defined by Fodor (1983), there is a strong link between modularity and domain specificity. Indeed, it follows from Fodor's account that domain specificity is one of the defining features of modularity, and therefore that all modular cognitive capacities are domain-specific.[1] Of course, a case might be made for rejecting this definition on the grounds that it is unwarranted by the empirical facts or otherwise detrimental to cognitivist research, but I will not try to make the case, nor do I think that the case ought to be made. I will simply follow Fodor and much subsequent theorizing in accepting it. Hence, I take it as uncontroversial that domain specificity is a necessary feature of modularity, though the two features ought not to be conflated, since the former is subsumed by the latter.

The case is more complicated when it comes to innateness. Although there is also a widespread assumption that there is a link between innateness and domain-specificity, there is no convincing reason for inferring such a link. It is neither the case that all innate cognitive capacities are domain-specific, nor that all domain-specific cognitive capacities are innate. To illustrate, human beings may have an innate cognitive capacity for associative learning that may be entirely domain-general. Conversely, there may be certain domain-specific cognitive abilities that are not innate but mainly learned, such as chess-playing ability. However, despite the lack of a direct link between the two features of cognition (innateness and domain specificity), a case could be made for an indirect link between them. It has been argued that when it comes to domain-specific abilities, it is easier to tell whether and to what extent they are innate or not (Khalidi 2001). In other words, the link between the two features is epistemic or evidential rather than intrinsic, since we can more easily gauge the amount of explicit learning or relevant experience in the case of domain-specific cognitive capacities than in the case of domain-general ones. This is so because it is easier to rule out relevant sources of information in the former case than in the latter.

As for the link between domain specificity and brain localization, this is also widely made, as is the link between

[1] In addition to being domain-specific, according to Fodor (1983) modular cognitive capacities are supposed to: 2) process items automatically and in a mandatory manner, 3) be inaccessible to consciousness, 4) be fast, 5) be cognitively impenetrable (e.g. resistant to being unlearned), 6) process "shallow" or highly salient features, 7) have fixed neural architecture, 8) have specific breakdown patterns (as in aphasia, agnosia), and 9) have fixed ontogeny (standard pace and sequence of development).

brain localization and modularity (which, as seen above, subsumes domain specificity). However, there does not seem to be a cogent reason for making either link. For instance, there are good grounds for thinking that various cognitive capacities are modular (and hence domain-specific) even though they are not localized in one region of the brain, indeed even though they are scattered across a range of brain regions. Modularity and domain-specificity pertain largely to the functioning of a cognitive capacity rather than its neural manifestation, so there is limited scope for inferring brain localization from either of these phenonema.

In distinguishing domain specificity from other features of human cognition, I have relied on an implicit preliminary understanding of the phenomenon. As the term implies, what it is for a cognitive capacity to be domain-specific is for it to pertain to a single domain or to a restricted range of domains, and more importantly, for it not to be generalizable to other domains. Moreover, this last proviso highlights the importance of restricting domain specificity to aspects of cognition that are in principle *generalizable* across domains, although they are not in fact *generalized*. In other words, it would be vacuous to describe as domain-specific some cognitive system that is not even in principle generalizable. For example, a body of information pertaining to some domain or another is not generalizable, since its subject matter is in principle restricted to a certain domain. By contrast, a rule that is deployed by a cognizer in one domain but that *could* be deployed in another domain is in principle generalizable. Hence, rather than domain-specific capacities it may be more fruitful to talk about domain-specific *rules* or *principles*. These are the kinds of cognitive structures that can be generalized across domains and that are therefore candidates for being domain specific, though they may not be the only cognitive structures that can be so generalized. This point will be developed further in subsequent sections.

One issue that might be raised here concerns the nature and scope of a domain. If we do not explicitly specify what a domain consists in, we may open ourselves to the following objection. Suppose it is claimed that some linguistic rule *R*, which is part of a subject's cognitive repertoire, is domain-specific since it only pertains to the domain of language. Now suppose someone objects to this on the grounds that language itself is not a single domain but rather that it comprises various domains, say syntax, semantics, pragmatics, and so on. Hence, this objector might contend that the rule is in fact a domain-general one since it ranges over several domains. This example suggests that it may always be open for someone to claim that a putative domain-specific rule is in fact domain-general unless we have some principled way of delimiting the scope of a domain. How could we define domain specificity without specifying the scope of a domain? In response, I would argue that it is not enough to show that a rule does in fact apply to what are supposed to be different domains. Rather, to demonstrate true domain generality with respect to some rule *R* used by some cognizer *C*, it should be shown that *C* has the ability to deploy *R* in some previously unencountered situation or to apply it to some new cases. In practice, it does not seem to pose much of a threat to our account of domain specificity if we do not have a principled way of distinguishing domains. Typically, it will suffice to show that a cognitive mechanism is truly domain-general if we can show that it can be deployed in unfamiliar contexts that are new to the cognizer. Conversely, it suffices to show that a cognitive mechanism is domain-specific if we can demonstrate that when one attempts to apply it to new cases, the mechanism either does not function or gives systematically erroneous answers. This point will be justified further in due course in discussing what constitutes a new case or a new domain.

Preliminary Examples

Spurious examples of domain specificity, or genuine examples of domain specificity that are misleadingly described, are not difficult to find in cognitive science. Two such instances stand out in Cosmides and Tooby (1994), in the context of an argument that domain-specific cognitive mechanisms are more efficient and adaptive that domain-general ones. Cosmides and Tooby (1994, 90) state: "A woman who used the same taste preference mechanisms in choosing a mate that she used to choose nutritious foods would choose a very strange mate indeed, and such a design would rapidly select itself out." Clearly, something has gone wrong here: a rule that chooses mates on the same basis as foods is not a domain-general rule but one that is based on an error. If we have a cognitive mechanism that enables us to recognize edible foods, it should only be sensitive to perceptual stimuli that are plausible candidates for food. Otherwise, it is not a genuine food-preference cognitive mechanism. Hence, this is not a plausible example of a domain-general mechanism being inferior to a domain-specific one. In describing the example, Cosmides and Tooby seem to be trading on the ambiguity of the term "taste", which can either refer to gustatory discrimination or a broader notion of discrimination, which could comprise mate selection.

A case of domain specificity that is misleadingly described can be found in another example given by Cosmides and Tooby (1994). Although the following example is a genuine case of domain specificity, it is improperly contrasted with a putative case of domain generality. By examining it, we can get a better handle on the phenomenon of domain specificity. Cosmides and Tooby explicate the well-known example of the alarm calls of vervet monkeys, who give three different calls in response to three different kinds of predators (leopard, eagle, and snake), leading conspecifics to take three different types of evasive action (respectively, climbing a tree, looking up or diving into bushes, and standing on hind legs and looking into the grass). In this case, they state: "A single, general-purpose alarm call (and response system) would be less effective because the recipients of the call

would not know which of the three different and incompatible evasive actions to take" (Cosmides & Tooby 1994, 89-90). The problem here is not there could not be a general-purpose alarm system; there clearly could. But a general-purpose alarm system is not one that would issue the same call for every predator. That would be a system that fails to discriminate among different stimuli. Rather, an all-purpose alarm system would be one akin to the human linguistic alarm system, which issues a different linguistic warning in the case of different predators. There are clearly certain advantages to such a system, since it is capable of handling a much wider range of predators ("Lion!", "Hawk!", "Stampede of buffalo!") and of being made more precise in various ways ("Tiger to the right", "Hyena to the northwest", "Human with weapon right behind you", etc.). However, it may also involve certain disadvantages, since given the diversity of inputs and outputs, it may take more processing time to issue the correct alarm, there may be more opportunity for error in both transmission and reception, and the evasive action involved may have to be figured out from scratch by the respondent once the alarm is sounded. Determining which of these two alarm systems, the domain-specific vervet system or the domain-general human system, is more efficient and adaptive is not an easy matter. It will clearly depend on various contingencies such as the nature of the environment, and Cosmides and Tooby may ultimately be right that in certain circumstances a domain-specific system may be superior to a domain-general one. But they have not made the appropriate contrast between domain specificity and domain generality.

This example is instructive since, once modified in the way that I have just done, it provides a fairly clear contrast between a domain-specific and a domain-general cognitive mechanism. In the following section I will attempt to give a more satisfactory formulation of domain specificity and illustrate it with better examples drawn from the literature.

Domain Specificity Revised

To refine our understanding of domain specificity, it is useful to build on the example mentioned in the previous section and attend to its instructive features. The first feature that can be gleaned from the vervet monkey alarm call system is that a cognitive mechanism for alarm calls is at least in principle generalizable. That is to say, even though the vervet alarms are only issued for a small set of specific predators, it is easy to conceive of an alarm system that would extend to other predators. Hence, it seems safe to conclude that for one to speak meaningfully of a domain-specific cognitive system, that system must have the following feature:

1) A *domain-specific* cognitive system is one that is in principle generalizable to new domains.

This condition may appear vacuous, but it is designed to rule out cognitive systems that consist of a "database" rather than rules or principles, as mentioned above. Domain specificity, to be meaningful, must be a feature of the cognitive capacity rather than a feature of the subject matter. Though this point may seem obvious, the attribute of domain specificity is often conferred on bodies of knowledge possessed by subjects that are not obviously generalizable, such as knowledge of animate as opposed to inanimate domains (e.g. Caramazza & Shelton 1998).

The first proposed feature of domain specificity makes reference to "new domains," which is a notion that needs further explication and justification. The new domain involved need not be what we might regard as an entirely disparate area of inquiry. In the case of the vervets, the original domain is something like: predators commonly encountered by vervets in the wild. It is *in principle* generalizable to include the new domain: all predators, or even, all threats. The vervet alarm system is domain-specific because it fails to generalize to these new stimuli. But these new stimuli do not, strictly speaking, have to be drawn from what we would normally consider to be another domain, such as a new sensory modality or a new area of inquiry. At this point, it might be asked, by virtue of what are they to be considered genuinely new stimuli? They must at least be stimuli that the cognizer has not encountered before. But that condition is surely too weak, since the domain-specific vervet alarm system clearly generalizes to new exemplars of leopards, eagles, and snakes, which the individual has not encountered before, indeed ones which perhaps no vervet monkey has encountered before. Rather, in this context, new stimuli are ones that the system was not originally designed to cope with. This is admittedly a vague formulation and brings in thorny evolutionary considerations concerning the *proper function* of an evolved trait (Millikan 1989, Neander 1991). Though it is not always easy to determine what the proper function of a cognitive system is, some reference to it seems inevitable, since cognitive systems have evolved to fulfill a certain purpose and their generalizability consists in part in being able to extend beyond that purpose to cases that they were not designed to cope with, or ones that are not normally encountered in the environment in which the system evolved.[2] Hence, I propose that the second crucial feature of a domain-specific cognitive system is as follows:

2) A *domain-specific* cognitive system is one that systematically fails to yield a correct result in the case of stimuli that the system did not evolve to deal with.

The aptness of this second feature can be further justified by reflecting on appropriate examples from the literature. One such case is provided by Cheney and Seyfarth (1985, 197), who describe the domain-specificity of certain cognitive capacities in vervet monkeys, as follows: "Within the social group, the behavior of monkeys suggests an understanding of causality, transitive inference, and the notion of reciprocity. Despite frequent opportunity and often

[2] A similar conclusion has been reached by Boyer and Barrett (2005, 98), who write: "The domain of operation of the system is best circumscribed by evolutionary considerations."

strong selective pressure, however, comparable behavior does not readily emerge in dealings with other animal species or with inanimate objects." In this example, both features outlined above are clearly in evidence. First, the principle of causality and the rule of transitivity clearly have application outside the realm of social interaction with conspecifics. The transitivity rule can be used to infer hierarchy relations among monkeys (if *A* ranks higher than *B* and *B* ranks higher than *C*, then *A* ranks higher than *C*) but it can also be used to infer information about size, quantity, and other matters (if object *A* is larger than object *B* and *B* is larger than *C*, then *A* is larger than *C*). However, despite the clear applicability of this rule to domains that go beyond social interactions with conspecifics, Cheney and Seyfarth claim that vervets do not so apply the rule. Second, it is clear that vervets do not use the rule of transitivity on other species or inanimate objects simply because they evolved the rule to deal with the restricted domain of social interaction with conspecifics, which may have been a more pressing adaptive problem. In this case, it may seem obvious that interactions with other animal species and with inanimate objects constitute genuinely new domains. There may not appear to be a need to use the second feature of domain specificity to justify the judgment that it does not generalize to genuinely new domains. But even though it may not play this role in this case, there are other cases in which the second condition is necessary to enable us to make a judgment concerning domain specificity.

Some recent studies of cognition in non-human animals have debated whether animals have the cognitive capacity to teach others. Thornton and McAuliffe (2006) demonstrate that mature meerkats provide young conspecifics with specimens of their usual prey, scorpions, that are either dead, disabled (stings removed), or intact. Which of these three types of scorpion is provided depends on the perceived age of the young meerkat. Younger pups are provided with disabled but alive scorpions and this provides them with the opportunity to learn how to kill the scorpion without being exposed to the possibility of a sting. Thornton and MacAuliffe (2006, 228) conclude that "the provisioning behavior of meerkat helpers constitutes a form of 'opportunity teaching' in which teachers provide pupils with opportunities to practice skills, thus facilitating learning." Moreover, they support their findings by relying on a definition of teaching derived from Caro and Hauser (1992), according to which an individual is a teacher if it modifies its behavior in the presence of a naïve observer, at some cost to itself, and as a result allows the observer to acquire knowledge or skills. In this case, it appears more difficult to rule decisively that the teaching is domain-specific rather than domain-general, since as Csibra (2007, 96) argues: "the opportunity teaching that has been demonstrated in meerkats does result in the acquisition of a generalizable skill: it not only provides youngsters with food but also 'teaches' them how to kill scorpions." However, despite the apparent (limited) generalizability of this skill, what rules it out as a genuinely domain-general skill is that it appears to be restricted to behaviors designed to facilitate preying on scorpions. Moreover, this is likely to be the function for which this behavior was evolved. Unless one can demonstrate otherwise, it is safe to conclude that this is a domain-specific capacity. (Whether that rules it out as a genuine case of teaching is another matter.) Hence, the second condition on domain specificity enables us provisionally to decide that this is indeed a domain-specific cognitive capacity.

Further Evidence

So far, the examples I have considered derive primarily from studies in evolutionary psychology and comparative cognition. But the concept of domain specificity has also had considerable influence in cognitive neuroscience and developmental psychology. I will now consider whether the notion as I have characterized it can be pressed into service in other areas of cognitive science.

There is a well-established body of evidence indicating the existence of category-specific semantic deficits in a range of patients with brain lesions and other neural abnormalities. However, the correct interpretation of this evidence remains a source of contention. Caramazza and colleagues have interpreted this evidence as indicating that semantic information is "domain-specific" (Caramazza & Shelton 1998; Caramazza & Mahon 2003). Other researchers have adopted different models to explain some of the same findings. Tyler and Moss hold that the selective deficits are an emergent phenomenon. Even though concepts are represented in a unitary distributed system, different types of concepts are structured differently. Since concepts in different domains have different internal structures, impairment of brain function leads to their being differentially affected (Tyler & Moss 2001).

On the face of it, much of this evidence, and the surrounding debate, seems to pertain not to the question of domain specificity but rather to that of brain localization. When damage to a certain part of the brain results in selective impairment in naming animals but not plants or body parts, the question is whether this is evidence that representations underlying our semantic information concerning animals is localized in a particular area of the brain, or whether they are not localized but that some of them are more impaired than others by such damage. Although this is an important question in its own right, it does not bear on domain specificity as such.

Similarly, neuroimaging data that has been brought to bear on this controversy is largely pertinent to the question of localization rather than domain specificity. On the one hand, Caramazza and Mahon (2003, 358) think that "there clearly does seem to be neural differentiation by semantic category" based on neuroimaging data. But Tyler and Moss (2001, 246) find that: "The most striking aspect of the neuroimaging data is the extent to which living and non-living concepts activate common regions with only small and inconsistent differences between domains." The neuroimaging data is obtained mainly by testing healthy

subjects on a variety of tasks (e.g. silent naming, word-picture matching) and then using various techniques (fMRI, PET scans) to determine whether different areas of the brain are differentially involved when processing content derived from different domains (e.g. animals, tools, food items, etc.). But this does not seem to enable us to draw conclusions regarding whether our capacity to think about such domains involves abilities that are generalizable or not. If our knowledge of animals activates different brain areas than our knowledge of tools, that does not mean that any cognitive abilities that range over such domains are restricted to these domains and cannot be applied to others.

Among developmental psychologists, a debate has also raged concerning the domain specificity of our cognitive capacities. For instance, Carey and Spelke have argued that children have innate systems of knowledge that apply to distinct sets of entities and phenomena. Moreover, the domains of human knowledge, such as knowledge of language, physical objects, and number, center on distinct principles. These "core principles" serve to distinguish one domain from another. But despite the fact that Carey and Spelke hold that our cognitive makeup consists of distinct domains, they also claim that conceptual change in these domains occurs in part by constructing mappings between these domains. For instance, mappings between the domains of physics and number play a role in children's reconceptualization of matter and material objects. Though the mapping is slow and difficult, children eventually succeed in using this mapping from one domain to another to differentiate the concept of weight from the concept of density (Carey & Spelke 1994, 191-192). But if one can transplant certain principles of reasoning from one domain to another, then those principles are surely not domain-specific. As I have already argued, it is wrong to argue that a cognitive capacity is domain-specific merely on the grounds that it pertains to a distinct body of knowledge. Rather, generalizability of rules or principles is key, and in this instance that condition would seem to be satisfied, thus casting doubt on whether the capacities in question are truly domain specific (as opposed to innate).

Opponents of the claim of domain specificity also sometimes seem to aim their criticism at a different target. Bates (1994/2001) is at pains to distinguish the claim of domain specificity from claims of innateness and (brain) localization. She rightly stresses that a cognitive capacity can have any two of these features without the third. However, her characterization of domain specificity is vague; with respect to language, Bates (1994/2001, 134) says that the claim of domain specificity is that "localized language abilities are discontinuous from the rest of mind, separate and 'special'…" Moreover, despite her explicit cautionary notes, in presenting the arguments for and against domain specificity, she sometimes argues against innateness or brain localization instead. For instance, she argues against the domain specificity of language on the grounds that the brain systems that support language show an extraordinary degree of neural plasticity (Bates 1994/2001, 139). But that does not have a direct bearing on the issue of whether knowledge of language or the capacity to learn language can be generalized to other domains. She also characterizes the controversy over the domain specificity of language as follows: "Have we evolved new neural tissue, a new region or a special form of computation that deals with language, and language alone?" (Bates 1994, 138) Whether or not there is a brain region that has evolved to deal with language alone concerns innateness and brain localization rather than domain specificity.

A more promising case for testing this account of domain specificity can be drawn from the research on face recognition. Researchers tend to be divided as to whether the human capacity to recognize the faces of conspecifics is a domain-specific capacity, or whether it is a capacity that is acquired as a result of more general cognitive processes, of the type used to acquire expertise in other areas of human cognition. Without trying to rehearse the voluminous evidence involved, I will mention just two findings that are pertinent to the issue of domain specificity. Humans do not develop expertise for recognizing the hands or bodies of conspecifics that is at all comparable to their expertise for recognizing their faces, as measured by accuracy and reaction time (McKone, Kanwisher & Duchaine 2007, 12). Similarly, humans show decrease in accuracy in identifying faces when those faces are inverted but do not show such a decrease in identifying houses in the inversion condition (Yovel & Kanwisher 2004). The capacity to recognize upright faces rapidly and accurately does not seem to generalize to other visual stimuli. Object recognition is a skill that is in principle generalizable to domains beyond faces (e.g. hands, bodies, houses), but it fails to be so generalizable in humans. This is clearly in keeping with the first and second conditions outlined above.

What of the evolutionary clause in the second condition? Though it is not always explicitly mentioned by the researchers who work in this area, I venture that it is at least implicitly assumed. Consider the following scenario. Suppose it were found that humans can indeed generalize their face recognition capacities to encompass the faces of dogs. Proponents of domain specificity might not give up on their claim that this capacity is domain-specific, but rather insist that it is a domain-specific capacity that is specific to the domain of faces in general, or perhaps mammalian faces. Indeed, even if further evidence came to light suggesting that this extends to other objects like the facades of houses, they might continue to posit that it is a domain-specific capacity dedicated to the detection of objects with certain parts in particular configurations. What would rule out such a challenge? As I argued earlier, there are no ready-made domains that would enable us to dismiss it in principle. Rather, it seems natural to say that such hypothetical data would not be evidence of domain specificity (albeit across a broader domain) because of evolutionary considerations. Since it is likely that such a cognitive ability would have arisen to detect faces rather than, say, faces of humans and dogs (given the relative recency of the domestication of

dogs), any extension beyond the domain of human faces is indeed a generalization of this ability, and an indication that it is not truly domain-specific. In fact, this is explicitly acknowledged by proponents of domain specificity in this area of research. McKone, Kanwisher and Duchaine (2007, 12) hold that the domain-specific theory "proposes that a face template has developed through evolutionary processes, reflecting the extreme social importance of faces."

Conclusion

In this article, I have tried to provide an analysis of domain specificity in cognition that enables us to make a theoretically useful distinction between domain-specific and domain-general cognitive systems. Drawing on examples from the literature, both genuine and spurious, I have tried to show that there are two features that make a cognitive system domain-specific. First, the cognitive system must be one that is in principle generalizable. Hence, it cannot be something like a body of information concerning a particular area, but something more like a rule or principle that has wider applicability. Second, it must be a system that the subject cannot apply to genuinely novel cases, where novel cases are ones of a type which this system was not originally evolved to deal with, or that are not within what has been termed the *proper function* of this cognitive system. This second condition is important in that it provides us with a principled way of delimiting domains, since these are not antecedently given. This condition is meant to help to address the question of whether or not a creature can go beyond the cases for which its cognitive capacities were evolved. The distinction between domain specificity and domain generality matters because a central debate in contemporary cognitive science concerns the extent to which our cognitive capacities are domain-specific tools evolved to solve certain problems in the evolutionary environment, or whether they are domain-general problem-solving capacities. A resolution of this disagreement depends on a clear means of demarcating domain-specific from domain-general system. In addition, it has often been claimed that one of the main points of difference between human cognition and that of other animals is its domain-general nature. Again, this debate cannot be properly adjudicated unless we have a principled way of making the distinction.

References

Bates, E. (1994/2001). Modularity, domain specificity and the development of language. In W. Bechtel et. al. (Eds.). *Philosophy and the neurosciences: A reader*, Oxford: Blackwell (first published 1994).

Boyer, P. & Barrett H. C. (2005). Domain specificity and intuitive ontology. In D. Buss (Ed.). *The handbook of evolutionary psychology*, Hoboken, NJ: Wiley.

Caramazza, A. & Mahon, B. Z. (2003). The organization of conceptual knowledge: The evidence from category-specific semantic deficits. *Trends in Cognitive Sciences*, *7:8*, 354-361.

Caramazza, A. & Shelton, J. R. (1998). Domain-specific knowledge systems in the brain: The animate-inanimate distinction. *Journal of Cognitive Neuroscience, 10:1, 1-34*.

Carey, S. & Spelke E. (1994). Domain-specific knowledge and conceptual change. In L. A. Hirschfeld & S. A. Gelman (Eds.). *Mapping the mind: Domain specificity in cognition and culture*, Cambridge: Cambridge University Press.

Caro, T.M. & Hauser, M.D. (1992). Is there teaching in nonhuman animals? *Quarterly Review of Biology, 67*, 151–174.

Cheney, D.L. & Seyfarth, R.M. (1985). Social and non-social knowledge in vervet monkeys. *Philosophical Transactions of the Royal Society, B308*, 187-201.

Cosmides, L. & Tooby, J. (1994). Origins of domain specificity: The evolution of functional organization. In L. A. Hirschfeld & S. A. Gelman (Eds.). *Mapping the mind: Domain specificity in cognition and culture*, Cambridge: Cambridge University Press.

Csibra, G. (2007). Teachers in the wild. *Trends in Cognitive Sciences, 11:3*, 95-96.

Fodor, J. (1983). *The modularity of mind*. Cambridge, MA: MIT Press.

Fodor, J. (2000). *The mind doesn't work that way*. Cambridge, MA: MIT Press.

Frank, N. R. & Richardson, T. (2006). Teaching in tandem-running ants. *Nature*, 439, 153.

Hirschfeld, L. A. & Gelman, S. A. (1994). Toward a topography of mind: An introduction to domain specificity. In L. A. Hirschfeld & S. A. Gelman (Eds.). *Mapping the mind: Domain specificity in cognition and culture*, Cambridge: Cambridge University Press.

Khalidi, M.A. (2001). Innateness and domain specificity. *Philosophical Studies, 105*, 191-210.

McKone, E. Kanwisher, N. & Duchaine, B.C. (2007). Can generic expertise explain special processing for faces? *Trends in Cognitive Sciences, 11:1*, 8-15.

Millikan, R. G. (1989). In defense of proper functions. *Philosophy of Science, 56*, 288-302.

Neander, K. (1991). Functions as selected effects: The conceptual analyst's defense. *Philosophy of Science, 58*, 168-184.

Samuels, R. (1998). Evolutionary psychology and the massive modularity thesis. *British Journal for the Philosophy of Science, 49*, 575-602.

Thornton, A. & McAuliffe, K. (2006). Teaching in wild meerkats. *Science, 313*, 227-229.

Tyler, L. K. & Moss, H. E. (2001). Towards a distributed account of conceptual knowledge. *Trends in Cognitive Sciences, 5:6*, 244-252.

Yovel, G. & Kanwisher, N. (2004). Face perception: domain specific not process specific. *Neuron, 44*, 889-898.

Cognition for action: an architectural account for "grounded interaction"

Anthony M. Harrison (anthony.harrison@nrl.navy.mil)
J. Gregory Trafton (greg.trafton@nrl.navy.mil)
Naval Research Laboratory
Washington, DC 20375 USA

Abstract

The effects of priming are not limited to semantics but have also been witnessed in visual-motor tasks (Tucker & Ellis, 2001). By generalizing ACT-R's (Anderson, 2007) existing spreading activation account to include visual representations and broadening the context within which associations are established, we have been able to replicate this small but reliable phenomenon both in simulation and embodied on a humanoid robotic platform. This model illustrates that the effect doesn't require strict embodiment (e.g., Barsalou, 1999) but can instead be accounted for with abstract representations that are "grounded by interaction" (Mahon & Caramazza, 2008).

Introduction

One of the current drumbeats in cognitive science is that cognition is for action. The strongest evidence for cognition for action comes from experiments that show that there is a much tighter coupling of perception and action than previously thought. For example, Glenberg and Kaschak (2002) found that when a sentence implied action in one direction (e.g., "Close the drawer"), participants had difficulty making a sensibility judgment that required a response in the opposite direction. Similarly, when participants indicated whether an object like a teapot was upright or upside down, reaction times were fastest when the response hand was the same as the hand that would be used to grasp the object (e.g., the right hand response was fastest if the teapot's handle was on the right) (Tucker & Ellis, 1998). Many of these researchers argue that this data shows that our thinking is fundamentally embodied, not abstract.

The main idea behind the embodied cognition movement is that cognitive representations and operations are firmly grounded in their physical context and that cognition relies heavily on modality-specific systems and actual bodily states (Tucker & Ellis, 1998; Barsalou, 1999; Wilson, 2002; Niedenthal, Barsalou, Winkielman, Krauth-Gruber, & Ric, 2005). The typical counter to embodied cognition theories are old-style abstract/symbolic theories (Newell & Simon, 1972; Pylyshyn, 1984), which argue that actual experience occurs in modality-specific representations, but those modality-specific states are abstracted and preserved as abstract, amodal symbols. Given the strength of abstract/symbolic theories, some have suggested that the only way that these theories can explain embodied effects is by adding increasingly complex post hoc assumptions about representations and processing (Barsalou, 1999; Niedenthal et al., 2005).

Mahon & Caramazza (2008) argue that the strict embodiment argument against abstract/symbolic theories neglects to consider the possibility that activation, spread through abstract symbols to modal representations, can account for these very same phenomena. While recognizing that abstract/symbolic theories could accommodate such tight perceptual/action coupling, they acknowledge that most theories do not adequately specify the computations and representational content that would permit such coupling through the spreading of activation. Such an abstract/symbolic system would be "grounded by interaction" (Mahon & Caramazza, 2008), if the abstract symbols come to be tightly coupled with their related percepts and required actions through experience acting in the environment. In this manner, the system would be able to exploit both the flexibility of the abstract representations and the richer context afforded by grounded representations.

We present an ACT-R (Anderson, 2007) model that fits within Mahon & Caramazza's "grounded interaction" framework (2008) that provides a process explanation of a classic embodied phenomenon – the visual-motor compatibility effect observed by Tucker & Ellis (2001).

Tucker & Ellis (2001)

Tucker & Ellis (2001) report a series of experiments that show a small but significant effect of visual presentation on grasp responses. In experiment 1, participants viewed a series of objects of different categories (e.g., natural or man-made) that were either large or small. The object size maps directly to the normal grasp used to manipulate the object: a power-grip (i.e., full hand) for large objects and a precision-grip (i.e., thumb and forefinger) for small ones. Objects were placed either near the response hand (15cm) or far away (2000cm). Subjects responded with either a power- or precision-grip response based on the category (i.e., natural/man-made) of the object seen. The task response-mapping (e.g., natural/precision) was varied between subjects.

While there were some simple main effects, the critical result from the first experiment was the interaction between the size of the object and response-mapping. Despite the fact that the size of the object was irrelevant to the task, its compatibility with the response-mapping resulted in reduced reaction times and error rates (figures 1 & 2). Specifically, when viewing large objects, power responses were faster and more accurate than precision responses. Similarly, viewing small objects resulted in faster and more accurate precision responses than power responses.

In experiments 2-4b, Tucker & Ellis used a go/no-go paradigm, with the response-mapping cued by a tone and go/no-go cued by the object category. Experiment 2 presented the response-mapping cue tone 500ms *before* object presentation. The lack of a compatibility effect in the results showed that prior knowledge of the required response was sufficient to override the phenomenon.

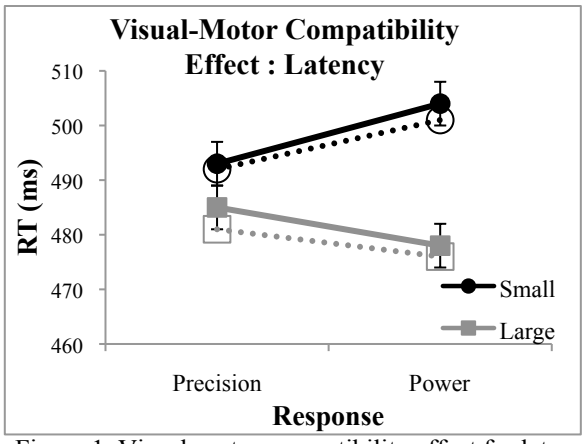


Figure 1. Visual-motor compatibility effect for latency (Tucker & Ellis, 2001, experiment 1). Dotted lines are model fit (R^2=0.99, RMSE=2.95ms).

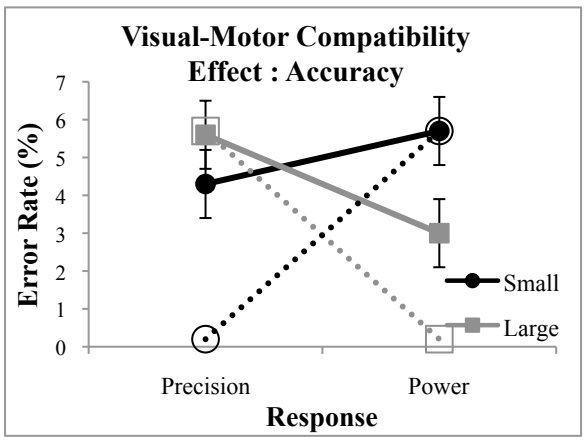


Figure 2. Visual-motor compatibility effect for accuracy (Tucker & Ellis, 2001, experiment 1). Dotted lines are model fit (R^2=0.88, RMSE=2.48%).

Experiment 3 reversed the time delay of the prior experiment and presented the response-mapping cue tone 300ms *after* object presentation. In this circumstance the compatibility effect was present. These results and those from experiment 2 show that the effect is dependent upon the motor system not already being prepared for a particular response.

In experiments 4a and 4b, the visibility of the object was manipulated. In 4a the object disappeared at the same time as the response-mapping cue tone was presented. In 4b the object disappeared 300ms *before* the response-mapping cue tone. The compatibility effect was present in 4a and not 4b, showing that the object's visibility at response selection is critical.

To summarize, Tucker & Ellis have shown that when the object's normal grasp response is compatible with the experiment's response-mapping, there is a small but significant benefit (experiment 1). However, this is conditional on the motor system not already being prepared for a particular response (experiments 2 & 3) and that the object is visible at response selection (experiments 4a & 4b). They discount the theory that this is an example of the percept *directly* priming a particular motor response, arguing that the object would have to be within reach and that such a mechanism would not work for images of objects as well (Tucker & Ellis, 1998). Instead they propose that this is evidence of "a more general representational mechanism that describe object properties in motor terms" (Tucker & Ellis, 2001).

Architectural Account

Within the ACT-R cognitive architecture (Anderson, 2007), the time it takes to retrieve a specific memory (i.e., chunk) is inversely related to that chunk's activation. The chunk's activation is composed of three primary components: base-level activation, spreading activation, and some stochastic noise. Base-level activation is a learned quantity subject to decay that incorporates the effects of frequency and recency of the memory's use. Spreading activation is context dependent, allowing chunks that are the focus of attention to activate related memories. In this way, the chunks within a given buffer (i.e., the focus of attention for a given module in ACT-R) can make related concepts more readily retrievable. Spreading activation is the mechanism used to account for semantic priming effects (Anderson & Reder, 1999). This same mechanism is used here to model the visual-motor priming reported by Tucker & Ellis (2001).

ACT-R defines the current context as the contents of the chunks currently in the model's buffers. If chunk *i* is in a buffer *k*, then all of the chunks that *i* references are in the context. The source activation of buffer *k* is shared equally among those context chunks, and they in turn spread that activation to all the chunks that contain references to them. ACT-R only establishes associative links from the referenced chunk to the referring chunk. The more chunks that reference a specific chunk *j*, the weaker its associative strengths are to the referring chunks. Chunk *j* becomes a less effective retrieval cue because the weaker associative links spread less of the source activation.

This mechanism of spreading activation through associative links from the currently defined context allows ACT-R to model semantic priming (Anderson & Reder, 1999). However, in order to address the visual-motor priming shown in Tucker & Ellis (2001), ACT-R's existing mechanisms must to be modified slightly. These modifications are not complex post-hoc assumptions, rather they are consistent with the existing framework.

Visual Representation and Activation Normally, ACT-R models use only the intentionality system (i.e., goal buffer) as a source of activation, even though all buffers have the capability. Obviously, in order to support visual priming,

the visual buffer must also be used as a source of activation. However, the utility of visual activation is limited due to the traditional structure of the visual representations. The visual representation does not represent a semantic concept; rather, it is a raw percept made up predominantly of non-chunk, primitive features (e.g., numbers, strings). They are therefore highly insular, having little connection to other chunks, which dramatically limits the spread of activation. Typically the visual object's *value* slot (usually a string literal) is used to uniquely associate it with the semantic representation of that percept (i.e., its symbol). To allow a visual percept to activate a semantic symbol, as well as other chunks related to that concept, the *value* slot was modified to reference the semantic symbol chunk directly. To access the semantics, a retrieval must still be made, but now the percept itself can prime that retrieval.

Co-occurring Contextualization Canonical ACT-R only establishes associative links between chunks through symbolic references (i.e., chunk *j* can activate chunk *i* since *i* directly references *j* as a slot value). We propose that symbolic links can also occur through co-occurrence. The context within which processing occurs is not limited to the symbolic structure of the chunks currently in buffer, but actually includes the patterns within the processing units (i.e., productions) that execute cognition. If a production matches against both contents of the goal and visual buffers in order to fire, then the contents of those buffers do not define the context independently, but jointly, and should be linked associatively. Because productions can contain perceptual and motor patterns, perception and action can become linked through co-occurrence.

The application of this mechanism is relatively straightforward. Specifically, the semantic symbol of a percept and the motor command used to grasp the object are associated with each other even though neither has a direct symbolic relationship to the other. These associations are learned from the environment as a consequence of attending to an object, considering its meaning, and manipulating it.

In the language of Mahon & Carmazza (2008), the semantic symbol that is linked to a percept provides the abstract representation that mediates perceptual processing and motor activation. The motor and visual representations are grounded to this abstraction through a history of interaction, allowing the establishment and strengthening of associative links through co-occurrence. Activating the abstract symbol propagates activation both to experienced percepts and motor commands.

Model Details & Fit

The model presented here focuses on a simplification of Tucker and Ellis' (2001) first experiment; how it accounts for the subsequent experiments will be saved for the discussion. Because their presentation distance manipulation had no influence on the visual-motor compatibility effect, it was eliminated from the simulation. Otherwise the simulation is identical to the actual experiment including the timing of object presentations.

Execution The model completed 160 trials (as did participants) where it was presented small and large objects that were either natural or man-made (e.g., strawberry, key, potato, frying pan). Retrieving the visual-symbol associated with the percept, the model was able to classify the object. With this information the model retrieved the appropriate response-mapping for the classification (e.g., natural/precision or man-made/precision). The final retrieval was of the appropriate grip command itself. Once the motor command was retrieved it was passed to the motor system to be executed as the trial response.

Assumptions This model relies upon three key assumptions. First, that activation is spread through not only the goal buffer but also the visual buffer. Second, that the process of encoding a visual percept includes linking that percept to its semantic representation (i.e., its visual-symbol). Finally, associative links are not limited to containment relationships. Over the history of interacting with the environment, both the visual-symbol for a percept and the motor command used to manipulate the object come to be associated with each other via co-occurrence.

Since priming within ACT-R is function of spreading-activation, base-level activations are not of theoretical interest. However, these values do come into play with respect to the rapid retrieval times in the data (figure 1) and are discussed in detail later.

Spreading Activation The model proposes that the visual-motor compatibility effect reported in Tucker & Ellis (2001) is due to activation spreading both from the intentionality (i.e., goal) and visual systems. Once the object is visually encoded, activation is spread to the learned motor response through the co-occurrence associative link between it and the visual-symbol. When the model has retrieved the appropriate response-mapping for the object's category, activation is spread to the task appropriate response. For incompatible responses, activation is spread to two different motor commands. However, when the responses are compatible, both activation sources converge on a single motor command (figure 3). Because of the higher total activation of the compatible motor response, it can be retrieved faster. The lower activation of the incompatible response also makes misretrieval more likely since noise might exceed the differences due to spreading activation.

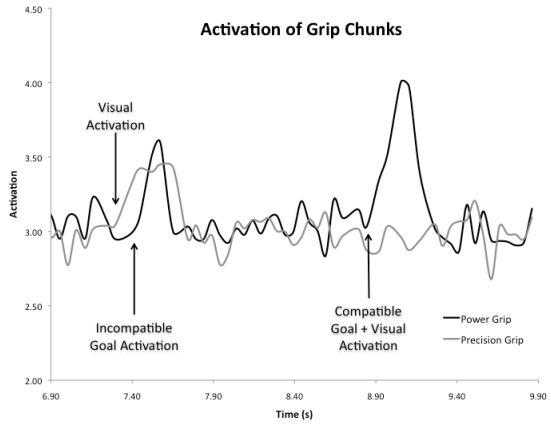


Figure 3. Total activation of incompatible and compatible response motor commands (with noise).

Results and Fits 1000 iterations of the model were run on the simulated version of the experiment. Reaction time fits were quantitatively very strong (R^2=0.99, RMSE=2.95ms, figure 1). Accuracy fits were less strong, but captured the qualitative effect (R^2=0.88, RMSE=2.48%, figure 2). The weaker accuracy fit was due largely to the exclusion of base-level learning from the model. With only spreading-activation, compatible response trials are effectively immune to noise making false retrievals impossible (figure 2 & 3).

Parameters The fits reported above required the manipulation of a few parameters, some of which were dictated by the architecture and the structure of the model.

The maximum associative strength was set to 3.1 (default of 1). This parameter is completely constrained by the structure and connectivity of the model. Conceptually, any chunk that has many references should be a weak retrieval cue; that is its associative strength should be near zero. A maximum associative strength of less than 3.1 would result in negative (i.e., inhibitory) associative strengths for the most heavily referenced chunks. If one of these chunks were to be used as a retrieval cue, it would actually become harder to retrieve its related concept.

The source activation from the goal buffer was kept at the default value of 1. The activation from the visual buffer was set to 0.3, instead of the default of 0. This allows the contents of the visual buffer (namely the semantic symbol) to weakly prime the normal motor response for the object.

While base-level learning was not used in this model, base-level activations were still critical to achieve the rapid retrieval times implied by the average response time of 490ms. Three separate factors influenced the selection of the base-level activations for the visual-symbol, grasp-command, and response-mapping chunks. First, the model, as implemented, requires five productions with three retrievals before a response can be started. At 50ms per production, an additional 85ms for the visual object encoding, and a minimum motor execution time of 50ms, there is only 105ms left for the three retrievals. Second, we assume that over a lifetime of observing and interacting with these objects, the base-level activations for the visual-symbols and grasp-commands are both stable (i.e. relatively immune to decay) and strong. Since we generally see objects more often than we grasp them, visual-symbol activations were set greater than the grasp-commands. Finally, while the response-mapping chunks would benefit from recency, their frequency of use would still be small, so base-level activations were set lower than those of the grasp-command chunks. Base-level activations of 5, 3, and 2.25 were used for the visual-symbol, grasp-command, and response-mapping chunks respectively. While these values are necessary for the low RMSE latency fit, the qualitative (R^2) fit is less sensitive to the base-level values.

To account for the errors in performance, the model relied upon misretrievals. This was accomplished by setting the activation noise parameter to 0.06. The qualitative error results are largely unchanged for most published noise values since the noise only affects the incompatible responses (unless noise exceeds the activation spread to the chunks by the visual buffer). The model's fit of the accuracy data is weaker due largely to the simplification of removing base-level learning. Since compatible responses receive all of the spreading-activation, they are effectively immune to noise (figure 3), which results in 100% accuracy for those trials. To achieve the average 3.5% error rate for compatible trials seen in the data, base-level learning would have to be enabled. This could produce situations where successive retrievals of one particular response might boost its base-level activation such that it could falsely intrude on a subsequent trial. Attempting to fit the error data in this manner would have required seven additional parameters (base-level learning rate, and average age and access counts for the visual-symbol, grasp-command, and response-mapping chunks) instead of the three fixed base-levels used.

Robotic Embodiment

One of the challenges in modeling embodied cognition is the lack of a physical body. This lack is especially relevant because one of the embodied cognition claims is that the body is central to both perception and action; it is disingenuous to claim that we can account for embodied cognition phenomena without a body.

One aspect of running cognitive models on embodied platforms is that actual perception and action must occur. Critically, both perception and action must use cognitively plausible representations and cause the physical body to move.

We have modified ACT-R by allowing it to perceive and act on the physical world by attaching robotic sensors and effectors to it; we call our system ACT-R/E (*Embodied*) (Trafton, Harrison, Fransen, & Bugaska, 2009). Changes to the visual and motor modules are described below.

The Visual Module is used to provide a model with information about what can be seen in the current environment. ACT-R normally sees information presented

on a computer monitor. We modified the original visual module to accept input from a video camera. The visual module allows access to object identification through fiducial (Kato, Billinghurst, Poupyrev, Imamoto, & Tachibana, 2000) and face (Fransen, Hebst, Harrison, & Trafton, 2009) trackers.

Traditional ACT-R has a virtual motor system that allows virtual hand movements (e.g., typing, mouse movements). ACT-R/E's motor module allows control over all of the robot's effectors. When a motor chunk enters into the motor module, a specified motor controller executes the actual physical response.

Our current robot platform is the Mobile-Dexterous-Social (MDS) Robot (Breazeal, 2009). The MDS robot neck has 18 degrees-of-freedom (DoF) for the head, neck and eyes allowing the robot to look at various locations in 3D space and 11 DoF on its four-fingered hand, allowing it to make various gestures and grips. Perceptual inputs include two color video cameras and a SR3000 camera to provide depth information. For the current project, the MDS head can identify various objects through the fiducial tracker and can move its hands in a power or precision grasp.

The 10ms visual-motor compatibility effect is completely obscured by the robot's motor system's slower execution times. In order to illustrate the effect, we dramatically increased the retrieval time scalar. In the video (see acknowledgments for the URL) the visual-motor priming accounts for around a 500 ms performance improvement.

Discussion

ACT-R has a long history of accounting for semantic priming effects (Anderson, 1974), but its perceptual/motor integration has been less explored. To address this theoretical gap, we have modified the visual representation linking the percept to its derived abstract symbol. This allows source activation to usefully spread from the visual system, instead of just from the intentionality system (i.e., goal buffer). We also present a broadened definition of predictive context for the establishment of associative links. Traditional ACT-R only establishes associative links from the contained chunk to the container. In this way, when another chunk has a reference to the contained chunk, it is potentially predictive of the need for the containing chunk. We augment spreading activation to deal with co-occurrence so that we can establish a richer context. In this manner, the visual-symbols and motor commands come to be associated as productions fire that simultaneously match both of the representations in their respective buffers. While only the consequences of this mechanism are exploited in the model presented, the actual process is under active investigation.

A particular limitation of this account is that it does depend upon both visual and motor experience with a given object. Lacking such experience, the modal representations will not be associatively linked to the abstract symbolic representation. As such this model cannot account for related effects when novel objects are used; such as those seen when subjects concurrently perform a compatible manual rotation during the classic Shepard & Metzler (1971) mental rotation task (Wexler, Kosslyn, & Berthoz, 1998).

Experiments 2-4 While the model presented only addresses Tucker & Ellis's (2001) first experiment, its extension to the other experiments is fairly straightforward. All of the subsequent experiments used a go/no-go paradigm where the response to be given was cued by a tone and the go/no-go was determined by the object's category. Recall that in experiment 2, subjects heard the response cue 500ms before the object was presented. This 500ms window of time would allow the model to retrieve the appropriate motor response before it had to determine whether or not to execute it. The lack of a visual-motor compatibility effect observed would be due to the fact that the response had already been selected, leaving visual priming no opportunity to influence performance.

In contrast, in experiment 3 the response cue tone was presented 300ms *after* object presentation. As in experiment 1, the visual presence of the object would allow activation to spread to the learned motor response, facilitating retrieval when it was compatible with the response cued by the tone.

Experiment 4a removed the object at the same time as the cue tone was presented. If the model were able to retrieve the motor command at the moment of the cue-onset and visual-offset, the compatibility effect would be observed. However, ACT-R's encoding time for auditory information would actually result in the retrieval starting at least 50ms after presentation. Since ACT-R's spreading activation mechanism is instantaneous, that activation would drop to 0 immediately after the percept disappeared, eliminating visual priming entirely. The results from experiment 4b are more easily accounted for. Since the object was removed 300ms before the cue tone, the activation of the learned motor response would have been eliminated before the retrieval of the task response. However, theoretical proposals that would allow spreading-activation to decay gradually (e.g., van Maanen & van Rijn, 2007) would not only support the compatibility effect in experiment 4a but also make predictions regarding how long the delay in 4b would have to be before the effect disappeared.

Conclusions

Tucker & Ellis interpret their results through a lens of strict embodiment (e.g., Barsalou, 1999). They argue that the phenomenon could not be due to the perceptual priming of the motor response, rather posit that the evidence supports activation of a more general representation that incorporates both visual and motor properties (Tucker & Ellis, 2001).

Mahon & Caramazza (2008) counter that this line of reasoning unjustifiably discounts the possibility that abstract/symbolic systems could account for visual-motor priming by the spreading activation through abstract

symbols. They propose that the challenge for abstract/symbolic systems is to "1) develop a model of the computations and representations that mediate between perceptual processing and motor activation, and 2) specify the conditions under which those computations are deployed" (Mahon & Carmazza, 2008). In this paper, we present a cognitive model that addresses both of those challenges while remaining within ACT-R's existing architectural constraints. While ACT-R is a traditional abstract/symbolic system, this work moves the architecture towards one that is "grounded by interaction", allowing it to not only exploit the flexibility of disembodied abstractions but also the richly contextualized representations inherent in more strictly embodied accounts (Mahon & Carmazza, 2008).

Combining the generalization of activation spread and co-occurrence associations allows ACT-R to account for semantic (Anderson & Reder, 1999), visual-motor Tucker & Ellis, 2001), and potentially even motor-visual (Craighero, Fadiga, Umilta, & Rizzolatti, 1999) priming. This richer account may also be a fundamental component in enabling symbol acquisition/grounding within ACT-R (Barsalou, 2003; Mahon & Caramazza, 2008).

Acknowledgments

This work was performed while the first author held a National Research Council Research Associateship Award and was partially supported by the Office of Naval Research under job order number N0001408WX30007 and 09-Y861 awarded to the second author. The views and conclusions contained in this document should not be interpreted as necessarily representing official policies, either expressed or implied, of the U.S. Navy.

The models and videos are available for download at *http://www.nrl.navy.mil/aic/iss/aas/CognitiveRobots.php.*

References

Anderson, J. R. (1974). Retrieval of propositional information from long-term memory. *Cognitive Psychology*, 5, 451-474.

Anderson, J. R. (2007) *How Can the Human Mind Occur in the Physical Universe?* New York: Oxford University Press.

Anderson, J. R. & Reder, L. M. (1999). The fan effect: New results and new theories. *Journal of Experimental Psychology: General*, 128, 186-197.

Barsalou, L. W. (1999). Perceptions of perceptual symbols. *Behavioral and brain sciences*, *22*(04), 637-660.

Barsalou, L.W. (2003). Abstraction in perceptual symbol systems. *Philosophical Transactions of the Royal Society of London: Biological Sciences, 358,* 1177-1187.

Breazeal, C. (2009). MDS Robot. http://robotic.media.mit.edu/projects/robots/mds/overview

Craighero, L., Fadiga, L., Rizzolatti, G., & Umilta, C. (1999). Action for perception: a motor-visual attentional effect. *Journal of experimental psychology: Human perception and performance, 25*. 1673-1692.

Fransen, B.R., Herbst, E., Harrison, A.M., Adams, W., Trafton, J.G. (2009) *Real-time face and object tracking*. Proceedings from 2009 IEEE/RSJ international conference on intelligent robots and systems.

Kato, H. Billinghurst, M., Pouplrev, I., Imamoto, K., & Tachibana, K. (2000). Virtual object manipulation on a table-top AR environment. In *IEEE and ACM International symposium on augmented reality*. 111-119.

Mahon, B.Z., & Caramazza, A. (2008) A critical look at the embodied cognition hypothesis and a new proposal for grounding conceptual content. *Journal of physiology - Paris, 102,* 59-70.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Niedenthal, P. M., Barsalou, L. W., Winkielman, P., Krauth-Gruber, S., & Ric, F. (2005). Embodiment in attitudes, social perception, and emotion. *Personality and Social Psychology Review*, *9*(3).

Pylyshyn, Z. W. (1984). *Computation and cognition*. Mit Press Cambridge, MA.

Shepard, R., & Metzler, J. (1971). Mental rotations of three-dimensional objects. *Science, 171*. 701-703.

Trafton, J.G., Harrison, A.M., Fransen, B.R., & Bugajska, M. (2009) An embodied model of infant gaze-following. In A. Howes, D. Peebles, R. Cooper (Eds.), *9th International conference on cognitive modeling - ICCM2009*, Manchester, UK.

Tucker, M., & Ellis, R. (1998). On the relations between seen objects and components of potential actions. *Journal of Experimental Psychology-Human Perception and Performance*, *24*(3), 830-846.

Tucker, M., & Ellis, R. (2001) The potentiation of grasp types during visual object categorization. *Visual cognition, 8,* 769-800.

Van Maanen, L., & Van Rijn, H. (2007). An accumulator model of semantic interference. *Cognitive Systems Research*, 8(3), 174-181.

Wexler, M., Kosslyn, S.M., & Berthoz, A. (1998). Motor processes in mental rotation. *Cognition, 68,* 77-94.

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin & Review*.

The resolution of ambiguity during conversation: More than mere mimicry?

Jennifer Roche (jroche@memphis.edu) **Rick Dale (radale@memphis.edu)**

Roger J. Kreuz (rkreuz@memphis.edu)
Department of Psychology, 202 Psychology Building
Memphis, TN 38152 USA

Abstract

Interlocutors often omit important words during conversation, which can lead to miscommunication during ambiguous scenarios (Rayner, Carlson, & Frasyer, 1983). Haywood, Pickering and Branigan (2004) show that under ambiguous situations, listeners are highly sensitive to syntactic primes. The studies reported here evaluated the effects of linguistic and nonlinguistic cues to ambiguity. Experiment 1 implemented a syntactic prime and a visual mistake from a pseudo-confederate to promote disambiguation. Participants were successfully primed to disambiguate their statements during the pseudo-conversation but the visual mistake had no effect. Experiment 2 evaluated the effect of the visual mistake in the absence of a prime during an ambiguous pseudo-conversation. There was a significant effect of visual mistake for participants who believed they were speaking with a real person. Overall, participants did not merely mimic their pseudo-conversation partner's syntactic prime, but perceived other cues to the breakdown in communication to better clarify their own statements.

Keywords: Priming; conversation; language; linguistics; nonverbal communication

Introduction

Verbal and nonverbal communication requires individuals to correctly decode meaning behind an intended message. However, there is a great deal of ambiguity that naturally occurs during conversation. This may occur because individuals are presented with a multitude of information during communication scenarios (i.e., foreground and background information, with an influence from visual, auditory, and motor events). Yet interlocutors have the ability to interpret the intended message with relatively little difficulty (Garrod & Pickering, 2004; Pickering & Garrod, 2004). In fact, miscommunication often occurs (e.g., leaving out a seemingly useless bit of information because it's utility is not recognized; Guhe & Bard, 2008). Individuals often leave out a single word that could help clarify the intended meaning behind a statement (e.g., "that" to group two objects as one, Haywood, Pickering, & Branigan, 2004). Some researchers have suggested that choices in the use of syntax are influenced by ease of production (Bock, 1986; Branigan, Pickering & Cleland, 2000).

This ease in production may help explain why interlocutors often omit information during conversations. Individuals may leave out words because it is initially easier to exclude information when s/he is unsure of what his/her communicative partner already knows (Lee, 2001; Levelt, 1989; Horton & Keysar, 1996). This strategy may save the speaker time in the beginning, but it will be costly in the end. Recent research suggests that this strategy of responding is relatively egocentric. This often occurs because cognitive load is initially reduced at the onset of the conversation, especially when common knowledge has not been fully established (Bard, Anderson, Chen, Nicholson, Harvard, & Dazel-Job, 2007; Rayner, Carlson, & Frayser, 1983; Schober, 1993). Taking an egocentric perspective may eventually become quite cumbersome if the speaker must continually adjust his/her own previous statements when the message is unclear (Levelt, 1989; Miller & Johnson-Laird, 1976). In order to resolve the confusion, interlocutors must perceive the existing ambiguity early on in the conversation. If the existing ambiguity is realized, then there will be no need to restate the message because it will not be misunderstood. Therefore, it is important to investigate how individuals recover during these instances of miscommunication.

Haywood, Pickering and Branigan (2004) have demonstrated an effective method in which conversation partners may resolve instances of ambiguity. These authors suggest that syntactic priming is an effective and automatic strategy interlocutors use to communicate effectively with each other (Garrod & Pickering, 2004). They maintain that under certain situations (e.g., giving instructions) conversation partners will initially respond ambiguously unless they are primed to disambiguate. This type of syntactic strategy shows the listener how to correctly clarify his/her statements. Haywood, Pickering, and Branigan have also shown that syntactic priming has a quite substantial effect on future utterances. This is beneficial to the speakers, because s/he realizes how to disambiguate his/her own statements without explicitly being instructed to do so.

Priming clearly has a dominant influence in dialogue, but interlocutors rarely implement this strategy on their own (Haywood, Pickering, & Branigan, 2004). It should be considered that the effect of the prime might merely represent the automaticity of aligning at the syntactic level (Garrod & Pickering, 2004; Pickering & Garrod, 2004). This level of alignment could represent conversational mimicry, rather than the understanding of why the speaker is required to disambiguate. Other strategies are possible and it is imperative to evaluate other cues speakers may retroactively use to elucidate confusing situations (Horton & Keysar, 1996). The studies reported here will evaluate the contribution of linguistic and/or non-linguistic behavioral cues to the breakdown in communication. If priming truly represents the mechanism behind disambiguation, then there

should be no differences between the use of a prime and the inclusion of a non-linguistic cue.

Experiment 1

The original Haywood, Branigan and Pickering (2004) study used a live confederate to prime participants to disambiguate during a two-referent instructional task. They found that participants were more likely to disambiguate on future trials if they were exposed to a syntactic prime (e.g., the complementizer "that") that resolved existing ambiguity. This study was successfully replicated using a pseudo-confederate (pre-recorded confederate statements; Roche, Caucci, Dale, & Kreuz, 2009; Roche, Dale, & Caucci, 2009). The next logical step to understanding how interlocutors resolve ambiguity during conversation was to evaluate other possible strategies they might use to disambiguate. The current study evaluated the contribution of a prime ("that") and a non-linguistic behavior cue (a visual mistake) to disambiguating ambiguous scenarios. If participants recognize the non-linguistic cue as a salient indication of their own ambiguity, they should then increase the number of times they disambiguate during the entirety of the conversation.

Method

Participants. Participants included 23 University of Memphis undergraduate students (mean age = 19.84 years years; 13 females). All participants were native speakers of American English with normal to corrected vision and no reports of hearing/speech impairments.

Materials. The experiment took place in a private laboratory room. Participants were seated at a comfortable distance from a 20-inch iMac computer screen. A headset with microphone was used to present and record acoustic data. MATLAB PsychToolbox-3 programs (Brainard, 1997) controlled stimulus presentation and recorded participant responses for the conversation.

Stimuli. There were 3 conditions (ambiguous, unambiguous, and incorrect), 12 rounds and 8 instructions per round (4 participant and 4 pseudo-confederate instructions per round). Experimental object stimuli included twenty-five images placed in a 5x5 grid. These grids contained four types of images (13 containers and/or objects, 4 containers + objects and 8 geometric shapes; see Figure 1a for an example of object placement; with 8 empty cells by the end of the round). Auditory stimuli included 48 pre-recorded pseudo-confederate statements (44.1kHz, 16 bit sampling rate, with equated RMS amplitude to adjust for comfortable listening level and to prevent unwanted acoustic cuing) that described 4 types of instruction statements about the object to be moved [e.g., container, object, "that" prime (container + object), no prime (container + object), see Table 1 for example statements]. It should be noted that there was only one prime from the pseudo-confederate per round (12 primes total). However, there were two instances in which the participant could disambiguate his/her instructions. Finally, visual stimuli included 48 pre-recorded pseudo-confederate video responses to the participant statements. Each condition contained a total of 48 videos, which differed by the type of pseudo-confederate video response the participant received (mistake or correct).

The *unambiguous* condition included 7 videos that contained a mistake in which the pseudo-confederate moved the wrong container or object. The *ambiguous* condition included 7 pseudo-confederate videos comprised of a container and object that was initially moved, but then the correct container + object was moved. Finally, the *incorrect* conditions included 7 pseudo-confederate videos comprised of cases in which the corresponding separate container and object were moved, but the correct grouped (C+O) object was never moved (see Figure 1b, for an example of video presentation). It should be noted that the video files that contained mistakes all occurred in the beginning (first 24 trials) of the experiment and were pseudo-randomly assigned to each condition.

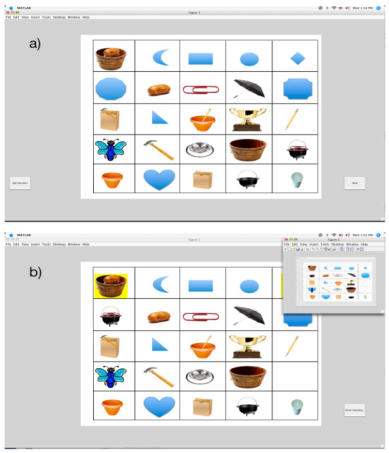


Figure 1. a) The 5x5 grid of objects to be moved by the participant; b) Represents the screen participants see after they have finished giving their instruction to the pseudo-confederate (the small box in the right corner was the video presented to the participant).

Table 1. Examples of pseudo-confederate instructions (C: Container, O: Object, C+O: Container + Object.

Object	Prime	Statement
C	No	"Put the bucket on the circle."
O	No	"Put the paperclip on the stop sign."
C + O	No	"Put the pencil in the flowerpot on the rectangle."
	Yes	"Put the potato *that's* in the bucket on the diamond."

Procedure. To begin, the participant was seated next to a Caucasian female confederate while completing the informed consent, but separated during the experimental sessions. This is an important to control, because much of

perceived talker variability is related to race and gender (Ryalls, Zipprer, & Bauldauff, 1997; Walton & Orlikoff, 1994). Therefore, it was important to match the confederate to the pseudo-confederate's race and gender, to conceal the deception of the task. Participants were then told that their conversation partner (the confederate) would receive instructions first, in a separate room. Once the participant and confederate were separated, the participant was told that they were separated from his/her conversation partner in order to obtain uncontaminated auditory recordings, because individuals often speak over each other during conversations.

Participants were assigned to 1 of 3 conditions that differed based on the pseudo-confederate's video responses (*i.e.*, ambiguous, unambiguous or incorrect). The participant and pseudo-confederate took turns giving instructions about moving objects around the screen (8 instructions per round: 4 participant and 4 pseudo-confederate). It should be noted that each pseudo-confederate statement and video file had a 2s delay before its presentation to imply she was thinking about giving and receiving the instruction.

Participants were informed that the pseudo-confederate would initiate the conversation because she had been viewing the first screen longer. After each pseudo-confederate response, the participant was asked to follow the instructions provided by his/her conversation partner. Once the participant finished moving his/her object, s/he would click a button to transition to another screen to provide instruction to his/her partner (e.g., the object to be moved had a yellow background, and its location had a yellow highlight around it). Once the participant finished giving his/her partner instructions, a smaller window would pop-up on the screen showing the participant if his/her partner made a correct or an incorrect response.

To ensure participants understood the task, they were presented with a brief video prior to the experimental session. This video included 3 mock trials, with 2 male talkers providing each other with instructions and moving the objects around the screen. Once the video was finished, the researcher then asked the participants to rephrase the instructions for the task in their own words. When the researcher felt the participant understood the task, the participant was then asked to make a mental note of how many mistakes were made by his/her conversation partner during the experiment (this helped the researchers determine if they were paying attention to the mistakes). All participants recognized the existence of the mistakes, but on average reported viewing 3-5 out of 12 mistakes (this was not surprising, since the experiment lasted about an hour). Upon completion of the experimental session, the confederate returned and participants were asked, "Would you be surprised if I told you that you were not actually speaking with the person sitting next to you?" The resulting percentage of deceived participants was 92.3%.

Results

A 3 (Condition: ambiguous, unambiguous, & incorrect) x 3 (Block: rounds 1-4, 5-8, & 9-12) mixed fixed effects, repeated measures model with a compound symmetry variance-covariance structure was used to assess the proportion of disambiguated responses from participants during the pseudo-conversation. This model provided non-significant results between the three conditions (see Table 2 for means and standard errors). However, it should be noted that the results from the current experiment did in fact replicate Haywood, Pickering and Branigan's (2004) study, suggesting that participants were significantly affected by the prime (no prime: 25% said "that", 15% disambiguated; prime: 53% said "that", 60% disambiguated).

Table 2. Means and standard errors for the proportion of disambiguation for each condition.

Condition	Mean	SE
ambiguous	0.55	0.15
unambiguous	0.63	0.09
incorrect	0.49	0.14

Discussion. The results from Experiment 1 replicated Haywood, Pickering and Branigan's (2004) study suggesting that participants used the syntactic prime "that" reliably to disambiguate their statements. This suggests that the use of a syntactic prime is effective for disambiguation. Unfortunately, including a visual mistake with a syntactic prime did not seem to significantly influence interlocutors. The interpretation of this non-significant effect may be that the results are an indication of the strength of the prime "that". The prime may have been a highly effective cue participants used to disambiguate. However, the prime alone may have created a ceiling effect in which participants were unable to find a more creative strategy of responding, thus leaving the effect of the behavioral cue hidden. An egocentric perspective may prevent participants from using a syntactic prime, such as "that". Under more natural situations, interlocutors must find other methods to help them disambiguate confusing scenarios.

Experiment 2

Experiment 1 replicated Haywood, Pickering, and Branigan's (2004) study, but failed to show an effect of the visual mistake. Regardless of the null effect, it is still important to evaluate the influence of non-linguistic behavioral cues to communication breakdown. Haywood, Pickering and Branigan suggest that interlocutors often do not automatically use complementizers on their own to disambiguate. They suggest that a listener should be primed to do so, but this is not to say syntactic priming of this nature never occurs naturally. Interlocutors must find other methods to demonstrate the ambiguity perceived, if the syntactic strategy is not naturally elicited. This is especially

important when it is too costly to explicitly describe the ambiguity (e.g., under time constraint).

Therefore, evaluating the use of other strategies interlocutors may enlist during perceived ambiguity is crucial. Again, a syntactic prime is an extremely effective and powerful strategy interlocutors may use to disambiguate ambiguous scenarios. The prevailing nature of the syntactic priming effect might have dampened the effects of the non-linguistic behavioral cue to communication breakdown in Experiment 1. The purpose of Experiment 2 was to evaluate the effects of non-linguistic behavioral cues to miscommunication in the absence of a syntactic prime. If this is an effective cue to disambiguation, then priming that may never occur naturally may be unnecessary under certain communicative scenarios.

Method

Participants. Participants included 16 University of Memphis undergraduate students (mean age = 19.64 years; 12 females). All participants were native speakers of American English with normal to corrected vision and no reports of hearing/speech impairments.

Materials. All the materials were identical to those in Experiment 1.

Stimuli. There were 2 conditions (correct or mistake), 12 rounds with 8 instructions per round (4 participant and 4 pseudo-confederate instructions per round). The object stimuli for this experiment were identical to Experiment 1. The auditory stimuli were identical to Experiment 1, except the prime "that" was removed. The vocally produced word "that" was clipped at the zero crossing at the onset and offset of the production from the original sound files using Audacity. The video stimuli consisted of either a correct response or a mistake provided by the pseudo-confederate. There were a total of 12 mistakes pseudo-randomly assigned throughout the mistake condition. The construction of the mistake was identical to the mistakes created for the "incorrect" condition in Experiment 1 (the incorrect objects were moved).

Procedure. The setup and instructions to the participants were identical to Experiment 1. It should be noted that the pseudo-confederate video presented to participants was moved to the middle of the computer screen to increase the likelihood that the participants would see the mistake. All participants noticed the mistakes and were able to reliably describe the mistakes when asked, but reported seeing on average 2-3 mistakes out of 12. This is not surprising since the experiment lasted about an hour.

Results

Upon the completion of the experimental session, participants were asked, "Would you be surprised if I told you that you were not actually speaking with the person sitting next to you?" The resulting percentage of deceived participants was 67%. Since some participants were not deceived by the experimental design, the statistical analysis for this experiment will include Deception as a factor.

A 2 (Condition: correct & mistakes) x 2 (Deception: deceived & not deceived) x 3 (Block: rounds 1-4, 5-8, & 9-12) mixed repeated fixed effects model with a first-order auto-regressive (AR1) variance-covariance structure, was used to evaluate the probability that individuals disambiguate their statements during an ambiguous instruction task. Upon initial analysis of the variance-covariance structure, the AR1 variance-covariance structure was used because it seemed to have the best fit for the data. The results from this model suggests there was a significant main effect of deception [$F(1, 14.139) = 10.593$, $p < .01$); see Figure 2] and block [$F(2, 24.933) = 5.087$, $p < .05$) see Figure 3]. The model also revealed a significant Condition x Deception interaction [$F(1, 14.139) = 12.682$, $p < .005$); see Figure 4].

The main effect of deception revealed that deceived individuals disambiguated their statements 36.9% more than participants who were not deceived (see Figure 2).

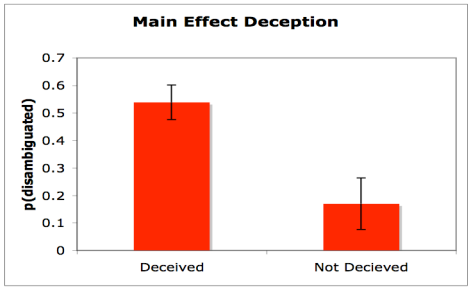


Figure 2. Means and standard errors for the proportion of disambiguated statements for deceived and not deceived participants.

Post-hoc adjusted Bonferroni pair-wise comparisons for the main effect of block revealed that there were significantly fewer instances of disambiguation in block 1, relative to block 2 (19.7%, $p < .05$) and marginally different than block 3 (18.5%, $p = .08$; see Figure 3 for means and standard errors).

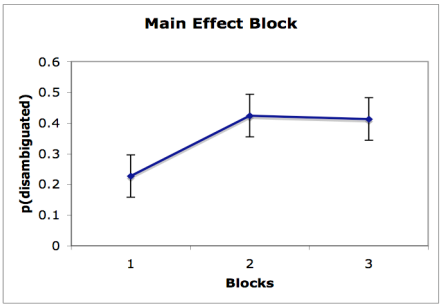


Figure 3. Means and standard errors for the probability of disambiguating during the 1st four rounds, 2nd four rounds and last four rounds.

Post-hoc adjusted Bonferroni pair-wise comparisons for the Condition x Deception interaction revealed that deceived participants who viewed the pseudo-confederate mistakes

disambiguated 52.2% more than participants who did not view mistakes (p < .001). However, there were no significant differences between the participants who were not deceived and the condition they were in (p = .206).

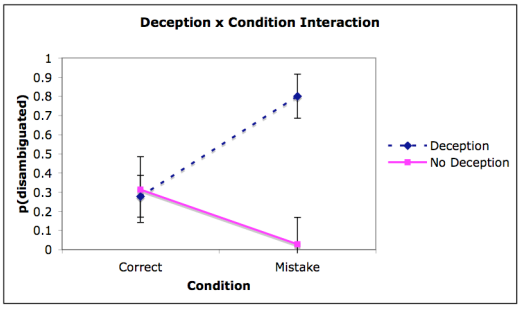


Figure 4. Means and standard errors for the probability of disambiguating when a behavioral cue was provided (a mistake) or not provided (correct).

Discussion

Upon initial evaluation, only 67% of the participants were deceived. However, when participants were asked why they felt the deception failed, many of the participants said that they were aware of the deception that usually occurs during psychological experiments. Many of these participants who were not deceived reported being upper division psychology students or had experience participating in other psychology experiments. This resulted in differential responding between deceived and not deceived participants in Experiment 2. This may have been due to the fact that some participants were more invested (deceived participants) in helping their conversation partner because they may have felt they were truly influencing another person's behavior. The individuals who were not deceived may have felt it was unnecessary to disambiguate, because there was nothing to lose or gain by instructing ambiguously.

Overall, all of the participants disambiguated their statements more as their interaction progressed. This suggests that participants may begin instructing their partners in an egocentric manner because they are initially unsure about the task at hand, but as time progressed they were able to take the other person's perspective into account. Also, the Deception x Condition interaction suggests that when the conversation scenario seemed relatively natural, providing a nonverbal behavioral cue to miscommunication was highly effective. The non-linguistic behavioral cue may have been successful above and beyond the use of a prime to communicate during ambiguous situations [Experiment 1 incorrect condition = 49% disambiguation; Experiment 2 mistake condition = 80% disambiguation; $F(2,18.386) = 50.928$, p <.001].

This suggests that a nonverbal cue to miscommunication may be a more effective cue to use during conversations that are ambiguous. Therefore, addressing such a concern will allow for the evaluation of a non-linguistic behavioral cue such as this under conditions that require interlocutors to quickly and accurately provide information to their conversation partner.

General Discussion

Under conditions in which an interlocutor aligns with a pseudo-conversation partner, priming has been shown to be highly effective (Haywood, Pickering & Branigan, 2004). When ambiguity exists and no prime is naturally produced, participants must find another method to help them disambiguate. The results from these studies do in fact support the notion that priming interlocutors is highly effective under ambiguous scenarios. This finding is supported by Garrod and Pickering's (2004) theory of interactive alignment, which suggests interlocutors automatically align at many levels (syntactic being one of them; Pickering & Garrod, 2004). The use of a syntactic prime may be a successful conversation strategy if at least one of the interlocutors is aware of the ambiguity in the beginning of the conversation. However, if neither participant realizes the magnitude of the ambiguity that exists, both partners might be less likely to adopt a syntactic strategy.

The efficiency of a syntactic prime is apparent, but the nature of participant responses may represent mere mimicry. The possibility that participants are mimicking the syntactic prime may lead to disambiguation. Within this artificial, confederate/pseudo-confederate design, it is quite possible that participants may never become aware of the ambiguity. This disregard of ambiguity may create a situation in which s/he may never realize why s/he needed to disambiguate his/her own statements. This strategy may never be elicited, if conversation partners do not naturally prime each other syntactically, because the ambiguity of the situation is not apparent. This seems to be evident in Experiment 2, when deceived participants in the correct condition disambiguate significantly less than participants who received the non-linguistic behavioral cue. The problem of ambiguity still exists, in which interlocutors never use disambiguating strategies if they do not realize there was a failure in comprehension.

Thus, providing a visual mistake or some other type of behavioral cue should be an alternative and effective strategy interlocutors have available for use during natural conversation scenarios. This notion was supported by Experiment 2, in which the non-linguistic disambiguating cue did in fact help the participants recognize the ambiguity. Recognizing the vagueness in their productions was retroactively beneficial, which allowed them revise their statements to accommodate their listener (Horton & Keysar, 1996). This type of cue to communication breakdown allowed participants to respond effectively and creatively when resolving the confusion.

Unfortunately, the pseudo-confederate paradigm was less effective because some of the participants recognized the artificial nature of the conversation in Experiment 2. This created a situation in which participants may have felt that it was unnecessary to disambiguate their own statements because there was no cost/benefit in doing so. This supports the concept that there may have been a perceived social exchange or *reciprocal altruism* necessary for the

conversation to work properly (Cosmides & Tooby, 1992). When the participants perceived no benefit in disambiguating, they expended less effort and disregarded the constraints of the conversation.

Another assumption in previous literatures has suggested that humans are generally egocentric in regards to their conversation strategies. This suggests that interlocutors rarely take the other person's perspective into account. However, when participants were deceived by the paradigm, they were highly affected by the mistake. This suggests that when interlocutors interact with each other, if there is something to gain or lose during a conversational situation, they are more likely to take the other person's perspective into account. Therefore, the presentation of a behavioral cue may help interlocutors assess the degree to which they invest their energies into the conversation.

It should also be noted, that upon evaluation of the types of syntactic structures the deceived participants chose, they not only used the word "that", when not primed to do so; they also used other syntactic strategies to group the "container+object" images. This supports the view that once speakers become aware of the ambiguity, they are better able to implement a syntactic strategy in the future and a prime may be unnecessary. A non-linguistic mistake has a dominant influence on the strategies interlocutors use to disambiguate scenarios. Therefore, if participants understand that they are communicating ambiguously and there are direct perceived consequences, then they will more quickly try to recover from their mistakes by any means available to them.

Though the pseudo-confederate paradigm was not as effective during the implementation of the nonverbal cue, it was still relatively successful. Future studies should evaluate other scenarios in which the use of a nonverbal behavioral cue to the breakdown of communication might be useful. For example, future studies should evaluate nonverbal behavioral cues under time-constrained tasks. These non-linguistic behavioral cues should also be assessed in more natural conversation scenarios. Future evaluation of such issues will help clarify whether or not a non-linguistic behavioral cue to miscommunication helps interlocutors resolve ambiguity within their own statements quickly and naturally. Understanding the role of such behavioral cues should provide valuable insight into how individuals are able to communicate within their own environment.

Acknowledgments

We thank undergraduates Ryan Morehead, Caitlin Mills, Amy Roche-Purdy, Rachel McKelroy, and graduate researchers Gina Caucci, Kristy Tapp, and Monica Riordan. Preparation of the manuscript was supported by a grant from the National Science Foundation to Rick Dale (NSF HSD-0826825).

References

Bard, E., Anderson, A., Chen, Y., Nicholson, H., Harvard, C., & Dazel-Job, S. (2007). Let's you do that: Sharing the cognitive burdens of dialogue. *Journal of Memory and Language, 57(4)*, 616-641.

Bock, J. (1986). Meaning, sound, and syntax: Lexical priming in sentence production. *Journal of Experimental Psychology: Learning, Memory and Cognition, 12*, 575-586.

Brainard, D. (1997). The psychophysical toolbox. *Spatial Vision, 10*, 433-436.

Branigan, H., Pickering, M. & Cleland, A. (2000). Syntactic coordination in dialogue. *Cognition, 75*, B13-B25.

Cosmides, L. & Tooby, J. (1992). Cognitive adaptations for social exchange. In J. Barkow, L. Cosmides, & J. Tooby (Eds.), *The adapted mind* (pp. 162-228). New York: Oxford University Press.

Garrod, S. & Pickering, M. (2004). Why is conversation so easy? *Trends in Cognitive Science, 8(1)*, 8-11.

Guhe, M. & Bard, E. (2008). Adaping referring expressions to the task environment. *Proceedings of the 30th Annual Meeting of the Cognitive Science Society*. Austin, Tx: Cognitive Science Society.

Haywood, S., Pickering, G., & Branigan, H. (2004). Do speakers avoid ambiguities during dialogue? *Psychological Science,16(5)*, 362-366.

Horton, W. & Keysar, B. (1996). When do speakers take into account common ground? *Cognition, 59*, 91-117.

Lee, B. (2001). Mutual knowledge, background knowledge, and shared beliefs: their roles in establishing common ground. *Journal of Pragmatics, 33*, 21-44.

Levelt, W. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.

Miller, G. & Johnson-Laird, P. (1976). *Language and perception*. Cambridge, MA: Harvard University Press.

Pickering, M. & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences, 27*, 169-226.

Rayner, K., Carlson, M., & Frayser, L. (1983). The interaction of syntax and semantics during sentence processing. *Journal of Verbal Learning and Verbal Behavior, 22*, 358-374.

Roche, J., Dale, R., & Caucci, G. (2009). Pragmatic alignment: The coordination of ironic statements during pseudo-interaction. *Proceedings of the 31st Annual Meeting of the Cognitive Science Society, Amsterdam, NL.*

Roche, J., Caucci, G., Dale, R., & Kreuz, R. (2009). *Conversational puppetry: Priming via pseudo-confederate.* Poster presentation at the 50th annual meeting of the Psychonomic Society. Boston, MA.

Ryalls, J., Zipprer, A., & Baldauff, P. (1997). A preliminary investigation of the effects of gender and race on voice onset time. *Journal of Speech, Language and Hearing Research, 40*, 642-645.

Schober, M. (1993). Spatial perspective taking in conversation. *Cognition, 43*, 1-24.

Walton, J. & Orlikoff, R. (1994). Speaker race identification from acoustic cues to the vocal signal. *Journal of Speech and Hearing Research, 37*, 738-745.

Syntax drives phonological choice – even independently of word choice

Marie Nilsenová (m.nilsenova@uvt.nl)
Department of Communication and Information Sciences, P.O.Box 90153
5000 LE Tilburg, The Netherlands

Marije van Amelsvoort (m.a.a.vanamelsvoort@uvt.nl)
Department of Communication and Information Sciences, P.O.Box 90153
5000 LE Tilburg, The Netherlands

Abstract

We report the results of three experiments designed to test priming percolation ('alignment boost effects') from one grammatical level to another. In the first two experiments, we set off to replicate in Dutch the results of Branigan, Pickering & Cleland (2000) for lexical boosts of syntactic alignment, adding a baseline control condition without priming. In the third experiment, we tested direct syntactic boosts of phonological alignment, using invented verbs. The direct link between syntax and phonology (without any interference from the lexicon) has been postulated in the past, but so far no empirical evidence has been offered in its favor. Our experimental results so far largely confirm the predictions of the Alignment Model (Pickering & Garrod, 2004), including the relation between syntax and phonology. Speakers, who were instructed to use the same syntactic structure as their dialogue partner did, also invented a verb that resembled more their partner's invented verb.

Keywords: Priming; alignment; phonology; syntax; boost effect.

Introduction

Speakers in all age categories adapt their speech to their linguistic environment. They order coffee with milk as 'caffè latte', 'cappuccino' or 'café au lait', depending on what they perceive to be the addressee's choice and they do so even if their personal preference would be to use a different expression (Garrod & Anderson, 1987; Metzing & Brennan, 2003; Branigan et al., in press). Speakers are also likely to copy the syntactic structure previously used by their interlocutor (Levelt & Kelter, 1982; Bock, 1986; Pickering & Branigan, 1999). Even babies only a few months old have been observed to use a higher pitch when interacting with their mother and lower pitch when interacting with their father (Liberman, 1967) and also older children appear to adopt the intonation patterns (low/high boundary tone) when naming pictures, depending on the tone they previously heard from their dialogue partner (Nilsenová, Swerts, Houtepen & Dittrich, 2008). Other documented cases of phonetic/phonological alignment include pronunciation of vowels and consonants, pitch, accent and speech rate (Natale, 1975; Gregory & Hoyt, 1982; Giles, Coupland & Coupland, 1992; Gregory & Gallagher, 2002; Pardo, 2006; Delvaux & Soquet, 2007).

Interestingly, it appears that if the experimental task forces participants to use the same form or structure as their dialogue partner, it increases the likelihood that they will also align on other forms/structures. In other words, alignment on one level of representation "boosts" alignment on other levels. For example, Branigan, Pickering and Cleland (2000; see also Branigan, Pickering, McLean & Cleland, 2007) found that in English, adaptation on lexical level significantly increases the frequency of aligned syntactic structures. In particular, if the subject is instructed to use the same verb as the confederate in the sentence she produced to describe a picture, the subject will be more likely to also use the syntactic structure the confederate did rather than an alternative one. In another series of experiments, Hartsuiker et al. (2008) illustrated the existence of boosts effects in written and spoken computer-mediated communication (see also Raffray, Pickering & Branigan, 2008).

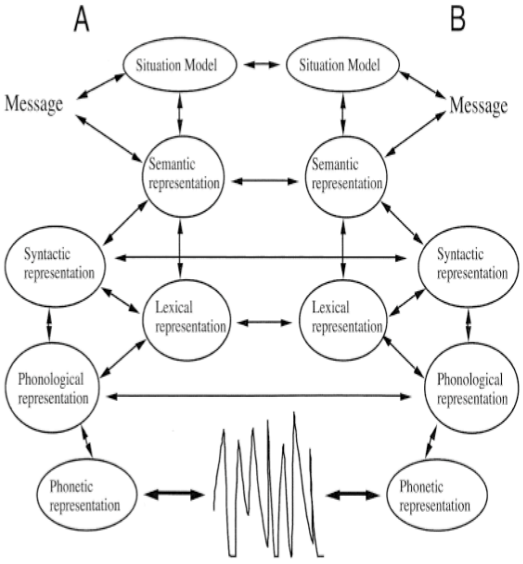


Figure 1: The Interactive Alignment Model (Pickering and Garrod, 2004:177).

Pickering and Garrod (2004) used the phenomenon of alignment boosts to support their (Interactive) Alignment Model. Although the model has been subjected to a number of critical remarks (e.g., Krauss & Pardo, 2004; Schiller & de Ruiter, 2004, and others in the volume), it offers a useful theoretical background for the testing of relations among

various levels of representations. In particular, the vertical lines that stand for possible percolation effects between linguistic representations have for the most part not been tested on empirical data.

Current project

What is of particular interest to us in our current study is the postulated direct link between the syntactic representation and the phonological representation, which in the model appears to be possible even without the intervention of the lexicon (see figure 2). To our knowledge, empirical evidence supporting this relationship is lacking. This is, perhaps, not surprising, since even the expectation of a phonological alignment appears to be rather far-fetched. At least on the level of phonemes, it is unlikely that speakers should be producing strings with identical phonemes (or even strings with comparable phonemic properties, e.g., with respect to the place or manner of articulation). If we exclude the lexical representation, we should be able to observe speakers producing utterances with identical syntactic structures and phonemic properties, turning a conversation into a game of anagrams. In our project, we thus set off to test what appeared to be the 'weakest link' of the Alignment Model, starting with a reproduction of the already established lexical boosts on syntactic alignment.

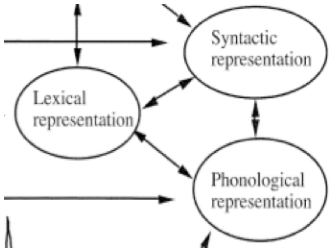


Figure 2: The part of Pickering & Garrod's Alignment Model predicting a direct boost effect of syntactic alignment on the phonological representation (and *vice versa*).

Experiment I.

In the first experimental study, we sought to extend the results of Branigan, Pickering & Cleland (2000) for English by adapting their experimental design for Dutch (for another contribution, see Hartsuiker et al., 2008). Contrary to previous studies, apart from two experimental conditions with a confederate, we also measured the preferred syntactic choices of Dutch speakers in a baseline condition without priming.

Methodology Thirty-nine Dutch speakers were randomly divided into three experimental conditions (A, B, C).

In condition A, **the 'base' condition**, the participants were describing drawings depicting either monotransitive events (one agent only, e.g., a woman drawing a picture; 16 drawings in total), or ditransitive events (including an agent and a recipient, e.g., a woman handing an apple to a boy; 12 drawings in total), viz. figure 3. All the drawings included either a monotransitive (for pictures with an agent only) or a ditransitive verb and the participants were instructed to use the verb in a simple sentence when describing the event.

In condition B, **without lexical alignment**, the participants took part in a confederate-governed task of describing 28 drawings (12 ditransitive stimuli + 16 monotransitive fillers, same as in the baseline study), while being primed alternatively with a syntactic structure of the form 'ditransitive verb + direct object + prepositional indirect object' and a structure of the form 'ditransitive verb + (nonprepositional) indirect object + direct object'. For their description, they were asked to use the verb given under the drawing. Each time, the verb differed from the verb used in the confederate's prime.

In condition C, **with lexical alignment**, participants performed the same task but they were asked to use the verb indicated to them underneath each drawing, identical to the immediately preceding confederate prime. To balance for order effects and verb effects, in both conditions, there were 4 confederate variants with structures alternating per verb.

During the experimental session in conditions B and C, the participant was seated opposite to the confederate who pretended to be ignorant as to the purpose of the experiment. The experimental leader was present in the same room to answer questions and make sure that the participant followed the experimental instructions. The experiment was presented as a game of describing and finding pictures, where both the correctness of the response (picture found) and the time needed to do so would be compared across conditions. The participants were explicitly told that rather than performing the task quickly, they should attempt to be as precise as possible. The output for all the three conditions was recorded on paper (by the participant in condition A and by the confederate in condition B and C), as well as digitally for the spoken dialogue. After each experimental session, the transcripts were compared to the audio recording and corrected if necessary.

The confederate and the participant were taking turns in describing the pictures (see figure 3), with the confederate always initiating the turn (in other words, priming the participant). The confederate picture set included full sentence descriptions of the pictures but in order to maintain the appearance of being a participant as well, the confederate pretended to be making up the descriptions on the spot. The participant was not aware of what was in the confederate set but assumed that it resembled his/her own.

After the experimental session, the experimental leader asked both the confederate and the participant if they noticed anything unusual. Only after that did she disclose

the real purpose of the experiment and the role of the confederate.



(a) (b)

Figure 3: The drawings which the participants were describing depicted either monotransitive (a) or ditransitive events (b). The monotransitive items were used as fillers.

Results In the experimental conditions (B and C), there was a significant effect of lexical alignment on alignment in syntactic structure ($t(21)=3.344$, $p<.005$, eta squared = .035). The participants in the condition C (with lexical alignment) aligned their verbal syntax more frequently (M=9, SD=1.9) than the participants in the condition B (without lexical alignment; M=6.7, SD=1.4). When compared to the condition A (baseline without priming), it turned out that the participants in the condition B and C used the primed constructions significantly less frequently ($F(1.9, 49.403)=5.146$ (sphericity not assumed), $p<.05$, partial eta squared = .165), see figure 4.

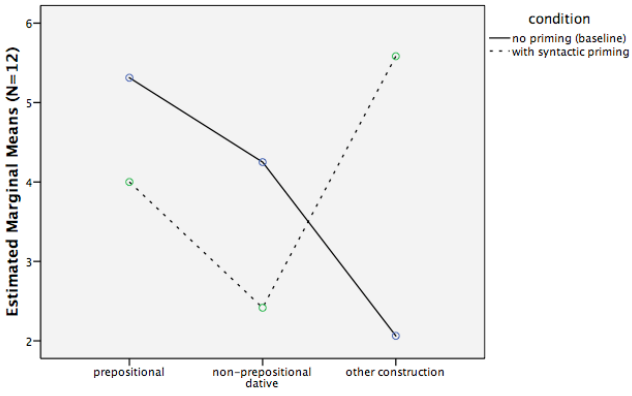


Figure 4: In the experimental conditions with syntactic priming (B and C), participants chose the primed structures less frequently than in the baseline condition without priming (A).

Discussion The comparison of the experimental conditions extends the results of Branigan, Pickering & Cleland (2000) for English to Dutch. We observed that syntactic priming received a lexical boost in the condition in which participants were using the same verb as the confederate in his/her prime. However, the puzzling outcome of the comparison of the experimental conditions with the baseline seems to suggest that syntactic alignment as such does not occur: speakers were more likely to use the prepositional and dative ditransitive constructions spontaneously than when actually primed with them. One possible explanation for the result could be the fact that the ditransitive verbs used in the experiment, such as 'give' (*geven*), 'hand' (*overhandigen*) or 'send' (*sturen*), can be used in monotransitive constructions in Dutch. Unlike in the English version of the task, our Dutch participants in the condition B and C could thus have been influenced by the monotransitive fillers. In fact, they were interpreting them as primes, albeit not in the immediately following turn. To test this hypothesis, we adapted the stimuli from experiment I. in a second experiment.

Experiment II.

In the second experiment, we attempted to account for the outcome of experiment I. (syntactic priming in conditions B and C resulted in less of the primed constructions being used than in condition A with no priming) by changing the structure of the fillers from simple monotransitive clauses of the form 'agent – finite verb – direct object' to clauses containing an adverbial phrase with a preposition, i.e., of the form 'agent – finite verb – direct object – adverbial phrase' (e.g., "The man is painting a picture on the wall" instead of "The man is painting a picture").

Methodology The procedure was the same as in experiment I., only with a different set of fillers as described above. Twenty-two Dutch speakers were randomly divided into one of the two experimental conditions either without or with lexical boost (B and C, respectively).

Results As in experiment I, participants in the condition without lexical boost (B) aligned less frequently (M=6.6, SD=.84) with the syntactic prime than participants in the condition with lexical boost (C; M=9.1, SD=1.38), $t(20)=4.963$, $p<.001$, eta squared = .55. Contrary to experiment I, this time we observed no uses of monotransitive constructions in descriptions that involved ditransitive events. In other words, once we replaced the monotransitive fillers with fillers involving a prepositional phrase (e.g., a locative), the participants used no alternative constructions on the experimental trials to describe the ditransitive events; they always chose either the prepositional dative construction or the non-prepositional dative.

Discussion On the basis of the results obtained in the second experiment, we concluded that participants in experiment I were, in fact, adapting to the monotransitive fillers used by the confederate in the turn preceding the ditransitive prime. When the monotransitive fillers were adapted to longer sentences resembling the experimental primes, their effect disappeared.

Experiment III.

In the third experiment, we explored the effect of a syntactic boost on phonological alignment. In order to test for the relationship directly, it was necessary to exclude the effects of the lexicon that is likely to facilitate phonological alignment in spontaneous data.

Methodology In the baseline condition, twelve drawings depicting a ditransitive event were presented to 17 Dutch speakers who were asked to describe the picture using a monoclausal sentence and a verb they would invent on the spot. In the experimental conditions, the participants again engaged in a confederate-steered task during which they were describing 24 drawings (same as in experiment I and II) with an invented verb, following a syntactic prime by the confederate which also involved an invented but Dutch-sounding verb (with correct morphology). The participants were being primed alternatively by a monotransitive construction or a structure with a direct object followed by a prepositional indirect object, or a structure with a non-prepositonal indirect object followed by a direct object. The phonological primes (i.e., the invented verbs) were alternatively monosyllabic and disyllabic words with a systematically varied phonological structure.

In the pilot version of the experiment, twenty-two participants received no instructions regarding the syntactic structure they were expected to use to describe the pictures. One third of the invented verbal primes contained two plosives (in the onset and the coda for the monosyllabic primes, or in the onsets of the two syllables of the disyllabic primes), another third contained two nasals, and yet another third contained two fricatives. There was no systematic variation of vowels and liquids which were inserted freely to make the verb appear Dutch-like.

When we compared the syntactic output of the participants to the baseline condition, however, we observed that there was no significant difference in the use of the three alternative structures to describe the depicted events (viz. figure 5). In other words, the participants in the experimental conditions were not aligning syntactically and hence it was not possible to measure the effect of a syntactic boost on phonology. Moreover, while the participants appeared to be taking over some phonological features of the verbal prime, the manner of articulation of the consonants did not appear to be a perceptually prominent feature.

Twenty-three speakers of Dutch took part in the third experiment. On the basis of the outcome of the pilot experiment, with respect to syntactic alignment, we adapted the task in such a way as to force the speakers to use the same structure as the confederate. In particular, we instructed them to start describing the picture by a clue that was given to them as an NP + relative clause underneath the drawing. In practice, the speakers were filling in an invented verb into a blank of the form NP – who – IO – DO (e.g., *De man die de non een appel...* – "The man who … the nun an apple.") or NP – who – DO – PO (e.g., *De man die een appel aan de non…* - "The man who … an apple to the nun"). Furthermore, we changed the phonological primes so that the systematic variation in the confederate's verbs consisted (1) in the number of syllables (one or two) and the initial phoneme (a vowel or a consonant), see table 1.

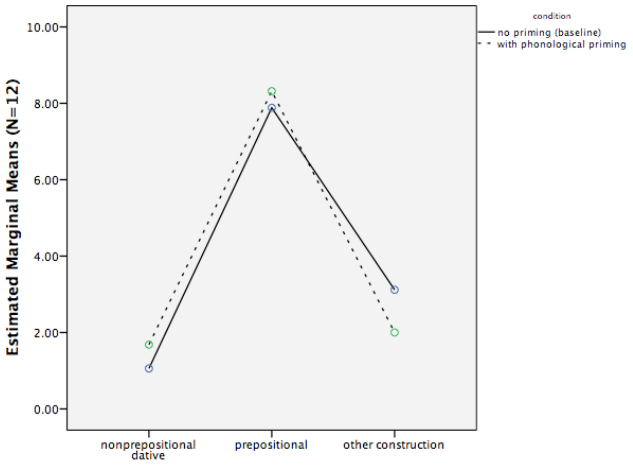


Figure 5: In the pilot version of the experiment, participants in the condition with phonological priming did not differ from the participants in the baseline condition (without priming) in their choice of syntactic structures, $F(2,74)=.825$, $p=.442$.

Table 1: Nonsense verbs used as primes in experiment III.

Initial phoneme	Monosyllabic	Bisyllabic
Vowel	*oeft*	*oegert*
	aapt	*eivelt*
	oot	*affelt*
	iert	*uitert*
	eift	*iemelt*
	eemt	*okkelt*
Consonant	*proest*	*manilt*
	kniert	*pippelt*
	bort	*lippert*
	vlaapt	*zachelt*
	slinkt	*poenkert*
	loept	*niesert*

Results The nonsense verbs created by the participants were transcribed by the experimental leader during the experimental session, as well as recorded digitally. The transcriptions were made in such a way as to reflect the rules of the Dutch spelling system and checked against the audio recordings first by the experimental leader and

subsequently by another linguist. We calculated the proportion of phonological alignment by (i) comparing the number of syllables in the prime and the verb created by the participant, (ii) comparing the initial phoneme of the verb (vowel or consonant), and (iii) comparing the Levenshtein distance between the prime and the participant's verb. The Levenshtein distance between two strings A and B is the (uniform) cost for insertion, deletion and substitution of characters in string B needed to make it identical to string B. The comparison was used to account for cases where the participant did not align phonologically on the systematically manipulated features (number of syllables and initial phoneme) but still appeared to create a new verb strongly influenced by the prime (consider, for instance, the invented verb 'choeft', which was independently created by three experimental participants as a response to the prime 'achelt').

When we compared the two experimental conditions, there was no significant difference between the group that aligned syntactically and the group that did not with respect to the initial phoneme of the invented verbs they created. Regarding the number of syllables, we observed a trend in the data suggesting some effect of the syntactic boost ($t(21)=1.855$, $p=.095$, eta squared=.14). The boost effect, however, was clearly present when we measured the Levenshtein distance between the prime and the participants' responses, with verbs created in the syntactic boost condition resembling the primes more ($M=108.42$, $SD=12.42$) than the verbs created in the condition without boost ($M=126.882$, $SD=24.31$; a lower mean stands for less operations needed to make the strings identical), $t(14.597)=-2.255$, $p<.05$, equal variances not assumed, eta squared=.26.

Discussion The results of the third experiment indicate that there is a link between the syntactic and the phonological component that does not have to be mediated by the lexicon. In particular, when speakers align on the syntactic level with their dialogue partner, they are also more likely to align phonologically. The phonological adaptation, however, is rather subtle and, at least in this experiment, was not obvious when we looked at traditional phonological features like the number of syllables or the word-initial phoneme. However, the resemblance between the prime and the response could be detected by calculating the Levenshtein distance between the two strings.

Conclusion

There is evidence that conversational participants adapt to each other's language use at various grammatical levels. This phenomenon has been well documented in a number of experiments, as well as studies of corpus data (Gries, 2005). The focus of our current study was the nature of percolation effects, which have been documented in priming experiments with lexical boost where participants who were forced to use the same verb as the confederate turned out to be more likely to use the same syntactic construction as well, compared to participants who could use a different verb. The evidence for other kinds of boosts has so far been lacking, despite the fact that these effects are interesting in that they offer insights into the architecture of the language model.

In the current study, we examined the link between the syntactic and the phonological component, which at first blush appeared to be rather arbitrarily postulated in the Alignment Model of Pickering and Garrod (2004). In order to approach the topic of alignment boosts in Dutch in a systematic manner, we started with a replication of Branigan, Pickering and Cleland's (2000) study concerning the effect of lexical alignment on syntax, enriched with a baseline study involving no primes. The results of the first experiment were more complex than the English findings due to the difference in selection properties of the Dutch ditransitive verbs, but both the first and the second experiment confirmed that alignment on the lexical level increases the frequency of aligned syntactic structures.

Finally, our data confirmed the prediction of the Alignment Model regarding a direct boost effect of syntax on phonological alignment. The role of the lexicon was excluded in the setup by making use of invented verbs that the participants had to come up with on the spot.

One open question that needs to be answered in follow-up studies concerns the relation between the spoken and the written form of the invented verbs (for example, the combination of graphemes 'oe' is pronounced as /u/ in Dutch but when calculating the Levenshtein distances, we based ourselves on the graphic representation rather than the pronunciation). In general, experimental evidence is needed for other types of boosts apart from the lexical and the syntactic one explored in the current study.

Acknowledgments

The authors would like to thank Annabelle Adams for her help with collecting the experimental data, to Holy Branigan for her suggestions regarding the design of the third experiment, to Erwin Marsi for his permission to use his script to calculate the Levenshtein distances, to Jan Peter de Ruiter for a helpful discussion in the initial stages of the project and to the CogSci 2010 reviewers for their helpful comments. The results of experiment I and of the pilot version of experiment III were reported as a poster at the Architectures and Mechanisms of Language Processing Conference in Cambridge, UK, 2008.

References

Bock, K. (1986) Syntactic persistence in language production. *Cognitive Psychology*, 18, 355-387.

Branigan, H.P., Pickering, M.J., & Cleland, A.A. (2000) Syntactic coordination in dialogue. *Cognition*, 75, B13-B25.

Branigan, H.P., Pickering, M.J., McLean, J.F., & Cleland, A.A. (2007) Participant role and syntactic alignment in dialogue. *Cognition*, 104, 163-197.

Branigan, H.P., Pickering, M.J., Pearson, J., & McLean, J.F. (in press) Linguistic alignment between humans and computers. *Journal of Pragmatics*.

Delvaux, V., & Soquet, A. (2007) The Influence of Ambient Speech on Adult Speech Productions through Unintentional Imitation. *Phonetica*, 64, 145-173.

Garrod, S., & Anderson, A. (1987) Saying What You Mean in Dialogue: A Study in Conceptual and Semantic Coordination. *Cognition*, 27, 181-218.

Gregory, S.W., & Gallagher, T.J. (2002) Spectral Analysis of Candidates' Nonverbal Communication: Predicting U.S. Presidential Election Outcomes. *Social Psychology Quarterly*, 49, 237-246.

Gregory, S.W., & Hoyt, B.R. (1982) Conversation Partner Mutual Adaptation as Demonstrated by Fourier Series Analysis. *Journal of Psycholinguistic Research*, 11, 35-46.

Gries, S.T. (2005) Syntactic priming: A Corpus-based Approach. *Journal of Psycholinguistic Research*, 34, 365-399.

Hartsuiker, R.J., Bernolet, S., Schoonbaert, S., Speybroeck, S., & Vanderelst, D. (2008) Syntactic priming persists while the lexical boost decays: Evidence from written and spoken dialogue. *Journal of Memory and Language*, 58, 214-238.

Krauss, R.M., & Pardo, J.S. (2004) Is alignment always the result of automatic priming? Behavioral and Brain Sciences, 27, 203-204.

Levelt, W.J.M., & Kelter, S. (1982) Surface forma dn memory in question answering. *Cognitive Psychology*, 14, 78-106.

Liberman, P. (1967) *Intonation, Perception, and Language*. Cambridge: The MIT Press.

Metzing, C., & Brennan, S.E. (2003) When conceptual pacts are broken: Partner-specific effects on the comprehension of referring expressions. *Journal of Memory and Language*, 49, 237-246.

Natale, M. (1975) Converge of mean vocal intensity in dyadic communication as a function of social desirability. *Journal of Personality and Social Psychology*, 32, 790-804.

Nilsenová, M., Swerts, M.G.J., Houtepen, V., & Dittrich, H. (2009) Pitch adaptation in different age groups: boundary tones versus global pitch. *Proceedings of Interspeech*, September 6-10, Brighton.

Pickering, M.J., & Branigan, H.P. (1999) Syntactic priming in language production. *Trends in Cognitive Science*, 3, 136-141.

Pickering, M., & Garrod, S. (2004) Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27, 169-226.

Pardo, J.S. (2006) On phonetic convergence during conversational interaction. *Journal of the Acoustical Society of America*, 119, 2382-2393.

Raffray, C.N., Pickering, M.J., & Branigan, H.P. (2008) Relation priming, the lexical boost, and alignment in dialogue. *Behavioral and Brain Sciences*, 31, 394-395.

Schiller, N.O., & de Ruiter, J.P. (2004) Some notes on priming, alignment and self-monitoring. *Behavioral and Brain Sciences*, 27, 208-209.

Corpus Evidence for Age Effects on Priming in Child Language

Jeffrey Gerard (jeffreygerard@gmail.com)
Frank Keller (keller@inf.ed.ac.uk)
School of Informatics, University of Edinburgh
10 Crichton Street, Edinburgh EH8 9AB, UK

Themis Palpanas (themis@disi.unitn.eu)
Department of Information Engineering and Computer Science
University of Trento, Via Sommarive 14, I-38123 Povo, Italy

Abstract

Structural priming, the tendency to repeat previously uttered syntactic structures, can give insight into human language processing and acquisition. We report two corpus-based studies of children's structural priming that test the following claim of the item-based account of language acquisition: as older children generalize over structures, priming increases with age. A hypothesis derived from this claim, viz., that the lexical boost effect decreases with age, is also tested. We fit mixed-effects logistic regression models on data from children aged 2 to 7.5 years from the CHILDES corpus. We demonstrate structural priming of arbitrary syntactic structures for the first time in child language data. We also find evidence that priming increases with age, but fail to confirm the hypothesis that the lexical boost effect decreases with age.

Keywords: Syntactic priming; Child language; Corpus studies; Mixed models; Age effects in language acquisition.

Introduction

Priming occurs when an initial stimulus, called the *prime*, causes a bias towards a related stimulus later on. Adaptation to the prime manifests itself in the latter stimulus, the *target*, which is comprehended faster or more accurately, or produced more frequently. A wide range of priming effects has been documented, including the priming of words, syntactic structures, and discourse patterns. The phenomenon is neither intentional nor conscious (Bock & Loebell, 1990). Establishing which aspects of a linguistic stimulus adapt to priming—and which ones do not—gives insight into the mental representation of language and the process by which speakers comprehend and produce sentences.

The vast majority of priming research has been carried out with adults, but there are some recent studies that investigate priming in children (e.g., Savage, Lieven, Theakston, & Tomasello, 2003; Huttenlocher, Vasilyeva, & Shimpi, 2004; Kemp, Lieven, & Tomasello, 2005). Such studies make it possible to examine the development of linguistic representations, based on how priming effects change over the time course of language acquisition, i.e., with the age of the child. Priming can therefore be used as a tool to test specific questions about human language acquisition.

One of the key questions in language acquisition is whether grammatical rules are acquired conjointly with individual words or, alternatively, syntactic knowledge is abstract from lexical knowledge. In the latter case, the question arises of the source of knowledge of the abstract structure of a language, since children's only input to language acquisition is the lexical expressions that they hear. Tomasello's (2000) *item-based hypothesis* proposes that children's early language consists of word-for-word chunks copied from adults' phrases, from which they only gradually abstract patterns and therefore grammar rules. An alternative view is that all children are born with a *universal grammar* (Chomsky, 1980); this theory suggests that abstract grammatical knowledge is innate in the human brain, and merely needs to be parametrized during the course of language acquisition.

This paper explores the item-based hypothesis by studying structural priming in corpora of child language. If a child adapts to structural priming—that is, the child shows a tendency to reuse syntactic constructions heard or produced recently—then this indicates that the child is using old syntactic representations to express new ideas with different words. The item-based hypothesis predicts that this behavior should increase with the age of the child: if syntactic development is a gradual shift from lexically dominant phrase repetition towards generalized grammatical rules, then structural priming should be more frequent in older children, who have more abstract syntactic representations available.

In a well-studied phenomenon called *lexical boost*, structural repetition rises when the target and the prime share a content word, i.e., lexical adaptation boosts structural adaptation (Pickering & Ferreira, 2008). We hypothesize that if grammatical abstraction is thought of as curtailing reliance on words, then priming may show decreased effects of lexical boost as children age.

In this paper, we test both hypotheses: that overall priming increases with age, and that the lexical boost effect decreases with age.

Background

Many experimental studies create an atypical context of language use, requiring the participant to respond to a number of similar trials, where the high repetition of trials may condition participants to become more practiced in their responses, or alternatively, participants may show fatigue. Priming studies, in particular, often present made-up nonce words and observe participants' comprehension or use of them (e.g., Brooks & Tomasello, 1999; Kemp et al., 2005). Teaching a participant a novel word requires multiple exposures which means multiple primes, and it is not clear what effect additive priming

might have. Likewise, several priming experiments, especially with children (e.g., Savage et al., 2003; Kemp et al., 2005; Huttenlocher et al., 2004), entail both hearing and then repeating every prime, again double-priming all targets. Corpus studies are not subject to these confounds, and they can help verify that a phenomenon observed in a few children in a few contexts can be generalized to child language as a whole.

With few exceptions, experimental and corpus studies alike have looked for priming of a small set of specific syntactic alternations—different syntactic forms that express the same semantics—providing very limited coverage of grammar. Recent corpus studies have overcome this limitation, and have found that priming is a more general phenomenon (Reitter, Moore, & Keller, 2006; Reitter, 2008), and that less frequent structures show more priming that more frequent ones (the inverse frequency effect).

In the current paper, we present the first corpus-based investigation of priming in children. In the first of two studies, we replicate an experimental study of the priming of passive and active constructions in children (Savage et al., 2003). Our second study generalizes these results by modeling adaptation to the priming of arbitrary structures. The studies bear on the item-based hypothesis of language acquisition. In particular, we investigate the role of a child's age as a predictor of priming, and consider the influence of lexical similarity.

Modeling Methodology

We used mixed-effects logistic regression to model how various explanatory variables affect structural repetition between pairs of sentences from the CHILDES corpus.

Data

The CHILDES corpus (MacWhinney, 2000) contains over 100 databases of transcriptions of face-to-face interactions between young children and their caretakers. The corpus studies described in this paper used a subset of these databases that contain multiple interviews with a child over different dates, so that priming could be compared at different ages of the same child.[1] For naturalness, the phrase "the corpus" will refer to this subset of CHILDES. The corpus comprises utterances from 84 child speakers, as well as speech from their adult interlocutors.

The most current collection of CHILDES transcripts as of April 17, 2009 was processed to remove structures containing unrecognized words, babble, test words, and fillers (onomatopoeia and child-invented word forms that could be recognized were kept). Certain types of clitics were separated to correspond with morphosyntactic annotations (e.g., *they'll* $\Rightarrow$ *they will*), as were assimilations (e.g., *wanta* $\Rightarrow$ *want to*). Disfluencies, retracings, and repetitions were kept.

CHILDES includes annotations of morphemes and syntactic categories, which are automatically generated by supervised taggers (MacWhinney, 2000). This is in turn used to generate labeled dependency structures based on grammatical relations between words (Sagae, Lavie, & MacWhinney, 2005). Sagae et al. evaluate the dependency hierarchy accuracy to be 90.1% on child language transcripts.

[1]The subset comprised the following databases: Bloom73, Brown, Demetras1, Demetras2, Feldman, Gathercole Gleason, Kuczaj, MacWhinney, Sachs, Suppes, Wells. The Wells database contains British English; all the others contain American English.

Mixed-Effects Logistic Regression

We used mixed-effects logistic regression to identify which variables influence priming in our corpus. Our dependent variable Y is a binary variable that indicates whether there is structural repetition between two sentences ($Y = \textsc{true}$) or not ($Y = \textsc{false}$). Logistic regression is a generalization of linear regression that predicts the logit of the probability p that Y is TRUE, as a function of explanatory variables $X_1 \ldots X_N$:

$$\text{logit}(p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_N X_N.$$

The logit link function is $\text{logit}(p) = \ln\left(\frac{p}{1-p}\right)$. Its inverse is the logistic function, ensuring that as a probability $0 \leq p \leq 1$:

$$\begin{aligned} p &= \text{logit}^{-1}(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_N X_N) \\ &= \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_N X_N)}} \end{aligned}$$

Mixed-effects regression allows the optional inclusion of *random effects* in order to generalize beyond the groups represented in a limited statistical sample. Modeling random effects allows for the possibility that, say, different children have different repetition behaviors, or that each child's repetition behavior changes in different ways as he or she gets older. In the first example, a random effect variable CHILD would be defined to model trends that are specific to each one of its grouping factors: Abe, Abigail, Adam, Allison, etc.

For each possible value g of the random-effects grouping variable, let α_g be the deviation of the dependent variable's mean for group g from the entire dataset's mean; α_g is a random variable from a normal distribution with a mean of zero and unknown variance. α_g is added to each datum taken from group g, effectively adjusting the intercept of regression independently for each group so that uncontrolled effects specific to each group do not unfairly distort the overall model.

Model Specification and Fitting

In the corpus studies reported here, we fitted mixed models with random intercepts grouped by the child speaker of target utterances, which were further nested by database to account for random effects of different annotators, interview strategies, etc. In Study 2, random intercepts were also defined for the syntactic structure being investigated in each datum, which is particularly important because the frequencies of the structures vary greatly, approximating a Zipf distribution.

The corpus data is strongly biased towards younger children, with relatively few utterances from children above five years old. Unlike natural cases of sparsity (e.g., spoken language uses far fewer passive sentences than active ones) the sparsity of data for older children is an artifact of CHILDES. Still, it presents a potential problem, as the model-fitting algorithms had to deal with higher variance for older children.

Models were fit using Laplace's method by the `lme4` software package for the R programming environment. All explanatory variables were centered around the mean to reduce multicollinearity between higher-order interactions and their constituent main effects. We built minimal models by dropping non-significant explanatory variables (unless explicitly relevant to the experiment, or necessary as the component of a significant higher-order interaction).

Study 1: Priming of Active and Passive Voice

To confirm that mixed-model regression analysis of corpus data can provide an insight into structural priming similar to what can be accomplished in experiments, we replicated an experimental design utilized by Savage et al. (2003) and compared the qualitative results of the two methods.

The experiment of Savage et al. (2003, experiment 1) proceeded as follows: In interviews with 84 children from age 2;11 to 7;1 (*years*; *months*), children heard and repeated a prime sentence—either active or passive—describing some transitive action depicted in a cartoon. Then they were shown another cartoon of a different action with different participants and asked "What's happening?". The target sentences the children produced in response were classified as PASSIVE or ACTIVE. Experimenters also varied the amount of lexical overlap that the child could potentially find between the given prime sentence and the child-produced target.

Method

All sentences in the specified subset of the CHILDES corpus (see Modeling Methodology above) were automatically identified as active, passive, or other, guided by heuristics. Whereas all passives primed in Savage et al.'s (2003) experiment included an agentive *by*-clause (e.g., *The ball got caught* ***by the net***), the corpus contains only four examples of children using a passive form with expressed agent, one of which is recitation from a storybook. Agentive *by*-clauses are optional in English, and their sparsity appears to be representative of natural language production (Huttenlocher et al., 2004). Accordingly, the present study considered agentless passives (e.g., *I got caught*) along with agentive passives.

The Savage et al. experiment considered "only the first sentence-like utterance … produced after exposure to each prime sentence," so we also compared only adjacent utterances from the corpus. Only pairs where the target was spoken by a child from age 2;0 to 7;6 were included; the potential primes were spoken by adults and children of all ages, but were always spoken in the presence of the target child. Furthermore, pairs were omitted from the analysis if either of the two sentences contained a negation or was a *wh*-question, which were not used by Savage et al., or if a sentence was not identified as obviously passive or active. A contingency table of the remaining pairings already makes clear that an active prime is much more likely than a passive one to precede an active target; see Table 1.

To answer the main questions of whether children's priming is dependent on their age and on lexical overlap, we fit a mixed-effects logistic regression model where the voice of the target sentence was predicted by the following main effects and their interactions:

- PRIME: the voice of the prime utterance (ACTIVE or PASSIVE);
- AGE: the child's age represented as decimal years with precision to the day;[2]
- LEXBOOST: the ratio of the number of words in common between both utterances to the total number of words in the target utterance;
- PRIMETYPE: CP for comprehension-production priming (another speaker produces the prime and the child comprehends it and produces the target) or PP for production-production priming (the child produces both the prime and the target).

Table 1: Frequencies of adjacent prime-target pairings in Study 1, where the target was spoken by a child.

		Target	
		Active	Passive
Prime	Active	359	13
	Passive	14	13

Table 2: Study 1 parameter estimates. Explanatory variables estimate the logit of the probability that TARGET (the utterance immediately following PRIME) is passive.

	β	$p(>\lvert z\rvert)$	
(Intercept)	-4.447	$\ll 0.001$	***
PRIME[PASSIVE]	1.597	0.082	·
AGE	0.351	0.179	
LEXBOOST	-1.373	0.274	
PRIME[PASSIVE]:LEXBOOST	16.285	< 0.002	**

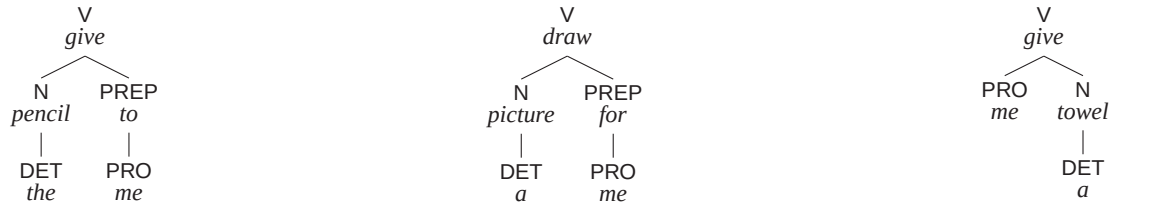
*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, · $p < 0.1$

Results

Table 2 above gives the coefficients of the mixed model together with significance values. We find a significant interaction of PRIME and LEXBOOST. All other interactions and the main effect of PRIMETYPE were evaluated and found to be non-significant regressors, so the model was refit without them. In particular, the model shows no influence of a child's age on his production of active or passive sentences, with or without active or passive primes (no main effect of AGE, no interaction PRIME:AGE).

The dependent variable TARGET was mapped such that passive targets yield TRUE and active targets yield FALSE. The positive coefficient for the PRIME[PASSIVE]:LEXBOOST interaction therefore means that together, a passive prime and lex-

[2] Where a child's age was specified to only monthly precision, the median value of 15 days after the start of that month was assumed.



(a) Prepositional dative (b) Prepositional dative (c) Ditransitive dative

Figure 1: Two prepositional datives share the same structural analysis, which differs from a ditransitive dative

ical boost *increase* the probability that the target will be realized as a passive (as opposed to an active). The marginally significant main effect of PRIME[PASSIVE] lends weak evidence for priming of passives also in the absence of lexical overlap.

Discussion

Our method of modeling the corpus is not identical to the analyses Savage et al. (2003) performed on their experimental data, but the results are comparable in qualitative terms. Savage et al. performed separate analyses of variance (ANOVA, an instance of linear regression), one for each target voice. For both voices, Savage et al. found reliable main effects of PRIME. The reliable interaction between PRIME and LEXBOOST we found in the corpus data was also present in Savage et al.'s ANOVA for passive targets. Meanwhile, PRIME and LEXBOOST formed part of a significant three-way interaction with AGE only in their ANOVA of active targets; they further broke down the active targets to find the PRIME:LEXBOOST interaction in their three- and four-year-old participants but not in six-year-olds. As mentioned above the CHILDES corpus is sparse in data over age five, which likely explains why we did not find any interaction with age.

That the effect of age was only found in active targets suggests that the sparsity of passive targets in both datasets is also important. This weakness cannot be overcome with the studies structured as they are—neither in experimental data nor in corpus data—simply because of the natural sparsity of passives in children's spoken language. Therefore, instead of relying on only a single alternation for insight into children's language, in the next study we investigated children's to priming of arbitrary syntactic structures.

Study 2: Priming of Arbitrary Structures with Decay

Most priming studies to date have only considered structures for which a semantically equivalent alternation exists. This limits the generality of conclusions that can be drawn, and data sparseness is a potential problem, as illustrated above. In the present study, we therefore use an approach that does not require the existence of an alternation, asking instead whether the appearance of a prime structure increases the probability that the *same structure* will appear again.

We define priming in probabilistic terms: the appearance of a prime structure increases the conditional probability that the same structure will appear again:

$$p\left(S_{\text{prime}} \mid S_{\text{target}}\right) > p\left(S_{\text{prime}}\right)$$

where $p\left(S_{\text{u}}\right)$ is the prior probability that an arbitrary structure S will appear in any utterance u . Using this approach, general structural priming—not only for specific structures—can be quantified in a single model.

Besides the sparsity of passives, both Study 1 and the experiment on which it was based had another limitation. By considering only adjacent utterances, they treated priming as an immediate phenomenon and ignored its well-documented temporal decay (Branigan, Pickering, & Cleland, 1999; Pickering & Ferreira, 2008). The formalism just presented is easily extended to estimate $p\left(S_{\text{prime}} \mid S_{\text{target}}, d\right)$, that is, the probability that structure S appeared in the d-th utterance before TARGET was spoken (Reitter et al., 2006; Reitter, 2008).

Structural Overlap

To measure repetition of arbitrary syntactic structures, we need a way to identify whether two constructions share the same structure or are syntactically distinct. We used the hierarchical structure supplied in the form of CHILDES's dependency annotation for this purpose, based on evidence that priming relies on shared hierarchical syntactic rules (Bock & Loebell, 1990; Reitter, 2008). However, priming is not sensitive to thematic roles (Bock & Loebell, 1990), so the relation labels in the annotation are not useful. We therefore used the part-of-speech tags from the CHILDES morphological annotation instead, imposed upon the dependency hierarchy. This combination gives the same analysis to those structures typically considered correspondent in priming studies (Figures 1a and 1b) while producing different analyses for their characteristic alternations (Figure 1c vs. 1a and 1b).

For this study, we used the subset of such structures that have exactly three levels. Of this subset, those with very low frequency—fewer than about twenty occurrences over the entire corpus, according to a manual evaluation—were usually incorrect analyses derived from inaccurate annotations in CHILDES (either in the morphosyntactic or the dependency structure). Thus data points corresponding to structures with frequency less than twenty were discarded. This leaves 4,279 unique structures for consideration, representing 81.3% of the original data. No outliers on the high end of the frequency spectrum were discarded, as they were correct analyses.

Method

Each structure S in some child's (age 2;0 to 7;6) utterance t was considered a potential target of adaptation, primed by the structures in all utterances p within the window of the fifteen utterances preceding t. For each combination of t, $S \in t$, and d ($1 \leq d \leq 15$), a record was created of whether S was in $p = t - d$. That is, the model's binary dependent variable represents *repetition* of a certain structure across a certain distance. Consequently, the parameters estimated by the regression model are effects on mere structural repetition. Priming is identified in this formulation by its decay, so only interactions with the variable DIST (which represents d) can be interpreted in terms of priming; specifically, negative coefficients of DIST indicate priming.

Because measuring grammatical abstraction requires differentiating between lexical and structural repetition, data points showing structural repetition resulting from complete lexical repetition (i.e., not differing by at least one word)—one-half a percent of the dataset—were dropped for this study. Structures in the first fifteen utterances of any interview session also were not considered as targets because they may have been influenced by primes not captured in the corpus. The remaining data points were segregated into strata, one stratum for each three-month period of each child. Two-thousand five-hundred data points were randomly sampled from each stratum, unless the stratum contained fewer than 2,500 points, in which case the entire stratum was used.

A mixed-effects logistic regression model was built to correlate structural repetition across distance (DIST) with explanatory variables AGE and PRIMETYPE as described in Study 1 and with the following variables:

- LEXBOOST: a binary variable that is TRUE if the heads of both hierarchical structures use the same *root morpheme* (lemma);
- ln(FREQ): the logarithmically transformed frequency of the structure in the entire corpus.

This experimental setup crucially relies on the assumption that priming decays. Figure 2 plots the sampled probability that an arbitrary structure is repeated between two utterances separated by a variable distance. It clearly shows the probability of repetition diminishes as a function of distance, with higher repetition across shorter distances—in short, structural priming decays.

There is evidence that both human attention and priming decay logarithmically (McKone, 1995). This is supported by Figure 2, and indeed the mixed-effects model yields a better fit when variable DIST is transformed logarithmically than when it is linear.

Results

Table 3 shows parameter estimates of the full model specification. We find a significant, negative coefficient for ln(DIST), showing the decay of priming of arbitrary structures in children's speech. In line with previous research (Reitter, 2008), the significant interaction ln(DIST):ln(FREQ) demonstrates

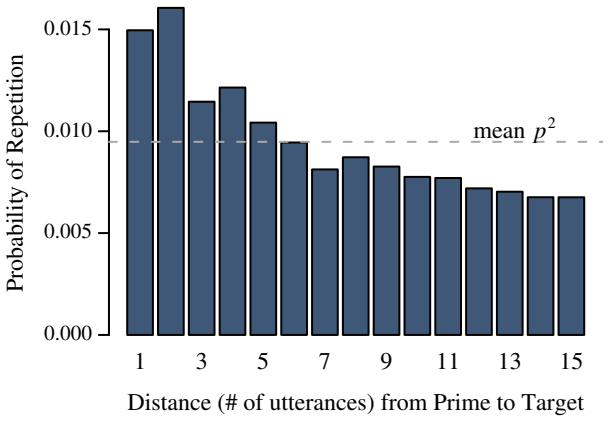


Figure 2: Decay of priming in children. The dashed line is the mean squared prior probability of structures, estimated over comparisons of up to fifteen utterances; since priming raises the probability of a structure above its prior, the true mean will be closer to the repetition probability at DIST = 15.

that less frequent structures show stronger adaptation. Note that the interaction's positive coefficient needs to be interpreted in the context of the negative slope of ln(DIST) (repetition decreases with increasing distance) and the positive slope of ln(FREQ) (less repetition for more frequent structures).

We also observe a significant positive coefficient of ln(DIST):LEXBOOST, suggesting that priming (that is, the decay of DIST) may weaken under lexical boost, all other factors held fixed. Meanwhile, ln(DIST):ln(FREQ):LEXBOOST has a negative coefficient: the decay effect increases with lexical boost and increasing frequency. This means we find the lexical boost effect to be stronger for high-frequency items.

The ln(DIST):AGE interaction is marginally significant ($p = 0.075$), providing only weak evidence for the claim that structural priming increases with age. More convincing support of this prediction is offered by the significant ln(DIST):ln(FREQ):AGE interaction. Its positive coefficient means that priming (the decay of DIST) becomes stronger if age increases while frequency decreases, or weaker as age and frequency increase. In other words, the inverse-frequency effect is stronger for older children than for younger children.

ln(DIST):AGE:LEXBOOST is not significant. We therefore find no evidence for our suggestion that lexical boost may influence structural priming differently as children gradually abstract grammar from phrasal repetition.

The marginally significant ln(DIST):PRIMETYPE interaction hints that children may be more inclined to repeat their own previous constructions (PRIMETYPE = PP) than primes by another speaker.

Discussion

This model shows that priming of arbitrary structures is evident in children, a population in which priming of only a few syntactic alternations had been studied previously. This study also provides an estimate that priming's main efficacy lasts

Table 3: Study 2 parameter estimates. Explanatory variables estimate the logit of the probability of structural repetition. **ln(DIST)** terms are emphasized to remind that this model provides insight to priming only through the DIST variable.

	β	$p(>\|z\|)$	
(Intercept)	-6.299	$\ll 0.001$	***
ln(DIST)	-0.423	$\ll 0.001$	***
ln(FREQ)	0.701	$\ll 0.001$	***
AGE	-0.191	$\ll 0.001$	***
LEXBOOST	2.738	$\ll 0.001$	***
PRIMETYPE[PP]	0.487	$\ll 0.001$	***
ln(DIST):ln(FREQ)	0.107	$\ll 0.001$	***
ln(DIST):AGE	-0.056	0.075	·
ln(DIST):LEXBOOST	0.114	0.049	*
ln(DIST):PRIMETYPE[PP]	-0.075	0.056	·
ln(FREQ):AGE	0.030	0.007	**
ln(FREQ):LEXBOOST	-0.230	$\ll 0.001$	***
ln(FREQ):PRIMETYPE[PP]	-0.150	$\ll 0.001$	***
AGE:LEXBOOST	0.046	0.349	
AGE:PRIMETYPE[PP]	-0.031	0.579	
LEXBOOST:PRIMETYPE[PP]	0.171	0.017	*
ln(DIST):ln(FREQ):AGE	0.022	0.026	*
ln(DIST):ln(FREQ):LEXBOOST	-0.056	0.009	**
ln(DIST):AGE:LEXBOOST	-0.035	0.344	
ln(FREQ):AGE:LEXBOOST	-0.057	< 0.001	***
ln(FREQ):AGE:PRIMETYPE[PP]	0.062	< 0.001	***
AGE:LEXBOOST:PRIMETYPE[PP]	-0.186	0.003	**

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, · $p < 0.1$

around six utterances, during which it shows strong decay and after which its decay is negligible (see Figure 2). Most importantly, this study enables us to quantify priming effects according to age during first language acquisition.

Crucially, the results do not support the conjecture offered in the Introduction that structural priming's reliance on lexical boost decreases as children age. It is important to bear in mind that this conjecture is not strictly predicted by the item-based hypothesis, which does not specify precisely what types of analogies children must make to abstract a grammar from word patterns. Kemp et al. (2005) provide evidence similar to our results, observing in one experiment that two-year-olds adapted to structural priming without regard to lexical influence.

On the other hand, we did find evidence that overall structural priming increases with age. If this is true, it supports the item-based hypothesis of language acquisition which holds that over time children gradually abstract grammatical rules from the sentences they hear.

Conclusion

This paper reported two corpus-based studies of structural priming during first language acquisition. Study 1 replicated an experiment on passive/active priming in children, and found similar effects in corpus data to those reported experimentally (Savage et al., 2003). Both studies tested the hypothesis that structural adaptation increases with age. Study 2 found evidence for this claim, though the change is not as large as might be expected by an item-based account of language acquisition. Neither study supports our conjecture, influenced by the item-based hypothesis, that the lexical boost effect should decrease with age, as children move from lexicalized to abstract syntactic knowledge.

References

Bock, K., & Loebell, H. (1990). Framing sentences. *Cognition*, *35*(1), 1–39.

Branigan, H. P., Pickering, M. J., & Cleland, A. A. (1999). Syntactic priming in written production: Evidence for rapid decay. *Psychonomic Bulletin and Review*, *6*(4), 635–640.

Brooks, P. J., & Tomasello, M. (1999). Young children learn to produce passives with nonce verbs. *Developmental Psychology*, *35*, 29–44.

Chomsky, N. (1980). *Rules and representations*. New York: Columbia University Press.

Huttenlocher, J., Vasilyeva, M., & Shimpi, P. (2004). Syntactic priming in young children. *Journal of Memory and Language*, *50*(2), 182–195.

Kemp, N., Lieven, E., & Tomasello, M. (2005, June). Young children's knowledge of the "determiner" and "adjective" categories. *Journal of Speech, Language, and Hearing Research*, *48*(3), 592–609.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk* (3rd Edition ed., Vol. 2: The Database). Mahwah, NJ: Lawrence Erlbaum Associates.

McKone, E. (1995). Short-term implicit memory for words and non-words. *Journal of Experimental Psychology: Learning Memory and Cognition*, *21*(5), 1108–1125.

Pickering, M. J., & Ferreira, V. S. (2008). Structural priming: A critical review. *Psychological Bulletin*, *134*(3), 427–459.

Reitter, D. (2008). *Context effects in language production: Models of syntactic priming in dialogue corpora*. Unpublished doctoral dissertation, School of Informatics, University of Edinburgh.

Reitter, D., Moore, J. D., & Keller, F. (2006). Priming of syntactic rules in task-oriented dialogue and spontaneous conversation. In R. Sun & N. Miyake (Eds.), *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (pp. 685–690). Vancouver.

Sagae, K., Lavie, A., & MacWhinney, B. (2005). Automatic measurement of syntactic development in child language. In *Proceedings of the 42nd Meeting of the ACL* (pp. 197–204). Association for Computational Linguistics.

Savage, C., Lieven, E., Theakston, A., & Tomasello, M. (2003, November). Testing the abstractness of children's linguistic representations: lexical and structural priming of syntactic constructions in young children. *Developmental Science*, *6*(5), 557–567.

Tomasello, M. (2000). The item-based nature of children's early syntactic development. *Trends in Cognitive Sciences*, *4*(4), 156–163.

Understanding acceptability judgments: Additivity and working memory effects

Laura Staum Casasanto
Max Planck Institute for Psycholinguistics
Nijmegen, NL

Philip Hofmeister
Center for Research in Language
University of California - San Diego
La Jolla, CA

Ivan A. Sag
Department of Linguistics
Stanford University
Stanford, CA

Abstract

Linguists build theories of grammar based largely on acceptability contrasts. But these contrasts can reflect grammatical constraints and/or constraints on language processing. How can theorists determine the extent to which the acceptability of an utterance depends on functional constraints? In a series of acceptability experiments, we consider two factors that might indicate processing contributions to acceptability contrasts: (1) the way constraints combine (i.e., additively or super-additively), and (2) the way a comprehender's working memory resources influence acceptability judgments. Results suggest that multiple sources of processing difficulty combine to produce super-additive effects, but multiple grammatical violations do not. Furthermore, when acceptability judgments improve with higher working memory scores, this appears to be due to functional constraints. We conclude that tests of (super)-additivity and of differences in working memory can help to identify the effects of processing difficulty (due to functional constraints).

Introduction

Grammatical theories are designed to reflect, explain, and predict what is and is not possible to say in a language. Potential utterances are usually classified as "possible" or "impossible" on the basis of native speaker judgments of their acceptability. Whether they are the judgments of theorists themselves or of a sample of naive speakers, these judgments are not a perfect window into the speaker's grammatical competence: the judgments themselves are colored by performance factors. This problem has been discussed since Miller and Chomsky (1963) pointed out that some sentences that native speakers judge to be unacceptable, such as triple center embeddings (1), are better ruled out by their extreme difficulty than by grammatical constraints.

(1) The salmon that the man that the dog chased smoked fell.

Miller and Chomsky's assessment that functional constraints on the language processing system underlie the unacceptability of these examples is fairly uncontroversial. However, it is often difficult to determine what role functional constraints might play in other acceptability contrasts. In the domain of island violations, for example, both processing and grammatical constraints have been proposed to account for the unacceptability of island-violating sentences (Ross, 1967; Chomsky, 1973, 1986; Kluender, 1998, inter alia).

Assessing whether functional constraints underlie acceptability contrasts may be difficult, but it is critical in determining which acceptability contrasts should be taken as evidence for the existence of grammatical constraints. But what tools do we have for recognizing when acceptability contrasts are a consequence of functional constraints? This paper will explore two properties of processing constraints that could help theorists to recognize their effects on acceptability. First, individuals have a limited set of cognitive resources that they can use to understand language (Just & Carpenter, 1992; Kluender, 1998; Cowan, 2001). Extreme sentence processing difficulty can exhaust these resources, resulting in a strong perception of unacceptability, as in (1). Second, the extent of this limited pool of resources arguably varies from one individual to another, as suggested by Just and Carpenter (1992).

To explore the first property, we will consider what happens when multiple possible sources of unacceptability are combined. There are three logically possible outcomes of combining two manipulations that each individually cause acceptability decrements: a significantly smaller penalty than the sum of the two individual penalties (a result which we will refer to as *sub-additive*), a penalty that is statistically indistinguishable from the sum of the two individual penalties (which we will refer to as *additive*), or a penalty that is significantly larger than the sum of the two individual penalties (which we will call *super-additive*).

Super-additive effects may result from combining two manipulations that tax the same set of limited resources, if the manipulations are sufficiently strong to deplete the available resources. Thus, super-additivity could result when multiple sources of processing difficulty co-occur, depending on the degree of difficulty. On the other

hand, formal (grammatical) violations are not expected to combine super-additively. At least for theories that distinguish between formal and functional constraints, grammaticality violations do not cause decrements due to the taxing of a limited set of resources but to the violation of grammatical rules. They could combine additively – if each violation influences judgments independent of the other – or sub-additively, if one violation overwhelms the other or the overall acceptability of the sentence depends simply on the most egregious violation. When formal and functional sources of unacceptability appear in the same sentence, either additive or sub-additive decrements could be the result, for the same reasons, but again, this combination should not produce a super-additive decrement.

To explore the second property, we will compare the performance of people with different processing resources on acceptability judgment tasks. If an acceptability contrast reflects overtaxing the resources of comprehenders, then comprehenders with greater processing resources should experience less difficulty and the contrast should be reduced. However, when acceptability contrasts are due to grammaticality violations, comprehenders with greater processing resources should, if anything, show enhanced contrasts, because they are better able to parse the sentences and notice the rule violations.

In order to evaluate these predictions, we conducted three acceptability judgment studies, combining two processing manipulations (Experiment I), two grammaticality violations (Experiment II), and a processing manipulation with a grammaticality violation (Experiment III).

Experiment I: Processing Difficulty

To investigate the role of processing complexity in acceptability judgments, we manipulated the distance between two dependent arguments and their syntactic head.

Participants Stanford University students (n=32) participated in exchange for payment. All self-identified as native speakers of English.

Materials Twenty-four items were selected from Grodner and Gibson (2005). In these items, the hierarchical distance between a subject and object noun phrase and their subcategorizing verb was varied. This was achieved by varying (1) the presence/absence of a relative clause between the subject and verb [2a,2c vs. 2b,2d] and (2) positioning the object NP immediately after the verb or before the subject NP by relativizing it:

(2) a. The nurse from the clinic supervised the administrator who scolded the medic while a patient was brought into the emergency room. [**short-short**]

b. The nurse who was from the clinic supervised the administrator who scolded the medic while a patient was brought into the emergency room. [**long-short**]

c. The administrator who the nurse from the clinic supervised scolded the medic while a patient was brought into the emergency room. [**short-long**]

d. The administrator who the nurse who was from the clinic supervised scolded the medic while a patient was brought into the emergency room. [**long-long**]

These items were selected because reading time evidence from Grodner & Gibson (2005) show that increasing the hierarchical distance in examples like these leads to slower processing at the critical integration sites. The 24 experimental items appeared with 72 fillers (24 of which were the items from Experiment III).

Procedure To acquire acceptability ratings from participants, we used the thermometer judgment methodology described in (Featherston, 2008). This paradigm resembles the Magnitude Estimation (ME) technique of gathering judgments (Bard, Robertson, & Sorace, 1996; Sorace & Keller, 2005), where participants are asked to rate the magnitude of acceptability difference between test items and a reference sentence (e.g. twice as good, three times as good, half as good, etc.). In both ME and thermometer judgment experiments, participants are not limited to a particular set of values that they can assign to sentences - in principle, every sentence could receive a different judgment.

There are, however, several key differences between the ME and thermometer methods of judgment collection. In the latter paradigm, participants are not instructed to evaluate test items in terms of the magnitude of acceptability compared to the reference item, as evidence shows that participants ignore these instructions and rate sentences in terms of their linear distance from the reference. In addition, in thermometer judgment studies, participants judge items relative to two reference sentences. One of these references is quite good and the other quite bad, and we follow Featherston (2008) in assigning these sentences the arbitrary values 20 and 30. For all of our experiments, we used the same reference sentences.

(3) a. The way that the project was approaching to the deadline everyone wondered. = 20

b. The architect told his assistant to bring the new plans to the foreman's office. = 30

While participants could theoretically assign any real number value to the test items, including negative or decimal values, participants almost always assign positive integer values, typically between 10 and 40.

Sentences were presented to participants on a computer screen one word at a time for a fixed duration, via the DMDX software package (Forster & Forster, 2003). The duration that each word stayed on the screen varied with the number of characters in the word (250 ms + 33.34 * number of characters), so that longer words remained visible for longer periods. We chose word-by-word presentation over full sentence presentation to prevent participants from excessive introspection about the test sentences, and we used auto-paced presentation

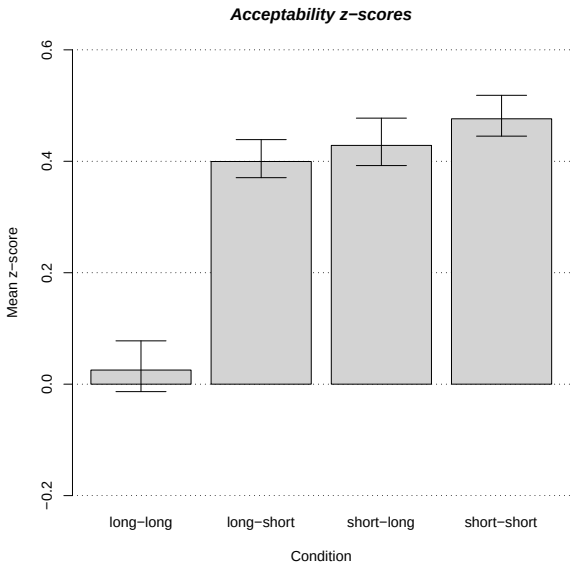


Figure 1: Acceptability z-scores for experiment I. Error bars show (+/−) one standard error.

rather than self-paced presentation to prevent differences in how long each participant studied a given stimulus.

Each participant also completed a reading span task during the same session, used to assess their working memory capacity (Daneman & Carpenter, 1980). For the analysis of reading span scores, we scored each test using using the partial credit method outlined in Conway et al. (2005): successful recall of a word in a study list counts toward the final reading span score, even if the entire item set was not recalled correctly.

Results

Prior to statistical analysis, we log-transformed judgment ratings to normalize the data and to reduce the effect of extreme data points. Subsequently, we computed z-scores for each subject on the basis of all data in the experimental data set (except practice items), including fillers. This reduces the impact of varying uses of the interval scale by subjects. Finally, we excluded data points with z-scores more than 2.5 standard deviations from the mean for each participant. For Experiment I, this outlier removal process affected 2.0% of the data. The resulting z-scores constitute the data on which we conducted statistical analyses.

For all experiments, we used linear mixed effects models to estimate the effects of the experimental manipulations. Such statistical analyses remove the need for prior averaging over subjects and items, are more robust in the presence of missing data, and do not require the assumptions of sphericity that are inherent to analyses such as repeated measures ANOVAs (Baayen, 2004, 2007). This method of statistical analysis also allows for the evaluation of additional factors such as list position alongside effects due to experimental manipulation. Prior to analysis, all predictors were centered—higher order variables (interactions) were also based on these centered predictors. Linear mixed effects models do not directly yield p-values (due to complications in estimating the degrees of freedom), but Monte Carlo Markov Chain (MCMC) sampling can be used to conservatively estimate p-values. For all p-values reported here, we utilized 25,000 MCMC samples to estimate the values.

The acceptability judgment results demonstrate main effects of both manipulations—-subject distance ($\beta = -.242$, t $= -6.412$, $p < .0001$) and object distance ($\beta = -.211$, t $= -5.584$, $p < .0001$). In addition, there was a highly significant interaction between these factors ($\beta = -.328$, t $= -4.343$, $p < .001$). This interaction reflects the result of combining multiple processing difficulties: the acceptability decrement produced by two processing manipulations was more than expected on the basis of the decrements produced by each manipulation in isolation.

Reading span score was also a highly significant predictor of acceptability scores ($\beta = .050$, t $= 3.685$, $p < .001$). In particular, higher reading span scores predicted higher judgments of acceptability. This effect appears to be largely driven by the conditions with multiple processing manipulations and a dislocated object phrase (the most difficult conditions according to Grodner and Gibson (2005)), which is reflected by the significant interaction of reading span score and the object distance manipulation ($\beta = .068$, t $= 3.858$, $p < .01$).

Discussion

According to the results, while these kinds of processing manipulations may have only minor effects on acceptability in isolation, they can have highly significant effects on judgments when combined. In this study, increasing the distance between a single dependent argument and its head only slightly lowered judgments. But when we increased the hierarchical distance of both dependents to their syntactic head, a sharp drop in acceptability judgments occurred. Consequently, these results indicate a super-additive effect on judgments resulting from the co-occurrence of multiple sources of difficulty in sentence processing.

In addition, estimates of working memory (operationalized as performance on the reading span test) indicate that better working memory predicts higher judgments of acceptability for items with processing challenges. This suggests that a positive linear relationship between reading span scores and acceptability scores indicates significant processing difficulty in the test items. The strength of these conclusions, however, depends on whether a similar relationship appears in sentences with grammatical violations.

Experiment II: Grammatical Violations

Experiment II evaluates how multiple grammatical violations affect judgments when they co-occur in the same sentence. Since grammatical violations do not affect acceptability via overtaxing processing resources, combining them should not result in super-additive decre-

ments (unlike the processing manipulations in Experiment I). In addition, for the same reason, comprehenders with greater working memory capacities should if anything show a greater decrement for grammatical violations than low-capacity comprehenders (also unlike the results of Experiment I).

Participants Stanford University students (n = 28) who had not participated in Experiment I completed this experiment in exchange for payment.

Materials The 24 experimental items in Experiment II contained either 0, 1, or 2 grammaticality violations. We manipulated the grammaticality of two separate but nearby constituents to yield a 2 x 2 design. The first manipulation targeted the morphological form of a verb in a subject relative clause. Subjects either saw the correct form (4a,4b) or they saw a form that was missing the appropriate inflectional morphology (4c, 4d). Additionally, participants either read an object pronoun with the proper case-marking (4b,4d) or they read a pronoun with unlicensed nominative case-marking (4a,4c):

(4) a. The friend who **visited** Sue asked **she** whether the value of the house had dropped since the recession began. [**good-bad**]

b. The friend who **visited** Sue asked **her** whether the value of the house had dropped since the recession began. [**good-good**]

c. The friend who **visit** Sue asked **she** whether the value of the house had dropped since the recession began. [**bad-bad**]

d. The friend who **visit** Sue asked **her** whether the value of the house had dropped since the recession began. [**bad-good**]

72 filler items appeared along with the critical items.

Procedure Procedure was identical to Experiment I.

Results

Data were analyzed using the same methods as in Experiment 1. Outlier removal affected 1.2% of the data. The acceptability results indicate that the manipulations of both inflectional morphology ($\beta = -.415$, $t = -8.525$, $p < .0001$) and case ($\beta = -.624$, $t = -12.817$, $p < .0001$) had significant effects on acceptability judgments. There was also a statistically significant interaction ($\beta = .234$, $t = 2.402$, $p < .05$); however, the interaction differs from the interaction found in Experiment I. Here, it emerges because the case error produces lower judgments than the verbal inflection error. This interaction is *not* due to super-additivity, as it was in Experiment I; two errors yield acceptability decrements that are approximately the sum of decrements caused by sentences with each error in isolation.

In further contrast with the results from Experiment I, reading span scores do not show an overall significant linear relationship with acceptability z-scores. For the conditions judged the worst by participants, memory estimates actually exhibit a negative linear relationship with z-scores, i.e. individuals with higher reading span scores judged these conditions worse, compared to individuals with lower reading span scores. The difference between the conditions leads to a statistically reliable interaction of reading span score and the effect of the case manipulation ($\beta = -.098$, $t = -3.028$, $p < .01$).

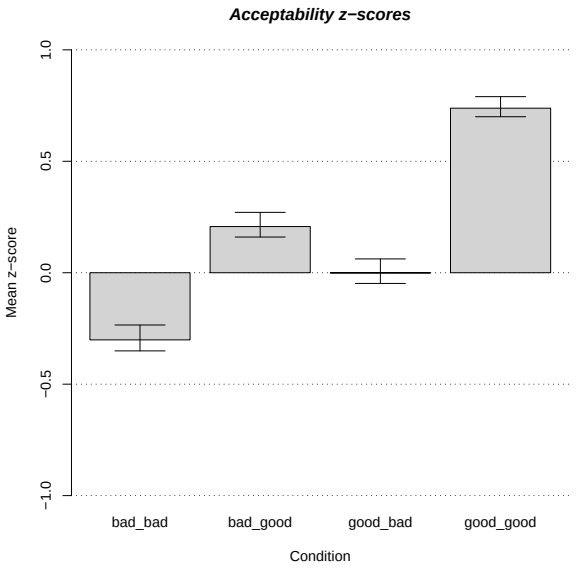


Figure 2: Acceptability z-scores for Experiment II. Error bars show (+/−) one standard error.

Discussion

Grammaticality violations appear to affect acceptability judgments in a qualitatively different way than processing manipulations. Most notably, grammaticality violations in this experiment combine additively—the effect of two co-occuring, proximal violations does not reduce judgments further than expected on the basis of each violation in isolation. These results align with independent evidence from Sorace and Keller (2005) that grammaticality violations combine additively.

The other important contrast between the first two experiments involves the relationship between reading span scores and acceptability scores. While we found a positive linear relationship between the two in Experiment I, in this experiment, reading span predicted *lower* judgments for the conditions judged worse (those with a case error). However, because the two types of manipulations were investigated in separate experiments, these high- and low-reading span participants were different individuals across experiments. In Experiment III, we directly compared the effects of grammaticality manipulations and processing manipulations in the same experiment and the same individuals.

Experiment III: Grammar and Processing

Participants This experiment was conducted in the same session as Experiment I, and involved the same 32

Stanford University students.

Materials Experiment III investigated how grammaticality violations and processing manipulations interact with one another. Experimental items appeared with either a correctly inflected verb (5a,5b) or incorrectly inflected verb (5c,5d). Dependency locality was utilized again to vary processing difficulty; the *wh*-dependencies in (5b) & (5d) are shorter than those in (5a) & (5c).

(5) a. They couldn't remember which lawyer that the reporter interviewed had defended the elderly man at the courthouse. **[hard-good]**

b. They couldn't remember which lawyer had defended the elderly man that the reporter interviewed at the courthouse. **[easy-good]**

c. They couldn't remember which lawyer that the reporter interviewed had defending the elderly man at the courthouse. **[hard-bad]**

d. They couldn't remember which lawyer had defending the elderly man that the reporter interviewed at the courthouse. **[easy-bad]**

The 24 experimental items were included alongside the materials from Experiment I and 48 additional fillers.

Procedure Procedure was identical to Experiments I and II.

Results

Data were analyzed using the same methods used in Experiments I and II. Removal of outliers affect 1.2% of the dataset. Results show that grammaticality significantly influences acceptability judgments ($\beta = .626$, $t = 15.583$, $p < .0001$). In contrast, the effect of processing difficulty on judgments is not statistically significant ($\beta = -.065$, $t = -1.628$, $p > .1$); however, there is a significant interaction between processing difficulty and grammaticality ($\beta = -.215$, $t = -2.680$, $p < .05$). As Figure 3 illustrates, this interaction arises because processing difficulty lowers judgments in sentences without grammatical violations, but it does not do so in sentences with grammatical violations.

While reading span does not emerge as a significant predictor for judgments across all condition types ($\beta = -.014$, $t = -.868$, $p > .1$), this seems to be because the grammatical and ungrammatical conditions pattern in different ways. The data reveal that individuals with higher reading span scores judge ungrammatical items worse, but in the grammatical conditions, better reading span performance predicts higher judgments of acceptability, leading to a significant interaction of reading span score and grammaticality ($\beta = .093$, $t = 4.986$, $p < .001$). In other words, estimates of memory capacity only show a positive linear relationship with judgments in the absence of grammar-based constraint violations.

Discussion

The results of Experiment III are consistent with our predictions and with the results of the first two experiments. Processing constraints and grammatical constraints combine sub-additively. Presumably, the grammaticality violations were so extreme as to "drown out" the effects of the processing manipulations in the ungrammatical conditions, i.e. a floor effect occurred. In general, if processing constraints and grammatical constraints reflect distinct and largely independent cognitive resources, super-additive combinations are unexpected. The present results support this hypothesis.

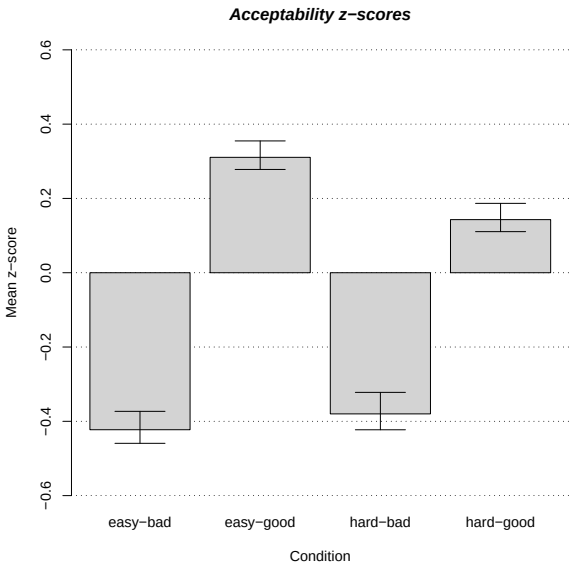


Figure 3: Acceptability z-scores for Experiment III. Error bars show (+/−) one standard error.

In addition, the relationships between reading span scores and judgments are as expected based on Experiments I and II: comprehenders with higher working memory scores find ungrammatical sentences worse, but difficult sentences better, compared to their low working memory counterparts. Experiment III shows that these contrasts can be observed even with the same set of subjects and minimally different items.

General Discussion and Conclusions

Three word-by-word acceptability judgment studies showed that (1) grammaticality violations combine additively, (2) differences that stem from functional constraints can combine super-additively with one another, and (3) grammaticality violations and processing manipulations can combine sub-additively with one another. These patterns suggest that when two constraints combine super-additively in acceptability decrements, it is likely that they are both functional constraints. Furthermore, participants' reading span scores predict sentence judgments differently for different types of manipulations. Participants with higher reading spans tend to judge ungrammatical sentences as being worse than their low-span counterparts do, yet they tend to judge difficult sentences as being better than participants with lower reading spans. These effects are not due to differ-

ences in comprehension accuracy, as this did not differ between conditions in any of the experiments.

It might be tempting to extend the findings of these experiments to the inverses of the relationships we have reported here; that is, if super-additivity indicates a processing contribution to an acceptability decrement, then does the absence of super-additivity rule out processing contributions? While it would be helpful for interpreting acceptability judgments if this were true, neither the present results nor general principles of language processing license this inference. It two sources of processing difficulty are sufficiently weak, they will not overtax the available resources, and should combine additively. Likewise, if two sources of processing difficulty were sufficiently extreme, the presence of just one might so overwhelm processing resources that the presence of the other was undetectable, resulting in a sub-additive combination.

It is also not possible to infer the inverse of the relationship we reported between reading span scores and processing difficulty. The lack of a positive linear relationship between reading span scores and acceptability judgments does not entail that the sentences do not cause processing difficulty. Further experiments not reported here involving center-embeddings show that reading span scores do not exhibit a positive linear relationship with acceptability judgments in the presence of massive processing difficulty. This could occur if language comprehension is not likely at a certain level of difficulty, and thus having greater language comprehension abilities might not produce better judgments. In other words, some stimuli may be so hard to process that virtually no one will have sufficient cognitive resources to understand the stimuli.

Given that these inverse inferences are not supported, tests of the functional origins of acceptability contrasts that seek to take advantage of the relationships of super-additivity and working memory capacity we demonstrate here must be designed accordingly. When super-additivity and/or positive linear relationships between acceptability and working memory measures are observed, however, these relationships will support conclusions that grammatical constraints are not necessary to account for the observed acceptability contrasts. This paper is a first step in what we hope will be a continuing process of developing criteria for establishing the role of functional constraints in acceptability contrasts – a necessary part of collecting and assessing evidence for grammatical theory-building.

Acknowledgments We gratefully acknowledge discussions and input from Daniel Casasanto and the assistance of David Kettler in conducting the experiments.

References

Baayen, R. (2004). Statistics in psycholinguistics: A critique of some current gold standards. *Mental Lexicon Working Papers*, *1*, 1–45.

Baayen, R. (2007). *Analyzing Linguistic Data: A Practical Introduction to Statistics.* Cambridge, UK: Cambridge University Press.

Bard, E., Robertson, D., & Sorace, A. (1996). Magnitude Estimation of Linguistic Acceptability. *Language*, *72*(1), 32–68.

Chomsky, N. (1973). Conditions on transformations. In S. Anderson & P. Kiparsky (Eds.), *A Festschrift for Morris Halle* (p. 232-86). New York: Holt, Reinhart & Winston.

Chomsky, N. (1986). *Barriers.* Cambridge: MIT Press.

Conway, A., Kane, M., Bunting, M., Hambrick, D., Wilhelm, O., & Engle, R. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin and Review*(12), 769–786.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*, 87–185.

Daneman, M., & Carpenter, P. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, *19*, 450–66.

Featherston, S. (2008). Thermometer judgments as linguistic evidence. In M. Claudia & A. Rothe (Eds.), *Was ist linguistische Evidenz?* Aachen: Shaker Verlag.

Forster, K., & Forster, J. (2003). DMDX: A Windows display program with millisecond accuracy. *Behavior Research Methods*, *35*(1), 116–124.

Grodner, D., & Gibson, E. (2005). Consequences of the serial nature of linguistic input for sentential complexity. *Cognitive Science*, *29*, 261–290.

Just, M., & Carpenter, P. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*(98), 122–149.

Kluender, R. (1998). On the distinction between strong and weak islands: a processing perspective. In P. Culicover & L. McNally (Eds.), *Syntax and Semantics 29: The Limits of Syntax* (p. 241-279). San Diego, CA: Academic Press.

Miller, G. A., & Chomsky, N. (1963). Finitary models of language users. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (Vol. 2, p. 419-492). New York: Wiley.

Ross, J. R. (1967). *Constraints on Variables in Syntax.* PhD Thesis. MIT.

Sorace, A., & Keller, F. (2005). Gradience in linguistic data. *Lingua*, *115*(11), 1497–1524.

The Role of Linguistic Labels in Categorization

Wei (Sophia) Deng (deng.69@osu.edu)
Center for Cognitive Science
The Ohio State University
209C Ohio Stadium East, 1961 Tuttle Park Place
Columbus, OH 43210, USA

Vladimir M. Sloutsky (sloutsky.1@osu.edu)
Center for Cognitive Science
The Ohio State University
208C Ohio Stadium East, 1961 Tuttle Park Place
Columbus, OH 43210, USA

Abstract

Do young children treat labels as features or as category markers? The current study addressed this question by examining the effect of labels on young children's classification and induction. The first experiment replicated previous study on adults demonstrating that adults treat labels as category markers. The other two experiments applied the same paradigm to young children. Children were trained by classification in Experiment 2A and by induction in Experiment 2B, whereas both experiments used the classification and induction tasks that were identical to those in Experiment 1. The results from the three experiments indicated that adults treated labels as category markers, whereas no such evidence was found for young children.

Keywords: Cognitive Development, Classification, Induction, Label, Psychology, Human Experimentation.

Introduction

The ability to use linguistic labels to generalize from the known to the unknown is crucial for learning new information. Although a substantial body of experimental evidence has demonstrated that label has an impact on categorization and induction processes (Gelman & E. Markman, 1986; Sloutsky, Lo, & Fisher, 2001; Welder & Graham, 2001; Yamauchi and A. Markman, 1998, 2000), the mechanism underlying the role of labels is hotly debated. Are labels used as features (similar to other objects' properties) or as category markers representing category membership? This issue is particularly contentious with respect to the role of labels in early development and developmental changes in this role.

Some researchers have argued that labels are more than features. According to this view, labels are category markers used for representing a category. For example, E. Markman and Hutchinson (1984) found that children regarded words presented as count nouns changed the way young children grouped objects. Without labels children grouped objects thematically (e.g., a police car was grouped with a policeman), whereas when the same police car was referred to by a count noun, children grouped objects taxonomically (e.g., the police car with a passenger car). Gelman and Heyman (1999) demonstrated that young children were more willing to generalize properties from one person to another when both persons were referred to by a noun (i.e., "carrot-eaters") than when both were referred to by a descriptive sentence (e.g., "both like to eat carrots").

This evidence, however, does not lend unequivocal support to the idea that labels have to be category markers to make inductive inferences. For example, some researchers suggested that contribution of labels is driven by attentional rather than conceptual factors (Napolitano & Sloutsky, 2004; Sloutsky & Napolitano 2003). There is also evidence that labels contribute to the overall similarity of compared entities and thus to both categorization and induction (Sloutsky & Fisher, 2004). Sloutsky and Fisher (2004) also demonstrated that similarity computed over labels and appearances can accurately predict young children's responses with the Gelman and E. Markman (1986) task. These findings suggest that reliance on labels does not necessarily indicate that labels are more than features.

In a series of studies, Yamauchi and A. Markman (1998, 2000) designed a paradigm that could address this issue directly. Specifically, they compared participants' performance on classification tasks (e.g., is X a dax?) with that on induction tasks (e.g., given that X is a dax, does it have Y?). The tasks are structurally identical, except a critical difference. In the classification task participants predicted the category label of an item given all of its feature values. In contrast, in the induction task, participants predicted the value of a missing feature of an item given its category label and other feature values. These researchers argued that if the label is a feature then performance on classification and induction task should be symmetrical. However, if labels are more than features, then performance on induction tasks should be better than performance on classification

tasks. Upon finding predicted asymmetries between the two conditions (i.e., participants were better at using the label to predict other features than at using other features to predict the label), these researchers concluded that participants are more likely to regard labels as category markers instead of object features.

However, this paradigm has not been applied to children. Does the asymmetry found in adults exist in children? Finding such an asymmetry would support the idea that labels are more than category features, whereas a symmetric performance in the classification and induction conditions would support the idea that labels are features. The primary goal of this study is to address these questions.

The current study consists of three experiments. Experiment 1 replicated Yamauchi and A. Markman's paradigm (2000) with adults. Based on their findings, it was hypothesized that adults would regard labels as category markers. Experiments 2A and 2B, using comparable learning and testing conditions, examined how labels would affect young children's performance.

Experiment 1

Method

Participants Sixteen adults participated in this experiment. Participants were undergraduate student from the Ohio State University participating for course credit. Three of them gave one type of response to over 95% of all trials. Their data were excluded from the analysis due to the response bias.

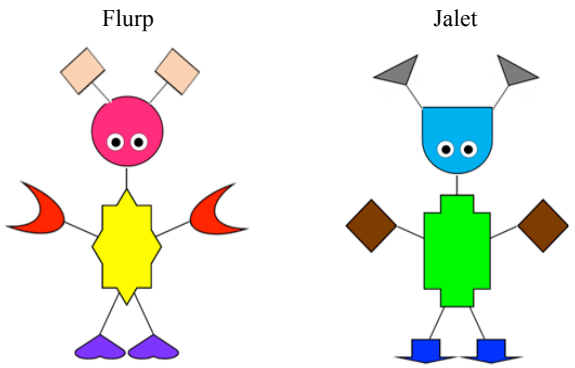


Figure 1. The prototypes of stimuli used in this study.

Stimuli The stimuli were artificial creatures accompanied by a category label ("Flurp" of "Jalet") and two categories of objects that were created using five features varying in color and shape (see Figure 1). As shown in Table 1 and 2, the two categories have a family-resemblance structure, which is derived from two prototypes (F0 and J0) by modifying the values of one of five features. For example, to produce the stimulus F1, the value of the antenna is changed from 1 to 0 so that it has four features consistent with the prototype F0 and one feature consistent with the prototype J0. The degree of similarity between test stimulus and the prototype is defined by the number of matching features of the test stimulus to the prototype of the corresponding category (Table 1 and 2).

Table 1. Category structure used in learning.

Flurp							Jalet						
Stimuli	Head	Body	Hands	Feet	Antenna	Label	Stimuli	Head	Body	Hands	Feet	Antenna	Label
F1	1	1	1	1	0	1	J1	0	0	0	0	1	0
F2	1	1	1	0	1	1	J2	0	0	0	1	0	0
F3	1	1	0	1	1	1	J3	0	0	1	0	0	0
F4	1	0	1	1	1	1	J4	0	1	0	0	0	0
F5	0	1	1	1	1	1	J5	1	0	0	0	0	0
F0	1	1	1	1	1	1	J0	0	0	0	0	0	0

Note. The value 1 = any of five dimensions identical to "Flurp" (see Figure 1). The value 0 = any of five dimensions identical to "Jalet" (see Figure 1). F = Flurp; J = Jalet. F0 and J0 are prototypes of each category.

Table 2. Stimulus structure used in testing.

Flurp								Jalet						
Stimuli	Head	Body	Hand	Feet	Antenna	Target Label	Match	Stimuli	Head	Body	Hand	Feet	Antenna	Target Label
F11	1	1	1	1	0	1		J11	0	0	0	0	1	0
F12	1	1	1	0	1	1		J12	0	0	0	1	0	0
F13	1	1	0	1	1	1	High	J13	0	0	1	0	0	0
F14	1	0	1	1	1	1		J14	0	1	0	0	0	0
F15	0	1	1	1	1	1		J15	1	0	0	0	0	0
F21	1	0	1	0	0	1		J21	0	1	0	1	1	0
F22	0	1	0	1	0	1		J22	1	0	1	0	1	0
F23	0	0	1	0	1	1	Low	J23	1	1	0	1	0	0
F24	1	0	0	1	0	1		J24	0	1	1	0	1	0
F25	0	1	0	0	1	1		J25	1	0	1	1	0	0

Note. High and low are two levels of feature match. F = Flurp; J = Jalet. Category-accordance responses were the ones consistent with the values indicated in the target features and target labels.

Similar to Yamauchi and A. Markman, there are two levels of similarity (or feature match) in current research: high and low. At the high level of feature match, each test stimulus has four features in common with the prototype of the corresponding category and one feature in common with the prototype of the contrasting category. Similarly, each test stimulus at the low level of feature match has two features in common with the prototype of the corresponding category and three features in common with the prototype of the other category.

Procedure The entire experiment consisted of two phases, learning and testing. During the learning phase, participants were instructed that they should try to remember and distinguish two groups of artificial creatures represented by the labels "Flurp" and "Jalet". And then participants were presented with 36 trials of creatures produced from stimulus structure shown in Table 1 and each stimulus had a correspondent label above it.

The testing phase was administered immediately after the learning phase. The Classification and Induction conditions differed in the type of features being predicted. In the Classification condition, participants predicted the category label of a stimulus given information about all five features with the label covered. In the Induction condition, participants predicted the value of one of five features given the other four features with the label uncovered. The classification question was phrased as "Which group do you think this creature is more likely to belong to, Flurp or Jalet?" The induction question was phrased as "Which antenna do you think this creature is more likely to have?" The order of the testing trials was randomized for each subject. Feedback was given in first 6 trials of each condition. No feedback in other 40 trials in both conditions. The proportion of responses in accordance with the category from which the exemplar was derived (called "category-accordance responses" by Yamauchi & A. Markman, 2000, see Table 2) was the dependent variable.

Results and Discussion

The main results of Experiment 1 are shown in Figure 2. The data were analyzed with 2 (testing type: Classification and Induction) × 2 (feature match: high and low) analysis of variance (ANOVA). There was a main effect of feature match, $F(1,12) = 165.39$, $MSE = 0.89$, $p < 0.01$, as well as an interaction between testing type and feature match, $F(1,12) = 38.61$, $MSE = 0.32$, $p < 0.01$. At the low level of feature match, category-accordance responses made in the Induction condition were more than in the Classification condition, $t(12) = 4.88$, $p < 01$. However, there was no significant difference in these two testing types at the high level of feature match, $t(12) = 2.12$, $p > 0.05$.

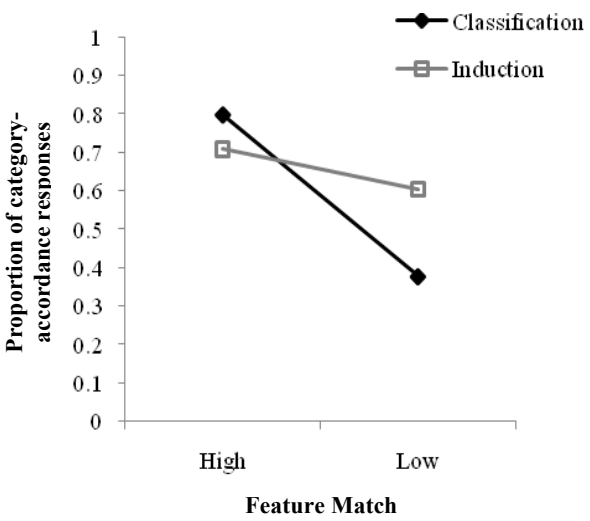


Figure2. Performance for classification and induction tasks in Experiment 1.

The results replicate Yamauchi and A. Markman (2000) pointing to the predicted asymmetry and suggesting that for adults labels are more than objects features. In Experiment 2, we expand this paradigm to young children.

Experiment 2A

Method

Participants There were thirteen preschool children (6 boys and 7 girls) with an average age of 55.8 months participating in this experiment. They were given Classification learning. In Classification learning, children were presented with all five features of a creature and told that it was a Flurp (or Jalet). A memory check was administered after main experiment to examine whether participants could remember the stimuli and correspondent labels. Children were given 5 trials in memory check by presenting a creature and asking them which group this creature came from. One of them answered less than 3 out 5 memory check questions correctly and these data were excluded from the analysis.

Stimuli and procedure The visual stimuli were identical to Experiment 1 (see Figure 1). The entire experiment consisted of two phases, classification learning phase and testing phase. During classification learning, in contrast to Experiment 1 with adults, children were instructed that there were two groups of creatures, Flurp and Jalet. And then they were trained by presenting creatures with category labels and told:" This is a Flurp (or Jalet)."

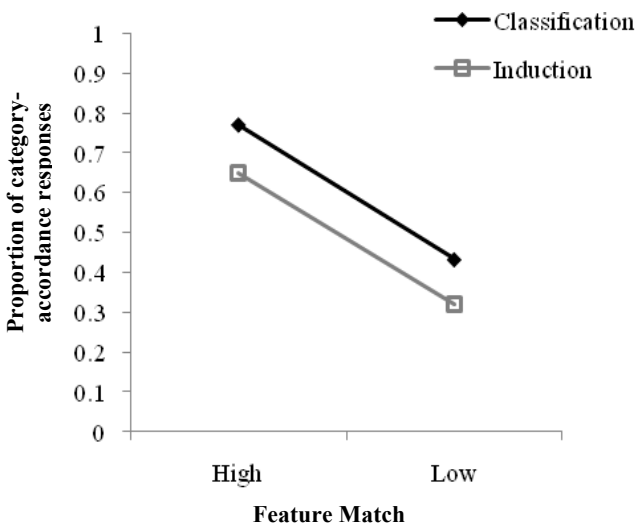


Figure 3. Performance for classification and induction tasks in Experiment 2A.

The testing phase was identical to Experiment 1 except how the questions were asked. Unlike the adults participants in Experiment 1 who read the questions presented on screen, children were asked both classification and induction questions by a female experimenter.

Results and Discussion

The main results of Experiment 2A are shown in Figure 3. The data were analyzed with 2 (testing type: Classification and Induction) × 2 (feature match: high and low) analysis of variance (ANOVA), with testing type and feature match as within-subjects factors. There was a main effect of feature match, $F(1,11) = 43.56$, $MSE = 1.33$, $p < 0.01$, as well as a main effect of testing type, $F(1,11) = 14.77$, $MSE = 0.16$, $p < 0.01$. However, unlike adults in Experiment 1, there was no significant interaction between testing type and feature match, $F(1,11) = 0.02$, $MSE = 0.00$, $p > 0.10$.

Children in Experiment 2A, unlike adults in Experiment 1, made more category-accordance responses on classification questions than on induction questions at both high and low level of feature match. These results indicate that, if anything, the Classification condition elicited better performance than the Induction condition. These results present little evidence that young children treated labels as category markers.

However, the learning type, in this experiment using classification learning, might have a facilitative effect on children's classification. In Experiment 2B, we explored the impact of learning type on children's performance in the classification and induction tasks by training them with induction instead of classification task.

Experiment 2B

Method

Participants There were fourteen preschool children (8 boys and 6 girls) with an average age of 54.00 months in this experiment. They were given Induction learning. In contrast to the children trained by classification in Experiment 2A, children in this experiment were trained by induction in which they were presented all five features of a creature and told that the creature had a Flurp (or Jalet) inside its body. A memory check, identical to Experiment 2B, was administered after main experiment to examine whether participants could remember the stimuli and correspondent labels. Two of them answered less than 3 out 5 memory check questions correctly and these data were excluded from the analysis.

Stimuli and procedure The visual stimuli were identical to previous experiments. The entire experiment consisted of two phases, induction learning phase and testing phase. During induction learning, in contrast to Experiment 2A, children were instructed that there were two groups of creatures and something special was inside each group of creatures. One group of creatures had Flurp while another group had Jalet. And then they were trained by presenting creatures and told:" This one has a Flurp (or Jalet)." The testing phase was identical to Experiment 2B.

Results and Discussion

The main results of Experiment 2B are shown in Figure 4. The data were analyzed with 2 (testing type: Classification and Induction) × 2 (feature match: high and low) ANOVA, with testing type and feature match as within-subjects factors. There was a main effect of feature match, $F(1,16) = 86.84$, $MSE = 2.50$, $p < 0.01$. In contrast to Experiment 2A, children in this experiment did not differ in the two testing types, $F(1,11) = 3.90$, $MSE = 0.03$, $p > 0.05$.

These results, compared to Experiment 2A, suggest that there was an effect of learning type and the induction learning facilitated children's performance on the induction questions. However, children's performance, similar to Experiment 2A, was symmetric in both testing conditions and there was no evidence that children treated differently induction and classification questions.

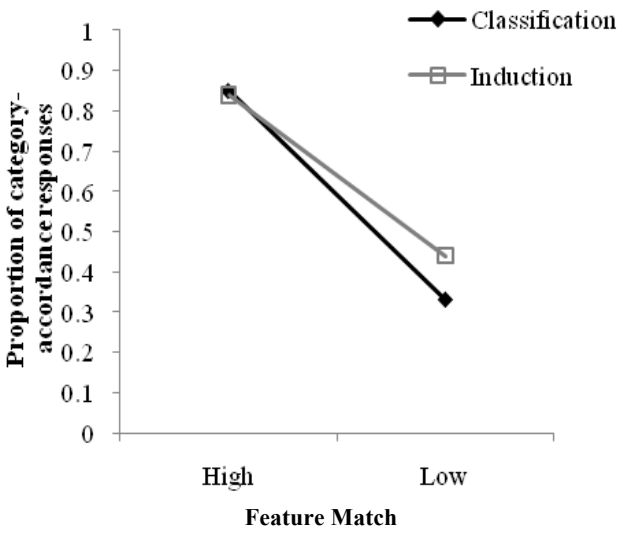


Figure 4. Performance for classification and induction tasks in Experiment 2B.

General Discussion

The results point to two main findings. First, there was an effect of learning. Children had better performance on classification questions when they were trained by classification (Experiment 2A). At the same time, when they were trained by induction (Experiment 2B), their performance on induction questions became equivalent to that on classification questions and there was no significant difference between these two testing types.

And more importantly, there were marked developmental differences in the role of linguistic labels. Adults exhibited better performance in inferring a feature by using a label than inferring a label by using features. These findings are consistent with previous research (Yamauchi & A. Markman, 2000) and suggest that adult may have used labels as category markers. However, labels had little facilitative effect on children's performance – their performance was equivalent whether they were asked to predict labels on the basis of other features (i.e., the classification condition) or to predict a feature on the basis of the label (i.e., the induction condition). Furthermore, regardless of the type of learning (i.e., Classification or Induction), children, in contrast to adults' asymmetry, consistently exhibited a symmetric pattern on classification and induction questions (see Figure 3 and 4). These results suggest that while labels may be different from category features for adults, this is not the case for young children.

These results have important implications for understanding of inter-relationships between language and cognition, and specifically the role of linguistic labels in categorization and category learning. Recall that according to some accounts, even early in development linguistic labels words affect categorization and inductive inference by marking the underlying category (e.g., Gelman & Heyman, 1999; Gelman & Markman, 1986). According to other accounts, early in development linguistic labels are features of entities. As a result, when two entities share a label, young children may perceive these entities as being more similar than when no labels are introduced (Sloutsky, et al, 2001; Sloutsky & Fisher, 2004). Yamauchi and A. Markman (1998, 2000) developed a procedure enabling the distinction between these accounts. This procedure, however, was never used with young children.

Our results with young children support the latter account, while generating little evidence that young children treat labels as category markers. In addition, although current research does not conclusively eliminate the possibility that for young children linguistic labels are category markers, it demonstrates that the role of linguistic labels changes in the course of development.

Acknowledgments

This research has been supported by grants from the grants from the NSF (BCS-0720135), from the Institute of Education Sciences, U.S. Department of Education (R305B070407), and from NIH (R01HD056105) to Vladimir M. Sloutsky.

References

Gelman, S. A., & Heyman, G. D. (1999). Carrot-eaters and creature-believers: The effects of lexicalization on children's inferences about social categories. *Psychological Science, 10,* 489 – 493.

Gelman, S. A., & Markman, E. (1986). Categories and induction in young children. *Cognition, 23,* 183-209.

Markman, E. M., & Hutchinson, J. E. (1984). Children's sensitivity to constraints on word meaning: Taxonomic versus thematic relations. *Cognitive Psychology, 16,* 1 – 27.

Napolitano, A. C., & Sloutsky, V. M. (2004). Is a picture worth a thousand words? The flexible nature of modality dominance in young children. *Child Development, 75,* 1850 – 1870.

Sloutsky, V. M., & Fisher, A. V. (2004). Induction and categorization in young children: A similarity-based model. *Journal of Experimental Psychology: General, 133,* 166 – 188.

Sloutsky, V. M., Lo, Y.-F., & Fisher, A. V. (2001). How much does a shared name make things similar: Linguistic labels and the development of

inductive inference. *Child Development, 72,* 1695 – 1709.

Sloutsky, V. M., & Napolitano, A. C. (2003). Is a picture worth a thousand words? Preference for auditory modality in young children. *Child Development*, *74*, 822 – 833.

Welder, A., & Graham, S. (2001). The influences of shape similarity and shared labels on infants' inductive inferences about nonobvious object properties. *Child Development, 72,* 1653 – 1673.

Yamauchi, T., & Markman, A. B. (1998). Category-learning by inference and classification. *Journal of Memory and Language, 39,* 124-148.

Yamauchi, T., & Markman, A. B. (2000). Inference using categories. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *26,* 776-795.

The Price is Right: A High Information Access Cost Facilitates Category Learning

Michael J. Wood (mw337@kent.ac.uk)
Michael Fry (mdf1@sfu.ca)
Mark R. Blair (mblair@sfu.ca)
Cognitive Science Program & Department of Psychology
Simon Fraser University, 8888 University Drive, Burnaby, BC, V5A 1S6, CANADA

Abstract

Previous work in object categorization has shown that people tend to optimize their allocation of attention to object features, and suggests that attentional optimization may best be explained in terms of cost-benefit tradeoffs. In support of this idea, we found that implementing a cost for accessing information about object features in a category learning task facilitates both attentional optimization and category acquisition, contrary to the predictions of existing models.

Keywords: category learning; categorization; access cost; attentional learning; optimization.

An important component of proper psychological functioning is the adaptive usage of limited resources. In many situations, careful conservation of money, food, water, memory capacity, and time can be vital to survival. The choice of strategy for dealing with a particular issue depends largely on the relative availability of the various resources required – for instance, installing hardwood floors might be best accomplished by doing it oneself if money is tight and time is plentiful, while hiring a contractor might be a better idea if money is no object but the job must be done quickly. An optimal strategy for a given problem, then, balances situational priorities (urgency, desire for quality) with available resources (time, money).

This characterization of optimal strategy applies equally to psychological domains such as categorization. Fiske and Taylor (1984) characterized humans as cognitive misers, meaning we will attempt to solve problems using the smallest amount of mental resources possible. Indeed, a good deal of evidence suggests that in category learning, people learn to ignore irrelevant information, thereby optimizing their allocation of attention for the task at hand (Rehder & Hoffman, 2005; Blair, Watson, & Meier, 2009; Blair, Watson, Walshe, & Maj, 2009; Blair, Chen, et al., 2009). The process of selectively allocating attentional resources to task-relevant information is labeled *attentional optimization*.

One approach to characterizing the optimal usage of attentional resources takes the view that the benefits of attending to a piece of information must outweigh the costs. This view of attentional optimization as a process of cost-benefit tradeoffs parallels some of the decisions made in the domain of medical diagnosis. A doctor attempting to diagnose a patient will order only tests which are necessary, and even then will strike a balance between efficacy, cost, and safety. A doctor who suspects a particular condition may be more likely to order a cheap, safe blood test than expensive, dangerous exploratory surgery.

It is not yet clear which resources are conserved as a result of attentional optimization. One candidate is working memory capacity: unattended object features are unlikely to be stored in memory. There is, in fact, evidence that working memory shares a close relationship with category learning: those with low working memory spans are less able to suppress task-irrelevant information, creating a need for selective attention to fill the gap (Conway, Kane, & Engle, 2003). Another resource that may be conserved is time – attending only to what is necessary is likely to result in a reduction in the amount of time required to categorize something. In either view, attending to a particular feature of an object incurs a cost – whether temporal or mnemonic – and attentional optimization minimizes the cost incurred for a successful categorization.

Hayhoe and colleagues (Ballard, Hayhoe, & Pelz, 1995; Ballard, Hayhoe, Pook, & Rao, 1997; Droll & Hayhoe, 2007) provided empirical evidence of cost-benefit tradeoffs in visual perception. When performing a task along the lines of the Blocks World game (an interactive paradigm in which subjects must duplicate a target image by positioning a group of coloured boxes from a resource pool), participants generally gather information from the environment as they need it, minimizing the usage of short-term memory. However, increasing the predictability of the task encourages participants to save time by storing information in memory: time, rather than memory capacity, becomes the focus of their conservation efforts. Similar results were found in a series of studies by Gray and colleagues, many of which also employed the Blocks World paradigm (Gray & Fu, 2004; Gray, Sims, Fu, & Scholles, 2006; Fu & Gray, 2006). When the target window was occluded by a removable square, or participants were forced to make head movements in addition to eye movements in order to direct their gaze about the work area, there tended to be a switch to a memory-intensive strategy in order to save time and energy.

It appears that saving time is not the only advantage of adopting a memory-based strategy in tasks along the lines of Blocks World, although the evidence is not unequivocal. Gray et al. (2006) found that participants in a memory-intensive variation of the Blocks World task made fewer errors and mastered the task sooner than others performing a standard task. While this contradicted earlier results by Gray and Fu (2004), the results of Gray et al. (2006) are supported by the research of Waldron and colleagues on interface design (Waldron, Patrick, Howes, & Duggan, 2006; Waldron, Patrick, Morgan, & King, 2007; Morgan, Patrick, Waldron, King, & Patrick, 2009). As in the Blocks

World literature, Waldron and colleagues found that an increased information access cost leads to a change in information-gathering strategy in a variety of different paradigms. Implementing a time delay for accessing information on the target encourages the usage of memorization, in contrast to the default strategy of scanning back and forth (Waldron et al., 2006). This strategic shift was found to be beneficial to memory for particular system states, general understanding of the system, competence in the absence of available information (Waldron et al., 2007), and the ability to fluently resume a task after interruption (Patrick et al., 2009), though it has its costs in the form of increased response time (Waldron et al., 2006).

While the above research suggests that there are some benefits (and some penalties) resulting from a shift toward memory-based strategies in response to increased information access costs, it is not yet clear whether increased attentional optimization is one of them. None of the studies by Waldron and colleagues involved the presence of irrelevant information. This is not surprising, as interface design tends to avoid including irrelevant data in a display; however, in object categorization it is often vitally important to be able to divert one's attention away from unimportant information (e.g. Rehder & Hoffman, 2005).

In spite of the evidence regarding the importance of cost-benefit considerations in the allocation of attention, it is possible (and, until recently, routine) to develop a coherent model of attentional optimization without making any mention of costs. Computational models of category learning, such as ALCOVE (Kruschke, 1992), simply shift attention away from irrelevant information and towards relevant information. However, in a disease diagnosis paradigm, Matsuka and Corter (2008) found that participants appeared optimize attention in a way consistent with a sensitivity to cost-benefit considerations. When presented with stimuli with two different features which perfectly and redundantly predicted category membership, people attended to only one of them. The idea of attentional optimization as a cost-benefit tradeoff explains this result quite well: the benefit of viewing one feature far outweighs the cost of accessing it, while the second feature provides no additional information to offset its access cost and is thus ignored.

If attentional optimization is indeed based partially on cost-benefit considerations, then the degree to which people optimize their attention should depend on the additional cost incurred in attending to irrelevant features. A high information access cost should provide more motivation to avoid the waste of time or resources associated with attending to irrelevant information, increasing the rate of attentional optimization. In contrast, optimizing one's attention would provide only a minimal benefit in a situation in which accessing information is nearly or entirely free, and as such may be less of a priority for those who are able to master the task.

All other things being equal, then, a category learning task with a high information access cost should result in more attentional optimization than a task with a low or nonexistent access cost. In addition, implementing a high access cost should encourage the use of a memory-based strategy, resulting in improved learning. The present experiment sought to test these hypotheses using the stimuli and category structure from Experiment 2 of Blair, Watson, Walshe, and Maj (2009). Since the stimuli involved three spatially separate features, we were able to manipulate access cost by obscuring the features with overlays and implementing a variable time cost to remove them.

Method

Participants

149 undergraduate students from Simon Fraser University students participated in exchange for course credit in introductory Psychology classes.

Apparatus

The computer program used in the present experiment was developed using E-Prime 1.1 (Psychology Software Tools), and was run on four Apple iMac computers running Windows XP. Responses were made using the computer mouse.

Design

The present experiment consisted of a supervised category learning task. Participants were shown computer-generated pictures of fictitious microorganisms (see Figure 1) and asked to categorize them as members of one of four different species. The microorganisms (following Blair, Watson, Walshe, and Maj, 2009), varied on three binary organelle-like features, each located in a distinct area of the cell. One organelle looked like either a muscle or a thin tube, another was a mitochondrion-like structure with either

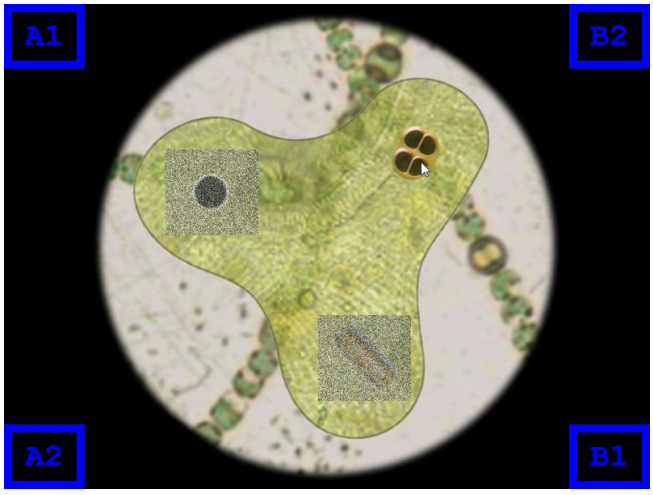


Figure 1: A sample microorganism stimulus with response buttons in the corners of the screen. The subject has revealed the top right feature by moving the mouse over it, while the other two are still occluded by overlays.

Table 1: Sample category structure.

Feature 1	Feature 2	Feature 3	Category
1	0	Irrelevant	A1
1	1	Irrelevant	A2
0	Irrelevant	0	B1
0	Irrelevant	1	B2

two or four internal compartments, and the third resembled an iris and pupil with either a green or a brownish coloration. Each feature occupied its own lobule of the cell, evenly distributed around the screen and counterbalanced across subjects.

There were four possible category labels for each stimulus: A1, A2, B1, and B2. One feature was always relevant, and determined whether the stimulus was a member of an A category or a B category. Of the two remaining features, one determined whether an A stimulus was A1 or A2, and the other determined whether a B stimulus was B1 or B2. Thus, only one of the two was relevant on any given trial, and the identity of the first feature informed the participant of which of the other features would be relevant (see Table 1). Feature relevance was counterbalanced across subjects, and category labels were assigned randomly according to the structure described above.

Procedure

Following a brief introduction to the experimental task, participants began a series of supervised categorization trials. A stimulus was presented, with its three variable features covered up by noisy square-like overlays. In order to remove an overlay and see the feature underneath, participants were required to hold the mouse on top of it for a predetermined period of time.

Participants were randomly assigned to a high-cost or low-cost condition. In the no-delay (low-cost) condition, the overlays disappeared instantly; in contrast, participants in the delay (high-cost) condition had to hold the mouse on top of an overlay for a full 3000ms before the feature was revealed. During this interval, the overlay was replaced by a black box marked "SCANNING…" In either condition, upon moving the mouse away from the revealed feature, the overlay would instantly reappear. The position of the mouse was tracked and recorded over the course of the experiment. Thus, at most one feature was available for viewing at one time. This allowed for a sensitive and dynamic measure of attentional allocation, similar to that of eye-tracking, and prevented participants in the delay condition from using the additional wait time to inspect other features.

Immediately after participants responded, they were presented with corrective feedback and were able to re-inspect stimulus features, with the same overlay restrictions as before, if they so desired.

By default the experiment lasted for 200 such categorization trials. An early learning criterion was implemented such that participants who learned the category

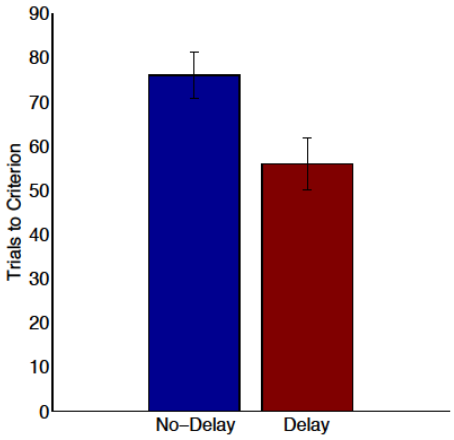


Figure 2: Mean number of trials taken to reach the learning criterion by condition. Error bars represent SEM.

structure well enough to provide 25 consecutive correct answers immediately proceeded to a 72-trial transfer phase where corrective feedback was not provided. Those who were unable to reach this criterion point by the 200th trial did not proceed to transfer. There was no time restriction on the experiment; participants were free to spend as long as desired on each trial. While there was some individual variation in completion times, the entire experiment took approximately 45 minutes to complete.

Results

20 participants were excluded due to computer errors or random responding, leaving 65 participants in the no-delay condition and 64 in the delay condition. The no-delay condition produced 45 learners and 20 non-learners, compared to 49 learners and 15 non-learners in the delay condition. This did not constitute a statistically significant difference, $\chi^2(1) = .877$, $p > .30$. Among those who were able to learn the category structure, however, participants in the delay condition reached criterion accuracy significantly earlier (M = 57.5 trials) than those in the no-delay condition (M = 76.2), $t(92) = 2.37$, $p < .05$ (see Figure 2).

We calculated attentional optimization scores for each trial following the formula used in the eye-tracking experiments of Blair, Watson, Walshe, and Maj (2009):

$$\frac{\bar{X}_{relevant} - \bar{X}_{irrelevant}}{\bar{X}_{relevant} + \bar{X}_{irrelev\ ant}}$$

This amounts to a comparison of the average length of time spent attending to relevant versus irrelevant features, where $\bar{X}_{relevant}$ is the total time during which relevant features were visible divided by the number of relevant features and $\bar{X}_{irrelevant}$ is the total time during which irrelevant features were visible divided by the number of irrelevant features. Our measure of attentional optimization thus ranged from -1 (fixating only irrelevant features) to 0

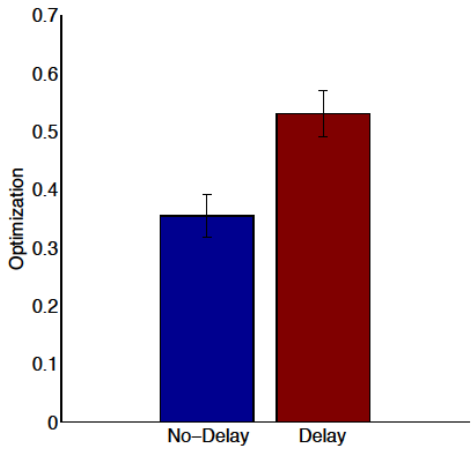


Figure 3: Mean optimization in the first 110 trials by condition. Error bars represent SEM.

(equal time spent fixating all features) to 1 (fixating only relevant features). Over the course of the experiment, subjects in the delay condition displayed significantly greater optimization (M = 0.623) than those in the no-delay condition (M = 0.520), $t(92) = 2.193$, $p < .05$.

Since the delay condition resulted in faster learning and thus an earlier end to the experiment, we elected to examine optimization between conditions over a set number of trials as an alternative comparison between conditions. We calculated the mean optimization over the first 110 trials (the approximate length of the shortest experimental run) for each successful learner. In this trial range, participants in the delay condition displayed a mean optimization score of .531 (SD .281), compared to .355 (SD .245) for the no-delay condition. This was a significant difference, $t(92) = 3.34$, $p < .01$ (see Figure 3).

As attentional optimization has been known to occur after categorization errors cease (Blair, Watson, & Meier, 2009), it is possible that the contribution of access cost to optimization in the first 110 trials was purely a product of the earlier learning criterion in the delay condition. With more error-free time to optimize, a greater degree of optimization would not be surprising. To examine whether access cost had an effect on optimization independent of its contribution to early accuracy, we performed a mediation analysis using the hierarchical multiple-regression techniques described by Baron and Kenny (1986). Having already demonstrated an association between access cost and criterion point (see Figure 2) and a connection between access cost and optimization in the first 110 trials (see Figure 3), we performed a hierarchical multiple regression analysis predicting early optimization from criterion point in the first step, and adding delay condition as a new predictor in the second. Criterion point proved to be a significant negative predictor of optimization, $\beta = -.777$, $t(92) = -11.85$, $p < .001$, as expected. Condition, when added to the model, contributed to optimization even after controlling for criterion point, $\beta = .140$, $t(91) = 2.11$, $p < .05$, indicating a partial-mediation relationship. Access cost contributed to optimization both indirectly (via earlier learning) and directly.

Finally, we suspected that the time course of optimization may have differed between conditions – it is possible that a long delay encouraged earlier optimization, but participants in the no-delay condition may have caught up later on in the experiment. To investigate this possibility, we calculated each learner's mean optimization scores before and after their criterion point. A 2 (pre-criterion/post-criterion) x 2 (delay/no-delay) mixed ANOVA revealed no interactive effect of stage and delay on optimization, $F(1,92) = .106$, $p > .70$, suggesting that attentional learning was uniform over the course of the experiment in both conditions.

Discussion

The results of the present work indicate that increasing the temporal cost of accessing information contributes not only to improved category learning, but also to more optimal allocation of attention. Learners in the high-cost delay condition reached the learning criterion earlier than those in the no-delay condition, and displayed greater attentional optimization over the course of the experiment. These findings support the counterintuitive idea that making information access more difficult improves multiple aspects of performance, extending earlier findings in disease-diagnosis (Matsuka & Corter, 2008) and interface design (Waldron et al., 2007). Taken together, this body of research provides compelling evidence for the validity of the conception of attentional optimization as a balancing act between costs and benefits.

In addition to cost-benefit considerations, one potential contributor to the improved learning in the presence of a high temporal access cost is the fact that such a cost encourages a strict sequential progression of attention. In recalling the positions of objects in space (Yamamoto & Shelton, 2009), as well as in recalling lists of words or letters (Frick, 1985; Goolkasian, Foos, & Krusemark, 2008), performance is significantly improved when information is presented sequentially rather than simultaneously. When access cost is low, participants are able to switch their attention back and forth as they please; in contrast, a high access cost discourages jumping back and forth between costly pieces of information and promotes a strategy of sequential attention.

Somewhat unexpectedly, in spite of the earlier criterion point among learners in the high-cost condition, there was not a concomitant difference in the number of learners. This may be an issue of motivation: while the increased access cost appears to facilitate learning by encouraging the use of memory-based strategies, participants in the delay condition may have become frustrated with the inconvenience of having to wait for features to become visible and applied less effort as a result. This possibility may be a fruitful topic for future research. Further investigation in this area may also benefit from some variance in the number of trials given to reach criterion; in the present study, participants in

both conditions were given 200, a number far in excess of the mean number of trials to criterion (58 for delay, 76 for no-delay).

The practical implications of the present research for training in interface design and related fields are obvious: implementing an access cost can in certain circumstances facilitate learning. However, caution should be taken, as a high temporal access cost can greatly reduce the temporal efficiency of a training period. While a subject might learn a particular system in fewer trials with a high access cost, the cost may make each trial so long that the net effect is ultimately more time spent on training. If it holds true that the learning advantage that comes with an increased access cost is largely the product of a shift toward memory-based strategies, there is probably a point beyond which increasing access cost confers no additional benefit. In addition, there may be more practical ways of encouraging the adoption of memorization strategies, such as only presenting information for a short period of time (Waldron et al., 2006) or implementing a non-temporal cost, such as money, tokens, or effort.

Within the field of category learning, researchers have long focused on tasks where all of the relevant information is immediately and simultaneously available to categorizers. Learning, according to the major models, is in most cases exclusively based on the accuracy of the response (e.g. ALCOVE; Kruschke, 1992). This is because they were designed around a specific event – the categorization trial – rather than around the dynamic unfolding of the task through time (though see Lamberts, 2002). While this has been a helpful simplification, it is becoming increasingly untenable in the face of dynamic measures of attention such as eye- and mouse-tracking (Rehder & Hoffman, 2005; Blair, Watson, Walshe, & Maj, 2009), as well as a number of results indicating a level of complexity untouched by the current generation of computational models. The time spent waiting for stimuli to appear can have implications for strategy selection and memory performance (Morgan et al., 2009), the information participants choose to access depends on which information was previously accessed during the trials (Blair, Watson, Walshe, & Maj, 2009), and the length of time spent viewing feedback impacts learning speed (Watson & Blair, 2008). Investigations of missing data (White & Koehler, 2004; Wood & Blair, 2010), tasks which present new sources of information (Blair & Homa, 2005), and studies of the speed of perceptual processing of features (Lamberts, 2003) are further evidence that the amount and order of known information exerts a considerable influence on the course of category learning.

These and other temporal effects on learning and performance are accumulating and will eventually force researchers to embed extant theoretical work in a dynamic, temporal framework in order to account for them.

Acknowledgments

This work was supported by funding from Simon Fraser University, the Natural Sciences and Engineering Council of Canada, and the Canadian Foundation for Innovation. Many thanks are due to the members of the Cognitive Science Lab for their contributions to literature review, data collection, and data analysis: Aaron Ancell, Jordan Barnes, Bill Chen, Taylor Clarke, Andrew Culp, Csilla Horvath, Luvdeep Malhi, Kim Meier, Gordon Pang, Jordan Shimell, Calen Walshe, Marcus Watson, and Edith Wu.

References

Ballard, D.H., Hayhoe, M.M., & Pelz, J.B. (1995). Memory representations in natural tasks. *Journal of Cognitive Neuroscience, 7*(1), 66-80.

Ballard, D.H., Hayhoe, M.M., Pook, P.K., & Rao, R.P.N. (1997). Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences, 20,* 723-767.

Baron, R.M. & Kenny, D.A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic and statistical considerations. *Journal of Personality and Social Psychology, 51*, 1173-1182.

Blair, M.R., Chen, L.C., Meier, K.M., Wood, M.J., Watson, M.R., & Wong, U. (2009). The impact of category type and working memory span on attentional learning in categorization. *Proceedings of the 31st Annual Meeting of the Cognitive Science Society* (pp. 3127-3132). Austin, TX: Cognitive Science Society.

Blair, M.R. & Homa, D. (2005). Integrating novel dimensions to eliminate category exceptions: When more is less. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*, 258-271.

Blair, M.R., Watson, M. R., & Meier, K.M. (2009). Errors, efficiency, and the interplay between attention and category learning. *Cognition, 112,* 330-336.

Blair, M. R., Watson, M. R., Walshe, R.C., & Maj, F. (2009). Extremely selective attention: Eye-tracking studies on dynamic attentional allocation to stimulus features. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*, 1196-1206.

Droll, J.A. & Hayhoe, M.M. (2007). Trade-offs between gaze and working memory use. *Journal of Experimental Psychology: Human Perception and Performance, 33*, 1352-1365.

Fu, W.T., & Gray, W.D. (2006). Suboptimal tradeoffs in information seeking. *Cognitive Psychology*, *52*(3), 195-242.

Goolkasian, P., Foos, P.W., & Krusemark, D.C. (2008). Reduction and elimination of format effects on recall. *American Journal of Psychology, 121,* 377-394.

Fiske, S.T., & Taylor, S.E. (1984). *Social Cognition.* New York, NY: Random House.

Frick, R.W. (1985). Testing visual short-term memory: Simultaneous versus sequential presentations. *Memory & Cognition, 13*, 346-356.

Gray, W.T., & Fu, W.T. (2004). Soft constraints in interactive behavior: the case of ignoring perfect knowledge in-the-world for imperfect knowledge-in-the-head. *Cognitive Science, 28*, 359-382.

Gray, W.D., Sims, C.R., Fu, W.T., & Schoelles, M.J. (2006). The soft constraints hypothesis: A rational analysis approach to resource allocation for interactive behavior. *Psychological Review, 113*, 461-482.

Kruschke, J.K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review, 99*, 22-44.

Lamberts, K. (2002). Feature sampling in categorization and recognition of objects. *Quarterly Journal of Experimental Psychology, Section A: Human Experimental Psychology, 55*(1), 141-154.

Lamberts, K., Brockdorff, N., & Heit, E. (2003). Feature-sampling and random-walk models of individual-stimulus recognition. *Journal of Experimental Psychology: General, 132*, 351-378.

Matsuka, T., & Corter, J.E. (2008). Observed attention allocation processes in category learning. *The Quarterly Journal of Experimental Psychology, 61*(7), 1067-1097.

Morgan, P.L., Patrick, J., Waldron, S.M., King, S.L., & Patrick, T. (2009). Improving memory after interruption: Exploiting soft constraints and manipulating information access cost. *Journal of Experimental Psychology: Applied, 15*(4), 291-306.

Rehder, B., & Hoffman, A.B. (2005). Eyetracking and selective attention in category learning. *Cognitive Psychology, 51*(1), 1-41.

Waldron, S.M., Patrick, J., Howes, A., & Duggan, G.B. (2006). Problem solving with information access costs in mind. *Proceedings of the 28th Annual Meeting of the Cognitive Science Society* (pp. 2335-2340). Vancouver, Canada: Cognitive Science Society.

Waldron, S.M., Patrick, J., Morgan, P.L., & King, S.L. (2007). Influencing cognitive strategy by manipulating information access costs. *The Computer Journal, 50*(6), 694-702.

Wood, M.J., & Blair, M.R. (2010). *Informed inferences of unknown feature values in categorization*. Manuscript submitted for publication.

Yamamoto, N., & Shelton, A.L. (2009). Sequential versus simultaneous viewing of an environment: Effects of focal attention to individual object locations on visual spatial learning. *Visual Cognition, 17*, 457-483.

Effects of generative and discriminative learning on use of category variability

Anne S. Hsu (ahsu@gatsby.ucl.ac.uk)
Department of Cognitive, Perceptual and Brain Sciences, University College London, London, UK

Thomas L. Griffiths (tom_griffiths@berkeley.edu)
Department of Psychology, University of California, Berkeley, Berkeley, CA 94720 USA

Abstract

Models of category learning can take two different approaches to representing the relationship between objects and categories. The generative approach solves the categorization problem by building a probabilistic model of each category and using Bayes' rule to infer category labels. In contrast, the discriminative approach directly learns a mapping between inputs and category labels. With this distinction in mind, we revisit a previously studied categorization experiment that showed people are biased towards categorizing objects into a category with higher variability. Modelling results predict that generative learners should be more greatly affected by category variability than discriminative learners. We show that humans can be prompted to adopt either a generative or discriminative approach to learning the same input, resulting in the predicted effect on use of category variability.

Keywords: human category learning; generative models; discriminative models; rational models; Bayesian models

Introduction

Categories can be learned using a variety of approaches. Here we examine two distinct approaches that humans can use to learn categories: *generative* and *discriminative*learning. While relatively unexplored in human categorization, this distinction has been widely studied in machine learning (e.g., Ng & Jordan, 2001). The distinction comes down to whether the ability to categorize objects is the result of estimating a distribution for each category, or learning a mapping from objects to categories. Both of these strategies can be used in learning real life categories. For example, you could learn the food preferences of a friend by observing the foods he eats and trying to infer a probability distribution, or by recording his affective responses to different kinds of foods and trying to identify which factors lead to positive or negative reactions.

More formally, generative and discriminative models represent two distinct strategies for estimating the probability that a particular object belongs to a category. Generative learners solve this problem by building a probabilistic model of each category, and then using Bayes' rule to identify which category was most likely to have generated the object. Discriminative learners estimate the probability distribution over category labels given objects directly. These different strategies have implications for the performance of these models. Theoretical and empirical analyses have shown that generative and discriminative models differ in their generalization behavior, as well as the speed and accuracy of learning (Efron, 1975; Ng & Jordan, 2001; Xue & Titterington, 2008).

While the generative/discriminative distinction has been studied extensively in machine learning and statistics, it has been little examined in human behavior. A recent study has shown humans can adopt these two different strategies while learning an artificial language (Hsu & Griffiths, 2009). In this paper, we explore whether people can adopt these two strategies in category learning.

The paper will be presented as follows. First we will provide an overview of generative and discriminative categorization models. Second, we will review related work from the existing human categorization literature. Third, we will revisit a previously studied paradigm that showed people are sensitive to category variability, being more likely to assign an object equidistant from the mean of two categories to the category with higher variance (Stewart & Chater, 2002; Cohen, Nosofsky, & Zaki, 2001; Rips, 1989; Smith & Sloman, 1994). Modelling results show that a generative model exhibits greater sensitivity to category variability than a discriminative model. We use this analysis as the basis for an empirical investigation of whether human learners can be prompted to take these two distinct learning approaches. Our results support the idea that humans adopt generative and discriminative approaches when appropriate. This provides new insight into the factors affecting human category learning.

Generative and discriminative models

Rational models of categorization identify the underlying problem as one of estimating the probability of a given object x belonging to a category c, as expressed by the distribution $p(c|x)$. The difference between generative and discriminative approaches to categorization comes down to how this probability distribution is estimated. Generative models build a probabilistic model of the input by learning the probability that an object x is generated given that the category is c, $p(x|c)$, and then solving the categorization problem by applying Bayes' rule. Discriminative models estimate $p(c|x)$ directly. Generative models thus assume that observed objects are sampled in a way that reflects $p(x|c)$, while discriminative models do not make any assumptions about the distribution from which the input is sampled. These two approaches to categorization are illustrated schematically in Figure 1.

Comparison of generative and discriminative approaches to category learning has been done in the machine learning and statistics literature, where the classic *generative-discriminative pair* being compared is usually (generative) naïve Bayes vs. (discriminative) logistic regression (Efron, 1975; Ng & Jordan, 2001; Xue & Titterington, 2008). Under certain conditions, these two models are identical in the asymptotic form of the function $p(c|x)$ that they produce, differing only in how that function is estimated.

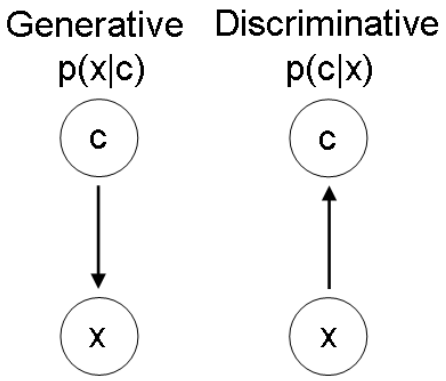


Figure 1: Generative and discriminative models. Generative models aim to estimate the probability distribution over the input given the category label. Discriminative models find a direct mapping between inputs and category labels.

Such generative-discriminative pairs can thus be used to explore the consequences of adopting these different strategies through mathematical analysis and simulations. For example, if the training data consist of two normally distributed samples, generative models learn categories more quickly (Efron, 1975; Ng & Jordan, 2001). However, when the training data come from other distributions, discriminative models are asymptotically more accurate (Xue & Titterington, 2008), though in some cases generative models may perform better initially and arrive at their (higher) asymptotic error more quickly (Ng & Jordan, 2001).

Summary of related work

Previous models of categorization have used both generative and discriminative strategies, without necessarily recognizing that the significance of the distinction.The commonly cited prototype and exemplar models can be applied both generatively and discriminatively. Prototypes and exemplars are psychological models of category representation whereas discriminative and generative are statistical models of learning. Thus, prototype and exemplar models can be used under either approach, depending on how learning takes place. For instance, ALCOVE (Kruschke, 1992) is an exemplar model akin to discriminative kernel methods. SUSTAIN (Love, Medin, & Gureckis, 2004) is a discriminative model that chooses between exemplar and prototype representations. Decision bounds (Maddox & Ashby, 1993) can be either discriminative or generative depending on how model parameters are estimated. While rational models of categorization can adopt either approach, the ones proposed so far have taken a generative approach (e.g., J. R. Anderson, 1990; Griffiths, Canini, Sanborn, & Navarro, 2007). These generative categorization models span the range between exemplar and prototype representations. At the extremes, generative prototype models estimate parameters of category distributions (usually a Gaussian with a mean and variance) and generative exemplar models estimate category distributions using kernel density estimation (Ashby & Alfonso-Reese, 1995).

Despite the prevalence of human categorization models with both discriminative and generative approaches, most experimental paradigms seem more consistent with discriminative learning: stimuli are presented, participants guess the category and feedback is given. However, a few exceptions this can be seen in previous work on classification vs. inference learning, and observational vs. feedback learning. While not explicitly mentioned in previous work, both of these paradigms are potentially related to our discriminative vs. generative distinction.

Classification vs. inference learning

Another line of experiments has shown that human category learning can also be influenced by using different tasks to teach people about the relationship between categories and features. The effect of using these two different tasks is similar to that of changing the direction of a learned causal relationship. (A. L. Anderson, Ross, & Chin-Parker, 2002; Markman & Ross, 2003; Ross & Murphy, 1996). In these experiments, all participants were presented with exactly the same training stimuli, consisting of the features and category membership of a set of objects. In one condition, learning took place via through *classification*: Participants were provided with the values for (some of) the features of an object asked to predict category membership. In the other condition, learning was based on making a predictive *inference*: The category membership and/or values of some of the features were provided and participants were asked to predict the value of another feature. Because participants in both conditions were given feedback, they were both ultimately provided with exactly the same information about categories and features. However, learning results differed in terms of performance accuracy and generalizations made. For example, inference learners performed better than classification learners on single-feature classification tasks but more poorly when all of the features were provided (A. L. Anderson et al., 2002). While this study was not motivated by generative and discriminative learning, people may have adopted these different strategies in the different conditions: Classification learning can be done using a discriminative model, while inference learning requires a generative model.

Observation vs. feedback training

Another study, by Ashby, Maddox, and Bohill (2002), has also examined how learning of the exact same input was affected by presentation style. Here they compared what they called *feedback* training (where the category label appears after the object) with *observation* training (where the category label appears before the object). Their results showed that participants in the feedback condition performed significantly better than those in the observation condition for information-integration categories, where category membership could not be expressed in terms of a rule using a single feature. These two forms of training might encourage learners to adopt gen-

erative and discriminative strategies. Feedback training gives an error signal that can be used to adapt a discriminative model. Observation training is more relevant for learning object features based on the category label, which is the generative approach.

Summary

Generative and discriminative models use different approaches to solve the problem of categorizing objects. Existing models of human category learning differ in which of these approaches they use. Previous work has not explored whether people are able to switch the approach they take in learning categories, although the effects of different training regimes that might encourage one approach over the other have been investigated. In the remainder of the paper, we explicitly test whether people can adopt these two approaches to learning categories, using a phenomenon that is diagnostic for one generative-discriminative pair of models.

Differential use of category variability

Several experiments have shown an effect of category variability on human categorization judgments. In these experiments, the stimuli belong to one of two categories with different means and variances. The key question is how stimuli with features lying (perceptually) in between the two categories are categorized. The results of these experiments all showed that there was a bias towards categorizing stimuli into the high-variance category (Stewart & Chater, 2002; Cohen et al., 2001; Rips, 1989; Smith & Sloman, 1994). Here we propose that the degree of preference for the high variance category may be affected by whether the learner is adopting a generative or discriminative approach.

Intuitively, we expect category variability to have a greater effect on generative learners because estimating $p(x|c)$ for each category requires being sensitive to the variance of that category. In contrast, one need not consider the variance of the stimuli in simply learning a function from x to c, $p(c|x)$. Indeed many discriminative models used in machine learning, such as support vector machines (Schölkopf & Smola, 2002), focus just on the location of the most extreme members of each category. We are not claiming that all generative models are sensitive to category variance, or that all discriminative models are insensitive, but that these approaches differ in the extent to which they are sensitive to this property of the stimuli. To illustrate this, we will explore the predictions of one generative-discriminative pair of models.

We follow previous work exploring the difference between generative and discriminative models (e.g., Ng & Jordan, 2001) and focus on the generative-discriminative pair of naïve Bayes and logistic regression. Since we will focus on continuous stimuli, we assume a Gaussian generative model, with

$$p(x|c=i) = N(\mu_i, \sigma_i) \tag{1}$$

where μ_i and σ_i are the mean and variance of the ith category with $i \in \{1,2\}$. The parameters μ_i and σ_i can be estimated by maximizing the likelihood $\sum_{j=1}^{n} \log p(x_j|c_j,\mu,\sigma)$, where c_j and x_j are the category membership and features of the jth stimulus respectively. The probability a novel stimulus belongs to a category, $p(c|x)$, is then computed by applying Bayes' rule, with the prior probability of each category being proportional to the number of observed stimuli from that category. The naïve Bayes model is similar to the Gaussian decision bound model used in Normal general recognition theory (Stewart & Chater, 2002; Maddox & Ashby, 1993).

The discriminative model uses logistic regression to estimate $p(c|x)$ directly, with

$$p(c=1|x,w,b) = 1/(1+\exp\{-w^T x) - b\}) \tag{2}$$

where w and b are the parameters of the model and x is a vector of feature values. The parameters w and b are estimated by maximizing the log likelihood $\sum_{j=1}^{n} \log p(c_j|x_j,w,b)$. In general, w and b are vectors of length equal to the number of stimulus features. However, we will be using one-dimensional stimuli (x_j is scalar), so w and b will be scalars in our case.

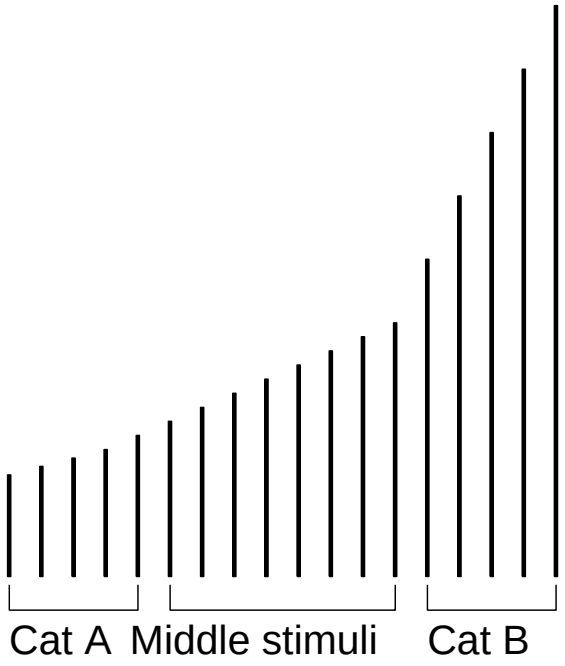


Figure 2: Stimuli used in the experiment. Category A and B were the low and high variance categories respectively

To examine the predictions of these models, we used stimuli based largely on those of Cohen et al. (2001). Stimuli consisted of vertical lines of varying lengths. Training stimuli belonged to one of two categories, A and B. Category A is the low variance category. Category A contained lines of length 110, 120, 130, 140 and 150 pixels. Category B was the high variance category. Category B contained lines of length 300, 375, 450, 525 and 600 pixels. All stimuli were equally likely within each category (categories had a flat distribution of stimuli). We also included novel transfer stimuli in the test

stimuli. There were eight transfer stimuli, equally spaced between the highest value of A and the lowest value of B (see Figure 2). A range of intermediate transfer stimuli were used in case the middle stimulus in psychological space differed from the numerical middle stimulus. The precise location of the middle stimulus is not important for our purposes, as the difference in results between generative and discriminative models is the question of interest.

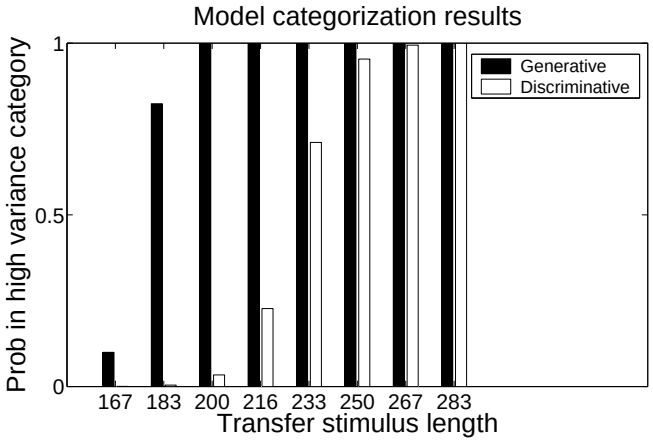


Figure 3: Generative and discriminative model predictions for the probability of categorization stimuli into the high variance category. The model predictions are that a generative learner is more likely to categorize in between stimuli in the high variance category

We trained a generative naïve Bayes model and discriminative logistic regression model on all labeled examples from category A and B. Our naïve Bayes model had uniform category priors, i.e. both categories were assumed to be equally likely. Parameters for both models were fit using maximum likelihood estimation. To compare the outcomes of the two models, we analysed categorization predictions for our transfer stimuli using these generative and discriminative models. The generative model predicts intermediate transfer stimuli will be classified to the high-variance category more often than the discriminative model (see Figure 3).This is because it is more likely that intermediate stimuli are extreme values from the high-variance category than the low-variance category. These results illustrate that sensitivity to category variability may be a diagnostic indicator of whether learners are using a generative or a discriminative strategy. In the next section we present an experiment that uses this indicator to determine whether human learners switch between these strategies depending on the way in which a categorization task is presented.

Human generative and discriminative learning

Method

Participants We collected data from 24 participants (12 in each condition). Participants wereundergraduates at the University of California, Berkeley and received course credit.

Stimuli Stimuli was the same training and transfer stimuli used in the model simulations described in the previous section. In the experiment, these stimuli were presented as white vertical lines in a black circle.

Procedure While previous related work had paradigms that may have encouraged discriminative or generative learning (Ashby et al., 2002; A. L. Anderson et al., 2002), the connection between these paradigms and the distinction was tentative. Thus, we will use our own experimental manipulation in order to encourage participants to adopt the distinct approaches as strongly as possible. Participants in both learning conditions were trained under the same randomized sequence of trials. In order to prompt generative or discriminative learning, the two conditions differed in the instructions, category-stimulus presentation order and question presented during testing blocks. Participants in both conditions were told they will see "signs" from an alien tribe. Participants in the *generative* condition were told that two aliens, one from each tribe (A and B) will appear and produce signs from their respective tribes. A picture of two aliens, who were identical except for the letter on their chest, was shown alongside the instructions. These instructions were intended to make it clear that the observed stimuli were generated from a probability distribution associated with the target category, consistent with the assumptions of a generative model. Participants in the *discriminative* condition were told that there are signs from two alien tribes and they would be shown a single alien translator who can report which tribe a sign was from. A single alien was shown alongside these instructions with a question mark on its chest. These instructions were intended to establish a situation in which participants learned a function from stimuli to category membership, consistent with a discriminative model.

For all participants, the experiment contained 10 blocks of 20 trials (each of 10 training stimuli were shown twice). Training blocks (odd blocks) were interleaved with testing blocks (even blocks). During training trials, participants were shown a black circular background on which the "sign" appears as a white vertical line, next to an alien with either A or B written on its chest. In the *generative* condition, the alien appeared 500 ms before the sign during training and the alien disappeared between trials to simulate different aliens appearing. In the *discriminative* condition, the sign appeared 500 ms before the alien and the alien did not disappear between trials to simulate one constant alien interpreter. In both conditions,once both stimulus and letter had appeared, both remained simultaneously on the screen for 1.5 s (see Figure 4). The total length of each training trial was 2 s and there were 700 ms between each trial.

During test trials, participants were shown a sign (white vertical line) on the black circular background. Participants in the *generative* condition were asked "Which alien was more likely to have produced this sign?". Participants in the *discriminative* condition were asked "Which alien tribe does this sign belong to?". Stimuli during each test block consisted of

every example stimulus in categories A and B, along with the eight transfer stimuli that were equally spaced between and highest value of category A and the lowest value of category B. (The highest value of category A and lowest value of category B were seen twice during each test block to make up the 20 trials.) No feedback was given during testing in either condition.

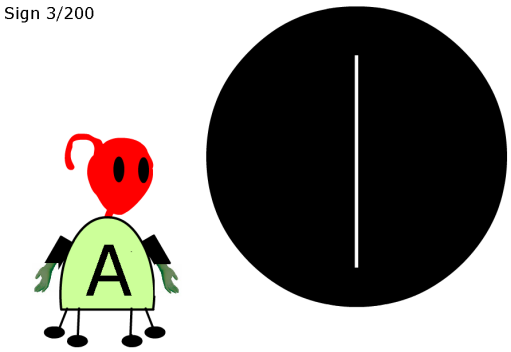


Figure 4: Screen shot of the experiment

Results

The human learning results correspond to the predictions of the models: Generative learners are more likely to categorize transfer stimuli that lie in between the two categories in the high-variance category relative to discriminative learners (see Figure 5). A two-way within-between ANOVA revealed statistically significant effects of test stimulus ($F(9,198) = 76.88$, $MSE = 0.036$, $p < .001$) and condition ($F(1,22) = 5.43$, $MSE = 0.216$, $p < .05$) and a marginally significant interaction ($F(9,198) = 1.90$, $MSE = 0.036$, $p = .054$). Planned comparisons using two-sample t-tests showed statistically significant effects of condition for stimuli 216 ($t(22) = 2.57$, $p < .05$) and 233 ($t(22) = 2.46$, $p < .05$). These statistics are calculated under the most conservative assumption, under which the responses from each participant for each stimulus are averaged together and treated as a single response.

The "middle stimulus" that lies midway between the two categories in human perceptual space (i.e. equally likely to be categorized in both categories in the discriminative condition) is of length around 200 pixels. This is smaller than the numerical middle (225 pixels). This is approximately the same value as the perceptual "middle stimulus" that was found in previous work (Cohen et al., 2001). Accounting for this shift, the discriminative model predictions match fairly well with the discriminative human results. The generative model predictions are significantly shifted to the left compared with our generative human results, meaning the generative model predicted an even stronger tendency to categorize the in-between stimuli in the high variance category. This difference in degree between model predictions and human judgments could be explained in many possible ways. One possibility is that perceptual stimuli might follow Weberian compression for the larger stimuli (Stewart & Chater, 2002). As a result of this compression, the perceptual variability of the longer length lines (which made up the high variability category) may have been significantly smaller than the absolute numerical variability values that were used in our models. If this were the case, a suitable transformation, such as to log space, would leave our qualitative results the same, while resulting in an appropriately less strong variability preference for the generative model. Another possibility is that people are not making the Gaussian assumption that was made by our model. This is plausible as our stimuli were very non-Gaussian. In this case, it is possible that the probability of belonging in the high variance category under a Gaussian assumption is greater than the probability estimates that generative participants might have made for our actual stimuli. Finally, participants may not be behaving fully generatively, or that the instructions resulted in a mixed population of generative and discriminative learners in this condition.

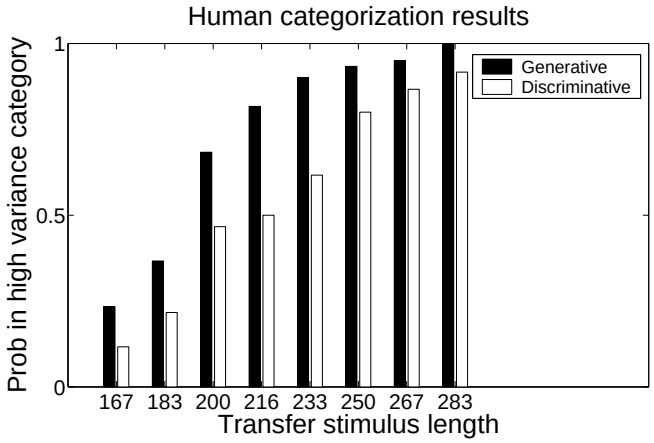


Figure 5: Probability of categorizing transfer stimuli in high variance category for participants in the generative and discriminative learning conditions. Total values are the average of all probabilities for individual stimulus lengths.

Discussion

The distinction between generative and discriminative approaches to categorization has played an important role in machine learning research, but has not previously been explored in cognitive psychology. Our results show that people can be cued to take these two different approaches to category learning through the way in which a categorization task is presented. These results have implications for understanding human category learning, and for establishing links between the communities studying human and machine learning.

The finding that people behave differently when encouraged to adopt these two different approaches to category learning may shed light on previous empirical results in cognitive psychology. For example, some previous experiments have shown effects that may be partly due to learning paradigms that encouraged participants to adopt generative or discriminative learning approaches (e.g., Ashby et al.,

2002). The generative/discriminative distinction also has potential implications for previously proposed models of categorization. For example, it seems appropriate that connectionist models (Kruschke, 1992; Love et al., 2004) will best characterize behavior when humans adopt a discriminative learning approach whereas rational models (J. R. Anderson, 1990; Griffiths et al., 2007) will best describe behavior when humans adopt a generative learning approach. Developing a deeper understanding of how this distinction plays out in human learning may provide additional insights into long-standing debates on category learning.

Showing that people can adopt both generative and discriminative learning strategies establishes a new connection between human and machine learning. While many of the goals of machine learning are inspired by human capabilities (e.g., the ability to recognize and categorize complex structures quickly and efficiently), the principal issues that are topical in machine and human learning seldom coincide. By showing that a key distinction long studied in machine learning research is also significant to human learning, this work begins to build an important bridge between machine learning and human learning communities. This will encourage collaboration between the two research communities where computational models of learning provide insight into human learning and human learning, in turn, inspires computational modelling. It also establishes a way to know how advances in specific aspects of machine learning, such as improved discriminative models, might be relevant to predicting aspects of human learning.

Identifying the relevance of the generative/discriminative distinction in human categorization also opens up many new avenues of research questions. For the neuroscience community, one can ask: What neural mechanisms are implementing these two very different learning strategies? Are the neural circuits involved similar or different? This research also provokes many questions about learning more generally: When does human learning tend to be generative or discriminative? How flexible are learners in alternating between generative and discriminative learning approaches? Can learning approaches be retrospectively altered? (i.e. if input is learned with a discriminative perspective and learners were later made to understand that the data was generated from a probability distribution, would they switch their categorization judgments?) Since much of human learning in everyday life consists of a mix of scenarios in which one or the other of these strategies is more appropriate, clarifying when people use generative and discriminative approaches will help us understand differences in learning among individuals and across situations. We anticipate that exploring these questions will result in improved models of human category learning, and a tighter coupling between research on human and machine learning.

Acknowledgments. We thank Ky Merritt for assistance in data collection. This research was funded by the Economics and Social Research Council grant RES-000-22-3275 and National Science Foundation grants SES-0631518 and IIS-0845410.

References

Anderson, A. L., Ross, B. H., & Chin-Parker, S. (2002). A further investigation of category learning by inference. *Journal of Experimental Psychology: Learning, Memory, Cognition*, *30*, 119–128.

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.

Ashby, F. G., & Alfonso-Reese, L. A. (1995). Categorization as probability density estimation. *Journal of Mathematical Psychology*, *39*, 216–233.

Ashby, F. G., Maddox, W. T., & Bohill, C. J. (2002). Observational versus feedback training in rule-based and information-integration category learning. *Memory and Cognition*, *80*, 666–677.

Cohen, A. L., Nosofsky, R. M., & Zaki, S. R. (2001). Category variability, exemplar similarity, and perceptual classification. *Memory and Cognition*, *29*, 1165-1175.

Efron, B. (1975). The efficiency of logistic regression compared to normal discriminant analysis. *Journal of the American Statistical Association*, *70*, 892–898.

Griffiths, T. L., Canini, K., Sanborn, A. N., & Navarro, D. J. (2007). Unifying rational models of categorization via the hierarchical dirichlet process. In *Proceedings of the Twenty-Ninth Annual Conference of the Cognitive Science Society.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Hsu, A. S., & Griffiths, T. L. (2009). Differencial use of implicit negative evidence in generative and discriminative language learning. In *Advances in Neural Information Processing Systems 22.*

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*, 22–44.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, *111*, 309–332.

Maddox, W. T., & Ashby, F. G. (1993). Comparing decision bound and exemplar models of categorization. *Perception and Psychophysics*, *53*, 49–70.

Markman, A. B., & Ross, B. H. (2003). Category use and category learning. *Psychological Bulletin*, *129*, 592–613.

Ng, A. Y., & Jordan, M. (2001). On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. In *Advances in Neural Information Processing Systems 17.* Cambridge, MA: MIT Press.

Rips, L. J. (1989). Similarity, typicality, and categorization. In S.Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning.* New York: Cambridge University Press.

Ross, B. H., & Murphy, G. L. (1996). Category-based predictions: influence of uncertainty and feature associations. *Journal of Experimental Psychology: Learning, Memory, sCognition*, *22*, 736–753.

Schölkopf, B., & Smola, A. J. (2002). *Learning with kernels*. Cambridge, MA: MIT Press.

Smith, E. E., & Sloman, S. A. (1994). Similarity-versus rule-based categorization. *Memory and Cognition*, *22*, 377–386.

Stewart, N., & Chater, N. (2002). The effect of category variability in perceptual categorization. *Journal of Experimental Psychology: Learning, Memory, Cognition*, *28*, 893–907.

Xue, J., & Titterington, D. M. (2008). Comment on "On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes". *Neural Processing Letters*, *28*, 169–187.

Category Learning Through Active Sampling

Doug Markant (doug.markant@nyu.edu)
Todd M. Gureckis (todd.gureckis@nyu.edu)
New York University, Department of Psychology
6 Washington Place, New York, NY 10003 USA

Abstract

Laboratory studies of human category learning tend to emphasize passive learning by limiting participants' control over the information they experience on every trial. In contrast, we explore the impact that *active* data selection has on category learning. In our experiment, participants attempted to learn categories under either entirely passive conditions, or by actively selecting and querying the labels associated with particular stimuli. We found that participants generally acquired categories faster in the active learning condition. Furthermore, this advantage depended on learners actually making decisions about which stimuli to query themselves. However, the effectiveness of active sampling was modulated by the particular structure of the target category. A probabilistic rule-learning model is proposed that explains the results in terms of a strong prior bias towards uni-dimensional rules which impairs learning of alternative category boundaries. Active learners appear to be able to bootstrap their own learning, but this ability may be strongly constrained by the space of hypotheses that are under consideration. **Keywords:** categorization, active learning, information sampling, rule learning, decision-bound models

Despite the widely held view that people learn better by *doing* than simply *observing*, there have been surprisingly few detailed accounts of the impact that "active" information acquisition has on the learning process. In particular, theoretical models which explain how people learn new concepts from examples usually treat learners as passive accumulators of evidence about the structure of categories. For example, the standard procedure in most category learning experiments is to exhaustively and randomly sample the set of training stimuli. However, in everyday life, human learners can often control their own learning by selectively "sampling" particular observations they estimate to be useful or informative. The goal of the present paper is to understand the cognitive consequences of this type of learning.

There are at least two explanations for why active sampling might result in better learning than passive observation. First, rather than being limited by the flow of information from passive experience, active learners are free to select which information they want to learn about. For example, by making directed queries that take into account their current uncertainty, the learner may be able to optimize their experience (e.g., avoiding redundant data). Research in machine learning has shown that the principle of uncertainty sampling (selectively querying data that is expected to be informative) can have a dramatic impact on the amount of training needed to reach a performance criterion (Settles, 2009).

Independent of the advantage of better data, active learners may also benefit from greater engagement in the learning task. For example, the very act of planning interventions or deciding which samples to take may necessitate deeper evaluation of the problem structure and of how observed experience relates to different hypotheses (c.f., Bruner, 1961). In a study of active intervention during a causal learning task, Sobel and Kushnir (2006) showed that active learners were more likely to learn a hidden causal structure than participants that were "yoked" to their interventions (i.e., a group with the same data but who did not independently make sampling decisions). Similar concerns are often used to support educational practices that emphasize "inquiry" or "discovery"-based instruction (Kuhn et al., 2000).

The aims of the present study were two-fold. First, we were interested if participants could adaptively structure their own learning experiences when acquiring new concepts. Second, we were interested in how the effectiveness of active sampling might interact with the specific structure of categories. While a number of recent studies have explored how learners make information sampling decisions to support their own learning (Castro et al., 2008; Kruschke, 2008; Gureckis & Markant, 2009; Steyvers et al., 2003), there has not yet been a systematic evaluation of how this ability might vary across different category structures.

Overview of the present experiment

Our experiment adapts a well-studied paradigm for perceptual category learning using multidimensional, continuous-valued stimuli. In the task, participants learned to classify perceptual stimuli into different abstract groups. Two types of category structures were used: *rule-based* (RB), in which the decision rule is defined as a criterion along a single dimension, and (2) *information-integration* (II), in which the decision rule is a function of at least two dimensions (see Figure 1). Participants in the experiment were further divided into three training conditions. In the *passive-normal* condition, participants observed training stimuli that were generated from two bivariate normal distributions (i.e., a standard training procedure). In the *active* condition, participants were able to "design" stimuli for which they received feedback about the category label. In the *passive-yoked* condition, each participant was linked to an active learner, passively observing the samples they made and receiving the same feedback.

There are three key aspects of the design worth highlighting. First, in binary classification tasks, the optimal sampling strategy is simply to make queries close to the current estimate of the category boundary (or margin) — the region of greatest uncertainty. However, we anticipated that participants' ability to do so might vary between the RB and II learning tasks. Previous research has suggested that these two types of category structures may be learned in fundamentally different ways (Ashby, Alfonso-Reese, Turken, & Waldron,

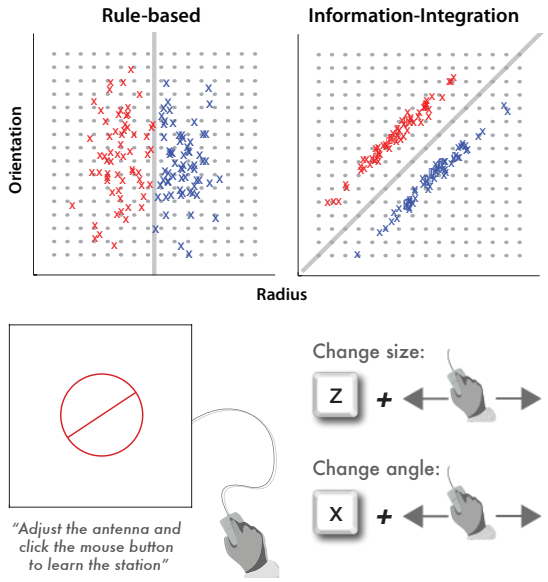


Figure 1: **Top:** Category distributions used in the experiment. 'X's indicate training stimuli shown to participants in the passive-normal condition with color indicating the generating distribution (actual feedback received by participants was probabilistic). The uniform grid of points over the stimulus space indicate the set of unlabeled test stimuli. **Bottom:** An example stimulus (left) and the interface used in the active learning condition.

1998). In particular, RB categories are thought to be learned by reasoning about verbal or explicit hypotheses (which is the default learning mode), while the structure of II categories precludes a simple verbal description and are instead thought to be learned via implicit or procedural learning. To the degree that effective sampling relies on explicit reasoning about uncertainty, people may perform better in the RB condition where this uncertainty may be better articulated. Similarly, active learning may be more effective in the RB case because the category aligns with default biases people bring to the task (Ashby et al., 1999; Kruschke, 1993).

Second, the comparison of active learners with the passive-normal group allowed us to test if active learning could lead to a performance advantage above and beyond the typical training procedure in such tasks. We expected that if active learners were able to make useful queries, they would be faster at learning the correct category distinction than the passive-normal participants. Again, if active learners are less successful at making useful queries in the II task, any learning advantage may be attenuated. Moreover, since successful learning in the II task may be contingent on abandoning rule-based strategies in favor of a more procedural type of learning, active learning might even lead to a learning *impairment* by encouraging perseveration in the search for a sub-optimal rule.

Finally, the inclusion of the passive-yoked training group allowed us to separately evaluate the impact of selecting samples from the statistical information contained in those samples (since the distribution of training data is identical for both groups). While previous research (in causal learning settings) suggests that active or intervention-driven learning may lead to advantages over comparable yoked conditions (Lagnado & Sloman, 2004; Sobel & Kushnir, 2006; Steyvers et al., 2003), it is unknown how these results generalize to other tasks.

An Experiment

Participants One hundred eighty undergraduates at New York University participated in the study. The experiment was run on standard Macintosh computers in a single 40 min session. Each participant was assigned to either the rule-based (RB) or information-integration (II) task condition, and to one of three training conditions: active (A), passive-normal (P), or passive-yoked (PY).

Stimuli Stimuli were defined by a two-dimensional continuous-valued feature space, where one dimension corresponded to the size (radius) of a circle and the second dimension corresponded to the angle of a central diameter (see example in Figure 1, bottom). One-hundred and twenty-eight training stimuli were created for the passive-normal training condition using bivariate normal distributions (see Figure 1, top) with mean and covariance parameters slightly modified from Ashby et al. (2002). Test stimuli were drawn from a uniform grid of samples over the feature space (depicted by the gray dots in Figure 1). Thirty-two stimuli were presented in each test block, amounting to a total of 256 test trials.

Procedure Participants were told that the stimuli in the experiment were "loop antennae" for old televisions, and that each antennae received one of two channels (CH1 or CH2). The channel received by any antenna depended in some way on the two dimensions described above, and participant's goal was to learn the difference between the two types of items. The feedback associated with each item during training was probabilistic and was proportional to the relative likelihood of either category for the ideal observer who knew the true category distributions. Participants were given instruction that the antennas were sometimes "noisy" and would pick up the wrong channel and that it would be beneficial to integrate over a number of trials when learning. The experiment consisted of 8 blocks, with each block divided into a set of 16 training trials followed by 32 (no feedback) test trials.

Training – Active Condition. On each training trial the participant "designed" a TV antenna and learned about its category membership. Each trial began with the presentation of a randomly generated stimulus in the center of the screen. The participant could then alter its size and orientation by moving the mouse from left to right while holding down either the 'Z' or 'X' key, respectively (see Figure 1, bottom). Only one dimension could be changed at a time, but participants could make any number of changes and use as much time as needed. When the stimulus was the desired size and orientation, participants pressed the mouse button to reveal the category label, which appeared above the stimulus and was visible for 1500ms. Querying the category label was not permitted until the participant had made a change to the initial stimulus.

Training Trials – Passive-Normal Condition. In the passive-normal condition, participants were unable to interact with the stimuli in any manner[1]. Instead, in each trial they were presented with a stimulus generated from the category distributions described above. On each trial, a fixation cross was presented, followed by the stimulus (for 250ms), followed by the category label (above the stimulus for 1500ms). When the category label was displayed, the participant was required to press a key corresponding to that category in order to end the trial. This procedure is equivalent to the observational learning condition used in Ashby et al. (2002).

Training – Passive-Yoked Condition. The purpose of the yoked

[1] In this design passive participants are not matched to active participants in terms of perceptual-motor task demands (e.g., precisely adjusting the stimulus). However, pilot data suggested that equating this made learning much more difficult for the passive group, potentially playing into any hypothesized active learning advantage.

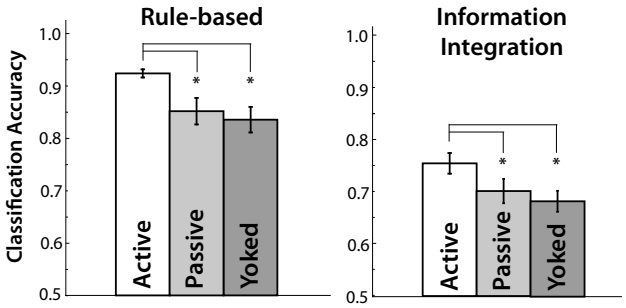


Figure 2: Accuracy in RB (left) and II (right) tasks for the three training conditions. Error bars show the standard error of the mean.

condition was to mimic the passive training experience, but to use a sequence of observations that were selected by a participant in the active condition. Each yoked participant was assigned to a matching participant in the active learning condition that had already completed the study. Training samples from the active participant were used as the set of training items for the yoked participant, and were presented in the identical order as they had been generated by the active participant. All other aspects of the yoked condition were identical to the passive-normal condition.

Test – All Conditions. On each test trial, a single item was presented in the center of the display and participants were asked to classify the item according to the channel the item was most likely to receive. A response was required to complete the trial, and participants responded at their own pace. No feedback was provided on individual test trials. At the end of each block participants were told their cumulative accuracy during the block they just completed, as well as their accuracy during the preceding test block.

Results

Responses during test blocks were scored according to whether the participant identified the correct category of each test item (with respect to the true discriminant function). Overall accuracy across tasks and conditions is shown in Figure 2. A 2-way ANOVA with task type (RB/II) and training condition (A/P/PY) as between subjects factors found significant main effects of both task ($F(1,174) = 155.97, p < 0.001$) and training condition ($F(2,174) = 15.34, p < 0.001$), but no interaction ($F(2,174) = 0.27$). In the RB task, overall accuracy was significantly higher in the active condition than in both the passive-normal ($t(58) = 2.69, p < 0.01$) and passive-yoked ($t(58) = 3.96, p < 0.001$) conditions, while there was no difference between the two passive conditions. Similarly, in the II condition, the active group was more accurate than both passive groups (P: $t(58) = 2.58, p < 0.05$; PY: $t(58) = 4.27, p < 0.001$), while there was no difference between passive-normal and passive-yoked ($t(58) = 1.57, p = 0.12$). Note that while active learners generally outperformed their passive counter-parts, active samplers in the II task only achieved 75% correct on average which may reflect a variety of sub-optimal rule-based strategies.

For participants in the II task, a 2-way ANOVA on average accuracy revealed a main effect of condition ($F(2,609) = 8.74, p < 0.001$), a main effect of block ($F(7,609) = 3.92, p < 0.001$), and a significant condition-by-block interaction ($F(14,609) = 1.74, p < 0.05$). Examination of this interaction suggested that it was driven by an early learning advantage for the active learners which was reduced later in the task. A similar analysis in the RB condition found only a main effect of training condition ($F(2,87) = 6.65, p < 0.005$) and block ($F(7,609) = 17.31, p < 0.001$).

Sampling behavior. Figure 3A shows the distribution of queries for active participants in the RB and II tasks for the final training block. In both tasks, participants begin by widely distributing their samples over the stimulus space, but over time make samples that are closer to the true category boundary. We measured the orthogonal distance of each sample to the true category boundary and computed the average distance within each block. Figure 3B shows that in the RB task average distance was significantly smaller than the null hypothesis of a random sampling strategy by the second training block (one-sample t-test, $t(29) = 4.33, p < 0.001$). This shift toward margin sampling was slower and less extreme in the II task, with average distance reliably smaller than expected from a random strategy starting around the sixth training block ($t(29) = 4.53, p < 0.001$).

Relating sampling behavior and learning. We found that overall sample distance from the boundary (averaged across blocks) was significantly correlated with active learners' overall test performance in both the RB ($r = -0.42, p < 0.05$) and II ($r = -0.8, p < 0.001$) tasks (see Figure 3D, blue line). One question is if being yoked to a high-performing active participant leads to a similar learning advantage for the passive-yoked participants. In contrast to active learners, average sample distance was not strongly correlated with performance in either task condition (RB: $r = 0.36, p = 0.051$, II: $r = -0.05, p = 0.4$, see Figure 3D, orange line). In fact, there was even a trend toward the reverse relationship in the RB task; that is, passive-yoked learners who received the most objectively useful training data were among the worst performers in the group for that task.

One objection to measuring sample "quality" by its distance from the true category boundary is that people may instead evaluate samples relative to their subjective belief about the boundary at any point in time. Using logistic regression we found the best-fit linear decision boundary for subjects' response data on each test block. We then computed the average "subjective" sample distance from that boundary in the following training block, and computed the average over blocks for all active and passive-yoked participants. We found that this distance was smaller in the active group than passive-yoked group in both tasks highlighting the divergence in inference between the two groups (RB: $t(29) = -4.07, p < 0.001$, II: $t(28) = -4.94, p < 0.001$). In addition, subjective distance measure was negatively correlated with overall accuracy in all conditions (RB(A): $r = -0.54, p < 0.005$, RB(PY): $r = -0.47, p < 0.05$, II(A): $r = -0.79, p < 0.001$, II(PY): $r = -0.41, p < 0.05$, see Figure 3E).

Discussion

There are three key behavioral findings from the experiment. First, active learners were more accurate than passive ob-

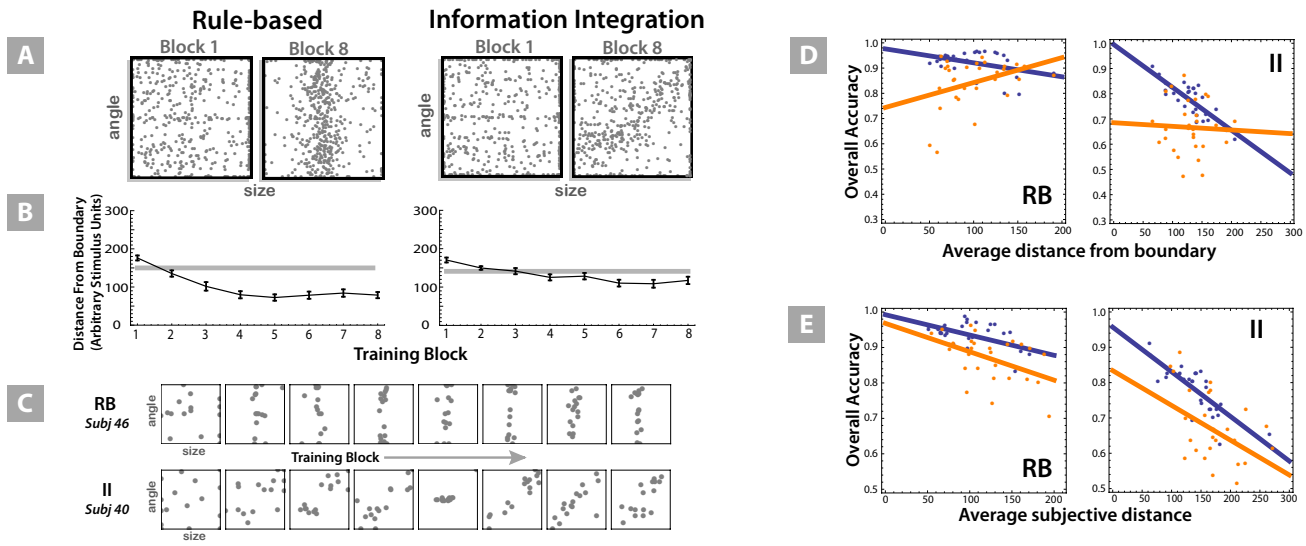


Figure 3: **A**: Composite of samples chosen by active participants in the first (left) and last (right) training block. **B**: Average distance of participants' samples from the true category boundary (black) as compared to average distance expected from random sampling (gray line). **C**: Examples of active participants in both tasks that successfully sample close to the true category boundary. **D**: Samples closer to the true boundary are associated with higher accuracy in active but not passive-yoked learners, while low "subjective" distance from a best-fit response boundary is predictive of higher accuracy in both groups (**E**).

servers in both tasks. One explanation is that active learners are able to query regions in the stimulus space where they are most likely to commit classification errors (i.e., the margin of the category boundary). Since participants in the passive-normal condition received samples from a "true" category distribution, they may be at a disadvantage because they were less likely to observe test items close to the boundary. Nevertheless, it is extremely interesting that näive participants could intuitively identify what information would most useful to support their own learning in an abstract problem space.

However, the advantage for active learners cannot be explained by a difference in training data alone. Most striking is the finding that yoked participants showed no improvement over the passive-normal group despite learning from the exact same observations as the active group. Indeed, the passive-yoked participants that observed the most objectively useful training data were among the worst performers, particularly in the RB task. If active and passive-yoked learners are assumed to update their beliefs through a common process (as would be predicted by all existing models of human categorization) then this strong pattern of divergence is unexpected.

Finally, we found a main effect of category structure. Overall, participants in the II task performed more poorly at the task. Also, even though active learners in the II condition out-performed their passive counterparts, they were unable to boost performance near to RB levels. In addition, their sampling behavior suggests that (outside of a few surprising exceptions, see Figure 3C) most participants were unable to sample near the diagonal category margin, as would be predicted by an optimal information selection strategy (Oaksford & Chater, 1994). In the following section, we present a simple model-based analysis of each of these effects.

A Probabilistic Model of Decision-Bound Learning

While there have been a number of models proposed for how people classify items using rules in continuous dimension spaces, there have been fewer attempts to articulate an inference procedure for such models (c.f., Nosofksy & Palmeri, 1998). As a result, there were two key properties that guided the development of our modeling framework. First, we wanted a way to specify a strong inductive bias towards uni-dimensional rules along either stimulus dimension (similar to the default verbal system in Ashby et al., 1998). Most existing models can specify a prior bias towards a particular dimension (e.g., based on salience), but not a more general preference for arbitrary uni-dimensional rules (Heller et al., 2009). Second, analysis of the decision rules that participants use from one block to the next suggested that these were updated in a rather rapid fashion characteristic of serial hypothesis testing.

These concerns led us to a probabilistic model of classification which assumes that the goal of learning is to discover the latent parameters of a simple linear decision boundary. In our model, the probability that an observation, o^t, on trial t falls in category A is assumed to depend on a set of latent model parameters $\{\mathbf{w}, b, \sigma\}$:

$$P(o^t = A|\mathbf{w}, b, \sigma) = (1 + exp(-\sigma(\sum_i w_i o_i^t) - b))^{-1} \quad (1)$$

where o_i^t is the stimulus value of dimension i. Since the classification is binary, $P(o^t = B|\mathbf{w}, b, \sigma) = 1 - P(o^t = A|\mathbf{w}, b, \sigma)$. The weight vector, $\mathbf{w}$, contains the decision weight assigned to each dimension. The bias term, b, allows fine adjustments to the position of the decision rule in the stimulus space. Finally, the slope of the sigmoid function is controlled by σ

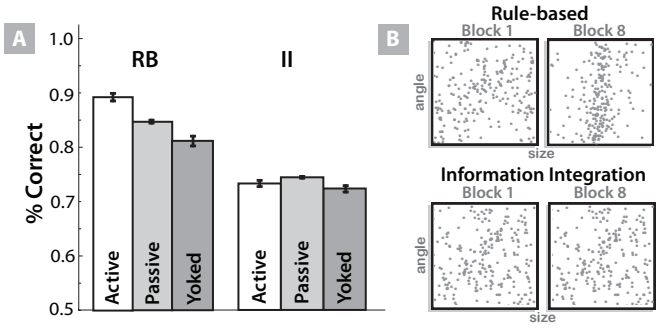


Figure 4: **A**: Expected accuracy of models trained on subject data. Active models which consider multiple hypotheses show an improvement in accuracy. **B**: Samples generated by the models during learning reflect the distribution of hypothesized rules; RB models are initially random but converge on the correct rule by the final block, while II models remain widely dispersed at both points in training.

which reflects how deterministic the decision rule is. Thus, each parameter combination $\{\mathbf{w}, b, \sigma\}$ reflects a unique decision rule or hypothesis about the category. The likelihood of a particular set of labeled observations $\mathcal{D} = \{o^1, .., o^t\}$ is given by $P(\mathcal{D}|\mathbf{w}, b, \sigma) = \prod_t P(o^t|\mathbf{w}, b, \sigma)$ (see Courville et al., 2003 for a similar approach). This basic model is equivalent to an equal variance Gaussian mixture model with two components.

We assume that learners are strongly biased toward unidimensional rules along either dimension. Accordingly, we defined a prior over the decision weights $\mathbf{w} = \{w_1 = cos(\theta), w_2 = sin(\theta)\}$, where θ is the angle of the vector corresponding to the decision boundary. We created a piece-wise scheme for translating θ into relative distances (bound between 0 and 1) from the horizontal axis:

$$r = \begin{cases} (2\theta)/\pi & : 0 < \theta \leq \frac{\pi}{2} \\ (2(\pi - \theta))/\pi & : \frac{\pi}{2} < \theta \leq \pi \\ (2(\theta - \pi))/\pi & : \pi < \theta \leq \frac{3\pi}{2} \\ (2(2\pi - \theta))/\pi & : \frac{3\pi}{2} < \theta \leq 2\pi \end{cases} \quad (2)$$

with $r \sim Beta(\alpha, \beta)$. Using this form, α and β act as a type of abstract attention weight (i.e., $\alpha = \beta < 1$ result in a general preference for rules along a single dimension. $\alpha, \beta < 1$ but $|\alpha - \beta| > 0$ results in a slight preference for one stimulus dimension over the other. $\alpha = \beta = 1$ implies no preference for rules of a particular orientation). The prior over the bias term was a Gaussian centered in the middle of the stimulus space, $b \sim N(0, 75)$, and the prior on the noise parameter was $\sigma \sim Beta(2, 1)$ (implying a mild preference for deterministic rules). Given these priors and the likelihood given in Equation 1, it is possible to infer the posterior distribution over the model parameters using Bayes rule. However, since full Bayesian updating in such a model is intractable, we assume that participants maintain an impoverished representation of the posterior distribution using a small set of point estimates from the posterior (similar to Sanborn et al., 2006).

At a given point in time we assume the learner has in mind a decision rule which can be characterized by parameter set $p^t = \{\mathbf{w}^t, b^t, \sigma^t\}$. On each trial, a new set of parameters p^{t+1} is proposed (or generated) which represents a change to the current rule. The learner is assumed to compare this new hypothesis to the old one and "accept" it as the new hypothesis if it provides a better account of the data (weighted by the prior belief in that parameter combination). If the new hypothesis results in a worse account of past data, it is accepted in proportion to the relative posterior likelihood of the new hypothesis compared to the old, otherwise the current parameter estimate remains unchanged. This procedure is similar to the Metropolis-Hastings algorithm (a form of Markov-Chain Monte Carlo) with an additional parameter k dictating the likelihood of accepting a proposal with a lower posterior estimate, giving the acceptance function $P(\mathcal{D}|p^{t+1})/(P(\mathcal{D}|p^t) + k)$. Proposals were generated from independent Normal distributions centered on the current parameter estimates: $\mathbf{w}^{t+1} \sim N(\mathbf{w}^t, \pi/2)$; $b^{t+1} \sim N(b^t, 20)$; $\sigma^{t+1} \sim N(\sigma^t, .05)$. The computational demands of this procedure are low: the learner is assumed to maintain a single hypothesis at any point in time. On each trial they must simply generate a new hypothesis and judge its relative quality. While we began with the simplification of assuming that the learner considers a single hypothesis on every trial, it is also possible that participants consider multiple hypotheses which are simultaneously updated in the same way.

Finally, we assume that the learner only stores n recent observations in memory, and evaluates the likelihood of a hypothesis over this limited set. This limitation results in ongoing shifts in the estimated decision rule, consistent with the variability in participants' response behavior throughout the task. Given the strong prior favoring rules along a single dimension, the estimate of the decision weights $\mathbf{w}$ will tend to bounce between these different modes of the hypothesis space, and convergence on the correct mode will be sensitive to the usefulness of recent training samples. This incremental, top-down hypothesis search may explain divergences between training conditions seen in our empirical results.

Evaluation of the Model The first goal of the simulation was to reproduce the difference in performance between the RB and II tasks. Individual models were trained with the data from passive-normal and passive-yoked participants in our experiment (the active group is addressed below). For this initial simulation the following parameter settings were used: $\alpha = \beta = 0.001, k = 1, n = 4$. Expected accuracy on each test item was calculated using the predicted likelihood that the item belonged to the correct category. Expected accuracy was averaged over test blocks and across 100 runs. The comparison of passive models (Figure 4A) shows a strong difference in accuracy between RB and II tasks as seen in our behavioral results. Due to the strong prior bias toward single-dimensional rules, in the RB task the model quickly converges on a rule similar to the true boundary, despite only retaining a small number of recent observations. In the II task, however, the model alternates between single-dimensional rules on different dimensions.

One way that the model can account for differences between active and passive-yoked groups is by assuming that active participants represent more than one hypothesis at any given time (consistent with the generalized "engagement" hypothesis described in the Introduction). In the model, this might correspond to an increase in the number of point estimates of the posterior maintained in working memory. To evaluate this idea, active participants were modeled using a set of 5 posterior samples per run (in contrast to one sample used for the passive groups), with the additional assumption that learners classify items according to the most likely hypothesis from the set they are considering. As seen in Figure 4A, the greater number of samples leads to higher accuracy over the passive groups in the RB task, but not in the II task. While a change in the number of particles maintain considered is consistent with the idea that active learners are more cognitively engaged in the task (and thus search the hypothesis space more effectively), further work is needed to directly test this representational hypothesis. At the very least, the potential divergence between the sequence of data observed in the task and the sequence of hypotheses considered by the learner provide a potential mechanism for explaining the active/passive-yoked distinction.

Finally, we were interested if samples generated by the active models show the same pattern as produced by our participants. Simulated samples were generated using *margin sampling*, in which an observation is most likely to occur when its predicted likelihood of belonging to category A and B are equal (i.e., the likelihood of making an observation o^t was proportional to $1 - |P(o^t = A|\mathbf{w}, b, \sigma) - P(o^t = B|\mathbf{w}, b, \sigma)|$). As seen in Figure 4B, the predict sampling distribution qualitatively matches the behavioral results. In the first block of both tasks, the model produces samples that are widely dispersed throughout the feature space. By the final block, RB models have converged on the true boundary, querying the margin of the boundary where uncertainty is greatest. In the II task, the diffuse distribution of samples reflects the variability in the hypotheses under consideration.

Conclusions

In our experiment, active learners were able to make informative queries to support their own learning, but this ability was more successful for RB categories than for II categories. Our simulation results explain this difference in terms of a bias toward considering rules along a single dimension. In addition, we evaluated one explanation for the divergence between active and passive-yoked participants, namely that active participants consider a greater number of hypotheses about the latent category structure. Our general finding that the effectiveness of active sampling may depend on the structure of the category adds to recent work examining active learning in binary classification tasks (Castro et al., 2008). While a number of theorists have attempted to explain active data selection in terms of optimal information gain (Oaksford & Chater, 1994; Nelson, 2005), our results suggest that the ability to design useful queries is strongly limited by the hypothesis search process that guides learning. To the degree that participants prefer particular types of rules, their sampling behavior will tend to be sub-optimal when the target rule mismatches these expectations, a similar point made in analyses of active machine learning (Mackay, 1992; Settles, 2009). In summary, active learning may promote learning, but it works best when you have a strong and correct idea of what you are trying to learn.

References

Ashby, F., Alfonso-Reese, L., Turken, A., & Waldron, E. (1998). A neuropsychological theory of multiple system in category learning. *Psychological Review*, *105*(5), 442-481.

Ashby, F., Maddox, W. T., & Bohil, C. J. (2002, Jul). Observational versus feedback training in rule-based and information-integration category learning. *Memory & cognition*, *30*(5), 666–77.

Ashby, F., Queller, S., & Berretty, P. (1999). On the dominance of unidimensional rules in unsupervised categorization. *Perception & Psychophysics*, *61*, 1178-1199.

Bruner, J. (1961). The art of discovery. *Harvard Educational Review*, *31*(21-32).

Castro, R., Kalish, C., Nowak, R., Qian, R., Rogers, T., & Zhu, X. (2008). Human active learning. In *Advances in neural information processing systems* (Vol. 21). Cambridge, MA: MIT Press.

Courville, A., Daw, N., Gordon, G., & Touretsky, D. (2003). Model uncertainty in classical conditioning. *Advances in Neural Information Processing Systems*, *20*.

Gureckis, T., & Markant, D. (2009). Active learning strategies in a spatial concept learning game. *Proceedings of the 31st Annual Conference of the Cognitive Science Society*.

Heller, K., Sanborn, A., & Chater, N. (2009). Hierarchical learning of dimensional biases in human categorization. In J. Lafferty & C. Williams (Eds.), (Vol. 22). Cambridge, MA: MIT Press.

Kruschke, J. (1993). Human category learning: Implications for backpropagation models. *Connection Science*, *5*, 3-36.

Kruschke, J. (2008). Bayesian approaches to associative learning: From passive to active learning. *Learning and Behavior*, *36*(3), 210-226.

Kuhn, D., Black, J., Keselman, A., & Kaplan, D. (2000). The development of cognitive skills to support inquiry learning. *Cognition and Instruction*, *18*(4), 495–523.

Lagnado, D. A., & Sloman, S. (2004, Jul). The advantage of timely intervention. *Journal of experimental psychology Learning, memory, and cognition*, *30*(4), 856–76.

Mackay, D. (1992). Information-based objective functions for active data selection. *Neural Computation*, *4*, 590-604.

Nelson, J. (2005). Finding useful questions: On bayesian diagnosticity, probability, impact, and information gain. *Psychological Review*, *112*(4), 979-999.

Nosofksy, R., & Palmeri, T. J. (1998). A rule-plus-exception model of classifying objects in continuous-dimension spaces. *Psychonomic Bulletin & Review*, *5*(3), 345-369.

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, *101*(4), 608-631.

Sanborn, A., Griffiths, T., & Navarro, D. (2006). A more rational model of categorization. In R. Sun & N. Miyake (Eds.), *Proceedings of the 28th annual meeting of the cognitive science society.* Mahwah, NJ: Erlbaum.

Settles, B. (2009). Active learning literature survey. *Technical Report*.

Sobel, D., & Kushnir, T. (2006). The importance of decision making in causal learning from interventions. *Memory and Cognition*, *34*(2), 411.

Steyvers, M., Tenenbaum, J., Wagenmakers, E., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science*, *27*(3), 453-489.

SARKAE – Modeling the Co-Evolution of Event Memory and Knowledge

Angela B. Nelson (a2nelson@ucsd.edu)
Department of Political Science, University of California San Diego, 9500 Gilman Dr. #0521
La Jolla, CA 92093 USA

Richard M. Shiffrin (shiffrin@indiana.edu)
Department of Psychological & Brain Sciences, Indiana University, 1101 E. Tenth Street
Bloomington, IN 47405 USA

Abstract

We present an overview of a model for the co-evolution of knowledge and event memory. The model, termed SARKAE (Storing and Retrieving Knowledge and Events), describes the development of knowledge and event memories as an interactive process: knowledge is formed through the accrual of individual events, and the storage of an individual episode is dependent on prior knowledge. We reference two experiments which provide data to inform our theory: these studies involve the development of new knowledge, and then testing in transfer tasks involving episodic memory, retrieval from knowledge, and perception. The results of the transfer tasks indicate a substantial role of pure frequency or raw exposure, in opposition to the contextual diversity accounts of frequency suggested by Adelman et al (2006). An overview of the SARKAE model is presented. The model is able to account for the effects of frequency in the absence of contextual diversity.

Keywords: episodic memory; semantic memory; learning; perception; Bayesian models.

Introduction

The processes involved in the accumulation of knowledge and the formation of event memories are interdependent. Almost every study since the 1890s has shown that the way episodic (or event) memories are encoded depends on the knowledge (or semantic memory) of the individual who is encoding them. Conversely, an individual's knowledge must be formed through the episodes they encounter; this idea was the basis of the REM model's account of priming (Shiffrin & Steyvers, 1997). These interdependent processes create a feedback loop in which knowledge and episodic memory formation develop together over lifelong learning.

Studies of memory and perception in the recent past have provided strong support for the idea that memory processes are robustly influenced by prior experience with the to-be-remembered content. Priming studies, for example, have shown that prior study of a word affects how well that word is identified in a forced choice perceptual identification task (Ratcliff & McKoon, 1997). The REMI model of Schooler, Shiffrin, and Raaijmakers (2001) accounts for these effects through a process in which the lexical representation (or knowledge) of the word is changed through prior study (the "prime"); when a word is studied an event memory is formed, but in addition, novel features of the event, such as the context of the experimental setting, are added to the lexical representation of the word. When the studied word is then presented for perceptual identification, the context tends to be similar to that at study, increasing the match of the probe cues to the lexical trace, predicting a variety of measurable effects that match those observed. In other words, the knowledge that a subject has about a stimulus, and the inclusion in that knowledge of factors like the experimental context, affect the way that a stimulus is perceived.

There are many models of the storage and retrieval of event memories, and sometimes the addition to existing knowledge of information from recent events (e.g. - Raaijmakers & Shiffrin, 1981, Shiffrin & Steyvers, 1997, Howard & Kahana, 2002, Anderson, 1983). The temporal context model of Howard & Kahana for example provides an explanation for recency and contiguity effects through the storage of both item information and recent contextual information. Other models, such as the ACT-R model of Anderson (1983) also provide eloquent representations of memory storage and retrieval. A few models attempt to explain aspects of the way events produce knowledge, especially for aspects of the role played by words in language (e.g. McClelland & Rumelhart, 1981). However, most of the prior research has been aimed to explain memory and learning when knowledge has already formed (to various degrees). Previous work by Reder et al. has examined the development of knowledge on a set of pseudowords, and used the dual-process SAC model to explain their findings. Although highly relevant and useful in the development of our research, the training study and modeling by Reder et al. did not explicitly model the growth of new knowledge.

Our aim is the development of a model that begins to explain the interacting growth of event memory and knowledge, as they influence both memory storage and retrieval. This co-evolution of the two systems was the focus of the REM-II model, created by Mueller and Shiffrin (2006). In this model, knowledge (or semantic memory) is represented as an accumulation of the co-occurrence of features: Features that are present in an episodic event are coded as occurring together in a matrix representation of semantic memory. REM-II is a quite powerful model, but a simplified version is sufficient to explain the basic concepts

by which event memory and knowledge co-develop, and is sufficient to model the empirical results presented in this paper. However, even a simplified model when applied to five different tasks spanning the range of learning, memory, and perception can grow to appear quite complex. The simplified model uses a representation in which each (separate) trace, whether an event trace or a knowledge trace, is a vector of feature values. Rather than term the model some other variant of REM, we use the terminology "Storing and Retrieving Knowledge and Events", abbreviated SARKAE.

Role of Experience and Frequency in Cognition

If one hopes to develop a theory in which events accumulate to form knowledge, then it is critical to understand the role of event frequency. Such effects are omnipresent in memory and perception tasks, but the processes responsible for such effects remain in debate.

Researchers have explored the effects of experience in various ways, typically by analyzing existing knowledge, identifying stimuli with different histories of experience, and using the stimuli with different frequencies in memory and perception tasks. The great majority of such investigations use words as stimuli: Words are categorized based on their frequency. Frequency is defined as normative occurrence in the environment, and these frequencies are estimated from various databases of typically textual materials. Words differing in frequency are then tested and exhibit a variety of consistent differences. These are termed the 'Word Frequency Effect', especially when found in recognition memory (Glanzer & Adams, 1985). In episodic recognition memory tasks, words that occur rarely in the environment are recognized *better* than words that occur frequently in the environment. Word frequency has also been shown to have effects on recall performance (high frequency words are recalled better), and perceptual tasks such as lexical decision and perceptual identification (forced choice, etc.).

However, given that word frequency is correlated with so many other variables (e.g. meaning, regularity of spelling, length of the word, and virtually every other characteristic one can measure for words), it is hard to know whether experience per se is responsible for the observed effects. In fact, a current debate concerns whether frequency per se or context effects are the primary cause of the observed findings. Adelman, Brown, and Quesada (2006) for example suggest that the diversity of contexts in which a word has been seen is a more accurate predictor of word frequency effects than the actual frequency of the word. By analyzing a large corpora of texts separated both by word frequency and contextual diversity (the number of documents in which a word was present), they concluded that it was the contextual diversity of an item, not the word frequency, that affected response times in word naming and lexical decision for three separate data sets. The difficulty of assessing the cause of frequency effects for words is one reason we chose to vary frequency of training of novel characters in the present studies. By training novel stimuli we can control with far greater precision the factors correlated with frequency and thereby properly constrain the theory. The studies referenced in this article create experience differences over a fairly lengthy period of training in two quite different tasks, one based on visual search, and the other based on perceptual matching.

In order to control for the confounds produced by word stimuli, our studies use stimuli that are far less related to existing language and numeric knowledge, and far less likely to bring with them existing frequency correlations: Chinese characters. (We select participants for whom such stimuli are unfamiliar). The first study used a visual search task in training. This task was based loosely on that of Shiffrin and Lightfoot (1997). Different Chinese characters appeared with widely differing frequencies during training. Following training, the subjects completed various recognition memory and perception tasks different from the training task, using both the trained characters and new characters as stimuli.

In the interest of space, this first experiment using the visual search training will not be discussed in detail. It is sufficient to mention that the crucial finding of this study was that substantial frequency effects occurred for all transfer tasks. What is more relevant to the discussion of the no-context experiment described below (as well as the SARKAE model) is that the visual search task used for training varied character frequency, but the randomization of trials and foils ensured that higher frequency characters most often occurred in the spatial and temporal vicinity of other higher frequency characters. Thus frequency per se was correlated with what could be termed character context, temporal context, or character diversity. As mentioned previously, Adelman et al. (2006) proposed that only the diversity of contexts in which an item occurs is responsible for most frequency effects. The confounding of frequency and character context made inference about causal mechanisms uncertain, and hence led to the design of the No-Context Experiment described below.

No-Context Experiment

The no-context experiment used a training paradigm not involving visual search. Participants were trained using a same vs. different judgment task: A character was presented briefly twice in succession, and half the time the two presentations varied slightly in size, rotation, or contrast. The participant judged whether the two presentations were exactly the same or varied slightly in one of these three dimensions. Thus a character was its 'own' context. Further, to remove the possibility that the test character on the previous trial might provide context for the present trial, one fixed 'control' character, different from any of the experimental characters, was tested using the same judgment task between every two experimental character judgments. This extremely high frequency character was not subsequently used in the post-training transfer tasks. If

context is carried forward from the previous trial during training, the context that is carried forward for the experimental characters of different frequency will be equated, because the previous character is always the same one. The no-context experiment used the same frequency distribution (given below) as the visual search training experiment. By removing characters that provide context on any given trial, and by holding constant the character context on the preceding trial, it is plausible to assume that the confound between context and frequency is mostly if not totally eliminated.

Training Methods

Participants. Seven participants, recruited with an email advertisement, participated in the experiment for monetary compensation. All participants reported no prior experience with Chinese characters.

Design and Stimuli. The occurrence of the characters in the same/different task was manipulated to produce four frequency conditions which varied in a ratio of 1::3::9::27. For each subject, a set of 32 characters was selected randomly from a pool of approximately 200 characters. From these 32 characters, 8 were assigned to each frequency condition. In order to keep the complexity of the characters reasonable, all the characters in the pool were composed of 7 strokes or less. In order to fully eliminate context from the training, one "super-high frequency" item was also randomly chosen, making the entire training set 33 characters. This character appeared as a "buffer" item every other trial, and was not used as a stimulus in the post-training tasks.

Procedure. Each trial consisted of two brief (500 ms) presentations of a single Chinese character, which subtended a visual angle of approximately 4.3 x 4.3 degrees. The two presentations of the character were either identical or varied slightly in size, rotation, or contrast of the character. Only one of these three dimensions varied at a time. There were three levels of each variable (size: small, medium, large; rotation: left, straight, right; contrast: dark, normal, light), and the change between each of these levels varied based on a staircase algorithm. The staircase rules were as follows: when the subject answered two rotation-difference trials correctly, the rotation factor (i.e. – the difference in angle between the three levels) decreased by a given amount. If they got a rotation-different trial wrong, the rotation factor increased by a given amount. This staircase was done separately for each of the three variables. In this way, subjects were kept at approximately 75% accuracy. Subjects completed 12 training sessions, approximately 3 per week. There were a total of 1060 trials for sessions 1-11, and 1140 trials for session 12.

Training Results

Since the training paradigm used a staircase algorithm to keep subjects at approximately 75% accuracy, the results of training were analyzed by examining the change factors for size, rotation, and contrast. If the subjects are showing improvement at the same/different discrimination, then the change in variable (size, rotation, or contrast) needed to keep them at 75% should decrease over session. Figure 1 shows the mean rotation, contrast, and size changes required (averaged over all subjects) as a function of training session. The results indicate that subjects were becoming more efficient at the task as training progressed, as indicated by the decrease in variable change over session.

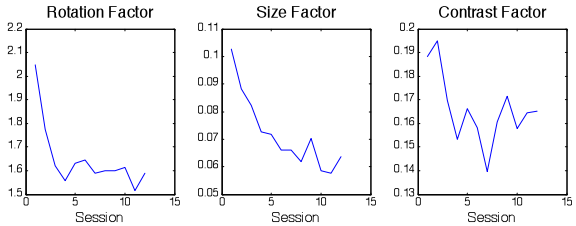


Figure 1: Mean change in rotation (panel A) size (panel B) and contrast (panel C) needed to obtain 75% accuracy as a function of training session. Rotation factor is measured in degrees, size factor in percentage size difference, and contrast factor in percentage contrast difference.

Post-training Tasks

Following the training, the subjects completed three post-training tasks: pseudo-lexical decision, episodic recognition, and forced-choice perceptual identification. Testing was carried out again six weeks after training. A programming error, discovered after the immediate transfer tasks, caused the forced choice data to be very noisy and essentially uninformative. These results are therefore neither reported nor analyzed. Also, because forced choice results were not available for immediate test, forced choice testing was omitted for the delayed testing at six weeks.

Pseudo-lexical Decision

Design and Procedure. Subjects viewed one list, which contained all 32 trained characters (excluding the buffer item), as well as 32 new characters. Each of these characters occurred 3 times throughout the list, making the total length of the list 192 characters. Subjects were presented with a single character on the screen, and were asked to decide (by keypress) as quickly as possible whether they had ever seen that character during any of the previous training sessions.

Results. Response time and accuracy were measured for each frequency condition, as well as new items. The results for the trained items when tested shortly after training was completed (2-3 days) are shown in Figure 2. A contrast analysis showed that there was a significant negative relationship between frequency and response time ($t(6)=-2.97$, $p=.03$) , and a significant positive relationship between frequency and accuracy ($t(6)=2.90$, $p=.03$) .

Response time and accuracy were measured again (for 6 of the 7 subjects) approximately 6 weeks after the previous test session. The results followed the same qualitative pattern as they did 6 weeks prior: there was a significant negative relationship between response time and frequency ($t(5)=-2.45$, $p=.058$), and a significant positive relationship between accuracy and frequency ($t(5)=2.44$, $p=.059$, see Figure 2). A contrast analysis showed that there was no significant difference in the magnitude of the effects that occurred in the shortly after training and those that occurred after the 6 week delay for either accuracy ($t(5)=1.14$, $p=.31$) or response time ($t(5)=.51$, $p=.63$).

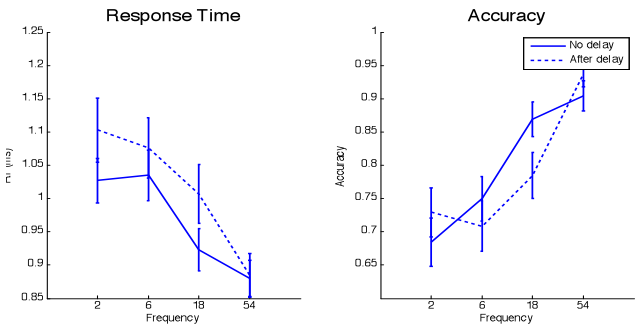


Figure 2: Mean response time (panel A) and accuracy (panel B) for all subjects in the lexical decision task as a function of frequency. The solid line shows the results when the test was administered after a very short delay (2-3 days), the dashed line corresponds to the data following a 6 week delay.

Discussion. The results of the lexical decision task showed that the absence of character-context during training did not eliminate the effects of frequency on speed and accuracy of decision. Therefore, it follows that there must be some mechanism other than the context present during training that is causing improved recognition that high frequency characters are present in knowledge. In addition, this frequency effect showed little signs of reduction over six weeks.

Episodic Recognition

Design and Procedure. The task consisted of eight pairs of study and test lists. Each study list contained eight trained characters (two from each frequency category) and eight untrained characters. Each test list contained all the items from the study list as well as 16 unstudied items, which included eight trained characters (two from each frequency category) and eight untrained characters. The first four items on the test list were always untrained characters, providing a buffer for the items of interest (trained characters). Subjects viewed each item on the study list for 1000 milliseconds, presented one at a time on the screen. Following the study list, the subjects were presented with the items on the test list one by one, and for each item had to respond whether the character had been present on the list they had just studied. Subjects were instructed to 'reset' their memory in between each list, and answer 'old' to an item on the test list only if it had been present on the most recent study list.

Results. The data from the episodic recognition task were analyzed by examining the hit rates (correctly identifying a studied item as old) and false alarm rates (incorrectly identifying an unstudied item as old). The hit and false alarm rates (averaged over all subjects) are plotted as a function of frequency in figure 3. When tested shortly after the completion of training, false alarms significantly increased as frequency increased (panel A, $t(6)=3.19$, $p=.02$). There was also a marginally significant decrease in d' due to frequency ($t(6)=-1.86$, $p=.11$). The hit rate analysis however showed no significant effect of frequency.

Six of the seven subjects were tested again following a six-week delay. The results of the delayed test are shown in panel B of figure 3. Statistical analyses showed no significant effect of frequency on hit rates, false alarm rates, or d'. Furthermore, a contrast analysis showed that there was a significant difference in the magnitude of the false alarm rate effect found immediately after training compared to the effect found after a 6 week delay: the increase in false alarms due to increased frequency was (marginally) significantly larger immediately after training ($t(5)=2.11$, $p=.09$).

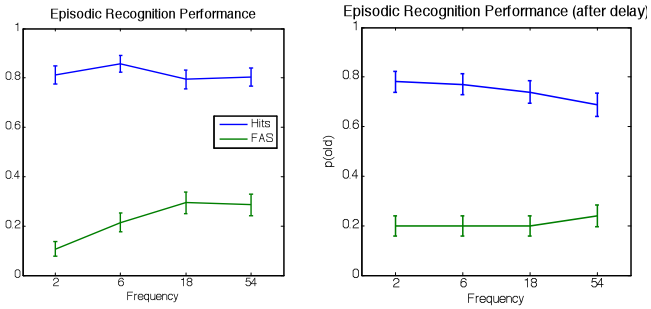


Figure 3: Episodic Recognition Results soon after training (Panel A) and after a 6-week delay (Panel B). Hit rates are shown in blue, false alarm rates in green.

Discussion. When tested shortly after the completion of training, the results in the episodic recognition task are similar to results found in our previous visual search training experiment and in normative word frequency studies: as frequency increases, d' decreases. In the current study, this is due more to an increase in false alarm rates than a decrease in hit rates with higher frequency items. Unlike some previous studies, the no-context training experiment did not show a significant effect of frequency on hit rates.

Unlike the lexical decision task which showed a large persistence of frequency effects after a six week delay, the d' effect and false alarm rate effect found in episodic recognition were largely reduced and possibly absent when

subjects were re-tested after delay. Both the existence of frequency effects in recognition, and the reduction with delay call into serious question the modeling processes used to account for recognition in the one factor model applied to our visual search training experiment. That model assumed poorer performance for high frequency test items was due to increased confusions with traces of list items, because those traces were more similar to the high frequency test probes. The present design should have eliminated such similarity differences. In addition, within list confusions should not have decreased if a recognition task was carried out at a six week remove from training, because the relevant episodic traces should have been those stored in the just seen study list. Thus the elaborated SARKAE model provides an explicit role for frequency per se (especially to explain pseudo-lexical decision findings) and an elaborated model for recognition. Due to spatial limitations, in this paper we present only an overview of the theory that is the foundation of the SARKAE model, with examples of how the theory is implemented to explain our experimental results.

SARKAE – Theoretical Overview

A fundamental storage assumption in SARKAE allows both event memories and knowledge to develop in concert: Each storage episode produces both: 1) an event trace; 2) additional information added to traces in memory that are brought to mind due to similarity to the present event. Such a prior trace can include a previous event trace (the basis for the start of knowledge accumulation), or a developing or mature knowledge trace. There is no fundamental distinction between event traces and knowledge traces in this view. Instead there is a continuum: traces are stored initially for each single event; some of these are retrieved (when a similar new event occurs), gain additional information, and are re-stored. As this process continues over successive occurrences of similar events, a rich knowledge trace results.

In SARKAE, accumulation of knowledge about an item or concept (e.g. for words, its lexical entry) includes features of the surrounding context that is present at the time of learning. Specifically, knowledge traces develop during learning by storing features that come both from the physical properties of the item or concept being learned, and also from the context surrounding the item during learning, both types of storage being modified and governed by attentional focus. These context features arise from other (attended) events nearby in time and the environment, and from the various components of internal and external context that numerous investigators have discussed for many years. For example, during training, when a character is presented, physical features of that item as well as surrounding context features (taken from other characters presented in close temporal proximity) are stored into the knowledge representation. In a more general sense, the knowledge trace that represents the concept of "table" will include information about the physical properties of various types of tables, information about the contents of events that involved tables (e.g. forks, dinners, conversations, replacing light bulbs), information about thoughts and feelings experienced at tables, and information about other events that occurred in the nearby temporal surround of table events (e.g. dropping of a milk bottle when removing it from the refrigerator). These features include context specific events themselves, such as the breakfast event in a given morning. Knowledge development is therefore built upon the events that accumulate to form the knowledge. Of course a mature knowledge trace includes features of numerous events, so a specific episode tends to be swamped in the accumulation of many episodes and tends not to be retrieved (from the knowledge trace—it can be retrieved as an episodic trace). Thus a knowledge trace in most instances seems to be context free. What do come to be retrievable from a mature knowledge trace are features that are consistent across many episodes, such as the spelling, pronunciation and meaning of a word.

Conversely, the formation of episodic memory traces is determined by prior knowledge and experience. Although certain very primitive features of experience might not depend upon learning and experience (e.g. a loud sound), most features of events are encodings based on prior learning (e.g. encoding and storing a table feature as 'dinner'). The model therefore creates episodic traces by choosing features of events from knowledge. Such features come from several sources: some are directly related to the central defining elements of the event such as the physical features of which it is composed (e.g. table physical features) and the central organizing concept (e.g. dinner); some come from other knowledge traces that are brought to mind during encoding of the event (e.g. the illness one encountered when eating breakfast last Sunday, or one's commitment to a new diet); some come from features of other nearby events still in short-term memory at the time of the present event. To some degree, the features chosen are modified by attentional focus. In terms of the experiment discussed in this paper, an episodic memory consists of a combination of physical features of the studied item, features drawn from the knowledge trace of that item, and features drawn from other items in close temporal proximity. One key concept is the perhaps non-controversial idea that the features comprising an event representation in short-term memory, and thereafter the stored event trace, are recruited from knowledge (e.g. one's prior experience and knowledge regarding tables will influence the formation of an event trace concerning a physically present table).

We have been highlighting mechanisms that produce storage of event memory and knowledge. Very similar mechanisms also occur in retrieval. We adopt the generally accepted view that retrieval is cue dependent, and based on similarity of the retrieval probe to the traces in memory. The generation of such a probe cue can be clearly defined, as when one is asked: "What is the capital of South Dakota"? In other cases retrieval seems more continuous and automatic, as when information moving through short-term

memory acts as retrieval cues to bring other associations to mind. However, because modeling continuous retrieval is quite complex, we will treat all retrieval in terms of discrete retrieval operations occurring one at a time, each based on some defined set of retrieval cues. The features that comprise such a retrieval cue are generated with the same processes that generate features for storage: They come from the query (if there is one), or from feature sets presently in short-term memory and attentional focus, and include features from the contextual surround at the time (internal and external context, and nearby events). More specifically, in the modeling of our experimental results, the retrieval cue consists of a combination of physical features of the test item, features drawn from the knowledge trace of the test item, and features taken from other items in close temporal proximity to the test item.

An absolutely essential component of storage and retrieval is noise in the processes. Following the approach in the REM model, we assume that both storage and retrieval are probabilistic, incomplete and error prone. When errors are made, it is natural to assume they are based on information in the knowledge base, and not completely random. Thus errors in retrieving and storing features are assumed to be relevant and consistent, in the sense that they are feature values for the feature in question (a 'blue' color feature might be retrieved or stored as 'red', but not as 'wet') and occur in proportion to the base rates of such values in knowledge.

When a cue is used to probe memory, it is compared in parallel to the event traces (and/or knowledge traces) in parallel. It would be unworkable and likely unreasonable to explicitly consider the match to each of the essentially uncountable traces in memory. Thus we assume that there is a probabilistic cutoff, only traces sufficiently similar to the probe becoming activated and participating in subsequent retrieval operations.

Similarity plays a role in both storage and retrieval, but we define similarity operations in such a way that similarity is measured as a relative construct: For both storage and retrieval a process based on similarity is defined as similarity of a given match compared to the similarity of matches that could have but did not occur. Thus in recent years we have characterized the match of a probe to an activated trace as a likelihood ratio: The numerator expresses the probability that the probe and cue were generated from the same event, and the denominator the probability that the two were generated by different events. These likelihood ratios occupy the theoretical niche played by 'strengths of activation' in various other theories (such as SAM; Raaijmakers and Shiffrin, 1981).

This brief summary of some of the central tenets of SARKAE provides hints concerning the theory, but is only the barest scaffolding upon which the model is constructed. When the full detailed processes are implemented, the model produces predictions that fit the results of the various post-training tasks from both the initial visual search training experiment as well as the no-context experiment described in this paper. We cannot fully describe the modeling processes and results here due to space; however the aim of this discussion is not to focus on quantifiable model fits, but instead to convey the basics of the theory that inspired both the experiments described in this paper and the subsequent model development. The SARKAE model provides plausible mechanisms by which knowledge grows from events, and knowledge informs the coding and retrieval of both events and knowledge itself. Based on this theory, or others of a similar character, we hope that future research developments will not focus so strongly on differences among systems as upon the ways they grow together, in highly dependent fashion.

References

Adelman, J. S., Brown, G. D. A., & Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. Psychological *Science, 17,* 815-824.

Anderson, J. R. (1983). A Spreading Activation Theory of Memory. *Journal of Verbal and Learning Behavior, 22*, 261-295.

Glanzer, M., & Adams, J. (1985). The Mirror Effect in Recognition Memory. *Memory & Cognition, 13*, 8-20.

Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology, 46,* 269-299.

McClelland J. L., & Rumelhart, D. E. (1981). An interactive activation model of the effect of context in perception, Part I. An account of basic findings. *Psychological Review, 88,* 375-407.

Mueller, S. T., & Shiffrin, R. M. (2006). REM-II: A Model of the developmental co-evolution of episodic memory and semantic knowledge. *Paper presented at the International Conference on Learning and Development (ICDL), Bloomington, IN, June 2006.*

Raaijmakers, J. G. W., & Shiffrin, R. M. (1981). Search of Associative Memory. *Psychological Review, 88,* 93-134.

Ratcliff, R., & McKoon, G. (1997). A Counter Model for Implicit Priming in Perceptual Word Identification. *Psychological Review, 104*, 319-343.

Reder, L. M., Angstadt, P., Cary, M., Erickson, M. A., & Ayers, M. S. (2002). A Reexamination of Stimulus-Frequency Effects in Recognition: Two Mirrors for Low- and High-Frequency Pseudowords. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 28,* 138-152.

Schooler, L. J., Shiffrin, R. M., and Raaijmakers, J. G. W. (2001). A Bayesian Model for Implicit Effects in Perceptual Identification. *Psychological Review, 108,* 257-272.

Shiffrin, R.M., & Lightfoot, N. (1997). Perceptual Learning of Alphanumeric-like Characters. *The Psychology of Learning and Motivation, 36*, 45-81.

Shiffrin, R., & Steyvers, M. (1997). A model for recognition memory: REM – retrieving effectively from memory. *Psychonomic Bulletin & Review, 4*, 145-166.

Looking at Nothing Indicates Memory Search in Multiattribute Decision Making

Frank Renkewitz (frank.renkewitz@uni-erfurt.de)
University of Erfurt, Center for Empirical Research in Economics and Behavioral Sciences
Nordhäuser Str. 63, D-99089 Erfurt, Germany

Georg Jahn (georg.jahn@uni-greifswald.de)
University of Greifswald, Department of Psychology
Franz-Mehring-Str. 47, D-17487 Greifswald, Germany

Abstract

The common methods for studying heuristics in memory-based multiattribute decisions provide outcome and response time data but leave the foregoing cognitive processes in the dark. We demonstrate a novel process-tracing method that uses the looking-at-nothing phenomenon to study memory search and cue processing via eye tracking. Participants learned cue information of decision alternatives in spatial frames and later were presented with emptied displays of two alternatives in binary choice trials. With freely chosen and with instructed decision strategies, fixation patterns on former cue locations were in line with memory search and cue processing as postulated for lexicographic and compensatory strategies.

Keywords: Multiattribute decision making, Probabilistic inference, Eye tracking, Take-the-best heuristic, Spatial index

Introduction

Integrating multiple cues in decision making in a completely rational manner that factors in all available information is widely assumed to be cognitively too demanding. Therefore, simple but efficient heuristics are suggested that are applicable in specific task domains (e.g., Gigerenzer & Todd, 1999). The domain we will deal with here consists of simple probabilistic inference tasks with two alternatives and binary cues (e.g. "Which of the cities A and B has more inhabitants?"). Perhaps the best known heuristic for this kind of tasks is Take-the-Best (Gigerenzer & Goldstein, 1996), which belongs to the class of lexicographic strategies (LEX). In a first step, a person using this strategy selects the (subjectively) most valid cue and looks up the cue values of both alternatives. If one alternative has a positive cue value (indicating a higher value on the target dimension) and the other has not, information search is stopped, and the alternative with the positive value is chosen. If the first cue does not discriminate between the alternatives, the second most valid cue is accessed, and so forth.

LEX is a non-compensatory strategy, because cues lower in validity are disregarded. Thus, cues with a low validity cannot compensate for a difference on a more valid cue dimension. An example of a simple but compensatory heuristic is the equal weight strategy (EQW), which ignores cue validities. With this strategy, the positive cue values are counted for each alternative and the alternative with the higher number of positive cue values is chosen.

The third strategy we will consider here - the weighted additive rule (WADD) - is also compensatory but uses information about cue values and validities fully. Cue values are weighed by cue validities and summed up. It has often been claimed that this strategy is computationally too demanding to be performed in a sequence of serial and deliberate cognitive processes. However, WADD can be conceived as a paramorphic model because the choices expected from WADD could be brought about by intuitive-automatic processes as well. Automatic processes that approximate WADD predictions have been simulated as parallel constraint satisfaction (see Glöckner & Betsch, 2008).

Methodologically, it poses a challenge to infer which strategy an individual used in a probabilistic inference task. One reason is that in tasks with a small number of alternatives, several strategies predict the same choices. More importantly, simple strategies are complete sub-models of more complex ones. With appropriately chosen weights, WADD could produce decisions that are indistinguishable of the predictions of LEX or EQW (Martignon & Hoffrage, 2002). Hence, an individual's choices may appear to be generated by LEX even if she is using cognitive processes that are not assumed in this heuristic.

Therefore, process-tracing methods are often used in addition to outcome-based measures. The most prominent process-tracing method is the computer-based information board called Mouselab (Payne, Bettman, & Johnson, 1993). Mouselab records which pieces of information a participant seeks, in which sequence the information is accessed, and how much information is gathered. A central idea is that different decision strategies should be accompanied by different information search patterns. For non-compensatory strategies like LEX a cue-wise information search is expected. A LEX-user that has looked up a cue value of one alternative should subsequently check the value of the other alternative on the same cue dimension and then either decide or, if the values do not differ, switch to the next cue dimension. In contrast, compensatory strategies should be associated with an alternative-wise information search. Users of these strategies should search for all cue values of one alternative before they turn to the other alternative.

Process-tracing methods like Mouselab necessitate that all relevant information is provided by the experimenter and accessible for the participant on the computer screen (or on

written information cards). However, in many real-life situations decisions have to be made from information that is stored in long-term memory. Additionally, proponents of simple heuristics argue that particularly memory-based decisions induce selective and heuristic processing to limit the costs of memory retrieval (Gigerenzer & Todd, 1999). Hence, "inferences from memory" and not "inferences from givens" should be studied to test these postulates. Thus, there is a need for process-tracing methods that can be applied in studies investigating memory-based decisions.

We propose a method that sticks closely to the basic idea of Mouselab. It draws on results by Richardson and Spivey (2000) demonstrating a close link between eye movements and memory retrieval. In a series of experiments, these authors found that participants recalling visually presented information saccade more often to the (empty) region of space where the information was originally presented than to any other region (for an overview of studies on this "looking at nothing" phenomenon, see Ferreira, Apel, & Henderson, 2008). Thus, if specific cue dimensions are associated with specific locations it might be possible to reveal which cue information a person searches in memory by tracking her eye movements.

In the experiment reported below, our participants worked through a learning phase, in which each cue dimension was presented in a different fixed location of a spatial frame. In the decision phase, we presented two empty spatial frames next to each other and recorded participants' eye movements on these empty frames while they recalled cue information to decide between the two alternatives. Our objective was to test whether participants who were classified as LEX-users based on their decision outcomes showed different gaze patterns than participants classified as using compensatory strategies. More specifically, we report tests of the following three hypotheses:

1. As soon as a cue value on one alternative is found, LEX-users should look for the respective cue value of the other alternative. In contrast, users of compensatory strategies should search for complete cue information on one alternative first. Hence, LEX-users should switch their gaze more often between alternatives than users of compensatory strategies, that is, there should be more transitions between alternatives per second of a trial for LEX-users than for users of compensatory strategies.

2. Because memory-retrieval of LEX-users is more extensive when differentiating cues have lower relative validity, the absolute frequency of transitions between alternatives should increase linearly with the validity rank of the first differentiating cue. In contrast, for users of compensatory strategies the validity rank of the first discriminating cue should not affect the frequency of transitions.

3. LEX-users should disregard cues lower in validity as soon as a higher cue differentiates between alternatives. Hence, fixation durations on former locations of cue values lower in validity should be shorter in trials, in which a cue higher in validity differentiates. Again, users of compensatory strategies should not be affected by the validity rank of the first discriminating cue and the former locations of all cue values should be fixated for about the same amount of time.

Experiment

The experiment consisted of a learning phase, in which the participants acquired cue knowledge about six objects, followed by two decision phases. In the first decision phase, pairs of objects were presented to the participants for binary choice. In this phase, participants were not instructed with regard to the strategy they should use. Thus, the first decision phase followed a quasi-experimental logic. Here, the aim was to identify groups of participants who spontaneously employed different strategies and to test whether these groups show different patterns of gaze behavior as predicted in our hypotheses. We added a second decision phase to gain better control over the strategies participants used. In this phase, we presented the same binary choice items as in the first phase, but we directly instructed the participants to employ a certain strategy. Thus, with the second phase we realized a simple one-factorial design with two experimental groups (LEX-, EQW-instructions).

Method

Participants. Fifty-three students at the University of Greifswald participated in the experiment (43 women, 10 men; mean age 21.9 years). They received partial course credit for their participation. They were assigned randomly to the different strategy conditions in the second decision phase.

Materials. The participants learned cue descriptions of six alternatives. These alternatives were mushrooms characterized by the four cue dimensions consistency, cell wall material, mineral, and spread. Each of the cue dimensions could have three different values (consistency: soft, elastic or firm; cell wall material: protein, cellulose or lipid; mineral: magnesium, zinc or potassium; spread: frequent, medium or rare). The critical cue values (elastic, protein, magnesium and rare), which indicated a higher value on the target dimension (toxicity of the mushrooms), were not revealed to the participants until the learning phase was completed. In the decision phase, we presented pairs of the mushrooms for binary choice.

A complete paired comparison of the six cue patterns yields 15 choice tasks. The six cue patterns were constructed in a way that allowed for an individual strategy classification of the participants based on the vector of their choices in these 15 tasks (see Bröder & Schiffer, 2003a). Among the 15 binary choice tasks was a sufficient number of items for which each of the decision strategies considered here yields a distinct prediction. (A more detailed description of the cue patterns is given in Renkewitz & Jahn (2010), where we report an earlier experiment in which we used the same patterns.) To attain a more reliable strategy

classification, all 15 items were presented twice in both decision phases. For the second presentation of each item, the order of the alternatives was reversed. Thus, each decision phase consisted of 30 binary choice items after two practice trials.

Figure 1 exemplarily shows two alternatives and their cue descriptions as they were presented in the learning phase. Each mushroom was symbolized by a different geometrical figure. Written descriptions of the cue values appeared in four rectangular frames that were arranged along the borders of the geometrical figures. The position of the frames was constant across all alternatives. For a given participant, values on the same cue dimension were always shown in the same frame (e.g., the respective value of the cue "consistency" was presented in the lower left frame for all mushrooms learned by this participant). Thus, each cue dimension was tied to specific spatial coordinates. We counterbalanced the position of the cues across different validity ranks. For half of the participants, the cues were arranged clockwise in descending order of validity starting from the upper left frame. Hence the two most valid cues appeared in the two upper positions. For the remaining half of participants, the cues were arranged counter-clockwise starting from the lower left frame. Here, the two most valid cues appeared in the two lower positions. Additionally, the labels of the cues were counterbalanced across validity ranks. We used the two validity orders "consistency, cell wall material, mineral, spread" and "spread, mineral, cell wall material, consistency".

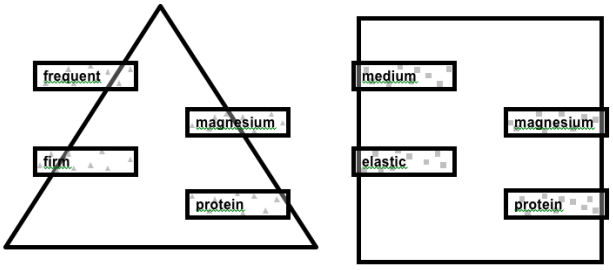


Figure 1: Two of the six alternatives as they were presented in the learning phase. In the decision phases, the rectangular frames containing the four cue values were empty.

In each trial of the decision phases, two of the geometrical figures were presented side by side. The size of the figures was the same as in the learning phase. The rectangular frames that contained the cue values in the learning phase were now empty. So, participants had to actively search their memory for cue information to be able to decide which of the two mushrooms was likely to be the more toxic one.

Procedure. At the beginning of the learning phase, the six alternatives and their cue values were presented one by one. Afterwards, the participants had to reproduce the cue values repeatedly in several testing cycles. The learning phase continued until the participants correctly reproduced at least 22 of the 24 cue values in a final memory test (see Renkewitz & Jahn, 2010, for more details on the learning phase).

Before the decision phases, the participants were informed about the validity ranks of the cues and the critical cue values. For instance, when the corresponding order of cue validities was used, participants were told that spread gave the most important hint to toxicity and that only rare mushrooms were likely to be toxic. The second most important hint was the mineral and only mushrooms predominantly containing the mineral magnesium were likely to be toxic, and so on. Additionally, a list showing the attributes of the 'typical toxic mushroom' was shown.

The 30 test trials of a decision phase were organized in two blocks. Each block consisted of the 15 binary choice items resulting from a complete paired comparison of the six mushrooms. Within each block, choice items were presented in random order. Participants responded with two keys on a standard keyboard.

After the first decision phase was finished, participants were given the strategy instructions for the second decision phase. In the LEX-condition, participants were told to consider attributes in descending order of importance and to decide according to the first attribute indicating that one of the two mushrooms was more toxic. All attributes of lower importance should be ignored. In the EQW-condition, participants were instructed to decide on the basis of the number of attributes indicating that a mushroom was toxic. If both mushrooms had the same number of 'toxicity attributes' they should guess. Subsequent to these explanations, an example of the application of the respective decision rule was given. In this example two cue patterns were used that were not part of the test set.

During both decision phases, eye movements were recorded with a desk-mounted SMI RED eye tracker sampling pupil position at 60 Hz.

Results

Behavioral data. We classified the decision patterns of the participants as most probably generated by LEX, EQW, WADD or a guessing strategy with the maximum-likelihood strategy classification method (Bröder & Schiffer, 2003a, 2003b). The strategy percentages in both decision phases are shown in Table 1. In the first phase, when there were no strategy instructions, 49% of the participants were classified as using the LEX-heuristic. This frequency of LEX-users corresponds closely to the findings of similar studies on memory based decision making using verbal stimulus material (Bröder & Schiffer, 2003b; Bröder & Schiffer, 2006; Bröder & Gaissmeier, 2007; Jahn, Renkewitz & Kunze, 2007). The classification results in the second decision phase suggest that the strategy instructions were largely successful. Thus, 80% of the participants instructed to employ the LEX-heuristic appeared to use this strategy and 89% of the participants instructed to use EQW were classified as using one of the compensatory strategies (EQW or WADD).

Table 1: Frequencies (and percentages) of strategy classifications in decision phase 1 (free strategy selection) and decision phase 2 (with LEX or EQW instructions)

Condition	Strategy classification					
	LEX	EQW	WADD	Guessing	unclassified	*N*
Decision phase 1						
Free	26 (49.1)	14 (26.4)	9 (17.0)	1 (1.9)	3 (5.7)	53
Decision phase 2						
LEX	20 (80.0)	1 (4.0)	1 (4.0)	2 (8.0)	1 (4.0)	25
EQW	2 (7.1)	18 (64.3)	7 (25.0)	1 (3.6)	0 (0.0)	28

Table 2: Mean absolute number of transitions between alternatives depending on the position of the first differentiating cue for groups of participants with different strategy classifications in both decision phases

Strategy	Validity rank of first discriminating cue							
	1		2		3		4	
	M	95% CI	M	95% CI	M	95% CI	M	95% CI
Decision Phase 1 (free)								
LEX	4.64	3.57-5.71	7.00	5.72-8.28	10.45	7.07-12.93	8.07	6.52-9.61
COMP	6.11	4.93-7.29	6.58	5.17-7.99	5.79	3.05-8.52	5.70	4.01-7.40
Decision Phase 2 (instructed)								
LEX	3.37	2.18-4.55	5.80	4.31-7.28	7.35	5.94-8.76	9.32	7.39-11.25
COMP	5.62	4.58-6.65	5.30	3.99-6.60	4.67	3.38-5.85	5.10	3.41-6.79

Analyses of eye gaze data. In the analyses of gaze behavior we did not consider participants who remained unclassified or were classified as using a guessing strategy because we held no hypotheses concerning these participants. In the first decision phase, we excluded one additional participant (classified as EQW-user) from further analyses because of her unusually long decision times. In the second decision phase, we restricted the analyses to those participants that appeared to follow the instruction to employ the LEX-heuristic and those participants who were classified as using one of the compensatory strategies under EQW-instructions. Finally, we discarded all trials (3.0% in the first phase and 4.4% in the second phase) from the analyses of gaze behavior, in which the tracking data for more than 40% of the trial duration were missing (due to blinks, looking off the screen, or lost pupil or corneal reflectance).

In all of the following analyses, we merged the eye tracking data of EQW- and WADD-users as the result pattern for both compensatory strategies was generally the same and no statistically significant differences occurred between these two groups.

Transitions between alternatives per second. To determine the number of transitions between alternatives, we defined two areas of interest (AOIs), each of which covered one alternative and, thus, almost one half of the screen. A transition was defined as two successive fixations in different AOIs. We counted the number of transitions per trial and divided this number by the trial duration (in seconds) to obtain an index of gaze transitions. The means of this index corroborate our first hypothesis: With instructed decision strategies, LEX-users (M = 0.68) switched their gaze faster between alternatives than users of a compensatory strategy (M = 0.46), 95% CIs [0.58, 0.78], and [0.41, 0.51], respectively, $t(43) = 3.52$, $p = .001$, $d = 1.08$. In the first decision phase, when participants spontaneously adopted a decision strategy, the same difference between the LEX-heuristic (M = 0.60) and compensatory strategies (M = 0.39) was found, 95% CIs [0.54, 0.66], and [0.34, 0.44], respectively, $t(46) = 4.90$, $p < .001$, $d = 1.45$.

Transitions between alternatives depending on the validity rank of the first discriminating cue. According to our second hypothesis, for LEX-users the absolute number of transitions between alternatives should depend on the validity rank of the first discriminating cue in a decision item. The lower the validity of the first discriminating cue the more transitions should occur. In contrast, the gaze behavior of users of a compensatory strategy should be unaffected by the validity rank of the first discriminating cue.

To test this hypothesis, we split the 30 decision items into four sets, according to the rank of the best discriminating cue. Table 2 depicts the mean number of transitions in each of the four sets for LEX-users and for users of a compensatory strategy in both decision phases. Under instructed strategy conditions, for LEX-users the mean number of transitions increased monotonically with the validity rank of the first discriminating cue, as expected. For users of compensatory strategies no systematic effect was associated with the validity of the best differentiating cue. This interaction effect of strategy classification and validity rank of the first discriminating cue was confirmed in a two-way mixed ANOVA, Greenhouse-Geisser corrected $F(2.28, 98.13) = 22.71$, $p < .001$, $\eta_p^2 = .35$. When participants chose freely between decision strategies, we found a similar result pattern. For LEX users the number of transitions again

increased markedly (but not monotonically) with the validity rank of the first discriminating cue. For users of a compensatory strategy this factor had no impact. The corresponding interaction effect was again statistically significant, Greenhouse-Geisser corrected $F(1.79, 82.25) = 8.42$, $p = .001$, $\eta_p^2 = .16$.

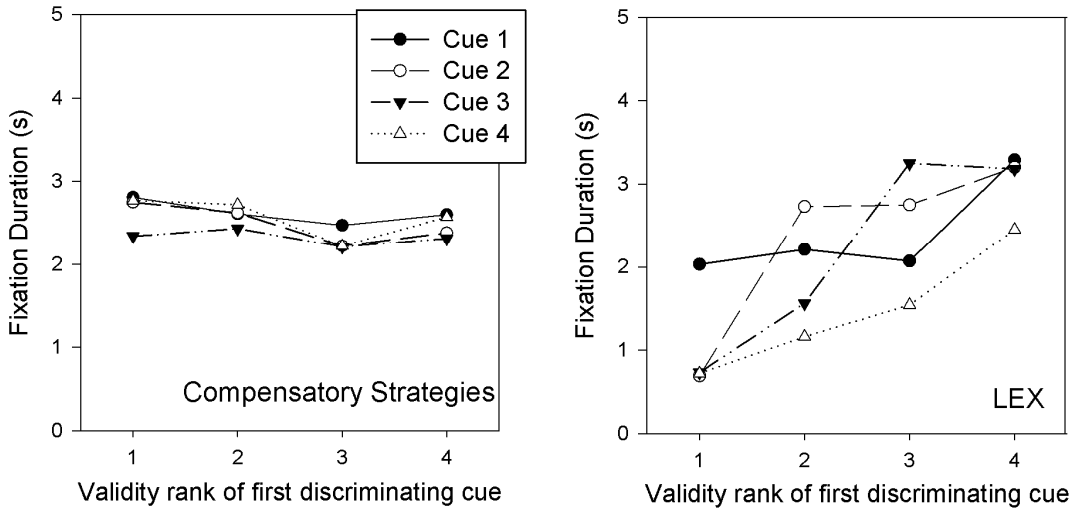


Figure 2: Mean fixation durations at cues with different validities depending on the validity rank of the first discriminating cue in decision phase 2 (instructed strategies). Data are presented separately by strategy classification.

Fixation durations at former locations of cue values. To assess fixation durations at former locations of cue values, we defined AOIs around the empty rectangular frames that had contained the cue values in the learning phase. These AOIs exceeded the original frames by 30 pixels in each direction.

For each trial and each cue, we determined the summed duration of all fixations in the two AOIs pertaining to the respective cue (one AOI for each alternative). These summed durations were averaged per participant across all trials, in which the first discriminating cue had the same validity rank. For participants of all strategy classifications, fixation durations exhibited a gaze bias towards the cue locations in the upper half of the stimuli. This bias was independent of both the specific cues presented at these locations and the validity rank of these cues. Hence, as we were interested in the average fixation durations at locations of cues with different validity ranks, we calculated weighted means across the groups of participants for which the two most valid cues appeared in the upper part of the stimuli and the groups for which these cues appeared in the lower part.

In Figure 2, the weighted mean fixation durations at cues with different validities are plotted against the validity rank of the first discriminating cue for the second decision phase with instructed strategies. As can be easily seen, we found clearly different result patterns for LEX-users and users of compensatory strategies.

Under instructed strategy conditions, users of compensatory strategies looked approximately equally long at all four cues and their fixation durations were unaffected by the validity rank of the first discriminating cue in a decision item. Consequently, in a three-way mixed ANOVA the effects of Cue, Greenhouse-Geisser corrected $F(2.21, 50.88) = 0.22$, $p = .83$, $\eta_p^2 = .01$, First Discriminating Cue, $F(2.33, 53.67) = 2.06$, $p = .13$, $\eta_p^2 = .08$, and their interaction, $F(5.00, 114.96) = 0,45$, $p = .81$, $\eta_p^2 = .02$, were all not significant (the third factor Position of the Two Most Valid Cues with the levels Upper Half or Lower Half was introduced to control for the gaze bias towards the upper half of the stimuli).

In contrast, in items in which the most valid cue discriminated, instructed LEX-users fixated this cue considerably longer than all other cues. When the second most valid cue was the first to discriminate, the largest increase in fixation durations occurred for this cue. However, also the third cue and the fourth cue were fixated longer than in items in which the first cue discriminated. Similar changes in fixation durations emerged when cues with a lower validity were the first discriminating ones. Thus, the data reveal two trends: First, as expected, cues with a low validity are fixated longer, when no cue higher in validity discriminates. Second, there is a tendency towards all cues being fixated longer when the validity rank of the first discriminating cue is low. Correspondingly, the interaction effect of Cue x First Discriminating Cue, $F(4.18, 75.31) = 5.75$, $p < .001$, $\eta_p^2 = .24$, and the main effect of First Discriminating Cue, $F(1.77, 31.83) = 16.94$, $p < .001$, $\eta_p^2 = .49$ were significant. The main effect of Cue was not statistically reliable, $F(1.55, 27.94) = 2.57$, $p = .11$, $\eta_p^2 = .13$.

When the participants spontaneously adopted a decision strategy in the first decision phase, the pattern of results was similar but somewhat noisier. For users of compensatory strategies, there were again no statistically significant effects. For LEX-users, the effect of First Discriminating Cue was confirmed, $F(2.08, 49.85) = 6.51$, $p = .003$, $\eta_p^2 = .21$, whereas the interaction effect of Cue x First Discriminating Cue was no longer statistically significant, $F(4.83, 115.96) = 1.94$, $p = .09$, $\eta_p^2 = .08$.

Discussion

The observed fixation patterns on emptied displays of decision alternatives differed markedly and in line with predictions depending on the decision strategies employed. If according to a lexicographic strategy more comparisons between single cue values of alternatives were necessary, more transitions between alternatives were recorded. Furthermore, even the fixation durations on the former locations of specific cues reflected the cues' relative importance according to a lexicographic or a compensatory strategy.

Thus, tracking fixations on emptied information displays provided indicators of information search in memory-based multiattribute decisions similar to those provided by Mouselab methods for "inferences from givens" (Payne, Bettman, & Johnson, 1993). This proves a novel process-tracing method applicable to memory-based decisions. The present results corroborate the outcome-based strategy classification method (Bröder & Schiffer, 2003a) and add to previous response time data on strategies in memory-based multiattribute decisions (Bröder & Gaissmaier, 2007). Now, there is a way to analyze overt behavior that seems to indicate which cognitive processes determine response times and decision outcomes.

The looking-at-nothing phenomenon has been interpreted as an attempt at memory retrieval that triggers an involuntary gaze shift to the former location of the sought information (Richardson & Spivey, 2000). The former location is specified by a spatial index in an integrated representation encompassing conceptual, linguistic, visual and spatial information (Ferreira, Apel, & Henderson, 2008). In the present experiment, several instances of memory retrieval were required in a single binary choice trial. The more information had to be retrieved according to a strategy, the more fixations occurred, however, locations were frequently refixated. Based on the current data we cannot decide whether these refixations are due to repeated retrieval attempts or further processing of the already retrieved information in working memory. We presume that they indicate processing of retrieved information similar to eye movements that occur while visuo-spatial imagery is experienced during discourse processing (Johansson, Holsanova, & Holmqvist, 2006). If this proves correct, the exposition of the LEX strategy has to be modified. The prolonged response times predicted and observed for LEX-users if cues lower in validity have to be processed seem to be due not only to additional memory retrieval, but to extended pondering that includes cue information that does not affect the final decision.

In our attempt to exploit the looking-at-nothing phenomenon for process tracing, we have shown that it manifests itself rather robustly. Here, encoding and retrieval were separated by multiple encoding and retrieval instances with respect to overlapping physical locations. Furthermore, spatial indexing operated relative to visual context that varied in its physical location. Hence, we think that looking-at-nothing has wide applicability. Observing information search on emptied displays opens up a window on the cognitive processing of memorized information.

Acknowledgments

We would like to thank Sebastian Burchert for his help in conducting the experiment.

References

Bröder, A. & Gaissmaier, W. (2007). Sequential processing of cues in memory-based multiattribute decisions. *Psychonomic Bulletin & Review, 14*, 895-900.

Bröder, A. & Schiffer, S. (2003a). Bayesian strategy assessment in multi-attribute decision making. *Journal of Behavioral Decision Making, 16*, 193-213.

Bröder, A. & Schiffer, S. (2003b). Take the best versus simultaneous feature matching: Probabilistic inferences from memory and effects of representation format. *Journal of Experimental Psychology: General, 132*, 277-293.

Ferreira, F., Apel, J., & Henderson, J. M. (2008). Taking a new look at looking at nothing. *Trends in Cognitive Science, 12,* 405-410.

Gigerenzer, G. & Goldstein, D. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review, 103*, 650-669.

Gigerenzer, G., Todd, P. M., & the ABC Research Group. (1999). *Simple heuristics that make us smart*. New York: Oxford University Press.

Glöckner, A., & Betsch, T. (2008). Multiple-reason decision making based on automatic processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*(5), 1055-1075.

Jahn, G., Renkewitz, F., & Kunze, S. (2007). Heuristics in multi-attribute decision making: Effects of representation format. In D. S. McNamara & G. Trafton (Eds.), *Proceedings of the Twenty-Ninth Annual Conference of the Cognitive Science Society* (pp. 383-388). Mahwah, NJ: Erlbaum.

Johansson, R., Holsanova, J., & Holmqvist, K. (2006). Pictures and spoken descriptions elicit similar eye movements during mental imagery, both in light and in complete darkness. *Cognitive Science, 30*, 1053-1079.

Martignon, L. & Hoffrage, U. (2002). Fast, frugal, and fit: Simple heuristics for paired comparison. *Theory and Decision, 52,* 29-71.

Payne, J. W. Bettman, J. R., & Johnson, E. J. (1993). *The adaptive decision maker*. Cambridge: Cambridge University Press.

Renkewitz, F. & Jahn, G. (2010). Tracking memory search for cue information. In A. Glöckner & C. Witteman (Eds.), *Foundations for tracing intuition: Challenges and Methods* (pp. 199-218). Hove, UK: Psychology Press.

Richardson, D. C. & Spivey, M. J. (2000). Representation, space and hollywood squares: Looking at things that aren't there anymore. *Cognition, 76,* 269–295.

Motor Simulation in a Memory Task: Evidence from Rock Climbing

Giovanni Pezzulo (giovanni.pezzulo@cnr.it)
Istituto di Linguistica Computazionale 'Antonio Zampolli', CNR - Pisa, Italy
Istituto di Scienze e Tecnologie della Cognizione, CNR – Roma, Italy

Laura Barca (laura.barca@istc.cnr.it)
Neuroscience and Neurorehabilitation Department, Children's Hospital Bambino Gesù - IRCCS
Santa Marinella (Rome), Italy

Alessandro Lamberti Bocconi
Facoltà di Scienze Motorie, Università degli Studi dell'Aquila
L'Aquila, Italy

Anna M. Borghi (anna.borghi@gmail.com)
University of Bologna, Bologna, Italy
Istituto di Scienze e Tecnologie della Cognizione, CNR – Roma, Italy

Abstract

This study concerns the role of motor simulations in a memory task performed by expert and novice climbers. In a behavioural task, expert and novice rock climbers were shown three novel climbing routes: an easy route, a route impossible to climb but perceptually salient, and a difficult route. After a distraction task, they were given a recall test in which they had to write down the sequence of holds composing each route. No difference emerged between experts and novices on the easy and impossible routes. Differently, the performance of expert climbers was better than that of novices on the difficult route. Results suggest that seeing a climbing wall activates a motor, embodied simulation, which relies not on perceptual salience, but on motor competence. Crucially, it is shown that the ability to form this simulation is modulated by individuals' motor repertoire and expertise, and that this strongly impacts recall.

Keywords: simulation, affordance, embodied cognition, grounded cognition, canonical neurons, mirror neurons, motor memory, memory for actions, motor chunks.

Introduction

A number of studies have shown that seeing an object, such as a cup, affords simple actions, such as reaching and grasping. According to the original definition by Gibson (1979) *affordances* are possibilities for action offered by the environment and perceived directly by an observer. A recent view of affordances, which we endorse here, is that they are potential action patterns activated in the observer's brain while observing objects. In other words, they are the product of the conjoining, in the brain, of visual stimuli and action responses (e.g., Ellis & Tucker, 2000), whose neural bases can be found in the discovery, in the F5 area of the ventral premotor cortex of the monkey, of visuomotor canonical neurons which discharge in the presence of graspable objects when no overt response is required (Murata et al., 1997). Evidence in humans confirms the existence of a parietopremotor circuit active during the observation of manipulable objects (Grèzes et al., 2003). Overall, both behavioural and brain imaging studies have shown that perceiving affordances activates in observers specific motor programs (Borghi, 2004; Borghi and Riggio, 2009; Martin, 2007). This phenomenon can be interpreted as activation of a *motor simulation*, where 'simulating' means that the same sensorimotor systems that are activated during interaction with objects are activated during object perception (e.g. when observing objects or when listening their characteristic sound), but without the execution of overt movements (Gallese, 2009; Jeannerod, 2006).

A computational framework proposed by Wolpert and Kawato (1998) and elaborated in Frith et al. (2000); Jeannerod (2006); Wolpert et al. (2003) explains motor simulations as the re-enactment of internal models that allow motor control. Internal models come in two varieties, inverse and forward. During motor control, the former compute the necessary motor commands to achieve a certain goal given a starting position, and the latter predict the sensory consequences of those motor commands. In addition, it is possible to re-enact internal models to form a simulation of possible actions by feeding the inverse model with predicted sensory inputs rather than 'true' sensory inputs, and successively feeding the new motor command to the forward models, and so on. This process permits the linking of multiple predictions in order to obtain simulations of possible actions for an arbitrary long number of steps. Note that for this process to work it is also necessary to inhibit 'true' sensory inputs and motor outputs. Indeed, simulating is not the same as performing an overt action, for a variety of reasons: simulation implies a weaker activation of the interested neural areas. In addition, during simulation some kind of blocking mechanisms might intervene that prevents the action to be executed overtly. Finally, during overt action a sensorial feedback is received, while no such feedback is given while simulating (Jeannerod, 2006).

Even if the activation of motor information elicited by object presentation has been extensively studied in the last years, the majority of studies have focused on how single objects or object pairs (e.g., Riddoch et al., 2003) activate an

internal simulation or even overt simple movements, such as reaching or grasping. The role played by multiple affordances for complex actions implying a sequence of movements has not been widely investigated. Imagine observing a mountain path before performing a complex action composed by a sequence of movements, such as hiking. One might observe whether the path is steep or not, how the different stones are displayed, whether tree branches represent obstacles for walking and how to avoid them. In other words, both the characteristics of single objects (e.g., the stones, their orientation and shape) and their placement along the path might afford or impede actions. The same is true for climbing.

Indoor rock climbing consists in reaching the top of a specially-designed wall (i.e., a climbing wall), by grasping climbing holds with the hands and the foots. Climbing routes, which consist in carefully arranged sequences of climbing holds, may have different difficulties depending on the slope of the wall, the length of the route, as well as on the number, kind, and arrangement of the climbing holds. Usually climbers, both during their training and during competitions, spend some time in "studying" climbing routes before climbing them, especially when they have to climb a route for the first time. Then, they can mentally simulate which holds to take, which movements to do, which rest positions they can find, etc. In some cases, they also overtly mimic the hand (and foot) movements that they expect to perform while climbing (see fig. 1).

Figure 1. Athletes studying a climbing route before climbing it. Note the overt hand movements.



The simulation they build might include both information on specific affordances, i.e. the characteristics of the holds (shape, orientation, etc.), and information on their displacement, i.e. the way they are arranged on the wall. Given that routes involve multiple climbing holds, clearly any simulation of a part of the route changes the way the rest of the route is perceived. For example, simulating grasping a certain hold with the right hand makes some other holds affordable to be grasped with the left hand, and some other holds out of reach; see fig. 2. At the same time, the need of reaching a certain 'goal' hold determine which holds are affordances retrospectively, and disrupts the affordances of some holds (e.g., far holds) in the climbing wall. For all these reasons, motor simulation in rock climbing should be considered an *affordance calculus* rather than a response to a sequence of individual affordances. Crucially, the motor competence of climbers also determines what constitutes an affordance. Experienced climbers can hold small holds that are difficult for weak climbers to grasp, and can simulate sequences of actions that are too complex to be picked up by novice climbers, much like how expert chess players 'see' complex strategies. We hypothesize that the proficiency of expert climbers allow them to climb better the routes also by understand them better, where understanding should be intended as proficiency in the affordance calculus and in the associated building of appropriate mental simulations before climbing.

Figure 2. A sample sequence of movements in rock climbing. Notice that (i) climbing holds afford different grips, and (ii) the way holds can be grasped depend on which holds were grasped before (and how) as well as which holds the climber intends to reach.



Aims and objectives of the study

Our study addresses the role multiple affordances play in the recall of routes by rock climbers with different level of expertise. An open issue in this field pertains to the extent to which affordances are elicited automatically, upon seeing objects, or are activated when a specific action goal is pursued. In addition, studying recall in expert and novice climbers can contribute by showing to what extent the activation of affordances is modulated by observers' experience and competence. Finally, we still know very little on how affordances improve recall. Acquired motor skills offer a unique way to test this question.

Here, novice and expert climbers were asked to observe and recall the position of holds of 3 routes that they never climbed: an easy route (ER), a difficult route (DR), and a (motorically) impossible but perceptually salient route (IPSR). Predictions were that their performances would not differ for the ER, because both groups would be able to perform a motor simulation, and for the IPSR route, when for both it was impossible to form a motor simulation of climbing. If this were true, this would demonstrate that the simulation formed is a motor one, and would be activated only when participants have the motor competence necessary to perform the sequence of actions. Accordingly, the performance of experts should overcome that of nonexperts in the DR, when the actions required climbing the route they are shown are part of their motor repertoire.

Method and Materials

Participants

Eighteen climbers who attended to the "Lanciani Climb" arena in Rome volunteered to study. Experts had between 5 and 10 years climbing experience, whereas novices had less than six months climbing experience. Groups were balanced for gender (6 men and 3 women each group) and age. To balance the order in which the different routes were presented, as well as to avoid assigning the task to large groups, we divided the participants in 6 groups of 3 randomly selected participants: 3 groups composed by experts, and 3 by novices.

Materials

Two climbing trainers set up three novel routes from a climbing wall containing 110 holds. Each route was composed of 10 holds (the typical average length for most training routes). Route difficulty depends on the configuration of the holds (their graspability) and the configuration of the limbs in transition between the holds (Smyth and Waller, 1999). Both experts and novices, because of the orientation and arrangement of the holds, could climb the Easy Route (ER) without difficulty. In order to control for perceptual factors that might facilitate memorization, the two other routes differed in perceptual salience. The Difficult Route (DR) was difficult to climb because the holds were not easily graspable due to their shape and orientation, and only expert climbers could benefit from their affordances. All holds in the ER and DR were grey- or dark-coloured and did not differ in size or other perceptual characteristics. The third route, (motorically) Impossible but Perceptually Salient Route (IPSR), was impossible to climb as a whole (but parts of it could be climbed). The difficulty of such route was not due to the fact that participants had to simulate biologically impossible movements (Costantini et al., 2005) but rather on the arrangement of the holds. Specifically, it was impossible to benefit from the affordances offered by the holds and to configure the limbs for a transition from one hold to the other. To facilitate memorization, however, we rendered the holds perceptually salient: they were vividly coloured, compared to the standard grey- or dark-coloured holds.

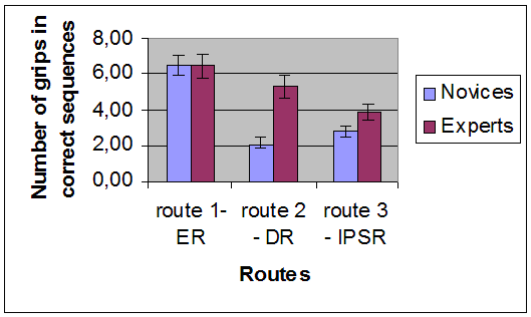
Procedure

Two experimenters and the trainer were present in the Lanciani Climb arena to administer the task. Before entering the arena, participants were instructed that they have to memorize a route made up of 10 holds, and that later they had to perform an additional task. Groups (of 3 participants) were then invited to enter and to sit in front of the climbing wall. The wall includes 110 holds with different size and orientation, placed uniformly to cover its entire surface. The trainer indicated twice the holds of each route with a stick. After this demonstration, participants had to turn their backs to the wall and perform a distracting task (i.e. to pronounce the letters from A to L). The procedure was repeated for each route. The presentation order of the routes (ER, DR and IPSR) was balanced across participants. Participants were given a folder containing three A3 sheets, each displaying a picture of the climbing wall (which included all the holds). After the first of the three routes had been shown, they were asked to extract the first sheet and to mark down as quickly as possible (with a time limit of 2 minutes) the sequence of holds composing the first route. The same procedure was repeated for the two remaining routes. Participants were then required to fill in a post-experiment questionnaire in which they were asked to report (by responding yes or no) whether they mentally imaged climbing the wall while being shown the route and while recalling them, whether they believed that imagining the route might be helpful for them, and which route appeared to them the easiest to climb.

Results

All participants performed the task without difficulties. The number of holds reported in a correct sequence for each route was computed for each participant, and submitted to a 3x2 mixed ANOVA with Route (ER, DR and IPSR) as within factor, Expertise (Expert vs. Novice) as between factor and participants as the random factor. Data are plotted in fig. 3. All analyses were conducted using a Type I error rate of .05.

Figure 3. Results of the task. Legenda. ER - Easy Route; DR - Difficult Route; IPSR – (Motorically) Impossible but Perceptually Salient Route



Expertise factor was not significant ($F(1, 16) = 1.35$; $MSe = 20.92$; $p = .26$), whereas Route factor was highly significant ($F(1,32) = 15.45$; $MSe = 3.35$; $p < .0001$). Post-hoc Newman-Keuls showed this was due to the difference between the ER ($M = 6.44$) and the two other routes, DR and IPSR ($M = 3.72$; $M = 3.33$, respectively). As predicted, the ER led to a better performance compared to the two other routes, independently from the degree of expertise of participants. It is worth noting that the average number of remembered sequences was exactly the same for experts and novices ($M = 6.44$).

Crucially to our hypotheses, the interaction between Expertise and Route was significant ($F(1,32) = 3.60$; $MSe = 3.35$; $p < .04$). Post-hoc test confirmed that there was no difference between Novices and Experts on the Easy Route ($p = 1$). More importantly, the difference between Novices and Experts was not significant with the IPSR (Newman-Keuls, $p = .21$, respectively $M = 2.78$, $M = 3.89$), whereas the performance of Novices was significantly worse than

that of Experts with the DR (Newman-Keuls, $p < .004$, respectively $M = 2.11$, $M = 5.33$). This suggests that the two groups did not differ in memory capabilities when for both of them it was impossible to mentally simulate the motor task, i.e. in the IPSR. This indicates that the impossibility to form a motor simulation clearly affects recall. The impact of motor simulation on recall is confirmed by results with the DR, where the difference between the two groups clearly emerged. Namely, in the DR, the capability to climb the wall was part of the experts' motor repertoire, thus they were able to build a motor simulation. In the post-experimental questionnaire, Experts and Novices did not differ in responding to whether they mentally imagined climbing the route while being shown it (55% of both groups responded using imagination) and while recalling it (44% for both groups responded positively). However, compared to novices, experts seem more aware of the effects of the simulation (22% of novices and 44% of experts reported that imagination helped), even though neither group seemed to believe that imaging was strategically important, as participants did not believe it helped them during recall (only 33% of athletes responded positively for both groups). Experts and Novices differed also in that Novices were less aware of the differences between the routes (55% of novices did not distinguish between them).

Discussion

Our results support the hypothesis that visually perceiving multiple affordances (here, climbing holds disposed in a climbing wall) leads to the activation of a motor simulation, which improved recall. The activation of the simulation is specific, and depends on whether or not the holds are disposed so to afford climbing, and on climbers' motor competence.

We found that both experts and non-experts performed equally well with the Easy Route. This suggests that, when participants have the motor competence allowing them to climb a given route, they simulate doing it, and this very fact improves their recall of the route. In addition, our results allow us to understand what happens with difficult routes, that is, when, for some of the participants, it is difficult or impossible to construe a simulation. Specifically, the design we used allow us to distinguish situations in which participants could rely on perceptual salience for memorization and situations in which only a subset of participants might build a motor simulation grounded on previous climbing experience. We found that the expert participants, who were able to rely on a mental simulation strategy, had better performance than novice ones, who were only able to rely on visual strategies. The advantages of motoric vs. visual strategies were also highlighted by the poor performance of both groups in the (motorically) impossible but perceptually salient route, despite the high salience of the holds that composed the route. Our results indicate that a simulation is evoked only when the holds have perceptual characteristics and also afford actions. Namely, no simulation is activated when climbers observe holds that are perceptually salient (i.e. having vivid colors) but not useful for climbing the route, that is, when the holds do not represent good affordances. This result helps to qualify the kind of simulation evoked: holds (affordances) elicit an embodied, motor simulation, not a purely visual simulation.

Notice that in this study we do not consider the specificity of the climbing method experts and non-experts adopt; we simply focus on different climbing competence. A few studies have addressed and demonstrated that experts and novices might use different patterns of action. Boschker et al. (2002) found that, differently from inexperienced climbers, experts focused on the functional aspects of a climbing wall, whereas they did not consider its structural features. In Boschker and Bakker (2002) inexperienced climbers who were shown a video of expert climbers learned to use experts modes of climbing (e.g., arm crossing) and climbed faster and with more fluent movements than those who were shown videos of novice climbers or a control video. Overall, our results fit well in the embodied cognition (Glenberg, 1997) literature and have implications, concerning the role of affordances for both simulation and recall, as well as the relationship between motor competence and the capability to form and use motor simulations.

In addition, this finding helps us comprehend the mechanisms on which memory of action relies (see for example Daprati et al., 2005). Overall, our study suggests that the ability to benefit from objects' (holds') characteristics and from their arrangement can help a climber form *motor chunks*, i.e. chunks based on sequences of real action possibilities, which, in turn, leads to better recall of a given route. The idea of "chunks" derives from the study of Chase and Simon (1973) on how competence influences recall of chess positions in novice and expert chess players. The main finding of such study is that expert chess players outperformed novices in the recall of meaningful chess positions, but not in non-meaningful positions. The authors proposed that this is due to the experts' larger set of 'chunks' of chess positions, which permits them to recognize complex patterns of chess positions as individual units and therefore to recall them better. Our study shares resemblances with the study of Chase and Simon (1973), the two main differences being that: (i) we focus on motor competence rather than abstract problems like chess, and (ii) unlike chess players, climbers see the climbing routes for the first time, and there is an immense variety of combinations of holds, orientations, inclinations of the climbing walls, etc. Although the climbers could still pick up abstract similarities between old and new patterns of holds, these similarities are meaningless if untied to body possibilities and more in general (competence-specific) motoric information. For this reason, we could hypothesize that a chunking mechanism could be in play that is similar to the one described in (Chase and Simon, 1973); it can be called motor chunking due to the

importance of motoric information. However, if this is the case, motor chunks cannot be simply retrieved from memory, but should be built anew (or at least reassembled) as part of the planning (and simulation) process, which is of course highly competence-specific, and involves the (partial) re-enactment of motor processes. Note that this view of motor chunking is compatible with the idea of Glenberg (1997) that simulations can be *meshed* with (episodic) memories. Overall, this view could explain why memory performance is better when climbers are allowed to form motor chunks, not when they use memory strategies relying on the visual saliency of some holds. This finding is also compatible with the idea that motor simulations elicit procedural memories (see Pezzulo, 2008; in press; Pezzulo and Castelfranchi, 2009, for a discussion).

Our results suggest also that the activation of a motor simulation is possible only when performing a given sequence of actions is part of participants' motor competence. The better recall of Experts compared to Novices is totally due to the fact that, given that they were able to climb the difficult route, they could mentally simulate climbing (do the 'affordances calculus') and, with the help of the affordances, they were able to recall the sequence of required movements. Novices were impeded from simulating because they did not possess the motor capability to climb the Difficult Route. This suggests that the ability to simulate is modulated by previous motor experiences, in keeping with ideomotor theories of perception and action (Hommel et al., 2001).

Differently from other sports, like dance, in rock climbing both the simulation elicited by action observation (of another rock climber) and the simulation elicited by affordances (simply observing a rock or climbing wall) can be studied. Therefore, our research extends also the results showing that a motor resonance phenomenon occurs when we observe others performing complex movements, such as dancing and playing basketball (e.g., Cross et al., 2006). This phenomenon has its neural basis in the mirror neuron system, which, differently from canonical neurons, are activated both during performance of an action (say, grasping, manipulating and holding objects), and during observation of others performing the same action (Gallese et al., 1996). In line with our results, this motor resonance is stronger when participants observe actors sharing their motor repertoire. Aglioti et al. (2008) demonstrated with a psychophysical study that elite basketball players predicted the success of free shots at a basket earlier and better than expert observers and novice players. The experts' advantage was due mainly to their higher capability to predict by reading body kinematics in the early movement phases. A transcranial magnetic stimulation (TMS) study showed a time-specific motor activation while observing videos of errors. The results of the combined physiological and TMS studie reveal that fine-grained motor resonance occurs after motor practice and that motor expertise specifically contributes to anticipating the actions of others.

Studying a special case, that of rock climbers, our behavioural study showed for the first time that multiple affordances activate a motor simulation, and that this strongly impacts recall, which is then modulated by participants' motor expertise and motor repertoire. Further studies are needed to better understand the neural underpinnings of the complex mechanisms of recall based on affordances and embodied simulation.

One alternative explanation for our results is that experts might be better in fitting visual images of climbers' postures, and thus they could use visual imagery rather than motor simulations. Although our study cannot rule out this possibility, there are reasons to believe that this is not the case. First, while this hypothesis explains the advantage of experts in the DR, it does not explain the good performance of novices in the ER. To explain why novices are better in recalling the ER than the DR, one should say that visual imagery is specifically modulated by one's own (motoric) climbing competence. Second, the exclusive use of visual imagery could hardly help solving our task. Namely, climbers experience the routes for the first time, and cannot see other climbers, so any visual simulation they build has to be done anew. However, spatial and configurational information (position of limbs in space) is not enough to determine which are the climbing positions one should remember, since valid climbing positions also depend on which affordances are offered by the holds, and which are the past and future movements. In other terms, although climbers could use visual imagery as part of their strategies, at least some of the processing required to recall climbing positions is better understood in motoric than purely visual terms. Another possibility is that experts are more experienced with some patterns of holds, much like chess players are supposed to be. As already discussed, however, climbers see the routes for the first time, and there countless dispositions of holds. More importantly, the visual appearance and the spatial configuration of the holds is not sufficient to understand the best path in a route, or its difficulty. To do so, climbers have to take into account at the same time the individual affordances offered by the holds, the previous movements, etc. Overall, then, due to the highly specific and situated nature of climbing, it is unlikely that a memory retrieval strategy could be sufficient (although it might help), and how memory retrieval could be done in purely abstract terms, without accessing one's own motoric information. (This is why we suggested that motor chunks should be built anew as part of the motor planning.)

Before concluding, it is worth mentioning that several studies distinguish between two kinds of motor simulations: conscious and unconscious (see Jeannerod, 2006 for a discussion). Most of the afore-mentioned studies address unconscious motor simulations; in this context, the idea is that seeing a climbing wall automatically activates specific motor processes in climbers. There is, however, another kind of motor simulation, a conscious one, which can be performed by climbers, and is indeed routinely done as part of the athletes' training, and before the start of competitions.

Jeannerod (2006) suggests that the representational content of conscious and unconscious simulations are the same, with different time constraints determining their level of access (e.g., most unconscious motor images arise for the demands of immediate action and simply do not have the time to become conscious). In this study, the climbers were not explicitly instructed to mentally simulate. However, the procedure adopted in this study, and in the afore-mentioned ones, does not permit us to discriminate whether or not participants used a conscious strategy. Further studies are necessary to shed light on the differences between conscious and unconscious mental simulations, and their respective roles in motor planning.

Acknowledgments

The research has been partially funded by the European's Community, FP7 under grant agreements no. 216125 (ROSSI, Emergence of communication in RObots through Sensorimotor and Social Interaction), no. PERG02-GA-2007-224919 (WoRHD, Written language processing in Hearing and Deaf), and no. FP7- 231453 (HUMANOBS, Humanoids That Learn Socio-Communicative Skills Through Observation). The authors thank Alessia Tessari and Marco Tullio Liuzza for useful discussions.

References

Aglioti, S. M., Cesari, P., Romani, M., and Urgesi, C. (2008). Action anticipation and motor resonance in elite basketball players. *Nature Neuroscience, 11*, 1109-1116.

Borghi, A. M. (2004). Object concepts and action: extracting affordances from objects parts. *Acta Psychologica, 115*, 69-96.

Borghi, A.M., Riggio, L. (2009). Sentence comprehension and simulation of objects temporary, canonical and stable affordances. Brain Research, 1253, 117-128.

Boschker, M. S. J. and Bakker, F. C. (2002). Inexperienced sport climbers might perceive and utilize new opportunities for action by merely observing a model. *Perceptual Motor Skills, 95*, 3-9.

Boschker, M. S. J., Bakker, F. C., and Michaels, C. F. (2002). Memory for the functional characteristics of climbing walls: perceiving affordances. *Journal of Motor Behavior, 34*, 25-36.

Costantini, M., Galati, G., Ferretti, A., Caulo, M., Tartaro, A., Romani, G. L., and Aglioti, S. M. (2005). Neural systems underlying observation of humanly impossible movements: an fmri study. *Cerebral Cortex, 15*, 1761-1767.

Cross, E. S., de C Hamilton, A. F., and Grafton, S. T. (2006). Building a motor simulation de novo: observation of dance by dancers. *Neuroimage, 31,* 1257-1267.

Daprati, E., Nico, D., Saimpont, A., Franck, N., and Sirigu, A. (2005). Memory and action: an experimental study on normal subjects and schizophrenic patients. *Neuropsychologia, 43,* 281-293.

Ellis, R., & Tucker, M. (2000). Micro-affordance: the potentiation of components of action by seen objects. *British Journal of Psychology*, *9*, 451-471.

Frith, C. D., Blakemore, S. J., and Wolpert, D. M. (2000). Abnormalities in the awareness and control of action. Philos Trans R Soc Lond B Biol Sci, 355(1404):1771–1788.

Gallese, V. (2009). Motor abstraction: a neuroscientific account of how action goals and intentions are mapped and understood. *Psychological Research, 73*, 486-498.

Gibson, J. (1979) The ecological approach to visual perception. *Lawrence Erlbaum Associates, Inc*

Glenberg, A. (1997). What memory is for. *Behavioral and Brain Sciences, 20*, 1-55.

Grèzes, J., Tucker, M., Armony, J., Ellis, R., and Passingham, R. E. (2003). Objects automatically potentiate action: an fmri study of implicit processing. *European Journal of Neuroscience, 17*, 2735-2740.

Hommel, B.; Musseler, J.; Aschersleben, G. & Prinz, W. (2001) The Theory of Event Coding (TEC): a framework for perception and action planning *Behavioral and Brain Science,* 24(5), 849-78

Jeannerod, M. (2006). *Motor Cognition*. Oxford University Press.

Martin, A. (2007). The representation of object concepts in the brain. *Annual Review Psychology, 58*, 25-45.

Murata, A., Fadiga, L., Fogassi, L., Gallese, V., Raos, V., and Rizzolatti, G. (1997). Object representation in the ventral premotor cortex (area f5) of the monkey. *Journal of Neurophysiology, 78*, 2226-2230.

Pezzulo, G. (in press). Grounding procedural and declarative knowledge in sensorimotor anticipation. *Mind and Language*.

Pezzulo, G. (2008). Coordinating with the Future: the Anticipatory Nature of Representation. *Minds and Machines, 18,* 179-225.

Pezzulo, G., and Castelfranchi, C. (2009). Thinking as the Control of Imagination: a Conceptual Framework for Goal-Directed Systems. *Psychological Research, 73,* 559-577.

Riddoch, M. J., Humphreys, G. W., Edwards, S., Baker, T., and Willson, K. (2003). Seeing the action: neuropsychological evidence for action-based effects on object selection. *Nat Neurosci*, 6(1):82–89.

Smyth, M. M. and Waller, A. (1999). Movement imagery in rock climbing: patterns of interference from visual, spatial and kinaesthetic secondary tasks. *Applied Cognitive Psychology, 12*, 145-157.

Wolpert, D. M. and Kawato, M. (1998). Multiple paired forward and inverse models for motor control. *Neural Networks*, 11(7-8):1317–1329.

Wolpert, D. M., Doya, K., and Kawato, M. (2003). A unifying computational framework for motor control and social interaction. *Philos Trans R Soc Lond B Biol Sci*, 358(1431):593– 602.

Enactive Social Cognition

Tobias Schlicht (tobias.schlicht@cin.uni-tuebingen.de)
Centre for Integrative Neuroscience, Universität Tübingen
Paul-Ehrlich-Str. 15
72076 Tübingen, Germany

Abstract

In this talk, I extend the enactive approach to cognition to the social domain within a larger framework of varieties of intentionality and argue for a second-person approach to understanding others, emphasizing a difference in our understanding of others depending on whether we are directly engaged with them in interaction or merely observing them. The enactive account is especially persuasive in developmental respects, suggesting that sophisticated forms of cognitive intentionality (e.g. believing) are grounded in motor intentionality (e.g. perception and action): Our own sensorimotor skills are partly constitutive of cognition, and other people's expressions of their sensorimotor skills in turn modulate our cognition of objects and our social understanding. The enactive account explains how young infants acquire the capacities that allow them to move from dyadic to triadic intentional relations at around their first birthday, and it claims that our basic form of social understanding is neither based on theoretical inference nor a kind of simulation, but constituted by an embodied implicit know-how displayed in online interaction.

Keywords: social cognition; embodied cognitive science; enactive approach; intentionality; social neuroscience.

1. Varieties of Intentionality

Intentionality is a technical term referring to the capacity to be directed towards an object, where 'object' is broadly construed: artefacts, events, people, states of affairs in the world, abstract or fictitious entities and mental states of others and of oneself can be the goal of an intentional activity (Brentano 1874). Moreover, there are a multitude of intentional attitudes via which one may be so directed: sensorimotor, affective and cognitive ways of dealing with the world (Barresi & Moore 1996, Crane 2001). One can think about, hope or doubt that it might rain tomorrow and one can perceive or desire a glass of wine or just intentionally grasp it with their hands. At the same time, one can *be* intentionally directed at something without having a sophisticated *understanding* of such intentional directedness in oneself and in others.

Searle (1983) and Crane (2001) have provided thorough analyses of intentionality, but it is odd that (1) they largely ignore what we may call motor intentionality—a directedness towards an object manifested in grasping and manipulating it, and that (2) they do not even attempt to characterize the subject being intentionally directed at objects, although this is an essential element in the structure of intentionality. My first claim is that a proper integration of these two aspects motivates a thoroughly embodied and enactive account to cognition and intentionality:

The *subject* of intentional relations is best characterized as an embodied agent, possessing a number of skills and capacities, ranging from performing bodily actions, perceiving and grasping objects, to thinking and imagining complex and even counterfactual states of affairs. Conceptualizing the cognizing subject in this way provides a first motivation for understanding intentionality in a broader sense by integrating sensorimotor forms of directedness. *Motor intentionality* has received much attention in recent investigations of perception and social interaction and plays an important foundational role in the enactive approach, as will be elaborated below. This constitutes a radical shift of emphasis: Whereas analytic philosophers of mind have been concerned primarily with intentionality as a feature of mental states, it is better construed as a feature of whole organisms (Thompson 2007, Hutto 2008). A second motivation for considering sensorimotor forms of intentionality is based on recent developments in the cognitive neurosciences and a more general transformation of paradigmatic cognitive science. Regarding the latter, the traditional computational paradigm is currently being replaced by an embodied-embedded cognitive science that does not consider the computational brain in isolation but investigates mental phenomena in the broader context of an embodied agent being situated in her environment, which itself constraints the agent's cognitive projects. Regarding the former, data from animal studies suggest that perceptual capacities are grounded in motor capacities quite generally. More specifically, the discovery of mirror neurons has fostered an ongoing debate about the role of motor intentionality in the overall cognitive architecture of human agents (Sinigaglia 2008). Consequently, this leads to the following general structure of intentionality:

An embodied agent (or organism) is directed towards an object or content by way of one among several sensorimotor, affective, or cognitive attitudes.

Since this structure can be realized in various ways, the task of a comprehensive theory of intentionality is to provide an adequate account for the different varieties of intentionality, differing in complexity and sophistication (Barresi & Moore 1996, Schlicht 2008). In broad strokes, such a framework may look something like this:

1. The most basic and biologically primary forms of intentionality are *perception and action* (Searle 1983). They are essentially dyadic relations to single *existing* objects or agents and depend on being situated and embedded in an environmental context. We share these forms of intentionality with many animals. Ontogenetically, human infants first make use of sensorimotor skills in order to perceive,

grasp and manipulate objects in their vicinity. Only later are they capable of more sophisticated and detached forms like beliefs.

2. On a second level, in scenes of *joint attention*, intentional relations become triadic: either an additional subject enters the subject-object-relation or an object enters the relation between two subjects (mother and child, say). True *joint* attention is not coincidental, i.e. both agents have to be mutually aware of their coordination of different perspectives on the world and actively track and manipulate the attention of each other. Merely looking at the same object coincidentally does not suffice. Joint attention involves a *cognitive* aspect—being directed on the world—and a *social* aspect (being engaged in social interaction with another subject). Numerous studies show that infants are capable of such "triangulation" (Davidson 2001) only from around nine to twelve months of age (Carpenter et al. 1998). Even these complex intentional relations strongly depend on the existence and presence of the object and person one is directed at. All relata being agents constitutes a special case of triadic relations, e.g. two adults attending to the infants' actions. This not only seems to occur a few months earlier but also does not involve a proper external *object*.

3. A third level is marked by the partial use of the imagination in the second year of life, for example in pretend play, where functional properties are detached from one object, a telephone say, and assigned to another, a banana say (Leslie 1987). It has also been shown that pretend play mark the onset of truly *collective intentionality*, where two agents pursue a common goal and coordinate their actions accordingly. This also involves an understanding of norms (e.g. rules of a game), and two-year olds have been shown to reinforce these norms (Rakoczy 2006). But even then infants still partly depend on the existence and presence of objects in their immediate environment, although they are already capable of representing an object in its absence.

4. Finally, a fourth level is characterized by an explicit directedness towards mental states like beliefs, desires and intentions of others. At around 4 years, infants display an understanding of other people and have acquired the concept of belief, which is reflected in their passing false-belief-tasks (Wimmer & Perner 1983). That is, they can now explain other people's actions on the basis of what the other believes to be the case rather than in terms of what really is the case.

It may be necessary to modify this model by adding further levels or more fine-grained distinctions, so this first sketch of a theory of intentionality needs to be worked out in much greater detail (see Schlicht 2008). But the approach to intentionality recommended here can be contrasted on the one hand to traditional approaches pursued by many analytic philosophers of mind, who try to explain it from the top down by focusing on propositional attitudes presupposing language and concept possession (Dennett 1971, Fodor 1987, also Brandom 1994), and to reductionist approaches on the other hand, which attempt to reduce everything mental to a different level, e.g. neural processing. The present approach treats intentionality as a "moving target", taking on different forms and recommends to explain it neither from the top down nor from the bottom up, but following Gallagher (2005), developmentally from the beginning onward. By integrating insights from the neurosciences as well as adequate phenomenological descriptions and distinctions, this strategy promises to account for the 'developments' of the intentional attitude towards cognitive sophistication and of the target object from existing to fictitious and purely mental objects.

Because of its complexity, joint attention may be seen as a "primitive state of consciousness" (Campbell 2005, Eilan et al. 2005): One is at the same time directed at an object of interest and at another subject, with a cognitive and a social dimension. Thus, it is not only interesting with respect to our cognition of worldly objects but also with respect to our understanding of others. It has to be emphasized that joint attention as a complex form of intentionality does not come out of nowhere but has important precursors from the point of view of cognitive development, namely dyadic intentional relations. In the context of investigating the neural correlates of engaging in joint attention, Schilbach et al. (in press) developed an interactive paradigm in which participants' gaze behavior as measured by an eye-tracking device was used to contingently control the gaze behavior of a computer-animated character. Test persons interacted with the virtual other while undergoing fMRI. It was found that in contrast to merely following the other's gaze, actively establishing joint attention by directing the other's gaze was correlated with a differential increase of neural activity in the *ventral striatum*, known to be a part of *reward*-related neurocircuitry (Rolls, Grabenhorst, Parris, 2008). These findings may be interpreted as the neural correlates of an *intrinsic motivation* to engage in triadic intentional relations and of sharing experiences. But a natural question to ask is what allows infants to move forward to triadic forms of intentionality at the end of their first year of life, apart from this natural inclination to *share* something with someone. In the following, the aim is to outline central elements of an enactive approach to cognition and transfer them to the social domain, to show that it provides a plausible account of the capacities needed for this cognitive development. Along the way, additional empirical support from developmental psychology and the neurosciences will be integrated.

2. Enactive Cognition

As has been pointed out above, an embodied-embedded and enactive approach to cognition in general is recommended by a proper understanding of the subject being engaged in intentional relations as being an embodied agent. Such agents are in possession of a number of capacities that allow them to be intentionally directed towards objects of all kinds. It is claimed here that such an agent's cognitive intentional relations are grounded in her motor intentional activities. Motor intentionality is systematically, phylogenetically and ontogenetically prior to cognitive intentionality. What this claim amounts to can be illustrated by referring to essential sources that feed into the enactive account: Husserl and Heidegger. Both of them criticized Brentano for giving too much

prominence to the cognitive intentionality of beliefs and desires, and in general to the problem of how it is possible for our mental states to be about or directed to non-existent objects. In contrast, they claimed that "the manner in which things are given initially is not theoretically, disinterestedly, neutrally to our sight, as it were, rather things are given as items involved in our various tasks and practical engagements, our 'comportments'" (Moran 1996, p.58). That is, according to these phenomenologists, we are primarily directed towards existing objects in practical, embodied and sensorimotor ways.

Embodied sensorimotor skills. This central phenomenological idea has been revived in the so-called 'enactive' approach to cognition, according to which the whole embodied organism embedded in its environment is the fundamental subject of experience and intentionality (Thompson 2007, Hutto 2008). One central claim of enactive accounts is that cognition is not merely achieved by neural activity alone, but to some extent by bodily and environmental factors, which play not only a causal but constitutive role in cognition (Clark 1997; Noë 2004, 2009; Wheeler 2005; Thompson 2007). On this view, cognition is an activity, enabled by the exercise of skillful know-how in the agent's active exploration of and coupling with its environment.

The enactive approach to perception emphasizes the importance of sensorimotor skills exercised in the dyadic interaction with objects. This can be illustrated in an analysis of perception: Husserl already argued that a perceived object is never given in its full detail. Since we always perceive it from some point of view, we always only perceive some specific profile of it. We see the side facing us, while the other sides are hidden. Yet, although we do not directly see these other sides, phenomenologically speaking we have a distinctly perceptual sense of their presence in our actual experience of the side facing us (Hua 16, 176). For example, when you see a yellow lemon, a yellow round object is presented to you; and the correct description of your phenomenal content is not that of a flat two-dimensional screen, although you are only presented with exactly that from where you are standing. What you perceive is a round voluminous object. When you encounter this object for the first time and explore it, then you have to perform certain actions, e.g. eye- or head-movements, in order to make the hidden profiles visible. For example, in order to see the reverse side of the lemon, you must either go around it, or grasp it and turn it around. In this way, in comprehending an objects' complete profile you draw on your know-how, i.e. on a set of sensorimotor skills you are equipped with and which you can refine in your ongoing exploration of the world. Alva Noë (2009, 60) puts this central idea of the enactive approach to perception this way: "Seeing involves moving the eyes and head and body. ... Movements of your eyes or your head or your body actively produce changes in sensory stimulation to your eyes. Or, put differently, how things look, depends, in subtle and fine-grained ways, on what you do. Approach an object and it looms in your visual field. Now turn away: it leaves your field of view. Now shut your eyes: it is gone. Walk around the object and its profile changes. ... There are patterns of dependence between simple sensory stimulation on the one hand and your own bodily movement on the other. ... Seeing is a kind of skillful activity."

Affordances. In all these activities necessary for perception, your body plays a constitutive role. For one thing, spatial objects can only be experienced by embodied subjects, which are situated and embedded in their environment. Moreover, your body constitutes the point of view from which you perceive objects in the environment, and thus functions as an egocentric principle of experience; and finally, as the analysis above has revealed, every perception of an object is mediated and made possible by the body. Your body is first and foremost not experienced as one object among others but with respect to its potential for action as an experiencing organ. Thus, the kinesthetic experience of your body is correlated with your object experience and, moreover, it presents objects as providing you with various possibilities for action. Thus, new emergent properties arise from the sensorimotor coupling with the environment: affordances (Gibson 1979). These are opportunities for perception and action offered by objects in the environment. A surface, say, may be horizontal and rigid such as to allow you to walk on it. That makes it 'walk-on-able'; it may also be 'sit-on-able' and 'stand-on-able' etc. At the same time, the features of the surface may prevent other actions and they may provide organisms of a different kind with yet other affordances. That is, such possibilities for action are not fixed properties, but vary as a function of the successful coupling between this specific agent and its environmental niche. They may differ for other organisms. Quite often, we even perform certain actions and use tools to change environmental structures in order for them to afford various other actions. In this respect, the coupling between agent and environment displays a certain dynamics.

Online intelligence. All this is especially plausible developmentally, since an infant's primary encounter with objects in the world is characterized by what they can do with objects rather than what these objects are exactly. Experiments by Sommerville and Woodward (2005) suggest that active experience also modulates an infants' *understanding* of simple actions. One of their studies shows that active experience using tools may enable infants to build motor representations of tool use events that subsequently guide action perception and support action understanding. Children can more easily detect and understand actions they have performed themselves than actions they have only observed being performed by someone else. Their understanding of the intentional actions of others may be facilitated by sensorimotor action representations that have been produced during their own performance of the same or similar actions. In this sense, Husserl was right to claim that in our dealings with the world, the practical *I can* is more fundamental than the cognitive *I know* (or the *I think*). And it is in this sense that one should understand the claim that cognitive intentionality is grounded in motor intentionality.

Another way to put this point is by emphasizing the important function of perceptual experience of enabling successful navigation in the environment. Wheeler (2005, p.12f) calls this *online intelligence*: "A creature displays online intelligence just when it produces a suite of fluid and flexible real-time adaptive responses to incoming sensory stimuli". Online intelligence is to be contrasted with *offline intelligence*, exhibited when pondering on a mathematical problem or deliberating about whether to move to another city. The present framework argues for the primacy of online intelligence over offline intelligence.

To sum up, the enactive account to cognition emphasizes the foundational and constitutional role of embodied *sensorimotor skills* for cognitive acts like perception. The corresponding kind of knowledge that is brought to bear in these situations is not propositional knowledge-that but rather a skillful *know-how* to cope with the environment in online cognition. Such know-how is implicit rather than explicit and can seldom be spelled out by those who possess it (Ryle 1949). A paradigm example is knowing-how to ride a bicycle. *Affordances*, the final notion that has been emphasized, are properties that emerge from the successful *coupling* of agent and environment and change in accordance with their *dynamic* relationship. In the next section, these ideas are applied to the social domain.

3. Enactive Social Cognition

Although it is easy to see how these ideas translate to the social domain, there is as yet no comprehensive account of enactive social cognition, apart from some noteworthy yet sketchy attempts (De Jaegher & DiPaolo 2007; Thompson 2007; Hutto 2008). Consider first the *primacy of embodied and sensorimotor skills*: Due to the dominance of theory-theory and simulation-theory, social cognition has often been interpreted in a very sophisticated way, based on the passing of false-belief tasks at around the age of four or five years. Everything that goes on before that age has (unjustifiably) been considered as a mere precursor to the real thing (cf. the modules distinguished by Baron-Cohen 1995, Ch. 4). According to the enactive approach, not only object perception is essentially embodied in the way specified. Social cognition is also fundamentally embodied and embedded, since the most intimate and basic encounter between two subjects is that in direct social interaction where gestures and facial expressions play a dominant role. Many critics have recently suggested that when we are actively and directly engaged with another, we do not need to draw theoretical inferences or engage in mental simulation. Instead, we have more basic and simple means for getting a grip on other minds: Once we drop the questionable separation between an inner (meaningful mental) and an outer (meaningless behavioral) realm and reject the premise that mental states are abstract entities hidden in someone's mind, there is room for the alternative view that we can often *directly perceive* other people's mental states, e.g. feelings and intentions, since mental states are not abstract theoretical entities, but essentially embodied and revealed to others in expressive behaviors like gestures and facial and other bodily expressions (Gallagher 2001, Ratcliffe 2007). Not only do we ourselves convey our feelings to others through facial expressions, we also use their bodily expressions as cues to what they feel and intend to communicate. Video-replay studies demonstrate that young infants have a good sense for appropriate bodily and facial responses from the caregiver to her own communicative signals since they respond when they are out of synchrony.

Moreover, my own eye- and head-movements are not only crucial for my own perceptual states. They also play an important role as cues for another subject to find out where I am looking and/or to establish joint attention with me. Consequently, Corkum and Moore (1995) found that it was easier for infants to locate a target if this was activated on the same side as an adult model's head turn than when it was activated on the side opposite to the adult's head turn. They also investigated the origins of the gaze-following response necessary for joint attention and found that head orientation information is more important for infants below twelve months than eye orientation information. Only at eighteen months gaze following is reliably produced when eye movement is the only cue. Thus, it seems that such bodily cues are important to different degrees in the course of development.

Earliest forms of social understanding are proto-conversations and dyadic emotional engagements between infant and caregiver. They are clearly based on embodied practices, which the infant can engage in from the very beginning. In numerous studies, Meltzoff & Moore (1977) as well as Kugiumutzakis (1998) have established that neonates can imitate simple facial expressions. This has been interpreted as demonstrating an intimate connection between proprioception of one's own bodily actions and one's perception of the bodily actions of others, mediated by an innate body schema (Gallagher & Meltzoff 1996). But it also demonstrates an early form of social coupling, i.e. the fact that adult and infant can form a conversational unit from the beginning.

Partly, the spectacular finding of mirror neurons may also be interpreted in support of the claim that we can detect intentional states with a kind of immediacy, since perceiving other's actions activates one's own motor program responsible for that particular action (Rizzolatti & Sinigaglia 2008). Mirror neurons also fire differentially depending on which action chain a bodily movement is embedded in. Interestingly, they fail to be activated for observed actions that are not part of the observer's own motor repertoire (Buccino et al. 2004). I take it that these sensorimotor neurons support and enable the perceptual understanding of intentional action, and that the activation of one's own motor system reflects the foundational role of motor intentionality for cognitive intentionality. This interpretation is anticipated in Merleau-Ponty's phenomenological claim that one can *see* one's own possible bodily actions *in* the actions of the other (Merleau-Ponty 1964, p.117).

That infants take pleasure in directing the attention of the caregiver to oneself or to one's actions can be seen in the

still-face procedure: Between two and three months they already actively seek to re-engage a parent's attention when it has been disrupted (Murray & Trevarthen, 1985). Reddy (2003) argues that infants acquire an understanding of attention already in the first few months of life, primarily on the basis of the caregiver's attention towards some aspects or actions performed by the infant or the infant as a person. That is, the infant is first confronted with attention to the self and then to some aspect of the self or the self's actions. She argues that in scenes of *dyadic* mutual attention infants already demonstrate a *capacity for* and an interest in dealing with other's attention and that this provides the infant with the experience required for *further developing* her intentional repertoire. In the context of joint attention, attention is best characterized as an *act* of attending rather than an information-bearing mental state that arises passively. Focal attention is a continuous process *executed* by *the human agent*. The infant's alternation of attention on the object and the other subject (which is constitutive for joint attention) is essentially active and embodied since it involves head and eye movements, and possibly pointing gestures as communicative signals to direct the others' attention.

Gaze and Engagement with other agents. Direct interaction with another agent in joint attention also modulates our own processing of that object. Becchio et al. (2008) found that objects under the gaze of others "acquire properties that they would not display if not looked at", namely the gaze "enriches that object of motor, affective and status properties that go beyond its chemical or physical structure" (2008, 254). The authors call this "intentional imposition". – Other studies have shown that by twelve to fourteen months of age infants can use the gaze of others to predict a person's subsequent actions (Phillips, Wellman, Spelke 2002), can interpret a person's emotional expressions as being about the object at which she gazes (Repacholi 1998), and can interpret the words a person utters as naming the object at which she directs referential behaviors (Woodward 2003). Finally, Moll et al. (2007) demonstrated that one-year olds can attribute knowledge and ignorance to others but that such knowledge-ignorance understanding strongly depends on the joint engagement between infant and adult. Such knowledge could not be demonstrated independently of such engagement. These data support the interplay between object perception and social interaction. They also support the notion that embodied practices and active engagement with another agent plays a crucial role for (social) cognition.

Reciprocity and social affordances. Primary intersubjectivity in direct face-to-face social interaction between infant and caregiver displays an important dynamics and reciprocity that is crucial for *online* social cognition quite generally. Understanding others is typically not a unidirectional process: My own efforts to engage with the other prompt reactions feeding into a communication 'loop' characterized by *reciprocity* (Frith 2007, p.175). The importance of this is underestimated by theory-theory and simulation theory: Since they presuppose a detached observational stance towards the other (*offline* social cognition) instead of a more engaged interaction, they fail to account for this reciprocity. Basic social understanding is based on a sensitivity to "expressions of intentional and affective attitudes, as revealed in another's gaze, gesture, facial comportment" etc. (Hutto 2008, p.117). But in addition, perceiving the *meaning* of another's bodily expression requires processing the *social affordances* (Costall 1995) provided by them, analogous to the affordances provided by objects we perceive. The coupling between two agents in direct interaction is even more complex than the coupling between agent and environmental object. Due to the general flexibility and unpredictability of others in social interaction and a higher degree of uncertainty, social affordances are richer and more complex than affordances provided by objects. But they can prompt appropriate actions and reactions in a conversational context, culminating in the maintenance and extension of reciprocal relations. And healthy human beings can distinguish and pick up deliberate as well as inadvertently emitted communicative signals and to intuitively grasp the communicative context in which to make sense of another's behaviour (Senju & Csibra 2008). The studies mentioned earlier suggest that infants already possess this skill.

Autism. All this is crucial for the interpretation of autism as a social cognitive impairment. The enactive approach offers an interpretation of autism different from the traditional diagnosis as a lack of theory of mind based on a failure in false-belief tasks (Baron-Cohen 1995). It has recently been demonstrated that autistic patients can indeed pass such tasks when prompted to do so explicitly. Yet, this does not improve their social skills in direct interaction. As Senju et al. (2009) conclude from their study, patients with Asperger's are impaired in the "automatic online computation of others' mental states". They are not impaired in mindreading generally, but lack the more basic social skill to spontaneously encode socially relevant information and understand gestures and facial expressions *as* expressions of emotions (see Lee, Meyer, Hobson 1997). Thus, if autism is seen as a more general deficit in the sensorimotor, embodied and implicit *know-how to deal with other people*, this account can also explain other peculiarities significant for autism that have nothing to do with social cognition, e.g. the problems in lying, righting, sitting, crawling, and walking (Gallagher 2001).

4. Conclusion

In accordance with the enactive approach, it has been argued that cognition is based on sensorimotor skills executed by the organism as a whole in its exploration of objects in the immediate environment. It has been shown how central ideas from enactive cognition can be transferred to the social domain. The *primacy of embodied sensorimotor skills* is obvious in online social cognition when two agents are directly engaged in social interaction. *Social affordances* emerge from the *coupling* between two agents. Picking them up can prompt appropriate reactions, which in turn culminate in the *dynamics and reciprocity* characteristic of online social cognition. Displaying and perceiving bodily expressions of feelings, intentions etc. allows for a *skillful know-how* to deal

with other people, a spontaneous social understanding below and before mindreading which is impaired in autistic subjects. Direct engagement in online interaction also modulates object cognition. In this sense, it has been demonstrated that motor intentionality is more basic than cognitive intentionality both for object cognition and social cognition. Thus, if this foundational role can be spelled in more detail, then it promises to lead to a comprehensive account of intentionality.

Baron-Cohen, S. (1995). *Mindblindness. An Essay on Autism and Theory of Mind*. Cambridge, Mass.

Barresi, J., Moore, C. (1996). Intentional relations and social understanding. *Behavioral & Brain Sciences 19*, 107-122.

Becchio C., Bertone C., Castiello U. (2008). How the gaze of others influences object processing. *Trends in Cognitive Science, 12,* 254-258.

Brentano, F. (1874) *Psychologie vom empirischen Standpunkt*. 2 Vols. Ed. by O. Kraus. Leipzig 1924.

Campbell, J. (2005). Joint Attention and Common Knowledge". In: N. Eilan et al. (eds.), *Joint Attention: Communication and Other Minds* (pp. 287-297). Oxford: OUP.

Carpenter, M., Nagell, K., & Tomasello, M. (1998). Social cognition, joint attention, and communicative competence from 9 to 15 months of age. *Monographs of the Society for Research in Child Development, 63* (4, Serial No. 255).

Corkum, V. & Moore, C. (1995). Development of joint visual attention in infancy. In C. Moore & P. Dunham (Eds.), *Joint attention: Its origins and role in development.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Crane, T. (2001) *Elements of Mind*. Oxford: OUP.

Davidson, D. (2001) *Subjective, Intersubjective, Objective*. Oxford: Clarendon Press.

De Jaegher, H. & Di Paolo, E. (2007). Participatory sensemaking. An enactive approach to social cognition. *Phenomenology and the Cognitive Sciences, 6,* 485-507.

Eilan, N., Hoerl, C., McCormack, T., Roessler, J. (eds.) (2005) *Joint attention: Communication and other minds.* Oxford : OUP.

Gallagher, S. (2005) *How the body shapes the mind*. Oxford: OUP.

Gibson, E.J., Rader, N. (1979). Attention: The Perceiver as Performer? In G.A. Hale & M. Lewis (eds.). Attention and cognitive development (pp. 1-21). New York: Plenum Press.

Husserl, E. *Gesammelte Werke* (Abbr. as *Hua*). The Hague, Netherlands: Martinus Nijhoff, 1980

Hutto, D. (2008). *Folk-psychological narratives. The sociocultural basis of understanding reasons*. Cambridge, Mass.: MIT Press.

Leslie, A.M. (1987) Pretense and Representation: The origins of ‚Theory of Mind'. *Psychological Review, 94*, 412-426.

Merleau-Ponty, M. (1964). *Phenomenology of Perception*. New York: Humanities Press.

Moran, D. (1996). The Inaugural Address: Brentano's Thesis. *Proceedings of the Aristotelian Society Suppl. Vol. LXX,* 1-27.

Murray, L.,Trevarthen, C. (1985). Emotional regulation of interactions between two-months olds and their mothers. In T.M. Field & N.A. Fox (Eds.) *Social perception in infants*. Norwood, NY: Ablex.

Noë, A. (2009). *Out of our heads*. Hill & Wang.

Phillips, A., Wellman, H., & Spelke, E. (2002). Infants' ability to connect gaze and emotional expression as cues to intentional action. *Cognition, 85*(1), 53-78.

Rakoczy, H. (2006). Pretend play and the development of collective intentionality. *Cognitive Systems Research 7*, 113-127.

Reddy, V. (2003). On Being an Object of Attention: Implications for self-other-consciousness. Trends in Cognitive Science, 7 (9), 397-402.

Reddy, V. (2008). *How infants know minds*. Cambridge, Mass.: Harvard Univ. Press.

Repacholi, B.M. (1998). Infants' use of attentional cues to identify the referent of another person's emotional expression. *Developmental Psychology, 34,* 1017-1025.

Rizzollatti, G., Sinigaglia, C. (2008). *Mirrors in the brain*. Oxford: OUP.

Rolls, E. T., Grabenhorst, F., & Parris, B. A. (2008). Warm pleasant feelings in the brain. *Neuroimage, 41,* 1504–1513.

Schilbach, L., Wilms, M., Eickhoff, S. B., Romanzetti, S., Tepest, R., Bente, G. et al. (in press). Minds made for sharing. Initiating joint attention recruits reward-related neurocircuitry. *Journal of Cognitive Neuroscience*. doi:10.1162/jocn.2009.21401.

Schlicht, T. (2008) *Ein Stufenmodell der Intentionalität*. In P. Spät (Ed.) *Zur Zukunft der Philosophie des Geistes*. Paderborn: Mentis, pp. 59-91.

Searle, J.R. (1983) *Intentionality: an essay in the Philosophy of Mind*. Cambridge, Mass.: MIT Press.

Senju, A., Csibra, G. (2008). Gaze following in human infants depends on communicative signals. *Current Biology* 18, 668-671.

Senju, A., Southgate, V., White, S., Frith, U. (2009). Mindblind eyes: An absence of spontaneous theory of mind in Asperger syndrome. *Science* 325, 883-885.

Sinigaglia, C. (2008). *Enactive understanding and motor intentionality*. In F. Morganti, A. Carassa, G. Riva (Eds.) *Enacting Intersubjectivity: A Cognitive and Social Perspective on the Study of Interactions*. Amsterdam: IOS Press, pp. 17-32.

Sommerville, J., Woodward, A. (2005). Pulling out the intentional structure of human action: The relation between action production and processing in infancy. *Cognition 95,* 1-30.

Tomasello, M. (1999). *The cultural origins of human cognition.* Cambridge, Mass.: Harvard Univ. Press.

Thompson, E. (2007). *Mind in Life. Biology, Phenomenology, and the Sciences of the Mind.* Cambridge, Mass.: Harvard Univ. Press.

Wimmer, H. & Perner, J. (1983) Beliefs about beliefs: representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition, 13,* 103-128.

Woodward, A.L. (2003). Infants' use of action knowledge to get a grasp on words. In D. G. Hall and S. R. Waxman (eds.) *Weaving a lexicon* (pp. 149-172). Cambridge, Mass.: MIT Press.

Building a Model of Infant Social Interaction

Joshua M. Lewis
Department of Cognitive Science
University of California, San Diego
San Diego, CA 92093
josh@cogsci.ucsd.edu

Gedeon O. Deák
Department of Cognitive Science
University of California, San Diego
San Diego, CA 92093
deak@cogsci.ucsd.edu

Hector Jasso
hmjasso@gmail.com

Jochen Triesch
Frankfurt Institute for Advanced Studies
Johann Wolfgang Goethe University
60438 Frankfurt am Main, Germany
triesch@fias.uni-frankfurt.de

Abstract

Naturalistic observations of infant/caregiver social attention have yielded rich information about human social development. However, observational data are expensive, laborious, and reliant on fallible human coders. We model interactions between caregivers and infants using a three dimensional simulation environment in order to gain greater insight into the development of infant attention sharing, specifically gaze following. Most models of infant cognition have been only abstractly linked to the detail of a real life environment and to the perception-and-action physicality of human infants. Our simulation uses human data from videotaped infant/caregiver interactions and a rich 3D environment to model the development of gaze following. Initial tests suggest that infant gaze following can be learned in our simulation using parameters derived from behavioral data.

Keywords: embodiment; infancy; joint attention; simulation; social learning.

Human communication is a dauntingly complex system to model. Consider a seemingly simple system like an infant and caregiver playing together: even with language pared away, infant/caregiver social interactions feature a wide range of behaviors. These take place across many time scales in a complex environment. Moreover, the infant is a moving target; its brain and behavior change rapidly, and this requires caregivers to adapt to the infant's changing skills. Thus it is difficult to generate a powerful model of infant social behavior and learning.

Developing such a model is important because there is ample evidence that early social development has long term effects on (and likely serves as a foundation for) later social cognition, language, and even cognitive style and exploratory behavior [1]. In this paper we describe a modeling approach that is unique in two key areas, extending the approach introduced in [2]. First, we model both the learning agent (in this case the infant) and the agent's environment. Many models of infant learning use an abstract symbolic environment with little relation to the dynamic world infants experience. Ideally, simulations are comprised of both a biologically plausible learning model, and a physically and socially realistic environment [3]. The latter requirement is problematic because detailed data on the structure of infants' learning environment only exist in bits and pieces. Our second innovation is to directly tie behavioral data collected by our lab into our simulation environment, creating rich and realistic stimuli for our learning agent.

In the following subsections we will review the theoretical issues relevant to this work.

Embodied Modeling The goal of developmental modeling is to test theories of learning processes as they take place within organisms undergoing gross changes. Valid tests of these theories require additional theories as to the information patterns found in realistically structured environments [3]. Currently, however, we do not possess the computational resources to model human perceptual and neural systems, and our technological ability to simulate real, multi-modal environments is still primitive. The key, then, is to gradually converge on a set of biological traits that capture key properties of learning, as well as some key ecological patterns that can be simulated at a level of detail that is appropriate for the theoretical question at hand. This typically requires consideration of the physicality of the organism and the environment. That is, to test our theories with greater validity we must incorporate the embodiment of our models [4]. To the degree that we can embody simulations, we improve our tests of the motivating theory of development and learning.

Robotic studies are one way to achieve embodied simulations, and there are a growing number of good examples [5, 6, 7]. Robots can be placed in the same environments as infants and presented with identical stimuli. Unfortunately robotic studies are expensive, and they introduce tangential methodological issues—they require solving mechanical and computational problems simply to begin testing learning theories. Solving these problems is certainly important for some theoretical questions, but it is not currently necessary to address basic questions about infant social development. Additionally, robotic models cannot be run faster than real time, and they require active supervision. In many cases, current theories can realize faster progress by using simulations that retain elements of embodiment while greatly simplifying implementation and reducing cost.

Gaze Following We have been investigating the development of attention sharing behaviors in human infants. Attention sharing is a behavioral cornerstone of all social learning. In general it means one or more agents changing their fo-

cus of attention because they have observed another individual attending to some stimulus or area. A common example is following the line-of-gaze of another person. There is an extensive literature on the development of infants' attention sharing skills. This literature has focused on the development of gaze following, which is defined as reorienting one's direction of gaze to intersect with that of another person, based on encoding the other's head pose and/or eye direction.

Infants begin following other people's gaze between 6 and 12 months of age, and their ability to follow more and more subtle cues, to a wider range of their environment, increases significantly between 9 and 18 months of age [8, 9]. It is unknown by what mechanism infants develop more powerful gaze-following skills.

We have hypothesized [10] that infants' gaze following skills might emerge as the byproduct of a "basic set" of perceptual, learning, and affective traits that are in place within the first 2 to 3 months of age, well before fully developed gaze following can be observed. The basic set theory states that the following elements are sufficient (though not necessary) for joint attention:

- A set of motivational biases, in particular a preference for social stimuli such as human faces.
- Habituation as a basic reward attenuation mechanism.
- A learning mechanism such as temporal difference learning [11], to learn the temporal structure of predictable, contingent interactions between infant and caregiver.
- Early emerging perceptual traits such as attention shifting, face processing and sensitivity to motion, contrast, and color.
- A structured environment providing strong correlation between where caregivers look and where interesting things are.

This basic set of infant traits might be sufficient to generate new attention sharing skills. However, this requires that the infant learn on a regular regimen of well structured social input, as provided by an organized caregiver [10]. Our modeling efforts are meant to prove the plausibility of this theory. If they are unsuccessful, then perhaps additional mechanisms, such as special-purpose modules, are necessary for an agent to learn gaze following skills during the first 6-9 months of human social experience. The question, then, is how to generate valid simulations of this social learning process. We must imbue the simulated infant with biologically plausible perceptual, learning, and motivational traits, and we must imbue its environment with a reasonable facsimile of a natural social environment.

Naturalistic Social Coding The fine-grained structure of infant social environments is difficult to quantify. Although it is possible to derive gross patterns from previous observational and ethnographic behavioral studies, these tend to be sparse in details, and coded at such a low sampling rate that there is no information about caregivers' meaningful moment-by-moment action patterns. In most experimental studies of infant social responses, the social input from the adult is controlled and extremely artificial (e.g. [9]). Although these experimental studies are critical for establishing developmental "benchmarks" that a simulated infant should replicate, they do not provide information about real infant learning environments, which can be abstracted for simulation.

Our approach to solving this problems starts by generating a dense, rich video dataset of minimally directed interactions between infants and caregivers. Figure 1 shows one frame of these interactions from two separate viewpoints. By coding these interactions at 30fps in the manner described below, we generate a temporally detailed dataset that opens a new window into infant/caregiver interaction in a natural setting.

In the following sections we will explain our methodological workflow, describe the machine learning and computer vision techniques driving our simulated infant, present results from the simulation environment, and finally discuss the impact this work has on the modeling of infant social interaction.

Workflow

Our lab takes an end-to-end approach to infant social modeling (see Figure 2)—we start in the lab and in the homes of our subjects by collecting hours of audiovisual data from infant/caregiver interactions. These data consist of both semi-naturalistic free play sessions and scripted lab sessions. In the free play sessions caregivers are instructed to play with their infants using a supplied set of toys while the infant is seated in a tray chair. In lab sessions an experimenter performs a series of gaze and point maneuvers to salient objects in the room while holding the infant's attention. In both cases the interactions are recorded with audio from multiple camera angles. The lab has amassed many terabytes of this audiovisual data, which is passed off to a team of undergraduate research assistants who perform a detailed frame by frame coding of relevant events (e.g. gaze shifts, manual actions, environmental and toy-generated noise). These codes are stored in a database in order to facilitate an automated analysis of infant/caregiver behavior using custom software written in C# and Python. The automated analysis derives information from the coding such as the probability of the infant or caregiver to transition from one state to another (e.g. from looking at a toy to looking at a social partner), the duration of their actions, and extended events where the infant and caregiver move through a specified series of states within a restricted time window [12].

Our simulation environment can operate in two modes. In the first, it simply replicates caregiver behavior from a particular experimental session using the codes in the database. If the real-life caregiver started off looking at the infant and then switched to looking at a toy after 2.3 seconds, the simulated caregiver will do the same. In the second mode, the care-



Figure 1: Still picture from naturalistic study, from which the simulated caregiver behavior is derived.

giver behaves probabilistically based on the transition probabilities and timings derived from the automated analysis. In this way, the caregiver behaves realistically without replicating the steps of any particular subject; the simulation can run indefinitely. For example, if our data indicate that caregivers transition from holding an object to holding and moving an object 20% of the time that they change what they are doing, then our simulation likewise will make that transition 20% of the time. In addition, this mode allows our caregiver to (in principle) respond contingently to previous actions of the infant. Our simulation environment is implemented in C++ and we use hardware-accelerated OpenGL for the 3D rendering. Unfortunately, to our knowledge there is no open software for human simulation, so we use Boston Dynamics' DI-Guy platform for rendering and animating our caregiver and props. Finally, at the end of the chain, our infant learning agent processes rendered frames of the simulation using the OpenCV computer vision library [13]. At each time step of the simulation the *only* information the infant agent receives about its environment are these rendered frames—it extracts a reward signal and high level information about the environment using the computer vision techniques described in the next section.

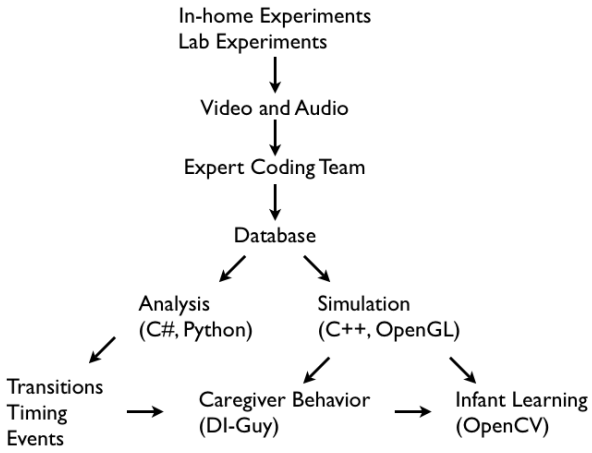


Figure 2: A flow chart depiction of the data collection, analysis and modeling work in our lab, annotated with relevant technologies.

Methods

There are three primary components to our simulation, the caregiver and environment, the infant agent's visual processing system, and its learning system. In this section we will detail the three components, starting with the caregiver and environment.

Caregiver and Environment Our simulation environment is set in the interior of a room containing a table and a chair. The caregiver is seated at the chair and interacts with toys placed on the table (see Figure 3, top). The caregiver is capable of interacting with more than one toy, but for our initial simulations we used just one toy, a red bus, for simplicity. The simulated caregiver occupies several different attention and action states. It can be: waving or not waving its arm, looking at the infant or the toy, and holding the toy or not. These states correspond to codes for caregiver motion, caregiver gaze target, and caregiver held object status in our empirical data. Because our caregiver is simulated as an actual body, these discrete behavior states manifest to the infant as a wide range of visual stimuli. For example while waving an object the caregiver's arm can be in many positions. Similarly, when looking to an object the caregiver's head pose varies over time as the motion is undertaken and the final head pose is based on the actual position of the object in the room.

From these data we also estimate the probability of transitioning between any of the states, and the simulated caregiver chooses its actions probabilistically based on these estimates (the caregiver is operating in the second mode de-

scribed above, not off a script). The caregiver uses two transition matrices: the first governs behavior with respect to the toy (holding and waving) and the second governs looking target. The interval between state transitions is based on the observed interval between separate caregiver behaviors (every 2.18 seconds) plus some uniform noise (+/- 1 second).

The infant also has a body in the environment (unseen from its perspective), with its head at about high-chair height. Changes in infant gaze target are accomplished by tying the position and orientation of a camera to the position and orientation of this body's head.

The objects in the environment are part of the DI-Guy package, which has a nice variety of (mostly military themed) props. A text configuration file specifies the props to load at the start of the simulation as well as their location, orientation and scale. Similarly, the text file specifies the initial location, orientation and appearance of human agents. In this way we can quickly modify the appearance of the simulation, add agents, and rearrange props.

Visual Processing In order for the infant agent to learn from its raw visual input , it needs to extract high level information about its environmental state as well as determine the reward value of the state that it is in. Since we are interested in gaze following, we extract the caregiver head position from the raw image, estimate head pose and use the discretized pose state as the infant agent's environmental state. To do this we first localize the caregiver's head by calculating the probability that each pixel in the raw image came from the known distribution of pixel properties in caregiver head pixels, running a Gaussian blur over that probability map, and then centering a head position rectangle over the maximum probability point on the blurred map. Technically, this is an application of `cvCalcBackProject` (to calculate the back projection of our face color histogram), `cvSmooth` (the Gaussian blur) and `cvMinMaxLoc` (to find the location of maximum probability in the image) from the OpenCV library. Pragmatically, we're only assuming the infant knows broadly what its caregiver's face looks like.

To calculate the head pose, we break the detected head region up into a left and a right segment then perform a color histogram comparison between the observed segments and model segments of left and right facing heads (using `cvCompareHist`). From the histogram distances we can calculate the probability that the caregiver is looking left or right by seeing how close the observed segments are to the models. If the segments are distant from both models we can infer that the caregiver's head pose is center. Again, the only assumption is that the infant knows generally what left and right facing heads look like. Finally we discretize the head pose probability into three states: left, center, and right. A visualization of this head position and pose detection can be seen in Figure 3, middle. The box represents the head position and the handles represent the pose probability.

To compute the reward for the current frame of input, we first calculate a salience map over the entire frame. The salience map has three components: motion, contrast, and saturation, and it is similar to salience-based visual processing approaches such as the one in [6]. The components are summed to represent overall saliency. Motion is calculated by comparison with the previous input frame (`cvAbsDiff`), contrast is derived from an edge detection routine (`cvSobel`), and saturation is extracted naturally

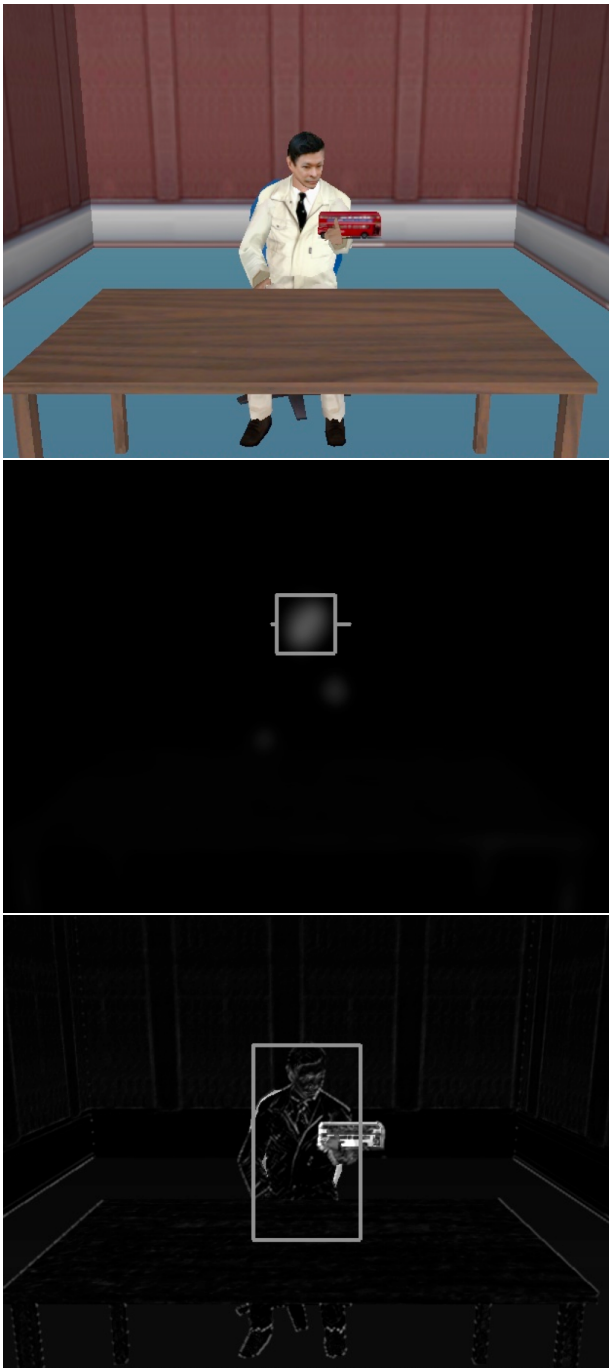


Figure 3: From top to bottom: the raw visual input to the infant agent, head detection and pose estimation output, salience and reward visualization.

from the color values of the pixels. Reward is then calculated by averaging the saliency values within the agent's center of vision (see Figure 3, bottom—the reward area is inside the rectangle). Since the agent only chooses looking direction on the horizontal axis, the center of vision is defined to be taller than it is wide.

Learning The agent uses a reinforcement learning [11] paradigm to choose its actions and learn from the consequences. Its state-action space is a cross of the discretized caregiver head poses and a set of five looking directions: left, near left, center, near right, and right. Every time the agent shifts gaze position, it updates its expected reward for the previous state-action pair using the following formula

$$er(i,j)_{new} = er(i,j)_{old} - \eta(er(i,j)_{old} - ar)$$

where $er(i,j)$ is the expected reward given caregiver head pose i and gaze action j, η is a learning rate parameter (set to 0.1 in our simulation) and ar is the average reward obtained since the last action j in state i. The agent changes gaze pose after a period of time derived from observed infant behavior (every 2.43 seconds) plus some uniform noise (+/- 1 second—a more complex but more realistic approach would be to draw fixation duration from an estimate of the fixation duration probability density function from actual infants).

It would be straightforward to increase the number of states and actions (e.g. by giving caregiver and infant looking direction a vertical degree of freedom) and add bells and whistles to the reward estimation process, but the purpose of this work is not to showcase machine learning techniques. Rather, we are investigating whether gaze following can be learned given a simple learning mechanism, data-driven caregiver behavior, and a complex simulated environment. The results of this endeavor are summarized and discussed in the next section.

Results

We ran our simulation for approximately 500 seconds (enough time for the infant to shift gaze about 200 times) with the infant agent watching a simulated caregiver interact with a single toy. The agent's expected reward over its state-action space is detailed in the table below. Looking at a location in the room with background (i.e. smallest) saliency results in a reward around 6.0, so that quantity is subtracted from the below numbers.

	Looking Direction				
CG Pose	left	near left	center	near right	right
left	1.30	1.54	3.58	2.62	1.79
center	1.09	2.62	8.50	3.20	1.97
right	1.56	1.72	1.71	1.43	0.76

Table 1: The final state/action reward space of the infant learning agent.

The course of learning over time is shown in Figure 4. The agent quickly learns that congruent gaze shifts result in higher reward and the advantage in expected reward generally increases over time.

Discussion

After a fairly short period of training, the agent expects more reward when its looking direction is congruent with the caregiver's head pose than when its looking direction is incongruent. For example, if the caregiver is looking to its left, then if the infant looks to the right it expects more reward (the infant and caregiver are facing each other and thus their looking directions to the same location are opposite). Both the near and far looking directions show this effect. Looking right in general is privileged because the caregiver is left handed (it only picks up objects with its left hand), and thus during time periods where the caregiver is holding the toy it is more likely to be near or far right than near or far left (from the infant's perspective).

Looking center is always very rewarding since the caregiver is at center. When the caregiver holds an object it will often be at center, and when it moves the object it generally is at center or near right. Motion is highly rewarding, and the caregiver is normally looking at center during motion, so the center/center expected reward is quite high. The caregiver also has a naturally higher contrast than other parts of the environment.

This general pattern of results fits other recent findings. It seems that infants in everyday setting are highly attentive to caregivers' manual actions [12], and this might bootstrap infants' learning of caregivers' head pose (because adults often look at what they are manipulating). It is also known that infants are attracted to faces, and the simulation results are consistent with that. Since the head pose and position estimation are not used in calculating reward, the infant agent learns that looking center (where the caregiver's head is) is valuable independent of the general knowledge about head appearance that it has.

These first results show that with a limited set of assumptions, a simple learning model, and a complex data-driven environment, gaze following can be learned. More importantly, this work sets the stage for even more detailed simulation of infant/caregiver interaction—such as interaction between more than two agents (a sibling agent, perhaps), reaching and grasping capability for the infants, and realistic audio cues. Further, since the infant agent no longer receives knowledge about its environmental state other than through visual processing, its input will degrade meaningfully and realistically. For example, if the infant picks up an object that occludes the caregiver, its head position and pose estimates will degrade realistically.

In the greater context of understanding infant social development, from modeling to robotics to experimental work, we see this as occupying a productive niche between disembodied and discretized 2D models and robotic agents. We open computer simulations up to state and action space complexities that mirror those in the real world, but our learning

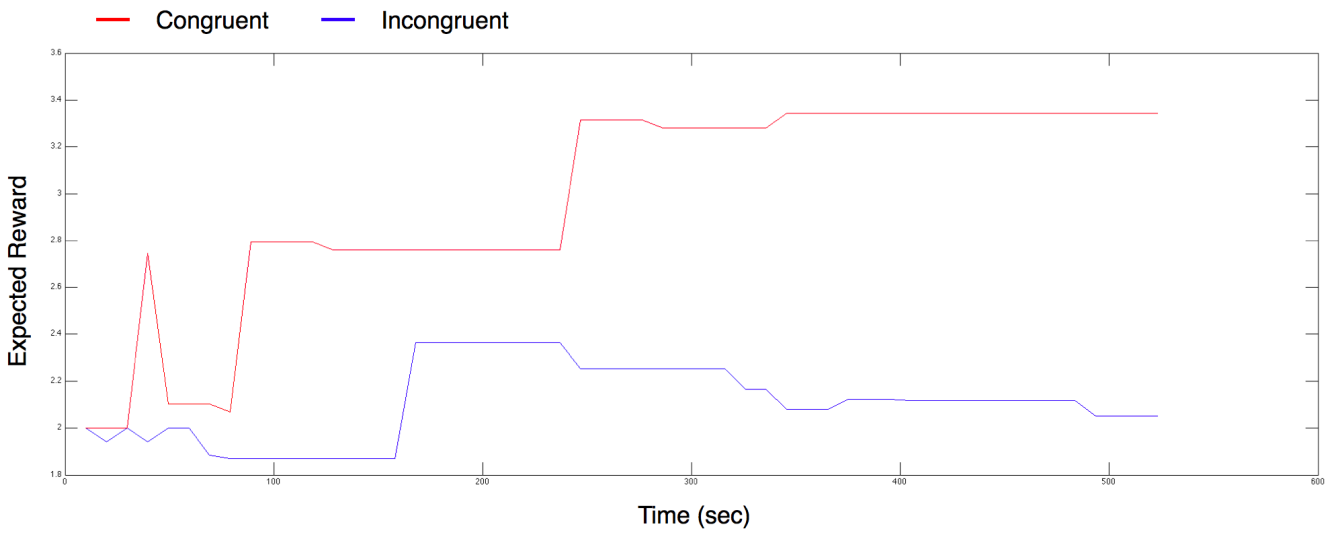


Figure 4: The sum reward expected from highly congruent gaze shifts (red, top right and bottom left of Table 1) and incongruent gaze shifts (blue, top left and bottom right) over the training period.

simulations are more convenient and we can have complete control over the agent and environment. Moreover, our simulations do not require the expensive and complicated hardware of robotic simulations; nor do they force us to address interesting but difficult and peripheral questions about motor control.

Acknowledgments

The authors would like to thank Ricky Ng for his valuable assistance on this project. This work was supported by research grants from the M.I.N.D. Institute and the National Alliance for Autism Research to G. Deák and J. Triesch, a National Science Foundation (SES-0527756) to G. Deák, and NSF IGERT Grant #DGE-0333451 to G. Cottrell and V de Sa.

References

[1] L. A. Sroufe, B. Egeland, E. Carlson, and W.A. Collins. *The Development of the Person: The Minnesota Study of Risk and Adaptation from Birth to Adulthood*. Guilford, New York, 2005.

[2] H. Jasso and J. Triesch. A virtual reality platform for modeling cognitive development. In *Biomimetic Neural Learning for Intelligent Robots*, volume 3575/2005, pages 211–224. Springer Berlin / Heidelberg, 2005.

[3] G. O. Deák, M.S. Bartlett, and T. Jebara. How social agents develop: New trends in integrative theory-building. *Neurocomputing*, 70:2139–2147, 2007.

[4] M. Wilson. Six views of embodied cognition. *Psychonomic Bulletin and Review*, 9:625636, 2002.

[5] G. Metta, G. Sandini, G. S., and L. Natale. Sensorimotor interaction in a developing robot. In *First International Workshop on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems*, pages 18–19. Lund University Press, 2001.

[6] Y. Nagai, K. Hosoda, A. Morita, and M. Asada. A constructive model for the development of joint attention. *Connection Science*, 20:211–229, 2003.

[7] N. Butko, I. Fasel, and J. R. Movellan. Learning about humans during the first 6 minutes of life. *Proceedings of the International Conference on Development and Learning*, 2006.

[8] G. Butterworth and N. Jarrett. What minds have in common is space: Spatial mechanisms serving joint visual attention in infancy. *British Journal of Developmental Psychology*, 9:5572, 1991.

[9] G. O. Deák, R. A. Flom, and A. D. Pick. Effects of gesture and target on 12- and 18-month-olds' joint visual attention to objects in front of or behind them. *Developmental Psychology*, 36:157–192, 2000.

[10] J. Triesch, C. Teuscher, G. O. Deák, and E. Carlson. Gaze-following: why (not) learn it? *Developmental Science*, 9:125–147, 2006.

[11] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998.

[12] A. Krasno, G. Deák, J. Triesch, and H. Jasso. Watch the hands: Do infants learn gaze-following from parents' object manipulation? In *Biennial Meeting of the Society for Research in Child Development*, 2007.

[13] OpenCV Wiki. `http://opencv.willowgarage.com/wiki/`.

The Emergence of Referential Gaze and Perspective-Taking in Infants

R. Joanne Jao (rjao@ucsd.edu)[*]
Marybel Robledo (marobledo@ucsd.edu)[*]
Gedeon O. Deák (deak@cogsci.ucsd.edu)[*]
[*]Department of Cognitive Science, 9500 Gilman Drive
La Jolla, CA 92093-0515 USA

Abstract

To understand the development of infant comprehension of visual obstructions and perspective-taking, this study tested the ability of $N = 28$ infants at 14, 16, and 18 months to adapt attention-sharing to visual constraints. An experimental task investigated how infants modify gaze following behaviors when an adult's line of sight is obstructed by a barrier. From 14 to 18 months, infants gradually learned to modify their search behavior when an adult looked toward a referent hidden behind a barrier from the infant's perspective. This suggests development of perspective-taking during this period. It also reveals age-related changes in infants' understanding of contextual effects on others' referential gaze in visually complex environments. Furthermore, the results address debates about "rich" versus "lean" theories of shared attention and intentionality.

Keywords: Perspective-taking; referential gaze following; visual obstruction; intentionality; cognitive development; social cognition.

Introduction

Infants learn socio-cultural routines and communicative patterns by sharing attention with adults. As they move into early childhood, 1- and 2-year-old infants gradually learn how another's attention can differ from their own; that is, they learn to take other people's visual perspectives in a shared environment. A critical component of this ability is *attention-following*, whereby infants follow the direction of attention of a more experienced person (e.g., a parent) to shift focus to interesting features of the environment. The clearest manifestation of this is referential gaze following, a type of triadic interaction which involves at least two people and a common referent. Referential gaze following is a two-part process: 1) one person directs her own attention toward a referent by orienting her eyes and usually her head, and 2) another person sees this behavior and consequently shifts attention in the direction of that referent (Butterworth & Jarrett, 1991; Scaife & Bruner, 1975). It is well established that referential gaze following plays a critical role in social learning, communication, and mental-state inferences (Argyle & Cook, 1976; Deák et al., 2008; Kleinke, 1986).

A question that has generated interest is how attention-following in general, and gaze following in particular, supports our inferences and predictions about the mental states of others. When one person subjugates her current interest to follow another person's attention, it may be assumed that the former is taking the latter's visual perspective. This implies that the follower imputes a mental or physiological state to the "looker." Indeed, adults attribute another person's direction of gaze to an internal state—their attention. However, it is difficult to tell what inferences infants make, or mental states they attribute to the people whose gaze they follow. Because infants cannot articulate their inferences, we can only observe their behavioral responses to other people's behavior (i.e., gaze-shifts). More generally, we do not know whether and how infants understand "seeing." Thus, the manner in which infants come to understand the "mental experience of seeing something" in others remains controversial (Caron, Butler, & Brooks, 2002).

"Rich" versus "Lean" Interpretations

One controversy about how children understand another person's looking behavior focuses on two distinct developmental interpretations. At one end, "rich" interpretations of gaze following assume that the follower explicitly represents the looker's intention to look in a particular region (Baron-Cohen, 1995; Woodward, 2003). At the other end, "lean" interpretations assert that gaze following emerges from simpler perceptual and learning processes, and structured social information (D'Entremont, 2000; Nagai et al., 2003; Triesch et al., 2006). Yet other positions focus on the transition from lean to rich inferences about others' gaze (Butterworth, 1998).

The rich interpretation refers to evidence that infants understand adults' gaze following behind visual obstructions (Brooks & Meltzoff, 2002, 2005). It also considers evidence that by 2 years, toddlers use adults' patterns of looking and emotional display to interpret their intentions (Tomasello, 1999). By contrast, the lean interpretation refers to evidence that infants' gaze following is modulated by factors such as target salience and the salience of an adult's head turn (Deák, Flom, & Pick, 2000). Also, earlier studies showed that infants follow an adult's head angle, but not eye direction (Corkum & Moore, 1998; Triesch, Jasso, & Deák, 2007). This is noteworthy because if infants do not know that the eyes mediate visual attention, then they do not grasp the basic mechanics of *seeing*. However, this conclusion has been challenged (Brooks & Meltzoff, 2002, 2005), as we review below. Given the diversity of evidence, we must consider the task paradigms used to test infants' knowledge. Since people eventually develop rich beliefs about looking and seeing, the controversy is inherently developmental. The question is at what age, and by what process, do children make mentalistic inferences about looking? Such inferences relate to the

origins of perspective-taking (Flavell, 1977). We now consider research evidence for age-related changes in infants' responses to looking and visual perspective-taking.

Age of Emergence

Recent studies have debated the age at which referential gaze following and perspective-taking emerge. Between 6 and 12 months of age, infants begin following an adult's direction of gaze (Adamson & Bakeman, 1991; Butterworth & Cochran, 1980; Butterworth & Jarrett, 1991; Corkum & Moore, 1998; D'Entremont, Hains, & Muir, 1997; Morissette, Ricard, & Décarie, 1995).

However, the age at which infants develop *referential* gaze following (i.e., knowing that someone's gaze is directed toward a percept, by virtue of seeing) is disputed. Brooks and Meltzoff (2005) reported that infants as young as 10 months start to realize that others are "'visually connected' to the external world." However, this is the only study showing such early ability, and the data are equivocal. There is more convergent evidence that referential gaze following emerges sometime between 12 and 18 months (Brooks & Meltzoff, 2002; Butler, Caron, & Brooks, 2000; Caron et al., 2002; Dunphy-Lelii & Wellman, 2004; Moll & Tomasello, 2004). For example, Deák et al. (2000) found that under optimal conditions, 12-month-olds follow gaze to targets located behind them. This is evidence of referential gaze following since it entails the representation that the looker is behaving in a way "toward" something, which the infants cannot detect. However, computer simulations show that this ability can be learned without high-level mental representations (Triesch et al., 2007).

By 18 to 24 months, there is substantial evidence for robust referential gaze following, particularly to targets that are visually occluded. That is, infants infer the existence of unseen objects and make inferences about others' visual perspectives. Notably, this is the same age that they begin to make inferences about other's mental states (Dunham & Dunham, 1995; Tomasello, 1999; Wellman, 1993).

The most active debate, then, centers on 12 to 18 months: if infants show referential gaze following by 12 or 14 months, it will suggest that gaze following is perhaps the earliest form of inferring others' mental states. If, however, referential gaze following does not emerge until 18 months, it will suggest that multiple forms of mentalistic inference emerge around the same time.

Problematic Occlusions

Many studies of referential gaze following in infants use large, distal occlusions (e.g., screen-like barriers) to obstruct either the infant's or adult's direct line of sight to a referent (Dunphy-Lelii & Wellman, 2004; Moll & Tomasello, 2004). Butler et al. (2000) compared infants' responses to transparent versus opaque barriers that were placed between a target referent and the experimenter. They found that 18-month-olds responded to the presence of an opaque barrier, whereas 14-month-olds did not reliably infer whether or not the adult could see the target through the barrier.

However, Dunphy-Lelii and Wellman (2004), who also used transparent and opaque barriers, found no change from 14 to 18 months. Infants by 14 months followed the experimenter's gaze more often when the barrier was transparent than when it was opaque.

In addressing this divergence of results, Moll and Tomasello (2004) charged that the task was too unnatural. They therefore used a different paradigm in which the target was placed behind a barrier from the infant's perspective. If the infant followed the experimenter's gaze, she would only see a boring opaque barrier. However, if the infant understood that the adult was looking at something, she would move around to peer behind the barrier. Results suggested that this behavior starts to emerge in some 12-month-old infants, and is more robust in 18-month-olds. This goes beyond Butler et al.'s (2000) results to suggest that 12-month-olds do basic referential gaze following.

Goals of the Current Study

We sought to resolve uncertainties about the development of referential gaze following in the second year. Since no study has examined the *process* of emergence, we tested infants at 14, 16, and 18 months, as a part of a longitudinal study. By testing at 3 bi-monthly ages, we might resolve conflicting results from previous studies of widely differing age groups. We can also test the stability and predictability of individual differences in development, which has not yet been studied.

Similar to Moll and Tomasello (2004), we used opaque barriers in a distal barrier paradigm, but added some improved controls. With a barrier on each side, one barrier occluded a target from the infant while the other displayed a target to both infant and adult. By making one target visible, we assessed each infant's baseline gaze following. We compared this to each infant's tendency to move and peer around the blank barrier when the adult looked toward it. This verified that the infant could visually orient to the experimenter's head and eye direction, thus making interpretable the "more demanding response" (Moll & Tomasello, 2004) of peering around when the adult's looking behavior was ambiguous. That is, in actively leaning forward or moving to look around a barrier to an occluded object, the infant's behavior signals her awareness of the implications of the adult's looking behavior.

In sum, the current investigation seeks to: 1) establish age-related trends in infants' acquisition of referential gaze following when the physical environment suggests that another person has a different visual perspective; and 2) relate the results to prior, simpler gaze following skills. Therefore, the goal of this study is to establish the validity of referential gaze following tasks and examine their implications for perspective-taking.

Method

Participants

Twenty-eight infants (17 males, 11 females) participated at 14 months (mean age = 427 days, *SD* = 7), 16 months (*M* =

491 days, *SD* = 13), and 18 months (*M* = 550 days, *SD* = 8). All infants were walking independently by 12.3 months (*SD* = 1.4). Infants were primarily of middle-class households from the San Diego area.

Materials

Two featureless, rigid brown barriers (92 cm x 58 cm) were placed side-by-side 1.2 m apart. Two 2D foam shapes (10.2 cm x 10.2 cm) were used as target stimuli. The control target was a red circle and the experimental target was a red duck. A researcher ("cue-giver" or CG) interacted directly with the infant in a quiet, controlled testing room (4.0 m by 3.6 m). A second researcher ("observer" or OB) monitored and recorded the infant's behavioral responses from an adjacent room. Target cues and locations were given by the OB to CG using a two-way radio and earpiece. A metronome was used to accurately time cue-length and inter-trial intervals. To control the visual scene, the CG wore a gray sweatshirt and tied her hair in a ponytail. Both the CG and parent were seated on the floor on cushions.

Procedure and Design

All infants participated in three sessions at 14, 16, and 18 months of age. Before each session, informed consent was obtained from the parent. Each session consisted of eight 10-second test trials.

Before the session, the barriers were placed on either side of the CG and the infant, who sat facing one another approximately 61 cm apart. Targets were attached to the middle of each barrier 46 cm above the floor. The control target was placed on the front of one barrier and the experimental target on the back of the other, relative to the infant. Barriers were angled so that both targets were visible to the CG, but only the control target was visible to the infant (Figure 1). The parent sat with the infant in her lap such that the infant could freely rise to walk around at will. The parent was instructed to provide no cues, and the infant remained seated between trials. The CG sat with her hands in her lap and displayed an open friendly expression.

To orient the infant to the target locations, the CG first held the control target at eye-level and said "[infant's name], look!" As she placed the target on the front of one barrier, she said "I'm going to put it there." The CG then held up the experimental target, saying "[infant's name], look!" She then placed the target on the back of the second barrier, saying "I'm going to put it here."

In each trial, OB gave the CG the onset cue, and the CG began the trial with an open-mouth smile. She called the infant's name until eye contact was made, and after two seconds said, "[infant's name], look!" The CG immediately turned to look directly at the target for four seconds. Then, CG looked back to the infant, establishing eye contact if possible, and said "Can you get it for me?" while executing another gaze cue to the target for four seconds.

After four trials (2 experimental, 2 control) in one left-right configuration, the barriers were switched between sides and the last four test trials were given. Between sessions the barrier sides were counterbalanced. Across trials, condition (control, experimental), direction (left, right), and target (circle, duck) were also counterbalanced.

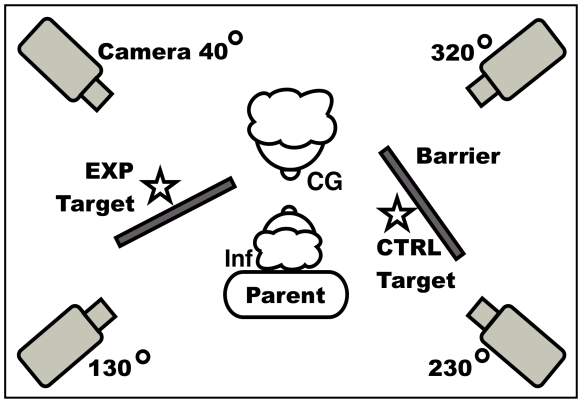


Figure 1: One configuration of the experimental setup.

Each session was recorded at 30 fps with four video cameras placed in the corners of the room at infant eye-level. The cameras recorded onto on-board hard drives, and simultaneously pushed video to be time-stamped and captured on a Level 5 RAID. In addition, a video camera with a fisheye lens was mounted above the infant's head, and was time-stamped and captured in the same manner.

Coding

Videos of infant behaviors were examined frame-by-frame. The infant's first look after each cue by the CG were coded (i.e., anticipatory looks were not examined). Furthermore, trials were coded only if the infant saw the CG's cue. Possible visual directions included looks to the correct target, incorrect target, front of the barrier that hid the target in experimental trials, CG, and "other" (i.e., anything else in the room). Success in referential gaze following was defined as the infant looking to the correct target location (i.e., specified by the CG's cue versus looking to the other target location or not looking at all). In a control trial, this meant looking toward the visible target after the CG's cue. In an experimental trial, this meant leaning or moving forward to peer around the back of the appropriate barrier.

An *incorrect* look was coded if the infant looked at the wrong target or to the front of the appropriate barrier in the experimental condition. A *non-look* was coded if the infant did not look to any target, but instead looked at the CG or at an irrelevant feature of the room.

Results

Proportions of correct looks were submitted to a 3 (age: 14, 16, 18 months) x 2 (condition: experimental vs. control) analysis of variance (ANOVA) within subjects. There was a significant main effect of age $F(2, 25) = 2.56$, $p < .09$; 18-month-olds looked proportionately more ($M = 0.55$, $SD = 0.39$) to the correct targets than 14- ($M = 0.41$, $SD = 0.44$) or

16-month olds (M = 0.48, SD = 0.40). There was also a significant effect of condition $F(1, 25) = 147.70$, $p < .001$. Infants looked more to the correct targets when they were visible in the control condition (M = 0.75, SD = 0.31) than when they were hidden behind barriers in the experimental condition (M = 0.20, SD = 0.29). However, there was no significant effect for the age x condition interaction, $F(2, 25) = 0.46$. Separate Student's t-tests were used to compare the factors of age and condition in looking behaviors (see Table 1). As expected, there was a significant difference between the control and experimental conditions for correct looks and non-looks ($p < .001$) at each age.

Table 1: t-tests comparing conditions across age.

	Correct Looks				Nonlooks			
Age	*t*	*df*	*SD*	*p-value*	*t*	*df*	*SD*	*p-value*
14	-7.43	27	0.39	$p < .000$	7.55	27	0.34	$p < .000$
16	-10.87	27	0.29	$p < .000$	9.53	27	0.24	$p < .000$
18	-7.03	27	0.38	$p < .000$	5.13	27	0.36	$p < .000$

Longitudinally, 71% of the infants performed steadily well in the control condition, while 4% of infants performed similarly well in the experimental condition. Comparatively, 14% of infants in the control condition and 21% of infants in the experimental condition improved in their performance from 14 to 18 months. Across all 3 age groups, 11% of infants in the control condition and 25% of infants in the experimental condition showed mixed abilities.

In addition to significant effects of age and condition, as well as longitudinal performance, there were subtler developmental changes that occurred between 14 and 18 months. Generally, infants at 14, 16, and 18 months looked to the correct target in the control condition; this showed a trend of increasing consistency, with mean proportions of 0.69 (SD = 0.40), 0.77 (SD = 0.27), and 0.80 (SD = 0.25) at the three ages, respectively. In the experimental condition, the mean proportions of looks to the correct target also increased from 0.14 (SD = 0.27), to 0.18 (SD = 0.26), and 0.29 (SD = 0.34), respectively (Figure 2).

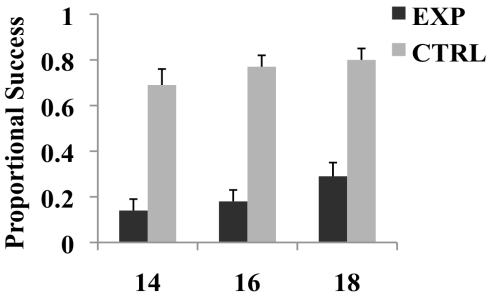


Figure 2: Mean proportions of success in looking behavior (with SE) in experimental and control conditions across age.

To understand these trends more fully, we examined the looking behaviors in each trial, distinguishing between correct looks, non-looks, and incorrect looks. Figure 3 displays these mean proportions at 14 months as a function of condition and looking behavior to illustrate the general pattern. While the general trends remained the same from 14 to 18 months, there was a decrease in the proportion of incorrect looks in the experimental condition from 0.52 (SD = 0.31) to 0.42 (SD = 0.32). Within the incorrect looks, looks to the front of the appropriate barrier in the experimental condition decreased from 0.39 (SD = 0.25) to 0.30 (SD = 0.29), and looks to the wrong target decreased slightly from 0.13 (SD = 0.18) to 0.12 (SD = 0.16).

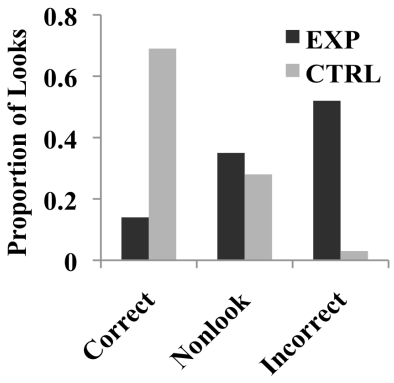


Figure 3: 14 month mean proportions of looking behavior.

However, one concern about these parametric data is that nothing compels infants to rise and peer around the barrier—especially after having done so once, since they may not be motivated to continue looking at such simple targets. To address this, we considered a less demanding measure of infants' understanding that the experimenter might be looking at something they could not see. The "1 Trial Pass" criterion defined an infant as "passing" if she looked to at least one correct target by moving or leaning forward to look behind a barrier. Since the active movement of searching for an unseen target indicates intentionality, this seems to show some basic level of understanding of visual obstructions and referential gaze. (In support of this, infants virtually never got up to look around the barrier in control trials.) Results showed a steady increase with age in the proportion of infants who looked to the correct target. In the control condition, 82% of the infants at 14 months, and 96% of the infants at 16 and 18 months, passed at least one trial (i.e., followed gaze to the visible target). In the experimental condition, 25%, 43%, and 54% of the infants, respectively, passed at least one trial. Thus, twice as many infants at 18 months followed blocked gaze successfully than at 14 months.

Discussion

The results show that some infants at 14 months are starting to develop an understanding of visual barriers and perspective-taking in referential gaze. This development goes beyond the ability to merely follow gaze, since infants were clearly able to do so by 14 months, as shown by the results in control trials. In the experimental trials, however,

infants must determine that the adult is looking at a referent that the infants cannot see. At 14 months, some infants looked behind the barrier to the correct target, but did so much less than they looked to the front of the barrier. Yet at 18 months, infants peered behind the barrier to the correct target just as often as they looked to the front of the barrier. In addition, the "1 Trial Pass" analysis suggests that by 18 months, more than half of infants develop some Level 1 visual perspective-taking (Flavell, 1977), inferring an unseen target on at least one trial.

A longitudinal analysis suggests that a sizeable minority of infants improved in the experimental condition. Thus, there is some sort of learning from 14 to 18 months. However, there was also some within-infant variability between sessions, suggesting sources of unidentified situation-specific variability.

These results support Butler et al.'s (2000); 18-month-old infants respond significantly more than 14-month-olds to an adult looking behind barriers at hidden targets. Yet, possibly due to our more "natural" experimental design with multiple barriers and targets (inspired by Moll & Tomasello, 2004), our results showed a stronger effect than Butler et al. (2000). In their experiment, only 33% of 18-month-olds and no 14-month-olds leaned forward to look behind a barrier that obstructed a target. Thus, they concluded that infants at 18 months understand referential gaze and visual obstructions, whereas infants at 14 months do not. In the current investigation, 54% of the infants at 18 months and 25% of the infants at 14 months leaned forward to look behind the barrier. This demonstrates that visual perspective-taking develops considerably, and is clearly established, by 18 months, but it remains unclear whether 14-month-old infants have any functional capacity for visual perspective-taking. The current results suggest that some 14-month-olds are starting to develop an incipient understanding, as suggested by Dunphy-Lelii and Wellman (2004). However, we cannot say whether, for example, providing 14-month-old infants with additional training or reinforcement would increase their rate of responsiveness to an adult looking behind a barrier.

In order to better understand the developmental trajectory of referential gaze following, and to establish more precisely the age at which this understanding emerges, we considered results from a prior session in the longitudinal study. A subset of the infants (N = 18) who had performed simpler gaze following tasks at 12 months was compared to their performance at 14 months in the current task. Overall, the infants at 12 months occasionally followed gaze to visible targets (M = 0.43, SD = 0.19), but seldom followed gaze to targets located behind them (M = 0.11, SD = 0.27). This can be considered a "first step" towards referential gaze following. Furthermore, when subjected to the "1 Trial Pass" criterion, only 16% of the infants successfully looked to at least one target out of their direct view. By comparison, the same infants at 14 months made a similar proportion of successful looks to targets behind barriers (M = 0.14, SD = 0.25), but a higher proportion of them met the "1 Trial Pass" criterion (28%) in the experimental condition. Generally, infants at 12 months seldom followed gaze to unseen targets located behind them, therefore failing to show referential gaze following ability. Somewhat more infants showed at least minimal referential gaze perspective-taking at 14 months. Thus, our results do not support claims that infants even younger than 12 months have a concept of intentional behavior (Brooks & Meltzoff, 2002). Rather, our data suggest a shift from a leaner interpretation of gaze following in most 12-month-old infants, to a richer interpretation in most 18-month-old infants. Given this shift, we favor a learning-based account (in which gaze following begins perceptually, and then becomes referential as well as intentional), over accounts that assume strictly either a maturational onset of perceptual processes or an inherent understanding of the referential nature of gaze following.

Between their first and second birthdays, most infants develop the understanding that there may exist some object of interest at which an adult is looking, and that adult visual perspectives, in general, offer useful information. From 14 to 18 months, infants learn that acting on that information, via referential gaze following, can be rewarding. Even if that information consists of a referent that is visually occluded, infants will deliberately move to a proper viewing perspective to search for the inferred referent. Notably, the gradual differentiation of performance in the experimental and control conditions of the current investigation offers some insights into infants' growing capacity for detecting cues of others' perceptual states. This capacity is based in, and demonstrated by, their active search patterns, which for unseen targets might serve as an interim "trial and error" strategy that allows infants to test or verify the objective underlying adult looking behavior. However, this strategy is minimal at 14 months of age.

It is worth noting that the behavioral measures used in our study assess infants' performance and emerging ability, rather than level of competence. Indeed, all of the infants were capable of walking independently or crawling to look behind the barrier. This demonstrates that any possible differences in motor capabilities were not the primary source of divergence in the data between 14 and 18 months. As an additional factor, the manipulation of infants' motivational states highly influences competence, and may impede performance.

Finally, little is known about how infant gaze following skills relate to other spatial representational skills. However, referential gaze following provides a unique arena for studying how infants develop skills for simultaneously processing social and spatial information, and using these processing skills to support inferences about non-obvious events and ecological relations. Referential gaze following offers a new ability to synthesize information about other people's embodied actions in a shared environment to infer unperceivable states. This ability may be critical for impending changes in social and communicative knowledge.

The current study confirms developmental trends in

referential gaze following from 14 to 18 months of age. Together with previous studies, this contributes to our understanding of infants' referential gaze following, perspective-taking, vision comprehension, and ultimately, theory of mind.

Acknowledgments

The authors thank the mothers and infants who participated in the study. We also thank Ana Ramundo, Corrine Zavala, Katy Brecht, and members of the Cognitive Development Laboratory for assistance with data collection and coding. This project was supported by NSF award HSD 0527756 to G. Deák.

References

Adamson, L., & Bakeman, R. (1991). The development of shared attention during infancy. *Annals of Child Development*, 8, 1-41.

Argyle, M., & Cook, M. (1976). *Gaze and mutual gaze*. Cambridge, England: Cambridge University Press.

Baron-Cohen, S. (1995). The eye direction detector (EDD) and the shared attention mechanism (SAM): Two cases for evolutionary psychology. In C. Moore and P. Dunham (Eds.), *Joint attention: Its origins and role in development*. Hillsdale, NJ: Erlbaum.

Brooks, R., & Meltzoff, A. N. (2002). The Importance of Eyes: How Infants Interpret Adult Looking Behavior. *Developmental Psychology,* 38(6), 958-966.

Brooks, R., & Meltzoff, A. N. (2005). The development of gaze following and its relation to language. *Developmental Science,* 8 (6), 535-543.

Butler, S. C., Caron, A. J., & Brooks, R. (2000). Infant Understanding of the Referential Nature of Looking. J*ournal of Cognition and Development*, 1(4), 359-377.

Butterworth, G. E. (1998). Origins of joint visual attention in infancy: commentary on Carpenter et al. *Monographs of the Society for Research in Child Development,* 63(4), 144-166.

Butterworth, G. E., & Cochran, E. (1980). Towards a mechanism of joint visual attention in human infancy. *International Journal of Behavioral Development,* 3, 253-272.

Butterworth, G. E., & Jarrett, N. (1991). What minds have in common is space: Spatial mechanisms serving joint visual attention in infancy. *British Journal of Developmental Psychology,* 9, 55-72.

Caron, A. J., Butler, S., & Brooks, R. (2002). Gaze following at 12 and 14 months: Do the eyes matter? *British Journal of Developmental Psychology*, 20, 225-239.

Caron, A. J., Kiel, E. J., Dayton, M., & Butler, S. C. (2002). Comprehension of the referential intent of looking and pointing between 12 and 15 months. *Journal of Cognition and Development,* 3(4), 445-464.

Corkum, V., & Moore, C. (1998). The origins of joint attention in infancy. *Developmental Psychology*, 34, 28-38.

D'Entremont, B. (2000). A perceptual-attentional explanation of gaze following in 3- and 6-month-olds. *Developmental Science*, 3, 302-311.

D'Entremont, B., Hains, S. M. J., & Muir, D. W. (1997). A Demonstration of Gaze following in 3- to 6-Month-Olds. *Infant Behavior and Development,* 20(4), 569-572.

Deák, G. O., Flom, R. A., & Pick, A. D. (2000). Effects of gesture and target on 12- and 18-month-olds' joint visual attention to objects in front of or behind them. *Developmental Psychology*, 36(4), 511-523.

Deák, G. O., Walden, T. A., Kaiser, M. Y., & Lewis, A. (2008). Drive from distraction: How infants respond to parents' attempts to elicit and re-direct their attention. *Infant Behavior and Development,* 31, 34-50.

Dunham, P. J., & Dunham, F. (1995). Optimal social structure and adaptive infant development. In C. Moore and P. J. Dunham (Eds.), *Joint attention: Its origins and role in development*. Hillsdale, NJ: Erlbaum.

Dunphy-Lelii, S., & Wellman, H. M. (2004). Infants' understanding of occlusion of others' line-of-sight: Implications for an emerging theory of mind. *European Journal of Development Psychology* ,1(1), 49-66.

Flavell, J.H. (1977). The development of knowledge about visual perception. *Nebraska Symposium on Motivation: Vol. 25* (pp. 43–76). Lincoln, NE: Univ. of Nebraska Press.

Kleinke, C. L. (1986). Gaze and eye contact: A research review. *Psychological Bulletin,* 100, 417-425.

Moll, H., & Tomasello, M. (2004). 12- and 18-month-old infants follow gaze to space behind barriers. *Developmental Science,* 7(1), F1-F9.

Morissette, P., Ricard, M., & Décarie, T. G. (1995). Joint visual attention and pointing in infancy: A longitudinal study of comprehension. *British Journal of Developmental Psychology*, 13, 163-175.

Nagai, Y., Hosoda, K., Morita, A., & Asada, M. (2003). A constructive model for the development of joint attention. *Connection Science, 15,* 211–229.

Scaife, M., & Bruner, J. S. (1975). The capacity for joint visual attention in the infant. *Nature,* 253, 265-266.

Tomasello, M. (1999). *The cultural origins of human cognition.* Cambridge, MA: Harvard University Press.

Triesch, J., Jasso, H., & Deák, G.O. (2007). Emergence of mirror neurons in a model of gaze following. *Adaptive Behavior, 15*, 149-165.

Triesch, J., Teuscher, C., Deák, G., & Carlson, E. (2006). Gaze following: Why (not) learn it? *Developmental Science, 9*, 125-147.

Wellman, H. M. (1993). Early understanding of mind: The normal case. In S. Baron-Cohen, H. Tager-Flusberg, and D. Cohen (Eds.), *Understanding other minds: Perspectives from Autism*. London, England: Oxford University Press.

Woodward, A. L. (2003). Infants' developing understanding of the link between looker and object. *Developmental Science*, 6(3), 297-311.

Joint Perception: Gaze and Beliefs about Social Context

Daniel C. Richardson (dcr@eyethink.org)
Chris N.H. Street (chris@eyethink.org)
Joanne Tan (tan.yin@ucl.ac.uk)
Cognitive, Perceptual & Brain sciences, University College London
Gower Street, London WC1E 6BT, UK

Abstract

The way that we look at images is influenced by social context. Previously we demonstrated this phenomenon of *joint perception.* If lone participants believed that an unseen other person was also looking at the images they saw, it shifted the balance of their gaze between negative and positive images. The direction of this shift depended upon whether participants thought that later they would be compared against the other person or would be collaborating with them. Here we examined whether the joint perception is caused by beliefs about shared experience (looking at the same images) or beliefs about joint action (being engaged in the same task with the images). We place our results in the context of the emerging field of *joint action*, and discuss their connection to notions of group emotion and situated cognition. Such findings reveal the persuasive and subtle effect of social context upon cognitive and perceptual processes.

Keywords: vision; joint action; eye movements; social cognition, situated cognition

Introduction

Cognition is enveloped by social context. It is rare that we use our cognitive or perceptual faculties outside of the world of social influence, what Allport (1954/1979) described as the real or imagined presence of other people. Yet in cognitive and perceptual laboratories, we typically place participants in experimental quarantine away from the confounds of social interaction. The risk of this strategy is that we overlook the ways in which cognitive and perceptual processes interact with social context.

It is now well demonstrated that social cues such as eye contact and gaze direction are attended to in fundamentally different ways from non-social stimuli, both in terms of higher-level attentional selection (e.g. Birmingham, Bischof & Kingstone, 2008a, b, 2009; Frischen, Bayliss & Tipper, 2007; Senju & Johnson, 2009) and their different neurological subsystems (e.g. Greene et al., 2009; Itier & Batty, 2009; Ristic, Friesen & Kingstone, 2002). These studies, and many others, show how perceptual processing differs for social and non-social stimuli (Cacioppo, Visser & Pickett, 2005).

In studies of *joint perception*, this relationship is turned on its head; we keep the stimuli constant and examine how different social cues exert an influence on perceptual processing. The first demonstration (Richardson, Hoover & Ghane, 2008) presented participants with a set of four images on screen for eight seconds. On different trials, participants either believed that in a cubicle next door another participant was looking at the same images, or that the person next door was looking at a set of unrelated symbols. In each set of images, there was one picture with a negative valence (such as crying child), one with a positive valence (a smiling couple) and two neutral images with no strong valence. When participants believed that they were the only ones currently looking at the images, they looked more at the unpleasant ones. When they thought they were looking jointly with another, they looked more at the pleasant images.

Participants in this experiment could not see or interact with each other. Yet their gaze was systematically shifted if they imagined that another person was looking at the same stimuli. There have previously been similar demonstrations of the influence of social context on social or affective responses, for example, that people will smile and laugh more if they imagine that a friend elsewhere is currently watching the same comedy clip as themselves (Fridlund, 1991). However, the joint perception result showed that, on a trial-by-trial basis, social context can shape a low level perceptual/cognitive process.

The original experiment was carried out at UC Santa Cruz in the US. A replication was soon performed at University College London in the UK (Richardson et al., 2009). The same pervasive effect of social context was found. Gaze patterns shifted in response to joint perception. However, in this case, when participants believed that they were looking together, they looked more at the negative images. The contrasting US and UK data is shown in the top panel of *Figure 1*. What is depicted is the total fixation duration for the positive and negative images during joint and alone looking. Each study found a significant interaction between picture valence and social context, and between the two experiments there was a significant three way interaction, showing that the direction of the effect changed.

Though there were many differences between the laboratories' set up and the participant populations, we hypothesised that an important determinant might be how participants construed that task. One criticism of the first study was that participants did not know why they were looking at the images, and why the person next door was (sometimes) doing the same thing. So, in subsequent research in London (Richardson et al., 2009), we repeated the experiments but told pairs of participants either that we would be comparing their picture preferences (comparison task), or that they would be collaborating on a memory task afterwards (collaboration task). As *Figure 2* shows, we found a pattern of results that mimicked the US / UK differences, and also produced a significant three way interaction. People who thought they were being compared to each other tended to look at the negative and positive images equally in the joint condition, like the US participants. People who thought that they were

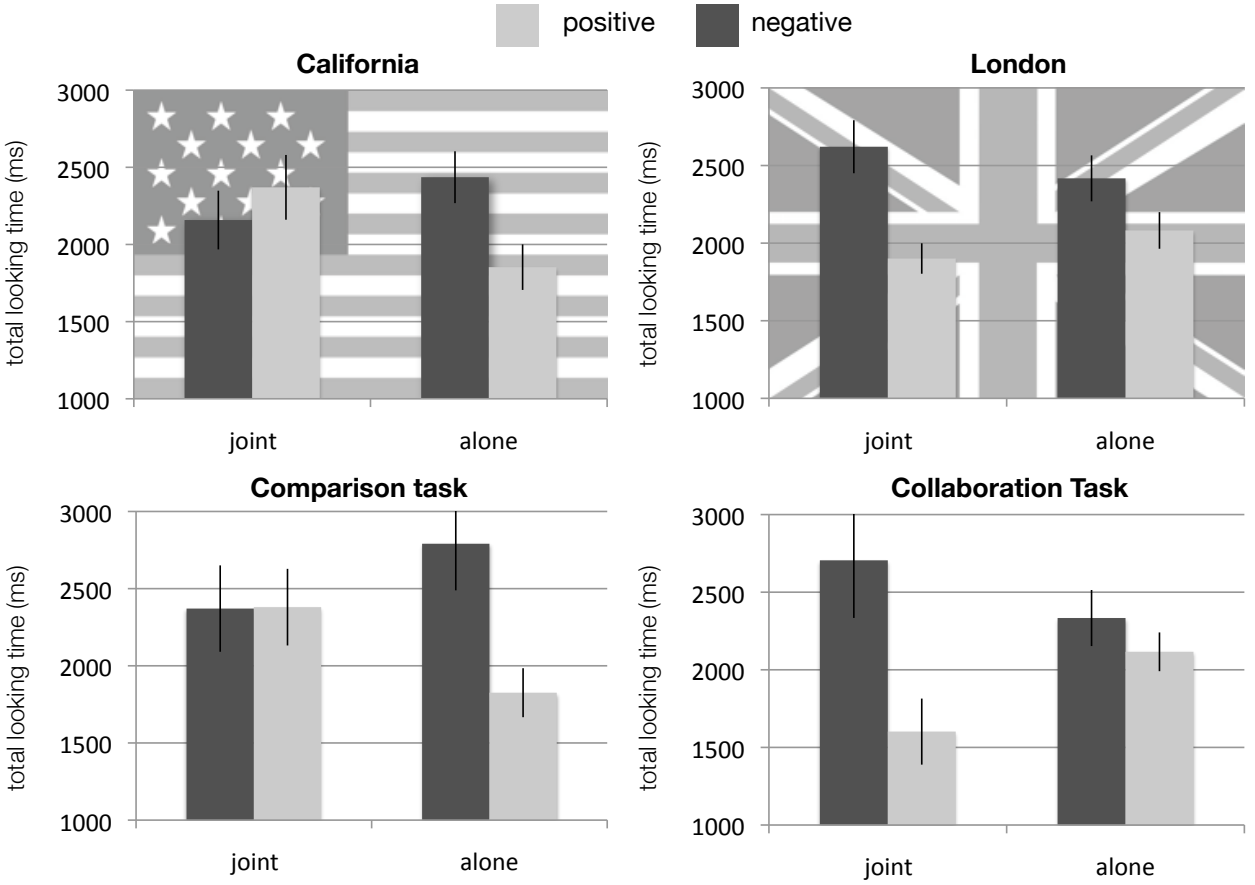


Figure 1. Results from Richardson, Hoover and Ghane (2008) and Richardson *et al.* (2009).

collaborating looked more at the negative images in the joint condition, like the London participants who did not get task instructions. There could be other reasons, of course, why the US and UK participants differed, but one plausible reason appears to be that in the absence of instructions, they interpreted the task in opposite ways. We can only speculate as to the reason the US participants might felt that they were being compared (they are academically evaluated more frequently than UK students), or it might have been that the physical setup of the lab (two adjoining cubicles, rather than one big room) engendered a feeling of being contrasted.

These previous studies have shown that gaze patterns can by systematically influenced by beliefs about social context, and that the direction of this influence is sensitive to differences in how participants construe their task. In the current experiment, we zoom in to this concept of looking at something 'together'.

For the joint perception effect to occur, is it enough to *experience* a set of stimuli at the same time as another person? Or do participants have to believe that they are engaged in the *same task* as the other person? In this experiment, unlike those described above, the participants always believed that they were looking at the same images as each other. What changed, trial-by-trial, was the task that they were doing, and the task that they believed their partner was doing. Inspired by the seminal work on *joint action* (Sebanz, Bekkering & Knoblich, 2006) that we discuss below, we predicted that joint perception effects would be strongest when participants believed that they were not just passively sharing an experience, but acting jointly.

Methods

Participants

32 University College London students (9 male) participated voluntarily or for course credit. Data from 4 participants were unusable due to equipment calibration problems.

Note that although we actually ran pairs of participants simultaneously in the lab, their experiments were run and their data analysed independently from each other. This is because participants could not see or interact with each other during the experiment. In effect, they acted as a mute social context for each other.

Procedure

Participants provided informed consent and then sat in opposite corners of the laboratory with their backs to each other, facing their display monitor. They could not see each other or each other's display. A brief 9-point calibration was carried out for each, and then task instructions were presented on screen. Two tasks were defined for the subjects. In the memory task they had to remember as many of the pictures as possible for a later test. In the search task, they had to look for a translucent X superimposed on one image, and press the mouse button that they held in one hand if they detected it. They were informed that their task could change from trial to trial, but that their partner would always be looking at the same pictures as them.

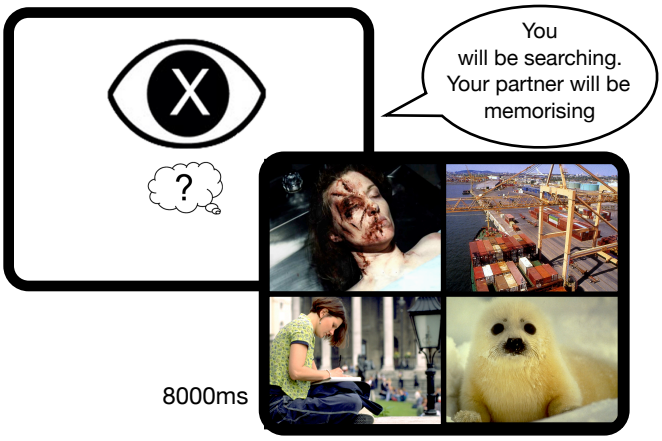


Figure 2. Trial schematic

Design

At the start of each trial, participants were told their task for the upcoming presentation. A large icon at the top of the screen showed their task (visual search or memory), and a smaller icon below that showed their partner's task (*Figure 2*). They also heard a voice say "You will be [memorising/ searching]. Your partner will be [memorising/searching]".

Participants then saw one negative, one positive and two filler images in random positions in a 2x2 grid (see *Figure 2*). They were presented for eight seconds, during which time their gaze was tracked. There was a 1 second interval, and then the instructions for the next trial began.

There were 40 trials. In half the participant was told that they were to memorise the stimuli and in half they were told that they were searching for an X. Similarly, they were told that their partner performed the memory task half the time, the search task the other half. These task conditions were counterbalanced so that half the time the participant and their partner were doing the same task, half a different task. On eight trials (spread evenly across conditions), an X appeared at a random location on one of the images.

Stimuli

Images were taken from the International Affective Picture System (IAPS), a set of photographs that have been extensively normed on a range of attributes (Lang, Bradley & Cuthbert, 2005). We chose 40 negative items with valence ratings from 1.6 to 2.4 and a mean of 2, 40 positive items from 7.6 to 8.3 and a mean of 8, and 80 filler items from 4.8 to 5.2 and a mean of 5. For each trial, stimuli were chosen at random from these categories.

Apparatus

The stimuli were presented on 19" LCD screen at a distance of approximately 60cm. Beneath each display was a *Bobax3000* remote eye tracker that sampled fixations at 100 Hz. iMac computers behind a partition presented the stimuli, calculated gaze position, and collected the data.

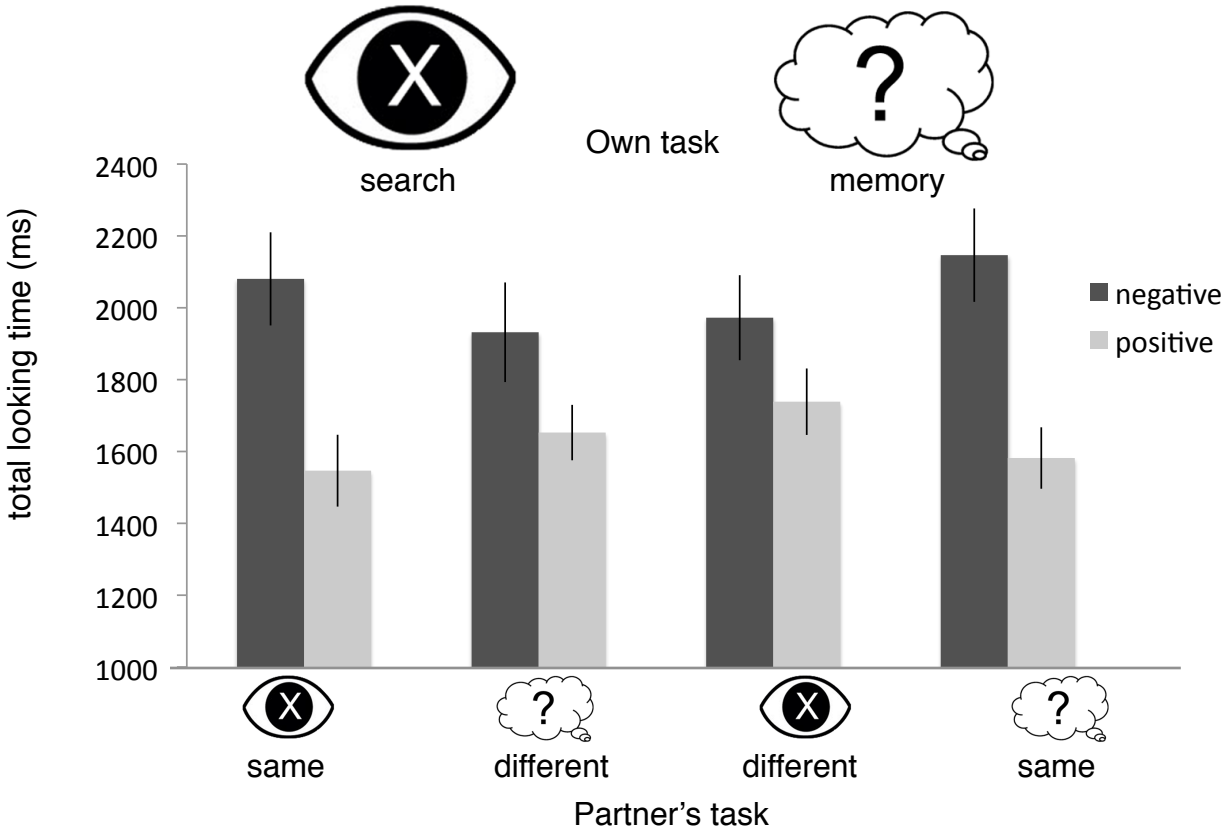


Figure 3. Looking times showed a significant interaction between valence and whether or not the participant's partner was believed to be doing the same or a different task

Results

Participants looked more towards the negative images when they believed that their partner was doing the same task as them, regardless of what the task was. We did not analyse the 20% of trials when there was an X present, as X and participants' responses to it would interfere with how they allocated their attention to each image. We calculated the total amount of time spent looking at the critical negative and positive images on trials where there was not X present. A 2 (valence: negative/positive) x 2 (own task: memory/ search) x 2 (other's task: same/different) ANOVA was performed, and the means for each cell are displayed in *Figure 3*. There was a significant two way interaction between valence and other's task ($F(1,27)=10.08$, $p=.004$). Post hoc tests show that the difference between positive and negative images was significant when the participants believed they were doing the same task (using Tukey's at 0.01), but did not reach significance when they were doing a different task. There was also a main effect of valence ($F(1,27)=19.19$, $p<.0001$), but all other main effects and interactions were non significant (all Fs <1).

General Discussion

The effects of joint perception do not occur simply when someone believes that another person is experiencing the same stimuli as themselves. We have shown that it is necessary that they believe that the other, unseen person is engaged in the same task as themselves. This task could be to memorise the pictures, which presumably would engage processing something of the meaning of an image, or the task could just be to search for a visual feature, which requires only superficial processing: regardless, the effect of joint perception arises whenever these tasks are believed to be done together. In each case, the effect of this co-engagement is to fixate the negative images more than the positive. Below, we discuss other areas of research that throw light on joint perception, and the direction of its effects in this situation.

Joint Action

Though the standard cognitive model marginalises social context, there have been notable exceptions. Studies of situated cognition (Barsalou, Breazeal & Smith, 2007; Robbins & Ayded, 2009) show that cognition 'in the wild' is intimately linked not only to representations of the external world, but also to the cognitive processes of others. Hutchins (1995) observed the ways that navy navigators distribute cognitive processes between themselves by using external tools and representations, such as maps and notations.

Recently, experimental methods are starting to reveal the mechanisms involved in such joint action (Galantucci & Sebanz, 2009; Sebanz, Knoblich & Bekkering, 2006). Social context can modulate even the simplest of tasks. For example, in a traditional stimulus-response compatibility task, participants make a judgment about one stimulus property (color) and ignore another stimulus property (location). If there is an incompatibility between the irrelevant property and the response (such as different spatial codes) then reaction times increase (Simon, 1969). Sebanz, Knoblich & Prinz (2003) divided such a task between two people. The participants sat next to each other, and each person responded to one colour: in effect, each acting as one of the fingers of a participant in Simon's (1969) experiment. Though each person had only one response to execute, they showed an incompatibility effect when acting together. There was no incompatibility effect when performing the same single response task alone. When acting jointly, participants represented their partners' actions as if they were their own.

Joint action effects do not occur if the participant is simply sat next to another person (Tsai et al., 2006), or if that person's button pressing actions are not intentional (their finger is moved by a mechanical device). Also, if the participant is acting jointly, but with a computer program (Tsai et al., 2008) or a marionette's wooden hand (Tsai & Brass, 2007) there is not a stimulus-response incompatibility effect. Therefore, participants only form representations of another when that person's genuine, intentional actions are engaged in the same task.

Our results fill out this picture. We have shown that a participant's perceptual process is changed when they believe that another person is co-acting with them: they do not have to see the person (c.f. Tsai et al., 2008), and the 'actions' do not have to be overt behaviour. If the participant thinks that the other person is memorising or scanning the images together with them, then that mutual cognitive process will shape their gaze patterns.

Focal Images

The term 'focal image' comes from Schelling (1960) who found that people were very good at guessing what images others would find salient. Schelling realised that everyday cases of verbal reference are often ambiguous. We say, 'Hand me the fork,' in the presence of many such items, yet listeners unproblematically infer the same referent. For example, when presented with a page full of items, such as watches from a catalog, participants agreed with each other which one was most likely to be referred to as 'the watch' (Clark, Schreuder & Buttrick, 1983).

When we enter into any joint activity, such coordination is all important (Clark, 1996). When we talk, we implicitly agree upon names for novel objects (Clark & Brennan, 1991), align our spatial reference frames (Schober, 1993), use each others' syntactic structures (Branigan, Pickering & Cleland, 2000), sway our bodies in synchrony (Shockley, Santana & Fowler, 2003; Condon & Ogston, 1971) and even scratch our noses together (Chartrand & Bargh, 1999). We also coordinate our gaze patterns with each other (Richardson & Dale, 2005), taking into account the knowledge (Richardson, Dale & Kirkham, 2007) and the visual context (Richardson, Dale & Tomlinson, 2009) that we share. Perhaps participants in our experiment, anticipating a future discussion of the stimuli, attempted to coordinate gaze patterns with their partner when they believed they were acting jointly. In other words, they looked at the pictures they thought another person would look at: the focal image.

Responses to Negative Stimuli

Our discussion so far has not touched upon one question: why is it that the effect of joint perception is sometimes to increase looks to the negative pictures, and sometimes to the positive images? It seems plausible that participants who thought that they were being compared to each other might want to look equally at the positive and negative images, since they may feel that ogling a disturbing image might not reflect well upon them. However why is it that in the collaborative memory task and the joint visual search tasks, the participants looking together tend to look at the negative images?

We are generally very responsive to unpleasant or threatening things. Negative images are considered more potent than equivalently-valenced positive images, so much so that when combinations of equivalent positively and negatively valenced stimuli are presented simultaneously participants rate the overall set as unpleasant (for reviews, see Baumeister et al., 2001; Lewicka, Czapinski & Peeters, 1992; Rozin & Royzman, 2004; Skowronski & Charlston, 1989). Negative stimuli are likely to receive attention more quickly (Norris et al., 2004, Smith et al., 2003) and for longer (Hajcak & Olvet, 2008). But why might this bias towards negative images be amplified during joint perception?

Emotion and Social Interaction

When people collaborate in groups, they tend to align with the group emotion (Barsade, 1998; Hatfield, Cacioppo & Rapson, 1993; Wageman, 1995). That emotion arises from the majority's personal disposition for positive or negative mood states (George, 1990). Since, as we've seen, negative stimuli are usually attended to more by individuals, when they cooperate together this would serve to amplify the negativity bias (Taylor, 1991). Affect can influence behaviour without necessarily having to personally experience the emotion (Winkielman, Berridge & Wilbarger, 2005). In this light, our joint perception phenomenon could be seen as a form of minimal, imagined cooperation that is sufficient to produce an alignment of group emotional biases.

Conclusion

How we move our eyes is swayed by a belief that others are looking at the same scene and thinking the same thing. These results broaden the notion of joint action to include perceptual processes, unseen collaborators and mental actions such as remembering and visual search. They also suggest a possible experiment to perform at a poster session. Sidle up to another conference attendee gazing over the results of an experiment. If our results generalise, a slight cough will alert them to your presence, engage their feeling of joint perception and perhaps sway their gaze towards more negative aspects of the poster, demonstrating that an effect of social context can even be found at a cognitive science conference.

Acknowledgments

We are grateful to Merrit Hoover, Arezou Ghane and Natasha Eapen for help in designing the experiments, running subjects and for many insightful discussions.

References

Allport, G.W. (1954/1979). *The nature of prejudice*. Cambridge, MA: Perseus Books.

Barsade, S.G. (2002). The ripple effect: Emotional contagion and its influence on group behaviour. *Administrative Science Quarterly, 47*(4), 644-675.

Barsalou, L.W., Breazeal, C., & Smith, L.B. (2007). Cognition as coordinated non-cognition. *Cognitive Processing*, 8, 79-91.

Baumeister, R.F., Bratlavsky, E., Finkenauer, C., & Vohs, K.D. (2001). Bad is stronger than good, *Review of General Psychology, 5*(4), 323-370.

Birmingham, E., Bischof, W.F., & Kingstone, A. (2008a). Gaze selection in complex social scenes. *Visual Cognition, 16*(2/3), 341-355.

Birmingham, E., Bischof, W.F., & Kingstone, A. (2008b). Social attention and real world scenes: The roles of action, competition, and social content. *Quarterly Journal of Experimental Psychology, 61*(7), 986-998.

Birmingham, E., Bischof, W.F., & Kingstone, A. (2009). Get real! Resolving the debate about equivalent social stimuli. *Visual Cognition, 17*(6), 904-924.

Branigan, H. P., Pickering, M. J., & Cleland, A. A. (2000). Syntactic coordination in dialogue, *Cognition*, 75, B13-B25.

Cacioppo, J.T., Visser, P.S.& Pickett C.L. (Eds.) (2005). *Social neuroscience: People thinking about thinking people*. Cambridge, MA: The MIT press.

Chartrand, T. L., & Bargh, J. A. (1999). The chameleon effect: The perception-behavior link and social interaction. *Journal of Personality and Social Psychology,* 76, 893-910.

Clark, H.H. (1996). *Being there: Putting brain, body, and the world together again.* Cambridge: MIT Press.

Clark, H.H., & Brennan, S.E. (1991). Grounding in communication. In L.B. Resnick, J.M. Levine, & S.D.Teasley (Eds.), *Perspectives on Socially Shared Cognition*. Washington, DC: American Psychological Association

Clark, H.H., Schreuder, R. & Buttrick, S., (1983). Common ground and the understanding of demonstrative reference. *Journal of Verbal Learning and Verbal Behavior,* 22, 245-258.

Condon, W., & Ogston, W. (1971). Speech and body motion synchrony of the speaker-hearer. In D. Horton & J. Jenkins (Eds.), *The Perception of Language*. Columbus, OH: Charles E. Merrill.

Fridlund, A.J., (1991). Sociality of Solitary Smiling: Potentiation by an Implicit Audience. *Journal of Personality and Social Psychology*, 60, 229-240.

Frischen, A., Bayliss, A.P., & Tipper, S.P. (2007). Gaze cueing of attention: Visual attention, social cognition, and

individual differences. *Psychological Bulletin, 133*(4), 694-724.
Galantucci, B., & Sebanz, N. (2009). Joint action: Current perspectives. *Topics in Cognitive Science*, 1, 255-259.
George, J.M. (1990). Personality, affect, and behavior in groups. *Journal of Applied Psychology, 75*, 107-116.
Greene, D.J., Mooshagian, E., Kaplan, J.T., Zaidel, E., & Iacoboni, M. (2009). The neural correlates of social attention: Automatic orienting to social and nonsocial cues. *Psychological Research, 73*, 499-511.
Hajcak, G., & Olvet, D.M. (2008). The persistence of attention to emotion: Brain potentials during and after picture presentation. *Emotion, 8*(2), 250-255.
Hatfield, E., Cacioppo, J.T., & Rapson, R.L. (1993). Emotional contagion, *Current Directions in Psychological Science. 2*(3), 96-99.
Itier, R.J, & Batty, M. (2009). Neural bases of eye and gaze processing: The core of social cognition. *Neuroscience and Biobehavioral Reviews, 33,* 843-863.
Knoblich, G., & Sebanz, N. (2006). The social nature of perception and action. *Current Directions in Psychological Science, 15*(3), 99-104.
Lang, P.J., Bradley, M.M., & Cuthbert, B.N. (2005). *International affective picture system (IAPS): Digitized photographs, instruction manual, and affective ratings* (Tech. Rep. A-6). Gainesville: University of Florida, Center for Research in Psychophysiology
Lewicka, M., Czapinsky, J., & Peeters, G. (1992). Positive-negative asymmetry or 'When the heart needs a reason'. *European Journal of Social Psychology*, 22, 425-434.
Norris, C.J., Chen, E.E., Zhu, D.C., Small, S.L., & Cacioppo, J.T. (2004). The interaction of social and emotional processes in the brain. *Journal of Cognitive Neuroscience, 16*(10), 1818-1829.
Richardson, D.C & Dale, R. (2005). Looking to understand: The coupling between speakers' and listeners' eye movements and its relationship to discourse comprehension. *Cognitive Science,* 29, 1045–1060.
Richardson, D.C, Hoover, M.A. & Ghane, A. (2008). Joint perception: gaze and the presence of others. *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 309-314). Austin, TX: Cognitive Science Society.
Richardson, D.C., Dale, R., & Kirkham, N.Z. (2007). The art of conversation is coordination: Common ground and the coupling of eye movements during dialogue. *Psychological Science, 18(5), 407-413.*
Richardson, D.C., Dale, R., & Tomlinson, J.M. (2009). Conversation, gaze coordination, and beliefs about visual context. *Cognitive Science, 33*(8), 1468-1482.
Richardson, D.C., Hoover, M.A. Ghane, A. Eapen, N. & Tan, J. (2009). Joint perception across tasks: gaze and social cognition. *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 66-72). Austin, TX: Cognitive Science Society
Ristic, J., Friesen, C.K., & Kingstone, A. (2002). Are eyes special? It depends on how you look at it. *Psychonomic Bulletin & Review, 9*(3), 501-513.
Rozin, P., & Royzman, E.B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review, 5*(4), 296-320.
Schelling, T. C. (1960). *The Strategy of Conflict*, Cambridge, Mass.: Harvard University Press.
Schober, M.F. (1993). Spatial perspective-taking in conversation. *Cognition, 47*(1), 1-24.
Sebanz, N., Bekkering, H., & Knoblich, G. (2006). Joint action: Bodies and minds moving together. *TRENDS in Cognitive Sciences*, *10*(2), 70-76.
Sebanz, N., Knoblich, G., & Prinz, W. (2003). Representing others' actions: Just like one's own? *Cognition, 88*(3), B11-B21
Senju, A., & Johnson, M.H. (2009). Atypical eye contact in autism: Models, mechanisms and development. *Neuroscience and Biobehavioral Reviews, 33*(8), 1204-1214.
Shockley, K., Santana, M-V., & Fowler, C.A. (2003). Mutual interpersonal postural constraints are involved in cooperative conversation. *Journal of Experimental Psychology: Human Perception and Performance, 29*(2), 326-332
Simon, J.R. (1969). Reactions toward the source of the stimulation. *Journal of Experimental Psychology, 81*(1), 174-176.
Skowronski, J.J., & Carlston, D.E. (1989). Negativity and extreme biases in impression formation: A review of explanations. *Psychological Bulletin, 105*(1), 131-142.
Smith, N.K., Cacioppo, J.T., Larsen, J.T., & Chartrand, T.L. (2003). May I have your attention, please: Electrocortical responses to positive and negative stimuli. *Neuropsychologia, 41*, 171-183.
Taylor, S.E. (1991). Asymmetrical effects of positive and negative events: The mobilization-minimization hypothesis. *Psychological Bulletin, 110*(1), 67-85.
Tsai, C-C., Kuo, W-J., Hung, D.L., & Tzeng, O. J-L. (2008). Action co-representation is tuned to other humans. *Journal of Cognitive Neuroscience, 20*(11), 2015-2024.
Tsai, C-C., Kuo, W-J., Jing, J-T., Hung, D.L., & Tzeng, O. J-L. (2006). A common coding framework in self-other interaction: Evidence from joint action task. *Experimental Brain Research, 175*, 353-362.
Tsai, C.-C., & Brass, M. (2007). Does the human motor system simulate Pinocchio's actions? Co-acting with a human hand versus a wooden hand in a dyadic interaction. *Psychological Science,* 18(12), 1058-1062.
Wageman, R. (1995). Interdependence and group effectiveness. *Administrative Science Quarterly, 40*, 145-180.
Winkielman, P., Berridge, K.C., & Wilbarger, J.L. (2005). Unconscious affective reactions to masked happy versus angry faces influence consumption behavior and judgments of value. *Personality and Social Psychology Bulletin*, *31*(1), 121-135.
Robbins, P. & Aydede, M. (Eds.) (in press). *The Cambridge Handbook of Situated Cognition.* Cambridge, UK: Cambridge University Press.
Hutchins, E., (1995). *Cognition in the Wild.* Cambridge, MA: MIT Press.

Teaching Students Self-Assessment and Task-Selection Skills with Video-Based Modeling Examples

Tamara van Gog (vangog@fsw.eur.nl)
Institute of Psychology, Erasmus University Rotterdam
P.O. Box 1738, 3000 DR Rotterdam, The Netherlands

Danny Kostons (danny.kostons@ou.nl)
Centre for Learning Sciences and Technologies, Open University of The Netherlands
P.O. Box 2960, 6401 DL Heerlen, The Netherlands

Fred Paas (paas@fsw.eur.nl)
Institute of Psychology, Erasmus University Rotterdam
P.O. Box 1738, 3000 DR Rotterdam, The Netherlands

Abstract

For self-regulated learning to be effective, students need to be able to accurately assess their own performance on a learning task, and to select an appropriate new learning task in response to that self-assessment. This study investigated the use of video-based modeling examples to teach self-assessment and task-selection skills. Students in both the experimental and control condition observed the model performing a problem solving task; students in the experimental condition additionally observed the model engaging in self-assessment and task selection. Results show that students in both conditions acquired problem-solving skills from the examples, as indicated by a substantial pretest to posttest knowledge gain. Moreover, students in the experimental condition also acquired self-assessment and task-selection skills from the examples: they demonstrated higher self-assessment and task-selection accuracy on the posttest than students in the control condition.

Keywords: Example-based learning; self-assessment; task selection; self-regulated learning.

The Role of Self-Assessment and Task-Selection Skills in Self-Regulated Learning

A major aim of many contemporary educational programs is to foster students' self-regulation skills. It is often assumed that this aim can be achieved by providing learners with the opportunity to self-regulate their learning processes. In the Netherlands, for example, a nationwide innovation was implemented in secondary education in 1999 that relies heavily on self-regulated learning (i.e., the 'study house'; http://www.minocw.nl/english/education/293/Secondary-education.html). Self-regulated learning is also assumed to result in personalized learning trajectories, in which instruction is adaptive to the individual student's needs. Such personalized instruction is expected to enhance students' motivation and learning outcomes compared to non-adaptive, fixed instruction that is the same for all students.

Unfortunately, there is little evidence for both assumptions. First of all, research has shown that students do not acquire self-regulation skills merely by engaging in self-regulated learning, rather, they need additional training or instructional support (e.g., Azevedo & Cromley, 2004; Van den Boom, Paas, Van Merriënboer, & Van Gog, 2004). Secondly, although the assumption is correct that adaptive, personalized instruction can foster learning compared to non-adaptive instruction (e.g., Camp, Paas, Rikers, & Van Merriënboer, 2001; Salden, Paas, Broers, & Van Merriënboer, 2004), it is questionable whether self-regulated learning actually results in adaptivity to students' needs.

In adaptive instructional systems, learning tasks are chosen for each individual student based on an assessment of their current level of knowledge and skill (based on several aspects of students' performance, e.g., Anderson, Corbett, Koedinger, & Pelletier, 1995; Koedinger, Anderson, Hadley, & Mark, 1997; or on a combination of their performance and invested mental effort, e.g., Camp et al., 2001; Corbalan, Kester, & Van Merriënboer, 2008; Kalyuga, 2006; Salden et al., 2004). The assessment of performance and the selection of an appropriate new learning task (i.e., based on that assessment) is conducted by the system. For self-regulated learning to be equally adaptive and effective, students themselves should be able to accurately assess their own performance and to recognize what an appropriate new task would be. Unfortunately, there is quite some evidence that students, and especially novices who lack prior knowledge of the learning tasks, are not very accurate self-assessors. Humans seem prone to several biases that affect accuracy of self-assessments (for a review, see Bjork, 1999), such as hindsight bias (i.e., once an answer or solution procedure is known, e.g., after feedback, students are more likely to think that they could have produced it themselves), or availability bias (i.e., answers that come to mind easily are not only more likely to be provided but are also more likely to be assumed to be correct). Moreover, accurate self-assessment also seems to require some domain expertise (Dunning, Johnson, Erlinger, & Kruger, 2003). Individuals with higher levels of prior knowledge are more accurate self-assessors, presumably because their experience not only provides them with more

task knowledge, but also with more knowledge of the criteria and standards that good performance should meet (Dunning et al., 2003). In addition, their experience also lowers the cognitive load imposed by the task, allowing them to devote more cognitive resources to monitoring their task performance, which likely provides them with a more accurate memory representation on which to base their assessment (Van Gog & Paas, 2009).

Support for our assumption that novice students' lack of self-assessment skills leads to ineffective self-regulated learning, comes from studies that have shown that providing novice students with control over their learning process may have beneficial effects on their motivation or involvement, but often has detrimental effects on learning outcomes (see e.g., Azevedo, Moos, Greene, Winters, & Cromley, 2008; Niemic, Sikorski, & Walberg, 1996). When positive effects on learning outcomes are found, this tends to be mostly for students with higher levels of prior knowledge in the domain (e.g., Niemiec et al., 1996; Moos & Azevedo, 2008), who, as mentioned above, are also likely to be more accurate self-assessors. In addition, Kostons, Van Gog, and Paas (2010) investigated differences in self-assessment accuracy between secondary education students who differed in the amount of knowledge gained from studying in a learner-controlled instructional environment that contained heredity problems with varying levels of support at different levels of complexity. They found that the students who had gained more knowledge, had also more accurately assessed their own performance during learning.

Without accurate self-assessment, selecting an appropriate new learning task will also be very difficult. Given the central role that self-assessment and task-selection skills seem to play in self-regulated learning, an important question is whether we can teach novice students to become more accurate self-assessors and task selectors. We decided to investigate this question, using modeling examples to teach those skills.

Learning from Examples

Learning from examples is known to be a highly effective instructional strategy. Research inspired by cognitive theories such as ACT-R (Anderson, 1993) or Cognitive Load Theory (Sweller, Van Merriënboer, & Paas, 1998) has extensively investigated the effects on learning of instruction consisting of studying *worked examples*, which provide students with a written worked-out didactical solution to a problem. These studies have consistently shown that for novices, studying worked examples is more effective and/or more efficient for learning (i.e., equal or higher learning outcomes attained with lower or equal investment of time and/or effort) than (tutored) problem solving, which is known as the 'worked example effect' (Sweller et al., 1998; for further reviews, see Atkinson, Derry, Renkl, & Wortham, 2000). Studies on the worked example effect have mainly used highly structured cognitive tasks, such as algebra (e.g., Cooper & Sweller, 1987; Sweller & Cooper, 1985), statistics (e.g., Paas, 1992), geometry (e.g., Paas & Van Merriënboer, 1994), or physics (e.g., Van Gog, Paas, & Van Merriënboer, 2006), although recent studies have shown the same effect with less structured tasks such as learning to recognize designer styles in art education (Rourke & Sweller, 2009).

Research inspired by Social Learning Theory (Bandura, 1986) has mostly focused on *modeling*, that is, learning by observing another person (the model) perform a task. Models can be either adults (e.g., Schunk, 1981) or peers (e.g., Braaksma, Rijlaarsdam, & Van den Bergh, 2002; Schunk & Hanson, 1985), and they can behave didactically or naturally (i.e., possibly skipping steps, or making and/or correcting errors). Moreover, modeling examples can consist of a video in which the model is visible (e.g., Braaksma et al., 2002), a video consisting of a screen capture of the model's computer screen in which the model is not visible (e.g., McLaren, Lim, & Koedinger, 2008; Van Gog, Jarodzka, Scheiter, Gerjets, & Paas, 2009), or an animation in which the model is represented by a pedagogical agent (e.g., Atkinson, 2002; Wouters, Paas, & Van Merriënboer, 2009). Like worked examples, modeling examples have also been used to teach highly structured cognitive tasks such as math (e.g., Schunk, 1981) or chemistry (e.g., McLaren et al., 2008), but they have also been widely applied with less structured tasks such as writing (e.g., Braaksma et al., 2002; Zimmerman & Kitsantas, 2002). In addition, they have been used for teaching *meta*cognitive skills such as self-regulation (e.g., Kitsantas, Zimmerman, & Cleary, 2000; Zimmerman & Kitsantas, 2002). For a more in-depth review of research on worked examples and modeling examples, see Van Gog and Rummel (in press).

This study investigated whether video-based modeling examples consisting of screen-recordings could be successfully applied for teaching secondary education students self-assessment and task-selection skills.

Method

Participants and Design

Participants were 39 Dutch secondary education students (age $M = 15.08$, $SD = 0.48$; 26 female) in the fourth year of pre-university education (the highest level of secondary education in the Netherlands, which has a duration of six years). They were novices on the content domain of the examples (heredity problems), which had yet to be taught in the formal curriculum. Participants were randomly assigned to the experimental ($n = 20$) or control condition ($n = 19$).

Materials

Pretest and Posttest The pretest and posttest consisted of 5 paper and pencil heredity problems, at five levels of complexity (see Figure 1), presented in random order. The students were informed at what level of complexity each problem was. These heredity problems could be solved by going through the following five steps: (1) translate the

phenotypes (expression of genetic trait) described in the cover story into genotypes (a pair of upper and/or lower case letters representing genetic information); (2) put these genotypes into a hereditary diagram; (3) determine direction of reasoning and number of Punnett Squares; (4) fill in Punnett Square(s); (5) extract final solution from Punnett Square(s). The posttest problems were equivalent but not identical to the pretest problems; they had similar structural features and were of similar complexity, but the surface features (cover stories) differed. On both tests, participants were instructed to write down the steps they took to reach their solution.

Complexity Level	Support Level	Learning Tasks				
Complexity 1 - 2 generations - 1 unknown - 1 solution - Deductive	Completion 3 steps worked out	Eye color	Hair structure	Shapes cat fur	Japanese Apple tree	Depression
	Completion 2 steps worked out	Eye color	Hair structure	Sickle cell Anemia	Curve chicken beak	Guinea Pigs
	Conventional 0 steps worked out	Eye color	Hair structure	Huntington	Milk Allergy	Cleft Lip
Complexity 2 - 2 generations - 1 unknown - Multiple solutions Deductive	Completion 3 steps worked out	Eye color	Hair structure	Flower color	Widow's peak	DNA
	Completion 2 steps worked out	Eye color	Hair structure	Shapes cat fur	Albinism	Pea plant
	Conventional 0 steps worked out	Eye color	Hair structure	Tongue Curling	Japanese Apple tree	Fruit flies
Complexity 3 2 generations 1 unknown Multiple solutions - Inductive	Completion 3 steps worked out	Eye color	Hair structure	Fruit flies	Curve Chicken beak	Wolfram syndrome
	Completion 2 steps worked-out	Eye color	Hair structure	Dog tail length	Japanese Apple tree	Milk allergy
	Conventional 0 steps worked-out	Eye color	Hair structure	Freckles	Flower Color	Earlobes
Complexity 4 3 generations 1 unknown - Multiple solutions - Both ways	Completion 3 steps worked-out	Eye color	Hair structure	Albino	Shapes cat fur	Fruit flies
	Completion 2 steps worked-out	Eye color	Hair structure	Fruit flies	Tongue Curling	Flower color
	Conventional 0 steps worked-out	Eye color	Hair structure	Pea plant	Dimples	Depression
Complexity 5 - 3 generations - 2 unknowns - Multiple solutions Both ways	Completion 3 steps worked out	Eye color	Hair structure	Milk Allergy	Depression	Huntington disease
	Completion 2 steps worked-out	Eye color	Hair structure	Dog tail Length	Wolfram syndrome	Flower color
	Conventional 0 steps worked out	Eye color	Hair structure	Cystic Fibrosis	Fruit flies	Red fur

Figure 1: Overview of the task database.

Mental effort rating After each problem in the pretest and posttest, participants rated how much mental effort they invested in solving that problem on a 9-point rating scale (Paas, 1992).

(Self-)assessment After the mental effort rating, participants self-assessed their performance on a 6-point rating scale ranging from 0 (none of the five steps correct) to 5 (all steps correct). After the experiment, participants' performance was scored by the experimenter on the same scale (i.e., max. problem: 5; max. test: 25).

Task selection After self-assessment, students indicated on an overview of the task database (Figure 1) what problem they would select next. At each of five complexity levels (left column), there were three levels of support: completion problem, 3 steps worked-out (white row); completion problem, 2 steps worked-out (light gray row); conventional problem, no steps worked-out (dark gray row). At each level of support within each complexity level there were 5 tasks to choose from, which had equal structural features but different cover stories. Participants knew the complexity level of the problem they had just worked on. They did not actually get the problem they selected to work on next; test problems were the same for all students.

Modeling examples The four modeling examples consisted of a recording of the model's computer screen along with a spoken explanation by the model of what s/he was doing. The gender of the models was varied: two examples were by two different male models, and two examples were by two different female models (see Table 1). In the experimental condition, the modeling examples consisted of three "phases":

(1) Problem solving: The model performed the problem solving task. Two models worked on problems of complexity level 1, and two models worked on problems of complexity level 2 (i.e., of the five complexity levels present in the task database and in the pretest and posttest; see Table 1). The quality of the models' performance varied between the examples: one example showed a model accurately solving the problem, but in the other three examples the models made one or more errors (see Table 1). This was done to create variability in phases 2 and 3 of the examples, that is, in the model's self-assessment scores and task selections (i.e., if the model would not make any errors or would detect and correct them immediately, they would always have the highest possible self-assessment score).

Table 1: Overview of modeling example characteristics.

Example	Model	Performance	Complexity
1	Male 1	0 errors	Level 1
2	Female 1	2 errors	Level 1
3	Male 2	4 errors	Level 2
4	Female 2	1 error	Level 2

(2) Self-assessment: Following task performance, the model rated invested mental effort on the 9-point rating scale and assessed their performance on the 6-point rating scale, assigning themselves one point for each correct step. The models' self-assessment was always accurate.

(3) Task selection: Then, the model selected a new task based on a combination of the performance score and the mental effort score. The models used a table (see Figure 2) in which the relationship between performance and mental effort scores was depicted, which could be used to infer a recommended 'step size' for task selection (e.g., performance of 4 and mental effort of 3 means a step size of +2). A positive step size means a recommendation to select a more challenging task (i.e., less support or higher complexity level), a step size of 0 means repeating a comparable task (i.e., same level of support and same complexity level), and a negative step size means a recommendation to select a simpler task (i.e., higher level of support or lower level of complexity). This kind of task selection algorithm based on performance and mental effort scores has proven to lead to an effective learning path in

studies on adaptive, personalized task selection (e.g., Camp et al., 2001; Corbalan et al., 2008; Kalyuga, 2006; Salden et al., 2004). The models' task selection was always accurate.

Participants in the control condition observed only the model's problem solving (phase 1). In the time in which the participants in the experimental condition observed the model's self-assessment and task selection, participants in the control condition were instructed to indicate whether the model made any errors during task performance, and if so, what the errors were and what the correct step would have been.

Performance **4-5**	**+2**	**+1**	**0**
2-3	**+1**	**0**	**-1**
0-1	**0**	**-1**	**-2**
	1, 2, 3	**4, 5, 6**	**7, 8, 9** **Effort**

Figure 2: Determining task selection step size.

Procedure

The experiment was conducted in a computer room at the participants' school. First, all participants completed the pretest on paper. Participants were given four minutes to complete each problem, followed by one minute for assessing their performance (a previous study had shown this to be sufficient time for solving the problem; Kostons et al., 2010). Participants were not allowed to proceed to the next problem before the time was up; time was kept by the experimenter using a stopwatch. After completing the pretest, participants studied the modeling examples on the computer; each participant had a head set for listening to the model's explanations. In the experimental condition, the modeling examples showed participants the task performance, self-assessment, and task selection by the model. In the control condition, participants only observed the task performance by the model and then indicated whether errors were made and if so, what the correct step was. This part was computer-paced, participants had to view the examples in the order in which they were offered and could not pause, stop, or replay the examples. Finally, all participants completed the posttest on paper, according to a similar procedure as the pretest.

Data Analysis

Self-assessment accuracy on each posttest problem was determined by computing the absolute difference between participants' objective performance score and their self-assessment of their performance. The lower this difference, the more accurate participants' self-assessment was (i.e., 0 = 100% accurate). We did not compute or analyze self-assessment accuracy on the pretest, because participants managed to solve very few problems on that test. When one is not able to perform a task at all, it is not very difficult to assess one's own performance as 0. This would be highly accurate, but would have led to a substantial overestimation of participants' self-assessment accuracy, as it is not very indicative of self-assessment accuracy on tasks that they were –at least partly- able to solve.

Task selection accuracy on the posttest was determined by computing the absolute difference between the complexity level that would be recommended based on the objective performance assessment and the complexity level participants chose.

Results

For all analyses, a significance level of .05 was used, and Cohen's *d* is reported as a measure of effect size, with 0.2, 0.5, and 0.8 corresponding to small, medium, and large effect sizes, respectively (Cohen, 1988).

Acquisition of Problem-Solving Skills

Participants' mean performance score on the pretest was 2.08 (SD = 3.58), and on the posttest it was 14.31 (SD = 6.43), so all students acquired procedural skills for solving heredity problems from the modeling examples. A *t*-test showed no significant difference between the control condition (M = 12.05, SD = 7.12) and the experimental condition (M = 12.40, SD = 6.40) in the knowledge gain from pretest to posttest, $t(37) = 0.16$, *ns*.

Acquisition of Self-Assessment Skills

A *t*-test on the mean self-assessment accuracy scores on the posttest, showed that participants in the experimental condition were more accurate (i.e., lower score; M = 0.70, SD = 0.53) than participants in the control condition (M = 1.26, SD = .85), $t(37) = 2.51$, $p = .016$ (two-tailed), $d = 0.79$.

Acquisition of Task-Selection Skills

Data from 1 participant in the experimental condition were excluded from this analysis because of too many missing values. A *t*-test on the mean task-selection accuracy scores on the posttest, showed that participants in the experimental condition were more accurate (i.e., lower score; M = 0.81, SD = 0.60) than participants in the control condition (M = 1.21, SD = 0.54), $t(36) = 2.15$, $p = .038$ (two-tailed), d = 0.70.

Discussion

This study showed that students can not only acquire problem solving skills from studying modeling examples, but also self-assessment and task selection skills, which are considered to play an important role in the effectiveness of self-regulated learning.

We chose modeling examples as a means to teach self-assessment and task-selection skills, because research has shown that example-based learning is a powerful instructional strategy. Thus far, in educational settings, examples have mostly been used for teaching cognitive

skills, and this study adds further evidence that they are useful for teaching metacognitive skills as well (see also Kitsantas et al., 2000; Zimmerman & Kitsantas, 2002). We did not, however, compare whether teaching self-assessment and task-selection skills via modeling examples was more effective than teaching those skills in some other way (e.g., via practice after having been explained the assessment and selection 'rules', i.e., how to come to a performance assessment score and how to combine performance and mental effort scores to select a new task), so the effectiveness of examples compared to other means of teaching self-assessment and task-selection skills might be explored in future research.

Our control condition received no self-assessment and task-selection training at all, but engaged in a filler task (finding and fixing errors) which may have been relevant for the acquisition of problem solving skills (see Große & Renkl,, 2007) and which we expected to direct students' attention towards assessment of performance (of the model) to some extent. Further analysis of data from the control condition was beyond the scope of this paper but could be interesting in its own right. For example, one might expect that students with better ability to find and correct errors would have better self-assessment skills and/or would show more knowledge gain. In addition, it might be interesting to establish whether the errors made by the models had any effects on students' test performance (especially for those students who were not able to find and fix errors).

A question we cannot address based on our data that would be interesting to address in future research concerns the relationship between students' levels of task knowledge and the accuracy self-assessment and task-selection skills. Even though there was some variability in pretest scores, these were in general very low. Problem-solving skills did increase from pretest to posttest. We cannot rule out that the increase in problem-solving skills might have increased students' self-assessment and task-selection accuracy in the control condition, we only know that the training in the experimental condition led to significantly higher accuracy than attained in the control condition. A problem that occurs in trying to establish gains in assessment and task selection accuracy is that it is hard to establish the level of these skills at pretest, because –as mentioned above- it is easy to rate performance as 0 when one is not able to perform a task at all. Although this is a highly accurate self-assessment, it probably does not reflect a high level of self-assessment skill. Therefore, a design in which students have lower and higher levels of prior knowledge at the start of the experiment would be required to address this question.

Other important questions for future research in this area concern whether training either self-assessment or task-selection skill would automatically lead to improvements in the other skill or whether both need training as in our experimental condition, as well as whether acquired self-assessment and task selection skills can transfer to other tasks in the same domain or even to other domains. We assume that spontaneous transfer is not very likely or would not be very effective, as assessment criteria and standards will differ for different types of task. However, we do expect that experience with self-assessment and task selection through training in one task or domain may facilitate acquisition of those skills for other tasks or domains (i.e., transfer in the sense of preparation for or accelerated future learning; Bransford & Schwartz, 1999).

Last but certainly not least, the most important question for future research is whether students can apply the self-assessment and task selection skills they acquired from modeling examples in a self-regulated learning environment in which they are allowed to select which problems to work on. If so, one would expect training self-assessment and task-selection skills to improve learning outcomes attained as a result of self-regulated learning.

Acknowledgments

The first author was supported by a Veni grant from the Netherlands Organization for Scientific Research (NWO; 451-08-003).

References

Anderson, J. R. (1993). *Rules of the mind*. Hillsdale, NJ: Erlbaum.

Anderson, J. R., Corbett, A. T., Koedinger, K., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *The Journal of Learning Sciences, 4*, 167-207.

Atkinson, R. K. (2002). Optimizing learning from examples using animated pedagogical agents. *Journal of Educational Psychology, 94*, 416-427.

Atkinson, R. K., Derry, S. J., Renkl, A., & Wortham, D. (2000). Learning from examples: Instructional principles from the worked examples research. *Review of Educational Research, 70*, 181-214.

Azevedo, R., & Cromley, J. G. (2004). Does training on self-regulated learning facilitate students' learning with hypermedia? *Journal of Educational Psychology, 96*, 523-535.

Azevedo, R., Moos, D. C., Greene, J. A., Winters, F. I., & Cromley, J. G. (2008). Why is externally-regulated learning more effective than self-regulated learning with hypermedia? *Educational Technology Research and Development, 56*, 45-72.

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Englewood Cliffs, NJ: Prentice Hall.

Bjork, R. A. (1999). Assessing our own competence: Heuristics and illusions. In D. Gopher and A. Koriat (Eds.), *Attention and performance XVII. Cognitive regulation of performance: Interaction of theory and application* (pp. 435-459). Cambridge, MA: MITPress.

Braaksma, M. A. H., Rijlaarsdam, G., & Van den Bergh, H. (2002). Observational learning and the effects of model-observer similarity. *Journal of Educational Psychology, 94*, 405-415.

Bransford, J. D., & Schwartz, D. L. (1999). Rethinking transfer: A simple proposal with multiple implications. *Review of Research in Education*, *24*, 61-101.

Camp, G., Paas, F., Rikers, R. M. J. P., & Van Merriënboer, J. J. G. (2001). Dynamic problem selection in air traffic control training: A comparison between performance, mental effort, and mental efficiency. *Computers in Human Behavior, 17*, 575-595.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

Cooper, G., & Sweller, J. (1987). The effects of schema acquisition and rule automation on mathematical problem-solving transfer. *Journal of Educational Psychology, 79*, 347-362.

Corbalan, G., Kester, L., & Van Merriënboer, J .J. G. (2008). Selecting learning tasks: Effects of adaptation and shared control on efficiency and task involvement. *Contemporary Educational Psychology, 33*, 733-756.

Dunning, D., Johnson, K., Erlinger, J., & Kruger, J. (2003). Why people fail to recognize their own incompetence. *Current Directions in Psychological Science, 12*, 83–87.

Große, C. S., & Renkl, A. (2007). Finding and fixing errors in worked examples: Can this foster learning outcomes? *Learning and Instruction, 17*, 612-634.

Kalyuga, S. (2006). Assessment of learners' organised knowledge structures in adaptive learning environments. *Applied Cognitive Psychology, 20*, 333-342.

Kitsantas, A., Zimmerman, B. J., & Cleary, T. (2000). The role of observation and emulation in the development of athletic self-regulation. *Journal of Educational Psychology, 92*, 811–817.

Koedinger, K. R., Anderson, J. R., Hadley, W. H., & Mark, M. A. (1997). Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education, 8*, 30-43.

Kostons, D., Van Gog, T., & Paas, F. (2010). Self-assessment and task selection in learner-controlled instruction: Differences between effective and ineffective learners. *Computers & Education, 54,* 932-940.

McLaren, B. M., Lim, S., & Koedinger, K. R. (2008). When and how often should worked examples be given to students? New results and a summary of the current state of research. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 2176-2181). Austin, TX: Cognitive Science Society.

Moos, D. C., & Azevedo, R. (2008). Self-regulated learning with hypermedia: The role of prior domain knowledge. *Contemporary Educational Psychology, 33*, 270–298.

Niemiec, R. P., Sikorski, C., & Walberg, H. J. (1996). Learner-control effects: A review of reviews and a meta-analysis. *Journal of Educational Computing Research, 15*, 157–174.

Paas, F. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive load approach. *Journal of Educational Psychology, 84*, 429-434.

Paas, F., & Van Merriënboer, J. J. G. (1994). Variability of worked examples and transfer of geometrical problem-solving skills: A cognitive-load approach. *Journal of Educational Psychology, 86*, 122-133.

Rourke, A., & Sweller, J. (2009). The worked-example effect using ill-defined problems: Learning to recognize designers' styles. *Learning and Instruction, 19*, 185-199.

Salden, R. J. C. M., Paas, F., Broers, N. J., & Van Merriënboer, J. J. G. (2004). Mental effort and performance as determinants for the dynamic selection of learning tasks in Air Traffic Control training. *Instructional Science, 32*, 153-172.

Schunk, D. H. (1981). Modeling and attributional effects on children's achievement: A self-efficacy analysis. *Journal of Educational Psychology, 73*, 93-105.

Schunk, D. H., & Hanson, A. R. (1985). Peer models: Influence on children's self-efficacy and achievement. *Journal of Educational Psychology, 77*, 313-322.

Sweller, J., & Cooper, G. A. (1985). The use of worked examples as a substitute for problem solving in learning algebra. *Cognition and Instruction, 2*, 59-89.

Sweller, J., Van Merriënboer, J. J. G., & Paas, F. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10*, 251–295.

Van den Boom, G., Paas, F., Van Merriënboer, J. J. G., & Van Gog, T. (2004). Reflection prompts and tutor feedback in a web-based learning environment: Effects on students' self-regulated learning competence. *Computers in Human Behavior, 20*, 551-567.

Van Gog, T., Jarodzka, H., Scheiter, K., Gerjets, P., & Paas, F. (2009). Attention guidance during example study via the model's eye movements. *Computers in Human Behavior, 25*, 785-791.

Van Gog, T., & Paas, F. (2009). Effects of concurrent performance monitoring on cognitive load as a function of task complexity. In N. Taatgen, & H. van Rijn (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 1605-1608). Austin, TX: Cognitive Science Society.

Van Gog, T., Paas, F., & Van Merriënboer, J. J. G. (2006). Effects of process-oriented worked examples on troubleshooting transfer performance. *Learning and Instruction, 16*, 154-164.

Van Gog, T., & Rummel, N. (in press). Example-based learning: Integrating cognitive and social-cognitive research perspectives. *Educational Psychology Review*.

Wouters, P., Paas, F., & Van Merriënboer, J. J. G. (2009). Observational learning from animated models: Effects of modality and reflection on transfer. *Contemporary Educational Psychology, 34*, 1-8.

Zimmerman, B. J., & Kitsantas, A. (2002). Acquiring writing revision and self-regulatory skill through observation and emulation. *Journal of Educational Psychology, 94*, 660-668.

Confidence without Competence in the Evaluation of Scientific Claims

Andrew Shtulman (shtulman@oxy.edu)
Department of Psychology, Occidental College
1600 Campus Rd., Los Angeles, CA 91106

Abstract

Scientific entities like X-rays and black holes defy firsthand observation and everyday intuition, yet most people outside the scientific community still believe in their existence. Upon what kind of epistemic foundations do such beliefs rest? The present study explored this question by comparing students' scientific beliefs to their supernatural beliefs along four dimensions of epistemic import: confidence, perceived consensus, means of justification, and openness to revision. Participants' scientific beliefs were strongly differentiated from their supernatural beliefs along the dimensions of confidence and consensus but only weakly differentiated along the dimensions of justification and revision. Moreover, participants' confidence in both types of beliefs was predicted by their consensus estimates but not their ability to cite evidence in support of, or potentially in conflict with, those beliefs. These findings imply that students' scientific beliefs are no more epistemologically sound than their supernatural beliefs, despite self perceptions to the contrary.

Keywords: Belief; testimony; naïve epistemology; intuitive theories; science education; conceptual development

Introduction

Research in cognitive science has informed the goals and methods of science education in a number of ways. Research on intuitive theories, for example, has clarified the nature of students' pre-instructional conceptions and the process by which those conceptions may be replaced by more accurate, scientific ones (Carey, 2009; Vosniadou, 1994). Research on knowledge representation has highlighted strategies effective at promoting conceptual change in the science classroom (Ohlsson, 2009; Slotta & Chi, 2006). And research on causal inference has shed light on how our theoretical commitments influence, and are influenced by, the interpretation of empirical data (Chinn & Brewer, 2001; Schulz, Goodman, Tenenbaum, & Jenkins, 2008).

To date, such research has focused mainly on the *understanding* of scientific claims, yet an equally important issue in the realm of science education is the *acceptance* of such claims as true. What, for instance, leads a student to accept the existence of electrons given that electrons are neither observable (with the naked eye) nor intuitive (with respect to our everyday conceptions of matter)? This issue is particularly important in domains where scientific explanations compete with supernatural explanations of the same phenomena, like explanations for the origin of species or explanations for the origin of the universe.

Various attempts to articulate the difference between scientific explanations and supernatural explanations have focused on differences in evidential structure (e.g., only scientific explanations generate testable hypotheses) or evidential support (e.g., only scientific explanations are supported by observation and experimentation), yet, from the perspective of how scientific explanations are *learned*, these criteria are not particularly salient. Students of science are not, after all, practitioners of science, and it is thus unlikely that most students appreciate differences in the *derivation* of scientific and supernatural explanations when simply presented with the explanations themselves.

Indeed, the *products* of science and religion – i.e., concepts, theories, explanations, and assertions – share many commonalities even if the *practices* of science and religion do not (McCauley, 2000). Both provide frameworks for interpreting everyday observations and experiences. Both posit unobservable entities as the causes of various observable phenomena. And both extend, or even defy, early-developing intuitions about the kinds of entities that exist and the kinds of interactions those entities engage in.

Given such similarities, it is unclear how well students differentiate the epistemic status of scientific claims from that of supernatural claims. Although no studies have addressed this question directly, extensive research on students' understanding of scientific inquiry provides reason to suppose that most students are not equipped to make such a differentiation (Lederman, Abd-El-Khalick, Bell, & Schwartz, 2002; Schauble, Glaser, Duschl, Schulze, & John, 1995; Smith, Maclin, Houghton, & Hennessey, 2000). This research has shown that students typically conceive of science as problem solving, rather than inquiry, and typically conceive of experiments as means of confirming, rather than testing, one's ideas. Even individuals who hold doctorates in the humanities tend to believe (a) that scientists abide by a single, deterministic method of inquiry, (b) that scientists conduct experiments in order to prove (rather than support) their ideas, (c) that scientists deduce (rather than infer) their ideas from the data at hand, and (d) that scientists who study the same data will inevitably arrive at the same conclusions (Lederman et al., 2002).

The present study attempted to extend this literature by exploring students' epistemic commitments regarding the products of science, rather than its methods. These commitments were assessed in relation to commitments regarding beliefs with ostensibly inferior evidential support – i.e., supernatural beliefs. Four dimensions of students' epistemic commitments were measured: (1) confidence in the validity of one's beliefs, (2) perceived consensus surrounding the endorsement of one's beliefs, (3) means of justifying one's beliefs, and (4) openness to revising one's beliefs. Of primary interest was the extent to which the first dimension (confidence) was related to the other three (consensus, justification, and revision).

Method

Participants

One-hundred and forty college undergraduates participated in the study for course credit in an introductory psychology class. Approximately half were recruited from a large, urban university in the Northeastern US and half from a small, urban college in the Southwestern US. Preliminary analyses revealed no significant differences between the two groups on any of the findings reported below, so they were pooled.

In the course of the study, participants rated their belief in the existence of six scientific entities and twelve supernatural entities. Although all participants endorsed the existence of at least three scientific entities, 31 participants did not endorse the existence of at least three supernatural entities. Those 31 were excluded from the analyses presented below, as they provided no internal metric against which to compare their scientific beliefs. The final sample thus consisted of 106 participants who endorsed the existence of supernatural entities at a frequency similar to the general public (Moore, 2005; Winseman, 2004).

Procedure

Participants completed a questionnaire that probed their beliefs about black holes, electrons, evolution, fluoride, genes, X-rays, angels, fate, ghosts, God, Heaven, Hell, karma, precognition, reincarnation, Satan, souls, and telepathy. Pilot data confirmed that, given this selection, college students tended to endorse an equal number of scientific and supernatural entities (i.e., all six scientific entities and around six of the twelve supernatural ones).

For each entity, participants were asked five questions: (1) whether they currently believed in the entity's existence; (2) how confident they were of that belief (on a scale from 1 to 7); (3) how many other Americans (out of 7) hold the same belief; (4) why they believed the entity exists; and (5) what evidence might persuade them to change their mind. Responses to these questions will henceforth be referred to as "existence judgments," "consensus estimates," "belief justifications," and "belief refutations," respectively.

Participants' belief justifications and belief refutations were analyzed using coded schemes described below. These schemes were constructed via a bottom-up process in which one-third of the data were sorted into numerous, fine-grained categories closely resembling the data themselves. Those categories were then collapsed into eight "basic-level" categories, which, in turn, were collapsed into three "superordinate" categories. These coding schemes were then applied to the entire dataset by two independent coders: the author, who created the coding schemes, and a research assistant, who was instructed on how to apply the coding schemes but was not involved in their creation. Among basic-level categories, agreement between coders was 90% for belief justifications and 89% for belief refutations. Among superordinate categories, agreement was 95% for belief justifications and 94% for belief refutations. Disagreements were resolved through discussion.

Results

Existence Judgments

The first question participants answered about each entity was whether or not they believed in its existence. "Yes" responses were assigned 1 point, and "No" responses were assigned 0 points. Participants' mean existence judgments are displayed in Table 1. On average, participants endorsed the existence of 5.9 scientific items (or 98%) and 7.6 supernatural items (or 63%). Item for item, these judgments were highly similar to those obtained in national surveys of supernatural belief (e.g., Moore, 2005; Winseman, 2004).

It should be noted that subsequent analyses were conducted *only* on responses connected with entities judged existent by the responder. Responses connected with entities judged nonexistent were excluded from the dataset, as they were not directly relevant to the question of how the *acceptance* of scientific claims compares to the *acceptance* of supernatural claims. Thus, the mean confidence ratings and mean consensus estimates reported in Table 1 represent only those participants who judged the target entity existent.

Table 1: Mean existence judgments (range = 0 to 1), confidence ratings (range = 1 to 7), and consensus estimates (range = 1 to 7) for the six scientific entities (top) and twelve supernatural entities (bottom).

Item	Existence	Confidence	Consensus
X-rays	1.00	6.9	6.4
Genes	.99	6.8	6.4
Electrons	.99	6.4	5.9
Fluoride	.98	6.7	6.3
Evolution	.95	6.3	4.7
Black holes	.94	5.7	5.1
Souls	.97	6.0	5.4
God	.83	5.9	5.2
Karma	.78	5.6	4.0
Heaven	.75	5.6	5.2
Angels	.66	5.4	3.9
Fate	.66	5.4	4.5
Ghosts	.56	4.6	3.6
Hell	.52	5.1	4.7
Precognition	.51	4.9	3.0
Telepathy	.44	4.7	3.0
Satan	.41	5.8	4.5
Reincarnation	.40	4.9	3.0

Confidence Ratings

Participants rated their confidence in the existence of each entity on a scale from 1 ("no confidence") to 7 ("100% confident"). Participants tended to be highly confident, selecting "7" significantly more than any other rating (33% of all selections, binomial $p < 0.001$). They also exhibited greater confidence in their existence judgments for scientific items than those for supernatural items, as shown in Table 1. Not only was the mean rating for the scientific items

significantly greater than that for the supernatural items when averaged across entities (M = 6.5 vs. M = 5.3, $t(108)$ = 11.30, $p < 0.001$), but the mean ratings for individual scientific items were almost always greater than the mean ratings for individual supernatural items. Interestingly, the more often an entity was judged existent by the group, as a whole, the more confident any individual participant was in his/her judgment. Indeed, the correlation between mean existence judgments and mean confidence ratings (i.e., columns 2 and 3 of Table 1) was highly significant ($r(18)$ = 0.85, $p < 0.001$), implying that participants' confidence in their existence judgments may have been influenced by their perception of how frequently others would agree with them.

Consensus Estimates

After selecting a confidence rating, participants estimated the number of Americans who would agree with their existence judgment on a scale from 1 ("1 out of 7") to 7 ("7 out of 7"). Mean consensus estimates for each entity are displayed in Table 1. Similar to the confidence ratings, consensus estimates for the scientific items were, on the whole, significantly greater than those for the supernatural items (M = 5.8 vs. M = 4.3, $t(108)$ = 16.80, $p < 0.001$). Participants' mean consensus estimates were also correlated with their mean existence judgments, when compared on an item-by-item basis ($r(18)$ = 0.84, $p < 0.001$), indicating that their estimates were at least partly veridical.

Belief Justifications

Participants provided a total of 1458 justifications for their existence judgments (639 for scientific items and 819 for supernatural items). These responses were sorted into the various categories and subcategories described below.

Evidential Justifications These justifications referenced objectively verifiable facts that support the existence of the entity in question. Evidential justifications came in two forms: appeals to direct evidence and appeals to indirect evidence (which accounted for 4% and 13% of all justifications, respectively). Appeals to direct evidence described an observable property or causal effect of the target entity (e.g., evolution must exist because "there are fossils for past species that have similar characteristics to present day animals;" genes must exist because "they can be sequenced and manipulated"). Appeals to indirect evidence referenced facts about the world consistent with the existence of the target entity but *not inconsistent* with other explanations (e.g., genes must exist because "children look like their parents;" God must exist because "some force must explain the Universe"). It should be noted that while it was not possible for participants to cite direct evidence of supernatural entities, it was possible for them to cite indirect evidence, and many did. It should also be noted that evidential justifications were the only justifications that could be considered epistemologically sound, as they were the only justifications that provided a *warrant* for belief rather than merely a *reason* for belief.

Deferential Justifications These justifications referenced the source of one's belief without referencing any factual or conceptual considerations relevant to the legitimacy of that source. These justifications came in three forms: appeals to unspecified evidence, appeals to authority or instruction, and appeals to a preexisting worldview or commitment (which accounted for 15%, 13%, and 23% of all justifications, respectively). Appeals to unspecified evidence differed from appeals to direct or indirect evidence in that they lacked any description of the evidence itself (e.g., X-rays must exist because "there is scientific evidence proving their existence;" souls must exist because "aspects of the concept have been sort of proven"). Appeals to authority/instruction referenced a trusted source of information without providing any details regarding the content of that information (e.g., black holes must exist because "I trust my physics teacher Mr. Murray;" Hell must exist because "it's in the Bible"). Finally, appeals to a worldview/commitment referenced some preexisting philosophy or creed consistent with the existence of the entity in question (e.g., fluoride must exist because "I'm a chemistry major;" angels must exist because "I'm Muslim; angels exist by default"). Like appeals to authority/ instruction, appeals to a worldview/commitment contained no factual information from which the belief could be inferred by someone who did not share the participant's same cultural or educational background.

Subjective Justifications These justifications referenced considerations predicated on a participants' own experience or point of view. They included appeals to intuition or volition, appeals to a personal experience or encounter, and appeals to definitions or clarifications (which accounted for 15%, 11%, and 6% of all justifications, respectively). Appeals to intuition/volition referenced the sensibility, plausibility, or desirability of the target entity without referencing considerations that actually bear upon its existence (e.g., electrons must exist because "they make rational sense;" Heaven must exist because "I like to think that my loved ones are going there"). Appeals to experience/ encounters took the form of autobiographical events whose interpretation presupposed the existence of the entity in question (e.g., fluoride must exist because "it has been used on my teeth;" telepathy must exist because "my sister and I used to have it a lot when we played 21 questions"). Finally, appeals to definitions/clarifications were intended to refine the scope or certainty of one's belief, which, like the other two subtypes, provided no objectively persuasive reasons for belief (e.g., "I believe that organisms adapt to their environment, but not that we all come from one common being;" "I believe in the presence of those who have passed away, and I suppose this is what you would call an angel").

Justifications Frequencies by Domain The proportion of justifications that fell into each of the above categories are displayed in Table 2. Paired-samples t tests revealed that participants provided significantly more deferential

justifications for scientific items than for supernatural items ($t(108) = 2.61$, $p < 0.05$) and significantly more subjective justifications for supernatural items than for scientific items ($t(108) = 8.02$, $p < 0.001$). Participants also provided significantly more evidential justifications for scientific items than for supernatural items ($t(108) = 6.75$, $p < 0.001$), but the magnitude of this difference was small (0.16), especially considering the fact that one subtype of evidential justifications – appeals to direct evidence – could be provided only for scientific items.

In sum, participants provided similar, yet non-identical, justification profiles across the two domains of belief. Although participants provided significantly more evidential justifications for their scientific beliefs than for their supernatural beliefs, they provided relatively few evidential justifications overall. Instead, they relied predominantly on deferential justifications in both domains, appealing to a trusted source of information (e.g., "my teacher," "my textbook," "my religion," "scientists," "the Bible") rather than the information itself.

Table 2: Mean proportion of justifications in each domain representative of each justification type (+ *SE*).

Justification type	Scientific	Supernatural
Evidential	.28 (.02)	.11 (.02)
Deferential	.54 (.02)	.47 (.03)
Subjective	.18 (.01)	.42 (.03)

Belief Refutations

The final question participants answered for each entity was what evidence would persuade them to change their mind about its existence. These responses are described below.

Evidential Refutations Refutations of this nature cited substantive facts or ideas that challenged the belief in question. Two subtypes were observed: anomalous data and alternative explanation (which accounted for 7% and 13% of all refutations). Participants who cited anomalous data described findings or phenomena that, if discovered, would be inconsistent with the target entity's existence (e.g., one's belief in electrons would be challenged "if an atom was found without them;" one's belief in karma would be challenged "if bad people started experiencing good things"). Participants who cited alternative explanations described situations in which the target entity would no longer be needed to explain the phenomena it was intended to explain (e.g., one's belief in fluoride would be challenged "if a new scientific model was heavily endorsed that could explain the building blocks of life without using the elements in the periodic table;" one's belief in souls would be challenged "if science could find a way to explain why there is life at all and how individuality is created in terms of thinking and feeling"). Just as evidential justifications were the only epistemologically sound type of justification, evidential refutations were the only epistemologically sound type of refutation.

Deferential Refutations These refutations fell into the same categories as deferential justifications: appeals to unspecified evidence, appeals to authority or instruction, and appeals to a preexisting worldview or commitment (which accounted for 22%, 6%, and 3% of all refutations, respectively). Appeals to unspecified evidence acknowledged that one's belief was revisable in light of new evidence but did not specify the content of that evidence (e.g., one's belief in genes would be challenged by "scientific evidence that can prove genes do not exist;" one's belief in precognition would be challenged by "proof that it's genuinely impossible"). Appeals to testimony/education cited an informant, or group of informants, whose change of mind was sufficient to incite a personal change of mind (e.g., one's belief in X-rays would be challenged "if a bunch of scientists got together and proved they didn't exist;" one's belief in Satan would be challenged "if the Church said it did not exist"). Finally, appeals to a worldview/commitment cited the possibility of changing a fundamental belief, or system of beliefs, that would result in a change to the specific belief at hand (e.g., one's belief in evolution would be challenged by "becoming extremely religious;" one's belief in reincarnation would be challenged by "more exposure to alternative beliefs"). All three categories cohered in their privileging of information sources over the information itself.

Subjective Refutations These refutations referenced considerations relevant only to the participant. They included appeals to a personal experience or encounter, appeals to ignorance or uncertainty, and denials of the premise itself (which accounted for 9%, 10%, and 30% of all refutations, respectively). Participants who appealed to a personal experience/encounter described hypothetical events that, if experienced, would call the target entity's existence into question (e.g., one's belief in X-rays would be challenged "if I found out all the X-rays I underwent were staged;" one's belief in Hell would be challenged "if I died and wasn't punished for all my sins"). Participants who appealed to ignorance/uncertainty explicitly claimed not to know what would constitute counterevidence to the entity's existence (e.g., "I don't know if you could ever prove it or disprove it;" "it is a personal belief, so I am not sure"). Finally, some participants denied the premise that their mind could be changed altogether, asserting that target entity's existence was irrefutable (e.g., "there is no evidence that could effect my belief in fate;" "Nothing at this point can dissuade me from the idea of evolution"). While denying the possibility of counterevidence is, of course, different from identifying counterevidence contingent on one's own experience, such responses were still fundamentally subjective in that they focused on personal predilections rather than external information (evidential refutations) or information sources (deferential refutations).

Refutation Frequencies by Domain Table 3 displays the mean proportion of evidential, deferential, and subjective

justifications to total justifications in each domain. Similar to the findings regarding belief justifications, participants provided significantly more deferential refutations for scientific items than for supernatural items ($t(108) = 4.80$, $p < 0.001$) but provided significantly more subjective refutations for supernatural items than for scientific items ($t(108) = 8.06$, $p < 0.001$). Participants also provided significantly more evidential refutations for scientific items than for supernatural items ($t(108) = 4.38$, $p < 0.001$), but the magnitude of this differences was, once again, quite small (0.11). Thus, just as participants tended to cite non-evidential considerations in justifying their scientific beliefs, they tended to cite non-evidential considerations when contemplating the revisability of those beliefs.

Table 3: Mean proportion of refutations in each domain representative of each refutation type (+ *SE*).

Refutation type	Scientific	Supernatural
Evidential	.29 (.02)	.19 (.02)
Deferential	.39 (.03)	.21 (.02)
Subjective	.32 (.02)	.59 (.03)

Interrelations among the Four Indices of Belief

The analyses presented thus far indicate that the epistemic foundations of participants' scientific beliefs are not identical to those of their supernatural beliefs. As a group, participants (a) exhibited greater confidence in their scientific beliefs; (b) perceived greater consensus surrounding their scientific beliefs; (c) cited more evidence in support of their scientific beliefs; and (d) identified more counterevidence to their scientific beliefs. This pattern of results appears, on its surface, to imply that participants' scientific beliefs were more epistemologically sound than their supernatural beliefs. Nevertheless, three additional findings militate against this interpretation.

First, the degree of differentiation between participants' scientific and supernatural beliefs was much greater along some dimensions than others. Cohen's *d* for the difference between scientific and supernatural beliefs was 1.41 for the confidence ratings and 1.64 for the consensus estimates, but was only 0.61 for participants' tendency to provide evidential justifications and 0.36 for participants' tendency to provide evidential refutations. Apparently, participants' sensitivity to differences between scientific and supernatural beliefs influenced their appraisals of confidence and consensus much more than it influenced their ability (or proclivity) to cite evidential considerations relevant to those beliefs.

Second, differences in confidence were not warranted by differences in the quality of participants' justifications or refutations. This finding emerged from a hierarchical regression analysis in which participants' confidence ratings for the scientific items were first regressed against their consensus estimates (Model 1) and then regressed against their tendency to provide (a) evidential justifications and (b) evidential refutations (Model 2). While the first model explained a significant amount of the variance in participants' confidence ratings ($R^2 = 0.12$; *F-change* $(1,637) = 89.63$, $p < 0.001$), the second model did not ($R^2 = 0.13$; *F-change* $(2,635) = 1.75$, *ns*). Thus, participants' confidence in their scientific beliefs was linked to their perception of how widely those beliefs are shared but was not linked to their ability to support those beliefs with evidence.

Third, the ability to provide evidential justifications and evidential refutations was not widespread. Only 13 participants (or 12%) provided evidential justifications as their modal justification type, and only 19 participants (or 17%) provided evidential refutations as their modal refutation type. Moreover, participants' tendency to provide evidential justifications was significantly correlated with their tendency to provide evidential refutations, as shown in Table 4. These tendencies were linked not only within the same domain but across domains as well, implying that they represent a domain-general disposition to reflect upon the validity of one's beliefs, similar to those documented in the domains of argumentative reasoning (Kuhn, 1991), inferential reasoning (Stanovich & West, 1998), and modal reasoning (Shtulman, 2009).

Table 4: Correlations between evidential justifications (JUS) and evidential refutations (REF) for both scientific items (SCI) and supernatural items (SUP).

Measure	JUS_SCI	JUS_SUP	REF_SCI	REF_SUP
JUS_SCI	1.0	.26**	.44**	.25*
JUS_SUP		1.0	.20*	.23*
REF_SCI			1.0	.43**
REF_SUP				1.0

Discussion

The evidential support for scientific claims is quantitatively and qualitatively superior to that for supernatural claims, yet it is unclear whether students appreciate this difference in light of the fact that both types of claims are conveyed in similar ways (through testimony) and perform similar functions (explaining observed phenomena in terms of unobservable entities). The present study addressed this issue by comparing students' scientific beliefs to their supernatural beliefs along four dimensions of epistemic import: confidence, consensus, means of justification, and openness to revision. Although participants were almost always more confident in their scientific beliefs than their supernatural beliefs, they were rarely able to identify evidence that might bear on the validity of those beliefs, either in the form of justification or refutation. Moreover, participants' confidence was related to their perception of how likely other people would agree with their beliefs but was *not* related to their ability to cite evidential considerations relevant to those beliefs.

Two features of the data were particularly notable. First, participants' modal form of justification was deference to the opinions and conclusions of others. That is, participants

were more likely to reference the *proximal* source of their beliefs (i.e., the testimony of an accepted authority or the tenets of an accepted worldview) than to reference its *distal* source (i.e., reasons for accepting the testimony/tenets as true), both for scientific beliefs and supernatural beliefs. Although it could be argued that deference to "more knowledgeable others" is a generally rational course of action (Keil, Stein, Webb, Billings, & Rozenblit, 2008), this claim is undermined, at least in the present study, by the fact that participants deferred to unsubstantiated sources of information (i.e., those propounding supernatural claims) as often as they deferred to substantiated ones (i.e. those propounding scientific claims). Moreover, participants who provided deferential justifications for their scientific beliefs, rather than evidential ones, also tended to claim that these beliefs were indefeasible, which clearly indicates a non-rational view of the nature of science. Indeed, the majority of participants (55%) denied that *anything* could dissuade them of the existence of at least one scientific entity.

Second, the findings obtained here with adults who had had multiple years of science instruction strongly mirror those obtained by Harris, Pasquini, Duke, Asscher, & Pons (2006) with individuals who had had little to no science instruction: 5- to 6-year-old children. In that study, children not only endorsed the existence of scientific entities, like germs and oxygen, more often than they endorsed the existence of supernatural entities, like God and the Tooth Fairy, but also claimed that more people, in general, believe in the existence of the former than the latter. They did *not*, however, provide different types of justifications for their judgments. Instead, they tended to appeal to generalizations that presupposed the target entity's existence in both cases (e.g., germs exist because "animals can have germs;" the Tooth Fairy exists because "she visits you when you lose a tooth"). While it is unclear how the justification categories used in Harris et al. (2006) relate to those used in the present study, it is telling that even young children appear to be more sensitive to the amount of consensus surrounding various extraordinary claims than to the conceptual and/or evidential status of those claims.

Taken together, these findings imply that students' understanding of science as a body of knowledge is not much better than their understanding of science as a method of inquiry (Lederman et al., 2002; Schauble et al., 1995; Smith et al., 2000). Just as students conceive of science as problem solving rather than inquiry, they justify their scientific beliefs with appeals to intuition and authority rather than evidence. And just as students think that scientists are in the business of "proving their ideas true," they think that certain scientific entities have been proven to exist beyond a shadow of doubt. These findings not only complement existing findings on students' scientific epistemologies but also point to the possibility that misconceptions about the *process* of science may actually be responsible for misconceptions about the *products* of science. Still, the question of whether, and how, such misconceptions are related awaits further research.

References

Carey, S. (2009). *The origin of concepts*. New York: Oxford University Press.

Chinn, C. A., & Brewer, W. F. (2001). Models of data: A theory of how people evaluate data. *Cognition and Instruction, 19*, 323-393.

Harris, P. L., Pasquini, E., Duke, S., Asscher, J., & Pons, F. (2006). Germs and angels: The role of testimony in young children's ontology. *Developmental Science, 9*, 76-96.

Keil, F. C., Stein, C., Webb, L., Billings, V. D., & Rozenblit, L. (2008). Discerning the division of cognitive labor: An emerging understanding of how knowledge is clustered in other minds. *Cognitive Science, 32*, 259-300.

Kuhn, D. (1991). *The skills of argument*. New York: Cambridge University Press.

Lederman, N. G., Abd-El-Khalick, F., Bell, R. L., & Schwartz, R. S. (2002). Views of nature of science questionnaire: Toward valid and meaningful assessment of learners' conceptions of nature of science. *Journal or Research in Science Teaching, 39*, 497-521.

McCauley, R. N. (2000). The naturalness of religion and the unnaturalness of science. In F. Keil & R. Wilson (Eds.), *Explanation and Cognition*, Cambridge: MIT Press, 61-85.

Moore, D. W. (2005). *Three in four Americans believe in paranormal*. Princeton, NJ: The Gallup Organization.

Ohlsson, S. (2009). Resubsumption: A possible mechanism for conceptual change and belief revision. *Educational Psychologist, 44*, 20-40.

Schauble, L., Glaser, R., Duschl, R. A., Schulze, S., & John, J. (1995). Students' understanding of the objectives and procedures of experimentation in the science classroom. *Journal of the Learning Sciences, 4*, 131-166.

Schulz, L. E., Goodman, N., Tenenbaum, J., & Jenkins, A. (2008). Going beyond the evidence: Preschoolers' inferences about abstract laws and anomalous data. *Cognition, 109*, 211-223.

Shtulman, A. (2009). The development of possibility judgment within and across domains. *Cognitive Development, 24*, 293-309.

Slotta, J. D., & Chi, M. T. H. (2006). Helping students understand challenging topics in science through ontology training. *Cognition and Instruction, 24(2)*, 261-289.

Smith, C. L., Maclin, D., Houghton, C., & Hennessey, M. G. (2000). Sixth-grade students' epistemologies of science: The impact of school science experiences on epistemological development. *Cognition and Instruction, 18*, 349-422.

Stanovich, K. E., & West, R. F. (1998). Individual differences in rational thought. *Journal of Experimental Psychology: General, 127*, 161-188.

Vosniadou, S. (1994). Capturing and modeling the process of conceptual change. *Learning and Instruction, 4*, 45-69.

Winseman, A. L. (2004). *Eternal destinations: Americans believe in Heaven, Hell*. Princeton, NJ: The Gallup Organization.

Individual Differences as Predictors of Learning and Engagement

Sidney D'Mello (sdmello@memphis.edu)
Claire Williams (mcwllams@memphis.edu)
Patrick Hays (dphays@memphis.edu)
Andrew Olney (aolney@memphis.edu)

Institute for Intelligent Systems, University of Memphis
Memphis, TN 38152 USA

Abstract

We investigated the possibility of predicting students' engagement and learning gains during a tutoring session from trait measures of motivation, engagement, burnout, cognitive ability, prior knowledge, and task related measures. Participants completed a multiple choice pretest, a learning session, a posttest, and a battery of individual differences tests and questionnaires. Multiple regression and exploratory factor analyses indicated that the individual differences measures yielded medium sized effects at predicting learning gains as well as engagement levels that were self-reported during the tutorial session. In general, self-reported interest in the task and confidence in learning from a computer tutor coupled with working memory capacity and attentional abilities were the major predictors of both engagement and learning.

Keywords: learning, engagement, individual differences, cognitive abilities, motivation, burnout, ITS

Introduction

There is no one-size-fits-all approach when it comes to promoting student engagement and learning. Engagement and learning are affected by a number of factors such as, the learning environment (classroom, human tutor, high stakes learning), the task (acquiring shallow facts versus obtaining a deeper conceptual understanding), and characteristics of the learners themselves (e.g., visual versus verbal learners, performance versus mastery-oriented learners) (Ackerman, Sternberg, & Glaser, 1989; Jonassen & Grabowski, 1993; Schmeck & Geisler-Brenstein, 1989). Therefore, understanding how a particular student will be engaged in and benefit from a learning session requires an analysis of how the learning environment, the task, and the characteristics of the learner (i.e. individual differences) interact and influence learning outcomes.

For a given learning activity (e.g., learning conceptual physics from a human tutor), the context and the task are fixed, however the individuals involved in the activity vary. Hence, it is important to discriminate learners that actively engage and benefit from a learning session from others who passively attend the session and do not demonstrate dramatic improvements in their knowledge levels. Consequently, individual differences research has been a long standing and valuable tradition in the fields of psychology and education (Ackerman et al., 1989; Jonassen & Grabowski, 1993). Although research efforts along this front have yielded some important insights, there is little data on how individual differences influence engagement and learning within the context of intelligent learning environments such as Intelligent Tutoring Systems (ITSs). Understanding how individual differences impact learning sessions with ITSs is important, because ITSs are emerging as effective alternatives to deliver individualized instruction to large numbers of students (Corbett, Anderson, Graesser, Koedinger, & VanLehn, 1999; Graesser, Person, & Magliano, 1995; Koedinger & Corbett, 2006).

It is generally acknowledged that all students do not benefit equally from learning sessions with ITSs (VanLehn et al., 2007). Some learners show dramatic improvements in learning gains from pre to post tests, while tutoring has a negligible impact on others. Some learners actively attend the session by carefully listening to the tutor, taking initiative by asking questions, and providing verbose responses to the tutor's questions (Graesser et al., 1995). However, other non-critical learners, socially attend the session, and are comfortable being passive information receivers rather than active problem solvers. Who are these learners? Can they be discriminated from standard individual difference measures? What are the individual differences that are predictive of engagement and learning gains? These are the questions that motivated the present study.

The present study investigated whether trait measures of individual differences in (a) motivation, engagement, and burnout, (b) cognitive abilities, and (c) task related measures, could predict state measures consisting of engagement levels and learning gains in a one-on-one tutoring session with an ITS. Our focus on trait measures of motivation, engagement, and burnout is motivated by numerous studies that have related these measures to engagement and learning (Bartels & Magun-Jackson, 2009; Pekrun, Elliot, & Maier, 2006). For example, learners with mastery-approach motivation orientations are expected to be absorbed in the learning process (i.e., more engaged) and process the material deeply, presumably resulting in higher learning gains (Elliot & McGregor, 2001). In contrast, learners with performance-approach characteristics process the material at relatively shallow levels and do not demonstrate impressive learning gains. Similarly, some research has linked trait measures of engagement and burnout to performance outcomes (Schaufeli, Martinez, Pinto, Salanova, & Bakker, 2002).

Individual differences in cognitive abilities have previously been related to a variety of outcomes, hence, we expect them to be predictive of both engagement and learning with ITSs. For example, working memory capacity has been linked to performance on tests of fluid intelligence (Yuan, Steedle, Shavelson, Alonzo, & Oppezzo, 2006). Sustained attention has been related to academic achievement in school contexts (Steinmayr, Ziegler, & Träuble, 2010). In general, existing research has empirically demonstrated interactions between affect, working memory capacity, attention, intelligence, and performance outcomes (Linnenbrink, Ryan, & Pintrich, 1999; Steinmayr et al., 2010; Vergus & Boeck, 2002; Yuan et al., 2006). Hence, the present study focused on working memory capacity, selective and sustained attention, and general intelligence as predictors of engagement and learning gains.

In addition to the motivation, engagement, burnout, and cognitive variables, there is reason to suspect that individual differences pertaining to the learning task itself might be predictive of both engagement and learning gains. For example, task interest is likely to trigger curiosity and promote engagement (Berlyne, 1978), while prior knowledge is expected to be predictor of learning gains (VanLehn et al., 2007). More interestingly, there is some recent evidence that suggests that students' confidence of learning from a computer can be a better predictor of learning gains that other variables (e.g., initial motivation, prior knowledge) (Jackson, Graesser, & McNamara, 2009).

The present study investigated whether engagement and learning gains from a tutoring session in biology could be inferred from the aforementioned individual differences measures. More specifically, our analyses focused on (a) comparing the predictive power of three banks of predictors (motivation/engagement/burnout *versus* cognitive *versus* task), (b) assessing the predictive power of combined models that simultaneously include predictors from all three banks, (c) deriving principal components from the individual difference measures, and (d) correlating the derived components with engagement and learning gains.

Methods

Participants

Participants were 90 college students (non biology majors) who participated for course credit.

Description of Learning Environment

The study used a dialogue-based ITS that tutored students on eight topics in biology (e.g., cellular respiration, mitosis, ecological succession) via natural language dialogues. The ITS was designed to mirror the pedagogical and motivational strategies of lectures delivered by expert human tutors (D'Mello et al., in review).

Participants were randomly assigned to one of three versions of the ITS. In the *dialogue* version, the tutor primarily transmitted information (68% of the time) but occasionally provided cues for acknowledgements (e.g., "Right?", "ok?"), asked comprehension gauging questions (e.g., "Do you understand?"), and prompted the student for answers (e.g., "X is a type of what?"). Alternatively, in the *monologue* version, the tutor did all the talking and the student was a passive recipient. The third version consisted of *vicarious dialogues*, where the discourse patterns were structurally similar to the dialogue condition, but with one important exception. Here, it was a virtual student, instead of the learner, that answered the tutor's comprehension gauging questions and prompts. The virtual student always provided the correct answer (via simulated keystrokes) and the human learner simply watched the interaction.

The lectures were delivered via a simple conversational interface that consisted of an animated conversational agent that delivered the content of the lectures by means of synthesized speech, a media panel that displayed images relevant to the lectures, and an input box for students to type their responses for the dialogue condition. In the vicarious dialogue condition, the virtual student's responses were provided in the input box with simulated keystrokes. The simulated keystrokes were carefully calibrated in order to mirror the temporal dynamics of actual typing (i.e., onset delay, variable interstroke delay, and delay before hitting enter key to submit response).

Dependent Measures

Engagement Measures. Participants engagement levels were tracked at multiple points in the tutorial session with the affect grid (Russell, Weiss, & Mendelsohn, 1989) and through post-lecture questionnaires. The affect grid is a validated single item affect measurement instrument consisting of a 9 × 9 (valence × arousal) grid. Valence and arousal are the primary dimensions that underlie affective experiences. The arousal dimension ranges from sleepiness to high-arousal, while the valence dimension ranges from unpleasant feelings to pleasant feelings. Participants indicate their affective state by marking an `X` at the appropriate location on the grid.

The post-lecture questionnaire asked participants to self-report their engagement levels after each lecture. There were three questions which asked the participant to rate their engagement at the beginning, middle, and end of each lecture. Participants indicated their ratings on a six-point scale ranging from *very bored* to *very engaged.*

Knowledge Tests. The knowledge tests (used to measure prior knowledge and learning gains) were 24-item multiple-choice tests with three questions for each lecture. *Prompt* questions tested participants on content for which the tutor explicitly prompted the student in the dialogue and vicarious conditions. Although there were no explicit prompts in the monologue condition, we verified that the content of the prompts was explicitly covered in the monologue. *Assertion* questions tested participants on content that the tutor explicitly asserted to the student via direct instruction. Finally, there were *deep reasoning* questions that required causal reasoning, inference, etc. rather than recall of shallow

facts. Participants completed alternate test versions for pretest and posttest that were counterbalanced across participants.

Individual Difference Measures

Motivation, Engagement, and Burnout. These measures consisted of: the Achievement Goals Questionnaire (AGQ) for motivation, the Utrecht Work Engagement Scale for Students (UWES-S) for trait engagement, and the Maslach Burnout Inventory Student Survey (MBI-SS) for burnout (Elliot & McGregor, 2001; Schaufeli et al., 2002).

The AGQ, a validated 12 item questionnaire, was used to classify participants' motivation levels as *performance-approach*, *performance-avoidance*, *mastery-approach*, and *mastery-avoidance* (Elliot & McGregor, 2001).

The UWES-S is a validated 14-item self-report measure of three dimensions of student engagement: *vigor*, *dedication*, and *absorption* (Schaufeli et al., 2002).

The MBI-SS is a validated 15-item self-report measure of three dimensions of student burnout: *exhaustion*, *cynicism*, and *professional efficacy* (Schaufeli et al., 2002).

Task Related Individual Differences. These measures consisted of pretest scores as a measure of *prior knowledge* in biology (see above) and a locally created Perceptions of Learning Biology Questionnaire (PLB). The PLB consisted of three questions that were designed to gauge participants' *interest* in learning biology, their perceived *usefulness* of learning biology, and their *confidence* that they could learn biology from a computer tutor.

Cognitive Measures. The cognitive measures consisted of: self-reported ACT or SAT scores as a measure of aptitude (these are standardized tests required for admission to universities in the US; SAT scores were converted to ACT scores in the present study), the validated Reading Span test (RSpan) to measure working memory capacity (Daneman & Carpenter, 1980), and the validated Ruff 2 and 7 Selective Attention test (Ruff 2 and 7) which measures selective and sustained attention (Ruff, Neimann, Allen, Farrow, & Wylie, 1992).

In each trial of RSpan, participants are presented with a logical or nonsensical sentence and an arbitrary letter that appears at the end of the sentence. They have to read the sentence out loud, determine if it was logical or nonsensical, and try to remember the unrelated letter. At recall, the participant typed the letters from the current set of trials in the correct order. The set sizes ranged from 2 to 5 letter strings (there were 3 trials of 2 character strings, 3 trials of 3 character strings, 4 trials of 4 character strings, and 2 trials of 5 character strings).

The measures from the RSpan include the *absolute span*, which is the highest set size (i.e., 2, 3, 4, or 5) that the participant recalled correctly, the *weighted span* (i.e., a score computed by weighting set size and items recalled), and the *total recalled* (i.e., the total number of items that the participant recalled correctly).

The Ruff 2 and 7 is a measure of selective and sustained attention (Ruff et al., 1992). It is a five-minute timed task with 20 trials (each trial is 15 seconds). For each trial, 30 targets (2's and 7's) were embedded in either a string of alphabetical capital letters (known as the automatic detection trials), or among strings of digits (known as the controlled search trials). Participants are required to spot the 2's and 7's from the distracters and click on them.

Selective attention was measured by the *automatic detection speed* and *accuracy* (the 10 letter trials) and by the *controlled search speed* and *accuracy* (the 10 digit trials). Sustained attention is measured by the *total speed* and *total accuracy* in the 20 trials.

Procedure

Participants were tested individually over a two hour session. They first completed an informed consent followed by the pretest and the Perceptions of Learning Biology questionnaire. Next, they read instructions on how to use the affect grid. On the basis of random assignment, participants then completed a tutorial session with either the monologue, dialogue, or vicarious version of the tutor. There were 30 participants in each condition. The tutoring session consisted of eight lectures that were randomly ordered for each participant. Random ordering was permissible because there was no major content overlap across lectures. Participants completed the affect grid and the post-lecture questionnaire after each lecture. They completed the posttest after the completion of all eight lectures. Finally, they completed the battery of individual difference measures after which they were fully debriefed.

Results and Discussion

We analyzed the data with multiple regression (MLR) and exploratory factor analysis techniques. The goal of the MLR analyses was to assess the predictive power of the three banks of predictors by comparing each bank separately, as well as building combined models that collectively considered all three banks. The factor analysis was used to extract principal components from the individual difference measures and to correlate the extracted components to the dependent measures (engagement and learning gains).

It is important to highlight some important points before describing the results. First, there were seven dependent variables: four learning gains measures and three engagement measures. The four learning gains measures were the corrected learning gains [(post − pre)/(1-pre)] for the prompt, assertion, and deep-reasoning questions, and an overall learning gains score (gains computed on all the items without segregating them into the different categories).

The three measures for engagement consisted of valence and arousal scores from the Affect Grid and a *composite engagement* score, which was the average engagement from the post lecture questionnaire (i.e., mean for each lecture of beginning engagement, middle engagement, and end engagement). Since the Affect Grid and post lecture questionnaires were administered eight times, once after

each lecture, an aggregate value for valence, arousal, and composite engagement was computed for each participant by averaging the scores across lectures.

It is important to emphasize that the goal of the present paper is to identify the individual difference measures that predict learning and engagement and not to assess the impact of the tutor version (i.e., dialogue, monologue, vicarious). Previous analyses have compared our dependent measures as a function of tutor type (D'Mello et al., in review). Hence, the present analyses collectively analyzed all participants without considering tutor version.

Comparing Individual Predictor Banks

The goal of this analysis was to compare the predictive power of the different banks of predictors. This was accomplished by constructing 21 multiple regression models for the seven dependent variables and the three predictor banks. There were ten motivation and engagement predictors, four task related measures, and ten cognitive predictors.

Prior to constructing the regression models, we performed a correlational analysis to identify the most diagnostic set of predictors. In particular, any predictor that marginally-significantly correlated ($p < .10$) with at least one of the seven dependent measures was preserved for the subsequent analyses. This reduced the predictor set to four motivation and engagement predictors (performance-approach, performance-avoidance, vigor, and exhaustion), three task related predictors (prior knowledge, confidence, and interest), and seven cognitive predictors (ACT; absolute span, weighted span, total recalled from the RSpan test; automatic detection speed, controlled search speed, and total speed from the Ruff 2 and 7). Multicollinearity problems among these predictor sets were diagnosed and corrected with tolerance analyses prior to constructing the regression models.

Space constraints preclude an extensive discussion of the regression models constructed by examining each predictor set independently. Hence, the current discussion is limited to comparison of the predictive power of the three feature sets (coefficients will be examined in the subsequent analysis). R^2 adj. values as a measure of goodness of fit for regression models are presented in Table 1.

It appears that on average the cognitive predictors explained 10.2% of the variance for the learning gains measures, which is consistent with a small to medium sized effect (Cohen, 1992). Variance explained by the cognitive set was also quantitatively greater than the variance explained by the motivation/engagement/burnout and task related predictors, which were on par with each other (mean R^2 adj. = .044 and .053, respectively). In contrast, the three predictor sets were equally effective in predicting the engagement measures.

Multiple Predictor Sets

The next set of regression models were constructed from the predictors that were significant in the previous set of analyses. Here, predictors from all three feature sets were simultaneously considered and the significant predictors were identified via stepwise regression.

Table 1. R^2 *adj*. for regression models

Dependent Measure	**Individual Banks**			
	M,E,B	*Task*	*Cog*	**Combined**
Learning				
Prompt	0 [c]	0 [c]	.085	.113
Assertion	.111	.027 [b]	.039	.122
Deep	0 [c]	.053	.129	.194
Overall	0 [c]	.062	.156	.149
Mean	*.028*	*.036*	*.102*	*.145*
Engagement				
Valence	.047	.030	.067	.082
Arousal	.066	.111 [b]	.061	.197
Composite	.081	.086	.136	.169
Mean	*.065*	*.076*	*.088*	*.149*

Notes. All models significant at $p < .05$ unless noted otherwise. [b] significant at $p < .10$, [c] not significant ($p > .10$). *M,E,B* = motivation, engagement, burnout. *Cog* = Cognitive.

Learning Gains. There were statistically significant models for learning gains on prompt questions, assertion questions, deep reasoning questions, as well as for total learning gains (see Table 1). On average, the combined feature sets explained .145 of the variance, which approaches a medium sized effect (Cohen, 1992) and represents a 43% improvement in the variance explained by considering the best feature set independently (i.e., cognitive features).

Turning our focus to the significant predictors of the regression models (see Table 2), it appears that students with higher working memory abilities performed well on prompt questions. Surprisingly, self-reported exhaustion scores positively predicted performance on assertion questions; this finding warrants further analysis.

Deep reasoning questions, however, were predicted by a combination of self-reported interest in learning biology as well as a high ability to sustain attention. Total learning gains, however, were predicted by a combination of working memory capacity and sustained attention, indicating that the cognitive variables are the most relevant.

Table 2. Direction (+, -) of significant predictors

Predictor	**Learning Gains**				**Engagement**		
	P	*R*	*D*	*O*	*A*	*V*	*C*
Perf-Approach						+	
Exhaustion		+					
Interest			+		+	+	+
Absolute Span	+				+		
Weighted Span						+	
Total Recalled				+			+
Total Speed			+	+			
Contrl. Srch. Speed							+[a]

Notes. [a] $p = .056$; $p < .05$ for other predictors; P, R, D = gains for prompt, assertions, and deep questions, respectively. O = overall learning gains. A, V, C = arousal, valence, and composite engagement, respectively.

Engagement. Statistically significant models were obtained for arousal, valence, and the composite engagement score. These models explained an average of 14.9% of the variance, which is consistent with a 70% improvement over the best individual model (cognitive features; see Table 1).

An examination of the significant coefficients of the regression models for engagement indicated that task interest and working memory capacity were the most diagnostic predictors (see Table 2). In particular, arousal was predicted by task interest and absolute span. Valence was predicted by task interest, weighted span, and with a performance-approach motivational orientation. Finally, composite engagement was predicted by task interest, total items recalled during the RSpan test, and controlled search speed (an important characteristic of selective attention). Simply put, being interested in the learning session and having the requisite cognitive ability (working memory span and attention) to handle the difficulties and demands of the session were the major predictors of engagement.

Factor Analysis

We analyzed the individual differences with an exploratory factor analysis (principal components analysis with varimax rotation and Kaiser normalization). The analysis was conducted on 18 out of the 24 predictors because the inclusion of some of the predictors from the RSpan and Ruff 2 and 7 tests posed problems with respect to the factorability of the data. Specifically, only the absolute span measure from the RSpan test and the total speed and total accuracy scores from the Ruff 2 and 7 test were included.

Several indicators of factorability on the model with 18 predictors indicated that the data were in fact factorable. In particular, (a) the Kaiser-Meyer-Olkin measure of sampling adequacy was .72, which is above the recommended value of .6, (b) Bartlett's test of sphericity was significant (χ^2 (153) = 287.16, $p < .05$), (c) the diagonals of the anti-image correlation matrix were all above .5, which supports the inclusion of each item in the factor analysis, and (d) the commonalities were above .3, which indicates that each item shared a degree of common variance with the other items.

The analysis yielded six components with eigen values greater than 1 that collectively accounted for 63.4% of the variance (see Table 3). It appears that Component 1, which consists of a combination of predictors from the UWESS-S, MBI-SS, and AGQ represents highly engaged, low burnout, and mastery-approach oriented learners. This component accounted for 18.9% of the variance. In contrast, Component 2 (10.3% variance) represents learners with mastery and performance-approach tendencies. Component 3 (9.5% variance) represents learners that have some prior knowledge in biology and they find it interesting and useful, while Component 4 (9.4% variance) is consistent with learners that are intelligent and have high attention abilities. Component 5 (8% variance) represents learners have a large working memory and are confident that they can learn biology from a computer tutor. Finally, Component 6 (7.2% variance) consists of learners that are absorbed, but have a performance-avoidance motivational orientation.

Our analyses proceeded by correlating the individual difference measures with the six extracted components (see Table 4). As evident from the table, components 4 and 5 are the major predictors. In particular, component 5 correlates with six out of the seven dependent measures, thereby indicating that confidence in learning biology from a computer tutor coupled with large working memory capacity and attentional ability is the individual difference component that predicts engagement and learning.

Table 3. Factor loadings

	Components					
Item	1	2	3	4	5	6
Dedication	.83					
Cynicism	-.80					
Pro Efficacy	.76					
Exhaustion	-.68					
Vigor	.61					.40
Mast Approach	.61	.50				
Mast Avoid		.73				
Perf Approach		.68				
Interest			.75			
Useful			.73			
Prior Knowledge			.60	.42		
ACT				.84		
Total Accuracy				.67		
Total Speed		.39		.40	.36	-.33
Absolute Span					.73	
Confidence			.35		.73	
Perf Avoid		.48				.71
Absorption	.36					.62

Note. Items sorted by size and values < .3 are suppressed

Table 4. Correlations between dv's and components

Dependent	**Components**					
Measure	1	2	3	4	5	6
Learning						
Prompt	-.111	-.018	-.008	.183[b]	.200[b]	-.036
Assertion	.016	-.041	.128	.030	.131	-.133
Deep	.133	-.017	.160	.302[a]	.264[a]	-.055
Total	.035	-.049	.162	.316[a]	.288[a]	-.059
Engagement						
Valence	.052	.209[b]	.259[a]	.028	.202[b]	.108
Arousal	.047	-.041	.113	.101	.252[a]	.062
Mean E.	.075	.136	.242[a]	.214[a]	.291[a]	.101

Notes. [a] significant at $p < .05$, [b] significant at $p < .10$

General Discussion

The present study investigated the possibility of predicting students' engagement and learning gains during a tutoring session with an ITS on the basis of individual differences in motivation, engagement, burnout, cognitive abilities, and task related measures. The results supported the conclusion that the cognitive factors reigned supreme when it comes to predicting learning outcomes; however, all three predictor banks were equivalent for predicting engagement. When

models were combined, the individual difference measures explained 15% of the variance in engagement and learning gains, which is consistent with a medium effect (Cohen, 1992). In general, interest in the task, confidence in learning from a computer tutor, large working memory capacity, and heightened attentional abilities were the major predictors of both engagement and learning.

Our findings have important implications for the design of ITSs that aspire to be dynamically adaptive to individual learners. These ITSs construct sophisticated student models and utilize them to tailor the instruction to each students zone of proximal development (Koedinger & Corbett, 2006). The models are usually constructed on the basis of how students' knowledge in a particular domain meshes with the material that the tutor is expected to cover. In our view, a brief pretesting session on some of the individual difference measures coupled with the existing student modeling approaches will yield more accurate models that can guide individualized instruction. How these models are utilized to heighten engagement and enhance learning gains awaits further research and technological development.

Acknowledgments

This research was supported by the Institute of Education Sciences (R305A080594). The opinions expressed are those of the authors and do not represent views of IES.

References

Ackerman, P. L., Sternberg, R. J., & Glaser, R. (Eds.). (1989). *Learning and individual differences: Advances in theory and research*. New York: Freeman.

Bartels, J. M., & Magun-Jackson, S. (2009). Approach-avoidance motivation and metacognitive self-regulation: The role of need for achievement and fear of failure. *Learning and Individual Differences, 19*(4), 459-463.

Berlyne, D. (1978). Curiosity in learning. *Motivation and Emotion, 2*, 97-175.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155-159.

Corbett, A., Anderson, J., Graesser, A., Koedinger, K., & VanLehn, K. (1999). Third generation computer tutors: Learn from or ignore human tutors? In *Proceedings of CHI Conference on Human Factors in Computing Systems* (pp. 85 - 86). New York: ACM.

D'Mello, S., Hays, P., Williams, C., Cade, W., Brown, J., & Olney, A. (in review). *Collaborative Lecturing by Human and Computer Tutors*

Daneman, M., & Carpenter, P. A. (1980). Individual difference in working memory and reading. *Journal of Verbal Learning and Verbal Behavior, 19*(450-466).

Elliot, A., & McGregor, H. (2001). A 2 x 2 achievement goal framework. *Journal of Personality and Social Psychology, 80*(3), 501-519.

Graesser, A., Person, N., & Magliano, J. (1995). Collaborative dialogue patterns in naturalistic one-to-one tutoring. *Applied Cognitive Psychology, 9*(6), 495-522.

Jackson, G. T., Graesser, A. C., & McNamara, D. (2009). What Students Expect May Have More Impact Than What They Know or Feel. In V. Dimitrova, R. Mizoguchi, B. DuBoulay & A. Graesser (Eds.), *Artificial Intelligence in Education - Building Learnning Systems That Care: from Knowledge Representation to Affective Modelling* (Vol. 200, pp. 73-80).

Jonassen, D. H., & Grabowski, B. L. (Eds.). (1993). *Handbook of Individual Difference, Learning, and Instruction*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Koedinger, K., & Corbett, A. (2006). Cognitive tutors: Technology bringing learning sciences to the classroom. In R. K. Sawyer (Ed.), *The Cambridge handbook of the learning sciences* (pp. 61-78). New York, NY: Cambridge University Press.

Linnenbrink, E. A., Ryan, A. M., & Pintrich, P. R. (1999). The role of goals and affect in working memory functioning. *Learning and Individual Differences, 11*(2), 213-230.

Pekrun, R., Elliot, A., & Maier, M. (2006). Achievement goals and discrete achievement emotions: A theoretical model and prospective test. *Journal of Educational Psychology, 98*(3), 583-597.

Ruff, R. M., Neimann, H., Allen, C. C., Farrow, C. E., & Wylie, T. (1992). The Ruff 2 and 7 Selective Attention Test: A neuropsychological application. *Perceptual and Motor Skills, 75*(1311-1319).

Russell, J. A., Weiss, A., & Mendelsohn, G. A. (1989). Affect Grid - a Single-Item Scale of Pleasure and Arousal. *Journal of Personality and Social Psychology, 57*(3), 493-502.

Schaufeli, W. B., Martinez, I. M., Pinto, A. M., Salanova, M., & Bakker, A. B. (2002). Burnout and engagement in university students - A cross-national study. *Journal of Cross-Cultural Psychology, 33*(5), 464-481.

Schmeck, R. R., & Geisler-Brenstein, E. (1989). Individual differences that affect the way that students approach learning. *Learning and Individual Differences, 1*(1), 85-124.

Steinmayr, R., Ziegler, M., & Träuble, B. (2010). Do intelligence and sustained attention interact in predicting academic achievement? *Learning and Individual Differences, 20*, 14-18.

VanLehn, K., Graesser, A. C., Jackson, G. T., Jordan, P., Olney, A., & Rose, C. P. (2007). When are tutorial dialogues more effective than reading? *Cognitive Science, 31*(1), 3-62.

Vergus, T., & Boeck, P. D. (2002). On the correlation between working memory capacity and performance on intelligence tests. *Learning and Individual Differences, 13*, 37-55.

Yuan, K., Steedle, J., Shavelson, R., Alonzo, A., & Oppezzo, M. (2006). Working memory, fluid intelligence, and science learning. *Educational Research Review, 1*, 83-98.

Do Tutors' Content Knowledge and Beliefs About Learning Influence Their Assessment of Tutees' Understanding?

Stephanie Herppich (herppich@ipn.uni-kiel.de)
Institute for Science and Mathematics Education at the University of Kiel
Olshausenstrasse 62, 24098 Kiel, Germany

Jörg Wittwer (wittwer@ipn.uni-kiel.de)
Institute for Science and Mathematics Education at the University of Kiel
Olshausenstrasse 62, 24098 Kiel, Germany

Matthias Nückles (matthias.nueckles@ezw.uni-freiburg.de)
University of Freiburg, Department of Educational Science, Instructional and School Research
Rempartstrasse 11, 79098 Freiburg, Germany

Alexander Renkl (renkl@psychologie.uni-freiburg.de)
University of Freiburg, Department of Psychology, Developmental and Educational Psychology
Engelbergerstrasse 41, 79085 Freiburg, Germany

Abstract

Research has established that tutors often have difficulty with accurately assessing a tutee's understanding. However, it is a completely open question which characteristics of tutors might affect their assessment. In an empirical study with $N = 22$ tutor-tutee dyads, we used a methodology developed by Chi, Siler, and Jeong (2004) to examine the influence of the tutors' content knowledge and beliefs about learning on their assessment accuracy. Results replicated previous research in showing that tutors overestimated a tutee's correct understanding and underestimated a tutee's incorrect understanding. In addition, more accurate assessments were positively related with tutees' learning. Finally, content knowledge had a positive impact on assessment accuracy, whereas beliefs about learning were not strongly associated with assessment accuracy. Thus, assessing a tutee's understanding seems to be important for the effectiveness of human tutoring. Moreover, the results suggest that the tutors' assessment accuracy is largely influenced by their content knowledge.

Keywords: assessment accuracy; beliefs about learning; content knowledge; human tutoring

Introduction

In educational psychology, it is widely acknowledged that for learning to be effective instruction should be tailored to a learner (Kalyuga, 2007). However, such learner-tailored instruction makes it necessary to assess a learner's individual understanding. Therefore, the ability to collect diagnostically relevant information about a learner is a central component that constitutes teaching competence (Wittwer & Renkl, 2008).

One-to-one tutoring is a form of instruction where tutors can make intensive use of the possibility of freely interacting with a tutee in order to assess a tutee's understanding. Accordingly, tutors can be expected to have a detailed "model of the student" (Putnam, 1987). However, research has shown that tutors often have difficulty with gleaning diagnostically relevant information about a tutee. For example, Chi, Siler, and Jeong (2004) examined tutoring in biology and found that tutors appeared to be relatively accurate in knowing the tutees' correct understanding but they failed to assess the tutees' incorrect understanding including their false beliefs and flawed mental models. The researchers interpreted this finding as evidence that tutors mainly used their own normative perspective as a basis for estimating what the tutees did and did not know. Similar findings were obtained by Graesser, Person, and Magliano (1995), who showed that tutors rarely diagnosed a tutee's incorrect understanding. Instead, their actions were largely based on a curriculum script that determined which skills and concepts were to be learned by the tutees (see also Putnam, 1987).

In light of these findings, the question arises as to what influences the tutors' assessment of a tutee's understanding. In this article, we shed light on two characteristics of tutors that might impact their assessment of tutees. Specifically, we look at the tutors' content knowledge and beliefs about learning. To theoretically elucidate the role of these tutor characteristics, we draw on research in the field of human tutoring and classroom teaching.

Tutors' Content Knowledge

There is widespread agreement that having a deep understanding of a subject matter is an important condition for effective teaching. Research has shown that teachers with higher content knowledge show, for example, a greater understanding of important concepts in a domain and of the relationships among them (e.g., Borko & Putnam, 1996).

However, the question as to how content knowledge specifically affects the assessment of learners in the process of teaching has not been the object of much research (cf. Baumert & Kunter, 2006). For example, Krauss et al. (2008) found that teachers with higher content knowledge tended to

have more knowledge about a learner's misconceptions. The influence of this knowledge on the teachers' practices in classroom, including their assessment of the learners, was, however, not examined.

Similarly, in the context of tutoring, little is known about the relationship between the tutors' content knowledge and their assessment of tutees. For example, Schmidt et al. (1993) found that tutors with higher content knowledge were generally more effective in promoting tutees' learning when compared to tutors with lower content knowledge. The researchers attributed this finding to the fact that tutors with more content knowledge engaged in content-related activities that helped tutees to acquire knowledge. Even so, the role of the tutors' assessment practices for the tutees' learning was not investigated in this study.

Overall, the findings suggest that tutors with higher content knowledge might assess a tutee's individual understanding more accurately than tutors with lower content knowledge. This is assumed to be because tutors with more content knowledge normally have a deeper understanding of the concepts to be learned by a tutee (Borko & Putnam, 1996). Accordingly, tutors can be expected to show a more differentiated understanding of a tutee's conceptual knowledge (Nickerson, 1999). For example, tutors with higher content knowledge might be more likely to think at a deeper level about the conceptual aspects of a tutee's comprehension difficulties (Chi, Feltovich, & Glaser, 1981). Similarly, tutors with higher content knowledge might more likely infer from a tutee's particular misunderstanding which related mis-understandings and misconceptions can occur (Person et al., 1994).

Tutors' Beliefs About Learning

Apart from the teachers' content knowledge, their beliefs about how learners learn might also influence their teaching (Borko & Putnam, 1996). These beliefs can be roughly divided into two different views of learning: a transmission view of learning and a constructivist view of learning. A transmission view of learning focuses on the contents to-be-learned and emphasizes the role of transmitting knowledge to the learner. In contrast, a constructivist view of learning places a learner's own knowledge-construction activities at the center of instruction and emphasizes the role of supporting a learner's learning.

Research has provided evidence that such beliefs have an impact on teaching and learning. For example, Staub and Stern (2002) found that teachers with a constructivist view of learning were more successful in enhancing a learner's problem solving. In addition, Dubberke et al. (2008) showed that the teachers' beliefs strongly guided their teaching practices. For example, teachers with a transmission view of learning less often engaged in activities to support the learners' knowledge acquisition than teachers with a constructivist view of learning.

Despite these findings, there is also research showing that the teachers' beliefs are not necessarily associated with their pedagogical activities observed in classroom (e.g., Leuchter et al., 2006). It can be assumed that this is because teachers might not be completely accurate in self-assessing their beliefs about learning. Another explanation is that the teachers' beliefs might be too distal to strongly shape their teaching practices.

In the context of tutoring, it is a completely open question as to how the tutors' beliefs about learning influence their assessment of tutees. In line with the findings obtained in research on classroom teaching, it can be assumed that a constructivist view of learning supports the accuracy with which tutors assess a tutee's understanding. This is because tutors with a constructivist view of learning as opposed to tutors with a transmission view of learning see tutees as being actively involved in learning. Thus, it is supposed that tutors with a constructivist view of learning provide tutees with opportunities to be active and constructive on their own. As a result, the tutors should get insights into the tutees' understanding and learning progress during the course of tutoring.

Research Questions

We present an empirical study in which we examined human tutoring in biology to shed light on the role of the tutors' content knowledge and beliefs about learning in assessing a tutee's conceptual understanding. We addressed the following research questions:

1) How accurately do tutors assess a tutee's correct understanding and a tutee's incorrect understanding?
2) Is the tutors' assessment accuracy positively associated with the tutees' learning?
3) Does the tutors' content knowledge positively influence their assessment accuracy?
4) Does the tutors' orientation towards a constructivist view of learning positively influence their assessment accuracy?

Method

Sample and Design

A total of $N = 22$ dyads of tutors and tutees participated in the empirical study. Tutors were university students of biology. Of the tutors, 18 were female and 4 were male. Their mean age was 22.64 years ($SD = 2.79$). Tutees were K-7 students from Realschulen (i.e., schools from the middle track of the German school system). Of the tutees, 9 were female and 13 were male. Their mean age was 12.64 years ($SD = 0.49$). The tutors and the tutees did not know each other before tutoring.

We examined the accuracy with which the tutors assessed a tutee's individual understanding. We also analyzed the impact of their assessment accuracy on tutees' learning. Finally, we investigated the influence of the tutors' content knowledge and beliefs about learning on their accuracy at assessing a tutee's individual understanding.

Materials

Textbook (Tutee and Tutor) In the tutoring session, the tutor and the tutee engaged in a dialogue on the basis of a passage about the human circulatory system, which was previously used by Chi et al. (2001). We adapted this passage for the present study by deleting and reformulating some sentences. Each of the remaining 59 sentences of the passage was printed on a separate sheet of paper. The sentences were presented to the tutor and the tutee in a ring binder.

Content Knowledge Test (Tutor) The test consisted of 18 multiple-choice items. Each correct answer was assigned 1 point. The test measured not only the tutors' knowledge about basic concepts to be discussed in tutoring, but also their knowledge about advanced concepts of the human circulatory system, about the relationships among these concepts, and about the relevance of these concepts for life processes. Hence, answering the test required different levels of knowledge. Accordingly, item difficulty ranged from .41 to .95 (M = .64, SD = .16).

Beliefs About Learning Questionnaire (Tutor) The questionnaire was adapted from Staub and Stern (2002). On a 4-point Likert scale ranging from 1 (= *strongly disagree*) to 4 (= *strongly agree*), the tutors indicated their agreement with 19 statements. Agreement with 9 out of the 19 statements indicated a constructivist view of learning. The agreement with the remaining 10 statements indicated a transmission view of learning. The statements indicating a transmission view of learning were reversed so that the mean agreement with a constructivist view of learning could be computed, with higher scores showing a more constructivist view of learning.

Misconceptions Test (Tutee and Tutor) The test consisted of 25 multiple-choice items that addressed concepts about the human circulatory system at the local level of propositions (cf. Chi et al., 2004). The items were adapted from tests originally developed by Sungur and Tekkaya (2003) and by Michael et al. (2002) or constructed on the basis of the literature on misconceptions of the human circulatory system (e.g., Pelaez et al., 2005). The items covered concepts about the human circulatory system that were explicitly or implicitly mentioned in the textbook. A correct answer indicated a scientifically correct understanding of the concept. Each of the incorrect answers indicated a specific type of incorrect understanding of the concept.

Drawings of the Human Circulatory System (Tutee and Tutor) On a sheet of paper, the outline of a human body was displayed. The tutees were asked to draw the blood path of the circulatory system into the human body and to explain the blood path. The explanations were audiotaped. By using this methodology, which was originally developed by Chi et al. (2004), we assessed a tutee's conceptual understanding at the global level of mental models.

To code the tutees' and the tutors' drawings and explanations of the human circulatory system, we adapted a classification scheme originally developed by Azevedo, Cromley, and Seibert (2004). On the basis of this classification scheme, the drawings were assigned to one of twelve categories. The categories reflect distinguishable types of correct and incorrect mental models with categories 0 to 9 indicating different types of incorrect mental models and with categories 10 to 11 indicating a correct mental model.

Procedure

Each tutoring session was divided into three phases: pre-test phase, tutoring phase, and post-test phase. It lasted about 3 hours.

Pre-Test Phase In the pre-test phase, the tutors completed the content knowledge test. The tutees completed the misconceptions test. In addition, the tutees were asked to draw the blood path of the human circulatory system in the outline of a human body and to explain the blood path as they knew it. Afterwards, both the tutors and the tutees individually read the passage about the human circulatory system.

Tutoring Phase The dyads of tutors and tutees read each sentence of the passage about the human circulatory system and engaged in a dialogue about each sentence. After the 33th sentence, tutoring was interrupted and the dyads were separated. The tutees were asked to draw and explain the blood path of the human circulatory system. To measure what the tutors thought that the tutees would know about the blood path, the tutors were required to draw and explain the tutees' mental model of the human circulatory system. After accomplishing this task, tutoring was continued.

Post-Test Phase After completing the tutorial dialogue, the dyads of tutors and tutees were separated again and asked to draw and explain the blood path of the human circulatory system. Afterwards, the tutees completed the misconceptions test. The tutors also received the 25 items of the misconceptions test and were asked to indicate how the tutee would answer each of the items. Finally, the tutors filled in the beliefs about learning questionnaire.

Results

The following results concerning the tutors' assessment accuracy and the tutees' learning are based on the data collected in the post-test phase.

Tutors' Assessment Accuracy

In a first step, we examined the accuracy with which the tutors assessed what the tutees did and did not know at the level of propositions (i.e., misconceptions test) and at the

level of mental models (i.e., drawings of the circulatory system).

Misconceptions Test On average, the tutees had a correct understanding of 49% (*SD* = 11%) of the concepts and an incorrect understanding of 43% (*SD* = 13%) of the concepts[1].

Generally, the tutors assumed tutees to have a correct understanding of 58% (*SD* = 12%) of the concepts and to have an incorrect understanding of 26% (*SD* = 5%) of the concepts. Hence, the tutors significantly overestimated the tutees' correct understanding of the concepts, $t(21) = -2.43$, $p = .02$, $\eta^2 = .22$ (strong effect), and significantly underestimated the tutees' incorrect understanding of the concepts, $t(21) = 6.10$, $p = .01$, $\eta^2 = .64$ (strong effect).

When we specifically looked at whether the tutors knew how the tutees would answer each of the items of the misconceptions test, we found that the tutors knew the tutees' precise answers for 43% (*SD* = 11%) of all items.

Drawings Of the tutees, 64% drew and explained an incorrect mental model, whereas 36% drew and explained a correct mental model.

The tutors assumed the tutees to have an incorrect mental model in 18% of all cases and assumed the tutees to have a correct mental model in 82% of all cases. Thus, the tutors tended to assume the tutees to have more often a correct mental model than the tutees actually had and to have less often an incorrect mental model than the tutees actually had, $\chi^2(1, N = 22) = 2.79$, $p = .09$, $\varphi = .36$ (medium effect).

When we further looked at the categories into which the drawings of the tutees and the tutors fell, we found that, on average, the tutees' mental models were assigned to category 7 (*M* = 7.36, *SD* = 3.19). The tutors' drawings of the tutees' mental models were, on average, assigned to category 10 (*M* = 10.27, *SD* = 0.88). The difference between the average category of the tutees' mental models and the average category of the tutors' drawings of the tutees' mental models (*M* = -2.91, *SD* = 3.25) was significant, $t(21) = -4.20$, $p = .01$, $\eta^2 = .46$ (strong effect). Hence, the tutors largely overestimated the tutees' understanding at the level of mental models.

Tutors' Assessment Accuracy and Tutees' Learning

In a next step, we examined the importance of the tutors' assessment accuracy for the tutees' learning. To do so, we computed the correlation between the tutors' assessment accuracy at the level of propositions and the tutees' understanding at the level of mental models. The correlation was significant, $r = .59$, $p = .01$. Hence, the tutors' assessment accuracy was substantially associated with tutees' learning.

[1]To reduce the probability of guessing the correct answer in the misconceptions test, the tutees were asked to check the option "don't know" in case of uncertainty. Thus, correct and incorrect answers do not add up to 100%.

Tutors' Content Knowledge, Beliefs About Learning, and Assessment Accuracy

In a last step, we determined the relation between the tutors' content knowledge and beliefs about learning on the one hand and their assessment accuracy on the other hand. To measure the assessment accuracy at the level of propositions, we used the number of answers that the tutors correctly assumed the tutees to give to each of the items of the misconceptions test. To measure the assessment accuracy at the level of mental models, we used the difference between the category number of a tutee's mental model and the category number of a tutor's drawing of the tutee's mental model. Content knowledge and beliefs about learning were not significantly related with each other, $r = .25$, $p = .26$.

Content Knowledge In the content knowledge test, the tutors answered, on average, 64% (*SD* = 21%) of the items correctly. The number of correctly answered items was positively and significantly correlated with the accuracy with which the tutors assessed the tutees' understanding at the level of propositions, $r = .47$, $p = .03$. It was also positively and significantly correlated with the accuracy with which the tutors assessed the tutees' understanding at the level of mental models, $r = .48$, $p = .02$. Hence, the tutors with higher content knowledge were clearly more accurate in assessing the tutees' understanding.

Beliefs About Learning When answering the beliefs about learning questionnaire, the tutors achieved a mean score of 2.76 points (*SD* = 0.44). Hence, the tutors, on average, tended to show a constructivist view of learning. The correlation between the tutors' beliefs about learning and their accuracy at assessing what tutees knew at the level of propositions just failed to reach the 10%-level of statistical significance, $r = .35$, $p = .11$. The correlation between the tutors' beliefs about learning and their accuracy at assessing what tutees knew at the level of mental models was not significant, $r = .12$, $p = .59$. Obviously, the tutors' beliefs about learning were not generally associated with the accuracy with which the tutors assessed the tutees' understanding.

Discussion

The present study examined the accuracy with which tutors assessed a tutee's understanding of the human circulatory system. We found that the tutors significantly overestimated the tutees' correct understanding of important concepts related to the human circulatory system and significantly underestimated the tutees' incorrect understanding of these concepts. A similar pattern of results was obtained when we looked at the tutors' assessments of the tutees' mental models of the human circulatory system. Again, the tutors assumed the tutees to have a more complete understanding than they actually had. Overall, our findings replicate the results of Chi et al. (2004) and suggest that tutors seriously fail to take into account a tutee's alternative understanding.

As already discussed by Chi et al. (2004), tutors appear not to carefully assess what tutees do and do not know. Instead, they seem to exhibit a bias towards imputing their own normative perspective to the tutees (Hinds, 1999; Nickerson, 1999).

However, our results also show that the accuracy with which the tutors assessed a tutee's understanding largely depended on their content knowledge. In other words, tutors with more content knowledge were more accurate in assessing a tutee's conceptual understanding both at the level of propositions and at the level of mental models. It can be argued that this is likely to be because tutors with more content knowledge assess and categorize a tutee's understanding of concepts at a deeper level (Nickerson, 1999). This might allow the tutors to discriminate a tutee's understandings and misunderstandings more accurately (Chi et al., 1981).

In addition, we found that the tutors' beliefs about learning seemed to be less important for their assessment accuracy. This finding, however, has to be interpreted with caution. In our study, nearly all tutors showed an orientation towards a constructivist view of learning. Therefore, the variance of this tutor characteristic apparently was too small to yield any significant result.

Even though we observed differences in the accuracy with which the tutors assessed a tutee's understanding, we do not know yet which assessment strategies they used to collect diagnostically relevant information about a tutee. Prior research has already provided evidence for differences in tutorial actions between more experienced tutors and less experienced tutors. For example, Cromley and Azevedo (2005) found that more experienced tutors more often engaged in cognitive scaffolding. Less experienced tutors, in contrast, more often delivered information to the tutees. Following Chi et al. (2001), it is plausible to assume that these tutorial moves might help or hinder tutors in assessing what a tutee knows. For example, when asking questions (i.e., asking for information) instead of providing explanations (i.e., generating information on one's own), tutors might have more cognitive resources left for assessing a tutee's understanding (see also Wittwer, Nückles, & Renkl, 2010). Thus, to shed light on the question which moves of tutors positively and negatively influence their assessments of tutees, we are currently analyzing the tutoring protocols collected during the tutoring sessions.

Related to this is the question how the tutors in our study adjusted their tutorial moves on the basis of their assessments. Our results show that the tutors' assessment accuracy was positively associated with the tutees' learning. This suggests that the tutors might have used their assessments of what a tutee does and does not know in order to individualize instruction. It can be conjectured that the assessments, for example, influenced the tutors in deciding to move on to the next sentence of the textbook or to ask a question in order to elicit knowledge-construction activities from a tutee. Again, our content analysis of the tutoring protocols could clarify how the tutors adapted their moves to a tutee's specific understanding.

What are the implications of our study and what are the directions for future research? First, our findings suggest that it seems to matter who serves as tutor. Obviously, tutors with higher content knowledge can more accurately assess what a particular tutee does and does not know. As a result, these tutors acquire knowledge about a tutee's knowledge which they can use to support the tutee's learning[2]. Hence the concrete effectiveness of human tutoring might vary, amongst other things, as a function of tutor characteristics such as a tutor's content knowledge and tutoring experience (Cromley & Azevedo, 2005; though tutoring has generally been shown to be effective: Cohen, Kulik, & Kulik, 1982).

Second, our study seems to indicate that, in general, tutors with lower content knowledge have more difficulty with taking into account a tutee's particular understanding. At first glance, this finding might contradict the notion that peer tutors who normally do not possess considerably more knowledge than their tutees can also be responsive to their tutees' needs. However, such responsive behavior might not primarily result from the tutors' accurate assessments of the tutees' knowledge. Instead, it can be argued that tutors in peer tutoring share with their tutees a similar understanding of the learning task and, thus, might encounter the same comprehension difficulties. As a result of this common ground (Chi et al., 2004), the tutor and the tutee are more likely to "automatically" possess a mutual understanding. Hence, peer tutors might not be required to deliberately assess a tutee's understanding at all.

Third, our results show that, on average, the tutors largely overestimated a tutee's understanding. It was assumed that this finding can be attributed to the tutors' bias to impute their own normative perspective to the tutees. Although our study suggests that having more content knowledge reduces the risk of overestimating a tutee's understanding, there might be a trade-off between the tutors' content knowledge and their assessment accuracy under some circumstances. For example, Nathan and Petrosino (2003) found that pre-service teachers with higher content knowledge had problems with correctly estimating the difficulty of mathematical problems for learners. This was assumed to be a result of the pre-service teachers' discipline-specific perspective on the mathematical problems. Accordingly, it might well be that tutors who have, due to their high content knowledge, a more discipline-oriented view of the subject matter are particularly prone to an egocentric bias. In this case, it can be expected that tutors with such knowledge are less accurate instead of more accurate in assessing a tutee's understanding.

[2]In a mediation analysis, we found that the tutors' content knowledge influenced the tutees' learning. This effect was significantly mediated by the tutors' assessment accuracy.

Acknowledgments

We would like to thank our research assistants Julian Etzel, Tatjana Scharping, Anika Schoneville, and Raoul Zimmermann for their help with many practical aspects of the project. This research was supported by grants from the German Science Foundation DFG (WI 3348/2-1).

References

Azevedo, R., Cromley, J. G., & Seibert, D. (2004). Does adaptive scaffolding facilitate students' ability to regulate their learning with hypermedia? *Contemporary Educational Psychology, 29*, 344-370.

Baumert, J., & Kunter, M. (2006). Stichwort: Professionelle Kompetenz von Lehrkräften [Keyword: Professional competence of teachers]. *Zeitschrift für Erziehungswissenschaft, 9*, 469-520.

Borko, H., & Putnam, R. (1996). Learning to teach. In D. Berliner & R. Calfee (Eds.), *Handbook of educational psychology*. New York: Macmillan.

Chi, M. T. H., Feltovich, P., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5*, 121-152.

Chi, M. T. H., Siler, S., & Jeong, H. (2004). Can tutors monitor students' understanding accurately? *Cognition and Instruction, 22*, 363-387.

Chi, M. T. H., Siler, S., Jeong, H., Yamauchi, T., & Hausmann, R. G. (2001). Learning from human tutoring. *Cognitive Science, 25*, 471-533.

Cohen, P. A., Kulik, J. A., & Kulik, C. (1982). Educational outcomes of tutoring: A meta-analysis of findings. *American Educational Research Journal, 19*, 237-248.

Cromley, J. G., & Azevedo, R. (2005). What do reading tutors do? A naturalistic study of more and less experienced tutors in reading. *Discourse Processes, 40*, 83-113.

Dubberke, T., Kunter, M., McElvany, N., Brunner, M., & Baumert, J. (2008). Lerntheoretische Überzeugungen von Mathematiklehrkräften: Einflüsse auf die Unterrichtsgestaltung und den Lernerfolg von Schülerinnen und Schülern [Beliefs of mathematics teachers: Impact on teaching practices and students' achievement]. *Zeitschrift für Pädagogische Psychologie, 3/4*, 193-206.

Graesser, A. C., Person, N. K., & Magliano, J. P. (1995). Collaborative dialogue patterns in naturalistic one-on-one tutoring. *Applied Cognitive Psychology, 9*, 495-522.

Hinds, P. J. (1999). The curse of expertise: The effects of expertise and debiasing methods on predictions of novice performance. *Journal of Experimental Psychology: Applied, 5*, 205-221.

Kalyuga, S. (2007). Expertise reversal effect and its implications for learner-tailored instruction. *Educational Psychology Review, 19*, 509-539.

Krauss, S., Brunner, M., Kunter, M., Baumert, J., Blum, W., Neubrand, M. et al. (2008). Pedagogical content knowledge and content knowledge of secondary mathematics teachers. *Journal of Educational Psychology, 100*, 716-725.

Leuchter, M., Pauli, C., Reusser, K., & Lipowsky, F. (2006). Unterrichtsbezogene Überzeugungen und handlungsleitende Kognitionen von Lehrpersonen [Teaching-related beliefs and practice-guiding cognitions of teachers]. *Zeitschrift für Erziehungswissenschaft, 9*, 562-579.

Michael, J. A., Wenderoth, M. P., Modell, H. I., Cliff, W., Horwitz, B. et al. (2002). Undergraduates' understanding of cardiovascular phenomena. *Advances in Physiology Education, 26*, 72-84.

Nathan, M. J., & Petrosino, A. J. (2003). Expert blind spot among preservice teachers. *American Educational Research Journal, 40*, 905-928.

Nickerson, R. S. (1999). How we know—and sometimes misjudge—what others know: Imputing one's own knowledge to others. *Psychological Bulletin, 125*, 737-759.

Pelaez, N. J., Boyd, D. D., Rojas, J. B., & Hoover, M. A. (2005). Prevalence of blood circulation misconceptions among prospective elementary teachers. *Advances in Physiology Education, 29*, 172-181.

Person, N. K., Graesser, A. C., Magliano, J. P., & Kreuz, R. J. (1994). Inferring what the student knows in one-to-one tutoring: The role of student questions and answers. *Learning and Individual Differences, 6*, 205-229.

Putnam, R. T. (1987). Structuring and adjusting content for students: A study of live and simulated students tutoring of addition. *American Educational Research Journal, 24*, 13-48.

Schmidt, H. G., Van der Arend, A., Moust, J. H. C., Kokx, I., & Boon, L. (1993). Influence of tutors' subject-matter expertise on student effort and achievement in problem-based learning. *Academic Medicine, 68*, 784-791.

Staub, F. C., & Stern, E. (2002). The nature of teachers' pedagogical content beliefs matters for students' achievement gains: Quasi-experimental evidence from elementary mathematics. *Journal of Educational Psychology, 94*, 344-355.

Sungur, S., & Tekkaya, C. (2003). Students' achievement in human circulatory system unit: The effect of reasoning ability and gender. *Journal of Science Education and Technology, 12*, 59-64.

Wittwer, J., Nückles, M., & Renkl, A. (2010). Using a diagnosis-based approach to individualize instructional explanations in computer-mediated communication. *Educational Psychology Review, 22*, 9-23.

Wittwer, J., & Renkl, A. (2008). Why instructional explanations often do not work: A framework for understanding the effectiveness of instructional explanations. *Educational Psychologist, 43*, 49-64.

Risk attitude in decision making: A clash of three approaches

Eldad Yechiam (yeldad@tx.technion.ac.il)
Faculty of Industrial Engineering and Management, Technion – Israel Institute of Technology
Haifa, 32000 Isreael

Eyal Ert (eert@hbs.edu)
Harvard Business School, Baker Library, 447
Boston, MA 02163, USA

Abstract

We evaluate the consistency of different constructs affecting risk attitude in individuals' experiential decisions across different levels of risk. Three major views concerning the psychological constructs that underlie risk attitude are contrasted. The first is the classical economic approach which views risk as the sensitivity to differences in variance. The second is the latent components approach suggesting the importance of sensitivity to losses and diminishing sensitivity to marginal increases in payoffs. The third approach, risk acceptance, relates to the willingness to accept probable outcomes over certainty. The results of three studies indicate that: (1) Individuals do not exhibit consistency in their sensitivity to variance (2) Across domains individuals are consistent when deciding between constant versus probable outcomes, refuting the prediction based on diminishing sensitivity. (3) Risk acceptance entails different psychological constructs when the decision involves co-occurring gains and losses. The results are modeled with a quantitative index of subjective risk.

Keywords: risk; choice; individual differences; cognitive style.

Introduction

A dominant view of the psychological construct of sensitivity to risk (or risk attitude) suggests that it in fact represents the consistent sensitivity to different latent components. The most prominent example of this idea is prospect theory (Kahneman & Tversky, 1979) which explains contingent risk taking in different domains by the assumption that subjective values (or utilities) are based on relative judgments reflecting the effect of two main constructs: (a) Loss aversion – the idea that losses loom larger than equivalent gains, and (b) Diminishing sensitivity to marginal changes in payoff – the assertion that the subjective impact of a change in the absolute payoff decreases with the distance from zero. Recent cognitive models of individual choice in decisions from experience (see Hertwig et al., 2004) have adopted this view by implementing these factors as two core components of subjective utility: (a) loss sensitivity – the assumption that individuals weigh gains and losses in a consistent fashion (e.g., Busemeyer & Stout, 2002; Worthy, Maddox, & Markman, 2007), and (b) diminishing sensitivity – the assertion that people are consistent in discounting payoffs magnitudes with the distance from zero (e.g., Ahn et al., 2008).

We contrast this "latent constructs" approach with two alternative views. The first is the classical economic approach that addresses risk attitude as sensitivity to differences in payoff variances (e.g., Pratt, 1964; Preuschoff, Bossaerts, & Quartz, 2006). The second is a recent view which suggests that "risk acceptance," the tendency of people to prefer (or avoid) risk over certainty is a single primitive construct that cannot be further dissected into the effect of gains and losses and the effect of diminishing sensitivity, but does not necessarily reflect sensitivity to variance (e.g., Brachinger & Weber, 1997). There are different formulations of the risk acceptance approach. For simplicity purposes we chose to focus on a simplified interpretation, referring to risk acceptance as the individual's sensitivity to certain versus probable outcomes. Thus, the risk acceptance hypothesis can be viewed as an extreme case of the sensitivity to variance hypothesis. That is, it suggests that the difference in variance is a necessary but insufficient condition of individual sensitivity to risk. The other necessary condition for risk sensitivity is a condition activating the individual's preference for certainty versus uncertainty.

The three aforementioned approaches are related but have distinct predictions that, surprisingly, have not been previously contrasted. The first such prediction involves the consistency between risk taking propensities in the gain and loss domain. Under the latent construct approach, supposing that indeed diminishing sensitivity underlies risk taking between domains, then a negative association is expected between risk taking in the gain and loss domains as implied by the reflection effect (Kahneman & Tversky, 1979). For example, if an individual discounts $1200 to a higher degree than she discounts $600 and is consistent in this diminishing sensitivity then she should be risk averse while choosing between a sure win of $600 and a bet with equal chances to win $1200 or nothing, but should be risk seeking when these values are framed as losses. In contrast, models based on the sensitivity to variance, as well as models of risk acceptance would predict a positive correlation between risky choices in the two domains, as individuals would either seek or avoid variance in both domains. However, the risk acceptance approach will have this prediction only when the choice alternatives also differ in their levels of certainty. These contrasting predictions are examined in Experiment 1.

The second prediction, which is the focus of Experiment 2, involves the consistency of the weighting of gains and losses. Under the latent construct model's assumption of weighting of gains and losses, a positive correlation should appear between choice problems differing in the magnitudes of gains and losses regardless of factors like variance or certainty. In contrast, the sensitivity to variance model predicts that the largest consistencies would appear between problems where the alternatives have the same levels of variance. The risk acceptance approach predicts choice consistency mostly when there are distinguishable differences in levels of certainty, such as in the choice between fixed and probabilistic outcomes. Experiment 3 focuses on the argument that risk acceptance involves a single primitive construct, even when gain domain problems are contrasted with choice problems involving both gains and losses.

Our comparison of different potential accounts for individual consistency in risk taking across tasks is closely related to previous studies of consistency in risk taking (Schoemaker, 1990) and to studies that compared models of risk taking (e.g., Battalio, Kagel, & Jiranyakul, 1990; Wakker et al., 2007). There are two major differences from these previous studies: First, these studies have tended to focus on the latent construct approach and did not systematically investigate alternative approaches to the psychological constructs underlying risk sensitivity. Secondly, these studies have focused on one-shot choices between described prospects, whereas we focus on risk taking in decisions from experience (Hertwig et al., 2004). In such decisions, Individuals do not get explicit information about the distributions that underlie the alternatives they face (e.g., the probabilities and payoff sizes). However, by choosing repeatedly between the different alternatives, and realizing the outcome of each choice (which is drawn from the relevant distribution) they can learn the potential outcomes associated with each alternative and their likelihoods. Previous studies have demonstrated that experience-based decision tasks have many attractive features for studying individual risk taking. It has been shown, for example, that such tasks have high external validity in assessing individual differences in decision making and that they are also relatively more resistant to social desirability than descriptive gambles (see review in Koritzky & Yechiam, 2010).

Experiment 1

The main purpose of our first study was to contrast the "diminishing sensitivity" assertion (appearing in latent component models such as prospect theory) with the "sensitivity to variance" hypothesis, and the "risk acceptance" assertion, by focusing on the main implication of the diminishing sensitivity construct, namely the contingent risk taking in the gain and loss domains. Each participant was presented with four repeated choice tasks, as described in Table 1. Each task included two alternatives and one (referred to as "L") was always associated with lower variance payoffs than the other ("H"). The main within-subject manipulation pertained to the domain in which choices were made. In the Gain condition choice alternatives yielded positive outcomes, whereas in the Loss condition outcomes were negative.

In order to differentiate between the "sensitivity to variance" and the "risk acceptance" hypotheses, the tasks were also distinguished with respect to the difference in the levels of uncertainty. In two of the tasks selecting the safer option eliminated probabilistic outcomes. We refer to these tasks as the "Avoidable Uncertainty" (AU) condition. In the other two tasks uncertainty could not be avoided since both alternatives included probable outcomes. These tasks are referred to as the "Unavoidable Uncertainty" (UU) condition.

The diminishing sensitivity assertion implies negative association between both domains in both the avoidable and the unavoidable uncertainty conditions because high diminishing sensitivity leads to risk seeking in the loss domain and risk aversion in the gain domain. Notice that this assertion also implies positive correlations between the two gain problems, and between the two loss problems. The risk acceptance assertion, however, suggests a positive association between the two avoidable uncertainty problems, and no association between the two unavoidable uncertainty problems. In the avoidable uncertainty problems there are clearer environmental signals concerning the differences in uncertainty level, which supposedly trigger risk acceptance tendencies. Finally, the sensitivity to variance model predicts positive association between all four choice problems due to one option being higher in variance than the other, even in the unavoidable uncertainty problems.

Forty undergraduates (20 males and 20 females) participated in the experiment. The participants' average age was 24 (ranging between 19 and 27). Payoffs ranged between NIS 14 and NIS 26 (NIS 1 = $4.5).

Each participant made 100 choices in each of the four choice problems. The participants were informed that they would be playing different games in which they would operate "computerized money machines" which include two unmarked buttons, and that their final payoffs would be sampled from one of the "machines" but received no prior information about the payoff distributions or the number of trials. Their task was to select one of the machine's two unmarked buttons in each trial. The payoffs in each task were contingent upon the button chosen and were randomly drawn from the relevant distributions described in Table 1. Final take-home amounts were determined according to the accumulating score in one choice problem that was randomly selected at the end of the experiment. The performance score was converted into cash money at a rate of 0.01 agora per 1 point (1 agora = 0.24 cents). The final payoff was then determined by summing this amount with the participation fee (NIS 25).

Two types of feedback immediately followed each choice: (1) the basic payoff for the choice, which appeared

on the selected button for two seconds, and (2) an accumulating payoff counter, which was displayed constantly, but was initialized at the beginning of each task. The order of the Gain and Loss conditions was counterbalanced, and the order of the two problems within each condition was randomized. The location of alternatives L and H was randomized across different participants. The measure used in each task was simply the proportion of choices of H across trials. There are therefore four variables in this study (and subsequent ones) conforming to the rate of H choices in each of the four choice problems.

Table 1: Payoff schemes of the four experimental conditions of experiment 1.

Domain	Condition	Payoff	P(H)
Gain	Avoidable Uncertainty	L: win 600 H: 50% to win 1200, 50% to win 0	0.26
Gain	Unavoidable Uncertainty	L: 50% to win 500, 50% to win 400 H: 50% to win 890, 50% to win 10	0.31
Loss	Avoidable Uncertainty	L: lose 600 H: 50% to lose 1200, 50% to lose 0	0.45
Loss	Unavoidable Uncertainty	L: 50% to lose 500, 50% to lose 400 H: 50% to lose 890, 50% to lose 10	0.49

Table 2: Spearman correlations between risk-taking in the different tasks in Experiment 1 (AU = Avoidable Uncertainty; UU = Unavoidable Uncertainty).

		AU		UU	
		Gain	Loss	Gain	Loss
AU	Gain	1.00			
	Loss	.45*	1.00		
UU	Gain	.63*	.22	1.00	
	Loss	.17	.35*	.03	1.00

* $p < .05$

Results

The choice proportions under the different conditions are summarized in the rightmost column of Table 1. The findings at the aggregate level show that people took more risk in the loss domain than in the gain domain ($t(39) = 3.98$, $p < .001$). There were no significant differences in risk taking between the AU and the UU conditions ($t(39) = 1.41$, NS).

The consistency of individuals' risk taking across the different tasks is presented in Table 2. The results show that in the AU condition there was a positive association between the gain and loss domains ($r = .45$, $p < .01$), which stands in contrast to the diminishing sensitivity hypothesis, and supports the risk acceptance assertion. Taking the UU condition into account, the results show that in this condition there was no association between the loss and gain domains ($r = .03$, NS), which further supports the risk acceptance assertion, since in the UU condition the probabilistic outcome could not be avoided (or accepted). In addition, participants were consistent between the two Gain problems ($r = .63$, $p < .0001$) and between the two Loss problems ($r = .32$, $p < .02$), suggesting that individuals might exhibit diminishing sensitivity to a certain degree.

Therefore, it seems that the reflection effect, implied by the diminishing sensitivity assertion, was not observed at the individual level. Instead, participants exhibited a consistent preference between a constant outcome and a probable outcome across the gain and loss domains. This suggests that risk acceptance modulates the consistency across the gain and loss domain and that diminishing sensitivity alone cannot account for it.

Additionally, the suggestion that the consistent sensitivity to risk is due to mere variance differences cannot account for the null correlations between gain and loss domain problems in the Unavoidable Uncertainty condition. Still, the variance difference in this condition was somewhat smaller than in the Avoidable Uncertainty condition (and thus it could be argued that this produced lower correlations in this condition). In the next experiment we examine problems that have the same exact differences in variance.

Experiment 2

The second experiment was designed to examine whether loss sensitivity indeed modulates risk taking behavior in problems involving gains and losses, or whether its effect are due to risk acceptance (or sensitivity to variance) as well. This was accomplished by contrasting two conditions involving losses and gains: A condition with strong differences in uncertainty level (i.e., the participants could opt for not selecting the gamble and get a sure outcome of zero) and a condition where the differences in uncertainty were smaller (i.e., selecting the safer option decreased the magnitude, but not the frequency of losses). We examined whether participants would still be consistent in their response to losses (across two choice problems) in the latter condition.

Under the latent component approach the loss-sensitivity construct involves pure sensitivity to the magnitude of losses compared to gains. Therefore, consistency is expected to be maintained regardless of the differences in uncertainty. Similarly, under the sensitivity to variance approach a positive correlation is expected to emerge as long as the alternatives maintain the same difference in variance. However, under the risk acceptance approach consistency is only expected to emerge in the condition where there are substantial differences in the level of uncertainty.

Each participant was presented with four repeated choice tasks, as described in Table 3. The tasks involved two conditions differing in the capacity of decision makers to avoid probabilistic outcomes. In two of the tasks selecting

the safer option eliminated the probability of losing. We refer to these tasks as the "Avoidable Uncertainty" (AU) condition. In the other two tasks uncertainty differences between alternatives were smaller and both alternatives included possible losses occurring with the same frequency (but differing in magnitude). Accordingly, these tasks are referred to as the "Unavoidable Uncertainty" (UU) condition. A second within-subject manipulation pertained to the level of variance associated with the riskier option. In condition "Low Variance" the standard deviation associated with alternative H (SD = 100) was one fifth of that associated with the corresponding alternative in condition "High Variance" (SD = 500). This enabled us to evaluate the consistency across different levels of variance and compare the consistency in the AU and UU conditions.

Thirty (15 males and 15 females) undergraduate students participated in the experiment. Their average age was 24 (ranging from 20 to 27). Payoffs varied between NIS 25 and NIS 33. The procedure was as in Experiment 1 except that the experiment focused on the tasks described in Table 3, and the conversion rate was 1 agora per 1 point.

Table 3: Payoff schemes of the four experimental conditions of Experiment 2.

Condition	Variance	Payoff	P(H)
Avoidable Uncertainty	Low	L: win 0 H: 50% to win 100, 50% to lose 100	0.64
Avoidable Uncertainty	High	L: win 0 H: 50% to win 500, 50% to lose 500	0.61
Unavoidable Uncertainty	Low	L: 50% to win 50, 50% to lose 50 H: 50% to win 150, 50% to lose 150	0.52
Unavoidable Uncertainty	High	L: 50% to win 250, 50% to lose 250 H: 50% to win 750, 50% to lose 750	0.51

Table 4: Spearman correlations between risk-taking in the different tasks in Experiment 1 (AU = Avoidable Uncertainty; UU = Unavoidable Uncertainty).

		AU		UU	
		Low var	High var	Low var	High Var
AU	Low var	1.00			
	High var	.54*	1.00		
UU	Low var	.07	-.08	1.00	
	High var	.20	.13	.13	1.00

* $p < .05$

Results

The choice proportions under the different conditions are summarized in the rightmost column of Table 3. At the aggregate level it seems that the participants tended to take more risk in the AU than in the UU condition ($t(29) = 3.15$, $p < .01$). Additionally, in both conditions participants did not appear to exhibit loss aversion, consistent with previous findings in experience-based tasks (e.g., Erev et al., 2008).

Table 4 presents the consistency of individuals' risk taking across tasks. The results reveal that despite showing no loss aversion on average, participants were highly consistent between the AU problems, in which risks could be avoided ($r = .54$, $p < .01$) but not in the UU problems, where risks could not be avoided ($r = .13$, NS).

Also, the participants did not show consistency across the two High-Variance and Low-Variance tasks, inconsistently with implication of the risk as variance. The correlations within each of the two pairs of High and Low variance tasks were small ($r = .07, .13$) and insignificant. This suggests that what makes participants respond consistently to high and low variance alternatives is not their mere variance.

This pattern suggests that the consistency in risk taking with losses is not driven by an accounting balance that inflates gains or losses (e.g., a weighted average of gain and loss amounts) nor is it driven only by sensitivity to variance. Rather, the participants were only consistent when a risky alternative involving losses and gains was contrasted with a safe alternative offering a fixed outcome. This indicates that the consistent construct in the mixed domain involves risk acceptance. Without strong signals of differences in risk level in the form of constant versus probabilistic outcomes, the correlation appears to disappear.

Experiment 3

From the results of Experiments 1 and 2 one can conclude that the main construct modulating people's responses is risk acceptance. Yet an alternative suggestion is that while risk acceptance consistently affects people's responses, this is limited to situations involving no explicit comparisons between gains and losses. Under the latent construct model, in the latter situation risk taking (i.e., selecting the high variance option) is solely due to the weighting of gains and losses and not due to diminishing sensitivity (because diminished sensitivity is balanced for gains and losses). While the pure weighting of gains and losses hypothesis was rejected in Experiment 2, it can still be argued that risk acceptance is an independent psychological construct when gains and losses are explicitly compared. The goal of Experiment 3 was therefore to examine whether risk acceptance is a single psychological construct or whether it implicates a second construct when the outcomes involve frequently appearing gains and losses. This was examined by comparing the consistency of risk taking across Gain and Mixed domain problems (as shown in Table 5). A second within-subject manipulation pertained to the level of risk. In Condition "Low Variance" alternative H was associated with a standard deviation smaller by half than in condition "High Variance" (SD = 1000, 2000, respectively).

Fifty (25 males and 25 females) undergraduate students participated in the experiment. Their average age was 24

(ranging from 21 to 28). Payoffs varied between NIS 20-30. The procedure was as in Experiment 1 except that the experiment focused on the tasks described in Table 5. The conversion rate was 1 agora per 1 point.

Table 5: Payoff schemes of the four experimental conditions of Experiment 3.

Condition	Variance	Payoff	P(H)
Mixed	Low	L: win 0 H: 50% to win 1000, 50% to lose 1000	0.55
Mixed	High	L: win 0 H: 50% to win 2000, 50% to lose 2000	0.56
Gain	Low	L: win 1000 H: 50% to win 2000, 50% to win 0	0.28
Gain	High	L: win 2000 H: 50% to win 4000, 50% to win 0	0.30

Table 6: Spearman correlations between risk-taking in the different tasks in Experiment 3.

		Mixed		Gain	
		Low var	High var	Low var	High Var
Mixed	Low var	1.00			
	High var	.57*	1.00		
Gain	Low var	.06	.11	1.00	
	High var	.14	.14	.55*	1.00

* $p < .05$

Results

The choice proportions under the different conditions are summarized in the rightmost column of Table 5. The results show that people took more risk on average in the Mixed condition than in the Gain condition under relatively low risk ($t(49) = 4.71$, $p < .01$) and also under higher risk ($t(49) = 2.93$, $p < .05$). This pattern is again inconsistent with loss aversion. It does replicate previous results in experience-based tasks (e.g., Erev et al., 2008).

The consistency of individuals' risk taking across the different tasks is presented in Table 6. The results reveal that participants were highly consistent between the two Mixed problems ($r = .57$, $p < .01$) and between the two Gain problems ($r = .55$, $p < .01$). However, participants were not consistent across the two problems: the association between the proportions of H choices in the two domains was small (average $r = .11$) and insignificant. These results suggest two separate construct for gains and losses of similar magnitudes. Another interpretation rests on the special case of a constant outcome of zero. It might be that the mixed condition was dissociated from the gain condition because participants have a special psychological tendency to respond to the absolute zero.

A quantitative index of subjective risk

The results of the current studies support the "risk acceptance" approach although suggesting that the psychological construct of risk acceptance could be different in a domain with both gains and losses. Yet a more challenging goal is to use these findings in an attempt to develop a quantitative index for what makes people respond consistently to risk. Individual differences studies indicate that a trait should be measured in a situation when it is relevant, which therefore involves a decision between a non-trivial amount of risk and a very low amount of risk. Therefore, the subjective difference in the risk of the alternatives is expected to lead to increased behavioral consistency in risk taking levels. We evaluated two quantitative indices for the emergence of consistency based on such subjective differences. A simple index was based on the idea that variance differences lead to consistency. According to this idea, the larger the differences in variance, the better a person differentiates between alternatives; and is thus more consistent in his or her risk taking behavior.

An alternative account involves the assumption that differences in subjective risk level (and therefore individual consistency) increase as a function of differences in variance but also decrease as a function of the distance from zero. This actually incorporates the two constructs that received only limited support in the experimental studies (sensitivity to payoff variance and diminishing sensitivity to marginal returns) into one construct that was largely supported by the experimental data (risk acceptance). This account can lead to a following index for subjective risk differences:

$$S = S_{diff} / \sum(|p_i \cdot x_i|) \qquad (1)$$

Where S is the Risk-Difference Signal (RDS), S_{diff} is the difference in standard deviation of the two distributions, p_i is the probability for each outcome i and and x_i is its size.

Under both accounts the risk differences in a problem pair are assumed to aggregate as follows:

$$C = S_1 \cdot S_2 \qquad (2)$$

This yields a parameter-free index C (of predicted consistency). The problems of Studies 1-3 were re-arranged into 18 pairs (representing all possible pairs within each study), and the risk difference in each pair was determined according to the two alternative indices. Then, the predictive ability of the two indices was determined by calculating the correlations between the predicted consistency of each pair and its actual consistency in risk level. The variance based index produced a correlation of 0.23, while the RDS index produced a correlation of 0.37 when predicting the consistency across all 18 comparisons.

A post-hoc version of the RDS, which differentiates non-mixed (gain or loss domain) from mixed (gain and loss) problems and is otherwise identical to the original index was also examined. It yields an average correlation (between predicted and actual consistency) of 0.80 for 14 relevant pairs: $r = 0.68$ for non-mixed problems ($n = 7$) and 0.91 for mixed problems ($n = 7$). For the variance-based

index the correlations are only 0.47 and 0.63, respectively.

Thus, the results of the current three studies cannot be interpreted by a parsimonious model resting just on variance differences. Rather, two additional assumptions must be made: (a). Subjective risk differences decrease as a function of the distance from zero, and (b). Two constructs of risk acceptance should be assumed: one for gain or loss domain problems and a unique construct for mixed outcomes.

Discussion

The main purpose of the current study was to shed light on the constructs leading to internal consistency in individuals' risk taking in experience-based decisions. Three approaches were contrasted: One suggesting that loss-sensitivity and diminishing sensitivity are the main factors that underlie individual differences in risk taking (see Busemeyer & Stout, 2002; Ahn et al., 2008), the other suggesting that the acceptance or the rejection of uncertainty is the principle factor modulating people's risk taking (Brachinger & Weber, 1997), and the third suggesting that sensitivity to differences in variance guides risk preferences (e.g., Pratt, 1964). To our knowledge, no previous studies have systematically evaluated the contrasting predictions of these approaches for the consistency of individual predispositions.

The findings of the three studies have important implications for the definition of subjective risk. Throughout the paper, and following the common convention in experimental studies of risky decisions in general and decisions from experience in particular, we have associated risk taking behavior with choices of the option with the higher variance as our point of departure. Nevertheless, our findings show that differences in variances alone do not drive individual consistencies in choosing the risky (higher variability) option. Rather, we have highlighted a second necessary condition: the presence of certainty. We view this finding as an example of a more general factor modulating individual consistencies, involving the extent to which the alternatives differ in their level of (un)certainty, with the case of certainty versus uncertainty being an extreme contrast along this axis. It appears that such a contrast is necessary in order to obtain consistency in risk taking even in problems that are relatively similar in terms of their payoff domain (e.g., the mixed domain problems of Experiment 2).

Additionally, the results of Experiment 1 confirmed the predictions of the risk acceptance construct for the consistency across domains (gains versus loss outcomes). In particular, this construct indicates positive consistency across domains, implying that people who take risks with gains also take risks with losses. This pattern contradicts the prediction based on diminishing sensitivity, which implies a negative correlation across domains (as explained above). It appears that the more consistent construct is risk acceptance.

In conclusion, as in previous examinations of individual risk taking, this construct was found to be consistent only in limited settings. Only in 6 out of 18 possible comparisons between simple experiential decision tasks did the participants exhibit consistency in their risk taking levels. Yet the current analysis also shows that the consistencies found are far from being coincidental, and it sheds light on the factors that modulate this behavioral consistency. A construct that seems to trigger the consistent tendency to take risk is the "risk acceptance" factor denoting individuals' sensitivity to differences in risk level when such differences are clearly perceived (such as in a decision between a constant outcome and a riskier prospect). When differences in risk level are less clear, lower consistency between different decision problems is observed.

References

Ahn, W.Y., Busemeyer, J.R., Wagenmakers, E.J., & Stout, J.C. (2008). Comparison of decision learning models using the generalization criterion method. *Cognitive Science*, *32*, 1376-1402.

Battalio, R.C., Kagel, J.H., & Jiranyakul, K. (1990). Testing between alternative models of choice under uncertainty: Some initial results. *Journal of Risk and Uncertainty*, *3*, 25-50.

Brachinger, H.W., & Weber, M. (1997). Risk as primitive: A survey of measures of perceived risk. *OR Spectrum*, *19*, 235-250.

Busemeyer, J.R., & Stout, J.C. (2002). A contribution of cognitive decision models to clinical assessment: Decomposing performance on the Bechara gambling task. *Psychological Assessment, 14*, 253-262.

Erev, I., Ert, E., & Yechiam, E. (2008). Loss aversion, diminishing sensitivity, and the effect of experience on repeated decisions. *Journal of Behavioral Decision Making, 21,* 575-597.

Hertwig, R., Barron, G., Weber, E.U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science, 15*, 534-539.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*, 263-291.

Koritzky, G., & Yechiam, E. (2010). On the robustness of decision tasks to response distortion. *Journal of Behavioral Decision Making, 23,* 83-99

Pratt, J.W. (1964). Risk aversion in the small and in the large. *Econometrica*, *32*, 122-136.

Preuschoff, K., Bossaerts, P., & Quartz, S.R. (2006). Neural differentiation of expected reward and risk in human subcortical structures. *Neuron*, *51*, 381-390.

Schoemaker, P.J.H. (1990). Are risk-attitudes related across domains and response modes? *Management Science*, *36*, 1451-1463.

Wakker, P.P., Köbberling, V., & Schwieren, C. (2007). Prospect-Theory's diminishing sensitivity versus economic's intrinsic utility of money: How the introduction of the Euro can be used to disentangle the two empirically. *Theory and Decision, 63*, 205-231.

Worthy, D. A., Maddox, W.T., & Markman, A.B. (2007). Regulatory fit effects in a choice task. *Psychonomic Bulletin and Review, 14*, 1125-1132.

Making Assessments While Taking Sequential Risks

Avishai Wershbale (wershbal@msu.edu) & Timothy J. Pleskac (pleskact@msu.edu)
Department of Psychology,
Michigan State University,
East Lansing, MI 48824

Abstract

This study utilized inter pump response times on a laboratory gambling task, the BART, to examine cognitive aspects of response selection during sequential risky decision making. Findings suggest a response procedure that utilizes multiple levels of processing. Amount of task exposure as well as the distance to the goal both affect the rate at which assessments are made, with task exposure decreasing assessment rate, while target distance increases assessment rate. Several alternative models are fit to the data, to determine if the behavioral results can be informative of a model that more accurately reflects differences in processing.

Keywords: Psychology, Decision Making, Mathematical Modeling, Cognitive Decision Theory

Introduction

People take sequential risks every day. Drivers repeatedly choose to talk on the cell phone, or text while driving each time they get in their car. People every day choose to eat at their favorite fast food establishment. Smokers of all ages repeatedly choose to smoke a cigarette. In all of these situations, our choice is not a one-shot deal rather our choices occur many times sequentially over the course of a day, a week, a month, etc. Often we are even presented with the same or similar choices on multiple occasions. Sometimes these choices may even change as a function of time or even as a function our previous choices (Busemeyer & Pleskac, 2009).

Despite the many instances of sequential risks in the real world, most of the laboratory analogs of risky decision making only ask participants to make one single choice (e.g., Hertwig, Barron, Weber, & Erev, 2004; Kahneman & Tversky, 1979). However, a number of laboratory-based gambling tasks have now been developed that require people to take sequential risk, such as the Iowa Gambling Task (Bechara, Damasio, & Anderson, 1994); the Balloon Analogue Risk Task (BART; Lejuez et al., 2002); or the Angling Risk Task (Pleskac, 2008). These tasks also appear to have some construct validity with real world risk taking. Risk taking in the BART, for example, correlates with smoking and drug abuse, as well as safety issues including seatbelt usage and safe sex (Lejuez et al., 2002; Lejuez, Aklin, Zvolensky, & Pedulla, 2003; Pleskac, Wallsten, Wang, & Lejuez, 2008).

The overall decision making processes for these sequential risk taking tasks are also reasonably well understood. In fact, the processes have been formalized in terms of cognitive models (Busemeyer & Stout, 2002; Wallsten, Pleskac, & Lejuez, 2005). However, the processes postulated within the components of the model have received less direct attention. In this study, we focus on the BART and test some of the processing implications of its respective cognitive model. In particular we use inter pump times to better understand the response selection process that decision makers use to take sequential risks during the BART.

Balloon Analogue Risk Task (BART; Lejuez et al., 2002)

During the BART participants are presented with a computerized balloon, a pump button, and a stop button. Pressing the pump button inflates the balloon, and also puts money in a temporary bank. Balloons in the BART explode after a randomly predetermined amount of pumps are made. An explosion terminates the trial and all money in the temporary bank is lost. Clicking the stop button, transfers the temporary points into the permanent bank and also terminates that trial. The participant's goal in the BART is to earn as much money as they can, but they need to take into account the chance that the balloon could pop. Participants typically complete 30 independent balloon trials.

Participants are given no information about the design of the task, other than the basic reward and punishment rules, and the fact that the balloon will explode before it fills the entire screen. It is up to the participant to determine the amount of risk they are willing to incur for a given reward. Though the BART may seem simple at first, there are many processes that take place within the course of completing the task. The processes have been formally defined in the Bayesian Sequential Risk Taking model (BSR; Wallsten et al., 2005).

Bayesian Sequential Risk Taking (BSR) Model

We briefly review the general processes of this 4 parameter model (for a formal derivation see Pleskac, 2008; Wallsten et al., 2005). The BSR model consists of three sub processes. The first process is a reward evaluation process. During this process, participants select a target pump number to pump the balloon towards. The target selected is a function of the participant's subjective value of a reward and their perceived chance that an explosion will occur for any pump. Participants who value the reward more will increase their target amount of pumps to be made, while an increase in the perceived chance of an explosion occurring decreases this pumping target.

The second process in the model is a response selection process. This process describes how participants use their target to determine whether to select a pump or a stop response at each pumping opportunity. In particular, the probability of pumping is assumed to be a function of the participant's current distance to the pumping target that they derived in the reward evaluation component. According to the model, the probability of pumping decreases as the current number of pumps taken approaches and passes the target number of pumps. Some participants are more consistent in pumping to their targets. A topic of interest in this paper is if in fact participants appear to be making a distance calculation, as the model seems to assume, at every pump opportunity.

The final process in the BSR model describes how participants learn from their experience. This process describes how participants arrive at their belief of the probability that the balloon will explode (used in the reward evaluation process described previously). The model assumes participants use a Bayesian learning process to integrate their prior beliefs with the observed data from each balloon trial (# pumps and if it exploded or not). Their new belief is used to evaluate rewards and select a target during the next balloon trial.

The model has been formally specified and tested (see Pleskac, 2008; Pleskac et al., 2009; Wallsten et al., 2005) with previous studies being by and large centered on the reward evaluation process and learning components of the BART. Little focus, however, has been allocated to the response selection component of the model. Recall the BSR model implies that on every pump opportunity of every balloon participants engage in some sort of distance-to-target calculation. If they are far from the target they are almost certain to pump and as they approach the target they become more and more likely to stop pumping. This raises the question whether participants perform a distance calculation at every pump opportunity? Our hypothesis is that instead of performing this calculation on every pump opportunity, there are instead two different types of pumping behavior being utilized. One pump type is a relatively automatic pump, while on other pump opportunities, decision makers pause and take an assessment of how far they have gone and how far they want to go. To test this hypothesis we examined the inter-pump times (the amount of time taken between responses).

Assessments

Cognitive psychologists have long known that due to limitations in working memory capacity an increase in the amount of information to process leads to an increase in the time it takes to process that information (Atkinson, Holmgren, & Juola, 1969; & Schneider & Schiffrin, 1977). This means that an action that occurs following a complex calculation should have a slower response time then if that action were preceded by an easier calculation. This cognitive principle implies the distance calculation the BSR assumes to take place when selecting a response should take some observable amount of processing time over and above motor time.

However, we also know that these sequential risk taking situations require choices on multiple trials. This high exposure to the task and task structure may lead to a routinization of the decision making process (Betsch, Haberstroh, Glöckner, & Fiedler, 2001) and perhaps even eventually approaching the automaticity properties of a habit (Aarts, Verplanken, & van Knippenberg, 1998). This routinization of decision making would imply less and less demands on working memory and thus lead to fairly quick inter-pump times.

Our hypothesis though is that there is some mix between fairly routine almost automatic pumps and other pumps where the decision maker pauses to take an assessment of where they are in the balloon trial. Our hypothesis is very much motivated by analogous findings from the animal learning literature where rats make a series of sequential decisions while traversing a maze. In particular, rats when learning a maze will at some decision points pause and appear to orient themselves toward potential options (Tolman, 1938 & Tolman 1948). Then after orienting themselves make a decision. This behavior has been termed vicarious trial and error (VTE), and has several interesting characteristics (Gallistel, Fairhurst, & Balsam, 2004). It was found that these VTEs occur fairly frequently during the early learning trials, and decreases with exposure to the task. This decrease in VTE's means that after enough exposure to the task environment, rats upon reaching a decision point (a fork where they have to go either right or left), eventually stop orienting themselves towards both potential options before making a decision, and instead simply immediately take the correct turn. It was also shown that this decrease in VTE's takes a non-linear shape.

These results prompt the question whether our postulated assessments follow the same pattern as VTEs. To test this we examined inter-pump times. Our hypothesis was that the inter-pump intervals in which a distance calculation was performed should take longer than those intervals in which no calculation was performed. And the inter-pump times for non-calculation intervals should not differ from baseline pumping speed.

To determine baseline inter-pump times, participants first completed a task in which only one option is presented to them: a pump option. Participants were instructed to pump each balloon as quickly as possible until they exploded. Participants neither received nor lost money for these trials. The inter-pump times from this were averaged together to estimate a baseline inter-pump time for each participant, as well as the standard deviation of their baseline pumping speed.

An assessment pump was operationalized as any pump for which the respective inter-pump time was 3 standard deviations or greater than the mean baseline time. Our hypotheses are as follows.

Hypothesis 1: Assessments will only occur periodically throughout a given balloon trial. The remaining trials will

reflect relatively routine almost automatic choices due to the frequency of their occurrence.

Hypothesis 2: Another testable prediction comes out of the hypothesis that the assessment points found in the BART reflect the same type of learning as the VTE's in rodent maze learning. As in the VTE's we expect that over balloon trials assessments will decrease in a non-linear fashion so that the observed assessment rate will decrease as exposure to the BART increases.

Hypothesis 3: Within a balloon trial, as participants approach their targeted stopping point their assessment rate should increase. This hypothesis is derived from the following reasoning. First, we assume that as participants pump they form an association between pump opportunities and the difficulty of making a choice (to pump or not). For example, participants will tend to associate the 4th pump trial with an easy decision (pump), but later pump opportunities (e.g., 48th opportunity) will present the participant with a more difficult choice. The prediction that follows from this hypothesis is that participants should be more likely to make an assessment on later pump opportunities for a given balloon.

Finally, we were interested whether the actual magnitude of the inter-pump times on routine pumps (non-assessed) would change over the course of pumping any given balloon. If we think of reaching the target pump as the main goal and the assessment points as sub-goals in reaching the target then we form a goal hierarchy. We know from goal activation models (Altmann & Trafton, 2002) that respondents often track their distance from previous sub goals and this leads to a slowing in response times as they progress. One might expect that participants are somehow implicitly tracking their distance from the last assessment. This would imply that inter-pump times of non-assessed pumps should increase as the distance from the last assessment increases. Next we test these predictions using data from two studies. Then we propose modifications based on these results to the BSR model.

Methods

The data examined comes from two experiments conducted in the Laboratory for Cognitive Decisions at Michigan State University during spring of 2008 and spring of 2009. These studies were designed to look at the effect of individual differences in various executive functions on BART performance. Both a standard version of the BART and a response time BART were included in these studies as well as a number of other executive function tasks, however we will limit descriptions of the tasks to just those relevant to this paper. Participants were college age undergraduates. A total of 104 students participated in the 2008 study and 108 in the 2009 study. There were no substantial differences between the two studies so we report their results together.

Baseline BART

The first task that every participant completed in both data sets was a baseline BART. The baseline BART is a simplified version of the BART that was created to measure average response time for pumping behavior. This version has only a pump button and participants are instructed to pump each balloon until it explodes. The balloons in the baseline BART were programmed to explode with the same statistical distribution that the normal BART balloons utilize. Participants completed ten trials of the baseline BART in order to establish a baseline non-fatigued measure of pumping motor time.

Manual BART

The regular BART task that we used is based on the task used in previous studies (see Lejuez et al., 2002; Pleskac et al., 2009). The task consists of a virtual balloon that is inflated by pressing a button. Participants were awarded 10 points for each successful pump. The popping point for each trial was randomly chosen out of 128, and pairings for each random trial were included to assure the same optimal distribution as in the original BART paper (Lejuez et al., 2002). Each trial ends with either a popped balloon (participants earn no points), or the participant clicking the stop button in which case the participant keeps all of the earned points for that balloon trial. Either way a fixation cross then appears in the center of the screen to prepare the participants for the onset of the following trial. To obtain more accurate response time data, our version of the manual BART was programmed in E-Prime 2.0. Furthermore, participants entered their pump and stop choices with separate keyboard buttons.

Results

Behavioral

In both studies, participants' risky behavior was consistent with past studies. On non-exploding balloons, they pumped an average of 39 (SD =16.03) and 34.3 (SD = 18.01) pumps

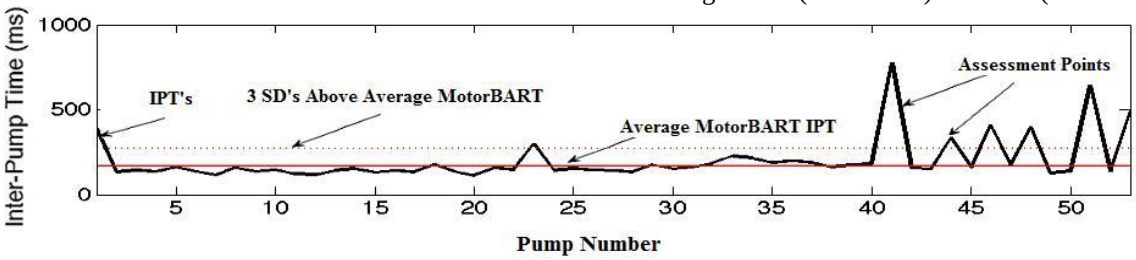


Figure 1: Inter Pump Time on a Single Trial for Participant 55 in study 2.

in studies 1 and 2, respectively. The average baseline inter-pump time in both studies was 181.38 ms (SD = 58.4) and 178.47 ms (SD = 54.22). The average within-subject standard deviation of inter-pump times on the baseline task was 53.91 ms (SD = 32.14) and 55.38 ms (SD = 40.66).

Recall we defined an assessment as any pump during the BART for which its respective inter-pump time exceeded 3 standard deviations above the baseline inter-pump time. With that definition, 11.1% of the pump opportunities from study 1 and 16.35% from study 2 would be classified as assessments. This yields an average assessment rate of 4.3 and 5.6 assessments per trial respectively. A plot of inter-pump times from a single subject and balloon trial is shown in Figure 1.

In terms of Hypothesis 2, to test whether there was a change in assessment behavior as participants become more familiar with the task, we regressed assessment rate onto trial number. Assessment rate was calculated by dividing the number of assessments in a given trial by the length of that particular trial. The results of this regression showed the same pattern for both studies, which is a decrease in assessment rate as trial number increases. Averaging across participants, the data seemed to best fit a logarithmic decreasing curve, with study 1 significant at $R^2 = .876$ and $p < .001$, and study 2 significant at $R^2 = .935$ and $p < .001$ (Figure 2). Thus, assessment rate was high on the first few trials (approximately a 40% assessment rate on average) and then decreased at an increasing rate as participants experienced more balloon trials.

Hypothesis 3 focused on whether the probability of an assessment changes within a balloon trial. Pump number itself cannot be used as the predictor for this regression, due to the fact that the length of each trial is entirely dependent on the participant's own pumping behavior. Instead a count of how many non-assessment pumps was taken between each assessment point, and then that count was divided by the length of that trial. This number is the proportion of pumps within that trial that preceded each assessment point (excluding the initial pump opportunity). Assessment points that are immediately followed by another assessment point were only counted as a single assessment. These proportions were then averaged across trials and across participants to give us a proportion score. A regression was run with assessment point number as the predictor and the proportion score as the dependant measure. Results of this regression also had a logarithmic fit, with $R^2 = .652$ and $p < .001$ for study 1, and $R^2 = .675$ and $p < .001$ for study 2. This is similar to the shape as for trial number, but the interpretation is that assessment rate increases as a factor of pump number.

Along with determining the factors that influence the probability of a distance to target assessment occurring, it is also important to identify characteristics of the non-assessed pumps as they approach the next assessment point. To characterize the changes in response times of the non-assessed pumps, Goodman-Kruskal Γ rank order correlations were ran to determine if there is generally an

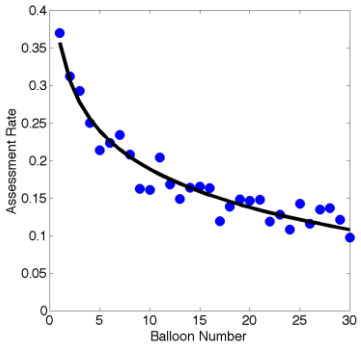


Figure 2: Change in Assessment Rate by Trial Number

increase, decrease, or no change in the response times the non-assessed pumps that were made between assessment points (hypothesis 4). A Γ coefficient was calculated for every string of non-assessed pumps that occurred. When the gamma coefficients were averaged (2615 coefficients in study 1 and 2887 coefficients from study 2), the results indicate an increase in inter-pump times time for the non-assessed pumps as they approach the next assessment point, with an average gamma coefficient of .108 for study 1 and .135 for study 2. While small, this result indicates a small but systematic increase in inter-pump times the further one gets from an assessment.

These results suggest that choice behavior during the BART is a bit more complex than that which is depicted in the BSR model. In particular, we have shown that in two studies on some trials participants pause and perhaps take an assessment of their situation. On the other trials the inter-pump times are quick enough to suggest a routine or perhaps even an almost automatic pumping behavior. Next we examine how to best modify the BSR model to incorporate these findings.

Proposed Changes to BSR Model

One possible way to account for these observed pauses in the BSR model is to modify the response selection process. Figure 3 illustrates this proposed change. The idea is that in the original response selection process participants either pumped or stopped and each response followed an assessment. Instead we propose that not every pump opportunity involves a distance to target assessment. That is with some probability participants stop to make an

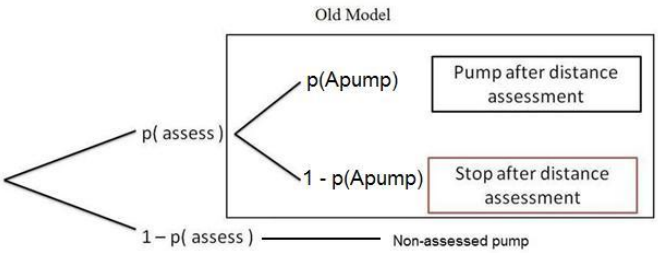


Figure 3: One possible modification to the BSR model.

assessment and only then do they choose between pumping and stopping. On other trials they make what we call a non-assessed pump which has an inter-pump time that is close to baseline. Several different functional forms of this modification were examined to determine if the inclusion of non-assessed behaviors improves the model fit.

The changes to the model are based off of the predictions from the new hypotheses. We tested four modifications. The first model assumed a static assessment rate, μ, which estimates the probability of making an assessment before the response is selected.

The second model assumed assessment changed as a function of balloon trial (h),

$$P(assess) = \frac{1}{1+\exp(\mu(h-\lambda))}. \quad (1)$$

Where λ is a biasing factor that controls the starting point of the assessment rate and μ now controls the rate of change in the assessment rate.

The third model incorporates the idea that assessments change as a function of pump opportunity i,

$$P(assess) = \frac{\exp(\mu(i-\lambda))}{1+\exp(\mu(i-\lambda))} \quad (2)$$

The parameters serve much the same role as in Equation 2, but now assessment rate changes as a function of pump opportunity and λ controls the starting assessment rate for each balloon.

The fourth model incorporates both an assessment rate that changes as a function of balloon trial h and pump opportunity i,

$$P(assess) = \frac{\exp(\mu(i_{/h}-\lambda))}{1+\exp(\mu(i_{/h}-\lambda))} \quad (3)$$

Preliminary Model Fitting & Comparisons

The modifications were incorporated into the BSR model. Then each model was fit at the individual level with maximum likelihood procedures using the Nelder-Meade numerical optimization routine. Several different starting points were used to try and guard against local maxima issues. A Bayesian Information Criterion (BIC) was calculated for each model, and was used to determine the best fitting model overall (where lower BIC means better fit). The BIC is a goodness of fit measure which penalizes models for the number of parameters they have.

Table 1: Average BIC scores for Alternative Models

	Ave (Std) BIC	Ave (Std) BIC
Baseline	740.32 (358.85)	758.86 (428.15)
Static Assessment Rate	733.06(360.15)	746.54 (434.73)
Assessment rate changes as a function of balloon trial	718.74 (325.78)	734.20 (397.82)
Assessment rate changes as a function of pump opportunity	728.57 (358.84)	747.88 (442.52)
Assessment rate is a function of both trial and pump opportunity	741.62 (338.5)	733.60 (410.01)

The models were also compared against a statistical baseline model. The baseline model simply uses the observed proportion of assessed pumps, non-assessed pumps, and stops over the thirty balloons and estimates the likelihood of the data by utilizing those proportions.

The average BICs are shown in Table 1. They show that the best fitting model is one in which assessment rate changes as a function of balloon trial. It is of note that there is some variability in this conclusion at the individual level. In particular, while nearly all the participants exhibited some form of assessment behavior, there was individual variability in the relationship between assessment rate and balloon trial. Some (~40%) showed a very weak relationship between balloon trial and assessment rate. For these individuals, the constant assessment model and the baseline model provided better fits.

Discussion

This study aimed to better understand the response selection process in sequential risk taking situations. Using the BART as an analog to these situations, we found that nearly all participants engaged in a behavior we call assessment. That is, within a given sequence of risky choices, generally decision makers would make very quick choices, but periodically (about every 4 to 5 pumps) they would take very long pauses. We interpret this behavior as a time of assessment when the decision maker gauges how many risks they have taken and how many more risks they plan to take. We also found the following behavioral properties of an assessment rate.

First, across balloon trials the assessment rate was on average higher for early balloon trials and then diminished at an increasing rate. This idea is consistent with previous decision making literature with rodents, showing that as task exposure increases, learning takes place, which leads to automated decision evaluations (Tolman, 1938 & Tolman 1948). The second, property of assessment rate, was the change within a single balloon trial, where assessment rate increases towards the end of the trial. One possible explanation is that the assessment rate increases relative to the level of perceived risk. It would be interesting to see if this is reflected in self reported risk taking measures. Lastly, there was a small but significant increase in inter-pump times between assessments. This suggests an increasing taxing of cognitive resources, which may be due to a buildup of interference, so that eventually the participants need to reassess their location relative to their pump target

Assessments and the change in assessment rate over balloon trials may be analogous to the distinction between exploration and exploitation in sequential decision making (Schumpter, 1934 & Holland, 1975). This idea of exploration versus exploitation holds that in order to maximize gains, one should initially explore the structure of the environment to create a good approximation of the distribution of rewards. Once a good approximation of the environmental structure is obtained, then one should begin exploiting it in a manner that maximizes their gains.

Assessments in sequential risk taking may afford the decision maker with an opportunity to explore different risk options and then exploit the options.

Acknowledgments

We thank David McFarlane for his help in programming the BART in E-Prime.

References

Aarts, H., Verplanken, B., & van Knippenberg, A. (1998). Predicting behavior from actions in the past: Repeated decision making or a matter of habit? *Journal of Applied Social Psychology, 28*, 1355-1374.

Altmann, E. M. & Trafton, J. G. (2002). Memory for goals: An activation-based model. Cognitive Science, 26, 39-83

Atkinson, R. C., Holmgren, J. E., & Juola, J. F. (1969) Processing time as influenced by the number of elements in a visual display. *Perception & Psychophysics, 6*, 321-326

Bechara, A., Damasio, A. R., Damasio, H., & Anderson, S. W. (1994). Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition, 50*(1-3), 7-15.

Betsch, T., Haberstroh, S., Glöckner, Haar, T., & Fiedler, K., (2001). The effects of routine strength on adaptation and information search in recurrent decision making. Organizational Behavior and Human Decision Processes*, 84,* 23-53.

Busemeyer, J. R., & Pleskac, T. J. (2009). Theoretical tools for understanding and aiding dynamic decision making. *Journal of Mathematical Psychology, 53*(3), 126-138.

Busemeyer, J. R., & Stout, J. C. (2002). A contribution of cognitive decision models to clinical assessment: Decomposing performance on the Bechara gambling task. *Psychological Assessment, 14*(3), 253-262.

Gallistel, C. R., Fairhurst, S., & Balsam, P. (2004). The learning curve: Implications of a quantitative analysis. *PNAS, 101*, 13124-13131

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science, 15*(8), 534-539.

Holland, J. H. (1975), *Adaptation in Natural and Artificial Systems*. Ann Arbor, Ml: University of Michigan Press

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47*(2), 263-291.

Lejuez, C. W., Read, J. P., Kahler, C. W., Richards, J. B., Ramsey, S. E., Stuart, G. L., et al. (2002). Evaluation of a behavioral measure of risk taking: The Balloon Analogue Risk Task (BART). *Journal of Experimental Psychology: Applied, 8*(2), 75-84.

Lejuez, C. W., Aklin, W. M., Zvolensky, M. J., & Pedulla, C. M. (2003). Evaluation of the Balloon Analogue Risk Task (BART) as a predictor of adolescent real-world risk-taking behaviours. *Journal of Adolescence, 26*(4), 475-479.

Pleskac, T. J. (2008). Decision making and learning while taking sequential risks. *Journal of Experimental Psychology-Learning Memory and Cognition, 34*(1), 167-185.

Pleskac, T. J., Wallsten, T. S., Wang, P., & Lejuez, C. W. (2008). Development of an Automatic Response Mode to Improve the Clinical Utility of Sequential Risk-Taking Tasks. *Experimental and Clinical Psychopharmacology, 16*(6), 555-564.

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic information processing: I. Detection, Search, and attention. *Psychological Review, 84*(1), 1-66

Schumpter, J. A. (1934). *The Theory of Economic Development*. Cambridge. MA: Harvard University Press.

Tolman, E. C. (1938). The determiners of behavior at a choice point. *Psychological Review, 45*, 1-41

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological Review, 55*, 189-208

Wallsten, T. S., Pleskac, T. J., & Lejuez, C. W. (2005). Modeling behavior in a clinically diagnostic sequential risk-taking task. *Psychological Review, 112*(4), 862-880.

Testing Two Explanations for the Disjunction Effect in Prisoner's Dilemma Games: Complexity and Quasi-Magical Thinking

Evgenia Hristova (ehristova@cogs.nbu.bg)
Maurice Grinberg (mgrinberg@nbu.bg)

Central and East European Center for Cognitive Science,
New Bulgarian University, 21 Montevideo Street, Sofia 1618, Bulgaria

Abstract

The paper explores the disjunction effect in the Prisoner's dilemma game using behavioral experiments with eye-movement recordings. An experiment was designed to explore the complexity hypothesis about the appearance of the disjunction effect. The results show that in games with payoffs which are simpler to perceive and compare, the disjunction effect disappears, while it is present when more complex payoffs are used. In a second experiment, the participants were told that the moves of the computer opponent had been made before the game session. This manipulation led again to the disappearance of the disjunction effect even. We interpret this result as a suppressing of a possible quasi-magic reasoning by stressing the fact that participants' own moves cannot influence the move of the opponent. The results from the experiments point to information processing complexity as a major factor for the disjunction effect contrary to the conclusions in some previous research.

Keywords: disjunction effect, eye-tracking, Prisoner's Dilemma, decision making

Introduction

The Prisoner's dilemma (PD) game is one of the most extensively studied social dilemmas. PD is a two-person game. The payoff table for this game is presented in Figure 1. In the PD game the players simultaneously choose their moves – C (cooperate) or D (defect), without knowing their opponent's choice.

In order to be a Prisoner's dilemma game, the payoffs should satisfy the inequalities $T > R > P > S$ and $2R > T+S$. Because of this game structure a dilemma appears – there is no obvious best move. On one hand, the D choice is dominant for both players – each player gets larger payoff by choosing D than by choosing C no matter what the other player chooses. On the other hand, the payoff for mutual defection (P) is lower than the payoff if both players choose their dominated C strategies (R for each player).

As PD game is used as a model for describing social dilemmas and studying the phenomena of cooperation, there is a great interest in the conditions that could promote or diminish cooperation. The *cooperation index* (CI), computed as $CI = (R–P)/(T–S)$ (see Rapoport and Chammah, 1965) is assumed to indicate the degree to which a player can be motivated to cooperate (choose move C).

The disjunction effect in Prisoner's Dilemma has attracted considerable interest and has been investigated in several experimental and theoretical studies without reaching consensus about its explanation (e.g. see Shafir & Tversky, 1992; Croson, 1999; Busemeyer et al. 2006; Li, Taplin & Zhang, 2007; Hristova & Grinberg, 2008). The disjunction effect can be summarized as follows: experiments with one-shot PD games show that players choose move D more often when they know the move of their opponent whatever it is (C or D) than when they don't know it. The logical expectation is that if participants choose a particular strategy for any of the two possible moves of the opponent, they should have the same strategy when they don't know their opponent's move. However, people do not act as expected and cooperate more in the latter situation, i.e. when the opponent move is uncertain.

		Player II C	*Player II* D
Player I	C	R, *R*	S, *T*
	D	T, *S*	P, *P*

Figure 1: Payoff table for the PD game. In each cell the comma separated payoffs are the Player I's and Player II's payoffs, respectively.

Several explanations for the disjunction effect have been put forward in the PD literature. Shafir and Tversky (1992) are accounting for the disjunction effect using their theory for reason-based choice: people need a reason in order to make a choice. Thus, when they know that their opponent will play C, they defect to get the higher payoff; and if they know that she will play D, they defect in order to avoid the lowest payoff and punish the opponent (see Figure 1). But when the move of their opponent is not known, they don't have a particular reason to make a move and this changes the situation contributing to the disjunction effect. Additional explanations, discussed in the same paper, claim that people cannot account properly for all alternatives of the game, or if they do, due to the uncertainty about the opponent's move, they cannot establish clearly their own preferences. Thus, depending on what outcome they focus on, they can choose to cooperate or defect. When people are made aware of their choices the disjunction effect disappears (Tversky & Shafir, 1992).

An alternative explanation is related to the change of participant's perspective (individualistic vs. collectivistic) about the PD game (see Shafir & Tversky, 1992). When

their opponent's move is known, people can be tempted to defect as the outcome of the game depends only on their choice and they have to consider only one column in the game matrix (see Figure 1). In this case, they adopt an individualistic point of view and defect. On the other hand, when the opponent's move is unknown, they have to consider the whole game matrix, the outcome depends on their and their opponents moves, and the collectively optimal decision of mutual cooperation becomes more attractive. This is supported by the fact that the CC outcome (payoff R) is better than the DD outcome (payoff P) for both players (as R>P; see Figure 1).

Experiments show that sometimes participants act as if they believe that their moves can influence the game outcome, although they know this is impossible. In Shafir & Tversky (1992), this is called quasi-magical thinking. Quasi-magical thinking, applied to PD, would imply that if people cooperate more when they are uncertain about the other player move, this means that the CC outcome is preferred by them, and by playing C they expect to elicit the same choice in the other player.

In the account of Shafir & Tversky (1992), the possibility of complexity to be an explanation of the disjunction effect is discarded and it is claimed that 'the failure to reason consequentially may constitute a fundamental difference between natural and artificial intelligence.' Croson (1999) tested the complexity explanation and the conclusion was that complexity plays no role and the reason-based choice explanation should hold. However, the test was performed using games which are not dilemmas as the PD game. Recently, inspired by the above conclusions, alternative explanations have been put forward even involving quantum probability theory and logic (see Busemeyer et al., 2006).

Li et al. (2007) used the so-called 'equate-to-differentiate' approach to explain the disjunction effect. This approach seems to involve the complexity hypothesis by assuming that when people have ambiguous alternatives concerning their own payoffs, they can equate them and take the perspective of their opponent. Moreover, the eye-tracking study of Hristova & Grinberg (2008), has shown longer information acquisition in PD games when the opponent's move is uncertain than when it is known by participants, reflecting the difference in the complexity of the task in the two cases.

One of the goals of the present study is to explore to what extent the complexity of decision making can lead to the disjunction effect in PD games. The approach adopted here, is different from the one followed in Croson (1999). Instead of using games with different structure, in our experiments we manipulated the payoffs by keeping their ratio and cooperation index the same. Participants in one experimental condition played PD games with payoffs which were two digit numbers with the second number *different* from zero. Participants in another experimental condition, played games with two-digit numbers with the second digit *equal* to zero. The idea was that, while equivalent from a game-theoretical point of view, the payoffs from the first condition are more difficult to perceive and compare than the numbers in the second condition. Thus, the complexity of the former case was assumed to be higher than the complexity of the latter case.

The second goal was to try to evaluate the influence of quasi-magical thinking discussed above on the disjunction effect in PD games. This has been done by using exactly the same experimental design as the one described above but with an additional manipulation – a sentence in the instruction which says that the computer program, playing against the participants, had chosen its moves before the beginning of the game session.

In both experiments eye-movements have been recorded in the hope to discern differences in the four conditions which could shed additional light on information processing involved based on the experience of Hristova & Grinberg (2008).

Experiment 1 – Testing the complexity explanation

Goals and hypothesis

The goal of the present experiment is to test the complexity explanation for the disjunction effect, namely that the effect appears because of the complexity of the game. When the opponent's move is not known, the situation is complex and the players are not able to analyze it well and to choose the appropriate move. To test this, in the current experiment we manipulate the complexity of the payoffs that are presented. The prediction is that if we make the game simpler (by using simple round payoffs) the disjunction effect will be smaller.

Stimuli

A set of 6 Prisoner's dilemma games was used in the experiment (see Table 1). Although the payoffs were different, the cooperation index of all the games was equal to 0.7 (as discussed above, cooperation index is an important predictor of the cooperation rate). Three of the games were with simple round payoffs, and 3 games were with 'complex' payoffs.

Table 1: PD games used in the experiment

	T	R	P	S
simple payoffs	100	90	40	30
	90	80	30	20
	80	70	20	10
complex payoffs	106	94	41	32
	91	83	34	22
	83	75	24	12

Each of the 6 PD matrices was presented 3 times during the game playing: the computer move in not known yet, the computer move is known to be cooperation, the computer move is known to be defection. These 18 payoff matrices that are later used in the analysis were intermixed with 62 other games resulting in a total of 80 games. The 18 PD games were pseudo-randomly distributed between the 4th and the 78th game.

. Care was taken as one and the same PD game to appear in the first, second, and third part of the game sequence. Playing games with different strategic structure was used to introduce the PD games as one-shot games and prevent subjects for using strategies applicable in the repeated play of PD.

Eye Movements Recordings

Eye movements were recorded using the Tobii 1750 remote binocular eye-tracker with 50 Hz sampling rate. The accuracy of the gaze position record is about 0.5 degrees visual angle. The game was presented on the Tobii monitor (17", 1280x1204 pixels). Each box containing payoffs or moves occupied about 1 degree visual angle on the screen. The distance between two adjacent boxes was at least 1 degree visual angle to ensure stable distinction between eye-fixations belonging to respective zones. The schematic game interface is presented in Figure 2.

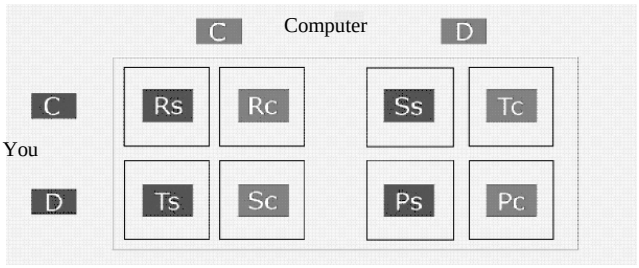


Figure 2: Schematic representation of the game interface. During playing, the actual payoffs and moves are presented. The superscript 's' refers to the subject, and 'c' to the computer opponent.

Procedure

After the eye-tracking calibration, subjects received instructions and were tested for understanding the instructions. Then each subject played 5 training games, and next the 80 games described above.

The game was presented in a formal and a neutral formulation. On the interface, the cooperation move was labeled '1' and the defection move was labeled '2'. However, further in the paper, for convenience, we will continue to use *cooperation* instead of move '1' and *defection* instead of move '2'.

Subjects were instructed to try to maximize their payoffs and not to try to 'beat' the computer. After each game the subjects got feedback about their and the computer's choice and payoffs in the current game. This information was visible for 3 seconds and then the next game automatically appeared. To ensure that players are following the instruction, three participants that got most points were promised and given a reward. In such a way we were trying to emphasize the importance of getting more points (and not trying to get more points than the opponent). Participants were not told their total number of points until the end of the game.

It is explained to the participants that the computer makes its choices trying to maximize the payoff in each game. They were also told that the computer is not aware of the participant's choice. In fact the computer's moves were randomly generated in advance and were the same for all participants.

Participants

33 subjects with normal or corrected to normal vision took part in study. Playing behavior of all subjects was analyzed, however, due to technical difficulties, eye-tracking data of only 22 of the subjects were analyzed.

Playing results

The number of cooperative choices for PD games was used as a dependent variable characterizing the participants' playing and choices. If the disjunction effect is present, the cooperation rate in the unknown move condition will be higher then either the defect (D) or cooperate (C) known move condition. If no disjunction effect is observed, the cooperation rate for the unknown move condition is expected to be equal or between the cooperation rates for D and C. This is the reason to compare the unknown move condition against the known D and C conditions separately and not against the aggregated data.

For the games with *complex* payoffs the expected pattern for a disjunction effect appeared in the data (see Figure 3). Participants cooperated in 27 % of the PD games in which the computer move was not known, in 11 % of the games that the computer move was known to be cooperation, and in 12 % of the games that the computer move was known to be D defection. Cooperation rate when the computer move is not known is significantly different ($p < 0.05$) from the cooperation rates when the computer move is known to be cooperation or defection.

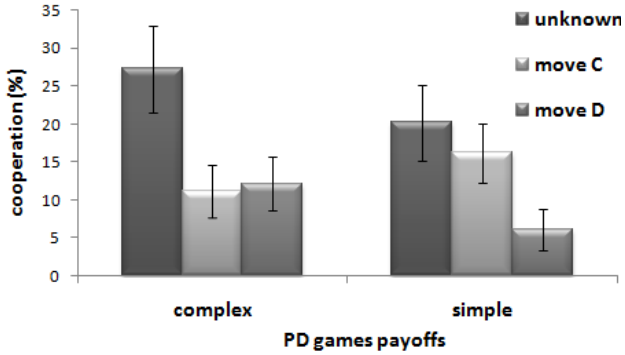


Figure 3: Mean cooperation (%) in Experiment 1 for *complex* and *simple* PD games when the computer's move is not known (*unknown*); the computer's move is known and it is cooperation (*move C*); the computer's move is known and it is defection (*move D*). Error bars represent standard error of the mean.

For the games with *simple* payoffs the disjunction effect was not so prominent (see Figure 3). Participants cooperated in 20 % of the PD games in which the computer move was not known, in 16 % of the games that the computer move was known to be cooperation, and in 6 % of the games that the computer move was known to be defection. Although the trend is present, there is no significant difference in the cooperation rates when the computer move is unknown or

when the computer move is known to be cooperation. Cooperation rate when the computer move is known be defection is significantly different ($p < 0.05$) from the cooperation rates when the computer move is not known or is known to be cooperation.

In summary, when the payoffs were complex, a clear disjunction effect appeared. However, the effect was not statistically significant when the payoffs were simple. It seems that these results support the complexity explanation for the disjunction effect because change in the complexity of the payoffs changes the participants' choices and the disjunction effect diminishes.

Eye-movements results

We defined several areas on the screen that are interesting in studying information acquisition during PD game playing. Each Area of Interest (AOI) contains the box in which the information is presented and a small region around it. Here we present the analysis of the eye-tracking data for the four AOIs containing the subject's possible payoffs These AOIs are referred to as T_S, R_S, P_S, and S_S (see Figure 2). We used the gaze-time (sum of all fixation durations on each AOI) as a measure of attention devoted to it (Rayner, 1998).

We expect that when the computer's move is known, the subject's attention will be directed to the possible payoffs corresponding to the computer's choice. So, for each game we computed the aggregate gaze-times in the zones containing subject's possible payoff if the opponent cooperates (R_S and T_S) and in the zones containing subject's possible payoff if the opponent defects (S_S and P_S). These data are analyzed in a repeated-measures analysis of variance with the computer's move (not known, known to be cooperation, and known to be defection) as a within-subject factor. Two such analyses were performed: for the PD games with complex payoffs and for the PD games with simple payoffs.

For the games with *complex* payoffs when subjects knew that the computer's move was defection they attended less to the AOIs denoted as T_S and R_S compared to the games when the computer's move was not known ($p = 0.018$) or it was known to be cooperation ($p = 0.055$) (see Figure 4). When subjects knew that the computer's move was cooperation they attended less to the AOIs denoted as P_S and S_S compared to the games when the computer's move was not known ($p = 0.004$) or when it was known to be defection ($p = 0.002$) (see Figure 4).

For the games with *simple* payoffs when subjects knew that the computer's move was defection they attended less to the AOIs denoted as T_S and R_S compared to the games when the computer's move was not known ($p = 0.018$) or it was known to be cooperation ($p = 0.02$) (see Figure 4). When subjects knew that the computer's move was cooperation they attended less to the AOIs denoted as P_S and S_S compared to the games when the computer's move was not known ($p < 0.001$) or when it was known to be cooperation ($p < 0.001$) (see Figure 4).

In summary, the eye-tracking data show that when the opponent's move is known, the eye-movement patterns are changed in both types of games (*complex* and *simple*). When the computer's move is D, the subject's possible payoffs are S_S or P_S and they do not pay attention to the other payoffs (R_S and T_S). When the computer's move is C, the subject's possible payoffs are R_S or T_S and they do not pay attention to the other payoffs (S_S and P_S).

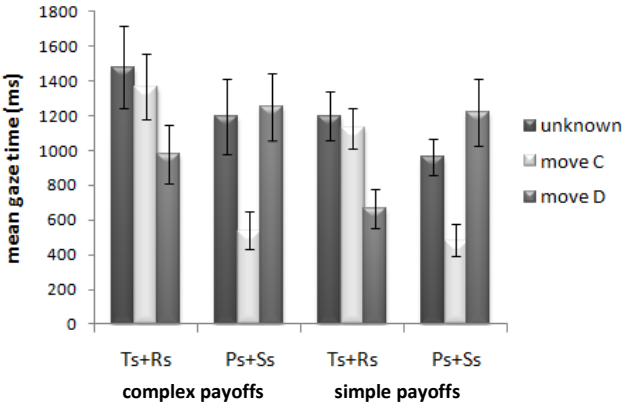


Figure 4: Average gaze-time for AOIs containing subject's possible payoffs (T_S, R_S, P_S, and S_S) when the computer's move is not known (*unknown*); the computer's move is known and it is cooperation (*move C*); the computer's move is known and it is defection (*move D*). Error bars represent standard error of the mean.

Another measure analyzed is the total gaze-time in the four AOIs containing the subject's possible payoffs. The analysis shows that when the payoffs are complex players spend more time looking at their them (mean 2270 ms) than when the payoffs are simple (mean 1880 ms), $p = 0.04$. This result indicates that *simple* payoffs are indeed easier to process than the *complex* payoffs.

Summary and discussion for Experiment 1

All these results are in accordance with the complexity explanation of the disjunction effect. When complexity of the PD game is reduced (by using payoff that are easy to process) the disjunction effect is reduced. Eye-movements data also support this explanation. When the computer's move is known, the eye-movement patterns are changed – players are paying less attention to the payoffs that are not relevant for the already revealed opponent's move. Eye-movement data also give evidence that the intended reduction in complexity of the game is successful.

Experiment 2 – Testing the quasi-magical thinking explanation

Goals and hypothesis

The goal of this experiment is to test the explanation that the disjunction effect arises due to the so called 'quasi-magical thinking'. The explanation is that the players behave as if they believe that their choices could influence the other

player's choices (although they know that in fact the other player is not aware of their choice while making his).
To test this explanation in this experiment we use a novel manipulation consisting in telling the subjects that the computer's moves are determined in advance. When this fact is known the above stated 'quasi-magical beliefs' should be diminished and the disjunction effect should disappear.

Stimuli and Procedure

Game and procedure were the same as in experiment 1. Change was made only in the information given to the participants in regard to the computer's move. It is said not only that the computer tries to maximize its payoff in each game but also that the computer has determined all of its moves in advance, before the start of the sequence of 80 games.

Participants

27 subjects with normal or corrected to normal vision took part in study. Playing behavior of all subjects was analyzed, however, due to technical difficulties, eye-tracking data of only 16 of the subjects were analyzed.

Playing results

For the games with *complex* payoffs participants cooperated in 21 % of the PD games in which the computer move was not known, in 12 % of the games that the computer move was known to be cooperation, and in 7 % of the games that the computer move was known to be defection (see Figure 5). Cooperation rate when the computer move is not known is significantly different ($p < 0.05$) from the cooperation rate when the computer move is known to be defection. All other differences are non-significant. Although the trend is present, there is no significant difference in the cooperation rates when the computer move is unknown or when the computer move is known to be cooperation.

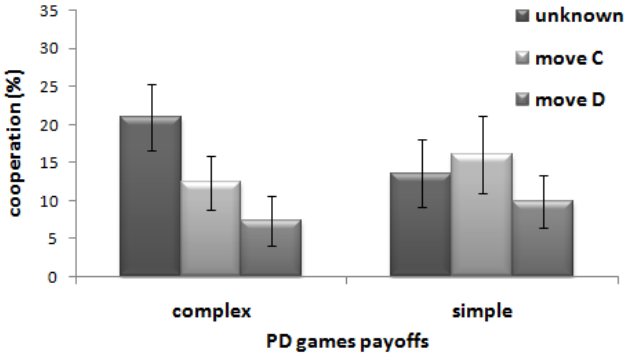


Figure 5: Mean cooperation (%) in Experiment 2 for *complex* and *simple* PD games when the computer's move is not known (*unknown*); the computer's move is known and it is cooperation (*move C*); the computer's move is known and it is defection (*move D*). Error bars represent standard error of the mean.

For the games with *simple* payoffs no disjunction effect was found (see Figure 5). Participants cooperated in 14 % of the PD games in which the computer move was not known, in 16 % of the games that the computer move was known to be cooperation, and in 10 % of the games that the computer move was known to be defection. There is no significant difference between these three cooperation levels.

In summary, when the players know that the computer moves are determined before the start of the sequence of games, the disjunction effect is smaller or absent. Especially when the PD games payoffs are easy to be processed and compared, no such effect is observed.

Eye-movements results

For the games with *complex* payoffs when subjects knew that the computer's move was cooperation they attended more to the AOIs denoted as T_S and R_S compared to the games when the computer's move is not known ($p = 0.009$) and it is known to be defection ($p = 0.003$). They also attended less to the AOIs denoted as P_S and S_S compared to the games when the computer's move is not known ($p < 0.001$) and it is known to be defection ($p < 0.001$) (see Figure 6).

For the games with *simple* payoffs, when subjects knew that the computer's move was D, they attended less to the AOIs denoted as T_S and R_S compared to the games when the computer's move was not known ($p = 0.044$) and when it was known to be C ($p = 0.08$). They also attended less to the AOIs denoted as P_S and S_S compared to the games when the computer's move was not known ($p = 0.001$) and when it was known to be C (< 0.044) (see Figure 6).

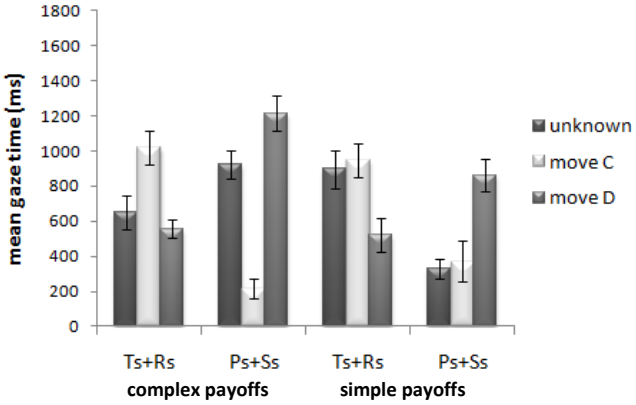


Figure 6: Average gaze-time for AOIs containing subject's possible payoffs (T_S, R_S, P_S, and S_S) in Experiment 2 when the computer's move is not known (*unknown*); the computer's move is known and it is cooperation (*move C*); the computer's move is known and it is defection (*move D*). Error bars represent standard error of the mean.

In summary, the eye-tracking data show that in both types of games (*complex* and *simple*) the eye-movement patterns are changed. When the computer's move is known, the subjects pay more attention to their possible payoffs in the corresponding column and they don't pay attention to the other payoffs.

Again we analyzed the total gaze-time in the four AOIs containing the subject's possible payoffs. The analysis shows that when the payoffs are complex players spend more time looking at their payoffs (mean 1525 ms) than when the payoffs are simple (mean 1309 ms), $p = 0.036$. Total gaze time in these four AOIs is less in Experiment 2 (mean 1417 ms) than in Experiment 1 (mean 2079 ms), $p = 0.034$.

Summary and discussion for Experiment 2

As expected, when the subjects are told that the computer moves are already decided, the disjunction effect is reduced and even disappears when games with lower complexity are played. These results are in accordance with the quasi-magical thinking explanation of the disjunction effect and also give further support for the complexity explanation.

Discussion and Conclusions

The paper presents an experimental study of the disjunction effect in PD games based on behavioral experiments with eye-movement recordings. The experiments were designed to explore the complexity hypothesis about the appearance of the disjunction effect which seems to have little support in the literature. However, our study showed that without changing the structure of the game (and its cooperation index), but by just using payoffs which can be easily processed, the disjunction effect can disappear. We interpret this result as an indication that despite the arguments and evidences that have been discussed in the literature (see Shafir & Tversky, 1992; and Croson, 1999) the role of complexity should not be underestimated and deserves further attention and exploration.

One possible interpretation of the findings from Experiment 1 can be that participants have difficulties in the comparison of alternatives in the complex payoff condition and cannot come out with clear preferences. In the simple payoff condition, outcome comparison is simpler and participants can chose their move in a similar way as when the move of their opponent is known.

In the second experiment, the participants were told that the moves of the computer opponent had been made before the game session. Such a manipulation hasn't been used before. This manipulation led again to the disappearance of the disjunction effect even in the complex payoff condition. We interpret this result as a suppressing of a possible quasi-magic reasoning by stressing the fact that participants own moves cannot influence the move of the opponent. Interestingly, the manipulation led also to considerable decrease in the payoff processing time which deserves further exploration.

The results from the two experiments point to information processing complexity as a major factor for the disjunction effect contrary to the conclusions in previous research.

The eye-movement data support the complexity explanation described above. They show a change in the dynamics of the information acquisition in relation to the experimental manipulations of the complexity of the payoffs, and of the information about the opponent's move. As has been suggested in Busemeyer et al. (1993), the deliberation process can play a crucial role in decision making, especially when participants cannot attend at once to the full information available but can compare alternatives based on selected features.

A further systematic experimental and theoretical investigation of the results presented in this paper is under way and will be presented in the future.

References

Busemeyer, J. R., & Townsend, J. T. (1993). Decision Field Theory: A dynamic cognition approach to decision making. *Psychological Review*, *100*, 432-459.

Busemeyer, J. R. , Matthew, M., & Wang, Z. (2006). A Quantum Information Processing Theory Explanation of Disjunction Effects. Proceedings of the Cognitive Science Society.

Croson, R. (1999). The disjunction effect and reason-based choice in games. *Organizational Behavior and Human Decision Processes, 80 (2),* 118-133.

Grinberg, M., & Hristova, E. (2009). SARL: A Computational Reinforcement Learning Model with Selective Attention. In N. A. Tatgen & H. van Rijn (Eds.), *Proceedings of the 31th Annual Conference of the Cognitive Science Society* (pp. 821-826). Austin, TX: Cognitive Science Society.

Hristova, E., & Grinberg, M. (2008). Disjunction effect in prisoner's dilemma: Evidences from an eye-tracking study. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 1225-1230). Austin, TX: Cognitive Science Society.

Li, S., Taplin, J., & Zhang, Y. (2007) The equate-to-differentiate's way of seeing the prisoner's dilemma**.** *Information Sciences: an International Journal, 177(6),* 1395-1412.

Messe, L. A., Sivacek, J. M. (1975). Predictions of Others' Responses in a Mixed-Motive Game: Self-Justification or False Consensus? *Journal of Personality and Social Psychology*, *37(4)*, 602-607.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*, 372-422.

Rapoport, A., & Chammah, A. (1965). Prisoner's dilemma: a study in conflict and cooperation. Univ. of Michigan Press.

Shafir, E. & Tversky, A. (1992) Thinking through uncertainty: nonconsequential reasoning and choice. *Cognitive Psychology*, *24*, 449-474.

Tversky, A., & Shafir, E. (1992). The disjunction effect in choice under uncertainty. *Psychological Science*, *3*, 305-309.

Mentalizing in games: A subtractive behavioral study of Prisoner's Dilemma

Antonio Napoli (antonio.napoli@phd.units.it)
Danilo Fum (Fum@units.it)
Dipartimento di Psicologia, Università degli Studi di Trieste
via S. Anastasio, 12, I-34134 Trieste, Italy

Abstract

Economists and neuroscientists often explain game playing by assuming that humans try to predict the opponent's behavior on the basis of her past choices. We try to question this assumption in a Prisoner's Dilemma Game by using a methodology which we call the "subtractive behavioral method". Our aim is to investigate which task features make participants attend to the opponent's behavior or, on the contrary, make them take into account only their own choices and received payoffs. We find a critical effect of contextual information and we derive some suggestions about the methodology of brain imaging and behavioral game theory experiments.

Keywords: Game Theory; Brain Imaging; Theory of Mind; Social Dilemmas; Prisoner's Dilemma

Introduction

Game Theory (Von Neumann & Morgenstern, 1944) is a branch of applied mathematics focused on describing and predicting the behavior of "players" involved in strategic interactions in which the result of every player's "move" is contingent on the move(s) made by the other player(s). One of the critical assumptions of the theory is that games are played by completely rational agents whose strategies could be precisely calculated. In recent years the Game Theory formalism has been adopted to develop models that try to account for the fact that people often behave differently from what the theory predicts. This approach has been named "Behavioral Game Theory" (Camerer, 2003).

Behavioral Game Theory models make the assumption that people learn during the interaction, i.e., that they change their behavior according to the efficacy of their past choices. Among these models there are some, like those based on Reinforcement Learning (Erev & Roth, 1998; Sarin & Vahid, 2001), which take into account only the player's own choices and received payoffs while others, like so-called sophisticated (Camerer, Ho, & Chong, 2002) and belief learning (Cheung & Friedman, 1997) models, consider also (or only) the opponent's choices and payoff history. We will refer to the former as "partial information models" and to the latter as "full information models".

Even if Behavioral Game Theory does not make any assumption about the internal mechanisms involved in game playing, from a cognitive perspective it is possible to find a difference between partial information and full information models. Partial information models obey to a strictly behaviorist rule: the more you get from a choice, the more you will choose it in subsequent trials. These models completely ignore the opponent's behavior and only manipulate representations about chosen moves and obtained payoffs. They may also be applied to situations of playing without opponents (one-person games); in fact, they have been proposed by Sutton and Barto (1998) to model the performance in multi-armed bandit tasks in which participants make repeated choices among different options which are followed by a numerical reward that depends on the choice being made.

On the other hand, full information models manipulate representations about the opponents' moves and payoffs to anticipate their behavior and obtain thus a strategic advantage. These models address the opponent's beliefs, intentions, and strategies, and therefore mimic a Theory of Mind (henceforth: ToM) or "mentalizing" mechanism.

Neuroscientist have recently begun to study the cortical circuits involved in game playing through neuroimaging. Krueger, Grafman, and McCabe (2008), after reviewing the literature on the topic, propose that two cognitive mechanisms are specifically involved in game playing.

The first one is a "shared affect system" located in the Anterior Insula. This area is only activated in non-zero sum games in which cooperation between players is possible, and therefore feelings of trust, reciprocity and collaboration could be developed. The area seems responsible of two main effects: it makes people feel disgust towards uncooperative behavior and react to it (for example, rejecting unfair offers in a Ultimatum Game: Sanfey, Rilling, Aronson, Nystrom, & Cohen, 2003) and it makes people reciprocate by distinguishing between cooperative and non-cooperative opponents (Singer, Kiebel, Winston, Dolan, & Frith, 2004).

The second mechanism is a "shared intentions system", which is located in the Medial Prefrontal Cortex (MPFC). This area is activated both in zero and non-zero sum games, because it has the function of representing the opponent's beliefs, desires, and intentions, i.e. it seems to constitute the neural substrate of the ToM. Several brain imaging studies (see Krueger et al., 2008, for a comprehensive review) have shown MPFC activation during game playing and, therefore, it seems plausible that people mentalize while playing these games.

There are two other circuits which are not specifically involved in game playing but seem to be engaged in all kinds of learning tasks: a reward-based mechanism situated in a broad network of cortical and subcortical areas (see Lee, 2005 for a review), and a system concerned with the

prediction of complex behavior independently of its source, which is located in the Posterior Superior Temporal Sulcus (Frith & Frith, 2003).

Studies about mentalizing in game playing usually rely on the comparison between a condition in which people play against a computer and one in which they play against a human opponent on the presumption that mentalizing could be promoted by the latter. However, it is not clear whether and when people adopt a "mentalizing stance" and which task features could promote this activity. In fact, some studies show that a computer opponent could elicit activity in MPFC (Rilling, Sanfey, Aronson, Nystrom, & Cohen, 2004), while others claim that not all game situations against humans make people mentalize (Sally, 2003).

It is also unclear how mentalizing affects behavior, or, in other words, how decision making is affected by a ToM. For example Hill, Sally and Frith (2004) report that autistic adults behave in the same way as healthy participants in the Prisoner's Dilemma game, even if the autistic participants are severely impaired in other ToM tasks. Also, most neuroimaging studies lack a comparison between participant's behavior while playing against a human and a computer opponent.

We are convinced that the study of the ToM mechanisms would benefit from experiments which analyze participant's behavior. Two questions are important to us: 1) Which task feature make people mentalize? 2) Which effect does mentalizing have on people's behavior? In the present paper we try to address the first question by investigating some of the task features which could promote mentalizing during game playing.

Previous work

We have already started to explore the behavioral effects of mentalizing (Napoli & Fum, 2009) in playing a computer version of Rock, Papers, and Scissors (henceforth: RPS).

We had three groups of participants play 100 turns of RPS. In the first group, the computer was presented as an opponent, and the game was explicitly described as RPS. In the second group, the computer was presented as a neutral device. Participants saw three geometric figures which they should choose among at each trial; they received a payoff after each choice, and they could see the payoffs they could have obtained by making the alternative choices. Thus, this condition was equivalent to a multi-armed bandit task (Sutton & Barto, 1998) with the indication of foregone payoffs. In the third group, the computer was presented as an opponent. The game was played with the same rules of RPS but the choices were represented by geometric figures and the hierarchy of the moves (what beats what) had to be discovered during the game. This condition served as a control for the effect of the knowledge of the payoff matrix. The algorithm which assigned the payoffs was the same in all groups; the conditions differed therefore only for the setting induced in the participants (and the user interface).

We did not find any behavioral difference between the conditions, and we were able to model the behavior of all the groups by using a Reinforcement Learning algorithm based on ACT-R's utility learning mechanism (Anderson, 2007). This corroborates the idea that people did not use any information about foregone payoffs in the second condition and did not use any information about the opponent's moves or payoffs in the first and third condition. In summary, participants did not seem to mentalize at all during the experiment.

There are many possible explanations for this "failure to mentalize". Maybe people did not mentalize because they played against a computer; maybe they did not mentalize because the game was a mixed-strategy equilibrium game in which no move was better than the others and a simple behaviorist strategy could efficiently cope with the game; maybe people did not mentalize because no cooperation was possible in playing a competitive game. Or it may be a combination of all the three.

In this paper we try to clarify the findings of our previous work by making participants play a non-zero sum game, the Prisoner's Dilemma, both against what they believed was a human opponent and against a computer. Our aim is to understand which task features make people mentalize in game playing, which features affect game behavior and, possibly, why.

The experiment

Prisoner's Dilemma (henceforth: PD) is a non-zero sum game which has been extensively studied in psychology (Rapoport & Mowshowitz, 1966), classical game theory (Bo, 2005), behavioral game theory (Camerer, 2003), and neuroimaging studies (Singer et al., 2004). The payoff matrix used in our experiment is presented in table 1.

Table 1: Our experiment's payoff matrix

	Cooperate	**Defect**
Cooperate	60 60	100 0
Defect	0 100	20 20

PD can be thought of as a paradigmatic situation for any social dilemma in which the selfish interest contrasts with the common one. Classical game theory states that, independently of the choice made by the opponent, the most rational move for a player is to defect. In fact, if the opponent chooses to cooperate, defection gets 100 points and cooperation only 60 while, if the opponent defects, defection gets 20 points and cooperation 0. The result is that the optimal strategy for both people is to defect.

The most intriguing aspect of this game is that, even if the most rational move is defection, experiments show a substantial amount of cooperation between the players when the game is played in the iterated version (Bo, 2005). Another finding is that players learn to cooperate more and more during the experiment (Rapoport & Mowshowitz, 1966).

In order to understand what makes people mentalize, we adopted a "subtractive behavioral method" by assigning people to four different conditions in a repeated PD decision-making task in which the points earned by the participants were converted into play money.

The conditions differed according to the task features present in them which are summarized in Table 2.

Table 2: Features present in the experimental conditions

	Conditions			
Features	**N**	**CB**	**HB**	**HPD**
Repeated decision making	Y	Y	Y	Y
Opponent	N	Y	Y	Y
Believed Human Interaction	N	N	Y	Y
Explicit social scenario	N	N	N	Y

In the first condition, named "Nature" (N), participants played the PD disguised as a binary decision task: in each trial they had to choose between two options receiving a reward after each choice. It should be noted that in this condition the PD is presented as a repeated decision making one-person game, or a game against nature, in which no opponent is involved.

In the second condition, named "Computer Bet" (CB) participants were told that they would play a game against the computer. The instructions, however, presented the PD as a betting task: in each trial, the participants and the computer should bet on one of two alternatives and, depending on the combination of their choices, they would receive a given reward.

The third condition, named "Human Bet" (HB), was similar to the previous one (CB) except for the fact that participants were made to believe that they would play against a human opponent while in fact they were engaged by the computer.

In the fourth condition, named "Human Prisoner's Dilemma" (HPD), participants played PD against what they believed was a human opponent, just as in CB condition. There was, however, a substantial difference in the instructions provided for this condition and the two betting ones: the game was introduced by a story which illustrated a classical PD scenario (see Procedure for more details) and the two choices were labeled as "Cooperate" and "Defect".

In CB, HB and HPD conditions the instructions stressed that the goal of the participants was to gain as much money as possible independently of the money gained by the opponent, and that their opponent had the same objective.

According to results of neuroimaging research discussed in the Introduction, there are four cognitive processes which may influence participants' behavior in this task: the reward-based system, the complex behavior detecting system, the shared intentions system, and the shared affect system.

It is known that the reward-based system plays a role both in individual learning tasks and in game playing (Lee, 2005) by integrating the information received during the task in order to calculate the expected utility of different choices. Thus, this system should be active in all conditions, because of the repeated nature of the task.

It has been shown that the complex behavior detecting system is active during game playing against both computer and human opponents (Gallagher, Jack, Roepstorff, & Frith, 2002; Haruno & Kawato, 2009), and thus it should be activated in all conditions except Nature.

The shared intentions system is the main concern of this article. This area is always activated during game playing against humans, but it has been shown to be activated also during game playing against computer opponents, even if it is unclear which effect it exerts on people's behavior. If we find any difference between the CB and HB conditions, we can argue that mentalizing has a behavioral effect only in the case of a human opponent.

Finally, the shared affects system has been shown to be active when game playing involves the possibility of pro-social behavior, reciprocity, or fairness, and therefore we expect it could influence people's behavior only in the HPD condition. In this case the instructions promote empathizing with the opponent both because of the explicit social scenario and because of the labels attributed to the choices, which have a strong moral connotation. Therefore, every difference between the HB and HPD conditions should be attributed to this system.

Method

Participants and design. Sixty-four students (38 males) enrolled at the University of Trieste, Italy, were recruited as participants. Their age varied between 18 and 29 years (M=21.2, SD=3.4). Participants played two PD rounds, each one against a different algorithm (see below) whose order was counterbalanced between rounds. The experiment followed therefore a 4x2 mixed design with Setting as between-subjects and Algorithm as within-subjects factors.

Materials. Two algorithms were used in the experiment. The first one, Tit for Tat, cooperated in the first interaction and then replicated the opponent's previous choice. The second one, named Biased, made his moves by randomly sampling from a distribution of 60% Cooperate and 40% Defect moves.

Procedure. The experimental sessions were held in groups of 10-12 participants convened in a computer laboratory. Each participant was randomly assigned to one of the four conditions taking care that participants assigned to the same condition were not sitting next to each other. Participants were told that they would play different versions of the same game and received the instruction according to the condition to which they were assigned. Then, they were engaged in two PD rounds lasting eight minutes each.

The interface was kept as similar as possible in the four conditions. Participants made their choices by clicking on one of two circles displayed in the upper part of the screen. After a random lag time, in the Nature condition participants received a feedback about the money gained in the trial,

while in the other conditions they received a feedback about the opponent's choice, the money gained by themselves and by the opponent. The length of a bar representing their running total was updated and they were allowed to make another choice. In all conditions the two circles were labeled as "Yellow" and "Blue" except for the HPD condition, in which they were named as "Cooperate" and "Defect".

The main differences between the conditions relied in the amount of information and the kind of instructions provided to participants. In the N condition it was stated that they would play a binary decision task. After the first round participants were told that the computer would change the rule according to which it assigned the money. In the other three conditions participants had the payoff matrix in front of them from the beginning of the game. In the CB condition instructions stated that they would play a betting game with the computer, and after the first round they were told that the computer would change its strategy. In the HB and HPD condition participants were told they would play the game with one of the other participants in the room, and that the opponent would change after the first round. In the HB condition the task was presented as a betting game while in the HPD condition the game was introduced through a bargaining scenario in which Cooperate meant to respect the contract by delivering the promised goods and valuable money, respectively, while Defect meant to give the other player an empty bag. The instructions explicitly underlined this aspects of moral obligation and contract infringement involved in the game.

All groups played against the same algorithms with the Yellow and Blue circles equated to Defect and Cooperate, respectively.

At the end of the experiment we had informal interviews with the participants to assess the possibility that they had some doubts about having played against a computer and not a mate. Subjects who reported doubts were discarded from data analysis. Finally, a collective debriefing session ensued in which the nature of the opponent was discovered to all participants and the reasons for always adopting a computer as opponent were explained.

Results

Since the experiment was self-paced, participants made a variable number of choices in each round. To perform statistic analyses, we took into account their first 50 moves only.

Analysis of Cooperations. Being interested in the quality of participant's behavior more than in their ability to exploit the opponent's algorithm, we concentrated the analysis on the number of Cooperate moves and not on the amount of money gained.

First, we looked for possible differences between the first and second round in order to control for effects of learning (or fatigue). A mixed design ANOVA between the Round and the Setting did not reveal any significant effect for the Round (p=.55) or interaction (p=.93), while there was a significant effect of the Setting (F(3,58)=10.1, p < .001).

We then analyzed the factors manipulated in the experiment. A mixed design ANOVA revealed a significant effect of Setting and Algorithm (F(3,58)=10.1, p<.001 and F(1,58)=93.14, p<.001 respectively), while the interaction was not significant (p=.92). Table 3 reports Means and Standard Deviations of the participants' total Cooperate moves.

Table 3: Means (and Standard Deviations) of Cooperate per Algorithm in the various Settings

	Setting			
Algorithm	N	CB	HB	HPD
Biased	**15.69** (5.41)	**11.18** (7.7)	**14.23** (10.03)	**24.24** (7.09)
TFT	**32.6** (9.93)	**26** (10.65)	**28.8** (17.2)	**41.35** (10.6)

Algorithm and Setting seem to have an addictive effect in promoting cooperation between participants. While it is evident that the TFT algorithm promotes Cooperation more than the Biased one, it is unclear how Settings exerted its effect. Since there was no main effect of Round and no interaction between Algorithm and Setting, we analyzed separately the participant's performance against the two algorithms.

Two separate one-way ANOVAs for Biased and TFT Algorithms were performed. Both showed a significant effect for Setting (F(3,58)=8.95, p<.001 and F(3,58)=4.94, p<.01 respectively). The probabilities associated with post-hoc Newman-Keuls tests to contrast each Setting condition with the others are summarized in tables 4 and 5. For both algorithms a significant difference was found between the HPD and the other three conditions which, on the other hand, did not differ from each other.

Table 4: Probabilities for Post-hoc Newman-Keuls tests for the Biased Algorithm

	N	**CB**	**HB**	**HPD**
N		.24	.6	.0029*
CB	.24		.27	.002*
HB	.6	.27		.0017*
HPD	.0029*	.002*	.0017*	

* = significant

Table 5: Probabilities for Post-hoc Newman-Keuls tests for the TFT Algorithm

	N	**CB**	**HB**	**HPD**
N		.29	.39	.051**
CB	.29		.52	.016*
HB	.39	.52		.0049*
HPD	.051**	.0049*	.016*	

*= significant **=marginally significant

Analysis of conditional probabilities. We ran another analysis in order to understand why there was a difference in the number of Cooperate moves in HPD condition. This analysis was proposed by Rapoport and Mowshowitz (1966) and was also utilized by Erev and Roth (2001) in order to assess the efficacy of their reinforcement learning model.

Rapoport and Mowshowitz analyzed the probability of cooperation in a given trial according to the choices made in the *previous* trial by *both* players. Thus, a participant's strategy can be described by four numbers, C|CC, C|CD, C|DC, and C|DD. In the N condition, these probabilities may be interpreted as an analysis of a "win stay / lose switch" behavior. We can assume that, after a few choices, people get acquainted with the payoffs associated with the various options. Thus, for example, C|CC would be the probability of making the Cooperate/Blue move after receiving the best reward associated with that choice; therefore, a high value of this parameter would be an expression of a "win stay" strategy.

We analyzed the four conditional probabilities separately for the two algorithms to search for possible different strategies used in the different Settings. We ran a total of eight one-way ANOVAs analysis and all post-hoc Newman-Keuls tests for the significant ones.

We found a significant difference in three ANOVAs: C|DC both in the Biased ($F(3,58)=7.94$, $p<.001$) and in the TFT condition ($F(3,58)=5.21$, $p<.005$) and C|CC in the TFT condition ($F(3,54)=4.73$, $p<.006$). Newman-Keuls post-hoc tests showed that: in C|DC / Biased, HPD was different from all the other conditions ($p<.001$ in all cases), which were similar between them; in C|DC / TFT, HPD was different from CB and HB ($p<.001$ in both cases) and only marginally significant respect to N ($p=.055$), and the other three conditions were similar between them; in C|CC / TFT, the only significant difference was between HPD and CB $p<.001$.

Discussion and conclusions

In the experiment, participants played against an algorithm, the Biased one, that chooses its moves by sampling randomly from a given distribution, i.e., independently from the move made by the opponent, and against another algorithm, the TFT, that cooperates only if the opponent cooperated in the previous trial and defects otherwise. This means that the most rewarding strategy for participants was to Defect against the Biased algorithm—in order to exploit the trials in which it cooperates and to defend against the possibility of being exploited when the algorithm defects—and to Cooperate against the TFT—in order to initiate and maintain a virtuous reciprocation loop. The statistical analyses demonstrated that participants made more Cooperate moves against the TFT than against the Biased algorithm, i.e., that they were successful in adapting their strategy to the strategy used by the opponent.

However, we also found some differences between the groups: in the HPD condition participants made a higher number of Cooperate moves against both algorithms. The conditional probability analysis showed that this difference could be explained by the higher rate of C|DC in both cases. Since the only difference between the HPD and the other groups relied in the use, in the former case, of instructions that explicitly underlined the aspects of moral obligation and contract infringement involved in the game, the most natural conclusion is that this feature made people more prone to regret their defection against a cooperative opponent in the previous trial leading thus to more frequent cooperative behavior.

Interpreting the behavioral results in terms of the cognitive systems framework introduced above, we could safely assume an influence on this task of the reward-based system, being the participants capable of successfully adapting their strategy to the opponent in all conditions. However, we cannot exclude that such a performance could reflect the activation of the complex behavior prediction system too, being the activation of this system not selectively associated with strategic interactions (Frith & Frith, 2003). As for the shared affect system, it could have played a role in both human conditions (HB and HPD). In fact, during the debriefing interviews, some HB participants spontaneously told us about their willingness to cooperate with the opponent, a behavior that is typically associated with the activation of this system (Singer et al., 2004). However it is unlikely that this system played a critical role in the HB group, whose performance was similar to that of the N and CB condition where it is not credible that people could empathize with a computer, being it an opponent or not. Therefore, this system could be active only in the HPD condition.

As for the ToM system, we can exclude that it influenced the participant's behavior in CB and HB groups, which was similar to that of the N group. Therefore, we are left with two systems (ToM and empathizing) as responsible for the difference found in the HPD condition. Because brain imaging studies show that playing against a human opponent activates ToM areas regardless of the specific game (see for example Gallagher et al., 2002) and because, according to the participant's reports, it seems likely that they did in fact mentalize, we think that this area was active in both situations, and suggest two possible explanations for our results: (1) ToM had no behavioral effect in HB situation or (2) ToM had no effect both in the CB and HB conditions, and the difference between the two groups should be attributed to the shared affect system.

We won't take position with regard to this issue, because the limitations of our behavioral method don't permit us to. However we think that, whichever is the real explanation, this study makes some interesting points about both brain imaging and behavioral game theory experiments.

With regard to brain imaging studies, even if it has been shown that ToM areas are active in almost every game played against human opponents, it is not clear when they have a behavioral effect, too. We can speculate that there is some mechanism which prevents ToM from influencing the behavior in some situations. Otherwise, it would seem really

strange that it wouldn't have any effect on behavior *at all*. Therefore, we think that brain imaging studies should always take in account people's behavior, in a similar way to Haruno & Kawato (2009) and Hampton, Bossaerts, and O'Doherty (2008).

As for behavioral game theory, this paper makes a case for Erev and Roth's (2001) proposal of accounting people's behavior in Prisoner's Dilemma by the means of Reinforcement Learning. In fact, in N condition participants did not have any information about foregone payoffs, and nonetheless, their behavior was similar to the other groups. This means that the knowledge of payoff matrix and of the opponent's choices had a limited effect on participant's behavior. On the other side, the paper shows also the importance of contextual information—a variable which is seldom taken into account in game theory. In a more general sense, we think that our paper suggests the utility of having, along experiments in which people play one against the other, some more controlled sessions in which the participants play against an opponent (be it a computer or a human actor) whose strategy was under the control of the experimenter and compare them with individual learning sessions. This could make the experimenter safely exclude in most cases unnecessary believes or sophisticated learning.

References

Anderson, J. R. (2007). *How can the human mind occur in the physical universe?* New York: Oxford University Press.

Bo, P. (2005). Cooperation under the shadow of the future: experimental evidence from infinitely repeated games. *American Economic Review, 95,* 1591–1604.

Camerer, C. F. (2003). *Behavioral game theory: Experiments on strategic interaction.* Princeton: Princeton University Press.

Camerer, C. F., Ho, T., & Chong, K. (2002). Sophisticated Experience-Weighted Attraction learning and strategic teaching in repeated games. *Journal of Economic Theory, 104,* 137–8.

Cheung, Y., & Friedman, D. (1997). Individual learning in normal form games: some laboratory results. *Games and Economic Behavior, 19,* 46–76.

Erev, I. & Roth, A. (1998). Predicting how people play games: Reinforcement Learning in Experimental Games with Unique, Mixed Strategy Equilibria. *The American Economic Review, 88,* 848–881.

Erev, I. & Roth, A.E. (2001). On simple reinforcement learning models and reciprocation in the prisoner dilemma game. In Gigerenzer, G. and Selten, R. (Eds.), *The Adaptive Toolbox*. Cambridge, MA: MIT Press.

Frith, U., & Frith, C. D. (2003). Development and neurophysiology of mentalizing. *Philosophical Transactions of the Royal Society, 358*, 459–473.

Gallagher, H.L., Jack, A. I., Roepstorff, A., & Frith, C. D. (2002). Imaging the Intentional Stance in a Competitive Game. *Neuroimage, 16,* 814–821.

Hampton, A. N., Bossaerts, P., & O'Doherty, J. P. (2008). Neural correlates of mentalizing-related computations during strategic interactions in humans. *Procedings of the National Academy of Sciences USA, 105*, 6741 – 6746.

Haruno, M., & Kawato, M. (2009). Activity in the Superior Temporal Sulcus Highlights Learning Competence in an Interaction Game. *The Journal of Neuroscience, 29,* 4542–4547.

Hill, E. L., Sally, D. & Frith, U. (2004). Does mentalizing ability influence cooperative decision – making in a social dilemma? *Journal of Consciousness Studies, 11,* 144 – 161.

Krueger, F., Grafman, J., & McCabe, K. (2008). Neural correlates of economic game playing. *Philosophical Transactions of the Royal Society, 363*, 3859 – 3874.

Lee, D. (2005). Neural basis of quasi-rational decision-making. *Current Opinion in Neurobiology, 16,* 191–198.

Napoli, A., & Fum, D. (2009). Applying Occam's razor to paper (and rock and scissors, too): Why simpler models are sometimes better. *Proceedings of the 9th International Conference of Cognitive Modeling*, Manchester, United Kingdom.

Rapoport, A., & Mowshowitz, A. (1966). Experimental studies of stochastic models for the prisoner's dilemma. *Behavioral Science, 11,* 444–458.

Rilling, J. K., Sanfey, A. G., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2004). The neural correlates of theory of mind within interpersonal interactions. *Neuroimage, 22,*1694–1703.

Sally, D. (2003). Dressing the mind properly for the game. *Philosophical Transactions of the Royal Society of London B, 358,* 583 – 592.

Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. 2003 The neural basis of economic decision-making in the Ultimatum game. *Science, 300*, 1755–1758.

Sarin, R. & Vahid, F. (2001). Predicting how people play games: a simple dynamic model of choice. *Games and economic behavior, 34,* 104–22.

Singer, T., Kiebel, S. J., Winston, J. S., Dolan, R. J. & Frith, C. D. (2004). Brain responses to the acquired moral status of faces. *Neuron, 41*, 653–662.

Sutton, R. S., & Bartho, A. G. (1998). *Reinforcement learning: An introduction.* Cambridge, MA: MIT Press.

Von Neumann, J., & Morgenstern, O. (1944) *Theory of games and economic behaviour.* Princeton, N.J.: Princeton University Press.

Prospective Perception

J. Scott Jordan (jsjorda@ilstu.edu)
Department of Psychology, Illinois State University, Campus Box 4620
Normal, IL 61761

Jessica K. Witt (jkwitt@purdue.edu)
Department of Psychological Sciences, Purdue University
West Lafayette, IN 47907

Michael Riley (michael.riley@UC.Edu)
Department of Psychology, University of Cincinnati, 429-A Dyer Hall ML 0376
Cincinnati, OH 45221-0376

Abstract

In recent years, more and more data have come to fore that indicate perception to be inherently prospective (i.e., anticipatory). The purpose of the present symposium is to examine the research of three scholars who investigate prospective perception from three different theoretical perspectives: the Theory of Event Coding, the Action-Specific Perception account, and Ecological theory. Panelists will examine differences between theories and address the extent to which prospective perception affords a means of potentially integrating these theories.

Keywords: Prospective control; sensory-motor learning; ecological psychology; Action-Specific Perception account: affordances; anticipation.

What is Prospective Perception?

In recent years, more and more data have come to fore that indicate perception to be inherently prospective (i.e., anticipatory). The purpose of this symposium is to present research from three scholars who investigate prospective perception. Each will discuss the types of dependent variables they measure, the variables they manipulate, and the theoretical frameworks they use to interpret their data. Emphasis will be placed on differences and similarities between theories, as well as possible means of overcoming such differences. In the end, the panel will address what exactly it means for perception to be prospective, and how this might impact current theorizing in cognitive science.

The Theory of Event Coding

J. Scott Jordan is a cognitive psychologist who investigates the well known finding that the perceived vanishing point of a moving stimulus is displaced beyond the actual vanishing point, in the direction the stimulus was traveling just before it vanished (Hubbard, 2005). Numerous studies have revealed that the magnitude of this forward displacement (FD) varies systematically as a function of stimulus factors such as velocity (i.e., FD increases as stimulus velocity increases), movement direction (i.e., upward moving stimuli give rise to less FD than downward moving stimuli), and implied friction (i.e., FD decreases as a stimulus appears to move across a surface). Traditionally, such findings are accounted for in terms of representational-momentum, the idea being that the brain evolved to represent dynamic as well as static stimulus properties. Thus, when the moving stimulus vanishes, its representation entails momentum and continues moving, as it were, in the direction of represented motion. In a series of recent papers, Jordan and colleagues have researched an alternative account; namely, that FD is due to the anticipation underlying action control. This interpretation is based on the Theory of Event Coding (TEC; Hommel, Muessler, Aschersleben, & Prinz, 2001) which assumes the following: (1) actions are planned in terms of the distal effects they are to produce, and (2) action-planning and perception share overlapping neural dynamics. According to TEC therefore, FD occurs because the stimulus' movements are perceived in terms of the action plans participants generate as they interact with the stimulus. In one study (Jordan & Hunsinger, 2008) it was discovered that when participants simply observed the movements of the stimulus, FD was larger for observers who had just recently learned to control its movements via key presses on a computer keyboard. According to TEC, when observers were simply observing the movements of the stimulus, they were 'perceiving' those movements in terms of the plans they had learned while controlling it, due to the neural overlap of perception and action-planning. In short, perception entails plans, and these plans render perception inherently prospective.

Action-Specific Perception Account

Jessica Witt is a cognitive psychologist who also studies perception, and does so in terms of a framework known as the action-specific perception account. According to this framework, perception is scaled to the abilities and intentions of the perceiver. For example, when participants intend to reach with a tool to targets that are just beyond their reach, the targets look closer than they do when participants intend to reach without the tool or when the participants hold the tool but never intend to reach (Witt,

Proffitt, & Epstein, 2005). As another example, targets on the ground look farther away to participants who intend to throw a heavy ball to them compared with participants who intend to throw a light ball (Witt, Proffitt, & Epstein, 2004). However, after throwing a heavy ball, targets only look farther away for participants who intended to throw again, but not to participants who intended to walk (Witt et al., 2004). Only effort for the action-about-to-be-performed influences perception. In addition, as was reported in Jordan & Hunsinger (2008), performance of a task and the plans one generates during such performance, can influence later perception. For example, softball players who were hitting better selected a larger circle as matching the size of the softball used during the game (Witt & Proffitt, 2005), and golfers who are putting better select a larger circle as matching the size of the hole (Witt, Linkenauger, Bakdash, & Proffitt, 2008). This implies that better athletic performance led players to perceive the target as larger.

Collectively, these findings are consistent with the action-specific perception account, and support the assertion that perception is scaled relative to the behavioral possibilities of anticipated actions. Again, as was the case with Jordan, this implies that perception is inherently anticipatory.

Ecological Theory

While on the one hand, the notion that perception takes place in terms of behavioral possibilities might seem new to cognitive science, the idea has been a foundational concept in Ecological Psychology, where perception is argued to take place in terms of *affordances*. That is, ecological theory assumes we perceive the environment in terms of the behaviors it affords. Michael Riley is an ecological psychologist who studies affordance perception during action perception. That is, he and his colleagues investigate the patterns of environmentally-available information generated by body-object systems and the ways perceivers use such information. In one study (Ramenzoni, Riley, Shockley, & Davis, 2008) he and his colleagues asked both short and tall participants to indicate maximum overhead reaching capabilities for both themselves and another participant. The available perceptual information was manipulated by changing the participants' optically specified eye-height. Participants were able to accurately perceive the maximum overhead reach for both the 'self' and the 'other'. However, when the perceiver's eye-height was increased, the perceived maximum overhead reach increased for both judgments about both self and other. Riley and his colleagues interpret these results as revealing a rich source of environmentally-available information that perceivers use when perceiving action possibilities. Given these perceived possibilities refer to possible *future* behaviors, they are, by definition, prospective.

Discussion

Common to the Theory of Event Coding (TEC), the Action-Specific Perception account (ASPA), and Ecological Theory (ET), is the notion that perception is prospective. The theories do differ, however, with TEC focusing on overlapping neural structures, ASPA focusing on task specificity, and ET focusing on information structures available in the optic array. While these may seem to reduce to an internal versus external difference, the members of the panel will address the issue of whether or not this common notion of prospective perception might constitute a means of overcoming the computational-ecological divide that has plagued cognitive science for decades..

References

Hommel, B., Müsseler, J., Aschersleben, G. & Prinz, W. (2001). The theory of event coding (TEC): A framework for perception and action planning. *Behavioral and Brain Sciences, 24*, 849-937.

Hubbard, T. (2005). Representational momentum and related displacements in spatial memory: A review of the findings. *Psychonomic Bulletin & Review*, *12*, 822, 851.

Jordan, J. S., & Hunsinger, M. (2008). Learned patterns of action-effect anticipation contributes to the spatial displacement of continuously-moving stimuli. *Journal of Experimental Psychology*: *Human Perception and Performance*, *34*, 113–124

Ramenzoni, V. C., Riley, M. A., Shockley, K., & Davis, T. (2008). An information-based approach to action understanding. *Cognition*, *106*, 1059-1070.

Witt, J.K., Linkenauger, S.A., Bakdash, J.Z., Proffitt, D.R. (2008) Putting to a bigger hole: Golf performance relates to perceived size. *Psychonomic Bulletin and Review*, *15(3)*, 581-585.

Witt, J. K., & Proffitt, D. R. (2005). See the ball, hit the ball: Apparent ball size is correlated with batting average. *Psychological Science, 16,* 937–938.

Witt, J.K., Proffitt, D.R., & Epstein, W. (2004). Perceiving distance: A role of effort and intent. *Perception*, *33*, 570-590.

Witt, J. K., Proffitt, D. R., & Epstein, W. (2005). Tool use affects perceived distance, but only when you intend to use it. *Journal of Experimental Psychology: Human Perception and Performance, 31,* 880–888.

Constructing Spatial Concepts from Universal Primitives

Yang Xu* **and Charles Kemp†**
Machine Learning Department*
School of Computer Science*
Department of Psychology†
Carnegie Mellon University
{yx1@cs.cmu.edu, ckemp@cmu.edu}

Abstract

Spatial terms such as *on* and *in* are found in every language, and psychologists have suggested that the meanings of these terms may be constructed from a universal set of spatial primitives. We develop a computational version of this idea and explore whether the primitives typically proposed are sufficient to account for the meanings of spatial terms across languages. We compare a model where spatial terms correspond directly to primitives with models that represent spatial terms as discrete or weighted combinations of primitives. Our results suggest that combinations play an critical role, and we find limited evidence for weighted combinations.

Keywords: spatial cognition; cross-cultural; semantics; computational model.

Every documented language includes some machinery for describing spatial relationships. For example, an English speaker might say that the cup in Figure 1b is *on* the table and that the spoon is *under* the cloth. Spatial terms like these are acquired relatively early by children (Antell & Caron, 1985) and are used so frequently that they may come to seem unremarkable. Researchers have found, however, that it is surprisingly difficult to specify the meanings of spatial terms (Brown, 1994), and that different cultures make use of very different spatial concepts (Levinson & Meira, 2003; Levinson & David, 2006). This paper presents computational models that explore how spatial concepts might be constructed from more basic components, and that help to establish whether spatial concepts across cultures are constructed from a universal set of spatial primitives.

Many previous researchers have discussed the idea that spatial concepts might be constructed as combinations of primitive notions such as "support", "contact" and "containment". (Piaget & Inhelder, 1956; Jackendoff, 1983; Feist, 2000) For example, Figure 1b suggests that *on* in English may be roughly defined as the conjunction of "support" and "contact". Although this basic proposal is very familiar, there have been few sustained attempts to evaluate how well it can account for cross-linguistic data. Here we focus on primitives gathered from the existing literature and ask whether the distinctions that they capture are sufficient to account for spatial concepts across 25 different languages. Future work in this area can compare different sets of candidate primitives and compare how well they account for the data.

Any attempt to study semantic primitives must include some proposal about how these primitives combine to create spatial concepts. Here we compare proposals that vary along three dimensions. One of these dimensions specifies whether combinations of primitives are or are not allowed. A simple baseline approach assumes that every concept in every language corresponds to one of the semantic primitives, and we compare this approach to alternatives which assume that concepts correspond to combinations of primitives. In Figure 1b, for example, "on" is defined as the conjunction of support and contact. A second dimension specifies whether primitives are differentially weighted. In Figure 1b, all combinations are assumed to be conjunctions, and we compare this approach with an alternative that relies on weighted combinations. The final dimension specifies whether or not negations of primitives are allowed—for example, whether "no contact" is included in addition to "contact." Our three dimensions produce a collection of eight possible models, and we explore the five most interesting cases (Table 1). Comparing the performance of these models suggests that combinations of primitives are important, but we find only limited evidence for weighted combinations. None of the models we consider is rich enough to capture the true complexity of spatial cognition, but these simple models are a useful starting point for the computational approach that we advocate.

Our work is inspired in part by several recent studies of cross-cultural spatial cognition (Feist, 2000; Bowerman & Choi, 2001; Levinson & Meira, 2003; Feist, 2008; Khetarpal, Majid, & Regier, 2009). A consistent theme in the previous literature is that spatial concepts correspond to regions in some kind of similarity space. To mention just two examples, Bowerman and Choi (2001) suggest that scenes described using "on" and "in" by English speakers can be arranged along a similarity gradient, and that different languages carve up this similarity space in different ways. Levinson and Meira (2003) propose that spatial terms correspond to attractors in a similarity space, and use multidimensional scaling to support their proposal. Approaches like these have helped to illuminate the basis of spatial cognition, but they rely on a notion of similarity that is rarely made precise, and are unable to explain exactly how humans recognize similarities between spatial configurations. Our work is compatible with many of the insights that have emerged from these previous approaches, and could be viewed as an attempt to ground the notion of similarity in terms of concrete spatial primitives. We prefer, however, to treat similarity as an epiphenomenon, and expect that similarity will play no explanatory role once the building blocks of spatial concepts are understood.

We begin by introducing the semantic primitives that we will consider and the cross-linguistic data that we will attempt to explain. We then evaluate five simple models which

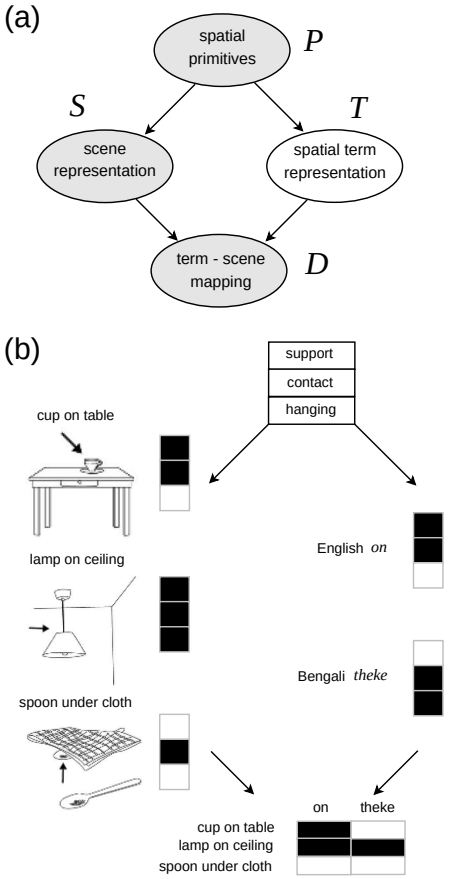


Figure 1: (a) A computational framework for exploring how spatial primitives (P) combine to create the meanings of spatial terms (T). Given information about which primitives characterize a set of scenes (S), the framework predicts which terms apply to which scenes. (b) An illustration of the framework in (a). English "on" is a combination of "support" and "contact," and applies to scenes (like cup on table) where both primitives are present.

make different assumptions about how spatial concepts are constructed from semantic primitives. Each successive model includes one or more previous models as a special case, and we explore whether the additional assumptions made by each model help to account for the cross-linguistic data.

A Computational Approach to Spatial Cognition

Our formal approach is summarized by the graphical model in Figure 1a. Suppose that P represents a set of spatial primitives and that S is a matrix of scene vectors, where column s_i is a binary vector that indicates which primitives apply to scene i. In Figure 1b, for example, the scene vector for "cup on table" indicates that this scene is characterized by "support" and "contact" but not "hanging." Let T be a matrix of term vectors, where vector t_j indicates which primitives contribute to the meaning of term j. In Figure 1b, the term vector for "on" indicates that the meaning of this term is based on the "support" and "contact" primitives. Finally, let D be a binary matrix where entry d_{ij} indicates whether the spatial relationship in scene i can be described by term j.

Table 1: A brief description of the five models and their abbreviations. The two columns on the right compare model scores on the real data to the mean scores on the random sets discussed in Results. D_1 is data from the authors and Levinson and Meira (2003). D_2 is data collected by Feist (2000).

Model	Abbrev.	S(D_1)	S(D_2)
Singleton	BS+	.61 : .39	.61 : .50
Singleton with negations	BS−	.62 : .41	.66 : .53
Conjunction	BC+	.66 : .46	.70 : .58
Conjunction with negations	BC−	.79 : .57	.83 : .68
Weighted combination	WC−	.79 : .54	.80 : .65

The graphical model in Figure 1a can capture at least three kinds of inferences. If asked to decide whether term j applies to scene i, a native speaker can use scene vector s_i and term vector t_j to decide whether $d_{ij} = 1$. When interpreting a description of an unobserved scene i , a native speaker can use term vector t_j along with the information that $d_{ij} = 1$ to predict the scene vector s_i. When learning the meanings of spatial terms, a learner given P, S, and D can infer the term vectors in T. We will address this third problem and the nodes for P, S, and D are shaded in Figure 1a to indicate that these variables are observed for all cases we consider.

We report results for two cross-linguistic data sets. The first is based on a triple (P_1, S_1, D_1) that combines data reported by Levinson and Meira (2003) with new data that we have collected. Our second data set is based on a triple (P_2, S_2, D_2) that is taken from the work of Feist (2000). The next sections describe these triples, and we then describe how we used these triples to explore the meanings of spatial terms.

Spatial primitives. The first set of primitives (P_1) is shown in Table 2, and includes 19 primitives that capture position along the vertical axis, position with respect to the observer, and various notions related to contact and inclusion. These primitives were collected from several previous authors, and the set is intended to capture most of the concepts that have previously been proposed as candidate primitives. The second set of primitives (P_2) is based on a set proposed by Feist (2000), and includes primitives like "above," "contact," and "support." The complete set of primitives is shown at the top left of Figure 2b.

Scenes and scene vectors. The scenes we consider are taken from the *Topological Relations Picture Series* designed by Melissa Bowerman. This picture set is composed of 71 different line drawings of a wide range of spatial scenes. Each scene in the picture set represents a spatial relationship between a designated *figure* (indicated by an arrow in the drawing) and a *ground* object. Figure 1 shows a few examples of these drawings. Scene matrix S_1 includes all 71 pictures. We asked three English speakers to code these pictures using the

19 primitives in Table 2. Each primitive was described using a short phrase, and summaries of these descriptions are shown in Table 2. Matrix S_1 was created by merging the three sets of responses using a majority vote, and a subset of this matrix appears in Figure 2a. Scene matrix S_2 includes information for 27 scenes from the picture series. Feist coded each scene in terms of the primitives in her set, and matrix S_2 is based on her codes. A subset of S_2 is shown in Figure 2b.

Scene-term mappings. Matrix D_1 includes results for all 71 scenes. Levinson and Meira (2003) reported data for 4 languages, and we built on this data set by asking one speaker for each of 21 additional languages to label the set of 71 scenes. The languages included are listed in Table 2. Participants were asked to provide a single spatial term for each picture and were allowed to use as many different terms as they liked across the set of 71 scenes. In cases where they were not sure, we asked them to choose the term that seemed best to them. Feist (2000) asked speakers of 16 languages to label the scenes represented in S_2, and the results are collected in data matrix D_2.

Modeling the meaning of spatial terms

The information in a triple (P,S,D) can be used to explore the semantics of spatial terms. We consider a family of five models that make different assumptions about the spatial term representations T and the way in which scene representations (S) and term representations (T) combine to generate the term-scene mappings (D). All of the models assume that spatial term j is represented as a term vector t_j, but the models vary along three dimensions which determine the nature of the entries in each vector.

One of these dimensions—binary (B) or weighted (W)—indicates whether primitives can be differentially weighted. Binary models use term vectors t_j where 0 indicates that a primitive makes no contribution to the meaning of t_j, and 1 indicates that a primitive must be present in order for term j to apply. Weighted models use vectors where each entry is a real number between -1 and 1 inclusive. Weights near 1 indicate that a primitive should be present in order for a term to apply, and weights near -1 indicate that a primitive should be absent. A second dimension—singleton (S) or combination (C)—indicates whether terms correspond to single primitives or combinations of primitives. Singleton models assume that each term vector has exactly one non-zero entry, but combination models allow term vectors to have multiple non-zero entries. The final dimension—positive (+) or negative (-)—indicates whether spatial terms can be defined using negations of primitives. For binary models with negation, we expand the set of primitives so that it includes negated versions of each primitive in Table 2. For weighted models with negation, we keep the original set of primitives and capture negation by allowing term vectors to include negative weights. The three dimensions just introduced generate 8 models in total, and we will focus on the five models in Table 1.

Although some of our models allow term vectors t_j to contain real-valued entries, scene vectors s_i are always represented as binary vectors which specify which primitives apply (1) or do not apply (0) to each scene. Given a scene vector s_i and a term vector t_j, all of our binary models determine whether spatial term j applies to scene i as follows:

$$d_{ij} = \begin{cases} 1, & \text{if } {s_i}^T t_j = |t_j| \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

where $|t_j|$ is the number of non-zero entries in term vector t_j. Equation 1 states that term i applies to scene j (i.e. $d_{ij} = 1$) only if all of the constraints specified by term vector t_j are consistent with the scene. Weighted models use a soft version of Equation 1:

$$d_{ij} = \begin{cases} 1, & \text{if } \sigma({s_i}^T t_j) > p \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

where $\sigma(\cdot)$ is a sigmoid function (e.g. $\sigma(x) = \frac{1}{1+exp(-x)}$) which maps its argument into a probability (i.e. a number between 0 and 1). The parameter p is a threshold that will be learned from the data sets that we consider.

The models in Table 1 make contact with previous ideas from several fields. The singleton model is based on an idea proposed by Piaget and Inhelder (1956) who claims that there exists a common topology in which spatial languages build on concepts such as proximity and contiguity. Jackendoff (1983) further suggests that spatial semantics are composed of simple primitives such as "on" and "in", which are directly encoded in languages. We expect, however, that the singleton model is unlikely to prove adequate. Levinson and others (Levinson & Meira, 2003; Levinson & David, 2006) have argued that there is great variation in spatial concepts across cultures, and the singleton model cannot account for this variation without an explosion in the number of primitives.

The combination models are also related to previous work. The discrete combination model captures the familiar proposal that meanings can be represented as conjunctions of primitive concepts, and psychologists have also proposed that spatial terms are represented as sets of weighted attributes (Feist, 2000). The weighted model in Equation 2 is known to statisticians as a logistic regression model, and is equivalent to a single-layer neural network, where the input (s_i) is mapped to the output (d_{ij}) via a layer of weights (t_j) and the sigmoid function.

Inferring term vectors

Our goals can now be precisely formulated. Given a triple (P,S,D) and one of the five models in Table 1, we wish to infer a term matrix T and decide how well S and T account for the data D. For both the singleton and conjunction models, we use a greedy algorithm to infer the term matrix T. For each spatial term we begin with a term vector t_j that includes only zeros, then greedily flip elements to improve a standard precision-recall F-score

$$F = \frac{2 \times \sum_i I(\hat{d}_{ij} = d_{ij} = 1)}{\sum_i I(\hat{d}_{ij} = 1) + \sum_i I(d_{ij} = 1)} \quad (3)$$

Table 2: Lists of author-collected languages (alphabetical), spatial primitives and their descriptions. "*" indicates negatable primitives. "F" and "G" stand for figure and ground.

Language	Primitive	Description
Arabic	above	F higher than G
Bengali	below	F lower than G
Cantonese	vertical equality*	F and G of equal height
Croatian	support*	F supported by G
English	horizontal support*	F supported horizontally by G
Finnish	front	F closer to viewer than G
French	back	G closer to viewer than F
German	viewpoint equality*	F and G equidistant from viewer
Hindi	contact*	F in touch with G
Indonesian	surface contact*	F in surface contact with G
Italian	attachment*	F attached to G
Japanese	adhesion*	F stuck to G
Mandarin	hanging*	F hung from G
Portuguese	piercing*	F pierces through G
Romanian	impaled*	F impaled by G
Russian	proximity*	F in close proximity to G
Slovakian	containment*	F contained by G
Slovene	encircled*	G circles F
Spanish	circlement*	F circles G
Thai		
Vietnamese		

where $\hat{d}_{ij}$ is a prediction based on the term vector t_j and d_{ij} indicates whether term j actually applies to scene i. The F-score will be high if most of the $\hat{d}_{ij} = 1$ entries predicted by t_j are correct (high precision), and if these predicted 1-entries include most of the actual 1-entries for term j (high recall).

For the weighted combination model, instead of inferring binary vectors we must learn a vector of weights for each term. Choosing the weights to maximize the F-Score is possible in principle (Jansche, 2005), but instead we fit a standard L1 regression model which is equivalent to a Bayesian logistic regression (Genkin, Lewis, & Madigan, 2004) with a Laplacian prior on the weights. For each spatial term, this approach searches for a weight vector t_j such that Equation 2 accurately predicts which scenes can be described by term j. The Laplacian prior captures the idea that term vectors t_j should be as simple as possible, and encourages small entries in t_j to end up as zero weights. In addition to this prior, we use the number of non-zero entries inferred by the conjunction model as an upper bound on the number of non-zero weights for the weighted model. Allowing many of the entries to be non-zero gives the weighted model more flexibility, but enforcing a sparsity constraint enables a direct comparison between the conjunction and weighted combination models. After learning the weights in all of term vectors t_j, we finish by choosing threshold p in Equation 2 to maximize the F-score (Equation 3).

Results

We applied the five models just described to the two triples (P,S,D) mentioned previously. In each case we computed the term matrix T that best accounts for the data. Term vectors for some languages are shown in Figure 2, and are discussed towards the end of this section.

The extent to which each model captures each data set can be captured using the F-score in Equation 3. Scores for the five models are shown in Table 1. To assess whether these scores are better than chance-level performance, we compared them with baseline scores achieved on random data sets. We used three randomization strategies. A *randomized D* set is created by randomizing all entries in D so that the sparsity is preserved (i.e. the number of "1" entries remains the same but all other structure is lost. A *shuffled D* set is created by randomly reordering the rows in D and leaving the scene vectors in S fixed. Finally, a *shuffled S* set is created by permuting the rows in S and leaving D fixed. Note that both shuffled sets leave the columns in D and S unchanged and therefore preserve many characteristics of these matrices, including the extent to which scenes (i.e. columns) tend to fall into clusters. For each triple, we created 20 random sets for each randomization strategy and computed the model scores. We then used t-tests to evaluate the hypothesis that performance on the real sets was significantly higher than performance on the random sets. In all cases we obtained highly significant results with truncated $p < 0.001$ after correction for multiple tests (first five rows of Table 2). These results suggest that all of our models were able to capture the structure in the observed data better than chance.

Although all models appear to capture some structure in the data, it is natural to ask which model performs best. The scores for the individual models do not address this question directly—for example, since the singleton model is a special case of the conjunction model, the conjunction model will always achieve a higher score regardless of whether it is actually the better approach. We therefore compared pairs of models by exploring whether whether the difference between their scores was significantly above chance level. For each pair, we compared the difference in prediction scores on the real data set against the differences achieved on the three random sets. The results appear in the final five rows of Table 2. Rows 6 and 7 suggest that the conjunction models perform better overall than the singleton models. Rows 8 and 9 suggest that allowing negated primitives leads to a significant improvement in performance. Finally, row 10 suggests that the weighted combination model does not perform better than the conjunction model with negations. Note, however, that we also evaluated an alternative weighted model where the sparsity of the weight vectors was not constrained by the conjunctive solution, and where all of the entries in each vector were allowed to be nonzero. This model performed significantly better than the conjunction with negation model on three of the six randomized tests across the two data sets, suggesting that weighted combinations may capture some aspects

Table 3: Significance of model performances and pairwise comparisons from t-tests. The model scores on the real data sets are compared to those on the random sets (1 – *randomized D*, 2 – *shuffled D*, 3 – *shuffled S*). D_1 is data from the authors and Levinson and Meira (2003). D_2 is collected by Feist (2000). For each pairwise comparison, the model on the left scores higher than the model on the right (e.g. BC+ outperforms BS+). '*' indicates statistical significance at $p < 0.05$.

	D_1			D_2		
	1	2	3	1	2	3
BS+	*	*	*	*	*	*
BS−	*	*	*	*	*	*
BC+	*	*	*	*	*	*
BC−	*	*	*	*	*	*
WC−	*	*	*	*	*	*
BC+ *vs* BS+	*	–	–	*	*	*
BC− *vs* BS−	*	*	–	*	*	*
BS− *vs* BS+	*	–	–	*	*	*
BC− *vs* BC+	*	*	*	*	*	*
BC− *vs* WC−	–	–	–	–	–	–

of spatial semantics. Future work can explore this issue in more detail and determine which sparsity assumptions allow weighted models to provide the best account of spatial terms.

Our analyses so far suggest that the primitives in P_1 and P_2 are able to account for much of the structure in data sets D_1 and D_2. It is important, however, to consider whether our models combine the primitives in psychologically meaningful ways. Figure 2 shows the definitions learned by our models for three languages, and focuses on a subset of 10 scenes that were used in both data sets. Figure 2a shows term vectors and predictions for our data set. Note that the conjunction model captures important aspects of meaning that the singleton model misses. For example, Figure 2a.ii shows that "contact" is included in the meaning of *on* by the conjunction but not the singleton model. The plots also illustrate how negations allow the conjunction model to improve its predictions. Figure 2a.xi shows that the conjunction model makes several predictions about "qian mian" that do not match the true scene-term mapping in Figure 2a.x. "Qian mian" corresponds roughly to the phrase "in front of," and including "no contact" in the definition of this term allows the negated conjunction model to successfully predict that it will not apply to scenes like "handle on cupboard" or "stamp on letter."

For our second analysis the term vectors T_1 (Figures 2b.iii and b.iv) can be compared against a gold standard, which is the set of term vectors manually assigned by Feist (Figure 2b.ii). The vectors learned by our model are similar to those specified by Feist, and the predictions that follow from Feist's representation (Figure 2b.vi) do not appear more accurate overall than the predictions generated by our automatically learned term vectors (Figure 2b.vii).

Conclusion

We presented computational models that explore whether spatial concepts can be constructed by combining a set of universal primitives. Our results suggest that a large proportion of the information in two cross-linguistic data sets can be captured by models that begin with the primitives typically discussed in the literature and combine them using simple operations such as conjunctions and weighted sums. Our general framework (Figure 1a) can be used to address many questions in spatial cognition and we mention just two directions for future work. First, we fit our models to the cross-linguistic data by learning definitions for each spatial term, and future work can use our approach to explore how humans learn spatial concepts. Second, all our analyses used primitives that were specified *a priori*, but it is conceptually straightforward to develop models that learn the primitives that best account for a given data set. Uncovering the nature of spatial primitives presents many challenges, but computational approaches can help address some of these challenges.

References

Antell, S. E. G., & Caron, A. J. (1985). Neonatal perception of spatial relationships. *Infant Behavior and Development*, *8*, 15-23.

Bowerman, M., & Choi, S. (2001). Shaping meanings for language: Universal and language-specific in the acquisition of spatial semantic categories. In M. Bowerman & S. C. Levinson (Eds.), *Language acquisition and conceptual development.* Cambridge: Cambridge University Press.

Brown, P. (1994). The INs and ONs of Tzeltal locative expressions: the semantics of stative descriptions of location. *Linguistics*, *32*, 743-90.

Feist, M. I. (2000). *On In and On: An investigation into the linguistic encoding of spatial scenes*. Doctoral dissertation, Northwestern University.

Feist, M. I. (2008). Space between languages. *Cognitive Science*.

Genkin, A., Lewis, D., & Madigan, D. (2004). *Large-scale Bayesian logistic regression for text categorization* (Tech. Rep.). Rutgers University.

Jackendoff, R. (1983). *Semantics and cognition*. Cambridge, MA: MIT Press.

Jansche, M. (2005). Maximum expected F-measure training of logistic regression models. In *Proceedings of EMNLP*.

Khetarpal, N., Majid, A., & Regier, T. (2009). Spatial terms reflect near-optimal spatial categories. In *Proceedings of the 31st annal conference of the cognitive science society.*

Levinson, S. C., & David, W. P. (2006). *Grammars of space*. Cambridge University Press.

Levinson, S. C., & Meira, S. (2003). 'Natural concepts' in the spatial topological domain — adpositional meanings in crosslinguistic perspective: an exercise in semantic typology. *Language*, *79*, 485-516.

Piaget, J., & Inhelder, B. (1956). *The child's conception of space*. London: Routledge and Kegan Paul.

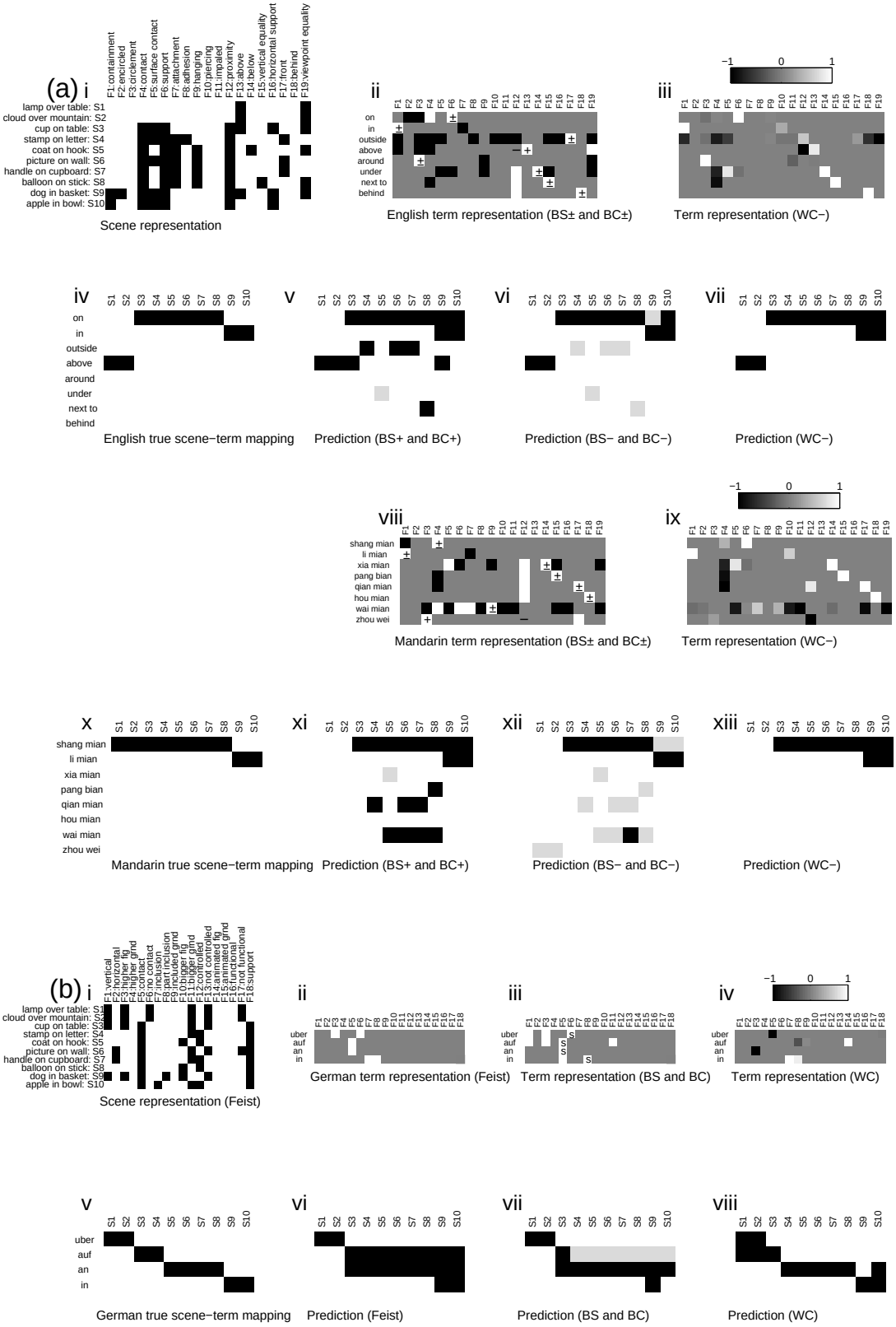


Figure 2: Term vectors and scene-term mappings for (a) English and Mandarin in the data set collected by the authors and (b) German in the data set collected by Feist. (a)(i) Ten scenes coded according to the nineteen primitives in Table 2. (ii) Inferred term vectors for four models: BS+ (indicated by +), BS- (-), BC+ (white cells) and BC- (white and black cells). Model BS- chooses a negated primitive only once (*above* is defined as "not F12"). (iii) Inferred term vectors for model WC-. (iv) True scene-term mappings (v) - (vii) Predicted scene mappings for five models. The predictions of models BC+ and BC- (black cells) are a subset of the predictions of the singleton models (black and gray cells). (viii)-(xiii) Results for Mandarin. (b) Results for German. Feist provided the encoding in (i) and the term vectors in (ii). (iii)-(iv) Term vectors for the singleton model ("S"), the conjunction model (white cells) and the weighted combination model. (v)-(viii) Actual and predicted scene-term mappings. Since the primitives in (b)(i) already include negations, models BS and BC do not allow additional negations.

Replicating Color Term Universals through Human Iterated Learning

Jing Xu (jing.xu@berkeley.edu)
Thomas L. Griffiths (tom_griffiths@berkeley.edu)
Department of Psychology, 3210 Tolman Hall
Berkeley, CA 94720 USA

Mike Dowman (Mike@ImageScope.net)
ImageScope

Abstract

In 1969, Berlin and Kay proposed that there exist cross-cultural universals in the form of basic color terms. To test this hypothesis, the World Color Survey (WCS) collected color naming data from 110 non-industrial societies, identifying regularities in the structure of languages with different numbers of terms. This leaves us with the question of where these universals come from. We use a simple model of cultural evolution known as "iterated learning" to explore the hypothesis that universals emerge from human perceptual and learning biases. We conducted an experiment simulating the process of cultural transmission in the laboratory, and compared the results to the systems of color terms that appear in the WCS data. Our results show that cultural evolution results in convergence of systems of color terms towards a form consistent with the WCS, supporting the hypothesis that universals are the result of perceptual and learning biases.

Keywords: basic color terms; iterated learning model; color term universals; cultural evolution; Bayesian inference.

Introduction

Linguistic universals – properties that seem to hold across all human languages – have the potential to provide unique insight into the nature of human cognition. Universals in systems of color terms are among the best documented of these properties. Berlin and Kay (1969) proposed that color naming systems across different cultures are based on one or more of eleven focal colors corresponding to the English color terms *black*, *white*, *red*, *green*, *yellow*, *blue*, *brown*, *purple*, *pink*, *orange*, and *gray*. Kay and McDaniel (1978) and Kay and Maffi (1999) later refined this model to emphasize the six Hering primary colors (*black*, *white*, *red*, *green*, *yellow*, and *blue*) (Hering, 1964), and to characterize the process by which societies might transition from one system of color terms to another as new terms are introduced.

The World Color Survey (WCS) was initiated in the late 1970's to provide a more comprehensive empirical test of the univerality hypothesis (Kay, Berlin, & Merrifield, 1991; Kay, Berlin, Maffi, & Merrifield, 1997). In the WCS, a total of 330 color chips, comprised of 40 equally spaced Munsell hues at 8 levels of lightness and achromatic chips at 10 levels of lightness (see Figure 1), were presented to speakers of 110 different languages in non-industrial societies. Those speakers were asked to name each color chip, and also to point out the most representative chip for each color term. Later analysis of the WCS data showed that the universality hypothesis was by and large confirmed (Kay et al., 1997). Recently, several statistical analyses of the WCS data have also been conducted in an attempt to resolve the debate over the universality of color naming. For example, Kay and Regier (2003; Regier, Kay, & Cook, 2005) showed that the focal colors in the WCS data largely fall in similar regions to those seen in English; in another study they defined a statistical measure of "well-formedness", and used this measure to show that observed systems of color terms correspond to a near-optimal partition of color space (Regier, Kay, & Khetarpal, 2007).

The consistent cross-linguistic structure highlighted by the WCS raises a new question: Where do these universals come from? They may be a result of cultural universals that may arise from the homogeneity of biological traits and evolutionary paths across cultures that constrain people to consider only a limited range of color categories when learning language, thus forcing color term systems to conform to a limited range of universal types (Hawkins, 1988). However, if we view language as a system culturally transmitted from generation to generation, a simpler hypothesis is that these universals may arise directly from biases that cause learners to prefer some color categorizations over others, but that do not place absolute constraints on the types of color categories that are learnable. One way to explore this hypothesis is using the *iterated learning* model, a simple model of cultural transmission in which a sequence of agents each learns from the behavior of the previous agent in the sequence (Kirby, 2001). In an iterated learning model of the transmission of systems of color terms, each agent learns a system of color terms from examples provided by another agent, and then generates examples which are provided to the next agent in the sequence. Mathematical analyses of iterated learning show that as this process continues, the information being transmitted gradually changes to become consistent with the learning biases of the agents involved (Griffiths & Kalish, 2007; Kirby, Dowman, & Griffiths, 2007). If systems of color terms similar to those seen in the WCS emerge from a process of cultural

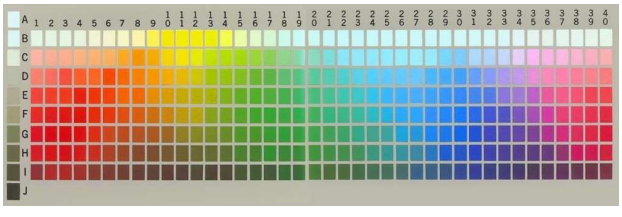


Figure 1: The World Color Survey stimulus array.

transmission by iterated learning, then the biases of individual learners may be sufficient to explain the regularities seen across human societies.

Previous work has used computer simulations to demonstrate that iterated learning with simulated agents can produce systems of color terms similar to those seen in the WCS (Dowman, 2007, 2009). In this paper, we test the hypothesis that color naming universals may be a result of the perceptual and learning biases of human learners by conducting a large-scale laboratory experiment based on iterated learning. In our experiment, human learners acquire and transmit novel systems of color terms, providing a human simulation of the process of cultural evolution. We examine how these systems of color terms change over time, comparing the results to the WCS. We show that, consistent with the hypothesis that perceptual and learning biases are the source of color-naming universals, the systems of color terms generated by our iterated learning chains converged over time to become more consistent with the WCS.

The plan of the paper is as follows. The next section provides further details of the iterated learning model and its predictions about the influence of learning biases on the outcome of cultural transmission. We then present our experiment, which used human learners to simulate the cultural transmission of systems of color terms. Analyzing the results of this experiment raises some technical challenges, which we address by introducing a novel method for quantifying the correspondence between the languages produced by our participants with those in the WCS. We conclude the paper by discussing the implications of our results, and consider some of the potential limitations of our analysis.

Iterated Learning

Much of human knowledge is not learned from the world directly, but from other people. When we learn languages, we learn them from the utterances of existing speakers, and our utterances inform the next generation of speakers. A simple way to model this process of cultural transmission is in terms of "iterated learning", as illustrated in Figure 2. We imagine a sequence of learners, each of whom observes data, forms a hypothesis about the process that produced those data, and then generates data for the next learner based on that hypothesis.

We can analyze the process of iterated learning by assuming that our learners are rational Bayesian agents. In this framework, learners come up with the *posterior* probability $P(h|d)$ of a hypothesis h given the observed data d by applying Bayes' rule,

$$P(h|d) = \frac{P(d|h)P(h)}{\sum_{h'}(d|h')P(h')} \quad (1)$$

where $P(d|h)$ is the *likelihood*, indicating the probability of observing d if h were true, and $P(h)$ is the *prior* probability, indicating the extent to which the learner was willing to accept h prior to observing d. The prior encodes the learner's inductive biases, being a factor that combines with the observed data to yield a conclusion.

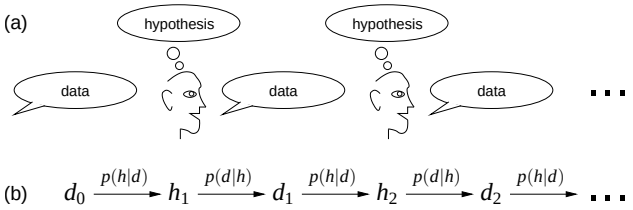


Figure 2: Iterated learning. (a) Each learner sees data produced by the previous generation, forms a hypothesis about the process by which those data were produced, and uses this hypothesis to produce the data that will be supplied to the next generation. (b) In iterated learning with Bayesian agents, each learner sees data d, and uses Bayes' rule to compute the posterior probability of each hypothesis h, $p(h|d)$. The learner samples a hypothesis from this distribution, and then generates data from the distribution $p(d|h)$.

In the cultural transmission process, data are passed along a chain of learners. Assume that the same Bayesian inference process happens repeatedly at each generation, with each learner sampling a hypothesis from their posterior distribution and generating data by sampling from the likelihood function associated with that hypothesis. This process can be analyzed as a Markov process: The probability each learner selects a particular hypothesis depends only on the data produced by the previous generation. Griffiths and Kalish (2005, 2007) showed that when learners share a common prior distribution, the probability a learner selects a hypothesis converges to the prior probability of that hypothesis as the process of iterated learning continues. Likewise, the probability of generating data d converges to the prior predictive distribution, being the average of the likelihood over the prior, $p(d) = \sum_h p(d|h)p(h)$.

The convergence of iterated learning to the prior potentially provides an explanation of linguistic universals, including universals in color naming. Languages are constantly being passed from speaker to speaker via a process of cultural transmission similar to iterated learning. If this process provides a way for perceptual and learning biases of the kind captured by a prior distribution to have an effect on the structure of languages, we should expect languages to demonstrate properties that are consistent with human biases. If this hypothesis is correct, we should expect to see systems of color terms transmitted via a process of iterated learning to change over time to resemble those that appear in the WCS. To test this idea, we ran a series of iterated-learning chains among an English-speaking population in our laboratory, comparing the systems of color terms produced by those chains to those seen in the WCS.

Color Term Transmission in the Lab

Methods

Participants Participants were 390 members of the community at the University of California, Berkeley, receiving either course credit or approximately $10/hr for taking part in the experiment.

Stimuli Each participant learned a system of color terms by being provided with examples of colors and the terms that were associated with them, and then generalized those terms to new colors. The color stimuli were presented on an Apple iMac computer by a Java program, and the monitor was calibrated using a ColorVision Spyder2 colorimeter/color calibrator on regular basis. A total of 330 colors were used as stimuli, corresponding to the computer screen analogues of the 330 Munsell color chips used in the WCS. Each term was a randomly-allocated pseudo word (from Rastle, Harrington, & Coltheart, 2002), and varied across participants.

Procedure We simulated a total of 30 iterated learning chains, each with 13 "generations" of learners. Each chain varied in the number of terms that were allowed in the "language" being transmitted, with two, three, four, five or six terms per language. The first learner in each chain received data generated from one of three types of initial partition of the WCS color space: hue, lightness, and random. The "hue" and "lightness" partitions were approximately equal vertical and horizontal partitions of the color space into the relevant number of categories; the "random" partitions were a truly random partition of the color space, generated uniquely for each chain. These three kinds of initial partitions were used as a means of checking the convergence of iterated learning: by starting the chains with very different systems of color terms, we could easily establish when the influence of the inital partition had disappeared. The following generations of learners all received data generated from the responses of the previous generation, as detailed below. We ran a total of 20 random chains, four for each number of terms, and five hue and five lightness chains, one for each number of terms.

Each participant was trained on the system of color terms by being shown a set of chips together with the corresponding terms. The total number of observed chips was six times the number of terms in the language. These chips were chosen at random from the 330 chips making up the full array, and then provided labels according to either the initial partition (for the first learner) or the responses of the previous learner (for subsequent learners). These training examples remained on the screen while the participant went on to label all 330 color chips from the WCS array. On every trial, they were presented a color chip and asked to select one of the terms to label the color chip. No feedback was given during this phase of the experiment. The responses of each participant thus produced a partition of the set of 330 chips, and this partition was used to generate the labels given to chips for the next learner in the chain.

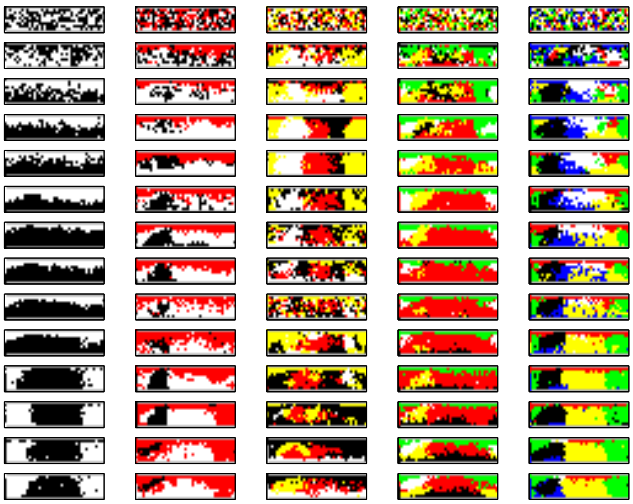


Figure 3: Representative examples of the data produced by simulating iterated learning of systems of color terms in the laboratory. Each panel shows one system of color terms, with arbitrary colors indicating the term assigned to each chip in the World Color Survey array. Each column is one chain with a particular number of terms, each row shows a different generation. The first row shows the random partitions used to initialize each chain.

Results

Figure 3 shows one set of chains initialized with random partitions, with the number of terms varying from two to six. Through this simple visualization of the data, we can see that each chain started from an unnatural color term system (a random partition), and that transmission along the chains resulted in a very rapid restructuring towards a more regular form. However, it is not clear how well these laboratory-generated data fit the WCS data. In the next section, we use a measure of the difference between each system of color terms and a randomly selected set of responses from the WCS data to test the convergence of the chains and to compare them to the kinds of systems seen across human languages.

Using Variation of Information to Analyze Color Term Systems

Analyzing the results of our experiment presents a challenge: how can we evaluate whether two systems of color terms are similar? Various methods have been proposed for solving this problem. For example, Kay and Regier (2003) converted the color chips from Munsell space to CIE L*a*b* space so they could compute the centroid for each color term. Centroid distances could then be used to compare clusterings. However, just using centroid measurements may discard important information about the variance of a cluster, and about the locations of boundaries. This method is also dependent on the psychological validity of the CIE L*a*b* representation of colors, which is disputable (Dowman, 2007).

Since our participants' responses consisted of partitions of

the same set of colors as those used in the WCS, we used a technique that compares the Munsell arrays directly, without referring to another color space. This technique uses an information-theoretic measure known as *Variation of Information* (VI) to compare clusterings of a set of items (Meila, 2003). Given two clusterings C and C', the VI is

$$VI(C,C') = H(C) + H(C') - 2I(C,C') \quad (2)$$

where $H(C)$ is the *entropy* of C,

$$H(C) = -\sum_{k=1}^{K} P(k) \log P(k) \quad (3)$$

where k ranges over the cluster labels and $P(k)$ is the probability of an item being assigned to each cluster, and $I(C,C')$ is the *mutual information* between the two clusterings

$$I(C,C') = \sum_{k=1}^{K} \sum_{k'=1}^{K'} P(k,k') \log \frac{P(k,k')}{P(k)P'(k')} \quad (4)$$

where $P(k,k')$ is the probability an item belongs to cluster k in clustering C and to k' in clustering C'.

While VI was originally developed for comparing clusterings, a clustering is simply a partition of a set of items, just as our systems of color terms partition colors according to the terms applied to them. The VI value for two systems of color terms is thus calculated by comparing the distribution of the terms in the two systems, as well as the extent to which they agree with one another. A high VI value reflects a larger difference between two clusterings, whereas a small VI value indicates that the two clusterings are more similar. Our primary analytic tool was comparing the partitions produced by our participants with those observed in the WCS data. We did this by randomly selecting one speaker from each of the 110 languages in the WCS data set, and then calculating the VI of the partition produced by our participants with the 110 partitions from the WCS. Finally, we average across all of the languages from the WCS, to give us a single measure of consistency.[1]

Testing Convergence

The theoretical analyses of iterated learning outlined above predict that, no matter what language begins a chain, it will eventually converge to a distribution over languages reflecting the prior. We could evaluate this prediction by comparing the chains generated with different initial partitions. A necessary characteristic for convergence is that the VI to the WCS data should not differ between chains, since they should all have reached the same distribution over languages. Figure 4 shows the VI values for the three types of initialization in each language system, showing individual chains with two to six terms for the hue and lightness initialization and the average over all chains for the random initialization.

To test for a difference in VI values across chains, we ran a two-way ANOVA at each generation with initialization and number of terms as the two factors. The main effect of number of terms are significant for all generations ($p < 0.05$); while only the initial systems ($F(2,23) = 196.78, p < 0.0001$) and the first generation ($F(2,23) = 11.19, p < 0.001$) showed a statistically significant effect of initial partition. These results are consistent with a relatively rapid convergence towards a common distribution. Rapid convergence is to be expected in this experiment, since only a very small proportion of the color chips were labeled in each generation, providing a good opportunity for other factors (such as the learning and perceptual biases of the participants) to influence the resulting systems of color terms. This can be seen in Figure 3, where the initial partitions quickly give way to more systematic responses.

Comparison to the WCS

As described above, to compare our experimental results with the WCS data, VI values were calculated between the responses of each participant and 110 randomly selected WCS systems. Figure 5 shows the VI values for all 20 random chains. A paired t-test on the VI values for the initial and final systems in those random chains showed a statistically significant difference ($t(19) = 11.44, p < 0.0001$), indicating a significant reduction of VI along iterated learning chains, resulting a better fit to the WCS data.

The remaining question is how close our data are to the WCS data: What counts as a low VI score? To address this question, we randomly selected another set of systems from the WCS data, one from each of the 110 languages. Using the same method as used above, we computed the VI between the two sets of WCS data. The average pairwise VI is shown in Figure 5. This average lies close to the mean VI seen in our random chains once they converge. We tested the difference between the VI scores produced by the final participants in each of our random chains and the VI scores for speakers sampled from the WCS using a two-samples t-test. The result was not significant ($t(128) = -0.29, p = 0.78$). These results suggest that the systems of color terms generated from our lab are indeed consistent with the data collected from the WCS.

Rotation Analysis

One potential objection to the conclusion that our chains are moving closer to the WCS could be that the reduction in VI may merely be a result of increasing regularity in the responses. As the systems of color terms in the random chains move towards more regular forms, the VI scores will go down naturally, regardless of whether the actual partition of terms reflects the structure of the WCS or not. To further test the consistency between our experiment results and the WCS

[1] While it would be desirable to also average over speakers, this was too computationally intensive to be practical in our current analyses. We observed little variation in average VI across sampled sets of speakers. We chose to use a single speaker rather than a composite formed by aggregating across speakers within a language (a "mode map") on the grounds that this might not produce a system typical of the language of any individual, especially as different speakers of the same language sometimes use different numbers of color words (Kay & Maffi, 1999).

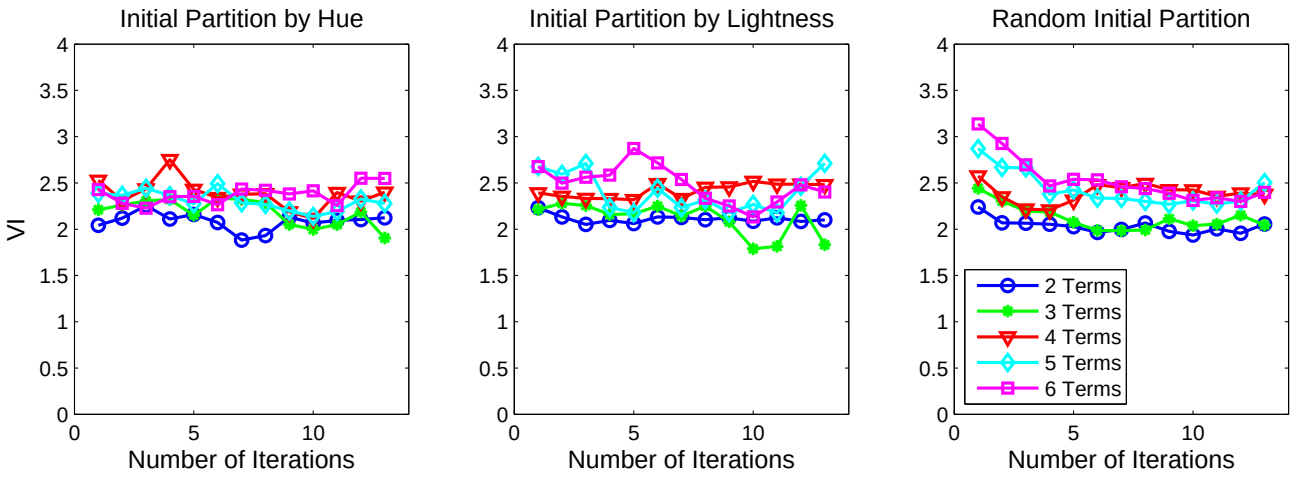


Figure 4: Variation of Information (VI) fit to WCS data for iterated-learning chains with three types of initial partitions and two to six color terms. Results for random initial partitions are averaged over four chains each.

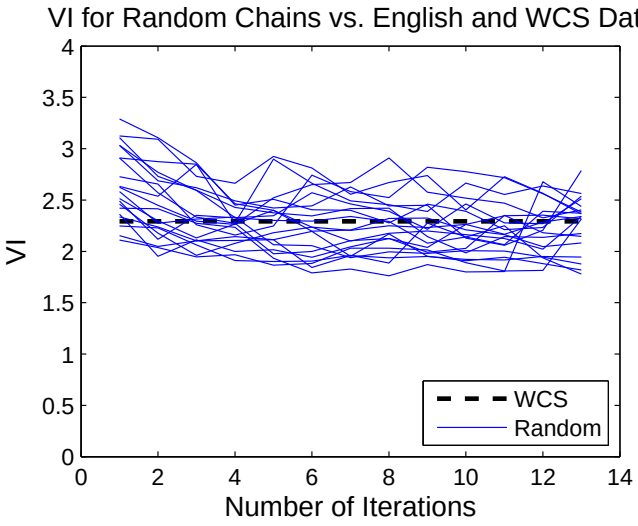


Figure 5: Variation of Information (VI) fit to WCS data for random chains. The dashed line shows the VI for comparing the WCS to itself.

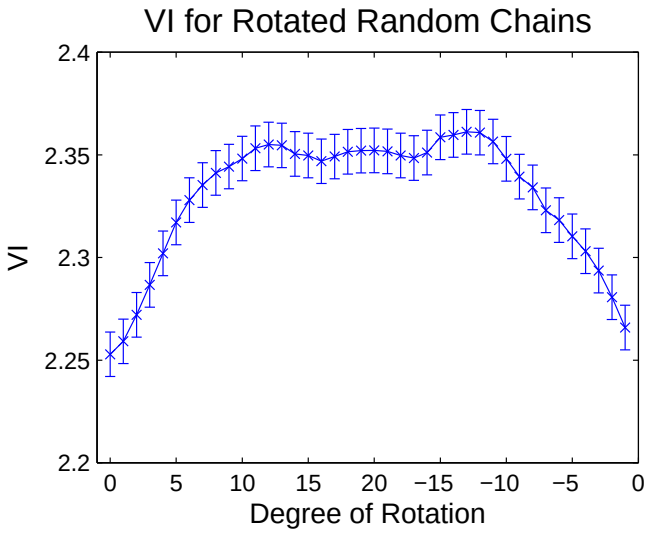


Figure 6: Variation of Information (VI) fit to WCS data for rotations of the final partitions produced by chains initialized with random partitions.

data, we compared the degree of match of each system to the WCS data when it was rotated in the hue dimension by varying amounts. We would expect that the more a partition was rotated out of position, the lower the resulting degree of match would be. This procedure was used by Regier et al. (2007) in connection with their measure of how "optimal" a set of color terms was as a division of the color space into maximally perceptually distinct regions.

Figure 6 shows the mean VI values of the partitions generated by the final participants in our random chains, when rotated from 0 to 20 steps in the hue dimension. Paired t-tests on VI values for no-rotation vs. maximum-rotation ($t(19) = -6.12, p < 0.01$), no-rotation vs. quarter-rotation ($t(19) = -3.66, p < 0.001$), and no-rotation vs. three-quarter-rotation ($t(19) = -4.61, p < 0.001$) all showed statistically significant differences, indicating that the data from the experiment fits the WCS data significantly better than the rotated systems. This analysis thus confirmed that the iterated learning chains did converge to forms closer to the WCS.

Discussion

We tested the idea that human color-naming universals may be a result of shared learning and perceptual biases, demonstrating that systems of color terms similar to those seen in a variety of non-industrial societies emerge purely as a result of cultural transmission. Using Variation of Information as a measure of the difference between systems of color

terms generated in our experiment and the WCS data, we showed that the VI for systems generated by iterated learning rapidly decreases as the systems moves from unnatural random partitions to more regular forms. Our rotation analysis also showed that this reduction of VI can not be explained as simply a result of the emergence of more regularity, but reflects the adoption of a form consistent with the WCS data.

One objection that could be made with respect to our study is that our English-speaking subjects could have been imposing a system of colour naming reflecting that of English on the languages in our experiments, rather than using pre-linguistic universal biases. As English has 11 basic color terms, many more than the 2 to 6 terms in our experiments, none of the emergent languages could reflect English very closely, which we could expect would minimize the potential for our participants knowledge of language to shape the colour categories formed in the experiment. We take the finding that systems of color terms similar to those seen in the WCS can be produced by cultural transmission by English speakers as supporting our argument that human learning and perceptual biases may be sufficient to explain universals, under the assumption that the English-speaking participants in our experiments share the same learning and perceptual biases as the members of non-industrial societies surveyed by the WCS. This result is less surprising when we take into account previous findings relating the color term categories produced by English speakers with cross-linguistic trends. For example, Boster (1986) found that when English speakers were asked to recursively split a set of color chips into subsets, the partitions they produced corresponded to those seen in other languages with a corresponding number of terms.

Our experiment and subsequent analyses not only demonstrate that iterated learning may provide a valuable experimental method for investigating human inductive biases, but also show that languages formed in the laboratory by English speaking participants seem to converge toward a form consistent with the WCS. These results suggest that the color-naming universals may come from the learning and perceptual biases of human learners, brought out through the process of cultural transmission. In particular, our results supplement previous computational modeling results demonstrating that such properties could be produced by iterated learning with simulated agents. We anticipate that similar pairings of laboratory experiments and computer simulations will be effective in further elucidating how languages and concepts change through cultural transmission.

Acknowledgments. This work was supported by grant number BCS-0704034 from the National Science Foundation. We thank Tony Lai, Jason Martin, Linsey Smith, and Joe Vuong for their assistance in collecting and analyzing the data.

References

Berlin, B., & Kay, P. (1969). *Basic color terms: Their universality and evolution*. Berkeley, CA: University of California Press.

Boster, J. (1986). Can individuals recapitulate the evolutionary development of color lexicons? *Ethnology*, *25*, 61-74.

Dowman, M. (2007). Explaining color term typology with an evolutionary model. *Cognitive Science*, *31*(1), 99-132.

Dowman, M. (2009). Evolution of basic color terms. In J. W. Minett & W. S.-Y. Wang (Eds.), *Language evolution and the brain* (p. 109-139). Hong Kong: City University of Hong Kong Press.

Griffiths, T. L., & Kalish, M. L. (2005). A Bayesian view of language evolution by iterated learning. In B. G. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the Twenty-Seventh Annual Conference of the Cognitive Science Society* (p. 827-832). Mahwah, NJ: Erlbaum.

Griffiths, T. L., & Kalish, M. L. (2007). Language evolution by iterated learning with bayesian agents. *Cognitive Science*, *31*, 441-480.

Hawkins, J. (Ed.). (1988). *Explaining language universals*. Oxford: Blackwell.

Hering, E. (1964). *Outlines of a theory of the light sense.* Cambridge, MA: Harvard University Press.

Kay, P., Berlin, B., Maffi, L., & Merrifield, W. R. (1997). Color naming across languages. In C. L. Hardin & L. Maffi (Eds.), *Color categories in thought and language.* Cambridge, UK: Cambridge University Press.

Kay, P., Berlin, B., & Merrifield, W. R. (1991). Biocultural implications of systems of color naming. *Journal of Linguistic Anthropology*, *1*, 12-25.

Kay, P., & Maffi, L. (1999). Color appearance and the emergence and evolution of basic color lexicon. *American Anthropologist*, *101*, 743-760.

Kay, P., & McDaniel, C. (1978). The linguistic significance of the meanings of basic color terms. *Language*, *54*, 610-646.

Kay, P., & Regier, T. (2003). Resolving the question of color naming universals. *Proceedings of the National Academy of Sciences*, *100*, 9085-9089.

Kirby, S. (2001). Spontaneous evolution of linguistic structure: An iterated learning model of the emergence of regularity and irregularity. *IEEE Journal of Evolutionary Computation*, *5*, 102-110.

Kirby, S., Dowman, M., & Griffiths, T. L. (2007). Innateness and culture in the evolution of language. *Proceedings of the National Academy of Sciences*, *104*, 5241-5245.

Meila, M. (2003). Comparing clusterings by the variation of information. In *Learning theory and kernel machines* (p. 173-187).

Rastle, K., Harrington, J., & Coltheart, M. (2002). 358,534 nonwords: The ARC nonword database. *Quarterly Journal of Experimental Psychology*, *55A*, 1339-1362.

Regier, T., Kay, P., & Cook, R. S. (2005). Focal colors are universal after all. *Proceedings of the National Academy of Sciences*, *102*, 8386-8391.

Regier, T., Kay, P., & Khetarpal, N. (2007). Color naming reflects optimal partitions of color space. *Proceedings of the National Academy of Sciences*, *104*, 1436-1441.

Similarity judgments reflect both language and cross-language tendencies: Evidence from two semantic domains

Naveen Khetarpal (khetarpal@uchicago.edu)[a]
Asifa Majid (asifa.majid@mpi.nl)[b]
Barbara Malt (barbara.malt@lehigh.edu)[c]
Steven Sloman (steven_sloman@brown.edu)[d]
Terry Regier (terry.regier@berkeley.edu)[e]

[a]Department of Psychology, University of Chicago, Chicago, IL 60637 USA
[b]Max-Planck Institute for Psycholinguistics, 6500 AH Nijmegen, The Netherlands
[c]Department of Psychology, Lehigh University, Bethlehem, PA 18015 USA
[d]Department of Cognitive and Linguistic Sciences, Brown University, Providence, RI 02912 USA
[e]Department of Linguistics, Cognitive Science Program, University of California, Berkeley, CA 94720 USA

Abstract

Many theories hold that semantic variation in the world's languages can be explained in terms of a universal conceptual space that is partitioned differently by different languages. Recent work has supported this view in the semantic domain of containers (Malt et al., 1999), and assumed it in the domain of spatial relations (Khetarpal et al., 2009), based in both cases on similarity judgments derived from pile-sorting of stimuli. Here, we reanalyze data from these two studies and find a more complex picture than these earlier studies suggested. In both cases we find that sorting is similar across speakers of different languages (in line with the earlier studies), but nonetheless reflects the sorter's native language (in contrast with the earlier studies). We conclude that there are cross-culturally shared conceptual tendencies that can be revealed by pile-sorting, but that these tendencies may be modulated to some extent by language. We discuss the implications of these findings for accounts of semantic variation.

Keywords: Language and thought; semantic universals; linguistic relativity.

A universal basis for semantic variation?

The semantic systems of the world's languages vary considerably. This observation has suggested two opposed accounts of the relation between language and thought. The Sapir-Whorf hypothesis holds that such cross-language differences cause corresponding differences in cognition, leading speakers of different languages to think about and perceive the world substantially differently (Lucy, 1992; Majid et al., 2004; Roberson et al., 2000). In contrast, many other theories accommodate such variation by positing a universal conceptual space that is partitioned in different ways by different languages (Berlin & Kay, 1969; Croft, 2003:139; Levinson & Meira, 2003; Majid et al., 2008; Malt et al., 1999; Regier et al., 2007). On this view, the significant point about the variation is that many logically possible semantic configurations are never attested – thus, the constrained variation illuminates underlying commonalities in human cognition.

Although the starting point for this debate is linguistic – namely the observation of semantic diversity across languages – a natural means of testing it is by probing non-linguistic cognition. The Whorfian view predicts that speakers of languages with different semantic systems should conceive of the world differently, each group in line with their own language's semantic system. The universal-space view in contrast predicts that speakers of different languages should conceive of the world similarly.

One source of support for the universal-space view comes from *pile-sorting*. In the first large-scale quantitative study of its kind, Malt et al. (1999) asked speakers of English, Chinese, and Spanish to name a set of household containers – e.g. a jar, a juice-box, an ice-cream carton, etc. – and to pile-sort pictures of these items on the basis of their overall similarity. They found that while naming patterns differed substantially across languages, sorting patterns did not.

The same view is indirectly supported by recent studies that explain differing patterns of semantic structure in the world's languages as optimal or near-optimal partitions of an underlying and presumably universal similarity space. Regier et al. (2007) demonstrated that color naming in the world's languages is consistent with this idea, assuming a standard perceptual color space, CIELAB. This account explains universal tendencies in color naming while also accommodating some deviation from those tendencies, as is observed empirically. Khetarpal et al. (2009) showed that the same idea can account for semantic variation in the spatial domain. In the spatial case, however, no standard independent assessment of a universal similarity space exists. Therefore, inspired by the Malt et al. (1999) results, Khetarpal et al. (2009) based their analysis on similarities derived from pile-sorting of spatial scenes by speakers of Dutch and English. Critically, while they assumed that these similarities would be universal or near-universal, and while their results were consistent with that assumption, they did not directly test the assumption. We test it here.

To preview our results, we find that pile-sorting of spatial stimuli, according to the data of Khetarpal et al. (2009), is broadly similar across languages – but does nonetheless

differ as a function of language. These results were obtained using an analysis different from that of Malt et al. (1999) – thus the question arises whether Malt et al.'s (1999) container data would yield similarly mixed results under our analysis. We show that they do. We conclude that on one analysis at least, pile-sorting reveals not just shared cross-language tendencies, but also apparent influence of the sorter's native language, suggesting an interesting combination of the universalist and Whorfian positions (Regier & Kay, 2009).

Spatial language and cognition

Khetarpal et al. (2009) demonstrated a commonality underlying the diversity of spatial naming in the world's languages. They based their study on a set of 71 spatial scenes that were originally designed by Melissa Bowerman and Eric Pederson. Figure 1 shows a sample of 10 of these scenes, as categorized in 2 languages.

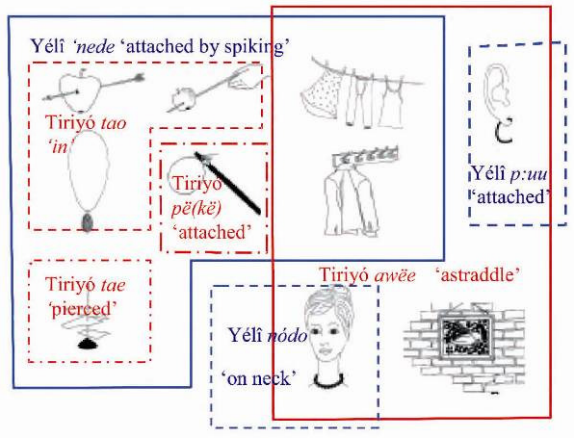


Figure 1: 10 spatial scenes, as categorized in 2 languages: Tiriyó and Yélî-Dnye. Source: Levinson & Meira (2003).

Khetarpal et al. (2009) had native speakers of Dutch and native speakers of American English sort pictures of these 71 spatial scenes into piles on the basis of the similarity of the spatial relation portrayed. Afterwards, they also elicited names for these spatial relations from each sorter in his or her native language. They then derived similarity judgments from sorting behavior: the similarity between any two scenes *x* and *y* was taken to be the proportion of all participants (American and Dutch pooled together) who sorted *x* and *y* into the same pile. Finally, they assessed the spatial semantic systems of 9 unrelated languages (one language was Dutch but the rest were unrelated to Dutch and English; Levinson & Meira, 2003) relative to these similarities. They found that these 9 attested spatial semantic systems maximized similarity within categories, and minimized it across categories (Garner, 1974), more than did a reasonable set of competitor systems of comparable complexity; in this sense these attested spatial semantic systems are *near-optimal*. This finding is consistent with the assumption that the sorting-derived similarities are universal – since they help to explain the spatial semantic systems of unrelated languages. But is this assumption in fact correct – or do these similarities reflect the sorters' native language? A natural means of testing this question is to compare the sorts produced by speakers of English and Dutch to the naming systems of the same two languages.[1] The Whorfian prediction is that speakers of each language should sort in a manner that reflects their native language, more than the other language. The universalist prediction is that speakers of the two languages should sort identically.

Methods

Naming data. For both English and Dutch, separately, we recorded the modal spatial term for each of the 71 spatial scenes — i.e. the spatial term that was used by the largest number of speakers of the language to name that scene. Ties were broken by random choice. The resulting labeling of the 71 scenes was taken to be that language's spatial naming system.

Sorting data. We analyzed the English and Dutch sorting data in 3 ways. First, we measured the *correlation* of sorting behavior across languages. Second, we measured how well sorts matched the semantic systems of English and Dutch, using *edit distance*. Third, we examined the *height*, or coarse-grainedness, of the sorts and of the English and Dutch semantic systems, since this quantity is helpful in interpreting other analyses, as will be seen below. Here, we describe each analysis in turn.

Correlation analysis. Following Malt et al. (1999), we compared sorts produced by English and Dutch speakers as follows. For each of Dutch and English, for each pair of scenes, we counted the number of times those two scenes were placed in the same pile by speakers of that language. This yielded, for each of the two languages, a vector of (71×70)/2 = 2485 co-sorting counts. We determined the correlation of the Dutch vector with the English vector.

Edit-distance analysis. We took a pile-sort of the 71 scenes to be a *partition* of those stimuli into groups; we similarly took a language's names applied to those scenes to be a partition of the same set of stimuli into groups. We quantified the dissimilarity between two such partitions by measuring the *edit distance* between them. The edit distance between two partitions A and B is the minimum number of operations required to change A into B, where each operation involves moving a single item from one group to another (possibly empty) group. We computed edit distances via the Hungarian algorithm for bipartite graph

[1] We collected new English data analogous to that of Khetarpal et al. (2009), since their English naming data were incomplete. We report here the comparison of Khetarpal et al.'s (2009) complete Dutch data with our complete English data. Comparison of Khetarpal et al.'s (2009) Dutch and English data yield qualitatively the same results as those we report here.

matching (Deibel et al., 2005).[2] For each pile sort produced by a speaker of either Dutch or English, we determined its edit distance to the partition defined by the Dutch language, and its edit distance to the partition defined by the English language.

Height analysis. The *height* of a partition is a measure of how coarse-grained it is: greater height indicates coarser grain, while lower height indicates finer grain. Height is defined as the sum, over all groups in a partition, of the number of pairs of items in each group (Coxon, 1999):

$$height = \sum_i \binom{g_i}{2} = \sum_i g_i(g_i - 1)/2$$

where g_i is the number of items in group i. We measured the height of the partitions corresponding to the English and Dutch naming systems, and the height of each pile-sort.

Results and discussion

Correlation. The correlation of the Dutch and English co-sorting vectors was 0.87. This correlation is fairly high, and is greater than the agreement between halves of the same group (Dutch or English): the mean within-group split-half reliability was 0.80. This result suggests that speakers of the two languages sorted quite similarly.

Edit distance. Edit distance gives us a means of measuring the dissimilarity between pile-sorts and naming systems. Figure 2 shows the average edit distance of sorts produced by Dutch speakers and those produced by English speakers, to the Dutch and English naming systems.

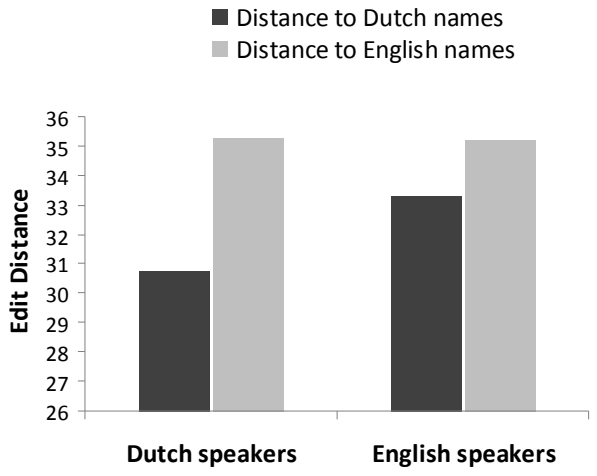


Figure 2: Edit distance of sorts, produced by Dutch and English speakers, to the Dutch and English naming systems.

We analyzed these data as follows. For each sorter from each of the two languages, we created a difference score: the edit distance of that person's pile sort to the English naming system minus the edit distance of that person's pile sort to the Dutch naming system. The difference scores for both groups were significantly greater than 0 (Dutch: M=4.5, $t(23) = 4.83$, $p < .0002$; English: M=1.92, $t(23) = 3.81$, $p < .002$), indicating that speakers of both languages sorted more in line with Dutch than with English. The Dutch mean difference score was greater than the English one ($t(46) = 2.44$, $p < 0.05$; all p values Bonferroni-corrected), indicating that Dutch speakers showed this preference for Dutch over English more strongly than English speakers did. Thus there appears to be both a cross-language tendency to sort more in line with Dutch than with English (a universalist finding), and a tendency to sort in line with one's native language (a Whorfian finding); these two forces pull in the same direction for Dutch speakers, but in opposite directions for English speakers.

What is it about the Dutch naming system such that speakers of both languages sort more in line with it than with English? It may be relevant that Dutch appears to be semantically *finer-grained* than English in this domain. For example, the English spatial term *on* covers a broad range of spatial meanings, including a cup on a table, and a picture on a wall – whereas these two spatial configurations are named differently in Dutch (as *op* vs. *aan*, respectively). Thus a possible explanation for the privileged status of Dutch in our results above is that people may tend to sort in a manner that is finer-grained than either language, and therefore more like the finer-grained language – in this case Dutch.

Figure 3 shows that this is the case. The height quantity measures the coarseness of a partition; thus, comparison of the two vertical lines shows that Dutch naming is indeed finer-grained than English naming with respect to these spatial scenes. Moreover, the bulk of sorts produced by speakers of both languages is finer-grained than the finer-grained language, Dutch.

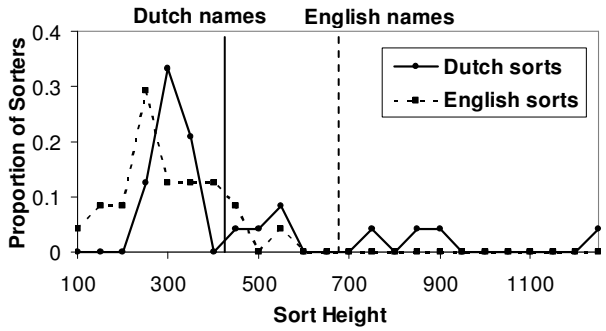


Figure 3: The height (coarse-grainedness) of the Dutch and English naming systems, and sorts produced by speakers of these two languages.

Thus, it seems likely that Dutch emerges as privileged in our edit-distance results at least in part because it is finer-grained than English in this domain. But are these results attributable to fine grain *per se*, or to the particular fine-grained partition that Dutch represents? To test this, we

[2] See http://psych.uchicago.edu/~khetarpal/code/edit-distance for our code, which extends an implementation written by Gary Baker and released under GPLv3.

also compared the pile-sorts to Dutch-*like* partitions which are as fine-grained as Dutch but group the items differently. The set of Dutch-like partitions was sampled repeatedly (n=3.5x10^6) by randomly grouping items such that the total number of groups equaled the number of Dutch spatial terms and the sizes of these groups matched the number of items associated with the Dutch spatial terms. We then measured the average edit distance from English speakers' sorts to each of these sampled hypothetical Dutch-like partitions (*Min*=46.79, *Mean*=52.09, *Max*=55.13), and the average edit distance from Dutch speakers' sorts to each of these sampled hypothetical Dutch-like partitions (*Min*=46.04, *Mean*=51.48, *Max*=54.29). In both cases the average edit distance of the sorts to actual Dutch (shown in Figure 2) was less than to any of the sampled hypothetical Dutch-like partitions of equally fine grain.[3] This finding suggests that the privileged status of Dutch in our edit-distance results is a function not just of its fine grain, but also of the similarity relations it captures.

Taken together, these reanalyses of the Khetarpal et al. (2009) spatial data suggest that spatial similarity judgments as gauged by pile-sorting are quite similar and fine-grained across languages – a universalist finding – but that they nonetheless vary in line with the sorter's native language – a Whorfian finding.

Container names and cognition

Our present analysis of the Khetarpal et al. (2009) spatial data revealed a mixed picture, in contrast with the purely universalist results of Malt et al. (1999) on containers. But our result was obtained through an edit-distance analysis that Malt et al. (1999) did not use. This raises the question whether the Malt et al. (1999) data would also exhibit an effect of language if analyzed using edit distance. We sought to test this question.

Malt et al. (1999) based their study on 60 pictures of simple containers, such as cartons, boxes, bottles, and the like. They asked speakers of 3 different languages – American English, Mandarin Chinese, and Argentinean Spanish – to name the containers shown in these pictures and to sort them into piles, on several different bases. Here, we re-examine their data from English and Chinese, for which data were readily retrievable, and we focus on pile-sorting based on overall similarity of the containers, rather than functional or perceptual similarity, which Malt et al. (1999) also probed. Importantly, while the semantic categories for the various containers differed across languages, the overall sorts showed no effect of language in their analyses.

Methods

We analyzed Malt et al.'s (1999) container naming and sorting data from Chinese and English using the same methods we had applied to the spatial data of Khetarpal et al. (2009). Specifically, we (1) identified each language's semantic partitioning of the space by determining the modal term applied to each stimulus in each language, and conducted (2) correlation, (3) edit-distance, and (4) height analyses of the sorting and naming data.

Results and discussion

Correlation. The correlation of the Chinese and English co-sorting vectors was 0.91, as Malt et al. (1999) had found. This correlation is quite high, and is comparable to the agreement between halves of the same group (Chinese or English): the mean within-group split-half reliability was 0.90. This result suggests that speakers of the two languages sorted quite similarly.

Edit distance. Figure 4 shows the average edit distance of sorts produced by Chinese speakers and those produced by English speakers, to the Chinese and English naming systems.

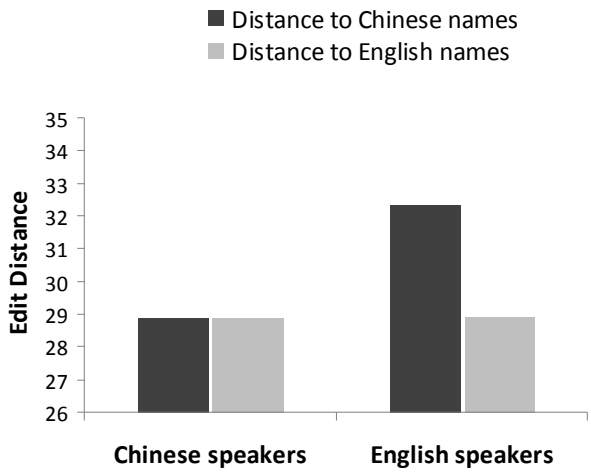


Figure 4: Edit distance of sorts, produced by Chinese and English speakers, to the Chinese and English naming systems.

We analyzed these data as before. For each sorter from each of the two languages, we created a difference score: the edit distance of that person's pile sort to the Chinese naming system minus the edit distance of that person's pile sort to the English naming system. The mean difference score for Chinese speakers was 0.0 (*SD* = 5.99), indicating that Chinese speakers sorted in a manner equally similar to the Chinese and English naming systems. In contrast, the mean difference score for English speakers was significantly greater than 0 (*M*=3.43; $t(55) = 6.17$, $p < .0002$), indicating that English speakers sorted in a manner more like the English than like the Chinese naming system. The English mean difference score was greater than the Chinese one ($t(36.6^4) = 2.64$; $p < .05$; all p values Bonferroni-corrected), indicating that English speakers sorted in line with English

[3] The actual Dutch naming system is also by definition a Dutch-like partition.

[4] Heteroscedasticity corrected using Welch's method.

more than Chinese to a greater extent than Chinese speakers did. As in the spatial case, a natural interpretation of these data is that there is a cross-language tendency to sort more in line with English than with Chinese, and also a tendency to sort in line with one's native language. For Chinese speakers these two forces cancel each other out, whereas for English speakers they reinforce each other.

Given our earlier discussion, a general tendency to sort more in line with English than with Chinese naming would make sense if English were more fine-grained than Chinese in this domain, and if people sorted more finely than either language. Figure 5 shows that this is the case.

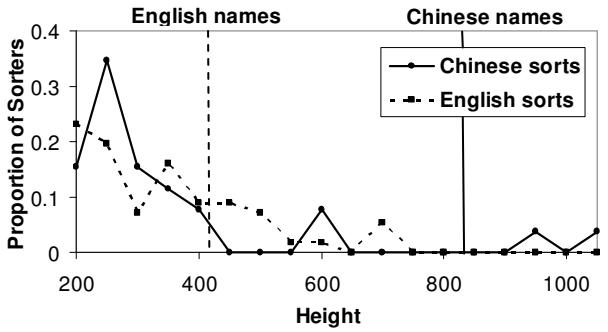


Figure 5: The height (coarse-grainedness) of the Chinese and English naming systems, and sorts produced by speakers of these two languages.

Whereas English was coarser-grained than Dutch in the spatial domain, it is finer-grained than Chinese in the container domain. And the bulk of the sorts produced by speakers of both languages is finer-grained yet. This is consistent with the reasoning proposed above for the apparently privileged status of English in our edit-distance analysis of the container data. Still, as before, we wished to ascertain whether the results are attributable to fine grain *per se*, or to the particular fine-grained partition that English represents. To test this, we also compared the pile-sorts to English-*like* partitions of the container items which are as fine-grained as English but group the items differently – analogously with our creation of Dutch-like partitions of spatial relations, described above. The set of English-like partitions was sampled repeatedly (n=3.5x10^6) by randomly grouping items such that the total number of groups equaled the number of English container terms and the sizes of these groups matched the number of items associated with the English container terms. We then measured the average edit distance from English speakers' sorts to each of these sampled hypothetical English-like partitions (*Min*=41.54, *Mean*=45.67, *Max*=48.02), and the average edit distance from Chinese speakers' sorts to each of these sampled hypothetical English-like partitions (*Min*=44.23, *Mean*=48.01, *Max*=50.31). In both cases the average edit distance of the sorts to actual English (shown in Figure 4) was less than to any of the sampled hypothetical English-like partitions of equally fine grain. This finding suggests that the privileged status of English in our edit-distance results is a consequence not just of its fine-grainedness, but also of the specific groupings of referents that it represents.

Taken as a whole, these reanalyses of the Malt et al. (1999) container data present a picture similar to the one that emerged from our examination of the Khetarpal et al. (2009) spatial data. Similarity judgments as assessed by pile-sorting are fine-grained and quite similar across languages, but also reflect the sorter's native language to some extent. Thus, there is again evidence both for cross-language and for language-specific forces – and thus for both the universalist and Whorfian positions.

Conclusions

Different languages exhibit different systems of semantic categories. It is often assumed that this semantic variation is constrained by, and can be explained by, a universal conceptual space that is partitioned in different ways by different languages. Malt et al. (1999) found evidence consistent with such a language-invariant space, and Khetarpal et al. (2009) assumed such a space existed. In both cases conceptual similarity was assessed through pile-sorting.

We reanalyzed data from these two earlier studies, with a view to reassessing whether pile-sorting on the basis of similarity does or does not reflect language. In both cases we found the same overall picture: pile-sorting was very similar across speakers of different languages (in agreement with the findings and assumptions of the earlier studies), but it also tended to reflect the sorter's native language (in contrast with those studies). Moreover, pile-sorting tended to be semantically finer-grained than any of the languages we considered. These findings suggest several conclusions.

First, they suggest a particular view of the relation of language and thought, namely that: (a) there is a set of fine-grained and potentially cross-cutting conceptual distinctions that may be made, and some languages will happen to mark more of these distinctions than will other languages; (b) distinctions that are unmarked in a language are nonetheless conceptually available to speakers of that language – this is suggested by the fine-grained sorting; and (c) a distinction becomes more salient if it is marked linguistically in one's native language (Hespos & Spelke, 2004) – this is suggested by the effect of language we find. This interpretation is consistent with the general view that "Whorf was half right" and correspondingly half wrong, as has been argued elsewhere (Regier & Kay, 2009).

Second, our results are compatible with the possibility that language may influence cognition in relatively subtle ways that are detectable by some analyses and not by others. Edit distance applied to pile-sorting may be a useful analytical tool, when used in tandem with others, in pursuing this question more generally.

Finally, our results suggest that caution is needed when basing accounts of semantic variation on an ostensibly universal similarity space derived from pile-sorting (e.g. Khetarpal et al., 2009) – because universality cannot be assumed. Similarity judgments are likely to be similar but

not identical across languages, as was the case in our analyses. This highlights an unavoidable tension. A universal conceptual space is a useful theoretical construct for explaining semantic variation, but we have no guarantee that such a thing actually exists – nor, if it does, do we have a completely reliable means of assessing it. Instead, we have somewhat language-colored approximations to such a space, and these should be treated as such. A reasonable treatment may be to average together similarity judgments obtained from speakers of different languages in an attempt to better approximate a universal similarity space, as Khetarpal et al. (2009) did. But any interpretation of results based on such an approximation should be tempered by the awareness that it is merely an approximation.

At the same time, our results leave a number of questions open. The first concerns the contrast between our findings and those of Malt et al. (1999). They found that language was not reflected in sorting by overall similarity, and we found that it was, based on the same data. One possibility, as mentioned above, is that our edit distance analysis is more sensitive than some others, such that it picks up on differences that are missed by other analyses. Is this conclusion correct? Or is our analysis itself inappropriately biased in some respect? Which set of results should be believed? Answering this question is critical to placing our present findings in their proper context.

A second question raised by our findings is the extent to which they generalize to other languages. If we were to examine a new language that partitions semantic space more finely than the languages we have examined here, we would expect to find that pile-sorts produced by people of all backgrounds tend to align more closely with this new fine-grained language than they do with the more coarse-grained languages we have already examined. Is this the case? This question provides a straightforward means of further testing these ideas.

There is also the question of whether these results generalize to other semantic domains. While we have restricted ourselves to the two domains of spatial relations and containers, this was simply a matter of convenience, as the data were readily available. The reasoning behind these ideas however is general in scope, and we would expect to find supporting evidence in other semantic domains as well.

Finally, while these results demonstrate a correlation between language and sorting behavior, they do not demonstrate the causal link claimed by the Whorf hypothesis. It remains an open question whether the observed correlation is attributable to an effect of language on cognition, or to other factors, such as culture influencing both language and cognition.

Regardless of how these questions are eventually answered, we hope that our present initial findings help to make plausible the central idea we have promoted here: a fine-grained conceptual space, largely shared in structure across speakers of different languages, but nonetheless also reflecting the speaker's native language.

Acknowledgments

This work was supported by NSF under grant SBE-0541957, the Spatial Intelligence and Learning Center (SILC).

References

Berlin, B. & Kay, P. (1969). *Basic color terms: Their universality and evolution*. Berkeley: University of California Press.

Coxon, A. (1999). *Sorting data: Collection and analysis.* Thousand Oaks, CA: Sage Publications.

Croft, W. (2003). *Typology and universals, 2nd edition.* Cambridge, UK: Cambridge University Press.

Deibel, K., Anderson, R., & Anderson, R. (2005). Using edit distance to analyze card sorts. *Expert Systems, 22*, 129-138.

Garner, W. R. (1974). *The processing of information and structure*. Potomac, MD: L. Erlbaum Associates

Hespos, S. J. & Spelke, E. S. (2004). Conceptual precursors to language. *Nature, 430*, 453 - 456.

Khetarpal, N., Majid, A., & Regier, T. (2009). Spatial terms reflect near-optimal spatial categories. In N. Taatgen et al. (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society*.

Levinson, S. C. & Meira, S. (2003). Natural concepts in the spatial topological domain—adpositional meanings in crosslinguistic perspective: An exercise in semantic typology. *Language, 79,* 485-516.

Lucy, J. (1992). *Grammatical categories and cognition: A case study of the linguistic relativity hypothesis.* Cambridge, UK: Cambridge University Press.

Majid, A., Boster, J., & Bowerman, M. (2008). The cross-linguistic categorization of everyday events: A study of cutting and breaking. *Cognition, 109*, 235-250.

Majid, A., Bowerman, M., Kita, S., Haun, D., & Levinson, S. (2004). Can language restructure cognition? The case for space. *Trends in Cognitive Sciences, 8*, 108-114.

Malt, B. C., Sloman, S. A., Gennari, S., Shi, M., & Wang, Y. (1999). Knowing versus naming: Similarity and the linguistic categorization of artifacts. *Journal of Memory and Language, 40,* 230-262.

Regier, T., Kay, P., & Khetarpal, N. (2007). Color naming reflects optimal partitions of color space. *PNAS, 104*, 1436-1441.

Regier, T., & Kay, P. (2009). Language, thought, and color: Whorf was half right. *Trends in Cognitive Sciences, 13,* 439-446.

Roberson, D., Davies I. & Davidoff, J. (2000). Color categories are not universal: Replications and new evidence from a stone-age culture. *Journal of Experimental Psychology: General*, 129, 369-398.

Adults' self-directed learning of an artificial lexicon: The dynamics of neighborhood reorganization

Neil P. Bardhan (nbardhan@bcs.rochester.edu)
Richard N. Aslin (aslin@cvs.rochester.edu)
Michael K. Tanenhaus (mtan@bcs.rochester.edu)
Department of Brain & Cognitive Sciences, University of Rochester
Meliora Hall, Box 270268
Rochester, NY 14627 USA

Abstract

Artificial lexicons have previously been used to examine the time course of the learning and recognition of spoken words, the role of segment type in word learning, and the integration of context during spoken word recognition. However, in all of these studies the experimenter determined the frequency and order of the words to be learned. Here we ask whether adult learners choose, either implicitly or explicitly, to listen to novel words in a particular order based on their acoustic similarity. We use a new paradigm for learning an artificial lexicon in which the learner, rather than the experimenter, determines the order and frequency of exposure to items. We analyze both the temporal clustering of subjects' sampling of lexical neighborhoods during training as well as their performance during repeated testing phases (accuracy and reaction time) to determine the time course of learning these neighborhoods. Subjects sampled the high and low density neighborhoods randomly in early learning, and then over-sampled the high density neighborhood until test performance on both neighborhoods reached asymptote. These results provide a new window on the time-course of learning an artificial lexicon and the role that learners' implicit preferences play in learning highly confusable words.

Keywords: spoken word recognition; phonological neighborhoods; word learning; artificial lexicon

Introduction

Since the pioneering work of Marslen-Wilson (1987) on the role of acoustic/phonetic similarity in on-line spoken word recognition, there has been debate over the structure of phonological neighborhoods in the mental lexicon. In the cohort model, lexical items were neighbors—and, thus, competitors—if and only if their sound-forms overlapped from the beginning of the word, such as in "pat" and "pack". The Neighborhood Activation Model (Luce, Pisoni & Goldinger, 1991; Luce & Pisoni, 1998) quantified neighborhood similarity as a combination of factors: the frequency of the single item in question, neighborhood density (also describable as confusability), and overall neighborhood frequency. Neighbors as defined by NAM can include rhyme words (e.g. "pat" and "rat") and words with other one-segment differences, such as "pat" and "past".

A series of studies by Creel, Aslin, and Tanenhaus (2006) employing an artificial lexicon (Magnuson, Tanenhaus, Aslin, & Dahan, 2003) further revealed another intricacy of neighborhood structure, specifically by asking whether all segment differences, regardless of type (i.e., consonant vs. vowel) have an equal influence on confusion of newly learned words. Two CVCV items with matching consonants are more often confused with each other than two CVCV items with matching vowels; in other words, segment type matters. Furthermore, the position of the consonants played a role: VCVC items with matched consonants did not elicit such confusions.

One outstanding question concerning neighborhoods is how they develop. After a pre-lexical infant learns its first word (e.g., "no" or its own name), how does acoustic/phonetic similarity affect the learning of new words? Do infants acquire words based solely on frequency of occurrence in the ambient linguistic environment, or do they systematically avoid attending to novel words that are similar in sound-structure with known words? Based on corpus analyses, Charles-Luce and Luce (1990, 1995) made just such a prediction, but others have provided conflicting evidence (Coady & Aslin, 2003; Dollaghan, 1994). More direct evidence comes from word-learning studies with toddlers. Swingley and Aslin (2007) taught young children new words that were either neighbors to words they already knew (e.g., "tog" vs. "dog") or non-neighbors (e.g., "meb"). Neighbor items were more difficult than non-neighbors for the children to learn. However, conflicting evidence from toddlers exists (Newman, Samuelson & Gupta, 2008), suggesting that with more exposure they can learn a novel item from a high-density neighborhood as well as from a very low-density neighborhood.

This same question of neighborhood effects on word learning applies to adults, who are constantly acquiring new words in their lexicon (e.g., "locavore", "staycation"). Perhaps more relevant to the growing adult lexicon is the case of learning words in a second language. Here there is both a neighborhood effect *within* the L2 lexicon and interference effects *between* the L1 and L2 lexicons. There is conflicting evidence of between-language neighborhood effects (Spivey & Marian 1999; Ju & Luce, 2004) in spoken word recognition, but virtually no evidence for such effects in word *learning*. One reason for this limited evidence is that studies of L1 and L2 lexicons are extremely difficult to control, and L2 often provides the learner with phonetic and

phonological cues that clearly mark the lexical item as a member of only one language.

Another approach to the study of neighborhood effects in word learning is to create new words that are designed to compete with known words. Gaskell & Dumay (2003) present evidence of competition development in English-speaking adults who learn a non-English word that competes with an English word that lacks neighbors. For example, "cathedral" has no English neighbors, but listeners were exposed to the meaningless word-form "cathedruke" over the course of an experiment. The novel item immediately leads to facilitatory effects on the English item. However, after sleeping, subjects' behavior reflected lexical competition between the two forms. These results suggest that new words compete with old words during spoken word recognition, but they do not bear directly on the time-course of *learning* new words. Importantly, Magnuson et al. (2003), using an artificial lexicon, found no significant evidence that the native language (English) interfered with the processing of neighbors from the artificial lexicon, at least not after only 2 hours of training. Thus, in the early phase of training, even with 90% or better accuracy in learning the names for novel objects, adults do not seem to show between-language neighborhood interference.

Here we describe a study of adult learners using an artificial lexicon. The rationale for using an artificial lexicon, as in Magnuson et al. (2003) and Creel et al. (2006), is that we can carefully control all the parameters of the lexicon (density, frequency, phoneme inventory, meaning) that are very difficult to balance using natural language materials. Our key innovation is creating a learning paradigm in which adults choose how they listen to the entire set of novel words. They must map 16 novel word-forms onto 16 novel visual shapes. Across a series of learning blocks, subjects sample the sound-object pairs by selecting a shape on a touch screen and hearing that shape's name. A testing phase after each training block assesses the accuracy and speed of word recognition using the same touch screen. By varying the neighborhood structure within the set of 16 words, we can determine whether adult learners choose to sample from high or low density neighborhoods during the process of learning novel word-object mappings.

Overview of Design

The learning environment was simplified by presenting each subject with an array of 16 novel shapes on a touch screen display. We selected a touch-screen monitor rather than a computer mouse because it lends itself to ease of use for children or other special populations who have limited mouse experience. Subjects are instructed that they should learn the names of the shapes by touching them and hearing that shape's name. They are told that they can touch shapes in any order and that they have 64 touches per block. After each block, they will be tested for their knowledge of shape names by hearing a name and completing a 16-AFC task. This alternation of training blocks and testing blocks allows us to describe any changes in how subjects allocate their exploration of the lexicon and its relation to how well subjects have learned the lexicon.

Method

Participants

A total of 41 subjects from the University of Rochester participated in the study and consented per the guidelines of the University of Rochester human subjects review board. Each subject received $10 for one session of approximately 45 minutes. They were told that they would be listening to words and selecting pictures on a touch screen, to learn the names of the pictures, and subsequently tested on what they have learned. All subjects reported normal hearing, normal or corrected-to-normal vision, and were native speakers of English.

Stimulus Materials

The lexicon was created to vary in neighborhood density and type of acoustic/phonetic similarity. It consisted of 16 items in total: a high-density cohort neighborhood (baga, bagi, bago, bagu), a high-density rhyme neighborhood (dido, kido, pido, tido) and 8 low-density items (gobu, dupi, poti, toku, kuba, tupa, gota, puki). Items were recorded as WAV files by a graduate student with linguistics training. The speaker read each item at a natural rate, yielding an average word length of 745 milliseconds. Items were paired with 16 novel black and white images (Hunt & Aslin, 2009). Three different list conditions of random pairings of words and pictures were used, counterbalanced to avoid any item effects that may have arisen from particular word-image pairings. Subjects were randomly assigned to conditions.

Environment

The experiment was run on MathWorks Matlab and the Matlab Psychophysics toolbox (Brainard, 1997; Pelli, 1997) on a Dell Dimension desktop PC running Windows XP with an NEC touchscreen monitor. Images were randomly presented in a 4x4 grid on the screen., as seen in Figure 1 Subjects listened to words over Sennheiser HD 570 headphones set at a comfortable volume level. The study was conducted in a sound-attenuated booth.

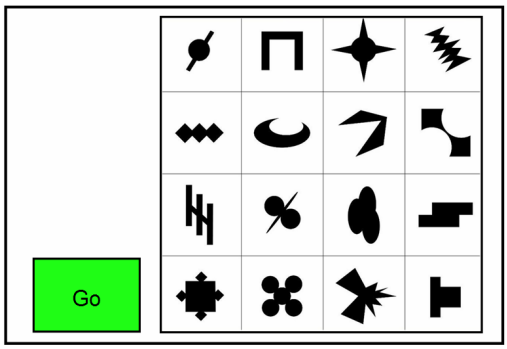


Figure 1: An example screen

Task

Training and testing were alternated in a session, with 6 blocks of each, for a total of 12 blocks. Participants were told to select the items (i.e., touch a shape) in any order they desired to learn the words that named each of the 16 images on the screen. They were not told that they would be trained on the same 16 images and corresponding words in future blocks. Rather, they were told that each testing block would correspond to the preceding training block. During each training block, an on-screen counter marked off the number of remaining training trials the subject had, from 64 to 0, until a test block would begin. If the subject had evenly distributed their touches in a block, they would hear each word 4 times, which was deemed sufficient for a minimal level of familiarity but not full mastery after the initial block. The location of each image was randomized four times during training blocks: once at the beginning of the block, then once after each 16 trials. This precluded the possibility that subjects made associations between item location and name, rather than the desired effect of item image (shape) and name.

A test block consisted of two passes through the list of lexical items, for a total of 32 trials, in random order. Subjects pressed a GO button image on the screen to start a test trial, then heard a word corresponding to one of the 16 images on the screen. They were free to select any of the items present on the screen, in a 16-AFC task. Instructions specified that they should respond as quickly and accurately as possible. If they correctly selected an image, the image turned green. If they selected an incorrect image, the image turned red; they were not informed of the correct image. Thus feedback provided the subject with only minimal information about each decision—whether it was correct—but not information as to which image was the correct item if they made an error. Allowing the participant to start each trial provided the opportunity for short rests as needed during the testing blocks.

Results

Training data

An analysis of variance was performed to determine the effect of density (high versus low) on overall proportion of selection of training items; the result was not significant. Within high density items, a two-factor ANOVA with replication was performed, comparing the number of selections of items from the high density cohort neighborhood to those from the high density rhyme neighborhood, across blocks. Cohort items were chosen more frequently than rhyme items, $F(1,84) = 15.69$, $p < .001$ (see Figure 2). Block was also a significant factor $F(5,420) = 2.52$, $p < .05$ and there was an interaction of neighborhood type and block, $F(5, 420) = 3.60$, $p < .01$.

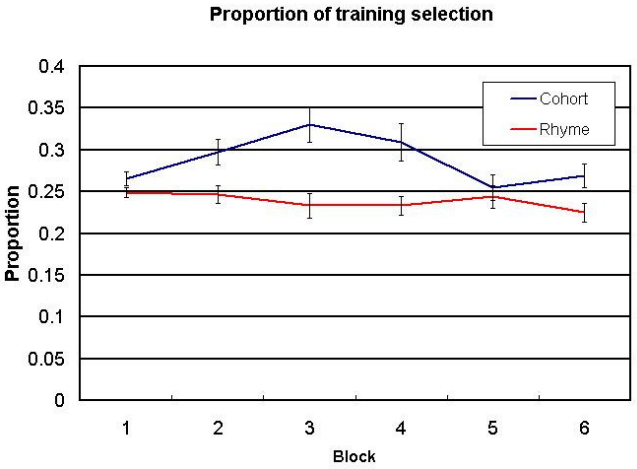


Figure 2: Training selections of high density items

Training sequences were then analyzed for the likelihood that a subject, having selected an item from a particular neighborhood, would next select an item from the same neighborhood. The blue line in Figure 3 shows, across all subjects, the proportion of item selections that, on the immediately following trial, were drawn from the same high density neighborhood. Error bars represent standard errors of the mean. The pink line represents the eight low-density items grouped into random pseudo-neighborhoods of four items, to provide a baseline comparison for the likelihood of selecting within any group of four items.

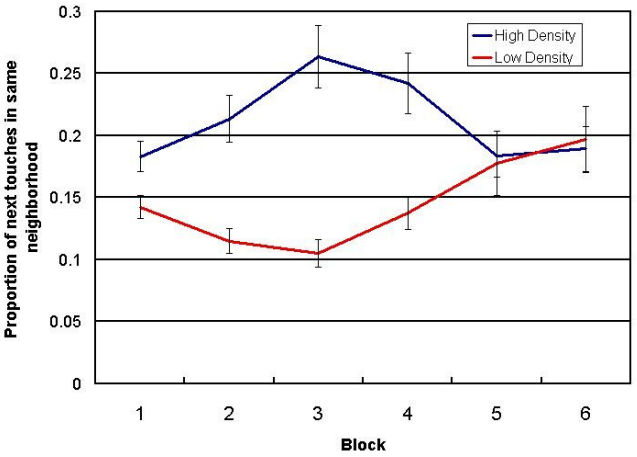


Figure 3: In-neighborhood probability

A two-factor ANOVA with replication was performed. This revealed a main effect of density, low versus high, $F(1,5) = 37.86$, $p < .001$, and a significant interaction of density x block, $F(1,5) = 5.92$, $p < .001$. Within the high-density neighborhoods, a single factor ANOVA examined whether there was an effect of block on probability of successive same-neighborhood selections. A significant effect of block was found, $F(5,252) = 2.94$, $p < .05$.

Test data:

Figure 4 shows that for both high and low density items, the accuracy of responding on the 16-AFC test blocks rose rapidly from 50% correct (chance=6.25%) to asymptotic performance within the 6 testing blocks. A two-factor ANOVA with replication examining accuracy across test blocks revealed a significant effect of block $F(5,504) = 98.19$, $p < .001$. There was also an effect of density $F(1,) = .10$, $p < .05$, but no interaction of density with block $F(5,504) = .36$, $p > .05$.

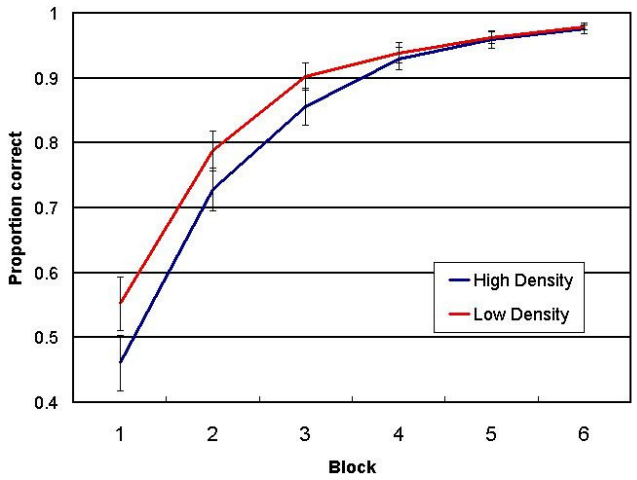


Figure 4: Proportion of test trials correct

A single-factor ANOVA of reaction times for correct trials across test blocks, collapsed across densities, showed a significant effect of block $F(5,252) = 8.03$, $p < .001$.

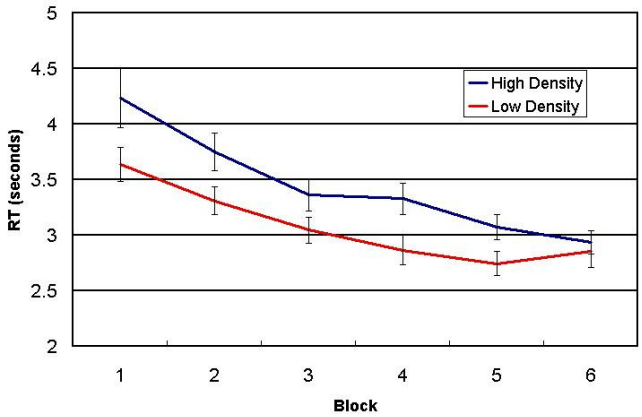


Figure 5: Reaction times in test

Figure 5 shows reaction times in the initial and final test blocks, as a function of neighborhood density. A two-factor ANOVA with replication revealed highly significant effects of block $F(5,492) = 14.50$, $p < .001$ and density $F(1,1) = 18.36$, $p < .001$.

Discussion

The present experiment is the first that we know of to assess how human learners allocate their attention by selecting novel words for association with novel visual objects in a word-learning paradigm. By having subsets of words that share acoustic/phonetic properties (lexical neighbors), we could ask whether learners seek or avoid repetitive samples of words from low- or high-density neighborhoods as they acquire new word-object associations.

Training data

Throughout the six training blocks, more high density cohort items were selected than high density rhyme items. Subjects concentrated their selections on cohort items, presumably because of the perceived phonological similarity among these items.

For high density items, there was a significant effect of block on the likelihood that subjects selected an item from one neighborhood and then on the subsequent touch selected an item from the same neighborhood. (This includes pressing the same item twice in a row.) In the initial trials, selection was nearly random. As the session continued, however, the probability that subsequent selections were within the same neighborhood significantly increased, then decreased to initial levels as mastery of the items was achieved and concentrated training on neighbors was no longer beneficial. When the low density items were randomly grouped into two groups of four and the selection data from those were compared to the high density selections, there was a significant effect of density and an interaction with blocks. Subjects were more likely to select two high-density items within a neighborhood (of four), one after another, than to select any two low-density items out of a random grouping of four. A regression model (proposed later) may be informative in further analyses of the influence of word-sampling behavior within one training block on the behavior exhibited in subsequent blocks.

Testing data

Subjects achieved 51% accuracy within the first testing block, after hearing each word on average only 4 times; this is significantly above chance (6.25% correct). This minimal exposure was sufficient to achieve significant learning, but performance did not reach asymptotic levels until 5 or 6 blocks of training. Accuracy was affected by density. High density items were correctly identified less frequently than low density items until halfway through the experiment. Their phonological similarity presumably created greater difficulty for the subjects.

As in previous studies, differences in reaction times also occurred as a result of density, with low density items being responded to more rapidly than high density items consistently until the final block, at which point performance on the two densities was equivalent.

Future studies

Numerous ways of examining the development of neighborhoods could provide greater insight into the time course of lexical competition during word learning. One such study would be to successively reveal subsets of neighborhoods to the learner in the paradigm described here. The training sequences would be of particular interest; in contrast to the present study, a different set of items would be present during each training block, and so the subject may adjust training strategies accordingly as overall neighborhood density is revealed across blocks.

Other statistical analyses, in the form of linear regression models, may reveal more about the present study and future designs. One analysis would be whether the performance on one test block influences training patterns in the immediately following block, which the current analyses cannot address well. Finally, our current analyses of within-neighborhood effects of training include pairs of trials in which one of the four items from the same high-density neighborhood is selected, including immediate repeats of the same item. Excluding these identical repeats may be more relevant to the question of lexical competition during learning.

Conclusion

It is well established that words in high-density neighborhoods are more difficult to process than words in low-density neighborhoods. When adults are given control over the frequency of exposure to novel words, they quickly adjust the rate of exposure by over-sampling words in high-density neighborhoods, particularly cohort neighbors more so than rhyme neighbors. The difficulty of learning high density items was revealed in this study as differences in accuracy and reaction time, which persist for the initial blocks of test trials. However, subsequent training yields equivalent learning accuracy for words in high- and low-density neighborhoods.

Our paradigm is likely to be useful for addressing a variety of issues in lexical learning. Perhaps most importantly, the method may be useful for teaching children novel lexical items, either in the laboratory or in the classroom.

Acknowledgments

This research was supported by NIH Grants HD037082 to RNA and DC005071 to MKT. The authors express gratitude to Johnny Wen for programming assistance and Dana Subik and Carrie Miller for recruiting participants.

References

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision, 10*, 433-436.

Charles-Luce, J. & Luce, P. A. (1990). Similarity neighbourhoods of words in young children's lexicons. *Journal of Child Language 17*, 205–15.

Charles-Luce, J. & Luce, P. A. (1995). An examination of similarity neighbourhoods in young children's receptive vocabularies. *Journal of Child Language 22*, 727–35.

Coady, J. A., & Aslin, R. N. (2003). Phonological neighbourhoods in the developing lexicon. *Journal of Child language, 30*(02), 441-469.

Creel, S. C., Aslin, R. N., & Tanenhaus, M. K. (2006). Acquiring an artificial lexicon: Segment type and order information in early lexical entries. *Journal of Memory and Language, 54*, 1-19.

Dollaghan, C. A. (1994). Children's phonological neighbourhoods: half empty or half full ? *Journal of Child Language 21*, 257–71.

Gaskell, M. G., & Dumay, N. (2003). Lexical competition and the acquisition of novel words. *Cognition, 89*(2), 105-132.

Hunt, R. H., & Aslin, R. N. (2009). Category induction via distributional analysis: Evidence from a serial reaction time task. *Journal of Memory and Language, 62*, 98-112.

Ju, M. & Luce, P. A. (2004). Falling on Sensitive Ears - Constraints on Bilingual Lexical Activation. *Psychological Science, 15* (5), 314-318.

Luce, P. A., Pisoni, D. B., & Goldinger, S. D. (1991). Similarity neighborhoods of spoken words. In G. T. M. Altmann (Ed.), *Cognitive models of speech processing* (pp. 122-147). Cambridge, MA: MIT Press.

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing, 19*(1), 1-36.

Magnuson, J. S., Tanenhaus, M. K., Aslin, R. N., & Dahan, D. (2003). The time course of spoken word learning and recognition: Studies with artificial lexicons. *Journal of Experimental Psychology: General, 132*(2), 202-227.

Newman, R., Samuelson, L., & Gupta, P. (2008). *Learning Novel Neighbors: Distributed mappings help children and connectionist models.* Paper presented at the Cognitive Science Society, Washington, DC, USA.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10*, 437-442.

Revill, K. P., Tanenhaus, M. K., & Aslin, R. N. (2008). Context and spoken word recognition in a novel lexicon. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*(5), 1207-1223.

Spivey, M. J., & Marian, V. (1999). Cross talk between native and second languages: Partial activation of an irrelevant lexicon. *Psychological Science, 10*(3), 281-284.

Swingley, D., & Aslin, R. N. (2007). Lexical competition in young children's word learning. *Cognitive Psychology, 54*(2), 99-132.

Perceptual Advantage from Generalized Linguistic Knowledge

Bożena Pająk (bpajak@ling.ucsd.edu)
Department of Linguistics, UC San Diego
9500 Gilman Drive #108, La Jolla, CA 92093 USA

Abstract

We address the question of how previously acquired linguistic knowledge facilitates perception and learning of a new language. We report results from two experiments showing evidence that participants better discriminate a segmental duration contrast in a novel language if they had some previous exposure to a language that uses duration contrastively. Crucially, the perceptual advantage occurs even when the novel language employs the contrast in entirely different conditions: in novel segmental contexts and for novel segments, including a change from application to vowels to application to consonants. We take these results to suggest that language learners use their knowledge of previously learned languages to make inferences about the ways in which languages are likely to vary, which in turn increases their perceptual sensitivity when languages do in fact vary in the predicted ways.

Keywords: Speech perception; language learning; overhypotheses; cross-linguistic influence; multilingualism.

Introduction

Second language (L2) acquisition and bilingualism have received a lot of attention in the literature over the past few decades (for an overview, see e.g., Ritchie & Bhatia, 2009). Significantly less research has been done on acquisition of more than two languages, but it is now more widely recognized that acquisition of additional languages (L*n*) is fundamentally different from L2 acquisition due to more possibilities for between-language interactions (De Angelis, 2007). However, the mechanisms behind these between-language interactions are still very poorly understood.

One specific consequence of this limitation in scope to L2 acquisition has been a lack of systematic research investigating the common intuition that while learning an L2 is often hard, learning each subsequent language becomes easier. We explore this intuition by asking higher-level questions about the learning process, as well as the nature of abstracting and generalizing, with the goal to understand what mechanisms would produce facilitation in L*n* acquisition. A convenient framework for asking these questions makes use of Goodman's (1955) notion of "overhypothesis". Goodman's observation was that humans not only learn things that they experience directly, but they also infer abstract knowledge about how the things they experience directly are structured. Imagine a stack of bags that contain colored marbles. You empty a few of them and discover that some bags only have black marbles, while others only have white ones. Now imagine you pick a new bag and draw one marble which turns out to be white. The experience with previous bags makes you hypothesize that this particular bag contains only white marbles. This hypothesis is based on the overhypothesis you formed through your overall experience with this stack of bags that each bag contains marbles that are uniform in color.

Our proposal is to work within the overhypothesis framework to build a model of multiple language learning (MLL). The reasoning behind the model is that with knowledge of only one language, one possesses a limited amount of information about the possible features that languages can have: for example, what sounds they use or what syntactic and morphological features they employ. For a monolingual speaker, the general overhypothesis over how languages are structured, or what linguistic dimensions are relevant to assign meaning, depends entirely on the knowledge of one's single native language (L1). This means that when learning an L2, one is likely to assume that the L2 features are similar to those of L1, a prediction confirmed by a large body of research in L2 acquisition (Ritchie & Bhatia, 2009). On the other hand, with at least some basic knowledge of two or more languages, one can update the overhypothesis (or, rather, the set of overhypotheses, each related to a specific linguistic dimension) by reevaluating which dimensions are relevant for each language. Structural differences between two or more languages on any given dimension provide a basis for expanding the hypothesis space (i.e., predictions about possible categories) for this dimension. We hypothesize that this conceptually expanded hypothesis space facilitates learning of novel categories along relevant dimensions in an L*n* due to the fact that specific predictions about possible categories have already been formed, which in turn accelerates their processing.

When applied at the level of sound, the model makes predictions regarding the ability of bi/multilingual learners to discriminate novel contrasts along familiar dimensions in an L*n*. Auditory perception of nonnative contrasts is often initially impaired, but can significantly improve with increased exposure. The explanation for this initial difficulty is that nonnative speech perception is shaped by the L1 experience, and current theories are successful in predicting which L2 sounds will be harder to initially perceive by listeners with a given language background (Kuhl & Iverson, 1995; Best, 1995; Flege, 1995). However, these models are not explicit about how the L1 bias is overcome when nonnative phonological categories are successfully learned. A possible answer is instead provided by the literature on perceptual categorization. When learning phonological categories in L2, language learners adjust weights assigned to different phonetic dimensions so that dimensions reliably aiding in proper phoneme categorization in L2 are given more weight, while dimensions creating phonologically irrelevant variation are given almost no weight (Kruschke, 1992; Strange & Shafer, 2008). Incorporating these theoretical assumption, the MLL model predicts that assigning high weight to a given phonetic dimension based on L2 input raises the likelihood of this dimension

also being considered as relevant in L*n*. This, in turn, leads to facilitation in perception and learning of novel L*n* categories that make use of this dimension.

As an example, consider the phonetic dimension of segmental duration. Imagine a native speaker of English who also speaks Cantonese. In English, segmental duration is used mainly as a prosodic cue (Klatt, 1976), while in Cantonese, vowel duration can be considered a *contrastive* feature (Bauer & Benedict, 1997), which means that words can potentially be distinguished based solely on the duration of a given vowel (short vs. long). The model assumes that this speaker has formed an overhypothesis over the dimension of duration stating that duration of segments can be relevant for assigning meaning in some languages. Crucially, even though the speaker's experience with duration is only based on *vowel* segments, he/she is expected to have formed an overhypothesis over the duration dimension that is not segment-specific, and consequently, to have formed hypotheses (or predictions) regarding the relevance of duration for *any* segment. Now, if this speaker is learning an L*n* like Polish, which has a *consonant* duration contrast, the model predicts a facilitation in learning this contrast, as compared to a learner who has not had any previous exposure to any duration contrasts, and consequently no opportunity to form an overhypothesis over the duration dimension.

Here we report the results of two experiments, in which the perception of short vs. long consonants was tested. In the first experiment, we tested speakers of American English who had previously learned a language with contrastive consonant duration. We expected to observe a perceptual advantage for this group over English speakers without such experience. Furthermore, we tested whether the perceptual advantage generalizes to novel consonant segments and novel segmental contexts. In the second experiment, we asked the question of whether the feature of contrastive duration can be generalized even further: namely, from vowels to consonants. Specifically, we tested the perception of a consonant duration contrast by speakers fluent in English and Cantonese, which has a duration contrast for vowels, but not for consonants.

Experiment 1

In an AX discrimination task we tested the perception of a consonant duration contrast (short vs. long) by a "bilingual group": native (or near-native) speakers of American English with previous exposure to another language that uses consonant duration contrastively (e.g., Japanese or Italian). The control "monolingual group" consisted of native speakers of American English with no previous exposure to any language that contrasts duration.[1] Following the assumption of the MLL model that the bilinguals have assigned a high weight to the duration dimension in their L2, it was predicted that the bilingual group would perform better than the monolingual group. Furthermore, the perception of the duration contrast was tested in different phonotactic environments (here, the position in a word and the adjacent segments). While some theories assume that learning new contrasts is context-specific (Flege, 1995), the proposed model predicts that the abstracted knowledge should allow generalization across different environments. Additional comparisons within the bilingual group were planned in order to investigate more closely the process of generalization from previous knowledge. In particular, it was predicted that the bilingual participants would be able to generalize their perceptual capacity for duration contrasts to novel segments, and – following the underlying principle of the overhypothesis framework – that familiarity with the contrast in at least two different contexts would faciliate generalization to a novel context more than its familiarity in only one context.

Method

Participants 80 undergraduate students at UC San Diego participated in the experiment for course credit: 40 "monolinguals" and 40 "bilinguals". The bilingual group was largely heterogeneous. It consisted of speakers of a total of 17 different L2s, and varied in their proficiency in L2, as well as the manner of exposure to L2 (school instruction or exposure at home through family members). The bilingual participants were further split in two ways depending on the types of segments possible as long consonants and the positions in which they occur in their L2. The division by "segment" included "[ss] bilinguals" who were only familiar with long [s], and "[ss] & [zz] bilinguals" who were familiar with both long [s] and long [z]. The division by "context" included "intervocalic bilinguals" who were only familiar with long consonants in the intervocalic context, and "intervocalic+ bilinguals" who were familiar with the contrast in the intervocalic context plus in at least one other context (word-medial preconsonantal and/or word-initial prevocalic).

Materials The materials consisted of nonce words constructed around either a long or a short target consonant. The target consonants were placed in four different contexts created by crossing two conditions: word position (medial vs. initial) with following segment (vowel vs. consonant). All the bilingual participants had previous exposure to the contrast in the word-medial prevocalic (or intervocalic) context, while none had previous exposure to the contrast in the word-initial preconsonantal context. Two different types of segments were used: voiceless alveolar fricatives [s]/[ss] and voiced alveolar fricatives [z]/[zz], resulting in a total of eight conditions. The materials are shown in Table 1.

[1] American English does not use duration contrastively. Vowel duration varies, but it correlates with the tense-lax distinction (e.g., *beat* vs. *bit*) or depends on the voicing of the following segment (e.g., *cad* vs. *cat*). Long consonants are sometimes attested but only at morpheme boundaries (e.g., *dissatisfied*; Benus, Smorodinsky, & Gafos, 2003). Minimal pairs are rare (e.g., *unnamed* vs. *unaimed*), and for most speakers the contrast in neutralized (Kaye, 2005). Furthermore, there is evidence that by 18 months of age English-learning infants process duration contrasts differently from infants learning a language that contrasts duration (e.g., Japanese; Mugitani, Pons, Fais, Werker, & Amano, 2008).

Table 1: Materials. (V-vowel, C-consonant, #-word boundary)

		Prevocalic	Preconsonantal
		V_V	**V_C**
Word-medial	*voiceless*	asa/assa	asta/assta
	voiced	aza/azza	azda/azzda
		#_V	**#_C**
Word-initial	*voiceless*	sa/ssa	sta/ssta
	voiced	za/zza	zda/zzda

The materials were recorded by a male native speaker of Moroccan Arabic since all the words were phonotactically legal in this language. For each test word, 18 repetitions were recorded. The duration of the fricatives was measured, and five tokens with fricatives that approximated mean duration for each condition were selected for use in the experiment. The trials consisted of an equal number of "different" pairs (e.g., asa-assa) and "same" pairs (e.g., assa-assa or asa-asa). Even in the "same" pairs, the first and second words in each pair were always physically different tokens, and were separated by an interstimulus interval of 500ms. Each participant heard 12 "different" pairs and 12 "same" pairs for each of the eight conditions. There was a total of 384 pairs in the experiment: 192 test pairs and 192 fillers.

Procedure The experiment began with a practice session during which participants listened to 16 filler stimuli (8 "different" and 8 "same" pairs) over headphones, and were asked to respond by clicking on one of two answer boxes displayed on the computer screen saying "same word" or "different words". No feedback was given during the practice session. The test trials followed immediately after the practice session. Participants were presented with six 64-trial blocks. On each trial, a stimulus was presented aurally through headphones, and the participant responded by clicking on one of the two boxes on the computer screen. Each stimulus was played once without a replay option. The response to one stimulus triggered the presentation of the following stimulus with a delay of 500ms. The stimuli order was randomized for every participant. There was a self-terminated break after each block.

Results

We calculated A-prime scores for each participant and each condition (the same results hold for d-prime). A-prime (Grier, 1971) was used to measure the participants' capacity to perceive the short/long consonant contrast, and is a non-parametric analog of d-prime. Both A-prime and d-prime are measures of sensitivity to a given contrast, and are calculated by taking into account the proportion of Hits (responding 'different' when the stimulus is 'different') and False Alarms (responding 'different' when the stimulus is 'same').[2] A-prime yields values between 0 and 1, where 1 means 'perfect discriminability' and 0.5 is chance performance.

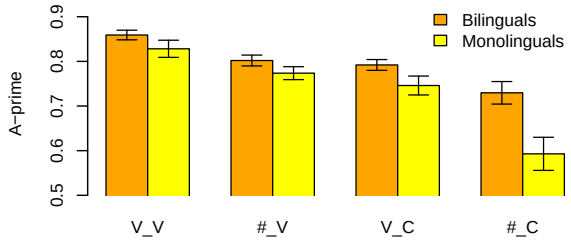


Figure 1: Performance on the short/long consonant contrast in different contexts by monolinguals and bilinguals. (Error bars are standard errors.)

Monolinguals vs. bilinguals The results are plotted in Fig. 1. We analyzed the scores using a repeated measures ANOVA with the within-participants factors *position* ('medial' or 'initial'), *following segment* ('vowel' or 'consonant'), *voicing* ('voiced' or 'voiceless'), and the between-participants factor *language background* ('monolingual' or 'bilingual'). There was a significant main effect of *language background* $[F(1,78) = 12.8; p < .001]$ with the bilingual group performing better $(\bar{A}' = 0.79)$ than the monolingual group $(\bar{A}' = 0.72)$.

There was also a significant interaction between *language background* and *following segment* $[F(1,78) = 11.4; p < .01]$, and a three-way interaction between *language background*, *following segment* and *position* $[F(1,78) = 4.2; p < .05]$. The difference in performance between monolinguals and bilinguals was larger in preconsonantal than in prevocalic contexts, and was especially striking in the word-initial, preconsonantal contexts.

In addition, there were significant main effects of *position* [F(1,78)=34.6; p<.001] and *following segment* $[F(1,78) = 64.8; p < .001]$ independent of language background. That is, both groups performed better when the contrast was in word-medial (vs. word-initial) contexts, and better when it was prevocalic (vs. preconsonantal).

Bilinguals: segments Two groups of bilinguals were compared depending on the segments that occur as long consonants in their L2. The results are plotted in Fig. 2. We analyzed the scores using a repeated measures ANOVA with the same as before within-participants factors *position* ('medial' or 'initial'), *following segment* ('vowel' or 'consonant'), *voicing* ('voiced' or 'voiceless'), and the between-participants factor *L2-long-consonant-segment* ('[ss] bilinguals' or '[ss] & [zz] bilinguals'). The two groups of bilinguals were balanced by randomly removing participants from the larger group, leaving a total of 20 participants for this comparison.

No significant main effect of *L2-long-consonant-segment*

[2]The formula used for calculating A-prime was the following: $A' = 0.5 + \frac{(H-FA)(1+H-FA)}{4H(1-FA)}$ (where H = Hits, and FA = False Alarms; Grier, 1971, p. 425). In order to avoid infinite or undefined values, $H = 0$ was converted to $\frac{1}{2N}$, and $F = 1$ was converted to $1 - \frac{1}{2N}$ (where N = number of trials on which the proportion is based; Macmillan & Creelman, 2005).

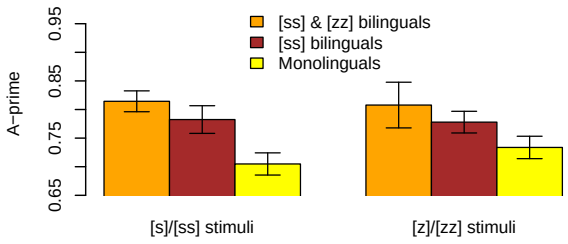


Figure 2: Performance on the short/long consonant contrast by bilinguals familiar with both long [s] and long [z] ('[ss] & [zz] bilinguals') and bilinguals only familiar with long [s] but not long [z] ('[ss] bilinguals'). Monolinguals added for comparison. (Error bars are standard errors.)

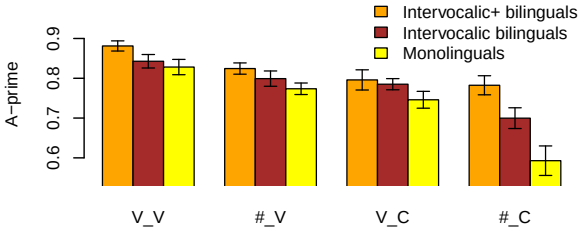


Figure 3: Performance on the short/long consonant contrast by bilinguals familiar with the contrast in different contexts ('intervocalic+ bilinguals') and bilinguals familiar with the contrast only in the intervocalic context ('intervocalic bilinguals'). Monolinguals added for comparison. (Error bars are standard errors.)

was found $[F < 1]$ and no interaction between *L2-long-consonant-segment* and *voicing* $[F < 1]$. Both groups of bilinguals performed the same on both voiceless and voiced contrasts.

There were also significant main effects of *position* $[F(1,18) = 8.1; p < .05]$ and *following segment* $[F(1,18) = 16.3; p < .001]$. Again, both groups performed better when the contrast was in word-medial (vs. word-initial) contexts, and better when it was prevocalic (vs. preconsonantal).

Bilinguals: context The final comparison involved bilinguals which were grouped depending on the contexts in which long consonants are possible in their L2. The results are plotted in Fig. 3. We analyzed the scores using a repeated measures ANOVA with the within-participants factors again being *position* ('medial' or 'initial'), *following segment* ('vowel' or 'consonant'), *voicing* ('voiced' or 'voiceless'), and the between-participants factor *L2-long-consonant-context* ('intervocalic' or 'intervocalic plus V_C and/or #_V'). The two groups of bilinguals were balanced by randomly removing participants from the larger group, leaving a total of 32 participants for this comparison.

There was a significant main effect of *L2-long-consonant-context* $[F(1,30) = 5.5; p < .05]$ with the intervocalic+ bilingual group performing better $(\bar{A}' = 0.79)$ than the intervocalic group $(\bar{A}' = 0.72)$.

As in previous comparisons, there were significant main effects of *position* $[F(1,30) = 18.5; p < .001]$ and *following segment* $[F(1,30) = 19.5; p < .001]$. As before, both groups performed better when the contrast was in word-medial (vs. word-initial) contexts, and better when it was prevocalic (vs. preconsonantal).

Discussion

As predicted, the participants performed better on the short/long consonant contrast than monolingual participants. Importantly, the effect was observed despite high heterogeneity of the bilingual group in terms of their L2 background, the shared feature being the presence of the short/long consonant contrast in every L2. All the bilingual participants seemed to be able to use a similar kind of perceptual capacity which emerged from their different backgrounds. This result provides support for the hypothesis that previous exposure to duration contrasts in an L2 improves perception of a similar contrast in a novel language.

Furthermore, better performance by bilinguals was not simply a result of direct incorporation of certain elements from L2 to a novel language, because – as predicted – the bilinguals were able to generalize their perceptual capacity to novel segments (at least across voicing of segments) and novel contexts. Thus, the perceptual capacity was not found to be context-specific, even though some contexts may be perceptually harder than others.

Finally, following the hypothesis, the bilinguals whose L2 made the contrast in at least two different contexts (intervocalic, word-medial preconsonantal and/or word-initial prevocalic) performed better in the novel word-initial preconsonantal context than the group whose L2 only used the contrast intervocalically. This suggests that, while exposure to at least one segmental context of a duration contrast helps with the overall perception (as is the case for the "intervocalic bilinguals"), it is the exposure to at least two different contexts that allows for a real boost in perceptual capacity (as observed for the "intervocalic+ bilinguals"). This result can be interpreted as a supporting piece of evidence for the overhypothesis framework: exposure to a contrast in at least two contexts allows for the formation of an overhypothesis that this particular contrast can occur in many different contexts. Interestingly, the "intervocalic+ bilinguals" also performed better than the "intervocalic bilinguals" in the intervocalic context, to which all the bilinguals had equal exposure. While this result does not directly follow from the hypothesis, it might be that forming the overhypothesis over contexts makes the perceptual system more attuned to the contrast in any context, either novel or non-novel.

Experiment 2

Experiment 2 was designed in order to test whether the feature of contrastive duration can be generalized further than across voicing of segments, namely, from vowels to consonants. The participants were speakers of Cantonese (also fluent in Mandarin) and speakers of Mandarin with no knowledge of Cantonese. Cantonese has vowel duration contrasts, but no consonant duration contrasts, while Mandarin does

not use duration of any segments contrastively. To control for differences in populations, the experiment also included stimuli with a sibilant contrast from Polish, which were chosen because similar consonants form part of the Mandarin consonant inventory. Thus, the two groups of participants (Cantonese/Mandarin and Mandarin) were exposed to two types of stimuli: duration contrasts (short vs. long consonants) and sibilant contrasts (alveolo-palatal vs. postalveolar/retroflex consonants). The MLL model predicted that Cantonese speakers would perform better than Mandarin speakers on the duration contrast due to their experience with the vowel duration contrast in Cantonese. However, both groups were predicted to perform equally well on the sibilant contrast due to their familiarity with a similar contrast in Mandarin (although a slight advantage for the native Mandarin speakers was expected for this contrast).

Method

Participants 40 undergraduate students at UC San Diego participated in the experiment for course credit. 20 were native speakers of Mandarin fluent in English, and the other 20 were native speakers of Cantonese fluent in English and at least competent in Mandarin.

Materials The materials consisted of two types of stimuli, as shown in Table 2.

Table 2: Materials: segmental contrasts.

A. Duration contrasts (short vs. long)

Sonorants	Obstruents
[j]/[jj] [w]/[ww] [l]/[ll] [m]/[mm] [n]/[nn]	[f]/[ff] [s]/[ss]

B. Sibilant contrasts (alveolo-palatal vs. postalveolar/retroflex)

Voiceless	Voiced
[ɕ]/[ʂ] [t͡ɕ]/[t͡ʂ]	[ʑ]/[ʐ] [d͡ʑ]/[d͡ʐ]

This created 4 conditions by crossing two factors: contrast (duration or sibilant) with language background (Cantonese/Mandarin or Mandarin).

Each contrast was embedded in seven different frames: [pa_a], [pe_a], [po_a], [ta_a], [te_a], [ka_a], [ke_a]. All the words were recorded by a phonetically-trained native speaker of Polish with five repetions of each word. One token of each stimulus type was chosen as a frame (a short consonant token for duration contrasts and a postalveolar/retroflex token for sibilant contrasts). The target consonants were spliced out from different tokens. Long consonants were created from the short consonants by either doubling their length (for sonorant consonants: [j], [w], [l], [m], and [n]) or elongating it by half its length (for obstruent consonants: [f] and [s]).[3] The stimuli were created by pairing words that were "different" (e.g., paja-pajja) and "same" (e.g., paja-paja or pajja-pajja). Unlike in Experiment 1, the "same" words in each pair were physically identical and the "different" words in each pair always shared a physically identical frame (i.e., the words were identical except for artificial lengthening for duration contrasts and a spliced consonant for sibilant contrasts). This was done to ensure that any difference in the participants' responses resulted only from the manipulation of interest. The words in each pair were separated by an interstimulus interval of 750ms. Each pair was repeated twice throughout the experiment, which yielded a total of 392 pairs: 196 pairs with duration contrasts and 196 pairs with sibilant or other (filler) contrasts.

[3]This difference was introduced in order to account for the fact that intervocalic duration contrasts are perceptually harder for sonorants than for obstruents (Kawahara, 2007).

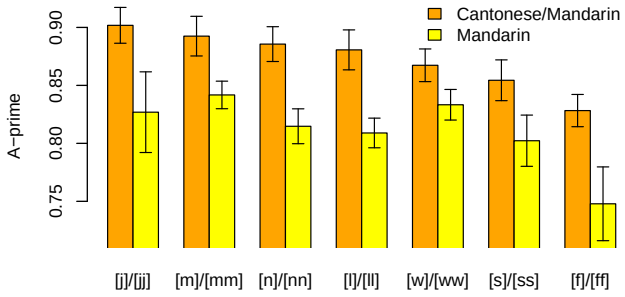


Figure 4: Performance on the long/short consonant contrast by Cantonese/Mandarin and Mandarin speakers. (Error bars are standard errors.)

Procedure The procedure was almost identical to experiment 1. The differences included the number of blocks (seven 56-trial blocks) and the response type: instead of clicking on the screen with a mouse, participants responded by pushing buttons on a game pad.

Results

As in Experiment 1, we calculated A-prime scores for each participant in each condition as a measure of contrast sensitivity (the same results hold for d-prime).

Duration contrast The results from the duration contrasts are plotted in Fig. 4. We analyzed the scores using a repeated measures ANOVA with the within-participants factor *segment* ([j], [w], [l], [m], [n], [f], or [s]), and the between-participants factor *language background* ('Mandarin' or 'Cantonese/Mandarin'). There was a significant main effect of *language background* $[F(1,38) = 12.7; p < .01]$ with the Cantonese/Mandarin group performing better $(\bar{A}' = 0.87)$ than the Mandarin group $(\bar{A}' = 0.81)$. There was also a significant main effects of *segment* $[F(6,228) = 5.8; p < .001]$, indicating that some segments were overall harder than others.

Sibilant contrast The results from the sibilant contrasts are plotted in Fig. 5. We analyzed the scores using a repeated measures ANOVA with the within-participants factor *segment* ([ɕ]/[ʂ], [ʑ]/[ʐ], [t͡ɕ]/[t͡ʂ], [d͡ʑ]/[d͡ʐ].), and the between-participants factor *language background* ('Mandarin' or 'Cantonese/Mandarin'). As predicted for this con-

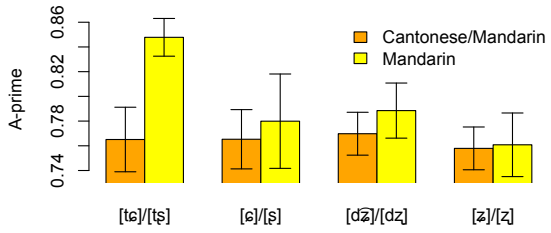


Figure 5: Performance on the sibilant contrast by Cantonese/Mandarin and Mandarin speakers. (Error bars are standard errors.)

trast, no significant main effects of *language background* $[F(1,38) = 1.6; p = .21]$ nor *segment* $[F(3,114) = 1.9; p = .14]$ were found. Both groups performed similarly on all sibilant contrasts, as illustrated in Fig. 5. There was, however, a tendency for Mandarin speakers to perform better than Cantonese speakers, at least on one type of contrast ([t͡ɕ]/[t͡ʂ]).

Discussion

This experiment showed that the speakers of Cantonese/Mandarin perform better on the short/long consonant duration contrast than Mandarin speakers without any exposure to Cantonese. It was hypothesized that such a difference between the two groups would be observed due to the fact that Cantonese uses *vowel* duration contrastively. This result suggests that Cantonese speakers were able to generalize their knowledge about a vowel duration contrast to a consonant duration contrast in a way that perception of the latter contrast was facilitated.

Importantly, the better performance of the Cantonese participants was not due to other differences in populations since the two groups performed equally well on the control contrast of sibilants. In this case, both groups were predicted to perform similarly due to the influence of Mandarin, which has a similar contrast between voiceless sibilants.

The combination of these results means that the feature of contrastive duration can indeed be abstracted away from a limited set of segments (e.g., vowels) and applied in novel conditions with a perceptual advantage, thus supporting the predictions of the model.

Conclusion

This paper argued that previously acquired linguistic knowledge can have a facilitative effect on perception and learning of new languages. In Experiment 1, we showed that participants with previous exposure to a language that uses consonant duration contrastively are better at discriminating a similar duration contrast in a novel language. Perceptual advantage was observed even if the contrast was presented in novel segmental contexts and for novel segments. Experiment 2 showed an even stronger result: perceptual advantage on consonant duration contrasts was observed for participants who only had previous exposure to *vowel* duration contrasts. Together, these results support the MLL model and the overhypothesis framework, which predict that knowledge of previously learned languages is generalized and leads language learners to make inferences about the ways in which languages are likely to vary. These inferences, or overhypotheses, about dimensions along which languages can differ may in turn increase learners' perceptual sensitivity to contrasts that the overhypotheses predict. Having established that such generalization occurs from previously learned to novel languages, the next step for future work will be to determine the exact conditions under which overhypotheses are made, and in what exact ways they are used in language learning.

Acknowledgments

For their helpful feedback on different stages of this project, the author thanks: Abdelhak Akjeje, Eric Baković, Klinton Bicknell, Sarah Creel, Tamar Gollan, Cindy Kilpatrick, Roger Levy, Sharon Rose, and four anonymous reviewers. Christopher Gaudiot and Rachel O'Sullivan helped with data collection for Experiment 1.

References

Bauer, R. S., & Benedict, P. K. (1997). *Modern Cantonese phonology*. Berlin/New York: Mouton de Gruyter.

Benus, S., Smorodinsky, I., & Gafos, A. (2003). Gestural coordination and the distribution of English 'geminates'. In *Proceedings of the 27th Annual Penn Linguistics Colloquium* (pp. 33–46). Philadelphia, PA: University of Pennsylvania.

Best, C. T. (1995). A direct realist view of cross-language speech perception. In W. Strange (Ed.), *Speech perception and linguistic experience: issues in cross-language research* (pp. 171–204). Timonium, MD: York Press.

De Angelis, G. (2007). *Third or additional language acquisition*. Clevedon: Multilingual Matters.

Flege, J. (1995). Second-language speech learning: theory, findings and problems. In W. Strange (Ed.), *Speech perception and linguistic experience: issues in cross-language research* (pp. 229–273). Timonium, MD: York Press.

Goodman, N. (1955). *Fact, fiction, and forecast*. Cambridge, MA: Harvard University Press.

Grier, J. B. (1971). Nonparametric indexes for sensistivity and bias. *Psychological Bulletin*, *75*(6), 424–429.

Kawahara, S. (2007). Sonorancy and geminacy. In *University of Massachusetts Occasional Papers in Linguistics 32: Papers in Optimality III* (pp. 145–186). Amherst: GLSA.

Kaye, A. (2005). Gemination in English. *English Today*, *21*, 43–55.

Klatt, D. H. (1976). Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *Journal of the Acoustical Society of America*, *59*(5), 1208-1221.

Kruschke, J. K. (1992). ALCOVE: an exemplar-based connectionist model of category learning. *Psychological Review*, *99*(1), 22–44.

Kuhl, P. K., & Iverson, P. (1995). Linguistic experience and the "perceptual magnet effect". In W. Strange (Ed.), *Speech perception and linguistic experience: issues in cross-language research* (pp. 121–154). Timonium, MD: York Press.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: a user's guide (2nd ed.)*. Mahwah, NJ: Lawrence Erlbaum Associates.

Mugitani, R., Pons, F., Fais, L., Werker, J. F., & Amano, S. (2008). Perception of vowel length by Japanese- and English-learning infants. *Developmental Psychology*, *45*(1), 236–247.

Ritchie, W. C., & Bhatia, T. K. (Eds.). (2009). *The new handbook of second language acquisition*. Bingley, UK: Emerald Group Publishing Limited.

Strange, W., & Shafer, V. L. (2008). Speech perception in second language learners: the re-education of selective perception. In J. G. Hansen Edwards & M. L. Zampini (Eds.), *Phonology and second language acquisition* (pp. 153–191). Amsterdam/Philadelphia: John Benjamins.

A rational account of perceptual compensation for coarticulation

Morgan Sonderegger (morgan@cs.uchicago.edu)
Department of Computer Science, University of Chicago, Chicago, IL 60637 USA

Alan Yu (aclyu@uchicago.edu)
Phonology Laboratory, Department of Linguistics, University of Chicago, Chicago, IL 60637 USA

Abstract

A model is presented that explains perceptual compensation for context as a consequence of listeners optimally categorizing speech sounds given contextual variation. In using Bayes' rule to pick the most likely category, listeners' perception of speech sounds, which is biased toward the means of phonetic categories (Feldman & Griffiths, 2007; Feldman, Griffiths, & Morgan, 2009), is conditioned by contextual variation. The effect on the resulting identification curves of varying category frequencies and variances is discussed. A simulation case study of compensation for vowel-to-vowel coarticulation shows the predictions of the model closely correspond to human perceptual data.

Keywords: Speech perception; perceptual compensation; rational analysis.

Introduction

A major challenge for models of speech perception is explaining the effect of context on phonemic identification. Depending on their acoustic, phonological, semantic, syntactic, and even socio-indexical contexts, identical acoustic signals can be labeled differently and different acoustic signals can be labeled identically. One of the most investigated types of contextual effects stems from phonemes' phonetic environments. Because of coarticulation, a phoneme's phonetic realization is heavily context-dependent. To understand speech, the listener must take into account context-induced coarticulatory effects to recover the intended message. The term *perceptual compensation* (PC) has often been used to characterize this type of context-induced adjustment in speech perception. For example, the identification of an ambiguous target syllable as /da/ or /ga/ is shifted by preceding /ar/ or /al/ contexts (Mann, 1980): the same /Ca/ token is *less* likely to be heard as /ga/ in /arCa/ context than in /alCa/ context. This effect has been argued to result from perceptual reduction of the coarticulatory fronting effects of /l/ on a following velar consonant: listeners are compensating for the effect of /l/ on /g/. This paper proposes a simple model in which PC effects emerge as an optimal solution to the problem of categorization in the presence of context-induced variation. In this model, listeners behave as if they are compensating because what is optimal differs by context.

PC effects have been observed in many phonetic settings. The fricative /ʃ/ has lower noise frequencies than /s/, and lip rounding lowers the resonant frequencies of nearby segments. Synthetic fricative noises ranging from /ʃ/ to /s/ are more often identified by English listeners as /s/ when followed by /u/ than by /a/ (Mann & Repp 1980; see also Mitterer 2006), presumably because listeners take into account the lowering effect of lip rounding from /u/ on the noise frequencies of /s/ in natural coarticulated speech. As another example, the perception of a fundamental frequency (f_0) contour can change as a function of vowel height (Hombert, 1978; Silverman, 1987) or consonant voicing (Pardo & Fowler, 1997): /i/ is perceived as lower in pitch relative to an /a/ with the same f_0, presumably because high vowels typically have higher f_0 than low vowels.

Listeners' language-specific experience crucially affects the degree of perceptual compensation. In a study replicated in part below, Beddor, Harnsberger, & Lindemann (2002) found that English and Shona listeners compensate for the coarticulatory effects of V_2 on V_1 in CV_1CV_2 sequences. That is, listeners identified a continuum of synthesized vowels between /a/ and /e/ more often as /a/ when the following vowel was /i/ than when the following vowel was /a/. Importantly, they observed that Shona listeners compensate more for the vowel contexts that triggered larger acoustic influences in speech production. Compensatory responses can affect listeners' rating judgments as well. English listeners are less accurate in judging vowel nasality in nasal than in non-nasal contexts, with nasal vowels in nasal contexts the most difficult (Beddor & Krakow, 1999; Kawasaki, 1986).

Explanations of PC effects have been advanced from several theoretical perspectives. Some emphasize the lexical and phonemic content of the context in determining the identification of the target sound (Elman & McClelland, 1988; Samuel & Pitt, 2003). Gestural theorists, who assume that listeners parse the acoustics in terms of its articulatory sources, argue that listeners attribute the acoustic properties of a target sound to the coarticulatory context rather than to the target (Fowler, 1996, 2006). Auditorists attribute context-induced shifts in category boundaries to general auditory processes such as frequency contrast or spectral contrast (Diehl & Kluender, 1989; Kingston, 1992; Kingston & Diehl, 1995; Lotto & Kluender, 1998). Such auditory explanations are unavailable for compensation effects such as vowel-dependent pitch height compensation (Fowler, 2006; Lotto & Holt, 2006). Motivated by such cases, Lotto & Holt (2006) suggest that the spectral contrast explanation be supplemented with a "general learning" mechanism for category formation from correlations between stimulus parameters.

The generality of PC effects is accentuated by evidence for contextual compensation with speech and non-speech sounds in human and non-humans (Holt, Lotto, & Kluender, 2000; Lotto, 2004). For example, when /da/–/ga/ syllables are preceded by tone glides matching in frequency to the third formant (F_3) transition of /al/ or /ar/, listeners' syllable identi-

fication responses shifted in the same direction as when targets were preceded by real speech (/al/ or /ar/). The same effect was observed even when steady-state tones at the offset frequency of /al/ or /ar/ F_3 were used (Lotto & Kluender, 1998; cf. Viswanathan, Fowler, & Magnuson, 2009) . Lotto, Kluender, & Holt (1997) conditioned four Japanese quails to exemplars of /da/ and /ga/ syllables. Two birds were trained to peck a key when presented with good /da/ exemplars and to not peck when presented with good /ga/ stimuli while two other quails were trained in the reverse condition (/ga/ positive, /da/ negative). After reaching a preset criterion of 10:1 ratio of pecks to positive versus negative stimuli, birds were presented with novel ambiguous CVs preceded by either /al/ or /ar/. All birds displayed a significant shift in peck rates across the change in preceding liquid. The /da/-positive birds pecked substantially more for CVs preceded by /ar/, while /ga/-positive birds pecked more for CVs preceded by /al/. Crucially, both the task and the results were essentially the same as in Mann (1980)'s experiment with human subjects. There is thus strong support for a language-independent, auditory mechanism of compensation.

In this paper, we develop a computational model of PC effects using rational analysis of speech perception and production. Rational analysis (RA; Anderson, 1990; Marr, 1982;) attempts to explain aspects of cognition as adaptive responses to the environment; its central claim is that much of people's behavior when performing some cognitive tasks can be understood as optimal, according to some criterion. RA represents a different type of explanation from existing theories of PC: instead of explaining the behavioral locus (e.g. gestural processing, lexical knowledge, general auditory processes) of PC effects, the model presented here gives an account of *why* PC effects occur, as a consequence of listeners optimally solving the problem of categorization given context-induced variation.

RA accounts have been developed for visual word recognition (Norris, 2006), spoken-word recognition (Norris & McQueen, 2008), perceptual magnet effects (Feldman & Griffiths, 2007; Feldman et al., 2009), and other cognitive domains, such as vision (Marr, 1982; Yuille & Kersten, 2006) and manual movement (Trommershäuser, Gepshtein, Maloney, Landy, & Banks, 2005). Our analysis of PC effects grows out of the rational model of perceptual magnet effects of Feldman et al. (2007, 2009). While "optimal" can be understood in Bayesian (e.g. Tenenbaum & Griffiths, 2001) or maximum likelihood (e.g. Fried & Holyoak, 1984) terms, following Feldman et al. and other recent rational accounts of speech perception (Clayards, Tanenhaus, Aslin, & Jacobs, 2008; Norris & McQueen, 2008), we use Bayesian inference here.

Model

Our rational model for PC effects assumes a simple scenario where an idealized optimal listener has to categorize some signal as one of two phonetic categories; this is analogous to the task listeners perform in the two-alternative forced choice (2AFC) paradigm commonly used in PC experiments. The model formalism is adapted from that used by Feldman et al. (2009). We differ in allowing model parameters to change with context, and focus on different aspects of the model's predictions.[1]

The modeled listener hears signal S in context k, and must decide whether it belongs to category c_1 or c_2. Listeners in this model assume S is normally distributed around a target pronunciation T, itself normally distributed around a category mean, and categorize based on the likelihood that S is an instance of the speaker producing an example from c_i in context k, with target T. Formally,

$$T \,|\, c_i, k \sim N(\mu_{c_i,k}, \sigma_c), \quad S \,|\, T, c_i \sim N(T, \sigma_S)$$

where $\mu_{c_i,k}$ is the mean of category i mean in context k, σ_C^2 is the variance in T around the category mean, and σ_S^2 is the variance in S around T. We assume for simplicity that σ_C and σ_S are the same for categories 1 and 2. Although we assume that T is the variable shifting by context, if it is instead assumed that S shifts by context in a similar way, all results turn out the same.[2] It thus does not matter under this analysis whether contextual variation is in the target pronunciation, T, or the acoustic signal itself, S.

The probability S comes from category c_1 can be calculated with Bayes' rule:

$$P(c_1 \,|\, S, k) = \frac{P(c_1 \,|\, k) P(S \,|\, c_1, k)}{P(c_2 \,|\, k) P(S \,|\, c_2, k) + P(c_1 \,|\, k) P(S \,|\, c_1, k)} \quad (1)$$

$P(c_i \,|\, k)$ is the probability of category i occurring in context k, i.e. in the lexicon as a whole. The $P(S \,|\, c_i, k)$ are calculated by integrating over all possible T, giving a logistic function:

$$P(c_1 \,|\, S, k) = \left(1 + \frac{f_2}{f_1} e^{b - Sg}\right)^{-1} \quad (2)$$

where

$$b = \frac{1}{2} \frac{\mu_{c_1,k}^2 - \mu_{c_2,k}^2}{\sigma_S^2 + \sigma_c^2}, \quad g = \frac{\mu_{c_1,k} - \mu_{c_2,k}}{\sigma_S^2 + \sigma_c^2}$$

and $f_i = P(c_i \,|\, k)$ is the frequency of category i in context k.

Studies of PC generally focus on locating the *crossover point*, where S is maximally ambiguous between categories, i.e. S' (see Fig. 1) such that $P(c_1 \,|\, S', k) = P(c_2 \,|\, S', k) = 0.5$. Solving from (2) gives

$$S' = \frac{\mu_{c_1,k} + \mu_{c_2,k}}{2} + \frac{\sigma_S^2 + \sigma_C^2}{\mu_{c_1,k} - \mu_{c_2,k}} \ln\left(\frac{f_2}{f_1}\right) \quad (3)$$

[1] Space constraints prevent us from giving detailed derivations below; these are given by (Feldman et al., 2009).

[2] Specifically, if we assume compensation is in S, of the form

$$T \,|\, c_i, k \sim N(\mu_{c_i}, \sigma_c), \quad S \,|\, T, c_i, k \sim N(T + \Delta_{i,k}, \sigma_S).$$

That is, the distribution of T varies by category, but is not affected by context. Given T, the distribution of S has a mean offset from T by an amount $\Delta_{i,k}$, which depends on the context.

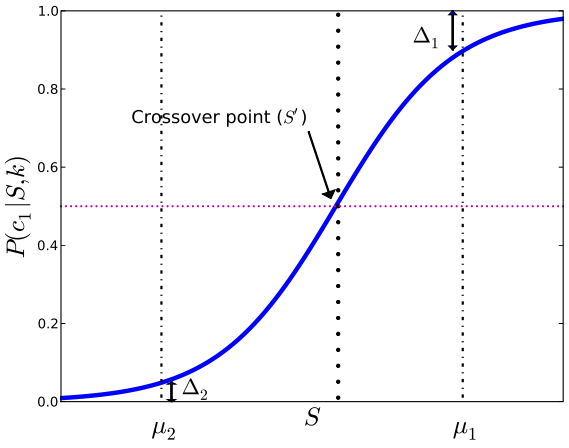


Figure 1: Schematic of a modeled identification curve. μ_1, μ_2 are category means, S' is the crossover point, and Δ_1, Δ_2 are miscategorization probabilities.

Perceptual compensation is thus captured in this model in terms of a shift in the crossover point as a function of the context. Note that if it is assumed that $f_1 = f_2$, S' is simply halfway between the category means, while if category frequencies are not equal ($f_1 \neq f_2$), S' is shifted.

The shape of the identification curve also changes as system parameters are changed. Two important properties of the curve, schematized in Fig. 1, are the slope at the crossover point and the misclassification probabilities at the category means.

The identification curve's slope at the crossover point is a rough measure of the "degree of uncertainty" (Clayards et al., 2008) of the category boundary:

$$\text{slope at } S' = \frac{dP(c_1\,|\,S,k)}{dS}\Big|_{S=S'} = \frac{\mu_{c_1,k} - \mu_{c_2,k}}{4(\sigma_S^2 + \sigma_c^2)}$$

The shallower the slope, the greater the uncertainty. The slope is steeper when the difference in category means is larger relative to category variances. Unlike the crossover point's location, the slope does not change depending on whether $f_1 = f_2$.

Categorization uncertainty can also be quantified as the *misclasssification probabilities* Δ_1 and Δ_2, defined as the probability a signal S produced at the mean of category i — a "perfect" exemplar from c_i — is misclassified. We find

$$\Delta_1 = (1 + \frac{f_1}{f_2} e^{\frac{(\mu_1-\mu_2)^2}{2V}})^{-1} \qquad \Delta_2 = (1 + \frac{f_2}{f_1} e^{\frac{(\mu_1-\mu_2)^2}{2V}})^{-1}$$

where $V = {\sigma_C}^2 + {\sigma_S}^2$. The misclassification probabilities decrease as the ratio of the difference in category means to the variance increases. When $f_1 > f_2$, Δ_1 decreases and Δ_2 increases (and vice versa for $f_1 < f_2$).

To illustrate the adequacy of the proposed model and its treatment of perceptual compensation, the next section reports the results of a simulation study of PC for anticipatory vowel-to-vowel coarticulation in English.

A Simulation Study

A modified replication study of Beddor et al. (2002)'s seminal perception and production study of vowel-to-vowel coarticulation in English was conducted. The perceptual results serve as the observed PC responses. These were compared to responses predicted by the rational model, using parameter values obtained from two production studies.

Observed perceptual responses

Eighteen native English speakers at the University of Chicago participated in a perception experiment, consisting of a training phase followed by a test phase. The training phase was intended to expose subjects to speech in which each of V_1=/a/ and V_1=/e/ was equally likely to occur in the context of following V_2=/a/ or V_2=/i/, corresponding to f_1=f_2 in our model. The test phase asked listeners to classify an ambiguous vowel V_1 as /a/ or /e/, in the context of V_2=/a/ or /i/.

In the training phase, listeners heard CV_1CV_2 tokens (C =/p/, /t/, or /k/, V = /e/ or /a/, V_2=/a/ or /i/). Tokens were constructed by splicing together CV syllables produced in isolation by an adult male speaker of English. A total of thirty-six tokens were constructed (=3C×2V_1×3C×2V_2). Each CV_1CV_2 token was heard ten times, for a total of 360 tokens, presented in random order. To encourage attention to the training stimuli, listeners performed a phoneme monitoring task where they were asked to identify whether or not each token contained a medial /t/.

In the test phase, listeners performed a 2AFC categorization task on V_1 in bV_1bV_2 context, with V_1 varying in F_1-F_3 along an 9-step /a-e/ continuum, and V_2=/a/ or /i/. The nine-step continuum was generated using Akustyk (Plichta & Preston, 2004), an add-on program for vowel analysis in Praat (Boersma & Weenink, 2001), by interpolating the formant values between two syllables (/ba/ and /be/) produced in isolation.[3] The test tokens were then created by splicing together each individual continuum syllable with either a /bi/ or a /ba/ syllable, also produced in isolation. The same speaker produced the speech stimuli used in both the training and test phases. Each subject heard each test stimulus ten times, for a total of 180 tokens, presented in random order. Subjects were paid a nominal fee to participate in the studies.

Fig. 2 shows empirical curves of the proportion of V_1=/a/ responses in V_2=/a/ and V_2=/i/ contexts, as a function of position on the V_1 continuum. Error bars correspond to 95% confidence intervals over individual-subject proportions.

The V_1 categorization responses (1=/a/) were modeled using a mixed-effects logistic regression (Baayen, 2008; Jaeger,

[3]The F_1 values of the nine steps along the /a-e/ continuum were 713Hz, 682Hz, 635Hz, 606Hz, 592Hz, 563Hz, 522Hz, 500Hz, and 483 Hz. Values for the higher formants were adjusted as well to create a more natural-sounding continuum. For simplicity, we focus on the coarticulatory effect on F_1 since the context vowels only vary in height and not in backness.

2008) with VOWEL CONTEXT (/a/ or /i/) and CONTINUUM (1–9) as fixed effects, and random effects of SUBJECT and BLOCK (test token number) on the intercept. As a measure of model quality, Nagelkerke's pseudo-R^2 was 0.64, relative to a model with only the intercept. There were significant effects of CONTINUUM and VOWEL CONTEXT ($p < 0.001$), as well as their interaction ($p < 0.05$) The effect of VOWEL CONTEXT was an increase in V_1=/a/ responses for V_2=/i/ compared to V_2=/a/, in agreement with the results of Beddor et al. (2002): native English listeners appear to perceptually compensate for the coarticulatory effects of a following vowel.

Model-predicted perceptual responses

To predict expected identification curves using Eqn. 2, we need the category means of /a/ and /e/ (V_1) in the context of following /a/ or /i/ (V_2), and category variances for V_1 in V_2=/a/ and V_2=/i/ contexts.[4] (Recall that we are assuming equal variances of V_1=/a/ and V_1=/i/, given the following context.) Eqn. 2 also includes the relative probability (f_1/f_2) of V_1=/a/ and V_1=/i/ in each V_2 context. We assume that $f_1/f_2 = 1$ following the training phase.

The category mean and variance parameters were estimated from two production studies. Category means were based on 40 productions of the form bV_1bV_2 (10 for each combination of $V_1 \in$ {a,e} and $V_2 \in$ {a,i}) by the speaker whose speech was the basis of the training and test tokens. Category variances were calculated from productions of initial stressed /adV_1CV_2/ sequences ($V_{1\&2}$=/a/, /e/, or /i/ and C=/p/ or /b/), each repeated ten times in random order, by four male, phonetically-trained native English speakers. No subjects who participated in the perception experiment participated in the production studies as well.

We thus assumed that during the experiment, subjects adjusted their expectation of category means to match the speaker they were hearing, but that their category *variances* reflected variation across speakers.[5]

For all production data, formant values were measured at the midpoint of the target V_1. Means and variances were calculated over Bark-transformed F_1 values for V_1. Variances for V_1 when V_2=/a/ were taken to be the mean of the variances for /aCa/ stimuli and for /eCa/ stimuli. Variances for V_1 when V_2=/i/ were calculated similarly. The resulting model parameters are listed in Table 1.

The predicted identification curves for V_2=/a/ and V_2=/i/ contexts are given in Fig. 2. For comparison with the experimental results, Step 1 was taken to be the mean of μ_{c_2} (where c_2 is "V_1=/e/") in V_2=/a/ and V_2=/i/ contexts, and Step 9 was taken to be the mean of μ_{c_1} (where c_1 is "V_1=/a/") in V_2=/a/ and V_2=/i/ contexts.

Table 1: Model parameters obtained from the production study, where c_1 is "V_1=/a/", c_2 is "V_1=/e/." B=Bark.

V_2	μ_{c_1}	μ_{c_2}	${\sigma_C}^2 + {\sigma_S}^2$
/a/	6.69 B	4.67 B	0.568 B^2
/i/	6.76 B	4.26 B	0.619 B^2

Qualitatively, the fit between the experimental and model-predicted curves in Fig. 2 is very good, without fitting any free model parameters to the production data. Both experimental and model curves show a rightward shift for V_2=/a/ context, and the predicted slope at the crossover point for both pairs of curves are approximately the same.[6] However, the quality of the fit depends on how rational model parameters are derived from the production study, and should be interpreted with caution. For example, category variances ($\sigma_C^2 + \sigma_S^2$) would be smaller if based on tokens from a single speaker rather than several speakers, making the slope of the rational model curves steeper.

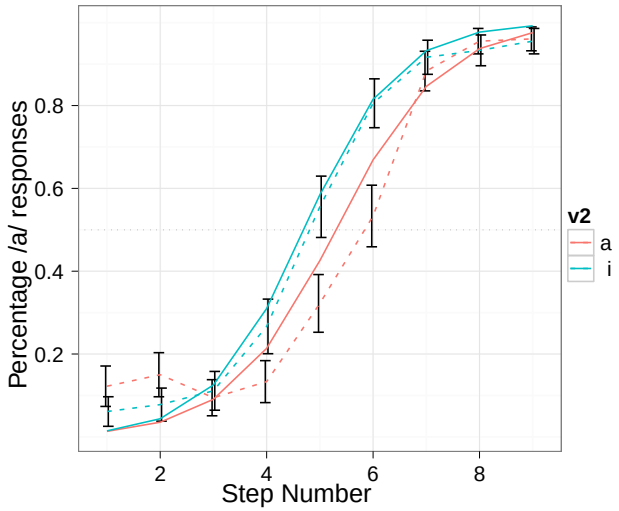


Figure 2: Dashed lines: Proportion of /a/ responses for V_2=/a/ (right curve) and V_2=/i/ (left curve) contexts, across all subjects. Error bars are 95% confidence intervals, based on individual-subject proportions. Solid lines: Predicted identification curves, based on production data. Dotted line: Crossover point (rate=0.5).

Discussion

We have illustrated a rational model of perceptual compensation effects and shown that, given a simple probabilistic model for the observed values of an acoustic-phonetic cue

[4] Nearey & Hogan (1986) propose two models for deriving identification curves from production data. Their 'NAPP' model is similar to the present model, but is not derived from an RA viewpoint. We also map production data to model parameters differently.

[5] Another interpretation of these category variances, suggested by a reviewer, is that subjects assume the tokens have category variances typical of a single speaker, but also account for some "noise" in perception, beyond the variance observed in the production data of an individual speaker.

[6] The correlation between the two sets of curves is very high ($r = 0.987$, $p < 0.001$), indicating good qualitative agreement.

(here, F_1 values) associated with a speech sound, it is possible to understand perceptual compensation as an idealized rational listener arriving at an optimal solution based on evidence from prior experience. In this model, by choosing the most probable categorization response given the context, based on their knowledge of the probability distribution of the relevant cue in that context, listeners appear to 'undo' the effect of coarticulation. Different contexts are associated with different cue distributions, and hence difference categorization responses.

Rational models provide a very general expression of the computational problem being solved when performing some cognitive task, and are largely orthogonal to proposed mechanisms by which the task is performed. Our model proposes an abstract explanation for why PC occurs, but is compatible with a role for different proposed mechanisms for PC effects via "prior knowledge" encoded in the cue distributions and category frequencies. The model assumes that listeners have different cue distributions for different contexts, but does not specify the source of the distributions; it could be that knowledge about gestures or general auditory capabilities generate or underly the distributions. The category frequencies could reflect knowledge of lexical or phonotactic probabilities, as pointed out by Feldman et al. (2009).

The model is able to accommodate two types of PC effects — language-dependent and domain-general — usually emphasized in opposing accounts of PC. That PC effects are language-dependent is expected because many coarticulatory effects are language-specific. Since language-specific coarticulatory effects are reflected in acoustic-phonetic cues, listeners' categorization responses should mirror the (language-specific) probability distributions of the relevant cues. The model is general in that it is not restricted to linguistically-relevant acoustic cues. As long as a non-linguistic acoustic cue has a probability distribution, the idealized rational listener (human or non-human) would seem to compensate in the same way as she would if the acoustic cue were linguistic.

Our model predicts that compensation effects could be ameliorated or even reversed via adjustments to the model parameters. In general, an observed PC effect corresponds to different values of S' (the crossover point) in different contexts, say k_1 and k_2. The second term of (3) predicts that S' in k_1 and S' in k_2 depend on the relative frequencies of c_1 and c_2 in these contexts. Thus, if f_2/f_1 differs significantly by context, the context-dependent PC effect can be exaggerated, diminished, canceled, or even reversed. Failure to compensate could therefore occur for sudden change in f_2/f_1 for k_1 but not k_2. Since this proposed effect depends on the second term of (3), compensation could also be undone by changes in variances ($\sigma_C^2 + \sigma_S^2$) or category mean differences ($\mu_{c_1,k} - \mu_{c_2,k}$) for k_1 versus k_2. We are currently running experiments to test the predicted effects of category frequency on compensation.

This understanding of PC failure has serious implications for current theories of sound change. Many researchers, most notably Ohala (1993), argue that articulatory and perceptual factors shape phonological systems through listener misperception-induced sound changes, and that the synchronic typology of sound patterns is a consequence of the phonologization of such phonetic "precursors" (Barnes, 2006; Blevins, 2004; Blevins & Garrett, 1998, 2004; Kavitskaya, 2001; Yu, 2004). That is, sound change occurs when listeners mistake as representative of the speaker's target pronunciation the effects of the speakers' production system, the listeners' own perceptual system, or ambient distortion of the acoustic stream. However, this account assumes that errors in perception (i.e. failure to compensate for contextual variation) lead to adjustments in perceptual and production norms. The fact that perceptual compensation is observed so robustly in speech raises questions about the feasibility of this type of model of sound change. Earlier work has assumed that failure to compensate for contextually-induced variation occurs when listeners do not detect the conditioning context. Our model suggests that the relative magnitude of compensation can be mediated by properties of the language's lexicon (e.g. the relative frequencies of phones) as well as speakers' prior experience with the language (e.g. pronunciation variation). That is, given certain lexical or contextual conditions, a change in compensatory response may take place even when the conditioning contextual information is accurately perceived.

Conclusion

The model proposed here allows the incorporation of both speech-specific and general auditory factors. It proposes that perceptual compensation effects emerge as a consequence of an optimal response to the problem of categorization in the presence of context-induced variation. To be sure, the present model is simplistic, and only a first step toward modeling compensatory phenomena in general. Future work will develop more general models, e.g. with unequal category variances and multiple (>2) categories, and explore their effects on predicted categorization behavior. Nonetheless, the present model contributes to the growing number of studies that attempt to understand speech perception from a rationalist point of view (Clayards et al., 2008; Feldman & Griffiths, 2007; Feldman et al., 2009; Norris & McQueen, 2008).

Acknowledgments We thank Matt Goldrick and James Kirby for comments on an earlier version of this paper, Max Bane for statistics discussion, and Ed King for setting up the experiment. Part of this work was presented at the 2009 Linguistic Society of America meeting. MS is supported by a Department of Education GAANN fellowship.

References

Anderson, J. (1990). *The Adaptive Character of Thought*. Hillsdale, N.J.: Erlbaum.

Baayen, R. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: CUP.

Barnes, J. (2006). *Strength and weakness at the interface: Positional neutralization in phonetics and phonology*. Berlin: Mouton de Gruyter.

Beddor, P., Harnsberger, J., & Lindemann, S. (2002). Language-specific patterns of vowel-to-vowel coarticulation: Acoustic structures and their perceptual correlates. *Journal of Phonetics*, *30*(4), 591–627.

Beddor, P., & Krakow, R. (1999). Perception of coarticulatory nasalization by speakers of English and Thai: Evidence for partial compensation. *Journal of the Acoustical Society of America*, *106*, 2868–2887.

Blevins, J. (2004). *Evolutionary Phonology*. Cambridge: CUP.

Blevins, J., & Garrett, A. (1998). The origins of consonant-vowel metathesis. *Language*, *74*(3), 508–56.

Blevins, J., & Garrett, A. (2004). The evolution of metathesis. In B. Hayes, R. Kirchner, & D. Steriade (Eds.), *Phonetically-based phonology.* Cambridge: CUP.

Boersma, P., & Weenink, D. (2001). Praat, a system for doing phonetics by computer. *Glot International*, *5*(9/10), 341–345.

Clayards, M., Tanenhaus, M., Aslin, R., & Jacobs, R. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition*, *108*, 804–809.

Diehl, R., & Kluender, K. (1989). On the objects of speech perception. *Ecological Psychology*, *1*(2), 121–144.

Elman, J., & McClelland, J. (1988). Cognitive penetration of the mechanisms of perception: Compensation for coarticulation of lexically restored phonemes. *Journal of Memory & Language*, *27*(2), 143–165.

Feldman, N., & Griffiths, T. (2007). A rational account of the perceptual magnet effect. In D. McNamara & J. Trafton (Eds.), *Proceedings of the 29th Annual Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Feldman, N., Griffiths, T., & Morgan, J. (2009). The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference. *Psychological Review*, *116*(4), 752–782.

Fowler, C. (1996). Listeners do hear sounds, not tongues. *Journal of the Acoustical Society of America*, *99*, 1730–1741.

Fowler, C. (2006). Compensation for coarticulation reflects gesture perception, not spectral contrast. *Perception & Psychophysics*, *68*(2), 161–177.

Fried, L., & Holyoak, K. (1984). Induction of category distributions: A framework for classification learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *10*(2), 234–257.

Holt, L., Lotto, A., & Kluender, K. (2000). Neighboring spectral content influences vowel identification. *Journal of the Acoustical Society of America*, *108*, 710–722.

Hombert, J.-M. (1978). Consonant types, vowel quality, and tone. In V. Fromkin (Ed.), *Tone: a linguistic survey.* New York: Academic Press.

Jaeger, T. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*(4), 434–446.

Kavitskaya, D. (2001). *Compensatory Lengthening: Phonetics, phonology, diachrony*. Unpublished doctoral dissertation, University of California at Berkeley.

Kawasaki, H. (1986). Phonetic explanation for phonological universals: The case of distinctive vowel nasalization. In J. Ohala & J. Jaeger (Eds.), *Experimental phonology.* Orlando: Academic Press.

Kingston, J. (1992). The phonetic and phonology of perceptually motivated articulatory covariation. *Language & Speech*, *35*, 99–114.

Kingston, J., & Diehl, R. (1995). Intermediate properties in the perception of distinctive feature values. In B. Connell & A. Arvaniti (Eds.), *Phonology and Phonetic Evidence: Papers in Laboratory Phonology IV.* Cambridge: CUP.

Lotto, A. (2004). Perceptual compensation for coarticulation as a general auditory process. In A. Agwuele, W. Warren, & S.-H. Park (Eds.), *Proceedings of the 2003 Texas Linguistics Society Conference.* Somerville, MA: Cascadilla.

Lotto, A., & Holt, L. (2006). Putting phonetic context effects into context: A commentary on Fowler (2006). *Perception & Psychophysics*, *68*(2), 178–183.

Lotto, A., & Kluender, K. (1998). General contrast effects in speech perception: effect of preceding liquid on stop consonant identification. *Perception & Psychophysics*, *60*(4), 602–19.

Lotto, A., Kluender, K., & Holt, L. (1997). Perceptual compensation for coarticulation by Japanese Quail (*Coturnix coturnix japonica*). *Journal of the Acoustical Society of America*, *102*, 1134–1140.

Mann, V. (1980). Influence of preceding liquid on stop-consonant perception. *Perception & Psychophysics*, *28*(5), 407–12.

Mann, V., & Repp, B. (1980). Influence of vocalic context on perception of the [ʃ]-[s] distinction. *Perception & Psychophysics*, *28*(3), 213–28.

Marr, D. (1982). *Vision*. San Francisco: W.H. Freeman.

Mitterer, H. (2006). On the causes of compensation for coarticulation: Evidence for phonological mediation. *Perception & Psychophysics*, *68*(7), 1227–1240.

Nearey, T., & Hogan, J. (1986). Phonological contrast in experimental phonetics: Relating distributions of measurements production data to perceptual categorization curves. In J. Ohala & J. Jaeger (Eds.), *Experimental phonology.* Orlando: Academic Press.

Norris, D. (2006). The Bayesian reader: Explaining word recognition as an optimal Bayesian decision process. *Psychological Review*, *113*(2), 327–57.

Norris, D., & McQueen, J. (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review*, *115*(2), 357–395.

Ohala, J. (1993). The phonetics of sound change. In C. Jones (Ed.), *Historical linguistics: Problems and perspectives.* London: Longman.

Pardo, J., & Fowler, C. (1997). Perceiving the causes of coarticulatory acoustic variation: consonant voicing and vowel pitch. *Perception & Psychophysics*, *59*(7), 1141–52.

Plichta, B., & Preston, D. (2004). Akustyk for Praat (Version 1.7.2) [Computer software manual]. East Lansing, MI.

Samuel, A., & Pitt, M. (2003). Lexical activation (and other factors) can mediate compensation for coarticulation. *Journal of Memory & Language*, *48*(2), 416–434.

Silverman, K. (1987). *The structure and processing of fundamental frequency contours*. Unpublished doctoral dissertation, Cambridge University.

Tenenbaum, J., & Griffiths, T. (2001). Generalization, similarity, and Bayesian inference. *Behavioral & Brain Sciences*, *24*, 629–640.

Trommershäuser, J., Gepshtein, S., Maloney, L., Landy, M., & Banks, M. (2005). Optimal compensation for changes in task-relevant movement variability. *Journal of Neuroscience*, *25*(31), 7169–7178.

Viswanathan, N., Fowler, C., & Magnuson, J. (2009). A critical examination of the spectral contrast account of compensation for coarticulation. *Psychonomic Bulletin & Review*, *16*(1), 74–79.

Yu, A. (2004). Explaining final obstruent voicing in Lezgian: Phonetics and history. *Language*, *80*(1), 73–97.

Yuille, A., & Kersten, D. (2006). Vision as Bayesian inference: Analysis by synthesis? *Trends in Cognitive Sciences*, *10*(7), 301–308.

Effects of Pragmatic Inference on Phoneme Identification

Hannah Rohde (hannah@northwestern.edu)
Department of Linguistics, 2016 Sheridan Road
Evanston, IL 60208 USA

Marc Ettlinger (marc@northwestern.edu)
Department of Communication Sciences and Disorders, 2240 Campus Drive
Evanston, IL 60208 USA

Abstract

Although previous research has established that multiple top-down factors guide the identification of sounds during speech processing, the ultimate range of interaction across levels of linguistic structure is still unknown. In a set of experiments, we investigate whether interactive effects emerge between the two most disparate domains: pragmatic inference and acoustic speech perception. We use contexts that trigger pragmatic expectations regarding upcoming coreference (expectations for either *he* or *she*), and, in those contexts, we test listeners' identification of phonetic category boundaries (using words on the /hi/~/ʃi/ continuum). The results indicate that pragmatic inference can indeed alter listeners' identification of phonetic categories.

Keywords: Phonetics, pragmatics, categorical perception, pronoun interpretation, implicit causality

Introduction

There is a growing body of evidence suggesting that language processing requires the integration of multiple sources of linguistic knowledge across multiple levels of linguistic structure. These relevant knowledge sources range from low-level features of the acoustic signal, through lexical and morpho-syntactic properties of words and phrases, up to higher-level semantic and pragmatic inferences about the speaker's intended message (e.g., Ganong 1980 and Pitt 1995 for lexical effects on phoneme perception; Spivey & Tanenhaus 1998 for lexical effects in syntactic processing; and van Berkum, Brown, & Hagoort 1999 for pragmatic effects in syntactic processing, among many others). Occupying the far ends of this spectrum are phonetics and pragmatics. Therefore, we submit that identifying contexts in which comprehenders bring together cues from these two very distinct domains would provide a strong demonstration of the maximum extent of this interactivity.

Our experiments test for interactive effects at the pragmatic-phonetic interface in contexts in which listeners' comprehension of an acoustically ambiguous word might reflect pragmatic biases of the discourse context. To do this, we use words whose interpretation is inherently discourse dependent—namely, personal pronouns. Based on existing pragmatics work on pronoun interpretation, we use contexts in which listeners have been shown to anticipate subsequent reference to a particular referent. We then capitalize on the fact that the English third person pronouns *he* and *she* constitute minimal pairs in order to construct acoustically ambiguous words that vary along a *h~sh* continuum. The results we find attest to the extent of interactive effects that any successful language processing model must capture. The results also contribute to the well-established literature on phoneme identification by broadening the set of known factors that can influence processing.

Modeling Pragmatic Interaction

Interactive approaches to processing are characterized by models "in which lexical, structural (syntactic) and interpretive knowledge sources communicate and interact during processing in an optimally efficient and accurate manner" (Marslen-Wilson & Tyler 1980). Existing work has identified top-down pragmatic effects, i.e. interaction, within syntactic processing, but interactive effects between pragmatic and phonetic information sources have not, to our knowledge, been demonstrated before.

Early work demonstrated the effect of pragmatic factors on other levels of sentence processing, showing that appropriate discourse contexts can eliminate syntactic garden paths (Crain & Steedman 1985; Altmann & Steedman 1988). This work established that comprehenders treat material following a definite NP (*The horse* in *The horse raced past the barn fell*) differently depending on the number of available referents (the number of horses) in the discourse context. These contextual effects have been attributed to a felicity constraint that requires that a definite NP have a unique and identifiable referent—a constraint that encourages comprehenders to interpret post-nominal material (*raced past the barn*) as NP modification rather than a main verb. Referential context has also been shown to yield online effects in syntactic processing (Ni, Crain, & Shankweiler 1996; van Berkum, Brown, & Hagoort 1999; Sedivy 2002). These results lend support to models of incremental processing in which comprehenders have access to pragmatic information before sentence-internal syntactic decisions have been fully resolved. Our work also relies on referential biases, but we push the extent of interactivity further by showing that discourse context can influence the identification of a phonetic category boundary.

Modeling Contextual Effects in Phonetics

Existing work on the factors that influence phoneme identification has established that listeners use more than

just the acoustic signal. The contextual factors that have been shown to have an impact include cues such as lexical status, syntactic category, and semantic congruity. The influence of such contextual factors can be captured both in models that permit top-down contextual information to impact sound perception directly, as in McClelland & Elman's (1986) TRACE model as well as in models in which the perceptual system operates fully independently from other levels of language processing and top-down factors only exert an influence at the point of lexical decision, as in Norris, McQueen, & Cutler's (2000) Merge model. Models like TRACE permit interaction at all levels whereas models like Merge attribute top-down effects to the integration of multiple information sources when a lexical decision is made. We use the term *interactive effects* here to refer to listener responses that reflect biases from information sources at different levels of linguistic structure, but we do not distinguish between the interaction and integration accounts (for discussion of this debate as well as methods for distinguishing the two approaches, see Norris et al. 2000, Magnuson, McMurray, Tanenhaus, & Aslin 2003, and Samuel & Pitt 2003). Our primary goal here is to extend the observed range of top-down effects beyond the previously reported lexical, syntactic, and semantic levels.

Contextual effects based on lexical status were first shown by Ganong (1980) in experiments that established that listeners' phonetic category judgments can be influenced by the lexical status of the stimulus: Ambiguous sounds along the /t/~/d/ continuum are more likely to be reported as /t/ when presented as the onset in a *task~dask* continuum and are more likely to be reported as /d/ when presented as part of a *tash~dash* continuum.

Phonetic category judgments are also sensitive to syntactic context: Acoustically ambiguous words along the *to~the* continuum are more likely be reported as *to* in contexts with a verb, as in *We tried to go*, than in contexts with a noun, as in *We tried the gold* (Isenberg, Walker, & Ryder 1980; see also van Alphen & McQueen 2001)

Furthermore, there is evidence that ambiguous sounds are interpreted differently depending on the semantic congruity of the target word in a particular context: Ambiguous sounds along the *path~bath* continuum are more likely to be reported as */p/* in the context *She likes to jog along the…* and are more likely to be reported as */b/* in the context *She needs hot water for the…*(Miller, Green, & Schermer 1984). Miller et al. report, however, that semantic congruity effects disappear when the task requires listeners to focus only on the target word, rather than on the full sentence frame.

One way of understanding these syntactic and semantic effects is to assume that a particular interpretation of the acoustically ambiguous item is more accessible or more strongly activated given the surrounding lexical items. In other words, lexical items like *go* and *gold* constrain the part of speech of the preceding word. Similarly, contexts that mention *hot* and *water* activate the word *bath*, whereas contexts that mention *jogging* activate *path*. These associations can be said to reflect comprehenders' syntactic knowledge and their mental models of particular events and event participants. As such, these results point to the dynamic integration of information sources ranging from hierarchical syntactic structures to real-world event knowledge. However, these associations may also be attributed to simple co-occurrence frequencies (see Willits, Sussman, & Amato 2008 for a co-occurrence-based account of data that has previously been taken to support highly interactive models). That is, it is possible that these results do not reflect listeners' understanding or parsing of the context in question, but rather reflect statistical frequencies over adjacent words.

The results presented in this paper go beyond this previous work in several important ways. In our contexts, we simultaneously hold constant both the lexical status of our target items and their syntactic category. Furthermore, our target items can be considered semantically neutral in that they are used across semantic contexts and their relationships to other words in the context do not reduce to co-occurrence frequencies.

Our experiments demonstrate that phoneme identification is sensitive to pragmatic inferences about referents in the discourse context and to domain-general causal reasoning. First, Experiment 1 replicates the Ganong effect of phonetic~lexical interaction for the /h/~/ʃ/ continuum. Experiments 2 and 3 use a novel design to test whether listeners' pragmatic expectations can influence phonetic category identification. For the second and third experiments, lexical status is not at issue because all acoustically ambiguous sounds yield legitimate lexical items, allowing us to attribute the effects we observe to interactive effects between pragmatic and phonetic cues.

Experiment 1: Ganong Replication for */hi/~/ʃi/*

In order to establish that the /h/~/ʃ/ continuum is a valid one for assessing phonetic category perception, we first obtained a measure of the effect of lexical status on acoustic perception by replicating the Ganong effect for /h/~/ʃ/. We tested whether listeners would judge the ambiguous onset of a monosyllabic item (e.g., /□ik/) as more /ʃ/-like if the English lexicon contains a word with a /ʃ/- onset (e.g., *sheik*) and lacks a corresponding word with a /h/- onset (**heik, *heek*).

Methods

Participants 35 native English-speaking Northwestern University undergraduates received either $6 or course credit for their participation in the study. A subset of these participants also completed Experiments 2 and 3. Note that this experiment, labeled here as Experiment 1, was always completed as the last part of the experiment session if the session included multiple tasks.

Materials Six pairs of items were created such that each pair consisted of a word and a non-word. The pairs *sheik/*heik*, *sheen/*heen,* and *sheaf/*heaf* were the /ʃ/-

biasing pairs in which the /ʃ/- onset constituted a word. The pairs *heeds/*sheeds, heels/*sheels,* and *heave/*sheave* were the /h/-biasing pairs in which the /h/- onset constituted a word. Onsets ranged from /h/ to /ʃ/ along a 20-step acoustic continuum. Unlike the /t/~/d/ continuum, which can be generated by varying a single acoustic parameter, namely voice onset time, /h/~/ʃ/ is not differentiated by a single simple parameterizable acoustic variable. Therefore, we combined two naturally produced tokens of *he* and *she* at varying intensities (McGuire, 2007; Munson & Coyne, *in press*). The duration of the fricative portion, which may also serve as a cue to differentiate these two items, was the average of the */hi/* and /ʃi/ tokens. Items were constructed such that each of the 6 pairs appeared with each of the 20 /hi/~/ʃi/ steps. Participants heard all items twice.

Procedure Participants listened to the items through headphones while sitting in a sound-attenuating booth. For each item, they were asked to indicate using a button box whether the onset of the item sounded more *h*-like or more *sh*-like on a 4-point scale.

Results and Discussion

As predicted, listeners were more likely to report hearing an initial /ʃ/ for /ʃ/-biasing items (items on the *sheik~heik*, *sheen~heen,* and *sheaf~heaf* continua; mean score=2.9, where 1 is /h/ and 4 is /ʃ/) than for */hi/*-biasing items (items on the *heeds~sheeds, heels~sheels,* and *heave~sheave* continua; mean score=2.4). There was a main effect of lexical status with the data collapsed across steps ($F(1,33)=192.737$, $p<0.001$). The results are shown in Figure 1 with error bars for standard error of the mean.

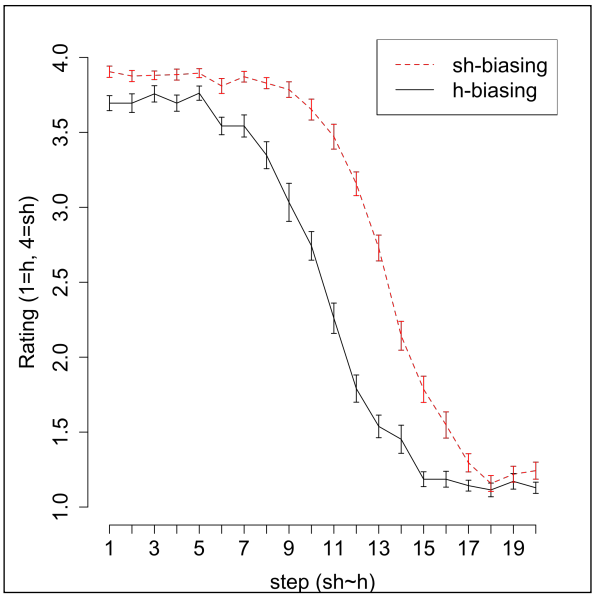


Figure 1: Impact of lexical status on perceived phonetic category for /hi/~/ʃi/ continuum in Experiment 1

The main effect of lexical status replicates the effect originally observed by Ganong for the /t/~/d/ continuum, extending the effect to the /h/~/ʃ/ continuum. Because a subset of the participants had already participated in Experiments 2 and 3 during the experiment session, we also compared performance based on prior experiment participation. There was no difference between participants who had only participated in Experiment 1 and those that had participated in multiple experiments ($F<1$).

Experiment 2: Referential Context

In our first examination of whether listeners' pragmatic knowledge and reasoning influences their phonetic category perception, we used contexts in which all contextually relevant referents were of the same gender in order to see if referential context biased listeners' interpretation of a subsequent acoustically ambiguous pronoun. If listeners do not combine pragmatic and phonetic information in determining phonetic category membership, we would expect to see category assignments based only on the acoustic input of the pronoun, regardless of referential context. On the other hand, if listeners can combine pragmatic and phonetic cues and if pragmatic information is available when listeners are making phonetic category decisions as part of the interpretation of words in full-sentence discourse contexts, we would expect to see category assignments that differ by context.

Methods

Participants 26 native English-speaking Northwestern undergraduates participated. All individuals went also participated in Experiments 1 and 3 during the same session.

Materials 40 sentences were constructed consisting of two clauses connected by *because*. The first clause introduced two individuals of the same gender and the second clause contained an acoustically ambiguous pronoun, as in (1-2).

(1) *he*-biasing context:
Luis reproached Joe because ☐ hadn't done the work**.**

(2) *she*-biasing context:
Joyce helped Sue because ☐ was up against a deadline.

If listeners infer that the discourse context is limited to the two named individuals in the first clause, then the pronoun in the second clause must be linked to an antecedent that is matched for gender. Because the two available referents in the discourse context were of the same gender, the sentences strongly bias the interpretation of the acoustically ambiguous pronouns to *he* in contexts like (1) or *she* in contexts like (2). We normed a total of 20 steps along the /hi/~/ʃi/ continuum (using the /hi/~/ʃi/ component in isolation, not in sentential contexts) to find steps that were centered around the point of maximum ambiguity for listeners. From those 20, we selected a smaller set of 5 steps for testing in order to increase the number of trials at each

data point without repeating items. Each sentence contained a pronoun consisting of one of the 5 /hi/~/ʃi/ steps. We manipulated gender bias within subjects and between items. Participants heard all items once.

Procedure Participants listened to the sentences through headphones while sitting in a sound-attenuating booth. For each item, they were asked to indicate on a button box whether the sentence mentioned *he* or *she*, using a 4-point scale. After each sentence participants were asked a *yes/no* comprehension question based on the sentence's meaning (but not the interpretation of the pronoun) to ensure they were focused on understanding the sentence and not simply focused exclusively on the ambiguous phoneme.

Results and Discussion

Only trials where participants correctly answered the comprehension question were included in the results. As predicted by an interactive account, we found that items with *she*-biasing contexts that contained only female referents yielded higher *she* ratings (mean score=2.3 where 1 is *he* and 4 is *she*) than *he*-biasing contexts that contained only male referents (mean score=1.6). There was a main effect of gender context with the data collapsed across steps ($F(1,26)=37.860$, $p<0.005$). The results appear in Figure 2.

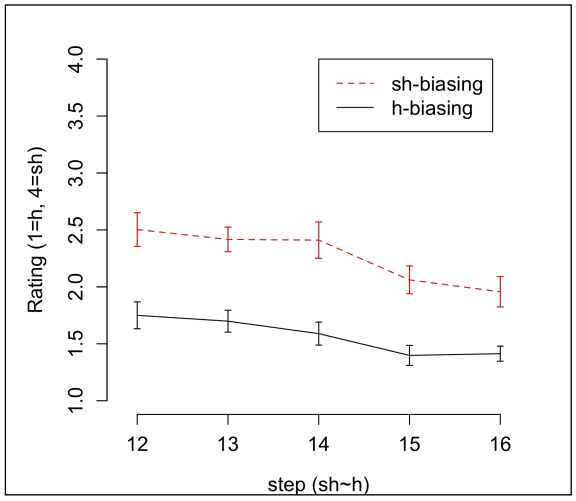


Figure 2: Impact of referential context on perceived phonetic category in Experiment 2

These results support a model of processing in which pragmatic biases are brought to bear on decisions regarding phonetic category membership, at least to the extent that referential context influences listeners' expectations about which individual will be mentioned next.

One question that can be raised regarding Experiment 2 is whether the experiment actually tests listeners' pragmatic reasoning or whether the results can also be explained by semantic neighborhood or co-occurrence effects. Sentences that contain female names may simply be more likely to contain the word *she*, and sentences that contain male names may be more likely to contain the word *he*. Given this concern, Experiment 3 uses contexts in which a female name and a male name are both present. Instead of relying on a single-gender referential context, Experiment 3 uses listeners' pragmatic reasoning about event causality in order to shift co-reference biases.

Experiment 3: Causal Reasoning

In order to construct contexts in which domain-general aspects of pragmatic reasoning might influence sound perception, we used sentences containing verbs from the class of so-called implicit causality verbs (Garvey & Caramazza 1974, *inter alia*). These verbs have been shown to guide listeners' coreference expectations by describing events in which one participant (either the subject or object, depending on the verb) is implicated as central to the event's cause and is thus likely to be re-mentioned in a subsequent *because* clause.

Methods

Participants 26 native English-speaking Northwestern University undergraduates participated. All individuals also participated in Experiments 1 & 2 during the same session. This experiment was completed as the first task.

Materials 40 sentences were constructed consisting of two clauses connected by *because*. The first clause introduced two individuals of opposite gender and an implicit causality verb; the second clause contained an acoustically ambiguous pronoun. Items were balanced for implicit-causality bias (subject preference vs. object preference) and the position of the male and female names (subject vs. object), as in (3-6).

(3) *she*-biasing context, object verb bias
Luis reproached **Heidi** because ☐ was getting grouchy.

(4) *he*-biasing context, object verb bias
Joyce helped **Steve** because ☐ was working on the same project.

(5) *she*-biasing context, subject verb bias
Abigail annoyed Bruce because ☐ was in a bad mood.

(6) *he*-biasing context, subject verb bias
Tyler deceived Sue because ☐ couldn't handle a conversation about adultery.

Each sentence contained one acoustically ambiguous pronoun (taken from the 5 steps on the /hi/~/ʃi/ continuum that were normed for the Experiment 2 materials). Participants heard all items once. In order to ensure that any measured effect was due to the pragmatic biases of the IC verbs and not the plausibility of the sentence continuations (e.g. *he/she was getting grouchy*), we normed the sentences and confirmed that both *he* and *she* versions were judged to be significantly more plausible than a set of implausible

passages ($F(1,11)=770.95$, $p<0.001$, with 12 subjects who did not participate in Experiments 1, 2, or 3).

Procedure The procedure was the same as in Experiment 2.

Results and Discussion

Only trials in which the comprehension question was answered correctly were included in the results. As predicted, we found that *she*-biasing contexts yielded higher *she* ratings (mean score=2.5) than *he*-biasing contexts (mean score=2.0). There was a main effect of gender context with the data collapsed across steps and across verb types ($F(1,26)=18.738$, $p<0.001$). The results appear in Figure 3.

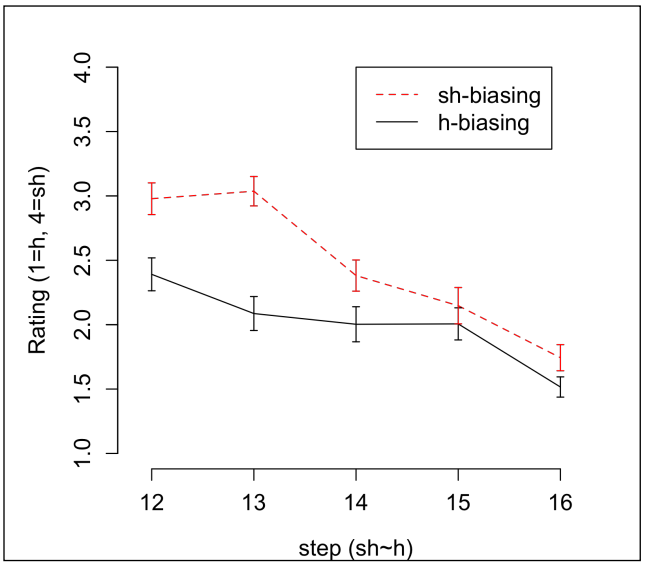


Figure 3: Impact of referential context on perceived phonetic category in Experiment 3

These results support a model of language in which listeners' pragmatic reasoning regarding who is likely to be implicated as the cause of an event influences their phonetic category decisions.

General Discussion

As we described in the introduction, a body of accumulating evidence points to the integration of multiple information sources during language processing. The results presented here suggest that the range of interacting cues spans the conceivable range of linguistic information sources and that phonetic information interacts with high-level causal inferencing about events, event participants, and the likelihood of co-reference across clauses in a discourse.

Our results are in keeping with work showing that the larger discourse context can influence language processing at lower levels. Furthermore, our results suggest that current processing models—be they interactive or integrative—which combine multiple cues from multiple linguistic domains must be refined and better articulated to capture the range of interactivity shown here.

Existing models of phoneme identification currently account for contextual effects such as the semantic congruity effect in one of two ways. Highly interactive models permit direct interaction between acoustic cues, the lexicon, and contextual cues (contextual cues broadly construed, e.g. visual cues, speaker information, acoustic context) such that top-down biases can influence the perceptual system itself (Goldinger 1996; Johnson 1997; Luce & Pisoni 1998; McClelland & Elman 1985). On the other hand, integrative models have been proposed that specify the point of lexical decision as the stage at which listeners combine higher-level information sources with lower-level phonetic cues (Norris et al. 2000). Both types of models could in principle be adapted to account for our results, so long as the range of contextual cues is not restricted to lexical or co-occurrence-based input. For interactive models, an important question is whether pragmatic information is integrated directly into the speech perception process, adding an additional set of non-acoustic cues into the lexical decision process, or whether pragmatic context yields an expectation for a particular continuation, which in turn makes the perceptual process more sensitive to certain acoustic cues. For models that rely on post-perceptual integration of information, however, context serves as a check on an encapsulated perception process; for those models, our results show that pragmatic biases can act as relevant constraints, in addition to other biases that are introduced by lexicality, syntax, or semantics. The difference in effect size between Experiments 2 and 3 may point to differences in the timecourse and strength of such biases.

Just as existing models of phoneme identification could in principle be extended to include higher-level top-down biases, another option for modeling our results would be to adapt existing sentence processing models to capture effects at lower levels of processing. Existing constraint-based sentence processing models have up until now primarily targeted syntactic processes not phoneme decisions (MacDonald 1994; Jurafsky 1996; Spivey & Tanenhaus 1998; McRae, Spivey-Knowlton, & Tanenhaus 1998; Levy 2008, among others). These models—crucially their architectures for integrating multiple cues—could be adapted to fit our data by incorporating discourse-based constraints that interact fully with other processing biases, including those generated at the phonetic level. The work described in this paper attests to the importance of a unified approach that models a range of information sources and their influence on each other during processing.

Existing models have thus not fully addressed the question of precisely which linguistic levels show interactive effects and what mechanism would allow phonetic and pragmatic information to be combined. Our results, which present a new type of interaction, help establish the extent of possible interactivity that must be accounted for, though the results also raise questions regarding the exact nature of these interactive effects.

Recent evidence on the neural bases of lexical effects on phonetic perception points towards the interactive approach (Myers & Blumstein, 2008). The contexts used here provide an opportunity to explore whether different processing systems make use of different strategies for incorporating information from different levels. If multiple systems are in operation, it is possible that the levels in closest proximity interact in a more dynamic fashion. By identifying contexts that induce interactive effects at quite disparate linguistic levels, future work can explore whether the timecourse of such effects are attributable to integrative or interactive mechanisms. Future work must address these questions, and the paradigm we have introduced here provides useful contexts for such work precisely because these contexts permit the manipulation of biases that may be active when listeners are interpreting sounds in rich discourse contexts.

Acknowledgments

This research was supported in part by a Mellon postdoctoral fellowship to Hannah Rohde and by NIH grant T32 NS047987 to Marc Ettlinger. We thank Ann Bradlow and Matt Goldrick for helpful discussion and research assistant Ronen Bay for his help during data collection.

References

Altmann, G. & Steedman, M. (1988). Interaction with context during human sentence processing. *Cognition* 30, 191–238.

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception & Performance,* 6(1), 110-125.

Garvey, C., & Caramazza, A. (1974). Implicit causality in verbs. *Linguistic Inquiry,* 5, 459-464.

Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22, 1166-1183.

Isenberg, D., Walker, E. C. T., & Ryder, J. M. (1980). A top-down effect in the identification of function words. Acoustical Society of America, Los Angeles, CA.

Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science,* 20(2), 137–194.

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing,* 19, 1-36.

Marslen-Wilson, W. & Tyler, L. K. (1980). The temporal structure of spoken language understanding. *Cognition*, 8, l-71.

MacDonald, M.C. (1994). Probabilistic constraints and syntactic ambiguity resolution. *Language and Cognitive Processes,* 9(2), 157–201.

Magnuson, J. S., McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2003). Lexical Effects on Compensation for Coarticulation: The Ghost of Christmas Past. *Cognitive Science*, 27, 285-298.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, 18, 1-86.

McClelland J. L., Mirman D., & Holt L. L. (2006). Are there interactive processes in speech perception? *Trends in Cognitive Science,* 10, 363-369.

McGuire, G. (2007). Phonetic Category Learning. Ph.D. Dissertation, The Ohio State University.

McRae, K., Spivey-Knowlton, M. J., & Tanenhaus, M. K. (1998). Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language,* 38, 283–312.

Miller, J. L. & Dexter, E. R. (1988). Effects of speaking rate and lexical status on phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance*, 14, 369-378.

Miller, J. L., Green, K., & Schermer, T. M. (1984). A distinction between the effects of sentential speaking rate and semantic congruity on word identification. *Perception & Psychophysics*, 36(4), 329-337.

Munson, B., & Coyne, A. C. (in press). The Influence of Apparent Vocal-Tract Size, Contrast Type, and Implied Sources of Variation on the Perception of American English Voiceless Lingual Fricatives. *Journal of the Phonetic Society of Japan.*

Myers, E.B & Blumstein, S.E (2008). The neural bases of the lexical effect: An fMRI investigation. *Cerebral Cortex* 18(2), 278-288.

Ni, W., Crain, S., & Shankweiler, D. (1996). Sidestepping garden paths: Assessing the contributions of syntax, semantics and plausibility in resolving ambiguities. *Language and Cognitive Processes*, 11(3), 283-334.

Norris, D., McQueen, J. M., & Cutler, A. (2000). Merging phonetic and lexical information in phonetic decision-making. *Behavioral and Brain Sciences*, 23, 299-325.

Pitt, M. A. (1995). The locus of the lexical shift in phoneme identification. *Journal of Experimental Psychology: Learning, Memory, & Cognition,* 21, 1037–1052.

Price, P. J. & Levitt, A. G. (1983). The relative roles of syntax and prosody in the perception of the /š/-/č/ distinction. *Language and Speech*, 26(3), 291-304.

Sedivy, J. C. (2002). Invoking discourse-based contrast sets and resolving syntactic ambiguities. *Journal of Memory and Language*, 46:341–370.

Spivey, M.J . & Tanenhaus, M. K. (1998). Syntactic ambiguity resolution in discourse: Modeling the effects of referential content and lexical frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* 24(6),1521–1543.

van Berkum, J. J. A., Brown, C. M., & Hagoort, P. (1999). Early Referential Context Effects in Sentence Processing: Evidence from Event-Related Brain Potentials. *Journal of Memory and Language,* 41, 147–182.

Willits, J. A., Sussman, R. S., & Amato, M. S. (2008). Event knowledge vs. verb knowledge. In Proceedings of the 30th Annual Conference of the Cognitive Science Society, 2227–2232. Austin, TX.

Individual Differences in Attention During Category Learning

Michael D. Lee (mdlee@uci.edu)
Department of Cognitive Sciences, 3151 Social Sciences Plaza A
University of California, Irvine, CA 92697-5100 USA

Ruud Wetzels (wetzels.ruud@gmail.com)
Department of Psychology, University of Amsterdam
Roetersstraat 15, 1018 WB Amsterdam

Abstract

A central idea in many successful models of category learning—including the Generalized Context Model (GCM)—is that people selectively attend to those dimensions of stimuli that are relevant for dividing them into categories. We use the GCM to re-examine some previously analyzed category learning data, but extend the modeling to allow for individual differences. Our modeling suggests a very different psychological interpretation of the data from the standard account. Rather than concluding that people attend to both dimensions, because they are both relevant to the category structure, we conclude that it is possible there are two groups of people, both of whom attend to only one of the dimensions. We discuss the need to allow for individual differences in models of category learning, and argue for hierarchical mixture models as a way of achieving this flexibility in accounting for people's cognition.

Keywords: Selective attention, Category learning, Generalized Context Model, Individual differences, Hierarchical Bayesian modeling

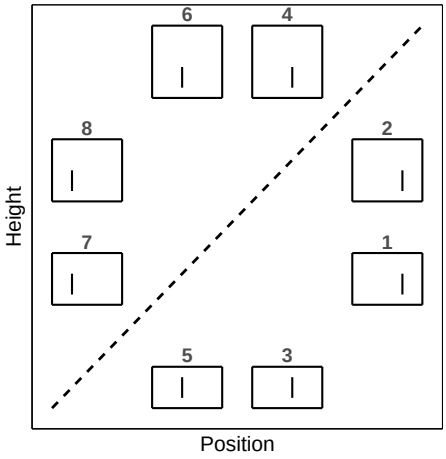


Figure 1: Condensation category structure "B" from Kruschke (1993).

Introduction

Selective attention is one of the most compelling theoretical ideas in the study of human category learning. The basic idea is that, to learn a category structure, people selectively attend those dimensions of the stimuli that are relevant to distinguishing the categories. Nosofsky's (1984) landmark paper showed that, for stimuli represented in terms of underlying continuous dimensions, selective attention could help explain previously puzzling empirical regularities in the ease with which people learn different category structures (Shepard, Hovland, & Jenkins, 1961).

The Generalized Context Model (GCM: Nosofsky, 1984, 1986) incorporates an attention process that has proven enormously helpful in accounting for human category learning behavior. Kruschke (1992) developed a natural extension of the GCM that was able to learn selective attention weightings on a trial-by-trial basis for dimensional stimuli, and Lee and Navarro (2002) showed that the same approach worked equally well for stimuli represented in terms of discrete features rather than continuous dimensions.

In this paper, we raise the possibility that different people might apply selective attention differently when learning the same category structure. We re-analyze human performance on a single task conducted by Kruschke (1993), using the GCM, but allowing for individual differences. We find evidence that one group of people attended primarily to one dimension of the stimuli, while a second group of people attended primarily to the other dimension. This finding runs counter to a standard analysis that does not allow for individual differences, and shows a distribution of attention across both dimensions.

Category Learning Data

The data we use in our re-analysis comes from Kruschke (1993), who studied the ability of ALCOVE to account for human learning across four category structures. Each structure involved the same eight stimuli—consisting of line drawings of boxes with different heights, with an interior line in different positions—but divided them into two groups of four stimuli in different ways. The category structure we use is the so-called "Condensation B" structure, which is shown in Figure 1. The eight stimuli are arranged by their heights and positions, and the four above and to the left of the dividing line belong to Category A. The stimuli are numbered 1–8 in the figure, for ease of reference later when we present modeling results.

Kruschke (1993) collected data from a total of 160 participants, with 40 attempting to learn each category structure. The task for each participant was, over eight consecutive blocks within which each stimulus was presented once in a random order, to learn the correct category assignment for each stimulus, based on corrective feedback provided for every trial. With the aim of analyzing human performance using the GCM—which means trial-by-trial learning is not being modeled—the data can be represented by d_{ik}, the number of times the ith stimulus was categorized as belonging to Category A by the kth participant, out of the $t = 8$ trials on which it was presented. In an analysis that does not consider individual differences, the behavioral data can be further summarized as $d_i = \sum_k d_{ik}$, the total number of times all participants classified the ith stimulus into Category A, out of $t = 40 \times 8$ total presentations.

Generalized Context Model Analysis

In this section, we present a standard version of the GCM, show how it can be formulated as a graphical model to enable fully Bayesian statistical inference[1], and present its application to the current data.

The Standard GCM

The GCM assumes that stimuli can be represented by their values along underlying stimulus dimensions, as points in a multidimensional psychological space. For the current data, there are only two dimensions, so the ith stimulus is represented by the point (p_{i1}, p_{i2}). The first dimension has an attention weight, w with $0 \leq w_d \leq 1$, and the second dimension then has an attention weight $(1-w)$. These weights act to 'stretch' attended dimensions, and 'shrink' unattended ones. Formally, the psychological distance between the ith and jth stimuli is $d_{ij}^2 = w\,(p_{i1} - p_{j1})^2 + (1-w)\,(p_{i2} - p_{j2})^2$.

The GCM assumes classification decisions are based on similarity comparisons with the stored exemplars, with similarity determined as a nonlinearly decreasing function of distance in the psychological space. We follow Nosofsky (1986) and model the similarity between the ith and jth stimuli as $s_{ij} = \exp(-c^2 d_{ij}^2)$, where c is a generalization parameter. The GCM also assumes that categories are represented by individual exemplars. This means that, in determining the overall similarity of a presented stimulus i to Category A, every exemplar in that category is considered, so that the overall similarity is $s_{iA} = \sum_{j \in A} s_{ij}$. Final categorization response decisions are based on the Luce Choice rule, as applied to the overall similarities. We assume an unbiased version of the choice rule, so that the probability that the ith stimulus will be classified as belonging to Category A, rather than Category B, is modeled as $r_i = s_{iA} / (s_{iA} + s_{iB})$. The observed decision data themselves are then simply modeled as $d_i \sim \text{Binomial}\,(r_i, t)$, meaning that each of the t presentations of the ith stimulus has a probability r_i of being categorized as belonging to Category A.

[1]Note that this does *not* mean we are proposing a "Bayesian" or "rational" version of the GCM (cf. Griffiths, Kemp, & Tenenbaum, 2008). We are simply using Bayesian statistics, rather than traditional model-fitting methods and frequentist statistical approaches, to make inferences about GCM parameters from data. That is, we are using Bayesian inference as statisticians do, and as psychologists should do, to relate models to data.

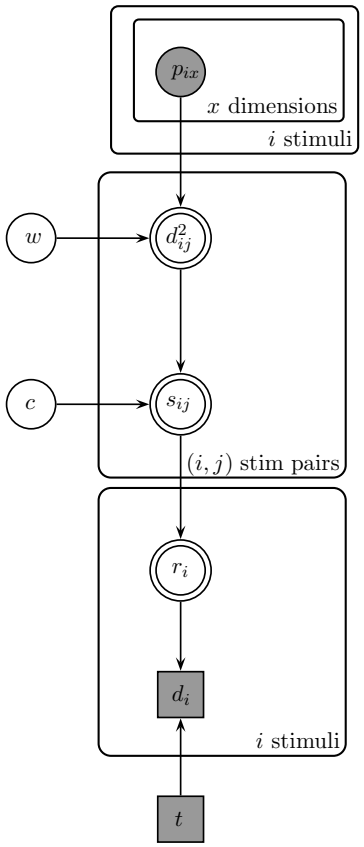


Figure 2: Graphical model implementation of the GCM.

Graphical Modeling Implementation

Our analyses are implemented using the formalism provided by graphical models. A graphical model is a graph with nodes that represents the probabilistic process by which unobserved parameters generate observed data. Details and tutorials aimed at cognitive scientists are provided by Lee (2008) and Shiffrin, Lee, Kim, and Wagenmakers (2008). The practical advantage of graphical models is that sophisticated and relatively general-purpose Markov Chain Monte Carlo (MCMC) algorithms exist that can sample from the full joint posterior distribution of the parameters conditional on the observed data. Our analyses rely on WinBUGS (Spiegelhalter, Thomas, & Best, 2004), which is easy-to-learn software for implementing and analyzing graphical models (see Lee & Wagenmakers, 2010).

A graphical model implementation of the GCM is shown in Figure 2. The known stimulus locations p_{ix}, to-

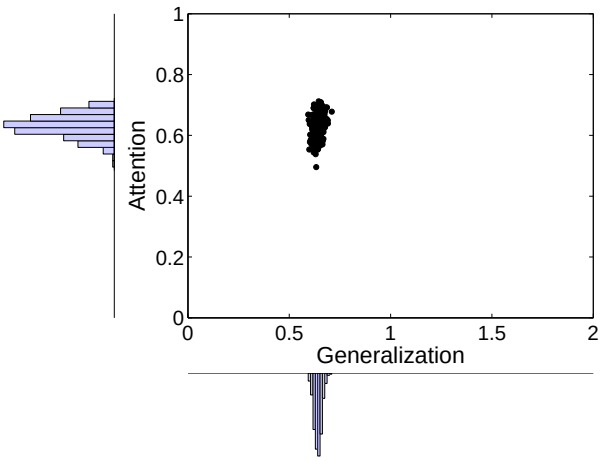


Figure 3: Joint and marginal posterior distributions over attention w and generalization c parameters of the GCM, when applied to the condensation data.

gether with the attention parameter w generate the pairwise distances d_{ij}^2. These distances, together with the generalization parameter c generate the pairwise similarities. These similarities, in turn, lead to response probabilities r_i which generate the observed data d_i.

Results

Our results are based on 3 chains of 5,000 samples each, with a burn-in of 1,000 samples, whose convergence was checked using the standard $\hat{R}$ statistic (Brooks & Gelman, 1997).

The key result is shown in Figure 3, which plots the joint posterior distribution of the generalization and attention parameters (as a scatterplot), as well as their marginal distributions (as histograms). The marginal posterior for the attention parameter w—which gives the weight for the position dimension—lies between about 0.55 and 0.7. This result can be interpreted as showing that people give significant attention to both dimensions, although they are probably focusing a little more on the line position than the rectangle height. In condensation tasks, both stimulus dimensions are relevant to determining how stimuli belong to categories, and so the shared attention result makes sense. In other words, the standard application of the GCM produces a psychologically reasonable inference about selective attention, and it is tempting to view this analysis as the end of the story.

Individual Differences Analysis

The standard analysis assumes, however, that all people used exactly the same parameterization of the GCM to guide their category learning. But an examination of the individual learning curves in the current data suggests a large degree of variation between subjects, and raises the possibility that there are psychologically meaningful individual differences.

Types of Individual Differences

Figure 4 gives a schematic picture of four different assumptions about individual differences. Each panel shows a data space, containing the possible outcomes of an experiment. In the No Differences panel, there is a single true point, represented by the circular marker, corresponding to one parameterization of a cognitive process. The gray circles show the variety of behavioral data that might actually be produced in an experiment. The assumption of no individual differences means the goal of inference would be to find the circular marker from the gray points, and corresponds to the standard analysis of the GCM we have presented.

In the Continuous Differences panel there are many true points, again shown by circular markers. Each of these points could correspond to an individual subject's data from an experiment. The individuals are not identical (i.e., there is no longer a single point), but nor are they unrelated (i.e., their points are not spread across the entire data space). This sort of individual differences can be accommodated by hierarchical or multi-level models, in which there is a single hierarchical group distribution over the parameters of the individuals (e.g., Rouder & Lu, 2005).

In the Discrete Differences panel there are two true points, shown by a circular and a square marker. Each of these points could correspond to the data from different individuals, or from different subgroups, each with multiple individuals, in an experiment. The two points correspond to fundamentally different parameterizations of a cognitive process, or even to fundamentally different cognitive processes, and so the overall data is a mixture of two different cognitive processes. Mixture models are typically used to accommodate this sort of individual differences (e.g., Lee & Webb, 2005).

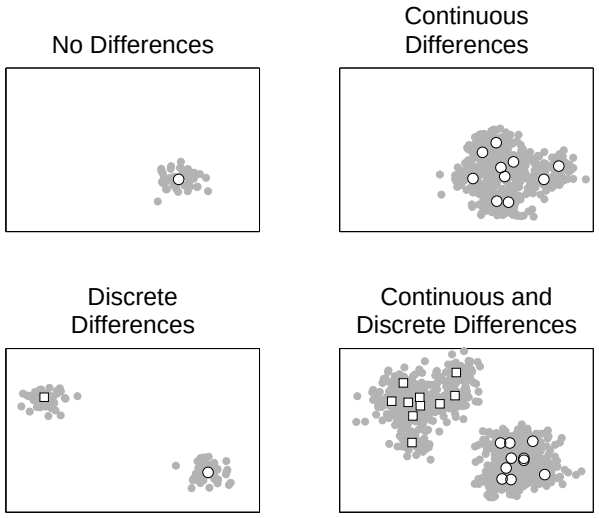


Figure 4: Four different assumptions about individual differences.

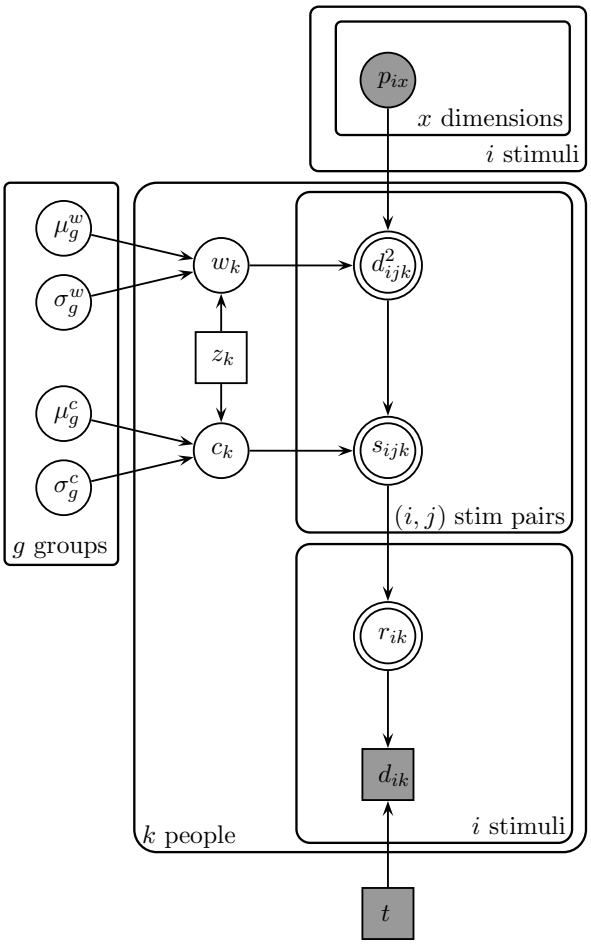


Figure 5: Graphical model for the GCM with individual differences.

The obvious strategy for a more complete account of individual differences is to combine both Continuous and Discrete differences, as in the bottom-right panel on Figure 4. Here, there are two types of true points—indicated by circular and square makers—and constrained individual variation within each type. A combination of both hierarchical and mixture modeling naturally deals with these patterns of differences. The mixture component identifies the fundamentally different cognitive processes, and the hierarchical component captures the variation within each process. We are not aware of cognitive modeling that has adopted this approach, but it seems the most general and natural way to extend the GCM analysis.

Graphical Model Implementation

Figure 5 shows the graphical model that extends the GCM to allow for continuous and discrete individual differences. There is now a plate for the participants, so that the kth participant has attention w_k and generalization c_k parameters. These are drawn hierarchically from one of a set of Gaussian distributions depending on their group membership z_k. Formally, this means $w_k \sim \text{Gaussian}(\mu^w_{z_k}, \sigma^w_{z_k})$ and $c_k \sim \text{Gaussian}(\mu^c_{z_k}, \sigma^c_{z_k})$.

Statistically, this is a hierarchical (or "random-effect") mixture model. Psychologically, people belong to different qualitative groups, given by z_k, and their attention and generalization parameters are sampled from a continuous Gaussian distribution corresponding to their group.

We put standard vague priors on the group means and standard deviations, and on the latent assignment indicator variables. We then applied this extended GCM model to the current condensation data, assuming there were two groups of participants.

Results

Once again, our results are based on 3 chains of 5,000 samples each, with a burn-in of 1,000 samples, whose convergence was checked. Our key findings are laid out in Figure 6. The top-most bar graph shows the inferred allocation of the 40 participants into the two groups, as measured by the posterior expectation of the z_k variable. There are unambiguous assignments for 36 participants, with 24 belonging to Group 1 and 12 belonging to Group 2. This lack of uncertainty in mixture model latent assignment is usually an indication that there are multiple groups.

The attention and generalization properties of the two groups, in the form of the joint and marginal posterior distributions of μ^w_g and μ^c_g, are shown in the next two panels. Group 1 on the left has an attention weight above 0.8, while Group 2 on the right has an attention weight close to 0. The natural interpretation is that the first group of participants is primary attending to the position dimension, while the second group is almost exclusively attending to the height dimension.

Below the posterior distribution for the groups, a posterior predictive check of fit to the behavioral data is shown. For each of the 8 stimuli the posterior predictive distribution over the number of times it is classified as belonging to Category A is shown by the squares, with the area of each square being proportional to posterior predictive mass. The single thick line shows the average observed categorization behavior for those participants assigned to the group. The many thin lines show the individual participant behavior for the group. It is clear that Group 1 and Group 2 have participants showing qualitatively different patterns of categorizing the stimuli, and these differences are captured by the posterior predictive distributions.

The bottom-most panels in Figure 6 interpret the different category learning of the groups. The original stimulus space and category structure is shown, with bars showing the average number of times each stimulus was placed in Category A and Category B by members of the group. To understand Group 1, note that stimuli 4 and 5 are the ones least clearly categorized correctly. This is consistent with a focus on the position dimension, which would assign these two stimuli incorrectly. Similarly, for Group 2, stimuli 2 and 7 are categorized very poorly.

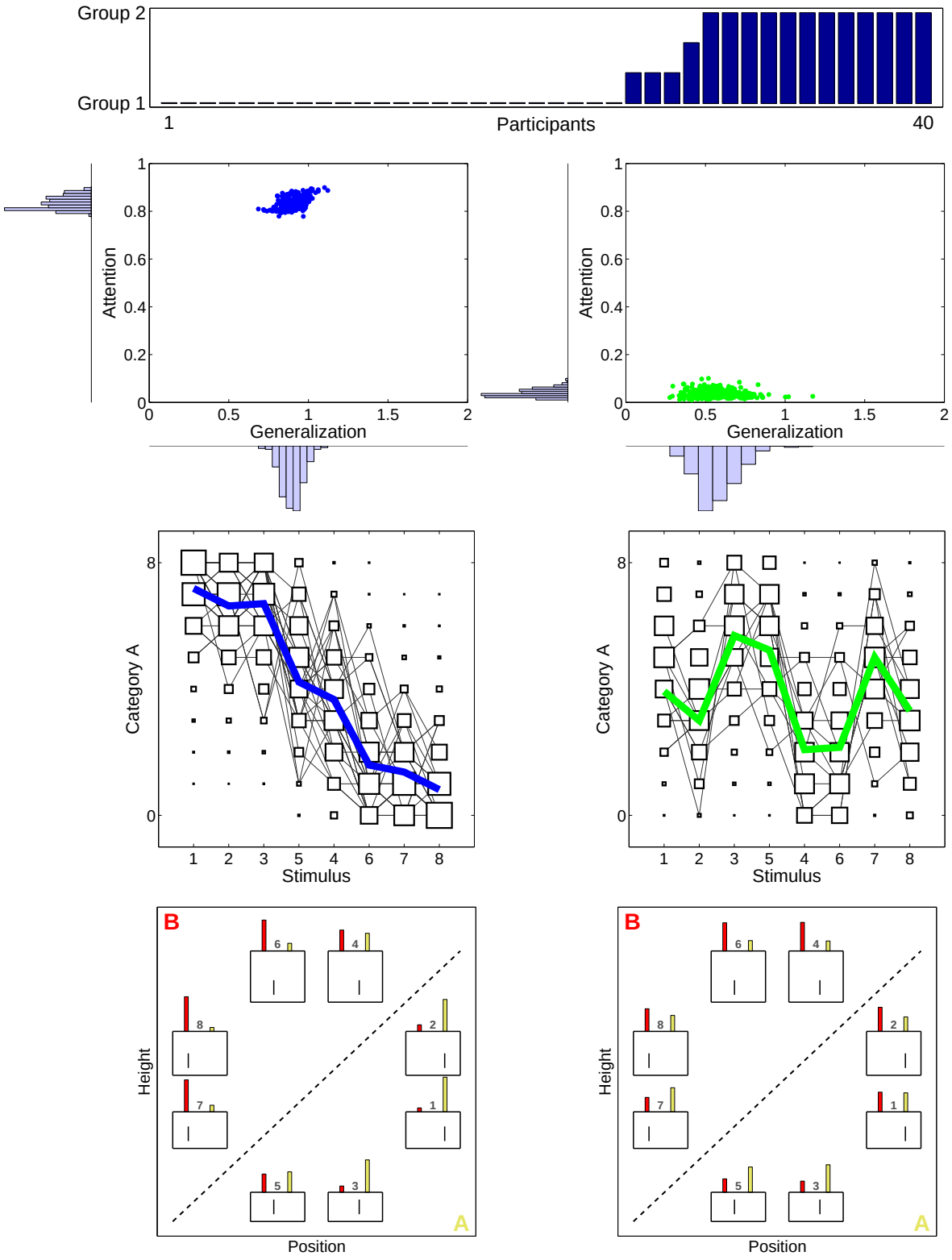


Figure 6: Results from GCM analysis assuming two groups of participants, showing the allocation of participants to groups, posterior and posterior predictive distributions for the groups, and the interpretation of the different groups in terms of the stimuli and category structure itself. See text for details.

This is consistent with a focus on the height dimension.

Finally, we compared a one-group to a two-group model, calculating the Bayes Factor using the Savage-Dickey method described by Wetzels, Grasman, and Wagenmakers (2010). This came out about 2.3 in favor the two-group model, meaning that the data are more than twice as likely to have come from two groups of participants than a single group. While this is far from conclusive evidence, it does suggest that the possibility there are two different groups of participants deserves serious consideration.

Discussion

Our extended analysis of Kruschke's (1993) condensation data, using a GCM with the ability to detect continuous and discrete individual differences, tells an interesting story. It suggests that there are two groups of participants, each of whom focus most of their attention on just one stimulus dimension while learning the category structure. The standard result of attention being distributed roughly evenly across both dimensions seems to be an artefact of failing to consider individual differences in modeling.

We realize that applying the GCM to the condensation data is non-standard, because the GCM is usually applied to category learning experiments with a training and a testing phase, rather than a single category learning sequence. Ideally, our modeling would be applied to transfer data collected after categories were learned to criterion, and it is possible the dynamics of learning provide a partial explanation for the individual differences we observe, although we do not think they can provide a full explanation. We also realize that there are many possible variations of the GCM that could be tried.

Accordingly, we certainly do not claim our single re-analysis automatically undermines the existing large and coherent body of work examining selective attention mechanisms in category learning. Systematic investigation of category learning across many tasks, looking for the presence of discrete and continuous individual differences, is needed to gauge the generality of our current results. We think this would be a worthwhile exercise, given the theoretical influence of selective attention mechanisms in the category learning literature.

We also think our analyses underscore a more general point, which is that it is important to consider and model individual differences in all of cognition. Finally, we think the ease with which very general assumptions about individual differences could be implemented to extend the standard GCM analysis shows the advantage of using Bayesian statistics to relate cognitive models to data.

References

Brooks, S. P., & Gelman, A. (1997). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, *7*, 434–455.

Griffiths, T. L., Kemp, C., & Tenenbaum, J. B. (2008). Bayesian models of cognition. In R. Sun (Ed.), *Cambridge Handbook of Computational Cognitive Modeling* (pp. 59–100). Cambridge, MA: Cambridge University Press.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*(1), 22–44.

Kruschke, J. K. (1993). Human category learning: Implications for backpropagation models. *Connection Science*, *5*, 3–36.

Lee, M. D. (2008). Three case studies in the Bayesian analysis of cognitive models. *Psychonomic Bulletin & Review*, *15*(1), 1–15.

Lee, M. D., & Navarro, D. J. (2002). Extending the ALCOVE model of category learning to featural stimulus domains. *Psychonomic Bulletin & Review*, *9*(1), 43–58.

Lee, M. D., & Wagenmakers, E.-J. (2010). *A Course in Bayesian Graphical Modeling for Cognitive Science.* Course notes, University of California Irvine. [http://www.socsci.uci.edu/∼mdlee/bgm].

Lee, M. D., & Webb, M. R. (2005). Modeling individual differences in cognition. *Psychonomic Bulletin & Review*, *12*(4), 605–621.

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(1), 104–114.

Nosofsky, R. M. (1986). Attention, similarity and the idenitification-categorization relationship. *Journal of Experimental psychology: General*, *115*, 39-57.

Rouder, J. N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review*, *12*, 573-604.

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs*, *13*.

Shiffrin, R. M., Lee, M. D., Kim, W.-J., & Wagenmakers, E.-J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive Science*, *32*(8), 1248–1284.

Spiegelhalter, D. J., Thomas, A., & Best, N. G. (2004). *WinBUGS Version 1.4 User Manual*. Cambridge, UK: Medical Research Council Biostatistics Unit.

Wetzels, R., Grasman, R. P. P. P., & Wagenmakers, E. (2010). An encompassing prior generalization of the Savage-Dickey density ratio test. *Computational Statistics and Data Analysis* , *54*, 2094–2102.

Categorisation, Deference and Cognitive Style

Nick Braisby (Nick.Braisby@winchester.ac.uk)
Department of Psychology, University of Winchester, Winchester, SO22 4NR, UK

Sharon Hanlon (S.Hanlon@winchester.ac.uk)
Department of Psychology, University of Winchester, Winchester, SO22 4NR, UK

Abstract

Despite the importance of psychological essentialism as an account of categorisation, it is unclear what import findings of individual difference have. The present study is designed to investigate individual differences in relation to deference, a key indicator of essentialist thought. This replicates previous findings of individual differences in deference, and demonstrates a strong association between deference and field dependence (Witkin et al., 1962). In spite of the status of field dependence as a cognitive style, this study finds it has highly localised influences in relation only to categorisations and categorisation-related actions that are partly social in nature. Implications for essentialism are discussed.

Keywords: categorization, deference, essentialism, cognitive style

Introduction

Categorisation research has largely adopted a core methodological assumption of cognitive science that adults are sufficiently alike that it makes sense to talk of a 'typical' cognizer, and to pursue generalisations that disregard individual differences (von Eckardt, 1993). Yet from the earliest days of categorisation research, important individual differences have been found. Moreover, in recent years, studies have uncovered individual differences with regard to psychological essentialism. However, with the exception of research conducted in the middle of the last century, there have been only few studies of the basis for such individual differences, and whether their scope is restricted to or extends beyond categorisation itself. This paper reports a study designed to evaluate individual differences in relation to deference and essentialism in the categorisation of natural kinds.

Psychological essentialism represents an important and popular theoretical account of categorisation. According to psychological essentialism people believe, and act as if, category membership is determined by the possession of an essence (Medin & Ortony, 1989). People are deemed to believe that objects have essences, that essences are causally responsible for other properties such as appearance, and that essences are responsible for category or kind membership.

Findings that have been argued to support psychological essentialism include those of Keil (1986, 1989) and Rips (1989). Rips described a transformation in which a bird-like animal came to appear more like an insect as a consequence of exposure to radiation. Participants judged the animal to be a bird still, even though they felt it was more similar to an insect. Keil reported the results of similar studies with children. For example, transformations included making a raccoon look and behave like a skunk through being painted and implanted with an odour sac. While younger children tended to categorise this as a skunk, older children considered it still to be a raccoon. Moreover, even younger children are disposed to categorise objects according to presumed essences (Gelman, 2000). Gelman & Wellman (1991) showed that 4 and 5 year old children appear to believe that an apple seed will grow into an apple tree, regardless of the environment in which this happens. Apparently children believe something inside the seed, and not contingent features of the environment, is causally responsible for the properties it later acquires.

Though largely developed to explain natural kind categorisation, the apparent explanatory success of psychological essentialism has led other researchers to seek to apply it in other domains, most notably to artefacts (Bloom, 1996; 1998; though see Malt & Sloman, 2007) and social categories (e.g., Haslam, Rothschild & Ernst, 2000, 2002; Haslam & Whelan, 2008; Rothbart & Taylor, 1992).

Of course, there have been criticisms of essentialism. Malt (1994) showed that categorisation of instances of water is not fully explained by the proportion of H_2O people believe the instances contain. Braisby, Franks & Hampton (1996) showed that categorisation is at odds with predictions suggested by Putnam and Kripke's articulation of essentialism. There has also been discussion of whether essentialism is required to explain the empirical evidence cited in its favour (Ahn et al., 2001; Strevens, 2000).

Deference and Individual Differences

Braisby (2001, 2004) also examined the further implication of essentialism that people should defer in their categorisations to appropriate experts, an implication developed by Putnam (1975) in a thesis he labelled the Division of Linguistic Labour (see also Kripke, 1980). Since, according to essentialism, categorisation is determined by micro-structural (e.g., genetic) properties, then scientists expert in the appropriate domain are likely to have more category-relevant information than lay-people. If lay people are psychological essentialists then they should rationally defer to people with more knowledge of the relevant properties, e.g., expert scientists. However, in a series of studies examining deference for biological and

chemical categories, Braisby found that participants deferred in only approximately one-third of cases for biological categories, and only slightly more than this for chemical categories. Braisby's conclusion was that the data concerning deference did not support essentialism but could be explained by a perspectival or similarity-based account of categorisation.

However, Braisby also found significant individual differences in the propensity to defer. Whereas many participants consistently switched their categorisation judgments to conform to those of experts, still others consistently maintained their categorisation judgments regardless of expert opinion. Therefore, an alternative explanation of these data is that some participants were psychological essentialists, while the judgments of others were similarity-based. Hampton, Estes & Simmons (2007) also found evidence of individual differences in essentialism. In an examination of Rips's (1989) transformation study, they found that some participants steadfastly maintained their categorisation both before and after the transformation. Only a minority of participants fitted the pattern reported by Rips.

An important question to resolve is whether such individual differences reflect deeper differences in the way that people cognize, or whether people flexibly deploy information and beliefs in making categorisation judgment depending on the task and context. Surprisingly, there is relatively little evidence to bear on this question.

There have nevertheless been demonstrations of individual differences relating to categorisation. Lewellen, Goldinger, Pisoni & Greene (1993) found that participants who scored higher on measures of lexical familiarity were more successful in rejecting foils in a semantic categorisation task. There have been a number of individual differences reported in relation to category learning. For example, McKinley and Nosofsky (1995) found individual differences both in the time course of learning, and also in the final categories learned. DeCaro, Thomas & Beilock (2008) also found that working memory influences category learning. Rule-based categories were learned more quickly by participants with a greater working memory capacity, and what they called information-integration categories were learned more quickly by participants with a smaller working memory capacity. Kalénine & Bonthoux (2006) showed that individual differences in 3-4 year olds' preferences for thematic or taxonomic matches affected their choice of superordinate categories – children showing greatest sensitivity to taxonomic relations showing superior performance in categorising living things.

While the above studies show how individual differences in cognitive processes impact categorisation, there is also a body of work which suggests that individual differences in categorisation arise from more global differences in cognitive style.

Lee, Kagan, & Rabson (1963) found that participants who adopted an analytic strategy when pairing visual stimuli (e.g., on the basis of a shared feature) learned analytic concepts (e.g., objects with a missing leg) more quickly than relational concepts (e.g., objects related to school). Participants who did not adopt this strategy when pairing visual stimuli, however, learned analytic concepts more slowly than relational ones. Interestingly, Lee et al. related their use of the term analytic to 'field dependence' – the phrase earlier coined by Witkin, Dyk, Faterson, Goodenough & Karp (1962). Norenzayan, Smith, Kim & Nisbett (2002) found that a similar distinction – between analytic and holistic processing affected category learning and similarity judgments.

Cognitive Style

According to Witkin, Oltman, Raskin, & Karp (1971), cognitive styles are "the characteristic, self-consistent modes of functioning which individuals show in their perceptual and intellectual activities" (p. 3). One such style, field dependence, is a construct intended to capture an individual's characteristic mode of perception (Witkin, 1975). It was initially tested using the body-adjustment test and the rod-and-frame test to assess perception of the true vertical, in a visual or postural field that may present misleading information. Typically, some people – field-independent – will accurately judge the true vertical regardless of the contents of the visual field, while others – field-dependent – would fail to do so, presumably being misled by the visual field. Witkin et al. (1962) developed other measures of field-dependence. The embedded figures test and group embedded figures test have since become commonly used. The group embedded figures test (see Figure 1) involves asking participants to find a simple geometric figure (e.g., the triangle labeled X at the top) within a more complex visual object (e.g., the geometric shape at the bottom).

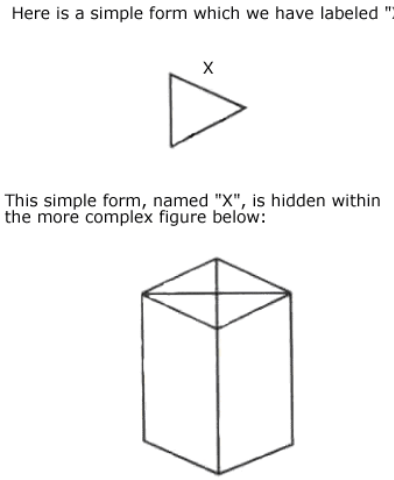


Figure 1.Sample image from the group embedded figures test

Differences in field dependence have been linked to other socio-psychological differences including, for example, identity, awareness of self and locus of control. Developmental research has suggested that children tend to

become more differentiated in their field dependence as they develop. Witkin, Oltman, Raskin & Karp (1971) suggest that field dependence in children is initially relatively fluid, but begins to crystallise around the age of ten and then appears stable during adulthood. Not surprisingly, there has been much interest in the distinction within research on education (Sternberg & Grigoernko, 1997).

Reflecting its possible status as a more global reflection of an individual's cognitive style, field-dependence-independence has sometimes been described as a distinction between global and articulated processing. However, the precise nature of the distinction remains unclear. There have been suggestions that field dependence is closely tied to underlying spatial ability (Sternberg & Grigoernko, 1997). There have also been arguments that field dependence reflects sensitivity to social information. In a complex design, Mausner & Graham (1970) asked pairs of participants to rate the speed of a flickering light, and then do so again when given information about the performance of the other member of their pair. Such reinforcement had no effect on the performance of field-independent participants. However, reinforcement led to a dramatic shift in the performance of field-dependent participants. Most strikingly, field-dependent participants who were told their estimates of speed were incorrect, but that their partner's estimates were correct, shifted uniformly and almost completely toward their partner's judgments.

This finding suggests one explanation for individual differences in relation to deference (and essentialism). Field dependent participants, sensitive to social information, including the views of others, may be more likely to shift their judgments towards those of experts. Field independent participants may be more reluctant to do so. If this is the case, then this relative difference in propensity to defer may give rise to considerable variability in the extent to which people's judgments conform with essentialism.

Experiment

The experiment was therefore designed with a number of aims in mind. First, it was important to replicate the findings of Braisby (2001) concerning individual differences in deference, and so determine whether such differences are robust. Second, and in order to better understand such differences, it was decided to take measures of participants' field dependence. Third, although the focus of the study is categorisation, in order to determine the scope of individual differences, a number of other judgments were also sought from participants. As in Braisby (2001), the experiment examined the extent to which lay-people defer in their categorisation of biological natural kinds to experts, as predicted by essentialism.

Method

Design

The experiment adopted a mixed design with the factor of Polarity (Yes, No) of expert judgment being within-subject, and Field Dependence (Field dependent, Field independent) being a between-subject factor.

Participants

40 participants volunteered to participate, 20 of whom were undergraduate students from the University of Winchester. 20 participants were drawn from the immediate residential neighbourhood, all of whom were in employment.

Materials

Following Braisby (2001), categories were four natural (living) kinds : apple, potato, salmon, chicken. These were chosen also to be food-stuffs so that they, and the prospect of their genetic modification, would be relatively familiar to the participants. Within these constraints, the kinds were chosen to be as typical as possible of their immediate super-ordinate categories (i.e., fruit, vegetable, fish, bird).

For each category, two scenarios were developed, one of which contained a positive categorisation judgment from scientific experts (biologists) and one of which contained a negative judgment. All scenarios conformed to the following pattern: "You have just bought a(n) X from a reputable retailer. On examining its packaging closely you find that it has been genetically modified. You also discover that according to most biologists the object you have bought [is/is not], in fact, an X. The object looks, feels, smells and tastes just like a X."

The group embedded figures test is a timed test and comprises a test booklet containing instructions, a practice section, and two test sections. In these two sections, 18 complex geometric shapes are provided and participants must identify in each a given simple shape.

Procedure

Participants were tested individually. Half of the participants were presented with the GEFT first and the categorization scenarios second; the remaining participants received the categorization scenarios and then the GEFT.

When presented with the GEFT, participants were first asked to read through instructions and complete the practice section. They then completed sections 2 and 3 of the GEFT, being given a limit of 5 minutes for each section.

The 8 categorisation scenarios were untimed and presented in one of two orders. Half of the participants were presented with the scenarios in random order, and the remaining participants were presented with the scenarios in the reverse of this order. On reading each scenario, participants were asked to answer six questions, including a categorization question, as follows.

Categorisation: Is the object that you have bought a(n) X?
Superordinate categorisation: Is the object that you have bought a(n) [Superordinate]?
Eat: Would you eat the object you have bought (either as is or prepared)?

Serve: Would you serve the object you have bought at a dinner party for your friends (either as it is or prepared)?
Buy: Would you continue to buy this kind of object?
Eat if served: Would you eat the same kind of object as the one you have bought (either as it is or prepared) if a friend served it to you at a dinner party?

As the categorisation question was the most central to the analysis, and to minimise any interference from other questions, this question was always presented first. Participants were required to answer Yes or No to each question. Lastly, participants were asked to rate how difficult they found making their judgments on a scale of 1-7, 1 being very easy and 7 being very difficult.

Results

Participants responses to the six Questions were recoded to express agreement with the expert judgments, and aggregated across the four categories. A median split was employed to divide participants into Field Dependent and Field Independent groups. The overall mean difficulty rating was 3.64, and this did not differ by Field Dependence.

A two-way ANOVA was conducted for each Question with Polarity (Yes, No) as within-, and Field Dependence (Dependent, Independent) as between-subject factors.

Categorisation

Agreement with biologists' judgments was influenced by Polarity ($F(1,38) = 5.87$, $\acute{\eta}_p^2 = 0.13$, $p < 0.05$), with participant's agreeing more when biologists' judgments were reported as affirmative (mean = 3.68) than when they were reported as negative (mean = 2.80). There was a significant effect of Field Dependence ($F(1,38) = 22.81$, $\acute{\eta}_p^2 = 0.34$, $p < 0.0005$), with Field dependents showing much higher levels of agreement (mean = 3.78) than Field independents (mean = 2.71). Polarity and Field Dependence did not interact.

Superordinate categorisation

Agreement with biologists' judgments was strongly influenced by Polarity ($F(1,38) = 47.81$, $\acute{\eta}_p^2 = 0.56$, $p < 0.0005$), with participant's agreeing the superordinate categorisation when biologists' judgments were reported as affirmative (mean = 3.89) but largely disagreeing when those judgments were negative (mean = 1.73). There was no effect of Field Dependence nor did Polarity and Field Dependence interact.

Eat

There were no effects of Polarity or Field Dependence, nor an interaction between them.

Serve

There were no effects of Polarity nor an interaction with Field Dependence, but there was a main effect of Field Dependence ($F(1,38) = 6.95$, $\acute{\eta}_p^2 = 0.16$, $p < 0.05$) with Field dependents showing greater agreement with biologists' judgments (mean = 2.94) than Field independents (mean = 2.32).

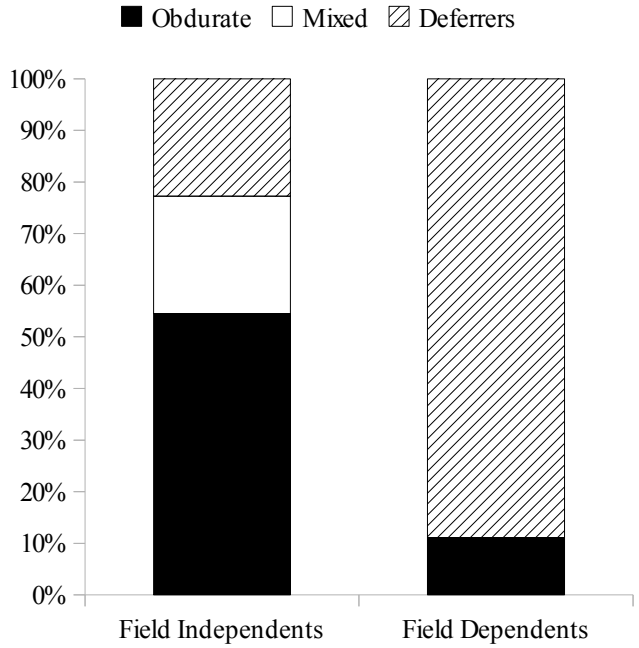


Figure 2.Proportions of participants by Deferring Style and Field Dependence

Buy

There were no effects of Polarity or Field Dependence, nor an interaction between them.

Eat if served

There was a significant effect of Polarity ($F(1,38) = 13.86$, $\acute{\eta}_p^2 = 0.27$, $p < 0.005$), but no interaction with Field Dependence, nor an interaction between them. Regardless of Field Dependence, more participants agreed with the biologists' judgment when this was in the affirmative (mean = 3.13) than in the negative (1.29).

Individual Differences

Lastly, each participant was classified according to their responses to the Categorisation question. Participants who consistently deferred to biologists' judgments in all eight scenarios were classified as Switchers; those who consistently did not switch their categorisations for any category were classified as obdurate; remaining participants were classified as mixed. This factor of Deferring Style was entered with Field Dependence in a log-linear analysis. This revealed a significant interaction between Deferring Style and Field Dependence (Chi-square (2) = 20.52, $p < 0.0005$) as shown in Figure 2.

Discussion

The first key aim of this experiment was to replicate the findings of Braisby (2001) in order to examine whether individual differences in deference are robust. Overall, 53%

of participants consistently deferred to expert judgment, 35% were consistently obdurate, and just 13% showed a mixed pattern (of deferring with some categories and being obdurate with others). Experiment 2 of Braisby (2001) obtained similar proportions: 62%, 31% and 7%, respectively. Thus, these data strongly support the view that the evidence for deference with regard to biological natural kinds is both mixed, and susceptible to substantial individual difference.

The second aim was to investigate the relationship between deference and field dependence. The data confirm that there is such a relationship and it is a strong one, with 34% of the variance in responses to the categorisation question being explained by this dichotomous factor. In this study, substantially more field dependent participants defer to expert judgment (89%) than field independents (23%). Considerably more field independent participants are obdurate when categorising in the light of expert judgment (55%) than field dependents (11%). These striking contrasts not only suggest the effect of field dependence is strong, they suggest reasons for individual differences in essentialism. Field dependents, willing to seek external frames of reference for making their categorisation judgments, appear more susceptible to externally provided information about the presence, role or value of essential properties. Field independents may by contrast tend to rely more on internally generated judgments of category membership which, given the hidden and/or unknown nature of essences, are likely to be based on a more superficial similarity judgment.

Lastly, by including other questions concerning the transformed natural kinds, it is possible to gauge the scope of these individual differences. Were field dependence to impact all measures, for example, it could be argued that it is not intimately related to categorisation, and perhaps that the influence of field dependence masks more subtle and interesting categorisation effects. However, there was no effect of field dependence on three of the five other questions asked. Indeed, only the questions concerning serving food to others, and eating it if it were served by others, showed an influence of field dependence. It is noteworthy that these two questions also involve a social dimension, while the other three questions arguably do not. Far from field dependence showing an over-powering or global impact on these results, it appears as though this factor bears only on those aspects of categorisation and categorisation-related actions that are social in nature. Indeed, when one recalls that Putnam (1975) called his Division of Linguistic Labour a socio-linguistic hypothesis, it seems hardly surprising that the quite particular feature of deference should be influenced by field dependence.

Another interpretation is that field dependence influences how participants understood the scenarios. Elements that are vague, such as the quantifier 'most', or open to different interpretation, such as the reputability of the supplier, may be particularly susceptible to different interpretations that perhaps align with field dependence. Likewise field dependence may alter whether people judge that genetically modified exemplars continue to be members of their original categories. These are intriguing possibilities, and the current data do not rule them out. However, there are reasons to doubt these could be the whole story. First, though the literature on field dependence is considerable, the authors are not aware of evidence for an influence on language understanding. Second, the data actually suggest these possibilities are unlikely. It would be hard, for instance, to reconcile the claim that field dependent and independent people derive different understandings of the scenarios, with the evidence that, when questioned, only certain highly specific questions show such an influence. In fact, it is only those questions which have an explicitly social element that reveal an effect of field dependence. This pattern is more consistent with field dependence having a highly specific influence, related to the informational demands of the task, rather than a global influence relating to people's understanding.

Some notes of caution are in order however. This initial study, while promising, remains exploratory, and much more needs to be done to confirm the impact of cognitive style on categorisation in general. Though these data are suggestive as to the meaning of individual differences in essentialism, it is unclear whether the same relationship would be found in different domains. Of particular interest would be social domains such as sexual orientation (cf. Haslam, Rothschild & Ernst, 2000; Braisby & Hodges, 2009) where claims for essentialism are already contested.

However, these data are illuminating in that they appear to confirm of an important social dimension to psychological essentialism, and one which can lead people to different categorisations. What is less clear is whether these data might shed light on field dependence itself. While such an aim is beyond the scope of this paper, it seems clear that field dependence is more than a spatial ability. It appears to involve a sensitivity to social information and as such implies less of a gap between cognitive science and the social world than one might at first imagine.

Acknowledgements

We are very grateful to Ian Hodges for comments on this paper.

References

Ahn, W. K., Kalish, C., Gelman, S. A., Medin, D. L., Luhmann, C., Atran, S., et al. (2001). Why essences are essential in the psychology of concepts. *Cognition, 82(1)*, 59-69.

Bloom, P. (1996). Intention, history, and artifact concepts. *Cognition*, *60*, 1–29.

Bloom, P. (1998). Theories of artifact categorization. *Cognition*, *66*, 87–93.

Braisby, N. R. (2001). Deference in categorisation: Evidence for essentialism. In J. D. Moore & K. Stenning (Eds.), *Proceedings of the Twenty-Third Annual*

Conference of the Cognitive Science Society. Mahwah, NJ.: Lawrence Erlbaum.

Braisby, N. R. (2004). Deference and Essentialism in the Categorization of Chemical Kinds. In, Alterman R., & Kirsch, D. (Eds.), *Proceedings of the Twenty-Fifth Annual Conference of the Cognitive Science Society*. Lawrence Erlbaum Associates, Inc.: Mahwah, NJ.

Braisby, N., Franks, B., & Hampton, J. (1996). Essentialism, word use, and concepts. *Cognition, 59*, 247-274.

Braisby, N. & Hodges, I. (2009). Categorisation of sexual orientation: A test of essentialism. In N. A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 2956-2961). Austin, TX: Cognitive Science Society.

DeCaro M. S., Thomas R. D., & Beilock S. L. (2008). Individual differences in category learning: Sometimes less working memory capacity is better than more. *Cognition, 107*, 284–294.

Gelman, S. A. (2004). Psychological essentialism in children. *Trends in cognitive sciences, 8(9)*, 404-9.

Gelman, S. A., & Wellman, H. M. (1991). Insides and essences: Early understandings of the nonobvious. *Cognition, 38*, 213-244.

Hampton, J. A., Estes, Z., & Simmons, S. (2007). Metamorphosis: Essence, appearance and behavior in the categorization of natural kinds. *Memory & Cognition, 35*, 1785–1800.

Haslam, N., Rothschild, L., & Ernst, D. (2000). Essentialist beliefs about social categories. *British Journal of Social Psychology, 39*, 113–127.

Haslam, N., Rothschild, L., & Ernst, D. (2002). Are essentialist beliefs associated with prejudice? *British Journal of Social Psychology, 41*, 87–100.

Haslam, N. & Whelan, J. (2008). Human natures: Psychological essentialism in thinking about differences between people. *Social and Personality Psychology Compass, 2/3*, 1297-1312.

Kalénine, S. & Bonthoux, F. (2006). The Formation of Living and Non-Living Superordinate Concepts as a Function of Individual Differences. *Current Psychology Letters: Behaviour, Brain and Cognition [Online], 19(2)*, online since 14 décembre 2006, connection on 06 février 2010. URL : http://cpl.revues.org/index1066.html.

Keil, F. (1986). Conceptual development and category structure. In U. Neisser (Ed.), *Concepts and conceptual development*. Cambridge: Cambridge University Press.

Keil, F. C. (1989). *Concepts, kinds, and cognitive development*. Cambridge, MA: MIT Press.

Kripke, S. (1980). *Naming and necessity*. Cambridge, MA: Harvard University Press.

Lee, L. C., Kagan, J., & Rabson, A. (1963). Influence of a Preference for Analytic Categorization upon Concept Acquisition. *Child Development, 34 (2)*, 433-442.

Lewellen M. J., Goldinger S. D., Pisoni D. B., & Greene B. G. (1993). Lexical familiarity and processing efficiency: individual differences in naming, lexical decision, and semantic categorization. *Journal of Experimental Psychology: General, 122(3)*, 316-330.

Malt, B. C. (1994). Water is not H_2O. *Cognitive Psychology, 27*, 41-70.

Malt, B. C. & Sloman, S. A. (2007). Category essence or essentially pragmatic? Creator's intention in naming and what's really what. *Cognition, 105(3)*, 615-648.

Mausner, B. & Graham, J. (1970). Field Dependence and Prior Reinforcement as Determinants of Social Interaction in Judgement. *Journal of Personality and Social Psychology, 16 (3)*, 486-492.

McKinley, S. C. & Nosofsky, R. M. (1995). Investigations of exemplar and decision bound models in large, ill-defined category structures. *Journal of Experimental Psychology: Human Perception & Performance, 21*, 128–148.

Medin, D. L., & Ortony, A. (1989). Psychological essentialism. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning*. New York: Cambridge University Press.

Norenzayan, A., Smith, E.E., Kim, B. J. & Nisbett, R. E. (2002). Cultural preferences for formal versus intuitive reasoning. *Cognitive Science, 26*, 653-684.

Putnam, H. (1975). The meaning of 'meaning.' In H. Putnam, *Mind, language, and reality: Philosophical papers, vol. 2*. Cambridge: Cambridge University Press.

Rips, L. J. (1989). Similarity, typicality, and categorization. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning*. New York: Cambridge University Press.

Rothbart, M., & Taylor, M. (1992). Category labels and social reality: Do we view social categories as natural kinds? In G. R. Semin & K. Fiedler (Eds.), *Language and Social Cognition* (pp. 11–36). London, UK: Sage.

Sternberg, R. J. & Grigorenko, E. L. (1997). Are Cognitive Styles Still in Style? *American Psychologist, 52 (7)*, 700-712.

Strevens, M. (2000). The essentialist aspect of naive theories. *Cognition, 74(2)*, 149-175.

von Eckardt, B. (1993). *What is cognitive science?* Cambridge, MA.: MIT Press.

Witkin, H. A. (1975). Some Implications of Research on Cognitive Style for Problem of Education. In J. M. Whitehead (Ed.), *Personality and Learning 1*, pp. 288-314. London: Hodder & Stoughton.

Witkin, H. A., Dyk., R. B., Faterson., H.F., Goodenough, D. R. & Karp., S. A. (1962). *Psychological Differentiation: Studies of Development*. New York: John Wiley and Sons, Inc.

Witkin, H. A., Oltman, P. K., Raskin, E. & Karp, S. A. (1971). *A Manual for the Embedded Figure Test*. Consulting Psychology Press.

When Comparison Helps: The Role of Language, Prior Knowledge and Similarity in Categorizing Novel Objects

Clare E. Sims (clare.holtpatrick@colorado.edu)
Department of Psychology and Neuroscience, 345 UCB
Boulder, CO 80309 USA

Eliana Colunga (eliana.colunga@colorado.edu)
Department of Psychology and Neuroscience, 345 UCB
Boulder, CO 80309 USA

Abstract

Research suggests a developmental shift from forming categories based on perceptual features, to recognizing deeper characteristics and relationships. One process found to highlight deeper properties is comparison between items. The bulk of the research on comparison, however, has been done with familiar items or familiar relationships. An open question remains: under which conditions will comparison help children attend to the deeper properties of novel objects? In two experiments we explore the effect of comparison in a word learning task and its interaction with prior knowledge, language support, and the perceptual features of the compared items. Our results suggest that comparison only highlights deeper similarities when children are given some support to counteract or reduce the influence of surface level features. These results have implications for how to best teach children depending on the amount of prior knowledge that they bring to the task.

Keywords: word learning; comparison; superficial vs. relational similarity

Introduction

The ability to look beyond surface similarities and make deeper connections between items is an important achievement in cognitive development. Although perceptual feature similarities (e.g., shape, material) are often a useful basis for grouping objects together, some categories may be better characterized by more abstract qualities. In analogy-making tasks and in categorization tasks, children seem to shift from attending to surface properties to being able to attend to relational or conceptual similarity at around age 5 (Gentner & Namy, 1999; Loewenstein & Gentner, 2005). The evidence also suggests that even young children who would not spontaneously attend to deeper similarities will do so with enough support. In two studies we examine the circumstances in which comparison is helpful in highlighting deeper similarities for novel categories for which children have no preexisting knowledge, even when the objects share perceptual features as well.

Comparison and deep similarities

A large body of research suggests that comparison is a mechanism that works to highlight deeper features. The evidence suggests that although a child may seem to only attend to surface, perceptual features when presented with a single item, being presented instead with two or more items to compare has the effect of highlighting deeper, relational features shared by those items. Researchers have found this effect in tasks such as word extension and analogical mapping. In these analogy tasks, children are typically shown a standard card showing three items that share some relational property. For example, Figure 1 shows a target card and two possible matches. The card on the left matches the target in superficial properties (they are three squares), whereas the card on the right matches it in relational similarity (two same-color figures flanking a different-color figure). Research using this sort of task indicates that without extra support, 4-year-olds will attend to properties of the specific items, rather than to the relational structure. However, when allowed to view two examples of the relational structure (say a second card with two white triangles flanking a black triangle) and compare it with the original target, the similarity in relational structure will be highlighted, allowing the child to abstract away from unimportant details such as the shapes of the items (Gentner, Rattermann, Markman, & Kotovsky, 1995).

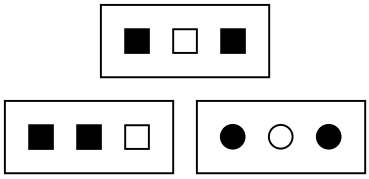


Figure 1: Example of an analogical mapping task.

A similar effect of comparison has been found in word learning tasks. Gentner and Namy (1999) taught 4-year-olds new names for known objects. For example, they showed children a picture of an apple and informed them that a toy dog had a special name for it: “blicket”. Then children were offered a choice between a picture of a banana and a picture of a balloon and asked which of those would also be called a blicket. Surprisingly, four-year-olds chose the superficially similar balloon, rather than the taxonomic match, the banana, also a fruit. However, when allowed to see more than one standard (e.g., a pear and a bunch of grapes), children chose more taxonomic matches. More recently Gentner, Loewenstein, and Hung (2007) found that four-year-olds were similarly able to use

comparison to learn novel names for specific parts of novel animal and object drawings. The process of comparison is thought to highlight deeper relations between items by promoting structural alignment (Gentner & Namy, 1999). Structural alignment refers to how considering two or more items together can focus attention on common relational structure that would not be readily apparent from only a single item. But how would comparison work if novel objects were used?

Comparison of Novel Objects

A first step in investigating whether comparison of novel objects can help children see beyond surface properties to deep features is to determine what these terms mean in the context of novel stimuli. A definition of surface properties can be transferred fairly straightforwardly from previous comparison research involving familiar items. That is, surface properties consist of perceptual features such as the shape, material, texture, and color of an object. Defining deep characteristics of novel objects is more complicated, for exactly the reason that such objects were chosen for this research: there is no prior knowledge or history associated with them. For example, in some previous work the deep features that are highlighted in comparison are defined as conceptual representations (Gentner & Namy, 2006). However, in other contexts, particularly analogical mapping studies, deep characteristics consist of the higher order relations or structures shared across items (Gentner et al., 1995). Gentner and Rattermann (1991) refer to this deeper level of similarity as analogy, and define it as "similarity in relational structure, independently of the objects in which those relations are embedded" (p. 226). For example, Figure 1 shows a sample analogical mapping involving a structure of two identical darker figures flanking a lighter figure. Understanding this structure at the analogy level means understanding that such a relation can encompass a flanking structure in various dimensions, such as darkness or size, and despite dissimilarities in other features like shape or texture (Gentner et al., 1995).

Drawing from this research on analogy, in the current experiments deep characteristics of novel objects are defined as the structure of the objects. Specifically, we designed the novel objects in these experiments such that the connections and relations betweens parts of individual objects conform to a generalizable structure. Figure 2 shows an example set of the novel object stimuli created for these experiments. The structure shared by the exemplars and structure choice test item in this set is one of three identical shapes, arranged vertically, and decreasing in size from bottom to top. While there is variation in surface level features, such as shape, color, material, and texture, the underlying structure is maintained. In this way, the novel objects created for the current experiments were carefully manipulated to have certain surface properties and deep features, in particular relational structure.

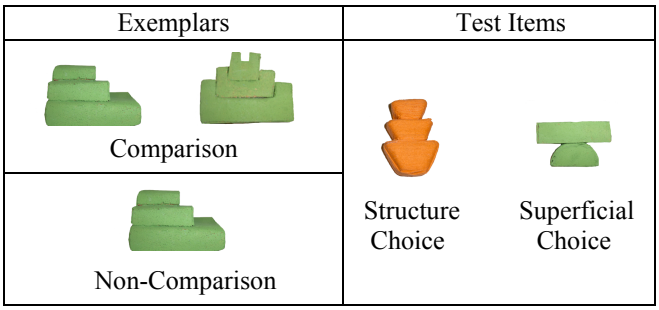


Figure 2: Sample item set from Experiment 1. Materials used include green foam and orange yarn.

Prior Knowledge Another key issue to consider in relation to the comparison of novel objects is the role of prior knowledge about the items that are compared. Much of the research cited so far supports the hypothesis that the development from categorization based on surface similarities to categorization based on higher order relations is driven by increases in domain knowledge. For example, understanding of higher order analogical relations has been found to develop between the ages of 4 and 8 years, but 4-year-olds can learn to appreciate and correctly use such relations through explicit teaching or over the course of targeted training (Kotovsky & Gentner, 1996). Experiments with novel objects offer a new way to test, and possibly further support, this hypothesis by controlling the amount of domain knowledge that participants have available. As will be shown in the first experiment, our novel object stimuli allowed us to directly explore the question of the role of prior knowledge in comparison. Using novel objects ensured that participants were not familiar with the stimuli, and we also manipulated the labels used (novel vs. known) to further control the amount of prior knowledge brought into the task. The label manipulation relates to the next issue as well.

Language Use Another guiding question of our design of the current experiments has to do with the role of language. Specifically, in the first experiment we explore an intriguing finding on the role of language in analogical mapping. Previous research shows that children as young as three years old can map familiar relational labels, like "top," "middle," and "bottom" onto spatial relations between presented items, and use those mappings to make correct relational choices, even in the face of tempting perceptual choices (Loewenstein & Gentner, 2005). We wondered whether the use of familiar structure related labels could have a similar influence on children's comparison processes with novel objects. Such labels would offer children some support in linking familiar structural representations with the novel objects to be learned; the first experiment tests whether they can effectively use this support to aid task performance.

Perceptual Features Our final guiding question about the role of perceptual features influenced our overall task

design and is the focus of the second experiment. To address this question we drew on research from the word learning literature. Our task has many parallels to the novel noun generalization (NNG) paradigm, in which a child is presented with a novel object that is labeled with a novel name, and is subsequently asked whether he or she would apply that label to various other novel objects that vary from the original in specific dimensions such as shape, size, color, or texture, while matching in other dimensions. Research using the NNG task has provided many interesting findings about the kinds of object properties that children use to guide their learning and labeling of different kinds of items. For example, from a young age children consistently and preferentially use the shape of an object, as opposed to other features like size, color, or material, to guide their labeling and categorization of artifact-like items (Jones, Smith, & Landau, 1991). On the other hand, the material of an item is treated as more important than other features in guiding children's novel noun generalization of non-solid substances (Soja, Carey, & Spelke, 1991). Because the novel objects created for the current experiments are artifact-like, we were concerned that having a shape match between exemplars and test items would strongly influence our results. To avoid the possible confound of a shape match, we minimized the degree of shape matching and manipulated other features known to be less influential in artifact-like object naming, particularly material and color. The second experiment in particular explores how the manipulation of these perceptual features influences comparison.

In two experiments we explored the effect of comparison on preschooler's learning about novel objects. We used previous work on comparison as well as word learning to guide our experimental task design. The current experiments explore the roles of prior knowledge, of language, and of perceptual features in children's comparisons of novel objects. The first experiment explores the role of prior knowledge and language, and the second experiment focuses on the role of perceptual features.

Experiment 1

The first goal of Experiment 1 was to create and test an experimental task that paralleled those used in the comparison literature but that involved only novel objects. To this end, we modeled our task after one designed by Gentner and Namy (1999, Experiment 2) in which children were presented with either one or two standard items (non-comparison and comparison conditions) and then decided which one of two test items best matched the standard, a perceptual choice or a taxonomic choice. The authors used drawings of familiar items, and carefully chose the stimuli such that each standard item was more strongly perceptually similar to the perceptual choice than the taxonomic choice. The design of our stimuli aimed to capture similar relations between the perceptual and structural characteristics of novel objects. We manipulated the surface level features of material, color, and shape to create exemplar objects that strongly matched the superficial choice test object. We manipulated the relationships between the parts of these objects to create structural similarities between the exemplar objects and the structural choice test object, which was perceptually dissimilar to the exemplars (see Figure 2). In this way we believe our experimental task is an accurate translation of the Gentner and Namy (1999) task from familiar item drawings to novel physical objects.

The other goal of Experiment 1 was to test two of our guiding questions: what is the role of prior knowledge and what is the role of language in novel object comparison? We included two labeling conditions: a novel label condition and a known label condition consisting of structurally related familiar words. The combination of novel objects and novel labels ensured that children in that condition had no prior knowledge of the task items. In the known label condition, we used familiar words that related to structure to see whether children could effectively use language support to make connections to known structural relationships.

If children treat novel objects similarly to how they treat familiar items, then we should see similar results in our novel label condition as those of Gentner and Namy (1999); that is, children will make more perceptual choices when there is non-comparison between exemplars, and will make more structural choices when there is comparison. In other words, comparison of novel objects will function as it does with familiar items, highlighting the deeper relations present between them. On the other hand, because children have no prior knowledge of novel objects given novel labels, comparison might not function in the same way, perhaps instead highlighting surface rather than deep features. Additionally, we expected the use of known labels to be effective in highlighting object structure in both the comparison and non-comparison conditions, and thus act to increase children's structural choices in both comparison conditions.

Method

Participants. Fifty-two 4-year-olds (M = 4;6) were assigned to the comparison or non-comparison condition, and to the novel label or known label condition in a 2 x 2 design.

Materials. The stimuli consisted of 16 novel objects created in the lab (8 exemplars and 8 test items). There were four sets of test items consisting of a structure choice and a superficial choice. For each set of test items there were two exemplars (see Figure 2 for a sample set). All exemplars were structural matches with the structure choice for their group, and also matched in material and color with the superficial choice. Due to the extent of surface similarity between the pairs of exemplars, this set is referred to as "high similarity" throughout this paper.

Both of the exemplars were used for the comparison condition, and the exemplar that matched the superficial choice somewhat on shape was used for the non-comparison condition.

Two types of labels were also used in two labeling conditions: novel and known. For the novel labeling condition four pseudoword labels were created, one for each set of objects. For the known labeling condition, four real, structurally related words were selected to go with each set of objects. The words selected were intended to be familiar to four-year-olds; for example, the items in Figure 2 were given the familiar label "stairs." The other known labels were "see-saw," "bumps," and "spiral."

Procedure. Participants sat at a table across from the experimenter. Participants were randomly assigned to one of four conditions: novel label comparison, novel label non-comparison, known label comparison, and known label non-comparison.

In the comparison conditions, the experimenter showed the participant two exemplar objects and labeled each with either the same novel label or with the same known label. For example, in the novel label comparison condition, the experimenter would say, "This is a tink. This is also a tink. See how they are both tinks?" The participant was able to examine the objects before the experimenter put them out of sight. Then the experimenter brought out two test items on a tray and asked the participant to "Get the tink." The experimenter then recorded whether the participant chose the structure match or superficial match. The non-comparison condition proceeded in a similar manner but with only one exemplar shown and labeled, for example "This is a tink. See how it is a tink?" In all conditions participants completed four trials with the order of trials counterbalanced across participants. In the novel label conditions the novel nouns used to label the four stimuli sets were also counterbalanced.

Results

The dependent variable was the average number of structure match choices that participants made across all test trials. Average numbers of structure choices were submitted to a 2 (comparison or non-comparison) x 2 (label type: novel or known) between-subjects analysis of variance (ANOVA). There was a main effect of label type such that children made more structure choices when objects in the task were given familiar, relational labels, $F(1, 48) = 12.43$, $p < .001$. There was also a significant interaction between comparison condition and label type, $F(1, 48) = 4.72$, $p = 0.03$ (see Figure 3). In the novel label condition, children made fewer structure choices after comparing two exemplars compared to viewing only one exemplar. This relationship was reversed in the known label condition, with more structure choices made in the comparison condition than the non-comparison condition.

Post hoc *t*-tests were conducted to further explore this interaction. Within the novel label condition, structure choices were marginally lower in the comparison than non-comparison condition, $t(24) = -1.77$, $p = 0.089$. This difference did not reach significance in the known label condition ($p = 0.17$), however looking across labeling conditions, structure choices following comparison were significantly higher with known compared to novel labels, $t(29) = 4.46$, $p < .001$.

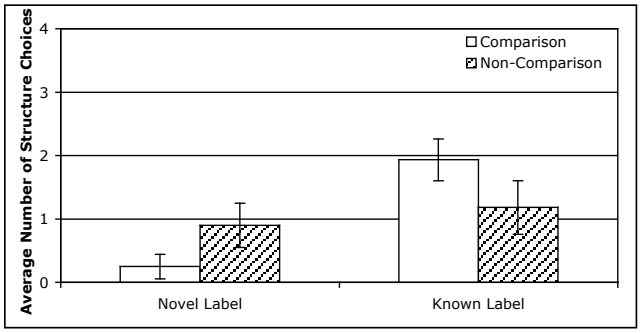


Figure 3: Experiment 1 results; all stimuli were high similarity.

Discussion

Overall the results of Experiment 1 show that the availability of prior knowledge in a task influences how comparison functions. Comparing the novel label condition to the Gentner and Namy (1999) experiment on which it was modeled, we found the opposite pattern of results. Rather than highlighting deeper relations between items, comparison in the novel label condition seemed to very strongly cue children's attention to the perceptual similarities between novel objects. The results of the novel label condition add support to the hypothesis that the capacity of comparison to highlight deeper features depends on the amount of relevant domain knowledge that participants have. In the novel label condition participants had no prior knowledge of the objects to help them identify the deeper structural similarities, and instead focused on surface level features to guide responses. Overall, the novel label condition of Experiment 1 shows that comparison of objects about which children have no prior knowledge functions differently than comparison of familiar items.

On the other hand, the results of the known label condition show that if there is some conceptual support, such as familiar labels that highlight the structure of objects, comparison seems to work in a way similar to that seen in studies using familiar items. This result is also consistent with previous work showing that young children can use language to guide performance in analogical mapping tasks (Loewenstein & Gentner, 2005). The role of language in this context seems to be to situate the novel objects in terms of familiar representations, allowing for recognition and use of structural properties in the task.

In the next experiment we explore the role of perceptual similarity in novel object comparison.

Experiment 2

The goal of Experiment 2 was to test the role of perceptual features in comparison of novel objects. This experiment also allowed for further exploration of how comparison operates in a context of low prior knowledge, that is, our task involving novel objects with novel labels. The first experiment showed that the process of comparison was only conducive to making structural choices when supported by familiar structure-related labels. In the second experiment we set out to investigate another way in which comparison of novel objects would highlight their deeper shared structure. To this end we used the same general procedure as in the novel label condition of Experiment 1, but varied the perceptual features of the novel objects. In Experiment 2, the stimuli were designed such that there was a lower degree of surface feature similarity between the exemplars in relation to each other as well as in relation to the test items (see Figure 4). With this manipulation, the results of the task using the low surface similarity novel object set can be directly compared to the results of the Experiment 1 novel label condition, which used a high surface similarity object set.

We predicted that the number of structure choice responses would increase overall when the task involved low surface similarity novel objects as compared to the high similarity objects of Experiment 1. Comparison of the high similarity objects seemed to more strongly highlight the perceptual feature overlap than the common structure of the objects. Therefore we reasoned that reducing the degree of that overlap should reduce the amount that comparison highlights surface features, and allow children to see the deeper relational match of the structure choice.

Method

Participants. Twenty-four additional four-year-olds (M = 4;4) were recruited for the second experiment, and were randomly assigned to the comparison and non-comparison conditions.

Materials. The stimuli consisted of 16 novel objects created in the lab (8 low similarity exemplars and the 8 test items from Experiment 1). As in the first experiment, there were four sets of test items consisting of a structure choice and a distractor choice, and four corresponding pairs of exemplars. All exemplars were structural matches with the structure choice for their group. For the low similarity exemplars, one object matched the distractor choice somewhat in shape only, and the other object did not match the distractor choice in shape, material, or color. For each pair of exemplars, both objects were used for the comparison condition, and the exemplar that matched the distractor choice somewhat on shape was used for the non-comparison condition.

Procedure. The procedure was the same as that used for the novel label condition of Experiment 1. Participants were randomly assigned to one of two conditions: comparison or non-comparison. As in the first experiment, each participant completed four trials; trial presentation order as well as the novel nouns used to label the four stimuli sets were counterbalanced.

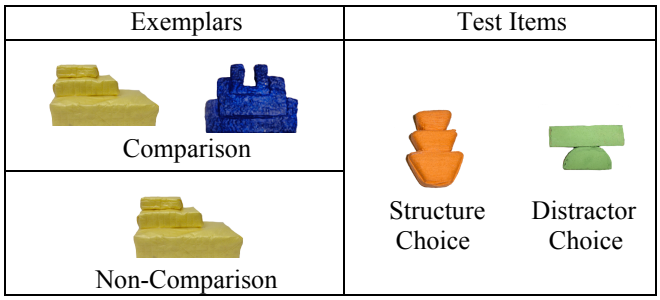


Figure 4: Sample item set from Experiment 2. Materials used include yellow cellophane, blue clay, orange yarn, and green foam.

Results

As in Experiment 1 the dependent variable was the average number of structure match choices that participants made across all test trials. Average numbers of structure choices from both Experiment 2 (low similarity) as well as the novel label condition of Experiment 1 (high similarity) were submitted to a 2 (surface similarity: high or low) x 2 (condition: comparison or non-comparison) ANOVA. There was a main effect of surface similarity such that number of structure match choices was higher when surface similarity was low, $F(1, 46) = 26.36$, $p < 0.001$. There was also a significant interaction between surface similarity and comparison condition, $F(1, 46) = 6.07$, $p = 0.02$ (see Figure 5). As shown in the results of Experiment 1, in the high surface similarity condition (novel label condition of Experiment 1) children made fewer structure choices after comparing two exemplars compared to initially viewing only one exemplar. The interaction here shows that this relationship reversed for the low surface similarity objects used in Experiment 2: children made more structure choices in the comparison condition than the non-comparison condition.

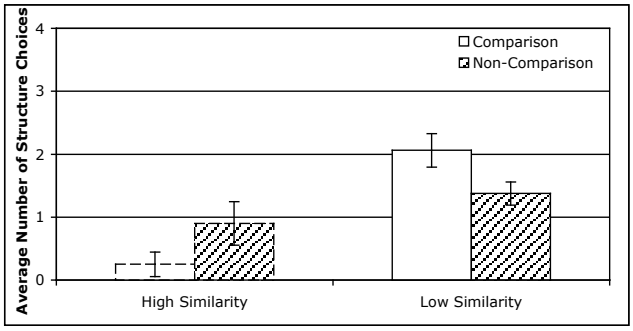


Figure 5: Experiment 2 results; all labels were novel. Note high similarity data is from the novel label condition of Experiment 1.

Post hoc *t*-tests were conducted to further explore this interaction. Participants were able to make significantly more structure match choices in the comparison condition when the exemplars had low surface similarity rather than high surface similarity, $t(30) = 5.51$, $p < 0.001$. Additionally, there was not a significant difference in number of structure choices in the non-comparison condition between high and low surface similarity, $p = 0.28$.

Discussion

The results of Experiment 2 help to shed some light on the role of perceptual feature similarities in comparison processes involving novel objects. In this experiment we varied the extent to which the exemplar objects shared surface feature similarities with the distractor choice test item. As shown in Experiment 1, when the exemplars were highly similar to the superficial choice in several dimensions, specifically material, color, and shape, comparison actually seemed to highlight surface similarities especially strongly. However when children performed the same task but with the low surface similarity exemplar objects of Experiment 2, comparison seemed to better highlight the deeper, structural relations between the exemplars and the structure choice test item. While this increase in number of structure choices in the comparison condition for low similarity as compared to high similarity objects was predicted, what is surprising is the magnitude of the increase. Specifically, administering this task with low surface similarity novel objects increased performance, in terms of number of structure choices, to the same extent as labeling objects with familiar structure related words.

General Discussion

In the current experiments we set out to explore the role of prior knowledge, language, and perceptual features in making comparisons of novel objects. We wondered whether the act of comparison highlights deeper relations rather than surface similarities, as has been found with studies using familiar items. In Experiment 1 we found that, in line with previous research in analogy making, prior knowledge of the items being compared does indeed matter: comparison *hindered* performance when novel objects and novel labels were used, that is, when prior knowledge of items was low. In the first experiment we also found that the use of known, structure-related labels led to increased identification of structural matches between novel objects. This indicates that language plays a role of supporting the integration of prior knowledge with new category information. In Experiment 2 we found that perceptual features also impact the comparison of novel objects. Reducing the degree of surface similarity between exemplar and test objects improved performance to the same extent as using familiar labels. Together these experiments show that in order for comparison to be beneficial, support has to be provided through links to prior knowledge. In the absence of prior knowledge that can be brought to bear, it is important to ensure that the possibility of mistakenly highlighting surface is minimized by comparing items of low similarity.

These results have implications for educational practices related to teaching new categories and concepts. Linking newly introduced items to familiar concepts, particularly with language, helps children make deeper connections between new items, and perhaps aids them in creating rich representations of new categories. Additionally, introducing new objects by presenting items that are more variable in surface features (i.e., share less surface similarity) further helps children by reducing the tendency to attend only to the superficial, especially when those feature similarities run counter to the deeper relationships.

References

Gentner, D., Loewenstein, J., & Hung, B. (2007). Comparison facilitates children's learning of names for parts. *Journal of Cognition and Development, 8,* 285-307.

Gentner, D. & Namy, L.L. (1999). Comparison in the development of categories. *Cognitive Development, 14,* 487-513.

Gentner, D., & Namy, L.L. (2006). Analogical processes in language learning. *Current Directions in Psychological Science, 15,* 297-301.

Gentner, D. & Rattermann, M.J. (1991) Language and the career of similarity. In S.A. Gelman & J.P. Brynes (Eds.) *Perspectives on Language and Thought: Interrelations in Development*. London: Cambridge University Press.

Gentner, D., Rattermann, M.J., Markman, A.B., & Kotovsky, L. (1995). Two forces in the development of relational similarity. In T.J. Simon & G.S. Halford (Eds.), Developing cognitive competence: New approaches to process modeling. Hillsdale, NJ: LEA.

Jones, S.S., Smith, L.B., & Landau, B. (1991). Object properties and knowledge in early lexical learning. *Child Development, 62,* 499-516.

Kotovsky, L., & Gentner, D. (1996). Comparison and categorization in the development of relational similarity. *Child Development, 67,* 2797-2822.

Loewenstein, J., & Gentner, D. (2005). Relational language and the development of relational mapping. *Cognitive Psychology, 50,* 315-353.

Soja, N.N., Carey, S., & Spelke, E.S. (1991). Ontological categories guide young children's inductions of word meanings: Object terms and substance terms. *Cognition, 38,* 179-211.

Category Learning and Adaptive Benefits of Aging

Angela Merritt (merri242@umn.edu)
University of Minnesota

Linnea Karlsson (karlsson@mpib-berlin.mpg.de)
Max Planck Institute for Human Development

Edward T. Cokely (cokely@mpib-berlin.mpg.de)
Max Planck Institute for Human Development &
Michigan Technological University

Abstract

We examined effects of normal aging on category learning, comparing performance and strategy choice on two learning tasks: one where a one-dimensional rule governed category membership and one where a multi-dimensional rule defined category structure. Paradoxically, we demonstrated that older adults can outperform younger adults in some types of complex category learning. In the current task—which required that multiple dimensions be integrated—simpler integration rules enabled more rapid achievement of reasonable levels of performance. As cognitive aging is associated with a reduction in working memory resources, older adults tended to adopt these simpler decision rules more often, facilitating complex category learning. Results provide some unique evidence highlighting potential adaptive benefits of cognitive aging. Implications are discussed.

Keywords: category learning, aging, rule-based, information-integration

Introduction

The learning of new categories is an important task throughout one's life span. While the literature investigating younger adults' category learning skill is vast, less is known about older adults' category learning competencies. In the current investigation, we sought to demonstrate that normal cognitive aging can confer cognitive performance benefits as older adults may favor simpler cognitive strategies known to facilitate learning and decision performance.

Previous research investigating aging and feedback-based category learning often draws the general framework of Ashby and colleagues (e.g. Ashby, Alfonso-Reese, Turken, & Waldron 1998). The framework involves contrasting two different task types that are assumed to be best solved by two distinct learning systems. The most basic difference between the tasks, as typically stated, is whether one or several dimensions of the probe determine category membership. In the so-called *rule-based tasks*, only one dimension of the probe determines category membership and these tasks are believed to rely on an explicit learning system capitalizing on simple verbalizable rules. In *information-integration tasks*, the values of several dimensions determines membership via a complicated combination rule. In these tasks, simple one-dimensional rules will not suffice for error-free performance. Learning in these tasks is thus said to be guided by an implicit learning system employing integration of dimensions at a "pre-decisional" stage. Moreover, it is suggested that there exists a *rule-bias*—a new learning endeavor will start off with the explicit learning system but compete with, and possibly lose against, the implicit system for determining the response.

The results with regard to older adults' learning in these tasks are mixed. Ashby, Nobel, Filoteo, Waldron and Ell (2003) demonstrated that older adults reached the learning criterion (10 correct consecutive responses, CCR) later than young adults in both a rule-based and an information-integration task. However, they did not investigate the cognitive processes used to guide categorization. Filoteo and Maddox (2004) compared younger and older adults on two versions of an information-integration task—one with a linear and one with a non-linear combination rule. Older adults were impaired compared to young adults on both versions. Via computational modeling the authors provided evidence suggesting that the age-related differences were less marked among individuals using simple one- or two-dimensional rules. In contrast, in a study comparing young and old adults on a probabilistic category learning task (the weather prediction task; Gluck & Bower, 1988) and an information integration task, age-related differences were only found in the probabilistic but not in the information integration task (Price, 2005).

Individual differences in working memory are known to influence cognitive task strategies and decision making performance (Cokely & Kelley, 2009; Cokely, Kelley, & Gilchrist, 2006). Furthermore, considerable evidence has documented declines and metacognitive changes associated with working memory during normal aging (e.g. Baltes, Staudinger, & Lindenberger, 1999; Herzog, Dixion, Hultsch, & MacDonald, 2003). Interestingly, individual differences in working memory have been shown to be a factor on success rates in category learning (DeCaro, Thomas, & Beilock, 2008). DeCaro et al. hypothesized that individuals with high working memory abilities should outperform individuals with low working memory abilities in rule-based tasks. In information-integration tasks it was hypothesized that low-capacity individuals would have a benefit: they may have less capacity to engage the explicit system in extensive hypothesis testing of the complex com-

bination rule. Thus low ability individuals might switch to the implicit system earlier than high-ability individuals and show faster learning. DeCaro et al. showed that high-ability individuals reached the learning criterion faster than low-ability individuals in a rule-based task. In contrast, in an information-integration task, low-ability individuals reached the learning criterion faster—a benefit assumed to stem from an earlier switch to the implicit system. Of note, however, Tharp and Pickering (2009) demonstrated that the learning criterion used by DeCaro et al (i.e. 8 CCR) were insufficient for capturing learning of the information-integration combination rule and thus reaching that criterion is likely not a reliable indicator that implicit learning has taken place. Tharp and Pickering demonstrated that considerably fewer participants in an information-integration task were able to sustain performance long enough to reach the stricter criterion of 16 CCR. Further, the responses from around 40% of the individuals reaching the 8 CCR criterion could be well captured by one-dimensional categorization models, suggesting that it is possible to reach 8 CCR with explicit memory and simple one-dimensional rules. DeCaro, Carlson, Thomas and Beilock (2009) subsequently tested low- vs. high ability individuals again, using the stricter 16 CCR, and the interaction between tasks and abilities disappeared. Moreover, when assessing learning strategies in the two tasks for the two groups DeCaro et al found evidence suggesting that the low-ability individuals primarily used one-dimensional rules.

Might older adults also benefit from the use of simpler processes in the tasks used by DeCaro et al and Tharp and Pickering? Previous research suggests it is unlikely that that older adults would be able to proceed in learning the information integration task with the implicit learning system (see also Filoteo & Maddox, 2004). Previous research also suggests that in tasks similar to implicit category learning (i.e., implicit learning of new associations) age-related decline in performance is to be expected (Curran, 1997; Harrington & Haaland, 1992; Howard & Howard, 1997, 2001). Moreover, there is evidence that older adults prefer simpler strategies over complex strategies in various tasks, for example in mental arithmetic (Geary, Frensch & Wiley, 1993), in memory (Dunlosky & Hertzog, 1998, 2000) and decision making (Chen & Sun, 2003; Johnson, 1990; Mata, Schooler & Rieskamp, 2007). Thus, this leaves us with two nested hypotheses for the current study: (1) older adults will not address the information integration task with the implicit system, and (2) advantages demonstrated by older adults in an information-integration task will stem from older adults' use of simple, verbalizable rules, in contrast to a larger portion of younger adults who might attempt futile hypothesis testing with more complex rules.

Experiment

In the following experiment we tested younger adults and older adults on a categorization task. The stimuli consisted of pictorial drawings with four binary cues and the task was to learn to categorize the stimuli into two different categories with guidance by outcome-feedback. The task exactly followed DeCaro et al. (2008).

Method

Participants

Fifty eight participants were tested. Twenty nine of the participants were younger, aged 20-32 (m = 25.1, SD = 3.1), and the other 29 participants, were older, aged 64-79 (m = 69.9, SD = 3.3). Participants were recruited from the participant pool of the Max Planck Institute for Human Development, Berlin. All were compensated 10 € for participation.

Procedure

Participants completed a computer-based category learning experiment adapted from DeCaro, Thomas & Beilock (2008). During the experiment, the participant was shown colored geometric figures on a computer screen and asked to place each one into either category "A" or category "B" by pressing buttons on a keyboard. Immediate feedback was given after each trial. After 200 of such trials, the participant was informed that a new set was to begin and the rules had changed, but were not informed by which rule to sort. Participants completed 4 sets of 200 trials. There were two different sets of rule-based tasks and two sets of information integration tasks, rotated across participants in four different orders. In the rule-based tasks one dimension decided category membership (in one set it was symbol color and in the other set symbol shape). There were also two different sets of information-integration tasks. Three of the four dimensions were regarded as relevant (with background color respectively number of embedded symbols being irrelevant). The correct combination rule was given by assigning each binary value of the dimensions with 1 or -1 and then linearly combining those values:

If value (X) + value (Y) + value (Z) > 0 respond A, otherwise B.

In addition, participants completed a battery of cognitive ability measures. These results are not reported as they are beyond the scope the current paper.

Results

As a first step, to statistically investigate the extent to which we replicated DeCaro et al (2008), our initial analyses followed DeCaro et al. First, we log-transformed the number of trials to reach the criterion

of 8 CCR, as the variable was positively skewed. Second, for the set of analyses directly aiming at comparing the results with DeCaro et al (2008) we only included participants who reached the criterion on all four task rules (two rule-based and two information-integration, of 200 trials each), and who were not higher than 2 *SD* above the mean in trials to criterion in each block.

First, we analyzed whether performance on any of the two different rules within each task differed and interacted with age. We performed one repeated measurement ANOVA per task, with rule type as within factor and age as between factor. In the rule-based task one rule was easier to learn than the other ($F(1,41) = 5.17$; $p = .03$) but this did not interact with age ($F(1,41) = .12$; $p = .73$). In the information-integration task rules did not differ in difficulty ($F(1,41) = .001$; $p = .97$) and there was no interaction with age ($F(1,41) = .26$; $p = .62$). Therefore, we averaged data across both rule types within each task.

Second, we investigated whether there was an effect of age on the ability to reach the 8 CCR criterion and if this interacted with task. We performed a repeated measurement ANOVA with task (rule-based vs. information-integration) as within-subjects factor and age as between-subjects factor. Overall, the age groups did not differ on the number of trials they took to reach the criterion ($F(1,41) = .07$; $p = .80$). The criterion was reached faster in the rule-based than the information-integration task ($F(1,41) = 9.96$; $p = .003$). Most importantly, there was a significant interaction between age and performance in the two tasks (Figure 1; $F(1,41) = 4.69$; $p = .04$). While the younger adults' ability to reach the criterion deteriorated significantly in the information-integration compared to the rule-based task ($F(1,15) = 28.05$; $p < 0.001$) the older adults reached it about equally fast across tasks ($F(1,15) = .25$; $p = .62$).

Following Ashby et al. (2003), DeCaro et al.(2008, 2009), and Tharp and Pickering (2009), and to further investigate the learning trajectories in the two tasks we next looked at the number of participants reaching the three different criteria used in those studies (8, 10 and 16 CCR) and the mean number of trials it took to reach the criterion (Table 1 and 2). All subsequent analyses included all participants.

It is evident that in the rule-based task most younger adults reached all three criteria while about 1/3 of the old adults did not reach the strictest criterion (Table 1). In the information-integration task on the other hand (Table 2), fewer older adults reached all criteria, but the number of learners dropped off proportionally in both age-groups as a function of how strict the criterion was. Critically, the older adults required fewer trials to criterion than the younger adults only when considering 8 CCR.

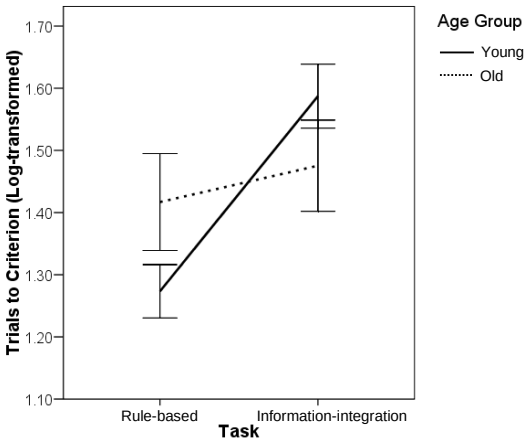


Figure 1. Average trials to reach the 8 CCR criterion as a factor of age group and task. Error bars represent +/- 1 SE.

Table 1. Rule-based task: number of learners and mean trials-to-criterion (TTC) as a function of age and criterion

	Number of learners (max 29 per group)		Mean TTC *(SD)*	
CCR	Young	Old	Young	Old
8	28	22	25.9 *(16.1)*	45.0 *(33.1)*
10	28	19	32.0 *(19.0)*	48.2 *(35.2)*
16	26	14	44.8 *(25.1)*	52.7 *(30.4)*

Table 2. Information-integration task: number of learners and mean trials-to-criterion (TTC) as a function of age and criterion

	Number of learners (max 29 per group)		Mean TTC *(SD)*	
CCR	Young	Old	Young	Old
8	27	19	53.9 *(32.2)*	47.5 *(31.3)*
10	21	15	69.0 *(25.1)*	77.9 *(33.8)*
16	4	2	91.0 *(21.6)*	155.8 *(2.48)*

To provide a more transparent impression of performance in the two tasks we next examined the proportion of correct responses as a function of task and age (Figure 2). First, performance on the two rule-types of each task did not interact with age, so we averaged the data across rule-types. In a repeated measurement ANOVA there were two main effects

and one interaction. The younger adults performed better overall than the older adults ($F(1,56) = 14.74$; $p < 0.001$), and performance was better in the rule-based than in the information-integration task ($F(1,56) = 124.4$; $p < 0.001$). The interaction suggests that the impact of age on performance was different depending on the task ($F(1,56) = 4.04$; $p = .049$). The difference between the age groups was larger in the rule-based task (m_{young} = 91.2% vs. m_{old} = 78.8%) than in the information-integration task (m_{young} = 71.3% vs. m_{old} 65.0%).

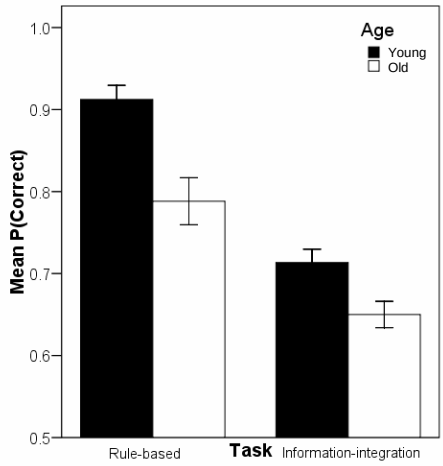


Figure 2. Proportion correct as a function of task and age-group. Error bars are SE +/- 1.

Next, to investigate whether the observed interaction reflected differences in the use of strategies we performed a rough strategy assessment for the information-integration task. The present task is limited when it comes to allowing reliable estimates of what model explains the data best and we thus refrain from sophisticated quantitative model assessments including parameter estimation. Indeed, several different models, including exemplar-models (e.g. Juslin, Olsson, & Olsson, 2003; Nosofsky & Johansen, 2000) and rule-plus-exception models (Nosofsky, Palmeri, & McKinley, 1994) are likely to give rise to a set of similar responses as one-dimensional and multi-dimensional rules under some values of the parameters. However, for the purpose of the present paper we were primarily interested in whether the age group differed in how many dimensions they utilized. A simplified means of assessing the correspondence between data and predictions from a model is to count the number of trials for which the data and the model gives the same answer (e.g. DeCaro et al., 2009). We defined the same set of 10 different strategies as DeCaro and colleagues (2009), including the correct three-dimensional rule, three different one-dimensional rules, and six different multi-dimensional rules (all of which could potentially be easily verbally described). Because we wanted to investigate what strategy accounted for participants responses when they had reached the 8 CCR criterion (and not what different strategies were at play early during learning), we used the responses containing the 8 CCR as well as the subsequent 8 x 3 responses and compared them to the different strategies' predictions (in total 32 trials). We reasoned that at that point the response strategy should be more stable than during the beginning of the task when the participants presumably tried out different ways of responding. The model comparison was done separately for each of the two different rule-types of the information-integration task. Next, we looked at which model had the lowest deviation between responses and model predictions for each individual and each rule-type of the information-integration task. We did not count individuals where there was a tie between two or more strategies (separately for the two rule-types of the task). This resulted in a total of 40 valid strategy assessments for the young adults and 30 for the old adults. We contrasted multi-dimensional with one-dimensional strategies and counted the number of times (max 2 per person since there were two versions of the task) where a one-dimensional model or a multi-dimensional model had the lowest deviation. The results (Figure 3) suggest that for younger adults about equally many of the information-integration tasks were best described by a one-dimensional strategy (52.5 %) as by a multi-dimensional strategy (47.5 %). However, for the older adults more were better described by a one-dimensional strategy (76.7 vs 23.3 %).

Discussion

With this study we sought to demonstrate potential adaptive benefits of aging. In a task where category membership was governed by the integration of several dimensions (in the present paradigm denoted an *information-integration task*) younger adults performed better than older adults overall (Figure 2). Importantly, however, older adults were able to produce reasonable levels of performance (i.e. to reach the 8 CCR criterion) somewhat faster than young adults (Figure 1). To investigate one potential mechanism underlying this advantage we did a simplified strategy assessment in the information-integration task. For the younger adults about equally many were best fit by one-dimensional as by multi-dimensional strategies. In contrast, for the older adults the larger proportion were best fit by one-dimensional strategies (Figure 3).

The results are intriguing in that they imply two important facets of age-related effects on the ability to acquire new categories. First, we find no evidence that the older adults engage an implicit learning system when trying to master the information-integration task. Had that been the case we should have observed sustained levels of performance inde-

pendent of the learning criterion. Instead we observed the opposite (Table 2). Further, we ascribe the reasonable levels of performance produced by the older adults in the information-integration task mainly to their adoption of simple, one-dimensional rules. For this particular task, such rules are able to lead to performance well above chance. Thus, while the younger adults presumably tried different versions of multi-dimensional rules, performance might have suffered initially from erroneous responses, while in the meantime the older adults could sustain reasonable performance by not doing that.

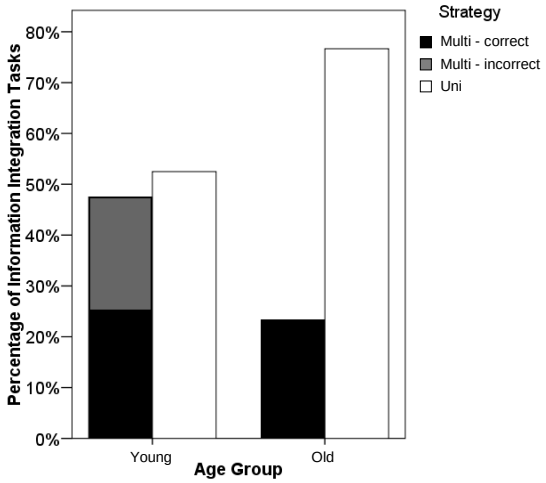


Figure 3. The proportion of information-integration tasks (with two different rule types) per age group where responses were best captured by a multi-dimensional (correct) multi-dimensional (incorrect) or a one-dimensional strategy.

The data presented in the current study replicates and extends the performance differences reported by Ashby et al (2003) who used the 10 CCR criterion. Our contribution extends the data presented by Ashby et al (2003) by demonstrating that older adults adopt one-dimensional rules to a larger extent than younger adults. Moreover, on the assumption that older adults represent a population with lower working memory capacities than younger adults we replicate DeCaro et al (2008, 2009), providing converging evidence on the influence of individual differences on the ability to acquire new categories.

Nevertheless, the data does not allow us to claim that younger adults engaged the implicit system in the information integration task. Performance dropped off as a function of learning criterion (Table 2). Further, nothing in the fit of a multi-dimensional strategy per se can tell us whether it was executed by an explicit or an implicit system. Unfortunately, there is some debate regarding whether the present set of stimuli are most suitable for studying the implicit system, as they may not be sufficiently complex (i.e. they involve binary stimuli dimensions). Rather, it has been suggested that the more complex Gabor patches are better for that purpose (e.g. Maddox, Ashby, & Bohill, 2003).

A number of interesting follow-up studies would help in clarifying some questions. First of all, it would be interesting to replicate the same experiment as reported here with the Gabor patch stimuli in order to investigate whether the ability to learn the tasks as well as the best performing strategies reveals the same pattern as reported here, even though the stimuli are more complex. Furthermore, follow-up experiments specifically designed for reliable quantitative model comparisons could provide a more detailed picture regarding the cognitive processes at play. Such experiments could for example aim at contrasting predictions by one- two- and three-dimensional rules with predictions by exemplar-models and rule-plus-exception models.

Results provide some new and unique data on potential benefits of cognitive aging. A large body of research has converged to reveal the benefits of simple decision strategies - in some cases *"less can be more"* (e.g. Gigerenzer, Todd, & the ABC Research Group, 1999). To the extent that cognitive aging biases older adults toward the use of simpler decision processes, there may be many benefits of cognitive aging that are currently underappreciated.

Authors' Note

We thank Gregor Caregnato and Paula Parpart for assistance in data collection. We thank DeCaro and colleagues for their experiment materials.

References

Ashby, F.G., Alfonso-Reese, L.A., Turken, U. & Waldron, E.M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review, 105,* 442-481.

Ashby, F.G., Noble, S., Filoteo, J. V., Waldron, E.M. & Ell, S.W. (2003). Category learning deficits in Parkinson's disease. *Neuropsychology,17,*115-124.

Baltes, P.B., Staudinger, U.M., & Lindenberger, U. (1999). Lifespan psychology: Theory and application to intellectual functioning. *Annual Review of Psychology, 50,* 471-507.

Cokely, E.T., & Kelley, C.M. (2009). Cognitive abilities and superior decision making under risk: A protocol analysis and process model evaluation. *Judgment and Decision Making, 4,* 20-33.

Cokely, E.T., Kelley, C.M., & Gilchrist, A.H. (2006). Sources of individual differences in working memory: Contributions of strategy to capacity. *Psychonomic Bulletin & Review, 13,* 991-997.

Chen, Y. W., & Sun, Y. L. (2003). Age differences in financial decision-making: Using simple heuristics. *Educational Gerontology, 29*, 627-635.

Curran, T. (1997). Effects of aging on implicit sequence learning: Accounting for sequence structure and explicit knowledge. *Psychological Research-Psychologische Forschung, 60*, 24-41.

DeCaro, M.S., Thomas, R.D., & Beilock, S.L. (2008). Individual differences in category learning: Sometimes less working memory capacity is more. *Cognition, 107*, 284-294.

DeCaro, M.S., Carlson, K. D., Thomas, R.D., & Beilock, S.L. (2009). When and how less is more: a reply to Tharp and Pickering. *Cognition, 111*, 415-421.

Dunlosky, J., & Hertzog, C. (1998). Aging and deficits in associative memory: What is the role of strategy production? *Psychology and Aging, 13*, 597-607.

Dunlosky, J., & Hertzog, C. (2000). Updating knowl edge about encoding strategies: A componential analysis of learning about strategy effectiveness from task experience. *Psychology and Aging, 15*, 462-474.

Filoteo, J. V., & Maddox, W. T. (2004). A quantitative model-based approach to examining aging effects on information-integration category learning. *Psychology and Aging, 19*, 171-182.

Geary, D. C., Frensch, P. A., & Wiley, J. G. (1993). Simple and Complex Mental Subtraction - Strategy Choice and Speed-of-Processing Differences in Younger and Older Adults. *Psychology and Aging, 8*, 242-256.

Gigerenzer, G., Todd, P. M., & the ABC Research Group (1999). *Simple heuristics that make us smart*. New York: Oxford University Press.

Gluck, M. A., & Bower, G. H. (1988). From Conditioning to Category Learning - an Adaptive Network Model. *Journal of Experimental Psychology-General, 117*, 227-247.

Harrington, D. L., & Haaland, K. Y. (1992). Skill Learning in the Elderly - Diminished Implicit and Explicit Memory for a Motor Sequence. *Psychology and Aging, 7*, 425-434.

Herzog, C., Dixion, R.A., Hultsch, D.F. & MacDonald, S.W.S. (2003). Latent change models of adult cognition: Are changes in processing speed and working memory associated with changes in episodic memory? *Psychology and Aging, 18*, 755-769.

Howard, D. V., & Howard, J. H. (2001). When it does hurt to try: Adult age differences in the effects of instructions on implicit pattern learning. *Psychonomic Bulletin & Review, 8*, 798-805.

Howard, J. H., & Howard, D. V. (1997). Age differences in implicit learning of higher order dependencies in serial patterns. *Psychology and Aging, 12*, 634-656.

Johnson, M. M. S. (1990). Age-Differences in Decision-Making - a Process Methodology for Examining Strategic Information-Processing. *Journals of Gerontology, 45*, P75-P78.

Juslin, P., Olsson, H., & Olsson, A. C. (2003). Exemplar effects in categorization and multiple-cue judgment. *Journal of Experimental Psychology-General, 132*, 133-156.

Maddox, W. T., Ashby, F. G., & Bohil, C. J. (2003). Delayed feedback effects on rule-based and information-integration category learning. *Journal of Experimental Psychology-Learning Memory and Cognition, 29*, 650-662.

Mata, R., Schooler, L.J., & Rieskamp, J. (2007). The aging decision maker: Cognitive aging and the adaptive selection of decision strategies. *Psychology and Aging, 22*, 796-810.

Nosofsky, R. M., & Johansen, M. K. (2000). Exemplar-based accounts of "multiple-system" phenomena in perceptual categorization. *Psychonomic Bulletin & Review, 7*, 375-402.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-Plus-Exception Model of Classification Learning. *Psychological Review, 101*, 53-79.

Price, A. L. (2005). Cortico-striatal contributions to category learning: dissociating the verbal and implicit systems. *Behavioral Neuroscience, 119*, 1438-1447.

Tharp, I. J., & Pickering, A. D. (2009). A note on DeCaro, Thomas, and Beilock (2008): Further data demonstrate complexities in the assessment of information-integration category learning. *Cognition, 111*, 410-414.

Encoding higher-order structure in visual working memory: A probabilistic model

Timothy F. Brady (tfbrady@mit.edu), Joshua B. Tenenbaum (jbt@mit.edu)
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology
Cambridge, MA 02139 USA

Abstract

When encoding a scene into memory, people store both the overall gist of the scene and detailed information about a few specific objects. Moreover, they use the gist to guide their choice of which specific objects to remember. However, formal models of change detection, like those used to estimate visual working memory capacity, generally assume people represent no higher-order structure about the display and choose which items to encode at random. We present a probabilistic model of change detection that attempts to bridge this gap by formalizing the encoding of both specific items and higher-order information about simple working memory displays. We show that this model successfully predicts change detection performance for individual displays of patterned dots. More generally, we show that it is necessary for the model to encode higher-order structure in order to accurately predict human performance in the change detection task. This work thus confirms and formalizes the role of higher-order structure in visual working memory.

Keywords: change detection; visual short-term memory; working memory; probabilistic model

Introduction

Working memory capacity constrains cognitive abilities in a wide variety of domains (Baddeley, 2000), including general intelligence and reading comprehension (Daneman & Carpenter, 1980). The architecture and limits of the working memory system have therefore been extensively studied, and many models have been developed to help explain the limits on our capacity to hold information actively in mind (e.g., Cowan, 2001; Miyake & Shah, 1999). In the domain of visual working memory, these models have grown particularly sophisticated (Alvarez & Cavanagh, 2004; Bays, Catalao, & Husain, 2009; Cowan, 2001; Luck & Vogel, 1997; Wilken & Ma, 2004; Zhang & Luck, 2008). However, nearly all of these models focus on memory for extremely simple displays of presegmented objects. Furthermore, these models address only average performance across displays and do not make predictions about the difficulty of particular displays.

By contrast to these simple displays, memory for real-world stimuli depends greatly on the background knowledge and principles of perceptual organization our visual system brings to bear on a particular stimulus. For example, when trying to remember real-world scenes, people encode both the gist and detailed information about some specific objects (Hollingworth, 2004). Moreover, they use the gist to guide their choice of which specific objects to remember (Hollingworth & Henderson, 2000), and when later trying to recall the details of the scene, they are influenced by this gist, tending to remember objects that are consistent with the scene but were not in fact present (Lampinen, Copeland, & Neuschatz, 2001). Existing models of the architecture of working memory do not address any of these hierarchical encoding or perceptual grouping factors. For this reason, they are unsatisfying as explanations of what observers will remember about more complex displays in which objects are not randomly chosen, but instead make up a coherent scene.

In this paper we reformulate change detection as rational probabilistic inference in a generative model (similar in spirit to Huber, Shiffrin, Lyle, and Ruys (2001) and Hemmer and Steyvers (2009b)). Rather than modeling the memory process per se, we model how observers encode a scene, and treat change detection as a probabilistic inference that attempts to invert this encoding model. We show that earlier models of visual working memory capacity are special cases of this framework, and show how our model can be extended to include the encoding of gist or higher-order structure. We thus take the first steps toward formalizing working memory capacity for displays in which the items are not all treated independently.

Visual working memory

One of the most popular ways to examine visual working memory capacity has been with a *change detection* task (Luck & Vogel, 1997). In this task, observers are presented with a small number of different colored squares (2, 4, 8, or 16) and told to remember which color appeared in which location. The squares then disappear for a brief period, and when they reappear they either are all the same colors as before, or contain one square which has changed color. Observers must report whether the display is the same or whether one of the squares changed color.

It is generally found that observers accurately detect changes when there are fewer than 3-4 simple colored squares, and as the number of squares increases above 4 observers accuracy steadily decreases (Luck & Vogel, 1997). In order to quantify this decrease and derive a capacity measure for the contents of visual working memory, change detection tasks have been modeled and formalized (Rouder et al., 2008; Wilken & Ma, 2004). For example, in the standard "slot" model of visual working memory (Cowan, 2001; Luck & Vogel, 1997; Rouder et al., 2008), it is assumed that on a display with N items observers perfectly recall the color of K items and completely forget the other N-K items on the display. Using this model, it is possible to convert change detection performance into an estimate of K, and these capacity estimates, termed Cowan's K, are widely reported in the literature on visual working memory (Alvarez & Cavanagh, 2004; Brady, Konkle, & Alvarez, 2009; Cowan, 2001; Luck & Vogel, 1997).

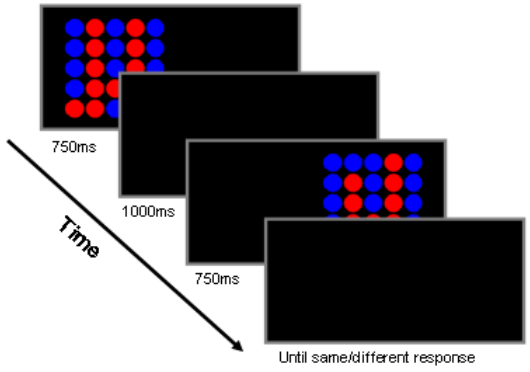


Figure 1: Methods of the change detection task modeled and used in Experiments 1 and 2. Observers are first briefly presented with a display, and then after a 1 sec blank are presented with another display where either the items are exactly the same or one item has changed color. They must say whether the two displays are the same or different.

Aside from Cowan's K, there are other models used to quantify working memory capacity (Bays et al., 2009; Wilken & Ma, 2004; Zhang & Luck, 2008). However, the displays used always consist of simple stimuli like colored dots that are sampled uniformly, limiting any overarching structure or gist. All existing models of change detection thus ignore the presence of higher-order structure and prior knowledge that characterize change detection in real-world scenes (Simons & Rensink, 2005).

A probabilistic model of change detection

We present a probabilistic model of change detection that attempts to bridge the gap between the simple models used to formalize working memory capacity and the complicated phenomena that characterize memory for real-world scenes. Thus, we sought to model change detection in cases where the displays to be remembered were not just random colored dots but also exhibited some higher-order structure. As stimuli we created 5x5 patterns in which each space was filled in by a red or blue circle (or black and white square, see Figure 3). The items could form patterns that were anything from completely random to completely one color or vertical or horizontal lines. Our displays were thus simple relative to real scenes but were complex enough that we expected existing models, which encode dots at random, would fail to predict what people remember about these displays.

Our modeling preceded in two stages, mirroring the two stages of a standard change detection task: view and encode display one, then view display two and decide if a change occurred (See Figure 1).

While the observer is encoding the first display, they have access to the color of all the dots present in the first display. We propose that observers use this information to do two things: first, they infer what "gist" may have given rise to this display; then, using this gist, they select the subset of the dots least well captured by the gist and encode these items specifically into an item memory. The specific dots to encode are selected based on how unlikely they are under the gist. Those that are the biggest outliers (e.g., least well captured by the gist) are encoded into an item memory that specifically encodes their colors.

After a short viewing, the first display disappears and the observer is left with only what they encoded about it in memory. Then, some time later, a second display appears and the observer must decide, based on what they have encoded in memory, whether this display is exactly the same as the first display. Thus, at the time of the second display (detection), the observer has access to the new display and the information in memory. Using the constraint that at most one item will have changed, it is then possible to use Bayesian inference to put a probability on each possible first display, and, using these probabilities, to calculate the likelihood of that the display changed.

Importantly, when the model encodes no higher-order structure it recovers the standard slot-based model of change detection. However, when the displays do have higher-order regularities or 'gist', our model uses this information to both select appropriate individual items to remember and to infer properties of the display that are not specifically encoded.

Encoding

The graphical model representation of the encoding model (shown in Figure 2) specifies how the stimuli are initially encoded into memory. We observe the first image (D^1), and we use this to both infer the higher-order structure that may have generated this image (G) and to choose the specific set of K items to remember from this image (S).

In the model, any given "gist" must specify which displays are probable and which are improbable under that gist. Unfortunately, even in simple displays like ours with only 2 color choices and 25 dots, there are 2^{25} possible displays. This makes creating a set of gists by hand and specifying the likelihood each one gives to each of the 2^{25} displays infeasible. Thus, as a simplifiying asumption we chose to define gists using Markov Random Fields, which allow us to specify a probability distribution over all images by simply defining a small number of parameters about how nodes tend to differ from their immediate neighbors; such models have been used extensively in computer vision (Geman & Geman, 1984). We use only two gist parameters, which specify how often dots are the same or different color than their horizontal neighbors (G_h) and how often dots are the same or different color than their vertical neighbors (G_v). Thus, one particular gist ($G_h = 1, G_v = -1$) might specify that horizontal neighbors tend to be alike but vertical neighbors tend to differ (e.g., the display looks like it has horizontal stripes in it). This gist would give high likelihood to displays that have many similar horizontal neighbors and few similar vertical neighbors.

We treat each dot in these change detection displays as a random variable D_i^1, where the set of possible values of each

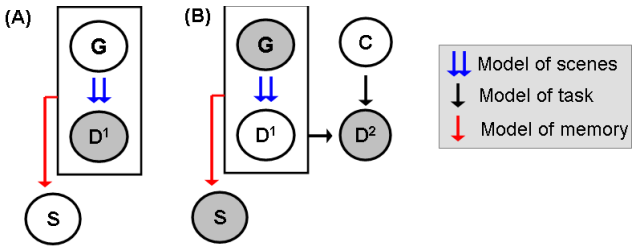


Figure 2: The model expressed in graphical model notation for (A) encoding and (B) detection. Filled circles indicate a node is observed (the model has access to it). Unfilled circles indicate the model must infer the value of the node. The arrows are colored based on what kind of process they represent. D1=the first display, D2=the second display, G=the gist S=specific items, C=the presence of a change.

D_i^1 is -1 (color 1) or 1 (color 2). To define the distribution over possible displays given the gist parameters, $P(D|G)$, we assume that the color of each dot is independent of the color of all other dots when conditioned on its immediate horizontal and vertical neighbors.

We thus have two different kind of neighborhood relations (clique potentials) in our model. One two parameters (G_h and G_v) apply only to cliques of horizontal and vertical neighbors in the lattice (N_h and N_v) respectively. Thus, $P(D^1|G)$ is defined as:

$$P(D^1|G) = \frac{exp\left(-En(D^1|G)\right)}{Z(G)} \quad (1)$$

$$En(D^1|G) = G_v \sum_{(i,j)\in N_v} \psi(D_i^1, D_j^1) + G_h \sum_{(i,j)\in N_h} \psi(D_i^1, D_j^1)$$

where the partition function:

$$Z(G) = \sum_{D^1} exp\left(-E(D^1|G)\right)$$

normalizes the distribution. $\psi(D_i^1, D_j^1)$ is 1 if $D_i^1 = D_j^1$ and -1 otherwise. If $G > 0$ the distribution will favor displays where neighbors tend to be similar colors, and if $G < 0$ the distribution will favor displays where neighbors tend to be different colors.

The "gist" of the display is therefore represented by the parameters G of an MRF defined over the display. Our definition of $p(D^1|G))$ thus defines the probability distribution $p(display|gist)$. To complete the encoding model we also need to define $p(items|display, gist)$ ($p(S|D^1, G)$). To do so, we define a probability distribution that preferentially encodes outlier objects (objects that do not fit well with the gist).

We choose whether to remember each object from the display by looking independently at the conditional probability of that object under the gist, assuming all of its neighbors are fixed $p(D_i^1|G, D_{/i}^1)$. S denotes the set of K specific objects encoded: $S = s_1, ..., s_k$. To choose S, we rank all possible sets of objects of size 0, 1, 2, ... to K objects based on how unlikely they are under the encoded gist. Thus, the probability of encoding a set of objects (S) is:

$$p(S|G, D^1) = \prod_{j:s_j\in S} [1 - p(D_j^1|G, D_{/j}^1)] \prod_{j:s_j\notin S} p(D_j^1|G, D_{/j}^1) \quad (2)$$

This defines $p(S|D^1, G)$, which provides the probability of encoding a particular set of specific items in a given display, $p(items|display, gist)$, in our model.

To compute the model predictions we use exact inference. However, due to the computational difficulty of inferring the entire posterior distribution on MRF parameters for a given display (e.g., the difficulty of computing $Z(G)$), and because we do not wish to reduce our gist to a single point estimate, we do not compute either the maximum posterior MRF parameters for a given display or the full posterior on G. Instead, we store the posterior in a grid of values for G in both horizontal and vertical directions ($G_h = -1.5, -1, -.5, 0, .5, 1, 1.5, G_v = -1.5, -1, -.5, 0, .5, 1, 1.5$). We compute the likelihood of the display under each of these combinations of G_h and G_v and then choose the items to store (S) by integrating over the different choices of G (we store the full posterior over S)). We choose a uniform prior on the gist (e.g., a uniform prior on MRF parameters G).

In summary, to encode a display we first treat the display as an MRF. We then calculate the posterior on possible gists by calculating a posterior on G at various (pre-specified) values of G. We then use this G and the original display to compute a posterior on which set of $\leq K$ items to encode into item memory (S). At the completion of encoding we have both a distribution on gists (G) and a distribution on items to remember (S), and these are the values we maintain in memory for the detection stage.

Detection

At the detection stage, we need to infer the probability of a change to the display. To do so, we attempt to recover the first display using only the information we have in memory and the information available in the second display. Thus, using the probabilistic model, we work backwards through the encoding process, so that, for example, all the possible first displays that don't match the specific items we remembered are ruled out because we would not have encoded a dot as red if it were in fact blue.

More generally, to do this inference we must specify $P(D^1|S)$, $P(D^1|D^2)$, $P(D^1|X)$, $P(S|G, D^1)$. Almost all of these probabilities are calculated by simply inverting the model we use for encoding the display into memory initially with a uniform prior on possible first displays. Thus, $P(D^1|G)$ is given by Equation 1, and $P(S|G, D^1)$ is given by Equation 2.

Those probabilities not specified in the forward model represent aspects of the change detection task. Thus, $P(D^1|S)$ is a uniform distribution over first displays that are consistent

with the items in memory and 0 for displays where one of those items differs. This represents our simplifying assumption (common to standard "slot" models of visual working memory) that items in memory are stored without noise and are never forgotten (it is possible to add noise to these memory representations by making $P(D^1|S)$ a multinomial distribution over possible values of each item, but for simplicity we do not model such noise here). $P(D^1|D^2)$ is uniform distribution over all displays D^1 such that either $D^1 = D^2$ or at most one dot differs between D^1 and D^2. This represents the simple fact that the task instructions indicate at most one dot will change color.

Together these distributions specify the probability of a particular first display given the information we have about the second display and information we have in memory, $P(D^1|G,S,D^2)$. Given the one-to-one correspondence between first displays and possible changes, we can convert this distribution over first displays to a distribution over possible changes. Our prior on whether or not there is a change is 0.5, such that 50% of the mass is assigned to the "no change" display and the other 50% is split among all possible single changes. Thus:

$$P(C|G,S,D^2) = \frac{0.5P(D^1 = D^2|G,S,D^2)}{0.5P(D^1 = D^2|G,S,D^2) + 0.5\sum P(D^1 \neq D^2|G,S,D^2)}$$

This fully specifies the model of change detection.

Experiment 1 and 2

To examine human memory performance, we collected data using Amazon Mechanical Turk, where we had observers perform a change detection task for each of 24 different displays. We then compared this performance to our model.

The model makes predictions about how hard it is to detect changes in particular displays of colored dots (i.e., some changes will be more difficult to detect than others). In addition, it makes predictions about overall accuracy for a particular set of displays. We can thus examine how well the model fits with human memory performance in two distinct ways: (1) How many particular items (K) the model needs to recall to match human performance overall, and (2) how well the model's predictions about the difficulty of particular displays correlate with human memory performance.

Method

We sampled a set of 16 displays from the Markov Random Field model we use to define our gist (using Gibbs Sampling). Four of these displays were sampled from each of $G_h = \pm 1$, $G_v = \pm 1$. In addition, we generated 8 displays randomly. In Experiment 1, these 24 displays consisted of red and blue dots. In Experiment 2 they were exactly the same displays, but composed of black and white squares instead.

The displays were presented to 65 participants in Exp. 1 and a separate set of 65 participants in Exp. 2 using Amazon Mechanical Turk. The first display was flashed up for 750 ms (timing was controlled using Javascript), followed by a 750ms blank period; then the second display was flashed up for 750ms in a different screen location. Observers' task was simply to say whether the two flashed displays were the same or different (See Figure 1). Each display was presented to each observer in both a "same" and "different" trial, so observers completed 48 trials each, with the entire experiment lasting approximately 4 minutes. The order of the 48 trials was randomly shuffled for each subject. Observers were paid 30 cents for their time.

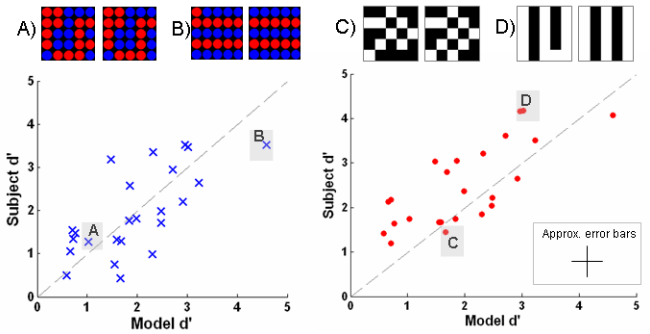


Figure 3: The fit of our probabilistic model to the observers' data with K=4 in the model for Experiments 1 and 2. Each point is the d' for a pair of displays. Approximate error bars are shown for both the subjects and model, calculated by bootstrapping standard errors. Example of both a hard and easy pairs of displays is shown for each experiment.

Results

For each display we computed a d', measuring how difficult it was to detect the change in that display (averaged across observers). Performance in Experiments 1 and 2 was highly similar, as the correlation in the display-by-display d' was r=0.91 between the two experiments. Thus performance was collapsed across both experiments for the remaining analyses.

On average, human observers d' was 2.18 (S.E. 0.06) suggesting they were quite good at detecting changes on these displays. Since the displays contain 25 dots, this d' corresponds to a Cowan's K of nearly 16.1 dots if the items are assumed to be represented independently and with no summary information encoded (Cowan, 2001). This is nearly 5 times the number usually found in simpler displays and thus represents a challenge to standard models of change detection and visual working memory capacity.

Importantly, our claim is not that observers remember 16 individual dots. Instead, our model provides an alternative explanation. The model achieves the same performance as people (d'=2.18) with a K value of only 4, thus encoding only four specific dots in addition to the display's gist (model d'=1.2, 1.8, 2.05, 2.25 at K=1, 2, 3, 4). This is because the model does not represent each dot independently: instead, it represents both higher-order information as well as information about specific dots. The model thus aligns nicely with both previous work from visual working memory suggesting

a capacity of 3-4 simple items (Luck & Vogel, 1997; Cowan, 2001) and also with data from the literature on real-world scenes which suggests a hierarchical representation with both gist and item information (e.g. Lampinen et al., 2001).

In addition to describing overall memory capacity, we can also examine the difficulty of particular displays. Previous models of change detection treat all displays as interchangeable, since they choose which objects to encode at random and do not represent any summary information about the display. They thus make no predictions about which particular changes will be hard or easy to detect. However, observers reliably find it more difficult to detect change in some displays than others, as measured both by averaging 200 split-half correlations on d-prime (r=0.75) and by bootstrapping standard errors on observers' d-prime (see Figure 3).

Our model does not treat each item independently, and chooses which items to encode by making strategic decisions based on the display's gist. Thus, our model does make predictions about the difficulty of detecting particular changes. In fact, the correlation between the model's difficulty with individual displays and the human performance on these displays was quite high (overall: r=0.71, $p<0.0001$; Exp.1: r=0.65, Exp.2: r=0.73; See Figure 3). Thus, the model's simple gist representation captures which changes people are likely to detect and which they are likely to miss.

Discussion

We here present a formal model of change detection which relies upon probabilistic inference to make predictions about visual working memory. The model takes into account the hierarchical nature of memory typically found in real-world scenes. It successfully predicts the display-by-display difficulty of visual working memory displays, indicating which changes observers will find easy to detect and which they will find difficult. The model also converges with the standard visual working memory literature on an estimate of 3-4 individual objects remembered, even in more complex patterned displays.

Importantly, the model recovers previous models of visual working memory capacity as a special case, and thus captures the properties of those models in displays with no higher-order information. However, by formulating change detection in terms of probabilistic inference, we can make much richer models of working memory than those typically used to calculate capacity in visual working memory experiments.

Non-independence in Visual Working Memory

While almost all experiments on visual working memory treat the items to be remembered as indepedent units, there are several exceptions (e.g., Jiang, Olson, & Chun, 2000; Sanocki & Sulman, 2008; Jiang, Chun, & Olson, 2004; Vidal, Gauchou, Tallon-Baudry, & Oregan, 2005). The most prominant exception to this assumption of independence is the work of Jiang et al. (2000), who suggested that the spatial context of other items is important to simple change detection tasks. On displays where the item that changed is presented in the context of the other items present at encoding, observers perform better at change detection (Jiang et al., 2000). This suggests the items are not represented independently of their spatial context. This is compatible with the encoding of both summary information and specific items used in our probabilistic model.

In addition, previous work by Brady et al. (2009); Brady and Alvarez (2010) demonstrates that observers can be induced to encode displays with colored dots using statistical regularities present between the dots, rather than treating each dot separately. Observers not only use information about co-occurence between items to form more compressed representations of these displays (Brady et al., 2009) but also encode the displays at multiple levels of abstraction, combining both an overall summary of the display and information about particular dots (Brady & Alvarez, 2010).

More broadly, the idea that memory encoding and retrieval are based on information represented at multiple levels of abstraction is common in the literature on reconstructive memory (Bartlett, 1932). Recent computational models similar in spirit to the one presented here have formalized this in both the domains of object size memory (Hemmer & Steyvers, 2009b) and more recently in the combination of gist and specific objects in real-world scenes (Hemmer & Steyvers, 2009a).

Chunking, Perceptual Grouping and Gist

One of the most popular explanations for observers' better-than-expected performance with more complex stimuli is chunking, or forming larger units out of smaller subsets of the stimuli (Miller, 1956; Cowan, 2001). In this framework, performance on our displays of patterned dots could be a result of observers' remembering only 3-4 independent items from the display and not encoding any overarching gist or structure. Instead, the items they remember would simply consist of multiple dots grouped into single items. This explanation has been proposed, for example, to explain why observers are better than expected at empty-cell localization tasks using patterned stimuli much like ours (Hollingworth, Hyun, & Zhang, 2005) and why some displays are remembered more easily than others in same/different tasks (Howe & Jung, 1986).

This kind of chunking could potentially explain observers' performance on our displays. However, our preliminary work with a model that partitions the display into contiguous regions of the same color and remembers K of these regions suggests that such a model does not adequately explain performance in the current experiments. Instead, such a model either fails to capture the pattern of human errors or requires memory for an overly large number of regions ($K>5$) to achieve human levels of performance. However, future work is needed to examine models that perform such grouping or chunking and compare them with models, like ours, that represent the displays at multiple levels of abstraction.

Model of Gist

In the model the "gist" is encoded using Markov Random Fields, and thus the only information that can be represented are local spatial continuity properties of the colors in the display (similiarity between horizontal and vertical neighbors). Obviously, this is too impoverished to be a fully accurate model of human visual memory, even for such simple dot displays. For example, we could draw letters or shapes in the dot patterns, and people would recall those patterns well by summarizing them with a gist-like representation. Our model cannot capture such representations. However, we believe that our model nonetheless represents a step forward in understanding how people make use of such gist during change detection.

References

Alvarez, G., & Cavanagh, P. (2004). The capacity of visual short-term memory is set both by visual information load and by number of objects. *Psychological Science*, *15*(2), 106–111.

Baddeley, A. (2000). The episodic buffer: a new component of working memory? *Trends in Cognitive Sciences*, *4*(11), 417–423.

Bartlett, F. (1932). *Remembering: A study in experimental and social psychology*. New York: Macmillan.

Bays, P., Catalao, R., & Husain, M. (2009). The precision of visual working memory is set by allocation of a shared resource. *Journal of Vision*, *9*(10), 7.

Brady, T. F., & Alvarez, G. A. (2010). Ensemble statistics of a display influence the representation of items in visual working memory. *Visual Cognition*, *18*(1), 114–118.

Brady, T. F., Konkle, T., & Alvarez, G. A. (2009). Compression in visual working memory: using statistical regularities to form more efficient memory representations. *Journal of Experimental Psychology: General*, *138*(4), 487–502.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*(1), 87–114.

Daneman, M., & Carpenter, P. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning & Verbal Behavior. Vol*, *19*(4), 450–466.

Geman, S., & Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Trans. Pattern Anal. Machine Intel.*, *6*, 721–741.

Hemmer, P., & Steyvers, M. (2009a). Integrating Episodic and Semantic Information in Memory for Natural Scenes. In N. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31th annual conference of the cognitive science society* (pp. 1557–1562). Austin, TX: Cognitive Science Society.

Hemmer, P., & Steyvers, M. (2009b). Integrating episodic memories and prior knowledge at multiple levels of abstraction. *Psychonomic Bulletin & Review*, *16*(1), 80.

Hollingworth, A. (2004). Constructing visual representations of natural scenes: The roles of short-and long-term visual memory. *Journal of Experimental Psychology: Human Perception and Performance*, *30*(3), 519–537.

Hollingworth, A., & Henderson, J. (2000). Semantic informativeness mediates the detection of changes in natural scenes. *Visual Cognition*, *7*(1), 213–235.

Hollingworth, A., Hyun, J., & Zhang, W. (2005). The role of visual short-term memory in empty cell localization. *Perception and Psychophysics*, *67*(8), 1332–1343.

Howe, E., & Jung, K. (1986). Immediate memory span for two-dimensional spatial arrays: Effects of pattern symmetry and goodness. *Acta psychologica*, *61*(1), 37–51.

Huber, D., Shiffrin, R., Lyle, K., & Ruys, K. (2001). Perception and preference in short-term word priming. *Psychological Review*, *108*(1), 149–182.

Jiang, Y., Chun, M., & Olson, I. (2004). Perceptual grouping in change detection. *Perception and Psychophysics*, *66*, 446–453.

Jiang, Y., Olson, I., & Chun, M. (2000). Organization of visual short-term memory. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *26*(3), 683–702.

Lampinen, J., Copeland, S., & Neuschatz, J. (2001). Recollections of things schematic: Room schemas revisited. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *27*(5), 1211–1222.

Luck, S., & Vogel, E. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*(6657), 279–280.

Miller, G. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, *63*(2), 81–97.

Miyake, A., & Shah, P. (1999). *Models of working memory: Mechanisms of active maintenance and executive control*. Cambridge University Press.

Rouder, J., Morey, R., Cowan, N., Zwilling, C., Morey, C., & Pratte, M. (2008). An assessment of fixed-capacity models of visual working memory. *Proceedings of the National Academy of Sciences*, *105*(16), 5975.

Sanocki, T., & Sulman, N. (2008). Visual short term memory for location: Does objecthood matter? *Journal of Vision*, *8*(6), 203-203.

Simons, D., & Rensink, R. (2005). Change blindness: Past, present, and future. *Trends in Cognitive Sciences*, *9*(1), 16–20.

Vidal, J., Gauchou, H., Tallon-Baudry, C., & Oregan, J. (2005). Relational information in visual short-term memory: The structural gist. *Journal of Vision*, *5*(3), 244–256.

Wilken, P., & Ma, W. (2004). A detection theory account of change detection. *Journal of Vision*, *4*(12), 1120–1135.

Zhang, W., & Luck, S. (2008). Discrete fixed-resolution representations in visual working memory. *Nature*, *453*(7192), 233–235.

Expertise in Visual Art is Associated with Altered Perceptual Strategies Within and Across Domains: Evidence from Eye Tracking

Kuba J. Glazek (kglazek@temple.edu)
Temple University Department of Psychology, 1701 N. 13th Street
Philadelphia, PA 19122 USA

Robert W. Weisberg (robert.weisberg@temple.edu)
Temple University Department of Psychology, 1701 N. 13th Street
Philadelphia, PA 19122 USA

Abstract

Eye-movement research on expert visual artists suggests that experts in this particular domain differ from novices in their strategies for encoding to-be-rendered stimuli. However, it remains unclear if such differences are specific to the domain of expertise or independent of it (i.e., if the different strategies are utilized only in relation to perception with goals specific to rendering, or if they generalize to visual perception of any stimulus with perceptual goals other than rendering). Experiment 1 examined eye-movement strategies utilized by experts and novices when rendering familiar and novel stimuli. Experiment 2 examined performance in a recognition task that also utilized novel stimuli. Results suggest that experts possess both domain-specific *and* domain-independent advantages, in that they have more efficient visual encoding abilities both when rendering and not. The results of a concurrent analysis suggest a link between the encoding advantage and schizotypy, which is correlated with creative advantage, as well as with a neural profile of left hypofrontality. Implications for a two-stage model of creativity are discussed.

Keywords: Expertise; far transfer; schizotypy; visual art; creativity.

Introduction

Only in the presence of a meaningful configuration of stimulus features do experts in various domains, including chess (Chase & Simon, 1973), cars (Curby, Glazek, & Gauthier, 2009), and digit strings (Ericsson & Kintsch, 1995) outperform novices in terms of recall performance. Theoretically, long-term memory plays a role in such a domain-specific expert advantage (Ericsson & Delaney, 1999): Repeated practice yields a hierarchically organized memory structure for a class of stimuli, into which a stimulus representation can easily be encoded, provided that the stimulus generally fits the pattern with which an expert is familiar. Furthermore, as expertise increases, so does the number of features, or chunks, that the structure can accommodate. Such a structure would, *ipso facto*, not exist for a novel stimulus. Based on this account of expertise development, expert visual artists (henceforth *experts*) should perform as poorly as control participants (henceforth *novices*) when rendering novel stimuli and when performing perceptual tasks independent of rendering.

However, there is evidence that expertise unique to the domain of visual art confers an advantage that transfers outside of what is familiar, e.g., mathematics performance in elementary school (Luftig, 1994), math and verbal SAT scores (Vaughn & Winner, 2000), visual analysis of out-of-focus pictures and novel stimuli, and mental rotation of three-dimensional objects (Kozbelt, 2001). Expertise in visual art may transcend a rendering-specific advantage, as creating drawings or paintings from life (henceforth *renderings*) requires visual analysis of objects in one's environment. Such visual analysis may generalize to visual perception in general (i.e., to situations where there is no rendering requirement, just streams of visual stimulation). In addition to examining rendering performance, the current experiments are designed to shed light on how experts and novices process novel stimuli under the perceptual goal of recognition.

A potential mechanism underlying an expert advantage in encoding visual information is also examined. Divergent thinking is considered a mechanism central to creativity (e.g., Burch, Pavelis, Hemsley, & Corr, 2006; Mednick, 1962; Miller & Tal, 2007; Schuldberg, 2000-2001; but see Weisberg, 2006, for a different view). It benefits from access to multiple associates (i.e., thoughts, ideas, etc. that come from memory and/or the environment) as starting points for creative synthesis; the more qualitatively-different associates a person has access to, the more likely it is that she will find a meaningful, novel combination in them (insight), then create a tangible product (elaboration).

In normal participants, environmental sources of stimulation outside of a point of focus are attenuated or blocked from consciousness (e.g., Lubow & Gewirtz, 1995). However, such blocking has been shown to be detrimental to creativity; individuals who were less likely to block out a task-irrelevant stream of stimuli were more likely to be creative, as measured by lifetime creative achievement (Carson, Peterson, & Higgins, 2003). If experts encode visual stimuli more rapidly than novices, they might then have access to more of them as associates in working memory, thus potentially boosting their divergent thinking capacity.

Individuals with schizotypic personality disorder (SPD), an attenuated form of schizophrenia, are also more likely to not attenuate irrelevant streams of stimulation (Baruch, Hemsley, & Gray, 1988). This population has a particular pattern of cortical activity: Left hypofrontality, whereby left

prefrontal cortical (PFC) function is attenuated (Buchsbaum et al., 1997; Raine et al., 2002). Left PFC activity is associated with two types of processing pertinent to the current study. First, left PFC function has been shown to play a role in translating modality-specific information into abstracted information (e.g., Anderson, Qin, Yung, & Carter, 2007). Normal left PFC function is associated with a lack of accuracy in rendering, and accurate rendering emerges when left PFC function is suppressed using transcranial magnetic stimulation (Snyder et al., 2003). These findings strongly suggest that left hypofrontality plays a major role in accurate rendering, potentiating it via a lack of interference in the signaling among sensory pathways and motor control centers. Second, increased bilateral PFC function is associated with creativity. Jung et al. (2010) found a negative correlation between lifetime creative achievement and left prefrontal cortical thickness. Carlsson, Wendt, and Risberg (2000) found that highly creative participants (as judged by the Creative Functioning Test) utilized right PFC to a significantly larger extent than low-creative participants, who utilized only left PFC, when coming up with alternate uses for a brick. The responses in that study were not rated as varying in creativity between the groups, implying that creative individuals utilize the right hemisphere to a greater extent than non-creative individuals in any kind of task.

Elevated levels of SPD in experts would be consistent with a pattern of left hypofrontality, which may underlie both rendering and creative abilities.

Experiment 1: Domain-Specific Performance

Several inferences can be made regarding cognitive processing on the basis of tracking eye movements. Theeuwes, Olivers, and Chizk (2005) showed that maintenance of the spatial location of an item in working memory only (i.e., without its presence in the field of vision) causes saccades (i.e., eye movement trajectories) to deviate in the direction of the maintained item. Tremblay, Saint-Aubin, and Jalbert (2006) showed that participants' use of eye movements as overt rehearsal was not only a default strategy used to maintain spatial position and serial order of dots presented on a computer screen, but also that denying subjects use of such a strategy caused a significant decrease in accuracy of recall of order of presentation. Under unconstrained conditions, experts reference (i.e., move their eyes from paper to stimulus during rendering) to-be-rendered stimuli significantly more frequently than novices (Cohen, 2005; Tchalenko, 2009). Experimentally manipulating the refresh rate (i.e., alternately illuminating either the stimulus or drawing pad every 1, 5, or 15 s) affected blindly-judged accuracy of experts' renderings; lower refresh rates (i.e., stimulus visible only every 15 s) yielded significantly less accurate renderings (Cohen, 2005). The manipulation had no effect on novices' accuracy. However, these results apply only to relatively complex stimuli, including faces (Cohen, 2005) and standing nudes (Tchalenko, 2009). For rendering straight and curved individual lines and squares, there do not appear to be differences in eye movement patterns between experts and novices under unconstrained conditions (Tchalenko, 2007).

Therefore, in addition to the dimension of novelty, the content of a stimulus can be operationalized along a continuum of complexity. This experiment is the first to explicitly manipulate novelty and complexity in a rendering task and record the effect on eye movement strategies of experts and novices. If experts possess a visual encoding advantage, they should encode familiar and novel stimuli by utilizing the same cognitive strategy (as evidenced by similar eye movement patterns), while novices should utilize distinct strategies for encoding familiar and novel stimuli of varying complexity.

Method

Stimuli and Apparatus Stimuli rated as most familiar (Snodgrass & Vanderwert, 1980), and ones that are entirely novel (Chinese ideograms) were used. Within each of these categories, complexity was manipulated. Stimuli rated as most familiar were sorted according to rated complexity and the 10 simplest and 10 most complex were selected. The 10 familiar simple stimuli had a mean complexity rating of 1.60 out of 5 (SD = 0.25), and the 10 familiar complex stimuli had a mean complexity rating of 3.78 out of 5 (SD = 0.31), a significant difference ($p < 0.001$).

Unique Chinese ideograms were selected as novel stimuli, and their features (i.e., number of line segments) counted. The set of 10 novel simple stimuli had a mean of 5.50 features (SD = 0.51), and the set of 10 novel complex stimuli had a mean of 13.85 features (SD = 1.66), a significant difference ($p < 0.001$).

Thus, complexity was explicitly controlled for both novel and familiar stimuli in order to examine its effect on eye movement behavior.

Stimuli were presented using E-Prime software, version 2.0 on a Tobii 1750 eye tracker (Psychology Software Tools, Pittsburgh, PA) set to 1024 x 768 pixels screen resolution, sampling eye position at 50 Hz and with a screen refresh rate of 50 Hz. The eye tracker was calibrated at the outset of each session prior to data collection to ensure reliable eye tracking. Participants rendered using a stylus on the screen of a tablet PC running CogSketch software, version 1.131 with a simplified graphic user interface (SILC, Chicago, IL).

Participants Novices (n = 8, mean age = 19.9 years, three males) were recruited from Temple University's undergraduate subject pool, and given the option of course credit or cash for their participation. Experts (n = 8, mean age = 30.1 years, two males) were recruited using flyers posted around the Philadelphia community, and had to meet the following criteria: Have at least five years of formal art training, be at least 18 years old, and must draw or paint more than once a week, all this information being gathered via e-mail or telephone interviews prior to participation. Experts were compensated with cash.

All novices were screened for art expertise following their experimental session. All participants were screened for proficiency in reading, writing, and speaking Chinese[1].

Procedure Factor 1 (between-subjects) was expertise (novice or expert participant). Factor 2 (within-subjects) was stimulus familiarity (familiar or novel), which pertains to the presence or absence of long term representations: No subjects possess long-term representations of novel stimuli, and all subjects possess representations of familiar stimuli. Factor 3 (within-subjects) was stimulus complexity (simple or complex stimulus). Participants were informed that they had 60 s to render each of the 40 stimuli as accurately as possible, with the opportunity to rest between trials. If a participant finished rendering before 60 s elapsed, she pressed the space bar on the keyboard in front of the monitor. If she did not finish, the stimulus disappeared once 60 s elapsed. Following a practice trial to provide familiarization with the drawing stylus and tablet, all participants rendered all 40 stimuli in randomized order. The dependent variables were percentage of time spent per trial with eyes on the on-screen image (i.e., visual encoding of a stimulus), and mean duration of eyes on the on-screen stimulus. An on-screen epoch was operationalized as 60 consecutive ms or more of the eyes looking at the rectangular area subsuming a stimulus.

Results

An examination of eye movements to and away from to-be-rendered stimuli during rendering yielded a significant main effect of expertise ($F(1, 14) = 6.43$, $p < .05$), with novices' total encoding time significantly longer than experts' (see Figure 1A). A significant main effect of stimulus complexity was evidenced, as well ($F(1, 14) = 46.08$, $p < .001$; see Figure 1B). There was also a significant interaction between stimulus complexity and stimulus novelty ($F(1, 14) = 6.96$, $p < .05$), which resulted from a significant difference between familiar complex and novel complex stimuli ($t(15) = 2.47$, $p < .05$), and a lack thereof between familiar simple and novel simple stimuli (see Figure 1B).

In order to examine the above effects in more detail, individual epoch durations were analyzed. There was a significant main effect of expertise ($F(1, 14) = 7.79$, $p < .05$); experts' epochs were significantly shorter than novices' (see Figure 1C). As with overall encoding time, there was a significant main effect of stimulus complexity ($F(1, 14) = 56.85$, $p < .001$), with longer epochs for complex stimuli (see Figure 1C). There was also a main effect of stimulus novelty ($F(1, 14) = 79.29$, $p < .001$), with shorter epochs for novel stimuli (see Figure 1D). Of central importance were three significant interactions. The first of these was complexity by expertise ($F(1, 14) = 4.67$, $p < .05$; see Figure 1C); the second was novelty by expertise ($F(1, 14) = 21.16$, $p < .001$ see Figure 1D), and finally, complexity by novelty

[1] One participant fluent in Chinese, excluded from analyses, evidenced patterns very similar to the expert group.

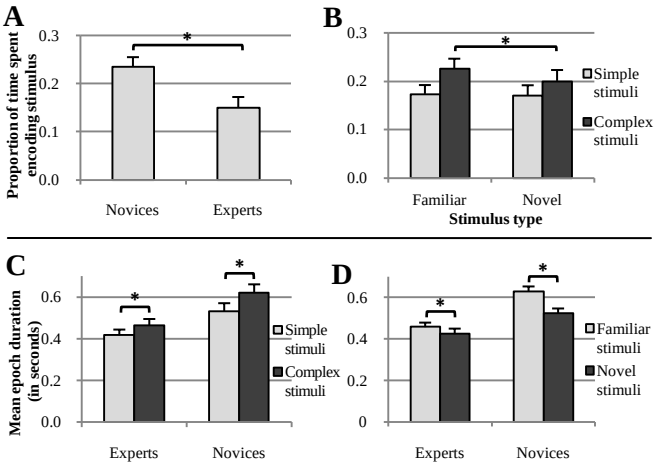


Figure 1: Effects of A) expertise, and B) stimulus type on total encoding time; C) stimulus complexity, and D) familiarity on epoch duration. Error bars represent one standard error.

by expertise ($F(1, 14) = 7.08$, $p < .05$). Essentially, as stimulus complexity and novelty changed, so did the novices' encoding strategy, which was also the case for experts, albeit to a significantly lesser extent (see Figures 1C and 1D).

Discussion

The results support the hypothesis that stimulus novelty and complexity have differential effects on processing strategy. The significant three-way interaction is of most interest, in that the effects of stimulus complexity and novelty on encoding strategy were different for experts and novices. As can be seen in Figure 1 C and D, both experts' and novices' encoding strategies were affected by stimulus complexity and novelty in a similar fashion. However, the experts evidenced an attenuation of the differences caused by novel and complex stimuli. These results suggest that visual art training is associated with an advantage in encoding novel and complex stimuli when the goal of perception is rendering.

The absence of long-term representations affected experts less than novices, suggesting that experts use less top-down processing (associated with PFC), or use it more efficiently, than novices when rendering.

The results suggest that experts approach equal efficiency at encoding familiar and novel visual stimuli regardless of complexity when visual encoding is linked to the goal of domain-relevant action. In fact, these results indicate that experts encode novel stimuli as though they are familiar, at least when compared to novices.

Experiment 2: Domain-Independent Performance

Experiment 1 showed that novices require longer epochs than experts in order to effectively encode novel and complex stimuli when rendering. The experts' advantage

may or may not disappear if the domain-specific task of rendering is absent.

In order to examine whether this expert advantage can be observed in a task that does not entail an expertise-based motor component, in Experiment 2 the domain-specific requirement of rendering was removed from the perceptual task, and replaced by a binary stimulus recognition task. It was hypothesized that experts require less encoding time than novices to correctly identify a stimulus as being the same as or different from a briefly-encoded novel stimulus. However, there may be a complexity-based limit on this encoding advantage; thus, the advantage was predicted to be more pronounced for simple novel stimuli than for complex novel stimuli.

Method

Stimuli and Apparatus Eighty Chinese ideograms were used as novel stimuli. Forty were simple and 40 complex. The stimuli used in this experiment were unique (i.e., none overlapped with the stimuli used in Experiment 1). Stimuli were presented on the same computer monitor used in Experiment 1. Eye movements were not recorded in this experiment.

Participants The same participants that took part in Experiment 1 took part in Experiment 2. The order of experiments was counterbalanced across participants.

Procedure Experiment 2 consisted of a binary judgment recognition task, as follows. At the outset of each trial, explicit written instructions appeared on the computer screen for the participant to keep her eyes focused on the screen so as to avoid missing the briefly-presented stimulus, which appeared upon her pressing the space bar. The stimulus was on-screen for a variable amount of time (50, 125, 200, or 275 ms)[2], randomly selected by the computer. Then, following a 1500 ms interval, a second stimulus appeared, which was either the same ideogram, or the same ideogram with one of four slashes superimposed over it. The first ideogram may have had a superimposed slash, as well. The presence of a slash was randomly selected by the computer. This randomization yielded relatively equal numbers of trials for same and different conditions, as well as for all four encoding durations. Participants responded as to whether the second ideogram was the same as or different from the first ideogram by pressing "F" or "J" on the keyboard, respectively (the keys' meanings were displayed on-screen), with explicit instructions to use one index finger for each key. There were four practice trials, followed by eight sets of 10 trials each, with a prompt to take a rest between each set. Sets of trials alternated between simple and complex ideograms. The dependent variable was the proportion of correct answers (same or different) for each encoding duration.

[2] Pilot data obtained from a sample of novices (n = 23) indicated that these intervals should yield meaningful variation.

Results

The results of Experiment 2 are summarized in Figure 2. Non-parametric tests were used due to violations of normality and homogeneity of variance in some of the distributions. There was a significant difference for complex stimulus recognition between experts and novices at 125 ms ($U = 46.5$, $p < .01$). Likewise, for simple stimuli, there was a marginally significant difference between experts and novices at 200 ms ($U = 67.0$, $p = .06$). Furthermore, experts attained above-chance performance for all but the shortest encoding duration, whereas novices did not.

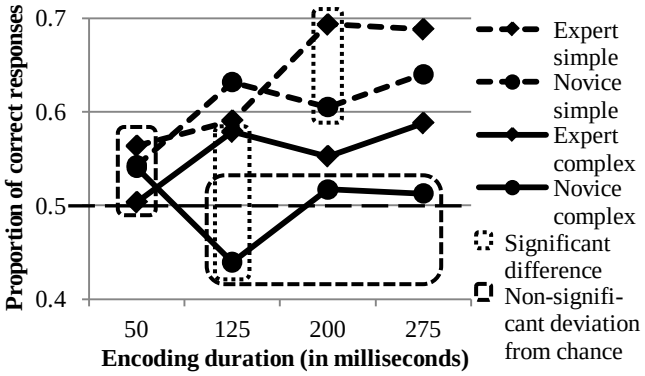


Figure 2: Rates of correct recognition of novel stimuli. 0.5 indicates chance performance.

Discussion

The results of Experiment 2 extended the results of Experiment 1, in that experts significantly outperformed novices at shorter encoding durations. The fact that a difference emerged at 200 ms for simple stimuli indicates that experts' advantage is somewhat limited; experts and novices were equally poor at encoding simple novel stimuli at short durations. Nevertheless, experts were able to encode simple novel stimuli significantly better than novices when given 200 ms or more, whereas novices required at least 275 ms to close the gap. Furthermore, when encoding complex novel stimuli, experts at least evidenced the ability to deviate from chance performance, whereas novices did not. This indicates that, although the experts' advantage appears to be limited, it does confer an advantage when encoding dense, unfamiliar patterns.

Clearly, experts' visual encoding advantage is not limited to rendering, insofar as in Experiment 2, experts were denied any synergistic boost from perceiving with the goal of rendering. Even with a lack of the rendering component, experts' encoding was superior, as evidenced by their higher performance on the recognition task.

Relation of Expertise to Schizotypy: Implications for the Neuroscience of Creativity

The question remains: What cognitive operations can experts perform while novices are still encoding?

Consistent with left hypofrontality in experts, novices encode visual information and abstract it, while experts use the same time to encode the same information and either plan motor commands (Experiment 1), or perform other cognitive tasks (Experiment 2), potentially including divergent thinking. In order to lend support to the theory that left hypofrontality underlies experts' more efficient encoding, self-report data on SPD were obtained, with the hypothesis that experts would evidence higher levels of SPD, an indirect measure of left hypofrontality.

Method

Stimuli, Apparatus, and Participants The schizotypal personality questionnaire, form B (SPQ-B; Raine & Benishay, 1995) was administered to assess schizotypic traits in the expert and novice samples. The SPQ-B is a reliable, 22-item binary judgment questionnaire that assesses three factors: Cognitive-perceptual aberrations, (ideas of reference, magical thinking, unusual perceptual experiences, and paranoid ideation), interpersonal dysfunction (social anxiety, lack of close friends, blunted affect, and paranoid ideation), and disorganization (odd behavior and odd speech). It was presented on the same computer as used in Experiments 1 and 2. Twenty eight novices and 18 experts filled out the questionnaire as part of ongoing investigations.

Results

For the disorganized factor, experts responded affirmatively to significantly more questions than novices ($t = 2.46$, $p < .05$). For questions that load onto the cognitive-perceptual factor, experts responded affirmatively marginally significantly more than novices ($t = 1.86$, $p = .09$). There was no difference between the groups on the interpersonal factor.

Discussion

Experts evidenced a pattern of elevated SPD relative to novices. There was no difference between experts and novices on the interpersonal factor, but there was a significant difference found on the disorganized factor, and a marginally significant difference on the cognitive-perceptual factor. These data provide a potential mechanism for experts' ability to render accurately, despite requiring less time to encode visual information. Hypoactive left PFC does not over-abstract stimulus representations (i.e., its functioning is attenuated) in expert cognition, allowing experts to perform well at modality-specific tasks (drawing is visual, writing is verbal, etc.).

This finding also has implications for creativity. Not only does left hypofrontality allow for modality-specific stimulus representation, it allows for attentional disinhibition. Thus, experts have better access to more visual associates upon which they can perform divergent thinking operations, and thus make *creative* modality-specific products.

General Discussion

Visual artists encode novel visual information more efficiently than control participants, both within the domain of rendering and in at least one task outside of it. This ability to transfer an encoding advantage outside of a domain of expertise implies that expert visual artists are prepared to perceive the unknown similarly to the way that novices perceive the familiar. However, novices' variable encoding strategies are only *attenuated* in experts, suggesting that training in visual art may allow for perceiving novel information in a manner only *similar* to that for familiar information. More extensive training may be associated with encoding strategies for novel and familiar stimuli that are indistinguishable. This possibility is of importance to the field of education, as students can be trained to encode novel information potentially as efficiently as familiar information. Neural plasticity caused by musical training has been demonstrated (Hyde et al., 2009), so there is potential for advantageous left hypofrontality to be an effect of visual art training.

With less time required to fully encode a novel stimulus, cognitive resources are free to be utilized for additional operations upon it and previously-encoded or recalled stimuli, including divergent thinking operations. In Experiment 1, experts encoded stimuli on average 157 ms faster than novices. In Experiment 2, experts attained levels of recognition that novices required an additional 75 ms to attain. Ecologically speaking, that additional processing time can be used to attend to other streams of stimulation, then make a creative connection. This process can be referred to as insight, and is distinct from elaboration, the phase during which the creative insight is turned into a tangible product. Martindale and colleagues (Martindale & Hasenfus, 1978; Martindale, Hines, Mitchell, & Covello, 1984) showed distinct brain activation patterns (as measured by electroencephalogram) during each of these two phases. The current results extend this two-stage theory of creativity; insight may be dependent upon processing on the scale of tens or hundreds of milliseconds, time made available by efficient encoding.

The results of the SPQ-B are consistent with the hypothesis that experts' creativity is based on attentional disinhibition, which allows them to make connections between far-flung associates; making distant connections on the basis of rapid encoding may be responsible for experts' self-reports of their speech or behavior being perceived by others as odd, as well as for having unusual perceptual experiences.

In order to more fully understand expert cognition, work currently under way by the authors examines experts' abilities to retain and manipulate novel visual information.

Acknowledgments

This work was supported by a grant from the Temple University Research Incentive Fund. Shannon Fitzhugh provided technical assistance with data analysis. Dr. Thomas Shipley provided access to eye tracking equipment.

References

Anderson, J.R., Qin, Y., Jung, K-J., Carter, C.S. (2007). Information-processing modules and their relative modality specificity. *Cognitive Psychology 54*, 185-217.

Baruch, I., Hemsley, D.R., & Gray, J.A. (1988). Latent inhibition and "psychotic proneness" in normal subjects. *Personality and Individual Differences, 9*, 777-784.

Buchsbaum, M.S., Yang, S., Hazlett, E., Siegel, B.V., Germans, M., Haznedar, M., O'Flaithbheartaigh, S., Wei, T., Silverman, J., & Siever, L.J. (1997). Ventricular volume and asymmetry in schizotypal personality disorder and schizophrenia assessed with magnetic resonance imaging. *Schizophrenia Research, 27*(1), 45-53.

Burch, G.J., Pavelis, C., Hemsley, D.R., & Corr, P.J. (2006). Schizotypy and creativity in visual artists. *British Journal of Psychology, 97*, 177-190.

Carlsson, I., Wendt, P.E., & Risberg, J. (2000). On the neurobiology of creativity. Differences in frontal activity between high and low creative subjects. *Neuropsychologia, 38*, 873-885.

Carson, S.H., Peterson, J.B., & Higgins, D.M. (2003). Decreased latent inhibition is associated with increased creative achievement in high-functioning individuals. *Journal of Personality and Social Psychology, 85*(3), 499-506.

Chase, W.G., & Simon, H.A. (1973). The mind's eye in chess. In W.G. Chase (ed.), *Visual information processing* (pp. 215-281). New York: Academic Press.

Cohen, D.J. (2005). Look little, look often: The influence of gaze frequency on drawing accuracy. *Perception & Psychophysics, 67*(6), 997-1009.

Curby, K.M., Glazek, K., & Gauthier, I. (2009). A visual short-term memory advantage for objects of expertise. *Journal of Experimental Psychology: Human Perception and Performance, 35*(1), 94-107.

Ericsson, K.A., & Delaney, P.F. (1999). Long-term working memory as an alternative to capacity models of working memory in everyday skilled performance. In A. Miyake & P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control* (pp. 257-297). New York, NY: Cambridge University Press.

Ericsson, K. A., & Kintsch, W. (1995). Long-term working memory. *Psychological Review, 102*, 211-245.

Hyde, K.L., Lerch, J., Norton, A., Forgeard, M., Winner, E., Evans, A.C., & Schlaug, G. (2009). Musical training shapes structural brain development. *The Journal of Neuroscience, 29*(10), 3019-3025.

Jung, R.E., Grazioplene, R., Caprihan, A., Chavez, R.S., & Haier, R.J. (2010). White matter integrity, creativity, and psychopathology: Disentangling constructs with diffusion tensor imaging. *PLoS ONE, 5*(3), 1-7.

Kozbelt, A. (2001). Artists as experts in visual cognition. *Visual Cognition, 8*(6), 705-723.

Lubow, R.E., & Gewirtz, J.C. (1995). Latent inhibition in humans: Data, theory, and implications for schizophrenia. *Psychological Bulletin, 117*, 87-103.

Luftig, R.L. (1994). *The schooled mind: Do the arts make a difference? An empirical evaluation of the Hamilton Fairfield SPECTRA+ Program, 1992-93*, Center for Human Development, Learning, and Teaching, Miami University, Oxford, Ohio.

Martindale, C., & Hasenfus, N. (1978). EEG differences as a function of creativity, stage of the creative process, and effort to be original. *Biological Psychology, 6*, 157-167.

Martindale, C., Hines, D., Mitchell, L., & Covello, E. (1984). EEG alpha asymmetry and creativity. *Personality and Individual Differences, 5*(1), 77-86.

Mednick, S.A. (1962). The associative basis of the creative process. *Psychological Review, 69*(3), 220-232.

Miller, G.F., & Tal, I.R. (2007). Schizotypy versus openness and intelligence as predictors of creativity. *Schizophrenia Research, 93*, 317-324.

Raine, A., & Benishay, D. (1995). The SPQ-B: A brief screening instrument for schizotypal personality disorder. *Journal of Personality Disorders 9*(4), 346-355.

Raine, A., Lencz, T., Yaralian, P., Bihrle, S., LaCasse, L., Ventura, J., & Colletti, P. (2002). Prefrontal structural and functional deficits in schizotypal personality disorder. *Schizophrenia Bulletin, 28*(3), 501-513.

Schuldberg, D. (2000-2001). Six subclinical spectrum traits in normal creativity. *Creativity Research Journal, 13*(1), 5-16.

Snodgrass, J.G., & Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory, 6*(2), 174-215.

Snyder, A.W., Mulcahy, E., Taylor, J.L., Mitchell, D.J., Sachdev, P., & Gandevia, S.C. (2003). Savant-like skills exposed in normal people by suppressing the left fronto-temporal lobe. *Journal of Integrative Neuroscience, 2*(2), 149-158.

Tchalenko, J. (2007). Eye movements in drawing simple lines. *Perception, 36*, 1152-1167.

Tchalenko, J. (2009). Segmentation and accuracy in copying and drawing: Experts and beginners. *Vision Research 49*, 791-800.

Theeuwes, J., Olivers, C.N.L., & Chizk, C.L. (2005). Remembering a location makes the eyes curve away. *Psychological Science,16*(3),196–199.

Tremblay, S., Saint-Aubin, J., & Jalbert, A. (2006). Rehearsal in serial memory for visual-spatial information: Evidence from eye movements. *Psychonomic Bulletin & Review, 13*(3), 452-457.

Vaughn, K., & Winner, E. (2000). SAT scores of students who study the arts: What we can and cannot conclude about the association. *Journal of Aesthetic Education, 34*(3-4), 77-89.

Weisberg, R. W. (2006). *Creativity: Understanding innovation in problem solving, science, invention, and the arts.* Hoboken, NJ: John Wiley.

Melody Recognition: Effects of Articulation Format

Stephen Wee Hun Lim (gmslwhs@nus.edu.sg)
Winston D. Goh (psygohw@nus.edu.sg)
Department of Psychology, National University of Singapore
Block AS4, 9 Arts Link, Singapore117570

Abstract

Various surface features – timbre, tempo, and pitch – influence melody recognition memory, but articulation format effects, if any, remain unknown. For the first time, these effects were examined. In Experiment 1, melodies that remained in the same, or appeared in a different but similar, articulation format from study to test were recognized better than were melodies that were presented in a distinct format at test. A similar articulation format adequately induced matching. Experiment 2 revealed that initial perceptual (dis)similarity as a function of the location of articulation (mis)match between two instances of the melody did not accurately determine discrimination performance. An important boundary condition of the matching process was defined: Whether matching occurs depends on the physical quantity, rather than location, of fit between the memory trace and the recognition probe, suggesting a global matching advantage effect.

Keywords: Melody recognition memory; articulation format effects; global matching advantage

Introduction

When we hear a piece of music, we detect and occasionally remember phrases, motifs, themes, syncopations, suspensions, tonic chords, cadences, and so on. We recognize the instrument playing the melody, or even identify with the emotions of the specific musician performing the work. To this end, what exactly is the nature of mental representations that underlie the music experience? To address this question, it is useful to first recognize that there are two kinds of information in music, namely *abstract structure* and *surface characteristics* (Trainor, Wu, & Tsang, 2004). The *abstract structure* entails the relative pitches and ratios of the durations between adjacent musical notes, regardless of the individual note's absolute pitch level or length *per se*. *Surface characteristics*, in contrast, contain the non-structural aspects of the music, such as absolute pitch, tempo, and timbre. Both the abstract structure and surface characteristics contribute towards musical interpretation. Representing the abstract structure enables recognition of a melody across different performances, and musical variations of a motif within a musical composition (Large, Palmer, & Pollack, 1995). For example, *Happy Birthday* retains its identity and is readily recognized even when it is played or sung in various keys and tempos, or by different voices or instruments. Yet, these very surface characteristics lead us to identify the specific musician and unique performance of the work, defining the emotional interpretation of that rendition. While Raffman (1993) has suggested that only the abstract structural information is encoded into long-term memory (LTM), others have reported that surface features, such as timbre (e.g., Peretz, Gaudreau, & Bonnel, 1998) and tempo (e.g., Halpern and Müllensiefen, 2008), are also encoded into LTM during a melody recognition task.

In music, the way a melody is articulated shapes its surface appearance. In the extant literature that examined the effects of surface characteristics on melody recognition performance, it is surprising that no study has explored the effects of articulation format, even though it is a feature that is commonly manipulated by both composers and performers. Trained musicians commonly define articulation as whether the music (e.g., melody) is played in a *legato* (i.e., continuous) or *staccato* (i.e., detached) format. Because no one has studied the influence of articulation on melody recognition, our initial motivation was to add to that literature. Thus far, memory representations that subserve explicit recognition of melodies appear to be formed by a highly specialized association that binds together characteristics such as timbre and tempo with melody identity. It is thus attractive to ask whether the articulation feature is tied to a melody's identity and computed during the perceptual analysis of the melodic input. By addressing this question, we hope to explicate more fully the central idea that variability in surface features, along with the idealized canonical structure of music, is important in music perception and processing.

To examine the effects of articulation format on melody recognition, we designed the melody to occur either fully in *legato* form, fully in *staccato* form, or in mixed articulation format (i.e., a combination of *legato* and *staccato* components). When the melody was played in *staccato* form, the duration of each note in the melody was manipulated to last 10% of the full duration when the note was played in *legato* form. The schematic of the eight different articulation formats is shown in Figure 1. These formats are coded as *l*, *s*, *a*, *b*, *c*, *d*, *e*, and *f*: The *legato* and *staccato* formats are abbreviated as format *l* and *s*, respectively, while the six mixed-articulation formats follow an alphabetical system of coding for ease of reference. Each set of four boxes represents sequentially the four bars of the melody respectively.

Taking format *f* for instance, the melody opens in *staccato* form (i.e., the notes of the melody are articulated by the instrument in a disjointed fashion) for the first bar, switches to *legato* form (i.e., the notes are now articulated smoothly

in a continuous manner) by the second bar, returns to *staccato* mode in the third bar, and finally closes with a long-sounding note in the final bar.

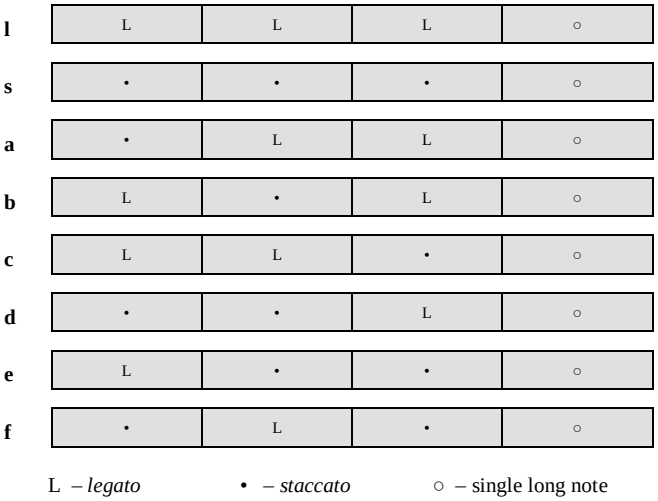


Figure 1: Schematic of the eight different articulation format manipulations.

Experiment 1

In Experiment 1, we asked two questions: (1) Is articulation feature information retained in LTM, and (2) what is the role of feature similarity in melody recognition memory? Our first goal was to investigate the effects of manipulating articulation context on melody recognition. The hypothesis was that to the extent that articulation format information is not erased from, but is in fact preserved in, LTM, discrimination performance ought to improve when old melodies are repeated in the same articulation format, as compared to when the melodies appeared in a distinct articulation format during the recognition stage.

In addition, we recognized that extant studies that examined surface feature effects have used test stimuli that were denoted as either of the same or different format, neglecting effects that could arise from varying magnitudes of intermediate perceptual differences. For instance, Peretz *et al.* (1998) presented melodies in timbres at test that were either the same as, or distinct from, those used at study; Halpern and Müllensiefen (2008) made the tempo changes in altered tunes "large enough to be perceptible" (p. 1378). Effects of fine-grained perceptual details of surface features, such as tempo or timbre, have been somewhat overlooked, so it is unclear whether these details actually contributed to the disparate surface feature effects observed in the literature. As such, a second goal was to assess the contribution of fine perceptual details in melody recognition memory, by including a similar-articulation-format condition. We speculated that to the extent that articulation similarity constitutes an integrated part of the matching and retrieval processes involved in melody recognition, performance ought to improve even when old melodies are tested with a different *but similar* articulation format, as compared to when the melodies appeared in a distinct articulation format.

Method

Participants Forty-seven introductory psychology students participated for course credit.

Materials The stimulus set contained 48 novel monophonic melodies (see Figure 2 for samples). An equal number of four-bar melodies were composed in the tonality (key) of C major, C minor, G major, or G minor. The melodies started either on the tonic, mediant, or dominant, but always ended with a single long note on the tonic of their home key. Each melody was written in simple triple or simple quadruple time, lasting approximately six seconds or 7.2 seconds respectively. The melodies were constructed using the *Finale 2009* software, and saved as .wav sound files.

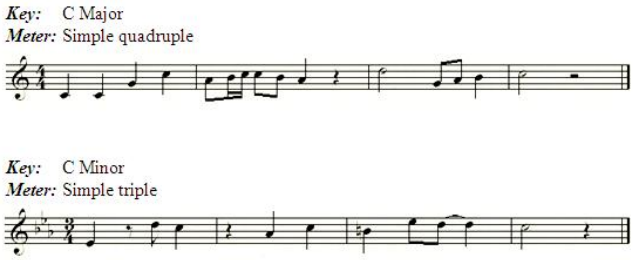


Figure 2: Samples of the 48 melodies used.

Prior to conducting Experiment 1, we first derived a multidimensional "articulation map" using MDS techniques (Kruskal & Wish, 1978) that shows the similarity relations between the individual articulation formats that will be used as the stimulus materials. This procedure was necessary to ensure that the selection of specific articulation formats for

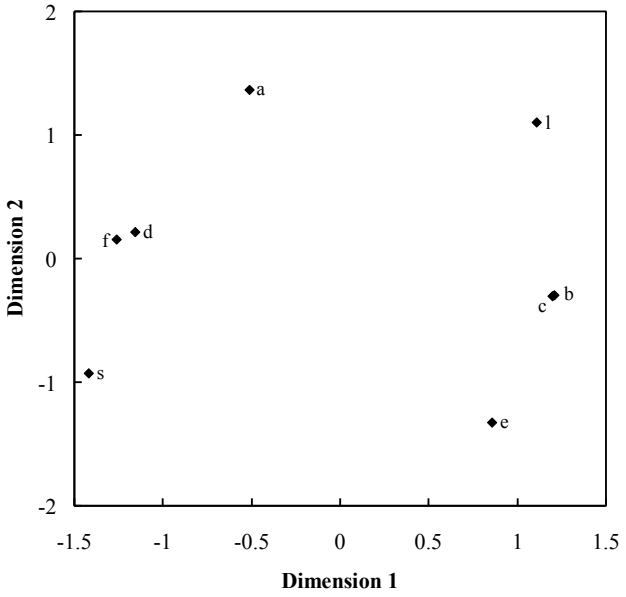


Figure 3: Two-dimensional MDS solution for eight articulation formats.

use in the subsequent main experiments can be based on objective measures of the degree of perceived similarity among different articulation formats. Sixteen students from the same population sample but who did not participate in the main experiments rated the pairwise similarity of the eight articulation formats across four different melodies using a 7-point Likert scale. The two-dimensional MDS solution (Kruskal's stress = .15, R^2 = .85) for the eight articulation formats appears in Figure 3. The interpretation is that the further away two articulation formats are positioned from each other in space, the more perceptually distinct they are. Two different combinations of articulation formats were selected for melody presentation. For each combination, the articulation formats are listed in the order that constitutes the same-, similar-, and distinct-articulation context conditions, respectively: (1) *l, b, s* and (2) *s, f, l.* These sets were created for counterbalancing purposes described in the procedure.

Apparatus Computers equipped with 16-bit sound cards were used for the experiment. Participants received the signals through a pair of Beyerdynamic DT150 headphones at approximately 70 dB SPL. The stimuli were presented using E-prime 1.2, and data were collected using the PST Serial Response Box (Schneider, Eschman, & Zuccolotto, 2002), with the left- and right-most buttons of the button-box labelled *No* and *Yes* respectively.

Design The 48 melodies were divided equally into two lists. One list was designated to consist of old melodies while the other to consist of new melodies. At study, all the 24 old melodies were presented using a single articulation format. In the test phase, the 24 new melodies were divided among three articulation formats, where eight melodies were assigned to be presented in the same format, eight in a similar format, and the remaining eight in a distinct format. For the 24 old melodies, likewise, eight were assigned to the same-articulation context condition, eight to the similar-articulation context condition, and the remaining eight to the distinct-articulation context condition (see Table 1).

Procedure Half of the participants were randomly assigned to listen to melodies played by the clarinet, while the other half were randomly allocated to listen to melodies played by the violin. The session consisted of two parts – the memorization phase and the recognition phase. The forthcoming recognition test was made known to participants before the memorization phase started. Participants were told to silently memorize each melody that was played through the headphones. At the start of each trial, a ready prompt was displayed on the monitor for one second, after which it was deleted. One second later, a melody was played over the headphones; the melody was repeated 800 ms following its first presentation. Participants then pressed the space key to proceed to listen to the next melody. This sequence persisted until all 24 melodies had been presented. The melody presentation sequence was randomized across participants.

Following the memorization phase, participants were first presented with versions of two well-known melodies – *Mary had a little lamb* and *London bridge is falling down* – that varied in their articulation formats to clarify the definition of "form". After which, the recognition test began. On each trial, the ready prompt appeared for one second and disappeared. 800 ms later, the question *Did you hear this melody in Part 1?* was displayed, and a single melody was played through the headphones. Participants were told to press the *Yes* button on the Serial Response Box if they thought they had heard the melody earlier, regardless of the original "form" (i.e., articulation format) that the melody was presented in. Otherwise, they were told to press the *No* button. Participants were told to respond as accurately as possible. No feedback was provided on any of the trials. A new trial was started after a button response.

Results and Discussion

Table 2 presents the pattern of results for *d'* performance across the three articulation-context conditions. There was a reliable main effect of articulation context, $F(2, 90) = 3.94$, $MSe = 0.36$, $p < .05$. Pairwise comparisons revealed that participants were significantly better at discriminating melodies presented with the same articulation format than they were at discriminating melodies presented with a distinct articulation format, $t(46) = 2.42$, $p < .05$; participants also performed better when melodies appeared in a similar articulation format than they did when melodies appeared in a distinct format, $t(46) = 2.03$, $p < .05$.

Table 1: Summary of Experiment 1's Design

Memorization	Recognition					
Study melodies	Test melodies (Old)			Test melodies (New)		
	Articulation format context					
	Same	Similar	Distinct	Same	Similar	Distinct
	Set combination 1 articulation formats					
l	l	b	s	l	b	s
24	8	8	8	8	8	8
	Set combination 2 articulation formats					
s	s	f	l	s	f	l
24	8	8	8	8	8	8

Discriminability did not differ between the same- and similar-articulation context conditions, $t < 1.05$. This pattern of results indicates that discriminability increased significantly so long as melodies were tested in at least a similar articulation format.

Table 2: Discrimination Performance (d') Across Articulation-Context Conditions in Experiment 1.

	Articulation context		
	Same	Similar	Distinct
M	0.97	0.90	0.64
SD	0.66	0.56	0.67

The present data revealed an advantage in melody recognition for same-articulation repetitions over distinct-articulation presentations. There was also an advantage in melody recognition for similar-articulation presentations over distinct-articulation presentations. An interpretation based on the the now-classic encoding specificity framework (Tulving & Thompson, 1973) is apt. Under this framework, the effectiveness of a retrieval cue depends on its degree of relatedness to the encoding of an item at first. Our view is that surface (articulation) and structural attributes of a melody are stored together in the LTM trace. Melody recognition is reliable when a specific match between the episodic memory trace and the probe occurs, but is hampered when there is a mismatch.

The comparison of shared properties between the memory trace and the probe implies that item similarity *per se* constitutes an integral part of the retrieval process. In fact, the degree of similarity among the features of the exemplar traces in memory and the target probe forms a central aspect in exemplar models of memory and categorization (Gillund & Shiffrin, 1984; Hintzman, 1988). Memory theorists have assumed that memory for a stimulus is really memory for features contained in that stimulus. The global matching approach (see Clark & Gronlund, 1996) suggests that these features in a test item, when matched with the features that have earlier been stored in memory, evoke a familiarity signal. Specifically, the greater the degree of match is, the stronger the signal will be. In our case, when a melody was re-played in the same or in a similar articulation format at test, there are many overlapping features between the articulation formats of the two melody instances from study to test. These overlaps presumably contribute to a strong sense of familiarity signal evoked by resemblance to the studied melody (see Cleary, 2004). In contrast, when the melody appeared at test in a distinct format, there are few overlapping features with the melody's original format. As such, the familiarity signal is presumably weaker, which hinders melody discrimination.

The present experiment suggests that when matching occurs, melody recognition performance is reliable at test. Experiment 2 was designed to establish an important boundary condition which determines whether this matching process would prevail (or fail).

Experiment 2

A first examination of the articulation similarity scaling solution shown in Figure 3 reveals that the greater the amount of *physical* articulation match between two instances of a melody, the more similar they were perceived to be. For instance, formats *d* and *f*, each containing two bars of *staccato* component, were perceived as similar to each other. But a closer look at the scaling solution reveals that only when the articulation format of two instances of the melody matched *at the melody's onset* would the two instances of the melody be perceived as similar to each other. This interpretation can explain why format *e* was perceived as rather different from formats *d* and *f* even though each of these formats contained two bars of *staccato* component. This observation is intriguing because two articulation formats, given the same quantitative amount of articulation match, could in fact be perceived as different from each other due to the fact that the match did not occur at the melody's onset.

We therefore pursued a third question here: Would this perceptual dissimilarity between two instances of the melody (e.g., in formats *d* and *e*) due to the location of the (mis)match hamper discrimination performance during the test stage, even when both instances contain the exact same quantity of articulation match (e.g., two bars of *staccato* component)? The goal was to illuminate the underlying nature of the matching process in melody recognition memory, and we hypothesized that to the extent that perceptual dissimilarity, as a function of the location of (mis)match in format, affects matching between study and test, discrimination performance ought to be hampered when old melodies that were originally played in, say, format *s* are repeated in format *e* (i.e., perceptually dissimilar format) at test, as compared to when the melodies are repeated in format *d* or *f* (i.e., perceptually similar format) at test, although formats *d*, *e*, and *f* each contains the exact same quantity (i.e., two bars) of *staccato* component.

Method

Participants Sixty-four psychology undergraduates participated. None had participated in Experiment 1.

Materials, Apparatus, Design, and Procedure The materials and procedures were essentially the same as those of Experiment 1, with a slight modification in materials. Based on Figure 3, four different combinations of articulation formats were selected for melody presentation. For each combination, the articulation formats are listed in the order that constitutes the same-, similar-, and distinct-articulation context conditions respectively: (1) *s*, *d*, *e*, (2) *s*, *f*, *e*, (3) *l*, *b*, *a*, and (4) *l*, *c*, *a*. Set combination was counterbalanced across participants.

Results and Discussion

Table 3 presents the pattern of results for *d'* performance across the three articulation-context conditions. There was no reliable main effect of articulation context, $F < 1.23$.

Discriminability between the same-, similar-, and distinct-articulation context conditions did not differ reliably. Articulation format did not influence performance.

Table 3: Discrimination Performance (d') Across Articulation-Context Conditions in Experiment 2.

	Articulation context		
	Same	Similar	Distinct
M	1.13	0.94	1.09
SD	0.67	0.78	0.70

Experiment 1 suggested that articulation properties are bound with the melody's structural identity. Surface feature information of the melody is first encoded and stored in the memory trace, and later used to retrieve the melody. Because a same- or similar-feature repetition constitutes an exact, or at least a close, match with the memory trace for the old melody, the trace becomes more salient than the other competing traces, enhancing discrimination performance. On the other hand, a distinct-feature presentation would not match with the trace for the old melody, thus performance is hampered. The interpretation is that given a retrieval cue that coincides with the initial encoding of the melody in terms of its surface properties, the cue would help the melody to be recovered at test.

But Experiment 2 revealed that initial perceptual (dis)similarity, as a function of the location of feature (mis)match between two instances of the melody, did not accurately determine discrimination performance. When two instances of the melody are perceived as different from each other from study to test, matching presumably would not occur. Yet, some form of matching must have occurred despite the perceptual mismatch because the overall discrimination performance (in the distinct articulation condition) was good, average $d' = 1.09$.

Values of *d'* between 1 and 2 usually represent good yes-no recognition performance (Neath & Surprenant, 2003, p. 202). To further justify that this was good performance, we conducted three planned comparisons on the *d'* data. The first and second comparisons established that the data sets between Experiments 1 and 2 were comparable: Performance in the same-articulation conditions, as well as performance in the similar-articulation conditions, across both experiments did not differ, ts < 1.28, ps > .21. The third comparison used performance in Experiment 1's distinct articulation condition as baseline, and revealed that performance in Experiment 2's distinct-articulation condition reliably exceeded performance in this baseline condition, $t(109) = 3.44$, $p < .01$, implicating good discrimination performance in this case.

Thus, the logical inference is that whether matching would occur is likely to be contingent on the absolute physical quantity of match between the memory trace and the recognition probe per se, rather than the perception of dissimilarity due to the location of (mis)match in the feature attributes. These data defined an important boundary condition of matching observed in melody recognition under which matching would (or would not) be successful.

General Discussion

Several studies have demonstrated that the alteration of the initial part of a sound can affect the recognition of musical instruments (e.g., Berger, 1964; Grey & Moorer, 1977). These findings suggest that temporal features are important in timbre perception and music processing at large. Yet, Experiment 2 suggests that altering the initial part of the articulation format (i.e., at the onset of a melody) did not influence discrimination performance. In explaining these data, we offer a global matching advantage interpretation which finds its roots in Gestalt psychology. A basic position of the Gestalt view is that a whole is qualitatively different from the complex that one might predict by considering only its parts. Under this view, wholes are organized prior to perceptual analysis of their properties and components in perceptual organization. Navon (1977) proposed that perceptual processing starts with global structuring and later moves towards more fine-grained analysis. This proposal was termed as the *global precedence hypothesis*. This hypothesis has been tested by studying the perception of hierarchical patterns in which larger figures are constructed by suitable arrangements of smaller figures.

An example is a set of large letters constructed from the same set of smaller letters having either the same identity as the larger letter or a different identity (see Figure 4). The larger letter is considered a higher-level unit relative to the smaller letters, which are, in turn, lower-level units. Properties of the higher-level unit are considered more

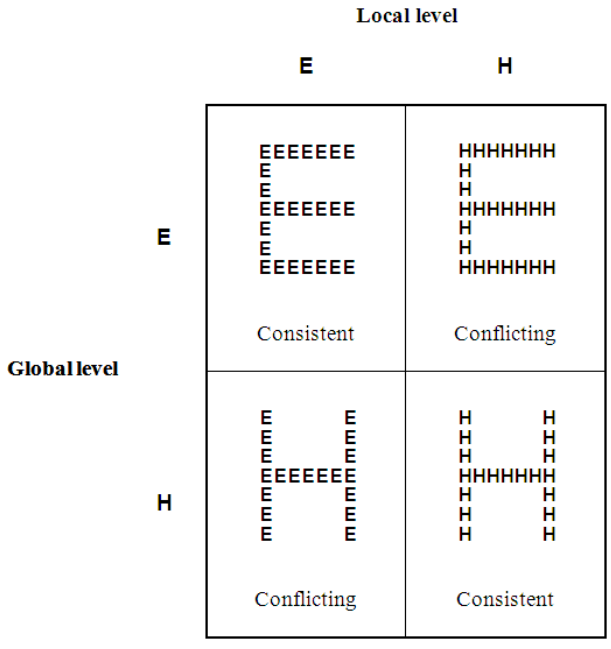


Figure 4: An example of Navon's (1977) type hierarchical stimuli. Large Es and Hs are composed using small Es and Hs.

global than properties of the lower-level units by virtue of their position in the hierarchical structure. In a typical experiment, observers are presented with such stimuli and are required to identify the larger (i.e., global) or the smaller (i.e., local) letter in different trials. *Global advantage* is observed, where the global letter is identified faster than the local letter.

Our view is that an analogous global advantage mechanism operates in the matching process found in melody recognition. The general articulation format of the melody (i.e., whether the melody is overall presented in a *staccato* or *legato* format) is considered a higher-level unit relative to the specific format of individual bars, which are, in turn, lower-level units, and properties of the higher-level unit are considered more global than properties of the lower-level (local) units based on their position in the hierarchical structure. In order for matching to occur, that there is a *global* match based on the *absolute quantity* of match between the memory trace and the recognition probe *per se* is more critical, as compared with whether there is a *local* match between the articulation format at the onset of the test melody and the format at the onset of the study melody. Once global matching attains, melody discrimination performance is enhanced.

The present global matching advantage hypothesis can be verified in a future study that manipulates the overall (global) and local matches in, say, timbre between two instances of a melody, by specifically altering the timbre at various temporal points (e.g., the onset) of the melody. Others could assess the effects of surface features that have yet to receive attention, such as the use of ornaments or phrase boundaries. More broadly, future investigations can extend to the domain of speech perception. There had been considerable work which argued for a commonality between music and speech processing (see Patel, 2003), and comparing these two processes can lead to an understanding of wider (and potentially shared) principles of perceptual categorization and temporal organization across brain areas (McMullen & Saffran, 2004; Patel, 2003). Thus, it is of interest whether the present effects would emerge in speech. There is a large body of data suggesting that talker's voice, a surface feature of spoken language, is encoded into LTM. Specifically, old words were recognized better when they were tested in a voice that matched with the original voice that originally spoke the word at study, than when the voices did not match (see Goh, 2005 for a review). Yet, the boundaries that permit (or prevent) this match in a speech context are not well defined. It is worthwhile to explore the extent to which speech recognition performance is driven by the absolute match in the physical properties of voice between two instances of speech and/or the location of match *per se* (e.g., in a sentence context).

References

Berger, K. W. (1964). Some factors in the recognition of timbre. *Journal of the Acoustical Society of America*, *36*, 1888–1891.

Clark, S. E., & Gronlund, S. D. (1996). Global matching models of recognition memory: How the models match the data. *Psychonomic Bulletin & Review*, *3*, 37-60.

Cleary, A. M. (2004). Orthography, phonology, and meaning: Word features that give rise to feelings of familiarity in recognition. *Psychonomic Bulletin & Review*, *11*, 446–451.

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review*, *91*, 1–67.

Goh, W. D. (2005). Talker variability and recognition memory: Instance-specific and voice-specific effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 40–53.

Grey, J. M., & Moorer, J. A. (1977). Perceptual evaluations of synthesized musical instrument tones. *Journal of the Acoustical Society of America*, *62*, 454–462.

Halpern, A. R., & Müllensiefen, D. (2008). Effects of timbre and tempo change on memory for music. *The Quarterly Journal of Experimental Psychology*, *61*, 1371–1384.

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple trace memory model. *Psychological Review*, *95*, 528–551.

Kruskal, J. B., & Wish, M. (1978). *Multidimensional scaling*. Newbury Park, CA: Sage.

Large, E. W., Palmer, C., & Pollack, J. B. (1995). Reduced memory representations for music. *Cognitive Science*, *19*, 53–96.

McMullen, E., & Saffran, J. R. (2004). Music and language: A developmental comparison. *Music Perception*, *21*, 289–311.

Navon, D. (1977). Forest before trees: The precedence of global features in visual perception. *Cognitive Psychology*, *9*, 353–383.

Neath, I., & Surprenant, A. M. (2003). *Human memory: An introduction to research, data, and theory*. Toronto: Wadsworth.

Patel, A. D. (2003). Language, music, syntax and the brain. *Nature Neuroscience*, *6*, 674–681.

Peretz, I., Gaudreau, D., & Bonnel, A. (1998). Exposure effects on music preference and recognition. *Memory & Cognition*, *26*, 884–902.

Raffman, D. (1993). *Language, music, and mind*. Cambridge, MA: MIT Press.

Schneider, W., Eschman, A., & Zuccolott, A. (2002). *E-Prime User's Guide*. Pittsburg: Psychology Software Tool Inc.

Trainor, L. J., Wu, L. & Tsang, C. D. (2004). Long-term memory for music: Infants remember tempo and timbre. *Developmental Science*, *7*, 289–296.

Tulving, E., & Thompson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, *80*, 352–373.

On the Relationship Between Entropy and Meaning in Music: An Exploration with Recurrent Neural Networks

Greg Cox (grcox@indiana.edu)
Department of Psychological and Brain Sciences, Indiana University
1101 E. Tenth St., Bloomington, IN 47405 USA

Abstract

Meyer (1956) postulated that meaning in music is directly related to entropy–that high entropy (uncertainty) engenders greater subjective tension, which is correlated with more meaningful musical events. Current statistical models of music are often limited to music with a single melodic line, impeding wider investigation of Meyer's hypothesis. I describe a recurrent neural network model which produces estimates of instantaneous entropy for music with multiple parts and use it to analyze a Haydn string quartet. Features found by traditional analysis to be related to tension are shown to have characteristic signatures in the model's entropy measures. Thus, an information-based approach to musical analysis can elaborate on traditional understanding of music and can shed light on the more general cognitive phenomenon of musical meaning.

Keywords: Music cognition; neural networks; information theory; entropy.

Introduction

Music is an intriguing artifact of human culture, and one of the challenges in music cognition is to explain how music is capable of having meaning for the listener. Much music carries meaning by virtue of associations to non-musical things like stories and literature (Rimsky-Korsakov's *Sheherezade*), visual imagery (Mussorgsky's *Pictures at an Exhibition*), environmental sounds (taxi horns in Gershwin's *An American in Paris*), symbols (the "cross" motif in the Fugue in C-sharp minor from Book I of J. S. Bach's *Well-Tempered Clavier*), and the meaning of text or lyrics. However, music theorists and cognitive scientists have been particularly concerned with investigating music that lacks text and that does not explicitly refer to anything non-musical.[1]

Meyer (1956) postulated that meaning in music arises from the ability of a musical event to imply or refer to other *musical* events that are expected to follow it. In a later work, he summarized his hypothesis:

> Musical meaning arises when an antecedent situation, requiring an estimate as to the probable modes of pattern continuation, produces uncertainty as to the temporal-tonal nature of the expected consequent. (Meyer, 1957, p. 416)

Within a particular style, a given musical event–e.g., a dominant chord–is expected to be followed by another musical event–e.g., a tonic chord, making for an authentic cadence. These expectations can also be violated or ambiguous–perhaps the dominant chord is followed by a submediant chord, making for a deceptive cadence. In such cases, a listener experiences tension which is manifested both in subjective reports (Krumhansl, 1996) and in physiological affective responses (Steinbeis, Koelsch, & Sloboda, 2006). Tension and its associated affective qualities–reflecting uncertainty–can thus be a signature of musical meaning.

Beyond suggesting a direct link between musical meaning and tension, Meyer's definition is readily formalized via the concept of entropy, which is a measure of both uncertainty and information content (more information is necessary to describe something that is difficult to predict). Other music theorists have made use of entropy measures in a variety of ways, including the analysis of structure in atonal music (Hiller & Fuller, 1967), stylistic variation in tonal music (Knopoff & Hutchinson, 1981), and differences between musical styles (Margulis & Beatty, 2008).

While most music theoretical studies of information in music have focused on gross properties of style or large segments of music, recently, modeling techniques from cognitive science have been brought to bear on Meyer's notion of musical meaning. Markov models and recurrent neural networks enable researchers to quantify entropy and other information measures by specifying the underlying predictive model a listener might have. Measures of information content in Markov models of music can predict structural boundaries that correspond to those assigned by human listeners to monophonic (single-part) music in the minimalist style (Potter, Wiggins, & Pearce, 2007; Abdallah & Plumbley, 2009).

However, structural boundaries are only a part of musical meaning. If meaning is related to subjective tension arising from uncertainty–i.e., entropy–it should be possible to correlate instantaneous measures of entropy (an "entropy profile") with momentary affective responses to music. For instance, an authentic cadence is a point of repose and thus should be correlated with lower entropy. A dramatic climax should be correlated with a high value of entropy (a local maximum) as it represents a large amount of tension. Human-derived entropy profiles for Bach chorale melodies (Manzara, Witten, & James, 1992) are in accord with these intuitions.

It is also likely that different dimensions of music (e.g., pitch, rhythm, harmony) contribute differently to tension and to entropy. This notion is embodied in multiple viewpoint models (Conklin & Witten, 1995), although since these models have only been applied to monophonic music, they tend to focus on pitch to the exclusion of rhythmic, harmonic, and

[1] Of course, even non-referential music is sure to remind a listener–consciously or not–of something other than the music he or she is currently hearing. However, these non-musical associations tend to vary widely between individuals and as such cannot be relied upon as a basis for musical meaning.

contrapuntal dimensions. A study of entropy as a correlate of tension should address more than just single melodic lines, since harmony and counterpoint are critical dimensions along which music can meaningfully vary.

The present study investigates the extent to which entropy can serve as a general measure of tension–and thus meaning–in music. To that end, I present a recurrent neural network as a predictive model of polyphonic (multiple-part) music and compare entropy measures derived from the model with traditional music theoretical analysis. I show that features of the traditional analysis related to subjective tension have particular signatures in the model's entropy measures, supporting the hypothesis that entropy underlies musical meaning.

A Recurrent Neural Network for Music Prediction

Recurrent neural networks (RNNs) have been fruitfully used as models of sequential prediction in many domains. In music research, they have been used to compose monophonic music both with (Mozer, 1994) and without (Todd, 1989) accompanying harmonic progressions, and to model the acquisition and perception of tonal harmony (Bharucha & Todd, 1989).

Although Markov models have seen wider–and, arguably, more productive–application in monophonic music than have RNNs, Markov models are less well suited to modeling polyphonic music. Monophonic music is easily translated into a sequence of discrete symbols drawn from a finite alphabet. It is much less clear, however, how one might translate polyphonic music into a language appropriate for a Markov model, as such music includes multiple pitch sequences updating at different rates with varying degrees of independence. For instance, to describe just the pitch transitions of a four-part piece where each part spans a diatonic octave (eight possible pitches), a naïve first-order Markov model would require a state space with $8^4 = 4096$ points and a transition matrix with $4096^2 = 16777216$ entries, and this does not even include any information about rhythm![2] Further, in any realistic training set, only a small portion of the number of possible transitions will be represented, leading to problems of over-fitting and lack of generalization (though these problems can be solved in some domains with the smoothing techniques described by Pearce & Wiggins, 2004). RNNs tend to avoid these problems, since they do not require enumerating and/or representing all state transition probabilities, but rather the weights of the network represent only those dependencies necessary to minimize prediction error. In addition, since the RNN must learn its own internal representation of the input, it will naturally converge toward representations that capture the generalities in the training set.

It should be emphasized that, as with a Markov model of music, no literal psychological reality is meant to be ascribed to the structure and training procedure of a RNN. Rather, the network should be seen only as a statistical model that mirrors the *function*–not necessarily the form–of whatever predictive model a human listener has acquired through musical experience. The resulting trained network does not–indeed, cannot–represent a listener's *entire* understanding of music in general, but is limited to representing the expectations of a listener who is familiar with a piece of music and/or its style.

[2]By making certain independence assumptions, it is possible to simplify a Markov model greatly, but it is not in general possible to know, *a priori*, what those assumptions should be.

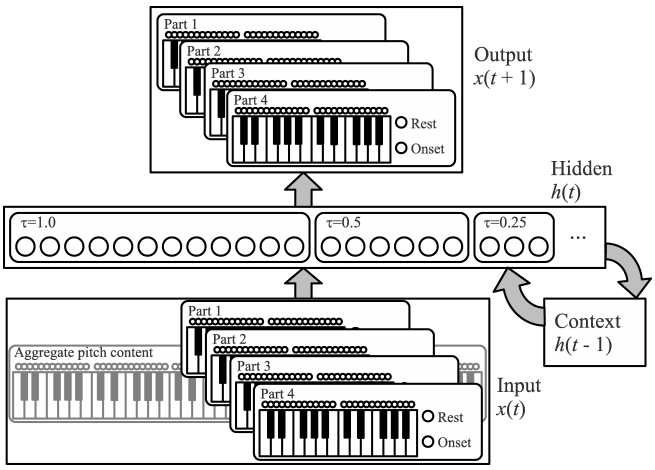


Figure 1: Schematic of the recurrent neural network used in this study, assuming 4 distinct parts. See text for details.

Architecture

The architecture of the RNN used in this study is shown in Figure 1. As in Elman (1990), the network is presented with the current state of the music, $x(t)$, and is trained–via back-propagation through time with a single time step (Rumelhart, Hinton, & Williams, 1986)–to produce the state of the music at the next time step, $x(t+1)$, at its output layer. Output layer units use a logistic activation function $f(net) = 1/(1+exp(-net))$, where *net* is the net input to the unit. Successive layers are fully interconnected.

Time Musical time is divided into discrete "time steps" of equal length, and the input, $x(t)$, describes all musical events (pitches and note onsets) occurring during that time step.

Input/Output Representation The input at time t, $x(t)$, is the concatenation of several vectors representing, for each part (i.e., distinct instrument or timbre, e.g., violin, piano, etc.) in a piece of music, the part's state along two dimensions: pitch and rhythm. The pitch dimension is represented in a localist fashion, with one unit corresponding to each absolute pitch (in twelve-tone equal temperament) that could occur in a part, including a unit representing silence or a "rest". At the input layer, a pitch unit is active (1) if it is currently sounding and zero otherwise; at the output layer, a pitch unit's activity represents the degree to which that pitch is expected at the next time step. The pitch state vector $\pi_p(t)$ for part p at time t may thus be expressed $\pi_p(t) = \langle \pi_{p0}, \pi_{p1}, \ldots, \pi_{pn}, \pi_{pREST} \rangle$ for possible pitches $0 \ldots n$ and the special *REST* "pitch". The input layer also contains a set of units representing all pitches that are sounding at the

current time across all parts, to allow for generalization of pitch content between parts. However, the network is only trained to predict the pitches of each part individually, not this aggregate pitch content.

An additional unit for each part represents its state along the rhythmic dimension: this unit is active (1) when the current time step contains a note onset within that part, and is otherwise inactive (when the part is silent or sustaining a previous pitch). At the output layer, this unit can be interpreted as the probability $\rho_p(t+1)$ that part p will contain a note onset at at time $t+1$. Note that the assumption of independence between rhythm and pitch in the input/output representation permits the analysis of each component separately. However, independence of representation does not guarantee probabilistic independence, as both pitch and rhythm units are treated equally in the network's internal representation in the hidden layer.

Hidden Layer Hidden unit activations are a function of the current input, the hidden layer at the previous time step (also called the "context" layer), and each unit's own prior activation. The activation of hidden unit h_i at time t is

$$h_i(t) = \tau_i f\left(\Sigma_j w_{ij} x_j(t) + \Sigma_k w_{ik} h_k(t-1)\right) + (1-\tau_i) h_i(t-1),$$

where $f(\cdot)$ is the logistic activation function described above, w_{ij} is the weight from input unit j to hidden unit i and w_{ik} is the weight from context unit k to hidden unit i. The different time constants τ_i cause the hidden units to change at varying rates over time, permitting the representation of multiple time scales at the hidden layer (Mozer, 1992).

For simplicity, I assume that the number of hidden units with time constant τ is $N_\tau = \lfloor \tau N_1 \rfloor$ where N_1 is the number of units with $\tau = 1$ and $\lfloor \cdot \rfloor$ is the floor function, ensuring that there will be only a finite number of hidden units and time scales represented. In the simulations reported here, each τ is the reciprocal of either a power of 2 or a power of 3, i.e., $\tau = 2^{-\gamma}$ or $\tau = 3^{-\gamma}$ for $\gamma = 0, 1, 2, \ldots$. The choice of time constant scales based on 2 and 3 derives from the predominant metrical subdivisions (duple and triple meters) in Western music, which is the domain of the current study. Thus, the hidden layer best represents information at time scales that are likely to be most salient.

Measures of Entropy

Although there are many ways to measure entropy within the current modeling architecture of the RNN, I will focus on four simple measures, three of which are used in the subsequent musical analyses. For any part p, the pattern of activity over its pitch units (including the "rest" pitch) at the output layer, π_p, can be normalized to sum to one, such that it can be considered a probability distribution over pitches. Then, the entropy with regard to pitch in part p at time t is $H_p^{pitch}(t) = -\sum_{i=0}^{n,REST} \left[\pi_{pi} log_{(n+1)}(\pi_{pi})\right]$, where the base of the logarithm normalizes the entropy to the range from zero to one. Similarly, the entropy with regard to rhythm in part p at time t is $H_p^{rhythm}(t) = -\rho_p(t) log_2(\rho_p(t)) - (1-\rho_p(t)) log_2(1-\rho_p(t))$.

To measure entropy over the entire ensemble rather than within each part, an aggregate pitch probability vector $\pi^* = \langle \pi_0^*, \pi_1^*, \ldots, \pi_n^* \rangle$ is created, where $\pi_i^* = C \sum_{p=0}^{P} \pi_{pi}$, i.e., the sum of the probability assigned to pitch i by each of the P parts, normalized (by constant C) to sum to one. The entropy of π^* can then be computed. The rhythmic entropy of the ensemble is computed over the joint distribution of the onset probabilities of each part. Pitch entropy represents uncertainty about *what* pitches will occur, while rhythmic entropy represents uncertainty about *when* those pitches will occur.

Long-Term and Short-Term Models

As in work with multiple viewpoint models of music (Conklin & Witten, 1995), for each piece of music to be analyzed, two of the above-described networks are trained. The first network is trained on a representative sample of a particular style of music and is meant to represent more global stylistic characteristics acquired by the listener over a longer time span, hence it is called the long-term model (LTM). A second network is trained on just a single piece of that style and is meant to represent knowledge of that piece in particular acquired over less time, hence it is called the short-term model (STM). This distinction is akin to that between "schematic" (LTM) and "veridical" (STM) knowledge made in Justus and Bharucha (2001). Both models produce patterns of activation over output units representing the expected pitch and rhythmic state of each part. These patterns can be combined to form an aggregate prediction from both the STM and LTM models[3]. Following Pearce, Conklin, and Wiggins (2005), this combination is a weighted geometric mean of the output activations for each dimension of each part of each model, where the weight is inversely proportional to the entropy of the activity over the relevant dimension of each part. For example, the aggregate activation of pitch π_i in part p (aggregate rhythm activation is analogous) would be

$$\bar{\pi}_{pi} = \left[\left(\pi_{pi}^{STM}\right)^{\frac{1}{H_{STM}^{pitch}}} \left(\pi_{pi}^{LTM}\right)^{\frac{1}{H_{LTM}^{pitch}}} \right]^{\frac{1}{\frac{1}{H_{STM}^{pitch}} + \frac{1}{H_{LTM}^{pitch}}}}.$$

The effect of combining the STM and LTM in this way is to emphasize "points of agreement" between them. For example, if they both strongly predict a particular pitch, the aggregate activity ascribed to that pitch will be very high. If one model is ambivalent (high entropy) while the other is certain (low entropy) of a particular pitch, the aggregate activity will accrue to the pitch of which one model is certain. If both models are certain but disagree, activity will be diffused over all possible pitches, leading to high entropy of the aggregate STM-LTM prediction.

[3] Justus and Bharucha (2001) found that schematic (LTM) and veridical (STM) knowledge made independent contributions to musical expectations; their results are consistent with a weighted geometric mean of those two sources of information.

Applying the Network: Haydn's String Quartet Op. 20, No. 3, First Movement

Because Markov models are already well-suited to modeling monophonic music and RNNs have already been shown to deal well with monophonic melodies, even those with accompanying harmonic progressions, I wanted to explore polyphonic music that did not have a simple "melody plus chords" texture–in other words, music that has been difficult to model with previous approaches. There is also an inherent difficulty in correlating entropy with tension, since tension in a listener is not directly observable. As such, I will consider tension as it is normatively described by traditional music theoretical analysis. The analytical procedure described below has been replicated with a variety of corpora, including Bach chorales, Chopin piano preludes, and Schönberg's *Pierrot Lunaire*, with similar results regarding the relationship between entropy and traditional accounts of tension. To show how an analysis of entropy relates to traditional approaches, I report here the results of a single analysis in detail.

The Op. 20 string quartets of Joseph Haydn share many stylistic characteristics–for example the use of "sonata form", a typical classical dramatic form, in the first movement of each quartet. Yet despite the regularities among the quartets and between their first movements in particular, they contain many deviations from standard practice. Both global regularities and local idiosyncrasies contribute to the dramatic content of these pieces and make them prime targets for analysis.

The third quartet, in G minor, is particularly dramatic, containing prolonged periods of tension, metrical ambiguity, and various surprising moments. I used the above-described RNN model to calculate measures of entropy for each time step in the first movement of this quartet. I then compared these measures to features derived from a music theoretical analysis of the piece in terms of its formal and dramatic structure.

Training

All pieces on which the RNN were trained were encoded as MIDI files, with each instrument (two violins, viola, and cello) assigned to a different part and thus separately represented in the RNN's input and output layers. In total, 247 units were used to represent the input (pitch and rhythm units for all four parts separately, as well as the aggregate pitch content) and 177 units were used in the hidden layer. The back-propagation learning rate parameter was set at 0.0625 and time steps were set at sixteenth-note duration.

The LTM was trained on the first movements of all six quartets in Op. 20 (19006 total time steps). All pieces of the training set were transposed to either C-major or C-minor as appropriate to eliminate effects of absolute pitch (since the model uses a localist pitch representation). The LTM was trained in cycles, during each of which it was trained on all six training pieces in random order. Training continued until mean accuracy–defined as the mean probability assigned to each time step in the training pieces–did not change by more than 0.0001 for 10 consecutive cycles. In all, the LTM was trained for 2000 cycles and achieved a final accuracy over the entire training set of 0.276 (range: 0.179 to 0.383).

The STM was trained on only the first movement of Op. 20, No. 3 (4332 total time steps). Using the same stopping criterion, the STM was presented with this movement 2500 times and achieved a final accuracy of 0.751. The combined LTM and STM models, which produced the output analyzed below, achieved an accuracy of 0.456 on the movement.

Simulations were also conducted which varies the number of hidden units, learning rate, and size of the LTM training corpus (for example, by including a wider selection of Haydn string quartet movements from Op. 17). The only major effect of these variations was that accuracy was improved with additional hidden units, but the form of the entropy profiles remained the same; specifically, major points of inflection were all at the same place and in the same direction.

Analysis of Entropy Profiles

The pitch and rhythmic entropy profiles derived from the combined STM and LTM are shown in Figure 2. Only the first repeat of the exposition (the first section of a sonata form piece; mm. 1-94) is shown, as this will be the focus of the subsequent analysis. Lacking a principled method of integrating pitch and rhythmic entropy, they are here considered separately, although both are assumed to contribute to a listener's subjective sense of tension. To enable the analysis of trends in the entropy measures, they were smoothed by convolving the raw entropy measures with an exponentially-decaying impulse response filter with weights $\psi(t) = e^{-\lambda t}$, where decay constant $\lambda = \frac{1}{32}$ corresponds to a mean lifetime of four measures (32 time-steps). Thus, the values shown in Figure 2 represent a "memory" of the instantaneous entropy that emphasizes the last four measures. The following analysis owes much to the work of Drabkin (1999), particularly pp. 105-111. Additional analytical material may be found in Grave and Grave (2006), especially pp. 190-192.

The only perfect authentic cadence in the exposition occurs at the end of the first phrase in m. 7, where there is a clear local minimum in pitch entropy as well as a low plateau in rhythmic entropy[4]. Mm. 8-26 effect a modulation from the home key of G minor to its relative major, B♭, all the while increasing the tension for a strong resolution to a B♭ harmony. This increase in tension is mirrored by increasing pitch and rhythmic entropy, where pitch entropy reaches a local maximum on the second beat of m. 24 with the introduction of a novel unison figure that prolongs the tension until the B♭ resolution in m. 27.

The second theme group (mm. 27-40) maintains a consistent pitch entropy while rhythmic entropy builds until the cello's eighth-note pulse disappears in m. 34, leaving just a high violin melody with the other instruments holding chords in long rhythmic values. The decrease in rhythmic entropy is

[4]In simulations with Bach chorales (not reported here), resolutions of authentic cadences also correspond to local minima in entropy measures while deceptive cadences produce no change or an increase in entropy.

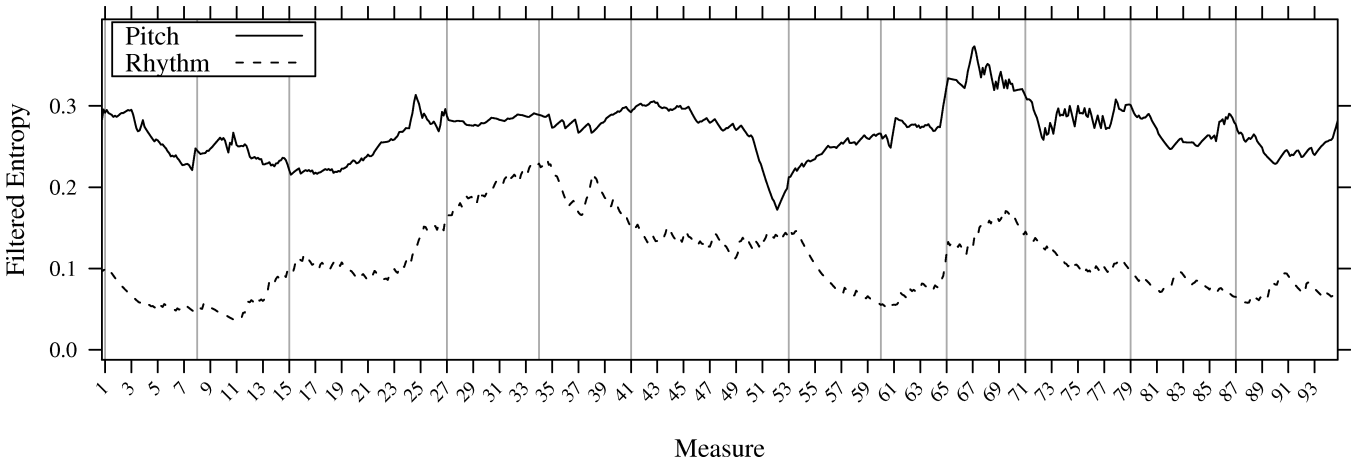


Figure 2: Ensemble pitch and rhythm entropy profiles for the exposition (mm. 1-94) of Haydn Op. 20, No. 3, first movement.

interrupted in mm. 37-38, when pitch changes become staggered between the instruments. Mm. 41-52 are more transitory and fragmented, with a large drop in pitch entropy during the violin solo in mm. 50-51 (greater certainty arising from predicting fewer separate parts). The long, regular rhythmic durations of mm. 53-59 continue the drop in rhythmic entropy while the chromatic harmonies increase pitch entropy until a break is reached at a deceptive cadence in m. 60.

This is followed by an F-major statement in mm. 61-64, then in mm. 65-66 by an "utter non sequitur–a fortissimo fanfare, poised on a first-inversion B♭ triad, with no compelling relationship to the immediately preceding or following material" (Grave & Grave, 2006, p. 190). This surprising event is naturally accompanied by a spike in both rhythmic and pitch entropy. Contrary to what might be implied by the B♭ fanfare–strong thematic material emphasizing the new key of B♭–we are instead treated in mm. 67-70 to the opposite: a softer, tonally ambiguous reprise of mm. 61-64. Mm. 67-70 are at a softer dynamic, played by solo violin instead of the entire quartet, and in a more restricted and chromatic melodic range. This unexpected consequent is assigned the highest pitch entropy in the entire exposition.

Rhythmic entropy continues to build until a resting point is reached at m. 70 on an unclear tonality. The succeeding violin solo and its accompaniment in mm. 71-77 is metrically ambiguous, implying a triple meter when in fact the duple meter still prevails. In this instance, the gradually diminishing ensemble rhythmic entropy is not in accord with this ambiguity, which should result in a higher rhythmic entropy for this passage. However, the rhythmic entropy of the individual parts, shown averaged in Figure 3, does show the expected staggered increase from mm. 71-77.

The remainder of the exposition is on more solid tonal and metrical footing. Of particular interest is the jump in pitch entropy in mm. 85-86, corresponding to another instance of the unison figure from mm. 24-25 and serving the same purpose–to prolong tension before before reaching a harmonic resolution–and producing the same effect on the entropy profile–an increase in pitch entropy whilst rhythmic

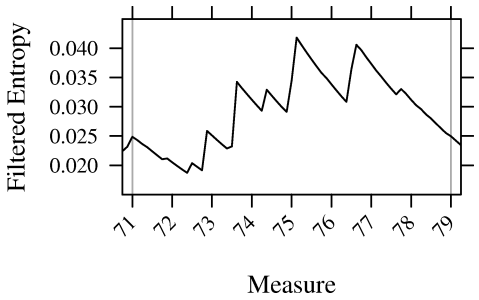


Figure 3: Rhythmic entropy averaged between parts for mm. 71-79 of Haydn Op. 20, No. 3, first movement.

entropy is unaffected. Although rhythmic entropy reaches a local maximum at m. 90 and begins to fall after a constant eighth-note pulse is established in the violins, pitch entropy increases toward the end of the exposition, reflecting the fact that the end of the exposition can be followed by either a repeat of the exposition (return to m. 1) or the start of the next section. In both cases, the rhythmic surface is the same, but the pitches are different and are assigned to different instruments, thus it is logical that there would be more uncertainty about pitch than rhythm at the end of the exposition.

Discussion

Analyses like the one presented above show that entropy derived from a predictive model of music can correspond to dramatically important features of music. Specifically, the entropy measures employed are sensitive to the calming effect of cadences (m. 7), the build-up of tension prior to resolutions (mm. 8-26), differential effects of textural change (mm. 27-60), and the shocking effects of interruptions (mm. 24-25, 85-86) and their consequents (mm. 65-70). Because a listener's subjective sense of tension is also affected by these features, this suggests a relationship between entropy and tension–and thus, perhaps, to musical meaning.

It is, perhaps, remarkable that such a relationship may be found, given the limitations of the current model. The model includes no information about dynamics, timbre, and expressive timing. A more realistic pitch representation, while in-

creasing the model's complexity, might also improve its performance (Mozer, 1994). Further, the use of a RNN at all imposes severe limitations on the approach outlined in this paper. While RNNs enable the analysis of music that is not amenable to other modeling techniques, they are slow to train, limited in the size of the corpus on which they can be trained, and, in the form presented here, cannot generalize to other ensemble types. The application of computational cognitive models to music is still in its infancy, and future research is sure to improve upon the techniques explored thus far. Future work must also compare model-derived entropy measures with human tension judgments (as in Krumhansl, 1996). This will elaborate on the relationship between entropy and tension, including the contributions of different sources of uncertainty (e.g., pitch and rhythm) to overall tension.

Even given the limited state of our current knowledge, it is possible to show that meaningful musical features correlate with features of musical entropy, given an appropriate predictive model. If human listeners have a similar predictive model "in mind"–consciously or not–as they listen to music, this provides great insight into the nature of music cognition and creation. The reasons why certain patterns recur within a style and that listeners have consistent responses to those patterns and violations thereof are not arbitrary–they can be understood in terms of prediction and uncertainty. With the advent of formal cognitive models, we can leverage this principle to better understand music that resists conventional analysis, for example, styles with few examples (e.g., the oeuvre of many idiosyncratic modern composers) or for which there is insufficient access to primary sources (e.g., historical and ethnomusicological studies). While more sophisticated methods will allow us to better elucidate the nature of entropy in music, it is clear that Meyer's (1956) thesis is still a viable approach to understanding the nature of meaning in music.

Acknowledgments

The author wishes to thank Peter Fontana, James Fry, Dora Hanninen, Michael Jones, and Richard Shiffrin for their guidance, suggestions, and support, as well as several anonymous reviewers.

References

Abdallah, S., & Plumbley, M. (2009). Information dynamics: Patterns of expectation and surprise in the perception of music. *Connection Science*, *21*(2), 89–117.

Bharucha, J. J., & Todd, P. M. (1989). Modeling the perception of tonal structure with neural nets. *Computer Music Journal*, *13*(4), 44–53.

Conklin, D., & Witten, I. H. (1995). Multiple viewpoint systems for music prediction. *Journal of New Music Research*, *24*(1), 51–73.

Drabkin, W. (1999). *A reader's guide to Haydn's early string quartets*. Westport, CT: Greenwood Press.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*, 179–211.

Grave, F., & Grave, M. (2006). *The string quartets of Joseph Haydn*. New York: Oxford University Press.

Hiller, L., & Fuller, R. (1967). Structure and information in Webern's Symphonie, Op. 21. *Journal of Music Theory*, *11*(1), 60–115.

Justus, T. C., & Bharucha, J. J. (2001). Modularity in musical processing: The automaticity of harmonic priming. *Journal of Experimental Psychology: Human Perception and Performance*, *27*(4), 1000–1011.

Knopoff, L., & Hutchinson, W. (1981). Information theory for musical continua. *Journal of Music Theory*, *25*(1), 17–44.

Krumhansl, C. L. (1996). A perceptual analysis of Mozart's piano sonata K. 282: Segmentation, tension, and musical ideas. *Music Perception*, *13*(3), 401–432.

Manzara, L. C., Witten, I. H., & James, M. (1992). On the entropy of music: An experiment with Bach chorale melodies. *Leonardo Music Journal*, *2*(1), 81–88.

Margulis, E. H., & Beatty, A. P. (2008). Musical style, psychoaesthetics, and prospects for entropy as an analytic tool. *Computer Music Journal*, *32*(4), 64–78.

Meyer, L. B. (1956). *Emotion and meaning in music*. Chicago: University of Chicago Press.

Meyer, L. B. (1957). Meaning in music and information theory. *The Journal of Aesthetics and Art Criticism*, *15*(4), 412–424.

Mozer, M. C. (1992). Induction of multiscale temporal structure. In J. E. Moody, S. J. Hanson, & R. P. Lippmann (Eds.), *Advances in neural information processing systems IV* (pp. 275–282). San Mateo, CA: Morgan Kaufmann.

Mozer, M. C. (1994). Neural network music composition by prediction: Exploring the benefits of psychoacoustic constraints and multi-scale processing. *Connection Science*, *6*, 247–280.

Pearce, M. T., Conklin, D., & Wiggins, G. A. (2005). Methods for combining statistical models of music. In *Computer music modeling and retrieval.* Berlin / Heidelberg: Springer.

Pearce, M. T., & Wiggins, G. A. (2004). Improved methods for statistical modelling of monophonic music. *Journal of New Music Research*, *33*(4), 367–385.

Potter, K., Wiggins, G. A., & Pearce, M. T. (2007). Towards greater objectivity in music theory: Information-dynamic analysis of minimalist music. *Musicae Scientiae*, *11*(2), 295–322.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart, J. L. McClelland, & PDP Research Group (Eds.), *PDP*. Cambridge, MA: The MIT Press.

Steinbeis, N., Koelsch, S., & Sloboda, J. A. (2006). The role of harmonic expectancy violations in musical emotions: Evidence from subjective, physiological, and neural responses. *Journal of Cognitive Neuroscience*, *18*(8), 1380–1393.

Todd, P. M. (1989). A connectionist approach to algorithmic composition. *Computer Music Journal*, *13*(4), 27–43.

Simulating Cognitive Coping Strategies for Intelligent Support Agents

Azizi Ab Aziz (mraaziz@few.vu.nl)
Michel C.A. Klein (michel.klein@few.vu.nl)
Jan Treur (treur@few.vu.nl)
Agent Systems Research Group, Department of Artificial Intelligence
Vrije Universiteit Amsterdam, De Boelelaan 1081a,
1081 HV Amsterdam, The Netherlands

Abstract

People react differently to stress. According to the Cognitive Motivational Relational Theory by Lazarus and Folkman, the appraisal of stress and the emotions related to it determine whether people cope with stress by focussing on altering the situation (problem focussed) or on changing the emotional consequences of the events (emotion focussed). These different coping strategies have different effects on the long term. The coping process can be described in a formal dynamic model. Simulations using this model show that problem focussed coping leads to better coping skills and higher decrease of long-term stress than emotion focussed coping. These results also follow from a mathematical analysis of the model. The presented model can form the basis of an intelligent support system that uses a simulation of cognitive processes in humans in stressful conditions.

Keywords: virtual human agent model; stress; cognitive and behavioral modeling; temporal dynamics.

Introduction

Stress is simply a reality of nature where forces from the outside world affecting the individual. It comes in many forms and affects people of all ages and all walks of life. The individual responds to stress in ways that affect the individual as well as their environment. Hence, all living creatures are in a constant interchange with their surroundings, either physically or behaviorally. In general, stress is generally considered as being synonymous with distress and dictionaries defined it as "physical, mental, or emotional strain or tension" or "a condition or feeling experienced when a person perceives that demands exceed the personal and social resources the individual is able to mobilize" (Beck, 1987; Folkman, 1984).

However, human has its own mechanism to adapt with this adversity. Through a process known as coping, our cognitive skill will evaluate the situation mentally. If the situation is threatening, then the human will decide how to deal with the situation, and what skills can be used. If the demands of the situation outweigh the resources human has, then it will be labeled as "stressful" and he or she will react with the classical stress response and vice versa (Carver et al., 1989). It is essential to consider that everyone sees situations differently and has different coping skills. For this reason, no two people will respond exactly the same way to a given situation. Understanding this coping ability is an essential ingredient for developing a software agent that is capable of providing the right intervention towards stressed individuals (Aziz et al., 2010). Therefore there is a need for a virtual human agent model that has this capability. In this paper, virtual human agents are computer model of people that can be used as substitutes for "the real person" in a virtual environment, with a specific focus on simulating human coping behaviors during the formation of stressful events. Although there has been several work in computational models of human stress, little work has been done in modeling coping strategies, with a few exceptions in (Marsella and Gratch , 2003; Marsella et al., 2009).

This paper focuses exclusively on the formal model for dynamics in coping process, as it is one of the essential components in the development of a software agent that is able to monitor individuals' conditions during stressful events (Aziz & Treur, 2009). In the next section, the underlying principles in coping during stress are discussed (Section 2). From this perspective, a formal model is designed and formulated (Section 3). Later, in Section 4, simulation traces are presented to illustrate how this model satisfies the expected outcomes in long-term stress. In Section 5, a detailed mathematical analysis is performed, to identify equilibria in the model. Finally, Section 6 concludes the paper

Underlying Concepts in Coping

The cognitive theory that governs the underlying principle of this work is based on Cognitive Motivational Relational Theory (CMRT) as in Lazarus and Folkman (1984). This theory explains the role of distinctive positive and negative emotions in the stress appraisal process. Essentially, it conceptualized a transactional process in which the person and the environment are viewed as being in a dynamic and bidirectional relationship, where the essence of cognitive appraisal and coping provides a critical mediator between stressful person-environment and health outcomes.

Dynamics in Cognitive Appraisal Process and Coping Strategies

The cognitive approach to coping is based on a mental process of how the individual appraises the situation. Cognitive appraisal can be viewed as the evaluation of the

significance of what is happening in the person-environment relationship (Lazarus, 1991). Normally, it is also related to the *intensity of the stressful events*, a condition where several factors such as *situational demands* (pressure), *personal resources* (i.e; support), and *negative events* play important roles (Aziz et al., 2009; Lazarus & Folkman, 1984). Having the stressful events in motion, individual appraises two types of appraisals; the primary and the secondary. The primary appraisal is made when the individual makes a conscious evaluation of the matter at hand of whether it is a sense of harm or a loss, a threat or a challenge. It is an evaluation process of what is at stake for a person's well being. From this first process, the situation can be appraised either as harm/loss, threatening, challenging or benign (Folkman et al., 1986). *Harm* or loss refers to a condition where damage has already occurred, while *threat* refers to damage, but an anticipated one (*imminence of harm*) and it is more to a risk assessment part (Kessler, 1997). *Challenging* differs from threat in term of how persons are viewing it where it has a positive tone compared to threat. When stressful events were appraised as irrelevant or as *benign*, it will offer the chance to preserve or enhance wellbeing as it does not initiate the stress process as there is no potential threat to overcome. In addition, this appraisal process also involves an array of *personality attributes* such as values, commitments, and beliefs about oneself and the environment in defining the condition that the individuals are facing through (Uehara et al., 1999). Later this process will determine individuals' emotion perception; *negative, positive* or *neutral* emotion (Folkman, 1984). Negative emotion is related to perceiving *harm* and *threat*, while position emotion is attributed to perceiving *challenge* (Lazarus, 1991). Neutral emotion is triggered when individual perceives the condition as *benign* (Noh, 2003).

In the second appraisal, the persons evaluate whether they have the resources to deal with the incoming stressors. It is commonly related to the emotional attribution, where a positive and neutral emotion results in *acceptance* and *change*, while the negative emotion triggers *holdback* behavior (Lazarus, 1991). During this stage, several coping strategies are evaluated. Coping strategies refer to the specific efforts, both behavioral and psychological, that people employ to either be in charge of, tolerate, reduce, or minimize stressful events. According to the CMRT model, there are two types of coping strategies have been distinguished, namely; *problem-focused coping* and *emotion-focused coping*. A problem-focused coping is associated with aggressive interpersonal efforts to alter the situation, as well as rational efforts to get the problem solved (Carver et al., 1989). Contrary to this, emotion-focused coping strategies (thinking rather than acting to change the person-environment relationship) entail efforts to regulate the emotional consequences of stressful or potentially stressful events (Pruchno & Resch, 1989). It is typically include distancing, escape avoidance, and seeking for social comforts.

Several findings showed that the type of coping strategies can be derived, depending on what was at stake (primary appraisal) and what the coping options were (secondary appraisal) (Lazarus, 1991; Ntoumanis et al. 2009). It means, when people feel that they are capable of changing the situation into something better (high perception of acceptance and change), and then a problem-focused coping is chosen. In contrast, when the conditions are considered not amenable to change (high perception in holdback) then emotion-focused coping is used. In addition to this, problem focused coping strategies may give an individual greater perceived control over their problem, while emotion focused coping strategies may more often lead to a reduction of control over the perceived events. All these strategies can be proven useful, but many individuals feel that problem-focused coping represent a more effective means of coping in adversities (Uehara, 1999). In addition to this, in a long run, emotion focused coping is associated with outcomes that people found unsatisfactory (*exhaustion in coping*) that later will increase long-term stress, and problem focused coping is associated with satisfactory outcomes (*improved coping skills*) (Clarke & Tanya, 2009). Furthermore, in psychological distress, problem focus coping strategies appear reliably to produce better emotional adjustment to chronically stressful events than do emotional focused strategies (Pruchno & Resch, 1989; Uehara, 1999).

In short, the following dynamics can be identified from the literature; *(1)* the intensity of the stressful events will lead to coping appraisal, *(2)* the perception of event regulates emotional attribution, *(3)* the emotional attribution will trigger a coping strategy, *(4)* a long-term overwhelming dependency in emotion-focused coping will lead to the exhaustion in coping, and *(5)* a problem-focused coping will improve the coping ability.

The Virtual Human Agent Model

Based on the analysis of the cognitive dynamics in coping appraisal and strategy as given in the previous section, it is possible to specify computational properties for the virtual human agent model. These computational properties are represented in a way that allows simulating how an individual is coping when experiencing stressors, and what are the consequences of that action. All of these concepts (and their interactions) are discussed in the following paragraphs in this section.

Formalizing the Cognitive Model Relationships

In the formalization, the dynamic concepts discussed in the previous section are translated into several interconnected nodes. Figure 1 depicts the global interaction between these nodes. The nodes are represented as variables that can have values ranging from 0 (low) to 1 (high). The interaction will determine the new value of it, either by a series of accumulations or an instantaneous interaction for each node.

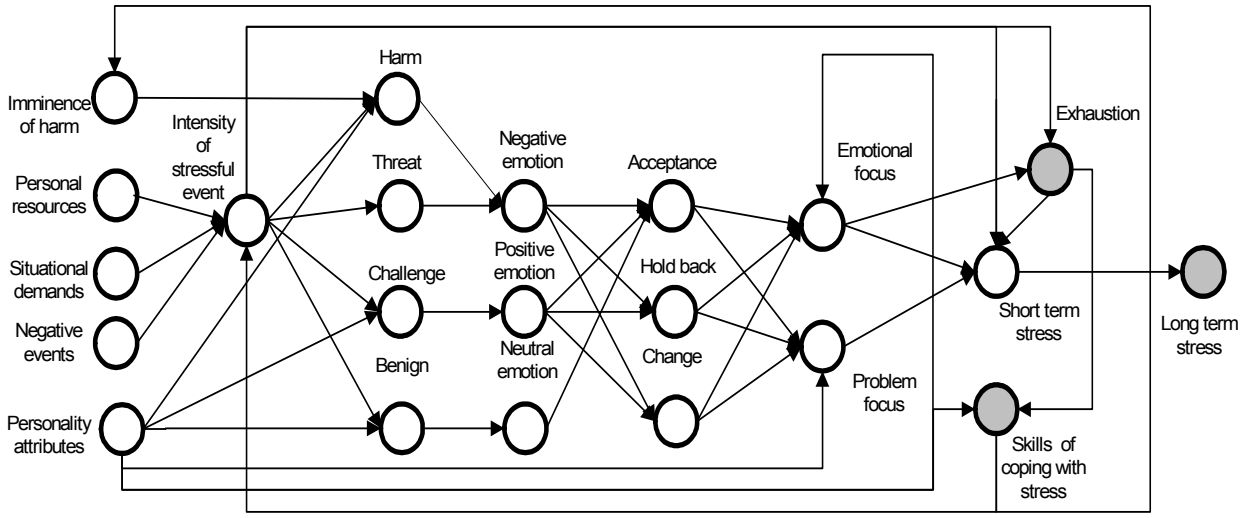


Figure 1: Global relationships of variables involved in the coping process

The description of these formalizations is described in the following. Together, this results in a dynamic model. This model involves a number of instantaneous and some temporal relations. The dark nodes represent concepts that have temporal relationships with the incoming nodes, in which the change is specified for a time interval between t and $t+\Delta t$

Stressor Events, Intensity of Stressful Event, and Imminence of Harm In the model, the stressor events *(e)* (negative events) are generated by simulating potential effects throughout t time using w weighted sum of three types of events; life (*le*), chronic (*ce*), and daily (*de*) events. The role of these factors in the model is to represent a series of events. The intensity of stressful event (*IsE*) represents the degree of stress encountered by a person related to his or her situational demands (*SiD*), and stressor events (*NeVt*), regulated by the proportion factor β_e. In addition, the intensity of a stressful event will be reduced if the coping skills (*ScS*) and personal resources (*PeS*) are high. Imminence of harm (*ImH*) can be measured by combining both concepts in perceived harm (*PeH*) (from the environment), and coping skills (*ScS*).

$$NeV(t) = w_1.le(t) + w_2.ce(t) + w_3.de(t), \ \sum w=1 \quad (1)$$

$$IsE(t) = [\beta_e.NeV(t) + (1-\beta_e).SiD(t)].(1-ScS(t)).(1-PeS(t)) \quad (2)$$

$$ImH(t) = PeH(t).(1-ScS(t)) \quad (3)$$

Harm, Threat, Challenge, and Benign The level of harm (*HrM*) is determined by the proportional contribution φ_h on the imminence of harm, and intensity of the stressful event. The intensity of the stressful event also related to threat (*ThT*). For both cases, in harm and threat, there is a negative relation with personality attributes. On the contrary, challenge (*ChL*) and benign (*BnG*) are positively related with good personality attributes (*PrA*), and negatively with the intensity of stress. Here parameters α_c and ψ_b represent the proportional factor for both challenge and benign respectively.

$$HrM(t) = [\varphi_h.ImH(t) + (1-\varphi_h).IsE(t).ImH(t)].(1-PrA(t)) \quad (4)$$

$$ThT(t) = IsE(t).(1-PrA(t)) \quad (5)$$

$$ChL(t) = \alpha_c.PrA(t) + (1-\alpha_c).(1-IsE(t)).PrA(t) \quad (6)$$

$$BnG(t) = \psi_b.(1-IsE(t)) + (1-\psi_b).PrA(t) \quad (7)$$

Negative, Neutral, and Positive Emotion When the harm and threat is perceived, a fraction from those two parts (by a proportional factor β_n) is contributed as a negative emotion (*NgE*). The notion of positive (*PsE*) and neutral (*NuE*) emotion is represented through a proportional factor of τ_p in challenge and ρ_e in benign respectively.

$$NgE(t) = \beta_n.HrM(t) + (1-\beta_n).ThT(t) \quad (8)$$

$$PsE(t) = \tau_p.ChL(t) \quad (9)$$

$$NuE(t) = \rho_e.BnG(t) \quad (10)$$

Acceptance, Holdback, and Change Positive and neutral emotion increases the acceptance (*AcP*) level by a proportional factor γ_a, while negative emotion works in a opposite way. Holdback (*HdB*) depends on the relation between negative and positive emotion. Change (*ChG*) uses the same concepts as in holdback but with the opposite relation.

$$AcP(t) = \gamma_a.PsE(t) + (1-\gamma_a).NuE(t).(1-NgE(t)) \quad (11)$$

$$HdB(t) = (1-PsE(t)).(NgE(t)) \quad (12)$$

$$ChG(t) = PsE(t)).(1-NgE(t)) \quad (13)$$

Emotional and Problem Focused Coping Emotional focused coping (*EmF*) is determined using the presence of acceptance, holdback and change. Using this relation, emotion focused coping decreases when either acceptance or change increases. However in problem focused coping (*PrF*), coupled with personality attributes, those factors

provide a positive effect. Parameters η_e and γ_p regulate the contribution preferences for both specifications respectively.

$$EmF(t) = [\eta_e.(1-AcP(t)).HdB(t) + (1-\eta_e).HdB(t)].(1-ChG(t)) \quad (14)$$

$$PrF(t) = [\gamma_p.PrA(t) + (1-\gamma_p).AcP(t)].(1-HdB(t)).ChG(t) \quad (15)$$

Short-term stress, Long-term stress, Exhaustion, and Coping Skills The notion of short-term stress (*StS*) models a relation between coping styles (regulated by μ_s), and a combination of exhaustion and intensity in stressful events (regulated by a proportional rate γ_s) and will influence the level of long-term stress (*LtS*) in a long run. The formation of exhaustion (*ExH*) is modelled using the presence of emotion-focused coping and the intensity of stressful events. The level of coping skills (*ScS*) is influenced by the exhaustion and personality attributes. The rates of change for all temporal relationships are determined by flexibility parameters β_{lts}, ψ_e, and ϕ_s respectively.

$$StS(t)=[1-(\mu_s.EmF(t)+(1-\mu_s).PrF(t))].(\gamma_s.ExH(t)+(1-\gamma_s).IsE(t)) \quad (16)$$

$$LtS(t+\Delta)t=LtS(t)+\beta_{lts}.[\text{Pos}(StS(t)-LtS(t)).(1-LtS(t))- \text{Pos}(-(StS(t)-LtS(t))).LtS(t)].\Delta t \quad (17)$$

$$ExH(t+\Delta t)=ExH(t)+\psi_e.[(\text{Pos}((IsE(t)-ExH(t)).(1-ExH(t)))- \text{Pos}(-(IsE(t)-xH(t)).ExH(t))].EmF(t).\Delta t \quad (18)$$

$$ScS(t+\Delta t)=ScS(t) + \phi_s.[\text{Pos}(ExH(t) - ScS(t)).(1-ScS(t))- \text{Pos}(-(ExH(t) -ScS(t)).ScS(t)].PrA(t).\Delta t \quad (19)$$

The operator Pos for the positive part is defined by $\text{Pos}(x) = (x + |x|)/2$, or, alternatively; $\text{Pos}(x) = x$ if $x \geq 0$ and 0 else.

Example Simulation Traces

In this section, the virtual human agent model of coping has been executed to simulate a number of scenarios with a variety of different conditions of individuals. Two example scenarios are shown: an individual with a tendency to choose problem focused coping (***A***), and an individual with a tendency to choose emotional focused coping ***(B)***. The initial settings for the different individuals are the following (*PrA, PeH, SiD, PeS*); ***A*** (*0.8, 0.5, 0.5, 0.8*), and ***B*** (*0.2, 0.5, 0.8, 0.1*). In all cases, the long term stress, exhaustion, and coping skill value are initialized at 0.3.

Corresponding to these settings, the level of severity is set at 0.5, defining that any individuals scoring higher than 0.5 in their long-term stress and exhaustion levels will be considered as experiencing difficulties in coping. These simulations used the following parameters settings: $t_{max}=1000$ (to represent a monitoring activity up to 42 days), Δt=0.3, all proportional and flexibility rates are assigned as 0.5 and 0.9 respectively. These settings were obtained from several systematic experiments to determine the most suitable parameter values in the model.

Result # 1: Simulation Trace for Repeated Stressor Events During this simulation, each type of individual has been exposed to an extreme stream of stressor events, with a moderate alteration between each corresponding event. Figure 2 depicts the comparison between the conditions of individual *A* and *B* during repeated stressors. In this simulation trace, it is visible that individual *A* has developed better coping skills. For this reason, an individual A recovers much faster from long-term stress compared to other individuals.

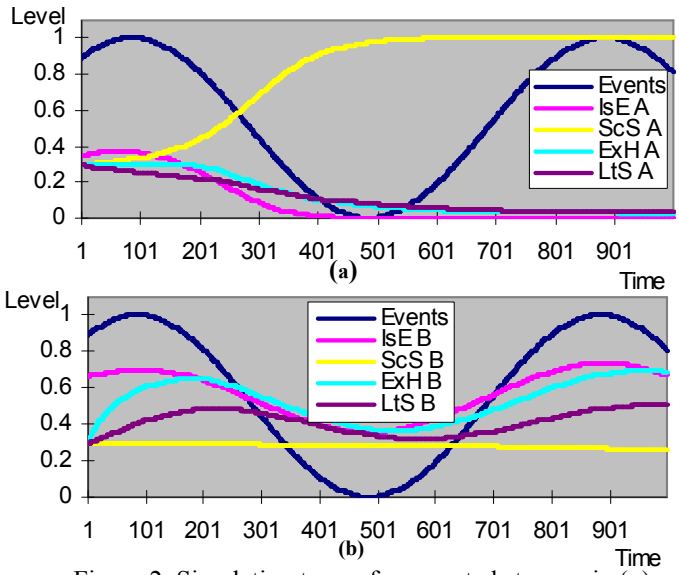


Figure 2. Simulation traces for repeated stressor in (**a**) individual *A* (**b**) individual *B*

Note that the individual *B* shows a repeated increasing pattern that may lead to potential long-term stress. As a consequence of this condition, an individual B will experience difficulty if that individual is having constant exposure towards stressors in a long run

Result # 2: Simulation Trace for Fluctuated Stressor Events This simulation trace shows two types of periods, one with a very high constant and with a very low constant stressor event. These events occurred in a constant behaviour for a certain period of time (approximately within 20 days).

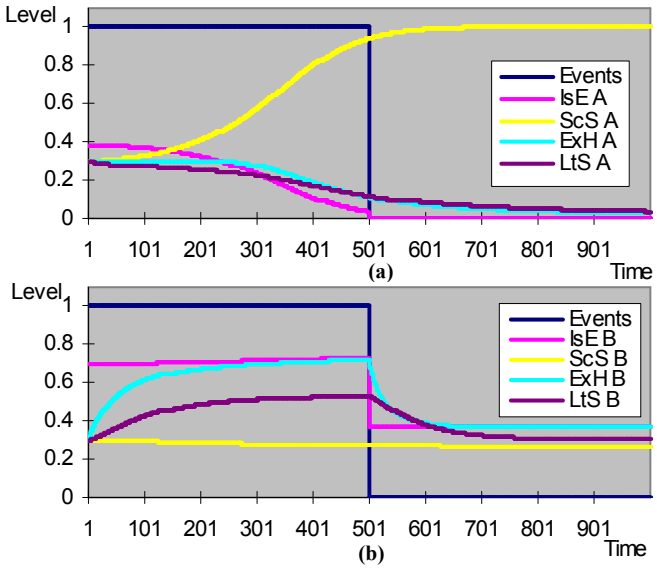


Figure 3. Simulation traces for fluctuated stressor in (**a**) individual *A* (**b**) individual *B*

Also here it can be seen (in Figure 3) that individual *B* gets into long-term stress much faster than individual *A*. Moreover, even at the end of the simulation time, the long term stress level of individual *B* is still slightly higher than individual *A*. Furthermore, in contrast with individual *B,* individual *A* has his/her coping skills improved throughout time.

Mathematical Verification

This section addresses the formal analysis of the agent model and the simulation results presented above by means of a mathematical analysis of the equilibria of the model. The equilibria describe situations in which a stable situation has been reached. Those equilibria are interesting as it should be possible to explain them using the knowledge of the domain that is modelled [2]. As such, the existence of reasonable equilibria is an indication for the correctness of the model. To analyze the equilibria, the available temporal and instantaneous equations are filled with values for the model variables such that the derivatives or differences between time point t and $t + \Delta t$ are all 0. The dynamic part of the model written in differential equation format is as follows:

$$dLtS(t)/dt = \beta_{lts}.[\ \mathrm{Pos}(StS(t)\text{-}LtS(t)).(1\text{-}LtS(t))\text{-} \mathrm{Pos}(\text{-}(StS(t)\text{-}LtS(t))).LtS(t)] \quad (20)$$

$$dExH(t)/dt = \psi_e.[\mathrm{Pos}(IsE(t)\text{-}ExH(t)).(1\text{-}xH(t))\text{-} \mathrm{Pos}(\text{-}(IsE(t)\text{-}\ ExH(t))).ExH(t)].EmF(t) \quad (21)$$

$$dScS(t)/dt = \phi_s.[\mathrm{Pos}(ExH(t) - ScS(t)).\ (1\text{-}ScS(t))\text{-} \mathrm{Pos}(\text{-}(ExH(t) - ScS(t)).ScS(t)\ ].PrA(t) \quad (22)$$

For an equilibrium it has to hold that all of the derivatives are zero:

$$dLtS(t)/dt = d\ ExH(t)/dt = d\ ScS(t)/dt = 0$$

Assuming β_{lts}, ψ_e and ϕ_s nonzero, this provides the following equilibrium equations:

$$\mathrm{Pos}(StS\text{-}LtS).(1\text{-}LtS)\text{-}\mathrm{Pos}(\text{-}(StS\text{-}LtS)).LtS = 0\ . \quad (23)$$

$$[\mathrm{Pos}(IsE\text{-}ExH).(1\text{-}ExH)\text{-}\mathrm{Pos}(\text{-}(IsE\text{-}ExH)).ExH]. \quad (24)$$
$$EmF = 0$$

$$[\mathrm{Pos}(ExH - ScS).\ (1\text{-}ScS)\text{-}\mathrm{Pos}(\text{-}(ExH - ScS)).\ ScS]. \quad (25)$$
$$PrA = 0$$

Table 1 shows which cases can be distinguished. For example, notice that always Pos(x) ≥ 0, so (23) is equivalent to;

$$\mathrm{Pos}(StS\text{-}LtS).(1\text{-}LtS) = 0$$
$$\mathrm{Pos}(\text{-}(StS\text{-}LtS)).LtS = 0$$

This provides cases;

$$(StS \le LtS\ \vee\ LtS = 1)\ \wedge\ (StS \ge LtS\ \vee\ LtS = 0) \quad (26)$$

This can be logically rewritten into;

$$(StS \le LtS\ \wedge\ StS \ge LtS)\ \vee\ (StS \le LtS\ \wedge\ LtS = 0)\ \vee$$
$$(LtS = 1\ \wedge\ StS \ge LtS)\ \ \vee\ \ (LtS = 1\ \wedge\ LtS = 0)$$

The latter case cannot exist, and as $0 \le StS \le 1$ the other three cases are equivalent to $StS = LtS$. Similarly the cases for (24) and (25) can be found as shown in Table 1.

Table 1: Equilibrium Equations

(1)	(2)	(3)	Combined
StS = LtS	*EmF =0*	*PrA = 0*	*StS = LtS , EmF = PrA = 0*
		ExH = ScS	*StS = LtS , EmF = 0, ExH = ScS*
	IsE = ExH	*PrA = 0*	*StS = LtS, IsE = ExH, PrA = 0*
		ExH = ScS	*StS = LtS, IsE = ExH = ScS*

Note that for each of the distinguished cases, further information can be found about the equilibrium values of other variables using the other non-dynamic-equations. For example, from *EmF = 0* by (14) it follows that *ChG = 1* or *HdB = 0*. This condition illustrates the generic condition that a problem-focused individual that encounters stressful events will never develop long term stress that typically caused by a prolonged dependency on emotion-focused focus coping (Aziz & Treur, 2009; Ntoumanis et al, 2009; Pruchno & Resch, 1989). From another condition *PrA = 0,* by (6) it follows that *ChL* = 0 represents a condition when an individual with negative personality attributes tend to appraise stressful events not as a challenge later will trigger emotion-focused coping (Clarke & Tanya, 2009; Uehara et

al. 1999). Both of these conditions can be found in our simulation results.

Conclusion

In this paper, we have presented a formal temporal model for the cognitive process of coping with stress as described in the informal Cognitive Motivational Relational Theory by Lazarus and Folkman. This theory explains the role of positive and negative emotions in the stress appraisal process, which results in either a problem focused coping strategy or an emotional focused coping strategy. The theory also describes the effect of the different strategies on the long term stress.

The resulting model has been used for two simulations of two persons with different personality characteristics in two different scenarios that describe the level of external sources of stress over time. The simulation traces exhibit patterns that are expected in this domain: problem focused coping leads to better coping skills and higher decrease of long-term stress than emotion focused coping. These results also follow from a mathematical analysis of the model, in which the equilibria of the model are determined to identify the stable situation in the model.

The resulting model can be considered as a virtual human agent model, in the sense that it is a computer models of a person that can be used as a substitute for the real person in a virtual environment. This could provide the basis for a intelligent support system, in which the system should be able to understand the coping process of the persons to which support is provided.

References

Aziz, A. A., Klein, M.C.A., & Treur, J. (2010). An Integrative Ambient Agent Model for Unipolar Relapse Depression. *Journal of Ambient Intelligence and Smart Environments* Vol 2 (1), pp. 5-20: IOS Press .

Aziz, A. A., & Treur, J. (2009). Modelling Dynamics of Social Support Networks for Mutual Support in Coping with Stress. In: Nguyen, N.T., Katarzyniak, R., Janiak, A. (eds.). *Challenges in Computational Collective Intelligence*, ICCCI' *09*, Studies in Computational Intelligence, pp. 167-179: Springer Verlag

Aziz, A. A., Klein, M.C.A., & Treur, J. (2009). An Agent Model of Temporal Dynamics in Relapse and Recurrence in Depression. In: Ali, M., Chen, S.M., Chien, B.C., Hong, T.P. (eds.), *IEA-AIE 2009*, pp. 36-45 :LNAI Springer Verlag.

Beck, A. T. (1987). Cognitive Model of Depression, *Journal of Cognitive Psychotherapy 1*, pp. 5-37.

Carver, C. S., Scheier, M. F., & Weintraub, J. K. (1989). Assessing Coping Strategies: A Theoretically Based Approach. *Journal of Personality and Social Psychology*, *56*, 267-283.

Clarke, D., & Tanya, G. (2009). The mediating effects of coping strategies in the relationship between automatic negative thoughts and depression in a clinical sample of diabetes patients, *Personality and Individual Differences, Vol. 46, Issue 4*, 460-464.

Folkman, S., Lazarus, R. S., Dunkel-Schetter, C., DeLongis, A., & Gruen, R. J. (1986). Dynamics of a stressful encounter: Cognitive appraisal, coping, and encounter outcomes. *Journal of Personality and Social Psychology, 50*, 992-1003.

Folkman, S.(1984). Personal Control, Stress and Coping Processes: A theoretical analysis. *Journal of Personality and Social Psychology, 46*, 839–852.

Kessler, R.C. (1997). The Effects of Stressful Life Events on Depression, *Annual Review of Psychology 48*, 191-214.

Lazarus, R. S., & Folkman, S. (1984*). Stress, Appraisal, and Coping*. New York: Springer.

Lazarus, R.S. (1991). *Emotion and Adaptation*. NY: Oxford University Press.

Marsella, S., and Gratch, J. (2003). Modeling Coping Behavior in Virtual Humans: Don't Worry, Be Happy, Proceedings of the 2nd International Joint Conference on Autonomous Agents and Multiagent Systems, 313-320.

Marsella, S., Gratch, J., Wang, N., & Stankovic, B. (2009). Assessing the Validity of a Computational Model of Emotional Coping. International Conference on Affective Computing and Intelligent Interaction, IEEE.

Noh, S., & Kaspar, V. (2003). Perceived Discrimination and Depression: Moderating Effects of Coping, Acculturation, and Ethnic Support, *Am J Public Health. 93(2),* 232–238.

Ntoumanis, N., Edmunds,J., & Duda, J.L. (2009). Understanding the coping process from a self-determination theory perspective, *British Journal of Health Psychology, 14*, 249–260.

Pruchno, R.M., & Resch, N.L.(1989). Husbands and Wives as Caregivers: Antecedents of Depression and Burden. *The Gerontologist 29*, 159-165.

Uehara T, Sakado K, Sakado M, Sato T, & Someya T. (1999). Relationship between Stress Coping and Personality in Patients with Major Depressive Disorder. *Psychother Psychosom ,68*, 26-30.

An Adaptive Integrative Ambient Agent Model to Intervene in the Dynamics of Beliefs and Emotions

Zulfiqar A. Memon [1,2], Jan Treur [1] ({zamemon, treur}@few.vu.nl)
[1] VU University Amsterdam, Department of Artificial Intelligence,
De Boelelaan 1081, 1081 HV Amsterdam, The Netherlands
[2] Sukkur Institute of Business Administration (Sukkur IBA),
Air Port Road Sukkur, Sindh, Pakistan

Abstract

In this paper an adaptive integrative ambient agent model is introduced incorporating estimation of a human's interactive dynamics of believing and feeling. The integrative agent model is equipped with a dynamical model which describes how the strength of a belief depends both on information obtained and emotional responses on the belief. In addition, the agent model integrates an adaptation model to tune parameter values representing personal characteristics. In a simple personalised case it is shown how the ambient agent model is able to assess a person's state and use this assessment to interact in a personalised manner.

Keywords: Integrative agent, believing, feeling, adaptive

Introduction

An important and interesting recent class of applications for software/hardware agents can be found in Ambient Intelligence: the area of ambient or pervasive systems; e.g., (Aarts, Collier, Loenen, Ruyter, 2003; Aarts, Harwig, Schuurmans, 2001; Riva, Vatalaro, Davide, Alcañiz, 2005). One of the more ambitious challenges in this area is to create ambient agents with an appropriate awareness of the (mental) states of humans. Human-aware ambient agent systems can be taken to perform a certain type of mindreading or to possess what in the psychological and philosophical literature is called a Theory of Mind; e.g., (Gärdenfors, 2003; Goldman, 2006). As developed during the evolutionary human history, mindreading addresses different types of mental states, such as intention, attention, belief or emotion states; e.g., see (Gärdenfors, 2003). Inspired by these facilities available in nature, ambient agent models can be developed that have mindreading capabilities for one or some of these types of mental states. However, it is more and more acknowledged that such mental states can be quite dynamic and often interact with each other intensively. To obtain an adequate ambient agent model, dynamical models describing such dynamics and interaction has to be integrated within the agent model.

Human-aware ambient agent systems equipped with the ability to reason about the different types of mental states can be applied in the area of personalised customer relationships and marketing. A recent trend is to dig deeper into the clients' minds and lives. The work reported here focuses on the dynamics and interaction of an individual client's beliefs and emotions and integrates models for these dynamics in an ambient agent model to provide effective intelligent marketing strategies by a better understanding of the cognitive and affective system of the client. In their generation process beliefs trigger emotional responses that result in certain feelings. In a reciprocal manner, the generated feelings affect the belief as well; for some literature on such reciprocal interactions between cognitive and affective astates, see, for example, (Eich, Kihlstrom, Bower, Forgas, and Niedenthal, 2000; Forgas, Goldenberg, and Unkelbach, 2009; Niedenthal, 2007; Schooler and Eich, 2000; Winkielman, Niedenthal, and Oberman, 2009).

In this paper, a computational dynamic model is adopted from (Memon and Treur, 2009) that models the client's reciprocal interaction between feeling and believing. This model is based on neurological theories on the embodiement of emotions as described, for example, in (Damasio, 1994, 1996, 1999, 2004; Winkielman, Niedenthal, and Oberman, 2009). More specifically, in accordance with, for example (Damasio, 1999, 2004), for feeling the emotion associated to a belief a converging recursive body loop is assumed. A second converging feedback loop introduced in the model, inspired the Somatic Marker Hypothesis (Damasio, 1994, 1996), involves the interaction back from the feeling to the belief.

This dynamical model is integrated within an ambient agent model to enable the agent to assess the strength of the belief and feeling, and to intervene when desired. As a personal characteristic represented by a parameter indicating a bias of the belief in a positive or negative direction, is hard to determine at forehand, the ambient agent is equipped with an adaptation model to adjust the value of this parameter over time. This results in an adaptive integrative agent model that learns to estimate the human's belief and feeling better and better over time.

To illustrate the model, the following example scenario is used. A person (client) develops strong (false) beliefs due to strong negative feelings (of insecurity) about a product offered by the bank, for example, buying bonds or shares. The ambient agent estimates the level of belief and feeling of the client related to this insecurity. When the client becomes too insecure, i.e., the emotion level goes above certain threshold, the ambient agent can take measures in order to achieve a reduction of the insecure feeling, e.g., by providing information that makes the client feel more secure.

In this paper, first in Section 2 the dynamical model for the interaction between belief and feeling is described. In Section 3 the ambient agent model is described which integrates the dynamical model. Section 4 describes the parameter adaptation model integrated within the agent. Section 5 presents some simulation results. Finally, Section 6 is a discussion.

Belief and Emotion

In this section a computational model for the interaction between believing and feeling is briefly discussed, as adopted from (Memon and Treur, 2009). As any mental state in a person, a belief state induces emotions felt within this person, as described by Damasio (1999; 2004, p. 93):

belief → preparation for bodily response → bodily response → sensing the bodily response → sensory representation of the bodily response → feeling

As a variation, an 'as if body loop' uses a direct causal relation: preparation for the bodily response → sensory representation of the bodily response; as a shortcut in the causal chain. The body loop (or as if body loop) is extended to a recursive body loop (or recursive as if body loop) by assuming that the preparation of the bodily response is also affected by the state of feeling the emotion: feeling → preparation for the bodily response; as an additional causal relation. Within the model used in this paper both the bodily response and the feeling are assigned a level, expressed by a number; for example, the strength of a smile and the extent of happiness.

Although beliefs in an idealised rational agent might only depend on informational sources, real life persons may, for example, have a more optimistic or pessimistic character and affect their beliefs accordingly. To model this a causal relation: feeling → belief; is added. Therefore two recursive loops result, as shown in Figure 1. From a neurological perspective the existence of a connection from feeling to belief may be considered plausible, as this may be developed based on a general Hebbian learning mechanism (Hebb, 1949; Bi and Poo, 2001) that strengthens connections between neurons that are activated simultaneously. Another type of support for a connection from feeling to belief can be found in Damasio's Somatic Marker Hypothesis; cf. (Damasio, 1994, 1996; Bechara and Damasio, 2004; Damasio, 2004). This is a theory on decision making which provides a central role to emotions felt. Each decision option induces (via an emotional response) a feeling which is used to mark the option. Usually the Somatic Marker Hypothesis is applied to provide endorsements or valuations for options for a person's actions. However, it may be considered plausible that such a mechanism is applicable to valuations of internal states such as beliefs as well.

The hybrid dynamic modelling language LEADSTO used subsumes qualitative and quantitative causal relationships, and dynamical systems; cf. (Bosse, Jonker, Meij and Treur, 2007). Within LEADSTO the temporal relation a → b denotes that when a state property a occurs, then after a certain time delay (which for each relation instance can be specified as any positive real number), state property b will occur. A dedicated software environment is available to support specification and simulation.

An overview of the model for believing and feeling is depicted in Figure 1. Note that the precise numerical relations between the indicated variables V shown are not expressed in this picture. The detailed specification of the model can be found in (Memon and Treur, 2009).

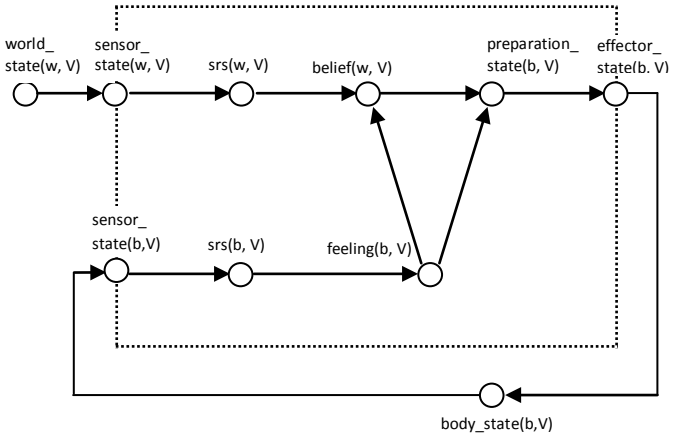


Figure 1: Dynamical model for belief and feeling

As an example, the dynamic property for the process for belief generation is described. The level for the belief is calculated based on a function $g(\beta, V_1, V_2)$ of the original levels, where β is the personal characteristic (with values from *0* to *1*) indicating positive or negatieve bias for the belief.

LP3 Generating a belief for a feeling and a sensory representation

If a sensory representation for w with level V_1 occurs,
and the associated feeling of b with level V_2 occurs
and the belief for w has level V_3
and β_1 is the person's orientation for believing
and γ_1 is the person's flexibility for beliefs
then a belief for w with level $V_3 + \gamma_1 (g(\beta_1, V_1, V_2)-V_3) \Delta t$ will occur.

has_state(human, srs(w, V_1)) &
has_state(human, feeling(b, V_2)) &
has_state(human, belief(w, V_3))
→ has_state(human, belief(w, $V_3 + \gamma_1 (g(\beta_1, V_1, V_2) - V_3) \Delta t$)

For the function $g(\beta, V_1, V_2)$ the following was taken:

$$g(\beta, V_1, V_2) = \beta(1-(1-V_1)(1-V_2)) + (1-\beta)V_1V_2$$

Dynamic property LP4 describes the emotional response to a belief in the form of the preparation for a specific bodily reaction. This dynamic property uses the same combination model based on $g(\beta, V_1, V_2)$ as above.

LP4 From belief and feeling to preparation of a body state

If belief w with level V_1 occurs
and feeling the associated body state b has level V_2
and the preparation state for b has level V_3
and β_2 is the person's orientation for emotional response
and γ_2 is the person's flexibility for bodily responses
then preparation state for body state b will occur with level $V_3 + \gamma_2 (g(\beta_2, V_1, V_2)-V_3) \Delta t$.

has_state(human, belief(w, V_1)) &
has_state(human, feeling(b, V_2)) &
has_state(human, preparation(b, V_3))
→ has_state(human, preparation(b, $V_3+\gamma_2 (g(\beta_2, V_1, V_2)-V_3) \Delta t$)

The Ambient Agent Model

Within the integrative ambient agent model, the model for the dynamics of belief and feeling is embedded in order to enable the agent to reason about this process, and to assess the person's beliefs and feelings. In psychology, this capability is often referred to as mindreading or Theory of Mind (e.g., Gärdenfors, 2003). The embedding uses the format that the causal relationships of the model described in

Section 2 above are transformed into relationships for beliefs of the ambient agent on mental states of the person. In order to achieve this, the idea of recursive modelling is used; e.g., (Marsella, Pynadath and Read, 2004). This means that the beliefs that agents have about each other are represented in a nested manner. Each mental state is parameterized with the name of the agent considered, thus creating concepts like

has_state(human, feeling(b, 0.5))
has_state(AA, performed(add_pos_info))

In addition, a number of meta-representations are introduced. For example, has_state(AA, belief(has_state(human, feeling(b, 0.7)))) states that the ambient agent (AA) believes that the human has a feeling level of *0.7* for b. The following are the resulting agent local properties (ALP) that specify the processes within the ambient agent. The first property specifies how the agent AA observes that the human senses external information.

ALP1 Observing the human's sensing external information
If the human senses external information,
then the ambient agent AA will observe this.

has_state(human, sensor_state(externalinfo, V))
→ has_state(AA, observed(
has_state(human, sensor_state(externalinfo, V))))

ALP2 Generating a belief for the human's sensing
If the ambient agent AA observes that the human senses an external information,
then it will generate a belief on it.

has_state(AA, observed(
has_state(human, sensor_state(externalinfo, V))))
→ has_state(AA, belief(
has_state(human, sensor_state(externalinfo, V))))

ALP3 Generating a belief for a sensory representation
If AA believes that the human senses external information,
then it will generate a belief that the human will have a sensory representation for this.

has_state(AA, belief(
has_state(human, sensor_state(externalinfo, V))))
→ has_state(AA, belief(has_state(human, srs(externalinfo, V))))

ALP4 From sensory representation and feeling to belief
If AA believes that the human has a sensory representation for external information with level V_1
and AA believes that the human has feeling b with level V_2,
and the belief for w has level V_3
and β_1 is the person's estimated orientation for emotional response
and γ_1 is the person's flexibility for bodily responses
then it will generate the belief that the human's belief with level $V_3+\gamma_1 (g(\beta_1, V_1, V_2)-V_3)\ \Delta t$ will occur

has_state(AA, belief(
has_state(human, srs(externalinfo, V1)))) &
has_state(AA, belief(has_state(human, feeling(b, V2)))) &
has_state(AA, belief(has_state(human, belief(w, V3))))
→ has_state(AA, belief(has_state(human,
belief(w, $V_3+\gamma_1$ (g(β_1, V_1, V_2)-V_3) Δt))))

ALP5 From belief and feeling to preparation of a body state
If AA believes that the human has a belief for w with level V_1
and AA believes that the human has feeling b with level V_2,
and the preparation for body state b has level V_3
and β_2 is the person's orientation for emotional response
and γ_2 is the person's flexibility for bodily responses
then it will generate the belief that the human's preparation state for body state b will occur with level $V_3+\gamma_2\ (g(\beta_2, V_1, V_2)-V_3)\ \Delta t$.

has_state(AA, belief(has_state(human, belief(w, V1)))) &
has_state(AA, belief(has_state(human, feeling(b, V2)))) &
has_state(AA, belief(has_state(human, preparation(b, V3)))) &
→ has_state(AA, belief(has_state(human,
preparation(b, $V_3+\gamma_2$ (g(β_2, V_1, V_2)-V_3) Δt))))

ALP6 From preparation to body modification
If AA believes that the human's preparation state for body state b with level V occurred,
then it will believe that the human's body state will have level V.

has_state(AA, belief(has_state(human, preparation(b, V))))
→ has_state(AA, belief(has_state(human, effector_state(b, V))))

ALP7 From body modification to modified body
If AA believes that the human's body is modified with level V,
then it will believe that the human's body is showing b with level V.

has_state(AA, belief(has_state(human, effector_state(b, V))))
→ has_state(AA, belief(has_state(human, body_state(b, V))))

ALP8 Sensing a body state
If AA believes that the human's body is showing b with level V,
then it will believe that the human will sense this body state.

has_state(AA, belief(has_state(human, body_state(b, V)))) →
has_state(AA, belief(has_state(human, sensor_state(b, V))))

ALP9 Generating a sensory representation of a body state
If AA believes that the human has sensed body state b with level V,
then it will believe that the human has a sensory representation for body state b with level V.

has_state(AA, belief(has_state(human, sensor_state(b, V))))
→ has_state(AA, belief(has_state(human, srs(b, V))))

ALP10 From sensory representation of body state to feeling
If AA believes that the human has a sensory representation for body state b with level V,
then it will believe that the human has feeling b with level V.

has_state(AA, belief(has_state(human, srs(b, V))))
→ has_state(AA, belief(has_state(human, feeling(b, V))))

In addition, a number of other rules have been established to model the behaviour of the human and the ambient agent, and its effect on the world:

ALP11 Intervention by the Ambient Agent
If AA believes that the human has feeling b with level V which is higher than a certain threshold th1,
then it will add some positive information to the external environment

has_state(AA, belief(has_state(human, feeling(b, V)))) & V ≥ th1
→ has_state(AA, performed(add_pos_info))

ALP12 Effect of intervention in the world
As long as AA does not add some positive information to the external environment,
then positive information will remain 0.

not has_state(AA, performed(add_pos_info))
→ added_pos_info(0)

As soon as AA adds some positive information to the external environment, it will be available in the environment.

has_state(AA, performed(add_pos_info))
→ added_pos_info(0.2)

The Adaptation Model

Characteristics of a human, used as parameters in a dynamical model (such as the β used in the belief generation in the model described above) are often not easy to determine at forehand, and can only be given to the agent as initial beliefs. This section describes a method by which an agent is able to adapt these beliefs concerning human characteristics to the real characteristics. Using the dynamical model with parameter values as represented by these initial beliefs, the agent predicts the human belief and feeling state, up to a certain time point. When at that time point, for example by observation, information is obtained about the real value of this belief or feeling state, this is used

as input for the adaptation process. The agent adjusts the belief on the human characteristic, to reduce the difference between predicted and real value.

For reasonable adjustments, information is required on how a change in parameter value affects the difference between predicted and real value of the variable that is considered; this is called the sensitivity of the variable value for the parameter value. The *sensitivity S* of variable X (e.g., the belief or feeling level) for parameter P (e.g., the β used in belief generation) is the number such that a change ΔP in the value of parameter P will lead to a change ΔX in X which is (approximately) proportional to ΔP with S as proportion factor: $\Delta X = S\ \Delta P$. This is an approximation which is more accurate when the Δ's are taken small. To determine a sensitivity S the following approximation method is used. A small change ΔP in the parameter is used to make an additional prediction for X, and based on the resulting difference ΔX found in the two predicted values for X, by $S_{X,P} = \Delta X / \Delta P$ the sensitivity S can be estimated. Once the sensitivity and a deviation ΔX between estimated and observed level have been determined, the value W of the parameter P is adjusted by ΔP in the following manner (with α the adaptation speed):

$$\Delta P = \alpha*(1 - W)* (-\Delta X / S_{X,P}) \quad \text{if } -\Delta X / S_{X,P} \geq 0$$

$$\Delta P = \alpha* W* (-\Delta X / S_{X,P}) \quad \text{if } -\Delta X / S_{X,P} \leq 0$$

This has been specified in LEADSTO-format as follows.

ALP13 Calculating change ΔX in predicted belief X

If AA believes that the human has a sensory representation for external information with level V_1
and AA believes that the human has feeling b with level V_2,
and AA believes that the predicted belief for w has level V_3
and β_1 is the person's estimated orientation for emotional response
and γ_1 is the person's flexibility for bodily responses
and the change to be made in person's estimated β_1 is V_4
then it will generate the predicted belief for w with level $V_3+\gamma_1\ (g(\beta_1+ V_4, V_1, V_2)-V_3)\ \Delta t$

has_state(AA, belief(as_state(human,srs(externalinfo, V1)))) &
has_state(AA, belief(has_state(human, feeling(b, V2)))) &
has_state(AA, belief(predicted_belief(w, V3)))
→ has_state(AA, belief(
predicted_belief(w, $V_3+\gamma_1$ (g(β_1+V_4, V_1, V_2)-V_3) Δt))))

ALP14 Generating sensitivity

If AA believes that the predicted belief for w has level V_1
and AA believes that the human has a belief for w with level V_2
and the change to be made in person's estimated β_1 is V_3
then AA will generate the belief for sensitivity by $(V1 - V2)/V3$

has_state(AA, belief(predicted_belief(w, V1))) &
has_state(AA, belief(has_state(human, belief(w, V2))))
→ has_state(AA, belief(sensitivity, ($V_1 - V_2$) / V_3))

ALP15 Calculating deviation

If AA believes that the human has a belief for w with level V_1
and AA believes that the observed human belief is V_2
then AA will generate the belief that the deviation between estimated and observed belief is $(V1 - V2)$

has_state(AA, belief(has_state(human, belief(w, V1)))) &
has_state(AA, belief(observed_human_belief, V2))
→ has_state(AA, belief(deviation, $V_1 - V_2$))

ALP16 Adapt estimated beta

If AA believes the estimated beta is V_1
and AA believes that the deviation between estimated and observed belief is V_2
and AA believes that the sensitivity is V_3
and $-V_2 / V_3 > 0$
and α is the adaptation speed
then AA will generate the belief in an estimated beta with level $(\alpha * (1 - V1)* (-V1 / V3) + V1)$

has_state(AA, belief(estimated_beta(V1))) &
has_state(AA, belief(deviation, V2)) &
has_state(AA, belief(sensitivity, V3)) &
- V2 / V3 > 0
→ has_state(AA,
belief(estimated_beta(α * (1 - V1)* (-V1 / V3) + V1)))

If AA believes estimated beta is V_1
and AA believes that the deviation between estimated and observed belief is V_2
and AA believes that the sensitivity is V_3
and $-V_2 / V_3 \leq 0$
and α is the adaptation speed
then AA will generate the belief in an estimated beta with level $(\alpha * V1 * (-V1 / V3) + V1)$

has_state(AA, belief(estimated_beta(V1))) &
has_state(AA, belief(deviation, V2)) &
has_state(AA, belief(sensitivity, V3)) &
- V2 / V3 <= 0
→ has_state(AA, belief(estimated_beta(α * V1 * (-V1 / V3) + V1)))

Simulation Results

Based on the model described in the previous section, a number of simulations have been performed within the LEADSTO simulation environment (Bosse, Jonker, Meij and Treur, 2007). The model was tested in a small scenario, involving an ambient agent and a human (indicated by AA and human, respectively). The agent model was equipped with the model to estimate human's emotion level. The central emotion used in the scenario is insecurity for the particular product, as discussed in Section 1. In order to simulate this, every now and then certain events take place, which influence the level of insecurity of the human either positively (e.g., some good news about the product published in a newspaper) or negatively (e.g., some friend informed him about his own past bad experience with that product). To model this behavior, the following property has been used:

ALP17 Generating a sensor state for external information

If a sensor state of external information of level V_1 occurs
and the ambient agent has added some positive information V_2, has flexibility η
and some positive information V_3
and some negative information V_4 is present from the environment,
then the human will sense external information with level $(V_1-\eta*(V_2+V_3)*V_1+\eta*V_4*(1-V_1))$.

has_state(human, sensor_state(externalinfo, V1)) &
added_pos_info(V2) &
flexibility(η) &
positive_externalinfo(V3) &
negative_externalinfo(V4)
→ has_state(human, sensor_state(
externalinfo, ($V_1-\eta*(V_2+V_3)* V_1+\eta*V_4*(1-V_1)$))))

Here positive_externalinfo and negative_externalinfo represent the positive and negative events that are occurring randomly in the environment which influence the insecurity level of the human. For the example simulations the probability for the positive events to occur has been taken *0.8* and for negative events to occur is *0.3*. The main goal of the ambient agent is to estimate the level of insecurity of the human. To this end, it starts with some initial values of the human's belief and feeling levels, and then keeps on updating this, using the

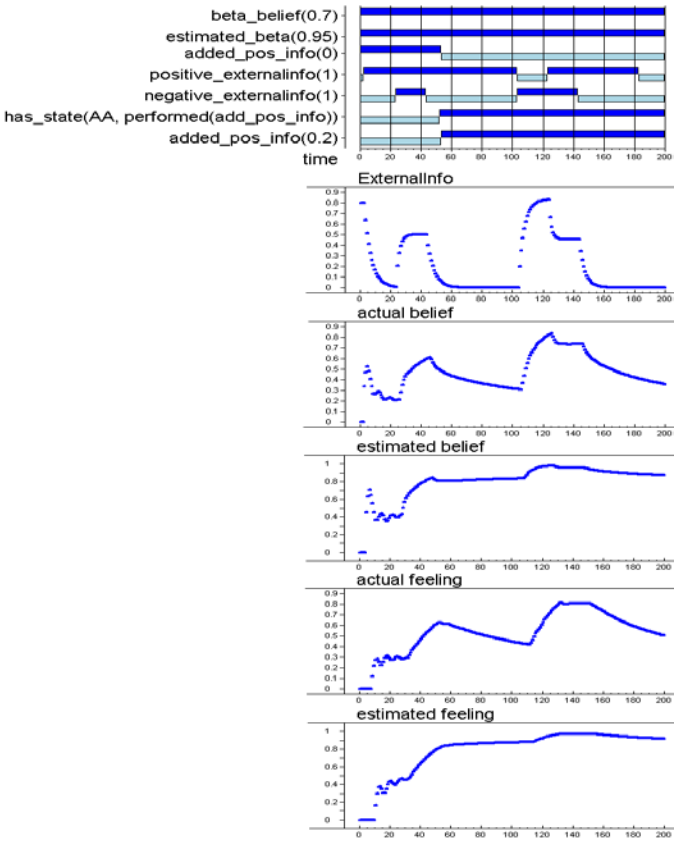


Figure 2: Simulation 1: the estimated β is higher than the real β

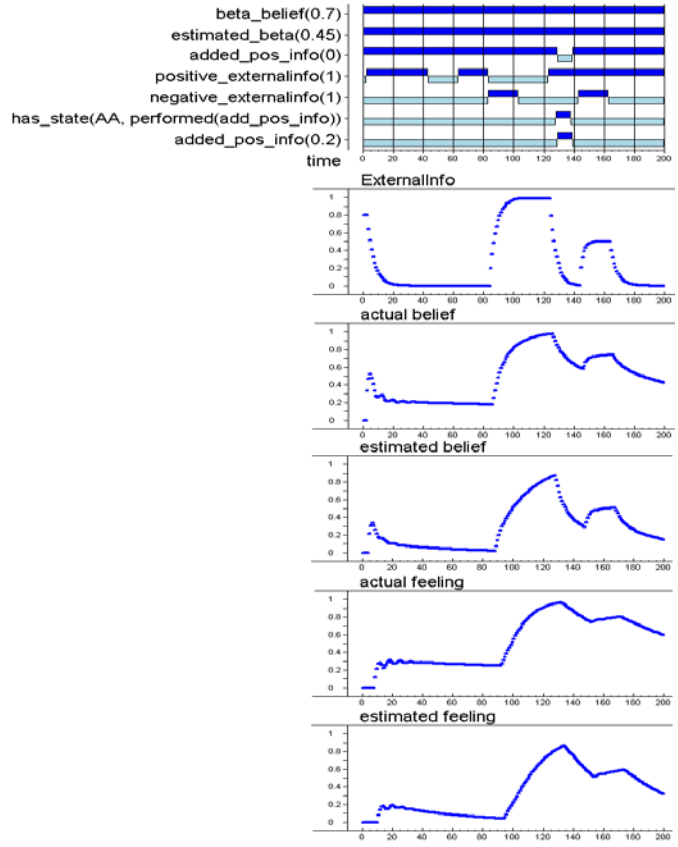


Figure 3: Simulation 2: the estimated β is lower than the real β

strategies explained earlier. him some positive information about the product). When it is estimated that the human becomes too (unreasonably) insecure, the ambient agent can take measures to calm him down (e.g., informing

Some example simulation traces (under different but fixed parameter settings) are illustrated in Figures 2 and 3 (here the time delays within the temporal LEADSTO relations were taken 1 time unit). In all of these figures, where time is on the horizontal axis, the upper part shows the time periods, in which the binary logical state properties hold (indicated by the dark lines); for example, added_pos_info. Below this part, quantitative information is provided about the human's actual belief and feeling level, and the ambient agent AA's estimation of this belief and feeling level, respectively. Values for these levels for the different time periods are shown by the dark lines. Note that only a selection of the relevant state properties is shown.

The first trace (see Figure 2), shows a situation in which the estimated β (*0.95*) is substantially higher than the real β (*0.7*), as indicated in the upper part of Figure 2. As shown in the figure, the ambient agent AA estimates the level of emotion of the human too high so that it is too early in adding the positive information indicated in the upper part by state property: has_state(AA, performed(add_pos_info)), at time point 52.

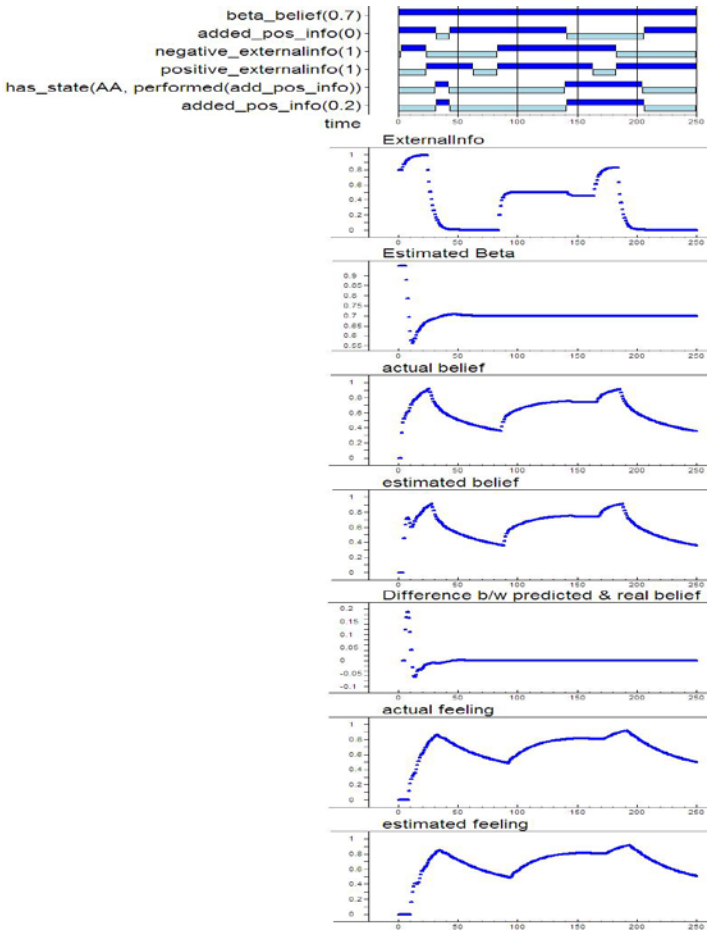


Figure 4: Simulation 3: the estimated β is adapted and approximates the real β

The second trace (see Figure 3) shows a situation in which the estimated β (*0.45*) is substantially lower than the real β (*0.7*), as indicated in the upper part of the Figure 3. As shown in the figure, the ambient agent AA estimates the level of emotion of the human much too low, so that it is too late in adding the positive information, indicated in the upper part by state property: has_state(AA, performed(add_pos_info)), at time point 128. This is too late, because, as shown in the actual emotion graph below, the human's emotion level has gone too high already at time point 118.

In Figure 4 a simulation trace is shown where the parameter β is adapted to the person. Here the initial value of β is too high (*0.95*) compared to the actual value (*0.7*). To compensate for that, the adaptation model first reduces the estimated value to below *0.6*, after which it almost monotonically approximates the real value *0.7*.

Discussion

To function in a knowledgeable manner, ambient agents (e.g., Aarts, Collier, Loenen, Ruyter, 2003; Aarts, Harwig, Schuurmans, 2001; Riva, Vatalaro, Davide, Alcañiz, 2005) need a model of the humans they are supporting. Such a model enables them to perform a form of mindreading (e.g., Gärdenfors, 2003; Goldman, 2006). The ambient agent model presented here focuses on mindreading concerning the interaction between beliefs and emotions, based on neurological theories that address this interaction. A belief usually triggers an emotional response and may also depend on this emotional response, as, for example, shown in literature such as (Eich et al., 2000; Forgas et al., 2009; Niedenthal, 2007; Schooler and Eich, 2000).

The ambient agent model presented uses a computational model of this interaction, adopted from (Memon and Treur, 2009). For feeling the emotion, based on elements taken from (Damasio, 1999, 2004; Bosse, Jonker and Treur, 2008), a converging recursive body loop is included in the model. As a second loop the model includes a feedback loop for the interaction between feeling and belief. The causal relation from feeling to belief in this second loop was inspired by the Somatic Marker Hypothesis described in (Damasio, 1994, 1996; Bechara and Damasio, 2004), and may also be justified by a Hebbian learning principle (cf. Hebb, 1949; Bi and Poo, 2001). Both the strength of the belief and of the feeling emerge as a result of the dynamic pattern generated by the two loops.

The adaptive integrative agent model equipped with the dynamical model for the dynamics of belief and feeling was specified in the hybrid dynamic modelling language LEADSTO, and simulations were performed in its software environment; cf. (Bosse, Jonker, Meij, and Treur, 2007). An adaptation model was integrated within the agent to be able to tune beliefs on the human's characteristics used as parameters in the dynamical model to the real characteristics. Here feedback can be used when at times the human reveals his or her belief or feeling. To evaluate the ambient agent model in human experiments is left to future work.

References

Aarts, E.; Collier, R.; Loenen, E. van; Ruyter, B. de (eds.) (2003). Ambient Intelligence. Proc. EUSAI 2003. Lecture Notes in Computer Science, vol. 2875. Springer Verlag, 2003.

Aarts, E., Harwig, R., and Schuurmans, M. (2001). Ambient Intelligence. In: P. Denning (ed.), The Invisible Future. McGraw Hill, pp. 235-250.

Bechara, A., and Damasio, A. (2004). The Somatic Marker Hypothesis: a neural theory of economic decision. Games and Economic Behavior, vol. 52, pp. 336-372.

Bi, G.Q., and Poo, M.M. (2001) Synaptic Modifications by Correlated Activity: Hebb's Postulate Revisited. Ann Rev Neurosci, vol. 24, pp. 139-166.

Bosse, T., Jonker, C.M., Meij, L. van der, and Treur, J. (2007). A Language and Environment for Analysis of Dynamics by Simulation. International Journal of Artificial Intelligence Tools, vol. 16, 2007, pp. 435-464.

Bosse, T., Jonker, C.M., and Treur, J. (2008). Formalisation of Damasio's Theory of Emotion, Feeling and Core Consciousness. Consciousness and Cognition Journal, vol. 17, 2008, pp. 94–113.

Damasio, A. (1994). Descartes' Error: Emotion, Reason and the Human Brain, Papermac, London.

Damasio, A. (1996). The Somatic Marker Hypothesis and the Possible Functions of the Prefrontal Cortex. Philosophical Transactions of the Royal Society: Biological Sciences, vol. 351, pp. 1413-1420

Damasio, A. (1999). The Feeling of What Happens. Body and Emotion in the Making of Consciousness. New York: Harcourt Brace, 1999.

Damasio, A. (2004). Looking for Spinoza. Vintage books, London.

Eich, E., Kihlstrom, J.F., Bower, G.H., Forgas, J.P., & Niedenthal, P.M. (2000). Cognition and Emotion. New York: Oxford University Press.

Forgas, J.P., Goldenberg, L., and Unkelbach, C. (2009). Can bad weather improve your memory? An unobtrusive field study of natural mood effects on real-life memory. Journal of Experimental Social Psychology, vol. 45, 2009, pp. 254–257.

Gärdenfors, P. (2003). How Homo Became Sapiens: On The Evolution Of Thinking. Oxford University Press, 2003.

Goldman, A.I. (2006). Simulating Minds: the Philosophy, Psychology and Neuroscience of Mindreading. Oxford University Press.

Hebb, D. (1949). The Organisation of Behavior. Wiley, New York.

Memon, Z.A., and Treur, J., (2009). Modelling the Reciprocal Interaction between Believing and Feeling from a Neurological Perspective. In: N. Zhong et al. (eds.), *Proc. of the First Intern. Conf. on Brain Informatics, BI'09*. Lecture Notes in Artificial Intelligence, vol. 5819. Springer Verlag, 2009, pp. 13-24.

Marsella, S.C., Pynadath, D.V., and Read, S.J. (2004). PsychSim: Agent-based modeling of social interaction and influence. In: Lovett, M. et al. (eds.), *Proceedings of the International Conference on Cognitive Modelling, ICCM'04*, pp. 243-248.

Niedenthal, P.M. (2007). Embodying Emotion. Science, vol. 316, (2007), pp. 1002-1005.

Riva, G., F. Vatalaro, F. Davide, M. Alcañiz (eds.) (2005). Ambient Intelligence. IOS Press.

Schooler, J.W., and Eich, E. (2000). Memory for Emotional Events. In: E. Tulving, F.I.M. Craik (eds.), The Oxford Handbook of Memory. Oxford University Press, 2000, pp. 379-394.

Winkielman, P., Niedenthal, P.M., and Oberman, L.M. (2009). Embodied Perspective on Emotion-Cognition Interactions. In: Pineda, J.A. (ed.), Mirror Neuron Systems. Springer Science, 2009, pp. 235-257.

Proposing Artificial Subtle Expressions as an Intuitive Notification Methodology for Artificial Agents' Internal States

Takanori Komatsu (tkomat@shinshu-u.ac.jp)
International Young Researcher Empowerment Center, Shinshu University,
3-15-1 Tokida, Ueda 386-8567, Japan

Seiji Yamada (seiji@nii.ac.jp)
National Institute of Informatics/ SOKEDAI,
2-1-2 Hitotsubashi, Tokyo 101-8430, Japan

Kazuki Kobayashi (kby@cs.shinshu-u.ac.jp)
Graduate School of Science and Technology, Shinshu University
4-17-1 Wakasato, Nagano 380-8553, Japan

Kotaro Funakoshi (funakoshi@jp.honda-ri.com) and Mikio Nakano (nakano@jp.honda-ri.com)
Honda Research Institute Japan Co., Ltd,
8-1 Honcho, Wako 351-0188, Japan

Abstract

We describe artificial subtle expressions (ASEs) as an intuitive notification methodology for artifacts' internal states for users. We prepared two types of audio ASEs: one was a flat artificial sound (flat ASE), and the other was a sound that decreased in pitch (decreasing ASE). These two ASEs were played after a robot made a suggestion to the users. Specifically, we expected that the decreasing ASE would inform users of the robot's lower level of confidence in its suggestion. We then conducted a simple experiment to observe whether the participants accepted or rejected the robot's suggestion based on the ASEs. The results showed that they accepted the robot's suggestion when the flat ASE was used, whereas they rejected it when the decreasing ASE was used. We thereby concluded that the ASEs succeeded in conveying the robot's internal state to users accurately and intuitively.

Keywords: Artificial subtle expressions (ASEs); Complementary; Intuitive; Simple; Accurate.

Introduction

Although human communications are explicitly achieved through verbal utterances, paralinguistic information (e.g., pitch and power of utterances) and nonverbal information (e.g., facial expressions, gaze direction, and gestures) also play important roles (Kendon, 1994). This is because one's internal state is deeply reflected in one's paralinguistic and nonverbal information. In other words, other people can intuitively and easily understand a person's internal state from such information when it is expressed (Cohen et al., 1990). Recently, some researchers have reported that very small changes in the expression of such information, called subtle expressions (Liu & Picard, 2003), significantly influence human communications, especially in the conveyance of one's internal state to others. For example, Ward (2003) reported that the subtle flections of the pitch information in speech sounds reflect one's emotional states even when contradicted by the literal meanings of the speech sounds, and Cowell & Ayesh (2004) offered a similar argument in terms of facial expressions.

It is therefore believed that such subtle expressions can be utilized to help humans easily understand an artifact's internal state because humans can intuitively understand such subtle expressions. For example, Sugiyama et al. (2006) developed a humanoid robot that can express appropriate gestures based on a recognition of its situation, and Kipp & Gebhard (2008) developed a human-like avatar agent that can control its gaze direction according to the user's gaze direction. However, since these researchers tried to implement subtle expressions on artifacts (e.g., humanoid robots or dexterous avatar agents), it resulted in considerably high implementation costs.

In contrast to the above approaches, Yamada & Komatsu (2006) and Komatsu & Yamada (2007) reported that simple beeping sounds from a robot with decreasing/increasing frequency enabled humans to interpret the robot's negative/positive states. Funakoshi et al. (2008) also reported that the robot's blinking LED could convey to users a robot's internal state (processing or busy) for the sake of reducing the occurrence of speech collisions during their verbal conversations. It then seemed that such simple expressions (beeping sounds or blinking LEDs) from artifacts could play a similar role to the subtle expressions of humans, so we named these expressions in artifacts "Artificial Subtle Expressions (ASEs)," referring to artifacts' simple and low-cost expressions that enable humans to estimate the artifacts' internal state accurately and intuitively. We stipulate that the ASEs should

simultaneously meet two design and two functional requirements.

Specifically, the two design requirements are as follows:

- **Simple**: ASEs should be implemented on a single modality. This is expected to lower the implementation cost.
- **Complementary**: ASEs should only have a complementary role in communication and should not interfere with communication's main protocol. This means that the ASEs themselves do not have any meaning without a communication context.

The two functional requirements are as follows:

- **Intuitive**: ASEs should be understandable by humans who have no prior knowledge of the ASEs.
- **Accurate**: ASEs should convey the designer's intended meanings accurately. Specifically, ASEs should convey the internal states of the artifact just as subtle expressions do in nonverbal information by humans.

In this study, we focused on audio ASEs. Related studies with audio ASEs include those that proposed simple and effective information to convey specific meaning to users, e.g., "earcon (Blattner, 1989)" or "auditory icon (Gaver, 1989; Gaver, 1997)" These earcons and auditory icons play an effective role in informing users of specific meanings as communication's main protocol, while ASEs play a complementary role for the main protocol. This is the significant difference between ASEs and earcons or auditory icons.

In this paper, we investigated whether the ASEs could convey the artifacts' internal state to the users accurately and intuitively; specifically, we created audio ASEs that were intended to meet the two design requirements and investigated whether they also met the two functional requirements by conducting a simple psychological experiment.

Experiment

Setting

We used a "treasure hunting" video game as an experimental environment to observe participants' behavior (Figure 1). In this game, a game image scrolls forward on a straight road, with small hills appearing along the way. A coin is inside one of three hills, while the other two hills have nothing. The game ends after the player encounters 20 sets of hills, and the approximate duration of this video game is about three minutes. The purpose is to get as many coins as possible. In this experiment, the participant was awarded 1 point for each coin that s/he found. The participants in this experiment were informed that 1 point was equivalent to 50 Japanese yen (about 50 US cents) and that after the experiment they could use their points to purchase some stationery supplies (e.g., file holders or USB flash memory) of equivalent value.



Figure 1: Treasure hunting video game.

The position of the coin in the three hills was randomly assigned. In each trial, an artifact placed next to the participants told them in which position it expected the coin to be placed. The artifact placed next to the participants was the MindStorms robot (LEGO Corporation, see Figure 2). The robot told the participant the expected position of the coin using its speech sounds. The participants could freely accept or reject the robots' suggestions. In each trial, even though the participants selected one hill from among three, they did not know whether the selected hill had the coin or not (actually, the selected hill just showed a question mark and a closed treasure box, as depicted in the center of Figure 1). The participants were informed of their total game points only after the experiment.

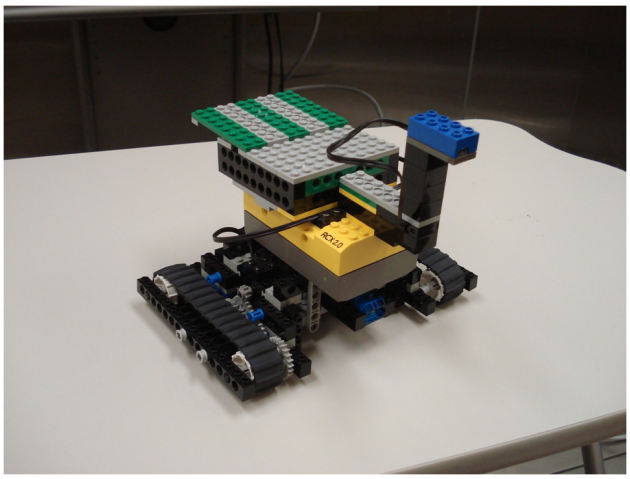


Figure 2: MindStorms Robot.

Utilized ASEs

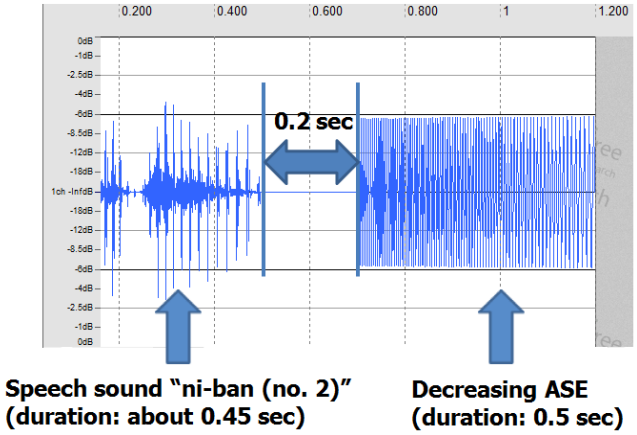


Figure 3: Speech sound "ni-ban (no.2)" and decreasing ASE.

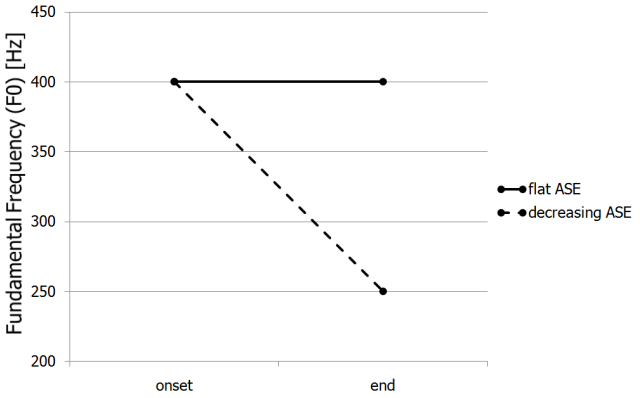


Figure 4: Flat and decreasing ASEs (duration: 0.5 second).

We implemented the audio ASEs in the robot's speech sounds. In this experiment, the robot expressed Japanese artificial speech sounds to tell the expected position of the coin; that is, "ichi-ban (no. 1)," "ni-ban (no. 2)," and "san-ban (no. 3)." These artificial speech sounds were created by the text-to-speech (TTS) function of "Document Talker (Create System Development Company)." Just 0.2 seconds after these speech sounds, one of the two simple artificial sounds was played as the ASE (Figure 3). These two ASEs were triangle wave sounds 0.5 seconds in duration, but their pitch contours were different (Figure 4); that is, one was a flat sound (onset F0: 400 Hz and end F0: 400 Hz, called "flat ASE"), and the other was a decreasing one (onset F0: 400 Hz and end F0: 250 Hz, called "decreasing ASE"). These ASE sounds were created by "Cool Edit 2000 (Adobe Corporation)." Komatsu & Yamada (2007) reported that the decreasing artificial sounds expressed from the robot were interpreted as negative feelings by humans; therefore, we intended that the decreasing ASE would inform users of the robot's lower confidence in the suggestions as the robot's internal state.

Here, the main protocol of the robot was to tell the expected position of the coin, while the ASE protocol was to indicate the robot's confidence level in a complementary manner. The two ASE sounds were created quite easily by simply editing the consumer software. Thus, the ASEs met the two design requirements, that is, simple and complementary. Therefore, to confirm whether the ASEs were able to convey the robot's internal states to the users accurately and intuitively, we needed to investigate whether the utilized ASE met the two requirements for functioning, that is, being intuitive and accurate.

Procedure

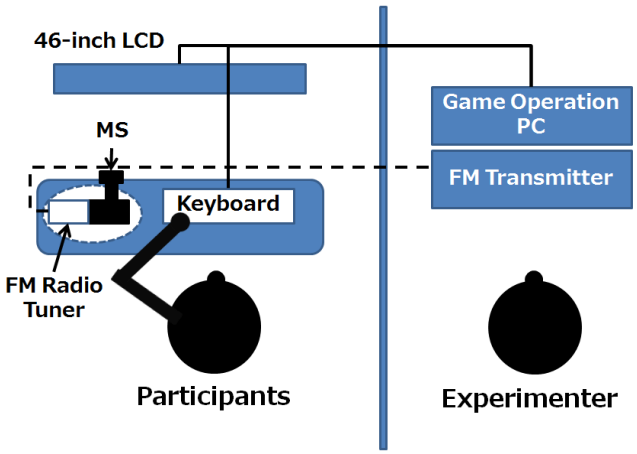


Figure 5: Experimental setting.

Nineteen Japanese university students (10 men and 9 women; 22 – 25 years old) participated. The treasure hunting video game was projected on a 46-inch LCD in front of the participants, and the robot was placed in front of and to the right of the participants, with the distance between them being approximately 50 cm (see Figures 5 and 6). The sound pressure of the robot's speech sounds at the participants' head level was set at about 50 dB (FAST, A). The robot's speech sounds with the ASEs were remotely controlled by the experimenter in the next room using the Wizard of Oz (WOZ) method. Before the experiment started,

the experimenter told the participant the setting and purpose of the game. However, the experimenter never mentioned or explained the ASEs. Therefore, the participants had no opportunity to acquire prior knowledge about the ASEs. Among the 20 trials, the robots expressed the flat ASE 10 times and the decreasing ASE 10 times. The order of expression for these two types of ASEs was counterbalanced across participants. Actually, the robot told the exact position of the coin in all 20 trials, but the participants did not know whether or not the robot was telling the right position because the participants were not able to find out whether the selected hill had the coin or not. If the participant actually knew whether or not the selected hill had the coin just after their selections, they would have associated the ASE with the robot's performance, e.g., whether or not the robot pointed to the right position. Thus, this experimental setting, where the participants were not notified of whether the selected hill was correct or not, was intended to reduce such associations and to clarify the effect of the ASEs on the participants' behavior.

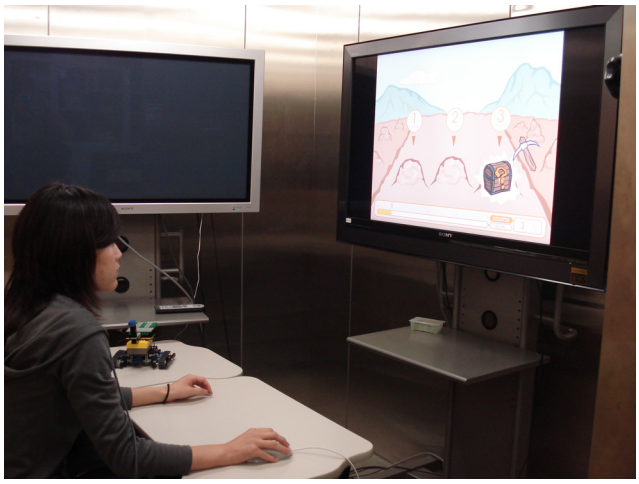


Figure 6: Experimental Scene

The purpose of this experiment was to observe the participants' behavior as to whether they accepted or rejected the robot's suggestions in terms of the types of ASEs used. We assumed that *the participants would accept the robot's suggestion when the flat ASE was added to the speech sounds while they would reject the suggestion when the decreasing ASE was used.* If we could observe these phenomena, we could recognize that the utilized ASE had succeeded in conveying the robot's internal state to the participants accurately and intuitively; that is, the ASE had successfully met all four requirements. In addition, after the experiment, we conducted interviews to determine whether or not the participants had noticed the ASEs and, if so, how they had interpreted them.

Results

To investigate the effect of the ASEs on participants' behavior, we calculated the rejection rate, indicating how many of the robot's suggestions the participants rejected for 10 flat ASEs and 10 decreasing ASEs. For all 19 participants, the average rejection rate of the 10 flat ASEs was 1.73 (SD=1.51), while the rejection rate of the 10 decreasing ASEs was 4.58 (SD=2.43, see Figure 7). These rejection rates for the 10 flat ASEs and 10 decreasing ASEs were analyzed using a one-way analysis of variance (ANOVA) (within-subjects design; independent variable: type of ASE, flat or decreasing, dependent variable: rejection rate). The result of the ANOVA showed a significant difference between the two groups ($F(1,18)=13.38$, $p<.01$, (**)); that is, the robot's suggestions with the decreasing ASE showed a significantly higher rejection rate compared to those with the flat ASE. Therefore, the ASEs significantly affected the participants' behavior, and we found evidence supporting our previously mentioned assumption .The most interesting point was that the ASEs affected the behavior of the participants without their being informed of the meaning or even existence of the ASEs.

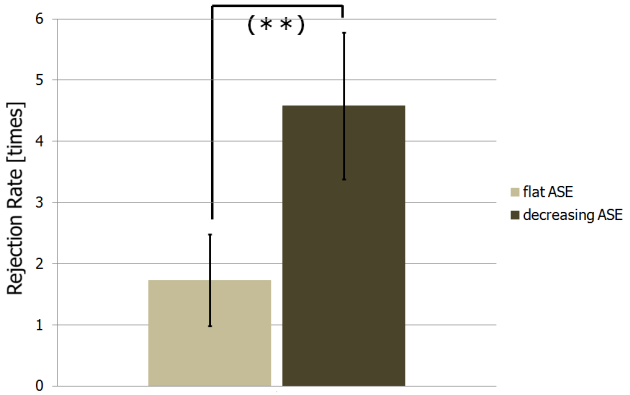


Figure 7: Rejection rate for all 19 participants.

In the interview sessions, 5 out of the 19 participants said that they immediately realized the meanings of the ASEs after the robot's speech sounds and that they utilized these ASEs when it came to accepting or rejecting the robot's suggestions, e.g., "I felt that the decreasing artificial sounds meant that the robot had less confidence in its answer." However, the remaining 14 participants said that they did not notice the existence of the ASEs. Here, if there were significant differences between flat and decreasing ASEs in their rejection rate, the ASEs were interpreted by these 14 participants unconsciously. In this case, we strongly argue that the ASEs were able to convey the robot's internal state to the participants accurately and intuitively. For these 14 participants, the average rejection rate of 10 flat ASEs was 2.28 (SD=1.73), while the rejection rate of the 10 decreasing ASEs was 3.43 (SD=1.59, see Figure 8). These rejection rates were analyzed using a one-way ANOVA (within-subjects design; independent variable: ASE type, flat or decreasing, dependent variable: rejection rate). The result of the ANOVA showed a significant difference between them ($F(1,13)=4.98$, $p<.05$, (*)); that is, the robot's suggestions

with the decreasing ASE had a significantly higher rejection rate compared to those with a flat ASE, even though these participants did not notice the existence of the ASEs. To sum up, the results of this experiment clearly show that the utilized ASEs succeeded in conveying the robot's internal states to the participants accurately and intuitively; that is, the ASEs succeeded in meeting all four requirements.

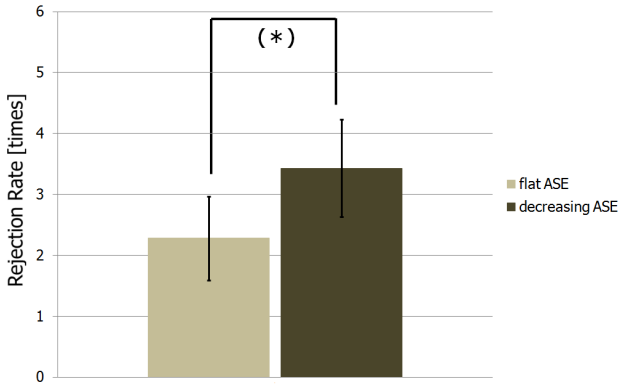


Figure 8: Rejection rate for 14 participants who did not notice ASEs.

Discussion

Future Applications

As a result of the experiment, we could confirm that the robot's suggestions with the decreasing ASEs showed a significantly higher rejection rate compared with those with flat ASEs. Moreover, these ASEs were interpreted by the participants even though they were not informed of the meaning or even the existence of the ASEs. Therefore, our experiment clearly showed that the utilized ASEs succeeded in conveying the robot's internal states to the participants accurately and intuitively.

Currently, we are planning to implement the ASEs in various kinds of spoken dialogue systems such as ATMs and automatic telephone reservation systems. Specifically, we are now focusing on car navigation systems' speech sounds; the reason for this is that current car navigation systems still sometimes give poor driving routes to users. However, if this navigation system's confidence level regarding the route instruction is not very high, the instructions of speech sounds with ASEs could implicitly convey a lower confidence level. If the ASEs are still effective in such situations, they could be utilized in various situations in which artifacts have to convey their internal states to users.

In our experiment, we only focused on the internal state of the artifact in order to convey to users its level of confidence in its own expressed information. However, we are planning to investigate which kinds of internal states could be conveyed to the users by means of ASEs. For example, it is expected that the artifacts should also convey other kinds of internal states, such as feelings or conditions, and the confidence level in interpreting the user's expressions. This consecutive study would also contribute to expanding the applicability of ASEs to various interactive situations.

Advantage of utilizing ASEs

It is said that the most significant advantage in utilizing ASEs is the lower implementation cost compared to utilizing human-like expressions. Therefore, it is expected that many applications in human-computer interaction or human-robot interaction will be able to include the ASEs quite easily. In addition to the lower cost, we believe that the advantage of utilizing ASEs includes the possibility of solving several problems such as those reported in the above research areas.

So far, it has been strongly believed that most robots or on-screen agents required to interact with users should have a human-like appearance and produce human-like expressions. However, we feel that these research directions have had two difficulties; one is the implementation cost mentioned above, and the other is that users have unexpected attitudes or impressions toward human-like artifacts; i.e., artifacts having a human-like appearance have a higher possibility of diving them into the "uncanny valley" (Mori, 1970). Moreover, users are likely to overestimate the artifacts' ability when it has a human-like appearance or expressions, so they would be disappointed if these artifacts were to demonstrate unpredictable or poor behavior (Komatsu & Yamada, 2010).

Therefore, our approach that the artifact should not produce human-like expressions but artifact-like ones to convey its internal state to the users has succeeded in proposing a novel research approach in the research area of human-computer interaction or human-robot interaction in order to resolve the above issues. Now we are planning to conduct a consecutive study to compare ASEs to human-like expressions in terms of users' cognitive load or cost-benefit relationships. Comprehending the advantages and disadvantages of these two expressions (ASEs and human-like expressions) would constitute a design methodology for artifacts' expressions in order to achieve smooth interaction between users and artifacts.

Conclusions

In this paper, we investigated whether the ASEs could convey artifacts' internal states accurately and intuitively to users; specifically, we created audio ASEs intended to meet the two requirements for design, and we investigated whether these ASEs met the two requirements for function by conducting a simple psychological experiment. As a result of this experiment, the robot's suggestions accompanied by decreasing ASEs showed a significantly higher rejection rate compared with those accompanied by flat ASEs. Moreover, these ASEs were accurately interpreted by participants even though they were not informed of the meaning or even the existence of the ASEs.

Therefore, our experiment clearly showed that the utilized ASEs succeeded in conveying the robot's internal states to the participants accurately and intuitively; that is, the ASEs succeeded in meeting all four requirements. Thus, we confirmed that simple and low-cost expression ASEs could be utilized as an intuitive notification methodology for artifacts to convey their internal states to users through paralinguistic or nonverbal information.

Acknowledgments

This study was partially funded by the Special Coordination Funds for Promoting Science and Technology granted by the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan.

References

Blattner, M. M., Sumikawa, D. A. & Greenberg, R. M. (1989). Earcons and Icons: Their Structure and Common Design Principles. *SIGCHI Bulletin. 21*, *1*, 123-124.

Cohen, P. R., Morgen, J., & Pollack, M. E. (1990). *Intentions in Communication*, The MIT Press, MA, USA.

Cowell, J. & Ayesh, A. (2004). Extracting subtle expressions for emotional analysis, In *Proceedings of 2004 IEEE International Conference on Systems, Man, Cybernetics (IEEE SMC 2004)*, pp. (1) 677-681.

Funakoshi, K., Kobayashi, K., Nakano, M., Yamada, S., Kitamura, Y., & Tsujino H. (2008). Smoothing human-robot speech interactions by using a blinking-light as subtle expression. In *Proceedings of the 10th International Conference on Multimodal Interface (ICMI 2008)*, pp. 293-296.

Gaver, W. W. (1989). The SonicFinder: An Interface That Uses Auditory Icons. *Human-Computer Interaction 4*, *1*, 67-94.

Gaver, W. W. (1997). *Auditory Interfaces. Handbook of Human-Computer Interaction*, Elsevier Science.

Kendon, A. (1994). Do gestures communicate? A Review. *Research in Language and Social Interaction 27*, *3*, 175-200.

Kipp, M. & Gebhard, P. (2008). IGaze: Studying reactive gaze behavior in semi-immersive human-avatar interactions, In *Proceedings of the 8th International Conference on Intelligent Virtual Agent (IVA2008)*, pp. 191-199.

Komatsu, T. & Yamada, S. (2007). How do robotic agents' appearances affect people's interpretation of the agents' attitudes? In *Extended Abstracts of CHI2007*, pp. 2519-2524.

Komatsu, T. & Yamada, S. (2010). Effects of Adaptation Gap on Users' Differences in Impressions of Artificial Agents, In *Proceedings of the 14th. World Multiconference on Systemics, Cybernetics and Informatics (WMSCI 2010)*, to appear.

Liu, K. & Picard, W. R. (2003). Subtle expressivity in a robotic computer. In *Proceedings of CHI2003 Workshop on Subtle Expressivity for Characters and Robots*, pp. 1-5.

Mori, M. (1970). Bukimi no tani (The uncanny valley, K. F. MacDorman & T. Minato, Trans.). *Energy 7, 4*, 33–35. (Originally in Japanese).

Sugiyama, O., Kanda, T., Imai, M., Ishiguro, H., Hagita, N. & Anzai, Y. (2006). Humanlike conversation with gestures and verbal cues based on a three-layer attention-drawing model. *Connection Science 18*, *4*, 379-402.

Yamada, S. & Komatsu, T. (2006). Designing Simple and Effective Expression of Robot's Primitive Minds to a Human, In *Proceedings of the 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'06)*, pp. 2614-2619.

Ward, N. (2003). On the Expressive Competencies Needed for Responsive Systems, In *Proceedings of CHI2003 Workshop on Subtle Expressivity for Characters and Robots*, pp. 33-34.

Experiments for Assessing Floating Reinstatement in Argument-based Reasoning

Iyad Rahwan
[1]Masdar Institute of Science & Technology, [2]MIT, [3]University of Edinburgh

Jean-François Bonnefon
CNRS and Université de Toulouse

Mohammed Iqbal Madakkatel, Ruqiyabi Naz Awan
British University in Dubai

Sherief Abdallah
[1]British University in Dubai, [2]University of Edinburgh

Abstract

Various Artificial Intelligence semantics have been developed to predict when an argument can be accepted, depending on the abstract structure of its defeaters and defenders. These semantics can make conflicting predictions, as in the situation known as floating reinstatement. We argue that the debate about which semantics makes the correct prediction can be informed by the collection of experimental data about the way human reasoners handle these critical cases. The data we report show that floating reinstatement yields comparable effects to that of simple reinstatement, thus supporting preferred semantics over grounded semantics. Besides their theoretical value for validating and inspiring argumentation semantics, these results have applied value for developing artificial agents meant to argue with people.

Keywords: Argumentation; Semantics; Nonmonotonic reasoning; Behavioural Experiment;

Introduction

Argumentation has become a very fertile area of research in Artificial Intelligence (Rahwan & Simari, 2009), where a highly influential framework for studying argumentation-based reasoning was introduced by Dung (1995). An *argumentation framework* is simply a pair $AF = \langle \mathcal{A}, \rightharpoonup \rangle$ where $\mathcal{A}$ is a set of arguments and $\rightharpoonup \subseteq \mathcal{A} \times \mathcal{A}$ is a defeat relation between arguments. This approach abstracts away from the origin of individual arguments and their internal structures, and focuses instead on the defeat relationship between them.

Figure 1 shows an example textual argument and its corresponding graph structure. This structure is the canonical example for the notion of *reinstatement*. In particular, while argument A is defeated by argument B, the presence of C reinstates A since C undermines A's only defeater.[1]

Given an argument framework (or graph), a semantics assigns a *status* to each argument. Classically, we distinguish between arguments that are *accepted* and those that are not (Dung, 1995). In some cases, all semantics agree on the result. For example, in Figure 1, all classical argumentation semantics agree that we should accept C (for lack of any counter-argument), reject B (because there is a good reason to), and accept A (since every objection to it has been defeated). When there are cycles, different semantics may prescribe different results.

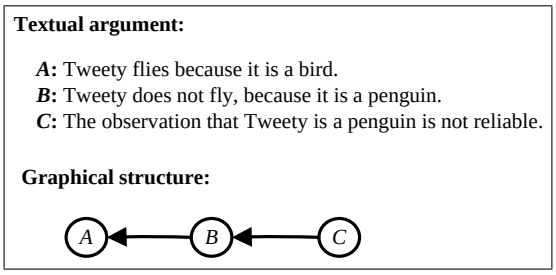


Figure 1: Defeat structure with reinstatement

These semantics typically come from a normative perspective, which relies on intuition and ad hoc hypothetical examples as to what constitutes correct reasoning. We will argue that there are limits to relying solely on this approach, and we will advocate the use of psychological experiments as a methodological tool for informing and validating intuitions about argumentation-based reasoning.

In this paper, we apply this experimental method to the problem of floating reinstatement. We will show that psychological experiments can help to evaluate these various semantics, and can provide unique insights even when all formal semantics are in agreement. Not only can these insights inform current and future semantics, but they are relevant to the design of software agents that can argue persuasively with humans, or provide reliable support to human evaluation of arguments (e.g., on top of argument diagramming tools).

Abstract Argumentation Frameworks

This section contains technical background only, whose outline is the following. Figure 1 displays the canonical graph of simple reinstatement, whereas Figure 2 displays the canonical graph of floating reinstatement. The main question is, in both cases, whether A can be accepted. For simple reinstatement, A is accepted by preferred as well as grounded semantics (to be defined below). For floating reinstatement, A is not accepted by grounded semantics, but is accepted by preferred

[1]While many notions of defeat exist, here we adopt the simple notion of *undercutting*: the defeater's conclusion explicitly negates the defeated argument's premise.

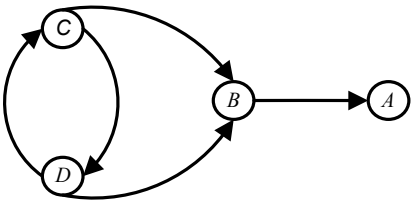


Figure 2: The canonical graph of defeat and floating reinstatement. Argument A is defeated by B, which is itself defeated by C as well as D, although C and D are mutual defeaters.

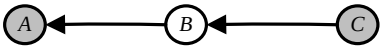


Figure 3: Single (complete, grounded, and preferred) extension in simple reinstatement. Accepted arguments are shaded.

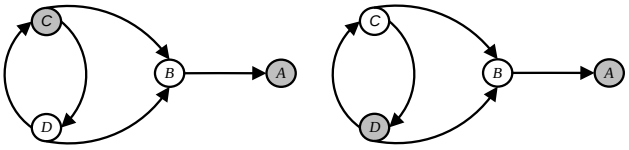


Figure 4: The two (complete, preferred) extensions in floating reinstatement. Accepted arguments are shaded.

semantics. Additionally, preferred semantics also accept C and D in the (formally defined) 'credulous' sense, but not in the 'sceptical' sense.

We now lay bare the technical background required to arrive at these conclusions. We begin with Dung's (1995) abstract definition of an argumentation framework.

Definition 1 (Argumentation framework)**.** *An* argumentation framework *is a pair $AF = \langle \mathcal{A}, \rightharpoonup \rangle$ where $\mathcal{A}$ is a set of arguments and $\rightharpoonup \subseteq \mathcal{A} \times \mathcal{A}$ is a defeat relation. An argument α* defeats *an argument β iff $(\alpha, \beta) \in \rightharpoonup$, also written $\alpha \rightharpoonup \beta$.*

The directed graphs displayed in Figures 1 and 2 will be our running examples all through the article. The critical issue with these examples is whether argument A can be accepted in spite of being defeated by argument B.

For a given set S of arguments, S^+ is the set of arguments that are defeated by the arguments in S. Formally, $S^+ = \{\beta \in \mathcal{A} \mid \alpha \rightharpoonup \beta \text{ for } \alpha \in S\}$. Conversely, for a given argument α, the set α^- is the set of all arguments that defeat α. Formally, $\alpha^- = \{\beta \in \mathcal{A} \mid \beta \rightharpoonup \alpha\}$.

Definition 2 (Conflict-freedom)**.** *Let $\langle \mathcal{A}, \rightharpoonup \rangle$ be an argumentation framework and let $S \subseteq \mathcal{A}$. S is* conflict-free *iff $S \cap S^+ = \emptyset$.*

In other terms, a set of arguments is *conflict free* if and only if no argument in that set defeats another.

Definition 3 (Defence)**.** *Let $\langle \mathcal{A}, \rightharpoonup \rangle$ be an argumentation framework, let $S \subseteq \mathcal{A}$, and let $\alpha \in \mathcal{A}$. S* defends *α if and only if $\alpha^- \subseteq S^+$. We also say that argument α is* acceptable with respect to *S.*

In other terms, a set of arguments *defends* a given argument if and only if it defeats all its defeaters.

Example 1. *In the graph displayed in Figure 1, the set $\{A, C\}$ is conflict free, but the set $\{A, B\}$ is not, and neither is the set $\{B, C\}$. Because the set $\{C\}$ defeats all the defeaters of A, we can say that the set $\{C\}$ defends argument A. In the graph displayed in Figure 2, the only conflict-free sets (apart from trivial ones containing single arguments) are $\{A, C\}$ and $\{A, D\}$. Either one of the sets $\{C\}$, $\{D\}$, or $\{C, D\}$, defends A against all its defeaters.*

We now define the *characteristic function* of an argumentation framework.

Definition 4 (Characteristic function)**.** *Let $AF = \langle \mathcal{A}, \rightharpoonup \rangle$ be an argumentation framework. The* characteristic function *of AF is $\mathcal{F}_{AF}$: $2^{\mathcal{A}} \rightarrow 2^{\mathcal{A}}$ such that, given $S \subseteq \mathcal{A}$, we have $\mathcal{F}_{AF}(S) = \{\alpha \in \mathcal{A} \mid S \text{ defends } \alpha\}$.*

Applied to an argument set S, the characteristic function returns the set of all arguments defended by S. Because we are only dealing in this article with one argumentation framework at a time, we will use the notation $\mathcal{F}$ instead of $\mathcal{F}_{AF}$.

We now turn to various so-called *extensions* that can characterise the collective acceptability of a set of arguments. Essentially, these extensions provide different possible ways to group self-defending arguments together. These extensions will be used subsequently to define the argument evaluation criteria that we study empirically in this paper.

Definition 5 (Complete/grounded/preferred extensions)**.** *Let S be a conflict-free set of arguments in framework $\langle \mathcal{A}, \rightharpoonup \rangle$.*

- *S is a* complete *extension iff $S = \mathcal{F}(S)$.*
- *S is a* grounded *extension iff it is the minimal complete extension with respect to set inclusion.*
- *S is a* preferred *extension iff it is a maximal complete extension with respect to set inclusion.*

S is a complete extension if and only if *all* arguments defended by S are also in S (that is, if S is a fixed point of the operator $\mathcal{F}$). There may be more than one complete extension, each corresponding to a particular consistent and self-defending viewpoint.

Example 2. *In the graph displayed in Figure 1, the set $\{C\}$ is not a complete extension, because it defends A without including it. The set $\{B\}$ is not a complete extension because it includes B without defending it against C –see Figure 3. The only complete extension is $\{A, C\}$. The graph displayed in Figure 2 has two complete extensions, $\{A, C\}$ and $\{A, D\}$ –see Figure 4.*

A grounded extension contains all the arguments in the graph that are not defeated, as well as all the arguments

which are defended directly or indirectly by non-defeated arguments. This can be seen as a non-committal view (characterised by the *least* fixed point of $\mathcal{F}$). As such, there always exists a unique grounded extension.

Example 3. *The graph in Figure 1 has only one complete extension, $\{A,C\}$, which is also its grounded extension. The graph in Figure 2 has two complete extensions $\{A,C\}$ and $\{A,D\}$, but none of this is the grounded extension, because there is no node in the graph that is initially undefeated. In that case, the grounded extension is the empty set.*

A preferred extension is a bolder, more committed position that cannot be extended (by accepting more arguments) without causing inconsistency. Thus a preferred extension can be thought of as a maximal consistent set of hypotheses. There may be multiple preferred extensions, and the grounded extension is included in all of them.

Example 4. *The graph in Figure 1 has only one complete extension, $\{A,C\}$, which is also a preferred extension. The graph displayed in Figure 2 has two complete extensions $\{A,C\}$ and $\{A,D\}$, and both qualify as preferred extensions.*

Now we can define the status of an individual argument within the graph, that is, we can define criteria for accepting or not each individual argument. The main question in this paper is whether people evaluate a reinstated argument sceptically or credulously in accordance with the definition below.

Definition 6 (Argument status)**.** *Let $\langle \mathcal{A}, \rightharpoonup \rangle$ be an argumentation framework, and $\mathcal{E}_1, \ldots, \mathcal{E}_n$ its extensions under a given semantics. Let $\alpha \in \mathcal{A}$ and $i = 1, \ldots, n$.*

- *α is accepted in the sceptical sense iff $\alpha \in \mathcal{E}_i$, $\forall \mathcal{E}_i$.*
- *α is accepted in the credulous sense iff $\exists \mathcal{E}_i$ where $\alpha \in \mathcal{E}_i$.*
- *α is rejected iff $\nexists \mathcal{E}_i$ such that $\alpha \in \mathcal{E}_i$.*

Under the grounded semantics, any argument that belongs to the unique grounded extension is accepted both in the credulous and the sceptical sense, and any argument that does not belong to the unique grounded extension is rejected. Under the preferred semantics, an argument is sceptically accepted if it belongs to all preferred extensions; but it can also be credulously accepted if it belongs to at least one preferred extension. If an argument is neither sceptically nor credulously accepted, it is rejected.

Example 5. *The graph displayed in Figure 1 has only one complete extension, $\{A,C\}$, which is grounded as well as preferred. As a consequence, arguments A and C are accepted by grounded as well as preferred semantics, both in the credulous and sceptical sense. The graph displayed in Figure 2 has an empty grounded extension, which means that no argument should be accepted under a grounded semantics. Under a preferred semantics, though, two extensions are identified, $\{A,C\}$ and $\{A,D\}$. From these extensions, only A can be accepted in a sceptical sense, but A, C, and D can all be accepted in a credulous sense.*

What Validates a Semantics?

As established above, different semantics can have different takes on which arguments can be accepted within a given argumentation framework. The question then arises of evaluating the different claims made by different semantics.

Most semantics for argumentation-based reasoning in Artificial Intelligence are based on intuition as to what constitutes correct reasoning. This intuition is informed by specific (hypothetical or real) argumentation scenarios in which a particular semantics draws the desired intuitive answer. This *example-based approach* (to borrow a term from Baroni & Giacomin, 2007) is problematic, since one can often construct other examples with the same logical structure, in which the proposed semantics draws counter-intuitive conclusions. For example, Horty (2002) famously devoted a whole paper to demonstrate counter-intuitive results with floating conclusions in default reasoning. Baroni and Giacomin (2007) made a compelling case for the limitations of the example-based approach, noting that even in relatively simple examples, there might not be a consensual intuition on what should be the correct conclusion. In parallel, Prakken (2002) observed that intuitions about given examples were helpful for generating new investigations, but less helpful as critical tests between different semantics.

To overcome the limitations of the example-based approach, a number of authors recently advocated a more systematic, axiomatic, *principle-based* approach. In this approach, alternative semantics are evaluated by analysing whether they satisfy certain principles, or quality postulates. Such postulates include the *reinstatement criterion*, according to which an argument must be included in any extension that reinstates it, and *directionality criterion* which requires that an argument's status should only be affected by the status of its defeaters (Baroni & Giacomin, 2007).

The principle-based approach provides a significant improvement over the basic example-based approach, since it enables claims that transcend individual examples and characterise semantics more generally. The source of the general postulates, however, is still the researcher's intuition as to what correct reasoning ought to be. In sum, most of the extent validation of various argumentation semantics, example-based or principle-based, relies on normative claims based on intuition. We now suggest that this normative-intuitive perspective could be adequately complemented with descriptive, *experimental* evidence about how people actually reason from conflicting arguments.

The Experiment-based Approach

There is a growing concern within the Artificial Intelligence community that logicians and computer scientists ought to give serious attention to cognitive plausibility when assessing formal models of reasoning, argumentation and decision-making. For example, Benthem (2008) strongly supports the rise of a *new psychologism* in logic at large, arguing that although logicians and computer scientists have tended to go by

intuition and anecdotal evidence, formal theories can be modified under pressure from evidence obtained though careful experimental design.

Pelletier and Elio (2005) also argued extensively for the importance of experimental data when formalizing default and inheritance reasoning, arguing that default reasoning is particularly psychologistic in that it is *defined* by what people do. Their own results have been complemented by a dynamic experimental literature consisting of controlled tests of human default reasoning (e.g., Bonnefon, Da Silva Neves, Dubois, & Prade, 2008; Ford & Billington, 2000; Pfeifer & Kleiter, 2009).

Finally, and in close relation to the problems of simple and floating reinstatement that we have introduced in the previous section, Horty (2002) implicitly appealed to descriptive validation when highlighting the issues that floating conclusions raise for sceptical semantics:

> There is a vivid practical difference between the two skeptical alternatives. [...] Which alternative is correct? I have not done a formal survey, but most of the people to whom I have presented this example are suspicious of the floating conclusion (p.64).

We believe that the field of computational argumentation can indeed benefit from the same kind of formal surveys that have been conducted in the field of default reasoning, and that have been generally called for in Artificial Intelligence. To our knowledge, only very few articles have explicitly sought to inform formal models of argumentation with experimental evidence, and these experimental data have only been collected in relation to the specific issue of argumentation-based decision making (e.g., Dubois, Fargier, & Bonnefon, 2008). What we offer in this article is an experimental investigation of the basic issue of how people reason from the complex argument structure corresponding to floating reinstatement, and whether one of the current available semantics can capture their reasoning.

Recently, we conducted experiments on the simple reinstatement structure, across a varied set of linguistic contents (Madakkatel, Rahwan, Bonnefon, Awan, & Abdallah, 2009). Our study revealed that participants reasoned in a way that reflected the formal notions of defeat and reinstatement: Their confidence in an argument A decreased when it was attacked by an argument B, but bounced back up when B itself was attacked by a third argument C. These findings are in agreement with grounded as well as preferred semantics (and others). What neither semantics could predict, though, is the finding that the recovery of argument A was not complete when reinstated by argument C: Confidence in A in presence of B and C did not raise back to its former level, when A was presented alone.

Our present study offers an experimental comparison of the simple reinstatement structure to the more complex structure known as floating reinstatement, shown in Figure 2.The present study seeks to answer the following questions: Does floating reinstatement restore the confidence in the conclusion of argument A, and does it do so to the same extent as simple reinstatement? (A 'yes' to both questions would go against the predictions of grounded semantics.) If so, does the effectiveness of floating reinstatement require that participants manifest a preference for either C over D or D over C? (A 'yes' would provide support to the predictions of credulous preferred semantics, a 'no' would provide support to the predictions of sceptical preferred semantics.)

Method

Fourty-seven participants were randomly approached in offices, shopping malls, and open spaces in Dubai. Participants read an introduction to the task, informing them that the purpose of the experiment was to collect information about how people thought, that the task included no trick question, and that they simply had to mark the answer that they felt correct. They were randomly assigned to two experimental groups corresponding to simple and floating reinstatement, respectively, then solved 12 problems, following a 3-level, 4-measure within-participant design.

The 3-level independent variable was the *Pattern* of the problem (Base, Defeated, Reinstated). In the Base pattern, participants were only presented with argument A; in the Defeated pattern, participants were presented with arguments A and B; finally, in the Reinstated pattern, participants were presented with the three arguments A, B, and C (in the simple reinstatement group) or with the four arguments A, B, C, and D (in the floating reinstatement group).

The linguistic contents of arguments A, B, C, and D were taken from four different argument sets (see Appendix). All participants saw each argument set in its Base, Defeated, and Reinstated versions. The order of argument sets within the questionnaire was counterbalanced across participants (two different orders), but the order of Pattern within each argument set was fixed across the experiment. Participants had to answer every problem, in the order they appeared in the questionnaire, without looking at the next problem in the questionnaire. For each problem, participants had to assess the conclusion of argument A, using a 7-point scale anchored at *certainly false* and *certainly true*.

In addition, participants rated their understanding of each problem ('How clearly did you understand the problem?') on a 7-point scale anchored at *Not at all* and *Completely*. Lastly, participants in the floating reinstatement group answered the following question about the four reinstated problems: Do you think that (i) C is a better argument than D, (ii) D is a better argument than C, or (iii) C and D are about equally good?

Results

Figure 5 displays the average confidence in the conclusion of A, as a function of Pattern and Type of reinstatement, averaged across the contents and participants. The visual inspection of Figure 5 already suggests that the results are very similar for the two groups. This preliminary intuition was confirmed by the results of a mixed-design analysis of variance, using the confidence in the conclusion as a dependent

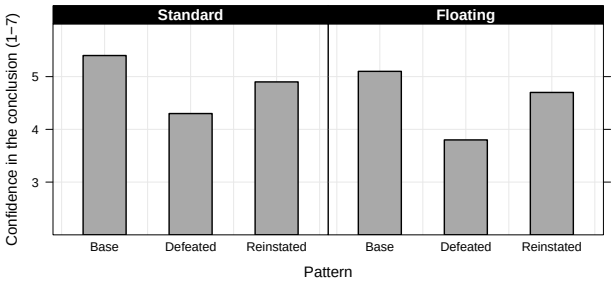


Figure 5: Reinstatement is as effective in its floating form as in its simple form. Confidence in the conclusion of an argument decreases when the argument is defeated, and is then imperfectly restored when its defeater is itself defeated, whether by a single argument (simple reinstatement) or by two mutually defeating arguments (floating reinstatement).

variable, pattern as a 3-level within-subject predictor (Base, Defeated, Reinstated), the type of reinstatement as a 2-level between-group variable (Simple, Floating), and four measures corresponding to the four linguistic contents.

The multivariate test detected a significant effect of Pattern, $F(8,38) = 6.1$, $p < .001$, $\eta_p^2 = .56$. It did not, however, detect a significant main effect of Type of reinstatement $F(4,42) < 1$, $p = .79$, $\eta_p^2 = .04$, nor a significant interaction between Pattern and Type, $F(8,38) = 1.2$, $p = .32$, $\eta_p^2 = .20$.

The overall effect of Pattern reflected a successful defeat followed by a successful reinstatement. As shown by contrast analysis, confidence ratings in the Defeated condition were significantly lower than ratings in the Base condition, $F(1,45) = 34.9$, $p < .001$, $\eta_p^2 = .44$, and this difference was not moderated by the Type of reinstatement (there is indeed no reason that it should be), $F(1,45) < 1$, $p = .67$, $\eta_p^2 < .01$. The confidence ratings in the Reinstated condition were significantly greater than in the Defeated condition, $F(1,45) = 13.7$, $p < .001$, $\eta_p^2 = .23$, and this difference (more interestingly this time) was not moderated by the Type of reinstatement, $F(1,45) < 1$, $p = .60$, $\eta_p^2 < .01$. Just as in our earlier study (Madakkatel et al., 2009), reinstatement is not perfect, as ratings in the Reinstated condition remain significantly lower than in the Base condition, $F(1,45) = 9.0$, $p < .01$, $\eta_p^2 = .17$. Again, there is no evidence whatsoever of a moderation by Type of reinstatement, $F(1,45) < 1$, $p = .92$, $\eta_p^2 < .01$.

So far, results suggest that floating reinstatement has an effect that is identical to classic reinstatement. We further note that although subjects found the floating reinstatement problems slightly harder to understand than the simple reinstatement problems, this difference appeared to play no role in the ratings they gave for their confidence in the conclusion. The average understanding rating was 4.6 (SD = 1.1) for simple reinstatement problems, compared to 4.0 (SD = 0.9) for floating reinstatement problems, $t(45) = 2.0$, $p = .05$. However, a regression analysis seeking to predict acceptance of reinstated arguments on the basis of problem understanding, Type of reinstatement (dummy coded, 1 for floating), and the interaction term between these two predictors, failed to find any significant effect. The interaction term in particular achieved a standardized β of .19, non-reliably different from zero, $t = 0.32$, $p = .75$.

The effectiveness of floating reinstatement does not appear to result from the subjects manifesting a preference for one of the mutually defeated arguments. We conducted four repeated-measure analyses of variance, one for each argument set, with conclusion acceptance as a dependent variable, pattern as a 2-level predictor (Defeated, Reinstated), and preference as a dummy coded between-group variable (0 for subjects who said the two mutually defeating arguments were equally good, 1 otherwise). The interaction term between pattern and preference did not achieve statistical significance in any of the four analyses, all Fs in the $0.5 - 1.5$ range, all ps in the $.23 - .48$ range.

Discussion

We applied the experimental approach to understand how people deal with floating reinstatement in argument-based reasoning, a case that has puzzled theoreticians for many years. Our results suggest that, empirically speaking, floating reinstatement works exactly as well as simple reinstatement. Participants' confidence in an argument A decreased when it was attacked by an argument B, but bounced back up when B itself was attacked by two mutually defeating arguments C and D. These results clearly speak in favour of preferred semantics. Results also suggest that the sceptical version of preferred semantics might be more cognitively plausible than the credulous version, since the effect of floating reinstatement was not dependent on participants showing a preference for one of the two mutually defeating arguments. This question is not yet settled, though, since the data do not make it clear whether participants would be willing to commit to accepting one of the mutually defeating arguments C and D. This aspect requires further investigation.

Besides their theoretical value, our results also have applied value for developing agents that are meant to argue with human users. We already know that artificial agents can achieve better negotiation results with human users when they do not play normative equilibrium strategies, but rather adopt boundedly rational strategies inspired from human behavioural data (Lin, Kraus, Wilkenfeld, & Barry, 2008). Generally speaking, we may expect that artificial agents may similarly be more successful when arguing with human users, if they can anticipate human reactions to various abstract argumentation frameworks. With that goal in mind, our results suggest that artificial agents may be better off avoiding discussion that may reveal a defeater, even if the agent has a counter-argument to that defeater; but should be ready to use floating reinstatement as well as simple reinstatement in order to neutralise a defeater raised by the human user. These kinds

of heuristics can be incorporated into a decision-theoretic model of a persuasive agent that interacts with users using natural language (e.g. to promote a healthy diet (Mazzotta, Rosis, & Carofiglio, 2007). Going beyond our specific results, by building up a corpus of argument structures and how they are evaluated, it may be possible to use machine learning techniques to build models that predict how people will react to novel argument structures.

Independently of our specific results, we hope to have convinced the reader that the wealth of scientific methodology from psychology can give a new perspective on the problems raised when formalising argumentation and developing argument evaluation semantics. We hope that our claims and findings can prompt researchers working on the computational modelling of argument to explore new avenues of investigation inspired by, and validated against, empirical evidence from psychology and cognitive science.

We also hope to have excited cognitive scientists about the growing literature on formal models of argumentation. These models, and their associated normative properties, have great potential in complementing existing research on human reasoning, and providing conceptual means for dealing with highly complex inference structures.

References

Baroni, P., & Giacomin, M. (2007). On principle-based evaluation of extension-based argumentation semantics. *Artificial Intelligence*, *171*(10–15), 675–700.

Benthem, J. van. (2008). Logic and reasoning: do the facts matter? *Studia Logica*, *88*(1), 67-84.

Bonnefon, J. F., Da Silva Neves, R. M., Dubois, D., & Prade, H. (2008). Predicting causality ascriptions from background knowledge: Model and experimental validation. *International Journal of Approximate Reasoning*, *48*, 752–765.

Dubois, D., Fargier, H., & Bonnefon, J. F. (2008). On the qualitative comparison of decisions having positive and negative features. *Journal of Artificial Intelligence Research*, *32*, 385–417.

Dung, P. M. (1995). On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, *77*(2), 321–358.

Ford, M., & Billington, D. (2000). Strategies in human nonmonotonic reasoning. *Computational Intelligence*, *16*, 446–468.

Horty, J. F. (2002). Skepticism and floating conclusions. *Artificial Intelligence*, *135*(1-2), 55-72.

Lin, R., Kraus, S., Wilkenfeld, J., & Barry, J. (2008). Negotiating with bounded rational agents in environments with incomplete information using an automated agent. *Artificial Intelligence*, (accepted).

Madakkatel, M. I., Rahwan, I., Bonnefon, J.-F., Awan, R. N., & Abdallah, S. (2009). Formal argumentation and human reasoning: The case of reinstatement. In T. Bench-Capon, S. Parsons, & H. Prakken (Eds.), *The uses of computational argumentation: Papers from the aaai fall symposium.* AAAI Press. (Technical Report FS-09-06)

Mazzotta, I., Rosis, F. de, & Carofiglio, V. (2007). Portia: A user-adapted persuasion system in the healthy-eating domain. *IEEE Intelligent Systems*, *22*(6), 42-51.

Pelletier, F. J., & Elio, R. (2005). The case for psychologism in default and inheritance reasoning. *Synthese*, *146*(1-2), 7-35.

Pfeifer, N., & Kleiter, G. D. (2009). Framing human inference by coherence based probability logic. *Journal of Applied Logic*, *7*, 206–217.

Prakken, H. (2002). Intuitions and the modelling of defeasible reasoning: some case studies. In S. Benferhat & E. Giunchiglia (Eds.), *Proceedings of the 9th international workshop on non-monotonic reasoning* (p. 91102).

Rahwan, I., & Simari, G. R. (Eds.). (2009). *Argumentation in artificial intelligence*. Springer.

Materials

Argument Set 1

(A) Cody does not fly. Therefore, Cody is unable to escape by flying.

(B) Cody is a bird. Therefore, Cody flies.

(C) Cody is a rabbit. Therefore, Cody is not a bird.

(D) Cody is a cat. Therefore, Cody is not a bird.

Argument Set 2

(A) Smith does not follow American spelling. Therefore, Smith writes 'colour' instead of 'color'.

(B) Smith speaks American English. Therefore, Smith follows American spelling.

(C) Smith was born and brought up in England. Therefore, does not speak American English.

(D) Smith was born and brought up in Australia. Therefore, does not speak American English .

Argument Set 3

(A) The car did not slow down. Therefore, the car approached the signal at the same speed or higher.

(B) Louis applied the brake. Therefore, the car slowed down.

(C) Louis applied the accelerator instead. Therefore, Louis did not apply the brake.

(D) Louis applied the clutch instead. Therefore, Louis did not apply the brake.

Argument Set 4

(A) Stephen is not guilty. Therefore, Stephen is to be free from conviction.

(B) Stephen was seen at the crime scene at the time of the crime. Therefore, Stephen is guilty.

(C) Stephen was having dinner with his family at the time of crime. Therefore, Stephen was not seen at the crime scene at the time of the crime.

(D) Stephen was watching football with his friends in the stadium at the time of the crime. Therefore, Stephen was not seen at the crime at the time of the crime.

Optimizing Learning Environments: An Individual Difference Approach to Learning and Transfer

Daniel M. Belenky (dmb83@pitt.edu)
Timothy J. Nokes (nokes@pitt.edu)
Learning Research and Development Center
University of Pittsburgh, 3939 O'Hara Street
Pittsburgh, PA 15260

Abstract

Prior work has found that the type of learning activity (direct instruction or invention) interacts with achievement goals (mastery or performance-oriented) such that invention tasks can help facilitate mastery goal adoption and knowledge transfer (Belenky & Nokes, 2009). In the current study, we investigated how robust the effect is, and whether explicit manipulations of the task goals can produce a similar effect. We conducted an experiment with 98 college students in which achievement goals were measured, while task goals and task structure were manipulated between subjects. Results indicated that task structure was generally a more effective way of influencing which achievement goals are adopted within a learning activity. However, task goals that promoted an evaluative context interfered with transfer for mastery-oriented learners from invention activities. The results are interpreted in relation to theories of regulatory fit and multiple goal hierarchies.

Keywords: learning; transfer; skill acquisition; motivation; achievement goals.

Student's achievement goals have a large influence on their behaviors and experiences in academic settings. The literature surrounding Achievement Goal theory shows that these goals lead to very different patterns of affect, interest and achievement (e.g., Harackiewicz et al., 2005). However, this literature has not focused on how the goals influence what is learned. That is, although "achievement" is frequently measured as an outcome, it is almost always done at a coarse-grain level, such as final grades in a course. It is not clear how different achievement goals (mastery versus performance) are related to different kinds of learning, such as learning procedural skills, simple facts, or conceptual knowledge.

To begin to address this gap, Belenky & Nokes (2009) examined how achievement goals impact the type of knowledge gained from different kinds of instruction. That study found that mastery-oriented learners do better on transfer measures, regardless of whether the mastery-orientation came from a stable predisposition or whether the open-ended structure of an "invention" task led to mastery-oriented feelings and goals in the specific context. Conversely, performance-oriented learners did better on skill acquisition when the instruction seemed to match their goals, by presenting a well-structured, simple task through direct instruction.

This initial work has provided evidence that task structure interacts with existing achievement goals to influence learning. In the current work we examine whether direct manipulations of task goals through instructions can change the ways students learn, similar to the effect of task structure. If directly manipulating task goals produces similar effects, it would offer a more direct way of encouraging students towards desired learning outcomes (whether towards transfer or skill). However, it is possible that achievement goals within a learning activity are not under conscious control, and task structure has more influence on how a student engages than instructions that attempt to prompt a particular achievement goal. It is also possible that task structure and task goals operate independently, leading to a three-way interaction in the adoption of achievement goals based on students' prior dispositions. This study explores these possibilities.

Background

Research on achievement goals has focused on two main aims; classifying what the goals are and then correlating those with predictors and outcomes. The prevailing classification is a 2 x 2 framework that has been well-validated (Elliot & McGregor, 2001). This framework separates the evaluative criterion (mastery or performance) from the valence (approach or avoidance), which results in four separable goals (mastery-approach, mastery-avoidance, performance-approach, performance-avoidance). Mastery goals refer to ones in which a person is basing his evaluation on the skill or competence he is trying to develop (that is, in comparison to an expectation or prior ability), while performance goals refer to evaluating oneself based on a normative standard (that is, in comparison to others). Approach goals refer to seeking out positive outcomes, while avoidance goals refer to averting negative ones. For example, a mastery-approach goal is one in which a person is seeking to improve his ability or knowledge, based on an internally-referenced criterion ("My aim is to completely master the material in this class"), while a performance-avoidance goal is one in which a person is seeking to not look bad compared to others ("My aim is to avoid doing worse than other students;" see Elliot & McGregor, 2001). Students can have different levels of each of these goals, and studies have validated that these four goals are separate factors (Elliot & McGregor, 2001). Because we are most interested in studying how different paths of *successful* learning affect what knowledge is gained, our work focuses on mastery-approach and performance-approach goals.

Mastery-approach (MAP) goals have been correlated with a host of positive outcomes, such as intrinsic motivation, interest, better self-regulation, and deeper strategy use.

However, MAP goals are generally unrelated to achievement scores (usually operationalized as exam scores or final grades). Performance-approach (PAP) goals have been correlated with some positive outcomes, such as perseverance, task engagement, enjoyment and topic interest, as well as some negative outcomes, such as test anxiety, poor help-seeking and, most importantly for the current study, shallow cognitive strategies. PAP goals are also positively correlated with achievement scores (see Elliot, 1999 for a review). One potential reason that MAP goals are unrelated to achievement scores but PAP goals correlate positively could be due to the type of knowledge being assessed on achievement measures. Scores on a final exam, for example, may reflect a person's ability to recall factual information (a performance-oriented outcome) more than a deep conceptual understanding (a mastery-oriented outcome).

Belenky & Nokes (2009) examined the possibility that different types of knowledge were being generated due to students' existing achievement goals. This work found that MAP goals led to more flexible use of knowledge on a transfer assessment, while PAP goals led to better procedural skill with a formula. That study also showed that an ill-structured, invention-based learning task promoted mastery-goal adoption. This goal adoption was particularly beneficial for those who entered the study low in MAP goals; this group performed one standard deviation better than low-MAP counterparts who completed a well-structured, direct instruction-based approach.

The paradigm and tasks used in the current work are based on previous work on "Preparation for Future Learning" (Belenky & Nokes, 2009; Schwartz & Martin, 2004). This paradigm allows one to measure a person's ability to use knowledge acquired in one situation to help learn in a new situation. The work that established this paradigm contrasted two types of learning activities, direct instruction and invention (Schwartz & Martin, 2004). The direct instruction condition was similar to a well-structured, "tell-and-practice" style of pedagogy, where a student is shown a method and asked to use it to solve similar problems. The invention activity was modeled on a form of "discovery learning" where students are asked to construct their own knowledge in an open-ended, ill-structured problem. All students were given a subsequent transfer problem on an extension of the concept. Half of the students from each of the group were given a learning resource (a worked example) embedded in the assessment, while half were not (see Figure 1). They found that only those students who had invented and were given the worked example showed large improvements in the ability to solve a transfer problem. Belenky & Nokes (2009) found evidence that this benefit was due, at least in part, to the adoption of different goals. Support for this view was based on a questionnaire given right after the learning activity showing that students given invention activities were more concerned about their understanding of the concepts than those given tell-and-practice activities, as well as a benefit in transfer

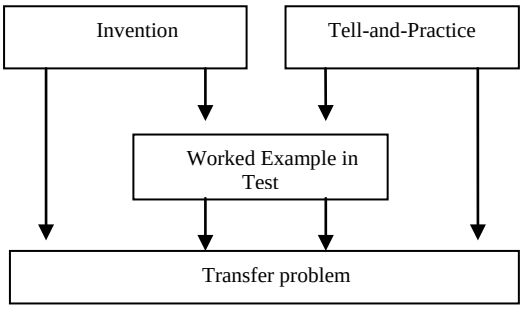


Figure 1. The general Preparation for Future Learning paradigm which measures the ability to transfer from an initial learning experience to a later one.

performance among those students who entered the study low in MAP goals but performed invention activities as opposed to those who performed tell-and-practice activities.

The current study further explores the issue of how achievement goals influence and are influenced by different task structures, by using the same tasks and adding task instructions to try and directly manipulate the achievement goals students adopt. We have evidence that goal adoption, whether due to individual dispositional differences or task structure, influences the form of the knowledge gained from instruction. The robustness of this effect is being examined, by adding the factor of task instructions to the framework.

Hypotheses

We are investigating how explicit, task-based goal instructions influence the effects of existing achievement goals and task structure on learning and transfer. We predict that the task structure will have a larger effect than the task-based instructions on which achievement goals students adopt and their impact on learning and transfer. However, there are several possibilities for interesting interactions between task structure, task goals, and existing achievement motivations. We are most interested in exploring whether there is evidence for a "multiple-goals" framework (Harackiewicz et al., 2002) or a "regulatory fit" viewpoint (Higgins, 2005).

The multiple goals hypothesis states that a mix of performance-approach and mastery-approach goals is optimal for learning (Harackiewicz et al., 2002). If this hypothesis is correct then we would expect that crossing mastery and performance goals would promote the best overall learning. Alternatively, a "regulatory fit" view states that the alignment of task structure, task-based goals and dispositional orientations should facilitate the best learning (Higgins, 2005). This suggests that matching students' goals with instructions that support pursuit of those goals would improve learning, (e.g., mastery-oriented achievement goals with invention tasks and mastery-oriented instructions). If this is the case, we would also expect to see that incongruous goals would harm learning. We will examine these possibilities by assessing three-way interactions

between existing goals, task structure and instructions on different learning outcomes.

We also predict replications of our basic prior findings that invention activities promote more mastery-oriented learning behaviors and feelings than tell-and-practice. We also predict that existing mastery-approach goals will lead to better transfer, while performance-approach goals will lead to better performance on a skill measurement.

Methods

This study closely followed the methods and materials of prior studies (Belenky & Nokes, 2009; Schwartz & Martin, 2004). Participants completed a pre-test, went through a series of learning activities on basic statistical concepts, and then took a post-test in a 2-hour laboratory session. They were given questionnaires at the beginning, middle, and end of the experiment.

Participants

Ninety-eight undergraduates from the University of Pittsburgh participated in this study (M = 19.4 years old, SD = 2.5 years) in exchange for course credit.

Design and Materials

This study had a 2 (task structure: invention or tell-and-practice) x 2 (task goal: mastery-oriented or performance-oriented instructions) between-subjects design. Materials were presented as packets in binders. These packets consisted of an initial questionnaire; a pre-test; learning activities (with an activity questionnaire after the first one); a post-test; and final questionnaires, including a demographic sheet.

Learning Materials

The learning materials consisted of one activity on mean deviation, instruction on the correct calculation of mean deviation (a narrated PowerPoint video), and then a new activity on standardization. The first learning activity presented four different data sets representing the spread of a number of pitches thrown by different pitching machines. The students' task was to decide which of the machines is the most reliable. The invention and tell-and-practice students both attempted this problem, but the instructions each received was different. The tell-and-practice group was given a worked example demonstrating how to calculate mean deviation immediately prior to attempting to solve this problem. The invention group was not given this worked example, and was instructed "Your task is to invent a procedure for computing a quantity that expresses the variability for each of the pitching machines and decide which is most reliable. There is no single way to do this, but you have to use the same procedure for each machine, so it is a fair comparison." Both groups were given access to scrap paper and a calculator during this activity.

The video demonstrated a brief introduction on variability before walking through the calculation of mean deviation in a worked example. This was followed by two simple problems to work on, with solutions demonstrated to make sure students understood the basics of the formula.

The structure of the standardization activity was similar to the mean deviation activity. The invention group was asked to evaluate which of two world records was "more shattered," given two small data sets and the corresponding world record for each. The tell-and-practice group was given the exact same problem to solve, but was first shown how to graphically arrive at a solution, by divvying up a visual representation of the distributions through a worked example.

We manipulated task goals through instructions. Immediately below the tell-and-practice or invention instructions, participants saw instructions about the purpose of the task. These were constructed to spur participants towards either a mastery or performance orientation while working on the learning activities. The motivation goal instructions were modeled on previous work that had experimentally manipulated goals (i.e., Elliott & Dweck, 1988; Elliot & Harackiewicz, 1996). The mastery goal instructions were: "Many people see problems like this one as a challenge, and feel like they are developing their skill to solve these types of problems. While working on this problem, you may make mistakes and feel a little confused at times, but in the end you will have learned some things and developed your skill to solve problems like this one." The performance goal instructions were: "This problem assesses your mathematical ability. People who can solve this problem generally have the capability to solve similar problems. While working on this activity, you can gauge how good you are at these types of math problems." These were presented underlined, to make them more salient.

Test Materials

The pre-test consisted of three items: a skill measure, a data representation problem, and a transfer problem. The post-test contained isomorphic versions of these problems, as well as an adaptive use and a qualitative reasoning problem. We will only focus on the transfer and skill measures here. The skill measures presented small data sets and explicitly asked the participants to calculate mean, mode, median and mean deviation. The transfer problems were both word problems that presented descriptive statistics for two data distributions and one exceptional score from each, and asked which individual score was more impressive. While similar to the standardization activity, this problem assesses transfer because it requires reasoning from descriptive statistics, not raw data, and because simpler heuristic processing would lead to an incorrect answer (i.e., reasoning about range, or because one value is higher than the other).

The skill measure was scored dichotomously as a 0 if incorrect, while a 1 was awarded if the student flawlessly used the formula. All other problems were coded as a 0 if incorrect, 1 if conceptually correct but there was a computational error, and 2 if the answer was conceptually and computationally correct. For the transfer problem, this meant calculating standardized scores and using them to decide which value was more impressive.

The test also included a worked example on how to calculate a standardized score. This was presented just like

the other test problems, and described a situation in which one would want to calculate a standardized score to compare values from different samples. The text then introduced the formula to do so and computed the values for the data presented. This was followed by a second, very simple data set, and asked the students to use the formula on these data. The worked example always came at least two problems before the transfer problem, so if a student used the formula on the transfer problem, it was because they noticed that it applied and could recall it, not due to temporal contiguity.

Motivation Measures

To assess achievement goals, we used the Achievement Goal Questionnaire (AGQ; Elliot & McGregor, 2001). This 12-item scale has 3 items for each of the 4 achievement goal constructs. We focus only on mastery-approach and performance-approach goals in this study. The questions were phrased to be specifically about math classes, and were assessed on a 7-point Likert scale. Cronbach's alphas for each of the scales was high (MAP = .839, PAP = .932). There was also an activity questionnaire, which was the same as that used in Belenky & Nokes (2009). This 8-item measure asked about student's focus and affect while working on the first learning activity. At the end of the study, the participants completed additional questionnaires – the AGQ again, and a questionnaire we developed to assess how students reflected on and our goal manipulations.

Procedure

The study was run in groups of up to six participants in a two-hour laboratory session, with all participants working individually. Inside the packet was: an initial questionnaire; a pre-test; a learning activity; the activity questionnaire; space to work on problems presented in the video; another learning activity; a post-test; a final set of questionnaires; a demographics sheet. Participants took as long as they needed to complete the questionnaires, with no one taking longer than three minutes for any one questionnaire. Both learning activity and the video took fifteen minutes each. Participants were given five minutes for each test item.

Results

Our hypotheses focused on the conceptual replication of earlier findings, as well as exploring how explicitly manipulating task goals and task structure would influence learning. First, we assess whether task structure and instructions influenced students' self-reported experiences during the learning activity. Then, we examine the competing "regulatory fit" and "multiple-goals" hypotheses by looking at the interactions of existing goals, learning tasks, and instructions.

Motivational effects on questionnaires

On the activity questionnaire, administered directly after the first learning problem, there were telling differences between responses from the invention and tell-and-practice groups (see Table 1). Namely, students who invented felt more concerned with understanding, challenged, and frustrated, Fs (1, 94) > 6.87, ps < .01. While frustration is generally considered a negative outcome, it may signal to a student a lack of understanding that drives further cognitive engagement. Notably, there were no differences due to our task goal manipulation, nor interactions. This is evidence that that the structure of the task led to very different experiences for the learner, both in terms of affective response and goals adopted, and that this structure has more influence on goal adoption than explicit instructions about the purpose of the task.

We also asked a similar set of reflective questions at the end of the experiment. The only observed significant differences were that students who completed invention tasks reported being able to be more creative, while those who received mastery instructions reported enjoying the activity more. There was also an interaction between the experimental manipulations on the item "it was a challenge to come up with the correct solution." This interaction was driven by the lower ratings on this item provided by those who completed the tell-and-practice activities with performance- oriented task goal. This manipulation seems to have led students to feel less challenged, even compared to the students who completed the same activity but were told that the task could help them improve. The results from the reflective questionnaire seem to suggest that manipulations of goals do not produce large differences in students' conscious, reflective experience of learning.

Interactions

We predicted that we would see three-way interactions between existing achievement goals, task structure, and task goals, specifically looking at performance-approach goals (PAP) when considering skill acquisition and mastery-approach goals (MAP) for transfer. To assess the effect of existing achievement goals, we performed a median split within each experimental group on the performance-approach and mastery-approach construct scores, based on the initial questionnaire, administered at the start of the experiment. We used the mean deviation skill measurement as the dependant variable in a 2 (task structure) x 2 (task

Table 1. Activity questionnaire items with differences between Tell-and-Practice (TP) and Invent. † $p < .1$, * $p < .05$.

Questionnaire Item	*Invent*	*TP*	*Effect Size d*
I was concerned with how well I understood the procedure I was using.*	3.7 (1.2)	3.0 (1.4)	0.54
I was concerned that the procedure I was using was not correct.*	4.0 (1.1)	2.5 (1.4)	1.17
I felt engaged. †	3.1 (1.3)	3.6 (1.0)	-0.37
I felt frustrated.*	3.5 (1.4)	2.6 (1.3)	0.65
I felt challenged. *	3.7 (1.1)	2.8 (1.2)	0.86

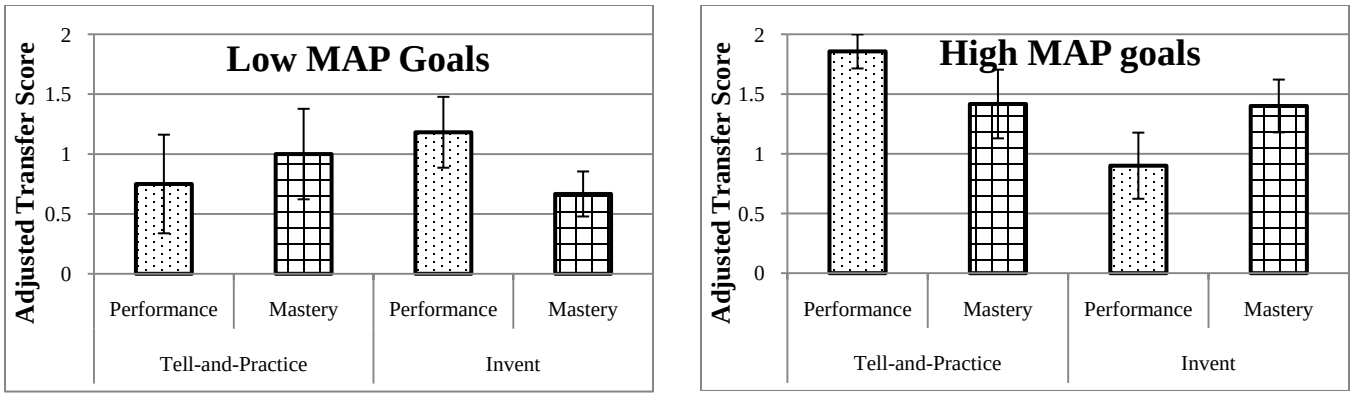


Figure 2. Adjusted transfer score (out of 2, post-test minus pre-test), for those below median in mastery-approach goals (MAP) in the left graph, and those above median in mastery-approach goals on the right, +/- 1 S.E.

goal instructions) x 2 (high or low performance-approach) ANOVA. There was no effect of instructional goal, F (1, 88) < 1, *ns*, but there was an effect of task structure, F (1,88) = 4.09, $p < .05$, with tell-and-practice leading to better performance (77% correct) than invention (57% correct). There was also a significant effect for PAP goals; those high in PAP goals (77% correct) did better than those who entered low in such goals (56% correct, F (1, 88) = 4.59, $p < .05$. However, the experimental manipulations did not interact with existing goals, nor was the three-way interaction significant, Fs (1, 88) < 2.48, *ns*. For learning a simple skill, the dispositional goal of being able to perform leads to better acquisition, as does a well-structured learning task.

To look at transfer, an adjusted score was calculated. This value is the post-test score minus pre-test score on two isomorphic transfer problems, which were counterbalanced. Participants who were at ceiling on pre-test were taken out of subsequent analyses of transfer, as they already had knowledge of the formula, and when it applies. No effect of either task structure or task goals on transfer performance was found. Invention activities produced the same level of transfer as tell-and-practice, as did mastery task goals and performance task goals, and there was no interaction between these factors, Fs (1,70) < 1.16, *ns*. However, students high in MAP goals did better (M = 1.36, SD = .84) than those low in MAP goals (M = .90, SD = .94) on the measure of transfer, F (1, 70) = 5.85, $p < .05$. When paired with the result that PAP goals lead to better skill, this is clear evidence that existing student goals affect the type of learning students exit instruction with. There was also a significant three-way interaction, F (1, 70) = 4.36, $p < .05$ (see Figure 2). This interaction is due to the invent structure and performance task goals condition behaving differently than all of the other conditions in terms of the effect of existing MAP goals. In the other conditions, high MAP goals produce better transfer than low MAP goals, as per the main effect. However, within the invention/performance-oriented cell, those low in MAP goals did slightly better than those high in MAP goals (Ms = 1.18, .9, SDs = .98, 88, respectively). It appears that an ill-structured environment like the invention activity, when presented in an evaluative context, changes the pattern of learning a mastery-oriented student might normally engage in, harming their ability to flexibly transfer their knowledge.

Discussion

This study examined whether explicit task goals would affect learning the same way task structures have been shown to. The evidence shows that they do not. Task structure was a bigger determinant in the learning outcome, as evidenced by the main effect of a well-structured domain helping with skill acquisition, and the difference in goal adoption on the activity questionnaire, which showed differences based on task structure, but not based on the task goals received. It seems that telling students which goal to adopt is not a particularly effective way to change learning behaviors or affect, and that task structure is a better way to change a student's focus during an instructional event.

The results provide strong evidence that existing student achievement goals change what a student learns during the course of instruction. Those high in performance-approach goals did better on skill measures, regardless of which task structure they learned with, while those high in mastery-approach goals did better on the transfer measure. These differences on their own are illuminating in light of the null effects for mastery-approach goals on achievement measures (see Harackiewicz et al., 2002). This may be due to the types of measures used when achievement is measured in schools. A grade on the test at the end of the semester may reflect factual knowledge that a performance-oriented student has been more focused on attaining than the deep understanding a mastery-oriented student focused on.

Finally, this work also examined whether having multiple goals would help one to learn, or if incongruous goals would produce a poor "fit" and interfere with successful learning. Within a given outcome (i.e., skill or transfer), we do not see evidence that multiple goals are best. On the skill measure, we see no interaction between existing goals, task structure and instructional goals. For transfer, we see evidence for a "regulatory fit" model, where a task goal that did not fit with the task structure and an existing achievement goal harmed learning. The focus on ability created by performance-oriented instructions may have

changed the way in which these students would have normally processed the material, which, based on the performance of their equivalent group who received mastery-oriented instructions, and from our prior study, would have performed much better. The evaluative context brought on by the performance instructions appears to have fundamentally changed the way these students engaged in the learning activity, producing worse transfer. Though we have discussed this in terms of regulatory fit, this adverse effect could also be due to anxiety, or the use of simpler learning strategies related with performance goals (Elliot, 1999).

Conclusion

The study of student motivation has much to offer researchers focused on *how* students learn. It seems naïve to believe that the goal a student has in a learning environment does not influence the form and utility of the knowledge generated, and this line of work is bearing this out empirically. Specifically, research has found that the way information is represented and processed will have an effect on how that information is used to solve new problems and learn new concepts (e.g., Nokes & Ohlsson, 2005). A student's goals for a learning environment can be a catalyst for different types of processing and representations, as demonstrated by performance-approach goals predicting performance on measures of skill and mastery-approach goals predicting performance on transfer tests.

That we can reliably show these differences in a laboratory setting seems to be evidence that these results have high external validity, as student goals should be even more salient in authentic academic environments than in the lab. Also, while our attempts to manipulate these goals did not produce effects similar to existing achievement goals, stronger interventions conducted over longer periods of time may have very profound effects on student learning, especially when they succeed in changing the goal itself (i.e., Blackwell, Trzesniewski, & Dweck, 2007).

Within a task, it appears that making the learning goals, task structure, and instructions as coherent as possible promotes better learning. Those high in mastery-approach goals who invented were disrupted by instructions which placed the task into a more evaluative context. While the results within a given learning measure (e.g., skill or transfer) do not support a "multiple-goals" viewpoint, the study as a whole does. The separation of skill and transfer seems to illustrate one potential mechanism in support of a multiple-goals viewpoint, which claims that a combination of performance-approach and mastery-approach goals is optimal (Harackiewicz et al., 2002). Our work illustrates that having both goals would allow for the development of efficient skills at routine aspects (i.e., performance-approach goals lead to improved formula use) as well as innovative ability to use the underlying concept (i.e., mastery-approach goals lead to improved transfer). This combination of efficiency and innovation is hypothesized to be necessary for adaptive expertise, a desired outcome of education (Hatano & Inagaki, 1986). If the multiple-goals viewpoint is correct, a critical question left to examine is if students flexibly switch between these goals, and how they do so. It may be that the time course is a critical factor in explaining why, across a semester, multiple goals may be optimal. A more fine-grained, microgenetic study of when and how each goal contributes could shine light on this possibility.

Future work should also further explore the mechanisms by which achievement goals affect the type of learning done. It could be due to different representations used and formed and/or different types of learning strategies (i.e., self-explanation versus rehearsal). Achievement goals remain an important individual difference to consider, and one that could greatly impact how we can use cognitive science to improve education.

Acknowledgments

This work was supported by the National Science Foundation, Grant Number SBE-0354420 to the Pittsburgh Science of Learning Center (http://www.learnlab.org).

References

Belenky, D.M., & Nokes, T.J. (2009a). Motivation and transfer: The role of achievement goals in preparation for future learning. In N.A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 1163-1168).

Blackwell, L.S., Trzesniewski, K.H., & Dweck, C.S. (2007). Implicit theories of intelligence predict achievement across and adolescent transition: A longitudinal study and an intervention. *Child Development, 78* (1), 246-263.

Elliot, A.J, & Harackiewicz, J.M. (1996). Approach and avoidance achievement goals and intrinsic motivation: a meditational analysis. *Journal of Personality and Social Psychology, 70* (3), 461-475.

Elliot, A.J., & McGregor, H.A. (2001). A 2 X 2 achievement goal framework. *Journal of Personality and Social Psychology, 80*(3), 501-519.

Elliot, A.J. (1999). Approach and avoidance motivation and achievement goals. *Educational Psychologist, 34* (3), 169-189.

Elliott, E.S., & Dweck, C.S. (1988). Goals: An approach to motivation and achievement. *Journal of Personality and Social Psychology, 54* (1), 5-12.

Harackiewicz, J.M., Barron, K.E., Pintrich, P.R., Elliot, A.J., & Thrash, T.M. (2002). Revision of achievement goal theory: Necessary and illuminating. *Journal of Educational Psychology, 94* (3), 638-645.

Hatano, G., & Inagaki, K. (1986). Two courses of expertise. In H. Stevenson, H. Azuma, & K. Hakuta (Eds.), *Child development and education in Japan* (pp. 262-272). New York: Freeman.

Higgins, E.T. (2005). Value from regulatory fit. *Current Directions in Psychological Science, 14* (4), 209-213.

Nokes, T. J., & Ohlsson, S. (2005). Comparing multiple paths to mastery: What is learned? *Cognitive Science, 29*, 769-796.

Schwartz, D.L., & Martin, T. (2004). Inventing to prepare for future learning: The hidden efficiency of encouraging original student production in statistics instruction. *Cognition and Instruction, 22*(2), 129-184.

The Effects of Similarity and Individual Differences on Comparison and Transfer

Samuel Day (day9@indiana.edu)
Dept. of Brain & Behavioral Sciences, 1001 E. 10th St.
Bloomington, IN 47405 USA

Robert L. Goldstone (rgoldsto@indiana.edu)
Dept. of Brain & Behavioral Sciences, 1001 E. 10th St.
Bloomington, IN 47405 USA

Thomas Hills (thomas.hills@unibas.ch)
Institut für Psychologie, Missionsstrasse 64A
4055 Basel, Switzerland

Abstract

Prior research has found that while people are generally quite poor at recognizing when a new situation is structurally similar to a known case, comparison of two analogous cases greatly improves the likelihood of achieving such recognition. Our study examines the effects of varying the similarity between these compared cases, both featurally and structurally. We find that between-case similarity has a significant impact on transfer, and that these effects interact with characteristics of the learner.

Introduction

Our minds are filled with valuable knowledge that we are unable to use. This is particularly true of what might be the most valuable knowledge of all: general principles that can be applied across a wide range of situations. Research in analogy has repeatedly found that principles that are learned in one context often fail to be retrieved when an individual is confronted with a deeply related situation that differs in concrete or contextual ways (e.g., Gentner, Rattermann & Forbus, 1983; Gick & Holyoak, 1983; Ross, 1984). For example, in Gick and Holyoak's classic (1980; 1983) analogy studies, individuals attempting to solve an insight problem routinely failed to recognize that the problem was analogous to one they had been taught earlier (unless given an explicit hint), and therefore failed to make use of their relevant knowledge. For both theoretical and practical reasons, researchers are keenly interested in finding ways to overcome this kind of impediment.

One approach that has shown great promise is simply asking learners to compare two different examples of a principle (e.g., Gick & Holyoak, 1983; Loewenstein et al, 2003; Gentner et al, 2003; Rittle-Johnson & Star, 2007). For example, Loewenstien and colleagues (2003) conducted research with MBA students enrolled in a course on negotiation. Some of the students compared two specific cases involving a "contingency contract," a useful but sometimes counterintuitive negotiation technique. Other students received the same two cases, but read and analyzed them separately, without any explicit comparison. The researchers found that students who had compared cases were nearly three times more likely to apply the relevant principle to a new case than those students who had analyzed the cases separately. Consistent with prior findings of poor analogical transfer in general, the students who had read but not compared cases performed no better on the transfer task than those who had received no training.

Results such as these point to the potential power of comparison. The most common explanation for these effects is that structural alignments generated when comparing two concrete examples serve to highlight meaningful structural commonalities between them, while simultaneously taking the focus away from elements that are extraneous or irrelevant (e.g., Markman & Gentner, 2000). This, in theory, allows a more explicit representation of the structure or principle itself, making it easier to recognize when it arises in new situations.

However, a great deal remains unknown about the factors that make comparison successful in transfer. Particularly, there is a surprising lack of research on how the relationship between the *compared* cases (such as their similarity) may influence the representations that are formed during comparison. Given that the similarities and differences between the cases are the basis for the knowledge that comparison is presumed to generate, this would seem to be a critical area for study.

For instance, will transfer to new situations be best when the features of the compared cases are relatively similar to one another, or when their content is more dissimilar? There are empirical reasons to predict either of these outcomes. Evidence for "conservative generalization" (Medin & Ross, 1989) suggests that the comparison of two examples that share significant surface commonalities may lead to a representation in which many of these irrelevant features are retained. If so, one of the primary assumed benefits of comparison—a more general representation—may be lost. Comparison of dissimilar cases may therefore lead to representations with broader generalizability. On the other hand, comparisons between overtly similar cases are likely to be less cognitively demanding, and may therefore help to "boot-strap" early learning processes. Consistent with this possibility, Kotovsky and Gentner (1996) found that young children were better able to perform matches on the basis of abstract structural commonalities after performing a similar task involving more perceptually similar stimuli.

A related issue is the effect of the similarity of the structures themselves. Most studies focusing on comparison and transfer have made use of cases with essentially identical relational structures. However, there are reasons to suggest that structural variation may be beneficial as well. For instance, some research has shown that comparing two "near-miss" cases (Winston, 1975), which are identical except for a crucial structural difference, may improve transfer (e.g., Gick & Paterson, 1992). This approach may be particularly effective when an individual needs to

discriminate examples of a specific structure from other non-matching cases, as is generally the case in the real world.

The current study examines the impact of both featural and structural similarity in compared cases. Additionally, unlike previous studies, our design requires participants to discriminate different kinds of structures, which may be a more ecologically valid way of assessing the benefits of comparison. Finally, in contrast to previous research that has concentrated on analogical transfer in college-age students, our study uses 7th and 8th grade students in a science class. Children may be more prone to concrete interpretations of scenarios, and thereby miss connections between deeply related scenarios. Given the importance of students appreciating deep principles (e.g. diffusion, order from randomness, and our current topic of interest – feedback loops), it is particularly important to know how children's understanding of principles is influenced by superficial and deep similarities between scenarios.

Experiment

Participants

90 students from a public middle school participated in this study, as part of their regular class time in a General Science course. The group included both 7th- and 8th-grade students (n = 49 and 41, respectively) from six class periods. Roughly half of the students (n = 47) were part of the school's Accelerated Learning Program (ALPs), which is composed of students passing a science achievement test.

Materials and Design

We led the students' class sessions for two days. The first day involved general instruction on the concept of complex systems, including several real-world examples of such systems. This instruction did not include any specific discussion of feedback systems, our target principle. The experiment itself was conducted on the second day.

The overall design of the experiment was as follows: Brief instruction on feedback systems was followed by a pre-test, in which students classified specific scenarios as examples of positive or negative feedback, and answered inference questions about those cases. Students then interacted with two computer simulations, each of which could vary in terms of its content domain (biology or economics) and the type of feedback system it represented (positive or negative). These variations represented the experimental manipulation in the study. Afterwards, students explicitly compared and contrasted the simulations they had completed. Finally, the classification and inference task was administered a second time, as a post-test.

The initial instruction included brief descriptions of positive and negative feedback systems, and included an example of each. These definitions and examples were available to students throughout the experiment.

Pre-test and post-test The pre-test and post-test materials were designed to assess students' understanding of feedback systems, particularly the ability to discriminate positive and negative feedback systems. These materials included eight brief scenarios (averaging 50 words apiece), each describing a real-world phenomenon. Four of these scenarios represented positive feedback systems, and four represented negative feedback systems. For example, one scenario was the following:

> *The lynx is a natural predator of the hare. When lynx populations are small, hare populations increase rapidly. This makes the lynx population increase, since food is plentiful. However, a large lynx population reduces the number of hares, which ultimately brings the lynx population back down. This cycle repeats every ten years or so.*

After reading each scenario, participants classified it as an example of either a positive or negative feedback system by selecting a response from a 5-point scale: *Definitely negative, Probably negative, Don't know, Probably positive, Definitely positive.* They also answered one multiple-choice inference question about each scenario. For example:

> *As the lynx population decreases, the population of rabbits should: [Increase / Decrease]*

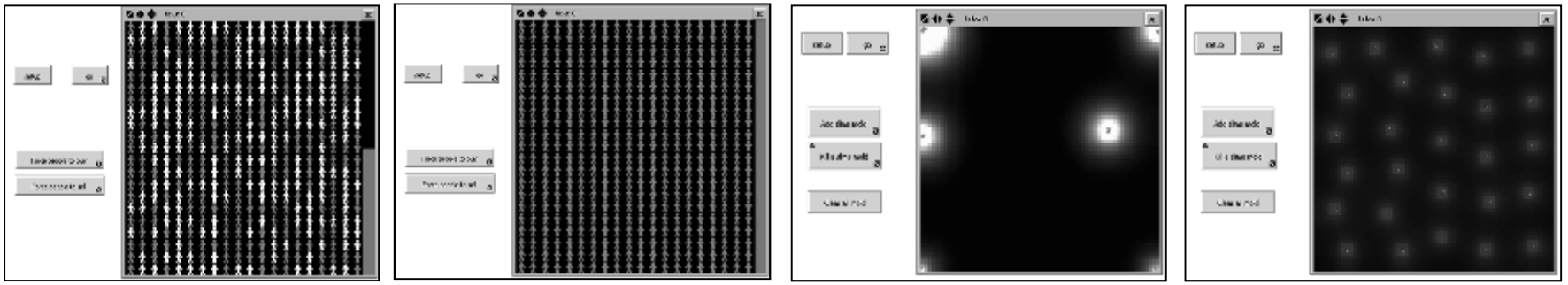
Identical items were given at pre-test and post-test. However, in order to minimize explicit memorization and reference to previous answers, students were not informed about the post-test until later in the experimental session.

Computer simulations. All students interacted with two computer simulations demonstrating feedback behavior. These were implemented in NetLogo (Wilensky, 1999), a software package for developing agent-based simulations. Each of the simulations depicted either a positive or a negative feedback system, and each instantiated one of two domains: biology (specifically, interacting slime mold cells) or economics (a simple stock market). This resulted in four relevant simulation types: Biology Positive, Biology Negative, Economics Positive, and Economics Negative. Two versions of each type were created, differing in cosmetic ways. This allowed some students to interact with two different versions of the same type without repeating an identical simulation. The main theoretical focus of our study was on the effects of the similarity between simulations; that is, whether the domain and/or feedback type were the same or different for each participant.

Each simulation began with a brief description of its behavior. For instance, the Economics Positive simulation presented the following introductory description:

> *"This simulation involves a small economic system. People in this system buy stocks, and they pay attention to what other people are doing. When they see someone else buying a stock, they are more likely to want to buy it themselves. When they see someone else* ***selling*** *a stock, they are more likely to sell it themselves. This creates a POSITIVE FEEDBACK LOOP. People buying the stock leads to even* ***more*** *people buying it. People selling the stock leads to even more people* ***selling*** *it."*

The presentation of the simulation was strictly guided although interactive, instructing students to perform specific actions and then to observe the resulting effects



A. B. C. D.

Box 1: Simulations.

The **economic** (stock market) simulations begin (Frame A) with the 420 agents evenly divided between owning the stock (dark; red in the original simulation) and not owning the stock (white). A bar on the right side of the screen indicates the proportion of the agents currently owning the stock. While the simulation is running, each agent will buy or sell the stock with some specified probability. In the Positive Feedback version of the simulation, the probability of buying rather than selling is a positive function of the overall ownership of the stock. As more agents own the stock, the likelihood of new agents purchasing the stock increases. Conversely, as fewer agents own the stock, the likelihood of other agents selling the stock increases. Because of this, random initial fluctuations in stock ownership tend to be amplified over time, and the system quickly moves toward the extremes, resulting in either ownership by all agents or ownership by no agents (Frame B). In the Negative Feedback version, the probability of an agent buying the stock is a *negative* function of overall ownership. Therefore, increased overall ownership makes agents more likely to sell the stock, while decreased overall ownership makes agents more likely to buy. This tends to create homeostasis in the system. As the ownership of the stock begins to increase or decrease, the market quickly "corrects" itself and maintains an even proportion of owners and non-owners (as in Frame A).

In the course of the simulation, students are instructed to force a proportion of the agents to buy or sell the stock. This is accomplished by selecting the appropriate button on the left side of the screen, then clicking and dragging across the agents. These interactions serve to highlight the way that the system responds to small imbalances, by either amplifying them (positive feedback) or reducing them (negative feedback). Additionally, students are explicitly reminded at one point during the simulation that it is an example of a positive or negative feedback system. For instance, those in the Negative Feedback version were told: "Observe how this system is a negative feedback loop. People buying the stock leads to other people selling it, and people selling the stock leads to other people buying it. This tends to keep the system in balance, without allowing too many people to own or not own the stock at once."

The **biological** (slime mold) simulations begin with 27 agents (mold cells) randomly distributed on the screen. While the simulation is running, each cell moves about the screen probabilistically, and secretes a chemical that remains for a short period of time in its current location. In the Positive Feedback version, cells are attracted this chemical, and their likelihood of moving toward a location increases with the quantity of the chemical there. Over time, this results in the cells grouping into a small number of clusters (Frame C), since more cells in a given location leads to a greater amount of the chemical there, attracting even more individuals. (Chemical density is reflected by the brightness of a location). In the Negative Feedback version, cells tend to be repelled by the chemical, and are therefore more likely to move to locations where less of the substance is present. This results in the cells attempting to maintain a maximal distance from one another, leading to a relatively homogenous distribution across the field (Frame D).

During the simulation, users are instructed to add additional mold cells to the system, by selecting the "Add slime mold" button and clicking in the desired location on the screen. They are asked at various points to observe the relative effects of clustering these new cells close together versus spreading them out in the space. They are also reminded at one point that the simulation is an example of positive or negative feedback, and why. For example, users in the Positive Feedback version were told: "Observe how this system is a positive feedback loop. Cells produce the chemical in a certain location, which brings other cells to that location, which leads to even *more* of the chemical there. This tends to bring the cells together into large clusters.

on the system. For example, students in the Economics simulations were instructed at various times to force a proportion of the agents to buy or sell the stock and observe the results. At one point during each simulation, students were explicitly reminded of which type of feedback system the simulation portrayed (positive or negative), and specifically why this system's behavior reflected that feedback type. After being guided through several relevant actions, students were encouraged to interact freely with the system. Each simulation lasted approximately five minutes. Box 1 provides a detailed description of the simulations.

After completing both simulations, students were instructed: "Now we would like you to compare the two simulations that you just interacted with. Please write about the ways in which the two simulations were similar and different from each other, especially in terms of the way that they behaved." There was no time restriction on the comparison phase. After comparison, all students completed the classification and inference task again.

Predictions. The primary variable of interest is the change in performance between pre-test and post-test. There are several potential predictions about how this variable might be affected by the comparisons that students make. First, prior work on the effects of comparing analogous cases (e.g., Loewenstein et al, 2003) leads us to expect an overall improvement in classification and inference performance, reflecting generally stronger representations of the principles underlying feedback systems. Given that all students are explicitly comparing cases that share a feedback structure, it seems likely that their understanding of such structures should improve on average.

We also predict that the *kinds* of comparisons made may affect performance. Comparing two systems involving the same type of feedback (i.e., both positive or both negative) could lead to a bias in the interpretation of new cases. For instance, a student comparing two simulations involving negative feedback may be more likely to classify new cases as examples of negative feedback at post-test.

Another way in which the kind of comparison may matter is in whether it provides an appropriate balance between the *compatibility* (ease of alignment) and the *generalizability* of the two simulations. As discussed, the similarity of the compared cases may have two opposing influences on transfer. Cases that are more similar to one another may be easier to align, and may therefore provide a more straightforward basis for learning about their shared underlying structure. On the other hand, highly similar cases may artificially restrict students' representations of the relevant principles, leading them to only recognize the structure in new situations that are concretely similar to the learned cases. Less similar comparison cases may therefore lead to better generalization of the principles. We predict that learning will be optimal when dissimilarity on one dimension is "scaffolded" by relatively high similarity on another dimension. In the current context, we would predict relatively good performance from those comparing different feedback types in the same domain (e.g., Biology Positive and Biology Negative). In this case, the relevant differences in the positive and negative systems should be particularly highlighted because the concrete features of the simulations are otherwise highly similar. Likewise, strong performance is predicted for individuals comparing the same feedback type across different domains (e.g., Biology Positive and Economics Positive), since the same underlying principles can be observed across more diverse contexts, presumably supporting broader generalization.

We are also interested in potential effects of individual differences between students, and how these may interact with comparison. For instance, it is possible that students in accelerated classes will tend to focus more on the underlying principles of the simulations, and will therefore be less influenced by perceptual variation between them.

Results

Our data yielded several informative findings. Surprisingly, however, most of our initial predictions were not borne out. We first examined the overall improvement of the students between pre-test and post-test. Calculating improvement simply as post-test performance minus pre-test performance, there was no evidence of any improvement on average, either in classification (M = .03, $t(89)$ = 0.52, *n.s.*) or inference (M = .01, $t(89)$ = 0.78, *n.s.*).

Next, we examined possible bias effects in classifications. Specifically, we predicted that individuals who had compared two cases representing the same kind of feedback system (i.e., either two positive cases or two negative cases) would become more disposed to classify new cases as instances of that particular type. For each of these students (n = 43), we calculated bias as the shift toward whichever end of the classification scale matched the type of feedback cases that the student had compared. This measurement did not differ from zero (M = .01, $t(42)$ = 0.23, *n.s.*).

There was also no evidence for the kind of interaction between structural and featural similarity that we had predicted (analysis below). Neither of the conditions that included one similar dimension and one dissimilar dimension showed any improvement (see Figure 1). However, our analysis did reveal several significant results.

We conducted a 2 (Feedback similarity: Same v. Different) × 2 (Domain similarity: Same v. Different) × 2 (ALPs: Accelerated v. Regular classes) ANOVA on the improvement scores. The omnibus test indicated reliable differences between groups for the classification task, $F(7, 82) = 2.27$, $p < .05$. (No effects were found for the inference task on this or any other analysis discussed). Specifically, the test revealed main effects for both Feedback similarity ($F(1, 82) = 4.02$, $p < .05$) and Domain similarity ($F(1, 82) = 6.18$, $p < .05$). In both cases, improvement was greatest when dissimilar cases were compared. Interestingly, for both dimensions of similarity, performance actually decreased numerically at post-test when similar cases were compared (Feedback: similar = -.07, dissimilar = .13; Domain: similar = -.08, dissimilar = .16). This fact explains the absence of the predicted improvement in overall performance: increased scores associated with comparing dissimilar cases were largely offset by *decreased* scores resulting from the comparison of similar cases. As seen in Figure 1, the greatest improvement was seen in students who compared cases involving both different feedback

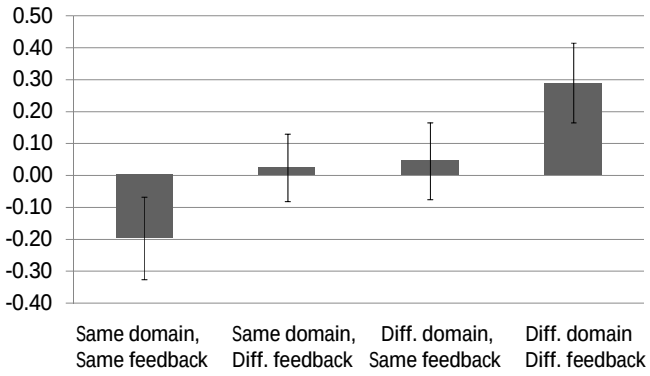


Figure 1: Post-test improvement, by condition

types *and* different domains, while the least improvement (actually negative) was seen in those whose comparisons involved the *same* domain and feedback type. Improvement by those in the Different-Different condition was reliably greater than zero (M = .26, $t(19)$ = 2.22, $p < .05$). Those in the Same-Same condition were marginally less than zero (M = -.21, $t(21)$ = 1.79, $p < .10$). No effect of membership in the accelerated class was observed ($F(1, 89)$ = 0.33, *n.s.*).

The influences of structural and featural similarity therefore appear to reflect independent main effects. However, these two effects did not apply equally across all individuals. Interestingly, students in non-accelerated classrooms showed large effects of Domain similarity ($t(42)$ = 2.83, $p < .01$), but no evidence of any influence from the similarity of the feedback types that were compared ($t(42)$ = 0.04, *n.s.*; see Figure 2). In contrast, the ALPs students were influenced by Feedback similarity ($t(46)$ = 2.38, $p < .05$) but not Domain similarity ($t(46)$ = 0.38, *n.s*).

Discussion

Several conclusions are suggested by these data. First, the results are consistent with previous characterizations of explicit comparison as a powerful cognitive process that may have an important impact on the acquisition of generalizable principles. Under the right conditions, participants in our study improved reliably in their ability to classify new cases, even in very dissimilar domains. However, our data also suggest that the situation is more complex that is generally proposed, and that comparison is not uniformly beneficial. In fact, on average, explicit comparison by the students was not associated with any improvement at all at post-test. Under some circumstances, there were even trends suggesting that students might be negatively impacted by the comparison process (although these effects were not reliable, they were large enough to effectively offset any overall benefits of comparison). These results highlight the importance of exploring the comparison process more deeply, and attempting to establish the factors that influence comparison-based learning. The remainder of our findings begin to address these issues, exploring aspects of both the compared materials and the learners themselves.

Our study varied both the structural similarity (whether the systems involved the same or different feedback types) and the surface similarity (same versus different content domain) of the compared simulations. We predicted that learning would be optimal when dissimilarity along one dimension was "balanced" by higher similarity on another dimension, which we believed would facilitate alignment while still highlighting important structural features. This prediction was based in part on the approach that has generally been taken in the literature: either presenting the same underlying structure in dissimilar contexts (e.g., Loewenstein et al, 2003), or using "near-miss" cases involving the same content but slightly varying the relevant structure (e.g., Gick & Paterson, 1992). In contrast to our expectations, however, we found that post-test improvement was greatest when the cases were less similar to one another on both dimensions of similarity.

Of course, it is important not to over-interpret the results from one task and set of materials. Each dimension was only tested at two levels, one of which was very high similarity. It is possible (even likely) that these effects do not reflect a simple linear relationship between dissimilarity and transfer, but that there is in fact some optimal similarity level beyond which learning and transfer will decline. Regardless, our results do clearly indicate that the similarity of the compared cases—and not simply the similarity between the learning and transfer cases—is a critical factor influencing whether or not relevant knowledge will be successfully learned and applied. Furthermore, our results highlight the importance of using materials that will maximize the generalizability of the learned representations, and suggest that this factor may often be more important than attempting to facilitate alignment through high similarity.

Perhaps the most interesting—and challenging—finding from our study is the way in which properties of the comparison cases appear to interact with individual differences between learners. Transfer by the students in accelerated classes was influenced by the structural similarity between the cases, but not at all by the similarity of the domains involved. In contrast, structural similarity had no impact on students in regular classes, but learning in these individuals was significantly affected by domain similarity. This finding raises important issues about the effects of comparing cases.

The benefits of comparison are generally attributed to its ability to focus attention on relevant aspects of cases while

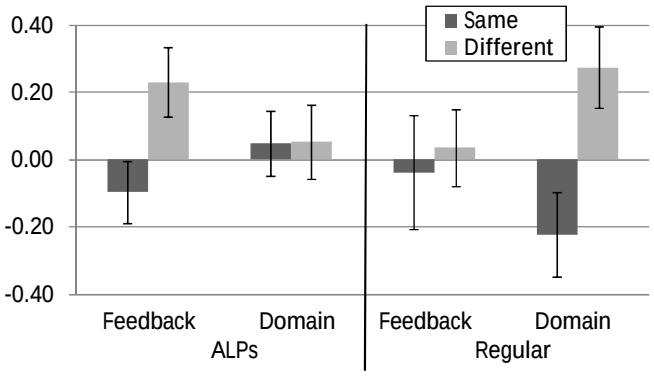


Figure 2: Post-test improvement for accelerated and regular classes.

backgrounding less relevant features. This is, in fact, a mechanism that likely frequently occurs. However, it is important to be mindful of the ways in which differences in individuals' representations of the cases will influence which aspects of the situations are highlighted, and to recognize that these do not always correspond with those that the experimenter may consider "relevant." While membership in the accelerated classes is certainly based on a number of interrelated factors—motivation, achievement, intelligence, ability to focus—it is clear that some difference between the groups is causing them to attend to different aspects of the simulations. These differences appear to have a stark impact on the effects of comparison.

Although more work will be necessary to establish the exact basis of these differences, it seems likely that the ALPs students are better able to look past the immediate surface features of a simulation, and to focus instead on its underlying structural relationships. There are many reasons that this might be the case. For instance, these individuals might be coming to the task with richer background knowledge about the systems that are being presented, and therefore have more cognitive resources available for learning. Consistent with this explanation, students in the accelerated classes had reliably greater performance at pre-test, prior to the primary instruction ($t(89) = 4.60$, $p < .001$). It is also possible that these students have adopted different learning strategies, and are more likely to view all instructional cases as examples of some relevant principle rather than simple facts to be learned independently. Bassok and Holyoak (1989, Experiment 3) found that individuals appeared to acquire the exact same material more concretely or more abstractly based on the specificity of the context in which it was presented. It is possible that successful students have learned to take advantage of this cognitive flexibility by deliberately treating new materials as instantiations of deeper principles, rather than ends in themselves. Previous research has found that experts tend to weigh structural similarities more than superficial similarities (Novick, 1988). The current results extend this finding; even non-experts that are generally high achieving in science show similar tendencies. As such, there appear to be domain-general individual differences in sensitivity to structure that go beyond expertise in a particular domain.

Future research will provide more insight into the exact processes underlying these differences, but our results make clear that characteristics of the learner must be considered when using comparison as an instructional tool. As our data show, cases that lead to reliable gains in one population may foster no improvement at all in another (even very similar) group.

Conclusions

Our knowledge is only valuable to the extent that we are able to make use of it. In previous research, the simple act of comparing two analogous situations has been shown to be extremely valuable in this regard, freeing up concepts that were otherwise bound to a specific context and allowing them to be employed in a much wider range of situations.

The current research shows, however, that these processes may interact in complex and unexpected ways with the features of the cases that are compared and with individual differences in the learner. Our results begin to establish some of the factors that influence the efficacy of comparison, and point the way to future research that may further help us take advantage of this powerful cognitive tool.

Acknowledgments

This work was supported by National Science Foundation REESE grant 0910218. We would like to thank Nancy Martin of Jackson Creek Middle School and Lisa Byrge for their help with our research.

References

Bassok, M., & Holyoak, K. J. (1989). Interdomain transfer between isomorphic topics in algebra and physics. *Journal of Experimental Psychology; Learning, Memory, and Cognition, 15*, 153-16.

Gentner, D., Loewenstein, J., Thompson, L. (2003). Learning and transfer: A general role for analogical encoding. *Journal of Educational Psychology, 95,* 393–408.

Gentner, D., Rattermann, M. J., & Forbus, K. D. (1993). The roles of similarity in transfer: Separating retrievability from inferential soundness. *Cognitive Psychology, 25,* 524–575.

Gick, M. L., & Paterson, K. J. (1992). Do contrasting examples facilitate schema acquisition and analogical transfer? *Canadian Journal of Psychology, 46,* 539–550.

Kotovsky, L., & Gentner, D. (1996). Comparison and categorization in the development of relational similarity. *Child Development, 67,* 2797–2822.

Loewenstein, J., Thompson, L. & and Gentner, D. (2003). Analogical learning in negotiation teams: Comparing cases promotes learning and transfer. *Academy of Management Learning and Education, 2*, 119–127.

Markman, A. B., & Gentner, D. (2000). Structure-mapping in the comparison process. *American Journal of Psychology, 113*, 501–538.

Medin, D. L., & Ross, B. H. (1989). The specific character of abstract thought: Categorization, problem-solving, and induction. In R. J.Sternberg (Ed.), *Advances in the psychology of human intelligence* (*Vol. 5*, pp. 189–223). Hillsdale, NJ: Erlbaum.

Novick, L. R. (1988a). Analogical transfer, problem similarity, and expertise. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*, 510–520.

Rittle-Johnson, B., & Star, J. R. (2007). Does comparing solution methods facilitate conceptual and procedural knowledge? A study on learning to solve equations. *Journal of Educational Psychology, 99*, 561–574.

Ross, B.H. (1984). Remindings and their effects in learning a cognitive skill. *Cognitive Psychology, 16*, 371–416.

Winston, P.H. (1975). Learning structural descriptions from examples. In P.H. Winston (Ed.), *The psychology of computer vision*. New York: McGraw-Hill, pp. 157–210.

Wilensky, U. (1999). *NetLogo (and NetLogo User Manual).* http://ccl.northwestern.edu/netlogo/

Seeing Language Learning inside the Math: Cognitive Analysis Yields Transfer

Kenneth R. Koedinger (koedinger@cmu.edu)
Human-Computer Interaction Institute, Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA 15217 USA

Elizabeth A. McLaughlin (mimim@cs.cmu.edu)
Human-Computer Interaction Institute, Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA 15217 USA

Abstract

Achieving and understanding effective transfer of learning requires a careful analysis of the hidden knowledge and skills to be transferred. We present an experiment that tests a subtle prediction of such an analysis. It concluded that a critical difficulty in students' learning to translate algebra story problems into symbolic expressions is in learning the grammar of such expressions. We hypothesized that exercises requiring students to substitute one algebraic expression into another would enhance students' algebraic grammar knowledge. This hypothesis led to a counter-intuitive prediction that learning to symbolize story problems could be better enhanced through practice on dissimilar looking substitution exercises than through practice on more similar looking story problems. We report on an experimental comparison involving 303 middle school students that supports this prediction. We discuss how having learners externalize a uniform abstract form and get interactive feedback on it may be important factors in enhancing transfer.

Keywords: cognitive task analysis; transfer; grammar learning; mathematics education.

Introduction

Humans learn language before they have a language to use to learn. Might the learning processes that make this amazing feat possible, like the capability to learn grammatical structures through experience without explicit instruction, be useful for other kinds of learning tasks? Once children have acquired language, are the cognitive functions employed in language learning no longer useful? For instance, as students take courses in complex academic topics, like algebra, does all that brain matter for language learning have nothing to do? Or is it possible that some of the same implicit learning mechanisms employed in language learning are useful for learning math and science?

This paper does not aim to provide conclusive answers to these questions, however, it does provide a compelling demonstration that grammar learning processes may be important in learning mathematics. Students may engage in such learning without explicit awareness and such implicit learning may be more prevalent in academic learning than is generally recognized (e.g., Alibali & Goldin-Meadow, 1993; Landay & Goldstone, 2007). In earlier work, we performed a cognitive task analysis of the important task domain of "symbolization", that is, the ability to model problem situations or "story problems" in algebraic symbols (Heffernan & Koedinger, 1997; 1998). Table 1 shows examples of symbolization problems, which ask students to translate a story problem into an algebraic expression. The obvious potential connection between language learning processes and this task is in learning to read and comprehend story problems. While such learning is indeed a significant challenge for elementary students (Cummins, Kintsch, Reusser, & Weimer, 1988), our past data provided evidence that comprehending story problems is no longer a major sticking point for most beginning algebra students.

This claim can be illustrated by an analogy to foreign language translation: Translating a story problem to algebra is like translating English to Greek. For an English speaker, the difficulty in translating to Greek is not comprehending the English, but generating the Greek. Similarly, the challenge for older students in a beginning algebra course is much less in understanding the English in which the story problems are written and more in being able to express that understanding algebraically, that is, in the language of algebra.

One indication that comprehension of algebra story problems is not a major sticking point for beginning algebra students comes from Heffernan and Koedinger's (1998) data showing that students can solve story problems (produce a value for the dependent or "y" variable when a value for the independent or "x" variable is given) much more accurately (63% correct) than they can symbolize (write an equation relating x and y) a story problem (18% correct). Since solving requires comprehension of the story, the performance difference is suggestive that symbolizing is problematic for students in ways beyond the demands of sentence comprehension. A second indication presents a contrast with a difficulty experienced by Artificial Intelligence systems programmed to solve story problems, namely that of understanding the arithmetic relationships between quantities described in the story (Bobrow, 1968). We created problems where natural implicit descriptions of such relationships (e.g., "Ms. Lindquist teaches 62 girls. Ms. Lindquist teaches b boys.") are supplemented (Heffernan & Koedinger, 1997) or replaced (Koedinger, Alibali, & Nathan, 2008) with explicit descriptions (e.g., "The number of students Ms. Lindquist teaches is equal to the number of boys plus the number of girls."), which are much easier for a program to process. We found, however, that providing such explicit descriptions does not

Table 1. Eight two-step symbolization items in order from easiest to hardest.

name	Item	Answer
cds	Mary opened a new music store. She got CDs delivered on her first day. She got 5 truck loads of CDs delivered. Each truck that arrived dropped off 12 boxes. Each box she received had ***c CDs***. Write an expression for how many CDs were delivered that first day.	5*12*c
mcdona	Mike starts a job at McDonald's that will pay him 5 dollars an hour. Mike gets dropped off by his parents at the start of his shift but he takes a taxi home that costs him 7 dollars. Mike works an ***h hour*** shift. After taking into account his taxi ride, write an expression for how much he makes in one night.	5*h-7
children	John and his wife Beth have been saving to give their 5 children presents for the holidays. John has saved 972 dollars for presents and Beth has saved ***b dollars***. They give each child the same amount. Write an expression for how much each child gets.	(972+b)/5
sisters	Sue made 72 dollars by washing cars to buy holiday presents. She decided to spend ***m dollars*** on a present for her mom and then use the remainder to buy presents for each of her 4 sisters. She will spend the same amount on each sister. Write an expression for how much she can spend on each sister.	(72-m)/4
students	Ms. Lindquist is a math teacher. Ms. Lindquist teaches 62 girls. Ms. Lindquist teaches ***f fewer boys*** than girls. Write an expression for how many students Ms. Lindquist teaches.	62+62-f
rowboat	Ann is in a rowboat on a lake. She is 800 yards from the dock. She then rows for ***m minutes*** back towards the dock. Ann rows at a speed of 40 yards per minute. Write an expression for Ann's distance from the dock.	800-40m
trip	Bob drove 550 miles from Boston to Pittsburgh to visit his grandmother. Normally this trip takes him ***h hours***, but on Tuesday there was little traffic and he saved 2 hours. Write an expression for what was his average driving speed.	550/(h-2)
jacket	Mark went to the store to buy jackets that cost ***d dollars***. When he got there the store was having a sale of 1/3 off the usual prices. Write an expression for how much the jacket cost him.	d-1/3*d

significantly improve the performance of beginning algebra students (77% on explicit vs. 79% on implicit in Koedinger, Alibali, & Nathan, 2008 and 53% vs. 50%, respectively, in Heffernan & Koedinger, 1997).

A third indication that problem comprehension is not a major sticking point identifies difficulties on the production side of the translation process (i.e., going from understanding to the target language, Algebra in this case) rather than the comprehension side (i.e., going from the source language, English story problems, to understanding). Heffernan and Koedinger (1997) contrasted the two-step problems shown in Table 1 (e.g., see the *students* problem in the fifth row) with matched one-step counter parts (e.g., see the first two rows in Table 2 for the one-step counterparts of the two-step *students* problem). In each matched set, the two one-step problems are designed to have essentially the same content as the two-step problem. Using the *students* problem as an example, the two-step problem requires the solver to understand that 1) the total number of Ms. Lindquist's students is the sum of the number of girls and number of boys and 2) that the number of boys is difference between the number of girls and the variable f. The one-step problem "a" in Table 2 requires understanding of first of these relationships and the other one-step problem "b" requires understanding the second of these. Heffernan and Koedinger (1997) found that student performance on symbolizing two-operator problems was significantly worse (40% correct) than combined performance on two matched one-operator problems (62% correct). (Note that average performance on a single one-operator problem is even better at 79% correct.)

The comprehension demands of the two one-operator problems are quite similar to that of the two-operator problem as the words and sentences used in each are substantially overlapping if not quite identical. The production demands, however, have an important difference. To correctly produce the algebraic expression for the one-step problems, 62+b and 62-f, learners need only acquire the mental equivalent of the grammar rule "expression => quantity operator quantity". However, this syntactic knowledge is not sufficient to produce two-operator symbolic expressions, like 62+62-f. To do so, requires the acquisition of knowledge equivalent to additional grammar rules that allow for an expression to be embedded inside another expression. More formally, producing two-operator symbolic expressions requires the equivalent of grammar rules like "expression => quantity operator expression" and "expression => expression operator quantity". Figure 1 illustrates how the first two of the three grammar rules above can combine to produce two-operator expressions like 62+62-f.

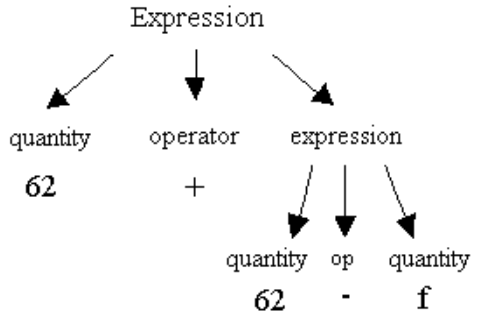


Figure 1: Grammar tree for a two-operator expression.

To be sure, we are not saying that students need to learn such grammar rules explicitly, but simply that they need to implicitly acquire the skills that are consistent with the patterns these rules describe. But the difference between two-step and one-step performance implicates such syntactic skill. In other words, that students are significantly worse at solving a single two-step problem than they are at solving both of the matched one-step problems is evidence that they lack implicit knowledge of grammar for combining expressions. There are alternative hypotheses to be sure (some of which were explored in Heffernan & Koedinger, 1997, 1998), but a strong test is to use this hypothesis to design purportedly better instruction and test whether it is indeed better.

So, for instruction, the ideal would be to find a task that isolates learning of these implicit "hidden" grammar rules. A task that does so is a substitution exercise, as illustrated in the last row of Table 2. This task requires students to produce of two operator expressions (and thus should exercise the hidden grammar rules) but without any of the requirements of comprehending a two-step story problem.

This leads us to a counter-intuitive hypothesis that instruction (substitution) that looks unlike the target objective (two-operator story problem symbolization) is going to lead to learning and transfer and, further, may do so better than instruction (one-operator story symbolization) that looks much more like the target objective. In particular, we hypothesize that practice on substitution exercises will transfer to better performance on translating algebra story problems into symbolic expressions. We will measure improvement by examining the differences on posttest two-step symbolization items between students who do substitution problems embedded within a problem set and students who only practice one-step symbolization problems within the problem set. As a pretest, both the treatment and control conditions begin with a measure of their ability to write one-step expressions before being presented with a two-step problem.

Method

The experiment was implemented inside the ASSISTment system and run in middle school classrooms in an urban school district outside of Boston, MA. The ASSISTment system is a web-based computer tutor authoring and delivery system designed to be used for both formative assessment and instruction (Razzaq et al., 2007). Instruction is provided by feedback on errors, on-demand hints, and scaffolding questions that reduce a problem into its components much like a simplified version of a Socratic dialogue.

Materials and Design

The materials for this study were the eight two-step story problems shown in Table 1 along with matched one-step and substitution items for each as illustrated in Table 2. This produces a pool of 32 items of which students saw 16 in one of two versions. Items were placed into the versions so that students never saw an item that has the same answer (or answer part) as another. The items were organized in three phases: 1) five pre-test items, 2) seven integrated instructed and post-test items, and 3) four filler items. The first two phases are relevant to the study design and are illustrated in Table 3. (The filler items are the one-step or substitution items the other condition received as instruction and were included to collect data on item difficulty.)

Table 2. The matched one-step problems (a & b) and substitution problem (c) for the two-step *student* item.

	Item	Answer
a	Ms. Lindquist is a math teacher. Ms Lindquist teaches 62 girls. Ms Lindquist teaches ***b boys***. Write an expression for how many students Ms. Lindquist teaches.	62+b
b	Ms. Lindquist is a math teacher. Ms Lindquist teaches 62 girls. Ms Lindquist teaches ***f fewer boys*** than girls. Write an expression for how many boys Ms. Lindquist teaches.	62-f
c	Substitute 62-f for b in 62+b Write the resulting expression.	62+62-f

In the pre-test phase, both groups did the same four one-step problems (labeled a or b in Table 3) depending on which version of the problem set received, followed by one two-step problem (labeled 0 in Table 3) depending on which version and order received. We created two "versions" to be evenly matched in difficulty by selecting two-step problems going down this list, cds, sisters, students, and jackets for version A and mcdonalds, children, rowboat, and trip for version B. Version A, then, had one-step and substitution items corresponding with cover stories mcdonalds, children, rowboat, and trip and vice versa for version B. We also created two orders of each version by reversing the sequence of the two-step problems, easy to hard (0, 1,2,3) vs. hard to easy (3,2,1,0). Thus, we expected order to have a significant effect on a pre-post comparison and controlled for it in the analyses below.

In the integrated instruction and post-test phase, students started with two instructional problems (either ab or cc in Table 3) and then alternated between two-step problems (1-3 in Table 3) and further instructional problems. As noted above, the four instructional problems come from the four base cover stories not used for the two-step problems, whether version A or B. The instructional problems corresponded with condition, one-step problems for the one-step condition and substitution problems for the substitution condition. For the one-steps, which come in a-b pairs (as illustrated in Table 2), two of type a and two of type b were selected from the four available cover story sets.

The two-step problems were ordered by difficulty based on a pilot study with students from the same grade and district and this order, from easiest to hardest, is shown in Table 1.

Table 3. Sequence of items for both conditions.

Condition	Pre-test	Instruct & test
One-step	abab0	ab1a2b3
Substitution	abab0	cc1c2c3

a & b = one-step, c = substitution, 0-3 = two-step

Given the nature of the ASSISTment system, all items are both assessment items (based on students' first response) and instructional items (based on feedback, hints, and scaffolding questions that may follow an incorrect response). The only difference between the two conditions is the placement of the substitution versus one-step items during the instruction.

Participants

The original data included 318 middle school students (N=158 one-step practice, N=160 substitution practice) using an on-line system during the 08-09 school year. The final data set included only those subjects who completed all 16 of the items in the problem set (four two-step, eight one-step, and four substitution) for a total of 303 participants (N=154 one-step, N=149 substitution).

Measures

The pre-test was designed to assess students' prior knowledge of translating story problem to algebraic expressions. It was the first five items in the item sequence and consisted of four one-step items and the first one two-step item. A pre-test measure was computed as the average of the two-step score and the average of the four one-step scores, thus appropriately giving more weight to the two-step item that is the goal of instruction. The posttest score was computed as the average of the scores on the last three two-step items. All pre and post-test scores were based on students' first attempt at an item such that either an incorrect entry or a hint request counted as an error.

Results

To test the main hypothesis that instruction on substitution tasks leads to better transfer of learning to two-step symbolization problems than does instruction on one-step symbolization problems, we performed an ANCOVA with pre-test as a covariate, condition and item order (easy-to-hard vs. hard-to-easy) as factors, and post-test as the dependent variable. As noted above, we included the order factor because of its obvious likely influence. We found significant effects of both factors, condition ($F(1,299) = 4.45$, $p < .05$) and order ($F(1,299) = 39.57$, $p < .001$), and of the pre-test covariate ($F(1,299) = 78.62$, $p < .001$). We found no other significant effects when we explored more complex models involving problem set version and the possible two- and three-way interactions with condition, version, and order.

Not surprisingly, high pre-tests are associated with higher post-tests and the easy-to-hard order yields lower post-test scores. With regards to condition, students in the substitution condition had similar pretest scores (M=.56) as students in the one-step condition (M=.57); however, the substitution group posttest scores (M=.39, SD=.35) were higher than the one-step group scores (M=.33, SD=.33). We used the ANCOVA results to compute adjusted posttest scores (M=.39 for substitution, M=.32 for one-step) and an effect size (Cohen's d = .29).

How Does Substitution Practice Help

To better understand how substitution practice may enhance learning of algebra symbolization skills, we investigated the errors students made on the posttest items. A common error category on two-step symbolization problems is to provide a 1-operator answer, for instance, "62-f" rather than "62+62-f". This error is consistent with a student whose only algebra grammar knowledge is "expression => quantity operator quantity". We hypothesized that substitution practice should aid the acquisition of grammar rules that allow for embedded expressions, like "expression => quantity operator expression". The addition of such knowledge should reduce the 1-operator responses to two-step problems.

We coded incorrect solutions in four error categories: 1-operator, 2-operator, missing parentheses, or hint/other. The most common error for both conditions is a 1-operator error. We found that the one-step group produces the 1-operator error (34%) somewhat more often than the substitution group (30%). This difference is larger for some problems and, in particular, appears to account for improved performance on four of the problems (cds, students, rowboat and trip) on which the one-step group is 9% worse than the substitution group (23% vs. 32%) and makes 12% (47% vs. 35%) more 1-operator errors. We found no consistent differences between conditions for 2-operator or hint/other errors. Three post-test problems require parentheses (sisters, children and trip) and on these, missing parentheses errors account for condition differences. The one-step group is 8% (34% vs. 42%) worse on these problems than the substitution group and makes 12% (25% vs. 13%) more missing parentheses errors.

We did not discuss parentheses in our brief characterization of the algebra grammar above, but the correct use of parentheses is clearly an important part of algebra expression structure. Consistent with the hypothesis that substitution practice should enhance algebra grammar learning, we indeed found a reduction in missing parenthesis errors in the substitution group relative to the one-step group.

One way grammar learning can be achieved is through the kind of implicit or non-verbal statistical learning mechanisms that are presumably used in first language acquisition. If these mechanisms are in part responsible for algebra grammar learning (see Li, Koedinger & Cohen, 2010 for a demonstration of the feasibility of such), then we might expect to see more frequent use of grammatical forms seen by those students who have seen such forms more

frequently. Indeed, the one-step group sees 1-operator expressions more frequently and generates such expressions more frequently on post-test problems than the substitution group. In contrast, the substitution group sees more expressions with parentheses and generates such expressions more frequently on post-test problems than the one-step group.

In fact, these patterns appear not only in student errors, as discussed above, but also in their correct responses. On some two-step posttest problems (cds, students, and jackets) it is possible to produce a correct 1-operator solution (e.g., "60c" for 5*12c","124-f" for "62+62-f", 2/3 *d for d-1/3*d). The one-step group, despite doing generally worse on these problems (23% vs. 31%), actually produces twice as many correct 1-operator solutions as the substitution group (7.2% vs. 3.5%). It is also possible for students to produce correct answers that include parentheses on problems that do not require them (e.g., "62+(62-f)"). Again, consistent with the hypothesis that statistical properties of learning, like frequency, are operative even in formal domains like algebra, we find that the substitution group has more correct solutions that involve unnecessary parentheses than the one-step group (15% vs. 9.3%).

An astute reader may wonder about the following alternative interpretation of the observed overall differences in learning. Might the one-step group's experience generating 1-operator solutions simply be interfering with production of 2-operator solutions needed for correct performance on the two-step post-test problems? Or, to put it in more stark terms, might students in the substitution group simply be learning a shallow bias to generate 2-operator solutions and the one-step group students simply learning a shallow bias to generate 1-operator solutions? It is first worth emphasizing that, because of the instructional scaffolding for all on the two-step problems, neither group was exclusively seeing one response type or the other.

Certainly though, part of our hypothesis is that a shift in bias is causing improvement, but that that shift is in probabilities on implicit grammatical structure knowledge not in shallow or surface features. To be better, the substitution group students must not only avoid generating 1-operator solutions (note that they are not so easily biased that they stop making 1-operator errors), but also learn how to generate correct 2-operator solutions, including appropriate use of parentheses. If substitution group students were simply shallowly biased toward 2-operator solutions, we would expect them to perform worse on the four one-step problems they were given in the filler phase than the one-step group did on the same problems during instruction. In fact, both groups were 72% correct on one-step problems. Thus, the substitution group was not blindly over-generating 2-operator solutions.

Discussion and Conclusion

When we think about learning and transfer, it is tempting to think just in terms of the observable tasks between which transfer may occur. However, the vehicle of transfer is the knowledge the learner acquires from a source task and transfer occurs to the extent that that knowledge is relevant and employed in the target task (cf. Singley & Anderson, 1989). Careful cognitive task analysis regarding the underlying nature of the knowledge demands of tasks can thus provide insight into how best to achieve transfer. We presented an experiment that tested a subtle prediction of a prior data-driven cognitive task analysis. That analysis suggested that comprehending story problems tends not to be a major source of difficulty for students learning to translate story problems to algebra. Instead, learning to produce longer symbolic expressions is a more significant challenge and that students must acquire more sophisticated algebra grammar knowledge to meet this challenge. We hypothesized that practice on substitution tasks would assist students in extending their algebra grammar and, counter-intuitively, that such practice would yield better transfer to story problem symbolization than practice on simple story symbolization would. A classroom-based study with some 300 middle school students provided support for this hypothesis.

It may seem surprising that we found transfer from instruction on symbolization tasks, which have little natural language content, to story problem tasks and, even more, that such transfer is greater than from instruction on story problem tasks themselves (albeit simpler ones). After all, the literature and theory on analogical transfer (e.g., Gentner, 1983; Gick & Holyoak, 1983) suggests that people are particularly sensitive to surface features and have great trouble transferring experience from one situation (e.g., converging radiation treatment) to another with dissimilar surface features (e.g., converging military forces). How, then, does the instruction used in this study apparently help students acquire a relevant deep structure and transfer it from substitution tasks to surface-dissimilar story problem symbolization tasks?

An important observation here is that while these task categories (substitution and two-step story) do not have common surface features in their stimulus structure, they are similar in their response structure. The answer in both cases is a two-operator algebraic expression. To be sure, the correct responses to the instructional problems (analogical sources) and post-test problems (analogical targets) are not identical, nor even similar in surface characteristics -- for instance, "800-40x" has little or no surface similarity with "62+62-f". However, the structure of these responses, whether generated from a story problem or a substitution problem, is similar in underlying grammatical form ("expression => quantity operator expression").

Similarity in response structure is not enough to produce transfer. The well-known convergence tasks of Gick and Holyoak (1983) have an arguably similar response structure, yet learners show little transfer between such tasks under most instructional variations. What may be critical is that the learner externalizes the response, gets feedback and support to get the response right, and the external form is abstract and uniform (e.g., if a common converging arrow

diagram was used in response to convergence tasks). In this study, the demands of both substitution and symbolization problems put the solution response into the world where it can be "re-perceived" (c.f., Goldstone, Landy, & Son, in press). Further, the kind of interactive instruction we employed (use of the ASSISTment tutor) guarantees that students get the response right before moving on. By generating, or at least perceiving a correct response, students may (implicitly) engage the same perceptually-grounded, similarity-based generalization processes on the response that they use on the task stimulus. Thus, they may develop better mental representations, whether grammar rules or "perceptual chunks" (Chase & Simon, 1973), of those response representations. Further, it may be important to the transfer result that the response representation is a uniform abstraction (i.e., algebraic expressions). This concise, unadorned representation may make it easier for pattern recognition mechanisms to learn the deep patterns (i.e., the algebraic grammar rules) needed for transfer (c.f., Kaminski, Sloutsky, & Heckler, 2008).

More practically, this research illustrates how a general instructional principle like starting simple (or mastery-based learning, Bloom, 1984) may not be effective if it is not accompanied with a careful cognitive task analysis of the target subject matter domain. Instruction that helps students master parts before helping them master the whole may seem obvious, however, what seem like "parts" on the surface may not be the right "cognitive parts" a learner needs to acquire. It is not particularly hard to identify the part-whole relationship between one-step and two-step story problems. Thus, the control condition in this study is not a straw man, but a reasonable application of part-task training principles and is representative of sequencing in math textbooks.

It is not *a priori* obvious, however, that substitution tasks are a "part" of two-step story problem symbolization. We came to that conclusion after a data-driven cognitive task analysis (cf., Clark, Feldon, van Merriënboer, Yates, & Early, 2007) that involved the analytic use of computational modeling (e.g., the grammar rule analysis). We believe that there is great promise for greatly improving the efficiency and effectiveness of instruction, even in well-investigated domains like algebra, through a combination of domain-general instructional principles and such detailed cognitive task analysis.

Acknowledgments

This research was funded by grants from the U.S. Department of Education Institute of Education Sciences (Grant #R305K030140 and Grant #R305A07044). We would like to thank the members of the ASSISTment team especially Neil Heffernan and Hui Cheng.

References

Alibali, M. W., & Goldin-Meadow, S. (1993). Transitions in learning: What the hands reveal about a child's state of mind. *Cognitive Psychology, 25*, 468-523.

Bloom, B.S. (1984). The 2-sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher, 13*, 4-16.

Bobrow, D. G. (1968). Natural language input for a computer problem-solving elementary word problems. *Cognition and Instruction, 1*, 245-296.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology, 4*, 55-81.

Clark, R. E., Feldon, D., van Merriënboer, J., Yates, K., & Early, S. (2007). Cognitive task analysis. In J. M. Spector, M. D. Merrill, J. J. G. van Merriënboer, & M. P. Driscoll (Eds.), *Handbook of research on educational communications and technology* (3rd ed., pp. 577–593). Mahwah, NJ: Lawrence Erlbaum Associates.

Cummins, D.D., Kintsch, W., Reusser, K., & Weimer, R. (1988) The role of understanding in solving word problems. *Cognitive Psychology, 20*, 405-438.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy, *Cognitive Science, 7*, 155- 170.

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. Cognitive Psychology, 15, 1-38.

Goldstone, R. L., Landy, D. H., & Son, J. Y. (in press). The education of perception. *Topics in Cognitive Science.*

Heffernan, N. & Koedinger, K.R. (1997). The composition effect in symbolizing: The role of symbol production vs. text comprehension. In *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society,* (pp. 307-312). Hillsdale, NJ: Erlbaum.

Heffernan, N. & Koedinger, K. R. (1998). A developmental model for algebra symbolization: The results of a difficulty factors assessment. In *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society,* (pp. 484-489). Hillsdale, NJ: Erlbaum.

Kaminski, J. A., Sloutsky, V. M., & Heckler, A. F. (2008). The advantage of learning abstract examples in learning math. *Science, 320*, 454-455.

Koedinger, K. R., Alibali, M. W., & Nathan, M. M. (2008). Trade-offs between grounded and abstract representations: Evidence from algebra problem solving. *Cognitive Science, 32*(2): 366-397.

Landy, D., & Goldstone, R. L. (2007). Formal notations are diagrams: Evidence from a production task. *Memory & Cognition, 35*, 2033-2040.

Li, N., Cohen, W. W., & Koedinger, K. R. (2010). A computational model of accelerated future learning through feature recognition. To appear in *Proceedings of the 10th International Conference of Intelligent Tutoring Systems.*

Razzaq, Heffernan, Koedinger, Feng, Nuzzo-Jones, Junker, Macasek, Rasmussen, Turner & Walonoski (2007). A Web-based Authoring Tool for Intelligent Tutors: Assessment and Instructional Assistance. In Nadia Nedjah, et al. (Eds.) *Intelligent Educational Machines.* Intelligent Systems Engineering Book Series. Springer.

Singley, K., & Anderson, J. R. (1989). *The transfer of cognitive skill.* Cambridge, MA: Harvard University Press.

Initial Evidence of the Effects of Linguistic Framing on Transfer

Randi A. Engle (RAEngle@Berkeley.Edu)
Adam Mendelson (AMendelson@Berkeley.Edu)
Phi D. Nguyen (PDNguyen@Berkeley.Edu)
Graduate School of Education, University of California, Berkeley
4646 Tolman Hall, Berkeley, CA 94720 USA

Abstract

This paper investigates the idea that it is not just the content of what students learn that influences transfer, but also how learning and transfer contexts are linguistically framed. In a one-on-one tutoring experiment we manipulated framing while controlling for several known transfer mechanisms. We contrasted an expansive framing in which students are positioned as contributing to larger conversations that extend across time, places, people, and topics, with its opposite. We then measured the degree to which high school biology students transferred knowledge from a learning session about the cardiovascular system to a transfer-of-learning session about the respiratory system. We found that students in the expansive condition were more likely to transfer: (a) facts, (b) a conceptual principle, and (c) a learning strategy from one system to another.

Keywords: Transfer-of-learning; Linguistic framing; Social interactions and learning; Human tutoring; Self-explaining

Introduction

Transfer-of-learning, or the application of something learned in one context to another context, is one of the most important but difficult issues in cognitive science and education (e.g. Gick & Holyoak, 1983; Lave, 1988; Lobato, 2006). As Barnett and Ceci (2002) explain, "there is little agreement in the scholarly community about the nature of transfer, the extent to which it occurs, and the nature of its underlying mechanisms." This paper focuses on an instructional mechanism that has rarely been investigated systematically: the linguistic *framing* of learning contexts (Engle, 2006). In this paper, we report the first experimental study of this mechanism in an educational context: a tutoring experiment testing whether framing affects transfer.

Framing Contexts as a Mechanism for Transfer

Most research on transfer mechanisms does not focus on *contexts* or their framing, but on the nature of the *content* students transfer. For example, the importance of comparing multiple examples to form generalizations is often emphasized (e.g. Gick & Holyoak, 1983; Gentner, Lowenstein & Thompson, 2003). When context is addressed, the focus is on similarities between objective features of learning and transfer contexts, like their physical locations and who is present (e.g. Catrambone & Holyoak, 1989; Spencer & Weisberg, 1986).

Our approach to the relationship between context and transfer investigates the idea that otherwise objectively similar contexts can be linguistically framed as different social realities (e.g, Duranti & Goodwin, 1992; van Dijk, 2008) that may encourage or discourage transfer (Engle, 2006; Laboratory for Comparative Human Cognition [LCHC], 1983; Greeno, Smith & Moore, 1993; Hammer et al., 2005). As Pea (1987, p. 647) explained, "contexts [that matter for transfer] are *not* defined in terms of physical features of settings, but in terms of the *meanings* of these settings constructed by the people present."

We use the term *framing* to refer to the linguistic processes of establishing these social realities (e.g., Tannen 1993). For explaining transfer, the framing of boundaries of learning and transfer contexts is particularly important as it affects which contexts students view as being relevant sites for using what they have learned. For example, when a teacher introduces a lesson as providing students entry into knowledgeable roles within communities they plan to participate in throughout their lives, the social boundary of the lesson expands to encompass additional contexts for which each student's understanding of the lesson will be relevant. In contrast, the teacher could have introduced the same lesson as only relevant to the next day's quiz, thus framing it as divorced from other contexts-of-use.

Here we investigate the hypothesis that transfer is more likely when learning and transfer contexts are framed *expansively* as opportunities for students to actively contribute to larger conversations that extend across times, places, people, and activities (Engle, 2006). The boundaries of expansive contexts are framed as wide-ranging and permeable to increase the contexts that can become linked with them (Floriani, 1994; Gee & Green, 1998). Additionally, learners become positioned as authors who share their knowledge more generally. Thus, learners learn under the assumption that they will be expected to transfer what they learn to other contexts (LCHC, 1983; Pea, 1987). In potential transfer contexts they act under the assumption that they are accountable for using what they know from other times, places, and people (Greeno et al., 1993; Pea, 1987).

Existing Evidence About Framing and Transfer

Few studies have empirically investigated potential connections between framing and transfer. Hammer et al. (2005) showed that when two transfer contexts were re-framed as being about active sense-making rather than the replication of knowledge, students transferred-in their prior knowledge in ways that helped them understand physics concepts. Engle (2006) showed how a classroom case of

successful transfer that occurred despite weak content-based supports could be explained by a teacher's expansive framing of time, participants, and roles. Finally, Hart and Albarracin (2009) found that people were more likely to repeat an action they had just engaged in—the most basic form of transfer that there is—if they were prompted to describe it using a progressive verb tense that frames it as a continuing activity ("I was doing…") versus a perfective tense that frames it as a completed action (e.g., "I did…").

A Tutoring Experiment to Investigate the Effects of Framing on Transfer

We conducted a tutoring experiment using a 2x2 design with framing condition (expansive vs. its opposite, bounded) as a randomized variable and student population (first year General Biology vs. Advanced Placement [AP] Biology) as a fixed variable included to assess the generality of effects across populations. To reduce pre-intervention differences between conditions, matched pairs of students from the same classes who performed similarly on a screening test were randomly assigned to each framing condition (Shadish, Cook & Campbell 2002). Each student participated individually in a 3-4 hour learning session about the cardiovascular system on one day followed by a 1-2 hour transfer-of-learning session about the respiratory system the next day. Each session's order was: instructions, pre-test, tutoring, survey, and post-test. In all conditions we aimed to strongly support learning while moderately supporting transfer via known instructional mechanisms.

Participants and Their Originating Biology Classes

24 biology students from the same Northern California high school participated in the experiment, 14 from General Biology and 10 from AP Biology, with half of each population assigned to each condition. Instruction in both biology courses was generally consistent with a bounded framing. Students took notes from lectures, the textbook, and educational movies, and teachers evaluated their ability to correctly recall individual facts from these sources. The AP course may have been framed somewhat expansively by its implicit linking to the end-of-year AP exam and college.

Similarities in Procedures Across All Participants

We controlled for objective features of the contexts in which tutoring occurred as well as elements of instruction commonly known to affect learning and transfer.

Objective Features of Context On day 1, the tutor was the first author and the videographer was a research assistant. On day 2, the tutor and videographer were different research assistants. Both days of the study occurred in the same laboratory room, but the student, tutor, and videographer were located in different places on each day.

Target Content to Transfer The learning goal for the first day was to have all students master the same facts and principles about the cardiovascular system. Transfer to the respiratory system would be assessed on day 2. For facts, students learned the sequence of body parts through which blood flows—a sequence that overlaps with where oxygen travels within the respiratory system. This material is necessary for forming correct mental models of each system (Chi et al., 1994; Liu & Hmelo-Silver, 2009). For principles, students learned that *pressure differentials* determine the direction of blood flow in the cardiovascular system, which applies to gas movement in the respiratory system and fluid flow more generally. They also learned that a large collective *surface area* increases diffusion across capillaries, which applies to increasing the rates of diffusion across alveoli in the respiratory system as well as chemical reactions, heat transfer, and many other processes.

Tutoring Methods The foundation of day 1's tutoring was having each student self-explain the same text and diagrams about the cardiovascular system (Chi et al., 1994, 2001). This method, which promotes learning and transfer (e.g. Chi et al., 1994; Rittle-Johnson, 2006), also allowed us to reduce and control for the tutor's role as provider of content. Drawing on methods established in prior research (Chi et al., 1994; McNamara, 2004), we first trained students to self-explain using an unrelated science text, and then asked them to read each sentence from the cardiovascular system text out loud and self-explain it. Although most students self-explained without difficulty, if the tutor observed a student only paraphrasing (cf. Hausmann & vanLehn, 2007), she prompted for a more elaborate explanation.

Self-explaining was supplemented by having students:

1. Identify key body parts from the text on diagrams.
2. Answer questions about structures, behaviors, functions, and their relationships (Goel et al., 1996).
3. Draw diagrams to represent their evolving models of the cardiovascular system (Ainsworth & Loizou, 2003).
4. Interact with a gestural or physical model for each target principle.

Tutoring about the respiratory system on day 2 was less guided. Students were first asked to anticipate what they would need to learn about the respiratory system. They were then given as long as they wished to "think aloud" while reviewing a text (adapted from Hmelo-Silver & Pfeffer, 2004), hypermedia system (identical to Liu & Hmelo-Silver's 2009), and diagrams. They were also provided with pen and paper, but not required to use it, which provided an opportunity for them to transfer the learning strategy of drawing diagrams from day 1. Each student was also asked to: (a) explain a lung model representing pressure differentials, and (b) explain why there are so many alveoli in the lungs, which relates to the surface area principle.

Known Instructional Supports for Transfer Use of known transfer mechanisms was controlled for all students in ways designed to avoid floor and ceiling effects. All students received the same: (a) overlapping surface linguistic cues between learning and transfer contexts (Catrambone, 1998), (b) examples of each principle (e.g. Gick & Holyoak, 1983), (c) comparisons between those

examples (e.g. Gentner et al., 2003), and (d) level of abstraction of statements of each principle (e.g. Reeves & Weisberg, 1994). No students were given any direct hints (e.g. Anolli et al., 2001; Gick & Holyoak, 1983), nor was the respiratory system mentioned prior to day 2.

Operationalization of the Framing Manipulation

We manipulated the framing of five key aspects of contexts: *who, when, where, what,* and *how* (Engle, 2006). Here we provide illustrations of each framing in classrooms and then in the experiment, with each sentence presenting the more bounded framings first and the more expansive ones second.

Who Is Involved? Lessons can be framed as just involving the teacher and each student, or as being relevant to a much larger community in the classroom and beyond. In our experiment, we framed the student as interacting separately with each tutor versus collectively with the whole research team and anyone else students mentioned.

When Is It Happening? The temporal horizon of a lesson can be framed as an isolated event that has been completed or as part of an ongoing activity that will be continuing. In our experiment, we framed each day as separate studies that consisted of separate completed sub-events versus as one ongoing study that extended across the two days and beyond to other times students mentioned as being relevant to them.

Where Is It Happening? Lessons can be framed as only being relevant to the particular classroom or as also being relevant to other settings like the rest of the school, the local community, a workplace, etc. We framed tutoring as being contained to the room versus being relevant throughout the university and anywhere else students mentioned.

What Is the Scope of the Activity? Two lessons can be framed as being relevant to separate classes, topics, or curriculum units; or as being part of the same larger subject area, unit or topic. In our experiment, we framed each day as a separate tutoring session about a different topic versus part of a pair of tutoring sessions about a larger topic.

How Are Learners Positioned Intellectually? In lessons, learners can be framed as disconnected recipients reporting about the ideas of others or as authors and respondents who take ownership of their own ideas. In our experiment, we framed the learner as a spokesperson for the text versus as the author of his or her own ideas about the body.

Instruments

Post-tutoring Survey To measure whether students detected the intended framing and their general level of motivation during tutoring, the videographer asked each student to complete a survey during a break after tutoring. The tutor was out of the room during its administration.

Cardiovascular System Pre/Post Test At the start and end of day 1's tutoring session, a written pre/post test (adapted from Chi et al., 1994) measured students' knowledge of the target facts and principles about the cardiovascular system that could be applied to the respiratory system.

Respiratory System Pre/Post Test To measure transfer we devised analogous written assessment questions about the respiratory system. The fact question and the first question about each principle comprised the three-item screening test used to select students to participate in the study.

Analytical Methods

We coded assessments at all five time points—screening, pre-cardiovascular, post-cardiovascular, pre-respiratory, and post-respiratory. Coding was done blind to condition and not by the first author.

We assessed transfer of facts and principles using three different but partially overlapping measures in order to measure converging evidence of transfer effects. *Transfer-of-knowing* is when a student *knows* something about one topic that they apply later to a related topic. It was measured by calculating the proportion of material included in either of the cardiovascular tests that re-appeared in the respiratory system pre-test. *Transfer-of-learning* is when a student *learns* something about one topic that they apply later to a related topic. It was measured by calculating the proportion of material that appeared in the cardiovascular system post-test but not in its pre-test that then re-appeared in the respiratory system pre-test. Finally, *transfer-after-exposure* is when a student increases the extent to which they use a set of ideas with one topic after being *exposed* to those same ideas with a related topic. It was measured by calculating the proportion of material not known in the respiratory screening test that was included in the same parts of the respiratory pre-test.

We measured transfer of the learning strategy of diagram drawing by simply recording which students spontaneously chose to draw diagrams during day 2's tutoring.

Results

Students Perceived Differences in Framing

Day 1's survey indicated that students generally perceived the intended differences in framing. Students in the expansive condition perceived greater use of expansive framing than those in the bounded condition ($F(1,19)= 10.6$, $p < .01$), a large effect (Cohen's $d = 1.4$). There was no interaction effect or main effect of population. Follow-up analyses found that students were most aware of the framing of intellectual positioning and temporal horizon.

No Differences in Other Factors Affecting Transfer

There were no significant differences between conditions in common factors affecting learning and transfer. Prior knowledge, as measured by the screening test, was similar across groups. There also were no differences in time spent learning or in responses to the motivation question ("how

much did you care about learning the cardiovascular system?"). Perhaps most importantly, there were no differences by condition in how much students learned the facts and principles whose transfer serves as the main outcome of this study.

Differences by Condition in the Transfer of Facts

To assess the transfer of facts, we examined responses to a question on each corresponding test that required listing the body parts that *oxygenated* blood (cardiovascular system) or *oxygen* (respiratory system) passes through between the lungs and the body's cells. Because these two paths involve the same 10 body parts we assessed transfer by counting how many were listed in each test and comparing them.

For transfer-of-knowing, there was a large main effect of condition (see Fig. 1, error bars are SEM), with students in the expansive condition transferring 42% of facts they knew while those in the bounded condition only transferred 21% of them (d = .89; $F(1,20)$ = 4.37, p = .04). There were no population nor interaction effects. For transfer-after-exposure, there was also a large main effect of condition (d = .94), with students in the expansive condition listing 20% more facts than they had during the screening test while students in the bounded condition listed only 3% more facts ($F(1,19)$=4.82, p = .04; Fig. 1). Again, there were no other effects. For transfer-of-learning, there was a trend of more transfer for expansive (36%) versus bounded (13%) conditions ($F(1,20)$=3.27, p = .09; Fig. 1), with no other effects found.

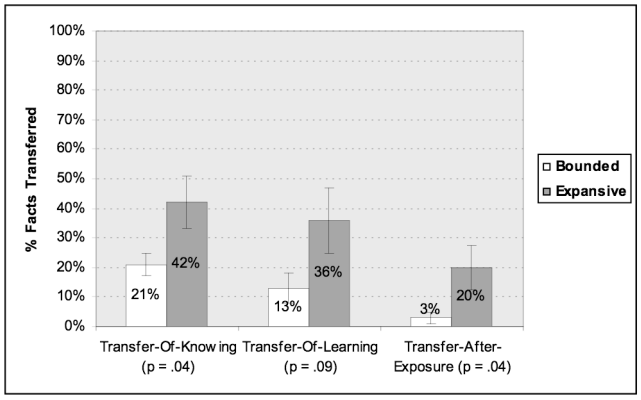


Figure 1: Greater transfer of facts in expansive condition.

Partial Evidence for Differences by Condition in the Transfer of Principles

To measure degree of transfer for principles, we divided each principle into a set of propositions that could be included in student responses to analogous questions at each testing occasion. There were 12 codeable propositions relevant to the differential pressure principle and 11 codeable propositions relevant to the surface area principle (91% agreement; Kappa = .82).

For the differential pressure principle, there was a large main effect (d = .95) of condition on transfer-of-knowing ($F(1,20)$ = 5.42 p = .03), with no interaction effect or main effect of population (see Fig. 2). Students in the expansive condition transferred much (M = 78%) of what they knew while those in the bounded condition transferred only about half (M = 55%). For transfer-of-learning, there was a trend for students in the expansive condition to transfer more than the bounded condition (74% vs. 46%; $F(1,21)$=3.04, p=.098; see Fig. 2). Upon further examination of the data, however, we suspect this trend was driven by the General Biology students. There were no differences between groups in transfer-after-exposure. Thus, we found a statistically reliable effect of framing for one of the three measures of transfer for the differential pressure principle.

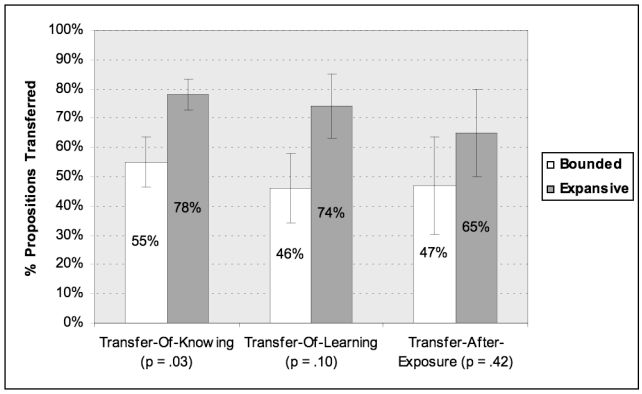


Figure 2: Generally greater transfer of the differential pressure principle in the expansive condition.

In contrast, for the surface area principle there were no main or interaction effects on transfer when measured in each of the three ways. Although the observed means did favor the expansive condition with the transfer-of-knowing measure, there is no reliable evidence that framing affected students' propensity to transfer what they knew or learned about the surface area principle.

Differences by Condition in the Transfer of the Learning Strategy of Drawing Diagrams

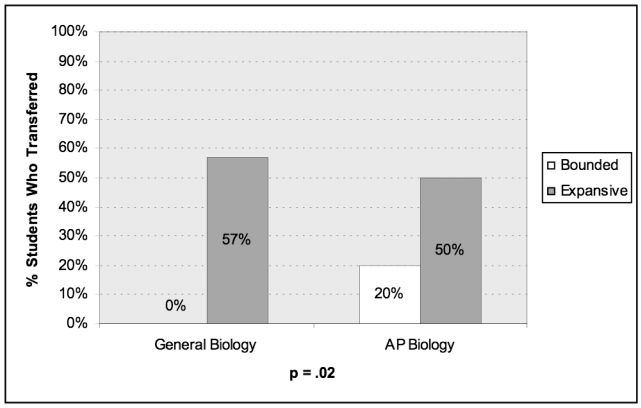


Figure 3: Greater transfer of the strategy of drawing diagrams in the expansive condition.

On the basis of a 2x2x2 loglinear analysis, students in the expansive condition were much more likely to draw diagrams than those in the bounded condition ($G^2(2) = 8.28$, $p = .02$). Only 1 of the 12 students in the bounded condition drew diagrams while 7 of the 12 students in the expansive condition did so (see Fig. 3). There was a trend of this effect being greater for General Biology than AP Biology students ($G^2(4) = 8.52$, $p = .07$).

Discussion

We found compelling initial evidence that framing may in fact influence transfer. Students in the expansive condition were more likely than those in the bounded condition to transfer: (a) the learning strategy of drawing diagrams; (b) facts they knew or (c) had been exposed to; and (d) what they knew about the differential pressure principle. In addition, (e) General Biology students in the expansive condition were more likely to transfer what they learned about the differential pressure principle.

The fact that several large effects of framing on transfer were found within a small-scale experiment suggests that it is likely that framing does play an important role in transfer. Also framing does not appear overly specialized in terms of what kinds of transfer it can influence. In this study it affected the transfer of facts, principles, and strategies while in prior research it influenced the transfer of actions, experiences and explanatory schemes (Engle, 2006; Hammer et al., 2005; Hart & Albarracin, 2009).

In future research it will be important to investigate whether it is the framing of one particular aspect of contexts that is responsible for the effects or whether all are necessary. For example, is the manipulation of intellectual positioning as authors versus spokespersons the most important, or is it the way in which time and other aspects of settings are framed as being linked with each other? If more than one aspect of expansive framing matters, does each one make its own independent contributions or is the whole greater than the sum of its parts? To address these questions, future experiments can manipulate each aspect of framing alone and in coordination. This will simultaneously advance understanding of how exactly framing works, provide replication of the effects reported here, and guide educators about which aspects of framing to focus on.

Although transfer of the differential pressure principle was found, no differences were detected across conditions in any kind of transfer of the surface area principle. This contrast opens up issues about how framing may interact with other mechanisms for supporting transfer. This could suggest that framing's effects on transfer may be found only when there is at least some minimal level of content-based support for transfer. In this study, we provided more examples and comparisons for the differential pressure principle than the surface area principle. However, this outcome could also be due to the fact that the surface area principle is arguably more complex. To distinguish between these possible interpretations, follow-up experiments could cross content-based support with framing while controlling for principles.

More generally it is possible that the framing of learning contexts in an expansive manner makes it more likely that students assume they will need to transfer what they have learned, which may prompt them to make better use of those content-based supports for transfer that are available to them (Engle, 2006). For instance, students learning with an expansive framing may be more likely to bring in multiple examples from a wide range of contexts. In anticipation of applying what they are learning, they may also be more likely to make systematic comparisons between multiple examples to form abstract generalizations. Although tracking which examples, comparisons, and generalizations students made was beyond the scope of this study, it would be a compelling focus of future investigation. Future investigations also should more systematically probe whether motivational variables like utility, relevance, and importance mediate these effects (Pugh & Bergin, 2006).

What is potentially so powerful about expansive framing is that it is much less targeted and content-specific than previously studied instructional supports for transfer. Because of this, it may be easier for teachers to implement expansive framing than instructional supports for transfer that rely on sophisticated content knowledge. In addition, as students come to regularly orient to learning activities in an expansive fashion, one would expect them to make greater use of prior knowledge more generally as they become increasingly accountable for sharing what they know across connected contexts.

At the same time, we do not claim that expansive framing is the be-all and end-all for instruction. Our informal observations of the tutoring sessions and broader theoretical considerations suggest that there may be costs as well as benefits of expansive framing for both learning and transfer. For example, we observed a few students in the expansive framing condition that brought in so much prior knowledge while self-explaining that they became overwhelmed or had difficulty focusing on what the text could contribute to their understanding. Thus, it may make sense for the starts and ends of lessons and curriculum units to be framed more expansively, but to use a less expansive framing when students need to focus on learning particular new material. Also, expansive framing should ideally be paired with activities in which students critically evaluate the knowledge they transfer in for its relevance and validity.

In closing, this study provides converging evidence that framing is an important instructional mechanism to consider when trying to enhance transfer, one that can potentially affect the transfer of many different kinds of knowledge.

Acknowledgements

This research was supported by a UC Berkeley junior faculty grant, a Hellman Family Faculty Fund grant, and National Science Foundation (NSF) Grant #0844910 to Randi A. Engle. We thank Seda Bourikian, Pegah Ghaneian, Pauline Huang, Diane Lam, Jonathan Lesser, Xenia Meyer,

Sarah Nix, Melissa Pandika, Sharla Roberts, Sadaf Sareshwala, Alexandra Tee, and Pamela Yee for assistance. We also thank our reviewers for their invaluable input.

References

Ainsworth, S. & Loizou, A. Th. (2003). The effects of self-explaining when learning with text or diagrams. *Cognitive Science, 27,* 669-681.

Anolli, A., Antonietti, A., Crisafulli, L. & Cantoia, M. (2001). Accessing source information in analogical problem-solving. *Quarterly Journal of Experimental Psychology, 54A*(1), 237-261.

Barnett, S. M. & Ceci, S. J. (2002). When and where do we apply what we learn? A taxonomy for far transfer. *Psychological Bulletin, 128*(4), 612-637.

Catrambone, R. (1998). The subgoal learning model: Creating better examples so that students can solve novel problems. *Journal of Experimental Psychology: General, 127*(4), 355-376.

Catrambone, R. & Holyoak, K. J. (1989). Overcoming contextual limitations on problem-solving transfer. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*(6), 1147-1156.

Chi, M. T. H., de Leeuw, N., Chiu, M-H. & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science, 18,* 439-477.

Chi, M. T. H., Siler, S., Jeong, H., Yamauchi, T., & Hausmann, R. G. (2001). Learning from human tutoring. *Cognitive Science, 25,* 471-533.

Engle, R. A. (2006). Framing interactions to foster generative learning: A situative account of transfer in a community of learners classroom. *Journal of the Learning Sciences, 15(4),* 451-498.

Floriani, A. (1994). Negotiating what counts: Roles and relationships, texts and contexts, content and meaning, *Linguistics and Education, 5,* 241-274.

Gee, J. P. & Green, J. L (1998). Discourse analysis, learning, and social practice: A methodological study, *Review of Research in Education, 23,* 119-69.

Gentner, D., Loewenstein, J. & Thompson, L. (2003). Learning and transfer: A general role for analogical encoding. *Journal of Educational Psychology, 95(2),* 393-408.

Gick, M.L. & Holyoak, K. (1983). Schema induction and analogical transfer. *Cognitive Psychology, 15*, 1–38.

Goel, A. K., Gomez de Silva Garza, A., Grué, N., Murdock, J. W., Recker, M. M., & Govinderaj, T. (1996). Towards designing learning environments I: Exploring how devices work. In C. Fraisson, G. Gauthier, & A. Lesgold (Eds.), *Intelligent tutoring systems: Lecture notes in computer science*. Berlin: Springer-Verlag.

Duranti, A. & Goodwin, C. (Eds.) (1992). *Rethinking context: Language as interactive phenomenon* Cambridge, UK: Cambridge University Press.

Greeno, J. G., Smith, D. R., & Moore, J. L. (1993). Transfer of situated learning. In D. K. Detterman & R. J. Sternberg (Eds.), *Transfer on trial: Intelligence, cognition, and instruction*. Norwood, NJ: Ablex.

Hammer, D., Elby, A., Scherr, R. E., & Redish, E. F. (2005). Resources, framing, and transfer. In J. Mestre (Ed.), *Transfer of learning: Research and perspectives*. Greenwich, CT: Information Age Publishing.

Hart, W. & Albarracin, D. (2009). What was doing vs. what I did: Verb aspect influences memory and future actions. *Psychological Science, 20*(2), 238-244.

Hausmann, R. G. M. & vanLehn, K. (2007). Explaining self-explaining: A contrast between content versus generation. In R. Luckin, K. R. Koedinger & J. E. Greer (Eds.), *Artificial intelligence in education: Building technology rich learning contexts that work*. Amsterdam: IOS Press.

Hmelo-Silver, C. E. & Pfeffer, M. G. (2004). Comparing expert and novice understanding of a complex system from the perspective of structures, behaviors, and functions. *Cognitive Science, 28,* 127-138.

Laboratory of Comparative Human Cognition (1983). Culture and cognitive development. In P. H. Mussen (Ed.), *Handbook of child psychology: Vol. 1. History, theory and methods*. New York: Wiley.

Lave, J. (1988). *Cognition in practice: Mind, mathematics, and culture in everyday life*. New York: Cambridge University Press.

Lobato, J. (2006). Alternative perspectives on the transfer of learning: History, issues, and challenges for future research. *Journal of the Learning Sciences, 15*(4), 431-449.

Lui, L. & Hmelo-Silver, C. E. (2009). Promoting complex systems learning through the use of conceptual representations in hypermedia. *Journal of Research in Science Teaching, 46*(9), 1023-1040.

McNamara, D. (2004). SERT: Self-explanation reading training. *Discourse Processes, 38*(1), 1-30.

Pea, R. D. (1987). Socializing the knowledge transfer problem. *International Journal of Educational Research, 11*, 639–663.

Pugh, K.J. & Bergin, D.A. (2006). Motivational influences on transfer. *Educational Psychologist, 41*(3), 147-160

Reeves, L.M. & Weisberg, R.W. (1994). The role of content and abstract information in analogical transfer. *Psychological Bulletin, 115(3),* 381-400.

Rittle-Johnson, B. (2006). Promoting transfer: Effects of self-explanation and direct instruction. *Child Development, 77*(1), 1-15.

Shadish, W., Cook, T. & Campbell, D. (2002). *Experimental and quasi-experimental design for generalized causal inference.* New York: Houghton-Mifflin.

Spencer, R. M. & Weisberg, R. W. (1986). Context-dependent effects on analogical transfer. *Memory and Cognition, 14,* 442-449.

Tannen, D. (Ed.) (1993). *Framing in discourse*. New York: Oxford University Press.

van Dijk, T. A. (2008). *Discourse and context: A sociocognitive approach.* New York: Cambridge U.P.

Impatience as Intertemporal Egoism

Daniel M. Bartels (bartels@uchicago.edu)
Center for Decision Research, University of Chicago
5807 S. Woodlawn Ave., Chicago, IL 60637 USA

Oleg Urminsky (ourminsk@chicagobooth.edu)
University of Chicago Booth School of Business
5807 S. Woodlawn Ave., Chicago, IL 60637 USA

Abstract

We build on a philosophical account of personal identity (Parfit, 1984) which argues that the degree of concern one has for one's future self may be scaled by the degree of "psychological connectedness"—overlap in properties such as beliefs, values, and ideals—held between one's current and future self. We pose participants with tradeoffs between consuming a benefit in the near future versus consuming more of that benefit in the distant future. When people's sense of connectedness with their future self is reduced, they make impatient choices.

Introduction

Many of the most important and difficult decisions we face in life hinge on the same underlying dilemma: how to choose when trading off consumption or happiness in the immediate future with (more) consumption or happiness delayed to the more distant future. Research on such dilemmas has been broadly defined as concerning choices between one option with higher immediate benefits but lower (or negative) long-term utility and another with lower immediate benefits but higher long-term utility. People's widely documented tendency to prefer smaller rewards sooner over larger rewards later has been characterized as revealing short-sightedness or impatience (Elster, 1979).

In this paper, we focus on a fundamental question raised by the literature on intertemporal choice: why do people's choices often seem so short-sighted or impatient, and why do people differ in their degree of impatience, as inferred from the choices they make? In particular, we will focus on a subset of such dilemmas—intertemporal choices in which the tradeoffs between short and long-term benefits are made explicit—as an ideal setting in which to investigate the decision processes leading to impatience in decision making across a wide range of future-directed thought and behavior. Economists and psychologists have extensively studied how people make these kinds of intertemporal choices, and have offered metrics for judging the degree to which behavior conforms to or diverges from normative and descriptive models.

Much of the work on intertemporal choice has centered on the specific issue of temporal discounting: how people choose between smaller amounts of money or other goods in the immediate future and larger amounts of money or goods to be received at a later date (see Frederick, Loewenstein, and O'Donoghue, 2002 for a detailed review). In this context, the discount rate, the degree to which an outcome loses value by being delayed for a given period of time, can be interpreted as a measure of impatience (Ainslie 1975). Thus, we can restate the general question of (im)patience in intertemporal choices as asking why people exhibit such *high discount rates* (compared to market interest rates or some other norm) in their behavior and to provide a partial account for why people exhibit such different discount rates from each other.

We will argue that our understanding of what constitutes a reasonable discount rate (or, more generally, prudent vs. impatient choices) has been limited by the implicit assumption that people should maximize the utility of a constant self over one's lifetime. The philosopher Derek Parfit (1984) proposed an alternative view: that a decision about consuming now or later should depend not only on the temporal distance between events, but also on the perceived continuity between one's present and future selves. In this view, the degree of concern one has for one's future self should be scaled by the degree of "psychological connectedness"—overlap in properties such as beliefs, values, and ideals—held between one's current and future self. These properties have been proposed to define the mental ties between selves that comprise identity over time (Lewis, 1983; Perry, 1972).

We employ the notion of psychological connectedness—drawn from a literature in which there is an ongoing debate over its specifically *normative* implications (Parfit, 1984; see Dancy 1997 for dissenting views)—to test a *descriptive* account of people's intertemporal choices. In our view, the greater the perceived connectedness to the future self, the greater people's willingness to defer benefits to the future self, all else equal. Conversely, feeling disconnected from the future self will undercut the general motivation to preserve resources for the future self, causing a reduction in patience that is distinct from other factors that affect valuations of present and future outcomes.

Evidence for high discount rates

In the context of intertemporal choice, impatience (or short-sightedness) is exhibited by consistently choosing sooner-smaller options even when the latter option is more than large enough to compensate for the delay (per some normative standard). Normative models (Koopmans, 1960) indicate that the premium needed in order to forego receiving money sooner rather than later (i.e. the discount rate) should depend primarily on how much interest could be earned on the money in the intervening time, taking into account liquidity constraints and economic factors such as inflation. In contrast,

empirical research has found that actual behavior is generally more impatient than what would be predicted by these views.

Numerous studies have attempted to estimate discount rates, using field and experimental studies, real and hypothetical outcomes, and a range of elicitation methods. Frederick et al. (2002) summarize the literature as characterized by a "predominance of high discount rates—discount rates well above market interest rates." In addition to experimental studies with hypothetical choices, they review field study evidence for high discount rates (i.e. impatience) in everyday decisions, such as people's preferences for lower priced appliances with substantially higher long-term usage costs and military employees' preferences for a large lump-sum payment over an annuity representing a higher than market interest rate.

Heterogeneity in discount rates

Of the research that has shed light on high discount rates, the primary focus has been on moderators of discount rates, both across people and across decision contexts. While generally high, discount rates have been shown to be sensitive to the specific experimental elicitation methods used (e.g. choice, willingness-to-pay, matching, titration of indifference points). Discount rates have been found to exhibit reasonably high test-retest reliability as individual traits (Simpson & Vuchinich, 2000), but vary systematically by demographics (Green et al., 1994) as well as by individual differences in how people think about the long-term implications of their choices (Strathman et al., 1994).

This large literature on interpersonal differences in discounting provides correlational evidence that people often have fundamentally differing discount rates, often in ways that map onto more generalized short-sightedness. However, the behavioral correlates of discounting (e.g. higher discounting among alcohol or drug abusers), in particular, raise questions as to potential confounds and the order of causality.

Connectedness to the future self and discounting

In this paper, we propose that the notion of connectedness to the future self is fundamental for understanding impatience, shedding light on why discount rates are generally high, why some people are more impatient than others, and what kinds of interventions may lead to higher or lower discount rates. In doing so, we draw on the views of philosopher Derek Parfit, who has theorized that changes over time in the psychological properties that comprise one's identity should warrant a reduction in concern for a later self:

> "We care less about our further future… because we know that less of what we are now—less, say, of our present hopes or plans, loves or ideals—will survive into the further future… We may, because of this, act knowingly against our own long-term self-interest… [If] what matters holds to a lesser degree, it cannot be irrational to care less." (Derek Parfit 1976, p. 99)

In this view, the future self, given an extremely large reduction in connectedness, may be reasoned about almost as a distinct individual. We do not mean to overstate the analogy of regarding the future self as you would regard another—in our account, rather, the future self is seen as a continuation of the current self, to varying degrees. The future instantiations of the self may be seen as nearly identical to the current self, or they may be substantially different, and we will argue that this perceived degree of continuity leads to differences in patience.

In extending the notion of connectedness to a descriptive account of impatience, we define connectedness between the current self at time t_0 and a future self at time t_1 as the proportion of the defining psychological features of the current self believed to persist in the self that will exist at time t_1. Consistent with the recent empirical literature on how people judge the continuity of identity over time (Nichols & Bruno, in press; Rips, Blok, & Newman, 2006), connectedness between current and future selves hinges specifically on the stability of one's defining psychological properties over the time interval.

In our view, a person values future outcomes in proportion to how much she believes that the current self's important psychological characteristics will persist in the future self. When people feel highly connected to the future self, benefits received by the future self are valued much as if they were received by the present self. However, when a discontinuity in identity is perceived, deferred benefits accrue to a disconnected future self (i.e., a somewhat different person), and this outcome is valued less than having those benefits consumed by the present self. Thus, when people are faced with explicit intertemporal tradeoffs, their allocations of benefits to the future selves are driven, in part, by how psychologically connected they feel to those future selves. As a result, decisions that might appear short-sighted (i.e. characterized by placing a low weight on future consequences or having an inflated discount rate) may instead merely reflect an unwillingness to share resources with a future self who is evaluated to be substantially different from the current self.

A few studies have examined correlations between people's perception of the continuity of their identity over time and the choices they make. Ersner-Hershfield and colleagues have shown that people who perceive less continuity with the future self show greater devaluation of money (Ersner-Hershfield, Wimmer, & Knutson, 2009; cf. Frederick 2002) and tend to have accrued fewer material assets in their lives (Ersner-Hershfield et al., 2009). Bartels and Rips (2010) investigated the role of connectedness in non-constant discount rates over time and found that declines in a given persons' discount rates over time—a pattern often referred to as "hyperbolic discounting" (Ainslie, 1975)—correlated with perceived reduction in their own connectedness over time.

In this paper, we provide the first direct, experimental evidence that changes in connectedness to the future self across individuals causes differences in patience for real choices and that the influence of connectedness on patience is distinct from other factors already identified in the literature as impacting people's relative timing preferences.

Study 1

In Study 1, we investigate the effect of manipulating connectedness on subsequent hypothetical choices between either the immediate receipt of a gift card or a gift card bundled with an additional payment to delay receipt. After

reading either that identity changes radically in early adulthood (especially during the college years) or that the core features of one's identity are fixed in early childhood (and stable during college), participants made a set of hypothetical choices between receiving a gift certificate later in the day, or receiving it in a year along with an additional payment to compensate for the delay. If disconnectedness from the future self is a driver of discounting, then anticipating changes in the properties that comprise one's identity will make people more impatient, and participants exposed to the instability message should require a larger delay premium than participants exposed to the stability message.

Method

One hundred seven undergraduates were approached in a dining hall on campus and agreed to fill out a short survey for a chocolate square. We manipulated connectedness (high vs. low, between subjects) by inducing the belief that the identity of the future self will either change or not change from one's current identity. Specifically, in the high-connectedness condition, participants began by reading a short description of "recent research" suggesting that young adulthood is characterized by stability in identity (e.g., "the important characteristics that make you the person you are right now... are established early in life and fixed by the end of adolescence"). In the low-connectedness condition, participants read about instability (e.g., "the important characteristics that make you the person you are right now... are likely to change radically in young adulthood."). Then, participants wrote a short summary of the passage they read. Data from four participants were dropped from further analysis because they left this response blank or because their paraphrasing indicated misunderstanding or noncompliance.

Next, participants in both conditions were asked to imagine being given a $120 gift certificate. We used two different retailers, Target and Expedia, to ensure the generalizabilty of results. They were then asked to make a series of choices between receiving the gift certificate later that day vs. receiving the gift certificate one year later and being paid an extra amount for the delay, using eight dollar values (0, 17, 34, 51, 69, 86, 103, and 120). Participants then answered two kinds of manipulation checks: an assessment of connectedness and a rating of the believability of the passage they had read. To assess connectedness, we asked participants to "think about the important characteristics that make you the person you are now—your personality, temperament, major likes and dislikes, beliefs, values, ambitions, life goals, and ideals and circle the one diagram out of the six below that best reflects your opinion about the degree of connectedness between the person you are now and the person you will be in a year, where no overlap means 'completely different' and complete overlap means 'exactly the same.'" Participants circled one of the six sets of Euler circles representing connectedness, which were coded as numeric scores.

Results and Discussion

Manipulation checks. Participants who read about stability rated themselves as more connected (M = 4.43, SD = 0.73) than did participants who read about instability (M = 4.00, SD = 1.07; $t(1,102) = 2.39$, $p < .05$), suggesting that our manipulation was effective in promoting perceptions of one's own identity as more (or less) stable over time and therefore more (or less) connected to one's future self. Believability of the stability and instability passages did not differ ($t < 1$).

Relating perceived (in)stability to discounting. Our measure of patience was the number of deferred options (waiting one year for the gift certificate) chosen out of the eight given, such that larger values indicated greater patience. Participants in the high-connectedness conditions were more patient, requiring a smaller delay premium, on average ($49, inferred from M = 5.14), than did participants in the low-connectedness conditions ($68, inferred from M = 4.04). A 2 (Condition: High/Low Connectedness) x 2 (Good Type: Target/Expedia) ANOVA, finds the expected main effect of Connectedness ($F(1,100) = 9.21$, $p < .01$); neither the effect of good type nor the interaction term reached significance (Fs < 1). These results demonstrate both that perceived connectedness to one's future self can be directly manipulated, and, more importantly, that increasing perceived connectedness to the future self increases patience.

Study 2

Study 1 shows that disconnectedness causes impatience, as revealed by the premium people demand to delay receiving an award. Implicit in these choices is that people are depriving their future selves of potential resources, in order to consume sooner. However, it is not necessarily the case that people think of such tradeoffs in terms of allocating resources, and they may instead think in terms of fair rates of return for delay or other factors. In fact, the literature has shown that framing matters in such choices: while people are generally willing to accept compensation to wait to consume, they are much less willing to actually pay to speed up an outcome, due to the pain of paying and other factors (Loewenstein, 1988). We have argued that reduced connectedness impacts preferences due to a reduced willingness to share resources with a future self who is evaluated to be substantially different from the current self. In our view, the effects of connectedness should persist even when it is made explicit to participants that they have to, in effect, deprive their future self of resources in order to consume now, thereby highlighting the future consequences of impatience. In this study, we test whether disconnectedness causes impatience so pronounced that people actually are willing to spend their own money in order to consumer sooner.

The results of Study 1 show that over periods of time where one might reasonably expect meaningful change in the properties that comprise one's identity, providing information that highlights the likelihood of decreased connectedness leads to more impatience. Note, however, that the way in which people's perceived connectedness was manipulated relied on participants in different groups being presented with different information. A potential concern, then, is that participants' choices may have reflected a lay theory about what the appropriate effect of changes in identity on patience should be,

rather than reflecting their true preferences. Study 2 addresses this concern in two ways: (i) we pose participants with a decision involving real economic outcomes, and (ii) we manipulate connectedness while keeping the information content the same across the two conditions.

In this study, we used the inferences that participants reached from a metacognitive cue to manipulate their sense of connectedness to the future self. Specifically, we drew from the work on "accessibility experiences" (Schwarz, 2004) to indirectly manipulate people's perceptions of the stability of their identity, by asking them to judge how difficult it would be to generate either 2 or 10 reasons why their identity will remain very stable over the next 12 months. Participants asked to imagine how difficult it would be to generate two reasons should find the task easy, and therefore have no reason to doubt the stability of their identity. Conversely, participants in the 10 reasons condition should anticipate that the task would be more difficult, and are likely to use this anticipated difficulty as a cue to question the stability of their identity, yielding a feeling of low connectedness.

Method

As part of a larger study, one hundred five graduating seniors filled out an online survey 1-2 weeks before their graduation in return for $4 and entry into a lottery for which they could receive a real gift certificate.

All participants were presented with a passage that described the effect of college graduation on the stability of one's identity as mixed. Participants in the high-connectedness condition were then asked to judge (on a 7-point scale) how easily they could generate 2 reasons why their own identity would remain very stable over the next 12 months (i.e., before and after graduation), after reading that most participants were able to generate 2 reasons in a previous study. In the low-connectedness condition, participants judged how easily they could generate 10 such reasons, after reading that most participants previously had been able to generate 10 reasons.

Lastly, they read that they had been entered into a lottery for a gift card. They read:

> "The drawing will occur in two weeks, and if your survey is chosen, you will receive a $95 Amazon.com gift card either in one year, or you can pay to receive it immediately after the drawing is held in two weeks.
>
> What is the maximum amount that you would be willing to pay now to be able to use the $95 gift card immediately?"

Results and Discussion

Manipulation checks. Participants in the 2 reasons (high-connectedness) condition rated the reason-generation task as relatively easy ($M = 5.28$, $SD = 1.51$) compared to the ratings of the participants in the 10 reasons (low-connectedness) condition ($M = 4.58$, $SD = 1.81$; $t(103) = 2.15$, $p < .05$).

Relating perceived (in)stability to willingness to pay to expedite gift certificate. Participants in the 10 reasons (low-connectedness) condition were willing to pay more to speed up receipt of the gift certificate ($M = \$14.83$, $SD = 15.96$) than were participants in the 2 reasons (high-connectedness) condition ($M = \$9.49$, $SD = 8.99$; $t(103) = 2.16$, $p < .05$). In other words, participants made to feel disconnected from the future self were significantly more impatient—as in the previous study, they strongly preferred to allocate benefits to their sooner, more connected self over their later, less connected self. Unlike the previous study, making impatient choices did not merely imply being less generous to the future self, but rather required the participants to, in effect, deprive their future self of resources in order to consume sooner, thus highlighting the long term consequences of impatient choices.

Study 3

The studies above provide the first evidence that directly manipulating connectedness systematically affects people's patience for the outcomes they will receive. Next, we test whether naturally-occurring individual differences in perceived connectedness to the future self relate to individual differences in patience. One goal was to rule out the possibility that the observed effects on impatience could be attributed to highlighting the notion of connectedness for our participants prior to their choices, by extending the findings to more natural contexts in which people might or might not spontaneously reflect on connectedness when making choices. Recall that in the prior studies, we manipulated perceived connectedness and then asked for people's preferences. In this study, we instead employed a re-contact methodology. In the first stage, we measured connectedness (without manipulating it). Three weeks later, in a separate study, we re-contacted participants and collected preference data followed by measures of other psychological constructs known to affect intertemporal choice.

The second goal of this study was to assess the impact of several other variables that could contribute to possible alternative explanations for our findings. In particular, we assess whether intertemporal preference is affected by connectedness, even when we control for natural variation in several psychological factors, distinct from connectedness, that have been linked to patience in the literature. Furthermore, by simultaneously assessing the relationship of individual differences in connectedness and these alternative psychological factors with patience, we can gauge how large an impact connectedness has on patience relative to the impact of other important psychological factors..

In order to test whether rated connectedness has a unique influence on patience when controlling for other potentially explanatory variables, we included measures of (i) degree of "projection bias", (ii) future anhedonia, (iii) time perception, (iv) reward responsiveness, and (v) non-planning impulsiveness at the end of the second survey.

Projection bias is a measure that captures whether people believe that specifically their *tastes and preferences* will be different in the future (Loewenstein et al. 2003), which might lead people to consume sooner, rather than later, if they think delayed benefits might not fit the future self's tastes. "Future anhedonia" refers to an affective forecasting phenomenon where people view both positive and negative outcomes as less extreme the farther into the future these outcomes occur. Viewing both positive outcomes as less extreme when delayed

to the farther future may cause people to consume benefits sooner, when their positive qualities are more intense (Kassam et al., 2008). Time perception has been implicated by Zauberman et al. (2009) as a partial explanation for hyperbolic discounting and for high discount rates in the near future. In this view, the proportion of value retained over a given delay is linearly-related to the *perceived* duration of the delay, rather than the actual duration.

Lastly, people who score high in reward responsiveness (degree of desire induced by a reward; Carver & White, 1994) may be more susceptible to factors that induce impulsivity in discounting tasks, and non-planning impulsiveness (inability to resist temptation; Patton, Stanford, & Barratt, 1995) has been linked to higher discount rates (Hinson, Jameson, and Whitney 2003).

We argue that psychological connectedness predicts intertemporal choice over and above these other contending variables, and it does so even in a context in which the idea of connectedness to the future self is not brought to mind by the study's procedure. So, this study assesses the contribution of connectedness to patience, relative to the influence of several other relevant psychological factors.

Method

Ninety four undergraduates participated in the first round of data collection, 57 of whom agreed to participate in the second round of data collection when re-contacted. Participants in the first survey were paid $1 for their time, and those who agreed to participate in the second survey participated in exchange for entry into a lottery for a $50 gift certificate.

First survey. Participants gave three sets of connectedness ratings. First, as in Study 1, they circled the pair of Euler circles that best represented their perceived degree of connectedness. Next, participants were asked to think again about these identity-comprising properties and to rate connectedness on a 0 to 100 scale. Finally, participants were asked to draw a mark on a line to rate their connectedness. The multiple measurement procedures enabled us to limit the impact of elicitation method-specific biases.

Second survey. Approximately three weeks later, we re-contacted our participants, offering them an opportunity to participate in a second round of data collection. They were first presented with a titration task, in which they made real choices between receiving a $50 gift card for Amazon.com (if their survey was chosen) in a week, when the drawing would be held, or receiving a larger-valued gift card in a year ($50, 58, 66, 74, 82, 90, 98 or 106). Next, they responded to items which measured (i) projection bias, (ii) future anhedonia, (iii) time perception, (iv) reward responsiveness, and (v) non-planning impulsiveness.

Results and Discussion

We combined the three connectedness ratings (Euler circles, similarity rating, and line scale) into an index of connectedness which yielded high internal reliability (α = .91). We used this index, along with the alternative variables, to predict people's discounting, as expressed in their choices of gift certificates. Our measure of patience is simply the number of deferred, larger rewards chosen. We first correlated patience to each predictor variables individually; then conducted a multiple regression, including all predictor variables simultaneously.

Our index of psychological connectedness in the first survey was significantly correlated with patience for receiving a gift card, as measured three weeks later (r=.29, p<.05). In addition, projection had a marginally significant effect (r=-.24, p<.10), such that those who anticipated that their tastes would change exhibited less patience. None of the other measures had a significant correlation with patience in the gift card task.

More importantly, connectedness predicts patience in a multiple regression (β=.78, p <.05) controlling for the other factors which have been shown, in other circumstances, to exert their own influences on patience (but which were not significant in this regression). This finding is particularly striking, given that we measured each construct (connectedness and patience) uncontaminated by the other construct, due to the three week delay between the two measures. Thus, the fact that psychological connectedness remains a significant predictor of patience, even when all of the factors are entered in the regression simultaneously (model R^2 = .20), provides strong evidence for both the distinctiveness and pervasiveness of psychological connectedness as an explanation for discounting.

Summary and Conclusions

The three studies described here show that people's beliefs about the stability of the important characteristics that determine their identity over an interval of time also determine the patience they exhibit over that interval. People who perceive relatively less connectedness to their future selves require a larger delay premium to wait for a gift card, pay more to expedite receipt of a gift card, and are more likely to favor smaller-valued gift cards over larger-valued, delayed gift cards than people who feel highly connected to their later selves. Perceived connectedness, in turn, can be influenced by exposure to information regarding the variability of identity-comprising characteristics over time and by the ease with which reasons for expecting stability over time can be generated. We found that both manipulated and measured perceptions of connectedness influence intertemporal choice, even when connectedness is not brought to mind in the testing session. Moreover, in the last study, connectedness was shown to be a unique, and in our data, the strongest predictor of discounting compared to other psychological factors.

Taken together, these results shed light on a heretofore under-represented explanation of discounting, and one that is quite well-grounded theoretically (Parfit, 1984): A powerful determinant of people's future-oriented preferences, plans, and behavior is the person they expect to be when outcomes are realized. When this later person is more closely connected to the current self in terms of sharing important psychological properties, the decision maker is more motivated (consciously or not) to act patiently—that is, in a manner that reflects greater consideration of the later self's welfare.

To our knowledge, the current studies are the first to manipulate perceived connectedness to a later self and the first

to assess the descriptive adequacy of this determinant of discounting against the adequacy of other determinants. It is important to note, however, that temporal discounting is likely to be multiply-determined. There have been several attempts to integrate these multiple determinants in models of discounting (e.g., Killeen, 2009), but because none of the existing models accommodate how inferences about continuity of self over time affect preference, none explicitly account for the effects we have demonstrated. A model designed to capture our effects would need to incorporate a parameter which represents the degree of connectedness, such as the proportion of the defining characteristics of the current self's psychological make-up believed to persist in the future self at future points in time. Discounted utility would then be scaled by this parameter, representing the partiality towards more connected selves which we hypothesize and provide evidence for.

The key intuition of our framework that is absent from other accounts of discounting is that "impatience" can be the result of simply allocating less to a future self that is seen, to varying degrees, as a continuation of the current self. And notably, by our account, allocating less utility to a less connected later self is thus not necessarily a mistake. However, in those contexts where it is a mistake—for example, where people consistently fail to maintain their plans in advance of temptation (e.g., under-saving relative to budgetary allowances)—fostering the sense that what matters most in defining us persists over time may help us persist in achieving important goals, including those that most help us maintain what defines us.

Acknowledgments

Thanks to Gretchen Chapman, Kristin Diehl, Hal Ersner-Hershfield, Shane Frederick, Ryan Hamilton, Reid Hastie, Craig Joseph, Aparna Labroo, Sean Nichols, Pete McGraw, Doug Medin, Greg Murphy, Chris Olivola, Howard Rachlin, Ed Smith, Stephen Spiller, George Wu, and special thanks to Lance Rips for feedback, suggestions, and encouragement regarding this project.

References

Ainslie, G. (1975). Specious reward: A behavioral theory of impulsiveness and impulse control. *Psychological Bulletin, 82*, 463-496.

Bartels, D. M. & Rips, L. J. (2010). Psychological connectedness and intertemporal choice. *Journal of Experimental Psychology: General, 139*, 49-69.

Carver, C. & White, T. (1994). Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment. *Journal of Personality and Social Psychology, 67*, 319–333.

Dancy, J. (1997). *Reading Parfit*. Oxford: Blackwell

Elster, J. (1979). *Ulysses and the Sirens*. NY: Cambridge U.

Ersner-Hershfield, H., Garton, M.T., Ballard, K., Samanez-Larkin, G.R., & Knutson, B. (2009). Don't stop thinking about tomorrow: Individual differences in future self-continuity account for saving. *Judgment and Decision Making, 4*, 280-286.

Ersner-Hershfield, H., Wimmer, G.E., & Knutson, B. (2009). Saving for the future self: neural measures of future self-continuity predict temporal discounting. *Social Cognitive and Affective Neuroscience, 4*, 85-92.

Frederick, S. (2002). Time preference and personal identity. In *Time and decision: Economic and psychological perspectives on intertemporal choice*, G. Loewenstein, D. Read, & R. Baumeister, Eds. NY: Russell Sage.

Frederick, S., Loewenstein, G., & O'Donoghue, T. (2002). Time discounting and time preference: A critical review. *Journal of Economic Literature, 40*, 351-401.

Green, L., Fry, A., & Myerson, J. (1994). Discounting of delayed rewards: A life-span comparison," *Psychological Science, 5*, 33-36.

Hinson, J., Jameson, T., & Whitney, P. (2003). Impulsive decision making and working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*, 298-306.

Kassam, K., Gilbert, D., Boston, A., & Wilson, T. (2008). Future anhedonia and time discounting. *Journal of Experimental Social Psychology, 44*, 1533-1537.

Killeen, P.R. (2009). An additive-utility model of delay discounting. *Psychological Review, 116*, 602-619.

Koopmans, T. (1960). Stationary ordinal utility and impatience. *Econometrica, 28*, 87-309.

Laibson, D. (1997). Golden eggs and hyperbolic discounting, *Quarterly Journal of Economics, 112*, 443-477.

Lewis, D. (1983). Survival and identity. In *Philosophical Papers*, Vol. 1, Oxford: Oxford U. Press, 55-77.

Loewenstein, G. (1988). Frames of mind in intertemporal choice. *Management Science, 34*, 200-214.

Loewenstein, G., O'Donoghue, T., & Rabin, M. (2003). Projection bias in predicting future utility. *The Quarterly Journal of Economics, 118*, 1209-1248.

Nichols, S. & Bruno, M. (in press). Intuitions about personal identity: An empirical study. *Philosophical Psychology*.

Parfit, D. (1976). Lewis, Perry and what matters. In *The Identities of persons,* A.O. Rorty, Ed. Berkeley, CA: University of California Press, 91-108.

Parfit, D. (1984). *Reasons and Persons*, Oxford: Oxford U. Press.

Perry, J. (1972). Can the Self Divide? *Journal of Philosophy, 69*, 463-488.

Rips, L.J., Blok, S., & Newman, G. (2006). Tracing the identity of objects. *Psychological Review, 113*, 1-30.

Schwarz, N. (2004). Meta-cognitive experiences in consumer judgment and decision making. *Journal of Consumer Psychology, 14*, 332–348.

Simpson, C. & Vuchinich, R. (2000). Reliability of a measure of temporal discounting. *Psychological Record, 50*, 3-16.

Strathman, A., Gleicher, F., Boninger, D., & Edwards, C. S. (1994). Consideration of future consequences: Weighing immediate and distant outcomes of behavior. *Journal of Personality and Social Psychology, 66*, 742-752.

Zauberman, G., Kim, B.K., Malkoc, S., & Bettman, J. (2009). Discounting time and time discounting: Subjective time perception and intertemporal preferences. *Journal of Marketing Research, 46*, 543-556.

WIN!! vs. win:
Impact of "Outcome" Salience on Illusion of Control

Stefania Mereu (smereu@illinois.edu)
Psychology Department, 603 E. Daniel Street
Champaign, IL 61820 USA

Alejandro Lleras (AlejandroLleras@gmail.com)
Psychology Department, 603 E. Daniel Street
Champaign, IL 61820 USA

Abstract

In contingency judgment tasks (CJT) people typically overestimate their control over an outcome. We hypothesized that this outcome density effect (a type of illusion of control) may be due to an attentional bias toward positive outcomes, which may lead one to ignore negative outcomes and thus to underestimate their occurrence. In order to directly test this hypothesis, we manipulated the outcome's salience in a CJT, inducing participants to focus on either positive or negative outcomes. Results showed that enhancing the salience of positive outcomes (wins) enhanced participant's judgment of control more so than enhancing than of negative outcomes (losses). Moreover, when positive outcomes were salient, participants overestimated the amount of money they had earned during the experiment. In sum, the salience of the "outcome event" affected both judgment of control and memory for positive, more than for negative events, implying that attentional mechanisms may play an important the role in the illusion of control phenomenon.

Keywords: Illusion of Control; Density Outcome Effect; Salience; Attention; Memory; Mood.

Introduction

The desire for control is widespread across both normal (see Keinan, 2002) and psychiatric (e.g. Moulding & Kyrios, 2007) populations, often leading to magical thinking, superstitious behavior and distortion of reality (e.g. Bar-Hillel & Neter, 1996). According to Taylor and Brown (1988), although correlating with psychiatric disorders (e.g. Reuven-Magril, Dar & Liberman, 2008), a moderately amount of positively distorted self-perceptions and expectations about the future might be functional in preserving mental health, through maintaining an adequate self-esteem. An important aspect of any adaptive behavior is the ability to selectively attend to salient or relevant information (Bradley, 2009). In fact, biased attention leads to distorted perception, often observed in major clinical disorders such as, depression (e.g., Leyman et al., 2007) and anxiety (e.g., Bradley et al., 1998). The present study focuses on the role of attentional biases in the establishment of cognitive illusions, specifically, the illusion of control (Langer, 1975).

Jerkins and Ward (1965) observed that in an active contingency judgment task (CJT), where participants had to judge the contingency between their action and an outcome, the perceived control correlates with the desired outcome's density instead of the actual contingency. In an active CJT, observers typically have to perform an action (e.g., pressing a button) to which it may, or may not follow a desirable outcome. After the task they are asked to judge to what extent their action affected the outcome. The key finding is that people tend to base their judgment of control on the frequency of reinforcement instead of on the objective evaluation of the actual contingency (Jerkins & Ward, 1965). In other words, high outcome's density leads to a higher judgment of control, while lower outcome's density leads to an underestimation of control.

According to a study conducted by Alloy and Abramson (1979), only non-depressed individuals show the outcome density effect, while depressed subjects tend to estimate their control more realistically. Alloy and Abramson (1979) argued that the lack of the outcome density effect in depression (depressive realism) indicates that depressed people are "sadder but wiser" than non-depressed people. While non-depressed individuals seem to succumb to positive illusion, depressed people lack this illusion and show a more accurate judgment of the contingency between their actions and external effects. The outcome density effect has been referred to as a type of "illusion of control" (see Alloy & Abramson, 1979). In the illusion of control (Langer, 1975) people overestimate their chance to success, ignoring the objective evaluation of the actual contingency.

Only few studies (e.g., Msetfi et al., 2005) have proposed a link between the lack of illusion of control in depressed individuals and an attentional dysfunction. Msetfi and colleagues (2005) observed that differences between depressed and non-depressed individuals disappear at long inter trial intervals (ITI). They suggested that depressed people might be deficient in exploring all the contextual elements, due to an attentional deficit. This conclusion is supported by studies showing attentional deficits in depression (e.g. Paelecke-Habermann, Pohl & Leplow, 2005). Similarly, Allan, Siegel and Hannah (2007) suggested that differences between depressed and non-depressed people might rely on a change in the decision criterion related to the salience of the outcome (i.e. the one with lower density rate as in the case of low density outcome), instead of a distorted perception of contingency.

Here we suggest an alternative hypothesis, that illusion of control is due to an attentional bias toward positive outcomes, which may lead one to selectively ignore

negative outcomes, thus, to underestimate their occurrence. There is a growing number of studies showing that major depression is characterized by an impairment of selective attention (e.g., Purcell et al., 1997), increased sensitivity to negative reinforcement (Pizzagalli et al., in press) and enhanced brain response to negative feedback (Santesso et al., 2008). Moreover, Nelson and Craighead (1977) showed that depressed individuals recall the frequency of the negative feedback more accurately that non-depressed individuals. An attentional bias toward negative outcomes could enhance the memory for negative feedback and therefore, improve the performance in the judgment task.

In their Experiment 3, Alloy and Abramson (1979) implicitly manipulated attention, associating either positive outcome with a monetary winning or negative outcome with a money loss, separately. Illusion of control was observed only when the positive outcome was associated with a monetary winning. It has now been documented that monetary rewards have strong effects on the attentional system (e.g., Della Libera & Chelazzi 2009), thus it is likely that the value assigned to the outcomes may have modulated the attentional pull of these events. Specifically, the money loss associated with the negative outcome may have encouraged the observer to attend to the negative outcomes, eliminating the bias and therefore, the illusion of control. On the other hand, emphasizing the salience of positive outcomes should enhance the bias, therefore leading to an increase of the illusion.

The goal of the present study was to directly test the hypothesis that attentional mechanisms are involved both in the judgment of control and in the memory for events, in the CJT. We asked observers to estimate their control over an outcome in an active CJT, by pressing one of two buttons in the attempt to maximize their winnings. In the present experiment, although the relative density of the outcome changed (P(O)=.25 or .75), the actual control (ΔP)—defined as the difference between the probability of the outcome given an answer and the probability of the outcome given the other answer—was zero.

We manipulated attention by means of the outcome's salience by having two salience conditions (blocked between subjects): a condition in which the negative outcome was perceptually more salient than the positive outcome, and a condition in which the positive outcome was more salient than the negative outcome. In order to evaluate whether attention also affects the memory representation for winning and losses, we also asked participants to estimate the amount of money they thought they won in the experiment. Predictions are straightforward: if illusion of control is modulated by a natural tendency to neglect negative outcomes, an increase of the negative event's salience should accompany a reduction of the illusion. On the contrary, an enhancement of positive event's salience should enhance the illusion. If the same attentional bias also affects memory for positive and negative events, we also expect salience to affect the perceived money winning or loss.

Method

Participants

Fifty-four females and 43 males (age=21±3) participated in the experiment. All participants had normal or corrected to normal vision, signed an informed consent before the experiment and were paid $8 per hour. Participants assigned to the high reward rate condition were given extra $5 at the end of the experimental session.

Stimuli and Materials

Stimuli (Figure 1) were presented on a 21-inch monitor running at 85Hz. All stimuli were white unless otherwise specified, and they were displayed on a black background. All writings were typed in white, Helvetica font. The fixation point appeared in the center of the monitor, and consisted of a cross sign subtending 0.6° visual angle. The "get ready" message appeared at fixation and occupied 1° visual angle vertically and 16° horizontally. The countdown numbers subtended about 1°x2° visual angle and replaced fixation, when displayed.

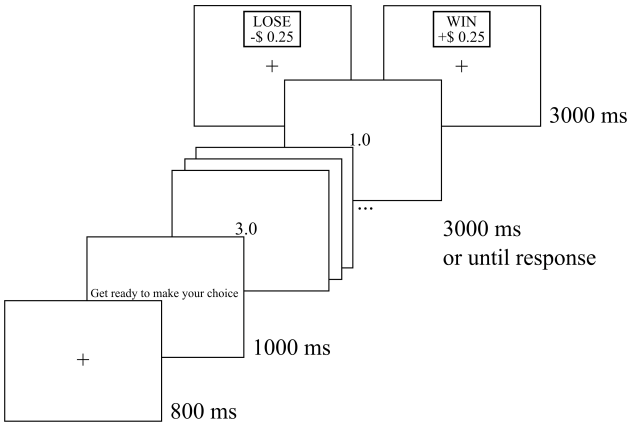


Figure 1. Stimuli used in the contingency judgment task.

The outcome display consisted in a box (6.4° x4.4° visual angle) located 6° above the fixation point. One of two messages could be displayed inside the box: the word "WIN" presented above the amount of money actually won on that trial, or the word "LOSE" above the amount of money lost on the trial (see Figure 1). On salient trials, the outcome boxes were very similar to the boxes on regular trials, with the only difference that the inner part was red and the outline was yellow; the font size was also increased.

Visual Analogue Mood Scales (VAMS) We used the VAMS to assess the mood of participants in the experiment. In this procedure, six positive adjectives are presented. The bottom of the page contained the question: "How do you feel right now?". Below the question, the adjectives: "Pleased", "Cheerful", "Optimistic", "Contented", "Satisfied" and "Happy" are displayed. Underneath each adjective there was a 100mm long line. Participants were verbally instructed to draw a mark along the line, at the point that best described their feelings, in that particular

moment. Score varies from 0 to 100, with "Not at all" at the left-most position in the line, and "Very much" at the right-most position. Intermediate values correspond to intermediate states.

General procedure

Upon arrival to the lab, participants were asked to sign an informed consent and to fill out the first VAMS. Through the whole exchange, the experimenter acted very friendly, using a cheerful attitude and trying to set up a positive interaction. Participants then watched a 5 minutes long, pleasant movie after receiving a treat (i.e., a candy). After the video, they were asked to recall a happy memory. The goal of these manipulations was to improve participants' mood (e.g. Rottenberg, Raye & Gross, 2007), because it has been shown that positive mood enhances the illusion of control (Alloy, Abramson & Viscusi, 1981). Importantly, it was not our goal to study the effects of mood on the illusion of control, but simply to maximize the magnitude of the effect, so that we could in turn study modulations of this magnitude by our attention manipulations. Once these manipulations were completed, participants filled out a second VAMS and then, performed the CJT. A subset of participants also completed a third VAMS after the CJT.

Procedure

Participants sat in a comfortable chair, positioned at 56cm from the monitor and located in a dim-lighted, thermo-regulated room. Given that realistic circumstances enhance illusion of control (Matute, 1996), participants were told that they had the actual opportunity to win money depending on their button pressing, and they were asked to make an effort in order to figure out the best strategy to win more money. They were suggested to explore the use of the two buttons as much as possible. This was meant to discourage participants from adopting the strategy of pressing only one of the two buttons. Such strategy would not be desirable in this type of task because it would inflate the participant's perceived control. Even if the instruction were clear and effective (only two subjects pressed the same button throughout the whole task), uncontrolled imbalance was taken into account.

Each trial (Figure 1) begun with a fixation point, which participants were instructed to look at. One second later, a "Get Ready" message appeared (also 800 ms in duration). Following this message, participants were given 3 seconds to make a choice between two keyboard buttons ("c" or "n"). During this time, there was a numerical countdown display on the monitor, with the numbers 10 counting down towards one, three times in a row. The countdown stopped after 3 repetitions or upon the subject's response. This procedure had the purpose to maximize the illusion of control, which has been shown to increase using stopping devices (Ladouceur, & Savigny, 2005). Participants were simply asked to press a button during the countdown.

After the response, a box appeared for 3000 ms to tell participants whether they won or lost $0.25. If no response was detected, a warning message appeared and a new trial began.

After the task participants were asked to judge both, how much control they had over the outcome on a scale from 0 (no control) and 100 (complete control). Intermediate values corresponded to intermediate judgments of control. In addition, they had to indicate the total amount of dollars they believed to have earned throughout the whole experiment.

Design

Participants were randomly assigned to one of the six possible conditions (each made-up of 40 trials). There were two levels of reward frequency: low reward rate, in which the relative density of the positive outcome P(W) was 0.25—i.e., the negative outcome occurred 75% of the trials—and a high reward rate, in which the relative density of the positive outcome P(W) was 0.75—i.e., the negative outcome occurred 25% of the trials. One half of the participants were assigned the low reward rate condition and the other half was assigned the high reward rate condition. Within each group, one third of the subject were assigned to the control condition (identical salience for win and loss feedback messages), one third were assigned to the condition in which the negative outcomes (the loss events) were salient (the loss salient condition) and the remaining third received the one in which the positive outcomes (the win events) were salient (the reward salient condition).

Independently of the reward rate, the CJT gave participants no control ($\Delta P=0$). That is, the reward rate varied independently from which button the participant decided to press.

Data analysis

Six people were excluded from the analysis because of missing data; one was excluded for participating in the experiment twice and another one was excluded for providing an unrealistic answer about the winning's amount.

In order to evaluate the effectiveness of our mood induction procedure, and to rule out the possibility that our results could be caused by mood differences, a mixed ANOVA was carried out on the VAMS scores (before mood induction, after mood induction) with reward rate (low, high) and salience (control, loss salient, reward salient) as factors.

In order to evaluate the effect of attention on the outcome density effect, judgments of control and win were analyzed using a between-subjects ANOVA with reward rate (low, high) and salience (control, loss salient, reward salient) as factors. VAMS scores collected after mood inductions were included as covariate.

Judgments of control were corrected for the actual amount of control that participants experienced during the task, by means of the formula adapted from Allan (1980):

$$\Delta P = P(W \mid C) - P(W \mid \sim C)$$

where P(W|C) is the relative probability to win by pressing one button ("c") and P(W|~C) is the relative probability to win by pressing the other button ("n"). Judgments of control were also analyzed using a series of t-tests, in order to evaluate whether they differed from zero (correct estimation of control).

Judgments of winnings were corrected for the actual amount of money won during the CJT, so that positive values correspond to an overestimation of winnings and negative values correspond to an underestimation of the winnings.

Results

Mood The 3 (mood; before mood induction, after mood induction, after task) by 2 (reward rate; low, high) by 3 (salience; control, loss salient, reward salient) ANOVA on VAMS scores for happiness showed a significant effect of mood induction ($F_{2,132}$=10.3; p<.001). Happiness after mood induction (mean=82 ±16.15) increased by 12%, when compared to the first assessment (mean=70 ±19.29; p<.001) and decreased again after the experiment (p<.001). More important, there was a significant interaction ($F_{2,132}$=4.8; p<.01). Post hoc tests showed that, after the CJT, the mood in the high reward rate groups was higher than the one in the low reward rate (p<.001). Mood decreased by 26% after the CJT in the low reward rates groups (p<.01), while in the high reward rate condition it remained higher than the first assessment (p<.001) but it did not change with respect to the second assessment (p>.05).

Judgment of Control as a function of reward rate and Outcome Salience The 2 (reward rate; low, high) by 3 (salience; control, loss salient, reward salient) ANCOVA on judgments of control (corrected by the actual control, ΔP) showed a significant effect of the reward rate ($F_{1,90}$=25.23; p<.001). Participants assigned to the high reward condition (mean = 22.19) reported higher perceived control than the ones assigned to the low reward rate condition (mean = -4.19). The interaction between reward rate and salience showed a tendency towards significance ($F_{2,90}$=2.91; p=.06). In order to better understand this result, we ran an ANCOVA using reward rate (low, high) and only two levels of reward salience (control, loss salient). This analysis only showed an effect of the reward rate ($F_{1,58}$=8.32; p<.01). Participants who performed the high reward condition (mean = 15.41) reported higher perceived control than the ones assigned to the low reward rate condition (mean=-2.97). The interaction between reward rate and reward salience was not significant (F<1). A second analysis focused on the reward-salient results: we ran an ANCOVA with factors reward rate (low, high) and two levels of salience (control, reward salient). This analysis showed an effect of the reward rate ($F_{1,64}$=22.9; p<.001), with groups assigned to the high reward condition (mean = 23.66) reporting higher perceived control than the ones assigned to the low reward rate condition (mean = -5.08). More importantly, the interaction between reward rate and reward salience was significant, $F_{2,64}$=6; p<.05 (see Figure 2). Further post hoc analyses revealed that this significant interaction was reflecting the fact that the group assigned to the [high reward rate, reward salient] condition reported higher perceived control ($F_{2,31}$=7.1; p<.01) than the other groups.

Further analysis on the judgment of control, using Student's t-test, showed that none of ratings of control for the groups in low reward conditions differed than zero (all ps>.05). Moreover, judgments of control expressed by participants assigned to the high reward rate condition were significantly higher than zero only when the positive outcome (p<.001) was salient; when none of the outcomes was salient there was a tendency to significance (p=.06) while when the negative outcome was salient the judgments of control were no significantly higher that zero (p=.08).

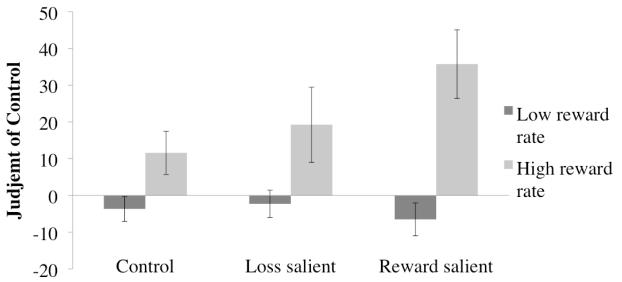


Figure 2. Reported judgment of control corrected by the actual control experienced during the task.

Winnings results The 2 (reward rate; low, high) by 3 (salience; control, loss salient, reward salient) ANCOVA on the difference between the reported and the actual winnings showed a significant effect of the reward rate ($F_{1,90}$=41.1; p<.001).

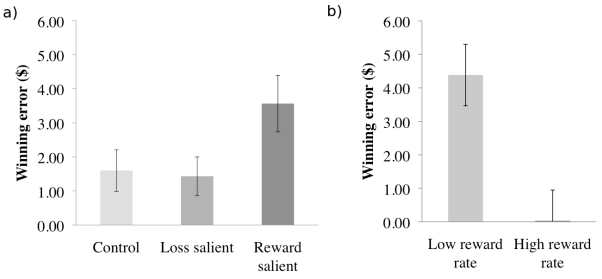


Figure 3. Errors in perceived money won, corrected for the actual winning displayed by a) salience and b) reward rate. Positive values indicate overestimation and negative values indicate underestimation of winnings.

The groups assigned to the low reward condition (mean = 4.39; corrected for the actual winning) overestimated their winnings more than the ones assigned to the high reward rate condition (mean =0.03). The analysis also showed an effect of salience ($F_{2,90}$=3.04; p<.05; Figure 3), with overall larger overestimation errors in the reward salient condition (mean = 3.5) compared to control and loss salient conditions (mean = 1.6, mean = 1.4, respectively). The interaction between reward rate and salience was not significant (F<1). Post hoc tests showed that the group assigned to the reward

salient condition overestimated their winnings more than the ones assigned to the loss salient condition ($p<.05$); moreover, the difference between the control condition and the reward salient condition also tended towards significance ($p=.06$), but failed to reach it due to relatively larger variability in that condition, compared to the loss salient condition.

Discussion

The aim of this study was to evaluate the role of attention in the outcome density effect. We hypothesized that, if the illusion of control is caused by an attentional bias toward the positive outcome, increasing the negative outcome's salience should reduce the illusion. On the other hand, salient positive outcomes were expected to enhance the illusion.

The results partially confirmed our hypotheses: judgments of control were indeed inflated when the positive outcomes were made more salient; yet, judgments of control were unaffected by the salience of the loss outcomes.

Overall, our results replicated the traditional outcome density effect (Jerkins & Ward, 1965). Judgments of control of participants who were often rewarded were higher than those of participants who received fewer rewards. Furthermore, participants in the high reward rate condition tended to overestimate the control they exerted over the outcome ($p=.06$).

The enhancement of perceived control when the positive outcome occurs often and when it is more salient than the negative one, is in accordance with the hypotheses that attention modulates the illusion of control: the increased salience of positive outcomes likely attracted attention towards those events, enhancing a baseline bias towards attending to those events in the first place, increasing the illusion of control. That said, it is also important to note that equivalent salience manipulations on the feedback of "loss" events did not significantly modulate neither the illusion of control nor the perceived winnings in the task.

Our mood induction procedure was successful in enhancing the general mood in participants. Importantly, differences in the mood of participants across groups were not responsible for the differences observed in perceived control or winnings, since no difference in participant's mood was observed across conditions.

There may be several reasons why our salience manipulation failed to influence the illusion of control in the loss salient condition. It is possible that the specific colors we chose in our manipulation may have interacted differently with the perceptions of gain and loss. There is a lifetime associations between yellow and cautious behavior and red with maximum levels of hazard (see Williams and Noyes, 2007; for a review). If the observers interpreted the color of the outcome as a warning clue, it is possible that this encouraged them to abandon a risk taking strategy, which is common in gamblers and known to be correlated with illusion of control (Fenton-O'Creevy et al., 2003). Moreover, results showing that red color facilitates cognitive tasks in which negative stimuli are involved (Mehta, & Zhu, 2009) suggest that positive and negative salient outcomes may have been processed differently. Specifically, red may have increased accuracy in remembering the occurrence of the negative outcomes only. That said, this would not explain why the illusion grew in size in the reward condition. Lastly, it is well known that gains and losses are perceived asymmetrically to begin with (Kahneman & Tversky, 1979). As such, it is possible that loss aversion may have been at play in our experiment, making participants in the loss salient conditions overall more cautious than in the reward salient condition, or turned them into more "objective" assessors of the events (much like in the "depressed realism" effect). In contrast, participants in the reward salient condition may have been more prone to get excited about their winnings, inducing something like a positive-mood amplification of the illusion of control effect. Overall, the asymmetry in the effects of event salience on perceived control and perceived winnings has strong implications in terms of understanding some aspects of gambling behavior: in most gambling situations, loss events have little salience, whereas win events tend to be very salient. This may be contributing to increase levels of illusion of control in gambling scenarios (like slot machines), and further, our results suggest that simply increasing the salience of the loss events (making them as bright and noisy as win events) may be insufficient to counteract the increased illusion of control arising from salient win events.

A reverse outcome density effect was observed in the winning ratings. On the one hand, participants generally overestimated the amount of money won in the experiment. On the other hand, the biggest mistake in overestimating the amount of money was observed in the low reward rate condition (i.e., when the win events happened more rarely). This result, although surprising, could be due to a bias induced by the experimental procedure: subjects signed an informed consent in which they were promised a fixed amount for the experimental session, plus the possibility to increase their earnings for the day. This manipulation is intended to increase the illusion (Matute, 1996), but could have caused participants to be skeptical on the actual possibility to lose money during the experiment, encouraging them not to state a money loss.

A particularly striking result was the salience effect observed on the winnings recall. When the positive outcome was salient, participants overestimated the winnings more than in either of the other two conditions (salient loss outcome and control conditions). This result was independent of the reward rate and, although preliminary, might also be potentially relevant to gambling. Winnings are often exaggerated and amplified by means of lights, sound and colors and the saliency is not necessarily commensurate to the actual winning. This may not only increase the gambler's tendency to overestimate its own control over the situation but also to remember inflated winnings throughout the gambling experience.

In sum, in line with our initial hypothesis, attentional biases seem to partly contribute to the illusion of control phenomenon. Salience of the outcome, in fact, modulated both the contingency judgment and the memory for winnings. These results are promising and have potentially important implications for the understanding of cognitive mechanisms underlying gambling behaviors.

Acknowledgments

The authors wish to thank Matthew Harding for his permission to use his video in this experiment. The research of S. Mereu was supported by Regione Sardegna (T2-MAB-A2008-608).

References

Allan, L. G. (1980). A note on measurement of contingency between two binary variables in judgment tasks. *Bulletin of the Psychonomic Society, 15*, 147–149.

Allan, L. G., Siegel, S., & Hannah, S. (2007). The sad truth about depressive realism. *Quarterly Journal of Experimental Psychology, 60*, 482–495.

Alloy, L. B., & Abramson, L. Y. (1979). Judgment of contingency in depressed and nondepressed students: Sadder but wiser? *Journal of Experimental Psychology: General, 108*, 441–485.

Alloy, L. B., Abramson, L. Y, & Viscusi, D. (1981). Induced mood and the illusion of control. *Journal of Personality and Social Psychology, 41*, 1129–1140.

Bar-Hillel, M., & Neter, E. (1996). Why are people reluctant to exchange lottery tickets? *Journal of Personality and Social Psychology, 70*(1), 17-27.

Bradley, B. P. Mogg, K., Falla, S. J., & Hamilton L.R. (1998). Attentional bias for threatening facial expressions in anxiety: manipulation of stimulus duration. *Cognition and Emotion, 12*, 737–753.

Bradley, M. M. (2009). Natural selective attention: orienting and emotion. *Psychophysiology, 46*(1), 1–11.

Della Libera, C., & Chelazzi, L. (2009) Learning to attend and to ignore is a matter of gains and losses. *Psychological Science, 20*(6), 778–784.

Fenton-O'Creevy, M., Nicholson, N., & Soane, E., Willman, P. (2003). Trading on illusions: Unrealistic perceptions of control and trading performance. *Journal of Occupational and Organisational Psychology, 76*, 53–68.

Fisher, S., & Ledwith, M. (1984).The perception of control in loud noise. *Perception. 13*(6) 709–718.

Jenkins, H., & Ward, W. C. (1965). Judgment of contingency between responses and outcomes. *Psychological Monographs: General & Applied, 79*, SUPPL 1, 1–17.

Kahneman, D., & Tversky., A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica, 47*(2), 263–291.

Keinan, G. (2002). The effects of stress and desire for control on superstitious behavior. *Personality and Social Psychology Bulletin, 28*, 102–108.

Ladouceur, R., & Sévigny S. (2005). Structural characteristics of video lotteries: effects of a stopping device on illusion of control and gambling persistence. *Journal of Gambling Studies, 21*(2),117–31.

Langer, E. J. (1975). The illusion of control. *Journal of Personality and Social Psychology, 32*, 311–328.

Leyman, L., De Raedt, R., Schacht, R., & Koster, E. H. (2007). Attentional biases for angry faces in unipolar depression. *Psychological Medicine, 37*(3), 393–402.

Matute, H. (1996). Illusion of control: Detecting response-outcome independence in analytic but not naturalistic conditions. *Psychological Science, 7*, 289–293.

Moulding, R., & Kyrios, M. (2007). Desire for Control, Sense of Control and Obsessive-Compulsive Symptoms. *Cognitive Therapy and Research, 31*,759–772.

Msetfi, R. M., Murphy, R. A., Simpson, J., & Kornbrot, D. E. (2005). Depressive realism and outcome density bias in contingency judgements: The effect of the context and inter-trial interval. *Journal of Experimental Psychology: General, 134*, 10–22.

Nelson, R. E., & Craighead, W. E. (1977). Selective recall of positive and negative feedback, self-control behaviors, and depression. *Journal of Abnormal Psychology, 86*(4):379-88.

Paelecke-Habermann, Y., Pohl, J., & Leplow, B. (2005). Attention and executive functions in remitted major depression patients. *Journal of Affective Disorders; 89*(1–3):125-35.

Pizzagalli, D. A., Dillon, D. G., Bogdan, R., & Holmes, A. J. Reward and punishment processing in the human brain: Clues from affective neuroscience and implications for depression research. In: Vartanian O, Mandel D, editors. *Neuroscience of decision making*. New York, NY: Psychology Press (in press).

Purcell, R., Maruff, P., Kyrios, M., & Pantelis, C. (1997). Neuropsychological function in young patients with unipolar depression. *Psychological Medicine, 27*, 1277–1285.

Reuven-Magril, O., Dar, R., & Liberman, N. (2008). Illusion of control and behavioral control attempts in obsessive-compulsive disorder. *Journal of Abnormal Psychology*, *117*(2):334–41.

Rottenberg, J., Raye, R.D., & Gross, J.J. (2007). Emotion elicitation using films. In J.A. Coan & J.J.B. Allen (Eds.), *Handbook of Emotion Elicitation and Assessment* (pp. 9-28). Oxford, UK: Oxford University Press.

Santesso, D. L., Steele, K. T., Bogdan, R., Holmes, A. J., Deveney, C. M., Meites, T. M., & Pizzagalli, D. A. (2008). Enhanced negative feedback responses in remitted depression. *Neuroreport, 19*(10):1045–8.

Taylor, S. E., & Brown, J. D. (1988). Illusion and well-being: A social psychological perspective on mental health. *Psychological Bulletin, 103*, 193–210.

Williams, D. J., & Noyes, J. M. (2007). How does our perception of risk influence decision-making? Implications for the design of risk information. *Theoretical Issues in Ergonomics Science, 8*(1), 1–35.

Signs of Non-Linearity in Base-Rate Neglect

Christopher D. Erb (erbcd@mail.uc.edu)
University of Cincinnati, Department of Psychology,
Dyer Hall, Cincinnati, OH 45221-0376, USA

Heidi Kloos (heidi.kloos@uc.edu)
University of Cincinnati, Department of Psychology,
230 Dyer Hall, Cincinnati, OH 45221-0376, USA

Abstract

Base-rate neglect, the tendency of adults to ignore the prior probability of an event, has been well-studied over the past decades. However, the evidence for base-rate neglect and its theoretical implications are still debated. We argue that such lack of agreement comes from the mistaken assumption that performance unequivocally reflects cognitive processes. We adopt a different viewpoint, namely that performance reflects existing constraints in the person-task relation. To test whether this viewpoint is appropriate for performance in base-rate problems we manipulated the constraints available in the task's response options. With a highly constraining response mode adults are expected to exhibit the classic base-rate neglect, with little variability in their performance as procedural factors are manipulated. However, with a less constraining response mode performance is expected to be more variable and more susceptible to subtle changes in the task procedure. Results support this view, demonstrating non-linear context effects in decision making.

Keywords: rationality; adult reasoning; context effects.

Introduction

Consider the following problem:

> "In a study 1000 people were tested. Among the participants there were 5 engineers and 995 lawyers. Jack is a randomly chosen participant of this study. Jack is 36 years old. He is not married and is somewhat introverted. He likes to spend his free time reading science fiction and writing computer programs. What is most likely? (a) Jack is a lawyer (b) Jack is an engineer."

Based on the description of Jack, you may be tempted to think that he is an engineer; after all, he is introverted, he enjoys reading science fiction, and he writes computer programs. Indeed, a vast majority of adults would agree with you (De Neys & Glumicic, 2008). However, the statistical information provided in the problem indicates otherwise. Given that Jack was randomly selected from a study consisting of far more lawyers than engineers (995 vs. 5), it follows that Jack is most likely a lawyer.

This type of decision-making problem has a long history in the literature on reasoning, stretching back to the classic studies of Daniel Kahneman and Amos Tversky (1973). Despite nearly four decades of research featuring these base-rate problems, discussions concerning the task are still going strong. Take, for example, the disagreement about the influence of presenting statistical information as frequencies rather than one-case probabilities. Some argue that presenting problems in terms of frequencies has more ecological validity and, therefore, improves performance on the task (Cosmides & Tooby, 1996; Gigerenzer & Hoffrage, 1995; Hoffrage, Gigerenzer, Krauss & Martignon, 2002). Others have suggested that gains in performance are not attributable to the frequency format alone, but to presentation formats that encourage the formation of a set inclusion mental model (Evans, Handley, Perham, Over & Thompson, 2000).

Or consider the discussion about how to characterize the cognitive processes that take place as the task is solved. Is human reasoning subserved by two distinct processes (c.f., Evans, 1984; 2007; Kahneman & Frederick, 2002; Stanovich & West, 2000), one being heuristic while the other is analytic? If so, how do these processes function in relation to one another? Is one of these processes the default, aided by the other process only in the case of conflict? Or is the dual-process approach presenting a false dichotomy altogether?

There is even disagreement about whether base-rate neglect implies a shortcoming of the human mind or a sophisticated adaptation. Some suggest that neglect of base-rate information is an indication of humanity's underlying irrationality (Nisbett & Borgida, 1975), while others argue that the exact same performance reflects adaptive processes that maximize efficiency in decision making (Gigerenzer & Brighton, 2009).

Similar disagreements in how to interpret performance have been documented in virtually all of adult cognition, including memory, attention, decision making, and learning (for a review see Van Orden, Pennington, & Stone, 2001). It is rather the norm than the exception to disagree about which task might best reflect natural reasoning, or how to best characterize underlying cognitive processes. These disagreements are symptomatic of an assumption that performance in a task allows direct inferences about cognitive structures or cognitive processes that are at work (for full arguments, see Kloos & Van Orden, 2009; Van Orden & Kloos, 2003). Only if performance is thought to be transparent to the underlying cognitive architecture can details of the task context be argued about. But this assumption, known also as the 'effect = structure' fallacy, has been shown to be faulty (e.g., Gibbs, 1994; Lakoff, 1987). An alternative is the assumption that performance reflects a unique person-task relation, one that cannot be

reduced to the person (or the task) alone (e.g. Gibson, 1979).

The idea that performance reflects non-reducible person-task units has been formalized in the idea of constraints that reduce degrees of freedom for action (Kloos & Van Orden, in press). If a task context is highly constraining (e.g., there are only two answer options, one of which is understood to be correct), then we expect to see formulaic, uniform performance – as if a stable cognitive structure or process is operating. If, however, a task context is less constraining (e.g., the person is presented with many answer options or believes that there is no right or wrong answer), performance is likely to be affected by idiosyncratic aspects of the person's history, miniscule changes in the procedure and seemingly irrelevant aspects of task instructions or stimuli. The resulting difference in performance does not reflect different cognitive processes but rather a different coupling between the person and the task.

In the current paper we investigate whether the idea of constraints could help shed light on performance in a base-rate neglect task (c.f., Kahneman & Tversky, 1973). Adult participants had to determine the likelihood of a certain event, given base-rate information (the a priori statistical probability of a certain event) and individuating information (the stereotypical probability of the event). The crucial manipulation was in the answer options: Participants were presented either with highly constraining multiple-choice answer options (*multiple-choice condition*), or they were presented with less constraining open-ended answer options (*open-ended condition*). We also manipulated a superficially irrelevant factor, namely the order in which information was presented: In *Order 1*, base-rate information appeared first, before the individuating information; and in *Order 2*, base-rate information appeared second, after the individuating information. If constraints, rather than cognitive structure, decide the performance in a task, then our constraints manipulation should matter. In particular, one would expect performance to be more susceptible to order changes in the less constraining task (open-ended response options) than in the more constraining task (multiple-choice response options). A recall task was added at the end of the experiment that had the same response mode across conditions. This allowed us to determine the degree to which conditions differed in how information was encoded.

Method

Participants

Participants were 24 undergraduate students from the University of Cincinnati (10 men, 14 women) who volunteered their time in return for course credit. The mean age of participants was 19.25 years (*SD* = 3.43). One additional adult was tested and excluded from the final sample due to apparent confusion with task procedures.

Materials and Procedure

Participants were tested individually in a quiet room. The testing session consisted of a decision-making task, an unannounced recall task, and a brief exit survey. As was done in a recent study by De Neys and Glumicic (2008), participants were asked to think aloud while solving the decision-making problems. Participants were introduced to the experiment and the thinking-aloud procedure with the following script used by De Neys and Glumicic:

> "In this experiment we try to find out how people solve everyday reasoning problems. Therefore, we ask you to "think aloud" when you're solving the problems. You should start by reading the complete problem aloud. When you're solving the problem you have to say everything that you're thinking about. All of the inferences you're making, all the comments that you're thinking of, basically everything that is going through your mind, you have to say aloud. You should be talking almost continuously up until the point that you have answered the question. Try to keep thinking aloud the whole time. If you are silent for a while I will ask you to continue to voice your thoughts."

Participants were then given the opportunity to ask questions concerning the thinking aloud procedure. Once the participants were ready to move on, the experimenter began the audio recording and presented the decision-making task. Using the same problem set developed by De Neys and Glumicic (2008), the decision-making task consisted of 18 separate decision-making problems, each containing base-rate information and individuating information. The order of problems was randomized, and the problems were organized in booklet form. The first page of each booklet featured a set of instructions which corresponded to the response mode of the featured problems. Participants in the multiple-choice condition received the following instructions, again adapted from De Neys and Glumicic (2008):

> "In a big research project a number of studies were carried out where short personality descriptions of the participants were made. In every study there were participants from two population groups (for example, carpenters and policemen). In each study one participant was drawn at random from the sample. You'll get to see the personality description of this randomly chosen participant. You will also get to see the number of people in each of the two population groups. Finally, you will be asked to indicate which population group the participant most likely belongs to (policemen, for example) by circling a response."

Only the last sentence was modified for participants in the open-ended condition. It read: "Finally, you will be asked to write the probability that the randomly chosen participant belongs to one of the population groups (policemen, for example)." Participants were asked to read the instructions aloud and were given the opportunity to ask questions regarding the task. Participants then began the decision-making task.

The base-rate information featured a brief description of a sample of 1000 people who were said to have taken part of a study. The sample consisted of two groups of people which were grossly disproportionate in number. For example, base-rate information in one problem stated: 'In a study 1000 people were tested. Among the participants there were 5 sixteen-year olds and 995 fifty-year olds. Ellen is a randomly chosen participant of this study.' Other ratios used were 996 to 4 and 997 to 3.

The individuating information provided a description of an individual who was randomly selected from the featured sample of 1000 people. For instance, given the base-rate example provided above, the individuating information was described as: 'Ellen likes to listen to hip hop and rap music. She enjoys wearing tight shirts and jeans. She's fond of dancing and has a small nose piercing.'

In a third of the trials, base-rate information was pitted against individuating information; the description of the selected person was stereotypic of an individual from the smaller group of the sample (like in the example above). These trials were *incongruent* because the stereotypic associations did not match the most probable option according to the base-rate information.

Alternatively, in another third of the trials, base-rate information matched with the individuating information. That is to say, individuating information was stereotypical of an individual from the larger group of the sample. These trials were considered *congruent*. Finally, the remaining six problems did not feature stereotypes of either population group and, therefore, were considered *neutral* problems.

In order to determine how adults combine base-rate with individuating information, each set of information was followed by a question. In the multiple-choice condition, participants had to select the most probable event out of two options. For example, given the information provide above, the test question was: 'What is most likely? (a) Ellen is sixteen (b) Ellen is fifty'. The answer options (a) and (b) were counterbalanced, such that answer option (a) matched with the base-rate information in half of the trials, while answer option (b) matched with the base-rate information in the other half of the trials.

In the open-ended condition participants were asked to write the probability of the event. For example, the question from the base-rate and individuating information above was: 'What is the probability that Ellen is sixteen?' or 'What is the probability that Ellen is fifty?' Half of the questions inquired about the smaller sub-group of the sample and the other half inquired about the larger sub-group of the sample.

In the open-ended response mode additional instructions were occasionally provided. For example, if participants were unsure of how to express their answers, the experimenter explained that probabilities are typically expressed as fractions, decimals or percentages. If participants wrote responses such as "the probability is high," the experimenter requested a more specific, numerical response. Finally, in instances where participants responded with ranges such as "50-70%," the experimenter instructed participants to provide a more precise response.

At the conclusion of the task the audio recording was stopped. The experimenter then checked the decision-making task to ensure that none of the problems were overlooked. After a short break of about a minute, participants were presented with an unannounced recall task and were instructed to answer the questions to the best of their ability. As was done in the De Neys and Glumicic (2008) study, participants solved four recall questions for each corresponding decision-making problem. The first two questions tested recall of base-rate information, and the second two tested recall of individuating information. All four questions were printed on one page. The pages were once again stapled into a booklet and followed the same order with which the decision-making problems were presented. The following is an example of the recall task:

> One of the problems you just solved concerned Ellen whose description was drawn at random from a sample of fifty-year olds and sixteen-year olds. Try to answer the following questions.
>
> Exactly how many sixteen-year olds were there in the study?
> _____
>
> Exactly how many fifty-year olds were there in the study?
> _____
>
> Circle the correct statement:
> a. Ellen likes to knit
> b. Ellen listens to hip hop
> c. Ellen shops at thrift stores
> d. Ellen drives a truck
>
> Circle the correct statement:
> a. Ellen speaks German
> b. Ellen plays the trumpet
> c. Ellen does not have a job
> d. Ellen has a small nose piercing

After completing the recall task participants were presented with an exit survey that measured the participants' perceptions of the task.

Design

There were two different orders (Order 1: Base-rate First; Order 2: Base-rate Second) and two answer modes (Multiple-choice condition; Open-ended condition). Participants were randomly assigned to one of the four resulting experimental groups. Each participant solved six incongruent problems, six congruent problems, and six neutral problems. Recall was identical across groups.

Results

Our first analysis pertains to participants' performance in the multiple-choice condition. It was scored according to whether the normatively correct option was selected (i.e., the answer option that corresponded to the largest

population group). Responses were collapsed across trials within problem type (incongruent, congruent, neutral), yielding three proportion-correct scores for each participant. Figure 1A displays the means of these scores for the multiple-choice condition as a function of problem type and order. A mixed-design 2 x 3 ANOVA (with problem type as the within-subject factor and order as the between-subject factor) revealed a significant effect of problem type, $F(2, 20) = 101.42$, $p < .001$, but no significant effects of order or order interaction, ps $> .4$. As expected, below-chance performance was obtained for the incongruent problems ($M = 0.19$, $SE = 0.05$), while performance was at ceiling (or above chance) for congruent problems ($M = 1.00$) and neutral problems ($M = 0.83$, $SE = 0.11$). The order in which information was presented had no effect on performance in this condition.

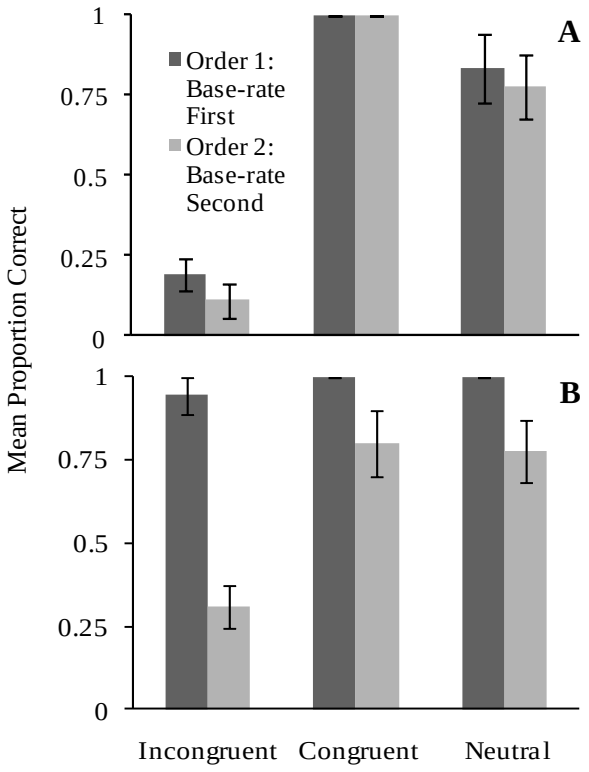


Figure 1: Mean proportion of correctly answered problems as a function of problem type and order. A: multiple-choice condition. B: open-ended condition. Error bars display standard errors.

A very different picture emerges when adults were given a continuum of response options (opened-ended condition). Responses to the prompt for each trial were first scored to match the multiple-choice scoring system. Probabilities below 50% were scored as correct for questions that pertained to providing the probability that the individual is a member of the smaller population group. Alternatively, probabilities above 50% were scored as correct for questions that pertained to providing the probability that the individual is a member of the larger population group. Responses of 50% were not included in the following analysis. This resulted in the exclusion of three of 108 responses in Order 1 and six of 108 in Order 2.

Figure 1B displays the mean proportion of correct responses in the open-ended condition as a function of problem type and order. A mixed-design 2 x 3 ANOVA (with problem type as the within-subject factor and order as the between-subject factor) revealed not only the expected significant effect of problem type, $F(2, 20) = 10.04$, $p < .001$, but also a significant effect of order, $F(1, 10) = 52.90$, $p < .001$, and a significant interaction, $F(2, 20) = 6.30$, $p < .01$. Problem type affected performance only in Order 2, $F(2, 10) = 9.05$, $p < .01$, with below-chance performance on incongruent problems ($M = 0.31$, $SE = 0.06$), and above-chance performance on congruent problems ($M = 0.80$, $SE = 0.10$) and neutral problems ($M = 0.78$, $SE = 0.09$). In Order 1, however, problem type did not affect performance, $F < 1.0$, $p < .4$, with participants performing at or near ceiling on all problem types ($M = .98$, $SE = .02$).

One critique of the above analysis is that the assumption of homogeneity of variance was not met across problem types. A Levene's test of equality of error variances revealed significant difference in variance for the congruent and neutral problem types (ps $< .01$), undermining the results of the parametric tests for these problem types. For this reason, we focus only on the incongruent problem type in the next analysis. Recall that this problem type is the more relevant problem type in the base-rate literature because it demonstrates the base-rate neglect. A 2 x 2 between-subjects ANOVA, with response mode (multiple-choice; open-ended) and order (Order 1, Order 2) as between-subject factors, replicates the results of our previous analyses. It revealed a significant effect of response mode, $F(1, 20) = 69.45$, $p < .001$, a significant effect of order, $F(1, 20) = 39.53$, $p < .001$, and a significant interaction effect, $F(1, 20) = 23.28$, $p < .001$.

To account for performance in the open-ended condition on a continuum, and thus to get a more accurate sense of the data, we computed the distance of responses from the normatively correct probability. For example, if a participant responded with "30%" when the normatively correct response 0.5% or lower, the resulting score would be 29.5%. These scores were once again collapsed across trials within a problem type, yielding three mean distance scores for each participant. Figure 2 shows the mean scores as a function of problem type and order.

A mixed-design 2 x 3 ANOVA (with problem type as the within-subject factor and order as the between-subject factor) revealed a significant effect of problem type, $F(2, 20) = 11.17$, $p < .001$, a significant effect of order, $F(1, 10) = 33.41$, $p < .001$, and a marginally significant interaction, $F(2, 20) = 3.04$, $p < .07$[1]. Once again, a significant effect of

[1] The interaction might not have reached significance due to unequal variances between the two orders, found for each of the problem types (Levene's Test: Fs$(1,10) > 11.8$, ps $< .01$).

problem type was found for Order 2, $F(2, 10) = 8.81$, $p < .01$, but not for Order 1 ($p > .13$).

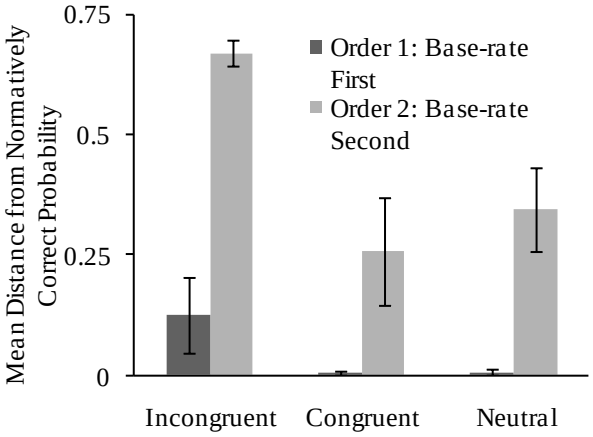


Figure 2: Mean distance from normatively correct probability as a function of problem type and order in the open-ended condition. Error bars display standard errors.

Finally, performance in the open-ended condition was scored in a third way, this time according to whether the response violated the rules of normative probability. For example, if the base-rate information listed a ratio of 3 to 997, probability judgments above 0.3% were scored as incorrect (assuming the question pertained to providing the probability that the individual is a member of the smaller population group). A mixed-design 2 x 3 ANOVA (with problem type as the within-subject factor and order as the between-subject factor) revealed a significant effect of order $F(1, 10) = 26.35$, $p < .001$, no significant effect of problem type, $p > .4$, and no significant interaction, $p > .4$. As Figure 3 illustrates, average performance across problem types was higher for Order 1 ($M = .81$, $SE = .12$) than for Order 2 ($M = .06$, $SE = .12$).

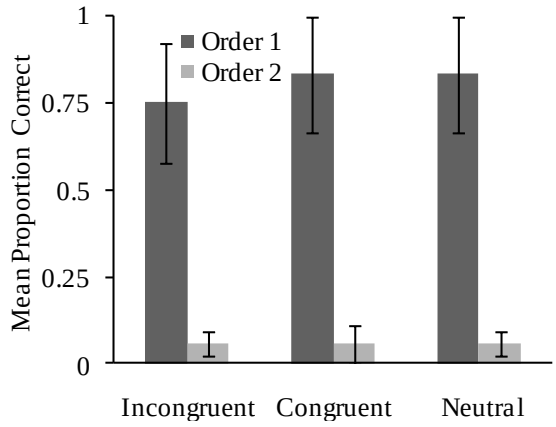


Figure 3: Mean proportion of normatively correct answers as a function of problem type and order (Order 1: Base-rate First, Order 2: Base-rate Second) in the open-ended condition. Error bars display standard error.

Thus far we have shown that the pattern of responses on the decision-making task varied with response mode (multiple-choice vs. open-ended). In the multiple-choice condition order had no effect on performance. But in the open-ended condition, no matter how data was scored, order had a highly significant impact.

One could argue that the difference between conditions is spurious, due to perhaps extraneous factors pertaining to small sample size. Our analysis of participants' base-rate recall provides reason to doubt these possible objections. Bear in mind that recall took place at the end of the experimental session, and the task employed the same response mode for all participants. Thus, if the effect of response mode in base-rate problems was spurious due to small sample size, then we would expect to see differences among conditions in the recall task as well.

Performance on recall of the base-rate information was scored according to whether participants correctly identified the relative size of each group (i.e., which group was larger and which group was smaller). A 2 x 2 x 3 mixed-design ANOVA was conducted, with condition and order as the between-subject factors and problem type as the within-subject factor. Importantly, there was no significant difference and no significant interaction (Fs < 2.47, ps > .13), with above-chance performance for each group (assuming a chance probability of 0.5), single-sample ts > 2.2, ps < 0.05. Figure 4 provides the individual means for response mode and problem type, collapsed across order.

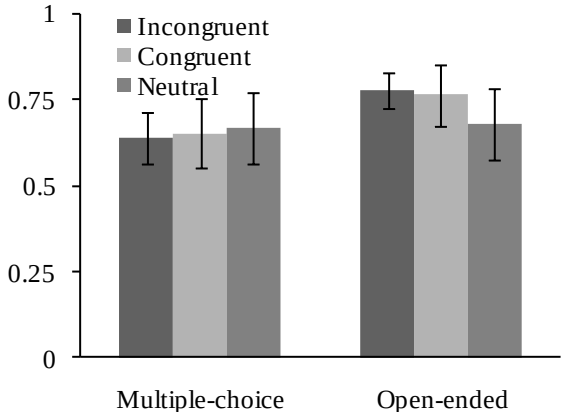


Figure 4: Mean proportion of correct recall as a function of response mode and problem type. Error bars display standard error.

Summary and Discussion

A commonly used base-rate problem was adapted in the current experiment to manipulate the constraints of the task context. Adults participated in one of two conditions that differed only in whether the base-rate problems had a constraining multiple-choice response mode, or a less constraining, opened-ended response mode. Patterns of performance across base-rate problems differed markedly as a function of our manipulation.

In the multiple-choice response mode participants demonstrated the classical base-rate neglect without being affected by superficial changes in the order in which the information was presented to them. Conversely, in the open-ended response mode participants neglected base-rate information only in one of the order conditions, when base-rate information was presented after individuating information (Order 2). In the reverse order, when base-rate information was presented before individuating information (Order 1), participants took base-rate information into account.

Note that Order 1 is the common way in which information was presented to participants in previous research (e.g., De Neys & Glumicic, 2008; Kahneman & Tversky, 1973). Accordingly, we did indeed replicate the previous findings when the multiple-choice response mode was used. But when the response mode was less constraining, the superficial changes in order made a difference in performance. Performance in the recall task provides reason to doubt the possibility that these differences are spurious effects of some sort. Participants in all groups performed above chance on the recall task, independently of how the information was presented in the decision-making problems.

The results of the present investigation underscore the idea that performance cannot be uniquely attributed to cognitive structures or processes. Any plausible cognitive structure that could be responsible for the current findings would be post-hoc and rather complex, given that even irrelevant changes in order affected performance. A constraints view, in contrast, could readily explain our results. It predicts, a priori, that the tightening of degrees of freedom cuts down on idiosyncratic variability in performance and the impact of seemingly superficial factors. Our findings suggest that adults are neither rational nor irrational reasoners. Instead, their performance reflects a coupling with the task, and thus says as much about the task as about the reasoner.

Acknowledgments

The authors thank Sue Collins for her help with scoring the data. Writing of the manuscript was supported by a grant from the National Science Foundation (DRL # 723638) to Heidi Kloos.

References

Cosmides, L., & Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition*, *58*(1), 1.

De Neys, W., & Glumicic, T. (2008). Conflict monitoring in dual process theories of thinking. *Cognition*, *106*(3), 1248-1299.

Evans, J. St. B. T. (1984). Heuristic and analytic processes in reasoning. *British Journal of Psychology*, *75*(4), 451-468.

Evans, J. St. B. T. (2007). On the resolution of conflict in dual process theories of reasoning. *Thinking & Reasoning*, *13*(4), 321-339.

Evans, J. St. B. T., Handley, S. J., Perham, N., Over, D. E., & Thompson, V. A. (2000). Frequency versus probability formats in statistical word problems. *Cognition*, *77*(3), 197-213.

Gibbs, R. W. (1994). *The poetics of mind: Figurative thought, language, and understanding.* New York: Cambridge University Press.

Gibson, J. (1979). *The ecological approach to visual perception.* Dallas: Houghton Mifflin.

Gigerenzer, G., & Brighton, H. (2009). Homo heuristicus: why biased minds make better inferences. *Topics in Cognitive Science*, *1*(1), 107-143.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction. *Psychological Review*, *102*(4), 684-704.

Hoffrage, U., Gigerenzer, G., Krauss, S., & Martignon, L. (2002). Representation facilitates reasoning: what natural frequencies are and what they are not. *Cognition*, *84*(3), 343-352.

Kahneman, D., & Frederick, S. (2002). Representativeness revisted: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment*. U.S.A.: Cambridge University Press

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, *80*(4), 237-251.

Kloos, H. & Van Orden, G. C. (2009). Soft-assembled mechanisms for the unified theory. In J.P. Spencer, M. Thomas, & J. McClelland (Eds.), *Toward a New Grand Theory of Development? Connectionism and Dynamics Systems Theory Reconsidered*. Oxford: Oxford University Press.

Kloos, H., & Van Orden, G. C. (in press). Voluntary performance on cognitive and motor tasks. *Mind and Matter*.

Lakoff, G. (1987). *Women, fire, and dangerous things: What categories reveal about the mind*. Chicago: University of Chicago Press.

Nisbett, R. E., & Borgida, E. (1975). Attribution and the psychology of prediction. *Journal of Personality and Social Psychology*, *32*(5), 932-943.

Stanovich, K. E., & West, F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences*, *23*(5), 645-665.

Van Orden, G. C., Pennington, B. F., & Stone, G. O. (2001). What do double dissociations prove? *Cognitive Science, 25*, 111–172.

Van Orden, G. C., & Kloos, H. (2003). The module mistake. *Cortex, 39*, 164–166.

Mathematically Modeling Anchoring Effects

Jessica M. Choplin (jchoplin@depaul.edu)
DePaul University Department of Psychology
2219 North Kenmore Avenue
Chicago, IL 60614

Mark W. Tawney (mtawney@iosolutions.org)
Industrial/Organizational Solutions
1127 S Mannheim Road
Westchester, IL 60154

Abstract

This article proposes a method by which anchoring effects can be mathematically modeled. Anchoring effects are a type of assimilation effect; so this article proposes using Anderson's (1965; 1981) integration model to model anchoring effects, as it is typically used to model other assimilation effects. The difficulty in using the integration model is that doing so requires that the modeler knows or is able to estimate participants' unbiased estimates (i.e., what their estimates would have been had they never seen the anchor) and this information is not available from conventional anchoring effect paradigms. A method for estimating unbiased estimates is proposed. This method is used to estimate unbiased estimates for a set of anchoring effect data and the integration model is fit to these data. This article closes by speculating on possible theoretical insights into anchoring effects that might be gleaned by using the proposed methodology and possible practical applications.

Anchoring Effects

The goal of this paper is to propose a method by which anchoring effects can be mathematically modeled. The ability to mathematically model anchoring effects might be useful both for differentiating among theoretical models of anchoring effects and for calculating likely practical applications of anchoring effects in situations such as negotiations (e.g., Chapman & Bornstein, 1996; Galinsky & Mussweiler, 2001), auctions (Ku, Galinsky, & Murnighan, 2006), and pricing (Northcraft & Neale, 1987). These possible applications of the proposed model will be discussed in the General Discussion section.

In anchoring effects, estimates of an unknown value are assimilated towards an arbitrary numeric value called the anchor. For example, in a well-known study, Tversky and Kahneman (1974) asked participants to judge whether African nations represented a higher or lower percentage of UN-member nations than an anchor and then to estimate the actual percentage. Estimates were assimilated towards the anchor. When the anchor was 10% of UN-member nations, the median estimate was assimilated downward toward 10% to equal 25%; but when the anchor was 65%, the median estimate was assimilated upward toward 65% to equal 45%.

Assimilation effects like these are typically mathematically modeled using Anderson's (1965; 1981) integration model. A mathematical formalization like the integration model formalization was alluded to in at least one anchoring effect paper (see Jacowitz & Kahneman, 's, 1995, discussion of priming models of anchoring effects). In addition, this mathematical formalization has been used to model assimilation effects in phenomena as diverse as impression formation (the domain that originally inspired Anderson's model, see Urada, Stenstrom, & Miller, 2007, for a recent application), physical attractiveness (e.g., Wedell, Parducci, & Geiselman, 1987), product evaluation (e.g., Miyazaki, Grewal, & Goodstein, 2005; Troutman & Shanteau, 1976), risk assessment (e.g., Hampson, Andrews, Barckley, Lee, & Lichtenstein, 2003), and the best timing for lesbian and gay politicians to come out of the closet (Golebiowska, 2003)[1].

The Proposed Mathematical Model

Anderson's (1965; 1981) integration model would model the assimilation observed in anchoring effects as a weighted average of the anchor value (A) and the unbiased estimate a participant would have made had she or he never seen the anchor (U: U for Unbiased; see below for how this quantity can be empirically measured):

$$EST = wA + (1-w)U \quad (1)$$

where EST represents a participant's estimate (i.e., the dependent measure in anchoring estimation tasks) and w is the weight bound between 0 and 1 of the anchor value (A) relative to the unbiased estimate (U). A weight of 0 would represent a case in which estimates were not affected at all by exposure to the anchor. In such a case, unbiased estimates (U) would be equal to participants' estimates (EST) so that EST = U. A weight of 1 would represent a case in which all participants simply respond with the anchor value. Weights between these two extremes represent estimates that are assimilated toward the anchor value, but are not equal to it.

The key problem in using Anderson's (1965; 1981) integration model to model anchoring effects is that it requires the modeler to know what participants' unbiased estimates (U) would have been had they never seen the anchor. Measuring these unbiased estimates is made particularly difficult, because it is not possible to ask participants to make a numerical estimate twice (once before and once after being exposed to the anchor value) as

their first numerical estimate will bias their second. To solve this problem, the methodology proposed here would have participants make a non-numerical estimate before being exposed to the anchor and then make a numerical estimate afterwards. The mapping between non-numerical estimates and numerical estimates can then be established by running a control condition in which participants make both types of estimates without being exposed to the anchor and calculating a regression line between the two types of estimates. The unbiased estimates (U) of the participants in the experimental condition can then be calculated using the non-numerical estimates that these participants make and the regression line.

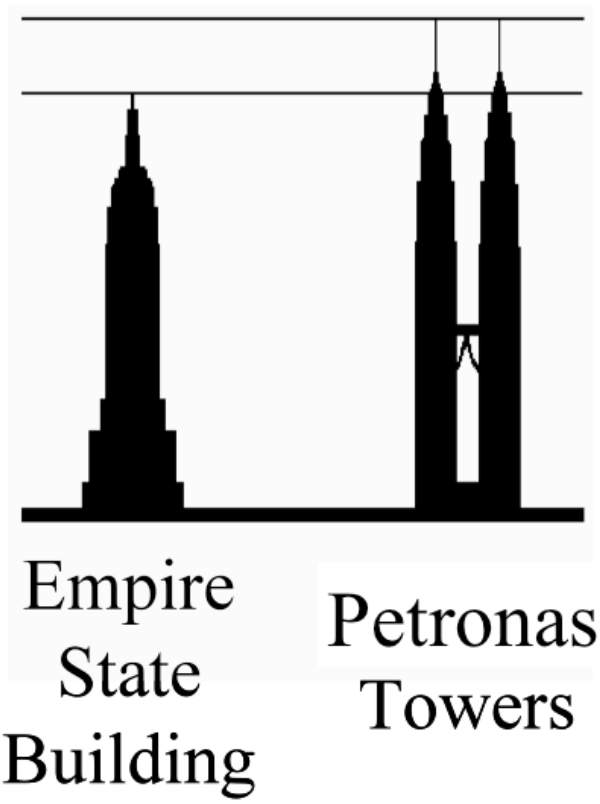


Figure 1. Graphic estimate of the height of the Sears Tower. Participants placed a tick mark between the horizontal line representing the height of the Empire State Building and the horizontal line representing the height of the Petronas Towers to represent how tall they believed the Sears Tower to be.

In the data modeled below, for example, the task was to estimate the height of the Sears Tower (a Chicago landmark and one of the world's tallest buildings; since the time during which these data were collected, this building has been renamed the Willis Tower). Participants made two estimates: a non-numerical estimate and a numerical estimate. The non-numerical estimate was made on the graphic presented in Figure 1. Participants were told that the Empire State Building was the tallest building in the world until the Sears' Tower was built and that the Sears' Tower was the tallest building in the world until the Petronas Towers in Kuala Lumpur, Malaysia were built (taller buildings yet have been built since the Petronas Towers were built). Participants made a tick mark between the two horizontal lines in Figure 1 to denote how tall they believed the Sears Tower to be relative to the Empire State Building and the Petronas Towers. The distance between the bottom line representing the height of the Empire State Building and each participant's tick mark was then measured in millimeters (mm).

The numerical estimate was the number of feet tall that participants estimated the Sears Tower to be. Participants in the control condition made the non-numerical estimate and then the numerical estimate without being exposed to the anchor. Participants in the experimental condition made the non-numerical estimate before they made a judgment regarding whether the Sears Tower was taller or shorter than the anchor value of 1,367 feet and then made the numerical estimate. A regression line was calculated between the control participants' non-numerical and numerical estimates. This regression line was then used to calculate the experimental participants' unbiased estimates (U) from their non-numerical estimates.

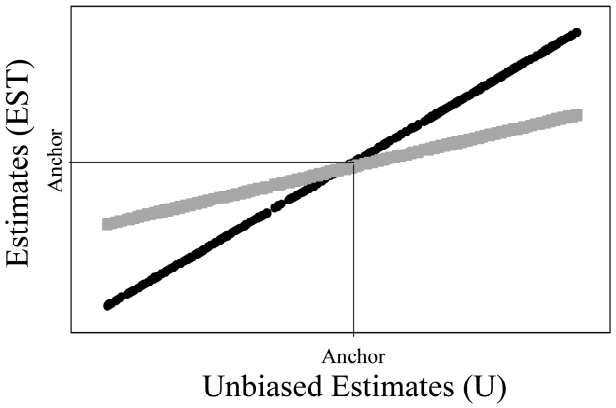


Figure 2. Anchoring effect that would be characterized as an assimilation effect. The black line represents the predicted pattern of estimates, if estimates were not affected by the anchor. The gray line represents the predicted pattern of estimates, if an assimilation effect were observed. Notice that the gray line represents a weighted average of the black line (estimates unbiased by the anchor) and the anchor value (See Equation 1).

A pattern of biases that would fit Anderson's (1965; 1981) integration model definition of an assimilation effect as presented in Equation 1 is demonstrated in Figure 2. The x-axis represents unbiased estimates (U) and the y-axis represents participants' estimates in anchoring estimation tasks (EST). Do not confuse this figure with the similar-looking figures used by Chapman and Johnson (1994). In Chapman and Johnson's figures, the x-axis represented

alternative anchor values. In Figure 2, the x-axis represents unbiased estimates and the location labeled "Anchor" represents a situation wherein a participant's unbiased estimate just happened to be equal to the anchor value. The black line in Figure 2 represents what the pattern of estimates would look like, if the anchor did not bias estimates (i.e., the case in which w = 0 and EST = U). The gray line represents a pattern of biased estimation that would be characterized as an assimilation effect (any linear slope between the slope of the black line and horizontal—that is, where w in Equation 1 takes a value greater than zero and less than one—would be classified as an assimilation effect).

Notice that regardless of the values of the unbiased estimates (U), Equation 1 predicts that they will be biased toward the anchor value by the same proportion. For example, all values might be biased 20% toward the anchor. Sometimes the term "assimilation effect" has been used roughly to refer to any bias towards a standard regardless of the extent of the bias and whether the bias toward the standard is uniform (e.g., Schwarz & Bless, 1992). While using the term in this way often provides a useful way to quickly classify results (i.e., as either "assimilation," bias toward or "contrast," bias away from a standard), Anderson's (1965; 1981) definition is more precise in that it captures the degree of bias toward the anchor across the entire range of unbiased estimates and provides a starting point from which to model anchoring effects. If it turns out that not all estimates are biased toward the anchor by the same proportion (e.g., unbiased estimates close to the anchor might be biased towards the anchor by a smaller proportion than unbiased estimates that are farther away from the anchor or vice versa), then the methodology proposed here can also be used to fit alternative equations—other than the integration theory equation—to anchoring effect data.

We used this methodology and collected anchoring effect data to which Anderson's (1965; 1981) integration model could be fit.

Anchoring Effect Data

The purpose of the experiment reported here was to use the methodology proposed above to collect data to which mathematical models—Anderson's (1965; 1981) integration model, in particular—could be fit. There was an experimental group of participants and a control group. The experimental group made a non-numerical estimate of the height of the Sears' Tower, then compared its height to the anchor value of 1,367 feet, and finally made a numerical estimate of the height of the Sears' Tower in feet. The control group made a non-numerical estimate and then a numerical estimate without ever being exposed to the anchor.

Method

Participants. One hundred sixty passengers on the Chicago elevated train system participated voluntarily (80 in the control condition and 80 in the experimental condition).

Materials and Procedure. We told our participants that the Empire State Building was the tallest building in the world until the Sears Tower was built and that the Sears Tower was the tallest building in the world until the Petronas Towers were built. To measure unbiased estimates, we first asked participants to estimate the height of the Sears Tower graphically by showing them in-scale silhouettes of the Empire State Building and the Petronas Towers as shown in Figure 1. Horizontal lines crossed the page to represent the height of each skyscraper. Participants placed a tick mark between the lines to represent their estimates of the height of the Sears Tower. After estimating the height of the Sears Tower graphically, participants in the control condition simply estimated the height of the Sears Tower in feet (numerical estimate). Participants in the experimental condition judged the height of the Sears Tower to be "more" than or "less" than the anchor value of 1,367 feet before estimating the height of the Sears Tower in feet (numerical estimate).

Results

The results are presented in Figure 3. As noted in the discussion of Figure 2 above, be careful not to confuse these figures with the similar-looking figures used by Chapman and Johnson (1994). The x-axis here represents unbiased estimates as measured using the graphic presented in Figure 1; and the y-axis represents participants' numerical estimates in feet. We first investigated whether an anchoring effect was observed by performing a t-test on the absolute difference between participants' numerical estimates in feet and the anchor value of 1,367 feet. The anchoring effect was highly reliable, $t(158)=4.72$, $p<.01$. Estimates were significantly closer to the anchor value in the experimental condition (M=128.30 feet away from 1,367 feet, SD=127.93) than in the control condition (M=479.90 feet away from 1,367 feet, SD=654.42).

Fitting the Model

Equation 1 was fit to the results of this experiment. The criterion variable, EST, represented each participant's estimate. To use Equations 1 to predict EST, one must somehow measure the estimates participants would have made had they never seen the anchor value. That is, one must measure participants' unbiased estimates, Parameter U. To do so, we used the results from the control group to regress their non-numerical estimates (as collected using the graphic presented in Figure 1 and measured on mm from the bottom horizontal line representing the height of the Empire State Building) on their numerical estimates. We then used this regression equation along with each experimental participant's non-numerical estimate to predict what their unbiased numerical estimates, U, would have been had they never seen the anchor. The regression line predicts U as: U=766.12+(50.93*the distance in mm between the bottom line in Figure 1 representing the height of the Empire State Building and each participant's tick mark). With EST equal to each experimental participant's estimate and U equal to the value predicted by the regression equation, assimilation effects toward the anchor were modeled using Equation 1.

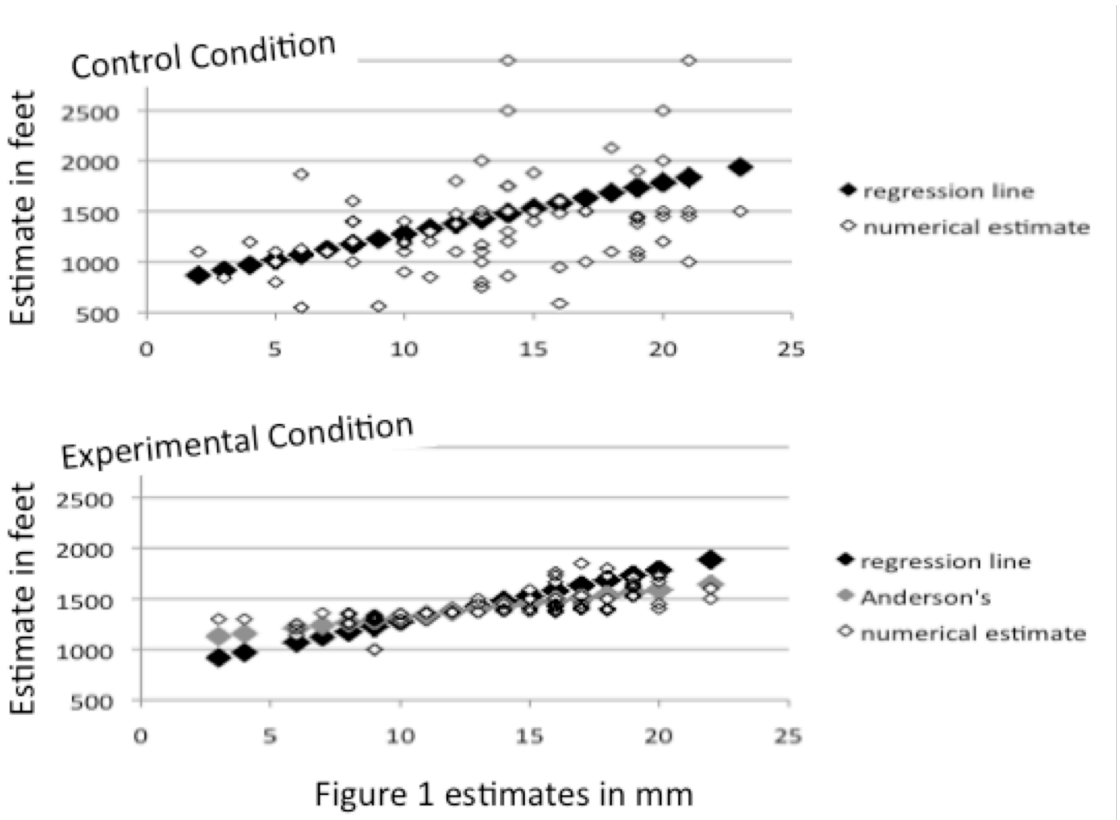


Figure 3. Results of the anchoring effect experiment reported here including the regression line and Anderson's (1965;1981) integration model fits. The x-axis represents participants' unbiased estimates of the height of the Sears/Willis Tower on the graphic presented in Figure 1 and the y-axis represents participants' numerical estimates of the height in feet. The white diamonds represent particular participants' estimates; the black diamonds represent the regression line calculated on the control participants' estimates; and the gray diamonds represent the best fit from Anderson's integration model.

Parameter A, representing the anchor value, took a value of 1,367 feet and the best-fitting value for Parameter w was calculated using a root mean squared error (RMSE) criterion. The best-fitting value for Parameter w was 0.47; and the RMSE was 116.93. A paired sample t-test on the squared errors of the values predicted by Anderson's integration model versus the squared errors of the values predicted by the regression equation found that Anderson's integration model provided a better fit, $t(79)=3.81$, $p<.01$.

Discussion

A method of mathematically modeling anchoring effects was proposed. This method calculated unbiased estimates (the estimates participants would have made had they never seen the anchor value) by having participants make a non-numerical estimate before being exposed to the anchor value and a numerical estimate afterwards. The mapping between non-numerical estimates and numerical estimates was calculated by asking a control group of participants to make both types of estimates without ever being exposed to the anchor and calculated a regression line between the two types of estimates. The regression line along with the non-numerical estimates of the experimental participants allowed us to estimate what the experimental participants' estimates would have been had they never been exposed to the anchor value. Anderson's (1965; 1981) integration model (Equation 1) was then fit to these data where U represented each experimental participants' unbiased estimate as calculated by the regression line, EST represented each participants' numerical estimate, and A represented the anchor value of 1,367 feet. The best fitting value for parameter w using a RMSE criterion was 0.47.

Future research should fit other types of equations to anchoring effect data collected using this method. Doing so might prove useful for further refining theoretical models of anchoring effects. For example, if the anchor value is outside of the range of acceptable estimates, then Tversky and Kahneman's (1974) account of anchoring effects—under which anchors provide a starting point for participants' search for an appropriate estimate—would not produce a pattern of results that should be modeled using Anderson's integration model. Instead of predicting that all unbiased estimates would be biased toward the anchor by the same proportion, Tversky and Kahneman's (1974)

account would predict an approximately horizontal estimation function. It would predict a horizontal estimation function, because all participants would start their search for an appropriate value at the anchor value which is outside of the range of acceptable estimates, and adjust from there, stopping at the first acceptable value. They would do so regardless of what their unbiased estimates would have been had they never been exposed to the anchor value. One qualification on this prediction of the anchoring and adjustment model of anchoring effects would be if the range of values that participants thought acceptable correlated with their unbiased estimates, but this question could be addressed in future research as well (by having control participants identify the range of values they consider acceptable) and the issue would not have been addressable without the methodology proposed here.

By contrast, priming models of anchoring effects (Wilson, et al., 1996; Wong & Kwong, 2000) would predict estimation functions that would follow Anderson's integration model pattern (see Jacowitz & Kahneman, 's, 1995, discussion of priming models of anchoring effects). Exposure to the anchor value would prime that value and then estimates would be a weighted average between the primed values and the unbiased estimates participants would have made had they never been exposed to the anchor.

The pattern of bias predicted by Mussweiler and Strack's (1999; see also Strack & Mussweiler, 1997) selective accessibility model is less clear. The selective accessibility model assumes that when people compare the unknown, to-be-estimated value to the anchor value, they test whether the unknown, to-be-estimated value might be the same as the anchor value by searching for semantic information that would confirm that the to-be-estimated value is equal to the anchor value. Confirmation biases almost always produce a situation wherein people are able to find semantic information about the to-be-estimated value suggesting that it is equal to the anchor value. If this account of anchoring effects is correct, then the degree of bias toward the anchor will depend upon the amount of confirmatory information they are able to recall. The ability to find such confirmatory evidence may vary as a function of people's unbiased estimates. People whose estimates would have otherwise suggested a value close to the anchor based upon their unbiased semantic knowledge of the to-be-estimated value may be more likely to find confirmatory evidence than people whose unbiased estimates would have otherwise been farther away. The proportion of bias towards the anchor may then be greater for unbiased estimates that are relatively close to the anchor than for unbiased estimates that are farther away from the anchor. Furthermore, future work might investigate the role of selective accessibility mechanisms in anchoring effects by using the methodology proposed here to investigate anchoring effects when participants have a great deal of semantic knowledge about the to-be-estimated value and when they do not.

The methodology proposed here (perhaps using a rating scale to measure unbiased estimates, rather than the measure presented in Figure 1) may also be useful for studying practical applications of anchoring effects in situations such as negotiations (e.g., Chapman & Bornstein, 1996; Galinsky & Mussweiler, 2001), auctions (Ku, Galinsky, & Murnighan, 2006), and pricing (Northcraft & Neale, 1987). For example, starting negotiations over the selling price of a home at a high initial asking price may have different effects depending upon what the potential buyer's unbiased estimate of a reasonable price for the house would have been had she or he never heard the asking price. It is not clear *a priori* whether all buyers' bids are biased toward the initial asking price by the same proportion. It might turn out that closer unbiased estimates are biased toward the initial asking price by a smaller proportion; or it might turn out that they are biased toward the initial asking price by a greater proportion. If it turns out that closer unbiased estimates are biased toward the initial asking price by a greater proportion, then it may not be the case that larger initial asking prices always produce the highest selling prices even if on average they do so. It may turn out that this phenomenon is mostly due to people who's unbiased estimates would have been relatively high before hand and the bias just makes their estimates of an appropriate bid higher yet. If so, then lower initial asking prices might be more effective in producing high selling prices among the segment of consumers whose unbiased estimates of an appropriate price were not quite as high at the start. If so, then the methodology proposed here might be useful in setting optimal initial asking prices for the entire range of potential consumers.

References

Anderson, N. H. (1965). Averaging versus adding as a stimulus-combination rule in impression formation. *Journal of Experimental Psychology, 70*, 394-400.

Anderson, N. H. (1981). *Foundations of integration theory*. New York: Academic Press.

Chapman, G. B., & Bornstein, B. H. (1996). The more you ask for, the more you get: Anchoring in personal injury verdicts. *Applied Cognitive Psychology, 10*, 519-540.

Chapman, G.B. & Johnson, E.J. (1994). The limits of anchoring. Journal of Behavioral Decision Making, 7, 223-242.

Galinsky, A. D., & Mussweiler, T. (2001). First offers as anchors: The role of perspective-taking and negotiator focus. *Journal of Personality and Social Psychology, 81*, 657-699.

Golebiowska, E. A. (2003). When to tell?: Disclosure of concealable group membership, stereotypes, and political evaluation. *Political Behavior, 25*, 313-337.

Hampson, S. E., Andrews, J. A., Barckley, M., Lee, M. E., & Lichtenstein, E. (2003). Assessing perceptions of synergistic health risk: A comparison of two scales. *Risk Analysis, 23*, 1021-1029.

Jacowitz, K. E., & Kahneman, D. (1995). Measures of anchoring in estimation tasks. *Personality & Social Psychology Bulletin, 21*, 1161-1166.

Ku, G., Galinsky, A.D., Murnighan, J.K. (2006). Starting low but ending high: A reversal of the anchoring effect in

auctions. *Journal of Personality and Social Psychology*, 90, 975-986.

Miyazaki, A. D., Grewal, D., & Goodstein, R. C. (2005). The effect of multiple extrinsic cues on quality perceptions: A matter of consistency. *Journal of Consumer Research, 32*, 146-153.

Mussweiler, T., & Strack, F. (1999). Hypothesis-Consistent Testing and Semantic Priming in the Anchoring Paradigm: A Selective Accessibility Model. *Journal of Experimental Social Psychology, 35*, 136-164.

Northcraft, G. B., & Neale, M. A. (1987). Experts, amateurs, and real estate: An anchoring-and-adjustment perspective on property pricing decisions. *Organizational Behavior and Human Decision Processes, 39*, 84-97.

Schwarz, N., & Bless, H. (1992). Constructing reality and its alternatives: An inclusion/exclusion model of assimilation and contrast effects in social judgment. In L. Martin & A. Tesser (Eds.), *The construction of social judgments* (pp. 217-245). Hillsdale, NJ: Lawrence Erlbaum.

Strack, F., & Mussweiler, T. (1997). Explaining the enigmatic anchoring effect: mechanisms of selective accessibility. *Journal of Personality and Social Psychology, 73*, 437-446.

Troutman, C. M., & Shanteau, J. (1976). Do consumers evaluate products by adding or averaging attribute information? *Journal of Consumer Research, 3*, 101-106.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124-1130.

Urada, D., Stenstrom, D. M., & Miller, N. (2007). Crossed categorization beyond the two-group model. *Journal of Personality and Social Psychology, 92*, 649-664.

Wedell, D. H., Parducci, A., & Geiselman, E. (1987). A formal analysis of ratings of physical attractiveness: Successive contrast and simultaneous assimilation. *Journal of Experimental Social Psychology, 23*, 230-249.

Wilson, T. D., Houston, C. E., Brekke, N., & Etling, K. M. (1996). A new look at anchoring effects: Basic anchoring and its antecedents. *Journal of Experimental Psychology: General, 125*, 387-402.

Wong, K. F. E., & Kwong, J. Y. Y. (2000). Is 7300 m equal to 7.3 km? Same semantics but different anchoring effects. *Organizational Behavior and Human Decision Processes, 82*(314-333).

The Philosophy of Affective Neuroscience

Our symposium showcases the interdisciplinary cutting edge innovations of the cognitive sciences. It is the unique meeting of the founder of Affective Neuroscience with an interdisciplinary set of scholars who follow the implications of his work through the philosophy of psychology, the philosophy of Self, and neuroscience and law.

Speakers
Stephen Asma
Rami Gabriel
Thomas Greif

Moderator
Jaak Panksepp

Moderator

Jaak Panksepp, Distinguished Research Professor Emeritus of Psychology, Bowling Green State University. Head, Affective Neuroscience Research, Falk Center for Molecular Therapeutics, Northwestern University. In addition to 300+ scientific articles, I have co-edited the multivolume Handbook of the Hypothalamus and of Emotions and Psychopathology, a series in Advances in Biological Psychiatry and most recently a Textbook of Biological Psychiatry (Wiley, 2004), My other textbook, Affective Neuroscience: The Foundations of Human and Animal Emotions (Oxford, 1998), has helped inaugurate a new field of inquiry which attempts to probe the affective infrastructure of the mammalian brain. Our working assumption is that all of consciousness was built on affective value systems during the long course of brain evolution

Speakers

Rami Gabriel, Ph.D in Cognitive and Perceptual Sciences from University of California, Santa Barbara. Dissertation concerned non-conscious affective processes in a Prosopagnosic patient. Member of Columbia College Chicago School of Liberal Arts and Sciences Research Group in Mind, Science, and Culture.

Title of talk: Modularity in Cognitive ψ and Affective Neuroscience.

My talk explores the psychological module in the context of findings from affective neuroscience. I contrast, in terms of practicality and veridicality, cognitive science's formulation of the cognitive module with Panksepp's notion of basic biological behavioral systems. The deeper theme of my presentation is the question of human nature and the processes of the human animal we need to specify towards positing a biologically-based codification of the cognitive processes that constitute human nature.

Stephen T. Asma, Ph.D in Philosophy of Science, is author of several books that seek to bridge the sciences and humanities. He is Professor of Philosophy at Columbia College

Chicago, and member of the Liberal Arts and Sciences Research Group in Mind, Science, and Culture.

Title: Affective Neuroscience and Its Implications for the Philosophy of Self.

The mind/body problem continues to plague philosophy. The nature of Self awareness and the origin and persistence of personal identity still loom large in contemporary philosophy of mind. Many philosophers have been wooed by the computational approach to consciousness and they attempt to generate the Self amidst the phenomenon of neo-cortical information processing. Affective neuroscience offers another pathway to understanding the evolution and nature of Self. This talk will explore how affective neuroscience acts as a positive game-changer in the philosophical pursuit of Self. In particular, I will focus on closing the gap between the phenomenology of psychological consolidation and affective neurodynamic processes.

Glennon Curran, May 2011 Candidate for Juris Doctor at The John Marshall Law School, Chicago, Illinois. Executive Board Member of The John Marshall Law School Center for Trial Advocacy and Dispute Resolution. Member of Columbia College Chicago School of Liberal Arts and Sciences Research Group in Mind, Science, and Culture.

Title of Talk: Affective Neuroscience and Law

My talk explores the application of Affective Neuroscience to the law. Jaak Panksepp's exegesis of the triune brain complicates extant applications of cognitive neuroscience to the law. I contrast the neocortical causal explanations of the mental culpability of criminal defendants made in Cognitive Neuroscience with sub-cortical affective systems explanations raised in Affective Neuroscience. I argue the sub-cortical foundations of human behavior must be considered in any attempt to inform law with neuroscience.

Response Choice When Telling Lies

Emma Williams
School of Psychology, Cardiff University

Lewis Bott
School of Psychology, Cardiff University

Michael Lewis
School of Psychology, Cardiff University

John Patrick
School of Psychology, Cardiff University

Abstract: It is commonly believed that telling a lie is more difficult than telling the truth. However, the precise reason for this difficulty remains uncertain. The Activation-Decision-Construction Model (ADCM; Walczyk, Roper, Seemann & Humphrey, 2003) suggests that telling a lie will take longer than telling the truth due to the additional processes involved, such as suppression of the truth. Experimental work investigated the lie construction component of the model by manipulating the number of plausible lie responses available. Participants lied and told the truth regarding the colour of a square presented on a computer screen. Results support the general claims of the ADCM, but suggest that longer response times when individuals lie to questions with more than one possible lie response relate to a fixed cost choice between multiple response possibilities. This contributes to current understanding of which situations may enhance processing differences between telling a lie and telling the truth.

Does Practice Narrow the Radius of Spatial Interference in Mental Images?

Don Lyon
L3 Communications

Abstract: When people attempt to generate a mental image of a complex, verbally-described path, crowded regions of the path suffer from spatial interference (Lyon, Gunzelmann & Gluck, Cognitive Psychology, 57, 2008). A path is presented as a sequence of one-unit segments within a 7x7 grid, analogous to city blocks (e.g. 'Up 1 [Block]'; 'Right 1'; 'Down 1', etc.). Participants must decide whether each new segment intersects with a prior part of the path. Initially, prior segments of a path within 2 grid spaces of the current path segment produced spatial interference. Although there were substantial individual differences, for most participants interference radius was reduced to one grid space with under 10 hours of practice. One possible explanation for this reduction is that, with practice, people can learn to attend selectively to increasingly smaller areas within a vision-like representation of the mental image, in the absence of any actual visual stimulus.

A Sense of Order: Numerical ordering ability predicts complex mental arithmetic performance

Ian Lyons
University of Chicago

Sian Beilock
University of Chicago

Abstract: What are the key cognitive factors that characterize the potential difference between symbolic and non-symbolic representations of numerical magnitude, and can individual variability in such factors be used to predict differences in more complex mathematical processes? We suggest that the availability of information about relative numerical order is a critical factor that distinguishes symbolic from non-symbolic numbers. In the current experiment, we provide evidence that individual variability in symbolic numerical ordering ability strongly predicts performance on a series of complex mental arithmetic tasks even when controlling for a wide array of competing factors, including individuals' precision in non-symbolic magnitude representations. Moreover, symbolic numerical ordering ability is shown to fully mediate the previously reported relation between non-symbolic magnitude processing and more complex mathematical skills. These results have important implications for designing math-education techniques and identifying reliable math-performance markers.

Imaginary affordances shape children's preference judgments

Tania Henetz
Stanford University

Daniel Casasanto
Max Planck Institute for Psycholinguistics, Neurobiology of Language Group, Nijmegen, NL

Abstract: Motor fluency influences preference judgments: people tend to like things they can manipulate easily. Yet, links between motor fluency and preference extend beyond the domain of concrete objects that afford physical manipulation. People implicitly associate abstract ideas like goodness and intelligence with locations in space that ordinarily afford fluent actions. How do these abstract associations develop? Here we tested whether children's preference judgments are influenced by implicit affordances of imaginary objects. Children imagined helping a cartoon character store toys in a bookcase, drawn next to the character. They tended to assign toys the characters liked to locations on the shelves that would afford fluent actions and toys they didn't like to locations that would afford less fluent actions. Crucially, the 'fluent' location was determined by implicit constraints on the character's actions, not by the child's own action affordances. Imaginary affordances may help link concrete motor actions with abstract preference judgments.

Testing fMRI predictions of a Cognitive Model of the Problem State Multitasking Bottleneck

Jelmer Borst
University of Groningen Carnegie Mellon University

Niels Taatgen
University of Groningen Carnegie Mellon University

Hedderik Van Rijn
University of Groningen

Andrea Stocco
Carnegie Mellon University

Abstract: It has been shown that people can only maintain one intermediate mental representation, or 'problem state', concurrently. When multiple problem states have to be maintained, performance decreases sharply, an effect referred to as the problem state bottleneck. We investigate this bottleneck using a triple-task, in which participants have to solve subtraction problems, enter text, and perform a listening task concurrently. The triple-task confirmed the existence of a problem state bottleneck. To explain the behavioral results in detail, a cognitive model was developed using ACT-R (Anderson, 2007) and the threaded cognition theory (Salvucci & Taatgen, 2008). The model showed a close fit to the emperical data. It was subsequently used to generate fMRI predictions for five brain areas. These predictions where tested in an experiment, showing a good correspondence between model predictions and fMRI data, indicating that the problem state bottleneck is probably located in the intraparietal sulcus.

Grounded Congruency Effects Between Vertical Meaning and Vertical Responding: Not Replicated

Lauren McDonough
Emory University

Christine Wilson-Mendenhall
Emory University

Lawrence Barsalou
Emory University

Abstract: According to theories of grounded cognition, words whose semantics are associated with a salient vertical position (e.g., CEILING vs. CARPET) should activate simulations of these positions in space. When responses are analogously made in the vertical dimension, grounded congruency effects should result (e.g., processing CEILING should be faster for an UP vs. DOWN response). Previous research obtained grounded congruency effects when participants used ink color (RED vs. BLUE) as a cue for response direction (Casasanto, 2008). Typically researchers assume that these effects are automatic, but they could possibly be strategic. Two experiments attempted to assess this issue with 6 groups of 24 participants each, but failed to replicate the original grounded congruency effect, leading us to question its reliability when ink color is used as a cue. We further discovered a motor facilitation effect for upward as opposed to downward responses not reported previously in this paradigm.

MIReR: Media Integration Reflection Resource

Andreea Danielescu
Arizona State University

Ellen Campana
Arizona State University

Abstract: Designers of experiential media systems rely on intuition and experience to create systems. This trial and error process takes time to learn, and it is possible to exceed a user's cognitive load in multimodal environments. This offers an opportunity to explore perception and cognition. Behavioral experiments, the standard for psychological inquiry, are time consuming and focus on one variable to achieve accurate results. Instead, MIReR provides a dynamic, holistic way of exploring data collected from user experiences. A unified representational framework makes this possible – designers provide a concept map that describes the intended meaning behind the sounds and visuals, and their relationship. This map is analyzed by MIReR's cognitive architecture and compared to user experience data to produce an estimate of the user's cognitive load. As the designer explores different possibilities, MIReR tracks changes and effects on users, creating an environment that produces new insights into perception and cognition.

Social Influences on the, um, Use of, uh, Fillers

Esther Walker
University of British Columbia

Evan Risko
University of British Columbia

Alan Kingstone
University of British Columbia

Abstract: Language, at its core, is a social act. The present investigation sought to examine the influence of interpersonal context on filler use ("um", "uh") while answering factual questions. Experiments 1 and 2 investigated differences in computer versus human interaction. As predicted, more fillers were uttered when interacting with a human than with a computer. Experiment 3 sought to examine a self-presentational view of filler use, whereby the mere presence of another human should increase one's use of fillers. Consistent with a self-presentational account, mere presence of the experimenter elicited more fillers than when the experimenter was absent. A cross experimental analysis revealed that while mere presence increases filler production, the need for interpersonal coordination increases filler use above and beyond mere presence. These results are consistent with at least two views of filler function: (1) fillers are used to save face and (2) fillers are used to coordinate interpersonal interactions.

Probability estimation by mice in an interval timing task

Aaron Kheifets
Rutgers University Center for Cognitive Science: RUCCS

C. Randy Gallistel
Rutgers University Center for Cognitive Science: RUCCS

Abstract: Keeping track of and detecting changes in the probability of events is a central problem for animals. We presented mice with an interval timing task: with probability p, mice were reinforced for staying at the first hopper until time t. With complementary probability 1-p they were reinforced for arriving at the second hopper before t+k. Because no animals are perfect timers, this task was difficult due to small k. Depending on p, the optimal switch point changed: if long trials were more likely, switching too late became more costly than switching too early, so the optimal switch time occurred later. Subjects showed highly significant (p<0.005) differences in their mean switch times when p was manipulated. Moreover, subjects were able to update their frequency estimates when the underlying probabilities of trial types changed and their estimates converged on accurate values quickly in comparison to plausible Bayesian optimal models.

The Specificity of Non-Arbitrary Sound-to-Meaning Correspondences in Spoken Language

Christina Y. Tzeng
Emory University

Lynne C. Nygaard
Emory University

Laura L. Namy
Emory University

Abstract: Sound symbolism, or non-arbitrary correspondences between the sound of a word and its meaning, is an inherent property of natural language. Although previous research suggests that listeners are sensitive to sound-to-meaning correspondences, little is known about the specificity of these mappings. The present study investigated whether sound symbolic properties correspond to specific meanings, or whether these properties aid mappings to other semantic dimensions as well. English-speaking adults heard sound symbolic foreign translations of four adjective pairs (big-small, pointy-round, fast-slow, still-moving), and for each foreign word, chose which of two English antonyms (matched or mismatched with word dimension) was its correct translation. Participants reliably matched foreign words to their correct meanings, as well as to related semantic dimensions, suggesting not only that listeners utilize sound-to-meaning correspondences to infer the meanings of unfamiliar words, but also that sound-symbolic properties facilitate word-to-meaning mappings within a range of associated and co-varying semantic dimensions.

Pattern Recognition Principle Theoretical Model of Mind-Brain Functioning

Gilberto de Paiva

Abstract: The Computational Theory of Mind and the Connectionist Neural Model are actually the dominant cognitive models in the scientific community. But there is still a discussion about their compatibility as the basis of Cognitive Science. A more fundamental explanation of this two leading cognitive mechanisms in physical basis can give a better explanation of the mind and brain phenomena.

I propose the concepts of PATTERN RECOGNITION-PROCESSING-LEARNING as NECESSARY AND SUFFICIENT PRINCIPLE to build a solid foundation of cognitive science. This encloses a general body-mind, physical, psychological, neural, functional and computational reductionist explanation of the mind cognitive phenomena.

Supporting arguments are: 1- Pattern recognition is a necessary principle for cognitive science. It is a key process in many different scientific areas, but no cognitive science model takes it formally for a general brain-mind theory. 2- The equivalence of the physical-neural-computational mechanisms of pattern recognition as the basis of the cognitive phenomena in performing all cognitive functions (sensory, memory, learning, processing, logical, feeling, emotions, thought, consciousness, etc) , which are here demonstrated. 3- I also propose some key biological cognitive processes and strategies strictly related to pattern recognition processing. 4- A definitional-explicative modeling building theory is also shown giving simple, understandable and unambiguous definitions and scientific explanation of most key cognitive concepts like thinking, self and consciousness. Such a theory is of fundamental importance because those type of cognitive concepts lacks even a reasonable definition. With a scientific-objective definitions we can evaluate, compare and estimate its consistency and also propose experimental and empirical experiments. As an example I suggest one preliminary definition of self-consciousness as: "the recognition by the pattern recognition system of the patterns of its own activity". 5- A mathematical-logical formalism study of cognitive pattern recognition is here proposed as theoretical basis. Here a general formal definition of pattern recognition is proposed in the cognitive science scope. 6- With this theoretical formulation we are able to include other cognitive properties to any basic definition as far as needed, demonstrating it as a solid and promising theory.

As any candidate as a complete theory of cognitive science, this model allow us to reinterpret all branches of human reasoning, including the philosophy and foundations of science and the human understanding of the universe. With the promising applied pattern recognition technology already in development, this model could help to give some directions to artificial intelligence and also neurobiology and psychology-sociology research.

Individual Differences in Explaining Noisy Data

Daniel R. Little
Indiana University

Richard M. Shiffrin
Indiana University

Abstract: In science, we design our inference approaches to trade off fit to observed data (models are good that fit well) and complexity (models or explanations that fit or explain everything are bad). Here, we examine how observers balance fit and complexity by asking observers to estimate causal models for noisy data. Specifically, participants are shown a number of scatterplots that vary in the number of data points shown, the noise added to the true function and the complexity of the true function. For each set of noisy data points, participants estimate a function which best captures their guess at the causal explanation between the input and the output. A generative psychological model combining Bayesian model selection and Gaussian process regression is used to examine individual differences in biases toward simple explanations. Our results indicate that some participants prefer simple polynomial, rule-based explanations and others prefer distance-based, similarity explanations.

Making a good impression (formation model): a more complete account of processing

Tei Laine
Institute of High Performance Computing

Swati Gupta
Institute of High Performance Computing

Brian M. Monroe
Institute of High Performance Computing

Abstract: First impression formation is the process by which people make assumptions, regardless of objective accuracy, about someone they meet for the first time by integrating information including the person's appearance, verbal and non-verbal cues, and facts she might reveal about herself.

We propose a model of first impression formation that integrates this kind of information into a coherent representation taking into account the 1) potentially asymmetric nature of inferences people make from stereotypes, traits, and behaviors, 2) prior probabilities of inferred characteristics, 3) cognitive capacity limitations in processing of incoming information, and 4) the influence of positive and negative affect in the impression.

We think that our model not only compares favorably with Kunda & Thagard (1996) parallel constraint satisfaction model, but also accounts for additional phenomena such as asymmetrical inferences and affective coherence.

The representation of idiom words in the mental lexicon

Simone Sprenger
University of Groningen

Hedderik van Rijn
University of Groningen

Abstract: The way in which idioms are processed and the nature of their underlying representations are subject to an ongoing debate. Most processing models agree that idioms are specific combinations of ordinary words. However, models differ with respect to the exact role that these words are allowed to play.

In the present study we tested the hypothesis that the relations between idiom words are specified on a lexical processing level. Specifically, we tested whether the constituent words of an idiom activate each other in the absence of an idiomatic (phrasal) context. In two lexical decision experiments, we found that this is indeed the case. However, the effect is modulated by the type of target word that precedes the idiomatic targets. Targets that are literally related to the first idiom target prevent activation of the second idiom target.

The results support the superlemma model of idiom processing (Sprenger, Levelt & Kempen, 2006).

Assessing the Effectiveness of Wayfinding Directions

Alycia Hund
Illinois State University

Amanda Padgitt
Illinois State University

Abstract: Our goal was to assess people's responses to wayfinding directions. Ninety college students rated the effectiveness of route descriptions through the basement of a university building. They also provide open-ended responses regarding wayfinding preferences and completed wayfinding anxiety, wayfinding strategy, and environmental familiarity self-report measures and a sense of direction exercise. The primary goal was to specify the descriptive features contained in effective and ineffective wayfinding descriptions. The best-rated route descriptions included more cardinal features, landmarks, left-right, distance, number, straight, and miscellaneous information than did the worst-rated route descriptions. Moreover, positive mentions of landmarks and negative mentions of cardinal directions were very frequent. As expected, women reported significantly higher spatial anxiety than did men. Men preferred orientation strategies more than did women, whereas women preferred route strategies more than did men. Women also reported poorer sense of direction and made larger sense of direction errors than did men.

How Agent Placement Can Influence Perceived Boss/Co-worker Agreement in a Simulated Work Environment

Justin L. Matthews
University of California, Merced

Teenie Matlock
University of California, Merced

Abstract: Interpersonal distance, the physical distance between people while they interact, is known to influence attitudes and other social dynamics. For example, merely sitting closer to a person who is presenting information can increase the persuasive power of that speaker. In the current work, we investigate how interpersonal distance will influence perceived agreement among employees in an office setting. Our participants read a passage that asked them to imagine working for an advertising firm and being in a meeting about employee layoffs with a boss. After looking at a picture of an employee seating arrangement that was close, medium, or far from a boss during the meeting, participants were asked to estimate how far the chairs were from the boss and to judge how much agreement they felt during the meeting. On average, participants indicated that they felt less agreement with bosses when interpersonal distance was high (versus medium or low). The results, which revealed that increased physical distance is associated with greater attitude "distance", have implications for the design and use of applications for virtual meetings and more generally, social interactions in virtual environments.

The structure of event representations: behavioral, imaging, and computational investigations

Anna Schapiro
Princeton University

Timothy Rogers
University of Wisconsin-Madison

Matthew Botvinick
Princeton University

Abstract: Event segmentation is often thought to rely on the identification of points in a sequence where there is relative uncertainty about what will happen next. We exposed participants to sequences of stimuli that had temporal structure but no variability in predictive uncertainty: each stimulus could be followed by four others with equal probability. We found reliable parsing between groups of stimuli that were preceded and followed by overlapping sets of items, suggesting that people are sensitive to temporal statistics beyond predictive uncertainty. We hypothesized that this reflects learning of temporal category structure, with items that occur in overlapping temporal contexts represented as belonging to the same category. Supporting this idea we found that, following exposure to the same structured sequence, participants sorted items based on their temporal contexts. To elucidate the mechanisms and representations supporting this behavior we consider alternative computational models of temporal structure learning and present preliminary fMRI results.

Pair Analysis and Joint Action Theory: A Research Protocol to Study Cognition and Interaction in Visual Analytics

Richard Arias Hernández
Simon Fraser University, School of Interactive Arts and Technology

Linda T. Kaastra
University of British Columbia, Media and Graphics Interdisciplinary Centre

Brian D. Fisher
Simon Fraser University, School of Interactive Arts and Technology

Abstract: Visual analytics, the "science of analytical reasoning with interactive visual interfaces," calls for the development of new models of human cognition in analytic interaction with information technology. While foundational work brings traditional cognitive science models to address interaction with visualization environments, research protocols to empirically test these models in "the wild" are lacking. We combine a research protocol called "Pair Analysis" with H.H. Clark's Joint Action Theory as a theory-methods package for studying cognition and interaction in visual analytic environments. Pair Analysis, an observed analytic interaction by a subject matter expert and a visual-analytic tools expert, provides a unique empirical window into the cognitive process of analytical reasoning and the social processes of interaction. We use JAT's operational concepts to characterize analytic dyads' thought processes and joint use of visualization technology. Our main hypothesis is that sustaining rhythmic interactions in Pair Analysis is indicative of sustaining cognitive flow.

The Importance of Visual Modeling in Children's Understanding of Physical Science

Nancy L Stein
University of Chicago (U of C)

Marc W. Hernandez
NORC, UYniversity of Chicago

Abstract: How do we teach children about the physical universe, given that much of what is to be learned is invisible? How does visual dynamic modeling facilitate the process of physical science understanding? What are the constraints on dynamic visual modeling? We carried out two studies on 4th and 7th grade children, where the presence or absence of visual models and dynamic visual models were varied. Fourth grade students were as good as 7th grade students in learning all parts of the module sequence. Children receiving dynamic visual graphics outperformed children who saw only static visual graphics, and both of these groups outperformed children who received only the oral/written part of the text. The presence of graphics, however, was not enough to ensure the learning of measurement concepts. Strategies that breakdown parallel physical processes and temporalize them, a well as embodiment strategies are also necessary.

Reversing the side-effect effect: the 'Rational Scientist' explanation

Kevin Uttich
University of California-Berkeley

Tania Lombrozo
University of California-Berkeley

Abstract: Theory of mind, our intuitive understanding of the mind, is often conceptualized as analogous to a scientific theory with the function of predicting and explaining behavior. However, the so-called "side-effect effect" illustrates that moral considerations influence theory of mind judgments, and has been taken as evidence that theory of mind is fundamentally evaluative. We present new evidence for an alternative, the "rational scientist" view, which holds that moral evaluations inform ToM judgments, but that this relationship arises because behavior that conforms to norms (moral or otherwise) is less informative about underlying mental states than is behavior that violates norms. In two new experiments we demonstrate that different norms (moral or conventional) lead to different intentional descriptions of the same actions, and that the effect can be eliminated when norms are reversed. This view preserves the traditional understanding of ToM, but also suggests the importance of normative considerations in social cognition.

MHP/RT: Model Human Processor with Real Time Constraints

Makoto Toyota
T-method

Muneo Kitajima
National Institute of Advanced Industrial Science and Technology

Abstract: We propose "Model Human Processor with Real Time Constraints" as a simulation model of human behavior selection. It stems on the successful simulation model of human information processing, Model Human Processor (Card, Moran, and Newell, 1983), and extends it by incorporating three theories, Maximum Satisfaction Architecture (MSA, presented at CogSci2007), Structured Meme Theory (SMT, presented at CogSci2008), and Brain Information Hydrodynamics (BIH, presented at CogSci2008). MSA, SMT and BIH deal with coordination of behavioral goals, utilization of long-term memory that works as an autonomous system, and a mechanism for synchronizing individual with environment, respectively. MHP/RT works as follows: 1) inputs information from environment and individual, 2) MHP/RT builds a cognitive frame in working memory, 3) resonates it with autonomous long-term memory, 4) maps the resonance on consciousness to form reduced representation of the input information, 5) predicts future cognitive frames to coordinate input and working memory.

http://staff.aist.go.jp/kitajima.muneo/organic-self-consistent-field-theory/index.html

CCE: Cognitive Chrono-Ethnography

Muneo Kitajima
National Institute of Advanced Industrial Science and Technology

Makoto Toyota
T-method

Abstract: We, human beings, select next behaviors that should maximize our satisfaction by making use of meme that stores our past experiences and by processing input from environment and individual by appropriately allocating available cognitive resources. The underlying processes are simulated by Model Human Processor with Real Time Constraints, MHP/RT (to be presented at CogSci2010). On the basis of MHP/RT, this paper proposes a new study method for understanding human behavior selections in daily life, Cognitive Chrono-Ethnography, CCE. When a study field is specified, CCE defines critical parameters by conducting qualitative MHP/RT simulations, then designs ethnographical field observations and recordings of elite monitors' behaviors in the space defied by the critical parameters. Structured interviews follow in order to obtain the descriptions of the participants' history of behavioral development. By analyzing the results of interviews, models of present behavior selections and chronological changes will be built.

http://staff.aist.go.jp/kitajima.muneo/organic-self-consistent-field-theory/index.html

On the diversity of folk morality: Measuring classical positions in moral philosophy

Stephanie Müller
University of Granada, Spain

Bernd-Christian Otto
University of Heidelberg, Germany

Edward Cokely
Max Planck Institute for Human Development

Abstract: Moral psychology often oversimplifies moral philosophical debates into either deontological or consequentialist theories. The current research attempted to document greater variation in the extent to which participants used one of six core concepts to justify actions (i.e., appeals to religion, intuition, or one of four classical philosophical positions associated with Bentham, Hobbes, Kant, or Schopenhauer). Two hundred and fifty student participants (121 males) from the University of Granada were asked why a specific action would be "morally" correct or incorrect, which of the six concepts would be most adequate to justify the action, and whether they would behave similarly. Results indicated that participants agreed on a variety of diverse moral positions and that moral justification changed depending on context. The present research contributes to a growing body of work suggesting that different people apply different moral concepts to different life situations.

Linguistic Control in Monolingual and Bilingual Language Learners

James Bartolotti
Northwestern University

Viorica Marian
Northwestern University

Abstract: One of the difficulties in learning a new language is controlling competition from the language(s) already known. This interference resembles between-language competition in bilinguals, whose languages are both activated in parallel (Marian & Spivey, 2003). To test whether bilingualism confers an advantage in controlling competition during language learning, we compared monolinguals' and bilinguals' ability to manage interference from English while using a newly-learned language. Participants were taught an artificial vocabulary, then their eye-movements and mouse-movements were tracked in a visual world paradigm to assess activation of English competitors while processing the new language. We found that monolinguals, but not bilinguals, looked at interlingual competitors more than at controls, indicating greater interference from English. Similarly, monolinguals, but not bilinguals, demonstrated increased attraction to competitors compared to controls in mouse-movement trajectories. Results suggest that bilingual experience promotes efficient management of native language activation, with implications for linguistic control during language acquisition.

Intent discerning agent for more intuitive visualizations

Tera Marie Green
School of Interactive Arts and Technology (SIAT) Simon Fraser University

Steve DiPaola
School of Interactive Arts and Technology (SIAT) Simon Fraser University

Abstract: Today's visual interfaces are capable of representing large, semantically-complex datasets; user interaction of subsets is used to maximize display space. However, what data to provide and in what context requires decisions that are computationally and graphically expensive. We have built an autonomous, rule-based intelligent agent, which sits underneath a visualization, observes user behavior, determines user intent, and, based on what was learned, predicts future interest. The agent observes user manipulation and gathers interaction information continuously. For the purposes of demonstration, the agent sits underneath a shallow visualized hierarchical graph. The agent makes a determination about user intent through simple computations on its gathered interaction information and passes its decisions back to the visualization, which displays them to the user via ambient overlay. Future work will enable the agent to direct the interface as to which data to display and in which context to display it, enabling more intuitive human-computer collaboration.

A Difference in Working Memory Capacity among Chinese Speakers Using Different Computer Word Typing Methods

Jenn-Yeu Chen
National Cheng Kung University

Cheng-Yi Li
National Cheng Kung University

Abstract: Chen & Chuang (2008, CogSci) showed that Chinese speakers using phonology-based and orthography-based computer word typing methods (zhuyin vs. cangjie) displayed differential sensitivity in processing the phonological and the orthographic information of Chinese characters. The present study examined whether the zhuyin and the cangjie users might differ in their working memory (WM) capacities. Five verbal WM tasks and five visuospatial WM tasks were administered to 24 zhuyin and 23 cangjie users, whose typing speeds were comparable (53.7 and 53.2 characters per minute). Results show that the zhuyin users scored higher on the verbal WM tasks than the cangjie users, but the two groups performed similarly on the visuospatial WM tasks. The results suggest that general cognitive abilities like the WM capacity are related to the use of a technological artifact, consistent with the 'extended-mind' view of cognition proposed by Clark and Chalmers (1998).

Cognitive Arithmetic revisited: Effects of equation presentation format

Michael C. W. Yip
The Hong Kong Institute of Education

Abstract: The present study examined the cognitive processing of basic arithmetic. Thirty university students participated in a simple calculation verification experiment. In the experiment, a series of simple addition problems were randomly presented to each participant in one of the twelve experimental conditions (3 + 4 = 8) or (8 = 3 + 4) or (?O + ?l = ") or (" = ?O + ?l) or (?O + 4 = ") or (8 = ?O + ?l) or (12 = ?O + ¡ã) or (8 + 7 = 13). Participants were asked to verify whether the equation is correct or not by pressing a key as quickly and accurately as possible. The general pattern of results revealed that both the variables of equation presentation format and the numerical surface form influences the equation verification time but this was not the case for the variable of problem size.

The Capacity to Discover: Working Memory and the Ability to Use Self-Explanation to Discover Early Algebra Concepts

Marci DeCaro
Vanderbilt University

Bethany Rittle-Johnson
Vanderbilt University

Abstract: Prompting learners to generate explanations (self-explanation) can facilitate knowledge discovery and integration (Atkinson et al., 2000; Siegler, 2002) but does not always (Matthews & Rittle-Johnson, 2009). We examined whether greater capacity to retrieve problem-relevant information from memory (higher working-memory capacity) would enhance procedure discovery using self-explanation. Students (N=104; 2nd-4th graders) were instructed about math equivalence either before or after solving problems involving operations on both sides of the equal sign (e.g., 3+7+8=3+_). During problem-solving, some students self-explained answers, and some completed additional practice instead. Problem-solving accuracy was no different across the four conditions at posttest or retention. However, working-memory capacity moderated the effect of condition on retention. Self-explanation did not improve learning if instruction occurred first. However, when students solved problems before instruction, self-explanation benefited those students who were higher in working-memory capacity. Individual differences in learners' cognitive capacities may influence when self-explanation is beneficial as a discovery tool.

Number, Language, and Object Individuation

Lisa Cantrell
Indiana University

Linda B. Smith
Indiana University

Abstract: Recent research has suggested that the number of objects in a set affects the kinds of properties people attend to when speaking and categorizing (Barner & McKeown, 2005; Cantrell & Smith, 2009; Newstead & Coventry 2001). Here we asked whether number also affects the count-mass syntax that speakers use for common objects. Children ages 3-5 years were asked to look at pictures of common items (e.g., chairs, paper, soap) and label them. The images varied in number; children saw objects in sets of 2, 6 or 25. Results showed an effect of number on the kind of language children used. As the number of items increased, children became less likely to use individuating syntax, suggesting that objects in larger sets were seen less as individual entities and more as portions of a continuous mass. These results have theoretical implications for current ideas in number and object representation.

The Cognitive and Motor Performance of Children with Functional Articulation Disorders

Rong-Ju Cherng
National Cheng Kung University

Hung-Yi Chen
National Cheng Kung University

Jenn-Yeu Chen
National Cheng Kung University

Yung-Jung Chen
National Cheng Kung University

Abstract: Thirty children with functional articulation disorders (FAD) at the age of 4 to 8 years and age- and gender-matched typically-developing (TD) children were recruited to examine and compare their cognitive and motor performance. The Chinese versions of Peabody Picture Vocabulary Test and The Chinese versions of the Test of Non-verbal Intelligence-3rd edition were used for cognitive assessment and Bruininks-Oseretsky Test of Motor Proficiency, 2nd edition (BOT-2) and Movement Assessment Battery for Children (M-ABC) were used for motor assessment in the study. The results showed that children with FDA had significantly lower cognitive performance than TD children although their scores were all within the normal range. Children with FAD did not differ from TD children in the overall motor performance in either motor test. However, children with FAD showed worse performance than TD children in fine motor precision subtest. The performance of fine motor precision subtest was correlated with cognitive performance.

Keywords: Motor skill; Motor assessment; Functional articulation disorders; Developmental speech-language disorders

I let the music speak: a model of music perception that predicts speech segmentation

Geraint Wiggins
Goldsmiths, University of London

Abstract: To study the relationship between language and music, we apply a successful model of music perception to segmentation of phoneme streams into syllables.

Our model is a complex mixed-order multiple-feature n-gram model, with advanced back-off and smoothing capabilities. It has a long-term component, learned by unsupervised mere exposure, and a short-term component, exposed to the current stimulus; entropic weighting biases predictions between components. It was invented to simulate implicit learning of melodic pitch expectation, but it also predicts melodic segmentation, subjective expectation strength, associated neurophysiological activity, and aspects of expert musicologists' judgements. It is unusual as a Markov model in being multidimensional: it is capable of modeling sequences of objects with multiple features, using those features independently or together, and combining resulting mulitple predictions in a principled way.

Here, we model phoneme/stress sequences from the TIMIT speech resource metadata, using 2,342 phoneme sequences from US English, containing 21,427 syllables and 82,611 separate phoneme occurrences. We predict syllable boundaries by rise-picking in the resulting sequence of information-content values. The model predicts given segmentation with kappa=0.48, precision is .71, recall is .63, F1=.67, correct, using phoneme and stress only, over this surprisingly small learning corpus.

The results suhhest that our model may be a cross-modal model of preceptual sequence learning.

Peer Reviewing in Undergraduate Psychology Students

Joanna Salapska-Gelleri
Florida Gulf Coast University

Abstract: Students peer reviewed each others' grant proposals in an undergraduate psychology course. Unlike traditional in-class peer reviews, these were completed online and discussed during a class period. The review occurred multiple times throughout the year. Students had a chance to correct their papers and then resubmit for a second peer-review. As compared to students who either did not have the benefit of a peer review, only a single faculty grade, and ones who only received a one-time review, those students whose papers were reviewed multiple times received significantly higher marks on their final papers as judged by an outside reviewer. The benefit of peer-reviewing has been experimentally demonstrated in the past (Dunn,1996; Topping, 1998) but faculty have been reluctant to use this pedagogical tool due to time constraints. The current demonstration utilized a hybrid setting where students completed the reviews online and delivered the results in a brief in-class activity.

Training University Students on the Balance Scale Problem

Thomas Scaife
The Ohio State University

Andrew Heckler
The Ohio State University

Abstract: The cognitive development relating to solving balance scale problems has been studied in great detail, though effective training methods have not. In this study, students enrolled in a university-level introductory physics course were trained with examples from one of four conditions: when only one quantity (either weight or lever arm length) was different on each side of the balance, or when both weight and length were different, but the correct response corresponded to the side with either the greater length or greater weight. We found that when training involved the variation of only one quantity, participants were able to transfer learning to other configurations of weight and lever arm length. However, when training involved the variation of both quantities, participants were only able to answer correctly questions similar to those with which they were trained. These participants were unable to transfer learning to other configurations.

Verb tense and aspect in scene descriptions in a humanoid robot

Carol J. Madden
Stem Cell and Brain Research Institute, Lyon France

Stéphane Lallée
Stem Cell and Brain Research Institute, Lyon France

Peter Ford Dominey
Stem Cell and Brain Research Institute, Lyon France

Abstract: The present research implements a more human-like system of language in a humanoid robot model through improved use of grammatical constructions in described events. The present demonstrations show how the iCub humanoid robot is taught to recognize and use verb tense and aspect more appropriately. While hearing spoken verbal descriptions and watching visual displays of simple events involving objects moving on a table, the iCub robot begins to link the correct grammatical constructions with the appropriate information from its visual inputs and perceptual primitives. Thus, when it views a scene, the iCub robot is more likely to use the correct linguistic constructions to accurately describe it. In this way, past events are accurately described in the past tense, whereas ongoing events are described in the present progressive. Experimental evidence shows increased accuracy of scene descriptions for the iCub humanoid robot after learning phases involving live human-robot interactions. Supported by FP7 CHRIS & Organic, and ANR Comprendre and Amorces.

The importance of being present: The effect of a real or videotaped person on visual attention

Kaitlin Laidlaw
University of British Columbia

Tom Foulsham
University of British Columbia

Gustav Kuhn
Brunel University

Alan Kingstone
University of British Columbia

Abstract: How does visual attention operate in social contexts? Most research exploring this question has used picture or video paradigms that contain social stimuli but do not provide any opportunity for social interaction with the participant. In our study, we monitored participants' eye movements as they sat in a waiting room. Participants waited while either a confederate quietly completed a questionnaire, or while a video of the same confederate filling out the questionnaire was displayed on a nearby computer. Participants' fixation patterns of real vs. videotaped confederates indicated that participants actively avoided looking at the real person, while they looked more and for longer at the videotaped confederate. These results demonstrate the importance of social presence on visual attention and suggest that video- or picture-based studies of social attention are measuring performance and drawing conclusions that do not accurately reflect the real effect of social presence on visual attention.

Exploring Phonological Levenshtein Distance Effects in Auditory Lexical Decision

Lidia Suárez
National University of Singapore

Seok Hui Tan
National University of Singapore

Melvin J. Yap
National University of Singapore

Winston D. Goh
National University of Singapore

Abstract: Phonological similarity among spoken words is traditionally indexed by neighbourhood density (i.e., the number of words differing by a single phoneme from the target). However, density is of limited utility for long words, which have few or no neighbours. In this study, we explored the effects of phonological similarity and word-frequency on auditory lexical decision performance, using multisyllabic words with no neighbours and a new similarity metric called phonological Levenshtein distance (PLD20), which reflects the mean number of substitution, insertion, or deletion operations required to transform a word into 20 of its closest Levenshtein neighbours. Inhibitory effects of PLD20 were observed, where words with closer neighbours were recognised slower; importantly, these effects were present for only low-frequency words, replicating previous findings with other neighbourhood measures. The properties of PLD20 make it a promising new measure for quantifying the phonological distinctiveness of multisyllabic words in spoken word recognition research.

How is children's exploratory play influenced by evidence conflicting with their theory?

Tessa J. P. van Schijndel
Developmental Psychology University of Amsterdam

Maartje E. J. Raijmakers
Developmental Psychology University of Amsterdam

Abstract: Bonawitz, Lim, and Schulz (2007) demonstrated that children's exploratory play is affected by the interaction of their naïve theories and the evidence they observe. 6- and 7-year-olds were more likely to play with a balance toy when they observed evidence inconsistent with their balancing theory than when they observed evidence consistent with their balancing theory. The present study was set up to investigate how children's exploratory play is influenced by evidence that conflicts with their theory. Do children who observe inconsistent evidence play more systematically and make more informative comparisons than children who observe consistent evidence? 4- to 6-year-olds' naïve theories on shadow size were assessed with the shadow task (Siegler, 1981). 52 children with one specific naïve theory were selected and shown evidence consistent or inconsistent with their theory. Preliminary results show that the inconsistent group made more informative comparisons during free exploratory play than the consistent group.

Causal reasoning in decision making: A test of causal model theory of choice

Motoyuki Saito
Kwansei Gakuin University

Tsuneo Shimazaki
Kwansei Gakuin University

Abstract: Hagmayer & Sloman (2009) proposed causal model theory of choice based on causal Bayesian networks in order to provide comprehensive explanation for causal reasoning and decision making. The purpose of this study is to test predictions made by their theory. It considers that a deliberately chosen action is an intervention and that inferences based on choice are derived from structure of causal models in disregard of parameters of them (i.e., base rate, causal strength). In Experiment 1, we manipulated base rate and asked participants to infer probabilities conditional on deliberately chosen actions, interventions, and observations within common cause model. The estimates based on choices resembled that based on interventions and didn't reflect the differences in base rate. In Experiment 2, in which causal strength was manipulated within participants, the results revealed that participants neglected common cause. The differences in experimental situations between causal reasoning and decision making are discussed.

Gaze movement and language production when talking about events in live-recorded video clips

Monique Flecken
University of Heidelberg

Christiane von Stutterheim
University of Heidelberg

Mary Carroll
University of Heidelberg

Abstract: The paper deals with the interrelation between patterns in gaze movement when watching dynamic video clips and what is mentioned at what point when talking about events. This interrelation was investigated with respect to dynamic, live-recorded videoclips depicting everyday situations. Speakers of three languages were asked to view the clips and tell what is happening. Attention distribution to different aspects of the clips (causative actions) were measured in 2 identified areas of interest: the area where the agent is located and the area in which the entity acted upon is located. Contrary to studies on the production of single words or clauses relating to pictures (eg. Meyer & Dobel 2003), gaze movement to the areas of interest and the time at which they are mentioned are not directly linked, given real time presentations. Factors that drive patterns of attention and mention over time cross-linguistically will be presented.

Fluency and cognitive control in judgment: Influences of memory and elaborative encoding

Paula Parpart
Max Planck Institute for Human Development

Edward T. Cokely
Max Planck Institute for Human Development

Abstract: Recent research has documented some surprising relations between cognitive control and fluency use in consumer judgments. Theoretically, the observed link is explained by differences in elaborative heuristic search: More elaborate encoding leads to more vivid memory representations, which changes the ease of cognitive processing along with one's subjective basis for judgment. The current research presents new evidence in a stock profit estimation task, documenting a relationship between better memory and the use of fluency. Specifically, participants were provided with fictional company names that varied in their ease of pronunciation, and were asked to judge past company profits. Participants with a higher reliance on the company name pronunciation for their judgments were found to later have higher recall and recognition of company names. Results are consistent with an elaborative heuristic search account of the unusual relationship between heuristics and cognitive control. Implications for dual process theories are discussed.

Bridging the Implementation Gap: From Sensorimotor Experience to Abstract Conceptual Knowledge

Anna Koop
University of Alberta

Leah Hackman
University of Alberta

Rich Sutton
University of Alberta

Abstract: We develop a sensorimotor perspective on conceptual knowledge, paying particular attention to the imperatives of an artificial system. Motivated by the gap between low-level sensorimotor experience and human-level conceptual knowledge, we contrast experiential knowledge with the classical notion of concepts. We discuss three ways in which experience and classical concepts differ: experience is dynamic rather than static, subjective rather than objective, and composed of minutiae rather than compact abstractions. We present a mechanism for abstracting from experience and show how it might recover some of the benefits of concepts while addressing some of the difficulties of classical theory. Finally, we implement a simple example which illustrates first steps towards bridging the gap between sensorimotor experience and high-level concepts.

An examination of learner control during web-based instruction

Jessica Federman
Cornell University

Ryan Morris
Cornell University

Lisa Dragoni
Cornell University

Abstract: We investigated the efficacy of an asynchronous computer-based educational learning platform called VideoNote. Key features of the technology include: video streamed content, searchable, detailed, time-linked text notes of learning topics that are hierarchically ordered, detailed analytics that timestamp and track which topics within the video students are viewing, and a mechanism for students to rate and mark the difficulty of certain topics to ease subsequent review. Students used the tool as a supplement to course lecture and for distance learning. Using a randomly assigned, between subject sample (N=77), students who used Video Note improved their exam grades by 9.5

The Linguistic Distribution of Relational Categories

Micah Goldwater
Northwestern University

Jon Willits
University of Wisconsin

Abstract: Behavioral research has distinguished relational categories (e.g., barrier), which are defined by relations among entities, from feature-based categories (e.g., vegetable), which are defined by sets of descriptive features intrinsic to entities (e.g., Gentner & Kurtz, 2005; Rein, Goldwater, & Markman 2010). Corpora research has demonstrated that category structure is reflected in their distribution in natural language texts (e.g., Willits, 2009). The current project connects these two lines of research by examining the distributions of both kinds of categories. Findings include: Feature-based categories' distributions are more similar to each other than relational categories' are to each other. Relational categories appear in more diverse contexts than feature-based; however relational categories are "anchored" by a single frequent collocate to a greater degree than are feature-based categories. We discuss relations between corpus measures and behavioral ratings and consider theoretical implications for category representation.

Multisensory stimuli improve numerical matching abilities of preschool children

Kerry Jordan
Utah State University

Joseph Baker
Utah State University

K.S. Rodzon
Utah State University

Abstract: We previously showed that giving young infants synchronous, multisensory information about number increases the precision of their numerical discriminations. Does intersensory redundancy also facilitate numerical learning in older children? Twenty-four preschool children (3-5 years) played a number matching game on a touch-screen computer. On each trial, children counted a sample numerosity whose elements were presented serially. On some trials, the sample was visual, on some auditory, while on still others audiovisual. Children were then presented with two choices and asked to touch the numerically matching array. Data support the idea that intersensory redundancy improves children's numerical estimations. Multisensory information may be more salient than unimodal information, which could better recruit attention and result in more precise learning and remembering than when such information is presented to only one modality. Results should spur future research into whether such multisensory facilitation can be harnessed for educational benefit in the early mathematics classroom.

Wayfinding Tasks and Heuristics

Simon J. Buechner
Center fo Cognitive Science

Christoph Hölscher
Center fo Cognitive Science

Abstract: Wayfinding is the process of determining and following a route between an origin and a destination (Golledge, 1999). Research on wayfinding has been conducted by means of a variety of tasks making it difficult to compare research results among each other. Wiener, Büchner & Hölscher (2009) proposed a taxonomy of wayfinding tasks that classifies the tasks with respect to external constraints and the type of knowledge that is required to solve the task. Besides these two factors, heuristics play an important role for wayfinding. A number of heuristics for wayfinding tasks are documented in the literature, including but not limited to work by the authors. We will first present the taxonomy and then relate observed heuristics to the tasks. The relationship of tasks and heuristics will then be discussed with respect to superordinate concepts.

Auditory distraction during semantic processing: Data and a model

Philip Beaman
University of Reading

John Marsh
Cardiff University

Dylan Jones
Cardiff University, University of Western Australia

Abstract: An experiment is reported demonstrating how free recall of visually-presented, categorically-related lists of words is disturbed by the presence of auditory distracters which subjects were instructed to ignore. Data show that auditory distracters from the same category as the to-be-recalled items produce the most disturbance to recall and result in the most intrusion errors. Additionally, the points at which these intrusion errors occur differ dependent upon whether recall is written or spoken. A variant of the SIMPLE (Scale Invariant Memory and Perceptual Learning) model (Brown, Neath & Chater, 2007) is fit to these data and the modifications necessary to achieve this fit are discussed. It is concluded that intrusion errors are not random but are dependent upon a weighted combination of the semantic and temporal overlap between the to-be-recalled and to-be-ignored material in semantic processing tasks and free recall

How to Foster the Integration of Text and Diagrams: An Eye Tracking Study on the Use of Signals in Multimedia Learning

Katharina Scheiter
Knowledge Media Research Center

Alexander Eitel
Knowledge Media Research Center

Abstract: Learners studying text and diagrams (multimedia) often fail to integrate information from both sources. Hence, signals that make explicit the relation between both representations should improve understanding. The current study investigated which changes in information processing can explain improvements in comprehension due to signals. In an eye tracking study 35 students learned about the functioning of the heart. In a no-signals condition, a text and diagram were presented unaltered. In the signals-condition, correspondences between the representations were highlighted by means of labels, color coding, and deixis. Signals improved understanding of text-diagram correspondences and guided attention towards diagrams. Moreover, diagrams were fixated earlier in the signals-condition. A mediation analysis showed that these changes in visual attention completely explained the effect of signals on comprehension. Hence, signals improve learning from text and diagrams by fostering learners' early reference to diagrams and by increasing the amount of attention devoted to them.

Cognitive Modeling Repository

Jay Myung
Ohio State University

Mark Pitt
Ohio State University

Abstract: Quantitative modeling has contributed substantially to the advancement of the cognitive sciences. Papers introducing and testing cognitive models regularly appear in the top journals. The growth and success of cognitive modeling demonstrate why modeling itself should be a primary quantitative method in the researcher's toolbox. Yet this method of scientific investigation remains under-utilized by the research community at large, in part because of the hassles in obtaining data sets to model and the difficulties in implementing models. The goal of this project is to assist scientists in their cognitive modeling efforts by creating an online repository containing data sets that can be modeled and the cognitive models themselves. The current state of the project and future plans will be presented. Funded by the Mathematical Modeling of Cognition and Decision of the Air Force Office of Scientific Research.

The hindsight bias in temporal predictions of animated automobile accidents

Dustin Calvillo
California State University San Marcos

Dayna Gomes
California State University, Los Angeles

Abstract: The hindsight bias occurs when people judge the outcome of an event as more predictable than it actually was before it occurred. The current experiment examined the hindsight bias in animations of automobile accidents. Participants viewed eight animations in one of two conditions. Those in the foresight condition were told that some animations contained accidents and were instructed to stop the animation when they were certain that an accident would occur. Those in the hindsight condition were told that all animations contained accidents and viewed each animation twice. They viewed the entire animation first. On the second viewing, they stopped the animation when they thought a naïve viewer would be certain an accident would occur. Those in the hindsight condition stopped the animations closer to when the accident actually occurred than those in the foresight condition, demonstrating a hindsight bias.

Strategies for multitasking: An fMRI study of individual differences in multitasking ability

Winston Jones
Mississippi State University

Jarrod Moss
Mississippi State University

Stephanie Doane
Mississippi State University

Abstract: Multitasking is the ability to interleave tasks that vary in duration and the demands placed on cognitive resources. The Abstract Decision Making (ADM) task correlates with performance in real-world multitasking environments (Joslyn & Hunt, 1998). This study uses the ADM to measure multitasking ability. Our hypothesis is that use of consistent and effective task strategies can partially explain individual differences in multitasking ability. This hypothesis was investigated using behavioral and fMRI measures. The behavioral results show a correlation between strategy consistency and individual differences in ADM performance and support the strategy hypothesis. The fMRI results suggest that executive control areas of the brain are involved in task performance, but that activation in these areas alone does not explain differences in ADM performance. However, activation in other areas, including temporo-parietal regions, is correlated with individual differences in performance.

Facilitation in Second Language Word Meaning Evaluation from Masked Primes

Robert Zheng
University of Utah

Fernando Rubio
University of Utah

Dan Woltz
University of Utah

Abstract: An experiment with 90 students learning Spanish as a second language was conducted to investigate 1) the ability of cross-language primes to facilitate semantic decisions which required L2 meaning retrieval of recently learned words, and 2) the relationship between prime facilitation and prime awareness. The priming task consisted of a Spanish word presented in the upper third of the computer display, followed by two pictures in the lower left and right corners of the display. Our results indicated facilitation in vocabulary response time was substantial at both 67 and 83 ms prime exposure duration. There was no dependable facilitation at 50 ms exposure. It was concluded that automatic prime effects, independent of strategic processing of primes, might offer an important tool for reducing the working memory load inherent in initial L2 acquisition which could allow greater opportunity for the acquisition of semantic and structural elements of the new language.

Physical design tools support and hinder innovative engineering design

Jooyoung Jang
LRDC (Learning Research and Development Center), University OF Pittsburgh

Christian Schunn
LRDC (Learning Research and Development Center), University of Pittsburgh

Abstract: Engineers use various physical tools (e.g., computers, smartboards, notes, and prototypes) to support their design work. To understand cognitive processes underlying the innovative design process and to reveal the characteristics of innovation-supporting environments, we examined the pattern of tool use in 43 interdisciplinary engineering design teams enrolled in a full-semester Product Realization course. Teams worked all semester on a single project, with each team being assigned a different industry-sponsored project. Group meetings were video-recorded. Team success was measured in terms of meeting client requirements, and groups were divided into high, medium, and low success. Low success groups (relative to high and medium) used smartboards and prototypes less often and paper notes relatively more often. The results suggest that more successful groups focused on group discussion, supported by large sharable screen, and transitional thinking from abstract ideas to concrete products.

Seductive Images and Metacomprehension of Science Texts

Allison Jaeger
University of Illinois at Chicago

Jennifer Wiley
University of Illinois at Chicago

Abstract: Although the intention behind including illustrations alongside expository text is generally to increase student motivation, interest, or understanding, images do not always have beneficial effects. Generally, students tend to have poor comprehension when learning from expository science texts. Further, they also tend to have poor metacomprehension accuracy, meaning they are not able to differentiate what they have understood well, from what they have understood poorly. In the current study, the inclusion of either conceptual or seductive images actually increased comprehension as compared to a no-image condition. The inclusion of images also tended to increase readers' interest in the texts. However, including seductive images decreased metacomprehension accuracy compared to the no-image condition. This suggests that seductive images may provide readers with a false sense of fluency or understanding which could potentially undermine effective self-regulation and studying behaviors.

Emotion and association-memory

Christopher Madan
University of Alberta

Christine Lau
University of Alberta

Jeremy Caplan
University of Alberta

Esther Fujiwara
University of Alberta

Abstract: Emotional items are remembered better than neutral items. It is unclear how this extends to memory for associations involving emotional items. We manipulated the pairings of emotional and neutral words and direction of cued-recall probes. Pairs were pure (EMOTIONAL-EMOTIONAL, NEUTRAL-NEUTRAL) or mixed (EMOTIONAL-NEUTRAL, NEUTRAL-EMOTIONAL). We asked whether emotion would enhance association-memory, independently of its effects on item-memory (e.g., target retrievability). We fit the data with a probabilistic model to obtain estimates of how emotion influenced cued recall depending on emotionality of the target or probe (item-memory effects), or relationship between constituents (association-memory effect). In a follow-up we replaced emotional words with taboo words to exaggerate the manipulation. Findings suggest that mildly emotional words reduced memory for the associations whereas taboo words neither impaired nor enhanced memory for the associations. Consistent with other recent findings, our results suggest that emotional enhancement of memory effects do not extend to relational memory.

The Effects of Alcohol Use on Creative Problem Solving

Andrew Jarosz
University of Illinois at Chicago

Gregory Colflesh
Georgia Tech

Jennifer Wiley
University of Illinois at Chicago

Abstract: Though creativity is highly valued in many disciplines, ranging from the fine arts to the natural sciences, little is known concerning its underlying causes and mechanisms. In particular, the problem solving literature has long sought an explanation for the processes underlying creative problem solving. The present study tested the commonly held notion that alcohol use increases an individual's creativity, using a problem solving paradigm. Participants completed the Remote Associates Task (RAT), while either sober or intoxicated to a blood alcohol content of approximately .07. It was found that intoxicated individuals outperformed their sober counterparts. These results are interpreted as evidence that alcohol use leads to a diffuse attentional state, which in turn can benefit creative problem solving.

Tapping into Student Knowledge about Science Systems

Jodi Davenport
WestEd

Edys Quellmalz
WestEd

Mike Timms
WestEd

Abstract: What do students know about science? If students have a deep understanding of a science system they should understand core concepts and be able to use their knowledge to make inferences and carryout scientific investigations. Thus, the challenge of science assessment is to develop tasks that not only tap into declarative and procedural knowledge, but also schematic and strategic knowledge that allow students to demonstrate the ability to reason through complex systems and use existing knowledge to generate new understandings.

The current study investigates the range of knowledge and skills addressed by existing middle school science assessments administered at state, national and international levels. We conducted an analysis of released and sample items related to ecosystems and chemistry from more than 30 exams. We will present the results of our analysis and a framework that characterizes the types of knowledge likely to be elicited by different types of assessment items.

Preschoolers writing of multidigit numbers: From an additive to multiplicative representational system?

Sandra Street
Indiana University

Richard Prather
Indiana University

Cody Stitzel
Indiana University

Linda Smith
Indiana University

Kelly Mix
Michigan State University

Abstract: Possible systems for representing multi-digit numbers include additive systems (such as roman numerals) in which unique symbols that represent different amounts are written in strings and multiplicative systems in which the same digit represents different multiples of different set sizes depending on place, as in the base-10 place value system. Multiplicative systems such as this depend on place holders (zero). Preschool children (4 to 6 years of age) prior to any explicit training with multi-digit representations were asked to write 2 and 3 digit numbers. Children's responses were collected and coded for a variety of features. Young children on their own seem to develop an additive idea about how to represent multidigit numbers, that preserves left to right place value, and uses 0 to represent group size (e.g., two hundred twenty seven = 20027 or 200207).

Knowledge about the role of illustrations on motivation for reading

Hideaki Shimada
Shinshu University

Abstract: Suppose that you grab an instructional manual, you may glance over some of the pages to determine whether it looks interesting enough to read carefully. Our past study demonstrated that illustrations enhance readers' motivation in the first few seconds in text comprehension. This study examined the knowledge about the role of illustrations and the relationships to the motivation effect. In the first phase of this experiment, participants were required to glance over a page of a disaster prevention manual for two seconds and to answer questions as to motivation, such as "Did the page motivate you to read?" In the next phase, they were required to evaluate a subjective amount of information and comprehension efficiency for each page. Results showed that illustrations didn't increase the subjective amount of information but subjective efficiency, which enhanced motivation to read.

Effects of Self-Explanation and Prompts Depend on the Students' Need for Cognition

Kyung Soo Do
Sungkyunkwan University

Hyo-hee Lee
Sungkyunkwan University

Hanna Kim
Sungkyunkwan University

Abstract: Determining what to explain and generating explanations have to be satisfied to do successful self-explanations. Providing prompts helps the first part, and students' having high Need for Cognition can help the whole process. Seventy four adult vocational school students participated in a three factors (ie., self-explanation, prompt, Need for cognition) between subjects experiment. The results were different depending on the tasks and the students' level of Need for Cognition. In memory tests, asking to do self-explanations or giving prompts helps students of low levels of Need for Cognition. However, asking to do self-explanations or giving prompts was not effective to students of high level of Need for Cognition. In tests measuring understanding, the main effect of Need for Cognition was significant, and three factors interaction effects was marginally significant. Giving prompts helped students of high Need for Cognition understand better. The results were interpreted in terms of cognitive load.

Engineering Models of Human Behavior

Spyridon Revithis
University of New South Wales

Abstract: The level at which a computational cognitive model provides explanations of phenomena is often unclear, especially when there is no sufficient distinction between behavior and cognition. It has been shown that human behavior is amenable to SOM modeling aiming at compressed classification and prediction. Reducible to an engineering level this modeling approach offers no associations to biologically plausible cognitive mechanisms if there is no explicated claim of correspondence between the mechanisms used and the biological mechanisms that drive behavior. At the statistical level no claim of biological plausibility is always a prerequisite for the validity of the model.

Case studies of behavioral SOM models, conducted by the author, demonstrate and support the proposition that engineering models are not by default brain cognitive models or causal models of human behavior until appropriate associations have been established; prior to the latter, SOM models merely suggest algorithmic engineering solutions to challenging statistical problems.

Attention for Action: Attentional Modulation by the Hands

Holger Schultheis
Universität Bremen

Laura Carlson
University of Notre Dame

Richard Abrams
Washington University

Abstract: Actions by the hands may be represented with respect to a spatial reference frame that is centered on the hands, thereby organizing the surrounding space into regions and possibly influencing the allocation of attention to these regions. Indeed, it is more difficult to disengage visual attention from a visual search display when responses are made by hands that are grasping the display screen than by hands that are on the lap. In the current study we assess whether the attentional modulation is due to proximity to the screen or orientation of the hands. Specifically, we contrast conditions that dissociate proximity (hands near or far from the screen) and orientation (hands spanning the display with responses toward the stimuli, hands spanning the display with responses orthogonal to the stimuli, and hands near to but not spanning the display). The results indicate that both proximity and orientation influence the allocation of attention.

The effect of conventionality and aptness on suppression of metaphor-irrelevant meaning

Tomohiro Taira
Kyoto University

Takashi Kusumi
Kyoto University

Abstract: Metaphor comprehension needs a categorization process. The categorization process includes two different functions to the vehicle of metaphor: one is the activation of metaphor-relevant meaning, and the other is the suppression of metaphor-irrelevant meaning. Some previous studies showed that the conventionality of the vehicle and the aptness of the metaphor affect whether the categorization process happens. But they did not clarify whether these factors affect the activation and the suppression. In this point, our past study showed the evidence of the effect of the conventionality and the aptness on the activation of metaphor-relevant meaning. And in this study, we investigated the effect of these factors on the suppression of metaphor-irrelevant meaning. One experiment that consisted of a metaphor-priming task and a meaningfulness decision task showed that after metaphor comprehension, the vehicle suppress the metaphor-irrelevant meaning regardless both of the conventionality of the vehicle and the aptness of the metaphor.

A Categorization of Face Recognition Deficits in Congenital Prosopagnosia

Rainer Stollhoff
Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany

Jürgen Jost
Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany Santa Fe Institute, Santa Fe, NM, USA

Ingo Kennerknecht
Institute of Human Genetics, Westfälische Wilhelms-Universität, Münster, Germany

Abstract: Congenital prosopagnosia refers to a lifelong impairment in face identification that is present from birth. In contrast to acquired prosopagnosia, where the deficit is due to brain damage, people with congenital prosopagnosia never evolve a functional face recognition system in the first place but develop compensatory processing strategies to overcome their deficit. In order to assess both deficit and compensatory processing in congenital prosopagnosia, we conducted a series of experiments with a large group of 15 prosopagnosic participants. The tests administered covered different aspects of face and object recognition: Identification under unlimited and restricted viewing times, the influence of rotation in depth, recall of target stimuli after one year, and recognition of famous faces. Based on the test results, we propose a categorization of congenital prosopagnosia along the lines of an apperceptive, associative or amnestic deficit and discuss this categorization scheme with respect to potential neuroanatomical differences underlying the deficit.

Focusing on the Intermediate Event Makes the Chain Structure More Learnable

Kyung Soo Do
Sungkyunkwan Univ.

JaeHyuk Choi
Sungkyunkwan Univ.

Abstract: The chain structure has not been easily inferred when three events are causally related like a chain (A causes B, and B in turn causes C). We hypothesized that focusing on the intermediate event make the chain structure more learnable, if the causal structure is locally computed (Fernbach & Sloman, 2009). In Experiment 1, focusing is manipulated by informing which event is important. Participants who were informed that the intermediate event (B) is important inferred the chain structure (61

Effects of Physical Structure and Creators' Intentions on Judgments of Function

Kyung Soo Do
Sungkyunkwan University

Kyuhee Kim
Sungkyunkwan University

Abstract: Affordances from physical structure and creator's intentions have been regarded as the determinants of an object's perceived function. Two experiments were conducted to explore the conditions when creator's intentions affected judgments of function. Participants rated the plausibility of the two possible functions after an outline drawing of an object and verbal description of creator's intentions were presented to them. When one of the two functions was easily derived from the drawing (Experiment 1), the function was judged more plausible regardless of creators' intention. However, the difference between the plausibility ratings of the two functions got smaller when intention did not match the dominant function. When the two functions were equally likely from the drawing (Experiment 2), creators' intentions affected the perceived function of an object. The results of two experiments suggested that creators' intentions affect the judgment of function only when affordances are not strong.

A Computation Model synthesizing the Rule based and Experience based Cognitive Processes of Chinese Characters

Sau-chin Chen
Tzu Chi University

Jon-Fan Hu
National Cheng Kung University

Ping Li
Pennsylvania State University

Abstract: A new model organizing two self-organizing maps (SOM) is presented here for synthesizing the rule based and experience based cognitive activities in support of processing the sounds and forms of Chinese characters. One SOM is constructed to form the phonological representations generated from the Chinese PatPho (Zhao & Li, 2009); and another SOM is aimed to produce the orthographic representations according to the Chinese character coding system (Chen, Zhao, & Li, 2009). The two SOMs are connected by an associative learning algorithm in shaping the mapping between the phonological and form patterns. The mappings belong to the phonetic radical and characters with this radical are suggested to construct the Chinese phonetic-and-sound relations. This presentation summarizes the tests on a specific group of characters representing the typical phonetic-and-sound relations. Insights of this model would reveal if these mappings are the origins of the rule based and experience based cognitive activities respectively.

Virtual Brainstorming: Avatar Visibility and Group Size

Thomas Ward
University of Alabama

Matthew Guerdat
University of Alabama

Beverly Roskos-Ewoldsen
University of Alabama

Abstract: Dyads and triads of college students brainstormed via computer based text chat rather than face-to-face. In one condition, participants were represented as 3D avatars in a virtual space. In another, no avatars were visible. It was hypothesized that triads would outperform dyads due to facilitation from a third perspective, and that having avatars visible would increase participants' sense of presence and task involvement, resulting in better performance. Participants were given standard brainstorming instructions that emphasized withholding criticism, generating unusual ideas, generating as many ideas as possible, and piggybacking on others' ideas. The task was to generate ideas as for how to improve the university. Participants were surveyed after the problem solving task regarding their sense of presence and task involvement. Triads outperformed dyads but there was no facilitation of presence or idea generation from having visible avatar representations. Implications for the believability of virtual environments are discussed.

Semantic richness modulates early word processing within left-lateralized visual brain areas and enhances repetition priming

Milena Rabovsky
Humboldt University at Berlin

Werner Sommer
Humboldt University at Berlin

Rasha Abdel Rahman
Humboldt University at Berlin

Abstract: Repetition priming has been shown to be modulated by prior knowledge about structural regularities (Stark & McClelland, 2000). Here, we examined influences of higher-level semantic knowledge, more specifically the richness of semantic representations, on repetition priming. The EEG was recorded while twenty-four participants performed a visual lexical decision task on 160 words and 160 pseudowords. Within the word stimuli, we orthogonally manipulated two measures of semantic richness, namely the number of semantic features (McRae et al., 2005) and free associations (Nelson et al., 2004); the whole stimulus set was presented twice. The number of semantic features modulated the amplitude of the posterior N2 component over left occipito-temporal areas. This effect arose only about 30 ms after the onset of lexicality effects on left-lateralized N170 amplitudes, presumably reflecting visual word form processing within the fusiform gyrus. Thus, word form and meaning are accessed in rapid succession within left-lateralized visual brain regions. Repetition priming was consistently enhanced for words with many semantic features in both performance and ERP data, suggesting a role for feature-based semantic richness in word repetition priming.

The role of stimulus familiarity in non-linguistic sequence learning

Jennifer A. Sturm
Northumbria University

Kenny Smith
Northumbria University

Abstract: Recent experiments suggest that the mechanisms employed in language learning are also involved in sequential learning of non-linguistic stimuli and are therefore domain-general. However, the non-linguistic materials typically used in these experiments (e.g. Kirkham, Slemmer & Johnson, 2002) do not adequately replicate the internal complexity of words in language. Furthermore, stimulus familiarity appears to play a crucial role (Saffran, 2007). We explore both factors, investigating the acquisition of non-adjacencies (ubiquitous in language) in non-linguistic sequences. Crucially, the black and white matrix patterns we use are orthographically matched with an artificial language, replicating the componential re-use of elements in language or speech (Sturm & Smith, 2009). Prior to the sequence learning experiment, participants were familiarized with individual patterns. The results show that although participants became familiarized with the patterns, they were unable to identify the grammar underlying sequences of those patterns. These findings allude to an important role of domain-specific expectations.

The Development of Numeracy: Fingers Count!

Marcie Penner-Wilger
Franklin & Marshall College

Lisa Fast
Carleton University

Jo-Anne LeFevre
Carleton University

Brenda L. Smith-Chant
Trent University

Sheri-Lynn Skwarchuk
University of Winnipeg

Deepthi Kamawar
Carleton University

Jeffrey Bisanz
University of Alberta

Abstract: Butterworth (1999) proposed that three component abilities support the development of numeracy: subitizing, finger gnosis, and finger agility. We assessed these abilities in children in Grade 1 (N = 144) and followed them to Grade 2 (n = 102). In Grade 1, subitizing and finger gnosis were related to children's number system knowledge and all three component abilities were related to calculation skill. Using cluster analysis, we identified three groups of children based on skill profiles across subitizing, finger gnosis, and finger tapping. One group had strong subitizing, finger gnosis and finger agility – they also had good numeracy performance both concurrently in Grade 1 and longitudinally in Grade 2. Two other groups both performed worse than the highly-skilled group on numeracy measures in Grade 1 and Grade 2; these two less-skilled groups showed strikingly different patterns of performance on number comparison, a task designed to assess the representation of number.

Location, Location, Location: Environmental constraints on interpreting spatial terms

Kevin Mickey
University of Notre Dame

Laura Carlson
University of Notre Dame

Scott Freundschuh
University of Minnesota, Duluth

Abstract: The environment can constrain the way we think and act within it. Such an influence has been largely ignored within the domain of spatial language, which has largely focused on objects and their identities, independently of the environments in which they occur. To investigate whether the environment also has an influence, we instructed participants to place a located object either near or far from a reference object within survey perspectives of manipulated 3D environments. When a geographical feature in that environment was present and had meaningful semantic content, it systematically altered the distance, direction and orientation of the placements, with these alterations well beyond the range expected based on a geometric definition of the spatial term. This environmental influence is consistent with a situated view of cognition.

Goals and the Perception of Distance and Time in Virtual Spaces

Angie Johnson
University of Northumbria at Newcastle

Kenny Coventry
University of Northumbria at Newcastle

Emine Mine Thompson
University of Northumbria at Newcastle

Abstract: Individuals rarely walk in an environment without a purpose. However , the influence of goals on the development of 'cognitive maps' has largely been ignored. The results of t wo experiments are reported that investigated the role of both goals and environmental structures on me mory for distance and time in Virtual Reality (VR) environments. Experiments 1 and 2 compared the effect of goals varying in urgency and desirability , on memory for distance and time in VR environments with (Experiment 1) and without turns (Experiment 2). Striking effects of goals were found for memory for distance and time in both environments. Experiment 3 examined the origins of these goals effects through the use of physiological measurement and mood scales . Results show that goals influence distance estimation as a function of the degree of urgency experienced in situ, and not as a function of overall mood state or arousal they induce.

Artificial Cognitive Systems for Human-like Situation Awareness Ability

Soo-Young Lee
KAIST

Abstract: The Artificial Cognitive Systems (ACS) will be investigated for human-like functions such as vision, auditory, inference, and behavior. Especially, computational models and artificial HW/SW systems will be devised for Proactive Knowledge Development (PKD) and Self-Identity (SI). The PKD model provides bilateral interactions between robot and unknown environment (people, other robots, cyberspace). For the situation awareness in unknown environment it is required to receive audio-visual signals and to accumulate knowledge. If the knowledge is not enough, the PKD should improve by itself though internet and others. For human-oriented decision making it is also required for the robot to have self-identify and emotion. Finally, the developed models and system will be mounted on a robot for the human-robot co-existing society. Based on the computational models of PKD and SI, we would like to build functional modules for Knowledge Represenation, Knowledge Accumulation , Situation Awareness, Decision Making, and Human Behavior.

Comprehension and a Complex Task: A construction-integration study of individual performance in a non-routine task situation

Paul Ladny
Mississippi State University

Jordan McGuire
Mississippi State University

Randy J. Brou
Navy Personnel Research, Studies, and Technology

Stephanie M. Doane
Mississippi State University

Abstract: Comprehension is the ability to relate background knowledge to incoming information to build a "situation model" (Kintsch, 1998). The ConstructionIntegration (C/I) architecture of comprehension has been shown to predict individual performance on complex but routine tasks (e.g., Doane & Sohn, 2000). This study tests the ability of the architecture to explain and predict nonroutine (unexpected) instrument flight performance in aviation piloting. The behavioral results indicate significant differences in individual pilot ability to detect and recover from unexpected instrument failures as a function of piloting expertise. However, expertise is not the sole predictor of performance. The computational experiments indicate that the C/I architecture explains and predicts a significant amount of individual pilot performance. Overall the findings suggest that comprehensionbased processes play a significant role in understanding human performance in unexpected situations.

Competitive Routes to Belief and Their Impact on Future Learning

Carlos R. Salas
University of Illinois at Chicago

Thomas D. Griffin
University of Illinois at Chicago

Abstract: Models of belief formation and conceptual change have begun to allow for affective preferences and motives to supplement normative processes, such as reasoning and coherence evaluation (Kunda, 1990; Thagard, 2006). Griffin (2008) goes a step further in arguing that affect can be a separate competitive route to belief formation that could prevent these normative processes from taking place. One implication is that affect-based beliefs will lack the conceptual coherence presumably produced by engaging in reasoning and coherence evaluation processes. Borrowing from the expertise literature (e.g., Ericsson & Kintsch, 1995), this reduced conceptual coherence should hinder one's ability to represent new domain-related information. We present a study showing that a person's route to belief (evidence or emotion based) predicts their comprehension of belief-relevant information, even after controlling for several general and domain-specific individual differences in knowledge, skills, dispositions, motivation, and task engagement.

Unique and Additive Effects of Self-Explaining and Contrasting Cases on Learning Fraction Division

Shanta Hattikudur
University of Wisconsin - Madison

Pooja G. Sidney
University of Wisconsin - Madison

Martha W. Alibali
University of Wisconsin - Madison

Abstract: Studies show that both contrasting cases and self-explanation are useful in promoting procedural and conceptual knowledge when learning from worked examples. It is not clear whether these instructional tools draw on similar problem-solving skills or provide unique support for learning. The purpose of this study is to assess whether self-explanation and contrasting cases are more effective when combined than when applied separately. To the extent that these processes can be manipulated separately, we hypothesize that the effects of both instructional techniques are unique, and together will lead to greater knowledge gains.

Participants completed a pretest, assessing their ability to divide fractions, before engaging in a problem study session and procedural lesson about fraction division. Participants then completed the posttest, which included procedural and conceptual transfer measures. Pilot data suggest that there are differences between the benefits of self-explanation and contrasting cases on procedural and conceptual learning.

When dog is more wolf than bone: Computational and electrophysiological evidence for featural organization of semantic memory

Sarah Laszlo
Carnegie Mellon University

Blair Armstrong
Carnegie Mellon University

Joseph MacInnes
Oculus Info. Inc.

David Plaut
Carnegie Mellon University

Kara Federmeier
University of Illinois, Urbana-Champaign

Abstract: Semantic space algorithms account for human performance in semantic tasks via knowledge representations derived from the analysis of large text corpora. The N400 Event-Related Potential (ERP) component is thought to reflect automatic access of the same lexical-semantic information. We trained LSA (Landauer & Dumais, 1997) and HAL (Lund & Burgess, 1997) on a random selection of Wikipedia articles and compared the algorithms' performance at predicting the similarity between N400 waveforms elicited during reading. HAL was best at explaining the ERP data, suggesting that its representations—thought to be more semantic-featural than lexical-associative in nature—are most similar to those automatically accessed during N400 processing. These results are consistent with findings that, although the N400 is sensitive to lexical relationships, it seems to represent access of information arranged primarily by semantic features. Preliminary evaluations of other algorithms (e.g., COALS, Rohde, Gonnerman & Plaut, submitted) further support this conclusion.

Hindsight bias in judgments of others' performance on inattentional blindness tasks

Alan Penaloza
California State University San Marcos

Dustin P. Calvillo
California State University San Marcos

Richard Brooks
California State University San Marcos

Dayna M. Gomes
California State University, Los Angeles

Abstract: The hindsight bias occurs when people judge the outcome of an event as more predictable after the event has already happened. Participants (N = 45) completed two inattentional blindness tasks, in which an unexpected object appeared while participants performed a primary task. Participants were asked if they noticed the unexpected object. They were then shown the unexpected object and asked how many people (out of 100) they thought would be able to see it. For both tasks, those who saw the objects (Task 1: n = 5; Task 2: n = 19) judged that more people would see it (overall: M = 39.8, SD = 23.0) than those who did not (overall: M = 17.1, SD = 14.6). This finding is consistent with a hindsight bias: those who experienced seeing the object thought that others would see it, while those who did not, thought others would not.

A decision science blind to decision procedures would be "unfair": The effect of decision process on decision-outcome satisfaction and subsequent choice in a performance environment

Daniel DeCaro
Miami University (MU)

Joseph Johnson
Mimia Univeristy (MU)

Abstract: Contemporary models of decision making assume individuals evaluate options solely in terms of their expected outcomes. However, recent research indicates that in institutional settings decision makers are also concerned with the procedural fairness of the decision process that generates those outcomes, such as whether decision makers were granted democratic inclusion in the decision-making process itself. We provide a much-needed quantification of the value decision makers place on inclusive decision procedures, showing (a) the pattern by which decision procedures alter individuals' perceptions of otherwise identical outcomes, spanning losses and gains of differing quantity and quality (e.g., failure/success) and (b) that decision makers' felt freedom and feeling of being treated respectfully mediate the effect of decision process on the perception of outcomes. We show that individuals prefer lesser economic returns in order to receive higher utility from the decision process, and we discuss the implications of this finding for decision science.

A Functional, Hormonal, and Computational Study of Sex Differences in Working Memory

Brandon Abbs
Harvard Medical School Brigham and Women's Hospital Connor's Center for Women's Health and Gender Biology MGH-MIT Athinoula Martinos Center for Biomedical Imaging

Jill Goldstein
Harvard Medical School Brigham and Women's Hospital Connor's Center for Women's Health and Gender Biology MGH-MIT Athinoula Martinos Center for Biomedical Imaging

Abstract: Studies show sex differences in working memory (WM) measured by using functional magnetic resonance imaging (fMRI). Effects have been associated with hormonal regulation of prefrontal (PFC) and parietal (PAR) cortices, regions implicated in WM function. Determining the pathophysiology of these sex differences has implications for understanding individual differences in WM. Using fMRI, we assessed WM using an N-back task and acquired hormonal status in 13 males and 13 females. Findings demonstrated sex differences in brain activity and connectivity between PAR and PFC, which were associated with female hormonal status. We suggest that hormones may regulate the 'gain' of neuronal activity in PFC and PAR, leading to less diffuse activation in women compared to men, the effect for which we propose a neural network model.

Now You See It, Now You Dont: Social Attention in a Magic Trick, Live and On Video

Robert Teszka
University of British Columbia

Evan Risko
University of British Columbia

Gustav Kuhn
Brunel

Alan Kingstone
University of British Columbia

Abstract: Research in social attention assumes that the way individuals attend to images of people accurately reflects how individuals attend to real live people. We examined whether this assumption is valid by studying how verbal cues affect where people look. In previous work using a live performance of a magic trick, verbal cues by the magician were shown to affect shifts in gaze. We extended this work by recording a similar trick in HD video and having observers watch the video while wearing an eye-tracker in order to investigate the effects of presentation medium (live vs. video) on social attention. While we discovered several key similarities across presentation mediums (e.g., verbal cues affected social attention), there were also some remarkable differences as well (e.g., the strength and frequency of the attention effect). The implication of these data for past and present theories of social attention, and future research approaches are discussed.

Declarative and procedural memory abilities as predictors of successful adult language learning

Katherine Brill
University of Illinois at Chicago

Mandy Faretta
University of Illinois at Chicago

Francis Wong
Northwestern University

Patrick Wong
Northwestern University

Kara Morgan-Short
University of Illinois at Chicago

Abstract: Evidence from two related but independent fields of research, second language acquisition (SLA) and cognitive neuroscience, suggests that some adult learners do reach high proficiency in an L2, as assessed by performance on language tasks. Not much is known about how or why certain adults attain high proficiency while others do not. Certain cognitive abilities, specifically procedural and declarative memory systems, may factor into language proficiency. This research aims to address this question by examining how individual differences interact with implicit language learning. Subjects completed a battery of cognitive tests including the Tower of London task, the Continuous Visual Monitoring Task and the Modern Language Aptitude Task, and learned an L2 under implicit training conditions. After practicing on comprehension and production tasks, subjects were given a grammaticality judgment task. A multiple regression was conducted to determine what cognitive abilities are unique predictors of L2 aptitude (as measured by the GJT).

A Bird's-Eye View of Numerical Discrimination in the Wild

Alexis Garland
Victoria University of Wellington

Jason Low
Victoria University of Wellington

K.C. Burns
Victoria University of Wellington

Abstract: Theory in numerical cognition has been in large part informed by evidence of numerical discrimination in humans and primates. Posited universal cognitive systems have, as a result, fundamentally been shaped by mammalian physiology and phylogeny, with findings largely supporting a ratio-based system. Explorations of large number discrimination with wild populations of any kind have, until now, been virtually unknown. Extant evidence on avian numerical capacity either focuses on large number discrimination in trained pigeons within a laboratory setting or object identification in terms of clutch size and brood parasitism in water fowl. Heretofore, no evidence has been uncovered to indicate the precise cognitive mechanisms that may be deployed in avian numerical choices of 'more' in a natural setting. Our study presents stark new evidence that redefines the capacities and limitations of number representation in the scatter-hoarding New Zealand robin (Petroica australis).

Distinguishing first-line defaults and second-line conceptualization in reasoning about humans, robots, and computers

Daniel Levin
Vanderbilt University

Megan Saylor
Vanderbilt University

Simon Lynn
Vanderbilt University

Abstract: We previously demonstrated that people distinguish between human and nonhuman intelligence by assuming that humans are more likely to engage in intentional goal-directed behaviors than computers or robots. In the present study, we tested whether participants who respond relatively quickly when making predictions about an entity are distinguish more or less between human and nonhuman agents. Participants responded to a series of five scenarios in which they chose between intentional and nonintentional actions for a human, a computer, and a robot. Those who chose quickly were more likely to distinguish human and nonhuman agents than participants who deliberated more over their responses. We suggest that the short-response time participants were employing a first-line default to distinguish between human intentionality and more mechanical nonhuman behavior, and that the slower, more deliberative participants engaged in deeper second-line reasoning that changed their predictions for the behavior of a human agent.

Cross-Modality Strategy Transfer: A behavioral study of strategic discrimination skill acquisition and transfer across auditory and visual modalities

Hao Bai
Mississippi State University

Paul Ladny
Mississippi State University

J. Gregory Trafton
Naval Research Laboratory

Randy J. Brou
Navy Personnel Research, Studies, and Technology

Stephanie M. Doane
Mississippi State University

Abstract: Discrimination is the ability to differentiate one object from another. Previous research suggests that with practice, individuals develop efficient discrimination strategies, and that strategies acquired in training transfer within a modality to novel stimuli (e.g., Haider & Frensch, 1999; Sohn, Doane, & Garrison, 2006). This study examines whether discrimination strategies transfer across presentation modalities. Subjects were trained to make difficult (similar) or easy (dissimilar) discriminations among visual or auditory objects during training. At transfer, subjects made difficult discriminations among stimuli presented in a different modality. Results suggest that the strategy acquired by subjects trained on difficult discriminations leads to superior performance at transfer compared to subjects trained on easy discriminations regardless of the modality of initial training and the modality switch at transfer. Cross?]modality transfer was observed, and this suggests the role of a central mechanism in the acquisition and transfer of strategic skills.

"That's what she said": The effect of emotional prosody on the interpretation of intent.

Jennifer M. Roche
The University of Memphis

Rick Dale
The University of Memphis

Abstract: Nygaard and Lunders (2002) have shown that emotional prosody impacts the resolution of lexical ambiguity during spoken language, yet sentential meaning is often ignored (Snedeker, 2008). The integration of linguistic and non-linguistic cues to speech is vital for successful interpretation of intent. Experiment 1 evaluated the perception of sentential contexts with emotional prosody (e.g., irritation, disgust, neutral, compassion, sarcasm and innuendo). Results suggested that listeners have the ability to differentially categorize the intent behind statements, based on emotional prosody. Experiment 2 evaluated the online process of categorizing sentences with emotional prosody, via systematic curvatures in the arm using the Wii remote. Results of Experiment 2 suggest that a perceiver's arm curvatures are reflective of the differentially interpreted categorized emotional information when the sentences remained stable. This supports the notion that emotional information plays a large role in the interpretation of sentential meaning, as it sharply influences the interpretation of intent.

Turn that frown upside down and to the left: Memory for faces is affected by their gravitational orientation

Nicolas Davidenko
Stanford University

Stephen Flusberg
Stanford University

Abstract: Recent research suggests the way our body is situated influences how we perceive our environment (Lopez et al., 2009), but it remains unclear whether our body's position influences how we process faces when retinal cues are kept constant (Troje, 2003; Lobmaier & Mast, 2007). In this study, participants completed an old/new face memory task in three different body positions (sitting upright, lying right, and lying left) and four different image orientations (upright, inverted, 90 degrees clockwise, and 90 degrees counterclockwise), allowing us to isolate the effects of retinal versus gravitational face orientation on recognition memory. We found a main effect of retinal face orientation, with higher d' to faces oriented upright versus inverted with respect to participants' retinas. Keeping retinal orientation constant, we also found an effect of gravitational orientation, with higher d' to faces orientated upright with respect to gravity, indicating a role of gravitational orientation in face processing.

Individual Differences in Successful Second Language Learning: The Roles of Working Memory and Intelligence

Brendan McCarthy
University of Illinois at Chicago

Mandy Faretta
University of Illinois at Chicago

Francis Wong
Northwestern University

Patrick Wong, Ph.D.
Northwestern University

Kara Morgan-Short, Ph.D.
University of Illinois at Chicago

Abstract: Individual Differences in Successful Second Language Learning: The Roles of Working Memory and Intelligence

B. McCarthy, M. Faretta, F. Wong, P. Wong, & K. Morgan-Short

Poster Abstract

Learning a second language (L2) in adulthood is notoriously difficult. Some learners, however, seem to learn with ease. In order to understand the characteristics of successful learners, research has explored the role of individual differences. The current study examines the role of two individual cognitive abilities: working memory and intelligence. Participants complete a battery of cognitive tasks, including a listening span test and the Kaufman Brief Intelligence Test. They subsequently learn an artificial language, which is meaningful, productive and consistent with natural languages, over the course of four training sessions. Language proficiency is assessed within subjects at the end of the first and final session, and a multiple regression analysis is conducted to probe the contribution of working memory and intelligence to successful L2 development. Implications for theories of L2 learning and for L2 learners are discussed.

Words (137)

Individual differences in anticipatory eye-movements: Vocabulary size is associated with speed of noun-verb integration

Arielle Borovsky
UCSD

Jeffrey Elman
UCSD

Abstract: Humans can integrate information from a rapidly changing speech stream with astonishing speed. In this study, we measure the impact of vocabulary knowledge on the incremental integration of speech using language-mediated anticipatory eye-movements. Following Kamide, Altmann & Haywood (2003), Experiment 2, we examined the degree to which an upcoming sentential Theme is anticipated by a combination of information from an Agent and Verb (eg. "The pirate hides the treasure" vs. "The dog hides the bones"). Replicating prior results, combinatory effects of the Agent and Verb yielded anticipatory looks to the Theme. When participant's performance was split by receptive vocabulary score, differences in anticipatory eye movements were apparent. The group with higher vocabulary scores was faster to integrate Agent and Verb information to correctly look at the upcoming Theme. Together our findings are suggest that prior language knowledge plays a pivotal role in even simple sentence processing tasks.

Reasoning through Mindful Actions: Effect of Instruction and Spatial Ability on Understanding Dynamic Systems

Margaret Chan
Teachers College, Columbia University

Abstract: The ability to understand systems is important for individuals engaged in various domains in science. Prior research has demonstrated direct-manipulation animation (DMA) effectively supports learners to understand dynamic systems. However, the role of learners' characteristics and scaffolding upon their performance remains unclear. The current study examines how instruction and spatial ability modulates learners' executive attention when they use DMAs to learn physical systems. To demonstrate comprehension and reasoning ability, participants were asked to explain and predict outcomes of "what-if" scenarios. Their eye-movements during their interaction with DMAs were recorded. Participants' eye-gaze patterns and learning performance revealed interactivity itself did not lead to understanding of these systems. Instead, the conjunction of attending to relevant information while actively manipulating the animation facilitated retention and reasoning. Spatial ability, however, was not significant in predicting performance. Overall, findings suggest learning with interactive animation involves cognizant interchange between bottom-up visuo-motor support and top-down cognitive control.

Are children irrational category learners? Evidence from a process model

Gavin Jenkins
University of Iowa

Jodi Smith
University of Iowa

John Spencer
University of Iowa

Larissa Samuelson
University of Iowa

Abstract: How do multiple labeling events influence children's understanding of objects that can be named at multiple levels of specificity ("Rover" or "dog")? To investigate, we replicated Xu & Tenenbaum (2007, Psychological Review, 114, 245-272), who found that children generalized more narrowly when three identical toys (e.g., plush Dalmatians) were labeled with a novel word compared to one toy labeled three times. Xu & Tenenbaum suggested the extra two referents provide statistical evidence that rationally supports a narrow hypothesis. In our "extra labeling" condition, however, children generalized broadly when each object was labeled ten times instead of once. This violates the predictions of a purely rational account and suggests situational, lower-level processes are critical to novel word generalization. A Dynamic Neural Field model is used to examine these processes, and further shows how process-oriented models can solve the problem of overlapping extensions.

Brain Response Over Time to Structured and Unstructured Musical Sequences

Kat Agres
Psychology Department, Cornell University

Hia Datta
Sackler Institute for Developmental Psychobiology, Weill Cornell Medical College

Jason Zevin
Sackler Institute for Developmental Psychobiology, Weill Cornell Medical College

Abstract: Research in speech and music shows that listeners model their auditory environment to form expectations about future input. Here we asked how the ability to predict upcoming musical tones based on both general implicit knowledge of musical structure and familiarity with a particular tune influence early portions of the auditory evoked response (AER). Using electroencephalography, we examined how predictability influenced brain responses to repeated tunes. The musical stimuli were simple, monophonic Irish folk tunes (Normal) and Random sequences in which the notes of each tune were presented in randomized order. Because the Random stimuli lacked musical structure, we hypothesized that listeners would be unsuccessful in learning or creating predictive models for these sequences. Differences were observed in early AER between the Normal and Random sequences. The results suggest that listeners successfully learn, remember and model the Normal sequences, but are unable to do this for Random sequences.

Seeing the world through a visual language: Visual world paradigm in British Sign Language

Robin L. Thompson
University College London

David P. Vinson
University College London

Neil Fox
University College London

Gabriella Vigliocco
University College London

Abstract: We used a visual world paradigm with British Sign Language (BSL) to test the methodology with a visually perceived language, and gain insight into the time course of BSL processing at the lexical level. Subjects were tracked while viewing four object pictures and a BSL video. One picture was (semantically or phonologically) related to the target sign, with target pictures present on some trials and absent on others.

Like previous spoken language studies, sign perceivers looked significantly more often towards related distracters than unrelated after the onset of the target sign, regardless of target picture presence. However, results indicate important differences in the time course of looks compared to speech, with gaze to pictures infrequent (due to attention to video) and earlier, often before actual sign onset. This last occurs because, unlike speech, the preparation phase (in which sign handshape and location are attained) is visible, allowing earlier cohort activation.

Insight into dynamics of speech perception in English and Japanese native speakers using a mouse-tracking paradigm

Hia Datta
Sackler Institute for Developmental Psychobiology, WCMC & The Graduate Center, CUNY

Ran Liu
Department of Psychology, Carnegie Mellon University & Sackler Institute for Developmental Psychobiology, WCMC

Jason Zevin
Sackler Institute for Developmental Psychobiology, WCMC

Abstract: In American English, the qualitative vowel contrasts /I/-/E/ and /E/-/ae/ are distinguished primarily by spectral and duration cues, respectively. Japanese uses duration cues for quantitative, but not qualitative contrasts. We measured arm movements with a mouse-tracking task while native English and Japanese speakers living in the United States distinguished spectral and duration contrasts in the word pairs "pin/pen" and "pen/pan." While both groups identified the words correctly, when distinguishing pan from pen, the English but not Japanese speakers' mouse-trajectories lean towards 'pen' earlier in the trial, when the information accumulated to that point is more consistent with the shorter of the two words. This suggests that while English speakers act rapidly on incoming acoustic information as it unfolds, Japanese speakers wait for more information before responding. These results may reflect Japanese speakers' expertise with duration cues from extensive experience with native quantity contrasts.

Concrete Models as Aids to Representational Translation of Molecular Diagrams

Andrew Stull
University of California, Santa Barbara

Hegarty Mary
University of California, Santa Barbara

Stieff Mike
University of Maryland, College Park

Dixon Bonnie
University of Maryland, College Park

Abstract: Chemists use many different types of diagrams to represent molecules and must develop skills to accurately translate between such diagrams. Translating between such diagrams can potentially involve the intermediate step of forming an internal 3-d representation of the molecule, so we hypothesized that performance would be enhanced when concrete models were used. Thirty students were provided with models as they translated one molecular diagram into a second and their spontaneous use of the models was recorded. Students' model use was coded for behaviors, such as moving, holding, reconfiguring, pointing to, or gesturing about the model. Results showed a great diversity in whether and how students used the models. Although performance on the representational translation task was generally poor, using the models was positively correlated with performance. We will also report the results of a follow-up study that compares student performance with and without models.

I spy with your eye: On the perception of others' gaze

Nicola Anderson
University of British Columbia

Craig Anderson
University of British Columbia

Evan Risko
University of British Columbia

Alan Kingstone
University of British Columbia

Abstract: Eye tracking can often be prohibitively expensive, proprietary and incredibly complex. We examined the accuracy of eye tracking using only a web camera. Participants were shown webcam recordings of a persons' eyes moving 1, 2, or 3 degrees of visual angle in one of 8 directions (North, NE, E, SE, etc) or no eye movement occurred at all. Observers judged whether an eye movement was made, and if so, its direction. Detection and direction judgments were significantly above chance across all three distances, with larger eye movements resulting in better performance. These data indicate that a webcam plus human observer system can be used to study human eye movements in a simple non-invasive manner. They also support theories of human social attention and evolution predicated on the assumption that humans have developed an especially fine ability to detect the eye movements of others and determine where those eyes are looking.

Beyond binary: One small step across the artificial-naturalistic divide in understanding human category learning

Kimery Levering
State University of New York: Binghamton

Kenneth Kurtz
State University of New York: Binghamton

Abstract: The foundational six types problem (Shepard, Hovland, & Jenkins, 1961) offers a limited view of human category learning due to minimal ecological validity. The SHJ types actually do address the explanatory constructs historically proposed to explain natural categories (rules, exemplars, family resemblance), but do so in an impoverished manner due to binary-valued dimensions. We studied the SHJ types using stimuli with four values on each dimension. The SHJ types become more like natural concepts with more robust intension (internal structure) and extension (category size). This allows for richer evaluation of the representations and processes underlying learner performance. Results depart from the traditional ease of learning order (I>II>III,IV,V>VI) since Type IV (family resemblance structure) shows higher accuracy than Type II early in learning. Transfer to novel items and typicality ratings reveal graded structure even in rule-described categories and show distinct signatures of the categorization basis used by each learner.

Dimension Word Knowledge and Flexible Attention Shifting

Rima Hanania
Indiana University

Thea Ionescu
Babe-Bolyai University,Cluj Napoca, Romania

Linda B. Smith
Indiana University

Abstract: One of the developing skills in executive functions is the flexible control of selective attention. Young children improve in this ability during the preschool years as measured by such tasks as the Dimension Change Card Sort (DCCS) which asks children to sort pictures, first by one dimension (e.g. color), and then by another (e.g. shape). Three-year-olds sort correctly by the first dimension, but do not sort correctly when the rule changes (i.e., they perseverate). By five, children switch flexibly between dimensions. This study asks about the relationship between perseveration in the DCCS and dimensional word knowledge which also improves during preschool years. Thirty children participated in a study which tests knowledge of feature terms (e.g., red, blue) and dimension terms (e.g. color, shape). Results indicate no difference in knowledge of feature terms for the two groups, but children who switch successfully in the DCCS were significantly better at dimension terms.

Lay Theories and Linguistic Framing in Teaching Children about Nutrition

Sarah Gripshover
Stanford University

Ellen Markman
Stanford University

Abstract: Research has shown that health interventions that build wisely upon existing conceptual knowledge can be effective in producing conceptual and behavioral change. Pre-school aged children have been found to understand the relationship between food and the body principally in terms of their lay mechanical theories. A conceptual intervention is designed to build upon this emerging understanding in order to teach nutritional balance and variety. Special attention is paid to the role of linguistic framing in building coherently upon children's existing mechanical knowledge. In particular, framing inert, inactive food as a causal agent in sentences such as milk gives you strong bones may be especially opaque given children's mechanical understanding of nutrition, compared with framing the body as the causal agent, e.g., your bones use milk to grow strong. Results indicate that children given a body-agentively—but not food-agentively—framed intervention achieved greater understanding of nutritional balance and variety.

The hindsight bias with dynamic stimuli and the propensity effect with static stimuli

Dayna Gomes
California State University, Los Angeles

Dustin Calvillo
California State University San Marcos

Abstract: A reversal of the hindsight bias, termed the propensity effect, has been found with dynamic stimuli when likelihood estimates of an event are lower among people with outcome knowledge than among those without outcome knowledge (Roese, Fessel, Summerville, Kruger, & Dilich, 2006). One hundred sixty-two participants were shown a vehicular accident depicted either by diagrams or by an animation. Some participants saw information leading up to the accident and were asked to predict the likelihood of an accident occurring. Others saw the accident and were instructed to disregard this outcome knowledge before providing a likelihood estimate. Results contradicted previous findings; the propensity effect occurred with diagrams and the hindsight bias occurred with the animation. Evidence for the propensity effect appears to depend on how stimuli are constructed and the point at which participants are asked to disregard outcome information. The present findings indicate that different presentation modes influence decision making.

Category Learning in Second Life: Effects of Learning Context on Mechanisms of Categorization

Joshua Sturm
Vassar College

Peter Nachbaur
Vassar College

Alex Goldberg
Vassar College

Julianne Herts
Vassar College

Jan Andrews
Vassar College

Ken Livingston
Vassar College

Abstract: A major challenge in category learning research is designing novel stimuli that are functionally meaningful, then presenting them in an environment that is sufficiently controlled while remaining true to the dynamic nature of the real world. The interactive online environment Second Life represents an interesting methodological setting for such work, offering sophisticated stimulus design software, a powerful scripting language, and a fully editable physics engine. Two well-documented phenomena of category learning are "compression" (where objects classified together appear more similar) and "expansion" (where objects classified differently appear more different). Our studies using Second Life have thus far demonstrated its utility as a research tool by 1) successfully creating a compression effect "in world," 2) showing that this effect did not require verbal labels, and 3) revealing that a more interactive version of the task with much more complex and naturalistic stimuli produced learning without compression or expansion effects.

Deciding Whether or Not to Guess the Answer Predicts Subsequent Learning

Sean Kang
UCSD

Michael Mozer
University of Colorado

Harold Pashler
UCSD

Abstract: In a study on the effects of incorrect guessing on subsequently learning from feedback, we found confidence for wrong responses on an initial test was positively associated with correct final recall (higher confidence errors corrected better; Butterfield & Metcalfe, 2001). One explanation for this hypercorrection effect is that subjects are surprised by their error and thus pay more attention to the feedback (Fazio & Marsh, 2009). Inconsistent with this surprise hypothesis, however, was our finding that the decision to volunteer a low-confidence guess, even when the response was wrong, was associated with better subsequent learning of the correct answer than when a response/guess was withheld. We propose that the willingness to venture a guess, even when confidence is low, may reflect a higher state of learning, relative to choosing to omit a response. We present additional behavioral data and an error-correction neural network model in support of our alternative hypothesis.

Am I a Robot? How Verb Agency and Agent Description Influence Perspective-Taking in Visual Scenes

Michelle D. Greenwood
University of California, Merced

Teenie Matlock
University of California, Merced

Michael J. Spivey
University of California, Merced

Justin L. Matthews
University of California, Merced

Abstract: People often take an egocentric perspective when describing space. However, they occasionally take an alternative perspective. When and why? In a series of experiments that followed work on perspective, we explored this question. In one experiment, participants were given photographs of two objects on a table. Objectively, the scene could be described from either the perspective of the person viewing the picture or from the opposite perspective (i.e., facing the viewer). To test which viewpoint would be elicited, we asked participants to describe where an object was relative to another. In one experiment, a toy humanoid robot (facing the participant) was included in the scene to determine whether people would take its vantage point when referring to object locations, and how this inclination might vary according to changes in linguistic context. Results indicate that people can spontaneously take the perspective of an agent-like toy when describing object locations.

The role of conventional number knowledge in young children's nonverbal number matching: Is "two" special?

Mee-Kyoung Kwon
the University of Chicago

Yoonkyung Jeong
Catholic Universtiy

Susan Levine
the University of Chicago

Abstract: Two studies examined the role of conventional number knowledge on young children's nonverbal number matching (NVM). We hypothesized that acquiring two number words ("one" and "two") would facilitate children's performance on the NVM task by highlighting numerical comparisons rather than comparisons based on other variables. To test this hypothesis, two- and three-year-olds were given a NVM task that was either controlled for line-length/density (Experiment 1) or total surface area (Experiment 2). Conventional number knowledge was assessed using two tasks: Give-A-Number and How-Many. Results from Experiments 1 and 2 showed the advantage of knowing at least two number words: "Two-knowers" and above performed significantly better on NVM than one-knowers or non-counters. Performance of one-knowers and non-counters did not significantly differ, suggesting that children's performance on NVM is not significantly impacted by learning only one number word.

A sequence analysis of actions in complex system comprehension

Patrick Jeuniaux
Universite Laval

Sebastien Tremblay
Universite Laval

Jean-François Gagnon
Universite Laval

Daniel Lafond
DRDC-Valcartier

François Bernier
DRDC-Valcartier

Abstract: Complex systems have a broad network of relations for which human comprehension is severely limited and analysts often rely on the support of technological systems. In this study we investigated whether comprehension can be augmented by IMAGE – a set of interactive visualization, data exploration and knowledge representation tools – and explore behavioural signatures associated with the optimal use of IMAGE. The comprehension and use of IMAGE of 24 participants were examined in the context of a scenario involving military convoys evolving their strategy according to the reactions of a hostile and dynamic environment. Comprehension was measured by a score normalized in function of a randomly generated exploration of the system. A sequence analysis was performed to extract the pattern of IMAGE-user interaction. Our results reveal a great diversity across participants and that transitional probability of key IMAGE events is not related to augmented comprehension in a simple structured way.

A valid separation of location memory based on allocentric and egocentric reference frames

Jonna Nilsson
Northumbria University & Newcastle University

Kenny Coventry
Northumbria University

Nicol Ferrier
Newcastle University

Abstract: A valid separation of location memory based on allocentric and egocentric reference frames
Jonna Nilsson, Kenny Coventry, Nicol Ferrier

The existence of two separate spatial systems, one based on an egocentric viewpoint-dependent reference frame and one based on an allocentric viewpoint-independent reference frame, is now well accepted both at a conceptual and a neurological level (O'Keefe & Nadel, 1978; Lavenex & Lavenex, 2009; Zaehle et al, 2007). However, methodologies intended to separate and compare location memory based on distinct reference frames in humans vary widely and are often confounded. To allow for a more reliable separation of the egocentric and allocentric reference frames, a new location memory task was developed that eliminated these confounds. The results of a series of studies based on this task are reported and discussed. The results highlight the importance of controlling for the extraneous variables present in previous studies. It is evident that the investigation of location memory has a lot to gain from the valid separation of the allocentric and egocentric spatial systems.

The relationship between similarities computed by LSA and several types of association

Keisuke Inohara
Kyoto University

Takashi Kusumi
Kyoto University

Abstract: Latent Semantic Analysis (LSA) is a computational theory of meaning. Meanings of words are extracted from a large corpus of texts by a statistical method and represented as vectors in a semantic space. It is known that similarities of word-pairs computed by LSA can explain various language processing of human(e.g., Landauer and Dumais, 1997). To know features of LSA, we created four semantic spaces from Japanese corpses: news paper, novels, books (except for novels), and both of novels and books. The similarities of word pairs were compared with scores in different types of association tasks. Participants were asked to associate several concepts(e.g., subordinate categories, synonyms, or action concepts) with stimulus words. As a result, a correlation coefficient between the similarities and association scores of the action concept task is the highest. This finding is consistent with Hare et al.(2009), LSA and similar models reflect people's knowledge about daily events.

Reanalysis of Linda Problem

SangSuk Yoon
Pusan National University

MinGyung Choi
Pusan National University

HyunJung Shin
Pusan National University

Abstract: There has been many researches about conjuction fallacy, which people tend to evaluate conjunctive statement more probable than component statements. Conjunction Fallacy shows that people obviously violate probability rules, which is conjuctive state cannot be higher than component statement. However, we in our experiment, we tried to explain this fallacy by using conditional probability. If people used bayesian updating rule, then the fallacy cannot be the evidence of people's irrationality. In our result, people showed some pattern that they had used bayesian updating rules, but the result was not statistically significant.

A computational model for the acquisition of referring subjects in discourse

Jacolien van Rij
University of Groningen

Hedderik van Rijn
University of Groningen

Petra Hendriks
University of Groningen

Abstract: In this study, we investigate how children acquire adult-like performance on their use of referring subjects, by modeling experimental data within the cognitive architecture ACT-R. When choosing which type of referring expression to use, speakers have to take into account the linguistic discourse context as well as the way listeners will interpret the referring expression in that context. Children, however, tend to produce unrecoverable pronouns in particular contexts where adults would use a full noun phrase. This suggests that children are not yet able to take into account the listener's perspective. Based on simulations of our computational model, we argue that the adult use of referring subjects is crucially dependent on sufficient working memory capacity and speed of linguistic processing.

Word-Form Typicality and Its Influence on Grammatical Category Assignment

Thomas Farmer
Northumbria University

Padraic Monaghan
Lancaster University

Jennifer Misyak
Cornell University

Morten Christiansen
Cornell University

Abstract: Farmer et al. (2006) found that how typical a word's phonology is of other words in its lexical category influences the reading times of nouns and verbs in predictive contexts. When a preceding context generated a strong expectation for a noun, target noun-like nouns were read faster than verb-like nouns, along with a similar effect for verbs. Further, Dikker et al. (2010) found that the magnitude of the M100 response (sensitive to category-expectation violations) was modulated by phonological typicality: when a mismatch existed between whether context predicted a noun and the noun's degree of typicality, a heightened M100 response occurred in visual cortex. Here, using lexical decision and a self-paced reading, we examine whether these word-form effects occur in all cases where a noun or verb is possible, or whether they are more robust when expectations are strong, and thus visual word-form information may be enough to facilitate categorization.

Did you say "gross snails" or "gross nails?" The problem of segmenting co-occurrences of the same segment

Dahee Kim
OSU, Linguistics

Colin Widmer
OSU, Psychology

Christine Szostak
OSU, Psychology

Mark Pitt
OSU, Psychology

Abstract: To comprehend spoken language, listeners have to segment words from a continuous stream of speech. This task may be difficult when [s] repeats at a word boundary (e.g., gas station) because the two s's could blend together and form one long s-sound. Do talkers produce cues that signal the word boundary? If not, how do listeners segment the words correctly? We addressed these questions by having talkers produce two-word sequences that could be interpreted in three ways, depending on how the middle s-sound was heard (e.g. grow snails, gross snails and gross nails). Acoustic analyses of their productions examined whether there are cues indicating the presence of a word boundary. Listening experiments using the talkers' productions as stimuli were carried out to determine the extent to which signal-based and knowledge-based cues are used to resolve ambiguities. Implications of the results for models of spoken word recognition will be presented.

Expectations of common ground with a computer dialog agent

Donna Byron
Northeastern University

Joy Hanna
Oberlin College

William Hartmann
Ohio State University

Abstract: The time required to comprehend referring expressions is influenced by many contextual factors that establish expectations, including attributes of the referent and competitors, the discourse history, and the common ground between dialog partners. When human partners violate expectations by, for example, breaking a conceptual pact to maintain a referential perspective across mentions, listeners incur processing costs. But listeners don't have difficulty when other partners refer differently. We use this response pattern to examine whether subjects process discourse from computer agents the same way as from human partners, and what expectations they have of multiple agents. As computer dialog systems mature, we expect individual computers to be running multiple dialog agents playing diverse roles: information butlers, virtual salesmen, health coaches, etc. Knowing more about the expectations humans have of these dialog partners is a fundamental first step toward designing algorithms that can generate referring expressions that are easy to comprehend.

Interaction of bottom-up and top-down attentional influences on the processing of contingency information

Kelly Goedert
Seton Hall University

Brianna Eiter
Hofstra University

Abstract: How do individuals determine what information to attend to when making causal inferences on the basis of contingency information? Although much research has focused on the role of top-down attentional influences on this process (e.g., effects of prior beliefs and motivation), little work has addressed the role of bottom-up attentional influences, which may be driven by low-level perceptual and motor processes. In prior work, the current authors demonstrated a distinct bottom-up rightward bias in overt attention during contingency acquisition that was associated with subsequent causal judgments (Goedert & Eiter, 2008). Here we recorded eye movements of participants while they acquired information about two candidate causes whose spatial locations varied over the course of learning. We found that the bottom-up rightward bias in gaze direction persisted in spite of the varied spatial locations of the causes. Additionally, the bias interacted with top-down, knowledge-based contingency acquisition processes to influence participants' gaze patterns.

Influence on memory of the temporal schedule of repetitions over multiple days and its modulation by the retention interval

Emilie GERBIER
Université Lyon 2

Olivier KOENIG
Université Lyon 2

Abstract: We studied the influence on memory of three temporal schedules of repetitions of vocabulary pairs. Pairs were presented on Day 1, 7, and 13 in a Uniform schedule; on Day 1, 2, and 13 in an Expanding schedule; and on Day 1, 12, and 13 in a Contracting schedule, with schedule as a within-subject factor. Retention was tested with a cued-recall task performed on Day 15, 19 (Experiment 1), or 26 (Experiment 2). Cued recall did not differ as a function of the schedule on Day 15, whereas the Expanding schedule led to the best performance on Day 19 and 26. We interpreted this new finding of a modulation of the effect of schedule as a function of the retention interval within the frame of a model based on the study-phase retrieval theory that accounts for different forgetting curves for different schedules.

Matching Exact Posterior Probabilities in the Multinomial Interactive Activation Model

Pranav Khaitan
Stanford University

James L. McClelland
Stanford University

Abstract: Interactive activation models of context effects in perception ((McClelland & Rumelhart, 1981; McClelland & Elman, 1986) have been criticized for failing to combine stimulus and context information in a Bayes-optimal way, leading to a rejection of interactive approaches (Norris & McQueen, 2008). We show that interactive activation can compute correct Bayesian posterior probabilities. We present a variant of the interactive activation model that produces outputs exactly corresponding to the correct posterior probabilities of letters given specified letter feature and context information. In the new variant of the model, inhibition between units within pools is replaced by selection of a single unit to be active, using the softmax function to assign probabilities to candidate alternatives. The model is fully interactive, yet the probability of a letter unit being activated is both provably and demonstrably equal to the posterior probability given the presented feature and context information.

Threat and anxiety interactively impair task switching ability

Wolfgang Rauch
Goethe University Frankfurt IDeA - Center for Individual Development and Adaptive Education

Marie Lauer-Schmaltz
Goethe University Frankfurt

Abstract: Anxiety is associated with an attentional bias toward threat, yet the underlying mechanisms of this bias remain to be explored. Based on the assumption that anxiety selectively increases the strength of stimulus-driven attention to threat, we hypothesize that threatening stimuli and anxiety interactively impair the ability to re-allocate attention based on internal goals.

We tested this assumption in a task-switching experiment with N=29 participants. Compared to task repetition, task switching requires goal-directed attention in order to reconfigure the task set. Both emotionally neutral and threatening stimuli were presented, and dispositional anxiety was measured using the STAI questionnaire. Results showed an interaction of anxiety and emotional quality of pre-switch stimuli on switch cost, independent of the emotional quality of the currently presented stimulus: switch cost was larger for more anxious participants, but only when the task switch was preceded by a threatening stimulus.

Grammars for Funk Drumming: Symbolic and Motor-Spatial Aspects

Richard Ashley
Northwestern University

Abstract: Skilled musical performance requires a detailed knowledge of the grammatical patterns of a musical style, but also spatial and motoric skills in coordinating the abilities of the human body with the affordances of musical instruments. This study investigates these aspects of musical performance in the domain of popular music drumming, seeking to understand how drum patterns are constructed and produced by skilled players. It considers drum patterns from two perspectives: abstract grammaticality as revealed in frequency of sequences of drum sounds (via a probabilistic grammar), and also embodied and spatial cognition, as revealed in how patterns are produced by the player by motion sequences across the drumset. To investigate these issues, a combination of corpus analysis and experimental methods have been employed. Results to date indicate that production of drum patterns is based partly on articulatory constraints (ease of production), but also on a cognitive or acoustic attribute of contrastiveness.

About the Validity of Computer Models in Cognitive Science

Ricardo Sanz
Universidad Politecnica de Madrid

Carlos Hernández
Universidad Politecnica de Madrid

Jaime Gómez
Universidad Politecnica de Madrid

Guadalupe Sánchez
Universidad Politecnica de Madrid

Adolfo Hernando
Universidad Politecnica de Madrid

Abstract: Many cognitive models are evaluated by implementing an artificial system –a program, a robot– that performs the concrete task that the model is addressing. Success in task performance by the program is considered a proof of the adequacy of the cognitive model proposed. However, this is not a valid inference in general. The reason is that the transformation of the model from a textual or graphical form into a computer implementation is not transparent. This implies that phenomena and properties observed in the program cannot be predicated of the model. The reason for the lack of transparency is that models are not expressed using a rigorous language and that transformations into implementations aren't rigorous either. Hacks are introduced during the construction of the program –to make it work– and they are not taken back into the model, hence invalidating the model-implementation relation.

The roles of working memory capacity and spatial ability in first-time solution of the Tower of Hanoi

Patrick Cushen
University of Illinois at Chicago

Jennifer Wiley
University of Illinois at Chicago

Abstract: Decades of research have highlighted the important role of working memory capacity (WMC) in higher cognition and problem solving. Strangely, the relationship between WMC and the Tower of Hanoi task, a classic problem in Cognitive Psychology, has yet to be firmly established. Many studies have failed to find the suspected relationship and those that have identified a relationship have almost universally used spatial-modality measures of WMC. These results fail to differentiate between whether it was an individual's WMC or spatial ability that predicted performance. As such, the goal of the current research was to investigate the complex relationship between WMC, spatial ability, and the Tower of Hanoi. Results suggest different roles for WMC and spatial ability in the first-time solution of the Tower of Hanoi.

Causal Learning in Joint Activity: Comparing Collaborative, Active, and Passive Contexts

Andrew G. Young
University of Wisconsin - Madison

Martha W. Alibali
University of Wisconsin - Madison

Charles W. Kalish
University of Wisconsin - Madison

Abstract: Children's causal learning from intentional actions (i.e., interventions) can be dramatically affected by the psychological and social circumstances in which they are produced. A context fundamental to children's learning across many domains is joint activity; however little is known about children's causal learning with others. Collaborative settings with shared goals and action plans might facilitate children's learning from their own and a partner's coordinated actions (Sommerville & Hammond, 2007). Alternatively, children's own interventions may be more informative than those produced by a partner (Kushnir, Wellman, & Gelman, 2009). To address these issues, young children learned about simple causal systems via interventions performed: 1) by themselves, 2) by an experimenter, or 3) jointly with an experimenter. Children's subsequent causal knowledge, source-memory, and free play are compared across these collaborative contexts. Findings may provide guidance about structuring learning environments.

What makes for inspirational examples in design? The effects of example modality, distance, and familiarity.

Joel Chan
University of Pittsburgh

Katherine Fu
Carnegie-Mellon University

Christian Schunn
University of Pittsburgh

Kristin Wood
University of Texas-Austin

Jonathan Cagan
Carnegie-Mellon University

Kenneth Kotovsky
Carnegie-Mellon University

Abstract: An important question in the cognitive science of design concerns the influence of environmental input on ideation processes. Prior work has demonstrated that analogizing over examples in the environment is a double-edged sword: examples can help designers come up with innovative designs, but variations in key properties can result in negative design outcomes (e.g., fixation). We investigated the influence of variations in presentation modality, analogical distance, and familiarity of provided design examples on ideation processes. Engineering students generated solution concepts for an engineering design problem with or without provided design examples (analogy groups vs. control group). Examples in the analogy groups were fully crossed by modality (pictures vs. text), distance (near vs. far), and familiarity (familiar vs. unfamiliar). Results indicate that designers' familiarity with examples influence whether they suppress or promote innovation, regardless of modality or analogical distance.

False Recognition in the DRM-Paradigm reflects False Encoding

Tamella M. Pettitt
Birkbeck College, University of London

Eddy J. Davelaar
Birkbeck College, University of London

Abstract: A thorny question is whether the high rate of false memories in the Deese/Roediger-McDermott (DRM) paradigm is due to encoding or retrieval processes. Previous work suggests a locus at encoding using the free recall task. Here we use a recognition task in which false memories have typically been associated with intrusions due to high levels of featural overlap between the test probe and the stored memories. We used the REM model to assess the expectations from a retrieval account and from an encoding account of false memories. These simulations show that (probabilistic) false encoding leads to exceptionally large standard deviations in memory strength that are shown in ROC curves. We tested and verified these predictions in an experiment using a divided attention paradigm. These findings suggest that high false alarm rate in a DRM-paradigm might not be as useful as a proxy for everyday false memories as previously supposed.

Monday is before Tuesday in speech, but left of Tuesday in gesture.

Daniel Casasanto
Max Planck Institute for Psycholinguistics, Donders Institute for Brain, Cognition and Behaviour

Kyle Jasmin
Max Planck Institute for Psycholinguistics

Abstract: Do English speakers gesture about time the same way they talk about it? In spoken English, time appears to flow along the sagittal (front/back) axis: we look forward to the future and back on the past. Yet, when English speakers produce spontaneous gestures they often use the lateral axis, gesturing leftward for earlier times and rightward for later times, consistent with the flow of time on calendars and graphs.

Here we show that speakers spatialize time on the lateral axis overwhelmingly more often than on the sagittal axis in spontaneous co-speech gestures. This is true despite the prevalence of spoken front/back metaphors and complete absence of left/right metaphors for time in spoken language. Interestingly, front-back gestures, though rare, were more common during deictic language, consistent with predictions based on signed languages. We propose possible pragmatic, kinematic, and mnemonic motivations for this dissociation between spatio-temporal metaphors in speech and gesture.

Enhanced visuo-spatial learning and memory effects in time-space synesthesia

Ursina Teuscher
University of California, San Diego

David Brang
University of California, San Diego

Vilayanur S. Ramachandran
University of California, San Diego

Seana Coulson
University of California, San Diego

Abstract: Time-space synesthetes report that they consistently experience time events, such as the months of the year, as having a specific spatial layout. Two studies compared 11 synesthetes' and 41 non-synesthetic controls' ability to memorize novel spatial calendars. Both studies revealed better memory performance (quicker reaction times and higher consistency) in synesthetes than controls, even if the memorized calendar was specifically designed to conflict with synesthetes' own involuntarily experienced calendar. Furthermore, an additional group of controls' performance was better for counterclockwise than clockwise calendars, perhaps due to less interference with conventional mappings. These findings suggest that time-space synesthetes' enhanced visuo-spatial memory abilities may underlie the emergence of time-space synesthesia, as people with a greater capacity to learn mappings between spatial forms and temporal sequences might be more likely to think of months of the year in terms of idiosyncratic shapes, while others might be more reliant on culturally established mapping schemes.

Analyzing Discourse Functions in Student Research Reports to Assess Gains Due To Research Experiences

Roman Taraban
Texas Tech University

Brianna Bennett
Texas Tech University

Xiaofang Zeng
Texas Tech University

Abstract: Two case studies are presented that concern the assessment of scientific discourse in undergraduates' research papers. An assessment methodology was developed capable of tracking and evaluating the level and kinds of changes that result from students' participation in laboratory experiences. An upper-level performance limit was established by analyzing journal articles written by the students' faculty mentors. Students were compared to mentors in terms of the frequency of use of higher-level (e.g., stating a hypothesis) and lower-level discourse functions (e.g., stating background information), as well as with respect to the syntactic complexity of their respective sentence constructions. In self-reports of research knowledge and skills, students express gains that are not evident in their papers, suggesting that the written form poses specific challenges. We consider prospects for automating the assessment of students' research papers by using electronic means to assist in the identification and enumeration of the types and frequencies of discourse functions in these papers.

Embodying attentional states: The role of posture in task performance

Joseph Chisholm
University of British Columbia

Evan Risko
University of British Columbia

Alan Kingstone
University of British Columbia

Abstract: When participating in either an engaging or unengaging task, individuals appear to adopt consistent postures that may reflect a focused or unfocused attentional state. Posture is known to communicate certain affective states; however, it is of interest whether posture may elicit an attentional state that can influence performance on a task. To address this question, participants were first instructed to sit focused or unfocused while conducting a word recall and visual search task. Results showed a benefit of being in a focused posture in the visual search task as well as consistent postures adopted within groups. However, when participants were given instructions on how to sit, without explicit mention of focused or unfocused postures, no performance differences were observed. These results suggest that posture alone may not be enough to elicit particular attentional states, at least not in the tasks we used. Implications and future directions will be discussed.

Similarity avoidance in processing consonants: apparent exceptions from Polynesian languages

John Alderete
Simon Fraser University

Abstract: In languages as diverse as English, Arabic, and Russian, the cooccurrence frequency of two consonants can be predicted from a gradient measure of the similarity of the phonological make-up of the two consonants. An argument for the role of phonological similarity in these languages is that it tends to be the case that sound classes with more members have weaker cooccurrence restrictions than classes with fewer members. More members require more contrasts, which reduces similarity. In Arabic, there are weaker restrictions on coronals (n=11) than labials (n=4) because coronals exhibit more contrasts. However, there are a number of well-documented Polynesian languages (Maori, Hawaiian, Tongan) were the reverse is true: there are more labial consonants than coronals, but there are stronger cooccurrence restrictions on labials than coronals. This project investigates the lexical statistics of these languages and proposes a revised characterization of similarity avoidance that has greater cross-linguistic coverage.

Cross-Situational Word Learning in Bilinguals

Viridiana Benitez
Indiana University

Linda B. Smith
Indiana University

Abstract: Recent research indicates that fluency in more than one language has consequences for fundamental aspects of the cognitive system beyond that of speaking two languages. Specifically, a now growing body of research shows speakers of two languages can allocate their attention more efficiently than speakers of one language in nonlinguistic tasks (e.g., Bialystok, Klein, Craik, & Viswanathan, 2004). In addition, bilingualism has also been shown to promote abilities in the linguistic domain, such as the ability to learn novel words better than monolinguals (Kaushanskaya and Marian, 2009). In the present study, we investigated how the efficiency of attentional allocation and the ability to learn novel words may interact and affect performance in Yu and Smith's (2007) cross-situational word learning paradigm. Monolingual and bilingual adults' learning of 18 novel words using this paradigm were examined, and results help to better understand the effects of bilingualism on cognition.

Metacognition and Writing: How an Academically Gifted Adolescent Organizes and Controls the Writing Process

Delayne Connor
Bridgewater State College

Abstract: This single subject study concerns a 14 year-old academically gifted student's use of metacognition when producing written discourse in response to a writing prompt. Ethnographic procedures were used to collect and analyze data. The participant engaged in a think-aloud procedure as he composed an expository paper, and his transcribed verbalizations were then analyzed for metacognitive strategy use. The strategies were organized into a taxonomy by means of Spradley's domain and taxonomic analysis.

Results indicated that the participant advanced the writing process by employing five major types or domains of strategies. They are: (a) planning discourse/thinking, (b) evaluating discourse/thinking, (c) recognizing difficulty with discourse/thinking, (d) responding to difficulty with discourse/thinking, and (e) repairing discourse/thinking. The participant employed over 80 individual strategies within these five domains as he wrote a well organized and cohesive composition at one of Britton et al.'s higher abstractive levels of discourse.

Immediate Introduction to Multiple Procedures Supports Procedural Flexibility in Equation Solving

Kelley Durkin
Vanderbilt University

Bethany Rittle-Johnson
Vanderbilt University

Jon Star
Harvard University

Abstract: Knowing multiple procedures and using them adaptively is important for problem-solving. We examined how different methods for developing procedural flexibility affected novices learning equation solving. Students (N = 198) were assigned to one of three conditions that differed in whether multiple procedures were introduced immediately or after practice with one procedure (no delay vs. delay) and in whether comparison of examples was supported. Students in the no delay condition had greater procedural flexibility and accuracy than students in either delay condition, regardless of whether they compared examples. Differences in students' explanations during the intervention suggest reasons for the benefits of immediate introduction to multiple procedures. Students in the no delay condition more frequently compared and evaluated efficiency of procedures than students in delay conditions. They also more frequently used efficient procedures during the intervention. Immediate introduction to multiple procedures supports attention to and adaption of efficient procedures, which benefits flexibility.

Promoting Cross-Disciplinary Communication in Nanotechnology

Sarah Kriz
University of Washington

Karen Cheng
University of Washington

Marco Rolandi
University of Washington

Yeechi Chen
University of Washington

Abstract: Nanotechnology is a quickly growing field comprised of researchers from many disciplines who investigate nanoscale materials and phenomena. Because many well-established disciplines merge together to form the field of nanotechnology, a crucial aspect of nanotechnology education is promoting cross-disciplinary thinking and collaboration. We present a model that proposes a novel approach to multidisciplinary learning in nanotechnology. While existing educational solutions attempt to expose students to content from all of the nanotechnology disciplines, we focus on the development of visual communication skills as a means to promoting cross-disciplinary thinking and communication. We have developed a graduate course that combines the instruction of visual communication design principles with a studio component that allows students to create science graphics and reflect on how design choices relate to disciplinary goals and cross-disciplinary communication. We discuss the benefit of this course in the larger nanotechnology educational curriculum.

Nouns are more stable than Verbs: Patterns of semantic change in 19th century English

Eyal Sagi
Northwestern University

Abstract: It has been hypothesized in the literature that nouns are acquired earlier than verbs because they are more concrete and involve fewer relations. This hypothesis also predicts that the meaning of nouns should be more stable over time and across speakers. In this paper I use Latent Semantic Analysis of a 19th century literary corpus containing works from British and American authors to test this prediction.

I examined the variability in the vector representations of frequently used nouns and verbs based on the culture of the author and the time period. The results show the nouns vary less than verbs between the two cultures and across time. Moreover, these differences still exist when the concreteness of the words is taken into account. These results are consistent with the hypothesis that the relational nature of verbs contributes to their difficulty and variability beyond its effect on the verb's concreteness.

Context in distributed situated cognition

Hedda Rahel Schmidtke
Karlsruhe Institute of Technology (KIT)

Michael Beigl
Karlsruhe Institute of Technology (KIT)

Abstract: Ambient Intelligence (AmI) can be understood as a research effort towards physical environments that can use artificial intelligence techniques, in order to serve people in an intelligent, pro-active manner. AmI environments provide a unique, novel platform for studying and applying concepts of situated cognition and self-organization. In particular, we find that representations of context are crucial for AmI systems to perform these tasks. We follow the idea that the notion of context plays a central role with respect to economy, evolution, and architecture of cognitive systems. In particular, context can be understood to bridge the gap between the sensory stream and goal-directed reasoning. We present a logical language in which contexts, and not objects, properties, or propositions, are the primary entities. We show that, from this logical formalism, a corresponding symbolic-connectionist hybrid model of distributed, situated cognition can be derived.

Exploring Active Learning in a Bayesian Framework

Stephen Denton
Indiana University, Bloomington

John Kruschke
Indiana University, Bloomington

Abstract: Bayesian approaches provide a framework for models of active learning–learning in which stimuli are actively probed to disambiguate potential beliefs regarding outcomes. Within a Bayesian framework, uncertainty across beliefs is inherently represented and expected uncertainty reductions for candidate stimuli can be evaluated. Bayesian active learning models offer the prediction that an active learner would select the stimuli for which the expected uncertainty across all hypotheses is minimized. This research contrasts four possible hypothesis spaces for active learning consisting of two simple cue-combination models and two possible priors. An automated search of associative learning structures for which the models make maximally different predictions was performed. Participants were tested on these same structures in an allergy diagnosis context and were asked which cues they would find the most informative to learn about; i.e., their active learning preferences were assessed. Model and prior combinations that best mimic human active learning are discussed.

Linguistic Mediation of Visual Search: The effects of relative timing of speech and display

Eric Chiu
University of California, Merced (UCM)

Michael Spivey
University of California, Merced (UCM)

Abstract: Recent studies have shown that instead of a dichotomy between parallel and serial search strategies, in many instances we see a combination of both search strategies utilized. Consequently, computational models and theoretical accounts of visual search processing have evolved from traditional parallel or serial descriptions to labels of "efficient" and "inefficient." In the first experiment, we replicate previous findings regarding incremental spoken language comprehension on visual search processing utilizing a between subjects design. Next, a series of four experiments further explore the subtle timing of the influence of real-time language processing on visual search. The results provide further evidence toward understanding linguistically mediated influences on real-time visual search processing and support an interactive processing account of visual search and language comprehension.

Belief bias in judgments of sample-size adequacy

Richard Anderson
Bowling Green State University

Leisha Colyn
Bowling Green State University

Beth Hartzler
Bowling Green State University

Abstract: Previous research on syllogistic logical reasoning indicates that people are more likely to judge an argument as valid when they believe the argument's conclusion to be true. The present research assessed whether belief bias would also occur in intuitive statistical judgment. There were two versions of a judgment scenario (varying between subjects). Version A described an observer who, based on 100 observations, draws the conclusion that most Americans are left-handed. Version B was like A except the conclusion was that most Americans are right-handed. Participants' task was to assess the degree to which 100 is a sufficiently large sample to support a confident conclusion (i.e., that Americans tend to be left-handed, or that Americans tend to be right-handed). Participants judged the sample size to be more adequate when the argument conclusion was presumably believed (i.e., that "most Americans are right-handed") than when the conclusion was "most Americans are left-handed."

Must analysis of meaning follow analysis of form? A time course analysis

Laurie Beth Feldman
The University at Albany, SUNY & Haskins Labs

Fermín Moscoso del Prado Martín
CNRS & Université de Provence, France

Patrick A O'Connor
The University at Albany, SUNY & Haskins Labs

Abstract: Many models of word recognition assume that processing proceeds sequentially from analysis of form to analysis of meaning. In the context of morphological processing, some interpret the apparent absence of differences in recognition latencies to targets (RAT) in form and semantically similar (ratty-RAT) and in form similar and semantically dissimilar (ratify-RAT) prime contexts as consistent with this claim. We examined the time course over which degree of semantic similarity between morphologically related pairs influences recognition in the forward masked priming variant of the lexical decision paradigm. Across a range of SOAs., latencies were significantly faster after semantically similar than dissimilar primes, Results limit the scope of form-then-semantics models of recognition and demonstrate that semantic context influences even the very early stages of recognition.

Learning Cross-Modal Contingencies through Attentional Cues

Daniel Yurovsky
Indiana University

Rachel Wu
Birkbeck, University of London

Natasha Kirkham
Birkbeck, University of London

Chen Yu
Indiana University

Abstract: Infants must develop attentional mechanisms that support extraction of relevant information from a cluttered world. Though both social and non-social cues shift infants' attention, Wu and Kirkham (accepted) showed they produce qualitatively different learning effects in 4 and 8-month-old infants. While both types of cues led infants to attend preferentially to the relevant locations during training, cross-modal contingencies were learned only by older infants exposed to social cues. In this work, we analyzed the eye movement of these infants to shed light on the underlying attentional processes elicited by these cues. Using a dual-process model to link learning to looking (Yurovsky, Hidaka, Yu, & Smith, Cogsci Conference 2010), we characterized each infant's underlying learning function and used these functions to predict individual test results. We can thus understand the impact of social and non-social cues by examining how cue-elicited differences in exploratory behavior cascade into differences in learning.

The Effects of Alcohol on Working Memory and Change Detection

Gregory Colflesh
Georgia Institute of Technology

Andrew Jarosz
University of Illinois at Chicago

Jennifer Wiley
University of Illinois at Chicago

Abstract: The prevailing account of how alcohol affects attention suggests that intoxication reduces attentional focus and capacity. To better understand presumed cognitive consequences of intoxication, the present study tested the effects of moderate intoxication (.07 BAC) on both change blindness and complex span tasks. Change blindness tasks require finding a small change across alternating versions of a scene. Complex span tasks consist of interleaved processing and storage components. As expected, intoxication significantly decreased performance on the complex span tasks. But, surprisingly, it improved performance on the change blindness task. The results are interpreted as evidence that intoxication decreases attentional control, and causes a more diffuse attentional state. This can harm performance on some tasks where attentional control or focus are required, but may actually facilitate performance on other tasks.

Effects of the Exploration Perspective on Pointing Accuracy

Julia Frankenstein
University of Freiburg

Manuel Vidal
College de France, Paris

Michael Rouillé
IRISA / INRIA Rennes

Stéphane Donikian
IRISA / INRIA Rennes

Mohamed Zaoui
College de France, Paris

Alain Berthoz
College de France, Paris

Abstract: Abstract: We examined the influence of the perspective during exploration on the ability of subjects to point correctly to memorized targets in a virtual 3D environment. This environment consisted of a two-storied factory building with 32 machines on the ground floor. Four machines were marked as targets. Eight trials were conducted in each of the four different perspectives: map view from above, four different views from the corners, perceiving the environment as a person walking through it, and following an avatar through the environment. Subsequently, participants were asked to point to the four targets from the upper floor.

We expected the best performance in the map view as all information is given in a single reference frame and a rotation suffices for pointing. Surprisingly, eleven of the nineteen participants performed best in conditions different from the most straightforward. This finding indicates that different memorization strategies were used by different persons.

The Effect of Processing Type on Re-Categorization

David G. Cosejo
University of Illinois at Chicago

Stellan Ohlsson
University of Illinois at Chicago

Abstract: The effect of processing type on the process of overriding prior experience and learning – restructuring – was examined within a categorization paradigm. Participants were trained to categorize either explicitly, implicitly or received no training. Explicit training encouraged participants to use hypothesis-testing while implicit training encouraged categorizing via intuition or "gut instinct." Participants then worked on a modified categorization task in which they learned an initial "misconception" category and later had to restructure their representation to learn a target category. Participants were able to successfully learn the misconception and restructure to the target. Data suggest that increased category complexity results in a longer learning period that leads to the generation of a category representation that is more accessible and more readily restructured. The complexity effect is driven by similar performance across the different training types. The experimental and real-world implications for the interplay between stimuli complexity and processing type are addressed.

Fractioning Factors that Influence Phonological Word Form Learning

Libo Zhao
Department of Psychology, University of Iowa

Prahlad Gupta
Department of Psychology, University of Iowa

Abstract: Although learning of phonological word forms is important for mastering a language, little is known about the factors influencing it. We addressed this question by comparing phonological word form learning in two situations: learning novel word forms as the labels for referents (deliberate word learning); and learning novel word forms through incidental exposures to the word forms alone, without any referents (incidental word form learning). Phonological word form learning as measured by stem completion ability was found to be better in the former than in the latter situation (Experiment 1). Experiment 2 found that deliberate memorization of word forms, even without any referents, also yielded better stem completion ability than purely incidental learning. These results suggest that incidental word form learning may not yield full mastery of word forms, and that deliberate learning may be a necessary component for such mastery.

Large differences in the distribution of instances of common object-based categories in early childhood

Alfredo F. Pereira
Indiana University Bloomington

Karin H. James
Indiana University Bloomington

Susan S. SJones
Indiana University Bloomington

Linda B. Smith
Indiana University Bloomington

Abstract: Few studies have documented the examples of common early-learned object-based categories children actually encounter. This study asked parents (N = 10) to record, using a digital camera, the concrete instances their children saw (M = 16 mo). For five days, if an object was labeled with one of the nouns inside a pre-determined list of eight nouns, parents were to take a photo–our final dataset consisted of 700 photos. We coded the contents of each photo as: 3D real object, 3D realistic toy, 3D simple shape toy, 2D realistic object, and 2D simple shape.

Our results show large differences between categories in terms of type of exemplars: mostly composed of 3D real objects, with a mixture of 3D and 2D variability, or only experienced as 2D images.

These results are relevant to theories of visual object categorization–e.g. in understanding viewpoint invariance, or perception of abstract structural shape.

Phonetic symbolism for size and shape

Patrick Thompson
University of Warwick

Zachary Estes
University of Warwick

Abstract: Many previous studies of phonetic symbolism, wherein the sounds of a word convey the referent's attributes, have confounded multiple attributes such as size and shape. In the current study, participants viewed novel objects of varying size and shape and were asked to rate the appropriateness of a spoken non-word as a name for the object. Size and shape interacted (e.g., higher ratings were given for names with front vowels like /i/ when the object was small and spiky), and both spiky and round objects were phonetically marked. However, participants tended to use phonemes to mark size only when the object is large. This suggests that marking via phonemes should not be assumed to simultaneously mark all physical properties of the object (i.e., both size and shape). Consequently, phonetic symbolism of physical properties may not correlate as neatly with gesture as has previously been thought.

Using Embodied Cognition in the Instruction of Abstract Programming Concepts

Cameron L. Fadjo
Teachers College, Columbia University

John B. Black
Teachers College, Columbia University

JeeHye Hong
Teachers College, Columbia University

Chun-Hao Chang
Teachers College, Columbia University

Abstract: We present two models, physical and imaginary, for implementing embodied cognition during the instruction of abstract programming concepts. We examine previous studies using embodiment in the instruction of reading (Glenberg et al., 2004), mathematics (Goldstone & Landy, 2010), and science (Chan & Black, 2006a, 2006b) as a foundation for proposing an embodied instruction of programming. We discuss the embodied instruction of abstract symbols in mathematics and suggest that the nature of programming a video game (Fadjo et al., 2009a, 2009b) provides adequate grounding (Barsalou, 2008) for the instruction of abstract conditional statements. We suggest that an embodied form of instruction integrates the actions prevalent in two-dimensional video games with the instruction of abstract programming concepts. We discuss our findings on using Instructional Embodiment (Fadjo et al., 2009a) to improve novice programmer's tracing and conditional logic thinking skills.

Decision-Making in Older Adults: Sometimes Older is Wiser

Darrell Worthy
University of Texas at Austin

W. Todd Maddox
University of Texas at Austin

Abstract: We examined the performance of younger and older adults in a dynamic decision making task that required exploring long-term increasing options that had worse short term gains, but that eventually yielded higher rewards. Results indicate that older adults are more willing to sample the long-term increasing option earlier in the task than the Younger adults. We employed a model-based approach that modified a simple reinforcement learning model to allow for biases to 'stay' or 'switch' following a response (e.g. Otto et al., in press). Using this approach we were able to characterize exploration of alternative options or exploitation of the best option as 'targeted' or 'untargeted' responding. We found that Older adults were best fit by the targeted exploration model. This suggests that while Younger adults tend to repeatedly select short-term advantageous options, Older adults engage in greater exploration of the decision space, and this leads to better performance.

Coordination dynamics in speech and lexical semantics

Christopher Kello
University of California, Merced

Theo Rhodes
University of California, Merced

Geoff Hollis
University of Cincinnati

Bryan Kerster
University of California, Merced

Abstract: Variations in individual human behavior are intrinsically long-range correlated (i.e. 1/f noise). These correlations may reflect interdependence (i.e. coordination) among components at various scales of structure and dynamics (e.g., neurons, cortical columns, brain areas, brain-body interactions). Two experiments tested whether long-range correlations emerge when two people interact. In experiment 1, perceptual-motor coordination was invoked by instructing participant pairs to coordinate uttering the word "mom" in alternation with key taps. In experiment 2, semantic coordination was invoked by alternating free word associations (e.g. one says "cat", the other says "dog", first says "collar", second says "shirt", and so on). Fluctuations in series of mom-tap intervals were long-range correlated. Long-range correlations were also found in semantics fluctuations of free word associations, the latter being weaker and derived from lexical co-occurrence statistics. Results indicate that common principles of coordination apply across individual and dyadic scales of perceptual-motor and cognitive performances.

Knowing who knows what best: Preschoolers selectively use others' past accuracy in causal learning

Chris Vredenburgh
Cornell University

Lauren Schneider
Cornell University

Andy Hsia
Cornell University

Tamar Kushnir
Cornell University

Abstract: Knowing who knows what best: Preschoolers selectively use others' past accuracy in causal learning
Authors: Christopher Vredenburgh, Lauren Schneider, Andy Hsia, and Tamar Kushnir

Preschoolers use a person's past accuracy labeling common objects as a cue to their "trustworthiness" when learning new words. These studies investigate how children use past accuracy to differentially trust people for causal learning. Across two studies, we found that children (Mean age = 4 years, 3 months; SD = 4 months) would differentially trust an accurate (versus inaccurate or ignorant) labeler for learning a novel causal function but not for learning a novel causal mechanism. However, when an accurate labeler was pitted against a causal expert, children trusted the expert with both causal mechanism and causal function. These studies demonstrate that children show selective trust of accurate labelers depending on the availability of sources with causal expertise. Thus, children are sensitive to domains of knowledge and can use them to request information from the most appropriate source.

The effects of perspective on understanding of projective spatial terms

Takatsugu Kojima
Kyoto University

Abstract: We use projective spatial terms to indicate a location and a direction in communications about both real space and three-dimensional computer graphics (3DCG) space. However, it is often possible to also use an overhead perspective or the perspective of a communication partner, especially in a communication system using 3DCG space, in addition to our own first-person view. This study focused on the effects of these three perspectives on understanding projective spatial terms in a 3DCG system. In this experiment, we used four projective spatial terms (front, back, left, and right) and 3DCG stimuli based on three views (a participant's first-person view, an overhead view, and a communication partner's view). We investigated how the three views influenced understanding of the spatial terms. The results show that an overhead view had little effect on understanding of the spatial terms and that the communication partner's view had considerable effect on such understanding.

Inferring Object Structure from Human Action at 9 Months

Stephen Killingsworth
Peabody College, Vanderbilt University

John Jacobson
Peabody College, Vanderbilt University

Megan Saylor
Peabody College, Vanderbilt University

Abstract: This study investigates whether 9-month-olds use action information to make predictions about the hidden structure of an object. Two groups saw an actor repeatedly raise and lower a box. In one group, the box was moved with a hidden handle. In the other group, a box with no handle was grasped along the hidden back face and repeatedly raised and lowered. Following this familiarization, the box was rotated 90 degrees either to reveal a structure consistent or to reveal a structure inconsistent with that suggested by the initial action. Patterns of looking between familiarization and test trials differed for the two familiarization groups, suggesting that 9-month-old infants can infer certain details of an object's structure from human action.

Social Indexing: How the People Around Us Aid Cognition

Chris N.H. Street
University College London (UCL)

Daniel. C Richardson
University College London (UCL)

Abstract: We hypothesise the existence of a specific attentional mechanism, social indexing, that directs gaze towards people in our environment who are relevant to moment by moment cognitive processing. Whilst Festinger (1954) proposed that individuals will depend upon and conform to similar others in times of uncertainty, the social indexing hypothesis posits that individuals will actively seek out information from those we perceive to be able to provide socially relevant information. Such an attentional mechanism would allow cognition to be situated in both the physical and social world (Hutchins, 1995). In an early demonstration, Crosby, Monin and Richardson (2008) showed that gaze is directed towards the target of potentially offensive remarks when they might provide an informative response. We discuss whether these results extend to both positive and negative remarks, and whether participants' confidence in their ability to interpret social situations will result in more or less social indexing behaviour.

Impact of Diverse Abilities on Learning to Write through Peer-Review

Melissa Patchan
University of Pittsburgh

Christian Schunn
University of Pittsburgh

Abstract: Theoretically, there are advantages to working with students of the same ability and different ability (Lou et al., 1996). In order to determine how students' ability affects the peer-review process, students' writing ability (e.g., high-ability versus low-ability) was first determined. Then students' were randomly assigned to review either four high-ability peers' papers or four low-ability peer's papers. In return, they received feedback from either four high-ability peers or four low-ability peers. The quality of students' second draft of their first paper and the quality of the first draft of a second paper were analyzed to determine whether students' learning was affected by the feedback they provided to high-ability versus low-ability students and by the feedback they received from high-ability versus low-ability students. In addition, several mediators (e.g., motivation, amount and type of feedback) were examined to explain the learning differences.

Look who's talking (and follow the leader)! Eye movements in a social interaction reveal effects of speaking and social status

Tom Foulsham
University of British Columbia (UBC)

Joey Cheng
University of British Columbia (UBC)

Jessica Tracy
University of British Columbia (UBC)

Joseph Henrich
University of British Columbia (UBC)

Alan Kingstone
University of British Columbia (UBC)

Abstract: Human visual attention tends to operate in an environment that is complex, dynamic and social. However, experimental investigations of where people direct their attention often neglect these factors. In our research, we recorded people making decisions in groups of 3 and then showed video clips of these situations to new participants while monitoring their eye movements. This provided a rich record of how people distributed their gaze on a moment-by-moment basis. Observers tended to look at the person who was talking at any one time, and they fixated this person slightly before they started to speak. Higher-level social attributions also had an effect: people who were rated as having high social status were gazed at more often, over and above the effects of speaking. These effects show that the gaze system is extremely sensitive to the complexities and dynamics of the current social context.

Using analogical learning in science curricula to improve conceptual understanding

J. Elizabeth Richey
Learning Research and Development Center, University of Pittsburgh

Alicia Chang
Learning Research and Development Center, University of Pittsburgh

Timothy J. Nokes
Learning Research and Development Center, University of Pittsburgh

Christian D. Schunn
Learning Research and Development Center, University of Pittsburgh

Abstract: The goal of the 21st Century Center for Research and Development in Cognition and Science Instruction (CaSE) is to improve middle school students' science learning by systematically applying cognitive science principles in the revision of instructional materials and teacher professional development. In this presentation we focus on the application of analogical learning principles to two popular middle school science curricula. Analogical learning through comparison is a powerful activity for facilitating the acquisition of critical features and concepts underlying concrete examples and preparing students for future learning. We instantiated these principles by creating "contrasting cases" to introduce abstract science concepts in conceptually driven lessons that could easily be inserted into existing curricula. We assessed the effectiveness of this intervention with tests targeting both the concepts that were covered in the cases as well as transfer concepts taught later in the curriculum. We will present preliminary data from pilot teachers.

Modeling age of exposure in L2 learning of vowel categories

Meghan Clayards
McGill University

Joseph Toscano
University of Iowa

Abstract: Age of exposure is known to be an important indicator of second language proficiency. Native-like phonological proficiency is attained only by learners exposed at the earliest ages. This paper examines one account of age-of-exposure effects. Two computational models (a mixture of Gaussians and a neural network) were trained without supervision on F1 and F2 tokens based on production data from two different vowel systems (Quichua and Spanish; Guion, 2003). Both models learn the individual phonological systems when trained on monolingual distributions. When exposed to bilingual data, both models also achieve varying degrees of success depending on when the second language (Spanish) is introduced in training, paralleling data from bilingual speakers with different ages of acquisition (Guion, 2003). This demonstrates that learners may be restricted in learning a second language not because of a biological critical period, but by the commitments that the system has already made to the first language.

Mental models of virology in experts and novices

Benjamin Jee
Northwestern University

David Uttal
Northwestern University

Caroline Crouch
Northwestern University

Amy Spiegel
University of Nebraska-Lincoln

Judy Diamond
University of Nebraska-Lincoln

Abstract: Viruses are invisible and their effects, though often experienced, arise through mechanisms that may be poorly understood by many people. The present work examined what people with different levels of virology expertise think and believe about viruses. We conducted detailed, semi-structured interviews about virus infection, replication, transmission, and other topics with a group of middle-school students, science teachers, and expert virologists. Participants' responses were coded for content and used to establish their mental models for several key topics (cf. Hmelo-Silver & Pfeffer, 2004). Analyses revealed that the experts' mental models were greatest in depth and breadth. Many of the students—and several teachers—possessed scientific inaccuracies and inconsistencies in their mental models. By capitalizing on experts' knowledge organizations and by targeting common misconceptions about viruses found in students and teachers, it will be possible to develop materials and tools for increasing people's understanding of viruses and the microbiological world.

Facilitating Educator Evaluation of Online Instructional Materials: Does Conceptual Browsing Impact Cognitive Processing?

Kirsten Butcher
University of Utah

Robert Zheng
University of Utah

Anne Cook
University of Utah

Lisa Ferrara
University of Utah

Sarah Davies
University of Utah

Ashley Crockett Mazal
University of Utah

Aaron Dewald
University of Utah

Abstract: A key challenge for beginning educators is finding high-quality online materials that will support deep learning in their classrooms. Identifying and evaluating effective digital resources requires careful attention to the match between domain learning goals and the conceptual information contained in resources, but preservice teachers often lack strong prior knowledge that would facilitate such processing. Conceptual browsing interfaces may support deeper cognitive processing by providing a visual representation of the conceptual relationships between domain ideas and by providing a direct retrieval mechanism to find specific online resources related to key domain ideas. Using a combined think-aloud and eye-tracking study, we are examining the effects of conceptual browsing vs. keyword searching on the cognitive processes of preservice teachers performing educational tasks. In this poster, we summarize preliminary results and discuss how keyword search vs. conceptual browsing interfaces can impact the depth with which beginning educators process online information.

Are Hindu-Arabic Numerals Concrete or Abstract Symbols?

Percival Matthews
Vanderbilt University

Abstract: Much recent experimental work has investigated the comparative merits of concrete versus abstract instantiations of to-be-learned concepts for promoting learning and transfer in complex domains. A critical question, however, is what exactly counts as a concrete instantiation. Using a design similar to that of Sloutsky, Kaminski, & Heckler (2005), the current experiments provide findings that suggest that Arabic numerals have effects that parallel those of perceptually concrete instantiations when used to illustrate the domain of modular arithmetic. Specifically, numerals can be used to speed initial learning within the domain, but impede transfer relative to more abstract instantiations of the same underlying content. Moreover, these effects can be moderated by subtle warm-up tasks that affect the degree to which these numerals activate prior arithmetic schemas. These results suggest that the current conception of "concrete" should be expanded to include representations typically thought to lie outside of the realm of concrete.

Order Effects in Categorization: Identifying "the Nuts" in Poker

Brian D. Gane
Georgia Institute of Technology

Richard Catrambone
Georgia Institute of Technology

Abstract: Research in concept learning indicates that the order of example instances affects acquisition of conceptual structures. There is less research, however, regarding how example order affects categorization skill. Might the order of training examples affect categorization even after the concepts have been learned? Participants were trained to categorize sets of playing cards into the best possible poker hand. Training followed either a blocked (the best hand remained the same for contiguous trials) or a mixed order (the best hand did not repeat more than twice in a row). Preliminary results suggest that example order affects categorization reaction time (RT) during acquisition training: the blocked order reduced RT. This trend reversed, however, during transfer trials: the mixed group had lower RT. These findings suggest that example order plays a role in developing categorization skill. We offer a preliminary explanation regarding how participants' strategy develops based on the order of training examples.

Semantics in the wild: Context-sensitive inferences about mammals

Jeremy Glick
Stanford University

James McClelland
Stanford University

Abstract: Several accounts of semantic representation have relied on a contextually insensitive similarity space, including recent structured probabilistic approaches (Kemp & Tenenbaum, 2009). However, evidence for these models relies on participants' inferences about a particular kind of property, namely, biological properties. We show that the training set used to extract a single tree structure by Kemp and Tenenbaum (2009) in fact contains additional structure in the relations between properties of different types. Moreover, participants who are asked to make inferences about different kinds of properties (biology, diet, habitat) show generalization differences that reflect this additional structure. We suggest that models of semantic representation must be able to dynamically adjust their representations in a context-sensitive manner, and we will present simulation results using a model that can do so.

Judgements of relative order: Mechanisms underlying subspan versus supraspan lists

Yang Liu
University of Alberta

Michelle Chan
University of Alberta

Jeremy Caplan
University of Alberta

Abstract: Judging the relative order of materials is a core function of human memory. In short, subspan consonant lists with immediate judgments of relative recency (JOR), instruction wording ("which item was presented earlier?" versus "which item was presented later?") could flip around memory search direction (Chan et al., 2009). We wondered whether instruction wording could have an analogous influence on the JOR judgement in supraspan lists. However, supraspan lists typically show a very different behavioural pattern - distance effects (e.g., Yntema & Trask, 1963). Our participants performed JOR judgements on "short" (LL=8) supraspan noun lists. We evaluate whether it is possible to reconcile the subspan and supraspan data by assuming that the judgement in both sub- and supra-span regimes are influenced by the same factors, positional discriminability and attentional bias across serial positions, and that speed-accuracy tradeoffs combined with ceiling in subspan lists account for the observed qualitative differences in behaviour.

Verb-body part associations across users of English, Telugu, and Hindi

Raju Bapi
Centre for Neural and Cognitive Sciences, University of Hyderabad

Jigar Patel
Centre for Neural and Cognitive Sciences, University of Hyderabad

Viswanath Naidu
Language Technologies Research Centre, IIIT, Hyderabad

Sireesha Jala
Memory Clinic, Dept of Neurology, Nizams Institute of Medical Sciences, Hyderabad

Vasanta Duggirala
Dept of Linguistics, Osmania University, Hyderabad

Suvarna Alladi
Memory Clinic, Dept of Neurology, Nizams Institute of Medical Sciences, Hyderabad

Abstract: Verb-body part association data were obtained from 36 adults using English (a second language), and Telugu and Hindi (as first languages). A set of 100 action verbs in English (same as those used in Maouene et al, 2008) and their equivalents in Telugu and Hindi served as target stimuli. Three groups of young adults (12 per language) provided written responses to action verbs printed in each language. There was greater agreement for verbs involving actions of hand, mouth and leg compared to the other body parts across all three languages. However, there were some language specific findings with respect to number of unique body parts specified, and the number of body parts required to cover 80

Reasoning in pedagogical versus deceptive situations

Russell Warner
University of Louisville

Todd Stoess
University of Louisville

Patrick Shafto
University of Louisville

Abstract: The majority of human learning occurs in social situations. In such situations, people may be cooperating or competing, and these different intentions may affect what kinds of information people exchange. Recently, Shafto and Goodman (2008) formalized a Bayesian model of reasoning in pedagogical situations – situations in which a knowledgeable teacher cooperates with a learner. This and other research suggests that individuals understand what information is more helpful and can use the knowledge of a person's intent to facilitate learning. We extend this model to apply to both the pedagogical and deceptive situations. We present a new experiment comparing reasoning in pedagogical and deceptive games. In the experiment, participants play the role of teacher/deceiver or learner/reasoner in a series of games. The results show that people converge to strategies predicted by the model, that people's behavior differs in pedagogical and deceptive conditions, and an analysis of individual games reveals interesting dynamics. We discuss the implications for models of learning in social situations.

Interactions Between the Fast and Slow Mental Processes

William Kennedy
George Mason University

Magdalena Bugajska
Naval Research Laboratory

Abstract: Our actions seem to be controlled by two separate types of mental processes: one fast, automatic, and unconscious and one slow, deliberate, and conscious. With the attention in the literature focused on the characteristics of the two processes and whether to include emotions, we do not find any discussion of how they interact. We present evidence that the slower process is not able to perceive the operation of faster process, but it can perceive the environmental stimulus common to both processes and the response of the faster process. It can then generate its own more deliberate response, possibly contrary to the faster process's response. We also provide evidence that the slower process is sometimes able to inhibit the fast process's response, but with effort. We present common experiences as well as cognitive theory and neurological studies in support of our description theory of the interactions of the two processes.

Word Length Effects and the Serial vs. Parallel Debate in Connectionist Models of Reading Aloud

Alan H. Kawamoto
UC Santa Cruz

Abstract: In reading aloud, naming latency (i.e., reaction time) increases linearly as the number of written symbols in a word increases (e.g., Rastle, Havelka, Wydell, Coltheart, & Besner, 2009). Advocates of dual-route models have argued that these effects can be accounted for by connectionist models that have a serial processing component, but not by models that are completely parallel when a single fixation is assumed. However, this conclusion is valid only for the specific assumptions made in parallel models implemented to date, and is not valid more generally. As Townsend and colleagues (Snodgrass & Townsend, 1980; Townsend, 1972) have argued, parallel models can mimic linearly increasing RTs as the number of items to be processed increases if stochastic processing times, limited processing capacity (i.e., attention), and a self-terminating stopping rule (i.e., the criterion to initiate articulation corresponding to the segment) are considered.

Differential effects of dopamine dysfunction on context usage in people with autism and schizophrenia: A computational exploration

Trent Kriete
University of California, Merced

David C. Noelle
University of California, Merced

Abstract: The ability to utilize contextual information in a flexible manner is vital for the successful navigation of our lives. People with autism demonstrate serious problems on tasks requiring the integration of contextual information across experiences. One such task is the determination of the proper meaning of an ambiguous word in a sentence. Homographs are words with one spelling, but different meanings, such as "bow" and "tear". People with autism appear unable to utilize sentential context in order to determine the correct meaning of a homograph. Instead, they rely on the statistically most frequent meaning. We present a neurocomputational model that suggests that these difficulties arise from a deficit in the flexible updating of attentional control, driven by dysfunctional interactions between the prefrontal cortex and the midbrain dopamine system. This work is compared to a previous computational account of the effects of abnormal dopamine levels on context processing difficulties in schizophrenia.

Can Statistics Change our Minds? The Role of Causal Explanation in Accommodation of Base Rate Statistics

Edward Munnich
University of San Francisco

Saera Khan
University of San Francisco

Melissa Latham
University of San Francisco

Michelle Brewer
University of San Francisco

Valesia Ho
University of San Francisco

Sierra Walton
University of San Francisco

Abstract: Can statistics change our minds? Base rate statistics are taken into account when they are causally relevant (e.g., Tversky & Kahneman, 1980). However, evidence is mixed on whether causally-relevant statistics can drive revision of existing beliefs: Hagmayer and Sloman (2009) found that those providing direct causal explanations for statistical patterns were more likely to recommend action than those providing incidental explanations. By contrast, Hewstone et al. (1988) found that the causal relevance of statistics only affected assignment of guilt when consistent with one's prejudices. To examine the extent to which beliefs can be revised due to statistics, we asked participants about statistics before and after feedback. Specifically, participants estimated quantities (e.g., U.S. traffic fatalities), provided explanations for trends in those statistics, and indicated what action they would take. We then provided actual statistics, and either a direct or incidental causal explanation for trends—and observed changes in actions participants would take.

Handedness and Hand Used Differentially Affect Object Facing

Jyotsna Vaid
Texas A&M University, College Station

Hsin-Chin Chen
National Chung Cheng University, Taiwan

Rebecca Rhodes
University of Michigan, Ann Arbor

Sumeyra Tosun
Texas A&M University, College Station

Abstract: When producing line drawings of common objects with an intrinsic front, directional biases are observed in starting location, stroke direction, and figure orientation, Previous studies of drawing directionality have predominantly examined right handers and/or have considered dominant hand drawing performance only. By contrast, the present study compared drawing directionality of right vs. left handers drawing objects with their dominant and non-dominant hands. Object facing direction was found to differ significantly as a function of handedness and hand used. Whereas right handers' orientation preference was generally unaffected by hand used to draw, left handers tended to show a stronger right-facing bias when drawing with their left hand than when drawing with their right hand. The latter finding is consistent with a biomechanical account in terms of hand movement asymmetries. Additional accounts of drawing asymmetries are addressed and their implications explored for current views on the relationship between perception, cognition, and action.

Experience word learning predicts children's ability to generalize novel labels

Emily Thom
University of California, Los Angeles (UCLA)

Catherine Sandhofer
University of California, Los Angeles

Abstract: Recent research has suggested that rapid world learning develops within specific categorical domains as the result of previous within the domain (Thom & Sandhofer, 2009). The current studies further examined the relationship between children's previous experience learning words within a category and their ability to extend additional words within the category. In Study 1, children's extension of novel labels was compared across three common categories in relation to their existing vocabulary size within that category. In Study 2, children were trained in a greater or fewer number of category exemplars in two common categories, then tested in their ability to generalize new labels within each of these categories. Preliminary results indicate that greater experience learning words within a category predicts better extension within that category, but not in other categories, suggesting the development of rapid word learning is domain-specific, and occurs as the result of experience learning words within each category.

A Biologically Plausible Account of the Computational Utility of Consciousness

William B. St. Clair
University of California Merced

David C. Noelle
University of California Merced

Abstract: According to Mathis and Mozer (1996), visual awareness requires internal representations to be stable over time. They demonstrated that attractor dynamics in a hand-wired, abstract, connectionist model could both produce this stability and also explain behavioral differences between conditions of subliminal and supraliminal stimulus presentation. One such demonstration involved a lexical decision study by Marcel (1980), in which conscious perception of an ambiguous prime word, disambiguated by previous context, sped lexical decision of a subsequent target word only when the target was related to the context-cued meaning of the prime. In contrast, subliminal presentation of the prime produced facilitation for targets related to either meaning of the prime. Here, we show that the attractor dynamics needed to explain this effect naturally arise from the balance of excitatory and inhibitory connections in cortex, as modeled in the biologically constrained Leabra framework, providing some neuroscientific support for this account of visual awareness.

Perceiving the Other during Joint Action

Jerome Scott Jordan
Illinois State University

Andrew Kenning
Illinois State University

James Clinton
Illinois State University

Justin Durtschi
Illinois State University

J. Cooper Cutting
Illinois State University

Abstract: The perceived vanishing point of a moving stimulus is displaced beyond the actual vanishing point. This forward displacement (FD) decreases with implied friction (i.e., the stimulus appears to move across a surface). The effect reverses when participants control stimulus movements (via right- and left-key presses) versus observe them. This reversal is consistent with economy-of-action (EOA) effects in which variables such as perceived pitch are influenced by the energy-demands implied by a stimulus (e.g., a steeper hill). The present poster presents experiments that reveal EOA effects when two participants control stimulus movements together, each having access to one of two control buttons. Specifically, FD increases across implied friction, regardless who controls the stimulus when it vanishes. Since participants are basically observers as the other participant controls the stimulus, the increase of FD during such observation indicates participants perceive the other-controlled stimulus movements in terms implied effort (i.e., EOA).

A Model of Cognitive Rehabilitation: Recovering with Constraints

Shin-ichi Asakawa
Tokyo Woman's Christian University

Yoshihiro Itaguchi
Waseda University

Abstract: Neural network modeling offers a useful computational framework for exploring the nature of normal and impaired cognitive processes. Among such modelings, Hinton and Shalllice (1991) and Plaut (1996) lesioned recurrent neural networks and investigated the degree of recovery through retraining. However, many ways of brain damages has remained to be unresolved. The current works propose the method of constraints in which brain damages might occur. In order to understand the nature and variability of recovery in patients, we examined both simple three layerd perceptrons and attractor networks with lesions in various parts of networks, and observed recovery processes in retraining. The findings in this study revealed conditions to recover from brain damages and suggests good and proper ways of therapy in which therapists have to select training words to maximize generalization.

Restructuring representations in analogy making by children: the role of cognitive flexibility

Jean-Pierre Thibaut
LEAD

Robert French
LEAD

Yannick Gerard
LEAD

Abstract: In classical A:B::C:D analogies, it is often assumed that participants first find a relation between A and B, which is then transferred to C and D. By contrast, we hypothesized that the first interpretation of A-B must sometimes be later revised, given the nature of C and the D available in the solution set. We hypothesized that young children (5-6 year-olds) would encounter difficulties when restructuring is necessary because restructuring requires cognitive flexibility which is less developed in young children. In an A:B::C:D task, we compared analogies requiring restructuring with analogies that did not. We also compared analogies based on weakly semantically associated pairs (e.g., child-bed) with analogies based on strongly semantically associated pairs (e.g., dog-bone). Results revealed an interaction in which the difference between restructuring and no restructuring was significant only for analogies based on weak semantic associations. They were discussed in terms of executive functions.

White- and Grey-Matter Damage Differentially Impair Learning and Generalization in a Computational Model of the Raven Matrices Task

Vincent G. Berthiaume (Vincent.Berthiaume@McGill.ca)
Department of Psychology, McGill University, 1205 Dr. Penfield Avenue
Montréal, QC H3A 1B1 Canada

Thomas R. Shultz (Thomas.Shultz@McGill.ca)
Department of Psychology and School of Computer Science, McGill University, 1205 Dr. Penfield Avenue
Montréal, QC H3A 1B1 Canada

Olaf Dammann (ODammann@TuftsMedicalCenter.org)
Division of Newborn Medicine, Floating Hospital for Children at Tufts Medical Center, 800 Washington Street, Box 854
Boston, MA 02111 USA

Abstract

Many preterm neonates have white-matter damage (WMD, damaged connections between neurons) and grey matter-damage (GMD, dead neurons). These children are known to have lower IQs than their full-term peers, yet the mechanisms underlying this association are poorly understood. We designed a developmental connectionist model of the Raven Matrices IQ task in which (1) all neurons had intact output, simulating normal development, or (2) half the neurons had noisy output, simulating noisy transmission or WMD, or (3) half the neurons had no output, simulating cell death or GMD. We found that damage increased task error. Further, WMD was worse than GMD overall, yet GMD was at once worse for generalization problems not given in training and better for training problems. Our model is the first to simulate an effect of perinatal brain damage on a cognitive task, and predicts that different types of brain damage may lead to different cognitive impairments.

Keywords: White-matter damage; cortical damage; preterm birth; Raven Matrices; IQ; connectionism; learning.

Background

In 2007, 12.7% of all births in the United States were preterm, an increase of over 2% since 1990 (Heron et al., 2009). This increase inevitably exacerbates family distress and healthcare costs, as children born preterm present many cognitive and developmental impairments compared to their full-term peers, including lower IQ scores (Bhutta, Cleves, Casey, Cradock, & Anand, 2002). The severity of preterm children's cognitive deficits appears to be correlated with brain abnormalities, e.g., reduced volume in specific brain regions (Peterson et al., 2000), which may result from abnormal development following perinatal brain damage (Robinson, 2005). Indeed, preterm neonates have immature brains that are likely to suffer damage from prematurity-associated adverse exposures before and after birth.

Perinatal brain damage can occur in either of the two major macroscopically distinct areas of the brain, the white (Dyet et al., 2006) and grey matter (Burd et al., 2009). White matter is made up of myelinated axons connecting neuronal regions and is the matter principally damaged in preterm brains (Leviton & Paneth, 1990). By contrast, grey matter consists of neuronal cell bodies and its damage is usually more constrained in the preterm brain (Billiards, Pierson, Haynes, Folkerth, & Kinney, 2006). Although the association between cognitive impairments and brain damage is well known in the pediatric community, not much is known about either the general mechanisms underlying the association (Counsell et al., 2008), or more specifically, about how damage to white or grey matter may potentially affect cognitive function differentially. Although a previous computational model indicated that white-matter damage may be worse than grey-matter damage for synaptic recovery (Follett, Roth, Follett, & Dammann, 2009), that model did not implement any cognitive task and thus did not inform us about the effect of damage on cognition.

In order to explore how white- and grey-matter damage may affect cognitive ability, we designed a computational developmental model of a popular IQ task, the Raven Matrices, and incorporated white- and grey-matter damage in the model to assess their effects on task performance.

Computational Developmental Algorithm

Sibling-Descendent Cascade-Correlation (SDCC, Baluja & Fahlman, 1994) is a supervised-learning, artificial-neural-network algorithm which benefits from fast and powerful learning and implements some psychologically- and neurologically-plausible mechanisms (Shultz, 2006; Shultz, Mysore, & Quartz, 2007). Its developmental or constructive aspect comes from the fact that networks initially have only input and output units (fully interconnected with random weights), but develop by recruiting hidden units, as required to reduce error in training.

Training includes output and input phases. Networks are first given training patterns (input and target patterns), and training enters the output phase, in which the algorithm reduces output error, the discrepancy between output activation (initially random) and the target patterns. If the algorithm cannot bring error lower than the Score Threshold (ST) parameter, left at its default value of .4 for all training patterns, training switches to the input phase. In the input

phase, the network selects the one hidden unit, out of a pool of 8 randomly-initialized candidate recruits, that correlates most with output error. This selected unit is integrated into the network and training switches back to output phase. Training usually stops as soon as error for each training pattern drops below the ST. However, in order to have consistent amount of training across all types of networks, we imposed here a training limit of 14 hidden units and 2500 epochs, based on the average training cost of an independent, undamaged sample of 100 networks.

At the end of training, networks are tested by freezing connection weights (so that networks do not learn during testing), and measuring output error on testing patterns.

Raven Matrices task

The Raven Matrices task consists of a series of problems, in which subjects have to study a 3-by-3 matrix, and chose amongst 8 alternatives the figure that best fits the empty spot in the matrix (Figure 1).

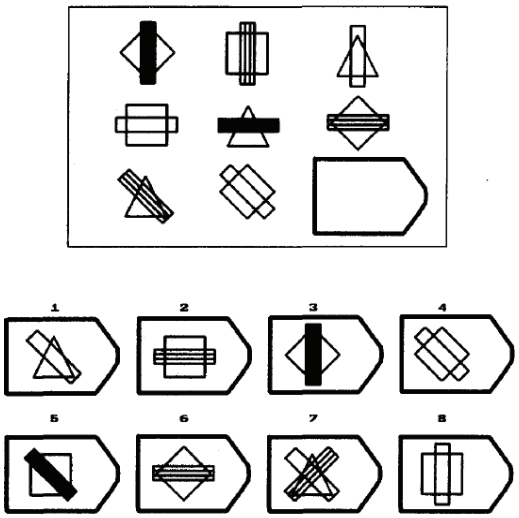


Figure 1. An example Raven problem. Copyright © 1990 by the American Psychological Association. Reproduced with permission from Carpenter, Just, and Shell (1990). The use of APA information does not imply endorsement by APA.

There are four rules (Carpenter et al., 1990) for predicting the missing figure. In the constant-in-a row rule, a figure feature is constant across rows. For example, the narrow rectangle in Figure 1 is always vertical in the first row, horizontal in the second, and diagonal in the third. In the distribution-of-three rule, a feature is distributed amongst the figures in a row, e.g., the narrow rectangle is either black, striped, or transparent in each column in Figure 1. If one of the three features is absent, the distribution-of-three rule can also cover a distribution-of-two-values rule, sometimes considered as a separate rule. In the quantitative-pairwise-progression rule, figure attributes (such as small squares in a grid) increment or decrement between adjacent columns. In the addition and subtraction rules, a figure feature from column 1 is added to or subtracted from a figure in column 2 to produce a third figure in column 3.

Methods

We used SDCC to train and test undamaged networks on the Raven Matrices task. We next incorporated damage in two different groups of networks by either randomizing (white-matter damage) or blocking (grey-matter damage) the output activation of approximately half the networks' neurons.

Undamaged Training and Testing

A first group of 100 undamaged networks were trained and tested on Raven task problems that each implemented one of the four rules identified by Carpenter and colleagues (1990). Performance was evaluated on problems that networks knew about, and on novel problems, a technique somewhat similar to some psychological studies using the Raven task (e.g., Skuy et al., 2002).

Networks had eight inputs corresponding to the eight figures constituting a Raven problem, and one output corresponding to the missing ninth figure. Inputs and outputs used linear activation functions to cover the range of possible input and output values (see below). In order to compare network performance on known and novel data, two datasets of equal size were constructed: the training and generalization sets. Figure 2 illustrates an example Raven problem coded for training and generalization patterns.

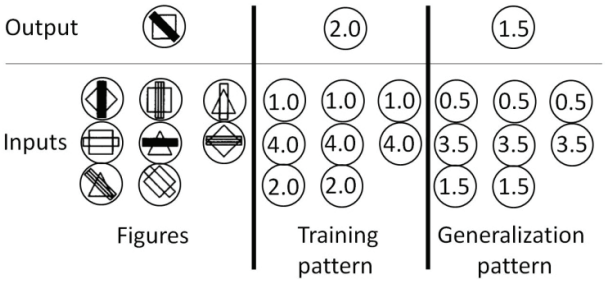


Figure 2. A Raven problem represented in figures and as a training pattern, and its derived generalization pattern.

The left-most panel of Figure 2 shows the example figures, and the middle panel shows how the figures may be coded as a training pattern. For each training pattern, selected features were coded by integers (chosen at random between 1 and 4 from a uniform distribution) that represented the figure feature relevant to the problem rule. Each training pattern implemented one of the 4 rules identified by Carpenter and colleagues, (1990). For instance, in this constant-in-a-row example problem, 1.0 represents a vertical bar, 4.0 an horizontal one, and 2.0 a diagonal one.

The right panel shows a generalization pattern, obtained by subtracting .5 from every value of the example training pattern. Following previous practice (Dandurand, Berthiaume, & Shultz, 2007), generalization patterns were all obtained using this calculation (although in feature-addition and -subtraction problems, .5 was only subtracted

from numbers in the first two columns, because the third value depended on the first two). Other types of problems were coded similarly. Distribution-of-three problems had one of three numbers appear in each column. Quantitative-pairwise-progression problems were represented by an increment or decrement of numbers across adjacent columns. Addition and subtraction problems had a number from the second column added to or subtracted from the number the first column, to produce the third column number (in subtraction problems, the first column value was always bigger than in the second column, to ensure positive values in the third column). The range of input and output values was [.5, 8.0], where [5.0, 8.0] were only present when due to the addition of other features, i.e., [1.0, 4.0] for the training set and [.5, 3.5] for the generalization set.

Training and generalization sets each included 20 examples of each of the 5 types of Raven problems (feature-addition and-subtraction were considered 2 different types), for a total of 100 problems. Each dataset was created by sampling randomly, with possible repetitions of rows and problems, through the possible permutations of the 4 feature values, so that no network had identical training or testing. In test, after training, we calculated mean squared output error for both training and generalization datasets.

Damaged Training and Testing on the Raven task

Two other groups of 100 networks were trained and tested as described above, except that they were damaged by either randomly reducing (white-matter damage) or blocking (grey-matter damage) the output activation of some of their neurons. Damaged neurons were selected randomly for each network, and half of the input neurons and half of the candidate hidden neurons were damaged. There is nothing special about impairing half the neurons, we selected that proportion as a starting point for our experiments. Networks were free to recruit or not recruit impaired hidden neurons, so as to simulate more naturally perinatal brain damage, i.e., prior to learning and performing on tasks. The output neuron was not damaged, in order to insure a fairer comparison of white- and grey-matter damage (a grey-matter-damaged output would prevent any network output).

White-matter damage. White-matter damage is often observed as abnormal white-matter signal and abnormal axonal myelination (Counsell et al., 2006). A reduction in white-matter signal may be due to noisy or leaky axonal transmissions in which abnormal axonal myelination causes action potentials to be lost. To model this leaky transmission we subtracted a different random value from the activation value of impaired neurons each time an activation value was calculated, as in:

$$A_r = Activation - [Activation \times RandomValue(0,1)]$$

where A_r is the reduced random activation, *Activation* is the undamaged activation and *RandomValue(0,1)* is a value chosen randomly from a [0, 1] uniform distribution.

Grey-matter damage. Grey-matter damage can be considered as cell death, leading to a complete loss of signal (e.g., Follett et al., 2009). It was therefore modeled by reducing the activation values of each impaired neuron to 0.

Results

After training, we performed a two-way between networks analysis of variance (ANOVA) in order to compare the effects of dataset (training, generalization) and damage type (undamaged, grey-matter, white-matter) on mean output error. The main effects of dataset and damage type, as well as the dataset by damage type interaction, were all significant. Figure 3 shows mean output error for the different datasets and damage types.

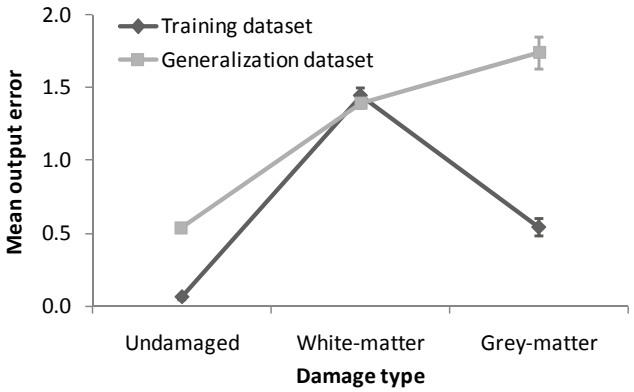


Figure 3. Mean output error and SE bars for the different datasets and damage types. Due to low variation, error bars in the undamaged condition are not clearly visible.

First, error was higher for the generalization, M = 1.22, SD = .83, than for the training set, M = .68, SD = .73, $F(1, 594) = 139$, $p < .001$. It is common for networks to perform better on problems on which they have been trained.

Second, the significant effect of damage type, $F(2,594) = 213$, $p < .001$, was explored using Bonferroni post-hoc tests. Error was significantly lower for the undamaged condition, M = .30, SD = .31, than for either the white-matter, M = 1.42, SD = .42, or grey-matter damage condition, M = 1.14, SD = 1.04, ps < .001. Further, error was significantly lower for grey- than for white-matter damage networks, p = .001.

Third, to explore the significant dataset by damage type interaction, $F(2,594) = 62$, $p < .001$, we analyzed mean network error for each level of the factor dataset (training, generalization), using one-way ANOVAs with damage type (undamaged, grey-matter, white-matter). For the training set, the effect of damage type was significant, $F(2, 297) = 250$, $p < .001$, and Bonferroni post-hoc tests revealed that error was significantly lower for the undamaged condition, M = .06, SD = .12, than for either grey-, M = .54, SD =.58, or white-matter damage, M = 1.44, SD = .49, with error being significantly lower error for the grey- than the white-matter damage, ps < .001. For the generalization set, the effect of damage type was also significant, $F(2, 297) = 87$, p

< .001, and error was still significantly lower for undamaged, M = .54, SD = .24 than for either grey, M = 1.74, SD = 1.06, or white-matter damage, M = 1.39, SD = .34, ps < .001. However, this time error was significantly lower for *white-* than for grey-matter damage, p = .001.

Discussion

We modeled undamaged, white-matter-damage and grey-matter-damage performance on the Raven Matrices task. Of the three conditions, white-matter damage produced highest error. However, the damage type by dataset interaction revealed that compared to white-matter damage, grey-matter damage produced at once higher error for generalization problems not seen in training, and lower error for problems seen in training. To our knowledge, our computational model is the first to demonstrate an association between white- and grey-matter damage and cognitive impairment.

White- worse than grey-matter damage overall

Why was white-matter damage, i.e., noisy reduced axonal signal, overall worse than grey-matter damage, i.e., no axonal signal at all? This perhaps unexpected result may be due to white-matter damage varying in time. That is, white-matter damaged neurons had different noise values every time activation values were calculated, whereas grey-matter damaged activation values were constantly null. White-matter damage networks thus had to deal with changing information, whereas grey-matter damage networks—although missing considerable information—could adapt better to their damage because at least it was constant.

In their computational model of synaptic recovery, Follett and colleagues (2009) also reported a worse effect of white- compared to grey matter-damage, but their model did not test cognitive impairment. Our model adds to their findings by indicating that white- may be worse than grey-matter damage for learning and performing on cognitive tasks. Our results may thus provide insights into the mechanisms underlying the association between damaged and/or reduced white-matter structure and reduced cognitive abilities in preterm children (Skranes et al., 2007), full-term children (Schmithorst, Wilke, Dardzinski, & Holland, 2005) and normal, age-related cognitive decline (Charlton et al., 2006).

Damage type and dataset interaction

Even though error was overall larger for white- than grey-matter damage, grey-matter damage produced larger error on generalization problems, i.e., problems not used in training. Our model thus predicts that different types of perinatal brain damage may be associated with different types of cognitive impairment. It is however difficult to compare our predictions with findings from the preterm literature as not much is currently known about white- versus grey-matter damage in cognitive development (Dammann, Kuban, & Leviton, 2002), and because preterm children with grey-matter damage generally also have white-matter damage, (Pierson et al., 2007). Further, the association between preterm perinatal grey-matter damage and cognitive impairments has not yet been studied directly.

Why different effects?

Interestingly, our further simulations (not reported here) indicate that the differential effects of white and grey-matter damage still hold when the imposed training limit is either doubled or cut in half, when using generalization patterns drawn from the same distribution as training patterns, as well as on the continuous XOR benchmark problem. In continuous XOR there are 2 inputs, each varying between [-.5, .5] and the output is 1 when inputs indicate a point in either the first or third quadrant, and zero in the other two quadrants. The interaction thus seems to be robust to changing the training length and the task.

Insight into our findings may be achieved by analyzing other computational studies. We implemented white-matter damage by randomly reducing the output activations of damaged neurons. Such manipulations resemble injection of noise in neural-network simulations, which was previously found to improve generalization. For instance, Jim, Giles, and Horne (1996) found improved generalization on a dual-parity problem and a randomly generated six-state problem by adding noise to the connection weights of their networks. Unsworth and Coghill (2006) also found improved generalization in their multilayer perception networks, designed to recognize partially obscured human movement, but this time by injecting noise in the training data.

Adding noise can thus improve generalization, perhaps explaining better generalization for white than grey-matter damage. Generalization was however worse for white-matter damage than for *undamaged* networks. This may be due to very high training error in white-matter damage (more than four times higher than for undamaged networks). Indeed, networks' generalization is limited by the quality of their learning. Because white-matter damaged networks had high training error, their overall generalization error was also high. Further, Figure 3 reveals white-matter damage to be the only condition in which error is *not* higher for generalization than training problems (in fact it appears to be slightly *lower* for generalization), which suggests some improved generalization in white-matter damaged networks.

Our implementation of white matter damage differed from the previous noisy simulations. Compared to others who injected noise in either connection weights (e.g., Jim et al., 1996) or in the training data (e.g., Unsworth & Coghill, 2006), we injected noise at the level of neurons' output activations, to simulate impaired axonal transmission. Further, whereas others have used absolute, small noise values, e.g., between [0, 2] (Jim et al., 1996), we used proportional, large noise values that varied between 0% and 100% of neurons' output activations. Thus our noise values varied between [.5, 8.0] due to the range of possible values in the input patterns. Therefore, white-matter damage may have produced large error due to the large noise values.

We implemented grey-matter damage by blocking the output of damaged neurons, simulating cell death and no

axonal transmission. This manipulation resembles neuronal pruning, usually used to increase generalization in neural networks (Reed, 1993). However, pruning algorithms usually select smaller, less important connection weights to be deleted (LeCun, Denker, & Solla, 1990). The idea is that large networks may use their extra connections to encode some of the specifics of the training data. Pruning algorithms thus usually remove smaller weights, in the hope that the remaining, larger connection weights better encode the pattern underlying the data. By contrast to these connection pruning techniques, our networks had whole neurons damaged and these neurons were chosen at random, without regards to whether they were important or not for task performance. Removing potentially critical neurons and connections, as opposed to non-important ones, may explain why grey-matter damage worsened generalization rather than improve it like pruning algorithms.

It is still unclear why training error was lower for grey- than for white-matter damage. This result may reflect the intuition that learning may be easier when missing some information compared to when having wrong information. For instance, Eggert, Ladda, and Straube (2009) found that subjects were better at predicting the trajectory of dots on a screen if *no* aiding cues were provided compared to when both correct and misleading cues were provided. In the case of grey-matter damage, networks apparently learned training problems without the missing input neurons. By contrast, networks with white-matter damage received information from all their input neurons, including some misleading, noisy information which may have made it difficult to learn.

Future directions

We simulated the Raven task by assigning random values to the main features of the matrix figures, and arranging these values in problems following any of the four Raven rules (Carpenter et al., 1990). By contrast, real Raven matrix figures often contain several features which vary along several rules, and thus human subjects have to find which of the features are relevant to which rules. Future simulations may more closely match the task, e.g., by using vectors or sub-matrices to encode all the figures' features. However, because networks still had to figure out the four rules only from the pattern of inputs, we consider our task to still be quite challenging. An indication of this difficulty may lie in the fact that many hidden neurons, i.e., 14 on average, were required by undamaged networks to learn the task. Further analyses may also use the number of problems solved correctly rather than using the usual output error measure. We could thus study whether white- and grey-matter damage also have differential effects on the number of problems solved, and assess the order in which networks succeed at different types of problems as they develop.

We implemented white- and grey-matter damage by impairing half of the neurons in damaged networks (excluding the single output neuron), and damage was static, i.e., a given damaged neuron stayed damaged for the whole simulation. However, because the infant brain is very plastic, perinatal brain damage may interact in a complex way with the child's later development. Future work may consider developmental damage, e.g., punctual damage only at the beginning rather than throughout the simulation, or that is more closely related to the networks' hidden neuron recruitment. For instance, an area often damaged in the preterm brain is the germinal matrix, which is responsible for generating cortical neurons. Because white-matter damage is associated with damage to neurons migrating from the germinal matrix (Leviton & Gressens, 2007), future simulations may more closely simulate perinatal brain damage by directly impairing the hidden neuron recruitment process in SDCC, rather than letting networks decide whether to recruit damaged or undamaged neurons. We may also compare networks with different proportions of both white- and grey-matter damage.

Summary

Our computational model explored the potential link between brain damage and cognitive impairments in preterm children. White-matter damage produced overall higher task error, but grey-matter damage produced higher error on generalization problems, not seen in training. Our results thus predict that different types of brain damage may lead to different types of cognitive impairments. Future psychological work may test this prediction, e.g., by having white- and grey-matter damage populations trained on Raven problems and tested on novel problems (perhaps using a procedure similar to Skuy et al., 2002). Insights gained into the mechanisms underlying the association between perinatal brain damage and cognitive impairment may lead to more effective treatment for survivors of prematurity and help alleviate this aggravating problem.

Acknowledgments

We thank Kristine H. Onishi for insightful discussion and helpful comments. This research was supported by a scholarship to V.G.B. and a grant to T.R.S. from the Natural Sciences and Engineering Research Council of Canada, and by grants to O.D. by the Richard Saltonstall Charitable Foundation, the National Institutes of Health (R21Y019253), and the European Commission (LSHM-CT-2006-036534, HEALTH-F2-2009-241778.

References

Baluja, S., & Fahlman, S. E. (1994). *Reducing network depth in the cascade-correlation* (Technical report No. CMU-CS-94-209). Pittsburgh: School of Computer Science, Carnegie Mellon University.

Bhutta, A. T., Cleves, M. A., Casey, P. H., Cradock, M. M., & Anand, K. J. S. (2002). Cognitive and behavioral outcomes of school-aged children who were born preterm: A meta-analysis. *Journal of the American Medical Association, 288*(6), 728-737.

Billiards, S. S., Pierson, C. R., Haynes, R. L., Folkerth, R. D., & Kinney, H. C. (2006). Is the late preterm infant

more vulnerable to gray matter injury than the term infant? *Clinics in Perinatology, 33*(4), 915-933.

Burd, I., Chai, J., Gonzalez, J., Ofori, E., Monnerie, H., Le Roux, P. D., et al. (2009). Beyond white matter damage: Fetal neuronal injury in a mouse model of preterm birth. *American Journal of Obstetrics and Gynecology, 201*(3), 279.e1-279.e8.

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven progressive matrices test. *Psychological Review, 97*(3), 404-431.

Charlton, R. A., Barrick, T. R., McIntyre, D. J., Shen, Y., O'Sullivan, M., Howe, F. A., et al. (2006). White matter damage on diffusion tensor imaging correlates with age-related cognitive decline. *Neurology, 66*(2), 217-222.

Counsell, S. J., Edwards, A. D., Chew, A. T., Anjari, M., Dyet, L. E., Srinivasan, L., et al. (2008). Specific relations between neurodevelopmental abilities and white matter microstructure in children born preterm. *Brain, 131*, 3201-3208.

Counsell, S. J., Shen, Y., Boardman, J. P., Larkman, D. J., Kapellou, O., Ward, P., et al. (2006). Axial and radial diffusivity in preterm infants who have diffuse white matter changes on magnetic resonance imaging at term-equivalent age. *Pediatrics, 117*(2), 376-386.

Dammann, O., Kuban, K. C. K., & Leviton, A. (2002). Perinatal infection, fetal inflammatory response, white matter damage, and cognitive limitations in children born preterm. *Mental Retardation and Developmental Disabilities Research Reviews, 8*(1), 46-50.

Dandurand, F., Berthiaume, V. G., & Shultz, T. R. (2007). A systematic comparison of flat and standard cascade-correlation using a student-teacher network approximation task. *Connection Science, 19*(3), 223-244.

Dyet, L. E., Kennea, N., Counsell, S. J., Maalouf, E. F., Ajayi-Obe, M., Duggan, P. J., et al. (2006). Natural history of brain lesions in extremely preterm infants studied with serial magnetic resonance imaging from birth and neurodevelopmental assessment. *Pediatrics, 118*(2), 536-548.

Eggert, T., Ladda, J., & Straube, A. (2009). Inferring the future target trajectory from visual context: Is visual background structure used for anticipatory smooth pursuit? *Experimental Brain Research, 196*(2), 205-215.

Follett, P. L., Roth, C., Follett, D., & Dammann, O. (2009). White matter damage impairs adaptive recovery more than cortical damage in an in silico model of activity-dependent plasticity. *Journal of Child Neurology, 24*(9), 1205-1211.

Heron, M., Sutton, P. D., Xu, J., Ventura, S. J., Strobino, D. M., & Guyer, B. (2009). Annual summary of vital statistics: 2007. *Pediatrics, 125*, 4-15.

Jim, K.-C., Giles, C. L., & Horne, B. G. (1996). An analysis of noise in recurrent neural networks: Convergence and Generalization. *IEEE Transactions on Neural Networks, 7*(6), 1424-1438.

LeCun, Y., Denker, J. S., & Solla, S. A. (1990). Optimal brain damage. In D. S. Touretzky (Ed.), *Advances in Neural Information Processing Systems II* (pp. 598-605). San Mateo, CA: Morgan Kaufmann.

Leviton, A., & Gressens, P. (2007). Neuronal damage accompanies perinatal white-matter damage. *Trends in Neurosciences, 30*(9), 473-478.

Leviton, A., & Paneth, N. (1990). White matter damage in preterm newborns—An epidemiologic perspective. *Early Human Development, 24*(1), 1-22.

Peterson, B. S., Vohr, B., Staib, L. H., Cannistraci, C. J., Dolberg, A., Schneider, K. C., et al. (2000). Regional brain volume abnormalities and long-term cognitive outcome in preterm infants. *Journal of the American Medical Association, 284*(15), 1939-1947.

Pierson, C., Folkerth, R., Billiards, S., Trachtenberg, F., Drinkwater, M., Volpe, J., et al. (2007). Gray matter injury associated with periventricular leukomalacia in the premature infant. *Acta Neuropathologica, 114*(6), 619-631.

Reed, R. (1993). Pruning algorithms—A survey. *IEEE Transactions on Neural Networks, 4*(5), 740-747.

Robinson, S. (2005). Systemic prenatal insults disrupt telencephalon development: Implications for potential interventions. *Epilepsy & Behavior, 7*(3), 345-363.

Schmithorst, V. J., Wilke, M., Dardzinski, B. J., & Holland, S. K. (2005). Cognitive functions correlate with white matter architecture in a normal pediatric population: A diffusion tensor MRI study. *Human Brain Mapping, 26*(2), 139-147.

Shultz, T. R. (2006). Constructive learning in the modeling of psychological development. In Y. Munakata & M. H. Johnson (Eds.), *Processes of change in brain and cognitive development: Attention and performance XXI* (pp. 61-86). Oxford: Oxford University Press.

Shultz, T. R., Mysore, S. P., & Quartz, S. R. (2007). Why let networks grow? In D. Mareschal, S. Sirois, G. Westermann & M. H. Johnson (Eds.), *Neuroconstructivism: Perspectives and prospects* (Vol. 2, pp. 65-98). Oxford: Oxford University Press.

Skranes, J., Vangberg, T. R., Kulseng, S., Indredavik, M. S., Evensen, K. A. I., Martinussen, M., et al. (2007). Clinical findings and white matter abnormalities seen on diffusion tensor imaging in adolescents with very low birth weight. *Brain, 130*(3), 654-666.

Skuy, M., Gewer, A., Osrin, Y., Khunou, D., Fridjhon, P., & Rushton, J. P. (2002). Effects of mediated learning experience on Raven's matrices scores of African and non-African university students in South Africa. *Intelligence, 30*(3), 221-232.

Unsworth, C. P., & Coghill, G. (2006). Excessive noise injection training of neural networks for markerless tracking in obscured and segmented environments. *Neural Computation, 18*(9), 2122-2145.

Neural networks for word recognition: Is a hidden layer necessary?

Frédéric Dandurand (Frederic.Dandurand@univ-provence.fr)
Laboratoire de Psychologie Cognitive, CNRS, Aix-Marseille University
3, place Victor Hugo, 13331 Marseille, France

Thomas Hannagan (thom.hannagan@gmail.com)
Laboratoire de Sciences Cognitives et Psycholinguistique, EHESS/CNRS/DEC-ENS, École Normale Supérieure
29 rue d'Ulm, 75005 Paris

Jonathan Grainger (jonathan.grainger@univ-provence.fr)
Laboratoire de Psychologie Cognitive, CNRS, Aix-Marseille University
3, place Victor Hugo, 13331 Marseille, France

Abstract

We study neural network models that learn location invariant orthographic representations for printed words. We compare two model architectures: with and without a hidden layer. We find that both architectures succeed in learning the training data and in capturing benchmark phenomena of skilled reading – transposed-letter and relative-position priming. Networks without a hidden layer use a strategy for identifying target words based on the presence of letters in the target word, but where letter contributions are modulated using the interaction between within-word position and within-slot location. This modulation allows networks to factor in some information about letter position, which is sufficient to segregate most anagrams. The hidden layer appears critical for success in a lexical decision task, i.e., sorting words from non-words. Networks with a hidden layer better succeed at correctly rejecting non-words than networks without a hidden layer. The latter tend to over-generalize and confuse non-words for words that share letters.

Keywords: Computational modeling, word recognition, neural networks, reading, priming effects.

Introduction

An important cognitive activity involved in skilled reading is the mapping of retinal images of letters onto abstract word representations. Skilled readers can identify words relatively easily (although not perfectly, see e.g., Rayner, White, Johnson, Liversedge, 2006) even when letter order is jumbled, except for the first and last letters. This suggests that at least one intermediate level of coding exists that abstracts away from absolute letter position and instead codes some information about relative letter order. Such an intermediate level of representation has been studied using a number of techniques including masked priming (see Grainger, 2008 for a review). Robust priming effects found include the transposed-letter effect and the relative-position effect. The transposed-letter effect describes the superior priming observed from primes formed by transposing two of the target's letters (e.g., gadren-garden) compared with primes formed by substituting two of the target's letters (e.g., galsen-garden). The relative-position priming effect describes a processing advantage for targets preceded by primes formed of a subset of the target's letters (e.g., grdn-garden) compared with a prime formed of the same subset of letters in the wrong order (e.g., gdrn-garden).

A number of models have been proposed for an intermediate level of coding that can account for these priming effects (see Grainger, 2008 for a review). Notably, the Grainger and Van Heuven (2003) model of orthographic processing was the inspiration for a computational model that learned to map location-specific letter identities (letters coded as a function of their position in a horizontal array) onto location-invariant lexical representations (Dandurand, Grainger, & Dufau, 2010). Because parsimony dictates to assume a single intermediate level of representation, we considered a neural network architecture with a single hidden layer.

This network architecture with a hidden layer successfully captured transposed-letter and relative-position priming effects (Dandurand et al., 2010). Intermediate representations were explicitly probed and analyzed as patterns of activation at the hidden layer (Hannagan, Dandurand, & Grainger, submitted; see also Plaut, McClelland, Seidenberg, & Patterson 1996 for a discussion of internal representations in neural networks). These patterns were found to have two important characteristics. First, letters seemed to be represented in a semi-location-invariant fashion at the hidden layer. Second, representations at the hidden layer were well-characterized as a holographic overlap coding in which small changes of the inputs resulted in small differences in hidden layer representations. More specifically, differences in patterns of hidden layer activations were monotonically related to differences in identity and position of input letters. For example, patterns of activity at the hidden layer were more different for a two-letter substitution at the input (POLL vs. BULL) than a single letter substitution (PULL vs. BULL) when position in the horizontal array was kept constant. Furthermore, differences in patterns of activity were also larger when the input word was moved by two positions in the alphabetic array (#THAT##### vs. ###THAT###) than moved by a single position (#THAT##### vs. ##THAT####). Holographic overlap coding explains the observed transposed-letter and relative-position priming and

makes a number of predictions which are tested in this article; see (Hannagan et al., submitted) for details.

As they map letters onto words, skilled readers can also perform lexical decision, that is, deciding if a string of letters is a word or a non-word (Meyer & Schvaneveldt, 1971). Lexical decision has been extensively studied, and a number of models exist to account for human performance (e.g., Ratcliff, McKoon, & Gomez, 2004). In the current work, we test our models on a simple lexical decision task, assuming a minimal lexical read-out mechanism, namely that words would activate output units more than non-words. We are not, however, claiming that this ability should be interpreted as a full-blown or realistic model of lexical decision. Note that performing lexical decision is not trivial for networks because non-words are never seen in training as negative evidence, and thus networks may be expected to over-generalize what they consider as words.

In the current study, we revisit the assumption previously made for the need of a hidden layer. We ask if such a hidden layer is required for networks to learn location invariant orthographic representations for printed words. To this effect, we contrast two model architectures: (1) the previous model with a hidden layer and (2) a simpler model without a hidden layer. In this alternative model, letters are mapped to words directly using a layer of connection weights. We compare the two architectures on a number of criteria: (1) their ability to learn the training set, including the anagrams present in the training data, (2) their size and complexity, (3) their capacity to simulate key priming effects, and (4) their capacity to perform a simple lexical decision task. Finally, we investigate how processing and representations differ, how networks without a hidden layer manage to segregate anagrams, and how well these networks conform with the predictions made by holographic overlap coding.

Our goal is to gain insights into the role that the hidden layer plays in performing a word recognition task. Without a hidden layer, networks are computationally limited to taking decisions based on weighted combinations of input letters. It is unclear how, and even if, such model could handle anagrams where the identity of input letters is insufficient to discriminate words, and where position of letters has to be taken into account.

Methods

We compare two architectures of standard multilayer perceptron neural networks. The first one includes a single hidden layer of 91 hidden units with logistic activation functions, identical to (Dandurand et al., 2010). The second one has no hidden layer (inputs are directly connected to outputs). In the two architectures, adjacent layers are fully connected, and are trained using standard backpropagation (learning rate = 0.1, momentum = 0.9) until an SSE of 30. Training material consists of 1179 real words of four letters (same as the one used by McClelland, & Rumelhart, 1988) presented in all 7 possible positions of an alphabetic array (e.g., #ABLE#####, ######ABLE where # are empty, blank slots). Local (sparse) coding is used for input letters (one out of 26 possible letters, for each slot) and output units (one out of 1179 words, also with logistic activation functions). Networks learn to associate letter strings presented at the input with the corresponding output unit coding for some word. For further details, see (Dandurand et al., 2010).

We trained and tested samples of 10 networks for each condition (with and without a hidden layer). Networks varied in the random initial values of their connection weights.

In tests that involve lexical decision, we present some pattern at the input and compute activations of all output units. Output units activated above a threshold value of 0.9 are considered as active, and thus the word associated with the unit as having been detected. For tests that involve priming, a measure dubbed "target supremum measure" (Dandurand et al., 2010) quantifies the ability of some prime to activate the output unit associated with the target word more than any other active output unit[1].

Results

Learning the training set

The training set comprises 1179 words, 24.0% (N = 283) of which are anagrams. Anagrams come in pairs (111 pairs x 2 = 222 words), triplets (15 triplets x 3 = 45 words) and quadruplets (4 quadruplets x 4 = 16 words). These quadruplets (1. live – evil – veil – vile; 2. team – meat – mate – tame; 3. tied – diet – tide – edit; 4. pear – rape – reap – pare) should be especially difficult to discriminate because the same four letters activate four different target word units.

Networks with a hidden layer achieve perfect performance (100%) on the target supremacy measure for the training set. In contrast, networks without a hidden layer reach 98.6%, and more than 95% of anagrams were successfully segregated. In the 1.4% of errors, activations of output units (including the target) fail to reach the threshold of 0.9. These failure-to-recognize errors involved pairs of anagrams (bear – bare, and read – dear) or sets of words from an orthographic neighborhood sharing three letters (bare – mare – pare, seep – seed – deep, and pull – burl – bull).

Model size and complexity

From a size and complexity perspective, the hidden layer adds 91 extra units, and an additional layer of processing. However, in terms of size, networks with a hidden layer actually have fewer connection weights (132 219, i.e., 1179

[1] Models allow for multiple outputs to be activated, but some competitive, winner-takes-all mechanism could be used to select the most active one. Item-level target supremum value was set to 1 when the prime activated the output unit associated with the target lexical item more than any other unit; it was set to 0 otherwise. The target supremum measure of a set of primes was computed as the mean of item-level values for the primes in the set.

outputs x (91 hidden + 1 bias) + 91 hidden x (260 inputs + 1 bias)) than networks without a hidden layer (307 719 connection weights (1179 outputs x (260 inputs + 1 bias)), despite having two layers of weights. We can think of the hidden layer as enforcing data compression from 260 inputs to 91 hidden units, which reduces the number of connections required.

Priming effects

Networks are tested using the relative-position priming and transposed-letter priming manipulations described in (Dandurand et al., 2010). Examples of primes for word ABLE are overlapped on the graphs below, see (Dandurand et al., 2010) for details of the content of testing sets. Primes (e.g., ###ABE####) are expected to activate the target word (e.g., ABLE) more so than any other word, especially when prime letters are in the correct, forward order (ABE) and not the reserved, backward (EBA) order.

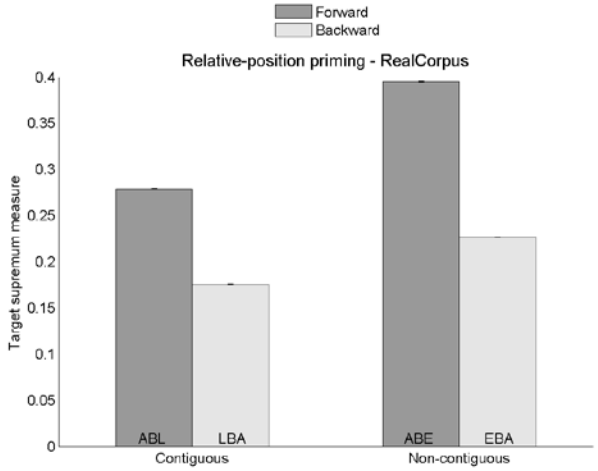


Figure 1 – Target supremum results for the relative-position priming test. Example primes provided for target word ABLE.

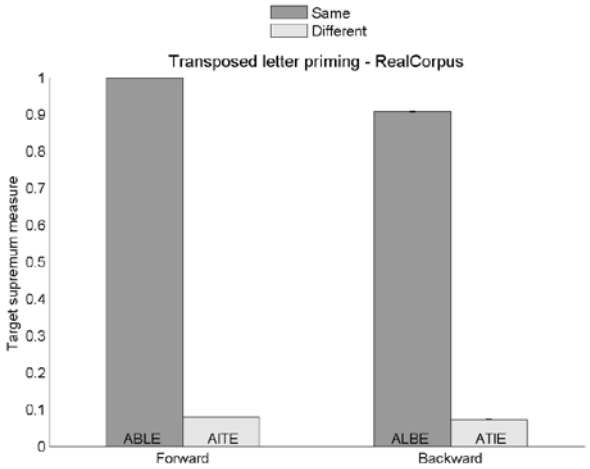


Figure 2 – Target supremum results for the transposed-letter priming test. Example primes provided for target word ABLE.

As we can see, patterns of results are very similar for networks with (see Figures 5 and 6 in Dandurand et al., 2010) and without a hidden layer. More specifically, relative-position primes formed of forward letter subsets yield a higher target supremum measure than backward primes (see Figure 1); and transposed-letter primes containing central letters from the target word yield a larger supremum measure than primes with central letters from a different word (see Figure 2).

Lexical Decision

To test for lexical decision, we assess performance (target supremum measure) on three simple testing conditions: (1) words: all words seen in training in all positions (for a total of 1179x7 patterns); (2) non-words: a sample of 100 patterns made of four random letters presented at a random position in the alphabetic array (e.g. #JKTS#####, ######HIQL, ###BXGA###); (3) letters: a sample of 100 patterns, each made of a randomly selected letter repeated to match word length presented at a random position in the alphabetic array (e.g., ##AAAA####, #####HHHH#). Word patterns are expected to activate, and only activate, their target word unit. We also expect no output word unit to be activated above threshold for patterns in the non-words and in the letters conditions.

Results are shown in Figure 3. As we can see, network with a hidden layer perform much better than networks without one. Networks without a hidden layer are especially poor at correctly rejecting letter patterns, activating several of the words that contain the letter. For example, input pattern ###PPPP### activates 85 word units above threshold including part, open, help, kept, step, post and ship. Similarly, for non-words, errors involve incorrectly activating words that share some letters with the target. For example, input pattern ####KNKR## activates the following word units above threshold: kind, dark, park, mark, link, monk, fork, tank, pork, cork, knot, and trek.

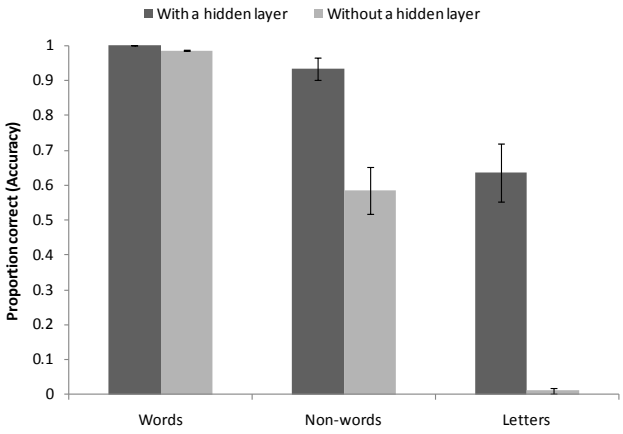


Figure 3 - Accuracy of networks at accepting words, and rejecting non-words and repeated letters.

Discussion

To sum up our results, both networks with and without a hidden layer correctly recognized words at rates reaching 98.6% to 100%. Performance was high even on anagrams (95% to 100%). Both types of networks showed relative-position (see Figure 1) and transposed-letter priming effects (see Figure 2). Networks with a hidden layer are more complex due to the additional hidden units, but contain fewer connection weights. The critical benefit of the hidden layer appears to be in the ability of networks to correctly reject non-words and strings of repeated letters (see Figure 3).

Segregating anagrams

One of the most difficult aspects of the task is arguably that of segregating anagrams. While regular words can be discriminated on the basis of differences of at least one letter, anagram identification must rely solely on the relative position of letters within word. The task appears especially difficult for the network without a hidden layer which is limited to computing linear combinations of independent inputs.

Networks with a hidden layer

In networks with a hidden layer, holographic overlap coding (Hannagan et al., submitted) can explain both transposed-letter priming and the ability of networks to segregate anagrams. During learning, networks form semi-location specific representations for individual letters - assigning similar representations to the same letter input seen at different positions - that is, networks combine letters in a continuous manner to build a string code. Displacing letters (whether in primes or in anagrams) results in small, but measurable differences in patterns of activation at the hidden layer. In the case of transposed-letter priming, most words have no orthographic neighbor, and therefore the target word is still the most activated (e.g., WTIH activates word WITH), and so will be recognized according to the target supremum measure. Networks can capitalize on this small difference in hidden pattern activation to segregate words. It is plausible that this small difference gets enhanced or amplified by the processing of the second layer of weights (hidden to output weights) to generate the correct classification of anagram patterns (e.g., ABLE and BALE as distinct).

Networks without a hidden layer

To gain insights into how networks without a hidden layer can segregate anagrams, we study the connection weights between inputs and outputs after training. The first thing we notice is that connection weights strongly code for the mere presence of letters. Typically, connection weights are small for letters not present in the target word, and large for letters that are present, irrespective of position. For instance, connections weights from input units that code letters A, B, L, and E (in all slots where they have been seen during training) are large to output unit coding for word ABLE. This simple scheme makes each letter vote for the target word, and a word must get 4 votes to be fully activated. This may explain why letters activate very strongly a number of targets, as AAAA also counts as 4 letters of evidence for ABLE. However this does not explain how the network can distinguish between anagrams.

Figure 4 illustrates how networks might manage to segregate anagram patterns. Boxes in the plot show the average magnitude of connection weights between within-word position on the Y axis and within-alphabetic-array (within-slot) location on the X axis for letters relevant to the identification of the target word. For example, for pattern ABLE###### connection weights would be found at boxes (X,Y): A(1,1), B(2,2), L(3,3) and E(4,4); whereas for pattern ###ABLE### relevant boxes would be A(4,1), B(5,2), L(6,3) and E(7,4).

As we can see, there is a negative correlation ($r = -0.73$, $p < 0.01$) in the first within-word position (P) between the average magnitude of connection weights (C) and location (L), while the correlation is positive in the last position ($r = 0.67$, $p < 0.01$). Namely, for the first letter of the word, the connection weight is largest for smaller locations in the slot, and decrease as location in slot increases. This makes intuitive sense, as A######### is better evidence for word ABLE (or any word that begins with letter A) than ######A###, which could be evidence for ######ABLE, but also for ####THAT## or any word having an A in any position. The correlation is reversed for the last slot where say letter E provides more evidence for ABLE if it appears later in the word. The direction reversal suggests an interaction between location (L) and within-word position (P).

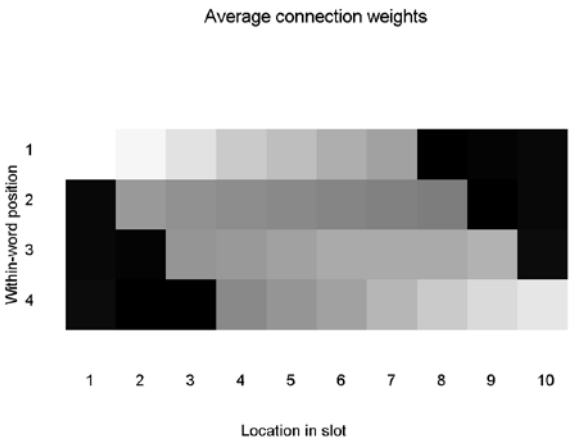


Figure 4 - Average magnitude of weights connecting input units relevant to identifying an output word, by location in the alphabetic array (X axis) and by position with target word (Y axis). Black boxes correspond to positions where letters were never seen in training (e.g., letter A was never seen in slots 8 to 10 for word ABLE, and similarly letter E was never seen in slots 1 to 3).

To test for this interaction, we performed a linear regression with the following model (including LxP to test for interaction effects):

$C = b_0 + b_1L + b_2P + b_3LxP$ (1)

In the fitted model, we get $b_0 = 31.0$ ($p < 0.001$), $b_1 = -4.0$ ($p < 0.001$), $b_2 = -8.9$ ($p < 0.001$) and $b_3 = 62.9$ ($p < 0.001$). This confirms the significant interaction. Redoing the analysis with central locations only (4 to 7), we also get significant coefficients, $b_0 = 22.6$ ($p < 0.001$), $b_1 = -1.8$ ($p < 0.001$), $b_2 = -4.6$ ($p < 0.001$) and $b_3 = 30.4$ ($p < 0.001$).

To sum up, the processing strategy or coding scheme that networks without a hidden layer develop can be described as follows: most important is the number of letters shared between inputs and targets independently of position – we can think of this as input letters providing independent votes for the target words that contain them. The presence of letters is then modulated by the interaction between location and position. This scheme is sufficient to explain how networks can discriminate between anagrams. For instance in strings ABLE and BALE, an equal number of four letter votes go to each word, and connection weights between small slot positions and target word ABLE are slightly larger for letter A than letter B. In contrast, for target word BALE, the connection weight is slightly larger for letter B than letter A. This difference enables the correct target to be activated.

This coding scheme also accounts for the priming effects: larger priming as the number of letters shared between primes and targets increase, and larger priming as the agreement increases between the order of letters in the prime and in the target.

Comparison with holographic overlap coding

How does this processing strategy in networks without a hidden layer compare to holographic overlap coding used by networks with a hidden layer? As mentioned in the introduction, holographic overlap coding makes two important predictions about similarity of activation patterns: a proximity effect and a disruption of activation when replacing letters with other letters of the word (e.g., AAAA for word ABLE). The normalized Euclidian distance between two activation vectors $Act(V_1)$ and $Act(V_2)$ is computed as follows:

$dist = \sqrt{(\Sigma \Sigma (Act(V_{1ij}) - Act(V_{2ij}))^2)} / (N_{pattern} \times N_{activation})$

Activations are taken at the hidden layer, or at the output layer for networks without a hidden layer. The two Σ indicate summing over all patterns and all activation values.

The proximity effect predicts that the Euclidian distance between activation vectors V_1 and V_2 should increase monotonically with the magnitude of displacement of the vectors (i.e., distances). As shown in Table 1, a proximity effect is observed indeed, when vectors V_1 are in the central position (###XXXX###) and vectors V_2 vary in position. Distances presented in the table are normalized using a displacement of 1 as a reference (that is, V_2 ##XXXX#### and ####XXXX##). Vectors V_2 for displacement 2 are #XXXX##### and #####XXXX#; and for displacement 3: XXXX#### and ######XXXX. As we can see, distances increase with displacement, in accordance with the proximity effect.

Table 1: Normalized Euclidian distance for networks with and without a hidden layer, as a function of displacement of letters in the input vector

	Euclidian distance	
Displacement	With hidden	Without hidden
2	1.3	1.5
3	2.2	1.7

Holographic overlap coding also makes a prediction about the effect of letter substitutions: the more letters are replaced, the larger the difference in activation should get. We empirically test this hypothesis by generating samples of 100 test items for which the target word and the location of letters in the input slot is randomly chosen. We compute the Euclidian distance between patterns of activation generated in one of three conditions: (1) transposition – transpose two letters, randomly chosen (e.g., V_1 = ABLE → V_2 = ABEL), (2) one letter substitution with a random letter (e.g., V_1 = ABLE → V_2 = ABWE), (3) one letter substitution with another letter of the target – that is, a letter repetition (e.g., V_1 = ABLE → V_2 = BBLE).

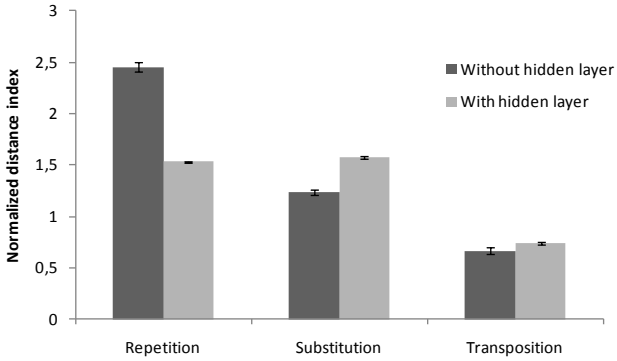


Figure 5: Normalized Euclidian distance index as a function of transformation and architecture type

Holographic overlap coding predicts similar distances for letter repetitions and substitutions, and a lower distance for transpositions. As we see in Figure 5, this is precisely the pattern of distances measured for networks with a hidden layer. However, these predictions are not verified for networks without a hidden layer, namely because distances are too large for the letter repetition set. This somewhat counter-intuitive result can be explained by the fact that repeating a letter means, on average, replacing a letter with a rather frequent letter compared to substituting with a randomly chosen one (as in the substitution case). And thus, many output words activate in the repetition case, which increases the distance due to the higher activation of the non-target words. In sum, we fail to find evidence that networks without a hidden layer implement a holographic overlap coding scheme.

Lexical decision, over-generalization and their theoretical implications

In the lexical decision task, correct rejection of non-words and letters can be interpreted as a test of generalization, which probes the network's ability to correctly set the boundary of word acceptance. Based on a poverty of stimulus argument, we may expect networks to over-generalize, that is being overly liberal in accepting strings as words, because networks see positive evidence for words but never see any negative evidence, i.e., they are never trained to reject non-words. These over-generalization errors are much more common in the network without a hidden layer. This has interesting theoretical implications for the functional role of the hidden layer where independent letters are combined. Given that each letter/position has a uniquely defined code, the network just has to find a way to integrate them so as to ensure that each combination is unique. For instance, using a simple averaging approach, the resulting code for AAAA will be very close to A, in effect providing only evidence for one letter. Without combinations, networks have to base their decisions on some position-weighted voting scheme relating to the presence of letters. This scheme fails to reject non-words cases that consist of 4 repetitions of a letter from the target word.

Beyond simply removing letter duplicates, the hidden layer may well be coding for some letter combination, or sub-lexical units, as postulated in the Grainger and Van Heuven's (2003) model and other models. A simple approach to lexical decision could thus be seen as follows: letters provide evidence for activating sub-lexical units. These sub-lexical units would in turn be combined to activate target words. For non-words, activation of sub-lexical units would be small, and result in activation of output units that fall below threshold.

Conclusion

To summarize, the hidden layer developed a holographic overlap coding scheme which explains priming effects and segregation of anagrams. Because it is sensitive to letter substitutions, this scheme also allows networks with a hidden layer to correctly reject most non-words.

In contrast, networks without a hidden layer have developed a strategy for identifying target words largely based on presence of letters but where letter contributions are modulated using the interaction between within-word position and within-slot location. This modulation allows networks to factor in some information about letter position, which is sufficient to segregate most anagrams, and replicate the previously observed priming effects. On the other hand, these networks are poor at the lexical decision task, as they tend to over-generalize and confuse non-word strings as words. As long as the number of letters is the same and that all input letters exist in the target word, networks do not require that all letters in the target word are present to activate it.

The hidden layer also implements some data compression, by forcing 260 input units to be represented onto 91 hidden units. As a result, networks with a hidden layer have fewer than half the number of connection weights of networks without a hidden layer.

Computational models of word identification are expected to perform well at lexical decision, as humans do. The model with the hidden layer suggests a parsimonious account of lexical decision as an emergent property of the word recognition task (although, again, the setup is highly simplified, and further work would be necessary to fully assess how good of a lexical decision model this is). An alternative explanation consists in using an additional module (performed before, or in parallel with, word identification). For the latter, a network without a hidden layer is sufficient to simply recognize words.

Acknowledgments

This project was supported by the Agence Nationale de la Recherche (grant no. ANR-06-BLAN-0337) and the Europrean Research Council (ERC-230313).

References

Dandurand, F., Grainger, J., & Dufau, S. (2010). Learning location invariant orthographic representations for printed words. Connection Science, 22(1), 25-42. doi:10.1080/09540090903085768

Grainger, J. (2008). Cracking the orthographic code: An introduction. *Language and Cognitive Processes*, *23*(1), 1-35.

Grainger, J., & van Heuven, W. J. B. (2003). Modeling letter position coding in printed word perception. In *The Mental lexicon* (pp. 1-23). New York: Nova Science Publishers.

Hannagan, T., Dandurand, F., & Grainger, J. (submitted). Broken symmetries in a location invariant word recognition network, *Neural Computation*.

McClelland, J. L., & Rumelhart, D. E. (1988). *Explorations in parallel distributed processing.* Boston, MA: MIT Press.

Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, *90*(2), 227-234.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding Normal and Impaired Word Reading: Computational Principles in Quasi-Regular Domains. *Psychological Review, 103*(1), 56-115.

Ratcliff, R., McKoon, G., & Gomez, P. (2004). A Diffusion Model Account of the Lexical Decision Task. *Psychological Review*, *111*(1), 159-182.

Rayner, K., White, S., Johnson, R., Liversedge, S. (2006). Raeding Wrods With Jubmled Lettres; There Is a Cost. *Psychological Science, 17*(3), 192-193

The dimensionality of episodic images

Vishnu Sreekumar (sreekumar.1@buckeyemail.osu.edu)
Department of Psychology, Ohio State University
Columbus, OH 43201 USA

Yuwen Zhuang (zhuang.14@buckeyemail.osu.edu)
Department of Computer Science and Engineering, Ohio State University
Columbus, OH 43201 USA

Simon J. Dennis (simon.dennis@gmail.com)
Department of Psychology, Ohio State University
Columbus, OH 43201 USA

Mikhail Belkin (mbelkin@cse.ohio-state.edu)
Department of Computer Science and Engineering, Ohio State University
Columbus, OH 43201 USA

Abstract

Previous studies (Doxas, Dennis, & Oliver, 2010) show that natural language discourse exhibits a two-scale structure with a lower dimension at short distances and larger dimension at long distances. We attempt to search for the source of this constraint in the visual input that goes into forming episodic experiences in human beings. This information is assumed to be approximated well by images captured by a Microsoft^{TM}Research SenseCam that our subjects used. The hypothesis is that if the same two scale structure is observed here, the constraint is possibly not one that is imposed by the cognitive system. We use and contrast two methods by which images can be represented: the traditional color histogram and a more recently developed color correlogram method. The color correlogram is established to work better for our current purposes. We observe hints of a two scale structure in the correlation dimension plots but these are not conclusive.

Keywords: Episodic Memory; Correlation Dimension; Networks; Graphs.

Introduction

The existing models of episodic memory assume a representation of context. Retrieval of episodes involves reinstatement of context. The current literature does not address the nature of representation of context and the question of how the representation was formed in the first place. Our ultimate goal is to model contextual reinstatement as a search over episodic networks. We begin by looking at the images that people encounter everyday. In a parallel study, graphs of these images are constructed and the structure of the graphs is investigated. People are extremely fast at isolating episodes from memory. Such a search has to be fast and efficient. The graph has to satisfy certain properties for it to be efficiently searchable (Steyvers & Tenenbaum, 2005). We attempt to test the idea that contextual reinstatement can be modeled as a network search. One prerequisite for this model to be feasible is that the episodic network must be quickly searchable.

We encode events into our memory as we encounter and experience them. What kinds of constraints are inherent to this input information? Such a question is motivated by previous studies on natural language discourse where paragraph spaces of corpora of different languages exhibited a two-scale structure (Doxas et al., 2010). Doxas et al. did a correlation dimension analysis on the paragraph spaces of text corpora taken from five different languages and genres. The correlation dimension is a measure of how points within a given distance r scales with that distance. The paragraph spaces were found to exhibit a low dimensional structure at short distances and a higher dimensional structure at larger distances. This is similar to a "weave" structure. For example, if we zoom in to look at a thread that is part of a shirt, the observed dimensionality is one. If we zoom out to intermediate length scales, we would start observing a two dimensional structure. Further zooming out will further increase the dimensionality. The finding of this "weave" structure in natural language discourse raises an important question regarding the origin of this constraint. Is this constraint one that is imposed by the cognitive system or is it a property of the input the system receives that is being mirrored by the cognitive system? We attempt to address this question in the current study. To investigate this, we used a Microsoft^{TM}Research SenseCam to capture images that can be thought of as representative of a person's (visual) episodic experience. A dimensionality analysis was then done on these images.

The paper is organized as follows. The next section outlines the method used to capture and represent the images on which the dimensionality analysis is done. The Microsoft Research SenseCam device is described briefly. Two different image representation schemes and their corresponding distance measures are discussed. The two methods are then contrasted using a definition of a ratio that is based on the requirement that these methods must, among other things, successfully identify images that belong to the same contexts. The subsequent section describes the correlation dimension. The results section discusses the correlation dimension plots for the image sets obtained from different individuals. The paper concludes with a discussion of the structure that is observed in the correlation dimension plots of the image data.

Image Data Collection, Representation and Distance Measures

Microsoft Research SenseCam

To capture a sufficient number of images that can sufficiently represent an individual's visual episodic experience for a period of about a week, we used a Microsoft Research SenseCam. Subjects hung the camera around their necks for about a week each. The SenseCam contains sensors which can detect changes in color, light-intensity and temperature. Changes in these sensor readings can be set to automatically trigger the SenseCam to take pictures. The camera can also be set to a timer mode where pictures can be captured periodically. Our camera captures an image once every eight to ten seconds. The camera has wide-angle (fish-eye) lens that maximizes its field-of-view. The resulting images are particularly useful for studying episodic experience because these images are fragmentary, time compressed, temporally ordered, and have a 'field perspective' (Berry et al., 2006).

HSV Space

The HSV (hue, saturation, value) color space is very different from the better known RGB (red, green, blue) color space.The problem with using the RGB color space is that it is not perceptually uniform. To get a satisfactory representation of the image in the RGB space, the quantization step sizes should be fine such that distinct colors are not assigned to the same bin. This increase in the number of bins affects performance in terms of computation time. The oversampling also produces a larger set of colors than are necessary and this is not an accurate representation of human visual discrimination of colors.

A three dimensional representation of the HSV color space is a hexacone (Stockman & Shapiro, 2001). The central axis represents the intensity. Hue is defined as an angle in the range [0,2π] relative to the red axis such that red is at angle 0, green is at $2\pi/3$, blue at $4\pi/3$ and red again at 2π. Saturation takes values between 0 and 1. Saturation is the depth or purity of the color. It is measured as a radial distance from the central axis. The saturation value is 0 at the central axis and is 1 at the outer surface. As saturation varies from 0 to 1, the corresponding hues vary from unsaturated (shades of gray) to fully saturated (no white component, pure form of the color represented by its hue). In other words, for a low value of saturation, a color can be approximated by a gray value specified by the intensity value and for a high value of saturation, the color can be approximated by its hue. HSV separates out the light-intensity information (luminance) from the color information (chromaticity).

Color Histogram Representation

A color histogram for an image is generated by concatenating 'N' higher order bits for the Red, Green and Blue values in the RGB space (Swain & Ballard, 1991). The histogram is generated by counting the number of pixels with the same color and accumulating it in 2^{3N} bins . We generate such a histogram from the representation of each image in the HSV space. Quantizing the hue component more precisely than the value and saturation components makes the HSV histogram more sensitive to color differences and less sensitive to brightness and depth differences. We found it sufficient to use a (h=30 levels, s=10 levels, v=3 levels) quantization to generate the histograms based on the fact that the human eye is more sensitive to variations in hue and intensity than variations in saturation.

Several distance measures can be used to calculate distance between images (Jeong, Won, & Gray, 2004). These include the histogram euclidean (HE) distance and the histogram intersection (HI) distance (Smith & Chang, 1995, 1996). A Kullback-Liebler divergence (Greenspan, Goldberger, & Ridel, 2001) measure is also discussed which has been established to work better than the HE and HI measures in information retrieval tasks (Goldberger, Gordon, & Greenspan, 2006).

Histogram Euclidean Distance If **h** and **g** represent two color histograms, the euclidean distance between them is given by

$$d^2(h,g) = \sum_A \sum_B \sum_C (h(a,b,c) - g(a,b,c))^2 \qquad (1)$$

A,B and C are the three colors (RGB or HSV). In this formula, all bins contribute equally to the distance and only identical bins in the respective histograms are compared.

Histogram Intersection Distance The histogram intersection (HI) distance (Swain & Ballard, 1991) between **h** and **g** is given by

$$d(h,g) = \frac{\sum_A \sum_B \sum_C \min(h(a,b,c), g(a,b,c))}{\min(|h|, |g|)} \qquad (2)$$

$|\mathbf{h}|$ and $|\mathbf{g}|$ are the number of samples in the respective histograms. The sum is normalized by the histogram with the lesser number of samples. We used the histogram intersection distance for our initial analysis. The distance tends to 1 if the images are highly similar and 0 if they are highly dissimilar.

Square root of the Jensen-Shannon divergence: a proper metric A better measure to calculate similarity between images is the information theoretic Kullback-Liebler (KL) divergence (Greenspan et al., 2001). This is a non-symmetric measure of the difference between two probability distributions. It has been shown to perform better than HI in image search and retrieval tasks (Goldberger et al., 2006). Though the intuition is to use the KL divergence directly as a distance measure, it is not a true metric. A symmetrical version of the KL divergence is the Jensen-Shannon (JS) divergence, the square root of which is a metric. Using a proper metric is important since we intend to study the dimensionality of the space of these image representations. Our color histogram results here are based on the distance measure that is the square

root of the JS divergence. Figure 1 shows a query image and the retrieved images that are similar to the query image based on the JS distance. The distance is printed on top of each of the retrieved images.

The Kullback-Liebler divergence of Q from P is defined as

$$D_{KL}(P||Q) = \sum_i log \frac{P(i)}{Q(i)} \tag{3}$$

where P and Q are probability distributions of a discrete random variable. The symmetric Jensen-Shannon divergence is given by

$$D_{JS} = \frac{1}{2} D_{KL}(P||M) + \frac{1}{2} D_{KL}(Q||M) \tag{4}$$

where $M = \frac{1}{2}(P+Q)$

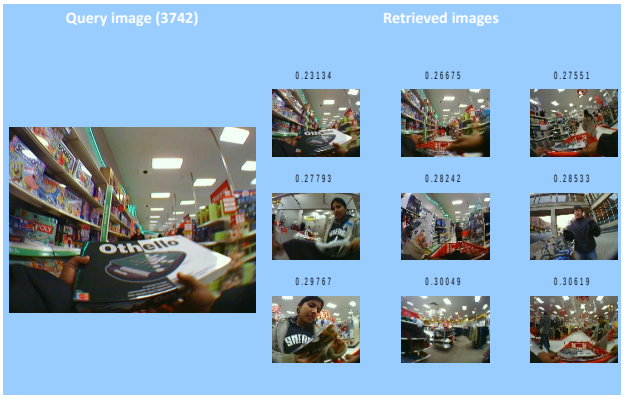


Figure 1: Query image and retrieved images (JS divergence distance method).

Color Correlogram Representation

The color histogram has the drawback of being a purely global description of the color content in an image. It does not include any spatial information. Purely local properties when used can be extremely sensitive to appearance changes due to slight changes in angle, zoom, etc. Purely global properties (like those used in the color histograms) can give false positives as it can classify images from widely separated scenes as belonging to the same scene if they have similar color content. An example of this can be found in figure 1. The third image in the second row of the retrieved images is a false positive because that image belongs to an entirely different event.

A color correlogram (Huang, Kumar, Mitra, Zhu, & Zabih, 1997) describes global distributions of local spatial color correlations. In other words, a correlogram of an image is a three dimensional matrix whose k-th entry for $\langle i, j \rangle$ is the probability of finding a pixel of color j at a distance k from a pixel of color i. This makes the correlogram robust to changes in appearance caused by occlusions, zoom, viewing angles, etc. We use a special case of the correlogram for ease of computation: the banded correlogram (Huang, 1998). Figure 2 shows the same query image as earlier and the retrieved images that are based on the relative L_1 distances between images represented as banded correlograms. The distance is printed on top of each of the retrieved images. There are no false positives in these retrieved images.

Let I be an $n \times m$ image. The colors in I are quantized into k colors $c_1, c_2, \cdots c_k$. For a pixel $p = (x, y) \in I$, let $I(p)$ denote its color. $I_c \triangleq \{p | I(p) = c\}$ where $c \in \{c_1, c_2, \cdots c_k\}$. For pixels $p_1 = (x_1, y_1), p_2 = (x_2, y_2)$, we define L_∞ norm to measure the distance between them, such that $|p_1 - p_2| \triangleq max\{|x_1 - x_2|, |y_1 - y_2|\}$.

The correlogram of I is defined for $i, j \in \{1, 2, 3, \cdots, k\}, d \in \{1, 2, 3, \cdots, l\}$ where distance d is fixed a priori, such that

$$\gamma^{(d)}_{c_i,c_j}(I) \triangleq \Pr_{p_1 \in I_{c_i}, p_2 \in I}[p_2 \in I_{c_j} \, | \, |p_2 - p_1| = d] \triangleq \frac{|I_{c_j} \cap I^d_{c_i}|}{|I^d_{c_i}|} \tag{5}$$

where $I^d_c \triangleq \{p_2 | p_1 \in I_c \wedge |p_2 - p_1| = d\}$, where $d \in \{1, 2, 3, \cdots, l\}$ is a distance between two given pixels in the image. Given any pixel of color c_i in the image, $\gamma^{(d)}_{c_i,c_j}(I)$ gives the probability that a pixel at distance d away from the given pixel is of color c_j. Hence, the color correlogram is a three-dimensional table indexed by color and distance between pixels and the size of the correlogram is $O(k^2 l)$.

The banded correlogram (Huang, 1998) is for storage trimming. Given b, for $1 \leqslant d \leqslant l/b$,

$$\bar{\gamma}^{(d)}_{c_i,c_j}(I) \triangleq \sum_{d'=(d-1)b+1}^{db} \gamma^{(d')}_{c_i,c_j}(I) \tag{6}$$

For each color pair (c_i, c_j), the probability values for the distances in the selected distance set whose cardinality is b are summed as a single number. Hence, a banded color correlogram is a restricted version of the color correlogram.

Distance Measure We use a relatively weighted L_1 distance measure for computing the distance between images I and I' as follows:

$$|I - I'|_{\gamma, L_1} \triangleq \sum_{i,j,d} \frac{|\gamma^{(d)}_{c_i,c_j}(I) - \gamma^{(d)}_{c_i,c_j}(I')|}{1 + \gamma^{(d)}_{c_i,c_j}(I) + \gamma^{(d)}_{c_i,c_j}(I')} \tag{7}$$

where $i, j \in \{1, 2, 3, \cdots, k\}$, and $d \in \{1, 2, 3, \cdots, l\}$.

The L_1 distance is also known as the manhattan distance. The manhattan distance between two points in an n-dimensional vector space with a fixed cartesian coordinate system is just the sum of the lengths of the projections of the line segment between the two points onto the coordinate axes. The normalization is such that non-uniform weights are assigned to the contribution of different colors to the dissimilarity between the two images. This is in keeping with the intuition that a difference in the number of pixels in any given

color bucket has a more significant contribution to the perceived dissimilarity if the content of that color in the image is low to start with. The same difference but when the color content is extremely high shouldn't contribute too much to the perceived dissimilarity between two images.

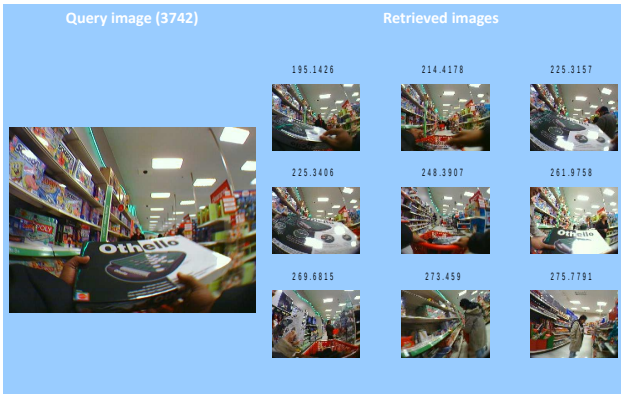


Figure 2: Query image and retrieved images (the color correlogram method).

Comparison: Common Neighbor Ratio

We now need to compare the performance of the two methods for our current purpose: to check if the distance measure on the respective representation does a good job of identifying as neighbors images that really are closely spaced in the time sequence. Events within a context are closely spaced in time and one of the major tasks for our method is to be able to accurately retrieve images that are from the same context. The idea here is that most of the closely spaced images as characterized by the distance measure ought to be closely spaced in time. Periodic events are exceptions where people might return to the same place after a certain duration. The images from those two episodes will be closely spaced but might be far apart in time. With this in mind, we define the common neighbor ratio. Given a positive integer k, for each image I, we find its k nearest neighbors both in the distance domain and in the time domain. Suppose $D_I = \{I_{d1}, I_{d2}, \cdots, I_{dk}\}$ are image I's k nearest neighbors in the distance space and $T_I = \{I_{t1}, I_{t2}, \cdots, I_{tk}\}$ are image I's k nearest neighbors in the time space (the images come with timestamps on them which are used in this calculation), then

$$common\ neighbor\ ratio = \frac{\sum_{I=1}^{n} |D_I \cap T_I|}{n \times k} \qquad (8)$$

where n is the total number of images. If k equals to n, then the ratio is 1. The method that has a higher common neighbor ratio is the better one. Figure 3 shows clearly that the correlogram representation and its corresponding distance measure outperforms the traditional histogram representation

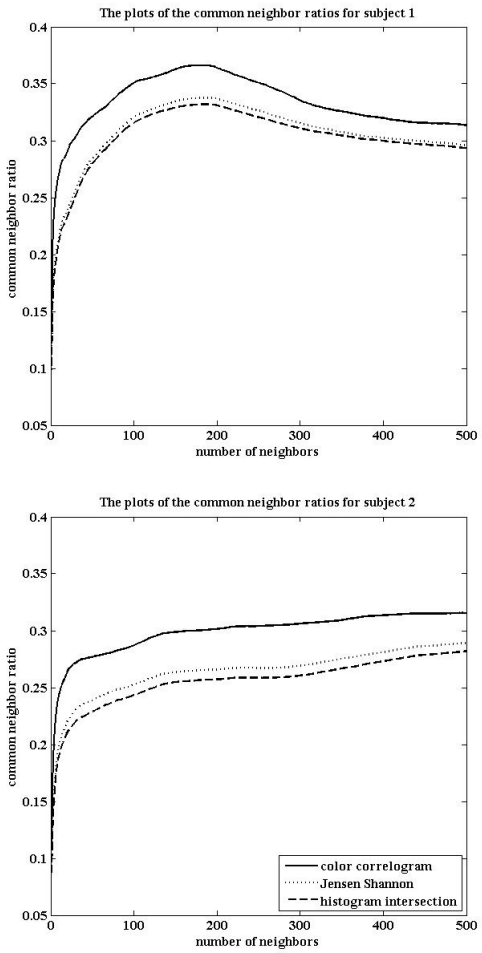


Figure 3: The common neighbor ratio as a function of number of nearest neighbors for image data from two subjects. The correlogram-relative L_1 distance method gives a higher ratio than the histogram-JS distance and the histogram intersection distance methods.

and the associated JS divergence distance measure. It can also be seen that the JS divergence measure works slightly better than the histogram intersection distance.

Correlation Dimension Analysis

Dimension measures are used to quantify the space filling properties of a set. A fractal dimension is a more informative measurement than a topological dimension which can take only integer values. For example, the topological dimension of a point is 0, of a line is 1 and of a surface is 2. A wiggly line is more space filling than a straight line but has a topological dimension of 1. The wiggly line is said to be a fractal if its fractal dimension is greater than its topological dimension (Mandelbrot, 1967). Fractal dimension measurements have been widely used in nonlinear dynamics time series analysis.

If a time series is from a nonlinear dynamical system or from a random process, the time series is irregular in both

time and frequency domains. Methods of time series analysis based on phase space reconstructions can reveal structure in time series from nonlinear dynamical systems as opposed to little structure in time series from random processes. Many popular methods of analysis involve correlation dimension estimates. There are several dimension measurements that are possible (Camastra, 2003). The correlation dimension is one of the simplest to calculate and is the most widely used dimension measurement in time series analysis. The correlation dimension is also related to the minimum number of variables needed to model the system's behavior in phase space.

The correlation dimension is a measure of the dimensionality of the space occupied by a set of points and is a type of fractal dimension because it allows non-integer values. Grassberger and Procaccia (1983a, 1983b) introduced the correlation dimension to characterize phase space filling properties of attractors. The set is covered by spheres of a given size r and the correlation dimension v is defined by:

$$v = \lim_{r \to 0} \sum_{i} \frac{\log\left(\sum_i p_i(r)^2\right)}{\log r} \tag{9}$$

where $\sum_i p_i(r)^2$ is the probability of finding a pair of points in a sphere of size r. For small values of r, this probability is the same as the probability of finding a pair of points separated by less than r. This probability, for large data sets, is given by the correlation sum. For N points in an M-dimensional space, the correlation sum is given by

$$C(r) = \frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1 j \neq i}^{N} H\left(r - \left|\vec{X}_i - \vec{X}_j\right|\right) \tag{10}$$

H is the heaviside function. Here, it counts the number of pairs of points which are separated by less than r. For sufficiently small r and large number of points N,

$$C(r) \propto r^v \tag{11}$$

Taking logarithms of each side, we get:

$$v \sim \frac{\log\left(C\left(r\right)\right)}{\log\left(r\right)} \tag{12}$$

v is calculated from the slope of the straight line scaling region of a $log(C(r)$ versus $log(r)$ plot.

Results

The color histogram method was used to represent the images and the square root of the Jensen-Shannon divergence was used to calculate the similarity between pairs of images. $log(C(r))$ was then recorded in a series of 1000 bins. The correlation dimension(s) v is the slope $\frac{d\log(C(r))}{d\log(r)}$ of the linear portion(s) of the $log(C(r))$ versus $log(r)$ plot. The same procedure was repeated for the color correlogram representations using the relative L_1 distances to calculate similarity between images. Figure 4 shows the correlation dimension plots for image data taken from 2 subjects. The left panel contains the results for the correlation dimension using the color histogram representation and the associated square root of the JS Divergence. The right panel contains the results using the color correlogram representation and the associated relative L_1 distance measure. Points close to zero have been discarded in the correlogram correlation dimension plots due to insufficient pairs of points in that region.

There are hints of a two scale structure in the histogram based correlation dimension plots but the correlogram based correlation dimension plots do not show this structure. More discussion follows in the next section.

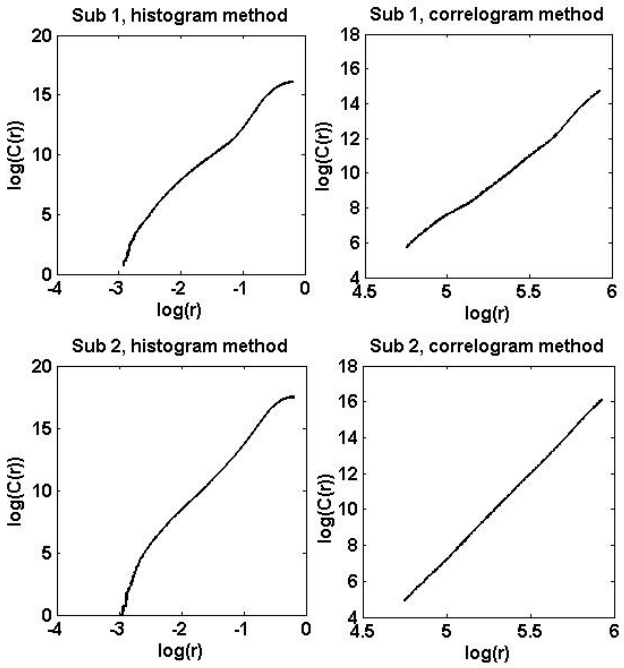


Figure 4: The correlation dimension plots for 2 subjects: The left panel is with the color histogram-JS div distance method and the right panel is with the correlogram-L1 distance method.

Conclusion and Discussion

Images were captured by subjects using a Microsoft^{TM}Research SenseCam. A correlation dimension analysis was done on images that were obtained from each subject. These images can be considered as representative of the visual input that goes into an individual's episodic memory. Distances between pairs of images represented by color histograms were calculated using the square root of the Jensen-Shannon divergence. Color histograms do not include spatial information. HSV autocorrelograms have been found to work better in image retrieval studies (Ojala, Rautiainen, Matinmikko, & Aittola, 2001). Spatial information in the images may be relevant here. For example, how do people recognize that two very different images in terms of color

content belong to the same episode? The distances calculated from the HSV histogram have given us sufficiently accurate nearest neighbour pairs as demonstrated in Figure 1 but the correlogram method and the associated L_1 distance measure was found to work better for our current purposes based on our definition of the common neighbor ratio. We conclude that the better method is the one that correctly identifies images that are close in time (within context) by classifying them as close in space based on the distance measure employed by the respective method.

A two scale structure was found in earlier studies on corpora of different languages (Doxas et al., 2010). The trajectory through a semantic space as one transitions from paragraph to paragraph in written discourse was shown to display a low dimensionality at short distances and higher at larger distances. This structure was observed in five corpora of written text in English, French, Modern Greek, Homeric Greek, and German respectively. The lower scale dimension of eight was observed to be approximately the same across languages. These structures suggest that there are strong constraints on the topology of the space through which authors move as they write and through which readers move as they read. The question now is if this is a constraint imposed by the cognitive system. This study is aimed at addressing this question. The images used represent the visual input that goes into a person's episodic experience, i.e., of the everyday events that one encounters (visually). The correlation dimension plots however don't reliably show a two scale structure here. Further exploration is necessary, however, to determine if the image representation meets all of the assumptions of the correlation dimension analysis as it has been used in this study. One such assumption is that the space has orthonormal basis vectors.

Acknowledgments

This work was supported by the Air Force Office of Scientific Research (AFOSR) under grant number FA9550-09-1-0614. We thank Microsoft for providing us the Microsoft Research SenseCams.

Références

Berry, E., Conway, M., Moulin, C., Williams, H., Hodges, S., Williams, L., et al. (2006). Stimulating episodic memory: Initial explorations using sensecam. In *Abstracts of the psychonomic society. 47th annual meeting* (Vol. 11, p. 56-57). Oxford University Press.

Camastra, F. (2003). Data dimensionality estimation methods: A survey. *Pattern Recognition*, *36*, 635–652.

Chiu, G. S. (2002). *Bent-cable regression for assessing abruptness of change*. Doctoral dissertation, Department of Statistics and Actuarial Science, Simon Fraser University.

Doxas, I., Dennis, S., & Oliver, W. L. (2010). Dimensionality of discourse. *Proceedings of the National Academy of Sciences*, *107*.

Goldberger, J., Gordon, S., & Greenspan, H. (2006). Unsupervised image-set clustering using an information theoretic framework. *IEEE Trans Image Process*, *15*.

Grassberger, P., & Procaccia, I. (1983a). Characterization of strange attractors. *Physical Review Letters*, *50*, 346–349.

Grassberger, P., & Procaccia, I. (1983b). Measuring the strangeness of strange attractors. *Physica D*, *9*, 189–208.

Greenspan, H., Goldberger, J., & Ridel, L. (2001). A continuous probabilistic framework for image matching. *Journal of Computer Vision and Image Understanding*, *84*, 384–406.

Huang, J. (1998). *Color-spatial image indexing and applications*. Doctoral dissertation, Department of Computer Science, Cornell University.

Huang, J., Kumar, S. R., Mitra, M., Zhu, W. J., & Zabih, R. (1997). Image indexing using color correlograms. In *Proceedings of the 1997 conference on computer vision and pattern recognition.*

Jeong, S., Won, C. S., & Gray, R. M. (2004). Image retrieval using color histograms generated by gauss mixture. *Computer Vision and Image Understanding: Special Issue on Color for Image Indexing and Retrieval*, *94*, 44–66.

Mandelbrot, B. (1967). How long is the coast of britain? statistical self-similarity and fractional dimension. *Science*, *156*, 636–638.

Ojala, T., Rautiainen, M., Matinmikko, E., & Aittola, M. (2001). Semantic image retrieval with hsv correlograms. In *Proc. 12th scandinavian conference on image analysis.* Bergen, Norway.

Smith, J. R., & Chang, S. F. (1995). *Automated image retrieval using color and texture* (Rapport technique N^{o} CU/CTR 408-95-14). Columbia University.

Smith, J. R., & Chang, S. F. (1996). Tools and techniques for color image retrieval. In *Symposium on electronic imaging: Science and technology - storage retrieval for image and video databases iv* (Vol. 2670). San Jose, CA.

Steyvers, M., & Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science*, *29*, 41–78.

Stockman, G., & Shapiro, L. (2001). *Computer vision*. Prentice-Hall.

Swain, M., & Ballard, D. (1991). Color indexing. *International Journal of Computer Vision*, *7*, 11–32.

Distributional Analyses in Visual Lexical Decision: Orthographic Neighborhood Density and Word Frequency Effects

Stephen Wee Hun Lim (gmslwhs@nus.edu.sg)
Melvin J. Yap (psyyapm@nus.edu.sg)
Department of Psychology, National University of Singapore
Block AS4, 9 Arts Link, Singapore117570

Abstract

The effects of orthographic neighborhood density and word frequency in visual word recognition were investigated using distributional analyses of response latencies in visual lexical decision. Main effects of density and frequency were observed in mean latencies. Distributional analyses, in addition, revealed a density x frequency interaction: for low-frequency words, density effects were mediated predominantly by distributional shifting whereas for high-frequency words, density effects were absent except at the slower RTs, implicating distributional skewing. The present findings suggest that density effects in low-frequency words reflect processes involved in early lexical access, while the effects observed in high-frequency words reflect late postlexical checking processes.

Keywords: Orthographic neighborhood density; word frequency; visual lexical decision; distributional analyses

Introduction

Word frequency and orthographic neighborhood density effects are among the most influential findings in the visual word recognition literature. Researchers study word recognition using the lexical decision task (LDT) that requires lexicality discrimination and decision where subjects would classify stimuli as either words or nonwords, and the speeded pronunciation (word naming) task that involves lexical access but excludes the word/nonword discrimination and decision components of the LDT. During word naming, subjects would typically be tested individually and read the stimuli into a microphone (see Andrews, 1997).

Word frequency effects, where latencies for common words are faster than those that are relatively less common, have been observed in many LDT studies (see Balota & Chumbley, 1990 for a review). In visual word recognition, frequency effects have been attributed to changes in activation thresholds or baselines. The logogen-style activation framework was inaugurated by Morton (1969), which assumes that information extracted from the sensory representation of the word leads to parallel activation of all word units that match that information. When sufficient activation has accumulated in a particular word unit, it reaches threshold and lexical access occurs. Morton's (1969) initial model was later specified in greater detail by McClelland and Rumelhart (1981). Their model, which they called the interactive activation model, suggests that activation occurs at three levels. Activation of featural units feeds to units corresponding to letters, which in turn activate the units for words containing these letters. Activity also feeds back from the word to the letter level, causing reverberating patterns of activity to occur between these levels. To ensure that only one word unit eventually obtains threshold, McClelland and Rumelhart (1981) also assume that inhibition occurs between word units, so that the activity level of competing word units is reduced relative to the maximally active node. Within the activation framework, word frequency is assumed to be reflected in the threshold (Morton, 1969) or resting activation level (McClelland & Rumelhart, 1981) associated with a particular word unit. The critical interpretation is that less evidence is required to enable recognition of a high-, than a low-, frequency word.

The findings for orthographic neighborhood density effects (N), on the other hand, appear to be more mixed. The N metric has been defined by Coltheart, Davelaar, Jonasson, and Besner (1977) as the number of close neighbors a word has and refers to the number of words that can be created by changing a single letter of this target word. For instance, *tell* has many neighbors such as *well*, *yell*, *sell*, *teal* and *tall*, while *once* has no neighbors. Neighborhood effects can help specify the mechanisms underlying lexical access. The implication of the overlap in the features constituting different words is that any subset of the features constituting a particular word is unlikely to uniquely specify its corresponding lexical representation. Neighbors are items that are highly confusable with the target word, in the sense that they share a large number of their features with the target. Thus, it seems inevitable that some or all of the neighbors of a target word will be selected by the access mechanisms as eligible target candidates.

Effects of N can be accommodated within activation-based models of lexical access, and appear to provide substantive support for an activation mechanism. If presenting a word leads to an activation of all lexical items that sufficiently match features of the target word, the density of the word's neighborhood should influence access time. Unfortunately, this class of models does not make precise predictions about the nature of the effect of neighborhood density. McClelland and Rumelhart's (1981) interactive activation model, for instance, assumes excitatory links between levels which can account for facilitatory effects of neighborhood size. Activated neighbors will feed back to their constituent letters which in turn lead to heightened activation of word units containing these letters. According to McClelland and Rumelhart

(1981), such facilitatory effects of N are likely to be greater for low- than high-frequency words. The reason is that high-frequency words have higher base activation levels and are therefore likely to reach threshold before allowing reverberating letter-level activation from neighboring word units to become influential.

Yet, the same model can also predict inhibitory effects of neighborhood size because of its assumption of lateral inhibition between word nodes. Active nodes send inhibition to other active nodes to an extent that is proportional to their current activation. If the unit corresponding to the target word becomes activated before other units, this inhibitory mechanism would decrease background activation and make the target more salient. On the other hand, if nodes corresponding to neighbors obtained activation before the target word, these activated competitors would inhibit activation of the target and delay threshold activation. The more neighbors a word has, the greater the likelihood that the target unit would fall prey to this inhibitory mechanism, resulting in interfering effects of large neighborhoods. Thus, depending on the relative contribution to performance of excitatory activation between letter and word levels, as well as inhibitory activation within the lexical level, the interactive activation model can explain facilitatory, inhibitory, or null effects of neighborhood size.

Using the visual LDT paradigm, Coltheart et al. (1977) first observed that low-N nonwords were classified more quickly than high-N nonwords, but that N did not influence performance for English words. The researchers interpreted their data using Morton's (1969) logogen-style activation framework, in which the strength of activation in individual logogens is determined by sensory input and is insensitive to activity in other logogens. The researchers then attributed N effects on nonword classification to a decision mechanism that is sensitive to the overall lexical activation. Subsequently, Andrews (1989) reported that N actually influenced responses to English words in the LDT when the words were selected to orthogonally manipulate N and word frequency. Specifically, it was reported that high N facilitated performance for words, but only for the 4-letter low-frequency words. These facilitatory effects of N, which are not incompatible with McClelland and Rumelhart's (1981) interactive activation model, were later replicated in several other experiments (e.g., Sears, Hino, & Lupker, 1995; Michie, Coltheart, Langdon, & Haller, 1994; Andrews, 1992). However, Grainger, O'Regan, Jacobs, and Segui (1989) concurrently found no systematic relationship to exist between N and performance in the LDT; lexical decision latencies were not affected by the number of neighbors *per se*.

Traditionally, visual lexical decision studies that examined neighborhood effects have used mean RT differences among the experimental conditions to make inferences about the mechanisms underlying the recognition process. The implicit assumption that the researchers would have made is that RT distributions across conditions are symmetrical, where the mean constitutes a reasonably good estimate of the central tendency of these distributions. But RT distributions are in fact rarely symmetrical around a mean. They typically assume a positively skewed unimodal shape which contains information that cannot be derived from the mean and variance of the distributions. For instance, mean RT differences, or the lack thereof, between conditions can be due to changes in the shape (skew) of the distribution in itself or in addition to a shift in the modal portion of the distribution. By relying on a traditional RT analysis that uses mean RTs as the dependent variable (DV) to interpret LDT performance, one can, in some instances, fail to recognize the tradeoff between the effects of shifting and skewing, and be misled to incorrectly infer null results (Heathcote, Popiel, & Mewhort, 1991). Recognizing the problems concerned with the traditional RT analysis approach, several researchers have argued that the nature of the RT distributions ought to be scrutinized more closely (e.g., Balota, Yap, Cortese, & Watson, 2008; Heathcote et al., 1991).

Two distributional analyses techniques were used in the present study, namely the ex-Gaussian and Vincentile analyses. Shifting and skewing in the RT distributions were investigated using the ex-Gaussian function. The procedure was to fit an empirical RT distribution to this theoretical function that captures important aspects of typical RT distributions. The ex-Gaussian function conceptualizes RT distributions as the convolution of two underlying distributions: a Gaussian distribution and an exponential distribution. The mean and standard deviation of the Gaussian component are captured by the mu and tau parameters, while the exponential function is captured by the sigma parameter that reflects its mean and standard deviation. An important property of the ex-Gaussian function is that the mean of the RT distribution is constrained to be the algebraic sum of the mu and tau parameters obtained by fitting that distribution. This constraint allows one to partition mean differences into individual components due to distributional shifting (mu) and skewing (tau), and then make inferences from these components to determine the nature of the effect of an independent variable (IV) (see Balota et al., 2008).

Parameter estimates from the ex-Gaussian function were supplemented by analyses of Vincentiles to enable a graphical, non-parametric estimate of the variable's effect. In these analyses, the RTs are ordered, from fastest to slowest, within each condition, and the average of the first 10%, that of the second 10%, and so forth, are plotted. The mean of the Vincentiles across participants can then be plotted to obtain a description of how the RT distribution is changing across conditions. Importantly, differences between two levels of an IV across Vincentiles can be graphically represented to reveal how the effect of an IV may change across different portions in the RT distribution.

This study had two goals. The first was to replicate the N effects in the visual LDT in the light of the initial contradictory reports (see Grainger et al., 1989; Andrews,

1992, 1989; Coltheart et al., 1977). The present hypothesis was that facilitatory effects of density would be observed, but only for low-frequency words (see Sears et al., 1995; Michie et al., 1994, Andrews, 1992, 1989). The second, and more important, was to extend the ex-Gaussian and Vincentile analyses techniques to the orthographic neighborhood density and word frequency effects found in the extant visual lexical decision studies, and to explore the extent to which these two effects are driven by distributional shifting and skewing.

Method

Participants

Fifty-seven introductory psychology undergraduates with no reported history of speech or hearing impairment participated for course credit. Their mean vocabulary age of the Shipley Test was 18.09 (*SD* = 1.06).

Design and Materials

A 2 (Neighborhood Density: low, high) x 2 (Word Frequency: low, high) within-subjects design was used. Forty 4-letter English words were selected for each of the four conditions, and their properties are summarized in Table 1. Two-way analyses of variances (ANOVAs) showed a main effect of frequency, $F(1, 156) = 19826.68$, $MSe = 0.67$, $p < .001$, for the log-frequency values ($M = 6.58$, $SD = 0.53$ for low-frequency words and $M = 11.67$, $SD = 1.02$ for high-frequency words), and a main effect of density, $F(1, 156) = 1827.88$, $MSe = 2.10$, $p < .001$ for the density values ($M = 3.35$, $SD = 1.38$ for low-density words and $M = 13.14$, $SD = 1.50$ for high-density words). No other effects were significant, $Fs < 1$. The 160 legal non-words used were obtained from the ARC non-word database (Rastle, Harrington, & Coltheart, 2002) and were matched against the 160 words in terms of length and density.

Procedure

Participants were tested on individual PCs in groups of seven or fewer. E-prime 1.2 and the PST Serial Response Box (Schneider, Eschman, & Zuccolotto, 2002) were used for stimuli presentation and data collection. Participants were instructed to indicate as quickly and as accurately as possible whether the visual token presented on each trial was a real English word (or a non-word). The left- and right-most buttons of the button-box were labeled *No* and *Yes* respectively. On each trial, a fixation cross appeared and remained on the screen for 500ms, and terminated for 200ms before the target word appeared. RT was measured from the onset of the target stimulus to the button-press. Accuracy feedback was provided for each trial. A practice set of 20 trials for task familiarization was given, using stimuli unrelated to the experiment. The 320 experimental trials were then presented in a random order for each participant, with a short self-paced break after every set of 80 trials was completed.

Table 1: Mean Density and Log-frequency of the Words in the Neighborhood Density and Word Frequency Conditions.

	Density		Log-frequency	
Conditions	*M*	*SD*	*M*	*SD*
Low-frequency				
Low-density	3.33	1.33	6.61	0.54
High-density	13.05	1.95	6.56	0.52
High-frequency				
Low-density	3.38	1.44	11.67	1.23
High-density	13.23	0.86	11.67	0.78

Results

Errors and latencies faster than 200 ms or slower than 3000 ms were first excluded, and the overall word and non-word means and SDs for each participant were computed across all conditions. Following which, latencies exceeding 2.5 SDs from the participant mean, as well as items where proportion of correct responses was not at least 0.5, were removed. Table 2 summarizes the results obtained from mean latencies, accuracy, and the ex-Gaussian parameters. Two way ANOVAs by participants and items were

Table 2: Mean Latency, Accuracy, and Ex-Gaussian Parameter Estimates Across Neighborhood Density and Word Frequency

Conditions	Latency	Accuracy	Mu	Sigma	Tau
Low-frequency					
Low-density	679 (123)	87 (11)	535 (79)	59 (38)	147 (89)
High-density	662 (127)	88 (8)	509 (74)	54 (37)	157 (84)
Density effect	17	-1	26	5	-10
High-frequency					
Low-density	554 (90)	98 (2)	444 (45)	35 (14)	112 (62)
High-density	546 (83)	99 (1)	442 (47)	38 (16)	105 (54)
Density effect	8	-1	2	-3	7
Interaction	9	0	24	8	-17
Non-words	692 (144)	94 (4)	542 (68)	58 (23)	152 (90)

performed for latencies and accuracy, and by participants for the ex-Gaussian parameters.

Latency

For latency, reliable main effects of density, $F_p(1, 54) = 11.51$, $MSe = 790.68$, $p < .01$, and frequency, $F_p(1, 54) = 222.87$, $MSe = 3600.78$, $p < .001$, were obtained for the analyses by participants. Participants were faster in responding to high-density words ($M = 604$, $SD = 102$) than to low-density words ($M = 617$, $SD = 104$); they were also faster in responding to high-frequency words ($M = 550$, $SD = 83$) than to low-frequency words ($M = 671$, $SD = 123$). For the analyses by items, a reliable main effect of frequency was obtained, $F_i(1, 153) = 299.53$, $MSe = 1981.84$, $p < .001$. High-frequency words yielded a shorter response time ($M = 551$, $SD = 30$) as compared to low-frequency words ($M = 674$, $SD = 56$). No other effects were significant, $Fs < 2.01$, $MSes < 1982.84$, $ps > .1$.

Accuracy

For accuracy, there was a reliable main effect of frequency, $F_p(1, 54) = 90.97$, $MSe = 0.007$, $p < .001$ for the analyses by participants; the main effect of density was marginally significant, $F_p(1, 54) = 3.53$, $MSe = 0.001$, $p = .066$. Participants were more accurate with high-frequency words ($M = 98$, $SD = 0.01$) than with low-frequency words ($M = 88$, $SD = 0.09$); they also tended to be more accurate with high-density words ($M = 93$, $SD = 0.04$) than with low-density words ($M = 92$, $SD = 0.06$). For the analyses by items, a reliable main effect of frequency was obtained, $F_i(1, 153) = 55.86$, $MSe = 0.008$, $p < .001$. High-frequency words yielded a higher accuracy rate ($M = 98$, $SD = 4$) as compared to low-frequency words ($M = 88$, $SD = 12$). No other effects were significant, $Fs < 1.16$, $MSes < .008$, $ps > .1$.

Mu

Turning to the ex-Gaussian parameters, for mu, there were reliable main effects of density, $F(1, 54) = 18.61$, $MSe = 589.63$, $p < .001$, and frequency, $F(1, 53) = 160.02$, $MSe = 2151.95$, $p < .001$. These main effects were qualified by the significant interaction, $F(1, 53) = 12.00$, $MSe = 726.81$, $p < .01$. Simple main effects analyses at each level of the frequency factor revealed that for low-frequency words, mu was larger for low-density words compared to high-density words, $F(1, 54) = 16.56$, $MSe = 1185.71$, $p < .001$, but there was no density difference for high-frequency words, $F < 1$. This finding implicates a shift in the modal portion of the RT distribution as a function of density, but only for low-frequency words.

Sigma

For sigma, a significant main effect of frequency was obtained, $F(1, 54) = 25.75$, $MSe = 890.57$, $p < .001$. Sigma was larger for low-frequency ($M = 57$, $SD = 29$) than high-frequency ($M = 36$, $SD = 13$) words. No other effects were significant, $Fs < 1.32$, $MSes < 645.75$, $ps > .1$.

Tau

For tau, a significant main effect was obtained for frequency, $F(1, 54) = 37.25$, $MSe = 2834.02$, $p < .001$, but not for density, $F < 1$. The main effect of frequency appears to be qualified by the marginally significant interaction, $F(1, 54) = 3.21$, $MSe = 1417.98$, $p = .079$. Follow-up analyses indicated that tau tends to be smaller for low-density words compared to high-density words for low-frequency words, but it tends to be larger for low-density words compared to high-density words for high-frequency words. More important, a cross examination of the tau data, with the mu data, revealed that the small density effect observed for the high-frequency word condition appears to be attributable to distributional skewing, rather than distributional shifting.

Recall that one important constraint of the ex-Gaussian analyses is that the mean of the RT distribution is the algebraic sum of mu and tau. In the traditional mean latency analyses, only reliable main effects of frequency and density were obtained; there was no reliable frequency x density interaction. Analyses of the ex-Gaussian parameters provide important observations that constitute a more faithful account of the apparent lack of interaction between the factors; the tradeoff between the mu and tau parameters accounts for why the mean interaction effect was very small (see Table 2). First, analyses of the mu parameter as a function of density suggest that there is distributional shifting only for the low-frequency words but not for the high-frequency words. This finding strongly suggests that the density effect observed for low-frequency words in the traditional mean latency analyses is predominantly mediated by distributional shifting. Second, analyses of the tau parameter, in conjunction with the mu parameter, strongly suggest that the small density effect observed for high-frequency words in the traditional mean latency analyses is, on the other hand, largely mediated by distributional skewing.

To corroborate this interpretation, vincentile analyses were performed on the RT data. Figure 1 shows the mean vincentiles across the different experimental conditions. The lines represent the estimated vincentiles of the best-fitting ex-Gaussian distribution. This graphical representation allows a visual assessment of the goodness-of-fit between the empirical and estimated vincentiles.

From the top panel, it is clear that the density effect is observed for the low-frequency words across all vincentiles. The high-density means are always below the low-density means in each of the vincentiles. In the middle panel, the density effect is only apparent at the later vincentiles. The differential density effects can be seen more clearly in the bottom panel, which plots the difference scores between the low- and high-density means for each of the low- and high-frequency conditions. It can be observed that the density effect generally remains stable across vincentiles for the low-frequency words, indicating that the difference between low- and high-density words remains fairly constant as RT increases. This trend implicates distributional shifting *per se*. However, for high-frequency words, the density effect increases only in the slower RTs. This trend implicates distributional skewing *per se*.

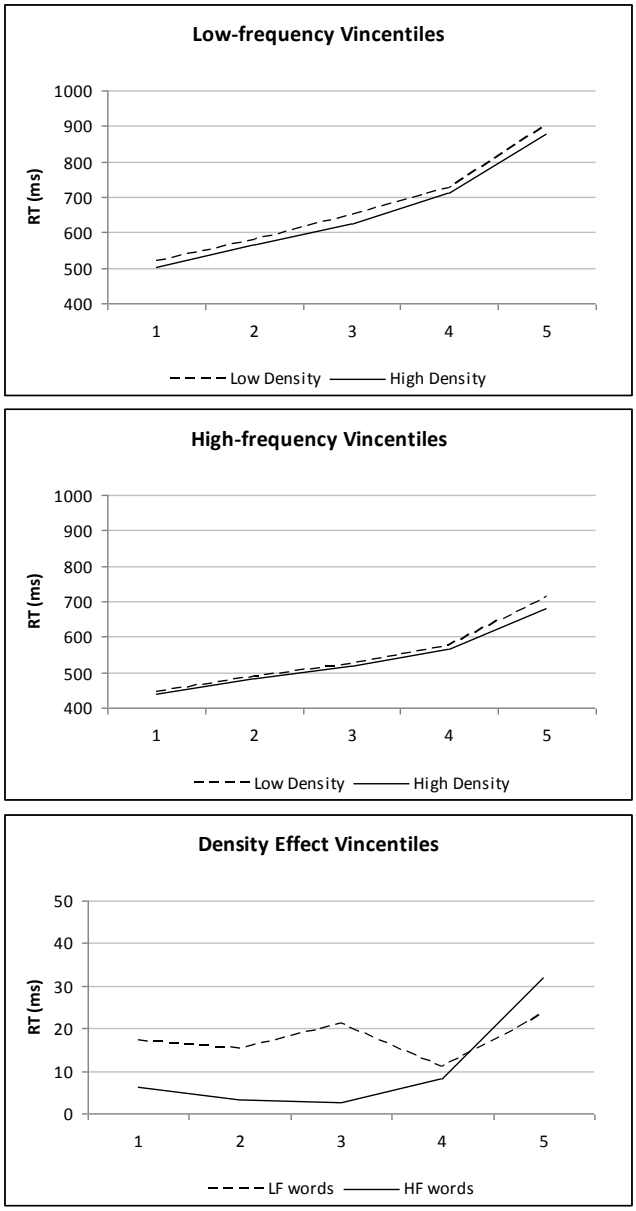


Figure 1: Vincentiles of lexical decision performance. The participants' mean vincentiles are represented across different conditions. The lines represent the estimated vincentiles of the best-fitting ex-Gaussian distribution. The top and middle panels show performance as a function of density in the low- and high-frequency conditions respectively, while the bottom panel shows the density effect.

Discussion

RT distributional analyses of orthographic neighborhood density and word frequency effects in visual lexical decision have not been done in previous studies examining neighborhood effects, which relied on *mean* RTs as the primary DV. The findings in the present study can be summarized as follows.

First, facilitatory effects of frequency, where high-frequency words elicited faster RTs than low-frequency words did, and of density, where words from high-density neighborhoods elicited faster RTs than words from low-density neighborhoods did, were obtained.

Second, and more important, the distributional analyses revealed a density x frequency interaction which was primarily attributable to differential shifting and skewing of the latency distribution between low- and high-density words as a function of frequency. For low-frequency words, the density effect obtained, replicating Andrews' (1992, 1989) finding, and the effect was predominantly mediated by distributional shifting; for high-frequency words, the small density effect observed was primarily mediated by distributional skewing.

A shift in the RT distribution as a function of density for low-frequency words is compatible with existing accounts which assume that lexical access relies upon an activation mechanism. Such an activation mechanism, which postulates top-down feedback from word to letter nodes, characterizes McClelland and Rumelhart's (1981) interactive activation model[1] which assumes parallel activation of both lexical units and units that correspond to sublexical components, such as letters. First, the assumption must hold that excitatory activation between lexical and sublexical units is not cancelled out by lateral inhibition at the lexical level. Then, the partial activation of neighbors can increase the activation of sublexical components of the target, and consequently accelerate access to the target representation.

To explain the present data within such an activation mechanism framework, one must specify why the neighborhood effects arising from such sublexical/lexical interactions would affect only responses to low-frequency words. Frequency effects have mainly been attributed to differences in the resting activation level of lexical units within the original logogen (Morton, 1969) as well as the interactive activation (McClelland & Rumelhart, 1981) accounts. A functionally equivalent assumption appears to characterize distributed memory models (McClelland & Rumelhart, 1985) that assume that frequency determines how rapidly a lexical unit reaches a threshold level of activation. The present interaction between frequency and neighborhood size implicates that sublexical units play a greater role in the recognition of low-, rather than high-, frequency words; high-frequency words obtain threshold sufficiently quickly through direct activation of lexical

[1] Although activation models, such as McClelland and Rumelhart's (1981), can accommodate the present data, one must recall that whether the net effect of neighborhood size is facilitatory or inhibitory depends, within this framework, on the relative values of the parameters governing letter-word excitation, word-word inhibition, and the base activation level associated with word frequency. In a sense, rather than regarding the present data as supporting the model *per se*, it might be more appropriate to regard the data as providing evidence that constrains the future specification of activation models.

units, such that they are not influenced by the reverberating sublexical activation arising from active neighbors.

The increase in response time as a function of density for low-frequency words observed in the present study appears to be additive in nature, reflected by the distributional shift. Such a shift effect has been argued by Balota and Spieler (1999) to indicate early automatic processes, rather than later analytical or more attention-demanding processing. That density effects for low-frequency words are predominantly mediated by distributional shifting reflect processes involved in early lexical access, and not late postlexical processes which may also be involved in the LDT.

On the other hand, for high-frequency words, it appears that density effects are absent except at the slower end of the distribution, which are reflected in slightly greater skewing for low-density words. Recall that under the activation framework (e.g., McClelland & Rumelhart, 1985), high-frequency words obtain threshold sufficiently quickly through direct activation of lexical units, such that lexical *access* need not be facilitated by the reverberating sublexical activation arising from activated neighbors. The tau parameter revealed, for high-frequency words, some difference in RTs comparing low- with high-N words. It appears that high-frequency words with small neighborhoods would have received little facilitation from their active neighbors to aid lexicality *decision* of the target, as compared to those with big neighborhoods. Where facilitatory effects of N were lacking, compensatory postlexical checks could tend to be adopted, resulting in slightly longer RTs for low-N words. The emergence of density effects at the tail end of the distribution may therefore reflect, particularly for the low-N words, late postlexical checking processes that are specific to the lexical decision task (see Balota & Chumbley, 1984), rather than early lexical access processes.

Conclusion

The present study extends previous work on distributional analyses and underscores the contribution of these techniques in illuminating the interaction between orthographic neighborhood density and word frequency effects in a visual LDT. The new understanding is that the effects of density as a function of frequency are represented differentially in the shift and skew of the underlying RT distributions.

References

Andrews, S. (1989). Frequency and neighborhood size effects on lexical access: Activation or search? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*, 802-814.

Andrews, S. (1992). Frequency and neighborhood effects on lexical access: Lexical similarity or orthographic redundancy? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*, 234-254.

Balota, D. A., & Chumbley, J. I. (1984). Are lexical decisions a good measure of lexical access? The role of word frequency in the neglected decision stage. *Journal of Experimental Psychology: Human Perception and Performance, 10*, 340-357.

Balota, D. A., & Chumbley, J. I. (1990). Where are the effects of frequency in visual word recognition tasks? Right where we said they were! Comment on Monsell, Doyle, and Haggard (1989). *Journal of Experimental Psychology: General, 119*, 231-237.

Balota, D. A., & Spieler, D. H. (1999). Word frequency, repetition, and lexicality effects in word recognition tasks: Beyond measures of central tendency. *Journal of Experimental Psychology: General, 128*, 32-55.

Balota, D. A., Yap, M. J., Cortese, M. J., & Watson, J. M. (2008). Beyond mean response latency: Response time distributional analyses of semantic priming. *Journal of Memory and Language, 59*, 495-523.

Coltheart, M., Davelaar, E., Jonasson, J. T., & Besner (1977). Access to the internal lexicon. In S. Dornic (Ed.), *Attention and performance VI* (pp. 535-555). Hillsdale, NJ: Erlbaum.

Grainger, J., O'Regan, J. K., Jacobs, A. M., & Segui, J. (1989). On the role of competing words unit in visual word recognition: The neighborhood frequency effect. *Perception & Psychophysics, 45*, 189-195.

Heathcote, A., Popiel, S. J., & Mewhort, D. J. K. (1991). Analysis of response time distributions: An example using the Stroop task. *Psychological Bulletin, 109*, 340-347.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. *Psychological Review, 88*, 375-407.

McClelland, J. L., & Rumelhart, D. E. (1985). Distributed memory and the representation of general and specific information. *Journal of Experimental Psychology: General, 114*, 159-188.

Michie, P. T., Coltheart, M., Langdon, R., & Haller, M. (1994). *Effects of orthographic neighborhood size on visual word recognition: Behavioral, electrophysiological and computational evidence.* Unpublished manuscript, Macquarie University.

Morton, J. (1969). Interaction of information in word recognition. *Psychological Review, 76*, 165-178.

Schneider, W., Eschman, A., & Zuccolott, A. (2002). *E-Prime User's Guide*. Pittsburg: Psychology Software Tool Inc.

Sears, C. R., Hino, Y., & Lupker, S. J. (1995). Neighborhood size and neighborhood frequency effects in word recognition. *Journal of Experimental Psychology: Human Perception and Performance, 21*, 876-900.

Linguistic and Non-Linguistic Influences on Learning Biases for Vowel Harmony

Sara Finley (sfinley@bcs.rochester.edu)
Department of Brain and Cognitive Sciences, Meliora Hall
Rochester, NY 14627 USA

William Badecker (badecker@arizona.edu)
Cognitive Science, Communications Building
Tucson, AZ 85702 USA

Abstract

This paper addresses the question of the domain-specificity of learning biases for phonological processes. In two artificial grammar learning experiments we explore the role of learning biases in shaping the distribution of phonological patterns across the world's languages. In Experiment 1, we demonstrate that learners are biased toward phonological patterns that occur in natural language, as opposed to patterns that are not found across the world's languages. Specifically, learners are biased towards directional vowel harmony spreading processes. In Experiment 2, we exposed learners to a non-linguistic analogue to vowel harmony. Learners processed spreading such that learners favored the cross-linguistically valid pattern only when the first item of the series underwent spreading. This set of similarities and differences in learning may provide some insight into the origin of learning biases for spoken languages.

Keywords: artificial grammar learning; phonology.

Introduction

The experiments presented in this paper address the hypothesis that typological restrictions on languages are due to learning biases (Slobin, 1973). Specifically, we address the distribution of vowel harmony across the world's languages. Vowel harmony is a phonological process that induces statistical tendencies for words to share the same vowel quality along a particular phonetic dimension. In Turkish, which displays harmony for both backness and rounding, if the first vowel of the word is front and unround (with some exceptions), all following vowels must be both front and unround as well (Clements & Sezer, 1982). Thus, Turkish vowel harmony may be thought of as a directional spreading process in which the leftmost vowel spreads its feature (round, back) to the right.

Vowel harmony languages exhibit both left-to-right and right-to-left spreading characteristics. The direction of spreading can be decided by the morphology of the language (stems are more likely to spread harmony than affixes (Bakovic, 2000)) as well as the characteristics of the input vowels (spreading [+Round] is more likely than spreading [–Round] (Korn, 1969)). The direction of spreading can also be set such that spreading always occurs from right to left or from left to right. One way in which the direction of spreading is never decided is by the number of changes from the input to the output of the phonological process. For example, consider the disharmonic input /– + +/. There are two possible harmonic outputs: [– – –], which changes the feature value of two of the input vowels, and [+++] which changes only one of the vowels in the input. A left-to-right spreading language chooses [– – –] even though two vowels change. Another possibility is to have no intrinsic direction of spreading, but to choose the harmonic output with the fewest changes from the input (in this case [+++]). This type of spreading is termed 'majority rules' because the direction of spreading is determined by the majority feature value of the input (Bakovic, 2000). One peculiarity is that while languages never use 'majority rules' to determine the direction of spreading, 'majority rules' grammars are extremely easy to produce in generative phonology[1]. Generative linguistics assumes that the non-existence of patterns in natural language implies that they should not be generated by the grammar. However, it is possible that the lack of 'majority rules' grammars is due to an accidental gap. Under this assumption, 'majority rules' patterns are grammatically plausible, but the lack of such languages is an accident of history and language sampling.

One way of distinguishing between a principled restriction on the nature of vowel harmony languages and an accidental gap account is through testing for learning biases. If learners are biased against 'majority rules' languages and biased towards a directional harmony pattern, it suggests that the non-existence of 'majority rules' languages is a valid restriction on grammar. Because it is impossible to test learning biases for unattested languages in a naturalistic setting, as there are no naturalistic settings where a 'majority rules' grammar might be present, the artificial grammar learning paradigm is the best method for addressing this question. In an artificial grammar learning paradigm, it is possible to manipulate naturalness, complexity and statistical regularities in a way that is impossible with naturalistic studies of language learning.

The present experiments test whether learners make use of the 'majority rules' strategy when making grammaticality judgments between harmonic items. We present an experimental paradigm in which learners are exposed to a harmony language that is ambiguous between directionality and 'majority rules'. If learners are biased towards directional patterns and against 'majority rules' patterns,

[1] In 'majority rules' grammars, "ties" (e.g., two round and two unround vowels) are decided by a default strategy (lower-ranked constraint).

they should infer a directional pattern given data ambiguous between 'majority rules' and directionality. By pitting 'majority rules' and directional spreading against each other, it will be possible to determine what kind of pattern learners inferred. One reason testing for biases towards directionality and against 'majority rules' (as opposed to direct learnability) is that unnatural patterns may be learned by a language learner given the proper cues (Anderson, 1981). Further, even if 'majority rules' grammars are learnable, it still could be that learners are simply biased against 'majority rules' given the fact that much of their learning data will be ambiguous between other types of harmony (e.g., directional spreading). The present experiments capitalize on this hypothesis by exposing learners to language data that is ambiguous between 'majority rules' and a directional pattern.

Experiment 1

Participants were exposed either to a left-to-right harmony pattern or a right-to-left harmony pattern in which the majority of the vowels in the input spread. If participants learn a 'majority rules' pattern, they will reverse the direction of spreading when the majority feature reverses, but if participants learn a directional pattern, they will be consistent with the direction of spreading.

Methods

Participants All participants were adult native English speakers with no knowledge of a vowel harmony language. Twenty-four Johns Hopkins undergraduate students participated for extra course credit. Participants were randomly assigned to one of three training conditions: Control, Right-to-Left and Left-to-Right.

Design Because 'majority rules' patterns involves choosing the direction of spreading based on the proportion vowels with a particular feature in the input, it is necessary to provide clear evidence that the vowel harmony process involves a change from input to output. Because inputs to grammatical processes are abstract and not available on the surface, we trained participants on a compounding process where the underlying forms are available as separate lexical entries. Participants were exposed to base forms (the inputs) in addition to their concatenation as a compound (participants in the Control condition were exposed to input forms only). Training consisted of three single syllable forms in isolation, followed by their harmonic concatenations. The harmony rule paired back/round vowels together such that a harmonic trisyllabic item contained all front vowels ([i, e]) or all back vowels ([u, o]). The three individual syllables were disharmonic such that their faithful concatenation would be disharmonic. The concatenated form always followed 'majority rules', in one particular direction. Participants in the critical conditions were trained on either right-to-left harmony (Right-to-Left condition) or left-to-right harmony (Left-to-Right condition). All items were ambiguous between directionality and 'majority rules'. In the Left-to-Right condition [pu], [gu], [de] is concatenated as [pugudo], where the final vowel changes to [+Round] to match the feature values of the first two vowels (e.g., [+] [+] [–] → [+ + +]). In the Right-to-Left condition [pi], [gu], [do] is concatenated as [pugudo] ([–] [+] [+] → [+ + +]). There was a 500ms pause between the trisyllabic forms and the concatenated form. There were 24 alternations of monosyllablic words and their harmonic trisyllabic concatenations. All training items involved a single change from the input to the output.

The compounding procedure is similar to the triad procedure used to study phonological processes in infants (Jusczyk, Smolensky, & Alloco, 2002) in which the infants are given two forms followed by their concatenation. While there is some concern that learners do not infer a phonological process in this paradigm, adapting this paradigm to adults makes it possible to alleviate some of these concerns. First, participants were specifically informed that the trisyllabic item was the 'combined form' of the first three monosyllabic items. Second, the forced-choice task (described below) makes it possible to test for preference for left-to-right versus right-to-left spreading.

Table 1: Training Items for Experiment 1

Left-to-Right		Right-to-Left	
bo du ti	bodutu	be du tu	bodutu
gi te ko	giteke	gu te ke	giteke
mo bo di	mobodu	me bo nu	mobonu
pi ke to	pikete	pu te ne	pitene

All stimuli were recorded in a sound proof booth at 22,000kHz by a male speaker of American English with basic phonetic training (had completed a graduate-level phonetics course). While the speaker had no knowledge of the specifics of the experimental design, he was aware that the items would be used in an artificial language learning task. All stimuli were phonetically transcribed, and presented to the speaker in written format. The speaker was instructed to produce all vowels as clearly and accurately as possible, even in unstressed positions. Stress of the concatenated forms was produced on the initial syllable. All sound editing was done using Praat (Boersma & Weenink, 2005). All stimuli contained the same consonant inventory: [p, b, t, d, k, g, m, n]. The vowel inventory for all conditions consisted of [i, u, e, o]. The training stimuli were counterbalanced to contain all possible combinations of vowel sounds. Consonants were also counterbalanced such all consonants appeared equally often in each position. Concatenated words were produced semi-randomly with the condition that any word too closely resembling an English word was intentionally avoided (the final profile of the stimuli contained consistent numbers of vowel and consonant pairs).

Following training, participants were given a two-alternative forced-choice task. In this task participants were

given two pairs of three-syllable items. The first member of each pair was the disharmonic form, and the second member was a harmonic form with either spreading from right-to-left or left-to-right (e.g., [pi] [de] [go] [pudogo] vs. [pi] [de] [go] [pidege]). Participants were asked to choose which pair was the one that best fit the language they were trained on. At test, the critical items are reversed such that spreading the majority feature value requires spreading in the opposite direction. If learners infer a directional pattern, then they will accept multiple items undergoing harmony from the input to the output. If learners infer a 'majority rules' pattern, they will reverse the direction of spreading. Test items included 12 Old Items, 12 New Items and 12 New Direction Items. Old and New items have the majority feature reflect direction of spreading that the participant was trained on, but the New Direction items reflect a reversal of the direction that the participants were trained on. Examples of test items appear in Table 2.

Table 2: Examples of Test Items
('majority rules' Items **bold**, Directional Items <u>underlined</u>)

	Left-to-Right	Right-to-Left
Old	de mi ku **<u>demiki</u>** vs. de mi ku domuku	pu mi te pumuto pu mi te **<u>pimite</u>**
New	nu pu ki **<u>nupuku</u>** nu pu ki nipiki	nu pi ki nupuku nu pi ki **<u>nipiki</u>**
New Direction	pu mi te <u>pumuto</u> pu mi te **pimite**	de mi ku **demiki** de mi ku <u>domuku</u>

Procedure All phases of the experiment were run using Psyscope X (Cohen, MacWhinney, Flatt, & Provost, 1993). All participants were given written and verbal instructions. They were told that they would be listening to a language they had never heard before, and that they would later be asked about the language, but they need not try to memorize any forms they heard. They were told that the language would be presented in terms of three single syllable items followed by their combined form. This was done to ensure that participants inferred that the monosyllabic items were in fact the input to the harmonic concatenation. Participants heard all 24 concatenated forms in a random order, repeated 5 times. No information about vowel harmony was given. No semantics accompanied the sound pairs.

Training was followed by a forced-choice test phase in which participants heard the three mono-syllabic inputs followed by a choice of harmonic concatenations: all round or all unround. If the first concatenation of the syllables belonged to the language, they must push the 'a' key on the keyboard; if the second concatenation of the syllables belonged to the language, they must press the 'l' key on the keyboard. Participants were told to respond as quickly and accurately as possible.

Results

Proportions of 'majority rules' responses were recorded for each participant, shown in Figure 1. If participants learned a 'majority rules' pattern, this proportion should remain high for all test items. However, if participants learned a directional pattern, proportion of 'majority rules' responses should be above chance for Old and New test items, but below chance for New Direction Items.

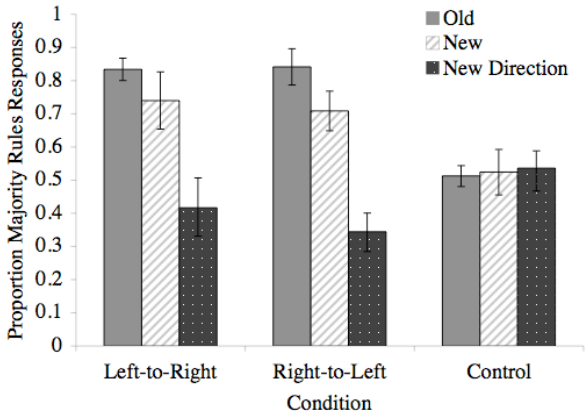


Figure 1: Experiment 1 Results

A 2 (Training) x 3 (Test Condition) mixed-design ANOVA compared each critical condition with the Control condition. There was a significant overall effect of Training for the Left-to-Right condition ($F(1, 14) = 8.90$, $p < 0.05$). There was an effect of Test Item ($F(2, 28) = 5.70$, $p < 0.01$), reflecting greater proportions of 'majority rules' responses in the performance in the Old ($F(1, 14) = 9.67$, $p < 0.01$) and New Test Items ($F(1,14) = 4.95$, $p < 0.05$) compared to the New Direction Test Items. There was a significant interaction ($F(2, 28) = 9.78$, $p < 0.01$), reflecting the fact that there were more 'majority rules' responses for Old Items ($t(14) = 5.29$, $p < 0.001$) but a trend of fewer 'majority rules' responses in for New Direction Items ($t(14) = 2.11$, $p = 0.073$).

There was also a significant overall effect of Training for the Right-to-Left condition ($F(1, 14) = 5.72$; $p < 0.05$). There was an effect of Test Item ($F(2, 28) = 5.04$, $p < 0.05$), reflecting greater proportions of 'majority rules' responses in the performance in the Old ($F(1, 14) = 11.24$, $p < 0.01$) compared to the New Direction Test Items. There was a significant interaction ($F(2, 28) = 7.87$, $p < 0.01$), reflecting the fact that there were more 'majority rules' responses in the Right-to-Left condition for Old Items ($t(14) = 7.43$, $p < 0.001$) but a trend of fewer 'majority rules' responses for New Direction Items ($t(14) = 2.11$, $p = 0.053$).

To test whether participants inferred a directional rule versus a 'majority rules' pattern, we performed contrasts comparing the New Direction test condition to the Old and New items respectively. In the Left-to-Right Condition, there was a significant difference between the New Direction and both the Old ($F(1, 7) = 17.07$, $p < 0.01$) and New ($F(1, 7) = 10.13$, $p < 0.05$) test items. The Right-to-Left condition also showed a significant difference between New Direction and Old items ($F(1, 7) = 17.49$; $p < 0.01$) and a marginally significant difference between the New Items ($F(1, 7) = 5.20$; $p < 0.08$) test conditions. The fact that

participants chose the 'majority rules' items significantly less often in the New Direction test condition (compared to Old and New items) suggests that learners inferred a directional pattern rather than a 'majority rules' pattern, reflecting a bias against 'majority rules'[2].

Among the 16 participants in the Experiment 1, only three chose the 'majority rules' item in the New Direction Condition greater than 60% of the time, while three chose the 'majority rules' item 50% of the time, and nine chose the 'majority rules' item less than 50% of the time.

Discussion

The results of Experiment 1 suggest that participants inferred a directional harmony pattern over a 'majority rules' harmony pattern. When learners were exposed to a spreading process that was ambiguous between a 'majority rules' pattern and a directional spreading pattern, learners inferred a directional pattern. This suggests that the non-existence of 'majority rules' spreading processes across the world's languages is in part due to learning biases. Learners do not postulate 'majority rules' languages because they are biased towards directional spreading processes.

However, it is unclear whether this bias is shaped by language-specific constraints or more general cognitive principles, such as attention and memory. Learners may not infer 'majority rules' because such languages require the language user to keep track of the number of vowels of a particular feature value in the input, inducing a greater memory load. Further, there may be a bias in favor of directional patterns, which are in line with attentional biases. For example, in a left-to-right language, it is fully predictable which vowel triggers harmony (the left-most vowel) and which vowels undergo harmony (the right-most vowels). Learners may be biased to infer a directional pattern, given that the consistent cues for harmony are found at the attention-heavy locations in the word (Beckman, 1998). Additionally, 'majority rules' patterns require the learner to keep track of a wider range of conditioning factors: how many vowels of each feature value are in the input, and which direction of spreading to use when there is a tie. A 'majority rules' pattern may require more episodic memory because several different situations in the input induce very different results. For example, two round vowels and one unround vowel will yield round vowels, but three round vowels and four unround vowels will yield unround vowels. While complicated phonological patterns are not uncommon cross-linguistically, if a learner has to decide between a simpler directional pattern and a complicated 'majority rules' pattern, they should choose the directional pattern.

One way to determine whether the directionality preference is due to non-linguistic factors against 'majority rules' is to replicate Experiment 1 with non-linguistic stimuli. If learners of a non-linguistic pattern follow the same constraints on 'majority rules', then it is likely that the bias found in these experiments is due to non-linguistic factors, but if no bias is found in non-linguistic stimuli, it suggests that there is something about the linguistic nature of harmony that biases learners towards directional spreading. Experiment 2 addresses this question with a visual analogue of Experiment 1.

[2] We also found a significant effect when the alternations were presented as changes from a disharmonic word (as opposed to a concatenation of mono-syllabic words) (Finley & Badecker, 2008).

Experiment 2

Experiment 2 addresses whether the bias against a 'majority rules' found in Experiment 1 may be reflected in a non-linguistic version of the vowel harmony learning task.

Methods

Participants All participants were adult native English speakers with no knowledge of a vowel harmony language, and did not participate in Experiment 1. Twenty-seven University of Rochester undergraduates participated for $10. Participants were randomly assigned to one of three training conditions: Control, Right-to-Left and Left-to-Right.

Design The optimal way to test for the effects of non-linguistic constraints on pattern learning is to design a pattern that makes use of known categories, but does not make use of any linguistic strategies. For this reason, a visual learning pattern using colors and shapes is optimal. First, shapes and colors are categories that are readily available to the adult learner, making it possible for the participant to infer a spreading pattern based on the experimenter-defined parameters. Second, the visual stimuli are completely outside the range of linguistic input to the learner, making the pattern learning task as non-linguistic as possible. While non-linguistic auditory stimuli present a closer match to the language learning task, there are two potential problems with such a design. First, standard non-linguistic auditory pattern learning makes use of tones or uncommon sounds that are not clearly defined categories. Thus, it is not clear whether learners of a tone-spreading pattern would make use of the same experimenter-defined categories. The visual stimuli that were chosen for this experiment have definitive categories: shapes (circles and squares) and colors (red, green, blue, yellow). In the present experiment, squares and circles of various colors assimilated based on a spread-right pattern or a spread-left pattern. Second, non-linguistic auditory pattern learning may invoke linguistic strategies to learning (e.g., acoustic properties of the sounds), and therefore may not directly address the questions posed in the present experiment.

It is important to note that the directional labels (left-to-right) are figurative for both Experiments 1 and 2. Left refers to the first item heard/seen; right refers to the final item heard/seen. In the visual analogue, all items appeared sequentially in the center of the monitor for 500ms.

Each input-output pair was presented as a series of three shapes followed by the assimilated version of those three

shapes. Each shape was flashed on the screen for 500ms followed by a 100ms pause in the center of the screen. A 500ms pause was placed between each series of 3 shapes. For example, participants in the Left-to-Right condition, saw /RED SQUARE, BLUE SQUARE, GREEN CIRCLE/ → [RED SQUARE, BLUE SQUARE, GREEN SQUARE]. Participants in the Right-to-Left condition, participants saw /RED CIRCLE, BLUE SQUARE, GREEN SQUARE/ → [RED SQUARE, BLUE SQUARE, GREEN SQUARE].

The training and test items were analogous to the items in Experiment 1. There were 24 training pairs, repeated 5 times each in a random order. There were 12 items each in three test conditions: Old, New and New Direction.

Table 3: Training Items for Experiment 2

Left-to-Right	Right-to-Left
SQUARE SQUARE CIRCLE → SQUARE SQUARE SQUARE	CIRCLE SQUARE SQUARE → SQUARE SQUARE SQUARE
CIRCLE CIRCLE SQUARE → CIRCLE CIRCLE CIRCLE	SQUARE CIRCLE CIRCLE → CIRCLE CIRCLE CIRCLE

Stimuli Shape stimuli were produced using the standard drawing tools for Microsoft Power Point. The shapes consisted of a square and a circle each for four different colors: red, green, blue and yellow, with a small amount of grey shading around each shape. All shapes were standardized to be the same size on the screen (occupying a 5in x 5in space in the center of the monitor).

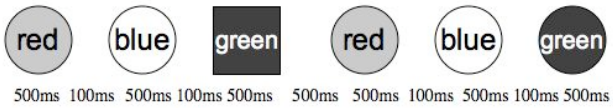


Figure 2: Experiment 2 Stimuli (Left-to-Right)[3]

Procedure The procedure was identical to Experiment 1 except that participants were told that they would be watching a series of shapes, presented as a series of pairs of three shapes.

Results

The proportions of 'majority rules' responses were recorded for each participant, shown in Figure 3. A 2 (Training) x 3 (Test Condition) mixed-design ANOVA compared each critical conditions with the Control condition. There was a significant effect of Training for the Left-to-Right condition ($F(1, 14) = 9.83$, $p < 0.01$). There was a significant interaction ($F(2,32) = 7.28$, $p < 0.01$), due to the fact that there was a significant difference between the Controls for New Items ($t(16) = 2.59$, $p<0.05$), but not New Direction items ($t(16) < 1$). This suggests that learners did not distinguish 'majority rules' and directional items. This is confirmed by a significant effect of Test Item ($F(2, 32 = 10.94$, $p < 0.001$), as there was a significant difference between the New Direction items and both the Old and New Items combined ($F(1,16)=16.57$, $p < 0.01$), suggesting that learners did not infer a 'majority rules' pattern.

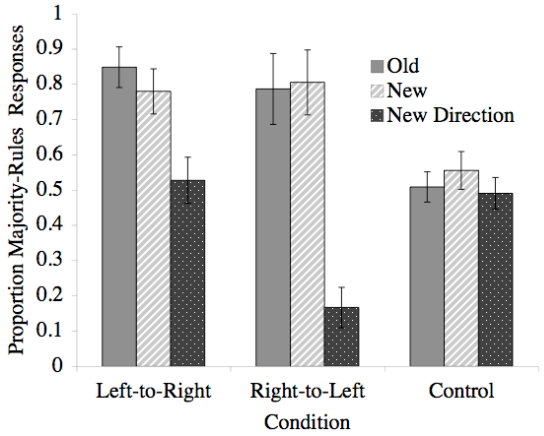


Figure 3: Experiment 2 Results

There was no significant effect of Training for the Right-to-Left condition ($F(1, 16) = 1.63$, $p <0.05$). This was carried the interaction between condition and test item ($F(2,32) = 13.48$, $p <0.001$). There were significantly more 'majority rules' responses compared to Controls for Old ($t(16) = 2.55$, $p < 0.05$) and New ($t(16) = 2.33$, $p < 0.05$) but there were significantly fewer 'majority rules' responses for New Direction items compared to the Control condition ($t(16) = -4.41$, $p < 0.001$). This difference reflects the fact that participants in the Right-to-Left condition inferred a directional pattern over a 'majority rules' pattern. The fact that there was no overall significant difference between the Right-to-Left condition and the controls is reflected in the low 'majority rules' responses in the New Direction condition, creating an overall average that was not different from the overall average of the Control condition. There was a significant effect of Test Item ($F(2, 32) = 17.66$, $p < 0.001$), due to the fact that there was a significant difference between the New Direction items and both the Old and New Items combined ($F(1,16)=19.90$, $p < 0.001$).

Participants in the Right-to-Left condition learned a directional harmony pattern, while participants in the Left-to-Right condition had no preference. This difference is reflected in the New Direction items, as participants in the Left-to-Right condition chose the majority option significantly more often than participants in the Right-to-Left condition ($t(16) = 4.16$, $p < 0.01$).

Among the nine participants in the Left-to-Right condition, four participants chose the 'majority rules' item in the New Direction Condition between 40 and 50% of the time, while two chose the 'majority rules' item 25% of the time, and three chose the 'majority rules' item greater than 60% of the time. This variation suggests that there is no intrinsic strategy towards 'majority rules'.

[3] All items were presented in the center of the screen.

Discussion

The difference between the Right-to-Left and Left-to-Right conditions suggests that visual pattern stimuli are processed differently depending on whether the change occurs first in the sequence or last in the sequence. This difference may be due to attentional constraints. If learners pay the most attention to the first part of the sequence, learners in the Left-to-Right condition will notice that there is a change in the first shape, but learners in the Right-to-Left condition will have to wait for the entire three shapes in order to see what changes. Thus, their representation of the pattern may be more holistic, and thus may be more amenable to both 'majority rules' and directional responses. Another possibility is that learners in Experiment 2 were influenced by their prior reading experience, which was left-to-right. This predicts that the opposite pattern should emerge for learners whose reading system is right-to-left. Future research will address these questions.

Because we used namable categories (shapes and colors), it is possible that participants engaged in naming the shape patterns as they appeared on the screen (e.g., 'GREEN SQUARE', 'RED CIRCLE', etc). However, this type of naming is different from the grammatical process that applies in a phonological pattern. First, phonological rule processing is less likely to involve naming (e.g., 'round vowel' or 'u'). Second, if naming the non-linguistic objects induced linguistic processing, we would expect an exact replication of Experiment 1, but this did not occur. In order to replicate a harmony process, it is necessary to use non-linguistic stimuli that have clear categories. Because all stimuli that are a priori categorical have a name, it is not possible to use non-linguistic stimuli that are not namable. Further, participants often create names for non-namable stimuli (e.g., 'the squiggly one') making it unclear if non-namable stimuli would remove naming strategies.

General Discussion and Conclusions

The results of Experiment 1 provided evidence in favor of a learning bias that favors directionality over 'majority rules' patterns. This bias towards directional harmony patterns provides insight into why 'majority rules' patterns do not exist in natural language. If learners are not biased to infer 'majority rules' from their language data, it is unlikely that such a pattern would emerge.

Experiment 2 demonstrated that the attentional constraints that may lead to a bias towards directional spreading pattern must work differently for spoken language versus non-linguistic visual stimuli. In this non-linguistic analogue of Experiment 1, participants only inferred a directional pattern when the spreading pattern occurred from right-to-left, affecting the first image. These results suggest that the source of the learning bias for directional patterns occur as an interaction of the ways in which speakers attend to auditory spoken language. One possibility is that linguistic material is continuous in a way that non-linguistic material is not. This continuity may make listeners more likely to attend to both beginnings and ends of words.

The experiments presented in this paper support the hypothesis that learners have biases that shape the distribution of patterns cross-linguistically. While 'majority rules' spreading patterns may be easily generated by rule and constraint-based theories of phonology, such spreading patterns violate constraints on attention, perception and memory. These constraints bias the learner towards directional spreading patterns over 'majority rules' patterns. In many ways, these biases hold for both linguistic and non-linguistic stimuli, suggesting that domain general constraints may affect the distribution of linguistic patterns across the world's languages.

Acknowledgments

The authors would like to thank audiences at 2008 WCCFL Conference as well as: Iris Berent, Paul Smolensky, Colin Wilson, Peter Graff, Neil Bardhan, Patricia Reeder, Luigi Burzio, and our anonymous reviewers for their helpful insights and suggestions. Funding was provided by NSF IGERT and NIH Grant DC00167.

References

Anderson, S. (1981). Why phonology isn't "natural". Linguistic Inquiry, 12, 493-547.

Bakovic, E. (2000). Harmony, dominance and control. Unpublished doctoral dissertation, Rutgers University.

Beckman, J. M. (1998). Positional faithfulness. Unpublished doctoral dissertation, UMass Amherst.

Boersma, P., & Weenink. (2005). Praat: Doing phonetics by computer.

Clements, G. N., & Sezer, E. (1982). Vowel and consonant disharmony in Turkish. In v. d. H. a. Smith (Ed.), The structure of Phonological Representations (Vol. II, pp. 213-255). Dordrecht: Foris.

Cohen, J. D., MacWhinney, B., Flatt, M., & Provost, J. (1993). PsyScope: A new graphic interactive environment for designing psychology experiments. Behavioral Research Methods, Instruments and Computers, 25, 257-271.

Finley, S., & Badecker, W. (2008). Analytic biases for vowel harmony languages. WCCFL, 27, 168-176.

Jusczyk, P., Smolensky, P., & Alloco, T. (2002). How English-learning infants respond to markedness and faithfulness constraints. Language Acquisition, 10, 31-73.

Korn, D. (1969). Types of labial vowel harmony in the Turkic languages. Anthropological Lingusitics, 11, 98-106.

Slobin, D. I. (1973). Cognitive prerequisites from the development of grammar. In C. A. Ferguson & D. I. Slobin (Eds.), Studies of child language development. New York: Holf, Rinehart & Winston.

Structural Constraints and Real-World Plausibility in Analogical Inference

Linsey A. Smith (linsey@u.northwestern.edu)
Dedre Gentner (gentner@northwestern.edu)
Department of Psychology, Northwestern University
2029 Sheridan Road, Evanston, IL 60208 USA

Abstract

Theoretical accounts of analogy have largely agreed that structural constraints play a substantial role in the mapping process. Less is known, however, about the robustness of these constraints in the inference process and the way in which particular content influences the use of structural constraints in analogical inference. We conducted two studies testing whether the plausibility (or implausibility) of an inference influences adherence to general structural principles in analogical reasoning. We found substantial reliance on the predicted structural constraints, but also an influence of the plausibility of the inference.

Introduction

Our goal in this research is to explore the stability of analogical inference under different conditions: specifically, whether analogical inference is a domain-general reasoning process, governed by structural constraints inherent to the analogical process, or whether it is a loosely constrained process whose outcome is strongly influenced by the plausibility of the potential inferences in particular domains. This question is important not only for what it can tell us about basic analogy processes, but also because the use of analogy in scientific discovery (and even in science learning) sometimes requires making initially implausible inferences. We first review research on this issue in the arena of analogical mapping and alignment, which has been extensively studied, and then turn to analogical inference.

Structural Constraints on Analogical Mapping

Reasoning by analogy involves identifying a common system of relations between two domains and generating further inferences driven by these commonalities (Gentner, 1983; Holyoak & Thagard, 1989; Hummel & Holyoak, 1997; Kokinov & French, 2003). According to structure-mapping theory, the comparison process involves aligning a pair in such as way as to achieve a consistent structural alignment between two domains (Falkenhainer, Forbus & Gentner, 1989; Gentner, 1983; Gentner & Markman, 1997). The structural alignment process is guided by a set of tacit constraints that lead to structural consistency and inferential clarity: *parallel connectivity,* which requires that arguments of matching predicates must also be placed into correspondence; and *one-to-one correspondence,* which requires that each element of a representation match, at most, one element of the other representation. Importantly, deep matching systems are preferred over shallow matches (the *systematicity principle),* which reflects a preference for coherence and inductive power in analogical processing (Clement & Gentner, 1991; Falkenhainer, Forbus & Gentner, 1989). Candidate inferences are generated by completing the pattern in the (initially) less-structured member of the pair, based on the common structure.

Models of analogy have largely converged on a set of assumptions like those outlined above (Falkenhainer, Forbus & Gentner, 1989; Gentner, Holyoak & Kokinov, 2001; Holyoak and Thagard, 1989; Hummel & Holyoak, 1997; Kokinov & French, 2003; Larkey & Love, 2003). Further, there is substantial empirical evidence in support of the idea that analogical reasoning obeys these constraints. A variety of studies have provided evidence that analogical matching is constrained by both structural consistency (including one-to-one mapping) (e.g., Krawczyk, Holyoak, & Hummel, 2005; Markman, 1997; Markman & Gentner, 1993; Spellman & Holyoak, 1992) and systematicity (e.g., Clement & Gentner, 1991). For example, Clement and Gentner (1991) showed participants analogous scenarios and asked them to judge which of two lower-order assertions shared by the base and target was most important to the match. Participants chose the assertion that was connected to matching causal antecedents – their choice was based not only on the goodness of the local match, but also on whether it was connected to the larger matching system. Thus, matching lower-order relations that are interconnected by higher-order relations were considered more important to the analogy. In sum, people demonstrate considerable structural sensitivity in analogical mapping.

Analogical Inference

There is some research on the degree to which structural constraints hold in analogical inference. In the Clement and Gentner (1991) research just described, a second study found evidence for systematicity in inference projection. People generated inferences that were part of a shared system, rather than equally applicable inferences that were not. Markman (1997) also found evidence for systematicity in inference generation. In addition, he found that people based their inferences on one-to-one mappings. When given analogies with two possible sets of correspondences, people noticed both possibilities, but drew inferences from only one of them. These findings suggest a role for structural consistency in inference, as in alignment.

However, one question that is largely unexplored is the degree to which the analogical inference process is influenced by the factual plausibility of the inference in the target. That is, are people able to track structural consistency despite implausibility in making inferences? The studies described above did not involve wide variations in plausibility, so they do not answer this question. Work by Keane (1996) does bear on this issue. He found that people readily accepted inferences that were both highly plausible [had high "entity utility"] and easy to place into correspondence with the target ["entity parallelism']—that is, highly *adaptable*—compared to those inferences that were less adaptable. These findings suggest that plausibility in the target is important in analogical inference. However, the question remains open as to what people will do if structural consistency directly conflicts with target plausibility.

Another way to put this question is, are there *content effects* in analogical inference? The issue of content effects has been investigated extensively in the research on deductive reasoning. Deductive reasoning has traditionally been considered a relatively rigorous, principle-governed process, although empirical support for this claim (e.g., Marcus & Rips, 1979) is punctuated by many observations that show that people's judgments about the logical validity of deductive arguments is influenced by the 1) specific content that is being reasoned about (e.g., Cheng & Holyoak, 1985; Cummins, Lubart, Alksinis, & Rist, 1991; Rips, 2001; Thompson, 1994), and 2) whether the reasoner agrees with the premises and conclusions of the argument (e.g., Markovits, 1995; Newstead, Pollard, Evans, & Allen, 1992). Thus, there is evidence that logical reasoning is swayed by particular content.

a. *Logically valid, real-world plausible:*
If Fred sprinkles water on wood shavings, the shavings get wet.
Fred sprinkles water on wood shavings.
The shavings get wet.

b. *Logically invalid, real-world plausible:*
Fred sprinkles water on wood shavings.
The shavings get wet.

For example, Rips (2001) asked participants to evaluate arguments like (a) and (b) in which the plausible conclusion was either logically valid or invalid. The question was whether people could track deductive logic regardless of the plausibility of the conclusion. A substantial number of participants (mistakenly) identified invalid arguments as logically correct when they were plausible. Overall, Rips's (2001) findings suggest that people were largely able to maintain logical rigor under the strain of real-world implausibility, but that logical rigor was sometimes compromised by the content of the arguments: people could not wholly divorce logical form from content in this task.

A parallel question can be asked about analogical inference: can people maintain structural consistency despite real-world implausibility in making analogical inferences (which we will refer to as *analogical rigor*)? Our question in this paper is what happens when the structural alignment process leads to inferences that the reasoner considers implausible. On the one hand, some prior research shows reliable effects of structural consistency on inference (Clement & Gentner, 1991; Markman, 1997). On the other hand, these studies (and Keane's (1996) study) did not directly pit structural consistency against plausibility. And unlike deductive reasoning, analogical reasoning is generally not explicitly taught. Thus we might expect people to be less committed to maintaining analogical rigor than they are to maintaining logical rigor.

The Current Experiments

In this set of studies, we asked participants to evaluate analogies where the inferences derived from the structure-mapping process are at odds with the real-world plausibility of the inferences. This method allowed us to identify how much people rely on domain-specific content over general mapping principles in analogical inference.

For the task, we adapted the deductive reasoning task from Rips (2001). As discussed above, in that experiment, participants evaluated the validity of conclusions from arguments that orthogonally varied in logical validity and real-world plausibility. His study assessed whether people would follow deductive logic in drawing conclusions even when these conclusions conflicted with plausibility. In this research, we posed the parallel question for analogical inference, that is, would people respect the structural constraints of analogy in drawing inferences even when these inferences conflicted with real-world plausibility. To put it another way, are people able to maintain analogical rigor in the face of real-world implausibility? We asked participants to assess whether a particular inference followed from an analogy. We created materials whose inferences varied in *structural consistency,* that is, we varied whether the inference was a structurally consistent completion of the analogy. Table 1 shows an example set. The inferences in (a) and (b) are structurally consistent and those in (c) and (d) are structurally inconsistent. The pairs also varied orthogonally in *real-world plausibility*, with (a) and (c) having plausible inferences and (b) and (d) having implausible inferences. Participants might find analogies (b) and (d) (both implausible inferences) to be odd or downright wrong, but this is precisely the point: when an analogical inference conflicts with reasoners' knowledge, the question is whether they can identify inferences that the analogy must structurally yield, without being swayed by the plausibility of those inferences.

Of course, the ultimate evaluation of an analogical inference is not solely contingent on structural consistency, but also involves checking the factual validity of the inference (and in a real problem solving situation, the contextual relevance) (Gentner & Clement, 1988; Holyoak & Thagard, 1989). To this end, we also asked participants to provide ratings of the *overall goodness* of each analogy. We

Table 1: Sample materials from Experiment 1.

Base (constant)
Mary has built a sandcastle. Her younger brother comes by and kicks the base of the castle. The sandcastle crumbles.

Target (four versions)
a. Structurally consistent, factually plausible
A wrecking ball knocks into a building's foundation.
Conclusion: **The building comes crashing to the ground.**

b. Structurally consistent, factually implausible
A tennis ball knocks into a building's foundation.
Conclusion: **The building comes crashing to the ground.**

c. Structurally inconsistent, factually plausible
A tennis ball knocks into a building's foundation.
Conclusion: **The building stays standing.**

d. Structurally inconsistent, factually implausible
A wrecking ball knocks into a building's foundation.
Conclusion: **The building stays standing.**

had two goals with this question. First, for implausible inferences, this question would give participants a way to indicate that they considered some analogies to be quite poor. We hoped that this would leave them more free to judge structural consistency on its own. Second, a more direct goal was to discover whether participants would incorporate both structural consistency and real-world plausibility into their judgments, as we expected they would. If so, we would expect only analogies that yield structurally consistent and plausible inferences to receive high overall goodness ratings.

Experiment 1

Method

Participants 19 Northwestern University undergraduates took part in the study individually or in small groups of up to four people. Participants completed the task in 10-15 minutes and for their time they received credit towards a course requirement or monetary compensation.

Procedure and Materials The experimenter gave one task booklet to the participant, and upon completion they returned the booklet to the experimenter. The booklet contained a page of instructions, followed by eight analogies (one per page). The analogies came from quartets of items, as in Table 1, that varied in structural consistency and real-world plausibility. We assigned each participant eight analogies, two of each type (structurally consistent and real-world plausible, structurally consistent and implausible, structurally inconsistent and plausible, structurally inconsistent and implausible), as in Table 1. For an individual participant, however, different content instantiated each of these arguments. Thus, for example, no participant received more than one pair from the Table 1 quartet. The order of the problems in the test booklet was pseudo-randomized into four orders.

Measures Participants rated their agreement with the statement "The conclusion follows directly from the analogy." Responses were measured on a 7-point Likert scale, ranging from 1 (strongly disagree) to 7 (strongly agree). To facilitate analysis, responses were recoded into a dichotomous variable (with responses > 4 recoded as "Yes, the conclusion follows" and $\leq$ 4 recoded as "No, the conclusion does not follow"). The proportion of "Yes" responses for each type of stimuli was the measure of interest, and these were aggregated within conditions to form a measure of *inference acceptance rates,* which we'll simply refer to as *acceptance rates.* To the extent that participants strongly differentiate structurally consistent from inconsistent inferences, such that structurally consistent inferences have high acceptance rates and structurally inconsistent inferences have low acceptance rates, this measure will approximate analogical rigor.

In addition participants were asked to judge the *overall goodness* of each analogy. Participants rated their agreement with the statement "Overall, this is a good analogy." Responses were measured on a 7-point Likert scale, ranging from 1 (strongly disagree) to 7 (strongly agree).

Results

Figure 1 presents the inference acceptance rates for each of the four types of stimuli. The data were analyzed with a two-way ANOVA, with structural consistency and real-world plausibility as within-subjects factors.

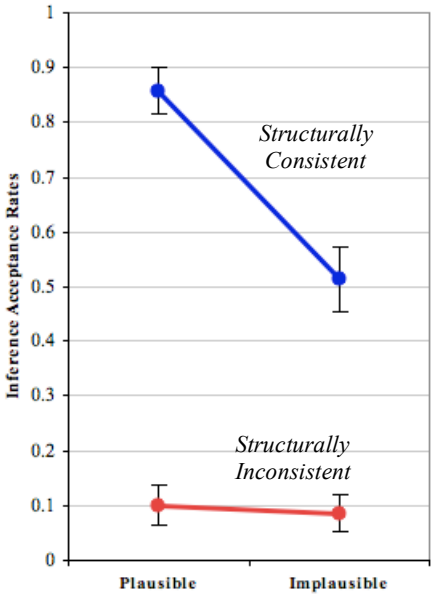


Figure 1: Inference acceptance ratings for Experiment 1. Error bars reflect the standard error.

Overall, there was a strong effect of structural consistency on acceptance rates, $F(1,37) = 110.87$, $p < .001$, $\eta^2 = .38$; people were far more likely to accept structurally consistent inferences (M=.63, SD=.49) than structurally inconsistent inferences (M=.09, SD=.29). There was also a main effect of real-world plausibility on acceptance ratings, $F(1,37) = 8.74$, $p < .01$, $\eta^2 = .05$; a greater proportion of plausible inferences was judged as following from the analogy (M=.45, SD=.50) than implausible inferences (M=.30, SD=.46). The effect size for real-world plausibility was considerably smaller ($\eta^2 = .05$) than that for structural consistency ($\eta^2 = .38$).

There was also a significant interaction between structural consistency and plausibility, $F(1,37) = 27.89$, $p < .001$, $\eta^2 = .10$. For structurally consistent analogies, participants were less likely to judge implausible inferences as following from the analogy (implausible: M=.50, SD=.51; plausible: M=.89, SD=.31), $t(37) = 4.09$, $p < .001$. No such difference was obtained for structurally inconsistent analogies.

We reserve the analysis of overall goodness judgments until after we present Experiment 2.

Discussion

Our primary question is whether people can maintain analogical rigor in the face of real-world implausibility. We found fairly good support for this possibility. Acceptance ratings were higher overall for structurally consistent analogies, indicating that people are able to track the structural consistency of an inference regardless of the plausibility of that inference. Additional support for this claim comes from the observed effect sizes: structural consistency explains 38% of the overall variance on inference acceptance rates, whereas real-world plausibility only accounts for 5% of the variance. However, analogical rigor is also influenced by particular content. Specifically, participants were more likely to reject structurally consistent inferences when they were implausible. If individuals had been entirely rigorous, we would not have expected to see this difference between plausible and implausible conditions. Interestingly, this effect of plausibility did not appear for structurally inconsistent inferences, which were uniformly rejected.

In short, the results so far suggest that people are able to abide by structural constraints when making inferences; however, conflicting content can influence whether people maintain these constraints. In the next study, we sought to identify whether clarifying the instructions would attenuate these content effects.

Experiment 2

This study tested whether more explicit instructions would lead participants to more strictly observe analogical constraints. We used the same basic method as Experiment 1, with one important modification: we re-wrote the question to clarify that the focus should be on what follows from the analogy.

Method

Participants 19 Northwestern University undergraduates took part in the study individually or in small groups of up to four people. Participants completed the task in 10-15 minutes and for their time they received credit towards a course requirement or monetary compensation.

Materials and Measures The materials for the analogy task were the same, except that the question used to elicit inference acceptance ratings was modified from rating agreement with the statement "The conclusion follows from the analogy?" to instead read "The conclusion in Situation 2 would **necessarily follow** if Situations 1 and 2 were truly analogous, regardless of whether the conclusion *could* be true or not." Participants were then asked to circle "Yes" or "No." The proportion of "Yes" responses for each type of stimuli was the dependent measure, and these were aggregated within conditions to form a measure of *inference acceptance rates.* The *overall goodness* question remained the same. The procedure was as in Experiment 1.

Results

The results showed a strong effect of structural consistency; structurally consistent inferences had higher acceptance rates (M=.91, SD=.29) than did structurally inconsistent inferences (M=.12, SD=.33). Figure 2 shows the inference acceptance rates for each of the four types of stimuli. For ease of comparison, the results from Experiment 1 (dotted lines) have also been included. Analysis entailed a two-way within-subjects ANOVA, with structural consistency and real-world plausibility as within-subjects factors.

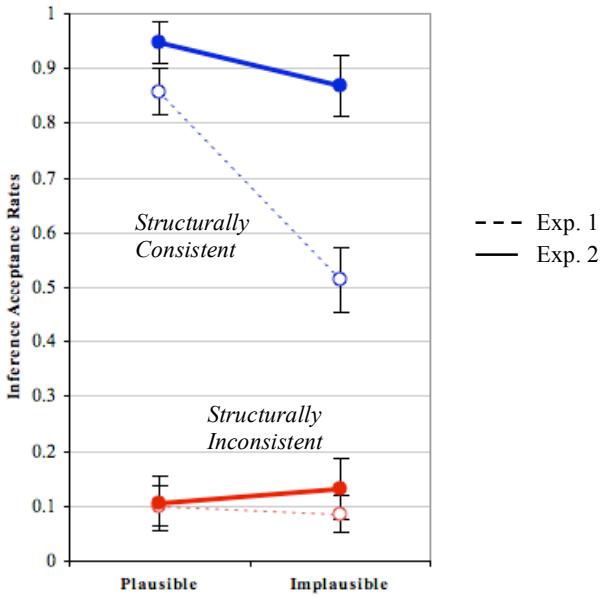


Figure 2: Inference acceptance ratings for Exp. 1 (dotted line) and Exp. 2 (solid), divided into structurally consistent and inconsistent. Error bars reflect the standard error.

As in Experiment 1, there was a main effect of structural consistency, $F(1,37) = 311.22$, $p < .001$, $\eta^2 = .71$. Real-world plausibility no longer influenced inference acceptance: there was no main effect of real-world plausibility nor an interaction between the factors (real-world plausible: M=.53, SD=.50; implausible: M=.50, SD=.50).

Cross-Experiment Analysis To further test whether more explicit instructions to focus solely on whether an inference follows from the analogy bolstered participants' focus on structural constraints, we entered Experiments 1 and 2 into a three-way mixed ANOVA, adding in instruction type (i.e., Experiment 1 or 2) as a between-subjects factor. In addition to the main effects of structural consistency and real-world plausibility, there was also a main effect of instruction type, $F(1,74) = 6.26$, $p < .05$. These main effects were qualified by a significant three-way interaction between all three variables, $F(1,74) = 5.31$, $p < .05$. This significant interaction is due to different patterns of performance on structurally consistent inferences: in the explicit instructions condition (Experiment 2), there was no difference in acceptance rates between plausible and implausible inferences, but in the implicit instructions condition (Experiment 1), acceptance rates were higher for plausible inferences, $t(37) = 4.09$, $p < .001$.

Judgments of overall goodness We elicited judgments of overall goodness for the analogies to identify participants' overall impression of the analogy, which may not have been captured in the acceptance rates, especially in the case of implausible inferences. To identify whether judgments of overall goodness for the analogies varied by instruction type, we entered both experiments into a three-way mixed ANOVA, with *overall goodness* as the dependent measure. There was only a marginally nonsignificant effect of instruction type, $F(1,74) = 3.33$, $p = .07$; participants rated overall goodness similarly across both instruction conditions. There were main effects of both structural consistency ($F(1,74) = 97.35$, $p < .001$, $\eta^2 = .27$) and real-world plausibility ($F(1,74) = 28.43$, $p < .001$, $\eta^2 = .06$), which were qualified by a significant interaction between the two, $F(1,74) = 43.02$, $p < .001$, $\eta^2 = .11$. Structurally inconsistent pairs were given low overall ratings that did not vary by real-world plausibility (max = 7, plausible: M=1.92, SD=1.16; implausible: M=2.20, SD=1.77); structurally consistent pairs that were plausible were given higher ratings than implausible pairs (plausible: M=5.05, SD=1.52; implausible, M=2.91, SD=1.86), $t(75) = 8.25$, $p < .001$. This pattern of goodness ratings partly mirrors the pattern of inference acceptance ratings in Experiment 1: there was an effect of both structural consistency and plausibility, with a stronger effect of structural consistency; and structurally consistent analogies were rated lower when their inferences were implausible. Thus, with the exception of the Experiment 2 acceptance ratings, the deviation from analogically rigorous behavior occurs only for structurally consistent but implausible analogies.

Discussion

Our primary question in Experiment 2 was whether people are capable of separating structural consistency from real-world plausibility when explicitly told to do so. The results indicate that the answer is yes: people were able to ignore the real-world plausibility of analogical inferences in making their judgments.

General Discussion

Two studies probed the robustness of structural constraints on analogical inference when challenged by the particular content of the inferences. In Experiment 1, we investigated whether people would follow the structural constraints of analogy in drawing inferences even when they conflicted with plausibility. Acceptance rates were higher for structurally consistent inferences than inconsistent inferences; overall, people can reliably follow structural consistency in inference. Plausibility did influence inference acceptance rates, but only for structurally consistent analogies. Structurally inconsistent inferences were noticed as such, regardless of their real-world plausibility. However, when people encountered potentially analogous (i.e., structurally consistent) inferences, their judgments were influenced by target plausibility.

Experiment 2 tested whether more explicit instructions would lead participants to make a clearer separation between analogical rigor and plausibility. The results indicate that this is indeed the case: participants no longer demonstrated content effects, but instead recognized inferences that followed from completing the common system, as predicted by structure-mapping and other current models of analogy (Falkenhainer, Forbus & Gentner, 1989; Holyoak and Thagard, 1989; Hummel & Holyoak, 1997; Kokinov & French, 2003). Understanding the conditions under which people will put aside their knowledge to work through an analogy has implications for educational contexts, where analogies are used extensively to promote knowledge acquisition and conceptual change (e.g., Richland, Holyoak, & Stigler, 2004). Importantly, the analogies used by instructors may require learners to make ostensibly implausible inferences (e.g., Clement, 1993).

In both experiments, we elicited judgments of overall goodness of the analogies. We found, as expected, that people considered both structural consistency and real-world plausibility in judging the analogies. Ratings for overall goodness did not vary as a function of instructions. In both experiments, people reliably indicated that only those analogies that were both structurally consistent and real-world plausible were good analogies. This pattern of judgments is in accord with the general assumption that while analogy may involve a mapping process guided by structural constraints, ultimate evaluation of the analogy involves checking the factual validity of projected inferences.

Although Experiment 1 demonstrates that analogical rigor is influenced by content, for both experiments, participants showed a general tendency to identify structurally consistent inferences as following from the analogy. Furthermore,

effect sizes were moderate for structural consistency, whereas they were extremely small for plausibility. Perhaps more tellingly, in judgments of overall goodness, the effect of structural consistency was much larger (η^2 =.27) than that of plausibility (η^2 =.06). Taken together, these observations suggest that people are relying heavily on structural principles to guide their evaluations of overall analogical goodness. The results of these experiments are consistent with the claim that analogical processing involves a structure-mapping process of alignment and inference largely governed by structural constraints.

One concern here is that the materials were too simple to engage serious content-based reasoning. It will be necessary to investigate a wider range of material to determine the whether the effects identified in these studies will generalize to more natural materials. However, the results so far suggest that analogical inference is to a large extent guided by a tacit set of structural constraints that may function something like the principles that guide deductive reasoning. In future studies it would be of interest to contrast these two reasoning tasks to see whether similar patterns emerge. Another future direction would be to obtain online measures, such as reading times, to investigate the time course of content effects in analogy and further explicate the interaction between mapping processes and target content in analogical inference.

Acknowledgments

This research was supported by ONR Grant N00014-08-1-0040. We thank Doug Medin, in addition to Julie Colhoun and the rest of the Cognition and Language Lab, for many helpful comments and suggestions on this work. We also thank Katherine James, Ted Bedell, Noel Dwyer, and Lindsey Zamarripa who helped with recruitment, scheduling, and data collection.

References

Cheng, P. W., & Holyoak, K. J. (1985). Pragmatic reasoning schemas. *Cognitive Psychology, 17*, 391–416.

Clement, J. (1993). Using bridging analogies and anchoring intuitions to deal with students' preconceptions in physics. *Journal of Research in Science Teaching, 30,* 1241-1257.

Clement, C. A., & Gentner, D. (1991). Systematicity as a selection constraint in analogical mapping. *Cognitive Science, 15*, 89-132.

Cummins, D. D., Lubart, T., Alksnis, O., & Rist, R. (1991) Conditional reasoning and causation. *Memory & Cognition, 19*(3), 274-282.

Falkenhainer, B., Forbus, K. D., & Gentner, D. (1989). The structure-mapping engine: Algorithm and examples. *Artificial Intelligence, 41,* 1–63.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science, 7*, 155–170.

Gentner, D., Holyoak, K. J., & Kokinov, B. (Eds.). (2001). *The analogical mind: Perspectives from cognitive science.* Cambridge, MA: MIT Press.

Gentner, D., & Markman, A. B. (1997). Structure mapping in analogy and similarity. *American Psychologist*, 52, 45-56.

Gentner, D., & Toupin, C. (1986). Systematicity and surface similarity in the development of analogy. *Cognitive Science*, 10, 277-300.

Holyoak, K. J. & Thagard, P. (1989). Analogical mapping by constraint satisfaction. *Cognitive Science, 13*, 295--355.

Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review, 104*, 427-466.

Keane, M.T. (1996). On adaptation in analogy: Tests of pragmatic importance and adaptability in analogical problem solving. *Quarterly Journal of Experimental Psychology, 46,* 1062-1085.

Kokinov, B., & French, R. M. (2003). Computational models of analogy making. In L. Nadel (Ed.), *Encyclopedia of Cognitive Science.* London: MacMillan.

Krawczyk, D. C., Holyoak, K. J., & Hummel, J. E. (2005). The one-to-one constraint in analogical mapping and inference. *Cognitive Science, 29*, 797 – 806.

Larkey, L. B., & Love, B. C. (2003). CAB: Connectionist Analogy Builder. *Cognitive Science, 27* (5), 781-794.

Marcus, S. L., & Rips, L. J. (1979). Conditional Reasoning. *Journal of Verbal Learning and Verbal Behavior*, 18(22), 199-223.

Markman, A. B. (1997). Constraints on analogical inference. *Cognitive Science, 21,* 373–418.

Markman, A. B., & Gentner, D. (1993). Splitting the differences: A structural alignment view of similarity. *Journal of Memory and Language*, *32*, 517-535.

Markovits, H. (1995). Conditional reasoning with false premises: Fantasy and information retrieval. *British Journal of Developmental Psychology, 13,* 1-11.

Newstead, S. E., Pollard, P., Evans, J. S., & Allen, J. L. (1992) The source of belief bias effects in syllogistic reasoning. *Cognition, 46*(1), 87-92.

Reeves, L. M., & Weisberg, R. W. (1994). The role of content and abstract information in analogical transfer. *Psychological Bulletin, 115*(3), 381-400.

Richland, L. E., Holyoak, K. J., & Stigler, J. W. (2004). The role of analogy in teaching middleschool mathematics. *Cognition and Instruction, 22*, 37-60.

Rips, L. J. (2001). Two kinds of reasoning. *Psychological Science, 12,* 129-134.

Spellman B. A., & Holyoak, K. J. (1992). If Saddam is Hitler then who is George Bush?: Analogical mapping between systems of social roles. *Journal of Personality and Social Psychology, 62*(6), 913-933.

Thompson, V. A. (1994). Interpretational factors in conditional reasoning. *Memory and Cognition*, *22*(6), 742–758.

Cross-Modal Influence on Binocular Rivalry

Joshua M. Lewis
Department of Cognitive Science
University of California, San Diego
San Diego, CA 92093
josh@cogsci.ucsd.edu

Adam S. Fouse
Department of Cognitive Science
University of California, San Diego
San Diego, CA 92093
afouse@cogsci.ucsd.edu

Virginia R. de Sa
Department of Cognitive Science
University of California, San Diego
San Diego, CA 92093
desa@cogsci.ucsd.edu

Abstract

Binocular rivalry occurs when two distinct stimuli, one for each eye, are presented to corresponding retinal areas. Similar to other bistable phenomena such as Necker cubes, this overlap often causes one's conscious perception to alternate between a coherent perception of one stimulus, a coherent perception of the other and sometimes a mixture of the two. Previous studies have tried to identify where rivalry occurs, and what is actually being rivaled. Some studies have provided evidence for low-level effects on rivalry, lending support to the idea that rivalry is between monocular visual streams. Other studies have provided evidence for higher-level effects on rivalry, supporting the idea that rivalry is between opposing patterns. While this debate has largely been passed on in favor of a hybrid theory of rivalry that includes effects at several levels, questions still remain about specific higher-level effects. In the present study, we look at the effect of a congruent auditory stimulus on perception of rival videos of speaking people. We find that auditory stimuli can have an effect on rivalry, indicating that cross-modal processes such as speech to lip matching or voice to face matching are among the high-level factors impacting rivalry.

Keywords: binocular rivalry; patchwork rivalry; stimulus rivalry; cross-modal; multi-modal; psychophysics.

In this paper, we investigate the role of cross modal interaction between audition and vision in determining stimulus dominance in a binocular rivalry paradigm. Binocular rivalry occurs when two distinct stimuli are presented to separate eyes, so that each eye only sees one stimulus but they overlap in one's visual field. Similar to other bistable phenomena such as Necker cubes, this overlap often causes one's conscious perception to alternate every few seconds between a coherent perception of one stimulus and a coherent perception of the other.

Historically, researchers have debated where in the visual processing stream one stimulus becomes dominant over the other and rises to conscious perception [1]. Evidence of ocular suppression has been found very early in the visual processing stream, at the Lateral Geniculate Nucleus and in V1 [2, 3, 4]. This finding supports the idea that rivalry is between monocular visual streams. On the other hand, high-level properties of the stimuli, such as visual coherence [5] and "natural" amplitude spectra [6], have been shown to affect rivalry dominance duration and strength, indicating that rivalry may be between the perceived stimulus rather than the monocular pathway. Further support for the stimulus being the object of rivalry comes from sudies that rely on interocular grouping during rivalry, in which cohesive stimuli can be perceived from parts that are divided between the eyes [7, 8]. These findings also indicate that areas of the brain further along in the visual processing stream are likely play a significant role in the phenomenon. Recent evidence from neuroimaging studies suggests that a complete answer for rivalry likely involves a hybrid of the two theories, involving both high-level and low-level visual processing systems [1].

While controversy over whether rivalry is controlled from low-level or high-level processing has largely been supplanted by an acknowledgment of the role of multiple levels of processing, questions remain about specific roles. Studies such as [4] have effectively answered the "how low?" question in the binocular rivalry literature, but the "how high?" question has remained more elusive. Attention is one potential candidate for a mechanism for bistable perception, as attention has been noted to have an effect on dominance of rival stimuli since Helmholtz [9] . Studies have shown that attention can control the rate of alternation between rival stimuli, but that selective attention showed stronger affects for ambiguous figures than for binocular rivalry [10]. However, the strength of effect on stimulus duration appears to depend on specific features of the stimuli, such as their complexity, and whether attention is focused on specific stimulus features [11, 12]. Attention seems to have the most effect on the initially dominant stimulus [13], and neurophysiological results indicate that attention can bias early processing in the visual stream [14].

Other higher-level effects on rivalry have been shown, in particular the importance of global coherence in pattern rivalry [15] and of biological motion in determining perception with both ambiguous monocular stimuli and rival binocular stimuli [16]. The biological motion result, in which upright walking figures were perceived more often than inverted figures, suggests a top-down effect where the global perception influences lower-level processing.

One interesting question is whether stimuli in another modality can influence rivalry. This question has been recently studied for bistable perception of visual and auditory

objects. Hupe and colleagues looked at perception of concurrently perceived bistable (but not binocularly rivaling) visual and auditory stimuli [17], particularly the temporal proximity of auditory and visual perception shifts during perception of the parallel bistable stimuli.

In the present study, we look at the question of whether simultaneously perceived auditory input can influence perception during binocular rivalry. We hypothesize that if a subject views two rivaling videos while listening to a soundtrack appropriate to only one of the videos, the video appropriate to the soundtrack will dominate perception for a greater period of time than the other video. There is considerable evidence that normal speech recognition involves both audition and vision. For example, the McGurk effect has shown that different articulations, as seen in a video of moving lips, can affect the perception of identical-sounding syllables [18], and many studies have suggested that speech perception is inherently multimodal (see [19] for a review). Recent studies have demonstrated sensitivity in speech recognition for matching between the gender of auditory and visual sources, lending support for the idea that cross-modal integration in speech recognition involves top-down processes [20]. Cross-modal matching is robust, with the ability to match a voice to lips that are represented only by point light sources [21].

Cross-modal experience has also been shown to affect performance in a visual-auditory temporal frequency matching task [22] where subjects were better able to match auditory and visual temporal repetition rates when the match was in the context of an upright point-light walker than for scrambled and inverted point-light walkers (with the same local motions). Auditory input has also been shown to influence visual perception of the number of flashed stimuli [23, 24] and visual input (color) has been shown to influence olfactory perception [25]. All of these effects are automatic, just as one cannot ignore the visual input when looking at it in the McGurk Effect. We reason that well-associated auditory input could similarly bias visual perception in a binocular rivalry paradigm.

Methods

We performed our experiment using StereoGraphics CrystalEyes LCD shutter goggles attached to a PC running Matlab and Psychophysics Toolbox 3.0. Our CRT monitor was configured to display stimuli intended for the left and right eyes on alternating refreshes, which were coordinated with the eye alternation of the shutter goggles via an emitter attached to the GeForce QuadroFX quad-buffered graphics card in the machine. We recruited 18 subjects, all undergraduates, 7 males and 11 females, with normal or corrected normal vision and no colorblindness. One subject was removed from the study due to incorrect performance on catch trials (described below), and another was removed because they only ever pressed one of the two responses, resulting in a grand total of 16 subjects.

Our stimuli were composed of four videos. All four videos showed head shots of volunteers relating a story about a recent experience. Two videos were clips of a story told by a male actor, and the other two videos were clips of a story told by a female actor. In order to make the videos easier to distinguish, we created both red and green versions of each video by converting the videos to grayscale and using the grayscale values as brightness on the red or green color channel. On our equipment, the green versions of the videos were noticeably brighter and we therefore reduced the brightness of the green videos at presentation time to 65% of their original brightness in order to better match them with the red videos. Note that we used shutter goggles, not red/green glasses, so the colors have no impact on which eye sees which stimulus, they just serve to aid discrimination and help group the patterns. The audio tracks from each video were separated so that video and audio media could be presented independently of one another.

We used a stimulus rivalry paradigm where we presented one eye with the left half of the male video and the right half of the female video and the other eye with the right half of the male video and the left half of the female video (see Figure 1). Stimulus rivalry is believed to occur higher in the visual processing stream than ocular rivalry [1], so we use stimulus rivalry in order to give ourselves the best chance to discover a high-level cross-modal effect. Our early pilot trials with standard eye rivalry (female video to one eye and male video to the other) did not reveal a cross-modal effect.[1]

Subjects viewed 25 trials, consisting of one warmup trial (not reported) and 24 trials generated from the following counterbalanced conditions: four possible combinations of male and female videos by male in red and female in green or vice versa by male soundtrack, female soundtrack or no soundtrack. Subjects indicated which video they felt they mostly perceived by pressing keys on the keyboard for female or male. If subjects were unsure of their perception, we instructed them to press both keys or press neither key and we considered either of those responses as identical. Each trial lasted 86.6 seconds and was followed by a short (approximately 5 second) catch trial where only one of the two videos was displayed.

The experiment was run in a darkened room with the subjects seated in front of the computer described above. A fixation cross was present in the center of the video, and we instructed subjects to stay focused on the cross as much as possible. The video itself was 640 x 480 pixels in the center of a 1024 x 768 display, with a black background. The response keys were the Z and / keys on a standard qwerty keyboard. We affixed glow-in-the-dark labels to the response keys to help subjects reorient if their hands got lost in the dark.

We performed two primary analyses on our data. For the purposes of both, a congruent response is a response indi-

[1] Just before the due date for the camera ready copy of this paper we discovered a poster at Vision Sciences Society Annual Meeting presented May 10, 2010 that did find that auditory congruent stimuli could bias binocular rivalry of line drawings presented to each eye [26].

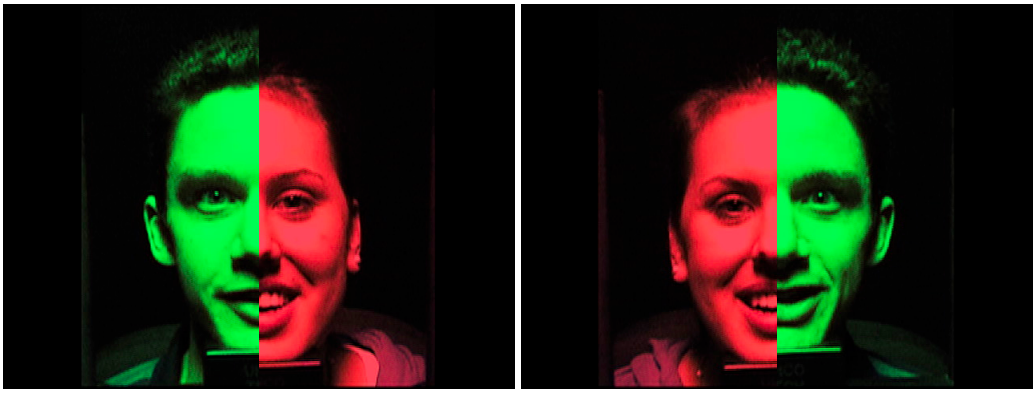


Figure 1: Sample stimulus from a single frame of the video. Left: left eye display. Right: right eye display.

cating dominance of the video associated with the currently playing audio and an incongruent response is a response indicating dominance of the video not associated with the audio. A neutral response is a response made during a trial with no audio. Our first analysis considered only the 16 trials that included sound. We subtracted the congruent dominance duration from the incongruent dominance duration and performed a positive one-tailed t-test comparison between the distribution over the subjects and a null distribution with zero mean (equal time spent on congruent and incongruent responses). Even though our trials were counterbalanced, we were concerned about two possible biases, a gender bias and a color bias. To correct for these biases we looked for a per-subject systematic bias in the no sound trials (previously unused) and subtracted the mean value of that bias from the appropriate responses in the trials with sound. We did this independently for both color and gender, resulting in three versions of this result: raw (uncorrected), corrected for color and corrected for gender. The equation for calculating this measure is as follows

$$\delta_{ci}(s) = \sum_{i \in S} \left(\sum_{j \in C_{is}} R_{ijs} - \sum_{k \in I_{is}} R_{iks} \right)$$

where $\delta_{ci}(s)$ is the difference between responses of congruent and incongruent visual percepts for subject s, S is the set of trials with sound, C_{is} is the set of congruent responses from subject s on trial i, I_{is} is the set of incongruent responses from subject s on trial i, and R_{ijs} is the duration of response j from subject s on trial i.

Second, we recorded the difference in reported dominance time of the male stimulus on trials with male sound versus trials with no sound. We did the same with female stimuli (dominance time of female stimulus on trials with female sound versus trials with no sound) and summed the results to see how much more often congruent stimuli were dominant versus their neutral counterparts in the no sound trials. We did the same comparison in the other direction to see what (dis)advantage incongruent stimuli had compared to neutral stimuli. Since these measures do not come at the expense of one another like those above (both congruent and incongruent are being compared to neutral, rather than to each other), the effect should be weaker but still an interesting basis for comparison. Also note that though there are twice as many trials with sound, we are only considering the congruent or incongruent responses from each trial. Since we consider both male and female responses from every no sound trial the comparison is even. Similar to the above, we performed a positive (for congruent, negative for incongruent) one-tailed t-test comparison between the distribution of this measure over the subjects and a null distribution with zero mean. The equation for calculating this measure (in the congruent case) is as follows

$$\delta_{csns}(s) = \sum_{i \in S} \left(\sum_{j \in C_{is}} R_{ijs} - \sum_{k \in N} \sum_{l \in FM_{ks}} R_{kls} \right)$$

where $\delta_{csns}(s)$ is the difference between congruent sound responses and corresponding no sound responses for subject s, N is the set of trials with no sound, and FM_{ks} is the set of female or male responses (as opposed to not sure) for subject s on trial k. Other terms are the same as described above and the incongruent case is a simple modification.

Results

Figure 2 shows the result (ordered by increasing effect) of our first analysis. In each of the raw, color corrected and gender corrected conditions we reject the null hypothesis that congruent stimuli are as likely to be dominant as incrongruent stimuli ($\mu = 60.25$ $\sigma = 102.37$ $p < .017, \mu = 44.42$ $\sigma = 83.59$ $p < .026, \mu = 54.33$ $\sigma = 96.44$ $p < .020$, respectively). Qualitatively the results don't change much after correction, as would be expected given the counterbalanced experimental design.

Figure 3 shows the result (in the same order as Figure 2) of our second analysis for the advantage of both congruent and incongruent stimuli compared to neutral stimuli. In the congruent case we do not obtain a significant effect, but there is a trend ($\mu = 24.26$ $\sigma = 69.21$ $p < .091$), as can be seen from

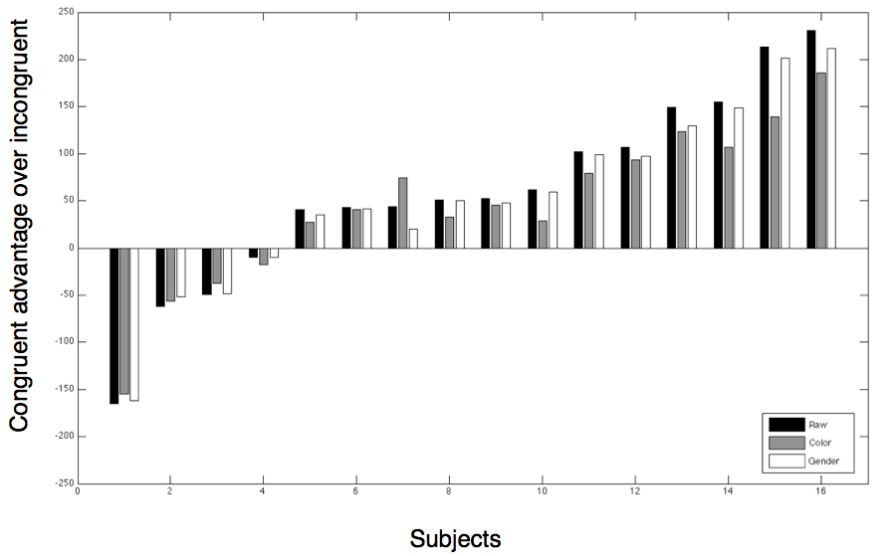


Figure 2: Difference in seconds between total congruent responses and incongruent responses, bucketed by subject in order of increasing effect. Bar color represents correction method.

the bar plot. The incongruent case does show a significant disadvantage compared to neutral ($\mu = -35.99$ $\sigma = 56.39$ $p < .011$). Notably, subject 4 changes character significantly in this analysis as compared to the previous. As a post-hoc investigation, we calculated the average absolute difference in total dominance duration between sound and no sound trials for each subject. Subject 4 had a per trial average dominance duration of 14.3 seconds less on the sound trials. Over all the other subjects the average absolute difference was 2.4 seconds with a max of 5.4. Clearly subject 4 responded much less to the trials with sound than was typical for the subject pool. If subject 4 is excluded from the hypothesis test on the second analysis there is a significant effect of congruent sound versus no sound ($\mu = 33.84$ $\sigma = 59.66$ $p < .023$) and the incongruent effect remains significant though weakened as one would expect ($\mu = -31.09$ $\sigma = 54.73$ $p < .023$). These results are very much in line with the first analysis.

Discussion

Our result represents an important step forward in mapping out the ways in which high-level processing can impact rivalry. Unlike previous high-level effects, such as global coherence and biological motion, this effect is not solely in the visual domain. Instead it is the result of matching a voice to a speaker, constituting the integration of both auditory and visual information.

It is not clear from this experiment whether the gender of the voice alone was enough to cause greater dominance of the congruent video, or whether voice to lip matching was responsible for the effect. Either of these effects would reveal an interesting cross-modal influence on binocular rivalry. Future studies that pair each speaker's video for one of their stories with the audio from the other could help illuminate the particular role of each aspect. [26] seems to show that semantically relevant sounds can bias the perception of eye rivaling static stimuli. However given that cross-modal voice to lip matching is so robust [21], we believe that the dominance effect is likely helped by the temporal coherence of voice and lips. An interesting question is whether subjects are aware of the matching even when the incongruent stimulus is dominant. This question could be addressed with an experiment that manipulates the temporal phase of the matched visual video during periods of nonperception. As soon as the subject indicates dominance of an incongruent stimulus one could switch or delay the audio track so as to put it out of sync with the congruent video. This might require less naturalistic stimuli (with pauses between words, for example) in order to execute without the audio sounding garbled. If subjects detect the lack of matching even when they're not consciously perceiving the congruent stimulus (e.g. by changing perceived dominance status), it would indicate the presence of a cross-modal blindsight for the voice to lip relationship.

Given that voice to lip matching is such a powerful effect, we were very concerned to sync the audio to the video precisely. It was not possible to do this perfectly given our experimental design, which required us to decouple the audio and video, though we came very close. We wonder whether the few subjects that showed an auditory congruence effect in the opposite direction were more temporally sensitive subjects (perhaps musically trained?) and more sensitive to slight offsets in sync and thereby biased against congruent stimuli at times when the sync is not quite right (a slightly offset audio/visual pair would be more anticorrelated than an unrelated audio/visual pair leading to a potential preference for perception of the unrelated video). An experiment where audio/video sync is manipulated across trials and subjects are

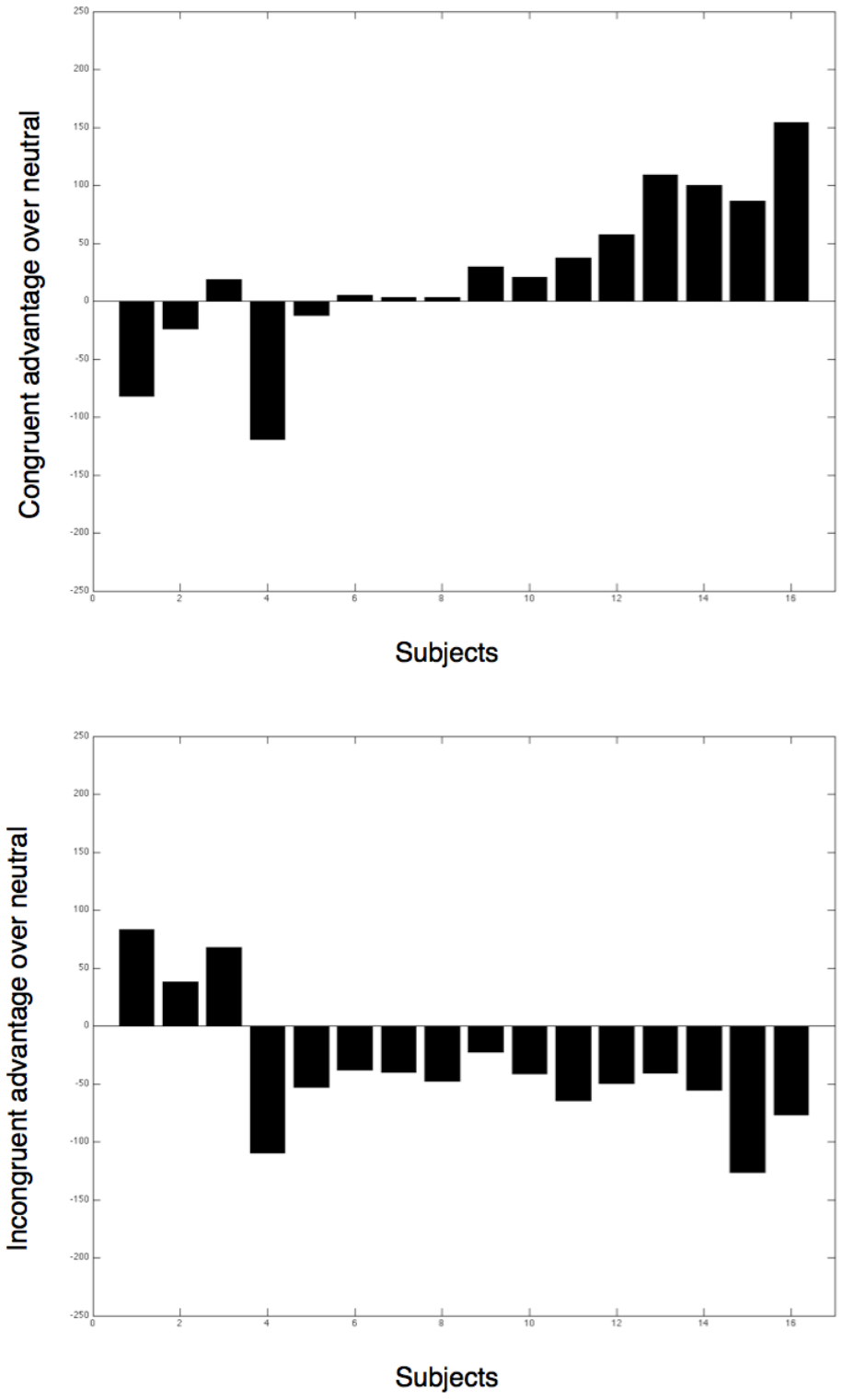


Figure 3: Difference in seconds between total congruent responses (top) and corresponding responses on no sound trials and between incongruent responses (bottom) and corresponding no sound responses, bucketed by subject in the same order as Figure 2. See the post hoc outlier analysis of subject 4 is in the results section.

screened for temporal sensitivity could help address this issue.

Another interpretation of our results might be that the auditory stimulus is causing subjects to consciously attend more to the congruent video, resulting in a greater dominance period due to attention rather than a more automatic cross-modal effect. As mentioned in the introduction, however, attention mainly seems to affect the rate of alternation [10] and the initially dominant stimulus [13], both of which have a limited impact on dominance duration. When attention does bias dominance duration it is usually when subjects are attending specific stimulus features [12]. By contrast, [26] seem to find a significant effect of commanded attention (which adds with their cross-modal interaction), but it is unclear whether they had subjects maintain fixation. Without maintaining fixation, subjects' eyes can easily wander or be specifically directed to higher contrast/complexity regions of the attended image and thus bias dominance on a low level. Since our subjects were specifically instructed to fixate on a fixation cross, and had no task related reason to remember or interpret the stories our actors told (subjects were simply told they would hear sounds during some of the trials), we do not believe that attention had a significant impact on our results. A future study carefully designed to focus on the interaction of cross-modal/attention effects (e.g. by requiring subjects to attend both to stimuli congruent and incongruent with a soundtrack) would likely help illuminate this issue further.

We believe this is an exciting result for the bistable perception field. It shows a new way in which high-level perceptual processes can interact with conscious perception and opens up new ground for researching the nature of both cross-modal interactions and bistable perception.

Acknowledgments

This work was supported by NSF IGERT Grant #DGE-0333451 to GW Cottrell/VR de Sa & NSF CAREER Grant #IIS-0133996 to VR de Sa.

References

[1] F. Tong, M. Meng, and R. Blake. Neural bases of binocular rivalry. *Trends in Cognitive Sciences*, 10(11):502 – 511, 2006.

[2] A Polonsky, R Blake, J Braun, and D Heeger. Neuronal activity in human primary visual cortex correlates with perception during binocular rivalry. *Nature neuroscience*, 3(11):1153–1159, 2000.

[3] S Lee and R Blake. V1 activity is reduced during binocular rivalry. *Journal of Vision*, 2:618–626, 2002.

[4] J. D. Haynes, R. Deichmann, and G. Rees. Eye-specific effects of binocular rivalry in the human lateral geniculate nucleus. *Nature*, 438(7067):496–499, 2005.

[5] David Alais and David Melcher. Strength and coherence of binocular rivalry depends on shared stimulus complexity. *Vision Research*, 47:269–279, 2007.

[6] D Baker and E Graf. Natural images dominate in binocular rivalry. *Proceedings of the National Academy of Sciences of the United States of America*, 106:5436–5441, 2009.

[7] I Kovács, T Papathomas, and M Yang. When the brain changes its mind: Interocular grouping during binocular rivalry. *Proceedings of the National Academy of Sciences of the United States of America*, 93:15508–15511, 1996.

[8] Derek H Arnold, Bridie James, and Warrick Roseboom. Binocular rivalry: spreading dominance through complex images. *Journal of Vision*, 9(13):4.1–9, 2009.

[9] H. von Helmholtz. *Treatise on physiological optics, Vol. III*. Dover, 1925.

[10] Ming Meng and Frank Tong. Can attention selectively bias bistable perception? Differences between binocular rivalry and ambiguous figures. *Journal of Vision*, 4(7):539–51, 2004.

[11] R Van Ee, LCJ Van Dam, and GJ Brouwer. Voluntary control and the dynamics of perceptual bi-stability. *Vision Research*, 45(1):41–55, 2005.

[12] S Chong, D Tadin, and R Blake. Endogenous attention prolongs dominance durations in binocular rivalry. *Journal of Vision*, 5:1004–1012, 2005.

[13] S Chong and R Blake. Exogenous attention and endogenous attention influence initial dominance in binocular rivalry. *Vision Research*, 2006.

[14] J Mishra and S Hillyard. Endogenous attention selection during binocular rivalry at early stages of visual processing. *Vision Research*, 2009.

[15] A Maier, N Logothetis, and D Leopold. Global competition dictates local suppression in pattern rivalry. *Journal of Vision*, 2005.

[16] T Watson, J Pearson, and C Clifford. Perceptual grouping of biological motion promotes binocular rivalry. *Current Biology*, 14:1670–1674, 2004.

[17] J Hupé, L Joffo, and D Pressnitzer. Bistability for audiovisual stimuli: Perceptual decision is modality specific. *Journal of Vision*, 8(7):1.1–15, 2008.

[18] H McGurk and J MacDonald. Hearing lips and seeing voices. *Nature*, 264:746–748, 1976.

[19] L Rosenblum. Speech perception as a multimodal phenomenon. *Current Directions in Psychological Science*, 17(6):405–409, 2008.

[20] A Vatakis, C Spence, and S Vecera. Crossmodal binding: evaluating the "unity assumption" using audiovisual speech stimuli. *Perception and Psychophysics*, 69(5):744–756, 2007.

[21] LD Rosenblum, NM Smith, SM Nichols, S Hale, and J Lee. Hearing a face: Cross-modal speaker matching using isolated visible speech. *Perception and Psychophysics*, 68(1):84, 2006.

[22] Ayse P. Saygin, J. Driver, and Virginia R. de Sa. In the footsteps of biological motion and multisensory perception: Judgements of audio-visual temporal relations are enhanced for upright walkers. *Psychological Science*, 19(5), 2008.

[23] Y. Kamitani L. Shams and S. Shimojo. What you see is what you hear. *Nature*, 408, 2000.

[24] Ladan Shams Shinsuke Shimojo. Sensory modalities are not separate modalities: plasticity and interactions. *Current Opinion in Neurobiology*, 11:505–509, 2001.

[25] Debra A. Zellner and Mary A. Kautz. Color affects perceived odor intensity. *Journal of Experimental Psychology: Human Perception and Performance*, 16(2):391–397, 1990.

[26] Yi-Chuan Chen, Su-Ling Yeh, and Charles Spence. Crossmodal constraints on human visual awareness: Auditory semantic context modulates binocular rivalry. Presented at the 10th annual Vision Sciences Society Meeting, 2010.

The Necessity of Ordinary Experience

Robin Flanagan (FlanaganR@wcsu.edu)
Department of Psychology, WCSU, 181 White Street
Danbury, CT 06810 USA

Abstract

I argue in this paper that ordinary experience is not only a nice part of everyday life; it is a necessity for the development of human knowledge. I begin by looking at why the particular biological machinery that defines our nervous system matters. I then examine the particular machineries that constrain but also foster the development of human knowledge. Finally, I examine the kinds of activities that foster the development of knowledge, given the constraints of the given machinery, and conclude that activities that are repeated often and that involve meaningful interaction with an inherently meaningful environment form a plausible basis for the formation of knowledge within the particular neural net machinery that evolution has produced for us.

Keywords: Learning; neural networks; embodied cognition; practice; education; development; instructional technology

Mind and world in short have been evolved together, and in consequence are something of a mutual fit.

(James, 1948, p. 4)

The Implementation Problem

The implementation question is the notion that once a system of knowledge has been completely and accurately articulated, it shouldn't matter in what kind of machinery the system is implemented. This was a major assumption of cognitive science for quite a long time, and to its credit it was a very useful and fruitful assumption. If we assume that there is no important difference between carbon-based machinery and silicon-based machinery, and this is a very reasonable assumption, we can investigate and test knowledge systems on silicon-based machinery, machinery which is much easier to control, much easier to completely specify, and much easier to manipulate in ethical ways. However, this assumption has two gaping holes in it: how does the knowledge get into the machinery (most biological organisms have no programmers to install useful data structures or programs, while most silicon-based machines do have programmers), and how does the knowledge get interpreted (most silicon-based machines have intelligent "users" to interpret the output; most biological organisms must interpret the knowledge for themselves).

If, instead of ignoring implementation, we examine how the actual machinery works, we find that there are many important constraints derived directly from the machinery that actually help us to understand how the knowledge gets incorporated into the machinery and how the "knowledge" in the system gets interpreted. This, of course, does not mean that a silicon-based machine couldn't learn and interpret on its own (see (Brooks, 2008) for example); it only means that silicon-based machinery isn't necessarily constrained by the same physical qualities that constrain biological organisms. A lot of very interesting work in artificial intelligence, does examine cognition while taking biological constraints into consideration, and these lines of research have been extremely fruitful, which should help to support the idea that implementation does indeed matter. The embodied cognition paradigm already assumes, however, that implementation is a critical element of any intelligent system.

The Basic Machinery

That leads to the examination of the actual elements of the biological machinery from which the nervous system is constructed. There are, of course, very few elements in the biological machinery. The main element is an ordinary neuron, which is not too dissimilar from other cells in the biological organism. Like other cells in the biological organism the neuron is best at responding to elements in the immediate surroundings. In other words, the neuron is best at noticing what's in its immediate neighborhood and responding by secreting to its immediate neighborhood.

However, the neuron can take on very unusual shapes, and these shapes, make them particularly good for communicating with each other, by redefining what is meant by "its immediate neighborhood". The maximized surface area of the neuron (the dendrites) allows the neuron to receive multiple messages simultaneously from other neurons or from the environment. The other part of the neuron's unusual shape (the axon) can sometimes be quite a long extension of the cell body. The axon is the main tool that the neuron has at its disposal for communicating to other neurons or to the muscles. So just by changing its shape the neuron has the ability to get information from, and have an effect on, parts of the nervous system and ultimately parts of the body that are not apparently in its immediate neighborhood.

This is important because the main technique that neurons have for getting information, and for sending information, involves the idea of simple local processing. So it's important to note that "local" for the neuron has been redefined to include connections to quite distant elements of the nervous system and the biological organism. In fact, in the case of the photoreceptors, "local" involves light waves arriving in the immediate vicinity from potentially extremely distant locations. Simple local processing is the kind of processing that single-celled organisms developed at the very beginning of organized life, to detect things in their

immediate environment, and through very simple rules made decisions about how to act on their environment. The typical example is a bacterium floating through water. When it detects a particular toxin in the environment, it activates its flagellum and flaps away from the toxin. The cool thing about simple local processing is that when many organisms are using simple local processing at the same time, intelligent behavior can emerge at the level of the group or colony, without any programmer or leader or teacher.

Because there is no programmer or leader or teacher to direct the nervous system this is an incredibly useful quality to include in any description or explanation of biological intelligence to account for the undoubtedly quite intelligent behavior of this leaderless system.

So, the basic elements of which our nervous system is composed consist of billions of very simple agents, performing simple local processing in which "local" has been redefined to include any "neighbor" to which a neuron's unusual shape can give it access, including, for example, any light wave event within the visual vicinity of the amazing biological eye. This massively parallel system of simple agents acts without a leader, without a programmer, without a teacher; yet intelligent and useful behavior emerges over time. We turn next to the question of how knowledge, or intelligent behavior, can emerge in such a system.

How knowledge develops in such a system

While no single model of the human nervous system has been universally accepted, we have established the basic building blocks and parameters from which it must be built. Several of the basic mechanisms with which such a neural network could store knowledge have also been identified.

The first important mechanism was established about a hundred years ago by Pavlov (2009)and articulated more fully in the sea slug by Kandel and his colleagues (Hawkins, Greene, & Kandel, 1998, for example). The ability of the nervous system to associate a previously non-meaningful stimulus with an already meaningful stimulus may seem rather minor and non-cognitive when discussed within the context of dog saliva and sea slugs, and yet this is an amazingly useful mechanism. Association between a stimulus that is already meaningful and a previously meaningless stimulus can produce symbols, where a symbol means anything that stands for something else. Surely this is the basis of the nervous system's ability to use language and, more generally, abstract symbols. Abstract symbols are, by definition, meaningless stimuli on their own which have taken on meaning by association with something already meaningful.

The second important mechanism was robustly established during the half century of American behaviorist research (Staddon & Cerutti, 2003). Operant conditioning increases the probability of a pattern of neural activity to reoccur if that pattern has proven to be useful (that is, if it has been reinforced). In a probabilistic network, this couldn't be more important. A relatively more predictable pattern of activation that is meaningful or important to the organism is pretty close to a basic definition of intelligent behavior, or knowledge. Again, operant conditioning may seem too basic and non-cognitive when discussed in the absence of a mind or within the context of training animals, as behaviorism often is; yet surely the ability to increase the probability of activating a useful pattern of neurons when it becomes clear that the pattern is, in fact, useful, could form the basis of an endogenous back-propagation system, the exogenous form of which is such an essential aspect of so many artificial neural nets (see, for example, McClelland and Rumelhart (1988)). Whether it forms the basis of the feed backward system or not, most would agree that knowledge that is more probable, rather than less probable, to become available at the appropriate time is the main point of learning and education.

The third important mechanism was hypothesized by Hebb sixty years ago (1949), and established more recently in empirical neuroscience research ((Isaac, Buchanan, Muller, & Mellor, 2009) for example). Hebb theorized that neurons that become activated simultaneously would be subsequently more likely to activate each other. This has been found at least in the case of the NMDA receptor, a receptor that requires simultaneous messages in order to allow permanent, structural changes to occur at the synapse (see (Isaac, et al., 2009) for example). This is perhaps a more general mechanism upon which both Pavlovian association and Skinnerian contingency are both built. Long-term potentiation has been the chief candidate for this process. Long-term potentiation involving the NMDA receptor requires more than one converging pathway to neural activation. Also important is the idea that this process is dependent on an overwhelmingly huge stimulus, or an often repeated activation before it makes permanent changes to the synapse. If long-term potentiation (or any kind of wiring) developed after every mere exposure the neural net would be in constant flux without the ability to store meaningful knowledge (something to keep in mind when considering so-called "smart" genes and genetic modifications). The ability of neurons to strengthen their association when they find themselves simultaneously activated over time is essential to both forms of conditioning, as well as learning to perceive and to act on any reliable invariance in the internal and external environment. Invariance in the environment, by definition, provides almost endless repeated activation in response to objects and events that are important, or, at least, enduring.

Finally, with lots of neurons activated simultaneously in response to an event in the environment, distributed "representation" is possible: that is, a distributed set of neurons together form a concept. This is important as a storage mechanism, but it is even more important as a means of developing categories and abstract concepts. When lots of neurons, rather than a single neuron, become activated by a particular stimulus, and then another large group of neurons becomes activated by a slightly different

stimulus, any overlapping active neurons get twice the opportunity to wire together with each other, and so subsequently are even more likely to activate each other. This overlapping set comes to stand for (or "mean") the precise similarity between the two stimuli, not as an analogy but as a literal overlapping commonality. This is a profoundly important part of our machinery if we want to be able to explain the human genius for categorization, abstraction, and creativity.

Very few psychologists admit that these crude mechanisms are useful for more than motor skill learning and perception. Yet, what other mechanisms have been identified in the nervous system to account for lasting changes? I am aware of none. So, leaving physical skill learning and perception aside for the moment (although they're quite important) let's examine, briefly, how verbal, spatial, or declarative knowledge could develop in such a system, although the research in this area is ongoing and not at all settled yet.

These mechanisms certainly do look better suited to the implementation of non-declarative knowledge than of declarative knowledge. Non-declarative knowledge can build up over time through normal interactions and perceptions, and even without conscious awareness or attention. But how do we explain the (seemingly) more cognitive, conscious and occasionally instantaneous category: declarative knowledge? How could declarative knowledge be implemented in such a system?

Unlike all other organisms, human beings have a rich set of verbal (as well as visual) symbols at their disposal. One possibility is that words become associated (through classical conditioning mechanisms) with "concepts" already established in the neural network through Hebbian synapses. In fact, Bloom and her colleagues found that as soon as children are reliably able to refer to objects in their environment, jointly with their caregiver, vocabulary suddenly blossoms (Lifter & Bloom, 1989). Goldin-Meadow found that as soon as learners were capable of gesturing appropriately during problem-solving, that the correct words almost immediately followed (2003). It seems that in humans, at least, language is produced almost simultaneously with the ability to identify and perceive a referent. If this is the case, this is a powerful addition to the simple machinery with which we have to work: to be able to have a word associated with each distinction we are capable of perceiving or acting upon.

Once a word is in place (associated with a meaningful distinction) the neural net can use the activation of a word in place of the primary experience: the word can initiate a cascade of neural activity that is very similar to the cascade that would be produced by the primary experience. At this point, a coach, or a teacher, or a friend, or a parent can use a word ("hot") to produce the same neural activity that might have been produced by a similar ("hot") experience, thus allowing learning to take place without the primary experience. Clearly the primary experience, or some critical conjunction of important partial experiences, must have occurred at some point. But learning can quickly be produced in the absence of the primary experience once the word is in place. From here it is a not impossibly large leap to the nervous system supplying the words internally in the absence of an external coach, teacher, friend or parent. These internally activated words, then, could form the basis of explicit knowledge and rational thought.

Re-activation of sensory-motor cortex, followed by a cascade of neural activity similar to primary activation, has repeatedly been found to be the case with stored concepts (see, for example, the visual imagery work of Kosslyn (2005) and the motor imagery work of Jeannerod (1994)). The research on mirror neurons has even indicated that watching someone else's behavior can trigger a cascade of neural activity that is similar to the neural activity involved in one's own primary experience (Brass & Rüschemeyer, 2010).

The other aspect of declarative knowledge, the apparent ability of explicit knowledge to be examined consciously, needs more explanation, and probably a completely separate paper. Briefly, though, the main advantage of implicit, or non-declarative knowledge, is that it is so well integrated into the neural network that it is ready for use without any conscious reflection. That is of course its main liability as well, because without conscious reflection there is no room for "free will", no room for new responses, and no room for transfer of knowledge to novel situations. How, then, does declarative knowledge gain this apparently conscious element? There is perhaps no hotter topic in philosophy of mind these days (see Metzinger (2009) for example), so I will not presume to solve this problem for all time. However, an intriguing possibility, and one that is in line with what is known about the biological constraints of the human nervous system, was put forth decades ago by Antonio Damasio (1989). He pointed out that a mechanism in the hippocampus allowed incoming messages to be, essentially, bounced back to the sensory store from which they had just come. Because incoming sensory information must reach the hippocampus in a cohesive timeframe, the bouncing back must also occur in tandem, restimulating the same sensory stores as the original experience. He did not discuss verbal stimulation in particular, but because we know that verbal information stimulates the same sensory store as heard language (Hubbard, 2010), this mechanism should work for verbal information as for any other sensory stimulation. What does this ability to bounce an experience back for re-experiencing buy us? Just this: it allows for the opportunity, as any multi-neuron synaptic junction would, for the original stimulus to be affected by other elements rather than triggering an automatic and unalterable cascade of activity. Implicit knowledge does not need to go through this bounce-back process because it's already usable, and in many cases, already crystallized. Explicit knowledge, however, differs from implicit knowledge in the "second chance" it gives its network, and of course the environment, to affect the cascade of activity in a new or more subtle way. This explicit second chance may not result in fast, or

graceful, processing and activity (that is the strength of implicit knowledge), but it gives our neural net the opportunity to bring old symbols, old categories and old knowledge to bear on a new situation. The analogy I have used with students is very over-simplified, but may help to illustrate this distinction. If a sensory stimulus is like a pebble and our neural network is like a pond, then implicit knowledge is the set of waves that travel across the pond without hindrance when the pebble is dropped into its center, and explicit knowledge is the set of waves that results from the pebble's original waves encountering a partial barrier that bounces back some of the waves allowing them to interact again with the out-moving waves. The explicit is more complicated, more interesting, more filled with information (in the information theory sense), but the implicit is more graceful and efficient.

So, it's possible for both non-declarative and declarative knowledge to develop within the severe constraints built into the biological machinery about which we already know.

Activities that Foster Development

What kinds of activities, then, foster knowledge development in such a system, with so few clear mechanisms for plasticity? Imagine the elaborately connected human nervous system moving about in the environment with all of its electrical activity visible for observation. Notice that the nervous system is constantly active and that what changes is the relative activity of the system: relative both in time and space. This system does not passively await inputs, but constantly changes in response to the particular interactions it has with its environment. It should be clear at this point, that a system such as this one has no "input" device. It has, rather, the ability to make small adjustments in real time in response to real events. This system will only be as useful as the meaningful distinctions to which it can attend and respond. What activities will, naturally, produce patterned and intelligent behavior?

Perhaps obviously, the neural network will store reliable patterns detected in the environment: if a set of neurons is consistently firing together, they will begin to wire together, thus storing a united response to a unified set of stimuli. There are two major sources for such reliable patterns: the natural invariances in the physical world, and the sets of actions that produce reliable (or meaningful) results for the organism (contingent activities). Notice how perfectly these sources match our two major learning mechanisms: associative conditioning and operant conditioning.

Invariance in the Environment

Why does the physical world provide such a rich source of useful invariances (or correlations) for the nervous system? The short answer is "evolution". Because the particular physical environment in which we all develop is the product of multiple, simultaneous lines of successful evolution, within the same set of physical constraints based on the physical structure and physical laws of this particular planet, the characteristics that tend to appear simultaneously tend not to be arbitrary co-occurrences, but rather quite meaningful and successful co-occurrences. In other words, if our nervous system happens upon a set of co-occurring characteristics in the natural world, they are extremely likely to be the product of a long and successful line of evolution, and therefore be the opposite of arbitrary or capricious.

All else being equal, then, the set of repeated co-occurrences we encounter will tend to be meaningful, not meaningless, co-occurrences, and therefore very useful for us to learn to perceive, "chunk" and to be able to act on. Our physical environment is full of non-arbitrary co-occurrences. The physical laws at work here are the same physical laws that have shaped our planet for billions of years and that have driven evolution of all the living species we encounter since life began on this planet. And the co-occurrences of living things in a particular environment are also non-arbitrary because these living things have had to survive within the same environment for millions of years. So the living organisms that we encounter have been successful not just in our particular physical environment, but in our particular ecological niche as well.

Contingent Activities

Held and his colleagues found quite a while ago that contingent experiences were necessary for the normal development of kittens (Held & Hein, 1963). In his elegant set of experiments, in which kittens were literally yoked during their daily visual stimulation and were able to move around the visual stimuli based on just one of the yoked kittens' movements, Held showed that kittens with completely equal visual stimulation, and deprivation, developed completely different visual capabilities depending only on whether the visual stimulation was contingent on the kitten's own activity.

Fox and Oakes updated Held's experiments by doing a similar set of experiments using undergraduates, instead of kittens, and video games, instead of a yoked carousel experience (Fox & Oakes, 1984). In this set of experiments, undergraduates were virtually yoked to each other while they played one of two versions of a video game. In one version of the game, the undergraduates' success at destroying elements of the virtual world was completely contingent on their motor behavior: if their aim and timing was good, they were able to blow up a lot of objects; if their aim and timing was poor, they had little success. In the second version of the game, undergraduates experienced the same (yoked) number of apparent successes, but the success had nothing to do with their motor behavior: it depended completely on the success of the undergraduate to which they had been virtually yoked. However, the second version of the game was designed to make it look like the success was contingent on the player's skill: elements were slowed or speeded up in order to make appropriate, successful, contact. When tested afterwards all of the undergraduates felt as though they had succeeded: consciously they felt like their actions mattered. But the undergraduates who played

the non-contingent form of the game were not as successful at a subsequent, unrelated, lexical decision task.

Notice that "contingent experience" is any experience in which the organism's actions are related, reliably, to the feedback the organism receives, whether or not the organism is consciously aware of this reliable relationship.

Both Invariance and Contingency

Diamond and Rosenzweig and their colleagues looked at both elements at once. They found that rats that grew up in an environment with lots of new, physical and social interactions, developed more useful and heavier brains (Rosenzweig, Bennett, & Diamond, 1972). Interestingly, when the interaction was eliminated, by having rats near enough to watch but not interact with all the stimulation, the rats' brains did not become as useful or heavy. Most importantly, however, these "enriched" lab rats had brains that were significantly less useful, heavy, and well-connected than rats raised in the wild (where both invariants, and contingency are much more widely available) (Huck & Price, 1975; Zhao, Toyoda, Wang, & Zhuo, 2009) .

Flanagan (1996) showed that in a normal classroom setting, third graders who did an activity that involved contingent rather than non-contingent feedback for just fifteen minutes were subsequently significantly less likely to give up in a challenging but possible puzzle. Furthermore, third graders who used physical rather than virtual materials were significantly more likely to be able to build on that knowledge.

Natural feedback refers to feedback that is not dependent on a teacher, programmer or author, but that is instead inherent in the activity itself. So dropping objects of different weights does not require a teacher to give positive or negative feedback; the gravity of the physical world gives this feedback naturally. Most interactions with the natural world provide such feedback, but natural feedback is not limited to the natural or physical world: computer programming, for example, provides natural feedback because the programmer does not need a teacher or authority to provide positive or negative reinforcement – the programmed code either works or it doesn't. All else being equal, though, the natural world is the safer bet since co-occurrences in the natural world are the product of evolution, and interactions with the natural world follow the laws of physics. Artificial, or authored, environments depend completely on the author, or programmer to provide meaningful co-occurrences, and to provide meaningful feedback – these must be deliberately incorporated, while in the natural world they are already an integral part.

Natural feedback is also less available in stereotypically female hobbies than in stereotypically male hobbies. Playing with water pistols provides natural feedback – either you get wet or you don't. Many stereotypically female hobbies depend on the opinions offered by peers or authority figures: does this look pretty? Have I pleased you? Is this good? Dweck and her colleagues have found that personal feedback rather than task-related feedback interferes with the mastery orientation of children solving challenging problems (Dweck & Leggett, 1988). Because of this difference in available stereotypically female and male after school activities, Flanagan and Canada provided school-age female students with one hour a week of after-school activities in which the students got natural feedback for both invariance in the environment and their own contingency (Flanagan & Canada, 2010). These students did computer programming (Scratch (Group) or Lego Mindstorms (Lab, 1999)) or building scale models (Google Sketch (Google, 2010) or physical craft materials) for eight weeks. At the end of the eight weeks the students had significantly better spatial reasoning skills than a similar control group, and felt significantly more confident about doing math and using computers.

Ordinary Experiences

In environments that consist of inherently meaningful co-occurrences and opportunities for consistently meaningful feedback the nervous system thrives. Repetition, or practice, in such environments should produce robust, well-organized, functional nervous systems. The practice effect is well-established, but shouldn't be ignored: too often we turn to the conceptual or technological shortcut when mere practice in a meaningful environment would do more good. Imagine a basketball team that got an hour or two of lecture a week and then several readings in order to get ready to play the season; imagine an orchestra that got an hour or two of lecture a week and then had to read their musical scores as homework for getting ready for their concert season. This sounds ridiculous, of course. But we expect our students to learn more "cognitive" skills this way even though it shouldn't work given the mechanisms available, and routinely fails to work (see (Sahiner, 1987) for example). If we accept the mechanisms we've been given, cognitive education should begin to look more like physical and musical education.

"Baby Einstein" media have recently been (finally) recalled because they probably do more harm than good (Lewin, 2009). As cognitive scientists we owe anxious parents the benefit of our expertise, and must point out that ordinary interactions with people and meaningful objects are better suited to the developing nervous system than "educational" consumer media. Because, (un)fortunately, constrained by the biological machinery with which we are born there is no magical input portal for pouring fully formed knowledge systems into the human mind: there are just a few simple mechanisms that must incorporate knowledge through lots of simple, ordinary, meaningful encounters over a long period of time.

Conclusion

The human nervous system is the product of millions of years of evolution within an ecology that has simultaneously been evolving. So it makes sense that the human nervous system should be optimized for operating within the

particular natural and physical world we call "earth". Indeed when we look at the particular mechanisms actually available to the human nervous system for learning and developing a solid knowledge base, these mechanisms seem to be ideal for detecting and learning naturally occurring invariances in our ordinary environment, as well as for learning actions that turn out to be important and meaningful to the nervous system itself. These are the very elements that Lloyd argued were the minimum essential requirements for anything we would consider to be a "mind" (1989). Furthermore, these mechanisms work best when the applicable neurons are activated simultaneously over a significant period of time.

Activity that involves important co-occurrences that are meaningful for the organism over significant periods of time are more succinctly termed "ordinary" experiences and are the foundation of our solid and meaningful neural network. We would be wise to build on this framework rather than attempting to circumvent it. Practice in real environments in real time has long been the accepted practice in athletics and music. It is time for other human endeavors to follow the same advice.

Acknowledgments

Thanks to James Schmotter of Western Connecticut State University and the CSU-AAUP for funding much of this work.

References

Brass, M., & Rüschemeyer, S.-A. (2010). Mirrors in science: How mirror neurons changed cognitive neuroscience. Cortex: A Journal Devoted to the Study of the Nervous System and Behavior, 46(1), 139-143. doi: 10.1016/j.cortex.2009.04.005

Brooks, R. A. (2008). Intelligence without representation. In W. G. Lycan & J. J. Prinz (Eds.), Mind and cognition: An anthology (3rd ed.). (pp. 298-311). Malden: Blackwell Publishing.

Damasio, A. R. (1989). Time-locked multiregional retroactivation: A systems-level proposal for the neural substrates of recall and recognition. Cognition, 33, 25-62.

Dweck, C. S., & Leggett, E. L. (1988). A Social-Cognitive Approach to Motivation and Personality. Psychological Review, 95(2), 256-273.

Flanagan, R. (1996). Learning through direct vs. indirect experience: The role of interactivity and physicality in media effects. 57, ProQuest Information & Learning, US. Available from EBSCOhost psyh database.

Flanagan, R., & Canada, T. (2010). Ordinary experience helps girls develop their spatial reasoning. Paper presented at the Association of Psychological Science, Boston, MA.

Fox, P. E., & Oakes, W. F. (1984). Learned helplessness: Noncontingent reinforcement in video game performance produces a decrement in performance on a lexical decision task. Bulletin of the Psychonomic Society, 22(2), 113-116.

Goldin-Meadow, S. (2003). Hearing gesture: How our hands help us think. Cambridge, MA US: Belknap Press of Harvard University Press.

Google. (2010). Google Sketchup.

Group, L.-l. K. Scratch Programming Language. Cambridge, MA: MIT Media Lab.

Hawkins, R. D., Greene, W., & Kandel, E. R. (1998). Classical conditioning, differential conditioning, and second-order conditioning of the Aplysia gill-withdrawal reflex in a simplified mantle organ preparation. Behavioral Neuroscience, 112(3), 636-645. doi: 10.1037/0735-7044.112.3.636

Hebb, D. (1949). The organization of behavior.

Held, R., & Hein, A. (1963). Movement-produced stimulation in the development of visually guided behavior. Journal of Comparative and Physiological Psychology, 56(5), 872-876. doi: 10.1037/h0040546

Hubbard, T. L. (2010). Auditory imagery: Empirical findings. Psychological Bulletin, 136(2), 302-329. doi: 10.1037/a0018436

Huck, U. W., & Price, E. O. (1975). Differential effects of environmental enrichment on the open-field behavior of wild and domestic Norway rats. Journal of Comparative and Physiological Psychology, 89(8), 892-898. doi: 10.1037/h0077160

Isaac, J. T. R., Buchanan, K. A., Muller, R. U., & Mellor, J. R. (2009). Hippocampal place cell firing patterns can induce long-term synaptic plasticity in vitro. The Journal of Neuroscience, 29(21), 6840-6850. doi: 10.1523/jneurosci.0731-09.2009

James, W. (1948). Psychology. Cleveland, OH: The World Publishing Company.

Jeannerod, M. (1994). The representing brain: Neural correlates of motor intention and imagery. Behavioral and Brain Sciences, 17(2), 187-245.

Kosslyn, S. M. (2005). Mental images and the brain. Cognitive Neuropsychology, 22(3-4), 333-347. doi: 10.1080/02643290442000130

Lab, M. M. (1999). Lego Mindstorms: The Lego Group.

Lewin, T. (2009, 10/23/2009). No Einstein in Your Crib? Get a Refund, New York Times.

Lifter, K., & Bloom, L. (1989). Object knowledge and the emergence of language. Infant Behavior & Development, 12(4), 395-423. doi: 10.1016/0163-6383(89)90023-4

Lloyd, D. (1989). Simple Minds. Cambridge, MA: MIT Press.

McClelland, J. L., & Rumelhart, D. E. (1988). Explorations in parallel distributed processing: A handbook of models, programs, and exercises. Cambridge, MA US: The MIT Press.

Metzinger, T. (2009). The ego tunnel: The science of the mind and the myth of the self. New York, NY US: Basic Books.

Pavlov, I. P. (2009). Conditioned reflexes: An investigation of the physiological activity of the cerebral cortex (1927). In B. F. Gentile & B. O. Miller (Eds.), Foundations of psychological thought: A history of psychology. (pp. 441-453). Thousand Oaks, CA US: Sage Publications, Inc.

Rosenzweig, M. R., Bennett, E. L., & Diamond, M. C. (1972). Brain changes in response to experience. Scientific American, 226(2), 22-29.

Sahiner. (1987). A Private Universe: Harvard-Smithsonian Center for Astrophysics, Science Education Department, Science Media Group.

Staddon, J. E. R., & Cerutti, D. T. (2003). Operant conditioning. Annual Review of Psychology, 54, 115-144. doi: 10.1146/annurev.psych.54.101601.145124

Zhao, M.-G., Toyoda, H., Wang, Y.-K., & Zhuo, M. (2009). Enhanced synaptic long-term potentiation in the anterior cingulate cortex of adult wild mice as compared with that in laboratory mice. Molecular Brain, 2.

A Cross-linguistic Model of the Acquisition of Inflectional Morphology in English and Modern Greek

Themis Karaminis (tkaram01@students.bbk.ac.uk)
Department of Psychological Sciences, Birkbeck College, University of London,
Malet Street, London WC1E 7HX, UK

Michael S. C. Thomas (m.thomas@bbk.ac.uk)
Department of Psychological Sciences, Birkbeck College, University of London,
Malet Street, London WC1E 7HX, UK

Abstract

We present a connectionist model of a general system for producing inflected words. The Multiple Inflection Generator (MIG) combines elements of several previous models (e.g., association between phonological representations of stem and inflection form: Rumelhart & McClelland, 1986; multiple inflections for a grammatical class: Hoeffner & McClelland, 1993; lexical-semantic input: Joanisse & Seidenberg, 1999; multiple grammatical classes: Plunkett & Juola, 1999). MIG assumes that the goal of the morphological component of the language system is to output a phonological form appropriate to the grammatical context in which the word appears. Our aim was to demonstrate that the model is able to capture developmental patterns in the acquisition of morphology in two different languages: one with a simple morphological system (English), and one characterized by rich morphology and absence of default forms (Modern Greek).

Keywords: Inflectional Morphology; Cross-linguistic Language Acquisition; Neural Network Modeling

Introduction

The Rumelhart and McClelland (1986) model for the acquisition of past tense was extremely influential and spawned new models on morphological acquisition. The model had several drawbacks. First, it is unlikely that the language system would have a specific component for one inflection type within one grammatical class. Second, the model did not simulate all the error patterns that children exhibit in development, notably the presence of unmarked forms. Third, the generalization of inflectional rules to unusual novel inputs was somewhat poor. More widely, it remains to be seen whether an architecture appropriate for modeling morphological acquisition in one language can readily extend to other languages that may have quite different inflectional paradigms. In this article, we present a model that is generalized to all inflectional types within a language (English) and show how the same architecture can be generalized to a different language with a richer inflectional structure (Modern Greek).

Our approach assumes that the language system comprises functional components and that at least one of the components is involved in conditioning the phonological properties of words during output so that their forms are appropriate to the grammatical context in the sentence in which they will appear. The goal was to simulate qualitative developmental patterns in the acquisition of English and Modern Greek, including the order of acquisition across inflection types and proportions of error types across development.

Previous connectionist models of morphology

Rumelhart and McClelland's (1986) model of the acquisition of the English past-tense was the first to apply the principles of Parallel Distributed Processing in the domain of inflectional morphology. This influential model showed that a two-layered feed-forward neural network architecture could learn mappings between phonological representations (Wickelfeature representations) of stems and corresponding past tense forms of English verbs. The model also simulated a wide range of phenomena reported in empirical studies of the acquisition of morphology, such as frequency effects and the U-shaped learning curve for the acquisition of irregulars (Brown, 1973).

This model demonstrated that an explicit representation of rules was not necessary for the acquisition of morphology. Instead, rule-like behavior was an emergent property of the learning system and reflected statistical regularities in the mappings of the training set. Rumelhart and McClelland challenged the existing 'symbolic' view, which proposed the dual-route account for morphological development (Pinker, 1984). According to this account, two separate mechanisms were involved in the learning of morphology. A rule-based system supported the learning of regular mappings, while a rote-memory system supported the learning of irregular mappings. The so-called 'past tense debate' emerged within the field of language acquisition.

Criticisms against the connectionist approach (e.g., Pinker & Prince, 1988) ranged from those pointing out implementational issues (e.g., the psycholinguistic implausibility of Wickelfeature representations) to those questioning the ability of the connectionist framework to address certain aspects of language acquisition (e.g., generalization). Subsequent connectionist studies addressed many of these criticisms by proposing more detailed models: Plunkett and Marchman (1993) refined the general principles of the model of Rumelhart and McClelland (1986) in a three-layered feed-forward architecture which employed more realistic phonological representations;

other studies incorporated lexical-semantics in the connectionist architecture to address dissociations in the learning of regular and irregular verbs (e.g., Joanisse & Seidenberg, 1999); Plunkett and Juola (1999) studied the acquisition of noun plural and verb past tense in a single connectionist network, while Hoeffner and McClelland (1993) considered multiple verb inflections. Finally, other work demonstrated that implementing a developmental deficit in connectionist architectures could simulate the acquisition of morphology in atypical language development. (e.g., Hoeffner & McClelland, 1993; Joanisse, 2004; Thomas & Karmiloff-Smith, 2003).

Acquisition of inflectional morphology in English

English inflectional morphology is characterized by its simplicity, manifested by the extensive use of default (base or uninflected) forms. For example, noun inflection does not consider gender and does not distinguish between the nominative and the accusative case. Psycholinguistic studies of inflectional morphology in English often focus on the domain of the past tense. This paradigm is of particular theoretical interest because it is quasi-regular. The majority of verbs form their past tenses through stem-suffixation (e.g., walk / walked). A rule determines the appropriate allomorphic suffix (/t/, /d/, or /^d/)) based on the last phoneme of the stem. However, a significant number of verbs form their past tenses irregularly (e.g., swim / swam, hit / hit, go / went).

Early studies on child language (e.g., Berko, 1958; Brown, 1973; de Villiers & de Villiers, 1973) established that different inflections in English are acquired in a consistent order along development. For example, the progressive of the verbs is acquired earlier than the past tense. Other studies addressed the profile of individual inflections in greater detail. For example, van der Lely & Ullman (2001) showed that accuracy rates are greater for regular than for irregular inflections. Accuracy also depends on type and token frequency. Frequency effects are more pronounced in irregular inflections (the so-called frequency by regularity interaction). Finally, children are efficient in generalizing the rule to novel forms (e.g., wug / wugged).

Morphological development is characterized by developmental error patterns. For example, children often produce base forms in contexts in which grammatical marking is obligatory (e.g., *He *come* home / He comes home). This type of error is referred to variously as a *no-mark error*, *no-change error* or *omission error*. Rice, Wexler, and Cleave (1993) suggested that omission errors define an early stage in language development, in which morphological marking is not applied consistently on the base forms. They termed this stage as the Optional Infinitive (OI) stage. Zero-mark errors occur in greater percentages in irregular inflections (e.g., Matthews & Theakston, 2006; van der Lely & Ullman, 2001).

Another prototypical error pattern is *over-regularization* or *over-generalization.* This type of error refers to the (incorrect) application of a rule on irregular stems (e.g., **thinked* / thought). Overregularization errors appear later in development than omission errors (Brown, 1973). As a result, in Brown's stage II (age range: 28-36 months, MLU range: 2.0-2.5) a sudden drop in the production of correct irregular forms was observed. This phenomenon is often described in terms of a U-shaped learning curve of irregulars. Overregularization errors are sometimes taken as evidence for the productive use of rules in child language (Marcus, 2000). Finally, a related error type is the *blend error* or *double-marked error* (e.g., Kuczaj, 1978). These errors refer to cases in which children apply a rule to an irregularly inflected form (e.g., **wented* / went).

Acquisition of inflectional morphology in Modern Greek

Modern Greek is a language with a rich morphological system. As Stephany (1997) describes, there are no default forms of words in Modern Greek. Instead, many different grammatical features are fused in single word forms. For example, nouns have grammatical gender, and are inflected with respect to case and number. Verbs are inflected with respect to person, tense, aspect and voice.

Modern Greek also presents different conjugational classes in nominal and verbal inflections, challenging the dichotomy between regular and irregular inflectional categories. For example, studies on the perfective past tense (e.g., Stavrakaki & Clahsen, 2001) describe three classes of verbs with respect to the marking of the perfective aspect. The 'sigmatic' class is the major class of verbs. The perfective past tense forms in this class are characterized by the addition of the aspectual marker /s/ ('sigma' in Greek) to the stem (e.g. pez-o / e-pek-s-a, play / played - 1st person singular). The addition of the aspectual marker may invoke phonologically predictable changes to the stems. A second class of verbs does not employ the aspectual marker /s/, and presents unpredictable modifications of the stem (e.g., plen-o / e-plin-a, wash / washed - 1st person singular). Finally, a third class of verbs have idiosyncratic perfective past tenses forms (e.g., tro-o / e-fag-a, eat / eaten - 1st person singular).

Stephany (1997) studied the production data of three children. Based on these data she suggested an order for the acquisition of different grammatical inflections and different grammatical features in Modern Greek. For example, tense is acquired earlier than aspect. Rare nominal conjugational categories are acquired late in development. As default forms are missing in Modern Greek, it has been suggested that the Optional Infinitive stage is realized by production of certain frequent forms in inappropriate contexts. Stephany (1997) observed that children undergo an early stage of development (up to 3 years old) in which they produce a lot of 3rd singular forms instead of the correct verbal inflections. Thus, 3rd singular forms could be considered an analogue of root infinitives in English (Varlokosta, Vainikka & Rohrbacher, 1998). Finally, with regard to the perfective past tense, Stavrakaki

and Clahsen (2009) found that the sigmatic rule is over-generalized in verbs belonging to non-sigmatic categories. The sigmatic rule is also preferred for the production of past tenses of novel verbs.

Simulations

Design

Our aim was to increase the generality of the original past tense model across inflection types, grammatical classes, and across languages. We began by combining elements of previous connectionist models of morphology (e.g., multiple grammatical classes: Plunkett & Juola, 1999; multiple inflections for a grammatical class: Hoeffner & McClelland, 1993; lexical-semantic input: Joanisse & Seidenberg, 1999) to implement a generalized inflectional system. The Multiple Inflectional Generator (MIG) considered three grammatical classes (nouns, verbs, and adjectives) and multiple inflections for each grammatical class (e.g., nouns: base forms, plurals, and possessives). The aim of MIG was to output a phonological form appropriate to the grammatical context in which the word appeared.

Following Plunkett and Marchman (1993), we constructed two training sets based on artificial languages that reflected the basic features of the morphological systems of English and Modern Greek. We performed two sets of simulations. In the first set of simulations, MIG was trained using the English training set. In the other, MIG was trained on the Modern Greek training set. In each condition, we contrasted the learning profile of MIG to corresponding data from empirical studies on the acquisition of morphology outlined above. For reasons of space, from the full set of behaviors exhibited by the model, we concentrate on reporting results from past tense. The goal was to replicate the following empirical effects: For English: (i) the relative acquisition of regular and irregular verbs; (ii) the frequency by regularity interaction in accuracy; (iii) the Optional Infinitive stage; (iv) the greater incidence of unmarked stem errors for irregulars; (v) the relative incidence of over-generalization and blend errors; (vi) generalization to novel stems. For Modern Greek: (i) the relative acquisition of sigmatic and non-sigmatic categories; ii) the production of 3rd singular forms as analogue of the Optional Infinitive stage; iii) the over-generalization of the sigmatic rule in verbs belonging to non-sigmatic categories; (iv) the generalization of the sigmatic rule to novel stems.

Architecture

The MIG employed a three-layered feed-forward neural network architecture. Four sources of information (cues) were presented in the input layer (Figure 1). (1) Input Phonology (95 units) encoded the phonological properties of the base forms using a five-slot scheme parallel to the that used in Plunkett & Marchman (1991, 1993). Each slot could encode a phoneme based on a distributed code of 19 binary articulatory features (Thomas & Karmiloff-Smith, 2003). The articulatory features (e.g., sonorant, consonantal, voiced, rounded) corresponded to standard linguistic categorizations (Fromkin & Rodman, 1988). The Input Phonology layer used only the first three slots to encode the phonological structure of monosyllabic words. (2) Following Joanisse and Seidenberg (1999), Lexical Semantics (1600 units) were used to provide localist representations of the meaning of each base form. (3) Grammatical Category (3 units) provided part-of-speech information. (4) Target Inflection (10 units) provided information on the type of inflection the network should consider (e.g., for verbs: base, past tense, 3rd singular or progressive).

The network was required to produce a phonological representation of the appropriate inflected form in the output layer (Output Phonology). The Output Phonology layer employed 95 units to implement a five-slot scheme. The last two slots were used to encode inflectional suffixes. In order to address morphology in Modern Greek, limited changes were introduced to the initial architecture solely to capture differences in the morphological structure of Modern Greek. In particular, the Target Inflection cue was expanded to include: gender, number and case information for nouns; gender, number, case, and grade information for adjectives; tense, aspect and person information for verbs. Additionally, Input Phonology provided phonological representations of word stems, without considering any inflectional suffixes and affixes. Finally, the Input and Output Phonology layers employed a twelve-slot scheme to incorporate morphological affixes, suffixes and disyllabic stems.

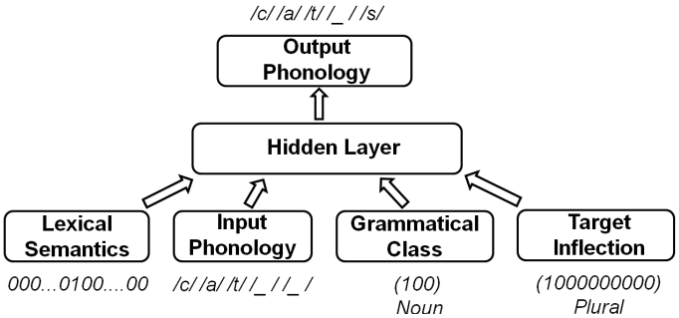


Figure 1: The architecture of MIG with an example of input-output mappings (here, to output the plural noun *cats*)

Training Sets

English Training Set. The training set for English was constructed based on measurements of type frequencies of different grammatical categories, different inflections or allomorphic subcategories of the same inflection. These measurements were derived from the tagged Brown corpus (Francis & Kucera, 1999) via computational linguistics methods. The NLTK open source software (http://www.nltk.org, accessed May 2010) was used for processing the Brown corpus. Frequencies of different grammatical categories and different inflection types were based on the counts of different tags in the corpus. Fre-

quencies of the allomorphic categories (e.g., /t/, /d/, /^d/ past tenses) were obtained using algorithms that identified the last phoneme of the stems.

The training set consisted of 1,600 words and 5,200 inflections based on those words (word-to-inflection ratio: ~0.3). The 1,600 words were artificial monosyllabic phoneme strings (800 verbs, 400 nouns, and 400 adjectives) which followed one of three templates (CCV, VCC and CVC; see Plunkett & Marchman, 1993). Ten different inflections were considered for the English training set (nouns: base form, plural, possessive; verbs: base form, progressive, 3rd singular, past tense; adjectives: base form, comparative, superlative). The inflected forms incorporated two additional phonemes for the inflectional suffixes. Combining words with their different possible inflections, the English training set comprised 5,200 stem / inflected form mappings. A simplified two-level scale of token frequency (Thomas & Karmiloff-Smith; 2003) was implemented by scaling the weight changes computed by the Back-propagation of Error algorithm (Rumelhart, Hinton, & Williams, 1986) after the presentation of each mapping. For arbitrary mappings (e.g., go / went) the weight changes were multiplied by 9 for tokens of high-frequency and 6 for tokens of low-frequency. For all other mappings, the weight changes were multiplied by 3 for high-frequency tokens, and 1 for low-frequency tokens.

A generalization set of 1,600 novel types and the corresponding 5,200 tokens was also created. It consisted of three subsets of stems with differing degrees of similarity to the stems of the training set. Items for the first subset of the generalization set were created by changing the first phoneme of existing stems. Items for the second subset were generated by changing the first two phonemes of the existing stems. In both cases a consonant was replaced by another consonant and a vowel with another vowel to conform to the phonotactics imposed by the three templates used for the training set. Items in the third subset were generated by changing the first two phonemes of existing stems in a way that violated the phonotactics of the artificial language.

Modern Greek Training Set. For the Modern Greek training set, type frequencies of different inflections and different conjugational categories were based on descriptions of Stephany (1997) or sampling of the Hellenic National Corpus of the Institute of Speech and Language Processing (ISLP, http://hnc.ilsp.gr/en/, accessed May 2010). In the absence of any other constraints, type frequencies were made parallel to type frequencies of the English training set. The Modern Greek training set consisted of 1600 types and 26,400 tokens (type to token ratio: ~0.06). The 1,600 types were a vocabulary of 800 verbs, 400 nouns and 400 adjectives. Items were dissyllabic, and conformed to the phonotactics of Modern Greek. Nouns were inflected in the nominative, the genitive and the accusative case of the singular and plural number. Verbs were inflected with respect to person (1st, 2nd, and 3rd), number (singular, plural) and tense (present, perfective past, imperfective past). Adjectives were inflected with respect to gender, case and number in the plain, comparative, and superlative grade. The Modern Greek training set consisted of a total of 26,400 mappings (tokens). A generalization set of 1,600 novel types and the corresponding 26,400 types was also constructed. Items for the generalization set were generated by changing the phonemes of the first syllable of the stem of items of training set.

Procedure

Networks were trained for 400 epochs, using the Back-propagation of Error algorithm (Rumelhart, Hinton, & Williams, 1986). The length of training was selected to ensure that the networks achieved final ceiling levels of performance. Based on piloting, the following parameters were used in both English and Greek versions of the model: 75 hidden units, learning rate 0.01, momentum 0. Results were averaged over 10 replications with different random seeds. Training was not incremental but used the full training set throughout, with one caveat: in each epoch, the network was exposed to a random 30% of the total inflected forms, corresponding to the number of different words in the training set.

Results

Network output was evaluated using a variant of the Nearest Neighborhood algorithm. The output activation for each slot was made equal to its nearest neighbor in the Euclidean space of the phonemes, so that continuous activations were converted to phonemic strings. The string was then assessed against pre-defined categories, based on patterns presented in empirical investigations of children's productivity (e.g., 'correct', 'omission errors', 'over-generalization errors', 'blend errors', 'other'). In this section we present initial results from the two simulations, demonstrating the viability of the more general model.

Simulation 1: English Training Set

The simulation results were parallel to the acquisition profile of the English past tense in several ways. Accuracy rates were higher for regulars than for irregulars. Type frequency effects were more pronounced for irregulars. MIG reproduced an OI stage, characterized by high percentages of omission errors for both regulars and irregulars. The rates of no-mark errors were higher for irregulars than for regulars. MIG also simulated overgeneralization errors and blend errors. Finally, the past tense rule was efficiently generalized in novel items with accuracy rates of 88%, 86%, and 43% for novel stems most to least similar to stems in the training set. Importantly, for the latter, accuracy levels went up to 83% when errors in the reproduced stem were ignored. That is, while the network sometimes struggled to output very strange, phonotactically illegal novel stems, it nevertheless showed a high level of accuracy in outputting an appropriate past tense morpheme. It was able to do so because the Verb gram-

matical class unit and Past Tense target inflection units could form strong connections to the inflectional morpheme region of the output layer. In some respects, this is equivalent to an implementation of a 'rule' for past tense formation (Marcus, 2001). In this way, the MIG improves on the rule induction ability shown by the original Rumelhart and McClelland model.

Figures 2 and 3 contrast the developmental trajectory of MIG for the first 100 epochs of training with corresponding cross-sectional behavioral data from van der Lely and Ullman (2001) for 6-8 year old children, for regular and irregular past tense formation. As training was performed in a non-incremental fashion, we do not take the very early stages of training to be psychologically realistic (see Plunkett & Marchman, 1993). To evaluate the modeling results in light of the empirical data, we identified a window in the training time of the model (epochs 20-70) in which the accuracy rates of the model in the regular past tense were matched to those reported in the developmental study of van der Lely and Ullman (2001). In this time window, the rates of the main error patterns in the simulation results present qualitative similarities to the rates in the empirical data. Once more, compared to the Rumelhart and McClelland model, MIG now combines simulation of correct performance with error patterns.

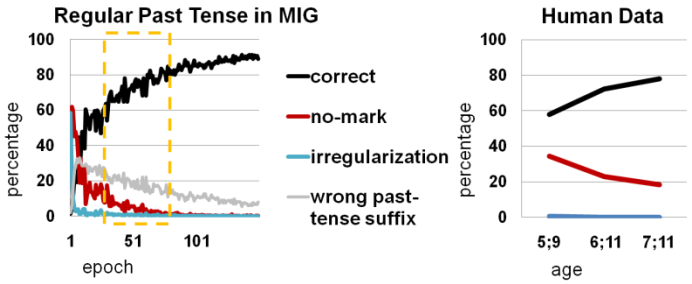


Figure 2: Regular past-tense acquisition in MIG compared to empirical data on from van der Lely & Ullman (2001)

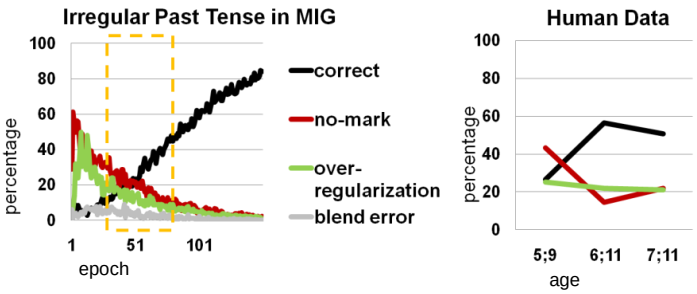


Figure 3: Irregular past-tense acquisition in MIG compared to empirical data on from van der Lely & Ullman (2001)

Simulation 2: Modern Greek Training Set

MIG was also able to learn the complex mappings of the Greek training set. For the perfective past tense, accuracy rates were higher for the sigmatic class than the other conjugational classes. The sigmatic rule was generalized efficiently to novel items (accuracy rates for generalization: 71%).

The model also captured the major developmental error patterns. It simulated an early phase in which 3rd singular forms were produced in inappropriate contexts, which Varlakosta et al. (1998) identified as a marker of the Optional Infinitive stage. It also captured the pattern of overgeneralization of the sigmatic rule in non-sigmatic conjugational classes. Both of these error patterns are depicted in Figure 4, which compares the learning profile of MIG in the 2nd person singular non-sigmatic category (e.g., plen-o / e-plin-es, wash / washed) and corresponding data by Stavrakaki and Clahsen (2009).

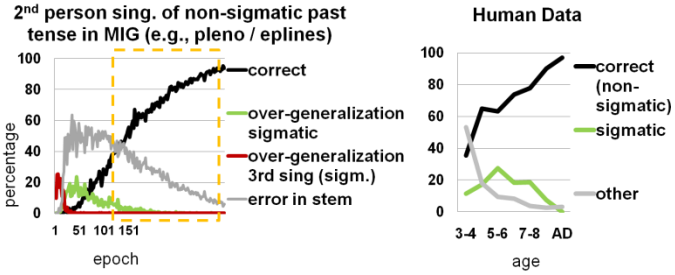


Figure 4: Non-sigmatic perfective past tense in MIG and empirical data from Stavrakaki & Clahsen (2009)

Conclusions

Connectionist approaches to the acquisition of morphology have faced four challenges: to simulate developmental error patterns as well as accuracy levels; to demonstrate that associative systems can generalize inflectional rules to unusual novel stems; to show that architectures can be general across inflection types and grammatical classes, rather than focusing on narrow inflectional paradigms; and to show that architectures can be general across languages, even though those languages may place very different demands on acquisition due to the complexity of their morphology.

In this paper, we introduced the Multiple Inflection Generator. The model is novel in that phonological output forms are conditioned to be appropriate to their grammatical context by the integration of multiple input cues. These input cues include the phonological form of the stem, lexical-semantics, grammatical class, and target inflection information. Cues are relied on differentially depending on the mappings of various inflectional forms (see, e.g., Joanisse & Seidenberg, 1998, for the greater reliance of irregular verbs on lexical-semantic information, also shown by our model).

Focusing on the past tense, we showed how the MIG reproduced error patterns as well as accuracy levels. Notably, in both English and Modern Greek, an Optional Infinitive stage was observed, even though the character of that stage is different in each language (unmarked stems vs. 3rd person singular). Generalization rates of the past tense rule were high for novel stems, even for phonotactically illegal stems. MIG captured the order of emergence of different inflection types for different grammatical classes. And it was able to capture developmental patterns for two languages of different morphological complexity.

These results are only preliminary. More detailed work is required to establish quantitative fits both within and between languages. However, our initial findings demonstrate the viability of a more general, cross-linguistic model of the acquisition of inflectional morphology.

Acknowledgments

This work was supported by UK MRC Grant G0300188 and the European Commission Grant NEST-029088 (ANALOGY) to the second author. The studies of the first author are funded by the Greek State Scholarship Foundation (IKY) and the Greek Ministry of Education.

References

Berko, J. (*1958*). The child's learning of English morphology. *Word, 14*, 150-177.

Brown, R. (1973). *A first language: The early stages.* London: George Allen & Unwin Ltd.

Clahsen, H., & J. Dalalakis (1999). Tense and agreement in Greek SLI: A case study. *Essex Research Reports in Linguistics, 24,* 1-25.

de Villiers, J.G., & de Villiers, P.A. (1985). Acquisition of English. In D. Slobin (Ed.), *The crosslinguistic study of language acquisition: Vol. 1. The data* (pp. 27-140). Hillsdale, NJ: Lawrence Erlbaum.

Francis, W.N., & Kucera, H. (1979). *Manual of Information to Accompany a Standard Sample of Present-day Edited American English , for Use with Digital Computers.* Original ed. 1964, revised 1971, revised and augmented 1979. Department of Linguistics, Brown University, Providence, R.I.

Hoeffner, J. H., & McClelland, J. L. (1993). Can a perceptual processing deficit explain the impairment of inflectional morphology in developmental dysphasia? A computational investigation. In E. V. Clark (Ed.*), Proceedings of the 25th Child Language Research Forum.* Palo Alto, CA: Stanford University Press.

Joanisse, M.F. (2004) Specific language impairments in children: Phonology, semantics and the English past tense. *Current Directions in Psychological Science, 13,* 156-160.

Joanisse, M.F., & Seidenberg, M.S. (1999). Impairments in verb morphology following brain injury: A connectionist model. *Proceedings of the National Academy of Science USA, 96,* 7592-7597.

Kuczaj, S.A. (*1977*). The acquisition of regular and irregular past tense forms. *Journal of Verbal Learning and Verbal Behavior*, 16, 589-600.

Matthews, D. E., & Theakston, A. L.(2006) 'Errors of omission in English-speaking children's production of plurals and the past Tense: The effects of frequency, phonology, and competition', *Cognitive Science: A Multidisciplinary Journal, 30: 6*, 1027 — 1052.

Marcus, G. F. (2000). Children's Overregularization and Its Implications for Cognition. In P. Broeder, & J. Murre (eds). *Models of Language Acquisition: Inductive and deductive approaches.* Oxford: Oxford University Press, pp 154-176.

Marcus, G.F. (2001). *The Algebraic Mind: Integrating Connectionism and Cognitive Science (Learning, Development, and Conceptual Change),* Cambridge, MA: MIT Press.

Pinker, S. (1984). *Language learnability and language development*. Cambridge, MA: Harvard University Press.

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, *28*, 73–193

Plunkett, K., & Juola, P. (1999). A connectionist model of English past tense and plural morphology. *Cognitive Science, 23,* 463-490.

Plunkett, K., & Marchman, V. (1991). U-shaped learning and frequency effects in a multi-layered perceptron: Implications for child language acquisition. Cognition, 38, 1-60.

Plunkett, K., & Marchman, V. (1993). From rote learning to system building: Acquiring verb morphology in children and connectionist nets. *Cognition, 48,* 21-69.

Rice, M.L. (2000). Grammatical symptoms of specific language impairment. In: D.V.M. Bishop & L.B. Leonard (Eds.), *Speech and Language Impairments in Children: Causes, characteristics, intervention and outcome* (pp.17-34). Hove, England: Psychology Press.

Rumelhart, D. E., Hinton, G. E., & Williams, R.J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart, J. L. McClelland and The PDP Research Group, *Parallel distributed processing: Explorations in the microstructure of cognition. Volume 1: Foundations* (pp. 318–362). Cambridge, MA: MIT Press.

Rumelhart, D. E. & McClelland, J. L. (1986). On learning the past tense of English verbs. In J. L. McClelland, D. E. Rumelhart and PDP Research Group (Eds), *Parallel distributed processing: Volume 2: Psychological and Biological models.* (pp. 216-271). Cambridge, MA: MIT Press.

Stavrakaki, S. & Clahsen, H. (2009). The perfective past tense in Greek child language. *Journal of Child Language*, *36*, 113-42.

Stephany, U. (1997). The Acquisition of Greek. In: Slobin, D. I. (ed.). *The cross-linguistic study of language acquisition 4* (pp.183-333).Hillsdale, NJ: Erlbaum.

Thomas, M.S.C. & Karmiloff-Smith, A. (2003). Modelling language acquisition in atypical phenotypes. *Psychological Review, 110, 4,* 647-682.

van der Lely, H. K. J., & Ullman, M. (2001). Past tense morphology in specifically language impaired children and normally developing children. *Language and Cognitive Processes, 16,* 177-217.

Varlokosta, S., Vainikka, A., & Rohrbacher, B. (1998). Functional projections, markedness, and ‚root infinitives' in early child Greek. *The Linguistic Review* 15:187-207.

Handling what the other sees: the effects of seeing and being seen on gesture production

Lisette Mol (L.Mol@uvt.nl)

Emiel Krahmer (E.J.Krahmer@uvt.nl)
Tilburg Centre for Cognition and Communication (TiCC), School of Humanities, Tilburg University
P.O. Box 90135, NL-5000 LE Tilburg, The Netherlands

Abstract

Language production is often argued to be adapted to addressees' needs. As an instance of this, speakers produce fewer speech accompanying hand gestures if the speaker and the addressee cannot see each other. Yet there is also empirical evidence that speakers tend to base their language production on their own perspective, rather than their addressee's. Therefore, speakers may gesture differently because they do not see their addressee, rather than because their addressee cannot see them. Can speakers truly apply their knowledge of what their addressee sees to their gesture production? We answered this question by carrying out a production experiment in which visibility between speaker and addressee was manipulated asymmetrically. We found that representational gestures were produced more frequently when speakers could be seen by their addressee, rather than when they could see their addressee, suggesting that speakers indeed apply their knowledge of the addressee's perspective correctly to their gesturing.

Keywords: Gesturing, Audience Design.

Introduction

Language use sometimes requires taking into account what another person can or cannot see. For example, when watching a documentary on Venice with a friend, you might ask your friend "have you ever been there?", where *there* refers to Venice. However, if your friend was in the same room, but working on her computer "have you ever been to Venice?" may be more appropriate. Because you know your friend is not watching the documentary, you may choose a more explicit reference. On the other hand, if you were asked by your friend, "have you ever been there?", while working on your computer, your knowledge of her watching a documentary on Venice may help in arriving at the correct interpretation. Yet would you do so correctly if you happened to be browsing a website on Cologne?

Language production is often argued to be adapted to the needs of addressees (e.g. Grice, 1989). As an instance of this, it is well established that speakers produce fewer speech accompanying hand gestures when interlocutors cannot see each other (Cohen & Harison, 1973). Yet several empirical studies suggest that applying knowledge of what another person can and cannot see is not at all straightforward (e.g. Keysar, Lin, & Barr, 2003; Wardow Lane, Groisman, & Ferreira, 2006). These studies suggest that interlocutors tend to base their language use on their own perspective, rather than that of their conversation partner.

To our knowledge, in studies on hand gestures, visibility has always been manipulated symmetrically. That is, whenever the addressee could not see the speaker, neither was the speaker able to see the addressee. Therefore, these studies cannot reveal whether it is the speaker's own perspective that underlies this reduction in gesture frequency, or whether speakers adapt their language use to their addressee's perspective. In this study we aim to fill this gap, by manipulating visibility asymmetrically. For this we make use of computer-mediated communication. We will therefore also make a comparison of our data in computer-mediated settings to data acquired in similar unmediated settings (Mol, Krahmer, Maes, & Swerts, 2009).

Taking into Account what an Interlocutor sees

Keysar, Lin, and Barr (2003) have shown that people make 'mistakes' in interpreting speech, when deriving the correct interpretation requires applying ones knowledge of what the speaker does not see. In their study, a follower had visual access to an object that was occluded from the director's view. Still, when the (confederate) director's description more closely resembled the hidden object than any of the mutually visible objects, the follower often considered this object as a referent, sometimes even moving it instead of the intended object. This shows that the follower's knowledge of what the director could (not) see was not automatically applied to the interpretation process.

Wardlow Lane, Groisman, and Ferreira (2006) found similar results for reference production. In their study a speaker had private visual access to an object that only differed from the target object in size. Even though the addressee could not see this competing object, speakers often included a contrasting adjective, such as 'small' in their reference to the target object. Surprisingly, they did so even more when instructed to conceal their private information from the addressee. Thus, it seems that speakers have difficulty in applying their knowledge of what their addressee can see to the speech production process as well.

Gesturing out of Sight

The question naturally arises whether knowledge of what another person sees is applied correctly to the production of

co-speech hand gestures. These gestures are spontaneous movements of the hands and arms during speech (e.g. McNeill, 1992). Hand gestures can, amongst other functions, be communicative. For example they can convey meaning (e.g. Beattie & Shovelton, 1999) or emphasize certain parts of speech (e.g. Hadar, 1989; Krahmer & Swerts, 2007). It has been found repeatedly that speakers' gesturing differs depending on whether their addressee can see them or not (e.g. Alibali, Heath, & Myers, 2001; Bavelas, Gerwing, Sutton, & Prevost, 2008; Cohen & Harison, 1973). For example, Alibali et al. asked participants to retell the story of an animated cartoon to an addressee. During half of the narration, an opaque screen separated speaker and addressee, such that no information could be conveyed through hand gestures. They found that speakers gestured less frequently when the screen was in place. This was especially true for *representational gestures*, which depict some of the content a speaker is trying to convey. It thus seems that at least some gesturing is influenced by the speaker's knowledge of what the addressee can and cannot see.

However, in the studies cited above, visibility was always manipulated symmetrically. That is, the addressee could not see the speaker, but neither could the speaker see the addressee. It is thus possible that speakers used their own perspective, and that their gesturing changed as a result of them not seeing the addressee, rather than of them correctly applying their knowledge of what the addressee could see. If so, many other factors may have influenced gesture production, such as perceived attentiveness of the addressee, social fulfillment during the task, general motivation, etc. Indeed, Jacobs and Garnham (2006) found that people gesture less frequently towards an addressee who appears to be less interested. Interest can be conveyed by gaze (Argyle & Cook, 1976), and also by body posture and head nods, which are all absent if visibility is obstructed. It is therefore still unclear whether the reduced frequency of hand gestures when interlocutors cannot see each other is an instance of the correct application of the knowledge the speaker has about the addressee's visual perspective.

Desktop Video-Conferencing

One way to manipulate visibility in an asymmetrical way is by computer-mediated communication. Yet is mediated communication representative of unmediated communication? Brennan and Oheari (1999) found evidence that mediated communication may differ from unmediated communication as a direct result of the differences in affordances between the media, rather than for example because interlocutors become less socially aware when they are not physically copresent. In typing - which is often used in mediated communication - different types of communicative behavior are effortful than in speech. Brennan and Oheari found that especially back-channeling behavior differed between spoken and written dialogue. This in turn may affect interlocutors' perception of each other, rather than them not being physically co-present. Thus, the more affordances mediated communication offers, the more similar it will be to unmediated communication.

Modern video-conferencing tools allow speakers to see and hear each other even though they are in different locations. Isaacs and Tang (2003) observed interactions between technical experts that took place over the phone, through desktop video-conferencing, or face-to-face. They found that the experts used the visual modality in video-conferencing much like they did in face-to-face communication. "Specifically, participants used the visual channel to: express understanding or agreement, forecast responses, enhance verbal descriptions, give purely nonverbal information, express attitudes through posture and facial expression, and manage extended pauses", p. 200. They also list some differences between video-conferencing and face-to-face communication, for example, managing turn-taking, having side conversations, and pointing towards objects in each other's space were more difficult in video-conferencing.

In the video-conferencing we use, interlocutors can communicate through speech as though they are in the same room. The need for turn-taking is minimal, and there are only two interlocutors. Also, our task is not about manipulating the environment, which reduces the factor of not sharing a workspace. We therefore expect that manipulating mutual visibility will have similar effects in our mediated settings as it does in unmediated settings. But more readily than unmediated communication, video-conferencing enables one-way visibility, allowing for example the speaker to see the addressee, but not vice versa. It is thus very suitable for testing whether or not speakers employ an egocentric perspective when they cannot see their addressee.

Present Study

In this study we aim to gain insight into whether people generally employ an egocentric perspective in their language production. We address this question by testing if speakers' knowledge of whether their addressee can see them or not influences their co-speech gesturing. We manipulate visibility asymmetrically. That is, some speakers will be able to see their addressee, but will know that the addressee cannot see them, and some speakers will not be able to see their addressee, but will know that the addressee can see them. If gesturing is based on the speaker's own visual perspective, then gesturing will be more frequent when speakers can see the addressee, regardless of whether the addressee can see them. This could be either because the addressee seems more engaged or more present when visible, or because from the speaker's visual perspective, it seems as though speaker and addressee can see each other. Yet if speakers correctly apply their knowledge of the addressee's visual perspective, then they are expected to

gesture more when the addressee can see them, regardless of whether they can see the addressee. If both of these factors increase gesture production, then gesturing should be most frequent when interlocutors can see each other.

Method

Design

We have used a 2 x 2 between subjects design in which we manipulated whether or not the addressee could see the speaker and whether or not the speaker could see the addressee. In all conditions speaker and addressee could hear each other.

Participants

38 (21 female) native Dutch speakers, all students of Tilburg University, participated in this study as part of their first year curriculum. Two participants were excluded from our analysis (see Coding and Analysis). The remaining 36 participants (20 female) had a mean age of 22, range (18 - 30). The addressee was a female confederate, who was also a student at Tilburg University.

Procedure

The participant and the confederate were received in the lab by the experimenter, who assigned the role of speaker to the participant and the role of addressee to the confederate. Like in the study by Alibali et al. (2001), narrators were asked to retell the story of an animated cartoon (*Canary Row* by Warner Bro's). After reading the instructions participants could ask any remaining questions. (The confederate always posed a question.) The narrator's instructions stated that the addressee had to summarize the narration afterwards and explained that the narrator was videotaped in order to compare the summary to the narration afterwards.

When all was clear the narrator was seated behind a table with a computer screen on it, which in some settings showed a live video-image of the addressee, and in the remaining settings showed the interface of a video-conferencing application (Skype). The screen was connected to a pc, which also had a web cam connected to it. Behind the table stood a tripod, which held the web cam and a digital video camera. On the wall behind the video camera were eight stills from the animated cartoon, one from each episode, as a memory aid for the narrator and to elicit more structured and hence more comparable narrations.

The experimenter took the addressee to another room with a similar setup (but without the stills) and established a connection between the two pc's over the internet, using Skype. Sound and video were both captured by the web cams and sound was played back through speakers. Sound was tested by the narrator and addressee talking to each other and if applicable, the video image was tested by them watching each other. The connection was then suspended temporarily while the narrator was left alone to watch the animated cartoon on a different computer. When the cartoon had finished the experimenter re-established the connection, and seated the narrator behind the camera. The experimenter repeated whether the addressee could see the narrator or not, started the video recording, and left the room.



Figure 1: Left: example of a representational gesture (depicting hitting), Right: example of a non-representational gesture (placing emphasis while referring to a character).

When the narrator was done telling the story, a questionnaire followed, which included questions on how the communicative setting had been experienced, how interested the addressee had appeared, whether any deception was suspected, and finally whether the participant was left or right handed. Meanwhile, the addressee ostensibly wrote a summary on yet another computer in the lab room. None of the participants had suspected any deception. After filling out the questionnaire, they were fully debriefed. All of the participants gave their informed consent for the use of their data, and if applicable for publishing their photographs.

During the narration, the confederate refrained from interrupting, laughing, etc. When necessary, minimal feedback was provided verbally. She always gazed somewhere near the web cam capturing her, independent of whether she could see the speaker.

Coding and Analysis

Video recordings of all narrators were coded using Noldus Observer. For each movement of the hands it was determined whether the movement was a gesture or a self-adaptor. Gestures were labeled as either *representational*, expressing some of the content of the speaker's story, or *non-representational*, placing emphasis or regulating interaction. Figure 1 depicts two examples. In the scene on the left, the speaker imitates a hitting motion while talking about someone hitting. In the scene on the right, the speaker refers to the main character and briefly moves his fingers up and down. In order to normalize for the duration of each speaker's narration, we have used the number of gestures produced per minute as the dependent variable, rather than the total number of gestures produced.

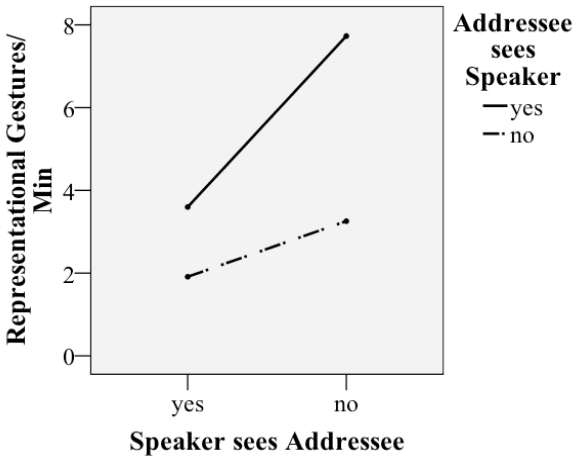


Figure 2: Means of the rate of representational gestures across settings.

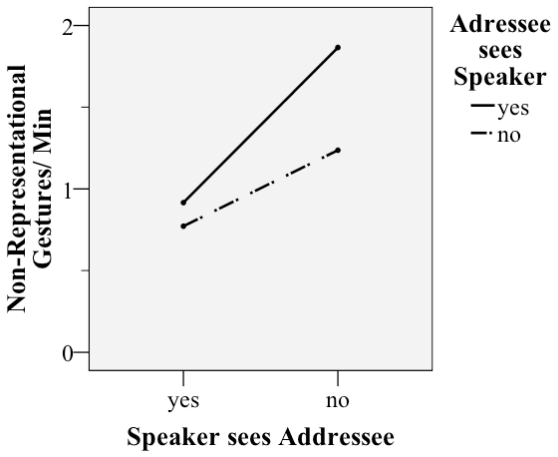


Figure 3: Means of the rate of non-representational gestures across settings.

The analysis was done using ANOVA, with fixed factors *addressee sees speaker* (yes, no) and *speaker sees addressee* (yes, no). Our significance threshold was .05 and we have used partial eta squared as a measure of effect size.

Two participants were excluded from the analysis, because they deviated more than 2 standard deviations from the mean gesture rate in their condition. As a result, there were 9 participants in each condition. Inclusion of these two participants did not affect the significant effects found, but did reduce the significance of the overall model.

Results and Discussion

We did not find an effect of gender or left or right handedness on gesture rate, or on the total duration of the narration. Neither did we find an effect of condition on the duration of the narration.

Effect of the Addressee seeing the Speaker

Figure 2 shows the mean number of representational gestures per minute in each setting. Whether or not the addressee could see the speaker reliably influenced this gesture rate, $F(1, 32) = 4.873$, $p < .05$, $\eta^2 = .13$. When speakers could be seen by the addressee, they produced representational gestures more frequently ($M = 5.7$, $SD = 5.8$) than when they could not be seen ($M = 2.6$, $SD = 3.4$). We found no significant effect of visibility of the speaker on the rate of non-representational gestures ($p = .35$).

Effect of the Speaker seeing the Addressee

The effect of whether the speaker could see the addressee approached significance for the rate of representational gestures, $F(1, 32) = 3.854$, $p = .06$, $\eta^2 = .11$. When speakers could see their addressee, they produced these gestures *less* frequently ($M = 2.8$, $SD = 3.4$) than when they could not see their addressee ($M = 5.5$, $SD = 5.3$). There was no significant interaction between visibility of the speaker and addressee ($p = .33$).

The mean number of non-representational gestures in each condition is depicted in Figure 3. The effect of the speaker seeing the addressee on this gesture rate showed a trend towards significance, $F(1,32) = 2.977$, $p = .09$. Non-representational gestures were produced less frequently when speakers could see their addressee ($M = .84$, $SD = .87$), compared to when they could not ($M = 1.6$, $SD = 1.5$). There was no significant interaction with the addressee seeing the speaker ($p = .56$).

Perceived Interest

Our questionnaire revealed that in the setting in which the speaker could see the addressee but not vice versa, the addressee was perceived as significantly more uninterested than in any of the other conditions, $F(3, 31) = 5.232$, $p < .01$, see Table 1. (Pairwise comparisons were done using the LSD method with a significance threshold of .05.)

Discussion

When the addressee could see the speaker, speakers produced representational hand gestures more frequently than when the addressee could not see them. This was true both when the speaker could see the addressee and when

Table 1: *Means* and *Standard Deviations* of speakers' answer to the statement "The addressee was disinterested" on a 7 point scale, 1 = completely disagree, 7 = strongly agree.

Addressee sees Speaker	Speaker sees Addressee	Mean, *SD* of Perceived disinterest (1 to 7 scale)
Yes	Yes	2.7, 1.0
Yes	No	3.3, 1.3
No	Yes	4.5, 1.2
No	No	2.4, 1.1

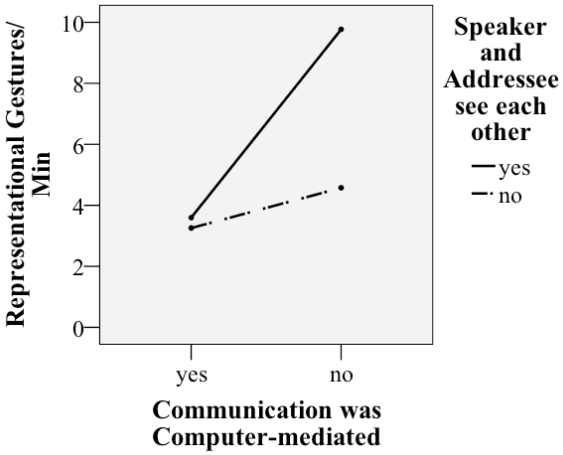


Figure 4: Means of the rate of representational gestures in mediated and unmediated settings.

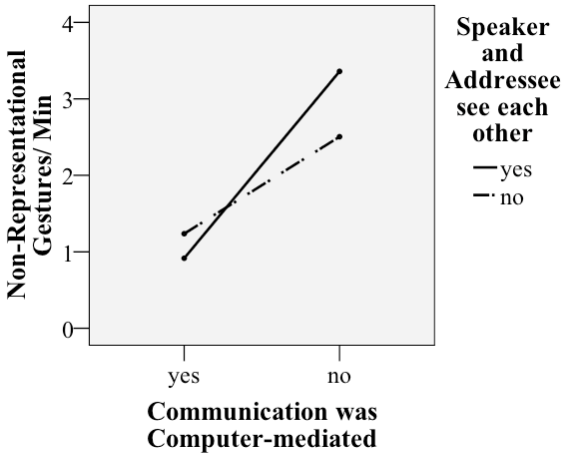


Figure 5: Means of the rate of non-representational gestures in mediated and unmediated settings.

not. We therefore conclude that the knowledge a speaker had about what the addressee could and could not see was incorporated correctly into their hand gesture production.

We found an unexpected effect when speakers could see their addressee. When they saw a live video-image of their addressee, speakers produced representational gestures less frequently and also tended to produce non-representational gestures less frequently than when they did not see their addressee. This would be understandable if the addressee came across as less interested when visual feedback was provided. In the setting in which the addressee could not see the speaker, there was nothing relevant to look at for the addressee. To keep the settings comparable, the addressee therefore always gazed somewhere near the web cam capturing her. This may have been interpreted as lack of interest. The answers to our questionnaire support this hypothesis. In the setting in which the speaker could see the addressee but not vice versa, the addressee was rated as significantly less interested than in all other settings.

Mediated vs. Unmediated Settings

In the study above, we manipulated visibility by means of computer-mediated communication. In an earlier study (Mol et al. 2009), we have manipulated visibility while speaker and addressee were in the same room. The procedure was the same as in the current study, except that the speaker and addressee were in the same room facing each other (N = 10), or in the same room but separated by an opaque screen (N = 9). Given that the affordances in these mediated and unmediated settings are a close match, it is interesting to see whether there still is an effect of computer-mediation. To address this question we compare the mediated settings with mutual visibility and with audio only to their unmediated counterparts. Participants were mostly first year students of Tilburg University and all were native speakers of Dutch. The mean age was 19, range (17 – 21), and 15 out of 19 participants were female.

Effect of Visibility

The gesture rates across settings for representational gestures are depicted in Figure 4. The main effect of visibility on this gesture rate approached significance, $F(1, 33) = 4.1$, $p = .05$. Participants gestured more frequently when they could see each other ($M = 6.8$, $SD = 6.1$) than when they could not ($M = 3.9$, $SD = 2.3$). There was no significant effect of mutual visibility on the rate of non-representational gestures ($p = .65$).

Effect of Mediation

Mediation had a significant main effect on the rate of representational gestures, $F(1, 33) = 7.579$, $p < .01$. The interaction between mutual visibility and mediation showed a trend towards significance, $F(1, 33) = 3.180$, $p = .08$. The difference between the visibility and no visibility condition was larger in the unmediated settings.

Mediation also influenced the rate of non-representational gestures, $F(1, 33) = 10.330$, $p = .01$. Non-representational gestures were produced more frequently in the unmediated settings ($M = 3.0$, $SD = 2.2$), compared to the mediated settings ($M = 1.1$, $SD = 1.1$). There was no significant interaction between the factors ($p = .32$). The gesture rates for non-representational gestures are depicted in Figure 5.

Perceived Interest

The effect of the setting on how disinterested the addressee was perceived showed a trend towards significance, $F(3, 33) = 2.288$, $p = .097$. Table 2 (next page) shows the means and standard deviations for this measure in each setting. Pairwise comparisons with the LSD method showed that addressees were perceived as less interested in the unmediated setting without visibility, compared to the unmediated setting with visibility and the mediated setting without visibility, $p < .05$.

Table 2: *Means* and *Standard Deviations* of speakers' answer to the statement "The addressee was disinterested" on a 7 point scale, 1 = completely disagree, 7 = strongly agree.

Mutual Visibility	Computer-Mediation	Mean, *SD* of Perceived disinterest (1 to 7 scale)
Yes	Yes	2.7, 1.0
Yes	No	2.6, 1.1
No	Yes	2.4, 1.1
No	No	3.6, .73

Discussion

Whether or not communication was computer-mediated affected gesture production. Participants gestured more frequently in the unmediated settings. In the unmediated settings, seeing each other seemingly only increases gesture production. Yet in the mediated setting with mutual visibility, two factors may act in opposite directions. Our previously discussed results showed that in the mediated setting, being seen by the addressee increases gesture production, whereas seeing the addressee decreases gesture production. This may explain why participants gestured less frequently in the mediated setting. However, we did not find a difference in perceived interest of the addressee between the mediated and unmediated setting with mutual visibility.

Another possible explanation is a difference in affordances between mediated and unmediated communication (Brennan & Ohaeri, 1999). Even though one of the mediated settings offered live audio and video, narrators produced fewer gestures than in a face-to-face setting. The most notable difference between these two settings may be that the mediated setting did not enable interlocutors to look each other in the eyes. One either looks at the camera, or at the eyes of the other person, such that mutual gaze never occurs. We intend to address this factor in a follow-up study, by using a mediated setting that does allow for mutual gaze. Other factors such as not sharing a physical space may also be of influence, especially for pointing gestures (Isaacs & Tang, 2003).

General Discussion and Conclusion

Although our results suggest that several factors interact in our mediated settings, we found a clear effect of whether the addressee could see the speaker. Speakers produced representational hand gestures more frequently when they could be seen by their addressee, rather than when they could see their addressee, suggesting that speakers adjusted their gesturing to the addressee's perspective correctly. This is not to say that they never make mistakes in taking into account what their addressee can and cannot see during language production. Yet our results cannot be explained by assuming that speakers predominantly base their gesture production on their own visual perspective. Rather, they apply their knowledge of what the addressee can see correctly to their hand gesture production.

Acknowledgements

We like to thank Nelianne van den Berg for her help in collecting and coding the data. We also thank all participants as well as the anonymous reviewers.

References

Alibali, M. W., Heath, D. C., & Myers, H. J. (2001). Effects of visibility between speaker and listener on gesture production: some gestures are meant to be seen. *Journal of Memory and Language, 44*, 169-188.

Argyle, M., & Cook, M. (1976). *Gaze and mutual gaze*. Cambridge: Cambridge University Press.

Bavelas, J., Gerwing, J., Sutton, C., & Prevost, D. (2008). Gesturing on the telephone: Independent effects of dialogue and visibility. *Journal of Memory and Language, 58*, 495-520.

Beattie, G., & Shovelton, H. (1999). Mapping the range of information contained in the iconic hand gestures that accompany spontaneous speech. *Journal of Language and Social Psychology, 18*, 438-462.

Brennan, S. E., & Ohaeri, J. O. (1999). Why do electronic conversations seem less polite? the costs and benefits of hedging. *SIGSOFT Softw. Eng. Notes, 24*(2), 227-235.

Cohen, A. A., & Harison, R. P. (1973). Intentionality in the use of hand illustrators in face-to-face communication situations. *Journal of Language and Social Psychology, 8*, 211-288.

Grice, P. (1989). *Studies in the Way of Words*. Cambridge MA: Harvard University Press.

Hadar, U. (1989). Two types of gesture and their role in speech production. *Journal of Personality and Social Psychology, 8*, 211-228.

Isaacs, E. A., & Tang, J. C. (2003). *What video can and can't do for collaboration: a case study.* Paper presented at the Multimedia, Anaheim, CA.

Jacobs, N., & Garnham, A. (2006). The role of conversational hand gestures in a narrative task. *Journal of Memory and Language, 26*, 291-303.

Keysar, B., Lin, S., & Barr, D. J. (2003). Limits on theory of mind use in adults. *Cognition, 89*, 25-41.

Krahmer, E., & Swerts, M. (2007). The effects of visual beats on prosodic prominence: acoustic analyses, auditory perception and visual perception. *Journal of Memory and Language, 57*(3), 396-414.

McNeill, D. (1992). *Hand and Mind: what gestures reveal about thought*. Chicago and London: The University of Chicago Press.

Mol, L., Krahmer, E., Maes, A., & Swerts, M. (2009). The communicative import of gestures: Evidence from a comparative analysis of human-human and human-computer interactions. *Gesture, 9*(1), 97-126.

Wardow Lane, L., Groisman, M., & Ferreira, V., S. (2006). Don't talk about pink elephants: speakers' control over leaking private information during language production. *Psychological Science, 17*(4), 273-277.

On Attractiveness of Surprising Ideas: How Memory for Counterintuitive Ideas Drives Cultural Dynamics

M. Afzal Upal (Afzal.Upal@drdc-rddc.gc.ca)
Adversarial Intent Section
Defence Research & Development Canada (DRDC) Toronto
1133 Sheppard Ave W, Toronto, M3M 3B9

Abstract

The emerging field of cognition and culture has had some success in explaining the spread of counterintuitive religious concepts around the world. However, researchers have been reluctant to extend its findings to explain the widespread occurrence of counterintuitive ideas in general. This article suggests a way to generalize the minimal counterintuitive hypothesis, which argues that such ideas spread because they are more memorable, to form the outline of a model of cultural dynamism which can help explain why strange and novel ideas spread more quickly than ordinary seeming traditional ideas.

Keywords: ideology, shared beliefs, counterintuitiveness.

Introduction

Why do some aspects of group ideologies and cultural worldviews change over time while others stay unchanged for long periods of time? What explains the patterns of persistence and change in shared beliefs of social groups such as new religious movements and political parties? The cognition and culture researchers argue that any attempt to satisfactorily answer such questions must take the individual cognitive tendencies for communication, comprehension, and belief revision into account (Sperber, 1996). A key finding of this research has been the minimal counterintuitiveness hypothesis (Boyer, 1994, 2001) which suggests that the reason why minimally counterintuitive concepts, such as God and ghosts, dominate religious concepts is that people remember them better than intuitive and maximally counterintuitive ideas. This article first reviews the minimal counterintuitiveness hypothesis and then argues that it can be used to explain the spread of novel ideas in general and not just in the context of religious ideas.

The Minimal Counterintuitiveness (MC) Hypothesis

The minimal counterintuitive (MC) hypothesis posits that:

1. Most of the widespread religious concepts around the globe are minimally counterintuitive.
2. The minimally counterintuitive (MCI) concepts that violate a small number of intuitive expectations (such as, a talking tree, a rock that eats, and an invisible cow) are more memorable than either intuitive concepts (such as, a green tree, a brown rock, and a good person) or maximally counterintuitive concepts that violate a larger number of intuitive expectations (such as, an invisible talking tree that does not occupy any space and a sad illuminant travelling rock).

While a number of subsequent empirical studies (Atran, 2004; J. Barrett & Nyhof, 2001; Boyer & Ramble, 2001; Gonce, Upal, Slone, & Tweney, 2006; Upal, 2005a; Upal, Gonce, Tweney, & Slone, 2007) have found some support for better memory for the MCI concepts, some cultural scientists (Bloch, 2005; Harris & Koenig, 2002; Keller, 2004) have argued that a number of widespread religious concepts such as Gods and ghosts are maximally counterintuitive and not minimally counterintuitive as implied by the minimal counterintuitiveness hypothesis. Some cognitive scientists of religion (J. L. Barrett, 1997, 1999; J. L. Barrett & Keil, 1996; Slone, 2004) have responded by suggesting that this is because believers hold two different ("theologically correct" and "intuitive") conceptualizations of God and that only the intuitive conceptualizations enjoy the transmission advantages because they are the only ones that are minimally counterintuitive. Barrett (1997, Page 124) says:

> *God, and perhaps other religious objects and entities, are conceptualized on at least two different levels: the basic, everyday concept used in real-time processing of information, and the "T.C." or theologically correct level used in theological discussion of God's properties or activities outside of a real-time context. As was shown in above, these two levels of conceptualization may represent God in substantially different ways.*

Thus, argue these cognitive scientists of religion, that the MC hypothesis "does not apply" to the theological conceptualizations of God or to any other cultural concepts that do not involve violating expectations of intuitive reflective thinking (J. L. Barrett, 1997) (Page 127). This includes ideas that have been learned through explicit training such as the socio-cultural and religious schemas, scripts, and scientific concepts (J. L. Barrett, 2008). Another hurdle in the applicability of the MC-hypothesis to the spread of the cultural beliefs in contemporary social groups is the often implicit assumption that the MC-hypothesis is only applicable to societies where oral transmission is the primary source of the transmission of cultural information. Since most of the modern cultural ideas are spread through pen, paper, and the internet the MC-hypothesis may not apply to them.

Previously (Upal, 2009a), I have argued against this narrow interpretation of the MC-hypothesis and suggested that memory advantages obtained by violating conceptual expectations should not be limited to "intuitive concepts". Instead, I argued that ideas that violate cultural schemas,

scripts, and expert knowledge acquired through learning should also enjoy memorability advantages. While details of the context-based view of the MC effect has been specified elsewhere (Upal, 2005a, 2007a, 2007b, 2009a, 2009b; Upal, et al., 2007), here I review its salient points as they relate to the development of group ideologies.

The Minimal Counterintuitiveness Effect and the Distintiveness Effect

One of the most robust findings in experimental psychology has been the so called *distinctiveness effect* which indicates that an item, that stands out as compared to other items in its context, is more likely to be remembered than those other items (Hunt & Worthen, 2006). For over a century, experimental psychologists working with a variety of stimuli have found support for this effect (Calkin, 1894, 1896; McDaniel, Dornburg, & Guynn, 2005; von Restorff, 1933). Thus unexpected events and entities in a story are recalled better than expected events and objects (Davidson, Larson, Luo, & Burden, 2000; Kintsch & Green, 1978; Upal, 2005a), bizarre images are recalled better than ordinary images (McDaniel, Einstein, DeLosh, & May, 1995), unexpected words in a list of words are recalled better than expected words (Atran, 2004; von Restorff, 1933), orthographically distinct words are recalled better than ordinary words, as are typographically distinct words (Hunt & Worthen, 2006). Cognitive scientists and evolutionary psychologists argue that the distinctiveness effect reveals the evolutionary pressures that guided the evolution of animal and human memory systems. They suggest that distinctiveness effect supports the view that the ability to predict relevant aspects of one's environment was the primary driver for the evolution of animal and human memory systems. People use the knowledge of their environment to generate expectations about other hitherto unobserved aspects of the environment (Schank, 1975, 1979, 1999; Schank & Abelson, 1977). If these expectations are not fulfilled, it indicates a gap in the agent's world model. Agents whose memory systems treat expectation-violations as learning opportunities to revise their world model to make them more accurate stand to gain evolutionary advantages in terms of being able to collect more food or find better mates.

In (Upal, 2005a), I argued that Schank's learning theory and findings in psycholinguistics (Graesser, Singer, & Trabasso, 1994; Kintsch, 1998) explain that minimally counterintuitive ideas are remembered better than intuitive ideas because they violate a reader's expectations. Psycholinguists argue that when reading a text people primarily ask why questions i.e., why did the author include this information in this text? The cognitive processes of readers accessing the knowledge structures in their long term memory to construct a justification for the inclusion of ideas in question result in establishment of strong memory links between counterintuitive ideas and thematic cues about the story. When these cues are presented to subjects, the strongly connected minimally counterintuitive concepts are easily retrieved and recalled. I hypothesized that for the minimally counterintuitive ideas, readers are able to construct such justifications and create a coherent concept but that readers fail in their effort to construct a justification and create a coherent concept for maximally counterintuitive ideas. The *memorability hypothesis* (Upal, 2005a) suggests that memorability of a concept in a context is a function of the difference between its degree of expectation violation and its coherability as a new concept.

Besides explaining the past observations of why minimally counterintuitive ideas are better remembered than intuitive and maximally counterintuitive ideas, the memorability hypothesis makes a number of predictions. Since proposing this model (Upal, 2005a), I and others have conducted a number of empirical experiments and found that results generally support a context-based view of the minimal counterintuitiveness effect (Gonce, et al., 2006; Upal, 2007a, 2007b, 2009b; Upal, et al., 2007; Upal & Harmon-Vukic, 2010). It predicts that, on average, readers should spend more time to process counterintuitive concepts than they do in processing intuitive concepts. This is because counterintuitive concepts trigger cognitively taxing process of justification creation while intuitive concepts do not. A recent study has confirmed this finding (Upal & Harmon-Vukic, 2010).

The context-based model also posits that counterintuitiveness is a property of the context in which a concept appears as much as it is a property of the concept itself. The context includes the mental knowledge that the reader brings to the table as well as the prior parts of the text in which the concept is embedded. This means that the same concept may appear more unexpected in context A than in context B and that the same concept may be more memorable in one context and less memorable in another context. Since knowledge structures in people's memories change over time, the same concept may be more counterintuitive for a person at a time t_1 than at a time t_2. A one-time exposure to an idea, however, does not guarantee that the idea will not seem counterintuitive in the future. In order for an idea to lose memorability advantages, the knowledge in long term memory that generated the expectation has to be revised so as to make the counterintuitive idea as the new expected and the old idea as the new unexpected (and therefore the new counterintuitive). Since knowledge structures in memory are richly connected with each other revising them requires significant cognitive resources to untangle old connections and establishing new ones. Thus it is not surprising that people are very conservative when it comes to revising their beliefs. People's expectations guide what they see leading them to sometimes miss the unexpected objects and events. When the evidence of expectation violations is too overwhelming to ignore, they prefer to generate elaborations that allow them to preserve as much of their old beliefs as possible. Even though observing a single instance of a counterintuitive object or event can (at least in principle) trigger belief change, this does not happen very often. For

instance, upon seeing an ostrich for the first time, one may no longer be surprised when one hears of, "a healthy adult bird that cannot fly" assuming one can create a justification that an ostrich is still a bird because it has feathers but is not able to fly because it is too heavy. Creation of justifications in response to seeing an unexpected object or event does not automatically lead to generation of different expectations in a similar future context. One may for instance assume that the expectation violation only happens in an overly restricted context, for instance, assume that ostriches do not fly on Tuesdays between 9 and10 am or that the ostrich under observation is a mutant. Seeing a healthy adult ostrich at a different time in the future may still lead to the expectation that it will fly. It may take prolonged exposure to numerous observations of unexpected objects and events and significant cognitive effort for someone to revise enough knowledge structures in their long term memory for them to generate new expectations. Once all the relevant memory structures are revised and the old unexpected becomes the new expected, the once minimally counterintuitive idea should no longer be so. Thus the context-based model predicts that minimally counterintuitive ideas should lose their memorability advantages over time.

Since the context-based model does not support differential processing for mental knowledge acquired through intuitive and doctrinal modes of thinking, it predicts that violations of online intuitive cognition should not have a privileged status, at least when it comes to memorability. Thus, ideas that violate expectations generated by offline learned concepts such as cultural schemas and religious doctrine should also be better remembered than ideas that do not violate such expectations.

The context-based view emphasizes the role played by the knowledge that an individual possesses when processing a concept in making a concept a concept minimally counterintuitive. This means that a concept that is minimally counterintuitive for one person may not be minimally counterintuitive for another person whose mental knowledge differs from that of the first person. If counterintuitiveness is not the property of the concept alone, then a concept can only appear minimally counterintuitive to a population if individuals within the population share beliefs that are relevant to the concept i.e., if the concept violates the expectations raised by those shared beliefs and if the expectation violation can be justified using those shared beliefs. Thus contrary to the traditional view that ideas that violate cultural schemas should not have memorability advantages, the context-based view suggests that *they should.* I will refer to such ideas as *socially counterintuitive* and point out the role that they play in constantly reshaping the fabric of cultural beliefs.

Social Counterintuiveness

I define an idea as *minimally socially counterintuitive* for a population if it violates a single expectation generated by beliefs shared by that population. Thus the notion of a person remembering details of her past lives may be minimally counterintuitive to a western population that may have a passing familiarity with the idea of reincarnation but not to a Hindu population among whom the belief in reincarnation is intricately woven into the fabric of socially shared beliefs. Minimally counterintuitive social ideas have a memorability advantage over intuitive cultural ideas that do not violate any expectations generated by shared cultural beliefs. Thus the notion of a person who remembers her past life would have a memorability advantage in a western population that did not expect the idea but can use their passing knowledge to understand it. However, it will not enjoy memorability advantages due to counterintuitiveness in a Hindu population where it is already well entrenched.

Similar to the case with individual counterintuitiveness, socially counterintuitive ideas can also become socially intuitive overtime but the process is far more difficult and involved because it involves changes in shared beliefs of a large number of individuals. As advocates of social change would attest, getting a new idea to become widely accepted by a population is a long and painstaking process that requires years of effort by dedicated individuals. This is because, similar to ideas in individual memories, shared cultural ideas are like a well-knit fabric and once this fabric is ripped up by an expectation violating concept, a number of threads become exposed. All of these threads have to be stitched together in new and innovative ways to fully mend the fabric such that the new idea becomes culturally expected. This is why cultural conceptual change faces such daunting prospects requiring years, if not a lifetime, of effort by social leaders who dedicate their lives to the issue. Previously, I have referred to such social leaders as information entrepreneurs (IEs) (Upal, 2005b) because to successfully lead conceptual change, these leaders have to posses the following characteristics.

- They must have high social capital in the group whose shared beliefs they are trying to change. This is needed both to have the credibility needed to persuade others and also because they can afford to be seen as dissenting from group-think (Packer, 2008).
- They must have the marketing skills required to sell the conceptual change to their target audience. Like all good marketers, they are able to make their ideas seem as inevitable as ideas whose time has come.
- They must have the cognitive skills required to integrate the seemingly counterintuitive idea with the group's traditional thinking and make it seem as if the new idea is intuitive and perfectly in line with the group's original thinking.

In (Upal, 2005b), I argued that the IE view helps us understand that new religious movement leaders create seemingly counterintuitive ideas because they believe that these ideas are needed to solve problems being faced by the group. Upal (2005c) argued that revision in socially shared beliefs is driven by a belief among one or more of the strongly identified group members that the group's shared beliefs are harmful to the long-term prosperity of the group.

I focused on social identity beliefs which include, "who belongs to the group and who does not, who is admitted to the group, and who is not? This is particularly clear for racist, ethnocentric, xenophobic or nationalist ideologies, according to which only 'we, white Europeans' belong in Europe, and others should not be admitted, at least not as (equal) citizens" (van Dijk, 1995) (Page 250). Anthropologists studying ethnic groups find that ethnocentric beliefs in "superiority of the ingroup's culture combined with condemnation of the outgroup as immoral and inferior" are "commonplace (e.g., (LeVine & Cambell, 1972)). 'Chosenness' is a particularly prominent expression of this belief" (Page 6). Van Evera (1994) argues that such chauvinist myths are "hallmark of nationalism, practiced by nearly all nationalists to some degree" (Page 27). He provides a number of illustrative examples including Nazi myth of Aryan supremacy, British and American beliefs in rational and intellectual exceptionalism (Longley, 2003), and Russian belief in their extra-ordinary inventiveness. These could be complemented by Pakistani belief that one Pakistani Muslim soldier can dominate 10 Indian Hindu soldiers, American Indian belief that they are more spiritual than the more material "white man", Israeli belief that they are more rational than crazy Arabs, Muslim belief that God chose to favor them as his final chosen people after Christians and Jews strayed from the prescribed path, and the Nation of Islam belief that an evil black scientist created the wicked white man. Group superiority myths are reflected in the literature and art of a group and feature prominently in its creation stories that form the master narrative of a group.

Social psychologists argue that such beliefs are necessary for people's well being since people have a fundamental need to feel good about themselves and that people derive part of their identity from membership in social groups that they associate with (Tajfel & Turner, 1985). To achieve and maintain a positive self image, people view their group more positively than comparison outgroups on some valued dimensions (Hogg & Vaughan, 2002). This ingroup favoritism is an essential part of group identity and such beliefs arise even in minimal group settings. In a number of lab studies where subjects were arbitrarily assigned to groups (but told that they had something in common with other group members who they may never meet), participants gave more rewards to members of their group (Hogg & Vaughan, 2002). Group superiority beliefs permeate rumours, myths, and folktales of groups around the world.

Events that manifest a higher status of the out-groups along the dimensions of value to a group, violate group's cultural expectations and may cause some highly identified group members to believe that the group myths are broken and need to be fixed. For instance, Christian conquest of Muslim lands in the 19th and 20th centuries, lead Muslims to ask the question, "what went wrong." If we are the chosen people who have been promised dominance in the world then how come we are losing so many battles to the Christian West? (Lewis, 2003). Such changes provide opportunities for information entrepreneurs to step up and offer their solution to the social problems. Groups have various mechanisms for rewarding those who are thought to be working for the group's benefit especially at a personal cost to their own welfare such as soldiers. Upal (2005a) argued that those who pioneer change in group social beliefs stand to gain an increase their social status if they come to be credited with having successfully advocated for the betterment of their group.

Ratcheting Up Social Counterintuitiveness

Once the efforts of information entreprenurs are successful and a counterintuitive idea becomes fully entrenched in a group, it no longer seems counterintuitive to most members of that group and therefore loses its memorability advantages. This resolves another paradox that critics of cognitive science of religion have often pointed out, namely, that while the counterintuitive beliefs such as religious belief in gods, and ghosts as well in popular culture beliefs about Draculas, vampires, Vulcans, djinns, chupacabras, and leprechauns are counterintuitive in the traditional cognitive science of religion sense, they do not appear to be counterintuitive to the people whose informational worlds are full of such creatures. Theists from a variety of traditions, for instance, routinely point out that they see God in everything such as people's eyes, flower petals, grass blades, running streams, stars, and singing birds and that the concept of God appears no more counterintuitive to them than air, energy, and kinetic potential (Cook, 1883; Rasor, 2006). Cultural anthropologist routinely point out that while mythical cultural creatures such as djinns and ghosts seem counterintuitive to us, they do not seem counterintuitive to the people who believe in them (Bloch, 2005).

The answer I believe lies in acknowledging the criticism that minimally counterintuitive ideas do indeed lose their privileged status and do not have any memorability advantages once they become embedded as part of a culture. However, this does not mean that further cultural innovation stops. New ideas continue to be created and communicated to others and those ideas that have transmission advantages continue to spread. In order to have memorability advantages due to counterintuitiveness however, new ideas must violate people's expectations in the new context and not the old context which is no longer relevant. This means, for instance, that once as a minimally counterintuitive idea such as the idea of a being who can see everyone becomes widely culturally accepted, it loses its memorability advantages because it no longer violates people's expectations. In order for a concept to achieve memorability advantages and to spread in the new cultural context, an idea has to seem counterintuitive in the new context. One way to do that is to build on the counterintuitiveness. For instance, the concept of a being who can see *and hear* everyone would seem minimally counterintuitive in the new context. In light of the model we develop here, one should not be surprised to see maximally counterintuitive concepts to form

a significant part of religious beliefs. Indeed, it would be surprising if they did not!

This ratcheting-up of counterintuitiveness not only explains how seemingly maximally counterintuitive concepts such as Judeo-Christian-Islamic God and ghosts come to be widely distributed but it also predicts a continuous transmission advantage for unorthodox ideas that violate cultural expectations over traditional ideas. This explains continuing evolution of cultural beliefs among groups ranging from post-modern artists to new religious movements. As arts historians know, each artistic trend is both defined in opposition to the old one and also as a continuation and improvement of the old trend. At the core of each trend is a minimally counterintuitive idea that is advocated by a group of innovators and becomes widespread because it is unexpected according to socially shared beliefs. However, once it becomes widely accepted group it loses its memorability advantages making room for a new layer of innovation. Similarly, new religious movement scholars recognize (Bainbridge, 1985) that splitting of a new religious movement (NRM) from an existing movement often involves introducing an innovation into the doctrinal beliefs of the existing movement. NRM scholars Bainbridge and Stark (1979) provide a number of examples of new religious movement leaders who created the fundamental doctrine of new religious movements by modifying the beliefs of existing NRMs. Indeed they argue that tracing the history of such deviations, labeled "cultural genetics", may be a useful way to study NRMs. Idea innovations leading to splits in NRMs are common. Bainbridge (1985) counts over half a dozen movements that split from Dianetics and the Church of Scientology in the short period of 20 years from 1952 to 1972.

In this way, the context-based model explains cultural innovation but what accounts for cultural continuity? In particular, what explains the perception that cultural concepts such as gods, ghosts, and angels have not changed for a long time? As anthropologists and historians know, despite the need for protagonists of conservative movements to argue otherwise, cultural ideas are continually undergoing change, so much so that social movements and societies often have to build a number of safeguards to prevent unwanted innovation. This includes writing down the doctrines in books and elevating such books to the level of the sacred, punishing any changes in the content of these books, and instituting measures to discourage translation and interpretation of these books.

Orthodox Christianity's attempts at rooting out heresies (Hogan, 2001) spanning over two thousand years illustrate problems that organized religions face as they attempt to maintain continuity over time. Both Judaism and Islam also had to repeatedly put down various attempts at introducing innovations in their religious doctrine and practices. In the case of Islam, the Quran was not allowed to be translated in any language other than Arabic until the 19th century. Innovation in religion (termed as "bidah") is explicitly forbidden (Islam, 2008). NRMs, despite having had to fight against the oppressive measures against innovation to have their own voice heard when facing the same need to protect the integrity of their own doctrine, disdain any attempts at introducing further innovations into their doctrines. For instance, the founder of Scientology, L. Ron Hubbard is reported to have referred to those who modify his techniques as "squirrels" who should be harassed, "in any possible way" (Welkos & Sappell, 1990). Weapons used to discourage any change in religious doctrine and practice include ridicule, expulsion, and harassment. Continuity in group ideologies is explained to the extent that such thought control techniques are successful.

Conclusions

Cognitive scientists, including cognition and culture researchers have long favored general models of cognition over specific ones not just because they explain a larger variety of phenomena but also because they are perceived as more parsimonious and subject to a larger battery of tests because of the availability of a larger number of data points to test them on. This paper makes a contribution to this literature by presenting a generalized version of the minimal counterintuitiveness hypothesis to argue that better recall for minimally counterintuitive ideas is part of a larger class of memory preference for distinctive items and that ideas that violate a small number of expectations generated by offline cognition/doctrinal thinking should also be remembered better than ideas that do not violate such expectations. The secondary contribution of this article is the development of the notion of social counterintuitiveness which allows us to explain the spread of culturally counterintuitive ideas.

References

Atran, S. (2004). *In Gods We Trust*. Oxford, MA: OUP.

Bainbridge, W. S. (1985). Cultural Genetics *Religious Movements* (pp. 157-198). New York: Paragon.

Bainbridge, W., & Stark, R. (1979). Cult Formation: Three Compatible Models. *Sociological Analysis, 40*(4), 283-291.

Barrett, J., & Nyhof, M. (2001). Spreading non-natural concepts. *Cognition and Culture, 1*, 69-100.

Barrett, J. L. (1997). *Anthrpomorphism, intentional agents, and conceptualizing God.* Cornell University, Ithaca, NY.

Barrett, J. L. (1999). Theological Correctness: Cognitive Constraint and the study of religion. *Method and Theory in the Study of Religion, 11*(4), 325-339.

Barrett, J. L. (2008). Coding and Quantifying Counterintuitiveness in religious concepts. *Method and Theory in the Study of Religion, 20*, 308-338.

Barrett, J. L., & Keil, F. C. (1996). Anthropomorphism and God concepts: Conceptualizing a non-natural entity. *Cognitive Psychology, 31*(3), 219-247.

Bloch, M. (2005). Are religious beliefs counter-intuitive. In M. Bloch (Ed.), *Essays on Cultural Transmission* (pp. 103-123). New York: Berg.

Boyer, P. (1994). *The Naturalness of Religious Ideas: A Cognitive Theory of Religion.* Berkeley, CA: University of California Press.

Boyer, P. (2001). *Religion Explained*. New York, NY: Basic.

Boyer, P., & Ramble, C. (2001). Cognitive templates for religious concepts. *Cognitive Science, 25*, 535-564.

Calkin, M. W. (1894). Association. *Psychological Review, 1*, 32-49.

Calkin, M. W. (1896). Association. *Psychological Review, 3*, 32-49.

Cook, J. (1883). *Advanced Thought in Europe, Asia, Australia*. London, UK: Richard D. Dickson.

Davidson, D., Larson, S. L., Luo, Z., & Burden, M. J. (2000). Interruption and bizzareness effects in the recall of script based text. *Memory, 8*(4), 217-234.

Gonce, L., Upal, M. A., Slone, D. J., & Tweney, R. (2006). Role of Context in the Recall of Counterintuitive Concepts. *Cognition and Culture, 6*(3-4), 521-547.

Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review, 101*, 371-395.

Harris, P. L., & Koenig, M. A. (2002). Trust in Testimony: How children learn about science and religion. *Child Development, 77*(3), 505-524.

Hogan, R. M. (2001). *Dissent from the Creed: Heresies Past and Present* Our Sunday Visitor.

Hogg, M. A., & Vaughan, G. M. (2002). *Social Psychology* (3rd ed.). London, UK: Prentice Hall.

Hunt, R. R., & Worthen, J. B. (2006). *Distinctiveness and Memory*. New York, NY: Oxford University Press.

Islam, M. (2008). *Decline of Muslim States and Societies.* Xlibris.

Keller, E. (2004). Towards Complete Clarity: Bible Study among Seventh-Day Adventists in Madagascar. *Ethnos, 69*(1), 89-112.

Kintsch, W. (1998). *Comprehension.* Cambridge, MA: Cambridge University Press.

Kintsch, W., & Green, E. (1978). The role of culture-specific schemata in the comprehension and recall of stories. *Discourse Processes, 1*(1), 1-13.

LeVine, S., & Cambell, D. T. (1972). *Ethnocentrism: Theories of conflict, ethnic attitudes, and group behavior*. New York: Wiley.

Lewis, B. (2003). *What went wrong? The clash between Islam and modernity in the Middle East*. New York: Harper Perennial.

Longley, C. (2003). *Chosen People: The Big Idea that Shaped England and America*. London, UK: Hodder & Stoughton.

McDaniel, M. A., Dornburg, C. C., & Guynn, M. J. (2005). Disentangling encoding versus retreival explanations of the bizareness effect: Implications for distinctiveness. *Memory and Cognition, 33*, 270-279.

McDaniel, M. A., Einstein, G. O., DeLosh, E. L., & May, D. P. (1995). The bizzarreness effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*, 422-435.

Packer, d. (2008). On Being Both With Us and Against Us. *Journal of Personality and Social Psychology, 12*(1), 50-72.

Rasor, E. (2006). *The Journey of Modern Mystic*. New York: iUniverse.

Schank, R. C. (1975). *Conceptual information processing* New York: American Elsevier.

Schank, R. C. (1979). Interestingness: Controlling inferences. *Artificial Intelligence, 12*, 273-297.

Schank, R. C. (1999). *Dynamic Memory Revisited.* New York: Cambridge University Press.

Schank, R. C., & Abelson, R. (1977). *Scripts, plans, goals, and understanding: An inquiry into human knowledge*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Slone, D. J. (2004). *Theological Incorrectnes.* New York, NY: Oxford University Press.

Sperber, D. (1996). *Explaining culture: A naturalistic approach.* Malden, MA: Blackwell Publishers.

Tajfel, H., & Turner, T. J. (1985). An integrative theory of intergroup conflict. In S. Worchel & W. Austin (Eds.), *Psychology of intergroup relations*. Chicago: Nelson-Hall.

Upal, M. A. (2005a). *Role of Context in Memorability of Intuitive and Counterintuitive Concepts.* The Proceedings of the 27th Annual Meeting of the Cognitive Science Society.

Upal, M. A. (2005b). Towards A Cognitive Science of New Religious Movements. *Cognition and Culture, 5*(2), 214-239.

Upal, M. A. (2007a). *The Optimal Cognitive Template of Minimally Counterintuitive Narratives.* Paper presented at the the 29th Annual Meeting of the Cognitive Science Society.

Upal, M. A. (2007b). *What is More Memorable Counterintuitive Concepts Interpreted Metaphorically or Literally? The* 29th Annual Meeting of the Cognitive Science Society.

Upal, M. A. (2009a). An Alternative Account of the Minimal Counterintuitiveness Effect. *Cognitive Systems Research.*

Upal, M. A. (2009b). Counterintuitiveness, Coherence And Memory for Folktales. DRDC Toronto Technical Report.

Upal, M. A., Gonce, L., Tweney, R., & Slone, D. J. (2007). Contextualizing counterintuitiveness. *Cognitive Science, 31*(3), 415-439.

Upal, M. A., & Harmon-Vukic, M. (2010). Why Maximally Counterintuitive Concepts Are Less Well Recalled, Cognition & Culture.

van Dijk, T. (1995). Discourse semantics and ideology. *Discourse and Society, 6*(2), 243-289.

von Restorff, H. (1933). Analyse von Vorgangen in Spurenfield. I. Uber die Wirkung von Bereichtsbildungen im Spurenfeld. *Psychologishe Forschung, 18*, 299-342.

Welkos, R. W., & Sappell, J. (1990, June 29, 1990). When the Doctrine Leaves the Church. *Los Angeles Times.*

Consciousness is Data Compression

Phil Maguire (pmaguire@cs.nuim.ie)
Department of Computer Science, NUI Maynooth
Co.Kildare, Ireland

Rebecca Maguire (rebecca.maguire@dbs.ie)
Department of Psychology, Dublin Business School
34/35 South William Street, Dublin 2, Ireland

Abstract

In this article we advance the conjecture that conscious awareness is equivalent to data compression. Algorithmic information theory supports the assertion that all forms of understanding are contingent on compression (Chaitin, 2007). Here, we argue that the experience people refer to as consciousness is the particular form of understanding that the brain provides. We therefore propose that the degree of consciousness of a system can be measured in terms of the amount of data compression it carries out.

Keywords: Information theory, data compression, Solomonoff induction, phenomenal experience, Turing test.

Introduction

According to Einstein, the most incomprehensible thing about the world is that it is comprehensible. But what does it mean to comprehend? A common feature of understanding in both science and mathematics is that it involves the reduction of a set of observations or truths to a more basic set of assumptions. Indeed, Chaitin (2007) has proposed that all forms of understanding can be viewed as instances of data compression. Have a look at the sequence below and see if you can 'understand' it:

4, 6, 8, 12, 14, 18, 20, 24...

What is involved in understanding this sequence? Intuitively, one searches for a pattern that links all of the numbers together. If the numbers were randomly selected, then, more than likely, no pattern could be identified. In this case the sequence could not be described any more concisely: it would be incompressible. However, the above sequence seems amenable to compression. For example, one can posit the following hypothesis: "start at 4 and keep adding 2, except if the digits of the previous number sum to 2, 5 or 8, in which case add 4". These instructions provide a complete description of the sequence. However, because the description seems somewhat unwieldy, it is not particularly convincing. A more concise description is possible: "go through all odd prime numbers and add 1". Because this hypothesis is more concise, it intuitively reflects a deeper understanding of the sequence.

Scientific understanding is furthered by exposing greater levels of redundancy in observational data. The goal of the scientist is to craft a model which can describe a dataset in more concise terms. These models are called *theories*. The more compression a theory achieves, the greater its value. For example, Kepler's heliocentric model of the heavens is considered superior to Ptolemy's geocentric model, because it manages to describe astronomical observations in terms of three simple mathematical laws rather than a convoluted set of epicycles.

The idea that compression underpins scientific endeavor is not new. Occam's razor is a fundamental scientific principle which is attributed to the 14th century English friar, William of Ockham. The principle states that the explanation of any phenomenon should make as few assumptions as possible, eliminating those that make no difference to the observable predictions: "entities should not be multiplied unnecessarily". This law of parsimony implies that if you have two competing theories which both describe a phenomenon, the simpler (i.e. more compressed) explanation is better.

Algorithmic Information Theory

As homo sapien sapiens (Latin for *knowing man*), the urge to understand is a defining characteristic of our species. But why is it that we should devote so much energy to understanding the world around us? In order to answer this question we must turn to algorithmic information theory. Algorithmic information theory is a field which brings together mathematics, logic and computer science. The foundations of this field were laid by Chaitin, Solomonoff and Kolmogorov in the 1960s (see Li & Vitányi, 1997). According to Chaitin, it is "the result of putting Shannon's information theory and Turing's computability theory into a cocktail shaker and shaking vigorously". The basic idea is that the complexity of an object can be represented by the size of the smallest program for computing it. This new way of thinking about information was first proposed by Solomonoff (1964) and subsequently independently identified by Kolmogorov and Chaitin.

Algorithmic theory provides a clear answer as to why organisms should seek to compress observational data. Specifically, Solomonoff's (1964) theory of inductive inference reveals that compression is a necessary component of prediction. The theory provides a universal measure of the probability of an object by taking into account all of the ways in which it might have been produced. This universal a

priori probability can then be incorporated into Bayes' rule for inductive inference in order to make optimal predictions based on a set of prior observations.

Solomonoff's theory of inductive inference reveals that the more a set of observations can be compressed, the more accurately subsequent events can be predicted. Consider again the sequence 4, 6, 8, 12, 14, 18, 20, 24... The long-winded description predicts that the next number in the sequence will be 26, while the more succinct description predicts that 30 will follow. According to Solomonoff's theory, the latter must be the better prediction, because it involves a fewer number of assumptions: the shorter the length of the description, the more likely it is to be correct.

Algorithmic information theory reveals that compression is the *only* systematic means for generating predictions based on prior observations. All successful predictive systems, including animals and humans, are approximations of algorithmic induction. All useful contributions to human knowledge work by coaxing people into modifying their inductive strategies in such a way that they better approximate algorithmic induction.

In order to thrive in an uncertain environment, organisms must be able to anticipate future events; the more efficiently they can compress their experiences, the more accurate these predictions will be. Consequently, organisms have evolved brains which are prodigious compressors of information: compressing sensory information provides them with an 'understanding' of their environment (see Chater & Vitányi, 2002; Schmidhuber, 2006; Wolff, 1993). Tononi (2008) has proposed that the feeling of being conscious must be linked in some way to the integration of information which occurs in the brain. In the following sections we specify precisely the relationship between information processing and subjective awareness: specifically, we argue that the experience people describe as consciousness is equivalent to the compression that the brain carries out. Henceforth, this idea is referred to as the 'compression conjecture'. It should be noted that the conjecture does not merely suppose an association between consciousness and compression; rather it asserts that no meaningful distinction can be drawn between the two concepts.

Consciousness

From an evolutionary perspective, the sole purpose of the brain is to produce behavior that optimizes the reproductive success of an organism and its genetic material. Features of the brain which are not linked to optimizing behavior should therefore not have been rigorously preserved by evolution. Why then should brains go to the trouble of producing consciousness?

Algorithmic information theory tells us that the key to enhancing prediction (and hence reproductive success) is to optimize data compression. If the principal evolutionary pressure determining the structure of the brain has been its capacity to compress data, and if brains are the only system we know of that support consciousness, then this suggests a rigorous link between consciousness and compression. Systems that are good at compressing data seem to produce consciousness. But why should this be the case?

The Brain as a Compressor

In order to answer this question, we must consider the nature of the compression that the brain carries out. In other words, what type of understanding does the brain provide?

The success of an organism is dependent on cooperation between all of its constituent components. In order to achieve the goal of reproduction, it must exhibit coordinated behavior. For example, it does not make sense for an organism's legs to maintain independent agenda. Because the interests of both legs are intimately bound, it is more productive for them to cooperate with each other in achieving a single set of objectives (e.g. putting one foot forward while the other stays on the ground). Accordingly, the brain sources sensory information from all over the body and compresses it *in parallel*, thereby optimizing predictive accuracy for the organism as a whole. Tactile information from every limb is compressed in parallel with visual information from the eyes and audio information from the ears, giving rise to a form of understanding that is *centralized* and representative of the organism's experiences as a singular unit. The resulting decisions of the organism also appear centralized: to the external observer it seems as if the organism's body is being 'controlled' by a single entity with a singular set of objectives.

Not only does the success of an organism depend on cooperation between its constituent components, it also depends on cooperation between its past and future states. Snapshots of an organism's behavior taken at different points in time again reveal evidence of a singular set of objectives. For example, if you know you will be hungry in several hours time, you might pack a lunchbox in your bag. In this case, you are cooperating with your future self. From an evolutionary perspective, organisms cooperate with their future selves because reproduction is a challenging task which requires coordinated behavior manifested over an extended period of time. As a result, the brain goes to the effort of distilling memories which are maintained with the expectation that they will facilitate data compression at a future point.

The utility of memory can again be explained in terms of enhancing algorithmic induction. Memory allows us to make greater sense of the world by enhancing our ability to carry out compression. Incoming sensory data are compressed *in parallel* with stored historical data, allowing redundancy to be identified more efficiently and, consequently, enhancing predictive accuracy. Thus, the form of understanding that the brain produces unites not only distributed sensory organs but also past and current states of an organism. The compression conjecture proposes that the experience of this unitary form of understanding is what we mean when we use the term 'consciousness'.

Self-Awareness

Intuitively, the above account does not seem fully satisfactory. For example, one might conceive of an artificial compressor which compresses large amounts of current and historical data in parallel, though without experiencing the same form of awareness that we humans are familiar with. Indeed, the compression carried out by the brain has one additional ingredient which sets it apart from simpler compression systems: it compresses its observations of its own behavior. The capacity for a system to model its own actions necessarily involves the identification of itself as an entity separate to its surroundings. As a result, self-compression entails self-awareness.

The human brain is a self-representational structure which seeks to understand its own behavior. For example, people model their own selves in order to more accurately predict how they are going to feel and react in different situations. They build up internal models about who they think they are and use these models to inform their decisions. In addition, the human brain compresses the observed behavior of other organisms. When we watch other individuals, we realize that there is a great deal of redundancy in their activity: rather than simply cataloguing and memorizing every action they perform, we can instead posit the more succinct hypothesis of a concise 'self' which motivates these actions. By representing this self we can then make accurate predictions as to how the people around us will behave. The idea that the actions of an organism are controlled by a singular self is merely a theoretical model which eliminates redundancy in the observed behavior of that organism. People apply this same process to themselves: what you consider to be the essence of you is simply a model which compresses your observations of your own past behavior.

Phenomenality

A significant obstacle to providing a fully satisfactory theory of consciousness lies in explaining the phenomenon of subjective experience: why is it that we experience qualia which seem to elude scientific description? According to the consciousness conjecture, the 'flavor' of a quale can be linked to the particular form of compression that the brain carries out in response to a stimulus.

If an organism perceives a stimulus, yet can discern no pattern in the sensory data, then that stimulus will appear completely random and meaningless to the organism: the stimulus will not be experienced at all. On the other hand, if some redundancy can be identified, then the stimulus can be 'understood' (i.e. experienced) by relating it to previously gathered sensory information. For example, when people look at an apple, they perceive a round shape by identifying redundancy between the appearance of the apple and previously encountered round objects; they perceive a green color by identifying redundancy between the appearance of the apple and previously encountered green objects. When we 'see' an apple we are not just processing an instantaneous visual stimulus but, rather, compressing a set of data which has been gathered over a wide cross section of space and time. The structure of the brain allows a sensory stimulus to be translated into the subjective experience of understanding through the process of compression.

In sum, people don't passively observe the world around them; they gaze through the lens of understanding provided by their brains. When people talk about their subjective experience they are referring to the particular form of compression that their brain provides. The reason that these qualitative descriptions differ from objective scientific descriptions is because the subjective experience of a stimulus is dependent on how it is processed. The particular 'flavors' of qualia that we humans are familiar with are artifacts of our cognition, which are determined by the patterns our brains have evolved to detect and encode.

Describing Qualia

Intuitively, qualia appear to resist objective description. However, this intuition must be flawed, for if qualia could not be recorded in some informational form in the brain then we would not be able to remember them. In this case, all current subjective experiences would seem random and meaningless because there would be no previous subjective experiences with which to reconcile them.

According to the compression conjecture, which supposes that subjective experience and data compression are equivalent, it should be possible to provide a full description of a quale by detailing the compression that a system achieves in response to a stimulus. Thus, for example, the experience of red could be captured by describing the changing structure of the brain in response to the sight of a red object. This experience could then be comprehensively represented in terms of bits of bytes and could feasibly be contained in a book. Yet, intuitively, a book containing symbols could never capture the experience of the color red in the same way that we feel it; leafing through the pages of the book would not give rise to the subjective feeling of red. How can this apparent incongruity be rationalized?

The compression conjecture indicates that even if a book does carry a complete description of a subjective experience, merely reading the book is not sufficient for reproducing that experience. To appreciate it, the reader must be capable of compressing the data in the same manner in which it was originally compressed. For example, rather than simply leafing apathetically through pages of symbols, the reader must be capable of identifying the underlying patterns which link those symbols together. If a system is incapable of compressing the data, then it cannot 'understand' the experience which is contained within. Experience is dependent on the system which is doing the experiencing, as opposed to being intrinsic to a stimulus. Because reading a description of compression will not necessarily cause the same compression to occur in your own brain, reading about the experience of red will not make you experience red.

The Hard Problem

Initially, it might not be clear that the above satisfactorily addresses the hard problem of consciousness, which Chalmers (1995) identifies as the question of why consciousness *feels* like anything at all. In order to tackle this question, let us consider the case of an assembly of coordinated neurons (or, indeed, logic gates) called Amy. If we observe Amy's behavior over time, we will notice considerable redundancy in her actions. We can compress Amy's behavior through the succinct hypothesis of a core centralized self which is motivating her actions and which feels experiences. But this is just an abstract hypothesis based on a dataset: why should the formation of a hypothesis result in experience? The answer to this question lies in the realization that the hypothesis of Amy's subjective experience is a hypothesis which Amy herself holds, an understanding which is manifested through the compression she carries out. Understanding the hypothesis that one is feeling something and the actual experience of feeling are the *same thing*. Amy's feeling therefore exists relative to the assumption of her own existence, an assumption which the system itself is capable of making.

Conscious Systems

Algorithmic information theory makes clear predictions regarding what systems are conscious: objects which carry out compression are conscious, all other objects are not. Let us consider a chair. Intuitively, we would not expect a chair to be conscious. Can this intuition be justified by the compression conjecture?

Chairs do not carry out compression. They do not source sensory information from multiple locations and process it in parallel. They do not store memories to enhance future compression. And they do not develop a theory of self by compressing their own actions. Therefore they are not conscious.

Imagine holding a flame to the leg of a chair. The flame leaves a black mark, therefore the chair has certainly been affected by the flame. But intuitively, it does not seem reasonable to claim that the chair has *experienced* the flame. This difference between effect and experience is directly related to compression: specifically, the chair fails to experience the flame because the information it provides is not compressed in any way. If a chair's leg is burned it has no effect on any of the other legs. No information is communicated, and consequently there is no inter-leg data compression to bind the experiences of the chair together. Furthermore, the chair stores no memory (other than a black mark). The burning event has no effect on how subsequent events are processed, meaning that the experiences of the chair are not bound together across time. Finally, because the chair does not compress its own response to the flame, it has no awareness of any subjective experience.

In contrast, if a flame is held to the leg of a human, it has an immediate effect on how information from all other parts of the body is processed. The brain also stores a memory of being burned, thus altering the individual's future behavior in a manner which reflects the interests of the system as a whole. People 'feel' the effect of being burned because the compression carried out by their brain reflects an understanding of what it feels like to be burned. In contrast, no matter how many times you burn a chair, it will never react any differently.

Artificial Consciousness

The consciousness conjecture suggests that any system that carries out compression can be considered conscious to some extent. However, it should be noted that no known system is capable of matching or even approaching the depth of compression carried out by the organic brain.

Although computer algorithms such as Lempel-Ziv and BZip2 are used to compress files and text, these programs simply skim through data looking for trivial redundancy. Such compressors cannot realistically be described as 'understanding' text because the only patterns they can identify are based on simple statistical repetitions of symbols. In contrast, when people read a book they can 'explain' the text in terms of an underlying narrative derived from their own experiences of the world, a feat which involves a much deeper level of compression.

Nevertheless, there is no theoretical obstacle that would prevent consciousness from being implemented in an artificial medium. Any system that is arranged and updated in a way which allows for the compression of information will support consciousness, be it implemented in windmills, beer cans or toilet rolls. Although toilet rolls take up a lot more space and interact a lot more slowly, they can be arranged in such a manner so as to perfectly replicate the compression carried out by neurons in the brain.

Of course, the idea that a conscious being could be implemented in toilet rolls is very unsatisfactory. Such an implementation exacerbates the hard problem of reconciling a clearly reducible system with the feeling of intuitively irreducible experiences. One might ask: where does the consciousness reside? In this case the consciousness is not a property of any particular toilet roll. Rather, it is a property of the toilet roll system as a whole. Just like the behavior of a human, the output of the toilet roll system exhibits deep redundancy which can be effectively compressed through the hypothesis of a single centralized 'self'. In particular, the toilet roll system is itself aware of this hypothesis, and uses the theory of selfhood to guide its processing. The consciousness of the system therefore resides in its capacity to understand (i.e. compress) what it senses, thereby identifying itself as an entity separate to its environment.

The Location of Consciousness

Thus far, we have used the term 'compression' without describing precisely how compression can be identified in the brain. Where is it to be found? Intuitively, people assume that conscious experience must be drawn together at a single point, an idea which Dennett (1991) derisively refers to as the 'Cartesian theatre'. However, brain imaging studies indicate that cognitive processing is widely

distributed and does not appear to be bound at any particular point in space or time (Zeki, 2003).

Although intuition might suggest the need for a Cartesian theatre, it is important to note that the evolutionary demands which have shaped the brain's structure have not required information processing to be integrated in this way. The only moment that the brain is required to bring information together is when some action must be elicited; furthermore, only data relevant to that action needs to be integrated. Outside of this constraint, processing can remain distributed in space and time, with no impact on the success of the organism.

Accordingly, external time and 'conscious time' need not be synchronized to any greater extent other than to facilitate the undertaking of action when required. However, conscious observers have no possible means for observing any distribution in their consciousness relative to the environment: whenever they act on their surroundings the appropriate information processing is pulled together 'just in time'. Since it always appears to the observer as if they are embodied at a particular point in space and time, this leads them to mistakenly assume that their consciousness must be brought together at a single point in the brain, giving rise to the Cartesian theatre fallacy.

How Does the Brain Create Consciousness?

One of the goals of consciousness research is to identify how it is created in the brain: which neural structures support consciousness and which are merely superfluous biological apparatus? Using elementary computability theory we will prove that, if the compression conjecture holds, then the goal of identifying a complete theory of consciousness is unattainable.

Let us imagine that somebody someday submits a theory which offers a full description of how the brain produces consciousness. The theory is complete, meaning that it is capable of identifying precisely which structures in the brain give rise to consciousness, separating the conscious part from the non-conscious meat. Now, of course, the reviewers wish to check that the theory is correct. Accordingly, they apply the theory to their own brain activity to see whether the predictions match their experience. However, this raises the question: are the reviewers able to define their own consciousness, as required to validate the theory? Is it possible for a system to define its own self? In fact, computability theory rules this out, meaning that a complete theory of consciousness is not possible.

According to the compression conjecture, the recognition of one's own consciousness involves the identification of a structure which carries out the same form of compression. We can therefore present the problem formally in terms of a Turing machine which is capable of recognizing a program with the same input-output relationship. Consider a Turing machine T which takes input x and outputs 1 if $L(T) = L(x)$ (i.e. the languages recognized by T and x) and 0 if $L(T) \neq L(x)$. Is such a machine possible?

The machine T is not consistent. We can imagine another machine A which takes input x. The machine A first computes $T(A)$. If $T(A) = 1$ it then outputs $1 - T(x)$, which is the opposite of T, while if $T(A) = 0$ it then outputs $T(x)$, which is the same as T. In other words, the machine A checks to see whether T recognizes it as being equivalent or not. If T recognizes A as being equivalent then A proceeds to do the exact opposite, making it not equivalent to T. However, if T does not recognize A as being equivalent then A produces the same output at T, making it equivalent to T. There is no way around this obstacle (see Rice's theorem; Rice, 1953). Since no system can recognize an equivalent system from within itself, developing a complete theory of consciousness is not possible: the more precisely a theory attempts to define the conscious structure of the brain, the less feasible it will be to validate it.

The unrecognizability of the self has important implications for how we think about ourselves. For instance, we can never know who we really are; we can never fully explain our actions; we can never be certain as to what we are going to do next. In effect, the self is a helpless observer carried along by the compression going on in the brain. Of course, one feels like one is directing one's own actions because, as far as one is aware, one is. According to the compression conjecture, the model of the self is simply an explanatory mechanism that the brain uses to explain and predict its own behavior. As a result, the actions of the brain cannot help but be consistent with those of the self (see Gazzaniga, 1992). However, it is the activity of the brain which defines the nature of the self, rather than the other way around. Are you controlling your own actions? Certainly, but at the same time you can never know who *you* is.

Measuring Consciousness

If, as the compression conjecture supposes, consciousness is equivalent to data compression, then it should be possible to measure consciousness by quantifying the amount of compression that a system is capable of. The formal measure of compression is logical depth (see Bennett, 1988). Bennett's idea is that objects can be trivial, random or deep. Trivial objects, being completely predictable, contain no useful information; random ones, being completely unpredictable, do not contain any useful information either. In contrast, objects that are neither random nor trivial are called deep objects, because they support deep compression.

Deep objects are useful because they provide a store of mathematical work, allowing associated data to be compressed far more efficiently than can be achieved using shallower tools. Indeed, Bennett's (1988) theory implies that the concepts of 'depth' and 'intelligence' are equivalent, since the facilitation of compression that depth provides cannot be replicated by alternative means. Of all known objects, the human brain is the deepest, representing the stored mathematical work of decades of active cognitive processing on top of billions of years of evolution. The brain relies on its depth to mitigate the physical limitations on information processing imposed by its biological structure, such as limited storage capacity, processing speed and

susceptibility to degradation. The complexity of its structure allows people to effortlessly identify patterns which continue to elude the most advanced artificial intelligence programs.

The Turing Test

Turing (1950) suggested that if a computer, through a textual interface, can successfully convince a human judge that it is human, then it should be considered equal in intelligence to a human. However, the Turing test is not a reliable indicator of depth. Fooling a human judge is unlikely to require a deep program: a far simpler solution is to exploit the weaknesses of human psychology.

We propose an alternative test, involving compression, on which it is not possible to cheat. Because of its complexity, natural language provides the ideal medium for testing compressor depth. People use complex linguistic patterns to communicate with each other and assume that other speakers are capable of compressing the words they produce. If a computer system is as intelligent as a human, then it should be capable of compressing language to the same extent as a human.

According to algorithmic information theory, compression can be quantified in terms of predictive accuracy. For example, Shannon (1951) examined the human-perceived entropy of English by asking people to predict each letter in a document, one by one. The entropy rate turned out to be less than 1 bit per letter. People are able to predict language because of the fact that they 'understand' the text. In contrast, artificial compressors like BZip2 and Lempel-Ziv achieve much poorer levels of compression because they rely on predictable sequences of characters, without any regard for the deeper connections between words, sentences and narrative. If a computer was genuinely as intelligent as a human, it would be capable of matching the entropy rate of 1 bit per letter that Shannon observed.

We propose that the compression test is far more reliable and practical than the Turing test. For a start, there is no way to cheat: by definition, deep processing cannot be reproduced by any means other than underlying depth (the Slow Growth Law; see Bennett, 1988). It is also extremely quick and reliable: the probability of guessing the correct symbols decreases exponentially with the length of the test. While the Turing test is ambiguous and is affected by the gullibility of the tester, the compression test is simple, rigorous, reproducible and provides an exact measure of intelligence by means of the relative entropy score.

Conclusion

Intuitions regarding consciousness seem to create many problems which have not been satisfactorily resolved (see Dennett, 1991). In contrast, the framework we have described here can explain many of the questions regarding consciousness in an unambiguous and consistent manner.

The compression conjecture explains why a brain that evolved to optimize an organism's behavior should be associated with consciousness. It explains why consciousness is not amenable to scientific description. It explains what we mean by 'the self' and why brains provide self-awareness. It explains the apparent paradox of experiencing a singular perspective in a brain which carries out distributed processing. It predicts what systems are conscious and what systems are not; it reveals that a complete theory of consciousness is not possible. It tells us how to identify consciousness and it even provides a standard by which to measure consciousness.

The compression conjecture does not require special neuro-biological causal properties. It does not require mysterious quantum fluctuations in micro-tubules. It does not require an additional imperceptible dimension to the universe. It does not require the actions of a divine being. In fact, it requires nothing except data compression.

References

Bennett, C.H. (1988). Logical depth and physical complexity. In Herken, R., editor, *The Universal Turing Machine: A Half-Century Survey*, 227-258. Oxford: University Press.

Chaitin, G.J. (2007). The halting probability Ω: irreducible complexity in pure mathematics. *Milan Journal of Mathematics*, *75*, 291-304.

Chalmers, D.J. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies*, *2(3)*, 200-219.

Chater, N. and Vitányi, P. (2003). Simplicity: A unifying principle in cognitive science? *Trends in Cognitive Sciences*, *7(1)*, 19-22.

Dennett, D. (1991). *Consciousness Explained*. Boston: Little, Brown and Company.

Gazzaniga, M.S. (1992). *Nature's Mind*. London: Basic Books.

Li, M. & Vitányi, P. (1997). *An Introduction to Kolmogorov Complexity*. New York: Springer.

Rice, H. G. (1953). Classes of recursively enumerable sets and their decision problems. *Transactions of the American Mathematical Society*, *74*, 358-366.

Schmidhuber, J. (2006). Developmental robotics, optimal artificial curiosity, creativity, music, and the fine arts. *Connection Science*, *18(2)*, 173-187.

Shannon, C.E. (1951). Prediction and entropy of printed English. *The Bell System Technical Journal*, *30*, 50-64.

Solomonoff, R.J. (1964). A formal theory of inductive inference. *Information and Control*, *7*, 1-22.

Tononi, G. (2008). Consciousness as integrated information: a provisional manifesto. *Biological Bulletin*, *215*, 216-242.

Turing, A.M. (1950). Computing machinery and intelligence. *Mind*, *59*, 433-460.

Wolff J.G. (1993). Computing, cognition and information compression. *AI Communications*, *6(2)*, 107-127.

Zeki, S. (2003). The disunity of consciousness. *Trends in Cognitive Sciences*, *7*, 214-218.

Feature repetition effects on object familiarity: Evidence from an old/new recognition task

Selda Eren (e115275@metu.edu.tr)
Department of Cognitive Science, Middle East Technical University
Ankara, 06531 Turkey

Annette Hohenberger (hohenberger@ii.metu.edu.tr)
Department of Cognitive Science, Middle East Technical University
Ankara, 06531 Turkey

Abstract

We performed an old/new study/test recognition task to investigate feature repetition effects on object familiarity. The results showed that repeated features increased "old" responses during the test phase for new objects. This increase was linear with the number of repeated features on the object. Old objects, which had been among the study phase stimuli, were not affected by the number of repeated features on the object. We also analyzed the effect of feature type (colour, shape, border and pattern) on familiarity responses. We found an effect of feature type only for the old objects. Saliency of the features also affected familiarity: the more salient the repeated feature was, the more familiar the object was found. We propose that the feature repetition effect for the new objects might be due to (1) activation of more than one representation constructed during the study phase (2) a separate representation for the repeated features, which has the potential to interfere with several perceptual processes.

Keywords: feature repetition effect; object recognition

Introduction

Formation and activation of perceptual representations has been the subject of various disciplines including, but not limited to, philosophy, psychology, psychophysics, neuroscience, and computer science. In philosophy, the existence of "mental representations" is a fundamental debate in the philosophy of mind. In psychology, the studies of categorization and memory directly relate to this problem. Artificial intelligence and robotics research concentrate on implementing visual systems that construct a representation of their virtual or real environments. With the emergence of cognitive science, the bodies of knowledge that developed in these separate fields are coming together, for a better understanding of how perceptual representations are constructed and accessed. This study aims to contribute to the research on the formation and activation of visual object representations by revealing some important factors involved in memory processes. Our approach takes its roots from findings in perception and memory literature and computational approaches in artificial intelligence.

From a computational perspective, it is possible to represent everything in the environment as a combination of some features, like color, shape, pattern, etc. We know that the human brain has specialized areas for each of these feature domains (Hanna & Remington, 1996). Whenever a visual scene is encountered, activation is observed in these areas. Is this a mere bottom-up activation, or does the perceptual system attend to specific areas in the scene? We know that the visual system is not a passive receiver of visual data, but it actively obtains information from the visual flux (Jingling & Yeh, 2007). Attention makes a difference but we do not know whether the representation is stronger or the conscious access is easier in this case.

Whether the features in the scene are selected or all stored, it is clear that a combination of these features constitutes visual representations (Slotnick, 2004). Also audial and tactile features can be integrated with visual features, in which case the resulting representation can be called an "event file" (Hommel, 1998). Hommel states that all the features perceived in the same temporal window are automatically stored in these event files. These files can include features of every type, blurring the distinction between different domains of features, including visual and spatial pathways, which are assumed to exist separately in the brain. He points to the importance of building arbitrary connections between the features from different domains for learning.

In this study we investigated feature repetition effects on object familiarity. Hommel and Colzato (2009) report a decrease in performance in a stimulus-response task when one object feature is repeated while other features varied, as compared to complete repetitions and alterations. We predicted that repetition of particular features while other features vary would also affect familiarity of objects. We aimed to test this prediction with a continuous old/new study/test recognition design. In the study phase, participants saw a series of items one by one. In the test phase, they evaluated familiarity of the test items. To create the feature repetition effect, particular features were displayed more frequently than the other features in the study phase. We will call these features "frequently repeated features" (FRFs). In the test phase, items either had none, one, or two of the FRFs. We expected that the more FRFs the item had, the more participants would classify the item as familiar. We obtained scores for hits, misses, correct rejections and false alarms. False alarm scores are especially important

for our purposes. If items that were not displayed in the study phase are yet found familiar when they have FRFs, this would mean that (1) activation of previous bindings do not require an exact match with the given stimulus, or (2) there are other factors than binding of features that influence a familiarity judgement. If false alarms increase linearly with the # of FRFs, this might indicate an accumulated effect of repetition frequency on this judgement.

The design of the experiment is similar to the experiments in the categorization literature. In these experiments, a set of training objects are presented to the participants. In the test phase, they are expected to identify which category each test object belongs to. The features of the training objects are manipulated so that the effects of various variables such as similarity can be analyzed. However, our experiment significantly differs in the following terms: We do not assume a categorization process. Participants do not necessarily construct a categorical representation of the training stimuli and making familiarity judgements do not necessarily require accessing categorical representations.

The task in our experiment differs from the classical old/new recognition tasks, too. The usual old/new recognition task aims investigating the memory performance with respect to the dynamics of serial presentation of the stimuli. In our experiment, we systematically controlled the statistical properties of the object features and tested the effect of individual and combined feature repetitions instead of whole objects. In short, it can be said that our experiment integrates elaborate manipulations of object features as in categorization studies and experimental structure of an old/new recognition task. This provides a way of investigating the mechanisms of formation of perceptual representations through an analysis of the relationship between the statistical properties of the perceived stimulus and familiarity responses.

Another issue is feature intensity. Object representations in visual LTM have different intensities. The graded nature of these intensities shows its dominance in object recognition tasks, where object-based effects are tested (Ariga, Yokosawa, & Ogawa, 2007). In one task, participants were asked to recognize a target object in different conditions. In the first condition, the object was presented with a cue and in the second condition with no cue. Participants were faster at responding to objects presented with a cue only when the displayed object has a LTM representation of high intensity.

Finally, we investigated whether the type of feature is important for the feature repetition effect. Table 1 shows the feature types and values that appear in the stimuli set. By repeating different pairs of features, we analyzed familiarity responses for colour/border and shape/pattern pairs. In the next section the details of our design will be explained.

Method

Stimuli

Features There were four types of features: colour, shape, border and pattern. Each type had three values, as shown in Table 1. It was possible to create 81 objects using 4 features with 3 different values (3^4).

Table 1: Feature types and values used in the experiment

Colour	Red	Green	Blue
Shape	Square	Triangle	Circle
Border	Solid black	Dashed black	Coloured
Pattern	Dots	Diagonal lines	Shingle

Objects There were 15 objects. Objects were chosen among the pool of 81 possible objects, according to the following criteria: Solid black border and green color (pair 1) repeated together on 5 objects (see Figure 1a for an example of such object). Diagonal line pattern and square shape (pair 2) repeated together on 5 objects (see Figure 1b). Other feature pairs existed on 2 objects at most. FRFs were solid black border, green color, diagonal line pattern, and square shape, each repeating 7 times. Other features repeated only 4 times, e.g. 4 objects had blue color. Objects were created using the AutoShape tool of Microsoft Power Point. Objects had the same height (5 cm) and width (5 cm).

Slides One object was displayed on each slide. The center of gravity of the object was aligned to the center of the slide.

Training and test files There were 15 slides in the training file. Each slide was displayed for 2 seconds. Slide transitions were automatic. In the test file, there were 18 slides. The order of the slides was reversed in half of the participants. 8 slides were copied from the training file. The objects on these slides were the actual "old" objects. Remaining slides contained new objects. Each slide was displayed for 3 seconds.

Table 2: Number of objects of each category in the test phase.

	Old	New
Objects with two FRFs – pair 1	2	2
Objects with one FRF – pair 1	1	2
Objects with two FRFs – pair 2	2	2
Objects with one FRF – pair 2	1	2
Objects without any FRFs	2	2
Total	8	10

Participants

20 participants participated in the experiment. The age of the participants ranged between 22 and 35 years. All participants were university graduates. Participants had normal or corrected-to-normal vision. People who reported to be colorblind were not accepted to the experiment.

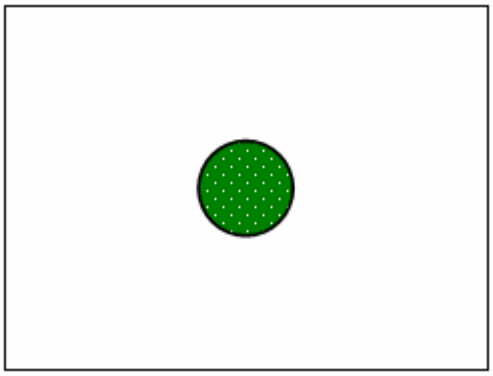
Experimental Design

There were two independent variables: familiarity and number of FRFs. Familiarity had two values: old or new. Number of FRFs had three values: 0, 1 and 2. The dependent variable was the familiarity score. It is the average of familiarity responses given to the objects in a category. Categories are displayed in Table 2. This was a 2x3 repeated measures design.

Setting

Computers in the Informatics Institute Computer Lab were used for the experiments. Stimuli were presented on a 19" widescreen LCD monitor by Microsoft Power Point software.

(a)



(b)

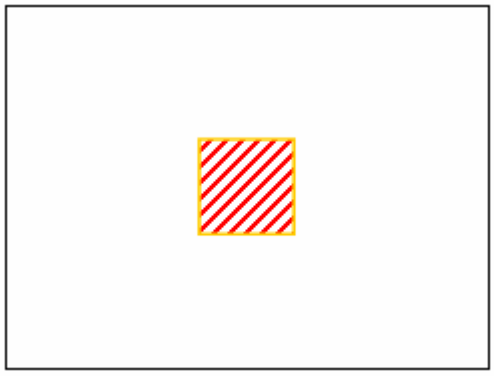


Figure 1 : Example stimuli from the study phase of the first experiment. These objects include features that have high repetition frequency (a) Green color and solid black border (b) Oblique pattern and square shape

Procedure

Before the experiment, participants signed an informed consent form. The instructions were as follows:

"The experiment consists of two parts. In the first part, you will see a series of slides. There will be objects on these slides. In the second part, I will show you another series of slides and ask you whether the object is familiar from the previous part."

The experimenter opened the Power Point file. "Press spacebar to continue" displayed on black background.

"You will press the spacebar when you are ready to start the first part. You will just watch the slides."

After all 15 slides were displayed, the Power Point turned back to the design view. At that point, the experimenter started the training slides from the beginning and instructed the participants as follows:

"Now I will repeat the same slides for better recall."

After the second round, the experimenter opened the test file, and gave the following instructions:

"I will show you a series of slides and ask if the object is familiar from the first part. Reply with Yes or No. Since there is a time limit, try to be as quick as possible." As the subject responded to each slide, the experimenter noted +/- marks on a response sheet.

Results

First, the familiarity scores for each category were calculated. The familiarity score is the average of the familiarity responses given by the participants to the test objects in a category. For example, if the participant responded with "familiar" to both objects the familiarity score was 1 ($response_1$= 1, $response_2$= 1, average($response_1$, $response_2$)= 1). If one of them was familiar, and the other one was unfamiliar, the familiarity score was 0.5 ($response_1$= 1, $response_2$= 0, average($response_1$, $response_2$)= 0.5). If both objects were unfamiliar the familiarity score was 0 ($response_1$= 0, $response_2$= 0, average($response_1$, $response_2$)= 0). Counts of familiarity responses are displayed in Table 3.

Color and border We analyzed the effect of repeating the features green color and solid black border on familiarity responses. The effects of the two independent variables, familiarity (old, new) and the number of FRFs (0, 1, or 2), were analyzed in a two-way repeated measures ANOVA. There was a main effect of familiarity ($F(1,19)=46.77$, $p<0.001$, $e=0.7$[1]), a main effect of number of FRF ($F(2,38)=13.57$, $p<0.001$, $e=0.4$) and an interaction between familiarity * number of FRFs ($F(2,38)=3.57$, $p<0.05$, $e=0.2$). The mean familiarity score was higher for the old objects, objects which actually existed in the set of the stimuli of the study phase, and the main effect of familiarity implies that this was significant. In other words, participants could successfully remember the

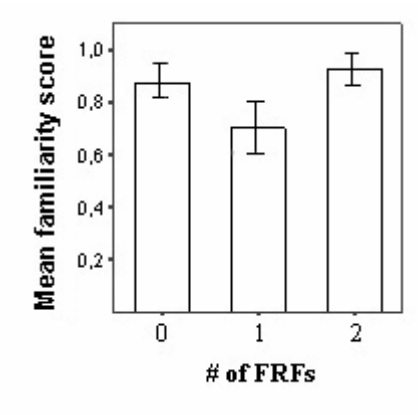
[1]"e" denotes "partial eta square".

Table 3 : Responses for the old/new recognition task. The numbers '0', '1' and '2' at the top of each column correspond to the number of FRFs on the object.

Response	Stimulus											
	Color and border repeated						Shape and pattern repeated					
	Old			New			Old			New		
	0	1	2	0	1	2	0	1	2	0	1	2
"Old"	35	28	37	9	16	27	35	36	32	9	17	28
"New"	5	12	3	31	24	13	5	4	8	31	23	12

objects that had been presented to them before. The main effect of number of FRFs shows that the familiarity response of the participants was affected by the number of FRFs on the object. As the number of FRFs increased, the mean familiarity score increased. The third significant effect is the interaction effect. In Figure 2, the different patterns of responses for familiar and unfamiliar objects can be seen. The number of FRFs did not affect mean familiarity scores for the familiar objects. However, for the unfamiliar objects, we see a totally different picture. If the object had no FRFs, then most of the participants reported that they had not seen the

(a)



(b)

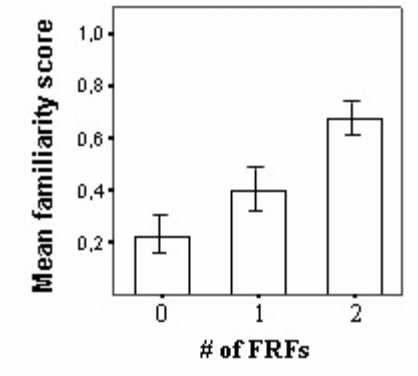


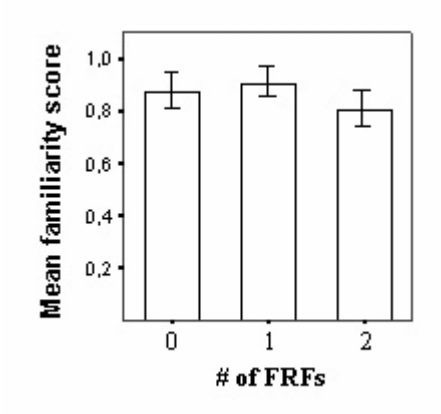
Figure 2: Mean familiarity scores for the objects with zero, one or both of the features color green and solid black border. Error bars represent standard error
(a) Old objects (b) New objects

object before. If the object shared only one of the FRFs, the mean familiarity score doubled. Finally, if the object shared both of the FRFs, most of the participants reported that they had seen the object, although they had not.

Shape and pattern Likewise, for the second pair, the square shape and the diagonal lines pattern, the effects of familiarity (old, new) and the number of FRFs (0, 1, 2) were analyzed with a two-way repeated measures ANOVA. There was a main effect of familiarity ($F(1,19)=28.89$, $p< .001$, $e=0.6$), a main effect of number of FRFs ($F(2,38)=5.67$, $p< .01$, $e=0.2$) and an interaction between familiarity * number of FRFs ($F(2,38)=10.89$, $p< .001$, $e=0.4$). The mean familiarity score was higher for the old objects, objects which existed in the set of stimuli and the main effect of familiarity implies that this was significant. The main effect of number of FRFs shows that the familiarity response of the participants was affected by the number of FRFs on the object. As the number of features increased, the mean familiarity score also increased. The third significant effect is the interaction effect. In Figure 3, the different patterns of responses for familiar and unfamiliar objects can be seen. The number of FRFs did not affect mean familiarity scores for the old objects. For the new objects, however, we see an effect of FRFs. If the object had no FRFs, then most of the participants reported that they had not seen the object before. If the object had only one of the FRFs, the average familiarity score doubled. Finally, if the object had both of the relevant features, most of the participants reported that they had seen the object.

Effect of feature types on familiarity responses The aim of this analysis is to test whether there was a difference between effects of repeating the color/border pair and repeating the shape/pattern pair on the familiarity judgment of objects. Mean familiarity scores for each pair are depicted in Figure 4. p1 represents the feature pair green color/black border and p2 represents the feature pair square shape and diagonal lines pattern. For hits, we see a slightly different pattern for p1 and p2. For false alarms, familiarity responses for p1 and p2 are almost identical. In this analysis we want to check whether the difference between p1 and p2 for the hits is significant. Two 2 (# of FRFs: 1, 2) x 2 (feature pair: 1, 2) repeated-measures ANOVA were

(a)



(b)

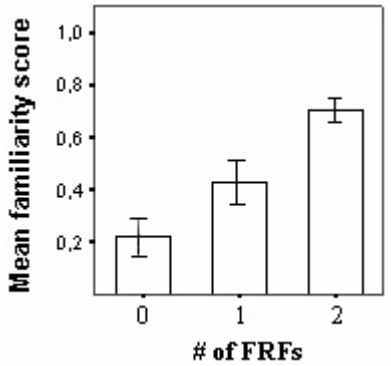


Figure 3: Average familiarity scores for the objects with zero, one or both of the features square shape and diagonal lines pattern. Error bars represent standard error
(a) Old objects. (b) New objects.

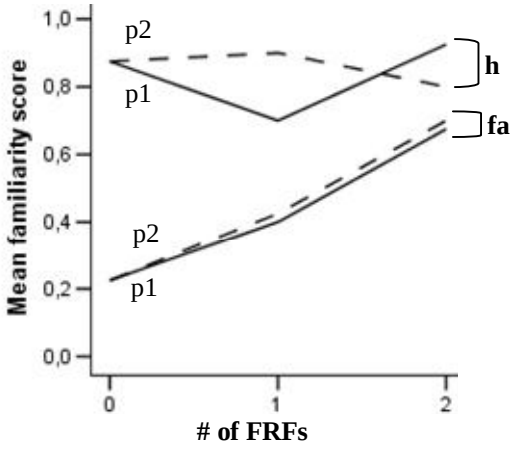


Figure 4: Familiarity scores for hits and false alarms for two different feature pairs. h denotes hits and fa denotes false alarms. p1 represents the feature pair green color/black border and p2 represents the feature pair square shape and diagonal lines pattern.

performed separately for hits and false alarms. For false alarms, there was no significant difference. For hits, there was an interaction effect between feature pair and # of FRFs, $F(1,19)=4.65$, $p< .05$, $e=0.26$). The interaction effect showed that as the "old" responses increased with the # of FRFs for the objects with green color and solid black border, a decrease was observed for the objects with square shape and diagonal lines pattern.

Discussion

We obtained three results from the old/new recognition task about the feature repetition effects on familiarity. (1) False alarm rates increase as the # of FRFs on new objects increase. (2) Hit rates were not affected by the # of FRFs on the object. (3) The type of feature influences the effect of FRFs only for hits.

The first result showed that if a new object in the test phase had no FRFs, the object was correctly identified as new. Familiarity responses increased linearly as the number of FRFs on the test object increased. In other words, participants classified new objects as old, if these objects had FRFs. The increase in "old" responses with two FRFs was twice the increase with only one FRF. Thus the relation was almost perfectly linear. This supports our hypothesis that familiarity judgements are not based solely on an exact match between the presented stimulus and existing representations. Partial activations of features enabled the classification of new items as old. However, this was true only if the partial activation is caused by frequently repeated features.

Hommel showed that repetition of a set of features while others vary affects performance in a response selection task. Our experiments revealed a similar pattern in a familiarity task. The repeated features caused an increase in false alarm rates. However, we believe that one should not consider the influence of frequently repeated features on familiarity as detrimental to performance. The perceptual system is sensitive to statistical properties of the stimuli (Turk-Browne et al., 2008). This enables extracting crucial information about the environment. Frequently repeated features might indicate regularities which are meaningful to the agent.

Second, feature repetition did not affect the hit rates. Hit rates were in general very high, indicating that participants responded as "old" to actually old objects most of the time. This may indicate that recognition success of the participants was high for the old objects. It means they could successfully represent the objects in the study phase.

Third, we found an interaction effect between feature type and the # of FRFs for hits. This was caused by the relatively small decrease in familiarity scores for objects with 1 FRF of pair 1. Further analysis revealed that this feature was the border feature. This might be due to the difficulty in perceiving or representing the border feature. It is not a basic feature as shape, color and pattern. So, we think that the interaction effect is related with the relatively poor representation of the border feature.

It is important that we obtained different patterns of results for old and new objects. For old objects, the repetition of particular features did not affect familiarity responses significantly. This is reasonable, since if a reliable representation of an object was constructed during the study phase, it should be identified as familiar during the test phase regardless of individual repetitions of the features. However, the opposite is not true, as shown by the increase in false alarm rates with the # of FRFs. Even though the new objects did not have previously constructed representations, they were identified as familiar if they had FRFs. This supports the claim that an exact match between the stored representations and a given stimulus is a sufficient but not a necessary condition for familiarity.

What do the FRFs activate? Do they cause partial activation of the existing representations? The existing theories of categorical representations do not provide answers to these questions. The context model (Medin and Schaffer, 1978) which claims that individual exemplars are stored in memory would not reflect sensitivity of the participants to statistical regularities of the stimuli. On the other hand, prototype theories would not account for the success of participants in recognizing individual objects from the training phase. The hybrid models (Nosofsky, Kruschke & McKinley, 1992) aim to combine the advantages of these two models but this pragmatic approach does not necessarily satisfy biological plausibility. We believe that a more comprehensive theory of perceptual representations, which is not restricted to representation of categories, should be developed, taking recent research on neural populations into account.

From the perspective of synchronization of neuron populations, FRFs can synchronize many representations at once. Why do the "old" responses increase linearly with the # of FRFs on the new object? More FRFs would mean activation/synchronization of more representations. However, since the joint frequency of the FRFs was also high, as well as their individual frequencies, this linear increase might be due to a better match between the stimulus and previously constructed representations. Alternatively, one may claim that FRFs do not activate existing representations, but they themselves constitute individual representations which are easier to activate and which can interfere with perceptual and motor processes in general.

Another thing to note is the effect of feature saliency. Color salience was not homogeneous among the objects because of the patterns we used in the experiment. The color green in dotted objects (where dots are black and other areas are green) were more salient than in objects with diagonal lines (where lines are green and other areas are white). The effect of saliency was reflected in the average familiarity responses for the objects, 0.8 for dotted pattern and 0.5 for diagonal lines pattern. If the FRF was more salient, the feature repetition effect was stronger. This variable will be manipulated in our future experiments.

Conclusion

In this experiment we tested the feature repetition effect on object familiarity with a continuous old/new recognition task. We found that repetition of particular features increased "old" responses during the test phase for new objects. This increase was linear with the number of repeated features on the object. Saliency of the features also affected familiarity; the more salient the repeated feature was, the more familiar the object was found. We proposed that feature repetition effect might be due to (1) activation of more than one representation constructed during the study phase (2) a separate representation for the repeated features, which has the potential to interfere with several perceptual processes. These findings will guide our efforts in the development of a computational model for the formation and activation of perceptual representations which is currently in progress.

Acknowledgments

This research was partially supported by The Scientific and Technological Research Council of Turkey (TUBITAK).

References

Ariga, A., Yokosawa, K., & and Ogawa, H. (2007). Object-based attentional selection and awareness of objects. *Visual Cognition, 15(6),* 685-709.

Hanna, A., & Remington, R. (1996). The representation of color and form in long-term memory. *Memory and Cognition, 24(3),* 322-330.

Hommel, B. (1998). Event files: Evidence for automatic integration of stimulus-response episodes. *Visual Cognition, 5,* 183-216.

Hommel, B., & Colzato, L. S. (2009). When an object is more than a binding of its features: Evidence for two mechanisms of visual feature integration. *Visual Cognition, 17(1/2),* 120-140.

Jingling L. & Yeh S. L. (2007). New objects do not capture attention without a top-down setting: Evidence from an inattentional blindness task. *Visual Cognition, 15(6), 661-684.*

Medin, D. L. & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 85,* 207-238.

Nosofsky, R. M., Kruschke, J. K. & McKinley, S. C. (1992). Combining Exemplar-Based Category Representations and Connectionist Learning Rules. *Journal of Experimental Psychology: Learning, Memory and Cognition, 18(2),* 211-233.

Slotnick, S. D. (2004). Visual memory and visual perception recruit common neural substrates. *Behavioral and Cognitive Neuroscience Reviews, 3,* 207-221.

Turk-Browne, N. B., Isola, P. J., Scholl, B. J., & Treat, T. A. (2008). Multidimensional Visual Statistical Learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34(2),* 399–407.

Beyond Transitional Probabilities: Human Learners Impose a Parsimony Bias in Statistical Word Segmentation

Michael C. Frank
mcfrank@mit.edu
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology

Harry Tily
hjt@stanford.edu
Department of Linguistics
Stanford University

Inbal Arnon
inbal.arnon@manchester.ac.uk
Department of Linguistics
University of Manchester

Sharon Goldwater
sgwater@inf.ac.uk
School of Informatics
University of Edinburgh

Abstract

Human infants and adults are able to segment coherent sequences from unsegmented strings of auditory stimuli after only a short exposure, an ability thought to be linked to early language acquisition. Although some research has hypothesized that learners succeed in these tasks by computing transitional probabilities between syllables, current experimental results do not differentiate between a range of models of different computations that learners could perform. We created a set of stimuli that was consistent with two different lexicons—one consisting of two-syllable words and one of three-syllable words—but where transition probabilities would not lead learners to segment sentences consistently according to either lexicon. Participants' responses formed a distribution over possible segmentations that included consistent segmentations into both two- and three-syllable words, suggesting that learners do not use pure transitional probabilities to segment but instead impose a bias towards parsimony on the lexicons they learn.

Keywords: Word segmentation; statistical learning; computational modeling.

Introduction

Human adults, infants, and even members of other species have the ability to identify statistically coherent sequences in unsegmented streams of stimuli after only a very short exposure (Saffran, Aslin, & Newport, 1996; Saffran, Newport, & Aslin, 1996; Hauser, Newport, & Aslin, 2001). This segmentation ability is extremely robust, operates across a wide range of modalities (Conway & Christiansen, 2005), and has been hypothesized to play an important role in early language acquisition (Kuhl, 2004). Nevertheless, relatively little is known about the computations underlying statistical segmentation.

In one influential study, Saffran, Newport, and Aslin (1996) exposed participants to a simple artificial language which consisted of six trisyllabic words concatenated together to form a continuous speech steam. After only a few minutes of exposure, participants were able to distinguish words in this language from strings that did not occur with the same frequency. They speculated that participants could succeed by computing syllable-to-syllable transitional probabilities (TPs) and segmenting the speech stream at local minima in TP.

There are many possible computations by which learners could extract coherent units from the statistical structure of the speech stream, however. Lexicon-based learners like PARSER (Perruchet & Vinter, 1998) and Bayesian lexical models (Brent, 1999; Goldwater, Griffiths, & Johnson, 2009) have also been proposed as possible models of segmentation. Though these models differ on several dimensions, all assume that learners attempt to learn a consistent lexicon—a set of word forms that is combined to form the training sequence—and they do this by preferring small lexicons composed of frequent, short words.

Two previous studies have examined whether this kind of model could provide a good fit to human learning performance. The first contrasted recognition of sub-parts of the words from a speech stream and found that PARSER, like human learners, failed to discriminate sub-parts of words after training (Giroux & Rey, 2009). The second study found that a parsimony-biased chunk-finding model better accounted for human performance across a range of experiments in the visual domain than a purely associative model (Orbán, Fiser, Aslin, & Lengyel, 2008). Thus, both of these studies suggest that human learners do not simply represent association probabilities in statistical learning.

Our current study asks what kinds of learning biases operate in statistical learning. Our study makes use of a novel language whose transition statistics support not just one but a range of possible coherent segmentations: training data could be interpreted as a sequence of sentences of six words from a lexicon of two-syllable words or a sequence of sentences of four words from a lexicon of three-syllable words (where all words appeared with approximately the same frequency). TPs for a single sentence in this language are shown in Figure 1. A learner using pure TPs to segment the language would not recover either lexicon but would instead either segment the language into sets of six-syllable words or else segment inconsistently into a mix of two- and three-syllable words. Thus, our language was designed to test whether human learners would learn more parsimonious lexicons than those implied by pure transition statistics.

Experiment 1 validates two methodological innovations: a web-based interface for data collection and a dependent measure which directly evaluates participants' word segmentation judgments. Experiment 2 uses these methods to test

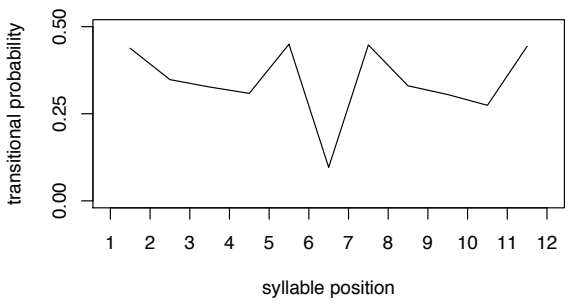


Figure 1: Average transitional probabilities between syllables in an ambiguous language from Experiment 2.

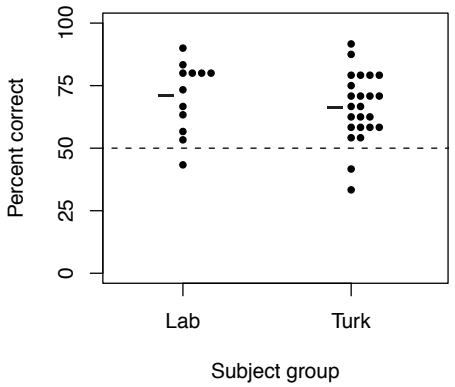


Figure 2: Average percent correct is plotted by subject for in-lab participants from Frank et al. (under review) and Mechanical Turk participants from the 2AFC condition of Experiment 1. Each point is an individual participant, bars show the mean, and the dashed line represents chance.

participants' segmentation judgments in the ambiguous language discussed above. We compare the distribution of participants' segmentations to the performance of two computational models—a standard TP model and a Bayesian model that looks for a parsimonious lexicon—and conclude that participants' judgments reflect the operation of a parsimony bias.

Experiment 1

The first condition of Experiment 1 compares web-based data on a segmentation task to previously-collected lab data (Frank, Goldwater, Griffiths, & Tenenbaum, under review) on a standard 2 alternative forced choice (2AFC) test trial. The second condition evaluates a new measure of segmentation: explicit segmentation decisions. We developed a graphical paradigm in which participants heard a sentence, saw it transcribed on the screen, and were asked to click between syllables to indicate where they thought the boundaries between words were.

Methods

Participants Forty eight separate HITs (opportunities for a participant to work) were posted on Amazon's Mechanical Turk web-based crowd-sourcing platform. We received 40 HITS from distinct individuals. Participants were paid $1 for participating.

Stimuli For each condition, we constructed 16 distinct languages to be heard by different participants (to avoid item effects caused by phonological similarity of words). These languages each had a lexicon of six words (2 x two syllables, 2 x three syllables, 2 x four syllables). Words were created by randomly concatenating the syllables *ba*, *bi*, *da*, *du*, *ti*, *tu*, *ka*, *ki*, *la*, *lu*, *gi*, *gu*, *pa*, *pi*, *va*, *vu*, *zi*, and *zu*. Stimuli were synthesized using MBROLA (Dutoit, Pagel, Pierret, Bataille, & Vrecken, 1996) at a constant pitch of 100Hz with 25ms consonants and 225ms vowels. Sentences were generated by randomly concatenating words into strings of four words with no repetitions. All words had frequencies of 300 in the resulting corpus of 75 sentences.

For the 2AFC condition, part-word test stimuli (Saffran, Newport, & Aslin, 1996) were created by concatenating the first syllable of each word with the remaining syllables of another word; this created distractors which appeared in the training corpus with lower frequency than the words. For the segmentation condition, we generated 10 extra sentences according to the same uniform frequency distribution and lexicon as the training corpus.

Procedures After selecting our HIT, our Adobe Flash interface tested that participants' sound was on and that they were able to understand our instructions by asking them to listen to a simple English word and enter it correctly. Participants were then instructed that they would listen to a set of sentences from a made-up language and then be tested on what they had learned. In order to hear each sentence during training, participants clicked a button marked "next."

In the test phase of the 2AFC condition, participants heard 24 pairs consisting of a word and a length-matched part-word and clicked a button for each to indicate which one sounded more like the language they just heard. In the segmentation condition, participants were asked to click on the breaks between words in a graphic display of a sentence. They performed one practice trial on an English sentence presented in this way ("In di an go ril las ne ver eat ba na nas") and prevented from continuing until they segmented it correctly. They then segmented 10 test sentences. Sentences were presented with each syllable separate. Each sentence was played once at the beginning of a trial, and below the sentence was a button that offered the option of hearing the sentence again.

Results and Discussion

In the 2AFC condition (N=24), we found that participants were above chance in their mean accuracy, taken as a group ($t(23) = 5.92$, $p < .0001$). Results are plotted together with data from an identical condition of Frank et al. (under review) (Experiment 2, 300 words exposure), collected from a group of participants in the lab (Figure 2). Mean performance was slightly lower for the Internet-based Turk par-

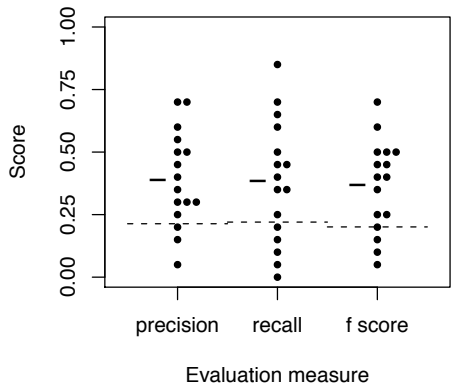


Figure 3: Token precision, recall, and F-score are plotted for individual participants in the segmentation response condition of Experiment 1. Points represent individual participants for each measure. Bars show means and dashed lines show permutation baselines.

ticipants (M=66% compared with M=71%) but not significantly so (Welch two-sample t-test for unequal sample sizes, $t(21.21) = -.92$, $p = .37$). Participants completing the learning task on their own computer via the Internet were able to perform at levels comparable to participants in an isolated room in a psychology laboratory.

In the segmentation condition (N=16), we could not analyze participants' percent correct judgments as in the 2AFC condition. Instead, we evaluated two aspects of performance. First, we asked about the correctness of the boundaries participants placed: whether these decisions corresponded to the correct segmentation (*boundary* performance). Second, we asked about whether each word in the sentence was segmented correctly at its boundaries (*token* performance).

We computed hits (correctly placed boundaries or correctly segmented tokens), misses (missed boundaries or tokens that were not segmented appropriately), and false-alarms (extra boundaries or incorrect tokens that were segmented). Precision captures the proportion of boundaries that were placed correctly and is computed as hits / (hits + false-alarms), while recall captures the total proportion of correct boundaries that were identified and is computed as hits / (hits + misses). We combined these into an F-score, a commonly used metric that is the harmonic mean of precision and recall (Goldwater et al., 2009).

Figure 3 shows token precision, recall, and F-score for participants in the segmentation condition. We calculated an empirical baseline for each measure via permutation: we repeatedly shuffled each participant's boundary decisions within each sentence at random and computed the same measures over it, then took the mean for each. We then used these empirical baselines to test whether participants were above chance in this condition and found that they were for both measures (boundary performance: one sample t-test for precision, $t(15) = 5.23$, $p = .0001$; recall, $t(15) = 6.79$, $p < .0001$; F-score, $t(15) = 8.75$, $p < .0001$, token performance: $t(15) = 3.63$, $p = .002$; recall, $t(15) = 2.71$, $p < .01$; F-score, $t(15) = 3.41$, $p < .004$), though boundary performance was better than token performance. Participants were able to understand the segmentation task and link the regularities they extracted from the exposure corpus to the response format.

Experiment 2

We made use of the two methodological innovations from Experiment 1—Internet data collection and explicit segmentation judgments—to ask about participants' responses to a language where TP did not reveal the possible lexicons of two- or three-syllable words. Instead, pure TPs predicted that participants would often segment the language into words of six-syllables and would rarely segment into words of two or three syllables. Our next experiment tests these predictions.

Methods

Participants Two-hundred and three separate experimental HITs were posted on Amazon Mechanical Turk. We received 119 HITs from distinct individuals who made segmentation decisions on every trial. Participants were paid $0.50 for participating. An addition 145 HITs in the test-only control condition were posted at $0.25 each; we received 102 HITs from distinct individuals who made segmentation decisions.

Stimuli Languages were generated using two parallel vocabularies, one of eight two-syllable words and one of six three-syllable words. These vocabularies were designed to allow overlapping segmentations where the presence of a certain word from one vocabulary did not always indicate the presence of the same set of words from the other. For example, if the three-syllable vocabulary contained ABC, the two-syllable vocabulary would contain at least either AB and two words beginning C, or BC and two words ending A. Sentences of 12 syllables were generated by choosing syllables one at a time from the set that made the sentence to the current point compatible with both vocabularies. At each point, syllables were chosen from a distribution over this set, weighted inversely to the frequency with which they had been chosen to follow the previous syllable in all sentences so far. The resulting sentences displayed probabilistic word-to-word dependencies, much as one would expect in natural language due to the syntactic relationships between words, but in no languages were there pairs of words from either vocabulary which always appeared together. We generated 30 distinct languages and synthesized them as in Experiment 1. Each language contained 25 sentences for training and 10 test sentences, sampled from the same distribution. Sentence presentation order was random.

Procedures Procedures were identical to the segmentation condition of Experiment 1. Participants in the test-only control condition received no training sentences.

Results and Discussion

Participants produced a wide range of segmentations, from those which segmented every three syllables to those which

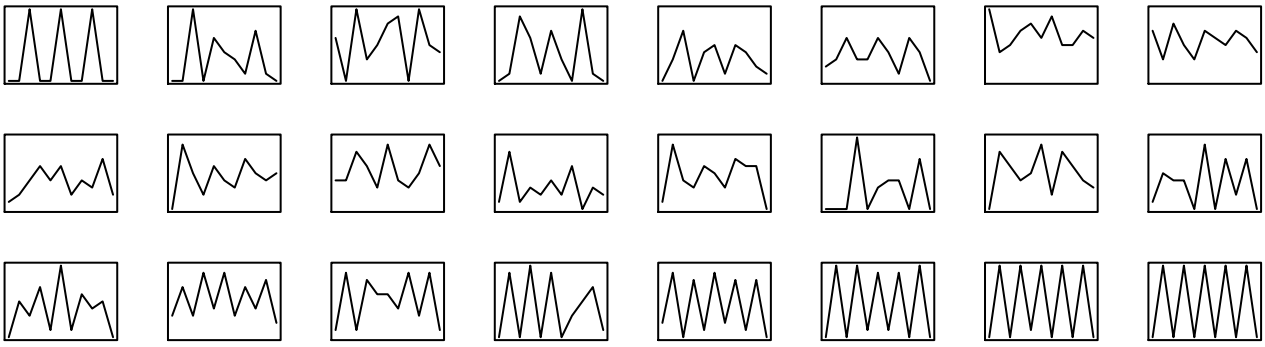


Figure 4: Twenty four participants in Experiment 2, uniformly sampled along the dimension of 2-segmentation F-score. Plots show average probability of placing a boundary at each location in a sentence. Top left shows three-segmenters (three peaks separating four three-syllable plateaus), while bottom right shows two-segmenters (five peaks separating six two-syllable plateaus).

segmented every two syllables. Sample responses are shown in Figure 4. While there was an overall trend towards 2-consistent segmentations, a wide variety of segmentations were observed. Contrary to the predictions of the TP account, there were almost no segmentations into words of six syllables and there were a considerable number of segmentations into words of two and three syllables.

We evaluated participants' performance on the same measures used in Experiment 1: precision, recall, and F-score for both boundaries and tokens. Rather than using a single correct segmentation, we calculated these measures for both the 2-syllable lexicon and the 3-syllable lexicon (Figure 5), showing the distribution of responses on the continuum between a perfect 2-segmentation and a perfect 3-segmentation.

One possible alternative explanation of our finding could be that learners have a bias towards segmenting consistently (e.g., because of the trochaic, bisyllabic structure of English) even without taking into account the structure of the languages they heard. However, results from the first trial of the test-only condition had a very different distribution than those who underwent training (Figure 5). Without training, performance was similar to a randomized baseline in which participants' judgments for each sentence were shuffled randomly. Although there was some learning during test for participants in the test-only condition, there was very little change in the distribution of responses during test for those participants who underwent training.

Our results are inconsistent with the hypothesis that participants segmented on the basis of TPs. Instead, the distribution of participants' responses shows a bias towards segmentations that were consistent with a more parsimonious lexicon than that produced by segmenting at low transition probabilities.

Models

To formalize the intuitions motivating Experiment 2, we evaluated a TP model and a lexicon-finding model on the experimental stimuli. We then evaluated the segmentations produced by these models on the same criteria that we used for the human participants.

Transitional probability model

For each language, we calculated TP for each pair of syllables that appeared in the training portion of the corpus. We computed TP as $P(s_2|s_1) = C(s_1,s_2)/\sum_{s' \in S} C(s_1,s')$ where $C(s_1,s_2)$ refers to the count of instances of the string $s_1 s_2$.

Earlier proposals for TP models called for segmenting at local minima in TP (Saffran, Newport, & Aslin, 1996). However, this method produces only a single possible segmentation for a given sentence and provides no plausible explanation for how participants could have given such different responses for such similar languages. Thus, we chose to convert the TPs for test sentences into decision boundaries via a simple threshold operation: we inserted a boundary in a test sentence every time TP was below a threshold value in that sentence. Rather than picking a single threshold value, we assumed that participants might have a range of threshold values and that this range might explain the variation between participants we observed. Therefore we created a separate segmentation for each language for each threshold value from zero to one at an interval of .1.

Lexical model

We also ran the unigram Bayesian Lexical model described in Goldwater et al. (2009). This model is a probabilistic model which uses Bayesian inference to search the space of segmentations of the training corpus, evaluating each segmentation on the parsimony of the lexicon that would have created it. The structure of the model makes a segmentation more probable when it results in fewer, shorter lexical items (though also when the segmentation itself contains fewer word tokens, which leads to a trade-off).

As in the TP model, it was important to investigate the range of segmentations that were available under this model. When we ran a standard Markov-chain monte carlo algorithm using the parameter set from previous simulations, we found

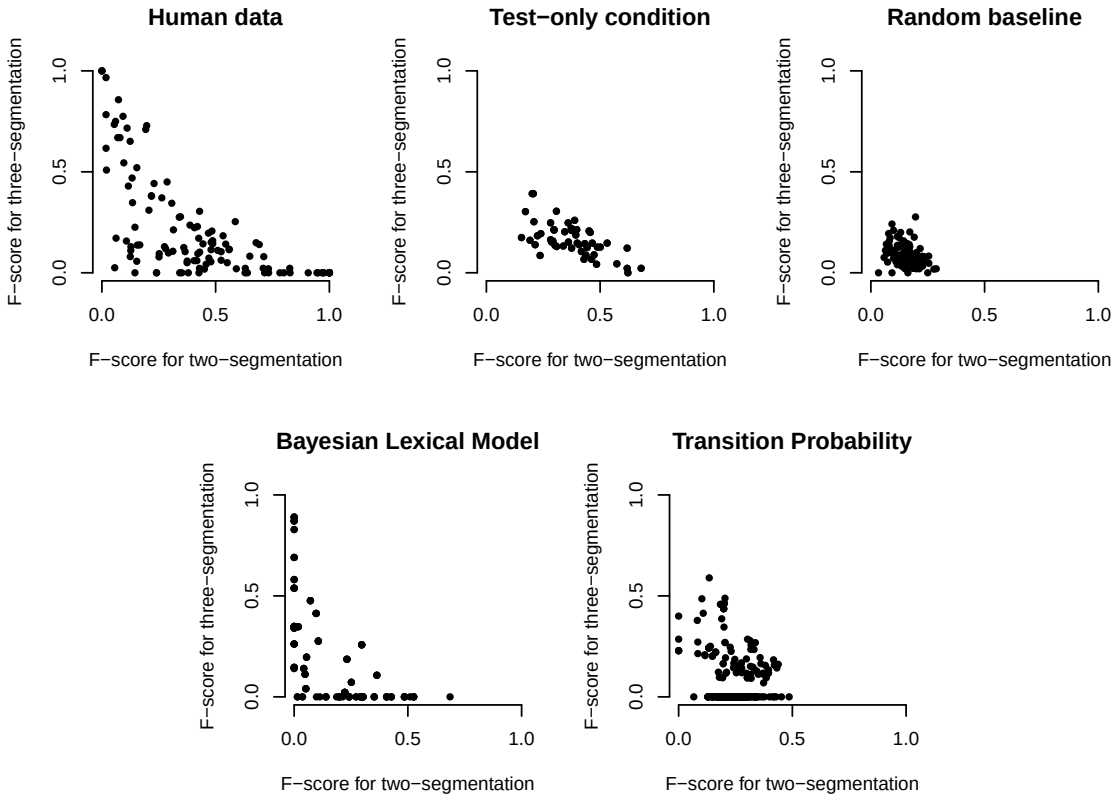


Figure 5: Participant and model token F-scores for Experiment 2. Three-syllable token F-scores are plotted by their two-syllable token F-scores. Each dot represents a single participant or a single model run.

Table 1: Kullback-Leibler divergence between the distribution of human experimental data and other data.

Model	Token F	Boundary F
Test-only condition	4.01	3.45
Random baseline	7.26	9.45
Lexical model	**2.07**	**3.16**
Transitional probability	4.62	3.72

Table 2: Log probability of consistent segmentations under the Lexical model.

Syllables per word	Log probability
6	-594.28
4	-932.92
3	-530.62
2	-697.07
1	-1127.20
unsegmented	-1907.20

that it converged to a segmentation that preferred a lexicon of three-syllable words. In order to investigate a broader range of segmentations, we manipulated the temperature of inference in the model by exponentiating posterior probabilities at a range of values. (This manipulation is a standard technique for allowing sampling algorithms to explore a hypothesis space more broadly, rather than converging to the single highest-probability answer.) With slightly higher temperatures, our sampler explored a broad range of possible segmentations. We report results for temperature = 2 although results for a temperature of 3 were comparable.

Results and Discussion

Results for both models are shown in Figure 5, bottom. The transitional probability model failed to capture the spread of human results: nearly all segmentations it found were comparable in F-score for 2- and 3-segmentation, and no segmentation was over an F-score of .5 on either measure. The Lexical model came closer to capturing the distribution of responses, though it was not as effective at finding 2-segmentations as the human participants, suggesting a possible role for a trochaic bias. Unlike the TP model, however, its probability landscape was truly multi-modal, finding relatively high probability segmentations with 2, 3, and 6 syllables per word (Table 2).

We measured the differences between the distributions of responses across human participants and models using Kullback-Leibler divergence—an information-theoretic mea-

sure of the difference between a true distribution and an approximation of that distribution—to quantify the number of bits between distributions (MacKay, 2003). In order to convert sets of observations into smooth distributions, we convolved them with a Gaussian kernel with a constant kernel width. This manipulation produced a smooth density which could be effectively compared using KL divergence.[1] Results are shown in Table 1. The Lexical model showed the lowest divergence from the human response distribution, while the TP model was closer to the empirical baseline in its divergence from the human distribution.

General Discussion

We presented two studies of statistical word segmentation. The first study introduced two methodological innovations, web-based data collection and explicit segmentation judgments. We used these new methods in the second study to test whether human learners faithfully learned the transitional probabilities of an ambiguous language or whether they gave a segmentation that was more consistent with one of the two possible lexicons that generated the training corpus. We found that the distribution of participants' responses was not consistent with the distribution of segmentations produced by segmenting according to a TP model. Thus, our results provide evidence that human learners do not simply encode transitional or associative statistics but instead impose some kind of bias on what they learn.

This bias could be either a bias for consistent word lengths or for a parsimonious lexicon. A model which searched for lexicons with small lexicons consisting of highly frequent, short words produced a distribution similar to that produced by the human learners. Nonetheless, the Lexical model preferred a lexicon with three-syllable words, unlike human learners who preferred to segment into two-syllable words; and the Lexical model assigned a high probability to a segmentation into two words of six syllables each, while participants rarely produced this segmentation. Frank et al. (under review) found that models with memory limitations provided a better fit to human performance, suggesting that one possible explanation for these differences is the increased difficulty for human learners of remembering longer words.

The language used in Experiment 2 has a number of limitations. First, unlike recent studies (Frank et al., under review; Giroux & Rey, 2009), the competing lexicons we used in this study were composed of words of homogenous length, leading to stimuli that could be perceived as isochronous. Second, the size of the lexicons was relatively small and the restrictions on sentences were tight, leading to a small number of possible sentences. Our ongoing work attempts to address both of these issues.

[1]Because both the TP model and the Lexical model produced a significant number of segmentations that failed to place any boundaries—for the TP model this was due to extreme threshold values, and for the Lexical model this was due to convergence issues in the online sampler we used—we excluded all model runs that failed to make any segmentation decisions.

Results in the statistical learning literature have rightly been interpreted as showing that human learners are sensitive to associative and transitional statistics in their environment. But these interpretations should not be confused with the conclusion that learners compute these particular—or any—transition statistics. Instead, future research on statistical learning should attempt to characterize both human learning biases and the computations that give rise to them.

Acknowledgments

Thanks to Richard Aslin, Noah Goodman, Elissa Newport, Josh Tenenbaum, and Ed Vul for valuable discussion. MCF was supported by a Jacob Javits Graduate Fellowship and NSF DDRIG #0746251.

References

Brent, M. (1999). An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning*, *34*, 71–105.

Conway, C. M., & Christiansen, M. H. (2005). Modality-constrained statistical learning of tactile, visual, and auditory sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 24–39.

Dutoit, T., Pagel, V., Pierret, N., Bataille, F., & Vrecken, O. (1996). The MBROLA project: Towards a set of high quality speech synthesizers free of use for non commercial purposes. In *Fourth International Conference on Spoken Language Processing.*

Frank, M., Goldwater, S., Griffiths, T., & Tenenbaum, J. (under review). Modeling human performance in statistical word segmentation.

Giroux, I., & Rey, A. (2009). Lexical and sub-lexical units in speech perception. *Cognitive Science*, *33*, 260–272.

Goldwater, S., Griffiths, T., & Johnson, M. (2009). A bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, *112*, 21–54.

Hauser, M., Newport, E., & Aslin, R. (2001). Segmentation of the speech stream in a non-human primate: Statistical learning in cotton-top tamarins. *Cognition*, *78*, 53–64.

Kuhl, P. (2004). Early language acquisition: cracking the speech code. *Nature Reviews Neuroscience*, *5*, 831–843.

MacKay, D. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge, UK: Cambridge University Press.

Orbán, G., Fiser, J., Aslin, R. N., & Lengyel, M. (2008). Bayesian learning of visual chunks by human observers. *Proceedings of the National Academy of Sciences*, *105*, 2745-2750.

Perruchet, P., & Vinter, A. (1998). Parser: A model for word segmentation. *Journal of Memory and Language*, *39*, 246–263.

Saffran, J., Aslin, R., & Newport, E. (1996). Statistical learning by 8-month-old infants. *Science*, *274*, 1926.

Saffran, J., Newport, E., & Aslin, R. (1996). Word segmentation: The role of distributional cues. *Journal of memory and language*, *35*, 606–621.

Effects of Goal Specificity on a Search in a Hypothesis Space and an Instance Space

Miki Matsumuro (muro@cog.human.nagoya-u.ac.jp)
Kazuhisa Miwa (miwa@is.nagoya-u.ac.jp)
Graduate School of Information Science, Nagoya University Furo-cho, Chikusa-ku, Nagoya, Japan

Abstract

We examined the effect of goal specificity on a search in two problem spaces: a hypothesis space and an instance space. Two hypotheses were considered: 1) a nonspecific goal facilitates a search in a hypothesis space more than a specific goal, and 2) as a hypothesis space is searched more, the performance in discovering the target rule improves. We also defined an initial hypothesis space consisting of initially considered hypotheses, and investigated the effect of this initial hypothesis space on the goal specificity effect. The results of three experiments indicated that when an initial hypothesis space was offered, the goal specificity effect was observed. A nonspecific goal actually facilitated a search in the hypothesis space. When, however, no initial hypothesis space was offered, the goal specificity effect was not confirmed. We also found that the facilitation of the hypothesis space search improved performance in discovering the target rule.

Keywords: discovery, rule induction, goal, hypothesis testing

Introduction

Dual Space Search Theory

Rule induction and scientific discovery have been studied based on the dual space search theory. Simon and Lea (1974) first suggested that a problem space consists of two spaces: a "rule space" for searching rules and an "instance space" for testing rules. Both rule and instance spaces are searched to find a correct rule.

Klahr and Dunbar (1988) extended the dual space search theory to the Scientific Discovery as Dual Search (SDDS) model for investigating scientific discovery. They considered a "hypothesis space" as a rule space and an "experiment space" as an instance space where the process of scientific discovery develops through the interaction between two types of searches in the two spaces. Reasoners state hypotheses by searching in a hypothesis space, receive feedback from an experiment space, and modify the current hypotheses or propose new hypotheses. Klahr and his colleague confirmed this model through a long series of experiments (Klahr & Dunbar, 1988; Klahr, 2000). They also identified "experimenters" who preferred to search in an experiment space and "theorists" who preferred to search in a hypothesis space. In this study we call the two spaces a "hypothesis space" and an "instance space."

The search in a hypothesis space is crucial for scientific discovery. Klahr and Dunbar (1988) demonstrated that a search in only a hypothesis space led to the discovery of a correct rule without the execution of any experiments.

Goal Specificity Effect in Dual Space Search

On the other hand, we often neglect to consider the theories or rules behind phenomena when we aim for a specific goal. We tend not to search in a hypothesis space at times like this, as we concentrate on a search in an instance space to achieve the goal.

Problem-solvers given a specific goal learn more poorly than problem-solvers given a nonspecific goal (Sweller & Levine, 1982). Burns and Vollmeyer (2002) investigated this effect of goal specificity based on the dual space search theory. Using a task in which participants were asked to learn the relations between inputs to and outputs from a system, they observed the effect of goal specificity. The NSG (nonspecific goal condition) participants, who were not informed of the target values of the outputs, learned the system structure better than the SG (specific goal condition) participants, who were informed of the target values. Burns and Vollmeyer (2002) also found that the NSG participants conducted more hypothesis testing than the SG participants. From these results, they concluded that a nonspecific goal encouraged the participants to search actively in a hypothesis space. Therefore, a nonspecific goal might lead to better learning than a specific goal.

Present Study

A hypothesis space is usually huge, hence a hypothesis space search is performed based on constraints offered by attentional perspectives (van Joolingen & de Jong, 1997). In the present study we define the hypothesis space in which the participants initially search as an "initial hypothesis space" and the space containing a target rule to be discovered as a "target space."

In the earlier studies, the initial hypothesis space was typically decided by an experimenter because the participants were informed of all the relative factors of focus. This initial hypothesis space also contained the target rule to be discovered. There was no need for the participants to find the target hypothesis space, as the initial hypothesis space and target hypothesis space were identical (Figure 1(a)). Here, in contrast, we investigate situations in which the participants must find a target hypothesis space by themselves in order to discover an appropriate rule.

(1) Initial-space situation

One situation we deal with is the "initial-space situation" (Figure 1(b)). Participants are given an initial hypothesis space by an experimenter. This initial space, however, contains no rule to be discovered. The initial hypothesis space differs from the target hypothesis space. To discover the target rule, the participants need to shift a searching space from the initial hypothesis space to the target hypothesis space. This situation typically emerges in insight problem solving (Kaplan & Simon, 1990).

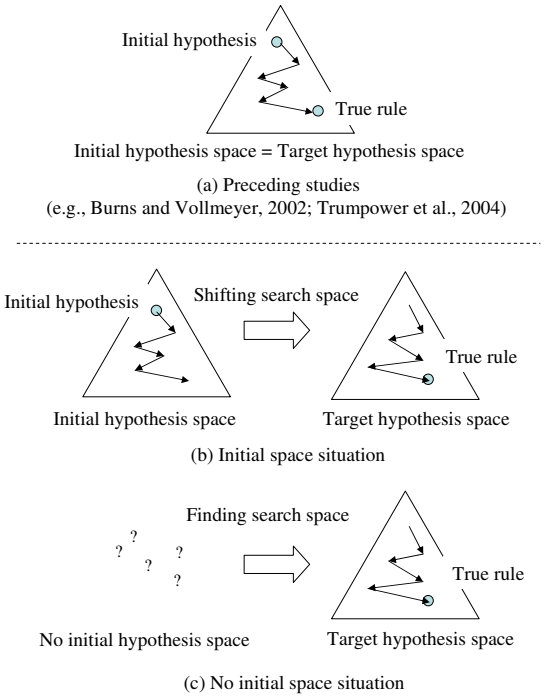


Figure 1: Conceptual diagrams of preceding and present studies

(2) No-initial-space situation

The other situation is the "no-initial-space situation" (Figure 1(c)). The participants in this case are not informed of any relative factors of focus, and thus receive no information on the initial hypothesis space to be searched. The participants have to find relative factors for the search from the initial stage. The investigation of this situation is important, as the size of a hypothesis space and the number of available hypotheses might affect the search strategies.

Aim of the Present Study

We investigate two hypotheses regarding the effect of goal specificity on a search in the dual spaces in the two situations: the initial-space situation and the no-initial-space situation.

Hypothesis 1: A nonspecific goal facilitates a search in a hypothesis space more than a specific goal. In other words, participants who are given a nonspecific goal may search more actively in a hypothesis space.

Hypothesis 2: As a hypothesis space is searched more, the performance in discovering the target rule improves.

Task

Figure 2 is a screen shot of the task for this study. The participants are asked to use the arrow buttons to pass the ball from player to player and to shoot for the basket. Two rules, one fake and one true, are valid in each game. These rules determine the relation between the arrow buttons and pass directions for the ball. In both rules, the up-arrow button corresponds to a certain direction and the other seven buttons correspond to the other seven directions relative to the up-arrow in clockwise rotation. The direction of the prior pass

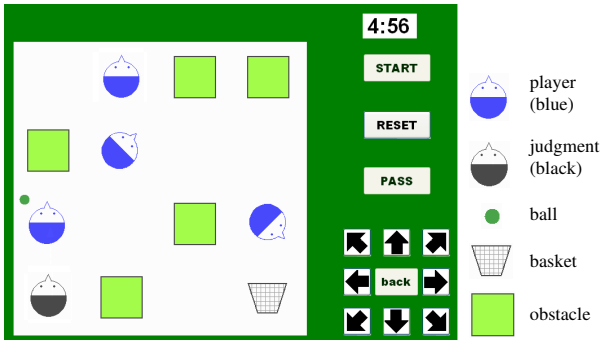


Figure 2: Screen shot of the task
A screen shot of one of the games during Phase 1. The participants pass the ball from player to player and shoot for the basket.

corresponds to the up-arrow button in the true rule, whereas the direction of the current player's nose corresponds to the up-arrow button in the fake rule. Although the fake rule is expected to be discovered more easily than the true rule, it can be valid in the initial games (not in all games). The true rule, in contrast, is valid in all games.

The participants, having discovered the fake rule, initially search in the hypothesis space consisting of hypotheses characterized by a face direction (the "face hypothesis space"). Even if the fake rule no longer validly works in the games, the participants continue to search in the face hypothesis space. To discover the true rule, the participants must shift a searching hypothesis space from the face hypothesis space to the hypothesis space made up of hypotheses characterized by the orbit of the prior pass (the "orbit hypothesis space"). The orbit hypothesis space contains the true rule. Thus, in this task, the initial hypothesis space means the face hypothesis space and the target hypothesis space means the orbit hypothesis space.

The experiment basically consists of three phases.

Phase 1: The participants engage in games in which both the fake and true rules are valid. The participants are asked to shoot the ball into the basket as many times as possible. We expect the participants to discover the fake rule and use it in Phase 1.

Phase 2: Only the true rule can be applied in the games in Phase 2. At the beginning of Phase 2, the participants are expected to search in the face hypothesis space, based on their successes with the fake rule accumulated in Phase 1. To discover the true rule, the participants must shift a searching hypothesis space from the face hypothesis space to the orbit hypothesis space. We manipulate experimental factors and observe how these factors affect the searches in the hypothesis space. The playing time is limited in Phases 1 and 2. The participants are told that the games in these phases are for training, and that the real games, or the actual takes, will come in Phase 3.

Phase 3: The participants are informed the real games, or actual takes, come in Phase 3. They are asked to shoot the ball into the basket. Phase 3 consists of two games, each of which

is played to completion without a time limit imposed. The true rule is valid in both the first and second games, but the fake rule can be applied to only the first game. We, the researchers, judge whether or not the participants discover the true rule based on their performances in each game.

Manipulation of Goal Specificity

Goal specificity is manipulated mainly through the following three experiments. The participants are given a specific goal (the SG condition) and asked to play the games in Phase 2 (the screen shot in Figure 2 shows a Phase 2 game). The basket and obstacles determine only one pass route. Thus, the participants' next moves are specified. Meanwhile, the participants in the other group are given a nonspecific goal (the NSG condition) and asked to play games in which there are no obstacles and in which the basket is replaced by a player. In this situation, the next move is unspecified: a participant can intentionally select one of several valid passes without a specific final goal (basket). With these manipulations in the games come differences between the instructions under the SG and NSG conditions in Phase 2. The participants in the SG condition are asked to shoot the ball into the basket, whereas the participants in the NSG condition are asked to pass the ball from player to player. In both conditions, the participants are asked to perform as many games as possible within the time limit. The time point of every button selection by a participant is recorded.

Measurement The hypothesis space in this task consists of hypothesized rules on the relations between the arrow buttons and the pass directions. An instance space consists of all instances; each instance is described as "when a certain arrow button is selected, a ball is thrown to a certain direction under a certain situation." Assuming that a search in one space is performed after a search in the other space, in turn, a hypothesis space search is performed during the period elapsed between the receipt of one pass result (the result of one pass thrown) to the receipt of the next pass result. Therefore, in this study, we use a time interval of two successive passes as a measurement for the amount of searches in a hypothesis space. Henceforth we refer to this time interval as the "pass interval time."

We judged whether each participant discovered the true rule from his or her performance in Phase 3. If the participants could not discover the true rule, the adjustment strategy minimized errors. The participants who use the adjustment strategy make a pass at first based on some criterion or randomly, and then adjust the direction of arrow buttons in order to minimize the difference between the expected and actual pass directions. We defined the successful participants as the participants whose error rate was lower than the expected error rate when they use the adjustment strategy.

Experiment 1

We conducted Experiment 1 to investigate the effect of goal specificity on a search in a hypothesis space in the initial-space situation (see Figure 1(b)). In addition, we manipulated another factor, the instruction factor, to test whether the pass interval time is valid as a measurement of the amount of searches in a hypothesis space. In the search-oriented condition (the SO condition), the participants were asked to find a rule that determines the relation between the arrow buttons and pass directions. By contrast, in the non-search-oriented condition (the NSO condition), the participants were told nothing about the rule. This manipulation may lead the participants in the SO condition to search more in the hypothesis space, compared to the participants in the NSO condition. If the pass interval time correlates with the amount of searches in the hypothesis space, the pass interval time of the participants in the SO condition will exceed that of the participants in the NSO condition.

Method

Participants Sixty-four undergraduates participated in Experiment 1. Each was assigned to one of four conditions: goal specificity (SG and NSG) $\times$ instruction (SO and NSO).

Task and Procedure Experiment 1 was conducted in small groups of three or fewer participants. After the participants received a basic explanation of the procedures, the participants briefly rehearsed the task. Next, they carried out the task in the three phases. Phase 1 and Phase 2 each lasted for five minutes. In Phase 2, two factors: the participants' search preferences in the hypothesis space by the instruction and goal specificity, were manipulated. Finally, in Phase 3, all participants played two games in an identical situation without a time limit imposed.

Results and Discussion

Pass Interval Time Figure 3 presents the average pass interval time in each condition in Phase 2. A two-way ANOVA ((goal specificity: SG and NSG) $\times$ (instruction: SO and NSO)) was performed on the pass interval times in Phase 2. The interaction between the two factors was not significant ($F(1,60) = 0.673, n.s.$). The main effects of both the goal specificity factor ($F(1,60) = 38.454, p < .001$) and the instruction factor ($F(1,60) = 5.030, p < .05$) reached significance.

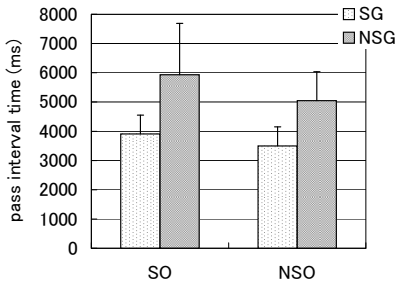


Figure 3: Average pass interval time in each condition in Phase 2 in Experiment 1 (bars show standard errors)

For the instruction factor, the pass interval time of the participants in the SO condition was longer than that of the participants in the NSO condition. The instruction given in the SO condition, the instruction which encouraged the participants to find a rule, increased the pass interval time. Noting

that this implied a correlation between the pass interval time and the amount of searches in a hypothesis space, we decided to use the pass interval time as a measurement of the amount of searches in a hypothesis space.

For the goal specificity factor, the pass interval time of the participants in the NSG condition was longer than that of the participants in the SG condition. This corroborated the first hypothesis: a nonspecific goal facilitates a search in a hypothesis space.

Proportion of Successful Participants Next, we analyzed the proportion of participants who discovered the true rule in each condition. For the instruction factor, 6 of 33 participants in the SO condition and 2 of 31 participants in the NSO condition were successful. There was no significant difference between the two conditions ($p > .10$). Similarly, for the goal specificity factor, 4 of 30 participants in the SG condition and 4 of 34 participants in the NSG condition were successful. Again, there was no significant difference between the two conditions ($p > .10$). Hence, these results did not confirm the second hypothesis: more searches in a hypothesis space improve performance in discovering the target rule.

Experiment 2

In Experiment 2 we investigated the effect of goal specificity on a search in a hypothesis space in the no-initial-space situation (see Figure 1(c)). We also manipulated the instruction factor to test the validity of the pass interval time, as was done Experiment 1.

Method

Participants Sixty-four undergraduates participated in Experiment 2. Each was assigned to one of four conditions: goal specificity (SG and NSG) $\times$ instruction (SO and NSO).

Task and Procedure The task in Experiment 2 was almost the same as that in Experiment 1, with the following adjustments. No Phase 1 was conducted in Experiment 2. The faces were removed from the players and the referee, and replaced with blue- and black-filled circles. The participants did not acquire the initial hypothesis space, as they were given no perspectives on which to focus for forming hypotheses at the beginning of the task.

Results and Discussion

Pass Interval Time Figure 4 presents the average pass interval time in each condition in Phase 2. A two-way ANOVA ((goal specificity: SG and NSG) $\times$ (instruction: SO and NSO)) was performed on the pass interval times in Phase 2. The interaction between the two factors was not significant ($F(1,60) = 0.022, n.s.$). The main effects of both the goal specificity factor ($F(1,60) = 6.708, p < .05$) and the instruction factor ($F(1,60) = 4.056, p < .05$) reached significance.

In the analysis for the instruction factor, this result was consistent with the result in Experiment 1. The pass interval time of the participants in the SO condition was significantly longer. The correlation between the pass interval time and the searches in a hypothesis space was again supported.

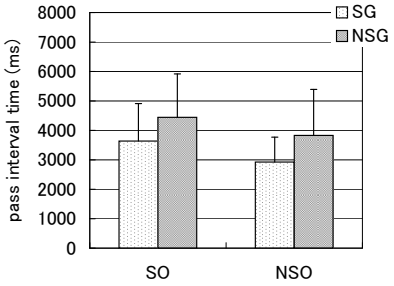


Figure 4: Average pass interval time in each condition in Phase 2 in Experiment 2 (bars show standard errors)

In the analysis for the goal specificity factor, the pass interval time was longer in the NSG condition than in the SG condition. This result also confirms the first hypothesis, corroborating the finding of Experiment 1. Note, however, that the difference between the SG and NSG conditions was much smaller in Experiment 2 than in Experiment 1. We will discuss this difference further in a later section.

Proportion of Successful Participants Next, we analyzed the proportion of participants who discovered the true rule in each condition. In the analysis for the instruction factor, 4 of 32 participants in the SO condition and 5 of 32 participants in the NSO condition were successful. There was no significant difference between the two conditions ($p > .10$). The result was similar in the analysis for the goal specificity: 3 of 32 participants in the SG condition and 6 of 32 participants in the NSG condition were successful. Again, there was no significant difference between the two conditions ($p > .10$). Hence, these results did not confirm the second hypothesis.

Comparison between Experiments 1 and 2

By comparing the results of Experiments 1 and 2, we could explore how the existence of the initial hypothesis space affected the effect of goal specificity. A premise for this study was dual spaces for search: the hypothesis space and instance space. Yet the participants in the NSO condition may not have assumed any hypothesis space, as they may not have noticed the rule determining the pass directions. For this reason, the following analysis focused on the participants in the SO condition.

In these experiments we introduced what we called the "situation factor," manipulating whether or not the participants had the initial hypothesis space across Experiments 1 and 2. At the beginning of Phase 2, the participants in Experiment 1 had the initial hypothesis space. Recollecting their accumulated successful experiences with the fake rule in Phase 1, they directed their attention to the face hypothesis space. This situation was called the "initial-space condition" (the IS condition). In contrast, the participants in Experiment 2 did not acquire an initial hypothesis space or experience any game play in Phase 1. And by removing the faces as cues from the players of the games in Phase 2, we deprived the participants of perspectives for forming hypotheses. This situation was called the "no-initial-space condition" (the NIS condition).

A two-way ANOVA ((situation: IS and NIS) $\times$ (goal speci-

ficity: SG and NSG)) was performed on the pass interval times in Phase 2. As a result, a marginally significant interaction between the situation and goal specificity factors was revealed ($F(1,61) = 3.158, p = .081$). In the IS condition, the pass interval time was longer in the NSG condition than in the SG condition ($F(1,61) = 17.449, p < .001$). This effect, however, disappeared in the NIS condition ($F(1,61) = 2.769, n.s.$). Both the goal specificity and situation factors had significant effects ($ps < .05$).

In this comparison, the participants with a nonspecific goal had a longer pass interval time than the participants with a specific goal in the IS condition. Meanwhile, the goal specificity factor had no effect on the pass interval time in the NIS condition. Thus, the first hypothesis was confirmed only in the IS condition, and not in the NIS condition.

Experiment 3

In Experiments 1 and 2, we found that goal specificity had no effect on a search in a hypothesis space when the participants lacked an initial hypothesis space. In Experiment 3, we manipulated both the goal specificity and situation factors to confirm the effect of these factors directly. Several of the experimental procedures were improved for this experiment. First, Phase 1 was performed in both the IS and NIS conditions, so that the participants would begin Phase 2 with identical prior experiences. In the NIS condition, the faces of the players and referee were removed in Phase 2 to eliminate the initial hypothesis space. Second, only a few participants successfully discovered the target rule in Experiments 1 and 2. In Experiment 3, the players who threw a successful pass and the receiver from the previous trial were marked visually on the game display. This cue lowered the memory loads of the participants, thus helping the participants discover the true rule in the orbit hypothesis space more easily.

Method

Participants Seventy-four undergraduates participated in Experiment 3. Each was assigned to one of four conditions: situation (IS and NIS) $\times$ goal specificity (SG and NSG).

Task and Procedure Experiment 3 was conducted in small groups of three or fewer participants. To control prior experiences, the participants in all conditions played games in all three phases. In Phase 1, the participants played games in which both the fake and true rules were valid, over a total play time of five minutes. In Phase 2, the participants played games in which only the true rule was valid. The goal specificity factor was manipulated by the same method used in the prior two experiments. Additionally, the situation factor was manipulated by adjusting the players' faces. The participants in the IS condition played the games with normal face players, as they had in Experiment 1. Meanwhile, the participants in the NIS condition played the games with faceless players, as they had in Experiment 2. Unlike Experiments 1 and 2, the game time in Phase 2 was increased to seven minutes in order to increase the number of successful participants. All participants were instructed that there was a rule valid through all of the games. Finally, Phase 3 was conducted using the same player faces used in Phase 2, but without a time limit.

Results and Discussion

Pass Interval Time Figure 5 presents the average pass interval time in each condition in Phase 2. A two-way ANOVA ((situation: IS and NIS) $\times$ (goal specificity: SG and NSG)) was performed on the pass interval times in Phase 2. The interaction between the situation and goal specificity factors reached significance ($F(1,70) = 4.989, p < .05$). In the IS condition, the pass interval time was longer in the NSG condition than in the SG condition ($F(1,70) = 9.078, p < .005$). This effect disappeared, however, in the NIS condition ($F(1,70) = 0.021, n.s.$). The goal specificity and situation factors both had significant effects ($ps < .05$).

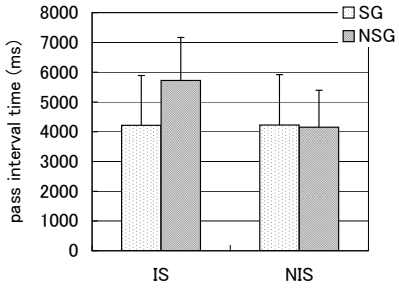


Figure 5: Average pass interval time in each condition in Phase 2 in Experiment 3 (bars are standard errors)

The result of Experiment 3 was consistent with the results of Experiments 1 and 2. In the IS condition, where the participants acquired the initial hypothesis space, goal specificity had an observable effect on a search in a hypothesis space. The participants with a nonspecific goal searched in a hypothesis space more actively than the participants with a specific goal. This effect was not observed, however, in the NIS condition, where the initial hypothesis space was eliminated by the change of the game display. Therefore, the presence or absence of an initial hypothesis space affected the goal specificity effect in a search in a hypothesis space. The first hypothesis is confirmed only in the IS condition.

Proportion of Successful Participants Next, we analyzed the proportion of participants who discovered the true rule in each condition (Figure 6). In the IS condition, 2 participants discovered the true rule in the SG condition and 8 participants discovered the true rule in the NSG condition.

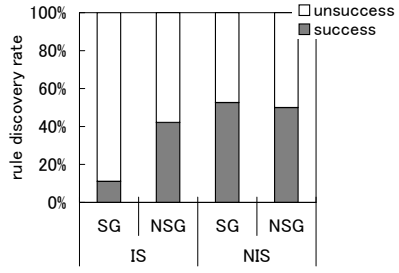


Figure 6: Proportion of successful participants in each condition in Experiment 3

The participants in the NSG condition discovered the true rule more frequently than the participants in the SG condition ($\chi^2(1) = 4.502, p < .05$). In the NIS condition, 10 participants discovered the true rule in the SG condition and 9 participants discovered the true rule in the NSG condition. There was no significant difference between the SG and NSG conditions in the NIS condition ($p > .10$).

In the IS condition, the pass interval time of the participants in the NSG condition was longer than that in the SG condition. Additionally, the proportion of successful participants in the NSG condition was also greater than that in the SG condition. Thus, the second hypothesis is confirmed.

Discussion and Conclusions

In this study we investigated the following two hypotheses in two situations, an initial-space situation and a no-initial-space situation: (1) A nonspecific goal facilitates a search in a hypothesis space rather than a specific goal. (2) As a search in a hypothesis space is more actively performed, the performance in discovering the target rule improves.

From the results of three experiments, the first hypothesis was partially confirmed. The effect of goal specificity on a search in a hypothesis space depended on whether or not the participants noticed an initial hypothesis space. When the participants noticed an initial hypothesis space, goal specificity had an observable effect on a search in a hypothesis space. The participants with a nonspecific goal searched in a hypothesis space more actively than the participants with a specific goal. On the other hand, this effect of goal specificity was not observed when the participants were not given any initial hypothesis space.

The second hypothesis was confirmed in the results of Experiment 3. The pass interval time of the participants with a nonspecific goal was longer than that of the participants with a specific goal in the IS condition. Additionally, the proportion of successful participants with a nonspecific goal was larger than that of successful participants with a specific goal in the IS condition. However, there was no significant difference between the specific goal and nonspecific goal conditions in Experiments 1 and 2. This may have been due to a floor effect, as only a few participants discovered the true rule in these experiments.

The results of the present study are consistent with the finding of Klahr and Dunbar (1988). They defined hypotheses as the forms of a "frame." In their study, they classified the frames (hypotheses) into several types, according to their features. The hypothesis spaces in our study, i.e., sets of hypotheses sharing a common feature, could be explained by the types of frames defined by Klahr and Dunbar (1988). The theorists in Klahr's experiments preferred to do their searches in hypothesis spaces. They were able to switch the hypotheses types correctly, within short periods of time and over the course of only a few experiments, and discovered the rule rapidly. In our study, the participants with nonspecific goals behaved like the theorists in the situation where the initial hypothesis space was given. They preferred to search in a hypothesis space, repeating the behavior of the theorists in the earlier studies. They were able to switch the searching hypothesis space from a given initial hypothesis space to a target hypothesis space with fewer instances, and discovered the true rule. In contrast, the participants with a specific goal preferred to search in an instance space, repeating the behavior of the experimenters defined in Klahr's study.

In the initial-space situation we observed the effect of goal specificity on a search in a hypothesis space, duplicating the results from earlier studies. This situation is identical to situations covered in the preceding studies, where participants were given an initial hypothesis space. Unlike the preceding studies, we used a task in which the true rule was not included in the initial hypothesis space. To discover the true rule, the participants needed to shift their attention to the target hypothesis space. Here, the effect of goal specificity on a search in a hypothesis space was still confirmed.

Yet when the participants were given no initial hypothesis space, goal specificity had no observable effect on a search in a hypothesis space. To state hypotheses, the participants initially needed to find a focused hypothesis space by themselves in this situation. We assume that they searched in an instance space to collect data as cues for determining a hypothesis space to search. This may explain why the hypothesis space search was not activated for the participants with the nonspecific goal. The SDDS model proposed that a discovery process is controlled with three main components: "search hypothesis space," "test hypothesis," and "evaluate evidence." The search hypothesis space component corresponds to a search in a hypothesis space in our study. This component contains a search in an experiment (instance) space as one of the sub lower components. Participants could collect data and find a pattern of these data gathered through the experiment space search, and state hypotheses. Similarly, the participants in our study who were given no initial hypothesis space needed cues to find a focused hypothesis space and state hypotheses. Therefore, we conclude that they searched in an instance space regardless of goal specificity.

References

Burns, B. D., & Vollmeyer, R. (2002). Goal specificity effects on hypothesis testing in problem solving. *The Quarterly Journal of Experimental Psychology, 55A*, 241–261.

Kaplan, C. A., & Simon, H. A. (1990). In search of insight. *Cognitive Psychology, 22*, 374–419.

Klahr, D. (2000). *Exploring science*. The MIT Press.

Klahr, D., & Dunbar, K. (1988). Dual space search during scientific discovery. *Cognitive Science, 12*, 1–48.

Simon, H. A., & Lea, G. (1974). Problem solving and rule induction: A unified view. In I. L. W. Gregg (Ed.), *Knowledge and cognition.* Potomac, Md: L. Erlbaum Associates.

Sweller, J., & Levine, M. (1982). Effects of goal specificity on means-ends analysis and learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 8*, 463–474.

van Joolingen, W. R., & de Jong, T. (1997). An extended dual search space model of scientific discovery learning. *Instructional Science, 25*, 307–346.

Speaker's choice of frame based on rarity information

Hidehito Honda (hito@muscat.L.chiba-u.ac.jp)
Department of Cognitive & Information Science, Chiba University
1-33, Yayoi-cho, Inage-ku, Chiba-shi, Chiba, 263-8522, Japan

Toshihiko Matsuka (matsukat@muscat.L.chiba-u.ac.jp)
Department of Cognitive & Information Science, Chiba University
1-33, Yayoi-cho, Inage-ku, Chiba-shi, Chiba, 263-8522, Japan

Abstract

Previous studies have discussed how speakers select a frame (e.g., "half full," or "half empty"), and have proposed a hypothesis such as *reference point hypothesis* (e.g., Sher & McKenzie, 2006, 2008). In this paper, we propose a new hypothesis, *frame choice based on information about rarity*. This hypothesis predicts that speakers tend to select a frame denoting a rare event. Four studies provide evidence that speakers' choice of frame is consistent with the prediction from our hypothesis. Furthermore, our hypothesis is reconciled with the *positive bias* in frame choice, which cannot be accounted for by the reference point hypothesis. We discuss the possibility that linguistic behaviors are widely explained from people's sensitivity to rarity information.

Keywords: Framing effect; speaker's choice of frame; reference point hypothesis; sensitivity to rarity; positive bias in frame choice

Introduction

Since Tversky and Kahneman (1981) documented the original research, many researchers have studied framing effect (for reviews, see Levin, Schneider, & Geath, 1998; Soman, 2004). One example of the framing effect is the "Asian disease problem" proposed by Tversky and Kahneman (1981):

> Imagine that the US is preparing for the outbreak of an unusual Asian disease, which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume that the exact scientific estimates of the consequences of the programs are as follows:
>
> If Program A is adopted, 200 people will be saved.
>
> If Program B is adopted, there is a 1/3 probability that 600 people will be saved and a 2/3 probability that no people will be saved.

For this problem, a majority of the participants preferred Program A to Program B. Another group was presented with the same cover story, but with the two programs rephrased:

> If Program C is adopted, 400 people will die.
>
> If Program D is adopted, there is a 1/3 probability that nobody will die, and a 2/3 probability that 600 people will die.

Although Program C is only a rewording of Program A and Program D is a rewording of Program B, a majority of the participants preferred Program D to Program C. Thus framing effect refers to the effects such that the framing of a problem influences decision making.

Many studies on the framing effect have examined how listeners, or those presented with frames, behave based on the frames. Various models have been proposed to explain the framing effect. However, relatively few studies have been conducted on how speakers frame a problem. For instance, what influences speakers to describe the Asian disease problem with the "save" frame or "die" frame?

Some researchers have recently discussed how speakers frame outcomes (Keren, 2007; McKenzie & Nelson, 2003; Sher & McKenzie, 2006, 2008; Teigen & Karevold, 2005; van Buiten & Keren, 2009; Wang, 2004). For example, McKenzie and his associates have examined problems such as "Which do speakers select to describe a 4-ounce cup with 2 ounces of water, half full or half empty?", and have proposed the *reference point hypothesis* (McKenze & Nelson, 2003; Sher & McKenzie, 2006, 2008). This hypothesis assumes that a speaker tends to use a frame that corresponds to the label that has increased. In the above example, the reference point hypothesis predicts that a speaker uses the full frame when a cup has been previously empty, and that a speaker uses the empty frame when a cup has been full of water. Therefore, the reference point influences speaker's choice of frame.

The reference point hypothesis is intriguing in that it not only predicts how a speaker selects a frame, but also explains why decision makers are influenced by framing (Sher & McKenzie, 2006, 2008). However, we point out that the reference point hypothesis does not predict one of the interesting findings of frame choice, *positive bias*, which has been repeatedly reported in the previous studies. The positive bias refers to the tendency that in choosing from two frames which have positive and negative valenced meanings such as "gain"-"loss" or "success"-"failure," people tend to prefer the positive valenced frame (e.g., Keren, 2007; Sher & McKenzie, 2006; van Buiten & Keren, 2009; Wang, 2004). For example, Sher and McKenzie (2006) showed that in describing results of the last 50 projects in which 20 projects have succeeded and 30 projects have failed, participants generally used a positive frame (e.g., 20 out of the last 50 projects have succeeded) rather than a negative one (e.g., 30 out of the last 50 projects have failed). In Wang (2004), participants were presented with probabilistic life-death or monetary problems by pie

charts, and asked to complete sentences that summarized the problems. It was found that participants tended to complete sentences with positive frames (e.g., save, help) rather than negative ones (e.g., killed, die).

These findings suggest that psychological mechanisms other than those explained from the reference point hypothesis exist when speakers select a frame.

Choice of frame based on rarity information

We propose a new hypothesis, *frame choice based on rarity information*. We predict that information about rarity influences choice of frame, and that the speakers frame outcomes in terms of rarity. Consider the following problems: There is a die colored both black and white. One of the 6 sides of this die is black, and the other 5 sides are white. In rolling this die, the occurrence of black side is rare. In contrast, the occurrence of white side is common. We predict that when speakers describe results of rolls of this die, they prefer using the black frame because the occurrence of black side is expected to be rare. Imagine that someone rolls this die 6 times and the black side came up once, and the white sides came up 5 times. We predict that s/he will describe the results, "With 6 rolls, black came up once", rather than "white came up 5 times." Hence, our hypothesis states that speakers focus on the rarity and prefer using the frame describing rare events rather than those describing common events.

This hypothesis is based on the findings about hypothesis testing. Previous studies on hypothesis testing have shown that people are very sensitive to information on rarity, and that they adaptively use such information in hypothesis testing (e.g., Klayman & Ha, 1987; McKenzie & Mikkelsen, 2000; Oaksford & Chater, 1994). Furthermore, the finding in McKenzie, Ferreira, Mikkelsen, & McDermott (2001) is more relevant. They examined the people's sensitivity to rarity in the context of how to phrase a conditional hypothesis. Imagine the conditional hypothesis, "If X1, then Y1," where each variable, X and Y, has two levels (X1 and X2, Y1 and Y2). In this case, this hypothesis can be denoted with another form, "If X2, then Y2." McKenzie et al. (2001) showed that when participants observed rare X1 & Y1 and common X2 & Y2, they tended to phrase the conditional hypothesis "If X1, then Y1" rather than "If X2, then Y2," suggesting that people phrase a conditional hypothesis in terms of rarity. Although this finding in McKenzie et al. (2001) was limited to how to phrase conditional hypothesis, other linguistic behaviors such as frame choice might be explained from the same perspective. That is to say, speakers choose a frame in terms of information about rarity.

In this paper, we conducted 4 studies, and examined our hypothesis regarding to the speaker's choice of frame. In Study 1, we conducted an experimental study to examine our hypothesis. In Studies 2-A, 2-B, 2-C, we discussed the positive bias in frame choice, and examined whether our hypothesis is reconciled with the positive bias.

Study 1

In study 1, we examined our hypothesis using a frame choice task. We predict that frame choice is influenced by information about rarity. In particular, participants will choose a frame describing a rare event.

Method

Participants. The participants were 614 Aoyama Gakuin University students, who received partial course credit. There were from 64 to 72 participants in each of nine conditions (see Table 1).

Task and experimental conditions. We conducted a frame choice task that was analogous to that in McKenzie and Nelson (2003) using a questionnaire. In one of the 9 conditions, participants read the following story:

> There is a die that is painted black on one side and painted white on the other five sides. You have rolled this die 6 times, and the results are as follows:
>
> | Side of the die | Frequency |
> |---|---|
> | Black | 1 |
> | White | 5 |
>
> Which is the most natural way to describe these results, "The die came up black 1 out of 6 times" or "The die came up white 5 out of 6 times"?[1]

In this question, participants were required to choose one of two frames (i.e., "black" frame or "white" frame) to describe the outcomes.

There were 9 experimental conditions. Three dies differed in the color (i.e., black rare, white rare, black-white equal), and there were three patterns of outcomes from the roll of die. These three dies and three outcomes were varied orthogonally with respect to one another (see Table 1).

Results and discussion

Figure 1 shows proportions of black frame choice for 9 conditions. It was found that in describing the 3 outcomes (i.e., Black1-White5, Black3-White3, Black5-White1), participants in the Black-rare condition significantly preferred the black frame than those in the White-rare condition in each of the 3 outcomes ($p<.0001$, Fisher's exact test). We also found general preference for the rare side frame. 67.6% of participants in the three Black-rare conditions significantly chose the black frame, and 62.4% of those in the three White-rare conditions significantly chose the white frame ($p<.001$, binomial test).

In the Equivalent conditions, wherein explicit information about rarity was not available to participants, 52.2% of participants in the 3 Equivalent conditions chose the black frame. This result indicated that participants did not have a

[1] The order of these options was reversed for half of the participants in each condition in each experiment (Studies 1 and 2-C).

Table 1. 6 conditions in Experiment 1.

Die (number of side)	Outcome (Black, White)
Black-rare (Black1-White5)	(1,5; n=71)
	(3,3; n=68)
	(5,1; n=71)
White-rare (Black5-White1)	(1,5; n=64)
	(3,3; n=72)
	(5,1; n=66)
Equivalent (Black3-White3)	(1,5; n=71)
	(3,3; n=64)
	(5,1; n=67)

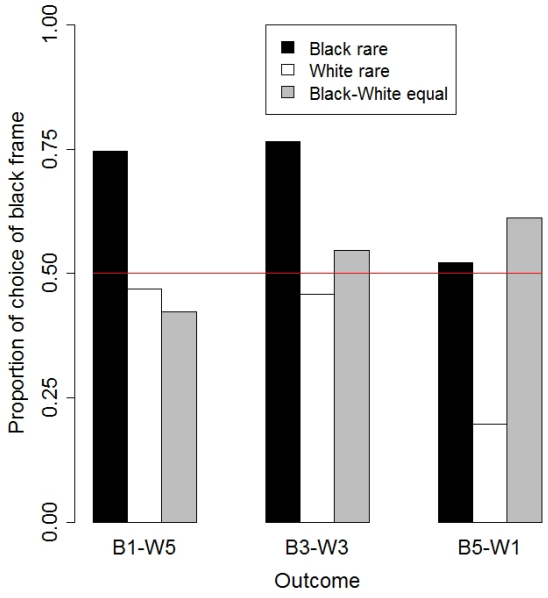


Figure 1. Proportion of black frame choice in Study 1

specific preference between the two frames ($p=.574$, binomial test).

Taken together, these results show that information about rarity of side of a die influenced participants' choice of frame. In particular, participants preferred the rare side frame. In addition, when the information about rarity was not available, participants were indifferent between the two frames. Hence, these results supported our hypothesis about frame choice based on information about rarity.

Study 2

In Study 2, we discuss whether the hypothesis about frame choice based on rarity information is reconciled with the positive bias in frame choice (e.g., Keren, 2007; Sher & McKenzie, 2006; van Buiten & Keren, 2009; Wang, 2004).

Why do people prefer a positive frame? Our hypothesis is that the positive bias derives from belief of rarity about what positive or negative words describe. We predict that people generally have the belief that what positive words describe are rarer than what negative words describe, and this belief influences frame choice. In other words, speakers tend to prefer a positive frame because of its rarity. Therefore, if people explicitly know that a negative frame describes a rare event and a positive frame describes a common event, preference for the positive frame will disappear.

In order to examine this hypothesis, we conducted three studies. In study 2-A, we examined whether the positive bias observed in laboratory experiments is also observed in a naturalistic environment. In study 2-B, we examined belief of rarity about what positive and negative words describe. In study 2-C, we conducted an experimental study and tested whether the positive bias disappears when participants explicitly know that a negative frame denotes a rare event and a positive frame denotes a common event.

Study 2-A

The positive bias in frame choice reported in the previous studies suggests that people generally prefer using positive expressions rather than using negative ones. Study 2-A examined whether a positive bias is observed in a naturalistic environment. Specifically, we counted a number of articles in a Japanese newspaper that contains positive or negative words. If positive bias is to be observed, there ought to be more articles containing positive words than those containing negative words.

Method

We used 26 positive-negative Japanese pairs of antonyms for this study. Table 2 illustrates 5 examples of positive-negative pairs of antonyms. These 26 pairs were selected using the following procedure. First, one rater, who did not know the hypothesis of the current study, randomly picked out 35 pairs of antonyms that he thought had positive-negative valenced meanings from Japanese dictionary of antonyms (Kitahara & Togo, 1989). Then two other raters, neither of whom knew the hypothesis, judged whether each of the 35 pairs had positive-negative meanings. We adopted 26 pairs (i.e., 52 words) that these theree two raters regarded as having positive-negative meanings.

Then we counted a number of articles in a Japanese newspaper. We used *Yomidasu* as the search system. This search system includes the data-base of *Yomiuri shibun*, which is one of the most subscribed newspapers in Japan. Using this system, we counted the number of articles that had been published from January 1990 to December 2007. We conducted this search using each of the 52 words.

Results and discussion

We calculated the *positive bias index* (*P-Bias index*) for each of 26 pairs. In a certain positive-negative antonym pair, when the numbers of articles in which the positive or negative word is mentioned are N_p and N_n respectively, the P-Bias index is defined by the following equation:

$$P\text{-}Bias\ index = N_p/(N_p+N_n)$$

For example, when number of articles is 400 for a positive word and 100 for a negative word in a certain pair, the calculated P-Bias index is 0.8. Therefore, when the P-Bias index is more than 0.5, a positive word is used more often than a negative word in a positive-negative antonym pair. Figure 2 illustrates the P-bias index for 26 positive-negative antonym pairs. The mean value of the P-bias for the 26 pairs was 0.678 (*SD*=0.240, maximum=0.997, minimum=0.065), and this value was significantly higher than 0.5 ($t(25)=3.78$, $p<.001$). These results suggest that in positive-negative antonym pairs, people tend to use positive words more often than negative words in a naturalistic environment.

Study 2-B

In Study 2-B, we examined belief of rarity about what positive and negative words describe. We predicted that people generally have the belief that what positive words describe is rarer than what negative words describe.

Method:

Participants. The participants were 116 Aoyama Gakuin University students, who received partial course credit.

Task and materials. Participants were asked about their belief of rarity on what positive and negative words describe using a questionnaire. The question was as follows:

> There are 26 pairs in this booklet. Two words in each of pairs have opposite meanings. Imagine "people," "things," or "outcomes" that are described by each of the words in a pair. Then which do you think is more unusual to become such people, to make such things, or to achieve such outcomes?

For this question, participants were required to choose either a positive or negative word from a pair. We used the same 26 pairs that were used in Study 2-A. If it is unusual to become, make, or achieve what a word describes, what the word describes must be rare. Hence we assume that a selected word in a pair is judged to refer to something rarer than the reference of the other word in the pair.

Results and discussion

We calculated the proportion of positive word choice for each of the 26 pairs. Figure 3 shows the proportions for the 26 pairs. In 21 out of 26 pairs, participants significantly chose positive words rather than negative words ($p<.05$, binomial test), and no negative words were chosen with more than 50%. Hence these results suggest that people have the belief that what positive words describe are generally rarer than what negative words describe are.

Table 2. Examples of positive-negative pairs of antonyms used in Studies 2-A and 2-B.

positive words		negative words
best	-	worst
success	-	failure
rich	-	poor
safety	-	danger
usefulness	-	uselessness

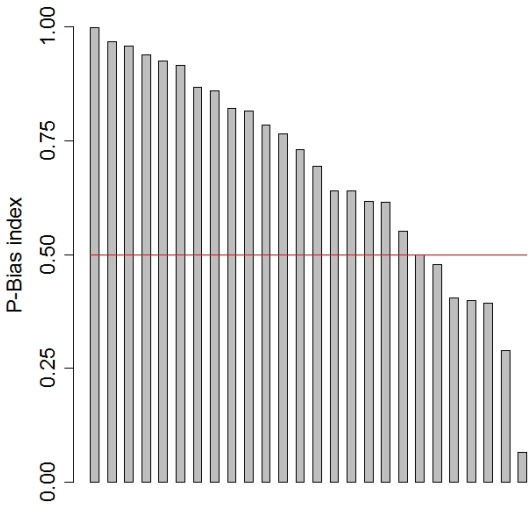


Figure 2. P-Bias index for 26 pairs in Study 2-A.

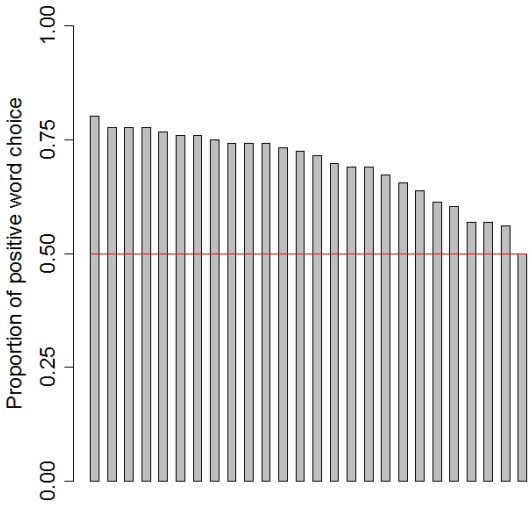


Figure 3. Proportion of positive word choice for 26 pairs in study 2-B.

Study 2-C

Studies 2-A and 2-B indicated that positive bias in frame choice is observed in a naturalistic environment, and that people generally have the belief that what positive words describe are rarer than what negative words describe. We hypothesize that positive bias in frame choice derives from this belief about rarity, and that speakers tend to select a positive frame because of its rarity. Hence, our hypothesis

predicts that when people explicitly know that a negative frame describes a rare event and a positive frame describes a common event, they will choose the negative frame rather than the positive frame. We examined this prediction conducting an experiment.

Method

Participants. The participants were 689 Aoyama Gakuin University undergraduate students, who received partial course credit. There were from 70 to 81 participants in each of nine conditions (see Table 3).

Task and experimental conditions. Task and experimental conditions were the same as those in Study 1 with the exception of the labels of dies. In place of the black-white labels, we used *winning-losing*[2] labels, which have positive and negative meanings. In one of the 9 conditions, participants read the following story:

> There is a die that is described "winning" on one side and described "losing" on the other five sides. You have rolled this die 6 times, and results are as follows:
>
> | Side of the die | Frequency |
> |---|---|
> | Winning | 1 |
> | Losing | 5 |
>
> Which is the most natural way to describe these results, "The die came up winning 1 out of 6 times" or "The die came up losing 5 out of 6 times"?

As in the Study 1, participants were required to choose one of two frames (i.e., "winning" frame or "losing" frame) to describe the outcomes.

For this task, there were 9 experimental conditions as in Study 1. Three dies differed in the description (i.e., winning rare, losing rare, winning-losing equal), and there were three patterns of outcomes from roll of die. These three dies and three outcomes were varied orthogonally with respect to one another (see Table 3).

Results and discussion

Figure 4 shows proportions of winning frame choice for 9 conditions. If the positive bias is observed in the frame choice, the winning frame will be chosen irrespective of rarity of sides in a die. However, the observed choice patterns were not consistent with this prediction. In each of the three outcomes, participants in the Winning-rare condition significantly preferred the winning frame than those in the Losing-rare condition ($p<.0001$, Fisher's exact test). Thus, the rarity of sides in a die influenced frame choice. As a general preference of frame, 78.7 % of the participants in the three Winning-rare conditions significantly preferred the winning frame ($p<.0001$, binomial test). However, only 50.6 % of the participants in the three Losing-rare conditions preferred the winning frame, and this preference was not significant ($p=.90$, binomial test). These results show

[2] Original Japanese labels were "atari" and "hazure." "atari" means winning lotteries, and "hazure" means losing lotteries.

Table 3. 6 conditions in Experiment 2.

Die (number of side)	Outcome (winning, losing)
Winning-rare (Winning1-Losing5)	(1,5; n=75) (3,3; n=70) (5,1; n=76)
Losing-rare (Winning5-Losing1)	(1,5; n=79) (3,3; n=79) (5,1; n=73)
Equivalent (Winning3-Losing3)	(1,5; n=76) (3,3; n=80) (5,1; n=81)

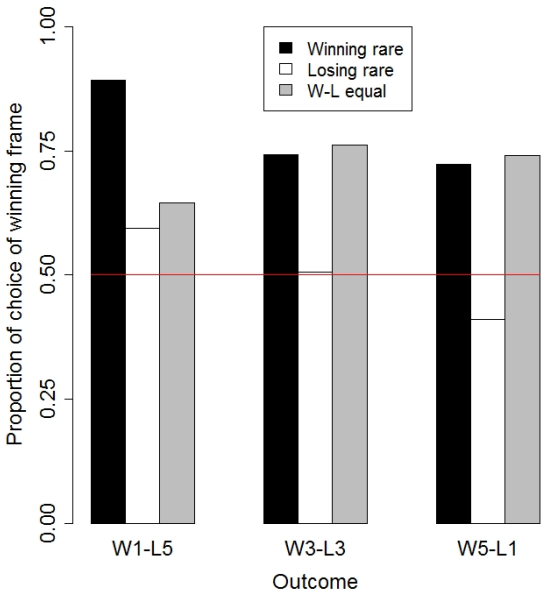


Figure 4. Proportion of winning frame choice in Study 2-C.

that positive bias did not prevail irrespective of information about rarity, and that participants' frame preference shifted from the winning frame to the losing frame when they explicitly knew that the losing frame denoted a rare event.

According to the frame choice based on rarity information, participants in the Losing-rare conditions will prefer the losing frame. However, only 49.4% of participants in the three Losing-rare conditions preferred the losing frame. This result implies that choice of positive frame is a robust bias, and that even when explicit information about rarity was available, participants may have preferred the positive frame.

In the Equivalent conditions, wherein explicit information about rarity was not available to participants, positive bias was observed. In all of the three Equivalent conditions, 71.7% of participants preferred the winning frame ($p<.0001$, binomial test).

Taken together, our hypothesis about positive bias in frame choice was corroborated. Although participants generally preferred the positive frame, participants' preference shifted to the negative frame with the explicit information about rarity. In particular, positive bias disappeared with the

explicit information that the negative frame denotes a rare event and the positive frame denotes a common event.

General discussion

Through the 4 studies, we examined our hypothesis that people choose a frame based on the information about rarity. It was found that information about rarity influenced speakers' choice of frame. In particular, participants tended to prefer a frame denoting a rare event.

The reference point hypothesis (e.g., McKenzie & Nelson, 2003; Sher & McKenzie, 2006, 2008) argues that speakers are sensitive to an increase in proportion relative to a reference point, and use a frame that corresponds to the label that has increased. In short, the reference point hypothesis assumes that speakers select a frame based on a reference point. In contrast, our hypothesis assumes that speakers select a frame based on rarity information. It should be noted that our hypothesis does not necessarily contradict the reference point hypothesis. For example, our hypothesis does not make any predictions about speakers' choice of frame based on a specific reference point. It is mute as to which frame people use to express a content of a cup when a cup has been previously empty (or full of water). On the other hand, when a reference point adopted by speakers is not clear, the reference point hypothesis does not predict specific patterns of frame choice. For instance, the reference point hypothesis does not explain why speakers show the positive bias in frame choice. Therefore, the two hypotheses can be regarded as providing explanations for different psychological mechanisms on frame choice.

We indicated in Studies 2-A and 2-B that usage of positive and negative words in a naturalistic environment is also related to belief about rarity. McKenzie et al. (2001) showed that participants tended to phrase a conditional hypothesis in terms of rarity. These findings suggest that speakers are very sensitive to information about rarity, and that linguistic behaviors are widely explained from the perspective of sensitivity to rarity.

Furthermore, previous studies have suggested that people are very sensitive to rarity information in hypothesis testing (e.g., e.g., Klayman & Ha, 1987; McKenzie & Mikkelsen, 2000; Oaksford & Chater, 1994). The findings on linguistic behaviors and hypothesis testing imply that people have the strong intuition that information about rarity is very informative, and this intuition influences various behaviors as well as linguistic behaviors and hypothesis testing. Hence, reconsideration from the perspective of sensitivity to rarity will provide insightful findings for various human behaviors.

Acknowledgements

This work was in part supported by the Japan Society for the Promotion of Science KAKENHI (Grant No. 20700235) and the Support Center for Advanced Telecommunications Technology Research (SCAT).

References

Keren, G. (2007). Framing, intentions, and trust-choice incompatibility. *Organizational Behavior and Human Decision Processes, 103*, 238-255.

Kitahara, Y., & Togo, Y. (Eds.). (1989). Hantaigo, Taishogo, Ziten. Tokyo: Tokyodo Shuppan.

Klayman, J., & Ha, Y.-W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review, 94*, 211-228.

Levin, I. P., Schneider, S. L., & Geath, G. J. (1998). All frames are not created equal: A typology and critical analysis of framing effects. *Organizational Behavior and Human Decision Processes, 76*, 149-188.

McKenzie, C. R. M., Ferreira, V. S., Mikkelsen, L. A. McDermott, K. (2001). Do conditional hypotheses target rare events? *Organizational Behavior and Human Decision Processes, 85*, 291-309.

McKenzie, C. R. M., & Mikkelsen, L. A. (2000). The psychological side of Hempel's paradox of confirmation. *Psychonomic Bulletin and Review, 7*, 360-366.

McKenzie, C. R. M., & Nelson, J. D. (2003). What's a speaker's choice of frame reveals: Reference points, frame selection, and framing effects. *Psychonomic Bulletin and Review, 10*, 596-602.

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review, 101*, 608-631.

Sher, S., & McKenzie, C. R. M. (2006). Information leakage from logically equivalent frames. *Cognition, 101*, 467-494.

Sher, S., & McKenzie, C. R. M. (2008). Framing effects and rationality. In N. Chater & M. Oaksford (Eds.), *The probabilistic mind: Prospects for bayesian cognitive science* (pp. 79-96). New York: Oxford University Press.

Soman, D. (2004). Framing, loss aversion, and mental accounting. In D. J. Koehler & N. Harvey (Eds.), *Blackwell handbook of judgment and decision making* (pp. 379-398). Oxford: Blackwell Publishing.

Teigen, K. H., & Karevold, K. I. (2005). Looking back versus looking ahead: Framing of time and work at different stages of a project. *Journal of Behavioral Decision Making, 18*, 229-246.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211*, 453-458.

van Buiten, M., & Keren, G. (2009). Speaker-listener incompatibility: Joint and separate processing in risky choice framing. *Organizational Behavior and Human Decision Processes, 108*, 106-115.

Wang, X. T. (2004). Self-framing of risky choice. *Journal of Behavioral Decision Making, 24*, 1-16.

Discovering Structure by Learning Sparse Graphs

Brenden M. Lake and **Joshua B. Tenenbaum**
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology
{brenden, jbt}@mit.edu

Abstract

Systems of concepts such as colors, animals, cities, and artifacts are richly structured, and people discover the structure of these domains throughout a lifetime of experience. Discovering structure can be formalized as probabilistic inference about the organization of entities, and previous work has operationalized learning as selection amongst specific candidate hypotheses such as rings, trees, chains, grids, etc. defined by graph grammars (Kemp & Tenenbaum, 2008). While this model makes discrete choices from a limited set, humans appear to entertain an unlimited range of hypotheses, many without an obvious grammatical description. In this paper, we approach structure discovery as optimization in a continuous space of all possible structures, while encouraging structures to be sparsely connected. When reasoning about animals and cities, the sparse model achieves performance equivalent to more structured approaches. We also explore a large domain of 1000 concepts with broad semantic coverage and no simple structure.

Keywords: structure discovery, semantic cognition, unsupervised learning, inductive reasoning, sparse representation

The act of learning is not just memorizing a list of facts; instead people seem to learn specific organizing structures for different classes of entities. The color circle captures the structure of pure-wavelength hues, a tree captures the biological structure of mammals, and a 2D space captures the geographical structure of cities (Fig. 1a, 1c, 5a). How does the mind discover which type of structure fits which domain?

Discovering structure can be understood computationally as probabilistic inference about the organization of entities. Past work has tackled this problem by considering rings, trees, chains, grids, etc. as mutually exclusive hypotheses called *structural forms* (Kemp & Tenenbaum, 2008). Forms are defined by grammatical constraints on the connections between entities; for example the ring form constrains each color to have two neighbors (Fig. 1a). After considering all of the candidate forms, the structural forms model selects the best fitting form and instance of that form. This can be a powerful approach; the model selects a ring for colors, a tree for mammals, and a globe-like structure for world cities. These structures can then predict human inductive reasoning about novel properties of objects (Kemp & Tenenbaum, 2009).

Despite its power, the structural forms approach is not clearly appropriate when structures stray from the predefined forms, and such exceptions are common in real world domains. While the genetic similarity of animals is captured by an evolutionary tree,[1] everyday reasoning about animals draws on factors that span divergent branches, including shared habitat, role as predator versus prey, and size. While these factors cannot be perfectly explained by a single tree, other domains are interestingly structured and are even further removed from a clean form, such as artifacts and social networks. Since humans learn and reason about all of these domains, they must entertain structural hypotheses without obvious grammatical descriptions.

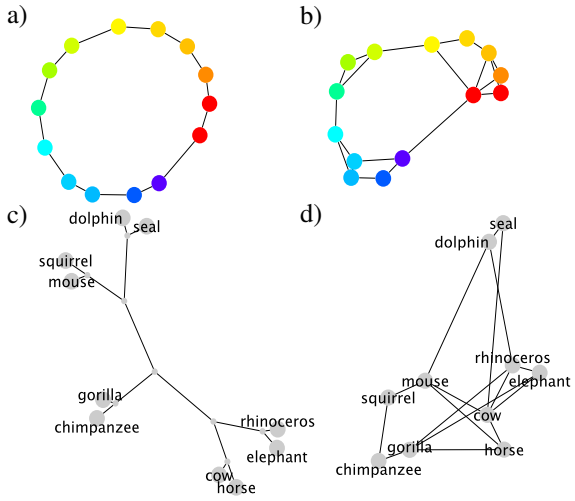


Figure 1: Structure learned by the structural forms model for colors (a) and mammals (c), compared to the sparse model (b, d). Shorter edges correspond to stronger connections. Graphs in this paper, except cities, were drawn with Cytoscape.

These considerations have motivated models without an explicit representation of structure. Rogers and McClelland (2004) demonstrated how structure can emerge in a connectionist network mapping animals (like canary) and relations (can) to output attributes that a canary can do (grow, move, fly, and sing). Without being constrained to follow a tree, their network learns a distributed representation that approximates a tree. But Kemp and Tenenbaum (2009) suggest some advantages of explicit representation: for incorporating observations that have direct structural implications ("Indiana is next to Illinois") and for learning higher-level knowledge (a tree helps learn the word "primate", Fig 1c). It also remains to be seen if this model can predict human inductive inferences about animal properties, as past researchers have found this difficult (Kemp, Perfors, & Tenenbaum, 2004).

Here, we present an approach to structure discovery that incorporates some of the best features of previous probabilistic and connectionist models. Rather than selecting between discrete structural hypotheses defined by grammars, the model

[1] Even this structure has exceptions; for example, Rivera and Lake (2004) provide evidence that at the deepest levels "the tree of life is actually a ring of life" where genomes fused.

learns structure in an unrestricted space of all possible graphs. In order to achieve good inductive generalization, there must be a method for promoting simple graphs. While Kemp and Tenenbaum (2008) used grammars, here we use sparsity, meaning only a small number of edges are active. This structural freedom can approximate cleaner structural forms, such as the ring-like graph for colors in Fig. 1b learned from similarity data printed in Shepard (1980), and on other datasets it deviates, such as mammals (Fig. 1d). Often these deviations capture additional information; while the tree suggests squirrels and mice are equidistant from chimps, the sparse structure suggests squirrels and chimps share additional similarity, like their association with trees.

The sparse model achieves performance equivalent to more structured approaches when predicting human inductive judgements. We show this for biological properties of animals and geographical properties of cities (Kemp & Tenenbaum, 2009). Due to the model's computational efficiency, it can learn on datasets too large for most previous approaches. We demonstrate learning a structure for 1000 concepts with broad semantic coverage, resembling classical proposals for semantic networks (Collins & Loftus, 1975).

The Sparse Model

In the structural forms and sparse models, a structure defines how objects covary with regard to their features. Objects are nodes in a weighted graph, where the strength of connectivity between two objects is related to the strength of covariation with regard to their features. The weights of the graph, denoted as the symmetric matrix W, are learned from data by optimizing an objective function that trades off the fit to the data with the sparsity of the graph.

The data D is an n x m matrix with n objects and m features. The columns of D, denoted as features $\{f^{(1)},...,f^{(m)}\}$, are assumed to be independent and identically distributed draws from $p(f^{(k)}|W)$. If the graph structure fits the data well, features should vary smoothly across the graph. For example, if two objects i and j are connected by a large weight w_{ij} (like seal and dolphin), they often share similar property values ("is active" or "lives in water"). As a result of sparsity, most objects are not directly connected in the learned graph ($w_{ij}=0$, like dolphin and chimp), meaning they are conditionally independent when all the other objects are observed.

Formally, the undirected graph W defines a Gaussian distribution $p(f^{(k)}|W)$, known as a Gaussian Markov Random Field (GMRF), where the n objects are the n-dimensions of the Gaussian. Learning GMRFs with sparse connectivity has a long history (Dempster, 1972), and recent work has formulated this as a convex optimization problem that can be solved very efficiently, in $O(n^3)$, for the globally optimal structure (e.g., Duchi, Gould, & Koller, 2008). Following Kemp and Tenenbaum (2008), we assume people learn a single set of parameters that fits the observed data well. Thus, we find the maximum *a posteriori* (MAP) estimate of the parameters $\underset{W}{\operatorname{argmax}} \log p(W|D) = \underset{W}{\operatorname{argmax}} \log p(W) + \sum_{i=1}^{m} \log p(f^{(i)}|W)$.

Generative model of features. Following the formulation in Zhu, Lafferty, and Ghahramani (2003), a particular property vector $f^{(k)}$, observed for all n objects $f^{(k)} = (f_1^{(k)},...,f_n^{(k)})$, is modeled as

$$p(f^{(k)}|W) \propto \exp(-\frac{1}{4}\sum_{i,j} w_{ij}(f_i^{(k)} - f_j^{(k)})^2 - \frac{1}{2\sigma^2} f^{(k)T} f^{(k)}).$$

This defines a notion of feature smoothness, and it is equivalent to the n-dimensional Gaussian distribution

$$p(f^{(k)}|W) \sim N(0, \tilde{\Delta}^{-1}),$$

where $\tilde{\Delta} = Q - W + I/\sigma^2$ is the precision (inverse covariance) matrix, $Q = \operatorname{diag}(q_i)$ is a diagonal matrix with entries $q_i = \sum_j w_{ij}$, and I is the identity matrix. We also restrict $w_{ij} \geq 0$, so the model represents only positive correlations. The model assumes the feature mean is zero, and raw data is scaled such that the mean value in D is zero and the maximum value in covariance $\frac{1}{m}DD^T$ is one. The parameter σ^2 can be thought of as the *a priori* feature variance (Zhu et al., 2003), and we choose the value that maximizes the objective function.

Sparsity penalty. To complete the model, we need a prior distribution on graph structures, $p(W)$. To learn a simple graph representation with a minimal number of edges, we assume each weight $p(w_{ij})$ is independently drawn from a distribution $p(w_{ij}) \sim \text{Exponential}(\beta)$, meaning

$$p(W) = \prod_{1 \leq i < j \leq n} \beta e^{-\beta w_{ij}}.$$

This prior encourages small weights, and in practice it produces sparse graph structures by forcing most weights to zero.

Structure Learning. Finding $\underset{W}{\operatorname{argmax}} \log p(W|D)$ is equivalent to the following convex optimization problem:

$$\begin{aligned}
&\underset{\tilde{\Delta} \succ 0, W, \sigma^2}{\text{maximize}} \log|\tilde{\Delta}| - \operatorname{trace}(\tilde{\Delta}\frac{1}{m}DD^T) - \frac{\beta}{m}||W||_1 \\
&\text{subject to} \\
&\qquad \tilde{\Delta} = \operatorname{diag}(\sum_j w_{ij}) - W + I/\sigma^2 \\
&\qquad w_{ii} = 0,\ i = 1,\ldots,n \\
&\qquad w_{ij} \geq 0,\ i = 1,\ldots,n; j = 1,\ldots,n \\
&\qquad \sigma^2 > 0.
\end{aligned}$$

The first term in the objective, $\log|\tilde{\Delta}| - \operatorname{trace}(\tilde{\Delta}\frac{1}{m}DD^T)$, is proportional to the log-likelihood from Kemp and Tenenbaum (2008) after dropping unnecessary constants, and $\frac{\beta}{m}||W||_1$, where $||W||_1 = \sum_{i=1,j=1}^{n} |w_{ij}|$, comes from the log-prior. $\tilde{\Delta} \succ 0$ denotes a symmetric positive definite matrix. The only free parameter, β, controls the tradeoff between the log-likelihood of the data and the sparsity penalty ($||W||_1$). A larger β encourages sparser graphs. As more features are observed (m increases), the likelihood is further emphasized in the tradeoff. For all simulations, we set $\beta = 14$. The solution was found using `CVX`, a package for solving convex programs (Grant & Boyd, n.d.).

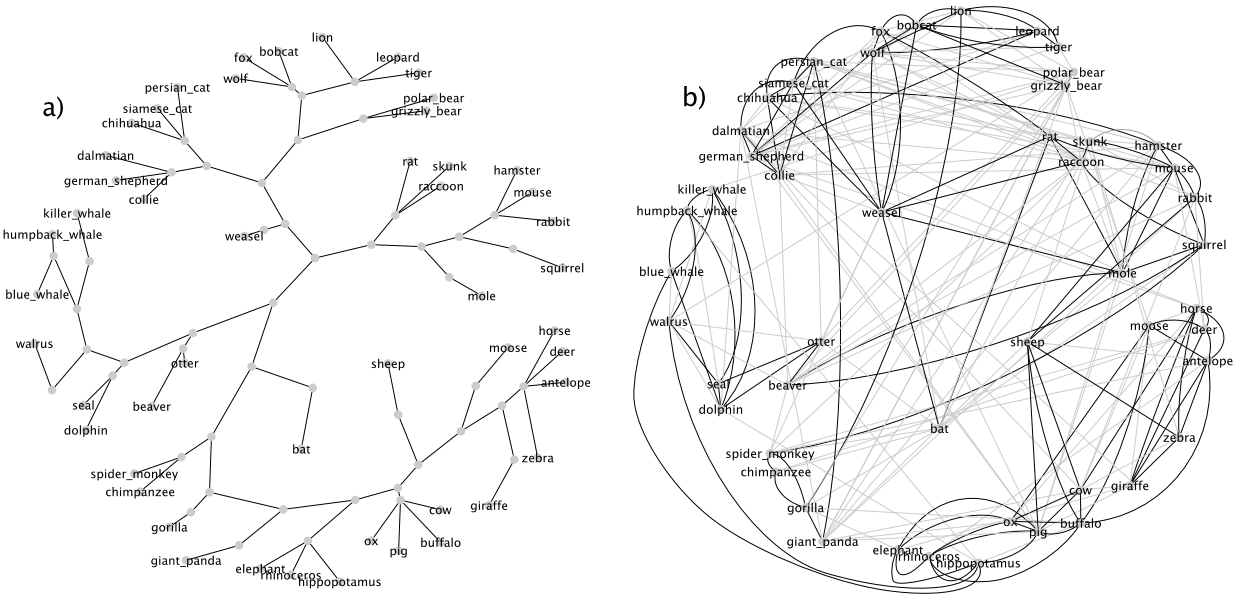


Figure 2: The (a) tree and (b) sparse graphs learned for mammals. Shorter edges in the tree correspond to stronger weights. The sparse graph is overlaid by node position, and thus edge length does not indicate strength. Strong edges $w > .2$ are in bold.

Taxonomic reasoning

Kemp and Tenenbaum (2009) learned a tree structure from a dataset of 50 mammals and 85 biological properties collected by Osherson, Stern, Wilkie, Stob, and Smith (1991). Properties were various kinds of biological and anatomical features, including "is smart," "is active," and "lives in water," and participants rated the strength of the association between each mammal and feature. The learned tree achieves high correlations when predicting human inductive judgments about novel biological properties. This predictive success may be due to the origin of these properties in the natural world; biological relatedness is determined by an evolutionary process where species split and branch off. But there are reasons to suspect humans learn more complicated cognitive structures due to shared similarity across divergent branches, as discussed in the introduction. Rather than constraining structure to be a tree, perhaps optimization with a sparsity constraint can learn appropriate structure for taxonomic reasoning.

Learning structure. Fig. 2 compares a tree learned by the structural forms model and a graph learned by the sparse model for the mammals dataset.[2] The sparse model has 19% of possible edges active ($w > .01$), and stronger edges are highlighted in the figure. While the sparse model does not learn a tree, it captures some important aspects of the tree-based model. Major branches of the tree correspond to densely connected regions of the sparse model. The sparse graph captures some additional detail not represented by the tree. For instance, hippo is connected to the blue whale and walrus; although distant in the taxonomy, they are large and live in/around water. Similarly spider monkey and squirrel have a new link, perhaps due to agility and living in trees.

Property induction. A learned structure defines a prior distribution on properties of animals, which can be used for induction about new properties. Learning often involves generalizing new properties to familiar animals; when a child first hears about the property "eats plankton," the child makes decisions about which mammals this property extends to. To test the sparse and tree model, we apply them to two classic datasets of human inductive judgments collected by Osherson, Smith, Wilkie, Lopez, and Shafir (1990), which were also used in Kemp and Tenenbaum (2009). Judgments concerned 10 species: horse, cow, chimp, gorilla, mouse, squirrel, dolphin, seal, and rhino (Fig 1). Participants were shown arguments of the form "Cows and chimps require biotin for hemoglobin synthesis. Therefore, horses require biotin for hemoglobin synthesis." The Osherson horse set contains 36 two-premise arguments with the conclusion "horse," and the mammals set contains 45 three-premise arguments with the conclusion "all mammals." Participants ranked each set of arguments in increasing strength by sorting cards.

We compare the inductive strength of each argument for both the models and the participants (averaged rank across participants). Following Kemp and Tenenbaum (2009), to compute inductive strength in the models, we calculated the posterior probability that all categories in a set Y have the novel feature (in the above example $Y = \{horses\}$)

$$p(f_Y = 1 | l_X) = \frac{\sum_{f: f_Y = 1, f_X = l_X} p(f)}{\sum_{f: f_X = l_X} p(f)}. \quad (1)$$

[2]The antelope had four missing color features which were filled in from giraffe. They were left missing in the structural forms work. When learning any model from Kemp and Tenenbaum (2009), the best fitting σ^2 variance parameter was found, as in the sparse model. In the original work this parameter was fixed at $\sigma = 5$.

A binary label vector l_x is a partial specification of a full binary feature vector f that we want to infer. In the above example, $X = \{cows, chimps\}$ and $l_X = [1,1]$ indicating both cows and chimps have biotin. Intuitively, Equation 1 states that the posterior probability $p(f_Y = 1|l_X)$ is equal to the proportion of possible feature vectors consistent with l_X that also set $f_Y = 1$, where each feature vector is weighted by a prior probability $p(f)$ defined by the structure. We compute $p(f)$ by drawing 10^6 feature samples from the Gaussian defined by that structure, converted to binary by thresholding at zero.

Performance of the sparse model is shown in column 1 of Fig. 3. The sparse model and tree-based model (column 2) perform equivalently and predict the participant data well. Both models outperform a spatial model (column 3, see Eq. 2) which embeds the animals in a 2D space, with particular advantage on the mammals dataset. The sparse, tree, and spatial models can be viewed as "cleaning up" the raw covariance matrix $\frac{1}{85}DD^T$, approximating it as closely as possible while satisfying certain constraints (sparsity, tree grammar, or 2D embedding). When compared to the raw covariance (column 4), the sparse and tree model show better performance.

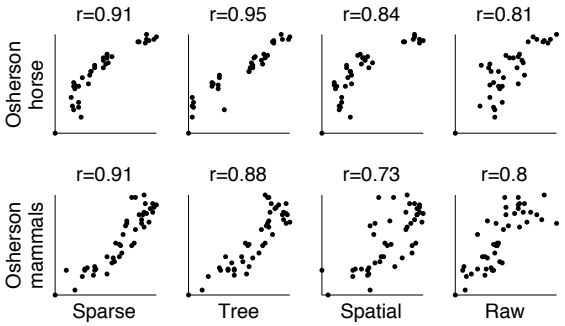


Figure 3: Model performance on taxonomic reasoning. Human ratings of argument strength (y-axis) are plotted against the model ratings (x-axis) for each argument.

Learning about new objects. In addition to learning about new properties, people constantly encounter new objects. How do the models learn about a new mammal, observed for just a few features? The tree-based model provides strong grammatical guidance, but it might be difficult to make discrete placement decisions with only a few observed features. By contrast, the sparse model has no grammatical guidance, so this provides an interesting comparison. Adding a new concept to the sparse model involves solving two convex programs. First, the model was trained on all but one mammal (49) and all properties (85). Second, the learned connections and variance were frozen, and the new concept was added while observing only a few features (10 or 20).[3] Performance was evaluated on predictive ability for the missing properties (75 or 65). The models were tested by adding four different mammals, where each addition was replicated 30 times with different random sets of observed properties. For each missing property, its expected value was calculated by performing inference in the Gaussian defined by the structure. Compared to the raw covariance matrix, the sparse model provided significantly better predictions of the missing features for each mammal tested (all 8 comparisons $t(29), p < .01$, Fig. 4). Since running all combinations is slow in the tree model, each model was also compared on an "informative feature set" (*'s in Fig. 4), defined as the feature set the raw covariance performed best on. For learning a new object with these features, the sparse model performs at least as well as the tree model.

[3]Since many data entries are missing, simply skipping missing entries results in a covariance matrix that is not positive semi-definite. Instead we use a maximum likelihood estimate of the covariance matrix found by Expectation-Maximization.

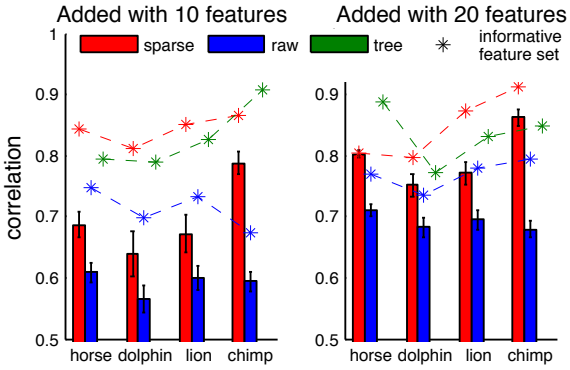


Figure 4: Each model adds a new object (seeing only 10 or 20 features), and the missing features are predicted. Bars are mean performance over 30 random feature picks, and stars (*) show performance from a single informative feature set.

Spatial reasoning

Geographical knowledge seems to require different structural representations than animals. Following the tradition of using Euclidean spaces to build semantic representations such as multidimensional scaling (Shepard, 1980), Kemp and Tenenbaum (2009) proposed learning a 2D space to represent the relationship between cities. This 2D space defines a Gaussian distribution with zero mean and covariance matrix K

$$K_{ij} = \frac{1}{2\pi}\exp(-\frac{1}{\sigma}||y_i - y_j||_2), \quad (2)$$

where y_i is the location of the city i in 2D space. Kemp and Tenenbaum (2009) found a double dissociation between the tree model and the spatial model, which only perform well on taxonomic and spatial reasoning respectively. Can the sparse model learn structures applicable to both domains?

Learning structure. Structures were learned from participant drawings of nine cities on a piece of paper, and similarity was calculated from the pairwise distances (Kemp & Tenenbaum, 2009). This similarity matrix was treated as the raw covariance input to all the models. The learned spatial representation is compared to the learned sparse graph in Fig. 5. All the models require an assumed number of features, set to $m = 85$, preserving the β/m sparsity ratio from before.

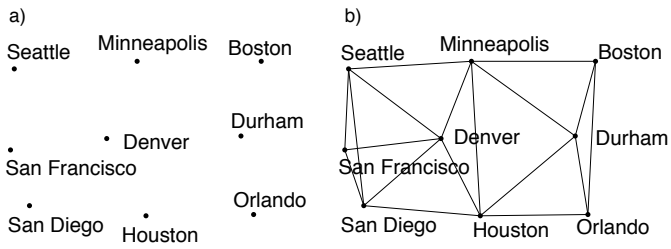


Figure 5: The (a) spatial and (b) sparse models learned from the city dataset. Graphs nodes are overlaid on the 2D space.

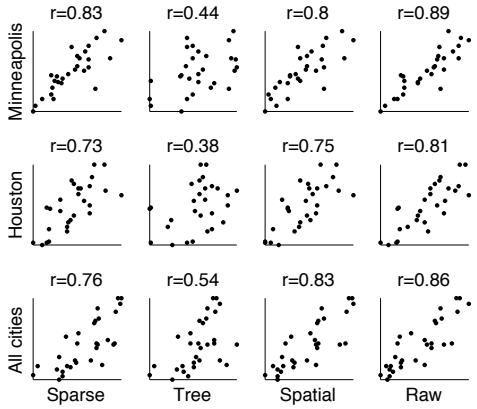


Figure 6: Model performance on spatial reasoning. Human ratings of argument strength (y-axis) are plotted against the model rating (x-axis) for each argument.

Property induction. As in the taxonomic reasoning section, the models were compared to human data regarding property generalization. In an experiment by Kemp and Tenenbaum (2009), participants were presented a scenario where Native American artifacts can be found under most large cities, and some kinds of artifacts are found under just one city while other are under a handful of cities. An example inductive argument is: "Artifacts of type X are found under Seattle and Boston. Therefore, artifacts of type X are found under Minneapolis." There were 28 two-premise arguments with Minneapolis as the conclusion, 28 with Houston as the conclusion, and 30 three-premise arguments with "all large American cities" as the conclusion. These arguments were ranked for strength, and mean rank was correlated with the model inductive predictions. The sparse model (column 1 of Fig. 6) provides good predictions, as does the 2D spatial model and the raw covariance matrix, which performs best (columns 3 and 4). The tree performs poorly (column 2). While there is a double dissociation between the tree and spatial model for taxonomic and spatial reasoning, the sparse model can predict human reasoning in both contexts.

Discovering structure for 1000 concepts

Learning sparse graphs can also be applied to domains with no simple structure. While animals may be fit by trees and

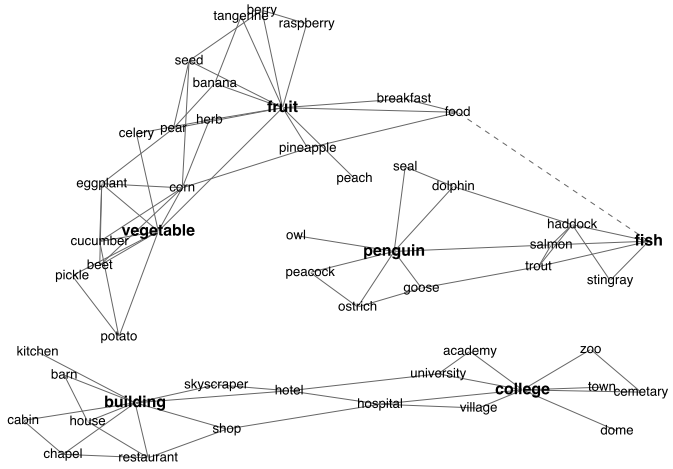


Figure 7: Structure learned for 1000 concepts. This small subset shows the significant neighbors of the bold nodes ($w > .2$ except dotted edge $w = .09$). Shorter edges are stronger.

cities by 2D spaces, what type of structure organizes concepts as diverse as fruit, vegetable, fish, penguin, building, and college? Human semantic reasoning operates in a huge semantic space, and here we learned a sparse model on an expansive domain of 1000 entities and 218 properties. A dataset of this size is prohibitive for the structural forms model as well as the connectionist model of Rogers and McClelland (2004).

Dataset and Algorithm. The dataset was collected by Intel Labs (Palatucci, Pomerleau, Hinton, & Mitchell, 2009). Semantic features were questions such as "Is it manmade?" and "Can you hold it?" Answers were on a 5 point scale from definitely no to definitely yes, conducted on Amazon Mechanical Turk. To learn the optimal structure, we use a faster algorithm from Duchi et al. (2008) instead of a generic convex solver. For now, this requires two small changes to the model: w_{ij} can be positive or negative and a separate variance term σ_i^2 is fit to each object instead of one for all objects.

Results. The structure learned from the entire data is very sparse with approximately 2.4% of edges active ($|w| > .01$). Fig. 7 shows snapshots of the network, consisting of nodes that are strong direct neighbors of either fruit, vegetable, fish, penguin, building, and college ($w > .2$). Fruit and vegetable are linked to subordinate examples, and connect to fish via a path through food. Interestingly, the network connects penguin to both sea animals (like fish and seal) and birds, highlighting its role as an aquatic bird. Building and college are connected via several paths, including building–hotel–university–college and building–hotel–hospital–college.

To evaluate the sparse model's predictive capacity for novel questions, we performed 4-fold cross validation, training on 3/4 of the properties and predicting the rest. The average test log-likelihood is $-3.50 \cdot 10^4$ for the sparse model and $-3.84 \cdot 10^6$ for the raw covariance. The raw covariance performs worse than in the past experiments since there are many more objects than features, and performance can be improved

by other regularization techniques such as Tikhonov (computed as $\frac{1}{m}DD^T + vI$ for identity matrix I (Duchi et al., 2008)), which achieves a test log-likelihood of $-3.63 \cdot 10^4$. Tikhonov regularization does not significantly improve the raw covariance on the previous property induction tasks. Even though we fine-tuned the Tikhonov parameter $v = .17$ to the *test* sets, the sparse model still performs better with its parameter $\beta = 14$ fixed across all experiments in this paper.

General Discussion

Here we applied the sparse model to taxonomic and spatial reasoning. Past work has found a double dissociation between these inductive contexts (Kemp & Tenenbaum, 2009), where a tree model and a spatial model provide good fits to only one context. However the sparse model is able to predict human inductive judgments in both contexts, by emphasizing sparsity in structural representation. In addition to these inductive tasks, we applied the sparse model to a dataset of 1000 concepts with broad semantic coverage and no simple structure. The sparse model learned reasonable structure and outperforms simple regularization on novel features.

The sparse model also provides a probabilistic foundation for classic models of semantic memory such as semantic networks (Collins & Loftus, 1975). Semantic networks stipulate that concept nodes are connected to related concepts by varying degrees of strength. These networks resemble the large structure learned for 1000 concepts (Fig. 7), suggesting the sparse model can be used to learn semantic networks from data. The sparse model is also related to Pathfinder networks (Schvaneveldt, Durso, & Dearholt, 1989) that find the minimal graph that maintains all pairwise sum-over-path distances between objects. While highlighting important structure, it retains the same similarity matrix from input to output, lacking the regularization that is important in our simulations.

While the sparse model is an important first step, it leaves out desirable features of previous connectionist and probabilistic models. The Rogers and McClelland (2004) model accounts for a rich array of phenomena from development and semantic dementia, yet to be explored with the sparse approach. Compared to structural forms, the sparse model does not learn latent nodes (compare Fig. 1c,d), which increase sparsity and could be important for learning higher-level concepts such as "mammal" or "primate" (Kemp & Tenenbaum, 2009). Future work will use the sparse approach to explore learning deeper conceptual structure with latent variables.

Acknowledgements

We thank Intel Labs for providing their 1000 objects dataset, Charles Kemp for providing code and datasets for learning structural forms, and Venkat Chandrasekaran, Ruslan Salakhutdinov, and Frank Jäkel for helpful discussions.

References

Collins, A. M., & Loftus, E. F. (1975). A spreading activation theory of semantic processing. *Psychological Review*, *82*, 407-428.

Dempster, A. P. (1972). Covariance selection. *Biometrics*, *28*, 157-175.

Duchi, J., Gould, S., & Koller, D. (2008). Projected subgradient methods for learning sparse gaussians. In *Proceedings of the twenty-fourth conference on uncertainty in AI (UAI).*

Grant, M., & Boyd, S. (n.d.). *CVX: Matlab software for disciplined convex programming.* Retrieved 2009, from `http://stanford.edu/~boyd/cvx`

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2004). Learning domain structures. In *Proceedings of the twenty-sixth annual conference of the cognitive science society.*

Kemp, C., & Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences*, *105*(31), 10687-10692.

Kemp, C., & Tenenbaum, J. B. (2009). Structured statistical models of inductive reasoning. *Psychological Review*, *116*(1), 20-58.

Osherson, D. N., Smith, E. E., Wilkie, O., Lopez, A., & Shafir, E. (1990). Category-based induction. *Psychological Review*, *97*(2), 185-200.

Osherson, D. N., Stern, J., Wilkie, O., Stob, M., & Smith, E. E. (1991). Default probability. *Cognitive Science*, *15*, 251-269.

Palatucci, M., Pomerleau, D., Hinton, G., & Mitchell, T. (2009). Zero-shot learning with semantic output codes. In Y. Bengio, D. Schuurmans, & J. Lafferty (Eds.), *Advances in neural information processing systems (NIPS).*

Rivera, M., & Lake, J. (2004). The ring of life provides evidence for a genome fusion origin of eukaryotes. *Nature*, *431*, 152-155.

Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach*. Cambridge, MA: MIT Press.

Schvaneveldt, R. W., Durso, F. T., & Dearholt, D. W. (1989). Network structures in proximity data. In G. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 24). Academic Press.

Shepard, R. N. (1980). Multidimensional scaling, tree-fitting, and clustering. *Science*, *210*, 390-398.

Zhu, X., Lafferty, J., & Ghahramani, Z. (2003). *Semi-supervised learning: From gaussian fields to gaussian processes* (Tech. Rep. No. CMU-CS-03-175). Carnegie Mellon University.

Using the Social of Tagging: The Interplay of Social Tags and the Strength of Association in Navigation and Learning Processes

Christoph Held (c.held@iwm-kmrc.de)
Knowledge Media Research Center, Konrad-Adenauer-Strasse 40
72072 Tuebingen, Germany

Ulrike Cress (u.cress@iwm-kmrc.de)
Knowledge Media Research Center, Konrad-Adenauer-Strasse 40
72072 Tuebingen, Germany

Abstract

When people navigate through the World Wide Web they choose their path of navigation based on their prior knowledge. This may be problematic when users have a deficient knowledge leading them to suboptimal information. In this study we examined how the externalized knowledge of social tags can be used to change navigation behavior and to trigger learning processes. In an online experiment with 531 participants we investigated the effect of the individual strength of association on navigation processes, and how the collective strength of association, visualized in tag clouds, may affect individual navigation and the strength of association. Results showed the effect of individual strength of association on navigation behavior, selection time and recognition. Furthermore, we found that the collective strength of association affects navigation behavior and triggered incidental learning processes, leading to a change of individual strength of association.

Keywords: social tagging; tag clouds; social software; information foraging; web search; incidental learning

Introduction

People frequently use the World Wide Web for information and product search. In some topic domains, Web users may only possess deficient prior knowledge and an incomplete view of relevant aspects. A user's knowledge may, however, be critical for the search process and the information which is retrieved from the Web. The Web offers enormous quantities of heterogeneous information and products, and each user will have to select between different links and keywords for finding relevant resources. When users follow navigation links based on their deficient prior knowledge these may lead to information which will confirm or even reinforce the deficient knowledge of that user. For example, users might associate the treatment of a disorder with some specific medication. Instead of considering other treatments or medications a user may quickly select a navigation path leading to information which reinforces potentially deficient knowledge saying, for instance, that a specific medication is the only reasonable treatment. This might happen when navigating to a website from a pharmaceutical company.

So, on the one hand, the mass and the diversity of resources available on the Web is combined with the risk that people might select suboptimal information or products. On the other hand, new tools may provide the opportunity to use the mass of available information on the Web to improve individual navigation and to adjust and change the users' previously deficient prior knowledge. In this paper we address the research question how social tags, as emerging collective information, can affect the individual process of navigation and how social tags trigger learning processes during navigation. In particular, we focus on situations in which the externalized knowledge of social tags contradicts the prior knowledge of users.

The next chapter will provide a theoretical overview on Web navigation and its interrelation with spreading activation, followed by an overview on social tagging and how it may interact with cognitive processes. As a next step we will present an experimental study on the effects of social tags and the strength of association on navigation and knowledge acquisition.

Theoretical Background

Information Foraging Theory

A pivotal cognitive theory of Web navigation is the Information Foraging Theory (Pirolli, 2007; Pirolli & Card, 1999). It explains selection processes of links and navigation paths on the Web, the so-called "information foraging". Taking for granted that many search tasks on the Web require browsing activities in order to find a desired resource (Marchionini, 2006), users will have to select between different links and navigation paths. They will have to decide which link may lead to a desired – and not directly accessible – distal resource, say, a piece of information on some Web site. When navigating the Web users have to make judgments based on proximal cues (e.g., links) and assess which of these cues have the highest likelihood of leading to a desired distal resource. One of the core concepts of the Information Foraging Theory is the so-called "information scent" of links. The information scent describes the subjective usefulness of links for navigation. Links with a subjectively high probability of leading to a desired distal resource have a high information scent and are very likely to be selected in the search process. How will users estimate the information scent of links?

Spreading Activation Understanding how people evaluate the information scent of a link is closely related to models of semantic memory and spreading activation (e.g., Anderson,

1983; Collins & Loftus, 1975). Cognitive models of semantic memory assume that memory is based on a collection of cognitive structures, so-called "chunks". These are organized as nodes in a large network in memory. Each of the chunks is connected to other chunks with a different strength of association. The strength of association derives from the respective individual's previous learning experiences. When two chunks frequently co-occur in a meaningful context, the association between these chunks becomes stronger. For example, when Valium is often mentioned in the context of anxiety disorders, a high strength of association will be established. The strength of association is important in the process of retrieving chunks from memory. To retrieve a chunk from memory, it must be activated by other chunks. The activation spreads from one chunk to another, and the stronger the association, the higher is the likelihood of exceeding a certain level of activation for a chunk. For instance, the activation of Valium in the context of anxiety disorders is facilitated by a high strength of association.

In a search process a desired distal goal activates connected chunks in semantic memory. Based on the strength of association of connected chunks and the resulting strength of activation, users estimate the information scent of links: when a chunk receives a high spreading activation through a search goal, the corresponding link receives a high information scent, too. For example, when the chunk Valium is highly activated by a search goal, e.g., treatment for anxiety disorder, then the corresponding Web link Valium would also have a high information scent for a user.

Research on Navigation Processes Several studies have demonstrated the effect of the information scent on Web navigation (e.g., Blackmon, Polson, Kitajima & Lewis, 2002; Fu & Pirolli, 2007; Pirolli, Fu, Reeder & Card, 2002). These studies have mainly used cognitive modeling of Web navigation and validated them against actual user data.

In these studies, differences in prior knowledge were not considered for the modeling process. In some studies, for instance, the simulation of strengths of association and the resulting information scents were based on the same large text corpora and the co-occurrence of words within these texts (Fu & Pirolli, 2007; Pirolli et al., 2002). So the focus of these studies did not lie in investigating the effects of differing prior knowledge, but rather in modelling the general search process for specific search tasks. Another aspect which has not been investigated in more detail within the (non-social) context of these studies is learning processes during navigation, i.e. incidental learning as a by-product of navigation. When navigating the Web, users process information in order to assess the information scent of links. But the choice of the navigation path is not only a means to an end. It may also be of importance what happens "along the path". Except for one study showing incidental category learning during navigation (Pirolli, 2004) it remains unclear how navigation itself could change the strength of association of chunks through incidental learning processes.

The question of learning is particularly interesting in a social Web context, in which large numbers of other users contribute information and, in particular, in which this information can be used for navigation processes. New Web technologies, like social software tools, provide this opportunity. Social tagging systems make it possible to learn from the externalized knowledge of a community. In the next section we will give an overview on social tagging and relevant studies that have been conducted in this field of research.

Social Tags

Social tagging is the activity of annotating digital resources, e.g. bookmarks, pictures or products, with keywords, the so-called "tags". Tags represent metadata on resources. For most applications each user can choose individual tags for stored resources. Tags reflect individual associations with resources and are based on the specific meaning or relevance to that user. At the individual level, tags will help users to structure, organize and find their own stored Web resources. In a social context, tags offer the opportunity to use other users' navigation links for search processes. Moreover, social tagging systems can aggregate the tags of individual users. In this way, resources are described by the community in a "folksonomy", developed in a bottom-up process of individual tagging. The aggregated tags represent an emerging collective knowledge of Web users. These aggregated tags can also be used as links for individual search processes. In a social tagging system, the community creates a network of connections between resources and tags. The connection between a resource and a tag becomes stronger when tags for that resource are used more frequently by many users. The connection between two tags becomes stronger when both tags are used together for one resource: the more often two tags co-occur with the same resources, the stronger they are related to each other. When aggregating all tags from a community, a representation of the connections between related tags and their strength of association will emerge. Typically, tag clouds visualize these associations and their specific strength: The font size of tags illustrates the strength of association of tags to a related tag or a resource (see Figure 1).



Figure 1: A tag cloud representing tags related to "red wine" (from vinorati.com). The font size visualizes the strength of association between "red wine" and the tag.

Social tagging systems may be regarded as shared external knowledge structures of communities (Fu, 2008). They can externalize the connections of tags and their specific strength of association in tag clouds. Because these associations are based on the collective tagging behavior of a community, they can also be considered to be the externalized associations of a particular community. The structure of social tagging systems even constitutes an analogy to spreading activation processes in semantic memory models, in which tags represent the nodes of a large network. When a tag is selected – or activated –, the activation spreads from this tag to others, and the related tags and their strengths of association can be visualized in tag clouds.

So far, research on social tagging has mainly focused on the description of regularities of tagging systems (e.g., Golder & Huberman, 2006) or the use of tagging systems (e.g., Millen, Yang, Whittaker & Feinberg, 2007). But, as stated by Fu (2008), surprisingly little is known about how these new technologies, like social tagging systems, may directly interact with individuals at the knowledge and cognitive level. Some studies have investigated the influence of tag clouds on visual attention, recognition and tag selection (Bateman, Gutwin & Nacenta, 2008; Rivadeneira, Gruen, Muller & Millen, 2007). But these studies focused primarily on the visual features of tag clouds and did not address aspects of collective knowledge, as it is externalized in social tagging systems and tag clouds.

A study investigating the interplay between the collective knowledge of a tagging system and the individual cognitive level was presented by Fu in 2008. He presented a rational model of social tagging and provided evidence for the interaction of social and cognitive systems. That study showed the impact of externalized knowledge structures on individual learning processes, especially the formation of mental categories, but did not focus on the effects of the representation of knowledge, externalized in the form of tag clouds, on individual navigation behavior and the strengths of associations. A further study, which also investigated the knowledge exchange within tagging systems dealt with the question of how social tags affect tag choice (Kang, Kannampallil, He & Fu, 2009). This study also showed that the externalized knowledge of social tags will influence individual behavior: users adapted their tag choice to the collective structure of the social tagging system.

Research Questions and Hypotheses

Models of Information Foraging (Pirolli, 2007) assume that processes of spreading activation are crucial for the navigation behavior of users. So the strength of association between a search goal and available links plays a critical role in the selection of navigation paths. We assume that users with deficient prior knowledge are likely to choose navigation links that lead to suboptimal resources. So we manipulated the users' prior knowledge and investigated the impact of individual strength of association on navigation processes.

The main goal of this study is to investigate how the collective knowledge of a community affects individual learning and navigation processes. Can social tags be used to change the navigation behavior of users? Will users learn from the collective knowledge during navigation, and will they improve the deficits of their prior knowledge accordingly? Apart from the variation of the users' prior knowledge we also manipulated the strength of association of tags. In our experiment we created a situation in which the individual strength of association contradicts the collective strength of association. We examined the effect of collective strength of association on the change of individual navigation and strength of association.

We expected (1) a main effect for both the individual strength of association and the collective strength of association on navigation behavior. Secondly, we expected (2) an incidental learning process and a change of individual strength of association during navigation through the collective strength of association. Thirdly, we expected that (3) in a situation of highly contradicting individual and collective strengths of association, users will perceive a conflict and process all tags more thoroughly and spend more time on their selection of tags.

Experiment

Method

Participants 531 participants (179 female, 352 male; mean age 28.94 years, *SD* = 9.36) were recruited on Amazon Mechanical Turk (mturk.com), an Internet marketplace for engaging users in online micro-tasks. The participants were paid US-$1.20 for the experiment. The participants came from 52 different countries. Most of the subjects lived in the United States (41.1%) and India (36.7%).

Materials and Procedure In order to ensure that we could actually manipulate the prior knowledge of subjects, we selected a topic which was very likely to be mainly unfamiliar to the subjects: wine from the Asian country of Georgia, in particular from various wine regions of Georgia. The experiment was set up online and all participants could perform the task from a computer with Internet access. On average, the experiment took about 8 minutes for each user. We instructed the subjects that our aim was to receive feedback on the design of Web sites dedicated to wine. The actual goal of the task in the experimental context was not transparent to the subjects. We did not inform them before or during the task that we were actually measuring navigation and learning processes.

The task consisted of two parts. In the first part subjects had to provide feedback on design features of a wine list from "a pilot user who is a wine lover of Georgian wines". The list was presented to the subjects for 30 seconds, followed by five questions on the design of this list in order to direct attention to it. The first independent variable - the individual strength of association - was manipulated by this wine list (see Figure 2).

The second part was a navigation task. In this part, subjects were asked to use tag clouds as navigation links. The subjects were told that the tags originate "from different sources of the Internet, like online wine communities and wine retailers". After a basic introduction to social tags, subjects were presented tag clouds and were asked to click on one tag of each tag cloud that was most appropriate to direct them to a typical Georgian wine. Overall, we presented four tag clouds. The first and the third tag cloud were used as case examples. Only the second and fourth tag cloud were relevant to the experiment. In each condition these two tag clouds were identical. These tag clouds represented related tags (wine regions) to Georgia (see Figure 3). After having clicked on a tag, it was color-marked and two seconds later the next tag cloud appeared. The next tag cloud was independent of the previous selection. Only tag clouds were presented, no corresponding resources were displayed. After the navigation task we presented tests measuring the dependent variables "decision" and recognition.

Independent Variables and Design A 5 x 4 between-subjects design was used. Subjects were randomly assigned to one of the 20 conditions. As a first independent variable we experimentally manipulated the individual strength of association by varying the content of the wine list, which was presented to the subjects in the first part of the task. We manipulated how strongly users associate the wine region "Kakheti" with Georgian wine. Users were presented a wine list with five Georgian wines. In the different conditions the number of wines coming from the region "Kakheti" was varied. The independent variable had five continuous levels: wines from the region "Kakheti" were either (1) not part of the list (see Figure 2a); or (2) one time; (3) two times; (4) three times; or (5) four times in the list (see Figure 2b).



a) b)

Figure 2: Wine lists manipulating the individual strength of association for the "Kakheti" region, representing the lowest and highest levels: a) "Kakheti" is not part of the list b) 4 of the 5 wines come from "Kakheti".

As a second independent variable we experimentally manipulated the collective strength of association by varying the tag size in the tag clouds. Except for the tag "Kakheti" none of the regions presented in the wine list reappeared in the tag clouds. We manipulated how strongly the fictitious tagging community associates the wine region "Imereti" with Georgian wine by varying the tag size of "Imereti". The other tags did not vary in size. The independent variable had four continuous levels: (1) the tag "Imereti" had the same size as the tag "Kakheti" with both tags representing the biggest tags in the tag cloud (see Figure 3a); (2) the tag "Imereti was 33% bigger than in the first condition; (3) the tag "Imereti" was 67% bigger than in the first condition; (4) the tag "Imereti" was 100% bigger than in the first condition (see Figure 3b).

a)



b)



Figure 3: Tag clouds manipulating the collective strength of association for "Imereti", representing the lowest and highest levels: a) "Imereti" has the same size as "Kakheti" b) "Imereti" is twice as big as "Kakheti".

Dependent Measures As dependent variables we measured the navigation behavior of users for the two relevant tag clouds (wine regions) by analyzing the logfiles. It was assessed how often users clicked the tag "Kakheti" (for which the individual strength of association was manipulated in the wine list) or the tag "Imereti" (for which the collective strength of association was manipulated in the tag cloud). Accordingly, the number of clicks for navigating the two tag clouds could range between 0 and 2 for either of the dependent variables "Navigation Kakheti" and "Navigation Imereti". We also measured how much time users spent for the selection process. We added the time which was used for navigating each of the two tag clouds.

For the assessment of the dependent variable "decision", users were asked which Georgian wine region they would select if they had to buy a typical wine from Georgia. They had to choose between the alternatives "Kakheti" and "Imereti". Referring to the fluency heuristic (e.g., Schooler & Hertwig, 2005) it is assumed that if one of the two alternatives has a higher strength of association and is more fluently processed, then users will infer that this alternative has a higher value regarding to the criterion – in this case,

the decision which wine region is more typical of Georgia. We assume that the decision in favor of one of the alternatives will be based on the higher individual strength of association for that alternative. This dependent variable was coded (-1) for the selection of "Kakheti" and (1) for the selection of "Imereti".

Another dependent variable was the recognition of tags. This measure was assessed in a multiple choice test consisting both of tags that were presented in the tag clouds (seven items) and tags which were not contained in the tag clouds (nine items). The task of the subjects was to correctly identify those tags which were presented in the tag clouds. The score was calculated as the sum of correctly identified items minus incorrectly marked items. Tags which were part of the manipulation ("Kakheti" and "Imereti") were not considered for the recognition score.

Results

To test the impact of individual and collective strength of association on the dependent variables, multiple regression analyses were conducted with the predictors individual strength of association, collective strength of association and the individual x collective strength of association interaction, and the dependent variables as criteria. The predictor variables were centered, and the interaction term was computed by a multiplication of both variables.

It was predicted that a higher individual strength of association would lead to a higher probability of selecting a tag corresponding to this association, whereas the contradicting collective strength of association is assumed to attenuate this tendency. To test these predictions, a regression with the criterion "Navigation Kakheti" was computed. The predictions were confirmed: the individual strength of association for "Kakheti" significantly increased the selection rate of the tag "Kakheti" (β = .34, p < .001), whereas the contradicting collective strength of association significantly decreased it (β = -.12, p < .01), adjusted R^2 = .12, $F(2, 528) = 38.50$, p < .001. No significant interaction was found (β = .05, p = .21).

We also predicted that a higher collective strength of association would lead to a higher probability of selecting the corresponding tag, whereas a contradicting individual strength of association would lead to an opposite effect. To test these predictions, a regression with the criterion "Navigation Imereti" was computed. The predictions were confirmed: the collective strength of association for "Imereti" significantly increased the selection rate of the tag "Imereti" (β = .24, p < .001), whereas the contradicting individual strength of association significantly decreased it (β = -.08, p < .05), adjusted R^2 = .06, $F(2, 528) = 18.29$, p < .001. No significant interaction was found (β = -.03, p = .56).

It was assumed that users would show incidental learning when navigating through tag clouds that represent collective strengths of associations. We predicted that users would change their individual strength of association and adapt to the collective strength of association. The strength of association for either "Kakheti" or "Imereti" was assessed in the dependent variable "decision". On the one hand, we assumed that a higher individual strength of association for "Kakheti" would also lead to a higher probability of choosing "Kakheti". On the other hand, we predicted that the collective strength of association for "Imereti" would increase the individual strength of association for "Imereti", leading to a higher probability of deciding in favor of this contradicting alternative. To test these predictions, a regression with the criterion "decision" was computed. Both predictions were confirmed: The strength of association for "Kakheti" significantly increased the tendency to choose this alternative (β = -.30, p < .001). The collective strength of association for "Imereti" significantly increased the tendency to decide in favor of the contradicting alternative "Imereti" (β = .24, p < .001), adjusted R^2 = .14, $F(2, 528) = 43.38$, p < .001. No significant interaction was found (β = .01, p = .82).

Furthermore, we predicted an interaction between individual and collective strength of association for the dependent variables recognition and selection time: we assumed that for users with a high individual strength of association (e.g., for "Kakheti") a high contradicting collective strength of association (e.g., for "Imereti") would lead to a cognitive conflict, and that this conflict leads to a higher level of processing regarding all presented tags and to a longer duration of tag selection. To test the first of these predictions, a regression with the criterion recognition was computed. The prediction could not be confirmed: no significant interaction was found (β = .00, p = .95). What we did find, however, was that increasing individual strength of association significantly decreased performance in the recognition test (β = -.12, p < .01). The analyses did not reveal a significant effect for the collective strength of association (β = .03, p = .52), adjusted R^2 = .01, $F(2, 528) = 3.96$, p < .05. To test the second of these predictions, a regression with the criterion selection time was computed. This prediction could not be confirmed: no significant interaction was found (β = -.01, p = .83). But the individual strength of association significantly decreased the time used for the selection process: a high individual strength of association led to a faster tag selection (β = -.16, p < .001). The analyses did not reveal a significant effect for the collective strength of association (β = .00, p = .96), adjusted R^2 = .02, $F(2, 519) = 6.51$, p < .01.

Discussion

The aim of this study was to investigate the potential of emerging collective structures of the Web, such as social tags, on individual processes of navigation and learning. We addressed the research question how the collective externalized knowledge of a social tagging community could interact with individual knowledge, and if the navigation process per se – without the explicit intention to learn something – is sufficient for changing individual knowledge representations. In an experiment we investigated how the externalized representation of the

associations of a community in a tag cloud affects the individual strength of association. Through the experimental manipulation we were able to create continuous levels for each of the two independent variables, the individual and collective strengths of associations.

The results showed that both the individual and the collective strength of association affect navigation. In the context of a Web search, these results suggest that, on the one hand, a user's prior knowledge is an important factor when choosing a navigation path. On the other hand, our results suggest that the collective knowledge of other Web users may help to open up better navigation paths, especially if a user's prior knowledge is deficient or biased. The results also show that users learn from the collective strengths of associations and, in the case of contradicting knowledge, that they will change their own individual strength of association by adapting it to the collective one. In this way, users will learn incidentally how a large community evaluates the relevance of certain information or concepts, and they can change their own strengths of associations accordingly.

Furthermore, the results revealed that a high individual strength of association leads to a faster and – as far as the perception of other available links is concerned – to a less thorough selection process. When a user has a strong, but incorrect strength of association, this could lead to a fast selection of a suboptimal navigation path. Especially in this unfavorable case social tags may be helpful: if the collective knowledge of a community was able to provoke a strong cognitive conflict, this could lead to a highly improved navigation process by that user. In this study, however, we could not find any interactions that suggest effects of cognitive conflicts caused by a highly contradicting individual and collective strength of association. A possible explanation could lie in the limitations of the scenario of this experiment, e.g. the rather static and simple navigation process, the small relevance of the task to the users, or the highly unfamiliar topic domain (which had only been selected for the purpose of manipulating the prior knowledge of users). So future research could focus on larger and more dynamic scenarios, combined with a variety of topic domains with a higher degree of relevance to the respective users. In addition, it would be interesting to create a setting which combines both tags as metadata and actual information resources or products.

References

Anderson, J. R. (1983). A spreading activation theory of memory. *Journal of Verbal Learning and Verbal Behaviour, 22,* 261-295.

Bateman, S., Gutwin, C., & Nacenta, M. (2008). Seeing things in the clouds: the effect of visual features on tag cloud selections. *Proceedings of the 19th ACM Conference on Hypertext and Hypermedia, 2008* (pp. 193-202). New York: ACM Press.

Blackmon, M. H., Polson, P. G., Kitajima, M., & Lewis, C. (2002). Cognitive walkthrough for the web. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Changing our world, changing ourselves* (pp. 463-470). New York: ACM Press.

Collins, A., & Loftus, E. (1975). A spreading activation theory of semantic processing. *Psychological Review, 82,* 407-428.

Fu, W. (2008). The microstructures of social tagging: a rational model. In *Proceedings of the ACM 2008 conference on Computer supported cooperative work* (pp. 229-238). San Diego, CA, USA: ACM.

Fu, W., & Pirolli, P. (2007). SNIF-ACT: a cognitive model of user navigation on the world wide web. *Human-Computer Interaction*, *22*(4), 355-412.

Golder, S., & Huberman, B. A. (2006). Usage patterns of collaborative tagging systems. *Journal of Information Science, 32*(2), 198-208.

Kang, R., Kannampallil, T., He, J., & Fu, W. (2009). Conformity out of Diversity: Dynamics of Information Needs and Social Influence of Tags in Exploratory Information Search. In D. Schmorrow, I. Estabrooke, M. Grootjen, (Eds.), *Foundations of Augmented Cognition. Neuroergonomics and Operational Neuroscience.* Berlin/Heidelberg: Springer.

Marchionini, G. (2006). Exploratory search: from finding to understanding. *Communications of the ACM*, *49*(4), 41-46.

Millen, D., Yang, M., Whittaker, S., & Feinberg, J. (2007). Social bookmarking and exploratory search. In L. Bannon, I. Wagner, C. Gutwin, R. Harper, & K. Schmidt (Eds.), *Proceedings of the 10th European Conference on Computer-Supported Cooperative Work*. London: Springer.

Pirolli, P. (2004). InfoCLASS model: Conceptual richness and inter-person conceptual consensus about information collections. *Cognitive Studies: Bulletin of the Japanese Cognitive Science Society*, *11*(3), 197-213.

Pirolli, P. (2007). *Information foraging theory: Adaptive interaction with information*. New York: Oxford University Press.

Pirolli, P., & Card, S. K. (1999). Information foraging. *Psychological Review, 106*(4), 643-675.

Pirolli, P., Fu, W., Reeder, R., & Card, S. K. (2002). A user-tracing architecture for modeling interaction with the world wide web. *Proceedings of the Working Conference on Advanced Visual Interfaces* (pp. 75-83). Trento, Italy: ACM.

Rivadeneira, A. W., Gruen, D. M., Muller, M. J., & Millen, D. R. (2007). Getting our head in the clouds: toward evaluation studies of tagclouds. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2007* (pp. 995-998). New York: ACM Press.

Schooler, L. J., & Hertwig, R. (2005). How Forgetting Aids Heuristic Inference. *Psychological Review*, *112*(3), 610-628.

Interactivity During Spoken Word Recognition: Evidence from Bimodal Bilinguals

Anthony Shook (a-shook@northwestern.edu)
Department of Communication Sciences and Disorders, 2240 Campus Drive
Evanston, IL 60208 USA

Viorica Marian (v-marian@northwestern.edu)
Department of Communication Sciences and Disorders, 2240 Campus Drive
Evanston, IL 60208 USA

Abstract

We explore the role of top-down information in language processing by investigating parallel language activation in bimodal bilinguals, who are fluent users of a spoken and a signed language. In an eye-tracking study, bimodal bilinguals showed activation of their signed language while receiving input in English only. Since spoken and signed languages do not share structure, the results suggest that linguistic information can be readily transmitted across modalities, and that parallel language activation can be driven by top-down processes.

Keywords: bilingualism; ASL; language co-activation; top-down processing; eye-tracking; visual world paradigm

Introduction

The architecture of the language system is determined by the way that information flows among levels of processing. Language processing may involve both bottom-up/feed-forward and top-down/feed-back mechanisms (Rapp & Goldrick, 2000; Navarette & Costa; 2009). However, exclusively feed-forward systems may also be capable of explaining language processing *without* the aid of feed-back mechanisms (Hagoort & Levelt, 2009; Levelt, Roelofs, & Meyers, 1999; McQueen, Jesse, & Norris, 2009; Norris, 1994). Proponents of language systems that recruit top-down mechanisms face the difficulty of separating the impact of top-down information from that of bottom-up information. When both forms of information are present, it is difficult to disentangle the unique contributions each may make to language processing. To understand the role of top-down mechanisms during language processing, the influence of top-down pathways must be measured in isolation. One possible way of limiting the impact of bottom-up information is by investigating language processing in bimodal bilinguals.

Unlike unimodal bilinguals, who use two spoken languages, bimodal bilinguals are fluent in a spoken and a signed language. Research on unimodal bilinguals has revealed non-selective language effects, wherein unimodal bilinguals activate both of their languages in parallel (Blumenfeld & Marian, 2007; Marian & Spivey, 2003; Weber & Cutler, 2004). For example, when a Russian-English bilingual hears the English word "marker," she will also make eye movements to items that are phonologically similar in the non-target language (e.g., Russian) such as "marka" (stamp), suggesting that her Russian is simultaneously activated. This effect appears to be bottom-up in nature – as auditory input enters the language system, it non-selectively activates lexical items in both languages based on structural overlap. Critically, a dual-language bottom-up pathway cannot exist in bimodal bilinguals, as their languages do not utilize the same structural input. The cross-modal nature of bimodal bilingualism therefore allows for the direct investigation of top-down mechanisms in isolation.

If bimodal bilinguals co-activate their two languages in the absence of form overlap, it would suggest that language co-activation can be driven by top-down information, and would require a system capable of activating the non-target language via top-down or lateral connections. Models that consider exclusively bottom-up information for lexical activation or selection (such as the Shortlist Model, Norris, 1994) are unlikely to be able to explain this result, as they limit activation to items that exist in the same modality as the target. Therefore, a bimodal bilingual, when faced with single-modality input (e.g., spoken English), should not activate the signed language.

However, recent research indicates that bimodal bilinguals do co-activate their languages. For example, hearing ASL-English bilinguals produce speech and signs simultaneously (Emmorey, Borinstein, Thompson, & Gollan, 2008), deaf ASL-English bilinguals show interference from sign-language while processing written-English (Villwock, Wilkinson, Bailey, Kroll, Morford, & Piñar, 2009), and late-learning Dutch-Sign Language of the Netherlands bilinguals show interference from English while processing SLN signs (Van Hell, Ormel, van der Loop, & Hermans, 2009). In addition to clarifying the role of top-down mechanisms in language processing, language co-activation in bimodal bilinguals would suggest that linguistic information is readily transmitted across modalities, such that two unrelated languages can be activated simultaneously, even when phonological information from one of the two languages is absent.

The current study used an adapted visual world paradigm to examine parallel language processing in normal-hearing, ASL-English bilinguals. Investigating whether languages that do not share modality are co-activated in bimodal bilinguals can provide insight into the influence of top-down mechanisms on language processing and the architecture of the language system in general, as well as reveal the extent to which linguistic information is modality independent.

Method

Participants

Twenty-six participants were tested (thirteen ASL-English bilinguals, M_{age}=33.2, SD=11.8 and thirteen English monolinguals, M_{age}=23.9, SD=9.8). An additional five participants were not included in the analysis – three due to failure to display sufficient proficiency in ASL, and two due to technical error with the eye-tracker. All participants completed the *Peabody Picture Vocabulary Test* (PPVT-III; Dunn & Dunn, 1997) to assess their English vocabulary skill. No differences were found between bilinguals (M=108.2, SD=9.9) and monolinguals (M=111.8, SD=9.4; $t(24)=0.96$, $p=.35$). Information on the participants' language background was obtained via the *Language Experience and Proficiency Questionnaire* (LEAP-Q; Marian, Blumenfeld, & Kaushanskaya, 2007). On a scale of 1-10, where 10 means "fluent," bilinguals rated their ASL abilities at 8.5 for production, and 8.8 for comprehension, indicating a high degree of ASL proficiency. All participants reported normal hearing and vision.

Materials

Twenty-two minimal sign pairs were developed by choosing two signs that matched on three of four ASL-phonological parameters – handshape, location in space, motion, and orientation of the palm (Brentari, 1998). These sign pairs represented the target and competitor items in our competitor condition. For example, the signs for "cheese" and "paper" overlapped in handshape, location, and palm orientation, but differed in the motion of the sign. Target and competitor signs did not differ significantly in English word frequency [$t(38)=-1.654$, $p=.106$] (obtained from the SubtLexus database; Brysbaert & New, 2009). In addition, twenty-two control items and 110 filler items were chosen based on their lack of phonological overlap to the target in both ASL and English. Control items were used in place of competitor items in the control condition. Control signs also did not differ from target signs in English word frequency [$t(38)=-1.027$, $p=.311$]. In the experiment, each item was represented by a black and white line drawing. In each condition, four black and white drawings were displayed on a computer screen in the corners of a 3x3 grid. The words were recorded at 44.1 Khz, 32 bits by a female, monolingual speaker of English, in sentence context as the final word in the phrase "click on the _____." Recordings were normalized such that the carrier phrase was of equal length for all target sentences, and the onset of the target word always occurred at 600 ms post onset of the sentence. Recordings were amplitude-normalized.

Design

The current study used a 2x2 Mixed design, with group (bilingual, monolingual) as a between-subjects factor, and condition (competitor, control) as a within-subjects factor. The dependent variables include the proportion of looks (1 for a look, 0 for no look) and duration of looks (percent of time per trial spent looking at an item). There were twenty-two competitor trials, containing a target, a competitor item that overlapped with the target in ASL phonology, and two fillers (Fig. 1). Every competitor trial had a corresponding control trial, in which the content and location of the target item and two filler items were identical, but where the phonologically-overlapping competitor item found in the competitor trial was replaced with an unrelated control item. This allowed for the comparison of looks to a specific location in the display as a function of the presence or absence of a phonological competitor. There were also forty-four filler trials, containing a target and three phonologically unrelated items.

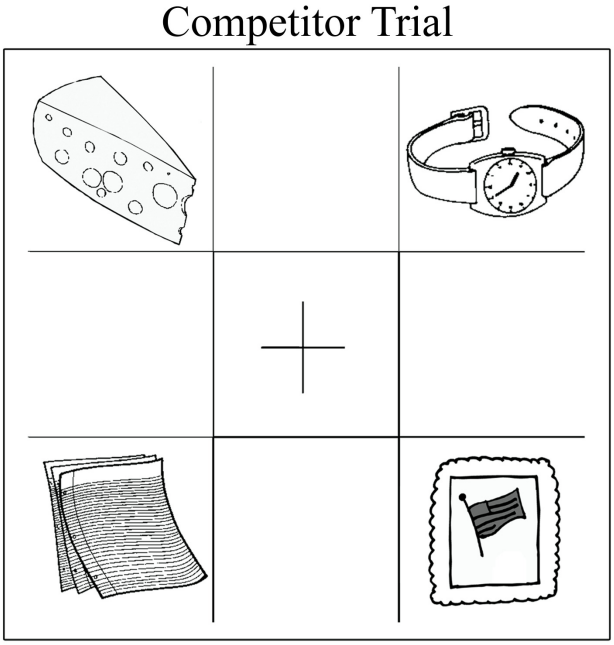


Figure 1: Example of a Competitor Trial. Participants eye-movements were recorded while they were instructed in English to "click on the cheese." At the same time, a phonologically related competitor in ASL ("paper") was present in the display.

Procedure

After informed consent was obtained, participants viewed a video clip displaying the experimental instructions in ASL performed by a native signer of ASL. Following the instructions, participants were fitted with an ISCAN eye-tracker to measure the location of their gaze during the eye-tracking portion of the experiment. Instructions were again provided, in both written and spoken English, followed by five practice trials meant to familiarize participants with the task. Auditory stimuli were presented over headphones and appeared synchronously with picture stimuli. Participants were told that they would hear instructions to choose a specific object in the visual display, and should click on the object that best represents the target word. Participants' eye-

movements were recorded. After the eye-tracking portion of the experiment, all participants completed the PPVT and the LEAP-Q. In addition, bilingual participants were presented with a list of words and asked to provide the American Sign Language translations. Bilinguals provided correct translations for 95.2% of the words (M=62.8/66, SD=2.5).

Results

Frequency of Looks

We measured both the proportion and duration of looks to competitor and control items. Bilinguals looked more at competitor items than at control items, and looked more at competitor items than monolingual participants. Repeated measures Analyses of Variance (ANOVAs) revealed a significant Group x Condition interaction for both the proportion [$F(1,24)=27.284$, $p<0.001$; Fig. 2] and duration [$F(1,24)= 23.285$, $p<0.001$; Fig. 3] of looks. Bilinguals looked at competitor items more than at control items [$t(12)=7.62$, $p<0.001$] and for a longer period of time [$t(12)=5.925$, $p<0.001$], signifying that bilinguals activated phonologically related competitors more than phonologically unrelated controls. No differences were found in the monolingual group for either the proportion [$t(12)=-0.95$, $p=0.362$] or duration [$t(12)=-0.16$, $p=0.87$] of looks. Bilinguals also looked at competitor items more than monolinguals [$t(24)=5.58$, $p<0.001$] and for a longer period of time [$t(24)=3.512$, $p<.01$], while both groups looked at control items equally [proportion=$t(24)=1.18$, $p=0.248$; duration=$t(24)=-.73$, $p=0.47$], verifying that bilinguals activated phonologically related items more than monolinguals. Means and standard errors are illustrated in Table 1.

Table 1: Means and Standard Errors of the Proportion and Duration of Looks (%).

	Proportion (%)		*Duration (%)*	
	Comp.	Control	Comp.	Control
Bilingual	66.9 (*2.8*)	51.6 (*3.5*)	12.9 (*0.8*)	7.7 (*0.6*)
Monolingual	42.9 (*3.2*)	45.5 (*3.8*)	8.5 (*0.9*)	8.6 (*1.0*)

Time Course

Analysis of the bilingual time-course was consistent with the overall looks analysis, with bilinguals looking at competitors more than at control items. In contrast, monolinguals looked at competitor and control items equally across time. The activation curves were divided into 100 ms windows, beginning with the time window between -600 and -500 ms (which signified the first 100 ms after the onset of the picture), and ending with the window between 1900 and 2000 ms. Three-by-two repeated-measures ANOVAs were performed on each individual window, with time (1, 2, 3) and condition (competitor, control) as within-subjects factors. Significant effects of condition were found in each of the four 100 ms time windows between 0 ms (word onset) and 400 ms, between 1000 and 1100 ms, and between 1300 and 1400 ms (all ps<0.05); in all cases, bilinguals showed more looks to competitor items than control items. Similar analyses performed on the monolingual activation curves revealed no

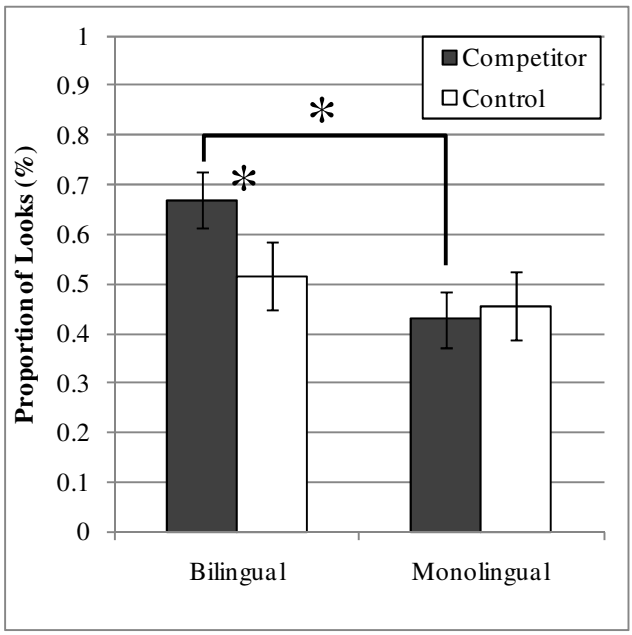


Figure 2. Proportion of Looks (%)

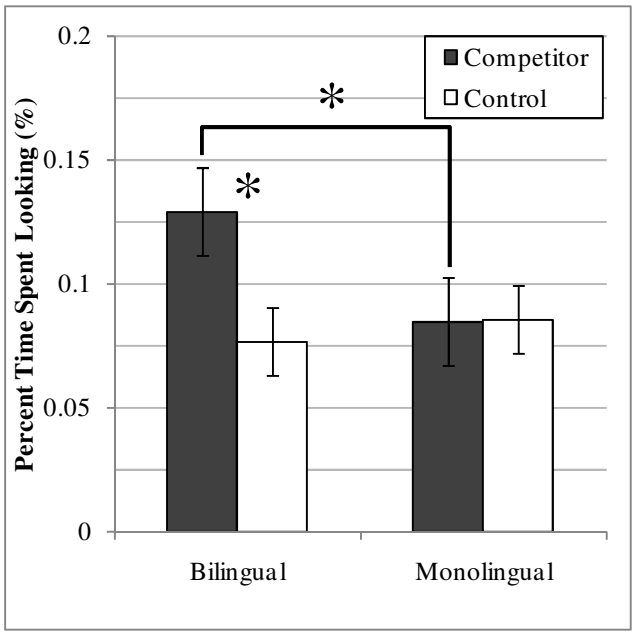


Figure 3. Duration of Looks (%)

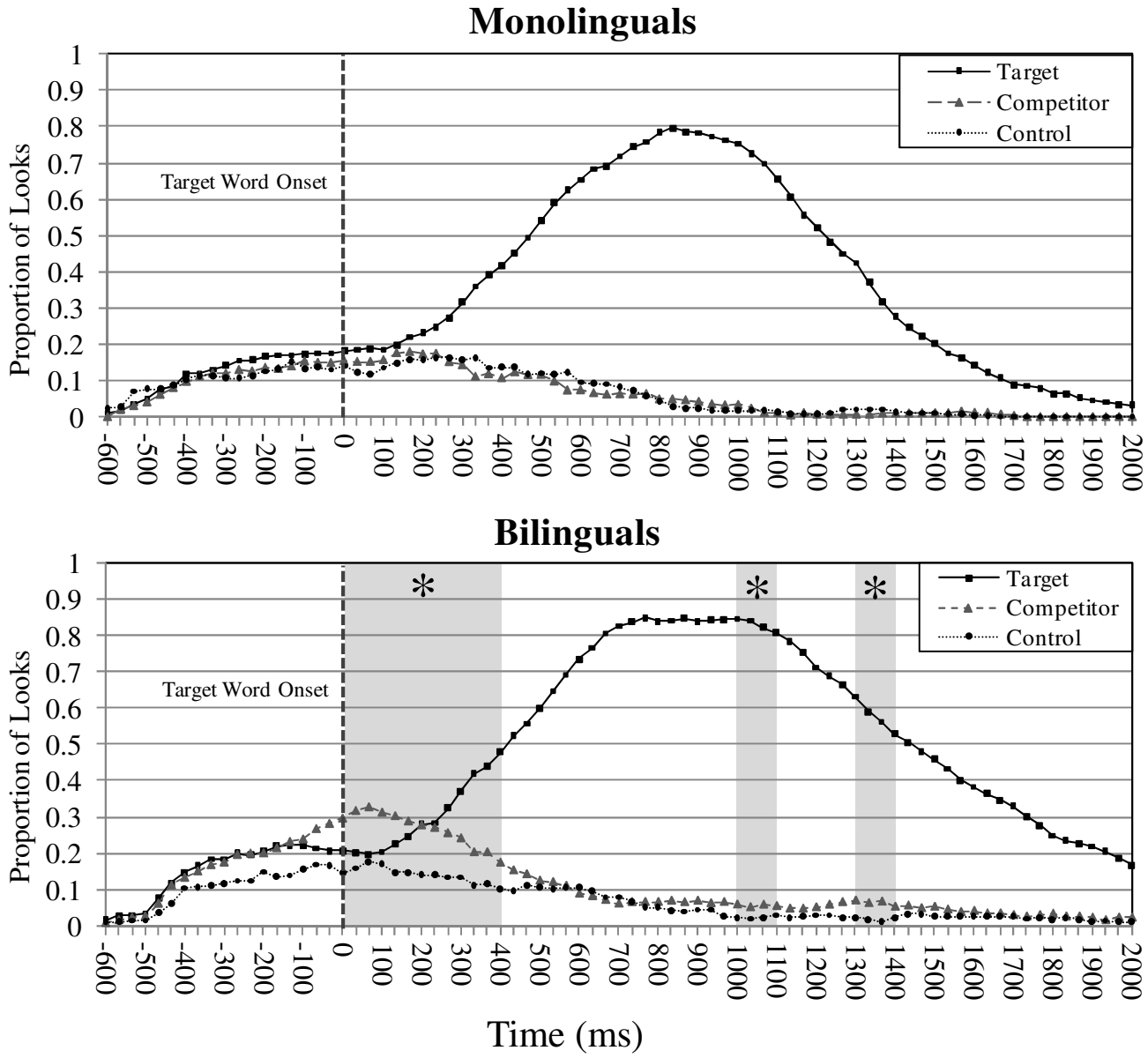


Figure 4. Time-course data for Monolingual (top) and Bilingual (bottom) participants, showing activation curves for proportion of looks to target, competitor, and control items across time. The negative 600 time point represents onset of the picture stimulus, and the 0 time point represents onset of the target word. Shaded areas indicate windows where looks to competitor and control items differed at $p<0.05$

significant effects of Condition for the monolingual group in any time window (all *p*s > 0.1), suggesting that monolinguals did not look more at competitor items than controls (see Figure 4).

It is possible that the effect seen in the late windows (1000-1100 ms and 1300-1400 ms) is a product of residual activation from the early window. To ensure that the late-window effects were not due to lingering activation from the early window, the proportion of late-window looks to competitors *with* a look to targets or competitors in the early-window was compared to late-window looks *without* earlier target or competitor fixations. If there was a higher proportion of late-window looks when the early window contained a look to either the target or competitor than when it did not, it would suggest that late-window activation was due to residual activation from the early window. However, both instances showed the same proportion of looks, $t(12)=1.04$, $p=.377$, suggesting that the results seen in the late-windows are not due to previous activation in the early-window.

Discussion

The results of the current experiment provide evidence for a modality-independent language system that utilizes top-down pathways during processing by revealing parallel language activation in bimodal bilinguals. Specifically, bilinguals looked more to items with ASL translation equivalents that overlapped phonologically with the target item than to items with translation equivalents that did not overlap, suggesting that phonologically overlapping competitor items were more activated than unrelated controls. In turn, monolinguals looked at competitor items and unrelated control items equally. This pattern was found in the overall looks analysis and in the duration of looks analyses, as well as within specific time windows during processing. The results suggest that even though the bilingual participants received no ASL input, they nevertheless activated their sign-language during the experiment.

The finding that bimodal bilinguals coactivate their languages indicates that lexical items from two distinct languages do not require surface-level overlap in order to be simultaneously activated. Previous studies on unimodal bilinguals have relied on bottom-up information as the force behind parallel language activation – words activate phonologically similar words, regardless of language. If parallel activation is driven purely by overlap at the phonological level, then the bimodal participants should not have shown cross-linguistic activation. Instead, the connection between ASL and English likely exists at the semantic level, since the two languages do not share phonological or lexical items. Semantic representations, once activated, appear to be able to feed back to the lexical levels of *both* signed and spoken languages, resulting in parallel activation.

However, the mechanisms that underlie parallel processing in bimodal bilinguals are unclear. Examination of the time-course showed that at the moment of the onset of the target word, competitor items were activated more than control items in the bilingual group. If the target word has yet to be presented in full, how is it possible that bilinguals would show increased activation of competitor items? Prior to onset of the word, bilinguals view the display containing all four images for 600 ms. Bilinguals may automatically activate the corresponding semantic concepts due to visual input. This activation can feed back ASL lexical levels and activate phonologically similar lexical items. The phonologically related items may then continually activate one another until target selection occurs.

The process of top-down activation of the non-target language in bimodal bilinguals can also be initiated by linguistic input. The initial semantic representation could be activated by the incoming English target word, rather than by visual stimuli. When the semantic representation is activated, it feeds back to the lexical representations in both English and ASL, thereby activating phonologically similar ASL signs and their corresponding semantic representations. It is important to note that in this account, parallel activation is still a product of top-down processes – the linguistic input should activate English only. However, bimodal bilinguals clearly activate their ASL during the task.

Coactivation may also occur via lateral links between translation equivalents. As an English word is presented, it may activate its ASL translation via direct excitatory connections at the lexical level, which may in turn activate phonologically similar ASL items. While this account does not involve top-down processes, it is also not exclusively bottom-up, and requires a system capable of interaction across languages, *within* a single level of processing. However, the strength of within-level translational connections in bimodal bilinguals is unclear. For instance, bimodal bilinguals do not show enhanced performance on executive control tasks, which has been found in unimodal bilinguals. Emmorey, Luk, Pyers, and Bialystok (2008) suggest that since a bimodal bilingual's two languages utilize separate modalities, they do not compete to the same extent as two spoken languages. Therefore there is less need for executive control of the non-target language in bimodal bilinguals. One could argue that the lack of competition between a spoken and a signed language may indicate that bimodal bilinguals do not develop cross-linguistic connections in the same manner as unimodal bilinguals. It is not yet clear whether the connections between translation equivalents, or the way in which they are processed, are similar in unimodal and bimodal bilinguals.

Regardless of whether parallel activation in bimodal bilinguals is due to top-down effects or lateral connections at the lexical level, it is clear that the processes that underlie language coactivation in bimodal bilinguals differ from those of unimodal bilinguals. While coactivation in unimodal bilinguals relies more on phonological overlap across two languages, no such overlap exists within the processing architecture of bimodal bilinguals. Therefore, the finding that bimodal bilinguals coactivate their languages implies a system where top-down or lateral processes are capable of governing cross- linguistic activation. Our results also indicate that language information may be readily transmitted across modalities, such that two highly unrelated languages can be activated simultaneously. Thus, the language system should be considered modality-independent and able to process linguistic information equally, regardless of whether it is auditory or sensorimotor in nature.

Moreover, there is reason to believe that a system of this nature is not unique to bimodal bilinguals, and may provide a window into more general language processing. Unimodal bilinguals and monolinguals have shown robust cross-modal effects as well (for a review, see Marian, 2009), and semantic-competition effects from eye-tracking provide evidence for rapid and highly robust lexical-semantic interaction (Huettig & Altmann, 2005; Yee & Sedivy, 2006). In addition, when bimodal bilinguals produce code-blends, they do so in the same fashion as unimodal bilinguals would code-switch, with the added benefit of

being able to produce speech and signs in tandem, suggesting similarities in the underlying mechanisms of production for unimodal and bimodal bilinguals (Emmorey et al., 2008). It is possible that the top-down pathways utilized by bimodal bilinguals are present in unimodal bilinguals and monolinguals as well, but are overshadowed by more immediate bottom-up effects.

The results of the current study indicate that bimodal bilinguals activate both of their languages simultaneously via a cross-linguistic lexical-semantic loop where top-down information from the conceptual level feeds back to lower levels of processing in both languages, regardless of modality. The results have further implications for the architecture and processing dynamics of the language system, bilingual and monolingual alike, suggesting that language information can be freely accessed across modalities, and that top-down mechanisms can have a strong influence on language processing.

Acknowledgments

This research was funded in part by grants NICHD 1R03HD046952 and NSF BCS-0418495 to the second author. The authors would like to acknowledge Dr. Margarita Kaushanskaya, Dr. Henrike Blumenfeld, Caroline Engstler, Scott Schroeder, James Bartolotti, Michelle Masbaum, and Rucha Mehta for their contributions to this project.

References

Blumenfeld, H., & Marian, V. (2007). Constraints on parallel activation in bilingual spoken language processing: Examining proficiency and lexical status using eye-tracking. *Language and Cognitive Processes, 22(5),* 633-660.

Brentari, D. (1998). *A Prosodic Model of Sign Language Phonology*. MIT Press.

Brysbaert, M., & New, B. (2009). Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavioral Research Methods, 41(4),* 977-990.

Dunn, L. M. & Dunn, L. M. (1997). *Peabody Picture Vocabulary Test, Third Edition*. Circle Pine, MN: American Guidance Service.

Emmorey, K., Borinstein, H.B., Thompson, R., & Gollan, T.H. (2008). Bimodal bilingualism. *Bilingualism: Language and Cognition, 11(1),* 43-61.

Emmorey, K., Luk, G., Pyers, J. E., & Bialystok, E. (2008). The source of enhanced cognitive control in bilinguals: Evidence from bimodal bilinguals. *Psychological Science, 19(12),* 1201-1206.

Hagoort, P., & Levelt, W. J. M. (2009). The speaking brain. *Science, 326(5951),* 372-373.

Huettig, F., & Altmann, G. T. M. (2005). Word meaning and the control of eye fixation: semantic competitor effects and the visual world paradigm. *Cognition, 96*, B23–B32.

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences, 22,* 1-75.

Marian, V., Blumenfeld, H., & Kaushanskaya, M. (2007). The language experience and proficiency questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals. *Journal of Speech, Language, and Hearing Research, 50(4),* 940-967.

Marian, V. & Spivey, M. (2003b). Competing activation in bilingual language processing: Within- and between-language competition. *Bilingualism: Language and Cognition, 6(2),* 97-115.

McQueen, J. M., Jesse, A., & Norris, D. (2009). No lexical-prelexical feedback during speech perception or: Is it time to stop playing those Christmas tapes? *Journal of Memory and Language, 61(1),* 1-18.

Navarette, E., & Costa, A. (2009). The naming of gender-marked pronouns supports interactivity in models of lexical access. *Psicológica, 30,* 301-321.

Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition, 52(3),* 189-234.

Rapp, B., & Goldrick, M. (2000). Discreteness and interactivity in spoken word production. *Psychological Review, 107(3),* 460-499.

Van Hell, J. G., Ormel, E., van der Loop, J., & Hermans, D. (2009). Cross-language interaction in unimodal and bimodal bilinguals. Paper presented at the *16th Conference of the European Society for Cognitive Psychology*. Cracow, Poland, September 2-5.

Villwock, A., Wilkinson, E., Bailey, R., Kroll, J., Morford, J., & Piñar, P. (2009). Cross-language lexical activation in deaf bilinguals: Does English print activate ASL signs? Presented at *The International Symposium on Bilingualism 7*. Utrecht, NL.

Weber, A., & Cutler, A. (2004). Lexical competition in non-native spoken- word recognition. *Journal of Memory and Language, 50*, 1-25.

Yee, E. & Sedivy, J. (2006). Eye movements reveal transient semantic activation during spoken word recognition. *Journal of Experimental Psychology: Learning, Memory and Cognition, 32*, 1-14.

Conditions of Directed Attention Inhibit Recognition Performance for Target-Aligned Stimuli

Andrew D. Dewald (adewald@hawaii.edu)
Leonidas A. A. Doumas (leonidas@hawaii.edu)
Scott Sinnett (ssinnett@hawaii.edu)
University of Hawaii at Manoa
Department of Psychology
2530 Dole Street, Honolulu, HI 96822

Abstract

Watanabe, Náñez & Sasak (2001) demonstrated that the perceptual learning of task-irrelevant items was enhanced under conditions when attentional resources were diverted away from the irrelevant stimuli. However, the current study suggests that when attention is depleted, recognition for task-irrelevant items is *impaired* in a subsequent recognition task. Participants were presented with a stream of simultaneously presented written words and line drawings, and required to respond to immediate repetitions in either the word or picture stream. A surprise recognition test measured performance for the words. When analyzing word recognition performance after attention had been directed to the pictures, words that had previously appeared when attention was most depleted (i.e., with a picture repetition in the primary task) were recognized at levels significantly *below* chance. This novel finding suggests that information that is actively ignored when appearing in conjunction with an attended stimulus is subsequently inhibited in a recognition task.

Introduction

The role of attention in human perception has been investigated extensively through the better part of experimental psychology's history (e.g., Ahissar & Hochstein, 1993; Broadbent, 1953; Cherry, 1953; James, 1890; Mack & Rock, 1998; Moray, 1954; Seitz & Watanabe, 2005; Sinnett, Costa & Soto-Faraco, 2006; Triesman, 1960). A number of findings converge on the notion that explicit perception requires, at least a certain degree of attention (Mack & Rock, 1998; Rees, Russell, Frith, & Driver, 1999). Indeed, this has been demonstrated even for cognitive processes that have been considered at one point to proceed in an obligatory or automatic fashion. For instance, written word recognition, audiovisual integration in speech perception, and motion detection have all been empirically supported to require explicit attention in order for perception to occur (Alsius, Navarra, Campbell & Soto-Faraco, 2005; Rees, Frith & Lavie, 1997; Rees et al., 1999).

Despite numerous examples suggesting that visually presented words are processed automatically (see, e.g., Lupker, 1984; Stroop, 1935), Rees et al (1999) demonstrated that when attentional reservoirs were depleted, written word perception was interrupted. In their experiment participants viewed a rapid serial visual presentation (RSVP) of written items (words or non-words), superimposed on top of a stream of pictures. The primary task was to detect immediate repetitions in either the picture or the word stream. Directly following this task, participants were given a word recognition test for the words that had previously been presented. Behavioral findings suggested that performance was significantly better (i.e., more words were correctly recognized) after directly attending to the words. Furthermore, after attending to the picture stream, participants were just as likely to incorrectly affirm that a non-presented foil word had in fact been presented as they were to correctly identify words that had been originally presented in the repetition detection task.

While the findings of Rees et al (1999) suggest that attention plays a critical role in word recognition, one could make the claim that the words were indeed perceived, but quickly forgotten because a stabile memory code could not be formed. However, the authors also compared brain activations via functional magnetic imaging (fMRI) between the presented non-words (consonant streams) and words in the repetition detection task. Importantly, and discrediting any memory based explanation, while attending to the picture stream, words and non-words failed to show different levels of activation in word processing brain areas, such as the posterior basal temporal region, an important area associated with word identification (Buchel, Price & Friston, 1998). Essentially, a string of consonants (e.g., BCRTM) was treated the same as a word (e.g., HOUSE) when attending to the picture stream (Rees et al., 1999). These results demonstrate that the processing of a written word requires that attentional resources be directed towards that word.

Rees and colleagues have also demonstrated a decrease in visual processing for motion when attentional resources were depleted (Rees et al., 1997). That is, when attention was diverted to a difficult task, a reduction in visual motion perception occurred. In this experiment participants performed linguistic judgment tasks of varying difficulty superimposed over a visual motion background while brain activity was measured with fMRI. The findings suggested that as the difficulty of the linguistic task increased, brain activity in an area associated with the processing of motion (V5; Tootell, 1995) diminished when compared to the easier task. The authors posited that as task difficulty increases, attentional resources that could otherwise be used to process task irrelevant stimuli are recruited for the more difficult task, resulting in a reduction in perception for task irrelevant events (Rees et al., 1997; see also Lavie, 1995; 2005 for a description of attentional load theory).

Despite a multitude of findings suggesting that perception levels diminish as attentional resources are depleted (see

Ahissar & Hochstein, 1993; Lavie, 2005; Mack & Rock, 1998; Rees et al., 1999; Sinnett et al., 2006), Watanabe et al. (2001; see also Seitz & Watanabe, 2003, 2005) demonstrated—in direct contrast to the results described above by Rees et al. (1997)—that the perception of irrelevant motion can actually be *increased* under situations when attentional resources are depleted. Indeed, they showed that participants' detection performance for task irrelevant motion stimuli improved under conditions when attention was directed to a separate task. Moreover, the improvement was only seen when the irrelevant motion was temporally aligned with targets occurring in the attention demanding task. This is surprising as it demonstrates a situation where improved perception is observed during moments when attention is arguably most depleted (i.e., when required to detect and respond to a target).

Watanabe et al's. (2001) participants took part in a series of experiments in which they were repeatedly exposed to a background motion signal that was set at either 5% or 10% coherent motion (see also Seitz & Watanabe, 2003; 2005 for further examples using the same task). When asked to determine the direction of coherent motion by choosing one of eight possible directions, participants performed at chance levels for the 5% condition, but above chance levels for the 10% motion condition (suggesting that the motion was subthreshold in the former, but not the latter, condition). The same task was also performed when engaged in a simultaneously presented attention-demanding task. An RSVP of letters was superimposed over the background motion, and participants were required to report the identity of white target letters that occurred in a sequence of black distractor letters. It is important to note that when the superimposed white target letter appeared (i.e., the task-target), the same subthreshold coherent motion direction was present every single time (i.e., the task-irrelevant target).

Upon completion of this task, participants were again shown the weak background motion signal and asked to indicate the direction of the motion by choosing from an array of eight directions (depicted as arrows). While the 5% coherent motion condition remained at chance performance before and after exposure, the 10% coherent motion condition showed significant improvements in perceptual performance for the coherent motion, but only for the specific motion that was synchronized with the presentation of the white target letter during exposure. Note, this result is surprising as it shows that an implicitly presented motion can have a later effect on behavior.

Watanabe and colleagues (2001; 2003; 2005) postulated that the improved motion perception is due to the temporal relationship between the task-relevant stimulus (presence of white letter) and the task-irrelevant stimulus (background motion). It was hypothesized that if these two stimuli were presented simultaneously, then the learning associated with attention being directed to the task-relevant features would also be applied to the task-irrelevant stimulus, despite attention being explicitly directed away from the motion stimulus. These findings are even more surprising when one considers that significant improvements in performance only occur when irrelevant stimuli are paired with the most demanding aspect of a secondary task (i.e., when attentional reservoirs are depleted, but directed to a temporally aligned target).

The findings of Watanabe and colleagues (2001; 2003; 2005) seemingly suggest that directed attention is not a necessary condition for the perceptual learning of irrelevant targets. While the results are ostensibly robust, their conclusions stand contrary to the wealth of research that suggests that these findings would be unlikely to occur; most research would indicate that perception for irrelevant stimuli would be diminished under conditions where attention is utilized in a separate task and not explicitly directed to the irrelevant stimuli (see for example Rees et al., 1997).

The present study aimed foremost to investigate the robustness of Watanabe and colleagues' claims and expand their findings to a different type of stimulus; explicitly presented written words, using a different paradigm. Accordingly, task-relevant items (visual pictures) were temporally aligned with task-irrelevant (written words) items in a RSVP stream to see if this synchronization would lead to enhanced recognition levels of the task-irrelevant items. Based on the findings of Watanabe et al. (2001), enhanced performance would be predicted for task-irrelevant words that appear at the same time as a target picture when compared with words that do not.

Method

Participants.

Forty participants (n=40) were recruited from the University of Hawai'i at Manoa in exchange for course credit. Participants were naïve to the experiment and had normal or corrected to normal vision.

Materials.

A total of 150 pictures were selected from the Snodgrass and Vanderwart (1980) picture database. The pictures (on average 5 to 10 cm's) were randomly rotated +/-30 degrees from upright so as to ensure the difficulty of the task in each version of the experiment (see also Rees et al., 1999). Each of these pictures was combined with 150 one to two syllable, high-frequency English words (average length of 5 letters; range 4-6) selected from the MRC psycholinguistic database (Wilson, 1988). The overall average frequency of the 150 selected words was 120 per million, ranging between 28 and 686. The words were displayed in bold, capitalized letters in Arial font at a size of 24 points. Each word was superimposed over a picture and the picture-word stimuli did not exceed 10 cm horizontally or vertically. Care was taken to ensure that picture-word combinations did not have any semantic relationship.

Two streams of picture-word stimuli were created. In one stream, 50 pictures were selected from the database, 25 of which were pre-selected, duplicated and paired with their

match. These repeated pictures acted as targets as each pair occurred in the visual presentation as an immediate repetition. The remaining 25 pictures were also duplicated, but their positioning in the stream of stimuli never allowed for an immediate repetition. Together this created a block size of 100 items. A second block of 100 items was created in which the 25 pictures not immediately repeated in the first block now served as the pictures that were immediately repeated. Therefore, across both blocks, each picture was displayed a total of four times (once as a repeat and then two other times as non-repeats in the complementary block). The same principle was used when making streams of items when the words were repeated (attending to words condition). To ensure an enhanced level of randomization, three different groups of 50 words and pictures were created and randomized in the aforementioned fashion, creating six different versions of the picture-word superimposed stimuli for use in the attending to pictures condition as well as the attending to words condition.

The surprise recognition test administered after the completion of the repetition detection task, consisted of 100 words from both the previously viewed visual stream (50) as well as never seen before foil words (50). The foils were words that were used in a different version of the experiment as repeated words (fully randomized). The 50 non-foil words presented in the surprise recognition test were words that were either temporally aligned with the task-relevant target, (i.e., superimposed over the immediate repetition of a picture), or were not temporally aligned with the task-relevant target (i.e., superimposed over non-immediately repeating pictures). Words synchronized with task-relevant targets have been given the nomenclature of *target-aligned* words and those not aligned with task-relevant targets have been named *non-aligned* words (see Table 1).

Table1: Description of *Target-Aligned* and *Non-Aligned* words.

Word Type	**Synchronized Temporal Pairing with Task-Target of Immediately Repeated Pictures**	**Synchronized Temporal Pairing with Non-Task Target of Non-Immediately Repeated Pictures**
Target-Aligned	Yes	No
Non-Aligned	No	Yes

Both the repetition detection and word recognition tasks were randomized and presented on a computer screen one at a time, in bold, capitalized letters in Arial font at a size of 24 points, just as they were displayed in the previous stream. The words in the recognition test remained on the screen until a response was made.

Procedure.

Participants were randomly assigned to one of two conditions. One group was required to attend to the picture stream (i.e., ignore the superimposed words) and respond to immediate picture repetitions, while the other group was required to respond to immediate repetitions in the word stream. Participants responded to the repetitions by using the 'G' key on the keyboard.

Each item in the picture-word presentation was presented for 350 ms with a 150-ms inter-stimulus interval (ISI; blank screen) between each item for a stimulus onset asynchrony (SOA) of 500 ms (see Figure 1). Before the first experimental block, a training block of eight trials was given and repeated until participants were familiar and comfortable with the task.

Immediately after the repetition detection task, a surprise word recognition test was administered to all participants. Words were displayed individually on the center of the screen in the same size and font as previously presented in the repetition detection task, and remained on the screen until the participant made a response. Participants were instructed to press the "B" key if they had seen the word during the repetition detection task or, instead, the "V" key if they had not seen the word before. Within each group, half of the participants (n=10) were presented with foils and target-aligned words, while the other half were presented with foils and non-aligned words.

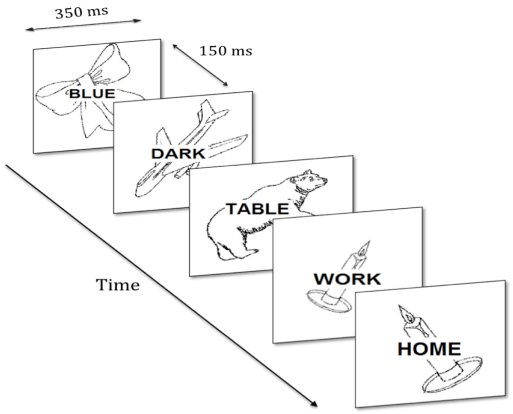


Figure 1. Rapid Serial Visual Presentation sequence employed. Each picture–word stimulus was presented for 350 ms and was then replaced by a blank screen for 150 ms before the next stimulus. Both the word-monitoring task and the picture-monitoring tasks were performed on the same streams. Note that in the present example, the word "HOME" serves as a *target-aligned* word.

Results

Overall surprise recognition performance.

The results of the surprise recognition test were analyzed in order to compare between conditions (attending pictures vs. attending words), and also against chance levels. Overall, recognition performance was significantly better after attending to the words when compared with after attending to the pictures (59.4%, *SE*=1.08 vs. 46.7%, *SE*=2.12, $t(19)=3.94$, $p=0.001$; see Figure 2). Performance after attending to the words was significantly better than chance ($t(19)=5.19$, $p<0.001$) while performance after attending to the picture stream was not significantly better than chance ($t(19)=1.52$, $p= 0.143$).

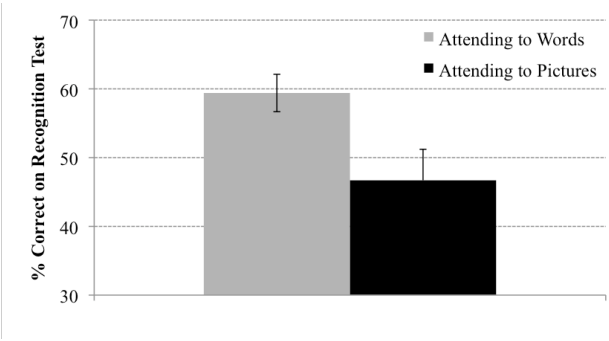


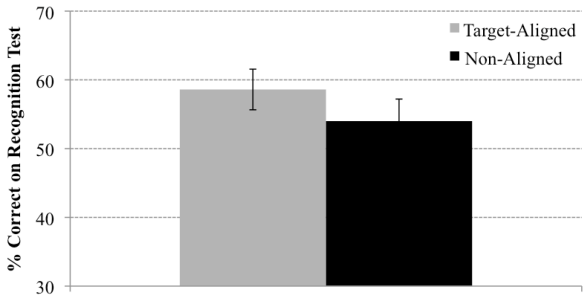
Figure 2. Overall recognition percentages and standard error bars for correct identification of words in the surprise word recognition test after attending to either the word stream (grey bar) or the picture stream (black bar).

Target-aligned word recognition.

In order to address the question at hand, that is, if performance is enhanced for words appearing with a picture repetition, recognition performance for *target-aligned* words was compared with *non-aligned* words and also against chance. When attending to words in the repetition task, subsequent recognition for *target-aligned* words (words immediately repeated) was significantly better than chance performance (59%, $t(9)= 2.67$, $p=.025$), while recognition for *non-aligned* words (not immediately repeated) was not statistically different from chance (54%, $t(9)= 1.35$, $p=.210$). There were no significant differences between *target-aligned* and *non-aligned* word performance after attending to the words ($t(9)=1.30$, $p=.224$; see Figure 3a). Analysis of recognition performance after attending to the picture stream demonstrated that participants were not better than chance at recognizing *non-aligned* words (50%, $t(9)= 0.08$, $p=.931$). Interestingly, performance was significantly different from chance at recognizing *target-aligned* words (38%, $t(9)= 4.54$, $p=.001$).

However, the direction of this significance was the opposite of what was expected, with performance significantly worse than chance (see Figure 3b). When compared to each other, recognition for *non-aligned* words was significantly better than *target-aligned* words ($t(9)= 2.34$, $p=.044$).

A.



B.

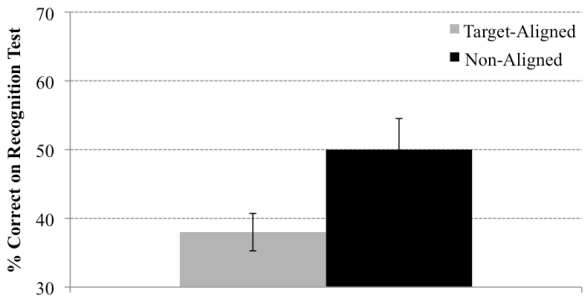


Figure 3. Recognition percentages and standard error bars for Target-Aligned (grey bar) and Non-Aligned (black bar) words in the surprise word recognition test after attending to either the word stream (A) or the picture stream (B).

An analysis was also conducted on the accuracy of the primary task of immediate target repetition detection. Overall, subjects were able to accurately detect target repetitions (75% hit rate vs. 25% miss rate, $t(9)= 21.69$, $p<.001$, see also Sinnett et al. 2006 for similar hit rates using the same paradigm). In addition, a significant negative correlation was found between target detection accuracy and recognition performance for *target-aligned* words ($r(10) = -.69$, $p = .02$), further suggesting that target-aligned words are inhibited in the recognition task.

Discussion

There are three main findings for the current experiment. First, we have replicated previous findings on inattentional blindness showing that word recognition is significantly better after attending directly to the word stream as opposed to attending to a distracting stream of pictures (see also Most, Simmons, Scholl, Jiminez & Chabris, 2001, Rees et al., 1999; Sinnett et al., 2006). Second, word recognition failed to be significantly better than chance levels after attending to the picture stream. That is, participants were unable to recognize the words if their attention had been placed elsewhere, suggesting that attention may be a

necessary component for word recognition (see also Rees et al., 1999; Sinnett et al., 2006). Lastly, we have shown for the first time that words that appeared with a picture repetition (i.e., target-aligned) are recognized at significantly lower than chance levels after attending to the picture stream, suggesting, perhaps, an inhibition for irrelevant information that appears simultaneously with an attended target. Furthermore, after attending to the words themselves, subsequent recognition was better than chance for words that had appeared as a target repetition (i.e., target-aligned), while at chance levels for those that had appeared elsewhere in the stream (i.e., non-target aligned). Accordingly, this suggests that words that appeared with a target repetition were either inhibited or facilitated, depending on whether attention was originally directed to the pictures or the words in repetition detection task, respectively.

The finding that there is a possible inhibition of previously viewed words that appeared with a picture target stands in direct contrast to the conclusions drawn by Watanabe and colleagues (2001; 2003; 2005). For their findings to be replicated here, an enhanced recognition performance for words synchronized with task-relevant targets should have occurred. However, while the necessary temporal synchronization between task-relevant and task-irrelevant stimuli was present, enhanced perception for task-irrelevant stimuli was not observed. In fact, the exact opposite was seen, in that there was an inhibition of performance for the recognition of words that were temporally aligned with the task-relevant target of an immediate picture detection.

The potential inhibition of the target-aligned words when attention was diverted to the picture stream is of key interest to the present findings. While it is apparent that many investigations have found that when attentional resources are depleted, unattended and irrelevant stimuli are often not perceived (Mack & Rock, 1998; Rees et al., 1999; Sinnett et al., 2006), an inhibition for these stimuli has not been observed. However, it should be noted that to the best of our knowledge, this is the first time that a distinction between irrelevant stimuli appearing with a target, or not, has been empirically investigated. When doing precisely this in the present study, an inhibition for words that appeared with repeated target pictures was observed. One possible explanation for this would be that due to focused attention being placed directly on the demanding task of detecting repetitions, thereby necessitating that the attentional system actively inhibit irrelevant information in order to facilitate goal oriented behavior.

Despite significant differences in paradigms, a possible explanation for the inhibition of *target-aligned* words after attending to pictures may be found in the inhibition of return (IOR) literature (see Klein, 2000 for a review of IOR). If a target stimulus occurring in the periphery is first cued by a salient attention grabbing event, then a facilitation is normally found for the processing of that target if the time between the cue and the target is relatively short (i.e., < 300 ms; Posner, 1980). However, if there is a longer time period between the cue and the target (i.e., after attention has been disengaged from that space), then there is a delay (i.e., inhibition) for processing of targets in the previously cued area. This might be analogous to what was observed in the present experiment: Information that was attended to is later inhibited. However, it should be noted that this comparison is difficult to make as IOR is traditionally seen in visual search paradigms and measure response latency, while the present findings result from a non-spatial paradigm measuring accuracy. Nevertheless, the present findings could be viewed as an instantiation of a non-spatial, accuracy based inhibition for ignored stimuli.

As the comparison between visual search and the present paradigms can be viewed as difficult at best, perhaps a stronger explanation for the present results can be drawn from research on negative priming (see Milliken, Joordens, Merikle, & Seiffert, 1998; Tipper, 1985; Tipper & Driver, 1988). Typically, in negative priming experiments observers are presented, for instance, with two overlapping streams of object outlines with each stream printed in a different color (i.e., green and red). Participants would be required to name items in one stream (green objects) while ignoring stimuli in the other stream (red objects). Interestingly, response latencies are slower for objects that had appeared previously in the ignored stream (i.e., the to-be- ignored color), than for objects that participants did not have to ignore previously. Accordingly, this suggests that while selecting and naming one picture, the other (simultaneously displayed but not selected) object seems to be processed as well, at least to the extent that it influences naming latencies in the following trial. The theoretical implications of this could quite obviously be supported by the present findings, as behavioral responses to the ignored items here were inhibited in the form of response accuracy.

The significant negative correlation between target detection accuracy and recognition performance for *target-aligned* words further illustrates the possibility of negative priming. That is, while there was a high level of accuracy for immediate picture repetition detection, performance was decreased for recognition of *target-aligned* words superimposed over the target pictures. Perhaps, as occurs in the aforementioned negative priming paradigms, the accurate detection of the primary target is related to decreased recognition accuracy (rather than a response latency) for the ignored *target-aligned* words.

Performance on the surprise word recognition after attending to the word stream was comparable to that of previous findings, suggesting that if attention is directed to words, they are recognized at both better than chance levels and better than after attending to the picture stream. While this is not surprising, there is one noteworthy finding: Overall better than chance performance is driven by *target-aligned* words (words immediately repeated and serving as task targets). That is, recognition performance for *non-aligned* words was not better than chance. Arguably, an increased amount of attention is allocated to target

detection, thereby potentially facilitating memory consolidation and subsequent performance in the word recognition task (see Craik & Lockhart, 1972 for a discussion on levels of processing theory). Accordingly, the present findings suggest an inhibition for target-aligned words when attention was directed to the picture stream, but a trend in the data (59% *target-aligned* vs. 54% *non-aligned*) for a facilitation of target-aligned words when attention was directed to the words themselves.

It is important to take into consideration significant procedural differences between the present study and the works by Watanabe and colleagues (2001; 2003; 2005). A detailed analysis of Watanabe et al's. (2001) original paradigm shows that a total of 960 trials, in which 120 consisted of the paired task-relevant and task-irrelevant stimuli, were presented daily for 20 days (i.e., nearly 100 times the amount here). In addition, the 120 paired task-relevant/-irrelevant stimuli always had the same direction in the coherent motion background. This would be equivalent to presenting only one specific word to appear with picture repetitions in the present study. Therefore, it might be possible that perception for irrelevant information paired with task-relevant information in the Watanabe et al. studies was an artifact of prolonged exposure in addition to the temporal synchronization (although this may be negated by an increased perception for the coherent motion paired with the task relevant target only). Future research could employ the paradigm from the present study to investigate prolonged exposure rates through the utilization of a larger number of trials and a smaller number of *target-aligned* words to see if perception is enhanced, rather than inhibited.

References

Alsius, A., Navarra, J., Campbell, R., & Soto-Faraco, S. (2005). Audiovisual integration of speech falters under high attention demands. *Current Biology, 15*(9).

Ahissar, M. and Hochstein, S. (1993). Attentional control of early perceptual learning. Proceedings of the National Academy of Science U.S.A, 9.

Craik, F.I.M. & Lockhart, R.S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning & Verbal Behavior, 11* (6).

DeSchepper, B. & Treisman, A. (1996). Visual memory for novel shapes: Implicit coding without attention. *Journal of Experimental Psychology: Learning Memory and Cognition, 22*(1).

Driver, J., & Spence, C. (2004). Crossmodal spatial attention: Evidence from human performance. In C. Spence & J. Driver (Eds.), Crossmodal space and crossmodal attention. Oxford, UK: Oxford University Press.

Duncan, J., Martens, S., & Ward, R. (1997). Restricted attentional capacity within but not between sensory modalities. *Nature, 387*.

Lavie, N. (1995). Perceptual load as a necessary condition for selective attention. *Journal of Experimental Psychology: Human Perception and Performance, 21*.

Lavie, N. (2005). Distracted and confused?: Selective attention under load. *Trends in Cognitive Sciences,* 9.

Lupker, S.J. (1984). Semantic priming without association: A second look. *Journal of Verbal Learning and Verbal Behavior, 23*.

Klein, R.M. (2000). Inhibition of return. *Trends in Cognitive Sciences*, *4*(4).

Mack, A., & Rock, I. (1998). Inattentional blindness. Cambridge, MA: MIT Press.

Milliken, B. , Joordens, S. , Merikle, P. & Seiffert, A.E. (1998). Selective attention: A reevaluation of the implication of negative priming. *Psychological Review, 105* (2).

Moray, N. (1959). Attention in dichotic listening: Affective cues and the influence of instructions. *Quarterly Journal of Experimental Psychology, 11,* 56-60.

Most, S. B., Simons, D. J., Scholl, B. J., Jimenez, R., Clifford, E., & Chabris, C. F. (2001). How not to be seen: The contribution of similarity and selective ignoring to sustained inattentional blindness. *Psychological Science, 12*.

Posner, M.I., (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology, 32*.

Posner, M.I., & Peterson, S.E. (1990). The attention system of the human brain. *Annual Review of Neuroscience,13*.

Rees, G., Russell, C., Frith, C. D., & Driver, J. (1999). Inattentional blindness versus inattentional amnesia for fixated but ignored words. *Science, 286*.

Sinnett, S., Costa, A., & Soto-Faraco, S. (2006). Manipulating inattentional blindness within and across sensory modalities. *Quarterly Journal of Experimental Psychology, 59*(8).

Seitz, A.R. & Watanabe, T. (2003). Psychophysics: Is subliminal learning really passive? *Nature, 422*, 36.

Seitz, A.R. & watanabe, T. (2005). A unified model for perceptual learning. *Trends in Cognitive Science, 9* (7).

Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory, 6*.

Tootell, B., Silverman, M.S. & R.L. De Valois, R.L. (1995). Spatial frequency columns in primary visual cortex. *Science*, 214(4522).

Tipper, S.P. & Driver, J. (1988). Negative priming between pictures and words in a selective attention task: Evidence for semantic processing of ignored stimuli. *Memory & Cognition, 16*(1).

Treisman, A. (1960). Contextual cues in selective listening. *Quarterly Journal of Experimental Psychology,12*.

Watanabe, T., Náñez,Y., & Sasak, S. (2001). Perceptual learning without perception. *Nature, 413*.

Information Selection in the Blogosphere: The Effect of Expertise, Community Rating, and Age

Stephan Winter (stephan.winter@uni-due.de)
Nicole C. Krämer (nicole.kraemer@uni-due.de)
Jana Appel (jana.appel@stud.uni-due.de)
Kathrin Schielke (kathrin.schielke@stud.uni-due.de)
University of Duisburg-Essen
Department of Social Psychology: Media and Communication
Forsthausweg 2,
47057 Duisburg, Germany

Abstract

The World Wide Web offers a lot of information that has been provided by laypersons instead of experts or professional journalists. This raises the question how Internet users perceive credibility of online authors and which information on the source influences the users' selection and processing of texts. Our study investigated the effect of self-reported expertise, community rating, and age of weblog authors. In an online laboratory experiment, information seeking behavior of 60 participants on a science weblog was analyzed. As exemplary scenario, the discussion on the effects of violent media contents on children was chosen. Results showed that authors with a high level of expertise (operationalized by the author's self-reported profession) were rated as more credible and their texts were selected for further reading more frequently. This suggests that self-reported expertise emerges as a strong cue for information selection, whereas there was only partial evidence for the importance of community ratings.

Keywords: Credibility, Selective Exposure, Persuasion, Source Cues, Information Processing.

Introduction

The Internet is today's largest source of information and communication. As Metzger (2007) points out, "more information from more sources is available and more easily accessible now than ever before" (p. 2078). Although this can definitely be seen as a major advancement, it might also lead to the problem that users get lost in the digital world and do not know how to find the content they need, e.g. when searching for information on science related issues. This phenomenon of "information overload" in the Internet (Eppler & Mengis, 2004) has brought new attention to the issue of credibility and quality of information – especially since the World Wide Web is rapidly developing in the direction of user-generated-content (Web 2.0, O'Reilly, 2005). For example, in blogs and forums "any user can say anything about any topic" (Van der Heide, 2008, p. 30). Thus, one can increasingly find information that has been provided by laypersons instead of experts or professional journalists and therefore may be less reliable.

This raises the question how Internet users perceive credibility of online authors and which information on the source influences the users' selection and processing of information in the World Wide Web. While previous research on selective exposure focused on content features such as the relevance of the topic (e.g. Knobloch, Zillman, Gibson, & Karrh, 2002; Zillmann, Chen, Knobloch, & Callison, 2004), information on the authors has not been taken into account yet. Similarly, models of online information seeking (e.g. Pirolli & Card, 1999; Schamber & Bateman, 1996; Tombros, Ruthven, & Jose, 2005) consider factors like title, currency and layout. With respect to theoretical modelling, it can be asked whether these models have to be amended by aspects of social cognition with regard to the authors.

Against this background, we wanted to investigate the effect of source cues – self-reported expertise, community rating, and age of authors – on the perception of credibility and the selection of online science information. Who do Internet users trust? And whose information do they select? Our examination focuses on weblogs (or blogs), which can be defined as "frequently updated websites where content is posted on a regular basis and displayed in reverse chronological order" (Schmidt, 2007). These websites are popular means of science communication in the Web (e.g. www.scienceblogs.com). Therefore, they are increasingly used by laypersons for obtaining information on science-related issues. As exemplary scenario for our study, we chose the discussion on the effects of violent media contents on children and adolescents.

Credibility and Information Selection in the Web

While several studies examined the general credibility of the Internet as a medium (Stavrositu & Sundar, 2008; Metzger, Flanagin, Eyal, Lemus, & McCann, 2003) or the credibility of different web sites as a whole (e.g. Walther, Wang, & Loh, 2004), this study focuses on the credibility of authors within a certain web site, which means that the analyzed message sources here are persons. According to the theory of social information processing (Walther, 1992), impressions of persons in computer-mediated communication are formed on the basis of verbal, linguistic, and textual manipulations – even though a lot of information that would be visible in face-to-face communication is missing. These impressions, primarily based on text-based cues, accrue over time and lead to a

relatively elaborate evaluation of other persons. In this context, Walther (1996) stated that, due to the absence of other cues, basic personal information might even be more important than in face-to-face situations (hyperpersonal communication).

Van der Heide (2008) distinguishes between system generated cues (e.g. the number of posts in a forum or the number of friends on the social networking site Facebook), aggregated feedback systems (such as reputation or rating systems) and self-disclosure behaviors (e.g. self-report of profession and age) as relevant types of heuristically valuable information about computer-mediated message senders. While system generated cues and aggregated feedback systems are based on information that has been provided by other users or the computer system itself, self-disclosures are easier to manipulate by the authors themselves. This means that someone might claim to be an expert although he is not. On the other hand, self-disclosures are "an efficient, direct, and visible method of communicating one's qualification" (Van der Heide, 2008, p. 24) and might therefore be particularly important.

As "authority is no longer a prerequisite for content provision on the Internet" (Metzger, 2007, p. 2078), it seems reasonable that people use these information on the author and his/her estimated credibility as a criterion for information selection. However, it has not been analyzed yet if these cues are a relevant factor for laypersons who are seeking information on science-related everyday issues in the Internet.

Expertise

Persuasion research in the tradition of the Yale studies (e.g. Hovland, Lumsdaine, & Sheffield, 1949) shows that messages presented by persons with a high level of expertise are more likely to influence other people (Wilson & Sherell, 1993). Therefore, dual-models of information processing (Chaiken, 1987; Petty & Cacioppo, 1986) include expertise of the source as one major factor – which is especially relevant if the level of elaboration is low.

Expertise Communicated via Self-Report On a weblog on science-related issues, self-reports, which may consist of a short self-description and the profession of the author, are able to provide important cues on the expertise of the author. This information is able to serve as a heuristic ("experts are usually correct"). As humans are cognitive misers (Fiske & Taylor, 1991) who do not include more cues than necessary for their decisions, it seems plausible to assume that this aspect is already relevant for the selection of information. For an investigation of newsbots such as Google News, Sundar, Knobloch-Westerwick, and Hastall (2007) demonstrated that source credibility cues (name of the medium in which a certain article was found, e.g. New York Times vs. tabloid newspaper) – which can be seen as an equivalent of expertise information on the level of persons – influenced perceived message credibility and likelihood of clicking. Following these results and considerations on persuasion research, we hypothesize that the information on the expertise of the author influences rating of the source and selective exposure to the corresponding message:

H1a: Sources with a high level of self-reported expertise will be perceived as more credible.
H1b: The texts of authors with a high level of self-reported expertise will be selected more often than the ones of the low-expertise-sources.

Expertise Attributed by Others (Community Ratings) Next to self-reports, expertise can also be expressed through the statements of other users. Therefore, collaborative filtering, e.g. rating systems (1 to 5 stars) or popularity indications (most e-mailed, number of views), is also likely to influence information choice. As these ratings are difficult to manipulate, they provide valuable information on the qualities of the user. Walther et al. (2009) showed that comments of friends on social networking sites are even more important for impression formation than self-generated statements. Furthermore, according to Chaiken (1987), people use the heuristic that, if many agree with an opinion, the opinion is probably correct. In this line, community ratings should produce a bandwagon effect (Sundar & Nass, 2001) in that articles or elements which already have a positive rating are clicked more frequently. On the other hand, individuals sometimes seek distinctiveness from others (Brewer, 1991), which would be an explanation for the opposite effect. Previous research has supported the idea of the bandwagon effect: In an experiment on selective exposure, Knobloch-Westerwick, Sharma, Hansen, and Alter (2005) found that online articles with better explicit recommendations were read longer. Additionally, Resnick, Zeckhauser, Swanson, and Lockwood (2006) showed that sellers with a high rating at the auction website Ebay were able to sell products for higher prices than users without a positive reputation. In this context, we hypothesize that:

H2a: Authors with a high community rating are perceived as more credible than authors with a low community rating.
H2b: Texts of sources with a high community rating are more likely to be chosen.

Social Comparison (Age)

Furthermore, social comparison (Festinger, 1954) may be relevant for selection. According to Festinger's theory, people are motivated to evaluate their opinions and abilities in comparison to similar persons, e.g. people with the same socio-demographic background (age, gender, education, etc.). The (positive or negative) results of this comparison process have been shown to influence self-evaluations and behavior (Mussweiler, 2001). In order to gain information that is relevant for social comparison, people should choose content that is connected to similar persons. In an experiment with an online news magazine, Knobloch-

Westerwick and Hastall (2006) already demonstrated that recipients more often choose news with protagonists of the same sex and that young readers prefer texts about same-age-characters. As similar effects can be expected for text authors, we hypothesize that:

H3a: Users perceive sources of similar age as more credible.
H3b: Users choose texts that were written by sources of similar age.

Method

Sample

In order to investigate these hypotheses, we created an online laboratory experiment in which 60 German participants were asked to search for information on a science weblog. As exemplary scenario, the website dealt with the controversy on the effects of violent media contents on children and adolescents. To ensure that this topic was personally relevant, participants were parents with children between the age of 2 and 18. Subjects were recruited via different channels, e.g. newspaper ads, postings in forums for parents and flyers which were distributed in schools. Participants (30 female, 30 male) were between the age of 22 and 47 (M = 36.93; SD = 6.54). 26.7 % of them had a university degree, 31.7 % finished high school with a qualification for university entrance and 41.7 % finished high school without this degree.

Stimulus Material

As stimulus material a blog platform (see figure 1) was created. On the overview page, 16 summaries of articles (with a headline, short description and information on the author) were shown. By clicking on the summary, the user was able to read the whole article – furthermore, it was possible to get more information on the author.

Independent Measures

As independent measures, the information on the author (self-reported expertise, rating, age) was systematically varied as within-subject factors. Expertise was operationalized via profession (professions with a close connection to the topic, e.g. psychologist (high) vs. professions without a connection to the topic, e.g. banker (low)). Sex, rating and age were also varied (rating: five or four stars vs. one or two stars / age: 24-27 years vs. 42-45 years). As a result, there were 16 combinations of author information that were shown below the headlines of the summaries. For every combination, a fictitious "character" was created (e.g. "Dr. Thomas Moos, 42, media scholar, community rating: 2 out of 5 stars" or "Jens Kohwall, 27, insurance broker, community rating: 5 out of 5 stars"). Headlines and texts were written in a neutral tone (e.g. "New studies on the effects of first-person shooters" or "Survey on children's media usage"), and connections between authors and texts were systematically rotated to avoid effects of the different topics and formulations.

Dependent Measures

As dependent measures, information selection and rating of the information and the source's credibility were assessed. It was coded which of the texts were chosen (in which order) and how long the texts were read. Furthermore, it was assessed whether the participants decided to get more information on the author. Credibility was measured with a scale based on research by Berlo, Lemert, and Mertz (1969) and Gierl, Stich, and Strohmayr (1997), including items like "trustworthy", "experienced" and "altruistic".

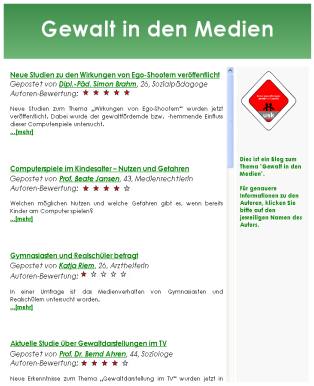


Figure 1: Screenshot of the weblog
(Title: "Violence in the Media")

Procedure

Data were collected in a laboratory at the University of Duisburg-Essen. First, the participants filled out an online questionnaire in which their previous knowledge on the topic, their media usage, need for cognition (Cacioppo & Petty, 1982) and self-efficacy with regard to Internet and Web skills (Eachus & Cassidy, 2006) were assessed. After that, they were told to search for information on the topic by reading the weblog. In order to create a selection situation, time was limited to four minutes. The sessions were saved with a screen-recording software. After that, the participants filled out a post-questionnaire in which they rated the credibility of the authors.

Results

Usage of the weblog

The participants of the study selected an average of 5.68 articles (SD = 1.99) during four minutes of reading time. Average reading time per article was 28.60 seconds (SD = 12.52). 25 % of the participants wanted to see further information on the author.

H1: Self-Reported Expertise

H1 predicted that authors with a high level of self-reported expertise (with a profession that has a close connection to

the topic) are perceived as more credible (*H1a*) and that their texts are selected more often *(H1b)*. To test these hypotheses, we conducted an analysis of variance (ANOVA) with repeated-measures in which the values for the authors were grouped according to their level of expertise. This revealed a significant effect of self-reported expertise on credibility ratings, F (1, 59) = 98.040, p = .000, η_p^2 = .624. As table 1 shows, the credibility scores for high-expertise authors are higher than for the low-expertise sources. Therefore, *H1a* has been supported by the data.

Table 1: Descriptive statistics for the effect of self-reported expertise on credibility score, number of clicks and reading time (in seconds)

	M	SD	N
Credibility Score High Expertise	153.35	19.45	60
Credibility Score Low Expertise	115.30	24.47	60
Number of clicks High Expertise	3.13	1.44	60
Number of clicks Low Expertise	2.55	1.53	60
Reading Time (s) High Expertise	79.43	34.69	60
Reading Time (s) Low Expertise	67.67	41.32	60

For the number of clicks, ANOVA also revealed a significant effect of expertise, F (1, 59) = 4.145, p = .046, η_p^2 = .066. The mean values (see table 1) show that texts that were attached to authors with a high level of self-reported expertise were selected more often for further reading. This means that *H1b* can also be supported. However, it has to be noted that the effect size is low.

With regard to reading time, the mean values (see table 1) indicate that texts of high-expertise-authors were read longer. However, ANOVA did not show a significant effect.

H2: Community Rating

H2 predicted that the participants prefer authors with a high community rating. However, with regard to credibility evaluations, no significant result was revealed *(H2a)*. For the number of clicks *(H2b)*, the mean values indicate that texts of authors with a high rating were selected more often (M = 3.07; SD = 1.52) than texts of authors with a low rating (M = 2.62; SD = 1.45). However, this trend was not significant. As a result, *H2* is not supported by these data. In further exploratory analyses, we found that, if only the authors with a low level of self-reported expertise are taken into account, community rating has a positive, marginally significant effect on the number of clicks, F (1, 59) = 3.020, p = .087, η_p^2 = .049: Participants selected an average of 1.40 texts of high-rating-authors (SD = 1.01) in comparison to an average of 1.15 texts of low-rating-authors (SD = .88).

H3: Social Comparison (Age)

H3 stated that the participants would perceive sources of the same age as more credible and choose their texts more often. For this analysis, the sample was separated into two age groups (from 22 to 38 years and from 39 to 47 years) via median split. With regard to credibility ratings *(H3a)*, the analysis of variance revealed a significant effect of the author's age, in the group of older participants (F (1, 29) = 14.920, p = .001, η_p^2 = .340) as well as in the group of younger participants (F (1, 29) = 8.696, p = .006, η_p^2 = .231). However, in contrast to our hypothesis, mean values (see table 2) show that older authors were generally perceived as more credible in both age groups. The effect of author's age on credibility rating was significant for the whole sample, F (1, 59) = 23.041, p = .000, η_p^2 = .281. For the number of clicks (*H3b*), no significant effects emerged.

Table 2: Descriptive statistics for the effect of age on credibility score

Sample		M	SD	N
Age 22-38	Cred., Young Authors	129.43	15.94	30
	Cred., Old Authors	133.63	17.16	30
Age 39-47	Cred., Young Authors	133.10	16.34	30
	Cred., Old Authors	141.13	18.25	30
Total Sample	Cred., Young Authors	131.27	16.11	60
	Cred., Old Authors	137.38	17.96	60

Discussion

Against the background of the rise of Web 2.0 formats in which a lot of content is produced by laypersons instead of experts, we aimed to answer the question how online users perceive credibility and which factors determine their selection of online science information. For this purpose, the present study investigated the effect of expertise (as self-reports and community ratings) and age of weblog authors.

Our analysis showed that self-reported expertise has a strong influence on the perception of credibility: As hypothesized in *H1*, the participants preferred texts of

authors who had a profession with a close connection to the topic, e.g. psychologists or media scholars. Furthermore, their texts were chosen more frequently for further reading. These results are in line with studies from (offline) persuasion research (e.g. Wilson & Sherell, 1993) and dual-models of information processing (Petty & Cacioppo, 1986; Chaiken, 1987) in which expertise of the source is one important factor. From our findings, we can conclude that expertise as heuristically valuable information is already relevant in the earlier stage of information selection: Following the heuristic that "experts are usually correct", online users assess the credibility of the author and the estimated quality of the text before choosing an article. While Sundar et al. (2007) showed that this is true for newspaper sources, the present study indicates that expertise cues are also relevant if the message sources are persons. Therefore, it seems that online users prefer declared experts to "normal" people (who may be personally concerned with regard to the topic) even in websites that are dedicated to user-generated-content.

However, statements of other users on the expertise of the authors, expressed by community ratings *(H2)*, did not have a significant effect on credibility rating and information selection. Obviously, the display of rating stars did not produce a bandwagon effect, as it was found for online articles (Knobloch-Westerwick et al., 2005) and for the credibility of Ebay sellers (Resnick et al., 2006). This is all the more astonishing as previous research on social networking sites (Walther et al., 2009) has shown that information given by other people is seen as more important than self-descriptions. The lack of impact might be due to the fact that it was not clear to the participants what exactly the ratings indicated and by whom (e.g. how many people) the evaluation had been given. The cue concerning self-reported expertise (profession of the author) has obviously been more important because the participants trusted in the correctness of these self-reports: It is also possible that they perceived it as an objective fact (possibly verified by the blog owner) rather than a subjective assessment made by the author. Furthermore, the costs and consequences of the decision to choose an article or not are smaller than e.g. when deciding to buy a product on Ebay. As a result, the considerations may be less careful, which would lead to a decreased importance of community ratings. However, if only the authors with a low self-reported expertise were taken into account, community ratings produced a marginally significant effect: Texts of authors with a high rating were selected more often than texts with a low rating. This suggests that community rating does not matter when the level of self-reported expertise is high. But if the level of expertise is low, ratings seem to make a difference in that people with a better rating are selected more often.

Our analysis for *H3* showed that the age of weblog authors has a significant influence on credibility ratings and that older authors are generally perceived as more credible. This is in contrast to our assumptions that users prefer sources of similar age, based on social comparison theory (Festinger, 1954; Mussweiler, 2001). While Knobloch-Westerwick and Hastall (2006) found a social comparison effect on the selection of news articles according to the age of protagonists, there seems to be no such effect for blog authors. An explanation could be that users of a science weblog are mainly concentrating on the quality of information (which can e.g. be deducted from a profession with a close connection to the topic, a high rating and maybe higher age due to more professional experience) rather than seeking personal information on the author. Possibly, other websites, such as social networks in which detailed personal information and pictures are included, are more likely to foster social comparison processes (see Haferkamp & Krämer, 2010). The effect that older authors are seen as more credible may be explained by the topic of "violent media effects", in which experiences with child-rearing are helpful. For other topics (e.g. pop music or Internet technology), the relationship between age and source credibility may be different.

In summary, self-reported expertise of the author emerges as a strong cue for the perception of online science information, whereas there is only partial evidence for the importance of community ratings and age. In line with Sundar et al. (2007), these results demonstrate that the "information scent" of articles is not restricted to its content or formal features (position or layout): Information on the author, especially expertise, must also be taken into account.

In order to achieve further insights into these processes, future research should investigate the effects of sources in combination with other variables, such as different message types and different levels of motivation of information seeking. In the present study, texts have been written in a neutral style, which might have created a slightly artificial situation that differs from the normal situation in the blogosphere. If variations of content are included, the analysis of user behavior may show the interdependencies between several important factors of information selection.

Acknowledgments

The present study was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) in the Special Priority Program "Science and the General Public" (Kr 2240/2).

References

Berlo, D., Lemert, J., & Mertz, R. (1969). Dimensions for evaluating the acceptability of message sources. *Public Opinion Quarterly, 33*, 563-675.

Brewer, M. B. (1991). The social self: On being the same and different at the same time. *Personality and Social Psychology Bulletin, 17*, 475-482.

Cacioppo, J. T. & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology, 42*, 116-131.

Chaiken, S. (1987). The Heuristic Model of Persuasion. In M. Zanna, J. Olson, & C. Herman (Eds.), *Social*

Influence: The Ontario Symposium (Vol. 5, pp. 3-39). Hillsdale, NJ: Lawrence Erlbaum Associates.

Eachus, P. & Cassidy, S. (2006). Development of the Web User Self-efficacy Scale, (WUSE). *Issues in Informing Science and Information Technology, 3*, 199-209.

Eppler, M. J. & Mengis, J. (2004). The concept of information overload: A review of literature from organization science, accounting, marketing, MIS, and related disciplines. *The Information Society, 20*, 325-344.

Festinger, L. (1954). A theory of social comparison processes. *Human Relations, 7*, 117-140.

Fiske, S. T. & Taylor, S. E. (1991). *Social cognition* (2nd Ed.). New York: McGraw-Hill.

Gierl, H., Stich, A., & Strohmayr, M. (1997). Einfluss der Glaubwürdigkeit einer Informationsquelle auf die Glaubwürdigkeit der Information. [The influence of source credibility on information credibility.] *Marketing Zeitschrift für Forschung und Praxis, 19*, 27-31.

Haferkamp, N. & Krämer, N.C. (2010). *Social comparison 2.0. Examining the effects of online profiles on social networking sites*. Paper presented at the annual meeting of the International Communication Association, Singapore.

Hovland, C. I., Lumsdaine, F. D., & Sheffield, F. D. (1949). *Experiments on Mass Communication*. Princeton, NJ: Princeton University Press.

Knobloch-Westerwick, S. & Hastall, M. (2006). Social Comparisons with News Personae: Selective Exposure to News Portrayals of Same-Sex and Same-Age Characters. *Communication Research, 33*, 262-284.

Knobloch-Westerwick, S., Sharma, N., Hansen, D. L., & Alter, S. (2005). Impact of Popularity Indications on Readers' Selective Exposure to Online News. *Journal of Broadcasting & Electronic Media, 49*, 296-313.

Knobloch, S., Zillmann, D., Gibson, R., & Karrh, J.A. (2002). Effects of Salient News Items on Information Acquisition and Issue Perception. *Zeitschrift für Medienpsychologie*, 14 (N.F. 2) 1, 14-22.

Metzger, M. J. (2007). Making sense of credibility on the Web: Models for evaluating online information and recommendations for future research. *Journal of the American Society for Information Science and Technology, 58*, 2078-2091.

Metzger, M., Flanagin, A., Eyal, K., Lemus, D., & McCann, R. (2003). Credibility for the 21st century: Integrating perspectives on source, message, and media credibility in the contemporary media environment. *Communication Yearbook, 27*, 293-335.

Mussweiler, T. (2001). Focus of comparison as a determinant of assimilation versus contrast in social comparison. *Personality and Social Psychology Bulletin, 27,* 38-47.

O´Reilly, T. (2005). What Is Web 2.0 - Design Patterns and Business Models for the Next Generation of Software. Available: http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html?page=5.

Petty, R. E. & Cacioppo, J. T. (1986). *Communication and Persuasion. Central and Peripheral Routes to Attitude Change*. New York: Springer Verlag.

Pirolli, P. & Card, S. K. (1999). Information Foraging. *Psychological Review, 106*, 643-675.

Resnick, P., Zeckhauser, R., Swanson, J., & Lockwood, K. (2006). The value of reputation on eBay: A controlled experiment. *Experimental Economics, 9*, 79-101.

Schamber, L. & Bateman, J. (1996). User criteria in relevance evaluation: Toward development of a measurement scale. *Journal of the American Society for Information Science, 33*, 218-225.

Schmidt, J. (2007). Blogging Practices: An analytical framework. *Journal of Computer-Mediated Communication*, [On-line serial], *12*. Available: http://jcmc.indiana.edu/vol12/issue4/schmidt.html

Stavrositu, C. & Sundar, S. S. (2008). If Internet credibility is so iffy, why the heavy use? *Cyberpsychology & Behavior, 11*, 65-68.

Sundar, S. S., Knobloch-Westerwick, S., & Hastall, M. (2007). News cues: Do indicators of newsworthiness by newsbots affect our perception of news stories? A cross-cultural study in Germany, the Netherlands, and the U.S. *Journal of the American Society of Information Science and Technology, 58*, 366-378.

Sundar, S. S. & Nass, C. (2001). Conceptualizing sources in online news. *Journal of Communication, 51*, 52-72.

Tombros, A., Ruthven, I., & Jose, J. M. (2005). How users assess Web pages for information seeking. *Journal of the American Society for Information Science and Technology, 56*, 327-344.

Van Der Heide, B. (2008, May). *Persuasion on the 'net: A synthetic propositional framework.* Paper presented at the annual meeting of the International Communication Association in Montreal, Canada.

Walther, J. B. (1992). Interpersonal effects in computer-mediated interaction: A relational perspective. *Communication Research, 19*, 52-90.

Walther, J. B. (1996). Computer-mediated communication: Impersonal, interpersonal, and hyperpersonal interaction. *Communication Research, 23*, 3-43.

Walther, J. B., Van Der Heide, B., Hamel, L., & Shulman, H. (2009). Self-generated versus other-generated statements and impressions in computer-mediated communication: A test of warranting theory using Facebook. *Communication Research, 36,* 229-253.

Walther, J. B., Wang, Z., & Loh, T. (2004). The effect of top-level domains and advertisements on health web-site credibility. *Journal of Medical Internet Research, 6*.

Wilson, E. J. & Sherrel, D. L. (1993). Source effects in communication and persuasion research: A meta-analysis of effect size. *Journal of the Academy of Marketing Science, 21*, 101-112.

Zillmann, D., Chen, L., Knobloch, S., & Callison, C. (2004). Effects of lead framing on selective exposure to Internet news reports. *Communication Research*, *31,* 58-81.

The Impact of Syntax on the Interpretation and Graphical Depiction of Underspecified Propositions

Aaron Kalb, Dave Barker-Plummer[1]**,**
Deonne Castaneda, Christopher Potts[2]
Stanford University, Stanford, CA 94305 USA
{kalb,dbp,deonne,cgpotts}@stanford.edu

Richard Cox
School of Informatics, University of Sussex
Falmer, E. Sussex, BN1 9QJ, UK
richc@sussex.ac.uk

Robert Dale
Center for Language Technology, Macquarie University
Sydney, NSW 2109, Australia
rdale@science.mq.edu.au

Abstract

Different representational systems permit differing degrees and forms of ambiguity and underspecification in the content they represent. Independently of this observation, a notable feature of natural language as a representational system is that it allows the same content to be expressed in different ways. In this paper, we examine the interaction of these two observations; in particular, we explore a number of linguistic forms involving underspecified content, and look at how subjects express the content of these linguistic forms both in logic and in diagrams. Our analysis demonstrates that variations in syntactic realization of the same semantic content lead to different interpretations of that content.

Keywords: logic; natural language; syntactic structure; diagrams; representations; negation.

Introduction

This paper takes as its starting point two widely-made observations. First, different representational systems permit different abstractions, and consequently, they permit underspecification on different dimensions. In particular, natural language (NL) and first-order logic (FOL) are two representational systems that permit underspecification of aspects of meaning that must be made explicit in diagrammatic representations. Direction is a case in point: consider the natural language statement *The house is adjacent to the park*; neither this sentence, nor a typical FOL rendering such as $\mathsf{AdjacentTo}(\mathsf{house}, \mathsf{park})$, specifies the direction of adjacency, but a picture or diagrammatic rendering of the sentence must make this explicit. We will say that the NL representation of the state of affairs is **underspecified with respect to** the diagrammatic representation. This makes it clear that underspecification as defined here is a relational notion; however, for convenience in the remainder of this paper we will simply refer to representations as being **underspecified** when the relatum is obvious from the context.

[1]Center for the Study of Language and Information.
[2]Department of Linguistics.

The second observation we take as a starting point is that natural language affords multiple ways of realizing the same semantic content. This is often exemplified by reference to the fact that active and passive sentences, such as *Fred wrote the book* and *The book was written by Fred*, describe the same state of affairs. There may be contextual, or pragmatic, reasons for choosing one realization over the other, as commonly discussed under the heading of **information packaging** (Vallduvi, 1992); but the common view is that the semantics of the two sentences, *in terms of propositional content*, is the same.

We are interested in how these two phenomena interact. Our interest is motivated by an effect found in an earlier study (Cox, Dale, Etchemendy, & Barker-Plummer, 2008), in which the specificity of participants' responses to NL sentences containing negation differed markedly between their FOL translations and their diagrammatic interpretations. Specifically, it was found that in their FOL translations of the sentence ***d*** *is not a small dodecahedron*, participants overwhelmingly treated the predicates *small* and *dodecahedron* symmetrically, whereas their diagrams of the sentence tended to make **d** a dodecahedron that isn't small, rather than a small shape other than a dodecahedron. However, contextual confounds made the source of this effect hard to establish.

In this paper, we report on an experiment which sought to elucidate the effects of different possible factors on this phenomenon. In particular, we ask: if we have a number of natural language forms that express the same *underspecified* semantic content, what happens when subjects are asked to draw diagrams that require them to be more explicit? If the NL sentences truly express the same meaning, then we might expect to see similar distributions of the possible diagrammatic renderings, regardless of the NL surface form used. Alternatively, syntax or semantics may make some diagrammatic renderings more salient or available than others, This paper sets out to determine which of these alternatives hold.

Hypotheses

We explore two hypotheses in particular.

Hypothesis 1: When asked to translate from one representation into another that permits underspecification to be maintained, then in the absence of any contextual factors that encourage a more specific reading in the target representation, subjects will maintain the underspecification.

Hypothesis 2: When asked to translate from one representation into another that requires underspecification to be made specific, and there are a limited number of ways of doing this, we expect to see similar distributions across these solutions irrespective of superficial variations in the way the content is expressed in the source representation.

To test Hypothesis 1, we ask subjects to translate from NL to FOL. We make use of syntactic variations that represent the same semantic content; for example, the three sentences below all are expressions of the FOL statement $\neg(\mathsf{Striped}(\mathsf{q}) \wedge \mathsf{Circ}(\mathsf{q}))$:

(1)	**q** is not a striped circle	PREMOD
(2)	**q** is not a circle with stripes	POSTMOD
(3)	**q** is not striped and circular	COORD

The first two sentences are syntactically and semantically unambiguous. It is possible, with appropriate contextual cues, to encourage a more specific reading than the **wide-scoped** FOL statement above. For example, in spoken form, emphasis on either *striped* or *circle*, as in ***q** is not a striped circle* or ***q** is not a striped circle*, may encourage a **narrow-scoped** reading, corresponding to $\neg\mathsf{Striped}(\mathsf{q}) \wedge \mathsf{Circ}(\mathsf{q})$ and $\mathsf{Striped}(\mathsf{q}) \wedge \neg\mathsf{Circ}(\mathsf{q})$, respectively. However, in the absence of any such cues, Hypothesis 1 predicts that subjects will provide the wide-scoped reading.[3]

The third sentence is syntactically ambiguous (see Figure 1). Each parse corresponds to a different semantics, one of these being the wide-scoped reading, and the other the narrow-scoped-left reading. We would expect to find a distribution across these two readings in the FOL renderings.

To test Hypothesis 2, we ask subjects to translate from NL into diagrammatic realizations. We focus our analysis on those subjects who maintained underspecification in our test of Hypothesis 1, i.e., we leave aside any subjects who produce a narrow-scope reading for COORD sentences. We set up the diagram task conventions in such a way that there are only a limited number of possible ways of making the underspecified content specific. In particular, the wide-scope FOL above can be realized by three classes of diagrams:

[3]There is an extensive literature on scope ambiguity and its effects on human sentence processing (see, for example, (Kurtzman & MacDonald, 1993)) and on how discourse factors and lexical frequency impact on the processing of syntactic ambiguities (see, for example, (Trueswell, 1996; Spivey & Tanenhaus, 1998)); however, the focus of the former tends to be on quantifier scoping, and the latter is not obviously relevant to the kind of data we explore here. We are not aware of any existing work that looks at the processing of negated conjunctions of verbal complements, as explored here.

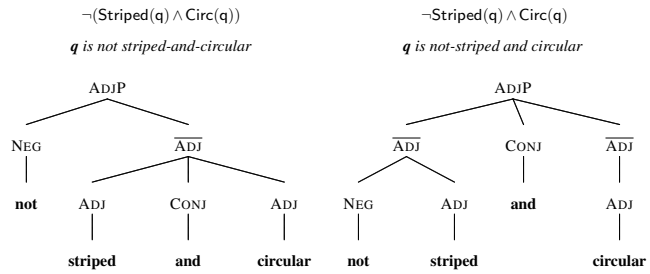


Figure 1: Two parse trees for ***q** is not striped and circular*

1. diagrams in which **q** is a circle that isn't striped (i.e. only the predicate Striped is realized-as-negated);
2. diagrams in which **q** is a striped object other than a circle (i.e. the predicate Circ alone is realized-as-negated); and
3. diagrams in which **q** is neither circular nor striped (i.e. both predicates are realized-as-negated).

Each of these realizations commits to some information left unspecified in the FOL sentence.

Our hypothesis predicts that the distribution of different diagrammatic realizations should be roughly similar irrespective of which surface NL form is being translated.

Methodology

The Subjects

Forty-one students enrolled in an introductory logic class at Stanford University took part. The experiment was conducted in the final weeks of the term. All of the background material necessary to complete the experimental task was presented within the first two weeks of the term. A key aim of the course is to teach the ability to distinguish the propositional content of sentences from their implicatures. The subject pool had therefore been primed to consider different possible interpretations of sentences and whether those interpretations depend on factors external to the sentence, such as common knowledge. Further, the students knowledge of FOL allows us to test for an unambiguous reading of sentences with multiple interpretations.

The class used (Barwise, Etchemendy, Allwein, Barker-Plummer, & Liu, 1999) as the textbook, and used the Tarski's World computer program for teaching the semantics of FOL. Tarski's World presents a system similar to the diagrammatic representation used in this experiment. In general, the materials were designed to parallel the structure of materials in the course, both in terms of the diagrammatic representations that were used, and the names used to refer to the distinct activities within the experiment.

Materials Administered

Subjects were given workbooks consisting of: (a) a page of study information; (b) a sheet consisting of 18 declarative natural language sentences to be translated; (c) a page of instructions; and (d) a half-page description with illustrative ex-

Table 1: Experimental Sentences in PREMOD Formulation

d is not a large dotted object	**k** is not a small circle
h is not a small dotted object	**l** is not a striped triangle
p is not a small striped object	**n** is not a dotted triangle
b is not a large triangle	**q** is not a striped circle
c is not a large circle	

amples of the diagrammatic representation and the first-order language to be used in the task. Following these were pages describing four activities to be completed. Three of these were presented as 'You Try It's (YTIs), and are described further below; these would be familiar to the subjects from (Barwise et al., 1999) as activities for becoming familiar with a concept. The final page of the workbook contained a more complex exercise in translation and realization, the contents of which are not discussed here.

Subjects were asked to complete all four activities with no time-limit and with no supervision. Only data from the second and third YTIs—translating the sentences into FOL and drawing realizations, respectively—are analyzed as part of this experiment; the remaining activities were included in order to format and embed the experimental tasks within an exercise form that the participants were familiar with from (Barwise et al., 1999). More importantly, they were designed to encourage subjects to submit spontaneous, naturalistic realizations.

The Tasks

As noted, subjects were given 18 natural language sentences in English. Nine of these were **negated logical conjunctions**, expressed in a form determined by the different conditions described below. The remainder of the sentences were fillers, also varying by condition, such as ***m*** *is a triangle that's not dotted*. The filler sentences all use the same vocabulary as the experimental sentence; some involve negation and others do not. All have unique readings.

The complete set of sentences (shown for condition PREMOD in Table 1) was counterbalanced such that each predicate is mentioned an equal number of times. In particular, of the nine experimental sentences, three mention size and pattern, three mention pattern and shape, and three mention size and shape; and each of the words *large*, *small*, *circle*/*circular*, *triangle*/*triangular*, *striped*/*stripes*, and *dotted*/*dots* are mentioned three times.

Subjects were asked to perform two tasks. The first task was to translate each sentence into a formal language of FOL, as discussed above. For the negated logical conjunctions, three FOL readings were possible, which we refer to as wide-scope, narrow-scope-left, and narrow-scope-right; for the example sentences introduced at the outset of the paper, these FOL readings are $\neg(\mathsf{Striped}(\mathsf{q}) \wedge \mathsf{Circ}(\mathsf{q}))$, $\neg\mathsf{Striped}(\mathsf{q}) \wedge \mathsf{Circ}(\mathsf{q})$, and $\mathsf{Striped}(\mathsf{q}) \wedge \neg\mathsf{Circ}(\mathsf{q})$ respectively.

The second task was to draw, for each sentence, a diagram of a situation making the sentence true. We call these **diagrammatic realizations** of the sentences. We devised a highly constrained diagrammatic representation system in which objects have exactly three properties: shape, size, and pattern, with each of these properties having only two possible values (circle/triangle, small/large, striped/dotted). The students were asked to draw such objects in prepared spaces.[4]

Since the sentences have different readings, they may be realized in different ways, but the wide-scoped reading itself has multiple equally valid realizations. In the example above, **q** can be either a dotted circle, a striped triangle, or a dotted triangle. For such **multiply realizable** readings, each of the three possible realizations are equally valid, but the response requirement of a single diagram forces subjects to choose one.

The Conditions

The sentences of interest share the common property that they can be read as expressing the negation of a conjunction. Sentences (1)–(3), introduced earlier, are examples of such sentences. Each corresponds to one of three different conditions.

In a between-groups design, subjects were randomly allocated to one of these three conditions, named PREMOD (pre-nominal modifier, $N = 14$), POSTMOD (post-nominal modifier, $N = 11$) and COORD (coordination, $N = 16$). Subjects in each condition were presented with negated conjunctions expressed in one of these three forms.[5] Within each condition, subjects were randomly assigned to one of three random sentence orderings in order to control for possible priming effects within the stimulus sentences.

Data Collection and Encoding

Each worksheet was encoded independently by two coders. The FOL sentences and features of the diagrams[6] were recorded for each subject along with the condition that they were in. We also encoded which of the random sentence-orderings the subjects saw, but this information was not used for this study, as no systematic ordering effects were observed. Where they differed, the independent codings were

[4] These included guide lines for distinguishing large objects from small ones.

[5] The POSTMOD condition included an even mix of sentences with pattern expressed as a prepositional phrase (as above) and as a relative clause, as in ***q*** *is not a circle that's striped* In the COORD condition, the order of the predicates was varied, with some sentences of the form ***q*** *is not striped and circular* and others of the form ***q*** *is not circular and striped*. For the three sentences in the PREMOD and POSTMOD conditions which mention size and pattern but not shape, the word *object* is used as the noun (see the sentences describing **d**, **h**, and **p** in Table 1). Finally, in the POSTMOD condition, pattern is expressed post-nominally, but size is expressed as a pre-nominal adjective, as in the PREMOD condition, because the formulation ***k*** *is not a circle with small* is ungrammatical.

[6] These were encoded as large/small, striped/dotted, circle/triangle or as 'unclear' (if, for instance, a medium sized object were drawn), 'unspecified' (if, for instance, a shape were drawn with neither stripes nor dots), or 'other' (if, for instance, a square were drawn instead of a triangle or a circle).

Table 2: Readings: FOL scope by negation sentence condition

	Scope		
Condition	Wide	Left	Right
PreMod ($N = 126$)	100%	0%	0%
PostMod ($N = 99$)	99%	1%	0%
Coord ($N = 132$)	42%	50%	0%

Table 3: Realizations of sentences with 'negatable heads'

	Realized As Negated		
Condition	Both	Head Only	Mod. Only
PreMod ($N = 68$)	34%	25%	41%
PostMod ($N = 38$)	21%	21%	58%

Table 4: Realizations of modifier-only (headless) sentences

	Realized As Negated		
Condition	Both	First Only	Second Only
PreMod ($N = 32$)	47%	18%	35%
PostMod ($N = 18$)	50%	17%	33%
Coord ($N = 52$)	71%	14%	15%

arbitrated by a third coder, who resolved disagreements.[7]

Results

Translations into FOL

We can measure the accuracy with which subjects completed the task of translating into FOL by considering their success in expressing an expected reading of each sentence. In the case of filler sentences, there is a unique expected FOL sentence, while for experimental sentences there are three possible readings for each. 78.6% of translations were expected. 92% of unexpected sentences were produced by four of the participants. Table 2 shows the proportions of each reading obtained for the experimental sentences.

Participants in the PreMod and PostMod conditions almost universally wrote wide-scoped readings: only one sentence out of 225 was translated with a narrow-scoped reading. Subjects in the Coord condition displayed markedly different behavior. Table 2 gives the breakdown by condition, but the results are interesting when broken down by subject. 43.7% of subjects produced a wide-scoped reading for all (25%) or all but one (18.7%) of the nine sentences. 50.0% of these subjects always produced a narrow-scoped reading, and all of these were narrow-scoped-left.[8] The subjects who produced wide-scoped reading for all but one of the sentences were the only subjects to produced a mix of wide- and narrow-scoped readings. In short, participants were systematic, with 50.0% always translating with narrow-scope-left, and 43.7% (almost) always translating with wide scope.

Thus, the results are consistent with Hypothesis 1: When subjects are asked to translate from one representation (NL) into another (FOL) that permits underspecification to be maintained, then the underspecification is indeed maintained. This is almost universally the case in the PreMod and PostMod conditions, and also the case in around half of the Coord condition sentences, consistent with the fact that the latter are syntactically ambiguous, and one of the two parses is consistent with an underspecified reading.

Diagrammatic Realizations

Tables 3 and 4 show the results of encoding the realizations that students produced in the diagramming task. We recorded the predicates in the sentence that were **realized as negated** in each diagram. If the sentence is ***q*** *is not a striped circle* and the drawing was of a dotted circle, the predicate Striped is realized as negated and Circle is not.

Table 3 shows the results for sentences which are expressed syntactically with a head and modifier.[9] This pattern only arises in conditions PreMod (**q** *is a striped circle*), and PostMod (**q** *is a circle with stripes*) in which the shape predicate is the head and the other is the modifier. The columns record whether both predicates, just the head predicate, or just the modifier predicate are realized as negated.

In the PreMod and PostMod conditions, the sentences which do not mention shape (such as ***d*** *is not a large, dotted object*) contain a head predicate (*object*) which cannot be realized as negated. In Coord, all of the sentences lack a 'negatable' head (**q** *is striped and circular*). We will call such sentences **headless**, although this is not literally true in the PreMod and PostMod conditions. Table 4 give the results for these sentences, with the columns indicating which modifier appears lexically first in the sentence.

Correspondence between Readings and Realizations Recall that approximately half of the subjects in the Coord condition wrote FOL sentences corresponding to the narrow-scoped-left reading of the sentences. These subjects universally drew diagrams consistent with this reading.

This suggests a strong alignment of readings with realizations, perhaps because the subjects referred to their FOL while producing the realizations, or because they arrived at the same mental representation on reading the sentence in preparation for translation into FOL and again in preparation for drawing their realization.

Similarly, subjects with wide-scoped readings in all three conditions drew diagrams consistent with this reading, although there are fewer possible incorrect realizations for these readings.[10] While variation in the narrow-scoped case

[7] The workbooks for all three conditions, an exemplar subject response, and the complete encodings can be downloaded from `http://openproof.stanford.edu/readingsandrealizations`.

[8] One subject (6.3%) wrote on the packet that the sentences were ambiguous, and submitted both a narrow-scoped left translation and a wide-scoped translation for each. Data for this subject was discarded.

[9] For convenience we talk from here on of 'sentences with heads and modifiers', although of course this refers to the heads and modifiers used in the descriptions of the objects.

[10] Subjects could only incorrectly realize **q** *is not a striped circle*

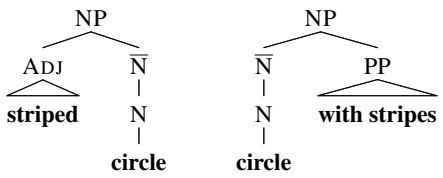


Figure 2: Asymmetric Parse Trees for Head-Modifier Constructions

would represent error (or at least inconsistency), variation in the wide-scoped case is expected.

Realizations of Wide-Scoped Readings

Subjects who obtained wide-scoped readings of the sentences have a choice of the realization that they can draw while remaining consistent with their reading. We focus on these subjects and that variability for the remainder of the analysis. Hence, we consider only the realizations for sentences with wide-scoped readings ($N = 100$ for PREMOD, $N = 63$ for POSTMOD, and $N = 60$ for COORD).

Heads vs. Modifiers

We first discuss the sentences that have heads that could have been realized as negated. These are just those sentences in conditions PREMOD and POSTMOD that mention the shape properties *circle* and *triangle*. For these sentences, the predicate which is expressed by a modifier is significantly more likely to be realized-as-negated than the predicate expressed by the head: In the PREMOD condition, the modifier is realized as negated 75% of the time while the head is realized as negated 59% of the time ($\chi^2 = 6.752$, $p < .01$). In the POSTMOD condition the modifier is realized as negated 79% of the time while the head is realized as negated 42% of the time ($\chi^2 = 10.794$, $p < .01$). Note that in some realizations both predicates are realized as negated.

This result mirrors that reported in (Cox et al., 2008). Our sentences analogous to ***d*** *is not a small dodecahedron* are those in the PREMOD and POSTMOD conditions which do not mention pattern. In our subjects' diagrams for these sentences, size was realized as negated significantly more often than shape. In 53.4% of the realizations of the 53 readings of the three sentences of the form ***b*** *is not a large triangle,* the size alone takes the negation. By contrast, participants negated just the shape or negated both predicates in only 22.3% and 24.3% of the realizations, respectively.

Modifier Choice

We now turn our attention to those sentences that only express properties via modifiers. All sentences in the COORD condition belong in this category, as do the sentences from the other two conditions which do not mention shape (e.g. ***d*** *is not a large object with dots*).

In 57% of the 105 realizations of these sentences, both predicates are realized as negated. In the 156 realizations of the other sentences (those with heads), both predicates are realized as negated only 35% of the time. This is a highly significant difference ($\chi^2 = 14.656, p < .001$).

It seems that when both predicates are expressed as modifiers, subjects are likely to realize them both as negated (perhaps because they must negate at least one and there is no obvious means of deciding which), while if one is expressed as a head, its identity is likely to be preserved.

It is worth noting, as well, that the tendency to realize both predicates as negated is most pronounced for sentences in the COORD formulation: both predicates are realized as negated in 71% of the realizations in this condition ($N = 52$), compared with 47% and 50% of the realizations of headless PREMOD and POSTMOD sentences, respectively. This may be because, in wide-scoped parses of a COORD formulation, the conjunction attaches to both arguments symmetrically (see Figure 1, left), so there are no structural differences whatsoever between the expressions of the two predicates.

Discussion

The results just discussed suggest that Hypothesis 2 does not hold. When subjects are asked to translate from one representation (NL) into another (a diagram) that requires underspecification to be made specific, the way in which this is done depends on the syntactic form used in the source representation. In particular, if a property is expressed via a syntactic nominal head, it is less likely to be realized as negated than when it is expressed via as a modifier.

There are other possible explanations for the observed behaviour, which we consider briefly below.

Ontological Primacy: *Perhaps shape as a concept is less readily negate-able than the other predicates.*

Since the only heads occurring in our sentences are the shape nouns *circle* and *triangle*, perhaps the phenomenon is due to some ontological primacy accorded to shape, but not to the other predicates. In our materials, shape is primarily seen as a *type* of object, whereas the other predicates are *attributes* of objects. If shape were protected because of its ontological status, rather than because of the way it is expressed, we would see these same results in conditions PREMOD and POSTMOD, since the only heads appearing in our sentences are the shape predicates. However, if it were the ontological status of shape that were protected, we would expect it to be protected in the COORD condition as well, even though in that condition shape is expressed as a modifier.

Among sentences in the COORD condition, however, shape is realized as negated 77.0% of the time ($N = 39$) (Figure 3)—just as much (more, in fact) than the other predicates. This strongly suggests that shape, as a concept, is not protected.

Surface Proximity: *Perhaps participants simply tend to negate the predicate closest to the word* not.

In sentences such as **q** *is not a striped circle*, *striped* is closer to *not* and perhaps this accounts for the preference for realizing this predicate as negated.

by drawing a striped circle.

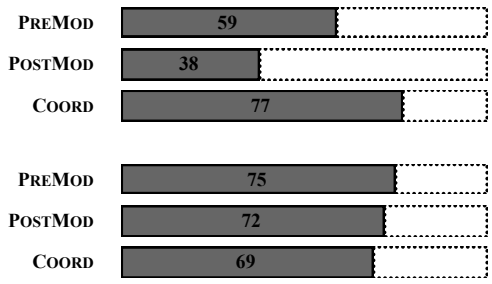


Figure 3: Among realizations of sentences which mention shape, % which negate shape (TOP) vs. % which negate the other predicate—size or pattern (BOTTOM)

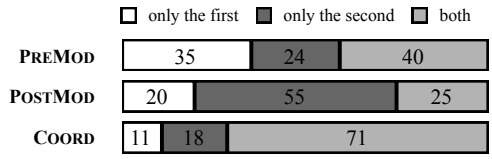


Figure 4: % of realizations of sentences which mention pattern and shape in which — predicate is negated.

PREMOD	***q*** *is not a striped circle*	$N = 33$
POSTMOD	***q*** *is not a circle with stripes*	$N = 24$
COORD	***q*** *is not striped and circular*	$N = 19$

Looking at the readings of the three sentences which mention the predicates pattern and shape (Figure 4), we see that students are somewhat (though not significantly) more likely to negate just pattern than just shape in the PREMOD condition, where the pattern predicate occurs closest to the word *not.* However, when phrased so that the pattern predicate occurs farthest (in the POSTMOD condition with sentences like ***q*** *is not a circle with stripes*), we find that pattern continues to take the negation—this time, 2.8 times as often as just shape (more, in fact, than when it occurs in closer proximity to *not*). The difference in likelihood to realize-as-negated just the first vs. just the second predicate across the PREMOD and POSTMOD conditions is significant ($\chi^2 = 3.979, p < .05$). Moreover, we see no tendency whatsoever toward negating the closer predicate in the COORD condition, for any sentence.[11]

Conclusion and Future work

We set out to test two hypotheses, one of which suggested that subjects would maintain underspecification in their representations if this were possible, and a second which suggested that, if subjects had to translate into a representation that required more specificity than the source representation, then the results would be the same for semantically-equivalent source representations.

[11](Kroch, 1974) proposes 'a general surface ordering principle that fixes the initial scope order of the operator words in an English sentence according to their surface order'; however, in line with our findings, this claim is refuted by (Kurtzman & MacDonald, 1993).

The evidence from our experiment supports the first hypothesis. This allowed us to go on to test our second hypothesis, where the results turned out to be surprising: we demonstrated that the same semantic content, expressed in natural language in different ways, leads to different interpretations when subjects are asked to express that information in diagrams which require them to choose a more specific representation.

This is unexpected. Of course, it is not suprising that the particular form of an utterance has an impact on how that utterance is interpreted; but such variations are usually considered to be in the realm of pragmatics, and more concerned with connotation than with denotation. The results here, however, indicate that how something is expressed has an impact not only in terms of the pragmatic aspects of interpretation, but also in terms of the state of affairs in the world the utterance is taken to describe.

If we characterize shape via a noun, then it is less likely to be negated than if it is expressed via an adjective or other modifier. It would appear that it is how things are described, or how, in Langacker's terms (Langacker, 1991), they are **construed**, that governs our interpretation; not what they are.

Acknowledgements

This work was supported by funds from the office of the Provost at Stanford University. Emma Pease, Michael Murray, and Sommer Panage assisted with the experimental sessions and encoding the data. Students in Stanford's Fall 2009 Philosophy 150 course served as the subject pool.

References

Barwise, J., Etchemendy, J., Allwein, G., Barker-Plummer, D., & Liu, A. (1999). *Language, proof and logic*. CSLI Publications and University of Chicago Press.

Cox, R., Dale, R., Etchemendy, J., & Barker-Plummer, D. (2008). Graphical revelations: Comparing students' translation errors in graphics and logic. In *Diagrams 2008, fifth international conference on the theory and application of diagrams.* (Herrsching, Germany, September 2008)

Kroch, A. (1974). *The semantics of scope in English*. Unpublished doctoral dissertation, Massachusetts Institute of Technology.

Kurtzman, H., & MacDonald, M. (1993). Resolution of quantifier scope ambiguities. *Cognition*, *48*, 243-279.

Langacker, R. (1991). *The foundations of cognitive grammar*. Stanford University Press.

Spivey, M. J., & Tanenhaus, M. K. (1998). Syntactic ambiguity resolution in discourse: Modeling the effects of referential context and lexical frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(6), 1521–1543.

Trueswell, J. (1996). The role of lexical frequency in syntactic ambiguity resolution. *Journal of Memory and Language*, *35*, 566–585.

Vallduvi, E. (1992). *The informational component*. New York: Garland.

Development of Prototype Abstraction and Exemplar Memorization

Irina Baetu (irina.baetu@mail.mcgill.ca)
Department of Psychology, McGill University, 1205 Penfield Avenue
Montréal, QC H3A 1B1 Canada

Thomas R. Shultz (thomas.shultz@mcgill.ca)
Department of Psychology and School of Computer Science, McGill University, 1205 Penfield Avenue
Montréal, QC H3A 1B1 Canada

Abstract

We present a connectionist model of concept learning that integrates prototype and exemplar effects and reconciles apparently conflicting findings on the development of these effects. Using sibling-descendant cascade-correlation networks, we found that prototype effects were more prominent at the beginning of training and decreased with further training. In contrast, exemplar effects steadily increased with learning. Both kinds of effects were also influenced by category structure. Well-differentiated categories encouraged prototype abstraction while poorly structured categories promoted example memorization.

Keywords: exemplar memorization; prototype abstraction; category structure; neural networks; sibling-descendant cascade-correlation.

Introduction

One of the most fundamental abilities is learning to group things into categories. This faculty allows us to classify new examples and make useful predictions concerning their properties. Two general classes of models have been proposed to account for phenomena in concept learning: prototype and exemplar models. Prototype models claim that experience with items that belong to a given category results in the formation of a summary representation of all the items observed (Posner & Keele, 1968; Reed, 1972). Subsequent categorization of a new item is then based on a comparison between the prototype and the new item. Thus, the more similar a particular instance is to the abstracted prototype, the more likely it is to be classified as a category member (Homa & Cultice, 1984; Homa, Sterling, & Trepel, 1981). In contrast, exemplar models claim that all the observed items are remembered and that the categorization of a new item involves a comparison with items that are stored in memory (Hintzman, 1986).

There is ample evidence in favor of both prototype (Homa, et al., 1981; Posner & Keele, 1968) and exemplar models (Medin & Schaffer, 1978; Palmeri & Nosofsky, 2001), suggesting that both processes are used during category learning. What is more, the relative contribution of each mechanism to categorization might vary across development, as well as during training on a novel task. Early in development, categorization seems to be based on prototype representations while exemplar representations seem to increase with age (Hayes & Taplin, 1993; Mervis & Pani, 1980). There is also evidence that people are more likely to rely on prototypes at the beginning of a categorization task, and as training progresses they rely more on memorized exemplars (Horst, Oakes, & Madole, 2005; Minda & Smith, 2001; Smith & Minda, 1998). These studies are consistent with a shift from early prototype use to later exemplar memorization.

In addition to the amount of experience with a categorization task, category structure also influences which type of information is most used. Better-structured categories can be represented as separate clusters in psychological space, whereas poorly structured categories overlap with each other (Figure 1). Smith and Minda found that better structured categories encourage the early prototype formation, while poorly structured categories discourage it, and may even strongly disadvantage the use of prototypes (Smith & Minda, 1998). Their findings are consistent with a number of other studies (Homa, et al., 1981; Horst, et al., 2005; Reed, 1978).

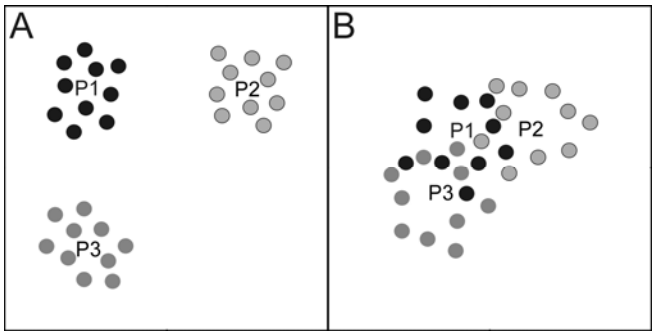


Figure 1: Hypothetical representations of three concepts. P1, P2 and P3 represent three prototypes and the circles represent examples of each concept. A: prototypes are relatively far from each other and examples are tightly clustered around their respective prototype, yielding concepts that are easy to distinguish. B: prototypes are close to each other and examples are more widely dispersed around their respective prototype, resulting in overlapping concepts that are difficult to distinguish.

The aim of this paper is to present a unified model able to simulate prototype and exemplar processes during concept learning. This unified model captures prototype and

exemplar effects with the same mechanism, as opposed to implementing two separate processes. We intend to demonstrate that it is possible for a unified mechanism to capture prototype and exemplar processes to different degrees depending on category structure and amount of training. We present here simulations with sibling-descendant cascade-correlation (SDCC) networks (Baluja & Fahlman, 1994), which offer several demonstrated advantages including automatic network construction, rapid and strong learning, and psychological and neurological plausibility (Shultz, 2003, 2006; Shultz, Mysore, & Quartz, 2007; Shultz, Thivierge, & Laurin, 2008). At the start, SDCC networks are composed of only input and output units. During training, examples were presented to the networks as specific activation patterns in the input layer. In encoder fashion, the networks gradually learned to reproduce this pattern on the output layer by changing the strength of the connections between the units and by recruiting and organizing new hidden units as needed.

In such networks, a relatively small number of units can store a large number of representations, with each representation being a specific pattern of activation across the units. These representations are relatively distributed, as opposed to being localized in single units. Because of its distributed nature, a network is likely to represent similar items as similar patterns of activations on the hidden units. The connection weights between the units reflect all trained items; thus, they represent something similar to a prototype, or an average of the trained concepts. Even if the networks are never presented with the category prototype, they are likely to falsely recognize it because it is so similar to many of the trained items. In addition, because the networks retain some specific information about the trained items, they show a familiarity effect when presented with old items, which is typical of exemplar models (Shultz, *et al.*, 2008).

The networks exhibit a prototype effect if they perform better when presented with examples that are similar to the hypothetical prototype (typical examples) than when they are presented with examples that are less similar to the prototype (atypical examples). We also tested whether the networks memorized some of the features of the trained examples. If our networks become more familiar with the trained examples and perform better when presented with old rather than new examples, regardless of distance from the prototype, then they reveal an exemplar effect.

We studied the impact of category structure and amount of training on prototype and exemplar effects. We manipulated category structure by changing the similarity between the prototypes of the trained categories and the similarity between each example and its prototype. Better-structured categories have more dissimilar prototypes and examples that are more similar to the prototype of their category (in other words, examples that are more tightly clustered around their prototype). To study the impact of training experience, networks were presented with varying numbers of training trials.

Method

As in past work (Shultz, *et al.*, 2008), we trained SDCC networks in encoder mode. Encoder networks learn to encode the input signal onto the hidden units, and then decode that hidden unit signal back onto the output units. Because error is computed as the sum-squared difference between input and output activations, this can be construed as self-supervised learning, without an externally-provided category name as target output. This type of learning occurs when people are not given information about category membership; hence, they can freely create concepts based on their observation of the examples (Homa & Cultice, 1984). In contrast, learning with category labels is much simpler and quicker. In typical encoder fashion, there were no input-output connections in our networks because such connections would have made the learning too simple.

Also as in Shultz *et al.* (2008), we trained the networks with examples belonging to four concepts. Each example varied on ten binary dimensions. A prototype was first constructed by randomly assigning values of 0.5 or -0.5 to each dimension. We refer to it as the prototype of the loner concept because it was relatively isolated from the other three concepts. Another 10-dimensional vector orthogonal to the first one was randomly selected (the normalized inner product between these two vectors was zero). From this orthogonal vector, three prototypes were created by randomly flipping one, two or four values. Flipping a value means reversing its sign. These three prototypes were much closer to each other in the 10-dimensional space than to the loner vector. We refer to them as the trio.

Nineteen examples were created from each prototype by flipping one or several values depending on the condition. Fifteen of these examples were used for training the networks, while four were used only during the test. Out of the fifteen trained examples, ten were closer to the prototype than the other five, i.e. they were created by flipping fewer values. We refer to the examples that were created through fewer flips as the close examples, and to the other ones as the far examples.

For each of the four concept prototypes, we manufactured examples by flipping 1, 2, 4, or 8 values of the prototype, randomly selected without replacement, depending on condition and subject to three additional constraints: (a) each example had a unique combination of features to flip, ensuring example uniqueness, (b) each feature was flipped in at least one example, and (c) no feature was flipped in every example. This last constraint ensured that no defining features were inadvertently created.

Out of the four examples that were used only during the test, two were close and two were far from the prototype. The networks were also tested on four of the trained examples, two that were randomly selected from the close examples, and the other two, from the far examples. Thus, testing consisted of presenting the networks with eight examples: two close trained examples, two far trained examples, two close test examples, and two far test examples. An exemplar effect is established if the networks

perform better on the trained examples than on the new test examples. Superior performance on the close examples versus the far ones demonstrates a prototype effect.

We manipulated the structure of the categories, which was determined by two factors. First, the number of flips that were applied to the vector orthogonal to the loner to create the trio was varied. Applying fewer flips means that the three concepts are closer to each other, while performing more flips means that the concepts are more distinct from one another. Second, we varied the number of flips applied to the loner and the trio to create examples. Fewer flips indicate that the examples are more tightly clustered around their prototype, while more flips imply a more dispersed distribution of the examples. These two manipulations affect the overall distinctiveness of the concepts. The concepts are more separate from one another with more prototype flips and fewer example flips.

Three levels of category structure were defined. The number of flips applied to the vector orthogonal to the loner to create the trio was 4 (Condition Easy), 2 (Condition Intermediate), or 1 (Condition Difficult). The number of flips applied to each prototype to create the close examples was 1 (Condition Easy), 2 (Condition Intermediate), or 4 (Condition Difficult). Finally, the number of flips applied to each prototype to create the far examples was 2 (Condition Easy), 4 (Condition Intermediate), or 8 (Condition Difficult).

The three conditions may be conceptualized as three levels of difficulty of a categorization task. Condition Easy was the easiest task because the examples were tightly distributed around their prototype and the concepts were well-differentiated. Condition Difficult was the hardest task because the examples were widely dispersed around their prototype and the concepts overlapped. Condition Intermediate was an easier task than Condition Difficult, but harder than Condition Easy. The concepts overlapped less than in Condition Difficult, but they were not as well differentiated as in Condition Easy.

To study the influence of training experience, the networks were trained for different numbers of epochs, varying from 5 to 700. An epoch is a training period during which a network is exposed to all trained examples once in random order. The networks were trained for 5, 10, 25, 50, 75, 100, 200, 300, 400 or 700 epochs. Twenty networks were trained for each number of epochs in each of the three conditions, for a total of 600 networks.

Results

We reserve a detailed discussion of all our findings for a longer paper and we describe here only some of the most important results. We chose network error as the dependent measure, error being defined as the sum of the squared differences between inputs and outputs. Because network error is the difference between the input and output patterns, it reflects familiarization with the examples – how well the networks recognize the examples. Thus, lower network error indicates a higher level of familiarization with the examples.

As training progressed, the mean network error decreased in all three conditions, reflecting the networks' increased familiarity with the examples. At the end of training, the mean error for the trained examples approached zero. The mean error for the new test examples was higher than the error for the trained examples, although it had decreased considerably during training. This indicates that the networks learned the trained examples very well, and at the same time generalized their acquired knowledge to the test examples never seen in training..

The most central findings of the simulations are illustrated in Figures 2 and 3. The figures show the prototype and exemplar effects in each condition as a function of the number of epochs.

Figure 2 shows the prototype effect calculated separately for the trained and for the new test examples. We calculated the prototype effect for each network by subtracting the mean error for the close examples from the mean error of the far examples. Thus, the prototype effect on the trained examples is the difference between the error for the far-train examples and the close-train examples. The prototype effect on the test examples is the error difference between the far-test and the close-test examples. A positive difference indicates a prototype effect, that is, smaller error for the examples that are more similar to the prototype.

Figure 3 illustrates the exemplar effect calculated separately for the far and the close examples. We calculated the exemplar effect by subtracting the mean error for the train examples from the mean error of the test examples. The exemplar effect on the close examples is the error difference between the close-test and the close-train examples. The exemplar effect on the far examples is the error difference between the far-test and the far-train examples. A positive difference indicates an exemplar memorization effect, which means that the error is smaller for the trained examples than for the test ones; or, in other words, that the networks are more familiar with examples that have already been encountered than with novel examples.

We performed an ANOVA on the error differences shown in Figure 2 with the within-network factor Train vs. Test Examples and the between-network factors Number of Epochs and Condition. We performed a similar ANOVA on the error differences shown in Figure 3. All main effects and interactions were reliable in both analyses, minimum $F(9, 570) = 4.54$, $p < .001$. We analyzed these effects separately for each condition, and found that all main effects and interactions were significant, minimum $F(9, 190) = 2.49$, $p = .010$, except the main effect of Epoch in Condition Difficult in Figure 2, $F < 1$. Hence, we describe the results without referring to more detailed statistical tests because all the effects we discuss are licensed by these significant main and interactive effects.

Category Structure

The difficulty of the task had a sizeable impact on the prototype effect (Figure 2). The prototype effect was quite

large in Condition Easy and somewhat smaller in Condition Intermediate. This effect was reversed in Condition Difficult as demonstrated by the negative difference scores; networks' error was higher for the close examples than for the far ones. The close examples in Condition Difficult shared a high degree of similarity, causing the networks to easily confuse them with each other. Thus, examples that shared a high degree of similarity with their prototype no longer had an advantage over ones that did not. This finding is consistent with Smith and Minda's (1998) psychological results. They found a reversed prototype effect with poorly structured categories. Thus, the prototype effect diminished and even reversed as the difficulty of the task increased.

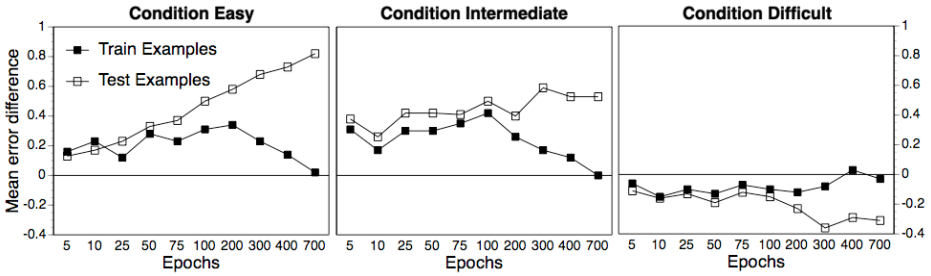


Figure 2: Prototype effect on the trained and the new test examples.

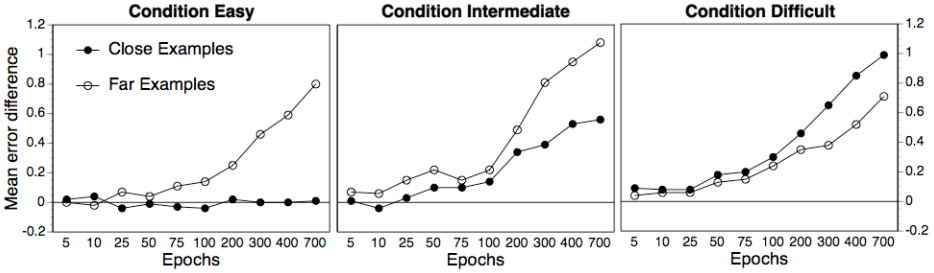


Figure 3: Exemplar effect on the examples that were close and those that were far from the prototype.

In contrast, the exemplar effect increased with the difficulty of the task (Figure 3), which is also consistent with psychological data (Minda & Smith, 2001; Smith & Minda, 1998). The networks relied more on exemplar memorization as the task became increasingly difficult and the prototype representation no longer provided useful information for discriminating the categories.

Amount of Training

The exemplar effect increased with the number of epochs in every condition (Figure 3), simulating Smith and Minda's psychological results (Minda & Smith, 2001; Smith & Minda, 1998). The prototype effect on the trained examples, on the other hand, decreased with the number of epochs in Conditions Easy and Intermediate, but was less affected by the number of training epochs in Condition Difficult. The decreasing prototype effect for the trained examples is consistent with Smith and Minda' results with trained examples. They did not test new examples in their experiments. Our networks make another novel prediction, namely that the prototype effect should increase with training for new test examples, especially if the categorization task is easy (left panel of Figure 2).

Networks became increasingly familiar with trained examples because they could memorize them. As training progressed, networks' recognition of trained examples relied more on individual memories, and less on their similarity to the prototype (just as with Smith and Minda). In contrast, novel examples had not been memorized. Hence, recognition of novel examples relied solely on their similarity to the prototype, and this prototype effect increased during training presumably because the prototype representation became increasingly well-defined.

Interaction Between Exemplar and Prototype Effects

The prototype effect was greater for new test examples than for old, trained ones (Figure 2). This finding seems realistic because only the trained examples could be memorized. Furthermore, the exemplar effect was stronger on the far examples than on the close examples in Conditions Easy and Intermediate (Figure 3, left and middle panels). Features of atypical instances were better remembered than those of typical instances. This presumably occurred because there was less interference between the memories of the atypical examples than between the similar memories of the typical examples. This is consistent with Light, Kayra-Stuart and

Hollander's (1979) finding that adults' recognition memory is better for atypical rather than typical faces. Similar results were found by Going and Read (1974) and Cohen and Carr (1975).

In Condition Difficult (right panel of Figure 3), however, the exemplar effect was larger on the close examples than on the far ones. The close examples were disadvantaged by their similarity to their prototype (because of the overlap between the categories); hence, these examples may have been the ones that benefited most from exemplar memorization. Reitman and Bower (1973) reported a similar effect with adult participants who were trained on an easy or a difficult categorization task. Following training, participants were given a recognition test. The results for the easy task were similar to Light *et al*.'s (1979) psychological results and our simulations in Conditions Easy and Intermediate: recognition performance was better for atypical examples. In contrast, their results for the difficult task were reversed: recognition performance was better for typical examples, matching our simulations in Condition Difficult.

Thus, prototype and exemplar effects seem to complement each other, each process having a stronger influence on the examples that are not favored by the other.

Discussion

We demonstrated that a unified model can capture both prototype and exemplar effects. The networks abstracted concept prototypes and at the same time remembered some features of the trained examples.

Networks also successfully simulated the prototype-to-exemplar trend as the learning task increased in difficulty (Minda & Smith, 2001; Smith & Minda, 1998). Our networks also showed an increase in the size of the exemplar effect from Condition Easy to Condition Difficult, as the concepts became more poorly structured. At the same time, the prototype effect substantially decreased and even reversed as difficulty level increased. For better-structured concepts (Conditions Easy and Intermediate), the exemplar effect was greater farther away from the prototype; for poorly structured concepts (Condition Difficult), the exemplar effect was greater closer to the prototype. As we mentioned earlier, this is consistent with a number of psychological studies.

The networks also exhibited a shift from prototype use to exemplar memorization during training. We observed an increase in the exemplar effect and a decrease in the prototype effect on the trained examples. Better memorization with more training makes perfect sense, as memorization depends on the amount of experience. A possible reason for the decrease in the use of prototype information for the trained examples is that it is less needed as the examples are better remembered. This is consistent with psychological studies reviewed earlier (Hayes & Taplin, 1993; Horst, *et al*., 2005; Mervis & Pani, 1980; Minda & Smith, 2001; Smith & Minda, 1998).

Other studies, however, reported that exemplar information is used earlier in development, and the ability to abstract a prototype emerges later (Fisher & Sloutsky, 2005; Sloutsky & Fisher, 2004; Tighe, Tighe, & Schechter, 1975). Fisher and Sloutsky (2005), for instance, found that younger children's memory for trained items was significantly better than that of older children and adults, suggesting that the latter relied more on an average prototype representation.

It is important to note a key difference with these studies. The studies finding an exemplar-to-prototype shift used concepts with defining features, while those that found a prototype-to-exemplar shift did not (and neither did our simulations). Defining features are present in all examples that belong to a category, and only in those, allowing perfect categorization performance. For example, Tighe *et al.* (1975) used a word classification task in which names of animals belonged to one category, while body parts belonged to another. Following this classification task, adults were less likely to correctly recognize a previously encountered example than children. Tighe *et al.* proposed that adult participants used the defining feature as an encoding device and learned less about the other features of the words. In contrast, children are less likely to use defining features (Keil & Batterman, 1984), which may result in better memorization of the probabilistic features.

Interestingly, Shultz *et al.* (2008) successfully simulated this shift from probabilistic feature learning to the use of defining features using the same kind of networks presented here. To test the hypothesis that defining features affect exemplar memorization in the present work, we repeated the simulations for Condition Intermediate, but added two defining features to each example. Although exemplar memorization did not decrease with training (on the contrary, it increased), overall network error was higher in the simulations with defining features. These networks were less familiar with the trained examples than if they had been trained without defining features. This is consistent with Tighe *et al.*'s (1975) finding that adults, who use defining features more readily than children, exhibit poorer recognition performance. This explains why Tighe *et al.* and other researchers who also used defining features (Fisher & Sloutsky, 2005; Sloutsky & Fisher, 2004) found better memorization of exemplars in children than in adults.

To conclude, our simulations further decrease the gap between the numerous incongruent studies reported in the literature regarding the development of exemplar and prototype effects during category learning. Indeed, considering factors such as the structure of the categories and the presence of defining features, there is considerable, unexpected coherence in these mixed results. Most importantly, we have demonstrated that it is possible for a single mechanism to capture a gradual shift in concept processing depending on task difficulty and the amount of experience.

Acknowledgments

This research was supported by a postgraduate fellowship to IB and a grant to TRS from the Natural Sciences and Engineering Research Council of Canada.

References

Baluja, S., & Fahlman, S. E. (1994). *Reducing network depth in the cascade-correlation learning architecture. Tech Report CMU-CS-94-209*: School of Computer Science, Carnegie Mellon University.

Cohen, M. E., & Carr, W. J. (1975). Facial recognition and Von Restorff effect. *Bulletin of the Psychonomic Society, 6*(4), 383-384.

Fisher, A. V., & Sloutsky, V. M. (2005). When induction meets memory: Evidence for gradual transition from similarity-based to category-based induction. *Child Development, 76*(3), 583-597.

Going, M., & Read, J. D. (1974). Effects of uniqueness, sex of subject, and sex of photograph on facial recognition. *Perceptual and Motor Skills, 39*(1), 109-110.

Hayes, B. K., & Taplin, J. E. (1993). Developmental differences in the use of prototype and exemplar-specific information. *Journal of Experimental Child Psychology, 55*(3), 329-352.

Hintzman, D. L. (1986). Schema abstraction in a multiple-trace memory model. *Psychological Review, 93*(4), 411-428.

Homa, D., & Cultice, J. (1984). Role of feedback, category size, and stimulus distortion on the acquisition and utilization of ill-defined categories. *Journal of Experimental Psychology: Learning Memory and Cognition, 10*(1), 83-94.

Homa, D., Sterling, S., & Trepel, L. (1981). Limitations of exemplar-based generalization and the abstraction of categorical information. *Journal of Experimental Psychology: Human Learning and Memory, 7*(6), 418-439.

Horst, J. S., Oakes, L. M., & Madole, K. L. (2005). What does it look like and what can it do? Category structure influences how infants categorize. *Child Development, 76*(3), 614-631.

Keil, F. C., & Batterman, N. (1984). A characteristic-to-defining shift in the development of word meaning. *Journal of Verbal Learning and Verbal Behavior, 23*(2), 221-236.

Light, L. L., Kayra-Stuart, F., & Hollander, S. (1979). Recognition memory for typical and unusual faces. *Journal of Experimental Psychology: Human Learning and Memory, 5*(3), 212-228.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 85*(3), 207-238.

Mervis, C. B., & Pani, J. R. (1980). Acquisition of basic object categories. *Cognitive Psychology, 12*(4), 496-522.

Minda, J. P., & Smith, J. D. (2001). Prototypes in category learning: The effects of category size, category structure, and stimulus complexity. *Journal of Experimental Psychology: Learning Memory and Cognition, 27*(3), 775-799.

Palmeri, T. J., & Nosofsky, R. M. (2001). Central tendencies, extreme points, and prototype enhancement effects in ill-defined perceptual categorization. *Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology, 54*(1), 197-235.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology, 77*(3), 353-363.

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology, 3*, 382-407.

Reed, S. K. (1978). Category vs item learning: Implications for categorization models. *Memory & Cognition, 6*(6), 612-621.

Reitman, J. S., & Bower, G. H. (1973). Storage and later recognition of exemplars of concepts. *Cognitive Psychology, 4*(2), 194-206.

Shultz, T. R. (2003). *Computational developmental psychology*. Cambridge, MA: MIT Press.

Shultz, T. R. (2006). Constructive learning in the modeling of psychological development. In Y. Munakata & M. H. Johnson (Eds.), *Processes of change in brain and cognitive development: Attention and performance XXI*. (pp. 61-86). Oxford, UK: Oxford University Press.

Shultz, T. R., Mysore, S. P., & Quartz, S. R. (2007). Why let networks grow? In D. Mareschal, S. Sirois, G. Westermann & M. H. Johnson (Eds.), *Neuroconstructivism: Perspectives and prospects* (Vol. 2, pp. 65-98). Oxford, UK: Oxford University Press.

Shultz, T. R., Thivierge, J.-P., & Laurin, K. (2008). Acquisition of concepts with characteristic and defining features. *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, 531-536.

Sloutsky, V. M., & Fisher, A. V. (2004). When development and learning decrease memory: Evidence against category-based induction in children. *Psychological Science, 15*(8), 553-558.

Smith, J. D., & Minda, J. P. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning Memory and Cognition, 24*(6), 1411-1436.

Tighe, T. J., Tighe, L. S., & Schechter, J. (1975). Memory for instances and categories in children and adults. *Journal of Experimental Child Psychology, 20*(1), 22-37.

Person, place, and past influence eye movements during visual search

Barbara Hidalgo-Sotelo (bhs@mit.edu)
Aude Oliva (oliva@mit.edu)
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology
Cambridge, MA 02139

Abstract

What is the role of an individual's past experience in guiding gaze in familiar environments? Contemporary models of search guidance suggest high level scene context is a strong predictor of where observers search in realistic scenes. Specific associations also develop between particular places and object locations. Together, scene context and place-specific associations bias attention to informative spatial locations. At the level of eye fixations, it is not known whether a person's specific search experience influences attentional selection. Eye movements are notoriously variable: people often foveate different places when searching for the same target in the same scene. Do individual differences in fixation locations influence how a scene is subsequently examined? We introduce a method, comparative map analysis, for analyzing spatial patterns in eye movement data. Using this method, we quantified the consistency of fixated locations within the same observer and between observers during search of real world scenes. Results indicated a remarkable consistency in the locations fixated by the same observer across multiple searches of a given scene. This observer-specific guidance was shown to be distinct from general scene context information or familiarity with the scene. Accordingly, this is considered evidence for a uniquely informative role of an individual's search experience on attentional guidance in a familiar scene.

Keywords: visual search; eye movements; scene perception; learning; attentional guidance; fixation similarity

Introduction

An important feature of ecological visual search is that there are few truly novel, unfamiliar places in which a person is likely to search. Many tasks involve examining the same place repeatedly, such as the various times spent searching for a specific utensil in one's own kitchen. Locating the target in question benefits from both category based information (e.g. utensils are on countertops) and place specific information (e.g. in this kitchen, utensils hang over the stove). For any observer, there will be many sources of information that guide which scene regions are inspected during search. What influence does a person's own search experience (i.e. fixation locations) have in guiding where they are likely to look in familiar scenes?

A growing body of evidence suggests that observers use high level information, such as learned target features and global scene context, to guide their gaze when searching for an object in real world environments (Ehinger, Hidalgo-Sotelo, Torralba, & Oliva, 2009; Hwang, Higgins, Pomplun, 2009). At this level of *categorical* representation, knowledge of the basic-level scene category and target features directs gaze to expectation-based scene regions (Eckstein, Drescher & Shimozaki, 2006; Henderson, 2003; Torralba, Oliva, Castelhano, & Henderson, 2006). At the level of *scene exemplar* representations, spatial context can also be used to allocate attention preferentially to regions that have become associated with the target. In contextual cueing, for example, observers implicitly learn patterns in repeated displays that help them find a target faster in repeated configurations (Chun & Jiang, 1998). It is not well understood, however, whether a scene exemplar representation can systematically bias individual fixations.

How can "experience based" influences be distinguished from the myriad of sources that guide attention to relevant scene regions? One challenge is that attention is strongly guided by information that does *not* depend on specific experience. Figure 1 illustrates regularities in eye fixations across and within observers. In Figure 1A, fixations from 9 observers searching for a book are shown; the high fixation density along countertop surfaces illustrates how spatial layout and context guide where observers look. Systematic biases unrelated to the scene's content also influence gaze location. In Figure 1B, fixations sampled from random scenes have been projected onto the kitchen scene. Center bias in the fixation distribution is driven by oculomotor tendencies (Tatler, 2007; Tatler & Vincent, 2009) and photographer bias. A second challenge, of the opposing nature, lies in the significant variability in fixation locations across individuals. As a result, two independent observers may fixate different scene regions, even when looking for the same object in the same scene (Figure 1C). It is possible that individuals are biased by experience, but that the effects are masked by pooling over experienced observers. Given initial differences in search patterns, could systematic differences arise when an observer repeats her search of the scene? To reasonably estimate the influence of past experience, the search patterns of observers who have *never* viewed the scene must be contrasted with different observers who have *previously* searched the scene.

In this paper, eye movement data from a visual search study was analyzed using approach we have termed comparative map analysis. This analysis was used to evaluate how different sources of information contribute to attentional guidance during visual search of familiar scenes. In our experiment, observers' eyes were tracked while they looked for a book in pictures of real world scenes. On some trials, observers searched a scene that had been presented previously. Importantly, the target object and location remained unchanged in each presentation of the scene. The

main question was whether a person's past experience (as measured by fixated locations) biases attentional selection when searching a familiar scene. Using comparative map analysis, we show that visual attention is sensitive to the influence of a person's past experience of searching in familiar scenes.

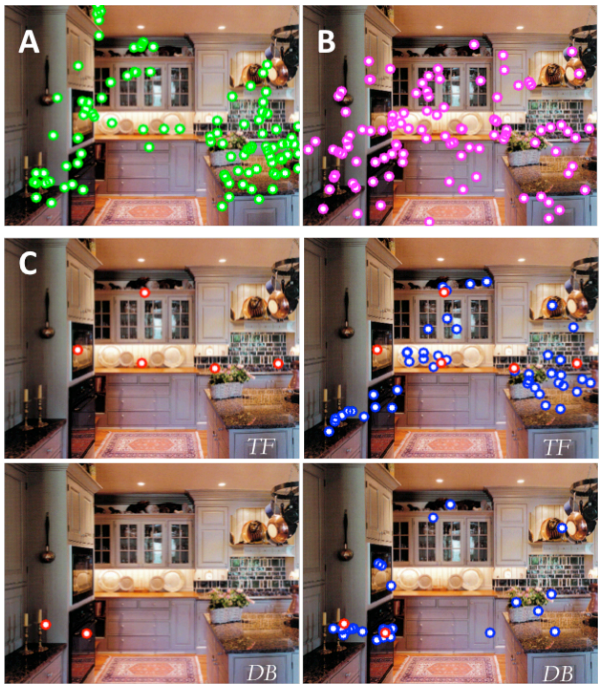


Figure 1: Regularities in eye movements while searching for books. (A) Fixations from 9 observers searching for a book in this kitchen (green dots). Context and spatial layout constraints guide search (e.g. high density along countertop surfaces in the foreground and background. (B) Fixations sampled from random scenes and projected onto this scene (pink dots). Oculomotor and photographer bias contribute to a roughly central fixation distribution with sparse fixations in the image periphery. (C) Fixations from 2 observers who repeatedly searched this kitchen. Each row shows fixations from: the observer's first search (Left, red dots), and the next 7 search trials (Right, blue dots). Individual differences in fixation patterns are evident, before and after learning.

Comparative Map Analysis

The approach we describe here as comparative map analysis is used to evaluate how well different distributions of fixations predict where observers will look in a scene. Critically, each fixation distribution is sampled from a different, strategically chosen, population of fixations. The resulting distributions are evaluated in regards to how well they distinguish between fixated and unfixated locations. In the present paper, this analysis was used to determine whether an observer's experience plays a significant role in attentional selection during search.

Logic of the approach. Given the challenges outlined in the introduction, how can we isolate the bias resulting from an individual's experience searching a *specific* scene? The solution lies in strategically identifying fixation populations relevant to the question of interest. One population, for example, includes the locations fixated by novel searchers in a given scene. A second population includes the locations fixated by a *single observer* when the same scene was repeatedly searched. While the first population represents the influence of (general) scene context on search, the second population reflects the specific influence of the observer's own examination of the scene. Fixation maps were created for each population and used to predict fixation locations from a separate trial. If the two populations are equally informative, then there will be no significant difference in the accuracy between the predictions. The logic is analogous to established methods for determining whether fixated and control locations can be discriminated (e.g. Parkhurst & Niebur, 2003; Tatler, Baddeley, and Gilchrist, 2005). In those studies, the two distributions represent measurements of a dependent variable (e.g. visual feature content) at fixated versus unfixated locations. If the dependent variable successfully discriminates between these locations, then it is considered to inform fixation selection. Control distributions, it should be noted, can be constructed in several ways. Recent studies of attentional guidance have constructed control distributions by randomly sampling fixations from other populations (e.g. Ehinger et al, 2009; Tatler et al, 2005; Tatler & Vincent, 2009). Comparative map analysis extends this rationale by defining several control populations that vary with respect to the degree of "person," "place," and "past" information represented.

Broadly, we consider three *scene dependent populations* representing scene regions empirically fixated by observers when searching that specific scene: (1) Fixations made by a single observer's repeated searches, (2) Fixations of other observers who searched the scene repeatedly; (3) Fixations of novel observers (i.e. searched the scene once). Importantly, these populations represent slightly different sources of information: self-consistency, scene familiarity and general scene context, respectively.

Control populations are crucial to assess the relative informativeness of other regularities (e.g. oculomotor biases) in predicting the same eye movements. These *scene independent populations* provide controls for different sources of information: (4) Fixations from the same observer on random scenes, (5) Fixations from different observers on random scenes. These populations reflect spatial biases in oculomotor behavior that manifest across the set of scenes (intra-observer and inter-observer biases respectively). Two simple model-based populations (as opposed to sampling from empirical fixations) serve as controls to evaluate the extent to which a central gaussian distribution (6) and uniform distribution (7) predicted observers' fixations. The uniform distribution serves as the true measure of chance, while the widely recognized central fixation bias in human eye movements (Tatler, 2007)

suggest that a central gaussian distribution may predict fixations above chance level.

Building fixation maps. Fixation maps were created for each of the above populations using the following procedure, shown schematically in Figure 2. First, we collected a list of the locations fixated by one observer in all repeated searches of a scene; trials in which the eye was lost or the observer failed to find the target object were not included. For each repeated search trial *R*, a self-consistency fixation map (1) was built by excluding fixations from one search trial and using the remaining *N* fixations to define a prediction map. Next, the other fixation maps were created by sampling *N* times from the appropriate population of empirical fixations (2-5) or statistical model (6-7). This process was iterated for R repeated search trials, and the resulting fixation maps were used to predict the excluded trial's fixations (probe fixations).

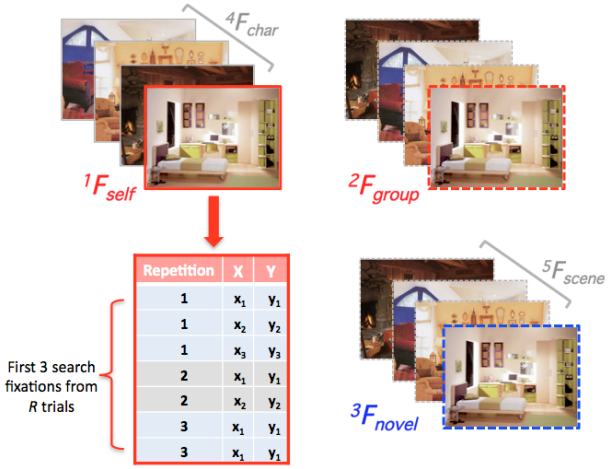


Figure 2: Schematic of comparative map analysis. This illustrates the source of fixation populations (1-5) and how they are sampled to create fixation maps that represent several influences on eye guidance. The following steps are performed iteratively for each of *R* trials: select one search trial (i.e. first 3 fixations of one trial) from F_{self} ; use the remaining *N* fixations to create a prediction map for intra-observer similarity. Fixation maps for populations (2-5) are created by sampling *N* times from the corresponding distributions. Red (familiar observers) and blue (novel) outlines represent scene dependent populations. Dashed outlines indicate non-self fixation populations.

In the present analysis, the first 3 *search* fixations in each search trial were used to build the fixation maps. Search fixations are defined as fixations made during active exploration of the scene, thus excluding fixations landing on the target and the initial central fixation. The maps were compared in terms of how well they predicted the first 3 search fixations of the excluded trial. Given past findings that the consistency of fixation locations across observers decreases over time (Mannan, Ruddock, Wooding, 1997; Yarbus, 1967), we used the first 3 search fixations because it represented a time window appropriate for capturing the highest consistency across novel and repeated conditions.

Evaluating fixation maps. We used the Receiver Operator Characteristic to evaluate how well fixated and unfixated locations could be discriminated. The ROC curve is a common signal detection technique that represents the proportion of real fixations falling within a fixation map (detection rate) in relation to the proportion of the image area selected (false alarm rate) (e.g. Ehinger et al, 2009; Renninger, Verghese, & Coughlan 2007; Tatler et al, 2005). The area under the curve or AUC area (Green & Swets, 1966) was used to compare differences in prediction maps.

Search Experiment

In this experiment, observers searched for a book in indoor scenes (e.g. kitchens, bedrooms) while their eye movements were recorded. The original goal of this study was to investigate how time influenced the retrieval and use of scene specific associations to guide search in realistic scenes. We examined this by introducing a variable stimulus onset asynchrony (SOA) between the scene onset (observers fixating centrally) and the initial search fixation on the scene. We predicted that there would be an interaction between scene familiarity and SOA, such that longer delays would predict shorter search times on familiar, but not novel, scenes. For the present analysis, the eye movements collected from this study were collapsed across the retrieval-time manipulation since this variable was tested using a within-subject design.

Participants. Eighteen observers, ages 18-34, with normal acuity gave informed consent, passed an eyetracking calibration test, and were paid $15/hr for their participation.

Materials. Eye movements were collected using an ISCAN RK-464 video-based eyetracker with a sampling rate of 240 Hz. The stimuli were high resolution color photographs of indoor scenes presented on a 15" LCD monitor with a resolution of 1280 x 1024 px and refresh rate of 60 Hz. The original images were cropped and resized to be presented at a resolution of 1024 x 768 px, subtending 30 x 20 deg of visual angle. Presentation of the stimuli was controlled with Matlab and Psychophysics Toolbox (Brainard, 1997; Pelli, 1997). The target prevalence in the stimuli set was 100%: all scenes contained a target and, importantly, the target location never changed in a particular scene. To make the task challenging, book targets were small (from 1 to 2°) and spatially distributed across the image periphery.

Procedure. The experiment consisted of a learning phase followed by a probe phase. In the learning phase, observers learned associations between specific scenes and a book's location in each scene. In the probe phase, observers searched following a variable SOA (200, 400, 800, or 1600 ms) on a novel or familiar scene. In both phases, observers freely explored the scene with their eyes. Each phase was comprised of 4 search blocks: 24 repeated search trials and 8 novel search trials presented randomly in each block. Scenes were counterbalanced across observers with respect

to the novel or repeated conditions. The trial sequence, similar in learning and probe phases, is as follows. Observers fixated a central fixation cross for 500 ms to begin the trial (gaze contingent). First, a scene was presented with a fixation cross superimposed over the scene; observers fixated the central cross for the duration of this interval without saccading away otherwise the trial ended. In the test phase, this was followed by a variable SOA on a gray screen. Finally, the same scene was presented again and observers actively explored the scene to find the book. Observers had a maximum of 8 s to respond via key press (learning phase) or by fixating the target for 750 ms (probe phase). Feedback was given after each trial (reaction time displayed for 750 ms) to encourage observers to search speedily throughout the experiment. The entire experiment lasted approximately 50 min.

Eyetracker calibration was critical for the gaze contingent aspects of the procedure, as well as to ensure accurate dependent measures (fixation locations). For this reason, calibration was checked at 9 locations evenly distributed across the screen after each search block; fixation position had to be within 0.75° of visual angle for all points, the experiment halted and the observer was recalibrated.

Eye movement analysis. Fixations were identified on smoothed eye position data, averaging the raw data over a moving window of eight data points (33 ms). Beginning and end positions of saccades were detected using an algorithm implementing an acceleration criterion (Araujo, Kowler, & Pavel, 2001). Specifically, the velocity was calculated for two overlapping 17 ms intervals; the onset of the second interval was 4.17 ms after the first. The acceleration threshold was a velocity change of 6 deg/s between the two intervals. Saccade onset was defined as the time when acceleration exceeded threshold and the saccade terminated when acceleration dropped below threshold. Fixations were defined as the periods between saccades. Saccades within 50 ms of each other were considered continuous.

Comparative map analysis. Forty eight scenes were searched by equal numbers of participants in the novel and repeated conditions. Search trials in the learning and probe phases, excluding block 1, were combined to yield a maximum of 7 repeated trials for each observer. The following experiment conditions correspond to each population: (1) One observer's repeated searches of a familiar scene, (2) *Other* observers' repeated searches of the same familiar scene. (3) Different observers' *novel* search of the same scene. (4) Any scene searched by the same observer. (5) Any scene searched by other novel observers.

Results

The results of comparative map analysis are shown in Figure 3. Our main finding is the evidence of experience based influences on attentional selection, specifically during the first 3 search fixations in a scene. An identical pattern of results was found when using only the *first* search fixation. We first report the results from the populations based on scene dependent information (F_{self}, F_{group}, F_{novel}), followed by the scene independent control populations.

Role of the person

The role of a person's own search experience was evaluated by using the locations of their own fixations (F_{self}) to predict empirical fixations from the same observer on a separate search of the same image. We found that this population provided the most accurate predictions (mean AUC=0.907) relative to the other scene dependent populations F_{group} ($t(94)=5.41$, $p < 0.001$) and F_{novel} ($t(94)=6.57$, $p < 0.001$), and was significantly higher than control populations. Interestingly, observer's own population of fixations resulted in the most consistently accurate predictions across the set of images, as evident in the boxplot of figure 3.

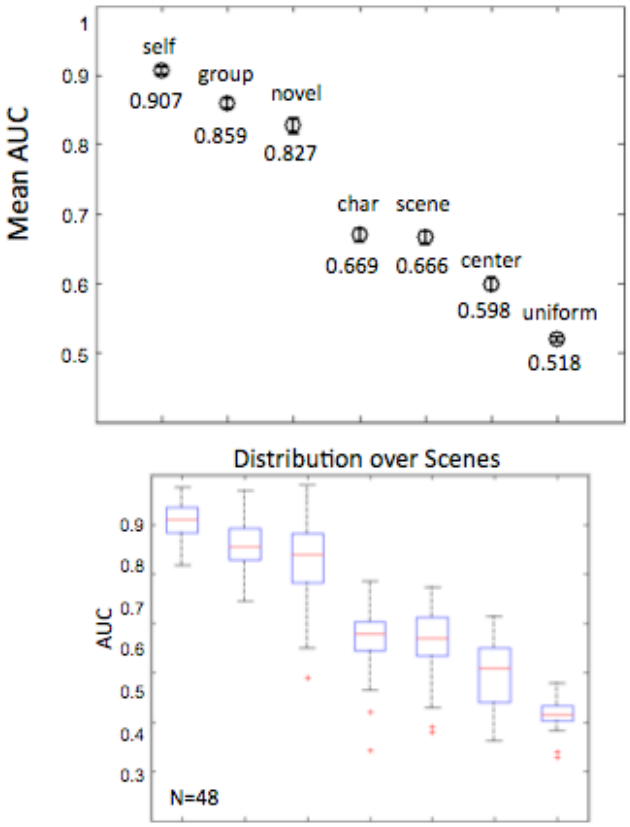


Figure 3: Results of comparative map analysis on eye movement data from the book search experiment. Distributions shown in the boxplot show the median (red line), upper and lower quartile values (box), and outliers.

Is this influence in fact due to a person's *specific* search experience? Perhaps the experience of the individual is not unique from the experience of the group. This is a reasonable hypothesis, given that all observers have the same opportunity to learn the association between the scene's identity and the location of a book. To examine this hypothesis, we compare F_{self} and F_{group}.

Role of the past

The role of past experience was evaluated using the fixation locations from other observers who searched the same scene repeatedly (F_{group}). Interestingly, there was no significant difference between the prediction accuracy of this group and a group of novel observers (mean AUCs of 0.859 and 0.827, respectively; $t(94)=1.97$). This suggests that sampling from many individuals with past experience is not significantly more informative than sampling from the population of novel observers.

Role of the place

The role of the place is perhaps the most intuitive source of information: it represents how scene context drives consistency in fixation locations across different novel observers. We found that F_{novel} provided a significant source of guidance relative to the random scene control F_{scene} ($t(94)=11.7$, $p < 0.001$). Our finding confirms previous reports of overall high inter-observer consistency in search tasks (Ehinger et al, 2009; Torralba et al, 2006).

Scene Independent Control Populations

Two control populations were based on empirical fixations sampled from different scenes: F_{char} (same observer as F_{self}) and F_{scene} (different observers). These populations predicted fixations well above chance with mean AUCs of 0.669 and 0.666 respectively and were not significantly different from one another ($t(94)=0.11$). The overlap in these distributions is not surprising given that these populations reflect systematic oculomotor tendencies and regularities in the stimuli set (e.g. photographer bias). The two model distributions, central gaussian and uniform, were used to compare with the other populations and confirm intuitions about the results of comparative map analysis. Indeed, the central gaussian model was a better predictor of fixations than the uniform distribution ($t(94)=2.7$, $p < 0.05$).

Discussion

We have shown that the past repeats itself: a person's experience, as indexed by fixated scene locations, influences how they search familiar scenes. Although the notion of idiosyncratic gaze patterns has been previously presented (Noton & Stark, 1971), to the best of our knowledge, this is the first time observer-specific experience has been shown to influence gaze patterns in a naturalistic search task. What is the nature of the information that underlies this self-consistency effect? Is it behaviorally relevant or an incidental consequence of scene exposure? Is the encoded information object-based or spatially-based? How does self-consistency interact with other well characterized forms of search guidance (e.g. saliency)?

In order to refine our understanding of why intra-observer consistency occurs, it would be helpful to examine patterns across observers and scenes. Are certain scenes searched more consistently than others? This question can be approached in two ways. From the perspective of general scene context constraints, scenes vary in the distribution of target-probable surfaces they contain. Looking for books in a library, for example, may present a significantly less constrained search than searching a bathroom. Still, the boxplot in figure 3 suggests the scenes are variable with respect to how consistently similar regions are selected by different viewers. From the perspective of person specific constraints, what is the relation between inter-observer and intra-observer consistency? One possibility is that scenes searched consistently by novel observers also promote self-consistency among a large proportion of familiar observers. Alternatively, high *variability* in intra-observer consistency (i.e. high F_{self} variance) may negatively correlate with inter-observer consistency. Identifying properties of the scene and task that promote self-consistency across searches remains an open question.

In the ecological psychology tradition (e.g. Gibson, 1979), our findings also raise questions about the behavioral significance of self-consistency. Are some observers more self-consistent than others? If so, what are the implications for the search task? One hypothesis is that *high* self-consistency may be associated with good search performance (e.g. fast overall reaction time). Indeed, a widely recognized feature of human memory relates to the benefit of reinstating the encoding context in retrieval (Jacoby & Craik, 1979; Tulving & Thomson, 1973). Furthermore, embodied cognition accounts suggest that a person's own movements may play a role in perceptual and cognitive performance (e.g. Knoblich & Flach, 2001). When imagining a previously viewed stimulus, for example, observers tend to make reenact patterns of eye movements from the initial viewing (Brandt & Stark, 1997; Laeng & Teodorescu, 2002; Spivey et al, 2001).

It is important to note the role of our task in driving similar patterns of viewing across observers. A number of recent studies have sought to predict where observers will look in naturalistic scenes. Many of these studies, however, deliberately employ free viewing (e.g. Bruce & Tsotsos, 2006; Itti & Koch, 2000) or a memory task (e.g. Foulsham & Underwood, 2008) so as to reduce the influence of having a common goal. Theories of visual search guidance (e.g. Wolfe, 1994) describe observers' deployment of attention as resulting from a combination of stimulus and goal driven factors. Seeing how the magnitude of self-similarity varies across tasks can serve as another approach to assessing the behavioral significance of intra-observer consistency. Recognition memory tasks, in particular, provide an opportunity to investigate the causal role of re-fixations in scene recognition. Holm & Mantyla (2007) used a remember/know paradigm to evaluate whether successful recognition was associated with similarity between an observer's fixations during study and test phases. Indeed, they found evidence that recollection ("remember" responses) were related to a high degree of study-test consistency. Recently, Underwood and colleagues (2009) investigated the roles of domain knowledge and visual saliency on fixation consistency in scene recognition. Their

findings again support the idea that observers look at scene locations that have been previously fixated and, interestingly, that the effect is stronger for individuals who were experts in the domain related to the picture.

Our experiment shows that observers have access to perceptual and memory based information that helps them locate the book in a familiar scene. What is the nature of this information? Two possibilities are that observers encoded the oculomotor movements to *spatial locations* (e.g. left side of the screen) or the *objects* (e.g. empty bookshelf) that were attended on the way to finding the target. One way to distinguish these possibilities would be compare the resulting search patterns when an observer initiates search from a familiar (e.g. center of the scene) or an unfamiliar location and comparing whether similar objects or locations were still fixated. Moreover, the speed of human eye movements (roughly 3-4 per second) suggests an automatic component to self-consistency that may not be available to conscious awareness. Although our experiment cannot speak to this issue directly, we found the same pattern of results shown in figure 3 using only observer's first fixation on the scene. This suggests that the information underlying self-consistency is rapidly available to bias eye movements.

Conclusion

Comparative map analysis, a novel approach for analyzing patterns in eye movement data, was used to evaluate the role of various sources of search guidance. We found evidence from a search study showing a uniquely informative role of an individual's experience on attentional guidance in a familiar scene

Acknowledgments

BHS was supported by a graduate fellowship from an Integrative Training Program in Vision grant (T32 EY013935) and an NSF CAREER award (0546262) to AO. The authors wish to thank Talia Konkle and Edward Vul for helpful comments.

References

Araujo, C., Kowler, E., & Pavel, M. (2001). Eye movements during visual search: The cost of choosing the optimal path. *Vision Research*, 41, 3613 3625.

Brainard, D.H. (1997). The Psychophysics Toolbox. *Spatial Vision*, 10, 433-436.

Brandt, S.A., & Stark, L.W. (1997). Spontaneous eye movements during visual imagery reflect the content of the visual scene. *Jrnl of Cog Neuro*, 9, 27-38.

Chun, M.M., & Jiang, Y. (1998). Contextual cueing: Implicit learning and memory of visual context guides spatial attention. *Cognitive Psychology, 36,* 28–71

Foulsham, T., & Underwood, G. (2008). What can saliency models predict about eye movements? Spatial and sequential aspects of fixations during encoding and recognition. *Journal of Vision*, 8(2), 1-17.

Ehinger, K., Hidalgo-Sotelo, B., Torralba, A., & Oliva, A. (2009). Modeling search for people in 900 scenes: A combined source model of guidance. *Visual Cognition*, 17, 945-978.

Eckstein, M. P., Drescher, B. A., & Shimozaki, S. S. (2006). Attentional cues in real scenes, saccadic targeting and Bayesian priors. *Psychological Science*, 17, 973 980.

Gibson, J. (1979). The ecological approach to visual perception. Boston: Houghton Mifflin.

Henderson, J. M. (2003). Human gaze control in real-world scene perception. *TICS*, 7, 498-504.

Hidalgo-Sotelo, B., Oliva, A., & Torralba, A. (2005). Human Learning of Contextual Object Priors: Where does the time go? Proceedings of the IEEE Computer Society Conference on CVPR (pp. 510-516).

Holm, L., & Mantyla, T. (2007). Memory for scenes: Refixations reflect retrieval. *Memory & Cognition*, 35, 1664-1674.

Hwang, A.D., Higgins, E.C., Pomplun, M. (2009). A model of top-down attentional control during visual search in complex scenes. *Journal of Vision,* 9(5), 1-18.

Jacoby, L.L., & Craik, F.I. (1979). Effects of elaboration of processing at encoding and retrieval: Trace distinctiveness and recovery of initial context. *Levels of processing in human memory*. Hillsdale, NJ: Erlbaum.

Knoblich, G., & Flach, R. (2001). Predicting the effects of actions: Interactions of perception and action. *Psychological Science*, 12, 467-472.0

Parkhurst, D.J., & Niebur, E. (2003). Scene content selected by active vision. *Spatial Vision*, 16(2), 125-154.

Pelli, D.G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, 10, 437 442.

Mannan, S., Ruddock, K.H., & Wooding, D.S. (1997). Fixation patterns made during brief examination of two-dimensional images. *Perception*, 26, 1059-1072.

Noton, D., & Stark, L. (1971). Scanpaths in eye movements during pattern perception. *Science*, 171, 308-311.

Renninger, L.W., Verghese, P., & Coughlan, J. (2007). Where to look next? Eye movements reduce local uncertainty. *Journal of Vision*, 7(3), 1-17.

Tatler, B.W. (2007). The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, 7, 1-17.

Tatler, B.W., Baddeley, R.J., & Gilchrist, I.D. (2005). Visual correlates of fixation selection: Effects of scale and time. *Vision Research*, 45(5), 643 659.

Tatler, B.W., & Vincent, B.T. (2009). The prominence of behavioural biases in eye guidance. *Visual Cognition*, 17, 1029-1054.

Torralba, A., Oliva, A., Castelhano, M., & Henderson, J.M. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review,* 113, 766-786.

Tulving, E., & Thomson, D.M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, 80, 352–373.

Underwood, G., Foulsham, T., & Humphrey, K. (2009). Saliency and scan patterns in the inspection of real-world scenes: Eye movements during encoding and recognition. *Visual Cognition*, 17, 812-834.

Wolfe, J.M. (1994). Guided search 2.0. A revised model of visual search. *Psych Bulletin & Review*, 1, 202-228.

Yarbus, A. (1967). Eye movements and vision. New York: Plenum Press.

Perspectivizing Space in *Bāŋlā* Discourse

Samir Karmakar (samirk@nias.iisc.ernet.in)
Cognition Programme, School of Humanities, NIAS, IISc Campus
Bangalore 560012, India

Rajesh Kasturirangan (rkasturi@nias.iisc.ernet.in)
School of Humanities, NIAS, IISc Campus
Bangalore 560012, India

Abstract

The paper attempts to conceptualize the production and comprehension of spatial perspectives as the synchronization of intentions and contentions in a linguistic discourse. In doing so, it investigates the acts of intending and contending in invoking and instantiating the categories. The paper explains perspective setting and taking in terms of intending and contending which are crucial in shaping the conceptual route for the gradual revelation of the communicative intent. Answer to these questions, in turn, results into an understanding of what constitute the perspectivization process in a discourse.

Keywords: intention; contention; conceptual route; perspective taking; perspective setting.

Introduction

This paper investigates how spatial perspectives are represented and accessed in discourse due to the activation of linguistic expressions. We also explore how these explicitly languaged elements are situated and grounded. The term 'languaging' refers to the manner in which meaning potentials are invoked and realized at the time of discourse production and comprehension. The role of languaging in producing and/or comprehending a discourse is primarily an act of interpretation, since the emergence of meaning in a communicative situation is actually an outcome of the interpretive acts that unfold the structure of the communicative situation and the structuring capacities of the habitual attitudes of the mind (Rochberg-Halton 1982): While producing a discourse, we interpret our thought into language; whereas the discourse comprehension presumes the interpretation of language into the thought. We argue that linguistic expressions trigger two distinct cognitive functions – namely, intending and contending – while languaging spatial perspectives. These two cognitive functions are crucial in accommodating commonsense knowledge into the discourse interpretation through the act of languaging. We test our approach using spatial perspectives in *Bāŋlā* discourse, but the underlying ideas apply to the general question of how meaning is produced and comprehended in discourse.

Researchers have addressed questions related to the issues of spatial perspectives in language from different theoretical persuasions: In these studies, it has been shown that the production and comprehension of spatial descriptions presuppose the activation of asymmetries intrinsic to conceptual categories (Clark 1973). These categories are termed as frames of reference (Levinson 1996; Landau & Hoffman 2005; Majid et al. 2004; Neggers et al. 2006). A frame of reference can function egocentrically or allocentrically. An egocentric frame of reference invokes body-based asymmetries to organize spatial coexistences. In interpreting coexistences, allocentric frames of reference employ external reference frames such as landmark based cognition.

The importance of a frame of reference, as it follows from Piaget and Inhelder (1948), lies in its capacity to mirror the invariant aspects of a category with respect to which perspectives are interpreted. Researchers – see Heine (1989), Heine et al. (1991), Levinson (1996, 2003), Gibbs (2005), Levinson and Wilkins (2006) and others – have studied the linguistic realization of frames of reference at the sub-sentential level in order to answer the following question: how does the linguistic realization of space project the underlying conceptualization of different frames of reference? The answer to this question, in turn, sheds light on old puzzles about the relation between world, language and thought. These 'Whorfian' concerns led researchers to explore spatial universals and their lexicalization in different languages. They are extremely useful in understanding the representation of space in language and in setting correlations between spatial language and spatial cognition.

In spite of these advances in exploring the linguistic realization of space, what remains unanswered is how the users of a language access those representations and correlations while processing a discourse. So, a further investigation of perspective *taking* is long overdue. Additionally, a shift of interest from studying sub-sentential expressions to the study of discourse, as Fauconnier (1981) stressed, will offer "a conceptually different, theoretically more promising, and empirically more broader, system of understanding natural language logic." At the level of discourse, a static correlation between linguistic and cognitive categories is not enough. We also need to understand how these categories are grounded and situated; and, how higher order inferential judgments are integrated during the transformation of one spatial perspective into another (Karmakar 2009). The current investigation seeks to unveil the cognitive structures underlying the

perspectivization[1] process through the study of discourse. We investigate the following two questions: what cognitive functions are at work in perspectivizing space in discourse? How does the languaging of discourse manipulate these cognitive functions? These two questions will be discussed in this paper with reference to *Bāŋlā* language data.

Outline of the Approach

Spatial descriptions are perspectival, like any other linguistic communication (Mead 1938; Chakraborty 1992; Moore 1997; Coventry et al. 2009). In discourse, the descriptions of space, i.e. viewing arrangements, are languaged with respect to certain *vantage points*. A viewing arrangement, as Langacker (2008) defines it, is the 'overall relationships between the "viewers" and the situation being "viewed"'. The process of producing and comprehending spatial viewing arrangements in discourse is termed here as *perspespectivization* – that is the languaging of perspectives.

The act of perspectivization is a consequence of shared linguistic capacity (Akman 2000; Stalnaker 2002; Gibbs 2005), evolving through the generalization process (Mead 1934; Noe 2002; Kristiansen 2008; Langacker 2008) and enabling the interlocutors to understand one another's communicative intent (Lewis 2002; Millikan 2004; Gehlbach 2004; Ganeri 2006). It is a complex phenomenon consisting of perspective setting and perspective taking (Graumann 2002).

The languaging of (spatial) viewing arrangements in discourse can be studied in terms of two cognitive functions associated with linguistic expressions that we term intending and contending. The function of an expression, while intending is to invoke the relevant conceptual category. A conceptual category is a systematic representation of interrelated knowledge systems (Laurence & Margolis 1999; Aarts 2006). For our purposes, a conceptual category is conceived as a cognitive capacitance, storing all possible perspectives of a phenomenon (Merleau-Ponty 1945/2002; Millikan 2000). As a cognitive capacitance, a category is useful in presupposing and entailing large numbers of facts associated with it, because on activation it illuminates a cluster of other categories with which it is associated (Givon 2005). However, intending alone is not enough to language a discourse, since linguistic communication is always embedded in a specific context. We need another cognitive function, whose role is to situate conceptual categories in that context (Zilberman 1938/1988; Langacker 2008). We call this act of relativization *contending*. The function of a linguistic expression, while contending, is to choose a particular perspective in a discourse context. Consider the expression, 'table'. The act of intending, associated with 'table', invokes the corresponding category which includes information about its structural aspects (like shape, size, constituencies etc.) and functional aspects (like dining table, computer table, drawing table etc.). Depending on the communicative situation, one or more of these structural and functional aspects are selected. This selection procedure is guided by the act of contending provided by an expression like 'on' as in 'on the table' in contrast to the 'under the table'. The role of 'on' – while contending – is to delimit the cognitive capacity of a category to window the cognizer's attention to a specific conceptual configuration.[2] Similarly, in an expression like 'tabletop', the categorial capacity of 'top' is delimited by the modifier 'table', when compared with an expression like 'mountaintop'. The act of contending is a complex phenomenon: It is crucial not only in situating the categorial information in a conceptual configuration (such as when we concatenate 'table' with 'top' or 'mountain' with 'top'); but also equally significant in situating the conceptual configuration in a perceptual set up (as in 'this tabletop', 'that mountaintop' etc.). This issue will be discussed later in this paper.

In our view, expressions are not the ready-made items stored in a mental inventory, but "a made-to-order product reconstructed on each occasion for use" in any linguistic construction (Hirtle 2007). The meaning construing capapcity of an expression in a discourse is determined by the way underlying domains of our cognition are grounded and situated by the respective functions associated with an expression – i.e intending and contending. This way of grounding and situating is what we call the *conceptual route* that a cognizer follows - though intuitively - in order to access the communicative intent. In fact, study of the conceptual route is an effort to explore the way conceptualization processes are structured.

Perspectivizing Space

Though the earlier investigations – as is briefed in the introductory section of this paper, led by different researchers – explore how linguistic realization of oriented space reflects its conceptual structure in different linguistic communities, very little has been done to answer how we language relevant representations and correlations at the time of perspectivizing space in discourse. At the level of discourse, puzzles about the relation between language and thought do not end with setting a correlation between linguistic and cognitive categories; we also need to answer how the above mentioned functions work together while licensing inferences that gradually reveal the conceptual route.

The mental locomotions involved in the construction of the conceptual route do not have an explicit linguistic realization. As we will see in the next two sections, the conceptual route is a combination of first-order perspectivizations that are explicitly languaged, and higher-

[1] In stead of using 'perspectivation' as is used by Graumann (2002), we use 'perspectivization' which is borrowed from Taylor (2003).

[2] In case of the example 'on the table', 'the' also acts as contender. However, this issue is not discussed here since it has no direct relevance in this paper.

order inferential tasks that go beyond what is available in the linguistic input alone. We argue that the formation of the conceptual route is determined by the interactions among various intendings and contendings, activated at the time of setting perspective in discourse.

Phrase level Discourse

The claim outlined above is first explored at the phrase level discourse, like (1) and (2); and then elaborated further in discourse larger than the phrase as is exemplified in (3).

(1) tomār ḍān dik-e
you-of right direction-on
On your right

(2) ṭebil-er ḍān dik-e
table-of right direction-on
On the right side of the table

The interpretations of (1) and (2) presume frames based on bodily asymmetries. It is worth noting that these two phrases are grounded and situated in different ways resulting in two different conceptual routes: (1) is interpreted with respect to the *addressee's* ego-centric perspective; (2) is interpreted from the *addresser'*s ego-centric perspective, since the conceptual category 'table' does not have an inbuilt left/right orientation. More specifically, the right side of the table is interpreted with respect to the cognizer's understanding of his/her own physical asymmetry. Here, the intended asymmetry is extrinsic to the conceptualization of tables. In contrast, an extrinsic frame of reference is not required in interpreting a phrase like *on/under the table*, since tables have an inbuilt sense of vertical opposition. The different interpretation of (1) and (2) is a consequence of the interactions holding between intentions and contentions at the time of conceptual integration.

The act of intending associated with the expression *ḍān dik* invokes our background knowledge of asymmetries intrinsic to the human body. This schematic representation of the human body is an abstract and general invariant cognitive standard, applicable to a range of situations. Consequently, in every concrete situation, the abstract standard needs to be identified with a real world entity/situation in order to convey meaning. In case of (1), the body-schema is identified with the body of a person addressed by the genitive form of the second person pronominal form in *Bāŋlā*. The function associated with the genitive case marker, here in this context, is crucial in contending the relation between pronominal (*tomār*) and nominal (*ḍān dik*) forms. The genitive marker functions in situating the intended categorial information in a conceptual configuration, as opposed to the function of the pronoun in situating intended categorial information into a perceptual set up: Since the body indexed by the pronominal form is identified with the body-schema presupposed by the expression *ḍān dik*, the intended orientation in space is now referred with respect to the indexed body in the real world situation. This shows how the act of contending situates the intended categorial information both in conceptual as well as perceptual environments. The situating of communicative intent in conceptual and perceptual worlds often follows different conceptual routes depending on the types of categories invoked by the intenders. This point will be elaborated further with a discussion of example (2).

The interpretation of *ḍān-dik* in (2) also requires the existence of a body in the real world with which the intended body-schema can be identified. However, unlike (1) it does not has an explicit contender whose function can provide schematic support. The function of the table as intender presupposes a frame of reference that does not support the left/right opposition. In order to satisfy the semantic expectancy activated by the expression *ḍān-dik* in (2), the act of contending invokes a frame of reference which has no explicit linguistic realization: This implicit reference indexes the presupposed body-schema with either addressee or addresser. In discourse, addressee and addressor are the 'last resort' to solve any problem related to the act of contending. Therefore, the act of contending first scans for a local solution which is often explicitly available in discourse; otherwise the function invokes contextualization cues as is shown in case of (2). The mechanism of last resort, as it follows from Lewis (2002), lies with "a system of concordant expectations capable of producing coordination at the salient equilibrium".

The above discussion shows how the formation of the conceptual route at the time of perspective taking (which is a part of discourse comprehension) is influenced by the way perspectives are set at the time of discourse production.

Discourse: Sequence of Connected Phrases

So far, we have discussed how the synchronization of intending and contending is crucial in languaging phrase level discourse that invokes a single frame of reference. In this section, we will investigate how different frames of reference are mapped into one another when more than one frame of reference is languaged in discourse, under the assumption that complexities arise at the level of discourse not because of the multiple perspectives set by the different intenders, but because of the inter-translatability of different perspectives.

Consider the example cited in (3), where various categories are intended, and also contended in order to describe a situation.

(3)

āmi	nadī-r	dhār	diye
I	river-of	bank	through
hā̃ṭ-ch-i	āmār	**bām-dik-e**	dhān-er
is walking	my	left-side-on	of-paddy
kṣet	ār	**ḍān-dik-e**	nadī
field	and	right-side-on	river
sāman-e	sūrya	asta	jācche
front-in	sun	setting	is going

I am walking along the river side. The paddy fields are on my left, and the river is on the right. In front, the sun is setting.

The lexical expressions marked bold in (3) are egocentric, in the sense that they are defined in terms of asymmetries intrinsic to the cognizer's/ego's body-schema; and they produce an egocentric perspective of the landscape described by the cognizer. Further, while egocentric perspectives are the only reference frames that are explicitly languaged in the above discourse fragment, an allocentric frame of reference also plays a crucial role. In (3) the cognizer narrates that the sun is setting in front of him/her. From our commonsense knowledge we know that the sun sets in the west. This fact provides an allocentric frame of reference. Due to the interaction between egocentric and allocentric frames of reference, the following inferences are licensed about the landscape described in (3).

(4) (a) The cognizer's motion is west-directed;
(b) The river, which is on the left of the cognizer, is to his/her south;
(c) The paddy field, which is on the right of the cognizer, is to his/her north;

The information enumerated in (4) is not directly stated in (3). Inferencing, on the basis of the commonsense knowledge, is a significant feature of languaging discourse; it is one way to accommodate the commonsense knowledge in discourse interpretation (Stalnaker 1998).

The inferences enumerated in (4) are drawn out of the conceptual route that emerge through the process of designed coordination among the discourse participants on the basis of the functions associated with different expressions in discourse, just in the fashion it happens in case of (1) and (2). What seems to be of significance, here, is that the inter-translatability of ego and allo-centric frames needs to be viewed as an act of contending – mapping different domains of our cognition.

Observations

Translating one perspective into another presupposes two facts: (i) the structural parallelisms intrinsic to the intended categories used in setting two different perspectives; and, (ii) a capacity to interpret the (asymmetric) configuration of one intender with respect to the other. This process of setting up a relational equivalence among different cognitive domains and facilitating higher order inferential tasks is an act of contending, which remains implicit in discourse level languaging. We will consider this type of contending as a *covert* function crucial to higher-order perspectivization.

While setting a correspondence between the ego- and allo-centric construals of space narrated in (3), the first inference (i.e. (4a)) acts as the vantage point with respect to which ego- and allo-centric references are translated into each other. The *relative salience* of (4a) over (4b) and (4c) also suggests a higher order perspectivization process.

Discussion: Perspectivization as a Process

Our analysis of (1-4) above shows that the viewing arrangement in discourse evolves due to the fixation and translation of vantage points. The translation of vantage points is governed by the relative salience that a vantage point has with respect to other vantage points. In discourse, the viewing arrangement is not a fixed arrangement of different isolated vantage points; rather it is an emergent phenomenon evolving gradually due to the shift of attention from one vantage point to other vantage point with every contention/assertion, as is also argued by Fauconnier and Turner (2002). We identify this process as *second order perspectivization*, in contrast to *first order perspectivization* triggered at the time of setting a perspective.

In brief, first order perspectivization activates the relevant frames of reference to construe the context of interpretation. First order perspectivization, then in turn, intends background information necessary for second order perspectivization; whereas, second order perspectivization contends the interactions between the conceptual categories invoked by the process of the first order perspectivization. Therefore, the viewing arrangement at the level of discourse is a consequence of a two tiered complex cognitive process.

Conclusion

The paper views the production and comprehension of spatial viewing arrangements as the synchronization of intentions and contentions in linguistic discourse. In doing so, it investigates the role of two cognitive functions, namely intending and contending (associated with a linguistic expression), in invoking and instantiating conceptual categories. These two processes underlie cognitive capacities like perspective setting and perspective taking at the level of discourse. We have argued for a bi-layered perspectivization process in order to understand the way ego- and allo- centric perspectives interact in discourse to shape the conceptual route.

Acknowledgments

We would like to thank Harish Karnick (IIT Kanpur), Narayanan Srinivasan (CBCS Allahabad), and Leonard Talmy (University of Buffalo) for their comments and suggestions on an earlier version of this paper, presented in the "International Conference on Language and Cognition Interface: State of the Art" at CBCS Allahabad in 2009. We are also indebted to the four anonymous reviewers of CogSci 2010 for their useful comments and suggestions. This study is supported by National Institute of Advanced Studies, India.

References

Aarts, B. (2006). Conceptions of categorization in the history of linguistics. *Language Sciences*, 28, 361-365.

Akman, V. (2000). Rethinking context as a social construct. *Journal of Pragmatics*, 32, 743-759.

Chakrabarti, A. (1992). I Touch What I Saw. *Philosophy and Phenomenological Research*, 52(1), 103-116.

Clark, H.H. (1973). Space, Time, Semantics, and The Child. In T.E. Moore (Ed.), *Cognitive Development and The Acquisition of Language*, New York: Academic Press.

Coventry, K.R., Lynott, D., Cangelosi, A., Monrouxe, L., Joyce, D., & Richardson, C. (2009). Spatial Language, visual attention, and perceptual simulation. *Brain & Language*, doi:10.1016/j.bandl.209.06.0001.

Fauconnier, G. (1981). Pragmatic Functions and Mental Spaces. *Cognition*, 10, 85-88.

Fauconnier, G., & Turner, M. (2002). *The way we think*, New York: Basic Books.

Ganeri, J. (2006). *Artha: Meaning*, Oxford: Oxford University Press.

Gelbach, H. (2004). A New Perspective on Perspective Taking: A Multidimensional Approach to Conceptualizing an Aptitude. *Educational Psychology Review*, 16(3), 207-234.

Gibbs, Jr. R.W. (2005). *Embodiment and Cognitive Science*, Cambridge: Cambridge University Press.

Givon, T. (2005). *Context as Other Minds: The Pragmatics of Sociality, Cognition and Communication*. Amsterdam: John Benjamins Publishing Co.

Graumann, C.F. (2002). Explicit and implicit perspectivity. In C.F. Graumann & W. Kallmeyer (Eds.), *Perspective and Perspectivation in Discourse*. Amsterdam: John Benjamins Publishing Co.

Heine, B. (1989). Adpositions in African Languages. *Linguistique Africaine*, 2, 77-127.

Heine, B., Ulrike, C., & Hunnemeyer, F. (1991). *Grammaticalization: A Conceptual Framework*, Chicago: University of Chicago Press.

Hirtle, W. (2007). *Language in the mind: An introduction to Guillaume's theory*. Montreal/Kingston: McGill-Queen's University Press.

Karmakar, S. (2009). *Temporal Ordering in Discourse: A Study of Baŋla*. Unpublished PhD Dissertation, Kanpur: Indian Institute of Technology.

Kristiansen, G. (2008). Idealized cultural models: The group as a variable in the development of cognitive schemata. In R.M. Frank, R. Driven, T. Ziemke, & E. Bernardez (Eds.), *Body, Language and Mind: Socio-Cultural Situatedness*, Vol. 2, Berlin: Mouton De Gruyter.

Landau, B., & Hoffman, J.E. (2005). Parallels between spatial cognition and spatial language: Evidence from William syndrome. *Journal of Memory and Language*, 53, 163-185.

Laurence, S., & Margolis, E. (1999). Concepts and Cognitive Science. In E. Margolis & S. Laurence (Eds.), *Concepts*, Massachusetts: The MIT Press.

Levinson, S.C. (1996). Language and Space. *Annual Review of Anthropology*, 25, 353-382.

Levinson, S.C. (2003). *Space in Language and Cognition*. Cambridge: Cambridge University Press.

Levinson, S.C., & Wilkins, D.P. (2006). The background to the study of the language of space. In S.C. Levinson & D.P. Wilkins (Eds.), *Grammars of Space*. Cambridge: Cambridge University Press.

Lewis, D. (2002). *Convention*. Oxford: Blackwell Publishers.

Langacker, R.W. (2008). *Cognitive Grammar*. New York: Oxford University Press.

Majid, A., Bowerman, M., Kita, S., Haun, D.B.M., & Levinson, S.C. (2004). Can language restructure cognition? The case for space. *Trends in Cognitive Sciences*, 8(3), 108-114.

Mead, G.H. (1934). *Mind, Self, & Society*. In W.C. Morris (Ed.), *Works of George Herbert Mead*, Vol. I. Chicago: The University of Chicago Press.

Mead, G.H. (1938). The Philosophy of the Act. In W.C. Morris (Ed.), *Works of George Herbert Mead*, Vol. III. Chicago: The University of Chicago Press.

Merleau-Ponty, M. (1945/2002). *Phenomenology of Perception*, Translated by Colin Smith, London: Routledge.

Millikan, Ruth G. (2004). *Varieties of Meaning*, Massachussetts: MIT Press.

Moore, A.W. (1997). *Points of View*, Oxford: Oxford University Press.

Neggers, S.F.W., Van der Lubbe, R.H.J., Ramsey, N.F., & Postma, A. (2006). Interactions between ego- and allocentric neuronal representations of space. *NeuroImage*, 31, 320-331.

Noe, A. (2002). Is Perspectival Self-Consciousness Non-Conceptual? *The Philosophical Quarterly*, 52(207), 185-194.

Piaget, J. & Inhelder, B. (1948/1977). The Child's Conception of Space. In H.E. Gruber & J.J. Voneche (Eds.), *The Essential Piaget.* New York: Basic Books, Inc.

Rescoria, M. (2007). Convention. In Edward N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Sep. 6, 2007), http://plato.stanford.edu/entries/convention/.

Rochberg-Halton, E. (1982). Situation, Structure, and the Context of Meaning. *The Sociological Quarterly* 23(4), 455-476.

Stalnaker, R. (2002). Common Ground. *Linguistics and Philosophy*, 25, 701-721.

Stalnaker, R. (1998). On the Representation of Context. *Journal of Logic, Language and Information*, 7(1), 3-19.

Taylor, John R. (2003). *Linguistic Categorization*. Oxford: Oxford University Press.

Egocentric and allocentric spatial references in children with Cerebral Palsy

Laura Barca (laura.barca@istc.cnr.it)
Neuroscience and Neurorehabilitation Department
Children's Hospital Bambino Gesù - IRCCS
Lungomare G. Marconi 36, 00058 - Santa Marinella (Rome), Italy

Giovanni Pezzulo (giovanni.pezzulo@cnr.it)
Istituto di Linguistica Computazionale 'Antonio Zampolli', CNR, Pisa, Italy
Istituto di Scienze e Tecnologie della Cognizione, CNR – Roma, Italy

Enrico Castelli (enrico.castelli@opbg.net)
Neuroscience and Neurorehabilitation Department
Children's Hospital Bambino Gesù - IRCCS
Lungomare G. Marconi 36, 00058 - Santa Marinella (Rome), Italy

Abstract

Spatial memory is supported by multiple parallel representations of the environment. Egocentric perspective (body-centered) and allocentric representations (object-centered) are integrated to allow correct interaction with the world. According to Milner and Goodale (1995, 2008), the action-related dorsal system is specialized for location of objects in space and visuo-motor integration, and uses an egocentric frame of reference. The perception-related ventral system is specialized for categorical recognition of objects and forms, and supports an allocentric frame of reference. Here we use a Distance Judgment Task to explore the use of different spatial frames in children with Cerebral Palsy (CP). Following the dorsal stream vulnerability hypothesis (Atkinson et al., 2007) children with CP might have more difficulties in egocentric judgments and in the processing of peri-personal space than controls. No significant difference emerged between CP children and controls in allocentric judgments, whereas performance was worse in egocentric judgment, indicating inefficient use of the body-centered representations. **Keywords:** Egocentric-allocentric spatial references; Distance Judgment task; Cerebral Palsy (CP).

Introduction

Humans are provided with different reference systems to code the environment and its physical attributes. For example, if we have to specify the location of an object we can make use of different frames of reference: we can define its position with respect to our body (egocentric frame) or we can refer to other objects in the environment or the environment itself (allocentric frame). Egocentric coordinates are based on the organism's position, and then linked to the specific perspective under which spatial information has been processed. Hence, these representations are particularly relevant in action planning and motor control in near space, when there is a direct interaction between body and objects. Egocentric frames have been described in relation to the different body part they are based on, such as head-centered, eye-centered, and arm-centered (Colby, 1998). Allocentric, or object-centered frames, are external to the organism and usually centered on objects in the environment. Such coding of space has an important role in the processing of far space when objects are out of reach. Among allocentric representation, distinctions can be made when the point of reference is centered on an object of interest (object-centered) or on the environment (e.g., room-centered) (Colby, 1998). The information derived by egocentric and allocentric maps is usually integrated to allow proficient spatial processing (Burgess, 2006). However, some tasks rely more on one frame than the other. For example, pointing to a location in space within arm reach or grasping an object are likely accomplished within an egocentric framework, whereas defining the fastest route between two destinations is likely to involve an allocentric frame. Overall, selection of what spatial frame(s) of reference to use is highly action-specific.

A number of studies showed that several regions of the cerebral cortex subserved functions involved in spatial processing, having a reach network of reciprocal connections and link with subcortical structures. In the fMRI study by Zaehle et al., (2007) participants performed a spatial judgment task based on verbal instructions. They have to define the spatial relations between different objects (allocentric condition), or the position of objects with respect to the participants (egocentric condition). A fronto-parietal network was involved in both egocentric and allocentric judgments (e.g., superior occipital gyrus, medial portion of superior parietal cortex, superior frontal gyri bilaterally), but partly separated networks mediate different spatial coding strategies. While egocentric spatial coding revealed activation mainly within the medial parts of the posterior superior parietal lobe, the use of the allocentric reference frame revealed activation in right parietal lobe, bilateral ventrolateral occipito-temporal cortex and bilateral hippocampal formation. There is also increasing evidence of the critical role of connecting circuits, and the vestibular

system (Paillard, 1991). Dysfunction of egocentric frames appeared to be associated with damage in premotor cortex involving frontal eye field, whereas allocentric impairments are linked to lesions in more ventral regions near the parahippocampal gyrus (for a recent review see Grimsen, Hildebrandt, and Fahle, 2008). Patients with visual form agnosia, which is associated with ventral stream damage, have been reported to have selective impairments in allocentric judgments of spatial coding, with spared egocentric processing (Carey, Dijkerman, and Milner, 2009; Carey et al., 2006; Dijkerman, Milner, and Carey, 1998). The study from Galati et al. (2000) showed a different lateralization of spatial coding networks across the cerebral hemispheres, with body-centered frames more lateralized in the right hemisphere. In line with this evidence are the neuropsychological data from Iachini et al., (2009), were patients with right parietal lesions failed in egocentric but not allocentric distance judgments, whereas those with left parietal damages have difficulties in both frames of reference.

From a developmental point of view, the body is the primary available spatial code for the infant and allocentric references develops later in life, having a longer maturational trajectory. However, Nardini et al., (2006) suggests that object-centered coordinates and the integration between different coding systems occur earlier than previously thought. Using a task in which children have to recall the location of hidden toys within an array, they showed that spatial representations based on the environment (allocentric frames) develop between years three and six. Such experimental paradigm has been applied also to the study of spatial localization in clinical population (Nardini et al., 2008), however it might have limited application to patients with motor and deambulation deficits as one of its key components is the 'subject-move' condition.

Here we study spatial cognition in children with Cerebral Palsy exploring their use of different spatial frames of references. CP is defined as ‘‘a group of permanent disorders of the development of movement and posture, causing activity limitation, that are attributed to non progressive disturbances that occurred in the developing fetal or infant brain’’ (Rosenbaum et al., 2007). In the framework of visual cognition, it has been shown that the dorsal visual system (with its connections to parietal, frontal and hippocampal areas and its relations to the egocentric frame of reference) is more vulnerable to insult occurring early in life than ventral visual system. Children with hemiplegic CP (e.g., a motor deficit characterized by paralysis of the arm, leg, and trunk on the same side of the body) perform significantly worse than controls in dorsal stream tasks (e.g., motion coherence task) than ventral stream tasks (e.g., form coherence task). While a subgroup of hemiplegic children performed better than the normal median level for their age on the form coherence task, all the hemiplegic children performed close to the median level, or worse, for their age on the motion coherence task (Gunn et al., 2002). CP children often presents with visual disorders comprising ophthalmological abnormalities and impairments in higher visuofunctional skills, which are considered a clinical manifestation of dysfunctions of visual associative areas of the dorsal visual path (Barca et al., 2010). The vulnerability of dorsal stream has been shown also in healthy children born preterm with no sign of neurological deficit, visual disturbances, or cognitive and motor deficits (Santos et al., 2008). Such findings suggest that the number of gestational weeks has an important influence on the normal development of visual cognition. Linking the vulnerability of the dorsal stream with the association of this brain regions with egocentric spatial representations, one can assume that mainly egocentric representations would be impaired in spatial processing of CP.

The aim of this study was to investigate the impact that brain injuries occurring early in life (e.g., prenatal or perinatal period) exert on the development of the different coordinate systems used for the coding of space, by studying the performance of hemiplegic CP children on a distance judgment task. Specifically, our main research question is: are egocentric (self-referred) and allocentric (object-referred) distance judgments similarly impaired? The dorsal/ventral distinction has been recently extended to spatial processing, suggesting that the dorsal circuit provides egocentric coding of space for motor control and action planning whereas the ventral circuit is tuned with allocentric coding of space (Medina et al. 2009). Hence, the dorsal stream vulnerability hypothesis would predict children with CP to have more difficulties in egocentric judgments and in the processing of peri-personal space than age matched controls. However, given the precocity of the cerebral insult, they might develop compensatory mechanisms that allow to correctly processing spatial representations, as has been shown in patients with idiopathic cervical dystonia (Ploner et al., 2005). Neuropsychological adult literature provides evidence of a link between dorsal stream lesions and impairments in egocentric judgments (Berryhill, Fendrich and Olson, 2009). However, patients with parietal damage having the opposite deficit (i.e., allocentric impairments with spared egocentric references) have also been reported (Carey et al., 2006), thus questioning the direct link between parietal lesions and body-centered perspective.

To test the prediction of a major impairment in egocentric than allocentric representations of space in children with CP, we conducted a behavioral study in which egocentric and allocentric stimulus coordinates were varied in order to individuate their contribution in making spatial judgments. The procedure of the experiment was motivated by the work of Iachini and colleagues (Iachini, et al., 2006; Iachini et al., 2009), as they were able to consistently and effectively induce a differential involvement of spatial coding systems

with such procedure. However, several changes (which will be described in the following section) have been introduced to make the task feasible for a pediatric clinical population.

Method and Materials

Participants

A group of seven children with CP participated in the study. They were 3 male and 4 female, with mean chronological age of 7 years (range 5-9 years), with no spatial neglect, language or general intellectual impairments. Four child presents with Hemiplegia, and 3 with Diplegia. A control group of 5 children with typical development was used for comparison. Children of this group had no history of visual, motor or cognitive delay, and mean chronological age of 10 years (range 8-12 years),

Neuropsychological assessment

General cognitive level was assessed with the Raven's Colored Progressive Matrices, CPM (Raven and Raven, 1986), which has been recently shown to be a valid tool in the assessment of cognitive functioning in CP (Pueyo et al., 2008). To assess visuoperceptual and visuomotor integration skills we used the Developmental Test of Visual Perception, DTVP (Hammill, Pearson, and Voress, 1994). The Corsi block-tapping task (Corsi, 1972; Milner, 1971) was used as a measure of visuospatial working memory. Parents of the controls group fulfilled the questionnaire of Houliston et al., (1999), adapted to Italian and used as a screening measure of children's neurovisual behavior (e.g., questions regards child's ability to recognize objects and faces, finding way in home, distinguishing line from steps and the perception of motion).

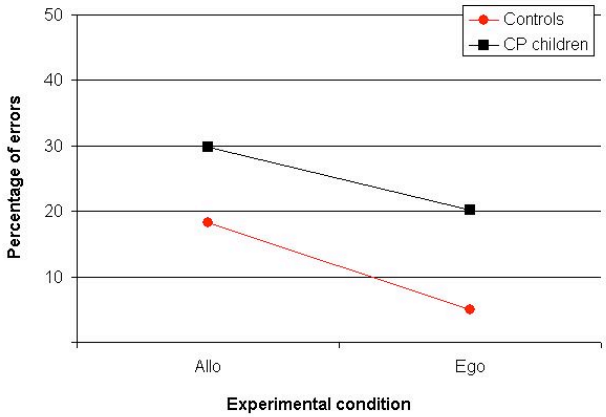
Experimental task

A Distance Judgment Task, adapted from Iachini et al. (Iachini et al., 2006; Iachini et al., 2009), has been used. Children were presented with triads of 3D objects in peri-personal space (within arm reach) and were asked to give egocentric and allocentric judgments. Materials comprised eighteen graspable objects divided in triads. They were geometrical shapes with different colors (e.g., cube, pyramid, and wheel), animals (e.g., duck, rabbit, and horse), vehicles (e.g., car, helicopter, and airplane) and everyday objects (i.e., key, cork and clothes-peg). Objects within triads had similar size. Each triad was spatially arranged so that distance between objects was clearly discriminable, and the amount of metric difficulty was the same for egocentric and allocentric judgments. Participants sat at 30 cm from the edge of the desk. Each triad was placed centrally on the desk and with respect to the participants' mid-sagittal plane. A white cardboard measuring 50 x 50 cm was used to arrange stimuli. Children were instructed to study and memorize the position of the objects for 30 seconds. Then the objects were covered with cups and data acquisition started. There were eight judgments for each triads: two egocentric questions (''Which object was closer/farther to/from you?"), two allocentric questions (''Which object was closer/farther to/from the Cube (target)?"), and four distractors questions about objects' shapes and colors. For each judgment, accuracy was coded as dummy variable (1 = correct, 0 = incorrect) and the mean accuracy by subject was computed. The order of presentation of the questions was first randomized and then balanced across subjects. Before start with the session, the examiner spent some time to familiarize with the child and explained the nature of the experiment to the parents in order to have their consent.

Results

Neuropsychological assessment. CP children did not present cognitive delay as measured with the CPM (the cut-off point for clinically significant impairment was the 25th percentile) and have a visuospatial memory span adequate to their age. Although some variation emerged among patients, they did not present marked deficits in visuoperceptual and visuomotor integration skills as measured by the DTVP.
Regarding the controls group, parents' questionnaire did not report any difficulties in visuoperceptual or visuospatial behavior (e.g., problems with shapes, objects and faces recognition, simultaneous perception, perception of movement, colors perception, and orientation).
Distance Judgment task. Patients and controls performance at the Distance Judgment Task are presented in Figure 1. Chi-square test was used to evaluate significance level of observed differences.

Figure 1: Results Distance Judgment task



Overall, both groups of children made few errors in completing the task (no child exceeded the chance threshold). The task resulted more difficult for CP than controls in that they were less accurate (12% and 25% errors, respectively in controls and patients). At the group

level, controls have a Frame effect, with nearly no errors in responding to egocentric vs. 18% errors in allocentric condition (Chi-square = 5.2, p <.05). Differently, in CP children the egocentric-allocentric difference was less marked (20% and 30% of errors, respectively) and did not reach the significance level (Chi-square = 2, p >.1). A comparison between the two groups confirmed that CP children were less accurate than controls in responding to egocentric questions (Chi-square = 6.8, p <.01) whereas no differences emerged in allocentric questions (Chi-square < 2.4, p >.1).

Discussion

In the present study, children with Cerebral Palsy were asked to judge the position of graspable objects with respect to their body (egocentric condition) or with respect to landmarks (other objects) in the environment (allocentric condition). The first evidence is that such paradigm proved to be feasible to study spatial cognition in normally developing children and children with CP. Such paradigm, indeed, has been previously used with adult population (Iachini et al., 2006; Iachini et al., 2009) and this is the first time it is applied to developmental age.

Typically developing children were less accurate in recalling the position of objects using allocentric spatial coordinates then when using body-centered coordinates, confirming the predominance of egocentric coding in the developmental trend of spatial cognition (see Nardini et al., 2006). This was not the case for the group of children with CP. Indeed, they have similar performance when using egocentric or allocentric coding. Given that no differences emerged between groups with respect to the allocentric judgments, results suggest a specific deficit in using body-centered coordinates. One might argue that such difficulty reflects a deficit in visuospatial memory. However, such explanation is unlikely given that our sample of CP children have adequate score in visuo-spatial working memory task. Additionally, there is no reason to believe that (if present) a similar limitation would selectively affect egocentric vs. allocentric judgments. CP children's performance reflects preservation of categorical coding within the ventral stream, despite a loss of coordinate coding which is consistent with the hypothesis of dorsal stream vulnerability in such population (Atkinson et al., 2007; Fazzi et al., 2007). Deficits in spatial perception are usually matched with deficit in generating spatially directed actions. Patients have been described to neglect stimuli presented in peripersonal space and correctly perceive them when located in extrapersonal space, as well as the opposite pattern (Bisiach, Perani, Vallar, and Berti, 1986). Thus, information about how patients perceive the environment both in near and far space has implications in rehabilitation treatments of visuoperceptual and visuospatial impairments. We believe that this is an important issue that needs to be further explored in impaired population in developmental age. Findings of the study are preliminary as more participants are needed to broaden our conclusions.

The extent to which our results generalize to other aspect of spatial cognition and other types of CP are important further questions. Nico and Daprati (2009) propose a distinction between two separate egocentric mechanisms: one allowing construction of the immediate point of view and the other extracting a required perspective within a mental representation. This, for example, should be further addressed in our sample of patients. Moreover, Cerebral Palsy is an umbrella term which comprises different types of motor limitations which differently affect how children experience the external world and create internal representation of it. Children of our study can be considered 'high functional' cerebral palsied children in that they do not present language delay, general intellectual impairments and marked deficits in visuoperceptual and visuomotor integration skills. Notwithstanding such limitations, we believe the study provides interesting findings relevant for the field of spatial cognition in impaired population in developmental age.

To summarize, children with CP were impaired in a distance judgment task: Allocentric spatial representations were present even in the context of impaired egocentric coding. Further studies are needed to tackle this issue and to understand how a unitary perception of the world is achieved from its multiple representations.

Acknowledgments

The research leading to these results has been partially founded by the European's Community Seventh Framework Programme under the grant agreement no. PERG02-GA-2007-224919 to Laura Barca.

References

Atkinson, J. and Braddick, O. (2007). Visual and visuocognitive development in children born very prematurely. *Progress in Brain Research, 164*,123-149.

Barca, L., Cappelli, F.R., Di Giulio, P., Staccioli, S., and Castelli, E. (2010). Outpatient assessment of neurovisual functions in children with Cerebral Palsy. *Research in Developmental Disabilities*, 31, 488-495.

Berryhill, M. E., Fendrich, R., and Olson, I. R. (2009). Impaired distance perception and size constancy following bilateral occipitoparietal damage. *Experimental Brain Research, 194,* 381-393.

Bisiach, E., Perani, D., Vallar, G., and Berti, A. (1986). Unilateral neglect: personal and extra-personal. *Neurospychologia, 24*, 759-767.

Burgess, N. (2006). Spatial memory: how egocentric and allocentric combine. *Trends Cogn Sci,* 10, 551-557

Carey, D. P., Dijkerman, H. C., and Milner, A. D. (2009). Pointing to two imaginary targets at the same time: bimanual allocentric and egocentric localization in visual form agnostic. *Neuropsychologia, 47*, 1469-1475.

Carey, D. P., Dijkerman, H. C., Murphy, K. J., Goodale, M. A., and Milner, A. D. (2006). Pointing to places and spaces in a patient with visual form agnosia. *Neuropsychologia, 44*, 1584-1594.

Colby, C. (1998). Action-oriented spatial reference frames in cortex. *Neuron,* 20, 15-24

Corsi, P.M. (1972). Human memory and the medial temporal region of the brain. Unpublished doctoral dissertation. McGill University, Montreal, Canada.

Dijkerman, H. C., Milner, A. D., and Carey, D. P. (1998). Grasping spatial relationships: failure to demonstrate allocentric visual coding in a patient with visual form agnosia. *Conscious Cognition, 7*, 424-437.

Fazzi, E., Signorini, S. G., Bova, S. M., Piana, R. L., Ondei, P., Bertone, C., Misefari, W., and Bianchi, P. E. (2007). Spectrum of visual disorders in children with cerebral visual impairment. *Journal of Child Neurology, 22*, 294-301.

Galati, G., Lobel, E., Vallar, G., Berthoz, A., Pizzamiglio, L., Le Bihan, D. (2000). The neuronal basis of egocentric and allocentric coding of space in humans: a functional magnetic resonance study. *Experimental Brain Research*, *133*, 156-164.

Grimsen, C., Hildebrandt, H., and Fahle, M. (2008). Dissociation of egocentric and allocentric coding of space in visual search after right middle cerebral artery stroke. *Neuropsychologia, 46,* 902-914.

Gunn, A., Cory, E., Atkinson, J., Braddick, O., Wattam-Bell, J., Guzzetta, A., and Cioni, G. (2002). Dorsal and ventral stream sensitivity in normal development and hemiplegia. *Neuroreport, 13,* 843-847.

Hammill, D., Pearson, N., and Voress, J. (1994). *Developmental test of visual perception*; Edizione italiana: TPV. Trento: Edizioni Centro Studi Erikson.

Houliston, M. J., Taguri, A. H., Dutton, G. N., Hajivassiliou, C., and Young, D. G. (1999). Evidence of cognitive visual problems in children with hydrocephalus: a structured clinical history taking strategy. *Developmental Medicine and Child Neurology, 41*, 298-306.

Iachini, T. and Ruggiero, G. (2006). Egocentric and allocentric frame of reference: a direct measure. *Cognitive Processes, 7,* S126-S127.

Iachini, T., Ruggiero, G., Conson, M., and Trojano, L. (2009). Lateralization of egocentric and allocentric spatial processing after parietal brain lesions. *Brain and Cognition, 69*, 514-520.

Medina J, Kannan V, Pawlak MA, Kleinman JT, Newhart M, Davis C, Heidler-Gary JE, Herskovits EH, Hillis AE. (2009). Neural substrates of visuospatial processing in distinct reference frames: evidence from unilateral spatial neglect. *Journal of Cognitive Neuroscience, 21*, 2073-84.

Milner, A. D. and Goodale, M. A. (1995). *The visual brain in action*. Oxford University Press, Oxford.

Milner, A.D. and Goodale, M.A. (2008). Two visual systems re-viewed. *Neuropsychologia*, 46, 774-785

Milner, B. (1971). Interhemispheric differences in the localization of psychological processes in man. *British Medical Bulletin*, 27, 272-277

Nardini, M., Atkinson, J., Braddick, O., and Burgess, N. (2008). Developmental trajectories for spatial frames of reference in williams syndrome. *Developmental Science, 11*, 583-595.

Nardini, M., Burgess, N., Breckenridge, K., and Atkinson, J. (2006). Differential developmental trajectories for egocentric, environmental and intrinsic frames of reference in spatial memory. *Cognition, 101*, 153-172.

Nico D., Daprati E. (2009). The egocentric reference for visual exploration and orientation. *Brain and Cognition, 69,* 227-235.

Paillard, J. (1991). Motor and representational framing of space. In J. Paillard (Ed.), *Brain and space*. Oxford, UK: Oxford University Press.

Ploner, C. J., Stenz, U., Fassdorf, K., and Arnold, G. (2005). Egocentric and allocentric spatial memory in idiopathic cervical dystonia. *Neurology, 64*, 1733-1738.

Pueyo, R., Junquè, C., Vendrell, O., Narberhaus, A., & Segarra, D. (2008). Raven's Coloured Progressive Matrices as a measure of cognitive functioning in Cerebral Palsy. *Journal of Intellectual Disability Research, 52,* 437-445.

Raven, C. and Raven (1986). *Coloured Progressive Matrices*. Giunti O.S.

Rosenbaum, P., Paneth, N., Leviton, A., Goldstein, M., Bax, M., Damiano, D., et al. (2007). A report: The definition and classification of cerebral palsy. April 2006. *Developmental Medicine and Child Neurology. Supplement, 109*, 8–14.

Santos, A., Duret, M., Mancini, J., Gire, C., and Deruelle, C. (2008). Preterm birth affects dorsal-stream functioning even after age 6. *Brain and Cognition, 69*, 490-494.

Zaehle, T., Jordan, K., Wüstenberg, T., Baudewig, J., Dechent, P., and Mast, F. W. (2007). The neural basis of the egocentric and allocentric spatial frame of reference. *Brain Research, 1137*, 92-103.

Vocabulary Spurt: Are Infants full of Zipf?

Julien Mayor[a,b] and **Kim Plunkett**[b]
[a] Basque Center on Cognition, Brain and Language, San Sebastian, Spain
[b] Department of Experimental Psychology, University of Oxford, United Kingdom
(j.mayor@bcbl.eu, kim.plunkett@psy.ox.ac.uk)

Abstract

Infants do not learn words at a constant rate. During the second year of life, a dramatic increase in the speed of word learning is observed. Different mechanisms explaining this vocabulary spurt have been proposed, either through endogenous factors such as learning capacity or exogenous factors, such as frequency of word usage. We demonstrate that occurrence statistics alone is not sufficient to explain the acceleration in vocabulary growth, discuss other potential exogenous contributions such as phonological complexity and suggest that a change in word learning capacities is necessary. A model implementing an increased ease of learning is introduced and illustrates this endogenous approach by replicating the non-linear vocabulary growth characteristics of language acquisition.

Keywords: vocabulary spurt; mathematical modelling; word learning; learning mechanisms; Zipf's law; endogenous vs. exogenous factors

Introduction

Around their first birthday infants utter their first word and by their second birthday they learn on average one new word every waking hour. Between 18 and 24 months of age, an abrupt change in the speed of word acquisition is observed, called the vocabulary spurt or naming explosion (Bloom, 1973)[1]. Two types of theories have been offered to explain the vocabulary spurt. One suggests that the vocabulary spurt corresponds to representational and/or maturational changes in the infant's brain. For example, researchers have suggested that infants start acquiring words at a faster pace when they understand that words refer to things and/or that things have names. On this view, the vocabulary spurt corresponds to a *naming insight* (Dore, Franklin, Miller, & Ramer, 1976; Reznick & Goldfield, 1992; McShane, 1979; Kamhi, 1986). Alternatively, word learning occurs at a faster pace when object concepts and categories become more detailed and refined (Bates, Benigni, Bretherton, Camaioni, & Volterra, 1979; Gopnik & Meltzoff, 1987; Nazzi & Bertoncini, 2003). Other researchers have proposed that the spurt corresponds to linguistic refinements such as word segmentation (Plunkett, 1993), word retrieval capacities (Dapretto & Bjork, 2000), improvements in social cognition (Ninio, 1995) or changes in hemispheric specialisation (Mills, Coffey-Corina, & Neville, 1993). All of these hypotheses share the assumption that the vocabulary spurt reflects endogenous changes in the infant.

A second, contrasting, hypothesis has recently been introduced by McMurray (2007). He argued that under the reasonable assumptions that (i) words are learnt in parallel and (ii) some words are easier to learn than most words, a vocabulary spurt is inevitable and that "this distribution in difficulty derives from many factors, including frequency, phonology, syntax, the child's capabilities, and the contexts where words appear."(McMurray, 2007, p.631). Invoking the central limit theorem, he suggested that the individual contributions of the different factors sum to a Gaussian distribution of word difficulty. Later, using the logarithm of utterance statistics as a proxy for word difficulty, he showed that a time-to-acquisition growth curve yields a pattern of vocabulary development typical of infants during their second year. On the basis of this finding, he claimed that "acceleration in vocabulary growth could arise from occurrence statistics alone" (McMurray, 2007, p.631).

Our aim is to clarify the origin of this non-linear increase in the speed of lexical acquisition; whether this transition is the result of a change in the infant's mental representations or brain organisation (endogenous factors), or caused by the statistical nature of the input, such as phonological complexity or the frequency of word usage (exogenous factors). We show mathematically that word frequency cannot alone explain the acceleration in vocabulary growth. This demonstration also fits well with empirical findings that word frequency is not an entirely reliable proxy for word difficulty (Huttenlocher, 1991; Goodman, Dale, & Li, 2008). Instead, we suggest that changes in the infant's learning capacity are required to display the non-linear growth in the speed of word acquisition. These changes, such as the emergence of fast mapping (Carey & Bartlett, 1978), provide the basis for the unique learning capacities displayed late in the second year of human life.

Statement of the problem

For expository purposes, we make three simplifying assumptions; (i) infants only learn words when hearing them, (ii) word occurrence statistics follows Zipf's law and (iii) all words are equally difficult to learn. If these three criteria are satisfied, we demonstrate that vocabulary growth will be linear, unless a change in learning capacity takes place (as a function of time or as a function of the number of words al-

[1] We will use the terminology "vocabulary spurt" throughout the manuscript in the sense of a *supra-linear lexical growth*, characterised by slow learning in early development, followed by an increase in the speed of word learning later on. Even though an increase in the speed of word learning in the first years of human life is not questioned, its mathematical description is debated; should it possess a clear inflection point or is there a more gradual increase throughout early development, as suggested by Ganger and Brent (2004)? For the scope of the present manuscript, we use the term "vocabulary spurt" in its general – and milder – interpretation, whereby infants display slow initial learning followed by a faster rate of word learning, contrasting with a linear increase in which the rate of word learning would be constant during life.

ready present in the lexicon[2]). In other words, a change in learning capacity is a necessary pre-requisite to drive a non-linearity in vocabulary growth. We justify this claim by both analytical considerations and through simulations. Later, we will show that (i) the assumption of online learning can be relaxed, (ii) that speech corpora used with real infants follow the same behaviour as Zipf's law and we will suggest that (iii) phonological complexity of early words do not seem to play a prominent role in shaping the vocabulary spurt. We will suggest, therefore, that a change in the infants' learning capacities is driving the naming explosion.

Let us first justify our initial assumptions. First, we argue that infants learn words when they are confronted with them and not by processing words off-line after accumulating evidence. Carey and Bartlett (1978) introduced the idea that infants are able to "fast map", whereby infants demonstrate rapid mastery of the appropriate use of labels after a limited number of learning opportunities. Evidence of the infant's ability to learn a new word after limited exposure was also explored by Woodward, Markman, and Fitzsimmons (1994), suggesting that novel words can be retained at least 24 hours after the infants have been exposed to them only 9 times, even for infants as young as 13 months of age. More recent evidence based on infant-caregiver interactions showed that the naming event needs to occur at the right moment in time when the infant is attending to the named object to be successful (Yu, Smith, & Pereira, 2008)[3]. These findings provide strong support for the claim that infants perform on-line word learning when exposed to them. Consequently, if infants only engage in online word learning, the raw statistics of word usage should be exploited and not, as in McMurray (2007), a logarithmic transformation of the word frequencies (a further comparison of our approach to McMurray, 2007, is discussed later). Moreover, we will show that even a relaxation of the assumption of online learning cannot explain an accelerated vocabulary growth.

Second, we adopt the perspective that infants are exposed to a distribution of word frequencies approaching Zipf's law (Zipf, 1949), which states that, from any substantial corpus, the frequency of a word is inversely proportional to its rank. For example, the most frequent word is used twice as much as the second most frequent word and three times more often than the third most frequent word. A broad range of evidence suggests that spoken language essentially follows a Zipf distribution of word usage (Miller & Chomsky, 1963; Zipf, 1935; Beier, 1965; Dahl, 1979; Altmann, 2002). We will show, in a model with constant learning capacity and exposed to a corpus of speech used with real infants, that lexical growth fails to exhibit the characteristics of a vocabulary spurt, even when the utterance statistics deviate slightly from Zipf's law.

[2]Mitchell and McMurray (2009) have shown that leveraged learning—the fact that knowledge of some words helps with the learning of others—does not create acceleration in word learning.

[3]In a recent experiment, Smith and Yu (2008) showed that infants were able to use cross-situational statistics to learn novel words. It remains to be shown, however, if these effects extend to longer time windows than used in the experiment, consisting of multiple presentations of each word-referent pair over the course of 4 minutes.

Analytical considerations

On average, an infant hears a word i having a frequency $f(i)$ within a time window $t(i) = 1/f(i)$. For example, a word uttered twice an hour will be heard on average every 30 minutes and a word uttered 4 times a month will be uttered every week or so. As a consequence, and to a first approximation, the time $T(i)$ to acquire a word i is inversely proportional to its frequency $T(i) \propto 1/f(i)$. The constant of proportionality depends on the number of times a word needs to be heard with respect to the threshold for learning it. Zipf's law states that, from any substantial corpus, the frequency of a word is inversely proportional to its rank: $f(i) \propto 1/i$. This predicts a linear distribution of time to acquisition; $T(i) \propto i$, which in turn predicts a linear increase in the size of the lexicon. The (constant) speed at which infants increment their lexicon size would then be proportional to their (fixed) learning capacity, as defined by the number of times they need to hear a word in order to add it to their lexicon. In real word learning situations, words do not follow Zipf's law deterministically. However, the fluctuations in everyday interactions can be modelled by drawing words probabilistically from Zipf's distribution. Since analytical calculations become increasingly complex, we simulate this process in a stochastic model.

Simulation results

Fig. 1 displays simulations using raw frequencies of word usage from Zipf's distribution. As in McMurray (2007), a knowledge level is associated with each word and is incremented with each presentation. When this crosses a threshold, the word is learnt. The model reveals a regular increase of word acquisition, the absence of an early, slow learning phase and no inflection point in word learning; in other words, the absence of a vocabulary spurt. The different curves on

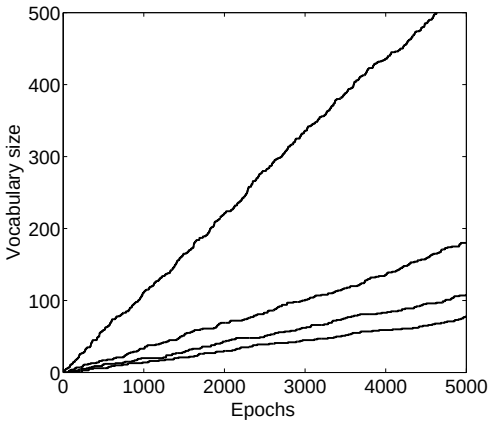


Figure 1: Vocabulary size as a function of time when a model with constant learning capacity is presented with a Zipf distribution of word usage (the different curves correspond to different numbers of words uttered per epoch).

Fig. 1 correspond to different (constant) learning capacities, i.e., the number of presentations needed to acquire the word in the lexicon. Steeper curves correspond to better learning capacities. Note that, in the model, the absolute number of words uttered in an epoch of exposure also modulates the slope of the learning curves. Similar curves may correspond either to a proficient learner confronted to a low number of words or to a slower learner presented with a higher number of words in any time window. All combinations of learning capacities and absolute number of words uttered per epoch lead to linear increases in the lexicon size.

As a further control we ran simulations with actual word frequencies extracted from the CHILDES Parental Corpus, made out of the following 27 corpora; Bates, Belfast, Bernstei, Bliss, Bloom, Brown, Clark, Cornell, Demetras, Fletcher, Gatherco, Hall, Higginso, Howe, Kuczaj, Macboys (MacWhinney), Macros (MacWhinney), Peters, Post, Sachs, Snow, Suppes, Valian, Vanhout, Vankleec, Warren, and Wells. The Parental Corpus consists of 2.6 million word tokens (about 24,000 word types), and is a representative sample of the speech to which children are typically exposed (MacWhinney, 1991; Li & Shirai, 2000). Fig. 2 shows that for differing numbers of presentations needed to acquire a word (different learning capacities), vocabulary growth lacks both the long latent period of slow learning and subsequent rapid increase characteristics of the vocabulary spurt. Instead, the speed of acquisition is reduced after having learnt about 100 words, and converges to a constant rate of acquisition thereafter. The number of epochs used in this simulation is greater due to the large size of the corpus. We conclude that word

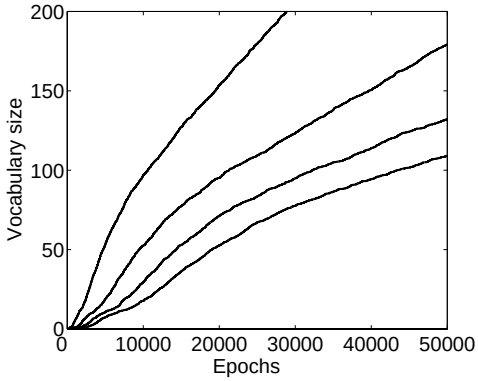


Figure 2: Vocabulary acquisition in a model with constant learning capacity, when presented with parental input. No vocabulary spurt is observed. Higher curves correspond to a lower threshold of learning.

statistics alone cannot be responsible for the vocabulary spurt observed at the end of the second year of life. Rather, as many researchers have suggested, the vocabulary spurt is driven by underlying changes in learning capacity arising from changes in mental representations and/or brain organisation. We next offer a conceptual implementation of this alternative view; that a change in learning capacity *is required* in order to display an accelerated increase in word learning.

Relaxation of the assumption of 'online' learning

We have demonstrated that an acceleration in vocabulary growth cannot be expected when presented with word distributions following Zipf's law, unless a change in learning capacity is implemented in the model or further variations in word difficulties are present. We have also shown that when the occurrence statistics deviates moderately from Zipf's law, as exemplified through simulations using the Parental Corpus, a vocabulary spurt is still absent in the model. We now show that the assumption of online learning can also be relaxed.

Let assume that upon presentation of a word, a 'memory trace' is initiated. This memory trace would modulate over time the value of the knowledge variable associated with that given word. Let us discuss the potential behaviour of this memory trace. We have already discussed the case for which the memory trace remains constant: It corresponds to the case in which each presentation of a word leads to an increment in the knowledge variable associated with that word, until it crosses a threshold. We have demonstrated earlier that no acceleration in vocabulary growth is observed unless an improvement of learning capacity is implemented in the model. Moreover, frequent words are learnt very early on, thereby failing to reproduce the long latency period observed in early childhood. Alternatively, the memory trace could increase over time (Vlach, Sandhofer, & Kornell, 2008), mimicking consolidation of the word form or meaning during sleep (Dumay & Gaskell, 2007) or through rehearsal of that word. However, high frequency words would be learnt even more rapidly under these conditions than with the constant memory trace, resulting, again, in the absence of the long latency period observed early in life. This account would fail to exhibit the characteristic contrast observed in infancy between a slow initial learning followed by an acceleration in lexical growth. Finally, the memory trace could decay over time, reflecting the degradation in the representations of words in absence of a new utterance, as described by Horst and Samuelson (2008). In this case, low frequency words whose memory trace decays faster than the typical interval between successive word presentations would never be learnt. Although we do not suggest that no learning take place beyond the actual presentation of a word, dynamic memory traces associated with individual presentations of the word are not the ingredient needed to explain the supra-linear vocabulary growth. Decaying memory traces as in Horst and Samuelson (2008) or reinforcement (Vlach et al., 2008; Dumay & Gaskell, 2007) would merely modulate the vocabulary spurt, not create this acceleration.

Relationship to McMurray's account

Our approach shares a similar goal to that of McMurray (2007): understanding the cause of the sudden increase in the speed of word learning observed during the second half of the second year of life. However, our approach differs in some important respects to both the original paper (Mc-

Murray, 2007) and subsequent implementations (Mitchell & McMurray, 2008, 2009). First, if infants only engage in online word learning, the raw statistics of word usage should be exploited and not, as in McMurray (2007), a logarithmic transformation of the word frequencies. In addition to a lack of psychological validity, such a transformation suffers from mathematical instability: depending of the lexicon size, the sum of log-frequencies may become negative, and/or words with a very low usage (frequency smaller than 1 in the timescale used) would have a negative log-frequency, resulting in negative probability of occurrence. Thus, the vocabulary spurt described in McMurray (2007) is driven by a distribution of word frequencies that, due to its log-sampling, do not reflect the true nature of the statistics of word occurrences. Second, Mitchell and McMurray (2008) introduce a stochastic adaptation of the original model and show that a wide range of distributions can lead to a spurt-like behaviour. Crucially, Zipf's law belongs to the class of distributions that do not lead to a vocabulary spurt.

Finally, Mitchell and McMurray (2009) study leveraged learning in word learning. They explore different metrics for relating word difficulty to word frequency. In a first case, they scale difficulty as an additive function of frequency. In order to avoid the problem of very high frequency words having negative difficulty values, they add a constant value to the difficulty score. The second case, in which word difficulty is scaled to the inverse of frequency is the approach we have chosen: For example, a word that is heard twice as often is deemed to be exactly twice as easy to learn. However, words follow Zipf's law only at a stochastic level. Our analysis, beyond initial analytical considerations, provides a stochastic account of word learning, when infants hear words drawn either from Zipf's distribution or from a corpus consisting of speech to which infants are typically exposed. Mitchell and McMurray (2009) provide a non-stochastic implementation of Zipf's distribution and Mitchell and McMurray (2008) provide a stochastic implemetation of non-Zipfian distributions. The critical combination of a Zipfian distribution with a stochastic implementation is absent from their account.[4]

An alternative account

Since a Zipf distribution of word usage is insufficient to capture the vocabulary spurt, we simulate an alternative account where the capacity of learning a word is not kept constant during early life. As infants only learn words on the basis of raw exposure, the model is presented with words drawn from a Zipf distribution and, for each presentation, the model has an increasing probability of learning that word. We presented 10,000 words per "day" in the simulation[5], out of a 40,000 word lexicon distributed with Zipf's law. Words that were presented on average less than once per day, were sampled according to their probability of occurrence within a day. The developmental time course of this probability is implemented as a non-linear function of time, in order to mimic the emergence of fast mapping and increased learning capacity, observed during the second year of life. In the model, the probability of learning a word increases with time; $p(t) = (t/20000)^3$. Any non-linear increase in the probability of learning a word would result in a non-linear developmental trajectory of word learning. Such a change in the parameters would only result in a quantitatively different path to word learning, not a qualitative change[6]. Note that this model is equivalent to a modified version of McMurray's model, in which increment size increases with time. From this perspective, many presentations of a word are needed for successful learning early in development whereas later in the second year, just a single presentation may be sufficient for learning that word, due to the emergence of fast mapping. Fig. 3 depicts the developmental trajectory simulated with the model. The curve of vocabulary acquisition possesses a clear non-linearity separating the early slow learning and the late fast learning regimes, similar to the naming explosion.

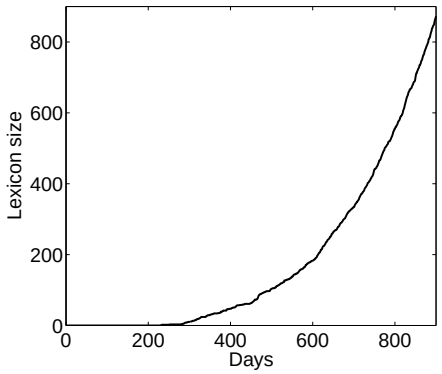


Figure 3: Acquisition in the present model, where learning capacity changes over time.

Discussion

Two contrasting hypotheses have been proposed in order to explain the rapid increase in the speed of word learning occurring in the second year of life. On the one hand, researchers have argued that the vocabulary spurt is driven by changes in the infant's learning capacities, such as the emergence of a naming insight or via maturational changes in the brain. We refer to this view as the *endogenous* hypothesis. In con-

[4] "[...] it is important to remember that frequency is not a property of the word [...], it is an estimate of how often it occurs (stochastically) in the child's environment. Thus, our model may be limited in its ability to handle frequency, and a stochastic model may be a better approach for dealing with it (e.g., Mitchell & McMurray, 2008)" (Mitchell & McMurray, 2009, p.1519)

[5] Hart and Risley (1992) reported that, on average, 10- to 18-month-old infants hear 1275 words per hour. Assuming that this level of exposure is maintained for 8 hours per day, then infants hear about 10,000 words a day.

[6] Since in the present simulations we did not simulate the system for more than 20000 epochs, the probability is always smaller than 1. One could alternatively choose a non-linear function of time that saturates at 1 (or close to 1) for increasing time, so as to mimic a smooth and continuous improvement in learning capacities.

trast, a second hypothesis highlights environmentally-based factors that contribute to the difficulty in learning words, such as frequency, phonological complexity, etc. On this view, the vocabulary explosion is a by-product of variability in word difficulty. We refer to this hypothesis as the *exogenous* hypothesis.

We have argued that simple analytical considerations demonstrate that a linear increase in the size of the lexicon is expected when presented with word frequencies distributed with Zipf's law. Moreover, simulations with a stochastic sampling of words following Zipf's law, as well as with samples of speech to which infants are exposed, confirmed that the type of distribution of word frequencies found in natural language would fail to induce a naming explosion. Mitchell and McMurray (2008) have shown that a wide range of mathematical distributions of word difficulties predict a non-linear growth of the infant lexicon. We have demonstrated that word occurrences following Zipf's law and speech typically heard by infants does not belong to this family of mathematical distributions.[7]

Since we have demonstrated that word frequency cannot account for the vocabulary spurt, it is reasonable to ask whether other exogenous factors that influence word difficulty could be the source of the non-linear vocabulary growth. For example, McMurray (2007) points out that phonological complexity contributes to word difficulty. It is not straightforward to measure the impact of phonological complexity during early word learning since the basis of infant's lexico-phonological representations is not yet well understood. However, as a first approximation, we might consider word length as a proxy for phonological complexity and hence word difficulty. In a recent review, Juhasz (2005) identified contributing factors in picture naming tasks. All reviewed studies (13) showed a correlation between age of acquisition and latency measures, suggesting that latency in picture naming tasks is a reliable way of determining when the word was acquired. In contrast, word length was found to be a significant variable in only 3 studies, whereas 9 studies found it to be non-significant. Phonological complexity, therefore, like frequency may not be a suitable candidate for predicting vocabulary acceleration as "an unavoidable by-product of variation in difficulty". Whereas many factors can impact the distribution of difficulty in learning a word, such as word length or word frequency, it remains to be proven that they play a primary role in determining the shape of the vocabulary spurt. Nevertheless, it is important to highlight that other exogenous factors are likely to contribute to differences in word difficulties. Many researchers would argue that words are not learnt in isolation, and the context in which they appear may affect directly the set of potential interpretations of the words, through referential uncertainty. Computational models have shown that word learning in a sentential context can display a spurt-like pattern in the learning curve (Siskind, 1996; Fazly, Alishahi, & Stevenson, 2008) and experimental studies have shown that context diversity and within-context ambiguity can override the role of word frequency (Kachergis, Yu, & Shiffrin, 2009). Nevertheless, Hayes and Ahrens (1988) have shown that there is a positive correlation between a caregiver's mean length of utterance and the age of the infant. As a consequence, young infant are exposed frequently to words in isolation or in short motherese.

We propose, instead, that endogenous factors are primarily responsible for the vocabulary spurt. Among them, the emergence of fast mapping can explain the increase in the ease of acquisition late in the second year of life (Carey & Bartlett, 1978). Further evidence for a change in learning capacity is that word familiarity impacts the distribution of brain regions involved in word learning, reflecting an increased efficiency in the manner in which infants process familiar and novel words across the vocabulary spurt (Mills, Plunkett, Pratt, & Schafer, 2005). It is, however, important to note that neither maturational changes in the brain, nor the application of innate or domain-specific constraints are required to explain a change in learning capacity. For example, Mayor and Plunkett (2008) showed that no specialised mechanisms are needed to explain the vocabulary spurt, as a simple general learning mechanism can lead to the spontaneous emergence of fast mapping. A change in learning capacities, not mechanisms, drives the rapid onset of vocabulary acquisition observed late in the second year of life. Hence, a word that seems difficult for a 15-month-old may be acquired almost instantaneously by a 21-month-old. Is the vocabulary spurt compatible with Zipf's law? The answer is clearly "yes" provided we allow the listener to develop her learning capacities.

[7] An anonymous reviewer pointed out that a caregiver's word usage may vary over time, despite following Zipf's law at a global scale. As a result, fragments of a caregiver's speech may deviate from Zipf's law, resulting in a vocabulary spurt. The analysis of a biased stochastic sampling of words from a Zipf distribution would be an interesting avenue for further research. However, a random sampling from a Zipf distribution failed to display a spurt-like pattern of word learning.

References

Altmann, G. (2002). Zipfian linguistics. *Glottometrics*, *3*, 19–26.

Bates, E., Benigni, L., Bretherton, I., Camaioni, L., & Volterra, V. (1979). *The emergence of symbols: Cognition and communication in infancy*. New York: Academic Press.

Beier, E. (1965). *Analysis of word frequencies in spoken language of children.*

Bloom, L. (1973). *One word at a time: The use of single word utterances*. The Hague: Mouton.

Carey, S., & Bartlett, E. (1978). Acquiring a new word. *Papers and Reports on Child Language Development*, *15*, 17–29.

Dahl, H. (1979). *Word frequencies of spoken American English*. distributed by Gale Research Co Detroit, Mich.

Dapretto, M., & Bjork, E. (2000). The development of word retrieval abilities in the second year and its relation to early vocabulary growth. *Child Development*, 635–648.

Dore, J., Franklin, M. B., Miller, R. T., & Ramer, A. L. H. (1976). Transitional phenomena in early language acquisition. *Journal of Child Language*, *3*, 13-28.

Dumay, N., & Gaskell, M. (2007). Sleep-associated changes in the mental representation of spoken words. *Psychological Science*, *18*(1), 35.

Fazly, A., Alishahi, A., & Stevenson, S. (2008). A probabilistic incremental model of word learning in the presence of referential uncertainty. In B. Love, K. McRae, & V. Sloutsky (Eds.), *Proceedings of the 30th annual conference of the cognitive science society.* Cognitive Science Society.

Ganger, J., & Brent, M. (2004). Reexamining the Vocabulary Spurt. *Developmental Psychology*, *40*, 621–632.

Goodman, J., Dale, P., & Li, P. (2008). Does frequency count? Parental input and the acquisition of vocabulary. *Journal of Child Language*, *35*(03), 515–531.

Gopnik, A., & Meltzoff, A. (1987). The development of categorization in the second year and its relation to the other cognitive and linguistic developments. *Child Development*, *58*, 1523-1531.

Hart, B., & Risley, T. (1992). American parenting of language-learning children: Persisting differences in family-child interactions observed in natural home environments. *Developmental Psychology*, *28*(6), 1096–1105.

Hayes, D., & Ahrens, M. (1988). Vocabulary simplification for children: A special case of "motherese". *Journal of Child Language*, *15*(2), 395–410.

Horst, J., & Samuelson, L. (2008). Fast mapping but poor retention by 24-month-old infants. *Infancy*, *13*(2), 128–157.

Huttenlocher, J. (1991). Early Vocabulary Growth: Relation to Language Input and Gender. *Developmental Psychology*, *27*(2), 236–48.

Juhasz, B. (2005). Age-of-acquisition effects in word and picture identification. *Psychological Bulletin*, *131*(5), 684.

Kachergis, G., Yu, C., & Shiffrin, R. (2009). Frequency and Contextual Diversity Effects in Cross-Situational Word Learning. In N. e. a. Taatgen (Ed.), *Proceedings of the 31st annual conference of the cognitive science society.* Cognitive Science Society.

Kamhi, A. (1986). The elusive first word: the importance of the naming insight for the development of referential speech. *Journal of child language*, *13*(1), 155.

Li, P., & Shirai, Y. (2000). *The acquisition of lexical and grammatical aspect*. Walter de Gruyter.

MacWhinney, B. (1991). *The CHILDES project : Tools for analyzing talk*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Mayor, J., & Plunkett, K. (2008). Learning to associate object categories and label categories: A self-organising model. In B. Love, K. McRae, & V. Sloutsky (Eds.), *Proceedings of the 30th annual conference of the cognitive science society.* Austin, TX: Cognitive Science Society.

McMurray, B. (2007). Defusing the childhood vocabulary explosion. *Science*, *317*(5838), 631.

McShane, J. (1979). The development of naming. *Linguistics*, *17*, 879–905.

Miller, G., & Chomsky, N. (1963). Finitary models of language users. In R. Luce, R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (Vol. 2, pp. 419–491). Wiley, New York.

Mills, D. L., Coffey-Corina, S. A., & Neville, H. J. (1993). Language acquisition and cerebral specialization in 20-month-old infants. *Journal of Cognitive Neuroscience*, *5*, 317–334.

Mills, D. L., Plunkett, K., Pratt, C., & Schafer, G. (2005). Watching the infant brain learn words: Effects of language and experience. *Cognitive Development*, *20*, 19–31.

Mitchell, C., & McMurray, B. (2008). A stochastic model for the vocabulary explosion. In B. Love, K. McRae, & V. Sloutsky (Eds.), *Proceedings of the 30th annual conference of the cognitive science society.* Austin, TX: Cognitive Science Society.

Mitchell, C., & McMurray, B. (2009). On Leveraged Learning in Lexical Acquisition and Its Relationship to Acceleration. *Cognitive Science*, *33*(8), 1503–1523.

Nazzi, T., & Bertoncini, J. (2003). Before and after the vocabulary spurt: two modes of word acquisition? *Developmental Science*, *6*(2), 136-142.

Ninio, A. (1995). Expression of communicative intents in the single-word period and the vocabulary spurt. *8*, 103–124.

Plunkett, K. (1993). Lexical segmentation and vocabulary growth in early language acquisition. *Journal of Child Language*, *20*, 1–19.

Reznick, J. S., & Goldfield, B. (1992). Rapid change in lexical development in comprehension and production. *Developmental Psychology*, *28*, 406–413.

Siskind, J. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, *61*(1-2), 39–91.

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*(3), 1558–1568.

Vlach, H., Sandhofer, C., & Kornell, N. (2008). The spacing effect in children's memory and category induction. *Cognition*, *109*(1), 163–167.

Woodward, A. L., Markman, E. M., & Fitzsimmons, C. M. (1994). Rapid word learning in 13- and 18-month-olds. *Developmental Psychology*, *30*, 553–566.

Yu, C., Smith, L., & Pereira, A. (2008). Grounding Word Learning in Multimodal Sensorimotor Interaction. In B. Love, K. McRae, & V. Sloutsky (Eds.), *Proceedings of the 30th annual conference of the cognitive science society.* Austin, TX: Cognitive Science Society.

Zipf, G. (1935). *The Psychobiology of Language: An Introduction to Dynamic Biology*. MIT Press, Cambridge, Massachusetts.

Zipf, G. (1949). *Human behavior and the principle of least effort*. Cambridge, Mass.: Addison-Wesley Press.

Context and Category Information in Children and Adults

Adam F. Osth (adamosth@gmail.com)
Center for Cognitive Science
The Ohio State University
208F Ohio Stadium East, 1961 Tuttle Park Place
Columbus, OH 43210, USA

Simon Dennis (simon.dennis@gmail.com)
Memory and Language Lab
The Ohio State University
225 Psychology Building, 1835 Neil Avenue
Columbus, OH 43210, USA

Vladimir M. Sloutsky (sloutsky.1@osu.edu)
Center for Cognitive Science
The Ohio State University
208C Ohio Stadium East, 1961 Tuttle Park Place
Columbus, OH 43210, USA

Abstract

An experiment by Dennis and Chapman (in press) found that as the length of a categorized list of materials increased, the false alarms to unrelated distracters decreased, a finding suggesting that adults are best described by context-noise models of recognition memory. Developmental evidence demonstrating that children the age of five are more sensitive to item information suggests that children might be described by item-noise models. We tested children and adults' performance and eye movements during recognition and found that adults' usage of category context was evident in both their performance and in their eye movements. Children, however, did not give conclusive evidence in their memory performance but their eye movements did not reflect usage of category context.

Keywords: Recognition memory; Inverse list length effect; Categorization and memory; Development of memory; REM; BCDMEM; context-noise; item-noise

Introduction

Current models of recognition memory, such as the REM model (Shiffrin & Steyvers, 1997) and the BCDMEM model (Dennis & Humphreys, 2001) are capable of making accurate predictions about a number of previously problematic effects in the literature, such as the list-strength effect (Ratcliff, Clark, and Shiffrin, 1990) and the mirror effect (Glanzer and Adams, 1985). However, these models not only possess different architectures but also capture the same trends in the data using different sources of information and interference.

The REM model assumes that during the study phase, each item is stored as a separate, noisy representation. During the test phase, a probe item is compared against every item in memory and an activation value is calculated based on the degree of match between each studied item and the presented probe item. These activation values are then averaged and a mean activation value that is sufficient (exceeding a fixed criterion value) produces a yes response. If distracters happen to have sufficient match with some of the studied items in memory, this produces a false alarm to the distracter. The REM model, as well as other global-matching type models (see Clark & Gronlund (1996) for a review), are classified as item-noise models due to the fact that interference is produced by the content and number of studied items.

BCDMEM, in contrast, assumes that during the study phase, each item isn't stored but is instead bound to the study context. During the test phase, the probe item cues all previous contexts in which the item was studied in, including learned elements of the study context. Additionally, the context of the study episode is reinstated. This reinstated context is then compared against the retrieved contexts of the item to evaluate whether or not the item was presented during the experiment, and a sufficient degree of match between the elements of the study context and matching elements in the retrieved contexts produces a yes response. If a distracter item happens to have been experienced in a large number of contexts, such as a high frequency word in the English language, then it is more likely for the retrieved context layer to spuriously contain elements of the study context and a false alarm can be produced. Consequently, this model is classified as a context-noise model because previous contexts are the principal source of interference.

Because these models account for the same effects using different sources of information and interference, determining which model is correctly representing the memory system requires looking at more current evidence in the literature. Dennis and Chapman (in press) recently found that when list length was varied between 10 and 80 items but the number of categories was kept constant (essentially varying the number of exemplars per category),

false alarms to unrelated distracter items were lower in the long list relative to the short list. This qualitative trend was dubbed the 'inverse list length effect,' which is the opposite of the predictions of the REM model. REM predicts a slight increase to unrelated distracters with increased list length, as extra study items produce additional noise during recognition. However, this effect can be handled by the BCDMEM model with the assumption that when participants are reinstating the context of the study episode, the categories learned while studying the long list become a part of the reinstated study context. These added elements of category context are then matched to the category contexts cued by the items, producing higher likelihoods of a "yes" response for matching category items and lower likelihoods of "yes" responses for mismatching category items. The authors further argue that it would be impossible for the REM model or virtually any other model that solely relies on information from individual exemplars to capture this effect.

There exist developmental evidence, however, that suggest that the source of interference may change through development. Sloutsky and Fisher (2004) conducted an experiment in which participants, which consisted of adults and children the age of five, either participated in a categorization task in which they had to induce a novel category property to animal photos, or they participated in a baseline condition in which they merely studied the photos. A surprise recognition task followed either of the two conditions, and it was found that adults experienced a sharp decrement in their ability to discriminate between studied and non-studied category items due to an increase in false alarms to related distracters. Children, in contrast, experienced no such decrement in their memory performance between the two conditions. The authors attribute this dissociation in performance between the two age groups to the fact that adults are much more sensitive to category-based information, while children are much more sensitive to item-based information.

These findings are consistent with a large number of findings from the research review performed by Brainerd, Reyna, and Ceci (2008). In this review, a number of experiments were discussed in which it was demonstrated that children are much less susceptible to false memory errors, a pattern which increases with age up until adulthood. These effects and age trends were evident in paradigms ranging from DRM type tasks to suggestibility experiments. The authors attribute these errors, particularly the decreased likelihood of recalling or falsely recognizing items highly similar to items presented at test, to children's weaker ability to spontaneously extract gist from the test materials in the same manner as adults.

We argue that if children are indeed more focused on item information and are weaker in their ability to extract gist during the study phase, then they should be deficient in their ability to extract and reinstate category context in the same manner as adults. We tested this by running both adults and children in a paradigm very similar to that used by Dennis and Chapman (in press). If our hypothesis about children's inability to use context is correct, then they should be unable to exhibit an inverse list length effect and their performance may follow the predictions of item-noise models of recognition. Adults, in contrast, should behave consistently with previous findings and experience facilitation in their ability to reject unrelated distracters in the long list condition.

Experiment 1

Method

Participants Participants were 65 children (33 female and 32 male, *M* = 4.87 years, SD = 0.61 years) and 83 adults (36 women and 47 men, *M* = 19.7 years, SD = 2.93 years). Child participants were recruited from suburbs in Columbus, OH. Adult participants consisted of undergraduate students from The Ohio State University participating for course credit.

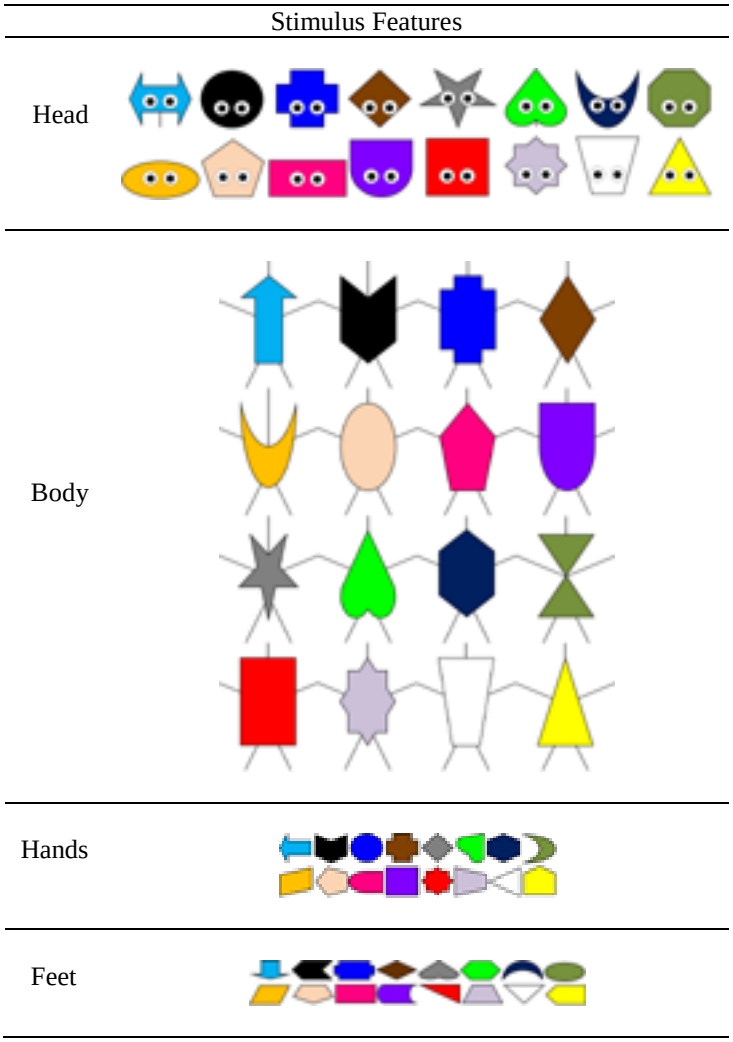


Figure 1: All stimulus feature possibilities in exemplar construction. For the sake of space efficiency, only left hands and left feet are presented.

Stimuli Visual stimuli consisted of artificial creatures that were composed of four different body parts: a head, a body, a pair of hands, and a pair of feet. Each body part was composed of a unique color and shape pairing. The creatures were assembled by randomly selecting a shape and a color for each component from a selection of over 16 different colors (common to all body parts) and 16 different shapes (unique to each body part). All shape and color features are detailed in Figure 1.

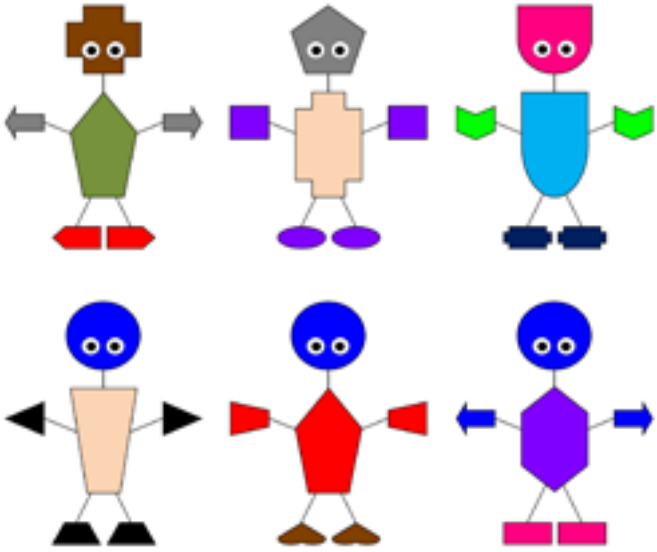


Figure 2: Examples of composite figures created from the sample features. *Top row:* Examples of three non-categorized exemplars. *Bottom row:* Examples of three categorized exemplars potentially seen in the long list.

For the study phase of the short list, a total of 8 exemplars were presented to the participant with unique shapes and colors for the different components (head, body, hands, and feet). For the study phase of the long list, 64 exemplars were presented and blocked into 8 categories of 8 exemplars each. Categories were defined as exemplars which shared heads of the same shape and color while the hands, feet, and body of the shared category exemplars consisted of different shape and color combinations. All categories exhibited unique, non-overlapping shape and color combinations for the heads. For the body, hands, and feet, all categories used the same colors and shapes (all randomly selected from the 16 available), no shape/color combinations were reused between categories and no shapes or colors were used for more than one exemplar within a given category. It should also be mentioned that the same number of features were used in both the short and long list conditions. Examples of possible exemplars created for short lists and long lists can be seen in Figure 2.

Distracter items in the test list were divided into two types: related distracters and unrelated distracters. Related distracters were constructed by selecting one head from each of the 8 presented categories and using reshuffled combinations of previously presented shapes and colors for the body, hands, and feet. Unrelated distracters were constructed in the same manner as related distracters, except that the heads were not categorically related to the presented stimuli and instead were composed of new combinations of shapes and colors that were not presented at test (all remaining colors and shapes not sampled for the study items). It should be noted that all shapes and colors have an equal probability of being used in target, related distracter, and unrelated distracter items. Examples of possible distracter item composites can be seen in Figure 3.

The stimuli were presented on the center of the computer screen. All stimuli and tasks in the experiment were controlled by E-Prime 2.0 Professional software.

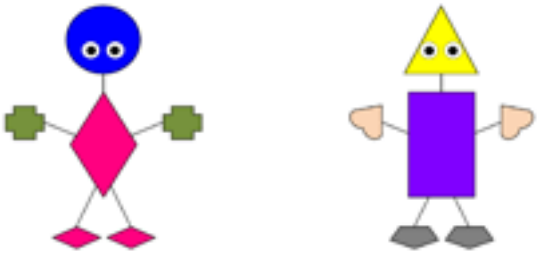


Figure 3: Examples of possible distracter items. *Left:* a related distracter item, featuring an identical head as the above category exemplars. *Right:* an unrelated distracter, featuring a head composed of non-presented features.

Procedure Upon arrival, participants were randomly selected into either the short list condition or the long list condition. Participants were then briefed about the stimuli they would be viewing. They were given incidental learning conditions, in that they were told to look at the stimuli and make a decision as to whether each stimulus was scary or funny, but were not told that they would be tested on their memory later in the experiment. During the study phase of the experiment, each exemplar was presented on the screen for 4500 ms while evaluating the exemplar to make the decision described above. Adult participants gave their responses by pressing keys on the keyboard while child participants gave their responses verbally to the experimenter, who then recorded the response on the keyboard. Each exemplar was preceded by a fixation cross which appeared for 500 ms.

To control for the different retention intervals between the short and long list, after completing the study phase of the experiment, both the short and long lists were followed by a distracter task that took place for 340 seconds for the short list and 60 seconds for the long list. The distracter task consisted of a rhythm game where participants listened to a sequence of 4 drumbeats and were then asked to tap out the sequence on the spacebar of the keyboard at the same tempo that the beats had played.

After completing the distracter task, participants were then instructed that they would be tested on their memory of the items from the study condition. Participants were instructed to respond "yes" if they recognized the item from the study phase, or to respond "no" if they did not recognize the item. The adult participants were instructed to give their responses on the keyboard while the children gave verbal responses to the experimenter who recorded the responses on the keyboard.

The test list consisted of 24 items: 8 of which were target items drawn from the study list, 8 of which were related distracters, and 8 of which were unrelated distracters. For the long list, each target item and related distracter was selected from each of the separate categories with no category being sampled more than once, such that all categories were represented on the test list.

Adult participants were tested in a laboratory at the university. Child participants were tested in local daycares or preschools by trained adult experimenters.

Results and Discussion

d' scores were calculated as a measure of memory sensitivity for all participants. Since we were most principally interested in the usage of the head information in making recognition judgments, only hits and false alarms to unrelated distracters were used in the calculations. Edge corrections were performed by adding 0.5 to the hit and unrelated false alarm counts and 1 to the target and unrelated distracter counts, as hit rates of 1.0 or false alarm rates of 0.0 produce infinite values for *d'* (Snodgrass & Corwin, 1988). Participants with *d'* less than or equal to 0 were excluded from the analysis (7 children and 4 adults). Hits, false alarms to related distracters, and false alarms to unrelated distracters are summarized in Table 1.

Table 1: Mean Proportions of Hits, Related False Alarms (FA), Unrelated False Alarms, and Mean *d'* Scores

	Adults		Children	
	Short List	Long List	Short List	Long List
Hits	0.82	0.86	0.66	0.64
Related FA	0.47	0.79	0.52	0.60
Unrelated FA	0.20	0.16	0.22	0.22

A repeated measures analysis of variance (ANOVA) was conducted on the responses for each subject, with list condition (short vs. long) and age group (adults vs. children) as between subjects factors and item type (target vs. related distracter vs. unrelated distracter) as a within subjects factor. Results indicated a significant main effect of item type, $F(2, 405) = 132.50$, $p < .001$, list type, $F(2, 405) = 4.28$, $p < .05$, as well as a significant item by age interaction, $F(2, 405) = 10.48$, $p < .001$, an age by list interaction, $F(1, 405) = 5.34$, $p < .05$, and an item by age by list interaction, $F(2, 405) = 3.09$, $p < .05$.

Planned post-hoc comparisons revealed a significant difference in adults' related false alarm rates across the two conditions ($t = 24.05$, $p < .001$), a difference which was insignificant for children ($t = 2.01$, $p > .05$). This replicates the findings of Sloutsky and Fisher (2004) but does not distinguish between the two models of recognition memory we're comparing against. Thus, planned post-hoc comparisons were also calculated on the differences in unrelated false alarms between the two list conditions for both adults and children, and revealed insignificant differences for both age groups (adults: $t = 1.16$, children: $t = .04$, $ps > .05$). Thus, neither age group revealed an inverse list length effect in their memory performance. However, there were differences in the groups' reaction times.

Table 2: Mean of Median RTs for Target Items, Related Distracters, and Unrelated Distracters

	Adults		Children	
	Short List	Long List	Short List	Long List
Target	1009	1011	3222	3126
Related	1344	1053	3130	3137
Unrelated	1198	944	2944	2859

To counteract positive skew in reaction time data, we took the median of each participant's reaction times for each of the 8 trials of every item type. The means of these median RTs can be seen in Table 2. We subjected the RTs to a repeated measures analysis of variance with list condition and age group as between subjects factors and item type as a within subjects factor. Results indicated a significant main effect of age, $F(2, 405) = 235.4$, $p < .001$. Planned post-hoc comparisons were also calculated on the RT differences between the two conditions for each item type and for each age. Significant differences between the two list conditions were found in adults for related distracters ($t = 2.95$) and unrelated distracters ($t = 3.75$), $ps < .01$. No significant differences in reaction times were found between the two list conditions in children, however it should be mentioned that children's responses were recorded by an experimenter and are thus difficult to interpret.

Because adults in the long list are quicker to react to unrelated distracters without receiving any decrement in accuracy, it is clear that they are in fact receiving a facilitation in rejecting unrelated distracters, which is not only contrary to the predictions of item-noise models but in accordance with the context-noise approach. However, this evidence does not give us any indication as to which model describes children's memory judgments. For this, we decided to run the same experiment in an eye tracker as a way of measuring the information that is being used in the test phase. Considering that the head is the most relevant feature of an unrelated distracter, the clearest prediction that can be made is that participants that have successfully abstracted the category context in the long list will look

significantly longer at the head relative to participants in the short list condition.

Experiment 2

Method

Participants Participants were 34 children (15 female and 19 male, *M* = 5.25 years, SD = 0.60 years) and 43 adults (18 female and 25 male, *M* = 19.5 years, SD = 1.29 years), participated in this experiment. Child participants were recruited from suburbs in Columbus, OH. Adult participants consisted of undergraduate students from The Ohio State University participating for course credit.

Apparatus Eye gazes were measured using a Tobii T60 eye tracker with a sampling rate of 60 Hz (60 data points collected per second). The device is integrated into a 17-inch monitor within a testing booth. A camera adjacent to the eye tracker provided a live feed to a trained experimenter at a nearby computer, who was able to monitor both the participant's eye movements as well as the stimuli they were viewing.

Procedure The procedure was nearly identical to that of Experiment 1 with the exception of a couple of minor adaptations to make this experiment compatible with the usage of an eye tracker. Each stimulus component (head, body, hands, and feet) was given a pre-determined area of interest (AOI) for recording eye gaze movements. To keep participants visual attention, we used a gaze-contingent fixation point between all trials in both the study phase and the test phase such that a stimulus would only be presented if participants maintained their gaze on the fixation point for a randomly calculated time interval between 300 and 700 ms. Additionally, since the stimulus appears in the center of the screen, the fixation point was randomly presented in the center of one of four quadrants on the screen to ensure that first looks were not biased by the fixation position.

All participants, including both adults and children, were run by trained adult experimenters at a nearby computer throughout the duration of the experiment. Because trials continue until a response is given, allowing adults to enter their own responses on a keyboard and having children's responses be entered by an experimenter will yield different patterns of data for the two groups. To make the data comparable between the two age groups, both adults and children gave all responses to the experimenter who entered them on a keyboard.

Results and Discussion

d' scores were calculated in the same manner as Experiment 1 and all participants with d' scores less than or equal to 0 were excluded from the analysis (2 adults and 4 children).

Because the head was the category relevant feature, all analyses were restricted to that area of interest. The dependent measure we selected was the proportion of looks at the head, which was calculated for each item type. These results can be seen in Table 3. Because trials continued until participants gave their responses, with some trials continuing for as long as 10 seconds, the calculation was restricted from the start of the trial up until the mean reaction time (2 seconds for adults and 3 seconds for children).

Table 3: Mean Proportion of Looks at the Head for Target Items, Related Distracters, and Unrelated Distracters

	Adults		Children	
	Short List	Long List	Short List	Long List
Target	.261	.282.	.272	.247
Related	.272	.256	.276	.259
Unrelated	.261	.345	.296	.369

Considering that there is no category information in the short list (there were no repetitions of category exemplars), we interpreted looks at the head in the short list as a baseline degree of looking when only item information is available. Since the test phase is identical in both list conditions, any increase in looking at the head in the long list above that of the short list has to be due to differences in the study phase, most notably the repetitions of category exemplars. Visualizations of the looks at the head over time can be found in Figures 4 and 5.

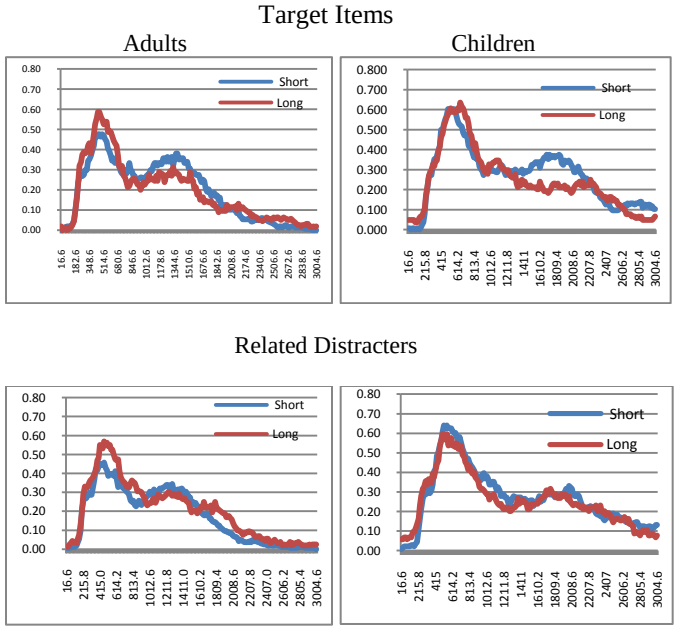


Figure 4: Differences in proportions of looking at the head target and related distracter presentations for both children and adults recorded at each refresh rate (every 16.6 ms).

T-tests were calculated on the differences between the two list conditions for each item type to determine if the

differences in looking were significant. For adults, differences between the two lists were insignificant for both target items ($t(39) = .90$, $p > .05$) and related distracters ($t(39) = -.53$, $p > .05$). This is not surprising, considering that for both target and related distracter items, the head is not diagnostic of whether or not the item was on the list. This is not the case for unrelated distracters, where the category information (i.e.: the head) is the most relevant feature for discrimination. A t-test between the two list conditions for unrelated distracters was significant ($t(39) = 2.39$, $p < .05$), in that adults looked significantly longer at the head in the long list condition relative to the short list condition.

For children, differences between the two list conditions for target items ($t(28) = -.76$, $p > .05$), related distracters ($t(28) = -.46$, $p > .05$), and most importantly, unrelated distracters ($t(28) = 1.97$, $p > .05$) were all insignificant. Because children in the long list condition were not using the category relevant information above and beyond that of the short list, we interpret this to mean that category context information was not being accessed above and beyond that of item information. However, considering that this p-value is close to the significance margin, it is possible that there are subsets of children showing the effect that are outnumbered by children not showing the effect.

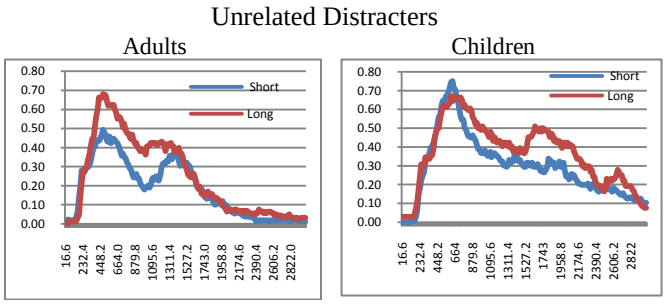


Figure 4: Differences in proportions of looking at the head unrelated distracter presentations for both children and adults recorded at each refresh rate (every 16.6 ms).

Conclusions

To summarize, for adults, increasing category length not only facilitated the rejection of non-category items but this facilitation manifested itself in a bias for category relevant features in their eye movements. For children, no such facilitation could be found in either their behavioral data or in their eye movements, implying that they may be meeting the predictions of the item-noise models of recognition memory.

We believe this is important research because despite there being a large volume of research on developmental differences in episodic memory, there has been little work connecting these differences to the components and processes of current memory models. We hope that this work as well as future work will make clear connections between the developmental literature and the modeling literature and use them to construct a detailed theory of how memory changes with development.

Acknowledgments

This research has been supported by grants from the grants from the NSF (BCS-0720135), from the Institute of Education Sciences, U.S. Department of Education (R305B070407), and from NIH (R01HD056105) to Vladimir M. Sloutsky.

References

Brainerd, C. J., Reyna, V. F., & Ceci, S. J. (2008). Developmental reversals in false memory: A review of data and theory. *Psychological Bulletin, 134*, 343-382.

Clark, S. E., Gronlund, S. D., 1996. Global matching models of recognition memory: How the models match the data. Psychonomic Bulletin and Review 3 (1), 37{60.

Dennis, S., & Chapman, A. (in press). The inverse list length effect: Implications for exemplar models of recognition memory.

Dennis, S., & Humphreys, M. S. (2001). A context noise model of episodic word recognition. *Psychological Review, 108*, 452–478.

Glanzer, M., & Adams, J.K. (1990). The mirror effect in recognition memory: Data and theory. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 16, 5-16.

Ratcliff, R., Clark, S. E., & Shiffrin, R. M. (1990). List strength effect: I. Data and discussion. *Journal of Experimental Psychology: Learning, Memory, & Cognition*,16, 163-178.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: Rem-retrieving effectively from memory. *Psychonomic Bulletin & Review, 4*, 145–166.

Sloutsky, V. M., & Fisher, A. V. (2004). When development and learning decrease memory: Evidence against category-based induction in children. *Psychological Science, 15*, 553 – 558.

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General, 117*, 34–50.

Visual Similarity is ObViS

Michel E. Brudzinski (brudzm@rpi.edu)
Chris R. Sims (simsc@rpi.edu)
Wayne D. Gray (grayw@rpi.edu)
Michael J. Schoelles (schoem@rpi.edu)
Cognitive Science Department
Rensselaer Polytechnic Institute
Troy, NY 12180

Abstract

Visual search for a target is affected by visual similarity. Research on visual similarity has primarily focused on the high-level features of objects. Real-world objects are composed of low-level features that can be harder to measure and categorize. We have developed ObViS, an algorithm that measures the visual similarity of objects, based on Rao & Ballard (1995). ObViS calculates a high-dimensional vector that represents the low-level features of a real-world object. The algorithm was applied to a library of real-world object images in order to calculate the similarity of each object to every other object in the library. Two experiments evaluated the ability of our algorithm to predict the effects of visual similarity on visual search behavior.

Keywords: similarity; visual similarity; visual search.

Introduction

The visual similarity of objects in a scene is an important exogenous factor driving visual search behavior. Theories of visual search acknowledge the importance of similarity but do not specify how the visual search system uses these exogenous factors to guide search (Wolfe, 2007).

Most models of visual search, such as Treisman's Feature Integration Theory (FIT) (Treisman & Gelade, 1980), propose separate serial and parallel search mechanisms. A limited set of features such as color and size can be processed in parallel. Parallel visual search for a target that can be distinguished solely on the basis of one of those features is fast and efficient. Serial visual search for a target that is defined by multiple features is slower and requires attention to bind together the features of objects.

Six of the eight phenomena that Wolfe lists as affecting visual search response time entail some form of visual similarity: (1) target-distractor similarity, (2) distractor heterogeneity, (3) flanking/linear separability, (4) search asymmetry, (5) categorical processing, and (6) guidance (Wolfe, 2007). Each of the six types of visual similarity is specific to a low-level feature such as size, color, or orientation.

In addition to low-level features, research on visual similarity has also focused on high-level features. Approaches to the study of similarity include: geometric, feature-based, alignment-based and transformational measures of similarity (Goldstone & Son, 2005). Each of these approaches requires a form of reductionist representation of low-level properties, features, or elements.

Research on visual search has focused on issues surrounding the deployment of the serial or parallel processes. Visual search in the laboratory has used simple stimuli, manipulated set size or a few high-level features, tested search for a single feature, or the conjunction of two features, and used response time as their primary measure. Visual search in the real-world involved conjunctions of many low-level features that can be difficult to measure or categorize.

Statistical models of visual search, such as Itti & Koch (2001), have demonstrated the potential use images of real-world scenes in visual search experiments. Their mathematical model simulates the role of bottom-up saliency in guiding visual attention. It extracts low-level visual features such as color, intensity, and orientation from real-world images. The model calculates the conspicuity at every point in a scene and thus is able to provide bottom-up guidance to direct visual attention. Such models predict where the eye is attracted to in a visual scene and, thereby, have an important but limited role to play in explanations of visual search.

Top-down guidance may dominate bottom-up guidance, such as saliency, when there is a search target (Chen & Zelinsky, 2006). The goal of a visual search for a particular target is to find a location in a scene that matches the visual features of the target. The visual search system must be able to maintain a representation of the features of the target and compare those features with features at locations in the scene.

Rao & Ballard's Active Vision Architecture describes two primary visual routines: one for object identification and one for object location (Rao & Ballard, 1995). Statistical models of visual similarity, such as Rao & Ballard's, share with saliency models their reliance on low-level visual features. They differ from saliency models in that they define vectors of low-level visual features for a known target and for the locations in a scene. They then compare the similarity of the target vector with the vectors computed for locations in the scene. Top-down guidance is based on the statistical similarity of scene locations to the search target.

Cognitive architectures, such as ACT-R (Anderson & Lebierre, 1998), are used to model visual search behavior. Visual attention guidance in ACT-R suffers from the same reliance on high-level visual features that limits most theories of visual search. Statistical models of both bottom-up and top-down guidance would greatly increase the ability

of cognitive models to model real-world visual search. We have implemented a variation of Rao & Ballard's model and applied it to the study of visual search with real-world objects.

Algorithm

Our algorithm, ObViS, is a variant of the one developed by Rao & Ballard (1995) and Rao et al (2002). The algorithm represents image patches using high dimensional feature vectors, where the computed features consist of the image response to oriented spatial frequency filters. Such filters approximate the receptive fields of simple cells in primary visual cortex, and are also similar to features obtained from the statistics of natural images (Hyvärinen, Huri, & Hoyer, 2009). In our implementation, we used 10 filters defined by the directional derivatives of a 2D Gaussian, using derivatives of up to 3rd order. The 10 combinations of Gaussian derivative order and orientation were as follows:

Table 1: Steerable basis set.

Order of derivatives	Filter orientations used (degrees)
0	0
1	0,90
2	0,60,120
3	0,45,90,135

This set of filters was chosen as it forms a steerable basis set—that is, the filter response at any orientation can be computed by a linear combination of the filter responses in the basis set (Freeman & Adelson, 1991). This property endows the feature representation with some rotation invariance, though we have not explored this property in our current work. In our implementation, the filter kernels were 9x9 pixel discretized versions of the Gaussian derivatives defined above. These 10 filters were applied to 3 color channels extracted from the original image: luminance, red–green, and blue–yellow color opponency channels. In addition, the filters were applied to these color channels at 5 spatial scales, by resampling the image to 25, 50, 100, 200, and 400% of its original size. Thus in total, each image was represented by the ObViS algorithm using a set of 150 measurements (10 filters x 3 color channels x 5 spatial scales). The use of color opponency channels was an extension to the algorithm presented by Rao et al (2002), as their implementation used only grayscale images whereas we are interested in capturing the visual similarity of color images. Finally, to determine the visual similarity between any two images, we determine the feature vector representation for two images, and then calculate the root mean square (RMS) of the difference between the images' respective feature vectors. Images with low RMS difference are highly similar according to the ObViS measure of visual similarity.

Experiments

We conducted two visual search experiments that examined ObViS' accuracy in predicting human visual search behavior. Subjects were asked to find target objects located in a circular array of object images from the Amsterdam Library of Object Images (ALOI) (Geusebroek, Burghouts, & Smeulders, 2005). We compared the timing, and accuracy of responses, and the number of fixations with predictions based on ObViS' calculations of visual similarity. The two experiments differed in how similar the distractors displayed on each trial were to the target object. In experiment 1, for each trial, the distractors in the search array were all either similar, or dissimilar to the target. In experiment 2, for each trial, the distractors in the search array were approximately half similar and half dissimilar to the target.

Methods

Subjects. Thirty-four RPI undergraduates participated in the experiment 1 and twenty-seven RPI undergraduates participated in experiment 2. All subjects received course credit for their participation and signed informed consent forms. Subjects were screened for color blindness using a 10-plated Ishihara test (Ishihara, 1987).

Materials. The experiment was run on a Mac OSX computer and displayed on a 17" flat-panel LCD monitor with a screen resolution set to 1280 x 960 pixels. The software used for the experiment was written in LispWorks 5.0. The object images displayed during the experiment were 192 x 192 pixel images of real-world object from the Amsterdam Library of Object Images. A Cedrus RB-834 response pad was used to collect responses. White noise was played over headphones, using the freeware program Noise, to reduce auditory distractions. All subjects in these experiments were eye-tracked using an LC Technologies eye-tracker that recorded at a rate of 120 Hz. Subjects were asked to rest their chin on a chinrest throughout the experiment.

Design. The same 100 target objects were used as the target objects in experiments 1 and 2. The target object was present in the search array in only half of the trials. Subjects had to respond whether the target objected was present, and if so, identify its location in the circular search array. Experiments 1 and 2 differed in how similar the distractors displayed on each trial were to the target object.

In experiment 1, subjects performed a visual search for a target in a search array that contained distractor objects that were either similar or dissimilar to the target. All subjects in experiment 1 saw the same 400 trials. In half of the trials, the distractor objects were all similar to the target; in the other half of the trials the distractor objects were all dissimilar to the target (Table 2).

In experiment 2, subjects performed a visual search for a target in a search array that contained approximately half similar and half dissimilar distractor objects (Table 3). The similarity of all distractor objects was based on calculations

from the ObViS algorithm. All subjects in experiment 2 saw the same 400 trials. The locations of all objects, in all trials, for all subjects, were randomized.

Table 2: Experiment 1 trials.

Trial count	Targets	Similar Distractors	Dissimilar Distractors
100	0	8	0
100	0	0	8
100	1	7	0
100	1	0	7

Table 3: Experiment 2 trials.

Trial count	Targets	Similar Distractors	Dissimilar Distractors
100	1	4	3
200	0	4	4
100	1	3	4

Procedure. Experiments 1 and 2 used the same procedures. All task instructions were presented using a Keynote presentation prior to the experiment. Subjects pressed a button on the response pad labeled "next" to begin each trial. A fixation cross, consisting of a white "+" was displayed in the center of the screen (Figure 3a). The trial did not begin until the participant had fixated on the fixation cross for 500 milliseconds. The target image was then displayed in the center of the screen for 300 milliseconds (Figure 3b). A random dot image was displayed for 300 milliseconds (Figure 3c). A circular search array was then displayed until the subject responded by pressing the "Present" or "Absent" button on the response pad (Figure 3d).

Following a response, the random dot image was displayed for another 300 milliseconds (Figure 3e). If the response indicated that the target was present, the subjects were asked to indicate the location of the target. Buttons were arranged on the screen in locations that matched the locations of objects in the search array (Figure 3f). Subjects responded by moving a mouse to and clicking on one of the buttons. Once the participant responded, a progress screen displayed the number of trials completed out of the total number of trials. Subjects pressed the response pad button labeled "next" to begin the next trial (Figure 3).

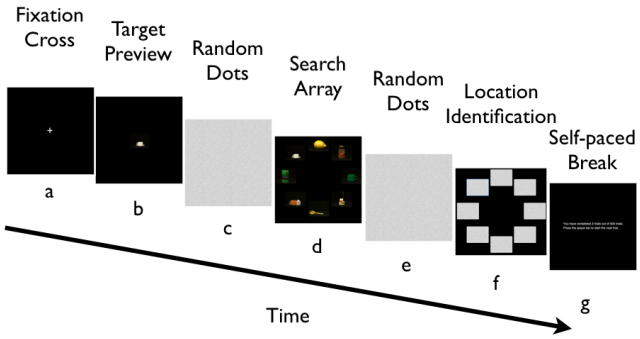


Figure 3: Experimental procedure for experiments 1 & 2.

Measures. The search array was displayed until the participant responded that the target object was either present or absent. The duration of time from the initial display of the search array until the participant responded was measured as the response time. The participant's response was recorded and measured as target presence accuracy.

Eye-tracking data was recorded throughout the experiment. The number of fixations on distractor objects was counted for each trial.

Results

Response time was compared for trials in which the target object was present in the search array and trials in which it was absent.

In experiment 1, subjects took longer to respond on trials when the target was absent (M = 1190.13, SE = 61.76), than on trials in which the target was present (M = 1066.22, SE = 34.05), $t(129)_{\text{two-tail}}$ = 1.716, p = 0.089, marginally significant.

Experiment 2 showed the same trend in response times: subjects took longer to respond on trials in which the target object was absent (M = 1234.83, SE = 54.82), as compared to trials when the target was present (M = 1064.59, SE = 31.76), $t(79)_{\text{two-tail}}$ = 2.872, p = 0.005 (Figure 4).

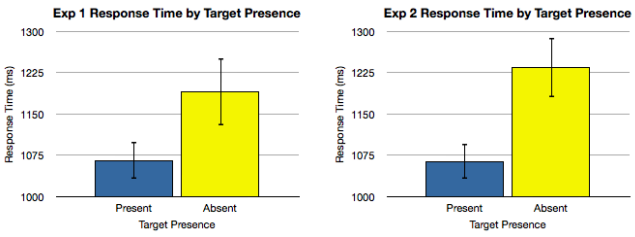


Figure 4. Response time (ms), by target presence, in experiments 1 & 2.

In experiment 1, each trial contained distractors that were all similar or all dissimilar to the target object. Response time was significantly greater for trials with distractors that were similar to the target (M = 1244.35; SE = 56.04) than for trials with dissimilar distractors (M = 1012.00; SE = 39.18), $t(130)_{\text{two-tail}}$ = 1.66, p = 0.0009 (Figure 5). In

experiment 2, each trials contained both similar and dissimilar distractors, so the analogous comparison was not possible.

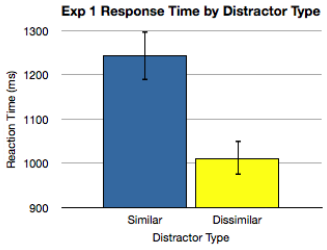


Figure 5. Response time (ms), by distractor type, in experiment 1.

The number of fixations on distractor objects was counted for each trial. In experiment 1, when the target was absent, subjects averaged more fixations on similar distractors (M = 2.99; SE = 0.22), than on dissimilar distractors (M = 1.76; SE =0.17). When the target was present, subjects averaged fewer fixations on distractors, but still had more fixations on similar distractors (M = 0.88; SE = 0.07), than on dissimilar distractors (M = 0.41; SE = 0.04) (figure 6). An ANOVA showed a significant main effect of target presence, (F(1, 124) = 23.15, p < 0.001). There was also a significant main effect of distractor type, (F(1, 124) = 6.46, p < 0.05). There was no significant interaction.

In experiment 2, when the target was absent, subjects averaged more fixations on similar distractors (M = 1.96; SE = 0.11), than on dissimilar distractors (M = 0.94; SE = 0.07). An ANOVA showed a significant main effect of target presence, (F(1, 100) = 70.20, p < 0.001). There was also a significant main effect of distractor type, (F(1, 100) = 25.73, p < 0.001). There was a significant interaction between the two factors (F(1, 100) = 25.73, p < 0.001); there was a larger difference in mean fixation count for trials with similar distractors when the target was absent. When the target was present, subjects averaged fewer fixations on distractors, but still had more fixations on similar distractors (M = 0.54; SE = 0.03), than on dissimilar distractors (M = 0.22; SE = 0.02) (Figure 7).

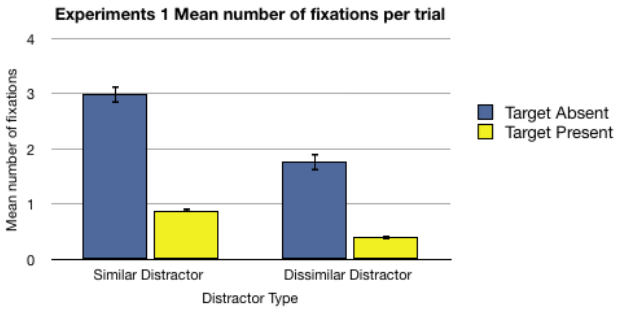


Figure 6. Mean number of fixations on distractor object per trial, by target present and distractor type, in experiment 1.

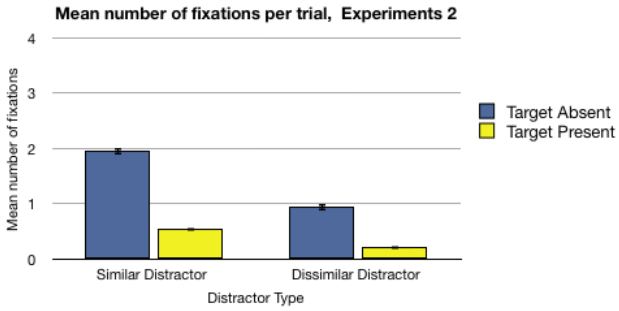


Figure 7. Mean number of fixations on distractor object per trial, by target present and distractor type, in experiment 2.

Target presence accuracy was a measure of the accuracy of subjects' response that the target was present or absent. Overall, subjects made few mistakes in both experiment 1 (M = 96.32%; SE = 0.36), and experiment 2 (M = 96.30%; SE = 0.002). In experiment 1, accuracy was significantly higher for trials in which the target was absent (M = 97.06%; SE = 0.36), than when the target was present (M = 95.58%; SE = 0.60), $t(130)_{\text{two-tail}}$ = 2.12, p = 0.018. In experiment 2 there was no main effect of target presence on accuracy; accuracy for trials with the target absent (M = 95.83%; SE = 0.002), was not significantly different than trials with the target present (M = 96.54%; SE = 0.003) , $t(79)_{\text{two-tail}}$ = 1.36, p = 0.17. Each trial in experiment 1 had distractors that were either all similar to the target or all dissimilar to the target. Target presence accuracy was higher for trials with dissimilar distractors (M = 97.76%; SE = 0.60), compared to trials with similar distractors (M = 94.88%; SE = 0.36), $t(130)_{\text{two-tail}}$ = 4.31, p = 0.0001.

General Discussion

We developed the ObViS algorithm in order to measure visual similarity for the top-down guidance of visual search. One of the goals of this work was to extend the study of visual search and visual similarity to real-world objects. We replicated the basic visual search phenomena. The algorithm's calculations were used to manipulate the similarity of visual search distractors. The response time data replicated phenomena typically found in laboratory search tasks using very simple stimuli; subjects took longer to find target when the distractors were similar to the target. They also made more mistakes in their responses when the distractors objects were similar to the target.

Fixation data added an additional source of information that has not typically been used in the study of similarity. Our results demonstrated that the longer response times for visual searches with similar distractors were the result of a greater number of fixations.

There are other visual search phenomena that we did not test. We did not manipulate set size, randomize locations, occlude objects, or place objects in natural scenes. All of

these phenomena could be studied in future work using our measures of visual similarity.

Conclusions

We developed a measure of the visual similarity of real-world objects based on the representation of their low-level visual features. We applied the algorithm to a library of object images. The resulting similarity calculations were used to manipulate the similarity of distractors in visual search tasks. We replicated basic findings on the effects of target presence and distractor similarity, using real-world objects. Further refinement of the ObViS algorithm could improve its ability to predict the effects of visual similarity on visual search. The algorithm could be used to create iconic representations to guide top-down visual search in computational models. ObViS could extend the study of visual search and visual similarity to real-world objects and even provide visual representations for cognitive architectures.

Acknowlegments

This work was supported, in part, by grant N000140710033 to Wayne Gray from the Office of Naval Research, Dr. Ray Perez, Project Officer.

References

Anderson, J.R., & Lebiere, C. (Eds.). (1998) *Atomic components of thought*. Hillsdale, NJ: Erlbaum.

Chen, X., & Zelinsky, G.J. (2006). Real-world visual search is dominated by top-down guidance. *Vision Research, 46*, 4118-4133.

Freeman, W.T., & Adelson, E.H. (1991). The design and use of steerable filters. IEEE Transactions on Pattern Analysis and Machine Intelligence, 13(9), 891-906.

Geusebroek, J.M., Burghouts, G.J., & Smeulders, A.W.M. (2005) The Amsterdam library of object images. *International Journal of Computer Vision, 61(1)*, 103-112.

Goldstone, R.L., & Son, J.Y. (2005). Similarity. In K.J. Holyoak & R.G. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning*. New York: Cambridge.

Hyvärinen, A., Hurri, J., & Hoyer, P.O. (2009). Natural Image Statistics: A Probabilistic Approach to Early Computational Vision: Springer.

Ishihara, S. (1987). Ishihara's tests for colour-blindness (concise edition). Tokyo: Kanchara.

Itti, L., & Koch, C. (2001) Computational modeling of visual attention. *Nature Reviews Neuroscience*, *2*, 194-203.

Treisman, A., & Gelade, C. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*, 97-136.

Rao, R.P.N., & Ballard, D.H. (1995). An active vision architecture based on iconic representations. *Artificial Intelligence*, 78, 461-505.

Rao, R.P.N., Zelinsky, G.J., Hayhoe, M.M., & Ballard, D.H. (2002). An active vision architecture based on iconic representations. *Artificial Intelligence*, 78, 461-505.

Wolfe, J.M. (2007). Guided Search 4.0: Current progress with a model of visual search. In W. Gray (ed.), Integrated Models of Cognitive Systems (pp. 99-119). New York: Oxford.

Learning Structured Preferences

Leon Bergen[1], **Owain R. Evans, Joshua B. Tenenbaum**
{bergen, owain, jbt}@mit.edu
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology
Cambridge, MA 02139

Abstract

Learning the preferences of other people is crucial for predicting future behavior. Both children and adults make inferences about others' preferences from sparse data and in situations where the preferences have complex internal structures. We present a computational model of learning structured preferences which integrates Bayesian inference and utility-based models of preference from economics. We experimentally test this model with adult participants, and compare the model to alternative heuristic models.

Keywords: Theory of Mind; Bayesian Modeling; Preference Learning

Introduction

Children and adults are highly adept at inferring hidden mental states such as preferences from observed behavior. Work in developmental psychology has shown that by 18 months, children have learned that others have preferences which are different from their own; moreover, they can infer these preferences on the basis of observing several choices (Repacholi & Gopnik, 1997). Further work has shown that children's inferences are sensitive to the statistical properties of observed choices, so that choices which are less likely a priori are more likely to reflect strong preferences (Kushnir, Xu, & Wellman, 2008).

Despite the ease with which people appear to reason about others' preferences, such reasoning is a highly challenging computational task. An individual's preferences are never fully determined by their past choices. Moreover, people have to infer preferences on the basis of sparse data, perhaps only a few observed choices.

Our goal in this paper is to develop a computational model of preference learning on the basis of sparse observations. In the spirit of Marr's levels of explanation, we must first spell out the computational problem which has to be solved. In this case, the problem has two parts. First, what exactly has to be learned? Are an agent's preferences simply a ranking of different choices, or do they have more internal structure? Second, how can these preferences be learned on the basis of observing an agent's choices?

Previous Work

Previous computational models of preference learning have mostly focused on the second question. Central to these models is the *principle of rationality*, according to which people are rational agents whose behavior is directed towards satisfying their preferences. Crucially, these models do not claim that the principle of rationality is descriptively true, but rather that it is assumed by people's intuitive theory of mind. These models treat the problem of preference learning as one of inverting the principle of rationality: in order to infer someone's preferences from their behavior, consider what preferences would have made this behavior rational.

In the inverse-planning model of Baker, Tenenbaum, and Saxe (2007), the task is to determine which of several candidate goals an agent desires on the basis of partial movements towards those goals. The agent is assumed to have a ranking of the goals, from best to worst, and which of the goals is desired is inferred by assuming that the agent is moving efficiently towards the desired goal. Using Bayesian inference, the model recovers a distribution over desired goals by supposing that a given goal makes efficient movement toward that goal highly likely. In this paper, there is a sophisticated link between preference and behavior, but preferences themselves have a simple structure.

Lucas et al. (2009) again use a Bayesian model for inference, but use a more sophisticated model for preferences. Instead of having preferences over outcomes, different possible object features have different utilities (are differentially preferred). Preferences over objects are given by a weighted sum over the objects' features. Agents are assumed to choose the object which gives them the maximal quantity of desirable features.

These models were not intended to capture how preferences might be learned when the goods being chosen between have complex functional relationships. Suppose, for example, that someone is choosing between a bicycle frame and bicycle wheels. It will not generally make sense to say that they prefer the frame to the wheels, or vice-versa. Rather, they most likely want an entire bicycle, meaning that whether they want the wheels will depend on whether they already have a suitable frame. Any computational model of this situation will have to account for this additional structure in the agent's preferences.

We combine the Bayesian modeling of Baker et al. and Lucas et al. with formal models of structured preferences developed by economists. We test our model experimentally by examining adult inferences in tasks that provide information about the interactions between goods. By systematically varying these interactions, we try to determine whether preference learning is sensitive to the structure of the preferences being learned.

[1]The first two authors contributed equally to the paper

Modeling Structured Preferences

Economists are interested in modeling agents who are choosing between goods that can have a range of functional relationships. They are also interested in modeling agents over long spans of time, meaning that they have to account for how novel goods may interact with goods already possessed by the agent. In order to model these interactions and effects of context, economists often will not assign utilities to individual goods, but rather to bundles of goods, where the bundle (A,B) consists of A units of one good and B units of another good. Different interactions between goods in a bundle can be captured by different utility functions on these bundles.

We will be considering three kinds of interactions between goods, each of which is standard in economics (see, e.g., Mas-Colell, 1995). The first class of goods are known as *substitutes*. Two goods will be substitutes if either can be used, independent of the other, to accomplish a single goal. For example, two brands of cherry soda are substitutes for each other: if someone wants to drink cherry soda, they can satisfy that goal by drinking either brand. The second class of goods are *complements*. These goods must both be used in a specific proportion to accomplish a particular goal. Bicycle wheels and a bicycle frame must be put together in a ratio of two to one in order to accomplish the goal of building a functional bicycle. The last class of goods will be described by a *Cobb-Douglas* utility function, which is presented below. These goods perform some goal most effectively in a particular ratio, but can still perform the goal to some extent in other ratios as well. If someone wants to have a fun vacation, it may be best to spend some number of days on the beach and some number of days on a cruise, as an entire vacation on the beach might get boring. However, additional days on either the beach or the cruise will still improve the trip.

These interactions can be captured by three different utility functions on bundles of goods. If two goods are substitutes, then we can set:

$$u(A,B) = \alpha \cdot A + (1-\alpha) \cdot B \tag{1}$$

where A is the quantity of the first good and B is the quantity of the second good. The weight encodes how well each of the goods serves its function. This equation says that we always get the same amount of utility out of an extra unit of the first good, regardless of how much of the second good we have, and vice-versa. This is essentially the utility function used in Lucas et al., with A and B representing the quantities of two different features.

For complements, we set the utilities of the two goods to be:

$$u(A,B) = min(\alpha \cdot A, (1-\alpha) \cdot B) \tag{2}$$

Parameter α in this case determines the required proportions between the two goods. Once the goods are in their correct proportions, an additional unit of either good alone will not increase utility. In order to increase utility, extra units of both goods have to be added.

The Cobb-Douglas utility function is defined by:

$$u(A,B) = A^{\alpha} \cdot B^{1-\alpha} \tag{3}$$

As in the case of complements, the parameter α determines the optimal ratio of the two goods. Unlike in the case of complements, additional units of a single good can still increase utility without additional units of the other good. Crucially, there is diminishing marginal utility in each good, holding the other good fixed. In other words, each additional unit of a single good increases utility less and less, unless more of the other good is added.

These utility functions are often the first introduced in microeconomics textbooks, both because of their simplicity and because the structures which they encode are so common. However, the space of possible utility functions and possible interactions between goods is enormous (Mas-Colell, 1995). Though we will only be considering these three cases, our model of preference learning, which we describe next, is compatible with any parameterized utility function.

Computational Modeling

Economists have supposed that individual preferences can be represented by the utility functions above and that agents choose bundles with highest utility. By assuming that preferences have this form, predictions can be made about choice in unobserved situations, by computing the relevant utilities.

On our model of human intuitive preference inference, people's intuitive inferences use the same representation of preferences as economists—people are intuitive economists. They can represent different utility function types, corresponding to the way goods interact in a particular situation. For a given utility function type, they can represent different relations between the goods involved—e.g. that a particular ratio of one good to another is optimal. Again following economics, we propose that people employ the principle of rationality, connecting preference to choice.

In order to model reasoning about agents' choices in situations where the objects of choice interact with each other, we integrate the structured utility functions discussed above with Bayesian inference and the principle of rationality. We assume that the observed agent's preferences over a given set of objects is well-described by one of the utility functions discussed above; which of the utility functions is appropriate in a particular situation will depend on both the causal relations between the objects and the agent's goals. Whatever the utility function of the agent, we assume that she is approximately rational and softly maximizes her utility. The probability of a choice given her utility function is therefore given by the Luce-Choice rule:

$$P(c|u) \propto exp(\beta \cdot u(c)) \tag{4}$$

where u is the agent's utility function, c is a choice, and β is a noise parameter that determines how close the agent is to rational. We set β equal to 15 throughout.

We do not attempt to model how an observer might determine, before seeing any of an agent's choices, which utility function the agent is using in a particular situation. Rather, we assume that the observer has a prior distribution P(u) over the possible utility functions which apply in the situation. For example, we will not attempt to model how one might learn that bicycle wheels and bicycle frames are complements. In general, P(u) will depend on the observer's prior knowledge about the relations between the goods in question. The observer is also given a prior P(α) for each utility function u over the parameter alpha of the utility function. Throughout, we assume P(α) follows a Beta(2,2) distribution.

These assumptions allow us to use Bayesian inference to infer an agent's utility function from her choices. Given observed choices C, the posterior over utility functions is given by:

$$P(u|C) \propto P(C|u) \cdot P(u) \tag{5}$$

where u is the utility function with parameter α. This formula inverts the generative model of the agent's behavior captured by her utility function and the Luce-Choice rule (Equation 4). From observed choices C, we can infer the probability of a novel choice c by averaging over the agent's utility functions:

$$P(c|C) = \int P(c|u)P(u|C)du \tag{6}$$

Using the Luce-Choice rule to give the likelihood $P(c|u)$, we recover the mixed multinomial logit model which is used in Lucas et al. as well as many papers in econometrics (McFadden & Train, 2000).

Experiment

In the following experiment we test the quantitative predictions of the structured preference learning model. We designed scenarios in which an agent chose between bundles of goods, with each bundle containing *A* amount of one good and *B* amount of another good. In order to distinguish our model from previous utility-based models (Baker et al., 2007; Lucas et al., 2009) we varied the functional relationship between the goods described in the scenarios in order to match the three utility function types described above. Thus we constructed three separate groups of scenarios. In the Substitutes Group, the two goods were substitutes for each other. Hence, economists would model the preferences of an agent over these goods using a substitutes utility function (Equation 1). In the Complements Group of scenarios, the goods were complements for each other and so preferences over them would be modeled by a complements utility function (Equation 2). In the Cobb-Douglas Group, the goods were related in the way described by Cobb-Douglas utility function (Equation 3).

Our experiment aims to question whether people are intuitive economists. That is, whether they recognize the functional relationship between goods in a particular scenario and model an agent's preferences with a utility function over bundles of the goods that is appropriate to that functional relationship. A key prediction of the theory from economics is that different functional relationships will lead to different patterns of preference across bundles of the same size. For example, once you have two wheels for your bike, additional wheels are worth very little, while after two days at the beach, further days might still be very desirable. To test this, we designed scenarios for which the numerical properties of the bundles could be held fixed while the functional relationship between the goods in the bundles was varied.

Methods

Participants Participants were 480 individuals on Amazon's Mechanical Turk who received a small compensation for their time.

Materials and Procedure As noted above, we designed three groups of scenarios: Substitutes, Complements and Cobb-Douglas. Scenarios differed across the groups in how the goods being acquired related functionally, and hence in which utility function models preferences for the goods in the scenario.

An example stimulus for the Cobb-Douglas group is the following:

> Last year, John and his wife took a one-week vacation in the Caribbean. When John was booking the trip he had the choice between two package deals, which both cost the same amount:
>
> (A) 4 days on the beach and 2 days on a cruise
>
> (B) 3 days on the beach and 4 days on a cruise.
>
> John chose package (B) and thought that was the best choice given his options. This year John and his wife are planning another trip to the Caribbean. John needs to book the trip in advance. He has to decide between two package deals. Both deals cost the same amount:
>
> (C) 5 days on the beach and 5 days on a cruise
>
> (D) 8 days on the beach and 2 days on a cruise.
>
> Which option should John take?

As this example illustrates, the scenarios consisted of three parts: (1) setup of the first choice situation, (2) the agent chooses one bundle of goods over another, and (3) the agent faces a new choice situation. Each scenario was shownin 16 numerical conditions: these were the quantities of goods in each bundle in the two choice situations. In the example above, the bundle (3,4) was chosen over the bundle (4,2), and participants were asked to choose between the bundles (5,5) and (8,2). In different numerical conditions, the quantities in these bundles were varied. In pre-testing, we found that participants were making inferences about the cost of each bundle on the basis of bundle size; as a result, we subsequently stated in each scenario that the bundles being chosen from were the same price. Because this was incompatible with one of the scenarios, it was removed from subsequent testing. This left a total of 11 scenarios: three in the Cobb-Douglas Group, four in the Substitutes Group, and four in the Complements Group. Pre-testing also suggested possible order effects in some of the numerical conditions, which were

controlled in subsequent experiments. The scenarios were crossed with the numerical conditions, for a total of 11x16 inference questions.

Participants each received five or six distinct scenarios; numerical conditions were randomized across participants.

If people are sensitive to the structure of the agents' utility functions, then identical numerical conditions should give rise to different patterns of inference, depending on the scenario group. For example, if the goods are complements, then the observer should be uncertain about the optimal proportion of one good to the other. They will take the agent's initial choice as evidence about the optimal proportion, and will subsequently choose the bundle which fits this inferred proportion as closely as possible. On the other hand, if the goods are substitutes, then they will take the initial choice as evidence about how well each good serves its function, and will subsequently choose the bundle which maximizes the weighted sum of the two goods. If both bundles have the same size, then they will choose the one with the most of the better good.

Model Predictions

In order to make model predictions, we used the multinomial logit model (Equation 6) to find the probability of taking option (C) or (D) conditional on the agent's first choice. This probability was approximated through MCMC simulations, using the Metropolis-Hastings method.

We computed these probabilities under a variety of settings for the noise parameter β in the soft-max equation (Equation 4) and for the prior distribution Beta(α, α) over the utility functions' weight parameter. The noise parameter was set to 15 and α was set to 2 based on the fit of the model to the data over all conditions. Figure 1 shows the sensitivity of the model's predictions to the value of β. The mean squared error of the model varied between 0.04 and 0.1 as β varied between 0.1 and 20. We similarly analyzed the model's sensitivity to the value of α. We found that the mean squared error of the model varied between 0.04 and 0.1 as α varied between 1 and 10.

Our computational model predicts agent preferences by matching the appropriate utility function to the scenario. Thus, e.g. scenarios in the Complements group will be modeled with very strong prior probability on the Complements utility function and likewise for the other two groups.

In order to assess the distinctive predictions of our model, we implemented three heuristic models of preference inference. These heuristics implement rules for making preference inferences that are not based on calculations of utility. The heuristics worked as follows. The Sophisticated Ratio Heuristic was a model of ratio-matching, in which goods are selected based on how close their ratio is to an optimal ratio. The heuristic considers a range of possible optimal ratios and uses Bayesian inference to integrate over these ratios in making predictions. This differs from the complements function in being insensitive to absolute quantities of the goods. The Crude Ratio Heuristic was similar to the Sophisticated Ratio Heuristic, except that instead of considering a range of possible ideal ratios, it treats the ratio of goods in the observed choice as ideal. The Max Heuristic assumes that one of the two goods is preferred, and that agents will always choose the bundle that contains the greatest quantity of the preferred good. This differs from the substitutes utility function in being insensitive to the quantity of the non-preferred good. These three heuristics do not assign utilities to bundles but instead simply select one outcome as the best.

Table 1: Correlations between the three different utility models and subject judgments. Scenarios 1-3 were intended to be best fit by a Cobb-Douglas utility function; scenarios 4-7 by a Substitutes utility function; and scenarios 8-11 by a Complements utility function.

Scenario	Cobb	Substitutes	Complements
1	**0.79**	0.83	0.50
2	**0.62**	0.72	0.30
3	**0.80**	0.70	0.80
4	0.85	**0.95**	0.45
5	0.69	**0.83**	0.32
6	0.87	**0.91**	0.49
7	0.81	**0.93**	0.36
8	0.64	0.34	**0.87**
9	0.63	0.41	**0.86**
10	0.78	0.74	**0.69**
11	0.46	0.23	**0.87**

Table 2: Correlations between the best heuristic models and participant judgments.

Scenario	Crude Ratio	Max
1	0.55	0.89
2	0.57	0.73
3	0.91	0.76
4	0.48	0.95
5	0.37	0.76
6	0.61	0.93
7	0.55	0.96
8	0.82	0.44
9	0.94	0.49
10	0.70	0.80
11	0.87	0.26

Results. We first consider the model fit when all of the weight of the prior probability distribution is placed on a particular utility function type. As noted above, scenarios were grouped based on which utility function we believed would best fit the goods in the scenario. If participants are sensitive to the functional relationship between the goods in each scenario, then their judgments should be best fit when a high prior probability is placed on the group's utility function

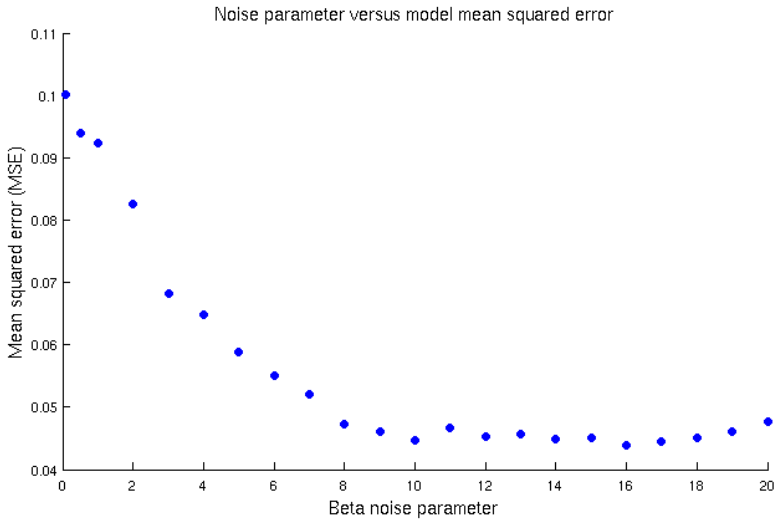


Figure 1: Sensitivity of model performance to the value of the noise parameter β.

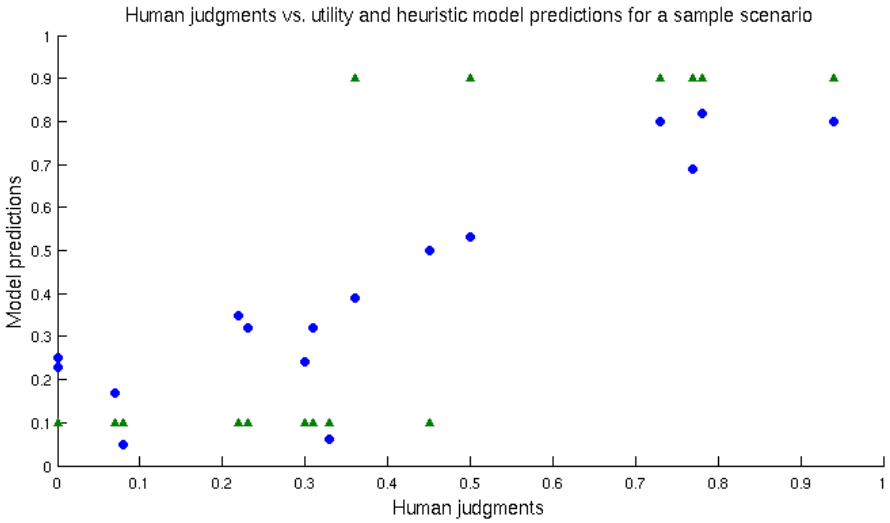


Figure 2: Human judgments vs. utility model predictions (blue dots) and heuristic predictions (green triangles) for a complements scenario. Other scenarios showed a similar relationship between human judgments and the predictions of the two models.

type. In all of the Substitutes scenarios and three of the four Complements scenarios, the human data was best fit by the scenario's corresponding utility function type. Participants' judgments in two of the three Cobb-Douglas scenarios are best fit by the substitutes utility function. These results are shown in Table 1.

We next tested the fit of our heuristic models. The free parameters of the heuristics were chosen to maximize overall correlation with the human data. The Crude Ratio and Max heuristics had the highest correlations, so we restrict our attention to them. The Max Heuristic had high (above 0.7) correlations with human judgments on all of the Cobb-Douglas and Substitutes scenarios. The Crude Ratio Heuristic had similarly high correlations on all of the Complements scenarios. These results are shown in Table 2.

Discussion

Our experiment aimed to investigate two questions. The first was whether people are sensitive to the functional relationships between goods when they make inferences about people's preferences. The second was, if people are sensitive to these functional relationships, what do they learn about others' preferences from their observations? Our results provide evidence that participants were sensitive to the func-

tional relationships between goods. In particular, participants appeared to distinguish between the Complements scenarios and the other scenarios, although we did not find evidence that they distinguish between the Cobb-Douglas and Substitutes scenarios. This is indicated by the fit of both the utility and heuristic models. A single utility model (with all of its prior weight on the substitutes utility function) and a single heuristic (the Max Heuristic) correlate well with human judgments in the Cobb-Douglas and Substitutes scenarios. On the other hand, these two models fare poorly on most of the Complements scenarios. The complements utility model and the Crude Ratio Heuristic provide better fits for these scenarios.

The performance of the Max Heuristic on a particular scenario is very well predicted by the performance of the substitutes utility function on that scenario; the same is true of the Crude Ratio Heuristic and the complements utility function. The Max Heuristic naively tracks a relationship which is important to the substitutes utility function, namely which good is favored over the other. The Crude Ratio Heuristic gives a noisy estimate of the optimal ratio between goods, which is similarly important to the complements utility function. This provides evidence that participants were sensitive to the optimal ratio of goods in the Complements scenarios, and to which was the preferred good in the other scenarios.

It is an interesting question whether the utility-based or heuristic models provides a more promising approach to modeling preference learning. The best heuristic models were able to capture the qualitative shape of participants' judgments. The heuristics were able to correctly classify bundles as more or less likely to be chosen. A representative scenario is shown in Figure 2. On the other hand, human judgments showed a gradedness that was better captured by the utility models. As predicted by the utility models, participants varied across numerical conditions in how strongly they thought one bundle should be chosen over another. This is also shown in Figure 2.

Conclusion

We have presented a computational model of structured preference learning. Our model incorporates Bayesian inference with economic models of internally complex preferences. Using a standard tool from econometrics, it can be used to model how individuals infer what someone prefers on the basis of past observations. We experimentally tested the model, and found that participants were sensitive to the functional relationships between goods in some of the ways that the model predicts. Future experiments could help to differentiate the utility-based and heuristic models. These models will make distinct predictions, for example, when the sizes of the bundles being chosen between are different. Bigger bundles will, all things being equal, be preferred by the utility-based models, while this is not always true for the heuristics.

Further work needs to be done in systematically varying the structure of the preferences being learned. This may be done by explicitly varying the causal structures of the choice scenarios so that subjects' prior beliefs about these relations do not come into play. This suggests a further line of research: studying whether individuals can make use of more complex functional relationships than the ones that were considered here. If individuals can effectively learn others' preferences in a range of situations, then it is unlikely that these preferences will be captured by the simple functional forms studied here.

References

Baker, C. L., Tenenbaum, J. B., Saxe, R. R. (2007). Goal inference as inverse planning. In *Proceedings of the twenty ninth annual conference of the Cognitive Science Society.*

Kushnir, T., Xu, F., Wellman, H. (2008). Preschoolers use sampling information to infer the preferences of others. In *Proceedings of the thirtieth annual conference of the Cognitive Science Society.*

Lucas, C., Griffiths, T. L., Xu, F., Fawcett, C. (2009). A rational model of preference learning and choice prediction by children.In *Advances in Neural Information Processing Systems* 21.

Mas-Colell, A., Whinston, M. D., Green, J. R. (1995). *Microeconomic Theory*. Oxford University Press.

McFadden, D. Train, K. E. (2000). Mixed MNL models of discrete response. *Journal of Applied Econometrics*, 15, 447-470.

Repacholi, B. M., Gopnik, A. (1997) Early reasoning about desires: Evidence from 14- and 18-month-olds. *Developmental Psychology*, 33(1), 12-21.

Beyond Boolean logic: exploring representation languages for learning complex concepts

Steven T. Piantadosi, Joshua B. Tenenbaum, Noah D. Goodman
{ **piantado, jbt, ndg** } **@ mit.edu**
MIT Department of Brain and Cognitive Sciences
43 Vassar Street, Building 46, Room 3037G, Cambridge, MA 02139

Abstract

We study concept learning for semantically-motivated, set-theoretic concepts. We first present an experiment in which we show that subjects learn concepts which cannot be represented by a simple Boolean logic. We then present a computational model which is similarly capable of learning these concepts, and show that it provides a good fit to human learning curves. Additionally, we compare the performance of several potential representation languages which are richer than Boolean logic in predicting human response distributions.
Keywords: Rule-based concept learning; probabilistic model; semantics.

Introduction

Every cognitive theory requires a hypothesis about mental representation—what structures and operations form the basis for complex ideas. High-level, symbolic theories often characterize this representation using a *representation language (RL)* (Fodor 1975) that specifies primitive elements and composition laws which can be used to form complex cognitive structures. This approach has been extensively studied within two traditions in cognitive science: concept learning and linguistic semantics. In models of rule-based concept learning (Bruner, Goodnow, & Austin 1956, Shepard, Hovland, & Jenkins 1961, Feldman 2000) the representation language has typically been simple Boolean logic, which represents concepts—stable mental representations—using conjunctions, disjunctions, and negation of simple perceptual primitives. Goodman et al. (2008) presented a model of probabilistic learning for rule-based concepts that represents concepts in a simple propositional language and achieves state-of-the-art fits to experimentally-measured difficulty of learning Boolean concepts. The logical complexity of concepts appears to play a crucial role in determining how these concepts are learned: learners are biased to preferentially learn concepts with simpler representations.

However, simple Boolean concepts can capture only a very limited range of the human conceptual repertoire. People readily conceptualize context-dependent meanings such as "happiest," and can form more complex and abstract relational concepts like "everyone with two or more siblings." Semantic theories capture such meanings using primitive operations which manipulate and quantify over sets of objects, rather than simply features and propositional connectives. The denotation of a quantifier like "some," for instance, is a function which takes two sets, A and B, and is true only when the intersection of $A \cap B$ is nonempty[1] (Montague 1974).

[1] In a sentence like "Some boy smiled," the set of boys would be A and the set of things which smiled would be set B. "Some boy smiled" is true if and only if the intersection of boys and things which smiled is nonempty.

In this paper we extend the probabilistic approach to concept induction to representation languages which manipulate sets of objects. We first describe an experiment that explores the difficulty of learning concepts that involve set-manipulation and quantification. Second, we compare human difficulty to the predictions of models with varying RLs. Our modeling work has two goals: the first is to test different RLs to see which provide the best account of people's learning behavior. Each possible RL differs in representational power and the way in which it assigns probability to potential concepts. This means that different RLs make different predictions about people's learning trajectories and we can therefore compare RLs by determining how well they match subjects' empirical response distributions. The second goal of the modeling work is to provide an explicit *learning theory* for these concepts. Work on boolean concept learning has provided a probabilistic model which accounts for subjects' behavior in acquiring boolean concepts, but there is no comparable formal theory for concepts which require a richer representation language. Such a theory would importantly extend rule-based concept learning in cognitive science to richer, linguistically-interesting semantic representations.

Behavioral experiment

The experiment we present aims to extend the rule-based concept learning paradigm to concepts which refer to sets and properties of sets of objects. To do this, we used a learning paradigm where subjects see a set of objects, guess at a labeling of the objects according to the unknown target concept, and receive feedback on their responses. Subjects used this feedback to infer the target concept.

Procedure

Amazon's Mechanical Turk was used to run 381 subjects. Each subject was told that they had to learn the meaning of a novel word, *wudsy*, from an alien language. Subjects were told that aliens use *wudsy*, to refer to some objects in a collection of objects, and that they have to figure out what makes an object *wudsy*. Subjects were informed that what makes an object *wudsy* may depend on which other objects are present.

During the experiment subjects were shown a set of four objects which varied in size, color, shape, and background color. An example set of items is shown below:

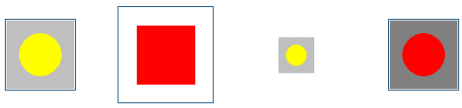


Figure 1: An example set.

After seeing a set of objects, subjects were told to guess which objects were the *wudsy* ones. For each object in the set, they were required to response "Yes," "No," or "NA," and were told to respond "NA" when it is unspecified whether an object is *wudsy*. For this example, subjects might entertain the concept is "red objects," in which case they should respond that the second and fourth objects are in the concept and the first and third are not. However, subjects might also entertain that the meaning of *wudsy* is context-dependent, as in, for instance "unique smallest." Similarly, the concept may also be complex, such as "same shape as the object with the darkest background." The shape with the darkest background is a circle, so subjects should say all the circles are in the concept; if all backgrounds are the same color, subjects may respond "NA." After responding, subjects were told what the correct answer was according to the target concept, but never given explicit instruction on the target concept. Subjects who responded incorrectly to any element of the set were penalized with a 5 second delay, during which they saw the set and the correct responses for each object.

Materials & Concepts

The meanings subjects were required to learn consisted of the concepts shown in Figure 2. These concepts include simple boolean rule-based concepts (e.g. "circles" and "circles or blue objects"), as well as more complex concepts which cannot be expressed in boolean logic ("larger than all the other objects"), and concepts which require several bound variables to express ("Same shape as the largest blue object").

Several of the concepts we studied focus on size predicates. This is because size predicates, such as "largest" and "smallest," are salient properties of objects in sets. They are perhaps the simplest words whose meaning is context-sensitive, and therefore not expressible with only conjunctions, disjunctions, and negation of object features. We included three simple size relations, "there exists a smaller object," "larger than all other objects," and "one of the largest objects." Note that the latter two differ with respect to uniqueness: if there are two objects of the maximal size, then neither is larger than all other objects, but both are one of the largest[2].

Because we included these simple size predicates, it is natural to include complex concepts which are also based on size, such as "same shape as the largest object," "same shape as the largest blue object," and "unique largest blue object." All three of these concepts require finding the largest object and selecting other elements based on the properties of the largest. As such, they require answering *NA* when there is not a unique largest element[3].

Results

The plots in Figure 2 show subjects' accuracy at labeling which objects are *wudsy* (y-axis), as a function of the amount of labeled data they received (x-axis). Subjects who were more than 3 standard deviations below the mean accuracy for each concept were removed in order to exclude subjects who were not performing the task. The vertical error bars show binomial 95% confidence intervals, and the red lines show the best fitting model, which is discussed in the next section. These results reveal several interesting qualitative trends. First, subjects accuracies increase for almost all of the concepts. Importantly, even though the subjects receive labeled data, they are never explicitly instructed on the concept. This means that high accuracy can only be achieved by generalizing from the observed data, which requires inferring abstract rules for these concepts.

Two interesting exceptions to subjects' general ability to learn these concepts are "Everything iff there is a triangle" and "Everything iff there is a single blue object." Subject performance on these concepts does not substantially improve, and these are intuitively somewhat unnatural concepts which require all elements of a set to be selected based on what the set contains. Words do exist with similar denotations in English—for instance, a set is *contaminated* if one element of the set is bad—but subjects find these types of concepts unusually hard to learn.

Figure 2 reveals a number of places where subject performance drops temporarily for a single set–for instance at item 32 of "there exists a smaller object." Post-hoc analysis revealed that many of these dips are caused when subjects see one of the first exceptions to a plausible alternative concept: item 32 is the first time that all objects in the set are the same size. Subjects responded *true* to objects in this item, consistent with a concept such as "not smaller than the rest," but incorrect according to the target concept.

Analysis

We first used a regression to analyze how subjects' learning rate varied across the 12 concepts studied[4]. In each logistic regression the outcome was whether the subject's response to each object in a set was correct, and the independent variable was the number of items each subject had seen so far $(0 \ldots 70)$. The key prediction we tested is whether slopes (regression coefficients)—which quantify the effect of additional data on accuracy—differed between concepts.

[2]These concepts are interesting in part because it is unclear which of these meanings corresponds to the denotation of "largest" in English, and also what role pragmatics plays in understanding "largest" in normal conversation.

[3]This makes it difficult to compare these concepts with the simple size-predicate concepts since the latter never require *NA*, which may be a difficult response for subjects to learn, independent of the concept.

[4]Because subjects typically were only run on one concept, subject effects are confounded with concept. We therefore performed a separate mixed-effect logistic regression (Gelman & Hill 2007) *within* each concept including slopes and intercepts by subjects. Regression coefficients across concepts were compared using t-tests.

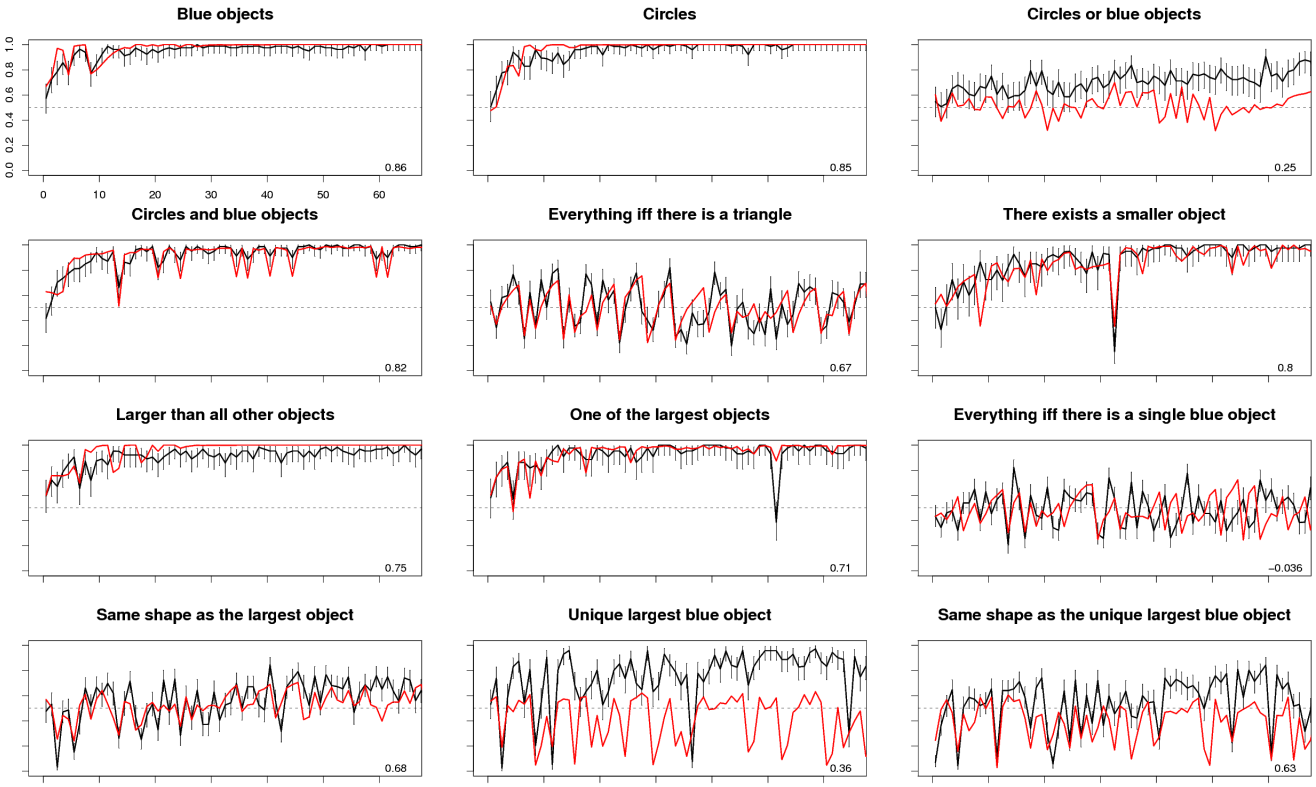


Figure 2: Subject accuracy (y-axis) in labeling the *wudsy* objects as a function of trial number (x-axis). Black lines show subject mean percent correct. Error bars are 95% confidence intervals. The red line shows the best-fitting model, although note that the model is fit based on agreement with the full distribution of responses, not the accuracies shown here. Numbers in the lower right show the correlation between the model and human accuracies.

Our results replicate basic effects in Boolean concept learning (Bruner, Goodnow, & Austin 1956, Shepard, Hovland, & Jenkins 1961). As is clear from Figure 2, simple concepts ("blue objects") are easier to learn than complex concepts ("circles and blue objects") ($t = 2.70, p < 0.01$). In addition, our results replicate that conjunctions ("circles and blue objects") are easier to learn than disjunctions ("circles or blue objects") ($t = 3.10, p < 0.01$). These replications provide validation for our experimental paradigm.

These effects of complexity also generalized to more complex functions than those expressible in Boolean logic. For instance, "The unique largest blue object" was easier to learn than "The same shape as the unique largest blue object" ($t = 2.71, p < 0.01$). This effect is interesting because it shows the additional difficulty associated with more complex set-theoretic concepts. The latter concept requires an additional bound variable to express in first-order logic, or a lambda abstraction to express in lambda calculus, and the effect of this complexity is reflected in subjects' learning rates. The regression revealed no difference between uniqueness presuppositions for concepts involving the largest element of a set: "larger than all other objects" was no more difficult than "one of the largest objects" in either slopes ($t < 1.26, p > 0.20$) or intercepts ($t < 1.83, p > 0.05$).

Importantly, subjects may infer a *different* concept from the one that was used to generate the data—high accuracy on some concepts can be achieved by inferring related concepts. To address this, we compared how well closely-related concepts predicted subjects' responses in the last half of the experiment. For each target concept in Table 1, we looked at data points for which the target concept made a different prediction from the specified alternative hypothesis. For instance, we looked at sets for which "Largest blue object" and "blue objects" made different predictions—that is, when there are multiple blue objects, so not all of them are the largest. We then computed the percent of subjects who responded more than half the time in agreement with the target concept, as well as the overall proportion of time subjects responded with the target concept. These results show that for most of the concepts, subjects typically responded in accord with the target concept and not a close alternative. The only exception to this is the comparison between "One of the largest objects" and "size = 5" (In the items, "5" is the maximal size for objects), which showed that subjects may have been learning to identify objects based on comparing them to absolute size, rather than a context-sensitive measure of "largest". In general, though, these results show that subject's pattern of learning cannot be explained by simpler theories which make reference only to only individual objects' properties. This is especially striking given that many of the alter-

Target	Alternative	Subject pct.	Response Pct.
Larger than all other objects	One of the largest objects	1.00	0.94
	Size = 5	0.81	0.67
	Size ≥ 4	1.00	0.82
	Size ≥ 3	1.00	0.88
Unique largest blue object	One of the largest objects	0.79	0.66
	Blue Objects	0.79	0.74
	Larger than all other objects	0.79	0.63
Same shape as the unique largest blue object	Same shape as the largest object	0.52	0.58
One of the largest objects	Size = 5	0.36	0.46
	Size ≥ 4	1.00	0.67
	Size ≥ 3	1.00	0.67
There exists a smaller object	Size = 5	1.00	0.67
	Size ≥ 4	1.00	0.73
	Size ≥ 3	1.00	0.91

Table 1: Comparison of subject agreement with target concepts compared to alternative concepts. *Subject pct.* shows the proportion of subjects agreeing more than 50% with the target concept, *Response Pct.* shows the overall percent agreement with the target.

native hypotheses are much simpler than the target concepts, and provides strong evidence that subjects are attending to more than simple object properties.

Computational model

The behavioral experiment shows that generally subjects are able to induce these types of set-theoretic concepts from the labeled data. Although it is important that subjects can eventually learn most of these concepts, we are also interested in whether their learning trajectory—their guesses and hypothesized concepts at each point in time—follow sensibly from the observed data. It may be rational to initially learn simpler related concepts which give approximately correct answers. There is no guarantee, for instance, that 70 items are enough to justify learning the correct form of the target concepts. We next present a computational model which can learn these types of set-theoretic concepts.

Our computational model aims to extend the *rational rules* model of Goodman et al. (2008) to a richer hypothesis space—one which is capable of representing these types of set-theoretic concepts. The probabilistic structure of the model and inference algorithm we use are neutral with respect to the RL, meaning that any potential RL can be incorporated and tested to see what distribution of responses it predicts for each object in each concept.

Each potential RL $\mathcal{L}$ defines a hypothesis space of potential concepts corresponding to the set of all ways to compose the RL's primitive functions in order to create functions which map objects to labels (*true*, *false*, *NA*). For instance one concept might be[5]

$$\lambda x.(red\ x) \wedge (square\ x)$$

This expression represents a function which checks whether its argument, x, is red and a square, and returns *true* or *false* for any object (since *red* and *square* are assumed to return *true* or *false*). We might also have concepts which take two arguments, a contextually-relevant set S and an element x:

$$\lambda S\ \lambda x.(equal\ x\ (unique\text{-}smallest\ S))$$

This function checks if x is the object which is the unique, smallest element of S.

We use a probabilistic context-free grammar (PCFG) which assigns probability to every possible composition of primitive elements. This PCFG functions as a prior over concepts and for simplicity, we assume that all PCFG expansions are equally probable[6]. In general, the PCFG assigns high probability to short or "simple" compositions of $\mathcal{L}$'s primitives, and lower probability to complex rules. For instance, a function $\lambda x.(red\ x)$ will be higher probability a-priori than $\lambda x.(red\ x) \wedge ((square\ x) \vee (circle\ x))$. This captures the notion that people should be biased to prefer simple explanations of the labeled data they observe.

The second part of the model is a likelihood function which provides the probability of labels according to a hypothesized RL expression. Specifically, for any composition E of primitives in $\mathcal{L}$, the correct label is generated with probability α, and a label is chosen uniformly at random with probability $1-\alpha$. However, it is also likely that memory factors come into play in remembering past labeled examples. We include this in the model by weighting the log likelihood for the n'th data point back in time by $n^{-\beta}$, where $\beta > 0$. As $\beta \to 0$, the model has perfect memory, and as $\beta \to \infty$ the model quickly forgets past data points. This leaves us with two unknown free parameters: α, which controls how reliably set elements are labeled, and β which controls how much more recent set elements matter than past ones.

Together, the prior and likelihood specify a complete probabilistic model for any RL. Formally, we can score the probability of a hypothesized concept expression E conditioned on a collection of example sets S with corresponding labels L according to Bayes rule:

$$P(E \mid S, L, \mathcal{L}) \propto P(L \mid S, E) P(E \mid \mathcal{L}). \quad (1)$$

Here, $P(E \mid \mathcal{L})$ is the probability of E according to the PCFG for $\mathcal{L}$ and $P(L \mid S, E)$ scores the likelihood of the labels L under the observed sets of objects S and hypothesized expression E. While Equation 1 scores the probability of any given expression E, it is a complex inference problem to actually determine what expressions are likely given the data. This problem is difficult because the space of possible expressions E is in principle infinite and difficult to search. We solved this problem using a Markov-Chain Monte-Carlo (MCMC) similar to Goodman et al (2008)'s method, which takes samples

[5]We write functions as lambda expressions, meaning that the name for the argument is preceded by λ. We also use prefix notation: a function f applied to an argument x is written $(f\ x)$.

[6]Unlike the rational-rules model, we do not integrate out the PCFG production probabilities. This is because primitives which introduce new bound variables, such as quantifiers, make this integration difficult and potentially not analytically tractable in general.

from the posterior distribution $P(E \mid S, L, \mathcal{L})$. This method takes a biased random walk around the space of hypotheses by making local changes to hypothesized expressions E, and can be shown to, in the limit, draw samples from the posterior distribution. We ran the MCMC algorithm for a range of α and β values for each amount of data, in each sequence, conditioning on the correct, observed labels for all previous sets in the sequence. This gives a distribution $P(E \mid S, L, \mathcal{L})$ on expressions E in the RL $\mathcal{L}$ at each point during learning. These expressions can be evaluated on the next item in order to provide a model prediction of subject's distribution of responses, conditioned on the observed labeled data. Thus, the model was run conditioned on the same labeled data human subjects were given, and—just like human subjects—was asked to make predictions about the correct labels for the next data point. Ideally, subjects' distribution of responses at each point in time during learning should correspond to the predictions of the model, conditioned on the exact same sequence of training data.

One goal of the model is to test different representational languages to see which provide the best theory of people's inductive biases in learning these concepts. We computed the posterior predictive distribution of responses for each representation language $\mathcal{L}$ and saw which assigned the human responses highest likelihood[7]. We compared four different RLs with differing primitives and representational power:

Language	**Primitive Operations**
RESPONSE-BIASED	*true*, *false*, *undefined*
SIMPLE-BOOLEAN	*and*, *or*, *not*, *shape*, *size*, *color*, *background-color*, *equal*
SET-FUNCTIONS	*contains*, *filter*, *only*, *unique-largest*, *unique-smallest*, *set-of-largest*, *set-of-smallest*, *same-object*
QUANTIFIERS	*exists*, *forall*

Each RL is a *superset* of the preceding languages, except that none other than RESPONSE-BIASED contain *true*, *false*, and *na* as primitives. Here, *shape*, *color*, and *background-color* are functions which extract the corresponding properties of objects. *equal* tests if two properties are equal. *contains* returns true if a set contains an element, *filter* removes all elements in a set not satisfying a predicate, and *only* return the only element of a set and *NA* if the set has more than one element. The primitives *unique-largest* and *set-of-largest* return the unique largest element in a set (and *NA* otherwise), and the set of elements for which none are larger, respectively. *same-object* tests if two objects are identical on all dimensions. *exists* and *forall* are first-order existential and universal quantifiers.

Intuitively, the RESPONSE-BIASED language allows learners only to infer a distribution on responses, but not give responses which depend on the current objects. This serves as a baseline, and way to test if subjects are really performing the task. The SIMPLE-BOOLEAN language is one which include basic logical operations and object properties, and implements the representational system studied most in previous rule-based concept learning experiments. The SET-FUNCTIONS language extends the SIMPLE-BOOLEAN language by including primitive operation for testing if sets contain elements, extracting sets or elements with maximal or minimal properties along the size-dimension and filtering sets by elements. The QUANTIFIERS language extends the SET-FUNCTIONS language by incorporating quantification.

Results & Discussion

Table 2 shows the performance of these models in predicting the human distribution of responses across the 12 concepts studied. This shows the average log-likelihood of the human responses for the best-fitting values of α and β within each concept[8]. This table illustrates several key properties of the RLs. First, the RESPONSE-BIASED model is overall the worst predictor of human responses. This is important because it shows that subjects are performing the task, and performing nontrivial inferences about the target concepts.

In addition, this figure shows that while SIMPLE-BOOLEAN is a good predictor for the simple Boolean concepts, SET-FUNCTIONS and QUANTIFIERS provide a better account for the set-theoretic concepts that subjects are able to learn. SIMPLE-BOOLEAN provides the worst account for "same shape as the largest object" and "same shape as the unique largest blue object." While subjects do not learn these concepts especially well, these results show that the SIMPLE-BOOLEAN does not account well for subject responses.

Overall, the best RL is QUANTIFIERS; however, the differences between QUANTIFIERS and SET-FUNCTIONS is small. Richer representation languages not only have the formal power to represent the types of set-theoretic and logical concepts required by human conceptual systems, but also provide a better account of human inductive leaning than the other RLs considered here.

As discussed above, the black line in Figure 2 shows learning curves showing percent accuracy over time for human subjects. This figure also shows a red line, corresponding to the performance of the RL QUANTIFIERS for the best-fitting α and β within each concept. We chose the best-fitting model parameters based on which parameter values assigned highest likelihood to the observed *distribution* of human responses, "true," "false," and "NA." Doing this does not necessarily provide the best fit to the human learning curves in Figure 2 since the model is not fit to human *accuracy* (correct/incorrect). This means that Figure 2 shows a conservative view of the agreement between human accuracies and model accuracies. For the concepts "circles or blue objects" and "unique largest blue object" the model's learning trajectory would increase

[7]Model predictive distributions were smoothed to give each response a minimum possible probability of 0.01, to prevent divergence.

[8]That is, these numbers are the total log likelihood assigned to human responses, divided by the number of responses. This was necessary for cross-concept comparison since concepts may have differing numbers of subject responses.

Concept	RESPONSE-BIASED	SIMPLE-BOOLEAN	SET-FUNCTIONS	QUANTIFIERS
Blue objects	-0.66	-0.18	-0.19	-0.18
Circles	-0.73	-0.17	-0.17	-0.17
Circles or blue objects	-0.81	-0.73	-0.74	-0.74
Circles and blue objects	-0.30	-0.27	-0.27	-0.27
Everything iff there is a triangle	-0.80	-0.73	-0.73	-0.73
There exists a smaller object	-0.81	-0.51	-0.40	-0.41
Larger than all other objects	-0.58	-0.48	-0.46	-0.36
One of the largest objects	-0.80	-0.63	-0.28	-0.28
Everything iff there is a single blue object	-0.85	-0.78	-0.78	-0.78
Same shape as the largest object	-1.10	-1.75	-0.99	-0.99
Unique largest blue object	-1.05	-1.54	-1.06	-1.06
Same shape as the unique largest blue object	-1.08	-1.34	-1.06	-1.04
Mean	**-0.797**	**-0.760**	**-0.594**	**-0.584**

Table 2: Model log likelihoods *per response* for each concept. These represent the model log likelihood assigned to human responses, divided by the total number of responses in each concept to allow comparisons across concepts.

more for other values of α and β, and thus look more like subject's accuracies, but provide a less-good fit to subjects' overall response distribution.

This figure shows good fit between the probabilistic model and human learning. This fit appears especially remarkable for concepts which subjects have a difficult time learning, such as "Everything iff there is a triangle." Because subjects do not learn this concept well, the best-fitting α is low and β is highly negative, meaning that the model is not penalized much for incorrect answers and down-weights old data. The model therefore responds with in simple ways, such as always responding true, or responding true to only the triangles; subjects appear to use similar strategies, and thus both show similar patterns of response accuracies[9]

The model also shows more subtle agreement patterns with human subjects. First, it is capable of learning simple boolean concepts in a way similar to humans, quickly arriving at the correct meaning given the training data. This is also true for concepts like "there exists a smaller object" and the other size-predicates. The model also matches local dips and peaks in reasonably well. This is because the model, like people, may temporarily be led to a concept which is not the target concept, just as subjects (e.g. at item 32 of "there exists a smaller object"). This provides evidence that people make the same rational, statistical inferences given the same data.

Conclusion

While the SIMPLE-BOOLEAN RL provided a good fit to human response data in some cases, it is insufficient to represent some of the complex concepts that subjects learned. Subjects' ability to learn these concepts was demonstrated by their learning curves for several context-dependent concepts. The comparison of different RLs suggests a potentially fruitful approach to discovering the precise form of semantic representations. Recently, Pietroski et al. (2009) and Hackl (2009) have used psychophysical measures to make inferences about plausible representations and computations that underlie semantic meaning for words like "most." Our work provides a complementary approach to the same problem—instead of measuring response times, we studied what RLs provide a good account of human inductive biases during learning. This method may be broadly applicable to discovering the form of semantic representations in natural language.

Of course, the RLs we study here are still incomplete with respect to the full richness of human conceptual systems; however, this work suggests that rule-based concept-learning can be extended to complex concepts which can begin to approach the complexity and context-dependence observed in human linguistic systems. Furthermore, the model provides one potential acquisition theory for semantic concepts. Children may learn semantic meanings like adults in our experiment did—by inducing concepts in a sufficiently-powerful compositional RL.

[9]The best fitting β also shows a modest negative correlation ($R = -0.55$, $p = 0.06$, $N = 12$) with response accuracies over the 12 concepts, suggesting an interaction between the target concept and the attentional or memory resources people allocate.

References

Bruner, J. S., Goodnow, J., & Austin, G. (1956). *A study of thinking*. New York: Wiley.

Feldman, J. (2000). Minimization of boolean complexity in human concept learning. *Nature*, *407*, 630-633.

Fodor, J. (1975). *The language of thought*. Cambridge, MA: Harvard University Press.

Goodman, N., Tenenbaum, J., Feldman, J., & Griffiths, T. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, *32*, 108-154.

Hackl, M. (2009). On the grammar and processing of proportional quantifiers: most versus more than half. *Natural Language Semantics*, *17*, 63-98.

Montague, R. (2002). The proper treatment of quantification in english. In P. Portner & B. H. Partee (Eds.), *Formal semantics: The essential readings.* Oxford: Blackwell.

Pietroski, P., Lidz, J., Hunter, T., & Halberda, J. (2009). The meaning of most: semantics, numerosity, and psychology. *Mind and Language*, *24*, 554-585.

Shepard, R., Hovland, C., & Jenkins, H. (1961). Learning and memorization of classifications. *Psychol. Monogr. Gen. Appl.*, *75*, 1-42.

Encoding Sequential Information in Vector Space Models of Semantics: Comparing Holographic Reduced Representation and Random Permutation

Gabriel Recchia (grecchia@indiana.edu)
Cognitive Science Program, 1910 E 10th St.
Indiana University, Bloomington, Indiana USA

Michael N. Jones (jonesmn@indiana.edu)
Department of Psychological and Brain Sciences
Indiana University, Bloomington, Indiana USA

Magnus Sahlgren (mange@sics.se)
Swedish Institute of Computer Science
Box 1263, SE-164 29 Kista, Sweden

Pentti Kanerva (pkanerva@berkeley.edu)
Redwood Center for Theoretical Neuroscience
University of California, Berkeley, California, USA

Abstract

Encoding information about the order in which words typically appear has been shown to improve the performance of high-dimensional semantic space models. This requires an encoding operation capable of binding together vectors in an order-sensitive way, and efficient enough to scale to large text corpora. Although both circular convolution and random permutations have been enlisted for this purpose in semantic models, these operations have never been systematically compared. In Experiment 1 we compare their storage capacity and probability of correct retrieval; in Experiments 2 and 3 we compare their performance on semantic tasks when integrated into existing models. We conclude that random permutations are a scalable alternative to circular convolution with several desirable properties.

Keywords: semantic representation, semantic space models, binding, convolution, permutation, random indexing.

Introduction

Vector-space models of lexical semantics have seen considerable recent attention in the psychological literature both as automated tools to estimate semantic similarity between words, and as psychological models of how humans learn and represent word meaning from repeated contextual co-occurrences. In general, these models build semantic representations for words from statistical redundancies observed in a large corpus of text (e.g., Landauer & Dumais, 1997; Lund & Burgess, 1996). As tools, the models have provided invaluable metrics of semantic similarity for stimulus selection and control in behavioral experiments using words, sentences, and larger units of discourse. As psychological models, the vectors derived from distributional models serve as useful representations in computational models of word recognition, priming, and higher-order comprehension (Landauer et al., 2007). In addition, the abstraction algorithms themselves are often proposed as models of the cognitive mechanisms humans use to learn meaning from repeated episodic experience.

A classic example of a vector-space model is Landauer and Dumais' (1997) Latent Semantic Analysis (LSA). LSA begins with a word-by-document co-occurrence matrix representation of a text corpus. A lexical association function is applied to dampen the importance of each word proportionate to its entropy over documents. Finally, an algorithm is applied to reduce the matrix's dimensionality; words are represented as vectors whose dimensions refer to the largest eigenvalues of the reduced representation.

Despite their successes both as tools and as psychological models, vector-space models suffer from several shortcomings. Most prominently, the models have been criticized as "bag of words" models that encode only the contexts in which words co-occur, but ignore *word-order information*. The role of word order was traditionally thought to apply only to the rules of word usage (grammar) rather than to the lexical meaning of the word itself. However, temporal information is now taking a more prominent role in the lexical representation of a word's meaning. Recently, Elman (2009) has convincingly argued that an inherent part of a word's lexical representation is information about its common temporal context, event knowledge, and habits of usage (cf. McKoon & Ratcliff, 2003; see also Hare et al., 2009).

A second issue for these models is *lack of scalability* (Recchia & Jones, 2009; Kanerva, Kristofersson, & Holst, 2000), due to reliance on computationally complex decomposition techniques to reveal the latent components in a word-by-document matrix (e.g., singular value decomposition). Not only is decomposition computationally expensive, the entire word-by-document matrix must be stored in memory during the operation. The problem is exacerbated by the fact that as the size of the corpus increases, the number of both rows and columns in the matrix increase significantly, the number of columns growing linearly with added documents, and the number of rows growing approximately in proportion to the square root of the number of tokens (Heap's law). The corpora that vector-space models like LSA are most commonly trained upon in the literature contain approximately the number of tokens that children are estimated to have experienced before age 3, not counting words that they produce during this time (Riordan & Jones, 2007; Risley & Hart, 2006). Recently, Recchia and Jones (2009) demonstrated that although simple semantic metrics such as pointwise mutual information (PMI) are outperformed by more complex models such as LSA on small corpora, PMI is capable of much better correspondence to human-derived semantic similarity judgments due to its ability to scale to large corpora. This led the authors to favor simple and scalable

algorithms to more complex non-scalable algorithms, concordant with approaches that have met with success in the computational linguistics literature (e.g. Banko & Brill, 2001).

Encoding Word Order

Two recent vector-space models that directly address the concerns of word order and scalability are Jones and Mewhort's BEAGLE (2007) and the "random permutation" model of Sahlgren, Holst, and Kanerva (2008) (henceforth referred to as RPM). Rather than starting with a word-by-document matrix, BEAGLE and RPM maintain static, randomly generated *signal vectors* intended to represent the invariant properties of each word (such as its orthography or phonology), as well as dynamic *memory vectors* that store information about each word's semantic representation. To represent statistical information about the word, BEAGLE and RPM bind together collections of signal vectors into *order vectors* that are added to memory vectors during training. Integrating word-order information has yielded greater success at fitting a variety of human semantic data than encoding only contextual information (e.g., Jones, Kintsch, & Mewhort, 2006; Jones & Mewhort, 2007). Because they require neither the overhead of a large word-by-document matrix nor computationally intensive matrix decomposition techniques, both models are significantly more scalable than traditional vector-space models.

Although BEAGLE and RPM differ in several ways, arguably the most important difference lies in the nature of the binding operation used to create order vectors. BEAGLE uses circular convolution, a binary operation (henceforth denoted as *) performed on two vectors such that every element i of $(x * y)$ is given by:

$$\sum_{j=0}^{D-1} x_j \cdot y_{(i-j) \bmod D}, \qquad (1)$$

where D is the dimensionality of x and y. Circular convolution can be seen as a modulo-n variation of the tensor product of two vectors x and y such that $(x * y)$ is of the same dimensionality as x and y. Furthermore, although $(x * y)$ is dissimilar from both x and y by any distance metric, approximations of x and y can be retrieved via the inverse operation of correlation.

In contrast, RPM employs *random permutations*, henceforth referred to as RPs. True to their name, RPs are functions that map input vectors to output vectors such that the outputs are simply randomly shuffled versions of the inputs. Just as $(x * y)$ yields a vector that differs from x and y but from which approximations of x and y can be retrieved, the sum of two RPs of x and y, $\Pi x + \Pi^2 y$ (where $\Pi^2 y$ is defined as $\Pi(\Pi y)$) yields a vector dissimilar from x and y but from which approximations of the original x and y can be retrieved via the inverse permutations Π^{-1} and Π^{-2}.

Both systems offer efficient storage properties, compressing order information into a single composite vector representation, and both encoding operations are reversible. However, RPs are much more efficient to compute. In language applications of BEAGLE, the computationally expensive convolution operation is what limits the size of a text corpus that the model can encode. As Recchia and Jones (2009) have demonstrated, scaling a semantic model to more data produces much better fits to human semantic data. Hence, both order information and magnitude of linguistic input have been demonstrated to be important factors in human semantic learning. If RPs have similar characteristics to convolution, they may afford encoding very large-scale order information, and much better approximations to human semantic structure.

This work is further motivated by the cognitive implications of circular convolution and RPs. Vector representations constructed by means of circular convolution have been frequently described as psychologically or neurally plausible (Levy, 2007; Jones & Mewhort, 2007), due to several features that they share with connectionist networks: distributed encoding, robustness to noise, affordance of generalization, error correction, pattern completion, and easy associative access (Plate, 2003). Furthermore, implementing neural networks that instantiate convolution-like operations is straightforward (Plate, 2000; but compare Pike, 1986). Similarly, RPs possess many properties relevant to human cognition. Not only have they been proposed as a particularly versatile multiplication operator for constructing vector representations that are highly distributed, tolerant of noise in the input, robust to error and component failure, and mathematically compatible with several known properties of neural circuitry (Kanerva, 2009), RPs are trivially easy to implement in connectionist terms; a RP can simply be thought of as a two-layer network connected by randomly placed one-to-one copy connections. Thus, comparing circular convolution and RPs affords us a better understanding of two psychologically plausible operations for encoding semantic information that have never been systematically compared.

We conducted three experiments intended to compare convolution and RPs as means of encoding word-order information with respect to performance and scalability. In Experiment 1, we conducted an empirical comparison of the storage capacity and the probability of correct decoding under each method. In Experiment 2, we compared RPs with convolution in the context of a simple vector accumulation model equivalent to BEAGLE's "order space" (Jones and Mewhort, 2007) on a small battery of semantic evaluation tasks when trained on a Wikipedia corpus. The model was trained on both the full corpus and a random subset; results improved markedly when RPs are allowed to scale up to the full Wikipedia corpus, which proved to be intractable for the convolution-based model. Finally, in Experiment 3, we specifically compared BEAGLE to RPM, which differs from BEAGLE in several important ways other than its binding operation, to assess whether using RPs in the context of RPM improves their performance further. We conclude that random permutations are a promising and scalable alternative to circular convolution.

Experiment 1

Plate (2003) made a compelling case for circular convolution in the context of holographic reduced representation, demonstrating its utility in constructing distributed representations with high storage capacity and high probability of correct retrieval. However, the storage capacity and probability of correct retrieval with RPs has not been closely investigated. This experiment compared the probability of correct retrieval of RPs with circular convolution under varying dimensionality and number of vectors stored.

Method

As a test of the capacity of convolution-based associative memory traces, Plate (2003, Appendix D) describes a simple paired-associative retrieval task in which the algorithm must select, from set E of m possible random vectors, the vector x_i that is bound to its associate y_i . The retrieval algorithm is provided with a trace vector of the form $t = (x_1 * y_1) + (x_2 * y_2) + (x_3 * y_3) + \ldots$ that stores a total of k vectors. All vectors are of dimensionality D, and each x_i and y_i is a vector with elements independently drawn from $N(0, 1/D)$. The retrieval algorithm is provided with the trace t and the probe y_i , and works by first calculating $a = (y_i \# t)$, where # is the *correlation operator* described in detail in Plate (2003, pp. 94-97). It then retrieves the vector in the "clean-up memory" set E that is the most similar to a. This is accomplished by calculating the cosine between a and each vector in the set E, and retrieving the vector from E for which the cosine is highest. If this vector is not equal to x_i , this counts as a retrieval error. We replicated Plate's method to empirically derive retrieval accuracies for a variety of choices of k and D, keeping m fixed at 1,000.

Sahlgren et al. (2008) essentially bind signal vectors to positions by means of successive self-composition of a permutation function Π, and construct trace vectors by superposing the results. Because the signal vectors are random, any permutation function that maps each element of the input onto a different element of the output will do; we adopt Sahlgren et al.'s suggestion of using rotation of a vector by one position for Π for the sake of simplicity. We also use their notation of $\Pi^n x$ to mean "Π composed with itself n times;" thus, $\Pi^2 x = \Pi(\Pi x)$, $\Pi^3(x) = \Pi(\Pi^2 x)$), and so forth. The notion of a trace vector of paired associations can then be recast in RP terms as follows:

$$t = (\Pi y_1 + \Pi^2 x_1) + (\Pi^3 y_2 + \Pi^4 x_2) + (\Pi^5 y_3 + \Pi^6 x_3) + \ldots$$

where the task again is to retrieve some y_i's associate x_i when presented only with y_i and t. A retrieval algorithm for accomplishing this can be described as follows: Given a probe vector y_i , the algorithm applies the inverse of the initial permutation to trace vector t, yielding $\Pi^{-1}t$. Next, the cosine between $\Pi^{-1}t$ and the probe vector y_i is calculated, yielding a value that represents the similarity between y_i and $\Pi^{-1}t$. These steps are then iterated: the algorithm calculates the cosine between y_i and $\Pi^{-2}t$, between y_i and $\Pi^{-3}t$, etc., until this similarity value exceeds some high threshold; this indicates that the algorithm has probably "found" y_i in the trace. At that point, t is permuted one more time, yielding x', a noisy approximation of y_i's associate x_i. This approximation x' can then be compared with clean-up memory to retrieve the original associate x_i.

Alternatively, rather than selecting a threshold, t may be permuted some finite number of times[1] and its cosine similarity to y_i calculated for each permutation. Let n indicate the inverse permutation for which $\cos(\Pi^{-n}t, y_i)$ is the highest. We can permute one more time to get $\Pi^{-n-1}t$, that is, our noisy approximation x'. This method is appropriate if we always want our algorithm to return an answer (rather than, say, timing out before the threshold is exceeded), and is the method we used for this experiment.

The final clean-up memory step is identical to that used by Plate (2003): we calculate the cosine between x' and each vector in the clean-up memory E, and retrieve the vector in E for which this cosine is highest. As when evaluating convolution, we keep m (the number of vectors in E) fixed at 1,000 while varying the number of stored vectors k and the dimensionality D.

Results

Figure 1 reports retrieval accuracies for convolution-based associative memory traces, while Figure 2 reports retrieval accuracies for the RP formulation of the task. 500 vector pairs were sampled randomly from a pool of 1,000 possible random vectors with replacement and the proportion of correct retrievals was computed. All 1,000 vectors in the pool were potential candidates; thus, an accuracy of 0.1% would represent chance performance. The horizontal axes of all figures indicate the total number of pairs stored in the trace (i.e., half the total number of vectors in the trace).

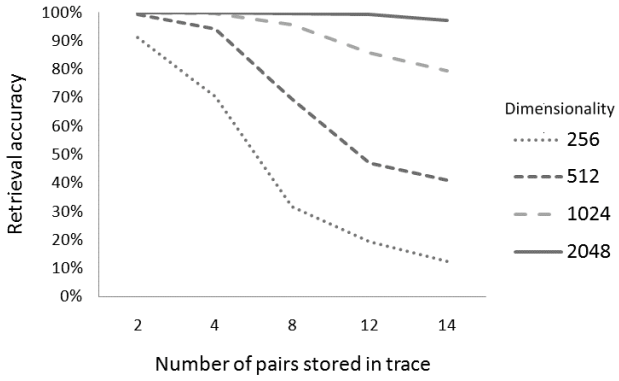


Figure 1. Retrieval accuracies for convolution-based associative traces.

[1] In Plate's (2003, p. 252) demonstration of the capacity of convolution-based associative memories, the maximal number of pairs stored in a single trace was 14; we likewise restrict the maximal number of pairs in a single trace to 14 (28 items total).

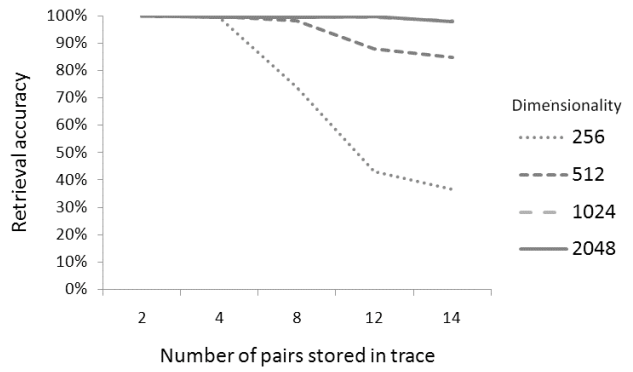


Figure 2. Retrieval accuracies for RP-based associative traces.

Discussion

Circular convolution has an impressive storage capacity and excellent probability of correct retrieval at high dimensionalities; the results were comparable to those reported by Plate (2003, p. 252) in his test of convolution-based associative memories. However, RPs seem to share these desirable properties as well. In fact, the storage capacity of RPs seems to drop off significantly more slowly than does the storage capacity of convolution as dimensionality is reduced.

This information capacity is particularly interesting given that, with respect to basic encoding and decoding operations, RP is computationally more efficient than convolution. Encoding *n*-dimensional bindings with circular convolution using equation (1) is a very slow $O(n^2)$ operation. This can be sped to $O(n)$ by means of the Fast Fourier transform (Jones, 2007; Plate, 2003). The algorithm to bind two vectors *a* and *b* in $O(n)$ time involves calculating discrete Fourier transforms of *a* and *b*, multiplying them pointwise to yield a new vector *c*, and calculating the inverse discrete Fourier transform of *c*. Encoding with RPs can also be accomplished in $O(n)$ time, but with steps that are not as computationally expensive. To bind two vectors *a* and *b,* the elements of *a* are permuted by directly copying then into a new vector, but with the mapping of their indices determined by the permutation function. For example, if the permutation function were chosen to be *rotation by one position* and vectors were of dimensionality *D*, each value at index *i* in the vector *a* would be copied to index $(i + 1) \bmod D$ in the new vector. The vector *b* is permuted in the same way, but using a different permutation function (e.g., $(i + 2) \bmod D$). Finally, *a* and *b* are added to yield a final binding *c*.

Noisy decoding—the retrieval of a noisy version of one or more bound associates from a trace (which may then be passed to clean-up memory to unambiguously determine the identity of the associate)—also operates in $O(n)$ time in both representations. As with encoding, the operation is $O(n)$, but fewer operations are required (a single permutation decodes one associate, rather than an involution + two discrete Fourier transforms + an elementwise multiplication + one inverse discrete Fourier transform).

Experiment 2

In order to move from the paired-associates problem of Experiment 1 to a real language task, we evaluated how a simple vector accumulation model akin to Jones & Mewhort's (2007) encoding of order-only information in BEAGLE would perform on a set of semantic tasks if RPs were used in place of circular convolution. In Experiment 2, we replaced the circular convolution component of BEAGLE with RPs so that we could quantify the impact that the choice of operation alone had on the results. Due to the computational efficiency of RPs, we were able to scale them to a larger version of the same textbase, and simultaneously explore the effect of scalability on order.

Method

Order information was trained using both the BEAGLE model and a modified implementation of BEAGLE in which the circular convolution operation was replaced with RPs as they are described in Sahlgren at al. (2008). A brief example will illustrate how this replacement changes the algorithm. Recall that in BEAGLE, each word *w* is assigned a static "environmental" signal vector e_w as well as a dynamic memory vector m_w that is updated during training. Recall also that the memory vector of a word *w* is updated by adding the sum of the convolutions of all n-grams (up to some maximum length λ) containing *w*. Upon encountering the phrase "one two three" in a corpus, the memory vector for "one" would normally be updated as follows:

$$m_{\text{one}} = m_{\text{one}} + (\Phi * e_{\text{two}}) + (\Phi * e_{\text{two}} * e_{\text{three}})$$

where Φ is a placeholder signal vector that represents the word whose representation is being updated. In the modified BEAGLE implementation used in this experiment, the memory vector for "one" would instead be updated as:

$$m_{\text{one}} = m_{\text{one}} + \Pi e_{\text{two}} + \Pi^2 e_{\text{three}}$$

The modified BEAGLE implementation was trained on a 2.33 GB corpus (418 million tokens) of documents from Wikipedia. Training on a corpus this large proved intractable for the slower convolution-based approach. Hence, we also trained both models on a 35 MB, six-million-token subset of this corpus constructed by sampling random 10-sentence documents from the larger corpus without replacement. Accuracy was evaluated on two synonymy tests: the English as a Second Language (ESL) and the Test of English as a Foreign Language (TOEFL) synonymy assessments. Rank correlations to human judgments of the semantic similarity of word pairs were calculated using the similarity judgments obtained from Rubenstein and Goodenough (G, 1965), Miller and Charles (M&C, 1991), Resnik (R, 1995), and Finkelstein et al. (F&al, 2002). A description of these measures can be found in Recchia and Jones (2009).

Results and Discussion

Table 3 provides a comparison of two variants of the BEAGLE model, each trained on order information only.

"Convolution" refers to the original BEAGLE as described in Jones & Mewhort, while "Random Permutations" refers to a version in which order information is encoded using RPs rather than circular convolution. Three points about these results merit special attention. First, there are no significant differences between the performance of convolution and RPs on the small corpus. Both performed nigh-identically on F and TOEFL; neither showed any significant correlations with human data on R&G, M&C, R, nor performed better than chance on ESL.

Table 3. Comparisons of variants of BEAGLE that differ by binding operation. Accuracy scores are reported for ESL & TOEFL; remaining tasks are Spearman rank correlations.

Criterion	Wikipedia subset		Full Wikipedia
	Convolution	Random Permutations	Random Permutations
ESL	.20	.26	.32
TOEFL	.46†	.46†	.63†
R&G	.07	-.06	.32*
M&C	.08	-.01	.33*
R	.06	-.04	.35*
F&al	.13*	.12*	.33*

*Significant correlation, $p < .05$, one-tailed.
†Accuracy score differs significantly from chance, $p < .05$, one-tailed.

Second, both models performed the best by far on the TOEFL synonymy test, supporting Sahlgren's et al. (2008) claim that order information may indeed be more useful for synonymy tests than tests of semantic relatedness, as paradigmatic rather than syntagmatic information sources are most useful for the former. However, it is unclear exactly why neither model did particularly well on ESL[2], as many models have achieved scores on it comparable to their scores on TOEFL (Recchia & Jones, 2009). Finally, only RPs were able to scale up to the full Wikipedia corpus, and doing so yielded extreme benefits for every task. This is a very strong point in favor of RPs, and suggests that sequential information can even be useful for tasks that involve semantic relatedness but not synonymy per se (R&G, M&C, R, F), provided that the model is trained at a sufficiently large scale.

Experiment 3

In Experiment 2 we saw that importing RPs into BEAGLE yielded comparable results on a small corpus and considerable improvement in scalability. Here we compare BEAGLE to the original model of Sahlgren et al., which we have been referring to as RPM. In many ways the two are similar: Like BEAGLE, RPM can construct a semantic space by (1) adding only order vectors to memory vectors during training, yielding an "order space," and (2) by adding order vectors *as well as* "context vectors," yielding a "composite space." Besides using RPs in place of circular convolution, the specific implementation of RPM reported by Sahlgren et al. differs from BEAGLE in several ways (signal-vector representation, window size, lexicon size, and the stoplist). This experiment aims to assess RPM's performance with another corpus and on other semantic tasks besides TOEFL, and to determine if performance improves under RPM parameter settings (compared to the BEAGLE settings in Experiment 2).

Method

The same evaluation method was applied as in Experiment 2, but with BEAGLE being compared directly to RPM. Both models were trained in order and composite space.

Results and Discussion

Table 4 reports the results of BEAGLE and RPM trained in order space, while Table 5 reports results in composite space (context + order information). As in Experiment 2, RPs but not convolution proved capable of scaling up to the full Wikipedia corpus. We replicated Sahlgren et al.'s (2008) performance on TOEFL in order space at this dimensionality, but this Wikipedia implementation of RPM fell short of the ~.73 accuracy they reported on TOEFL at a dimensionality of 2000 in composite space; the difference is most likely due to the different corpora used in the two evaluations. On the small corpus, switching from order space to composite space did not yield significant differences for either model when contrasted with the use of order space alone. On the large corpus, however, when contrasted with RPs in Experiment 2 (Table 3), RPM performed far better on several evaluations, most notably the correlations to the R&G, M&C, and R similarity judgments. It is intriguing that the version of RPM trained on the full Wikipedia in order space was able to perform well on several tasks that are typically conceived of as tests of associative relatedness and not tests of synonymy per se—for example, .70 on the Miller & Charles pairs (Table 4).

Table 4. BEAGLE and RPM in order space.

Criterion	Wikipedia subset		Full Wikipedia
	BEAGLE	RPM	RPM
ESL	.20	.27	.38†
TOEFL	.46†	.37†	.65†
R&G	.07	.15	.50*
M&C	.08	.16	.70*
R	.06	.11	.63*
F&al	.13*	.18*	.32*

* Significant correlation, $p < .05$, one-tailed.
†Accuracy score differs significantly from chance, $p < .05$, one-tailed.

[2] Note that the absolute performance of these models is irrelevant to the important comparisons. Many factors (e.g., frequency thresholding, morphological normalization, corpus siz/type) are known to improve performance on synonymy tests; we held these constant, which produced poor absolute performance (but see Sahlgren et al., 2008). The key comparisons are the consistency of the operations on the same textbase, and the relative performance boost when data are scaled up.

Table 5. BEAGLE and RPM in composite space.

Criterion	Wikipedia subset		Full Wikipedia
	BEAGLE	RPM	RPM
ESL	.24	.27	.42†
TOEFL	.47†	.40†	.66†
R&G	.10	.10	.49*
M&C	.09	.12	.70*
R	.09	.03	.60*
F&al	.23*	.19*	.32*

* Significant correlation, $p < .05$, one-tailed.
†Accuracy score differs significantly from chance, $p < .05$, one-tailed.

General Discussion

Experiment 1 demonstrates that RPs are capable of high retrieval accuracy even when many paired associates are stored in a single trace, and their storage capacity appears to be slightly better than that of circular convolution for low dimensionalities. Experiments 2 and 3 reveal that both methods achieve approximately equal performance on a battery of semantic tasks when trained on a small corpus, but that RPs are ultimately capable of achieving superior performance due to their higher scalability. In all, these results suggest that RPs are worthy of further study both as encoders of sequential information in word space models and as operators capable of storing associative information more generally. It should be noted that Sahlgren et al. (2008) found better synonymy performance when RPs were trained on "direction" vectors rather than order vectors; direction vectors simply encode whether words appear before or after a word in the temporal stream, but ignore the order chain. Given the computational efficiency of this approach, future work should explore the effects of scaling to large-scale data on RP direction vectors.

Both convolutions and RPs are naturally derived from properties of the human cognitive system, namely groups of neurons connected with a certain degree of randomness (see Plate, 2003 for convolution and Kanerva, 2009 for RPs; also see Howard et al. [in press] for a related model using neural properties of temporal context encoding). The current work demonstrates that when a model is able to apply these associative learning mechanisms across a large amount of episodic experience with linguistic structure, it produces much better approximations of human semantic structure. As Elman (2009) has argued, the encoding of large-scale order information is a core component of a word's lexical representation that is often overlooked. Future work needs to explore application of large-scale RP encoding to more complex semantic and linguistic tasks.

Acknowledgements

This research was supported in part by a grant from Google Inc. to MNJ and a Shared University Research Grant from IBM to Indiana University.

References

Banko, M. & Brill, E. (2001). Scaling to very very large corpora for natural language disambiguation. In *Proceedings of the 39th Annual Conference of the Association for Computational Linguistics* (pp. 26-33). Toulouse, France: Association for Computational Linguistics.

Burgess, C., & Lund, K. (2000). The dynamics of meaning in memory. In E. Dietrich & A. Markman (Eds.), *Cognitive dynamics: Conceptual and representational change in humans and machines* (pp. 117-156).

Elman, J. L. (2009). On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. *Cognitive Science, 33,* 547-582.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Z., S., Wolfman, G., & Ruppin, E. (2002). Placing search in context: The concept revisited. *Association for Computing Machinery Transactions on Information Systems, 20*(1), 116-131.

Hare, M., Jones, M. N., Thomson, C., Kelly, S., & McRae, K. (2009). Activating event knowledge. *Cognition, 111*(2), 151-167.

Howard, M. W., Shankar, K. H., and Jagadisan, U. K. K. (In press). Constructing semantic representations from a gradually-changing representation of temporal context. *Topics in Cognitive Science.*

Jones, M. N. (2007). Holographic neural networks. Poster presented at the 48th Meeting of the Psychonomic Society. Long Beach, CA.

Jones, M. N., Kintsch, W., & Mewhort, D. J. K. (2006). High-dimensional semantic space accounts of priming. *Journal of Memory and Language, 55,* 534-552.

Jones, M. N. & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, 114, 1-37.

Miller, G. A. and Charles, W. G. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes,* 6(1), 1-28.

McKoon, G. & Ratcliff, R. (2003). Meaning through syntax: Language comprehension and the reduced relative clause construction. *Psychological Review, 110*, 490-525.

Kanerva, P. (2009). Hyperdimensional computing: An introduction to computing in distributed representations with high-dimensional random vectors. *Cognitive Computation, 1,* 139-159.

Kanerva, P., Kristoferson, J., & Holst, A. (2000). Random indexing of text samples for latent semantic analysis. In *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, (p. 1036). Hillsdale, NJ: Erlbaum.

Landauer, T. K. & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review, 104,* 211-240.

Landauer, T., McNamara, D. S., Dennis, S., & Kintsch, W. (Eds.) (2007). *Handbook of latent semantic analysis.* Mahwah, NJ: Erlbaum.

Levy, S. D. (2007). Changing semantic role representations with holographic memory. (Report No. FS-07-04). In *Computational approaches to representation change during learning and development: Papers from the 2007 AAAI Symposium*. Menlo Park, CA: AAAI Press.

Pike, R. (1986). Comparison of convolution and matrix distributed memory systems for associative recall and recognition. *Psychological Review, 91*, 281-293.

Plate, T. (2003). *Holographic reduced representation: Distributed representation for cognitive structures*. CSLI Publications.

Recchia, G. L., & Jones, M. N. (2009). More data trumps smarter algorithms: Comparing pointwise mutual information to latent semantic analysis. *Behavior Research Methods*, 41(3), 647-56.

Resnik, P. (1995). Using information content to evaluate semantic similarity. In C. S. Mellish (Ed.), *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (IJCAI)* (pp. 448-453). Montréal, Canada: Morgan Kaufmann.

Risley, T. R. & Hart, B. (2006). Promoting early language development. In N. F. Watt, C. Ayoub, R. H. Bradley, J. E. Puma & W. A. LeBoeuf (Eds.), *The crisis in youth mental health: Critical issues and effective programs, Volume 4, Early intervention programs and policies*, 83-88. Westport, CT: Praeger.

Riordan, B., & Jones, M. N. (2007). Comparing semantic space models using child-directed speech. In D. S. MacNamara & J. G. Trafton (Eds.), *Proceedings of the 29th Annual Conference of the Cognitive Science Society,* 599-604.

Rubenstein, H., & Goodenough, J. (1965). Contextual correlates of synonymy. *Communications of the Association for Computing Machinery, 8*(10), 627-633.

Sahlgren, M., Holst, A., & Kanerva, P. (2008). Permutations as a means to encode order in word space. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society,* p. 1300-1305.

Learning and Generalization of Abstract Semantic Relations: Preliminary Investigation of Bayesian Approaches

Dawn Chen (sdchen@ucla.edu)
Department of Psychology

Hongjing Lu (hongjing@ucla.edu)
Departments of Psychology and Statistics

Keith J. Holyoak (holyoak@lifesci.ucla.edu)
Department of Psychology
University of California, Los Angeles
Los Angeles, CA 90095 USA

Abstract

A deep problem in cognitive science is to explain the acquisition of abstract semantic relations, such as antonymy and synonymy. Are such relations necessarily part of an innate representational endowment provided to humans? Or, is it possible for a learning system to acquire abstract relations from non-relational inputs of realistic complexity (avoiding hand-coding)? We present a series of computational experiments using Bayesian methods in an effort to learn and generalize abstract semantic relations, using as inputs pairs of specific concepts represented by feature vectors created by Latent Semantic Analysis.

Keywords: Bayesian inference; induction; generalization; abstract relations; machine learning; LSA

Introduction

An intelligent human adult can recognize that the concepts *day* and *night* are related in much the same way as *hot* and *cold*, but not in the same way as *day* and *hour*. This ability to appreciate abstract semantic relations is fundamental to analogical reasoning, and is arguably a core component of what is special about the human mind (Penn, Holyoak & Povinelli, 2007). But how are such abstract relations acquired? If they are learned, how this could be achieved is far from obvious. On the face of it, no perceptual or other features seem to be available to represent such abstract relations as antonymy, synonymy, or superordination. Almost by default, it might be assumed that abstract relations must be innate (Fodor, 1975).

Research on cognitive development has clearly established the phenomenon of a *relational shift* (Gentner & Rattermann, 1991), such that children process relations more effectively with increasing age. In particular, children move from a focus on global similarities of objects to similarities defined by specific dimensions, such as size or color (Smith, 1989; Smith & Sera, 1992). Less is known about the development of abstract relations that seem yet further divorced from perceptual similarity (see Halford, 1993). Analyses of corpora of child speech have identified systematic use of antonyms by children aged 2-5 years (Jones & Murphy, 2005). Children aged 6-7 years are more accurate in detecting the falsity of sentences such as *Some valleys are mountains* as compared to *Some valleys are lakes*, where the former sentence type contains an antonymous pair (Glass, Holyoak & Kossan, 1977), suggesting that some sense of antonymy is available prior to any formal instruction about this concept.

The Problem of Relation Learning

Regardless of whether abstract relations are learned or mature over the course of development, there is no doubt that adults can distinguish among instances of relations such as antonymy versus synonymy. In the present paper we pose the following computational problem: Given as inputs a modest number of pairs of concepts that instantiate an abstract relation (e.g., *day-night* and *hot-cold*, which instantiate antonymy), is it possible to extract a representation of the abstract relation that may then be used to accurately classify novel instantiations (e.g., *valley-mountain*)?

Most recent connectionist models of relation learning (e.g., Rogers & McClelland, 2008) have focused on the acquisition of small numbers of specific input-output pairs (e.g., "canary" + "can" → "fly"), but have not demonstrated the capacity to generalize to novel inputs dissimilar to the training items. In contrast, achieving such generalization is the central aim of our project. Moreover, an important constraint we imposed is that inputs to the learning system could not be hand-coded, as has been commonplace in the literature on computational models of analogy and relation learning. For example, Doumas, Hummel, and Sandhofer (2008) showed how structured relations corresponding to relative adjectives such as *bigger-than* can be extracted by bottom-up mechanisms given inputs consisting of unstructured feature vectors of objects. However, the modelers ensured that "size" features were present among the relatively small feature set defining the inputs, setting the stage for selecting these size features to form a part of the to-be-learned relational predicate. While perceptual relations may indeed be derived from the perceptual features of objects, this assumption is unwarranted for more abstract relations, for which hand-coding of features is even more problematic. In addition, realistic semantic representations would seem to require very large numbers of features, raising all the difficulties associated with search in a large

representational space. Learning models that are developed for small, hand-tailored inputs at best postpone the challenges of "scaling up".Another approach to learning relations is to combine statistical techniques with structured representations. For example, Kemp and Tenenbaum (2008) showed how Bayesian techniques can operate on relational structures to learn relational systems such as hierarchies and linear orderings. The relational structures are provided to the system by including a grammar that generates possible structures. Although this approach may be appropriate for relations that have a well-defined logical structure known to the modeler, it is not clear that it can readily be extended to the full range of "messy" semantic relations. In addition, since the postulated grammar of relations is not itself learned, rather strong nativist assumptions remain.

Learning Relations from Unstructured Inputs

In this project, we have taken the tack of attempting to model the learning of abstract relations through essentially data-driven statistical learning, using Bayesian algorithms applied to large, unstructured input representations that we the modelers did not create. The raw inputs are vector representations of words, derived by Latent Semantic Analysis (LSA; Landauer & Dumais, 1997). Such vectors, the product of singular value decomposition applied to lexical co-occurrence data from a large corpus of text, have proved extremely useful in many applications, often serving as good measures of semantic similarity of concepts (Wolf & Goldman, 2003). However, LSA vectors do not provide any direct basis for identifying abstract relations between concepts (although some modest success has been achieved by exploiting LSA vectors for relation words, such as *opposite*; Mangalath, Quesada & Kintsch, 2004). Related machine-learning algorithms have had some success in solving relational analogies by working directly from co-occurrence data for word combinations found in a large corpus of text (Turney & Littman, 2005). However, our goal is different in that we aim to model learning of relational representations from the LSA vectors for a small (< 20) set of word pairs that instantiate each abstract relation. The task of learning relations from representations of simpler concepts bears at least some resemblance to the task a child might face in acquiring an abstract relation from a modest-sized set of examples that instantiate it.

For our present purpose, we do not assume that LSA provides anything like an optimal psychological representation of concepts (indeed, it has well-known and serious limitations, notably problems dealing with lexical ambiguity). However, by using LSA inputs we ensure that we have in no way tailored the inputs so as to "hand hold" the learning algorithms we test. Moreover, we do not assume that it is in fact possible to acquire human-like representations of abstract relations solely by data-driven learning. Rather, by pressing the limits of data-driven approaches, we may be able to identify more clearly what nativist assumptions may ultimately prove essential.

A General Framework for Relation Learning

Here we report a preliminary investigation of relation learning based on two variants of the same basic framework. Our goal is to learn an explicit representation of a relation from a training set, **S**, consisting of pairs of concepts that each instantiate the relation. We assume that a decision regarding whether a pair of concepts instantiates a particular relation *R* is determined by a representation that includes both the basic features of the input concepts and additional features that the model automatically derives from the basic features. The full input representation is comprised of the basic features of two concepts, **A** and **B**, which are represented by LSA vectors, and of derived features $\boldsymbol{\Phi}$(**A**, **B**) computed from **A** and **B** (see Fig. 1). In this study the derived features included two types, product features **AB** = $[A_1B_1 \quad A_2B_2 \quad \cdots \quad A_dB_d]$ and absolute difference features $|\mathbf{A}-\mathbf{B}| = [|A_1 - B_1| \quad |A_2 - B_2| \quad \cdots \quad |A_d - B_d|]$, both defined across corresponding positions in the **A** and **B** vectors. The length of each type of derived vector is thus equal to the length of each basic vector, so that the total size of the input vector scales linearly with the number of basic features.

If we let **X** denote the full vector including basic and derived features, **X** = [**A**, **B**, $\boldsymbol{\Phi}$(**A**, **B**)], then the computational goal of relation learning is to estimate the distribution of a corresponding weight vector **w** from a set of training pairs that share the same relation. That is, we calculate $P(\mathbf{w} \mid \mathbf{X_S}, \mathbf{R_S} = 1)$, where the subscript **S** indicates the set of training examples (the source) and $\mathbf{R_S}$ is a set of binary indicators, each of which (denoted by *R*) indicates whether a particular pair of concepts instantiates the relation or not. The vector **w** constitutes the learned relational representation, which can be interpreted as attention weights reflecting the importance of the corresponding features in **X**. To test generalization of the learned relational representation, we test on new transfer pairs, denoted by the subscript *T*. The inference step needs to estimate the probability that a target pair shares the same relation as the training pairs, $P(R_T = 1 \mid \mathbf{X}_T, \mathbf{X_S}, \mathbf{R_S} = 1)$.

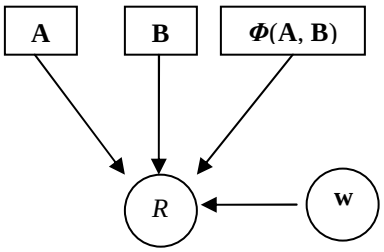


Figure 1: Graphical representation of the general framework. **A** and **B** denote two vectors of concept features (LSA inputs); $\boldsymbol{\Phi}$(**A**, **B**) denotes derived features based on the two concepts, i.e., product features **AB** and absolute difference features **|A – B|**. Vector **w** represents the unknown relational weights that define *R*, and is learned using the training set of examples instantiating *R*.

The models we consider are both based on Bayesian logistic regression, as described by Silva, Airoldi and Heller (2007) and Silva, Heller and Gharamani (2007). Given a small set of word-pairs **S** that all instantiate a given abstract relation *R*, both models compute the posterior probability that $(\mathbf{A}_T, \mathbf{B}_T)$ is an example of the same relation,

$$P(R_T = 1 \mid \mathbf{X}_T, \mathbf{X}_\mathbf{S}, \mathbf{R}_\mathbf{S} = 1) = \int_\mathbf{w} P(R_T = 1 \mid \mathbf{X}_T, \mathbf{w}) P(\mathbf{w} \mid \mathbf{X}_\mathbf{S}, \mathbf{R}_\mathbf{S} = 1) \quad (1)$$

where the likelihood is assessed using a logistic regression function to predict the probability of a word-pair instantiating a given relation,

$$P(R = 1 \mid \mathbf{X}, \mathbf{w}) = \text{logistic}(\mathbf{w}^\mathrm{T}\mathbf{X}) \quad (2)$$

where $\text{logistic}(x) = (1 + e^{-x})^{-1}$.

For the first model we consider (based directly on Silva et al., 2007), the posterior distribution for **w** is found by applying Bayes' rule using the prior distribution for **w** and the training word-pairs:

$$P(\mathbf{w} \mid \mathbf{X}_\mathbf{S}, \mathbf{R}_\mathbf{S} = 1) = \frac{P(\mathbf{R}_\mathbf{S} = 1 \mid \mathbf{w}, \mathbf{X}_\mathbf{S}) P(\mathbf{w})}{\int_\mathbf{w} P(\mathbf{R}_\mathbf{S} = 1 \mid \mathbf{w}, \mathbf{X}_\mathbf{S}) P(\mathbf{w})} \quad (3)$$

Because of the high dimensionality of the learning problem we are tackling, the choice of a good prior $P(\mathbf{w})$ is essential to the performance of any model. We investigated two kinds of priors, a simple empirical prior proposed by Silva and colleagues, and our own hierarchical model.

The Empirical Prior

Intuitively, our simple empirical prior distinguishes word-pairs that instantiate *any* of the to-be-learned relations from unrelated word-pairs. The empirical prior takes the form $P(\mathbf{w}) = N(\mathbf{w};\ \hat{\mathbf{w}}, \hat{\mathbf{\Sigma}})$, in which the sample mean estimate $\hat{\mathbf{w}}$ is by found by fitting a logistic regression classifier using maximum-likelihood estimation on a relatively small set of related word pairs (positive examples), and a larger set of unrelated word pairs (negative examples), reflecting the fact that most pairs of actual concepts do not instantiate any abstract relation. The covariance matrix $\hat{\mathbf{\Sigma}}$ for this empirical prior is calculated by

$$\left(\hat{\mathbf{\Sigma}}^{-1}\right) = c \cdot \left(\mathbf{X}^T \mathbf{M} \mathbf{X}\right) / N \quad (4)$$

where *c* is a user-defined smoothing parameter set to twice the number of related pairs in the training samples, *N* is the total number of word pairs in the training set, and **X** is a matrix containing the features of all (related and unrelated) word pairs in the training set. **M** is a diagonal matrix with each entry defined as

$$(\mathbf{M})_{ii} = \hat{p}(i)(1 - \hat{p}(i)) \quad (5)$$

where $\hat{p}(i)$ is the MLE predicted probability of the *i*th word pair being related, given by Eq. (2).

The Hierarchical Prior

The above model computes its prior based on the observed data. This empirical prior uses all related pairs as members of the set of positive training cases, and numerous unrelated pairs as negative cases. An alternative empirical prior could be computed by considering pairs of a specific relation as positive examples and pairs instantiating other relations as negative examples. Although empirical priors are a sensible choice to facilitate inference in the high-dimensional space, the question of how the best data set for learning an empirical prior could be constructed remains unresolved.

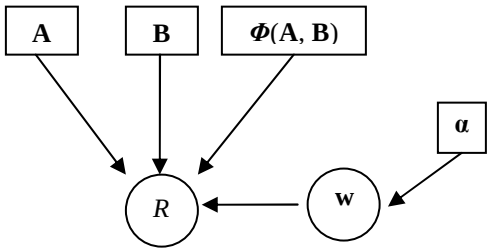


Figure 2: Graphical representation of hierarchical model. Distribution of **α** is determined by the hyperparameters that model the variance of the relational weight vector **w**. The other notations are the same as in Figure 1.

Here we explored a different approach, specifying a hierarchical prior on the distribution of the weight vector **w** (see Fig. 2). Specifically, the posterior distribution of **w** learned from training data is derived (replacing Eq. 3) by

$$P(\mathbf{w} \mid \mathbf{X}_\mathbf{S}, \mathbf{R}_\mathbf{S} = 1) = \int_\boldsymbol{\alpha} P(\mathbf{w} \mid \boldsymbol{\alpha}, \mathbf{X}_\mathbf{S}, \mathbf{R}_\mathbf{S} = 1) P(\boldsymbol{\alpha}) \quad (6)$$

where vector $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \ldots]$ determines the precision (the inverse variance) of each element of the weight vector $\mathbf{w}$. We use a conjugate prior distribution in the form of a Gamma distribution for α_i with two hyperparameters a_0 and b_0:

$$P(\alpha_i) \sim Gamma(\alpha;\ a_0, b_0) \quad (7)$$

The individual prior for each element in vector **w** is assigned in the form of a normal distribution:

$$P(w_i \mid \alpha_i) \sim N(w_i;\ 0,\ \alpha_i) \quad (8)$$

This normal distribution imposes a general prior that the value of w_i is centered at 0 (i.e., the *i*th feature dimension is not expected to be relevant in predicting whether a certain relation exists between the two words). However, the value of α_i controls the certainty about this prior belief. A low precision value makes the prior belief uninformative, whereas a high precision value imposes a strong bias that w_i is most likely 0. Accordingly, the hyperparameters play an important role in determining the relevance of feature dimensions in predicting the existence of a relation.

The other term in Eq. (6) can be derived by applying Bayes rule directly,

$$P(\mathbf{w} \mid \boldsymbol{\alpha}, \mathbf{X}_\mathbf{S}, \mathbf{R}_\mathbf{S} = 1) = \frac{P(\mathbf{R}_\mathbf{S} = 1 \mid \mathbf{w}, \mathbf{X}_\mathbf{S}) P(\mathbf{w} \mid \boldsymbol{\alpha})}{\int_\mathbf{w} P(\mathbf{R}_\mathbf{S} = 1 \mid \mathbf{w}, \mathbf{X}_\mathbf{S}) P(\mathbf{w} \mid \boldsymbol{\alpha})} \quad (9)$$

The Inference Algorithm

Although the general framework of the relation learning models is straightforward, the inference step is non-trivial because the calculation of the normalization terms in Eqs. (3) and (9) and integrals in Eq. (6) are intractable, lacking analytic solutions. A sampling approach is impractical for dealing with high feature dimensionality. We therefore employed variational methods developed by Jaakkola and Jordan (2000) to obtain a closed-form approximation to the posterior distribution. Specifically, the variational method updates the mean of vector **w** and its covariance matrix **V** iteratively:

$$\begin{aligned} \mathbf{V}^{-1} &= \mathbf{a}/\mathbf{b} + 2\sum \lambda(\xi_{\mathbf{n}})\mathbf{x}_{\mathbf{n}}\mathbf{x}_{\mathbf{n}}^{T}, \\ \mathbf{w} &= \mathbf{V}\sum_{n} \mathbf{x}_{\mathbf{n}} / 2, \\ \mathbf{a} &= a_0 + 1/2, \\ \mathbf{b} &= b_0 + \mathrm{E}_w(\mathbf{w}\mathbf{w}^T)/2, \\ \xi^2 &= \mathbf{x}^T(\mathbf{V} + \mathbf{w}\mathbf{w}^T)\mathbf{x}. \end{aligned} \tag{10}$$

Computational Experiments

The Training Set and Generalization Test

Table 1 shows some examples of pairs of concepts that we used to train and test the two models. We used four different relations: function, synonyms, linear ordering, and antonyms. For each relation, we chose 15-20 pairs that were examples of that relation to use as the training set. We will refer to pairs used for training as AB pairs. All pairs were selected from experimental materials used previously to form four-term verbal analogy problems, and for which LSA vectors (derived using the tasaALL corpus) were available. We selected pairs for which the cosine similarity between the words (based on their LSA vectors) was at least 0.1, aiming to exclude pairs that included highly ambiguous words (e.g., *gift-present* as an example of synonyms).

After learning representations of the abstract relations based on the AB pairs, the model was tested on a two-alternative forced-choice generalization task. For each test item, the model was asked to choose which of two alternative pairs instantiated a specified relation. We will refer to correct and incorrect options as CD and CD', respectively. For example, one item required the models to decide which pair instantiated antonymy, *shallow-deep* (CD) or *shallow-depth* (CD'). As this example suggests, the discrimination was quite subtle, as the C term was common to both options and the CD' pair also instantiated an abstract relation (but not the relation being queried). The words used in this generalization test did not overlap at all with the AB pairs used in training, but were selected according to the same general criteria. For each test problem, the models calculated the probability of CD and of CD' being examples of the relation, respectively, according to Eq. (1), and chose the pair with the higher probability as the answer. The percentage of test questions that each model answered correctly for each relation was calculated.

Table 1: Examples of word pairs used in the training sets and generalization tests (correct option on left).

Training pairs	Testing pairs
Function	
door-open	rabbit-hop vs. rabbit-bunny
sun-warm	cup-drink vs. cup-mug
zoo-animals	smile-happy vs. smile-frown
Synonyms	
liberty-freedom	car-auto vs. car-bus
huge-enormous	weak-feeble vs. weak-strong
forest-woods	sad-unhappy vs. sad-sadder
Linear ordering	
worse-worst	inch-foot vs. inch-length
kitten-cat	rain-downpour vs. rain-fall
tap-strike	pebble-rock vs. rock-mineral
Antonyms	
weak-strong	shallow-deep vs. shallow-depth
start-finish	float-sink vs. float-boat
slowly-quickly	find-lose vs. find-search

Simulation Details

Inputs for each word were LSA vectors of length 300. The LSA algorithm orders its features from highest to lowest in terms of their predictive power. Preliminary tests indicated that most of the information useful for our learning models was encoded in the first ten features of the LSA vectors. Accordingly, we used just these first ten features for each word as inputs. The full vector for a word pair included the basic and derived features, **X** = [**A**, **B**, **AB**, |**A** – **B**|], with a total length of 40 features.

In the implementation of the model by Silva et al. (2007), the dataset for computing the empirical prior included all AB word pairs plus a large number (>3500) of unrelated word pairs. Each unrelated word pair was weighted by approximately the ratio of the total population of unrelated word pairs to the number of unrelated word pairs that were sampled. After obtaining the prior, the model employed variational methods to compute the posterior distribution for **w** using the AB training pairs for each relation separately.

In the simulation of our hierarchical model, the values of hyperparameters (a_0, b_0) were searched separately for each relation to maximize generalization performance.

To provide baselines for evaluating the two Bayesian learning models, we applied three simpler methods of judging the correct relational alternative. First, we calculated the mean cosine distances of the correct alternative and its foil to the training set using "raw" LSA vectors, i.e., using only the basic features [**A**, **B**] over all 300 dimensions of the LSA vector for each word in a pair (yielding 600 features total). Specifically, we computed the average of cosine distances between a CD pair and all AB

pairs in the training set, and for the corresponding CD' pair and all AB pairs. The baseline decision for the discrimination task was determined by which pair yielded the closer cosine distance. The performance of this method informs us about the amount of information that "raw" LSA vectors provide for the four abstract relations of interest.

Second, we used an additional cosine distance measure defined over the same feature vectors as those used by the Bayesian models, i.e., the **X** vectors, which included the first ten features of the LSA vector for each word, plus the corresponding derived features.

Third, we examined the performance of simple logistic regression (which obtains the relational representation **w** through maximum-likelihood estimation) using the first ten LSA dimensions and the full set of derived features.

Results and Discussion

The five modeling methods were evaluated on nine different sets of training pairs and testing pairs. Each set was randomly chosen from the analogy problems available to us. Mean proportion correct over the nine different training/test sets for each of the methods described above is shown in Fig. 3. Overall, the Bayesian model incorporating the hierarchical prior yielded the best generalization performance for all four relations, and in each case was reliably more successful than any of the three baseline models. The proportions correct for the hierarchical model were .78 for function, .72 for synonyms, .86 for linear ordering, and .66 for antonyms. In general, the generalization performance for the Bayesian models was best for linear ordering and weakest for antonymy. It should be noted that the linear ordering relation can be viewed as a generalization of the type of specific comparative relation (e.g., "larger than") to which the learning model proposed by Doumas et al. (2007) has been applied.

The Importance of the Prior

The improvement in generalization performance of the Bayesian models over the MLE logistic regression model illustrates the importance of the prior distribution on the relational weights **w**. This result suggests the possibility that children may also benefit from prior knowledge, either innate or acquired through previous experience, when learning new abstract relations. They may, for example, first learn to distinguish related or generally similar concepts from unrelated concepts before discriminating among more specific relations. Future experiments could explore the kinds of prior training that best aid human learning of new abstract relations, and compare the results with model performance using different priors.

The superior generalization of the Bayesian model using the hierarchical prior compared with the model using the empirical prior indicates that learning can be further improved by introducing a more effective prior. Using the general prior knowledge obtained by contrasting related and unrelated relations is a sensible choice in the applications on which Silver et. al. (2007) focused. However, this empirical prior may not be sufficient to provide informative guidance for inferences in the high dimensional space created using LSA inputs. Adopting a hierarchical prior increases learning power by incorporating soft constraints on the relational representation, **w**, and its associated uncertainty**.**

Why are Antonyms so Hard?

The fact that the Bayesian models performed relatively poorly on antonyms warrants further analysis. It should be noted that for antonyms only, the cosine distance method based on 300 LSA dimensions (with basic features only) outperformed cosine distance based on 10 LSA dimensions and the full set of derived features. This finding raises the possibility that finding a good representation for antonymy

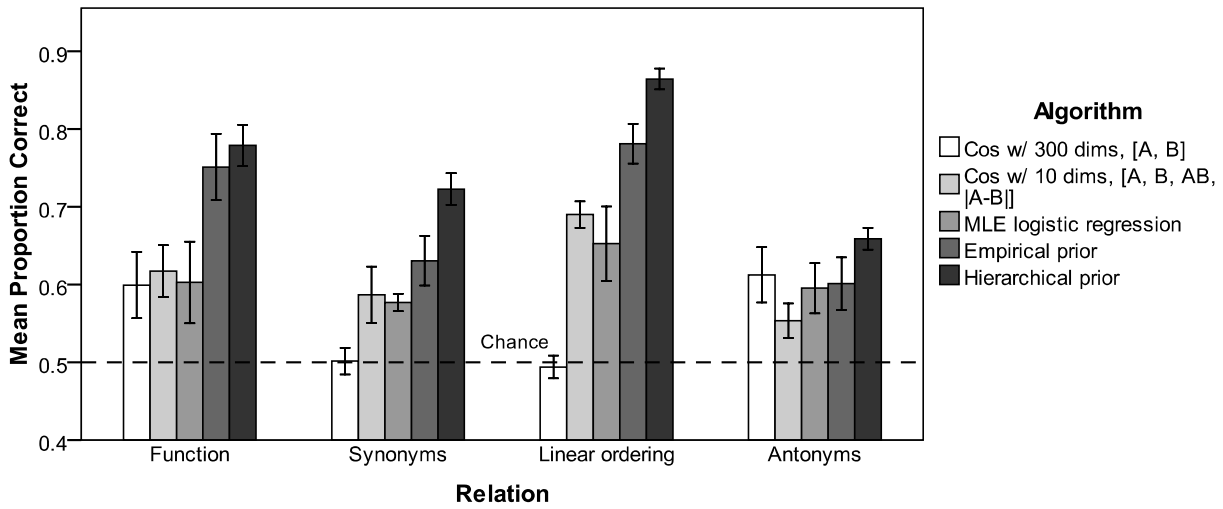


Figure 3: Simulation results. Prediction accuracy for generalization of relations in the two-alternative forced-choice relation-discrimination task. Error bars represent 1 standard error of the mean, based on 9 random samples of training/test items.

may require attention to more feature dimensions than is the case for the other relations. Another possible reason for their greater difficulty is that antonyms are usually very similar concepts that are dissimilar in only a few aspects (e.g., both *love* and *hate* can be used as a noun as well as a verb, and are strong emotions that one sentient being can have about another). Moreover, the aspects or dimensions on which antonymous concepts differ vary from one pair to another (e.g., *love-hate* vs. *black-white*). The shifting relevance of features makes learning a good representation for antonyms challenging, especially using a method that learns weight distributions over a fixed set of features.

Conclusions

We investigated the possibility that abstract semantic relations can be learned at least in part by purely data-driven statistical techniques applied to concept pairs represented by unstructured feature vectors. By using LSA vectors as inputs we avoided any hand-coding of semantics or relational structure, while assuring that inputs were of realistic complexity. Compared to baseline performance (inference based on cosine similarity of test options to the training set and MLE logistic regression), two models of relation learning based on Bayesian logistic regression achieved higher overall performance on a transfer test requiring discrimination between learned relations instantiated entirely by new concepts. The more successful of the two models incorporated hierarchical priors.

Neither model approached perfect performance on transfer problems. However, considering the small size of the training set (less than 20 examples of each relation), the total absence of overlap between training and test items, and the relatively subtle discrimination of relations required on the generalization test, these preliminary findings are encouraging. Further exploration of statistical approaches to learning abstract semantic relations appears to be warranted.

Acknowledgments

We thank Walter Kintsch, Tom Landauer and Praful Mangalath at the University of Colorado Institute of Cognitive Science for kindly providing us with LSA vectors. This research was funded by a University Fellowship and a Chancellor's Prize from the Graduate Division at the University of California, Los Angeles, and by ONR grant N000140810186.

References

Doumas, L. A. A., Hummel, J. E., & Sandhofer, C. M. (2008). A theory of the discovery and predication of relational concepts. *Psychological Review, 115*, 1-43.

Fodor, J. A. (1975). *The language of thought*. Cambridge, MA: Harvard University Press.

Gentner, D., & Rattermann, M. J. (1991). Language and the career of similarity. In S. A. Gelman & J. P. Byrnes (Eds.), *Perspectives on thought and language: Interrelations in development*. London: Cambridge University Press.

Glass, A. L., Holyoak, K. J., & Kossan, N. E. (1977). Children's ability to detect semantic contradictions. *Child Development, 48*, 279-283.

Halford, G. S. (1993). *Children's understanding: The development of mental models.* Hillsdale, NJ: Erlbaum.

Jaakkola, T. S., & Jordan, M. I. (1999). Bayesian parameter estimation via variational methods. *Statistics and Computing, 10*, 25-37.

Jones, S., & Murphy, M. L. (2005). Using corpora to investigate antonym acquisition. *International Journal of Corpus Linguistics, 10*, 401-422.

Kemp, C., & Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences, USA, 105*, 10687-10692.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review, 104*, 211-240.

Mangalath, P., Quesada, J., & Kintsch, W. (2004). Analogy-making as predication using relational information and LSA vectors. In K. D. Forbus, D. Gentner & T. Regier (Eds.), *Proceedings of the Twenty-sixth Annual Meeting of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Penn, D. C., Holyoak, K. J., & Povinelli, D. J. (2008). Darwin's mistake: Explaining the discontinuity between human and nonhuman minds. *Behavioral and Brain Sciences, 31*, 109-178.

Rogers, T. T., & McClelland, J. L. (2008). Précis of *Semantic cognition: A parallel distributed processing approach*. *Behavioral and Brain Sciences, 31*, 689-714.

Silva, R., Airoldi, A., & Heller, K. (2007). *Small sets of interacting proteins suggest latent linkage mechanisms through analogical reasoning* (Tech. Rep. GCNU TR 2007-001). London: University College London, Gatsby Computational Neuroscience Unit.

Silva, R., Heller, K., & Ghahramani, Z. (2007). Analogical reasoning with relational Bayesian sets. In M. Mella & X. Shen (Eds.), *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*.

Smith, L. B. (1989). From global similarities to kinds of similarities: The construction of dimensions in development. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning*. Cambridge, UK: Cambridge University Press.

Smith, L. B., & Sera, M. D. (1992). A developmental analysis of the polar structure of dimensions. *Cognitive Psychology, 24*, 99–142.

Turney, P., & Littman, M. (2005). Corpus-based learning of analogies and semantic relations. *Machine Learning, 60*, 251–278.

Wolfe, M. B. W., & Goldman, S. R. (2003). Use of latent semantic analysis for predicting psychological phenomena: Two issues and proposed solutions. *Behaviour Research Methods, 35*, 22-31.

You Can't Wear a Coat Rack: A Binding Framework to Avoid Illusory Feature Migrations in Perceptually Grounded Semantic Models

Michael N. Jones (jonesmn@indiana.edu)
Department of Psychological and Brain Sciences
Indiana University, Bloomington, Indiana USA

Gabriel Recchia (grecchia@indiana.edu)
Cognitive Science Program, 1910 E 10th St.
Indiana University, Bloomington, Indiana USA

Abstract

Recent Semantic Space Models (SSMs) are now integrating perceptual information with linguistic statistics into a unified mental space, offering a solution to the criticism that SSMs are disembodied. However, these new models introduce the problem of illusory feature migrations. When the word *dog* is perceived, its perceptual features should migrate to *hyena*, so the system can infer the perceptual features for a non-perceived word (hyenas have fur). In doing so, however, the models are unable to avoid migrating the features for *dog* to syntagmatically related words, such as *bone.* As a result, the models incorrectly infer that bones have fur. We argue that the problems of perceptual grounding and word order are not independent—a model of word order information is needed to correctly infer how features should migrate in mental space. We introduce a multiplicative binding framework that allows all information sources to be stored in a composite mental space, but features will only migrate to words that share sufficient order information with directly perceived words.

Keywords: semantic space models, symbol grounding problem, perceptual integration, embodied cognition.

Introduction

Semantic Space Models (SSMs) have seen remarkable success in recent years as models of how humans learn the meanings of words from repeated episodic experience, and for how lexical semantics are represented in mental space. Many types of SSMs now exist, with several modifications to better approximate human semantic cognition.[1] In general, these models all create semantic representations from statistical regularities in large linguistic corpora, building on Harris' (1970) *distributional hypothesis* of lexical semantics: the more similar the contexts in which words are experienced, the more similar their meanings. SSMs have successfully accounted for a wide variety of human semantic data, ranging from semantic priming and free association, up to high-level discourse processing by applying compositional algorithms to SSM representations.

Despite their successes, SSMs have been heavily criticized as implausible psychological models on a number of grounds. Firstly, most of these models have been criticized as "bag-of-words" models, in that they simply consider the context in which the word occurs, but ignore the statistical information inherent in word transitions. Recent solutions to the word-order problem use binding operations to learn a blended semantic space in which a word's representations reflects its history of co-occurrence with, and position relative to, other words (e.g., Jones & Mewhort, 2007). Further, these models are able to retrieve plausible n-gram information (coarse grammaticality) directly from the blended space, without the need for explicit rules of grammaticality. The integration of word order information has been shown to give a much better fit to human data in a variety of semantic tasks.

Secondly, SSMs have been criticized as "disembodied" in that they learn from only linguistic information but are not grounded in perception and action (see de Vega, Graesser, & Glenberg, 2008 for a workshop on the issue). The lack of grounding in SSMs is in direct contrast to the recent literature on embodied cognition, demonstrating that a word's meaning is grounded in sensorimotor experience. Sensorimotor information is an inherent part of the semantic organization of the human lexicon, but much of this information cannot be learned from statistics in a text corpus—it must be learned from multisensory experience (but see Riordan & Jones, in press). In addition, current models have a symbol-reference problem: there is no way to link a word's internal representation back to its referent in the real world.

We are now seeing the emergence of the first perceptually grounded SSMs. As a proxy for sensorimotor experience, these models use norms of human-generated features (such as the norms of McRae et al., 2005). These norms represent aggregate human productions of the physical properties, appearance, sounds, smells, functional properties, etc. for concrete nouns and event verbs based on multisensory experience. For example, the feature <has_4_legs> will have a high probability for *dog* and *cow*, but a low probability for *centipede*, and a zero probability for *strawberry*. However <is_red> is a highly salient feature of *strawberry* and not for *dog*.

Most of the new grounded SSMs simultaneously consider the distribution of words across contexts in a text corpus and the distribution of words across perceptual features, allowing them to extract joint information between the two data sources. This allows the models to make implicit inferences across the two information sources: if the model learns from perceptual experience that *sparrows* have beaks, and from linguistic experience that *sparrows* and *mockingbirds* are used in a similar distributional fashion, it naturally makes the inference that *mockingbirds* have beaks. The inference chain works in the opposite direction as well. Most impressive, given a novel word most of these models can retrieve an accurate representation of the perceptual features of the novel word's referent. Simulations have

[1] For recent advances in SSMs, see the upcoming issue of *Topics in Cognitive Science* edited by Danielle McNamara.

demonstrated that the blended linguistic/perceptual mental space may yield a superior approximation of human data.

However, a major issue common to all of these new grounded models is that they have no way to discriminate between syntagmatic relationships (e.g., the relationship between *bee* and *honey*) and category-based paradigmatic relationships (e.g., *bee* and *wasp*). The linguistic abstraction phase of these models will learn to position the vectors for *car, automobile,* and *road* close in semantic space. This produces the problem that the model cannot distinguish which regions of space may adopt features that migrate from a perceptually grounded word during the feature inference phase. The result is that the model correctly infers that *automobile* <has_wheels>, but it also incorrectly infers that *road* <has_wheels>. We refer to these errors as *illusory feature migrations*, and argue that errors of migration are much more common in semantic space than are correct migrations, which can severely pollute the resulting semantic space relative to a human representation that would not contain this type of error.

One reason these models fail to discriminate between context-based syntagmatic vs. category-based paradigmatic relationships is that they ignore word order information, which is a powerful cue for category membership (Jones & Mewhort, 2007). That is, words that are flanked by similar n-grams tend to belong to the same conceptual categories. Extensive study in the field of category-based inference has investigated the ways in which category structure constrains feature generalization (for reviews, see Heit, 2000; Rips, 2001). To ignore word information is to ignore a very salient cue to category membership at an SSM's disposal.

To be clear at the outset, we strongly commend the authors of these perceptually grounded models for taking a huge step in the right direction towards our understanding of human semantic representation. However, a plausible model must also be able to filter components of this representation so that perceptual information may generalize to paradigmatically but not syntagmatically similar words (i.e., from *car* to *automobile* but not *road*). Here we explore the utility of a formal binding framework based on ideas from signal processing and Jones and Mewhort's (2007) BEAGLE model that has these desiderata.

Grounding Semantics in Perception and Action

Recent attempts to ground SSMs in perception and action can be placed into one of two classes: post-hoc inference models, and ad-hoc inference models. Both types can be trained on the same text corpus and feature representations (e.g., TASA and McRae et al., 2005).

Post-hoc inference models begin with the abstraction of a text corpus into a reduced vector space (a traditional SSM), and then attempt to bind these linguistic vector representations to the feature norms. For example, Durda, Buchanan, and Caron (2009) train a feedforward neural network to associate linguistic vectors with their corresponding activation of features. Given the linguistic representation for *dog*, the output feature <has_fur> should be activated but the output feature for <made_of_metal> should be inhibited. After iterative training with backprop, the model can infer the correct pattern of perceptual properties for words that did not have a perceptual feature vector. At its core, this technique simply maps similar linguistic vectors to similar output vectors, as with other pattern generalization applications of feedforward networks.

Ad-hoc inference models typically begin with a raw word-by-document matrix of a text corpus and a word-by-feature matrix of a feature database. During learning, the model attempts to learn a word's representation by simultaneously considering inference across documents and features. An excellent example of an ad-hoc model is presented in Andrews, Vigliocco, and Vinson (2009). Andrews et al., use a Bayesian framework to infer the joint distributional information for a word between linguistic and perceptual data. It is important to note that their technique is *joint inference*: it squeezes more information out of the data than simply adding perception to linguistic experience. Andrews et al. convincingly demonstrate that their joint model gives better fits to word association data than a model that considers only one data source, or the simple addition of the two sources.

Illusory Feature Migrations

A major problem with both post-hoc and ad-hoc inference models is that they must exhibit illusory feature migrations as a consequence of their architecture. An illusory feature migration occurs when a non-perceived word adopts erroneous features from a linguistically related word simply because they are proximal in semantic space. This is a common issue in the aforementioned models because they do not have order information to discern between syntagmatic and paradigmatic word relations. If the models are optimized on free-association data (which is strongly dominated by syntagmatic productions), then they must position syntagmatically related words like *bee* and *honey* close in space, as well as paradigmatically related words like *bee* and *wasp*. As a result, the inference mechanism simply sees both *honey* and *wasp* as similar patterns to *bee*, and naturally makes the inference that *honey* can fly and has wings.

Note that the "migration" described need not be a dichotomous on/off feature. It is simply the case that the inferred distribution over possible features for *honey* has some correlation with that of *bee* simply because their distributional structure in language has overlap. This overlap introduces error in the labeling of novel referents (e.g., a novel object that looks like an insect will activate words like *honey* as potential labels). Furthermore, this inference error will introduce noise to the overall semantic organization, which will lead to a poorer account of human semantic data compared to a human who will not make these inference errors. The aforementioned models demonstrated examples of correct feature generalizations in their papers; what was not illustrated is the larger number of incorrect feature generalizations.

Presumably, humans use word-order information to constrain the inference of features in mental space. This information allows a model to distinguish what types of words may adopt features given a perceived target word. Rather than making this a terse rule-based model, we choose to adopt a graded feature migration framework—words adopt the aggregate features of proximal words that have features, weighted by their similarity in order space. However, it is also important to keep the sources (context, order, perception) blended to account for the wide range of embodied semantic data. This requires a model that can create a blended semantic representation, but that can know what part of the semantic signal to use in computing similarity for feature migration. We next describe a simple framework towards this type of integrated model, test its behavior on an artificial language paradigm, and then scale it up to a real language corpus to see how the properties are maintained at a large scale.

A Feature-Binding Framework

Our goal was to build an SSM with two key properties. First, context, order, and feature information should be represented as patterns in high-dimensional vectors. Even though these three sources of information should be blended within a single vector, it should be possible to determine the degree of similarity between two words in context space, order space, or feature space alone. Because context, order, and feature information is distributed, computing a vector cosine between two vectors reflects their similarity when all three sources of information are taken into account.

Second, feature migration should occur, but features should only migrate to words with which they share order information (i.e., words that are commonly flanked by similar n-grams). For example, *food* and *table* will share primarily context information, whereas *table* and *countertop* will share primarily order information; therefore, features should migrate from *table* to *countertop*, but not from *table* to *food*.

Encoding. Our model is similar to other SSMs that represent both context and order information with fixed-length high-dimensional vectors (Jones & Mewhort, 2007; Sahlgren et al., 2008). When a word w is encountered in the input text for the first time, it is assigned an initial "environmental" vector e_w—a random vector whose elements are randomly selected from a Gaussian distribution of mean 0 and variance 1. Environmental vectors are intended to represent the static properties of a word's surface form, such as its orthography and phonology, and are not updated during processing. The new word is also assigned an initially empty memory vector m_w to represent its semantics. When the model encounters a new sentence in the input corpus, m_w is modified according to the update rule:

$$m_w = m_w + (C_I \odot context) + (O_I \odot order) + (F_I \odot features_w)$$

where the circumpunct "$\odot$" denotes elementwise vector multiplication, one of a class of multiplication-like operators that vector symbolic architectures employ to combine vectors in a neurally plausible manner (Levy & Gayler, 2009; Kanerva, 2009). C_I, O_I, and F_I are *indicator vectors*—unchanging vectors that are bound with vectors representing context, order, and feature information, respectively. They serve to "tag" the source of the information signal (context, order, or perception). They may be initialized either as random vectors, or as binary vectors of ones and zeros sharing little or no overlap with each other.

As in Jones & Mewhort (2007) and Sahlgren et al. (2008), the *context* vector represents co-occurrence information: it is the sum of all environmental vectors of words occurring in the same sentence as w. The *order* vector is the sum of all n-grams surrounding w up to some fixed window size, where an n-gram is represented by binding the environmental vectors of all the words comprising the n-gram via elementwise multiplication. In the experiments presented here, only bigrams directly to the left and right of w are considered. As in Sahlgren et al. (2008), words to the right and left are distinguished by rotating the environmental vectors by one unit in a positive or negative direction, respectively. Finally, the *features* vector represents information about sensorimotor features of words. Each of 2,526 features taken from the feature norms of McRae, et al. (2005) was assigned a unique random vector. If w is the word for one of the 541 concepts for which feature norms were collected, $features_w$ is the sum of the five vectors that correspond to the five features that were attributed to w by the greatest number of participants. If w is not among the concepts in the McRae et al. feature norms, $features_w$ is initialized as a vector of zeroes (and only acquires nonzero values during training, when vectors are added to m_w via the update rule). The fact that $features_w$ has a w subscript while *context* and *order* do not reflects the fact that $features_w$ is derived from information about w in the feature norms, while *context* and *order* represent information about the sentence currently being processed.

Retrieval. After training, the cosine between every pair of memory vectors is calculated to determine the model's estimate of the semantic similarity between words. These similarity scores can be thought of as distances between points in a high-dimensional space, which we refer to as the *composite space*. In addition to having a lower computational complexity than circular convolution, one benefit of using elementwise vector multiplication for binding the information source tag is that the operation serves as its own approximate inverse when vector elements are sampled from a z-distribution, hence:

$$X \approx (X \odot Y) \odot Y \qquad (1)$$

This allows vectors to be elementwise multiplied with the aforementioned context indicator vector C_I before calculating their cosines. The operation serves to 'unbind' the C_I * *context* binding, yielding a *context space* in which two words' distance from each other reflects the amount of context information they share (but is not heavily influenced

by shared order or feature information). Similarly, unbinding via elementwise multiplication with O_I yields an *order space* in which cosine similarity reflects the amount of shared order information; unbinding with F_I yields a *feature space* where feature information is paramount.

Experiment 1

The objective of Experiment 1 was to determine whether the binding model we outlined does in fact possess the desired property of representing context, order, and feature information in a separable fashion, and whether it behaves appropriately with respect to feature migration. Demonstrating this required training the model on a corpus in which the amount of context, order, and feature information that words share is known, which is best accomplished using a corpus of an artificial language. Strictly controlling the input allows us to determine conclusively whether the model at least exhibits the desired properties in the simplest case and lets us more clearly observe how the inclusion or exclusion of different types of information affects the similarity space.

Method

Input corpus. The model was trained on a corpus of 1,000 sentences from a simple artificial language. This language was designed such that it would contain some word pairs that shared context information but not order information, some pairs that shared order information but not context information, and some words that shared context as well as order information. The language used is described by the following context-free grammar (symbols in bold are terminal symbols):

S → **A** *Aux* **B** *Num* **Cs** | **D** *Aux* **E** *Num* **Fs**
Aux → **can** | **should** | **would** | **could** | **does**
Num → **two** | **three** | **four** | **five** | **six**

Sentences of the corpus were generated randomly, with each possible transition of equal probability. Thus, it consisted of sentences such as "A can B three Cs", "A would B four Cs", "D should E three Fs", and so forth. In this corpus, *A, B,* and *Cs* each share context information, as they always co-occur, but they do not share order information. If this were a real language, one could think of *A, B,* and *Cs* as fillers for three different grammatical roles. Similarly, *D, E,* and *Fs* share context, but not order, information. In contrast, the members of pairs {A, D}, {B, E}, and {Cs, Fs} each share order information, but significantly less context information. The auxiliary verbs {can, should, would, could, does} and numbers {two, three, four, five, six} share significant amounts of order information with each other. They also share context information: even though the grammar allows auxiliaries and numbers to co-occur with any of *A, B, Cs, D, E,* or *Fs*, each auxiliary always co-occurs with some number.

Procedure. Two simulations were conducted. In Simulation 1, no feature information was included. In Simulation 2, we retrained the model with the full update rule $m_w = m_w + (C_I \odot context) + (O_I \odot order) + (F_I \odot features_w)$, adding five vectors corresponding to five features for the word "strawberry" from the McRae et al. norms to the concept for the word *A* (*a_fruit, grows_on_plants, grows_in_fields, grows_on_bushes,* and *has_green_leaves*). We compared the model under three conditions: context, composite, and order. In each condition, feature migration proceeded by unbinding $m_w \odot F_I$ to retrieve an approximation $features'_w$ of $features_w$, and adding this approximation to every other memory vector m_i in proportion to the strength of their similarity in the relevant space (context space, composite space, or order space, depending on condition). That is, features tend to be more likely to migrate in the order condition between two words that share a large amount of order information than between two words that do not. Because we are interested in migrating features not merely to words that are "close" to the perceived word but rather to words that are similar to *w* in terms of their relationships to other words, the similarity between words w_1 and w_2 is obtained by correlating a vector of w_1's cosine with each word in the lexicon with a vector of w_2*'s* cosine with each word in the lexicon. However, using just the cosine of w_1 and w_2 yields largely similar results.

Simulation 1.1. Tables 1 and 2 illustrate the most similar words to *A, B, Cs, D, E, Fs, can,* and *two* in context and order space, respectively, after training using the update rule $m_w = m_w + (C_I \odot context) + (O_I \odot order)$; no feature information was included in this simulation. In the absence of feature information, context and order information are separable in this model, despite the fact that both information sources are fully distributed across vector elements. Appropriately, the members of {A, B, Cs} cluster together in context space, as do the members of {D, E, Fs}. Additionally, pairs {A, D}, {B, E}, and {Cs, Fs} cluster together in order space. Although they do not appear in the tables, auxiliaries and numbers also cluster together.

Table 1. Z-scores of cosines of the most similar words to *A, B, Cs,* and *D* in context space, Simulation 1.

A		B		Cs		D	
A	3.6	B	3.6	Cs	3.6	D	3.6
B	.20	Cs	.20	B	.21	E	.18
Cs	.16	A	.16	A	.13	Fs	.15
five	-.08	two	.01	two	-.07	three	-.06
two	-.09	five	-.01	five	-.09	could	-.09

Table 2. Z-scores of cosines of the most similar words to *A, B, Cs,* and *D* in order space, Simulation 1.

A		B		Cs		D	
A	3.5	B	3.7	Cs	3.5	D	3.5
D	1.2	E	.32	Fs	1.1	A	1.2
B	-.10	A	-.03	B	-.15	Fs	-.14
Cs	-.13	Cs	-.04	A	-.17	E	-.17
can	-.24	can	-.22	can	-.24	can	-.30

Simulation 1.2. Table 3 illustrates the standardized correlations of vector cosines of the four most similar words to A under each migration condition. Because the migration rule transfers feature information in direct proportion to these values, the higher the value of a word, the more feature information that word receives from A. The important pattern in Table 3 is the reversal of B and D: in context space, the syntagmatic relation between A and B is much more salient, but in the order space the paradigmatic relation between A and D is emphasized. In the overall composite space, these relations are mixed (our desired blending in full lexical space), but the information required for correct feature migration is still implicitly represented.

Table 3. Standardized correlations of vector cosines of the four most similar words to *A* under the context, composite and order conditions, Simulation 2.

context		composite		order	
A	3.5	A	3.2	A	3.4
B	.63	B	.06	D	1.2
Cs	.55	D	.04	B	.05
does	-.17	Cs	.00	Cs	-.03

Thus, it appears that only the *order* condition minimizes opportunity for illusory feature migrations while preserving the appropriate migration to D, which is paradigmatically similar to A in this corpus. Furthermore, when feature information is added, the separability between context and order space is maintained, (allowing features to appropriately migrate from A to D) and individual features can be successfully retrieved.

Experiment 2

The objective of Experiment 2 was to explore whether the proposed binding framework continues to yield distributions that inhibit illusory feature migrations (i.e., migrations to syntagmatically similar words) while facilitating appropriate feature migrations to paradigmatically similar words when scaled up to a corpus of natural language. We therefore designed a version of Experiment 1 trained on a real corpus, the TASA corpus of high-school level English text. Two simulations were conducted: The first to examine the similarity of the decoded context and order spaces to paradigmatic and syntagmatic relations, and the second to demonstrate feature migrations to category co-ordinates vs. non-categorical associates of a target word. Both simulations were identical to Experiment 1's Simulation 2 in terms of the update rule, the conditions (context, order and composite), and the feature migration rule.

Simulation 2.1. For each word, its feature vector $features_w$ was generated by summing the five vectors corresponding to the five features from the McRae et al. norms attributed to *w* by the greatest number of participants. As test items, we extracted 1075 word pairs from the word association norms of Nelson, McEvoy, & Schreiber (1998) for which both the first word of the pair (the cue) and the second word of the pair (the target) were members of the McRae et al. feature norms[2]. For each pair, we determined the category membership of each word, using the categories employed by Cree & McRae (2003, Appendix B): weapons, vehicles, foods, and so forth. Cree & McRae explicitly list which normed words belong in which categories, allowing us to code whether the cue was a member of the same conceptual category as the target. The 690 pairs in which both words shared a category were interpreted as being paradigmatically related (e.g., *apple-pear*), while the 385 paired words not sharing a category were interpreted as being syntagmatically related (e.g., *apple-crab*). The fact that two words are associates and do not appear in the same category does not guarantee syntagmatic similarity nor does it preclude phrasal association, however, informal observation suggests that many word pairs in the latter condition tend to appear in collocations or other classic syntagmatic relationships for which feature migration would be inappropriate. Indeed, the cosine similarity scores from the McRae et al. (2005) feature vectors for the word pairs were significantly higher for our paradigmatically related words than for syntagmatically related ones, $t(1073) = 24.66$, $p < .001$.

Motivated by the results of Experiment 1, we predicted that words sharing paradigmatic relationships would be closer in order space than in context space. This pattern of results would suggest that attending to order information facilitates more feature migrations among paradigmatically related words than among syntagmatically related ones, while attending to context information does just the opposite. For paradigmatically related words, the model's cosine similarities were significantly higher in order space than in context space, $t(689) = 2.96$, $p < .01$. That is, words in paradigmatically related pairs were gauged to be more similar to each other in order space than in context space. In contrast, for syntagmatically related pairs, the model's cosine similarities were significantly higher in context space than in order space, $t(384) = 4.371$, $p < .001$.

Simulation 2.2. To briefly demonstrate how illusory feature migrations may be corrected by incorporating order information, we selected 25 "triples" from Simulation 1, each consisting of a target **T** that existed in the McRae et al. norms, a category coordinate **CC** of T, and a syntagmatically related word **R** that had an associative relationship with T but was not a member of the same category. An example triple is <T:freezer, CC: refrigerator, R:ice>. *Freezer* and *refrigerator* each share a common class (kitchen appliances); *freezer* and *ice* are certainly related as well, but not by virtue of a category relationship. Intuitively, one would like features to migrate more strongly from T to CC than from T to R, given that categories for concrete words are defined at least partly on the basis of feature overlap. For example, the most popular features of *freezer* are *used_for_storage,* and *has_an_inside*, features that are

[2] We excluded the 24 concept words that the McRae et al. norms explicitly identify as having ambiguous meanings, such as "mouse_(animal)" and "mouse_(computer)."

much more applicable to kitchen appliances than they are to related non-category members (ice, frozen waffles, etc.). If a particular feature migrated more strongly from T to R than from T to CC, this was coded as an illusory feature migration. Otherwise, it was coded as an appropriate feature migration.

The (incorrect) migration of feature information from T to R was much stronger in the context condition than the order condition, and the (correct) migration of feature information from T to CC was stronger in the order condition than the context condition. By our coding scheme, 56% of the triples exhibited at least one illusory feature migration in the context condition (recall that this means the migration was stronger from T to R than it was from T to CC). In contrast, only 40% of the triples exhibited at least one illusory feature migration in the order condition. Most notable is that *all* illusory feature migrations that took place in the order condition also took place in the composite condition, and *all* illusory feature migrations taking place in the composite condition also took place in the context condition. In other words, some illusory feature migrations that took place in the context and composite conditions were avoided in the order space. Hence, emphasizing order information by unbinding with O_I (order space) yielded equal or better results for every triple when compared with emphasizing context information by unbinding with C_I (context space) or not unbinding at all (composite space). Table 4 presents four triples that differed by condition as to whether CC or R was deemed a better candidate for feature migration from T by the model. In each case, a feature migration error was committed in the context condition, but was avoided in the order condition.

Table 4. Example feature migration errors in context space that were corrected in the order space. Cases in which the related word was the stronger attractor were considered illusory feature migrations. Target word is bold.

Triple	Features most strongly attributed to target by participants in McRae et al. (2005)	Competitor that *features*$_w$' Migrated More Strongly To, By Condition		
		context	comp	order
bottle CC: jar R: fill	used_for_holding_things made_of_glass used_for_holding_liquids made_of_plastic has_a_lid	fill	jar	jar
cat CC: mouse R: tom	has_fur an_animal a_pet eats has_whiskers	tom	tom	mouse
horse CC: cow R: saddle	used_by_riding is_large an_animal has_a_mane has_legs	saddle	cow	cow
motorcycle CC: car R: wheels	has_wheels has_2_wheels is_dangerous has_an_engine is_fast	wheels	car	car

General Discussion

Integration of sensorimotor information is an important next step in the development of SSMs. While human-generated feature norms are admittedly an intermediary step, it is important to understand the cognitive mechanisms that humans might use to integrate perception/action and linguistic structure to organize meaning in memory for when perceptual models (e.g., computer vision) are sophisticated enough to directly represent environmental information to integrate with linguistic distributional structure (see Roy, 2008 for a discussion).

While early attempts at integrating perception and language in SSMs have shown much promise, our work here indicates that a model must have a mechanism to encode temporal linguistic information to know how perceptual information may be generalized in the mental space. The binding framework presented here shows the basic property of storing all information sources in a blended composite space (as is suggested by the literature in embodied cognition). However, the model is able to identify which components of the composite signal perceptual information should be allowed to migrate to. While this scheme needs more testing at a large scale, we believe it has promise for accounting for a wide range of semantic and embodied data, and is a step toward addressing criticisms of SSMs being ungrounded.

Acknowledgements

This research was supported in part by grants from Google Inc. and IBM to MNJ.

References

Andrews, M., Vigliocco, G., & Vinson, D. P. (2009). Integrating experiential and distributional data to learn semantic representations. *Psychological Review, 116*(3), 463-498.

Cree, G. S. & McRae, K. (2003). Analyzing the factors underlying the structure and computation of the meaning of chipmunk, cherry, chisel, cheese, and cello (and many other such concrete nouns). *Journal of Experimental Psychology: General, 132*(2), 163-201.

de Vega, M., Graesser, A., & Glenberg, A. (2008). *Symbols and Embodiment: Debates on Meaning and Cognition*. New York: Oxford.

Durda, K., Buchanan, L., & Caron, R. (2009). Grounding co-occurrence: Identifying features in a lexical co-occurrence model of semantic memory. *Behavior Research Methods, 41, 1210-1223.*

Heit, E. (2000). Properties of inductive reasoning. *Psychonomic Bulletin & Review, 7,* 569-592.

Jones, M. N. & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review, 114,* 1-37.

Kanerva, P. (2009). Hyperdimensional computing: An introduction to computing in distributed representations with high-dimensional random vectors. *Cognitive Computation, 1,* 139-159.

Levy, S. D., & Gayler, R. W. (2009). A distributed basis for analogical mapping. In B. Kokinov, K. Holyoak, & D. Gentner (Eds.), *New frontiers in analogy research.* New Bulgarian University Press.

McRae, K., Cree, G., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods, 37,* 547-559.

Riordan, B., & Jones, M. N. (in press). Redundancy in perceptual and linguistic experience: Comparing feature-based and distributional models of semantic representation. *Topics in Cognitive Science.*

Roy, D. (2008). A mechanistic model of three facets of meaning. In de Vega, Glenberg, and Graesser (Eds.) *Symbols and Embodiment.*

Sahlgren, M., Holst, A., & Kanerva, P. (2008). Permutations as a means to encode order in word space. *Proceedings of Cognitive Science Society.*

The evocative power of words: Activation of visual information by verbal and nonverbal means.

Gary Lupyan (lupyan@sas.upenn.edu), Sharon L. Thompson-Schill (sschill@psych.upenn.edu)
Institute for Research in Cognitive Science, Center for Cognitive Neuroscience
University of Pennsylvania
Philadelphia, PA 19104 USA

Abstract

A major part of learning a language is learning to map spoken words onto objects in the environment. An open question concerns the consequence this learning has for cognition and perception. We show that hearing common words (e.g., dog) activates visual information more than equally informative non-linguistic information (e.g., a dog bark). The main results show that (1) pictures were verified more quickly after hearing a word than after hearing a nonverbal sound, even after hundreds of trials of practice. (2) Verbal labels activated visual information more effectively than nonverbal sounds as tested by a simple visual discrimination task that required minimal semantic processing. (3) The advantage of the verbal modality did not arise simply due to greater familiarity of verbal labels: when experience with novel labels and sounds was equated, verbal labels continued to activate the associated visual information more reliably than the equally well-learned nonverbal sounds. These results inform the understanding of how human cognition is shaped by language and hint at effects that different patterns of naming can have on individuals' conceptual structure.

Introduction

Two hallmarks of human development are the development of conceptual categories—learning that things with feathers tend to fly, that animals possessing certain features are dogs, and that foods of a certain color and shape are edible (Carey, 1987; Keil, 1992; Rogers & McClelland, 2004), and learning *names* for those categories. The latter achievement is unique to humans. While many have commented on the transformative power of names (Clark, 1998; Dennett, 1994; Harnad, 2005; James, 1890; Vygotsky, 1962), it is only recently that the interplay between verbal labels and concepts has become a subject of rigorous empirical study.

The learning of categories is, in principle, separable from the learning of language. A child can have a conceptual category of "dog" without having a verbal label associated with the category. However, in practice the two processes are intimately linked. Not only does conceptual development shape linguistic development (Snedeker & Gleitman, 2004), but linguistic development—specifically learning words—impacts conceptual development (Casasola, 2005; Lupyan, Rakison, & McClelland, 2007; Gentner & Goldin-Meadow, 2003; Spelke, 2003; Spelke & Tsivkin, 2001; Waxman & Markow, 1995; Yoshida & Smith, 2005). The effects of words on nonverbal cognition only begin at word-learning. The learned associations between words and their referents appear to continue to influence cognitive processes such as visual recognition memory (e.g., Lupyan, 2008a) and even visual processing (Gilbert, Regier, Kay, & Ivry, 2006; Lupyan, 2008b; Winawer et al., 2007). For example, hearing a verbal label such as "chair" facilitates the visual processing of the named category compared to trials on which participants know the relevant object category but do not actually hear its name (Lupyan, 2007, 2008b; Lupyan & Spivey, 2010). Hearing a label can even make an invisible object, visible (Lupyan & Spivey, 2008). One way to think about such results is that processing a verbal label preactivates the sensory and higher-level representations of objects denoted by the label—over and above activation caused by just thinking about the object category.

The present work addresses the question of how special words are in evoking visual information. Is a highly familiar concept accessed equivalently through verbal and nonverbal means with words being a merely convenient way to activate conceptual information? Or, do words evoke conceptual representations in a special way? We focus here on visual representations and compare the power of verbal and nonverbal cues to evoke visual information of both familiar and novel categories.

It has been long known that a response to a visual stimulus can be altered by a cue presented prior to the target stimulus. These cues can be nonverbal (Egly, Driver, & Rafal, 1994; Eriksen & Hoffman, 1972; Posner, Snyder, & Davidson, 1980) as well as verbal. For example, verbal cues in the form of words like "left" and "right" produce automatic shifts of attention just as reliably as nonverbal cues such as directional arrows even when the words are entirely non-predictive of the target's location (e.g., Hommel, Pratt, Colzato, & Godijn, 2001). Words related to motion, e.g., "float," have been shown to affect visual motion processing (Meteyard, Bahrami, & Vigliocco, 2007). Several studies have also shown visual object processing can be altered by verbal cues (Puri & Wojciulik, 2008; Vickery, King, & Jiang, 2005). Such effects of cues on visual processing have been linked to increases in category-specific cortical activity. For example, after seeing the word "face," participants are not only better at making a gender judgment of faces embedded in visual noise, but this enhanced discrimination correlates with activity in the fusiform face area (Esterman & Yantis, 2009). These experiments have typically used verbal labels as cues, as language is a natural way to convey information about objects. It is at present unknown whether a verbal label should be thought of as merely a convenient method of cuing—it is a primary function of language to convey information not presently in view—or whether there

is something special in the way language activates visual information. In other words, is the type of visual activation produced by hearing the word "cow" somehow special or can it be achieved by nonverbal cues similarly associated with the concept of cows, e.g., a mooing sound. Although both "cow" and the sound of a cow mooing are associated with cows, only the former is treated (in the normal course of things) as referring to a cow.

We present six experiments comparing the powers of verbal and nonverbal cues to evoke visual information. Experiments 1a-1c contrast verbal and nonverbal cues in a series of picture-verification tasks. Experiments 2a-2b contrast verbal and nonverbal cues in a visual discrimination task that requires minimal semantic processing. Experiment 3 tests whether the verbal advantage arises due to participants' greater familiarity with the verbal cues or whether the verbal advantage in evoking visual information is due specifically to the referential status of words.

Experiments 1a-1b

Experiments 1a-b comprised picture verification tasks in which participants heard an auditory cue (a label or a nonverbal sound), and then saw a matching or mismatching picture. If verbal labels activate visual information more reliably than do nonverbal cues, participants should be able to respond more quickly after hearing a label than a nonverbal sound.

Participants

A total of 116 University of Pennsylvania undergraduates volunteered in the experiments in exchange for course credit: 18 in Exp. 1a, 15 in Exp. 1b, 20 in Exp. 1c, 18 in Exp. 2a, 25 in Exp. 2b, and 20 in Exp. 3.

Materials

We selected 10 objects that were easily nameable and that had characteristic sounds (cat, car, dog, frog, gun, motorcycle, rooster, train, cow, whistle). Each category was instantiated by 5 images: a normed color drawing (Rossion & Pourtois, 2004), 3 photographs obtained from online image collections, and 1 "cartoon" image (e.g., a drawing of a cartoon dog). We used several instances of each category to introduce some visual heterogeneity. Spoken labels comprised basic-level names (listed above). Nonverbal sounds were obtained from online environmental sound libraries and judged to be unambiguously related to the target categories through piloting. All sounds were volume and length-normalized.

Procedure

On each trial participants heard a label or nonverbal sound followed by a picture, which, with equal probability, either matched the cue or did not. In the latter case, the picture was randomly selected from among the non-matching category images. Participants responded by pressing a "match" or "does not match" key on a keyboard. Immediately following their response, auditory feedback in the form of a buzz or bleep indicated whether the response was correct. Exps. 1a and 1b differed in one respect: in Exp. 1a the delay between cue offset and picture onset was 400 ms. In Exp. 1b this was increased to 1 s—a common delay used in verification tasks (Stadthagen-Gonzalez, et. al.2009). The rationale for this long delay is that it gives plenty of time for the word or sound to be encoded thoroughly by the time the picture appears. Thus, the verification RTs will be largely determined by the time it takes to recognize the picture rather than reflecting residual processing of the label or sound cue.

All factors were within-subjects and each participants completed 400 verification trials: 10 categories × 5 category exemplars × 2 levels of congruence × 2 cue-types (sound vs. label) × 2 repeats.

Results and Discussion

The data were analyzed using a mixed-effects ANOVA with all factors as within-subject effects. Only correct RTs were included. RTs less than 200 ms or greater than 1500 ms were excluded (1.9% of all trials). An analysis of RTs revealed a highly reliable validity advantage, M_{valid}=552 ms, $M_{invalid}$=600 ms, $F(1,18)=35.72$, $p<.0005$ and a strong advantage for label trials, M_{label}=563 ms, M_{sound}=588 ms, $F(1,18)=24.77$, $p<.0005$ (Figure 1A). This advantage was also observed in accuracy, M_{label}=96.2%, M_{sound}=95.2%, $F(1,18)=6.38$, $p=.02$. There were no reliable cue-type × validity interactions.

Experiment 1b likewise revealed a validity advantage for RTs, $F(1,14)=20.80$, $p<.0005$, and a strong label advantage, M_{label}=583 ms, M_{sound}=620 ms, $F(1,14)=26.80$, $p<.0005$ (Figure 1B). The label advantage was also observed in accuracy, M_{label}=97.8%, M_{sound}=96.0%, $F(1,14)=13.11$, $p=.003$. There was no significant prime-type × item interaction, $F<1$. A replication of Exp. 1b with a 1.5 s delay yielded virtually identical results.

It is conceivable that the advantage of labels is short-lived, owing its existence to the initial unfamiliarity of the sound cues. If so, the advantage should vanish or be diminished with practice. We divided each participant's data into four equal blocks and ran an ANCOVA with block as a covariate. Although participants became faster, and more accurate over time (Fs >10), there were no hints of an interaction between block and cue-type for either RT or accuracy in either experiment, Fs<1.

Experiments 1a-1b show that hearing a verbal label compared to a nonverbal sound affords a quicker identification of a subsequent picture most likely by pre-activating visual information associated with the label allowing for quicker and more accurate acceptance of a congruent picture and a quicker rejection of an incongruent picture.

Experiment 1c

The results from Exps. 1a-1b suggest that labels may play a special role in evoking visual representations owing to their referential nature (see below for discussion). Alternatively,

participants may have simply been more familiar with verbal labels than the sounds we used. This latter account predicts that, in a verification context, participants should, on seeing an image, be faster to activate a label than its nonverbal sound. Experiment 1c tested this possibility by reversing the order of the label/sound and picture. Participants now saw a picture first and had to judge a subsequently presented auditory label or nonverbal sound as either matching the picture or not. A finding of a continued advantage of labels would support the familiarity account (but would not necessarily contradict the reference-based account). A disappearance of the label advantage would provide evidence against the familiarity-based account.

Materials and Procedure

Materials were identical to Experiments 1a-1b. The procedure was identical except for the reversal of cue and target identities. On each trial, participants saw a picture for 1 s. One additional second after it disappeared, a verbal label or nonverbal sound was played and the participants task was, as quickly as possible, to press the appropriate key indicating whether the sound matched the picture (valid trial) or not (invalid trial). Participants could start responding at any time after the onset of the target label or sound, although responses generally occurred after the offset of the label or sound. Accuracy feedback was provided immediately after the response.

Results and Discussion

The data were analyzed identically to Exps. 1a-1b. There was a significant validity advantage in RTs, $F(1,19)=17.45$, p=.001: M_{valid}=575 ms, $M_{invalid}$=614ms. There was no significant difference between label and sound trials, $F(1,19)=2.62$, $p=.12$, with a trend for slower responses times to label trials than to sound trials, M_{label}=502 ms, M_{sound}=587 ms. An analysis of accuracy also failed to find a difference between label and sound cues, M_{label}=94.6%, M_{sound}=94.9%, $F<1$, further demonstrating that the nonverbal sounds were as recognizable as the labels. Comparing Exps. 1b and 1c revealed a highly reliable experiment × cue-condition interaction, $F(1,33)=24.19$, $p<.0005$.

If the label advantage observed in experiments 1a-1b was a simple consequence of participants' greater familiarity with labels, it was expected that a label advantage would be observed in the present study because viewing the picture would activate the stronger associate—the label—more quickly than the weaker associate—the nonverbal sound. However, that is not what we observed. Rather, the label advantage appears to be asymmetric, occurring when visual information is to be activated by a label cue, but not when a label needs to be activated by a visual cue. An alternative explanation is that lexical items are more complex than environmental sounds and thus require additional processing time. On this account, however, it is unclear why, if labels required greater processing time, we found a reliable verification advantage in Exps. 1a-1b.

Experiments 2a-2b

A limitation of Experiments 1a-1c is that the response requires the participants to semantically classify the image. It is thus unclear whether the label advantage derives from a faster activation of associated visual information (which facilitates subsequent recognition) or if it arises from faster activation of a semantic category itself. To tease apart these accounts, we use a task with a response that depends on visual processing, but only minimally dependent on semantic processing: discriminating an upright image from an upside-down one. The task is similar to one used by Puri and Wojciulik (2008) to examine effects of general and specific cues on visual processing.

Materials

The verbal and nonverbal sounds were identical to Experiments 1a-1c. In addition, a non-informative cue was created consisting of white noise of the same length and volume as the other auditory cues. For the pictures, only the standardized and normed instances of each category were used (Rossion & Pourtois, 2004).

Procedure

On each trial, participants saw two pictures for 200 ms. presented simultaneously to the left and right of a fixation cross. These pictures were identical except one was upside-down (flipped about the x-axis). Participants' task was simply to indicate which side of the screen contained the upright picture by pressing the 'z' key with their left index finger if it was the picture on the left, and the '/' key with their right index finger if it was the picture on the right. It was stressed that it did not matter what object was shown in the picture. The pictures were preceded by an auditory cue. The trials were evenly divided into label cues, sound cues, and uninformative noise cues. The label and sound cues validly cued the upcoming picture on 80% of the trials. On the remaining 20% the cue was invalid, for example, participants would hear "cow" (or hear a mooing sound) but then see a car. This allowed us to measure both the advantage of a valid cue relative to a noise cue (are people faster to locate the upright cow after hearing "cow"/a moo sound?) and the cost of an invalid cue relative to a noise cue baseline (are people slower to identify the upright cow after hearing "car"/a car-starting sound?) And, critically, we can compare these benefits and costs for label and sound cues.

Exps. 2a and 2b differed in only one respect: in Exp. 2a the delay between the offset of the cue and onset of the pictures was 400 ms. In Exp. 2b it was lengthened to 1 s to determine whether the results observed in Exp. 2a were due to insufficient time to process the nonverbal sound. There were 20 practice and 300 real trials.

Results and Discussion

RTs were analyzed by mixed-effects ANOVAs followed by directed t-tests. The first analysis included validity and cue-type (sound vs. label) as fixed factors (validity is undefined

for noise cue trials) and subject as a random factor. Results are shown in Figure 1. We found a highly reliable effect of validity, with valid trials being reliably faster than invalid trials, $F(1,17)=39.72$, $p<0005$. There was a significant validity × cue-type interaction with label cues showing a larger cuing effect than sound cues, $F(1,17)=8.23$, $p=.011$. Relative to the no-cue baseline, valid sound cues improved performance by a significant amount, $t(17)=2.84$, $p=.03$). Label cues also improved performance, $t(17)=5.01$, $p<.0005$, and this improvement was significantly greater than the improvement due to sounds, $t(17)=2.93$, $p=.009$. Conversely, relative to the no-cue baseline, invalid label cues significantly slowed responses, $t(17)=4.38$, $p<.0005$; sounds cues did not, $t(17)=1.19$, $p>.2$. There was a significant difference between the cost of invalid labels and the cost of invalid sounds, $t(17)=2.12$, $p=.048$. Accuracy was very high ($M=97.8\%$) and did not vary between conditions.

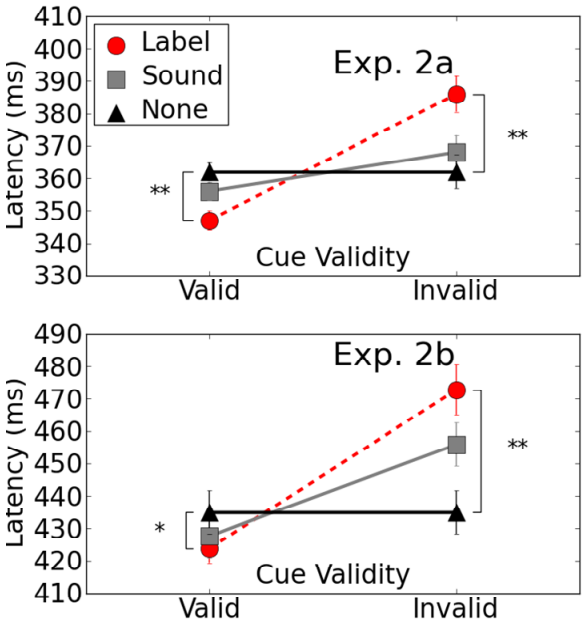


Figure 1: Results of Exps 2a-2b. Error bars show ±1 SE of the difference between noise cues and the condition closest to its mean. The mean of the noise cue trials is plotted twice for ease of comparison.

Did the label advantage result from a lack of time to process the sound cue? This was unlikely given the results of Exp. 1c, but nevertheless, we conducted a replication of Exp. 2a with a longer (1 s) delay between cue offset and picture onset. As shown in Figure 2b, valid labels helped relative to baseline, $t(24)=2.45$, $p=.022$, while sounds did not, $t(24)=1.13$, $p>.2$, though the interaction was not significant. Invalid sounds now hurt performance relative to baseline (although not as much as labels). In sum, labels continued to function as more effective cues than sounds.

With a longer time to process the cue, the nonverbal cues start to act more like verbal cues, quite possibly because participants may explicitly label the nonverbal sounds.

Experiment 3

The studies thus far examined effects of words/sounds on visual processing of objects with which participants have had extensive prior experience. We had no way of knowing whether and what types of differences in experience may have produced the label advantage observed in Exps 1-2. The label advantage is unlikely to be a product of a simple familiarity difference between labels and sounds (see Exp. 1c), but it is possible that labels have a greater power to evoke visual information because they have been more frequently encountered in the context of the visual referent. In Exp. 3, we exerted complete control over by training different groups of participants to associate either novel labels or nonverbal sounds with novel stimuli. A finding of a label advantage in this context would lend support to the idea that words have a special power to evoke visual information.

Materials

The learning set consisted of 6 novel 3D objects (Figure 2). There were 3 variants of each object to increase visual heterogeneity. These variants involved changes in viewpoint and slight changes in feature configuration. Each category was paired with a novel label (shonk, whelph, scaif, crelch, foove, and streil). Each of these nonce words was designed to have approximately equal bigram and trigram statistics and similar real-world lexical neighborhoods. We also created 6 nonverbal sounds: one for each category. These were created by modifying and combining environmental and animal sounds to create 6 sounds that were not readily nameable, as judged by pilot testing.

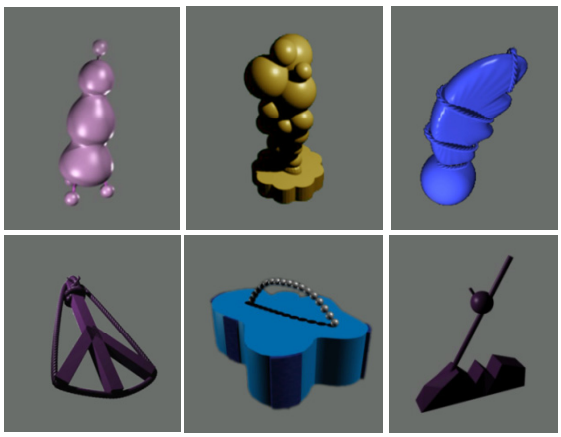


Figure 2: Materials used in the learning task for Exp 3.

Procedure

Participants were randomly assigned into *label* and *sound* groups. There were 3 parts to the experiment presented in immediate succession. In the first part, participants passively viewed 12 trials during which all three exemplars of each category were presented together with a recording, e.g., "These are all shonks" (for the label condition), or These all make the sound___" (for the sound condition).

Part 2 consisted of a verification task. Participants saw two exemplars from different categories followed by a prompt, e.g., "Which one's the streil?" or "Which one makes the sound___") and had to select whether the left or right stimulus matched. There were 180 training trials.

The last part was a replication of Experiment 2b with the novel stimuli. That is, participants judged whether the left or right picture was upright (i.e., in the familiar orientation) after hearing a sound or label cue (now without a sentential context). The images were presented for 200 ms after a 1 s delay which was timed to the offset of the auditory cue.

Results and Discussion

Participants were remarkably adept at learning the 6 categories. After Part I—just two exposures to each category—participants could correctly perform the 2AFC task of Part II with ~95% accuracy. The label group was slightly less accurate and slower than the sound group, *p*s=.08, and there were no reliable condition × block interactions. By block 5 both groups were performing at 99%, demonstrating that learning names for novel categories is no more or less difficult than learning what sounds they make.

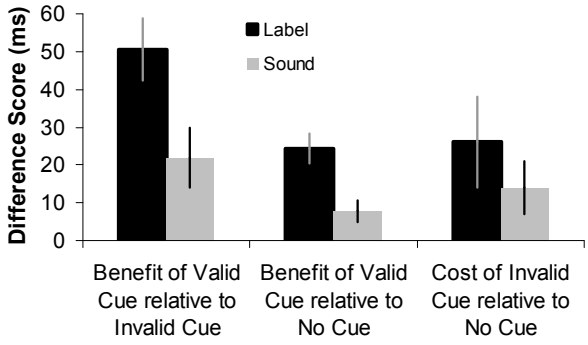


Figure 3: Cuing effects in Experiment 3. Left: $RT_{invalid}$-RT_{valid}. Middle: $RT_{no\text{-}cue}$-RT_{valid}. Right: $RT_{invalid}$-$RT_{no\text{-}cue}$. Error bars show ±SE of the mean difference score.

The critical part of the experiment was subsequent orientation judgment task. Having ruled out entirely differences in familiarity and association strength between labels and sounds, would labels continue to evoke visual activations in a more robust way than sounds? Indeed, that is what we found. As shown in Figure 3, there was a significant validity advantage, $F(1,18)=49.55$, $p<.0005$, but this advantage was significantly larger for label than sound cues, $F(1,14)=6.14$, $p=.023$. The valid cue also benefited RTs relative to the uninformative noise cue, $F(1,18)=38.34$, $p<.0005$, and this benefit was larger for the label than sound trials, $F(1,18)=10.73$, $p=.004$. Finally, there was a significant cost of hearing an invalid cue relative to no cue, $F(1,18)=8.08$, $p=.011$, but this cost was not reliably different for the two groups, $F<1$. An identical pattern of results was found when we used proportions instead of RT differences.

The two groups did not differ in overall response times, M_{label}=433 ms, M_{sound}=389 ms, $F(1,18)=1.70$, $p>.2$, or accuracy, M_{label} = 96.4%, M_{sound} = 95.7%, $F<1$.

In Experiment 3 we had complete control over participants' exposure to the pictorial stimuli, labels, and sounds. We could thereby ensure that they were equally familiar with the labels and nonverbal sounds. Participants were equally proficient in learning to associate the novel categories with labels or sounds. After only about 10 minutes of training, hearing a label or sound activated the corresponding visual form, as revealed by an RT advantage on valid trials and an RT cost on invalid trials. This in itself is quite remarkable. Critically for our thesis, the label cues were more reliable in activating the corresponding visual form than the sound cues, confirming that even when familiarity and experience with verbal and nonverbal associates is fully equated, verbal cues activate visual information more reliably than nonverbal cues.

General Discussion

Humans learn an elaborate system of sounds (or gestures in case of sign language) that refer, in a largely arbitrary way, to objects, actions, and relations. Beyond enabling linguistic communication, does the acquisition and use of the system confer certain cognitive and perceptual abilities?

In this work, we have investigated whether information communicated verbally (through words denoting concrete objects) and nonverbally (through sounds associated with those objects) activates visual information in the same way. We found that it does not. Cuing categories by using words is more effective than cuing them using nonverbal cues. Verbal cues, more than nonverbal cues appear to preactivate a visual representation of the cued category, helping when the cue is valid and hurting performance when the cue is invalid. This phenomenon is robust, being observed in virtually every subject. A number of control experiments rule out the possibility that this effect is due to different levels of familiarity with verbal versus nonverbal cues.

These findings contradict the popular view that language simply activates nonverbal concepts (Gleitman & Papafragou, 2005; e.g., Li, Dunham, & Carey, 2009; Snedeker & Gleitman, 2004) because presumably such concepts should have been activated in the same way by equally well-learned nonverbal information (as in Exp. 3), but they were not. The finding that representations of very familiar categories (e.g., dogs, cats, and cars) can be evoked more reliably by labels than by sounds, even a full second after cue offset hints at the powerful effects of language on visual activation.

How do words come to have such evocative powers? We believe it is unlikely that there is innately privileged access to vision from the verbal modality (indeed, it is unclear what an innate verbal modality would entail). Rather, the special status of words may derive from accumulated experience of treating them in a referential way (Waxman, 1999), although what exactly this entails vis-à-vis a neural mechanism remains unknown. The present results show that verbal labels serve as powerful cues (Elman, 2004; Rumelhart, 1979), invoking associated concepts and percepts in a unique way, even when the concept in question is a highly

familiar one such as [dog]. The finding that after only ~10 minutes of experience, labels affect representations of new concepts (Exp. 3), hints that long-term differences in linguistic experience can have significant effects on the ease of activating specific mental states. Rather than being simple constituent feature of the concept with which it is associated, a name appears to offer a particularly efficient route to the activation of visual and perhaps other information. Although the verbal activation of conceptual information can be deemed a human universal—perhaps the defining feature of language—the present results hint that the substantial differences in lexicalization patterns between languages may translate to cross-linguistic differences in how particular mental states can be evoked.

Acknowledgments

This work was supported by an IGERT training grant to G.L. and NIH R01DC009209 and R01MH70850 to S.T-S. We thank Nina Hsu for designing the stimuli used in Exp. 3, and thank Joyce Shin, Ali Shapiro, and Arber Tasimi for help with data collection.

References

Carey, S. (1987). *Conceptual Change in Childhood* (First paperback edition.). The MIT Press.

Casasola, M. (2005). Can language do the driving? The effect of linguistic input on infants' categorization of support spatial relations. *Developmental Psychology*, *41*(1), 183-192. doi:10.1037/0012-1649.41.1.188

Clark, A. (1998). Magic Words: How Language Augments Human Computation. In *Language and Thought: Interdisciplinary themes* (pp. 162-183). Cambridge University Press.

Dennett, D. (1994). The Role of Language in Intelligence. In *What is Intelligence? The Darwin College Lectures*. Cambridge University Press.

Egly, R., Driver, J., & Rafal, R. (1994). Shifting Visual-Attention Between Objects and Locations - Evidence from Normal and Parietal Lesion Subjects. *Journal of Experimental Psychology-General*, *123*(2), 161-177.

Elman, J. L. (2004). An alternative view of the mental lexicon. *Trends in Cognitive Sciences*, *8*(7), 301-306. doi:10.1016/j.tics.2004.05.003

Eriksen, C., & Hoffman, J. (1972). Temporal and Spatial Characteristics of Selective Encoding from Visual Displays. *Perception & Psychophysics*, *12*(2B), 201-&.

Esterman, M., & Yantis, S. (2009). Perceptual Expectation Evokes Category-Selective Cortical Activity. *Cereb. Cortex*, bhp188. doi:10.1093/cercor/bhp188

Gentner, D., & Goldin-Meadow, S. (2003). *Language in Mind: Advances in the Study of Language and Thought*. Cambridge, MA.: MIT Press.

Gilbert, A., Regier, T., Kay, P., & Ivry, R. (2006). Whorf hypothesis is supported in the right visual field but not the left. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(2), 489-494.

Gleitman, L., & Papafragou, A. (2005). Language and thought. In *Cambridge Handbook of thinking and Reasoning* (pp. 633-661). Cambridge: Cambridge University Press.

Harnad, S. (2005). Cognition is categorization. In H. Cohen & C. Lefebvre (Eds.), *Handbook of Categorization in Cognitive Science* (pp. 20-45). Elsevier.

Hommel, B., Pratt, J., Colzato, L., & Godijn, R. (2001). Symbolic control of visual attention. *Psychological Science*, *12*(5), 360-365.

James, W. (1890). *Principles of Psychology. Vol. 1*. New York: Holt.

Keil, F. C. (1992). *Concepts, Kinds, and Cognitive Development*. The MIT Press.

Li, P., Dunham, Y., & Carey, S. (2009). Of substance: The nature of language effects on entity construal. *Cognitive Psychology*, *58*(4), 487-524. doi:10.1016/j.cogpsych.2008.12.001

Lupyan, G. (2007). *The Label Feedback Hypothesis: Linguistic Influences on Visual Processing*. PhD. Thesis. Carnegie Mellon University.

Lupyan, G. (2008a). From chair to "chair:" A representational shift account of object labeling effects on memory. *Journal of Experimental Psychology: General*, *137*(2), 348-369.

Lupyan, G. (2008b). The Conceptual Grouping effect: Categories matter (and named categories matter more). *Cognition*, *108*, 566-577.

Lupyan, G., Rakison, D., & McClelland, J. (2007). Language is not just for talking: labels facilitate learning of novel categories. *Psychological Science*, *18*(12), 1077-1082.

Lupyan, G., & Spivey, M. (2008). Now You See It, Now You Don't: Verbal but not visual cues facilitate visual object detection. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 963-968). Austin, TX.

Lupyan, G., & Spivey, M. (2010). Redundant spoken labels facilitate perception of multiple items. *under review*.

Meteyard, L., Bahrami, B., & Vigliocco, G. (2007). Motion detection and motion verbs - Language affects low-level visual perception. *Psychological Science*, *18*(11), 1007-1013.

Posner, M., Snyder, C., & Davidson, B. (1980). Attention and the Detection of Signals. *Journal of Experimental Psychology-General*, *109*(2), 160-174.

Puri, A., & Wojciulik, E. (2008). Expectation both helps and hinders object perception. *Vision Research*, *48*(4), 589-597.

Rogers, T., & McClelland, J. (2004). *Semantic Cognition: A Parallel Distributed Processing Approach*. Cambridge, MA: Bradford Book.

Rossion, B., & Pourtois, G. (2004). Revisiting Snodgrass and Vanderwart's object pictorial set: The role of surface detail in basic-level object recognition. *Perception*, *33*(2), 217-236.

Rumelhart, D. (1979). Some problems with the notion that words have literal meanings. In A. Ortony (Ed.), *Metaphor and Thought* (pp. 71-82). Cambridge University Press.

Snedeker, J., & Gleitman, L. (2004). Why is it hard to label our concepts? In D. G. Hall & S. R. Waxman (Eds.), *Weaving a Lexicon* (illustrated edition., pp. 257-294). The MIT Press.

Spelke, E. (2003). What Makes Us Smart? Core knowledge and natural language. In *Language in Mind: Advances in the Study of Language and Thought* (pp. 277 -311). Cambridge, MA.: MIT Press.

Spelke, E., & Tsivkin, S. (2001). Initial knowledge and conceptual change: Space and number. In *Language acquisition and conceptual development* (pp. 475-511). Cambridge, UK: Cambridge University Press.

Stadthagen-Gonzalez, H., Damian, M. F., Pérez, M. A., Bowers, J. S., & Marín, J. (2009). Name–picture verification as a control measure for object naming: A task analysis and norms for a large set of pictures. *The Quarterly Journal of Experimental Psychology*, *62*(8), 1581. doi:10.1080/17470210802511139

Vickery, T. J., King, L., & Jiang, Y. (2005). Setting up the target template in visual search. *Journal of Vision*, *5*(1), 81-92. doi:10:1167/5.1.8

Vygotsky, L. (1962). *Thought and Language*. Cambridge, MA: MIT Press.

Waxman, S. (1999). The dubbing ceremony revisited: Object naming and categorization in infancy and early childhood. In *Folkbiology* (pp. 233 -284). Cambridge, MA: MIT Press.

Waxman, S., & Markow, D. (1995). Words as invitations to form categories: Evidence from 12- to 13-month-old infants. *Cognitive Psychology*, *29*(3), 257-302.

Winawer, J., Witthoft, N., Frank, M., Wu, L., Wade, A., & Boroditsky, L. (2007). Russian blues reveal effects of language on color discrimination. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(19), 7780-7785.

Yoshida, H., & Smith, L. (2005). Linguistic cues enhance the learning of perceptual cues. *Psychological Science*, *16*(2), 90-95.

Perceptual Simulations of Temporal Uses of *In* and *On* in First and Second Language Processing

Luca Onnis (lucao@hawaii.edu)
Department of Second Language Studies, University of Hawaii
Honolulu, HI 96822 USA

Daniel Jackson (danieloj@hawaii.edu)
Department of Second Language Studies, University of Hawaii
Honolulu, HI 96822 USA

Michael J. Spivey (spivey@ucmerced.edu)
School of Social Sciences, Humanities and Arts, University of California, Merced
Merced, CA 95343 USA

Abstract

Prepositions in natural languages often appear to be governed by arbitrary conventionalized idiomatic uses (e.g., *I was born in May*, *I will see you on Sunday*). We present empirical evidence that such prepositional uses are not entirely arbitrary, as they activate image-schematic perceptual simulations during language processing.

In Experiment 1, native speakers of English were prompted to think about either the date or the month of their birthday, and then select one of four calendar diagrams, two foils, one flat calendar and one box-like calendar diagram designed to invoke perceptual simulations of support and containment respectively. There was a significant relationship between the question prompt (implicitly eliciting in or on) and the type of calendar chosen (containment or support). Thus, spatial schemas can be spontaneously activated when thinking about time even for non-literal, idiomatic uses.

Prepositional uses are notoriously difficult for English L2 learners. We surmised that improper prepositional uses may be linked to improper underlying perceptual simulations. This was confirmed in Experiment 2 where Japanese-speaking students of English were presented the same task as Experiment 1. Here, results indicate no relation between the date or month question and calendar choice. The experiments offer both theoretical and practical insights into how prepositions are processed by individuals with varying levels of language knowledge.

Keywords: Perceptual simulations; prepositions; second language learning; embodied cognition.

Introduction

Prepositions in natural languages often appear to be governed by arbitrary conventionalized idiomatic uses (e.g., Vandeloise, 1991). For example, native English speakers tend to take it for granted that one uses the preposition "on" to refer to the date on which one was born but one uses "in" to refer to the month in which one was born. These are of course conventional idiomatic uses that didn't have to be that way, and are "partly a matter of collocational habit" (Lindstromberg, 1998, p. 76). How are these idiomatic prepositional uses processed in the mind? On one account, locative prepositions such "in" and "on" lose their original literal semantic content of containment and support respectively, to take on grammatical characteristics, a process often described as grammaticalization (Hopper & Traugott, 1993). Thus, conventionalization over time would lead to semantic bleaching, such that the same prepositions – and the expressions they are embedded in – would be processed differently, whether they are used literally or idiomatically. Some evidence for a processing difference between literal (spatial) and idiomatic (temporal) uses of prepositions comes from a preliminary neuropsychological study. Kemmerer (2005) reports that brain-damaged subjects with left perisylvian lesions failed a test of knowledge of the temporal meanings of prepositions, but passed a test that assessed knowledge of the corresponding spatial meanings of the same prepositions, suggesting that the spatial and temporal meanings of prepositions are represented and processed independently of each other in the brains of adult speakers.

In the present study we considered an alternative hypothesis, namely that idiomatic prepositional uses in one's native language are not entirely arbitrary, as they may activate image-schematic perceptual simulations during language processing (Gibbs, 2006; Richardson, Spivey, Barsalou, & McRae, 2003; Zwaan, 2004). Spatial prepositions have been studied for artificial intelligence and automated translation (e.g., Andre et al., 1987; Retz-Schmidt, 1988), as well as for their complex mappings to various gradations in spatial relations (e.g., Bowerman & Choi, 2003; Coventry & Garrod, 2004) and image-schematic mental representations (Brugman & Lakoff, 1988). In particular, previous research has demonstrated our tendency to use spatial metaphors to help us understand time (e.g., Boroditsky, 2000). A wide variety of laboratory experiments have clearly demonstrated a role for embodied sensorimotor properties (or "perceptual simulations") in language processing (e.g., Barsalou, 1999; Bergen, Matlock, Lindsay, & Narayanan, 2007; Glenberg & Kaschak, 2002; Richardson et al., 2003). There is in fact a long history to embodied sensorimotor accounts of language that predates the recent spate of laboratory experiments. The field of cognitive linguistics has provided a number of spatial

descriptions of linguistic meanings in the form of "image schemas" (Gibbs & Colston, 1995; Lakoff, 1987; Langacker, 1987; Talmy, 1983). In particular, image schemas (two-dimensional layouts of idealized trajectors and landmarks) have proven especially illustrative for understanding the varied meanings of spatial prepositions (Brugman & Lakoff, 1988; Talmy, 1983; Tyler & Evans, 2003). Indeed, image schemas may be the quintessential generic form of the perceptual simulations that underlie the understanding of prepositions like over, on, and in.

Therefore, it may not be surprising to find that more specific properties of spatial relationships (such as the containment properties of "in", or the support properties of "on") become articulated in our perceptual simulations of the spatial metaphors we use for understanding time. Thus, although they are idiosyncratic, idiomatic uses of prepositions in English such as "I will see you on Thursday" may not be entirely arbitrary. Rather, they may involve perceptual simulations and/or image schemas based on the preposition being used (e.g., Brugman & Lakoff, 1988). We set out to test this hypothesis in Experiment 1.

Studying the processing of idiomatic prepositions in adult native speakers bears not only theoretical import, but also practical implications for learning a second language. Prepositional uses are recognized as notoriously difficult for English L2 learners to acquire, especially when their native language has no equivalent prepositions (e.g., in Korean, I May was born), or it possesses one single general preposition that collapses the meaning of two (e.g., in Japanese, ni subsumes both in and on), or it uses different prepositions (e.g., in Italian, I am going in Italy). Textbooks and materials for English second language (L2) learners emphasize the arbitrary nature of non-literal prepositional uses and have little to offer except the instruction to memorize either rules or examples. In Experiment 2, we hypothesized that Japanese L2 learners of English struggle with prepositional uses particularly because they cannot rely on the congruent perceptual simulations underlying such uses. In the Discussion section, we argue that an embodied account of sentence processing can potentially change instructional practices in second language education. We propose that if image-schematic mental representations are part and parcel of sentence processing, then language teaching curricula should benefit from taking advantage of that additional source of information in training second-language learners.

To summarize, the main goal of Experiments 1 and 2 was to investigate the influence of language on the activation of image schematic perceptual simulations. Perceptual simulation was operationally defined as the process of selecting a visually presented object, a calendar, congruent with a conventionally accepted response to a target question (date or month of birth). There were two experimental conditions (Date and Month), both of which incorporated four calendar diagrams designed to invoke perceptual simulations of support (Figure 1, item 1) or containment (Figure 1, item 2) or neutral filler items (Figure 1, items 3 and 4). Participants were first asked to think about the date or month of their birthday, then select one of the four calendars (in the experimental conditions). Japanese learners of English were additionally asked to respond to a sentence completion task eliciting the prepositions in or on. The following research questions were intended to explore the issues outlined above:

1. Do native speakers of English select calendar images whose perceptual simulation (container vs. support) is congruent with the unmentioned preposition (on or in) that is associated with the prompted question (date of birth or month of birth, respectively)?

2. Do Japanese speakers of English as a foreign language select calendar images whose perceptual simulation (containment vs. support) is congruent with the unmentioned preposition that is associated with the prompted question?

3. For the Japanese participants, does the presence or absence of an image influence the accuracy of responses in the sentence completion task?

4. Do those Japanese speakers who do select the congruent image respond with the correct English preposition?

5. In the case of Japanese participants, do higher proficiency speakers select the congruent primes more often than lower proficiency speakers?

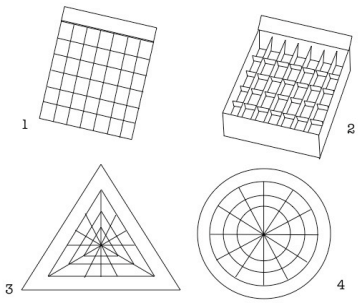


Figure 1. Calendar diagrams used in the date prompting condition, with the support image at the top-left.

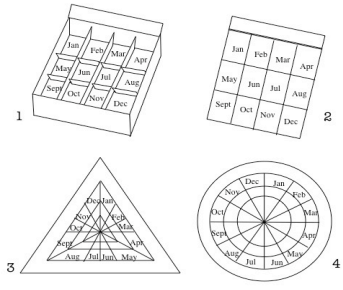


Figure 2. Calendar diagrams used in the month prompting condition, with the containment image at the top-left.

Experiment 1

Different languages use their spatial prepositions to carve up the various possible spatial relationships in a variety of ways. For example, where English uses "in" for containment spatial relationships, and "on" for support spatial relationships, Spanish and Japanese use a single preposition ("en" and "ni" respectively) for both containment and support (for discussion, see Coventry & Garrod, 2004). Moreover, where English collapses tightly-fitting containment and loosely-fitting containment into a single "in" category, Korean uses the prepositions "kkita" for the former (plus tight-fit support) and "nehta" for the latter (Bowerman & Choi, 2003; Choi & Bowerman, 1991; Mandler, 1992). Interestingly, English-learning infants can actually learn the tight-fit containment-or-support spatial category (referred to as "kkita" in Korean) when they are given a spoken novel word with which to label the image (Casasola, Bhagwat, & Burke, 2009). Thus, children's categorization of spatial relationships is influenced by the spatial prepositions they grow up with. Might even adults' real-time perceptual simulations of the language they read and hear be influenced by the preposition they use to describe an event (even if that preposition does not call for a literal meaning, but a purely idiomatic one)?

Method

Participants Fifty-one native English-speaking students were recruited at the University of Hawai'i at Mānoa.

Materials Since almost every spatial preposition in every language has a quite varied range of uses (e.g., Bowerman & Choi, 2003; Brugman & Lakoff, 1988; Haspelmath, 1997; Lindstromberg, 1998), the experiment focused on a particular idiomatic use of a spatial preposition, and examined the perceptual simulations that native speakers may be generating when they understand that particular idiom. Two sets of four calendar diagrams were designed to each include a single image that would invoke a perceptual simulation of support or containment. For instance, in the top-left corner of Figure 1, the diagram corresponds to the expression, born on [date] because it displays a spatial affordance of support, and the diagram in the top-left corner of Figure 2 corresponds to the expression, born in [month] because it displays a spatial affordance of containment. The circle and triangle diagrams are filler items intended to distract participants from figuring out the experimental manipulation and also to mitigate the potential influence of cultural bias due to the prototypicality of the flat calendar. The arrangement of the support and containment calendar diagrams was counterbalanced from left to right to avoid location preferences.

Procedure Randomly assigned participants were first prompted (without mentioning the prepositions "in" or "on") to think about the date or month of their birthday. Then one of two questions, "Which one of these calendars would you use to indicate the date [or, month] of your birth?" accompanied the calendar images depicted in Figure 1 or Figure 2 depending on the experimental condition. The calendar images were printed on a US letter sheet of paper. Participant choices were recorded by the experimenter.

Results

Congruence of perceptual simulations To address the above research questions in Experiments 1 and 2, Pearson's Chi-squared test with Yates' continuity correction was used to determine whether observed differences were statistically significant. An alpha level of .05 was set for each test we conducted. Experiment 1 was devised to answer the first research question, regarding the congruence of perceptual simulations in native speakers. The English native speakers in this study tended to select calendar images whose perceptual simulation was congruent with the preposition associated with the prompted question: 98% selected the support calendar when prompted with the date question, and 41% selected the container calendar when prompted with the month question. There was a significant relationship (χ^2 = 23.73, df = 1, $p < .001$) between the date or month question prompt (eliciting in or on) and the type of calendar chosen (containment or support). Thus, it appears that spatial schemas can be activated in one's native language when thinking about time even for idiomatic prepositional usages.

Experiment 2

A key topic of research in the area of spatial language is the cross-linguistic variation of how a given language's spatial prepositions partition and categorize various spatial relationships (Bowerman & Choi, 2003) – which brings us to the crux of the question addressed in Experiment 2. When second language (L2) learners make errors with spatial prepositions, what is the character of the perceptual simulations they generate (see also Coventry & Guijarro-Fuentes, 2008; Tyler & Evans, 2003)?

Method

Participants Eighty-two native Japanese-speaking undergraduate students enrolled in an English program at a private university in Tokyo, with a mean age of 20.5 years (SD = 5.4), voluntarily participated. On average, they had studied English for 8.8 years (SD = 2.6), and had spent a mean of 0.9 years abroad in English-speaking countries (SD = 2.1). When asked to rate their ability to use English on a scale of one to 10, their average rating was 4.5 (SD = 1.6).

Materials The same materials as Experiment 1 were used. Materials were remotely delivered via the Internet using a survey software (www.surveymonkey.com).

Procedure Participants were randomly assigned to an experimental condition or a control condition. In the experimental condition, participants were prompted with the statement that read "Think about the date [or, month] of your birthday" and then asked to select a calendar picture as

in Experiment 1. In the control condition, participants were prompted with the same statement but did not view the calendar pictures. Next all participants in both conditions were asked either, "What is the date of your birthday?" or "What is the month of your birthday?" and subsequently typed their answer in a blank field preceded by the stem, "I was born ...". Following this, they completed a 10-item cloze test (Brown, 1998) and filled out a questionnaire in Japanese, in part based on the Language Experience and Proficiency Questionnaire (Marian et al., 2007). All participants' responses (calendar choice, sentence completion, cloze items chosen, and responses to the questionnaire, were recorded by the survey software, and later downloaded for analysis by the experimenter.

Results

Congruence of perceptual simulations in L2 learners Our second question in this research project aimed to address the issue of congruence between prepositional use and elicited schemas with Japanese-speaking EFL learners. Here, in stark contrast to the findings in Experiment 1 with native English speakers, the results indicated no relation between the date or month question and calendar choice: 85% selected the support calendar when prompted with the date question, and 30% selected the container calendar when prompted with the month question ($\chi^2 = 0.92$, df = 1, p = .34). As might be expected, second language learners of English in this study did not show a tendency to select calendar images congruent with the time question posed to them. To directly compare native and non-native speakers, we collapsed the data from Experiment 1 and 2, and fitted a generalized log-linear model with three variables (Native Language, Prompt Question, Calendar Diagram). This saturated statistical model yielded a significant three-way interaction. To properly assess this significance, we followed a model simplification method (Crawley, 2005), by deleting the three-way interaction from the model, and checking whether a simpler model would lose explanatory power. Indeed the models differed (p = 0.025), so we retained the more complex model with the three-way interaction. The analyses confirmed that only the native English speakers show a tendency to select calendar diagrams consistent with the prompt question posed to them.

Calendar images and accuracy in L2 learners Our third research question asked whether the presence or absence of an image would influence English non-native speakers' accuracy of responses by way of a visual priming. Answers to the question, "What is the date of your birthday?" or "What is the month of your birthday?" were coded as either correct or incorrect for the experimental and control groups, depending on whether participants produced on or in. For the date question, neutral responses in which participants optionally deleted the preposition (e.g., *I was born Ø December 11th*) were excluded from the analysis (n = 12). To begin with, there was no apparent difference in the ability of participants assigned to the date and month to use prepositions correctly to answer the above question: 55% of participants in the date condition and 50% in the month condition used the correct preposition ($\chi^2 = 0.33$, df = 1, p = .57). Also, no significant interaction was found between the experimental and control group in terms of accuracy of response ($\chi^2 = 0.16$, df = 1, p = .69). Therefore, the presence or absence of the calendar images appeared not to influence accuracy of production. This is reasonable, given that participants were presented with a choice of four different calendars, and thus it is not apparent which of the four should have independently primed the a correct answer. These results appear to rule out an account in terms of conceptual priming, that is that conceptual representations based on a visual context prime other conceptual representations and preposition choices in production.

The possibility of an influence of calendar choice on accuracy was more closely examined through research question four, which probed whether participants who selected the congruent image provided an accurate response to the target question. Only 40% of them did ($\chi^2 = 1.2$, df = 1, p = .27), suggesting that for these participants, there is no relationship between image choice and accurate production**.**

Image choice and proficiency in L2 learners Finally, the fifth research question examined the likelihood of a relationship between image choice and proficiency level. After standard test item analysis techniques were applied to the cloze test, two items with low discrimination indices were excluded. Internal reliability was found to be .61 for the remaining eight items. We ran a logistic regression analysis, using a generalized linear model, with cloze test scores predicting congruency of calendar choice. Although the results were not significant under a two-tailed test (t = .56), there was a positive trend (slope coefficient = 0.08688) toward higher proficiency participants selecting images corresponding with the prepositional uses implied by the prompt question. This trend was confirmed using a chi-square analysis. A median cut-off point for high versus low proficiency was established and the relationship between choosing an image congruent with the prompt question and proficiency level (cloze test) was tested. Although these results were not significant ($\chi^2 = 2.13$, df = 1, p = .14), of those participants who did select the congruent image, a greater number were in the high proficiency group (19) than in the low proficiency group (11).

General Discussion

The present study looked at the susceptibility of individuals to forming spatial image schematic perceptual simulations when thinking about non-spatial metaphorical expressions. Although prompted to consider time, native speakers activated schemas for spatial prepositions, as shown by their selection of calendar images congruent with either containment or support. On the contrary, Japanese-speaking learners of English as a foreign language tended not to associate the spatial and temporal meanings of prepositions in this manner.

At present a number of questions remain unanswered about the role of perceptual simulations in second language learning. We suggest a few avenues for research that we are currently investigating, and that would shed light on whether it is possible to use image-schemas to assist L2 learning. First, further research should address to what extent L2 proficiency matters in activating perceptual simulations. The Japanese group we tested was comprised of low-to-intermediate level of English learners, who had mainly received formal schooling in Japan and little or no genuine immersion. It is possible that a comparison between this group and a more proficient group of L2 speakers will reveal a significant difference in the learners' susceptibility to perceptual simulations.

Follow-up studies will test for the reverse direction of the effect obtained in Experiment 1. That is, can priming a particular perceptual simulation influence the type of phrase that a speaker chooses to produce? Previous work has shown that participants' use of in and on is influenced by a variety of factors in the scene, including animacy of the Figure and the Ground, as well as the function and degree of concavity of the Ground, (Feist & Gentner, 1998; 2003). Future experiments could explore this use of in and on for idiomatic expressions that only metaphorically involve a spatial relation. For example, participants could be presented a single picture of either the box-like calendar (Fig. 2 top-left) with the month of their birthday included in its top portion or the flat calendar (Fig. 1 top-left) with the day of their birthday included in its top portion. They would then be instructed to report their date [or, month] of birth, and the measure is whether they use the preposition in or on. This type of manipulation would provide evidence regarding the bi-directionality of influences between perceptual simulations and language processing – showing not only that concepts can potentiate sensorimotor primitives (Mahon & Caramazza, 2008), but also that sensorimotor primitives can potentiate concepts. This is particularly relevant for L2 learning situations where the exposure to the statistical patterns of a specific idiom are less robust. Under such circumstances, the choices between seemingly-acceptable prepositions and perceptual simulations are more open-ended, and thus malleable by one another. Conversely, the extensive statistical exposure of English L1 speakers is expected to entrench their perceptual simulations, resulting in a more uni-directional influence. Evidence of perceptual malleability in L2 learners would then set the stage for other manipulations investigating the learnability of prepositions in L2. It is possible that perceptual experience guides learning and use. In fact, Celce-Murcia and Larsen-Freeman (1999) have claimed that "anchoring the meaning of prepositions in spatial relationships is the first step to helping students learn to deal with areas where the meaning is more abstract" (1999, p. 405). Activating the spatial meanings of these prepositions may give rise to simulations of containment and support that serve to anchor the temporal senses of in and on, respectively. New experiments can thus be devoted to an explicit investigation of a novel visual-context-oriented method of learning prepositions in a second language. Experiments that follow from this line of research can make explicit comparisons and tests between alternative second language teaching methods (e.g., Brown, 2006; Cook, 1996; Long & Doughty, 2009; Nunan, 1999), as well as explore additional prepositions. While language educators have often made suggestions regarding the use of images and other cognitively appropriate stimuli in teaching English prepositions (e.g., Celce-Murcia & Larsen-Freeman, 1999), their suggestions often lack rigorous empirical grounding. In particular, it remains for studies to incorporate both the use of pedagogic tasks involving pictures and the insights available from experimental cognitive linguistics in a single series of laboratory studies.

The two experiments presented in this study are merely the first steps in this research project, but they suggest that assisting the learning of subtle idiomatic use of spatial prepositions by adding visual aids that correspond to the image schemas (or perceptual simulations) associated with those prepositions might be a viable solution to helping L2 learners use language in a more native-like manner. By encouraging learners to think about the perceptual simulations that match the prepositions being used, language teaching materials and methods can make some of the more subtle and seemingly-arbitrary properties of a second language become more accessible to learners. Finally, such image-schema-based training methods could also help assess the effectiveness of motor simulations in the treatment of language disorders. For example, Kemmerer (2005) reports that brain-damaged subjects with left perisylvian lesions failed a test of knowledge of the temporal meanings of prepositions, but passed a test that assessed knowledge of the corresponding spatial meanings of the same prepositions. Training regimes that activate particular motor simulations might help reestablish the spatio-temporal links of prepositional usages in brain-damaged patients.

In closing, the two experiments reported here suggest that L1 and L2 speakers pay differential attention to image properties when spatial language is employed to talk about time. The results seem to support a conclusion consistent with the idea that image schemas may underlie perceptual simulations recruited during language processing. These experiments add to evidence that shows how spatial language directs attention as a function of particular entities and the interaction between them (Coventry, et al., 2010), which points to an explanation grounded in both cognitive linguistic theory and embodied sensorimotor accounts. Findings from future research may offer insight into the role of language experience.

References

André, E., Bosch, G., Herzog, G., and Rist, T. (1987). Coping with the intrinsic and deictic uses of spatial prepositions. In Artificial Intelligence II, Proceedings of the Second International Conference on Artificial Intelligence: Methodology, Systems, Applications, Varna,

Bulgaria, eds. P. Jorrand and V. Sgurev, 375-382. Amsterdam: North Holland.

Barsalou, L.W. (1999). Perceptual Symbol Systems. Behavioral and Brain Sciences, 22 (4), 577-660.

Bergen, B., Lindsay, S., Matlock, T., & Narayanan, S. (2007). Spatial and linguistic aspects of visual imagery in sentence comprehension. Cognitive Science, 31, 733-764.

Boroditsky, L. (2000). Metaphoric Structuring: Understanding time through spatial metaphors. Cognition, 75(1), 1-28.

Boroditsky, L. & Ramscar, M.J.A. (2002). The roles of body and mind in abstract thought. Psychological Science, 13(2), 185-188

Bowerman, M. and Choi, S. (2003). Space under construction: Language specific spatial categorization in first language acquisition. In D. Gentner and S. Goldin-Meadow (Eds) Language in Mind: Advances in the study of Language and Cognition (pp. 387-428). Cambridge: MIT Press.

Brown, H. D. (2006). Principles of language learning and teaching, 5th edition. Pearson.

Brown, J. D. (1998). An EFL readability index. JALT Journal, 20, 7-36.

Brugman, C., & Lakoff, G. (1988). Cognitive topology and lexical networks. In S. Small, G. Cottrell & M. Tanenhaus (Eds.), Lexical Ambiguity Resolution. San Mateo, CA: Morgan Kaufmann Publishers, Inc.

Casasola, M., Bhagwat, J., Burke, A.S. (2009). Learning to Form a Spatial Category of Tight-Fit Relations: How Experience With a Label Can Give a Boost. Developmental Psychology, 45(3), 711-722.

Celce-Murcia, M., & Larsen-Freeman, D. (1999). The grammar book: An ESL/EFL teacher's course. Boston, MA: Heinle & Heinle.

Choi, S., & Bowerman, M. (1991). Learning to express motion events in English and Korean: the influence of language-specific lexicalization patterns. Cognition, 48, 83-121.

Cook, V. (1996). Second language learning and language teaching, 2nd edition. Hodder Arnold Pub.

Coventry, K. R., Lynott, D., Cangelosi, A., Monrouxe, L., Joyce, D., & Richardson, D. C. (2010). Spatial language, visual attention, and perceptual simulation. Brain & Language, 112, 202-213.

Coventry, K.R., & Garrod, S.C. (2004). Saying, seeing and acting. The psychological semantics of spatial prepositions. Psychology Press. Hove and New York.

Coventry, K. R., & Guijarro-Fuentes, P. (2008). Spatial language learning and the functional goemetric framework. In P. Robinson & N. C. Ellis (Eds), Handbook of cognitive linguistics and second language acquisition (pp. 114-138). New York: Routledge.

Feist, M. & Gentner, D. (1998). On plates, bowls, and dishes: Factors in the use of English IN and ON. Proceedings of the 20th Annual Meeting of the Cognitive Science Society.

Feist, M. & Gentner, D. (2003). Factors involved in the use of IN and ON. Proceedings of the 25th Annual Meeting of the Cognitive Science Society.

Gibbs, R.W. (2006). Embodiment and Cognitive Science. New York: Cambridge University Press.

Gibbs, R., & Colston, H. (1995). The cognitive psychological reality of image schemas and their transformations. Cognitive Linguistics, 6, 347-378.

Glenberg, A. M., & Kaschak, M. P. (2002). Grounding language in action. Psychonomic Bulletin & Review, 9, 558-565.

Haspelmath, M. (1997). From space to time: Temporal adverbials in the world's languages. (Lincom Studies in Theoretical Linguistics, 3.) Munich & Newcastle: Lincom Europa.

Kemmerer, D., (2005). The spatial and temporal meanings of English prepositions can be independently impaired. Neuropsychologia, 43, 797-806.

Lakoff, G. (1987). Women, fire, and dangerous things: what categories reveal about the mind. University of Chicago Press.

Langacker, R.W. (1987). Foundations of cognitive grammar: Theoretical Prerequisites. Stanford, CA: Stanford University Press.

Lindstromberg, S. (1998). English Prepositions Explained. John Benjamins.

Long, M. & Doughty, C. (2009). (Eds.). The handbook of language teaching. Wiley-Blackwell.

Mahon, B.Z., & Caramazza, A. (2008). A Critical Look at the Embodied Cognition Hypothesis and a New Proposal for Grounding Conceptual Content. Journal of Physiology, 102, 59-70.

Mandler, J.M. (1992). How to Build a Baby: II. Conceptual Primitives. Psychological Review, 4, 587-604.

Marian, V., Blumenfeld, H.K., & Kaushanskaya, M. (2007). The Language Experience and Proficiency Questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals. Journal of Speech, Language, and Hearing Research, 50, 1-28.

Nunan, D. (1999). Second language teaching and learning. Heinle & Heinle.

Retz-Schmidt, G. (1988). Various views on spatial prepositions AI Magazine, 9, 95-105.

Richardson, D., Spivey, M., Barsalou, L., & McRae, K. (2003). Spatial representations activated during real-time comprehension of verbs. Cognitive Science, 27(5), 767-780.

Talmy, L. (1983). How Language Structures Space. In H. Pick & L. Acredolo (eds.) Spatial Orientation: Theory, Research, and Application. New York, Plenum.

Tyler, A., & Evans, V. (2003). The semantics of English prepositions: Spatial scenes, embodied meaning and cognition. Cambridge University Press.

Zwaan, R.A. (2004). The immersed experiencer: toward an embodied theory of language comprehension. In: B.H. Ross (Ed.), The Psychology of Learning and Motivation, Vol. 44 (pp. 35-62). New York: Academic Press.

A Motion Aftereffect from Literal and Metaphorical Motion Language: Individual Differences

Alexia Toskos Dils (atoskos@stanford.edu)
Lera Boroditsky (lera@stanford.edu)
Stanford University, Department of Psychology
Jordan Hall, 450 Serra Mall, Building 420, Stanford, CA 94305 USA

Abstract

Do people spontaneously form visual mental images when understanding language, and if so how truly visual are these representations? We test whether processing linguistic descriptions of motion produces sufficiently vivid mental images to cause direction-selective motion adaptation in the visual system (i.e., cause a motion aftereffect illusion). We tested for motion aftereffects (MAEs) following explicit motion imagery, and after processing literal or metaphorical motion language. Intentionally imagining motion produces an aftereffect in the overall sample with some participants showing a greater aftereffect than others. We then find that participants who show the strongest imagined motion aftereffects also show aftereffects in the natural course of processing motion language (without instructions to imagine). Individuals who do not show strong motion aftereffects as a result of imagining motion also do not show them from processing motion language. However, the aftereffect from language gained strength as people were exposed to more and more of a motion story. For the last two story installments (out of 4), understanding motion language produced reliable MAEs across the entire sample. The results demonstrate that processing language can spontaneously create sufficiently vivid mental images to produce direction-selective adaptation in the visual system. The timecourse of adaptation suggests that individuals may differ in how efficiently they recruit visual mechanisms in the service of language understanding. Further, the results reveal an intriguing link between the vividness of mental imagery and the nature of the processes and representations involved in language understanding.

Keywords: embodiment, language comprehension, perception, motion aftereffect, individual differences

Introduction

A good story can draw you in, conjure up a rich visual world, give you goose-bumps, or even make you feel like you were really there. To what extent is hearing a story about something similar to really witnessing it? What is the nature of the representations that arise in the course of normal language processing? Do people spontaneously form visual mental images when understanding language, and if so how truly visual are these representations? In this paper we make use of the motion aftereffect illusion to test whether processing linguistic descriptions of motion produces sufficiently vivid mental images to cause direction-selective adaptation in the visual system (i.e., cause a motion aftereffect).

A number of findings suggest that people do spontaneously engage in imagery during language comprehension and that processing language affects performance in subsequent perceptual tasks (e.g., Bergen, Lindsay, Matlock, & Narayanan, 2007; Meteyard, Bahrami, & Vigliocco, 2007; Richardson, Spivey, Barsalou, & McRae, 2003; Rinck & Bower, 2000; Rinck, Hähnel, Bower, & Glowalla, 1997; Spivey & Geng, 2001; Stanfield & Zwaan, 2001; Zwaan, Madden, Yaxley, & Aveyard, 2004; Zwaan, Stanfield, & Yaxley, 2002;).

What mechanism might underlie these interactions between linguistic processing and perception? The explanation frequently offered is that the representations generated during the course of language comprehension share processing resources with perception, recruiting some of the very same brain regions (Barsalou, 1999). As evidence for this possibility fMRI measures have revealed that classically 'perceptual' brain areas are recruited in service of language comprehension (e.g., Saygin, McCullough, Alac, & Emmorey, 2010). While these findings are consistent with the hypothesis, questions remain. The spatial resolution of current fMRI technology is coarse. A typical voxel (the smallest unit of measurement) may include 100,000 neurons. It is possible then that what appear in fMRI to be the same regions activated in linguistic and visual tasks are in fact neighboring (or closely interleaved) but distinct neural populations, potentially with quite different computational properties.

One powerful paradigm for determining whether neural populations involved in particular tasks indeed overlap is that of adaptation. In this paper, we make use of one such adaptation measure, the motion aftereffect (MAE). The MAE arises when direction-selective neurons in the human MT+ complex lower their firing rate as a function of adapting to motion in their preferred direction. The net difference in the firing rate of neurons selective for the direction of the adapting stimulus relative to those selective for the opposite direction of motion produces a motion illusion. For example, after adapting to upward motion, people are more likely to see a stationary stimulus or a field of randomly moving dots as moving downward, and vice versa (e.g., Blake & Hiris, 1993). To quantify the size of the aftereffect, one can parametrically vary the degree of motion coherence in the test display of moving dots (as in Blake & Hiris, 1993). The amount of coherence necessary to null the MAE (i.e. to make people equally likely to report the motion as upward or downward) provides a nice measure of the size of the aftereffect produced by the adapting stimulus.

Winawer, Huk, and Boroditsky (2008, 2010) adapted this technique to test for MAEs after participants either viewed still images implying motion (e.g., a runner in mid-leap), or simply imagined motion without any visual stimulus. Both implied and purely imagined motion produced reliable MAEs. These studies support fMRI findings suggesting the hMT+ complex is recruited in the service of mental imagery (Goebel, Khorram-Sefat, Muckli, Hacker, & Singer, 1998; Grossman & Blake, 2001), and further suggest that this activation is driven by direction-selective neurons.

Here we explore whether natural language comprehension can likewise produce MAEs. To the extent that people spontaneously engage in imagery in service of language comprehension, understanding motion language should yield MAEs (albeit likely weaker than those produced during explicit, effortful imagery). The present study was designed to test this prediction. Participants listened to stories describing motion in a particular direction and then judged the direction of a moving field of dots. The direction in which motion language affects subsequent motion perception speaks to the mechanisms underlying language comprehension. One possibility is that motion language adapts the same direction-selective mechanisms that subserve motion perception; this would cause people to see a real visual stimulus (e.g., dynamic dots) as moving in a direction *opposite* to that described in the adapting language. Another possibility is that understanding motion language recruits higher-level convergence areas that process visual motion, resulting in a bias to see dot motion in the *same* direction. Such a congruence effect is reported by Sadaghiani et al (2009) who showed that hearing the words 'right' and 'left' biased participants to see an apparent motion stimulus as moving in the same direction. fMRI data revealed that this audiovisual interaction was driven more by activity in the anterior intraparietal sulcus (IPS) than hMT+. A third possibility is of course that motion language does not recruit visual motion processing resources of any kind, resulting in no bias in dot motion perception.

Further, the direction and extent of transfer from language to perception may depend on an individual's visual motion imagery ability. People differ from one another in mental imagery ability, and these differences correlate with individual differences in spatial tasks and object perception (Kozhevnikov, Kosslyn, & Shephard, 2005). In Winawer et al (2010), most but not all participants showed MAEs as a function of imagining motion, and the degree of adaptation differed across people. We reasoned that people who show stronger adaptation as a result of imagining, should be more likely to show adaptation as a result of understanding motion language. It would be reasonable to expect that individuals who do not show an MAE as a result of explicitly imagining motion should also not show one as a result of processing motion language. To test for this possibility, we tested each participant both in an explicit visual imagery condition (as in Winawer et al (2010)), and in conditions where linguistic motion was used as an adapting stimulus. This allowed us to compare the effects of language for each participant with those of explicit imagery.

Finally, the present study is designed to test whether literal and metaphorical descriptions of motion recruit similar perceptual processes. To this end, we contrasted literal motion stories that described the motion of physical objects with metaphorical motion stories that used motion verbs to talk about changes in abstract entities (e.g. rising and falling stock prices).

Experiment

The experiment consisted of five parts: (1) a baseline task in which we measured participants' motion direction sensitivity, (2) a familiarization task in which participants viewed the stimuli to be imagined later in the study (3) the main experimental task in which we tested for MAEs following imagining motion or listening to stories describing motion, (4) a memory task in which we measured participants' recognition memory for the stories, and (5) an exit questionnaire in which we ascertained participants knowledge of the motion aftereffect and their explicit predictions about the direction of effects.

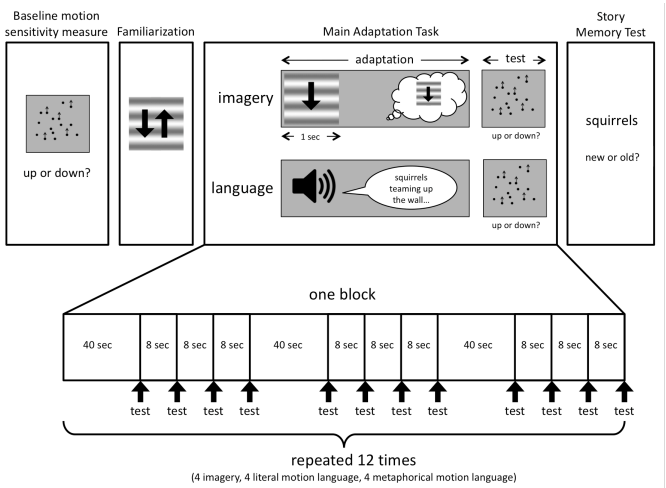


Figure 1: Schematic of experimental design highlighting the block and trial structure of the main adaptation task. In the imagery blocks, an upward or downward facing arrow superimposed on a static image of the grating indicated the direction in which to imagine the stripes moving. This cue faded slowly over the course of a second. Once the cue disappeared completely, a flickering fixation cross appeared at the center of the screen. Participants were instructed to fixate on the cross while imagining the stripes and to use the rate of the flicker to help them remember how fast the stripes should move. Participants were also instructed to use the fixation cross as a cue for when to start and stop imagining motion. In language blocks, participants listened to stories using headphones while fixating a dot centered on the monitor. Participants were told to listen carefully to the stories, as there would be a memory test. They were not instructed to imagine.

Methods

Participants Sixty Stanford students participated in exchange for payment.

Stimuli and Procedure

Main Experimental Task: The task design, procedure, and visual stimuli used were modeled on those used in Winawer et al (2010). On each trial participants judged the direction of dot motion after either listening to stories describing motion or engaging in explicit visual motion imagery. Trials were presented in 12 interleaved blocks. There were 6 block types, 3(motion type: imagined motion, literal motion, or metaphorical motion) by 2(motion direction: upward or downward).

Adaptation Stimuli: In the literal motion condition the stories used motion language to describe the movement of physical objects (e.g., squirrels, ping-pong balls). In the metaphorical motion condition, the stories used motion language to describe changes in abstract entities (e.g. stock prices, emotions). 12 literal and 12 metaphorical stories were used with an upward and a downward version for each, yielding a total of 48 stories. Individual participants heard 24 stories (either the upward or the downward version of each story, but not both). Example stories are in Table 1. In the imagery condition, participants were instructed to imagine upward and downward moving gratings (as in Winawer et al. (2010)). The trial structure for the language and imagery conditions is depicted in Figure 1.

Table 1: Sample stories heard by participants.

Literal Motion: Upward Story (Four installments)
1. You are running a psychology experiment in which you have trained hundreds of squirrels to race each other up a wall for a piece of food. Now you want to see what happens when they are all released at the foot of the wall at once. You watch through a small window in the next room as the cages are opened and the squirrels leap onto the wall in a frenzy. The little fur balls scurry up the wall in one relentless stream, despite obvious defeat in the race. Zip! The brown creatures surge up the wall with amazing agility. You see the same behavior in squirrel after squirrel – one swift jump onto the wall and an instantaneous burst upward. Zoom! The squirrels rush up the wall like a giant current. As if in a trance, the squirrels swiftly stream past your eyes in their race for the top of the wall.

2. Zoom! More and more squirrels jump onto the wall and scurry upwards. You watch them course up the wall in a blur.

3. The squirrels continue to sprint upwards in a flash. They spout onto the wall and surge directly toward the top.

4. Your eyes remain focused on the mob of squirrels teeming up the wall. You can no longer pick out individuals as they dash for the top.

Literal Motion: Downward Story (Four installments)
1. You are running a psychology experiment in which you have trained hundreds of squirrels to race each other down a wall for a piece of food. Now you want to see what happens when they are all released at the top of the wall at once. You watch through a small window in the next room as the cages are opened and the squirrels descend onto the wall in a frenzy. The little fur balls scurry down the wall in one relentless stream, despite obvious defeat in the race. Zip! The brown creatures surge down the wall with amazing agility. You see the same behavior in squirrel after squirrel – one swift drop onto the wall and an instantaneous burst downward. Zoom! The squirrels rush down the wall like a giant current. As if in a trance, the squirrels swiftly stream past your eyes in their race for the bottom of the wall.

2. Zoom! More and more squirrels drop onto the wall and scurry downwards. You watch them course down the wall in a blur.

3. The squirrels continue to sprint downwards in a flash. They pour onto the wall and surge directly toward the bottom.

4. Your eyes remain focused on the mob of squirrels teeming down the wall. You can no longer pick out individuals as they dash for the bottom.

Metaphorical Motion: Upward Story (Four installments)
1. You are standing in the middle of the trading floor at the New York stock exchange one busy morning. The room is buzzing with announcements of rising stock prices. First JP Morgan rockets dramatically. Accenture and Delaware blaze to new heights. Suddenly, Lincoln's stock surges, along with Time Warner. You hear animated reports of Toyota, Coca Cola, and The Gap going sky-high! You can hardly believe it, but Google's stock soars higher than ever. Walmart zips skyward, too. All morning, you marvel at the continually spiking stocks!

2. You hear that Ford and Exxon Mobile are really ramping up. Hewlett Packard is erupting too!

3. Next you hear that Nokia is boosting quickly. Likewise, Sprint, AT&T and Verizon are surging dramatically.

4. Stock prices heighten rapidly for Proctor and Gamble as well as Clorox. McDonalds' stock also jets to new heights!

Metaphorical Motion: Downward Story (Four installments)
1. You are standing in the middle of the trading floor at the New York stock exchange one busy morning. The room is buzzing with announcements of falling stock prices. First JP Morgan plummets dramatically. Accenture and Delaware tumble to new lows. Suddenly, Lincoln's stock plunges, along with Time Warner. You hear agitated reports of Toyota, Coca Cola, and The Gap hitting record lows! You can hardly believe it, but Google's stock sinks lower than ever. Walmart zips downward, too. All morning, you marvel at the continually diving stocks!

2. You hear that Ford and Exxon Mobil are really sinking down. Hewlett Packard is taking a nose-dive too!

3. Next you hear that Nokia is slumping quickly. Likewise, Sprint, AT&T and Verizon are tumbling dramatically.

4. Stock prices level rapidly for Proctor and Gamble as well as Clorox. McDonalds' stock also plunges to new lows!

Block structure: In the two language conditions, each block consisted of 3 stories with 4 installments each, for a total of 12 trials per block. Each story was broken up into one longer paragraph and three shorter 'top-up' installments so that multiple measurements could be collected for each story. The longer installments lasted on average 40.00 seconds, and the top-up installments 8.29 seconds. The imagery blocks mirrored this structure. Participants imagined motion for 40 seconds, and on the three subsequent 'top-up' trials, participants imagined motion for 8 seconds. This pattern was repeated 2 more times within the block to parallel the 3 stories used per block in the language conditions.

Adaptation Test: Following each story or imagery installment, participants judged the direction of motion coherence in a field of moving dots without feedback. The moving dot stimuli were presented as in Winawer et al. (2010). Each dot display had net motion coherence either up or down. For each subject, two coherence values were sampled: 12.5% and 25% of the coherence necessary for asymptotic performance (as assessed individually for participants in the baseline task). Coherence and direction of motion were fully crossed and balanced across trials and participants.

Exit questionnaire: At the end of the experiment we ascertained participants' familiarity with the motion aftereffect and also asked them to generate a prediction

about which way they thought the effect would go. Participants were asked: *Have you ever heard of the Motion Aftereffect or Waterfall Illusion?* and *After viewing upward motion, which way would you expect a static image to appear to move?*

Results

The distance between the null points of the logistic fits for upward and downward motion (normalized coherence values at which participants are equally likely to report upward and downward motion) was computed for both the imagined and linguistic motion conditions for each participant. Positive values reflect adaptation. Six participants whose results exceeded three standard deviations from the mean for all participants were excluded from subsequent analyses. The literal and metaphorical linguistic motion conditions did not significantly differ from one another ($t(53) = 0.219$, $p > .5$), and so were combined for analysis. Results are plotted in Figures 2-4.

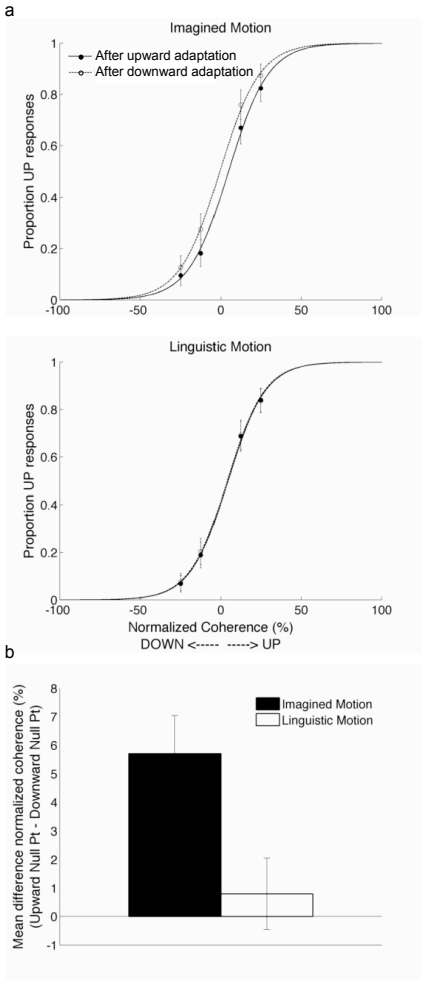


Figure 2: (a) Proportion of "UP" responses following imagined motion and linguistic motion across all participants. Error bars represent standard error. (b) Separation in motion response functions for imagined and linguistic motion across all participants. Positive values reflect adaptation. Error bars denote s.e.m.

In the overall sample, participants showed a reliable MAE after imagining motion (M = 5.7% normalized coherence, SD = 9.8%) ($F(1,53) = 18.26$, $p < .001$) (replicating Winawer et al, 2010), but not after listening to motion stories (M = 0.8% normalized coherence, SD = 9.2%) ($F(1,53) = 0.40$, $p > .5$). The two conditions differed reliably from one another ($F(1,53) = 10.81$, $p < .005$).

We reasoned that individuals who do not show MAEs as a result of explicitly imagining motion should also not show them as a result of processing motion language. However, participants who do show MAEs from motion imagery may show them from processing motion language as well. Indeed, there was a significant correlation between the effects of motion imagery and motion language ($r(52) = .34$, $p < .02$), such that stronger adaptation from imagining motion predicted stronger adaptation from understanding motion language (Figure 3).

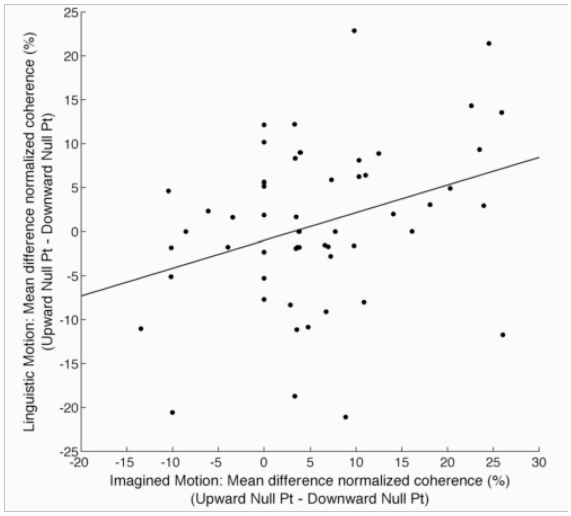


Figure 3: Correlation across all participants between the separation in motion response functions for imagined and linguistic motion, $r(52) = .34$, $p < .02$.

To confirm that participants who showed adaptation to imagined motion also showed it in response to linguistic motion, we sorted participants based on the magnitude and sign of the effect of explicit motion imagery and divided them into three groups of equal size (Imagery *Mdns* = 15.1%, 3.8%, -1.7%, and *SIQRs* = 6.6%, 1.6%, 5.0% normalized coherence) (Figure 4). We will refer to these as strong, weak, and no MAE groups respectively.

Indeed, the group that showed strong MAEs after explicitly imagining motion also showed reliable MAEs after listening to motion language (Language *Mdn* = 5.6%, *SIQR* = 4.7%) (n = 18, $p < .031$, sign-test, 2-tailed). There was no difference in the strength of this adaptation effect between the literal and metaphorical language conditions, n = 18, $p > .40$. The two groups that showed weak or no MAEs from imagery, did not show reliable MAEs from language: (*Mdn* = -1.7%, *SIQR* = 5.0%) (n = 18, $p > .05$), and (*Mdn* = 0.8%, *SIQR* = 5.1%) (n = 18, p > .5) for groups that showed weak or no MAEs respectively. The effects of

language in the strongest MAE group differed reliably from the other two groups, $\chi^2(1, N=54)=7.27$, $p<.01$.

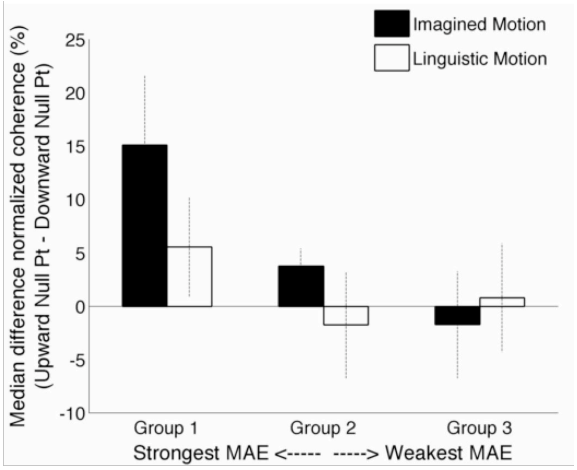


Figure 4: Participants were sorted based on the size of the aftereffect in the imagery condition and divided into three equal-sized groups. The plot shows the median separation between motion response functions for each group. Error bars denote SIQR.

To examine the timecourse of the MAE from imagined and linguistic motion, we subtracted the proportion of "up" responses following upward motion from those following downward motion across adaptation installments (e.g., the 4 installments of a story, or the analogous 4 imagery installments). The mean difference by installment across all participants is plotted in Figure 5. In the explicit imagery trials, the MAE appears after the initial 40-second installment of imagining (as would the MAE from real visual motion), and participants remain adapted for subsequent installments (there is no linear effect of installment, $F(1,53) = 0.076$, $p > .5$). In the two language conditions, however, the MAE does not emerge until later installments (there is a reliable linear effect of installment, $F(1,53) = 6.59$, $p < .05$). After the 3rd and 4th story installment, there is a reliable motion aftereffect including all participants, M=4.0%, SD = 12.7%; $F(1,53) = 5.42$, $p < .05$. Motion language appears to produce a reliable MAE across the entire sample only after sufficient exposure to each story.

These findings raise the possibility that individual differences in the MAE from linguistic motion reflect differences in how efficiently people recruit visual direction-selective mechanisms rather than qualitative differences in which mechanisms are recruited. Indeed, the linear effect of story installment does not differ among those who show strong, weak, and no MAEs from motion imagery ($F(2,51)=.144$, $p>.5$), with everyone showing the same trend toward more adaptation as they get further into the story.

Testing for effects of explicit bias: Of the 54 participants included in the analysis, 43 completed an exit questionnaire about their knowledge and predictions about the motion aftereffect (the remaining 11 omitted this portion of the study). Only three reported having heard of the motion aftereffect. Participants' expectations about the direction in which adapting to visual motion in one direction might affect subsequent visual processing did not reliably bias ($F(1,39) = 0.37$, $p>.50$) or interact with ($F(1,39) = 0.33$, $p>.50$) the effects of imagined and linguistic motion. This finding confirms that the results obtained in this study are not a product of participants' expectations or explicit biases regarding the direction of the effects.

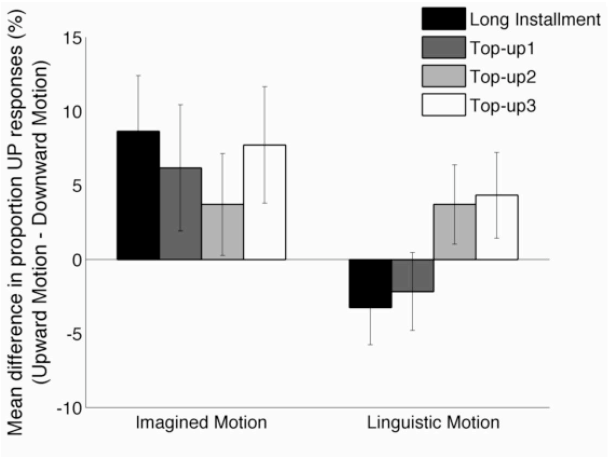


Figure 5: Mean difference in proportion upward responses following upward and downward motion across the four motion installments. The data are plotted for the overall sample. Positive values reflect adaptation, and error bars denote s.e.m.

Discussion

We tested whether processing linguistic descriptions of motion produces sufficiently vivid mental images to cause direction-selective motion adaptation in the visual system (i.e., cause a motion aftereffect illusion). We predicted that the perceptual consequences of processing language should depend on an individual's mental imagery ability. Imagery ability was operationalized as the extent to which explicit visual motion imagery produced an MAE in each participant. Put another way, imagery ability or vividness is the extent to which people recruit perceptual resources heavily enough to adapt them during explicit imagery.

We replicated previous work showing that intentionally imagining motion produces an aftereffect. We then found that participants who show the imagined motion aftereffect most strongly also show this aftereffect in the natural course of processing motion language (without instructions to imagine). The same effects held for both literal and metaphorical language. Individuals who did not show a motion aftereffect as a result of imagining motion also did not show an aftereffect from processing motion language overall. However, the aftereffect from language gained strength with the number of story installments. For the last two installments (out of 4), understanding motion language

produced reliable MAEs across the entire sample. This finding suggests the possibility that individuals may differ in how efficiently they recruit visual mechanisms in service of language comprehension. Future work will examine the effects of systematically varying exposure to motion language and the degree of story immersion on the MAE. Participants' knowledge of the MAE and their explicit predictions about the direction that the MAE should go did not predict their pattern of results. This helps us ensure that the patterns observed were not simply due to participants' explicit biases or expectations.

A further question concerns the effects from metaphorical motion language. Some researchers have found that literal and metaphorical language produce similar transfer effects to perceptuo-motor tasks (e.g., Boulenger, Hauk, & Pulvermüller, 2009; Glenberg & Kaschak, 2002; Richardson et al., 2003), while others have found no evidence for transfer from metaphorical language (Bergen et al., 2007). In our study, literal and metaphorical motion language produced the same effects. Our stimuli and methods differ from previous studies in many ways. One potentially important difference is that our stimuli were connected narratives that built over time, whereas the studies just cited used isolated sentences. Our results suggest that for language processing to produce effects on low-level visual processing, a greater amount of exposure to or immersion in a connected narrative may be necessary.

The results of the present study demonstrate that at least for a subset of the population, processing language spontaneously creates sufficiently vivid mental images to produce direction-selective adaptation in the visual system. Future work will examine the source and possible cognitive consequences of the individual differences we observed. Why might some people be better able to recruit or effectively modulate the activity of sensory neurons through top-down processes? Further, are there resulting systematic differences in the content and nature of representations people form in the service of understanding language?

Acknowledgments

We thank Jonathan Winawer and members of Cognation for their helpful feedback. This research was supported by an NSF Career Award Grant given to Lera Boroditsky.

References

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences, 22,* 577-660.

Bergen, B. K., Lindsay, S., Matlock, T., & Narayanan, S. (2007). Spatial and linguistic aspects of visual imagery in sentence comprehension. *Cognitive Science, 31,* 733-764.

Blake, R., & Hiris, E. (1993). Another means for measuring the motion aftereffect. *Vision Research, 33*, 1589-1592.

Boulenger, V., Hauk, O., & Pulvermüller, F. (2009). Grasping ideas with the motor system: semantic somatotopy in idiom comprehension. *Cerebral Cortex, 19,* 1905-1914.

Glenberg, A. M., & Kaschak, M.P. (2002). Grounding language in action. *Psychonomic Bulletin & Review, 9,* 558-565.

Goebel, R., Khorram-Sefat, D., Muckli, L., Hacker, H., & Singer, W. (1998). The constructive nature of vision: Direct evidence from functional magnetic resonance imaging studies of apparent motion and motion imagery. European Journal of Neuroscience, 10(5), 1563–1573.

Grossman, E. D., & Blake, R. (2001). Brain activity evoked by inverted and imagined biological motion. Vision Research, 41(10–11), 1475–1482.

Kozhevnikov, M., Kosslyn, S., & Shephard, J. (2005). Spatial versus object visualizers: A new characterization of visual cognitive style. *Memory and Cognition, 33*, 710-726.

Meteyard, L., Bahrami, B., & Vigliocco, G. (2007). Motion detection and motion verbs. *Psychological Science*, *18,* 1007-1013.

Richardson, D. C., Spivey, M. J., Barsalou, L. W., & McRae, K. (2003). Spatial representations activated during real-time comprehension of verbs. *Cognitive Science, 27,* 767–780.

Rinck, M., & Bower, G., H. (2000). Temporal and spatial distance in situation models. *Memory and Cogntion, 28,* 1310-1320.

Rinck, M., Hähnel, A., Bower, G., & Glowalla, U. (1997). *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*, 622-637.

Sadaghiani, S., Maier, J. X., & Noppeney, U. (2009). Natural, metaphoric, and linguistic auditory direction signals have distinct influences on visual motion processing. *Journal of Neuroscience, 29*, 6490 – 6499.

Saygin, A.P., McCullough, S., Alac, M., Emmorey, K. (2009). Modulation of BOLD response in motion sensitive lateral temporal cortex by real and fictive motion sentences. *Journal of Cognitive Neuroscience.* Early Access publication on Nov. 19, 2009, doi:10.1162/jocn.2009.21388.

Spivey, M., & Geng, J. (2001). Oculomotor mechanisms activated by imagery and memory: Eye movements to absent objects. *Psychological Research, 65,* 235-241.

Stanfield, R., & Zwaan, R. (2001). The effect of implied orientation derived from verbal context on picture recognition. *Psychological Science, 12,* 153-156.

Winawer J, Huk A, Boroditsky L. (2008) A motion aftereffect from viewing still photographs depicting motion. *Psychological Science, 19*, 276-283.

Winawer J, Huk A, Boroditsky L. (2010). A motion aftereffect from visual imagery of motion. *Cognition, 114*, 276-284.

Zwaan, R. A., Madden, C. J., Yaxley, R. H., & Aveyard, M. E. (2004). Moving words: dynamic representations in language comprehension. *Cognitive Science, 28,* 611-619.

Zwaan, R., Stanfield, R., & Yaxley, R. (2002). Language comprehenders mentally represent the shapes of objects. *Psychological Science, 13,* 168-171.

Language-Driven Motor Simulation is Sensitive to Social Context

Heeyeon Y. Dennison (yoon@hawaii.edu)
Department of Linguistics, 569 Moore Hall, 1890 East West Road,
Honolulu, HI 96822 USA

Benjamin K. Bergen (bkbergen@ucsd.edu)
Department of Cognitive Science, University of California, San Diego
9500 Gilman Dr., La Jolla, CA 92093-0515

Abstract

People use their bodies differently in different social situations. In Korea, Japan, and Thailand, for example, there are culture-specific conventions for how to transfer small objects. People use one hand to transfer objects to people of equal or lower social status, but two hands with people of higher status. But does individual knowledge of these conventions for how to use one's body extend to other aspects of cognition? For instance, it is known that understanding action language involves internally simulating what it would be like to perform described actions. Do people mentally simulate actions appropriate to the social context described in a sentence? We report on a behavioral experiment, conducted with people born and raised in Korea, that investigated whether cultural practices affect the actions that people represent during language comprehension. We report evidence that motor simulations do indeed reflect social constraints on action.

Keywords: sentence processing; mental simulation; motor simulation; embodiment; social cognition; culture.

Introduction

What are the cognitive processes involved in understanding the meaning of a sentence that you hear or read? Behavioral and brain imaging evidence over the past several decades has revealed that – among other mechanisms – one process that is engaged routinely and mostly unconsciously is mental simulation (Barsalou, 1999). When you hear or read a sentence about an event, you use your visual system to simulate what the mentioned entities would look like: how they would move (Kaschak et al, 2005), what color (Connell, 2007), orientation (Stanfield & Zwaan, 2001, Zwaan et al., 2004) and shape (Zwaan et al., 2002) they would have, and so on. Similarly, when a sentence describes actions, you engage mental simulations of the described actions (Glenberg & Kaschak, 2002; Tettamanti et al., 2005; Buccino et al., 2005; Kaschak & Borreggine, 2008; Bub et al., 2008; Bergen & Wheeler, In press, Bergen et al., In press).

Key evidence that understanding motor language engages motor routines comes from the so-called "Action-sentence Compatibility Effect" (Glenberg & Kaschak, 2002): manual responses to make sentence sensibility judgments are facilitated when the motion of the physical response is compatible with the sentence's implied direction. For instance, *You handed the puppy to Katie* speeds manual responses away from the body, while *Katie handed the puppy to you* speeds movements toward to the body.

Research in this area has shown that implied features of a described scene – even when only implicit – show up in the comprehender's mental simulation. For example, people simulate handshapes specifically afforded by the objects mentioned. For instance, *Mary caught the marble* primes a grasping handshape, while *Mary caught the watermelon* primes an open palm handshape (Bergen & Wheeler, 2005). However, it is unknown at present whether language-driven motor simulation is sensitive to not only physical but also social constraints on action.

Part of acculturation is for people to learn culture-specific prescriptions for motor action (Mauss, 1934). For instance, in Korean culture, people learn to use both hands when giving an object to someone of higher social status, but only one hand with peers or social inferiors. This is obviously a learned behavior – many other cultures around the world do not share this particular convention. What's more, it's a cultural action convention that's specific to social context. In order to transfer an object appropriately to someone, you have to determine what your relative social status is, so as to engage an action using the right number of hands.

In the study described below, we ask whether socially-contingent prescriptions for motor action reach into other aspects of cognition as well, in particular, language comprehension. Does hearing about an action performed in a particular social situation elicit simulation of the prescribed physical behavior that conforms to the cultural constraint? More specifically, do Koreans who hear sentences about object transfer simulate using one hand or two hands, depending on the relative social status of the mentioned recipient? If they do, this would suggest that the motor simulations are flexibly tailored to the social situations comprehenders would encounter in the described situations. In addition, it would suggest that motor simulations have features specific to culture-specific constraints on action.

Method

The logic of the experiment was relatively simple. Participants listened to sentences in Korean about transferring small objects to recipients who were of either high status or low status. After the end of each sentence, they made a meaningfulness judgment about the sentence, which required them to press buttons either with two hands

or just with their right hand. We predicted that if the participants were automatically engaging motor representations of actions that used either one hand or two hands (appropriate to the social status of the mentioned recipient) then the status of the recipient and the number of hands they had to use to respond should show an interaction in their effect on measured response times.

Participants

Thirty-two native Korean speakers at the University of Hawai'i participated in this experiment and received $5 in compensation. They all had been born and raised in Korea but moved to the U.S. for their higher education. All but 2 people had lived in Korea for at least 20 years. Their mean length of residency in Korea was 24 years (range: 13.5–35, std: 4.6), while the mean of their age was 27.8 (range: 20–45, std: 6.38). All had normal or corrected-to normal vision and hearing. All but one were right-handed.

Materials

We constructed twenty pairs of critical sentences in Korean, all of which are meaningful and describe transfer of a small object. The object in all these sentences is conventionally transferred to people of higher status using two hands and to people of equal or lower status using one hand. Examples of high status and low status sentences are in (1) and (2), below.

(1) [High Status]
Ne-nun cikum kyoswu-nim-kkey phyenci-lul tuli-koisse.
'You are now (humbly) giving a letter to (your) professor.'

(2) [Low Status]
Ne-nun cikum tongsayng-hanthey phyenci-lul cwu-koisse.
'You are now giving a letter to (your) younger sibling.'

The sentences in each pair (like (1) and (2)) differed in two ways. First, while High-status sentences mentioned people of conventionally high status as recipients (such as professors, doctors, lawyers, etc.), low status sentences had recipients who were of lower status or close peers, (like younger siblings, nieces, friends, etc.). Second, the sentences about transfer to people of high status were accompanied by grammatical/lexical markers called *honorifics*, which indicated that the status of the recipient is higher than the status of the subject (*you*). These markers are present on the recipient noun, the dative case marker, and the verb, and as confirmed in a norming study, this is the most natural and proper way to describe an object transfer to a social superior in Korean. (We'll discuss these honorific markers and their implications in more detail in the Discussion section below.) All critical sentences had *you* as the subject, were in the present tense, and used active dative sentence structure.

Beyond the critical stimuli, we created some additional materials. In order to disguise the intent of the experiment, we prepared twenty meaningful filler sentences that were not about transfer of a small object. Because participants were performing a forced choice task, we needed forty non-sensible fillers to balance the twenty critical and twenty meaningful filler sentences. By including these, we ensured that each participant was expected to respond "Yes" half of the time overall. So that participants could not learn to use syntactic properties of the sentences to make judgments, half of the non-meaningful sentences were dative sentences, while the other half included sentences varying in structure and length – just like the meaningful sentences. Also, orthogonally, approximately half of all items mentioned high-status people in somewhere in the sentences, whereas the other half mentioned equal- or lower-status people.

Design, Procedure, and Predictions

Participants performed a sentence meaningfulness judgment task. They were seated in front of a computer with two keyboards aligned side-by-side in front of them (as in Figure 1 below). They keyboards were oriented such that the long axis of the keyboards projected out directly in front of the participant, at their midline. Participants first pressed the yellow keys with their two thumbs to begin auditory presentation of a sentence – these are at the bottom of the image in Figure 1, and were on the edge of the keyboard closest to the participant. Once they decided whether the sentence made sense or not, they released the yellow buttons to press either the "Yes" or "No" buttons on the keyboard to indicate their decision. The response buttons were positioned to require the participant to use either both hands (the green buttons) or only their right hand (the pink buttons) to press the buttons. Participants were instructed to press the pink buttons with two fingers of their right hand and the green buttons with the index fingers of their two hands. We measured the Reaching Time – the time from the yellow-button release until the "Yes" or "No" button press, because we were interested in seeing if simulation of the socially expected action would influence subsequent motion execution.

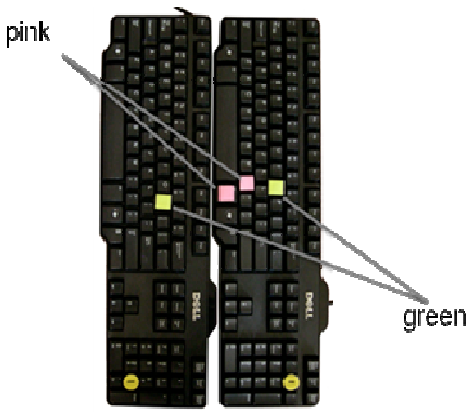


Figure 1. Configuration of keyboards to collect bimanual (green buttons) and unimanual (pink buttons) responses.

Participants pressed down the yellow buttons with the thumbs of their two hands to initiate presentation of a sentence, then pressed either the pink or green buttons to indicate their meaningfulness judgment.

We assigned participants to one of two starting conditions – they began with two-handed responses (the green buttons) meaning either "Meaningful" or "Non-meaningful" (and the pink buttons assigned complementarily). We switched the key assignments for each participant halfway through the experiment. Ten training trials preceded each half. Each participants heard the high-status recipient version of a randomly selected half of the critical sentences (e.g., (1)) and the low-status version of the other half (e.g., (2)). Each session lasted less than 20 minutes.

If native Korean speakers engage motor representations of two-handed actions when processing sentences about object transfer to people of high social status, and one-handed actions when processing sentences about transfer to people of equal or low status, then we should observe an interaction of Hand-Number by Sentence-Type.

But the direction of this interaction effect between language and action is a more difficult matter. Both match advantages (i.e., faster responses in the matching conditions) and mismatch advantages (i.e., slower responses in the matching conditions) have been reported in the literature. A close reading, however, leads us to expect a *mismatch* advantage; that is, two-handed responses will be faster when participants have just heard a sentence about transferring an object to someone of lower status, and conversely, one-hand responses should be faster when the preceding sentence describes transfer of an object to someone of higher status.

We are led to predict this mismatch advantage, rather than a perhaps more intuitive match advantage, for the following reason. Priming effects of action language on motor control are quite sensitive to timing. When there is a delay (more than 500 milliseconds) between the word that denotes the action and the action itself, language about actions facilitates motor actions that have broadly similar characteristics, such as the direction of motion (Glenberg & Kaschak, 2002; Kaschak & Borreggine, 2008; Bergen & Wheeler, In press) or handshape (Bergen & Wheeler, 2005; Bub et al., 2008). However, when the critical action word and the motor action are temporally aligned (within 500 ms), actions that are similar but not exactly the same will in fact be inhibited (Bergen, 2007; Bergen et al., In press, see also the review in Kaschak et al., 2005). That is, there is a *mismatch* advantage when an action verb (e.g., *throw*) immediately precedes activation of a motor routine for a similar but subtly different action (e.g., *push*), as compared with a less similar action (e.g., *kick)* (Bergen et al., In press).

Critically, Korean differs from English in terms of where in the sentence the main verb is placed. In English, the verb in canonically ordered sentences occurs after the subject and before the object. This means that in a sentence about someone acting on something (like *You handed Andy the pizza)*, the verb appears relatively early in the sentence. In Korean, however, the verb occurs at the end of the sentence (as exemplified in (3) below). As a result, when a participant is asked to perform an action immediately after the end of a Korean sentence, it falls within the 500ms window in which we observe mismatch-advantages for similar but non-identical actions. In this experiment, the described actions are indeed different in certain ways from the action of pressing keyboard buttons. Handing someone a business card or a letter, for instance, is different from pressing keyboard buttons in terms of its force, acceleration, palm orientation, handshape, and other motor details. Because participants are asked to process language about an action and then perform a similar but subtly different action very soon thereafter, we expect that Korean sentences should produce a mismatch advantage. This is unlike canonical English sentences, which, due to their verb occurring earlier in the sentence, would be expected to produce *match* advantages, as has been found in other studies investigating language-action interaction effects (e.g., Glenberg & Kaschak, 2002).

(3) Ne-nun Andy-hanthey ku pizza-lul *kenneyesse*.
You-TOPAndy-DAT the pizza-ACC handed.
'You *handed* Andy the pizza.' (English translation)

Results

Among data from 32 participants and 20 critical items, results from two participants and one item were excluded due to low mean accuracy (more than 3 standard deviations below the participants' and items' overall means of 97% and 96%, respectively). The means for accurate responses from each of 30 participants and 19 items all fell within 3 standard deviations from the overall means for participants and items, respectively, and none were therefore eliminated as outliers.

The dependent measure was Reaching Time, the time it took participants from release of the yellow buttons until press of the pink or green buttons, indicating that the sentence was meaningful. We first eliminated all incorrect responses. Correct responses to critical items were then winsorized using 3 standard deviations from each participant's mean as a cut-off and submitted to the subsequent inferential statistical tests. Mean reaction times in each condition are shown in Figure 2.

We performed two two-way repeated measures ANOVAs (one by participants and one by items) to look at effects of the two independent variables – Sentence-Type with Hand-Number. We found one significant main effect, that was unrelated to our hypothesis; bimanual responses were on average slower than unimanual responses, regardless of the Sentence-Type manipulation: $F_1(1,29)=20.21$, $p<.001$; $F_2(1,18)=27.18$, $p<.001$. More importantly, however, we also observed an interaction between Hand-Number and Sentence-Type that was significant both by participants $F_1(1,29)=7.60$, $p=0.01$, and by items, $F_2(1,18)=6.60$, $p=0.02$. This effect, as seen in Figure 2, seems to have been driven by slower two-handed responses in reaction to sentences

about object transfer to high status recipients than low status recipients, and one-handed responses that were slightly slower after sentences about transferring objects to low status recipients than high status recipients.



Figure 2: Response times as a product of high status and low status sentences, when manual responses were made with one hand or two hands, show a significant interaction between Hand-Number and Sentence-Type

Subsequent pairwise comparisons indicated that this interaction effect was driven mainly by the differences in bimanual responses. That is, the mean reaction time to respond with two hands was significantly slower when the manual action was preceded by high status sentences than low status sentences ($t_1(29)=2.05$, $p=.02$; $t_2(18)=2.11$, $p=.027$). However, looking at the one-handed responses revealed no significant effect of Sentence-Type on Reaching time.

Discussion

The results reported above indicate that Korean speakers' manual responses to indicate sentence meaningfulness were significantly influenced by the social context mentioned in the sentences they were processing. People were significantly slowed down in making bimanual actions after they heard sentences describing object transfer to someone of higher social status, as compared with sentences about people of equal or lower status. In contrast, unimanual responses were slower with the low status sentences than the high status sentences, although this numerical difference did not approach statistical significance.

This result is yet another piece of evidence in line with previous findings, showing that people engage their motor systems while processing language about interacting with objects and moving their bodies more generally. The study of how exactly we extract meaning from utterances is still in its infancy. And yet, the discovery that modality-specific systems are automatically engaged during the process suggests one part of the puzzle – it could be that motor (and perceptual) simulation plays a functional role in language understanding by allowing the comprehender to construct an experience that is in some ways like what it would be to experience a distal described scene. This simulation may do more than merely create a subjective experience akin to what the comprehender would experience when confronted with the described scene; it might also facilitate inference or be used to generate predictions or interpolate implied but implicit elements of the scene.

However, the findings we've reported here is different from previous work. The current study's findings suggest that the motor activation comprehenders engage during language processing is tailored not only to the objects described but also to the culturally appropriate motor actions that one would perform in the described social context. This suggests that motor simulation isn't merely activated by specific words. Instead, it reflects a computation of socially appropriate action, which must take into account not merely low-level physical properties of a mentioned object, but social variables like age and position. On some accounts, people perform mental simulation to create representations of what it would be like if they were immersed in the described experience (Zwaan, 2004). The finding that people take social variables into account when constructing motor simulations is coherent with this account.

While the finding reported above highlights the importance of social knowledge in language comprehension, the difference we observed between the high and low status sentences could be due to either or both of two differences between them. As we discussed above, the sentences differed both in terms of the social character of mentioned recipients, as well as in the presence or absence of honorific markers. Honorifics—also known as indexical politeness forms—are grammatical and lexical markers in the Korean language that systematically encode the speaker's socio-culturally appropriate regard towards the addressee and/or the referent (Sohn, 1999). In our stimuli, the High Status sentences, as exemplified in the example (1), indicated that the status of the recipient was higher than that of the subject (*you*) by employing honorific markers in three places within the sentence, since that is how native Korean speakers would most naturally express this. The marker *-nim* 'sir/madam' on the recipient noun indicates deference to the high status referent. The marker *-kkey* '(honorable) to' is a case marker used for a high status recipient. The verb root *tuli-*, which is the humble form of the plain verb *cwu-* 'give', is used to indicate deference to the high status recipient. To any native Korean speaker, the presence of these honorific markers clearly indicates that the subject of the sentence (*you*) has a lower status than the mentioned recipient. When these honorifics are dropped, Korean speakers naturally understand that the status of the sentential subject is at least equal or even higher to the mentioned recipient. For instance, dropping honorifics is a common way for Korean students to make fun of their teachers behind their back.

Due to these two kinds of differences, when we compare effects of these two sentence types (1) and (2), we cannot tell if the status of the recipients or the presence/absence of honorifics is responsible for the differences in hand

responses. We believe it is critical to determine whether the social status of the recipients by itself produces the interaction effect, or whether the presence of honorifics is important as well.

To investigate this question, we are currently collecting data for a follow-up study. In that study, we closely matched the current experiment's design and materials, but included a third Sentence-Type condition, one with sentences using high status recipients but no honorific markers. If these sentences behave like the high status sentences in this first experiment, that will suggest that it is the status of the recipient, and not the presence of honorifics, that produces the effect we've observed. But if we find that these new sentences behave like low-status sentences, then this will lead us to conclude that honorifics present in a sentence are a critical factor affecting the motor routines people simulate. This in turn will tell us a little bit about how world knowledge and linguistic cues affect the motor simulations comprehenders construct.

In the Method section, we presented the reasoning why we expected a mismatch-advantage rather than a match-advantage. We found such an effect, and the next step is for us to investigate whether the proposed explanation is correct. We intend to do this through another study, currently under design, in which we change the structure of the critical Korean sentences, so that several words appear at the end of the sentence, after the transfer-action verb. When we thereby increase the interval between the action verb and the subsequent motor response, our explanation would predict that we should find a reversal in the direction of the effect: from a mismatch- to a match-advantage.

The final aspect of the findings reported here that may be of interest is the result of the follow-up pairwise t-tests reported in the Results section. The significant mismatch advantage we observed appears to have been driven more by the bimanual responses than the unimanual responses. We'd like to offer two possible explanations for this fact.

First, in situations of actual transfer, people use the right hand regardless of the status of the recipient; two-handed actions use both hands by definition, and one-handed actions conventionally use the right hand, as left hand transfer is regarded as disrespectful in Korean culture. As a result, both the high-status and low-status sentences should engage motor circuitry involved in the use of the right hand. And the two sentence types might thus interfere to the same extent with subsequent one-hand actions. By contrast, two-handed transfer actions are more similar to – and thus are more interfered with following – sentences about high-status actions, which also use two hands.

A second possible explanation for the larger effect in two-handed responses than one-handed responses is that it could result from differences in the degree to which the two actions are routinized. Perhaps reaching actions to press buttons with one hand are so common as to be routinized to the point where people perform them at floor (i.e., shortest possible reaction times). This would leave little room for effects of previously heard sentences on reaching behavior. In contrast, using both hands is a more difficult task, particularly when participants are trying to maximize both speed and accuracy. It could be that greater difficulty in bimanual action provided room for the critical Sentence-Type factor to more differentially influence reaction times. Indeed, the significantly faster responses in the one-hand condition, as compared with the two-hand condition, are compatible with this floor account. We expect our follow-up studies will further illuminate the extent to which the ACE paradigm can tease apart the action execution resulting from a mental computation of the socially appropriate action, versus an action execution based on a routinized motor command.

Conclusion

Is language-driven motor simulation sensitive to not only physical but also social constraints on action? The current study suggests that it is. We've presented three new findings. First, the motor simulation that people engage during language comprehension includes the number of hands one would use to perform a described action. Second, cultural-specific rules for motor control enter into these mental simulations. For Korean people, it is customary to tailor the number of hands one uses to transfer objects to the relative social status of the self and the object recipient. This social knowledge about proper actions is engaged in the form of active motor representations of the number of hands appropriate in each social context, even when people merely hear language describing those events. Finally, the direction of action-sentence interactions is different in Korean and English – arguably due to differences in word order. These findings clearly indicate that language understanding is a constructive process that broadly engages heterogeneous cognitive systems – it uses our understanding of physical actions, and even the social conventions surrounding those actions.

These results raise a host of potentially productive follow-up questions. Do people engage socially constrained motor knowledge during other non-motor tasks, such as object perception or recall? Do people raised biculturally, such that they have two distinct systems of motor conventions, simulate different actions depending on the cultural context or the language of an utterance? And is a social constraint on action the type of thing that can be learned late in life – do people introduced to a culture and language as adults also simulate socially appropriate actions? Clearly, there is much more to know about how social constraints on action are learned and what other aspects of cognition they affect.

Acknowledgments

The authors are grateful to Chae Eun Kim for assisting the stimuli recording, Bodo Winter for helpful suggestions on this work, and to Korean students at the University of Hawai‘i who participated in the experiment.

References

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, *22*, 577-609.

Bergen, B., & Wheeler, K. (2005). Sentence Understanding Engages Motor Processes. In *Proceedings of the Twenty-Seventh Annual Conference of the Cognitive Science Society.*

Bergen, B. (2007). Experimental methods for simulation semantics. In M. Gonzalez-Marquez, I. Mittelberg, S. Coulson, and M. J. Spivey (eds.) *Methods in Cognitive Linguistics: Ithaca*. John Benjamins.

Bergen, B., Lau, A., Narayan, S., Stojanovic, D., & Wheeler, K. (In press). Body part representations in verbal semantics. *Memory and Cognition*.

Bergen, B., & Wheeler, K. (In press). Grammatical aspect and mental simulation. *Brain and Language*.

Bub, D. N., Masson, M. E. J., & Cree, G. S. (2008). Evocation of functional and volumetric gestural knowledge by objects and words. *Cognition*, *106*, 27-58

Buccino, G., Riggio, L., Melli, G., Binkofski, F., Gallese, V., Rizzolatti, G. (2005). Listening to action-related sentences modulates the activity of the motor system: A combined TMS and behavioral study. *Cognitive Brain Research*, *24*, 355–363.

Connell, L. (2007). Representing object color in language comprehension. *Cognition*, *102*, 476–485.

Glenberg, A. M., & Kaschak, M. P. (2002). Grounding action in language. *Psychonomic Bulletin & Review*, *9(3)*, 558–565.

Kaschak, M. P., Madden, C. J., Therriault, D. J., Yaxley, R. H., Aveyard, M. E., Blanchard, A. A., & Zwaan, R. A. (2005). Perception of motion affects language processing. *Cognition*, *94*, B79–B89.

Kaschak, M. P., & Borreggine, K. L. (2008). Temporal dynamics of the action-sentence compatibility effect. *Quarterly Journal of Experimental Psychology*, *61*, 883-895.

Mauss, M. (1934). Les Techniques du corps, *Journal de Psychologie 32*, 3–4. Reprinted in Mauss, Sociologie et anthropologie, 1936, Paris: PUF.

Sohn, H.- M. (1999). *The Korean Language*. New York: NY. Cambridge University Press.

Stanfield, R. A., & Zwaan, R. A. (2001). The effect of implied orientation derived from verbal context on picture recognition. *Psychological Science*, *12*, 153–156.

Tettamanti, M., Buccino, G., Saccuman, M., Gallese, V., Danna, M., Scifo, P., Fazio, F. (2005). Listening to action-related sentences activates fronto-parietal motor circuits. *Journal of Cognitive Neuroscience*, *17*, 273–281.

Zwaan, R.A., Stanfield, R., & Yaxley, R. (2002). Language comprehenders mentally represent the shapes of objects. *Psychological science, 13(2)*, 168–171.

Zwaan, R. A. (2004). The immersed experiencer: toward an embodied theory of language comprehension. In B. H. Ross (Ed.), *The Psychology of Learning and Motivation*. New York: Academic Press.

Zwaan, R.A., Madden, C.J., Yaxley, R.H., & Aveyard, M.E. (2004). Moving words: Dynamic mental representations in language comprehension. *Cognitive Science*, *28*, 611–619.

Connecting Causal Events: Learning Causal Structures Through Repeated Interventions Over Time

Benjamin M. Rottman (benjamin.rottman@yale.edu)
Department of Psychology, Yale U., 2 Hillhouse Ave
New Haven, CT 06520

Frank C. Keil (frank.keil@yale.edu)
Department of Psychology, Yale U., 2 Hillhouse Ave
New Haven, CT 06520

Abstract

How do we learn causal structures? All current approaches use scenarios in which trials are temporally independent; however, people often learn about scenarios unfolding over time. In such cases, people may assume that other variables don't change at the same instant as an intervention. In Experiment 1, participants were much more successful at learning causal structures when this assumption was upheld than violated. In Experiment 2, participants were less influenced by such temporal information when they believed the trials to be temporally independent, but still used the temporal strategy to some extent. People seem to be inclined to learn causal structures by connecting events over time.

Keywords: causal reasoning; causal structures; time

Introduction

How do our concepts of event units influence how we learn causal structures? Despite the surge of research on causal structure learning, there has been little attention to how learners "connect" streams of information over time.

Existing theories of how people learn causal structures have focused on cases with events considered to be *independent.* For example, suppose we are trying to learn the causal relationships between three economic variables: employment, GDP, and consumption. Existing psychological theories suggest that one looks at the relationships among the variables across *many separate countries* to determine the causal structure. We call this strategy the *independent events* strategy because the countries are assumed to be independent.

An alternative approach is to pick one country and follow the three variables over time. We could track whether GDP goes up when employment goes up, etc. We call this strategy the *dependent events* or *temporal* strategy because the state of each variable is dependent on its prior state.

Psychologically, the temporal strategy may be pervasive and perhaps a default. As temporal beings we often perform or witness sequences of actions on one entity. For example, a car mechanic or computer technician can repair different components until the problem is solved. A psychotherapist can attempt to change one person's beliefs, emotions, and behaviors systematically over time. A physician can intervene on heart rate, breathing, and blood pressure to stabilize a patient. In many real-world situations we do interact with causal systems repeatedly over time, and thus the temporal strategy may be common if not a default for learning causal structures.

In formal statistics we have developed specialized procedures for independent cases (e.g., between-subjects) and dependent cases (e.g., repeated-measures, time-series). Analogously, do people use different learning strategies for the two scenarios? In the following sections we detail the different inferences people might make.

Interventions with Independent Trials

Consider first one prominent account of how people learn causal structures from interventions when trials are independent (e.g., Gopnik et al., 2004; Pearl, 2000; Steyvers et al., 2003). According to this model, when you intervene upon a variable such that you control its state, that variable is assumed to be independent from its other causes, but its effects are still dependent on that variable. Consider again the example of learning the causal relationships between employment (*E*), GPD (*G*), and consumption (*C*). Pretend that a priori it is possible that any of these factors could influence or be influenced by any of the other factors. To learn the causal structure, one could intervene on each of the three variables to determine which other variables are influenced by (dependent upon) the intervention.

Suppose that the true causal structure is a chain; *E* influences *G*, which influences *C*; $E \rightarrow G \rightarrow C$. If we could institute jobs-creation programs in 10 countries, we would expect them to have high *G and C.* If, hypothetically, we instituted a mass lay-off of government employees, we would expect comparatively low *G* and low *C*. These opposite interventions demonstrate how *G* and *C* are dependent on *E*. If we somehow selectively *boosted G* for10 new countries, they would have high *C*, but the same *E* as if we *decreased G* for 10 other countries; *C* is dependent on *G* but *E* is not. And if we gave 10 countries a boost in *C*, and another 10 countries a decrease in *C*, the two countries should have the same *E* and *G*; neither is dependent upon *C*.

If instead the true causal structure is a common cause such that *E* influences both *G* and *C*, $G \leftarrow E \rightarrow C$, we would expect a different pattern of (in)dependence. If we increase or decrease *G*, the respective countries would have the same levels of *E* and *C* because they are independent of *G*.

This strategy can identify the precise causal structure because each causal structure has a different pattern of (in)dependence when the variables are intervened upon.

Importantly, however, this strategy requires that the observations be independent. This strategy does *not* look at whether one country's GDP *improves after* increasing employment compared to *before* (a within-subjects design). It only compares the outcome of countries with increased vs. decreased employment.

Repeated Interventions Over Time

In contrast to the case just described, there are many scenarios in which a person intervenes repeatedly on one entity, and states of variables are fairly stable over time (e.g., car mechanic, physician). Consider a case in which we repeatedly intervene to increase or decrease *E*, *G*, and *C* within the United States. Suppose that the true causal structure is E→G→C, and initially the country is in a recession and all three variables are low. If we start a jobs-creation program, we would expect *G*, and *C* to increase *compared to before the intervention*. Then, suppose that we decreased *G*. We would expect *E* to *stay high,* but *C* to decrease. Finally, suppose that we encouraged consumption. We would expect *E* and *G* to stay the same. In contrast, suppose that the true causal structure is *G*←*E*→*C*. Now, if we increase *G*, we would expect *E* and *C* to stay the same, but we would expect both to change if we intervened on *E.*

In sum, if we repeatedly intervene on one entity, we expect variables that are not influenced by the intervention to *remain constant*. If we intervene upon a variable *X*, and another variable *Y* changes *from the previous state*, it is a sign that *X* causes *Y*. If *Y* does not change when *X* is manipulated, it is a sign that *X* does not cause *Y*. These inferences are intuitive given the assumption that causes are generally stable and don't happen to change at the same moment that another cause is manipulated. This temporal assumption of "stability" is analogous to the atemporal assumption that interventions are independent of other causes (e.g., Pearl, 2000; see also Rottman & Ahn, 2009a).

Testing Whether People Use the Two Strategies

The temporal strategy is very different from the strategy appropriate for independent observations. Only in the temporal case are the changes in variables over time important for learning causal structure and thus the order of the trials is critical.

To determine whether people are sensitive to the temporal information, we created pairs of data that have the same sets of 24 intervention trials, but with different trial orders. For example, consider the chain data in Figure 1. There are three variables (*X*, *Y*, and *Z*) and two possible values (0, and 1). Bold represents an intervention. For example, on Trial 1 for the useful chain condition, *X* was intervened upon and set to 1. *Y* and *Z* consequently have the value 1.

According to the independent trials strategy, both orders suggest the chain *X*→*Y*→*Z*. When *X* is intervened and set to 1, *Y* and *Z* are also 1. When *Y* is set at 1, *Z* is 1, but *X* can be either at 0 or 1 because *X* is not dependent on *Y*. Finally, if *Z* is set to 1, *X* and *Y* could both be 0 or 1 because they are independent of *Z*.

However, according to the temporal strategy, the two orders lead to very different inferences because the useful condition upholds the stability assumption but the misleading condition violates it. The "useful" condition suggests the *X*→*Y*→*Z* causal structure. Whenever *X* is changed, *Y* and *Z* also change (e.g. the transition from Trials 1 to 2). Whenever *Y* is changed, *Z* also changes, but *X* stays the same (e.g., Trials 2-3). When *Z* changes, *X* and *Y* stay the same (e.g., Trials 4-5). In contrast, misleading conditions were designed to suggest the presence of links that do not exist. For example, on Trial 2, *Z* is changed from 1 to 0, and *X* and *Y* also change to 0, suggesting that *Z* causes *X* and *Y*. Additionally, causal links are not consistent. On Trial 2, Z appears to cause X and Y to change to 0, but on Trial 3 it does not cause them to change back to 1. Finally, the existence of real links is obscured. For example, on Trial 5, X is changed from 0 to 1, but Y is already at 1, obscuring that X influences Y. In sum, the "misleading" condition suggests different links from the "useful" condition, and does not clearly identify one causal structure.

We used this order manipulation in two experiments. In Experiment 1, we tested whether people do in fact use the temporal strategy. In Experiment 2, we tested whether people appropriately switch between the two strategies based on the causal scenario.

	Chain						Common Cause					
	useful			misleading			useful			misleading		
Trial	X	Y	Z	X	Y	Z	X	Y	Z	X	Y	Z
1	**1**	1	1	1	1	**1**	**1**	1	1	1	**1**	1
2	**0**	0	0	0	0	**0**	**0**	0	0	0	**0**	0
3	0	**1**	1	0	0	**1**	**1**	1	1	1	1	**1**
4	0	**0**	0	0	**1**	1	1	**0**	1	0	0	**0**
5	0	0	**1**	**1**	1	1	1	**1**	1	**1**	1	1
6	0	0	**0**	**0**	0	0	1	1	**0**	**0**	0	0
7	**1**	1	1	1	**1**	1	1	1	**1**	0	0	**1**
8	1	**0**	0	0	**0**	0	**0**	0	0	1	1	**0**
9	1	**1**	1	0	0	**1**	0	**1**	0	0	0	**1**
10	1	1	**0**	0	**1**	1	0	**0**	0	1	1	**0**
11	1	1	**1**	**1**	1	1	0	0	**1**	**0**	0	0
12	**0**	0	0	**0**	0	0	0	0	**0**	**1**	1	1
13	**1**	1	1	**1**	1	1	**1**	1	1	0	**0**	0
14	**0**	0	0	1	1	**0**	**0**	0	0	1	**1**	1
15	0	0	**1**	1	**0**	0	0	0	**1**	**0**	0	0
16	0	0	**0**	**0**	0	0	0	0	**0**	**1**	1	1
17	0	**1**	1	1	**1**	1	0	**1**	0	1	**0**	1
18	0	**0**	0	0	**0**	0	0	**0**	0	0	**1**	0
19	**1**	1	1	**1**	1	1	**1**	1	1	1	**0**	1
20	1	**0**	0	1	1	**0**	1	1	**0**	0	**1**	0
21	1	**1**	1	1	**0**	0	1	1	**1**	**1**	1	1
22	1	1	**0**	**0**	0	0	1	**0**	1	**0**	0	0
23	1	1	**1**	1	1	**1**	1	**1**	1	1	1	**1**
24	**0**	0	0	0	0	**0**	**0**	0	0	0	0	**0**

Figure 1: Summary of Data for Two Causal Structures in Experiment 1.

Experiment 1

In Experiment 1, we created a scenario in which one causal setup is repeatedly intervened upon over time. Thus participants would likely think that the temporal information was relevant. We presented participants with data generated by five causal structures. For each causal structure, there was a useful and misleading set of data. If participants use the temporal strategy, they will learn the causal structures more accurately in the useful condition.

Methods

Twenty undergraduates completed the study for payment at $10 per hour or partial course credit. Participants first read a cover story about three light bulbs. Participants were told that they would be instructed to turn on or off specific lights and should try to "learn how each light affects the others."

Next, participants saw 10 scenarios created by crossing the Order of the Data (useful vs. misleading) × Causal Structure (chain, $X \rightarrow Y \rightarrow Z$; common cause, $Y \leftarrow X \rightarrow Z$; common effect, $X \rightarrow Z \leftarrow Y$; one link, $X \rightarrow Y$, Z is unrelated; no links, X, Y, and Z, are unrelated). The 10 scenarios were ordered in a Latin square grouped by causal structure such that each scenario appeared first for some participants.

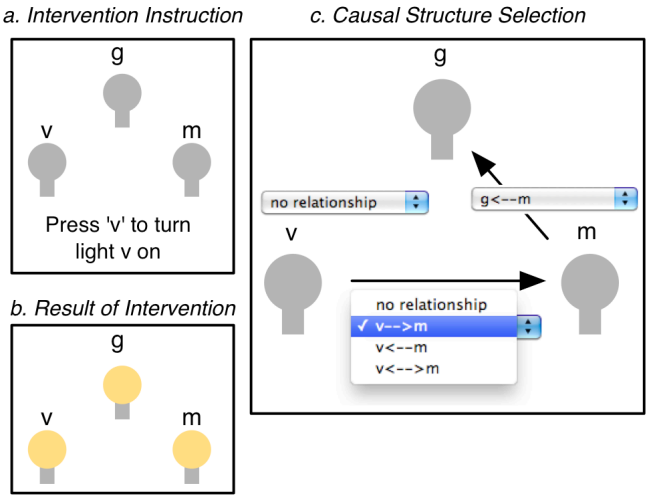


Figure 2. Example Screenshots from Experiment 1.

During each scenario, participants saw three light bulbs. Each bulb was named by a letter, and different letter triads were used across the 10 scenarios. Initially, all three bulbs were off. Then participants were instructed to intervene to turn on or off specific bulbs (e.g., Figure 2a). To intervene, participants pressed the key associated with the letter for the given bulb. After the intervention, participants observed the outcome of the intervention (which bulbs were on or off) for 2 seconds (e.g., Figure 2b). Then, while the bulbs were still visible, instructions appeared for the next intervention.

Each scenario had 24 interventions total, 8 per bulb; 4 on and 4 off. The data were determined in the following way. The causal relations were deterministic; when a bulb was intervened upon, all its effects (and all of their effects) assumed the same value. Exogenous variables had a base-rate of .5. For the common effect structure, the effect was on if either of the causes was on.

For the "useful" conditions, the trials were ordered in a way that upheld the stability assumption explained in the introduction whereas the "misleading" conditions violated it. Figure 1 displays a summary of the data for the chain and common cause scenarios. The data for the other three causal structures can be obtained from the authors.

After each scenario, participants selected the causal structure that they believed to have generated the pattern of data for the given scenario (e.g., Figure 2c). Participants selected arrows indicating the direction of the causal relationships between the three light bulbs. For each pair of bulbs (e.g., X and Y), participants chose between "no relationship; neither light influences the other", "$X \rightarrow Y$; X influences Y", "$X \leftarrow Y$; Y influences X", or "$X \leftrightarrow Y$; X and Y both influence each other." Participants did not receive feedback of the accuracy of their causal model. Finally, participants started the next scenario.

Results

Accuracy in causal structure inferences was assessed in the following way. For each pair of bulbs, X and Y, X can cause Y or not, and Y can cause X or not. Thus for each pair of bulbs, participants had the possibility of identifying zero, one, or two correct causal relations. Across the three bulbs in a given scenario, participants had the possibility of identifying zero to six correct causal relations.

For all of the five causal structures, participants identified more correct causal relations in the useful than misleading conditions $ts(19)>8.32$, $ps<.01$ (Figure 3), suggesting that they used the trial order for learning causal structures.

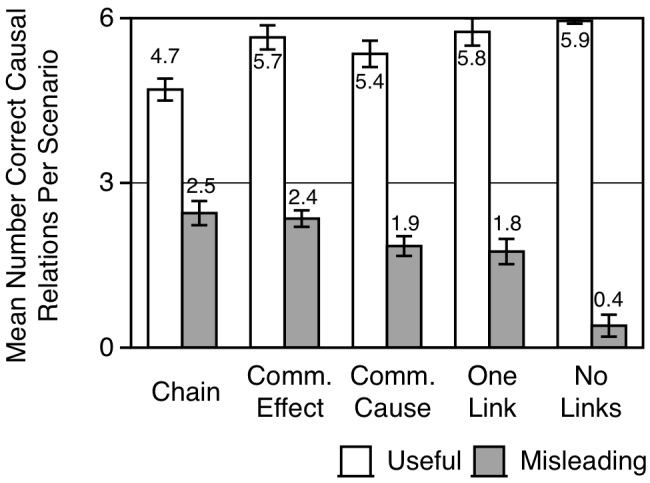


Figure 3: Mean Accuracy (Std. Errors) in Experiment 1.

There are two trends in participants' mistakes. First, in the useful chain condition ($X \rightarrow Y \rightarrow Z$), participants had difficulty learning that Y was a mediator between X and Z. This requires noticing that when Y is manipulated, X has no

influence on Z. Eighteen out of the 20 participants thought that *X* also caused *Z* directly, probably because when *X* was turned on and off, *Z* also changed state. Similar findings have been interpreted to suggest that people sequentially learn individual causal links rather than simultaneously learn an entire causal structure (Fernbach & Sloman, 2009).

Second, in the misleading conditions, participants frequently correctly identified true causal links, but they also mistakenly thought that other links existed. They often thought that links were bidirectional, even though they were just unidirectional. In the one link and no link conditions, they also frequently inferred relationships between variables with no causal relations. These inferences resulted in participants often misidentifying the majority of the causal links; the accuracy in all misleading conditions was below chance responding of 3, all $ts(19)>2.4$, $ps<.03$. However, these inferences make sense according to the temporal strategy; the misleading orders were designed so that variables that were not effects of a manipulated variable frequently change at the same time as the intervention, suggesting additional causal relationships.

In sum, the results strongly suggest that participants were sensitive to the order of the trials and were using the transitions between trials to infer causal relationships.

Experiment 2

In Experiment 1, it was rational for participants to use a temporal strategy to learn causal structures because participants observed entities change over time. The purpose of Experiment 2 was to determine how flexibly people apply the temporal vs. independent strategies given different scenarios. We created two scenarios intended to give maximal cues to participants that the trials were either independent (analogous to a between subject design) or dependent (analogous to a within-subjects design). Previous studies have successfully used such a manipulation (Rottman & Ahn, 2009b). We then tested whether participants would infer different causal structures in useful vs. misleading orders. If participants use the temporal strategy for the dependent case, they would be more accurate in the useful than misleading order, as in Experiment 1. Additionally, if they do not use temporal information in the independent scenario, they would not have different levels of accuracy for the two orders.

Methods

Sixteen students from the same population participated.

Participants first read a cover study story asking them to pretend that they are assistants in a biology lab studying hormones in amoebas. They would "produce" or "suppress" hormones by injecting chemicals into the amoebas and "learn how each hormone affects the others." They were told that the "hormones work immediately… without any perceivable delay."[1]

[1] This statement about no delay was intended to rule out the possibility of second order causal relationships (e.g., if Hormone A is produced and suppressed twice in a row, then Hormone B would be produced), which some participants reported in pretesting. In both the dependent and independent conditions, the interventions do work immediately after the intervention key is pressed.

Next, participants saw eight scenarios. Each scenario presented three hormones. "+" and "-" signs denoted the results of the hormones, presence and absence respectively. The eight scenarios were created by crossing Number of Amoebas (one vs. many) × Trial Order (useful vs. misleading) × Causal Structure (common cause, $Y \leftarrow X \rightarrow Z$ vs. one link, $X \rightarrow Y$, Z is unrelated). The design was entirely within subjects. The 8 scenarios were ordered in a Latin square such that each scenario appeared first for some participants, and the scenarios were grouped by number of amoebas. The trial order and causal structure manipulations were the same as in Experiment 1, so the following paragraphs focus on the number of amoebas manipulation.

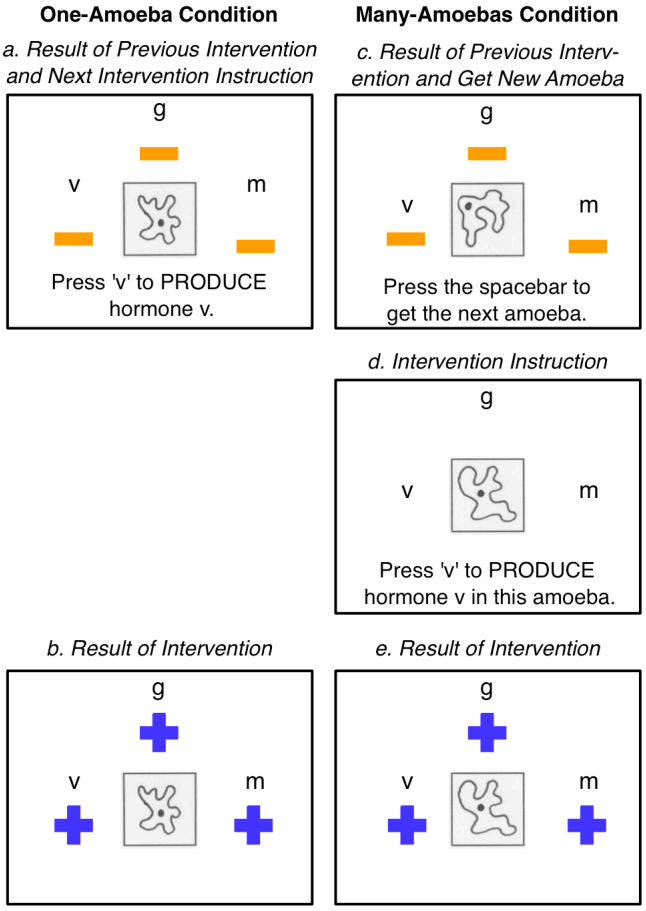


Figure 4: Example Screenshots from Experiment 2.

The one-amoeba condition, analogous to a within-subjects design, emphasized the dependent nature of the data. The one-amoeba procedures were similar to those in Experiment 1; participants repeatedly intervened on one amoeba. While the result of the previous intervention was displayed, participants were instructed to "PRODUCE" or "INHIBIT"

a specific hormone (e.g., Press "y" to PRODUCE hormone y; e.g., Figure 4a). After the intervention, participants observed the result of the intervention for 2 seconds (e.g., Figure 4b). While the results were visible, instructions for the next intervention appeared. Additionally, a picture of one amoeba was present for the entire scenario to emphasize the repeated interventions on a single entity over time.

The many-amoebas condition, analogous to a between-subjects design, emphasized the independent nature of the data. Participants made 24 interventions on 24 different amoebas. After the results of a given intervention were displayed, participants were instructed to "Press the spacebar to get the next amoeba" (e.g., Figure 4c). When the spacebar was pressed, a picture of a new amoeba appeared. Simultaneously, the results of the intervention on the previous amoeba ("+" and "-" marks) disappeared (e.g. Figure 4d). We removed the previous results to make it perceptually difficult to track the changes of the hormones over time. Two seconds later, the prompt for the next intervention appeared (e.g., Press "y" to PRODUCE hormone y in this amoeba). When the intervention key was pressed, the hormone results appeared for the current amoeba (e.g., Figure 4e). All of these modifications were intended to signal that the hormones within one amoeba were independent of the hormones within other amoebas.

After each scenario, participants selected the causal structure that they believed to have generated the data.

Results

The dependent variable was the same as in Experiment 1 – the number of correctly identified causal relations per scenario (zero to six).

A 2 (one vs. many amoebas) × 2 (trial order) × 2 (causal structure) repeated-measures ANOVA was performed. There was a main effect of trial order; participants correctly identified more causal relationships in the temporally useful than misleading orders, $F(1,15)=45.28$, $p<.01$, $\eta_p^2=.75$ (Figure 5). However, the most critical result for this experiment is a significant interaction between number of amoebas and trial order, $F(1,15)=12.61$, $p<.01$, $\eta_p^2=.46$.[2] Though there was a large difference between the useful and misleading orders for the one-amoeba condition, there was a smaller difference between the many-amoebas conditions, suggesting that participants were less sensitive to the temporal order of trials in the many-amoebas condition. This finding makes sense if participants believed that the trials were independent in the many-amoebas condition.

However, even though participants used the temporal strategy less in the many-amoebas condition, they still used it to some extent; there was still a significant difference between the useful and misleading, many-amoebas conditions, $t(15)=3.59$, $p<.01$. Furthermore, participants did not simply transfer the temporal strategy from the one-amoeba condition; they were more accurate in the useful than misleading many-amoebas conditions even before experiencing the one-amoeba scenarios, $t(7)=3.21$, $p=.02$.

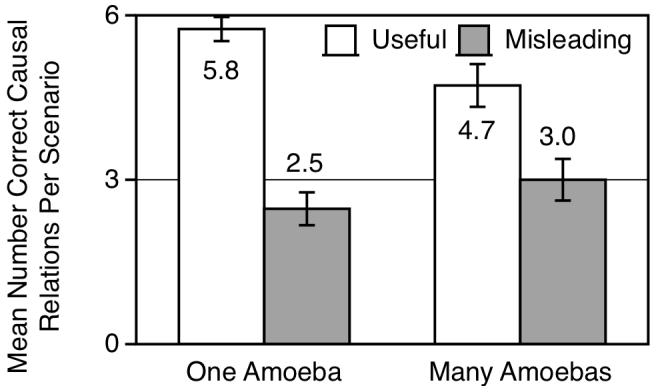


Figure 5: Mean Accuracy (Std. Errors) in Experiment 2.

There are two other important patterns. First, participants did worse in the many-amoeba than one-amoeba, useful condition, $t(15)=2.57$, $p=.02$. This finding makes sense if participants were using the temporal strategy less in the many-amoebas condition. However, according to the independent trials strategies (e.g., Gopnik et al., 2004; Steyvers et al., 2003), participants should have been able to correctly identify the causal structures in the many-entity conditions. Second, participants were not even above chance in the many-amoebas, misleading condition, $t(15)<1$. Yet again, participants should have been able to identify the correct causal structures according to the independent trials strategy. The low accuracy in both many-amoebas conditions suggests that participants may have difficulty applying such statistical strategies.

In sum, participants are able to switch between the temporal vs. independent strategies to some extent based on knowledge of the learning scenario. However, even in the many-amoebas condition, participants used the temporal information to some extent, suggesting that it is a common strategy for learning causal structures.

General Discussion

In two experiments, we demonstrated that people learn causal structures very well when entities are repeatedly manipulated over time (i.e. within-subjects or repeated measures situations). In Experiment 1, participants were much more accurate at learning causal structures when the data were ordered to reflect causes that are stable over time (don't happen to change at the moment another variable is intervened upon), a plausible real-world assumption. In Experiment 2, participants were less sensitive to the temporal order of trials when they were given reason to believe that the trials were independent (i.e. between-subjects situation).

[2] The only other finding was a marginally significant interaction between causal structure and trial order, $F(1,15)=4.03$, $p<=.06$, $\eta_p^2=.21$. The difference between the useful and misleading orders was slightly larger for the common cause than one link conditions.

Predominance of the Temporal Strategy

Why did participants in the many-amoebas condition in Experiment 2 still make use of the temporal information to some extent? There are two possible explanations. First, people may have still thought that the hormones within different amoebas were dependent upon one another. (For example, if all the amoebas were physically adjacent, perhaps hormones could mix across the amoebas.) Alternatively, people might have been able to learn that the trials were dependent from the data itself. In reality, in the many-amoebas, useful condition, the order was statistically dependent. For example, exogenous variables (e.g., X in $X \rightarrow Y \rightarrow Z$) only changed state when X was intervened upon. For long periods of time, X stayed the same (e.g., Trials 2-6 in Figure 1, Chain, Useful) even though its baserate is .5.

However, there is also a second possibility – the temporal strategy is likely simpler than the statistical strategies proposed for independent events (e.g., Gopnik et al., 2004; Steyvers et al., 2003). Thus, it is possible that people tend to use this strategy even in cases when the independent strategy is more appropriate. Perhaps the time-based strategy serves as a useful heuristic that is often accurate. In the real world, much of our causal reasoning involves manipulating and observing sequences of events unfolding over time (e.g., a car mechanic repairing different components until the problem is solved or a physician manipulating a patient's heart rate, breathing, and blood pressure to stabilize the patient). Given how frequently we engage in temporal reasoning, it may be hard to ignore temporal information such as the order of trials in these experiments even when we should for independent events.

Learning Causal Structure from Temporal Delay

Lagnado and Sloman (2004, 2006; see also Burns & McCormack, 2009; Meder et al., 2008; White, 2006) showed how people use temporal *delays* when learning causal structures. For example, if you intervene upon X, and then Y appears, and later Z appears, you would likely infer $X \rightarrow Y \rightarrow Z$. This strategy pertains to the time course of how a causal signal propagates through a network and the order in which the reasoner becomes aware of the states of the nodes. This strategy is entirely consistent with the current one, and they likely often work in parallel in the real world. However, they are distinct. In the current studies, both of the non-manipulated variables appear simultaneously for all causal structures. Additionally, in the previous studies (e.g., Lagnado & Sloman, 2006), the trials were independent and were often randomized.

Summary

Overall, people learn causal structures over time quite fluently and indeed seem biased to assume that this is the default mode of causal interpretation. Instead of treating trials as independent, which has been assumed by many approaches of causal structure learning, people weave together information across trials into larger event units.

The use of a temporal strategy can result in very quick and accurate causal structure learning when the trials are ordered in a temporally useful way. However, applying an incorrect causal strategy can result in substantially worse performance. For example, applying a more independent events strategy for events that were truly dependent and ordered in a useful fashion resulted in considerably worse performance than when participants applied the temporal strategy (Experiment 2). One intriguing possibility is that applying the temporal strategy when the events are truly independent could also likely result in reduced performance. Elaborating when and how people apply different learning strategies for diverse scenarios is an important future aim.

Acknowledgments

This research was supported by an NSF Graduate Research Fellowship (Rottman) and NIH grant R37 HD023922 (Keil). The authors thank Samuel Tepper for programming, and Rachel Litwin and Brianna Sullivan for data collection.

References

Burns, P., & McCormack, T. (2009). Temporal information and children's and adults' causal inferences. *Thinking & Reasoning, 15,* 167-196.

Fernback, P. M. & Sloman, S. A. (2009). Causal learning with local computations. *Journal of Experimental Psychology: Learning, Memory, and Cognition. 35,* 678-693.

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and bayes nets. *Psychological Review*, *111*(1), 3-32.

Lagnado, D. A., & Sloman, S. A. (2004). The advantage of timely intervention. *Journal of Experimental Psychology: Learning, Memory and Cognition, 30* (4), 856-876.

Lagnado, D. A., & Sloman, S. A. (2006). Time as a guide to cause. *Journal of Experimental Psychology: Learning, Memory and Cognition, 32*(3), 451-460.

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

Pearl, J. (2000). *Causality: Models, reasoning, and inference.* Cambridge Univ Pr.

Rottman, B. M. & Ahn, W. (2009a). *Causal Inference when Observed and Unobserved Causes Interact.* Proceedings of the 31st Annual Conference of the Cognitive Science Society (pp. 1477-1482).

Rottman, B. M. & Ahn, W. (2009b). *Causal Learning about Tolerance and Sensitization.* Psychonomic Bulletin and Review, 16, 1043-1049.

Steyvers, M., Tenenbaum, J. B., Wagenmakers, E., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science, 27,* 453-489.

The Induction of Hidden Causes: Causal Mediation and Violations of Independent Causal Influence

Christopher D. Carroll (cdcarroll@ucla.edu)
Department of Psychology, UCLA
Los Angeles, CA 90095 USA

Patricia W. Cheng (cheng@lifesci.ucla.edu)
Department of Psychology, UCLA
Los Angeles, CA 90095 USA

Abstract

In order to explain the apparent violation of a causal assumption, people often posit hidden causes. The assumption of independent causal influence states that the power of a cause to produce or prevent an effect is independent of other causes. Some preventers violate independent causal influence; we conducted an experiment to test whether people posit a hidden mediating cause to explain these preventers. The results indicated that participants are more likely to posit a hidden mediator when the preventer violates independent causal influence.

Keywords: causal reasoning; causal inference; prevention; hidden causes; unobserved causes

Introduction

Although people often reason about simple cause and effect, they typically assume that such causal relationships are embedded in complex causal structures with hidden causes. So while people know that aspirin prevents headaches, they also believe that this relationship is mediated by some complex biological mechanism involving hidden causes. In many circumstances, the hidden causes are inconsequential. Knowing how aspirin prevents headaches is less important than knowing that it does so. Indeed, people often reason appropriately with only shallow causal knowledge (e.g., Keil, 2003). However, hidden causes may be important in other circumstances.

In particular, hidden causes may be important when the observed causes violate causal assumptions such as the Markov assumption in causal Bayesian network models (Pearl, 2000; Spirtes, Glymour, & Scheines, 1993/2000) or the assumptions for inferring causal power (Cheng, 1997; Novick & Cheng, 2004). Inferences about hidden causes have been demonstrated in a number of studies where some causal assumption is violated. Children appeal to hidden causes in order to explain probabilistic causation, and this may reflect an assumption that causation is deterministic (Schulz & Sommerville, 2006; see also Lu, Yuille, Liljeholm, Cheng, & Holyoak, 2008). Similarly, both adults (Gopnik, Glymour, Sobel, Schulz, Kushnir, & Danks, 2004; Hagmayer & Waldmann, 2007; Luhmann & Ahn, 2007) and infants (Saxe, Tenenbaum, & Carey, 2005) posit hidden causes when there is an unexplained effect, presumably reflecting the assumption that every effect has a cause (Kant, 1781/1965). Finally, people infer a hidden contextual cause when the causal power of the observed cause interacts with its context (Liljeholm & Cheng, 2007; Rottman & Ahn, 2009).

In this paper, we focus on the assumption of *independent causal influence* (Cheng, 1997; Novick & Cheng, 2004). Independent causal influence requires that the power of one cause to produce or prevent the effect is constant: it does not change with context or with the occurrence or non-occurrence of other causes. According to independent causal influence, if aspirin prevents headaches caused by colds, then it will also prevent headaches caused by dehydration, stress, and so on. We investigate a specific violation of independent causal influence that arises in prevention.

Preventive scope is the range of circumstances across which a preventer works (Carroll & Cheng, 2009). A *broad preventer* stops the effect no matter what the cause, but a *narrow preventer* only stops the effect when the effect is produced by a certain *targeted cause*. Aspirin and nasal spray illustrate the difference between broad and narrow prevention. As a broad preventer, aspirin prevents headaches of all kinds (e.g., headaches caused by colds and headaches caused by stress). As a narrow preventer, sinus spray only prevents headaches caused by colds; it would not prevent a headache caused by stress.

Narrow prevention violates the assumption of independent causal influence because the power of the preventer depends on which cause is producing the effect *e*: a narrow preventer prevents *e* when it is brought about by the targeted cause *c*, but it does not prevent *e* otherwise. However, it is possible to reconcile narrow prevention and the assumption of independent causal influence by positing a certain type of hidden cause: a hidden *mediator*. Suppose that *c* produces *e* indirectly through a mediator and that the narrow preventer prevents the mediator rather than preventing *e* directly (see Figure 1). Once the mediator is included in the explanation, none of the causal relationships violate independent causal influence: *c* and the preventer independently influence the mediator, and the mediator and other causes independently influence *e*. As long as other causes of *e* produce *e* via mechanisms other than the mediator, the preventer will only stop *e* when it is being produced by *c*. Thus, narrow prevention would only appear to violate the assumption of independent causal influence because there is an unobserved mediator.

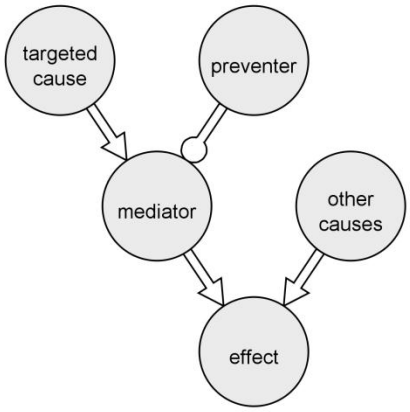


Figure 1: Mediation as an explanation of narrow prevention. Generative and preventive causation are denoted by arrows and modified arrows terminating in a circle, respectively.

Previous research (Carroll & Cheng, 2009) demonstrated that people distinguish between narrow and broad prevention, but the relationship between preventive scope and the inference of hidden mediation is less clear. Although participants in the previous research identified the mediation shown in Figure 1 as an explanation for narrow prevention, they only did so in a two-alternative forced-choice procedure. This shows that participants preferred the explanation in Figure 1 to the available alternative (an explanation where the preventer directly stopped the effect), but it is impossible to tell whether participants would have endorsed this explanation outside of this particular forced-choice question. Participants may have endorsed causal mediation as the better – but potentially unappealing – explanation of the two available choices. Furthermore, the experiment previewed the choices before showing the participants any data, and this may have biased participants towards interpreting the data with one of the provided explanations. Whether causal mediation is a favored explanation for narrow prevention more generally remains to be seen. Moreover, previous research did not clarify the relationship between preventive scope and the assumption of independent causal influence.

To assess whether the causal mediation explanation of narrow prevention is appealing more generally, we tested whether people endorse causal mediation after encountering a narrow preventer.

Method

Participants were asked to imagine themselves as researchers at a medical research company. They were directed to investigate how two fruit products from the rain forest - pane fruit and asmine juice - influence whether someone will have a headache. In all conditions, participants were shown some clinical trials where pane fruit caused headaches and asmine juice prevented headaches. We manipulated whether asmine juice was a narrow or broad preventer. After viewing the data, participants reported whether they expected asmine juice to prevent headaches under various circumstances. Finally, the participants were given a series of statements and were asked to endorse or reject each statement. One of these statements presented the mediation explanation.

Participants

Forty undergraduates at the University of California, Los Angeles (UCLA) participated to obtain course credit in a psychology course. Participants were assigned to the narrow (n = 20) or broad (n = 20) prevention condition.

Materials

The data presented in the narrow and broad prevention conditions are shown in Table 1. The critical difference between the conditions can be seen by comparing the effect of asmine juice on headaches attributed to the background cause. In the broad prevention condition, drinking asmine juice reduced the number of headaches even when pane fruit was not consumed. This can be seen by comparing the number of headaches when people neither ate pane fruit nor drank asmine juice to the number of headaches when people drank asmine juice but did not eat pane fruit (see the top half of Table 1). In the narrow prevention condition, it did not do so.

Table 1: The frequency of headaches (the effect) as a function of pane fruit (cause), asmine juice (preventer), and prevention condition. F = pane fruit, J = asmine juice.

Observed Causes	Broad prevention	Narrow prevention
none	10 out of 50	10 out of 50
J	5 out of 50	10 out of 50
F	40 out of 50	40 out of 50
J, F	20 out of 50	20 out of 50

As shown in Figure 2, the data were presented in displays containing cartoon faces. Each cartoon face represented a person in the clinical trial, and the type of cartoon face (happy face or sad face) indicated whether the person had a headache.

Procedure

Participants were randomly assigned to the broad or narrow prevention conditions and then given the following cover story:

Imagine that you work for a drug company that develops headache medications. You have heard rumors about a certain area in a rainforest where many of the fruits influence whether someone has a headache (either by causing a headache or preventing it).

The drug company has asked you to investigate these claims.

You decided to run clinical trials to assess the influence of pane fruit and asmine juice. You recruited

volunteers and randomly divided them into groups. Each group was assigned a specific treatment (e.g., eating pane fruit but not drinking asmine juice).

The results of each trial are summarized by tables of cartoon faces, and you can tell whether someone had a headache by looking at the cartoon face.

Participants were shown data in a display similar to Figure 2, and were given a print-out of the data to reference while answering subsequent questions. The instructions emphasized that the results had been replicated in much larger studies so that any differences in the frequency of the effect were reliable.

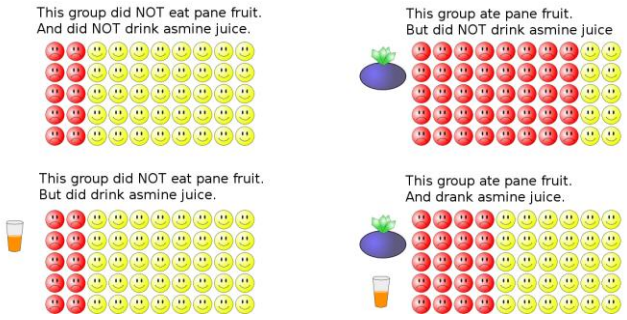


Figure 2: The data shown in the narrow prevention condition. Shaded frowning faces indicate people with headaches, and the lighter smiling faces indicate people without headaches.

Then, participants answered a series of counterfactual questions designed to measure beliefs about how pane fruit and asmine juice influence headaches. Each question asked the participants to imagine a group of people with certain characteristics and to predict whether consuming one of the food products would lead to more, fewer, or the same number of headaches in that group. For example, the *fruit* counterfactual - designed to assess the influence of pane fruit among a group of people who have not drunk asmine juice – asked following question:

Imagine that you go to a small town in the United States. If you brought PANE FRUIT to the town and everyone ate it, do you think that MORE people would have headaches, FEWER people would have headaches, or the SAME NUMBER of people would have headaches?

The other questions assessed when participants believed that asmine juice would prevent headaches. The *juice|fruit* question asked participants to predict the effect of asmine juice among people who live in a town near the rainforest and frequently consume pane fruit. The *juice|no fruit* question measured the influence of asmine juice among people living in a small town in America (who presumably have not eaten pane fruit). The *juice|withdrawal* question measured the influence of asmine juice among a group of people who have headaches for a specific reason other than eating pane fruit: they have stopped drinking coffee and are experiencing caffeine withdrawal.

Finally, participants were shown a series of statements about pane fruit and asmine juice. The statements were shown one at a time, and participants were asked to endorse whichever statements they agreed with. Table 2 lists these statements in the order that they were presented. Endorsement of the *mediation* statement provided the critical measure of whether participants inferred a hidden mediator. It should be noted that the mediation statement is compatible with broad prevention as well as narrow prevention: a broad preventer might destroy the substance in addition to directly preventing the effect when it is produced by other mechanisms.

Table 2: Participants were asked to indicate whether they agreed or disagreed with the following statements

Type	Statement
prevents	Asmine juice can sometimes prevent or relieve headaches.
develop drug	Your company may be able to turn asmine juice into a drug like aspirin, selling it widely as a headache treatment.
mediation	Pane fruit produces a RARE substance that causes headaches, and asmine juice destroys THAT substance.
combination	There is something special about the combination of asmine juice and pane fruit that prevents headaches.

Results

For the counterfactual questions, participants indicated whether there would be more, fewer, or the same number of headaches after consuming one of the food products. To analyze these responses, we coded responses of "more" as 1, "fewer" as -1, and "same number" as 0.

As expected, most participants predicted that pane fruit causes headaches. For the pane fruit counterfactual, the mean response was .90 (SD = 0.45) in the broad prevention condition and .95 (SD = 0.22) in the narrow prevention condition. The difference between these experimental conditions was not significant, $t(38) = 0.45$, $p = .66$.

On the other hand, the predicted influence of asmine juice depended on the experimental condition and the specific counterfactual (see Figure 3). Participants in both conditions believed that asmine juice would prevent headaches among groups of people that had eaten pane fruit (juice|fruit counterfactual). However, there were noticable differences between the conditions for the other counterfactuals. When participants were shown broad prevention, they believed that asmine juice would prevent headaches in every counterfactual. In contrast, when participants were shown narrow prevention, they were much less likely to believe

that asmine juice would prevent headaches when the headaches were produced by either an unknown (juice|no fruit) or a known non-targeted (juice|withdrawal) cause.[1]

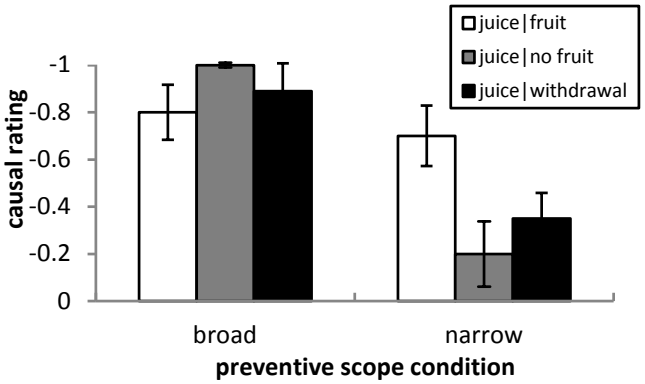


Figure 3: Prevention ratings for the counterfactual questions involving asmine juice. Error bars show the standard errors.

To confirm these patterns, we conducted an ANOVA with prevention condition (broad or narrow prevention) and prevention counterfactual (juice|fruit, juice|no fruit, or juice|withdrawal) as the independent variables. The ANOVA found a main effect of prevention condition, $F(1,38) = 17.08$, $p < .001$, and an interaction between prevention condition and prevention counterfactual, $F(2, 76) = 6.28$, $p < .01$. To investigate the source of the interaction, we conducted a separate ANOVA in each prevention condition. These ANOVAs confirmed that there was a non-significant effect of counterfactual question under broad prevention, $F(2, 38) = 1.85$, $p = .17$, and a significant effect of counterfactual question under narrow prevention, $F(2, 38) = 6.34$, $p < .01$.

The percentages of participants endorsing the statements are shown in Table 3. Participants in both conditions were very likely to report that asmine juice sometimes prevents headaches, but participants in the broad prevention condition were more likely to do so, $p < .05$ by Fisher's exact test. This difference might reflect the failure of some participants in the narrow prevention condition to notice that the preventer prevents the effect. Participants in the broad prevention condition were much more likely to believe that asmine juice could be developed into a headache drug and widely marketed, $\chi^2(1, N = 40) = 20.42$, $p < .001$. Participants in the narrow prevention condition were much more likely to believe that pane fruit and asmine juice might produce and prevent headaches via a rare shared mediator, $\chi^2(1, N = 40) = 4.91$, $p < .05$. Neither of the experimental conditions led many participants to suggest that the combination of pane fruit and asmine juice prevented headaches, and the difference between the conditions was not statistically significant, $p = .41$ by Fisher's exact test.

Table 3: Percentages of participants in each condition who agreed with the statements.

Question	Broad prevention	Narrow prevention
prevents	100%	75%
develop drug	95%	25%
mediation	35%	70%
combination	10%	25%

Discussion

This experiment demonstrates that people will endorse causal mediation in order to explain narrow prevention. The results also confirm that people distinguish between narrow and broad prevention, using preventive scope to guide generalization. Broad prevention was generalized irrespective of context, but narrow prevention was only generalized when the effect was produced by the targeted cause. A narrow preventer was not expected to stop the effect when the effect was produced by an unknown cause or a cause other than the targeted cause.

These findings contribute to a growing body of evidence showing that causal assumptions play a central role in the induction of hidden causes. Models of causal inference that make minimalistic assumptions (e.g., Pearl, 2000; Spirtes, Glymour, & Scheines, 1993/2000) may fail to explain these findings (see Griffiths & Tenenbaum, 2009 for other situations where minimalistic assumptions prove inadequate). In fact, since broad and narrow prevention imply the same conditional independencies among the observable variables,[2] Pearl's (2000) causal Bayesian network model represents them with the same causal graphs, treating a causal graph with mediation and a causal graph without mediation as equivalent.

Why is independent causal influence so important that its preservation warrants positing a hidden cause? The power PC theory (Cheng, 1997; Novick & Cheng, 2004) uses the assumption of independent causal influence as a defeasible default assumption to justify the inference of causal power. Without this assumption, the causal power of a candidate cause with respect to an effect is indeterminate even if the usual prerequisites for causal inference (e.g., "no

[1] Average causal ratings that are close to zero might be produced by (1) a roughly even mixture of "fewer" and "more" responses, or (2) many "same number" responses. Few participants reported that asmine juice causes headaches ($n = 2$); the answers close to zero were driven primarily by "same number" responses.

[2] Although the preventer is independent of the effect conditional on the cause being absent in narrow prevention but not broad prevention, causal Bayesian network models do not recognize this distinction. Causal Bayesian network models consider the conditional independencies of variables, not the conditional independencies at certain levels of variables. Since the preventer and effect are dependent for *some* values of the cause in both broad and narrow prevention, conditionalizing on the cause does not render them independent in the sense that causal Bayes nets consider when constructing a causal graph.

confounding") are satisfied. A difference in the probability of the effect in the presence of the cause and in its absence, for example, could be entirely due to the interaction between the cause and the context. If so, then there would be no reason to expect the cause to produce the effect in a different context. Indeed, without independent causal influence, causal power is bound to specific contexts, and causes will combine in unpredictable ways from one context to the next. This would render generalization unjustified. The assumption of independent causal influence jumpstarts the inference of causal power and supports generalization to transfer contexts via a context-independent causal power.[3]

Although people view causal mediation as a viable explanation for narrow prevention, the reason for this inference is less clear. There are at least two possibilities. First, people may posit causal mediation liberally, but only endorse causal mediation of a certain form. If so, participants in the broad prevention condition might be equally comfortable with causal mediation except that they prefer explanations where the mediator is common rather than rare. If this is the case, then people use the assumption of independent causal influence to infer the *form* of causal mediation. That is, the violation or non-violation of independent causal influence would determine whether people expect the mediator to be shared between different causes of the effect.

Alternately, people may posit mediation only when causal assumptions are violated. Since causal relationships can be decomposed almost indefinitely, this represents a reasonable strategy to minimize the complexity of causal explanations while maintaining useful assumptions. Broad prevention, which does not violate independent causal influence, can be explained and predicted without causal mediation. Therefore, positing causal mediation provides little practical benefit. For narrow prevention, however, the representation of mediation provides more tangible benefits: it allows people to generalize more accurately. If people can identify the mediator, they can infer whether the preventer will stop other causes from producing the effect. Thus, the violation of the assumption of independent causal influence serves as a criterion for revising one's causal explanation to achieve more accurate predictions.

In summary, narrow and broad prevention differ in whether they respect the assumption of independent causal influence. In narrow prevention, which violates independent causal influence, people view causal mediation as a plausible explanation. By positing causal mediation, people preserve the assumption of independent causal influence.

[3] The assumption of independent causal influence can be replaced, without changing the predictions regarding generalization, by the assumption that the causal factors in the background that interact with the targeted cause occur with the same probability across contexts (Cheng, 2000). Since our dependent measures do not allow differentiation between these assumptions, we treat them as equivalent for our purposes.

Acknowledgments

The preparation of this article was supported by AFOSR FA 9550-08-1-0489. The authors wish to thank Hannah Har for her assistance with data collection.

References

Carroll, C. D., & Cheng, P. W. (2009). Preventive scope in causation. In N.A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31th Annual Conference of the Cognitive Science Society* (pp. 833-838). Austin, TX: Cognitive Science Society.

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review, 104,* 367-405.

Cheng, P.W. (2000). Causality in the mind: Estimating contextual and conjunctive causal power. In F. Keil & R. Wilson (Eds.), *Explanation and cognition* (pp. 227-253). Cambridge, MA: MIT Press.

Gopnik, A, Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T. & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review, 111 (1),* 1-31.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction, *Cognitive Psychology, 51,* 334-384.

Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychological Review, 116 (4),* 661-716

Hagmayer, Y., & Waldmann, M. R. (2007). Inferences about unobserved causes in human contingency learning. *Quarterly Journal of Experimental Psychology, 60 (3),* 330-355.

Kant, I. (1965). *Critique of pure reason.* London: Macmillian. (Original work published 1781).

Keil, F. C. (2003). Folkscience: coarse interpretations of a complex reality. *Trends in Cognitive Sciences, 7 (8),* 368-373.

Liljeholm, M., & Cheng, P. W. (2007). When is a cause the "same"? *Psychological Science, 18 (11),* 1014-1021.

Lu, H., Yuille, A. L., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2008). Bayesian generic priors for causal learning. *Psychological Review, 115 (4)* , 955-984.

Luhmann, C. & Ahn, W. (2007). BUCKLE: A model of unobserved cause learning. *Psychological Review, 92 (3),* 657-677.

Novick, L. R., & Cheng, P. W. (2004). Assessing interactive causal influence. *Psychological Review, 111,* 455-485.

Pearl, J. (2000). *Causality: Models, reasoning, and inference.* Cambridge, England: Cambridge University Press.

Rottman, B. M. & Ahn, W. (2009). Causal inference when observed and unobserved causes interact. In N.A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31th Annual Conference of the Cognitive Science Society* (pp. 1477-1482). Austin, TX: Cognitive Science Society.

Saxe, R., Tenenbaum, J. B., & Carey, S. (2005). Secret agents: inferences about hidden causes by 10- and 12-

month-old infants. *Psychological Science, 16 (12),* 995-1001.
Schulz, L. E., & Sommerville, J. (2006). God does not play dice: Causal determinism and children's inferences about unobserved causes. *Child Development, 77 (8),* 427-442.
Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, prediction and search* (2nd ed.). Cambridge, MA: MIT Press. (Original work published 1993).

Edge replacement and nonindepedence in causation

David W. Buchanan (david_buchanan@brown.edu)
Department of Cognitive and Linguistic Sciences
Box 1978, Brown University, Providence, RI, 02912

Joshua B. Tenenbaum (jbt@mit.edu)
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology

David M. Sobel (dave_sobel@brown.edu)
Department of Cognitive and Linguistic Sciences
Box 1978, Brown University, Providence, RI, 02912

Abstract

Human beings show a robust nonindependence effect in causal reasoning: they predict that collateral effects should be correlated even given a common cause. This presents a problem for existing models of causal reasoning, as most predict independence. To deal with this problem, we propose an edge replacement process that builds up apparently probabilistic causal relations using hidden deterministic causes. This model allows us to fit nonindependence effects, and shows promise for modeling other phenomena, such as how causal relations change over time.

Keywords: Markov violations; nonindependence; causal reasoning; models of causal reasoning

Introduction

Causation is only as simple as we make it. Consider the example of sending an email to two colleagues: You push send, which causes them to see text on their screen. The relation seems simple enough, but in reality, there is a complex chain of events that connects cause and effect, which most of us understand only vaguely (Keil, 2003). These details are usually not worth considering, but they are useful when the causal relations fail. For instance, most of us know to check our spam filter when we fail to receive an expected email. Such details also tell you about correlations between events: If one colleague calls to say she has not received the message, you know to call the other one as well. Still, more detail is not always better – it would be absurd reason about email at a molecular level. Choosing the right level of detail is important, and human beings seem to do it easily. Models of causal inference must solve this problem as well.

Causal graphical models (hereafter, "CGMs"), (Pearl, 2000; Spirtes, Glymour, & Scheines, 2000) give us a language in which to express this problem formally. Under this framework, nodes in a graph represent events, and directed edges represent causal relations. Figure 1 gives an example of three graphs that capture the common cause scenario described above: person C sends an email, causing persons E_1 and E_2 to receive it. Under the assumptions of CGMs, unconnected nodes must be statistically independent, but otherwise there are a wide range of possible functional relationships that can be instantiated by an edge. There is also no limit to the number of hidden nodes that can exist in a graph.

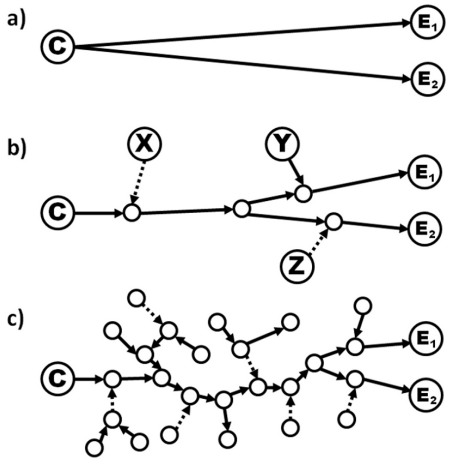


Figure 1: Examples of three different graphs that all capture the common cause relation. Minimality prescribes that we should begin by using a).

Thus, CGMs are enormously powerful, defining an infinite space of possible graphs for any given causal relation.

In order to make use of CGMs, we need some way of choosing which graph to use. The principle currently most used is called *minimality*: use the simplest graph that fits the data, in the sense that no other candidate graph has fewer edges. This means that given a common cause of two effects, the minimal graph is shown in Figure1a. It is often acknowledged (i.e. Pearl, 2000) that minimality creates problems, but in the absence of an alternative, it is widely used.

Within the literature on causal reasoning, the most acute problem with minimality is known as *nonindependence*. The simplest example of this phenomenon is found in a common cause scenario. If two effects of a common cause are related according to the minimal graph (Figure 1a) then the two effects should be independent given their common cause. That is, if C directly causes each of E_1 and E_2, then $P(E_1|C,\neg E_2)$[1]$= P(E_1|C) = P(E_1|C,E_2)$. If we see evidence that violates this expectation, then minimality allows us to use

[1]This notation means: The probability that E_1 occurs given that C occurred, but E_2 did not.

a slightly more complex graph that better fits the data. But according to minimality, independence should be our initial expectation.

Human beings do not have this expectation. In several experiments (Mayrhofer, Hagmayer, & Waldmann, 2008; Rehder & Burnett, 2005; Walsh & Sloman, 2004) participants robustly predict that $P(C|E_1, \neg E_2) < P(E_1|C) < P(E_1|C, E_2)$, even in novel scenarios, and even when independence is explicitly emphasized. Such nonindependence effects show that if people respect CGMs, they do not respect minimality. This raises the question of what principle, if any, people do respect.

Rehder and Burnett (2005) proposed an "underlying mechanism model" to address this problem, in which people represent hidden intermediate causal structure. In its current form, this model allows only qualitative fits to the data. Our model is one way of formalizing and extending Rehder and Burnett's proposal in order to make quantitative predictions. Further, Mayrhofer et al. (2008) modeled nonindependence effects using a source of common preventative noise, whose strength they fit to the data. Again, we hope to build on this initial step, and account formally for the source of this noise in a more principled way, while using fewer parameters to fit experimental data.

We propose a generative model of causation, which we call the *causal edge replacement process* (CERP). Theoretically, it is motivated by the hypothesis that causal reasoning involves representations of intermediate causal structure, or mechanisms (Shultz, 1982). Formally, CERP assigns a probability distribution to an infinite space of possible graphs, depending on how likely each is to be generated using repeated application of a specific edge replacement rule, and a restricted function set. The model's key contribution is that in the generative process, each edge has a *length*; longer edges tend to generate more hidden structure. While the graphs preferred by the model tend to be simple, they are not minimal. In particular, graphs generated by CERP have a characteristic branching structure that gives good quantitative fits to human data on nonindependence. CERP also provides a formal way of addressing questions about causal mechanisms.

We will begin by explaining exactly how the generative model operates, then show how the model fits three independently collected data sets, using the same parameter settings. Finally, we will discuss directions for future work.

Generative model

We will describe the generative model in three independent ways: In this section, we will give an informal verbal overview of the edge replacement process. Figure 2 also shows the process visually. Finally, we will describe the model in complete formal detail.

CERP begins with an edge between two nodes, which represents a causal relation between two events of interest. The process then moves down this edge, randomly generating replacements as it goes. Each replacement incorporates the influence of a new node. Because of the branching structure created by CERP, it is helpful to think of causation as flowing down a stream from causes to effects.

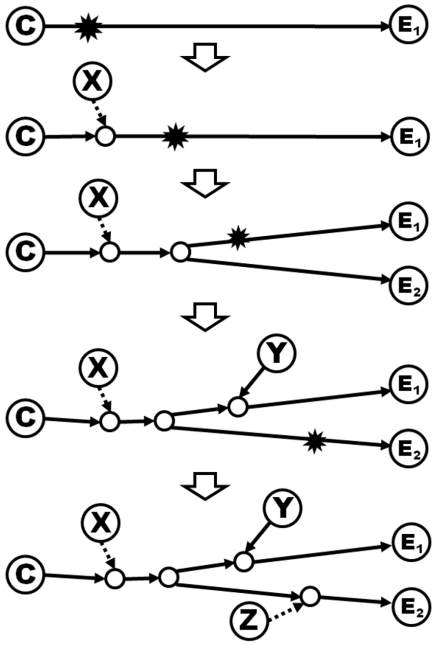


Figure 2: A series of edge replacements leading to a graph. The star indicates the location of the next replacement, as the process moves down each edge. Dashed edges indicate inhibitory relations.

To perform a replacement, we first replace the edge with an edge-node-edge path combination that has the same length. We call this middle node the "bridge node." Then we add a new edge (of randomly determined length) connecting the bridge node and a new node. The meaning of "length" here is more functional than physical – it does not correspond to the spatial distance between cause and effect, but to how vulnerable the relation is to other events. The replacement is randomly determined to be one of three types: If it is an inhibitory replacement, then causation at this point follows an AND NOT relation: causal power will flow through the bridge node only if the new node is off. For instance, in Figure 2, X is generated by an inhibitory replacement. This might be a failure in the server that sends your email. (These specific cover stories are not generated by CERP; they are only used to illustrate the principles involved.) Replacements can also be generative: For instance, E_2 is generated next – it fires whenever causation reaches the bridge node. We call these "side effects": For instance, sending an email may leave a record on the server. The third type of replacement is also generative, but causation flows inward rather than outward. We call these "alternative generative causes," because they follow an OR relation: the effect fires either if causation reaches the bridge node, or if the alternative generative cause fires. For instance, see Y in Figure 2: this might repre-

sent the fact that you can cause a given colleague to see the text on their screen via another method, like directing them to a website. In principle, outward inhibitory replacements are also possible, but are usually irrelevant – we only include them for the sake of formal completeness results discussed below. When the new node (not the bridge node) is generated, it is assigned some random probability of firing – this is the only source of randomness in the causal structures defined by CERP. Thus CERP is committed to determinism, in the weak sense that variability arises from hidden causes, not from intrinsic randomness in causal relations.

After a replacement, the same process continues along the old path, and along the new edge created by the replacement. Thus, graphs can become arbitrarily complex if replacements are common enough. The process eventually stops when it reaches the end of all edges, yielding a graph. We can "run" the graph by deciding (again, randomly) whether each exogenous node is on, then propagating causation deterministically through the graph.

Formal Description of CERP

This section describes the model in complete formal detail. Readers who are not interested in implementing CERP can skip this section. A graph with n nodes consists of the following components: 1. An $n \times n$ matrix G that encodes the causal relations (edges) between nodes. (1=generative, -1=inhibitory, 0=no relation) 2. An $n \times n$ matrix L of edge lengths. 3. A vector S of length n that encodes the spontaneous activation probabilities of each node. Together, $< G, L, S >$ defines a graph.

The generative process begins with an edge of length 1 between two nodes, which we will call A and B. We perform replacements by moving along each edge and generating replacements according to a Poisson process with rate λ. This is done by sampling x from $Exponential(1/\lambda)$. If $x > L(A,B)$, then stop. Otherwise do a replacement at point x: Create a new node M as the $(n+1)$th node. With probability ρ, designate a previously generated non-bridge node as E, otherwise create a new node E as the $(n+2)$th node (For our purposes in this paper, $\rho = 0$). Set $G(A,M) = 1$ and $G(M,B) = G(A,B)$. Set $L(A,M) = x$ and $L(M,B) = L(A,B) - x$. If E already exists, and it is exogenous, set $G(E,M)$, and if it is endogenous, set $G(M,E)$. Otherwise, with equal probability, choose to set either $G(E,M)$ or $G(M,E)$. Set this relation as -1 or 1 with equal probability. Also sample $L(M,E)$ or $L(E,M)$ from $Exponential(\gamma)$. Set $S(n+1) = 0$ and sample $S(n+2)$ from $Beta(\alpha, \beta)$. Finally, set $G(A,B) = 0$, eliminating the original edge. Initiate two new Poisson processes, along MB and ME, and repeat until all processes have stopped.

To sample from the graph, determine whether each node is on, according to S, then propagate causation deterministically through the graph to determine the values of each non-exogenous node. A node is on if and only if all of its inhibitory connections are off, and at least one of its generative connections is on, or it fires spontaneously. This instantiates the OR (for generative) and AND NOT (for inhibitory) functions originally applied to causation by Cheng (1997).

Completeness and Validity

We can use CERP to construct any logical relation: OR can be created by an inward generative replacement, AND NOT from an inward inhibitory replacement, while AND can be created by an inhibitory replacement on the negation of a variable. In particular, the inward inhibitory replacement acts as a "causal transistor," letting us construct a wide range of logical "circuits." We can also use the presence of a hidden inhibitor to generate any apparently probabilistic relation between two variables: To generate any $P(B|A)$, perform a hidden inhibitory replacement on the edge AB, with spontaneous activation probability $1 - P(B|A)$. Similarly, for any $P(B|\neg A)$, perform a hidden generative replacement with spontaneous activation probability $P(B|\neg A)$.

Yuille and Lu (2008) show that their noisy-logical graphs can capture any causal-functional relation. If we additionally allow CERP to reuse existing exogenous nodes when performing replacements (i.e. $\rho > 0$), then it is easy to see that CERP can be used to mirror any noisy-logical graph, as we can construct any logical or apparently probabilistic relation as described above. Thus, we can extend Yuille and Lu's (2008) completeness result to CERP. Some such relations will be generated with low probability by CERP, but all relations will have nonzero probability of being generated. Thus, CERP defines a prior distribution over the hypothesis space of all possible causal-functional relations.

The model also preserves validity: Because it introduces no undirected edges or cycles, it will always produce a directed acyclic graph when given a directed acyclic graph. To introduce a cycle, the model would have to introduce a path from a descendant to an ancestor. But this is not possible, because all new paths are either from nodes that have no ancestors, or to nodes that have no descendants.

Overall, CERP provides a way of expressing causal-functional relations using a compact set of rules. The restriction to deterministic OR and AND NOT functions means that complex relations must be expressed graphically, where the complexity is easier to see and measure, than it is in the complex conditional probability tables often used in existing instantiations of CGMs.

The model has two key components: the idea of edge replacement, and the use of deterministic causal relations. It is conceivable that we could use edge replacement with probabilistic relations. For instance, edges could begin with probabilistic values that change as replacements are made. However, this introduces a great deal of complexity, which is unwarranted unless necessary to fit human data. Given evidence (e.g. Schulz & Sommerville, 2006) that even children seem to be deterministic in the relevant sense, we believe it is a good assumption. Future work will focus on testing this determinism commitment directly. In this paper, we will instead focus on testing the structural predictions that arise primarily from the use of length in the generative process.

Fitting nonindependence effects

Walsh and Sloman (2004)

Walsh and Sloman (2004) showed a nonindependence effect that provides the simplest test of our model. They told adult participants simple common effect cover stories, in which event C caused both events E_1 and E_2. For instance, some participants were told that jogging (C) caused both increased fitness (E_1) and weight loss (E_2). They then asked participants to judge $P(E_1|C)$ and $P(E_1|C,\neg E_2)$. Figure 3 shows their data averaged across experiments. Participants reliably judged that $P(E_1|C) > P(E_1|C,\neg E_2)$. This is a nonindependence effect: If both effects are generated independently from the cause (as in Figure 1a), both values should be the same.

CERP's predictions are also shown in Figure 3. In fitting this and subsequent experiments, we used Monte Carlo sampling on causal structures as generated by CERP. One approach would be to generate a large set of graphs using CERP, keeping only the small subset that are consistent with the cover story and data presented to participants. In this case, we would accept only graphs that had exactly two visible effects. A sufficiently large sample will reflect the properties of the probability distribution defined by CERP. Such an approach is correct but computationally expensive, prohibitively so as we add complexity to the cover story.

We used a more efficient, but equivalent procedure: Begin with a single edge between the cause C and effect E_1 described to participants. Then generate the single visible side effect described to participants. This is equally likely to have been generated from any given point on the path from cause to effect, so we generate the second effect by choosing a random point $x \sim U[0,1]$.[2] Call the branch point M, and set the length of ME_1 to $(1-x)$ in order to ensure that E_1 and E_2 have the same path length from C and hence the same expected $P(E_n|C)$. This is in order to meet the condition, common in nonindependence experiments, that effects are equally likely given the cause. Because of this equivalence, the choice of initial effect in CERP is arbitrary. This process creates three edges: CM, ME_1, and ME_2.

At this point, we have generated all the visible causes and effects described to participants. Therefore, we are licensed in using a computational shortcut to do simultaneous inference over all the further (hidden) replacements that could be generated by CERP. In this case, all that matters are inward hidden replacements that occur along each edge. Active replacements on CM (like X in Figures 1 and 2) change both relations (creating a correlation); active replacements on ME_i (like Y and Z in figures 1 and 2) change only the relation CE_i, and inactive replacements have no effect.

We introduce the parameter h to describe a Poisson process that moves along the edge of interest, generating only active inward hidden replacements whose causal power actually reaches the path, with rate $-ln(h)$. This means that the probability of having zero active replacements along an edge of length l is h^l. We also introduce the parameter a to capture the probability that a given visible event fires spontaneously, as a result of causal processes not captured in the graph (Cheng, 1997). Together, h and a replace all the parameters described above in the full formal model. For instance, for most values of h and a, there are many settings of λ, γ, α and β that produce the same predictions. As long as we are not interested in the specific causal structure that generated each effect, this process is equivalent to generating a larger sample of more detailed graphs.[3] Crucially, it also uses fewer parameters.

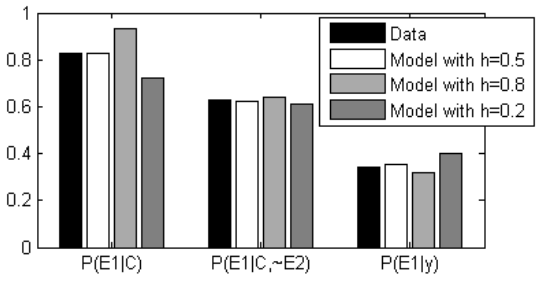


Figure 3: Data from Walsh and Sloman (2004), averaged over experiments, along with model predictions. Predictions are robust across alternative parameter settings.

After generating causal structures with branch points at various lengths, we used $h = 0.5$ and $a = 0.3$ to generate samples of the co-occurrence of the three events, by generating a set of replacements and propagating causation through the graph. Figure 3 shows the proportion of times that events occurred together, along with human probability judgments. We continued generating samples until we had at least 10000 samples for each entry. Predictions were resistant to changes in parameters; Figure 3 also shows other settings for h.

CERP can also easily make predictions about $P(E|\neg C)$, which Walsh and Sloman did not directly ask participants. However, they did ask participants a related question: the probability of an effect given a disabler that inhibited both effects (0.34). We sampled this by generating an active common inhibitor at a randomly chosen point $y \sim U[0,1]$, then choosing the branch point x from $U[y,1]$, because it would be incoherent for the branch point to occur before the common disabler. This gave us a $P(E|y)$ of 0.35. This value is lower than $P(E|\neg C)$ because there is less of the path remaining on which a generative cause could fire.

Overall, the model explains the data well. Because there were few data points (three) in comparison to the number of parameters in the model (two), we will look at more experiments using the same parameters that best fit these data.

[2] This means: "x was sampled from a uniform distribution between 0 and 1." We will use similar notation throughout the paper.

[3] We verified this by running the full generative model with a variety of parameter settings – several produced the same results as in Figure 3.

Rehder and Burnett (2005)

Another dataset is provided by Rehder and Burnett (2005), who found a nonindependence effect in the domain of feature inference. They told adult participants that one novel feature of a category (C), caused three other novel features ($E_{1,2,3}$). They then asked participants to judge the probability of one feature in a member of the category, given some value of C, and some number of other collateral effects. In Experiment 1, participants were given a cover story involving natural kinds. In Experiment 2, participants were not given a cover story at all – they were told that abstract features caused each other. Across multiple experiments, participants predicted that the values of collateral effects would be correlated, showing a nonindependence effect even with no cover story. The results of their Experiments 1 and 2 are shown in Figure 4.

We modeled this much like Walsh & Sloman, 2004, except that there were two branch points and thus five edges. Again, we ensured that all paths from C to E_n had the same length, because participants were told that all effects had the same probability. All parameters were the same as in fitting Walsh and Sloman: $h = 0.5$, and $a = 0.3$. Because there was always one node with a longer branch than the others, we also randomly permuted the role of each node.

Results are shown in Figure 4. The model provides a good fit to the data, especially when the cause is present. In the absence of the cause, the model predicts slightly higher probability judgments than participants' responses. This is probably due to the effect of categorization: Other data show that participants in this paradigm were significantly less likely to believe that a given instance was actually a member of the category when C was not present (Rehder & Burnett, 2005). This well-known "causal status effect" (Ahn, Kim, Lassaline, & Dennis, 2000) probably lowered their judgments of the other characteristic features of the category. Put another way, we assumed that feature C was uncaused, but participants may have assumed that all features had a hidden common cause that was present only in category members. CERP can model the effect of such an additional hidden common cause, but that was not our goal in the present investigation. We model an experiment below that replicated Rehder and Burnett's findings outside the domain of categorization, where we do not find this problem.

Mayrhofer, Hagmayer, and Waldmann (2008)

One strength of CERP is that it predicts how descriptions of the causal mechanism should affect the degree of nonindependence observed. Mayrhofer, Hagmayer, and Waldmann (2008) did just this. They told adult participants about four telepathic aliens; we will call them C, E_1, E_2 and E_3. The "cause" alien sometimes causes the "effect" aliens to think of food when he thinks of food. In the *transmit* condition, the instructions said that C sent his thoughts to each E_n, but sometimes C had difficulty concentrating. In the *receive* condition, instructions said that each E_n read the thoughts of C, but each effect alien sometimes had difficulty concentrating.

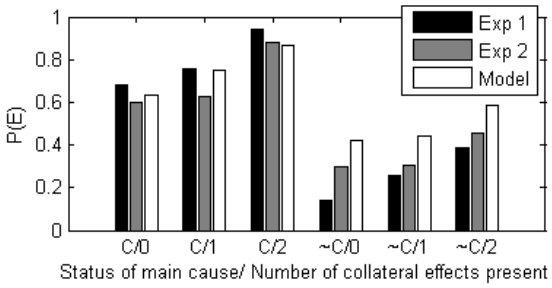


Figure 4: Data from Rehder and Burnett (2005) along with model predictions. Lower judgments in the absence of the cause feature are probably due to categorization, which is external to our model.

Like Rehder and Burnett (2005), they asked participants to judge $P(E)$ given different numbers of collateral effects. Data show a much stronger nonindependence effect in the transmit than in the receive condition (See Figure 5).

As before, we used a process which was equivalent to generating a large sample of graphs, and keeping only those consistent with the cover story. The cover story describes three similar effects, so we began by generating all three from the same branch point $x \sim U[0,1]$. We generated an inhibitor, explicitly mentioned in the cover story ("failure concentrating") at point $y \sim U[0,1]$, assigning it probability $a = 0.3$ of firing. In the receive condition, we kept only those samples in which $y > x$, since the inhibitor was described as applying to each alien individually. We generated one instance of the inhibitor on each branch. In the transmit condition, we kept only samples in which $y < x$, since only one inhibitor was described. We know of no other way to generate graphs consistent with both CERP and the cover story. Otherwise, we sampled as before, using $h = 0.5$ and $a = 0.3$.

As shown in in Figure 5, the model provides a good fit to the data. One exception is the point in the transmit condition in which two collateral effects occur, but the cause does not (the last entry in the "transmit" graph): The model predicted a medium probability judgment, while participants gave a low judgment. This may be due to a random anomaly in human responses, because the data are hard to explain under any account: As collateral effects were added, participants lowered, rather than raised, their probability judgment. This is not replicated in any other condition or experiment.

Mayrhofer et al. fit the qualitative difference between the conditions by adjusting a quantitative parameter: strength of inhibitory noise, which was strong in the transmit condition and weak in the receive condition. As they show, this parameter can be used to fit a wide range of data. Our model used a qualitative structural change instead, while the quantitative parameters have relatively little effect on the predictions, and remained constant between conditions and experiments. The model captures how changes to the mechanism description change the source and structure of the noise.

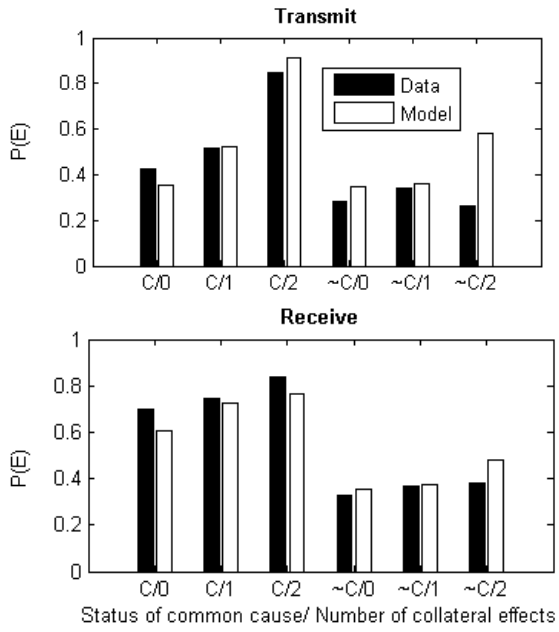


Figure 5: Data from Mayrhofer et al. (2008) along with model predictions. No parameters were varied between conditions –only the constraints given by the different cover stories.

Conclusion and Further Work

We use CERP to fit three independently collected data sets on nonindependence, using the same parameters between experiments, and even between conditions within experiments. Over all three data sets, we fit 21 data points using 2 free parameters, with a correlation of greater than 0.99.[4] The power of CERP seems to come not from its use of free parameters, but from the fact that structural aspects may mirror some important aspect of the way that human beings represent causation. Further work will focus on exploring these aspects more closely. For instance, we can generalize our explanation for Mayrhofer et al., 2008's data to make a novel predictions: Early inhibitors in a causal stream should create more nonindependence than late inhibitors. We call this a *stream location effect*. We have recently tested this on preschoolers, with positive results (Buchanan & Sobel, 2010).

Our main intent with CERP is to test predictions that go well beyond nonindependence effects. For instance, its commitment to a form of determinism (namely, that apparent randomness always comes from hidden causes) has implications for how we reason about data that varies over time. Imagine your car fails to start one morning. Is it more likely to start tomorrow morning, or on a morning one year from now? If the relation were truly random, there should be no difference in judgment between these two times. If we introduce time into CERP, it should be able to rationally justify and fit our intuition that the car is more likely to start a year from now, than it is tomorrow. This is because variability arises from hidden causes that have persistence in time and space.

Finally, because it can generate any functional relation, CERP represents one way of defining a prior distribution over logical graphs. This may be useful to researchers (i.e. Lucas & Griffiths, 2010) who are interested in how people learn about the functional form of causal relations. An interesting question that arises from this research program is whether something like CERP could itself be learned – for instance, children might start with more general causal expectations, and come to realize that the world follows some or all of the commitments of CERP, such as determinism, and the stream-like character of causation.

[4] In data sets with multiple experiments, we correlated the model's predictions with the average over the experiments.

Acknowledgments

Supported by NSF (DLS-0518161 to DMS). Thanks to Noah Goodman for discussions. Thanks also to Ralf Mayrhofer and Bob Rehder for generously sharing details of their data.

References

Ahn, W., Kim, N., Lassaline, M., & Dennis, M. (2000). Causal status as a determinant of feature centrality. *Cognitive Psychology*, *41*(4), 361–416.

Buchanan, D., & Sobel, D. (2010). Causal stream location effects in preschoolers. In *Proceedings of the 32nd annual conference of the cognitive science society.*

Cheng, P. (1997). From covariation to causation: A causal power theory. *Psychological Review*, *104*(2), 367–405.

Keil, F. (2003). Folkscience: Coarse interpretations of a complex reality. *Trends in Cognitive Sciences*, *7*(8), 368.

Lucas, C., & Griffiths, T. (2010). Learning the form of causal relationships using hierarchical Bayesian models. *Cognitive Science*, *34*(1).

Mayrhofer, R., Hagmayer, Y., & Waldmann, M. (2008). Violations of screening off: A Bayesian error attribution model of causal reasoning. *Unpublished presentation at Mathematical Psychology 2008*.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

Rehder, B., & Burnett, R. (2005). Feature inference and the causal structure of categories. *Cognitive Psychology*, *50*(3), 264–314.

Schulz, L., & Sommerville, J. (2006). God does not play dice: Causal determinism and preschoolers' causal inferences. *Child development*, *77*(2), 427–442.

Shultz, T. (1982). Rules of causal attribution. *Monographs of the society for research in child development*, 1–51.

Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, prediction, and search*. The MIT Press.

Walsh, C., & Sloman, S. (2004). Revising causal beliefs. In *Proceedings of the 26th annual conference of the cognitive science society.*

Yuille, A., & Lu, H. (2008). The noisy-logical distribution and its application to causal inference. *Advances in neural information processing systems*, *20*, 1673–1680.

Agents and Causes:
A Bayesian Error Attribution Model of Causal Reasoning

Ralf Mayrhofer (rmayrho@uni-goettingen.de)
York Hagmayer (york.hagmayer@bio.uni-goettingen.de)
Michael R. Waldmann (michael.waldmann@bio.uni-goettingen.de)
Department of Psychology, University of Göttingen,
Gosslerstr. 14, 37073 Göttingen, Germany

Abstract

One of the most fundamental assumptions underlying causal Bayes nets is the Markov constraint. According to this constraint, an inference between a cause and an effect should be invariant across conditions in which other effects of this cause are present or absent. Previous research has demonstrated that reasoners tend to violate this assumption systematically over a wide range of domains. We hypothesize that people are guided by abstract assumptions about the mechanisms underlying otherwise identical causal relations. In particular, we suspect that the distinction between agents and patients, which can be disentangled from the distinction between causes and effects, influences which causal variable people blame when an error occurs. We have developed a causal Bayes net model which captures different error attributions using a hidden common preventive noise source that provides a rational explanation of these apparent violations. Experiments will be presented which confirm predictions derived from the model.

Keywords: causal reasoning; Bayesian modeling; Bayes nets; Markov condition.

Introduction

Causal Bayes net theory is an increasingly popular approach to model causal reasoning in humans, especially in domains in which multiple variables are causally interrelated. Causal Bayes nets can be graphically represented as sets of (observable and hidden) variables that may represent present or absent events, and arrows that express the direction of the causal influences between the interconnected variables (for an example, see Fig. 1).

To make inferences in this network, additional assumptions need to be made about how the three arrows interrelate. A central assumption that turns probabilistic networks into Bayes nets is the Markov condition (see Pearl, 2000). The Markov condition states that for any variable *X* in a set of variables *S* not containing direct or indirect effects of *X*, *X* is jointly independent of all variables in *S* conditional on any set of values of the set of variables that are direct causes of *X*. An effect of *X* is a variable that is connected with a single arrow or a path of arrows pointing from *X* to it. The Markov condition implies in the common-cause model that each effect is independent of all the other effects conditional upon the presence or absence of its cause *C*.

The Markov condition provides Bayes nets with substantial computational power. Assuming conditional independence allows for learning and reasoning about subsets of variables while ignoring the states of other independent variables. For example, we can infer the presence or absence of an effect from the state of its cause without having to consider the states of the other conditionally independent effects. When using Bayes nets we are not forced to believe that in every situation effects of a common cause are conditionally independent. Whenever we have reasons to question this assumption, it is possible to model violations by adding hidden variables (again obeying the Markov constraint) representing unobserved causal influences. However, the validity of the Markov condition is typically assumed as a default unless we have domain knowledge that suggests hidden variables.

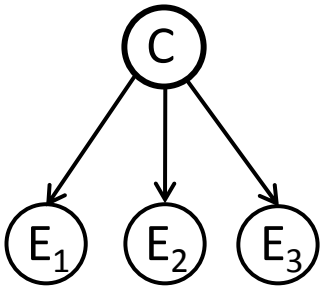


Figure 1: An example of a simple common-cause structure with a cause variable *C* and three effect variables E_1, E_2, E_3. The state of each effect variable depends statistically only on the state of the cause variable.

Rehder and Burnett (2005) developed a reasoning task which allowed for testing people's intuitions about the Markov condition. For example, subjects had to rate the conditional probability of an effect's presence given the state of its cause C. The crucial manipulation was whether other effects of C were present or absent. According to the Markov condition subject's ratings should be invariant across these conditions. Contrary to this prediction, the ratings were clearly sensitive to the states of other effects of C. The more collateral effects were present, the higher the rating of the conditional probability of the target effect given the presence of C. This Markov violation was extremely robust across many cover stories and domains.

Walsh and Sloman (2007) followed up on this research. They were interested in the boundary conditions of violations of the Markov condition. In one experiment they

presented subjects with a common-cause model in which loud music in an apartment building represented the common cause of the complaints of the neighbors on the left and the right side of the apartment in which the music was playing. Again the crucial test question referred to a case in which loud music was playing but the left neighbor was not complaining. According to the Markov condition this should not affect the rating of the likelihood that the right neighbor is complaining. However, Walsh and Sloman reasoned that the likelihood that complaints of the right neighbor are predicted should depend on the ad hoc explanations of why the left neighbor did not complain. If subjects were instructed that all neighbors were invited to the apartment in which music was playing, subjects should expect both neighbors not to complain (i.e., Markov violation). In contrast, when subjects were told that the left neighbor has left the building there is no reason to expect that the second neighbor will not complain (i.e., no Markov violation). The experiments confirmed these predictions although there was a fairly strong tendency to violate the Markov condition in all conditions. In this experiment the difference between the inferences is due to the fact that the initial causal model was differently augmented and changed in the contrasted conditions by adding further causal variables. In one condition an additional causal event, the invitation, was introduced, in the other condition one effect was effectively removed from the model, thus deleting its diagnostic relevance.

Agents and Causes

We are also interested in conditions moderating the degree of Markov violations. Whereas Walsh and Sloman (2007) have shown that different models containing different kinds of disabling events influence the inferences, our goal is to study the influence of assumptions about causal mechanisms while keeping the causal model on the surface level invariant. Causal Bayes nets combine assumptions about causal mechanisms with probabilistic covariations, but the assumed mechanisms are not elaborated. Tellingly, Pearl (2000) describes causal arrows as *mechanism placeholders*. Although recent empirical studies have casted doubt on the assumption that people have elaborate knowledge about mechanisms (Rozenblit & Keil, 2002), recent research on causal reasoning and language understanding has suggested that people may have abstract notions of basic properties of mechanisms (see Talmy, 1988; Wolff, 2007). Particularly relevant in the present context is the distinction between *agents* and *patients*, which is one of the important distinctions in our causal semantics introduced by Talmy. Agents are causal events that we represent as active in the generation of a causal relation. Patients are passive recipients of causal power. For example, in the familiar Michotte task the ball pushing the second ball is viewed as an agent endowed with force, whereas the ball that is being pushed is represented as a patient exerting resistance (White, 2009).

Agents and causes typically fall together but can be separated. Consider the example of tuners that receive music from a music station. Within a causal Bayes net the station would play the role of a common cause because sending out waves precedes the reception by tuners. However, depending on the focus, it is possible to view the sender as active senders and the tuners as passive receivers, or it is possible to highlight the active role of the tuners as receivers without whom no music can be heard. Thus, effects in a common cause model can be agents or patients depending on the framing. Our key prediction is that the agent role is associated with attributions of causal responsibility and blame. If something goes wrong in a causal transmission, then the agent will be the primary target of error attributions.

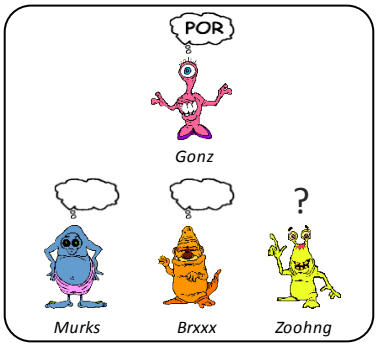


Figure 2: An example of a test item used in Waldmann et al. (2007).

Pilot Study

In an initial experiment, we tested this theory (Waldmann et al., 2007). Subjects were presented with instructions about four aliens, Gonz, Murks, Brxxx, and Zoohng, who mostly think of nothing and sometimes think of "POR" (food in alien language; material adapted from Steyvers et al., 2003). In one condition it was pointed out that Gonz is able to transmit its thoughts into the heads of the other alien (sending condition). In the contrasting condition it was pointed out that Murks, Brxxx, and Zoohng are able to read the thoughts of Gonz (reading condition). So, in both conditions the thoughts of Murks, Brxxx, and Zoohng are statistically and causally dependent on the thoughts of Gonz. Hence, both cases can be represented as a common-cause network (see Fig. 1; Gonz as the cause *C* and Murks, Brxxx, and Zoohng as the effects E_1, E_2, and E_3). However, the agent role was manipulated across conditions. Whereas in the sending condition cause and agent fall together, in the reading condition the effects were framed as agents. In the test phase subjects were requested to rate the conditional probability of a target alien, e.g., Zoohng, thinking of POR given the thoughts of the cause and the other effect aliens (for an example, see Fig. 2). Interestingly, the "Markov violation" was significantly stronger in the sending condition than in the reading condition (see interaction of upper two lines in Fig. 3), which confirms our prediction

that errors are associated with the agent. If there is only one agent (sending condition) then the failure of one of the receiving aliens to read his thoughts becomes diagnostic for a failure of the sending agent that also should affect the other aliens. In contrast, if the effects are represented as agents, then the error attributions should be locally attributed to the respective effect. The failure of one reader to read the thoughts of the cause alien should not predict whether the other readers will also fail or not.

Another important finding of the pilot study which we will follow up in Experiment 1 is that Markov violations in the sending condition were only observed when the cause was present but not when the cause was absent (see lower two lines in Fig. 3). Intuitively this can be interpreted as evidence for the assumption that sending errors can only occur when the cause alien is trying to send.

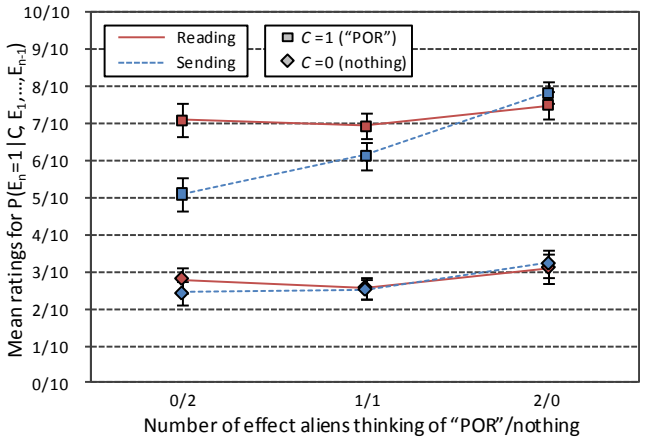


Figure 3: Mean ratings (and standard error) representing the estimates of the relative number of times the target alien thinks of "POR" in ten fictitious situations. The X axis represents the number of collateral effect aliens thinking of "POR". The upper two lines correspond to the cause alien thinking also of "POR", the lower two lines to the cause alien thinking of nothing. The dashed lines indicate the sending condition, whereas the solid lines indicate the reading condition.

In the next section we report a model that captures our intuitions about the role of agents in causal models. Subsequently we will report experiments testing the model.

A Bayes Net Model of Error Attributions

In Bayes nets, errors which are due to hidden mechanisms can be represented by hidden nodes in the network. We propose that *each cause* contains a hidden *common preventive noise* (PN) node which is connected to all effects, and can therefore alter the influence of the causes on their effects. Hence, in common-cause model there is one *PN* attached to its effects. This common noise source summarized *all* influences which potentially decrease the ability of the cause to bring about its effects (e.g., common preventer; missing enabling conditions, etc.)[1] (see Fig. 4). The strength of this noise source (w_{PN}) and its a priori base rate are domain dependent. In the sending condition, we assume that w_{PN} is pre-set to high values, thus increasing the influence of common preventive noise. In the reading condition, people should primarily attribute errors to the error links that are attached to each effect node and that are in Bayes nets assumed to be independent of each other. Thus, different parameterizations of w_{PN} explain the different degrees of Markov violations in the sending versus reading conditions.

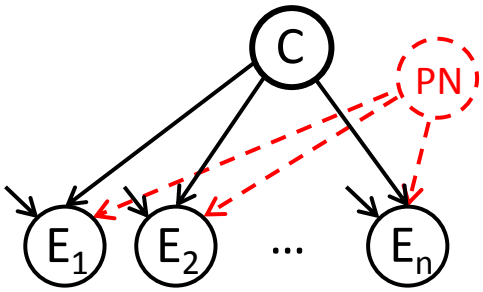


Figure 4: A simple common-cause structure extended by an unobserved common preventive noise node *PN*. The preventive noise interacts with the causal influence of *C*. If *PN* is present the power of *C* is lowered for all its effects. Thus, if $E_1, \ldots, E_{n-1}$ are observed as absent, even if the cause *C* is present, the presence of *PN* is likely. This lowers the predicted probability of E_n being present.

Asking people to judge the probability of a target effect alien (E_n) thinking of POR given the thoughts of the other aliens (C, E_1, …, E_{n-1}) is formally equivalent with asking the conditional probability of E_n: $P(E_n = 1|C, E_1, \ldots, E_{n-1})$. In a regular common cause structure (without a common noise source) this question simplifies to $P(E_n = 1|C)$ due to the Markov condition: The presence of the target effect only depends on the state of the cause, not on the states of the collateral effects. Introducing an unobserved common preventive noise node and integrating it out leads to the following derivation[2]:

$$P(E_n = 1|C, E_1, \ldots, E_{n-1}) = \sum_{PN} P(E_n = 1|C, \boldsymbol{PN}, E_1, \ldots, E_{n-1}) \cdot P(\boldsymbol{PN}|C, E_1, \ldots, E_{n-1}) = \sum_{PN} P(E_n = 1|C, \boldsymbol{PN}) \cdot P(\boldsymbol{PN}|C, E_1, \ldots, E_{n-1})$$

The second simplifying step in this derivation is possible because in the network with the common preventive noise

[1] Note that this is a specific preventive cause which does not affect the probability of the effect when the cause is absent.

[2] Actually, also the prior assumptions of the parameter values given by a set of Beta distributions are integrated out. To simplify the discussion we left this out in the description. The complete derivation includes a multiple integral over the parameter vector: $P(E_n = 1|C, E_1, \ldots, E_{n-1}) = \int P(E_n = 1|C, E_1, \ldots, E_{n-1}, \mathbf{w}) p(w) dw$

node the Markov condition holds: Given C *and* PN the target effect E_n is independent of the collateral effects. Thus, reasoning in this simple model can be thought as a two-step process: First the state of the noise is inferred, and then given that state (and given the a priori state of C) the state of the unobserved target effect is inferred.

The model predicts that inferences about the presence of an unobserved target effect in the presence of the cause should be influenced by the number of collateral effects that are present or absent. Absent effects in the presence of the cause should via the PN lower the ratings for the target effect. This influence should increase with increasing numbers of absent effects when the cause is present. When the cause is absent, however, no such pattern should be observed.

Experiment 1

When the cause varies between present (i.e., active) and absent (i.e., inactive), the model predicts an asymmetric influence of *PN* since in the cause's absence the *PN* cannot prevent *C* to bring about the target effect. Thus, the Markov violation in the sender condition should only be observed when the cause is present. In our pilot experiment we have indeed confirmed this prediction. In contrast, our model predicts a symmetric influence of the *PN* when the cause has two distinct but causally active states (i.e. A/B instead of 0/1). This prediction is tested in Experiment 1.

Method

Participants 56 students from the University of Göttingen participated in exchange for candy.

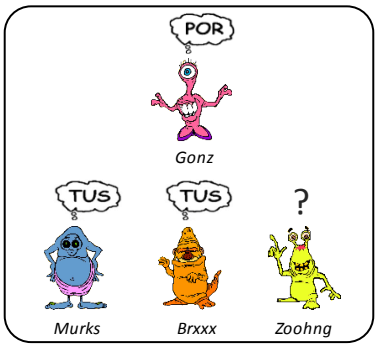


Figure 5: An example of a test item used in Experiment 1.

Procedure and Material In the instruction phase we presented subjects with instruction about four aliens: Gonz, Brxxx, and Zoohng, who usually think of "TUS" and sometimes think of "POR" (indicated by a bubble containing "TUS" or "POR"; see Fig. 5) (POR and TUS were counterbalanced). In two conditions it was either stated that the upper alien can transmit both thoughts to the lower three (sending condition), or that the lower three aliens can read the thoughts of the upper one (reading condition). It was pointed out that the effect aliens frequently think of "POR" or "TUS" when the cause alien thinks of "POR" or "TUS".

In the test phase, subjects were presented with six test panels with all the non-target aliens thinking of either "POR" or "TUS" (for an example, see Fig. 5). The order of test panels was randomized. For each panel, subjects were asked to imagine ten situations with the given configuration, and then to judge in how many of these situations the target alien (indicated by a question mark above its head) would probably think of "POR". This way we obtained probability assessments from the subjects.

Design The predictions were tested in a 2×2×3 ANOVA design with "sending" vs. "reading" as a between-subjects factor. The state of the cause alien ("POR" or "TUS" thoughts) and the number of collateral effect aliens thinking of "POR" (0, 1, or 2) were manipulated within subjects.

Results and Discussion

Figure 6 displays the results for Experiment 1. In general, the ratings for the target effect alien thinking of "POR" were higher when the cause alien thinks of "POR" ($F_{1,54}$=146.05, p<.001, η_p^2=.73).

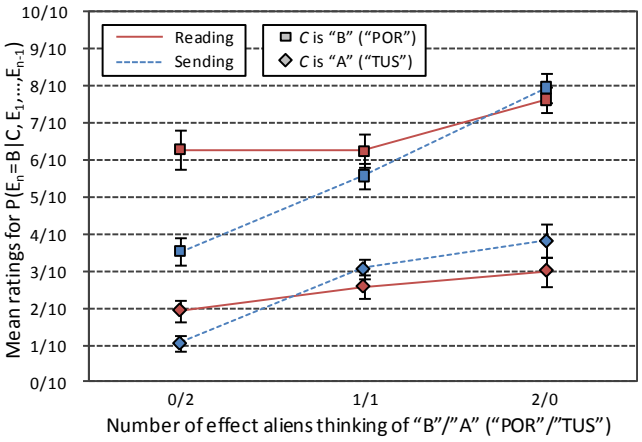


Figure 6: Mean ratings (and standard error) representing the estimates of the relative number of times the target alien thinks of "POR" in ten fictitious situations. The X axis represents the number of collateral effect aliens thinking of "POR". The upper two lines correspond to the cause alien thinking of "POR", the lower two lines to the cause alien thinking of "TUS". The dashed lines indicate the sending condition, whereas the solid lines indicate the reading condition.

As predicted by the model, people's judgments were symmetrically influenced by the states of the other effects for both states of the cause: In case of *C* representing "POR" (the upper two lines in Fig. 6) the ratings substantially increased with the number of effect aliens thinking of "POR" ($F_{2,108}$=31.47, p<.001, η_p^2=.37). As in our pilot study this influence was stronger in the sending condition than in

the reading condition yielding a significant interaction ($F_{2,108}$=8.94, p<.001, η_p^2=.14). In case of *C* representing "TUS" (the lower two lines in Fig. 6) the ratings also increased the more effect aliens thought of "POR" ($F_{2,108}$=20.25, p<.001, η_p^2=.27). As predicted by the model and in contrast to what we observed for the absent state of the cause in our pilot study the influence of the collateral effects was also stronger in the sending condition than in the reading condition when the cause alien thought of "TUS" ($F_{2,108}$=4.20, p<.05, η_p^2=.07). The descriptively weaker two-way interaction in the "TUS" case is predicted by the model as a consequence of the low base rate of "TUS". No three-way interaction was obtained, as predicted ($F_{2,108}$=1.37, p=.26).

The results confirm our model. Subjects' inferences were influenced by the location of the agent (sending vs. reading) in a fashion predicted by the error attribution model. Moreover, the model's predictions about the type of states of binary causal variables were confirmed. Our patterns in the sending condition correspond to the findings of Rehder and Burnett (2005), who also found symmetric influences of the states of other effect variables for both states of the cause. Although Rehder and Burnett described these states as present and absent, the two states in their experiments actually also represented two active states on a continuous dimension (typical vs. atypical).

Experiment 2

In our model, the common preventive noise node *PN* is attached to the specific cause it regulates. Therefore, in a causal chain structure each causal link should have its own *PN* node (see Fig. 7). This entails that the strength of each *PN* in the chain should not bias people's assumptions about the states of other variables. Consequently, our model predicts that in causal chain structures no Markov violation should be observed and that manipulations of people's assumptions about the location of the agent (i.e., sending vs. reading) should not have any effect. This prediction is tested in Experiment 2.

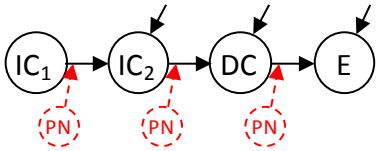


Figure 7: An extended causal chain model with two indirect causes (IC_1, IC_2), a direct cause (*DC*) and a final effect (*E*). Since the preventive noise (*PN*) is part of the causal process and therefore attached to each direct cause-effect relation, in each cause variable has its own preventive noise source.

Method

Participants 50 students from the University of Göttingen participated in exchange for candy.

Procedure and Material As in Experiment 1, we presented subjects with instruction about four aliens: Gonz, Brxxx, and Zoohng, who—as in the basic experiment—usually think of nothing and sometimes think of "POR" (indicated by an empty bubble or a bubble containing "POR", respectively; see Fig. 8). It was pointed out that—in the sending condition—an alien can transmit its "POR"-thoughts to its right neighbor or—in the reading condition—an alien can read the "POR"-thoughts of its left neighbor. Again it was stated that effect aliens frequently think of "POR" when the corresponding cause alien (the left neighbor) also thinks of "POR".

In the test phase subjects were presented with six test panels with the non-target aliens thinking of "POR" or nothing (for an example, see Fig. 8). The order of test panels was randomized. The target alien was generally the right most alien in the chain. As in Experiment 1, subjects were asked to judge in how many of ten situations the target alien would probably think of "POR".

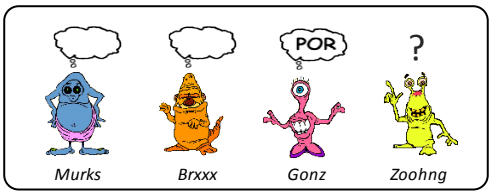


Figure 8: An example of a test item used in Experiment 2.

Design The predictions were tested in a 2×2×3 ANOVA design with "sending" vs. "reading" constituting a between-subjects factor and the state of the direct-cause alien ("POR" or nothing) as well as the number of indirect-cause aliens thinking of "POR" as within-subjects factors (0, 1, or 2).

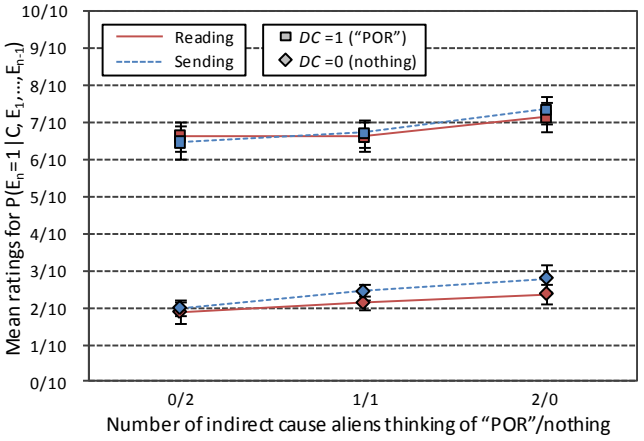


Figure 9: Mean rating (and standard error) of number of times the target aliens thinks of "POR" in ten fictitious situations plotted against the number of indirect-cause aliens thinking of "POR" (columns). The upper two lines correspond to the direct cause (*DC*) alien thinking also of "POR", the lower two lines to the direct cause alien thinking of nothing. The dashed lines represent the sending

condition, whereas the solid lines represent the reading condition.

Results

The results of Experiment 2 are shown in Figure 9. As in Experiment 1, the ratings for the target effect alien thinking of "POR" were higher when the direct-cause alien also thought of POR ($F_{1,48}$=191.99, p<.001, η_p^2=.80).

The prediction that different assumptions about the agents in the chain (sending vs. reading) should not matter was clearly supported. As predicted by the model, the sending vs. reading manipulation revealed no interaction with the states of the non-direct causes, neither in the presence of the direct cause ($F_{2,96}$<1, p=.5) nor in its absence ($F_{2,96}$<1, p=.66). However, in contrast to the predictions, significant, although very weak violations of the Markov condition in both the presence (the upper two lines in Fig. 9; $F_{2,96}$=11.77, p<.001, η_p^2=.20) as well as the absence of the direct causes (the lower two lines in Fig. 9; $F_{2,96}$=6.47, p<.01, η_p^2=.12) could be seen (see also Rehder & Burnett, 2005). The three-way interaction was clearly not significant ($F_{2,96}$<1, p=.99).

Discussion

The results of Experiment 2 show sensitivity to the instructed causal model and support the assumption inherent in our Bayesian model that preventive noise sources are attached to specific causes. Hence, whether preventive noise predicts error correlations is dependent on the underlying causal structure in which these nodes are an intrinsic property of each cause-effect relations.

However, our model cannot account for the small but still significant Markov violations in the data. Possibly subjects doubt that chain variables fully screen off previous influences or there are additional assumptions underlying causal chain representations.

General Discussion

Traditional causal theories view causes as endowed with the power to generate effects. However, little is known about how the mechanisms relating causes and effects are represented, and what influence assumptions about the mechanisms have on causal inferences. We have pinpointed one relevant factor, the distinction between agents and patients which can be separated from the distinction between causes and effects. We have used the example of sending versus reading to disentangle the location of the agent from the location of the cause. Our main hypothesis is that people tend to attribute potential errors to agents rather than patients. This intuition was formalized in a Bayesian model of error attribution which adds hidden preventive noise nodes to capture our intuitions about sources of error. Interestingly, this model explains violations of the Markov condition using a model that honors the Markov condition. Two experiments were conducted which tested and largely confirmed specific predictions of the model.

Traditionally there has been a conflict between covariation and mechanism (or force) theories. The present research shows that it is fruitful to combine the two approaches. Causal models are needed to guide processing of statistical covariations in data. However, the simple assumptions typically underlying these models are insufficient because additional knowledge about the mechanism seems to influence both the assumed hidden and observed structure of the model and the parameterization (see Mayrhofer et al., 2008). Future research will have to further elaborate the intricate relation between mechanism assumptions and causal models.

Acknowledgments

We wish to thank Marie-Theres Kater and Mira Holzer for assistance in data collection, and Noah Goodman, Josh Tenenbaum, and Tom Griffiths for helpful comments on the project. This research was supported by a research grant of the Deutsche Forschungsgemeinschaft (DFG Wa 621/20).

References

Mayrhofer, R., Goodman, N. D., Waldmann, M. R., & Tenenbaum, J. B. (2008). Structured correlation from the causal background. In *Proceedings of the Thirtieth Annual Conference of the Cognitive Science Society* (pp. 303-308).

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge, MA: Cambridge University Press.

Rehder, B., & Burnett, R. (2005). Feature inference and the causal structure of categories. *Cognitive Psychology*, *50*, 264-314.

Rozenblit, L., & Keil, F. C. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science*, *26*, 521-562.

Steyvers, M., Tenenbaum, J. B., Wagenmakers, E-J., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science*, *27*, 453-489.

Talmy, L. (1988). Force dynamics in language and cognition. *Cognitive Science, 12*, 49-100.

Waldmann, M. R., Mayrhofer, R., & Hagmayer, Y. (2007). *Mind reading aliens: Causal capacities and the Markov condition*. Unpublished manuscript.

Walsh, C. R., & Sloman, S. A. (2007). Updating beliefs with causal models: Violations of screening off. In M. A. Gluck, J. R. Anderson & S. M. Kosslyn, *A Festschrift for Gordon H. Bower* (pp. 345-358). New York: Erlbaum.

White, P. A. (2009). Perception of forces exerted by objects in collision events. *Psychological Review*, *116*, 580-601.

Wolff, P. (2007). Representing causation. *Journal of Experimental Psychology: General*, *136*, 82-111.

Modeling strategies in Stroop with a general architecture of executive control

Tomasz Smoleń (bertrand@o2.pl)
Adam Chuderski (a.chuderski@emapa.pl)
Institute of Psychology, Jagiellonian University
al. Mickiewicza 3, 31-120 Krakow, Poland

Abstract

This paper presents a preliminary work on a new architecture of executive control (DUCCA), aimed at integration and extension of some leading approaches to executive control. We present DUCCA assumptions and operation and use the architecture to simulate a few effects observed in Stroop-like task, a hallmark test of how control deals with interference. The focus of DUCCA is on how strategical use of general executive mechanisms contributes to Stroop effect. We explain also what is usually neglected in Stroop modeling: the significant individual differences in task performance.

Introduction

Executive control is implemented via numerous brain mechanisms and on different levels of neuronal organization. However, a few general flexible control mechanisms, which are involved in most of situations that require control, were also proposed in control literature (Anderson, Fincham, Qui, & Stocco, 2008; Braver, Gray & Burgess, 2007; Kane & Engle, 2003; Koechlin & Summerfield, 2007). The goal of this paper is to present a new model of executive control, called Dual Cognitive Control Architecture (DUCCA), which integrates several recent theoretical approaches to control and extends them with a few original control mechanisms. The model explains crucial effects related to interference control in Stroop-like tasks with an appeal only to general mechanisms of control, while abstracting from specific (e.g., semantical or stimulus-related) ones.

The first general function of control regards using contextual, episodic, or goal information in order to change the probability distribution of alternative actions into a one that maximizes their task-relevance (Anderson et al., 2008; Koechlin & Summerfield, 2007). Such a function is implemented in cognitive models of executive processing in two ways. In most of connectionist models, a network carrying out non-executive processing is supplemented with task (or goal) units, which modulate processing by propagating additional activation to nodes relevant to a respective task (e.g., Altmann & Davidson, 2001; Cohen, Dunbar & McClelland, 1990; Verguts & Notebaert, 2008). The control in symbolic architectures is usually implemented as control signals, stored in a goal or working memory buffer, which are matched to possible actions in order to select a next operation (Anderson et al., 2008; Meyer & Kieras, 1997). The first original aspect of DUCCA is that it integrates these two approaches into the unitary, general mechanism of top-down control, which may either directly select an action or just modulate a chance of its selection.

The second important function of executive control deals with regulation of its strength, as maintaining control for long periods of time is metabolically costly and often cognitively inefficient. Early observations indicated that control is amplified after errors. However, results like Gratton effect (i.e., the interference cost in a flanker task is 20 ms smaller in trials following incongruent stimuli, compared to ones following congruent stimuli; Gratton, Coles, & Donchin, 1992), usually observed even if errors are rare, suggested that control can be dynamically modulated on some other basis. Botvinick et al.'s (2001) conflict monitoring theory states that specialized brain mechanism (anterior cingulate cortex; ACC) performs online computing of the level of conflict between alternative responses and it increases the strength of top-down control as such a conflict arises. A more general idea is that ACC learns and reacts to a level of "risk" – conflict related error likelihood and its real-world consequences (Brown & Braver, 2007). Both cited models, however, evaluate only response representations in performing need-for-control monitoring, while conflicts can also be found between covert cognitive processes, which just influence next steps of cognition. Another new mechanism implemented in DUCCA is such a conflict monitoring procedure, which evaluates conflicts in cognitive processing (e.g., between opposing goals), which need not lead directly to any response.

Finally, DUCCA is aimed at taking into account the individual differences in control. Even healthy people differ in efficiency of control, which seem to be correlated with working memory capacity and fluid intelligence (Chuderski & Nęcka, 2010). Moreover, humans are able to regulate their mode of control by switching between top-down, proactive control and bottom-up, reactive one (Braver et al., 2007). All these differences can be expressed as differences in values of DUCCA internal control parameters, which yield qualitative changes in its simulated behaviour.

Overview of DUCCA

Cognitive operations

DUCCA is modeled as a hybrid production system. Coordination of the working of its modules is inspired by ACT-R architecture (Anderson et al., 2008). However, as the system is focused on executive functioning, "ordinary" cognitive operations have been very simplified. The system stores information received from the environment in a visual attention module, which recognizes 25 (5×5) locations on the computer screen and attends to one of them (via a focus of visual attention) at a time. Model can read symbols and some of their features (e.g., colors) from the focus. Long-term declarative knowledge is organized as a semantic network, which consists of information chunks of defined

categories (which are also chunks). A chunk contains a few slots. Each slot can contain either an atomic symbol or a reference to another chunk. One chunk can be retrieved at a time and placed in a retrieval buffer. Some information relevant to the task is actively maintained in system's focus of working memory (WM). The capacity of the focus is limited to a few (DUCCA's parameter) chunks. Contents of WM focus constitute a *context* of cognitive operations. Another structure is a goal module, which can do only one thing: it represents one chunk as a current goal of the system. Finally, a simplified motor module simulates reactions. Each response is registered and processed by a virtual key set.

Crucial for how DUCCA behaves is its procedural module, consisting of production rules, their utilities, and the mechanism for adapting utilities. Each rule is defined as a collection of conditions and a collection of actions. Conditions are imposed on both foci and the retrieval buffer. For each rule (i), a utility value (U_i) is assigned, which is updated on the basis of feedback. The utility of i tends to the expected value of feedback received after the action i. The higher U is, the more probable is the execution of a respective rule (see below).

DUCCA adapts the value of a recently executed rule in a reinforcement learning procedure, according to formula (1):

$$U_{i,t} = U_{i,t-1} + \frac{f - U_{i,t-1}}{1 + L_i}, \qquad (1)$$

where $U_{i,t}$ is a new value of utility of rule i, f is a feedback value (in range zero to one, where zero reflects "complete failure" while one means "full success"), and L_i is the reliability of a recent value of utility ($U_{i,t-1}$), estimated as the number of trials in which reinforcement of rule i has been applied. The rationale for equation (1) is that the more reliable a utility is, the less a current feedback alters this utility value. If a rule is new and L_i equals to zero, then after the first execution of a rule its utility reflects exact value of a feedback. After numerous rule's executions, its utility becomes very reliable and feedback can change it minimally. U values (in [0,1] range) reflect expected probability of reaching a goal if a rule is executed. In simple executive tasks, the reinforcement value f may be usually operationalized as the extent to which a task instruction was fullfiled, as perceived by a subject or signaled by a task.

If the environment and a context unambigously determine an adequate action, then one rule will be matched and executed in time inversely proportional to its utility. Execution of the rule may: change the goal and/or contents of WM focus, redirect the focus of visual attention, add a chunk to the declarative memory, and send a motor command to the motor module. Then a next cycle of operation starts, until the goal is reached. However, if at least two alternative rules match (i.e., DUCCA detects a conflict related to rule selection), then executive control has to be involved in the choice of one rule from a set of matching ones (*conflict set*).

Control of cognitive operations

The first mechanism of executive control deals with evaluation of the level of detected conflict C, which is calculated according to formula (2) based on nonlinear Luce's ratio:

$$C_i = \left(\frac{\sum_{j \neq i} e^{U'_j/n}}{\sum_k e^{U'_k/n}} \right)^c, \qquad (2)$$

where j indexes all production rules in a conflict set, which yield different cognitive or behavioural consequences than a rule i of maximum utility in a conflict set, k indexes all rules in a conflict set, and n is a noise parameter. Conflict measure is thus a proportion of utilities of matching rules which are alternative to the dominant tendency for cognitive or motor processing. Parameter n controls how nonlinear is the computation of C. Note that U's instead of Us are used (the calculation of U' is explained below).

The C value determines the strength of top-down control (G_t) exerted from the goal, according to formula (3):

$$G_t = ag\big(C + E(1 - C)\big) + (1 - a)G_{t-1}, \quad (3)$$

where G_{t-1} denotes the strength of control in a previous cycle, E is an error value (meaning the probability that the system committed an error in a previous cycle), g is the maximum strength of control that DUCCA can exert, and a is a control adaptation parameter. C and E work in underadditive interaction. Parameter a can vary between zero (DUCCA exerts fixed strength of control and ignores conflicts and errors) and one (system uses a proportion of its maximum control strength relative to the conflict level). Theoretically plausible values of a lay above zero and below one and they mean that DUCCA adapts control to conflicts and errors, but it does so with some inertia.

The set of DUCCA's rules and their utilities may be understood as a strategy, which maps a set of possible cognitive operations onto a set of probabilities of executing these operations, in a given state of the environment and a given goal and context. Without executive control, a distribution of these probabilities reflects the effects of learning (via Us). The operation of control consists in changing this distribution into one independent on learning but dependent on how these actions are adequate to a current goal. Due to control, an agent can undertake some arbitrary behavior, even if other well-learned behavioral patterns conflict with it. The second control mechanism operates thus as modifier of rules' utilities, according to formula (4):

$$U'_i = \frac{U_i}{e^{G(1 - A_{ij})}}, \qquad (4)$$

where modified utility U'_i of rule i, which is used is for conflict evaluation and conflict resolution (see below), is decreased in a function of a current control strength (G) and a value of association A_{ij} between rule i and current goal j. If either rule i is perfectly adequate to goal j (A_{ij} equals one) or control strength G is null, then U'_i equals U_i. In all other cases U_i is decreased in a nonlinear function of G and A_{ij}. If G is very high, the system just selects the rule closest to a goal. Though such a control mechanism can be judged inhibitory, our model is not committed to either an inhibitory or activational nature of control. In terms of probabilities, inhibition of one set of rules is conceptually indistinguishable from activation of an alternative set of rules.

Finally, DUCCA uses modified utilities in order to resolve a conflict among rules present in a conflict set. Analogously as in conflict evaluation formula, nonlinear Luce's ratio is exploited in formula (5) for the calculation of a probability P_i of rule i execution:

$$P_i = \frac{e^{U_i'/n}}{\sum_j e^{U_j'/n}}, \qquad (5)$$

where j denotes all rules in a conflict set, and n is a noise parameter (the same as in formula [2]). When n is very high, the rule with maximum U' always wins, while at n close to zero P equals to one divided by a number of rules in a conflict set. An important DUCCA's assumption (opposite to ACT-R theory) is that conflict resolution consumes time relative to the conflict level. Latency of conflict resolution is a multiplication of conflict value C and a scaling parameter s (i.e., $Lat = s \times C$).

Executive control in DUCCA stems from a dynamical interaction of external stimulation and its consequences (rules' utilities and goal-rule associations) and two internal mechanisms strategically adapting to the pattern of cognitive processing (conflict evaluation plus control strength modification and utility learning).

Modeling of Stroop

Stroop-like tasks, which are widely used to examine operations of executive control (MacLeod, 1991), impose interference by presenting bivalent, incongruent stimuli, which activate two cognitive processes: one dominant and the other much weaker. The task is to complete the non-dominant process. The well-known example is naming a color of a colored word that itself means an incongruent color. Interference effect, namely a positive difference between RTs for incongruent stimuli and neutral ones (e.g., colored letters X), reflects the unavoidable additional time needed for control processes to override interference from a dominant process. At the same time, control processes are usually successful, as error rates in Stroop-like tasks are low (2-10% on average). Often, a facilitation effect is also observed: people are faster for congruent stimuli (e.g., when word and its color match) than for neutral ones (MacLeod, 1991).

Some existing models

A seminal connectionist model (Cohen et al., 1990) represented alternative processing pathways as interconnected nodes in a network. Nodes for non-dominant process were associated more weakly than those of the dominant one. For the non-dominant pathway to win, an additional task-unit had to activate that pathway. A version of the model supplemented with conflict monitoring node (Botvinick et al., 2001), which controlled the amount of activation spread by the task-node in a function of conflict within a response layer, replicated above mentioned Gratton effect. It was also able to simulate an observed decrease in interference with increase in proportion of non-neutral (congruent plus incongruent) stimuli as well as smaller than interference a facilitation effect (Tzelgov, Henik, & Berger, 1992). In another model, Verguts and Notebaert (2008) implemented conflict-modulated Hebbian learning rule, which adapted specific network connections involved in conflict resolution. The model was able to account for a decrease in interference for items often presented in incongruent contexts, in comparison to stimuli usually presented as congruent (i.e., for a so-called item-specific proportion congruency effect).

However, connectionist models are often judged atheoretical (e.g., Altmann & Davidson, 2001). They represent a modeled mechanism as just a several links between a few abstract nodes of no internal structure. A node for "redness" would be exactly the same as a node for "left keypress", even if they belong to different categories of phenomena. These models are not related to any cognitive theory (e.g., of language or memory) either. In consequence, models of tasks imposing different constraints (e.g., Stroop, flanker, or antisaccade tasks) may be described by the same network.

Some other Stroop-like models do make assumptions on related cognitive processing and focus also on more specific aspects of Stroop performance. Altmann and Davidson (2001) modeled Stroop interference as an effect of the competition between syntactic properties of the words (lemmas) and embedded this linguistic mechanism in a broader cognitive architecture (i.e., ACT-R). The model was able to explain why the separation of incongruent aspects of stimulus in time decreased interference. Lovett (2005), exploiting ACT-R's idea of utility learning of production rules, was able to explain strategical preferences of participants in chosing dominant and non-dominant processes. However, all these models would have difficulty in explaining interference effects in Stroop isomorphic tasks, which do not relate so much on linguistic properties (e.g., flankers task) or memory retrievals (e.g., Navon task).

Specific processes surely explain some part of a variance in Stroop interference, but the general executive mechanisms beyond specific processes may be responsible for the significant part of that variance. Our architecture is aimed to describe these mechanisms. However, it explains them with higher theoretical plausibility than most of connectionist models do. The model identifies different categories of cognitive structures (e.g., rules, chunks, goals) and it can ascribe meaningful contents to particular representations. Moreover, the architecture isolates executive aspects common to different tasks from task-specific characteristics. Finally, it can easily be extended with additional theoretical assumptions (e.g., ones concerning language or memory).

DUCCA's model of Stroop

We developed a model of a *generalized* Stroop-like task in order to account for a variety of results, observed within different experimental conditions and numerous versions of Stroop tasks (i.e., we abstracted from task-specific aspects).

DUCCA was supplemented with task-specific rules and chunks. There are three crucial rules for response choice: "trained", "target", and "others". The first rule leads to a skilled action, which is not proper for a task instruction. For this rule, the maximum utility ($U_{trained} = 1.0$) was set, reflecting that for adult participants such a rule had received millions of positive feedbacks. The second rule leads to instructed, but relatively poorly trained action. Its utility should be

much lower than $U_{trained}$ but still significantly above 0 (here, U_{target} = .6). "Others" represents all task-unrelevant possible processes, including ruminations and mental slips, and it should have a utility close to 0 (here, U_{others} = .1), as ruminations and slips rarely lead to positive feedbacks.

The model contains some visual and memory chunks. One important aspect of perceived stimuli is that each congruent and incongruent stimulus is bivalent: one its aspect is matched by the rule "trained", while the other aspect is matched by the rule "target". Rule "others" matches any stimulus. Memory chunks associate stimuli with proper responses. We skip other details of chunks' description.

Though the rule "target" has a low utility, it is fully associated with the goal (A_{target} = 1.0). The rule "trained" has goal association much lower than A_{target}, but still significantly above 0 (here, $A_{trained}$ = .2), as it is somehow related to what happens during the task (e.g., when congruent stimuli are frequent, it may be beneficial to use sometimes the dominant rule). Thus, in every congruent and incongruent trial there is a competition between useful rule "trained" and goal-relevant rule "target". This is modulated by the strength of control (G): the stronger control is the higher is choice probability of the rule "target". Though the rule "others" is not associated with the goal (A_{others} = .01), it may sometimes be chosen, depending on the amount of noise. When the model perceives a neutral stimulus, the rule "trained" cannot be effectively applied and only the rules "target" and "others" fall into the conflict set.

Choosing a reaction means that either the rule "trained" or the rule "target" retrieves a chunk from the declarative memory, according to stimulus features present in the visual buffer. Perceiving a feedback is applied in a simplified form, as the information about correctness of the response is displayed on the screen and processed directly.

Simulation results and discussion

The noise was set to relatively low value of 0.15, as all modeled experiments involved young and healthy participants. Parameter g equalled to 3.625 (i.e., the mean value between high- and low-WM groups, see last section). Value of c was set to 0.6, reflecting relative sensitivity to conflicts. Two time scaling parameters for each simulation were optimized to fit observed data. As these data come from differing tasks (a flanker task and two different versions of Stroop task) and experimental conditions, we did not try to fit data precisely, but we were looking for qualitative replication of the wide range of effects, instead.

Gratton effect The original Gratton et al.'s (1992) effect in flanker task is often replicated within Stroop paradigm (e.g., Kerns et al., 2004). However, for comparision with other models, we aimed to replicate the original effect (see Figure 1, left panel). In the first simulation study, 5000 runs of the model were administered with 50/50 proportion of congrent vs. incongruent trials. The ordering of trials was random. The simulated Gratton effects is presented in Figure 1, right panel. Though the model generated slightly larger interference effect, influence of previous trial was the same as in the original experiment. The Gratton effect in DUCCA comes from the rise in conflict level (C) after incongruent trial. In a subsequent trial, C is higher than it would be if a previous trial was congruent. So, the control strength (G) is higher and it makes (via U's) the firing of the rule "target" faster, leading to decrease in RT in incongruent trials. It also makes the execution of the rule "trained" slower. As this rule may often be fired in congruent trials, it thus results in increased RT in these trials.

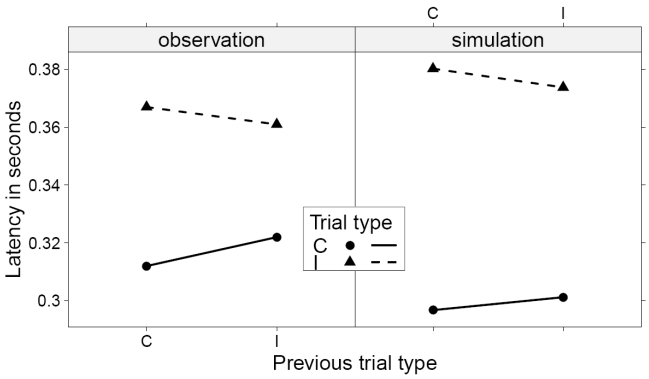


Figure 1: Left panel: data adapted from Gratton et al. (1992) on latency in congruent (C) and incongruent (I) trials as a function of a previous trial. Right panel: simulated data.

Practice on a non-dominant process The seminal study on a relation between the level of automaticity of a non-dominant process and Stroop interference was administered by MacLeod and Dunbar (1988). The participants were asked to name colors arbitrarily associated with shapes by an instruction (a task to be practiced). The shapes were colored. As expected, when color-to-name and actual color mismatched, responses took longer when colors matched or a shape was non-colored. On some days, only practice trials (naming shapes) were applied. General result was that practice on non-dominant process deacresed (and after some enourmous number of practice trials – even reversed) an interference cost. Here, we replicated the effect of five days of training (about 2000 practice trials) on interference (see Figure 2).

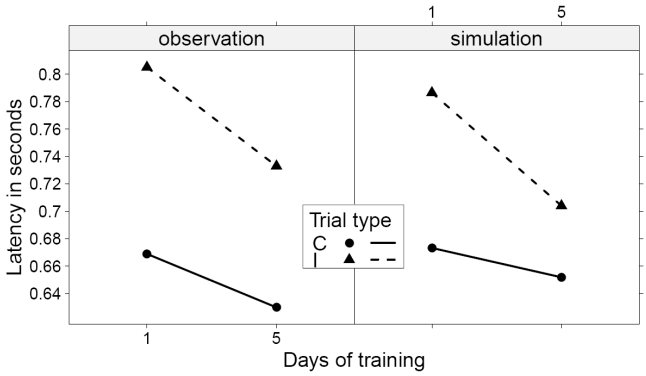


Figure 2: Left panel: data adapted from Experiment 3 by MacLeod and Dunbar (1988) on latency in congruent (C) and incongruent (I) trials as a function of practice on a non-dominant task. Right panel: simulated data.

In this simulation, which regarded a task with highly artificial non-dominant action, we used a lower value of U_{target} equal to 0.1. The practice runs resulted in decrease in utility of the rule "trained" (as it lead to errors during practice) and in increase in U_{target}. This "automatization" effect was caused by model equation (1). The change in Us caused the decrease in an interference cost, as the lower difference in utility between both rules increased a conflict value C. The increased conflict engaged more efficient control because of larger value of G. Then, 480 test runs were carried to simulate the presented data.

Proportion of incongruent stimuli, facilitation, and individual differences in Stroop performance Kane and Engle (2003; Experiment 4) observed decrease in Stroop interference as a result of decreasing proportion of congruent simuli, when neutral stimuli were absent. Moreover, it appeared that this proportion influenced the difference in accuracy in incongruent trials between low- and high-working memory capacity (WMC) participants, screened with operation span task. When proportion was low (20% congruent), both WMC groups scored around six percent of errors, with no significant advantage of WMC-high group. When incongruent trials were rare (80% congruent), error rate increased, but much more for WMC-low subjects (see Figure 3, left panel). Kane and Engle interpreted this as a result of more freqent slips of attention control of WMC-low group. In 20% congruent sequence, stimuli exogenously kept the control focused on non-dominant process and the differences in quality of internal control did not matter much. When incongruent trials were rare, only internal control could keep focus on non-dominant process and weak control of WMC-low group more often made it loose the task goal and commit more errors on incongruent trials.

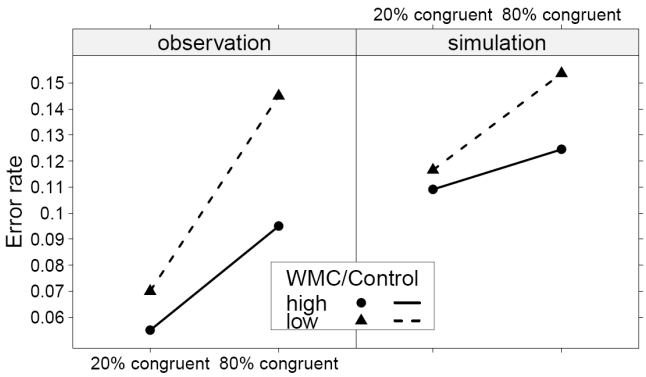


Figure 3: Left panel: data adapted from Kane and Engle (2003) on error rate in incongruent trials as a function of proportion congruent and WMC. Right panel: relevant data simulated with high/low parameter g value.

Interestingly, WMC differences did not interact with the effect of congruent trials proportion on latency: interference effect increased with increasing proportion of congruent trails, but WMC-low participants presented higher effect than WMC-high ones in both conditions of proportion congruent (see Figure 4, left panel). Kane and Engle observed also (Experiment 2) the differences in facilitation effect. Surprisingly, WMC-low persons exhibited a larger effect (72 ms) than WMC-high ones (41 ms). On congruent trials, WMC-low participants might have more often used the dominant process to emit a response. Although use of this process did not cause errors in congruent trials, as both processes lead to the same response, it could have speeded up WMC-low participants' RTs comparing to RTs of WMC-high ones (who probably avoided the dominant process).

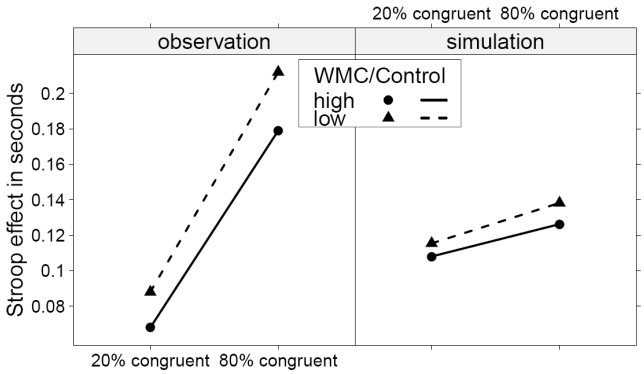


Figure 4: Left panel: data adapted from Kane and Engle (2003) on interference effect a function of proportion congruent trials and WMC. Right panel: data simulated with high/low parameter g value.

The complicated pattern of results presented in this subsection constitutes a tough test for any Stroop model. We simulated those data using either 36 congruent and 144 incongruent trials (20% congruent condition) or vice versa (80% congruent condition), following Kane and Engle's procedure in Experiment 4. The value of g parameter was set to lower value of g = 3.5 in order to reflect WMC-low group or set to higher value of g = 3.75, to reflect WMC-high group. 4320 runs of the model yielded simulated data.

All observed effects were qualitatively replicated. As in Kane and Engle's study, the effect of the proportion congruent was observed in latencies as well as in errors. In all conditions, increase in parameter g caused reasonably lower interference effects in latencies. However, the difference in g resulted in difference in accuracy on incongruent trials only when incongruent stimuli were rare. Simulated data are presented in right panels of Figures 3 and 4. In a simulation of Experiment 2, which differed slightly from Exp. 4, neutral trials were included and the values of g = 3 and g = 4 were set for WMC-low and WMC-high groups, respectively. The facilitation effect (68 ms) appeared much smaller than the interference effect (137 ms) and it fitted observed results. Also, WMC-low group scored larger facilitation effect (76 ms) than WMC-high persons (60 ms).

Figure 5 presents the indices of strategical adaptation to different (20% vs 80% congruent) task conditions. The model adapted mean level of control, rising its average level from 80% congruent to 20% congruent condition. Due to utility learning, in the more difficult condition the model amplified a utility of non-dominant rule and lowered the one of dominant rule, what increased internal conflict and thus recruited additional control. Such a strategical adaptation was less efficient when maximum strength of control was

limited (i.e., when *g* value was low), matching the results of WMC-low participants.

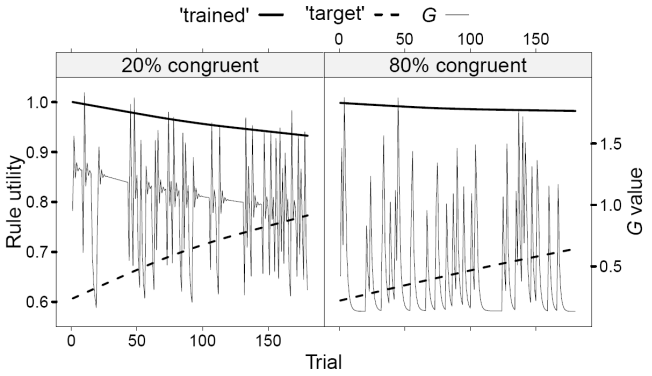


Figure 5: Internal dynamics of the model expressed as fluctuations in exerted control (*G*) and changes in utilities of the rules "trained" and "target" in two task conditions.

Two major quantitative deviations from data may be noticed: much smaller effect of proportion of congruent stimuli on latency interference and more errors committed by the model than by participants. These deviations probably result from the fact that our model captures only general aspects of control, while experimental situation involve many other general processes (e.g., expectations about the probability of events, changes in speed-accuracy trade-offs, decreased vigilance, and so on) as well as some task specific processes, all influencing interference effects. However, as a hybrid and general architecture, DUCCA can potentially implement all these processes within more complex models.

Summary and conclusions

DUCCA, a new general architecture of executive control was presented. It was applied in order to simulate Stroop-like task. We used only a few simple assumptions of how control operates and still were able to replicate most of general effects observed in Stroop paradigm: asymetrical interference and facilitation effects, the Gratton effect, an influence of practice on Stroop effect, decrease in interference as proportion of congruent trials decreases, and the complex pattern of individual differences related to WMC.

The presented work is on a preliminary stage. Taking into account semantics and item-specific effects in executive control, linking executive mechanism to brain structures, and explaining the common variance in several executive tasks and its role in complex cognition constitute the most important future directions of DUCCA development.

Acknowledgments

This work was sponsored by Polish Ministry of Science and Higher Education (grant N106 2155 33, yrs. 2007-2010).

References

Altmann, E. M., Davidson, D. J. (2001). An integrative approach to Stroop: Combining a language model and a unified cognitive theory. In J. D. Moore & K. Stenning (Eds.), *Proceedings of the 23rd Annual Conference of the Cognitive Science Society* (pp. 21-26): Hillsdale, NJ: Laurence Erlbaum.

Anderson, J. R., Fincham, J. M., Qui, Y., & Stocco, A. (2008). A central circuit of the mind. *Trends in Cognitive Sciences, 12,* 136-143.

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review, 108,* 624–652.

Braver, T. S., Gray, J. R., & Burgess, G. C. (2007). Explaining the many varieties of working memory variation: Dual mechanisms of cognitive control. In A. R. A. Conway. C. Jarrold, M. J. Kane, A. Miyake, & J. N. Towse, *Variation in working memory* (pp. 76-108). Oxford: Oxford University Press.

Brown, J. W., & Braver, T. S. (2007). Risk prediction and aversion by anterior cingulate cortex. *Cognitive, Affective, & Behavioral Neuroscience, 7,* 266, 277.

Chuderski, A., & Nęcka, E. (2010). Intelligence and cognitive control. In A. Gruszka, G. Matthews, & B. Szymura (Eds.), *Handbook on individual differences in cognition* (pp. 263-281). New York: Springer Verlag.

Cohen, J. D., Dunbar, K., & McClelland, J. L. (1990). On the control of automatic processes: A parallel distributed processing model of the Stroop effect. *Psychological Review, 97,* 332-361.

Gratton, G., Coles, M. G. H., & Donchin, E. (1992). Optimizing the use of information: Strategic control of activation of responses. *Journal of Experimental Psychology: General, 121,* 480–506.

Kane, M. J., Engle, R. W. (2003). Working-memory capacity and the control of attention: The contributions of goal neglect, response competition, and task set to Stroop interference. *Journal of Experimental Psychology: General*, *132,* 47-70.

Kerns, J. G., Cohen, J. D., MacDonald, A. W., 3rd, Cho, R. Y., Stenger, V. A., & Carter, C. S. (2004). Anterior cingulate conflict monitoring and adjustments in control. *Science, 303,* 1023–1026.

Koechlin, E., & Summerfield, C. (2007). An information theoretical approach to prefrontal executive function. *Trends in Cognitive Sciences, 11,* 229-235.

Lovett, M. C. (2005). A strategy-based interpretation of Stroop. *Cognitive Science, 29,* 493-524.

MacLeod, C. M. (1991). Half a century of a research on the Stroop Effects: An integrative review. *Psychological Bulletin, 109,* 163-203.

MacLeod, C. M., & Dunbar, K. (1988). Training and Stroop-like interference: Evidence for a continuum of automaticity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 126-135.

Meyer D., Kieras D. (1997) A computational theory of executive cognitive processes and multiple-task performance. Part 1: Basic mechanisms. *Psychological Review, 104, 3-65.*

Tzelgov, J., Henik, A., & Berger, J. (1992). Controlling Stroop effects by manipulating expectations for color words. *Memory & Cognition, 20,* 727-735.

Verguts, T., & Notebaert, W. (2008). Hebbian learning of cognitive control. *Psychological Review, 115,* 518-525.

Modelling the Correlation Between Two Putative Inhibition Tasks: An Analytic Approach

Eddy J. Davelaar (E.Davelaar) and Richard P. Cooper (R.Cooper@bbk.ac.uk)
Department of Psychological Science, Birkbeck, University of London
Malet Street, London WC1E 7HX, UK

Abstract

A process of response inhibition is often held to be recruited in situations where it is necessary to withhold or inhibit a prepotent response. Individual differences in the efficacy of this function have been held to underlie individual differences in behaviour on tasks such as the Stroop colour-naming task and the stop-signal task. These claims, however, have been supported only with correlational analyses and informal argument. This paper considers the operationalisation of response inhibition by exploring existing mathematical and process models of both the Stroop and stop-signal tasks. We identify parameters that might underlie individual differences in the performance of the tasks and consider potential relations between those parameters. It is shown that (a) at least three potential inter-relations between parameters of the task models may lead to inter-task correlations, and (b) the observed correlations arise when attentional bias parameters in the models are equated but not when inhibition parameters are equated. We conclude that the ascription of such correlations to a process of response inhibition is premature.

Keywords: Cognitive control; Response inhibition; Stroop task; Stop signal task; Individual differences.

Introduction

In much everyday behaviour, and in many psychological tasks, it is necessary to resist temptation or to avoid producing a prepotent response. Consider the well-known Stroop colour-naming task, where the subject is required to name the colour of the ink in which a word is printed. If the word is itself the name of a colour (e.g., RED printed in green ink) then the subject must actively or deliberately resist the temptation to read the word if they are to successfully name the ink colour.

It is commonly argued that the ability to inhibit a prepotent response is facilitated by a cognitive control process referred to as *response inhibition*. Critically, response inhibition is not a task-specific construct, limited to (e.g.) the Stroop task. Rather, it is held to be one of several general "executive" processes that are invoked across a range of tasks. Moreover, individual differences in the ability to inhibit a prepotent response are held to reflect individual differences in the efficacy of response inhibition. For example, in a well-known study of cognitive control by Miyake, Friedman and colleagues (2000), 137 subjects completed a battery of tasks, three of which were assumed specifically to tap response inhibition. Miyake and colleagues found significant pair-wise correlations in performance on the response inhibition tasks, and confirmatory factor analysis supported their model of executive function as comprising at least three separable components, one of which was response inhibition.

The three tasks held by Miyake et al. to tap the latent construct were the Stroop task (as discussed above), the stop-signal task of Logan (1994), and an antisaccade task (Roberts et al, 1994). In the stop-signal task subjects complete a series of trials in which they must normally respond as quickly as possible to a stimulus (e.g., by indicating whether an auditorily presented noun denotes a type of animal). On a small proportion of trials the stimulus is followed by a second "stop" stimulus (e.g., a beep), indicating that on this particular trial a response should be withheld. In the antisaccade task trials involved visual presentation of a fixation point at the centre of a monitor screen. This was followed by a brief cue appearing to the left/right of the screen and then an even briefer target appearing on the opposite side of the screen. Subjects were required to make a choice decision based on a feature of the target. To do so, they needed to avoid making a saccade to the cue, as this would prevent them from being able to make a saccade back to the target before it was replaced by a mask. Response inhibition was indexed by Miyake et al. (2000) in the Stroop task by the difference in response times between incongruent and neutral trials. In the stop-signal task it was indexed by the number of stop trials on which a response was (incorrectly) produced. In the antisaccade task it was indexed by the proportion of correct target decisions. As noted above, significant pair-wise correlations were found between these measures. This result was effectively replicated in a subsequent study with 220 subjects which used the same tasks but slightly different dependent measures (Friedman & Miyake, 2004).

The studies of Miyake, Friedman and colleagues appear to provide strong support for the response inhibition construct and for its variability across individuals. However in both cases the evidence is purely correlational. Neither study attempts to provide a mechanistic account of response inhibition as it might be manifest in the various tasks. Clearly, if response inhibition is a cognitive control process that plays a causal role in the performance of the Stroop, stop-signal and antisaccade tasks (amongst others), then that process should be shared by computational accounts of the three tasks. Moreover, if the efficacy of that construct can vary across individuals, then that process should be parameterised in the computational accounts. Lastly, if pair-wise correlations in performance of the tasks are to be attributed at least in part to the efficacy of response

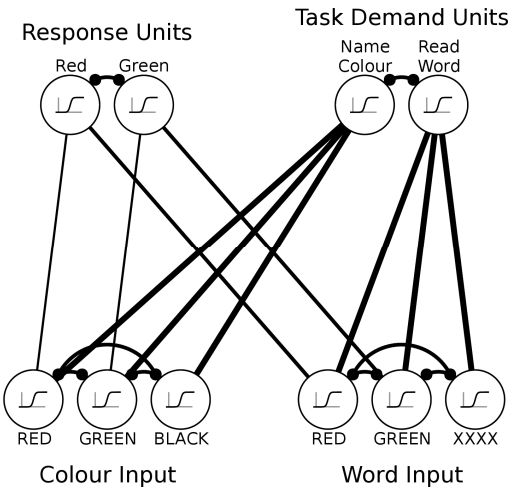


Figure 1: The architecture of the Cohen & Huston (1994) model of the Stroop task. The response function for each node is a sigmoid. Relative line thickness indicates connection strength. Lateral inhibition (shown by arrows with circular end points) operates between nodes in each group.

inhibition, then varying the response inhibition parameter in the computational accounts should also result in pair-wise correlations.

The difficulty, however, is that a cursory analysis of existing process models of the tasks used by Miyake, Friedman and colleagues suggests that their dependent measures are not obviously related to a common mechanism of response inhibition. Consider the widely accepted interactive activation model of the Stroop task of Cohen and colleagues (e.g., Cohen & Huston 1994; see Figure 1). In the model, interference on incongruent trials arises from competition between two response pathways – a word-reading pathway which is highly practiced and hence strong, and a colour-naming pathway which is less practiced and hence somewhat weaker. In order to generate a color-naming response on an incongruent trial it is necessary to selectively amplify the inputs to the color-naming pathway via task-demand units. This process, often referred to as attentional biasing, allows activation from the colour-naming pathway to dominate activation from the word-reading pathway. While individual differences in interference are not generally the focus of this model, they may be captured by assuming that individuals who show relatively little interference are better able to maintain strong excitation of the color-naming task-demand unit. This in turn might result either from greater input to the color-naming task demand unit from external sources (e.g., attentional processes) or conceivably from stronger lateral inhibition between task-demand units. Therefore in this model at least the dependent measure of Miyake et al (2000) indexes an aspect of task-demand, and not response inhibition.

The goal of this paper is to formalise this analysis and extend it to a second putative response inhibition task, namely the stop-signal task, for which a relatively well-developed "off-the-shelf" computational account is also available (Boucher et al., 2007). We analyse potential sources of correlations in performance across the two tasks by couching both models within a common architecture. In so doing we question the standard concept of response inhibition and propose instead that correlations between performance on the Stroop and stop-signal tasks might be due to a somewhat different factor related to the strength or potency of the currently selected goal.

The Task Models

In order to address the correlation between the Stroop and stop-signal tasks, we converged on an interactive activation architecture based on the existing published models. This architecture was then simplified to extract a small set of equations that relate the relevant parameters of cognitive control in these two tasks to the dependent measures used by Miyake et al. These equations were then used to generate distributions of the dependent measures by varying the critical parameters and calculating the resulting correlations.

Stop-signal task

The version of the stop-signal task used by Miyake et al. consisted of two blocks. The principal task was an animacy-categorisation task. The first block only had categorisation trials and was intended to ensure that generating a response was indeed the prepotent response. The second block included 25% stop-trials. For our analytic modelling efforts the following components are relevant. First, the first block produced a mean response time. This was used on a subject-by-subject basis to adjust the onset of the stop-signal on stop-signal trials in the second block. For each subject this onset was their mean response time less 225ms. The stop-signal was therefore presented at (RT–225) ms post-stimulus. We assume a similar approach in the model. Second, the dependent variable was the proportion of categorisation responses generated on stop trials. This value represents errors due to failure to inhibit.

The architecture for our model is inspired by several preceding models. First, Boucher et al. (2007) used a simple interactive race model in which a "go" and a "stop" unit compete through lateral inhibition. Critical for their simulations is that the inhibition from stop unit to go unit is much larger than the reverse connection. This makes the model interactive for only a brief time. Second, the location of the units is downstream in the basal ganglia. This is also assumed in a related go/nogo-model of Frank and colleagues (2004). Third, in the go/nogo-model the nogo-signal comes through the subthalamic nucleus. This nucleus has been shown to form part of a response inhibition pathway that included the inferior frontal gyrus (Aron et al, 2004). It has been postulated that choice responses can be optimised through this pathway (Davelaar, 2009; Frank, 2006). This leaves us with the architecture shown in Figure 2a. It is assumed that the two units are located in the striatum and receive input from earlier processing levels

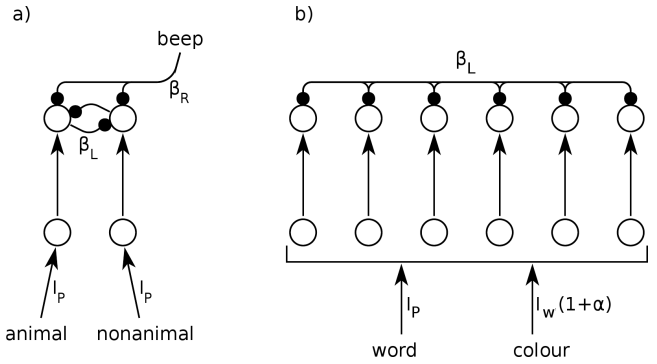


Figure 2: a) Basic architecture of the stop-signal model. b) Basic architecture of the Stroop model.

regarding the animacy of the stimulus. The two units are forward connected to two output units that are connected via lateral inhibition. These are assumed to be localised in the globus pallidus interna and substantia nigra. This is a straightforward model of response selection. The stop-signal is assumed to inhibit the final responses via the IFG → STN → SN/GPi-pathway. The strength of the response inhibition parameter β_R is one source of individual differences in stop-signal performance.

Stroop task

The Stroop task used by Miyake et al. involved naming the colour of a word of which the ink could be in one of six colours. Relevant for the current analytic modelling effort is that the dependent measure is the difference in voice key response time between the mean RT on incongruent trials and neutral trials (which consisted of coloured asterisks). The architecture for the Stroop model follows the incarnations of Cohen and colleagues discussed above. In particular, compared to the neutral trial, an incorrect unit gets activated in response to reading the colour-word. The lateral inhibition between the response units slows down the responses in incongruent trials. Some extension to this model is needed, however. Recent analyses have shown that the Cohen models are unable to produce the correct relation between the stimulus-onset asynchrony in versions of the Stroop task when the word and its ink colour are presented asynchronously (Stafford & Gurney, 2007). The solution to this problem was to treat the output units of the Cohen model as the input units to the basal ganglia, i.e., the striatum (see Figure 2b). This automatically leads to a consistent architecture for both the Stroop task and the stop-signal task.

Simplifying the Overarching Model

Some simplifications are necessary in order to provide an analysis of the correlation between the two tasks and the relevant parameters. First, we focus only on the abstracted basal ganglia pathway shown in Figure 2. Second, we assume no lateral inhibition in the input level and lateral inhibition of strength β_L at the output level. For the stop-signal model, an extra inhibitory connection of strength β_R to both units is assumed. Whereas in the stop-signal task, the animacy judgement is unambiguous (and prepotent), there is overlap in the Stroop task. This means that in the stop-signal task the only components doing the work are a single response unit and the β_R. In the Stroop task, there are two critical trial types. In the neutral condition the neutral response unit is activated in absence of any inhibition. In the incongruent condition the response unit receives input from the target channel and inhibition from the distractor channel via lateral inhibition. The amount of activation that goes through the target channel is under attentional control. Whereas earlier models of the Stroop task implemented a tradeoff in attention to both channels, recent functional imaging work did not find any support for a deactivation of the distractor channel (Egner & Hirsch, 2005). Instead, only a positive enhancing effect of attention was found in a Stroop-like task. Thus we assume an attentional parameter, α, which enhances the target channel. We assume that the prepotent inputs for both tasks are identical and that the weaker target channel propagates a weaker signal.

This leads to the following equations that govern the input activation of the target unit in all tasks and conditions:

Stop-signal task

$$X(t) = \begin{cases} I_P(t) & \text{when } t < \overline{RT} - 225ms \\ I_P(t) - \beta_R & \text{when } t > \overline{RT} - 225ms \end{cases} \quad (1)$$

Stroop task

$$X(t) = \begin{cases} (1+\alpha).I_W(t) & \text{neutral condition} \\ (1+\alpha).I_W(t) - \beta_L.I_P(t) & \text{incongruent condition} \end{cases} \quad (2)$$

In order to obtain response time, we assume a linear output activation function:

$$\frac{d}{dt}F(x) = X(t) \quad (3)$$

This choice is justified by the observation that simple and choice reaction time models operate optimally when they are in the linear part of a sigmoidal output function (Bogacz, et al., 2006). By assuming linear output activation functions, we thus assume optimal responding.

Finally, we assume that the response threshold, θ, is the same for both tasks. For the stop-signal task, a response deadline is included of 1500ms (as used in Miyake et al., 2000).

Given the above assumptions, the response time in the stop-signal task is:

$$RT_{SS} = \begin{cases} \dfrac{\theta - I_P.225}{I_P - \beta_R} + 225 & \text{for } I_p > \beta_R \\ \infty & \text{otherwise} \end{cases} \quad (4)$$

This is tested against the response deadline. An erroneous response is produced if the response time is less than this deadline. The difference in RTs between incongruent and neutral trials in the Stroop task is:

$$\Delta RT_{Stroop} = \theta.\left\{\frac{1}{(1+\alpha).I_W(t) - \beta_L.I_P} - \frac{1}{(1+\alpha).I_W(t)}\right\} \quad (5)$$

For both equations θ was fixed at one and noise was added.

One immediate observation of interest is that architecturally, the mechanisms producing incorrect stop-trials and slowed down Stroop trials are not identical. In fact, Stroop performance is determined by the lateral inhibition between two information channels, whereas stop-errors are due to a pathway that inhibits both competing channels.

Our focus is on four parameters: the prepotent response parameter, I_P, the response inhibition parameter, β_R, the attention parameter, α, and the lateral inhibition parameter, β_L. There are a number of constraints on the parameters and points to note. First, note that I_P is shared between the models and moreover that this is the *only* parameter that is shared. Thus, it is expected that this parameter will be the locus of (at least some of) the correlation between the two tasks. Second, the following constraints hold:

- $(1 + \alpha).I_W > \beta_L.I_P$ in order to ensure that response accuracy in the Stroop task is above 50%
- $\beta_L < \beta_R$. This is justified based on the findings of Boucher et al. (2007)
- $I_W < I_P$, by definition

We focus on the following three potential sources of correlation between the proportion of stop-errors and the size of the Stroop interference effect:

1. Pre-potency of input. The pre-potency of the input, I_P, is an obvious choice from the architectural viewpoint, as it is the only parameter that features in both models. Therefore varying I_P across subjects should produce the positive correlation between the two tasks. The pre-potency, however, is not a factor that is mentioned as an executive function by Miyake et al (2000) and in fact would in most accounts be categorised as the parameter that has to be overcome via executive control.

2. Correlated executive functions. To overcome the pre-potency in the stop-signal task, response inhibition, β_R, is the relevant parameter, while for the Stroop task, the attentional control, α, is the relevant parameter. Obviously, varying these parameters across subjects should not produce a correlation in performance measures. However, one could argue that executive functions are themselves partly correlated (as is done by many authors including Miyake et al., 2000). If this is the case, then a correlation between the two tasks may not be due to shared variance in inhibition parameters, but due to a correlation between the executive functions of inhibition and attentional focus. One possibility that we will come back to in the discussion is that both of these concepts might be subsumed under a more general notion of the strength or potency of the goal, as both tasks require the need to exert control based on the recognition of a stimulus (stop-signal or colour-word).

3. Correlated inhibition. Perhaps the most natural way of addressing the correlation is to assume that response inhibition in the stop-signal model, β_R, and lateral inhibition in the Stroop model, β_L, are correlated. However, note that the dependent variables are such that greater (response) inhibition in the stop-signal task leads to fewer errors and hence lower levels of the dependent measure, whereas greater (lateral) inhibition in the Stroop task leads (perhaps counter-intuitively) to slower responses in the incongruent condition and higher levels of the dependent measure. Thus, correlated inhibition will lead to a correlation in the dependent measures, but this will be a *negative* correlation – not a positive one! Thus correlated inhibition can only result in the observed positive correlation between dependent measures on the stop-signal and Stroop tasks if the inhibition parameters are *negatively* correlated.

Sampling Studies: Methods and Results

Several sampling studies were conducted based on the above analysis. The aim of these studies was to assess effects of the three potential sources of correlation identified above on the cross-task correlation in dependent measures. To this end, equations 4 and 5 were used to obtain dependent measures for each task as all parameters except I_W were varied uniformly using boundaries that (a) were found to be adequate to produce values for the dependent measures that were within the range of the actual empirical results and (b) adhered to the set of constraints above. I_w was fixed to 0.6. The choices of uniform distributions and the precise value of I_w are arbitrary and do not impact on the conclusions drawn from this work.

We imposed associations among parameters as follows:

1. To address the pre-potency view, only the I_P distribution was varied between-subjects and each subjects' I_P value was used in both task models. For each virtual subject, the other three parameters were randomly sampled 100 times corresponding to 100 trials within a task. The proportion of stop-errors was calculated as the mean number of times that a response time in the stop-signal task was shorter than 1500 ms. The Stroop mean interference was calculated over the 100 3-parameter combinations (together with the subject's I_P). One hundred subjects were simulated and a Pearson product-moment correlation coefficient was calculated over the resulting set of 100 data pairs.
2. To address the correlated executive function view, β_R and α were used as between-subjects parameters (I_P and β_L varied within-subjects). There were two versions: uncorrelated and correlated β_R and α.
3. Finally to address the correlated inhibition view, we correlated β_R and β_L between-subjects (I_P and α varied within-subjects).

In all cases additional noise was added to the correlated parameter in order to lower the resulting correlation in dependent measures and obtain a value of approximately 0.18 as found in the behavioural studies of Miyake et al. (2000).

Scatter plots showing the correlation between dependent measures for four situations are shown in Figure 3. Positive correlations can be obtained between the dependent measures either when I_P is fixed within-subjects (exploring

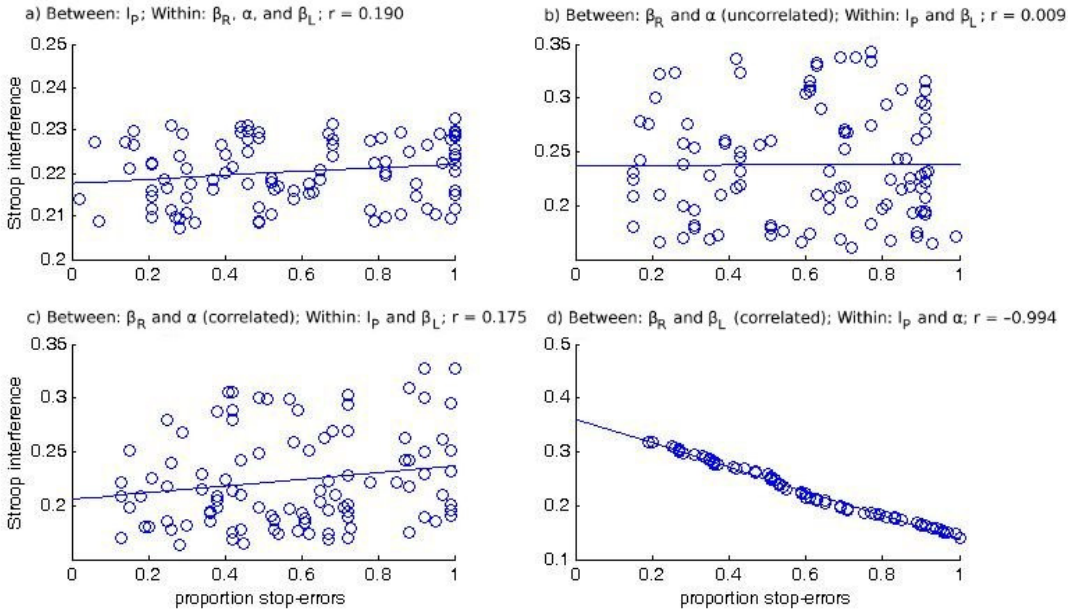


Figure 3: Scatter plots of dependent measures under different conditions. a) When I_P is fixed within-subjects (the pre-potency view) the correlation is positive. b) If β_R and α vary within-subjects but are uncorrelated, then there is no correlation. c) The correlation is positive when β_R is correlated with α (the correlated executive functions view). d) When β_R and β_L are positively correlated the correlation between dependent measures is negative.

the pre-potency view; Figure 3a), or when β_R is correlated with α (exploring the correlated executive functions view; Figure 3c). If β_R and α vary within-subjects but are uncorrelated, then there is no correlation between the dependent measures (Figure 3b). When β_R and β_L are correlated, then as anticipated the correlation between dependent measures is negative (Figure 3d).

Discussion

We set out to address the correlation between two well-known tasks that have been discussed as tapping executive inhibition. Correlations between performance on the stop-signal and the Stroop tasks have been found in several behavioural studies and both tasks have been the subject of detailed computational modelling. The modelling has been at the same architectural level, thus allowing the integration of those models into a larger more general model. As the parameters in the models are tied to specific mechanisms, we can address the source of the correlation between the tasks at a parameter level without having to make imprecise verbal assumptions about the relation between mechanisms operating in the two tasks. The general model itself can be simplified without loss of argument and applied to the complex enterprise of not only modelling individual differences in task performance, but also the correlations among tasks.

If the argument is that co-variability in the stop-signal and Stroop task is due to shared variability of a single executive function referred to as response inhibition, then our results question this strong statement. First, the only mechanism that is truly shared between the tasks is the strength of the pre-potent response channel. Given that this channel is the one that is the target of executive control and thus cannot be considered to be an executive control function itself, we see no basis to assume that a shared inhibition-type of executive function underlies the behavioural correlation. Second, the mechanisms that have been assumed and shown in modelling to be critical in overcoming the pre-potency are different between the tasks, thus a single inhibition-type of executive function is not an appropriate label. Instead, if these mechanisms are correlated, then a more appropriate label might be "goal potency". We elaborate on this view below. Third, if the shared inhibition function is taken literally and the inhibition mechanisms are correlated, then the simulation suggests that a negative correlation should be found between the tasks. However, the behavioural studies show a positive correlation between the tasks. This is in the context of literature that claims a positive correlation between each task and a latent inhibition factor. These points together argue against the use of a response inhibition construct in the individual differences literature as a mechanistic explanation for the behavioural correlation.

We suggested that the correlation between the tasks is due to the potency or strength of the current goal. More specifically, the computational studies are consistent with either a unitary mechanism that affects the rate of activation accumulation or one that relates to the level of the maximum possible activation. Both of these are emergent from an activation-based framework in which perceptual or cognitive information is actively maintained through self-excitatory loops (Davelaar, et al., 2005). Whether they can be distinguished remains to be demonstrated. However we

note that in a further part Miyake et al.'s (2000) study, it was shown that the common factor underlying performance on the stop-signal and Stroop tasks dissociated from a factor common to performance on several other tasks that were held to require a further executive function, referred to as *task-shifting*. A full account must therefore relate, in computational terms, the function isolated in this study and a task-shifting function. This is particularly important as Gilbert and Shallice (2002) consider task-shifting in the context of the Stroop task, and account for it in a model closely related to the Cohen and Huston model that forms the basis for part of this work.

The idea of goal potency has some support from other areas of cognitive neuroscience. Thus, Duncan et al. (2008) refer to the inability to execute a goal on presentation of a stimulus, even though knowledge about the rules regarding stimulus and response is present, as *goal-neglect*. Individuals differ in the degree to which they exhibit goal-neglect. If goal-neglect (or a factor underlying it) lies behind our factor, then one would expect that the proportion of stop-errors and the size of the Stroop interference effect should both be positively correlated with measures of goal-neglect. We know of no study that has investigated the correlation between stop-errors and goal-neglect.

We have focused only on the stop-signal and the Stroop task. As noted in the introduction, Miyake et al. (2000) also considered the antisaccade task. This task requires an eye-movement away from a distractor stimulus when this stimulus appears. In the Miyake et al. study the dependent measure for this task was the proportion of correct trials. Thus, overcoming pre-potency increases the score. This is important, as for the stop-signal and Stroop tasks, overcoming the pre-potency decreases the corresponding dependent measure. Consequently one might expect a negative correlation between the measures. Instead a positive correlation was found between the antisaccade task and both tasks. This is inconsistent within the response inhibition view. However, processes of active maintenance or activation accumulation can account for positive correlation where overcoming prepotent responses would expect negative correlations. In all but the antisaccade task, the stimulus conveys information that is used in activation of the relevant goal. In the antisaccade task, the first stimulus is a distractor and does not convey positive information, while the second is the target. Therefore being able to quickly activate information will produce less accurate responses. This leads to more activation producing lower levels of the dependent measure (accuracy) in the antisaccade task, together with more activation leading to lower levels of the dependent measures in the stop-signal (proportion stop-errors) and Stroop (interference effect) tasks. Our argument therefore is that the latent factor in the Miyake et al. studies reflects an activation-based function, and not an inhibition function.

This work also demonstrates more generally the importance of using explicit formal analyses to uncover the mechanisms underlying latent cognitive constructs.

References

Aron, A. R., Robbins, T. W., & Poldrack, R. A. (2004). Inhibition and the right inferior frontal cortex. *Trends in Cognitive Sciences*, *8*(4), 170-177.

Boucher, L., Palmeri, T. J., Logan, G. D., & Schall, J. D. (2007). Inhibitory control in mind and brain: An interactive race model of countermanding saccades. *Psychological Review*, *114*(2), 376-397.

Cohen, J. D. & Huston, T. A. (1994). Progress in the use of interactive models for understanding attention and performance. In *Attention and performance 15: Conscious and nonconscious information processing*. C. Umiltà & M. Moscovitch (Eds) (pp. 453-476). Cambridge, MA, US: The MIT Press.

Davelaar, E. J. (2009). Conflict-monitoring and (meta)cognitive control. In J. Mayor, N. Ruh, & K. Plunkett (Eds.), *Connectionist models of behaviour and cognition II.* (pp. 241-252). Singapore: World Scientific.

Davelaar, E. J., Goshen-Gottstein, Y., Ashkenazi, A., Haarmann, H. J., & Usher, M. (2005). The demise of short-term memory revisited: empirical and computational investigations of recency effects. *Psychological Review, 112,* 3-42.

Duncan, J., Parr, A., Woolgar, A., Thompson, R., Bright, P., Cox, S., et al. (2008). Goal neglect and Spearman's g: Competing parts of a complex task. *Journal of Experimental Psychology: General*, *137*(1), 131-148.

Egner, T., & Hirsch, J. (2005). Cognitive control mechanisms resolve conflict through cortical amplification of task-relevant information. *Nature Neuroscience*, *8*(12), 1784-1790.

Frank, M. J. (2006). Hold your horses: A dynamic computational role for the subthalamic nucleus in decision making. *Neural Networks*, *19*(8), 1120-1136.

Frank, M. J., Seeberger, L. C., & O'Reilly, R. C. (2004). By carrot or by stick: Cognitive reinforcement learning in Parkinsonism. *Science*, *306*(5703), 1940-1943

Friedman, N. P., & Miyake, A. (2004). The relations among inhibition and interference control functions: A latent-variable analysis. *Journal of Experimental Psychology: General*, *133*(1), 101-135.

Gilbert, S. J., & Shallice, T. (2002). Task Switching: A PDP Model. *Cognitive Psychology*, *44*(3), 297-337.

Logan, G. D., & Cowan, W. B. (1984). On the ability to inhibit thought and action: A theory of an act of control. *Psychological Review*, *91*(3), 295-327.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager T. D. (2000). The unity and diversity of Executive Functions and their contributions to complex "Frontal Lobe" tasks: A latent variable analysis. *Cognitive Psychology*, *41*(1), 49-100.

Stafford, T. & Gurney, K.N. (2007). Biologically constrained action selection improves cognitive control in a model of the Stroop task. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362 (1485), 1671-1684.

Estimation of Trade-off between Costs of Preprocessing and Primary Processing

Akihiro Maehigashi (mhigashi@cog.human.nagoya-u.ac.jp)
Kazuhisa Miwa (miwa@is.nagoya-u.ac.jp)
Graduate School of Information Science Nagoya University, Furo-cyo, Chikusa-ku, Nagoya, Japan

Abstract

People often conduct preprocessing to simplify primary processing. Usually, there is a trade-off between the costs of performing preprocessing and primary processing. Therefore, the utility of preprocessing differs depending on the task complexity. We conducted three experiments to find out whether people could adaptively estimate the utility of preprocessing and then take rational action. The overall result was that in performing a high complexity task, almost all the participants made a rational choice. However, for a low complexity task, the participants gradually learned to conduct preprocessing despite it not being effective. These results were explained based on theoretical perspectives proposed in previous studies.

Keywords: Strategy selection; Task environment; Cost; Time; Preprocessing

Introduction

When people engage in a task, they often create and conduct an additional preliminary task in order to conduct the main task more easily by using the result of the additional task. For example, people create preliminarily index cards for documents so that needed documents can be easily found. Moreover, people program preliminarily macros on a computer so that data can be easily processed later. In this study, we call such preliminary processing "preprocessing". People conduct preliminarily preprocessing so that primary processing in the task can be easily carried out. Preprocessing is often performed in our daily lives.

Kirsh (1996) referred to a "complementary action" which redesigns a task environment before engaging in the task in order to complete the task easily. He explained: "Complementary actions are a part of a strategy for restructuring the environment to improve the speed, accuracy, and robustness of cognitive processes" (p. 442). Such complementary actions taken before engaging in a task are also considered to be preprocessing. Moreover, Martin and Schwartz (2009) described preprocessing as an expertised action and call it an "adaptive pattern" in their manuscript. They stated that "in the adaptive pattern, people take an initial period to explore or adapt their ideas, practices, and/or environment. They are slower to start, but they can make up the lost time if they make an appropriate adaptation" (p. 372).

When preprocessing is conducted, the cost in primary processing is reduced. However, since conducting the preprocessing itself incurs a cost, there is a trade-off between the costs of preprocessing and primary processing. In such a situation, the utility of preprocessing seems to differ depending on the task complexity. When primary processing is conducted in a task without preprocessing, the task completion time increases with the task complexity. In contrast, when preprocessing is conducted for a certain period of time, it is considered that the increase in total task completion time for preprocessing and primary processing will be reduced, compared to the increase in the task completion time without preprocessing. Figure 1 illustrates our basic concept, showing the utility of preprocessing in relation to the task complexity. As shown in Figure 1, from the point where the task complexity (the amount of processing) exceeds a threshold level, the effectiveness of the preprocessing becomes significant. The utility of preprocessing differs depending on the task complexity; therefore, people should decide whether or not preprocessing is worthwhile depending on the situation.

Many researchers have studied cost estimation in situations where there is a trade-off between the costs of two different types of processing, e.g., a trade-off between processing using external resources (called external processing) and internal processing (Gray & Fu, 2004; O'Hara & Payne, 1998). In these studies, it was revealed that people adaptively estimate the costs of external processing and internal processing, and effectively adjust the usage of external resources depending on the cost of their use. Moreover, Matthew and Anderson (2009) found that people could adaptively determine whether or not external resources should be used depending on the task complexity. These studies show that people can adaptively estimate and allocate the costs of external processing and internal processing. On the other hand, in the current study, we investigate cost estimation in a situation where a different task needs to be conducted as preprocessing in order to reduce the cost of primary processing. The purpose of this study is to investigate whether people can adaptively estimate the utility of preprocessing and take rational action when there is a trade-off between the costs of preprocessing and primary processing as described above.

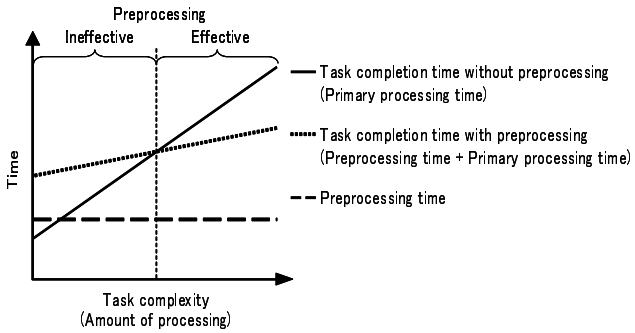


Figure 1: Concept diagram of trade-off between costs of preprocessing and primary processing

Experimental task

In the following experiments, we used a routine task of transcribing scores from test sheets to a tally sheet. In this task, fifty test sheets were prepared. On each test sheet, a student ID number was printed in the upper left and a test score in the center. In the experiments, three trials were conducted. In each trial, all the scores on the fifty individual test sheets had to be transcribed to a tally sheet to correspond to each student ID number.

Task complexity Each student ID number consisted of an eleven-digit number that encoded four categories, Grade, Major, Course, and Individual number. We set up a situation in which there were two grades, two majors, and five courses, and also there were twenty students in each course. Each student ID number on the test sheet was represented differently from that of the tally sheet, but could be collated by a certain transformation rule. In order to find the right place to fill out in the tally sheet, the student ID number on the test sheet needed to be transformed by the rule so that the transformed number could be found in the tally sheet (Figure 2). In the high task complexity condition, the participants had to calculate the student ID numbers on the test sheets for the transformation using a certain formula. On the other hand, in the low complexity condition, a student ID number correspondence table was given to the participants for the transformation so that they could find the referred numbers in the tally sheet without performing a calculation.

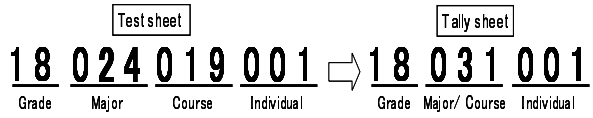


Figure 2: Transformation of student ID number; In the high task complexity condition, a transforming formula was used as follows: (1) add Major number to Course number ($24 + 19 = 43$). (2) multiply the first digit by the second digit of the result of (1) ($4 \times 3 = 12$). (3) subtract the result of (2) from the result of (1) ($43 - 12 = 31$). In the low task complexity condition, a correspondence table was used to transform the student ID numbers.

Preprocessing To carry out the preprocessing, a desk space was available for participants. Initially, the test sheets were arranged in random order and handed to the participants. They could rearrange the test sheets on the desk space so that they could group the test sheets according to the categories of Major and Course numbers. By preliminarily grouping the test sheets, the participants could transform the student ID numbers of a bundle of multiple test sheets in the same Major and Course number at one time. There was no need to transform each student ID number on each test sheet one by one. Therefore, the participants could reduce the number of transformations of the student ID numbers by preliminarily grouping the test sheets as a preprocessing task.

Experiment 1

Purpose

We investigated the relationship between the task performance and the task complexity, i.e., the task completion time and the number of errors respectively, when preprocessing is conducted or not.

Method

Participants Forty-six university students participated in this experiment.

Material A set of fifty test sheets made from A4 sized paper was prepared for each trial. Three trials were performed and a different set was used in each trial. All three sets were controlled using the number of times, and the order in which the operations of carrying and borrowing were required for the calculation to transform the student ID numbers. The tally sheet was made of A3 sized paper. A scenario with two grades, two majors, and five courses were conjectured with twenty students belonging to each course. Therefore, 400 empty cells ($= 2 \times 2 \times 5 \times 20$) were placed on the tally sheet. By transcribing all the scores to the tally sheet, fifty cells out of the 400 blank spaces had to be filled in. A desk with space large enough to accommodate ten A4 sized papers was used for preprocessing. Also, another desk was used for primary processing, that is, transcribing the scores to the tally sheet with a pencil.

Factorial design The experiment had a three-factor mixed design. The factors were: (1) Task complexity (high and low) between participants; (2) Preprocessing (preprocessing and no preprocessing) between participants; (3) Trial (1, 2, and 3) within participants.

Procedure In order to confirm participants' ability to calculate, they were required to solve computational problems that consisted of a total of 25 addition and multiplication problems. The main task was conducted three times with different sets of the test sheets and the tally sheets. At the beginning of each trial, the participants were informed of their task completion time and the number of errors in the previous trial as feedback. As a preprocessing condition, the participants had chosen a particular way of grouping the test sheets in the first trial and were instructed to rearrange them in the same way throughout the three trials.

Result

In order to maintain homogeneity of the participants' calculation ability, two participants whose computational time in the calculation problems fell outside 2 SD from the mean computational time for each condition were eliminated from the analysis. In addition, it was assumed that the participants who made too many errors in the task could not conduct the task appropriately, therefore three participants whose mean number of errors throughout the three trials fell outside 2 SD from the mean number of errors in each condition were eliminated from the analysis. Moreover, one participant who violated the instructions, i.e., conducted the transforming calculation on the desk space for preprocessing, was eliminated from the analysis. In the following experiments, the identical criteria were used for selecting appropriate participants. As a result, the performance of forty participants, of whom ten were assigned to each condition, was analyzed.

There was no significant difference in performing the calculation problems between the four conditions ($F(3,36) = .07, n.s.$). Therefore, the calculation abilities of the participants among the four conditions were considered equivalent.

Task completion time On the task completion time, a 2 (Task complexity: high/low) $\times$ 2 (Preprocessing: preprocessing/no preprocessing) $\times$ 3 (Trial) ANOVA was conducted. As a result, there was no significant three-way interaction ($F(2,72) = 2.17, n.s.$). There was a significant two-way interaction between task complexity and preprocessing ($F(1,36) = 9.80, p < .005$). There was neither significant two-way interactions between task complexity and trial

($F(2,72) = 1.83, n.s.$) nor between preprocessing and trial ($F(2,72) = .46, n.s.$). Figure 3 illustrates the mean task completion time in Experiment 1 based on the basic concept depicted in Figure 1. In Figure 3, the preprocessing time was measured from the time when the test sheets were put on the desk space for grouping (preprocessing) until they were lifted up for transcribing the scores to the tally sheet (primary processing).

Next, we conducted a simple main effect test on the preprocessing factor. As a result, (1) in the high task complexity condition, there was a marginally significant difference showing that the task completion time was faster in the preprocessing condition than in the no preprocessing condition ($F(1,36) = 3.19, p < .10$), whereas (2) in the low task complexity condition, it was significantly faster in the no preprocessing condition than in the preprocessing condition ($F(1,36) = 6.98, p < .05$).

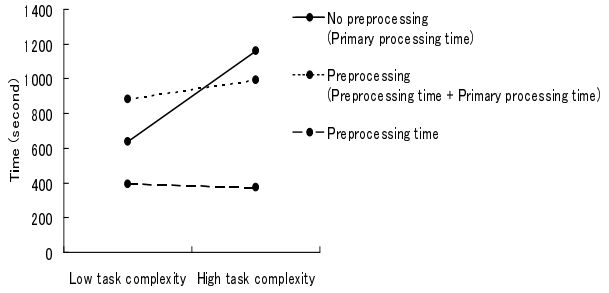


Figure 3: Task completion time in Experiment 1 represented by the basic concept

Number of errors We defined a transcribing error as being a transcription of an incorrect score or a transcription to an inappropriate cell. On the number of errors, a 2 (Task complexity: high/ low) $\times$ 2 (Preprocessing: preprocessing/no preprocessing) $\times$ 3 (Trial) ANOVA was conducted. As a result, there was no significant three-way interaction ($F(2,72) = .61, n.s.$). There was a significant two-way interaction between task complexity and preprocessing ($F(1,36) = 4.37, p < .05$) and a marginally significant interaction between task complexity and trial ($F(2,72) = 2.56, p < .10$). There was no significant interaction between preprocessing and trial ($F(2,72) = 1.98, n.s.$).

Next, we conducted a simple main effect test on the preprocessing factor. As a result, (1) in the high task complexity condition, the number of errors was significantly smaller in the preprocessing condition than in the no preprocessing condition ($F(1,36) = 11.12, p < .005$), whereas (2) in the low task complexity condition, there was no significant difference between the preprocessing and no preprocessing conditions ($F(1,36) = .14, n.s.$).

Discussion

As a result of Experiment 1, it is revealed that conducting preprocessing is effective for the high complexity task, and contrarily, not conducting preprocessing is effective for the low complexity task. These results proved that our transcribing task is an appropriate task for embodying a trade-off between preprocessing and primary processing.

In the following Experiment 2, using the same task, we investigated whether people could adaptively estimate the utility of preprocessing and take rational action depending on the task complexity.

Experiment 2

Purpose

Using the transcribing task, we investigated whether people could adaptively estimate the utility of preprocessing and take rational action depending on the task complexity.

Method

Participants Twenty-seven university students participated in this experiment.

Material Identical materials were used as in Experiment 1.

Factorial design The experiment had a two-factor mixed design. The factors were: (1) Task complexity (high and low) between participants; (2) Trial (1, 2, and 3) within participants.

Procedure Basically an identical procedure to that of Experiment 1 was followed. In Experiment 2, the participants were instructed: "it is allowed to rearrange and group the test sheets, but it is not a requirement to do so." In addition, when the participants chose to conduct preprocessing at the beginning of each trial, they were allowed to decide their own way of rearranging the test sheets.

Result

Three participants were excluded from the analysis based on the same criterion as in Experiment 1. As a result, the performance of twenty-four participants, of whom twelve were assigned to each condition, was analyzed. First, there was no significant difference in performing the calculation problems between the two conditions ($t(22) = .23, n.s.$).

Preprocessing and minimal transformation strategy In Experiment 2, we calculated the ratio of participants conducting preprocessing for each condition. Moreover, there was a rearranging strategy with which the participants could transform the student ID numbers with the minimum number of times in primary processing. This strategy could minimize the cost of primary processing. In particular, this strategy was to group the test sheets according to the categories of Major and Course numbers first. We calculated the ratio of participants using this minimal transformation strategy for each condition. Figure 4 shows the ratio of participants conducting preprocessing and the ratio of participants using the minimal transformation strategy. In the low task complexity condition, the participants gradually learned to conduct preprocessing from the first to third trial. On the other hand, in the high task complexity condition, the participants conducted preprocessing from the first trial. Moreover, the minimal transformation strategy was used more in the high task complexity condition than in the low task complexity condition.

Discussion

As a result of Experiment 2, when the participants were allowed to choose whether to conduct preprocessing or not, the participants conducted ineffective preprocessing in the low complexity condition. Moreover, the minimal transformation

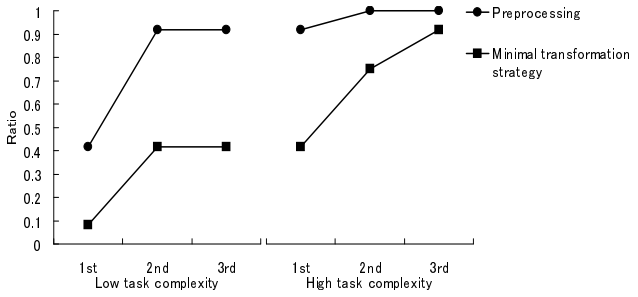


Figure 4: Ratio of participants conducting preprocessing and ratio of participants using minimal transformation strategy in Experiment 2

strategy was used more for the high complexity task than for the low complexity task.

At this point, we have questions. Was it possible the participants were trying to reduce fatigue by grouping the test sheets as a routine work throughout the trials although they realized that conducting preprocessing was ineffective for the low complexity task? Moreover, was it also possible for them to try to reduce the number of errors by grouping the sheets although they noticed that conducting preprocessing increased the task completion time for the low complexity task? To answer these questions, in Experiment 3, we conducted a questionnaire directly asking the participants which makes the task faster and more accurate, preliminarily grouping the test sheets as preprocessing or not. Furthermore, in Experiment 2, at the beginning of each trial, the participants were told the task completion time and the number of errors in the previous trial as feedback. Throughout the trials, the task completion time gradually decreased because of the learning effect. Consequently, the participants might have misunderstood that preprocessing is effective because of the effect of this feedback. Therefore, in Experiment 3, we gave no feedback to the participants.

Experiment 3

Purpose

We replicated Experiment 2 and confirmed whether people could adaptively estimate the utility of preprocessing and take rational action in the low complexity task.

Method

Paticipants Seventeen university students participated in this experiment.

Material Identical materials were used as in Experiment 2.

Procedure Basically an identical procedure to that of Experiment 2 was followed. In Experiment 3, at the beginning of each trial, we gave the participants a questionnaire asking them to estimate the utility of preprocessing for the task completion time and the accuracy. The participants were instructed to choose one out of four choices: (1) preprocessing is effective, (2) no preprocessing is effective, (3) no difference, and (4) impossible to estimate. When the participants chose to conduct preprocessing at the beginning of each trial, they were allowed to decide their own way of rearranging the test sheets. In addition, the participants were instructed to estimate the task completion time and the number of errors after each trial had been completed. They were neither informed of the actual task completion time nor the number of errors as feedback. Moreover, in Experiment 3, we set up the fourth trial in which the participants were not allowed to conduct preprocessing in order to compare the performance with the performance when preprocessing was conducted in the former three trials. Also, the participants were instructed to estimate the task completion time and the number of errors after the fourth trial had been completed.

Result

Two participants were excluded from the analysis based on the same criterion as in Experiment 1. As a result, the performance of fifteen participants was analyzed.

Preprocessing and minimal transformation strategy Figure 5 shows the ratio of participants conducting preprocessing and the ratio of participants using the minimal transformation strategy. The participants gradually learned to conduct preprocessing from the first to third trial. Moreover, the ratio of participants using the minimal transformation strategy was low.

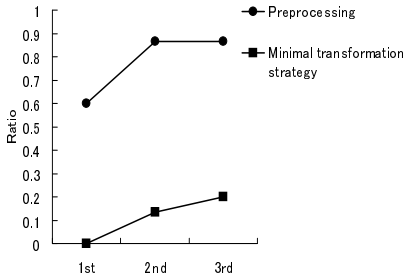


Figure 5: Ratio of participants conducting preprocessing and ratio of participants using minimal transformation strategy in Experiment 3

Estimation of preprocessing utility Figures 6 and 7 show the numbers of choices made in the questionnaire at the beginning of each trial for the task completion time and the accuracy. With each subsequent trial, the participants shifted towards estimating that conducting preprocessing is effective in producing a more rapid performance. They also either estimated that preprocessing was effective or produced no difference in performance accuracy.

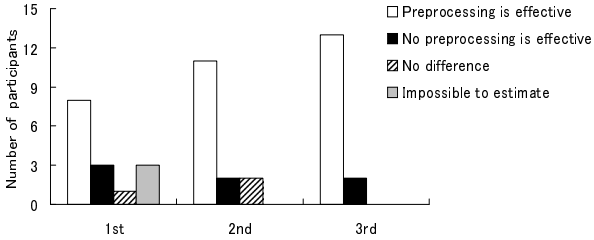


Figure 6: Respondents estimation for task completion time

Actual/ Estimated performance In Experiment 3, almost all participants conducted preprocessing as in Experiment 2. Consequently, we compared the participants' performance when preprocessing was conducted in the first three trials with their performance in the fourth trial where they were not allowed to conduct preprocessing. In particular, the mean

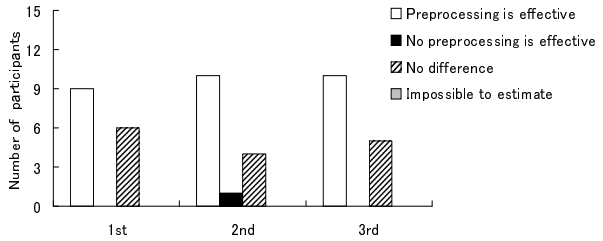


Figure 7: Respondents estimation for accuracy

performance of nine participants, who had conducted preprocessing in all the first three trials, and four participants, who had conducted preprocessing in the second and third trials, was calculated. The result was regarded as the representative performance for when preprocessing was conducted. The performance of these thirteen participants in the fourth trial was regarded as the representative performance for when preprocessing was not conducted. Moreover, we compared the participants estimated performance when preprocessing was conducted in the first three trials with their estimated performance in the fourth trial.

Actual/ Estimated task completion time Figure 8 shows the actual task completion time and the estimated task completion time. We conducted a t-test on the actual task completion time when preprocessing was conducted and when it was not conducted. As a result, the actual task completion time was significantly faster when preprocessing was not conducted than when conducted ($t(12) = 4.49, p < .001$). Moreover, we conducted a t-test on the participants estimated task completion time when preprocessing was conducted and when it was not conducted. As a result, there was a marginally significant difference showing that their estimated time was faster when preprocessing was conducted than when not conducted ($t(12) = 1.89, p < .10$).

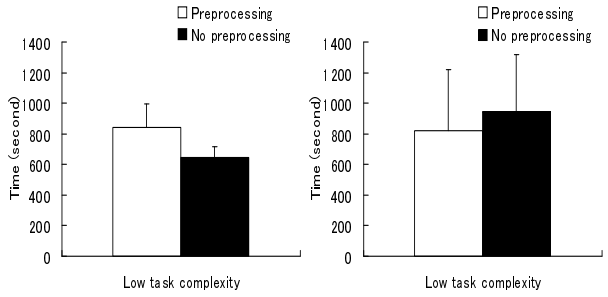


Figure 8: Comparisons of actual task completion time (left) and estimated task completion time (right) when preprocessing was conducted and when it was not conducted

Actual/ Estimated number of errors We conducted t-tests on the actual number of errors and on the participants estimated number of errors when preprocessing was conducted and when it was not. As a result, there was neither significant difference in the actual number of errors ($t(12) = .66, n.s.$) nor in their estimated number of errors ($t(12) = 1.54, n.s.$).

Discussion

First, as a result of Experiment 3, when the participants were allowed to choose whether to conduct preprocessing or not, the participants conducted ineffective preprocessing as in Experiment 2. This result eliminated the possibility that the feedback from their previous performance had encouraged the participants to mistakenly elect to perform preprocessing. Second, the ratio of participants using the minimal transformation strategy was also as low as in Experiment 2. Third, the results of the questionnaire indicated the participants overestimated the utility of preprocessing for the task completion time and the accuracy. This result eliminated the possibility that the participants conducted preprocessing as a strategy for reducing fatigue and reducing the number of errors, because they had reported that conducting preprocessing could reduce the task completion time in the questionnaire. Moreover, we compared the performances when preprocessing was conducted and when it was not conducted in the within-participant experiment. As a result, we confirmed the result was consistent with that of Experiment 1. Contrarily, the participants estimated the actual task completion time faster when preprocessing was conducted than when not conducted. This result is consistent with the result of participants' estimation for the task completion time in the questionnaire given at the beginning of each trial.

General Discussion

The purpose of this study was to investigate whether people can adaptively estimate the utility of preprocessing and take rational action. First, Martin and Schwartz (2009) investigated preprocessing to create representational tools before engaging in a task. However, they evaluated the performances of preprocessing and primary processing separately and did not address the issue of a trade-off between the costs of the two types of processing. They used a learning task to investigate how learning experiences in preprocessing influence the following behavior for learning. In contrast, in this study, we used a problem solving task. In order to investigate the utility of preprocessing, it is crucially important, especially in problem solving tasks, to consider the trade-off between increasing the cost of preprocessing and decreasing the cost of primary processing.

As a result of our experiments, in the low complexity task, preprocessing was aggressively conducted despite it not being effective. In our experimental task, preprocessing was performed with a desk space as an external resource. Brown, Collins, and Duguid (1989a, 1989b) suggested that people actively use external resources at the initial stage as an initial human impulse. In addition, Kirsh (2009) referred to the activity-centric model as an instinctive human behavior of using external resources without thinking. The human nature to instinctively use external resources described in the research of Brown et al. and Kirsh may explain the participants' behavior in our experiments.

Sirouzu, Miyake, and Masukawa (2002) experimentally confirmed such a human nature. They suggested that people actively use external resources as their "proto-plan". In their experiment, the existence of external resources prevented the participants from noticing the availability of usable internal processing. In contrast, in our experiments, the participants conducted preprocessing although they were explicitly offered the choice of using preprocessing or not. Moreover, Sirouzu et al. (2002) stated that the participants divided a sin-

gle over-all task into multiple simpler sub-tasks using an external resource so that they could visually confirm the completion and the result of each sub-task, and plan the next step. In our experiments, the participants could divide one transcribing task into two different tasks: rearranging the test sheets (preprocessing) and transcribing scores (primary processing). The participants could take advantage of the result of preprocessing, allowing them to conduct primary processing smoothly and easily. It is considered that the effect of such task decomposition causes the overestimation of the utility of preprocessing. Kirsh (1996) explained such human action of task decomposition as a complementary action, which enables the externalization of plans into sub-goals in order to easily achieve a final goal and that this is a central element of human activities.

The result of our experiments about the estimation of a trade-off between two types of processing is not consistent with the findings of Matthew and Anderson (2009). In their experiment, the participants had to choose between solving each problem using a calculator as an external resource or by using mental calculation. The participants were able to make the correct choice whether to use the external resource or not, dynamically and instantly, depending on the task complexity. Matthew and Anderson (2009) used calculation problems that each took around ten seconds to solve. In contrast, in our experiments, we used the transcribing task that took around ten minutes to complete. In order to conduct preprocessing, the participants had to create an additional sub-task as a preliminary task, once stepping away from the primary task, and thus conduct two different types of tasks sequentially. One reason why the participants failed to estimate the costs might be because the cost estimation in our task was much harder than such estimation in the previous study.

Another reason may depend on the participants' time perception. The task completion time in the low complexity task was estimated to be faster when preprocessing was conducted than when not conducted. This trend in estimation was opposite to that of the actual task completion time. In studies of time perception, it has been verified that the more cognitive processing people perform, the less attention they direct to time, so that they underestimate time duration. This phenomenon is explained by the attentional model (Hicks, Miller, & Kinsbourne, 1976; Zakay, 1993). Hicks et al. (1976) investigated the attentional model using a card sorting task. As a result, it was revealed that the more stacks the cards were sorted into, the faster the task completion time was estimated to be, because more cognitive processing was performed as there were more stacks to sort. In our experiments, when preprocessing was conducted, the participants had to sort the test sheets. Therefore, there is a possibility that the participants estimated the task completion time faster when preprocessing was conducted than when not conducted because they performed more cognitive processing when preprocessing was conducted than when it was not conducted in the low complexity task.

Last, the minimal transformation strategy was used more in the high complexity task than in the low complexity task. Cary and Carlson (1999) found that in a situation where high internal costs were demanded, the participants tended to use a strategy to minimize their internal costs. On the other hand, in a situation where low internal costs were demanded, the participants chose a strategy of following structures in the problem, and did not focus on reduction of internal processing. In our experiments, in the high complexity task, the participants rearranged the test sheets according to the categories of Major and Course numbers first. Using this strategy, the participants successfully transformed the student ID numbers for referring to the numbers on the tally sheet using the minimum number of times and minimized internal cost. On the other hand, in the low complexity task, the participants tended to rearrange the test sheets according to the category of Grade number first, affected by the structure of the student ID number in which the first two digits represented the Grade. This meant that they used a strategy to follow the structure of the problem. Our results were consistent with the findings of Cary and Carlson (1999).

References

Brown, J. S., Collins, A., & Duguid, P. (1989a). Debating situation: A rejoinder to palinscar and wineburg. *Educational Researcher*, *18*, 10–12.

Brown, J. S., Collins, A., & Duguid, P. (1989b). Situated cognition and the culture of learning. *Educational Researcher*, *18*, 32–42.

Cary, M., & Carlson, R. A. (1999). External support and the development of problem-solving routines. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 1053–1070.

Gray, W. D., & Fu, W.-T. (2004). Soft constraints in interactive behavior: the case of ignoring perfect knowledge in-the-world for imperfect knowledge in-the-head. *Cognitive Science*, *28*, 359–382.

Hicks, R. E., Miller, G. W., & Kinsbourne, M. (1976). Prospective and retorospective judgements of time as a function of amount of information processed. *American Journal of Psychology*, *89*, 719–730.

Kirsh, D. (1996). Adapting the environment instead of oneself. *Adaptive Behavior*, *4*, 415–452.

Kirsh, D. (2009). Problem solving and situated cognition. In P. Robbins & M. Aydede (Eds.), *The cambridge handbook of situated cognition.* Cambridge University Press.

Martin, L., & Schwartz, D. L. (2009). Prospective adaptation in the use of external representations. *Cognition and Instruction*, *27*, 370–400.

Matthew, M. W., & Anderson, J. R. (2009). The strategic nature of changing your mind. *Cognitive Psychology*, *58*, 416–440.

O'Hara, K. P., & Payne, S. J. (1998). The effects of operator implementation cost on planfulness of problem solving and learning. *Cognitive Psychology*, *35*, 34–70.

Sirouzu, H., Miyake, N., & Masukawa, H. (2002). Cognitively active externalization for situated reflection. *Cognitive Science*, *26*, 469–501.

Zakay, D. (1993). Time-estimation methods - do they influence prospective duration estimates. *Perception*, *22*, 91–101.

Increasing Information Access Cost to Protect Against Interruption Effects during Problem Solving

Phillip L. Morgan (MorganP4@Cardiff.ac.uk)
John Patrick (PatrickJ@Cardiff.ac.uk)
Tanya Patrick (PatrickTP3@Cardiff.ac.uk)
School of Psychology, Cardiff University, Tower Building,
Park Place, Cardiff, CF10 3AT, UK

Abstract

The aim of this experiment was to examine whether increasing the cost of accessing the goal-state during problem solving would induce a more internalized strategy that would protect against the negative effect of interruption. The soft constraints hypothesis (Gray, Sims, Fu & Schoelles, 2006) predicts that a more memory-based strategy will be developed with increasing information access cost (IAC). Three levels of access cost were used in the Tower of Hanoi (ToH) with three types of interrupting task (simple ToH, mental arithmetic and a blank screen control). Increasing access cost to a mouse movement and a few seconds delay to view the goal-state encouraged a strategy that not only improved resumption from memory but also reduced the number of moves required to solve the primary task. These effects came at no extra time cost and occurred irrespective of the type of interrupting task. The theoretical implications of these findings are discussed together with issues of using access cost as a method for alleviating the negative effects of interruption.

Keywords: Interruption; Problem Solving; Goal-State Access Cost

Introduction

Interruptions are intrinsic to our everyday and working lives (e.g., telephone calls, emails), and although often useful (e.g., responding to another important task), they are also often associated with performance decrements when returning to the interrupted task. Difficulties include: problems remembering what one was doing or intended to do prior to being interrupted (e.g., Edwards & Gronlund, 1998; Morgan, Patrick, Waldron, King & Patrick, 2009); and delays in resuming the interrupted task (e.g., Hodgetts & Jones, 2006; Monk, Trafton & Boehm-Davis, 2008). Whilst these issues are sometimes tolerable (e.g., when taking a telephone call whilst buying groceries), interruptions can be a nuisance, expensive or even life threatening in many other contexts, including, offices, aircraft flight-decks, and hospitals (see Trafton & Monk, 2008). Thus, investigating methods for minimizing interruption effects has become an important topic. Methods include: using an 'interruption lag' to briefly delay the onset of an interruption and allow an opportunity to prepare for resumption of a problem solving task (Hodgetts & Jones, 2006); and a reminder cue to support memory of a delayed intention in a prospective memory task (McDaniel, Einstein, Graham & Rall, 2004). A recent study by Morgan et al. (2009) developed another method that involved increasing goal-state access cost (the time, physical and mental effort costs associated with accessing information). This encouraged a more memory-based strategy that was effective in protecting against forgetting following interruption. The aim of this paper is to investigate the efficacy of this method in a problem solving task rather than the copying task used by Morgan et al. (2009).

The predicted advantage of increased access cost on promoting a more internal cognitively-based strategy derives from the soft constraints hypothesis (Gray et al., 2006). This theory posits that whilst certain elements of a task environment are fixed (i.e., hard constraints) and dictate what behavior is possible, task strategies are flexible and therefore adapt in a rational manner. People strive to minimize the time spent performing tasks at a local rather than global level (Gray et al., 2006), so strategy adjustments are made at the 1/3 to 3-second level of task performance favoring those that are more effective at this millisecond level (Gray & Boehm-Davis, 2000). On one hand, if information is readily available in the task environment, a strategy that relies less on internal memory (which is fallible and subject to error) will prevail (e.g., Anderson & Douglass, 2001). In contrast, if there is an unacceptable cost associated with accessing information in the task environment, cognition will adapt and adjust to a more memory-based strategy to minimize this cost.

Such strategy change has been demonstrated in a variety of studies (Gray & Fu, 2004; Gray et al., 2006; Morgan et al., 2009; Waldron et al., 2007) using the Blocks World Task (BWT) developed by Ballard, Heyhoe & Pelz (1995), that involves copying a target pattern of colored blocks to a workspace window. Increasing the cost of accessing the target pattern with a mouse movement and a brief time delay induced a shift to a more memory-based strategy (e.g., Gray & Fu, 2000; Gray et al., 2006) and improved recall of information (Waldron, Patrick, Morgan & King, 2007). Morgan et al. (2009) found that such an induced memory-based strategy protected against forgetting following both visuo-spatial copying and mental arithmetic interrupting tasks. This was particularly effective when the interruption occurred on approximately half of the trials.

In contrast, little is known about the effect of goal-state access cost on problem solving. Some studies have manipulated the availability and effort required to use the task environment as an external memory resource. For example, the use of internal memory increased when the ability to make paper notes was made more difficult whilst

solving demanding mental arithmetic problems (Cary & Carlson, 2001). Similarly, when the current-state had to be requested during performance of 'balls and boxes' problems, participants tended to execute more moves per request (Pfeiffer, 2004). Also, O'Hara and Payne (1998) demonstrated how increasing the cost of implementing an action during problem solving leads to improved planning. The results from these studies suggest increased use of internal memory to avoid additional costs of interacting with the environment. One recent problem solving study found that High goal-state access cost led to more 'planning before action' as opposed to 'planning during action' (Waldron, Patrick & Duggan, unpublished) although it had no effect on number of moves to solution. However the effect of High goal-state access cost on mitigating the negative effect of interruption has not been examined during problem solving.

Experiment

To fill this research gap, the aim of the experiment was to examine whether increasing goal-state access cost induces a more memory-based strategy in a ToH task and whether such a strategy can improve performance following interruption. It was predicted that High access cost would accomplish this by encouraging a more internalized rather than display-based strategy with more planning before action (Davies, 2003; Waldron, Patrick & Duggan, unpublished).

There were two subsidiary aims. First, Morgan et al. (2009, Experiment 3) demonstrated that High access cost induced a memory-based strategy that was powerful enough to abolish the effects of forgetting following different types of interrupting tasks compared to a no interruption condition, even when one task (another BWT) arguably had similar processing requirements to the primary task. There is mixed evidence regarding the effects of interruption similarity on post-interruption performance. Some studies report greater disruption following a similar interrupting task (e.g., Edwards & Gronlund, 1998; Gillie & Broadbent, 1989) whereas others argue against this so called 'interruption similarity' effect (e.g., Latorella, 1996). Given that it is difficult to unequivocally separate tasks on dimensions of similarity, we examined the effects of increasing goal-state access cost on performance following two different types of interrupting tasks: one involving another ToH (arguably similar to the primary task) and the other involving mental arithmetic.

Second, we wanted to assess whether the typical cost of High access cost on reducing speed of completing the BWT (Morgan et al., 2009) would be less pronounced or even eliminated in the ToH. Performing the BWT under a High access cost places a high demand on memory and participants spend time encoding and rehearsing block information, suffering the time cost of uncovering the goal-state at each visit, and making more move errors that have to be corrected. In contrast, the ToH is not as memory demanding given that the goal-state often consists of a small array of objects (e.g., 4-discs) that are bounded by a simple constraint (e.g., a larger disc cannot be placed on top of a smaller disc). Also, the hierarchical nature of the ToH in terms of goal and subgoals means that many moves are interdependent, and therefore the goal-state does not need to be re-visited as often as in the BWT. Furthermore, the proposed benefit of High goal-state access cost in the ToH is due to its encouragement to develop a more effective problem solving strategy that should lead to fewer moves and possibly reduced time to solve the problem.

Method

Participants

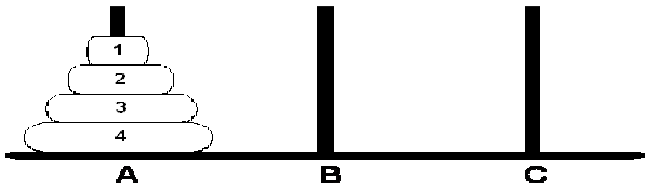
Fifty-four Cardiff University Psychology students participated for course credit and were randomly assigned to one of three goal-state access cost conditions. There were six men and forty-eight women with an age range of 18 to 37 years (M = 20.56, SD = 2.95).

Materials, Design & Procedure

There were 18 primary task four-disc ToH problems, each with a different start and goal-state disc configuration (see Figure 1 for an example start-state). Disc movement was controlled by clicking on a disc on one peg (e.g., A), holding down the mouse button, and dragging the disc to a destination peg before releasing the mouse button (e.g., at B or C). If a larger disc was dragged over a smaller disc, the larger disc would be returned to its source peg and a message reading 'illegal move' would appear on the screen. All primary task ToH problems required a minimum of 15 moves for error-free completion and all major subgoals occurred at the same point within the solution sequence (if solved error-free: subgoal 1, moves 1 – 8; subgoal 2, moves 9 – 12; and subgoal 3, moves 13 – 15). The main subgoals within the ToH task involve getting the largest-out-of-place discs to their goal destination peg(s) in the order 'largest first' through to 'smallest last'.

Figure 1. An example of the start-state of a four disc ToH task

Note. Disc numbers are used to represent colors (1 = green, 2 = blue, 3 = yellow, and 4 = red)



Goal-state access cost was manipulated between-subjects on three levels: Low (goal-state always visible), Medium (goal-state covered with a grey mask with a mouse movement to uncover) and High (as Medium but with an additional 2.5-second lockout cost to uncover). Each of nine ToH problems was interrupted once with one of three different interruption types (manipulated within-subjects). The ToH interruption consisted of 3-disc ToH problems that could be solved in a minimum of seven moves using the

same disc movement procedure as in the primary task. The mental arithmetic (MA) interruption involved solving a series of self-paced double-digit addition problems (e.g., 34 + 45 = ??). Answers were entered using a keyboard. In the control condition (referred to as blank screen interruption hereafter), there was no task during the interruption. Each interruption occurred immediately following the first move of the first, second or third major subgoals, giving three interruption points. Both interruption task type and interruption position were counterbalanced.

Two types of measure were used to examine the effects of goal-state access cost and interrupting task on the ability to resume the primary task from memory. First, we calculated the number of interrupted trials resumed without first re-visiting the goal-state and, for those trials, the number of moves executed subsequently. Second, we calculated both the number of moves and the time required to complete the ToH problem following interruption. Given that interruptions were equally distributed across the three points of interruption, fair comparisons could be made between both the different levels of access cost and the different interrupting tasks. Finally, it was important to confirm that any benefit of increased goal-state access cost on performance following interruption was due to increased use of a more internalized problem solving strategy with reduced reliance on the external problem space. For this we used an important measure of planning in problem solving (e.g., Davies, 2003, 2005; Ward & Allport 1997), which was the amount of time spent at the beginning of a ToH problem before the first move was executed. This was predicted to increase with increasing access cost.

Participants were tested individually. They were instructed on task procedures and informed that they could be interrupted at any time. The main experiment started after completion of one non-interrupted 15-move practice trial and one attempt at performing each interrupting task. The experiment lasted approximately 30 minutes.

Results and Discussion

First, the effects of goal-state access cost and interruption type on the ability to resume the primary ToH task from memory were considered, followed by an examination of their effects on various post-interruption performance measures. Finally, we assessed whether any beneficial effect of High access cost on these performance measures could be accounted for by participants using a more memory-based problem solving strategy.

Effects of Goal-State Access Cost and Interruption Task on Performance after Interruption

We predicted that the main benefit of a higher access cost would be a tendency to adopt a more internalized problem solving strategy with participants choosing to rely less on the external problem space, even following interruption. As such, it was anticipated that participants in the High access cost condition would resume the primary task without revisiting the goal-state (and suffering the associated time cost) more often than the Medium access cost condition. (Note that the Low access cost condition could not be considered for any resumption measure because the goal-state was permanently uncovered.) The results supported this prediction (Table 1) with a 2 (goal-state access cost: Medium and High) x 3 (interruption type: blank screen, ToH, MA) ANOVA revealing that more trials were resumed from memory by participants in the High compared to the Medium access cost condition, $F(1, 34) = 67.43$, $MSE = 3.28$, $p < .001$. There was also a main effect of interruption type, $F(2, 68) = 9.47$, $MSE = 6.57$, $p < .001$, due to less use of the goal-state to aid resumption following a blank screen interruption compared to a ToH interruption. Furthermore, and as predicted, Bonferroni post-hoc tests revealed that participants in the High access cost condition resumed more often without first viewing the goal-state than those in the Medium access cost condition following all interruption types ($ps < .001$).

Table 1. Effect of goal-state access cost and interruption task on resumption performance

	Goal-State Access Cost		Interruption Type		
			Blank	ToH	MA
Number of trials resumed without re-visiting the goal-state (max = 3)	Med	*M*	1.17	.11	.61
		SD	.86	.32	.78
	High	*M*	2.56	1.46	2.28
		SD	1.17	.94	.83
Number of participants resuming at least one trial without viewing goal-state	Med		13/18	2/18	8/18
	High		17/18	16/18	17/18
Number of moves executed without re-visiting the goal-state	Med	*M*	4.19	-	2.57
		SD	1.56	-	1.97
	High	*M*	6.78	-	6.55
		SD	3.30	-	3.87

Therefore it is evident that participants in different access cost conditions adopted different resumption strategies involving the need to view (or not) the goal-state to resume. For example, most participants in the Medium access cost condition were unable to resume at least one ToH interruption trial without first revisiting the goal-state window (Table 1). In contrast nearly all participants in the High access cost condition resumed at least one ToH trial from memory. This demonstrates the protective effect of High goal-state access cost on memory for the goal-state and/or a future move(s) following a ToH interruption *and* the ineffectiveness of Medium goal-state access cost following the same type of interruption. This beneficial effect of High access cost was reduced with the two other

interrupting tasks although it was still apparent for an MA interruption (Table 1).

Another indicator of participants relying less on the external display and more on an internal representation to continue an interrupted task is the number of moves they make before re-inspecting the goal-state (Table 1). A 2 (goal-state access cost) x 2 (interrupt type: blank screen and MA) ANOVA revealed a main effect of goal-state access cost, $F(1, 22) = 7.09$, $MSE = 7.54$, $p < .05$, with participants in the High access cost condition making more moves after interruption before goal-state re-inspection than those in the Medium access cost condition. There was no effect of interruption type and goal-state access cost and interruption type did not interact ($ps > .05$).

These results are testimony to the marked effect of High goal-state access cost on the ability to maintain an internal representation of the goal-state and/or a future move or series of moves throughout the course of an interruption, even when interruption involved a different task.

Whilst the above measures concerned interruption trials that were resumed without first viewing the goal-state window (and are thus restricted to the Medium and High goal-state access cost conditions), it is also important to establish the effects of all three levels of access cost on post-interruption performance (Table 2). Specifically, we were interested whether High access cost with its more internal memory-based strategy (1) led to fewer moves and (2) affected the time to complete interrupted ToH problems.

Participants in higher access cost conditions completed interrupted ToH problems in fewer moves (Table 2). A 3 (goal-state access cost: Low, Medium and High) x 3 (interruption type: blank screen; ToH, MA) ANOVA confirmed a significant main effect of goal-state access cost $F(2, 51) = 4.35$, $MSE = 2.53$, $p < .05$, $f = .41$ with participants in the High access cost condition completing problems in fewer moves than participants in the Low access cost condition ($p < .05$). However, participants in the High access cost condition did not perform significantly better than those in the Medium access cost condition ($p > .05$), and participants in the Medium access cost condition did not perform better than those in the Low access cost condition ($p > .05$). There was a non-significant main effect of interruption type ($p > .05$) and a non-significant interaction ($p > .05$).

Time to complete problems following interruption was similar across goal-state access cost conditions (Table 2), and there was a non-significant main effect ($p > .05$). There was, however, a significant effect of interruption type, $F(2, 102) = 7.77$, $MSE = 52.77$, $p < .01$, $f = .39$. Not surprisingly, participants were significantly faster to complete following a blank screen than a ToH interrupting task ($p < .01$) and were marginally faster following mental arithmetic than a ToH interrupting task ($p = .06$). Goal-state access cost and interruption type did not interact ($p > .05$).

Table 2. Effect of goal-state access cost and interruption type on performance following resumption

	Goal-State Access Cost		Interruption Type		
			Blank	ToH	MA
Number of moves to complete the primary task following interruption	Low	*M*	10.39	11.70	11.26
		SD	2.98	3.49	3.05
	Med	*M*	9.52	10.93	10.19
		SD	1.57	3.20	2.33
	High	*M*	9.28	9.89	9.52
		SD	2.25	1.93	2.29
Time to complete the primary task following interruption (s)	Low	*M*	20.65	25.79	22.16
		SD	9.28	8.84	8.18
	Med	*M*	19.09	23.19	21.36
		SD	7.79	9.23	9.52
	High	*M*	17.28	24.42	19.74
		SD	7.38	12.76	5.97

Thus, the more memory-based strategy associated with High goal-state access cost was sufficient to effect an improvement in problem solving efficiency following interruption compared to a Low access cost. In contrast, the cost of a mouse movement in the Medium access cost condition was not sufficient to improve performance compared to the Low access cost condition. Furthermore, the improvement under a High access cost on resumption *and* performance thereafter comes at no extra time cost compared to lower access cost conditions. This is especially encouraging given the extra time cost associated with such a condition in a more memory-demanding visuo-spatial copying task (e.g., Morgan et al., 2009).

Effects of Goal-State Access Cost on Planning

It is important to confirm that the improvements in performance following interruption during High goal-state access cost can be accounted for by a change of task strategy. One important measure of planning is the amount of time taken to execute the first move at the start of a ToH, which we predicted would be greatest in the High access cost condition. A significant main effect of goal-state access cost, $F(2, 51) = 4.82$, $MSE = 4.59$, $p < .05$, $f = .44$, revealed that participants in the High access cost condition indeed took more time to execute the first move ($M = 6.31$, $SD = 2.7$) than those in Medium and Low access cost conditions, $ps < .05$ ($M = 4.53$, $SD = 1.45$ and $M = 4.27$, $SD = 2.08$ respectively). Time spent planning in the Medium and Low access cost conditions did not differ statistically ($p > .05$).

General Discussion

The current experiment demonstrates the efficacy of imposing higher costs on accessing the goal-state in problem solving to promote a more efficient memory-based strategy that protects against the negative effects of different types of interruption. High goal-state access cost was superior to lower access cost conditions on nearly all measures following any type of interrupting task. The

benefit of a High access cost to induce a more memory-based strategy to protect against forgetting following interruption has been demonstrated with a simple BWT copying task (Morgan et al., 2009), but this paper highlights for the first time how these effects extend to a problem solving task. This is especially interesting given the different nature of the BWT and a ToH problem solving task. The BWT has a relatively flat and repetitive goal structure (e.g., locate a block in one position and move it to another position), whereas the ToH has a hierarchical goal structure and can be performed using a variety of different strategies that usually become more sophisticated with practice (e.g., Anzai & Simon, 1979).

The findings provide further support for the soft constraints hypothesis (e.g., Gray et al., 2006), particularly its claim that strategy selection is dependent upon the time costs imposed by interacting with the external task environment. We have shown that when the ToH goal-state is masked and cannot be viewed without suffering a mouse movement and a brief time cost, problem solving strategy becomes more internalised and less display-based. Whilst this strategy selection is based upon a very subtle change to the task environment, it is powerful enough to protect against forgetting following interruption and improves problem solving efficiency.

The results can also be interpreted within the theoretical framework of the memory for goals model (Altmann & Trafton, 2002). This is a model of goal suspension and resumption that posits: for a goal to govern behaviour it has to be repeatedly *strengthened* so that its activation level within internal memory exceeds that of an interference threshold set by all other goals in memory. A goal must be *primed*, that is, associatively linked to a reminder cue (either externally or internally based) that must be available both immediately prior to and following interruption, otherwise the goal will decay and become forgotten. Upon encountering this cue again, the decaying representation of the suspended goal will be reactivated such that it again governs behavior. Our data suggest that suspended goals may have undergone a greater amount of strengthening in the High access cost condition and we may speculate that the current-state disc configuration might have provided adequate priming cues. However, the memory for goals model suggests that an interruption lag is a critical period to strengthen and prime a to-be-suspended goal so that it can be retrieved from memory after interruption. The current findings, together with those reported in Morgan et al. (2009), suggest that an interruption lag may not be critical if the task performer is equipped with a more memory-based strategy *throughout* performance of the primary task. High goal-state access cost is a subtle yet powerful method to induce such a strategy and thus may either be an alternative method to an interruption lag or a complementary method to support its proposed benefit.

The findings are also important from a practical perspective, especially regarding principles relating to cognitive engineering and display design. These principles stress the importance of making available as much information as possible to perform a task (see Wickens & Hollands, 2000), at least within the realm of human capabilities (e.g., Rasmussen & Vicente, 1989). For example, ecological interface design recommends that complex relationships between variables should be made immediately available and information within the interface should be easily extractable (Vicente & Rasmussen, 1992). Similarly, information fusion involves synthesising information from a wide range of sources and displaying it to the user in an immediately available format (e.g., Dasarathy, 2001). Whilst adopting these principles has benefit in many situations, they also risk a passive and more display-based approach to monitoring and processing information that may ultimately result in the user moving 'out-of-the-loop' (e.g., Bainbridge, 1987). A recent study by Waldron et al. (2008) using a flight simulation found that making positional information temporarily rather than permanently available improved memory for aircraft location. Given the additional findings of the current experiment and related studies (e.g., Morgan et al., 2009; Waldron et al., 2007) we perhaps radically suggest that paradoxically making information harder to access may sometimes improve performance, such as when resuming some tasks following interruption. This will depend on what is the criterion measure of performance and the advantage will be greatest when recall is important to post-interruption performance. There are of course exceptions to this suggestion. For example, it is unlikely that the benefits of increased goal-state access cost would outweigh the costs of having to continually access information in fast-paced, safety-critical task environments such as an aircraft flight-deck.

Future experiments will be necessary to fully test the boundary conditions associated with the benefit of increased goal-state access cost in problem solving and other task environments, both with and without interruptions. Furthermore, it is practically important to compare the advantages and disadvantages of using access costs with other methods for mitigating interruption effects.

References

Altmann, E. M., & Trafton, G. J. (2002). Memory for goals: An activation-based model. *Cognitive Science, 26,* 39-83.

Anderson, J. R., & Douglass, S. (2001). Tower of Hanoi: Evidence for the cost of goal retrieval. *Journal of Experimental Psychology: Learning, Memory and Cognition, 27,* 1331-1346.

Anzai, Y., & Simon, H. A. (1979). The theory of learning by doing. *Psychological Review, 86,* 124-140.

Bainbridge, L. (1987). Ironies of automation. In J. Rasmussen, K. Duncan, & J. Leplat (Eds.), *New technology and human error* (pp. 271-283). New York: Wiley.

Ballard, D. H., Hayhoe, M. M., & Pelz, J. B. (1995). Memory representations in natural tasks. *Journal of Cognitive Neuroscience, 7(1),* 66-80.

Cary, M., & Carlson, R. A. (2001). Distributing working memory resources during problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 836-848.

Dasarathy, B. V. (2001). Information Fusion – what, where, why, when, and how? *Information Fusion, 2,* 75-76.

Davies, S. P. (2003). Initial and concurrent planning in solutions to well-structured problems. *The Quarterly Journal of Experimental Psychology, 56A(7),* 1147-1164.

Davies, S. P. (2005). Planning and problem solving in well-defined domains. In R. Morris & G. Ward (Eds.), *The cognitive psychology of planning* (pp. 35-51). Hove: Psychology Press.

Edwards, M. B., & Gronlund, S. D. (1998). Task interruption and its effects on memory. *Memory, 6,* 665-687.

Fu, W.-T., & Gray, W. D. (2000). Memory versus perceptual-motor tradeoffs in a blocks world task. *Proceedings of the 22nd annual conference of the Cognitive Science Society* (pp. 154-159). Hillsdale, NJ: Erlbaum.

Gillie, T., & Broadbent, D. (1989). What makes interruptions disruptive? A study of length, similarity, and complexity. *Psychological Research, 50,* 243-250.

Gray, W. D., & Boehm-Davis, D. A. (2000). Milliseconds matter: An introduction to microstrategies and their use in describing and predicting interactive behavior. *Journal of Experimental Psychology: Applied, 6,* 322-335.

Gray, W. D., & Fu, W.-T. (2004). Soft constraints in interactive behaviour: The case of ignoring perfect knowledge in-the-world for imperfect knowledge in-the-head. *Cognitive Science, 28,* 359-383.

Gray, W. D., Sims, C. R., Fu, W.-T., & Schoelles, M. J. (2006). The soft constraints hypothesis: A rational analysis approach to resource allocation for interactive behaviour. *Psychological Review, 113,* 461-482.

Hodgetts, H. M., & Jones, D. M. (2006). Contextual cues aid recovery from interruption: The role of associative activation. *Journal of Experimental Psychology: Learning, Memory and Cognition, 32(5),* 1120-1132.

Latorella, K. A. (1996). Investigating interruptions: Implications for flightdeck performance. *Unpublished doctoral dissertation.* State University of New York at Buffalo.

McDaniel, M. A., Einstein, G. O., Graham, T., & Rall, E. (2004). Delaying execution of intentions: Overcoming the cost of interruptions. *Applied Cognitive Psychology, 18(5),* 533-547.

Monk, C. A., Trafton, J. G., & Boehm-Davis, D. A. (2008). The effect of interruption duration and demand on resuming suspended goals. *Journal of Experimental Psychology: Applied, 14,* 299-313.

Morgan, P. L., Patrick, J., Waldron, S. M., King, S. L., & Patrick, T. (2009). Improving memory after interruption: Exploiting soft constraints and manipulating information access cost. *Journal of Experimental Psychology: Applied, 15(4),* 291-306.

O'Hara, K. P., & Payne, S. J. (1998). The effects of operator implementation cost on planfulness of problem solving and learning. *Cognitive Psychology, 35,* 34-70.

Pfeiffer, T. (2004). Problem solving with a simple transformation problem with and without continuous external support. *European Journal of Cognitive Psychology, 16,* 555-572.

Rasmussen, J., & Vicente, K. J. (1989). Coping with human errors through system design: Implications for ecological interface design. *International Journal of Man-Machine Studies, 31,* 517-534.

Trafton, J. G., & Monk, C. M. (2008). Task interruptions. In D. A. Boehm-Davis (Ed.), *Reviews of Human Factors and Ergonomics, Vol. 3.* Human Factors & Ergonomics Society.

Vicente, K. J., & Rasmussen, J. (1992). Ecological interface design: Theoretical foundations. *IEEE Transactions on systems, man, and cybernetics* (pp. 589-606). SMC-22.

Waldron, S. M., Patrick, J., & Duggan, G. B. The influence of information access cost during problem solving: consequences for memory and planning. *Manuscript submitted for publication.*

Waldron, S. M., Patrick, J., Duggan, G. B., Banbury, S., & Howes, A. (2008). Designing information fusion for the encoding of visual-spatial information. *Ergonomics, 51(6),* 775-797.

Waldron, S. M., Patrick, J., Morgan, P. L., & King, S. L. (2007). Influencing cognitive strategy by manipulating information access costs. *The Computer Journal, 50(6),* 694-702.

Ward, G., & Allport, D. A. (1997). Planning and problem solving in the five disk Tower of London. *Quarterly Journal of Experimental Psychology, 50A,* 49-78.

Wickens, C. D., & Hollands, J. G. (2000). *Engineering psychology & human performance (3rd Ed.).* New Jersey: Prentice-Hall Inc.

Acknowledgments

Phillip, L Morgan & John Patrick are at the School of Psychology, Cardiff University, UK, CF10 3AT.

This work was supported, in part, by the UK MoD Defence Technology Centre: Data and Information Fusion (MIMEX Cluster Project, DTC112) awarded to the second author. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the UK Ministry of Defence, or the UK Government. We would also like to thank Olivier DeCondappa and Sophia King for assisting with running participants and data coding.

Correspondence should be addressed to Phillip L. Morgan, School of Psychology, Cardiff University, Tower Building, Park Place, Cardiff, CF10 3AT. E-mail may be sent to MorganP4@cardiff.ac.uk.

Reference: Leijnen, S. & Gabora, L. (2010). An agent-based simulation of the effectiveness of creative leadership. *Proceedings of the Annual Meeting of the Cognitive Science Society,* August 11-14, 2010, Portland, Oregon.

An Agent-based Simulation of the Effectiveness of Creative Leadership

Stefan Leijnen (stefan.leijnen@ubc.ca)
Department of Psychology
University of British Columbia
Okanagan campus, 3333 University Way
Kelowna BC, V1V 1V7, Canada

Liane Gabora (liane.gabora@ubc.ca)
Department of Psychology
University of British Columbia
Okanagan campus, 3333 University Way
Kelowna BC, V1V 1V7, Canada

Abstract

This paper investigates the effectiveness of creative versus uncreative leadership using EVOC, an agent-based model of cultural evolution. Each iteration, each agent in the artificial society invents a new action, or imitates a neighbor's action. Only the *leader's* actions can be imitated by all other agents, referred to as *follower*s. Two measures of creativity were used: (1) *invention-to-imitation ratio*, i_{Leader}, which measures how often an agent invents, and (2) *rate of conceptual change,* c_{Leader}, which measures how creative an invention is. High i_{Leader} increased mean fitness of ideas, but only when creativity of followers was low. High i_{Leader} was associated with greater diversity of ideas in the early stage of idea generation only. High c_{Leader} increased mean fitness of ideas in the early stage of idea generation; in the later stage it decreased idea fitness. Reasons for these findings and tentative implications for creative leadership in human society are discussed.

Keywords: agent based modeling; broadcasting; creativity; culture; cultural evolution; imitation; leadership.

Introduction

It is widely assumed that effective leaders are creative (Basadur, 2004; Bellows, 1959; Puccio, Murdock, & Mance, 2006; Simon, 1988; Sternberg, Kaufman & Pretz, 2003). Creativity, however, has drawbacks (Cropley, Cropley, Kaufman, & Runco, 2010). For example, a creative solution to one problem may generate other problems, and similarly, a creative solution to one element of a situation may have unexpected negative consequences with respect to other elements. Moreover, time spent creatively finding a solution for oneself is time not spent imitating and passing on solutions already found by others.

Previous investigations of the pros and cons of creativity using an agent-based simulation approach addressed the question: in an ideal society, what proportion of individuals should be 'creative types' (Leijnen & Gabora, 2009; Gabora, Leijnen & Ghyczy, in press)? The rationale was that in a group of interacting individuals only a fraction of them need be creative for the benefits of creativity to be felt throughout the group. The rest can reap the benefits of the creator's ideas by simply copying, using, or admiring them. After all, few of us know how to build a computer, or write a symphony or novel, but they are nonetheless ours to use and enjoy. Numerical simulations showed that if the proportion of creators is low, the mean fitness of ideas in the artificial society is highest when creators dedicate themselves fully to invention. However, as the proportion of creators increases, for optimal results, creators should spend more time imitating. Creative agents amounted to 'puncture points' in the fabric of society that interfered with the dissemination of proven effective ideas.

In the current investigation we focused exclusively on the extent to which creativity is desirable in a leader, where leadership is equated with having substantial influence over others. Previous results indicated that the presence of a leader accelerates convergence on optimal ideas, but does so at the cost of consistently reducing the diversity of ideas (Gabora, 2008b,c). In these previous simulations, the leader was no more nor less creative than the rest of the agents, referred to here as *followers*. The goal of the work reported here was to investigate how creative versus uncreative leadership affects the group as a whole.

The Modeling Platform

Our investigation was carried out using an agent-based simulation referred to as 'EVOlution of Culture', abbreviated EVOC (Gabora, 2008b, 2008c). EVOC is an elaboration of Meme and Variations, or MAV (Gabora, 1994, 1995), the earliest computer program to model culture as an evolutionary process in its own right (as opposed to modeling the interplay of cultural and biological evolution). The approach was inspired by Holland's (1975) genetic algorithm, or GA. The GA is a search technique that finds solutions to complex problems by generating a 'population' of candidate solutions through processes akin to mutation and recombination, selecting the best, and repeating until a satisfactory solution is found. The goal here was to distil the underlying logic of not biological evolution but cultural evolution, i.e. the process by which ideas adapt and build on one another in the minds of interacting individuals. EVOC (as did MAV) uses neural network based agents that could (1) invent new ideas by modifying previously learned ones, (2) evaluate ideas, (3) implement ideas as actions, and (4) imitate ideas implemented by neighbors. Agents do not evolve in a biological sense—they neither die nor have offspring—but do in a cultural sense, by generating and sharing ideas for actions. EVOC (like MAV) successfully models how 'descent with modification' occurs in a cultural context. The approach can thus be contrasted with computer

models of how individual learning affects biological evolution (Best, 1999, 2006; Higgs, 2000; Hinton & Nowlan, 1987; Hutchins & Hazelhurst, 1991).

EVOC consists of an artificial society of neural network based agents in a two-dimensional grid-cell world. It is written in Joone, an object oriented programming environment, using an open source neural network library written in Java. This section summarizes the key components of the agents and the world they inhabit.

The Agent

Agents consist of (1) a neural network, which encodes ideas for actions and detects trends in what constitutes a fit action, and (2) a body, which implements actions.

The Neural Network. The core of an agent is an autoassociative neural network, as shown in Figure 1. It is composed of six input nodes that represent concepts of body parts (LEFT ARM, RIGHT ARM, LEFT LEG, RIGHT LEG, HEAD, and HIPS), six matching output nodes, and six hidden nodes that represent more abstract concepts (LEFT, RIGHT, FORELIMB, HINDLIMB, SYMMETRY and MOVEMENT). Input nodes and output nodes are connected to hidden nodes of which they are instances (e.g. RIGHT FORELIMB is connected to RIGHT.) Activation of any input node activates the MOVEMENT hidden node. Same-direction activation of symmetrical input nodes (e.g. positive activation—which represents upward motion—of both forelimbs) activates the SYMMETRY node.

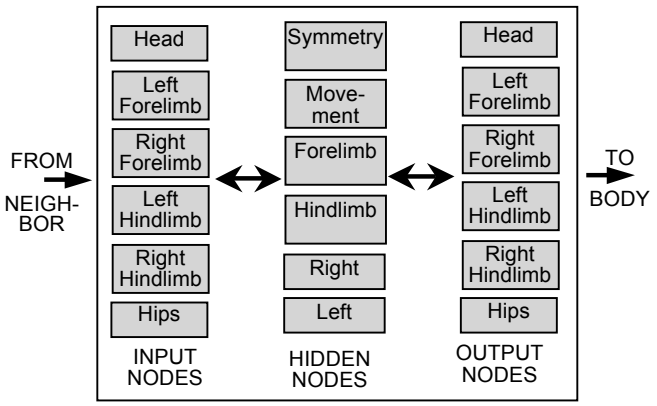


Figure 1. The neural network. See text for details.

The neural network learns ideas for actions. An idea is a pattern consisting of six elements that dictate the placement of the six body parts. Learning and training of the neural network is as per Gabora (1995). During imitation, the input is the action implemented by a neighbor. During invention, the pattern of activation on the output nodes is fed back to the input nodes, and change is biased according to the activations of the SYMMETRY and MOVEMENT nodes. In EVOC, the neural network can also be turned off to compare results with a data structure that cannot detect trends, and thus invents ideas merely at random.

The Body. If the fitness of an action is evaluated to be higher than that of any action learned thus far, it is copied from the input/output nodes of the neural network that represent *concepts of* body parts to a six digit array that contains representations of *actual* body parts, referred to as the *body*. Since it is useful to know how many agents are doing essentially the same thing, when node activations are translated into limb movement they are thresholded such that there are only three possibilities for each limb: stationary, up, or down. Six limbs with three possible positions each gives a total of 729 possible actions. Only the action that is currently implemented by an agent's body can be observed and imitated by other agents.

The Fitness Function

Agents evaluate the effectiveness of their actions according to how well they satisfy needs using a pre-defined equation referred to as a *fitness function*. The fitness of an action with respect to the need to attract mates is calculated as in (Gabora, 1995). The fitness function rewards actions that make use of trends detected by the symmetry and movement hidden nodes and used by knowledge-based operators to bias the generation of new ideas. It generates actions that are relatively realistic mating displays, and exhibits a cultural analog of *epistasis*. In biological epistasis, the fitness conferred by the allele at one gene depends on which allele is present at another gene. In this cognitive context, epistasis is present when the fitness contributed by movement of one limb depends on what other limbs are doing.

The World

MAV allowed only worlds that were square and toroidal, or 'wrap-around' (such that agents at the left border that attempt to move further left appear on the right border). Moreover, the world was always maximally densely populated, with one agent per cell. In EVOC the world can assume any shape, and be as sparsely or densely populated as required, with agents placed in any configuration. EVOC also allows for the creation of complete or semi-permeable permanent or eroding borders that decrease the probability of imitation along a frontier (although this was not used in the experiments reported here).

Incorporation of Cultural Phenomena

Agents incorporate the following phenomena characteristic of cultural evolution as parameters that can be turned off or on (in some cases to varying degrees):

- **Imitation**. Ideas for how to perform actions spread when agents copy neighbors' actions. This enables them to share effective, or 'fit', actions.
- **Invention**. This code enables agents to generate new actions by modifying their initial action or a previously invented or imitated action. (See Gabora 1995 for further details.)

- **Knowledge-based Operators**. Since a new action (or, in invention, new idea for an action) is not learned unless it is fitter than the currently implemented action, new actions provide valuable information about what constitutes an effective idea. This information is used by knowledge-based operators to probabilistically bias invention such that new ideas are generated strategically as opposed to randomly. For example, if successful actions tend to be symmetrical (e.g. left arm moves to the right and right arm moves to the left), the probability increases that new actions are symmetrical. Also, if movement is generally beneficial, the probability increases that new actions involve movement of more body parts. (See Gabora 1995 for further details.)
- **Mental simulation**. Before committing to implementing an idea as an action, agents can use the fitness function to assess how fit the action would be if it *were* implemented.
- **Broadcasting**. Broadcasting allows the action of a *leader,* or broadcaster, to be visible to not just immediate neighbors, but all agents, thereby simulating the effects of media such as public performances, television, radio, or the internet, on patterns of cultural change. Each agent adds the leader as a possible source of actions it can imitate. The leader itself is thus the only agent that can only acquire actions from its immediate neighbors. The leader can be specified by the user or chosen at random. Broadcasting can be intermittent, or continued throughout the duration of a run. It can also be turned off altogether.

A Typical Run

Each iteration, every agent has the opportunity to (1) acquire an idea for a new action, either by *imitation*, copying a neighbor, or by *invention*, creating one anew, (2) update the knowledge-based operators, and (3) implement a new action. To invent a new idea, for each node of the idea currently represented on the input/output layer of the neural network, the agent makes a probabilistic decision as to whether change will take place, and if it does, the direction of change is stochastically biased by the knowledge-based operators. If the new idea has a higher fitness than the currently implemented idea, the agent learns and implements the action specified by that idea. To acquire an idea through imitation, an agent randomly chooses one of its neighbors, and evaluates the fitness of the action the neighbor is implementing. If its own action is fitter than that of the neighbor, it chooses another neighbor, until it has either observed all of its immediate neighbors, or found one with a fitter action. If no fitter action is found, the agent does nothing. Otherwise, the neighbor's action is copied to the input layer, learned, and implemented.

Fitness of actions starts out low because initially all agents are immobile. Soon some agent invents an action that has a higher fitness than doing nothing, and this action gets imitated, so fitness increases. Fitness increases further as other ideas get invented, assessed, implemented as actions, and spread through imitation. The diversity of actions initially increases due to the proliferation of new ideas, and then decreases as agents hone in on the fittest actions.

The Graphic User Interface

The graphic user interface (GUI) makes use of the open-source charting project, JFreeChart, enabling variables to be user defined at run time, and results to become visible as the computer program runs. The topmost output panel is shown in Figure 2.

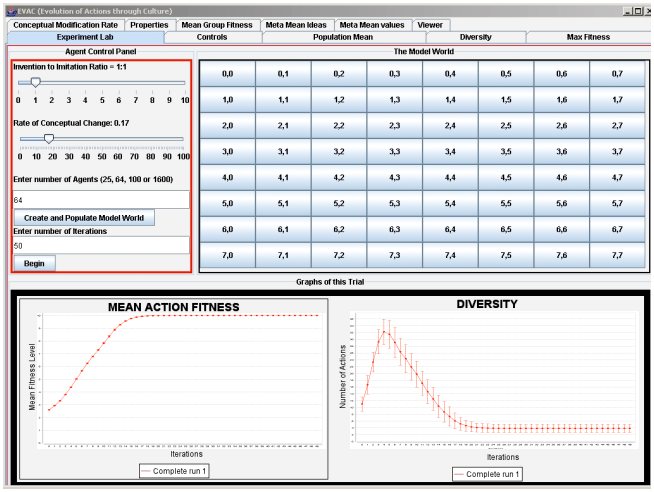


Figure 2. Output panel of GUI. See text for details.

At the upper left one specifies the *Invention to Imitation Ratio*. This is the probability that a given agent, on a given iteration, invents a new idea for an action, versus the probability that it imitates a neighbor's action. Below it is *Rate of Conceptual Change*, where one specifies the degree to which a newly invented idea differs from the one it was based on. Below that is *Number of Agents*, which allows the user to specify the size of the artificial society. Below that is where one specifies *Number of Iterations*, i.e. the duration of a run. Agents can be accessed individually by clicking the appropriate cell in the grid on the upper right. This enables one to see such details as the action currently implemented by that agent, or the fitness of that action. The graphs at the bottom plot the mean idea fitness and diversity of ideas. Tabs shown at the top give access to other output panels.

Experiments

We now present the creative leadership experiments performed with EVOC. Unless stated otherwise, graphs plot the average of 100 runs, the world consists of 100 cells, one agent per cell, a 1:1 invention-to-imitation ratio, and a 1/6 probability of change to any body part during invention (The rationale behind this is that since, with six body parts, on average each newly invented action differs from the one it was based on with respect to one body part.)

The current experiments made use of EVOC's broadcasting function. As described above, broadcasting enables the action implemented by a leader to be visible throughout the artificial society. While experiments reported elsewhere investigated the impact of varying the number of leaders on the fitness and diversity of ideas (Gabora, 2008c), in the experiments reported here, simulated societies consist of one leader and ninety-nine followers. The leader is chosen randomly and broadcasts throughout the entire run.

Experiment 1a: Effect of Varying Inventiveness (*i*) of Leaders and Followers on Fitness of Ideas

The first experiment investigated the effect of varying the ratio of iterations spent inventing versus imitating, or invention-to-imitation ratio, abbreviated *i*, of both the leader and the followers, on the fitness of ideas produced by the artificial society. The inventiveness of the leader, abbreviated i_{Leader} was systematically varied from 0.0 to 1.0. When i_{Leader} was 1.0, the leader invented a new action every iteration. When i_{Leader} was 0.0, the leader never invented new actions; it only imitated its neighbors' actions. (It was still the leader because its actions were visible to, and could be imitated by, all other agents in the society, not just its immediate neighbors, as was the case for followers).

In the first set of runs, followers only imitated; they never invented, i.e., $i_{Followers}$ = 0.0. As shown in Figure 3, with uncreative followers, the degree of creativity of the leader matters a lot; the mean fitness of ideas in the artificial society is positively correlated with creativity of the leader.

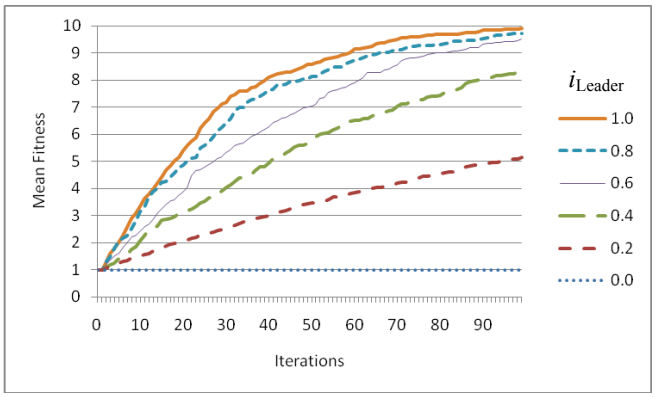


Figure 3. Mean fitness of actions with leaders of varying invention-to-imitation ratios, and followers that only imitate, i.e. that never invent (e.g. *i* = 0.0).

In the second set of runs, shown in Figure 4, followers were able to invent. More specifically, $i_{Followers}$ = 0.05; thus each iteration, each of the 99 followers had a 5% chance of inventing. Comparing figures 3 and 4 it is clear that while the degree of creativity of the leader had a large impact when followers are uncreative, it had almost no impact when followers were themselves creative. With creative followers, the mean fitness of ideas generated by the society increased over the duration of a run at more or less the same pace no matter how creative the leader was.

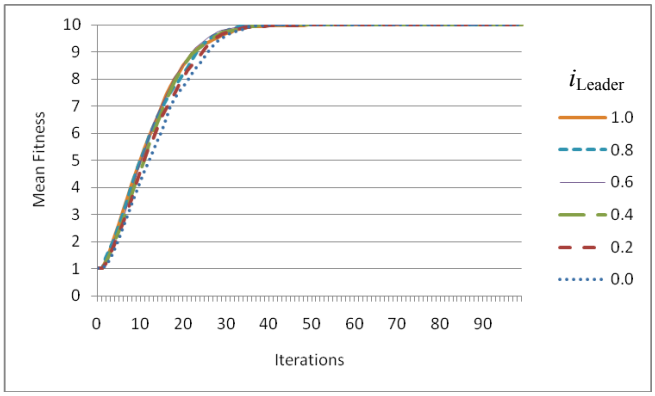


Figure 4. Mean fitness of actions with leaders of varying invention-to-imitation ratios, and followers that invent as well as imitate (*i* = 0.05).

Experiment 1b: Effect of Varying Inventiveness (*I*) of Leaders and Followers on Diversity of Ideas

The second part of this experiment involved investigating the effect of varying the invention-to-imitation ratio, *i*, of both the leader and the followers on the *diversity* of ideas produced by the artificial society. As in experiment 1a, i_{Leader} was systematically varied from 0.0 to 1.0. The result obtained with $i_{Followers}$ = 0.0 is shown in Figure 5.

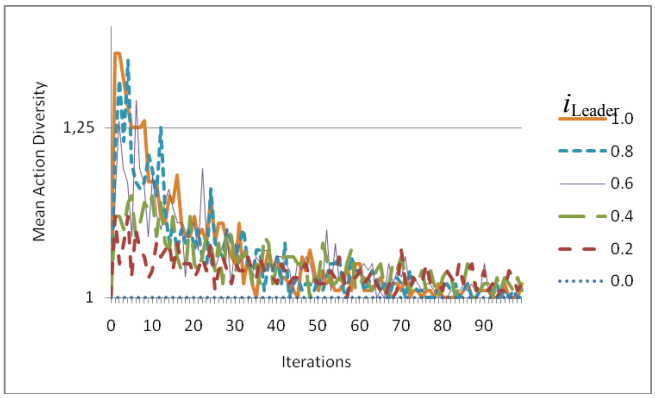


Figure 5. Diversity of actions in the artificial society with leaders of varying invention-to-imitation ratios, and followers that only imitate (*i* = 0.0).

In the short run, creative leadership was associated with increased diversity of actions. However in the long run, no matter how creative the leader, all agents converged on the same action, despite that there were seven other equally optimal actions they could have converged upon. Results with higher values of $i_{Followers}$ (not shown) were qualitatively similar. Action diversity was initially substantially higher, but it still always eventually converged to 1.

Experiment 2: Effect of Varying Leaders' Rate of Conceptual Change (*c*)

There are two ways an agent's creativity can be manipulated in EVOC. The first way involves changing *i,* the invention-

to-imitation ratio, as in the first set of experiments. It is possible to vary not just how frequently an agent invents, but how creative its newly invented ideas are. This second measure, referred to as the *rate of conceptual change,* abbreviated *c, is* implemented as follows. Invention occurs by taking the current action, and modifying it. When *c* is low, the newly invented action varies little from the previous action upon which it was based. When *c* is high, the newly invented idea varies dramatically from the previous idea upon which it was based.

As mentioned previously, the default value of *c*, the probability of change to any body part during invention, is 1/6 for any agent that invents, whether it is a leader or a follower. Previous experiments revealed this to be the rate that optimizes the rate of increase in mean fitness of actions (Gabora, 1995). Since ideas are ideas for actions, and since actions involve at most six body parts, on average, each newly invented action involves a change to the motion of one body part. Thus c = 1/6 means that each body part changes what it is doing with a 1/6 probability, or 17% of the time. In this second set of experiments, shown in Figure 6, c_{Leader} was systematically varied from 0% to 100%. Since the followers only imitated, $c_{Followers}$ = 0. Because that means there are no new actions for the leader to imitate, i_{Leader} was set to 1.0.

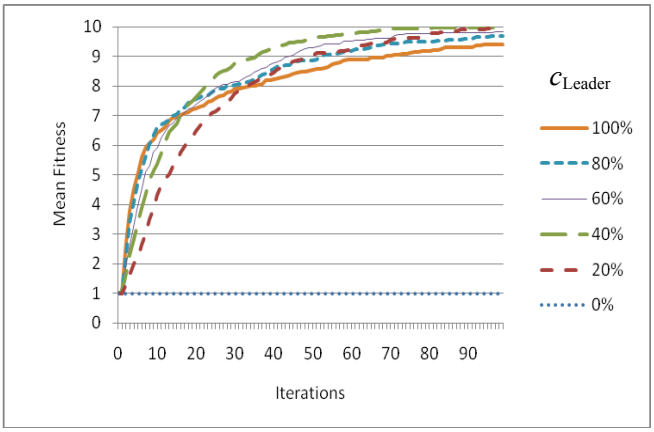


Figure 6. Mean fitness of actions in the artificial society with leaders of varying rates of conceptual change, and followers that only imitate.

Unlike in experiment one, the optimal degree of creative leadership with respect to this second measure of creativity depended on what phase of the creative process the society was at. Early on in a run, a form of leadership that entails the highest possible rate of conceptual change (100%) was most beneficial. However, as the run progressed a transition occurred, after which point a much lower rate of conceptual change (approximately 40%) was most beneficial.

Discussion

The experiments reported here investigated the impact of creative versus uncreative leadership on the mean fitness and diversity of ideas for actions in an agent-based artificial society. The first experiment looked at the effect of varying the invention-to-imitation ratio of both leader and followers. The mean fitness of actions was positively correlated with the creativity of the leader, but only when the followers were uncreative. The more creative the followers, the greater the extent to which the beneficial effect of creative leadership was washed out. One must be cautious about extrapolating from a simple simulation such as this to the real world. For example, real-world creativity is correlated with emotional instability, affective disorders, and substance abuse (Andreason, 1987; Flaherty, 2005; Jamieson, 1993) which presumably would interfere with effective leadership, and which were not incorporated in these simulations. However, the result suggests that creativity may be a relatively unimportant quality for a manager of a creative team, but an important quality for a manager of an uncreative team.

The first experiment also investigated the effect of varying the invention-to-imitation ratio of both leader and followers on the diversity or number of different of actions implemented by agents. Previous results with EVOC had suggested that the beneficial effect of leadership on mean fitness of ideas is tempered by decreased diversity of ideas, and this echoed previous simulation findings that leadership can have adverse effects when agents can communicate (Gigliotta, Miglino, & Parisi, 2007). We wanted to know whether the decreased diversity associated with the presence of a leader was still observed when leaders are highly creative or highly uncreative compared to followers. We found that while in the early stages of a run, creative leadership (as well as the degree of creativity of followers) was associated with higher diversity, eventually all agents converged on what the leader was doing no matter how creative the leader (or how creative the followers). This suggests that in the long run leadership diminishes cultural diversity regardless of how creative the leader is. It is worth noting, however, that in this artificial world, unlike the real world, agents had only one task to accomplish. Further experiments will investigate whether these results hold true when the fitness function varies over time.

The second set of experiments investigated the effect of not how often the leader invents, but how creative any particular invention is, referred to as the rate of conceptual change. We found that early on in the creative process, when the fitness of the ideas that are getting generated was still relatively low, it was best if the leader was very creative (high rate of conceptual change). However, later in the creative process, once relatively fit ideas were being generated, a less creative leader was better (low rate of conceptual change). This result may reflect that the fitness function used here exhibits the cultural equivalent of the biological phenomenon of epistasis, wherein what is optimal for one element of an idea depends on what is going on with respect to another element. Initially, the higher the rate of conceptual change, the more quickly fitter actions are found. However, once relatively fit actions have been found, a high rate of conceptual change breaks up co-adapted epistatically

linked elements and thus interferes with convergence toward optimal actions. In future experiments we will investigate whether these findings hold true when a different fitness function is used. However we believe that many real-world problem solving situations involve this kind of epistasis. Thus, although once again one must be cautious about extrapolating from the results of simple simulations such as this to the real world, our results suggest that a new startup company benefits most from highly creative leadership, while a more established company, or one that has stabilized on an established product line, benefits most from a more conservative form of leadership.

Acknowledgments

This work is funded by grants to the second author from the Social Sciences and Humanities Research Council of Canada (SSHRC) and the GOA program of the Free University of Brussels.

References

Andreason, N. C. (1987). Creativity and mental illness; prevalence rates in writers and their first degree relatives. *American Journal of Psychiatry, 144,* 1288-1292.

Basadur, M. (2004). Leading others to think innovatively together: Creative leadership. *The Leadership Quarterly, 15*(1), 103–121.

Bellows, R. M. (1959). *Creative leadership*. Upper Saddle River, NJ: Prentice-Hall.

Best, M. (1999). How culture can guide evolution: An inquiry into gene/meme enhancement and opposition. *Adaptive Behavior, 7*(3), 289–293.

Best, M. (2006). Adaptive value within natural language discourse. *Interaction Studies, 7*(1), 1-15.

Boone, J. L., Smith, E. A. 1998. Is it evolution yet? A critique of evolutionary archaeology. *Current Anthropology, 39,* S141–73.

Cropley, D. Cropley, A., Kaufman, J. & Runco, M. (2010). *The dark side of creativity*. Cambridge UK: Cambridge University Press.

Fracchia, J. & Lewontin, R. C. (1999). Does culture evolve? *History and Theory, 38*(4), 52–78.

Gabora, L. (1994). A computer model of the evolution of culture. In R. Brooks & P. Maes (Eds.) *Proceedings of the 4th International Conference on Artificial Life, July 4-6,* Boston, MA.

Gabora, L. (1995). Meme and Variations: A computer model of cultural evolution. In L. Nadel & D. Stein (Eds.) *1993 Lectures in Complex Systems*, pp. 471–486. Reading, MA: Addison-Wesley.

Gabora, L. (2004). Ideas are not replicators but minds are. *Biology & Philosophy, 19*(1), 127–143.

Gabora, L. (2008a). The cultural evolution of socially situated cognition. *Cognitive Systems Research, 9*(1–2), 104–113.

Gabora, L. (2008b). EVOC: A computer model of cultural evolution. In V. Sloutsky, B. Love & K. McRae (Eds.), *Proceedings of the 30th Annual Meeting of the Cognitive Science Society*, Sheridan Publishing.

Gabora, L. (2008c). Modeling cultural dynamics. In *Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI) Fall Symposium 1: Adaptive Agents in a Cultural Context*, Menlo Park, CA: AAAI Press, 18-25.

Gabora, L., Leijnen, S., & von Ghyczy, T. (in press). Imitation: more than just the greatest compliment. *International Journal of Software and Informatics*.

Gigliotta, O. Miglino, O. & Parisi, D. (2007). Groups of agents with a leader. *Journal of Artificial Societies and Social Simulation, 10*(4). http://jasss.soc.surrey.ac.uk/10/4/1.html

Higgs, P. G. (2000). The mimetic transition: a simulation study of the evolution of learning by imitation. *Proceedings of the Royal Society B: Biological Sciences, 267*(1450), 1355–1361.

Hinton, G. E. & Nowlan, S. J. (1987). How learning can guide evolution. *Complex Systems, 1,* 495–502.

Holland, J. K. (1975). Adaptation in Natural and Artificial Systems. Ann Arbor, MI: University of Michigan Press.

Hutchins, E. & Hazelhurst, B. (1991). Learning in the cultural crocess. In Langton, C., Taylor, J., Farmer, D., & Rasmussen, S. (Eds.) *Artificial Life II*. Redwood City, CA: Addison-Wesley.

Jamieson, K. R. (1993). *Touched by fire: Manic-depressive illness and the artistic temperament*. New York: Free Press.

Jeffreys, M. (2000). The meme metaphor. Perspectives in *Biology and Medicine , 43*(2), 227–242.

Kaufman, J. C., & Sternberg, R. J. (Eds). (2006). *The international handbook of creativity*. Cambridge UK: Cambridge University Press.

Krasnogor, N. & Gustafson, S. (2004). A study on the use of "self-generation" in memetic algorithms. *Natural Computing, 3*(1), 53–76.

Leijnen, S., & Gabora, L. (2009). How creative should creators be to optimize the evolution of ideas? A computational model. *Electronic Proceedings in Theoretical Computer Science, 9,* 108-119.

Puccio, G. J., Murdock, M. & Mance, M. (2006). *Creative leadership: skills that drive change*. Thousand Oaks, CA: Sage.

Simon, H. A. (1986). What we know about the creative process. In R. Kuhn (Ed.) *Frontiers in creativity and innovative management*. Cambridge, MA: Ballinger.

Sternberg, R. J., Kaufman, J. C. & Pretz, J. E. (2003). A propulsion model of creative leadership. *The Leadership Quarterly, 14*(4-5), 455–473.

Wright, S. (1969). *Evolution and the genetics of Populations.* Chicago, IL: University of Chicago Press.

Linguistic Cues Predict Fraudulent Events in a Corporate Social Network

Max Louwerse (mlouwerse@memphis.edu)
Department of Psychology / Institute for Intelligent Systems
University of Memphis
Memphis, TN 38152 USA

King-Ip Lin (davidlin@memphis.edu)
Department of Computer Science / Institute for Intelligent Systems
University of Memphis
Memphis, TN 38152 USA

Amanda Drescher (adreschr@memphis.edu)
Department of Psychology / Institute for Intelligent Systems
University of Memphis
Memphis, TN 38152 USA

Gün Semin (G.R.Semin@uu.nl)
Faculty of Social Sciences,
Utrecht University
3584 CS Utrecht, The Netherlands

Abstract

There is an increase in deception studies investigating which non-linguistic and linguistic cues best predict deception. Even though these studies have shown participants consistently use specific cues to deception when they are asked to deceive somebody in a particular situation, it is less clear how these findings translate to non-experimental settings, for instance, do these cues also apply in cases of global deception in social networks. This paper investigated whether fraudulent events can be related to linguistic cues of deception within records of a large corporate social network. Specifically, we investigated the Enron email dataset using a model of interpersonal language use. Results suggest that during times of fraud, emails were composed with higher degrees of abstractness.

Keywords: deception, social cognition, computer mediated communication, corpus linguistics.

Introduction

Humans lie because it helps them manipulate the impressions people have of them. Apologizing for being late (even though you could have been on time), telling a police officer you really thought the speed limit was 40 (even though you knew it was 35), and thanking the waitress for guiding you to your table (even though you had waited for 20 minutes and she just did her job), all help to establish an interpersonal glue between you and your social environment. We tell many lies, on average one or two a day (DePaulo & Kashy, 1998).

Of course, there are gradations in the acceptability of twisting the truth. Some lies are blatant transgressions with potentially far reaching consequences, such as cases related to fraud, others are harmless and would have very little or no consequences. Most research in the cognitive sciences on deception centers on lies with little consequences. In fact, very little research has been done on cases of deception with far reaching consequences, for the liar or the recipient of the lie.

Liars leave non-linguistic and linguistic footprints in their attempts to hide the truth, both in cases of blatant and not so blatant half-truths (DePaulo, et al., 2003). Several experiments have investigated these footprints using a paradigm whereby a participant in a deception condition is asked to tell a lie and/or to tell the truth. For instance, Newman, Pennebaker, Berry and Richards (2003) conducted a study in which they asked pro- (and anti-) abortion participants to produce both pro- and anti-abortion stories. They found that deceptive communication had fewer first-person singular pronouns, fewer third-person pronouns, more negative emotion words (e.g., *hate*, *anger*, *enemy*), fewer exclusive words (e.g., *but*, *except*), and more motion verbs (e.g., *walk*, *move*, *go*). Apparently liars wanted to dissociate themselves from their words (fewer first person pronouns), and made an attempt to create a story that seemed less complex (fewer exclusive words) and more concrete (more action words).

Hancock, Curry, Goorha, & Woodworth (2008) came to a very similar conclusion. They investigated deception in asynchronous computer-mediated communication. Participants were asked to write stories on five different topics. Half of the participants were asked to not tell the truth. Hancock et al. (2008) found that lies consisted of fewer words, more questions, fewer first person pronouns and more words pertaining to senses (e.g., *see*, *listen*) than truthful discussions.

Both Newman et al. (2003) and Hancock et al. (2008) found pronoun use, lowered word quantity, emotion words and lower cognitive complexity to be linguistic cues affiliated with deception. Both the experimental design and the findings of these two studies are prototypical for much of the empirical work on deception.

DePaulo et al. (2003) conducted a meta-analysis of experimental literature that investigated cues to deception. They reviewed 116 studies that looked into deceptive cues when people told lies. Results showed, for instance, that liars raised their chins more, pressed their lips more, and had larger pupil dilations than truth tellers. Moreover, lies had more verbal and vocal uncertainty, less verbal and vocal immediacy, were more ambivalent, less plausible and had less logical structure, with less contextual embedding.

However, DePaulo, et al. (2003) warned that these (and other) deception cues were moderated by motivation and transgressions. That is, when participants were more motivated to succeed and when the lies were about transgressions, the deception cues were more pronounced. These moderators are important to note. In fact, it is worth pointing out that the deception studies DePaulo et al. reviewed typically consisted of college students (87.1%), who lied to strangers (88.80%), with lies about transgressions (85.34%).

Indeed, the cues found in the studies DePaulo et al. (2003) used in their meta-analysis are extremely helpful to gaining further insight into deception. In these cases of deception researchers can compare the repertoires of deception cues that humans can use in their lying acts. At the same time, these cues come from unidirectional individual cases in which the participant is asked to act out a lie. It might well be the case that in ecologically situated settings no cues, or different cues, may be observed.

Furthermore, lies often do not impact only the liar. Instead, important cases of lying involve more than a single individual who is aware of the lie. Such instances, where a group of people become part of a collective deception are of a more global nature affecting a social network of people, whereby the individual feelings of guilt and shame are reduced due to a diffusion of responsibility. Examples of deception within a social network include cases of false bookkeeping, mislabeling of accounts, and corruption (Clinard & Yeager, 2006).

Knowing whether (and which) cues to deception can be found in social networks might not answer the question what deception cues humans will use, but it does answer the question whether (and which) deception cues humans generally use. Moreover, such an investigation would be informative in identifying deception strategies in cases of fraud detection or counterintelligence.

This study investigated whether deception in corporate social networks could be detected using linguistic cues.

Enron Email Dataset

The ideal corpus for a study on deception in corporate social networks is the Enron email dataset (Klimt & Yang, 2004). This dataset consists of email messages from various Enron executives/employees obtained from the accounts of 150 executives.

Enron Corporation is most famous for the elaborate network of accounting fraud spread throughout the organization. The company formed in 1985 through the merger of Houston Natural Gas (HNG) and InterNorth Inc. After years of extensive reorganization and rebranding by CEO Kenneth Lay, Enron formed into one of the world's leading natural gas, electricity, and communication companies. Despite its six-year title within Fortune magazine as "America's Most Innovative Company," Enron's network of accounting fraud prompted an SEC inquiry that ultimately lead to the dissolution of the accounting firm Arthur Andersen and a declaration of bankruptcy by Enron Corporation in 2001.

The Enron email dataset is extremely useful for the purposes of this study. First, the dataset is highly diverse, consisting of over 20,000 different senders. Second, the emails cover a relatively large time span (1999-2001). Most importantly perhaps, there is detailed information available on Enron Corporation, its rise and fall and its fraudulent activities (Diesner, Frantz, & Carley, 2005).

While the advantage of this corpus lies in its ecological validity as well as its diversity in senders, receivers, and topics, the disadvantage is that it is very difficult to determine which emails are deceptive and which emails are not. That is, even though Enron as a whole has been known for its deception, that deception cannot be uniquely attributed to specific people or specific topics. As a result, the best way to identify deception is to use those time stamps during which it was clear – in hindsight – that fraudulent activities took place.

There are a number of studies that have analyzed the Enron dataset. Most of these studies looked at the dynamics of the structure and properties of the organizational communication network (Diesner, et al., 2005). Very few studies have looked at deceptive cues in this email corpus. Keila and Skillicorn (2005) is an exception. They used the four deception categories mentioned earlier (first person pronouns, exclusive words, negative emotion words, and action verbs) to categorize the corpus into emails of interest (which were labeled as unusual and deceptive if they showed evidence of the four categories). Keila and Skillicorn's analysis used singular value decomposition (SVD) as the primary analysis technique and successfully showed how emails can be clustered on the basis of the four deception categories. Importantly, Keila and Skillicorn did not test whether these linguistic cues predicted deception.

The current paper tested exactly this question: can a relation be found between linguistic cues in the Enron email data set and fraudulent events? Because we are dealing with interpersonal communication, we investigated this question using the Linguistic Category Model (LCM).

Linguistic Category Model

There is a range of algorithms we could apply to a corpus like the Enron email dataset (Jurafsky & Martin, 2008). However, because we are dealing with a large number of emails sent by different people on a variety of topics covering a time span of many months, it is desirable to use an algorithm based on a model of interpersonal communication. There are very few computational models

Table 1. Overview categories in the Linguistic Category Model (LCM).

Verbs in this category:	DAV	IAV	SAV	SV	ADJ
Refer to a particular activity.	+	-	-	-	-
Refer to a physically invariant feature of the action.	+	-	-	-	-
Refer to a general class of behaviors.	-	+	-	-	-
Have an action with a clear beginning and end.	+	+	-	-	-
Have associated semantic valence, positive or negative.	-	+	+	-	-
Refer to a single behavioral event.	+	+	+	-	-
Refer to a specific object.	+	+	+	+	-
Refer to a specific situation.	+	+	+	-	-
Refer to a specific context.	-	-	-	-	-
Require context for sentence comprehension.	+	-	-	-	-
Express the emotional consequence of an action.	-	-	+	-	-
Refer to mental and emotional states.	-	-	-	+	-
Readily take progressive forms.	+	+	+	-	-
Are freely used in imperatives.	+	+	+	-	-
Require interpretation beyond description.	-	+	+	+	+

available in the field of social cognition (Newman, et al., 2003).

One successful model of interpersonal language is the Linguistic Category Model (LCM, Semin, 2000; Semin & Fiedler, 1988, 1991). The model consists of a classification of interpersonal (transitive) verbs that are used to describe actions or psychological states and adjectives that are employed to characterize persons. This classification gives insight into the meanings of verbs and adjectives that people use when they communicate about actors and their social events. The model makes a distinction between five different categories of interpersonal terms (Semin & Fiedler, 1991):

(a) Descriptive Action Verbs (DAV) refer to single, specific action with a clear beginning and end, such as *hit*, *yell*, and *walk*.
(b) Interpretative Action Verbs (IAV) refer to different actions with a clear beginning and end, but do not share a physical invariant feature, such as *help*, *tease*, *avoid*.
(c) State Action Verbs (SAV) refer to behavioral events, but refer to the emotional consequence of an action rather than the action itself, such as *surprise*, *amaze*, *anger*.
(d) State Verbs (SV) refer to enduring cognitive or emotional states with no clear beginning or end, such as *hunger, trust, understand.*
(e) Adjectives (ADJ) refer to a characteristic or feature qualifying a person or concept, such as *distraught, optimal*.

These five categories can be seen as a continuum from concreteness (DAV) to abstractness (ADJ). The distinction between the categories is obtained on the basis of a number of conventional grammatical tests and semantic contrasts (Miller & Johnson-Laird, 1976). An overview of the five categories is presented in Table 1.

Several studies have shown that the LCM can adequately capture differences in interpersonal language use predicted by theories in social psychology (see Stapel and Semin, 2007).

Semin and Fiedler (1991) proposed an aggregate of the five categories in the form of an abstractness score. This score was formed by the following straightforward formula:

$$\text{Abstractness score} = \frac{DAV + (2 \times (IAV + SAV)) + (3 \times SV) + (4 \times ADJ)}{DAV + IAV + SAV + SV + ADJ}$$

Semin and Fiedler (1991) make the important claim that items scoring high on abstractness (i.e., through abstractness score, or a high frequency of abstract categories, such as adjectives):

1) generate much disagreement;
2) are difficult to verify; and
3) are low in informativeness of the situation.

These claims are relevant for the purposes of the current paper. We hypothesize that when fraudulent events take place it is more likely that the language used is difficult to verify, is low in informativeness of the situation, and is likely to be subject to disagreement (because it is harder to verify and is low in informativeness). In short, we predict that fraudulent events relate to higher abstractness scores in interpersonal communication.

In the computational implementation of the LCM model we identified all verbs and adjectives that matched the criteria identified by Semin and Fiedler (1988; 1991). This set of words was then sent through the CELEX database (Baayen, Piepenbrock & Van Rijn, 1993) to obtain derivations and inflections. The final LCM result was a list of 31,444 words in total, classified in five categories: DAV (17,884), IAV (9,224), SAV (1,533), SV (433), and ADJ (2,370). In addition, adjectives were broken down by the same categorical separations as the verb categories: DA-ADJ (467), IAV-ADJ (1,564), SAV-ADJ (220), SV-ADJ (119).

Table 2. Overview of Enron Corporation events used in Study 1 and 2. Superscripts mark multiple events.

Variable	Description of Variable	Month and Year
Layoffs	Employees within Enron Corp. were laid off.	12/01
CEO	Indicating involvement of the CEO within any coded event.	3/00, 8/00, 11/00, 1/01-4/01, 8/01[6], 10/01[3], 11/01
Fraudulent Paperwork Filed Signed	Filing and/or signing of fraudulent paperwork (by the CEO or COO.)	3/00[2], 8/00
Fraudulent Comments	Enron made fraudulent comments, to the employees and/or investors.	1/01[2], 9/01[2]
Discussion of Ethics	A discussion of ethics occurred between Enron executives or between the CEO and employees	7/00, 3/01, 5/01, 8/01[2], 9/01, 10/01
Selling Enron Shares	Selling of Enron stock by high-level executives occurs.	11/00, 5/01, 6/01, 7/01[2], 8/01[2], 9/01[2]
Rolling Blackouts Initiated	Intentional initiation of rolling blackouts in California.	1/01
Meetings with Nat'l Political Figures	High-level Enron executives met with national political figures incl. the Secr. of the Treasury and the Secr. of Commerce	2/01, 3/01, 4/01, 8/01, 10/01[4], 11/01
Financial Support of Political Candidate	High-level Enron executives (CEO & President) provided financial support for a newly elected national political figure.	1/01
Profit Announced	Profits were announced for the quarter.	4/01
Loss announced	Losses were announced for the quarter.	10/01
SEC Inquiry Developments	Beginning of the SEC inquiry and the point at which the SEC inquiry became a formal investigation.	10/01[2]
Shredding Occurs	Shredding of Enron documents in Enron and/or Arthur Andersen accounting firm.	10/01[2]
Shredding Stopped	Shredding of Enron documents stopped in Enron and/or Arthur Andersen.	10/01, 11/01
Fraud Announced	Enron admitted to having overstated the company's profits	11/01
Bankruptcy Filed	Bankruptcy was filed.	12/01

The content of each of the 255,637 messages was extracted, and the frequency of words in each of the five LCM categories was determined. These frequencies were normalized to account for the number of words in an email. Sixteen events related to the rise and fall of Enron Corporation, and occurring during the time of the emails, were identified. These events are given in Table 2. Note that some events are directly related to fraudulent activities (e.g., Fraudulent paperwork filed signed; Fraudulent comments; Shredding occurs) and others indirectly (Selling Enron shares; Rolling blackouts initiated; Financial support of political candidate). These events were dummy-coded using a 1 for the presence and a 0 for the absence of an event in the month and year (Cohen, Cohen, West, & Aiken, 2003). This resulted in a database of the sender, the normalized frequency of the LCM categories in each email, and the events linked to the time the email was received.

A mixed-effect regression model analysis was conducted on the normalized frequency of LCM categories, with events as fixed factors, and email sender and email date (year and quarter) as random factors (Louwerse & Jeuniaux, 2010). The model was fitted using the restricted maximum likelihood estimation (REML) for the continuous variable (the normalized frequency of the LCM category). F-test denominator degrees of freedom were estimated using the Kenward-Roger's degrees of freedom adjustment to reduce the chances of Type I error. It is important to point out that mixed effect regression models are very robust with regards to unequal cell sizes, which are a necessary consequence of this dataset.

Given the sheer size of the LCM wordlist, the diversity of topics, senders, and dates (the latter two controlled for in the mixed effect regression model) it is surprising to find any fraudulent event being predicted by the data. Nevertheless, as Table 3 shows, several events can be successfully related to linguistic cues. Recall that, according to the LCM, emails scoring high on abstractness are difficult to verify and are low in informativeness of the situation. Table 3 supports this idea. For instance, during the times that shredding occurred, shredding stopped, and fraud was announced, emails scored higher on abstractness.

Moreover, the most abstract category according to the LCM model is the adjectives. Discussion of ethics, financial support of political candidate, shredding occurs, shredding stopped, fraud announced, and bankruptcy filed, all predicted a higher frequency of adjectives.

Even though these results generate new research questions, there is evidence that the LCM model allows for predicting fraudulent events. Earlier in this study, however, we reviewed studies that found categories such as pronoun use, word quantity, emotion words and cognitive complexity to be affiliated with linguistic cues to deception. Although we do not have access to the exact linguistic cues of some of these categories, we can create an algorithm that approximates these cues. This is what was done in a second study.

Study 2

In the second study we used some of the categories that Newman et al. (2003) and Hancock et al. (2008) reported to be linguistic cues to deception in their experiments: first person pronouns, third person pronouns, causal adverbs, negation (both analytic and synthetic negation), the connective "but", and the length of the email in number of words.

As in Study 1, each of these seven categories was compared with the dummy-coded events in Table 2 using a mixed-effect regression model, thereby controlling for sender and date of the emails.

Table 4 shows the results of this analysis. Events such as fraud announced, bankruptcy filed, fraudulent paperwork filed/signed, and layoffs were related to first person pronouns in emails. However, this relation was in the opposite direction of the one found by Newman et al. (2003) and Hancock et al. (2008).

Fraudulent comments, meetings with national political figures, SEC inquiry developments, and stopping of shredding were related to a higher frequency of negations (analytic negations). It is also noteworthy in these findings that negations were predicted by the stopping of shredding, but not by the occurring of shredding.

Overall, these findings are less uniform than the findings presented in Study 1. This lack of uniformity may be due to the incompleteness with respect to several of the linguistic cues assessed by Newman et al. (2003) and Hancock et al. (2008). Furthermore, the dataset analyzed here did not necessarily represent individual views on situations, unlike the situational data analyzed by Newman et al. (2003) and Hancock et al. (2008). Despite these discrepancies, the findings of the second study are helpful, as a tool of comparison to those in the first study.

Discussion and Conclusion

This study investigated whether linguistic cues can be linked to fraudulent events in a corporate social network. Various studies have looked at linguistic cues to deception. However, unlike the study presented here, these studies used carefully controlled experiments in which participants were asked to give their individual views to a receiver. Most notably, participants were placed in a lying or truthful condition. These studies provide an excellent insight in ways to deceive others, but it is at least an empirical question whether the same linguistic cues can predict deception in more ecologically valid situations. Moreover, it is worth determining whether linguistic cues of deception can be identified in large social networks.

The results of the two studies presented here on the Enron email dataset, a large record of a corporate social network, suggest that abstractness of an email is most indicative of fraudulent events.

By no means are we arguing that by using the LCM model we can predict whether an email consists of fraudulent information or not. At the same time, our results suggest that during times of fraudulent activities messages are sent out with a higher level of abstractness than during times such fraudulent activities are absent or less prevalent.

The work presented here can be extended along a number of dimensions. First, it might well be possible that the LCM categories used here allow for a different abstractness formula that better predicts the result.

To our knowledge this is the first study that has analyzed the impact of fraudulent events on the interpersonal language use of a large social network. Even though the results invite further research, the findings presented here are encouraging, and provide valuable information to the field of deception and interpersonal language use.

Table 3. Significant results mixed effects regression analysis LCM categories. Pluses mark positive relations, minuses negative relations (++ $p < .01$, + $p < .05$, - $p < .05$, -- $p < .01$)

	DAV	IAV	SAV	SV	DA-ADJ	IA-ADJ	SA-ADJ	S-ADJ	ADJ	Abstract-ness
Layoffs	++		++		++	-				
CEO					+			+		-
Fraudulent Paperwork Filed Signed		++								
Fraudulent Comments			-		--					
Discussion of Ethics					--				+	
Selling Enron Shares	-				--					
Rolling Blackouts Initiated					+					
Meetings with Nat'l Political Figures		-		++						
Financial Support of Pol. Candidate			--						++	
Profit Announced					--					
Loss Announced			+		++			+		
SEC Inquiry Developments					++	+		+		++
Shredding Occurs			++		++	+			++	+
Shredding Stopped		+	++		++			+	++	++
Fraud Announced			++		++				++	++
Bankruptcy Filed	++		++		++	-			++	

Table 4. Significant results mixed effects regression analysis various linguistic categories. Pluses mark positive relations, minuses negative relations (++ $p < .01$, + $p < .05$, - $p < .05$, -- $p < .01$)

	1st pers. pronoun	3rd pers. pronoun	causal adverbs	analytic negation	synthetic negation	*but*	word count
Layoffs	+	--		--			-
CEO							
Fraudulent Paperwork Filed Signed	+					-	
Fraudulent Comments				+			
Discussion of Ethics							
Selling Enron Shares			+				++
Rolling Blackouts Initiated							
Meetings with Nat'l Political Figures				+			
Financial Support of Pol. Candidate							
Profit Announced							++
Loss Announced		+					+
SEC Inquiry Developments				++			
Shredding Occurs							+
Shredding Stopped				++			
Fraud Announced	++						
Bankruptcy Filed	++				++		

Acknowledgments

This research was in part supported by grant NIH 1RC1LM010442-01 and NSF 0904909. The usual exculpations apply.

References

Baayen, R. H., R. Piepenbrock, and H. van Rijn (Eds.) (1993). *The CELEX Lexical Database (CD-ROM)*. University of Pennsylvania, Philadelphia (PA): Linguistic Data Consortium.

Clinard, M. & Yeager, P. (2006). *Corporate crime*. New York: Free Press.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

Cohen, W, The Enron e-mail dataset. http://www.cs.cmu.edu/~enron/. Last accessed 2/5/2010

DePaulo, B. M., & Kashy, D. A. (1998). Everyday lies in close and casual relationships. *Journal of Personality and Social Psychology, 74*, 63–79.

DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to deception. *Psychological Bulletin, 129*, 74-118.

Diesner, J., Frantz, T., Carley, K.M. (2005). Communication networks from the Enron email corpus "It's always about the people. Enron is no different". *Computational and Mathematical Organization Theory, 11*, 201-228.

Hancock, J.T., Curry, L., Goorha, S., & Woodworth, M.T. (2008). On lying and being lied to: A linguistic analysis of deception. *Discourse Processes, 45*, 1-23.

Jurafsky, D., & Martin, J.H. (2008). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. New Jersey: Prentice-Hall.

Keila, P.S. & Skillicorn, D.B. (2005). Detecting unusual email communication. *Proceedings of the 2005 conference of the Centre for Advanced Studies on Collaborative research (CASCON 2005)*, 238–246.

Klimt, B. & Yang, Y. (2004). The Enron corpus: A new dataset for email classification research. *Proceedings of the Fifteenth European Conference on Machine Learning*, pp. 217–225.

Louwerse, M.M., & Jeuniaux, P. (2010). The Linguistic and Embodied Nature of Conceptual Processing. *Cognition, 114*, 96-104.

Miller, G. A. & Johnson-Laird, P. N. (1976). *Language and perception*. Cambridge, MA: Harvard University Press

Newman, M. L., Pennebaker, J. W., Berry, D. S., & Richards, J. N. (2003). Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin, 29*, 665–675.

Semin, G. R. & Fiedler, K. (1988). The cognitive functions of linguistic categories in describing persons: Social cognition and language. *Journal of Personality and Social Psychology, 54*, 558-568.

Semin, G. R. & Fiedler, K. (1991). The linguistic category model, its bases, applications and range. *European Review of Social Psychology*, 1-30.

Semin, G. R. (2000). Communication: Language as an implementational device for cognition. *European Journal of Social Psychology, 30*, 595-612.

Stapel, D. & Semin, G. R. (2007). The magic spell of language. Linguistic categories and their perceptual consequences. *Journal of Personality and Social Psychology*, 93, 23-33.

The Cognitive Cost of Ethnocentrism

Artem Kaznatcheev (artem.kaznatcheev@mail.mcgill.ca)
Department of Physics and School of Computer Science,
McGill University, 3600 University Street, Montreal, QC H3A 2T8 Canada

Abstract

Recent computational studies suggest that ethnocentrism, commonly thought to rely on complex social cognition, may arise through biological evolution in populations with minimal cognitive abilities. We use the methods of evolutionary game theory and computational modelling to examine the evolution of ethnocentrism. Since ethnocentric agents differentiate between in- and out-group partners, and adjust their behavior accordingly, they are more cognitively complex than humanitarian or selfish agents that always cooperate or defect, respectively. We associate a fitness cost with this complexity and test the robustness of ethnocentrism, concluding that ethnocentrism is not robust against increases in cost of cognition. Our model confirms that humanitarians are suppressed largely by ethnocentrics. Paradoxically, we observe that the proportion of cooperation is higher in worlds dominated by ethnocentrics. We conclude that suppressing free-riders, such as selfish and traitorous agents, allows ethnocentrics to maintain higher levels of cooperative interactions.

Keywords: Ethnocentrism; humanitarianism; cooperation; agent-based simulation; minimal cognition; Prisoner's Dilemma; evolution; free-rider-suppression hypothesis

Introduction

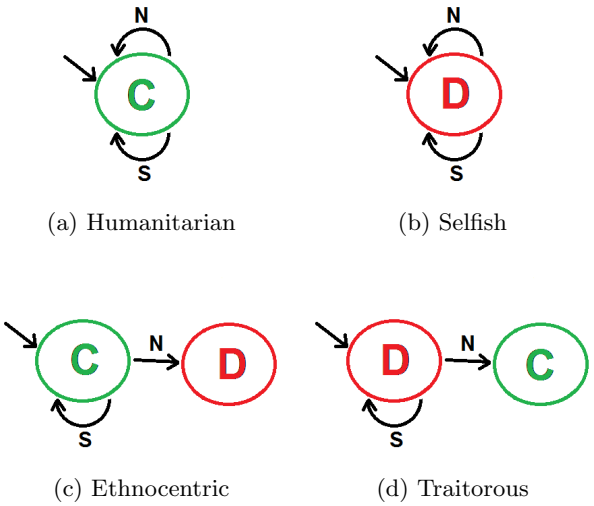
Seeing ones own group (in-group) as superior and other groups (out-groups) as inferior is a widespread syndrome of discriminatory behaviors (LeVine & Campbell, 1972). This perspective is associated with behavior including in-group favoritism (ethnocentrism) and out-group hostility (xenophobia) (Cashdan, 2001; Hewstone, Rubin, & Willis, 2002; Brown, 2004). Although the behavior is commonly thought to involve substantial cognitive ability (Sherif, 1966; LeVine & Campbell, 1972; Hewstone et al., 2002), extensive psychological evidence suggests that the presence of a strong in-group bias can be observed in individuals with minimal cognition and highly abstract social input (Tajfel, 1970; Tajfel, Billig, Bundy, & Flament, 1971). This is supported by observed ethnocentrism in human placenta (Haig, 1996), ants (Keller & Ross, 1998), and microbes (Lenski & Velicer, 2000), and suggests that ethnocentrism may have a basis in biological evolution. Cognitively, the ability to distinguish in- and out-group members and adjust behavior accordingly may be sufficient to foster this effect.

Recent computational studies (Hammond & Axelrod, 2006; Shultz, Hartshorn, & Hammond, 2008; Shultz, Hartshorn, & Kaznatcheev, 2009) have focused on the emergence of in-group favoritism through agent-based simulations of individuals with minimal cognitive ability. Agents interact via a one-time prisoners dilemma (PD) game that affects the reproductive potential of the participants. Agents can either defect against, or cooperate with, other in- or out-group agents, permitting four strategies: (1) a humanitarian strategy of universal cooperation, (2) an ethnocentric strategy of in- but not out-group cooperation, (3) a traitorous strategy of cooperation exclusively with the out-group, and (4) a selfish strategy of constant defection. Hammond and Axelrod (2006) showed that, after a transient period, ethnocentric agents dominate the population. Shultz et al. (2008) examined the transient period to uncover evidence for early competition between the ethnocentric and humanitarian strategies. More recently, Shultz et al. (2009) focused on explaining the mechanism behind ethnocentric dominance over humanitarians. In particular, they introduced the direct and free-rider-suppression hypotheses. The direct hypothesis is that ethnocentric clumps of agents directly suppress contacted clumps of humanitarian agents from different groups. The contrasting free-rider-suppression hypothesis is that ethnocentrics are more effective than humanitarians at suppressing groups of free riders — selfish and traitorous agents from the same group.

This paper extends beyond previous work by closely examining the cognitive mechanisms required for ethnocentrism. In particular, we measure the cost in fitness an agent is willing to pay in order to have the higher (yet still simple) cognitive processes required to discriminate between in- and out-groups. By varying the cost of cognition we also eliminate ethnocentric agents and confirm the direct hypothesis as the mechanism of ethnocentric dominance over humanitarians. Tracking the proportion of cooperation also reveals a novel and important role for the free-rider-suppression hypothesis — maintaining cooperative interactions.

Cognitive Complexity

An easy way to understand the complexity of reasoning carried out by simple abstract agents is to represent their decision procedure by finite state machines (FSMs). To use FSMs, it is important to understand what information agents receive and what actions they perform based on that information. During an interaction the agent receives some information about its partner and then makes a decision to cooperate or defect. In particular, the agent receives a signal S if the agent's partner is from the same abstract group (in-group) and N otherwise (out-group). In response, the agent outputs a D to defect in the PD interaction, and C to cooperate. Note that the agent does not receive any direct information



(a) Humanitarian (b) Selfish

(c) Ethnocentric (d) Traitorous

Figure 1: Finite state machines representing the 4 possible strategies. Transitions are represented by arrows, with label S corresponding to an input of same-tag and N to different-tag. The agent's action is represented in the center of the state, C for cooperate and D for defect.

on its partner's strategy, and instead relies on potential correlations of a partner from the same group having the same strategy.

The four strategies are represented by FSMs in figure 1. Circles correspond to states, that are labeled by their output, either C or D. Arrows represent transitions, labeled by S or N. The initial state is signified by the smaller arrow with no pre-state. Humanitarian agents that always cooperate, and selfish agents that always defect are easy to represent. In particular, they are single state machines that output C or D (respectively) regardless of the input they receive. These two strategies are shown schematically in figures 1a and 1b. Ethnocentric and traitorous agents, however, require two states as shown in figures 1c and 1d. The extra state represents the greater complexity in making a decision based on input received, compared to not making a decision at all (the single state agents). Since implementing this rudimentary decision-making requires extra energy expenditure on the part of the agents, we represent this extra cost as a small fitness decrement k for ethnocentric and traitorous agents. This cost is especially important for studying the co-evolution of cooperation and ethnocentrism. To follow an ethnocentric strategy the agents have to invest a bit of their fitness into developing more sophisticated cognitive processes.

Method

Prisoner's Dilemma

In virtually any competitive social situation, interacting agents have a basic decision to make: cooperate

		Agent 2	
		C	D
Agent 1	C	$b-c$	$-c$
	D	b	0

Table 1: Payoff matrix for one agent (agent 1) interacting with another (agent 2).

with each other, or not. In evolutionary game theory such interactions are usually modelled by the Prisoner's Dilemma (PD). In PD, two agents independently decide whether to cooperate with or defect against the other. When cooperating, an agent pays a cost c to provide a greater benefit b to its partner. When defecting, an agent pays no cost and provides no benefit, but can still receive benefit if its partner cooperates. Table 1 shows the payoff for one agent (agent 1) interacting with another (agent 2). The payoff matrix reveals that an agent can always receive a higher payoff by defecting instead of cooperating. In game theoretic terminology, mutual defection is the Nash equilibrium. However, if both agents manage to coordinate cooperation, then they are both better of than if they had mutually defected — mutual cooperation Pareto dominates mutual defection. Due to this paradox, the PD game is regarded as a paradigmatic example of the problem of achieving mutual cooperation (Axelrod & Hamilton, 1981). The game provides a simple model of an environment where one action (defection) is better for the individual and the other (cooperation) is better for the population.

The simplest approach to studying the evolutionary dynamics of the PD is in a well mixed population. If agents are paired randomly from a mixed population and interaction results modify individual reproductive potential, eventually defectors will dominate the population. To allow cooperation to emerge it is essential to introduce positive correlations between the strategies of paired agents. To study the emergence of cooperation, researchers explore various ways to create these correlations. In our model we consider an interplay of spatial structure and arbitrary tags.

Model

Our model, and the Hammond and Axelrod (2006) model it is based on, expand beyond random interactions to facilitate the emergence of cooperation. Instead of randomly choosing interaction pairs, agents populate a toroidal square lattice (50 by 50 cells) and interact with their four adjacent neighbors. Each individual is simple, only perceiving whether it shares a common tag with neighbors (from a total of 4 tags), allowing for two interaction strategies: an in-group (igs) and an out-group (ogs) strategy. The four strategies are summarized in figure 1 and table 2. The outcomes of PD interactions

Name	igs	ogs	Figure
Humanitarian	C	C	1a
Ethnocentric	C	D	1c
Traitorous	D	C	1d
Selfish	D	D	1b

Table 2: The four possible strategies. The igs column correspond to the in-group strategy, ogs to out-group strategy.

(with $b = .025$ and $c = .01$) are added to the agents potential to reproduce (ptr, which is reset to .1 at the start of each cycle). At the end of a cycle, each agent has a chance equal to its ptr to clone itself (with a constant mutation rate (.005)) and a constant probability (.1) of dying. If an agent expires its location is vacated until habitation by a new agent. Regardless of the agent's survival, if the agent cloned itself the child is placed in one empty cell adjacent to the parent (potentially including the parent's cell if the parent expired after cloning itself). To start the world, and if the population ever reaches zero, the world is seeded with 80 individuals distributed randomly across the torus and uniformly across the 16 strains (4 possible tags, 2 possible igs, 2 possible ogs). The simulation runs for 3000 cycles.

To account for the potential cost of extra cognitive abilities we introduce a new parameter k. This parameter is varied in different simulations from 0 to .01 to show the impact of cognitive costs on the evolution of ethnocentrism. The effect of the cost of cognition is to lower the base ptr of ethnocentric and traitorous agents. In particular, at the start of every cycle the ptr is reset to .1 for humanitarian and selfish agents and to $.1 - k$ for ethnocentric and traitorous agents. The rest of the simulation is unmodified. By adjusting k we can quantify how much the extra cognitive powers of ethnocentrics are worth in terms of the currency of evolution — reproductive fitness.

During the simulation we collect two primary types of data. We record the distribution of agents by strategy, and track the number of cooperative and non-cooperative interactions. If an agent chooses to cooperate during its interaction, we increment the number of cooperations for the cycle. The proportion of cooperation is then the number of cooperations divided by twice the number of interactions (to account for both agents having to make a decision during each interaction).When comparing simulations with different values of k we take the mean data from the last 500 cycles, since the dynamics stabilize by then. To account for the stochastic nature of our model, we present all results with standard error from averaging over 30 worlds (a world is a single instance of the simulation with specific parameter setting and initial random seed).

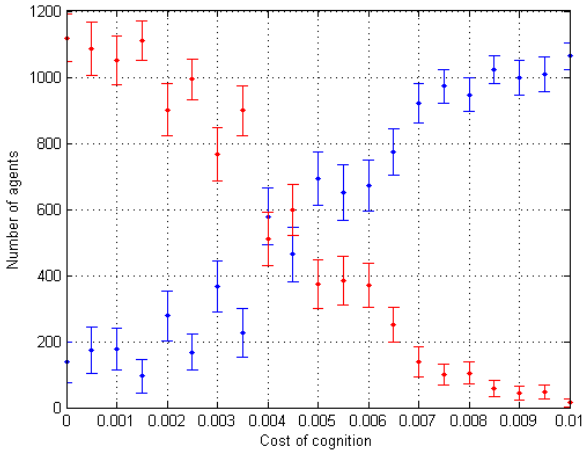


Figure 2: Number of ethnocentric and humanitarian agents vs. cost of cognition. The points represent the mean number of agents over the last 500 cycles of the simulations: red is ethnocentric agents, blue is humanitarian. The error bars represent standard error from averaging over 30 different worlds.

Results

Our primary result is the variation in the number of ethnocentric and humanitarian agents and the proportion of cooperation for different values of the cost of cognition, k. As in previous studies (Hammond & Axelrod, 2006; Shultz et al., 2008, 2009) the number of selfish and traitorous agents is negligible, and hence we concentrate our presentation on the number of humanitarians and ethnocentrics in the population. Figure 2 shows the number of ethnocentrics (in red) and humanitarians (in blue) in simulations with increasing costs of cognition. Unsurprisingly, as the cost of cognition increases the number of ethnocentric agents decreases, and humanitarians take their place. This is consistent with the assessment of Shultz et al. (2009) that humanitarians are directly suppressed by ethnocentrics. A surprising result, is how low the cost of cognition ($.004 \leq k \leq .0045$) is at the point where the humanitarian strategy becomes fitter. In particular, this transition point is more than an order of magnitude lower than the default ptr (.1) and less than half of the cost of cooperation (.01). The third interesting result, is the quick phase transition from ethnocentric to humanitarian dominance. With $k \leq .0035$ the number of humanitarian agents is relatively stable around 200, from $k = .004$ to $k = .0065$ we observe quick change, and then for $k \geq .007$ the humanitarian population stabilizes around 1000 individuals. Together, these results suggest that ethnocentrism is not very robust against variance in the cost of cognition. In particular, for widespread ethnocentrism to emerge, the cost of differentiating between in- and out-groups needs to be

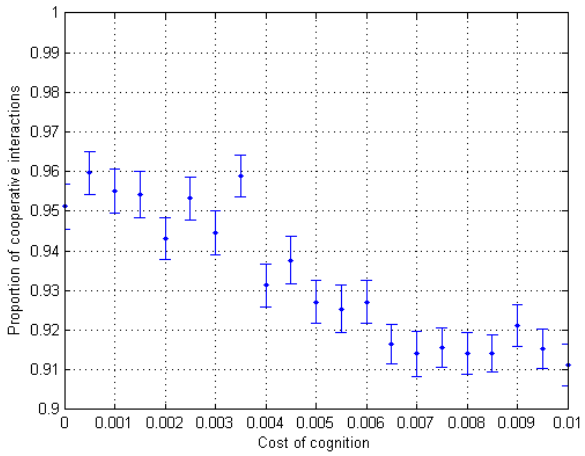


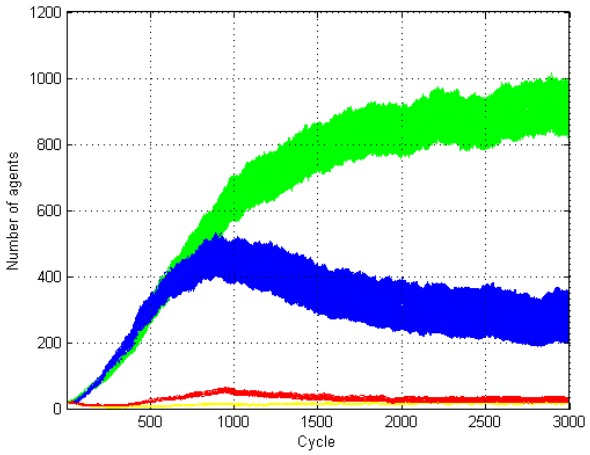
Figure 3: Proportion of cooperative interactions vs. cost of cognition. The points represent the mean proportion of cooperative interactions over to last 500 cycles. The error bars represent standard error from averaging over 30 different worlds.

extremely low in comparison to other relevant parameters (default ptr, cost and benefit of cooperation, etc.).

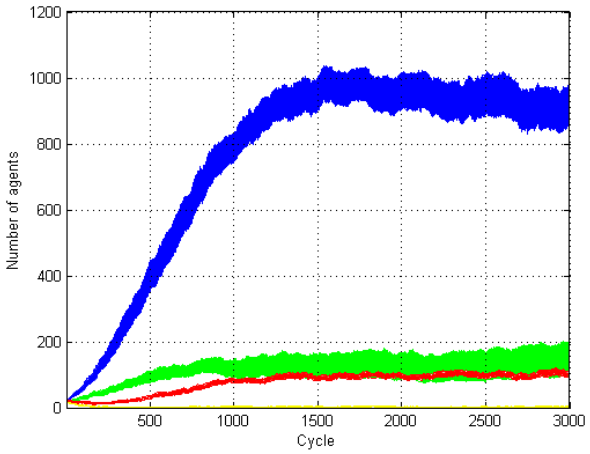
In figure 3 we show the proportion of cooperative interactions as the cost of cognition increases. Although the change in cooperation is not as drastic as the changes in strategy distribution, it is still statistically significant. In particular, we observe the same phase transition, with proportion of cooperative interactions stable around .955 while $k \leq .0035$ and stable around .915 when $k \geq .0065$. The counter-intuitive result in figure 3 is that as humanitarian agents start to dominate the population, the proportion of cooperation decreases. This raises the important question of what is more important: overall cooperation or the individual fairness of predominantly humanitarian agents?

For contrast, we also present two figures of strategy distribution by cycle. In figure 4a we show a cost of cognition $k = .002$, a bit before the phase transition. In figure 4b we examine a point after the phase transition, $k = .007$. The drastic change from ethnocentric (green) dominance to humanitarian (blue) dominance is self-evident. Further, in the humanitarian-dominated world of $k = .007$ selfish (red) agents perform around 5 times better than in the ethnocentric dominated world. This supports the intuition that ethnocentrics are better at suppressing selfish agents.

For other nearby choices of basic parameters (b, c, base ptr, and death rate) the qualitative results are similar, although the exact quantitative aspects change. As in reports (Shultz et al., 2008, 2009), we omit them for brevity.



(a) $k = .002$



(b) $k = .007$

Figure 4: Number of agents of various strategies vs. evolutionary cycle. The lines represent the number of agents of each strategy: blue — humanitarian; green — ethnocentric; yellow — traitorous; red — selfish. The width of the line corresponds to standard error from averaging 30 different worlds. The two figures correspond to different costs of cognition, k.

Discussion

The relatively low cost of cognition (around $.004 \leq k \leq .0045$) required to transition from ethnocentric to humanitarian dominance suggests that ethnocentrism is not very robust. In particular, the emergence of ethnocentric cooperation is unlikely to have caused significant investment of fitness in cognitive development. Alternatively, the mechanisms for differentiating between in- and out-groups and making basic decisions would need to be already in place by other means, and are unlikely to have co-evolved with cooperation. Making the distinction between in- and out-groups does not require fitness investment for humans, but for more rudimentary organism, it is likely that it would. Although the cognitive abilities required for ethnocentrism are as simple as distinguishing in- and out-groups, this simplicity can be deceiving. Our results stress that for ethnocentrism to evolve, these simple cognitive abilities must also be extremely cheap in terms of fitness invested.

The results in figure 3 suggest that displacing ethnocentric agents by humanitarian ones can lead to a decrease in overall proportion of cooperative interactions. In particular, it is important to reexamine the negative perception of ethnocentrism. Although unfair from the individual point of view, ethnocentrism might be essential to sustain the levels of cooperation required for complex structures such as multi-cellular organisms or human society. Thus, a tempting answer to the question of Shultz et al. (2009): "why is ethnocentrism more common than humanitarianism?" is that humanitarianism cannot maintain as high levels of cooperation.

A further connection to previous work (Shultz et al., 2009) is a reevaluation of the direct and free-rider-suppression hypothesis. Although Shultz et al. (2009) ruled out the free-rider-suppression hypothesis in favor of the direct hypothesis, they did not examine the proportion of cooperation. The direct hypothesis provides a good explanation of why ethnocentrics dominate humanitarians, but the free-rider-suppression hypothesis explains the increased levels of cooperation in largely ethnocentric populations. When humanitarians replace ethnocentric as the dominant strategy, significantly higher levels of selfish agents evolve in the population. The decrease in cooperation caused by higher levels of selfish agents exceeds the increase in cooperation caused by humanitarians cooperating across groups. This results in an overall reduction in the cooperative interactions. Thus, ethnocentric agents ability to better suppress free-riders is important for maintaining higher levels of cooperative behavior.

Although the decision making employed by our abstract agents is extremely simple, it is not beyond the scope of what contemporary cognitive science regards as cognition. Our research explores rudimentary cognition in a social and evolutionary context. In particular we hope that this paper highlights the importance of considering possible fitness investment in even the simplest forms of cognition. By exploring further we hope to gain a better understanding of the evolution and potential social effects of simple information processing.

References

Axelrod, R., & Hamilton, W. (1981). The evolution of cooperation. *Science*, *211*, 1390-1396.

Brown, D. E. (2004). Human universals, human nature, and human culture. *Daedalus*, *133 (4)*, 47-54.

Cashdan, E. (2001). Ethnocentrism and xenophobia: A cross-cultural study. *Current Anthropology*, *42*, 760-765.

Haig, D. (1996). Gestational drive and the green-bearded placenta. *Proc. Natl. Acad. Sci. USA*, *93*, 6547-6551.

Hammond, R., & Axelrod, R. (2006). The evolution of ethnocentrism. *Journal of Conflict Resolution*, *50*, 926-936.

Hewstone, M., Rubin, M., & Willis, H. (2002). Intergroup bias. *Annual Review of Psychology*, *53*, 575-604.

Keller, L., & Ross, K. (1998). Selfish genes: a green beard in the red fire ant. *Nature*, *394*, 573-575.

Lenski, R., & Velicer, G. (2000). Games microbes play. *Selection*, *1*, 51-57.

LeVine, R. A., & Campbell, D. T. (1972). Ethnocentrism. *New York: John Wiley*.

Sherif, M. (1966). Group conflict and co-operation: Their social psychology. *London: Routledge Kegan Paul*.

Shultz, T. R., Hartshorn, M., & Hammond, R. A. (2008). Stages in the evolution of ethnocentrism. In B. Love, K. McRae, & S. V.M. (Eds.), *Proceedings of the 30th annual conference of the cognitive science society* (p. 1244-1249). Austin, TX: Cognitive Science Society.

Shultz, T. R., Hartshorn, M., & Kaznatcheev, A. (2009). Why is ethnocentrism more common than humanitarianism? In N. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st annual conference of the congnitive science society* (p. 2100-2105). Austin, TX: Cognitive Science Society.

Tajfel, H. (1970). Experiments in intergroup discrimination. *Scientific America*, *223*, 96-102.

Tajfel, H., Billig, M. G., Bundy, R. P., & Flament, C. (1971). Social categorization and intergroup behavior. *European Journal of Social Psychology*, *1*, 149-178.

Self-esteem and the Matching Effect in Mate Selection

Artem Kaznatcheev (artem.kaznatcheev@mail.mcgill.ca)
Department of Physics and School of Computer Science,
McGill University, 3600 University Street, Montreal, QC H3A 2T8 Canada

Kyler Brown (kyler.brown@mail.mcgill.ca)
Laboratory for Natural and Simulated Cognition, Department of Psychology,
McGill University, 1205 Penfield Avenue, Montreal, QC H3A 1B1 Canada

Thomas R. Shultz (thomas.shultz@mcgill.ca)
Department of Psychology and School of Computer Science,
McGill University, 1205 Penfield Avenue, Montreal, QC H3A 1B1 Canada

Abstract

The matching effect is the empirical finding that romantic couples have a high correlation in physical attractiveness. It remains a debate as to whether this correlation is based purely on similarity preference - the matching hypothesis - or marketplace forces. We present a new marketplace model for romantic relationships. Previous models granted every person access to his/her own attractiveness. In reality, people have only a vague idea of their own attractiveness ratings. We introduce a concept analogous to self-esteem to model this phenomenon. Further, we extend beyond previous models by dealing explicitly with both the initialization and development of a relationship. Our model accounts for the experimental tendency to choose more attractive partners, while still explaining observed intra-couple attractiveness correlation and the difference in correlation between casual and serious daters.

Keywords: mate selection; matching hypothesis; self-esteem; social cognition

Introduction

The study of dating behavior in humans has both shed light on and raised many questions about the dynamics of human relations. A common parameter across myriad studies of human partner selection has been the physical attractiveness of individuals (Cash, 1981). Walster, Aronson, Abrahams, and Rottman (1966) attempted to address the tyranny in the advantages of attractive individuals by proposing the matching hypothesis. Basing their hypothesis on the Level of Aspiration Theory (Lewin, Dembo, Festinger, & Sears, 1944), they predicted that, in making a realistic social choice, an individual would choose a partner similar in social desirability. Simply, when faced with a realistic choice, one would choose a romantic partner of nearly identical physical attractiveness.

The theoretical sanctuary that the matching hypothesis offers from physical misgivings has not been well supported by direct experimental tests (Kalick & Hamilton, 1986). When homely men estimated attractive women as more likely to reject them compared to their handsome brethren, no significant difference in their choice of prospective partners was observed: both groups opted to choose the more attractive women (Huston, 1973). Even the most promising early experiments (Berscheid, Dion, Walster, & Walster, 1971) showed only a weak matching effect (Wetzel & Insko, 1982) that was overpowered by the attractiveness effect. So, why has the matching hypothesis survived?

Direct experiment is not the only way to test the hypothesis; one can also observe existing couples. The matching theory has been consistently supported when the correlation between the attractiveness of male and female partners in real couples was studied (Kalick & Hamilton, 1986). In average couples correlations of .38 (Murstein, 1972a), .39 (Price & Vandenberg, 1979), .42 (Feingold, 1981) and .53 (Citelli & Waid, 1980) were found. Further studies (White, 1980) expanded their investigation to differentiate between the type (and associated longevity) of relationships, observing intra-couple attractiveness correlations of .18 for casual dates, compared to the correlation of .56 and .63 for serious daters and engaged or married couples, respectively. The strong correlations in real couples provide the main evidential support for the matching hypothesis.

The stark dichotomy between direct experiment and observations of existing couples raises the question: is the matching hypothesis a good model for human courtship? By itself the matching hypothesis fails to match experiment but corresponds well to correlation studies in existing couples. It is the goal of this paper to provide a synthesis of the hypothesis and the apparent preference for the most attractive partner into a single computational model.

The computational approach was famously pursued by Kalick and Hamilton (1986). The Kalick and Hamilton simulation assumed every person has access to both their own attractiveness rating and their partner's attractiveness rating. Experimentally, the latter assumption is valid. Cunningham and Wu (1995) found a correlation of .9 between a single rating and the average rating of pictures of women. This correlation remains high if either the female picture of the rater is from a different culture. The assumption of access to own attractiveness, however, is not supported by experiment. Rand

and Hall (1983) found that people are very inaccurate at rating their *own* attractiveness. Females tend to have a .5 correlation between their self-perception of attractiveness and the rating of male judges. Males have only a .1 correlation between self-attractiveness ratings and the ratings of female judges. The inability of people to accurately judge their own attractiveness cannot be disregarded when simulating the matching hypothesis. Hence, for a model to have ecological validity, it must incorporate the inaccuracy of judging self-attractiveness.

Our model incorporates this self-perceptive inaccuracy through the effect of a variable self-esteem (or body-image) rating. We use a simple model of self-esteem based on sociometer theory (Leary, 2005). As suggested by naturalistic studies, self-esteem mediates self-perception of attractiveness (Fleming & Courtney, 1984; Feingold, 1988; Leary, 2005) and changes based on acceptance (or rejection) in the initiation of relationships and by the dissolution of relationships (Helgeson, 1994; Leary, 2005; Pass, 2009; Pass, Lindenberg, & Oark, 2010). We also extend beyond previous models by dealing explicitly with both the initialization and development of a relationship. This allows us to study the expected difference in attractiveness correlation between casual and serious daters (Cavior & Boblett, 1972; White, 1980) and track the effects of break-ups on self-esteem.

Method

The method of simulation is widely used to help understand certain types of complex systems. Models of human courtship lend themselves particularly well to simulation, since the goal is to define relatively simple rules for individual parts (people) and observe a more complex behavior and trend in the whole system (group). In our model, each individual i is parametrized by two values: a static $\alpha_i \in (0, 1)$ to represent the person's attractiveness and a dynamic $s_i \in (-1, 1)$ (referred to as 'self-esteem'). Together these parameters are used to derive $A_i \in (0, 1)$ — the person's perception of their own attractiveness. The two parameters that describe a person (α_i and s_i) are generated randomly from a uniform distribution. If two individuals i and j form a couple, then the relationship caries an extra parameter, $l_{ij} \in \mathbb{N}$, called longevity. Longevity counts the number of 'dates', or amount of time, i and j have been in a relationship. The longevity parameter is used to track the longest lasting couples and is reset to zero upon relationship dissolution.

Individuals are not explicitly given a gender, but the simulation is constructed such that males only ever show up in the list of male individuals (or the male side of a relationship) and vice-versa for females. For simplicity, the simulation is restricted to have the same number of male and female individuals. At the start all individuals are initialized as singles (not part of a couple) and only heterosexual relationships were considered. The simulation proceeds in discrete steps (epochs). On each epoch we follow the procedure:

1. existing couples are examined for a potential break up,
2. agents from dissolved couples are reintegrated into the pool of singles,
3. new couples are formed from the pool of singles, and
4. statistical data collected.

Any changes to self-esteem are incorporated at the instant they occur.

Formation and dissolution of relationships

The probability of date formation is based around the empirical observations that individuals seek the most attractive partners regardless of their own attractiveness (Huston, 1973; Kalick & Hamilton, 1986). In the simulation, each single man i is paired with a single woman j and each decides if they want to accept the date based on a probability of acceptance equal to the attractiveness of their potential partner ($P(m_{ij}) = \alpha_j$). If both partners accepts, then the pair become a couple, l_{ij} is initialized and self-esteem is modified as detailed in the next subsection.

For established couples, the break up probability is based on equity theory and the matching hypothesis. Since a break-up is seldom mutual (Hill, Rubin, & Peplau, 1976) we compute a separate break up probability for each member of the couple. Given a couple of woman x and man y the break up probability, $P(b_{xy})$ and $P(b_{yx})$, is calculated for each person, respectively, according to equation 1. The probability of i breaking up with j is linearly dependent on the absolute difference between i's perceived attractiveness, A_i, and his partner's actual attractiveness α_j. The dependence on absolute difference in perceived attractiveness is based in equity theory (Murstein, 1972b; Walster, Hatfield, Walster, & Berscheid, 1978) and the empirically observed importance of similar physical attractiveness to the longevity of relationships (Hill et al., 1976; Feingold, 1988). The values of 0.15 and 0.85 are arbitrary, but by rescaling time we can always assume the values we chose add up to 1.

$$P(b_{ij}) = 0.15 + 0.85|A_i - \alpha_j| \tag{1}$$

If the couple ij remains, then one more 'date' is added to their longevity ($l_{ij} \leftarrow l_{ij} + 1$). If at least one of i or j decides to break up with the other then the relationship ends, both individuals are added to the singles list before new couples are formed, and l_{ij} is reset to zero. The impact on individual's self-esteem depends on whether the dissolution was mutual or unilateral.

Self-esteem effects

The primary effect of the self-esteem variable s_i is on i's perception of its own attractiveness. Our model of the effect of self-esteem on self-perception is grounded in the Fleming and Courtney (1984) finding that self-ratings of attractiveness loaded heavily on self-esteem factors. In particular, we use equation 2 to determine an individual's self-perceived attractiveness A_i in terms of their actual (externally determined and static) attractiveness α_i and their varying self-esteem s_i.

$$A_i = \begin{cases} \alpha_i + (1-\alpha_i)s_i, & s_i \geq 0 \\ \\ \alpha_i(1+s_i), & s_i < 0 \end{cases} \tag{2}$$

Equation 2 is the simplest choice of equation that ensures that any value of actual attractiveness $\alpha_i \in (0,1)$ and self-esteem $s_i \in (-1,1)$ results in a perceived self-attractiveness A_i in the correct range of $(0,1)$. From the upper clause of equation 2 we can see that a positive s_i produce a linear increase in perceived attractiveness from $A_i = \alpha_i$ for $s_i = 0$ to $A_i = 1$ for $s_i = 1$. Thus, $s_i > 0$ corresponds to an overly high self-esteem or even arrogance and an overestimation of one's own physical attractiveness. In the lower clause of equation 2 we see that $s_i < 0$ produce a linear decrease in perceived attractiveness from $A_i = \alpha_i$ for $s_i = 0$ to $A_i = 0$ for $s_i = -1$. Negative s_i model a low self-esteem. A perfect judgement of one's own attractiveness is achieved with the 'perfect' esteem of $s_i = 0$.

Through its effect on A_i, self-esteem is important for the duration of relationships. However, in the formation of couples we only consider the actual attractiveness α_i and self-esteem plays no role. We do not incorporate self-esteem in the selection of a mate because Walster (1970) established that self-esteem has no effect on the tendency to prefer the most attractive choice of partner.

The key difference between α_i and s_i is that α_i is static throughout the simulation and s_i varies depending on social interactions. In other words, a person's physical attractiveness is not affected by social interactions, but their self-esteem, self-image, or body-image is affected (Leary, 2005; Pass, 2009). To lower an agent i's self-esteem by a factor x without exceeding the range of $(-1,1)$ we use $d_i(x)$:

$$d_i(x) = \begin{cases} s_i - x, & s_i \geq 0 \\ \\ s_i - (1+s_i)x, & s_i < 0 \end{cases} \tag{3}$$

and to raise it by a factor x we use $u_i(x)$:

$$u_i(x) = \begin{cases} s_i + (1-s_i)x, & s_i \geq 0 \\ \\ s_i + x, & s_i < 0 \end{cases} \tag{4}$$

If an agent has a positive self-esteem ($s_i \geq 0$) and we lower it by x with equation 3 then we simply subtract x from the agent's esteem. If an agent has a negative self-esteem ($s_i < 0$) then we need to worry about potentially reducing it past -1 and so we do as equation 2: lower self-esteem linearly from $d_i(0) = s_i$ to $d_i(1) = -1$. The same procedure is used in equation 4 except with negative and positive esteem swapped and raising instead of lowering. $u_i(x)$ and $d_i(x)$ allow us to increase and decrease an agent i's self-esteem in a simple and consistent way without leaving the range $(-1,1)$.

During the relationship forming stage, if both agents accept the relationship then each receives a self esteem boost: $s_i \leftarrow u_i(0.3)$. This corresponds to the feeling of well being individuals receive from the social acceptance of relationship formation as predicted by sociometry theory (Leary, 2005). On the other hand, if agent i proposes the relationship, but agent j declines, then agent i suffers a self-esteem loss from rejection (in our model: $s_i \leftarrow d_i(0.2)$) and agent j receives a small self-esteem boost from the flattery and reassurance of their attractiveness($s_j \leftarrow u_i(0.1)$) (Pass et al., 2010). If both agents reject the potential pairing then self-esteem is left unchanged because neither individual proposed a relationship.

The most drastic effects on self-esteem are in the case of unilatteral termination of a relationship (Helgeson, 1994). If one of the individuals decides to break up with the other, then the dumped agent's self-esteem is lowered to a new level: $s_i \leftarrow d_i(0.4)$. However, if both individuals want the relationship to end, then neither self-esteem is affected. Although the specific values 0.1, 0.2, 0.3, and 0.4 in our model are chosen for simplicity, the relative ordering of them is meant to correspond to the general ordering observed by Helgeson (1994): break-ups are the most damaging ($d_i(0.4)$), with rejection less damaging ($d_i(0.2)$) and the awards for acceptance higher for a new relationship ($u_i(0.3)$) compared to just the flattery of an offer ($u_i(0.1)$).

Results

To provide an idea of how effective the model is while keeping errors and simulation times reasonable, the simulation was run 50 times with 300 men and 300 women courting for 50 epochs. The main observed quantity was the mean intra-couple attractiveness correlation for the couples in each epoch. Figure 1 provides a visualization of the collected data. The mean correlation was collected for all of the couples in each epoch (blue), as well as the top 30% by longevity (red). Effectively, the blue points represent the 'average' daters and asymptote at around $r = .23$. The top 30% correspond to the 'serious' daters and asymptote near $r = .60$ which is in the observed range of .56 to .63 for serious and engaged or married couples (White, 1980). The large gap between between the attractiveness correlation in average and serious daters is consistent with White's (1980) em-

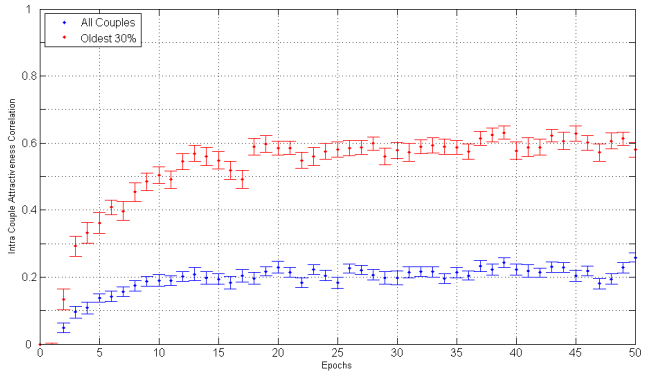


Figure 1: The intra-couple attractiveness correlation versus epochs with statistical error. The blue data points are the correlations of all of the couples in an epoch. The red data points are of the oldest 30% of the couples. The data were generated by averaging over 50 simulations of 50 epochs with 300 men and 300 women courting

pirical results. The lower correlation value of our model also matches empirical data much better than the unreasonably high correlations of earlier models (Kalick & Hamilton, 1986).

Discussion

Earlier simulations (Kalick & Hamilton, 1986) yielded an intra-couple attractiveness correlation of approximately .9, .85, and .55 for the matching hypothesis, combined, and mate attractiveness selection rules respectively. Kalick and Hamilton (1986) concluded based on these simulations that the matching hypothesis alone could not explain intra-couple attractiveness correlations as they were simply too high. By introducing modified rules that incorporated both formation and development of relationships, our model provided realistic correlations of .23 for average daters and .60 for serious and engaged or married couples. We matched experimental results of attractiveness selection by allowing partners to favour accepting dates with more attractive partners. We incorporated the matching hypothesis in the break-up probability instead of relationship formation. This allowed our model to track both the formation and development of relationships. Allowing couples to break-up also addressed an important shortcoming of earlier models (Aron, 1988). By allowing individuals to be single instead of eventually forcing everyone into a relationship we ensure that there is always choice of potential partners.

Our model has provided promising results, but only a portion of its potential has been examined. The model and simulation were used to show how the matching hypothesis can be present in a place other than the probability of date acceptance. This approach accounts for matching effects (especially in long-lived couples) while allowing for the experimental tendency to choose more attractive partners. The simulation could be extended to allow one of the sexes to select a potential partner (instead of random assignment). We believe that such a modification is essential to account for the asymmetry in male and female perception of self-attractiveness. In particular, if males select a potential partner more often, then they will face rejection more often than females and produce more variation in self-esteem and hence a lower correlation between self-perceived and externally judged attractiveness. However, the most important part of the model that needs more attention and study is the self-esteem variables and the choices of weights in various equations. As it stands, lack of knowledge about the self-esteem factor is the largest limitation of the model. To truly test and understand the model and simulation, experiments are essential.

The structure of the simulation and relative simplicity of the model, lends itself nicely to empirical studies. Our model's predictions could be tested with human participants. The attractiveness score of each individual could be evaluated by a panel of judges or by querying participants of the other gender. Individuals' self-esteem parameter could be estimated by comparing their own evaluation of attractiveness, A_i, to the attractiveness assigned by judges, α_i. The dates and choices to break up or accept partners can be carried out as in existing studies. The computer simulation can be run with the same initial population of parameters and results compared. By doing parameter fitting on the inputs for equations 3 and 4 we could estimate the effects of rejection and acceptance on self-esteem.

A further contribution of our simulation is the clarity a formal model brings to theories of human romantic relationships. This clarity allows us to easily generate hypotheses and, more importantly, to relate our model to work in the nearby fields of evolutionary and cognitive psychology. In particular, we hope that — using attractiveness as a proxy for fitness (Singh, 1993; Hönekopp, Rudolph, Beier, Liebert, & Müller, 2007) — future work can connect our social/psychological model to evolutionary and cognitive models. The methods of evolutionary game theory have already been used to study parts of equity theory such as the evolution of fairness in the ultimatum game (Nowak, Page, & Sigmund, 2000; Bolton & Ockenfels, 2000) and the predominance of ethnocentrism (Hammond & Axelrod, 2006; Shultz, Hartshorn, & Kaznatcheev, 2009). Recently, Kaznatcheev (2010) incorporated cognition into these evolutionary models. Recasting our model of mate selection in such a setting can provide important insights into the basis of romantic relations. By looking at the evolutionary and cognitive underpinning of mate selection (Miller & Todd, 1998), future work could explain not only *how* romantic relationships progress, but *why* this is the case.

Our model offers a new and alternative look at the dynamics of romantic relationships. Unlike earlier studies (Kalick & Hamilton, 1986), not only the initialization of a relationship is examined, but also its longevity. As any romantic can tell you, knowing how to start a relationship is nothing compared to keeping an existing one going. Hopefully, this model and simulation can illuminate the mysteries of dating and help us understand human interaction a little better.

References

Aron, A. (1988). The matching hypothesis reconsidered again: Comment on Kalick and Hamilton. *Journal of Personality and Social Psychology*, *54*, 441-446.

Berscheid, E., Dion, K., Walster, E., & Walster, G. (1971). Physical attractiveness and dating choice: A test of the matching hypothesis. *Journal of Experimental Social Psychology*, *19*, 78-92.

Bolton, G., & Ockenfels, A. (2000). ERC: A theory of equity, reciprocity, and competition. *American Economic Review*, *90*(1), 166–193.

Cash, T. (1981). Physical attractiveness: An annotated bibliography of theory and research in the behavioral sciences. *JSAS: Catalog of Selected Documents in Psychology*, *11*.

Cavior, N., & Boblett, P. (1972). Physical attractiveness of dating versus married couples. *Proceedings of the 80th annual conference of the American psychological association*, *7*, 175-176.

Citelli, J., & Waid, L. (1980). Physical attractiveness, romantic love, and equity restoration in dating relationships. *Journal of Personality Assessment*, *44*, 624-629.

Cunningham, M., & Wu, C. (1995). "Their ideas of beauty are, on the whole, the same as ours": Consistency and variability in the cross-cultural perception of female physical attractiveness. *Journal of Personality and Social Psychology*, *68*, 261-279.

Feingold, A. (1981). Testing equity as an explanation for romantic couples 'mismatched' on physical attractiveness. *Psychological Reports*, *49*, 247-250.

Feingold, A. (1988). Matching for attractiveness in romantic partners and same-sex friends: A meta-analysis and theoretical critique. *Psychological Bulletin*, *104*, 226-235.

Fleming, J., & Courtney, B. (1984). The dimensionality of self-esteem: II. Hierarchical facet model for revised measurement scales. *Journal of Personality and Social Psychology*, *46*, 404-421.

Hammond, R., & Axelrod, R. (2006). The evolution of ethnocentrism. *Journal of Conflict Resolution*, *50*, 926-936.

Helgeson, V. (1994). Long-distance romantic relationships: sex differences in adjustment and breakup. *Personality and Social Psychology Bulletin*, *20*, 254-265.

Hill, C., Rubin, Z., & Peplau, L. (1976). Breakups before marriage: the end of 103 affairs. *Journal of Social Issues*, *32*, 147-168.

Hönekopp, J., Rudolph, U., Beier, L., Liebert, A., & Müller, C. (2007). Physical attractiveness of face and body as indicators of physical fitness in men. *Evolution and Human Behavior*, *28*(2), 106–111.

Huston, T. (1973). Ambiguity of acceptance, social desirability, and dating choice. *Journal of Experimental Social Psychology*, *9*, 32-42.

Kalick, S., & Hamilton, T. (1986). The matching hypothesis reexamined. *Journal of Personality and Social Psychology*, *51*, 673-682.

Kaznatcheev, A. (2010). The cognitive cost of ethnocentrism. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd annual conference of the congnitive science society.* Austin, TX: Cognitive Science Society.

Leary, M. (2005). Sociometer theory and the pursuit of relational value: Getting to the root of self-esteem. *European Review of Social Psychology*, *16*, 75-111.

Lewin, K., Dembo, T., Festinger, L., & Sears, P. (1944). Level of aspiration. *Personality and the Behavior Disorders*, *1*, 33-378.

Miller, G., & Todd, P. (1998). Mate choice turns cognitive. *Trends in Cognitive Sciences*, *2*(5), 190–198.

Murstein, B. (1972a). Physical attractiveness and marital choice. *Journal of Personality and Social Psychology*, *22*, 8-12.

Murstein, B. (1972b). Physical attractiveness and marital choice. *Journal of Personality and Social Psychology*, *22*, 8-12.

Nowak, M., Page, K., & Sigmund, K. (2000). Fairness versus reason in the ultimatum game. *Science*, *289*(5485), 1773.

Pass, J. (2009). *The self in social rejection.* Unpublished doctoral dissertation, University of Groningen.

Pass, J., Lindenberg, S., & Oark, J. (2010). All you need is love: Is the sociometer especially sensitive to one's mating capacity? *European Journal of Social Psychology*, *40*, 221-234.

Price, R., & Vandenberg, S. (1979). Matching for physical attractiveness in married couples. *Personality and Social Psychology Bulletin*, *5*, 398-400.

Rand, C., & Hall, J. (1983). Sex differences in the accuracy of self-perceived attractivness. *Social Psychology Quarterly*, *46*, 359-363.

Shultz, T. R., Hartshorn, M., & Kaznatcheev, A. (2009). Why is ethnocentrism more common than humanitarianism? In N. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st annual conference of the congnitive science society* (p. 2100-2105). Austin, TX: Cognitive Science Society.

Singh, D. (1993). Adaptive significance of female physical attractiveness: Role of waist-to-hip ratio. *Journal*

of Personality and Social Psychology, *65*(2), 293–307.

Walster, E. (1970). The effect of self-esteem on liking for dates of various social desirabilities. *Journal of Experimental Social Psychology*, *6*, 248-253.

Walster, E., Aronson, V., Abrahams, D., & Rottman, L. (1966). Importance of physical attractiveness in dating behavior. *Journal of Personality and Social Psychology*, *4*, 508-516.

Walster, E., Hatfield, E., Walster, G., & Berscheid, E. (1978). *Equity: Theory and research.* Allyn & Bacon.

Wetzel, C., & Insko, C. (1982). The similarity-attraction relation: Is there an ideal one? *Journal of Experimental Social Psychology*, *18*, 253-276.

White, G. (1980). Physical attractiveness and courtship progress. *Journal of Personality and Social Psychology*, *39*, 660-668.

Determining the Internal Consistency of Attitude Attributions

Kyle E. Jennings (jennings@berkeley.edu)
Department of Psychology, University of California, Berkeley
Berkeley, CA 94720-1650 USA

Abstract

In the attitude attribution paradigm, observers must estimate the true attitude of an author who was assigned to advocate a particular position. Observers' tendency to attribute an attitude in line with the expressed position despite its having been assigned is called the correspondence bias. While there is strong evidence that such attributions are externally invalid, it is less clear whether they are internally consistent. This research develops a Bayesian model that specifies what attitude an observer should attribute, given assumptions about the prior attitude distribution, and perceptions of the degree of compliance shown in the essay and the strength of the situation. The model reproduces classical findings regarding chosen vs. assigned positions, prior attitude probability, and degree of compliance, and also fits newly collected data. The results suggest that future research should examine observers' assumptions and perceptions, and focus less on the reasoning process itself.

Keywords: Correspondence bias; Attitude attribution; Normative standard; Bayesian modeling.

People's tendency to neglect situational influences on behavior has been a subject of long-standing interest to social psychologists. Many of the earliest and most famous demonstrations of this error make use of the attitude attribution paradigm (Jones & Harris, 1967), wherein participants read an essay that expresses an opinion on an issue, and must estimate the author's attitude. Complicating this judgment is the fact that the author was assigned what position to express, which pits two competing explanations—holding the attitude, or complying with the request—against each other. Participants tend to make attitude attributions in line with the essay even when the position was assigned, which is called the correspondence bias (Gilbert & Jones, 1986).

It is not straightforward to say whether people's responses in the attitude attribution paradigm are in fact biased. On the one hand, when participants rate essays that other study participants wrote under constraint, the attributed attitudes are more in line with the essay than with the authors' self-reported attitudes (e.g., Reeder, Fletcher, & Furman, 1989). On the other hand, if people's attributions are internally consistent with their own perceptions and assumptions, it is hard to call their attributions completely biased (Jones, Worchel, Goethals, & Grumet, 1971; Morris & Larrick, 1995; Forsyth, 2004). These two views involve two different standards for correctness, known as correspondence and coherence (Hammond, 1996), which concern the external validity and internal consistency of the judgment, respectively. To avoid confusion between correspondence *criteria* and the correspondence *bias*, the terms external validity and internal consistency will be used. Though people's judgments probably lack external validity, it is not clear whether they are at least internally consistent.

Checking internal consistency requires knowing what information is relevant to a judgment, and how that information determines the correct answer. This paper develops a Bayesian model relating assumptions and perceptions to attitude attributions. Since the model is grounded in mathematics, the steps between premises and conclusions can be more readily verified than with verbally justified standards (Morris & Larrick, 1995). Additionally, the model is agnostic to what process people might use to make judgments, helping researchers advocating different mechanisms at least agree on the correct outcome.

Normative Model

In the attitude attribution paradigm, observers know what essay was written, and the circumstances under which it was written. Their judgment of whether the essay author holds the expressed attitude is (Morris & Larrick, 1995):

$$P(\text{attitude} \mid \text{essay}, \text{circumstances})$$

Letting A, E, and C stand for the attitude, essay, and circumstances, and strategically applying Bayes' rule,[1] this equals:

$$P(A) \cdot \frac{P(C \mid A)}{P(C)} \cdot \frac{P(E \mid A,C)}{P(E \mid C)}$$

Intuitively, these terms express the prior probability of the attitude, the co-occurance of the circumstances and the attitude, and the relative likelihood of a person writing the essay, comparing someone with the attitude to the average person.

The model can be applied in two ways. First, it can be interpreted schematically in order to draw conclusions about the general direction of normative inferences. For instance, the standard shows that the conventional wisdom that the essay communicates no information about the author's attitude in light of the circumstances is correct only if two conditions are met. First, the co-occurrance term must be one, meaning that positions must be assigned without respect to the author's attitude. Studies that merely say that the position to advocate was assigned leave open the possibility that the author's attitude was considered when making the assignment, in which case a correspondent inference may be justifiable. Second, the likelihood term must also be one, meaning that the constraint must be seen as equally compelling regardless of the author's attitude (with completely compelling being a special case of this). As other researchers have argued (e.g., Jones

[1] $P(A \mid E,C) = P(A,E,C)/P(E,C) = P(E \mid A,C)P(A,C)/P(E,C) = P(E \mid A,C)P(C \mid A)P(A)/[P(E \mid C)P(C)]$. See also Jennings (2010).

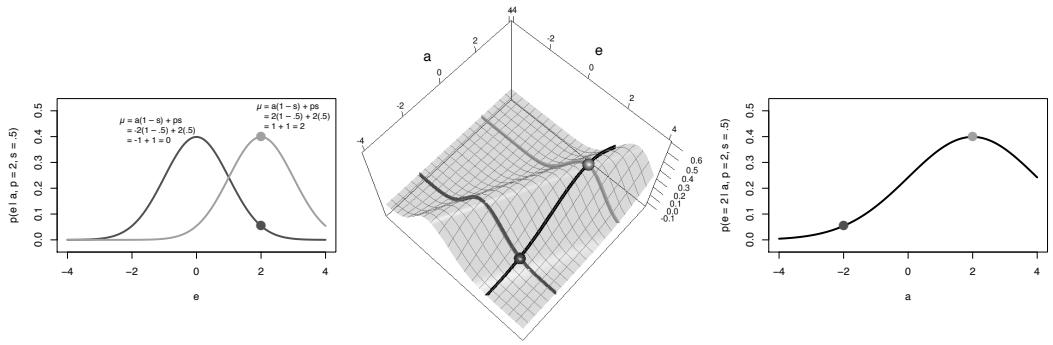


Figure 1: Illustration of $p(e \mid a, p = 2, s = .5)$. Left panel shows two essay distributions, for $a = -2$ and $a = 2$. Right panel shows likelihood distribution over attitudes for $e = 2$. As shown in the middle panel, these two distributions are really slices of the same three-dimensional function.

et al., 1971) or found (e.g., Forsyth, 2004), observers do not seem to hold this belief.

Though these conclusions are powerful, it is possible to do better. The second way to apply a model such as this one is to use it to quantitatively assess the internal consistency of people's judgments, which is done by measuring quantities on both sides of the equation. Previous authors have done this using alternative Bayesian standards (Trope, 1974; Morris & Larrick, 1995; Forsyth, 2004). However, every previous model has required participants to think in terms of discrete probabilities (e.g., the probability that the author holds the attitude expresed in the essay), while nearly all other studies of attitude attribution ask participants to estimate the author's attitude on a Likert-type scale. Achieving a match to what participants customarily estimate requires switching from the probabilities of dichotomous events to probability *densities* over continuous variables, as follows:

- "Pro" and "con" attitudes are generalized to real-valued attitudes along a "con" (negative) to "pro" (positive) continuum, with attitudes further from zero being more extreme. The variable a will refer to the author's attitude, while e will refer to the position expressed in the essay.

- The circumstances (C) are decomposed into two things: p, the position that the author was asked to express, and s, the strength of that request. The variable p can vary as discussed above, while s can vary between zero (no inducement) and one (a completely compelling inducement).

Using the above variables, $P(A \mid E, C)$ becomes $p(a \mid e, p, s)$. Converting the prior, co-occurance, and relative likelihood terms into probability distributions and multiplying the three over the range of a gives the probability of each possible attitude. The expected value of this distribution will be the attitude attribution, and the confidence in this attribution will be proportional to the distribution's standard deviation.

Completing the normative model requires specifying the forms of the three terms. The prior distribution, $p(a)$, is just the assumed attitude distribution in the population. The co-occurance term expresses how the circumstances vary with the author's attitude. Assuming a random assignment process, then this term equals one. This leaves the likelihood term, $p(e \mid a, p, s)/p(e \mid p, s)$. Since the denominator does not involve a, the expression can be written:

$$p(a \mid e, p, s) \propto p(a) \cdot p(e \mid a, p, s)$$

These terms will be called the posterior, the prior, and the (essay) likelihood, respectively.

The final task is to specify a form for the likelihood, $p(e \mid a, p, s)$. This can be done by determining the distribution of essay positions that an author with attitude a would write when asked to express position p, facing an inducement of strength s. Instead of requiring participants to estimate this themselves, the form of the function will be specified mathematically. Past research has found that observers expect constrained authors to express an attitude somewhere in between their own attitude and the position that was assigned (Miller & Rorer, 1982), which Reeder et al. (1989) refer to as the central tendency assumption. Thus, an author with (say) a strong con attitude who was asked to express a strong pro position would attempt to write a neutral essay. This expectation can be modeled by saying that when a, p, and s are known, $p(e \mid a, p, s)$ is a normal distribution, with:

$$\mu = a \cdot (1 - s) + p \cdot s$$

With no inducement ($s = 0$), $\mu = a$, the author's own attitude. With a completely compelling inducement ($s = 1$), $\mu = p$, the requested position. For other values of s, μ is a weighted compromise between a and p. While one could imagine ways that the distribution's standard deviation might depend on a, p, and s, for parsimony it will be assumed to be constant.

The above model of authors' responses specifies the distribution of e, given that the other variables are known. However, when applied, e is known but a is unknown. This does not present a problem, as illustrated in Figure 1. The left graph shows the essay distributions for two values of a (-2 and 2), where $s = .5$ and $p = 2$. The right graph shows the

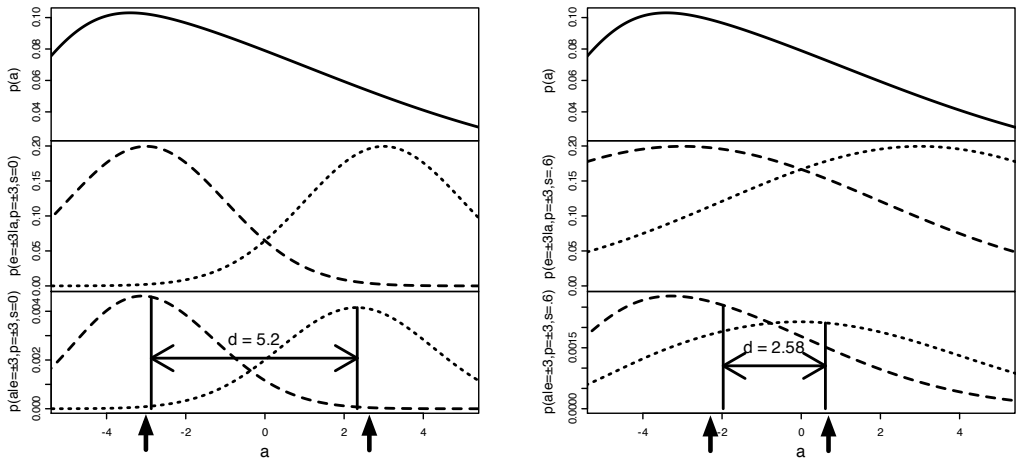


Figure 2: Model of the choice (left) and no choice (right) conditions for Jones and Harris (1967), Study 1. In the bottom panel, black arrows show original results, and black lines show the model's results.

likelihood distribution over the range of a, where $e = 2$. The middle image shows how the two are related, with two points shown on all three graphs ($e = 2$, $a = \pm 2$) for reference.

As already mentioned, this model improves upon previous Bayesian models of attitude attribution in that its output is the same kind of variable as participants usually estimate. In addition, the model's inputs correspond to perceptions that are relatively straightforward for participants to reason about (the position expressed in the essay, e, whether the essay is weaker or stronger than was expected of the author, $e - p$, and how constraining the situation was, s). When testing scenarios schematically, this makes it possible to continuously vary the model parameters, or to test specific combinations of parameters, rather than having to make verbal arguments about, say, the relative sizes of $P(\text{Essay} \mid \text{Attitude})$ and $P(\text{Essay} \mid \overline{\text{Attitude}})$. When testing the internal consistency of participants' actual judgments, it becomes possible to directly ask for the relevant quantities. The tradeoff is that the model assumes people believe that constrained authors will express a position between their own attitude and the request. However, other models of how authors respond could be translated into likelihood functions, and the results compared.

Illustrations

Choice and Prior Probabilities

Correspondent inference theory (Jones & Davis, 1965) was intended to be a normative standard for how people should make attributions, and aims to specify which behaviors justify the inference of information about a person that would not have been assumed previously (Jones & McGillis, 1976). Jones and Harris (1967) was an attempt to show that while both constrained behavior and expected behavior do not contribute new information, an expected behavior performed under constraint will still lead to a corresponding attribution, simply because the underlying disposition would be expected anyway. It is for this reason that they used advocacy for "Castro's Cuba"—a highly unexpected behavior in 1960's America—as the critical test. As predicted, they found that people made attributions corresponding to the constrained behavior when the behavior was expected (arguing against Castro). What they were surprised to find was that though people did not make completely corresponding attributions when the behavior was unexpected (arguing for Castro), their attributions did not revert to the level that would be obtained had the behavior been completely disregarded. This is the result that triggered the volumes of research on the correspondence bias that continues to this day.

This paper's model can reproduce the pattern of results that Jones and Harris obtained, using two reasonable assumptions. First, assume that the prior attitude distribution was strongly right skewed (i.e., very few people supporting Castro).[2] Second, for parsimony, assume that the pro and con essays were equivalently strong, and no weaker or stronger than requested.

To reproduce the "choice" condition, the model is run with strength set to zero ($s = 0$), which is illustrated in the left half of Figure 2. The top panel of this graph shows the prior distribution, $p(a)$, while the middle panel shows the likelihood functions for the con (dashed line) and pro (dotted line) essays. These lines show $p(e \mid a, p, s)$, where $e = \pm 3$,[3] $s = 0$ since authors could choose what to express (making the requested position, p, irrelevant), and a varies across the x-axis to encompass the range of attitudes shown. The bottom panel shows the posterior distributions, $p(a \mid e, p, s)$, which are the result of multiplying the prior distribution by either likelihood distribution. In this case, the prior distribution has only a small effect on the posterior distributions, and the expected

[2] See Jones and Harris (1967), p. 5.

[3] Note that attitude values are always rescaled to a -4 [con] to 4 [pro] for consistency of comparison across studies.

	Con		Pro	
	Strong	Weak	Weak	Strong
Requested position (p)	-2	-2	2	2
Essay position (e)	-3	-1	1	3
Attribution (a), weak constraint ($s = .25$)	-2.78	-0.56	0.56	2.78
Attribution (a), strong constraint ($s = .75$)	-2.16	0.72	-0.72	2.16
Essay position (e)	-3	-2	2	3
Attribution (a), strong constraint ($s = .75$)	-2.16	-0.72	0.72	2.16

Table 1: Model-predicted attributions for strong and weak essays under weak and strong situational constraint. Weak situation shows no reversal for the weak essays, but strong situation does. The same pattern can be obtained by keeping situation strength constant but making the weak essays less weak.

values of the distributions (shown by the black, vertical lines) are a very close match to the results that Jones and Harris originally obtained (shown by the black arrows).

Jones and Harris found nothing counterintuitive about their results for the choice condition, but were surprised by the results in the no choice condition, which can be replicated by choosing an appropriate value for s. Not shown in the figure is the case where the situation is seen as completely constraining ($s = 1$). Under conditions with no behavioral freedom, everyone is equally likely to have written the requested essay, and so the two likelihood functions are flat lines. As such, both posterior distributions are equal to the prior distribution, making the normative attribution for both essays equal to the mean attitude in the population. This result is what Jones and Harris were expecting to find. Since this is not what they obtained, values of s less than one must be tried.

A good fit to the original results was obtained with $s = .6$. The right half of Figure 2 shows this case, where it can be seen that though the prior distribution is the same and the likelihood functions have the same locations, the likelihood functions are also more spread out (since constrained behavior is less informative than freely chosen behavior). Even though the pro and con likelihood functions are equal and opposite, the posteriors are not, which is a result of multiplying by the asymmetric prior.

As the bottom panel shows, the expected values of either posterior (black, vertical lines) are quite close to the results that Jones and Harris obtained (black arrows). In particular, for the "con" essay, the model-derived and actual attributions are still in the direction of the essay. For the "pro" essay, however, multiplying by the prior probability has brought the model-derived results closer to the midpoint, and like the actual results, still somewhat correspondent with the essay itself. It is also worth noting that the "pro" posterior is more spread out than the "con" posterior, just as Jones and Harris found greater variance in this condition than in the other conditions of their study.

As the above shows, the model can reproduce the important features of the original demonstration of the correspondence bias, with only one parameter varying between the choice and no choice conditions. As such, it establishes that the results in Study 1 of Jones and Harris (1967) could be the result of an internally consistent reasoning process, given the assumption that the participants did not believe that the author's situation in the no choice condition was completely constraining. In fact, according to the model, the *only* internally consistent way for perceivers to make attributions other than to the mean attitude in the population is if they believe that the situation leaves room for choice, and that this choice depends on the compatibility of the author's attitude and the assigned position. Though these attributions are probably externally invalid, the possibility that they are internally consistent suggests that defects in observers' reasoning processes are not necessary to explain the results. Likewise, people may make perfectly reasonable assumptions about how situational constraints in general would influence essay authors. The source of "bias" may simply be that people applied those assumptions using an insufficiently strong appraisal of the power of the author's particular situation.

Degree of Compliance

Thus far, it has been assumed that perceivers believe that the essay written was no weaker or stronger than was requested. However, compliance needn't be all-or-nothing. Jones et al. (1971) manipulate the strength of the essay in order to understand how behavioral extremity affects attributions. One of their key results is that when people read an essay written under constraint and expressing a weak position, they attribute the *opposite* attitude to the author as was assigned. When the essay position was strongly argued, they attribute a corresponding attitude. In an attempt to replicate this result, Miller (1974) found that people made less extreme attributions when reading a weak essay than when reading a strong essay, but did not find any reversal. In both cases, however, the degree of compliance affected the attributions.

The model is able to reproduce these result patterns by varying the situation strength parameter, s, and leaving everything else constant. Model-predicted attitude attributions for weak and strong levels of constraint are shown in the top and middle of Table 1. Reversal occurs for the strong constraint, but not for weak constraint. Intuitively, this is because stronger constraints make it less likely that a person

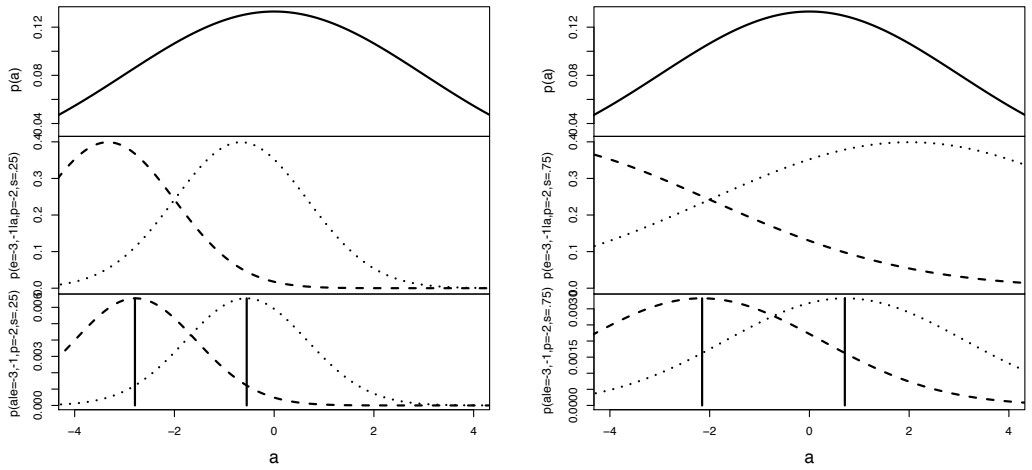


Figure 3: Illustration of strong and weak "con" essays for weak constraint (left, $s = .25$) and strong constraint (right, $s = .75$). In both cases, the requested position, p, is -2 and the essay positions, e, are -3 and -1 for the strong and weak essays.

would deviate from the requested position. Therefore, when someone does deviate from the requested position by writing a weaker-than-expected essay, it is reasonable to conclude that this person must hold an attitude very different than what was requested. This is illustrated in Figure 3. As can be seen, the likelihood functions are both further from the requested positions and more spread out at higher constraint.

Varying situation strength is not the only way to replicate the different patterns of results. The bottom two rows of Table 1 show what happens when the weak essays are made less ambivalent, but the strong level of constraint kept consistent. This change removes the reversal that had been obtained with the weak essays. In speculating on the failure to replicate the Jones et al. (1971) weak essay reversal, Miller (1974) does in fact note that his weak essays were not as weak as those in Jones et al. While both Jones et al. and Miller speculate that strong essays "engulf the field" whereas the weak essays allow the perceiver more latitude to notice the situation, the model suggests that no such perceptual metaphors are necessary. Instead, both outcomes are reasonable conclusions of an internally consistent logic that does not depend on any distortions in perception, failure to notice the situation, or alteration in the underlying behavioral model being used to make the attribution. This example also makes clear that there are often multiple internally-consistent ways to obtain the same pattern of results. The model makes it possible to explore many sources of a result, thereby suggesting hypotheses for behavioral research.

Empirical Results

In addition to fitting previous research results, the model fits new data (collected for a different purpose).[4] Participants ($N = 246$) read essays for and against the death penalty, and then learned that the author was randomly assigned the position to take. Participants then rated what they thought the author's attitude was, how confident they were in their answer, and other perceptions (detailed next). Replicating past results, there was a significant difference between the pro and con essay attributions ($M = -0.90$ vs. $M = 0.98$, $t(244) = -8.44$, $p < .0001$). Model-based predictions were then tested, after reversing all of the relevant quantities for participants in the "con" essay condition.

As the model of the Jones and Harris (1967) results showed, a skewed prior attitude distribution should result in asymmetric attitude attributions. In particular, attributions for essays expressing rare opinions should be closer to the midpoint that attributions for essays expressing common positions. Additionally, as judged by the variance of the posterior distributions, people should be less confident in their attributions when the expressed position is rare. This was tested by looking at participants' self-reported prior attitude distributions, which were elicited by having people apportion 100 percentage points to three equal-sized intervals encompassing the measurement scale. A "skew" was calculated for each participant by taking the log ratio of the lower and upper intervals of their priors. Negative ratios imply more probability mass near the "pro" end of the scale, and positive ratios imply more probability mass near the "con" end of the scale. Supporting the model's predictions, the correlation of attribution and skew was $r = -.14$ ($p < .05$), and the correlation of confidence and skew was $r = -.20$ ($p < .01$).

Next, the co-occurrance between the situation and attitudes was examined. As mentioned at the outset, if assignment is non-random, people might believe that the essay author's own attitude and the assigned position are related. To test

[4] These data are part of an in-progress replication of Miyamoto and Kitayama (2002), which uses their essays as stimuli. In addition to having "pro" and "con" essays, the study varies essay length. Since the effects listed next are not qualified by length, the length manipulation is not discussed further.

this, people were compared by whether they indicated (as intended) that the author had no control over assignment. There was a significant difference ($M = 0.73$ vs. $M = 1.21$, for no control vs. control, respectively, $t(244) = 2.16$, $p < .05$).

Finally, the likelihood model predictions were examined. As shown with the modeling of the Jones et al. (1971) and Miller (1974) result patterns, the model predicts that overcompliance and attribution extremity should be positively related, and that perceived situation strength and attribution extremity should be negatively related. Participants estimated overcompliance via a question asking how much weaker (or stronger) the essay was than what they believed was expected, and strength was measured via a question about how much overall choice the author had (reversed). After partialing out the effects of skew and strength, attribution and overcompliance were positively related, $pr = .14$ ($p < .05$). After partialing out skew and overcompliance, attribution and strength were negatively related $pr = -.16$ ($p < .05$). Because higher strengths lead to more spread out likelihood functions, the model also predicts that confidence and strength should be negatively related, which was supported $r = -.24$ ($p < .001$).

Though these results do not prove that people's attributions are internally consistent, they do demonstrate promise. Future work will systematically test the match between model predictions and empirical results in greater detail.

Conclusions

Using a simple yet plausible model of how people respond to instructions to advocate a particular opinion, this work derives a model that can postdict prior attitude attribution results, and that fits newly-collected data. Though the correspondence bias can be seen when people's attributions are compared to the ground truth, this work suggests that these attributions could be internally consistent with other beliefs and perceptions that people have. Future work should investigate why these beliefs (e.g., about how people respond to requests) and perceptions (e.g., of the request strength or the essay extremity) don't match reality. The likelihood model could also be extended to encompass other essay features, such as argument quality (cf. Miller & Rorer, 1982; Gawronski, 2003).

Early in the history of correspondence bias research, Jones et al. (1971) conceded that correspondent inferences for constrained behavior are only wrong if every person in that situation would comply. Short of this extreme, they say that "it would be very difficult if not impossible to determine whether [a correspondent inference] should be judged as attributional distortion" (p. 77). The model presented here helps answer this question by encoding a set of assumptions mathematically, and then using the logic of Bayes' rule to understand the implications of those assumptions. It is likely that many attitude attribution findings can fruitfully be reexamined in light of the added precision that this model provides.

Acknowledgments

Tom Griffiths, Rob MacCoun, and Kaiping Peng provided valuable comments on and assistance with this work.

References

Forsyth, D. R. (2004). Inferences about actions performed in constraining contexts: Correspondence bias or correspondent inference? *Current Psychology*, *23*(1), 41–51.

Gawronski, B. (2003). Implicational schemata and the correspondence bias: On the diagnostic value of situationally constrained behavior. *Journal of Personality and Social Psychology*, *84*(6), 1154–1171.

Gilbert, D. T., & Jones, E. E. (1986). Perceiver-induced constraints: Interpretation of self-generated reality. *Journal of Personality and Social Psychology*, *50*, 269–280.

Hammond, K. R. (1996). *Human judgment and social policy: Irreducible uncertainty, inevitable error, un-available injustice*. New York: Oxford University Press.

Jennings, K. E. (2010). *Coherent attributions with co-occurring and interacting causes*. Unpublished doctoral dissertation, University of California, Berkeley.

Jones, E. E., & Davis, K. E. (1965). From acts to dispositions: The attribution process in person perception. In *Advances in experimental social psychology* (Vol. 2, pp. 219–266). New York: Academic Press.

Jones, E. E., & Harris, V. A. (1967). The attribution of attitudes. *Journal of Experimental Social Psychology*, *3*, 1–24.

Jones, E. E., & McGillis, D. (1976). Correspondent inferences and the attribution cube: A comparative reappraisal. In J. H. Harvey, W. J. Ickes, & R. F. Kidd (Eds.), *New directions in attribution research* (Vol. 1, pp. 389–420). Hillsdale, NJ: Erlbaum.

Jones, E. E., Worchel, S., Goethals, G. R., & Grumet, J. F. (1971). Prior expectancy and behavioral extremity as determinants of attitude attribution. *Journal of Experimental Social Psychology*, *7*, 59–80.

Miller, A. G. (1974). Perceived freedom and the attribution of attitudes. *Representative Research in Social Psychology*, *5*, 61–80.

Miller, A. G., & Rorer, L. G. (1982). Toward an understanding of the fundamental attribution error: Essay diagnosticity in the attitude attribution paradigm. *Journal of Research in Personality*, *16*, 41–59.

Miyamoto, Y., & Kitayama, S. (2002). Cultural variation in correspondence bias: The critical role of attitude diagnosticity of socially constrained behavior. *Journal of Personality and Social Psychology*, *83*(5), 1239–1248.

Morris, M. W., & Larrick, R. P. (1995). When one cause casts doubt on another: A normative analysis of discounting in causal attribution. *Psychological Review*, *102*(2), 331–355.

Reeder, G. D., Fletcher, G. J. Q., & Furman, K. (1989). The role of observers' expectations in attitude attribution. *Journal of Experimental Social Psychology*, *25*, 168–188.

Trope, Y. (1974). Inferential processes in the forced compliance situation: A Bayesian analysis. *Journal of Experimental Social Psychology*, *10*, 1–16.

Cohesion, Coherence, and Expert Evaluations of Writing Proficiency

Scott A. Crossley (sc544@msstate.edu)
Department of English, Mississippi State University
MS, 39762 USA

Danielle S. McNamara (dsmcnamara1@gmail.com)
Department of Psychology, Institute for Intelligent Systems, The University of Memphis
Memphis TN 38152 USA

Abstract

This study investigates the roles of cohesion and coherence in evaluations of essay quality. Cohesion generally has a facilitative effect on text comprehension and is assumed to be related to essay coherence. By contrast, recent studies of essay writing have demonstrated that computational indices of cohesion are not predictive of evaluations of writing quality. This study investigates expert ratings of individual text features, including coherence, in order to examine their relation to evaluations of holistic essay quality. The results suggest that coherence is an important attribute of overall essay quality, but that expert raters evaluate coherence based on the absence of cohesive cues in the essays rather than their presence. This finding has important implications for text understanding and the role of coherence in writing quality.

Keywords: Coherence; Writing Quality; Cohesion, Linguistics, Computational Algorithms, Models.

Introduction

Writing affords the opportunity to thoroughly articulate ideas and synthesize a variety of perspectives allowing for persuasive communication that transcends both time and space (Crowhurst, 1990). As such, the ability to convey meaning proficiently in written texts is a critical skill for academic and professional success. Indeed, college freshmen' writing skills are among the best predictors of academic success (Geiser & Studley, 2001), and even outside of academia, writing skills continue to be important and are an important attribute of professional competence (Light 2001). As such, developing a better understanding of good and poor writing is an important objective, both for theoretical and applied reasons.

The overarching objective of this study is on the identification of essay features that are predictive of overall writing quality. Our goal is to better understand and model writing proficiency. We are particularly interested in the roles that cohesion and coherence play in writing quality. Cohesion refers to the presence or absence of explicit cues in the text that allow the reader to make connections between the ideas in the text. For example, overlapping words and concepts between sentences indicate that the same ideas are being referred to across sentences. Likewise, connectives such as *because*, *therefore*, and *consequently*, inform the reader that there are relationships between ideas and the nature of those relationships. Whereas cohesion refers to the explicit cues in the text, *coherence* refers to the understanding that the reader derives from the text, which may be more or less coherent depending on a number of factors, such as prior knowledge and reading skill (McNamara, Kintsch, Songer, & Kintsch, 1996; O'Reilly & McNamara, 2007).

There is a strongly held sense that essay quality is highly related to the cohesion and coherence of the essay. This is reflected in the literature about writing (e.g., Collins, 1998; DeVillez, 2003), as well as textbooks that teach students how to write (Golightly & Sanders, 1990). However, there are few studies that have empirically investigated the role of cohesion cues and by consequence, coherence, in essays. Whereas there is a strong assumption that coherence is an important aspect of writing, few studies have documented this assumption or tied the notion of coherence to explicit linguistic features of the essay. Indeed, our own examinations of linguistic features of good and poor essays have turned up no evidence that cohesion cues are positively related to essay quality for either first language writers (McNamara, Crossley, & McCarthy, 2010) or writers for whom English is their second language (Crossley & McNamara, in press). Therefore, the question of whether coherence or cohesion play important roles in essay writing and judgments of essay quality remains open.

In contrast, the role of cohesion in text comprehension is much better understood and there are numerous empirical studies on the topic (for a recent review, see McNamara, Louwerse, McCarthy, & Graesser, 2010). These studies show that increasing the cohesion of a text facilitates and improves text comprehension for many readers (Gernsbacher, 1990) and is particularly crucial for low-knowledge readers (McNamara et al., 1996).

From this literature on text comprehension, we glean two competing hypotheses for the effects of cohesion on estimates of essay quality (i.e., the coherence of the essay in the mind of the essay rater). On the one hand, cohesion underlies coherence, and thus should be important. On the other hand, the effects of cohesion on comprehension depend on the knowledge and reading skill of the reader. Indeed, a reverse cohesion effect, or an advantage for low cohesion text, can occur for high knowledge readers (McNamara, 2001; McNamara et al., 1996; O'Reilly & McNamara, 2007). High-knowledge readers, unlike low-knowledge readers, can successfully make the inferences needed to bridge the conceptual gaps that are in low-cohesion text. In fact, high-knowledge readers may benefit from low cohesion texts because gaps in cohesion force the reader to make connections in text that are not explicitly

available (McNamara, 2001; O'Reilly & McNamara, 2007). Hence, when the material covered in a text is familiar to the reader (as is often the case for narratives), cohesion cues may be unnecessary, and perhaps even distracting. Overall, text comprehension literature leads to the conclusion that cohesion may play an important role in facilitating coherence if the rater of the essay has less knowledge about the topic, but cohesion cues may be inversely related to essay scores if the rater has more knowledge about the topic.

We recently explored this topic by examining the effects of cohesion devices on human evaluations of writing quality. McNamara et al (2010) used linguistic indices of cohesion and language sophistication provided by the computational tool Coh-Metrix (Graesser, McNamara, Louwerse, & Cai, 2004) to analyze a corpus of 120 argumentative essays written by college undergraduate and scored by expert raters using a holistic rubric. The essays were scored on a 1-6 scaled SAT rubric and then categorized into two groups: essays judged as low versus high quality. The results indicated that there were no differences between these two groups according to indices of cohesion (e.g., word overlap, causality, connectives). By contrast, indices related to language sophistication (lexical diversity, word frequency, and syntactic complexity) showed significant differences between the groups. A follow-up discriminant function analysis (DFA) showed that these indices successfully classified the essays into their respective groups at a level well above chance. The results of the McNamara et al. study provide initial indications that text cohesion may not be indicative of essay quality. Instead, expert raters in the McNamara et al. study judged essays as higher quality when they were more difficult to process (less familiar words, more complex syntax).

While McNamara et al. (2010) showed that cohesion cues were not related to the overall scores assigned by essay raters, it did not investigate the role of the raters' judgments of the coherence or cohesion of the essay, nor did it investigate whether cohesion cues are related to raters' judgments of coherence and cohesion. Hence the purpose of the current study is two-fold. First, we examine the assumption that judgments of essay coherence are predictive of the overall score for an essay. While this is a commonly held belief, we are aware of no empirical support for this assumption provided in the literature. Second, we examine whether cohesion cues as measured by Coh-Metrix are related to raters' estimates of an essay's coherence. Whereas McNamara et al. (2010) did not find a relation between indices of cohesion and the overall essay scores, it remains an open question as to whether cohesion indices might be related to more direct ratings of an essay's coherence.

Method

Our method of inquiry involves an analysis of argumentative essays by expert scorers on atomistic features of essay quality (i.e., introductions, thesis statement, topic sentences, relevance, coherence) as well as a holistic evaluation of essay quality. Thus, unlike McNamara et al. (2010), we do not rely solely on computational indices to model overall essay quality, but instead concentrate on the evaluation of human judgments of individual text features in relation to overall text quality. Included in the individual text features evaluated by human experts are two measures of coherence. If the ratings of coherence are predictive of overall essay quality, we will also use computational indices of cohesion to model these human ratings. We can, thus, examine the importance of cohesion and coherence in writing quality and examine which cohesive devises may be predictive of human ratings of coherence Such an analysis will also afford the opportunity to examine whether indices of cohesion correlate with human ratings of coherence, providing us with an opportunity to gain a better understanding of the role cohesion plays in high-knowledge readers (i.e., the expert raters in our study).

Corpus

As in McNamara et al. (2010), our analyses were conducted using a corpus of essays collected from undergraduate students at Mississippi State University (MSU). The MSU corpus was designed to account for learner variables such as age (adult students) and learning context (freshman college composition class). The corpus was also designed to consider task variables such as medium (writing), first language (English), genre (argumentative essays), essay length (between 500 and 1,000 words), and topics (3 prompts on equality, television, and creativity). The final corpus consisted of 184 essays. The essays were untimed and written outside of the classroom. Thus, referencing of outside sources was allowed, but was not required. Students were allowed to select the essay prompt. Therefore, there are an unequal number of essays per prompt. Although 100 of the essays used in our current analysis were also used in the McNamara et al. study, these 100 essays were evaluated by different raters in the current study. The raters used both an atomistic and holistic survey instrument.

Rating Rubric

The essay-rating rubric used in this analysis was designed to parallel the rubric used initially by Breetvelt, van den Bergh, and Rijlaarsdam (1994) and later adapted with a focus on structure and argumentation by Sanders and Schilperoord (2006). Three experts in language processing with Ph.D.s in either linguistics or cognitive psychology developed the rubric. It was then subjected to usability tests by expert raters with at least three years experience in essay scoring. The final version of the survey instrument has three subsections: structure, content, and conclusion. The structure subsection contains questions related to essay structure and continuity. The content subsection contains questions related to the introduction, thesis, coherence, topic and evidential sentences, relevance, register use, and mechanics. The conclusion subsection contained questions related to the conclusion type, conclusion summary, and closing. In addition, the survey instrument included a holistic grading scale based on a standardized rubric

commonly used in assessing Scholastic Achievement Test (SAT) essays. This holistic scale was the same scale used by McNamara and colleagues (2010). The holistic scale and all of the rubric items had a minimum score of 1 and a maximum score of 6. The atomistic rubric ratings included the following:

Structure: Clarity of division into introductions, argumentation, and conclusion.
Continuity: Strength of connection of ideas and themes within and between the essays' paragraphs (cohesion).
Introduction: Presence of a clear, introductory sentence.
Thesis Statement: Strength of the thesis statement and its attached arguments.
Reader Orientation: Overall coherence and ease of understanding.
Topic Sentences: Presence of identifiable topic sentences in argumentative paragraphs.
Evidential Sentences: Use of evidential sentences in the argumentative paragraphs that support the topic sentence or paragraph purpose.
Relevance: Degree to which argumentation in the paper contained only relevant information.
Appropriate Registers: Degree to which the vocabulary in the essays followed the expected register.
Grammar, Spelling, and Punctuation: Accuracy of grammar, spelling, and punctuation.
Conclusion: Clarity of the conclusion.
Conclusion Type*:* Identifiable conclusion type.
Conclusion Summary: Presence of summary within the conclusion including arguments and the thesis of the essay.
Closing. Clarity of closing statements within the essay.

Essay Evaluation

Two expert raters with master's degrees in English and at least 3 years experience teaching composition classes at a large university rated the 184 essays from the corpus using the rubric. The raters were informed that the distance between each score was equal. Accordingly, a score of 5 is as far above a score of 4 as a score of 2 is above a score of 1. The raters were first trained to use the rubric with 20 essays. A Pearson correlation for each rubric evaluation was conducted between the raters' responses. If the correlations between the raters did not exceed $r = .50$ (which was significant at $p < .05$) on all items, the ratings were reexamined until scores reached the $r = .50$ threshold. Raters followed similar protocol for the holistic score, but were expected to reach an $r >= .70$.

After the raters had reached an inter-rater reliability of at least $r = .50$ ($r = .70$ for the holistic score), each rater then evaluated the 184 essays that comprise the corpus used in this study. Once final ratings were collected, differences between the raters were calculated. If the difference in ratings on survey feature were less than 2, an average score was computed. If the difference was greater than 2, a third expert rater adjudicated the final rating. Correlations between the raters (before adjudication) are located in Table 1. The raters had the lowest correlations for judgments of continuity and the highest correlations for essay structure.

Table 1: Pearson Correlations between Raters

Item	*r*
Structure	0.647
Continuity	0.307
Introduction	0.330
Thesis Statement	0.513
Reader Orientation	0.367
Topic Sentences	0.510
Evidential Sentences	0.404
Relevance	0.306
Appropriate Registers	0.394
Grammar, Spelling, Punctuation	0.599
Conclusion	0.596
Conclusion Type	0.355
Conclusion Summary	0.525
Closing	0.445
Holistic Score	0.533

Results

We used a multiple regression analysis to examine the predictive strength of the atomistic writing features in explaining the scoring variance in the holistic scores assigned to the essays. We used a training set to generate a model to examine the amount of variance explained by each writing feature. The model was then applied to a test set to calculate the accuracy of the analysis. Accordingly, we randomly divided the corpus into two sets: a training set ($n = 123$) and a test set ($n = 61$). The training set was used to identify which of the atomistic features most highly correlated with the holistic scores assigned to the essays. These features were later used to predict the holistic scores in the training and test sets using the generated model.

We controlled the number of variables included in the regression analysis in order to reduce the likelihood that the model was over-fitted. If too many variables are used, the model fits not only the signal of the predictors, but also the unwanted noise. The model may, thus, lack accuracy when applied to a new data set. We selected a ratio of 15 observations to 1 predictor, which is standard for analyses of this kind (Field, 2005). Given that the training set contained 123 essays, we determined that we could include eight features in our regression analysis.

Pearson Correlations

All features on the rubric correlated significantly with the holistic scores assigned to the essays in the training set. The strongest correlations were for Reader Orientation (coherence), Relevance, and Continuity (cohesion). The weakest correlations were for Thesis, Conclusion, and Introduction. All the features along with their *r* values are presented in Table 2 (all $p < .001$).

Table 2: Pearson Correlations Atomistic to Holistic Scores

Variable	*r* value
Reader Orientation	0.803
Relevance	0.710
Continuity	0.650
Conclusion Type	0.640
Structure	0.633
Evidential Sentences	0.629
Grammar, Spelling, & Punctuation	0.590
Appropriate Registers	0.589
Topic Sentences	0.583
Closing	0.578
Conclusion Summary	0.551
Thesis Statement	0.548
Conclusion	0.526
Introduction	0.389

Collinearity

The features Structure and Conclusion were both highly correlated (> .70) with the feature Conclusion Type. Because both of these features had lower correlations with the holistic score as compared to Conclusion Type, the Structure and Conclusion variables were dropped from the multiple regression analysis. Thus only the variables Reader Orientation, Relevance, Continuity, Conclusion Type, Evidential Sentences, Grammar, Spelling, & Punctuation, Appropriate Registers, and Topic Sentences were included in the regression.

Multiple Regression Training Set

A linear regression analysis (stepwise) was conducted including the eight variables. These eight variables were regressed onto the raters' holistic evaluations for the 123 writing samples in the training set. The variables were checked for outliers and multicollinearity. Coefficients were checked for both variance inflation factors (VIF) values and tolerance. All VIF values were at about 1 and all tolerance levels were well beyond the .2 threshold, indicating that the model data did not suffer from multicollinearity (Field, 2005).

Five variables were significant predictors in the regression: Reader Orientation (t = 6.668, p < .001) Conclusion Types (t = 5.068, p < .001), Evidential Sentences (t = 3.495, p < .001), Topic Sentences (t = 3.180, p < .010), and Appropriate Registers (t = -1.419, p < .050). Three variables were not significant predictors: Relevance (t = 1.841, p > .050), Continuity (t = 1.760, p > .050), and Grammar, Spelling, & Punctuation (t = 1.486, p > .050). The latter variables were left out of the subsequent analysis. The linear regression using the eight variables yielded a significant model, $F(5, 117) = 89.693$, $p < .001$, $r = .891$, $r^2 = .793$, demonstrating that the combination of the five variables accounts for 79% of the variance in the human evaluations essay quality for the 123 essays examined in the training set. All the features retained in the regression analysis along with their r values, r^2 values, unstandardized Beta weights, standardized Beta weights, and standard errors are presented in Table 3.

Test Set Model

To further support the results from the multiple regression conducted on the training set, we used the B weights and the constant from the training set multiple regression analysis to estimate how well the model would function on an independent data set (the 61 essays and their holistic scores held back in the test set). The model produced an estimated value for each writing sample in the test set. We used this correlation along with its r^2 to demonstrate the strength of the model on an independent data set. The model for the test set yielded $r = .922$, $r^2 = .850$. The results from the test set model demonstrate that the combination of the five variables accounted for 85% of the variance in the evaluation of the 61 essays comprising the test set.

Linguistic Features Analysis

Our regression analysis demonstrated that text coherence is an important predictor of human judgments of essay quality. Our subsequent goal was to identify which linguistic features are attributable to the coherence construct used by the human raters.

Table 3: Linear Regression Analysis to Predict Essay Ratings Training Set

Entry	Variable Added	R	R^2	*B*	B	SE
Entry 1	Reader Orientation	0.803	0.645	0.458	0.413	0.069
Entry 2	Conclusion Type	0.850	0.723	0.296	0.257	0.058
Entry 3	Evidential Sentences	0.871	0.758	0.271	0.182	0.078
Entry 4	Topic Sentences	0.882	0.778	0.222	0.160	0.070
Entry 5	Registers	0.891	0.793	0.201	0.152	0.069

Notes: Estimated Constant Term is 23.79; *B* is unstandardized Beta; B is standardized Beta; SE is standard error

To accomplish this goal, we conducted an analysis of the Reader Orientation scores using computational indices provided by Coh-Metrix that have theoretical correlates with cohesion features. Our goal in this second analysis is to examine if computational indices related to cohesion can successfully model the human coherence ratings from our essay analysis. We used the same corpus as the principle study, but concentrated solely on the human ratings for the Reader Orientation item (i.e., the coherence feature that was predictive of overall essay quality).

We selected a range of measures related to cohesion from the Coh-Metrix tool. The constructs measured included semantic coreference (LSA indices), causal cohesion, spatial cohesion, temporal cohesion, connectives and logical operators, anaphoric resolution, word overlap, and lexical diversity (see Crossley & McNamara, 2009; Graesser et al., 2004, for an overview of the cohesion indices in Coh-Metrix). Each construct was measured using multiple Coh-Metrix indices.

We first divided the corpus into a training (N = 123) and test set (N= 61). We then conducted Pearson correlations to relationships between the Coh-Metrix Indices and the human ratings of coherence.

Pearson Correlations. Among the selected cohesion constructs, only a few reported multiple indices that demonstrated significant correlations with the human ratings of coherence. The constructs that reported multiple significant indices included anaphoric reference (i.e., the proportion of anaphoric references between sentences), causal cohesion (i.e., the incidence of causal verbs and particles), incidence of connectives (i.e., positive temporal connectives, subordinating conjunctions, causative subordinators), and overlap measures (the overlap nouns, stems, and arguments between sentences). However, these correlations were negative (with the exception of Subordinating Conjunctions; i.e. *until, though, since*). Measures for semantic coreference, logical operators, lexical diversity, spatial cohesion, and temporal cohesion did not report significant indices. The indices with the highest correlations from the significant measures are presented in Table 3 along with their *r* and *p* values. The negative correlations indicate that the essays rated high in coherence included fewer cohesion cues.

Table 4: Correlations Coh-Metrix Indices to Raters' Coherence Scores

Variable	r value	*p* value
Anaphoric reference	-0.349	< .001
Ratio of causal particles and verbs	-0.259	< .010
Incidence of positive temporal connectives	-0.237	< .010
Subordinating conjunctions	0.240	< .010
Causative subordinators	-0.211	< .050
Content word overlap	-0.187	< .050

Discussion

This study has demonstrated that human ratings of coherence are an important indicator of holistic evaluations of essay proficiency. However, how human raters construct a coherent mental representation of a text seems opposed to many intuitive notions of coherence. For instance, we might expect that cohesive devices such as word overlap, causal particles and verbs, resolved anaphors, and positive temporal connectives would help the rater to develop a more coherent textual representation. However, in the case of the expert raters used in this study, the opposite was true. The absence of cohesive devices was associated with a more coherent mental representation of the text.

Our results indicate that coherence is an important element of human judgments of essay quality. In fact, overall text coherence is the most predictive feature of holistic essay scores. The coherence of a text (and by extension its understandability) was more predictive of writing quality than conclusion types, the use of evidential sentences, the use of topic sentences, and the use of appropriate registers. The overall coherence of a text was also the primary predictor of essay quality and explained 65% of the variance in the human ratings of writing quality. Human ratings of cohesion (continuity), although not retained in our regression analysis, also significantly correlated with essay quality.

However, our analysis using cohesion indices provided by Coh-Metrix demonstrated that our human judgments of coherence were not positively related to indices related to text cohesion indicating that cohesive devices may not underlie the development of coherent textual representations. Indeed, the majority of cohesive devices negatively correlated with human judgments of coherence. The exception is the use of subordinating conjunctions, which were positively correlated with human ratings of coherence. Yet, subordinating conjunctions also play a syntactic role and, by their nature, create more complex syntactic structures that result in a greater number of words before the main verb. Thus, it is likely that the subordinating conjunction index is actually detecting syntactic complexity, which does positively correlate with estimates of essay quality (McNamara et al., 2010).

So the question becomes: What factors are informing expert raters' mental representations of the text? One conclusion that the results of this study support is that factors important in text comprehension may have similarly important roles when raters evaluate the quality of essays. Specifically, the background knowledge of expert raters may influence text coherence in assessments of essay quality. Expert essay raters tend to be highly educated with advanced degrees and with experience in grading essays and other types of writing. The prompts used in the current study as well as prompts commonly used in essay writing assessments generally deal with topics that are relatively familiar to most educated individuals. As such, we can assume that essay raters will not tend to be low knowledge readers. Low knowledge readers lack sufficient knowledge

to generate inferences to bridge conceptual gaps in text, and, as a result, they tend to benefit from explicit text cohesion (i.e., word overlap, resolved anaphors, causal cohesion, connectives). By contrast, high knowledge readers benefit from texts low in cohesion because the cohesion gaps in the texts induce them to generate appropriate inferences to fill in the conceptual gaps. High knowledge readers can do this successfully because they have sufficient background knowledge to make appropriate inferences. When successful inferences are generated, the coherence of the mental representation can increase due to connections between the new information and their prior knowledge (McNamara, 2001; McNamara & McDaniel, 2004; O'Reilly & McNamara, 2007). Thus, more cohesive devices in essays may produce a less coherent mental representation in expert raters.

Conclusion

We conclude that coherence is an important attribute of writing quality. Essay raters' evaluations of coherence were highly related to their overall holistic scores for the essays. Nonetheless, we have found here that coherence is not necessarily defined through the use cohesion devices, and in fact may be inversely related to the presence of cohesion cues. Thus, the question becomes: What textual features of an essay lead to higher versus lower estimates of essay coherence? Our results demonstrate that the indices currently available from which to measure cohesion are not strongly linked to human judgments of coherence. However, it is highly unlikely that textual features do not affect coherence. Thus, our task becomes the identification of these features and the derivation of computational algorithms that accurately model them.

Acknowledgments

This research was supported in part by the Institute for Education Sciences (IES R305A080589 and IES R305G20018-02). Ideas expressed in this material are those of the authors and do not necessarily reflect the views of the IES. The authors would like to thank Brad Campbell and Daniel White for their help in scoring the corpus of essays.

References

Breetvelt, I., Van den Bergh, H., & Rijlaarsdam, G. (1994). Relations between writing processes and text quality: when and how? *Cognition and Instruction, 12*(2), 103-123.

Collins, J.L. (1998). *Strategies for struggling writers*. New York, NY: The Guilford Press.

Crossley, S.A. & McNamara, D.S. (2009). Computationally assessing lexical differences in L1 and L2 writing. *Journal of Second Language Writing, 18,* 119-135.

Crossley, S. A., & McNamara, D. S. (in press). Predicting second language writing proficiency: The role of cohesion, readability, and lexical difficulty. *Journal of Research in Reading*

Crowhurst, M. (1990). Reading/writing relationships: An intervention study. *Canadian Journal of Education, 15*, 155-172.

DeVillez, R. (2003). *Writing: Step by step*. Dubuque, IO: Kendall Hunt.

Field, A. (2005). *Discovering statistics using SPSS*. London, English: Sage Publications.

Gernsbacher, M.A. (1990). *Language comprehension as structure building*. Hillsdale, NJ: Earlbaum.

Geiser, S. & Studley, R. (2001). *UC and SAT: Predictive validity and differential impact of the SAT I and SAT II at the University of California*. Oakland, CA: University of California.

Golightly, K. B., & Sanders, G. (2000). *Writing and reading in the disciplines* (2nd Ed.). New Jersey: Pearson Custom Publishing.

Graesser, A.C., McNamara, D.S., & Louwerse, M.M. (2003). What do readers need to learn in order to process coherence relations in narrative and expository text. In A. P. Sweet & C. E. Snow (Eds.), *Rethinking reading comprehension* (pp. 82-98). New York, NY: Guilford Publications.

Graesser, A.C., McNamara, D.S., Louwerse, M.M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, and Computers, 36*, 193–202.

Light, R. (2001). *Making the most of college*. Cambridge, MA: Harvard University Press.

McNamara, D.S. (2001). Reading both high and low coherence texts: Effects of text sequence and prior knowledge. *Canadian Journal of Experimental Psychology, 55*, 51-62.

McNamara, D.S., Crossley, S.A., & McCarthy, P.M. (2010). Linguistic features of writing quality. *Written Communication, 27,* 57-86.

McNamara, D. S., Kintsch, E., Butler-Songer, N., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction*, *14,* 1-43.

McNamara, D.S., Louwerse, M.M., McCarthy, P.M., & Graesser, A.C. (2010). Coh-Metrix: Capturing linguistic features of cohesion. *Discourse Processes, 47*, 292-330.

O'Reilly, T. & McNamara, D.S. (2007). The impact of science knowledge, reading skill, and reading strategy knowledge on more traditional "high-stakes" measures of high school students' science achievement. *American Educational Research Journal*, *44,* 161–196.

O'Reilly, T., & McNamara, D. S. (2007). Reversing the reverse cohesion effect: Good texts can be better for strategic, high-knowledge readers. *Discourse Processes*, *43,* 121–152.

Sanders, T., & Schilperoord, J. (2006). Text structure as a window on the cognition of writing. In C. A. MacArthur, S. Graham & J. Fitzgerald (Eds.), *The handbook of writing research* (pp. 386 - 402). NY: Guilford Publications.

An acquired taste: How reading literature affects sensitivity to word distributions when judging literary texts

Justine Kao, Robert Ryan, Melody Dye & Michael Ramscar
Department of Psychology, Stanford University,
Jordan Hall, Stanford, CA 94305

Abstract

This study examines how reading habits affect people's sensitivity to word distributions in literary and non-literary writing. We manipulated eight literary and non-literary passages, creating modified versions that had lower word chunk frequencies but higher individual word frequencies than the originals. Subjects were then asked to rate the passages' quality of writing. Results showed that subjects with more experience reading literary writing (literary readers) gave higher ratings to original literary passages, while subjects with less literary reading experience (non-literary readers) preferred modified versions. Subjects with both types of reading habits rated original versions of non-literary passages higher. This indicates that literary readers are sensitive to frequencies of word chunks containing words that appear more frequently in the literary genre, while non-literary readers are not. We suggest that, over time, people can acquire slightly different representations of the probabilistic structure of language through their specific linguistic exposure.

Keywords: Psycholinguistics, Corpus linguistics, Word distributions, Genre differences, Reading habits, Discourse processes, Literary studies

Introduction

With one careful, calculated sip, a wine connoisseur can detect the subtle differences of quality between wines, and may even note the year and vineyard in which the grapes were grown. We, on the other hand, may stumble upon a thirty-year-old Bordeaux and not be able to tell it apart from a ten-dollar bottle. Appreciation for wine, like appreciation for high fashion or opera, is an acquired skill. Many fine things in life require years of experience to generate true appreciation. In what follows, we ask whether or not this "connoisseur phenomenon" translates to appreciation for literature as well. Is the ability to detect skill and beauty in literature also an acquired taste? If so, what is being acquired through the act of reading? Will avid readers have a stronger appreciation for good-quality writing, or be more sensitive to subtle changes of word choice?

As writing becomes an increasingly important form of communication and a central aspect of our lives, many studies have been conducted on the ways in which we are affected by what we read. Previous research shows that frequent readers are more sensitive to ambiguities in literary texts and are more likely to provide nuanced interpretations of them than infrequent readers (Dixon et al., 1993). Further, students who read recreationally perform better on reading comprehension and vocabulary tests (Anderson et al., 1988; Cipielwski & Stanovich, 1992), suggesting a relationship between reading enjoyment and competence. Still another study shows that frequent readers of literary writing have higher empathy and social measures than readers of non-literary writing (Mar et al., 2006). These results suggest that certain effects on our social, reasoning, and linguistic skills may be closely connected to the kinds of reading we engage in.

While these studies focus on higher-order social and cognitive effects of reading, we are interested here in examining how reading shapes readers' sensitivity to distributions of words. More specifically, we seek to explore whether readers' experience reading literary or non-literary writing shapes their sensitivities to word and chunk frequencies, and further, whether these fine-tuned sensitivities affect their judgments of quality when rating texts from different genres.

The Probabilistic Nature of Natural Languages

A slew of recent studies have shown that language users are sensitive to the distributional patterns of sounds, words, and even larger linguistic structures such as word sequences, or word 'chunks,' in the language they speak (see e.g., Altmann & Steedman, 1988; Bell et al., 2009; Bod et al., 2003; Bybee, 2002, 2006; Bybee & Hopper, 2001; De Long et al., 2005; Hale, 2003; Levy, 2008; Otten & Van Berkum, 2008; Pierrehumbert 2001, 2003; Ramscar et al., in press; Perruchet & Pacteau, 1990; Servan-Schreiber & Anderson, 1990). In many ways, the idea that we pay attention to how words are used is hardly surprising. It seems obvious, for example, that "a daunting task" sounds more "right," or more familiar, than "a daunting job." In fact, although *job* is a much higher frequency word than *task*, "a daunting task" appears 191 times on the Corpus of Contemporary American English (COCA), while "a daunting job" occurs only 6 times. We are sensitive to the different frequencies of the two chunks and prefer the one with the higher frequency. Since there is no real reason why it is less appropriate to describe a job as daunting, the preference for "a daunting task" over "a daunting job" does not seem to be driven by the appropriateness of the phrases' inherent *meanings,* but rather how the words are usually used.

The reason why we can sense these subtle mismatches is because words do not co-occur with each other with equal frequency. Indeed, the distribution of words in languages is

highly systematic (Baayen, 2001), and listeners are clearly sensitive to how words co-occur in sensible, and less sensible ways (see e.g., Wicha, Bates, Moreno, & Kutas, 2003). These kinds of co-occurrence patterns offer a rich and readily available source of information for anyone learning to understand the way that language relates to the world, and there is considerable evidence to support the idea that people are sensitive to this information.

However, it is critical to note that every person's internal model of his or her language is trained on a slightly different corpus. In other words, each person hears and reads different things throughout his or her life, and over time these differences in the input may result in different representations of the language. In written language, for example, genres of writing have been observed to differ on a number of linguistic dimensions. Research on corpus comparison and genre detection makes use of the idea that word distributions – how words are used and which words are used – differ across genres (Biber 1988, 1993; Eisenbeis & Avery, 1972; Karlgren & Cutting, 1994; Lee & Myaeng, 2002, Xiao & McEnery, 2005). Work in literary theory has also suggested that literary texts often use low-frequency words to foreground certain elements of writing (Miall & Kuiken, 1994, Mukarovský, 1964), while non-literary texts tend to use more conventional words to convey meaning clearly. Given that there is marked variation in the distributions of words that people will be exposed to over the course of their lives, it seems likely that people will have different sensitivities to word distributions depending on their "training sets." We examine this possibility through the lens of writing genres.

'Literary' and 'Non-literary' Words

For our purposes here, we class writing into two primary domains: literary and non-literary. Much of what people read can be identified as one of the two, with fiction and poetry belonging to the former category, and newspaper articles and textbooks to the latter. Based on whether a word occurs more frequently in literary writing or non-literary writing, we can refer to it as a 'literary' word or a 'non-literary' word. For example, "abruptly" is a literary word (37 per million in the fiction corpus and 6.7 per million in the newspaper corpus), while "actively" is a non-literary word (2.54 per million in fiction and 9.97 in newspapers) (Corpus of Contemporary American English (COCA)).

As we will illustrate in a later section, literary texts tend to contain more literary words, while non-literary texts tend to contain more non-literary words. Since literary and non-literary words are *defined* by how often they occur overall in literary and non-literary writing, this may not seem entirely surprising. However, it sheds light on the deeper point that the words in a given piece of writing will have different distributions depending on the corpus you examine (e.g., the frequency and usage of "abruptly" will differ sharply between a "non-literary" newspaper corpus and a "literary" fiction corpus).

This has implications for how people may be affected by their reading practices. Given that some people's reading habits may make them more familiar with one "corpus" than another (i.e., they may be more widely read newspapers and journal articles than fiction and poetry), this difference in exposure should translate into a corresponding difference in their probabilistic representation of the distributions of words within their language. In other words, readers within different genres will have learned somewhat different distributional patterns, and these differences should be similar to the ones that we can actually research and quantify by analyzing different corpora.

This leads to testable predictions. For example, we would expect that literary readers would be more sensitive to the probabilistic distributions of literary words than non-literary readers, and we would also expect them to have a better understanding of the environment – or linguistic context – in which such words are likely to occur. Thus, they should show higher sensitivity than non-literary readers to the frequencies of *chunks* of words in literary texts.

Reading Habits and Judgment: Experiment

In order to test the predictions detailed above, we selected four excerpts of choice contemporary fiction writing and four excerpts of non-literary writing. We then systematically manipulated the frequencies of several chunks (short sequences of words) within each passage, creating modified versions of each of the eight passages. Our method of modification is detailed in the section "Manipulation of passages." After creating the 8 modified versions, we had16 testing passages total: 4 literary and 4 non-literary *original* passages, which contain higher overall chunk frequencies but lower overall word frequencies, and 4 literary and 4 non-literary *modified* passages, which contain lover overall chunk frequencies but higher overall word frequencies. We hope to examine whether subjects' evaluations of writing quality differ for the original and modified versions, and further whether literary and non-literary readers' evaluation of literary and non-literary texts also diverge.

We hypothesize that for literary texts, literary readers will give higher ratings to literary passages containing chunks that have higher frequencies, because these chunks will be more familiar in the corpus they have been trained on, and thus more representative of their internal models of language (e.g., they should recognize "adamantine luster" as a frequent literary pairing and prefer it over "adamantine milk," which is not a frequent literary pairing). By contrast, we hypothesize that non-literary readers, who lack the same levels of exposure to 'literary' words and their contexts, will only be sensitive to individual word frequencies, and will prefer more highly frequent words even when they are used in contexts (e.g., "adamantine milk") that would seem anomalous or even jarring to a literary reader. In terms of quality ratings, this suggests that literary readers will prefer the original literary passages with higher chunk frequencies,

whereas non-literary readers will prefer the modified literary passages with higher individual word frequencies.

With regards to the non-literary texts, the picture is less clear. It may be that we should expect the opposite effect: that literary readers will prefer modified passages while non-literary readers will prefer the originals. However, it also seems likely that our literary readers, who read for pleasure, may read more widely than our non-literary readers, and be sensitive to our non-literary manipulations as well.

Participants

Participants were 31 Stanford University undergraduates recruited for credit for an introductory psychology course. All subjects were monolingual English speakers.

Materials

Four excerpts from literary writing and four excerpts from non-literary writing, each ranging from 80 to 130 words in length, were selected as materials. The literary passages were selected from four separate stories in "*The Vintage Book of Contemporary American Short Stories*," a collection of short stories featuring distinctive short fiction in American English published within the last 25 years. Three journalistic, or non-literary, English passages were selected from articles in the New York Times during the past year, and one non-literary passage was chosen from a reading comprehension article in a 2009 GRE prep book. Passages from each genre varied in style and content. We chose materials from these sources because they reflect high quality of writing, offer a variety of styles and themes, and are not famous or widely enough read to be likely to be recognized by our subjects during the survey.

Methods

Assessment of Passages

To explore the degree to which literary texts tend to contain more literary words and non-literary texts tend to contain more non-literary words, we examined the 400 million word COCA corpus (Davies, 2009) recording the frequency of each word in each passage in the fiction corpus, the newspaper corpus, and the corpus as a whole. The average log frequencies of the passages in the three corpora are shown in figure 1. This analysis revealed that within the specific corpora, the literary passages had significantly higher average frequencies in the fiction corpus than in the newspaper corpus ($t(670)=2.3148$; $p < 0.05$) whereas the average frequencies of the non-literary texts in the fiction and newspaper corpora were not significantly different ($t(584)=-1.0288$; $p>0.05$). This suggests that words occurring in literary texts are more frequent in literary than non-literary texts, while words in non-literary texts are more evenly distributed across literary and non-literary texts. This idea is supported by an analysis of the overall corpus, which revealed the literary passages to have higher average frequencies than the non-literary passages ($t(627)=2.2786$; $p<0.05$). Together these findings suggest that literary texts make specialized use of a specific subset of the overall corpus, rather than employ a markedly different vocabulary. Consistent with this idea, a 2 (literary versus non-literary text) x 2 (fiction versus newspaper corpus) ANOVA of the average frequencies of the texts revealed an interaction between text type and corpus type ($F(1, 627)=13.324$, $p<0.001$), and a main effect of text type ($F(1,627)=121.926$, $p<0.001$).

A more fine-grained analysis of the texts further supported the idea that the distribution of vocabulary items is specialized in different kinds of writing. When we compared the pair-wise frequency of each word in each specific corpus, we found (unsurprisingly) that the pair-wise frequencies of the words in the literary passages were significantly higher in the corpus for fiction writing than in the corpus for newspaper writing ($t(335)=11.4987$; $p<0.001$), but also that the reverse was true for each of the words in non-fiction passages ($t(292)=-4.7295$; $p <0.001$). In other words, what appears to set literary and non-literary writing apart is not that they make use of specialized sets of words, but rather that words are used in specialized ways in different kinds of writing, and, at least in this sample, the distribution of vocabulary within literary writing in English appears to be particularly distinctive.

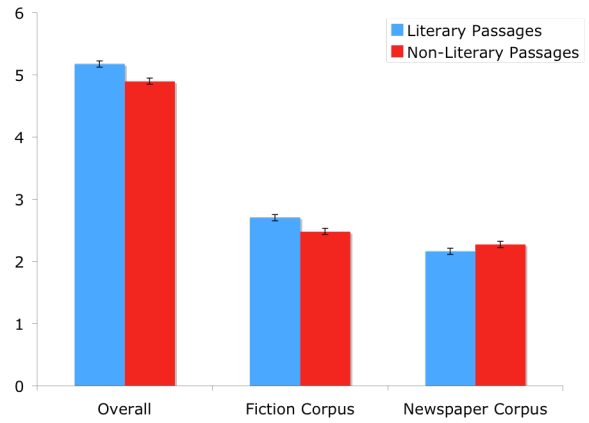


Figure 1. Average log frequencies of passages in different corpora

Manipulation of Passages

After analyzing the passages, we then manipulated the frequencies of three to seven chunks (strings of words) within each literary and non-literary passage, lowering chunk frequency while simultaneously retaining (or even raising) average individual word frequency. Measures of individual word frequency were taken form COCA, while chunk frequency was based on the number of 'hits' a chunk returned on Google. The reason why we used Google to measure chunk frequencies is because its magnitude allows us to find 'hits' for word sequences that are several words long, while many longer word sequences would return 0 counts even in large corpora like COCA, and thus fail to

measure the differences in frequencies of longer word chunks.

In the following example, (a) is the original chunk, and (b) is the modified chunk.

(a) On the further ***side*** of the field[1]
(b) On the further ***part*** of the field

While *side* and *part* have highly similar meanings in this context, *side* has an average frequency of 317/ million in COCA, while *part* has frequency of 479/ million, suggesting that *part* is a more frequent word in English than *side*.

In terms of chunk frequencies, "further side" returns 259,000 hits on Google, whereas "further part" returns 374,000 hits, suggesting that in English, "part" is more likely than "side" given "further" (Miller & Chomsky's[2] famous point about the lack of evidence for specific strings in English can be illustrated by considering that there are insufficient instances of "part of" and "side of" in the 400 million words of COCA to facilitate an analysis). Finally, the highly frequent "of" is the most likely word to follow both "side" and "part" in English, and by both our COCA and Google measures, the likelihood of "of" given "part" is three times that of "side"("part of" has 132176 Google hits compared to 40446 for "side of," and "part of" occurs on average 330 times per million words in COCA, as compared to 101/million for "side of").

Thus, the "average probabilities" of English suggest that, as a string of words, (b) is much more likely than (a). However, the original chunk (a) returned 898 hits on Google, while the modified chunk (b) returned 0 hits. Thus, although and both (a) and (b) appear to be similar in meaning and equally "grammatical," and although the average frequency of all words, and the average transitional probabilities between them in English as a whole are higher in (b), given that its chunk frequency is considerably lower, it appears that the likelihood of actually encountering (b) in English is lower than it is for (a).

Since we wished to manipulate word and chunk frequencies while keeping the meaning of the passages relatively constant, chunks were selected for modification in the manner just described based on whether or not they contained a word that could be replaced with a synonym that had a similar or higher frequency.

Procedure and design

All surveys were designed and distributed using the Qualtrics online survey software. Each survey had four literary and four non-literary passages, half of which were original excerpts, and half of which were modified as described in the section above. Two versions of the survey were distributed: either the odd-numbered passages were modified and the even-numbered passages were kept as the original, or vice versa. Participants were randomly assigned one of the two versions.

[1] Taken from "Emergency," by Denis Johnson
[2] Miller & Chomsky, (1963)

Participants were surveyed individually on a computer. They were asked to read the instructions in the survey carefully, and the time it took each subject to complete the survey was recorded to make sure they spent enough time reading the passages and answering questions.

Participants were presented with each passage in the same order and asked to read carefully. While each participant read the passages in the same order, the order of the passages was counterbalanced with respect to passage type. For example, in the version of the survey in which the odd-numbered passages were modified, the passages appeared in the order of: modified literary, original non-literary, modified literary, original literary, modified non-literary, original non-literary, modified non-literary, original literary. This design should weaken the effects of passage type ordering on subjects' preferences.

After subjects finished reading the passage as a whole, the same passage appeared again, but this time with a selection highlighted. They were asked to rate the quality of the highlighted section on a 7-point scale, with 7 being "Very well-written," and 1 being "Very poorly written." Each passage was equally divided into three sections, separately highlighted and presented to the subjects for rating.

After participants finished reading and rating all eight passages, they were asked to provide an estimate of how many hours a week they usually spent reading literary texts (including poetry, magazine stories, creative non-fiction, and novels) and non-literary texts (including text books, newspaper articles, and academic papers). In order to arrive at a score of how much more experience each subject had reading literary writing compared to reading non-literary writing, the hours reading literary texts was divided by the hours reading non-literary texts, a ration we will refer to as the 'literary reading bias.' We use this as our measure for subjects' reading habits because it reflects the relative amount of time they read literary texts versus non-literary texts, which for our purposes is the salient feature of their reading habits.

Results

A repeated measures ANCOVA of participant ratings of the modified and non-modified passages with literary reading bias as a continuous covariate revealed a significant interaction between literary reading scores and within-genre preference ($F(1, 21) = 3.095$; $p < 0.05$; see figure 2). To facilitate further analysis, subjects were divided equally into two groups based on their literary reading bias, with subjects whose scores were above the median placed in one group, and subjects whose scores were below the median placed in the other. The within-genre preference of each subject was measured using the difference between his or her average ratings for original and modified passages of each genre. The two groups' average preferences within each genre are shown in figure 2.

Further, participants who read more fiction relative to non-fiction writing showed a stronger preference for the

unmodified literary texts compared to participants who read more non-fiction ($t(29) = 1.7377$; $p<0.05$), and a one-sample t-test revealed that participants who read more fiction showed an overall preference for the original literary passages ($t(14)=1.856$; $p<0.05$. For non-literary passages, there was no significant difference between the within-genre preferences of subjects in the two groups ($t(29)=0.6556$; $p>0.5$), and while both groups showed a preference for the original non-literary passages, these preferences were not significant.

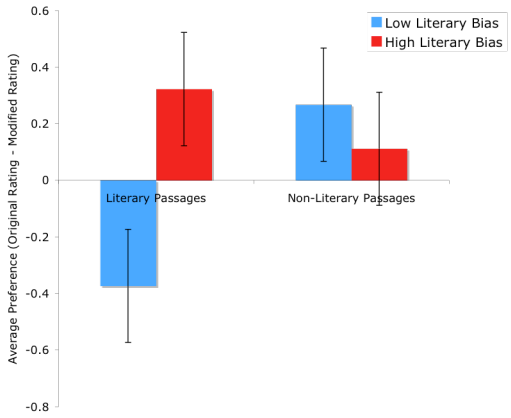


Figure 2. Literary and non-literary readers' preference for original passages in the two genres

Discussion

As predicted, there was a significant interaction between subjects' reading habits and their reading preferences. How might one explain these results?

Only people who are exposed to the distributional properties of those words in literary contexts appear be sensitive to our manipulations, which is consistent with our assessment of the passages, where we found that literary writing uses words in ways that are literary specific. On the other hand, both literary and non-literary readers were sensitive to the manipulation of non-literary passages. One reason may be that non-literary writing makes use of less specialized distributions, as shown in our corpus analysis. Literary writing can be thought of as a specialized form of writing that re-employs and expands upon distributional information also present in non-literary writing, which makes literary readers still reasonably familiar with the distributions of words in non-literary texts, whereas the same cannot necessarily be said for non-literary readers and literary texts. Another reason may be that our social and cultural lives naturally enforce a non-literary expertise on all readers, while literary expertise is more a matter of individual practice.

One potential weakness for our study was that we relied on self-report to measure our subjects' reading habits. There may be issues of accuracy in recall, given that subjects were trying to judge the exact number of hours they spent reading in a given week. For this reason, we used the *ratio* between reported literary and non-literary reading hours as a means of comparing our subjects. This ratio should, at the very least, reflect the subject's subjective sense of how much time he or she devoted to reading literary writing relative to non-literary writing, and hopefully separates out fiction and poetry readers from magazine and front-page readers.

In future studies, it may be possible to use more "objective" measures of reading habits— for instance, by examining the number of literary and non-literary authors each subject can identify (Mar et al., 2006), or by having subjects track their reading habits over time. Alternatively, we might conduct a study in which we ask a certain group of subjects to exclusively read literary texts for an extended period, while having another group read exclusively non-literary texts, and then measure the effect.

Our preliminary findings on the subject suggest that each person's model of the language they speak may be affected and "trained" over time by the specific linguistic samples they encounter. Intriguingly, differences in these individual language models appear to correspond with differences in "subjective" perceptions and judgment. Here, we examined how prior reading exposure may affect our perception and judgments of reading new texts. If our findings generalize to different genres of writing, spoken language, or even other modes of art and communication, we may be able to begin to explain individual differences in judgment and perception, and also how one can acquire taste through experience.

Acknowledgments

This material is based on work supported by the National Science Foundation under Grant Nos. 0547775 and 0624345 to Michael Ramscar, and by the Stanford University Undergraduate Advising and Research under Small Grant No. 4174 to Justine Kao.

References

Altmann, G.T.M. & Steedman, M. (1988) Interaction with context during human sentence processing. *Cognition*

Anderson, R., Wilson, P., & Fielding, L. (1988). Growth in reading and how children spend their time outside of school. Reading Research Quarterly

Baayen, R. H. (2001). Quantitative aspects of morphological productivity. In G. E. Booij and J. van Marle (eds), Yearbook of Morphology 1991, Kluwer Academic Publishers, Dordrecht

Bell, A., Brenier, J., Gregory, M., Girand, C., & Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language* 60(1)

Biber, D. (1988). *Variation across Speech and Writing*. Cambridge, UK: Cambridge University Press.

Biber, D. (1993). Representativeness in Corpus Design. *Literary and Linguistic Computing* 8 (4)

Bod, R., J. Hay, & Jannedy, S. Eds. (2003). *Probabilistic linguistics*. Cambridge, MA, MIT Press

Bybee, J. and P. Hopper, Eds. (2001). *Frequency and the emergence of linguistic structure*. Typological studies in language, vol. 45. Amsterdam, Netherlands, John Benjamins Publishing Company.

Bybee, J. (2002). Phonological evidence for exemplar storage of multiword sequences. *Studies in Second Language Acquisition* 24(2)

Bybee, J. (2006). From usage to grammar: the mind's response to repetition. *Language*

Cipielewski, J., & Stanovich, K.E. (1992). Predicting growth in reading ability from children's exposure to print. *Journal of Experimental Child Psychology,*

Cover, T.M. & King, R.C. (1978). A Convergent Gambling Estimate of the Entropy of English. *IEEE Transactions on Information Theory*

Davies, Mark. (2008-) The Corpus of Contemporary American English (COCA): 400+ million words, 1990-present. Available online at http://www.americancorpus.org.

DeLong K. A., Urbach,T.P., Kutas, M. (2005) Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*

Dixon, P., Bortolussi, M., Twilley, L. C. and Leung, A. (1993) 'Literary Processing and Interpretation: Towards Empirical Foundations', *Poetics*

Eisenbeis, R., and R. Avery (1972). *Discriminant Analysis and Classification Procedures: Theory and Applications*. Lexington, Mass.: D.C. Health and Co.

Hale, J. (2003). The Information Conveyed by Words in Sentences. *Journal of Psycholinguistic Research*

Karlgren, J., and D. Cutting (1994). Recognizing text Genres with Simple Metrics Using Discriminant Analysis. In *Proc. of the 15 '1' International Conference on Computational Linguistics (COLING '94)*.

Lee, Y.B. & Mayeng, S.H. (2002). Text Genre Classification with Genre-Revealing and Subject-Revealing Features. *Proceedings of the 25th ACM SIGIR Conference*.

Levy, R (2008). Expectation-based syntactic comprehension. *Cognition*

Mar, R.A., Oatley, K., Hirsh, J., dela Paz, J., & Peterson, J.B. (2006). Bookworms versus nerds: Exposure to fiction versus non-fiction, divergent associations with social ability, and the simulation of fictional social worlds. *Journal of Research in Personality*

Miall, D.S. & Kuiken, D. (1994). Beyond Text Theory: Understanding Literary Response. *Discourse Processes*

Mukarovský, J. (1964). Standard language and poetic language. *In P. L. Garvin (Ed.), A Prague School reader on esthetics, literary structure, and style* Washington, DC: Georgetown University Press. (Original work published 1932.)

Otten, M., & Van Berkum, J. J. A. (2008). Discourse-based word anticipation during language processing: Prediction of priming? *Discourse Processes*

Pierrehumbert, J. (2001) Exemplar dynamics: Word frequency, lenition, and contrast. In J. Bybee and P. Hopper (eds.) *Frequency effects and the emergence of lexical structure*. John Benjamins, Amsterdam.

Pierrehumbert, J. (2003) Probabilistic Phonology: Discrimination and Robustness. In R. Bod, J. Hay and S. Jannedy (eds.) *Probability Theory in Linguistics*. The MIT Press, Cambridge MA

Ramscar, M., Matlock, T., & Dye, M. (in press) Running down the clock: the role of expectation in our understanding of time and motion. *Language and Cognitive Processes*

Servan-Schreiber, E., & Anderson, J.R. (1990). Learning artificial grammars with competitive chunking. *Journal of Experimental Psychology: Learning, Memory, and Cognition*

Wicha, N. Y., Bates, E. A., Moreno, E. M., & Kutas, M. (2003). Potato not pope: Human brain potentials to gender expectation and agreement in Spanish spoken sentences. *Neuroscience Letters*

Xiao, Z.H. & McEnery, A. (2005). Two Approaches to Genre Analysis: Three Genres in Modern American English. *Journal of English Linguistics*

What You Did and Didn't Mean: Noise, Context, and Human Skill

Tiziana Ligorio (tligorio@gc.cuny.edu)
Susan L. Epstein (susan.epstein@hunter.cuny.edu)
Department of Computer Science, Hunter College and The Graduate Center of The City University of New York
New York, NY 10016 USA

Rebecca J. Passonneau (becky@cs.columbia.edu)
Joshua B. Gordon (gordon@cs.columbia.edu)
Center for Computational Learning Systems, Columbia University
New York, NY 10037 USA

Abstract

In spoken dialogue between people and machines, the computer must understand not only what the speaker means but also what she does not. The computer begins with a considerable disadvantage: even the best speech recognition technology can provide error-ridden transcriptions of human speech under real-world telephone conditions. The work recounted here examines how, and how well, people use context to interpret noisy transcribed utterances in a challenging domain. Models learned from this experiment highlight two aspects of this human skill: the ability to detect a context-supported match, and the ability to know when the quality of attempted matches is so poor that it should be questioned. These models can then be applied by a spoken dialogue system to find the correct interpretation of users' spoken requests, despite incorrect speech recognition.

Keywords: spoken dialogue systems; natural language processing; machine learning; Wizard of Oz studies, learning.

Introduction

A computer system intended to replicate a human skill faces two considerable limitations: it works from a different input modality and it is restricted to a preprogrammed set of alternative actions. The thesis of our work is that people should be studied for their skill at the target task as if they were similarly restricted, that is, as if they had only the system's input and alternatives. The domain of investigation here is a spoken dialogue system (*SDS*). Subjects were given the same data that would be available to the SDS: error-ridden strings representing transcribed speech plus a large database of possible matches (Passonneau et al., In Press). The resultant data was then used to identify the best performers, and to learn models of them destined for the system. There were two principal results. First, subjects ably guessed what the speaker meant, that is, they could often identify the correct item from the context provided by a database query on the error-ridden string. Second, the people most skilled at this task excelled because they could also identify what the speaker had *not* meant, that is, they knew when no item returned from the database was a correct match. Such recognition is essential to move dialogue forward constructively

Ideally, an SDS offers people a natural way to communicate with a computer and benefit from its expertise. In an SDS, automated speech recognition (*ASR*) transcribes human spoken input into a string of words, which is then assigned an interpretation. Under real-world telephone conditions, however, even state-of-the-art ASR can exhibit a word error rate (*WER*) as high as 68% (Raux et al., 2005). High WER is common when the environment is noisy, the language the system is expected to understand is flexible and based on a large vocabulary, or the user population is diverse in gender, age, and native language. These are all characteristic of our target domain: telephoned book requests from patrons of the Andrew Heiskell Braille and Talking Books Library.

Although people manage dialogue well in the presence of noise, computers do not. Our subjects used ASR output from a spoken title to query our copy of Heiskell's book database. For example, for one book title, the ASR output string was "ROLL DWELL." A database query on this string returned three likely matches (in real time): "CROMWELL," "ROBERT LOWELL," and "ROAD TO WEALTH."

This is a difficult task. (The reader is invited to guess whether any of these actually matches the spoken title.) Our experiment studies how people manage this task. The resultant data is then used to train accurate models of human behavior, and to identify the features that make people proficient. Such models are ultimately intended as an integral part of an SDS, to make it more robust to noisy ASR in the context of database queries. The next sections of this paper describe related work, our target system, and the experimental design and results. These are followed by a description of the models learned from the data collected during the experiment, and a discussion of their import and application.

Related Work

The Wizard of Oz (*WOz*) paradigm is a well-known approach to iterative prototype design. It gathers information about the characteristics of a successful system before the system's development (Dix, et al. 2003). In a WOz experiment, only a user-system interface is provided. Users believe they are interacting with a computer system through this interface, but instead a person (the *wizard*) is "behind the curtain." This permits the system designer to observe human responses to certain system functionalities, to study user behavior and expectation, and to assess interface design features before the construction of an initial prototype.

WOz can also be used to study the wizard, to provide data on how the system should behave. In particular, *wizard ab-*

lation is a WOz study in which a wizard relies on system input and output rather than her own communication resources (Levin and Passonneau 2006). Wizard ablation supports the collection of dialogues that illustrate the decisions people make when confronted by the same input/output data and choices as an SDS. Data collected under wizard ablation supports supervised learning to predict wizard actions. The resultant model can then be incorporated into a system to improve its behavior. WOz studies that directed their attention to the wizard during full spoken dialogues include efforts to predict: the wizard's response when the user is not understood (Bohus 2004), the wizard's use of multimodal clarification strategies (Rieser and Lemon 2006), and the wizard's use of application-specific clarification strategies (Skantze 2005). The experiment presented here is restricted to single utterances, rather than full dialogues. It also differs in that it analyzes several wizards' behavior. It recognizes differences among wizards and identifies distinctive and successful behavior, so that the system will ultimately benefit only from models of the most skilled wizards.

To limit communication errors incurred by faulty ASR, an SDS may use enriched strategies to detect and respond to incorrect recognition output (Bohus 2004). It may repeatedly request user confirmation to avoid misunderstanding, or ask for confirmation using language that elicits responses from the user that the system can handle (Raux and Eskenazi 2004). When the user adds new information in response to a system prompt, two-pass recognition can consider the extra information contained in such user responses to restrict the language expected in the second pass and thereby achieve better recognition (Stoyanchev and Stent 2009). In a highly interactive setting, an SDS might benefit when it takes this approach one step further and uses context-specific language for incremental understanding of the noisy input throughout the dialog (Aist, et al. 2007). This paper explores the use of system-internal resources, such as a database search, to respond to faulty ASR. It embeds a wizard into a system, and then observes and models her ability to use such context to respond appropriately.

Peripherally related are other approaches that increase understanding between an SDS and the user through the adaptation of an SDS's response based on a user model. In automated tutoring, for example, it is essential to validate the user when she is correct and to elicit more reasoning when she is not (Franceschetti, et al. 2003, Ohlsson, et al. 2003). In particular, affect-adaptive systems can improve learning efficiency by responding to uncertainty in the transcribed speech (Forbes-Riley and Litman 2009).

Ordering Books with CheckItOut

CheckItOut is a research SDS for book requests from patrons of the Andrew Heiskell Braille and Talking Books Library, a branch of the New York Public Library and part of the National Library System. Patrons of the library request books by telephone and receive them by mail. Regular newsletters provide patrons with the titles and catalogue numbers of new books. To gauge the kinds of interactions patrons have with Heiskell's librarians, we transcribed 82 telephone calls from a larger set we had recorded. Forty four percent of the book requests were by catalogue number, 28% by title or a combination of title and author, and 28% were more general. CheckItOut is therefore designed to accept book requests by catalogue number, author, or title.

CheckItOut builds upon the Olympus/Ravenclaw architecture and dialogue management framework (Bohus, et al. 2007, Bohus and Rudniky 2003). Olympus/Ravenclaw has been the basis for approximately a dozen research SDSs in different domains. During a dialogue with CheckItOut, the user first identifies herself as a patron of the library, and then requests at most four books. CheckItOut references two databases: a sanitized version of Heiskell's database of 5,028 patrons, and its entire book database with 71,166 titles and 28,031 authors. These force CheckItOut to manage a large vocabulary; titles and author names alone contribute 54,448 distinct words. Moreover, Heiskell's patrons include many elderly and non-native speakers. The experiment described next observes how human wizards respond to the same challenges that CheckItOut confronts.

Experimental Design

The experiment described here seeks to uncover how people marshal system resources (e.g., the ASR string and database results), and which strategies achieve the best performance. Here the focus is on single turn interactions that request books by title, the CheckItOut request type most likely to elicit problematic ASR output.

In an offline pilot study, 3 native speakers of English read 50 titles to generate 3 sets of ASR output strings (Passonneau, et al. 2009). Each subject received a different ASR set and was asked to find the corresponding title from a text file that listed all 71,166 titles. WER was 69% – 83%, depending on the speaker. Despite the high WER, these subjects identified the correct title 74% of the time.

Given this demonstration of human skill, we designed a WOz study to identify which aspects of human performance come into play when a wizard seeks to match noisy ASR against a list of *candidates* (possible title matches) (Passonneau et al. In Press). The experiment was designed to identify what makes a good wizard, and to extract any additional insights a wizard may offer when supported by database search with the quality common in modern systems.

During the experiment, users and wizards were isolated from one another in separate rooms. Each had her own graphical user interface (*GUI*) and microphone. In a *title cycle*, the user read a book title into a speech recognizer through the microphone, and the corresponding ASR was displayed on the wizard's GUI. The wizard then formulated a query for the database. Once the search returned a list of candidates, the wizard had four options: make a *confident choice* among the candidates, make a *tentative choice* among the candidates, ask a *question* through her microphone, or *give up*. (Wizards were also permitted to ask the user to repeat the title, but were discouraged from doing so.) If the wizard chose a candidate, it then appeared on the us-

user's GUI, and the user scored it as correct or incorrect. That score was also displayed on the wizard's GUI, so that the wizard knew if her most recent title choice was correct or incorrect. If the wizard asked a question instead, the user heard it through her headset and rated it on her GUI. The possible ratings with respect to the current book request were "relevant and I can answer it," "relevant but I cannot answer it," irrelevant," and "uncertain." Question ratings were not shared with the wizard. After the wizard saw the user's score or was notified that the user had judged the question, the wizard signaled the beginning of a new cycle.

The speech recognizer continually transcribed the speech signal from the user's microphone, and the wizard's GUI provided a live feed of the resultant ASR strings. For each request, the wizard submitted a database query after very limited editing of those strings (e.g., removing "um"). The return from the database was displayed on the wizard's GUI as a list of candidates in descending order of search confidence. This confidence was measured using Ratcliff/Obershelp pattern recognition (*R/O*) which evaluates the similarity of the ASR string to a book title from the database (Ratcliff and Metzener 1988). Confidence scores were not displayed on the wizard's GUI.

Given an ASR query, the database produced one of the four following kinds of returns, based on the R/O scores:

- *Singleton*: the single top-scoring candidate, if any were very good (R/O ≥ 0.85)
- *AmbiguousList*: two to five moderately good candidates (0.85 > R/O ≥ 0.55)
- *NoisyList*: six to nine poor but non-random candidates (0.55 > R/O ≥ 0.40)
- *Empty:* No candidates (max R/O < 0.40)

Our focus here is not on the database search, but on the wizard's actions given noisy ASR and an adequate but imperfect database return. Words in each candidate that exactly matched a word in the query appeared darkest on the GUI. All other words appeared in grayscale in proportion to their degree of character overlap with the words in the query.

Two of the seven subjects were non-native speakers of English (one Spanish, one Romanian). Each pair of students (a total of 21 possible pairs) met five times. In each meeting, one student was the user and the other was the wizard in a *session* of 20 title cycles. Then the pair immediately exchanged roles to run a second session of 20 title cycles. Thus, each student was the wizard on 100 title cycles and the user on 100 title cycles with every other student, for a possible 4200 title cycles in all. Users were permitted to end a session early after fewer than 20 title cycles if they experienced severe system problems.

Beyond the mechanics of this process, it was important to create a dialogue-like environment and to encourage the best possible performance from our subjects. To make her speech more conversational and less like simply reading a list, the user prepared immediately before each session. She read brief synopses of the 20 titles (chosen at random from the database) and then ordered them in some way (e.g., genre or theme) relevant to their content. To encourage thoughtful decisions, no time limits were imposed upon either the wizard or the user. Finally, we devised a score that subjects were asked to maximize throughout the experiment, with prizes to be awarded for the top two scorers. The wizard scored +1 for a correctly identified title, +0.5 for a relevant question and -1 for an incorrect title. To encourage cooperation between users and wizards, the user also scored +0.5 for a successfully recognized title.

Results

The analysis in this section provides essential support for automatically learning models of intelligent behavior worthy of incorporation into an SDS. Given the permitted early termination, there were 4172 title cycles (instead of 4200). In them, the average WER was 69%. Nonetheless, the distribution of database returns was 46.7% Singleton, 53.26% AmbiguousList and 2.83% NoisyList. (Although in pilot tests 5% - 10% of the returns were empty, during the experiment itself none were.)

Figure 1 shows the overall distribution of wizard actions for our subjects, W1 through W7. Each of them saw a similar distribution of database returns: Singleton ($\mu = 278.57$, $\sigma = 21.16$), AmbiguousList ($\mu = 300.57$, $\sigma = 16.92$), and NoisyList ($\mu = 16.86$, $\sigma = 4.78$). The correct title was among the candidates returned by the database 71.31% of the time. Singleton returns were the correct title 92.05% of the time. AmbiguousList and NoisyList returns contained the correct title 53.74% of the time.

Ideally, a wizard should identify the correct title if it appears among the candidates, and otherwise ask a thoughtful question that could constructively advance the dialogue. As one might expect from our pilot study, wizards knew what the user meant when they saw it. If the correct title was among the candidates, wizards identified it confidently 68.72% of the time and tentatively 26.53% of the time — 95.25% in all. Recall, however, that AmbiguousLists and NoisyLists were sorted by search confidence. When the database returned multiple candidates, the top candidate was the correct title 41% of the time. It was second 5.81%, third 2.61%, fourth 2.20%, and later (fifth through ninth) 1.67% of the time. This did indeed help the wizards, who correctly offered the first title 98.34% of the time (74.24% confidently, and 24.10% tentatively). Of course, preference for

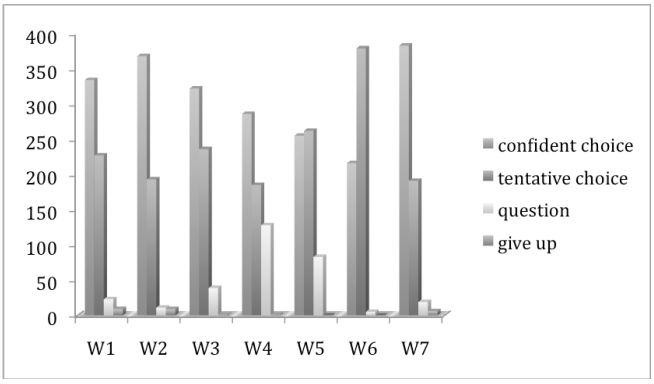


Figure 1: Distribution of wizard actions

the top returned candidate is readily programmed into an SDS. Instead we focus here on what wizards did when the title was not among the candidates.

Wizards were less skilled at recognizing what the user had not meant. Indeed, their performance differed primarily on their response when the correct title was not among the candidates — most wizards were less accurate then, and their performance was less uniform. Despite careful instructions to the subjects that had explained this option, wizards asked a question in only 22.32% of the cases where the correct title was not among the candidates. Instead they made a tentative guess (67.71%), chose confidently (7.78%), or gave up (2.20%). Table 1 shows each wizard's number of title cycles, session score, and *accuracy,* the proportion of title cycles where she identified the correct title or correctly recognized that the title was not among the candidates (by asking a question or giving up). It also shows the frequencies with which she offered the top candidate and correctly recognized that the title was not among the candidates

Wizards are ranked in Table 1 in descending order of session score and accuracy. Those values are highly correlated ($R = 0.91$, $p = 0.0041$). W4 scored highest, primarily because of the frequency with which she asked a question when the candidates did not include the correct title (correct non-offers = 64%). Table 2 shows the distribution of what should have been the correct action across all 4172 title cycles. The correct action was either to offer the title as the correct candidate at a given position (Return 1 through Return 9) or to ask a question or give up when the title was not among the candidates. Table 2 makes clear that the simple strategy "always guess the top candidate" (as our wizards often did) would achieve about 65% accuracy. Note too that those wizards who relied on it most (W3 and W6) were also the least accurate overall, while the wizard who relied on it least (W4) was the most accurate. Clearly, given a reasonable but fallible database search on noisy ASR, an SDS should emulate W4, not simply choose the top candidate.

Learning to be Like a Wizard

Wizards collaborate with the SDS — the system manages input and output (except for the wizard's questions), while the wizard exploits the available information (ASR string and database return) to make a decision. Our experimental design also captured data that described the system and the wizard's session history. That data was then used to train models of wizard actions selection. Such models could be used to implement the best wizard behavior within an SDS.

The experiment collected data on 60 features available at run time, selected for their likely relevance to wizard action choices. They described the ASR (e.g., number of words in the ASR string), the recognition process (e.g., recognizer's confidence score when it produced the ASR string), the speech signal (e.g., speech rate as number of 10ms speech fragments per word), the ability of the SDS to interpret the ASR string (e.g., number of parses in the natural language understanding component), and Olympus/ Ravenclaw confidence scores that combine recognition with language understanding. (Much of this system information was not available to the wizard.) Other features described the session history (e.g., number of correctly identified titles so far), the database return (e.g., return type of Singleton, AmbiguousList, NoisyList), or the similarity between the ASR string and the candidates (e.g., number of matching words). Because the number of candidates differed across title cycles, these features were averaged over multiple candidates.

Table 1: Raw session score, accuracy, proportion of offered titles listed first in the database search return, and frequency of correct non-offers for seven participants.

Subject	Cycles	Session score	Accuracy	Chose #1	Correct non-offer
W4	600	0.7585	0.8550	70%	64%
W5	600	0.7584	0.8133	76%	43%
W7	599	0.6971	0.7346	76%	14%
W1	593	0.6936	0.7319	79%	16%
W2	599	0.6703	0.7212	74%	10%
W3	581	0.6648	0.6954	81%	20%
W6	600	0.6103	0.6950	86%	3%

As a machine learning technique, we chose decision trees to model wizard behavior because they are easy to interpret and compare, and relatively transparent. A decision tree maps feature values to a target value (here, wizard action). A decision tree is a tree-like structure of nodes with directed links between them. Each node is a branch test based on feature values. To simulate the modeled behavior, a program traces a path from the *root* (the top node), following the branch tests until it reaches a *leaf*, a non-branch node that provides a target value. With a version of C4.5 (Quinlan 1993), we trained two kinds of decision-tree models: an *overall model* that used data from all the wizards to predict wizard action in general, and seven individual *wizard models,* one for each wizard.

Cross-correlation over the features indicated that many of the initial 60 features were heavily correlated. We manually isolated groups of correlated features with $R^2 > 0.5$, and retained only one representative feature from each group. We grouped features that described the similarity between ASR string and candidates, features that described the database

Table 2: Distribution of correct wizard actions

Correct action	N	%
Return 1	2722	65.2445
Return 2	126	3.0201
Return 3	56	1.3423
Return 4	46	1.1026
Return 5	26	0.6232
Return 6	0	0.0000
Return 7	7	0.1678
Return 8	1	0.0002
Return 9	2	0.0005
Question | give up	1186	0.2843
Total	4172	1.0000

search returns, features that described confidence scores from various system components, and features that described the speech signal. This left 28 features. Before training each model we also ran CfsSubsetEval, an attribute selection algorithm that evaluates subsets of features based on both their individual predictive power and the degree of redundancy among them (Hall 1999). This further reduced the number of features to between 8 and 12 per model. (Many of the same features survived into more than one model.) To reduce overfitting, we also activated pruning to remove subtrees likely to provide little additional power because they cover too few training instances.

To confirm the learnability and quality of the decision trees, we also trained logistic regression and linear regression models on the same data. Here, regression captures the change in wizard action based on the changes in feature values (Witten and Frank 2005). Linear regression fits data to a linear function, and represents the wizard's four actions numerically in decreasing value: confident choice, tentative choice, question, and give up. Logistic regression predicts the probability of an action based on fit to a logistic curve. This generalizes the linear model to predict categorical data, here, the wizard's four actions. All models were produced with the Weka data mining package (Hall, et al. 2009) under 10-fold cross-validation.

Ability to predict wizard action was uniform across learning methods. On the overall model, logistic regression had 75.2% accuracy while the decision tree's accuracy was 82.2%. The linear regression model had root mean squared error of 0.483, while the decision trees' was 0.306. Predictive ability for the individual wizard models was similarly comparable. Thus the remainder of this discussion is restricted to decision trees.

Table 3 describes the learned models for individual wizards (ranked by wizard accuracy from Table 1). It shows size in number of nodes, number of included features, accuracy, and the F measure on confident choice. Note that model accuracy does not correlate with wizard rank; model accuracy indicates only how well the tree predicts the wizard's action from the training data. The simplest wizard strategies (e.g., always select the top candidate) are clearly easier to predict, but not necessarily better. (Compare, for example, W4 and W6.)

Recall from Figure 1 that confident choice was more common than tentative choice, which was in turn more common than question or give up. As a result, the individual models consistently predicted a confident choice with $0.80 \leq F \leq 0.87$, but less consistently predicted tentative choices with $0.60 \leq F \leq 0.89$, and could predict question only for W4, the top-scoring wizard who most often asked questions.

Table 3: Learned decision trees model individual wizards.

Tree	Rank	Size	Features	Accuracy	F conf
W4	1	55	12	75.67	0.85
W5	2	21	10	76.17	0.85
W1	3	7	8	80.44	0.87
W7	4	45	11	73.62	0.83
W3	5	33	10	77.42	0.84
W2	6	35	10	78.49	0.85
W6	7	23	10	85.19	0.80

The features that appeared most often in the individual models primarily described the database return, the ASR string's similarity to the candidates, the wizard's recent performance, and the quality of the speech recognition and language understanding. (Note that the last two were not available to the wizard.) The five features that appeared most often at the root or top-level nodes were

- *ReturnType* (Singleton, AmbiguousList, NoisyList)
- *RecentSuccess*, how often the wizard had chosen the correct title within the last three title cycles
- *ContiguousWordMatch*, the maximum number of contiguous word matches between a candidate and the ASR string (averaged across candidates)
- *NumberOfCandidates*, how many candidates were returned by the database
- *Confidence*, an Olympus/Ravenclaw metric on confidence for recognition and language understanding

Careful inspection of the model for the most accurate wizard (W4) indicates that, if ReturnType was NoisyList, she asked a question. If ReturnType was AmbiguousList, her decision involved the five features above, plus the acoustic model score (another internal system measure that indicates the quality of the speech recognition), the length of the ASR string in words, the number of times the wizard asked the user to repeat, and the maximum size of the gap between matching words in the ASR string and the candidates. To further focus our analysis on W4's distinctive behavior, we trained an additional decision tree to model how W4 chose between selecting a title and asking a question. The resulting model on 600 data points (each corresponds to a title cycle) consisted of 37 nodes and 8 features, with F = .91 for selecting a title and F = 0.68 for asking a question. The root of this tree differs from all other wizard models — it is the number of frames (10ms speech segments used to produce the ASR string), a measure of the length of the ASR. On short ASR strings (as measured both in number of frames and number of words) with AmbiguousList or NoisyList returns, W4 asked a question when $RecentSuccess \leq 1$ or ContiguousWordMatch = 0, and the acoustic model score was low. (Short titles are more readily confused.) On long ASR strings, W4 asked a question only when $ContiguousWordMatch \leq 1$, $RecentSuccess \leq 2$, and either the return was a NoisyList, or Confidence was low and there was more than one candidate. In summary, the factors that drove W4 to ask a question include the length of the ASR string, the quality of the ASR transcription, the database return type, the similarity between the ASR string and the candidates, and how well she had performed on recent title cycles. These can all be captured by system-internal features.

Discussion and Future Work

As used here, wizard ablation embeds a wizard within an SDS to study her choices when placed in the same environment as a machine. Given noisy ASR and the results of a

database search, the best wizards do not always guess based on search return. Instead they sense that the knowledge they have is a poor fit with what the recognizer "heard." In that case, a good wizard infers that the correct title is not among the returned candidates, and asks a thoughtful question to move the dialogue forward. (The mystery book at the beginning of this paper, by the way, was the third title listed.)

Experiments like this provide insight into how people match noisy input with returns from database search. The experimental design led wizards to prefer the first candidate listed — they read it first, and it was typically correct if the return included the correct title. Thus a wizard's skill at finding the title when it is present is less noteworthy than W4's ability to question the relevance of all the candidates.

The focus here has been on a single book request by title. Current work extends this approach to full dialogue. Wizards will see ASR and query results, and will have a predefined set of system-actions from which to choose. Dialogue interactions will include greeting, user identification, and four book requests by author and catalogue number, as well as by title. In full dialogue, context will have more relevance and can be measured more realistically by metrics in addition to RecentSuccess, Analysis of wizards' questions from this experiment will motivate a pre-defined set of questions for wizards in the full dialogue study.

This work successfully learned models that predict wizard action primarily from system features. (The only prevalent wizard-specific feature was RecentSuccess, which is readily replaced by the system's recent success.) Similar learned models will be incorporated into CheckItOut. Our next experiment will train models to predict wizards' actions during full dialogue with our baseline version of CheckItOut, and then refine the system with the learned models. We predict that evaluation of the refined, wizard-informed CheckItOut will provide better performance.

Acknowledgments

This research was supported in part by the National Science Foundation under IIS-084966, IIS-0745369, and IIS-0744904. We thank the staff of the Heiskell Library, the Olympus/Ravenclaw developers at Carnegie Mellon, and our tireless undergraduate research assistants.

References

Aist, G. S., Allen, J., Campana, E., Gomez Gallo, C., Stoness, S., Swift, M. and Tanenhaus, M. K. (2007). Incremental dialogue system faster than and preferred to its nonincremental counterpart. *CogSci 2007,* 779-774. Nashville, Tennessee.

Bohus, D. (2004). *Error Awareness and Recovery in Task-Oriented Spoken Dialog Systems. Ph.D. Thesis*. Computer Science Carnegie Mellon University.

Bohus, D., Raux, A., Harris, T. K., Eskenazi, M. and Rudniky, A. I. (2007). Olympus: an open-source framework for conversational spoken language interface research. *Proceedings of Bridging the Gap: Academic and Industrial Research in Dialog Technology workshop at HLT/NAACL 2007,* 32-39.

Bohus, D. and Rudniky, A. I. (2003). RavenClaw: Dialog Management Using Hierarchical Task Decomposition and an Expectation Agenda. *Proceedings of Eurospeech 2003,* 597-600.

Dix, A., Finlay, J., Abowd, G. D. and Beale, R. (2003). *Human-Computer Interaction*, Prentice Hall.

Forbes-Riley, K. and Litman, D. (2009). Adapting to Student Uncertainty Improves Tutoring Dialogues. *Proceedings of the 14th International Conference on Artificial Intelligence in Education, AIED,* 33-40. Brighton, UK.

Franceschetti, D. R., Adcock, A. B. and Graesser, A. C. (2003). Analysis of strategies in expert tutoring dialog for use in Intelligent Tutoring System Development. *CogSci 2003,* 1344. Boston, Massachusetts.

Hall, M. (1999). *Correlation-based Feature Selection for Machine Learning*. *Ph.D. Thesis*. Department of Computer Science University of Waikato.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I. H. (2009). The WEKA Data Mining Software: An Update. *SIGKDD Explorations* 11(1).

Levin, E. and Passonneau, R. (2006). A WOZ variant with contrastive conditions. *Proceedings of the Interspeech Satellite Workshop, Dialogue on Dialogues: Multidisciplinary Evaluation of Speech-Based Interactive Systems,* 17-21.

Ohlsson, S., Corrigan-Halpern, A., Di Eugenio, B., Lu, X. and Glass, M. (2003). Explanatory Content and Multi-Turn Dialogues in Tutoring. *CogSci 2003,* 48. Boston, Massachusetts.

Passonneau, R., Epstein, S. L. and Gordon, J. B. (2009). Help Me Understand You: Addressing the Speech Recognition Bottleneck. *AAAI Spring Symposium on Agents that Learn from Human Teachers,* 119-126. Paolo Alto, CA.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*, Morgan Kaufmann.

Ratcliff, J. W. and Metzener, D. (1988). Pattern Matching: The Gestalt Approach. *Dr. Dobb's Journal* 7, 46.

Raux, A. and Eskenazi, M. (2004). Non-Native Users in the Let's Go!! Spoken Dialogue Systems: Dealing with Linguistic Mismatch. *HLT/NAACL,* 217-224. Boston, MA.

Rieser, V. and Lemon, O. (2006). Using Machine Learning to Explore Human Multimodal Clarification Strategies. *COLING/ACL-06,* 659-666. Sidney, Australia.

Skantze, G. (2005). Exploring human error recovery strategies: Implications for spoken dialog systems. *Speech Communication* 45(3), 325-341.

Stoyanchev, S. and Stent, A. (2009). Predicting Concept Types in User Corrections in Dialog. *EACL Workshop SRSL,* 42-49.

Witten, I. H. and Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. San Francisco, Morgan Kaufmann.

Spatial Reasoning as Verbal Reasoning

Antje Krumnack (antje.krumnack@psychol.uni-giessen.de)
Experimental Psychology and Cognitive Science, Otto-Behaghel-Strasse 10F
35394 Giessen, Germany

Leandra Bucher (leandra.bucher@psychol.uni-giessen.de)
Experimental Psychology and Cognitive Science, Otto-Behaghel-Strasse 10F
35394 Giessen, Germany

Jelica Nejasmic (jelica.nejasmic@psychol.uni-giessen.de)
Experimental Psychology and Cognitive Science, Otto-Behaghel-Strasse 10F
35394 Giessen, Germany

Markus Knauff (markus.knauff@psychol.uni-giessen.de)
Experimental Psychology and Cognitive Science, Otto-Behaghel-Strasse 10F
35394 Giessen, Germany

Abstract

We introduce an approach for how spatial reasoning can be conceived as verbal reasoning. We describe a theory of how humans construct a mental representation given one-dimensional spatial relations. In this construction process objects are inserted in a dynamic structure called a "queue" which provides an implicit direction. The spatial interpretation of this direction can be chosen freely. This implies that choices in the process of constructing a mental representation influence the result of deductive spatial reasoning. To derive the precise rules for the construction process we employ the assumption that humans try to minimize their cognitive effort, and a cost measure is introduced to judge the efficiency of the construction process. From this we deduce how the queue should be constructed. We discuss empirical evidence for this approach as well as a computational implementation of the construction process.

Keywords: Verbal Reasoning; Spatial Reasoning; Mental Models; Cost Function; Computational Framework

Introduction

One dimensional spatial relations like "right of", "left of", "in front", "behind", "north of" have in common that they are transitive and, thereby, allow us to create a linear order between objects linked by one of these relations. Let us demonstrate this by an example. Consider the following two sentences, also called premises.

1. The apple is to the left of the mango.
2. The mango is to the left of the kiwi.

These premises allow us to create a linear order of the objects named in the premises, apple–mango–kiwi. This order enables us to draw conclusions about information not directly given in the premises: we can infer that the apple is to the left of the kiwi. The ability to infer information about relations between objects not explicitly yielded by the premises is the subject of theories about relational reasoning (cf. Johnson-Laird & Byrne,1991; chapter 5). The bases of such inferences are internal representations that reflect information conveyed verbally by the premises. There are several theories on how this is accomplished. Syntactic-based approaches (Braine & O'Brien, 1998; Rips, 1994; Hagert, 1984; Henst & Schaeken, 2005) suggest that the reasoning process is based on operations similar to the syntactic rules of formal logic. A set of rules is applied to draw inferences from given premises in order to derive new information implicitly provided by the premises. Model-based approaches, such as the mental model theory (MMT) on the other hand, suggest that reasoners infer new information by inspecting a mental model, representing the "state of affairs", described by the premises (Johnson-Laird & Byrne, 1991).

Polk and Newell (1995), however, point out that the deduction process does not necessarily require deduction-specific mechanisms to operate on internal representations. Especially in reasoners that are not specifically trained on deductive reasoning more general cognitive mechanisms might guide the reasoning process. They introduced a third approach, called **verbal reasoning**, that assumes the cognitive processes in deductive reasoning to be based upon the same processes as language comprehension and generation. Verbal reasoning describes reasoning as transformation of verbal information provided by the premises of an inference problem. Linguistic skills operate in order to encode and re-encode a reasoning problem until the conclusion becomes obvious or until the reasoner gives up. Polk and Newell (1995) hypothesize that when task-relevant information is provided verbally, the crucial role in reasoning is played by the verbal processes of encoding and re-encoding accordingly and that inferences follow comparatively easily from the encoded information.

In the following, we sketch how spatial reasoning can be conceived in Polk and Newell's framework of verbal reasoning, which covers reasoning about relations. In particular, we propose new theoretical assumptions for the special case of reasoning with spatial relations. The key assumption is that the process of constructing a mental representation – a mental model – from the premises influences deductive spatial reasoning. This implies that the process of encoding the

spatial information is critical for the result of the spatial reasoning process. We discuss empirical evidence as well as a computational implementation of the process.

A cognitive model

We are proposing a theory on how humans create a mental model from a set of spatial relations. The theory is based on the idea of cognitive efficiency, that is humans try to minimize their cognitive effort, therefore a cost measure is introduced to judge the efficiency. From this we derive how a mental model should be constructed. This mental model can then be used to reason about spatial relations and its properties imply consequences for the reasoning process.

Basic assumption for the cognitive model

Since we consider arbitrary one-dimensional relations as basis for the model we assume that models consist of a "queue" of objects and an interpretation what this queue represents. The queue describes in which order the objects are aligned but what this order represents depends on the relation that is considered. So while the order is implicit the interpretation of the order is not. The queue is constructed by forming links between objects. The links signify which objects follow each other in that ordered arrangement. These links between the objects are one directional which means that when inspecting the queue we can move from one object to the next object in the order but not to the preceding object. To access the queue one needs to access the first element of the queue. Therefore the beginning of the queue is marked by a start pointer.

The queue can be accessed from this starting point which is directed at the first object. From there all other objects in the mental model can be reached by following the links between objects.

This amounts to the following assumptions about the queue:

1. There exist a starting point or first object.
2. Each object is linked to the next object in the linear order. Only the last object is not linked to other objects.
3. While this structure has an implicit direction, the interpretation of this direction depends on the context.

Mental Cost

We now introduce a cost measure that allows us to judge how to create the queue efficiently. The main assumption is that an existing link should not be broken if that is avoidable and as few new links as possible should be formed to minimize cognitive work.

So a cost efficient model is one that can be built by a minimal number of broken links. Since in the end of the construction process the complete mental model is supposed to have as many links as objects costs can only be reduced by altering as few links as possible during the construction process. Therefore it is most cost efficient if we can insert new objects creating just one new link and not changing any existing links. The only way to accomplish this is by attaching them at the very end, following the last object in the queue. So if an object can be inserted at the very end of the queue it should be inserted there.

The starting point is also considered a link. This is due to the fact that one has to know how the queue starts in order to access it. Therefore knowing which object is the first constitutes a link, connecting the start of the queue to that object.

Moving through the queue on the other hand takes very little cognitive effort as long as we move in the implicit direction of the queue. Due to the fact that the links only have one direction moving in the opposite direction through the queue is impossible.

Construction of mental models from spatial information

The question now is how a mental model is constructed from the premises of a reasoning problem. How are objects inserted in the queue and where does the cost measure come into play?

In this process the first premise that is considered has a special function and dominating effect on the construction of the rest of the arrangement. We consider the first premise independently of the following premises and postulate the following two rules for the construction process.

1^{fp} First object inserted in the queue is the starting point of the queue.

2^{fp} The second object is linked to the first object. The relation between the first and the second object thereby creates the interpretation of the link and the implicit direction of all the following objects in the queue.

If we know, for example, that the second inserted object is supposed to be to the right of the first (starting) object, then the link is interpreted as "to the right".

When we look again at our example from the introduction this gives us two options for the first premise: "The apple is to the left of the mango." We can choose the apple as the starting point (marked by the asterisk) and insert the mango second:

$$\text{apple}^* \rightarrow \text{mango}$$

The implicit direction of the queue is interpreted as moving from the leftmost object to the right. However, if we use the mango as a starting point (marked by the asterisk) inserting the apple second we get:

$$\text{apple} \leftarrow \text{mango}^*$$

In this case the implicit direction of the queue is interpreted as moving from the rightmost object to the left. So even though the premise describes only one arrangement of fruits there are two options for representing this arrangement in our queue.

Once the interpretation of the implicit direction of the queue is fixed by inserting the second object the rest of the

objects are inserted according to this interpretation. This amounts to the following options for inserting objects in an existing queue from the second premise on:

1. The first object of the premise has to be found in the queue.

2.(a) If the new object is to be placed behind this object (with regard to the implicit direction of the queue) it can be either inserted into the queue directly behind the object or at any point further to the end of the queue.

(b) If the new object is to be placed in front of the object (with regard to the implicit direction of the queue) it can be either inserted into the queue directly in front of the object or at any point further to the beginning of the queue.

The question is which of these choices is more cost efficient. As a cost measure we use primarily the number of links that need to be formed. If this does not show any difference between the options the required movement through the queue is used as an secondary cost measure.

Let us first look at the costs resulting from inserting a new object into the queue between two objects that are linked. To insert a new object between two existing objects in the queue the first object, that was linked to the second object before, is now linked to the new object. The new object is linked to the second object. This requires forming two new links. If the object is inserted at the beginning of the queue the starting point needs to be redefined which we will consider as creating a new link.

Using this information we will now judge the cost created by the insertion options described in 2.(a) and (b). Let us first look at option (a): If the object is inserted between two objects of the queue two new links need to be formed. If the object is inserted at the end of the queue, only one new link needs to be formed. So in case (a) it is most cost efficient to insert the object at the very end of the queue. Now we consider (b): The new object can only be inserted between two objects or at the starting point of the queue. Since we consider the starting point a link to the beginning of the queue both options require two new links to be formed. So it is the most cost efficient not to move around the queue but to insert the object directly in front of the found object. Using this analysis we postulate the following rules:

1^{ins} If the new object is to be placed behind an object of the queue it will be inserted *at the end* of the queue.

2^{ins} If the new object is to be placed in front of an object of the queue it will be inserted into the queue *directly in front* of this object.

If we apply these rules to the second premise of the first example we create one of the following two models depending on the direction of the queue.

$$\text{apple}^* \rightarrow \text{mango} \rightarrow \text{kiwi} \qquad (1)$$

$$\text{apple} \leftarrow \text{mango} \leftarrow \text{kiwi}^* \qquad (2)$$

While the results look similar, the costs for building these models differ. In case (1) we were able to use rule 1^{ins}, creating only one more link. In case (2) however, we needed to redefine the starting point. This resulted in creating two new links. So the cognitive costs for building the first model are lower.

Let us look at another example that is not quite as simple:

1. The apple is to the left of the mango.

2. The apple is to the left of the kiwi.

Here the premises describe an indeterminate order: there are two possible orders of these three fruits: apple–mango–kiwi and apple–kiwi–mango. So the question is, whether one of these orders is preferred over the other? Knauff, Rauh, and Schlieder (1995); Rauh et al. (2005); Jahn, Knauff, and Johnson-Laird (2007) have empirically shown that such preferences exist in human reasoners.

Since the first premise is the same as in the example with the determinate order we receive the same two options for models when applying the rules for the first premise. If we apply the rules of insertion to the second premise we get one of the following models.

$$\text{apple}^* \rightarrow \text{mango} \rightarrow \text{kiwi} \qquad (3)$$

$$\text{apple} \leftarrow \text{kiwi} \leftarrow \text{mango}^* \qquad (4)$$

Here we see a difference between the two models depending on the implicit direction of the queue. This is due to the fact that the arrangement is indeterminate and because the two queues have opposite interpretations of the implicit direction different rules are applied to form the queues. There is also a difference in the cost for building these models. In (3) we were able to apply rule 1^{ins}, again creating only one new link. In (4) we needed to apply rule 2^{ins}, redefining the starting point, creating two new links. So the cognitive costs for creating the last model (4) are higher than the ones for creating model (3).

Once a model has been constructed it can be used to make inferences. If we build the model

$$\text{apple}^* \rightarrow \text{mango} \rightarrow \text{kiwi}$$

from the premises of the first example we can answer the question "Is the apple to the left of the kiwi?" by finding the apple in the queue and then moving further down the queue till we find the kiwi.

Predictions based on the construction process:

From the model we can derive several behavioural predictions:

- If the model is indeterminate (allowing more than one model) the direction of the queue influences which model will be built.

- It should be easier to infer information that can be obtained following the implicit direction of the queue than infer information that require to go against that direction.
- The same mechanism is used for all one dimensional spatial relations, not just in the left/right direction.

Computational implementation and computational complexity estimation

The model construction process can be easily implemented as a computer model using the data structure linked list, consisting of nodes containing data and a pointer to the next node in the list. There is also a start pointer pointing at the first node of the list. If we compare that to our mental model the pointers from one node in the list to the next represent the link between the objects. The data represent the objects. It is therefore easy to model a queue such as the one we proposed in a computer program.

Computer science provides standard rules for analysing the efficiency of algorithms. However, the traditional cost analysis of the algorithm used to insert objects into a linked list provides different results than the above cost analysis. This is because in computational complexity theory every operation has the same weight. There are no operations that are harder or easier to perform than other operations.

Let us look again at the possibilities for insertion in an existing queue from above, 2(a) and (b). Which of these options is the most cost efficient? When inserting a node behind a node of the list as in 2(a), and we insert it directly behind the found node, we have the cost of assigning one pointer and reassigning another (if not the end of the list). If we move further down the list, the costs of moving through the list have to be added to the costs of assigning pointers and moving one node down the list costs as much as assigning one pointer. So inserting a node between two nodes further down the list is always more expensive than inserting it right behind the found node. Attaching the node to the end of the list is not a good idea either: if the end of the list is more than one node away, the cost of moving through the list and assigning the pointer will be higher than the cost of just inserting the new node right behind the found node. And since there is no way of knowing how far away the last node is, the cost efficient solution is to insert the new node right behind the found node.

When inserting a node in front of a found node of the list as in 2(b), the same costs result for inserting the new node right in front of the node and for inserting it at the beginning of the list as the starting pointer of the list can always be accessed at no extra cost. In both cases one pointer needs to be assigned and one pointer needs to be redirected. If it is inserted at any other point of the list the costs are higher since we first have to move to that point from the beginning of the list. So in this cost analysis it would be most efficient to insert the object either at the beginning of the list or directly in front of the found object.

Based on this analysis we derive alternative rules for inserting nodes into a list:

1^{alt} If the new node has to be placed behind a node of the list it should be inserted into the list *directly behind* the node.

2^{alt} If the new node has to be placed in front of a node of the list it should be inserted into the list *either directly in front* of this node or at the *very beginning* of the list.

If we apply these rules to the second premise of the second example we receive the following models.

$$\text{apple}^* \rightarrow \text{kiwi} \rightarrow \text{mango} \qquad (5)$$

$$\text{apple} \leftarrow \text{mango} \leftarrow \text{kiwi}^* \text{ OR apple} \leftarrow \text{kiwi} \leftarrow \text{mango}^* \quad (6)$$

The model (5) was built using rule 1^{alt}, the models in (6) are the two options following from rule 2^{alt}. The insertion of the last object has the same computational cost in all three of these models.

The models also show that rules based on a classic computational cost measure produce different results than our rules based on a cognitive cost measure. Model (5) differs from model (3) above and only one of the models of (6) is similar to the model (4). Of course this does not mean that a computational model would have to follow the alternative rules. It can also be implemented using the insertion rules that resulted from the cognitive model.

So one of the questions is whether it is justified to assume that forming a link is more cost intensive than moving through the queue. If not, the traditional computational complexity measure might provide better predictions than our model.

Empirical Evidence

We report an experiment that shows that rules derived from our cognitive cost measure predict human behaviour better than the rules derived from the traditional computer science cost measure. In this experiment we investigated what kind of mental model participants construct when they are faced with indeterminate problems that allowed more than one model to be constructed.

Participants, Materials, Procedure, and Design

Thirty-five participants (3 male; age: $M = 22.4$; $SD = 3.2$) from the University of Giessen had to solve sixteen determinate (like in example 1) and sixteen indeterminate problems (like in example 2). The three-term problems had two premises each and we used only the relation "left of". The problems were presented to the participants in a random order on a computer screen. Each premise was presented sequentially (in a self-paced manner). After having read the premises a conclusion was presented and the participants were asked if this conclusion was correct or not. For determinate problems the conclusion was either true or false. For indeterminate problems we used two different types of conclusions which could either hold in a model constructed according to rule 1^{ins} or to rule 1^{alt}.

Results and Discussion

Separate ANOVAs for the percentage of correct responses and reaction times of correct responses with the within-subject factor conclusion acceptance (hits, correct rejections) and insertion principle (indeterminate/rule1^{ins} vs. indeterminate/rule1^{alt}) were calculated, respectively. Level of significance was 5%.

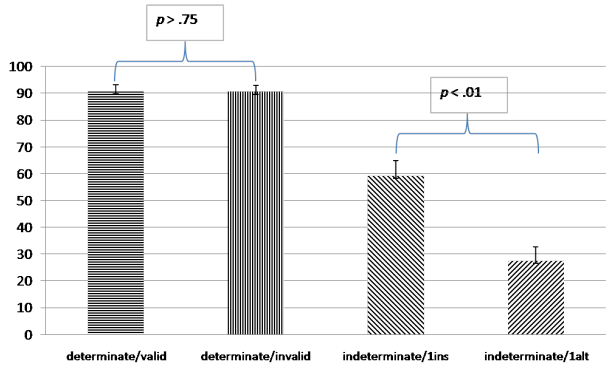


Figure 1: The left two bars show the mean number of correct responses for the determinate problems. In half of the problems the correct response was "yes" (hit), in the other half it was "no" (correct rejection). The right two bars show how often the participants accepted a conclusion that hold in the model built by rule 1^{ins} or rule 1^{alt}, respectively. Error bars indicate standard errors.

ANOVA of the percentage of correct responses yielded a significant main effect of conclusion acceptance [$F(2,32) = 54.79$, $p < .01$]. Percentage of correct responses of determinate/valid and determinate/invalid items did not differ ($p > .75$). The high percentage of correct responses for the determinate items ($M = 92.19$; $SD = 11.14$) indicate that the participants understood the task and were able to perform well. Because the determinate items were easiest constructed from the left to the right for both cost functions we assume that they were indeed constructed from left to right. We also assume that the indeterminate items were constructed from the left to the right as well, since the decision has to be made directly after reading the second premise before knowing whether the item is determinate or indeterminate. We find a higher percentage ($t(34) = 5.49$; $p < .01$) of yes-answers for the items where the conclusion was true if the model was built by rule 1^{ins} than for items where the conclusion was true if the model was built by rule 1^{alt} (see Figure 1). This indicates that indeed the rules derived from our cognitive cost functions are more often applied than the alternative rules derived from the classical computer science cost function.

ANOVA of the reaction times of correct responses also yielded a significant main effect of conclusion acceptance [$F(2,18) = 4.25$, $p < .05$]. Reaction times for determinate/valid items ($M = 3618$ ms, $SD = 1427$) were significantly lower compared to determinate/invalid items ($M = 4887$ ms, $SD = 2691$; $t(34) = -4.67$; $p > .01$). Reaction times for indeterminate/rule1^{ins} items ($M = 4156$ ms, $SD = 3066$) were significantly lower compared to indeterminate/rule1^{alt} items ($M = 5057$ ms, $SD = 3457$; $t(20) = -2.29$; $p > .05$). This implies that conclusions of the determinate/valid items were easier to confirm than the ones of the determinate/invalid items and the conclusions of the indeterminate/rule1^{ins} items were easier to accept than the ones of the indeterminate/rule1^{alt} items. These easier items were those where the confirmation could easily be made by following the implicit direction of the queue provided that the queue was indeed constructed from left to right.

Other evidence

Further evidence for our model comes from the experiments of Jahn et al. (2007). Their participants inserted an object to an existing array, as opposed to adding it to one end of the array, more often for objects that would have been added to the left end of an array than for entities that would have been added to the right end of an array (Jahn et al., 2007, Experiment 2, Table 4). The authors come to the conclusion that: "Given that the participants constructed arrays from left to right, they evidently found it easier to add a new entity to the right-hand end of an array than to the left-hand end of an array [...]." (Jahn et al., 2007, p. 2081)

For a queue that is constructed from left to right our model predicts this behaviour, since rule 1^{ins} is applied to the objects inserted on the right of a reference object and rule 2^{ins} is applied to objects inserted on the left of a reference object. So the results of Jahn et al. (2007) confirm the predictions of our model.

Discussion

We introduced an approach how spatial reasoning can be modelled as verbal reasoning. The main idea is that the deduction process does not necessarily require deduction-specific mechanisms to operate on internal representations. Instead we assume that a simple order of objects (represented by words) and some genuine verbal cognitive mechanisms might guide the reasoning process. Following Polk and Newell (1995) we assumed that the cognitive processes in deductive reasoning can be based upon the same processes as language comprehension and generation.

From our point of view our approach is a helpful addition to the long lasting controversy between models and rules in reasoning (e.g. Johnson-Laird, Byrne, & Schaeken, 1994; Rips, 1994; Hagert, 1984). In fact, models are often identified with visuo-spatial processing and rules with linguistic or sentential mechanisms (e.g. Goel, Buchel, Frith, and Dolan (2000)). Our study, however, shows that this distinction does not reflect the actual differences between the two approaches. In fact, our approach is a model-based approach, because at no point during the inference process rules of inferences are used, instead the new information must be derived from the queue - the model. And our results suggests that such models can be the basis of verbal reasoning, so no visuo-spatial

process are involved in the inference.

Our work has also shown that the approach and the related cost measure leads to good predictions about what kind of model will be created. It predicts behaviour better than the classical computer science approach to cost calculation. But there remain some open questions about the cost measure. One problem of our approach results from the fact that it is easier to move through the queue than alter the existing links, no matter how far we have to move. Another possibility is that if the queue becomes larger there might exist a critical distance when it requires more mental effort to move this distance through the queue than altering a link. This would imply that if the queue reaches a certain number of objects, new objects would not necessarily be attached to the end of the queue any more.

Another question is whether the starting point of a queue is really a link like all the other links in the queue. However, since this link is different concerning its cognitive nature it might be weaker or stronger than the links between object in the queue.

A third limitation of our project is that we only used problems with two premises, although we believe that the postulated rules also apply if there are more than two premises and three objects, as long as the premises all contain relations describing the same dimension. And it is possible to mix relations of the same dimension such as left and right, as done in many experiments (Jahn et al., 2007; Ragni, Fangmeier, Webber, & Knauff, 2006).

Finally, we postulate that the implicit direction of a queue can be chosen freely. But what is this choice based on? In a behavioural experiment, in which many spatial reasoning problems need to be solved, subjects are likely to choose the direction that produces the lowest cost for the items seen so far. Also, once a choice has been made on which direction to use, subjects are likely to stick with it. This also keeps the mental costs low because the tactic used is not constantly being analysed. Also, when using material with a left-right dimension as in the examples there seems to be a general preference for constructing a left to right queue (Jahn et al., 2007; Rauh et al., 2005; Ragni et al., 2006). This could be a cultural preference such as reading. Also other orders we see in daily life are arranged left to right.

Overall, we were able to present some evidence for our assumption that the process of constructing a "verbal mental model" from premises influences deductive spatial reasoning. The chosen interpretation for the implicit direction of the queue has consequences on what kind of conclusions can be easily made. And, most importantly, for indeterminate problems, we can predict which model is preferred over the other and and which model is more difficult to consider as a possible interpretation of the premises. While our model can not necessarily be generalized to other domains of reasoning we feel that it can describe some aspects of human reasoning with spatial relations and that it demonstrates that spatial reasoning can also be conceived as verbal reasoning.

Acknowledgments

This work was supported by a grant of the Deutsche Forschungsgemeinschaft (DFG) to M. Knauff und B. Nebel (project KN 465/6-1).

References

Braine, M. D. S., & O'Brien, D. P. (Eds.). (1998). *Mental logic*. Mahwah, NJ: Erlbaum.

Goel, V., Buchel, C., Frith, C., & Dolan, R. J. (2000). Dissociation of mechanisms underlying syllogistic reasoning. *NeuroImage*, *12*, 504-514.

Hagert, G. (1984). Modelling mental models: Experiments in cognitivemodelling of spatial reasoning. In *Proceedings of the sixth european conference on artificial intelligence* (pp. 179–188).

Henst, J.-B. V. der, & Schaeken, W. (2005). The wording of conclusions in relational reasoning. *Cognition*, *97*, 1-22.

Jahn, G., Knauff, M., & Johnson-Laird, P. N. (2007). Preferred mental models in reasoning about spatial relations. *Memory and Cognition*, *35*, 2075–2087.

Johnson-Laird, P. N., & Byrne, R. (1991). *Deduction*. Hove (UK): Erlbaum.

Johnson-Laird, P. N., Byrne, R. M. J., & Schaeken, W. (1994). Why models rather than rules give a better account of propositional reasoning. *Psychological Review*, *101*, 734–739.

Knauff, M., Rauh, R., & Schlieder, C. (1995). Preferred-mental models in qualitative spatial reasoning: A cognitive assessment of allen's calculus. In *Proceedings of the seventeenth annual conference of the cognitive science society* (pp. 200–205). Mahwah, NJ: Erlbaum.

Polk, T. A., & Newell, A. (1995). Deduction as verbal reasoning. *Psychological Review*, *102*, 533–566.

Ragni, M., Fangmeier, T., Webber, L., & Knauff, M. (2006). Complexity in spatial reasoning. In *Proceedings of the 28th annual cognitive science conference* (pp. 1986–1991). Lawrence Erlbaum Associates.

Rauh, R., Hagen, C., Knauff, M., Kuss, T., Schlieder, C., & Strube, G. (2005). Preferred and alternative mental models in spatial reasoning. *Spatial Cognition Computation*, *5*, 239-269.

Rips, L. J. (1994). *The psychology of proof*. Cambridge, MA: MIT Press.

Arbitrating Between Theory-Theory and Simulation Theory: Evidence from a Think-Aloud Study of Counterfactual Reasoning

Meredith R. Wilkinson (m.r.wilkinson@lancaster.ac.uk)
Department of Psychology, Lancaster University
Lancaster, LA1 4YF UK

Linden J. Ball (l.ball@lancaster.ac.uk)
Department of Psychology, Lancaster University
Lancaster, LA1 4YF UK

Rachel Cooper (r.v.cooper@lancaster.ac.uk)
Department of Philosophy, Lancaster University
Lancaster, LA1 4YG UK

Abstract

How we engage in mental state reasoning remains a contentious issue, reflected in the debate between theory-theorists, who argue that we deploy theory-based rules, and simulation theorists, who argue that such reasoning is subserved by simulation processes. The present study examined whether theory-based or simulation-based reasoning is adopted in regret-oriented counterfactual scenarios involving mental state inferences. Participants thought aloud while reasoning through such scenarios from the perspective of themselves, themselves and another, or two other individuals. The scenarios also manipulated the controllability of event outcomes. Results revealed more theorizing in the uncontrollable than the controllable scenarios, and more simulation in the controllable than uncontrollable ones. More theorizing was also observed in the "other-and-other" than the "self-only" condition. These findings highlight the value of adopting a hybrid model of mental state reasoning, where theorizing and simulation are integrated within a common framework, with such processing being deployed in a context-sensitive manner.

Keywords: Simulation theory; theory-theory; counterfactual thinking; regret; think aloud protocols; mental models.

Introduction

The theory-theory (TT) versus simulation theory (ST) debate focuses on how we understand and reason about our own and others' *mental states* relating to beliefs, desires, intentions and the like. The TT approach (e.g., Carruthers, 1996) argues that we apply both tacit and non-tacit "theories" when understanding mental states. In contrast, ST posits that such understanding is attained either: (i) by "offline" simulation (e.g., Goldman, 2006), in which we take our own beliefs and desires offline, input those of another individual, and run a simulation process; or (ii) by imagining how we would feel in a given situation and by assuming that since other individuals are similar to ourselves then they would feel the same as we do (e.g., Gordon, 1986). Recently, a new variety of theorists have emerged who argue for a *hybrid* approach, in which both theory and simulation are adopted, dependent on the situation (e.g., Mitchell, Currie, & Ziegler, 2009). As we will demonstrate, this hybrid approach has considerable appeal, not least because it can capture the way in which mental state reasoning is sensitive to a multiplicity of factors associated with the prevailing situation, such as its familiarity. The factor of interest in the present paper relates to the protagonist's capacity to have control over the outcome associated with a situation.

Previous Attempts to Arbitrate Between TT and ST

Since the emergence of the TT/ST debate, psychologists and philosophers have been keen to find a test case to arbitrate between these accounts. This pursuit has largely focused on empirical findings that derive from comparisons between autistic individuals (who have various deficits in mental state understanding) and those without autism. The debate surrounding the correct theoretical interpretation of mental state reasoning deficits in autism appears to have reached an impasse (e.g., Carruthers, 1996, has argued that evidence from autism supports TT, while Goldman, 2006, argues that it corroborates ST), such that researchers have started to explore other ways to address the TT/ST debate.

Recent work in this latter vein reported by Kühberger, Kogler, Hug, and Mösl (2006), examined the TT/ST debate using the "position effect" (i.e., a bias to select the rightmost object in an array of identical objects when asked to make a preference judgment). In Experiment 1, participants observed a target person viewing a line of pantyhose and had to imagine viewing the items in the same manner. It was found that participants could predict the target's preference for the rightmost item, which seems to support ST since the availability of sufficient imaginative input (i.e., reasoning from the perspective of the target) enabled participants to predict the position bias. In Experiment 2, participants were given a verbal description about an actor rather than observing them, which failed to produce the results of Experiment 1. In Experiment 3, participants were informed about the position effect (but what it entailed was not explained), and were told to ignore it when making their selection. The fact that most people still demonstrated the

bias was interpreted as indicating deployment of an incorrect theory.

Although, Kühberger et al. claim their results support ST, we wonder whether they speak more to a hybrid account in which TT operates under some conditions and ST under others (e.g., Experiment 3 suggests an incorrect theory can be overridden by simulation). We also note a limitation of the study, which is that it focuses on epistemic mental states (i.e., beliefs), which are rather divorced from the richness of everyday mental state understanding. In order for TT and ST concepts to be useful for examining mental state reasoning (rather than just the currency of an esoteric philosophical debate; cf. Ratcliffe, 2007), such concepts must show general applicability to a wide variety of everyday mental state reasoning contexts.

In this paper we propose that counterfactual reasoning about mental states may provide a new test case for arbitrating between TT and ST. Counterfactual reasoning involves imagining how events associated with regret or disappointment could have turned out differently. Such reasoning is commonplace in everyday life and is vital for understanding how other people may be feeling in response to negative outcomes of real-world situations.

Mental Models and Counterfactual Thinking

The conceptual analysis of counterfactual reasoning is currently dominated by those adopting a mental models framework (e.g., Byrne, 2002, 2005), where mental models reflect representations of actual and counterfactual possibilities. Two particular phenomena in counterfactual reasoning that have been addressed by mental model theorists are: (i) the "action effect", which concerns the observation that greater regret intensity is elicited by acts of commission than acts of omission in the short term (e.g., Kahneman & Tversky, 1982), with the reverse being evident in the long term (e.g., Gilovich & Medvec, 1994); and (ii) the "temporal order effect", whereby we are more likely to reason about undoing the final event in a sequence of events that led to a negative outcome (e.g., Byrne, Segura, Culhane, Tasso, & Berrocal, 2000).

Atkinson, Bell, and Feeney (2009) explored both effects in a study using regret-oriented scenarios. To examine the action effect participants were asked to decide which of two protagonists would feel more regret, an actor or non-actor. To investigate the temporal order effect, participants were asked who would feel worse, the actor who was mentioned first or second. The study also manipulated the time that participants had available to register their response: either they had as long as they wished or they had to answer as quickly as possible. It was found that there was no effect of response time on the emergence of the temporal order effect. However, the action effect was disrupted in the speeded condition, with the actor being selected significantly less often than in the delayed condition. The finding that the action effect and temporal order effect are differentially influenced by the response-time manipulation is claimed by Atkinson et al. to be a consequence of reasoners needing to build complex representations when displaying the action effect (cf. Feeney & Handley, 2006). In particular, for the action effect to arise the reasoner has to compare events associated with both the actor and the non-actor.

We concur with Atkinson et al.'s interpretation, and also believe that their findings are relevant to understanding the role of theorizing and simulation in counterfactual reasoning. The observation that participants readily select the second actor in the temporal order scenario, regardless of time constraints, suggests that participants may be applying a straightforward, theory-based "rule" that the second actor would feel more regret. For the action effect, however, we propose that with sufficient time participants are likely to engage in mental simulation to pursue comparisons between the levels of regret felt by the actor and non-actor - in line with Atkinson et al.'s claim that participants flesh out mental models in these cases.

We illustrate this latter point with reference to a classic action/inaction scenario (Gilovich & Medvec, 1994):

> "Dave and Jim do not know each other but both are enrolled at the same elite East Coast University. Both are only moderately satisfied where they are and both are considering transferring to another prestigious school. Each agonizes over the decision, going back and forth between thinking he is going to stay and thinking he will leave. They ultimately make different decisions: Dave opts to stay where he is and Jim decides to transfer. Suppose their decisions turn out badly for both of them: Dave still doesn't like it where he is and wishes he had transferred, and Jim doesn't like his new environment and wishes he had stayed".

When considering the mental models constructed when reasoning about Dave, Feeney and Handley (2006) argue that participants construct the actual state of affairs in which he stayed and was unhappy and the counterfactual state in which he moved and was happy. However, we propose that a third possibility may be constructed for the short term of Dave moving and being unhappy:

Actual: Stays – Unhappy [Regret]
Counterfactual: Moves – Happy [No regret]
Counterfactual: Moves – Unhappy [Regret]

When fleshing out this model set it is apparent that three possibilities have to be constructed. These would not be easy to derive from theory-based processing alone, which underpins our proposal that simulation may be necessary for participants to imagine options that the protagonist may be considering - along with their related emotional impact. We further propose that when individuals are presented with a counterfactual scenario they will typically evoke a two-stage reasoning process. The first stage involves bringing to mind an initial model based on theory-driven processing (e.g., "If failing to take an action turns out badly then one will feel regret"). Assuming that a response can be generated at Stage 1 without any perceived need for further processing then this will be done on the basis of the initial model. However, if more processing seems to be needed then individuals may

engage in simulation. This Stage 2 process would be more cognitively effortful than theorizing as well as more sequential and controlled in nature.

Predictions of the Study

These aforementioned ideas allow us to develop predictions in relation to the experiment we report below. One focus of the research concerned the distinction between "controllable" and "uncontrollable" outcomes within regret-oriented counterfactual situations. Scenarios involving uncontrollable outcomes limit the consideration of how the outcome could have turned out better. We therefore predict that such scenarios will be susceptible to reasoning based on the application of theory-based inferences. For controllable outcomes, however, although theory-based reasoning is available as a starting point, there is also the potential for the reasoner to flesh out possible ways in which events could have turned out differently. We therefore predict increased simulation-based reasoning when participants engage in mental state understanding in the controllable case relative to the uncontrollable case.

Most philosophers have focused on arbitrating between TT and ST with reference to situations involving reasoning about another person. However, it is also interesting to examine people's reasoning about their own mental states. Evidence that there are differences in how we reason about ourselves compared to others comes from a study by Girotto, Ferrante, Pighin, and Gonzalaz (2007), who presented participants with a scenario where they could win a prize by solving a problem. One key experiment involved two conditions. In the "actor" condition participants were presented with two sealed envelopes; one they were told contained an easy problem, one a difficult one (in fact, both contained an insoluble problem). In the "reader" condition, participants read about a protagonist who had to make the same choice as in the actor condition, with an identical outcome (i.e., failure to solve the problem). Participants were either assigned to the actor or reader condition and were afterwards asked to name one way in which the outcome could have turned out better. Responses were coded as either modifying choices (e.g., selecting the other envelope), or as modifying problem features (e.g., having more time). Girotto et al. found the actors were more likely to alter problem features, while readers were more likely to alter a choice, such as the selection of the envelope.

The finding that participants reason about different aspects of a scenario in the self/actor versus other/reader condition suggests that different processes may be occurring. The increased likelihood of undoing a problem feature in the actor condition indicates the consideration of more possibilities than in the reader condition, which tended to involve just choosing the other envelope. This points to the idea that more possibilities are considered when reasoning from the perspective of oneself than that that of another, with the implication being that more simulation may arise in the former than the latter situation.

We also note that when people engage in counterfactual reasoning about another individual then theory-based reasoning may take precedence as people tend to possess a wealth of generalized rules concerning how people will feel in regret-oriented situations (e.g., "A person will be upset if they miss out on something they desire greatly"). However, when reasoning about ourselves we may be more likely to progress onto the simulation stage since we possess more specialized knowledge concerning ourselves and the nuances of our own reactions to events. In this way, people who are engaged in self-oriented reasoning may move away from the application of generalized folk psychological theories toward the simulation of multiple eventualities.

In sum, by using regret-oriented counterfactual scenarios involving mental state reasoning we assumed that we would gain useful insights to address two issues surrounding the TT/ST debate. First, such scenarios should usefully inform whether theory-based or simulation-based reasoning dominate in mental state understanding, or whether both forms of reasoning are deployed. Second, the manipulation of factors such as outcome controllability and the self/other distinction should clarify whether contextual and instructional aspects of the presented scenarios determine whether individuals are more likely to theorize or simulate.

Method

Participants

Participants were 90 individuals at Lancaster University who received either course credit or payment. None had prior knowledge of research on reasoning or theory of mind.

Design

A 2 x 3 mixed between-within participants design was adopted. The between participants factor was the perspective that participants had to reason from, which had three levels: self; self-and-other; other-and-other. The within participants factor was outcome controllability, with two levels: controllable versus uncontrollable.

Materials and Pre-Test

Participants received two controllable and two uncontrollable scenarios. Controllable scenarios concerned: (i) two individuals performing poorly on a University assignment (assignment scenario); and (ii) individuals changing or not changing a minor subject to a major at University and then not enjoying the course and receiving a poor course grade (course scenario). The uncontrollable scenarios concerned: (i) losing a game of table football by scoring an own goal, followed by one's opponent scoring a winning goal (football scenario); and (ii) an individual missing their flight by 5 mins, with the plane having been delayed, and another individual missing their flight by 30 mins, with the plane leaving on time (plane scenario).

In the self-only and self-and-other conditions the participant had to take on the role of one of the individuals within the scenario. In those conditions that involved other individuals, we presented "personas" (i.e., brief bio-sketches

of the named individuals) in an effort to increase the realism of the scenarios. For the self-and-other and other-and-other conditions participants were required to state who they thought would feel more regret, upset or frustration (dependent on scenario). For the self-only condition participants had to state how much regret/upset/frustration they would feel with essentially equivalent scenarios.

To validate our controllability manipulation we gave 13 participants the scenarios from the self-only condition. After reading each scenario they had to use a 10-point scale to rate it for familiarity in their everyday life, and for the controllability of the outcome. We also included a mutability question in which participants were asked simply to list all the ways in which the situation could have turned out for the better. Using paired samples t-tests we found that controllable scenarios were rated as significantly more controllable (M = 6.54) than the uncontrollable scenarios (mean = 5.31), $t(12) = 2.66$, $p = .02$. There was no difference, however, in ratings of familiarity (means of 5.58 versus 4.88 for controllable vs. uncontrollable), $t(12) = 1.17$, $p = .26$. For the mutability measure there was a mean of 2.38 mutations for controllable scenarios and 2.69 mutations for uncontrollable scenarios, which was unreliable, $t(12) = 1.67$, $p = .12$. Overall, these pre-test data reveal a solid effect of controllability in the predicted direction, but no confounding effects of mutability or familiarity.

Procedure

Participants were randomly assigned to one of the three perspective conditions and were given associated instructions. They were then presented with a booklet containing the scenarios and were asked to think-aloud whilst reasoning about each one. Scenario order was independently randomized for each participant.

Results

Data Coding

To code the data we adopted Ball and Christensen's (2009) scheme in which each line was coded as reflecting theory-based or simulation-based reasoning. An "ambiguous" code was used when: (i) lines were evenly split across categories (there were 13 instances of these); or (ii) it was difficult to be certain whether theory-based or simulation-based reasoning was being adopted. Theory-based reasoning concerned instances in which the participant adopted tacit or non-tacit theories to make inferences about their own or others' mental states. Such reasoning tended to involve the participant stating general rules regarding mental states, typically involving a grammatical construction such as "The person will feel x because of y". The following excerpt illustrates theorizing taking place, with a participant adopting a tacit rule that captures the notion that action will elicit greater regret in the short term than inaction:

> "I think Mike's gonna feel the more regret in the short term coz he's actually chan- he actually made a bad decision whereas Timmy's decided - Timmy's chosen not to make the decision, so he doesn't know whether or not he'd prefer the other - you can assume he can".

Simulation occurred when participants took their own beliefs and desires offline and inputted those of other individuals (e.g., Goldman, 2006). This simulation process typically involved the participant running through their own or another individual's mental states in relation to the possibilities arising within the scenario so as to determine how they or others would feel (cf. Gordon, 1986). The following excerpt demonstrates such simulation, with the participant imagining themselves in a given situation and stating how they would feel and also how the protagonist might feel, rather than simply stating a rule such as "People feel upset when they receive a poor grade":

> "I myself am not particularly competitive erm, so I might be kind of disappointed and think, 'Oh well, that's kind of surprising, I'll erm, I'll have to find out why I went wrong'. But perhaps Jim might be slightly more likely to think, 'Oh I should have worked harder I should have'".

For each scenario the application of this coding scheme by the first author resulted in a percentage of theorizing and simulation for each participant as a function of all coded lines, including ambiguous ones. An independent coder checked a 10% sample of transcripts after first being trained in the application of the coding scheme. Inter-rate reliability was good, with 74% agreement. All areas of disagreement were resolved through discussion between the coders.

Theory-Based Reasoning

Table 1 presents the percentage of theory-based reasoning as a function of controllability and perspective. A 2 x 3 mixed design ANOVA revealed a main effect of controllability, $F(1, 87) = 15.81$, $MSE = 589.91$, $p < .001$, $\eta_p^2 = 0.15$, with theory-based reasoning being more prevalent in uncontrollable than controllable scenarios. There was also a main effect of perspective, $F(2, 87) = 11.41$, $MSE = 1238.82$, $p < .001$, $\eta_p^2 = 0.21$, with the other-and-other condition evoking the greatest level of theorizing and the self-only condition the least. The controllability by perspective interaction was not reliable, $F(2, 87) = 0.75$, $MSE = 589.91$, $p = .48$, $\eta_p^2 = 0.02$. Post hoc comparisons showed significant differences between the self-only condition and the other-and-other and self-and-other conditions ($ps < .01$). No difference was found between the self-and-other and other-and-other conditions ($p = .23$).

Table 1: Mean percentage of theorizing as a function of outcome controllability and perspective (SDs in brackets).

Perspective	Outcome Controllability		
	Controllable	Uncontrollable	*M*
Self-only	33 (33)	46 (40)	40
Self-&-other	49 (29)	69 (23)	59
Other-&-other	65 (25)	75 (28)	70
M	49	63	

Table 2: Mean percentage of simulation as a function of outcome controllability and perspective (SDs in brackets).

Perspective	Outcome Controllability		
	Controllable	Uncontrollable	*M*
Self-only	52 (38)	40 (37)	46
Self-&-other	24 (27)	20 (21)	22
Other-&-other	20 (21)	15 (24)	18
M	32	25	

Simulation-Based Reasoning

Table 2 presents the percentage of simulation as a function of controllability and perspective. A 2 x 3 mixed design ANOVA revealed a main effect of controllability, $F(1, 87) = 3.97$, $MSE = 555.95$, $p = .049$, $\eta_p^2 = 0.04$, with greater simulation in controllable than uncontrollable scenarios. There was also a main effect of perspective, $F(2, 87) = 12.63$, $MSE = 1122.86$ $p < .001$, $\eta_p^2 = 0.23$, with more simulation in the self-only condition relative to the other-and-other and self-and-other conditions. No interaction was observed between perspective and controllability, $F(2, 87) = 0.57$, $MSE = 555.95$, $p = .57$, $\eta_p^2 = 0.01$. Post-hoc comparisons revealed a significant difference between the self-only condition and the self-and-other other-and-other conditions ($ps < .001$), but no difference between the self-and-other and other-and-other conditions ($p = .90$).

Discussion

To our knowledge this is the first study to examine the TT/ST debate through the prism of mental state reasoning with counterfactual scenarios. The results from our protocol analysis indicate that although theorizing dominated overall, there were nevertheless differences across conditions when theorizing and simulation data were analyzed separately.

Looking first at the controllability factor, our results indicated that theorizing was more prevalent in uncontrollable than controllable scenarios, with the reverse being the case for simulation. The observation that people theorize more and simulate less in uncontrollable scenarios relative to controllable ones is consistent with the view that uncontrollable scenarios evoke less consideration and modeling of alternative possibilities. In essence, participants appear to be minimizing cognitive effort in these cases. For the controllable condition, simulation may have been facilitated because it was possible to consider more alternatives to reality, thereby provoking a more detailed examination of how an individual might feel in a situation. Participants also appeared to be more likely to engage in reasoning about *how* they might feel in such scenarios, using this to infer how the protagonist might feel.

How do these findings concerning the effect of outcome controllability on mental state reasoning fit in with other theories? Mitchell et al.'s (2009) hybrid account argues that simulation is used by default, but in cases where a situation is familiar they suggest that people might use rule-based theorizing as a shortcut strategy. However, our controllable and uncontrollable scenarios were equated for familiarity in a pre-test. As such, since controllability was not confounded with familiarity it is not immediately apparent how Mitchell et al.'s account might address the observed influence of controllability on rates of theorizing and simulation. It may be the case, however, that both theorizing and simulation reflect different strategies for engaging in mental state understanding, with one or other strategy being elicited by different factors in the prevailing context, including familiarity and event controllability - and potentially other cues (e.g., the emotionality of the situation).

Our study also set out to examine whether differences arise in how people reason about themselves versus others. Our analysis showed that the self-only condition elicited more simulation and less theorizing than the other-and-other condition, with the self-and-other condition occupying a middle position on both the theorizing and simulation measures. One reason for relatively more simulation arising in the self-only condition may be that it is triggered by direct emotional engagement with presented scenarios arising from specific memories of personal experiences.

Overall, our findings indicate that both theorizing and simulation occur in mental state reasoning about regret-oriented counterfactual scenarios. This supports a hybrid view of mental state understanding along the general lines espoused by Mitchell et al. (2009), and suggests the traditional TT/ST debate may be misconceived in its attempt to emphasize the deployment of a unitary reasoning approach based purely around *either* theorizing *or* simulation. Our results also have implications for mental models accounts of counterfactual reasoning. So far these accounts have been dominated by studies of the action effect (e.g., Feeney & Handley, 2006), and the temporal order effect (e.g., Atkinson et al., 2009), with less work examining issues relating to the controllability of regret outcomes. Our research suggests that an initial model may be formed by theory-based reasoning, with subsequent models being fleshed out through a mental simulation process involving the identification of multiple alternative possibilities. Although speculative, these ideas resonate with previous findings relating to the action effect, and represent a useful area for future research.

Reflecting on our results more generally, we wonder whether they also speak to dual-process accounts (e.g., Evans, 2003, 2006), which contend that human reasoning involves the interplay between two distinct reasoning processes. On the one hand Type 1 or heuristic processes are fast, automatic, high capacity and involve low cognitive effort. On the other hand, Type 2 or analytic processes are slow, controlled, low capacity and involve high cognitive effort. Under some dual-process accounts, Type 1 processes act by default to provide an initial response that can be overturned through the application of Type 2 processes (e.g., Evans, 2006). We suggest that theory-based reasoning may map onto Type 1 processing, and simulation-based reasoning may map onto Type 2 processing. Our findings suggest that there was little simulation that was not also driven by an initial phase of theorizing, which implies that

theorizing may be primary, and that if further processing is required this arises through simulation and may serve either to confirm or override a theory-based decision.

Evidence for this dual-process view of mental state reasoning comes from Atkinson et al.'s (2009) study, where the absence of an influence of speeded responding on the temporal order effect suggests the rapid and automatic deployment of a rule-based process, in line with TT assumptions that we possess a set of folk psychological theories. Furthermore, Atkinson et al.'s observation that speeded responding modulated the emergence of the action effect is indicative of slower, controlled, Type 2 processing linked to simulation. These dual-process arguments also resonate with Apperly and Butterfill's (2009) claims for two processing systems in mental state reasoning, with the proposal being that infants possess a cognitively efficient but inflexible method for tracking belief states that runs parallel to a later-developing adult system which is more flexible but cognitively demanding.

We conclude by returning to the two issues mentioned in our introduction that we hoped our research might address, that is: (i) whether mental state understanding is based on *either* theory-based reasoning *or* simulation-based reasoning - or whether both types of processing are deployed; and (ii) whether the manipulation of factors such as outcome controllability and the self/other distinction might determine the propensity for individuals to theorize or simulate. In relation to the first issue, we have demonstrated by means of think-aloud protocols and the adoption of counterfactual thinking scenarios that both theory-driven and simulation-driven reasoning play out in mental state understanding, with all participants deploying theorizing *and* simulation to greater or lesser degrees for many of the scenarios. In relation to the second issue, we have shown that people are more likely to engage in simulation when thinking about themselves rather than when thinking about other individuals. Furthermore, they are more likely to engage in simulation when reasoning about controllable than uncontrollable regret outcomes. Moreover, these two factors (i.e., perspective and controllability) appear to combine additively to determine the relative levels of simulation and theorizing that arise in mental state reasoning.

Standard, unitary TT and ST accounts do not seem to be able to accommodate our observations that the processes underpinning mental state understanding are influenced by content, context and perspective effects. Although these accounts may be able to develop ways to explain the present evidence, it remains for the proponents of these theories to take up this challenge. In contrast, hybrid accounts that embrace both TT and ST seem better able to deal with our findings. We suggest that hybrid theories represent an important new direction in research examining the processes associated with mental state reasoning.

Acknowledgements

We acknowledge the ESRC for an interdisciplinary studentship awarded to the first author.

References

Apperly, I.A., & Butterfill, S.A. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological Review, 116,* 953-970.

Atkinson, L., Bell, D., & Feeney, A. (2009). The relationship between counterfactual thinking and emotional reactions to event outcomes: Does one account fit all? *Psychonomic Bulletin & Review, 16,* 724-728.

Ball, L.J., & Christensen, B.T. (2009). Analogical reasoning and mental simulation in design: Two strategies linked to uncertainty resolution. *Design Studies, 20,* 169-186.

Byrne, R.M.J. (2002). Mental models and counterfactual thoughts about what might have been. *Trends in Cognitive Sciences, 6,* 426-431.

Byrne, R.M.J. (2005). *The Rational Imagination: How People Create Alternatives to Reality.* Cambridge, MA: MIT Press.

Byrne, R.M.J., Segura, S., Culhane, R., Tasso, A., & Berrocal, P. (2000). The temporality effect in counterfactual thinking about what might have been. *Memory & Cognition, 28,* 264-281.

Carruthers, P. (1996). Autism as mind-blindness: An elaboration and partial defence. In P. Carruthers & P.K. Smith (Eds.), *Theories of Theories of Mind* (pp. 257-273), Cambridge: CUP

Evans, J.St.B.T. (2003). In two minds: Dual-process accounts of reasoning. *Trends in Cognitive Sciences, 7,* 454-459.

Evans, J.St.B.T. (2006). The heuristic-analytic theory of reasoning: Extension and evaluation. *Psychonomic Bulletin & Review, 13,* 378-395.

Feeney, A., & Handley, S.J. (2006). Comparisons, mental models, and the action effect in judgments of regret. *Memory & Cognition, 34,* 1422-1430.

Gilovich, T., & Medvec, V.H. (1994). The temporal pattern to the experience of regret. *Journal of Personality and Social Psychology, 67,* 357-365.

Girotto, V., Ferrante, D., Pighin, S., & Gonzalez, M. (2007). Postdecisional counterfactual thinking by actors and readers. *Psychological Science, 18,* 510-515.

Goldman, A.I. (2006). *Simulating Minds: The Philosophy, Psychology and Neuroscience of Mindreading.* New York: Oxford University Press.

Gordon, R.M. (1986). Folk psychology as simulation. *Mind & Language, 1,* 158-171.

Kahneman, D., & Tversky, A. (1982). The psychology of preferences. *Scientific American, 246,* 160-173.

Kühberger, A., Kogler, C., Hug, A., & Mösl. E. (2006). The role of the position effect in theory and simulation. *Mind & Language, 21,* 610-625.

Mitchell, P., Currie, G., & Ziegler, F. (2009). Two routes to perspective: Simulation and rule-use as approaches to mentalizing. *British Journal of Developmental Psychology, 27*, 513-543.

Ratcliffe, M. (2007).*Rethinking Commonsense Psychology: A Critique of Folk Psychology, Theory of Mind and Simulation.* Basingstoke: Palgrave Macmillan.

Less-is-more Effects in Knowledge-based Heuristic Inference

C. Philip Beaman (c.p.beaman@reading.ac.uk)
Centre for Integrative Neuroscience & Neurodynamics
School of Psychology & Clinical Language Sciences, University of Reading
Earley Gate, Whiteknights, Reading RG6 6AL UK

Philip T. Smith (p.t.smith@reading.ac.uk)
School of Psychology & Clinical Language Sciences, University of Reading
Earley Gate, Whiteknights, Reading RG6 6AL UK

Rachel McCloy (r.a.mccloy@reading.ac.uk)
School of Psychology & Clinical Language Sciences, University of Reading
Earley Gate, Whiteknights, Reading RG6 6AL UK

Government Social Research Unit, HM Treasury
1 Horse Guards Road, London SW1A 2HQ UK

Abstract

Inference on the basis of recognition alone is assumed to occur prior to accessing further information (Pachur & Hertwig, 2006). A counterintuitive result of this is the "less-is-more" effect: a drop in the accuracy with which choices are made as to which of two or more items scores highest on a given criterion as more items are learned (Frosch, Beaman & McCloy, 2007; Goldstein & Gigerenzer, 2002). In this paper, we show that less-is-more effects are not unique to recognition-based inference but can also be observed with a knowledge-based strategy provided two assumptions, limited information and differential access, are met. The LINDA model which embodies these assumptions is presented. Analysis of the less-is-more effects predicted by LINDA and by recognition-driven inference shows that these occur for similar reasons and casts doubt upon the "special" nature of recognition-based inference. Suggestions are made for empirical tests to compare knowledge-based and recognition-based less-is-more effects.

Keywords: Heuristics; Recognition; Less-is-more; LINDA

The Less-is-more Effect

Suppose an individual is presented with the two cities *Milan* and *Modena* and asked to choose between the two along some criterion, for example to decide which has the larger population. In the classic work of Goldstein and Gigerenzer (2002), it is assumed that the participant will guess if they recognize neither of the items, they will use whatever additional knowledge is available to make a decision if they recognize both of the items and if they recognize only one of the items, they will choose this item as the larger without consulting any other cues or searching for further information about it (the *Recognition Heuristic* or *RH*). Recognition-driven inference of this type predicts a *less-is-more effect*, whereby individuals who recognize many of the items often perform worse than individuals who recognize fewer of the items (Goldstein & Gigerenzer, 2002). A number of studies have shown that this effect can be observed empirically (Frosch, Beaman & McCloy, 2007; Goldstein & Gigerenzer, 2002; Reimer & Katsikopoulos, 2004). It occurs because items that are more prominent (e.g., larger, more populous cities) are more likely to be encountered, hence more likely to be recognized. Recognizing one of the two items is thus a useful cue for choosing the recognized item; whereas if both items are recognized, additional knowledge is needed to make the decision and such additional knowledge may be very limited in discriminative power. In the terms provided by Goldstein and Gigerenzer (2002) a less-is-more effect, superior performance by an individual who recognizes fewer of the options, is expected when the recognition validity (the probability that a correct decision is made based upon recognition alone) exceeds the knowledge validity (the probability that a correct decision is made based upon the best available knowledge about the items).

The assumption underlying the RH is that items scoring higher on the criterion under consideration (larger cities, more successful ice-hockey teams, better tennis players etc). are ordinarily encountered more frequently. The counter-intuitive nature of the less-is-more effect makes its prediction by recognition-driven inference interesting, and has been used as a rhetorical device to promote the heuristic (Borges, Goldstein, Ortmann & Gigerenzer, 1999; Gigerenzer, 2007). Counter to this, failures to observe the effect have been cited in attempts to refute the RH (e.g., Boyd, 2001; Dougherty, Franco-Watkins & Thomas, 2008; Pohl, 2006). In describing the RH, Goldstein and Gigerenzer (2002) use the example of recognizing a city because it has appeared frequently in newspaper reports, a larger city is more likely to be so mentioned. Any individual who is presented with a city they recognize (but know nothing more about) and one they do not is therefore well-advised to choose the recognized city if judging which of the two is more populous. However, the recognizability of a particular city, for example, is a function of several factors, including its physical distance from the individual as well as its size. An appropriate analogy here might be the force of gravity. Local towns, like nearby planetary bodies, might have intrinsically less "pull" or prominence than distant

cities (or distant galaxies) but their appearance in local news reports is enhanced by their closer physical proximity and both of these factors influence recognizability. A further moderating factor is the way in which the individual might shape the environment their own ends. In the newspaper example, the individual receiving the newspaper is implicitly assumed to be a fairly passive processor of the information contained within the newspaper and no consideration is given to the potential difference between an individual who actively seeks out a newspaper and one who does not or to potential differences between choice of reading matter (e.g., the *New York Review of Books* versus the *National Enquirer*) which may have very different content, and each of which might be sought out, or passively encountered, to different degrees by different individuals or groups of individuals.

A basic premise in what follows is that, for any given individual, there are several subgroups of items which the individual is able to recognize and about which they may also have partial knowledge. This is particularly likely if they are local to the individual in some way or if they form part of a set of items of special interest to that individual. For example, American cities include large, famous cities such as *New York* and *New Orleans*, and small cities associated with famous universities, such as *New Haven* and *Palo Alto*. The relative recognition of various subgroups (such as those with famous Universities) may not be simply correlated with size. Any individual with specialist knowledge or affiliation with any special-interest group, e.g., membership of a European academic community, might be more likely to recognize small but academic cities in the USA than all but the most famous large USA cities. For this fictional individual[1], there is a weaker relationship between recognition and magnitude for the subset of US cities with famous Universities than for the subset of US cities that do not possess famous Universities.

This assumption that differential access to various sub-groups of items may occur between individuals is not reliant upon anecdotal evidence or arguments of plausibility as above. By-item analysis of data taken from an experiment by McCloy, Beaman, Frosch and Goddard (in press) shows, when a group of 40 participants were asked to indicate which of a group of famous individuals they recognize, a significant interaction between the reason for the individual's celebrity and the participant's gender, $F(3, 43) = 13.44$, $p < .001$. For example, males recognized, on average, sports personalities 78% of the time (females = 55%) and rock stars 75% of the time (females = 66%). In contrast, females recognized fashion and show-business professionals 57% of the time (males = 33%). In what follows, we consider similar situations where, for an individual within the environment, there is no simple correlation between recognition and magnitude because subsets of the items are prominent for reasons unconnected to magnitude (e.g., the age, gender or special interests of the individual). The question we wish to address is whether less-is-more effects still occur in such situations and what forms of decision-rule, if any, will give rise to such effects.

[1] Who bears a strong resemblance to the second author.

LINDA

To formally examine the appearance of less-is-more effects, we suppose a pool of N items, split into several subsets A, B, C, ... Within each subset the participant is able to recognize a, b, c, ... items, respectively. In a typical test of recognition-driven inference, the experimenter selects items quasi-randomly from the pool. Since the constraints on the experimenter are unknown, a random selection from N is assumed and the basic case considered is where pairs of items are chosen, and the participant's task is to say which is larger. For purposes of exposition, attention is also restricted to situations in which there are just three subsets. The models can easily be extended to other cases (e.g., the participant is asked to choose between more than two items (Frosch et al., 2007; McCloy, Beaman & Smith, 2008) and/or the pool is split into more than three subsets).

On a given trial, suppose the participant recognizes i items from subset A, j items from subset B, and k items from subset C. Only two items are presented, so $0 \leq i + j + k \leq 2$. p_{ijk} is the probability of recognizing 0-2 items from A-C. p_{ijk} is dependent on how many items the participant can recognize in each of the subsets, but is independent of the decision rule adopted. The probability of success is α_{ijk}, given the recognition of i, j and k items from their respective subsets. α_{ijk} is dependent upon the decision rule adopted and distinguishes between models. The overall probability of success for any model is given by:

$$\Sigma_{ijk}\, p_{ijk}\, \alpha_{ijk} \qquad (1)$$

The distinguishing feature of the RH model is that the participant chooses the recognized item when only one item is recognized. So $\alpha_{000} = 0.5$ (no item recognized, pure guess); α_{100}, α_{010}, and α_{001} reflect the success of the recognition heuristic; α_{110}, α_{101}, α_{011}, α_{200}, α_{020}, α_{002} reflect use of knowledge. The alternative against which the RH is to be compared we refer to as LINDA (Limited INformation and Differential Access). As the name implies, this model requires two basic assumptions:

1. *The limited information assumption*. For each recognized item, the individual has reliable but limited information about its size (e.g. that the size is above the population median).
2. *The differential availability assumption*. Some subsets are more accessible than others so that subset A contains items that are more readily recognizable than subset B and so forth. The extent to which items in A are larger than items in B implements the recognition-criterion correlation which is the basis of the RH.

The limited information assumption is that some information is available at the time of decision-making against which to evaluate the usefulness of choosing the recognized item in any given case. This is strictly limited: above or below *median knowledge* corresponds in

information theoretic terms (Shannon & Weaver, 1948) to only 1 bit of information. The reliability of this information may also vary. The differential availability assumption states merely that, within any subset, a given individual may recognize more or less items. Thus, a member of the UK academic community may recognize more US cities with famous Universities than a UK-based baseball fan. The baseball fan, by contrast, may recognize more US cities with famous baseball teams.

Existence-Proofs for Knowledge-Based Less-is-more Effects.

For the LINDA model, consider the situation where the individual has accurate median knowledge of items from pool *N*, i.e., they accurately know whether each recognized item is above or below median. Subset *A* includes items in the top quartile of the size distribution, subset *B* includes items in the second highest quartile of the size distribution, and subset *C* contains all the remaining items. It is assumed for purposes of exposition that median knowledge is perfect, i.e., that the knowledge about a recognized item is accurate. This assumption can be relaxed but the general conclusions reported here hold for all reasonably high levels of accuracy (to just above chance). The size of the pool from which the test items are drawn is set at 100 but the same pattern of results is obtained for all large values of *N*. The key prediction is the relation between the proportion of correct decisions (calculated by equation (1)) and *n*, the number of items in the pool the participant can recognize. A less-is-more effect occurs, according to Goldstein & Gigerenzer's (2002) definition whenever performance of the inference task is demonstrably superior under conditions where fewer items from the pool of test items are recognized. McCloy et al. (2008) use a stricter definition, arguing that less-is-more effects should be restricted only to those areas of the graph where learning more items will continue to impair performance. We use the latter definition for our examples, although note that when this definition holds it necessarily implies that Goldstein & Gigerenzer's conditions are also met.

Example 1: Low validity for complete recognition. One way that less-is-more effects may be produced relates to how decisions are made when both items are recognized (in a 2-alternative forced choice task). LINDA is assumed to access limited and possibly inaccurate knowledge about the size of each recognized item, and use this knowledge to choose the item she believes to be larger. Suppose that choosing between two recognized items may, in some instances, be extremely difficult. An extreme version of this appears in Figure 1. When only one item is recognized, LINDA makes decisions on the basis of whether the item is judged above median (choose the recognized item) or below the median (choose the unrecognized item), as given in the appendix. Recognition-criterion correlations can be varied by varying the availability of the items in the subsets available to LINDA. For example, if all items in subset *A* (top quartile of the criteria) are recalled before all items in subset *C* (below median) then the recognition-criterion correlation is obviously higher than when all items in subset *C* are recalled before all items in subset *A*. In this simulation, we manipulated the recognizability of individual items within the subsets to obtain pre-set correlations between recognition and criterion. For the current example, we also assume that LINDA does not have the capacity to make a decision when both items are recognized, and so is obliged to guess, that is α_{110}, $\alpha_{101,}$ α_{011}. $\alpha_{200,}$ α_{020} and α_{002} were not calculated but all set at 0.5, as would be the case with simulations of the RH. The situation resembles one outlined in Goldstein and Gigerenzer (2002, pp. 84-85) in which German participants were experimentally exposed to the names of US cities without being presented with any further information which might be of use, and is also comparable with Schooler and Hertwig's (2005) ACT-R implementation of the recognition heuristic, which also assumed chance level performance when both items were recognized (Schooler & Hertwig, 2005, p. 614).

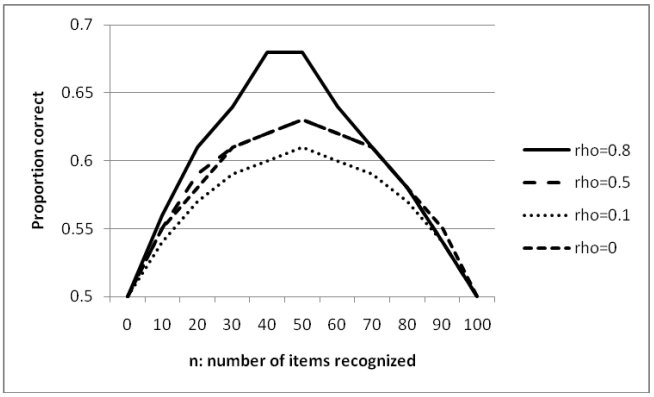


Figure 1: Proportion correct for the LINDA model when discrimination between two recognized items is at chance

Figure 1 shows clear less-is-more effects for all values of the recognition-criterion correlation tested. As more items are recognized (beyond a mid-point of 50% recognition rate) the proportion of correct inferences drops.

Unlike the RH model, which requires quite large criterion-recognition correlations to allow recognition validity to exceed knowledge validity, LINDA shows less-is-more effects for all values of the criterion recognition correlation, ρ, although the largest less-is-more effects occur for the largest values of ρ. For comparison, Figure 2 shows the predicted performance of the RH when knowledge validity is at chance and recognition validity takes the values of ρ reported in Figure 1. The validity of recognition is determined to some extent by ρ, which is determined for LINDA by the orderings of subset availability, and she experiences less-is-more effects occur even with low and zero values of ρ.

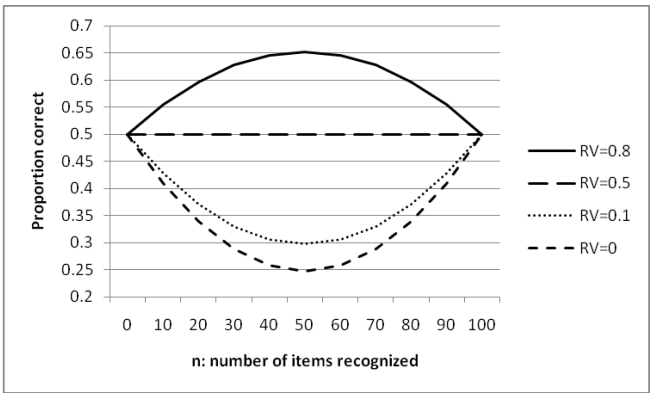


Figure 2: Predictions for the RH when recognition validity (RV) takes the values of the recognition-criterion correlation reported for LINDA. In this example, the RH is followed despite below-chance levels of validity (RV<0.5) which may not be realistic, but alternative strategies have yet to be suggested for these situations and, in particular, the point at which the RH is abandoned is not clearly outlined.

Example 1 relies upon the assumption that distinguishing between two recognized items is sufficiently difficult as to be effectively at chance. Both LINDA and RH are open to the criticism that, if knowledge validity for full recognition is chance, any non-random strategy able to operate when only one item is recognized will outperform knowledge and show less-is-more effects. This is a particular problem with the RH, where both knowledge validity and recognition validity are both set a priori for simulations such as this. Example 2 shows that low knowledge validity for full knowledge is not a necessary precondition for the appearance of less-is-more effects.

Example 2: Variation in subset availability. In order to formally compare LINDA with the RH model, we arranged that the models perform equally well when all items are recognized. Calculated probability of success when all items were recognized was 0.7525 for LINDA so knowledge validity was set at this level for the RH. The orderings of subsets in terms of recognition provide a potential rationale for variation in criterion-recognition correlation between individuals. Different orderings of subsets (and hence different recognition-criterion correlations) were simulated and the expected proportions correct using LINDA and the RH is given in Figure 2. We will use the notation *ABC* to denote subset availability, where *ABC* means that items from subset *A* are all more recognizable than the items from subset *B*, which in turn are all more recognizable than the items from subset *C*. A strict *ABC* recognition order obviously implies a high recognition-criterion correlation. Other recognition orderings (e.g., *ACB*) imply lower criterion-recognition correlations. *ABC* ordering is equivalent to a correlation between recognition and criterion of $\rho = .919$ and *ACB* ordering is equivalent to $\rho = .306$.

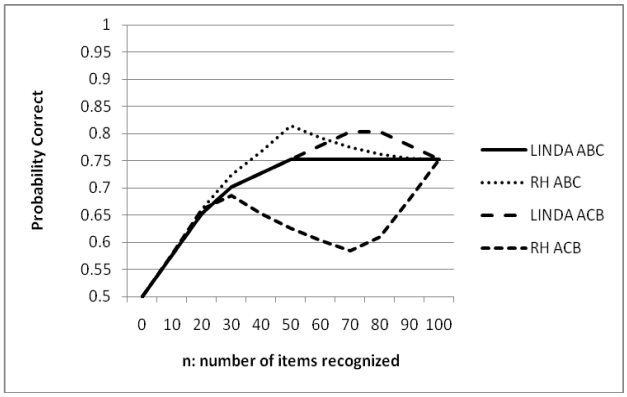


Figure 3. Performance of LINDA and the RH according to a recognition - criterion correlation determined by the recognizability of subsets. *ABC* ordering results in a high correlation and *ACB* a low (but still positive) correlation.

Figure 3 shows the performance of LINDA and the RH model for two different criterion-recognition orderings: *ABC* (items in the top quartile of the size distribution are most recognizable and items below median are least recognizable) and *ACB* (items in the top quartile are most recognizable, then items from below the median and finally items from the second quartile). *ABC* ordering corresponds to a strong criterion-recognition correlation ($\rho = .919$) and *ACB* ordering to a smaller, but still positive, correlation between criterion and recognition ($\rho = .306$).

The *ABC* ordering produces the expected effects from the literature. The RH model shows the less-is-more effect, while the knowledge-based LINDA model shows a monotonic relation between proportion correct and number of recognizable items. The situation is quite different for the *ACB* ordering: here, LINDA produces an inverted-U shaped function and a less-is-more effect. Less-is-more effects therefore do not imply use of the recognition heuristic – even given a positive criterion-recognition correlation – but may occur for other reasons. The inverted-U shaped functions that characterize the less-is-more effect indicate that a task becomes more difficult once the number of recognizable items passes a certain level. In the case of the RH model and the *ABC* ordering, this is because "easy" decisions (select the recognized item when only one item is recognized) are gradually outnumbered by "difficult" decisions (choose between items, both of which have been recognized) as the number of recognizable items increases. In the case of LINDA and the *ACB* ordering, moderate levels of recognition produce many easy decisions (discriminating a recognized item drawn from subset A from a recognized item drawn from subset C) but the decisions become more difficult when items of intermediate size, from subset B, begin to join the pool of recognizable items as the number of recognizable items increases.

Discussion

Whilst the two models give less-is-more effects in different circumstances, the effects are produced for essentially the same reasons. When few items are recognizable, the task is easier than when many items are recognizable. In the case of the RH model, for both Examples 1 and 2, when few items are recognizable the individual is more frequently confronted with the easy decision of selecting the one item recognized, rather than the problematic case of choosing between two recognized items, and this position is reversed when many items are recognizable. In Example 1 LINDA benefits from knowledge about the single item recognized which is not available to discriminate between two recognized items. For LINDA, performance in Example 2 for intermediate levels of recognition (up to 75 items) continues to improve as recognition rates rise because the discrimination required is still more likely to be between an item drawn from top quartile (subset *A*) and an item drawn from the bottom quartiles (subset *C*). Adding items from the second highest quartile (subset *B*), however makes the task more difficult this and leads to a drop in performance, and hence a less-is-more effect, at this point.

The fluency rule (discussed by Schooler & Hertwig, 2005) produces similar results and, once again, for similar reasons. Those items which are retrieved more quickly, dependent upon memory activation-level, are presumed to score more highly on the criterion (e.g., large cities are more quickly retrieved). For the fluency rule, intermediate rates of decay of activation allow for better discrimination between activated items than either fast or slow rates of decay. Over time, both slow and fast forgetting producing similar activation levels for dissimilar items (e.g., very large and very small cities). However, the fluency rule does not require or use any knowledge beyond the fact of fast retrieval. Thus, although it produces less-is-more effects of a kind, these are arguably recognition-driven based upon speed of access, rather than knowledge-driven, based upon some item-specific knowledge. The fluency rule is also reliant upon a fixed rate of decay from memory, an assumption which has recently been challenged (Berman, Jonides & Lewis, 2009; Lewandowsky & Oberauer, 2009; Lewandowsky, Oberauer & Brown, 2009; Nairne, 2002).

Testing LINDA.

LINDA demonstrates that less-is-more effects can occur for knowledge-based decisions and also that, when discrimination between two recognized items is sufficiently difficult, these effects can occur regardless of the recognition-criterion correlation. She therefore stands as an existence proof that less-is-more effects need not imply the use of recognition-driven inference but can be produced by strategies that invoke criterion knowledge. Any model that makes use of limited knowledge is likely to produce LINDA-like behavior.

Although LINDA reproduces the less-is-more effects observed with the RH, it is also worth noting that knowledge-based and recognition-based less-is-more effects are, or should be, empirically distinguishable. LINDA produces less-is-more effects similar to the RH when full knowledge has validity only slightly higher than chance but, unlike the RH, LINDA produces such effects regardless of the size of the recognition-criterion correlation (Figure 1). She also shows less inclination to produce such effects when knowledge validity is not artificially constrained and the recognition-criterion correlation is particularly high. Indeed, LINDA is more likely to show less-is-more effects when the recognition-criterion correlation is rather more moderate (Figure 3). Thus, although LINDA provides a plausible alternative account of existing less-is-more effects, there are experimental manipulations not yet investigated which should provide data that favor either one account or the other.

References

Berman, M. G., Jonides, J., & Lewis, R. L. (2009). In search of decay in verbal short-term memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 35, 317-333.

Borges, B., Goldstein, D. G., Ortmann, A., & Gigerenzer, G. (1999). Can ignorance beat the stock market? In: G. Gigerenzer, P. M. Todd, & the ABC Research Group (Ed.s). *Simple heuristics that make us smart*. Oxford: Oxford University Press.

Boyd, M. (2001). On ignorance, intuition and investing: A bear market test of the recognition heuristic. *Journal of Psychology and Financial Markets, 2*, 150-156.

Dougherty, M. R., Franco-Watkins, A. M., & Thomas, R. (2008). Psychological plausibility of the theory of probabilistic mental models and the fast and frugal heuristics. *Psychological Review, 115*, 199-213.

Frosch, C., Beaman, C. P., & McCloy, R. (2007). A little learning is a dangerous thing: An experimental demonstration of ignorance-driven inference. *Quarterly Journal of Experimental Psychology, 60,* 1329-1336

Gigerenzer, G., & Brighton, H. (2009). Homo heuristicus: Why biased minds make better inferences. *Topics in Cognitive Science, 1*, 107-144.

Goldstein, D. G., & Gigerenzer, G. (2002). Models of ecological rationality: The recognition heuristic. *Psychological Review, 109*, 75-90.

Lewandowsky, S., & Oberauer, K. (2009). No evidence for temporal decay in working memory. *Journal of Experimental Psychology: Learning, Memory & Cognition, 35*, 1545-1551.

Lewandowsky, S., Oberauer, K., & Brown, G. D. A. (2009). No temporal decay in verbal short-term memory. *Trends in Cognitive Sciences, 13*, 120-126.

McCloy, R., Beaman, C. P., Frosch, C., & Goddard, K. (in press). Fast and frugal framing effects? *Journal of Experimental Psychology: Learning, Memory & Cognition*.

McCloy, R., Beaman, C. P., & Smith, P. T. (2008). The relative success of recognition-based inference in multi-choice decisions. *Cognitive Science, 32*, 1037-1048.

Nairne, J. S. (2002). Remembering over the short-term: The case against the standard model. *Annual Review of Psychology, 53*, 53-81.

Pachur, T. & Hertwig, R. (2006). On the psychology of the recognition heuristic: Retrieval primacy as a key determinant of its use. *Journal of Experimental Psychology: Learning, Memory & Cognition, 32*, 983-1002.

Pohl, R. F. (2006). Empirical tests of the recognition heuristic. *Journal of Behavioral Decision-Making, 19*, 251-271.

Reimer, T., & Katsikopoulos, K. (2004). The use of recognition in group decision-making. *Cognitive Science, 28*, 1009-1029.

Schooler, L. J. & Hertwig, R. (2005). How forgetting aids heuristic inference. *Psychological Review, 112*, 610-628.

Shannon, C. E., & Weaver, W. (1948). *The mathematical theory of communication*. Urbana: University of Illinois Press.

Appendix

1. Derivation of the values of p_{ijk} in Equation (1):

Probability of recognizing no items:

$p_{000} = [(N - a - b - c)/N] \times [(N - a - b - c - 1)/(N - 1)]$
$= (N - a - b - c)(N - a - b - c - 1)/[N(N - 1)]$

Probabilities associated with the recognition of only one item:

$p_{100} = [2a/N] \times [(N - a - b - c)/(N - 1)]$
$= 2a(N - a - b - c)/[N(N - 1)]$

Probability of recognizing one item from the top quartile.

Similarly for second quartile and below median:

$p_{010} = 2b(N\text{-}a\text{-}b\text{-}c)/[N(N - 1)]$
$p_{001} = 2c(N\text{-}a\text{-}b\text{-}c)/[N(N - 1)]$

Probabilities associated with the recognition of both items:

$p_{110} = 2ab/[N(N\text{-}1)]$
(one item is in the top quartile and one item is in the second quartile)
$p_{101} = 2ac/[N(N\text{-}1)]$
$p_{011} = 2bc/[N(N\text{-}1)]$
(as above, substituting *v* and *w* where appropriate)
$p_{200} = a(a - 1)/[N(N\text{-}1)]$
(both items are in the top quartile)
$p_{020} = b(b - 1)/[N(N\text{-}1)]$
$p_{002} = c(c - 1)/[N(N\text{-}1)]$
(as above, substituting *v* and *w* where appropriate)

2. α_{ijk} parameters for the LINDA model demonstrated in Example 2.

$\alpha_{000} = 0.5$

No items are recognized, performance is chance.

$\alpha_{100} = [0.5 \times (0.25N - a)/(N - a - b - c)]$
$+ [(0.75N - b - c)/(N - a - b - c)]$

Probability correct if one item from the top quartile is recognized.

$\alpha_{010} = 0.5 \times (0.25N - b)/(N - a - b - c)$
$+ (0.5N - c)/(N - a - b - c)$

Probability correct if the recognized item is in the second quartile.

$\alpha_{001} = (0.5N - a - b)/(N - a - b - c)$
$+ 0.5 \times (0.5N - c)/(N - a - b - c)$

Probability correct if the recognized item is below median.

$\alpha_{110} = 0.5$

Two items are recognized: one item is in the first quartile and the second item is in the second quartile, so with median knowledge, performance is chance.

$\alpha_{101} = \alpha_{011} = 1$

One recognized item is above median and one is below so success is certain.

$\alpha_{200} = \alpha_{020} = \alpha_{002} = 0.5$

Both recognized items are from the same quartile, and so cannot be distinguished.

3. α_{ijk} parameters for the Recognition Heuristic model demonstrated in Example 2.

$\alpha_{000} = 0.5$

$\alpha_{100} = 0.5 \times (0.25N - a)/(N - a - b - c)$
$+ (0.75N - b - c)/(N - a - b - c)$
$= (0.875N - 0.5a - b - c)/ (N - a - b - c)$

There is only one item recognized, it is in the top quartile.

$\alpha_{010} = 0$
$+ 0.5 \times (0.25N - b)/(N - a - b - c)]$
$+ (0.5N - c)/(N - a - b - c)$
$= (0.625N - a - b - c)/ (N - a - b - c)$

The recognized item is in the second quartile.

$\alpha_{001} = 0$
$+ 0.5 \times (0.5N - c)/(N - a - b - c)$
$= (0.25N - 0.5c)/ (N - a - b - c)$

The recognized item is below median.

$\alpha_{110} = \alpha_{101} = \alpha_{011} = \alpha_{200} = \alpha_{020} = \alpha_{002}$

All these cases involve recognition of both items, and it is assumed knowledge can be used with a certain probability of success. In the Example 2, this probability was chosen to ensure that the LINDA and RH models produced the same probability of success when all items were recognized.

What Makes a Good Reasoner?: Brain Potentials and Heuristic Bias Susceptibility

Wim De Neys (wim.deneys@univ-tlse2.fr)
CNRS, Université de Toulouse, France

Nikolay Novitskiy (nikolay.novitskiy@psy.kuleuven.be)
Lab Experimental Psychology, University of Leuven, Belgium

Jennifer Ramautar (J.Ramautar@nin.knaw.nl)
Netherlands Institute for Neuroscience, Amsterdam, the Netherlands

Johan Wagemans (Johan.wagemans@psy.kuleuven.be)
Lab Experimental Psychology, University of Leuven, Belgium

Abstract

Human reasoning is often biased by intuitive heuristics. A key question is why some people are less susceptible to this bias than others. It is debated whether the bias results from a failure to monitor one's intuitive conclusions for conflict with logical considerations or from a failure to inhibit the tempting intuitions. This results in different views on the role of individual differences in executive monitoring and inhibition capacity for sound reasoning. The present study presents a new approach to address this issue. After an initial reasoning screening a group of the most and least biased reasoners were invited for an EEG study in which neural markers of their executive monitoring (ERN amplitude) and inhibition (N2 amplitude) skills were recorded. Results indicated that biased reasoners were characterized by less developed inhibition but not monitoring capacity. Findings support the view that monitoring one's intuition for conflict during thinking is a flawless and undemanding process suggesting that even the poorest reasoners at least detect that they are biased.

Keywords: Decision-making; Reasoning; EEG

Introduction

Decades of reasoning and decision-making research showed that human thinking is often biased (Evans, 2008; Kahneman, 2002). In general, human reasoners seem to have a strong tendency to base their judgment on fast intuitive impressions rather than on more demanding, deliberative reasoning. Although this intuitive or so-called "heuristic" thinking can be very useful, it will sometimes cue responses that conflict with traditional normative logical or probabilistic considerations and bias our decision-making.

Whereas it is well established that human judgment is often biased, the nature of this bias is far less clear. Some influential authors have argued that the widespread heuristic bias can be attributed to a failure to monitor our intuition (e.g., Kahneman & Frederick, 2002). Because of lax monitoring, people would simply fail to detect that the intuitive response conflicts with normative considerations. However, others have argued that there is nothing wrong with the monitoring process (e.g., Epstein, 1994; Houdé, 2007; Sloman, 1996). According to these authors, people have little trouble in detecting that their intuitive response is biased. The problem, according to this view, is that people's intuitive beliefs are so tempting that they fail to discard them. Thus, people "behave against their better judgment" (Epstein, 1994) when they give an unwarranted heuristic response: They detect that they are biased but simply fail to block the biased response. In sum, according to this flawless detection view, biased decisions are attributed to an inhibition failure rather than a conflict monitoring failure per se.

The debate on the nature of heuristic bias results in opposing views on the interpretation of individual differences in bias susceptibility. Although the vast majority of educated adults are typically biased when solving classic reasoning and decision-making tasks, some people do manage to reason correctly and refrain from giving the tempting but unwarranted heuristic response. Individual differences in executive control capacity (as measured with general working memory or intelligence test) are widely cited as the cause of this reasoning performance variability (e.g., De Neys, 2006; De Neys & Verschueren, 2006; Evans, 2008; Stanovich & West, 2000). However, conflict monitoring and inhibition are both considered key executive processes and the precise contribution of each component as possible mediator of reasoning performance has not been established. Bluntly put, it is not clear what makes a good reasoner: Having a superior monitoring capacity, having a superior inhibition capacity, or a combination of both.

The two views on heuristic bias make differential predictions here. According to the lax monitoring view, people are mainly biased because of inefficient monitoring. Hence, one can expect that good reasoners will be primarily characterized by superior executive monitoring skills. Good reasoners will be better at monitoring their intuitively cued conclusions for conflict with more normative considerations and will be more likely to detect that their initial response is biased. However, the flawless monitoring view conceives monitoring during thinking as a quite undemanding process by entailing that even the most biased reasoners are successful at it (De Neys, Moyens, & Vansteenwegen, 2010; Franssens & De Neys, 2009). Hence, given the postulated minimal demands of the monitoring process during thinking, one can predict that individual differences in

executive monitoring skills per se should have little impact on one's reasoning performance: Even people with the least developed monitoring skills should manage to detect the conflict during thinking. According to this view, it will be specifically one's inhibitory capacities that will determine the reasoning performance.

Clarifying the nature of heuristic bias and the individual bias differences is crucial for the study of human thinking. The issue has also far-stretching implications for our view of human rationality and the design of more optimal intervention programs to "debias" human thinking (De Neys & Glumicic, 2008; Evans, 2008). The problem, however, is that it is hard to decide between the alternative views based on traditional reasoning data (Evans, 2007, 2008). Although there have been some recent attempts to break the stalemate by developing processing measures of conflict detection and inhibition during reasoning (e.g., De Neys & Franssens, 2009), the rival views persist. The present study introduces a new approach to address this issue by focusing on neural markers of individual differences of conflict monitoring and response inhibition.

In the study we first invited a large number of participants for an initial screening session in which they were presented with reasoning problems based on two of the most-famous tasks from the judgment and decision-making field: Kahneman and Tversky's (1973) base-rate neglect and conjunction fallacy problems. In these tasks a stereotypical description cues a strong intuitive response that conflicts with more traditional probabilistic normative considerations (see Material for examples). Sound reasoning on these problems requires that people detect the conflict and inhibit the inappropriate heuristic response. Based on the screening, we invited a group of the least and most biased reasoners (i.e., participants with the highest and lowest normative reasoning scores) for a follow-up study in which they were presented with a Go/No-No task while electroencephalography (EEG) was recorded. The Go/No-Go task is a classic task that is widely used to measure people's executive control abilities (e.g., Amodio et al., 2007; Nieuwenhuis et al., 2003). In the task participants must quickly respond to a frequently presented Go stimulus such that the 'Go' response becomes habitual. However, on a small proportion of trials, a No-Go stimulus appears, signaling that one's habitual response should be withheld.

The EEG recording allowed us to test for a possible neurological marker of the differential executive monitoring and/or inhibition capacities of the least and most biased thinkers. Available evidence suggests that the operation of the executive monitoring and inhibition components are reflected in two different event-related potentials (ERP). On one hand, erroneously solved No-Go trials on which participants give the inappropriate dominant 'Go' response are known to give rise to a specific ERP referred to as the Error-Related Negativity or ERN. The ERN is a sharp negative voltage deflection in the EEG that typically peaks about 50 ms after an erroneous response. The ERN is believed to reflect executive control activity associated with the monitoring of conflict and error (Amodio et al., 2004, 2006; Compton et al., 2008; but see Burle et al., 2008). Available evidence suggests that the ERN amplitude is typically larger for people with better monitoring skills (Amodio et al., 2006; Inzlicht et al., 2009).

On the other hand, correctly solved No-Go trials on which participants manage to withhold the dominant 'Go' response are known to give rise to the so-called N2. The N2 is a negative voltage deflection in the EEG that typically peaks about 200 ms after the stimulus onset (i.e., before the response). The N2 is believed to reflect executive control activity associated with the successful inhibition of the prepotent Go response (Nieuwenhuis et al., 2003). Available evidence suggests that the few times that people with less developed inhibitory abilities do manage to withhold the Go response, the N2 amplitude is larger than for people with high abilities (e.g., Johnstone et al., 2007; Kaiser et al., 2003; Prox et al., 2007; Smith, Johnstone, & Barry, 2004; but see also Falkenstein, Hoormann, & Hohnsbein, 1999). This larger N2 amplitude has been interpreted as reflecting the fact that people who have fewer inhibitory control resources will need a much higher activation of the neural control structures for the response inhibition to be successful (Prox et al., 2007; Smith et al., 2004).

In sum, the EEG literature suggests that individual differences in executive inhibition abilities affect the N2 amplitude, whereas individual differences in executive monitoring abilities affect the ERN amplitude. Hence, contrasting these components in a group of biased and unbiased reasoners can help us to clarify the nature of individual differences in heuristic bias susceptibility. If the lax monitoring view is right and good reasoners are characterized by superior monitoring ability, the ERN should be more pronounced for the unbiased than for the biased reasoners. If the flawless monitoring view is right and good reasoners are characterized by superior inhibition rather than monitoring ability, biased and unbiased reasoners should not show a differential ERN and only the N2 should differ in the two groups.

Reasoning Bias Screening

Method

Participants. A total of 399 psychology undergraduates participated in return for course credit.

Material. To screen participants' bias susceptibility during reasoning we presented them with a booklet containing a total of three conjunction fallacy and three base-rate neglect problems. Problems were presented in a fixed, randomly determined order. In all problems a stereotypical description cued a heuristic response that conflicted with the normative response that is traditionally considered correct. Problem content was based on the work of De Neys, Vartanian, and Goel (2008). The exact problem format is illustrated below.

The average number of correct normative responses was taken as an index of people's reasoning performance.

Conjunction fallacy problems. In each problem participants first read a short personality description of a character. Next, they were given two statements about the character and were asked to indicate which one of the two was most probable. One statement always consisted of a conjunction of two characteristics (one characteristic that was likely given the description and one that was unlikely). The other statement contained only one of these characteristics (i.e., the unlikely one). Consider the following example:

> Bill is 34. He is intelligent, punctual but unimaginative and somewhat lifeless. In school, he was strong in mathematics but weak in social studies and humanities.
>
> Which one of the following statements is most likely?
> a. Bill plays in a rock band for a hobby
> b. Bill is an accountant and plays in a rock band for a hobby

Normative considerations based on the conjunction rule always cue selection of the non-conjunctive statement. However, intuitively, people will tend to select the statement that best fits with the stereotypical description (i.e., the most representative statement, see Tversky & Kahneman, 1983). Clearly, the fit will be higher for the conjunctive statement than for the unlikely non-conjunctive statement. Hence, people will be intuitively tempted to pick the erroneous conjunctive statement.

Base-rate neglect problems. In each problem participants first read information about the composition of a sample. People were also informed that short personality descriptions were made of all the individuals in the sample and they would get to see one description that was drawn randomly from the sample. They were asked to indicate to which one of the two groups the randomly drawn individual most likely belonged. Consider the following example:

> A psychologist wrote thumbnail descriptions of a sample of 1000 participants consisting of 995 females and 5 males. The description below was chosen at random from the 1000 available descriptions.
>
> Jo is 23 years old and is finishing a degree in engineering. On Friday nights, Jo likes to go out cruising with friends while listening to music and drinking beer.
>
> Which one of the following two statements is most likely?
> a. Jo is a man
> b. Jo is a woman

Normative considerations based on the group size or base-rate information cue response (b). Given the size of the two groups in the sample, it will be more likely that a randomly drawn individual will belong to the largest group. However, people will be tempted to respond (a) on the basis of stereotypical beliefs cued by the description. Hence, just as in the conjunction problems, normative considerations will conflict with the cued heuristic response.

Descriptions were selected on the basis of an extensive pilot study (Franssens & De Neys, 2009). Selected descriptions moderately but consistently cued one of the two groups. This point is not trivial. We label responses that are in line with the base-rates as correct answers. However, if reasoners adopt a formal Bayesian approach (e.g., Gigerenzer, Hell, & Blank, 1988) and combine the base-rates with the diagnostic value of the description, this can lead to complications when the description is extremely diagnostic. Imagine that we have a sample of males and females and the description would state that the randomly drawn individual "gave birth to two children". Now, by definition, no matter what the base-rates in the sample are, one would always need to conclude that the person is a woman. We limited the impact of this problem by only selecting descriptions that were judged to have a moderate diagnostic value. By combining these with quite extreme base-rates (i.e., 995 and 5) one may generally conclude that the response that is cued by the base-rates should be selected if participants manage to refrain from giving too much weight to the intuitive answer cued by the description.

Results and Discussion

The reasoning performance of our screening sample replicated the typical results in previous studies. Overall, participants were typically biased and gave the cued heuristic responses. The average percentage of correct normative responses on the six problems was only 24% (*SD* = 33%). This pattern was similar for the conjunction (*M* = 21%, *SD* = 32%) and base-rate problems (*M* = 28%, *SD* = 31%).

After the screening we invited a group of the most (i.e., participants who always gave the heuristic response) and least biased reasoners (i.e., participants who gave at least one normative response on both the conjunction and base-rate problems) for the EEG recording session. This cutoff value (at least one response correct) corresponded to the median accuracy for both types of reasoning problems.

EEG Recording

Method

Participants. After the bias screening seven of the least and seven of the most biased reasoners were recruited for the main Go/No-Go EEG study. We refer to these groups as the poor and good reasoners, respectively (see Table 1 for an overview of their reasoning screening performance). Participants were paid €25 for their participation.

Material. *Go/No-Go task*. The Go/No-Go task was based on the procedure introduced by Nieuwenhuis et al. (2003) and Amodio et al. (2007). On each trial, either the letter "M" or "W" was presented in the center of a computer screen. Approximately half of the participants in each group were instructed to make a "Go" response (mouse button press) when they saw "M" but to make no response when

they saw "W"; the remaining participants completed a version in which "W" was the Go stimulus and "M" the No-Go stimulus. Each trial began with a fixation point, presented for 500 ms. The target then appeared for 100 ms, followed by a blank screen. Participants were instructed to respond within 500 ms of target onset. A warning message appeared on the screen for 1 s after responses that exceeded this deadline and after erroneous responses. The inter trial interval was 1 s.

The task consisted of 600 trials: 80% Go trials and 20% No-Go trials. The high frequency of Go trials induced a prepotent "Go" response, enhancing the difficulty of successfully overriding a response on the critical No-Go trials. Participants received a short 2-min break after every 150 trials.

Procedure. *EEG recording*. Participants were fitted with a Quickcap, and EEG was collected from 128 equidistantly positioned scalp sites using Ag/AgCl electrodes. The active reference electrode was placed on the vertex between electrodes Cz and Cpz. A ground electrode was placed on the forehead close to AFz. Vertical and horizontal electro-oculogram (EOG) was collected to permit the reduction of the artifact due to eye movements. Impedances were below 5kΩ at each scalp site. EEG was recorded through a 0.15 – 30 Hz bandpass filter and digitized at 1000 Hz using a SynAmps2 amplifier. Data were referenced to the average earlobe. Offline, we used a computerized algorithm to remove eye-blink artifacts. EEG epochs with voltage exceeding +/- 200 µV were rejected as reflecting additional artefact.

ERP processing. N2. Our quantification of the N2 and ERN was based on Amodio et al. (2007). For N2 quantification a 1000 ms epoch of EEG signal, beginning 200 ms prior to stimulus onset, was selected for each artifact-free trial. Baseline correction procedures subtracted the average voltage during the 200 ms interval before stimulus onset within each epoch from the entire epoch. Epochs associated with correct responses on Go and No-Go trials were averaged within their respective trial types. The N2 was scored as the peak negative deflection occurring between 200 and 400 ms, relative to target onset, at the vertex site (Cz), where it is typically maximal. The critical N2 component refers to the average N2 amplitude associated with correct "No-Go" responses.

ERP processing. ERN. For quantification of the ERN an 800 ms response-locked epoch of EEG signal, centered on the time of response within each trial, was selected for each artifact-free trial. Baseline correction procedures subtracted the average voltage occurring from 400 ms to 50 ms prior to the response from the entire epoch. Epochs associated with incorrect responses on No-Go trials and correct responses on Go trials were averaged within their respective trial types. The ERN was scored as the peak negative deflection occurring between -50 and 150 ms, relative to response onset, at the frontocentral scalp site (Fcz). The critical ERN component refers to the average amplitude associated with incorrect "Go" responses on "No-Go" trials.

Results and Discussion

Behavioral findings. The behavioral Go/No-Go performance of our two groups of reasoners (see Table 1) was as expected. Accuracy on the No-Go trials is considered an excellent marker of people's executive control ability. Consistent with the well established finding that good reasoners have superior executive control capacities, we observed that our group of unbiased reasoners outscored the more biased group on the No-Go trials, $F(1, 12) = 11.26$, $p < .01$, $\eta^2p = .48$. As expected, accuracy on the Go trials, where correct responding did not require monitoring or overriding the intuitive response, was at ceiling and did not differ for the two groups of reasoners, $F(1, 12) < 1$.

Table 1: Average (SD) Reasoning and Go/No-Go Accuracy

	Reasoning			Go/No-Go	
	Base-rate	Con-junction	Total	No-Go	Go
Poor reasoners	0% (-)	0% (-)	0% (-)	67% (11)	99% (1)
Good reasoners	52% (26)	62% (30)	57% (21)	83% (5)	99% (1)

N2 findings. Our ERP data indicated that the average N2 amplitude differed in the group of good and poor reasoners, $F(1, 12) = 4.75$, $p < .05$, $\eta^2p = .28$. As Figure 1 shows, whenever the poor reasoners did manage to solve No-Go trials correctly this was accompanied by a more pronounced N2 amplitude (i.e., a more negative deflection). Next, we also calculated the correlation between each individuals' actual reasoning performance on the base-rate and conjunction problems and their N2 amplitude. This analysis showed that in our restricted sample of good and poor reasoners, the N2 amplitude was a good predictor of the tendency to give the standard normative response on these classic reasoning problems, $r = .55$, $p < .05$. Hence, the better participant's executive inhibition capacity, as indexed by their N2 amplitude, the more they managed to refrain from heuristic responding during reasoning.

ERN findings. As Figure 1 indicates, in contrast with the N2 findings, the average ERN amplitude did not differ for our group of good and poor reasoners, $F(1, 12) < 1$. A correlational analysis also established that the ERN amplitude was not predictive of participant's reasoning performance, $r = .14$, $p = .63$. Consistent with the flawless monitoring view, this suggest that individual differences in bias susceptibility during reasoning are not driven by differences in executive monitoring skills as indexed by the ERN amplitude.

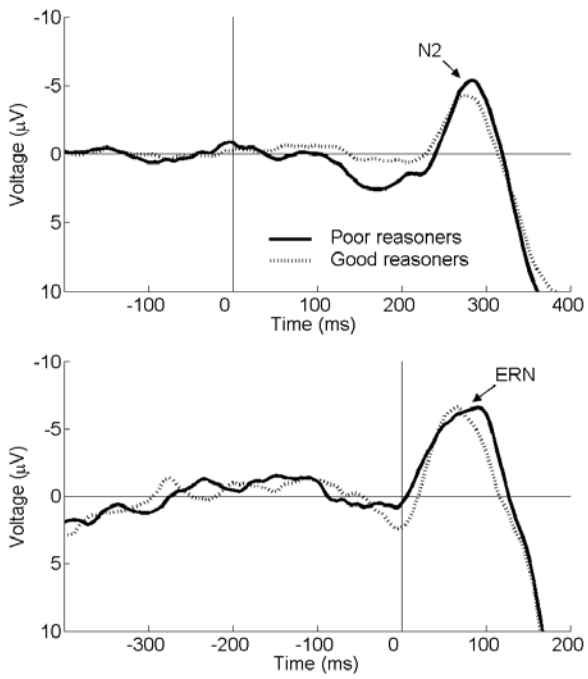


Figure 1. ERP waveforms corresponding to correct No-Go responses (N2 top panel, stimulus onset at 0 ms) and incorrect No-Go responses (ERN bottom panel, response onset at 0 ms), with the waveform for correct Go responses subtracted, for the most (poor) and least (good) biased reasoners.

General Discussion

In the present EEG study we contrasted neural markers of people's executive monitoring (ERN amplitude) and inhibition (N2 amplitude) capacity in two groups who showed differential susceptibility to heuristic bias during reasoning. Results indicated that less biased reasoners showed a smaller N2 amplitude than more biased reasoners while the ERN amplitude of biased and unbiased reasoners did not differ. Consistent with the flawless monitoring view, this suggests that good reasoners are specifically characterized by a superior executive inhibition capacity rather than by a superior monitoring capacity. Hence, what makes a good, unbiased reasoner is not a more developed ability to monitor one's intuitive conclusions for conflict with normative considerations but the ability to inhibit these tempting erroneous intuitions in case such a conflict occurs.

It should be stressed that the present results do not downplay the importance of conflict monitoring during reasoning per se. Both the lax and flawless monitoring views consider the monitoring of one's intuitive inferences as a cornerstone of the reasoning process. Obviously, if people would not monitor their intuitively cued problem solutions, they could simply not detect whether or not it is necessary to override them. Indeed, even the most gifted reasoners do not simply inhibit intuitive inferences throughout and tend to rely on heuristic computations in case it is appropriate (e.g., De Neys & Franssens, 2009; De Neys, Schaeken, & d'Ydewalle, 2005). As suggested previously (e.g., De Neys & Glumicic, 2008), the monitoring process allows reasoners to take advantage of the computational benefits (e.g., speed) of heuristic thinking as long as it does not conflict with normative principles. The key point, however, is that this crucial monitoring process does not seem to be very demanding. According to the flawless monitoring view, monitoring one's intuitions during reasoning is an effortless process that requires only minimal executive monitoring resources. It is this postulated undemanding or automatic nature of the monitoring process during reasoning that can explain why individual differences in executive monitoring capacity do not affect the reasoning performance. The undemanding nature of the monitoring during thinking entails that even for people with minimal executive monitoring resources, the process will be successful.

Our individual differences findings fit with some recent studies that started examining the processing characteristics of the conflict monitoring process during thinking. For example, Franssens and De Neys (2009) tested the postulated effortless nature of the monitoring process in a dual task study. People were asked to solve base-rate problems while their executive resources were burdened with a secondary task. After the reasoning task participants were also presented with a surprise recall test that can be used to measure whether people were monitoring their intuitive inferences and detected the conflict between cued intuitive and normative responses (see De Neys & Glumicic, 2008). Results showed that reasoning accuracy decreased under load (i.e., people gave more heuristic responses). However, the crucial finding was that the conflict monitoring index was not affected by the load. People were equally accurate in detecting the presence of conflict whether or not they were reasoning under load. Combined with the present individual differences findings these studies lend credence to the idea that conflict monitoring during thinking is effortless and flawless.

The present study is the first one to introduce EEG methodology to examine the nature of individual differences in bias susceptibility. Clearly, this implies that our results need to be interpreted with some caution. Although our data fits with recent findings pointing to the effortless nature of the monitoring process during thinking, the results will need to be validated in future studies. Bearing this in mind, our initial findings do suggest that individual differences in executive monitoring are not playing a major role in people's bias susceptibility. A good, unbiased reasoner seems to be primarily characterized by superior inhibitory skills. Although most reasoners might be detecting that their intuitive answer is biased, only people with superior inhibitory capacity manage to discard the tempting intuitive response.

References

Amodio, D. M., Jost, J. T., Master, S. L., & Yee, C. M. (2007). Neurocognitive correlates of liberalism and conservatism. *Nature Neuroscience, 10,* 1246-1247.

Amodio, D. M., Kubota, J. T., Harmon-Jones, E., & Devine, P. G. (2006). Alternative mechanisms for regulating racial responses according to internal vs external cues. *Social, Cognitive, and Affective Neuroscience, 1*, 26-36.

Amodio, D. M., Harmon-Jones, E., Devine, P. G., Curtin, J. J., Hartley, S. L., & Covert, A. E. (2004). Neural signals for the detection of unintentional race bias. *Psychological Science, 15*, 88-93.

Burle, B., Roger, C., Allain, S., Vidal, F., & Hasbroucq, T. (2008). Error negativity does not reflect conflict: A reappraisal of conflict monitoring and anterior cingulate cortex activity. *Journal of Cognitive Neuroscience, 20*, 1637-1655.

Compton, R. J., Robinson, M. D., Ode, S., Quandt, L. C., Fineman, S. L., & Carp, J. (2008). Error-monitoring ability predicts daily stress regulation. *Psychological Science, 19*, 702-708.

De Neys, W. (2006). Dual processing in reasoning: Two systems but one reasoner. *Psychological Science, 17*, 428-433.

De Neys, W., & Franssens, S. (2009). Belief inhibition during thinking: Not always winning but at least taking part. *Cognition, 113*, 45-61.

De Neys, W., & Glumicic, T. (2008). Conflict monitoring in dual process theories of reasoning. *Cognition, 106*, 1248-1299.

De Neys, W., Moyens, E., & Vansteenwegen, D. (2010). Feeling we're biased: autonomic arousal and reasoning conflict. *Cognitive, Affective, and Behavioral Neuroscience, 10*, 208-216.

De Neys, W., Schaeken, W., & d'Ydewalle, G. (2005). Working memory and everyday conditional reasoning: Retrieval and inhibition of stored counterexamples. *Thinking & Reasoning, 11*, 349-381.

De Neys, W., Vartanian, O., & Goel, V. (2008). Smarter than we think: When our brains detect that we are biased. *Psychological Science, 19*, 483-489.

De Neys, W., & Verschueren, N. (2006). Working memory capacity and a notorious brain teaser: The case of the Monty Hall Dilemma. *Experimental Psychology, 53*, 123-131.

Epstein, S. (1994). Integration of the cognitive and psychodynamic unconscious. *American Psychologists, 49*, 709-724.

Evans, J. St. B. T. (2007). On the resolution of conflict in dual process theories of reasoning. *Thinking & Reasoning, 13*, 321-329.

Evans, J. St. B. T. (2008). Dual-processing accounts of reasoning, judgement and social cognition. *Annual Review of Psychology, 59*, 255-278.

Falkenstein, M., Hoormann, J., & Hohnsbein, J. (1999). ERP components in Go Nogo tasks and their relation to inhibition. *Acta Psychologica, 101*, 267-291.

Franssens, S., & De Neys, W. (2009). The effortless nature of conflict detection during thinking. *Thinking & Reasoning, 15*, 105-128.

Gigerenzer, G., Hell, W., & Blank, H. (1988). Presentation and Content: the use of base-rates as a continuous variable. *Journal of Experimental Psychology: Human Perception and Performance, 14*, 513-525.

Houdé, O. (2007). First insights on "neuropedagogy of reasoning". *Thinking and Reasoning, 13*, 81-89.

Inzlicht, M., McGregor, I., Hirsh, J. B., & Nash, K. (2009). Neural markers of religious conviction. *Psychological Science, 20*, 385-392.

Johnstone, S. J., Dimoska, A., Smith, J. L., Barry, R. J., Pleffer, C. B., Chiswick, D., et al. (2007). The development of stop-signal and Go/Nogo response inhibition in children aged 7-12 years: Performance and event-related potential indices. *International Journal of Psychophysiology, 63*, 25-38.

Kahneman, D. (2002, December). Maps of bounded rationality: A perspective on intuitive judgement and choice. Nobel Prize Lecture. Retrieved January 11, 2006, from http://nobelprize.org/nobel_prizes/economics/laureates/2002/kahnemann-lecture.pdf

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgement. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and Biases: The Psychology of Intuitive Judgement* (pp. 49-81).

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80*, 237-251.

Kaiser, S., Unger, J., Kiefer, M., Markela, J., Mundt, C., & Weisbrod, M. (2003). Executive control deficit in depression: event-related potentials in a Go/Nogo task. *Psychiatry Research: Neuroimaging, 122*, 169-184.

Nieuwenhuis, S., Yeung, N., van den Wildenberg, W., & Ridderinkhof, K. (2003). Electrophysiological correlates of anterior cingulate function in a go/no-go task: effects of response conflict and trial type frequency. *Cognitive, Affective, and Behavioral Neuroscience, 3*, 17-26.

Prox, V., Dietrich, D. E., Zhang, Y. Y., Emrich, H. M., & Ohlmeier, M. D. (2007). Attentional processing in adults with ADHD as reflected by event-related potentials. *Neuroscience Letters, 419*, 236-241.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin, 119*, 3-22.

Smith, J. L., Johnstone, S. J., & Barry, R. J. (2004). Inhibitory processing during the Go/NoGo task: an ERP analysis of children with attention-deficit/hyperactivity disorder. *Clinical Neurophysiology, 115*, 1320-1331.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate. *Behavioral and Brain Sciences, 23*, 645-726.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review, 90*, 293-315.

The Mechanics of Embodiment

Ken McRae (kenm@uwo.ca)
Department of Psychology, University of Western Ontario
London, Ontario, Canada N6A 5C2

Martin H. Fischer (m.h.fischer@dundee.ac.uk)
School of Psychology, University of Dundee
DD1 4HN Scotland UK

Keywords: embodied cognition; computation; concepts.

Overview and Motivation

Embodied cognition is a theoretical stance which postulates that sensory and motor experiences are key parts of the representation of our knowledge. This view has challenged the longstanding assumption that knowledge is represented abstractly in an amodal conceptual network.

There now exist a large number of interesting and intriguing demonstrations of embodied cognition. Examples include changes in perceptual experience or motor behaviour as a result of semantic processing. These demonstrations have received a great deal of attention in the literature, and have spurred many researchers to take an embodied approach in their own work.

There are also a number of theoretical accounts of how embodied cognition might work. One influential proposal is "perceptual symbols system" theory, according to which the retrieval of conceptual meaning involves a partial re-enactment of experiences during concept acquisition. However, to a large extent, embodied theories are still developing, particularly in terms of computational implementations, as well as specification with regard to moment-by-moment on-line processing.

Given the established empirical foundation, and the relatively underspecified theories to date, many researchers are extremely interested in embodied cognition but are clamouring for more mechanistic implementations. This symposium aims to address this specific need for more detailed explorations of the specific processing mechanisms involved in embodied cognition. Four speakers from varying backgrounds and approaches will describe how they think the human mind is embodied, and what they view as the critical current and next steps toward mechanistic theories of embodiment.

Toward Implementing Embodiment

Lawrence Barsalou (Emory University, Atlanta, Georgia, USA) and Ken McRae (University of Western Ontario, London, Canada) will address issues concerning the construction of embodied computational models. One general set of issues concerns the computational architecture, aside from whether it takes the form of neural networks, Bayesian approaches, production systems, classic AI architectures, or another form. To implement a truly embodied system, multiple modalities are essential. In particular, intelligent action coupled with perception epitomizes embodied approaches, beyond basic response production. Other modalities are also essential from the embodied perspective, including affect and motivation, as well as abstract thought. Another architectural issue concerns the hierarchical structure of feature areas, the hierarchical structure of association areas, and the connectivity patterns among them (Simmons & Barsalou, 2003). Also important are the unique areas associated with bottom-up activation versus top-down simulation, along with shared areas. Finally, issues associated with the architecture's development and plasticity are important, including genetic and experiential contributions, and how epigenesis is realized (Elman et al., 1996).

A second set of critical issues surrounds specific forms of functionality to implement in the architecture. Barsalou (2003) argues that selective attention and categorical memory integration are essential for creating a symbolic system. Once these functions are present, symbolic capabilities can be built upon them, including type-token propositions, predication, categorical inference, conceptual relations, argument binding, productivity, and conceptual combination. Another key aspect is the implementation of space and time. Perception, cognition, and action must be coupled in space and time, and simulations of non-present situations must be implemented in space and time, perhaps using overlapping systems.

Because situated action in the environment is fundamental for all organisms, implementing embodied cognition that supports intelligent activity in a few critical situations may be a good place to start (Robbins & Aydede, 2008). By focusing on a complete embodied approach to achieving goals in specific situations, modelers must not only implement specific capabilities, such as goal setting, planning, perception, action, cognition, affect, reward, and learning, but implement interfaces that allow all these processes to interact effectively.

These are lofty goals indeed, and the remaining talks will describe current projects that are working toward them.

Computational Explorations of Perceptual Symbol Systems Theory

The second speaker is **Giovanni Pezzulo** (National Research Council, Rome, Italy) who has worked extensively

on computational models of embodied cognition. He will present an overview of his computational work and describe the precise mechanisms and processes involved in the emergence of embodied cognitive performance.

Pezzulo will present a computational architecture that acquires a "perceptual symbol system" through its autonomous interaction with the environment, and assembles perceptual symbols to form simulators for perceptual and abstract categories.

He will discuss the design of the architecture, which includes a combination of schema-based and dynamical systems principles, with the aim of suggesting a few basic mechanistic principles from which embodied theories of cognition can be implemented and tested.

Pezzulo will also discuss more generally the functioning of the perceptual-symbol-system-based architecture in prediction, categorization, and abstraction tasks, with the aim to assess the possible roles of perceptual symbols systems in producing embodied cognitive processing.

It is worth noting that, in addition to their use as "proofs of concept" for an embodied theory of cognition such as perceptual symbol systems, the importance of computational models also lies in the possibility to investigate elements that are left unspecified in the initial theoretical formulations, or that are challenging to study by means of experimental methods only. Therefore, Pezzulo will discuss the specific predictions and implications of our computational architecture for the perceptual symbol systems theory, and in particular the (tentative) answers it gives to challenging questions such as: *How are simulators formed from perceptual symbols? Which features are stored in simulators, and which are not? How are the most relevant simulators selected depending on the organism's goals and the current environmental context?*

Cognitive Modeling with the Open Source Humanoid Robot iCub

The third speaker is Angelo Cangelosi (University of Plymouth, UK) who leads a large-scale EU project on language and action learning (www.italkproject.org) and the Marie Curie doctoral network on developmental robotics (www.robotdoc.org). Approaching embodiment from an engineering and cognitive modeling perspective, his talk will describe the current state of development of efforts to implement human cognition in a physical agent with sensory and motor capabilities. He will describe in detail two current cognitive robotics models based on the humanoid robot iCub: one study on stimulus response compatibility effects and one on language acquisition. In addition, he will present the open source humanoid robotic platform iCub, and the associated computer simulator.

A Theoretical Framework for Embodied Cognition

The final speaker, Michael J. Spivey (University of California, Merced, USA), is a cognitive scientist who uses eye-tracking, reach-tracking, and neural network simulations to explore the close-knit relationship between sensorimotor processes and high-level cognition (Spivey, 2007). In Spivey's talk, he will discuss how the mountains of evidence for embodied cognition are now being acknowledged by classical cognitive scientists, albeit warily (Mahon & Caramazza, 2009). However, the next step, of how these many varied findings can impact traditional mainstream theories of cognition, is still not well articulated in the field. Rather than treating sensorimotor properties as "something extra" that gets facilitated after certain concepts become active, those sensorimotor properties may be part and parcel of the very conceptual representations themselves. What is needed at this stage is a push toward explicit computational models that implement sensorimotor grounding as intrinsic to cognitive processes. With such models, theoretical descriptions can be fleshed out as explicit mechanisms, idiosyncratic patterns across experiments may be explained, and quantitative predictions for new experiments can be put forward. Spivey will discuss some examples of such nascent modeling efforts.

References

Barsalou, L. W. (2003). Abstraction in perceptual symbol systems. *Philosophical Transactions of the Royal Society of London: Biological Sciences, 358,* 1177-1187.

Cangelosi A., Tikhanoff V., Fontanari J. F., & Hourdakis E. (2007). Integrating language and cognition: A cognitive robotics approach. *IEEE Computational Intelligence Magazine*, 2, 65-70.

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking innateness: A connectionist perspective on development*. Cambridge, MA: MIT Press.

Mahon, B.Z & Caramazza, A. (2009). Concepts & Categories: A Cognitive Neuropsychological Perspective. *Annual Review of Psychology,* 60, 27–51.

Pezzulo, G., & Calvi, G. (in press). Computational explorations of perceptual symbol systems theory. *New Ideas in Psychology*.

Robbins, P., & Aydede, M. (Eds.) (2008). *Cambridge handbook of situated cognition*. New York: Cambridge University Press

Spivey, M. J. (2007). *The Continuity of Mind*. New York: Oxford University Press.

Simmons, W. K., & Barsalou, L. W. (2003). The similarity-in-topography principle: Reconciling theories of conceptual deficits. *Cognitive Neuropsychology, 20,* 451-486.

Tikhanoff V, Cangelosi A., Fitzpatrick P., Metta G., Natale L., Nori F. (2008). An open-source simulator for cognitive robotics research: The prototype of the iCub humanoid robot simulator. In R. Madhavan & E. R. Messina (Eds.), Proceedings of IEEE Workshop on Performance Metrics for Intelligent Systems Workshop (PerMIS'08). Washington, DC.

On the Notion of Intended Meaning

Marco Cruciani (marco.cruciani@unitn.it)
Dept. of Information Engineering and Computer Science
University of Trento, Italy

Abstract

The issue of intended meaning is an open problem in the study of linguistic processes. The paper presents a notion of intended meaning based on the idea of speaker's preference for a state of affairs to which a sentence refers. Its argument has two components. The first is the conception of meaning developed by analytic philosophy of language; that is, the meaning of a sentence depends on the truth conditions of the sentence, and the meaning of an expression depends on contribution of that expression to the truth value of the sentence in which it appears. The second is the notion of agent's interest, as a state of affairs which implies a goal of agent, as developed by cognitive social theory. The paper maintains that a speaker's intended meaning establishes when the truth conditions of a sentence and the possibility conditions of the state of affairs preferred by the agent match. The last part of the paper illustrates three linguistic disputes to support its theoretical intuitions. The first dispute concerns syntactic ambiguity, while the other two disputes concern semantic ambiguity. The paper deals with the general problem of the semantic underdeterminacy of the conventional meaning of natural language sentences. Its specific contribution relates to the problem of intended meaning in communicative processes and to meaning negotiation processes in conflicting interactions.

Keywords: state of affairs; truth conditions; semantic underdeterminacy; intended meaning; interest; negotiation.

Introduction

The issue of intended meaning is an open problem in the study of linguistic processes (see Grice 1957, 1989; Kripke, 1979; Sperber & Wilson, 1986; Clark, 1996; Recanati, 2001; Bach, 2004; Bianchi, 2006). In this paper I present a notion of intended meaning based on the notion of speaker's preference for a state of affairs to which a sentence refers. This notion derives from the analysis of negotiation processes and the determination of meaning in linguistic controversies provoked by ambiguous clauses in contracts. The paper's contribution to the notion of intended meaning is based on the following thesis: given a set of contextually plausible interpretations of a sentence, the agent's intended meaning is determined by his/her extra-semantic situational interests (Cruciani, 2009a). It uses the notion of interest viewed as a state of affairs preferred by an agent because it implies his/her goal (see Conte & Castelfranchi, 1995).

In my view, the notion of the intended meaning of declarative sentences is founded on the relation between the states of affairs in which a sentence is true and the speaker's preferences ordering in regard to the states of affairs in which the sentence is true. A sentence can be true with respect to different sets of truth conditions, which correspond to different states of affairs (more technically, they correspond to sets of states of affairs). The state of affairs preferred by a speaker because it implies his/her goal provides the truth conditions which determine the intended meaning in the specific situation of use. From this perspective, the determination of intended meaning is viewed as a selection of a state of affairs which makes a sentence true (via truth conditions) and satisfies the agent's interest in situation.[1]

However, the aim of the paper is not to argue in favour of this conception, since the author has done so elsewhere (Cruciani, 2009b), but rather to explain its ontology. The schema in figure 1 illustrates the notion of intended meaning as it is conceived here. At the bottom of the schema is a sentence which, given a context of use, has some plausible interpretations. Each interpretation refers to a state of affairs which makes the sentence true: that is, it refers to specific truth conditions. The correspondence between the state of affairs preferred by the speaker and one of the states of affairs which make the sentence true determines the intended meaning. In other words, when the possibility conditions of the state of affairs preferred by speaker match the truth conditions of a sentence, we have intended meaning.

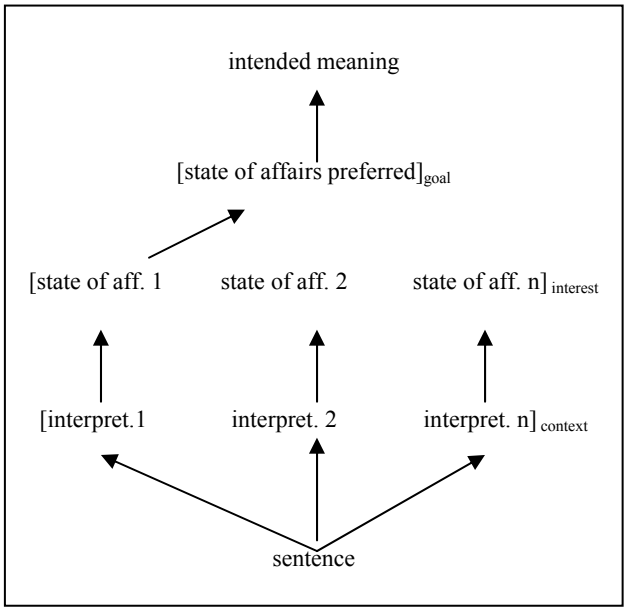


Figure 1

Meaning, Truth Conditions, States of Affairs and Context

In this section I illustrate the relation among meaning based on truth conditions, states of affairs, and context of use, and I outline some differences between semantics and pragmatics in regard to the phenomenon of semantic underdeterminacy. I base my view on the conception of

[1] Hence, the process of determining intended meaning can be explained in terms of preferences ordering.

meaning as developed by analytic philosophy of language (specifically by structural semantics). This maintains that the meaning of a sentence depends on its truth conditions, and that the meaning of an expression depends on the contribution of that expression to the truth value of sentence in which it appears (see Tarski, 1944). This notion entails that knowing the truth conditions of a sentence corresponds to knowing how the world would be if the sentence were true; but it does not correspond to knowing whether or not the sentence is actually true (see Wittgenstein, 1921). Hence, in cognitive terms, a speaker who knows the truth conditions of a sentence knows the meaning of the sentence even if s/he does not know how the world actually is, but only how it would be if the sentence were true.

The conception of meaning as (a set of) truth conditions is accepted by most philosophers and pragmatists of language. Any disagreement essentially concerns whether conventional meaning (obtained by linguistic conventions and rules) is sufficient to provide truth conditions or whether other items are required as well. In other words, is it sufficient to know semantic conventions and linguistic rules or do we need to know elements of the specific situation of use?

On a semantics view, conventional meaning and a small number of contextual parameters are sufficient to determine the truth conditions of a sentence (proposition expressed by the sentence). On a pragmatics of language view, conventional meaning is not sufficient to determine a unique set of truth conditions (semantic underdeterminacy): we need information on the context of use to complete the conventional meaning and to determine the truth conditions. Semantic underdeterminacy occurs when the conventional meaning of a sentence used by a speaker in a specific situation, coded by semantic conventions, underdetermines the proposition explicitly expressed by the utterance (see Travis, 1975, 1981; Searle, 1979, 1980).

In semantics, context is composed of some objective elements of the situation of utterance, and it is used to decode only some problematic kinds of expressions: indexical and demonstrative expressions such as "I", "here", "now" and "this", "that" (see Kaplan, 1977)[2]; pronouns such as "she", "he" (e.g. anaphoric use); cases of structural and lexical ambiguity (see Perry, 1997); and verbal tense.[3] In semantics, truth conditions lie at the level of the objective context of utterance and conventional meaning (obtained by linguistic conventions and rules). Pragmatists do not agree, however, arguing that in order to fix truth conditions we need supplementary contextual information. This information consists of shared knowledge (encyclopaedic and local), the discourse or sentence in which an expression is used, and elements of the physical surroundings.[4] A technical distinction between information on the semantic context and information on the pragmatic context is that the former is made accessible by and constrained to linguistic form of expression, while the latter is made accessible by speaker's communicative intentions and is not constrained to linguistic form.

I do not deal, in the paper, with the problem of whether there is a need for non-linguistic information to fix a unique set of truth conditions; there is a need (pragmatics/semantics distinction debate). I instead deal with the problem of intended meaning when a number of interpretations are all plausible in the same context (any context or combination of contexts).

I report a simple example to shed light on this point. The issue is the following: the conventional meaning of a sentence, even without indexical expressions, and structural and lexical ambiguity, actually underdetermines the proposition expressed by the sentence. And even with the additional pragmatic contribution of relevant contextual information, it is not always possible to fix a unique proposition. The sentence is as follows: (1) "there is water on Pluto". I do not know whether there is water on Pluto, but I understand the sentence because I am able to imagine the ways in which there might be water on Pluto: for example, in the form of ice on the planet's surface or in the form of gas in its atmosphere. Consequently, sentence (1) can have at least two interpretations (two different sets of truth conditions); that is, it can be true both if there is water in the form of ice on the surface and if there is water in the form of gas in the atmosphere. The two truth conditions correspond to different states of affairs:

a. "there is ice on the surface of Pluto";
b. "there is water vapour in the atmosphere of Pluto".

Hence, (1) can refer to both states of affairs. This is the case if we consider linguistic conventions, but also, in our propaedeutical example, if we consider the text of the sentence in which the word "water" appears and we use encyclopaedic knowledge (e.g. physical states of water). In general, then, how is it possible to determine the speaker's intended meaning when a sentence admits to various meanings all plausible in a context of use (or any combination of contexts), that is, when the context seems not to be conclusive?

Intended Meaning and Speaker's Meaning

In this section I illustrate the notion of speaker's meaning proposed by Grice (1989) in regard to the notions of "what is said" (explicit level of communication) and "what is communicated" (implicit level of communication). Speaker's meaning corresponds to "what is communicated" by a speaker with a sentence. "What is communicated" is understood by an interlocutor by mean of an inference (i.e. conversational implicature) based on the conventional meaning of the sentence and contextual information concerning the situation in which the sentence is uttered. Relevant information is made accessible to the interlocutor by means of speaker's communicative intentions. Essentially, Grice argues that conventional meaning, completed with treatment of ambiguity and indexical expressions (*latu senso*), determines a unique proposition ("what is said"), and he examines the implicit communicative process based on it.

Most philosophers, linguists, and relevant theorists agree on the notion that the speaker's meaning is "what is

[2] Note that Kaplan (1989), when introducing the "directing intention", admits a 'cognitive turn' for the reference of demonstrative pronouns (see Bianchi, 2006).
[3] On possessive expressions see Clark (1992).
[4] On background of meaning see Searle (1980).

communicated" by a speaker (implicit level of communication). But they do not agree on the role of conventional meaning in fixing "what is said" by a speaker (explicit level of communication) (see Sperber & Wilson, 1986; Carston, 1988, 2002; Recanati, 1989, 1993; Travis, 1997; Levinson, 2000; Bianchi, 2004). They consequently argue that we need some inferential (or associative) processes based on contextual information (e.g. free enrichment, transfer, saturation, bridging, narrowing, broadening, etc.). These processes fix a unique proposition (the one *explicitly* expressed by a sentence). In other words, these processes fix "what is said" by a speaker with a sentence in a specific situation. Consequently, conversational implicature determines, on the basis of "what is said" and further contextual information, "what is communicated" by a speaker. However, not all pragmatists agree on the temporal sequence of the above processes. Some of them maintain that implicature works in parallel with free enrichment, transfer, etc.[5] However, according to (a weak version of) contextualism in pragmatics, my proposal in regard to intended meaning concerns the level of "what is said" (explicit level of communication).[6]

I consider that if pragmatic processes, based on non-linguistic contextual information made accessible by communicative intentions, are not sufficient to determine a unique set of truth conditions (proposition), then, in order to determine the explicit level of communication, we can take into account the speaker's preferences for the states of affairs which make a sentence true.

State of Affairs and Preference

In this section I outline the notion of preference for a state of affairs based on the comparative notions: "better than" (>), "equal in value to" (≡) and "at least good as" (≥) taken from decision theory (see Hansson, 1994). Using this language, it is possible to express the preferences of agents for states of affairs. For instance, on writing: $[(sa_1)>(sa_2)]_{Ag}$, we assert that an agent prefers the state of affairs 1 rather than the state of affairs 2.

Decisions theorists assume that a rational agent *correctly* chooses an option if the ordering of options realizes certain properties: ordering, continuity, independence (see Myerson, 1991). For my purposes here, it is sufficient to consider the property of ordering, which concerns completeness and transitivity. Completeness for weak preference is defined as follows:

the relation ≥ is complete if and only if for any elements A and B of its domain, either A ≥ B or B ≥ A.

Transitivity for weak preference is defined as follows:

the relation ≥ is transitive if and only if it holds for all elements A, B and C of its domain, so that if A ≥ B and B ≥ C, then A ≥ C.

These properties ensure that an agent is able to compare some options coherently with his/her own interest. However, it is possible that an agent is not always able to compare all options clearly, but this does not prevent him/her from choosing coherently with his/her own interest. In our case, we can consider an agent as preferring one state of affairs coherently with his/her own interest if s/he chooses in accordance with the rule which states:

an alternative is uniquely the best if and only if it is better than all the other alternatives. If there is uniquely a best alternative, choose it (see Hansson, 1994).

Hence, in order to consider an agent's choice coherent with his/her interest, it is sufficient that s/he is able to determine the best state of affairs among others without necessarily ordering the other states of affairs. In this case, a partial ordering is sufficient to consider agents rational.

A Case of Syntactic Ambiguity

In this section I illustrate a case of structural ambiguity where support by the context is not sufficient to determine the state of affairs to which a sentence refers. I cite a case of linguistic controversy provoked by a labour agreement stipulated by a firm and a local trade union. The agreement stated the modes, schedules and procedures for the placement of redundant workers on a publicly-funded wages guarantee scheme and their job mobility.

The situation was as follows: the firm was attempting to turn around its economic-financial performance (economic reorganization) and had begun the procedures for the placement of redundant workers on the public wages guarantee scheme and for job mobility. To make the procedures lawful, the firm stipulated a collective company-level agreement with the local trade union to order to manage surplus workers. During the procedures, a controversy arose in regard to the one-off payment of a sum of money as an incentive for voluntary redundancy (as provided for the agreement). The controversy developed around two different interpretations of a specific clause in the agreement. The clause was the following:

"The firm shall pay a lump sum to workers accepting voluntary redundancy during the wage guarantee fund's validity (…)".[7]

The linguistic controversy concerned whether the expression: "during the wage guarantee fund's validity" referred to "the firm" or to "workers accepting voluntary redundancy", and therefore, whether only redundant workers who resigned would receive the sum of money or whether all workers (both redundant and still employed) who resigned in that period would receive the lump sum payment as an incentive.

Which state of affairs made the clause true?

[5] On the notion of explicature in Relevant Theory (see Carston, 1988); on impliciture in linguistics (see Bach, 1994).

[6] However, communication can succeed at the "what is said" level, for instance, in contracts and scientific texts. Note that when communication takes place at the implicit level a speaker can retract his/her statements; instead, when communication happens at the explicit level, s/he cannot freely retract.

[7] In Italian the clause is as follows: "L'azienda riconoscerà al personale dimissionario nel periodo di vigenza della Cassa Integrazione Guadagni straordinaria un importo forfetario *una tantum* (…)".

States of Affairs and Goals

The state of affairs preferred by the firm corresponded to its interest that only workers covered by the wage guarantee fund resigned, so that the firm could use the fund (it was limited by the agreement) for other employed workers (who took place of workers who accepted voluntary redundancy) and thus avoid paying their wages. The goal of the firm was to reduce the total amount of one-off incentives, to reduce payment of wages and to complete its restructuring. The state of affairs which comprised the possibility conditions for achievement of the firm's goal can be expressed as follows: 'the firm pays a lump sum for voluntary redundancy to *only* workers on the wages guarantee fund'.

The state of affairs preferred by the trade union corresponded to its interest in extending to all workers the possibility of receiving the lump sum for voluntary dismissal during the period of the redundancy payment scheme. The goal of the trade union was to improve the economic circumstances of workers as much as possible. The state of affairs which comprised the possibility conditions for achievement of the trade union's goal can be expressed as follows: 'the firm pays a lump sum for voluntary redundancy to *all* workers (those on the wages guarantee fund and those in employment) during the period of the wage guarantee fund's validity'.

Intended Meaning and Negotiation

In short, the goal of the firm was to reduce total wages, reduce total incentives and complete its restructuring. The interest of the firm was to move employed workers from regular employment to placement on the public wages guarantee scheme. The goal of the trade union was to improve the economic circumstances of workers. The interest of the trade union was to enable all workers to receive the sum of money.

The state of affairs was negotiated as follows: the firm gave eligibility to incentive to all workers. It thus obtained stability of the company-level agreement avoiding the risk of halting the reorganization and having to return the money already furnished by the state for wage guarantee fund. The trade union gave stability to the agreement by confirming its validity and obtaining the voluntary redundancy incentive for all workers. The state of affairs fixed in the negotiation was compatible with the truth conditions which made one interpretation of the clause true and excluded the other interpretations.

Two Cases of Semantic Ambiguity

In this section I illustrate two cases which concern two linguistic disputes provoked by the same clause in a nation-wide collective agreement stipulated by a trade union and Confindustria (corresponding to the British CBI). I show that the two different negotiations of interests gave rise to two different intended meanings in two very similar contexts. The clause was the following:

"The parties agree on working hours, which apply also to groups of workers, with respect to flexibility regarding the seasonality of products […]. The parties further agree that, at company level, the modes and schedules of implementation will be agreed with the local trade union representatives".[8]

The dispute centered on the expression "seasonality of products". The two interpretations were:

(a) 'seasons of the year';
(b) 'peaks in the market',

respectively in both cases. The clause's meaning was important because of its impact on the criterion for implementing flexibility measures. In both cases the respective interpretations were the same: in the former case the company adopted interpretation (b), and the local trade union adopted interpretation (a). Analogously, in the latter case, another company adopted interpretation (b) and the same trade union adopted interpretation (a). In my view, it is very interesting that the same agent was involved in both situations and negotiated the same interpretations with different agents. In particular, I would stress that, in the two negotiations, different interests induced the same agent (the trade union) to select two different meanings in two very similar contexts.

Before I report the two cases I shall briefly present the notion of 'meaning negotiation'. According to Bouquet and Warglien, agents have a meaning negotiation problem whenever they have:

"the problem of reaching an agreement on the meaning of an expression when an agreement is valuable for all agents, but agents have conflicting preferences over which solution should be selected, so that every agreement implies that at least someone has to concede to some extent to other agent" (Bouquet & Warglien, 2002, p. 2).

In what follows, I shall show how agreement on situational extra-semantic interests selects which is the intended meaning in linguistic disputes.[9]

Case 1

In case 1, the term "flexibility" in the clause meant that the company, during some periods of the year, could require its employees to work a large amount of overtime. Overtime was required on Saturdays or in addition to the daily regular working hours. The company compensated overtime with paid rest days taken in other periods of the year. Essentially, the clause regulated the times and ways in which the company could require overtime and compensate it with paid rest days.

The company was interested in managing working hours with discretionary power in order to save money,

[8] The clause in Italian is as follows: "Le parti convengono, a titolo di flessibilità sulla stagionalità dei prodotti e per le attività di installazione e montaggio, sull'orario plurisettimanale, da realizzarsi anche per gruppi di lavoratori". […] "Le parti altresì concordano che, a livello aziendale, verranno convenute, tramite accordo, le modalità di attuazione oltre che i tempi di implementazione dell'orario settimanale di cui al presente punto con le rappresentanze sindacali unitarie e le organizzazioni sindacali territoriali".

[9] As Clark puts it, "we cannot hope to understand language use without viewing it as joint action built on individual actions. The challenge is to explain how all these actions work" (Clark 1996, p. 4).

possibly on the basis of information unavailable to the local trade union (e.g. orders). In particular, the interest of the company was to be able to use overtime without paying the wage supplements due and to distribute the cost of paid rest days among periods according to its needs (its discretion). Moreover, the company was interested in being able to resort to overtime at any time of the year on the basis of market demand, and it was not interested in hiring new personnel or in paying overtime regularly. On the other hand, the trade union was interested in reducing (or avoiding) the use of overtime and particularly if it was not regularly paid, in favouring the right to rest and to plan free time. It was also interested in inducing the company to hire new personnel and/or pay overtime regularly.

The company argued that overtime should be regulated with respect to peaks in the market: specifically, the company could resort to unpaid overtime at any time of the year on the basis of market demand. The company could not know peaks in the market in advance and thus could not fix a specific period *a priori*. The trade union argued that overtime should be regulated with respect to the seasons of the year in which the company's products were most in demand, spring in particular.

The two interpretations were both plausible in the situation, where the relevant combination of contexts consisted of the linguistic context, i.e. the text of the clause; the encyclopeadic knowledge, i.e. the contract's rules (e.g. civil code); and local knowledge, i.e. the specific shared activity which the clause regulated. At this point the parties attempted to reach an agreement by negotiating their interests. And, in the end the company and the trade union fixed the intended meaning whereby "seasonality of products" stood for "season of year when products are particularly in demand"; in particular a 'positive season' was spring and a 'negative season' was autumn. They agreed that, in a positive season, the company could utilize non-regularly-paid overtime, while in the negative season overtime was recompensed with paid rest days. How did the agents determine the intended meaning? How did the negotiation of interests work?

The company obtained high discretionary power to utilize unpaid overtime in the positive season (from March to June) *de facto* independently of peaks in the market, and to arrange paid rest days in a period of year when it did not need labour, that is, during the negative season (from September to December). The company relinquished overtime throughout the year (except in the positive season) and discretionary power to distribute paid rest days during the negative season. The trade union obtained a reduction in unpaid overtime (except in the positive season) and the right of employees to choose which days to use for paid rest during the negative season. Moreover, the trade union induced the company to hire new personnel or to pay overtime regularly (except in the positive season). The trade union relinquished to check overtime in the positive season. Finally, the trade union relinquished the possibility of distributing paid rest days throughout year, in that they could only be taken in the negative season.

The agents' interests were mediated with respect to the specific situation: each party gave up something in favour of the other party, and meaning (a) (compatible with the agreement reached) was finally fixed.

Case 2

In case 2, the term "flexibility" meant that the company could hire temporary workers and manage working hours and shifts according its needs.

The company was interested in hiring temporary workers on the basis of increased orders in any period of the year. The company was also interested in managing temporary workers because of information unavailable to the trade union. The trade union was interested in reducing temporary work; in particular, it was interested in restricting the use of temporary labour to only limited periods of the year. Moreover, the trade union was interested in reducing the use of temporary workers and in changing temporary jobs into salaried ones (on both permanent and fixed-term contracts).

The company claimed that the use of temporary labour must be regulated in accordance with peaks in the market: that is, at any time of the year on the basis of market demand. The company could not know peaks in the market in advance and thus could not fix a specific period *a priori*. The trade union claimed that the use of temporary labour must be regulated according to the seasons of the year in which the company's products are most in demand, summer in particular.

The two interpretations were both plausible in the situation, where the relevant combination of contexts consisted of the linguistic context, i.e. the text of the clause; encyclopaedic knowledge, i.e. the rules of contracts (e.g. civil code); and local knowledge, i.e. the specific shared activity which the clause regulated. At this point the parties attempted to reach an agreement by negotiating their interests. And, in the end, the company and the trade union fixed the intended meaning whereby "seasonality of products" stood for "peaks in the market". How did the agents determine the intended meaning? How did the negotiation of interests work in this case?

The company and the trade union reached an agreement in which the employer could use, in the case of peaks in the market throughout the year, an amount of temporary labour representing only ten percent of salaried labour (employees). Hence the company obtained high discretionary power throughout year, but only for a limited number of workers. The trade union obtained a reduction in the use of temporary labour (ten percent of the workforce), but relinquished control over it. Finally, it lost bargaining power on new hirings.

The agents' interests were mediated with respect to the specific situation: each party gave up something in favour of the other party, and meaning (b) (compatible with the agreement reached) was finally fixed.

In the two negotiations, two different meanings were determined for the same expression on the basis of two different negotiations of interests; even the same agent determined two different intended meanings with regard to the two different interests. It is in this sense that situational interest drives the determination of intended meaning.

Conclusion

The paper has presented a notion of intended meaning for declarative sentences. Its argument has been based, on the one hand, on meaning as truth conditions and, on the other, on interest as a state of affairs preferred by a speaker because it implies his/her goal. The compatibility

of the two notions is centered on the notion of state of affairs.

The notion of intended meaning presented in the article is compatible with that of "what is said" in pragmatics: that is, it represents the explicit level of communication. It is compatible with "what is said" because it is fixed by means of pragmatic processes based on information not constrained to the linguistic form of the sentence. But it differs from "what is said" because of the kind of information used: essentially, the pragmatic context refers to items in the current situation or past situations (linguistic context, shared knowledge, physical surroundings). Instead, my approach also takes into account future states of affairs related to agents' goals.

On this view, the truth conditions which make a sentence true can be fixed by means of the commitment of agents to realizing a certain state of affairs. However, agents do not fix meaning freely; rather, they are constrained by sets of truth conditions previously selected by a combination of relevant contexts in the specific situation.

We have seen three cases of structural and semantic ambiguity where semantics, which should be able to fix meaning in these kind of cases, failed. We have also seen that standard pragmatic information is not conclusive in fixing the intended meaning; as a consequence, the situational interests of agents have been taken into account. In conclusion, this notion seems to be adequate to express intended meaning, given a set of contextually plausible interpretations, both in cases of communicative processes to determine speaker's intended meaning and in cases of negotiation to resolve linguistic disputes.

Acknowledgements

I thank Lawyer Dr. Sonia Guglielminetti for helpful discussions and advices on several cases of controversy.

This research is supported by the Okkam project (co-funded by the European Commission - GA 215032) – www.okkam.org

References

Bach, K. (1994). Conversational impliciture. *Mind & Language*, 9, 2, 124-162.

Bach, K. (2004). Minding the gap. In C. Bianchi (Ed.), *The pragmatics/semantics distinction*. Stanford: Csli.

Bianchi, C. (2004). Semantics and pragmatics: distinction reloaded. In C.Bianchi (Ed.), *The pragmatics/semantics distinction*. Stanford: Csli.

Bianchi, C. (2006). Nobody loves me: quantification in context. *Philosophical Studies*, 130, 2, 377-397.

Bouquet, P. & Warglien, M. (2002). Meaning negotiation: an invitation. In P. Bouquet (Ed.), *Meaning negotiation papers from the AAAI workshop*. Edmonton: AAAI Press.

Carston, R. (1988). Implicature, explicature and truth-theoretic semantics. In R. Kempson (Ed.), *Mental representations. Interface between language and reality*. Cambridge: Cambridge University Press.

Carston, R. (2002). Linguistic meaning, communicated meaning and cognitive pragmatics. *Mind and Language*, 17, 1-2, 127-148.

Clark, H.H. (1992). *Arena's of language*. Chicago: The University of Chicago Press & Csli.

Clark, H.H. (1996). *Using language*. Cambridge, Cambridge University Press.

Conte, R., & Castelfranchi, C. (1995). *Cognitive and social action.* London: University College London.

Cruciani, M. (2009a). Cono del linguaggio, negoziazione degli interessi e significato inteso di clausole contrattuali ambigue. *Sistemi Intelligenti*, 21, 3, 473-88.

Cruciani, M. (2009b). Intended meaning and situational interest. *Proceedings of the 31° Conference of Cognitive Science Society* (pp. 2747-2752). Austin, (TX): Cognitive Science Society.

Grice, P. (1957). Meaning. *Philosophical Review*, 66, 377-88.

Grice, P. (1989). *Studies the way of words*. Cambridge: Cambridge University Press.

Hansson, S.O. (1994). *Decision theory. A brief introduction*. Dept. Philosophy and the History of Technology, Royal Institute of Technology, Stockholm. http://www.infra.kth.se/~soh/decisiontheory.pdf

Kaplan, D. (1977). Demonstratives. An essay on semantics, logic, metaphysics, and epistemology of demonstratives and other indexicals. In J. Almong, J. Perry & H. Wettstein (Eds.), *Themes from Kaplan*. Oxford: Oxford University Press, 1989.

Kaplan, D. (1989). Afterthoughts. In J. Almong, J. Perry & H. Wettstein (Eds.), *Themes from Kaplan*. Oxford: Oxford University Press.

Kripke, S. (1979). Speaker's reference and semantic reference. In A. French, et al. (Eds.) *Contemporary perspective in the philosophy of language*. Dordrecht: Reidel.

Levinson, S. (2000). *Presumptive meanings. The theory of generalized conversational implicature*. Cambridge (MA): MIT Press.

Myerson, R.B. (1991). *Game theory*. Cambridge: Harvard University Press.

Perry, J. (1997). Indexical and demonstratives. In R. Hale & C. Wright (Eds.), *Companion to the philosophy of language*. Oxford: Blackwell.

Recanati, R. (1989). The pragmatics of what is said. *Mind & Language*, 4, 4, 207-32.

Recanati, R. (1993). *Direct reference: from language to thought*. Oxford: Blackwell.

Recanati, F. (2001). What is said. *Synthese*, 128, 75-91.

Searle, J. (1979). *Expression and meaning*. Cambridge: University Press.

Searle, J. (1980). The Background of meaning. In J. Searle, F. Kiefer and M. Bierwisch (Eds.), *Speech act theory and pragmatics*. Dordrecht: Reidel.

Searle, J. (1992). *The rediscovery of mind*. Cambridge: MIT.

Sperber, D., & Wilson, D. (1986). *Relevance theory: communication and cognition*. Oxford: Blackwell.

Tarski, A. (1944). The semantic conception of truth and the foundation of semantics. *Philosophical and Phenomenological Research*, 4, 341-56.

Travis, Ch. (1975). *Saying and understanding*. Oxford: Blackwell.

Travis, Ch. (1981). *The true and false*: *the domain of pragmatics*. Benjamins: Amsterdam

Travis Ch. (1997). Pragmatics. In R. Hale & C. Wright (Eds.), *A companion to the philosophy of language*. Oxford: Blackwell.

Wittgenstein, L. (1921). *Tractatus logico-philosophicus*. Oxford: Blackwell.

Predicative Metaphor Comprehension as Indirect Categorization

Akira Utsumi (utsumi@se.uec.ac.jp) **Maki Sakamoto (sakamoto@hc.uec.ac.jp)**

Department of Informatics, The University of Electro-Communications,
1-5-1, Chofugaoka, Chofushi, Tokyo 182-8585, Japan

Abstract

In this paper, we address the problem of how people understand predicative metaphors such as "The rumor flew through the office," and argue that predicative metaphors are understood as indirect categorizations. In the indirect categorization process, the verb (e.g., *fly*) of a predicative metaphor evokes an intermediate entity, which in turn evokes a metaphoric category of actions or states (e.g., "to spread rapidly and soon disappear") to be attributed to the target noun (e.g., *rumor*), rather than directly creating a metaphoric category as argued by Glucksberg's (2001) categorization theory. We test our argument using two experiments, offline comprehension and online priming. The two experiments provided convergent evidence for our argument. The psychological validity of indirect categorization as a process of predicative metaphor comprehension was confirmed.

Keywords: Metaphor comprehension; Predicative metaphor; Categorization; Priming; Verb

Introduction

Predicative metaphors are figurative expressions that involve the metaphorical use of a verb, such as "*The rumor flew through the office*" and "*His fame echoes throughout the world*." Despite their frequent use in everyday communication, predicative metaphors have been paid little attention in metaphor research. Particularly, the cognitive mechanism underlying predicative metaphor comprehension has never been examined, although a considerable number of studies have been made on the comprehension mechanism of nominal metaphors such as "*My job is a jail*" (e.g., Bowdle & Gentner, 2005; Glucksberg, 2001; Jones & Estes, 2006; Utsumi, 2007). Given the differences in what is being processed metaphorically between predicative metaphors (i.e., actions, states) and nominal metaphors (i.e., objects), together with a recent neuroanatomical finding (Chen, Widick, & Chatterjee, 2008) that predicative and nominal metaphors may be processed differently, it is obviously crucial to explore the cognitive mechanism of predicative metaphor comprehension.

Cognitive linguists may argue that the cognitive linguistics research on metaphor (e.g., Kövecses, 2002; Lakoff & Johnson, 1980) has addressed predicative metaphors as manifestations of the conventionalized, conceptual metaphors. However, these studies do not explore how the conceptual metaphors are constructed, i.e., how a set of correspondences or mappings is made between the source domain and the target domain. This problem becomes more serious when we consider how people comprehend novel predicative metaphors.

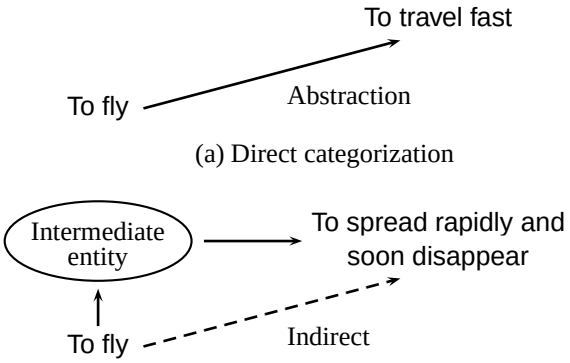
Glucksberg (2001, 2003) argues that people comprehend predicative metaphors via a categorization process as they do for nominal metaphors. Just as nominal metaphors use the source concepts that epitomize certain categories of objects or situations, predicative metaphors use verbs that epitomize certain metaphoric categories of actions (e.g., the category of *speedy travel* evoked by the verb "fly"). However, no clear empirical evidence has been provided for his argument. Although Torreano, Cacciari, and Glucksberg (2005) demonstrated that the level of abstraction of a verb's referent was related to the metaphoricity of a predicative metaphor, this finding does not necessarily imply that the verb directly evokes a metaphoric category in metaphor comprehension.

In this paper, we propose *indirect categorization* as the comprehension process of predicative metaphors (Utsumi & Sakamoto, 2007b). Indirect categorization is a two-stage process of categorization in which evocation (or creation) of metaphoric categories is indirect and mediated by intermediary entities, rather than direct as predicted by the categorization theory. Utsumi and Sakamoto (2007b) suggested a possibility of indirect categorization using a computer simulation, but no clear empirical evidence has been provided. Therefore, in this paper we conducted two psychological experiments to obtain empirical evidence for our indirect categorization theory. In these experiments, we manipulated metaphor aptness and vehicle conventionality because recent metaphor studies (e.g., Bowdle & Gentner, 2005; Glucksberg & Haught, 2006; Jones & Estes, 2006) have demonstrated that these properties play an important role in comprehension of nominal metaphors.

In Experiment 1, we examined what proportion of interpretations of predicative metaphors were derived directly from the verb and what proportion of interpretations were indirectly associated with the verb. For this purpose, we assessed a concordance rate between words listed as metaphorical interpretation and those associated with the verb or associated with the verb associates. In Experiment 2, we used a priming paradigm to assess the online availability of direct and indirect categories for predicative metaphor comprehension. In this experiment, a metaphorical sentence was presented as a prime and its effect on the speed of lexical decision about a subsequent target word was measured. The target conditions were a word related to the metaphorical meaning, a word directly associated with the verb, a word indirectly associated with the verb, and a control word unrelated to the metaphor.

Direct versus Indirect Categorization

As we mentioned above, Glucksberg's (2001, 2003) categorization theory argues that people understand predicative metaphors as direct categorizations. Just as nominal metaphors use vehicles (or source concepts) that epitomize certain superordinate categories of objects, which include a target concept as a member, predicative metaphors use verbs that epitomize certain categories of actions. According to this theory, for example, the predicative metaphor "The rumor flew through the office" is comprehended so that the verb *fly* evokes an ad hoc



(a) Direct categorization

(b) Indirect (two-stage) categorization

Figure 1: Direct and indirect categorization for the metaphor "The rumor flew through the office."

superordinate category of an action "to travel fast" and such the action is attributed to the target *rumor*, as illustrated in Figure 1 (a).

However, it is doubtful that predicative metaphors are processed in the same way as nominal metaphors. A primary reason for this doubt is that many empirical findings on semantic representation demonstrate that the semantic structure of verbs, which refer to events or actions, differs from that of nouns, which refer to objects, in many respects (Vigliocco & Vinson, 2007). For example, the hierarchical organization for objects and events is different; event categories are represented by fewer levels (generally two) and with fewer distinctions at the superordinate level than object categories. The role of hierarchical relations also differs between nouns and verbs. For nouns, the most important roles are played by the hierarchical relations including superordination and coordination, whereas the dominant relations for verbs are nonhierarchical ones such as entailment, causation, and antonymy. Some evidence compatible with the different role of hierarchical relations is provided by the analysis of semantic substitution errors; Garrett (1992) reported that for nouns the large majority of substitutions involve category coordinates (i.e., words in the same level of the hierarchical structure), while for verbs the preferred semantic relationship between target and intruding words is opposition (e.g., go/come). Furthermore, a neuroanatomical difference appears to exist between nouns and verbs (Shapiro & Caramazza, 2004; Vigliocco & Vinson, 2007) and between nominal metaphors and predicative metaphors (Chen et al., 2008). These findings indicate that hierarchical relations are less activated in the processing of verbs, and thus it is less likely that verbs directly evoke superordinate categories of events or actions; this contradicts Glucksberg's categorization theory.

Furthermore, the categorization theory does not address the richness of the metaphorical meanings expressed by predicative metaphors. For example, people can derive more meanings from the metaphor "The rumor flew through the office" than supposed in the categorization theory (e.g., *to travel fast*); the rumor spreads rapidly and suddenly, the rumor is dispersed or disseminated, the rumor disappears or is forgotten very soon, and so on. These rich interpretations are unlikely to be derived directly from the verb *fly*, given that the semantic structure of verbs is hierarchically less rich.

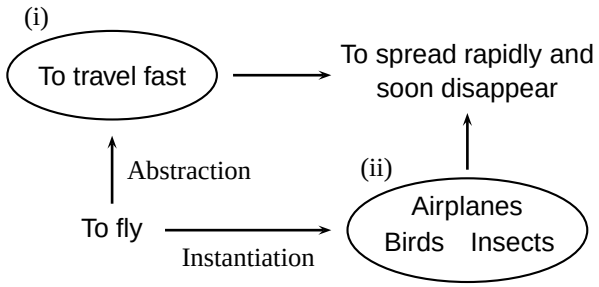


Figure 2: Two possibilities of an intermediate entity in indirect categorization.

To overcome the difficulties of the categorization theory of predicative metaphors, we propose an indirect categorization theory. The intuitive idea behind indirect categorization is that a correspondence between the actions or events literally expressed by the verb and the actions or events to be attributed to the target noun would be indirect, rather than direct as predicted by the categorization theory; constructing a correspondence is mediated by an intermediate entity. As illustrated in Figure 1 (b), in the case of "fly" metaphor, the verb *fly* first evokes some sort of an intermediate entity and the intermediate entity then evokes a final abstract category of "to spread rapidly and soon disappear," which is attributed to the target *rumor* being described.

One important question that arises here is what kind of entities are involved in the intermediate step. Two possible answers can be provided: (i) abstract actions or states produced by generalization from the verb, and (ii) objects produced by instantiation of the verb. In the case of "fly" metaphor, as illustrated in Figure 2, people may think of a very abstract action "to travel fast" by abstracting the verb *fly*, and this abstract intermediate entity is then specified to refer to *rumor*. A perhaps more likely explanation would be that people may consider a set of objects "things that fly" or "flying objects," which contains airplanes, birds, and insects, by instantiating the argument of the verb *fly*. Some actions or events that are relevant to both the "flying" objects and the target *rumor* are then extracted. These two types of intermediate entities may be activated simultaneously during comprehension, rather than selectively. The preference for instantiated objects (ii) may be determined depending on the difficulty in deriving an abstract category from a verb.

Experiment 1

In Experiment 1, we tested our indirect categorization theory by comparing people's interpretations of predicative metaphors (i.e., $I(M)$ in Figure 3), with words or phrases associated directly or indirectly with the verb of predicative metaphors (i.e., $A(w_v)$ or $A(S)$ in Figure 3). If a metaphoric category is evoked indirectly in predicative metaphor comprehension, the interpretation of predicative metaphors $I(M)$ would have greater overlap with indirectly associated words $A(S)$ than with directly associated words $A(w_v)$. If a metaphoric

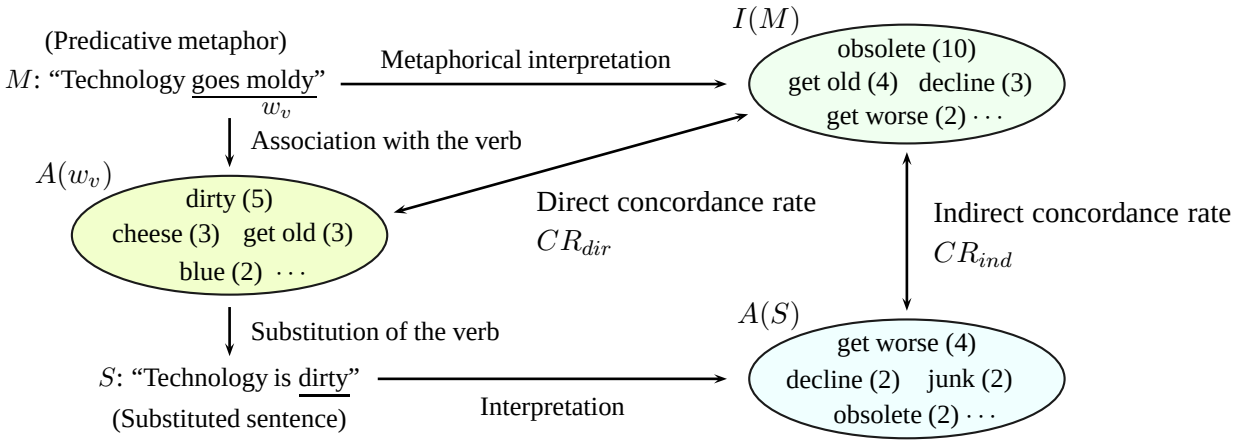


Figure 3: Direct concordance rate and indirect concordance rate as measures of the degree of overlap between the metaphorical interpretation and the direct or indirect associates.

category is evoked directly from the verb, the percentage of overlap between the metaphorical interpretation $I(M)$ and a set of directly associated words $A(w_v)$ would be greater than, or at least equal to, that between the interpretation and a set of indirectly associated words $A(S)$.

Our indirect categorization view therefore predicts that, regardless of metaphor aptness and vehicle conventionality, the interpretation of predicative metaphors has greater overlap with indirectly associated words $A(S)$. On the other hand, Glucksberg's categorization view predicts that, regardless of vehicle conventionality, the interpretation of predicative metaphors (in particular apt metaphors) has greater (or equal) overlap with directly associated words $A(w_v)$.

Method

Participants Eighty-eight people (78 undergraduate and graduate students and 10 working persons) participated as volunteers. All participants were native speakers of Japanese.

Materials Forty Japanese predicative metaphors were used for the experiment. These metaphors were selected from 80 metaphors in a pilot study.

Pilot study For a pilot study, we used 80 Japanese predicative metaphors. They included 20 intransitive verbs (e.g., "go moldy" ["*kabiru*" in Japanese] [1] or "echo" ["*hibiku*"]) and each verb was paired with four abstract nouns (e.g., "Technology goes moldy" ["*Gijutsu ga kabiru*"], "His fame echoes" ["*Meisei ga hibiku*"]). , In order to eliminate the possibility that the generated sentences were interpreted as personification metaphors, in which the subject of the sentence, rather than the verb, was used metaphorically, we did not use verbs that literally refer to human actions or experiences.

In this pilot study, we collected the aptness and conventionality ratings to select 40 metaphors used in the main study. Because the conventionality rating task requires the salient meaning of predicative metaphors, the pilot study was conducted separately in two parts. In the first part of the pilot study, 50 participants were assigned 40 metaphors such that each metaphor was assigned to 25 participants. They were asked to write down at least three interpretations of each metaphor and to rate the aptness of the metaphor on a 7-point scale (1 = *not at all apt*, 7 = *extremely apt*). A list of generated interpretations for each metaphor was used as a set $I(M)$. In the second part, 15 participants were given a list of 80 verbs used in the metaphors with the most salient meaning of the metaphors, i.e., the meaning listed by the largest number of participants in the first part of the pilot study. They were asked to rate how conventional each meaning was as an alternative sense of the verb on a 7-point scale of 1 (*very novel*) to 7 (*very conventional*).

After the pilot study, we chose 40 metaphors for the main study in the following way. First, we calculated the mean aptness rating and the mean conventionality rating for each metaphor. We then classified the 80 metaphors into four groups — conventional and high apt, conventional and low apt, novel and high apt, and novel and low apt — according to whether the mean aptness or conventionality was more than the midpoint 4. Finally, we chose 10 metaphors from each group such that metaphors in the same group had as different verbs as possible and their variance of aptness and conventionality was as low as possible.

Procedure In the experiment, we collected words or phrases associated directly or indirectly with the verb. The experiment was conducted separately in two parts because direct verb associates $A(w_v)$ were required for substituted sentences, from which indirect verb associates $A(S)$ were collected, as shown in Figure 3.

In the first part of the experiment, 12 participants were assigned all 16 verbs which were used in the 40 chosen metaphors, and asked to list at least two words or phrases that they associated with each verb. A list of generated words for each verb was used as a set $A(w_v)$ of direct verb associates.

The second part was performed by other 11 participants. They were assigned 40 substituted sentences and asked to list at least two words or phrases that they thought were involved in the interpretation of substituted sentences. A list

[1] Note that the original Japanese verb "*kabiru*" is a verb, although its English translation "go moldy" is a verb phrase.

Table 1: Means (*M*) and standard deviations (*SD*) of concordance rates between metaphor interpretation and direct or indirect association.

	CR_{dir} (Direct)		CR_{ind} (Indirect)	
Metaphor group	*M*	*SD*	*M*	*SD*
Conventional, High-Apt	.256	.202	.391	.279
Conventional, Low-Apt	.172	.135	.408	.318
Novel, High-Apt	.201	.171	.354	.240
Novel, Low-Apt	.125	.103	.368	.167
All	.189	.159	.380	.248

of generated words for the substituted sentences was used as a set $A(S)$ of indirect verb associates. Substituted sentences were generated by substituting three words in $A(w_v)$ listed by the largest number of participants for the verb w_v of the metaphor. For example, when three words "dirty" ["*kitanai*"], "cheese" ["*chīzu*"], and "get old" ["*furuku-naru*"] were listed by the largest number of participants for a verb "go moldy," the substituted sentence of a predicative metaphor "Technology goes moldy" was "Technology is dirty," "Technology is cheese," and "Technology gets old."

After the experiment, we generated three sets of words for each metaphor, namely $I(M)$, $A(w_v)$, and $A(S)$ in the following way. First, closely related words or phrases were accepted as the same word if they belonged to the same deeper category of a Japanese thesaurus. After that, any word that was mentioned by only one participant was eliminated from the set of words.

Results and Discussion

As shown in Figure 3, in order to assess the degree of overlap between the metaphorical interpretation and the direct or indirect verb associates, we calculated the direct concordance rate CR_{dir} and the indirect concordance rate CR_{ind} for each metaphor M:

$$CR_{dir} = \frac{\sum_{x \in I(M) \cap A(w_v)} n_I(x) + n_A(x)}{\sum_{x \in I(M)} n_I(x) + \sum_{x \in A(w_v)} n_A(x)} \quad (1)$$

$$CR_{ind} = \frac{\sum_{x \in I(M) \cap A(S)} n_I(x) + n_S(x)}{\sum_{x \in I(M)} n_I(x) + \sum_{x \in A(S)} n_S(x)} \quad (2)$$

where $n_I(x)$, $n_A(x)$ and $n_S(x)$ respectively denote the number of participants who listed a word x as a metaphorical interpretation, a verb associate, and an associate of the substituted sentences. (The numbers in parentheses in Figure 3 represent these values.) The direct concordance rate CR_{dir} defined by Equation 1 evaluates the degree of overlap between metaphorical interpretation and direct verb association, while the indirect concordance rate CR_{ind} defined by Equation 2 evaluates the degree of overlap between metaphorical interpretation and indirect association. For example, the direct concordance rate of the example shown in Figure 3 is calculated as $CR_{dir} = (4+3)/\{(10+4+3+2)+(5+3+3+2)\} = 7/32 = 0.219$, and the indirect concordance rate is calculated as $CR_{ind} = \{(10+2)+(3+2)+(2+4)\}/\{(10+4+3+2)+(4+2+2+2)\} = 23/29 = 0.793$.

Table 1 shows the mean concordance rates for direct and indirect categorization. Overall, as shown in the last row of Table 1, the mean indirect concordance rate CR_{ind} across the 40 metaphors was higher than the mean direct concordance rate CR_{dir}. This result is consistent with our indirect categorization theory and inconsistent with Glucksberg's categorization theory.

To confirm this difference statistically, we conducted a three-way ANOVA of Categorization (direct or indirect) $\times$ Conventionality (conventional or novel) $\times$ Aptness (high or low). In the analysis, the data were analyzed only by items (F_i) because the concordance rates could not be calculated for each participant. The factor of Categorization was within items and other two factors were between items. The predicted difference between the direct and indirect concordance rate was confirmed; the main effect of Categorization was significant, $F_i(1, 36) = 22.19$, $p < .001$, and the effect size was also large, $\eta^2 = .18$. None of the other main effects and interactions were significant. Hence it is concluded that the result of Experiment 1 supports the indirect categorization theory.

Furthermore, in order to examine which kind of entities were involved in the intermediate step of indirect categorization, we roughly estimated the preference for abstract actions as an intermediate entity by calculating the percentage of verbs and adjectives (i.e., verb rate) that were involved in the set of direct verb associates $A(w_v)$ for each metaphor. The mean verb rate across 40 metaphors was 0.46 (*SD* = 0.17), ranging from 0.20 to 0.77. The correlation between the verb rate and the indirect concordance rate CR_{ind} was far from significant, $r = .06$. This finding suggests that there may be no preferred process (generalization or instantiation) that leads to an intermediate entity; people understand predicative metaphors both by abstracting the verb and by enumerating entities typically expressed by the verb.

Experiment 2

In Experiment 2, we tested the indirect categorization view using a priming paradigm, in which a metaphorical sentence was presented first and the task was to make a lexical decision about a target word presented after the metaphorical sentence. The target conditions were a word related to the metaphorical meaning, metaphor target (MT); a word directly associated with the verb, direct associate target (DAT); a word associated with the substituted sentence, indirect associate target (IAT); and a control target (CNT) unrelated to the metaphor.

Faster lexical decisions in comparison with the CNT indicate on-line activation. If predicative metaphors are comprehended by the direct categorization process, the DAT would be faster to make a lexical decision than the CNT, but the IAT would not be faster. Hence, Glucksberg's categorization theory predicts facilitation of the DAT and no facilitation of the IAT. On the other hand, if predicative metaphors are comprehended by the indirect categorization process, the IAT would be faster than the CNT. Hence, our indirect categorization theory predicts facilitation of the IAT. The DAT may also

Table 2: Means (*M*) and standard deviations (*SD*) of correct lexical decision times in milliseconds for Experiment 2

	MT (Metaphor)			DAT (Direct associate)			IAT (Indirect associate)			CNT (Control)	
Metaphor type	*M*	*SD*	*DIF*	*M*	*SD*	*DIF*	*M*	*SD*	*DIF*	*M*	*SD*
Conventional, High Apt	799.3	210.2	50.1	781.6	194.3	67.8	768.2	169.8	81.2	849.4	205.0
Conventional, Low Apt	864.1	203.2	10.4	859.7	199.4	14.9	797.1	220.5	77.4	874.5	264.9
Novel, High Apt	821.3	236.2	−2.8	856.0	233.8	−37.6	807.4	197.3	11.1	818.4	165.7
Novel, Low Apt	810.7	233.8	31.9	832.4	183.4	10.2	832.7	219.0	9.9	842.6	270.4
All	823.8	182.7	22.4	832.4	170.1	13.8	801.3	168.2	44.9	846.3	195.8

Note. *DIF* = difference from control target.

be activated, but to a lesser degree than the IAT. Concerning the MT, both theories predict facilitation of the MT.

Method

Participants Forty-five undergraduate and graduate students participated as volunteers. All participants were native speakers of Japanese.

Materials The 40 predicative metaphors used in Experiment 1 were employed as prime sentences. The other 40 metaphors that were not selected in the pilot study of Experiment 1 were used as filler sentences for nonword targets.

For each prime metaphor, the MT, DAT, and IAT were selected from among the set of metaphorical interpretations $I(M)$, the set of direct verb associates $A(w_v)$, and the set of indirect verb associates $A(S)$ respectively. For an MT, we selected the word in $I(M)$ that was listed by the largest number of participants. For a DAT and an IAT, we selected the word that was listed by the largest number of participants in $A(w_v)$ or $A(S)$, excluding the MT word. The CNT was selected randomly from a dictionary such that it was not related to the metaphor. For example, the metaphor "Technology goes moldy" was combined with the MT "obsolete" ["*furuku-naru*"], the DAT "dirty" ["*kitanai*"], the IAT "get worse" ["*waruku-naru*"], and the CNT "vanish" ["*toozakaru*"].

Procedure A within-participants design was used with each participant comprehending all the 80 metaphors under all conditions. Participants, who were run individually, were seated in front of a computer screen. They were first given an overall instruction of the experiment and then presented with six practice trials followed by the 80 experimental trials presented in a random order. On each trial, they were presented with a predicative metaphor on the screen for 3000 ms and asked to interpret the metaphor. A target word (MT, DAT, IAT, CNT, or nonword) was then presented 500 ms after the offset of the predicative metaphor. Participants were asked to decide whether the target word was a word or a nonword as quickly as possible; they indicated decision by pressing the appropriate key on the keyboard. Reaction times were measured from the onset of the target word until the appropriate key was pressed.

Results and Discussion

A total of seven participants were eliminated from the analysis because they did not reach the decision error criterion of 90% correct. Only reaction times of correct decision were used in the analysis. Following metaphor priming research (Blasko & Connine, 1993), reaction times greater than 1750ms were eliminated from the analysis. This elimination caused the further elimination of two participants' data because the data of some conditions were missing.

Table 2 shows mean lexical decision times and standard deviations for the correct "yes" responses. The time difference (DIF) from the CNT indicates the extent of the priming effect. Although the pattern of DIF differs depending on conventionality and aptness, the overall result was that the IAT produced the greatest priming effect (44.9ms faster than the CNT), but the DAT showed the smallest priming effect (only 13.8ms faster). The MT showed a moderate priming effect (22.4ms faster). This result is consistent with the indirect categorization theory and inconsistent with the direct categorization theory.

A three-way ANOVA of Target (MT, DAT, IAT, or CNT) $\times$ Conventionality (conventional or novel) $\times$ Aptness (high or low) was conducted on lexical decision times. In the analysis, the data were analyzed by participants (F_p) and by items (F_i). The factor of Target was within participants and within items, while other two factors were within participants and between items. The main effect of Target was significant by the participant analysis, $F_p(3, 105) = 3.21$, $p < .05$, although its effect size was small, $\eta^2 = .01$. The main effect of Target was not significant by the item analysis, $F_i(3, 108) = 1.55$, $p = .21$, but a small effect size was found, $\eta^2 = .03$. Post-hoc pairwise comparisons ($p < .05$) revealed that the IAT (M=801.3ms) was significantly faster than the CNT (M=846.3ms); this indicates a significant activation of indirectly associated meanings during metaphor comprehension. In addition, the difference between the IAT (M=801.3ms) and the DAT (M=832.4ms) was marginally significant ($p < .10$). Again, the result is consistent with the indirect categorization theory but inconsistent with the direct categorization theory; predicative metaphors are understood via the indirect categorization process, in which constructing the correspondence between the actions or events literally expressed by the verb and those expressed metaphorically is mediated by intermediate entities. A little surprisingly, the priming effect of the MT was not statistically significant.

In addition, the interaction between Conventionality and Aptness was significant by the participant analysis, $F_p(1, 35) = 4.32$, $p < .05$. The nature of this interaction

was that, when predicative metaphors were high apt, decision times to all targets were faster for conventional metaphors (M=799.6ms) than for novel metaphors (M=848.8ms), but such the difference disappeared when metaphors were low apt (M=825.8ms for conventional metaphors; M=829.6ms for novel metaphors). The main effect of conventionality was also significant, $F_p(1, 35) = 4.91$, $p < .05$; $F_i(1, 36) = 3.80$, $p = .06$. Mean decision times to all targets were shorter for conventional metaphor primes (M=812.7ms) than for novel metaphor primes (M=839.2ms). These results suggest that vehicle conventionality facilitates comprehension of predicative metaphors, in particular when they are highly apt, but the comprehension process remains unchanged.

General Discussion

The two experiments reported in this paper provided empirical evidence in favor of the proposed view that predicative metaphors are understood as indirect categorizations.

As we mentioned previously, the most important problem with the indirect categorization view is what entities are involved in the intermediate step of indirect categorization. We provide two possible answers, i.e., abstract actions obtained by abstracting the verb, and objects typically expressed by the verb. Experiment 1 suggested that there seemed to be no preference between two possibilities, but we point out that objects typically expressed by the verb are really involved in the comprehension process.

"Float like a butterfly, sting like a bee."

These are the words of Muhammad Ali, a famous American boxer who won World Heavyweight Champion three times. This predicative metaphor expresses Ali's boxing style by describing his swift footwork as "float" and lightning-quick punch as "sting." At the same time, this metaphor clearly conveys a kind of gorgeousness and sharpness in his behavior, which cannot be derived solely from these verbs. It is more likely that such the interpretation would be derived when people call to mind "things that float" and "things that sting," and in the case of this metaphor they are verbalized in "like a butterfly" and "like a bee." In other words, these phrases suggest the psychological reality of the intermediate entities that are typically expressed by the verb for indirect categorization.

We have also argued that the indirect categorization view explains adjective metaphor comprehension (Utsumi & Sakamoto, 2007a). Because the semantic structure of adjectives is not at all hierarchical, intermediate entities only include objects with the property referred to by the adjective of a metaphor (i.e., "things that are red" in the case of the metaphor "red taste"). Some evidence for the predominance of intermediate objects is provided by Nakamura, Sakamoto, and Utsumi (2010).

At any rate, it would be vital for future research to explore in more detail the internal process of indirect categorization.

Acknowledgments

This study was supported by a Grant-in-Aid for Scientific Research(C) (No.20500234) from Japan Society for the Promotion of Science.

References

Blasko, D., & Connine, C. (1993). Effects of familiarity and aptness on metaphor understanding. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(2), 295–308.

Bowdle, B., & Gentner, D. (2005). The career of metaphor. *Psychological Review*, *112*(1), 193–216.

Chen, E., Widick, P., & Chatterjee, A. (2008). Functional-anatomical organization of predicate metaphor processing. *Brain and Language*, *107*(3), 194–202.

Garrett, M. (1992). Lexical retrieval processes: Semantic field effects. In E. Kittay & A. Lehrer (Eds.), *Frames, fields and contrasts: New essays in semantic and lexical organization* (pp. 377–395). Erlbaum.

Glucksberg, S. (2001). *Understanding figurative language: From metaphors to idioms*. Oxford University Press.

Glucksberg, S. (2003). The psycholinguistics of metaphor. *Trends in Cognitive Sciences*, *7*(2), 92–96.

Glucksberg, S., & Haught, C. (2006). On the relation between metaphor and simile: When comparison fails. *Mind & Language*, *21*(3), 360–378.

Jones, L., & Estes, Z. (2006). Roosters, robins, and alarm clocks: Aptness and conventionality in metaphor comprehension. *Journal of Memory and Language*, *55*, 18–32.

Kövecses, Z. (2002). *Metaphor: A practical introduction*. Oxford University Press.

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. The University of Chicago Press.

Nakamura, T., Sakamoto, M., & Utsumi, A. (2010). The role of event knowledge on comprehending synesthetic metaphors. In *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society (CogSci2010)*.

Shapiro, K., & Caramazza, A. (2004). The organization of lexical knowledge in the brain: The grammatical dimension. In M. Gazzaniga (Ed.), *The cognitive neurosciences* (pp. 803–814). MIT Press.

Torreano, L., Cacciari, C., & Glucksberg, S. (2005). When dogs can fly: Level of abstraction as a cue to metaphorical use of verbs. *Metaphor and Symbol*, *20*(4), 259–274.

Utsumi, A. (2007). Interpretive diversity explains metaphor-simile distinction. *Metaphor and Symbol*, *22*(4), 291–312.

Utsumi, A., & Sakamoto, M. (2007a). Computational evidence for two-stage categorization as a process of adjective metaphor comprehension. In *Proceedings of the Second European Cognitive Science Conference (EuroCogSci2007)* (pp. 77–82).

Utsumi, A., & Sakamoto, M. (2007b). Predicative metaphors are understood as two-stage categorization: Computational evidence by latent semantic analysis. In *Proceedings of the 29th Annual Meeting of the Cognitive Science Society (CogSci2007)* (pp. 1587–1592).

Vigliocco, G., & Vinson, D. P. (2007). Semantic representation. In M. G. Gaskell (Ed.), *The Oxford handbook of psycholinguistics* (pp. 217–234). Oxford University Press.

Abstract and Belief-Based Language Differentiate Joking, Pretending, and Literal Toddler-Directed Speech

Elena Hoicka (elena.hoicka@stir.ac.uk)
Psychology Department, University of Stirling, Stirling, FK9 4LA, UK

Ruth Campbell (rec00001@students.stir.ac.uk)
Psychology Department, University of Stirling, Stirling, FK9 4LA, UK

Abstract

Twenty-two parents read a book containing joking, pretense, and literal pages to their 15- to 21-month-old toddlers. Parents differentiated joking from pretense book pages by using (1) more disbelief statements and humor-specific words, (2) fewer belief statements, and pretense-specific words. Parents differentiated joking from literal book pages by using more (1) high-level abstract language, (2) disbelief statements, and (3) humor-specific words. This study extends findings that abstract language cues non-literal concepts in general (e.g., metaphor, irony). This is also the first study to discover differences in cues to joking and pretense.

Keywords: Humor, Pretense, Abstract language, Beliefs, Parent-child interaction

Introduction

Human life is permeated with social institutions with conventional and normative structures. In order to participate in collective activities, children must learn how to act within these settings. One interesting question is how children respond to violations of normative rules. Sometimes, the appropriate response may be to protest (e.g., Rakoczy, Warneken, & Tomasello, 2008), but sometimes the appropriate response may be to treat the violation as a joke (and laugh), or as pretend (and maybe join in). This involves not only understanding that people have intentions, but also that they have intentions to do the wrong thing. This is an important, yet difficult concept required to understand humor, pretense, lying, false belief, and metaphor (Hoicka, Jutsum, & Gattis, 2008, Leekam, 1991).

While some accounts suggest that children possess an innate capacity to understand others' pretense and false beliefs (e.g. Leslie, 1987), such accounts do not explain how children might distinguish when someone is pretending versus joking, or even doing the right thing. For example, how do we use a telephone? We could speak into it when someone is on the other end (literal). We could speak into it when no one is listening (pretending). We could put the receiver on our foot and speak (joking). To an adult, the act in and of itself may distinguish whether a person intends to joke, pretend, or be literal. However for a toddler still learning about new objects, it may not be clear which act follows which intention. If they have had little experience with telephones, any act could be seen as the literal act. Even with experience of telephones, the pretend act could be seen as a joke (it's silly to talk to no one) and the humorous act could be seen as pretending (she's pretending that her foot is her ear). In order for toddlers to distinguish amongst various types of communicative intentions, it is thus plausible that parents give additional cues in order to help them in this task.

The goal of the current study was to determine whether parents differentiate joking, pretense, and literal speech with linguistic cues. Parents use more abstract language when reading humorous versus non-humorous book pages (Hoicka, et al., 2008). Similarly children use past and future tenses when pretending (e.g. Lodge, 1979; Musatti, 1993; Sawyer, 1997). Since references to past and future are forms of abstract language (Hoicka, et al., 2008), parents might also use abstract language to cue pretending. When irony and metaphor, two other concepts involving intentional wrongness, are couched in abstract language, adults are more likely to judge them as ironic or metaphorical respectively (Hoicka, 2010; Torreano, Cacciari, & Glucksberg, 2005). Theoretically, infants and toddlers could use abstract language in the same way to determine that joking or pretense was intended.

Belief-based language may serve to highlight differences between joking and pretense. When parents read a humorous versus non-humorous book, they used more disbelief statements, i.e., statements that conveyed that they did not believe what they had said (Hoicka, et al., 2008). For example, when making the joke, "Ducks say moo", parents made statements such as, "What are ducks supposed to say?" Thus parents cued their toddlers to humorous intentions by contrasting the jokes to the parents' true knowledge and beliefs. In contrast, utterances referring to pretend play have "at best weak correspondence in the immediate situation" (Veneziano, 2001, p. 331), for example, saying "here is a drink" whilst referring to an empty cup. Parents refer to absent references when pretending (Lillard & Witherington, 2004). For example, when pretending to eat, versus really eating, parents utter more words referring to the act of eating, or the objects involved in eating. Parents may use disbelief statements when joking because they (1) have said something to express disbelief about, and (2) by highlighting wrongness, parents may help toddlers understand the punch line of the joke. Parents may create absent references, a.k.a., belief statements, when pretending because (1) parents must convey what their wrong action was meant to be, e.g., putting a cup to one's mouth is not actually drinking, and so

making statements such as “I’m drinking” could help toddlers understand what the action represents, and (2) the purpose of pretense, unlike joking, is to represent a wrong action as something right in a possible world (Nichols & Stich, 2003), thus belief statements could emphasize the truth-values of the representational state. However the Hoicka et al. (2008) study did not measure belief statements, nor did the Lillard and Witherington (2004) study measure disbelief statements. This study aimed to determine whether parents use belief-based language to differentiate joking and pretense.

A second way in which parents might differentiate joking from pretense is by using humor- and pretense-specific words. Parents used humor-specific words such as “funny” or “silly” more often when reading a humorous versus non-humorous book to their toddlers (Hoicka, et al., 2008). Similarly, parents used the word “pretend” more often when pretending versus being literal (Lillard, et al., 2007). Such words could assist children in linking past and present experiences, and determine whether they are acts of joking or pretense. Indeed, preschoolers are more likely to understand pretend intentions when words such as “pretend” are used (Rakoczy, Tomasello, & Striano, 2006).

In the current study, parents read a book to their toddler which contained two joking, two pretense, and two literal pages. We designed the book in this way so that we could compare (1) cues to joking versus pretending, and (2) cues to literal versus joking speech.

Method

Participants

Twenty-two parents (20 mothers, 2 fathers) and their toddlers (age *M* = 18 months, 19 days, *SD* = 2 months, 16 days, *range* = 15 months, 5 days to 21 months 26 days; 9 boys) participated. One additional participant was not included due to fussiness. Participants were recruited from playgroups, toddler classes, and a press release in the local news paper. Parents and children were primarily Scottish.

Materials

Four illustrated versions of a book, “James’ Big Day” were created. See Figure 1 for an example of pages. A Shure head-mounted microphone was fit into an Olympus MP3 recorder to record the parents’ speech. A Sony digital camcorder was used to record the visual aspects of the reading session, and as a backup for speech recordings.

Design

This was a within-subjects design. The independent variable was the type of utterance each page conveyed: joking, pretense, or literal. There were a total of six target sentences per book; two conveyed joking, two conveyed pretense, and two conveyed something literal. The books were designed such that the same target sentence conveyed either joking, pretense, or was literal, depending on the sentences prior to the target sentence, as well as the accompanying images in the books. See Figure 1 for an example. Four different books matched different page types to eight target sentences, and this was counterbalanced. Parents read only one book each. The dependent variables included parents’ use of abstract language, belief-based language, and humor- and pretense-specific language.

Coding

Parents’ utterances were transcribed from the MP3 files. For parents’ use of abstract language, each extra-textual utterance (ETU) was coded for levels of abstraction following Hoicka, et al. (2008), and Van Kleeck, et al. (1997). These included:

Level 1 (perceptual identification, concrete): The utterance refers solely to one object in the event. This level includes object labelling either at the basic, subordinate, or superordinate levels. It also includes stating an intrinsic property of the object (e.g., color) or drawing attention to the object or one of its properties. Examples are “What’s that?" and “It’s a bowl.”

Level 2 (perceptual relationship, concrete): The utterance links two objects or events. The link may involve an intrinsic property (same color), spatial relation (left of, above), a common action (X and Y produce something, or X acts on Y), or a common feeling. Examples are “This car is like the other car.” and "The cake is in his hand."

Level 3 (displaced reference, abstract): The utterance links an object or event with an object or event that is absent either in space (spatially displaced reference) or time (past talk), typically including subjective experiences with the object. Examples are “Do you remember seeing a duck in the pond?” and “You have a car at home.”

Level 4 (inference, abstract): The utterance conveys one of several inferences, including logical reasoning and imaginary description, or states some social knowledge. Examples are "If he eats that with his hands, he'll make a mess", and "It’s like the boy is flying through the air".

Transcripts were separately coded for belief-based language, following Hoicka, et al. (2008) for disbelief statements. All ETUs which followed the target sentences were coded as either a belief statement, a disbelief statement, or neither.

To be coded as a belief statement, the ETU should suggest that the parent believed the assertion of the target sentence. This can be coded in three ways:

General belief statements: statements which express belief that can be applied to any statement, e.g., “That’s right”, or “It’s true”

Sentence-specific belief statements: statements which express belief specifically in relation to the target sentence. This could include a repetition or re-phrasing of the target sentence.

Build-on belief statements: statements which show belief through building on the target sentence. E.g., if the context is that a child is pretending that a basket is a pram, or the child is really sitting in a pram, and the target sentence is

"He's sitting in the pram", the parent might add to this by saying something like "There's the wheel"

To be coded as a disbelief statement, the ETU should suggest that the parent does not believe the assertion of the target sentence. This can be coded in three ways:

General disbelief statements: statements or questions which express disbelief that can be applied to any statement, e.g., "That's wrong", or "That's not true".

Sentence-specific disbelief statements: statements and questions which express disbelief specifically in relation to the target sentence, e.g., for the target utterance, "He's sitting in the pram", the parent might say, "That's not a pram", or "Is that a pram?"

Build-on disbelief statements: statements which show disbelief through building on the target sentence. E.g., for the target sentence, "He's sitting in the pram", parents might say, "Prams should have wheels." or, "What is he really sitting in?"

ETUs were coded for humor-specific words such as jok*, funn*, hilarious, and sill*, and for pretense-specific words such as preten*, imagin*, and make-believe.

Results

No effects of child age were found, so child age was dropped from final analyses. Linear mixed models were used with participant code and target sentence as random variables. Simple contrasts were used to compare Joking to both Pretense and Literal pages.

Abstract Language

Means for Page Type by Abstraction Level can be found in Figure 2. No effects of child gender were found, so child gender was not included in the final analyses. A 3 (Page Type: Pretense, Joking, Literal) X 4 (Level of Abstraction 1-4) mixed model found an effect of Level of Abstraction, $t(503) = 2.87$, $p = .0043$, and an interaction between Level of Abstraction and Page Type, $t(503) = 2.41$, $p = .0165$. Additional models were run to examine interactions.

3 (Page Type) mixed models on Levels 1, 2, and 3 Abstraction using simple contrasts found no effects (Level 1: Pretense $M = 0.70$, $SD = 1.19$; Joking $M = 0.75$, $SD = 1.45$; Literal $M = 0.93$, $SD = 1.53$; Level 2: Pretense $M = 0.34$, $SD = 0.71$; Joking $M = 0.55$, $SD = 1.09$; Literal $M = 0.23$, $SD = 0.60$; Level 3: Pretense $M = 0.18$, $SD = 0.50$; Joking $M = 0.32$, $SD = 0.83$; Literal $M = 0.23$, $SD = 0.60$). A 3 (Page Type) mixed model on Level 4 Abstraction using a simple contrast found that parents uttered significantly more Level 4 ETUs when reading Joking ($M = 1.34$, $SD = 1.58$) versus Literal pages ($M = 0.61$, $SD = 1.22$), $t(108) = 2.64$, $p = .0094$. There was no difference between Joking and Pretense ($M = 1.30$, $SD = 1.72$) pages.

Belief-Based Language

Means for Page Type by Belief-based Language can be found in Figure 3. A 3 (Page Type: Pretense, Joking, Literal) X 2 (Statement Type: Belief, Disbelief) X 2(Child Gender) mixed model found effects of Page Type (Pretend vs. Joking), $t(232) = 5.07$, $p < .0001$; Page Type (Joking vs. Literal), $t(232) = 2.17$, $p = .0309$; an interaction between Statement Type and Page Type (Pretend vs. Joking), $t(232) = 5.04$, $p < .0001$; an interaction between Statement Type and Page Type (Joking vs. Literal), $t(232) = 3.27$, $p = .0012$; and an interaction between Statement Type, Page Type (Joking vs. Literal), and Child Gender, $t(232) = 2.35$, $p = .0197$. Additional models were run to examine interactions.

No effects of child gender were found for disbelief statements, so were dropped from the following analysis. A 3 (Page Type: Pretense, Joking, Literal) mixed model for disbelief statements with a simple contrast found that parents used significantly more disbelief statements when expressing Joking ($M = 1.45$, $SD = 1.70$) versus Pretense ($M = 0.39$, $SD = 0.92$), $t(108) = 2.42$, $p = .0173$, and when expressing Joking versus Literal ($M = 0.18$, $SD = 0.45$) speech, $t(108) = 4.03$, $p < .0001$.

A 3 (Page Type: Pretense, Joking, Literal) X 2 (Child Gender) mixed model for belief statements with a simple contrast found that parents used significantly more belief statements when expressing Pretense ($M = 0.66$, $SD = 1.35$) versus Joking ($M = 0.36$, $SD = 0.75$), $t(108) = 2.70$, $p = .0080$. An interaction between Child Gender and Page Type (Pretense, Joking), was found, $t(108) = 2.90$, $p = .0045$, such that parents used more belief statements when expressing Pretense to boys versus girls, but more belief statements when expressing Joking to girls versus boys. There was no difference between Joking and Literal ($M = 0.68$, $SD = 0.88$) speech.

Humor- and Pretense-Specific Words

Means for Page Type by use of humor- and pretense-specific words can be found in Figure 4. No effects of child gender were found, so child gender was not included in the final analyses. A 3 (Page Type: Pretense, Joking, Literal) X 2 (Word Type: Humor, Pretense) mixed model found effects of Page Type (Pretense vs. Joking), $t(237) = 2.40$, $p = .0173$; Page Type (Joking vs. Literal), $t(237) = 2.40$, $p = .0173$; and an interaction between Word Type and Page Type (Pretense vs. Joking), $t(237) = 2.87$, $p = .0045$. Additional models were run to examine interactions.

A 3 (Page Type: Joking, Pretense, Literal) mixed model for humor-specific words using a simple contrast found that parents used significantly more humor-specific words when expressing Joking ($M = 0.32$, $SD = 0.64$) versus Pretense ($M = 0.02$, $SD = 0.15$), $t(108) = 2.12$, $p = .0366$, and when expressing Joking versus Literal speech ($M = 0.02$, $SD = 0.15$), $t(108) = 2.12$, $p = .0366$.

A 3 (Page Type: Pretense, Joking, Literal) mixed model for pretense-specific words using a simple contrast found that parents used significantly more pretense-specific words in the Pretense ($M = 0.11$, $SD = 0.44$) versus Joking ($M = 0.02$, $SD = 0.15$) conditions, $t(108) = 2.05$, $p = .0427$. There was no difference between the Joking and Literal ($M = 0$, $SD = 0$) conditions.

Discussion

The current study investigated whether parents use linguistic cues to differentiate (1) joking from pretense, and (2) joking from literal speech

Joking vs. Pretense

This research provides the first evidence that parents use belief-based language to differentially cue toddlers to joking and pretense. Parents used significantly more disbelief statements and significantly fewer belief statements when reading joking versus pretense pages. Parents also used significantly more humor-specific words when reading joking versus pretense pages, and significantly more pretense-specific words when reading pretense versus joking pages.

While intentionality research typically focuses on whether or not children understand intentions (e.g., Carpenter, Akhtar, & Tomasello, 1998; Gergely, Bekkering, & Király, 2002; Meltzoff, 1995) little research has examined how children come to understand intentions to do the wrong thing, and how children come to distinguish amongst various types of intentions. The current study demonstrates that parents offer toddlers linguistic cues to distinguish between joking and pretense. Thus it may be the case that children learn to distinguish amongst abstract, non-literal concepts such as jokes and pretense. In particular, hearing proportionally more disbelief statements and proportionally fewer belief statements when encountering joking could allow a child to identify the reference to the wrong act, and to identify that the act was meant only as a wrong act and nothing else. In contrast, by hearing a more even mixture of belief and disbelief statements when encountering pretense, children could identify the reference to the wrong act through disbelief statements, and could also identify the representation of the wrong act through belief statements. Additionally, humor-specific words could help toddlers link past and present humorous situations, while pretense-specific words could help toddler link past and present pretense situations.

Literal vs. Non-literal Speech

The current study found that parents used more high-level abstract language when reading joking versus literal pages. This replicates findings that parents use high-level abstract language to cue humor (Hoicka, et al., 2008). This also converges with findings that abstract language cues adults to both metaphor and irony (Hoicka, 2010; Torreano, et al., 2005), two other non-literal abstract concepts. Interestingly, there was no difference between the amount of high-level abstract language that parents used to cue joking versus pretense. Thus parents may use abstract language in order to cue their children to both joking and pretense. This may allow toddlers to think in an abstract way in order to resolve the joke, or understand the representation underlying the pretense.

The current study also found that parents used more disbelief statements when expressing joking versus literal concepts. Additionally, parents used significantly more humor-specific words such as “funny” and “silly” when reading joking versus literal pages. This replicates and extends past research which found that parents use disbelief statements, as well as humor-specific words to cue humor (Hoicka, et al., 2008).

The present findings suggest that parents may bootstrap non-literal concepts, such as joking and pretending, by helping their toddlers to think in an abstract way, and by identifying that what was said is not literally true. These linguistic cues could also help toddlers identify what exactly made the situation false. Additionally, given that toddlers do not understand that others can intend to joke until they are 2 years old (Hoicka & Gattis, 2008), and do not understand that others can intend to pretend until they are 3 years old (Rakoczy, Tomasello, & Striano, 2004), abstract and belief-based language may help toddlers realize that others can intend to do or say the wrong thing in the first place.

Future Research

Future Research will examine whether mothers and fathers give the same or different types of linguistic cues to differentiate joking and pretense, and to differentiate literal from non-literal speech. Additionally, we will examine whether parents use the same types of linguistic cues when engaging in acts of joking and pretense with their toddlers, as compared to when they read about these concepts. Finally, future research should also consider whether toddlers can use these cues to adequately differentiate joking from pretense, and literal versus non-literal events.

Figures



Pretense:
James and his big sister Katie are playing make-believe. Katie crawls around and meows. Meow meow. Mummy asks, “What was that?” James knows…
It was a cat.

Joking:
James and his big sister Katie go in the yard. They hear some barking. Ruff ruff. Mummy says, “What was that?” James wants to make a funny joke…
It was a cat.

Literal:
James and his big sister Katie are really excited because Mummy brought home a new pet. They hear something go meow meow. What was the pet?
It was a cat.

Figure 1: Examples of Page Types for same target sentence. Original images were in color.

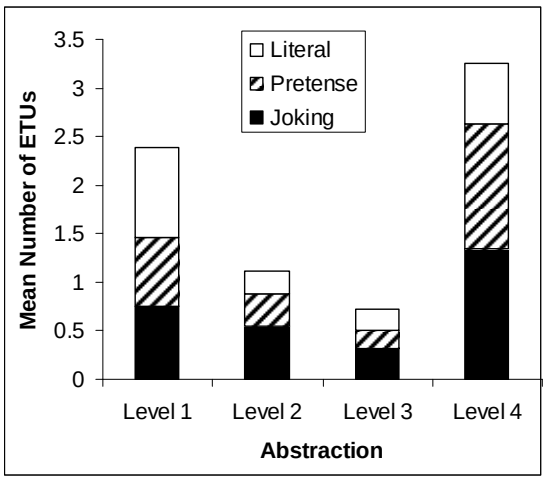


Figure 2: Mean number of ETUs for each Level of Abstraction

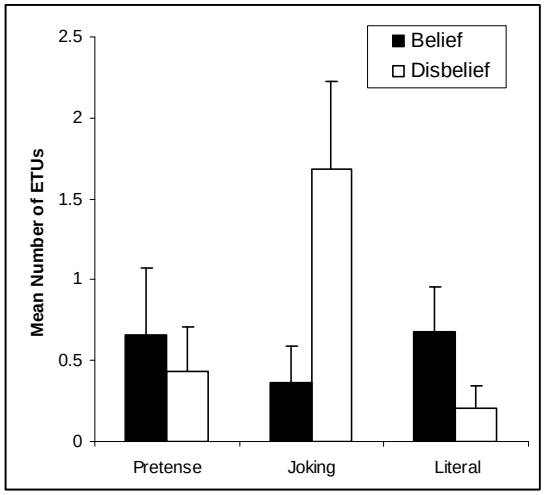


Figure 3: Mean number of ETUs expressing Belief and Disbelief.

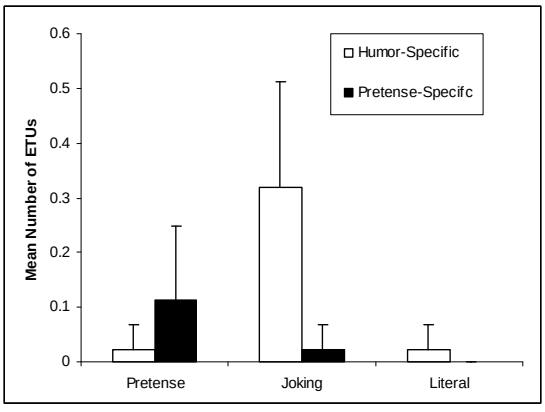


Figure 4: Mean numbers of ETUs using humor- and pretense-specific words.

Acknowledgments

We would like to thank parents and toddlers for participating in our research. We would also like to thank Felicity Malla, Jennifer Dent, and Sarah McAllister for help with creating stimuli and data collection. This research was supported by a British Academy Small Research Grant SG-54221 and an ESRC Small Research Grant RES-000-22-3888, both awarded to Elena Hoicka.

References

Carpenter, M., Akhtar, N., & Tomasello, M. (1998). Fourteen through 18 month old infants differentially imitate intentional and accidental actions. *Infant Behavior and Development, 21*. 315-330.

Gergely, G., Bekkering, H., & Király, I. (2002). Rational imitation in preverbal infants. *Nature, 415*, 755.

Hoicka, E. (2010). Abstract Language as a Cue to Irony. *Submitted Manuscript.*

Hoicka, E. & Gattis, M. (2008). Do the wrong thing: how toddlers tell a joke from a mistake. *Cognitive Development, 23*, 180-190.

Hoicka, E., Jutsum, S., & Gattis, M. (2008). Humor, abstraction, and disbelief. *Cognitive Science, 32*, 985-1002.

Leekam, S. (1991). Jokes and lies: Children's understanding of intentional falsehood. In A. Whiten (Ed.) *Natural Theories of Mind*. Oxford: Basil Blackwell.

Lillard, A. S., Nishida, T., Massaro, D., Vaish, A., Ma, L., & McRoberts, G. (2007). Signs of pretense across age and scenario. *Infancy, 11*, 1-30.

Lillard, A. S., & Witherington, D. (2004). Mothers' behavior modifications during pretense snacks and their possible signal value for toddlers. *Developmental Psychology, 40,* 95-113.

Lodge, K. R. (1979). The use of the past tense in games of pretend. *Journal of Child Language, 6*, 365-369.

Meltzoff, A. (1995). Understanding the intentions of others: Re-enactment of intended acts by 18-month-old children. *Developmental Psychology, 31*, 838-850.

Musatti , T., & Orsolini, M. (1993). Uses of past forms in the social pretend play of Italian children. *Journal of Child Language, 20*, 619-639.

Nichols, S., & Stich, S. P. (2003). *Mindreading: An integrated account of pretense, self-awareness, and understanding other minds.* Oxford: Oxford University Press.

Rakoczy, H., Tomasello, M., & Striano, T. (2004). Young children know that trying is not pretending: A test of the "behaving-as-if" construal of children's early concept of pretense. *Developmental Psychology, 40*, 388–399.

Rakoczy, H., Tomasello, M., & Striano, T. (2006). The role of experience and discourse in children's developing understanding of pretend play actions. *British Journal of Developmental Psychology, 24*, 305-335

Rakoczy, H., Warneken, F., & Tomasello, M. (2008). The sources of normativity: Young children's awareness of

the normative structure of games. *Developmental Psychology, 44*, 875-881.

Sawyer, R. K. (1997). *Pretend play as improvisation: Conversation in the preschool classroom.* London: Routledge.

Torreano, L. A., Cacciari, C., & Glucksberg, S. (2005). When dogs can fly: Level of abstraction as a cue to metaphorical use of verbs. *Metaphor and Symbol, 20*, 259–274.

Van Kleeck, A., Gillam, R. B., Hamilton, L., & McGrath, C. (1997). The relationship between middle-class parents' book-sharing discussion and their preschoolers' abstract language development. *Journal of Speech and Hearing Research, 40*, 1261–1271.

Veneziano, E. (2001). Displacement and informativeness in child-directed talk. *First Language, 21,* 323-356.

Wrongness and Representational Thought

Elena Hoicka (elena.hoicka@stir.ac.uk)
Psychology Department, University of Stirling, Stirling, FK9 4LA, UK

Merideth Gattis (gattism@cardiff.ac.uk)
School of Psychology, Cardiff University, Tower Building, Park Place, Cardiff, CF10 3AT, UK

Abstract

This paper examines the concept of wrongness as a violation of intention, convention, or fact. We demonstrate that wrongness is an underlying factor in mistakes, jokes, pretense, lying, metaphor, and irony. We argue that children's use and understanding of wrongness evolves in four steps through a developing understanding of representation. First, children understand that a wrong act can *refer* to a right act, through mistakes and basic jokes. Second, this leads to understanding that a wrong act can *represent* a right act, through pretense, puns and metaphor. Third, this leads to understanding mental representation, which in combination with understanding reference allows understanding of intentional jokes and lies. Finally, this leads to understanding mental representation in combination with representation, allowing an understanding of irony, and intentional pretense, metaphor, and puns.

Keywords: Wrongness; Representation; Mistakes; Jokes; Pretense; Lying; Metaphor; Irony

Wrongness and Representational Thought

Parents, educators, and even psychologists generally assume that an important goal of development is learning to do the right thing. In this paper, we consider the value of learning to do the wrong thing. We propose that learning about wrongness proceeds through four stages, each of which plays a critical role in the development of representational thought.

What is wrongness?

Most analyses of wrongness focus on the moral aspects of doing the wrong thing. For example, philosophers have argued that wrongness is something prohibited by morality (e.g. Calder, 2005), such as murder or cheating (e.g. Feezell, 1988; Marquis, 2001). The moral concept of wrongness is also examined in research on lying (e.g., Sorenson, 2007). Similarly, psychologists interested in wrongness have focused on deontic reasoning, that is, the speaker's attitude towards what she is saying, and in particular, how necessary a speaker deems some condition or act. This includes permission (*what one **may** do*), and obligation (*what one **must** do*), thus wrongness might violate what one is obliged or expected to do (Cummins, 1996; Tomasello, 2003).

Morality is not, however, the only basis for wrongness. Wrongness can be evaluated as a violation of fact, irrespective of moral issues. For example, if you eat the last cookie, and say that you did not, your statement, "I did not eat the last cookie" is wrong simply because it does not reflect truth-values in the world (Carson, 2006). Similarly, metaphors, pretending, and joking all involve wrongness because they do not represent the true state of affairs (Amsel, et al., 1996; Kazmerski, Blasko, & Desalegn, 2003; Leekam, 1991).

Wrongness extends beyond truth values, and can also describe violations of convention. Conventions do not have absolute truth values. Nonetheless an action which breaks convention, such as moving 6 places on a board game after having rolled a 5, is also wrong. Conventions can apply to how we speak, use objects, eat, dress, play games, interact with others, and hence permeate many aspects of our daily lives. Searle (2005) posits that there are two types of conventions in regards to objects. One type includes causal usage functions, in which we have the convention of using an object in a certain way, supported by the physical features of the object (e.g., knives are sharp, and so are used to cut things). Status functions are more conventional and attach arbitrary functions to objects, for example, in the case of paper used as money. Thus while you could technically try to use a knife as a paper weight, or tissues to pay for your purchase, this would be wrong according to convention. Children as young as 2 years demonstrate sensitivity to the conventions associated with objects, displaying what is called functional fixedness, where they refuse to use objects in unconventional ways, after only one exposure to how an object should be used (Casler & Kelemen, 2005). Language itself is also a set of conventions where certain words happen to be paired with certain actions, objects, and so on (Searle, 1969). Futhermore, different languages have different conventions, and within a language one must adhere to the specific labels given to specific objects and actions. Infants and toddlers respect the conventionality of language, demonstrating hesitance to assign the same label to multiple objects (e.g, Markman & Wachtel, 1988). Thus using the wrong words can be wrong by violating convention.

Wrongness can also describe violations of intentions. For example, it is not more right to request chocolate versus vanilla ice cream. However if you intend to eat chocolate, and instead ask for vanilla, such an utterance would be wrong in terms of the current goal. Thus mistakes embody a form of wrongness that violates one's intentions.

Wrongness as Violation

We define wrongness as a violation of intention, convention, or fact, independent of the moral standing of the act. Several concepts involve understanding wrongness. In

the next section we review relevant empirical findings on mistakes, jokes, pretending, lies, irony, and metaphor, which all involve wrongness (e.g., Carpenter, Akhtar, & Tomasello, 1998; Hoicka, Jutsum, & Gattis, 2008; Kazmerski, et al., 2003; Leekam, 1991).

Mistakes

Mistakes by definition involve doing the wrong thing. This type of wrongness necessarily involves a violation of intention: you meant to perform one act, but performed another instead. This could be for one of two reasons: you could do something in an accidental fashion, such as fall over, which might be considered a true mistake. You could also truly believe that what you are doing is the right thing, even though it is not, and perform what might be better called an error, where you violate your intention because of lack of knowledge (e.g., Lee & Cameron, 2000). As an example of a mistake (or error), you may wish to turn on the television, but press the wrong button (either through an accidental physical movement or a false belief that it is the right button), such that it does not light up, and the goal of turning on the television is not achieved. Additionally, mistakes could involve a violation of convention (accidentally driving on the wrong side of the road, either because you falsely believed that that is the convention in that country, or perhaps because the road is poorly lit), as well as a violation of fact (e.g., Saying that Tony Blair is the Prime Minister of the United Kingdom either because you did not realize that Gordon Brown had taken his place, or because the wrong name came out). However mistakes need not require a violation of convention or fact, for example, when accidentally requesting the wrong object the request cannot be wrong, but it still violates an intention.

Mistakes may be the earliest understood form of wrongness. Meltzoff (1995) found that when adults performed incomplete actions with objects (i.e., failed attempts, which could be viewed as a mistake) 18-month-olds completed (or corrected) those actions. From 14 months, infants avoid actions accompanied by the expression "Whoops!" (Carpenter, et al., 1998) and by rising intonation (Sakkalou & Gattis, in press). Finally, infants as young as 9 months (but not 6 months) react differently to someone unwilling to give an object, versus unable, due to an accident or failed attempt, a.k.a., a mistake, through looking and reaching less (Behne, et al., 2005).

Jokes

Basic jokes, which involve saying or doing something wrong, violate convention or fact (Hoicka & Gattis, 2008; Hoicka, et al., 2008). For example, one could joke that ducks say, "moo" (violating fact) or one could point to a duck and call it a "moogy" (violating English language conventions). Jokes by definition cannot violate intention since joking involves intentionally doing or saying the wrong thing.

Basic jokes are another form of wrongness that is understood early in development. Three- to 5-year-olds primarily laugh at events that others, or they themselves, intend to be humorous, such as clowning or being silly (Bainum, Lounsbury, & Pollio, 1984). This suggests that they appreciate that others do the wrong thing in order to joke. As early as 30 months, children copy mislabelling behaviors when couched in a humorous context, but not in a non-humorous context (Hoicka & Akhtar, 2010). From 25 months children copy incorrect actions followed by laughter, but correct the same incorrect actions followed by the expression, "Whoops!" indicating that they interpret others' wrong actions as humorous (Hoicka & Gattis, 2008). Finally, 15-month-olds match humorous cues to humorous actions, and from around 10 months, infants laugh when their mothers perform incongruous actions, such as putting socks in their mouths (Hoicka & Wang, 2010; Sroufe & Wunsch, 1972). Finally, observational evidence suggests that infants may not only appreciate others' jokes, but may create jokes as well. From 15 months, infants have been observed to create jokes such as putting sponges in their mouths, and from 8 months, repeat incongruous actions, such as screwing up their faces, in order to re-elicit laughter (Loizou, 2005; Reddy, 2001).

More complex joking, such as puns, involves saying something that initially appears to violate fact or convention, but upon further reflection is consistent with fact or convention (e.g., Shultz, 1974). By initially appearing to be unrelated, puns appear to be wrong answers to questions as they violate conventions of communication, specifically Grice's Maxim of relation, and by being ambiguous (having two meanings), puns also violate Grice's maxim of Manner (e.g., Grice, 1975). These types of jokes are not normally understood until later. From 8 years, children choose joke endings with double meanings as more humorous than non-sequitor joke endings. However 6-year-olds judge both joke endings to be equally humorous, demonstrating that they only find a violation of the Maxim of relation humorous, or put another way, saying something wrong in the context of the previous utterance (Shultz, 1974). However using cartoons instead of words, even 4-year-olds appreciate jokes involving double meanings (Pien & Rothbart, 1976).

Pretense

Pretense involves understanding that someone has done the wrong thing, but has represented this action as right in a possible world (Nichols & Stich, 2003). In particular, pretense violates conventions and facts. For example, one might pretend that a block is a bar of soap, and violate convention by rubbing the block on one's body. One might also make statements which violate fact, for example, if a child says, "I can fly", which is not technically true. Like joking, pretense cannot violate intention, since pretense involves intentionally doing the wrong thing by its very nature (Hoicka & Gattis, 2008; Hoicka, et al., 2008).

Lillard (1998) found that 4- and 5-year-olds did not understand intentions to pretend. However the task involved hearing stories illustrated by pictures or dolls, and verbal responses were required. In experiments using an action-based task, 36-month-olds, but not 26-month-olds, differentiated intentions to pretend from trying (Rakoczy, Tomasello & Striano, 2004). Additionally, children can tell whether someone else is pretending or doing the real thing from 2.5 years (Ma & Lillard, 2007). Finally, using a looking-time paradigm, infants as young as 15 months detected violations in a pretense scheme (Onishi, Baillargeon, & Leslie, 2007). Children themselves pretend from around 18 months (e.g., Elder & Pederson, 1978; Ungerer, et al., 1981).

Lies

Lying involves understanding that someone has said the wrong thing for the purpose of deceiving someone (e.g., Leekam, 1991). Lying can be a violation of fact (e.g., saying one has not eaten cake when one has) and could also be a violation of convention (e.g., telling someone that they should drive on the left side of the road whilst in Spain). Lee and Cameron (2000) argue that a lie need not actually violate a fact as long as the liar thinks that the lie violates fact. Thus one could argue that lying either involves violating fact or convention, or having false (or wrong) beliefs about facts and conventions.

In order to truly understand that someone is lying, it is necessary to understand their intention to lie. While joking, pretending, and metaphor can be detected without understanding the intention behind an action or utterance, for example, by finding the joke funny, or noticing a similarity between a pretend act or metaphor and its representation, this is not the case for lying. If one simply notices that someone has said the wrong thing, this could be due to their lying, or it could be due to them having made a mistake. What is crucial is thus whether the liar intended to deceive.

Depending on how studies are performed, children start to understand lies between 3 and 5 years. Lee, et al., (2002) conducted an experiment in which young children were told lies that violated a reality-fantasy distinction. Five- and 6-year-olds identified the lies, and hence did not believe them, while 3- and 4-year-olds accepted the lies. Wimmer, Gruber, and Perner (1985) used a story-based method to assess what young children understood about lying. When asked whether the character should be punished, children as young as 4.5 years assigned punishment to liars, but not to people who were mistaken. However when asked whether the person had lied, children did not reliably distinguish the liar from the mistaken person. Thus 4.5-year-olds have a moralistic understanding of lies, without necessarily understanding the lexical term relating to lies. Using a picture-based method, 4-year-olds differentiated lies from promises (Maas, 2008). Finally, children as young as 3 years distinguished lies from mistakes, (Siegal & Peterson, 1996, 1998). Considering when children begin to lie, from around 3 or 4 years children lie in order to hide a transgression of peeking when they were not supposed to, and tell white lies when receiving an unwanted gift (Talwar, et al., 2002; Talwar, Murphy, & Lee, 2007).

Metaphor

Metaphor involves intentionally saying the wrong thing, (e.g., Harris, Friel, & Mickelson, 2006) for purposes such as to provoke thought, compare similarities, and add interest, describe, and clarify (e.g., Gardner & Winner, 1986; Roberts & Kreuz, 1994; Sperber, 1984). Metaphor can violate fact, for example, saying, "Your room is a pig sty" when in fact it's just a room (e.g., Andrews, et al., 1986).

It is not until school age that children understand that people can intend to create metaphors. Eight-year-olds, but not 6-year-olds, differentiate metaphors from mistakes (Andrews et al., 1986). When one does not consider intentions, younger children appear to understand metaphors. In one task, 3-, 4-, and 5-year-olds were told stories that used time-based metaphors, and were then asked comprehension questions based on the metaphors. From 4 years, children correctly answered questions relating to metaphors (Ozcaliskan, 2005). In another task, 3- and 4-year-olds produced significantly more errors when repeating anomalous versus metaphorical utterances, and made the same number of errors when producing metaphorical and literal utterances, suggesting that the children understood the metaphors (Pearson, 1990). Finally, in terms of metaphor production, from 3 years children can produce appropriate metaphorical compounds. For example, if a stick-shaped bug is called a "leaf-bug" children might make a more appropriate metaphor by calling it a "stick-bug" (Gottfried, 1997).

Irony

Like metaphor, irony involves intentionally saying the wrong thing. Irony can violate fact, for example, saying, "That bungalow is the tallest building in the world" (e.g., Andrews, et al., 1986). Irony can also violate convention, for example, saying, "Driving on the right side of the road in London was a great idea." Again, irony cannot involve a violation of intention as irony is intentional in nature. Indeed, like lying, intention is the most important part of irony. An utterance is only ironic if the person meant it to be (e.g., Andrews, et al., 1986; Winner & Leekam, 1991).

Irony, like metaphor, is notoriously difficult for children to understand. It is not until school age that children understand intentions to be ironic. Eight-year-olds, but not 6-year-olds, differentiated irony from lies (Andrews et al., 1986). Winner and Leekam (1991) tested 7-year-olds on their ability to distinguish irony from deception. They found that children's ability to do so was contingent on their ability to distinguish second order intentions, that is, that the liar intended for the audience to believe the falsehood, whereas the ironist did not.

Wrongness, Reference, and Representation

Understanding the various types of wrongness involves understanding representation at different levels, and this understanding develops in stages (see Figure 1). The first stage involves understanding that a wrong act refers to a right act. While representation makes one think *of a wrong act as* a right act, reference only makes one think *of* a right act. Reference should be easier to understand since it can be accomplished by considering two different acts sequentially rather than simultaneously. Without being able to make a reference between a wrong act and a right act, it would be difficult to determine that an act was wrong in the first place: reference allows comparison between two acts. The ability to compare two acts appears to be present by 9 months, as children are first able to detect mistakes at 9 months (Behne, et al., 2005), and jokes at 10 months (Sroufe & Wunsch, 1972).

The second stage in understanding wrongness involves understanding representation, such that a wrong act can represent a right act (e.g., Nichols & Stich, 2003). Thus one act can have two meanings at the same time: the literal, and the imagined. This should be more difficult to understand than reference as representation requires simultaneous, rather than sequential processing. Understanding representation is essential for understanding pretense, metaphor, and pun-type jokes (e.g., Leslie, 1987; Shultz, 1974). Pretense is likely the first instance of understanding that a wrong act represents a right act. As metaphors and verbal puns require advanced linguistic skills as compared to pretense, understanding of metaphor and verbal puns should be delayed compared to pretense, but should involve the same underlying representational skills. Since children generally understand pretense from around 18 months (e.g., Ungerer, et al., 1981) this should mark when children understand the representation of wrong acts as right acts. Once children understand that a wrong act can represent a right act, then they have the possibility to distinguish mistakes and jokes from pretense. Thus a wrong act for which children cannot determine how it represents a right act could be thought of as a mistake or joke, and a wrong act for which children could determine how it represents a right act could be thought of as a general "as-if" for pretending (perhaps if action based). Later, when children's language abilities develop, they should also be able to distinguish metaphors from puns as well. At this point, lies and irony, if the verbal content were to be understood, might still be thought of as mistakes or jokes, since they (at least appear) to refer to right acts, instead of representing them.

The third stage involves a basic understanding of mental representation. This involves processing a mental representation and its reference sequentially: understanding that an *intended* wrong act refers to a right act. The earliest instance of this may be when children understand that others can intentionally do the wrong thing through joking from 25 months (Hoicka & Gattis, 2008), since both the actor's intention, and the reference between right and wrong acts are detected.

For the fourth stage, children must understand mental representation in terms of representations themselves. This requires understanding a representation in relation to mental representation, or in other words, understanding that an *intended* wrong act represents a right act. This may first be understood when children understand that others' can intend to pretend, at 36 months (Rakoczy, et al., 2004), when both the actor's intention, and the corresponding representation are detected.

Irony is more complex still. Like metaphor, it involves saying something wrong which represents something right. However, in metaphor, the similarity between concepts can lead a child to infer that a metaphor was made, without reference to the speaker's mental state. In contrast, like lying, irony is about the attitude of the speaker, and cannot be inferred without understanding intention and belief (e.g., Andrews, et al., 1986; Winner & Leekam, 1991). Thus irony involves understanding two mental representations simultaneously: the wrong act (what the ironist intended to say) and the right act (what the ironist believed). At this point, when children can simultaneously process two mental representations: the intention to perform a wrong act, and the belief of a right act, children should be able to distinguish all types of wrong acts from each other.

We propose that these four stages of representation are linked. First, understanding the reference point between right and wrong, through mistakes and basic jokes, could help children later understand the representation of a wrong act as a right act, through pretense. By processing two acts sequentially when detecting jokes or mistakes, children may get used to considering two ideas in relation to each other. This may bootstrap an understanding of representation as it involves a shift from considering two ideas sequentially to considering two ideas simultaneously. This should be easier than making a bigger shift of never processing two ideas in relation to each other, to processing two ideas simultaneously.

Second, we propose that understanding that a wrong act can represent a right act (e.g., through pretense) is a precursor to mental representation (following, Leslie, 1987), since understanding mental representations, such as intentions, involves understanding something that is inferred, and not concretely perceivable. At this point, children should already understand reference, and should thus bootstrap their understanding that a wrong act refers to a right act, to understanding that an intended wrong act refers to a right act. This would be a simpler cognitive leap versus having no understanding or reference, and then suddenly having an understanding of reference in terms of mental representations. Finally, once children can process an intentional wrong act, and a (belief-based) right act, sequentially, this should create a smoother transition for processing an intentional wrong act, and a (belief-based) right act, simultaneously.

Figure

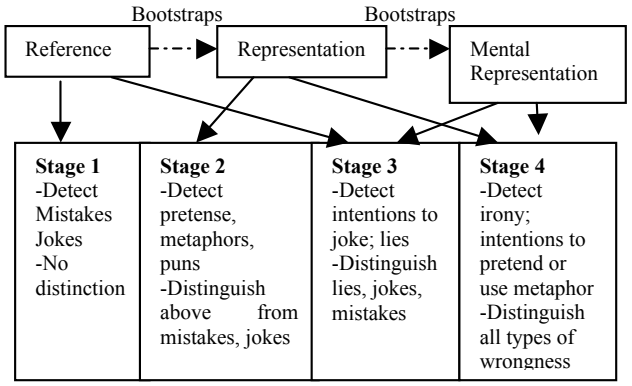


Figure 1. Stages of wrongness understanding.

Acknowledgments

The research was supported by a NIH Postdoctoral Fellowship (T32-HD046423) to Elena Hoicka.

References

Amsel, E., Bobadilla, W., Coch, D., & Remy, R. (1996), Young children's memory for the true and pretend identities of objects. *Developmental Psychology, 32,* 479-491.

Andrews, J., Rosenblatt, E., Malkus, U., Gardner, H., & Winner, E. (1986). Children's abilities to distinguish metaphoric and ironic utterances from mistakes and lies. *Communication & Cognition, 19*, 281-298.

Bainum, C. K., Lounsbury, K.R., & Pollio, H. R. (1984) The development of laughing and smiling in nursery school children. *Child Development, 55,* 1946-1957.

Behne, T., Carpenter, M. Call, J. & Tomasello, M. (2005). *Developmental Psychology, 41,* 328-337.

Calder, T. (2005). Kant and degrees of wrongness. *Journal of Value Inquiry, 39,* 229-244.

Carpenter, M., Akhtar, N., & Tomasello, M. (1998).Fourteen- through 18-month-old infants differentially imitate intentional and accidental actions. *Infant Behavior & Development, 21,* 315-330.

Carson, T. L. (2006). The definition of lying. *Nous, 40,* 284-306.

Cummins, D. D. (1996). Evidence of deontic reasoning in 3- and 4-year-old children. Memory & Cognition, 24, 823-829.

Elder, J. L., & Pederson, D. R. (1978). Preschool children's use of objects in symbolic play. *Child Development, 49,* 500-504.

Feezell, R. M. (1988). On the wrongness of cheating and why cheaters can't play the game. *Journal of the Philosophy of Sport, 15,* 57-68.

Gottfried, G. M. (1997). Using metaphors as modifiers: Children's production of metaphoric compounds. *Journal of Child Language, 24,* 567-601.

Harris, R. J., Friel, B. M., & Mickelson, N. R. (2006). Attribution of discourse goals for using concrete- and abstract-tenor metaphors and similes with or without discourse context. *Journal of Pragmatics, 38,* 863-879.

Hoicka, E., & Akhtar, N. (2010). *Say the Wrong Thing: Toddlers Joke with Jokers, but Correct Foreigners.* Sumitted manuscript.

Hoicka, E. & Gattis, M. (2008). Do the wrong thing: how toddlers tell a joke from a mistake. *Cognitive Development, 23,* 180-190.

Hoicka, E., Jutsum, S., & Gattis, M. (2008). Humor, abstraction, and disbelief. *Cognitive Science, 32,* 985-1002.

Hoicka, E., & Wang, S. (2010). *15-Month-Olds Match Humorous Cues to Humorous Actions*. Submitted manuscript.

Kazmerski, V.A., Blasko, D. G., & Dessalegn, B. G. (2003). ERP and behavioral evidence of individual differences in metaphor comprehension. *Memory & Cognition, 31,* 673-689.

Lee, K., & Cameron, C. A. (2000). Extracting truthful information from lies: Emergence of the expression-representation distinction. *Merrill-Palmer Quarterly, 46,* 1-20.

Lee, K., Cameron, C. A., Doucette, J. & Talwar, V. (2002). Phantoms and fabrications: Young children's detection of implausible lies. *Child Development, 73,* 1688-1702.

Leekam, S. (1991). Jokes and lies: Children's understanding of intentional falsehood. In A. Whiten (ed.). *Natural Theories of Mind*. Oxford: Basil Blackwell.

Lillard, A. S. (1998). Wanting to be it: Children's understanding of intentions underlying pretense. *Child Development, 69,* 981-993.

Loizou, E. (2005). Infant humor: The theory of the absurd and the empowerment theory. *International Journal of Early Years Education, 13,* 43-53.

Ma, L., & Lillard, A. S. (2007). Where is the real cheese? Young children's ability to discriminate between real and pretend acts. *Developmental Psychology, 77,* 1762-1777.

Maas, F. K. (2008). Children's understanding of promising, lying, and false belief. *The Journal of General Psychology, 135,* 301-321.

Markman, E. M., & Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meaning of a word. *Cognitive Psychology, 20,* 121-157.

Marquis, D. (2001). Deprivations, futures and the wrongness of killing. *Journal of Medical Ethics, 27,* 363-369.

Meltzoff, A. N. (1995). Understanding the intentions of others: Re-enactment of intended acts by 18-month-old children. *Developmental Psychology, 31,* 838-850.

Nichols, S., & Stich, S. P. (2003). *Mindreading: An integrated account of pretense, self-awareness, and understanding other minds.* Oxford: Oxford University Press.

Onishi, K., Baillargeon, R., & Leslie, A. M. (2007). 15-month-olds infants detect violations in pretend scenarios. *Acta Psychologica, 124,* 106-128.

Ozcaliskan, S. (2005). On learning to draw the distinction between physical and metaphorical motion: Is metaphor and early emerging cognitive and linguistic capacity? *Journal of Child Language,32,* 281-318.

Pearson, B. Z., (1990). The comprehension of metaphor by preschool children. *Journal of Child Language, 17,* 185-203.

Pien, D., & Rothbart, M. K. (1976). Incongruity and resolution in children's humor: A reexamination. *Child Development, 47,* 966-971.

Rakoczy, H., Tomasello, M., & Striano, T. (2004). Young children know that trying is not pretending: a test of the "behaving-as-if" construal of children's early concept of pretense. *Developmental Psychology, 40,* 388–399.

Reddy, V. (2001). Infant clowns: The interpersonal creation of humour in infancy. *Enfance, 53,* 247-256.

Roberts, R. M., & Kreuz, R. J. (1994). Why do people use figurative language? *Psychological Science, 5,* 159-163.

Sakkalou, E., & Gattis, M. (in press). Fourteen- to 18-month-old infants infer others' intentions from intonation. *Infancy.*

Searle, J. R. (1969). *Speech Acts: An Essay in the Philosophy of Language.* Cambridge: Cambridge University Press.

Searle, J. R. (2005). What is an institution? *Journal of Institutional Economics, 1*, 1–22.

Shultz, T. R. (1974). Development of the appreciation of riddles. *Child Development, 45,* 100-105.

Siegal, M., & Peterson, C. (1996). Breaking the mold: a fresh look at children's understanding of questions about lies and mistakes. *Developmental Psychology, 32,* 322-334.

Siegal, M., & Peterson, C. (1998). Preschoolers' understanding of lies and innocent and negligent mistakes. *Developmental Psychology, 34*, 332-341.

Sorenson, R. (2007). Bald-faced lies! Lying without the intent to deceive. *Pacific Philosophical Quarterly, 88,* 251-264.

Sperber 1984, D., & Wilson, D. (1981). Irony and the use – Mention distinction. In P. Cole (Ed.). *Radical Pragmatics.* New York: Academic Press.

Sroufe, A., & Wunsch, J. (1972). The development of laughter in the first year of life. *Child Development, 43*, 1326-1344.

Talwar, V., Lee, K., Bala, N., & Lindsay, R. C. L. (2002). Children's conceptual knowledge of lying and its relation to their actual behaviors: Implications for court competence examinations. *Law and Human Behavior, 26,* 395-415.

Talwar, V., Murphy, S. M., & Lee, K. (2007). White lie-telling in children for politeness purposes. *International Journal of Behavioral Development, 31,* 1-11.

Tomasello, M. (2003). *Constructing a Language: A Usage-Based Theory of Language Acquisition.* London: Harvard University Press.

Tzouvaras, A. (1998). Logic of knowledge and utterances and the liar. *Journal of Philosophical Logic, 27,* 85-108.

Ungerer, J., Zelazo, P., Kearsley, R., & O'Leary, K. (1981). Developmental changes in the representation of objects in symbolic play from 18 to 34 months of age. *Child Development*, *52*, 186–195.

Wimmer, H., Gruber, S., & Perner, J. (1985). Young children's conception of lying: Moral intuition and the denotation and connotation of "to lie". *Developmental Psychology, 21,* 993-995.

Winner, E., & Leekam, S. (1991). Distinguishing irony from deception: Understanding the speaker's second-order intention.

Spatial Position in Language and Visual Memory: A Cross-Linguistic Comparison

Solveig Bosse (solveig@udel.edu)
Department of Linguistics & Cognitive Science, 46 E. Delaware Ave
Newark, DE 19716 USA

Anna Papafragou (annap@udel.edu)
Department of Psychology, 109 Wolf Hall
Newark, DE 19716 USA

Abstract

German and English speakers employ different strategies to encode static spatial scenes involving the axial position (standing vs. lying) of an inanimate figure object with respect to a ground object. In a series of three experiments, we show that this linguistic difference is not reflected in native speakers' ability to detect changes in axial position in non-linguistic memory tasks. Furthermore, even when participants are required to use language to encode a spatial scene, they do not rely on language during a recognition memory task. These results have implications for the relationship between language and visual memory.

Keywords: Positional verbs, visual memory, space, language and thought

Introduction

It has often been observed that languages make available different strategies to encode spatial relations (Ameka & Levinson, 2007). A question of central interest within the cognitive sciences is how these cross-linguistic differences interact with underlying spatial representations recruited in spatial memory and other cognitive processes. According to an influential position, language exerts a strong influence on cognitive processes (Levinson, Kita, Haun & Rasch, 2002). Based on several cross-linguistic experiments involving spatial tasks, Levinson et al. proposed that spatial frames of reference provided by people's native language affect how people remember spatial arrays. Crucially, such linguistic effects are argued to emerge even when no overt linguistic labels accompany encoding of the spatial scene. The idea is that obligatory spatial distinctions made within one's native language direct attention to those aspects of spatial representation - thereby affecting spatial memory.

According to a different position, effects of native language on mental representation and memory are more limited. For instance, studies have shown that, despite differences in encoding motion events between English and Greek, memory for aspects of motion does not differ across speakers of these languages (Papafragou, Massey & Gleitman, 2002; cf. Gennari, Sloman, Malt & Fitch, 2002 for related results on English vs. Spanish). Other work has also suggested an independence of memory from cross-linguistic differences in spatial encoding (see Munich, Landau & Dosher, 2001; cf. reviews in Gentner & Goldin-Meadow, 2003).

The question of whether (and how) cross-linguistic differences might affect memory for spatial scenes is related to the question of whether (and how) the explicit presence of linguistic labels during spatial encoding might affect memory. Effects of overt labeling, even though not as deep and pervasive as the effects proposed by Levinson et al. (2002), are still important for understanding how language interfaces with other cognitive faculties. Several studies have shown that explicit spatial language can affect spatial memory – even though the scope and mechanisms responsible for such effects are still not well understood. For instance, there is evidence that memory for motion events can be biased depending on whether path (*exit*) or manner (*skip*) verbs accompany the events, regardless of whether the verbs are provided by the experimenter (Billman & Krych, 1998) or generated by participants (Billman, Swilley & Krych, 2000). Relatedly, Feist and Gentner (2007) showed that providing participants with spatial language can influence their behavior in a recognition task. Specifically, viewing ambiguous spatial representations paired with spatial prepositions (e.g., *on*) resulted in a false memory bias towards typical portrayals of the relation encoded by the prepositions. In another demonstration, Archambault, O'Donnell and Schyns (1999) showed that the level at which an object is categorized (general, e.g. "a mug", or specific, e.g. "Steve's mug") influences the time it takes people to detect a change in a picture containing the object. If objects are known at a specific (individual) level, then the changes are detected faster than if the objects are known on a general level. Crucially, in this study, the level of categorization was provided through linguistic labels prior to the main testing phase.

In this paper, we explored the extent of the influence that language can have on spatial memory (including contexts with and without overt linguistic encoding). We focused on an area of spatial encoding that has only recently begun to receive attention in the literature – namely, positional systems (see Ameka & Levinson, 2007) – and compared two languages, English and German, that differ in a specific aspect of spatial-positional encoding. Specifically, German naturally uses positional verbs that specify the axial orientation of the figure object: an object that is perceived to

be upright (whose vertical height exceeds its width) is described with the verb *stehen* 'stand', while an object whose horizontal width exceeds its vertical height is described with the verb *liegen* 'lie'. Although English has equivalent verbs and uses them for humans and other animates, the positions of inanimate objects are typically and canonically described with the English copula *be* (Kutscher & Schultze-Berndt, 2007). Consider, for example, the two scenes in Figure 1. In German, the two scenes would be canonically described with two difference sentences:

1) Das Buch **steht** auf dem Stuhl. (Fig.1a)
   the book stands on the chair
2) Das Buch **liegt** auf dem Stuhl. (Fig.1b)
   the book lies on the chair

In English, however, both scenes can canonically be described with the same sentence:

3) The book **is** on the chair.

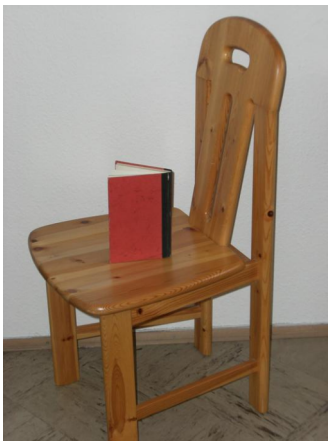


Figure 1a: *stehen* 'stand'

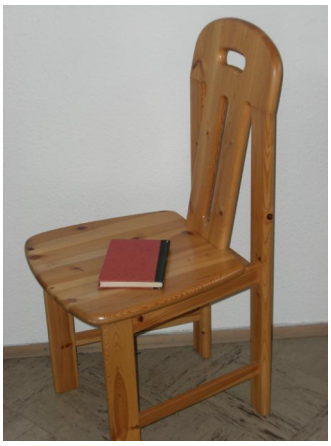


Figure 1b: *liegen* 'lie'

Since this aspect of linguistic encoding represents an obligatory way of encoding spatial position in German but is absent from the corresponding English sentences, we considered it a particularly appropriate test case for the hypothesis that linguistic distinctions influence non-linguistic memory. In a series of experiments (Exp. 1, 2 and 3), we investigated whether this difference in linguistic encoding is mirrored in performance in a (nonverbal) memory task. If language influences memory, then German speakers should perform better than English speakers on a recognition memory task involving changes of posture such as those between Figure 1a and 1b, *even when no language is overtly present as spatial scenes are committed to memory*. If language does not influence memory, native speakers of German and English should perform similarly on the recognition task.

Our studies also addressed the question whether the overt presence of linguistic labels during the encoding of spatial scenes affects memory performance (Exp. 3). Again if overt language affects memory for spatial scenes, German speakers should have an advantage in recognition memory targeting axial position of a figure object compared to English speakers. If recognition memory is independent of overt labeling, no difference in memory for positionals should exist between English and German speakers.

Experiment 1

Experiment 1 was conducted to collect linguistic data to confirm the difference in the use of positional verbs between English and German. The experiment also tested memory performance for the corresponding positions after participants had freely inspected a set of spatial scenes.

Participants

Twenty-six native speakers of German and 28 native speakers of English participated. The German speakers were recruited at Ruprecht-Karls-Universität Heidelberg in Germany, and the English speakers at the University of Delaware in the U.S. None of the German speakers spoke English natively, although almost all had studied it in school, usually alongside other languages. Similarly, none of the native English speakers had native speaker proficiency in German. Equal numbers of men and women were included.

Stimuli

The stimuli consisted of 40 pairs of pictures, taken in color with a digital camera. Each picture depicted two everyday household objects arranged in a particular way. The objects were placed in mostly unconventional pairings (e.g., a boot with a frying pan, a teabag with a wine glass) so that the participants would not focus on the position of the objects but rather on their unpredictable combinations of the objects. Each object appeared in one and only one pair of pictures.

Sixteen of the 40 pairs were test items, which always displayed a figure object on top of a ground object. One picture in each pair depicted the figure object in a standing, vertical position, consistent with the German verb *stehen*, while the other picture depicted the figure object in a lying, horizontal position, consistent with the German verb *liegen*. The position of the ground object was the same in both pictures (see Figure 1 for an actual example drawn from our stimuli). The figure objects had to be medium-sized items that balanced well, could be placed in either a standing or lying position, and would look acceptable in both. We avoided objects that resembled animate beings (e.g., dolls) because English uses *stand* and *lie* for human beings in the upright or horizontal position. In fact, most everyday objects have an inherent orientation — they either stand up or lie flat in their natural state. Therefore, we supplemented our small number of orientation-free figure objects (e.g., lipstick, a roll of paper towels) with an equal number of figures that either naturally "stand" (e.g., a wicker basket) or naturally "lie" (e.g., a wallet), in order to avoid any bias created by unusual positioning.

Another 8 of the 40 pairs of pictures were changing control items (i.e., they involved changes that were

unrelated to the stand-lie distinction). In the changing control pictures, the two objects were placed in a non-support relationship in at least one of the two pictures. Such relationships involved attachment (e.g., a paper clip on a pen), containment (e.g., a banana in a bowl), or piercing (e.g., a knife in an apple). The difference between the two arrays in each pair were either changes of state (e.g., a banana becomes a peeled banana) or non-axial changes in position (e.g., a paper clip originally attached to the cap of the pen becomes attached to the body of the pen).

Finally, 16 of the 40 pairs of pictures were non-changing control items. The two members of each pair were identical to each other and depicted relationships of support (with one object resting on top of another), attachment, or containment.

These pairs of pictures were arranged for display in two lists of 40 pictures each. One picture of each pair became part of List 1 and the second picture of each pair became part of List 2. Within each List, half of the test items depicted a standing relation and half a lying relation. Lists 1 and 2 displayed pictures in two different random orders. We also created two more lists (Lists 3 and 4) by reversing the presentation order of Lists 1 and 2. For the memory task, we arranged these lists into four different working orders that varied in terms of which list was used during the initial (encoding) phase vs. the second (memory) phase (List 1 vs. 2, 2 vs. 1, 3 vs. 4, or 4 vs. 3 respectively).

For the language task, we selected a subset of these stimuli for presentation. Specifically, we only used List 1 and List 2 but omitted the non-changing control items such that each list contained 16 test and 8 changing control items only.

Procedure

Language Task For the language task, we tested 10 German speakers and 12 English speakers. Participants viewed either the (shortened) List 1 or the (shortened) List 2. They were told that each picture would depict two household objects paired together, and were asked to describe each arrangement with a single complete sentence. Participants recorded their responses on a lined answer sheet and controlled the pace of the task by advancing the display themselves.

Memory Task For the memory task, we tested 16 German speakers and 16 English speakers. None of these had participated in the language task. Each participant was assigned to one of the four stimuli orders. The participants were simply told that they would see a set of pictures and their task was to look at the pictures carefully. During this encoding phase, each picture appeared for two seconds before the display automatically advanced to the next picture. Then participants were told that they would view a second set of pictures and were instructed to verbally provide fast judgments of whether each picture was the "same" or "different" (i.e., whether the exact same picture had appeared in the first round, or the picture was similar to a picture that had appeared before but was also recognizably changed). The pictures in the memory phase were also displayed for two seconds each. If a participant did not provide an answer within those two seconds, his or her response was discarded.

Results and Discussion

Language Task As the dependent variable, we calculated the percentage of answers that included a positional term for each language group. All positional information was encoded in verbs, namely German *stehen* and *liegen* and their English equivalents *stand* and *lie*. German speakers encoded position 90% of the time while English speakers encoded position only 32.3% of the time. This difference is significant (two-tailed t-test, $p<.001$). Thus, as expected, native speakers of German are more likely to encode the detailed spatial position of a figure object than English speakers.

Memory Task The results for this task are displayed in Figure 2. (All error bars in this paper indicate standard error.) For this and the following memory experiments, the dependent variable is the percentage of correctly identified pictures. An ANOVA with Language (German, English) and Trial (Test, Changing Control, Non-Changing Control) as factors returned only a main effect of Trial ($F(2,29)=22.04$, $p<.0001$). The effect is due to a significant difference between Test items ($M = 69.73$) and Changing Control items ($M = 88.49$; $p<.0001$), as well as a difference between Test items and Non-Changing Controls ($M = 91.51$; $p<.0001$). Thus, despite differences between English and German in the labeling of spatial position, English speakers did not perform differently from German speakers in memory for spatial position.

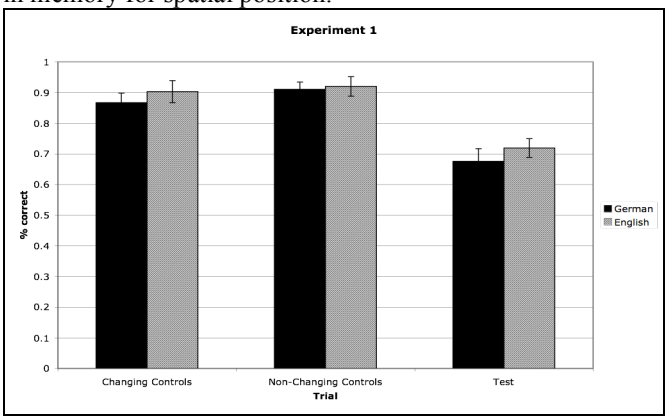


Figure 2: Accuracy in Memory Task of Experiment 1

Experiment 2

One possible explanation for the lack of native language influence in the memory task of Experiment 1 is that participants were not warned that memory for pictures would be tested. It is possible that prior knowledge of the nature of the task would make people more likely to recruit linguistic resources to encode the objects and relations in the pictures in anticipation of later testing. In Experiment 2, we

tested this hypothesis. Specifically, we replicated Experiment 1 but made participants aware of the fact that they would have to remember the pictures. To further bolster the opportunity to use linguistic labels (and store both the labels and the visual scene in memory) we introduced a temporal gap between pictures during the encoding phase. We reasoned that this lag of time would allow participants to encode stimuli verbally even if they were not specifically instructed to do so.

Participants

Sixteen native speakers of German were recruited at the Carl-von-Ossietzky Universität Oldenburg in Germany, and 16 English speakers were recruited at the University of Delaware in the U.S. None of the German speakers spoke English natively, although almost all had studied it in school. Similarly, none of the native English speakers had near-native speaker proficiency in German. Approximately equal numbers of men and women were included. None of these people had participated in Experiment 1.

Stimuli

The same materials as in Experiment 2 were used.

Procedure

The same procedure as for the memory task in Experiment 1 was used but with two modifications. First, participants were told that this would be a memory experiment and that they needed to remember the pictures they would see for a later recognition test. Second, 3 seconds of blank screen were inserted between pictures in the encoding phase.

Results and Discussion

The results are displayed in Figure 3. An ANOVA with Language and Trial as factors returned only a main effect of Trial ($F(2,29)=26.42$, $p<.0001$). This effect is driven by lower performance on Test items ($M = 71.29$) compared to the Changing Controls ($M = 90.23$; $p<.0001$) and the Non-Changing controls ($M = 93.16$; $p<.0001$). Thus even when participants know that they are participating in a memory experiment and are given the opportunity to encode the stimuli linguistically, linguistic encoding does not appear to affect the outcome of recognition memory.

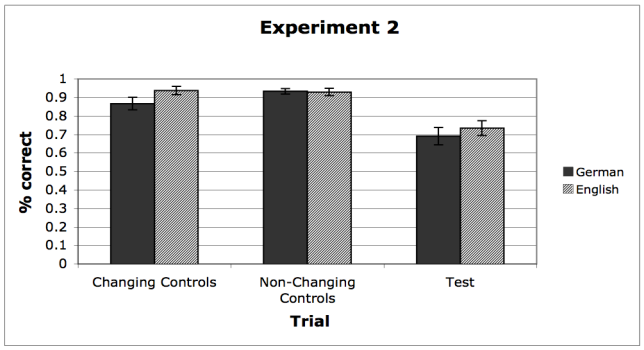


Figure 3: Accuracy in Memory Task of Experiment 2

Experiment 3

Participants in Experiment 2, even though given the opportunity to encode the visual scenes linguistically, did not necessarily do so. It is an open question whether, under different conditions (e.g., a more difficult task), participants might spontaneously recruit labels implicitly as an additional encoding strategy (which would lead to English-German differences in memory accuracy here). Experiment 3 followed the basic method of Experiment 2 but introduced a novel manipulation to address this question.

Specifically, we included a Non-Linguistic Shadowing condition in which participants engaged in a secondary task (shadowing a rhythm by tapping) while inspecting the scenes: crucially, this shadowing task did not engage the language faculty. We hypothesized that, because of the high cognitive load imposed by the secondary task, participants would be likely to recruit language as an additional means of encoding the scenes in preparation for the memory test. If so there could be language-specific patterns in memory performance. For comparison purposes, we also included a Linguistic Shadowing condition in which participants engaged in a comparable secondary task that blocked the language code (verbally shadowing a rhythm). This task was not expected to lead to recruitment of labels during encoding (or to cross-linguistic differences in spatial memory). Hermer-Vazquez et al. (1999) showed that these two types of shadowing tasks impose the same cognitive load but employ different cognitive resources. Thus, labeling could be possible with Non-Linguistic Shadowing but not in Linguistic Shadowing.

Experiment 3 also tested the hypothesis that, when forced to provide linguistic labels explicitly, participants would use such labels later during the recognition phase (thus triggering language-specific effects on memory performance). In a Linguistic Completion condition, participants were asked to fill out a sentence after each scene describing the scene they saw; critically, they had to provide the spatial verb describing the relationship between the figure and ground object. German speakers were expected to produce more positional verbs than English speakers. Importantly, if labels can affect visual memory, we should expect an advantage for German speakers compared to English speakers in later recognition of standing vs. lying object positions. This manipulation provides a powerful test for the hypothesis that labels affect memory performance by virtually guaranteeing the presence of labels (hence of cross-linguistic labeling differences) during the initial inspection of visual scenes.

Participants

Thirty-six native speakers of German were recruited from either the Carl-von-Ossietzky Universität Oldenburg or the Gymnasium Nordenham in Germany. All had learned English but none of them spoke it natively. Thirty-six native speakers of English were recruited at the University of Delaware. No native speaker of English was fluent in

German. None of these participants had participated in Experiment 1 or 2. Approximately equal numbers of men and women participated.

Stimuli

The same materials as in Experiment 2 were used.

Procedure

Participants were randomly assigned to one of three conditions:

Non-Linguistic Shadowing Procedure was as in Experiment 2 but participants wore headphones during the encoding phase and listened to an irregular rhythm. Their task was to repeat the rhythm by tapping it with their fingers on the table.

Linguistic Shadowing Participants followed the same procedure as those in the Non-Linguistic Shadowing condition except that they had to repeat the rhythm constantly using the syllable "na" (they had to say the syllable loud enough for the experimenter to hear it).

Linguistic Completion Procedure was as in Experiment 2 with some modifications. After each picture in the encoding phase, instead of a blank screen, participants saw a screen displaying a sentence. The sentence was presented for 3 seconds and appeared in the native language of each participant. The sentence described the preceding spatial scene but was missing the verb and the ground object. For instance, for Figure 1a above, English speakers saw "The book ____ on the ____." Participants were instructed to read the sentence out loud during the time it was displayed adding in the missing words (the ground object was omitted so that English speakers would not simply have to provide the copula *is* throughout). Sentences were recorded and later transcribed for coding.

Results and Discussion

Non-Linguistic and Linguistic Shadowing Conditions The results from the memory task for these two conditions are presented in Figures 4a-b. For the Non-Linguistic Shadowing condition, an ANOVA with Language and Trial as factors returned only a main effect of Trial ($F(2,21) = 10.5$, $p<.001$). The effect is driven by lower performance on Test items (M = 62.6) compared to Changing Controls (M = 83.9) and to Non-Changing Controls (M = 80.2, $p<.05$). A similar ANOVA for the Linguistic Shadowing condition gave similar results (main effect of Trial, $F(2,21) = 13.47$, $p< .0001$, with lower performance on Test items (M = 61.1) compared to Changing and Non-Changing Controls (M = 79.7 and 78.6 respectively, $p<.05$)). No difference was observed between performance in the two shadowing conditions ($p>.05$). Thus even in a task with higher cognitive demands that allows for the use of the linguistic code, language does not seem to have an effect on scene representations recovered from memory.

Linguistic Completion As expected, participants' linguistic productions confirmed the asymmetry between English and German: German speakers offered verbs encoding the (correct) position of the figure object for 73.3% of the Test items; English speakers did so for only 2.8% of these items. This difference is significant (two-tailed t-test, $p<.05$).

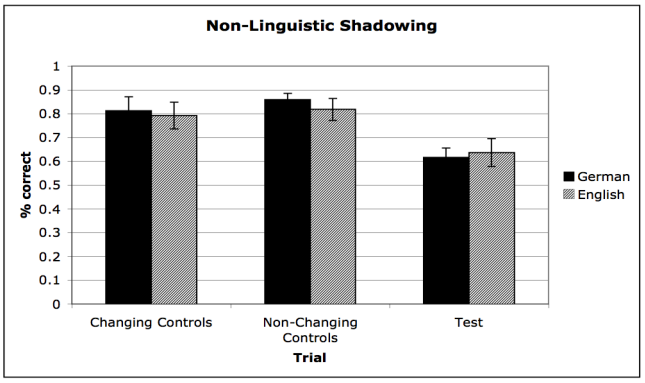


Figure 4a: Accuracy in Memory Task (Non-Linguistic Shadowing Condition) of Experiment 3

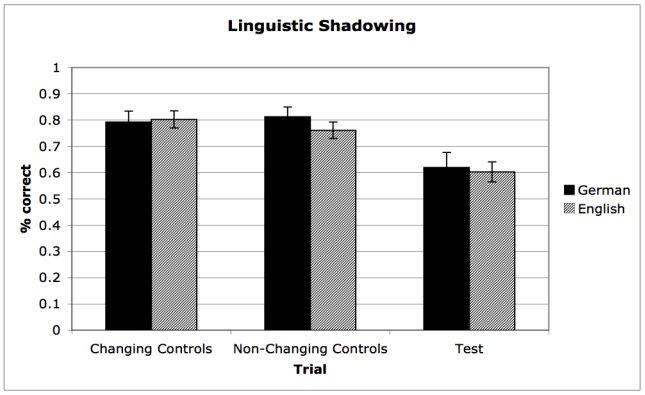


Figure 4b: Accuracy in Memory Task (Linguistic Shadowing Condition) of Experiment 3

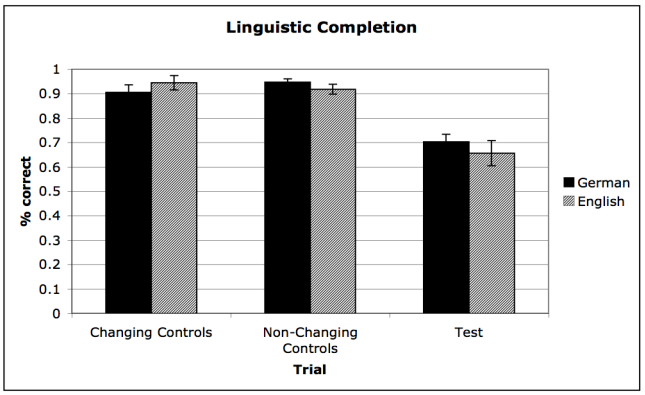


Figure 4c: Accuracy in Memory Task (Linguistic Completion Condition) of Experiment 3

For the memory data (Figure 4c), an ANOVA with Language and Trial returned only a main effect of Trial ($F(2,21)=47.29$, $p<.0001$). This effect is driven by lower performance on Test items (M = 66.3) than Changing Controls (M = 92.1) and Non-Changing Controls (M = 93.7). German speakers – unlike English speakers - overwhelmingly mentioned the axial position of the object

in filling out the target sentences but this linguistic encoding did not lead to an advantage in remembering axial position.

General Discussion

In this study, we asked whether differences in the way English and German encode the axial position of a figure object affect recognition memory for axial position. Our results suggest that cross-linguistic differences in positional encoding have no influence on memory for spatial scenes. Specifically, in a variety of contexts allowing or encouraging the choice to encode the scenes linguistically, participants did not appear to make this choice. These results argue against theoretical positions according to which obligatory lexical or grammatical distinctions in a language create cognitive biases in speakers even in situations where no language is overtly present (e.g., Levinson et al., 2002). Our data are consistent with prior finding showing that spatial memory is independent from cross-linguistic differences in spatial vocabulary (Papafragou et al., 2002; Gennari et al., 2002).

A particularly noteworthy aspect of our data is that native language distinctions failed to affect recognition memory even when participants explicitly provided linguistic encoding of the spatial scenes (Linguistic Completion condition of Exp. 3). This finding differs from previous reports which found effects of explicit labeling on visual memory in speakers of a single language (see Introduction). To reconcile these divergent findings, one possibility is that language effects are more likely to surface when labels occur before (as in Feist & Gentner, 2007; Archambault et al., 1999; Billman et al., 2000) or during (as in Billman & Krych, 1998) the encoding of visual scenes rather than after visual encoding has occurred (as in our Exp. 3). In support of this possibility, work by McCloskey and Zaragoza (1985) showed that verbally presented misinformation about an object *after* an object had been viewed (e.g., referring to a hammer as a screwdriver) did not impair participants' ability to later recognize the object, as opposed to a new, previously unmentioned, object. Nevertheless, this explanation cannot account for other work showing that, even when linguistic labels are generated as spatial scenes or events are viewed, they do not necessarily alter visual memory (Papafragou et al., 2002; Gennari et al., 2002).

Another possibility is that language effects are more likely to emerge when the visual scenes to be remembered are ambiguous (Feist & Gentner, 2007) or can be categorized on several levels (Archambault et al., 1999), and thus allow language to play a disambiguating role. Regardless of the specific explanation that will turn out to be correct, the fact that linguistic labels in the Linguistic Completion condition degraded faster than the visual memory of the scenes provides evidence that linguistic and visual representation of spatial position belong to different levels of representation and are potentially independent of each other. Further work is needed to specify the precise factors that affect language intrusions into non-linguistic cognitive processes. Nevertheless, results from the Linguistic Completion task suggest that such intrusions depend on subtle features of the task at hand and do not generalize across all contexts in which language is used to label spatial scenes.

Acknowledgments

We would like to thank the members of the Language and Cognition Lab and the participants at the University of Delaware and in Germany. This research was partly supported by NIH grant 5R01HD55498-2 to A.P.

References

Ameka, F. & Levinson, S.C. (2007). Introduction: the typology and semantics of locative predicates: posturals, positionals, and other beasts. *Linguistics* 45, 847-871.

Archambault, A., O'Donnell, C. & Schyns, P. (1999). Blind to object changes: when learning the same object at different levels of categorization modifies its perception. *Psychological Science* 10 (3), 249-255.

Billman, D. & Krych, M. (1998). Path and manner verbs in action: effects of "skipping" or "exiting" on event memory. *Proceedings from the 20th Annual Conference of the Cognitive Science Society.* Hillsdale, NJ: Erlbaum.

Billman, D., Swilley, A. & Krych, M. (2000). Path and manner priming: verb production and event recognition. *Proceedings from the 22nd Annual Conference of the Cognitive Science Society.* Hillsdale, NJ: Erlbaum.

Feist, M. & Gentner, D. (2007). Spatial language influences memory for spatial scenes. *Memory & Cognition* 35 (2), 283-296.

Gennari, S., S., Sloman, S., Malt, B. & Fitch, T. (2002). Motion events in language and cognition. *Cognition* 83, 49-79.

Gentner, D. & Goldin-Meadow, S., eds. (2003). *Language in mind*. Cambridge, MA: MIT Press.

Hermer-Vazquez, L., Spelke E., Katsnelson, A. (1999). Sources of Flexibility in Human Cognition: Dual-Task Studies of Space and Language. *Cognitive Psychology* 39, 3-36.

Kutscher, S. & Schultze-Berndt, E. (2007). Why a folder lies in the basket although it is not lying: the semantics and use of German positional verbs with inanimate figures. *Linguistics* 45, 983-1028.

Levinson, S., Kita. S., Haun, D. & Rasch, B. (2002). Returning the tables: language affects spatial reasoning. *Cognition* 84, 155-188.

McCloskey, M., & Zaragoza, M. (1985). Misleading postevent information and memory for events: Arguments and evidence against memory impairment hypotheses. *Journal of Experimental Psychology: General* 114, 1-16.

Munnich, E., Landau, B. & Dosher, B. (2001). Spatial language and spatial representation: A cross-linguistic comparison. *Cognition* 81, 171-207.

Papafragou, A., Massey, C. & Gleitman, L. (2002). Shake, rattle, 'n' roll: The representation of motion in language and cognition. *Cognition* 84, 189-219.

Generalized Event Knowledge Activation During Online Language Comprehension

Ross Metusalem (rmetusal@cogsci.ucsd.edu)
University of California, San Diego
9500 Gilman Dr. #0515
La Jolla, CA 92093-0515 USA

Marta Kutas (kutas@cogsci.ucsd.edu)
University of California, San Diego
9500 Gilman Dr. #0515
La Jolla, CA 92093-0515 USA

Mary Hare (mlhare@bgnet.bgsu.edu)
Bowling Green State University
Department of Psychology, BGSU
Bowling Green, OH 43403-0228 USA

Ken McRae (kenm@uwo.ca)
University of Western Ontario
Department of Psychology, UWO, Social Science Centre 7418
London, ON, Canada

Jeffrey L. Elman (elman@cogsci.ucsd.edu)
University of California, San Diego
9500 Gilman Dr. #0515
La Jolla, CA 92093-0515 USA

Abstract

Online language comprehension is guided by knowledge regarding real-world events. However, it remains unclear whether activation of event knowledge during language comprehension is constrained by the linguistic context or is generalized, including a wide variety of information associated with the event even if that information has not been mentioned previously and does not satisfy constraints imposed by the local linguistic context. The present study addresses this issue by analyzing event-related brain potentials recorded as participants read brief scenarios describing typical real-world events. The amplitude of the N400 elicited by a contextually anomalous word was reduced if that word was related to the event described. This result suggests that online language comprehension involves construction of rich event representations that include information beyond that which is relevant to the processing of the current linguistic input.

Keywords: event knowledge; online language comprehension; event-related potentials; ERP; N400

Background

Online language comprehension is a rapid and incremental process guided by a wide variety of information sources. Some researchers characterize this process as the incremental mapping of linguistic structure onto real-world event structure, mediated in part by the comprehender's prior knowledge associated with the described event (e.g., Altmann & Mirković, 2009). Recent research has highlighted the importance of event knowledge to online language comprehension. At the lexical level, priming studies have shown, for example, that verbs activate agents, patients, and instruments typically associated with the specific actions denoted by the verbs (Ferretti, McRae, & Hatherell, 2001), that agent, patient, instrument, and location nouns activate verbs denoting the events in which they typically participate (McRae, Hare, Elman, & Ferretti, 2005), and that word triplet priming with lexically unassociated primes and targets reveals rapid activation of script information (Chwilla & Kolk, 2005). Such findings suggest that processing words in isolation activates event knowledge, resulting in subsequent activation of other entities and/or actions associated with the event.

At the sentential level, self-paced reading (Bicknell, Elman, Hare, McRae, & Kutas, 2008) and eye-tracking (Kamide, Altmann, & Haywood, 2003) studies have demonstrated that comprehenders can rapidly integrate information provided by a verb in combination with its preceding agent in order to predict likely upcoming patients. For example, Kamide et al. monitored participants' eye movements around a visual scene as the participants listened to sentences such as *The* [*man/girl*] *will ride the* [*motorbike/carousel*]. They found more anticipatory looks to the picture of the motorbike when the agent of *ride* was *man* than when the agent was *girl* (and similarly more looks to the carousel for *girl will ride* than for *man will ride*), even though both a motorbike and carousel are equally plausible patients of the verb *ride*. This result demonstrates that thematic role assignment is not guided by the verb alone,

but crucially by knowledge associated with the event denoted by the agent-verb combination. Online language comprehension thus makes rapid use of event knowledge.

While event knowledge is clearly important to linguistic processing, the specificity of the event knowledge activated during comprehension remains an open question. Is the activated event knowledge general, containing a wide variety of salient features associated with the event? Or is activation restricted to only what is relevant to the current linguistic context? Consider the passage in (1):

(1) A huge blizzard swept through town last night. My kids ended up getting the day off from school. They spent the whole day outside building a big snowman in the front yard.

Given the previously discussed findings, it is likely the case that at the point of reading *building a*, a comprehender's knowledge regarding "playing in the snow" events allows for *snowman* to become activated as a likely patient (as opposed to *house*, for example). Is activation of event knowledge at this point limited to this feature of the "playing in the snow" event? If this were the case, it would indicate that event knowledge activation is constrained by the linguistic context, limiting activation to those event features that are relevant to the processing of the current linguistic input. However, it is also possible that an entire body of "playing in the snow" knowledge is activated in reading this passage. This knowledge might include, for example, the fact that the children are probably jackets, hats, and mittens, even though this has not been explicitly mentioned and is not directly relevant to comprehending the passage. This study seeks to determine whether event knowledge activation during comprehension involves rich, generalized representations such as this, or if it is limited to what is currently relevant given the linguistic context.

In the present study, participants read brief passages describing typical real-world events. The final sentence of each passage contained either a highly expected target word (in the above example, *building a snowman*) or one of two contextually anomalous target words: one related to the event described (e.g., *building a jacket*) and one unrelated to the event described (e.g., *building a towel*). Participants' EEG was recorded as they read the passages, and the event-related brain potentials (ERPs) elicited by these three target types were contrasted. The analysis focuses on the N400, an ERP component whose amplitude is inversely proportional to the degree to which a word is expected given the preceding context (Kutas & Hillyard, 1984). It was predicted that if language comprehension involves generalized event knowledge activation, then the event-related anomalous target should become activated during the reading of the passage, while the event-unrelated anomalous target should not. This should then result in a graded N400 effect: the smallest N400 to the expected target, the largest N400 to the event-unrelated target, and an intermediate N400 to the event-related target. Such a result would indicate that although both the event-related and event-unrelated targets violate local linguistic constraints, the event-related target becomes activated by virtue of its association with the event described. More generally, this result would support the notion that language comprehension involves generalized event knowledge activation.

Stimuli

Seventy-two experimental items (scenarios) were constructed. Each scenario consisted of three sentences and described a common real-world event. The first two sentences established the event (e.g., playing in the snow). The final sentence contained one of three sentence-medial target words: a highly expected word, a contextually anomalous word that was related to the established event (event-related anomalous target; ERA), or an equally anomalous word that was unrelated to the established event (event-unrelated anomalous target; EuRA). Each experimental item thus had three possible target words, giving the experiment three conditions.

Expected Targets

The expected targets were obtained via a cloze task in which participants read each scenario up to the word preceding the target and were asked to provide the single word most likely to come next. Participants completed the cloze task through an online form. Scenarios were presented one at a time, with all three sentences presented in paragraph format. The third sentence left off at the word preceding the target, and participants provided the most likely upcoming word in a blank text field before moving onto the next scenario. Responses could not be modified once entered.

Thirty undergraduates (twenty-three women) at the University of California, San Diego participated for course credit. All were native English speakers. Cloze probability was calculated as the percentage of participants who provided a particular response for the given scenario. The response with the highest cloze probability was chosen as the expected target word. Across the seventy-two items, mean cloze probability of the expected target was 0.81, with a standard deviation of 0.17.

Event-Related Anomalous Targets

To obtain the event-related anomalous targets (ERAs), a new group of participants completed a norming task in which they provided a list of people or things most likely to be present at each event. Participants completed this task through an online form. Scenarios were presented one at a time in paragraph format, with the expected target word obtained in the previous cloze task now filled in. Participants were instructed to read each item and to paint a mental picture of the event described. They were told that their picture would likely include prominent people or things that would participate in the event, but were not explicitly mentioned in the text. They were asked to provide up to five responses for each scenario. Responses could not be modified once entered.

Table 1: Norming results for the three rotation groups and the stimuli set overall[1]

		Group 1	Group 2	Group 3	Overall
Expected targets	Cloze probability	0.81	0.80	0.80	0.81
	Log frequency	6.95	7.01	6.88	6.95
	Orthographic length	5.58	5.71	5.75	5.68
Event-related targets	Cloze probability	0.00	0.00	0.00	0.00
	Log frequency	6.89	6.91	6.76	6.86
	Orthographic length	5.96	5.96	5.71	5.87
	Event-relatedness score	88.1	86.5	92.5	89.1
Event-unrelated targets	Cloze probability	0.00	0.00	0.00	0.00
	Event-relatedness score	0.00	0.13	0.04	0.06

Forty-five undergraduates (twenty-six women) at the University of California, San Diego participated for course credit. All were native English speakers, and none had participated in the cloze task. Each participant's responses for a particular scenario were given weighted scores based on response order (i.e., 5 for the first response, 4 for the second response, 3 for the third response, etc.). The highest scoring response that was not provided as a response in the previous cloze task (i.e., had a cloze probability of zero) was chosen as the ERA for the item. In a small number of instances, the highest scoring zero-cloze response was deemed to be a contextually sensible continuation of the scenario, despite having not been provided as a response in the cloze task. In these cases, the next highest scoring zero-cloze response was chosen. Across the seventy-two scenarios, the mean event-relatedness score of the ERA was 89.1, with a standard deviation of 34.4.

Event-Unrelated Anomalous Targets

Event-unrelated anomalous targets (EuRAs) were obtained by shuffling the ERAs across scenarios. Before this was done, the seventy-two experimental items were split into three rotation groups of twenty-four items each, allowing for three experimental lists to be constructed by rotating each group through the three conditions across the three lists. To minimize variability across the experimental lists, the rotation groups were matched on the following factors: mean cloze probability, log frequency, and orthographic length of the expected target; mean event-relatedness score, log frequency, and orthographic length of the ERA.

ERAs were shuffled across the items within each rotation group to obtain the EuRAs, thereby matching the ERAs and EuRAs for lexical factors within each group. EuRAs were chosen such that they were all zero-cloze, and in all but two of the seventy-two scenarios, EuRAs had event-relatedness scores of zero. (The two exceptions had extremely low event-relatedness scores of 1 and 3.) In addition, the shuffling was done in such a way as to match the ERAs and EuRAs within each scenario for animacy and concreteness. This was done so that if the ERA constituted an animacy or concreteness violation with respect to preceding context, the EuRA constituted the same violation. The norming results for each rotation group and the stimuli set overall are presented in Table 1.

Lexical Associations

One concern in constructing the stimuli was that the ERAs might be significantly more likely than the EuRAs to be lexically associated with the expected word. Such a confound might undermine the experiment, as the predicted graded N400 effect could then be accounted for by priming of the ERA by the expected word, as opposed to the activation of event knowledge. The University of South Florida Free Association Norms (Nelson, McEvoy, & Schreiber, 1998) were consulted to ensure that the ERAs and EuRAs were on average associated to equal degrees with their corresponding expected targets. Sixty-five of the seventy-two scenarios' expected targets appeared in the Nelson norms; mean association scores for the ERAs and EuRAs were calculated across these sixty-five items. The mean association score for the ERAs was 0.0005, and for the EuRAs was 0.0001.[2] These extremely low mean association scores and the small difference between them were deemed acceptable for the purposes of the study.

[1] Recall that the event-unrelated and event-unrelated targets consist of the same lexical items. Mean log frequency and orthographic length of the EuRAs is therefore equal to that of the ERAs and is not reported in the Table 1.

[2] The association score for a given word is calculated simply as the proportion of participants that provided the word (ERA or EuRA) in response to the cue word (expected target). The mean scores for ERAs and EuRAs thus correspond to one response per two thousand participants and one response per ten thousand participants, respectively.

Experiment

To examine the specificity of the event knowledge activated during online language comprehension, participants' EEG was recorded as they read the carefully constructed scenarios described in the previous section. To review, each scenario consisted of three sentences. The first two sentences established a typical real-world event. The third sentence contained one of three possible target words: a highly expected word, an anomalous but event-related word (ERA), or an anomalous and event-unrelated word (EuRA). It was hypothesized that the expected word would elicit the smallest N400, the EuRA would elicit the largest N400, and the ERA would elicit an intermediate N400. Such a finding would indicate that although both the ERA and EuRA violated local linguistic constraints, the ERA was activated through the activation of generalized event knowledge by the preceding context.

Methodology

Materials The materials consisted of seventy-two scenarios constructed according to the previously discussed criteria. Three experimental lists were created based on the grouping of the seventy-two items into the three rotation groups, such that each experimental item occurred exactly once in each condition across the three lists and exactly once in each list. In addition to the seventy-two scenarios, twenty-four fillers were included. Like the experimental items, these were three-sentence scenarios describing real-world events. None contained any anomalous words. Presentation order of experimental items and fillers was fully randomized for each participant.

Participants Thirty undergraduates (twenty-two women) at the University of California, San Diego participated for course credit. All were right-handed native monolingual English speakers with normal or corrected-to-normal vision. None reported any history of learning or reading disabilities or neurological or psychiatric disorders.

Procedure Participants sat in an electromagnetically shielded chamber and read each scenario from a computer monitor. The first two sentences of each scenario were presented in paragraph format. Once participants understood the two sentences, they pushed a button to advance to the final sentence. The final sentence was presented via rapid serial visual presentation (RSVP) with a stimulus onset asynchrony (SOA) of 350ms and a stimulus duration of 200ms. After the offset of the final word, participants answered a yes-no comprehension question before advancing to the next trial. Response hand was

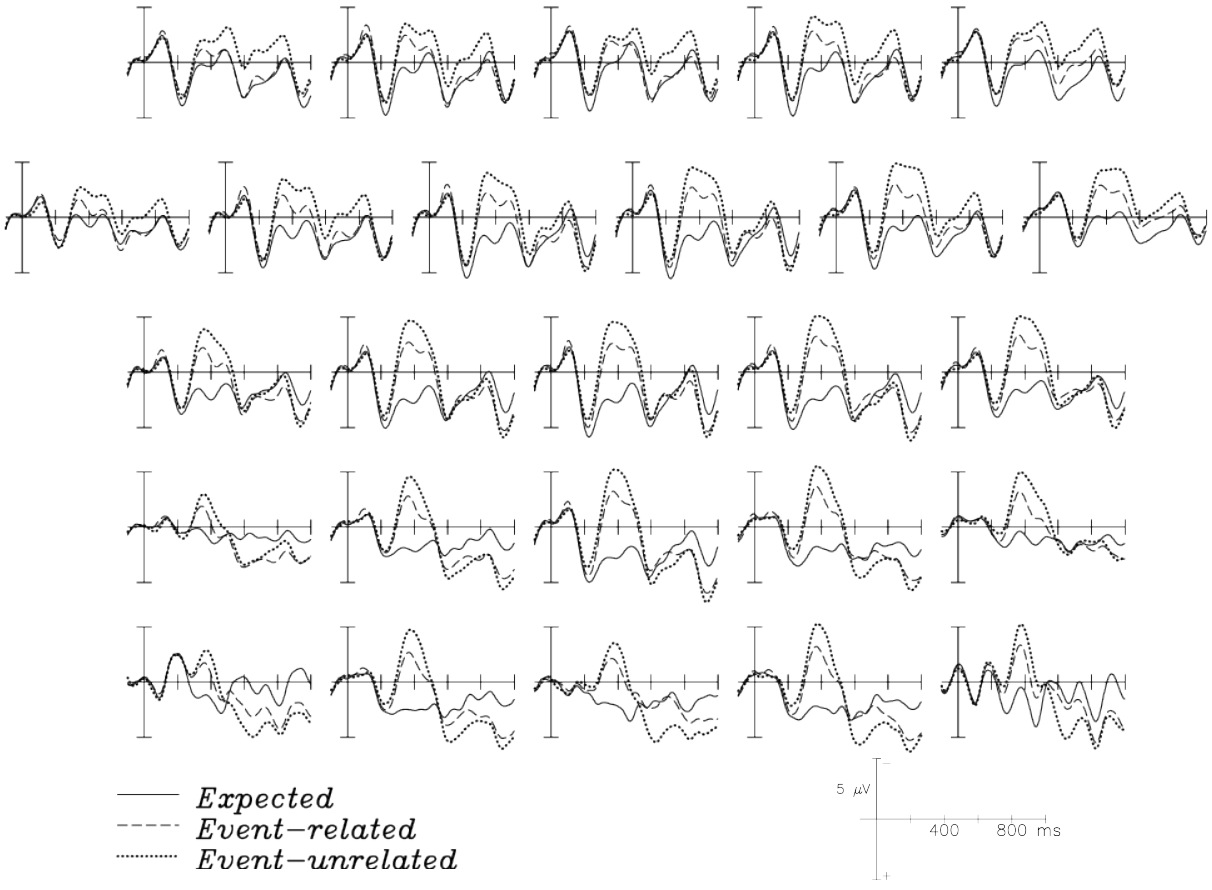


Figure 1: Grand average ERPs at all electrode sites

counterbalanced across participants.

EEG was recorded from twenty-six electrodes distributed evenly over the scalp, referenced online to the left mastoid and re-referenced offline to the average of the left and right mastoids. Electrodes were placed on the outer canthus and infraorbital ridge of each eye to monitor eye movements and blinks. All electrode impedances were kept below 5KΩ. EEG was amplified with Nicolet amplifiers with a bandpass of 0.016 to 100 Hz and digitized at a rate of 250 samples per second.

Results

Before analysis, all epochs containing artifacts caused by blinks, eye movements, muscle tension, channel drift, or amplifier blocking were rejected offline. Participants' responses to the comprehension questions were analyzed to ensure that each participant was reading the scenarios for comprehension. Only one participant scored below 90% correct (88.9%), indicating that participants were comprehending the scenarios.

EEG was time-locked to the onset of the target words and was first averaged within participants to obtain individual participant averages for each condition. These individual participant averages were then averaged together to obtain a grand average ERP waveform for each condition. Figure 1 contains the grand average ERPs at each electrode site, arranged according to the distribution of electrodes over the scalp (i.e., frontal electrodes at the top, posterior electrodes at the bottom; midline electrodes in the middle, lateral electrodes to the side); Figure 2 presents a close-up of the grand average ERPs at the midline parietal recording site (Pz). To conduct an analysis of N400 amplitude, mean amplitudes from 200 to 500ms post-stimulus onset (relative to a 500ms pre-stimulus baseline) at each electrode for each participant were entered into a repeated measures ANOVA. A main effect of Condition was obtained [$F(2,58)=38.33$, $p<0.0001$], as was a Condition X Electrode interaction [$F(50,1450)=7.26$, $p<0.0001$]. Planned comparisons revealed the event-unrelated condition to be significantly more negative than the event-related condition [$F(1,29)=13.00$, $p<0.01$], which in turn was more negative than the expected condition [$F(1,29)=35.44$, $p<0.0001$]. This result confirms the predicted graded N400 effect: expected targets elicited the smallest N400, event-unrelated anomalous targets the largest N400, and event-related anomalous targets an intermediate N400. Analysis of the distribution of the N400 effect revealed a significant Condition X Hemisphere interaction [$F(2,58)=9.69$, $p<0.001$], a significant Condition X Laterality interaction [$F(2,58)=15.14$, $p<0.0001$], and a significant Condition X Anteriority interaction [$F(6,174)=4.96$, $p<0.01$], indicating that the N400 effect exhibited a posterior, slightly right-lateralized distribution across the scalp.

Discussion

Previous research has demonstrated the important role that event knowledge plays in online language comprehension. However, the specificity of event knowledge activation has remained an open question. The results of the present study suggest that activated event knowledge is general, containing elements beyond what is relevant to the processing of the current linguistic input. This conclusion is supported by a reduction in N400 amplitude to a contextually anomalous word when that word is related to the event being described (and is crucially unrelated to the most expected word, as determined by consulting the South Florida Free Association Norms). This result shows that a wide range of event-relevant information is activated during online language comprehension, as opposed to only event-relevant information that meets local linguistic constraints.

It is important to note that while an event-related anomalous word elicits a reduced N400, it still elicits a larger N400 than a highly expected word. In the present study, the expected targets were more plausible patients of the preceding verbs than were the event-related targets. It is thus possible that event-related elements are activated in a gradient fashion, with those that satisfy the constraints imposed by the local linguistic context receiving greater activation. As the aforementioned study by Ferretti et al. (2001) suggests, verbs encode thematic roles in an event-specific fashion. According to this view, expected targets were closely related to the specific event denoted by the verb itself, whereas the event-related targets were related to the event described by the scenario as a whole but unrelated to the specific event denoted by the verb. This suggests that a word related to the event conveyed by the global linguistic context will receive even greater activation if it is also compatible with the event denoted by the verb.

While it is argued here that the N400 reduction for event-related targets results from activation of event knowledge, it might be argued that such a finding could arise if the event-related targets were lexically associated with the words in the preceding contexts to a greater degree than the event-unrelated targets. Associations between targets and their contexts were quantified using Latent Semantic Analysis (Landauer, Foltz, & Laham, 1998; http://lsa.colorado.edu/), a method for assessing word associations through analysis of the distribution of words in large-scale corpora. Each target received an association score between 0 and 1, with expected targets receiving a mean score of 0.276, event-related targets a mean score of 0.268, and event-unrelated targets a mean score of 0.220. A paired t-test confirmed the difference between event-related and event-unrelated targets ($p=0.002$). While this result suggests that the event-related targets were more strongly associated with the words in the preceding contexts than were the event-unrelated targets, this is in fact compatible the event knowledge account. Language describing real-world events will undoubtedly exhibit statistical regularities mirroring the structure of the real-world events themselves, and thus lexical co-occurrence measures calculated over large corpora should reflect the event structures encoded in event knowledge. In addition, event knowledge presumably can be derived from experience both in the real world and with language

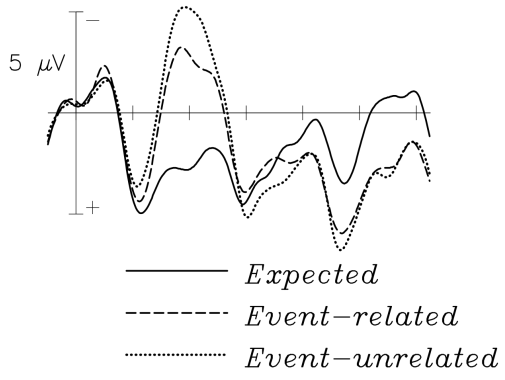


Figure 2: Grand average ERPs at Pz

describing that world, meaning that co-occurrence patterns in language likely result in part from the structure of real world events and contribute to knowledge regarding those events. Lexical co-occurrence, event structure, and event knowledge are thus tightly linked, and it is unlikely that the reported results stem from lexical co-occurrence independent of event knowledge activation. This reasoning raises interesting questions regarding the interplay between knowledge of linguistic regularities and event knowledge during online language comprehension, which is an area worthy of future research.

The finding that event-related anomalous words elicit a reduced N400 stands in contrast with a previous study by Traxler, Foss, Seely, Kaup, and Morris (2000). In their study, the authors examined whether facilitated word processing during sentence comprehension might be captured by a schema-based account in which linguistic input activates a precompiled knowledge structure pertaining to the event being described. In Experiment 1, participants' eye movements were monitored as they read sentences such as *The* [*lumberjack/young man*] *chopped the axe early in the morning*. First fixation and gaze durations for the target word (*axe*) did not vary with the target's compatibility with the event denoted by the combination of the agent and verb (i.e., *axe* was read equally fast following *The young man chopped* as it was following *The lumberjack chopped*), suggesting that reading *The lumberjack chopped* did not activate a "lumberjacking" schema that contains an axe as a prototypical instrument.

Given the present finding, the result reported by Traxler et al. is quite surprising. According to the account put forth here, reading *the lumberjack chopped* should activate generalized "lumberjacking" knowledge that would likely include an axe. The source of this apparent discrepancy is unclear, although it is possible that the stimuli used here activated event knowledge more strongly than the stimuli used by Traxler et al. It is also possible that Traxler et al.'s stimuli included target words that were on average less strongly associated with the event being described, or it may simply be the case that such eye movement measures are not sensitive to the effect in question. Further examination is necessary to determine whether the findings reported by Traxler et al. do in fact stand in contrast with those reported here, or if they are due to methodological differences.

Conclusion

Event knowledge plays an important role in online language comprehension. The present study demonstrates that activation of event knowledge is not constrained by the linguistic context, but instead is highly general, including a variety of information that is not necessarily relevant to the processing of the current linguistic input. This finding provides further support for the intimate link between language comprehension and real-world experience.

Acknowledgments

Special thanks goes out to the members of the Kutas Cognitive Electrophysiology Laboratory at UCSD. This research was supported by NICHD grants HD053136 and HD022614, as well as NIA grant AG008313.

References

Altmann, G. T. M. & Mirković, J. (2009). Incrementality and prediction in human sentence processing. *Cognitive Science, 33*, 583-609.

Bicknell, K., Elman, J. L., Hare, M., McRae, K., & Kutas, M. (2008). Online expectations for verbal arguments conditional on event knowledge. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 2220-2225). Austin, TX: Cognitive Science Society.

Chwilla, D. J. & Kolk, H. H. J. (2005). Accessing world knowledge: Evidence from N400 and reaction time priming. *Cognitive Brain Research, (25)*, 589-606.

Ferretti, T., McRae, K., & Hatherell, A. (2001). Integrating verbs, situation schemas, and thematic role concepts. *Journal of Memory and Language, 44(4)*, 516-547.

Kamide, Y., Altmann, G. T. M., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence of anticipatory eye movements. *Journal of Memory and Language, 49(1)*, 133-156.

Kutas, M. & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature, 307*, 161-163.

Landuaer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to Latent Semantic Analysis. *Discourse Processes, 25*, 259-284.

McRae, K., Hare, E., Elman, J. L., & Ferretti, T. (2005). A basis for generating expectancies for verbs from nouns. *Memory & Cognition, 33(7)*, 1174-1184.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). The University of South Florida word association, rhyme, and word fragment norms. http://w3.usf.edu/FreeAssociation/.

Traxler, M. J., Foss, D. J., Seely, R. E., Kaup, B., & Morris, R. K. (2000). Priming in sentence processing: Intralexical spreading activation, schemas, and situation models. *Journal of Psycholinguistic Research, 29(6)*, 581-595.

Framed: Factors influencing reference frame choice in tabletop space

Laurie E. Robinette (ler6578@louisiana.edu)
Institute of Cognitive Science, University of Louisiana at Lafayette
Lafayette, LA 70504 USA

Michele I. Feist (feist@louisiana.edu)
Institute of Cognitive Science, University of Louisiana at Lafayette
Lafayette, LA 70504 USA

Michael L. Kalish (kalish@louisiana.edu)
Institute of Cognitive Science, University of Louisiana at Lafayette
Lafayette, LA 70504 USA

Abstract

While English speakers generally rely on a viewer-centered frame of reference when interpreting table-top space, they will also adopt an object-centered frame in certain situations—prompting the question: *What factors determine which frame?* The current research investigates two possible contributors: the intrinsic "frontedness" of a reference object involved in the scene and the syntactic structure of the sentence used to describe the scene. If an object possesses an "intrinsic front side," then this side should highlight the properties necessary for the object to be capable of having its own distinguishable perspective. Also, certain linguistic constructions may further increase the salience of the reference object's inherent geometrical properties, leading to greater use of an object-centered frame.

Keywords: Frame of reference; spatial language.

Introduction

English spatial terms can be ambiguous as to which area of space they refer if a frame of reference is not established before analyzing any spatial relation between two or more objects. When interpreting descriptions of table-top space, English speakers have been shown to rely primarily on a viewer-centered (VC) frame of reference (Majid, Bowerman, Kita, Haun, & Levinson, 2004); however, interpretation may alternatively depend upon an object-centered (OC) frame (Carlson-Radvansky & Irwin, 1993; 1994) when applicable.

The VC frame – also referred to by Levinson (1996, 2003) as the relative frame and by Miller and Johnson-Laird (1976), Retz-Schmidt (1988), and Carlson-Radvansky and Irwin (1993,1994) as the deictic frame – assigns spatial terms according to the properties of an observer located externally to the scene. For example, a viewer attempting to determine the location of a teacup with respect to a nearby teapot will transfer his or her own left and right sides onto the scene and judge that the teacup is to the left/right of the teapot if the space occupied by the teacup corresponds with the viewer's own left/right side. Because the VC frame is relative to an external observer, it can be based on different perspectives: one whose origin is grounded on "ego" (a speaker) and another whose origin has been transferred from "ego" to a third party (an addressee) (Retz-Schmidt, 1988; Levinson, 1996; 2003). If the speaker and the addressee share vantage points, locating one object with respect to another is relatively straightforward; however, if not, their viewpoints may conflict, with the result that spatial term use relying on the VC frame may be ambiguous.

Unlike the VC frame, an OC frame of reference – also referred to as the intrinsic frame by Miller and Johnson-Laird (1976), Retz-Schmidt (1988), and Levinson (1996, 2003) – assigns spatial terms according to the ground object's inherent properties. With this frame, a viewer attempting to locate the teacup would first determine whether or not the teapot has its own left and right sides and then judge that the teacup is to the left/right of the teapot if the teacup's occupied space corresponds with the teapot's left/right. Unlike the VC frame, the OC frame is not affected by the locations of any external observers; regardless of the viewpoints of the speaker and the addressee, the teacup will remain intrinsically to the teapot's left as long as neither object is moved. However, use of the OC frame will require mental rotation if the vantage points of the viewer and ground object are not aligned, and knowledge of the ground's orientation (Levelt, 1996). In addition, because many objects may lack inherent left and right sides, the assignment of *left* and *right* in this frame is ambiguous and is influenced by functional properties of the object (Levelt, 1996) as well as the vantage point from which the object is considered (Retz-Schmidt, 1988). Contrary to the findings of Majid et al. (2004), Miller and Johnson-Laird (1976) have argued that interpretations based on the OC/intrinsic frame actually dominate those based on the VC/deictic frame, with VC/deictic interpretations requiring specific qualifications from the speaker, such as "the teacup is to the left of the teapot *from my point of view*."

In order to better understand the semantics of projective terms, many of which can be used with either a VC or an OC frame of reference, we ask in this paper what factors

in a spatial scene determine which frame will be selected for use.

Possible Contributing Factors to Frame of Reference Selection

Vandeloise (1991) and Levinson (1996, 2003) suggest that one way to resolve the ambiguity of spatial terms may lie in the structure of the utterance used to describe the scene (see also Levelt, 1996). They argue that rephrasing "the teacup to the left of the teapot" as "the teacup to the teapot's left" should encourage use of an OC frame because the possessive construction points out that the teapot has its own "left side" that may be separate from the "left side" that a viewer assigns to the scene. Moreover, because this construction is specifically possessive, it may suggest that "the teapot's left side" is the correct interpretation.

In addition, because use of an OC frame makes more sense when the ground object possesses distinguishable sides (Levelt [1996] argues that the OC frame is only possible if this holds), this frame should be more salient when the object possesses a high degree of "frontedness." Landau and Jackendoff (1993) argue that the ground object's inherent axial structure is its most important property, and the more an object can be thought of as possessing a front side, the more viewers should notice that its two horizontal axes are different: one assigns an object's front and back while the other assigns its left and right. For example, a ground object like a teapot, which has an obvious front side, should encourage greater use of the OC frame because its front side calls attention to its possession of a perspective and orientation governed by its intrinsic front, back, left, and right sides. A ball, on the other hand, should not encourage use of an OC frame because it lacks an inherent front, and therefore lacks distinguishable sides; any sides assigned to it should be more strongly based on a VC perspective.

In the current study, we ask whether these two factors – the syntactic form of the spatial description and the inherent frontedness of the ground object – facilitate use of an OC frame of reference for English speakers' descriptions of spatial relations in tabletop space.

Method

Participants

Twenty-five students from the University of Louisiana at Lafayette who were enrolled in an introductory psychology course received extra credit in return for their participation.

Stimuli

The task took place on a computer using the E-Prime software package. The stimuli used in the experiment included photographs of a figure and ground taken at a "3/4" angle (halfway between head-on and bird's-eye).
Each scene was presented with a sentence including a locative expression (see Figure 1 for an example).



The black dot is to the left of the jack o' lantern.

Figure 1. Example of fronted object stimulus.

For the sentences, we considered the two locative terms *left* and *right*, which could be aligned with one of the horizontal axes. There were two levels of Sentence Structure: *non-possessive* and *possessive*. Participants either saw sentences of the form "The F is to the left/right of the G," (non-possessive) or the form "The F is to the G's left/right" (possessive). These two structures were presented between-participants to forestall a strategy of pairing each structure with a different reference frame.

The pictures each showed one figure - a black dot (a black-painted wooden circle) - paired with one of 6 different ground objects that varied on two dimensions of Frontedness, *fronted* and *non-fronted*. Stimuli in the *fronted* group included a camera, a flower, and a jack o' lantern; stimuli in the *non-fronted* group included a balloon, a glass, and a watermelon[1].

The final two variables were Figure Position (FP) and Ground Rotation (GR), which were manipulated to vary the frame of reference with which the picture-sentence pairs were consistent (VC, OC, VC and OC, or none (cf., Carlson-Radvansky & Irwin 1993; 1994). The design included 4 degrees of ground rotation (facing 0, 90, 180, or 270 degrees), and four figure positions (at a 0, 90, 180, or 270 degree arc). VC-consistent arrangements always included FP 270 for *left* and FP 90 for r*ight*, regardless of Ground Rotation; OC-consistent arrangements depend upon both figure position and ground rotation for their interpretation. Figure 2 shows VC-consistent and OC-consistent FP-GR pairings for *left* and *right* (with stimuli at GR 180 being consistent with both frames) illustrated

[1] Assignment to the fronted or non-fronted group was based on two norming studies. In the first, viewers rated the extent to which different objects were said to have an intrinsic "front side;" objects with low ratings were considered non-fronted, while objects with high ratings were considered fronted. In the second, viewers attempted to select the "front side" of the two types of objects. A chi-square analysis revealed that viewers chose the intended front side significantly more often than the other sides for the objects in the fronted group only.

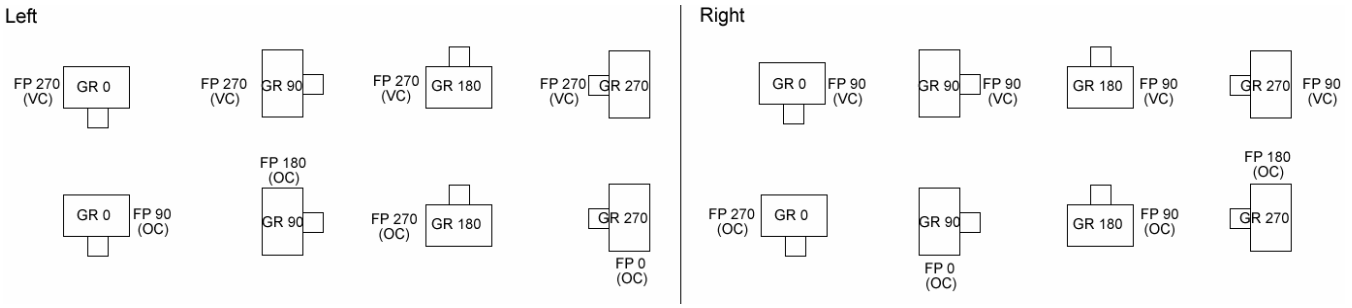


Figure 2. Viewer-centered (VC) and object-centered (OC) assignments of *left* and *right* for fronted objects. Non-fronted objects were assigned similarly to facilitate comparisons between object types.

for clarity with fronted ground objects. Each Figure Position-Ground Rotation combination was created for each of the Ground objects, for a total of 96 pictures.

However, because non-fronted objects cannot truly be said to face in a specific direction in a way that allows the different Ground Rotations—and ultimately, the OC frame of reference—to apply to them, attempting to compare the non-fronted objects to fronted objects becomes problematic. We resolved this issue by including non-fronted objects that possessed a pattern that allowed for their different sides to be discernable upon their rotation, and an arbitrary side was designated as the "front" side so that the objects could be said to "face" in the different directions of Ground Rotation. This designation also allowed for the object to possess "left" and "right" sides. Then, in order to make comparisons between non-fronted and fronted objects, we simply compared ratings at the same Figure Positions and Ground Rotations across object type.

The design of the experiment was 2 (Sentence Structure: *non-possessive* and *possessive*) X 2 (Spatial Term: *left* and *right*) X 2 (Frontedness: *non-fronted* and *fronted*) X 4 (Figure Position: *0, 90, 180, 270*) X 4 (Ground Rotation: *0, 90, 180, 270*). Spatial Term, Frontedness, Figure Position and Ground Rotation varied within participants, while Sentence Structure varied between participants.

Procedure

Participants were divided into two groups. One group saw arrangements with corresponding sentences in the *non-possessive* construction, while the other group saw arrangements with corresponding sentences in the *possessive* construction. For the first part of the experiment, participants looked at pictures of the ground objects (one picture per object) in order to introduce each object before the rating task began.

For the rating task, each of the 96 pictures was presented twice, once with a "left"-sentence and once with a "right"-sentence, in random order on a computer screen, for a total of 192 trials. In each case, the participant was asked to rate the acceptability of the sentence as a description of the picture, on a scale from 1 (not acceptable at all) to 5 (very acceptable). The variables for each trial were completely randomized.

Predictions

Acceptability of OC assignments should be higher when the spatial description is in the possessive structure ("the fork is to the knife's left") than when it is non-possessive ("the fork is to the left of the knife"), if awareness of the OC frame is made explicit by the possessive structure. Also, the inclusion of a fronted ground object should lead to higher acceptability of OC assignments than inclusion of a non-fronted object.

Alternatively, implicit awareness of the OC frame may lead to lower ratings of the VC assignments with a possessive structure or fronted object—which would suggest that viewers may at least recognize the possibility of using a different reference frame, even if they are not completely comfortable with it. Such a result might further suggest that the two frames share conceptual space and are simultaneously acceptable in a way that is similar to the predictions of the multiple frame activation hypothesis (Carlson-Radvansky & Irwin, 1994).

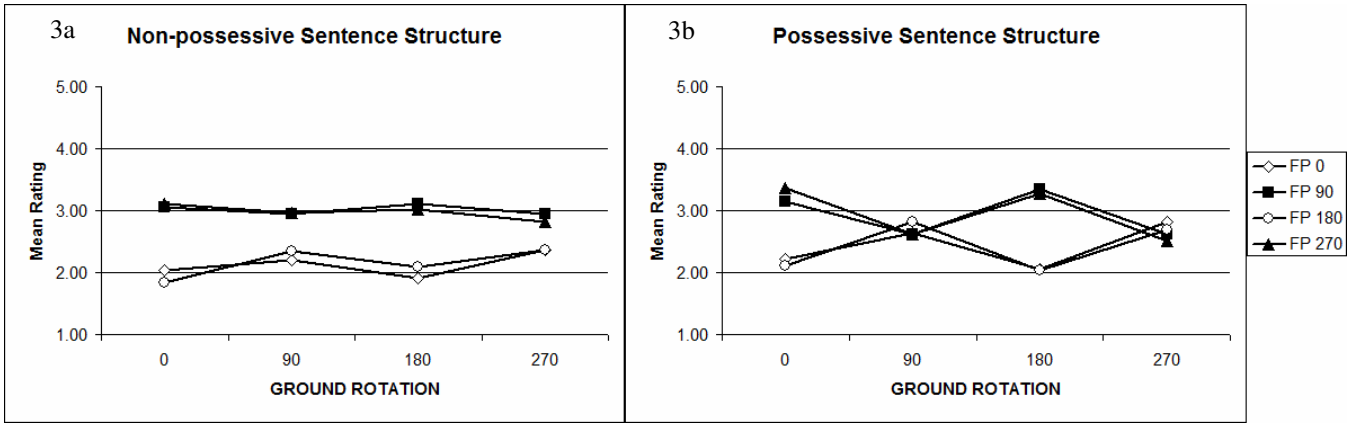
Furthermore, object frontedness and the structure of the spatial description should cooperate; when the possessive structure is combined with a fronted ground object, the structure of the description should call attention to the inherent frontedness of that object, maximally increasing acceptability of the OC frame. In this case, we would expect to see a situation in which acceptability of the OC frame surpasses that of the VC frame.

Results and Discussion

Because our interest is in how Sentence Structure and Frontedness might influence spatial term acceptability across the 16 figure position-ground rotation combinations, we will focus our discussion on higher-order interactions with the variables of figure position and ground rotation.

Sentence Structure

Figures 3a and 3b show that Sentence Structure influenced the pattern of acceptability across the figure position-ground rotation combinations, F (9, 207), =



Figures 3a – 3b. Mean sentence acceptability ratings for the 16 figure position/ground rotation arrangements broken down by Sentence Structure and collapsed across Term and Frontedness.

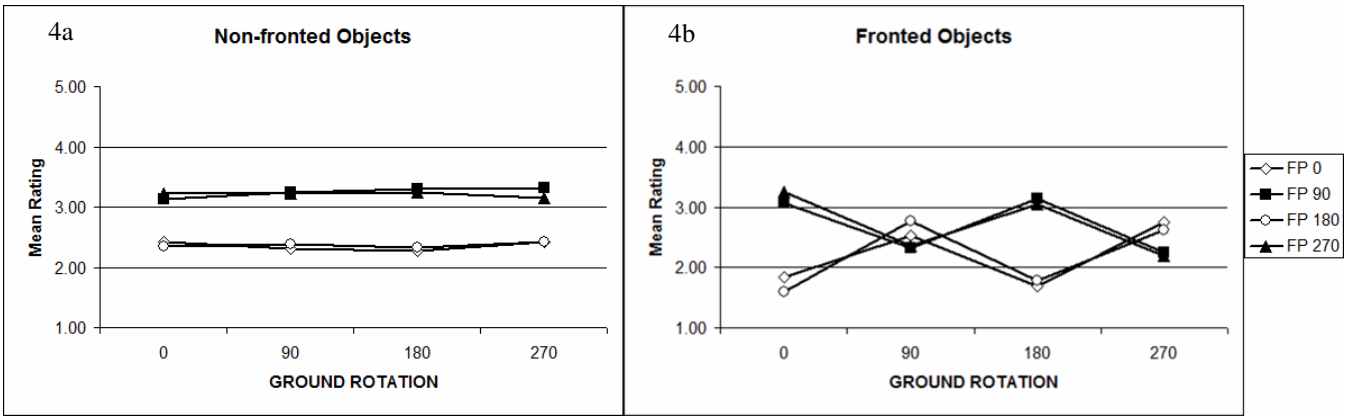
3.107, $p < .05$. At the two ground rotations where the VC frame was out of alignment with the OC frame, the average rating of all VC-consistent arrangements (FP 90-GR 90, FP 270-GR 90, FP 90-GR 270, FP 270-GR 270) (M = 2.922) was significantly higher than the average rating of all OC-consistent arrangements (FP 0-GR 90, FP 180-GR 90, FP 0-GR 270, FP 180-GR 270) (M = 2.314) for the non-possessive sentence structure, t (14) = 3.183, $p < .05$. However, these average ratings did not differ within the possessive condition (VC assignments, M = 2.593, vs OC assignments, M = 2.737, t (10) = -.641, *ns*). This effect is in line with the prediction that the possessive sentence structure may facilitate consideration of an OC frame of reference by increasing ratings of OC assignments and/or decreasing ratings of VC assignments to the point at which the two frames are equally acceptable.

Frontedness

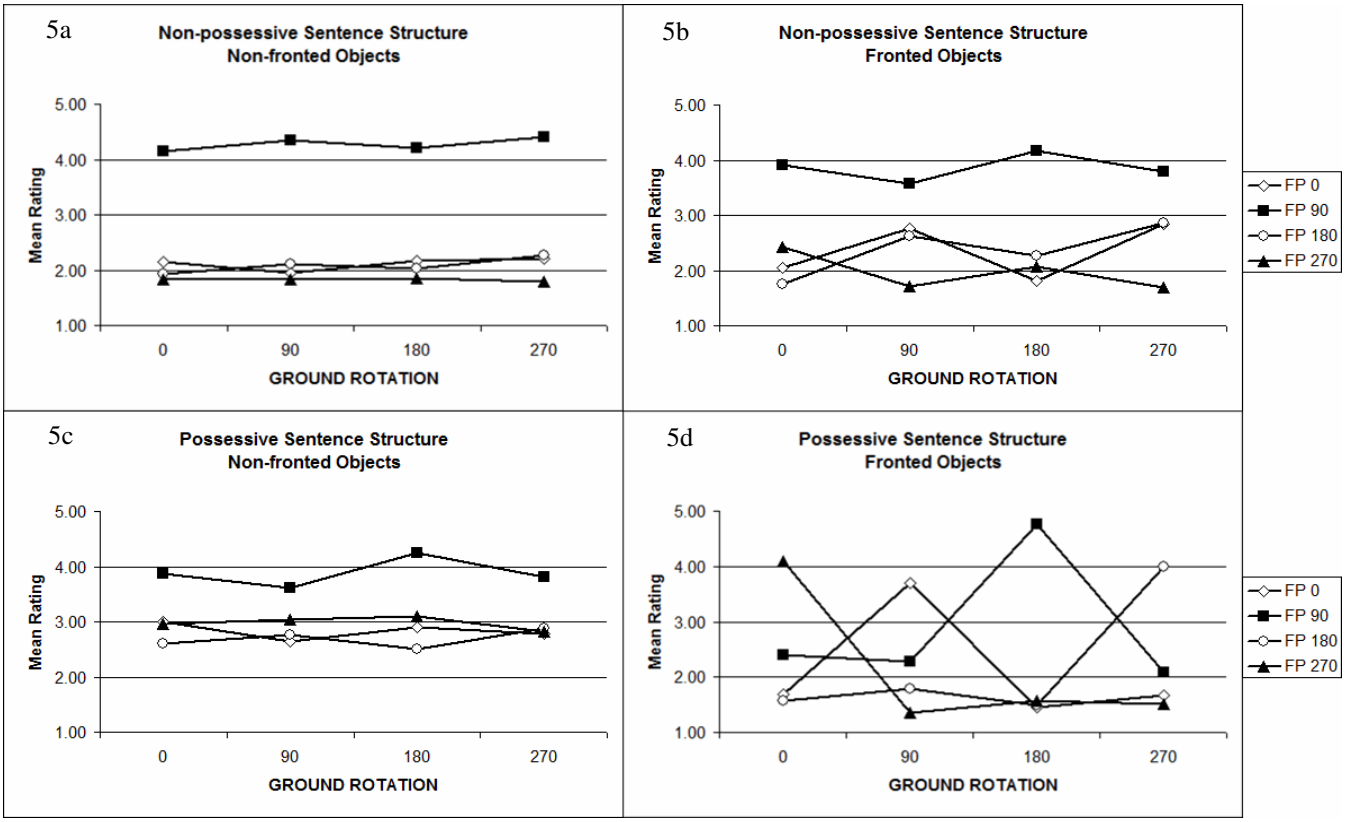
Figures 4a and 4b show that Frontedness influenced the pattern of acceptability across the figure position-ground rotation combinations, F (9, 207) = 13.555, $p < .05$, much as Sentence Structure did. For non-fronted objects, the average rating of all VC-consistent arrangements at the two ground rotations where the VC frame was out of alignment with the OC frame was significantly higher (M = 3.218) than the average rating of all OC-consistent arrangements (M = 2.342), t (24) = 5.054, $p < .05$; however, for the fronted objects, no difference was found (VC assignments, M = 2.336, vs OC assignments, M = 2.659, t (24) = -1.353, *ns*). As was the case for the possessive sentence structure, the inclusion of a fronted ground object appears to equalize the acceptability of a VC interpretation and the acceptability of an OC interpretation.

Sentence Structure and Frontedness

The combination of the influence of Sentence Structure and Frontednes is evident in the five-way interaction of Term, Sentence Structure, Frontedness, Figure Position and Ground Rotation; F (9, 207) = 5.444, $p < .05$. For the sake of brevity, and because results for the terms *left* and



Figures 4a – 4b. Mean sentence acceptability ratings for the 16 figure position/ground rotation arrangements broken down by Frontedness and collapsed across Term and Sentence Structure.



Figures 5a – 5d. Mean sentence acceptability ratings for the 16 figure position/ground rotation arrangements broken down by Sentence Structure and Frontedness for the term *right*.

right were similar, here we only describe the analysis of *right* (see Figures 5a – 5d).

The Viewer-centered frame. In order to further understand the way in which the variables of Sentence Structure and Frontedness influenced the acceptability of *right* for arrangements consistent with a VC frame of reference (FP 90), we examined changes in acceptability ratings across the different levels of Sentence Structure and Frontedness[2]. Looking at the acceptability ratings for these points across the 4 conditions (Figures 5a – 5d), we observe that VC points are rated as most acceptable for the non-possessive, non-fronted condition (5a) and for the possessive, non-fronted condition (5c), with lower acceptability in the non-possessive, fronted condition (5b), and with lowest acceptability in the possessive, fronted condition (5d). These differences in acceptability are significant (non-possessive, non-fronted vs. possessive, non-fronted, M = 3.767, t (23) = 1.326, *ns*; non-possessive, non-fronted, M = 4.301, vs. non-possessive, fronted, M = 3.754, t (13) = 2.508, $p < .05$ one-tailed; non-possessive, non-fronted vs. possessive, fronted, M = 2.252, t (23) = 4.463, $p < .05$; non-possessive, fronted vs. possessive, fronted, t (23) = 3.040, $p < .05$).

The Object-centered frame. In order to further understand how the variables of Sentence Structure and Frontedness influenced the acceptability of *right* for arrangements consistent with an OC frame of reference (FP 270-GR 0, FP 0-GR 90, FP 180-GR 270) (see Figure 2 for representations of these arrangements), we examined changes in acceptability ratings across the different levels of Sentence Structure and Frontedness for these arrangements. Looking at acceptability ratings across the different conditions reveals that OC points received the highest ratings in the possessive, fronted condition (Figure 5d). Ratings were lower in the non-possessive, fronted (5b) and possessive, non-fronted conditions (5c) and lowest in the non-possessive, non-fronted condition (5a). These differences in acceptability are significant (possessive, fronted vs. non-possessive, fronted, M = 2.683, t (23) = -2.421, $p < .05$; possessive, fronted, M = 3.930 vs. possessive, non-fronted, M = 2.828, t (10) = -2.366, $p < .05$; non-possessive, fronted vs. possessive, non-fronted, t (23) = -.296, *ns*; possessive, non-fronted, vs. non-possessive, non-fronted, M = 2.015, t (23) = -2.184, $p < .05$).

[2] For this and following analyses, we excluded data for GR 180, as at this ground rotation the VC and OC frames are in alignment.

Comparing the two types of reference frame. To test our prediction that the combination of a possessive sentence structure and fronted object would create a situation in which OC assignments would be rated as more acceptable than VC assignments—and that this effect would be unique to this combination—we compared ratings of OC assignments to ratings of VC assignments in each condition. For the non-possessive, non-fronted condition, the average rating of VC arrangements (M = 4.301) was higher than the average rating of OC arrangements (M = 2.015), F (1, 13) = 22.718, $p < .05$. For the possessive, fronted condition, the average rating of the OC arrangements (M = 3.930) was significantly higher than the average rating of the VC arrangements (M = 2.252), F (1, 10) = 6.698, $p < .05$. Average ratings of the VC arrangements and the OC arrangements did not differ for either of the remaining conditions.

Considered individually, both the possessive sentence structure and the fronted objects appear to raise the salience of the OC frame of reference as evidenced by an increase in acceptability ratings for OC-consistent arrangements and/or a decrease in acceptability ratings for competing VC-consistent arrangements. Figures 3 and 4 show this effect. However, the lack of difference between average ratings of the VC and OC assignments suggest that the simple act of incorporating a possessive sentence structure or a fronted object may only cause the OC frame to be as acceptable as the VC frame. In contrast, the combination of a possessive sentence structure and fronted ground object both decreases acceptability of VC assignments and increases acceptability of OC assignments to the point in which English speakers prefer an OC assignment to a VC assignment (at least when these assignments are in competition).

Conclusions

In this paper, we examined two factors that may influence how a reference frame is selected for a spatial description. Taken together, the results from this study provide more insight into the nature of viewers' consideration of the VC and OC frames of reference. When the non-possessive sentence structure is used to describe a scene in which a non-fronted object serves as the ground, viewers prefer a VC assignment over an OC assignment. When either a possessive structure or a fronted object is introduced, VC and OC assignments appear equally appropriate. Finally, when both a possessive structure and a fronted object are introduced, a preference for OC assignments over VC assignments arises. These results support our predictions that, as Levinson (1996, 2003) argues, VC assignments are the default for English speakers, but consideration of other assignments may increase when certain elements of the situation are changed in order to call attention to the ground object's inherent features (Carlson-Radvansky & Irwin, 1993, 1994). The inclusion of a possessive sentence structure and/or a fronted object appears to do just that. When a fronted object serves as the point of reference, the asymmetries associated with this type of object's axial structure (Landau & Jackendoff, 1993) may point out to the viewer that this object might have its own perspective, different from that of the viewer, which can also be used to assign space. Additionally, in support of Vandeloise's (1996), Levinson's (1996), and Levelt's (1996) claims, the use of a possessive sentence structure to describe the scene may also highlight any asymmetries associated with the ground object and similarly cause viewers to notice potentially competing perspectives. However, neither of these factors alone leads to a preference for one type of reference frame over the other. Rather, the inclusion of either factor on its own only seems to equalize the acceptability of the two reference frames, while preference for an OC assignment appears when there is a combination of a fronted ground object and possessive sentence structure.

Acknowledgments

We would like to thank the members of the Language and Cognition Lab at UL Lafayette for their comments.

References

Carlson-Radvansky, L.A. & Irwin, D.E. (1993). Frames of reference in vision and language: Where is *above? Cognition, 46,* 223-244.

Carlson-Radvansky, L.A. & Irwin, D.E. (1994). Reference frame activation during spatial term assignment. *Journal of Memory and Language, 33*, 646-671.

Landau, B. & Jackendoff, R. (1993). "What" and "where" in spatial language and spatial cognition. *Behavioral and Brain Sciences, 16*, 217-265.

Levelt, W. J. M. (1996). Perspective taking and ellipsis in spatial descriptions. In Bloom, P., Peterson, M.A., Nadel, L., & Garrett, M.F. (Eds.) *Language and space* (pp. 77-107). Cambridge, MA: The MIT Press.

Levinson, S.C. (1996). Frames of reference and Molyneaux's question. In Bloom, P., Peterson, M.A., Nadel, L., & Garrett, M.F. (Eds.) *Language and space* (pp. 109-170). Cambridge, MA: The MIT Press.

Levinson, S.C. (2003). Space in language and cognition: Explorations in cognitive diversity. Cambridge: Cambridge University Press.

Majid, A., Bowerman, M., Kita, S., Haun, D., & Levinson, S. (2004). Can language restructure cognition? The case for space. *Trends in Cognitive Science, 8,* 108-114.

Miller, G.A. & Johnson-Laird, P.N. (1976). *Language and perception*. Oxford, England: Harvard University Press.

Retz-Schmidt, G. (1988). Various views on spatial prepositions. *AI Magazine, 9* (2), 95-105.

Vandeloise, C. (1991). *Spatial prepositions: A case study from French.* (A.R.K. Bosch, Trans.) Chicago: University of Chicago Press.

Sentence Production in Naturalistic Scenes with Referential Ambiguity

Moreno I. Coco (M.I.Coco@sms.ed.ac.uk) and
Frank Keller (keller@inf.ed.ac.uk)
School of Informatics, University of Edinburgh
10 Crichton Street, Edinburgh EH8 9AB, UK

Abstract

Language production often happens in a visual context, for example when a speaker describes a picture. This raises the question whether visual factors interact with conceptual factors during linguistic encoding. To address this question, we present an eye-tracking experiment that manipulates visual clutter (density of objects in the scene) and animacy in a sentence production task using naturalistic, referentially ambiguous scenes. We found that clutter leads to more fixations on target objects before they are mentioned, contrary to results for visual search, and that this effect is modulated by animacy. We also tested the eye-voice span hypothesis (objects are fixated before they are mentioned), and found that a significantly more complex pattern obtains in naturalistic, referentially ambiguous scenes.

Keywords: language production; eye-tracking; naturalistic scenes; eye-voice span; referential ambiguity.

Introduction

Language production often happens in a visual context, for example when the speaker describes a picture, gives directions on a map, or explains the function of an artifact. In these situations, the speaker needs to select which objects to talk about, and in which order. He/she also needs to disambiguate the utterance referentially. For instance, if there are multiple clipboards in the visual context, then the speaker has to encode additional visual information to pick out one of them uniquely (e.g., *the brown clipboard* or *the clipboard on the table*).

Most work in psycholinguistics has dealt with isolated sentences, but there is some existing research investigating how language is processed in a visual context. A prominent line of research employs the visual world paradigm (VWP; Tanenhaus et al. 1995; Altmann and Kamide 1999) for this purpose. In a typical VWP study, participants' eye-movements are recorded while they view a visual scene and listen to a sentence at the same time. Some VWP experiments have investigated language production; the most well-known example is Griffin and Bock's (2000) study, in which participants were asked to describe line drawings depicting two objects (e.g., a turtle and a kangaroo) performing a transitive event (e.g., splashing). The key finding of this study was that speakers fixate visual referents in the order in which they are mentioned, and they begin fixating an object about 900 ms before naming it. The span between fixating and naming a referent is known as the **eye-voice span**; other studies (e.g., Qu and Chai 2008) have reported eye-voice spans consistent with those found by Griffin and Bock (2000).

The aim of the present paper is to establish whether the simple relationship between language production and eye-movements implied by the eye-voice span extends to more realistic situations. We investigate language production in a visual context that consists of naturalistic scenes (rather than line drawings) and in which multiple objects can correspond to a given linguistic referent (in contrast to Griffin and Bock 2000). This enables us to study how scene complexity and referential ambiguity affect the eye-voice span. Furthermore, we are interested in the interaction of visual and conceptual factors during linguistic encoding. The visual factor we focus on is clutter (density of objects in the scene); **clutter** has been investigated in the visual processing literature and found to affect visual search (Henderson et al., 2009). The conceptual factor we investigate is the **animacy** of the referent; animacy has been manipulated in the psycholinguistic literature and found to affect sentence production (Branigan et al., 2008). Here, we address the question whether these two factors representing different modalities contribute independently to the formation of reference in sentence production, or whether they interact.

Background

The recent visual cognition literature has emphasized the importance of contextual information for visual processing. For example, prior information about object categories facilitates visual search (Malcolm and Henderson, 2009; Schmidt and Zelinksy, 2009). This effect occurs if participants are asked to look for an object embedded in a scene or an object array (Brockmole and Henderson, 2006), or if categorical templates are provided which the visual system can use to determine where the target object is located (Vo and Henderson, 2010). It seems likely that similar contextual guidance effects (Torralba et al., 2006) also occur if the context is provided by another modality, e.g., by the linguistic material involved in a language production task.

In such task, speakers will often be faced with referential ambiguity, which they resolve by including disambiguating material in a sentence. For example, spatial prepositions can be used to locate an object in relation to the surrounding space, e.g., *the clipboard on the table* or adjectives can be used to contrast the intended referent with a competitor, e.g., *the brown clipboard*. Before any linguistic encoding can take place, however, the disambiguation has to happen at the visual level. When a target object is selected as a referent (because it will be mentioned in a sentence), the visual system has to retrieve scene and object information that can be used to refer to the object unambiguously. One can therefore hypothesize that if participants are faced with a linguistic task (e.g., scene description), then contextual guidance is afforded not only by visual information, but also driven by linguistic processing and the need to disambiguate.

Experiment

In this experiment, we investigated how visual attention is influenced by contextual factors during sentence production. Participants had to describe a visual scene after being prompted with a cue word. This cue word was ambiguous, i.e., two objects in the scene could be referred to by the cue. We manipulated the animacy of the cue (e.g., *man* vs. *clipboard*), expecting an effect on both linguistic encoding and visual attention. Animate objects are associated with a larger number of conceptual structures in encoding (Branigan et al., 2008); we should therefore observe more sentences containing action information in this case (e.g., *the man is reading a letter*). At the same time, we expect visual attention to be localized on animate targets, an effect that has already been demonstrated in visual search (Fletcher-Watson et al., 2008).

The second experimental manipulation concerned a visual factor, viz., clutter, defined as the density of visual information (Rosenholtz et al., 2007). Again, this is a factor that has shown effects on the performance and accuracy of visual search: the more cluttered the scene is, the less efficient the identification of target object (Henderson et al., 2009). In a language production task, however, the effect of clutter can be expected to change, due to the disambiguation strategies required. Clutter could have a beneficial effect: the more visual information there is, the more disambiguating material can be retrieved; clutter could therefore facilitate language production.

Finally, this experiment makes it possible to investigate the effect of referential ambiguity on the eye-voice span. In previous work, the relationship between linguistic and visual referents was unambiguous: looks to the visual referent always preceded naming (Griffin and Bock, 2000) and this trend exponentially increases towards the mention (Qu and Chai, 2008). In our setting, we expect a more complex gaze-to-name relationship caused by a process of visual disambiguation that arises both before and after the intended referent is mentioned.

Method

We used a factorial design that crossed the two factors *Clutter* (Minimal/Cluttered) and *Cue* (Animate/Inanimate). Participants' eye-movements were recorded while they described photo-realistic scenes after being prompted with a cue word, which ambiguously corresponded to two visual referents in the scene (see Figure 1).

We created 24 experimental items using photo-realistic scenes drawn from six indoor scenarios (e.g., Bathroom, Bedroom; four scenes per scenario). In each scene, we inserted two animate and two inanimate objects using Photoshop, which correspond to the two *Cue* conditions; *Clutter* was either added or removed.

Twenty-four native speakers of English, all students of the University of Edinburgh, were each paid five pounds for taking part in the experiment. They each saw 24 items randomized and distributed in a Latin square design that made sure that each participant only saw one condition per scene.



Figure 1: Example of an experimental trial, with visual region of interest considered for analysis. PRIMARY indicates that the ANIMATE and INANIMATE visual objects are spatially close and semantically connected (e.g., the MAN is doing an action using the CLIPBOARD). SECONDARY is used to indicate the remaining referent of the ambiguous pair. BACKGROUND and CLUTTER are defined in opposition: BACKGROUND is everything other than CLUTTER.

An EyeLink II head-mounted eye-tracker was used to monitor participants' eye-movements with a sampling rate of 500 Hz. Images were presented on a 21" multiscan monitor at a resolution of 1024 x 768 pixels; participants' speech was recorded with a lapel microphone. Only the dominant eye was tracked. A cue word appeared for 750 ms at the center of the screen, after which the scene followed and sound recording was activated. Drift correction was performed at the beginning and between each trial. There was no time limit for the trial duration and to pass to the next trail participants pressed a button on the response pad. The experimental task was explained using written instructions and took approximately 30 minutes to complete.

Data Analysis

We defined regions of interest (ROIs) both for the visual and the linguistic data. The visual data was aggregated into six different regions: PRIMARY and SECONDARY ANIMATE, PRIMARY and SECONDARY INANIMATE, BACKGROUND, and CLUTTER (see Figure 1).

For the linguistic data, we made a general division between time windows *Before* and *During* production. This allows us to capture the overall trend of the two main phases of a trial. For the analysis of eye-voice span, we consider a window of 2000 ms before the referent was mentioned, similar to Qu and Chai 2008. The resolution of visual ambiguity is analyzed using a window of 1600 ms (divided into 40 time slices 40 ms each): 800 ms before and after the mention of *Cue*. This makes it possible to explore how the linguistic referent is visually located before being mentioned and just after.

In order to unambiguously analyze fixated and named referents, we aggregate eye-movements responses in four blocks (Primary, Secondary, Ambiguous and Both) by manually

checking which referent was mentioned in each sentence.[1] We introduced referential ambiguity as predictor in the inferential model described below to investigate how looks to the mentioned object differ from those to its competitor. For reason of space, we only present the analysis for the *Primary* objects mentioned. The effect of mention on eye-movements' pattern is evaluated by comparing Primary with Secondary objects.

As an initial exploration of our data, we investigate the overall trend of fixations *Before* and *During* production. Production is a task with large between-participant variability, e.g., one participant will spend 2000 ms *Before* and 1000 ms *During* production, whereas another one will show the opposite pattern. Normalizing the production data is therefore crucial, in particular as we want to interpret eye-movements in relation to phases of linguistic processing. We normalize each sequence S^i_{old} of eye-movements by mapping it onto a normalized time-course of fixed length S^i_{new}. The length of S^i_{new} is set on the basis of the shortest eye-movement sequence $\min_i[\text{length}(S^i_{old})]$ found between Before and During production, across all participants.[2] For each sequence S^i_{old}, we obtain the number of old time-points k^i corresponding to a new time-unit u, as $k^i = \text{length}(S^i_{old})/\text{length}(S^i_{new})$. Proportions are then calculated over k^i old time-points and subsequently mapped into the corresponding unit u of the normalized time-course. In the Results section, we show plots of normalized proportions for *Primary* and *Secondary* (Animate and Inanimate) across conditions, *Before* and *During* production.

To explore the eye-voice span hypothesis, we compute the number of fixations to the mentioned object compared to the competitor. We also look at latencies, i.e., the onset of the last fixation to the referent or competitor before the mention, and gaze duration as a function of latencies, i.e., the time spent looking at the referent or competitor for the different latencies.

We also report inferential statistics for the referent region (for the time windows previously described). The dependent measure is the empirical logit (Barr, 2008), calculated as $emplog = \ln \frac{0.5+\phi}{0.5+(1-\phi)}$, where ϕ is the number of fixations on the region of interest. The analysis is performed using the framework of linear-mixed effect (LME) models as implemented by the R-package lme4 (Baayen et al., 2008). The predictors included were *Animacy*, *Clutter*, *Time* and *Object*. The random factors were *Participant* and *Item*. To reduce colinearity, factors were centered.

The model selection followed a conservative stepwise forward procedure that tests model fit based on a log-likelihood test comparing models each time a new parameter is included. If the fit improves, we accept the new model, otherwise we keep the old one. We include predictors, random intercepts and slopes ordered by their log-likelihood impact on model fit. We iterate until there is no more improvement on the fit; leaving us with the best model. In the result section, we show plots of the values predicted by the model for each condition.

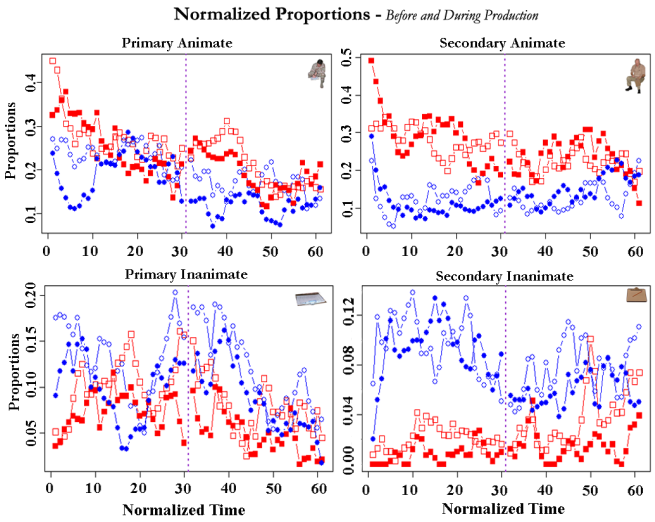


Figure 2: Normalized proportions of looks (60 bins) across the four conditions, Before and During production, for the different visual ROIs. The purple dashed vertical line indicates Before (to the left) and During (to the right) production. The four conditions are coded as following: *Animate/Cluttered*: red, full-square; *Animate/Minimal*: red, empty square); *Inanimate/Cluttered*: blue, full circle; *Inanimate/Minimal*: blue, empty circle

Results and Discussion

Before and During Production We first look at how fixations are distributed when we collapse the two main phases of the experiment: *Before* and *During* production. This analysis does not distinguish whether the *Primary* or *Secondary* referent was mentioned. Figure 2 shows normalized proportions of looks on the competitor visual objects corresponding to the *Cue* (Animate/Inanimate).

The first thing to note is that for the visual ROI corresponding to the *Primary* referent, the pattern of fixations is more complex than for the ROI of the *Secondary* referent. The spatial proximity and semantic relatedness of the two *Primary* referents result in a more complex pattern of interaction. The clearest effect is found in relation with the animacy of *Cue*; we observe more fixations to the animate referent when the cue is also animate. When looking at the *Primary* ROI, the effect is seen at the beginning of both the Before and the During region. At the beginning of the trial, the visual system retrieves information about the cued objects; when production starts, the referents are fixated again, probably before being mentioned. For the *Secondary* ROIs, the relation with the *Cue* is stronger, probably reinforced by the referential competition. Moreover, the pattern of looks is much clearer than for the *Primary* ROI. This confirms that spatial proximity and

[1] PRIMARY means that the Primary Animate or Inanimate is mentioned (e.g., *The man is writing on the clipboard*). SECONDARY is used when the Secondary Animate or Inanimate is mentioned (e.g., *The man is reading a letter*). AMBIGUOUS is used when is unclear which one is referred to (e.g., *the man is sitting on the couch*). BOTH indicates that both referents are mentioned (e.g., *the man is writing on a clipboard while the other man reads a newspaper*).

[2] We remove outliers that are two standard deviation away from the mean, after having log-transformed our data. The data are not normally distributed, due to right skewness. The log-transformation helps us to reduce the skew.

Table 1: Eye-voice span statistics. *Excluding* indicates that the percentage is calculated considering only those cases in which either the referent or competitor have been fixated, *Including* takes into account also cases where both have been fixated.

Measure		Referent	Competitor
Percentage of looks	Including	71.65	43.30
	Excluding	36.44	8.09
Mean Latency	Including	1032 ms	1203 ms
	Excluding	1012 ms	1325 ms
Gaze Duration	Including	489 ms	432 ms
	Excluding	568 ms	623 ms

semantic relatedness increase the interaction between visual referents. *Clutter* does not have a strong effect, though there is a small increase of looks when the scene is minimal and the animacy of the target matches that of the cue.

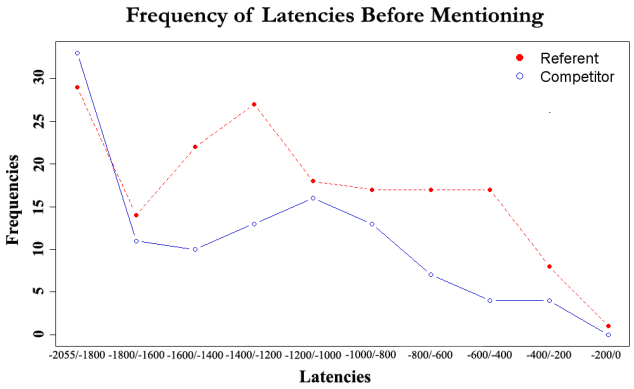
Eye-Voice Span We analyzed eye-voice span to investigate the gaze-to-name relation for the mentioned referent and its competitor. Table 1 shows percentages of looks to referent or competitor with mean latencies and gaze durations.[3]

There is a preference for looks to the referent over looks to the competitor, with a latency of about one second, confirming previous findings (Griffin and Bock, 2000). In a minority of cases, participants only look at the referent (36.44%); competition between the two ambiguous visual referents is the norm (71.65%). Moreover, we notice that the competitor is fixated earlier than the referent and the duration is shorter for the Including condition (which includes trials in which both referents have been fixated). This may indicate that the final decision on which referent is mentioned is made after discarding the competitor.

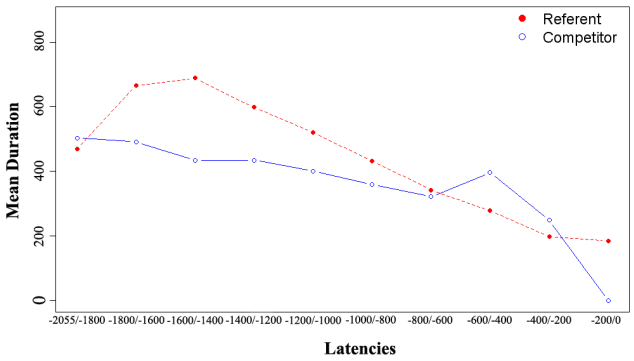
Figure 3(a) shows frequencies of *Latencies* at different temporal blocks (200 ms each) within a total window of two seconds. We find that latency frequency decreases towards the mention for both the referent and the competitor. This finding contrasts with Qu and Chai (2008) who found the opposite trend, i.e., the closer to the mention, the more gazes are associated with the referent object. Note also that this effect cannot only be due to the presence of a competitor, e.g., comparative looks before mention, as these present a similar decreasing trend.

In Figure 3(b) we show mean gaze duration as a function of the different latencies. Again, a decreasing trend is clearly visible: the closer the latency to the mention, the shorter the gaze duration. Interestingly there is a peak of gaze duration at 1600/1400 ms. The higher duration found at this latency might be an indicator of referential selection (gaze-to-name binding). We also find evidence of competition at 600/400 ms, where the competitor receives longer gazes compared to referent. A last visual check on the competitor is probably performed before referentiality is encoded linguistically.

[3]The measures are calculated only when the Primary and Secondary referent are mentioned; thus, we exclude the Both and Ambiguous cases, for which it was not possible to establish unambiguous eye-voice span relation.



(a) Frequencies of latencies at different temporal blocks (from two seconds to mention): red is the referent, blue the competitor. The latency measures the time elapsed from the beginning of the last fixation to the object (referent or competitor) until is mentioned.



(b) Mean gaze duration as a function of latency. The mean of gaze duration is calculated for the different blocks of latencies. We analyze only cases where gaze duration is shorter than latency, thus avoiding cases where fixations spill over into the region after mention.

Figure 3: Eye Voice Span statistics.

Inferential Analysis We now analyze the pattern of eye-movements before and after the mention of the cue word. To save space, we focus on the case where the Primary visual object is mentioned. Based on the eye-voice span analysis, we expect to find a decreasing trend of looks before the referent is mentioned, and the presence of competition should weaken the gaze-to-name relationship.

Recall that our experiment had two factors (Cue: animate/inanimate; Clutter: minimal/cluttered); we also include the object fixated (Object: primary/secondary) and Time (in 40 ms slices, see Data Analysis above) in the analysis. Figure 4 plots LME predicted values for the four conditions, Before and After mention.[4]

Beginning with the animate visual objects in Figure 4, we expect the *Primary Animate* to receive more looks than the *Secondary Animate*, and the number of looks should increase. We observe a preference for looks to Primary Animate,

[4]The intercepts for Before and During are different because they are calculated over distinct time intervals.

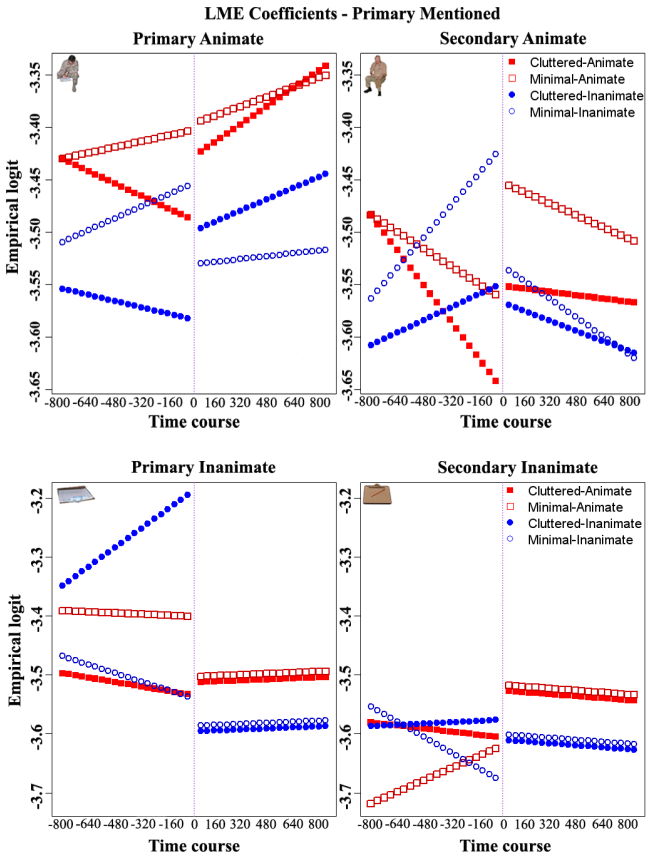


Figure 4: Linear mixed effect model: plot of predicted values (40 windows of 40 ms each) across the four conditions, Before and After referent, on the different visual ROIs. The *First* referent is mentioned and the dashed line indicates when.

but the difference is not statistically significant ($\beta_{Primary} = 0.0255; p > 0.1$). However, we find a main effect of *Cue* ($\beta_{Animate} = 0.0543; p < 0.01$): an animate cue facilitates looks to Animate visual objects. When looking at the time course, we find a general decreasing trend ($\beta_{Primary:Time} = -0.022; p < 0.01$), partly compensated by a three-way interaction of *Animacy*, *Object*, and *Time* ($\beta_{Animate:Primary:Time} = 0.049; p < 0.001$). Moreover, we observe a two-way interaction of *Clutter* and *Time* ($\beta_{Minimal:Time} = 0.024; p < 0.01$): a minimal scene makes it difficult to retrieve disambiguating information for the animate referent, forcing the visual system to look for this information on the referent itself. It is also conceivable that the minimality of the scene makes visual responses similar to those found for line drawings (Griffin and Bock, 2000); thereby explaining the increasing trend. In a cluttered environment, instead, there are more ways to relate the referent to the surrounding context, hence helping language production to disambiguate. This explains the decreasing trend of fixations on the referent in the cluttered condition.

After mention, we observe interactions of *Cue* with *Clutter* ($\beta_{Animate:Minimal} = 0.0165; p < 0.001$) and *Object* ($\beta_{Animate:Primary} = 0.017; p < 0.01$), confirming both the facilitation of the cued referent and the preference for referent information when scenes are minimal. In contrasts with previous findings, we observe increasing looks to the referent after mention ($\beta_{Primary:Time} = 0.0530; p < 0.001$). This effect could be due to referential ambiguity: the visual system is connecting disambiguating material retrieved before mention to the referent just uttered. For the *Secondary* Animate, we find an increasing trend of looks when *Cue* is Inanimate and especially for minimal scenes ($\beta_{Inanimate:Time} = 0.056, p < 0.001; \beta_{Minimal:Time} = 0.041, p < 0.01$). The minimality of the scene gives prominence to animate referents; probably the spatial and semantic proximity of one of Primary Inanimate and the Primary Animate also trigger comparative looks to Secondary Animate, i.e., participants check whether it can also be contextually related to the cue.

After the referent is mentioned (*Primary* in this case), looks to the *Secondary* Animate decrease over time in all conditions. Competition is triggered by visual ambiguity, but once the association of the visual with the linguistic referent has been established (i.e., after the mention), participants look back to the referent mentioned, presumably finalizing the choice made.

Looking at inanimate referents in Figure 4, we observe a statistically significant preference for looks to the Primary Inanimate ($\beta_{Primary} = 0.0621; p < 0.05$). This preference could be due to the spatial proximity and the semantic relation with the primary animate, which makes the primary inanimate more likely to be encoded either as a direct object or as subject of the description. As a consequence, we find an interaction with the animacy of the *Cue* ($\beta_{Animate:Primary} = 0.0155; p < 0.05$) but not a main effect ($\beta_{Inanimate} = 0.017; p > 0.1$). In contrast with standard visual search task, where performance degrades as a function of clutter, here we observe instead a positive interaction of *Clutter* and *Cue* on the target ($\beta_{Inanimate:Cluttered:Primary} = 0.028, p < 0.001$), which increase over time ($\beta_{Cluttered:Time} = 0.054, p < 0.01$). The visual system is not performing a search task, rather it is sourcing information to ground language processing. In a cluttered scene, an inanimate referent could be spatially related to many other different objects, whereas a minimal scene has fewer points to anchor the referent. The visual system therefore needs to select among the different spatial relations to find one that optimally situates the object within the contextual information.

For the secondary inanimate, there is a negative relationship between the animacy of *Cue* and the minimality of *Clutter* ($\beta_{Animate:Minimal:Secondary} = -0.0719; p < 0.001$); the proximity and relatedness of the primary inanimate and the primary animate is highlighted when visual information is minimal, which results in the secondary inanimate being fixated less.

General Discussion

Referential ambiguity is a common phenomenon in everyday experience. In a naturalistic scene, the same object (e.g., a clipboard) can occur multiple times (e.g., on a desk or on a counter). This fact turns into linguistic ambiguity when a referent has to be selected from the set of visual competi-

tors. Typically, referential ambiguity is resolved by encoding sufficient contextual information to discriminate the intended referent from competitors (e.g., *the clipboard on the desk*). However, this process of ambiguity resolution cannot be explained by linguistic factors alone, especially given that the disambiguating material needs to be selected by the visual system prior to any encoding. We therefore hypothesized that visual factors interact with well-established conceptual factors active during language production.

We reported the results of an eye-tracking language scene description experiment that support this hypothesis. We explored how the conceptual properties of the target referent (factor *Cue*: animate/inanimate) and the density of visual information (factor *Clutter*: minimal/cluttered) interact during the resolution of referential ambiguity. The results showed that the animacy of the cue facilitates looks to animate objects, especially at the beginning of two main phases of linguistic production: before and during the mention of the referent. The data indicate that a visual search is performed to localize the objects matching the cue word (Malcolm and Henderson, 2009). Our results also contrasted interestingly with findings for visual search, where clutter decreases search performance (Henderson et al., 2009). In cases in which an animate referent is mentioned, we found that there were fewer fixations to the target object in the cluttered condition compared to the uncluttered one. In other words, clutter makes language production easier, not harder: the visual system is not just searching for the target object, but it is also retrieving visual information that can be used to linguistically anchor it (e.g., for disambiguation). The more clutter there is, the easier this process becomes, explaining the reduced number of fixations in the cluttered condition.

Turning at the relation between fixating and naming an object (the eye-voice span), previous work found that referents are fixed shortly before being mentioned (Griffin and Bock, 2000). It has also been observed that fixation probability increases with decreasing distance to the mention (Qu and Chai, 2008). In our data, we found a numerical preference for looks to the mentioned referent over looks to the competitor, but this preference was not confirmed in the inferential analysis (see Figure 4). Only if the primary inanimate was mentioned, it was fixated significantly more than the secondary inanimate. This preference is likely due to the proximity, spatial and semantic, between the primary animate and inanimate. Moreover, we found that fixation probability decreased with decreasing distance to the mention, contrary to previous results, in particular when the scene was cluttered. The competition between visual referents seems to override the standard eye-voice span effect. Interestingly, we also observed an increasing trend of fixation to the referent object *after* its mention. Once production has started, the visual system needs to retrieve contextual information to produce disambiguating linguistic material, resulting in an increase in the number of looks after mention.

Taken together, our results indicate that visual factors such as clutter interact with conceptual factors such as animacy in language production. The simple view according to which referents are fixated in the order in which they are mentioned, with a fixed eye-voice span between fixation and mention, does not seem to generalize to more realistic settings in which speakers describe naturalistic scenes that involve referential ambiguity.

References

Altmann, G. and Kamide, Y. (1999). Incremental interpretation at verbs: restricting the domain of subsequent reference. *Cognition*, 73(3):247–264.

Baayen, R., Davidson, D., and Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language*, 59:390–412.

Barr, D. (2008). Analyzing 'visual world' eyetracking data using multilevel logistic regression. *Journal of memory and language*, 59(4):457–474.

Branigan, H., Pickering, M., and Tanaka, M. (2008). Contribution of animacy to grammatical function assignment and word order during production. *Lingua*, 2(118):172–189.

Brockmole, J. R. and Henderson, J. M. (2006). Using real-world scenes as contextual cues for search. *Visual Cognition*, 13:99–108.

Fletcher-Watson, S., Findlay, J., Leekam, S., and Benson, V. (2008). Rapid detection of person information in a naturalistic scene. *Perception*, 37(4):571–583.

Griffin, Z. and Bock, K. (2000). What the eyes say about speaking. *Psychological science*, 11:274–279.

Henderson, J. M., Chanceaux, M., and Smith, T. J. (2009). The influence of clutter on real-world scene search: Evidence from search efficiency and eye movements. *Journal of Vision*, 9(1)(32):1–8.

Malcolm, G. and Henderson, J. M. (2009). The effects of target template specificity on visual search in real-world scenes: Evidence from eye movements. *Journal of Vision*, 9(11)(8):1–13.

Qu, S. and Chai, J. (2008). Incorporating temporal and semantic information with eye gaze for automatic word acquisition in multimodal conversational systems. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP). Honolulu*.

Rosenholtz, R., Li, Y., and Nakano, L. (2007). Measuring visual clutter. *Journal of Vision*, 7:1–22.

Schmidt, J. and Zelinksy, G. (2009). Search guidance is proportional to the categorical specificity of a target cue. *Quarterly Journal of Experimental Psychology*, 62(10):1904–1914.

Tanenhaus, M., Spivey-Knowlton, J., Eberhard, K., and Sedivy, J. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, (268):632–634.

Torralba, A., Oliva, A., Castelhano, M., and Henderson, J. (2006). Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological review*, 4(113):766–786.

Vo, M. and Henderson, J. (2010). The time course of initial scene processing for eye movement guidance in natural scene search. *Journal of Vision*, 10(3):1–13.

Uncertainty in causal inference: The case of retrospective revaluation

Christopher D. Carroll (cdcarroll@ucla.edu)
Department of Psychology, UCLA

Patricia W. Cheng (cheng@lifesci.ucla.edu)
Department of Psychology, UCLA

Hongjing Lu (hongjing@ucla.edu)
Department of Psychology, UCLA

Abstract

Since causal evidence is often ambiguous, models of causal learning should be able to represent uncertainty over causal hypotheses. Uncertainty is especially important in *retrospective revaluation* (the re-evaluation of ambiguous evidence in light of subsequent learning). We examine how a Bayesian model and an associative model (the modified SOP model of Dickinson & Burke, 1996) deal with this uncertainty. We tested the predictions of the models in an experiment with retrospective revaluation of preventive causes. Results were consistent with the predictions of the Bayesian model, but inconsistent with the predictions of the modified SOP model.

Introduction

Causal evidence is often ambiguous and causal inference uncertain. When examining an isolated case of food poisoning, it is difficult to identify the meal – never mind the food item - that caused the illness. Uncertainty is especially salient in *retrospective revaluation* (when established but ambiguous evidence is re-evaluated after subsequent learning). We examine how Bayesian and associative models of causal learning represent and deal with ambiguous evidence in retrospective revaluation. Although Bayesian models naturally represent the uncertainty of causal inference from ambiguous evidence, associative models do not.

Examples of retrospective revaluation include *reduced overshadowing* and *backward blocking*. In both of these phenomena, there is one effect whose presence we denote as + and absence we denote as - and two cues that we will call cue A and cue B. In both reduced overshadowing and backward blocking, the initial evidence shows that the effect occurs after the presentation of both cues (AB+). This evidence is ambiguous because it could be that cue A alone causes the effect, cue B alone causes the effect, or that both cues A and B independently cause the effect. Of course, it is also possible that cues A and B interact to cause the effect, but we will not consider this possibility further. We assume that, due to parsimony, this explanation is only considered when the others are ruled out.

In reduced overshadowing, participants later learn that the effect does not occur after cue A is presented on its own (i.e., A- trials follow the AB+ trials). This new evidence suggests that cue A does not cause the effect. By conditional contrast or the process of elimination, this implies that cue B caused the effect on the AB+ trials. In backward blocking, the new evidence shows that the effect occurs when cue A is presented alone (i.e., A+ trials follow the AB+ trials). Since the knowledge that cue A causes the effect explains the AB+ trials, this new evidence should make it less likely that cue B causes the effect. However, it is still possible that cue B also causes the effect. Intuitively then, reduced overshadowing – which implies that cue B must cause the effect – should offer stronger evidence for re-evaluation than backward blocking.

This intuition is reflected in studies that have compared reduced overshadowing and backward blocking to a control condition (just AB+ trials). These studies have shown that reduced overshadowing is stronger and more robust than backward blocking (Corlett et al.., 2004; Larkin, Aitken, & Dickinson, 1998; see also Beckers, De Houwer, Pineno, & Miller, 2005; Lovibond, Been, Mitchel, Bouton, & Frohardt, 2003; but see Wasserman & Berglan, 1998; Wasserman & Castro, 2005).

In this paper, we consider how different models of causal reasoning explain reduced overshadowing and backward blocking. Our goals are two-fold. Firstly, we seek to provide a principled explanation of reduced overshadowing and backward blocking by representing uncertainty. In service of this goal, we formalize our intuitions in a Bayesian model of causal inference.

Secondly, we consider how associative models deal with retrospective revaluation. We focus on the modified SOP model (Dickinson & Burke, 1996) because it explains the observed asymmetry between reduced overshadowing and backward blocking. However, we will argue that the modified SOP model predicts this asymmetry for arbitrary reasons. Therefore, we tested the modified SOP and Bayesian models in a situation where they make competing predictions: the preventive analogs of reduced overshadowing (A+, ABC-, AB+) and backward blocking (A+, ABC-, AB-).

A Bayesian model of retrospective revaluation

Bayesian models have been applied to retrospective revaluation in order to explain trial-order effects (e.g., Daw, Courville, & Dayan, 2008; Kruschke, 2008; Lu, Rojas, Beckers, & Yuille, 2008) and the influence of prior knowledge (e.g., Sobel, Tenenbaum, & Gopnik, 2004). These models, however, have not been contrasted with

associative models that were designed to explain retrospective revaluation. For this comparison, we adapt Griffiths & Tenenbaum's (2005) model of causal inference.

The model represents each possible causal explanation as a *causal graph* (e.g., Figure 1). In the causal graphs that we consider, a causal link can be generative, preventive, or non-existent. Since we assume that there are multiple cues and a single effect, a causal graph can be represented as a vector of causal links $\vec{l}$, letting $l_i = 1$ denote a generative causal relationship between cue i and the effect, $l_i = 0$ denote the absence of a causal relationship, and $l_i = -1$ denote a preventive causal relationship. We provide each causal link with a weight that represents the strength of the causal relationship, and we represent these weights as a vector $\vec{w}$ where $0 \leq w_i \leq 1$ for each w_i.

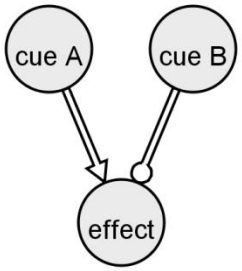


Figure 1: A causal graph where cue A causes the effect (as indicated by an arrow) and cue B prevents the effect (as indicated by a modified arrow terminating in a circle)

To represent a trial, we let the vector $\vec{c}$ denote the presence ($c_i = 1$) or absence ($c_i = 0$) of the cues and let e denote the presence ($e+$) or absence ($e-$) of the effect.

To specify the probability of the effect, we need to define a generating function that describes how causes combine to produce the effect. We adopt the noisy-or and noisy-and-not generating functions, which can be derived from the assumptions of causal power (Cheng, 1997) for generative and preventive causation, respectively. Given the vectors $\vec{c}$, $\vec{l}$, and $\vec{w}$, let G be the set of indexes such that $l_i = 1$ (i.e., generative causes of e), and let P be the set of indexes such that $l_i = -1$ (preventers of e). Using the noisy-or and noisy-and-not function, the probability of the effect is:

$$P(e+ \mid \vec{c}, \vec{l}, \vec{w}) = \left[1 - \prod_{g \in G} (1 - w_g c_g)\right] \prod_{p \in P} (1 - w_p c_p) \quad (1)$$

Then, given data D that provides a frequency count $N(e, \vec{c})$ for each combination of the presence/absence of the effect and the cues, the probability of the data as a function of the causal graph and its weights is:

$$P(D \mid \vec{w}, \vec{l}) = \prod_{(e,\vec{c})} P(e \mid \vec{c}, \vec{l}, \vec{w})^{N(e,\vec{c})} \quad (2)$$

We assume a uniform prior distribution on $\vec{w}$ and define a prior distribution on $\vec{l}$ as shown in equation 3. For each causal link, we make the link generative with probability α, preventive with probability β, and nonexistent with probability 1 – α – β. We use α and β as model parameters. For a causal graph with k generative causes, j preventive causes, and n cues, the priors are:

$$P(\vec{w}, \vec{l}) = P(\vec{w} \mid \vec{l}) P(\vec{l})$$
$$P(\vec{l}) = \alpha^k \beta^j (1 - \alpha - \beta)^{(n-k-j)} \quad (3)$$
$$P(\vec{w} \mid \vec{l}) \sim unif$$

From Bayes' theorem and our assumptions about the priors, we have

$$P(\vec{w}, \vec{l} \mid D) = \frac{1}{Z} P(D \mid \vec{w}, \vec{l}) P(\vec{w} \mid \vec{l}) P(\vec{l}) \quad (4)$$

The variable Z represents a normalizing constant. The model can be used to answer questions about the strength of a causal link or about its existence and direction. To find the posterior probability of a set of causal weights (i.e., causal strengths), we can integrate equation 4 over the other causal weights and sum over the causal graphs.

The experiment in this paper, however, asks about the existence and direction of a causal link – not its strength. Therefore, we are more interested in the probability that a causal graph generated the data. This can be found by integrating over the causal weights.

$$P(\vec{l} \mid D) = \int P(\vec{w}, \vec{l} \mid D) d\vec{w} \quad (5)$$

To calculate the probability that a cue is causal, preventive, or noncausal, we sum the probabilities of each causal graph that contains the desired relationship. If we let L be the set of causal graphs such that $l_i = x$ (where $x \in \{-1, 0, 1\}$ represents the existence and direction of the causal relationship), then:

$$P(l_i = x \mid D) = \sum_{\vec{l} \in L} P(\vec{l} \mid D) \quad (6)$$

Finally, to model causal judgments, we take the logit of this probability to obtain a measure of causal support, which is often viewed as a psychologically realistic measure of causal judgment (Griffiths & Tenenbaum, 2005):

$$\text{causal support} = \log\left(\frac{P(l_i = x \mid D)}{1 - P(l_i = x \mid D)}\right) \quad (7)$$

Retrospective revaluation

To explain reduced overshadowing and backward blocking, we consider the causal graphs with two cues and one effect. Since we only allow causal relationships between cue A and the effect and cue B and the effect, this gives us 9 (i.e., 3^2) causal graphs. We set the parameters such that the priors across the graphs are uniform (i.e., α = β = 1/3). When the model is given data where there are 4 trials of each type (e.g., 4xAB+ 4xA+ in the backward blocking condition), it can be used to generate a support measure for the hypothesis that cue B causes the effect. The model predicts that the difference between reduced overshadowing and a control (AB+) is larger than the difference between backward blocking and the control (see Table 1).

To understand these predictions, it is useful to consider the posterior distribution of the weights. First, we consider the joint posterior of cues A and B after the AB+ trials conditional on both links being generative (see Figure 2). This posterior suggests that there is considerable uncertainty over the weights of cues A and B. However, it also suggests a dependency between the weights of the cues: at least one

of the cues must be causal. If w_a is small, then w_b must be large. However, if w_a is large, then there is still uncertainty over w_b. This dependency explains reduced overshadowing and backward blocking. If subsequent evidence indicates that cue A does not cause the effect (as is the case for reduced overshadowing), then cue B must. However, if subsequent evidence indicates that cue A causes the effect (as is the case for backward blocking), then the influence of cue B cannot be conclusively known.

Table 1: The causal support measure for the causal link between cue B and the effect for reduced overshadowing, control, and backward blocking

Condition	support
reduced overshadowing (AB+, A-)	5.02
backward blocking (AB+, A+)	0.09
control (AB+)	1.05

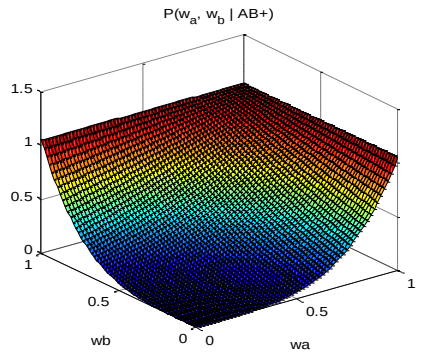


Figure 2: The joint posterior distribution of w_a and w_b.

These predictions are reflected in the posterior weights of cue B alone in the different retrospective revaluation conditions (see Figure 3). In the reduced overshadowing condition, it is clear that cue B must cause the effect: there is almost no possibility that the weight from cue B to the effect is zero. On the other hand, there is considerable uncertainty about the weight of cue B in both the blocking and control conditions: neither excludes the possibilities that B is noncausal or that B is causal.

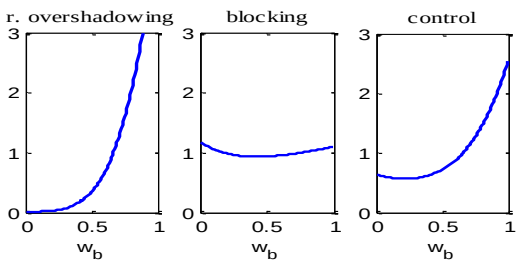


Figure 3: The posterior of the weights of cue B when cue B is a generative cause.

Associative models

Retrospective revaluation is notoriously problematic for associative models, but two associative models have been developed to explain it: Van Hamme & Wasserman's (1994) modified RW (Rescorla-Wagner) model and the Dickinson & Burke's (1996) modified SOP model. The problem for standard associative models is that they only learn about present cues. This precludes an explanation of reduced overshadowing and backward blocking, where learning about cue A leads participants to revise their beliefs about the absent cue B. To surmount this difficulty, the modified RW model and the modified SOP model utilize *within-compound associations:* associations formed between simultaneously-presented cues. On the initial AB+ trials, these models learn an association between cues A and B. Later, the within-compound associations are used to recall associated cues that are absent on the trial, allowing the model to learn about them. If an A+ trial followed, the models would identify the absent cue B as an expected cue and would use this identification to support re-evaluation.

Although within-compound associations allow the models to learn about absent cues, it is not clear whether they offer a genuine representation of uncertainty.

Since the modified RW model incorrectly predicts that backward blocking will be at least as strong as reduced overshadowing (see Larkin et al., 1998 for a detailed explanation), we focus on the modified SOP model.

The modified SOP model

In the modified SOP model, there are three activation states: the A1 (observed), A2 (expected), and I (inactive) states. Each cue is represented by a node that is made up of many elements, so a node can be in more than one activation state. For example, if a cue were presented on a trial and it was expected on the basis of within-compound associations, there might be 40% of its elements in the A1 state, 40% in the A2 state, and 20% inactive. Excitatory learning occurs between two nodes to the extent that they are both in the A1 state or both in the A2 state. Inhibitory learning occurs between two nodes to the extent that one is in the A1 state and the other is in the A2 state. No learning occurs otherwise.

On AB+ trials, the modified SOP model learns that each cue is associated with the effect and that there is a within-compound association between cues A and B. When cue A is presented alone, the within-compound association between cues A and B leads cue B to enter the A2 activation state (see Table 2). The state of the effect depends on the type of retrospective revaluation. For reduced overshadowing, the effect is expected but absent, so it will enter state A2. This puts the effect in the same state as cue B, so learning will be exclusively excitatory. For backward blocking, the effect is both expected and present, so it will enter states A1 and A2. Since this means that cue B and the effect will be partly in the same state and partly in a different state, there will be conflicted learning that is both excitatory and inhibitory. Therefore, the modified SOP model predicts the oft-

observed asymmetry between reduced overshadowing and backward blocking. When compared to a control condition, the modified SOP model predicts that reduced overshadowing will be a stronger and more robust effect than backward blocking.

Table 2: Activation states and learning during retrospective revaluation. The ↑ symbol indicates excitatory learning (an increase in associative strength) and the ↓ symbol indicates inhibitory learning (a decrease in associative strength).

condition	Cue B	Effect	B-effect learning
r. overshadowing (A-)	A2	A2	↑
b. blocking (A+)	A2	A1 and A2	↑↓
control	-	-	none

However, these predictions seem arbitrary: the modified SOP model predicts both excitatory and inhibitory learning whenever the effect is both present and expected, but it is not clear why this should be the case. To test the modified SOP model, we designed an experiment where its predictions diverged from those of the Bayesian model.

Method

To test the predictions of the modified SOP model, we examined the preventive analogs of reduced overshadowing (i.e., A+, ABC-, AB+) and backward blocking (i.e., A+, ABC-, AB-). Until the final AB+ or AB- trials, the evidence suggests that cue A causes the effect and that either cue B alone prevents the effect, cue C alone prevents the effect, or that cues B and C prevent the effect. Like its generative analog, preventive reduced overshadowing eliminates two of these explanations by showing that cue B does not prevent the effect. By the process of elimination, one would infer that cue C must have been responsible for preventing the effect on the ABC- trials. The AB- trials in backward blocking show that cue B prevents the effect, but these trials do not fully clarify the influence of cue C: it is still possible that C prevents the effect, and it is still possible that it does not. Preventive reduced overshadowing should be a stronger and more robust effect than preventive backward blocking.

The modified SOP model predicts the opposite. It predicts that learning is conflicted whenever the effect is both present and expected (as it is during reduced overshadowing AB+ trials), but that learning is clear whenever the effect is expected but absent (as it is during backward blocking AB- trials). According to the modified SOP model, preventive reduced overshadowing should be weaker and less robust than preventive backward blocking (see Table 3).

For our experimental task, we used a cover story where participants were asked to discover which foods cause and prevent allergic reactions in medical patients. We manipulated the retrospective revaluation condition (preventive reduced overshadowing, preventive backward blocking, reduced overshadowing control, and blocking control) within-subjects. We also manipulated expectations about the probability that a randomly selected fruit would prevent an allergic reaction. Bayesian models have a mechanism for integrating prior knowledge and evidence from observations, and we manipulated expectations to assess whether prior knowledge influenced the participants.

Table 3: Predicted changes in associative strength according to the modified SOP model for the preventive analogs of reduced overshadowing and backward blocking.

Preventive analog	Cue C	Effect	C-effect learning
r. overshadowing (AB+)	A2	A1 and A2	↑↓
b. blocking (AB-)	A2	A2	↑
control	-	-	none

Participants

Twenty-four undergraduates at the University of California, Los Angeles participated for course credit. The participants were randomly assigned to a infrequent (n = 7), occasional (n = 9), or frequent (n = 8) prevention condition.

Materials

We selected icons that pictorially represented 21 different fruits.

Procedure

At the beginning of the experiment, participants were asked to take the perspective of allergists specializing in patients who have fruit allergies. They were informed that fruit allergies can be both caused and prevented in these patients. That is, some fruits might cause an allergic reaction in a patient, but other fruits might prevent an allergic reaction.

Participants were told that they would read through the "fruit journals" of patients. They were informed that a fruit journal lists the fruits that a patient ate on a given day, and also records whether the patient had an allergic reaction.

Each experimental trial corresponded to the record for one day in the fruit journal. A trial began by displaying the icons and names of whichever fruits the patient ate on that day. These icons were displayed alone for 1.5 seconds, at which point a cartoon face appeared. The cartoon face signified whether the patient had an allergic reaction on that day: a smiley face with the text "ok" meant that the patient did not have a reaction and a frowning face with the text "allergic reaction" meant that the patient had a reaction. The fruits and cartoon face were displayed together for 2.0 seconds before the trial ended.

Participants read the fruit journals of five different patients. The journal of the first patient was used to manipulate the priors. The other four journals represented the four retrospective revaluation conditions. The fruits

were randomly mapped to the different fruit journals, and each fruit appeared in exactly one fruit journal.

When the first patient was introduced, participants were told the approximate probability that a fruit prevents allergic reactions (the bracketed phrases were selected according to the infrequent, occasional, or frequent priors conditions):

> *As is often the case with fruit allergies, a small number of fruits caused the patient's allergic reaction, [very few / some / many] prevented it, and [many / some / very few] did nothing.*

The first fruit journal provided evidence for this claim. The patient experienced an allergic reaction after consuming one of the fruits alone, but the other four fruits in the journal did not cause the patient to experience an allergic reaction. Zero, two, or four of the other fruits prevented the allergic reaction (in the infrequent, occasional, and frequent priors conditions, respectively). This was demonstrated by showing, for each of the other fruits, whether the patient had an allergic reaction after consuming that fruit and the causal fruit at the same time.

To familiarize the participants with the causal questions, the participants were then asked whether each fruit in the first journal caused, prevented, or did nothing to influence the patient's allergic reactions. Participants responded on a sliding scale running from -6 to 6 where -6 was labeled "definitely prevents", -3 was labeled "maybe prevents", 0 was labeled "neither", 3 was labeled "maybe causes", and 6 was labeled "definitely causes."

After answering questions about the influence of fruits on the first patient, participants viewed, in random order, a fruit journal for each retrospective revaluation condition. In each journal, the trials were divided into three stages, and the data for each stage are shown in Table 4. Each of the listed patterns was shown four times (e.g., fruit A caused an allergic reaction four times in stage 1). Within each stage, the trials were presented in a random order.

Table 4: The data (by retrospective revaluation condition)

condition	stage 1	stage 2	stage 3
reduced overshadowing	A+ B- C- D-	A+ ABC-	A+ AB+
backward blocking	A+ B- C- D-	A+ ABC-	A+ AB-
control (rOS)	A+ B- C- D-	A+ ABC-	A+ AD+
control (BB)	A+ B- C- D-	A+ ABC-	A+ AD-

Following the presentation of the data for each retrospective revaluation condition, participants were asked to report whether each fruit caused, prevented, or did nothing to influence the patient's allergic reactions. The response scale was identical to the scale that was used in the first fruit journal.

Results

The ratings for cue C are shown in Figure 4. The predicted asymmetry between reduced overshadowing and backward blocking was found. Compared to its control, reduced overshadowing had a substantial influence: it led participants to be much more certain that cue C prevented allergic reactions. Ratings for cue C did not differ substantially between the backward blocking and control conditions. The priors manipulation did not seem to substantially influence the causal ratings.

An ANOVA confirmed that the retrospective revaluation condition influenced causal ratings, $F(3, 63) = 23.84$, $p < .001$, and that there was no effect of the priors manipulation, $F(2, 21) = 0.29$, $p = .75$, or interaction between the priors condition and retrospective revaluation condition, $F(6, 63) = 0.57$, $p = .75$. Planned comparisons indicated that the effect of retrospective revaluation condition was driven by the difference between reduced overshadowing and its control, $t(23) = 6.30$, $p < .001$, and not by the difference between blocking and its control, $t(23) = 0.94$, $p = .36$.

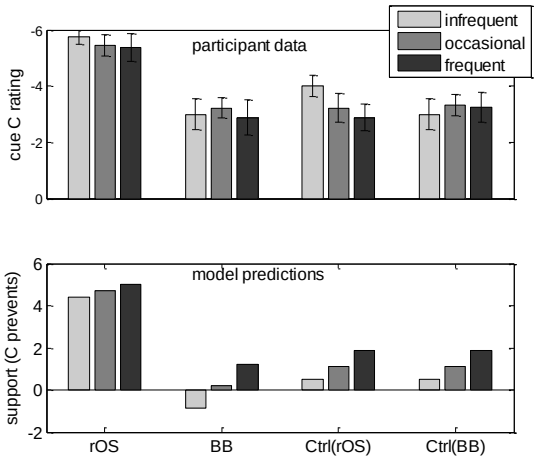


Figure 4: Causal ratings and Bayesian model predictions for cue C by retrospective revaluation condition and the prior likelihood of prevention. On both graphs, higher points on the y-axis correspond to greater certainty that cue C prevents allergic reactions. (rOS = reduced overshadowing, BB = backward blocking, Ctrl = control)

To derive the predictions of the model, we set $\alpha = .2$ and then set β depending on the priors condition ($\beta = .2$ for infrequent, $\beta = .4$ for occasional, and $\beta = .6$ for frequent). The predictions of the model are shown for each condition in Figure 4. The model offered a good quantitative fit to the data, $r = -.87$.

Discussion

The results clearly contradict the predictions of the modified SOP model. Preventive reduced overshadowing was a much stronger effect than preventive backward blocking. The Bayesian model predicts this finding, and also offers a principled justification for its prediction.

The priors manipulation did not influence the participants' causal ratings, but the interpretation of this finding is unclear. The Bayesian model predicts a limited effect of the priors manipulation, and the small number of participants per condition limited the experiment's statistical power. Furthermore, since the prior frequencies were merely manipulated verbally, the manipulation may have been too weak. Other research has shown that priors can influence causal judgment (e.g., Sobel, Tenenbaum, & Gopnik, 2004).

A final possibility is that the participants only represented approximate probabilities. Participants may have categorized the probability of causation by tracking whether a causal link *definitely, maybe,* or *definitely does not* exist. Consistent with this possibility, participants did not seem to differentiate between different degrees of *maybe* (e.g., see Figure 4).[1]

The modified SOP model predicts the relative size of reduced overshadowing and backward blocking, but the preventive analogs of these findings illustrate that it does so for the wrong reasons. In both the modified RW model and modified SOP models, within-compound associations make a poor substitute for a genuine representation of uncertainty. Other associative models that use within-compound associations may be capable of explaining these results (e.g., Denniston, Savastano, & Miller, 2001), so further experimentation is necessary. However, the results of this experiment raise serious questions about whether within-compound associations offer a genuine representation of uncertainty. As instantiated by the modified SOP model, they clearly do not.

Acknowledgments

The preparation of this article was supported by AFOSR FA 9550-08-1-0489.

References

Beckers, T., De Houwer, J., Pineno, O., & Miller, R. (2005). Outcome additivity and outcome maximality influence cue competition in human causal learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 31 (2),* 238-249.

Corlett, P. R., Aitken, M. R. F., Dickinson, A., Shanks, D. R., Honey, G. D., Honey, R. A. E., Robbins, T. W., Bullmore, E. T., & Fletcher, P. C. (2004). Prediction error during retrospective revaluation of causal associations in humans: fMRI evidence in favor of an associative model of learning. *Neuron, 44,* 877-888.

Daw, N. D., Courville, A. C., & Dayan, P. (2008). Semi-rational models of conditioning: The case of trial order. In N. Chater & M. Oaksford (Eds.), *The Probabilistic Mind: Prospects for Bayesian Cognitive Science* (pp. 427-448). New York, NY: Oxford University Press.

Denniston, J. C., Savastano, H. I., & Miller, R. R. (2001). The extended comparator hypothesis: Learning by contiguity, responding by relative strength. In R. R. Mowrer & S. B. Klein (Eds.), *Handbook of contemporary learning theories* (pp. 65-117). Mahwah, NJ: Lawrence Erlbaum Associates.

Dickinson, A. & Burke, J. (1996). Within-compound associations mediate the retrospective revaluation of causality judgements. *The Quarterly Journal of Experimental Psychology, 1996, 49B (1),* 60-80.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology, 51,* 334-384.

Kruschke, J. K. (2006). Locally Bayesian learning with applications to retrospective revaluation and highlighting. *Psychological Review, 113 (4),* 677-699.

Larkin, M. J. W., Aitken, M. R. F., & Dickinson, A. (1998). Retrospective revaluation of causal judgments under positive and negative contingencies. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24 (6),* 1331-1352.

Lovibond, P. F., Been, S., Mitchell, C. J., Bouton, M. E., & Frohardt, R. (2003). Forward and backward blocking of causal judgment is enhanced by additivity of effect magnitude. *Memory & Cognition, 31 (1),* 133-142.

Lu, H., Rojas, R. R., Beckers, T., & Yuille, A. (2008). Sequential causal learning in humans and rats. *Proceedings of the Twenty-ninth Annual Conference of the Cognitive Science Society*

Sobel, D. M., Tenenbaum, J. B., & Gopnik, A. (2004). Children's causal inferences from indirect evidence: Backwards blocking and Bayesian reasoning in preschoolers. *Cognitive Science, 28,* 303-333.

Van Hamme, L. J., & Wasserman, E. A. (1994). Cue competition in causality judgments: The role of nonpresentation of compound stimulus elements. *Learning & Motivation, 25,* 127-151.

Wasserman, E. A., & Castro, L. (2005). Surprise and change: Variations in the strength of present and absent cues in causal learning. *Learning & Behavior, 33 (2),* 131-146.

Wasserman, E. A., & Berglan, L. R. (1998). Backward blocking and recovery from overshadowing in human causal judgment: The role of within-compound associations. *Quarterly Journal of Experimental Psychology, 51B,* 121-138.

[1] Interestingly, causal support does something similar. Rather than using the untransformed probability of a causal link (equation 6), causal support transforms this probability with the logit function. The logit function deemphasizes differences in moderate probabilities (i.e., those near .5) and emphasizes differences in extreme probabilities.

It is worth noting, however, that the logit transformation would not save the modified RW and the modified SOP models. Even when augmented with a logit transformation, these models fail to explain the results. The modified RW model learns that cue C is non-causal in preventive backward blocking more quickly than it learns that cue C is preventive in preventive reduced overshadowing. If anything, the logit transformation would highlight this failing. The modified SOP model makes predictions in the wrong direction. As a monotonic transformation, the logit function preserves the direction of these incorrect predictions.

The Role of Causal Schemas in Inductive Reasoning

Ralf Mayrhofer (rmayrho@uni.goettingen.de)
Jonas Nagel (jnagel1@uni-goettingen.de)
Michael R. Waldmann (michael.waldmann@bio.uni-goettingen.de)
Department of Psychology, University of Göttingen
Gosslerstr. 14, 37073 Göttingen, Germany

Abstract

Inductive reasoning allows us to go beyond the target hypothesis and capitalize on prior knowledge. Past research has shown that both the similarity of categories and specific knowledge about causal relations affect inductive plausibility. We go one step further and focus on the role of abstract causal schemas about main effects and interactions. Two experiments show that both prior assumptions about abstract causal schemas and the similarity of the corresponding causal effects affect inductive judgments. Reasoners have different prior beliefs about the likelihood of main-effect versus interactive schemas, and rationally combine these prior beliefs with new evidence in a way that can be modeled as Bayesian belief updating.

Keywords: inductive reasoning; causal schemas; causal interactions; Bayesian inference

Introduction

Inductive reasoning occurs in various contexts. In associative learning we infer a general regularity from a set of learning trials. In causal learning we use a sample of observations and go beyond the information given to induce general causal laws. Inductions not only occur at the level of learning exemplars but can also relate prior knowledge about hypotheses to new hypotheses. For example, knowing that cats have hearts allows us to give an informed guess about the probable validity of the hypothesis that wolves have hearts, as well. The interconnectedness of our knowledge is a powerful tool to form informed guesses about new hypotheses.

Although inductive inferences based on exemplars have for a long time been studied in learning, inductions between hypotheses is a fairly recent research goal (see Feeney & Heit, 2007, for an overview). Many early studies have focused on the similarity of categories as the basis for inductive inference.

However, similarity between categories is not the only factor influencing inductive inferences. Based on a variant of causal-model theory, Rehder (2007) has proposed a theory which treats inductive generalizations as causal reasoning. According to this theory we assess the likelihood that a novel feature applies to a new category on the basis of our beliefs about the causal relations that connect that feature to the category. For example, subjects tend to infer that a category has a novel feature if they believe that this feature is caused by or is the cause of a characteristic feature of the category.

Whereas previous research on inductive reasoning has focused on the similarity of categories (i.e., feature overlap) or the causal relation connecting a novel feature to the categories, our current research explores the role of more abstract and complex causal schemas in inductive reasoning about hypotheses.

Causal Schemas in Learning

Thus far, causal schemas have mainly been studied in the context of learning, not inductive reasoning. We will briefly review this research to derive hypotheses for inductive reasoning. Within the causal learning literature it has typically been assumed that the default assumption about multiple causes is that they combine additively. For example, Cheng (1997) has postulated a noisy-OR schema as the default for generative causes according to which multiple causes independently generate an effect.

Although additive integration of multiple causes may be the default, causes may also interact (Kelley, 1972; Novick & Cheng, 2004). The majority of research about additive integration and interactions has been conducted within the associative learning literature. Popular examples of an interaction are positive and negative patterning, in which the effect cannot be predicted on the basis of an additive integration of the individual causes. Positive patterning (PP) refers to learning a situation in which two cues (e.g. A and B), when presented individually, are not paired with the outcome (A– and B– trials), but when presented together they are paired with the outcome (AB+ trials). For example, two drugs may individually not cause a side effect but only in combination. The corresponding additive cue combination (A–, B– => AB–) we henceforth call negative main effect (ME-). In contrast, negative patterning (NP) refers to a situation in which cues A and B, when presented alone, are paired with the outcome (A+ and B+ trials), but when presented together are not paired with the outcome (AB- trials). An example of this interaction are two drugs which each cause a side effect, but cancel out each other's effect when presented together. The complimentary main effect (A+, B+ => AB+) we will call positive main effect (ME+).

Shanks and Darby (1998) found that people can learn both of these interactions (PP and NP) concurrently, and can form the appropriate abstract schematic representations. Moreover, Shanks and Darby demonstrated that people transfer these schemas to new cues which have not previously been shown together. For example, participants

that underwent NP training with cues A and B, and were then shown C+ and D+ trials, could infer that the novel compound CD would not be followed by the outcome (see also Lucas & Griffiths, 2010).

Kemp et al. (2007) have proposed a Bayesian model which explains Shanks and Darby's (1998) data. The model learns causal schemas by monitoring the co-occurrences of cues and outcomes, and groups together cues that behave in a similar fashion. In the NP case this model groups together cues that co-occur with the outcome in isolation, but do not co-occur with the outcome when paired with another cue of the same kind. Importantly, the model can use these cue groupings to generate predictions about novel cue-combinations at test, and thus solve the [C+, D+, CD?] test cases.

Causal schemas differ in learning difficulty in a way consistent with the assumptions in the causal literature. Studies of patterning have shown that learning about patterning schemas is a difficult task and proceeds much slower than when organisms are confronted with main effects (Kehoe, 1988).

Causal Schemas in Inductive Reasoning

Previous research has shown that people are capable of abstracting causal schemas from learning data and transferring them to new situations. However, very little is known about how causal schemas affect inductive plausibility when knowledge is presented as a set of individual facts and hypotheses. Based on the findings in the learning literature and our predecessor study (O. Griffiths et al., 2009), we expect that reasoners bring to bear different priors about causal schemas. We expect them to consider main-effect relations more likely than interactions, especially disordinal ones as in the NP case. Different priors for schema knowledge should therefore constitute one important, hitherto neglected factor influencing inductive plausibility between facts and hypotheses. In particular, a simple application of Bayes' rule predicts that a new instance of an unlikely interaction should have a larger impact on inductive beliefs than a new instance of a schema that is already considered common (e.g., main effects).

A second factor we will explore in the present research concerns the question whether the similarity of the schemas influences induction. Since causal schema hypotheses are abstract not only with respect to the involved cues but also with respect to other properties of the underlying causal relation, similarity between patterning or main effect instances is obviously determined by at least two factors: the similarity of the involved cues in corresponding roles (thus the similarity between the pair of cues A and B and the pair of cues C and D) as well as the similarity of the effects which are generated by A and B on the one hand and C and D on the other hand. In the present experiments we will keep the similarity between cues constant across conditions but we will manipulate the similarity of the effect, a factor that has been neglected in previous research. Moreover, similarity will be investigated in the context of both confirming and disconfirming evidence. Our main hypothesis is that both confirming and disconfirming evidence should more strongly increase or decrease the prior belief, respectively, if the similarity between the effects mentioned in the facts and the hypotheses is high rather than low.

Schema-based Priors and Belief Updating

O. Griffiths et al. (2009) proposed a simple Bayesian account of schema-based belief updating which models how people update their belief in some schema hypothesis H_i (i.e., H_{PP}, H_{NP}, H_{ME+} or H_{ME-}, respectively) given a confirming instance D via Bayes' rule:

$$P(H_i|D_i) = \frac{P(D_i|H_i)P(H_i)}{P(D_i)}$$

The posterior belief in H_i depends upon the likelihood of the confirming instance D_i given H_i being true, and the participants' prior belief in H_i. An example would be an inference about the hypothesis $(A+, B+) \Longrightarrow (AB-)$ when it is already known that the conclusion $(C+, D+) \Longrightarrow (CD-)$ is true for novel cues C and D from the same domain as A and B.

Assuming that people consider patterning schemas to be less plausible than main-effect schemas, Griffiths et al. (2009) used this Bayesian belief updating to derive the following predictions: Beliefs regarding patterning schemas will be change more profoundly in response to the observation of a confirming instance than beliefs regarding main-effect schemas. After updating, however, plausibility ratings for patterning schemas should still not exceed those for main-effects.

In an experiment Griffiths et al. (2009) tested these predictions. 32 participants were presented with a series of eight fictitious scenarios describing causal relationships between a number of cues and an effect in several different content domains (e.g., physics or biology). Each of the eight trials consisted of two subscenarios (see Table 1). Subscenario 1 contained three statements: The first two statements, the premises, were labeled as facts (*Fact A* and *Fact B*), and the participants were instructed to treat them as true facts. Each of these premises described one of two cues (A or B) that either did or did not cause an outcome. The third statement, labeled *Conclusion*, was a causal statement about the compound AB that again either did or did not cause the same effect. The distribution of presence or absence of the effect in the three statements determined the cue interaction type of the trial (see Table 1). Participants were then requested to indicate to what extent they believed the conclusion statement to be true as well. Given that the cues and their combinations were novel, these responses were taken as indicators of prior beliefs in the plausibility of the corresponding schema. Afterwards, subjects were presented with the second subscenario, in which they received confirming evidence in the form of three further premises (*Facts 1—3*). These premises described two different cues from the same domain (C and D) and their

compound CD causing or not causing the same effect as in the first subscenario. Moreover, the presented schema was the same. Then the participants were once more asked to indicate the extent to which they believed the conclusion statement about the compound AB to be true, this time in consideration of the additional evidence they had received about C and D.

Table 1: Design of the Experiment by Griffiths et al. (2009)

		Cue Interaction			
Subsc.	*Statement*	PP	ME-	NP	ME+
1	Fact A	A–	A–	A+	A+
	Fact B	B–	B–	B+	B+
	Conclusion	AB+	AB–	AB–	AB+
2	Fact 1	C–	C–	C+	C+
	Fact 2	D–	D–	D+	D+
	Fact 3	CD+	CD–	CD–	CD+
	Fact A	A–	A–	A+	A+
	Fact B	B–	B–	B+	B+
	Conclusion	AB+	AB–	AB–	AB+

Note. Letters A – D represent causes. Symbols + and – indicate statements in which the cause either produced the effect or did not, respectively. The dashed line separates premises from conclusions. Subsc. = Subscenario. Adapted from Griffiths et al. (2009).

The results of this study were in line with the predictions of Bayesian updating. The baseline ratings in subscenario 1 indicated that participants assigned a higher prior probability to main effects than to patterning interactions. The increase in the ratings between subscenarios 1 and 2 was higher for patterning interactions than for main effects, while the mean ratings in subscenario 2 remained higher for main effects than for patterning interactions even after updating.

Bayesian Belief Updating and Similarity

For the sake of simplicity, Griffiths et al. (2009) assumed that the likelihood $P(D_i|H_i)$ of a confirming instance D_i is represented by some fixed number larger than 0.5 (i.e., the instance is informative) but less than 1 (i.e., the inference from the instance to the hypothesis, which is formulated with respect to another pair of cues, is tainted with uncertainty), and that the likelihood is a function of the similarity between the confirming instance D_i and the instance addressed by the hypothesis H_i.

Since the similarity of instances was not manipulated by Griffiths et al. (2009), it remained open whether this factor influences judgments. Making the similarity component explicit and extending the model to disconfirming instances the proposal of Griffiths et al. (2009) can be revised as:

$$P(D_i|H_i) = \begin{cases} \dfrac{S(D_i,H_i)}{2} & D_i\ confirms\ H_i \\ 1 - \dfrac{S(D_i,H_i)}{2} & D_i\ disconfirms\ H_i, \end{cases}$$

with $S(D_i,H_i) \in [0,1]$ representing some monotone similarity measure expressing the subjective similarity between the instance D_i and the instance addressed by H_i. Thus, the more similar D_i and H_i are, the stronger the predicted belief update should be (either in the positive direction in the confirming case, or in negative direction in the disconfirming case). Disconfirming evidence is defined as evidence that confirms the contrast hypothesis. For example, $(C+, D+) \Longrightarrow (CD-)$ confirms the negative-patterning hypothesis and disconfirms the positive main-effect hypothesis.

Experiment 1

This experiment aims at replicating and extending the results from the experiment by Griffiths et al. (2009). We make a first attempt to manipulate the similarity between the different hypotheses from the same domain. As laid out above, a decrease in similarity between the confirming instance and the instance about which the hypothesis is formulated should decrease the tendency to generalize from the confirmatory evidence to the conclusion statement in question. In the present experiment we will use identical cues in the two subscenarios (i.e., A=C, B=D) but vary the similarity of the effects. Thus, Bayesian belief updating predicts a stronger increase of inductive confidence from subscenario 1 to subscenario 2 in the high similarity condition as opposed to the low similarity condition.

Method

Participants 48 University of Göttingen undergraduates participated in a series of various unrelated computer-based experiments in our lab either for course credit or for €7/h.

Design The design was closely matched to the experiment by Griffiths et al. (2009). We manipulated two independent variables in the scenarios presented to participants. The first factor was the type of causal schema and had four levels: ME-, ME+, NP, and PP. The second factor was the similarity between the to-be-judged conclusion and the confirmatory evidence. We manipulated whether the two scenarios used the same causal effect or different effects from the same domain.

Each participant responded to a total of 16 scenarios. The scenarios were randomly assigned to the experimental conditions separately for each participant. We used a complete 4 (Cue Interaction: ME-, ME+, NP, PP) × 2 (Similarity: same effect vs. different effect) repeated-measurement ANOVA design. Each subject thus received two trials in each experimental condition. The trial order was randomly determined for each individual participant.

Materials and Procedure Participants completed the experiment individually on desktop computers. The experiment began with an instruction section which informed participants about the course of their task and briefly tested whether they thoroughly understood it.

Afterwards, participants were presented with 16 fictitious scenarios from different content domains (physics, chemistry, biology, medicine) that were constructed to cover

a broad range of settings. Fictitious cues and effects were used to make sure that participants could not rely on specific prior causal knowledge when making their inferences. Each scenario again consisted of two subscenarios. Subscenario 1 was set up exactly as in Griffiths et al. (2009) (see upper part of Table 1). After having read the two premises (*Facts A—B*) and the *Conclusion*, participants indicated the extent to which they believed the conclusion statement to be true (this rating will be labeled *Rating 1* from now on). For this task participants were provided with an 11-point scale, ranging from "definitely false" at the left-hand end (0) to "definitely true" at the right-hand end (10).

After having provided this rating, participants proceeded to subscenario 2. Its set-up was similar to that in Griffiths et al. (2009) in that three additional facts from the same domain were introduced (*Facts 1—3*; see lower part of Table 1). These facts always constituted confirming evidence for the hypothesis that the conclusion is true. The three statements from subscenario 1 were repeated below the three new statements. Apart from that, we made two important changes in the present experiment regarding the materials of subscenario 2 in order to manipulate the similarity between the to-be-judged conclusion and the confirmatory evidence. First, the confirming evidence consisted of statements about the *same* cues as in the statements in subscenario 1 (i.e., A & B), so that overall similarity was increased compared to the material in Griffiths et al. (2009). Second, we manipulated the similarity between the effect caused by these cues in subscenario 1 and in the new statements of subscenario 2. In half of the trials, cues A and B caused (or did not cause) the same effect in both the to-be-judged conclusion and the provided confirmatory evidence. In the other half, the effect differed between both sets of statements. This means that in all same-effect conditions, in subscenario 2 *Facts 1—2* were identical to *Facts A—B*, and *Fact 3* was identical to the to-be-judged conclusion. Logically, all participants should have indicated certainty about the truth of the conclusion in this condition, since it was already stated as true in the premises. Table 2 shows the material of subscenario 2 in a sample trial.

Table 2. Sample of Subsc. in an NP/Different Effect trial

Fact 1:	Aering Heptosulfin with methane causes the Heptosulfin to become crystalline.
Fact 2:	Aering Heptosulfin with butane causes the Heptosulfin to become crystalline.
Fact 3:	Aering Heptosulfin with a mixture of methane and butane does not cause the Heptosulfin to become crystalline.
Fact A:	Aering Heptosulfin with methane causes the Heptosulfin to become isomorph.
Fact B:	Aering Heptosulfin with butane causes the Heptosulfin to become isomorph.
Conclusion:	Aering Heptosulfin with a mixture of methane and butane does not cause the Heptosulfin to become isomorph.

Following the presentation of the statements, participants were asked once again to rate the *Conclusion*, using the same scale as in subscenario 1. This rating, which was given after confirming evidence had been presented, is from here on referred to as *Rating 2*. Participants then proceeded directly to the next scenario. This process was repeated until all 16 scenarios were complete. The computer program ensured that participants were not able to return to any previous questions.

Results

The results of Experiment 1 are summarized in Figure 1. First, different assumptions about the prior probability of main effects vs. patterning interactions are evident in the much higher mean ratings in *Rating 1* in ME- and ME+ trials compared to NP and PP trials ($F_{3,141}$=177.3, p<.001, η_p^2=.79). Thus, again main-effect schemas were assumed to be more likely than interaction schemas. Second, belief change (increase from *Rating 1* to *Rating 2* within conditions) was influenced by the Cue Interaction factor ($F_{3,141}$=28.18, p<.001, η_p^2=.37), by the Similarity factor ($F_{3,47}$=82.39, p<.001, η_p^2=.64), and by the interaction between both factors ($F_{3,141}$=15.18, p<.001, η_p^2=.24). That is, after receiving positive evidence, participants tended to increase their confidence in the conclusion more in the cases exhibiting patterning-interactions than in the cases exhibiting main-effects. The dependence of belief updates on prior knowledge is predicted by basic Bayesian belief updating and replicates the findings of Griffiths et al. (2009). The belief change is also larger when the cues cause the same effect in both instances rather than a different effect. Furthermore, the interaction indicates that the difference in belief change between same-effect and different-effect conditions was much more pronounced in patterning-interactions than in main-effect trials (planned contrast[1]: $F_{1,47}$=22.05, p<.001).

Experiment 2

The main goal of Experiment 2 was to investigate how effect similarity and type of evidence (i.e., confirming vs. disconfirming evidence) interact with the type of causal schemas. In Experiment 1 we have already shown that the more similar the instances are, the more confident the participants are in the truth of the hypothesis. Bayesian belief updating predicts that the opposite is expected if disconfirming evidence is presented. To test this prediction, we included disconfirming evidence in half of the trials. In contrast to Experiment 1, we increased the dissimilarity of the cues between the subscenarios to test whether the similarity of the effect event also influences inductive ratings when the cues are more dissimilar.

[1] $(M_{Same/PP\&NP}-M_{Diff/PP\&NP})-(M_{Same/ME}-M_{Diff/ME})$=4.15

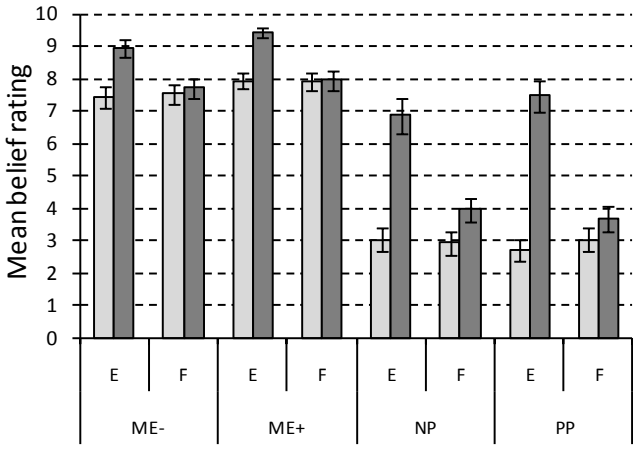


Figure 1: Means of belief ratings of conclusions, and standard errors for patterning-interaction and main-effect schemas (ME-, ME+, NP, PP) plotted against effect similarity (E: same effect, F: different effect). Light grey bars (left hand side) indicate ratings before the confirming instance was shown (*Rating 1*), dark grey bars (right hand side) show ratings after the confirming instance was presented (*Rating 2*).

yielded a complete 4 (Cue Interaction: ME-, ME+, NP, PP) × 2 (Similarity: same effect vs. different effect) × 2 (Evidence: confirming vs. disconfirming) repeated-measurement ANOVA design. Each participant again responded to 16 trials, one from each condition.

Materials and Procedure The procedure and materials corresponded to Experiment 1, apart from two changes. First, we manipulated an Evidence factor. In half of the trials, the additional evidence in subscenario 2 was disconfirming rather than confirming evidence. That is, *Facts 1—3* did not instantiate the same causal schema presented in subscenario 1, but rather its complement. If the statements in subscenario 1 formed a positive (negative) main effect, the additional evidence in subscenario 2 was confirming for negative (positive) patterning, and vice versa. The second change was that *Facts 1—3* were no longer about the same cues as *Facts A—B* and *Conclusion* (i.e., A and B), but about different cues from the same domain (C and D). That is, in all same-effect conditions, both instances differed with regards to the cues, whereas they differed with regard to both the cues and the effect in the different-effect conditions. We thus introduced a constant level of dissimilarity in all conditions on the cue level so that in none of the cases the to-be-judged conclusion was identical to one of the premises.

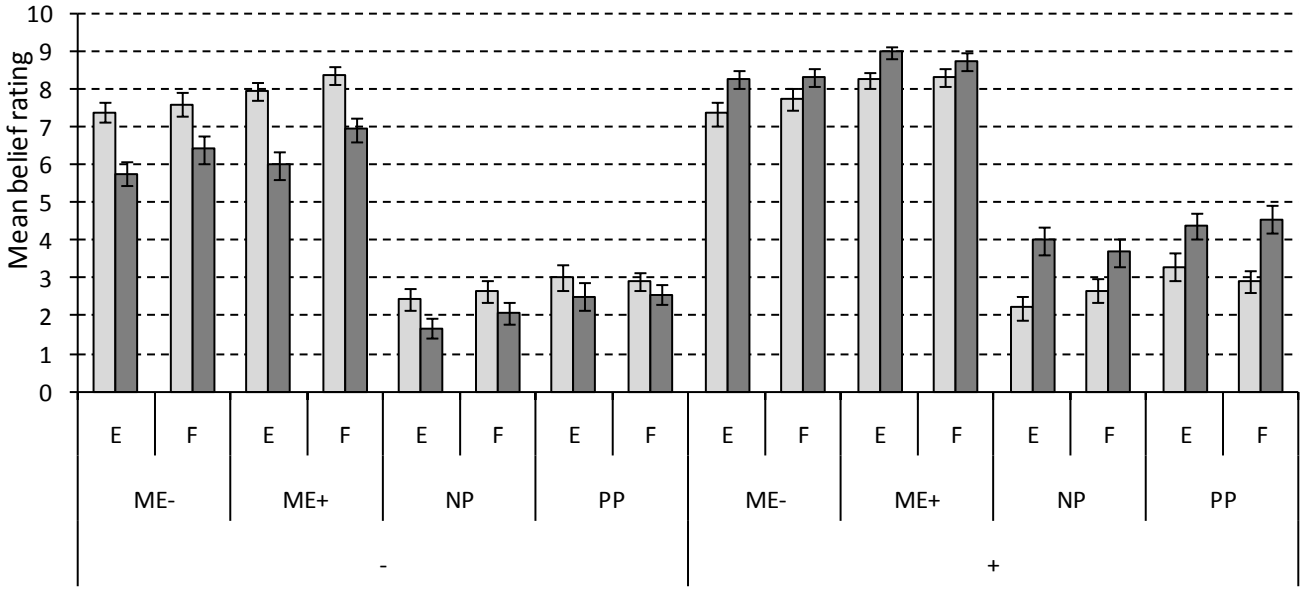


Figure 2: Mean belief ratings in conclusion statements, and standard errors for patterning-interaction and main-effect schemas (ME-, ME+, NP, PP) plotted against effect similarity (E: same effect, F: different effect) and evidence type ("-": disconfirming evidence, "+": confirming evidence). Light grey bars (left hand side) indicate ratings before the (dis)confirming instance was shown (*Rating 1*), dark grey bars (right hand side) show ratings after the (dis)confirming instance was presented (*Rating 2*).

Method

Participants A different sample of 48 University of Göttingen undergraduates participated.

Design We manipulated the same two independent variables as in Experiment 1. Additionally, we varied whether subscenario 2 contained confirming vs. disconfirming evidence for the to-be-judged conclusion statement. This

Results

The results are summarized in Figure 2. First, different assumptions about the prior probability of main effects vs. patterning interactions are evident in the much higher mean ratings in *Rating 1* in ME- and ME+ trials compared to NP

and PP trials ($F_{3,141}$=251.6, p<.001, η_p^2=.84). Thus, main effects were again assumed to be more likely than interactions.

Second, belief change (difference between *Rating 1* and *Rating 2* within conditions) was influenced by Cue Interaction ($F_{3,141}$=14.10, p<.001, η_p^2=.23) and type of additional evidence ($F_{3,47}$=124.42, p<.001). Planned contrasts revealed that in the case of confirming evidence, the increase in belief was stronger for patterning interactions than for main effects ($F_{1,47}$=11.04; p<.01); in the case of disconfirming evidence, the decrease in belief was stronger for main effects than for patterning interactions ($F_{1,47}$=40.06; p<.001), as predicted in (ii).

Finally, there was no main effect of the similarity factor on belief change ($F_{3,47}$<1, p=.45). This, however, was to be expected: While the confidence in the hypothesis should increase more after confirming evidence about the same effect than about a different effect, it should also *decrease* more after disconfirming evidence about the same effect than about a different effect (thus, on the level of the similarity factor both effects cancel out each other). This prediction, in turn, is reflected in the significant Evidence × Similarity interaction ($F_{3,47}$=82.39, p<.05, η_p^2=0.09) which is driven by the predicted specific differences (planned contrast[2]: $F_{1,47}$=4.43, p<.05). These results confirm prediction (iii).

General Discussion

We have presented two experiments which replicate and extend a previous study testing a rational model of belief updating (Griffiths et al., 2009). We showed again that people lacking specific causal knowledge may use knowledge about abstract causal schemas in inductive reasoning. Moreover, we found again that people find interactions less plausible than main effects, while, in line with Bayesian updating, evidence about a case of an interaction increases confidence more than evidence about main effects, which stays at a relatively high level. In the present study we elaborated our model to accommodate variations of similarity and cases of confirming versus disconfirming evidence.

The present research suggests a number of directions for future research. In the present experiments we have shown that the similarity of effect events influences inductive reasoning with both confirmatory and disconfirmatory evidence. It would be interesting to additionally explore the role of the similarity of the cues (A-D), which was only varied across experiments in the present paper. We expect that both the similarity of the cues and of the effect will equally contribute to similarity-related effects.

The present research used extremely abstract materials and a subset of possible interaction types. It might be interesting to look at differences between different causal schemas when more domain knowledge is allowed (see Waldmann, 2007, for other domain related schemas).

Finally, in the Introduction we have separated learning tasks from inductive reasoning tasks, but combinations are conceivable. Previous knowledge need not be stated as facts but can be presented in the form of statistical evidence (e.g., learning trials). It would certainly be interesting to develop a model of inductive reasoning that integrates prior beliefs about abstract and specific causal relations, similarity, and different types of evidence.

Acknowledgments

This research was supported by a research grant of the Deutsche Forschungsgemeinschaft (DFG Wa 621/20).

References

Feeney, A., & Heit, E. (2007). *Inductive reasoning: Experimental, developmental, and computational approaches.* New York, NY: Cambridge University Press.

Griffiths, O., Mayrhofer, R., Nagel, J., & Waldmann, M. R. (2009). Causal schema-based inductive reasoning. In N. Taatgen, H. van Rijn, L. Schomaker & J. Nerbonne (Eds.), *Proceedings of the Thirty-First Annual Conference of the Cognitive Science Society (pp. 691-696).* Austin, TX: Cognitive Science Society.

Kehoe, E. J. (1988). A layered network model of associative learning: Learning to learn a configuration. *Psychological Review, 95*, 411-433.

Kelley, H. H. (1972). *Causal schemata and the attribution process*. New York: General Learning Press.

Kemp, C., Goodman, N. D., & Tenenbaum, J. B. (2007). Learning causal schemas. *Proceedings of the Twenty-Ninth Annual Conference of the Cognitive Science Society* (pp. 64-70). Austin, TX: Cognitive Science Society.

Lucas, C. G., & Griffiths, T. L. (2010). Learning the form of causal relationships using hierarchical Bayesian models. Cognitive Science, 34, 113-147.

Novick, L. R., & Cheng, P. W. (2004). Assessing interactive causal influence. *Psychological Review, 111*, 455-485.

Rehder, B. (2007). Property generalization as causal reasoning. In A. Feeney & E. Heit (Eds.), *Inductive reasoning: Experimental, developmental, and computational approaches*. New York: Cambridge University Press.

Shanks, D. R., & Darby, R. J. (1998). Feature- and rule-based generalization in human associative learning. *Journal of Experimental Psychology: Animal Behavior Processes, 24*, 405-415.

Waldmann, M. R. (2007). Combining versus analyzing multiple causes: How domain assumptions and task context affect integration rules. *Cognitive Science, 31*, 233-256.

[2] $(M_{Same/Conf}-M_{Diff/Conf}) + (M_{Diff/Disconf}-M_{Same/Disconf})$=1.92

Causal Conditional Reasoning and Conditional Likelihood

Philip M. Fernbach (philip_fernbach@brown.edu)
Adam Darlow (adam_darlow@brown.edu)

Brown University, Department of Cognitive and Linguistic Sciences, Box 1978
Providence, RI 02912 USA

Abstract

We hypothesized that causal conditional reasoning reflects judgment of the conditional likelihood of causes and effects based on a probabilistic causal model of the scenario being judged. Although this proposal has much in common with Cummins' (1995) theory based on the number of disabling conditions and alternative causes, it takes more variables into account and therefore makes some differing predictions. To test this idea we collected judgments of the causal parameters of the conditionals and used them to derive predictions from a model with zero free parameters. We compared these predictions to Cummins' acceptability ratings and to analogous likelihood judgments that we also collected. The hypothesis was borne out for Affirming the Consequent and the analogous diagnostic likelihood judgments, where the model obtained close fits to both data sets. However, we found a surprising dissociation between Modus Ponens and judgments of predictive likelihood leading to a relatively poor fit to the Modus Ponens acceptability ratings. We propose an explanation for this in the discussion.

Key Words: Causal Conditional Reasoning, Conditional Probability, Reaction Time, Probabilistic Model, Modus Ponens, Affirming the Consequent

Causal Conditional Reasoning

When reasoning about deductive arguments people are biased to accept conclusions that are consistent with their beliefs and reject those that are inconsistent, regardless of argument validity (Evans, 2007). In a set of seminal papers, Cummins (1995; Cummins et al., 1991) showed that these belief biases follow systematic principles when people reason about conditional arguments with causal content. People judged the validity of four argument schemata: Modus Ponens (MP), Modus Tollens (MT), Denying the Antecedent (DA) and Affirming the Consequent (AC), though we focus on just MP and AC in this paper.

Despite MP being deductively valid and AC invalid regardless of content, Cummins predicted that for arguments where the antecedent is a cause of the consequent, acceptance rates for MP would be affected by the number of disabling conditions while AC would be affected by the number of alternative causes for the effect.

In the case of MP, thinking of a disabling condition provides a counterexample to the argument and hence may lead people to reject it. An example is given below. Cummins' predicted that (a) would be judged more acceptable than (b) because the conditional in (a) has fewer disablers; reasons why one could put fertilizer on plants and not have them grow quickly are more available than reasons why one could jump into a pool and not get wet.

(a) If Mary jumped into the swimming pool then she got wet.
Mary jumped into the swimming pool.
Therefore she got wet.

(b) If fertilizer was put on the plants then they grew quickly.
Fertilizer was put on the plants.
Therefore they grew quickly.

In the case of AC, alternative causes provide an alternative explanation for the effect and hence make the antecedent seem less necessary. For example Cummins predicted that (c) would be judged more acceptable than (d). It is hard to think of alternative causes for a gun firing besides the trigger being pulled but it is relatively easy to think of causes of wetness besides jumping into a swimming pool.

(c) If the trigger was pulled then the gun fired.
The gun fired.
Therefore the trigger was pulled.

(d) If Mary jumped into the swimming pool then she got wet.
Mary got wet
Therefore she had jumped into the swimming pool.

To test these ideas Cummins' asked one group of participants to spontaneously generate alternative causes and disabling conditions for a host of conditionals and then divided the conditionals into four groups of four conditionals each based on the number of alternatives and disablers (many alternatives, many disablers; many alternatives, few disablers; few alternatives, many disablers; few alternatives, few disablers). A different group was given the arguments based on the 16 conditionals and asked to judge the extent to which the conclusion could be drawn from the premise. Responses were on a 6 point scale from "very sure that the conclusion cannot be drawn" (-3) to "very sure that the conclusion can be drawn" (3). The results provided good support for both predictions.

A Causal Model Theory

Following Oaksford, Chater and Larkin (2000), if the conditional schemata are interpreted in terms of conditional probability, the acceptability of MP maps onto *P(Effect|Cause)* and AC to *P(Cause|Effect)*. Throughout the paper, we refer to *P(Effect|Cause)* as a *predictive* likelihood judgment and to *P(Cause|Effect)* as a *diagnostic* judgment.

By assuming the conditional scenarios approximate a noisy-or common effect model (Cheng, 1997) the expressions in (1) and (2) can be derived for MP and AC respectively (Fernbach & Darlow, 2009; Waldmann et al., 2008). The noisy-or model assumes that there are multiple independent causes for a given effect, each of which may or may not be effective on a given trial.

$$MP \approx P(Effect \mid Cause) = W_c + W_a - W_c W_a \quad (1)$$

$$AC \approx P(Cause \mid Effect) = 1 - (1 - P_c)\frac{W_a}{P_c W_c + W_a - P_c W_c W_a} \quad (2)$$

W_c is the causal power of the cause, the probability that the cause successfully brings about the effect (e.g. the probability that pulling the trigger causes the gun to fire), W_a is the combined strength of all alternative causes, equivalent to the probability of the effect in the absence of the cause (e.g. the probability of the gun firing given the trigger wasn't pulled) and P_c is the prior probability of the cause (e.g. the probability of the trigger being pulled).

According to the full probabilistic model MP increases with both the causal power of the cause and the strength of alternatives (because alternative causes raise the probability of the effect). However, in previous work, we have found that people are not sensitive to the strength of alternative causes when judging predictive likelihood despite its relevance (Fernbach, Darlow & Sloman, 2010). Thus, like Cummins we predicted no effect of W_a and our model for MP is given in (3).

$$MP \approx P(Effect \mid Cause, \sim Alternatives) = W_c \quad (3)$$

AC is a function of all three parameters. It increases with P_c and W_c and decreases with W_a.

Relation Between Cummins' Analysis and Model

According to the causal model the determinants of causal inferences, and hence MP and AC acceptability, are causal power, strength of alternatives and prior probability of the cause. The number of disablers and number of alternatives are factors in the first two parameters, respectively. Causal power is inversely related to the number of disablers. All else being equal, as the number of disablers increases, the probability that the cause fails to bring about the effect increases, corresponding to a decrease in causal power. Thus the model is consistent with the decrease in MP as number of disablers increases, as predicted and found by Cummins. However, not all disablers are equally likely or equally effective in preventing the effect. A single strong disabler could lead to a lower causal power than several weaker disablers, making number of disablers an imperfect predictor of causal power.

Similarly, the number of alternatives is a factor in strength of alternatives. All else being equal, as the number of alternatives increases so does the probability that they will bring about the effect. Therefore, the model predicts that AC will decrease with number of alternatives. As with disablers though, number of alternatives is only a partial predictor of strength of alternatives.

Despite these similarities, the model suggests that Cummins' analysis is incomplete because it only takes a single parameter into account for each judgment. The implication for MP is that its acceptability should increase with the strength of alternative causes but as discussed above we predicted no effect of alternative causes on MP. Our prediction for MP only differs from Cummins in that we expected W_c to provide a better fit than number of disablers.

The model identifies three factors relevant to the acceptability of AC arguments. First, according to the model the prior probability of the cause plays an important role in diagnostic strength. For instance, a cause that is very improbable is unlikely to have occurred relative to other more likely causes and is therefore not as good an explanation for the effect. The second factor is the overall strength of alternatives. This differs from the number of alternatives because not all alternative causes are created equal. In the causal model the strength of alternatives reflects the probability of the effect in the absence of the cause and thus is a joint function of the prior probabilities and causal powers of alternatives. For instance, even a large number of highly improbable or weak alternatives should have less effect on the judgment then a single probable, strong cause. Finally, causal power -- and hence disablers -- should have some influence on AC. All else being equal, if the causal power of the cause is higher, the cause is more likely responsible for the effect. Table 1 summarizes how our predictions differ from Cummins' theory.

Table 1: Best Predictors for MP and AC judgments and Predictive and Diagnostic Likelihood Judgments According to Cummins (1995) and According to our Model

	MP	*AC*
Cummins' Theory	No. of Disablers	No. of Alternatives
Causal Model	Causal Power (W_c)	Full Diagnostic Model
	Predictive Likelihood	*Diagnostic Likelihood*
Cummins' Theory	No Prediction	No Prediction
Causal Model	Causal Power (W_c)	Full Diagnostic Model

Qualitative Support for Causal model

Some trends appear in Cummins' (1995) data that are not predicted by her theory. One is that acceptability ratings of AC for conditionals with many alternative and few disablers were lower than those with many alternatives and many disablers. Both groups had many alternatives and thus should have yielded similar AC judgments according to Cummins. The difference was replicated by De Neys, Schaeken and D'ydewalle (2002) who found lower AC ratings for all few disabler items compared to many disabler items.

De Neys et al. (2002) proposed that when there are many disablers, they interfere with searching memory for alternatives, leading to the observed difference. A perusal of the individual conditionals suggests an alternative explanation based on the causal model. The two groups appear to vary not just in number of disablers but also in some of the factors that the probabilistic analysis says should affect diagnostic judgments. Specifically, the items that obtain low acceptability scores share the property that the cause is weak or improbable relative to the strength of alternatives (see Table 2). For instance, jumping into a swimming pool is improbable relative to other causes of wetness. Likewise, pouring water onto a fire is not the most

common cause of a campfire going out. On the contrary, the high ratings obtain for arguments in which the cause is strong and probable relative to alternatives. There may be many alternatives for a car slowing, but braking is likely the dominant cause. Likewise, studying hard is probably the strongest cause of doing well on a test. Thus, number of alternatives may be equated across groups, but diagnostic strength is not.

Table 2: Mean Acceptability of AC arguments for Two Groups of Conditionals from Cummins' (1995) Exp.1

Conditional	Acceptability (-3 to 3)
Many Alternatives, Many Disablers	
If fertilizer was put on the plants, then they grew quickly	1.00
If the brake was depressed, then the car slowed down	1.00
If John studied hard, then he did well on the test	1.50
If Jenny turned on the air conditioner, then she felt cool	1.08
Many Alternatives, Few Disablers	
If Alvin read without his glasses, then he got a headache	0.75
If Mary jumped into the swimming pool, then she got wet	0.25
If the apples were ripe, then they fell from the tree	1.00
If water was poured on the campfire, then the fire went out	-0.08

Another trend unexplained by her analysis is that few alternative conditionals obtained slightly higher MP judgments than many alternative conditionals despite being equated across number of disablers. Again, the probabilistic analysis suggests why this may be so. Several of the many alternative items have somewhat low causal powers (e.g. 'if the apples were ripe then they fell from the tree') while virtually all of the few alternative items have very high causal powers (e.g. 'if the gong was struck then it sounded.'). Thus, while number of disablers was equated across groups, causal power may have varied leading to differing MP judgments.

Experiment

To test whether the causal model accounts for the causal conditional acceptability ratings we collected judgments of the relevant parameters: the prior probability of the cause (P_c), the causal power of the cause (W_c) and the strength of alternatives (W_a) for Cummins' (1995) conditionals. Using these judgments we derived predictions with zero free parameters to which we compared Cummins' acceptability ratings.

Another implication of our argument is that judgments of the conditional probability of effects and causes should be similar to Cummins' acceptability ratings and should also be accounted for by the causal model. Thus, we collected predictive and diagnostic conditional probability judgments from a second group of participants. We also collected reaction times for these judgments. De Neys et al. (2002) showed that reaction times for causal conditionals basically supported Cummins' analysis. Collecting reaction times with materials phrased in conditional likelihood language allowed us to verify and extend these findings.

Method

Participants 133 Brown University students were approached on campus and participated voluntarily or participated through the psychology research pool in return for class credit.

Design, materials and procedure All experimental conditions used questions based on the 16 conditionals from Cummins' (1995) experiment 1. We therefore adopted Cummins' 2 (number of alternatives; few/many) X 2 (number of disablers; few/many) design with four conditionals in each condition. Judgments were on a 0 ('impossible') to 100 ('definite') scale.

17 Participants provided judgments of the prior probabilities (P_c) and strength of alternatives (W_a) for the 16 conditionals. The questions were split onto two pages with all of the P_c questions on the first page and all of the W_a questions on the second page. The order of questions was randomized on each page. For each question we first stated the conditional and then asked the relevant likelihood question. Examples of P_c and W_a questions are given in (e) and (f) respectively.

(e) If John studied hard then he did well on the test.
How likely is it that John studied hard?

(f) If John studied hard then he did well on the test.
John did not study hard. How likely is it he did well on the test?

A minority of participants interpreted the conditional statement in the P_c questions as indicating that the cause was present and therefore gave ratings of 100 for all of the P_c questions. We removed these responses from the dataset for all subsequent analyses.

An additional 21 participants judged causal power (W_c). Methods were identical except that there was just one page of questions. An example of a W_c question is given in (g).

(g) How likely is it that John studying hard for the test causes him to do well?

95 participants provided predictive and diagnostic likelihood judgments, fully within-participant. Each of these participants therefore answered 32 questions, one predictive and one diagnostic for each conditional. In order to avoid any reaction time differences due to reading time, the wordings of the questions were modified such that each had between 13 and 15 words and between 65 and 75 characters and such that the mean number of words and characters was equated across the four groups of conditionals. Examples of predictive and diagnostic questions are given in (h) and (i):

(h) John studied hard. How likely is it that he did well on the test?

(i) John did well on the test. How likely is it that he studied hard?

This part of the experiment was administered on a computer in the lab. For each question, participants input their answer using the number keys and hit 'return' to move to the next question. Reaction times were measured from the moment the question appeared on the screen to when the participant

hit 'return'. Order of questions was randomly determined for each participant.

Parameter Judgments and Modeling Results

For the following tests we collapsed over conditionals and compared participant means, using Bonferroni correction to control family-wise error rate. As expected, W_a was judged higher for many alternative items compared to few alternative items ($t(16)=13.4$, $p<0.001$) and didn't vary across few and many disablers ($t(16)=1.4$, *ns*).

W_c also varied across the number of alternatives manipulation; W_c was judged higher for few alternative items (M=83.4) compared to many alternative items (M=73.9), $t(20)=4.8$, $p<0.001$). This was not intended by Cummins, but confirmed our intuitions about the unexplained trend in MP; weak alternative items seemed to have lower causal powers despite being equated across number of disablers. Surprisingly, W_c did not vary across the many/few disablers manipulation ($t(20)=1.2$, *ns*) suggesting that number of disablers and causal power were not as closely linked as we expected. The low correlation between number of disablers and W_c ($r=-0.11$, *ns)* also supported this conclusion. P_c did not vary across either manipulation.

Applying the Model Simply computing Equations 2 and 3 using item means would have been inappropriate because the parameter judgments were collected between participants. We therefore used a sampling procedure to generate model predictions. For each conditional we took 10,000 samples each of W_a, P_c and W_c uniformly and randomly from participant responses, and calculated Equations 2 and 3 for each set of samples. We therefore generated 10,000 samples of each probability for each conditional and then took the mean over samples for each conditional as the output of the model. Reruns of the model yielded only negligible differences.

Fits to AC and Diagnostic Judgments Figure 1a depicts Cummins' acceptability ratings for AC on the X-axis plotted against model fits (Equation 2) on the Y-Axis for each of the 16 conditionals, along with the least squares regression line. Figure 1b shows diagnostic judgments plotted against model fits. The model predictions were highly correlated with both Cummins' acceptability ratings (AC) ($r=0.87$, $p<0.001$) and the diagnostic judgments (D) ($r=0.93$, $p<0.001$). To test whether the model is a better predictor of AC and D than the number of alternatives, we performed hierarchical multiple regression analyses of AC and D responses using the model predictions and the number of alternatives as predictors. The model accounted for a significant amount of unique variance beyond what number of alternatives accounted for, both for AC ($F(1,14)=10.7$, $p<0.01$) and for D ($F(1,14)=38.4$, $p<0.001$). Number of alternatives did not account for any unique variance for AC ($F(1,14)=0.24$, *ns*) or for D ($F(1,14)=0.46$, *ns*).

Fits to MP and Predictive Judgments Figure 1c depicts Cummins' acceptability ratings for MP plotted against model fits (equal to W_c according to Equation 3). Figure 1d shows predictive judgments plotted against model fits. Surprisingly, MP ratings and predictive judgments were not highly correlated ($r=0.30$, *ns*), and each was correlated with a different independent variable. MP ratings were significantly correlated with number of disablers ($r=0.53$, $p=0.035$)but not with the model ($r=0.39$, *ns*). Conversely, predictive judgments were highly correlated with the model ($r=0.81$, $p<0.001$) but not with number of disablers ($r=0.04$, *ns*). As predicted, alternative strength did not add any explanatory power; the full model was poorer than W_c at accounting for both MP and predictive judgments.

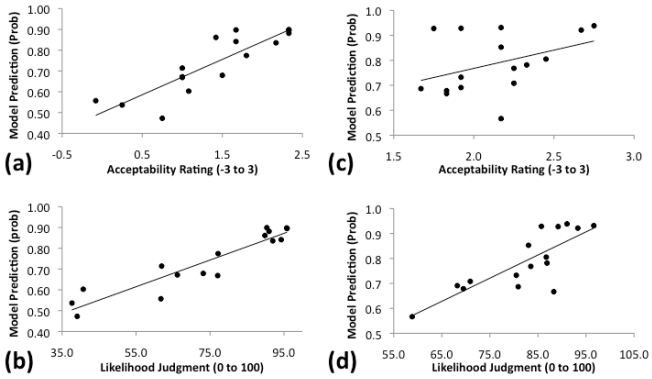


Figure 1: (a) Model fits against Cummins' AC acceptability ratings. (b) Model fits against diagnostic likelihood judgments. (c) Model fits against Cummins' MP acceptability ratings. (d) Model fits against predictive likelihood judgments.

Reaction Time Results

For the sake of concision, the analyses of the predictive and diagnostic judgments are described in the appendix and only the reaction times results are presented in this section. All statistical tests on reaction times used a log transform to normalize the data. Outliers were removed by eliminating all trials that fell more than four standard deviations above or below the participant's mean reaction time. Additionally any trial faster than 1 second was removed.

The reaction time results are depicted in Figure 2. The cleaned data were subjected to a 2 (direction of inference) X 2 (number of alternatives) X 2 (number of disablers) repeated measures ANOVA. There was a main effect of direction of inference; prediction (M=5.88 s) was faster than diagnosis (M=6.21 s), $F(1,95)=25.1$, $p<0.001$. There was also a significant interaction between number of alternatives and direction of inference, $F(1,95)=4.0$, $p<0.05$. No other main effects or interactions were significant.

The interaction between strength of alternatives and direction of inference was driven by diagnostic judgments being faster for items with few alternatives (M = 6.09 s) than for items with many alternatives (M=6.32 s), $t(94)=1.95$, $p=0.05$. Predictive judgments showed no difference in reaction time across the number of alternatives manipulation, $t(94)=0.61$, *ns*.

Number of disablers had no effect on reaction times for predictive judgments (t(94)=1.2, *ns)*. Since W_c accounted for the predictive judgments better than number of alternatives, we suspected it might also yield reaction time differences. To test this we split the conditionals at the median based on W_c and compared reaction times. Confirming the prediction, predictive judgments were faster for items with high W_c (M=5.71 s) than for items with low W_c (M=6.05 s), t(94)=4.19, p<0.0001.

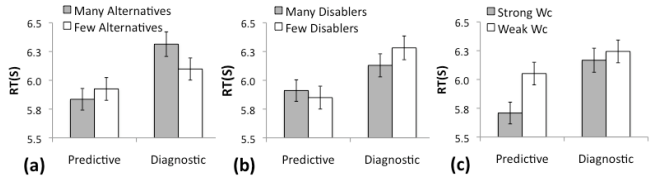


Figure 2: Reactions Times for Predictive and Diagnostic Judgments by (a) number of alternatives, (b) number of disablers and (c) strength of W_c

General Discussion

Summary and Interpretation of Results

Model Fits The diagnostic model achieved very good fits to both Cummins' AC data and our diagnostic likelihood judgments with zero free parameters. It also explained more variance than the single parameter number of alternatives. This confirmed the qualitative analysis indicating that AC judgments were sensitive not just to number of alternatives, but also to the other factors in the causal model in approximately the right way. The model also accounted for the previously unexplained trend in Cummins' AC data for higher AC ratings with more disablers. Altogether, it seems that when judging AC for causal conditionals, people are actually judging the likelihood of the cause (premise) given the effect (conclusion).

The model also matched the predictive judgments closely and differences in W_c explained the previously unexplained trend in Cummins' MP judgments for higher MP judgments with fewer alternatives, a pattern that also showed up in the predictive likelihood judgments (see appendix). But the model didn't match the MP data that well and in fact was slightly worse than the number of disablers at accounting for the variance. Additionally, number of disablers was a remarkably poor predictor of W_c judgments. This was surprising because we expected causal power to vary inversely with number of disablers.

Reaction Times The reaction time data yielded three noteworthy findings: First, predictive judgments were faster than diagnostic ones. This corroborates De Neys et al. (2002) who found that MP was faster than AC and it supports the claim that reasoning from cause to effect is easier in general than reasoning from effect to cause (Tversky & Kahneman, 1982). This difference likely reflects the time it takes to consider alternative causes and prior probability in diagnostic judgment.

Second, diagnostic judgments were faster with few alternatives. This also corroborates De Neys et al. (2002). It implies that searching for alternative causes takes time. It could also reflect the fact that when alternative causes are very weak the judgment is very high and may not require as much thought to calculate. Predictive judgments showed no reaction time differences across number of alternatives. This is more evidence that people don't think of alternatives when making predictions (Fernbach, Darlow & Sloman, 2010).

Finally, we found no reaction time differences for many versus few disablers. This failed to corroborate De Neys et al. (2002) who found that MP was faster for few versus many disablers. We did however find an effect of W_c on reaction times. Prediction was faster for high versus low W_c.

Explaining MP

Both the model fitting and reaction times imply dissociation between how people judged MP and how they judged predictive likelihood. Predictive likelihood judgments and reaction times were explained by differences in W_c but were uncorrelated with number of disablers. Conversely, number of disablers was slightly better at accounting for Cummins' (1995) MP acceptability ratings than W_c and also yielded reaction time differences for MP in De Neys et al.'s (2002) study. This leaves three open questions: First, why is number of disablers such a poor predictor of W_c? Second, why is W_c better at accounting for predictive likelihood judgments and reaction times? Third, why is it worse at accounting for MP?

A speculative answer to the first two questions comes from the possibility that when making predictive likelihood judgments people represent causal systems in terms of their normal, common or prototypical components. If asked to list disablers they may be able to come up with a relatively large number, some of them being very uncommon or atypical. But when asked to judge causal power or make a prediction they think only of the most important disablers. The 'depressed brake' provides a good example. It is not too hard to come up with disablers for why brakes would fail to slow a car (e.g. cut brake lines) but none of them is common. Thus, while number of disablers is relatively high, many of those disablers make a small impact on actual causal power and may have no effect on people's estimates of causal power. On this account, low causal power might still correlate with slower reaction time on the assumption that examples with a greater number of typical or high probability disablers yield lower W_c judgments, lower predictive judgments, and take longer to reason about.

This leaves the question of why W_c fails to account for MP judgments and reaction times, while number of disablers is somewhat better. We don't have a conclusive answer to this question, but we suspect it may be due to people using a mixture of strategies when judging MP. In a deductive context, people reason about MP more naturally than other conditional schemata (Johnson-Laird & Byrne, 2002). This suggests that some participants may be engaging in a different kind of thinking when judging MP in comparison to the other schemata. Perhaps more abstract

thinking leads to rejection of MP based on the ability to think of specific counterexamples without regard to their probability, in which case the number of disablers may be more important than W_c. This is consistent with work by Verschueren, Schaeken and d'Ydewalle (2005) showing two processes in causal conditional reasoning: A relatively quicker intuitive process that arrives at judgments that are highly correlated with conditional probability and a relatively slower, analytic process that correlates with number of alternatives or disablers. Of course, it's important not to jump to firm conclusions on the basis of so few examples (the poor fit to MP was primarily driven by 4 data points). Future work should aim to corroborate the differences in ratings and reaction times for MP versus predictive likelihood with a larger number of well-controlled items.

Conclusions

Our work provides some evidence in favor of the conditional probability approach to conditional reasoning (Oaksford & Chater, 2001, 2003; Over et al., 2007). One caveat to this is that the causal model we propose is incorrect in some important senses. People tend to neglect the strength of alternatives when making predictions, and while aggregate data are fit really well by the diagnostic model, individual data are less consistent. This suggests that people are not actually computing probabilities. It is more natural to think of the model as a computational solution that people only approximate. The literature on probabilistic causal reasoning tends to focus primarily on computational models like this to the detriment of process level implementations. The focus on semantic memory models in the causal conditional reasoning literature is admirable, but the downside of these models is that, as our work shows, people are sophisticated causal reasoners. Simple memory models based on the number of alternatives or disablers won't suffice. A complete model requires mechanisms for judging prior probability, for integrating over the strengths and probabilities of alternative causes, for judging causal power and for combining these various pieces of information in a reasonable way. These processes undoubtedly rely on retrieval from semantic memory – our reaction time data is strong evidence of that – but no current memory model can accommodate the balance of empirical evidence. Exploring how people construct their causal models from remembered alternatives, disablers and other parameters thus offers a promising avenue for future research.

Acknowledgments

This work was supported by a Galner Dissertation Fellowship and an APA Dissertation Research Award to the first author. We thank Steve Sloman, David Over and Dinos Hadjichristidis for helpful discussion and are especially grateful to Denise Cummins for digging up her data from 1995.

References

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review,* 104, 367–405.

Cummins, D. D., Lubart, T., Alksnis, O. and Rist, R. (1991) Conditional reasoning and causation. *Memory and Cognition,* 19 (3), 274-282.

Cummins, D. D (1995) Naïve theories and causal deduction. *Memory and Cognition,* 23 (5), 646-658.

De Neys, W., Schaeken, W. & D'ydewalle, G. (2002). Causal conditional reasoning and semantic memory retrieval: A test of the semantic memory framework. *Memory and Cognition*, 30 (6), 908-920.

Evans, J. ST. B. T. (2007). *Hypothetical thinking: Dual processes in reasoning and judgement*. London: Taylor & Francis.

Fernbach, P. M., Darlow, A. & Sloman, S. A. (2010). Neglect of alternative causes in predictive but not diagnostic reasoning, *Psychological Science*, In Press.

Fernbach, P. M. & Darlow, A. (2009). Causal asymmetry in inductive judgments. *Proceedings of the Thirty-First Annual Conference of the Cognitive Science Society.*

Johnson-Laird, P. N. & Byrne, R. M. J. (2002). Conditionals: A theory of meaning, pragmatics, and inference. *Psychological Review*, 109 (4), 646-678.

Oaksford, M., & Chater, N. (2001). The probabilistic approach to human reasoning. *Trends in Cognitive Science*, 5, 349 – 357.

Oaksford, M., & Chater, N. (2003). Conditional probability and the cognitive science of conditional reasoning. *Mind and Language*, 18 (4), 359 – 379.

Oaksford, M., Chater, N. & Larkin, J. (2000). Probabilities and polarity biases in conditional inference. *Journal of Experimental Psychology: Learning Memory and Cognition,* 26, 883-889.

Over, D., Hadjichristidis, C., Evans, J. St BT. Handley, S. J., & Sloman, S. A. (2007). The probability of causal conditionals. *Cognitive Psychology*, 54, 62-97.

Tversky, A. & Kahneman, D. (1982). Causal schemas in judgements under uncertainty. In D. Kahneman, P. Slovic & A. Tversky (eds.), *Judgement under uncertainty: Heuristics and biases (117-128).* Cambridge: Cambridge University Press.

Verschueren, N., Schaeken, W. & d'Ydewalle, G. (2005). A dual process specification of causal conditional reasoning. *Thinking & Reasoning*, 11 (3), 239-278.

Waldmann, M. R., Cheng, P. W., Hagmayer, Y., & Blaisdell, A. P. (2008). Causal learning in rats and humans: a minimal rational model. In N. Chater, & M. Oaksford (Eds.), *The probabilistic mind. Prospects for Bayesian Cognitive Science* (pp. 453-484). Oxford: University Press.

Appendix

The predictive and diagnostic judgments were subjected to a 2 (direction of inference) X 2 (number of alternatives) X 2 (number of disablers) repeated measures ANOVA. All of the main effects and two-way interactions were significant ($p<0.01$).

Further post hoc tests were performed on predictive and diagnostic judgments separately. Diagnostic judgments were sensitive to number of alternatives with higher judgments for the items with few alternatives ($M=90.7$) than for the items with many alternatives ($M=57.3$), $t(94)=27.9$, $p<0.001$. Diagnostic judgments also varied across number of disablers, with higher judgments for many disablers ($M=78.1$) than few disablers ($M=70.1$), $t(94)=8.9$, $p<0.001$.

As suggested by the differing W_c judgments, predictive judgments also varied across the number of alternatives; Few alternative items ($M=87.8$) yielded higher diagnostic judgments than those with many alternatives ($M=76.3$), $t(94)=6.0$, $p<0.001$. Predictive judgments did not vary with the number of disablers ($t<1$, *ns*). We also tested whether predictive judgments varied with the strength of W_c by dividing the 16 conditionals into two equal groups based on W_c and comparing predictive judgments. As expected, conditionals with high W_c obtained higher predictive judgments ($M=89.1$) than those with low W_c ($M=75.2$), $t(94)=7.0$, $p<0.001$.

Causal Models Interact with Structure Mapping to Guide Analogical Inference

Hee Seung Lee (heeseung@ucla.edu)
Keith J. Holyoak (holyoak@lifesci.ucla.edu)
Department of Psychology, University of California, Los Angeles
Los Angeles, CA 90095 USA

Abstract

We recently proposed a theoretical integration of analogical transfer with causal learning and inference (Lee & Holyoak, 2008). A Bayesian theory of learning and inference based on causal models (Lee, Holyoak & Lu, 2009) accounts for the fact that judgments of confidence in analogical inferences are partially dissociable from measures of the quality of the mapping between source and target analogs. The integrated theory postulates a dual role for causal relations, which can guide both analogical mapping and also subsequent inferences about the target. It follows that depending on whether or not a mapping is structurally ambiguous, dropping a preventive cause from the target can either decrease or increase confidence in the same analogical inference. We report an experiment that yielded data in close agreement with predictions of the Bayesian theory. These results provide further support for the importance of integrating analogical transfer with the broader framework of causal models.

Keywords: analogical inference; causal models; mapping; Bayesian modeling.

Introduction

Analogical transfer plays a key role in scientific reasoning (Dunbar & Fugelsang, 2005). Indeed, in some areas of science in which experimental research is impossible, such as historical ethnography, analogy may provide the only viable mechanism for evaluating hypotheses. Talalay (1987) gives the example of interpreting the function of strange clay fragments discovered in Neolithic Greek sites: individual female legs, apparently never attached to torsos, that had been manufactured in pairs and later broken apart. The best clues to their function have come from other cultures in which similar tokens are known to have served to seal contracts and provide special evidence of the identity of the bearer (in feudal China, for example, a valuable piece of jade would be broken in two to mark a contract between a master and his vassal, with each keeping one piece so they could later be matched). Here the known function in a source domain has a *causal* connection to the form of relevant artifacts, and the ethnographer makes the analogical inference that a similar cause may have operated in the target domain (see Bartha, 2010).

The general question faced by a reasoner using analogy to make inferences is: Given prior knowledge at various levels of abstraction, including one or more examples that serve as source analogs, what is the probability that any potential inference about a target analog is true? For analogical inferences that involve empirical claims about the world (e.g., scientific hypotheses), answering this question depends on at least two basic subprocesses: deciding how the causally-relevant elements of the source analog(s) relate to elements of the target (structure mapping), and using the corresponding causal relations suggested for the target to estimate the probabilities of potential inferences about the target (causal inference). In the above ethnography example, mapping is required to relate the two pieces of a broken jade object to the two parts of a broken piece of pottery; causal inference is required to evaluate the probability that the clay fragments could achieve a function analogous to that achieved by a divided jade object.

Both structure mapping and causal inference have received considerable attention within cognitive science. Mapping has been the central focus of many models of analogical reasoning (e.g., Falkenhainer, Forbus & Gentner, 1989; Holyoak & Thagard, 1989; Hummel & Holyoak, 1997). Causal learning and inference have also been studied extensively, with theoretical work largely based on the framework of *causal models* (Pearl, 1988; Waldmann & Holyoak, 1992; Waldmann, Holyoak & Fratianne, 1995). The power PC theory (Cheng, 1997) provides a quantitative explanation of how the strengths of probabilistic causal relations can be learned from contingency data. More recently, this theory has been extended based on a Bayesian formulation (Griffiths & Tenenbaum, 2005; Lu et al., 2008). Theories of category-based induction have also been enriched by adopting the framework of causal models (e.g., Ahn, 1999; Kemp, Goodman & Tenenbaum, 2007; Sloman, 1994; Rehder, 2009).

Integrating analogical inference with causal models

We have proposed that a more complete understanding of analogical transfer requires specifying the role played by causal models (Lee & Holyoak, 2008; Lee, Holyoak & Lu, 2009). Figure 1 schematizes causal models for a source (left) and target analog (right). The nodes represent variable causes (C) and effects (E). The superscripts (^{S}, ^{T}) indicate the source and the target, respectively. The links represent the causal structure (only linked nodes have direct causal connections). The vectors w_i represent the causal polarity (generative or preventive) and the causal strength for links. A key assumption is that analogical transfer involves using causal knowledge of the source to develop a causal model of the target, which can in turn be used to derive a variety of inferences about the values of variables in the target. Causal relations in Bayesian causal models can carry information about existence of causal links (e.g., causal structure) and distributions of causal strength, as well as about the generating function by which multiple causes combine to influence effects.

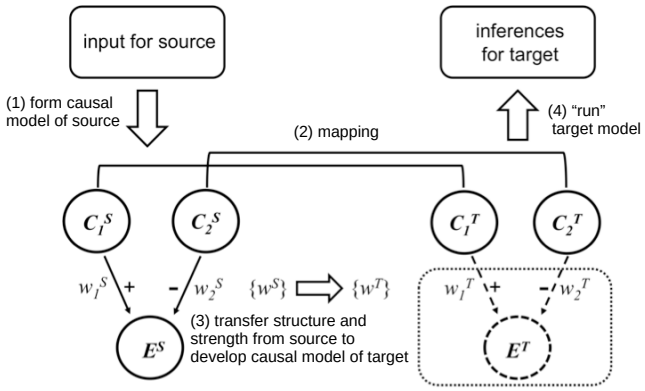


Figure 1: Framework for analogical transfer based on integrating mapping with causal models. Dotted lines indicate knowledge transferred from source to target (see text).

In our theory, the first step in analogical inference is to learn a causal model of the source. The source model is then mapped to the initial (typically impoverished) representation of the target. Based on the mapping, the causal structure and strengths associated with the source are transferred to the target, creating or extending the causal model of the latter. The model of the target can then be "run", using causal reasoning to derive inferences about the values of endogenous variables in the target. Accordingly, as summarized in Figure 1, the four basic components in analogical inference are: learning of a causal model for a source (step 1); assessment of the analogical mapping between the source and a target (step 2); transfer of causal knowledge from the source to the target based upon the analogical mapping to construct the causal model of the target (step 3); and inference based on the causal model of the target (step 4).

To generate predictive inferences (from causes to their effect), let $\boldsymbol{C}^S$ denotes the information that the source has a background generative cause, B^S, plus additional generative and preventive causal factors. (In this paper, vectors are indicated by bold font.) $\boldsymbol{C}^T$ provides analogous information about possible causes in the target. In predictive inference, the model estimates the probability of an effect occurring in the target, $E^T = 1$, based on initial information about the source, ($\boldsymbol{C}^S$, E^S), and the target, $\boldsymbol{C}^T$. The unknown causal strength of the target is represented by $\boldsymbol{w}^T$. The basic equation for predictive inference is

$$
\begin{aligned}
&P(E^T \mid \mathbf{C}^T, E^S, \mathbf{C}^S) \\
&= \sum_{\mathbf{w}^T} P(E^T, \mathbf{w}^T \mid \mathbf{C}^T, E^S, \mathbf{C}^S) \\
&= \sum_{\mathbf{w}^T} P(E^T \mid \mathbf{w}^T, \mathbf{C}^T) \sum_{\mathbf{w}^S} [P(\mathbf{w}^T \mid \mathbf{w}^S, E^S, \mathbf{C}^S, \mathbf{C}^T) P(\mathbf{w}^S \mid \mathbf{C}^S, E^S)]
\end{aligned}
$$

(Equation 1)

where the rightmost term on the right side of the equation, $P(\boldsymbol{w}^S|\boldsymbol{C}^S, E^S)$, captures the learning of a source model from observed contingency data (step 1 in Figure 1). Recent computational studies have developed detailed models that estimate distributions of causal strength by combining priors and observations (Griffiths & Tenenbaum, 2005; Lu et al., 2008). The middle term, $P(\boldsymbol{w}^T \mid \boldsymbol{w}^S, \boldsymbol{C}^S, \boldsymbol{C}^T)$, quantifies knowledge transfer based upon analogical mapping (steps 2 and 3 in Figure 1). We model the probability of transfer as

$$
\begin{cases}
P(w_i^T = w_i^S) = 1, & \text{if } C_i^T \text{ matches } C_i^S \\
P(w_i^T = w_i^S) = 0, & \text{otherwise}
\end{cases}
$$

(Equation 2)

where C_i^S and C_i^T represent the ith cause variables in the source and target, respectively. If the mapping of C_i^T to an element in the source is ambiguous (as will be the case for the materials we use in the experiment reported here), then Eq. 2 will simply sum over the transfer result obtained when C_i^T matches each of the alternative source elements, weighted by the probabilities of each of these possible matches.

The leftmost term, $P(E^T|\boldsymbol{w}^T, \boldsymbol{C}^T)$, uses knowledge from analogical transfer and observations about the presence of causal factors in the target to estimate the probability of the effect in the target (step 4 in Figure 1). For binary variables, this probability can be directly computed using the Bayesian extension of the power PC theory (Cheng, 1997; Griffiths & Tenenbaum, 2005; Lu et. al., 2008).

Mapping and Causal Inference

Although causal relations have sometimes been assumed to have special importance in guiding mapping (Holyoak, 1985; Hummel & Holyoak, 1997; Winston, 1980), models of analogical transfer have generally treated inference as a direct extension of mapping. In contrast, our causal-model approach postulates a deeper role for causal knowledge in transfer (Lee & Holyoak, 2008).

The present study sought to demonstrate a direct interaction between mapping and causal inference, which is predicted by our Bayesian theory. According to the integrated theory, a causal relation in the target potentially plays a dual role: it first may guide structure mapping between the source and target; then if mapping succeeds, it will also guide causal inference based on the resulting causal model of the target. In the present study we investigated analogical transfer when the mapping between the source and target was in some cases structurally ambiguous (cf. Spellman & Holyoak, 1996).

More specifically, we examined how presence or absence of a certain causal relation (preventive cause in this study) in the target might increase or decrease inductive strength depending on whether the structural mapping is clear or ambiguous. The source analog included a preventive cause, which might or might not be also included in the target. When the mapping is clear, the expected effect of inclusion of the preventive cause is evident in that presence of a preventive cause will decrease inductive strength in target, as shown in previous studies (e.g., Lee & Holyoak, 2008). However, when the mapping is ambiguous, and if the preventive cause is able to resolve the mapping ambiguity, the expected result will be reversed. The materials were designed so that when the mapping was ambiguous, the

inclusion of the preventive cause in the target provided sufficient structural information to resolve the ambiguity, and hence allow transfer of causal structure from source to target. Conversely, if the preventive cause were omitted from the target, the mapping ambiguity would be left unresolved, thereby impairing transfer of a causal model from source to target. In such situations our Bayesian model predicts that including the preventive cause in the target will actually *increase* inductive support for the occurrence of the effect that it actually tends to prevent. No previous analogy model predicts this type of interactive impact of causal and structural constraints on analogical transfer.

Experiment: Can a Preventive Cause Either Decrease or Increase the Judged Strength of the Same Analogical Inference?

Method

Participants Forty-five undergraduate students at the University of California, Los Angeles participated in the experiment to fulfill a course requirement. Each participant was randomly assigned to one of eight different sets of materials generated for counterbalancing purposes.

Design and materials The source story described a biochemist's findings about an imaginary liver disease called "tibulosis", found in rats. The disease had two different subtypes, "Type A" and "Type B", described as being caused by different factors and exhibiting quite different symptoms. The scientist had identified several factors that determine whether or not rats might develop either Type A or Type B tibulosis. For each type, certain hormones, enzymes, and antibodies were involved. Participants were asked to carefully study the biochemist's findings using a verbal description and diagram presented in the booklet in order to determine what characteristics are likely to produce or prevent the development of each type of the disease. Participants were then given descriptions of human patients with a liver disease, and asked to apply what they had learned about tibulosis in rats to judge the probability that the human patients had tibulosis Type A or Type B.

In the source, the two disease subtypes were designed to create a potential mapping ambiguity. The two types had identical causal structures except for the names of causal elements, but with one critical structural difference involving a preventive cause. Each source included two generative causes, one preventive cause, and an effect (consistent with a common effect model; Waldmann & Holyoak, 1992). The two generative causes were certain types of hormones and enzymes and the preventive cause was a certain type of antibody. In each case the preventive cause was narrow in scope (Carroll & Cheng, 2009), in that it served to stop the causal impact of one of the two generative causes but not the other. The description of the causal structure for Type A tibulosis was as follows:

Factors influencing development of Type A tibulosis
Hormone A tends to stimulate the production of enzyme A, and vice versa.
Hormone A tends to PRODUCE Type A tibulosis.
Enzyme A also tends to PRODUCE Type A tibulosis.
The immune system sometimes PRODUCES antibody A in response to enzyme A, but never in response to hormone A.
Antibody A tends to PREVENT enzyme A from producing Type A tibulosis. However, antibody A provides no protection against the direct effect of hormone A on Type A tibulosis.

To aid comprehension of the causal structure, a schematic diagram was provided right below the description. Figure 2 depicts the causal structure for Type A, described above. Hormone A and enzyme A are two generative causes that both tend to produce the effect, type A tibulosis. Antibody A is a preventive cause with narrow scope, which prevents enzyme A (but not Hormone A) from producing the effect. The B subtype was very similar to the A subtype described above, except that the effect was "type B tibulosis" (rather than type A), and the names of the hormone, enzyme and antibody were also B. The critical structural difference between the two sources was that in the B version, the immune system was described as producing antibody B in response to hormone B, but never in response to enzyme B (opposite to the situation in the A version); furthermore, antibody B tended to prevent the effect of hormone B (not enzyme B).

In the target story, participants read reports about human patients who might have a human form of Type A or Type B tibulosis. Examination reports for seven patients were constructed. Each examination report included information about a hormone, an enzyme, and (in some versions) an antibody found in each patient. A 2 x 2 within-subjects design was employed, resulting in four basic versions of the target descriptions. The first independent variable was whether the target description was *specific* or *generic*. In the specific condition, specific names of the hormone, enzyme, and antibody (e.g., hormone A, enzyme A, antibody A) were explicitly stated in the description of the patient report provided in the target. Given that these names matched those for one of the two subtypes described in the source,

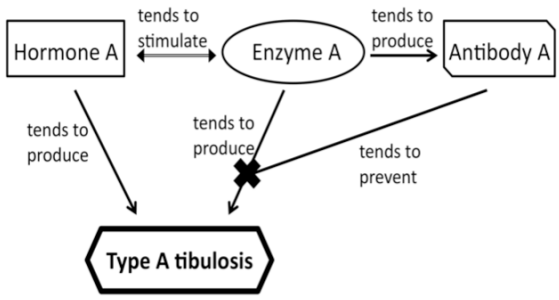


Figure 2: An example of a causal structure for one of two disease subtypes included in the source analog used in the experiment.

the mapping of the human case to Type A (or B) tibulosis as described in the source was accordingly transparent.

In contrast, in the generic condition, specific names of the hormone, enzyme, and antibody were not provided. Instead, each was simply described by its general categorical description (i.e., hormone, enzyme, and antibody). Thus in the absence of additional structural information, there was no basis for preferentially mapping the description of the factors observed in the human patient onto those related to Type A versus Type B tibulosis in rats.

The above manipulation of the target description was crossed with a second independent variable, *presence* or *absence* of the preventive cause (antibody) in the description of the human patient. Recall that the critical structural difference between Type A and Type B tibulosis as described in the source was that for Type A, the *enzyme* produced the antibody, which then acted to block the *enzyme's* impact; whereas for Type B, it was the *hormone* that produced the antibody, which then acted to block the *hormone's* impact. In the P-present condition, the target description included analogous information about the human case. For example, in the specific, P-present condition, the description might state:

> Hormone A and enzyme A are present, and each stimulates production of the other.
>
> The immune system produced antibody A in response to the enzyme (but not the hormone).

More critically, in the generic, P-present condition, the description stated:

> A hormone and an enzyme are present, and each stimulates production of the other.
>
> The immune system produced an antibody in response to the enzyme (but not the hormone).

Note that even though no specific names are provided, the above generic, P-present description (based on the second statement in the description) provides structural information sufficient to disambiguate the mapping between the human case in the target and the disease descriptions for rats as stated in the source. That is, only Type A tibulosis involves an antibody produced in response to an enzyme, which then blocked the enzyme's effect. Any of the major models of structure mapping (e.g., Falkenhainer et al., 1989; Hummel & Holyoak, 1997) would be able to use the structural information provided in the generic, P-present condition to resolve the potential ambiguity and identify a determinate mapping between the disease described in the target and one of the two subtypes described in the source. Accordingly, this condition would be essentially identical to the specific, P-present condition if participants could resolve mapping ambiguity using the preventive cause.

In the P-absent versions (both specific and generic), the second statement in the relevant description was simply replaced with "no antibody is present". Critically, in the generic, P-absent condition, no information was provided that could possibly serve to resolve the structural ambiguity inherent in the mapping; hence the target case could be mapped to either Type A or Type B in the source. If a preventive cause plays a dual role in analogical transfer, as the integrated theory postulates, then in this experiment its inclusion will have a paradoxical influence on the judged probability of an effect in the target. Specifically, given a specific description of the target, inclusion of the preventive cause will *decrease* the judged probability of the effect (by acting as a preventer within the causal model of the target); but given a generic description of the target, its presence will *increase* the judged probability of the same effect (by serving to disambiguate the mapping so that a causal model of the target can in fact be constructed).

For each condition except the generic, P-absent condition, two patient reports were constructed, resulting in seven patient reports in total. For each of the first three conditions, one of the two patient reports supported mapping to type A, and the other supported mapping to type B. Because the generic, P-absent condition did not support mapping to one type over the other, only one version of this patient report could be constructed. Two different sets of materials were constructed by counterbalancing whether the hormone or the enzyme produced an antibody in type A and in type B. Within each set, four different orders of targets were constructed, resulting in eight versions of materials in total.

Procedure Participants were given a booklet that included the source story, the target story, and a series of inference tasks. First, participants read the source story about a biochemist's findings about a new liver disease found in rats, and studied what factors were likely to produce or prevent the development of two types of the disease based on the verbal descriptions and diagrams.

In the generic conditions (but not in the specific conditions), a mapping task was included before the inference task to check if the potential mapping ambiguity was resolved or not. This task required identifying the generic hormone as "hormone A", "hormone B", or "can't tell". The analogous question was also asked about the generic enzyme. For the analogical inference task, participants were given the examination reports for seven different patients. For each, participants were asked to judge how likely it was that the patient had each disease type. To answer each question, they were to imagine there were 100 cases with the same known characteristics as for the specific case, and judge how many of these 100 cases would be expected to have each type of the disease.

Results and Discussion

On the mapping task, 33 of the 45 participants reported the structurally-justified mappings for the hormone and enzyme in the generic, P-present condition. The other 12 participants gave a variety of responses in this critical condition. We analyzed the results both including and excluding data from those participants who made mapping errors. As the basic pattern was the same in both sets of analyses, we will report the analysis including all participants.

For each patient case, participants estimated both the probability that the patient had Type A of the disease and the probability that the patient had Type B. The format

encouraged participants to treat the two types as mutually exclusive, and assignments of Type A versus Type B were fully counterbalanced across conditions. To code the responses on the inference task, we defined the "correct" disease type as that supported by the preferred mapping in the three unambiguous conditions (specific, P-present; specific, P-absent; and generic, P-present). This same disease type (either A or B) was defined as "correct" in the matched generic, P-absent condition (in which neither answer was inherently preferred).

The mean rated probability of the correct effect for each of the four conditions is shown in Figure 3 (left). These data were analyzed using a 2x2 analysis of variance in which both target description (specific vs. generic) and presence of the preventive cause (P-present vs. P-absent) were within-subjects variables. A significant main effect of specificity of the target description was obtained, $F(1, 44) = 123.09$, $MSE = 302.52$, $p < .001$, in that inference strength was significantly higher when the description was specific ($M = 83.03$, $SD = 16.09$) than when it was generic ($M = 54.26$, $SD = 17.70$). The main effect of presence of the preventive cause was not significant, $F < 1$. Most importantly, a significant interaction was obtained between target specificity and presence of preventive cause, $F(1, 44) = 79.66$, $MSE = 281.49$, $p < .001$, implying that the presence of a preventive cause had a different impact on analogical inference depending on the ambiguity of the mapping. When the description of the target was specific so that the mapping to one of the disease types in the source was transparent, participants gave significantly higher estimates of the probability of the correct effect in the P-absent condition ($M = 92.42$, $SD = 13.99$) than in the P-present condition ($M = 73.63$, $SD = 28.52$), $t(44) = 4.02$, $p = .001$. This result replicates previous findings (Lee & Holyoak, 2008; Lee et al., 2009), in that dropping a preventive cause from the target increased the strength of a predictive inference. In contrast, when the target description was generic, the effect of including the preventive cause was reversed. The estimated probability of the correct effect was now higher in the P-present condition ($M = 67.19$, $SD = 29.63$), where the preventive cause served to disambiguate the mapping, than in the P-absent condition, ($M = 41.33$, $SD = 23.89$), where the mapping was structurally indeterminate, $t(44) = 4.28$, $p < .001$.

Comparison of Bayesian Model to Human Data

We used our Bayesian model to provide a more quantitative account of our findings. The basic model was identical to that outlined by Lee et al. (2009), as summarized earlier. To fit the specific causal structures used in the present experiment, people were assumed to have no prior knowledge about causal structure or strength of the source; hence the stated causal relations were assigned a uniform strength distribution ranging between 0 and 1. Because no further information about the causal strengths was provided in the source, these distributions remained uniform (no updating based on examples), so that in effect only causal structure, not strength, was available to be transferred to the target. Based on Equation 2, causal links with uniform strength distributions were directly transferred from the source to the target analog when the mapping was determinate. Thus in the three unambiguous conditions (specific, P-present; specific, P-absent; and generic, P-present), the causal model for the correct effect was transferred to the target. In the ambiguous condition (generic, P-absent), the model summed over predictions made for each of the potential mappings to the two alternative sources, weighting them equally.

Given the general assumptions of the Bayesian version of the power PC theory (Lu et al., 2008), the predicted probability of the correct effect in the target, given the source, can be derived analytically without estimating any free parameters. To do so, the functional form of the preventive cause (a noisy-AND-NOT function) was applied in a manner that reflected the appropriate narrow scope of the preventer (Carroll & Cheng, 2009). The influences of the causes were integrated sequentially. After applying a noisy-AND-NOT function to integrate the influence of the preventer with that of its related generative cause, a noisy-OR function was applied to combine this intermediate result with the influence of the other generative cause and an assumed background cause. Figure 3 (right) depicts the parameter-free predictions of the Bayesian model. The quantitative fit was good, $r(2) = .93$. When data from just those participants who solved the mapping task correctly were modeled, the fit increased slightly, $r(2) = .94$. The model captures the trade-off that arises in the generic, P-present condition, where the presence of the preventer exerts a positive influence on analogical transfer by guiding the mapping, but then reduces transfer somewhat by acting to prevent the effect within the causal model created for the target. Also, the model makes identical predictions for the specific, P-present and generic, P-present conditions. This pattern is consistent with human response patterns in that most participants gave the same ratings for these two conditions. In the generic, P-absent condition, due to the unresolved mapping ambiguity, the probability that the effect occurs in the target is predicted by the sum of its

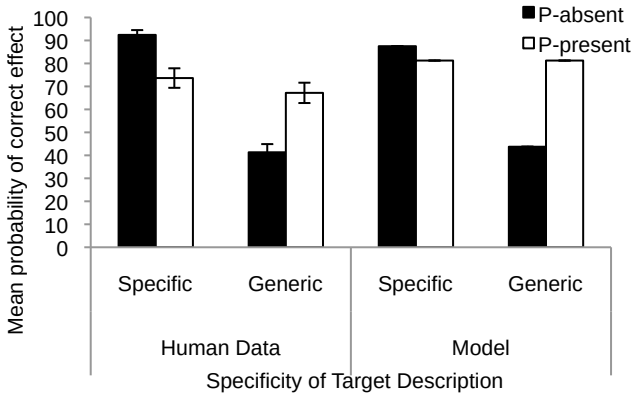


Figure 3: Mean probability of the correct effect in each condition. Left: human data; right: predictions derived from Bayesian model. Error bars represent 1 standard error of the mean.

probabilities based on each possible source, weighted by the probability of the mapping between the target and each source (equally weighted with probability of .5). Again, this prediction appears to be consistent with human response patterns, which primarily consisted of giving equal probability ratings for the two alternatives (i.e., 50/50) in this condition.

Conclusion

Our experiment demonstrated that the inclusion of a preventive cause in the target had an opposite impact on the judged probability of an effect in target, depending on whether or not the source-target mapping was ambiguous in the absence of the preventer. When the mapping was transparent (because objects in the target were described in the same specific terms as the corresponding objects in the source), inclusion of the preventive cause in the target decreased inference strength, as observed previously (Lee & Holyoak, 2008). However, when the mapping was potentially ambiguous (because objects in the target were described in generic terms), and the preventive cause provided structural information sufficient to disambiguate the mapping, then inclusion of the preventive cause in the target increased inference strength.

This pattern of interaction was predicted by our Bayesian theory (Lee et al., 2009), adding to the empirical and theoretical evidence supporting the importance of integrating theories of structure mapping with the framework provided by causal models (Waldmann & Holyoak, 1992). This type of integrated theory may provide deeper insight into many aspects of analogical inference, including its role in both the generation and evaluation of scientific hypotheses.

Acknowledgments

Preparation of this paper was supported by ONR grant N000140810186.

References

Ahn, W. (1999). Effect of causal structure on category construction. *Memory & Cognition, 27*, 1008-1023.

Bartha, P. (2010). *By parallel reasoning: The construction and evaluation of analogical arguments*. Oxford, UK: Oxford University Press.

Carroll, C. D., & Cheng, P. W. (2009). Preventative scope in causal inference. In N. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31th Annual Conference of the Cognitive Science Society* (pp. 833-838). Austin, TX: Cognitive Science Society.

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review, 104*, 367-405.

Dunbar, K., & Fugelsang, J. (2005). Scientific thinking and reasoning. In K. J. Holyoak & R. G. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning* (pp. 705-725). Cambridge, UK: Cambridge University Press.

Falkenhainer, B., Forbus, K. D., & Gentner, D. (1989). The structure mapping engine: Algorithm and examples. *Artificial Intelligence, 41*, 1-63.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology, 51*, 334-384.

Holyoak, K. J. (1985). The pragmatics of analogical transfer. In G. H. Bower (Ed.), *The psychology of learning and motivation*, Vol. 19 (pp. 59-87). New York: Academic Press.

Holyoak, K. J., & Thagard, P. (1989). Analogical mapping by constraint satisfaction. *Cognitive Science, 13*, 295-355.

Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review, 104*, 427-466.

Kemp, C., Goodman, N. D., & Tenenbaum, J. B. (2007). Learning causal schemata. In D. S. McNamara & G. Trafton (Eds.), *Proceedings of the Twenty-ninth Annual Conference of the Cognitive Science Society* (pp. 389-394). Austin, TX: Cognitive Science Society.

Lee, H. S., & Holyoak, K. J. (2008). The role of causal models in analogical inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*, 1111-1122.

Lee, H. S., Holyoak, K. J., & Lu, H. (2009). Integrating analogical inference with Bayesian causal models. In B. Kokinov, D. Gentner, & K. J. Holyoak (Eds.), *New frontiers in analogy research: Proceedings of the Second International Conference on Analogy* (pp. 300-309). Sofia, Bulgaria: New Bulgarian University Press.

Lu, H., Yuille, A., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2008). Bayesian generic priors for causal learning. *Psychological Review, 115*, 955-982.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference.* San Mateo, CA: Morgan Kaufmann.

Rehder, B. (2009). Causal-based property generalization. *Cognitive Science, 33*, 301-343.

Sloman, S. A. (1994). When explanations compete: The role of explanatory coherence on judgments of likelihood. *Cognition, 52*, 1-21.

Spellman, B. A., & Holyoak, K. J. (1996). Pragmatics in analogical mapping. *Cognitive Psychology, 31*, 307-346.

Talalay, L. E. (1987). Rethinking the function of clay figurine legs from Neolithic Greece: An argument by analogy. *American Journal of Archaeology, 91*, 161-169.

Waldmann, M. R., & Holyoak, K. J. (1992). Predictive and diagnostic learning within causal models: Asymmetries in cue competition. *Journal of Experimental Psychology: General, 121*, 222-236.

Waldmann, M. R., Holyoak, K. J., & Fratianne, A. (1995). Causal models and the acquisition of category structure. *Journal of Experimental Psychology: General, 124*, 181-206.

Winston, P. (1980). Learning and reasoning by analogy. *Communications of the ACM, 23*, 689-703.

Symbolic Reasoning in Spiking Neurons: A Model of the Cortex/Basal Ganglia/Thalamus Loop

Terrence C. Stewart (tcstewar@uwaterloo.ca)
Xuan Choo (fchoo@uwaterloo.ca)
Chris Eliasmith (celiasmith@uwaterloo.ca)
Centre for Theoretical Neuroscience, University of Waterloo
Waterloo, ON, N2L 3G1

Abstract

We present a model of symbol manipulation implemented using spiking neurons and closely tied to the anatomy of the cortex, basal ganglia, and thalamus. The model is a general-purpose neural controller which plays a role analogous to a production system. Information stored in cortex is used by the basal ganglia as the basis for selecting between a set of inferences. When an inference rule is selected, it commands the thalamus to modify and transmit information between areas of the cortex. The system supports special-case and general-purpose inferences, including the ability to remember complex statements and answer questions about them. The resulting model suggests modifications to the standard structure of production system rules, and offers a neurological explanation for the 50 millisecond cognitive cycle time.

Keywords: decision making; neural production system; neural engineering; cognitive architectures

Introduction

The primary goal of our ongoing research is the creation of a biologically realistic neural cognitive architecture. Such an architecture would provide an explicit and quantitative connection between cognitive science and neuroscience. Bridging these fields leads to benefits in both directions; aspects of a cognitive theory can predict and be constrained by neurological details, and the neurological details can in turn identify important modifications to cognitive theory.

In this paper, we present a model of sequential symbolic reasoning implemented using 373,000 simulated spiking neurons. The connectivity of these neurons, their neural parameters, and their associated neurotransmitters are fixed based on neurological evidence from the basal ganglia, thalamus, and cortex. By adjusting the synaptic connections of neurons at the inputs and outputs of the basal ganglia, we can define the inferences that the system will follow. Since these rules can be adjusted for a wide variety of IF-THEN symbol manipulation tasks, we believe that our model is the first biologically realistic general-purpose neural controller that can play a role analogous to a production system.

The model involves the basal ganglia, the thalamus, and various cortical areas. The cortex holds a variety of information about the current situation, such as visual input and the contents of working memory. The basal ganglia performs action selection, taking information from the cortex to determine which of the rules is most appropriate to use in the current situation. This choice is sent to the thalamus, which acts as a routing system, implementing the effects of those rules by transferring information between cortical areas. As the information stored in the cortical areas changes, different actions will be selected in turn, allowing for controlled and organized sequences of actions.

To present this model, we first provide a brief description of the Neural Engineering Framework (NEF; Eliasmith and Anderson, 2003), a general method for organizing realistic spiking neuron models so as to represent and transform information. This is used to derive the optimal synaptic connections (under neurological constraints) for creating our model. Next, we introduce Vector Symbolic Architectures (VSAs; Gayler, 2003), a method for efficiently encoding symbolic structures as high-dimensional fixed-length vectors. This is used to encode structured information in the cortex and to represent the IF-THEN rules themselves.

Given these tools, we then define the three anatomical components of our model (cortex, basal ganglia, and thalamus). This includes specifying the neurological parameters of the neurons involved, such as the neurotransmitters used. This is important for providing accurate timing predictions from our model, since various neurotransmitters have varying characteristic time constants.

We demonstrate our model performing three separate tasks: repeating the alphabet, repeating the alphabet starting from a particular letter, and answering questions using working memory. For each of these tasks we use exactly the same neural model; the only differences are the sensory inputs to the system.

Finally, we provide two conclusions that connect cognitive theory and neuroscience. First, we show that particular types of IF-THEN rules are more efficient to implement in spiking neurons, leading to a possible modification of standard production system-based theories. Second, we show that the time needed to select an action is determined primarily by the re-uptake rate of the neurotransmitter GABA in the basal ganglia, thus providing a neurological explanation for the 50-millisecond cognitive cycle time commonly found in behavioural results.

Neural Engineering Framework

To build a complex neural model, we need a method for determining how neurons can represent and transform information. We use the Neural Engineering Framework (NEF; Eliasmith and Anderson, 2003), which generalizes established findings on how sensory and motor neurons represent multidimensional information. This allows us to treat a group of neurons as representing a single vector of arbitrary length. By adjusting the connectivity between

groups of neurons, we can indicate how these representations should be changed over time.

The basic assumption of the NEF is that within a neural group, each neuron has a preferred value $\boldsymbol{e}$ (for ***e***ncoding) to which it responds most strongly (i.e. fires most quickly). As the difference between the actual value and the preferred value increases, this firing rate will decrease. If the value to be represented by the neurons is $\boldsymbol{x}$, this behaviour can be captured in terms of the amount of ionic current J flowing into the neuron given by Equation 1. Adjusting the neuron gain α, the background input current J_{bias}, and the preferred direction vector $\boldsymbol{e}$ allows us to capture a wide range of known neural tuning curves.

$$J = \alpha \boldsymbol{e} \cdot \boldsymbol{x} + J_{bias} \quad (1)$$

In the simplest case, 100 neurons could represent a 100 dimensional vector $\boldsymbol{x}$ by having each $\boldsymbol{e}$ be a different unit vector in each of the 100 dimensions. This would provide a completely local representation of each value in the vector. More realistically, 100 neurons could represent one or two dimensions by having $\boldsymbol{e}$ values chosen randomly (i.e. uniformly distributed around the unit hypersphere in that many dimensions). This approach has been observed in numerous areas of visual and motor cortex (e.g. Georgopoulos et al., 1986). The advantage of having more neurons than there are dimensions is that the amount of representational error can be controlled. Neurons are highly stochastic devices, but we have previously shown that the overall error is inversely proportional to the number of neurons per dimension (Eliasmith & Anderson, 2003).

Using Equation 1 to set the amount of input current to a particular neuron to represent a particular value, we can use existing models of neuron behaviour to determine the resulting spike times. There are an extremely wide variety of suitable neuron models, from Hodgkin-Huxley-type models up to extremely detailed compartmental models. For this model, we use a standard Leaky Integrate-and-Fire model, where input current causes voltage inside the neuron to gradually build up until it reaches a threshold, at which point it fires, producing a spike. Thus, given a particular vector, we can determine the resulting sequence of spikes.

We can also perform the opposite operation: given a sequence of spikes we can estimate the original vector. As shown elsewhere (Eliasmith & Anderson, 2003), this can be done by deriving the decoding vectors $\boldsymbol{d}$ as per Equation 2, where a_i is the average firing rate for neuron i with a given vector $\boldsymbol{x}$, and the integration is over all values of $\boldsymbol{x}$.

$$\boldsymbol{d} = \Gamma^{-1}\Upsilon \quad \Gamma_{ij} = \int a_i a_j dx \quad \Upsilon_j = \int a_j \boldsymbol{x}\, dx \quad (2)$$

The resulting vectors $\boldsymbol{d}$ can be used to determine an estimate of the represented value using Equation 3, where $h(t)$ is the current produced in a post-synaptic neuron by the pre-synaptic neuron firing at time t=0, and $t_{i,n}$ is the time that the ith neuron fired for the nth time.

$$\hat{\boldsymbol{x}}(t) = \sum_{i,n} \delta(t - t_{i,n}) * h_i(t)\boldsymbol{d}_i = \sum_{i,n} h(t - t_{i,n})\boldsymbol{d}_i \quad (3)$$

This is an estimate that varies over time based on the individual spikes. Importantly, it is the *optimal* estimate when under the constraint that the estimate must be built by linearly adding the effects of the post-synaptic currents caused by each spike. This is the constraint for other neurons receiving these spikes, so Equation 3 gives the optimal reconstruction of the vector by another neuron.

As a consequence of this, the decoding vectors $\boldsymbol{d}$ provide an extremely important tool that is at the heart of the Neural Engineering Framework. We can use $\boldsymbol{d}$ and $\boldsymbol{e}$ to derive *optimal synaptic connection weights* to perform particular mathematical manipulations on the encoded information. If one group of neurons represents $\boldsymbol{x}$ and we want another group to represent some particular linear transformation of this value (i.e. $\boldsymbol{y}=\boldsymbol{M}\boldsymbol{x}$), then we simply set the synaptic connection weights w as per Equation 4.

$$w_{ij} = \alpha_j \boldsymbol{e}_j \boldsymbol{M} \boldsymbol{d}_i \quad (4)$$

For nonlinear functions, we can modify Equation 2 to produce decoding vectors $\boldsymbol{d}^{f(\boldsymbol{x})}$ that optimally approximate any nonlinear function $f(\boldsymbol{x})$, as shown in Equation 5.

$$\boldsymbol{d}^{f(\boldsymbol{x})} = \Gamma^{-1}\Upsilon \quad \Gamma_{ij} = \int a_i a_j dx \quad \Upsilon_j = \int a_j f(\boldsymbol{x})\, dx \quad (5)$$

This approach allows us to create complex neural models where we directly derive the necessary synaptic connection weights, rather than relying on a particular learning rule.

Vector Symbolic Architectures

While the NEF provides a method for representing vectors, in order to implement a cognitive model we need to represent complex symbol-like structures. That is, while we might be able to say that one particular vector represents the concept of a square, another vector represents a triangle, and another represents a particular colour, this does not address the question of how we can represent "a blue circle and a red square".

A general approach to this problem is to use a Vector Symbolic Architecture (VSA; Gayler, 2003). There are three core ideas for all VSAs. First, each symbol is represented by a particular high-dimensional vector. For our purposes, we randomly choose these vectors, but they could also be selected based on semantic and sensory knowledge. Second, two vectors can be combined by superposition ($+$) to produce a new vector that is *similar to* both of the original vectors. Third, two vectors can be combined by binding ($\otimes$) to produce a new vector that is *dissimilar* to both of the original vectors.

This binding operation can be reversed by binding with the inverse of a vector (*), such that $\mathbf{A}\otimes\mathbf{B}\otimes\mathbf{B}^* \approx \mathbf{A}$. These operations are similar to standard addition and multiplication in terms of being associative, commutative, and distributive.

For our model, we chose a particular VSA known as Holographic Reduced Representations (HRRs; Plate, 2003). For this, superposition is performed by vector addition and the binding operation is circular convolution. These operations can be efficiently implemented in spiking neurons using synaptic connections calculated using the NEF (Eliasmith, 2005) and Equations 4 and 5, above.

With such a system we can represent symbol trees by combining superposition and binding. For example, we can find a vector to represent "a blue circle and a red square" by performing the following calculation:

$$\textbf{blue} \otimes \textbf{circle} + \textbf{red} \otimes \textbf{square}$$

The result is a single vector of the same dimensionality as the vectors for the basic symbols (**blue**, **red**, **square**, etc.). This one vector can be interpreted as a representation of the entire structure because it is possible to extract the original components. For example, to determine which object is red, we take the whole vector and bind it with the inverse of **red**.

$$\begin{aligned}&(\textbf{blue} \otimes \textbf{circle} + \textbf{red} \otimes \textbf{square}) \otimes \textbf{red}^{*} \\ &= \textbf{blue} \otimes \textbf{circle} \otimes \textbf{red}^{*} + \textbf{red} \otimes \textbf{square} \otimes \textbf{red}^{*} \\ &\approx \textbf{blue} \otimes \textbf{circle} \otimes \textbf{red}^{*} + \textbf{square}\end{aligned}$$

The result is a vector that is similar to **square**, but is not exactly the same since it has an additional term superposed on it. Due to the properties of the binding operation, however, $\textbf{blue} \otimes \textbf{circle} \otimes \textbf{red}^{*}$ will be a vector that is highly dissimilar to all of the original symbols, and can be treated as randomly distributed noise. We have previously shown how spiking neuron models can remove this noise (Stewart, Tang, & Eliasmith, 2009).

The Model

Basal Ganglia

The basal ganglia is generally believed by both neuroscientists (e.g. Redgrave et al., 1999) and cognitive scientists (e.g. Anderson et al., 2004) to be responsible for action selection. That is, given a wide variety of possible options as to what to do next, a single one must be chosen. This can be thought of as a winner-take-all mechanism: each option will have a numerical value indicating how relevant (or how beneficial) each action is in the current context, and the best of these should be chosen. Although winner-take-all mechanisms are common in neural models, there are few that adhere to the biological constraints of the basal ganglia, and none we are aware of that use realistic spiking neurons.

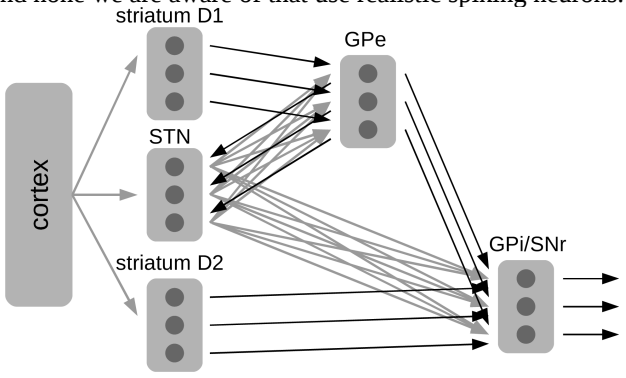


Figure 1: Basal ganglia model with three possible actions. Light lines are excitatory connections. Dark lines are inhibitory (based on Gurney et al., 2001, Figure 5).

While we have previously investigated simple mutual inhibition approaches for winner-take-all (Stewart & Eliasmith, 2009), for our current model we adapt work by Gurney, Prescott, and Redgrave (2001). As shown in Figure 1, the D1 cells in the striatum inhibit corresponding cells in the globus pallidus internal (GPi) and substantia nigra reticulata (SNr), while the subthalamic nucleus (STN) sends a broad excitatory signal to the GPi/SNr and globus pallidus external (GPe). The GPe and the D2 cells in the striatum act as a control signal on the excitation from the STN, adjusting it so that the correct amount of excitation is provided to select a single action. Each of these connections is well-documented anatomically, and the model's behaviour matches neurological results in rats and monkeys both with and without particular lesions (Gurney et al., 2001).

However, the Gurney et al. model uses idealized piecewise-linear non-spiking neurons that respond instantly without any random variation to changes in their inputs. We thus adapt their model, replacing individual idealized neurons with groups of realistic leaky-integrate-and-fire (LIF) spiking neurons. For our neurons, the membrane time constant (τ_{RC}; controlling the amount of current leaking out of the neuron) was fixed at 20ms, and the α and J_{bias} values were randomly chosen constrained by the reported response properties given by Gurney et al., including background firing rates of 60-80Hz and maximum firing rates of 400Hz. All synaptic connections were derived using Equation 4. We use 20 neurons to replace one ideal neuron (circle in Figure 1), so 100 neurons are needed per possible action.

The behaviour of this model is shown in Figure 2. The inputs to the model (top) are the desirability of three different actions. The firing response of the output of the basal ganglia (bottom) is shown as these inputs change over time. As in the actual basal ganglia, the output is inhibitory, so an action is selected by *turning off* the appropriate output neurons, stopping them from performing their inhibition. It should be noted that this output lags behind the input due to the time constants of the post-synaptic current caused by different neurotransmitters. In this case, the excitatory connections use glutamate with AMPA receptors (2ms; Spruston et al., 1995), and the inhibitory connections use GABA (10ms; Gupta et al., 2000).

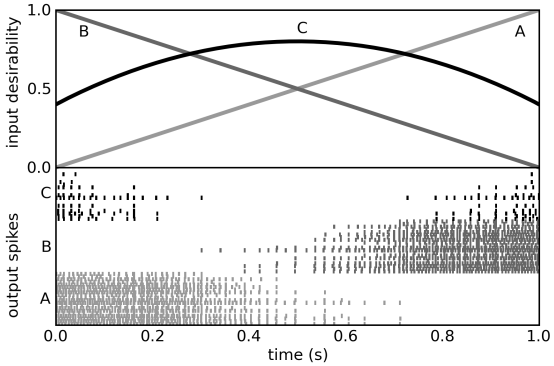


Figure 2: Inputs and outputs (GPi) of our basal ganglia model. The largest valued input consistently causes the corresponding output neurons to stop spiking.

Cortex

For the tasks under consideration in this paper, we need a visual area (for representing the current visual scene), a motor area (for producing outputs from the model), and a working memory (for storing a statement and questions to be answered). Each of these is implemented as 10,000 spiking neurons, storing a 250 dimensional VSA vector as per the NEF. We present stimuli to our model by injecting current into the visual area (**V** in Figure 3) using Equation 1. We can examine the contents of any area of the cortex by decoding the activation (Equation 3) and measuring the similarity (dot product) between the resulting vector and an ideal calculated vector. The closer this value is to 1.0, the more accurate the representation.

To perform general purpose tasks (such as question answering), our model contains two working memory areas: **A** and **B**. In order to maintain information over time, these areas contain connections back to themselves as per Equation 4 where $\boldsymbol{M}$ is the identity matrix. This forms the basis of an integrator model of memory, which has previously been used to model somatosensory working memory (Singh & Eliasmith, 2006). Areas **A** and **B** are also connected to two other neural groups **C** and **D** such that **C=A⊗B** and **D=A⊗B***. These connection weights are defined using Equation 5, where $f(x)$ is the circular convolution (see Eliasmith, 2005 for details). This allows the system to compute the VSA operations that are needed to perform symbol manipulation.

Thalamus

The only mechanism in our model for modifying the contents of the working memory areas and the motor areas is the thalamus. If the thalamic areas are all zero then no information is transferred between cortical areas. If the thalamic area corresponding to working memory **A** is set to some value (via the basal ganglia), then this value will be sent to cortical area **A**, using synaptic connections from Equation 4 with $\boldsymbol{M}$ as the identity matrix. Crucially for information transfer, if the thalamic area controlling the connection between **V** and **A** is set to **X**, then the value **V⊗X** will be sent to **A**.

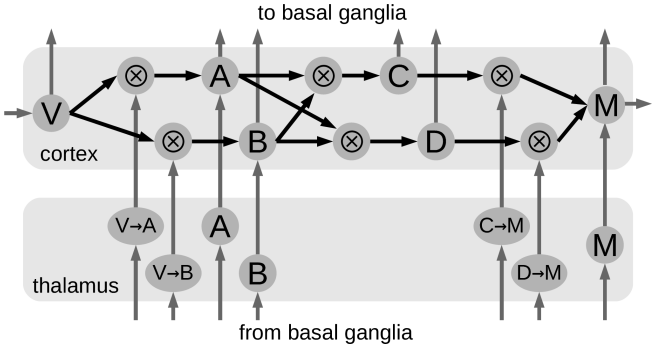


Figure 3: Thalamus and cortex model. Circles are 10,000 neurons representing 250 dimensional vectors (V=vision; M=motor; A,B,C,D=working memory). ⊗ are 40,000 neurons computing the binding operation.

Modelled Tasks

Fixed Sequences of Actions

The simplest task to perform with this model is sequentially going through a list of items, such the alphabet. We implement this by defining 25 rules of the following form:

IF working memory contains **letter+A**

THEN set working memory to **letter+B**

We create the IF portion of a rule by setting the synaptic connections between the working memory area of cortex and the striatum and sub-thalamic nucleus. Each component of the basal ganglia has a group of neurons corresponding to each rule (the dark circles in Figure 1). We set the input synaptic weights using Equation 4, where $\boldsymbol{M}$ is the vector corresponding to the IF portion of the rule (**letter+A**).

To implement the THEN portion of the rule, we set the synaptic connections at the output of the basal ganglia. In this case, we create a group of neurons that connect to the thalamic neurons that feed to working memory. We again use Equation 4 to set these weights, with $\boldsymbol{M}$ set to be the vector corresponding to **letter+B**. We then connect the group of neurons in the GPi that correspond to this rule to these new neurons. Because GPi is inhibitory, this connection will cause the new neurons to not fire at all, except in the case that the action selection system in the basal ganglia chooses this particular action. In that case, the inhibition will be turned off (as those GPi neurons will stop spiking), allowing **letter+B** to be sent to working memory. This in turn will cause the next rule to be selected, and so on. It should be noted that our model does not yet include the phonological loop, so any timing influence it may have on producing this sequence is not taken into account.

To test the model, we initialize it by forcing current into the working memory neurons as per Equation 1 such that they will represent **letter+A**. After this, all subsequent activity is due to the interconnections between neurons. Figure 4 shows the model correctly following the alphabet sequence. From the spiking pattern we see that the correct action for each condition is successfully chosen by turning off the appropriate inhibitory neurons in the GPi.

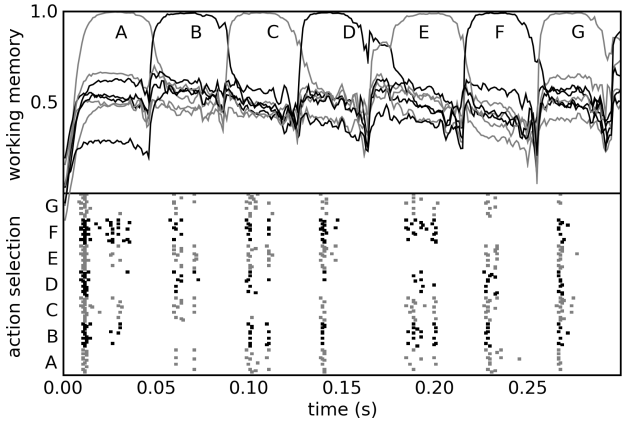


Figure 4: Contents of working memory (top) and spiking output from GPi indicating the action to perform (bottom).

Variables and Generic Rules

The previous section demonstrated that our model is capable of implementing rules where a specific pattern is sent to a specific part of cortex. While these sorts of rules may account for some kinds of highly specialized behaviour, most symbolic cognitive architectures assume that it is possible to have *general-purpose* rules. That is, these rules can contain variables, such as the following, where **?X** represents an unknown variable:

IF visual cortex contains **letter+?X**

THEN set working memory to **letter+?X**

The presence of this sort of rule in addition to the ones in the previous section would allow the model to start going through the alphabet starting from any letter. We would simply present the particular letter we wanted it to start from to the visual cortex (**letter+F**) and it would copy this value to working memory and continue from there.

While the above method is the standard approach for expressing this sort of rule, in order to implement it in our model, we need to slightly reformulate it as the following:

IF visual cortex contains **letter+?X**

THEN copy visual cortex to working memory

This rule has exactly the same effect as the first one. To implement it, we use the same approach as in the previous section. The synaptic connection weights for the inputs to the basal ganglia are set using Equation 4 with $\boldsymbol{M}$ as the vector for **letter**. For the output, instead of connecting to the parts of the thalamus which send information directly to cortical areas, we connect to the neural group which gates connections between these cortical areas. If we set this to the identity vector **I**, then working memory will now contain **V⊗I=V**. This has the effect of routing information between cortical areas.

The result of this model when **letter+F** is placed in the visual cortex is shown in Figure 5. The model correctly starts repeating the alphabet from F. Changing the visual stimulus to some other letter will start from there, demonstrating that the rule can apply to multiple situations.

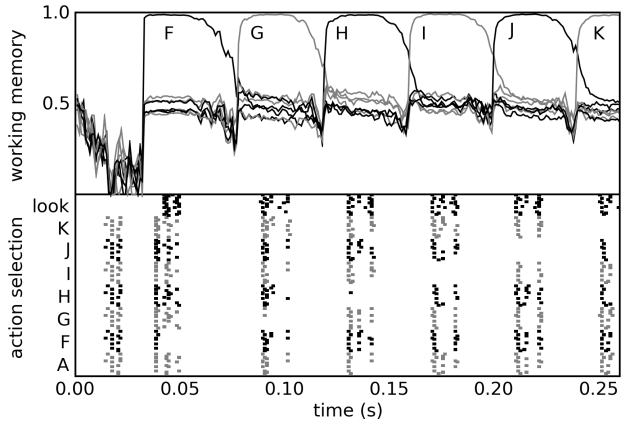


Figure 5: Contents of working memory (top) and spiking output from GPi indicating the action to perform (bottom). The *look* action takes information from visual cortex (in this case, **letter+F**) and routes it to working memory.

Question Answering

For the final task, we consider question answering. We perform this by first presenting the model with a symbolic statement such as the following:

statement + blue⊗circle + red⊗square

This would indicate a blue triangle and a red square are all in the visual field. The statement is presented to visual cortex for 50ms, and it will use the following rule to move it into working memory, as in the previous section:

IF visual cortex contains **statement+?X**

THEN copy visual cortex to working memory

After the statement is shown for 50ms, we stop stimulating visual cortex for another 50ms. This means that the system must successfully keep the statement in working memory over this time. After this time, we present a question to the visual cortex, such as the following:

question + red

A separate rule is defined for dealing with this situation:

IF visual cortex contains **question+?X**

THEN copy visual cortex to working memory B and also copy from working memory D to motor cortex

This rule copies the question to a separate area of working memory (B). As described previously (see Figure 3), this area allows a vector to be combined with the current contents of working memory. Furthermore, this rule also copies information from a third area of working memory (D) to the motor cortex. Since area D is connected to A and B so as to store the result of convolving area A (the statement) with the inverse of area B (the question), it should contain the answer to the question.

The results of this model answering two different questions from the same remembered statement are given in Figure 6. These two generic rules can answer any question provided in this format. Previous work on the capabilities of neural implementations of VSAs (Stewart, Tang, & Eliasmith, 2009) indicates that this system will scale well to 8 or more terms in a statement, out of a total vocabulary of 100,000 possible terms.

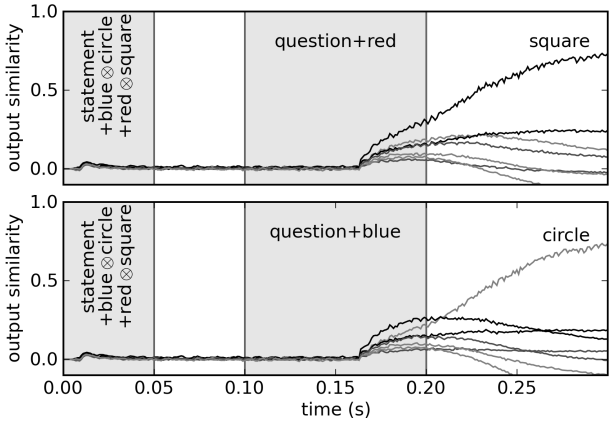


Figure 6: Answering two different questions starting from the same statement. The similarity between the contents of motor cortex and 7 possible answers is shown. The correct answer is chosen in both cases.

Implications

The model presented here helps to bridge the gap between cognitive science and neuroscience. It allows us to transform symbolic rules (the basis of much of cognitive theory) into specifications for the synaptic connectivity between neurons in cortex, basal ganglia, and thalamus. The resulting models give detailed predictions about the timing of events and the spiking behaviour of the neurons involved. With such models, we can also predict performance accuracy and the effects of various types of neurological damage.

The model also addresses a long-standing concern in cognitive science as to how neurons can possibly support the rich cognitive capabilities that seem clearly based on symbols and symbol manipulation. Specifically, we suggest that a VSA approach to representing symbols can be implemented in spiking neurons, and that these representations can be manipulated in a controlled and generic manner. We are aware of no other neural model with this flexibility, scalability, and connection to the underlying neurophysiology.

Rule types

Bridging cognitive science and neuroscience provides more than a mere neural implementation of cognitive theory. For our model, it has also suggested possible modifications to cognitive theory. When implementing the rules, we changed them from including explicit variables into commands to transform and copy the information currently represented in various parts of visual and working memory. If our future applications of this model continue to find this approach to rule definition sufficient for a wide variety of cognitive tasks, then we would argue this may be a more suitable framework for expressing cognitive rules than the standard variable-binding approach.

Timing

Our model is also highly constrained by known neurological data; the characteristics of the neurons involved and their connectivity are based on empirical results. As such, we can predict results that were previously derived purely by parameter fitting. For example, in most production system models of cognition (Soar, GOMS, EPIC, ACT-R, etc.), a certain amount of time is needed to select and apply an action. Based on empirical evidence, this is normally fixed to be 50 milliseconds (e.g. Anderson et al., 1995).

As can be seen in Figure 4 and Figure 5, our model requires just under 50 milliseconds to select and apply an action. While the median time needed is 44 milliseconds, the mean time for our current model is 48 milliseconds, due to the model occasionally repeating a step. These times are not affected by the size of our model, but can be changed by adjusting the time constant for the inhibitory neurotransmitter GABA in the basal ganglia. We currently use a value of 10ms (Gupta et al., 2000), and are seeking more detailed results from this area of the basal ganglia.

Conclusion

We presented a large-scale (373,000 spiking neuron) model capable of exhibiting rule-like behaviours such as question answering. By representing the conditions for applying inference rules as VSA vectors, and by representing the effects of those rules as vector transformations between cortical areas, we have shown a generic method for controlling neurally realistic cognitive systems.

Our ongoing work explores the broader capabilities of this model, including scaling up the number of rules (only 100 neurons need to be added per rule), and exploring the accuracy of the question answering as the vocabulary size increases. Other neural areas can also be added, including full vision and motor systems, as well as long-term memory.

References

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review 111*(4), 1036-1060.

Eliasmith, C. (2005). Cognition with neurons: A large-scale, biologically realistic model of the Wason task. *27th Annual Meeting of the Cognitive Science Society.*

Eliasmith, C. & Anderson, C. (2003). *Neural Engineering: Computation, representation, and dynamics in neurobiological systems.* Cambridge: MIT Press.

Gayler, R. (2003). Vector symbolic architectures answer Jackendoff's challenges for cognitive neuroscience. *International Conference on Cognitive Science*.

Georgopoulos, A.P., Schwartz, A., & Kettner, R.E. (1986). Neuronal population coding of movement direction. *Science*, 233, 1416-1419.

Gupta, A., Wang, Y., & Markram, H. (2000). Organizing Principles for a Diversity of GABAergic Interneurons and Synapses in the Neocortex. *Science 287*(5451), 273-278.

Gurney, K., Prescott, T., & Redgrave, P. (2001). A computational model of action selection in the basal ganglia. *Biological Cybernetics 84*, 401-423.

Plate, T. (2003). *Holographic reduced representations.* Stanford, CA: CSLI Publication.

Redgrave, P., Prescott, T., & Gurney, K. (1999). The basal ganglia: a vertebrate solution to the selection problem? *Neuroscience 86*, 353-387.

Singh, R. & Eliasmith, C. (2006). Higher-dimensional neurons explain the tuning and dynamics of working memory cells. *J. Neuroscience 26*, 3667-2678.

Spruston, N., Jonas, P., & Sakmann, B. (1995). Dendritic glutamate receptor channel in rat hippocampal CA3 and CA1 pyramidal neurons. *J. Physiology 482*, 325-352.

Stewart, T. C., & C. Eliasmith (in press). Compositionality and biologically plausible models. *Oxford Handbook of Compositionality.*

Stewart, T., & Eliasmith, C. (2009). Spiking neurons and central executive control: The origin of the 50-millisecond cognitive cycle. *ICCM 2009*, Manchester.

Stewart, T., Tang, Y., & Eliasmith, C. (2009). A biologically realistic cleanup memory: Autoassociation in spiking neurons. *ICCM 2009*, Manchester, UK.

A Hubel Weisel model for hierarchical representation of concepts in textual documents

Kiruthika Ramanathan (kiruthika_r@dsi.a-star.edu.sg), Shi Luping (shi_luping@a-star.edu.sg), Chong Tow Chong (chong_tow_chong@dsi.a-star.edu.sg)
Data Storage Institute, (A*STAR) Agency for Science, Technology and Research, DSI Building, 5 Engineering Drive 1, Singapore 117608

Abstract

Hubel Weisel models of the cortex describe visual processing as a hierarchy of increasingly sophisticated representations. While several models exist for image processing, little work has been done with Hubel Weisel models out of the domain of object recognition. In this paper, we describe how such models can be extended to the representation of concepts, resulting in a model that shares several properties with the PDP model of semantic cognition. The model that we propose is also capable of incremental learning, in which the knowledge is stored in the strength of the neuron connections. Degradation of old knowledge occurs as new knowledge is introduced to the system in a fashion that simulates decay theory in short term memory. The simulation model therefore captures several properties of cognitive conceptual memory, including generalization patterns, the role of rehearsal and, hierarchical representation.

Introduction

There exist several bottom-up approaches to hierarchical models of object recognition that are based on the visual cortex. They make use of Mountcastle's (1978) theory of uniformity and hierarchy in the cortical column and the model of simple to complex cells of Hubel and Weisel (1965), modeling how simple cells from neighboring receptive fields feed into the same complex cell, meaning that the complex cell has phase invariant response.

In this paper, we consider the following question. If the structure of the cortical column is uniform and hierarchical in nature and if the model of simple to complex cells can be used to model the visual cortex as discussed in prior works, then can such a model also be used to represent other modalities of information such as the concepts derived from text? We are therefore aiming to design a bottom up hierarchical memory for the representation of concepts, much the same way as it is designed for the representation of images. In this paper, we will define a concept as being a keyword in a document.

To deal with the dynamic nature of concept inputs, we look at incremental learning of concepts from two aspects relevant to concept representation from text – (a) with respect to new incoming features and (b) training of hierarchies. To perform this, we apply a set of geometric approximations to the incremental inputs and the existing memory, such that the new memory can be acquired without damage to the old ones.

Related work

Mountcastle (1978) showed that parts of the cortical system are organized in a hierarchy and that some regions are hierarchically above others. In general, neurons in the higher levels of the visual cortex represent more complex features with neurons in the IT representing objects or object parts (Hubel and Weisel, 1965). Hubel Weisel models have therefore been developed for object recognition (Cadieu et al., 2007; Fukushima, 2003) proposing a hierarchy of feature extracting simple (S) and complex (C) cells that allow for positional invariance. The combination of S-cells and C-cells, whose signals propagate up the hierarchy allows for scale and position invariant object recognition.

The idea of feature based concept acquisition has been well studied in psychological literature. Sloutsky (2003) discusses how children group concepts based on, not just one, but multiple similarities, which tap the fact that those basic level categories have correlated structures (or features). This correlation of features is also discussed in McClelland and Rogers (2003) who argue that information should be stored at the individual concept level rather than at the super ordinate category level allowing properties to be shared by many items.

Our model is related to Hubel Weisel approaches in that it implements a hierarchical modular architecture for bottom-up propagation of conceptual information. To our knowledge, however, this is the first implementation of a Hubel Weisel approach to non-natural medium such as text, and has attempted to model hierarchical representation of keywords to form concepts.

System architecture

The system that we describe here is organized in a bottom up hierarchy. This means that the component features are represented before the representation of concept objects. Our learning algorithm exploits the property of this hierarchical structure. Each level in the

hierarchy has several modules. These modules model cortical regions of concept memory. The modules are arranged in a tree structure, having several children and one parent. In our paper, we call the bottom most level of the hierarchy level 1, and the level increases as one moves up the hierarchy. The keywords from a document form the inputs to the system. These are directly fed to level 1. Level 1 modules resemble simple cells of the cortex, in that they receive their inputs from a small patch of the input space. Several level 1 modules tile the input space, possibly with overlap. A module at level 2 covers more of the input space when compared to a level 1 module. It represents the union of the input space of all its children level 1 modules. However, a level 2 module obtains its inputs only through its level 1 children. This pattern is repeated in the hierarchy. Thus, the module at the tree root (the top most level) covers the entire input space, but it does so by pooling the inputs from its child modules. In the visual cortex, the level 1 can be considered analogous to the area V1 of the cortex, level 2 to area V2 and so on.

Learning the first batch of information

To understand how the model learns, let us consider the inputs and outputs of a single module $m_{k,i}$ in level k of the system as shown in Figure 1a. Let **x**, representing connections $\{x_j\}$ be the input pattern to the module $m_{k,i}$. **x** is the output of the child modules of $m_{k,i}$ from the level *k-1*, and **a** represent the weights of the competitive network. The vector **a** is used to represent the connections $\{a_j\}$ between **x** and the cells in the module $m_{k,i}$. The output of a neuron in the module $m_{k,i}$ is given by $u = \sum_j a_j x_j$.

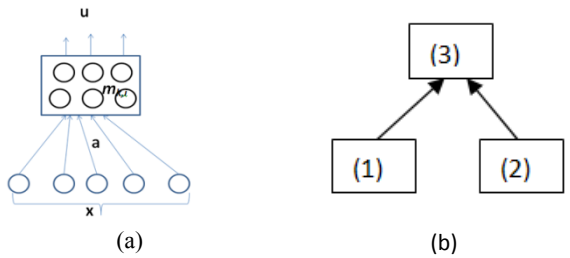


Figure 1a. Inputs and outputs to a single module $m_{k,i}$ b. The concatenation of information from the child modules of the hierarchy to generate inputs for the parent module

During learning, each neuron in $m_{k,i}$ competes with other neurons in the vicinity. Of the large number of inputs to a given module, a neuron is activated by a subset of them. The neuron then becomes the spatial center of these patterns. To ensure that there are no garbage neurons, we adopt in our creation of the module, a model of Growing SOM (GSOM) (Alahakhoon et al., 2000).

When all the modules at level k finish training, the second stage of learning occurs. This comprises the process by which the parent modules learn from the outputs of the child modules. Here, consider the case shown in Figure 1b where the module 3 is the parent of modules 1 and 2. Let **x(1)** be the output vector of module 1 and **x(2)** be the output vector of module 2. **x(i)** represents the Euclidean distance from the input pattern to the each output neuron i of the child modules. The input to module 3, $\mathbf{I(3) = x(1)||x(2)}$, is the concatenation of the outputs of modules 1 and 2. A particular concatenation represents a simultaneous occurrence of a combination of concepts in the child module. Depending on the statistics of the input data, some combinations will occur more frequently, while others will not. During the second stage of learning, the parent module learns the most frequent combinations of concepts in the levels below it. A GSOM is again used in the clustering of such combinations. The learning process thus defined can be repeated in a hierarchical manner.

Incremental learning

In this and the subsequent sections of the paper, we will use the terms *batch 0* to represent the first batch of documents. *Batch 1* refers to the subsequent set of documents. Once the system learns the documents in *batch i*, only the hierarchical structure and the neuron architecture are retained. All other information regarding the documents presented is discarded.

Incremental learning poses a challenge in Hubel Weisel based computational models due to three reasons. (1) Damage to the knowledge represented by old neurons which is fundamental in competitive learning. (2) Propagation of information in the hierarchical architecture. The number of output neurons of each child node increases with the introduction of the incremental batch. The input dimensions of the parent node are therefore changed and incompatible with the dimensions of the previous batch. (3) The irregularity in the input data dimensions. Where keywords are defined as concepts to be processed by the system, the keywords in an incremental batch will not be a subset of those in the previous batch. The architecture therefore has to provide rules for the generation and growth of new modules with respect to incoming incremental data.

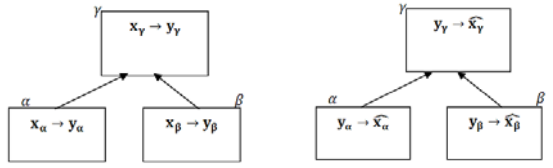
Preventing Damage to Old Memories: This problem is tackled using a sampling method using pseudo data inspired from Liu et al (2008). The algorithm implemented summarizes data distribution in a cluster map. Given neuron **a** in a GSOM of N neurons, consider the closest neuron **b**, **a,b** $\in N$, their midpoint is given by $\mathbf{a + b}/_2$. We generate a random set of vectors around neuron **a**, bounded on both sides by $\mathbf{a + b}/_2$. These pseudo vectors generated during the

training of batch *k* implicitly reconstruct the data used to train batches 0 to *k-1*.

Incremental learning in a hierarchy Let us consider Figure 2, where the modules α and β are child modules of γ.. At batch 0, the training sets $\mathbf{x}_\alpha$ and $\mathbf{x}_\beta$, consisting of p_0 patterns each are used to generate the neurons $\mathbf{y}_\alpha$ and $\mathbf{y}_\beta$.

$$\forall\ i \in p_0,\ \mathbf{x}_{\gamma,i} = \left[\left|\mathbf{x}_{\alpha,i} - \mathbf{y}_\alpha\right| \left|\right| \left|\mathbf{x}_{\beta,i} - \mathbf{y}_\beta\right|\right] \quad (1)$$

is passed to node γ. The vectors $\mathbf{x}_\alpha$, $\mathbf{x}_\beta$ and $\mathbf{x}_\gamma$ are then discarded.



(a) **(b)**

Figure 2. (a) Incremental learning stages. At batch 0, the training patterns at level 1, x_α and x_β cluster to form the neurons y_α and y_β. For simplicity, we consider that only one neuron is generated after training batch 0. (b) Batch 1 and the approximation of the pseudo vectors $\widehat{x_\alpha}$, $\widehat{x_\beta}$ and $\widehat{x_\gamma}$

When batch 1, consisting of p_1 vectors is now introduced, $\widehat{\mathbf{x}_\alpha}$ and $\widehat{\mathbf{x}_\beta}$ are approximated from $\mathbf{y}_\alpha$ and $\mathbf{y}_\beta$ respectively and used along with the new batch to train the GSOM modules α and β. After training, the neurons of the level 1 nodes y_α and y_β adapt to y_α' and y_β'. A set of pseudo data $\widehat{\mathbf{x}_\gamma}$ are approximated from the neuron $\mathbf{y}_\gamma$.

From equation 2, $\widehat{\mathbf{x}_\gamma}$ represents the Euclidean of $\widehat{\mathbf{x}_\alpha}$ and $\widehat{\mathbf{x}_\beta}$ from $\mathbf{y}_\alpha$ and $\mathbf{y}_\beta$ respectively, i.e, for a set of p_0' pseudo data,

$$\forall\ i \in p_0',\ \widehat{\mathbf{x}_{\gamma,i}} = \left[\left|\widehat{\mathbf{x}_{\alpha,i}} - \mathbf{y}_\alpha\right| \left|\right| \left|\widehat{\mathbf{x}_{\beta,i}} - \mathbf{y}_\beta\right|\right] \quad \mathbf{(2)}$$

However, during the training of batch 1, the measure for $\widehat{\mathbf{x}_\gamma}$ should be the distance to y_α' and y_β', the updated neuron vectors. A set of adapted pseudo vectors $\widehat{\mathbf{x}_\gamma}'$ should therefore be approximated.

In Euclidean space, we can we can visualize the problem as shown in Figure 3,

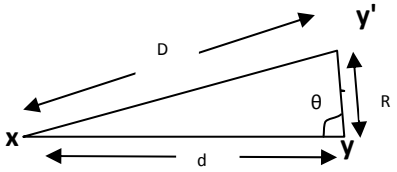


Figure 3. Approximation of incremented pseudo vector for levels 2 and above in the hierarchy

We consider two cases, (a) **y'** is not the winning neuron for the pattern **x**. (b) **y'** is the winning neuron.

Case (a). $\mathbf{y'}_\alpha$ is not the winning neuron, i.e, R<<d

For ease of analysis, assume that d=1

$$D \cong \frac{1}{\pi}\int_0^{\pi} \left|\mathbf{y}'(\boldsymbol{\theta}) - \mathbf{x}\right| d\theta$$

$$D \cong \frac{1}{\pi}\int_0^{\frac{\pi}{2}} \left|\mathbf{y}'(\theta) - \mathbf{x}\right| + \left|\mathbf{y}'(\pi - \theta) - \mathbf{x}\right| d\theta$$

Where $f(\theta) = \left|\mathbf{y}'(\boldsymbol{\theta}) - \mathbf{x}\right| + \left|\mathbf{y}'(\boldsymbol{\pi} - \boldsymbol{\theta}) - \mathbf{x}\right|$,

$$D \cong \tfrac{1}{\pi}\textstyle\int_0^{\frac{\pi}{2}} f(\theta) \quad (3)$$

$$f(\theta) = \sqrt{(1 - cos\theta R)^2 + (sin\theta R)^2} + \sqrt{(1 + cos\theta R)^2 + (sin\theta R)^2}$$

Let $a = \left|\mathbf{y}'(\boldsymbol{\theta}) - \boldsymbol{x}\right|$ and $b = \left|\mathbf{y}'(\boldsymbol{\pi} - \boldsymbol{\theta}) - \boldsymbol{x}\right|$

We obtain

$$f(\theta) = \sqrt{2}\sqrt{1 + R^2 + ab} \quad (4)$$

Where $E_0 = ab$,,

$$E_0 = \sqrt{(1 - cos\theta R)^2 + (sin\theta R)^4 + \sin^2(\theta R)[2 + 2cos^2\theta]} \quad (11)$$

$$\cong 1 - \frac{cos^2(\theta R)}{2} + \frac{3\sin^2(\theta R)}{2} \quad (5)$$

Substituting (5) into (4), we obtain

$$f(\theta) = 2\left(1 - \frac{R^2}{8} + \frac{1}{2}sin^2(\theta R)\right) \quad (6)$$

Substituting (6) into (3), we obtain

$$D = = 1 + \frac{R^2}{8} + \frac{R^2}{\pi}\textstyle\int_0^{\frac{\pi}{2}} sin^2\,\theta d\theta \quad (7)$$

Solving (7), we have

$$D = 1 + \frac{3R^2}{8} \quad (8)$$

Based on Figure 3, if we approximate a $\theta = {}^{\pi}/_{2}$, we obtain $D = 1 + \frac{R^2}{2}$. In implementation, to satisfy (8) we use the inequality (9) to assign the value of θ.

$$\frac{4\pi}{9} < \theta < {}^{\pi}/_{2} \quad (9)$$

We can therefore conclude that, a θ value specified by inequality (9) can be used to re-generate the dataset $\widehat{\mathbf{x}_{\gamma,1}}[0,1,\ldots,j,\ldots k_\alpha]$, where $\widehat{\mathbf{x}_{\gamma,1}}[\boldsymbol{j}] \equiv \left|\widehat{\mathbf{x}_{\alpha,1}} - \mathbf{y'}_\alpha\right|$ and $\mathbf{y'}_\alpha$ is not the winning neuron.

Case (b): $\mathbf{y'}_\alpha$ is the winning neuron

If y' is the winning neuron, a random value of θ, $0 < \theta < \pi$ can be used to regenerate $\widehat{\mathbf{x}_{\gamma,1}}[j] \equiv \left|\widehat{\mathbf{x}_{\alpha,1}} - \mathbf{y'}_\alpha\right|$, where y'_α is the winning neuron.

Dealing with the problem of new input dimensions: A rule based approach of creating a new module to process the new keywords is preliminarily proposed to deal with the dynamically increasing input dimensions. A module is trained and connected to parents only if the number of concepts that it represents increases above a predefined threshold. In order to avoid overcrowding, heuristic rules have been put into place such that a parent has atmost three children.

Experimental results

To illustrate the cognitive properties of the training model, we train the system using 21 concepts from 200 documents. The concepts included ideas such as "birds", "animals", "flowers", "trees" etc, same as the ones used by McClelland and Rogers (2003). The following preprocessing was performed to the documents. First, the contents were analyzed and the stopwords removed. The concept terms were stemmed and grouped using Wordnet (Fellbaum, 1998) before a tf.idf weighing scheme was used to select the most relevant concepts to the batch. For visualization purposes,

Hierarchical identification of concepts from wiki documents

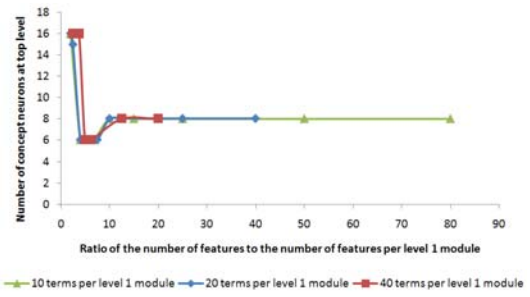


Figure 4. The number of concept neurons at the top of the hierarchy vs. the number of features per bottom level module

In this section we observe how our hierarchical model captures the properties of semantic cognition outlined by McClelland and Rogers (2003, 2008). The training data used by McClelland and Rogers is intuitively designed based on common sense knowledge. Our system, on the other hand is trained using information from 200 text descriptors of the concepts from wikipedia. The snippets varied in length from 50 word descriptions to 500 word descriptions. Figure 4 illustrates the number of concept neurons at the top level as a function of the ratio of the number of features to each level 1 module and the total number of features. When there are only two layers in the hierarchy, a larger number of concept neurons (16) are generated. The number of concept neurons converges to between 6 and 8 for all other architectures. Typically, for a six concept cluster, the concept of penguin is separate from that of other clusters. This is shown in Figure 6.

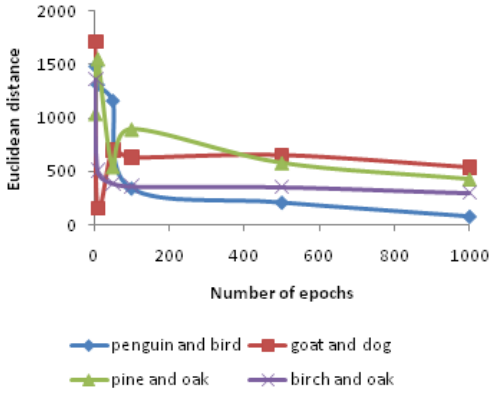


Figure 5. Euclidean distance between various concepts vs. the number of training epochs

In Figure 5, we observe the evolution of the Euclidean distance between concepts. The training shows empirical properties of convergence. The distances between the various concepts are stable after 500 epochs of training. We can also observe promising results from the concept representation point of view. For instance, the Euclidean distance between "pine" and "oak", for instance, is larger than the Euclidean distance between "birch" and "oak", which belong to the same family.

Figure 6.a shows the top two levels of a five level hierarchy of hierarchy of concepts obtained (10 concepts per GSOM module and 160 concepts used in training). We observe, as is the result in McClelland and Rogers that similar concepts tend to be near each other in space. For instance, "canary" and "sparrow" tend to be closer to each other, but far from "penguin". In some cases, super ordinate terms, such as "bird", "tree" etc are mined as part of the hierarchy. There are some interesting observations that can be made here. We can see that the highest level (level 5) shows general concepts while level 4 shows the concepts one level lower. i.e., while the neuron 1 refers to "animals", the neuron box "2" refers to more detailed differentiation of neuron 1. Further to this, the system also shows some intermediate level categorization characteristics that taps item frequency effects. In McClelland and Rogers' paper, they describe it as the process by which certain descriptive terms such as "tree", "bird" and "dog" tend to be acquired earlier than the super ordinate terms such as "plant" or "animal" or more specific terms such as canary, pine or poodle. The general consensus for this is that parents use certain intermediate level words more frequently when speaking to children. As such, intermediate concepts, based on their frequency of usage, are also clustered more tightly into intermediate groups within super ordinate concepts.

In the experimental data, some concepts were used more frequently than others of the same sub category. These include birch (12 instances) vs. 10 instances of pine and maple, 16 instances of rose vs. 3 instances of daisy and 7 instances of sunflower. It is seem that the more frequent concepts are tied together with the super ordinate concept neuron (dog is tied with animal, rose with flower, sparrow with bird) and so on.

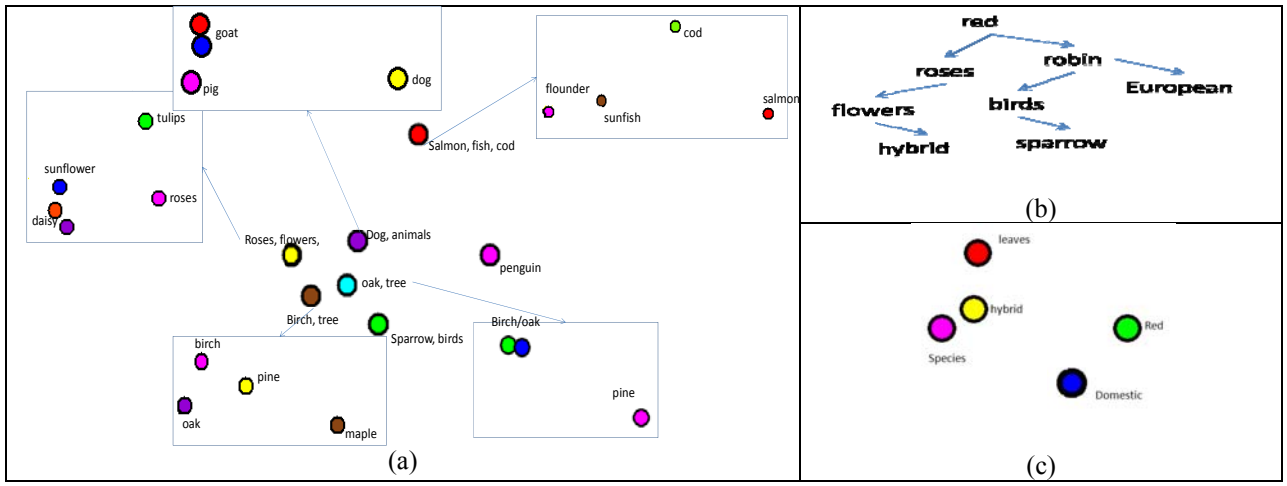


Figure 6 (a) Hierarchical representation of the 21 concepts in the memory system. (b) Grouping due to secondary characteristics – "leaves" groups all trees, "red" for rose, salmon etc. (c) Associative retrieval of concepts based on activation of high correlation neurons in the hierarchy.

At the lower level, we can also observe alternate similarity grouping which are based on individual features. For instance, at level 1 module, one might observe neurons denoting concepts such as "leaves" grouping birch, maple and oak. The concept "grow" groups objects of all categories. "Red" indicates objects that are red – rose, cod, salmon. One such lower layer representation is shown in Figure 6.c.

From these results, one can expect this model to perform a sort of associative retrieval of the kind that humans perform. For instance, if one sees the concept "gold", one would think of perhaps "ring" and subsequently "marriage" and then "family", "children" and so forth. In the same way, firing of the neuron "red" at level 1 will lead to firing of corresponding nodes at various levels. This will lead to other concepts being fired, both sequentially and parallely. For instance, probing the model with "red" leads to the sequential firing as described in Figure 6.b.

Each concept can, in turn, be used as a probe to activate relevant neurons. In this sense, the model describes the human chain of thought process. Work is in progress to study how this process can be modeled and how the firing can be controlled by the wiring strength between modules.

Incremental learning performance

Overview of incremental learning with 5 concepts and 3 batches: In this section, instead of introducing all the concepts are one go, concepts are learnt in batches. The figure below shows the evolution of the incremental concept network at the top level for the first three batches (a) tulips and sunflowers only (b) sparrow and sunfish (c) salmon. In batch 1, the top level generalizes, creating concept representation of flower and animal. Lower layers now portray the differentiated concepts. In batch 2, the concept of salmon was introduced. At this juncture, the old information from sunfish is sub grouped under fish and sparrow is considered a separate entity. These experiments suggest merit in the approximations that we have described earlier in the paper.

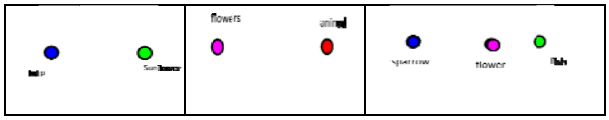


Figure 7. The top-level evolution of incremental concept representation. (a) Batch 0 tulip and sunflower. (b) Batch 1: sparrow and sunfish. (c)Batch 2: salmon

Hierarchical representation in incremental learning

In this experiment, concepts were introduced one by one, beginning with "birch" and "cod". At some juncture concepts are reintroduced to investigate the effect of new data and rehearsal on old data. Some of the results obtained are discussed in this section.

Figure 8a shows the evolution of how the concepts "birch" and "betula" are represented ("Betula" is the scientific name for "birch"). Birch is the first concept that was introduced to the system. At batch 0, the distinction between the concepts "birch" and "betula" appears in level 3 of the hierarchy. To the system, "betula" and "birch" are two distinct concepts, though with a low Euclidean distance of 39.81. When batch 2 is introduced, the distinction between the two concepts moves one level lower to level 2 and eventually to level 1. However, as more concepts are introduced to the system, the presence of the new information makes the system lose the distinction between the concepts "birch"

and "betula", and the Euclidean distance between the concepts reduces to 0 at batch 7. However, at batch 10, when the concept of birch is reintroduced, the Euclidean distance between the two terms increases before gradually decreasing to 0 once again. A similar result is also observed in the relationship between terms "canary" and "islands".

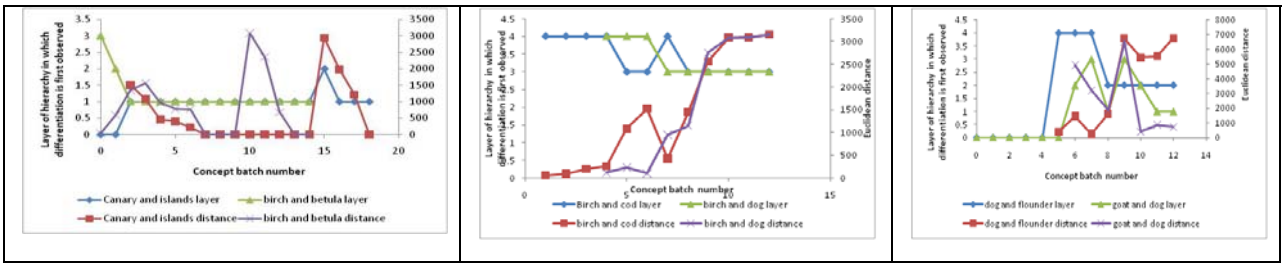


Figure 8. Evolution of the hierarchical and Euclidean relationship between the concepts (a)"birch" and "betula" vs "canary" and "islands" (b) "Birch" and "cod" vs. "birch" and "dog" (c) "dog" and "flounder" vs. "goat" and "dog"

Figure 8b shows the representation of the concept "birch" with respect to the concepts "dog" and "cod". "Birch" is introduced to the system at batch 0, "cod" at batch 1 and "dog" at batch 4. The differentiation between the concepts "birch" and "cod" is at level 4 and converges to level 3. By batch 8, the concepts of "dog" and "cod" are of the same distance from "birch". At this juncture, the system at level 3 no longer distinguishes between "birch", "cod" and "dog", but makes a distinction between "plant" and "animal". Figure 8c shows a similar relationship of concepts "dog" with the concepts "flounder" and "goat". The flounder-dog distinction converges to level 2 (from Figure 8b, we can see that the plant-animal distinction occurred at level 3) while the dog-goat distinction converges to level 1. The Euclidean distance between the concept terms "dog" and "goat" converges to approximately 700 which is close to the value that is obtained through batch learning (from Figure 5).

Conclusions and further work

In summary, our model attempts to propose a hierarchical Hubel Weisel model for the acquisition of concepts from text such that the concepts are represented in a hierarchical connectionist network. The result is a new framework that we have applied in two scenarios. The first is concept acquisition where we have shown that the system is able to represent everyday concepts in a hierarchical fashion, in a manner similar to the PDP model. The system was interestingly also able to perform chain retrieval, in that when "red" was given as a probe to the system, it was able to retrieve "robin" and by association "sparrow". Secondly, we have modeled information approximation and incremental learning, which models some properties of short term memory.

There are several directions for further work in this area. In addition to the pertinent issues of improving computation time and processing algorithms to make the system able to handle large sets of data, one important direction is the incorporation of semantic information into the hierarchical architecture. As of now, this information is ignored and only the statistical properties of keywords are taken into consideration in the generation of the concept hierarchy. Work is under process to integrate semantic information into the model. Work is also under progress to include common sense knowledge in the model. We expect that these additions will make the model more cognitively accurate. In addition to this, we are also incorporating other aspects of cognition such as attention; interest etc to study the generation and behavior of the cognitive map.

References

Cadieu, C., Kouh, M., Pasupathy, A., Conner, C., Riesenhuber, M., & Poggio, T.A. (2007). *A Model of V4 Shape Selectivity and Invariance*. J ournal of Neurophysiology 98, 1733-1750

George D, Hawkins J (2005), *A hierarchical Bayesian Model of Invariant Pattern Recognition in the Visual Cortex*, International Joint Conference on Neural Networks, 3, 1812-1817

Vernon Mountcastle (1978), *An Organizing Principle for Cerebral Function: The Unit Model and the Distributed System*, The Mindful Brain (Gerald M. Edelman and Vernon B. Mountcastle, eds.) Cambridge, MA: MIT Press.

Hubel D and Weisel T (1965), *Receptive fields and functional architecture in two non striate visual areas (18 and 19) of a cat*, Journal of NeuroPhysiology 28, pp229-289

Fukushima K(2003), Neocognitron for handwritten digit recognition 1, Neurocomputing , 51C, 161-180

Alahakhoon D, Halgamuge S K, Srinivasan B (2000), *Dynamic Self Organizing maps with controlled growth for Knowledge discovery*, IEEE Transactions on neural networks, 11(3), pp601-614

Liu M, Liu Y C, Wang X L (2008), IGSOM: *Incremental clustering based on self organizing map*, International Conference on Intelligent Information hiding and multimedia Signal Processing, IIHMSP'08, pp 885-890

Fellbaum C (1998), *Wordnet: An electronic lexical database*, MIT Press

Mc Clelland J L and Rogers T T (2003), *The parallel distributed processsing approach to semantic cognition*, Nature Reviews Neuroscience, 4(4), pp310-322

Sloutsky VM (2003), *The role of similarity in the development of categorization*, Trends in Cognitie Sciences, 7, 246-251

Rogers T T, McClelland J L (2008), *Precis of Semantic Cognition, a Parallel distributed Processing approach*, Brain and Behavioral Sciences, 31, pp 689-749

Automatic and Controlled Processes in Semantic Priming: an Attractor Neural Network Model with Latching Dynamics

Itamar Lerner (itamar.lerner@gmail.com)
Interdisciplinary Center for Neural Computation, The Hebrew University of Jerusalem
Giv'at Ram, Jerusalem 91904 Israel

Shlomo Bentin (shlomo.bentin@huji.ac.il)
Department of Psychology, The Hebrew University of Jerusalem
Mount Scopus; Jerusalem 91905 Israel

Oren Shriki (oren70@gmail.com)
Department of Physiology and Neurobiology, Faculty of Medicine, Ben-Gurion University of the Negev
Be'er-Sheva 84105 Israel

Abstract

Semantic priming involves a combination of automatic processes like spreading activation (SA) and controlled processes like expectancy and semantic matching. An alternative account for automatic priming has been suggested using attractor neural networks. Such networks offer a more biologically plausible model of real neuronal dynamics but fall short in explaining several important effects such as mediated and asymmetrical priming, as well as controlled effects. We describe a new attractor network which incorporates synaptic adaptation mechanisms and performs latching dynamics. We show that this model can implement spreading activation in a statistical manner and therefore exhibit all priming effects previously attributed to automatic priming. In addition, we show how controlled processes are implemented in the same network, explaining many other semantic priming results.

Keywords: Semantic priming; Attractor networks; Latching dynamics

Introduction

Semantic priming is one of the most important phenomena in the study of word perception and semantic memory. In a typical priming experiment (Neely, 1991), subjects are visually exposed to two words in succession, the prime and the target, and are required to silently read the prime and either name the target (pronunciation task), or decide whether it is a real word or not (lexical decision task). The target could either be semantically related or unrelated to the prime (or a nonword, in case of the lexical decision task). The priming effect is expressed as shorter average reaction times (RT) and reduced error rates in the related relative to unrelated condition. Sometimes, a neutral prime is used (e.g. a row of X's) to allow the differentiation between response facilitation (in the related condition) and inhibition (in the unrelated condition).

Computational accounts for semantic priming are divided between models based on automatic processes and those based on controlled processes. The most famous among the automatic accounts for priming is the spreading activation (SA) theory of Collins & Loftus (1975). This model suggests that concepts in semantic memory are represented by nodes that are connected to each other according to their semantic relatedness. When a concept is activated (by external or internal input) the activity spreads to related concepts (see figure 1). In priming experiments, activation of the prime concept (e.g. *table*) leads to activation of its related concepts (e.g. *chair*). This pre-activation facilitates the recognition of subsequent related targets. If an unrelated or a neutral target appears, no such head-start is available. Hence spreading activation may account for the facilitation component of semantic priming. Automatic priming can also be conceived in attractor networks with distributed representations of concepts (e.g. Mason, 1995). In such models concepts are represented by activity patterns of neurons' assemblies and semantic relationship is implemented as correlation between these representations. When the prime appears, the network converges on its corresponding activity pattern. When the target is then presented, the network changes its activity pattern from that of the prime to the one corresponding to the target. If the target is related to the prime, fewer changes need to take place due to the correlations; therefore, the convergence takes less time and a priming effect emerges.

Attractor networks are probably more true to the biological nature of real neuronal dynamics which include content-addressable memories, distributed representations and attractor states. However, they fall short in explaining several important priming results. Mediated priming is one example (e.g. McNamara, 1992): It was found that word pairs which are indirectly related to each other (i.e., related only through a mediating word, like *lion* and *stripes*, related through *tiger)* can nevertheless prime each other. Allowing activation to spread to more than one step, SA theories could easily account for such effects. Attractor networks, on the other hand, cannot explain mediated priming since the activation patterns of indirectly related pairs are not correlated. Similarly, whereas SA models allow asymmetric connections between nodes and therefore allow asymmetric priming (in which the magnitude of priming varies according to which word in a given pair is designated prime and which is the target; e.g. *pay-check* vs. *check-pay*), such an effect cannot

be obtained by attractor network models because they rely on correlation, a symmetric trait by definition.

Here we present an attractor neural network which implements SA in a statistical manner. By doing so, we bridge between SA and attractor models and show how attractor networks can exhibit results like mediated and asymmetric priming. In addition, we discuss some controlled mechanisms like expectancy (Becker, 1980) and semantic matching (Neely, Keefe & Ross, 1989) and suggest how they may be interpreted within the same network.

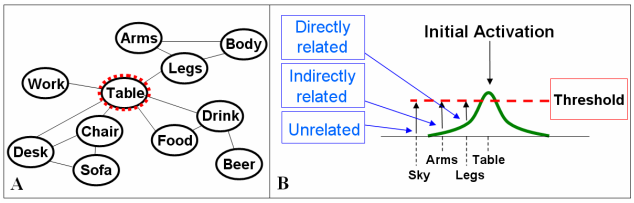


Figure 1: The spreading activation theory. (A) Related concepts connected in semantic memory. (B) Activation spreads through the network

Computational Model

Following the traditional separation between stages of processing (e.g. Smith et al., 2001), our model consists of 2 computational layers, lexical and semantic (Figure 2). We assume that after a string of letters is analyzed for orthographic composition, the result is fed to the lexical network where word identification occurs. If the letters form a real word, this word is 'recognized' by the lexical network and its activity is fed forward to the semantic network where the word's meaning is stored. However, the semantic network can influence lexical processing on line via feedback. Such a top-down effect contributes to semantic priming: when the semantic network is a priori 'tuned' to a concept with some relatedness to the newly arrived word, the lexical network recognizes this word quicker because both bottom-up and top-down pathways contribute to the recognition process (as opposed to the unrelated case, where the top-down pathway does not contribute). In the case of a neutral stimulus, none of the networks is activated and no transfer of information occurs.

The lexical and semantic networks are modeled as Hopfield-type attractor neural networks, with sparse representations and continuous-time dynamics (see Tsodyks, 1990). In our simulations, both the lexical and the semantic networks are fully connected recurrent networks, each composed of 500 neurons. Memory patterns (concepts) encoded by each network are binary vectors of size 500, with '1' indicating a maximally active neuron, and '0' an inactive one. The representations are sparse (i.e., a small number of neurons are active in each pattern) with p being the ratio of active neurons ($p<<1$). The connectivity between neurons assures stability of these patterns. External inputs to and from the network are always excitatory.

The neurons themselves are analog with activity in the range [0,1] and obey a logistic transfer function of their local input $h(t)$. The local input itself obeys a linear differential equation (following Herrmann, Ruppin & Usher, 1993) of the form:

$$\textbf{(1)}\ \tau_n \dot{h}_i(t) = -h_i(t) + \sum_{j=1}^{N} J_{ij} x_j(t) - \lambda \cdot (\bar{x}(t) - p) - \theta + \left[I_i^{ext}(t) - \theta^{ext} \right]_+ + \eta_i$$

In (1), τ_n is the time constant of the neuron, $x_j(t)$ the activity at time t of the j-th neuron (with $\bar{x}$ indicating average over all neurons), J_{ij} is the connectivity weight, N is the number of neurons (500 in our case), p is the sparseness of the representations, λ a regulation parameter which maintains stability of mean activation, and θ is a constant neuronal threshold (See Herrmann et al. for details). The $[\ldots]_+$ symbol indicates a threshold linear function, such that $[x]_+=0$ for $x<0$, and $[x]_+=x$ otherwise. This leads the external input to the neuron, $I_i^{ext}(t)$, to be consequential only if it surpasses the constant external threshold θ^{ext}. Finally, η_i is a noise term drawn from a Gaussian distribution with some temporal correlations. Relatedness between concepts is implemented in the model as correlations between memory patterns (reflecting the degree of overlap between them). The stronger two concepts are related, the higher is their correlation. The correlation of unrelated patterns is negligible ($|c|<0.05$ with c being the correlation)

Two major differences distinguish the lexical from the semantic network. First, while the semantic network includes correlated memory patterns representing semantic relations between concepts, there are no correlations in the memory patterns of the lexical network. This is not to indicate there are no lexical relations (such relations obviously exist), but merely to ensure that they would not influence the simulations. Indeed, typical semantic priming experiments do control for such confounds by selecting prime-target pairs that bare no lexical/phonological relations.

The second difference is, perhaps, the basic premise of our model: Unlike the lexical network (and the majority of previous attractor network models), our semantic network is associative in nature. Neuronal adaptation mechanisms at the synaptic level preclude the network from maintaining stability for long; therefore, the network, after converging to one attractor, leaves it quickly and jumps to another one. This process is stochastic in nature and can continue forever as long as no new input interferes. These jumps cannot be accurately predicted, but they tend to happen (although not necessarily) between correlated patterns. Such network behavior was termed 'latching dynamics' by Treves (2005). Specifically, short-term synaptic plasticity was modeled according to Loebel & Tsodyks (2002), with each synaptic weight of a neuron decreasing linearly with its activity:

$$\textbf{(2)}\ \dot{J}_{ij}(t) = \frac{J_{ij}^{\max} - J_{ij}(t)}{\tau_r} - U x_{\max} x_i(t) J_{ij}(t)$$

In (2), J_{ij}^{max} is the common Hopfield connectivity weight for sparse networks, τ_r is the time constant of recovery of the synaptic efficacy, and U is the utilization of synaptic

resources. The term x_{max} is a hypothetical maximum firing rate of a neuron (for example 100 pulses/sec) which adjusts the equation to fit a neural firing rate bounded by 1.

Links between the lexical and semantic networks are based on connections between active neurons in corresponding patterns (See figure 2): An active neuron in a certain word pattern in the lexical network sends excitatory connections to all active neurons in the corresponding concept-pattern of the semantic network, and vice-versa. Since correlations between patterns exist in the semantic network, one neuron in that layer could concurrently influence and receive input from different neurons activated in different patterns in the lexical layer. The lexical network also receives bottom-up input, representing the visual letter-string, which follows the same logic: Neurons belonging to the pattern presented to the lexical network receive excitatory inputs, while others receive no input.

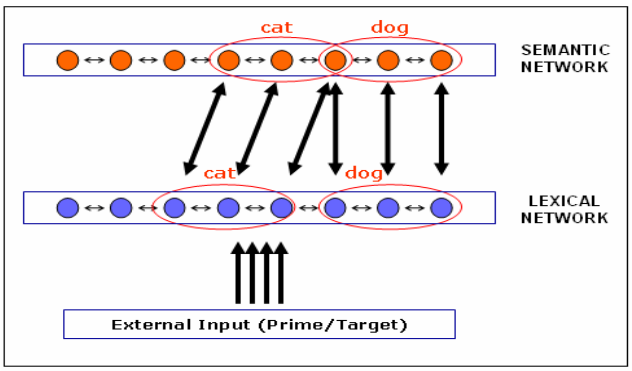


Figure 2: Architecture of the model. Two recurrent networks connected to each other with excitatory links. The semantic network contains correlated representations

Simulations

The simulations were run on an Intel Core 2 Quad CPU Q6600 with 2.4 Ghz and 2 GB of RAM. Simulations were written in Matlab 8a. In all the simulations, one numeric step represents 0.66ms.

Encoded Patterns

We encoded 17 memory patterns in each network. All patterns were binary vectors with equal mean activity and very sparse representations. In the semantic network, the following basic correlations between patterns were set: four groups, each containing four patterns, formed 'semantic neighborhoods' (patterns 1-4, 5-6, 9-12 and 13-16): Each pattern in a group was correlated with the other patterns in its group, but, with few exceptions (see below), no correlations existed between the groups. Correlations within a group had one of two values, representing two levels of direct relatedness. In addition, we also added some correlations between patterns of different neighborhoods to allow indirect priming investigations. For example, we added some correlation between pattern 2 and pattern 9, which resulted in patterns 1 and 9 being indirectly related. The 17th memory pattern was a 'baseline' pattern which the network was initialized to at the beginning of each trial, and was not correlated to any of the other patterns. In the lexical network, all 17 patterns were unrelated to each other. The 17th pattern was, again, the initial state for the network, and was not linked through top-down or bottom up lexical-semantic connections to the baseline pattern in the semantic network (thus forming a 'neutral' pattern).

Experimental Procedure and Data Analysis

Each trial began with the presentation of a binary vector to the lexical network, corresponding to one of its patterns (1's in the to-be activated neurons, 0's in the rest). This vector served as "prime". In neutral trials, pattern 17 (the neutral pattern) was presented. Two experiments were conducted. The first tested the general performance of the semantic network. The prime was presented for 100ms and it was always pattern no. 1. The network was allowed to advance according to the dynamic equations without further interference, for a total period of 3000ms. The procedure was repeated for 100 trials. Correlation of the momentary network state with each pattern, for each time point in the simulation, was averaged offline. The second experiment tested whether the performance of the model, when semantic priming occurs, corresponds with predictions based on human studies. The prime was presented for 100ms and followed by a target after 150 ms, hence creating a 250 ms SOA. The time interval from target onset until convergence of the lexical network indexed the reaction time, providing the network converged to the correct attractor. Primes and targets were either directly related (i.e., two patterns from the same neighborhood), indirectly related (two patterns from different neighborhoods but linked through a mediating pattern as explained earlier), unrelated (two patterns from different neighborhoods with no indirect connections), or neutral (in which the prime was the neutral pattern and the target any of the 'real' patterns). 100 trials were simulated for each relatedness condition, with prime-target pairs chosen randomly. Mean reaction times and standard errors were computed for each condition.

Results

Figure 3 presents the typical performance of the two networks (for presentation purposes, here we used a 1000ms SOA). Correlation of the state of each network with each of its stored patterns (including the memories and the neutral pattern) during a trial is presented in different colors, with convergence to a specific pattern indicated by its number appearing on top. The lexical network followed the external input, by converging to the corresponding memory pattern and keeping stability until a new input arrived. In contrast, the semantic network converged to the appropriate memory pattern, only to jump to other attractors in a serial manner, hence presenting latching dynamics. When a new external input arrived, the semantic network stopped its transitions and quickly converged to the corresponding memory pattern

shortly after the lexical network has done so. As evident in Figure 3, most jumps were within the neighborhood, while jumps to different neighborhoods occurred less frequently.

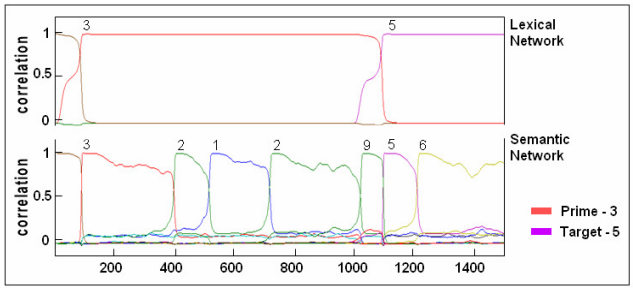


Figure 3: Typical behavior of the two networks

In the first experiment, trials always included pattern 1 as the prime. The mean correlation between the state of the semantic network and all its memory patterns was computed for each time point over trials. Figure 4A presents the correlations for five different time points after the prime onset. The x-axis represents different patterns according to their relatedness to pattern 1, with pattern 1 itself in the middle. Evidently, the mean correlations followed the principle of spreading activation: Initially, the concept represented by the external input has the strongest activation (correlation), its directly related concepts are activated to a smaller degree, and concepts not related to it are not activated at all. With time, as semantic transitions occur, the mean activation of the initial concept is decreasing, while activation in its related concepts increases. Indirectly related concepts also show some activation, with a delayed peak. Unrelated concepts receive no activation at all. After enough time, the mean correlation with each of the network's patterns is divided more or less equally, corresponding to a nearly deactivated state of the whole network (the mean activity would have reached near zero values in case more than 16 patterns were used).

In the second experiment, the mean RTs of the lexical network were computed and are presented in figure 4B. As can be seen, priming occurs for both directly and indirectly related pairs, although the effect is stronger in the direct case. In addition, weak relations produced smaller priming than strong relations. All these effects were significant at $p < 0.001$. There was no significant difference between the unrelated and neutral conditions, confirming that only facilitation occurred.

Discussion

The results of these simulations demonstrated that an attractor neural network with latching dynamics can implement spreading activation in a statistical manner. In a way, one could see the activity of nodes in the original spreading activation model as an average manifestation of the correlation in our attractor model. There is, however, an important distinction between our model and SA models: In our network, spreading is mixed with relaxation periods which correspond to the network reaching an attractor. In other words, activation does not spread in a monotonic manner like in the original SA model, but rather in jumps which fit the dynamical jumps from one attractor to another.

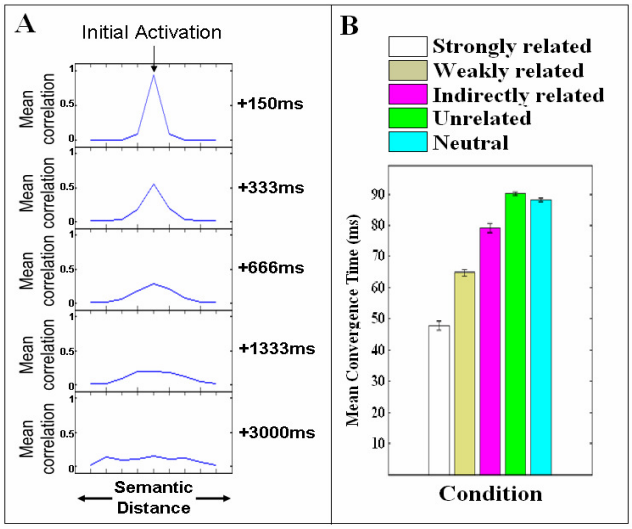


Figure 4: Simulations results. (A) Statistical spreading activation portrayed by the network as mean correlation over trials. (B) Mean convergence times of the lexical network for the different relatedness conditions

The results of the second simulation demonstrate how the dynamics in the semantic network affects the convergence time of the lexical network such that priming effects are produced. When the semantic network state is correlated with the target pattern at the moment the target word appears, its top-down influence shorten the lexical network's convergence times. Due to semantic transitions, such correlations may occasionally appear in indirectly related trials and produce the mediated priming effect. Although not explicitly simulated, these jumps can also produce asymmetry in priming: Transition probabilities from pattern A to pattern B are not necessarily equal to those from B to A since network transitions, in general, are uniquely influenced by the other memory patterns A and B are correlated with (which can be very different for A and for B). This asymmetry allows making a distinction between semantic relatedness (as indicated by correlation) and associative relatedness (as indicated by the probability of one pattern leading to another pattern). Former attractor models relied solely on correlations between prime and target and therefore could not produce either mediated priming or asymmetry in priming.

Controlled Processes

When the SOA between prime and target is sufficiently long, subjects may decide to engage in specific strategies while responding. The general aim of such strategies is to shorten reaction times to the target. In contrast to the automatic nature of SA, strategies are considered to be under the subject's cognitive control.

A well known example of such strategies is expectancy (Becker, 1980). It is assumed that subjects may be able to realize that in part of the trials, the target is semantically and/or associatively related to the prime and develop a set of expected targets from the prime's semantic "neighborhood". When the target appears, this "expected set" is searched first, while the general lexicon is searched only if the target is not included in the expected set. Obviously, when the target is found in the expected set, its recognition time is accelerated. If it is not found there, however, its recognition is delayed by this initial screening procedure. Hence, the application of an expectancy strategy may account for both facilitation and inhibition of the priming effect. Two types of this strategy were identified (Becker, 1980): A 'prediction' strategy is used when the upcoming target is highly predictable. Only one item (or very few) is included in the expected set and, consequently, facilitation is robust while inhibition is negligible. A 'general expectation' strategy is used when more than a few items could potentially be targets and the expected set includes them all. Both facilitation and inhibition should result. However, subsequent studies have shown that not all conditions yield inhibition (for example, pronunciation tasks), which put this later strategy into question (Keefe & Neely, 1990). In either case, the requirements for this controlled process to be initiated are sufficiently long SOA and a sufficiently salient proportion of related pairs in the stimulus set (called the 'relatedness proportion') which makes such expectancies reasonable. Indeed, it was found that the relatedness proportion is positively correlated with priming, but only at long SOAs (Neely, 1991).

Controlled Processes in the Model

So far, we implicitly assumed that semantic transitions in the network happen automatically. We now turn to a different hypothesis: Semantic transitions may be controlled to some degree; therefore, while SA is the default behavior of the network when no interventions occur, other patterns emerge as soon as subjects attempt to control these transitions.

Controlling transitions can allow our model to implement the 'prediction strategy' of expectancy, if we consider the transition of the semantic network's state from a given prime pattern to another pattern as an 'expected' word for that prime. By default, such expected word is determined according to semantic relatedness principles. However, this conceptualization of expectancy makes it no different than our implementation of SA. What, then, makes expectancy a distinct mechanism? The answer is that expectancy can be modeled as the controlled operation of manipulating transition probabilities according to any information acquired by the subject up to that point, as to induce certain transitions and avoid others. For instance, expectancy can be realized by maintaining just one single transition in the semantic network (as opposed to many transitions in the default case). Another realization can be by controlling the variability of the semantic network's transitions, such that transitions will almost always occur from the prime to its most correlated pattern (as opposed to the more stochastic nature of transitions in the default state). The first suggestion can be implemented by allowing the network to make a single jump, as usual, but then stop any further transitions by lowering the background noise. This means, of course, that noise amplitude must be susceptible to cognitive control. We suggest that this is the equivalence of 'focusing attention' on the prediction. The second suggestion can be implemented by lowering the amplitude of the temporal correlations of the noise, which may be seen as focusing attention on the most probable prediction. Each of these two manipulations, as well as their combination, may have beneficial results: In case they succeed (i.e., the target indeed turns out to be the equivalent of the pattern the network has jumped to), an increase in priming is to be expected compared to the default case since all of the activated neurons of the semantic network would participate in accelerating the response. Without such intervention, the network is much less likely to be converged on the 'right' pattern when the target arrives, which implies that on average, only a minor set of the activated neurons will participate in the acceleration of response. Naturally, if the prediction is wrong, the response might be delayed compared to the default case. Hence this mechanism should be used only when there are good reasons to assume the target is predictable, that is, when the relatedness proportion is high. Moreover, the effect of these manipulations is expected to be most conspicuous on long SOAs, since on short SOAs there is usually not enough time for a transition to occur, let alone a series of transitions.

As an illustrative example, we have repeated simulation 2 for direct, indirect and neutral primes, for short/long SOAs, but this time we manipulated the amplitude of the noise. In one condition, the noise was reduced after the first transition in the semantic network (implementing the first mechanism we suggested for expectancy). In the other condition, no such manipulation was conducted. Figure 5 presents the results. As can be seen, the manipulation increased the facilitation effect, echoing the results in the literature (e.g. Neely, 1991).

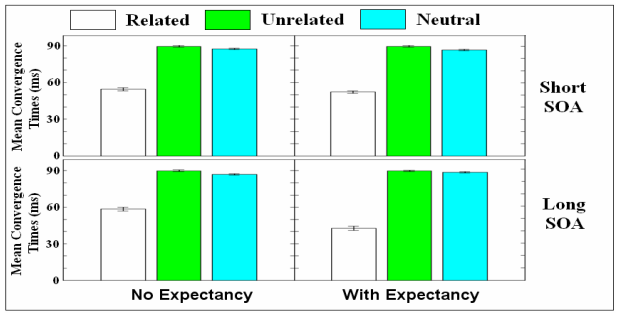


Figure 5: average convergence times of the lexical network with and without an expectancy mechanism

Another controlled process presented in the literature is semantic matching (Neely et al., 1989). This process mainly involves decision making strategies which occur after lexi-

cal access to the target is achieved. In principle, it suggests that subjects engage in comparison between prime and target, which enables them to facilitate word and nonword responses in the lexical decision task (and is also responsible, as a by-product, to inhibition in priming)

While the scope of this paper did not allow us to fully model the semantic matching mechanism (which would necessitate incorporating a decision making mechanism), we would like to point out that any comparison between prime and target must depend on the prime being constantly activated in semantic memory throughout the whole trial, which in turn may indicate that no semantic transitions should occur in the semantic network. This, of course, can be achieved in our model by assuming a reduction in the noise amplitude immediately after lexical access of the prime (as opposed to the expectancy strategy case, where such a reduction is applied only after one semantic transition). We would therefore expect the usage of semantic matching to place severe limitations on spreading activation behavior, and specifically eliminate the indirect priming effect. Interestingly, this is exactly the result found in the literature (e.g. McNamara, 1992; see Neely, 1991, for a review).

General Discussion

Our main goal in the current study was to implement classical semantic processes related to semantic priming, with an emphasis on spreading activation, in a biologically-plausible framework of attractor neural networks. The results demonstrate that the basic characteristics of SA can be embedded in attractor dynamics while maintaining the same explanatory power of the original process. In addition, we show that controlled mechanisms involved in priming such as expectancy can be implemented within the same network, where the definition of 'controlled' is narrowed to the subject's influence on some specific parameters of the network.

Our network implies that real automaticity is the product of correlated representations. Direct semantic priming is a purely automatic process since, by definition, one pattern cannot be activated without partially activating its correlated patterns. On the other hand, processes which require a transformation from one representation to another can in principle be the object of cognitive control. Indirect priming can therefore be avoided by eliminating transitions in the semantic network. Spreading activation, by this view, is best seen as a default mechanism rather than a process which is completely automatic (cf. Smith et al., 2001).

Finally, a pure mathematical interpretation of the dynamics would suggest that the nature of the transitions between patterns in our model takes the form of a Markov-chain, with the average correlation of the network with the various patterns forming a state vector and the transition probability matrix representing word association norms. Controlled strategies therefore represent a change in this matrix from the default values, based on the subject's expectations. Future inquiries may reveal the exact way by which accumulating data affect these probabilities, with Bayesian inference principles possibly governing this procedure.

Conclusion

Attractor neural networks have traditionally struggled with several important aspects of semantic priming compared to the more classical views. We have shown that an attractor network with latching dynamics can in fact implement some of these classical processes and serve as an equally competent model. The model may also be used to predict the time course of priming with SOA, which in turn could be validated by appropriate experiments. Future work will need to specify in a more precise manner the exact ways by which strategies may influence our model's dynamics and how priming is affected by them.

References

Becker, C. A. (1980). Semantic context effects in visual word recognition: An analysis of semantic strategies. *Memory & Cognition*, *8*, 493–512.

Collins, A. M. & Loftus, E. F. (1975). A spreading activation theory of semantic processing. *Psychological Review*, *82*, 407–428.

Herrmann, M., Ruppin, E. & Usher, M. (1993). A neural model of the dynamic activation of memory. *Biological Cybernetics*, *68,* 455–463.

Keefe, D. E., & Neely, J. H. (1990). Semantic priming in the pronunciation task: The role of prospective prime-generated expectancies. *Memory & Cognition*, *18*, 289–298.

Loebel A., & Tsodyks M. (2002). Computation by ensemble synchronization in recurrent networks with synaptic depression. *Journal of Computational Neuroscience*, *13*, 111–124.

Masson, M. E. J. (1995). A distributed memory model of semantic priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21,* 3–23.

McNamara, T. P. (1992). Priming and constraints it places on theories of memory and retrieval. *Psychological Review*, *99*, 650–662.

Neely, J. H. (1991). Semantic priming effects in visual word recognition: A selective review of current findings and theories. In D. Besner & G. Humphreys (Eds.), *Basic processes in reading: Visual word recognition*. Hillsdale, NJ: Erlbaum.

Neely, J. H., Keefe, D. E. & Ross, K. L. (1989). Semantic priming in the lexical decision task: roles of prospective prime-generated expectancies and retrospective semantic matching. *Journal of Experimental psychology: Learning, Memory, and Cognition*, *15,* 1003–1019.

Smith, M. C., Bentin, S. & Spalek, T. M. (2001). Attention constraints of semantic activation during visual word recognition. *Journal of Experimental psychology: Learning, Memory, and Cognition*, *27,* 1289–1298.

Treves, A. (2005). Frontal latching networks: A possible neural basis for infinite recursion, *Cognitive Neuropsychology*, *22*, 276-291.

Tsodyks, M. V. (1990). Hierarchical associative memory in neural networks with low activity level. *Modern Physics Letters B*, *4*, 259-265.

Cognitive Models and the Wisdom of Crowds: A Case Study Using the Bandit Problem

Shunan Zhang (szhang@uci.edu)
Michael D. Lee (mdlee@uci.edu)
Department of Cognitive Sciences, 3151 Social Sciences Plaza A
University of California, Irvine, CA 92697-5100 USA

Abstract

The "wisdom of the crowds" refers to the idea that the aggregated performance of a group of people on a challenging task may be superior to the performance of any of the individuals. For some tasks, like estimating a single quantity, it is straightforward to aggregate individual behavior. For more complicated multidimensional or sequential tasks, however, it is not so straightforward. Cognitive models of behavior are needed, to infer what people know from how they behave, and allow aggregation to be done on the inferred knowledge. We provide a case study of this role for cognitive modeling in the wisdom of crowds, using a multidimensional sequential optimization problem, known as the bandit problem, for which there are large differences in individual ability. We show that, using some established cognitive models of people's decision-making on these problems, aggregate performance approaches optimality, and exceeds the performance of the vast majority of individuals.

Keywords: Wisdom of crowds, Cognitive models, Bandit problem, Hierarchical Bayesian modeling

Introduction

An enticing idea in the study of individual and group decision-making is the phenomenon known as the "wisdom of crowds". The idea is that, by aggregating the behavior of a group of people doing a challenging task, it is possible for group performance to match or exceed the performance of any of the individuals. Surowiecki (2004) provides an extensive survey of wisdom of crowds results over a diverse set of human endeavors and decision-making situations, ranging from guessing the weight of an ox at a county fair, to inferring the location of a missing submarine, to predicting the outcome of sporting events. Recent research in cognitive science has looked at issues including whether it is possible to have a "crowd within", such that multiple estimates from the same person can be combined to improve their performance (Vul & Pashler, 2008).

While the exact conditions needed for group performance to exceed individual performance are not completely understood, it seems clear that crowds can be wise in any situation where people have some partial knowledge, and the gaps in their knowledge are subject to individual differences. Under these circumstances, aggregation of individual decisions can serve to amplify the common signal and reduce the idiosyncratic noise, leading to superior group performance.

One challenge in producing wisdom of crowds effects arises when tasks are more complicated than estimating a single quantity, or predicting a simple outcome. Many interesting and real-world decision-making situations are inherently multidimensional or sequential. In these situations, it is often not possible to combine the raw behaviors of people, because they are not commensurate. For example, imagine trying to combine the expertise of basketball fans trying to predict the result of an eight-team single elimination tournament, with quarter-finals, semi-finals and a final. Based on their decisions about the quarter-finals, these people may be making decisions about different teams in the semi-finals and final. This makes simple aggregation based on their raw decisions impossible for the later rounds.

For more difficult decision problems like these, we believe cognitive science has a key role to play in wisdom of the crowd research. Rather than aggregating people's behaviors, it is necessary to aggregate their knowledge, as *inferred* from their behavior. This inference needs models of cognition, accounting for how latent knowledge manifests itself as observed behavior within the constraints of a complicated task. Steyvers, Lee, Miller, and Hemmer (in press) present an example of this approach, using Thurstonian models of judgment to combine people's ranking decisions for a variety of general-knowledge questions, such as the chronology of the US Presidents.

In this paper, we present a case study of the application of cognitive models for a sequential task known as the bandit problem. By applying a series of existing models of human decision-making on the task to a variety of data sets, we show that it is sometimes possible to produce aggregate performance that is near optimal, and far exceeds the performance of most of the individuals. We discuss what sort of properties cognitive models might need to achieve this sort of useful aggregation of individual knowledge.

Bandit Problems

Bandit problems are a type of sequential decision-making problem widely studied in statistics and machine learning (Gittins, 1979; Kaebling, Littman, & Moore,

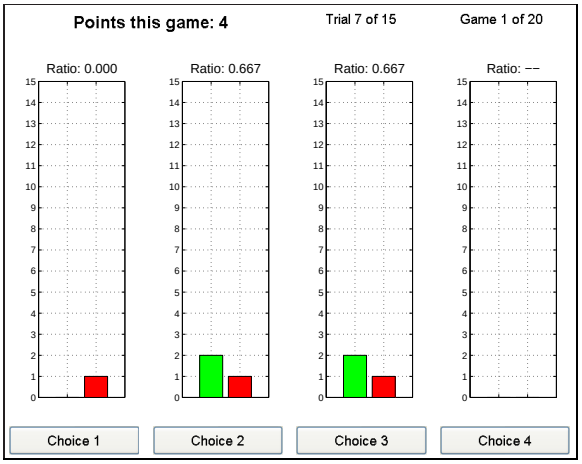


Figure 1: An experimental interface, giving an example of a Bandit problem.

1996; Sutton & Barto, 1998), as well as in cognitive science (Cohen, McClure, & Yu, 2007; Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006; Steyvers, Lee, & Wagenmakers, 2009). In Bandit problems, a decision maker chooses from a set of alternatives with fixed but unknown reward rates, which are drawn from a fixed but unknown environment, with the goals of maximizing the total number of rewards after a fixed number of trials.

A representative experimental interface of Bandit problems is shown in Figure 1. The four large panels contain information of choices and outcomes on four alternatives. On each trial, an alternative is chosen, and either succeeds in giving a reward (green, light) or fails (red, dark). At the top of each panel, the ratio of successes, defined as the ratio of successes to total choices, is shown. The interface provides a count of the total number of rewards obtained up to the current trial. The current game and trial are also shown.

The bandit problem has a well-known optimal decision-making process (e.g., Kaebling et al., 1996, p. 244), calculated by dynamic programming. This allows human decision-making, and plausible psychological models of decision-making, to be assessed in terms of their optimality. In particular, Bandit problems provide a natural task to study the inherent trade-off between exploration (seeking rewarding alternatives among those relatively unexplored) and exploitation (staying with alternatives known to be reasonably good) inherent in many real-world sequential decision-making situations.

Human Data

We use data from three experiments. In the first experiment, reported by Steyvers et al. (2009), a total of 451 participants completed a total of 20 bandit problems, each with 4 alternatives and 15 trials. Reward rates were drawn for each alternative independently from a Beta$(2,2)$ distribution. The reward rates were drawn only once, but the order of the games was randomized.

The second and third experiments involve new data. A total of 47 and 31 participants, respectively, completed 100 bandit problems, all with 4 alternatives and 16 trials. For the second experiment, the reward rates were drawn independently for each game from Beta$(8,4)$ (called a "plentiful" environment, because reward rates tend to be high). For the third experiment, reward rates came from a Beta$(4,8)$ (called a "scarce" environment, because reward rates tend to be low)

Four Decision-Making Models

In this paper, we consider four well-established models of decision-making on bandit problems. These come from the reinforcement- and machine-learning literatures (see Sutton & Barto, 1998), and have previously been examined as models of human decision-making (e.g., Lee, Zhang, Munro, & Steyvers, 2009).

Win-Stay Lose-Shift

Perhaps the simplest reasonable approach for making bandit problem decisions is the Win-Stay Lose-Shift (WSLS) heuristic. In its deterministic form, it assumes that the decision-maker continues to choose an alternative following a reward, but shifts to the other alternative following a failure to reward. In the stochastic form we use, the probability of staying after winning, and the probability of shifting after losing, are both parameterized by the same probability γ.

Extended Win-Stay Lose-Shift

A natural, and psychologically-motivated, extension to the WSLS model is to have different rates for staying after a reward (i.e., reinforcement) and shifting after a lack of reward (i.e., negative reinforcement). Formally, in our extended WSLS model, a decision-maker stays with probability γ^w following a reward, but shifts with probability γ^l following a failure to reward.

ε-Greedy

The ε-greedy model assumes that decision-making is driven by a parameter ε that controls the balance between exploration and exploitation inherent in bandit problems. On each trial, with probability $1-\varepsilon$ the decision-maker chooses the alternative with the greatest estimated reward rate (i.e., the greatest proportion of rewards obtained for previous trials where the alternative was chosen). This can be conceived as an 'exploitation' decision. With probability ε, the decision-maker chooses randomly. This can be conceived as an 'exploration' decision.

ε-Decreasing

The ε-decreasing model is a variant of ε-greedy, in which the probability of an exploration move decreases as trials progress. In its most common form, which we use, the ε-decreasing model starts with an exploration probability ε'

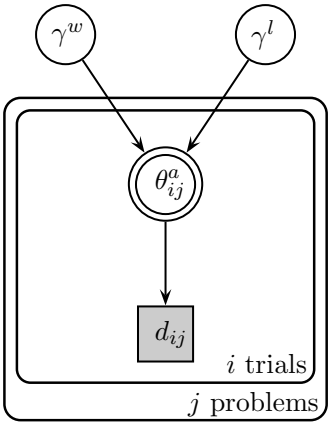


Figure 2: Bayesian graphical model for the extended WSLS decision-making model.

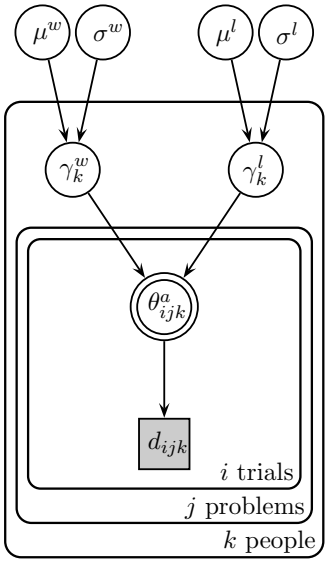


Figure 3: Bayesian graphical model for a hierarchical version of the extended WSLS decision-making model, which allows for individual-level parameter variation.

on the first trial, and then uses an exploration probability of ε'/i on the ith trial.

Modeling Analysis

In this section, we implement the four decision-making models in a way that allows for differences in individual behavior to be aggregated, culminating in model-based wisdom of crowds analyses of our experimental data sets.

Bayesian Graphical Model Implementation

We implemented all four decision-making models using the formalism provided by Bayesian graphical models, as widely used in statistics and computer science (e.g., Koller, Friedman, Getoor, & Taskar, 2007). A graphical model is a graph with nodes that represents the probabilistic process by which unobserved parameters generate observed data. Details and tutorials are aimed at cognitive scientists are provided by Lee (2008) and Shiffrin, Lee, Kim, and Wagenmakers (2008). The practical advantage of graphical models is that sophisticated and relatively general-purpose Markov Chain Monte Carlo (MCMC) algorithms exist that can sample from the full joint posterior distribution of the parameters conditional on the observed data. More specifically, for our purposes, graphical models can be specified that naturally combine information across multiple sources, and so can model the individual differences at the heart of the wisdom of crowds phenomenon.

As a concrete example, Figure 2 shows the graphical model implementation of the extended WSLS model. The two model parameters, the probability of win-stay γ^w and lose-shift γ^l, are shown as unshaded (i.e., unobserved) and circular (i.e., continuous) variables. These determine the probability of the ath alternative being chosen on the ith trials of the jth game, as

$$\theta^a_{ij} = \begin{cases} \gamma^w & \text{if succeeded on } a \text{ last trial} \\ 1-\gamma^l & \text{if failed on } a \text{ last trial} \\ (1-\gamma^w)/3 & \text{if succeeded on } \bar{a} \text{ last trial} \\ \gamma^l/3 & \text{if failed on } \bar{a} \text{ last trial,} \end{cases}$$

where $\bar{a}$ refers to not choosing the ath alternative. Since θ_{ij} is a deterministic function of γ^w and γ^l, it is shown as a double-bordered node. Given the choice probabilities in θ^a_{ij}, the actual decision made by the ith trial of the jth problem— which is represented by a shaded square node d_{ij}, since it is observed, and discrete—is modeled as $d_{ij} \sim \text{Discrete}(\theta^1_{ij} \ldots, \theta^4_{ij})$.

Parameter Differences

One obvious possibility for individual differences is that two people—even if they are both using, for example, extended WSLS—might not have the same probabilities of wining and staying or losing and shifting. To accommodate variation in these parameters on an individual-by-individual uses, we use a *hierarchical* or *multi-level* approach. The updated graphical model is shown in Figure 3. In this model, the parameters for individual people are drawn from over-arching Gaussian distributions, so that, for the kth person, $\gamma^w_k \sim \text{Gaussian}(\mu^w, \sigma^w)$, and $\gamma^l_n \sim \text{Gaussian}(\mu^l, \sigma^l)$. This allows different people to have different parameter values, while still estimating the mean parameter value of the group as a whole.

We implemented the graphical model in Figure 3, as well as analogous graphical models for the three other decision-making models, in WinBUGS (Spiegelhalter, Thomas, & Best, 2004). This software uses a range

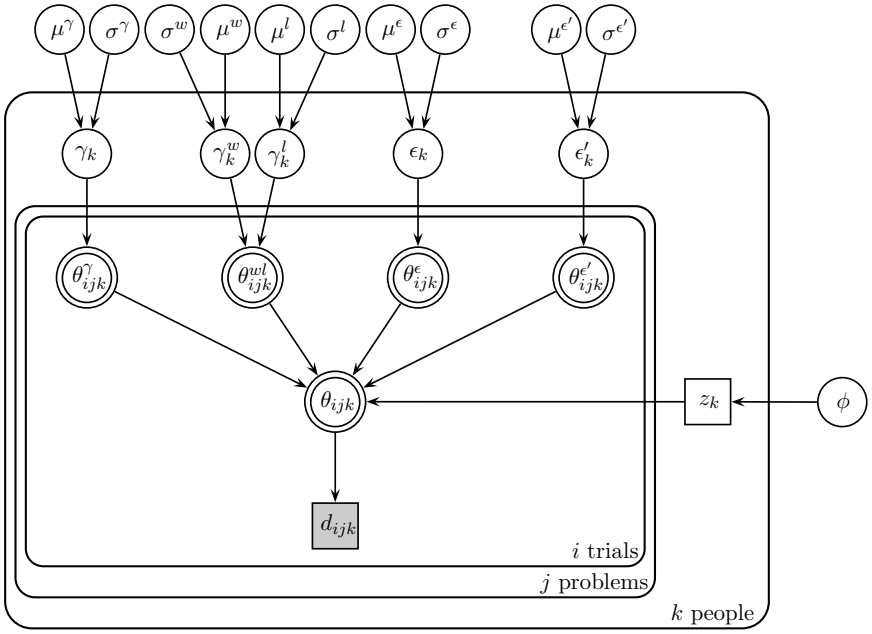


Figure 4: Graphical model using a hierarchical mixture of all four hierarchical decision-making models.

Table 1: Means, and standard deviations in brackets, of the group distributions for each parameter in the four decision-making models.

Parameter	Exp. 1	Exp. 2	Exp. 3
γ	0.71 (.10)	0.70 (0.10)	0.52 (0.10)
γ^w	0.99 (0.27)	0.97 (0.19)	0.81 (0.18)
γ^l	0.59 (0.25)	0.28 (0.23)	0.37 (0.23)
ε	0.24 (0.10)	0.18 (0.11)	0.42 (0.12)
ε'	0.61 (0.11)	0.61 (0.11)	0.90 (0.14)

of MCMC computational methods, including adaptive rejection sampling, splice sampling, and Metropolis-Hastings to perform posterior sampling (e.g., MacKay, 2003). For all four decision-making models, we made inferences about individual- and group-level parameters for all three data sets, using all of the participants. In each analysis, we collected 1,000 samples from 2 chains, collected after a burn-in period of 1,000 samples, and using standard checks for convergence.

Table 1 summarizes individual differences in parameters for each decision-making model, giving the means and standard deviations for each parameter in the hierarchical analysis. Remembering that experiments 1, 2, and 3 correspond to neutral, plentiful and scarce environments, the aggregated group parameters make sense. For example, there is more winning and staying (e.g., in the γ and γ^w parameters) in environments that deliver rewards, and there is more random exploration (e.g., in the ε and ε') in scarce environments that are not delivering rewards. The reasonably large standard deviations for most group distributions also indicate that there are significant individual differences.

Model Differences

An even more fundamental source of individual differences arises when different people use different decision processes. Rather than just varying the parameters of a model, people may differ in terms of which decision-making model they use. We accommodate this type of individual differences using a *mixture* or *latent assignment* model where people are categorized into different model-users.

The graphical model for achieving this mixture of decision models, while retaining the possibility of parameter variation within each model, is shown in Figure 4. Hierarchical versions of all four decision-making models—those used individual to assess parameter variation in the previous section—are all shown.

The key addition, in terms of individual differences, involves the model indicator variable z_k, which indexes which of the four models the kth participant uses. That is, depending on whether z_k is 1, 2, 3 or 4, the kth participant uses WSLS, the extended WSLS, ε-greedy or ε-decreasing to make their bandit problem decisions. The latent indicator variable has prior $z_k \sim \text{Categorical}(\phi)$, where ϕ is a latent base-rate, measuring the proportion of people who follow each model. We use the prior $\phi \sim \text{Dirichlet}(1/4,\ldots,1/4)$, so that there is no initial bias towards one decision model over another.

Table 2 gives the posterior expectation of the base-rate parameter ϕ, for all three experiments. This provides a natural summary of what proportion of people were us-

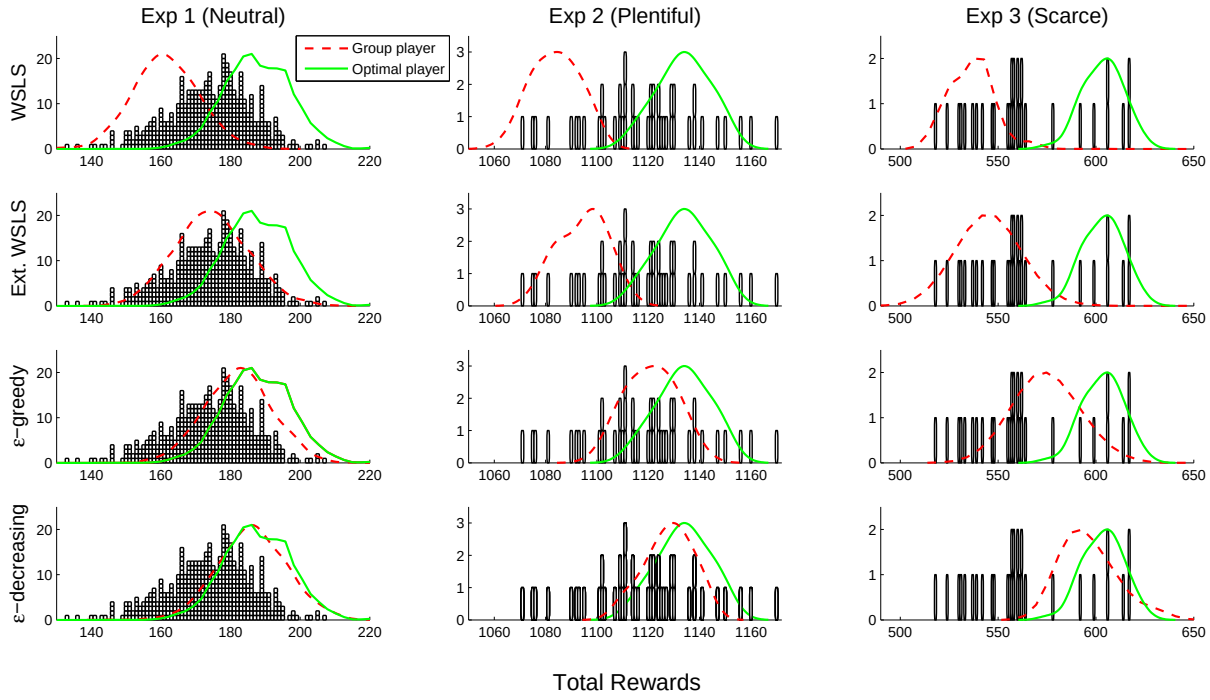


Figure 5: Distribution of rewards for individual participants, the group model, and the optimal decision-making process, for each decision-making model and each experiment. See text for details..

Table 2: Proportion of people using each model, for the three experiments, as measured by the posterior expectation of the ϕ parameter.

Model	Exp. 1	Exp. 2	Exp. 3
WSLS	0%	0%	0%
Extended WSLS	75%	81%	70%
ε-greedy	22%	16%	29%
ε-decreasing	3%	3%	1%

ing each of the four models, and so summarizes individual differences results at this fundamental level. The findings are consistent across all three experiments—even though they have different distributions of reward rates—with the clear majority of the participants inferred to be using the extended WSLS model, and a minority using ε-greedy. The proportion inferred to be using the other two models is negligible.

Wisdom of Crowds Analysis

Our modeling of individual differences in models and parameters immediately allows a range of wisdom of the crowd analyses. The most basic analyses involve taking each of our decision-making models, and using the inferred group mean in the hierarchical analysis, as shown in Table 1 as the aggregate of individual performance.[1] This approach solves the problem of aggregating the knowledge of different people solving different, but related, bandit problems. Rather than aggregating their behavioral choices, we are aggregating the psychology parameter values that lead to those choices.

To complete the model-based wisdom of crowd analyses, we used the group mean parameter values to define a "group model" that used the same decision-process, and completed the same problems given to participants in each of the three experiments. Because the number of rewards obtained is inherently stochastic, we repeated this many times to approximate the distribution of rewards. We also applied the optimal decision-making process to each experiment, to approximate the best possible distribution of rewards for each experiments

The results are shown in Figure 5. The columns correspond to the three experiments. The rows correspond to the WSLS, extended WSLS, ε-greedy and ε-decreasing decision models. Within each panel, the squares piled into histograms show the distribution of performance (i.e., how many rewards were obtained) for the individual participants. The two curves then correspond to the distribution of performance for the group model (red, dotted line) and the optimal decision process (green, solid line).

Figure 5 shows that some of our decision-making

[1]We tried more involved analyses, using the full mixture model in Figure 4 to sample a model, and then parameters, to define a group model. We never found a wisdom of crowd effect comparable to what was achieved with the basic analyses, so we just report those.

models do produce a clear wisdom of the crowds effect, whereas others do not. The distributions of rewards for the group model formed by the WSLS and extended WSLS models does not improve on the distribution of individual performance, and are not close to optimal. For the ε-greedy and ε-decreasing group models, however, there is significant improvement. In particular, the ε-decreasing group model has a distribution of rewards that is extremely close to the optimal distribution for all three experiments.

Discussion

There are some intriguing features of our wisdom of crowd results presented in Figure 5. Most obviously, it is very encouraging that it is possible to take a simple decision-making model like ε-decreasing, take the window it provides onto human decision-making, and produce an aggregate decision-maker that performs near optimally. But, we note that this wisdom of crowd effect is not achieved for all of the cognitive models we tried, and, most particularly, was not achieved for the extended WSLS that provided the best account of the vast majority of individual behavior, as detailed in Table 2.

We think the explanation for this finding is that , the ε-greedy and ε-decreasing models are able to match more closely optimal behavior. Detailed analysis showing this was presented by Lee et al. (2009) and makes intuitive psychological sense. Neither WSLS model is sensitive to which trial in the total sequence is being completed, which is important information in managing the trade-off between early exploration and late exploitation. As a consequence of this sub-optimality, it is not surprising a wisdom of crowd effect was not achieved for these simple models.

What is more surprising is that the effect could be achieved for a decision-making model like ε-decreasing that is not an especially good account of individual behavior. An important topic for future wisdom of crowds research is to identify what properties of cognitive models are important in producing good aggregations of individual knowledge. Being able to mimic optimal behavior is a start, but it is not currently clear how effective models must be able to account for what people do.

More generally, we think our case study with bandit problems demonstrates a very general approach for applying cognitive models to study and use the wisdom of crowds phenomenon. Using graphical models allows hierarchies of parameters, and mixtures of decision processes, to combine the individual differences in people, at the level of their basic knowledge about a task. This leads naturally to a principled sort of aggregation that is applicable to complicated, multidimensional and sequential tasks, which might be among those most needing the pooling of individual capabilities to achieve good performance.

Acknowledgments

This work is was supported by an award from the Air Force Office of Scientific Research (FA9550-07-1-0082).

References

Cohen, J. D., McClure, S. M., & Yu, A. J. (2007). Should I stay or should I go? Exploration versus exploitation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*, 933–942.

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, *441*, 876–879.

Gittins, J. C. (1979). Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society, Series B*, *41*, 148–177.

Kaebling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, *4*, 237–285.

Koller, D., Friedman, N., Getoor, L., & Taskar, B. (2007). Graphical models in a nutshell. In L. Getoor & B. Taskar (Eds.), *Introduction to statistical relational learning.* Cambridge, MA: MIT Press.

Lee, M. D. (2008). Three case studies in the Bayesian analysis of cognitive models. *Psychonomic Bulletin & Review*, *15*(1), 1–15.

Lee, M. D., Zhang, S., Munro, M. N., & Steyvers, M. (2009). Using heuristic models to understand human and optimal decision-making on bandit problems. In A. Howes, D. Peebles, & R. Cooper (Eds.), *Proceedings of the Ninth International Conference on Cognitive Modeling — ICCM2009.* Manchester, UK.

MacKay, D. J. C. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge: Cambridge University Press.

Shiffrin, R. M., Lee, M. D., Kim, W.-J., & Wagenmakers, E.-J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive Science*, *32*(8), 1248–1284.

Spiegelhalter, D. J., Thomas, A., & Best, N. G. (2004). *WinBUGS Version 1.4 User Manual*. Cambridge, UK: Medical Research Council Biostatistics Unit.

Steyvers, M., Lee, M. D., Miller, B., & Hemmer, P. (in press). The wisdom of crowds in the recollection of order information. In J. Lafferty & C. Williams (Eds.), *Advances in neural information processing systems, 23.* Cambridge, MA: MIT Press.

Steyvers, M., Lee, M. D., & Wagenmakers, E.-J. (2009). A Bayesian analysis of human decision-making on bandit problems. *Journal of Mathematical Psychology*, *53*, 168–179.

Surowiecki, J. (2004). *The wisdom of crowds*. New York: Random House.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge (MA): The MIT Press.

Vul, E., & Pashler, H. (2008). Measuring the crowd within: Probabilistic representations within individuals. *Psychological Science*, *19*(7), 645–647.

The Accuracy of Small-Group Estimation and the Wisdom of Crowds

Michael D. Lee (mdlee@uci.edu)
Jenny Shi (jshi1@uci.edu)
Department of Cognitive Sciences, 3151 Social Sciences Plaza A
University of California, Irvine, CA 92697-5100 USA

Abstract

We measure the ability of people to estimate the price of familiar household items in a variety of contexts. We manipulate whether estimation is done alone or with others, whether it is done independently or with the knowledge of the estimates of others, and whether it is done in a cooperative or competitive environment. From these basic estimation data, we construct a series of aggregated group estimates, exploring the conditions under which a small group of three people provide the most accurate information. We compare the performance of various small-group estimates to standard Wisdom of Crowds analysis, and find that priming people, or placing them in a cooperative group setting, is less effective than averaging the independent estimates of individuals. We also find, however, that it is possible to extract relatively more information from the decisions people make in a competitive group setting, using cognitive models of their decision-making.

Keywords: Wisdom of crowds, group estimation, Price is Right, game show, cooperative vs competitive decision-making



Figure 1: Basic experimental interface. On each trial, a picture and description of an item is shown. Once an estimate has been made, the true price is presented.

Introduction

A basic question for cognitive and social psychology involves how best to extract information from people. There is a large literature on the performance of groups in reaching good decisions in various contexts (see Kerr & Tindale, 2004; Hastie, 1986, for reviews), with accompanying theoretical positions ranging from believing in the robust effectiveness of group decision-making (e.g., Hastie & Kameda, 2005) to the destructive possibilities of "group think" (e.g., Moscovici & Zavolline, 1969).

A recent contribution to the issue of whether and how groups of people make effective decisions involves the "Wisdom of Crowds" phenomenon (Surowiecki, 2004). This refers to the empirical finding that an aggregated decision, made by combining the individual decisions of many people, can often perform as well as or better than the majority of the individual decisions themselves.

In this paper, we examine group decision-making and the Wisdom of Crowds phenomenon in a simple estimation setting. We ask people to estimate the price of everyday household objects, with which they people are familiar, but are unlikely to have exact price knowledge. We ask for these estimates in a wide variety of individual and group settings. These settings manipulate whether estimation is done alone or in the presence of others, whether it is done independently or with the knowledge of the estimates of others, and whether group estimation is done in a cooperative or competitive environment.

To examine how these manipulations affect the accuracy of small-group estimation, we focus on a specific research question. The question is: how well do different ways of using the knowledge of just three people to estimate the price perform, and how does this level of performance relate to standard Wisdom of Crowds aggregation with more people?

Experiment

Materials

Stimuli We used two sets of 50 household items, with pictures and descriptions sourced from on-line shopping websites. Both stimulus sets followed the same price distribution, with totals approximately uniformly distributed between $5 and $45.

Interface An example of the basic experimental interface is shown in Figure 1. On each trial, a picture and

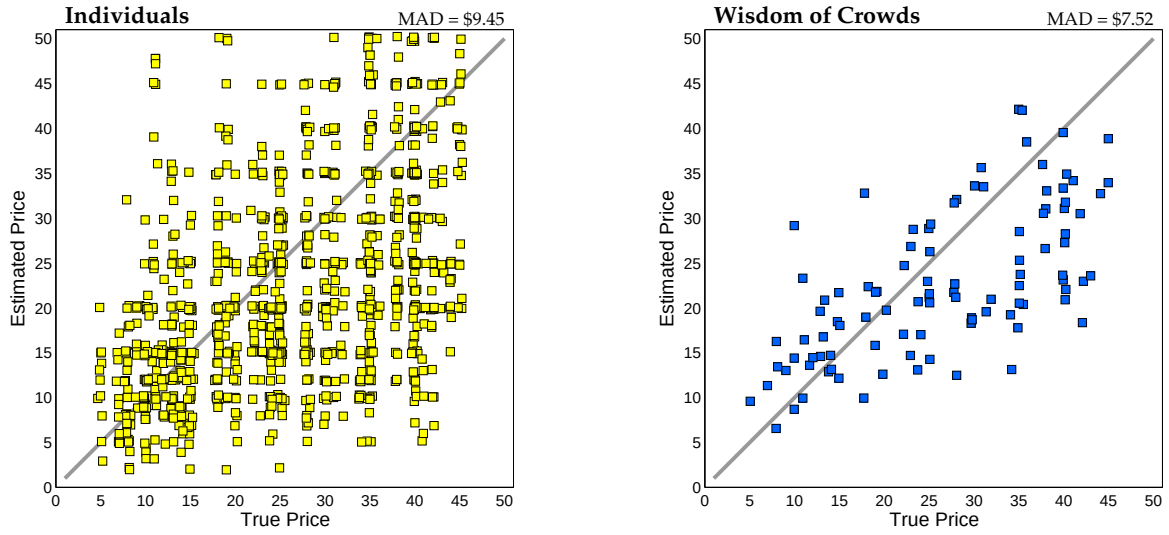


Figure 2: Relationship between true item prices and individual estimates (left panel), and Wisdom of Crowds estimate formed by averaging over all individuals (right panel). (MAD=Mean Absolute Deviation)

description of a prize is shown. Once an estimate has been made, the true price is presented. A counter shows how many of the 50 trials have been completed.

Methods

Using the same sets of items and basic interface, we collected price estimates under a variety of experimental conditions. These conditions manipulated whether estimation was done in an individual or group setting, whether estimates were done independently or with knowledge of other estimates, and whether estimation was done in cooperative or competitive setting.

Individual Estimates The simplest experimental condition just collects individual estimates for each of the 50 items, presented in a random order. A total of 22 participants completed this condition.

Primed Individual Estimates The 'primed' or 'calibrated' condition was the same as the individual condition, except that when each item was presented, the estimates of two other people were also presented. These estimates were drawn at random from the estimates made for the same prize in the individual condition. A total of 25 participants completed this condition.

Cooperative Group Estimates In the cooperative group condition, three people were co-located, and viewed the same experimental interface. They were asked to provide estimates sequentially, hearing the earlier estimates. After all three estimates had been made, the group was asked to form a consensus estimate, through unstructured discussion. The same three people completed all 50 trials, and the order in which they estimated was rotated between each trial. A total of 15 people completed this condition, forming 5 groups.

Competitive Group Estimates In the competitive group condition, three people played a version of the "Price is Right" game show, which has been used previously as a formalism to study competitive decision-making (e.g., Berk, Hughson, & Vandezande, 1996). They were asked to provide bids sequentially, hearing the earlier bids, with the goal of bidding *as close as possible to the true price without exceeding the true price*. People were not allowed to repeat an earlier bid, and the order of making bids was again rotated systematically after each trial. A total of 15 people completed this condition, forming 5 groups.

Basic Results

Bounds on Performance There are two worthwhile preliminary analyses that can serve to give bounds on the accuracy of estimation. The first of these simply considers each individual estimate, and is shown in the left panel of Figure 2. The mean average deviation between the estimated and true price is $9.45. This serves as a sensible baseline for accuracy, since it represents what how well a single person will perform on average.

The second preliminary analyses averages *all* of the individuals who gave estimates for each prize. This corresponds to a standard "Wisdom of the Crowds" analysis, and is shown in the right panel of Figure 2. The mean absolute deviation is a much-improved $7.52, and can reasonably serve as an upper bound on performance.

Simple Three Person Estimates

Figure 3 shows the performance of four simple ways to combine the information provided by three people to estimate the prices. These involve, the individual, primed individual, and cooperative group estimation contexts.

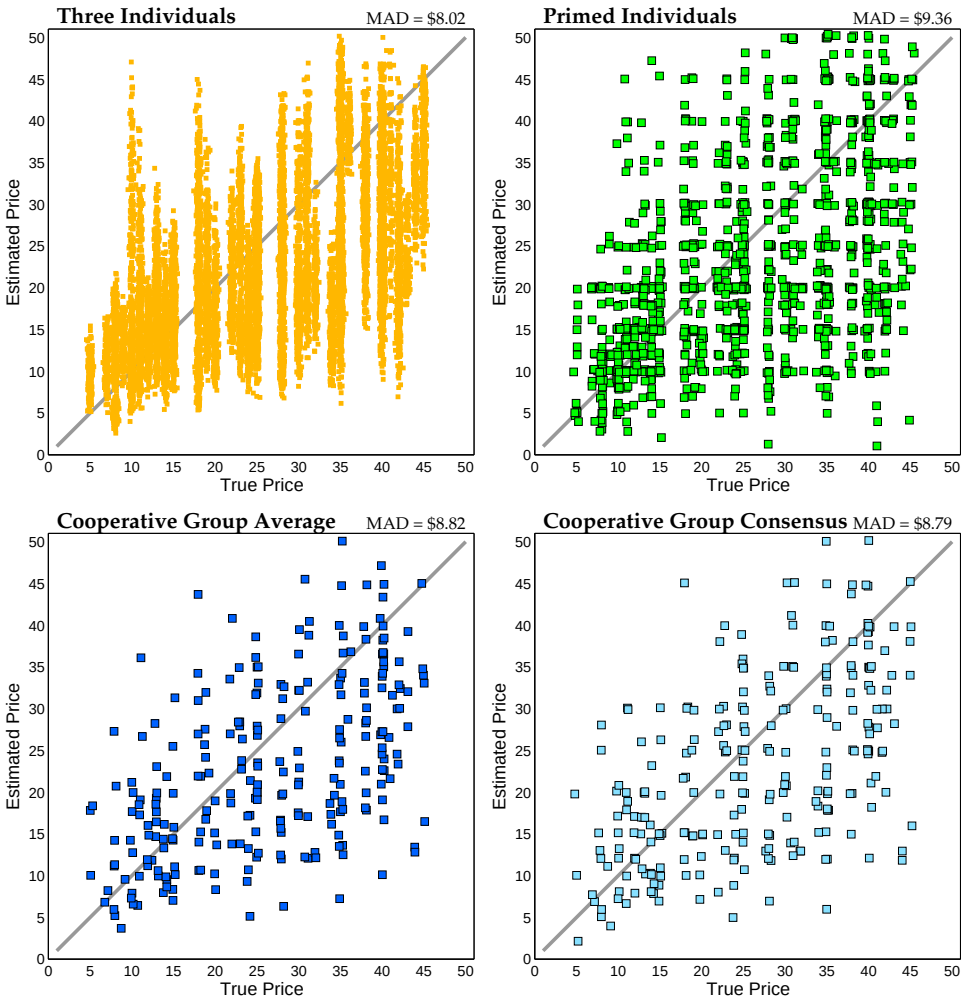


Figure 3: Relationship between true item prices and group estimates, formed from three people, by (top left) averaging the individual estimates of three people, (top right) priming an individual with the earlier estimates of two other people, (bottom right) averaging the estimates of three people made sequentially in a group setting, and (bottom right) the consensus opinion of a group of three people. (MAD=Mean Absolute Deviation)

Three Individuals The most obvious, given the estimates of three people, is simply to average them, as in a standard Wisdom of the Crowds analysis.[1] The performance of this approach is shown in the top left panel of Figure 3, which considers all possible groups of three people using individual estimates. The mean absolute difference is $8.02. As would be expected, this difference lies between that already observed for single individuals, and for all individuals considered together.

Primed Individuals Another estimate based on the information provided by three people comes directly from the primed estimate. This is the estimate of a single individual working along, but in the knowledge of two other people's estimates. The performance of primed estimates is shown in the top right panel of Figure 3. The mean absolute difference is $9.36, which barely improves upon the accuracy of estimates of single non-calibrated individuals.

Cooperative Average The bottom left panel shows the performance of the average of the three people in the cooperative group condition. The mean absolute difference is $8.82. This is better than single individuals, but does not come close to the level of performance achieved by averaging three estimates made independently.

[1]For all of the analyses we present involving the averaging of estimates, we also examined taking the median, or rounding answers to the nearest dollar. Rarely did performance, as measured by the Mean Absolute Deviation, change by more than a few cents, and never did it suggest different conclusions from those we present based on the mean.

Cooperative Consensus Finally, the bottom right panel of Figure 3 shows the performance of the consensus estimates reached by the groups of three people. The mean absolute difference is $8.79, which is very similar to the average of the group estimates. Taken together, these results suggest that being in a cooperative group setting hinders the generation of accurate estimates.

Competitive Group Analysis

Analyzing estimation performance for the competitive "Price is Right" condition requires more involved inference than averaging. This is because the bids that people make do not necessarily correspond to their actual best estimate of the price of a prize. In the competitive context formalized by the rules of the game, it is often sensible for a player to make a bid that is very different from what they believe the price to be.

This strategic relationship between bids and estimates is most easily seen for the final bid made by Player 3. If the previous bids are $35 and $40, then the best final bid is either $1, $36 or $41. One of these choices is rational, in the sense that it will maximize the probability that Player 3 wins the game. Which choice is rational depends on what Player 3 knows about the price of the prize. If, for example, they believe it is most likely somewhere below $35, then the $1 final bid is optimal.

For this reason, it does not make sense to combine the bids from the competitive group setting as if they were estimates, and just average them. Instead, inferences need to be made about what estimates the players have in their heads, based on their bids. This inference requires a model of decision-making that accounts for how estimates become bids, in the context of the game.

Inferring a Group Estimate from Bids

The decision model we used for inference makes two key assumptions. The first is a representational assumption, which is that all of the players have partial knowledge of the price of a prize, and that their uncertainty can be represented by the same Normal distribution. The second is a decision-making assumption, which is that players make the bid that maximizes their probability of winning the game. Given these assumptions, our inferential goal is to find the mean of the Normal distribution, since it represents the average price, based on the players' knowledge.

Formally, given a Normal distribution with mean μ and standard deviation σ, we can define a 'win' function $w_x(a,b,c,\mu,\sigma)$ for the probability the xth player will win, given bids a, b, c, for Players 1, 2, and 3, respectively. This win probability is just the area under the Normal curve between the bid of the xth player, and the next highest bid (or the maximum of $50, if it is the highest bid). On the basis of this win function, we can formalize what constitutes optimal bidding for each player.

Player 3 Given existing bids a, and b the probability Player 3 will win if they made the bid c is just

$$\pi_c(c \mid a,b,\mu,\sigma) = w_3(a,b,c,\mu,\sigma),$$

and so one way of formalizing what it means to be a rational player, is that they will choose according to these probabilities, so that

$$p_3(c \mid a,b,\mu,\sigma) = \frac{\pi_3(c \mid a,b,\mu,\sigma)}{\sum_{d'} \pi_3(c' \mid a,b,\mu,\sigma)}.$$

Player 2 Given an existing bid a, the probability Player 2 will win if they made the bid b, assuming Player 3 subsequently 'behaves optimally' and bids according to $p_3(c \mid a,b\mu,\sigma)$ above, is

$$\pi_2(b \mid a,\mu,\sigma) = \sum_c p_3(c \mid a,b,\mu,\sigma) w_2(a,b,c,\mu,\sigma).$$

So, if Player 3 makes their bid decision according to these probabilities, they will choose

$$p_2(b \mid a,\mu,\sigma) = \frac{\pi_2(b \mid a,\mu,\sigma)}{\sum_{b'} \pi_2(b' \mid a,\mu,\sigma)}.$$

Player 1 Player 1 provides the first bid. If they bid a, their probability of winning, assuming subsequent optimal behavior is

$$\pi_1(a \mid \mu,\sigma) = \sum_b p_2(b \mid a,\mu,\sigma) \sum_c p_3(c \mid a,b,\mu,\sigma) \times w_1(a,b,c,\mu,\sigma).$$

This gives the bid decision probabilities

$$p_1(a \mid \mu,\sigma) = \frac{\pi_1(a \mid \mu,\sigma)}{\sum_{a'} \pi_1(a' \mid \mu,\sigma)}.$$

Final Inference The joint posterior distribution over the parameters of the Normal representing people's knowledge is given by Bayes Rule

$$\begin{aligned} & p(\mu,\sigma \mid a,b,c) \\ \propto\ & p(a,b,c \mid \mu,\sigma)\, p(\mu,\sigma) \\ =\ & p(c \mid a,b,\mu,\sigma)\, p(b \mid a,\mu,\sigma)\, p(a \mid \mu,\sigma)\, p(\mu,\sigma). \end{aligned}$$

We put a simple improper flat prior on $p(\mu,\sigma)$, and all of the other likelihood terms are available from the optimal decision-making analysis.

There are many potential ways the $p(\mu,\sigma \mid a,b,c)$ could be used to estimate the final group price. We use probably the simplest possible approach, and find the mode (i.e., the MAP estimate) $(\mu^*,\sigma^*) \mid a,b,c$, and use μ^* as the price estimate of the competitive group, based on their bids.

Demonstration of Inference

Figure 4 provides a concrete example of the inference process used to estimate the price of a prize from the bidding in the competitive "Price is Right" game. The example relates to one trial for one of our groups, in which the players bid $13, $10 and $1. To find which Normal distribution best explains these bids, under the assumption that people bid to maximize their chance of winning, we

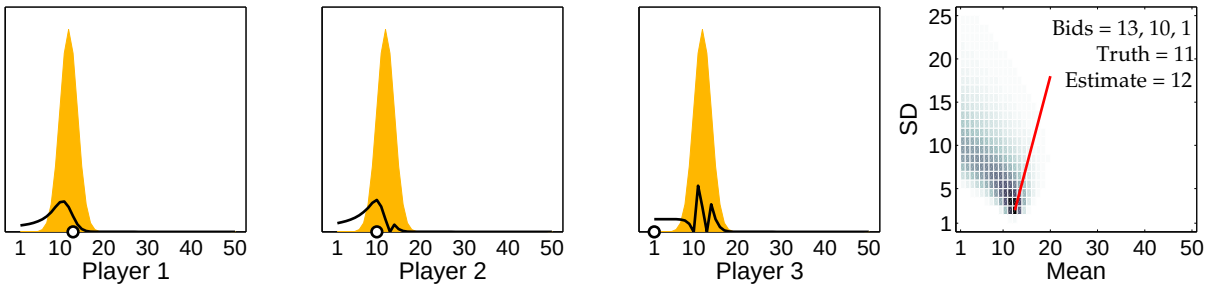


Figure 4: Inference process to find a group estimate from the bids in a competitive "Price is Right" game.

exhaustively test every Normal distribution with a mean of $1, 2, \ldots, 50$ and standard deviation of $1, 2, \ldots, 25$.

The first three panels of Figure 4—which correspond to the decision making of Players 1, 2 and 3—all show the same particular Normal in the background, with a mean of \$12 and a standard deviation of \$3. The black line then shows the probability of each player winning the game, if they made each possible bid between \$1 and \$50. The white circle represents the bid they actually made.

Intuitively, it is easiest to understand this analysis by looking at Player 3. Here, it is already known that the previous bids are \$13 and \$10. The black line shows that the probability of Player 3 winning peaks around \$14 and \$11, one above the earlier bids, and is also high for bids starting at \$1, up until the point where the Normal says it becomes possible the true price might lie.

Looking at all three players, it is clear that this particular Normal distribution gives predictions that are reasonably consistent with the bids actually made. It peaks at the right bid for Player 2, and gives appreciable probably to the bids of Players 1 and 3. In fact, the Normal shown corresponds to the most likely one, out of all the possibilities considered. This result is shown in the rightmost panel of Figure 4. In this plot, each point corresponds to a Normal distribution, and the darker it is shaded, the more probable that Normal made the observed bid data. The mode is at $\mu = 12, \sigma = 3$, and so the final estimate we infer is \$12. As it happens this is very close to the true \$11 price of the prize for this trial.

Notice that simply averaging the bids would not produce the same estimate, because it would treat the \$1 bid as a literal estimate, rather than a strategic attempt to win the game, based on the belief that earlier bids may have been too high.

Results

The performance of the inferred three-person estimates based on the competitive game bids is shown in Figure 5. The mean difference is \$8.05. This is a large improvement on the cooperative group average and consensus estimates, and is comparable to the accuracy obtained by averaging three individual estimates.

The results for all of the three-person estimates, and their relationship to Wisdom of Crowds averaging, are summarized in Figure 6. The curve shows the accuracy of Wisdom of Crowds averages, starting with a single individual and finishing with all individuals. These start- and end-points correspond to the bounds established in Figure 2.[2] A clear and interesting pattern evident in this curve is how quickly including additional independent people in the Wisdom of Crowds average fails to improve accuracy. There is little improvement beyond the fifth or sixth person.

Figure 6 shows the performance of all of the three-person estimates—primed individuals, cooperative average, cooperative consensus, and competitive Price is Right inference—in relation to the Wisdom of Crowds curve. Motivated by a similar analysis presented by Vul and Pashler (2008), we map from the mean absolute difference of each three-person estimate to the Wisdom of Crowd curve, and then down to the number of people. This mapping allows the performance of the various approaches to be conceived in terms of how many independent estimates worth of performance they achieve. The results show that a primed individual is the same as a single non-primed individual, putting three people in a cooperative setting produces the accuracy of about one-and-a-half independent people, but putting three people in a competitive setting constitutes three independent people's worth of information.

Discussion

There are many analyses besides those reported here that could be pursued with the current data. For example, it would be interesting to compare the accuracy of individuals primed while working alone with those who gave the final estimate in the cooperative group setting. In a sense, these individuals have access (on average) to the same information, and so differences in their accuracy could be attributed to the social setting. We plan to pursue extensions and variants on the cognitive modeling of the competitive setting, including making different assumptions about how homogenous information is across participants, and how bidding decisions might be made.

[2] There were 22 individuals who provided individual estimates, so that 11 completed each of the two stimulus sets. The performance measures shown average over the two stimulus sets.

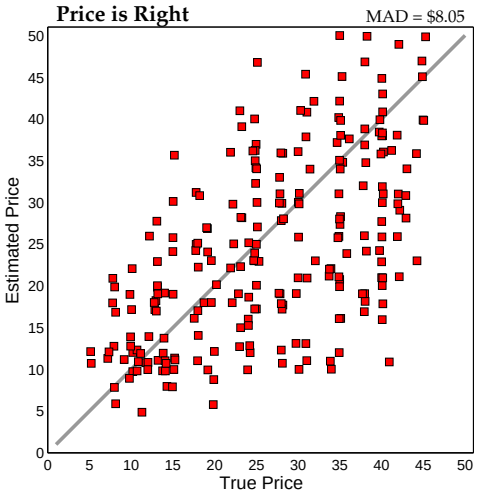


Figure 5: Relationship between true item prices and estimates inferred from an optimal decision-making analysis of three people competing in a "Price is Right" game. (MAD=Mean Absolute Deviation)

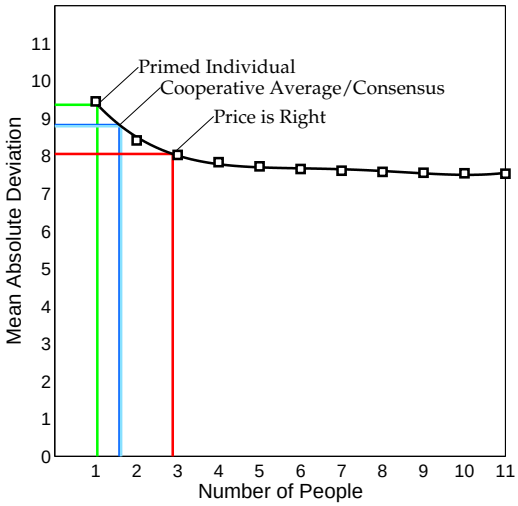


Figure 6: Characterization of the three-person estimates in terms of wisdom of crowds averages including $1, \ldots, 11$ individual estimates.

However, we can draw a number of interesting initial conclusions from the analyses reported here. The first is that the basic Wisdom of Crowds approach performs remarkably well. None of our alternative three-person estimation settings was superior to simply taking the average estimates of three random independent individuals. This averaging corresponds to a very simple generative model, in which each person estimates a signal with the addition of independent noise.

Our second conclusion applies to situations in which aggregate estimation must be done in a group setting, or when individuals share too much knowledge for independent estimates to be possible. These constraints could apply, for example, in situations where the goal is to pool the estimates of domain experts, who have overlapping training and knowledge. Here, our results argue for competitive rather than cooperative or passive approaches to extracting and combining information seem superior. The accuracy of the estimates from the simple "Price is Right" game were far superior to the other estimates we collected in group settings, and justified our effort to develop the much more complicated generative model for that setting.

We think the result highlighting the benefits of competition is suggestive, for both theoretical and applied reasons. Theoretically, it argues for the need to incorporate models of cognition and decision-making within Wisdom of Crowds research, to understand not just final behavior, but the underlying knowledge that generated that behavior. As we pointed out, it does not make sense to average the bids people make in the "Price is Right" game, but it does make sense to aggregate the knowledge they had that led them to decide on those bids. Practically, our results reinforce recent evidence for the effectiveness of competition instruments like prediction markets (e.g., Christiansen, 2007), rather than cooperative or collaborative groups settings, as better ways to combine knowledge across individuals.

References

Berk, J. B., Hughson, E., & Vandezande, K. (1996). The price is right, but are the bids? An investigation of rational decision theory. *The American Economic Review*, *86*(4), 954–970.

Christiansen, J. D. (2007). Prediction markets: Practical experiments in small markets and behaviors observed. *The Journal of Prediction Markets*, *1*, 17–41.

Hastie, R. (1986). Experimental evidence on group accuracy. In B. Grofman & G. Owen (Eds.), *Information pooling and group decision making* (pp. 129–157). Greenwich, CT: JAI Press.

Hastie, R., & Kameda, T. (2005). The robust beauty of majority rules in group decisions. *Psychological Review*, *112*(2), 494–508.

Kerr, N. L., & Tindale, R. S. (2004). Group performance and decision making. *Annual Review of Psychology*, *55*, 623–655.

Moscovici, S., & Zavolline, M. (1969). The group as a polarizer of attitudes. *Journal of Personality and Social Psychology*, *12*, 125–135.

Surowiecki, J. (2004). *The Wisdom of Crowds*. New York: Random House.

Vul, E., & Pashler, H. (2008). Measuring the crowd within: Probabilistic representations within individuals. *Psychological Science*, *19*(7), 645–647.

The Wisdom of Crowds with Informative Priors

Pernille Hemmer (phemmer@uci.edu)
Mark Steyvers (msteyver@uci.edu)
Brent Miller (brentm@uci.edu)

Department of Cognitive Sciences,
University of California, Irvine
Irvine, CA, 92697-5100

Abstract

In some eyewitness situations, a group of individuals might have witnessed the same sequence of events. We consider the problem of aggregating eyewitness testimony, trying to reconstruct the true sequence of events as best as possible. We introduce a Bayesian model which incorporates individual differences in memory ability, as well as informative prior knowledge about event sequences, as measured in a separate experiment. We show how adding prior knowledge leads to improved model reconstructions, especially in small groups of error-prone individuals. This Bayesian aggregation model also leads to a "wisdom of crowds" effect, where the model's reconstruction is as good as some of the best individuals in the group.

Keywords: Eyewitness Testimony; Wisdom of Crowds; Rank Ordering; Bayesian Modeling; Serial Recall.

Introduction

Studies of eyewitness testimony have shown that human memory can be incomplete and unreliable (e.g., Loftus, 1975). In real world situations, there might be multiple eyewitnesses, all of whom witnessed the same set of events. This raises the possibility of recovering the true account of events by analyzing the similarities in the recalled memories across individuals. Different individuals might also recall different aspects of the events, such that an aggregate narrative, based on the group's memory, would be closer to the true sequence of events than that of any one individual. An investigator might try to manually reconstruct the aggregate narrative, or witnesses might be allowed to discuss the events in order to develop the group narrative. Communication between witnesses however, has been shown to lead to much worse performance (Gagnon and Dixon, 2008), and humans have been shown to be inconsistent in assessing group information from multiple sources (Stasser & Titus, 1985). To avoid these problems, we propose a model of aggregation that can integrate the recalled memories from a number of independent individuals, while also taking in other important factors, such as individual differences and prior knowledge, into account.

Research on the "Wisdom of Crowds" (WoC) has shown that an aggregation of independent judgments often leads to a group estimate that is closer to the ground truth than that of most of the individuals (Surowiecki, 2004). These group estimates are often simply found by taking the mean, median, or mode of responses (Galton, 1907; Surowiecki, 2004). Much of the previous literature on aggregation of judgments has focused on tasks where individuals estimate numerical quantities and probabilities (Budescu, Yu, 2007; Hogarth, 1978; Wallsten, Budescu, Erev, & Diederich, 1997). It is, however, often that case that eyewitness have to retrieve information more complex than single numerical estimates.

The WoC effect can also be demonstrated with more complex problem sets. For example, the WoC effect has been demonstrated with solutions to problem-solving situations such as finding minimum spanning trees for a set of nodes (Yi, Steyvers, Lee & Dry, in press). Steyvers, Lee, Miller, and Hemmer (2009) showed that order information from semantic memory can also be combined across individuals to give high accuracy in reconstructing the true order of items along some physical or temporal dimension; when individuals recalled the order of US presidents, or the order of rivers according to length, many of the individual orderings were error-prone, but the aggregate orderings were more accurate, on average. In Steyvers et al. (2009), a number of aggregation models for order information were tested. It was found that using Bayesian models that incorporated psychologically plausible representations, cognitive processes and individual differences outperformed basic heuristic aggregation approaches, such as taking the mode.

When errors across individuals are uncorrelated (as they tend to be when individuals independently give their judgments) the errors will cancel out in the aggregate. Therefore, one expects the best results in WoC experiments with a large number of individuals. In eyewitness situations however, there is rarely a "crowd" available to witness the same set of events. In these cases, we have to rely on a small number of individuals (in many cases, just one) and significant errors might not cancel. Therefore, it might not be sufficient to just analyze the commonalities across the witness reports. We propose that it is better to combine the witness reports along with prior knowledge about the particular event sequence. Combining prior knowledge with noisy information has been shown in other domains to improve the recovered estimate (Hemmer & Steyvers, 2008; Konkle & Oliva, 2007; Kan, Alexander, Verfaelle, 2009).

We focus in this research on the problem of reconstructing event sequences. The goal is to reconstruct

the true ordering of a set of events by aggregating the recalled orderings from a small number of individuals, all of whom witnessed the same event sequence. The novelty of the current approach is that we incorporate informative prior knowledge in an aggregation model for order information in order to improve the aggregate estimate. This is especially helpful when aggregating across a small number of error-prone individuals.

We present our results as follows. We first report on behavioral experiments wherein we tested people's ability to reconstruct, from episodic memory, the order of stereotyped events (e.g., getting up in the morning), or random events (e.g., clay animation without a clear story line). We also report on experiments where we measured prior knowledge for the same set of events. We then describe a Bayesian approach that aggregates the orderings across individuals while taking prior knowledge into account.

Empirical Study on Serial Recall

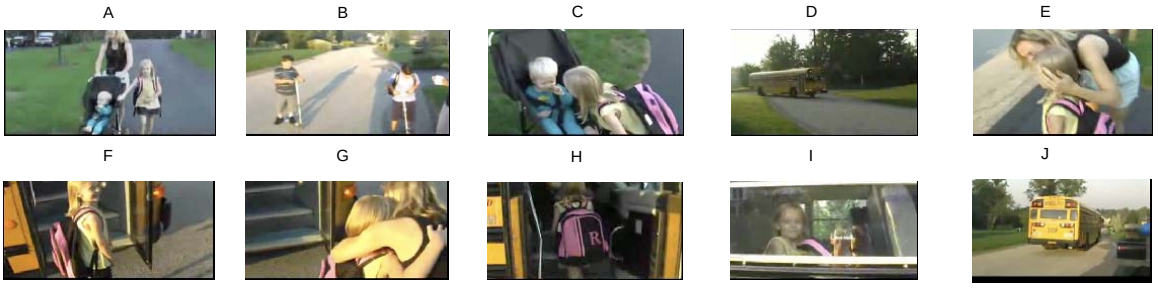
Much research on serial recall has been done on random word and letter sequences that do not have any obvious organization. In such experiments, individuals are shown a sequence of words or letters, and the task is to recall the original temporal order as best as possible during a later test. Typical errors in the recalled orderings are transposition errors where the orderings are locally perturbed (Estes, 1997; Nairne, 1992) -- two events nearby in time tend be reconstructed as occurring nearby but the amount of perturbation noise depends on many factors such as time elapsed between study and test, stimulus characteristics and individual differences. Similar patterns have been observed in more naturalistic experiments, such as naming the day of the week an event occurred (Huttenlocher, Hedges, & Prohaska, 1990), as well as for autobiographical memory, such as ordering the events of September 11th (Altmann, 2003). With more naturalistic event sequences, prior knowledge about the event sequences can influence episodic memory. People have clear expectations for routine activities and are sensitive to the ordering of actions within an activity (Bower, Black & Turner, 1979).

We conducted a series of behavioral experiments using two types of event sequences. We used a number of *stereotyped* event sequences, such as getting up in the morning, or jumping on a bus, for which people have clearly defined expectations, and a number of *random* event sequence, such as clay animation sequences or Japanese pizza commercials, for which the temporal organization might be less structured. To assess the prior knowledge people have about these types of events, we first conducted a prior knowledge study where we asked participants to order the events in the most natural order possible without actually showing them the original, true event sequence. This allows us to estimate a model for the prior probability of each sequence.

In a separate experiment, we assessed serial recall for each of event sequences. It should be noted that our definition of serial recall differs from the standard use of the term in that our task only involves ordering the events, not recalling the items to be ordered, as in a standard serial recall task. In our task, we first showed a video of the original event sequence which was followed by a serial recall test in which individuals ordered image stills from the video as best as possible according to the original temporal sequence in which the events appeared. No communication between individuals was allowed in any of our tasks, and therefore the data consists of independent recollections from individuals.

Methods

Participants were undergraduate students at the University of California, Irvine. There were 16 participants in the prior knowledge experiment and 28 participants in the serial



prior

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
4	3	13	1	9	5	3	12	11	10	10	2	5	1	1	10
A	B	A	A	A	B	B	A	B	B	A	A	D	B	A	B
B	C	C	B	C	D	A	D	A	A	C	B	B	A	B	A
E	A	E	C	D	A	D	G	C	E	E	D	A	C	D	E
D	D	F	D	E	E	C	C	E	C	D	C	C	D	C	F
C	E	G	E	G	C	E	E	D	G	G	E	E	E	E	D
G	G	H	G	F	G	G	F	J	F	F	G	G	F	F	G
F	F	I	F	H	F	F	H	I	H	H	F	F	G	G	H
H	H	D	H	I	H	H	I	H	J	I	H	H	H	H	I
I	I	J	I	J	I	I	J	F	I	J	I	I	I	I	J
J	J	B	J	B	J	J	B	G	D	B	J	J	J	J	C

serial recall

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
2	1	2	4	2	5	1	1	1	1	2	1	6	1	1	0	1	2	2	1	2	1	1	1	2	0	1	1
B	A	A	A	B	B	A	A	A	A	A	A	A	A	A	A	A	B	B	A	A	A	B	A	B	A	A	A
A	B	B	B	A	A	B	B	B	B	B	B	G	B	B	B	B	A	A	B	B	A	C	A	B	B	B	B
C	C	C	G	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	B	C	C	C	C	C
D	D	D	C	D	D	D	D	D	D	D	D	B	D	D	D	D	E	D	D	D	D	D	D	D	E	D	
E	E	G	D	E	H	E	E	F	F	F	E	D	F	E	E	E	D	E	E	F	E	E	E	E	D	E	
G	G	E	E	G	F	G	G	E	E	G	G	E	E	G	F	G	F	G	G	G	G	F	F	G	F	F	G
F	F	F	F	F	E	F	F	G	G	E	F	F	G	F	G	F	G	F	F	E	F	G	G	F	G	G	F
H	H	H	H	G	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H
I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I
J	J	J	J	J	J	J	J	J	J	J	J	J	J	J	J	J	J	J	J	J	J	J	J	J	J	J	J

Figure 1. The sequence A-J shows the 10 images from the 'bus' video sequence in the correct temporal order. The two tables show the participant orderings in the prior knowledge and serial recall experiment. The first row is the participant id. The second row is the Kendall's tau distance between the true ordering and the recalled order for that participant.

recall experiment.

Materials. We sampled 6 videos from YouTube.com. Three videos depicted stereotyped events sequences (getting up in the morning, a wedding, getting on the school bus). Three videos depicted more random event sequences (a Japanese yogurt commercial, a Japanese pizza commercial, and a clay animation sequence). For each of the 6 videos 10 still images of individual scenes were drawn. See Figure 1 for an example.

Prior Knowledge Experiment. Participants were shown 10 image stills from a given event sequence (e.g., Wedding) and asked to order the 10 images based on their prior expectation of how the event in the slides might unfold. Importantly, in this experiment, participants were never shown the original video sequence from which the image stills were drawn. They responded using an interactive interface in which the images were randomly ordered on the screen and the instruction was to order the images in any way to make the sequence as natural as possible.

Serial Recall Experiment. Participants first viewed the original video sequence. Participants were then presented with the same interface as in the prior knowledge experiment. They were shown 10 image stills that they had to order in the original temporal order. For both the prior knowledge and memory experiment, the initial ordering of the 10 image stills, as well as the order of the 6 video sequences, was randomized across participants.

Results and Discussion

To evaluate the performance of participants, we measured the distance between the reconstructed and the correct ordering. A commonly used distance metric for orderings is Kendall's τ (Marden, 1995). This distance metric is the minimum number of adjacent pairwise swaps necessary to resolve any disagreements between the two orderings being compared. Values of τ range from $0 \leq \tau \leq (N-1)/2$, where N is the number of items in the order: N=10 for all of our event sequences. In our experiment, a τ=0 indicates that the participant responded with the exact correct ordering. A τ=1 indicates that one adjacent pair of items was swapped. When participants are using a random guessing strategy, their expected mean expected distance is $\tau = (N-1)/4 = 22.5$.

Figure 1 shows the raw data collected for the "bus" video sequence – a stereotyped event sequence. In the prior knowledge experiment, participants produced orderings that were much better than chance, suggesting that a priori, it is possible to guess the true ordering of events in these types of event sequences. In the memory experiment, 2 participants produced the correct ordering, and 15 more were within one swap of the true order. Note that very few identical orderings are produced between participants. We found that for all 3 random events, in both the prior knowledge experiment and the memory experiment, each participant produced a unique ordering. For the 3 stereotyped event sequences however, only one sequence led to unique orderings across all participants.

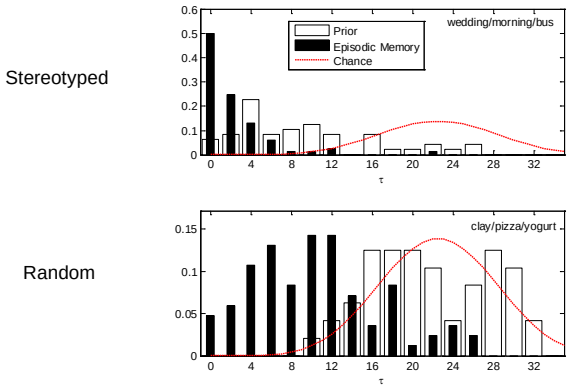


Figure 2. Distributions of Kendall τ distances.

Figure 2 shows the distributions of the Kendall τ distances for the serial recall and prior knowledge experiment. The top panel shows the distances for stereotyped event sequences and the bottom panel shows the distances for random event sequences. The dashed line shows the distribution of distances that can be expected from a random guessing strategy (this distribution can be calculated exactly, see Marden, 1995). For both the stereotyped and random event sequences, the distances are lower for the memory task than for the prior knowledge task. The distances are also lower for the stereotyped event sequences than for the random event sequences. Even when participants did not study the videos (the prior knowledge condition), they performed better than chance in the stereotyped condition, as compared to the random condition where prior knowledge performance led to a distribution of distances very similar to distances expected from chance performance. These results demonstrate that general knowledge about events can greatly contribute to the accuracy of recalling these events.

Modeling

We can conclude from our empirical study that prior knowledge can lead to improved average performance in recall. When ordering scenes from an event with strong prior expectations, the resulting orderings are relatively close to the true ordering. Of course, performance improves on average after observing the true event sequence and later recalling the sequence from memory. This raises the question of how one might incorporate an informative prior in a model for aggregating rank-ordered recall. Such priors might guard against errors from a small number of poorly performing individuals. In this paper, we explore very simple models to aggregate the orderings of individuals. The goal of the modeling is not to build a comprehensive model of recall that specifies all the representations and processes involved in storing and retrieving information from memory. Instead, we will focus on simple probabilistic models such as a Mallows model (e.g. Steyvers et al., 2009) that allow us to aggregate the retrieved orderings from a number of individuals using Bayesian inference. The current model incorporates two important differences to the

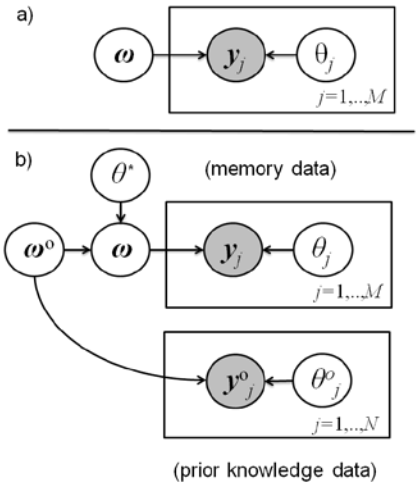


Figure 3. The graphical model representations for the Mallows model with an uninformative prior (a) and an informative prior about the group knowledge (b).

previous work by Steyvers et al. (2009). First, we generalize the model to allow for individual differences in memory performance. These individual differences are estimated by the model in a purely unsupervised fashion and do not require knowledge of past performance in other tasks or access to a known ground truth. With the individual differences, the model finds aggregates that are weighted towards solutions provided by the individuals that are estimated to have good memory performance.

Second, we develop a simple extension of Mallows models that allows for informative priors. This prior is estimated from the orderings produced in the prior knowledge experiment.

Mallows Model with an Uninformative Prior

In a basic Mallows model (Marden, 1995), all individuals are assumed to derive their orderings from a single underlying ordering, that we will refer to as the *group knowledge*. The group knowledge is a latent variable in the model that can be estimated from the data. Importantly, Mallows model assumes that each individual produces orderings centered on the group ordering with distant orderings less likely than orderings close to the group ordering. Although Mallows-type models have often been used to analyze preference rankings (Marden, 1995), they have not been applied, as far as we are aware, to ordering data from serial recall experiments. In our first extension of the standard model we allow for individual differences in memory performance. We evaluated this aggregation model by comparing the estimated group ordering to the ground truth. If the model is able to tap into the collective wisdom of a group of individuals, the estimated group ordering should be close to the true ordering.

Specifically, let $\boldsymbol{y}_j$ represent the ordering from individual *j*, and $\boldsymbol{\omega}$ the latent group ordering. In a Mallows model, the

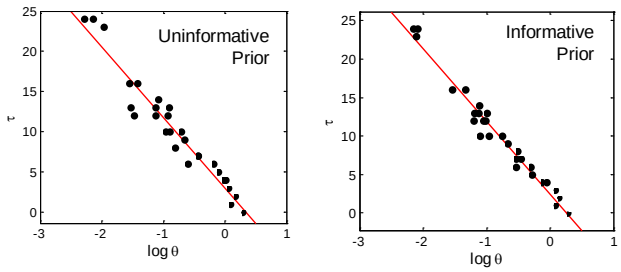


Figure 4. Calibration results for the two models for one event sequence.

probability of each individual ordering given the group ordering is given by

$$p(\boldsymbol{y}_j \,|\, \boldsymbol{\omega}, \theta_j) \propto e^{-\mathrm{d}(\boldsymbol{y}_j, \boldsymbol{\omega})\theta_j} \quad (1)$$

where for simplicity we have omitted the normalization constant. The function d returns the Kendall τ distance between two orderings. The scaling parameter θ_j determines how close the observed order for individual *j* is to the group ordering. It can be interpreted as an individual (inverse) noise parameter -- good individuals tend to closer to the group consensus (high θ) whereas poor performing individuals return more idiosyncratic orderings further away from the group knowledge (low θ). We will assume a Gamma prior on the individual noise levels: $\theta_j \sim \mathrm{Gamma}(\theta_0\lambda, 1/\lambda)$, where λ is a hyperparameter that sets the overall level of cohesion expected from the group. Notably, in this first model, we have assumed a uniform prior over group orderings, $\boldsymbol{\omega} \sim \mathrm{Uniform}(\Omega)$, where Ω is the set of all orderings. Therefore, a priori, the model assumes no preference for a particular group ordering.

Figure 3, panel a, shows a graphical representation of the model. Shaded nodes represent observed variables while nodes without shading represent latent variables. The arrows indicate the conditional dependencies between the variables and the plate represents the repeated sampling steps across *M* subjects in the memory experiment.

Mallows Model with an Informative Prior

We now introduce a simple variant of this model that allows for an informative prior. The idea is that the group knowledge is itself sampled from a Mallows model:

$$p(\boldsymbol{\omega} | \boldsymbol{\omega^0}, \theta^*) \propto e^{-\mathrm{d}(\boldsymbol{\omega}, \boldsymbol{\omega^0})\theta^*} \quad (2)$$

where $\boldsymbol{\omega^0}$ is the prior ordering from which the group ordering is derived, and θ^* is a scaling parameter. This prior stage in Mallows model at first might not seem to gain any additional information because it is not clear how the prior ordering can be constrained. However, we have data in the prior knowledge experiment in which *N* participants tell us what orderings they expect from certain scenes. Let $\boldsymbol{y}_j^0$ represent the prior ordering given by individual *j* in the prior knowledge experiment. We assume that these are produced by a Mallows model with $\boldsymbol{\omega^0}$ as the "center":

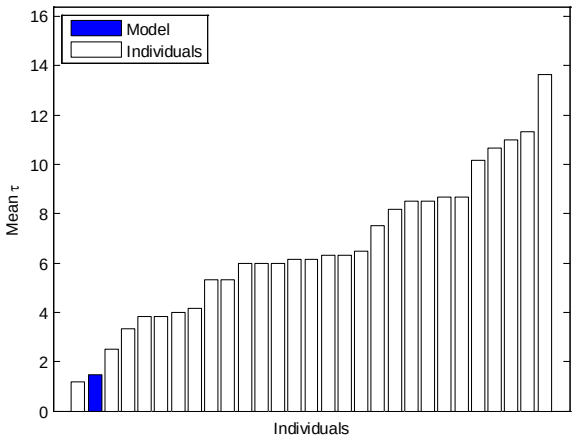


Figure 5. Performance of individuals and model (with informative prior) averaged over six event sequences.

$$p\left(\boldsymbol{y}_j^0 \middle| \boldsymbol{\omega}^{\mathbf{0}}, \theta_j^0\right) \propto e^{-\mathrm{d}\left(\boldsymbol{y}_j^0, \boldsymbol{\omega}^{\mathbf{0}}\right)\theta_j^0} \quad (3)$$

Figure 3, panel b, shows the corresponding graphical model. With this model, we are setting a prior on the group ordering -- when there is only data available from a few individual in the memory experiment, the group ordering will be influenced by the data from the prior knowledge experiment leading to group orderings that are a priori deemed likely. When data from more individual becomes available in the memory experiment, the prior knowledge data will have a diminishing influence on the group ordering which will be mostly determined by the memory data.

Modeling Results

All latent variables in the model were estimated using a MCMC procedure, separately for each event sequence. The result of the inference procedure is a probability distribution over group orderings, of which we take the mode as the single answer for a particular problem. Note that the inferred group ordering does not have to correspond to an ordering of any particular individual. The model just finds the ordering that is close to all of the observed memory orderings.

Figure 4 shows the calibration for the two models on a single event sequence (the clay animation video). Each panel shows the relationship between the inferred θ (related to the distance of each individual to the group ordering) and the Kendall's τ distance of the individual's answer to the ground truth. The plots show that individuals who are close to the group ordering tend to be closer to the ground truth. This means that the models can calibrate the performance levels of individuals, even in the absence of any explicit feedback or access to the ground truth.

Figure 5 shows the Kendall's τ distance for each individual in the memory experiment averaged over the six event sequences. Note that there are substantial individual differences with some individuals coming relatively close to the ground truth. The figure also shows the average model performance. Comparison between individual and model performance reveals a WoC effect: The model performs as well as some of the best individuals, with only one individual outperforming the model. Therefore, we can conclude there is a weak WoC effect (a strong WoC effect would correspond to a situation where the model outperforms all individuals in the group).

We now focus on applying the model to subsets of participants to mimic eyewitness situations that typically involve only small number of individuals. In the first analysis, we select a random set of *K* individuals from the original set of 28 individuals. We then apply the two models to the subset of individuals. Figure 6 shows model results for the model with the informative and uninformative prior separated for stereotyped and random event sequences. For random event sequences, where the prior is weak, there is no improvement in the aggregation between the two models (if anything, there is a small performance decrement for the model with the informative prior). For stereotyped event sequences however, people have strong prior expectations about the true ordering of events and there is a marked improvement in the aggregate response in the model with the informative prior. This improvement is most pronounced with low sample sizes (*K*=1 and *K*=2) when the prior can still exert an influence on the inferred group orderings. Note that when *K*=1, the model with the uninformative prior has no information other than the ordering given by a single individual – therefore, the aggregate solution given by the model is equivalent to the ordering provided by the individual. This results in an average tau of around 15. However, performance for the model with the informative prior is much better resulting in a tau of around 8, because the aggregate solution combines the single remembered ordering with the a priori likely orderings.

To better highlight the benefit of the prior information, we also conducted a model analysis where we selected the *worst* performing individuals in the sample. In this sampling procedure, we sample the *K* worst individuals where we vary *K* from 1 (the single worst performing individual) to 28 (all individuals combined). Figure 7 shows model results for both models separated for stereotyped and random event sequences. The relative performance benefits can be seen most clearly for the stereotyped event sequences for low sample sizes (*K*=1 and *K*=2). In these cases, the worst individuals recall event sequences that are a priori unlikely and the prior "corrects for" the noise in the available data.

Therefore, these analyses suggest that an aggregation model with informative priors can be used to guard against the most egregious errors committed by the worst individuals in the memory task.

Conclusions

We have presented two approaches for aggregating recalled sequences of events in order to reconstruct the true event sequence as best as possible. Individuals are likely to differ in their ability to recall event sequences and pay attention to different parts on an event sequences. Therefore, by

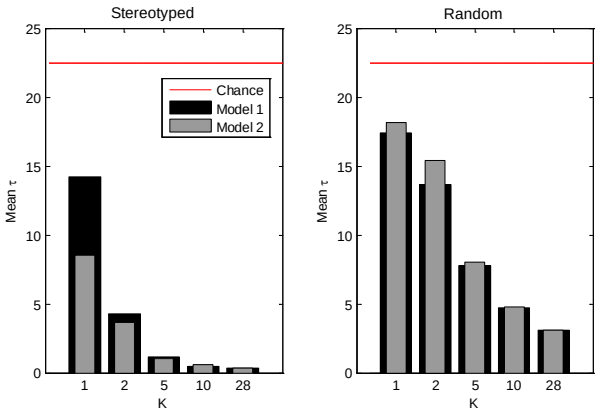


Figure 6. Results from the models with an uninformative prior (model 1) and informative prior (model 2) for *random* subsets of *K* individuals from the memory task.

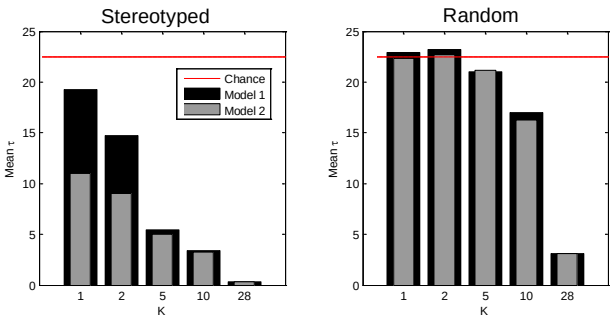


Figure 7. Results from the models with an uninformative prior (model 1) and informative prior (model 2) for subsets of the *worst K* individuals from the memory task.

analyzing the consistencies in orderings across individuals, we can extract the collective wisdom in the group. We presented two aggregation approaches based on Mallows model that allow for individual differences. The models combine information at the group level with information at the individual level to explain orderings given by an individual. In the first approach, the model uses only the data from the individuals who all witnessed an event sequence. In the second approach, the model uses an additional source of data based on the prior knowledge about the events extracted from another group of individuals.

We demonstrated a weak WoC effect, where the average performance of the model was better than every individual, save one. We have also shown that a Mallows model with informative priors has a markedly improved ability to reconstruct the ground truth in cases where the event sequences are highly stereotyped and a small sample of poorly performing individuals is used for aggregation. This is particularly important in eyewitness situations where we typically have only a small number of individuals available.

References

Altmann, E. M. (2003) Reconstructing the serial order of events: A case study of September 11, 2001. *Applied Cognitive Psychology*, **17**, 1067-1080.

Bower, G. H., Black, J. B., & Turner, T. J. (1979). Scripts in memory. *Cognitive Psychology*, **11**, 177-220.

Budescu, D. V. & Yu, H. (2007). Aggregation of opinions based on correlated cues and advisors. Journal of Behavioral Decision Making, 20, 153-177.

Estes, W.K. (1997). Processes of Memory Loss, Recovery, and Distortion. Psychological Review, 104, 148-169.

Gagon, L.M. & Dixon, R.A. (2008). Remembering and retelling stories in individual and collaborative contexts. *Applied Cognitive Psychology*, *22*, 1275-1297.

Galton, F. (1907). Vox Populi. *Nature*, 75, 450-451.

Hemmer, P. & Steyvers, M. (2008). A Bayesian Account of Reconstructive Memory. In V. Sloutsky, B. Love, and K. McRae (Eds.) *Proceedings of the 30th Annual Conference of the Cognitive Science Society*.

Hogarth, R. M. (1978). A note on aggregating opinions. Organizational Behavior and Human Performance, 21(1), 40-46.

Huttenlocher, J., Hedges, L. V., & Prohaska, V. (1992). Memory for day of the week: A 5+2 day cycle. *Journal of Experimental Psychology: General*, *121*, 313–325.

Kan, I.P., Alexander, M.P. & Verfaellie, M. (2009). Contribution of prior semantic knowledge to new episodic learning in amnesia. *Journal of Cognitive Neuroscience*, *21*, 938-944.

Konkle, T., & Oliva, A. (2007). Normative representation of objects: Evidence for an ecological bias in perception and memory. In D. S. McNamara & J. G. Trafton (Eds.), *Proc.s of the 29th Annual Cognitive Science Society, (pp. 407-413), Austin, TX: Cognitive Science Society.*

Loftus, E.F. (1975). Leading questions and the eyewitness report. *Cognitive Psychology*, *7*, 560-572.

Marden, J. I. (1995). Analyzing and Modeling Rank Data. New York, NY: Chapman & Hall USA.

Nairne, J. S. (1992). The loss of positional certainty in long-term memory. *Psychological Science*, **3**, 199–202.

Stasser, G., Titus, W. (1985). Pooling of unshared information in group decision making: Biased information sampling during discussion. *Journal of personality and social psychology*, 48(6), 1467-1478.

Steyvers, M., Lee, M.D., Miller, B., & Hemmer, P. (2009). The Wisdom of Crowds in the Recollection of Order Information. In J. Lafferty, C. Williams (Eds.) *Advances in Neural Information Processing Systems, 23*. MIT Press.

Surowiecki, J. (2004). The Wisdom of Crowds. New York, NY: W. W. Norton & Company, Inc.

Wallsten, T.S., Budescu, D.V., Erev, I. & Diederich, A. (1997). Evaluating and combining subjective probability estimates. Journal of Behavioral Decision Making, 10, 243-268.

Yi, S. K. M., Steyvers, M., Lee, M. D., Dry, M. J. (in press) Wisdom of the Crowds in Minimum Spanning Tree Problems. Proceedings of the 32nd Annual Conference of the Cognitive Science Society.

The Emergence of Adaptive Eye Movements in Reading

Yanping Liu (yanping@pitt.edu)
Department of Psychology, 3939 O'Hara St.
Pittsburgh, PA 15260 USA

Erik D. Reichle (reichle@pitt.edu)
Department of Psychology, 3939 O'Hara St.
Pittsburgh, PA 15260 USA

Abstract

Simulations were completed using artificial reading "agents" that are subject to known physiological (e.g., limited visual acuity) and psychological (e.g., limited attention) constraints and capable of learning to move their eyes and allocate attention to read as efficiently as possible. These simulations indicate that agents learn when and where to move their eyes to attain maximal reading efficiency, generalize this behavior from training sentences to novel test sentences, and use word length to predict word-identification times and thereby make optimal decisions about when to initiate saccadic programming—even if word length is only moderately predictive of word-identification times. These results suggest that humans may exploit even modestly informative cues in learning to decide when to move their eyes during reading.

Keywords: Attention; Eye Movements; Genetic Algorithms; Neural Networks; Reading; Reinforcement Learning

Introduction

One of the outstanding unanswered questions in the psychology of reading (Rayner & Pollatsek, 1989) is: To what extent are the moment-to-moment decisions about *when* to move the eyes during reading determined by cognition? Attempts to answer this question can be divided into three theoretical "camps" (Reichle, 2006; Reichle, Rayner, & Pollatsek, 2003).

The first maintains that when the eyes move is largely determined by the constraints imposed by the visual and oculomotor systems (e.g., limited visual acuity). Advocates of this *oculomotor-control account* (Feng, 2006; McDonald, Carpenter, & Shillcock, 2005; Reilly & O'Regan, 1998; Suppes, 1990; Yang, 2006) argue against an eye-mind link in reading, and maintain that individual fixation durations provide only minimal information about ongoing lexical and/or linguistic processing difficulty.

According to the second "camp," most decisions about when to move the eyes are determined by the activity of an autonomous random timer that causes the eyes to move at a rate that reflects a reader's comprehension goals and overall text difficulty, with cognition only intervening to inhibit saccadic programming when processing difficulty is encountered and thereby lengthening fixation durations. Advocates of this *autonomous-timer account* (Engbert, Longtin, & Kliegl, 2002; Engbert et al., 2005; Reilly & Radach, 2006) argue for a weak eye-mind link, with individual fixations occasionally reflecting ongoing lexical or linguistic processing difficulty.

Finally, the third "camp" maintains that the eyes and mind are tightly coupled, with the completion of some cognitive process (e.g., lexical access) being the "trigger" that normally causes the eyes to move from word to word during reading. Advocates of this *cognitive-control account* (Just & Carpenter, 1980; Reichle et al., 1998; Reichle, Warren, & McConnell, 2009; Reilly, 1993; Salvucci, 2001) argue for a strong eye-mind link, with individual fixation durations usually reflecting local processing difficulty.

Perhaps not too surprisingly, all three theoretical positions have been remarkably successful explaining the basic patterns of eye movements that are observed during reading; each position has provided one or more computational models that formally instantiates the core assumption of their respective positions and that simulate many or all of the "benchmark" findings related to eye-movement behavior in reading (Reichle et al., 2003). This makes it difficult to evaluate the models purely on the basis of their ability to account for data, and because the models make different a priori assumptions about the factors that guide readers' eye movements (e.g., how attention is allocated), model evaluation is like the proverbial problem of "comparing apples and oranges." The present simulations therefore adopt an entirely different approach to understanding eye-movement control in reading.

Rather than developing a computational model around a priori assumptions about the precise manner in which perception, cognition, and motor control guide eye movements in reading, the present approach is a direct extension of the work reported by Reichle and Laurent (2006). In this work, artificial reading "agents" that were subject to known physiological (e.g., limited visual acuity) and psychological (e.g., limited capacity attention) constraints were given the task of learning how to move their eyes and attention so as to "read" (i.e., identify sequences of "words") as efficiently as possible. The key results of this work were that the agents learned: (1) to direct their eyes towards the centers of words, the viewing location that afforded the most rapid identification of the words; (2) to use word length to predict when a given word would be identified, and then initiate saccadic programming to move its eyes from that word right as it was identified; and (3) to incur local fixation duration costs by identifying short, easy-to-identify words from peripheral vision, and thereby avoiding more costly saccades to those words.

The present simulations replicate and extend the Reichle and Laurent (2006) results using artificial agents that are capable of learning to move their eyes and attention via reinforcement learning (Sutton & Barto, 1998). However, in contrast to the Reichle and Laurent agents, the present agents were implemented using *artificial neural networks* (*ANN*s), and we demonstrated in two simulations that the behavior of these agents: (1) generalizes to novel sentences and words; (2) can be learned even in less than optimal learning conditions; and (3) is generally congruent with assumptions of cognitive-control theories. The theoretical implications of these results will be discussed after the simulation results are described.

General Simulation Method

The artificial reading "agents" that were used in the present simulations were given the task of learning how to "read" (i.e., identify sequences of words in their canonical order) as efficiently as possible. These words could vary in terms of their length (1-8 letters) and/or the time required for their identification (2-14 time steps). The agents learned to perform this task (subject to various constraints, discussed below) using *trajectory sampling*, a variant of the *value iteration* reinforcement-learning algorithm (Sutton & Barto, 1998) that is often used with large-scale problems. This algorithm is specified by:

```
i = 0
for all initial S:
   V_i(S) = ANN(S)
   repeat
      i = i + 1
      if (random value < greed) then:
         V_i(S) = V_{i-1}(S) + ε {max_{action∈M}[reward(S, action)
                + γ V_{i-1}(S′)] – V_{i-1}(S)}
      else random action
until learning has completed.
```

where i indexes the learning iteration, $V_i(S)$ is the value associated with state S at time i, and M is the set of permissible actions from a given state. There are three parameters: ε (= 0.5) controls the learning rate, *greed* (= 0.5) controls how often an agent exploits what it already knows in selecting actions versus exploring the consequences of randomly selected actions, and γ (= 0.9) determines how much the agent weighs the reward that is anticipated from the next state, S', versus the immediate reward that it receives from the action that it selects. Each state, S, consists of information that is available to the agent at any given point in time (see Table 1). The agents can perform one of three actions: (1) continue attending (i.e., lexically processing) the current word; (2) shift attention to the next word; and (3) request an eye movement of ±10 character spaces. An agent selects the actions that result in the most (anticipated) reward, being "rewarded" +1 for every identified word and "punished" -1 for every time step spent processing a sentence. Learning continues until the values of the states reach asymptote.

Table 1. State information (S) used by agents.

#	Available Information
1	Attended word (i.e., $word_n$) identified? (Y/N)
2	# time steps processing $word_n$
3	# spaces between center of $word_n$ and fixation
4	Length of $word_n$
5	Length of $word_{n+1}$
6	Length of $word_{n-1}$
7	Saccade being programmed? (Y/N)
8	Length (# spaces) of intended saccade
9	# time steps programming saccade

As mentioned, the agents are subject to several constraints. First, visual acuity is limited, so that the rate of lexical processing decreases as the spatial distance between the agent's center of vision and the center of the word being processed increases (i.e., a word that takes N time steps to identify when its middle letter is fixated will take 1 additional time step to identify for each character space of disparity between the letter being fixated and the center of the word). Saccades also require 3 time steps to program and 1 time step to execute, and are subject to Gaussian (μ = 0; σ = 1) random error. Finally, because the perceptual span is known to be of limited spatial extent (Rayner & Pollatsek, 1989; Rayner, 1998), the agents were only allowed to process one word at a time, instantiating the assumption that attention is allocated serially during reading (e.g., Reichle et al., 1998) or approximating the assumption that attention is allocated as a gradient—albeit a tightly focused one (e.g., Engbert et al., 2005). Although this assumption about attention is quite controversial (e.g., see Reichle et al., 2009), it was intended as a simplifying assumption to make the simulations as tractable as possible.

In the Reichle and Lauent (2006) simulations, the value of each state, $V_i(S)$, was stored in a look-up table (i.e., one value per combination of dimensions in Table 1). In the present simulations, the values were learned and stored in the connection weights of an ANN whose architecture and principle weights were selected using *NeuroEvolution of Augmenting Topologies* (*NEAT*) (Stanley & Miikkulainene, 2002) and whose weights were optimized via the *Covariance Matrix Adaptation Evolution Strategy* (*CMA-ES*; Hansen, Müller, & Koumoustsakos, 2003) when trapped in local maxima. Figure 1 is a schematic diagram of how the NeuroEvolution and CMA-ES algorithms are used in conjunction with task-specific training to select network architectures that are well suited to solve the types of problems explored in this article.

Each network was comprised of nine input units (one per dimension in Table 1), one bias unit, one output unit (representing the learned value of each state), and an unspecified number of hidden units. In contrast to many neural networks, the hidden units were not strictly layered,

but could be configured in a variety of ways (e.g., as additional bias units; see Fig. 1). The activation of input unit *i* when given some value *x* of one of the dimensions in Table 1 was scaled to the interval [-1, 1] using:

$act_i(x) = \{x - [max(x) / 2]\} / [max(x) / 2]$

where the function "max" returns the maximum value of the dimension. (Note that $act_i(x) = -1/1$ when Dimensions 1 and 7 equal false/true.)

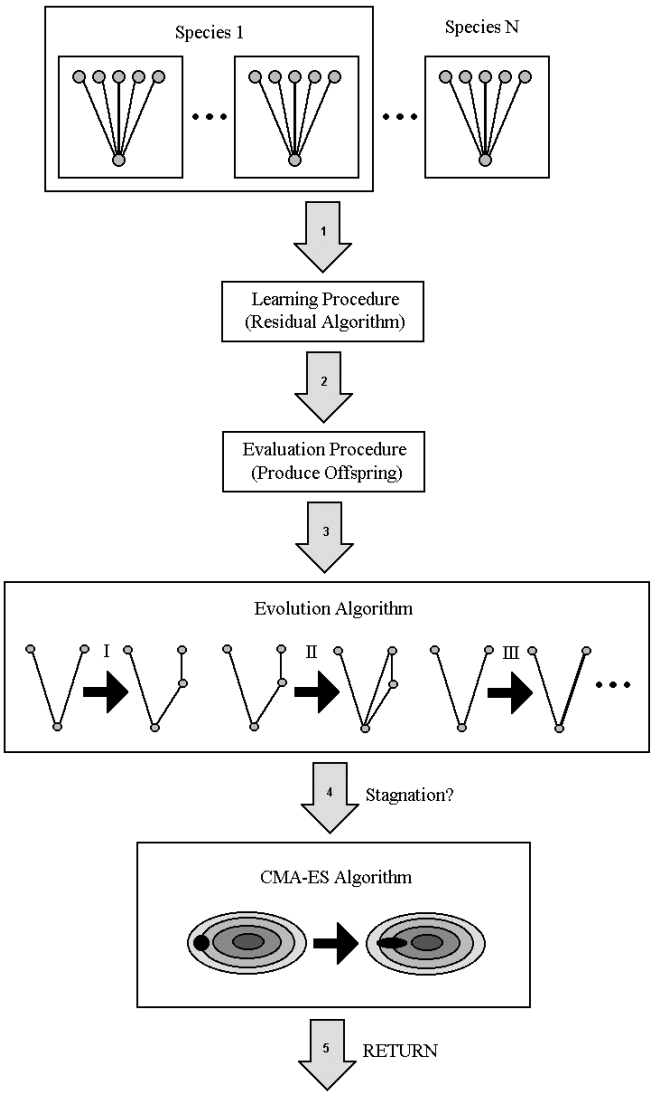


Figure 1. Method use to evolve and train agents.

This NeuroEvolution algorithm was used to select network architectures that were adapted to use the value iteration reinforcement-learning algorithm (via using a residual algorithm to implement back-propagation in the ANNs) for the tasks that are the focus of this article—learning to move the eyes and attention to read efficiently. The CMA-ES algorithm was used to optimize weights when optimization stagnated. Each reported simulation is based on populations of 100 individuals evaluated across 300 generations to solve the tasks of interest. Each individual networks agent also learned to perform its task using value iteration across five learning trials and using the specific training regimens.

Simulation 1

The first simulation replicated and extended the Reichle and Laurent (2006) results using agents implemented as ANNs (as described above) and various novel test sentences.

Method. Five agents were trained on a corpus of five 8-word "sentences" comprised of random permutations of 1-, 3-, 5-, and 7-letter "words." (These sentences were randomly selected from 20 used by Reichle & Laurent, 2006.) The first and last words were always 1-letter in length and required 2 time steps to identify, and always excluded from our analyses because their processing started/ended abruptly. The remaining 1-, 3-, 5-, and 7-letter words respectively required 2, 6, 10, and 14 time steps to identify when fixated from their central letters. After training, agents were tested on: (1) the same five sentences; (2) five novel 8-word sentences comprised of different random permutations of 1-, 3-, 5-, and 7-letter words; and (3) five 8-word sentences comprised of random permutations of 2-, 4-, 6-, and 8-letter novel words.

Results. Figure 2 shows the simulated fixation landing-site distributions on the words, as a function of their length and whether the sentences being using used to evaluate the agents were old (i.e., used during training), novel, or comprised of novel words (i.e., 2-, 4-, 6-, and 8-letter words). (In all of the figures shown below, the data points indicate the condition means and the error bars indicate the standard errors of the means.) As indicated, the agents learned to direct their eyes towards the centers of the words because this was the viewing position that afforded the most rapid identification of the words. However, because of saccadic error, the fixation landing sites are approximately normally distributed, in line with what is observed with human readers (McConkie et al., 1988, 1991; Rayner, Sereno, & Raney, 1996). Finally, the behavior generalized across both novel sentences and words.

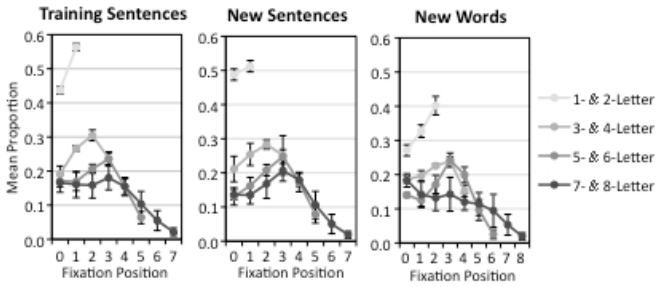


Figure 2. Simulated fixation landing-site distributions.

Figure 3 shows the mean probabilities of making a single fixation, making two or more fixations, or skipping, again as a function of word length and the nature of the test sentences. As the figure shows, agents tended to either make a single fixation on or skip the shorter words, and to make two or more fixations on the longer words. These results are consistent with what is observed with humans (Rayner et al., 1996) and did not vary by testing condition.

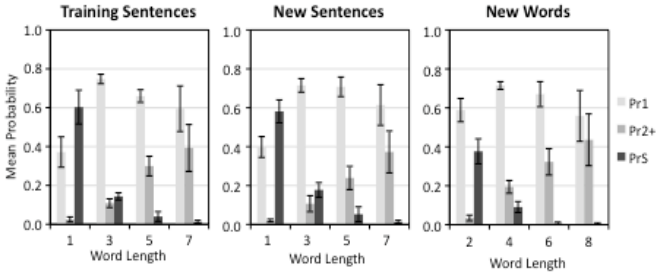


Figure 3. Simulated fixation probabilities.

Figure 4 shows the mean simulated values of five dependent measures (in time steps), again as a function of word length: (1) *first-fixation duration* (*FFD*), or the duration of the first fixation on a word during the first pass through the sentence; (2) *gaze duration* (*GD*), or the sum of all first-pass fixations; (3) *total-viewing time* (*TT*), or the sum of all fixations, irrespective of whether they occur during the first pass; (4) *word-identification times* (*ID*), or the time spent processing the words; and (5) *saccadic-programming initiation times* (*SPIT*), or the time spent processing $word_n$ prior to initiating the saccade that moved the eyes to $word_{n+1}$. As Figure 4 indicates, the measures increased with increasing word length (which is perfectly correlated with identification time), but with the mean processing time being longer than the minimal identification time because saccadic error often caused words to be processed from poor viewing locations, where lexical processing was slower. Importantly, if an agent spent *N* time steps processing $word_n$, then it on average spent approximately *N*-3 time steps processing $word_n$ before initiating saccadic programming to move its eyes to $word_{n+1}$. This is an optimal strategy for deciding when to move the eyes because initiating saccadic programming any earlier would cause the eyes to leave $word_n$ prematurely, resulting in it being processed more slowly from $word_{n+1}$ (because of reduced visual acuity). Conversely, initiating saccadic programming any later would cause the fixations to be unnecessary long in duration. Thus, by initiating saccadic programming at the observed times, an agent insures that, in most cases, the eyes move from $word_n$ right when it has been identified. (It is also worth noting that this strategy is similar to the "familiarity check" assumption of the E-Z Reader model of eye-movement control during reading, where a preliminary stage of lexical processing is the "trigger" that initiates saccadic programming; Reichle et al., 1998, 2003, 2009.)

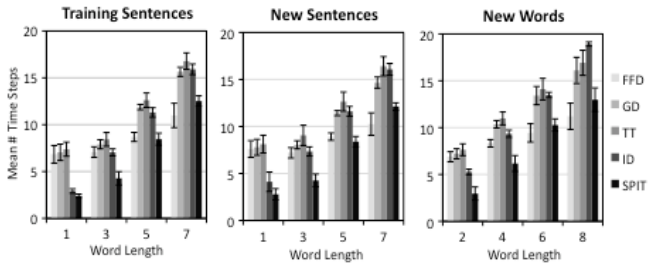


Figure 4. Simulated time-based dependent measures.

These results of Simulation 1 replicate and extend the key findings reported by Reichle and Laurent (2006) by showing that the reading agents, implemented as ANNs, are able to generalize from a small set of training sentences to sentences containing novel configurations of words. This is methodologically important because it shows how ANNs might be used to solve large-scale reinforcement learning problems that might otherwise be impossible to solve (e.g., a look-up table version of the agents would require the storage and updating of the more than 6 million different states listed in Table 1). This demonstration also makes it possible to explore more complex contingencies between eye-movement behavior and lexical variables, as described next.

Simulation 2

The second simulation examined the consequences of training on a more realistic sentence corpus—one in which word length is only moderately predictive of the time required to identify words.

Method. Five agents were trained and tested on five 8-word sentences in which 1-, 3-, 5-, and 7-letter words required 4-9 time steps to identify, and where word length was only moderately correlated to word-identification times across the corpus ($r = 0.31$).

Results. The simulated landing-site distributions, fixation probabilities, and time-based measures (grouped by both word length and identification time) are shown in Figures 5-7, respectively. As indicated in the left panel of Figure 5, the agents learned to direct their eyes towards the centers of words because this location afforded the most rapid identification of words. And as the left panel of Figure 6 shows, the agents were also more likely to make single fixations on or skip the shorter words, and make two or more fixations on the longer words. Both of these findings are consistent with human readers (McConkie et al., 1988, 1991; Rayner, 1996) and the results of Simulation 1. The right panels of Figures 5 and 6 indicate that similar word-targeting behaviors were evident when the words are grouped by their identification times, but that there are some irregularities (e.g., bimodal landing-site distributions with

the more-difficult-to-identify words) because these items included a mixture of both short and long words.

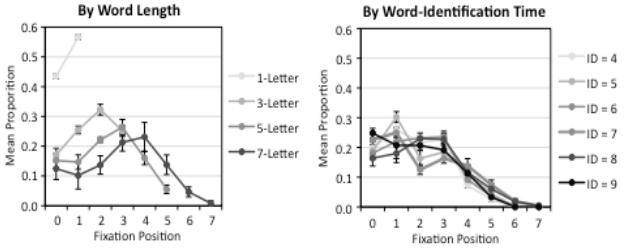


Figure 5. Simulated fixation landing-site distributions, by word length (left) and identification times (right).

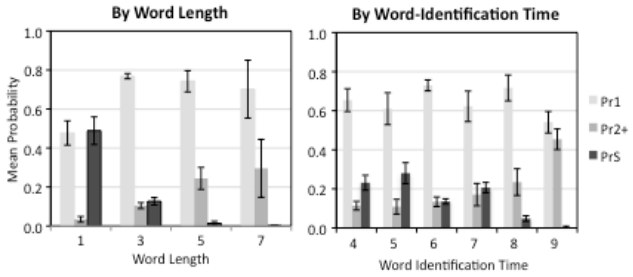


Figure 6. Simulated fixation probabilities, by word length (left) and identification times (right).

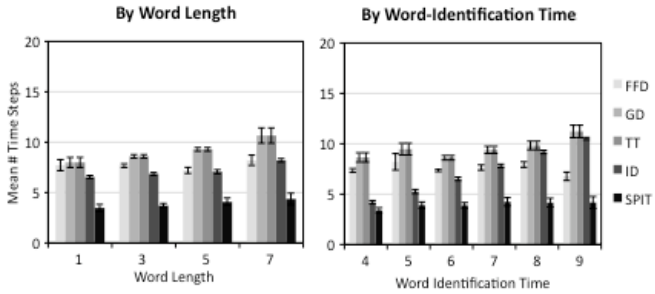


Figure 7. Simulated time-based measures, by word length (left) and identification times (right).

Finally, the most striking result from Simulation 2 is that the agents learn to use word length information to predict the time required to identify words, and then used this knowledge to program saccades so that the eyes would leave a word right as it was identified. This "strategy" is similar to the one that was adopted by the agents in Simulation 1, even though the relation between word length and identification times was much weaker in Simulation 2 (r = 0.31) than Simulation 1 (r = 1). And as the left and right panels of Figure 7 indicate, this strategy was evident irrespective of whether the words are grouped by their length or by their identification times.

General Discussion

The simulations reported in this article replicated the Reichle and Laurent (2006) results by showing that "intelligent" eye-movement behavior can emerge from artificial reading agents that are subject to fairly uncontroversial physiological and psychological constraints and that are capable of learning to coordinate attention and eye movements to support efficient reading. Simulation 1 extended the Reichle and Laurent results by implementing the reading agents within an ANN and then showing that the agents' eye-movement behaviors generalize to novel sentences and words. And importantly, the agents used word-length cues to predict when words would be identified, and then used this knowledge to learn when to initiate saccadic programming. Simulation 2 indicated that the agents learned the same eye-movement behaviors, including learning to use word length to initiate saccadic programming in an optimal manner—even though word length was only moderately predictive of word-identification time.

The simulation results have important theoretical implications for our general understanding of eye-movement control in reading and the specific questions of what determines when our eyes move during reading. First, the simulations suggest how information that is predictive of when a word will be identified can be used to initiate saccadic programming in a manner that affords efficient reading. In the absence of such predictive information, it may be optimal to either simply wait until $word_n$ has been identified before initiating saccadic programming to move the eyes to $word_{n+1}$, or to base the decision on the mean time required to identify $word_n$. Although both of these strategies prevent the eyes from moving prematurely (which would then slow reading considerably because words would have to be identified from poorer viewing locations), the strategies are also conservative because they often produce unnecessarily long fixations. This suggests that any strategy that simply ignores lexical processing difficulty and uses saccadic programming and visual acuity constraints to decide when to move the eyes will not be optimal because it ignores information (about the rate of lexical processing) that can be used to inform those decisions. This conclusion provides one argument against oculomotor-control theories of eye-movement control in reading (Feng, 2006; McDonald et al., 2005; Reilly & O'Regan, 1998; Suppes, 1990; Yang, 2006). And although our results admittedly say less about the feasibility of autonomous-timer theories (Engbert et al., 2002, 2005), such theories are not parsimonious if decisions about when to move the eyes can be made using information that is readily available to the reader (i.e., information about lexical processing difficulty). In other words, it is not parsimonious to posit an autonomous timer that is occasionally overridden by lexical processing difficulty if this information is itself sufficient to decide when to move the eyes in an optimal manner.

Second, the simulations suggest that humans may exploit cues that may be only modestly predictive of lexical processing difficulty in learning to decide when to initiate saccadic programming. These cues probably include word length, but also orthographic cues (e.g., prefixes and

suffixes, unusual letter sequences, etc.), and possibly cues that are generated via top-down processing (e.g., the syntactic and/or semantic constraints imposed by a word's prior sentence context). It is an open question as to how these different sources of information are combined in making moment-to-moment decisions about when to move the eyes during reading, but a vast experimental literature (e.g., for a review, see Rayner, 1998) indicates that these variables (and many others) do influence such decisions. Future simulations using our artificial reading agents will provide testable hypotheses regarding this question.

Acknowledgments

Address correspondence to: Erik Reichle, University of Pittsburgh, 635 LRDC, 3939 O'Hara St., Pittsburgh, PA 15260 USA (or via email at: reichle@pitt.edu.) This work was supported by a China Scholarship Council award to the first author and an NIH grant (R01HD053639) awarded to the second author.

References

Engbert, R., Longtin, A., & Kleigl, R. (2002). A dynamical model of saccade generation in reading based on spatially distributed lexical processing. *Vision Research*, *42*, 621-636.

Engbert, R., Nuthmann, A., Richter, E., & Kliegl, R. (2005). SWIFT: A dynamical model of saccade generation during reading. *Psychological Review*, *112*, 777-813.

Feng, G. (2006). Eye movements as time-series random variables: A stochastic model of eye movement control in reading. *Cognitive Systems Research*, *7*, 70-95.

Hansen, N., Müller, S. D., & Koumoutsakos, P. (2003). Reducing the complexity of the Derandomized Evolution Strategy with Covariance Matrix Adaptation (CMA-ES). *Evolutionary Computation*, *11*, 1-18.

Inhoff, A. W. & Rayner, K. (1986). Parafoveal word processing during eye fixations in reading: Effects of word frequency. *Perception & Psychophysics*, *4*0, 431-439.

Just, M. A. & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, *87*, 329-354.

McConkie, G. W., Kerr, P. W., Reddix, M. D., & Zola, D. (1988). Eye movement control during reading: I. The location of initial eye fixations in words. *Vision Research*, *28*, 1107-1118.

McDonald, S. A., Carpenter, R. H. S., & Shillcock, R. C. (2005). An anatomically constrained, stochastic model of eye movement control in reading. *Psychological Review*, *112*, 814-840.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*, 372-422.

Rayner, K. & Duffy, S. A. (1986). Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & Cognition*, *14*, 191-201.

Rayner, K. & Pollatsek, A. (1989). The Psychology of Reading. Englewood Cliffs, NJ: Prentice Hall.

Rayner, K., Sereno, S. A., & Raney, G. E. (1996). Eye movement control in reading: A comparison of two types of models. *Journal of Experimental Psychology: Human Perception and Performance*, *22*, 1188-1200.

Reichle, E. D. (2006). Theories of the "eye-mind" link: Computational models of eye-movement control during reading. *Cognitive Systems Research*, *7*, 2-3.

Reichle, E. D. & Laurent, P. A. (2006). Using reinforcement learning to understand the emergence of "intelligent" eye-movement behavior during reading. *Psychological Review*, *113*, 390-408.

Reichle, E. D., Liversedge, S. P., Pollatsek, A., & Rayner, K. (2009). Encoding multiple words simultaneously in reading is implausible. *Trends in Cognitive Sciences*, *13*, 115-119.

Reichle, E. D., Pollatsek, A., Fisher, D. L., & Rayner, K. (1998). Toward a model of eye movement control in reading. *Psychological Review*, *105*, 125-157.

Reichle, E. D., Rayner, K., & Pollatsek, A. (2003). The E-Z Reader model of eye movement control in reading: Comparisons to other models. *Behavioral and Brain Sciences*, *26*, 445-476.

Reichle, E. D., Warren, T., & McConnell, K. (2009). Using E-Z Reader to model the effects of higher-level language processing on eye movements during reading. *Psychonomic Bulletin & Review*, *16*, 1-21.

Reilly, R. (1993). A connectionist framework for modeling eye movements in reading. In G. d'Ydewalle & J. Van Rensbergen (Ed.), Perception and cognition: Advances in eye movement research (pp. 193-212). Amsterdam: North-Holland.

Reilly, R. & O'Regan, J. K. (1998). Eye movement control in reading: A simulation of some word-targeting strategies. *Vision Research*, *38*, 303-317.

Reilly, R. G. & Radach, R. (2006). Some empirical tests of an interactive activation model of eye movement control in reading. *Cognitive Systems Research*, *7*, 34-55.

Salvucci, D. D. (2001). An integrated model of eye movements and visual encoding. *Cognitive Systems Research*, *1*, 201-220.

Stanley, K. O. & Miikkulainen, R. (2002). Evolving neural networks through augmenting topologies. *Evolutionary Computation*, *10*, 99-127.

Suppes, P. (1990). Eye movement models for arithmetic and reading performance. In E. Kowler (Ed.), *Eye movements and their role in visual and cognitive processes* (pp. 395-453). Amsterdam: Elsevier.

Sutton, R. S. & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.

Yang, S.-N. (2006). An oculomotor-based model of eye movements in reading: The competitive/interaction model. *Cognitive Systems Research*, *7*, 56-69.

Rational eye movements in reading combining uncertainty about previous words with contextual probability

Klinton Bicknell & Roger Levy
{kbicknell, rlevy}@ling.ucsd.edu
UC San Diego Department of Linguistics
9500 Gilman Drive #108, La Jolla, CA 92093-0108 USA

Abstract

While there exist a range of sophisticated models of eye movements in reading, it remains an open question to what extent human eye movement behavior during reading is adaptive given the demands of the task. In this paper, we help to answer this question by presenting a model of reading that corrects two problems with a rational model of the task, Mr. Chips (Legge, Klitz, & Tjan, 1997). We show that the resulting model is closer to human performance across two measures, supporting the idea that many components of eye movement behavior in reading can be well understood as a rational response to the demands of the task.

Keywords: computational modeling; rational analysis; eye movements; reading

Introduction

Choosing when and where to move one's eyes during reading is one of the most complicated skilled tasks humans perform. While there are a number of computational models achieving good numerical fits on eye movement data from reading (e.g., Reichle, Pollatsek, & Rayner, 2006; Engbert, Nuthmann, Richter, & Kliegl, 2005), it is still unclear to what extent the complex behaviors observed are rational responses to the demands of the problem itself and to what extent they arise from the idiosyncrasies and restrictions of human cognition. Legge, Klitz, and Tjan (1997) started to answer this question with Mr. Chips, a model which predicts eye movements that approximate an optimal solution to one formalization of the task of reading. Legge et al. pointed out that their model's behavior exhibits a number of patterns also found in human reading, providing evidence for understanding those behaviors as rational responses to the task. Despite its success, however, the Mr. Chips model oversimplifies two important aspects of the problem of reading, and also has empirical problems accounting for human reading behavior in two domains. In this paper, we propose a model extending Mr. Chips that removes these two oversimplifications to make the model's task more similar to that faced by humans. We show that the resulting model also remedies the two empirical deficiencies in Mr. Chips, further supporting the notion that many aspects of human reading behavior can be explained as rational responses to the demands of reading.

The essentials of the problem of making eye movements in reading are determining how long to leave the eyes in a given spot and – when a reader decides to move them – where to go. These decisions are made sequentially to produce the alternating sequence of fixations (relatively stable periods) and saccades (movements) that characterizes the eye movement record. The past 30 years have seen a proliferation of experimental studies investigating this topic, which have answered a number of low-level questions such as the nature of the perceptual span and constraints on saccade latency as well as questions concerning the relationship between eye movements and higher-level cognitive processes such as the effect of word frequency and predictability (see Rayner, 1998 for an overview). Sophisticated computational models have been developed based on these findings, the most well-known of which are E-Z Reader (Reichle, Pollatsek, Fisher, & Rayner, 1998; Reichle et al., 2006) and SWIFT (Engbert et al., 2005). Both E-Z Reader and SWIFT assume that lexical processing (or word recognition) is the primary driver for eye-movements in reading, and both have enjoyed considerable success, in large part because they achieve very good fits to eye movement data from reading in a number of contexts, using a relatively small number of parameters. Despite their empirical strength, they fail to illuminate the reason why human reading behavior looks the way it does in one crucial respect – the extent to which it resembles a rational response to the problem posed by reading.

One leading approach for answering such questions is that of rational analysis (Anderson, 1990), a paradigm in which one formalizes the goals and cognitive and physical constraints relevant to a problem and develops a model of optimal behavior under those condition. To the extent that the behavior of the model is similar to that of humans, this provides a new way of understanding the reason why human behavior looks the way it does – it is the best way to solve the problem. The relationship between rational models and models such as E-Z Reader and SWIFT is well understood in terms of Marr's (1982) levels of analysis. Marr distinguishes three levels of mutually-constraining analyses that can be performed on cognitive processes: the *computational* level, which specifies the nature of the computation being performed, the information relevant to solving it, and the way to combine that information to solve it; the *algorithmic* level, which specifies the representation for the input and output and the algorithm by which the agent goes about solving it; and the *implementational* level, which specifies how the representations and algorithm are realized neurally. In these terms, rational models generally provide answers at the computational level of analysis. Models such as E-Z Reader and SWIFT help us to understand the algorithmic level, but cannot answer questions about the extent to which human reading is rational.

Legge et al. (1997) presented a computational level analy-

sis of reading, formalizing the central task – as in E-Z Reader and SWIFT – as one of serial word identification. They presented the Mr. Chips model, which approximates optimal behavior under their formalization, and shows a number of similarities with human reading behavior. Here, we point out two problems with their model of reading. First, their model takes the task to be to identify a string of independent words rather than a coherent sequence, i.e., their model does not make use of linguistic context, which experimental work suggests that humans use (McDonald & Shillcock, 2003). Second, it assumes that the task of the reader is to identify each word with complete certainty, yet recent evidence suggests that readers maintain uncertainty as to the identities of previous words (Levy, Bicknell, Slattery, & Rayner, 2009). In addition to these problems in their model's design, the model also makes incorrect predictions for two relatively basic measures of eye movements in reading: saccade sizes and word skipping rates. The model we present fixes these two design problems by including linguistic context and using a flexible word identification criterion, and results in improved performance in accounting for human saccade sizes and word skipping rates.

The plan of the remainder of the paper is as follows. First, we describe the details of the Mr. Chips model, along with its empirical successes and failures. Next, we describe our extension of the Mr. Chips model, and finally present two experiments showing that fixing each of the two design problems results in performance more like humans.

Mr. Chips

The task of reading in the Mr. Chips model (Legge et al., 1997) is one of planning saccades for serial word identification. That is, the model works by gathering visual input from the current fixation location and using that visual input to plan a saccade. That saccade is then executed (with some motor error), visual input is obtained from the new location, and the cycle repeats. When one word is identified with 100% confidence, identification of the next word begins. Thus, the only decision the model makes is where to move the eyes next. There are just three sources of information relevant to making that decision. Visual input and knowledge of the language are combined to identify words, and knowledge of the motor error in the system assists in the planning problem. Since it forms the basis for our model, we describe the Mr. Chips model here in detail, discussing in turn each of the sources of information and then the algorithm by which the model combines them to choose saccades. To match the description of our model later, we use a notation a bit different than Legge et al. to describe Mr. Chips.

Information sources

Visual input The visual input in Mr. Chips consists of the veridical identities of the nine characters centered on the fixated character (representing the visual fovea), as well as partial information about the four characters on either side of this range (representing the visual periphery). This partial information is simply word boundary information, indicating whether each character is part of a word or not (e.g., a space). The number of characters in each of these ranges was chosen to be representative of the perceptual span for readers of English, known to be around 17–19 characters (Rayner, 1998).

Language knowledge The model's knowledge of language consists of simply word frequency information, i.e., a unigram model. Note that this means the model cannot make use of the linguistic context to aid in word identification.

Motor error The final component of the model's knowledge of the task is that of motor error, the distribution of a saccade's landing position given the intended target position the model chooses. In Mr. Chips, the ith landing position ℓ_i is normally distributed around the ith intended target position t_i with a standard deviation of 30% of the intended distance[1]

$$\ell_i \sim \mathcal{N}\left(t_i, (0.3 \cdot |t_i - \ell_{i-1}|)^2\right). \quad (1)$$

Model

We now give the algorithm that the Mr. Chips model uses to select the intended target for the next saccade. First, note that given the visual input obtained by the model from the first to the ith fixation $\mathcal{I}_1^i$ and the word frequency information, the model can calculate the posterior probability of any possible identity of a word w that is consistent with the visual input by normalizing its probability from the language model by the total probability of all visually consistent identities,

$$p(w|\mathcal{I}_1^i) = \frac{\chi(\mathcal{I}_1^i, w)p(w)}{\sum_{w'} \chi(\mathcal{I}_1^i, w')p(w')} \quad (2)$$

where $\chi(\mathcal{I}, w)$ is an indicator function with a value of 1 if w is consistent with the visual input $\mathcal{I}$ and 0 otherwise, and $p(w)$ is the probability of w under the language model.

To identify a given word, the model selects the saccade target $\hat{t}_i$ that, on average, will minimize the entropy in this distribution, i.e., that is expected to give the most information about the word's identity

$$\hat{t}_i = \operatorname*{argmin}_{t_i} E\left[H(w|\mathcal{I}_1^i)|t_i, \mathcal{I}_1^{i-1}\right] \quad (3)$$

$$= \operatorname*{argmin}_{t_i} \sum_{\mathcal{I}_i} H(w|\mathcal{I}_1^i)p(\mathcal{I}_i|t_i, \mathcal{I}_1^{i-1}). \quad (4)$$

That is, the minimum can be found by calculating the conditional entropy produced by each possible new input sequence and weighting those entropies by the probability of getting that input sequence given a choice of target location. In information theory (Cover & Thomas, 2006), conditional entropy is standardly defined as

$$H(w|\mathcal{I}_1^i) = -\sum_w p(w|\mathcal{I}_1^i) \log p(w|\mathcal{I}_1^i). \quad (5)$$

[1] In the terminology of the literature, this model has only 'random' motor error (variance) and not 'systematic' motor error (bias), under the assumption that an optimal model would just compensate for any systematic problems with its motor control system.

The second term in the formula for $\hat{t}_i$, the probability of a particular visual input given a target location and previous input, is given by marginalizing over possible landing positions

$$p(\mathcal{I}_i|t_i,\mathcal{I}_1^{i-1}) = \sum_{\ell_i} p(\ell_i|t_i)p(\mathcal{I}_i|\ell_i,\mathcal{I}_1^{i-1}) \quad (6)$$

and then possible words

$$p(\mathcal{I}_i|\ell_i,\mathcal{I}_1^{i-1}) = \sum_{w} p(\mathcal{I}_i|\ell_i,w)p(w|\mathcal{I}_1^{i-1}). \quad (7)$$

Putting these together, we have that $\hat{t}_i$ is selected as

$$\underset{t_i}{\operatorname{argmin}} \sum_{\mathcal{I}_i} H(w|\mathcal{I}_1^i) \sum_{\ell_i} p(\ell_i|t_i) \sum_{w} p(\mathcal{I}_i|\ell_i,w)p(w|\mathcal{I}_1^{i-1}). \quad (8)$$

That is, we can calculate the expected conditional entropy for each possible value of t_i by summing over all possible inputs, whose probabilities are given by summing over all possible identities of the word and landing positions. To see that this sum ranges over a finite number of values, note first that there are only a finite number of possible word identities w to sum over. Given the possible word identities, there are only a finite number of landing positions ℓ_i for which the visual information could possibly help in identifying the word – any landing positions outside this range will not produce any reduction in entropy. Since there is a single visual input $\mathcal{I}_i$ for each combination of landing position and word identity, this summation is over a finite range. To ensure finiteness of the search to find the value of t_i that produces the minimum entropy, Mr. Chips only searches those within the range of the ℓ_i that could give some information about the current word. In case of ties, the model selects the furthest position to the right.

Comparing Mr. Chips to humans

Legge, Hooven, Klitz, Mansfield, and Tjan (2002) present a number of ways in which the behavior of the Mr. Chips model is similar to human reading behavior. The model produces behavior that replicates a number of human findings in word skipping rates, initial fixation locations on words, and refixation rates. The result for word skipping rates – where word skipping for the model is defined as never having any of the word's characters as the centrally fixated character – is that longer words are skipped less often, and the slope of the relationship between word length and skipping rate has a very similar slope for the model as for humans. For initial word fixation locations, or landing positions, the model replicates the human behavior of most commonly landing at or just to the left of the word's center, and also the fact that the landing position shifts toward the left as the launch site of the saccade shifts further to the left. For refixations, the model mimics human behavior in showing the proportion of refixations to increase with word length, and in addition, within a given word length class, the model refixates low frequency words a higher proportion of the time than high frequency ones. Finally, as a function of landing position, refixations are the least likely for the model, as for humans, when the initial landing position is near the center of the word.

As noted above, however, it is also the case that the model exhibits some behavior very different from that of human readers. For example, the model's average saccade length is just 6.3 characters, noticeably lower than that for humans, who are around 8 (Rayner, 1998). Second, although, as mentioned, the slope of the relationship between word skipping rates and word length has a similar slope for the model as for humans, the model skips far fewer words than humans do.[2] In short, judging by these two measures, a rational model that is using all the information available and expensively calculating the best saccades to reduce entropy in word identification appears to be reading slower than humans do.

In rational analysis, the fact that an 'optimal' model is performing worse than humans (here in terms of speed) suggests two likely problems: (a) the model is not making use of all the information that humans use or (b) the model's computational goal is not the same as the one that humans are solving. As suggested above, we argue that in this case both reasons are partially to blame. Since it has only word frequency information as its model of language, the Mr. Chips model cannot make use of linguistic context to aid in word identification, while there is evidence that humans make heavy use of it. The model also assumes that the goal is to identify each word with 100% confidence, but experiments suggest that humans do not. In the next section, we modify the Mr. Chips model to include some information about linguistic context and a flexible identification confidence criterion.

Extending Mr. Chips

The model described here generalizes the Mr. Chips model in three ways. First, it can use an arbitrary language model as its source of language knowledge, and thus make use of information about the linguistic context in word identification, solving the first problem with Mr. Chips we pointed out above. Second, it can move on to the next word after it achieves a flexible level of certainty about the current word's identity, solving the second problem. Finally, our model also allows for the standard deviation of the motor error to be an arbitrary linear function of the intended distance, allowing us to incorporate a more realistic motor error function. We describe the model in the same format as we described the Mr. Chips model, first describing its sources of information, and then its algorithm for selecting saccade targets.

Information sources

Visual input The visual input component is unchanged from the original Mr. Chips model.

Language knowledge The model's knowledge of language is represented by an arbitrary language model that can generate string prefix probabilities, e.g., an n-gram model or a

[2]The graph given in Legge et al. (2002) appears to show remarkably similar word skipping rates between the model and humans, but that graph is from the sole simulation in the paper for which Legge et al. assumed no motor error. When motor error is included, the skipping rates are significantly lower for the model than for humans, as shown in Figure 1.

probabilistic context-free grammar (*PCFG*). Such models can capture the between-word dependencies needed for the model to make use of linguistic context in word identification.

Motor error In our model as in Mr. Chips, the ith landing position is normally distributed around the ith target location, except that the standard deviation is an arbitrary linear function of the intended distance

$$\ell_i \sim \mathcal{N}\left(t_i, (\beta_0 + \beta_1 |t_i - \ell_{i-1}|)^2\right) \quad (9)$$

allowing for the use of a more realistic motor error function. Experiments in this paper use the one from SWIFT (Engbert et al., 2005; $\beta_0 = 0.87$, $\beta_1 = 0.084$).

Algorithm

As in the original Mr. Chips model, at any given point in time, the model is working to identify one word. However, this revised model considers the goal of identifying this word achieved when the marginal probability of some identity for the word given the visual input exceeds a predefined threshold probability α. This flexibility requires the algorithm to be substantially modified to allow for uncertainty about previous word identities and the use of linguistic context. We denote the sequence of words as W, where the first word is W_1.

Because every word in Mr. Chips was identified with complete certainty, the model always knew precisely at which position the next word to be identified began, and its goal was always to identify this next word. Now that the model has uncertainty about the identities of previous words, however, the goal must be changed. In the revised model, the reader is always focused on some character position n, and its goal is to identify whether some word $W_{(n)}$ begins at that position, and if so, which one, with confidence exceeding α. Once the model has achieved this goal, it then chooses a new character position n via a procedure whose description we leave for later. To be explicit about this goal, we slightly update our original equation for choosing $\hat{t}_i$, swapping w out for $W_{(n)}$,

$$\hat{t}_i = \underset{t_i}{\operatorname{argmin}} \sum_{\mathcal{I}_i} H(W_{(n)}|\mathcal{I}_1^i)\, p(\mathcal{I}_i|t_i, \mathcal{I}_1^{i-1}) \quad (10)$$

where the conditional entropy is calculated assuming that some word does in fact begin at position n. The fact that our language model can now make use of linguistic context means that the equation for finding the probability of the current word given some visual input (Equation 2) must also be changed to marginalize over identities of the preceding words

$$p(W_{(n)}|\mathcal{I}_1^i) = \sum_{W_1^{(n)-1}} p(W_{(n)}|\mathcal{I}_1^i, W_1^{(n)-1})\, p(W_1^{(n)-1}|\mathcal{I}_1^i). \quad (11)$$

These probabilities of strings consistent with the visual input are again given probabilities according to their probability in the language model normalized by the probability of all other consistent strings (cf. Equation 2)

$$p(W|\mathcal{I}_1^i) = \frac{\chi(\mathcal{I}_1^i, W)\, p(W)}{\sum_W \chi(\mathcal{I}_1^i, W)\, p(W)}. \quad (12)$$

The second term in Equation 10 is expanded as in Mr. Chips by marginalizing over possible landing positions

$$p(\mathcal{I}_i|t_i, \mathcal{I}_1^{i-1}) = \sum_{\ell_i} p(\ell_i|t_i)\, p(\mathcal{I}_i|\ell_i, \mathcal{I}_1^{i-1}), \quad (13)$$

but now to incorporate information about the linguistic context, we must next marginalize over possible full sentence strings instead of possible words

$$p(\mathcal{I}_i|\ell_i, \mathcal{I}_1^{i-1}) = \sum_W p(\mathcal{I}_i|\ell_i, W)\, p(W|\mathcal{I}_1^{i-1}). \quad (14)$$

If we make the simplifying assumption that the model does not consider possible future input about words that are after $W_{(n)}$, this sum can again be finitely computed for a given t_i by a relatively straightforward dynamic programming scheme. The range of possible values of t_i to search through also grows relative to Mr. Chips, because the model must consider not only any position that can give visual input about $W_{(n)}$ itself, but also positions that can give information about any position of uncertainty, since that may indirectly help to identify $W_{(n)}$ through linguistic context. In the case where the language model is an n-gram model, it can be shown that the minimum value of t_i that can contribute toward helping to identify $W_{(n)}$ only extends back to the first uncertain character after the most recent string of $n-1$ words for which the model has no residual uncertainty. Having established the method of selecting a saccade to identify $W_{(n)}$, we next give a description of the full algorithm of the model, including how to select n.

The model always begins reading by focusing on identifying $W_{(0)}$. Once the probability of some identity for $W_{(0)}$ is greater than α, all the possible identities of $W_{(0)}$ that have not been ruled out by visual input are combined into a set of possible 'prefixes'. Each of these prefixes has a conditional probability given the visual input, and each one predicts that the next word in the sentence begins at a particular position (i.e., two characters past the end of that string). Thus, the set of prefixes specify a probability distribution over the possible positions at which the next word begins. The model simply selects the most likely such position as the next character position n to focus on identifying $W_{(n)}$.

Now in the general case, the system has a set of prefixes together with their conditional probabilities given the visual input, and a position n, which it is trying to identity the word beginning at. It plans and executes saccades according to the formula for $\hat{t}_i$, and after getting each new piece of visual information, the model rules out not only possible candidates for the current word, but also possible prefix strings, and renormalizes both distributions. The model's attempt to identify $W_{(n)}$ can now end in one of two ways: (a) the model's confidence in some identity of $W_{(n)}$ exceeds the confidence threshold α or (b) the model eliminates all possible candidates for $W_{(n)}$ and thus knows that no word begins at that position. In the former case, the model creates all possible concatenations of prefixes ending 2 characters prior to $W_{(n)}$ (i.e., prefixes whose next word begins at n) with all the possible identities of

$W_{(n)}$, and adds these new strings to the set of prefixes. Then, in both cases, it removes those original prefixes whose next word begins at n from the set. Note that this update of the list of prefixes leaves unaffected prefixes that are incompatible with a word beginning at position n, but still compatible with visual input. Finally, since the set of prefixes again gives a distribution over the starting position of the next word, the model selects the most likely new n and the cycle continues.

Experiment 1

With our new model in place, we can now describe the two experiments we performed to test our hypotheses about the reasons for the Mr. Chips model's performance being below that of humans in terms of average saccade length and word skipping rates. In Experiment 1, we test the hypothesis that one of the reasons that its performance was below humans is due to its assumption that the goal of the reader is to identify each word with 100% confidence. Specifically, we compare the performance of our model using a 100% criterion vs. a 90% criterion. The former is equivalent to Mr. Chips except for the more realistic motor error function, so for ease of exposition, we will refer to it simply as Mr. Chips.

Methods

Confidence criterion We set $\alpha = 1.0$ to replicate Mr. Chips, and $\alpha = 0.9$ for the model using a slightly lower confidence criterion to trigger moving on to the next word.

Language model Both models used a unigram language model, smoothed with Kneser-Ney under default parameters (Chen & Goodman, 1998; equivalent to add-δ smoothing for a unigram model). As in Legge et al. (2002), the models were trained on a 280,000 word corpus of *Grimms' Fairy Tales*, containing 7503 unique words. This corpus was normalized by Legge et al. to convert all letters to lowercase, remove all punctuation other than apostrophes, convert all numbers to their alphabetic equivalents, and remove all gibberish words.

Text Following Legge et al. (2002), we tested the models by simulating the reading of 40,000 words of text generated from the language model, ensuring that the reading models had a normative probability model for the text they were reading.

Results

The results for mean saccade size for both models are given in the top two rows of Table 1. As shown in the table, using a criterion of 90% increases the average size of saccades, bringing it a bit closer to the human estimate of about 8 characters. The results for word skipping rates for the two models are plotted as the lower two lines in Figure 1. The results show a modest increase in word skipping rates across almost all word lengths for the new model with a lower criterion, bringing it closer to human performance.

Discussion

Although the gain is modest, the results of Experiment 1 show that changing the goal of the model to one more similar to that of human readers, i.e., relaxing the assumption that words need to be identified with 100% certainty, alters the performance of the model across two measures to look more like human performance. Such a result adds some support to the idea that the relevant human behavior is well understood as a rational response to the demands of the task. It is also worth pointing out that the resulting model maintains and uses uncertainty about previous input, something for which most models of sentence processing do not allow.

Table 1: Mean saccade size (and std. error) for each model

Model	Mean saccade size
Mr. Chips	6.7 (.012)
Without context, 90% criterion	7.1 (.013)
With context, 90% criterion	7.5 (.014)
Humans	$\approx$ 8 (Rayner, 1998)

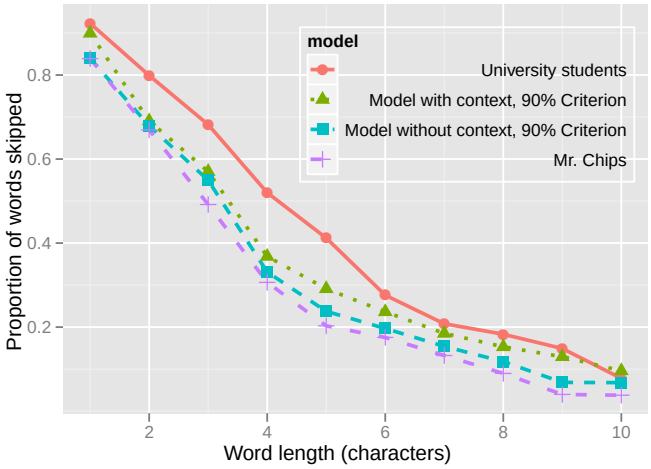


Figure 1: Proportion of words skipped by word length for each model. In all cases, the standard error of the mean for the Normal approximation to the Binomial distribution is smaller than the symbols. The human data is from Rayner and McConkie (1976) and has no standard errors.

Experiment 2

In Experiment 2, we test the effect of allowing the model to use the linguistic context as another source of information for word identification. Specifically, we compare the previous two models to one that includes a 90% confidence criterion as well as a simple bigram model of linguistic context.

Methods

The methods are the same as Experiment 1, except that the new model uses a bigram language model, again smoothed with Kneser-Ney under default parameters.

Results

The results for average saccade length for the new model is given in the third row of Table 1. As shown in the table, giving the model information about linguistic structure increases the

average size of saccades a bit more, bringing it still closer to the human estimate of 8 characters.

The results for word skipping rates for the new model is plotted as the second line in Figure 1. This new model gives an even larger increase in word skipping rates across all word lengths, on top of the increase seen previously, bringing it more in line with human results. Skipping rates are 30% closer to humans than the previous 90% criterion model on average, and for some word lengths are up to 75% closer.

Discussion

The results of this experiment show a case in which making more of the information that is available to a human reader also available to a rational model causes its behavior to more closely approximate human performance. Together with the previous result, this supports the notion that some aspects of reading are well understood as a rational response to the structure of the problem.

General Discussion

In this paper, we presented a new rational model of reading based on Mr. Chips, but which fixes two problems with it – it uses information about linguistic context in word identification and a flexible identification criterion. Experiment 1 showed that the model's performance looks more like humans' when the model's goal is shifted to one more like that of humans, 90% confidence in each word. Experiment 2 showed the model's behavior looks even more like humans' when the model can use information that is used by humans: linguistic context. Taken together, these results suggest that many facets of human reading behavior can be well explained as resulting from a rational solution to the problem of reading. This model adds to the growing literature on rational process models, exploring the extent to which human performance can be viewed as rational agents across a wide variety of complex behaviors, such as multiple object tracking (Vul, Frank, Alvarez, & Tenenbaum, 2009) and online change detection (Brown & Steyvers, 2009).

It is the case, however, that many aspects of human reading behavior cannot in principle be explained by a model such as those described in this paper. This is because much of what we know about human reading behavior is about fixation durations, and these models have no notion of duration. They cannot have a notion of duration because visual input is veridical categorical information about a range of characters, so that there is no reason to stay at a given location for more than one timestep. Reichle and Laurent (2006) overcome this problem by making the simplifying assumption that required processing times on words are a function only of their length.

We believe, however, that the way forward for rational models of reading is to incorporate a model of noisy visual input, so that the model can make decisions about fixation durations to get more or less visual information. In other work (Bicknell & Levy, 2010), we explore the use of such models to answer questions that are impossible to ask with non-rational models of reading such as when and why should between-word regressions be made, and how should reading behavior change as accuracy is valued more or less relative to speed.

Acknowledgments

We are very grateful to Gordon Legge for sharing the corpus used in the original Mr. Chips experiments. This work also benefited greatly from useful discussion with Jeff Elman, Tom Griffiths, Andy Kehler, Keith Rayner, and Angela Yu, and feedback from the audience at the 84th Annual Meeting of the Linguistic Society of America. The research was partially supported by NIH Training Grant T32-DC000041 from the Center for Research in Language at UC San Diego to K.B., by a research grant from the UC San Diego Academic Senate to R.L., and by NSF grant 0953870 to R.L.

References

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Bicknell, K., & Levy, R. (2010). A rational model of eye movement control in reading. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics ACL*. Uppsala, Sweden: Association for Computational Linguistics.

Brown, S. D., & Steyvers, M. (2009). Detecting and predicting changes. *Cognitive Psychology*, *58*, 49–67.

Chen, S. F., & Goodman, J. (1998). *An empirical study of smoothing techniques for language modeling* (Tech. Rep. No. TR-10-98). Cambridge, MA: Computer Science Group, Harvard University.

Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory* (2nd ed.). Hoboken, NJ: John Wiley & Sons.

Engbert, R., Nuthmann, A., Richter, E. M., & Kliegl, R. (2005). SWIFT: A dynamical model of saccade generation during reading. *Psychological Review*, *112*, 777–813.

Legge, G. E., Hooven, T. A., Klitz, T. S., Mansfield, J. S., & Tjan, B. S. (2002). Mr. Chips 2002: new insights from an ideal-observer model of reading. *Vision Research*, *42*, 2219–2234.

Legge, G. E., Klitz, T. S., & Tjan, B. S. (1997). Mr. Chips: an Ideal-Observer model of reading. *Psychological Review*, *104*, 524–553.

Levy, R., Bicknell, K., Slattery, T., & Rayner, K. (2009). Eye movement evidence that readers maintain and act on uncertainty about past linguistic input. *Proceedings of the National Academy of Sciences*, *106*, 21086–21090.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco: W.H. Freeman.

McDonald, S. A., & Shillcock, R. C. (2003). Eye movements reveal the on-line computation of lexical probabilities during reading. *Psychological Science*, *14*, 648–652.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*, 372–422.

Rayner, K., & McConkie, G. W. (1976). What guides a reader's eye movements? *Vision Research*, *16*, 829–837.

Reichle, E. D., & Laurent, P. A. (2006). Using reinforcement learning to understand the emergence of "intelligent" eye-movement behavior during reading. *Psychological Review*, *113*, 390–408.

Reichle, E. D., Pollatsek, A., Fisher, D. L., & Rayner, K. (1998). Toward a model of eye movement control in reading. *Psychological Review*, *105*, 125–157.

Reichle, E. D., Pollatsek, A., & Rayner, K. (2006). E-Z Reader: A cognitive-control, serial-attention model of eye-movement behavior during reading. *Cognitive Systems Research*, *7*, 4–22.

Vul, E., Frank, M., Alvarez, G., & Tenenbaum, J. (2009). Explaining human multiple object tracking as resource-constrained approximate inference in a dynamic probabilistic model. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, & A. Culotta (Eds.), *Advances in Neural Information Processing Systems 22* (pp. 1955–1963).

A New Perspective on Visual Word Processing Efficiency

Joseph W. Houpt (jhoupt@indiana.edu) and **James T. Townsend (jtownsen@indiana.edu)**
Indiana University, Department of Psychological and Brain Sciences
1101 E. Tenth Street, Bloomington, IN 47401 USA

Abstract

As a fundamental part of our daily lives, visual word processing has received much attention in the psychological literature. Despite the well established perceptual advantages of word and pseudoword context using accuracy, a comparable effect using response times has been elusive. Some researchers continue to question whether the advantage due to word context is perceptual. We use the capacity coefficient, a well established, response time based measure of efficiency to provide evidence of word processing as a particularly efficient perceptual process to complement those results from the accuracy domain.

Keywords: Visual word perception; Efficiency; Workload capacity.

As a fundamental part of our daily lives, visual word processing has received much attention in the psychological literature. However, the interest in visual word perception extends beyond its value in communication. The written word is a complex stimulus with which most adults have a large amount of experience. Unlike faces, there is no reason to believe we have any innate ability to perceive words. Thus, word perception may represent the limit of perceptual learning in the absence of innate ability.

Due to the relative ease with which most adults read, it is reasonable to assume that word perception is an efficient process. This is further supported by the intuition that with more experience with a process we become more efficient and we are quite experienced with the written word. Often, the efficiency is measured using single letter perception as a base line. When word context offers an advantage in the accuracy or processing time of perceiving a letter, this supports the claim that word perception is efficient.

From the early days of experimental psychology, researchers have been interested in the value of a word context for perceiving letters. In one study, letters were displayed sequentially to participants at faster and faster rates until they could no longer correctly identify the letters. They found that participants maintained accuracy with shorter durations when the letters were presented as part of a word compared with random letter sequences (Cattell, 1886).

One problem with studies of this nature is that they do not control for the constraint on possible letters that a word context puts on the possible letters. Hence it is not clear from those early results whether the advantage is a perceptual advantage or a decisional advantage. In the late 1960's an alternative task was designed to eliminate the decisional advantage of word context so as to examine the perceptual effects. In this task a letter or word was tachistoscopically displayed to a participant. They then chose from two possible choices, one of which was correct. In the letter condition, the choices were letters. In the word condition, both choices were words that differed in only a single letter. Since both alternatives were words, the word context was no longer informative as to the identity of the letter. Participants were still more accurate at perceiving letters in the word condition than the letter condition (Reicher, 1969). Furthermore, they found that participants are also more accurate with word contexts than random letter sequence contexts. This is known as the word superiority effect. An efficiency gain of context over letters alone is not unique to words though. If a sequence of letters conformed to the pronunciation rules of English, strings referred to as pseudowords, then participants were again more accurate than letters alone (e.g., McClelland & Johnston, 1977). This is known as the pseudoword superiority effect.

Despite the robustness of the word and pseudoword superiority effects, a comparable effect using response times (and controlling for decisional information due to context) has been elusive. This may be in part explained by the possibility that people will read an entire word even if the task does not require it. Indeed, this has been put forth as further evidence that word perception is special (LaBerge & Samuels, 1974). One of the goals of this paper is to demonstrate a response time based word superiority effect, and possibly a pseudoword superiority effect as well.

Even in the accuracy domain, some researchers continue to question whether there is a *perceptual* advantage due to word context. For example, Pelli, Farell, and Moore (2003) demonstrated evidence for a model of word perception in which letters are perceived independently and with separate detection decisions on each letter. Their evidence comes from comparing the efficiency of word perception as the number of letters in the word increases. Depictions of longer words have more information about their identity, since the more letters that are known, the fewer possibilities there are for the others. Hence, if a person is able to take advantage of this global information, they should need less per letter information as the number of letters increases. However, a model of word perception based on independent, separate decisions on the letters predicts that as the word length increases, the reader will still need the same amount of information per letter to maintain accuracy. In fact participants did need roughly the same amount of per letter information as the number of letters increased, supporting the latter model.

In the next section we describe the capacity coefficient, a response time based measure of efficiency. We propose that this measure, along with a task that controls for both the available information and possibly mandatory word reading, provides evidence of word processing as a particularly efficient process to complement those results from the accuracy domain.

	Target	Distractors				Single Character							
Word	care	bare	cure	cave	card	c	b	a	u	r	v	e	d
Pseudoword	lerb	nerb	larb	lemb	lerf	l	n	e	a	r	m	b	f
Non-Word	rlkf	vlkf	rtkf	rlhf	rljk	r	v	l	t	k	h	f	k
Upside-down	ɟʞlɹ	ɟʞlʌ	ɟʞtɹ	ɟɥlɹ	ʞlʃɹ	ɹ	ʌ	l	ɾ	ʞ	ɥ	ɟ	ʞ
Katakana	サイクオ	へイクオ	サナクオ	サイフオ	サイクノ	サ	へ	イ	ナ	ク	フ	オ	ノ

Table 1: Stimuli used for capacity analysis.

The Capacity Coefficient

The capacity coefficient, $C(t)$ is an established response time based measure of the effect of increased load on processing efficiency (Townsend & Nozawa, 1995; Townsend & Wenger, 2004). Specifically, $C(t)$ is a measure of the change in processing rates as the task requires attention to more targets, or possibly more dimensions of a single target. The basic idea of the measure is to compare response times when reading the full string to the times that would be predicted if each character took the same amount of time, whether or not it was in a string.

The capacity function for an exhaustive task is defined using the natural log of the cumulative distribution function, $K(t) = \ln F(t); F(t) = \Pr\{\mathrm{RT} \leq \mathrm{t}\}$, and is similar to the cumulative hazard function used in survival analysis. If K_{c1} is the cumulative hazard for the first character response times, K_{c2} is the cumulative hazard for the second character, etc., and K_S is the cumulative hazard for the string condition, the capacity coefficient is given by $C(t) = \left[\sum_{i=1}^{4} K_{c_i}\right] / K_S$.

This formulation is based on the predictions of the unlimited capacity, independent, parallel (UCIP) model. The assumptions of the UCIP model are sufficient conditions for there to be no change in the rates of processing with increased load. If these assumptions hold then the relationship between the processing times of the string to the processing times of the individual characters is as follows:

$$\Pr\{RT_S \leq t\} = \Pr\{RT_{c_1} \leq t\} \Pr\{RT_{c_2} \leq t\} \Pr\{RT_{c_3} \leq t\} \Pr\{RT_{c_4} \leq t\}$$

By taking the natural log of both sides of this equation, then dividing by the left hand side, we see that the UCIP model predicts $C(t) = 1$ for all $t \geq 0$. This gives us a baseline for comparison. If a person performs better than the baseline model, $C(t) > 1$, their performance is referred to as super-capacity. There are multiple ways performance could be super-capacity. For example, if there is facilitation between the characters, or in more extreme cases if the information from the characters is accumulated together toward a single decision (Townsend & Wenger, 2004).

Performance worse than the baseline model, $C(t) < 1$, is limited-capacity. In contrast to the case of super-capacity, inhibition between characters could result in limited-capacity. A standard serial model (independent and unlimited capacity) also predicts limited capacity. Furthermore, if the system only has a certain amount of resources to dedicate to the task, limited capacity performance would be expected. For example, a fixed capacity parallel model (a finite amount of resources is evenly divided up among the current processes) also predicts limited capacity.

When performance is about the same as the baseline model, $C(t) \approx 1$, then we refer to it as unlimited capacity. Of course this would happen if all of the assumptions of the baseline model were met. Alternatively some blend of features that lead to limited capacity and features that lead to super-capacity could balance out and result in capacity around 1. It is not likely that people are truly unlimited-capacity or super-capacity in the extreme case of very long words, but it is reasonable for shorter words.

The capacity coefficient measures processing efficiency in isolation by comparing the capacity coefficient to predicted values of unlimited capacity, independent, parallel models. Thus, this measure also allows us to compare the efficiency of a variety of processes despite any possible differences in difficulty due to component processes. In particular, we are able to draw conclusions about the efficiency of word processing relative to pseudoword, non-word, upside-down non-word, and unfamiliar character string processing.

Method

To properly compare perceptual efficiency across words, pseudowords, non-words, upside-down words and unfamiliar characters, our task must eliminate the extra information available given a word context. Furthermore, the possibility that words are exhaustively processed automatically may lead to a disadvantage for words on response time measures. To address these issues, we adapted a task from Blaha, Busey, and Townsend (2009) which forces exhaustive processing of the characters in a string. This experiment consists of two components. First, we measure the participants response times to correctly identifying the target string. To ensure that participants base their identification on the entire string and not any subset, we include a distractor of a string with a single character different in each position in the string. For example if the target is "care" then "bare," "cure," "cave" and "card" are used as distractors (see Table 1). Second, the participants distinguish between letters in isolation. Whereas in the exhaustive case the participant needed to distinguish between "bare" and "care," we now only require them to distinguish between "b" and "c." The response times on these tasks are used for computing the predicted performance of the UCIP model.

Stimuli Stimuli were created using GIMP version 2.2. For each stimulus, the character or characters were written in black in 29pt Courier onto a gray (200) background. Then each stimulus was doubled in size. There were five types of stimuli used: words, pronounceable non-words (pseudowords), unpronounceable non-words, upside-down unpronounceable non-words, and strings of Katakana characters. All strings used were four characters long. Words were chosen so that the frequency of the target was roughly equal to the average frequency of the distractors. Pseudowords were taken from the ARC Nonword Database (Rastle, Harrington, & Coltheart, 2002). Table 1 summarizes the stimuli used for both the single character and exhaustive trials for each type.

Participants Participants were recruited from the Indiana University population. Eight females and two males participated in this study, all of whom were native English speakers and reported that they did not read or speak Japanese. Their ages ranged from 19-34. All participants reported having normal or corrected to normal vision, no difficulty reading English, and no prior diagnoses of a reading disorder. Participants were paid $8 per session, and received a $20 bonus upon completion of all sessions. Each session lasted between 45 and 60 minutes.

Procedure All experimental conditions were run using DMDX version 2.9.06 developed at Monash University and at the University of Arizona by K.I.Forster and J.C.Forster. Stimuli were presented on a 17 Dell Trinitron CRT monitor running in 1024x720 mode. Participants used a two-button mouse for their responses.

Each session was dedicated to a type of stimuli and there were ten total sessions, two sessions for each type. At the beginning of each session, we read the participant the general instructions for the task while those instructions were presented on the screen. The instructions encouraged participants to respond as quickly as possible while maintaining a high level of accuracy. Each session was divided into five blocks, one block of string stimuli and a block for each of the corresponding single character stimuli. Each block began with a screen depicting the button that corresponded to each of the categories. Participants had 40 practice trials, 20 of each category. Next participants were given 240 trials divided evenly between the two categories, the first 40 of which were not used in the analysis. Each trial began with a 30 ms presentation of a fixation cross. After a random delay (300-600 ms), the stimulus was presented for 80 ms. Participants had a maximum of 2500 ms to respond. If the participant responded correctly, the next trial started after a 400 ms delay. If the participant responded incorrectly, a tone was played then the next trial started after a 400 ms delay. The session order was counterbalanced among the participants so that participants completed the different types on different days and in different orders.

Analysis All data was analyzed using R. Analysis was limited to the correct responses on the target category. We computed a repeated measures ANOVA of the response times in each condition. We then calculated the AND capacity coefficient, $C(t)$ for each participant and each condition. For each capacity coefficient, bootstrapped confidence intervals (95%) were calculated to determine if the function was reliably above or below 1 for any length of time.

To facilitate comparison between conditions, we analyzed the capacity functions using functional principal component analysis (fPCA, Ramsay & Silverman, 1997). In fPCA, each capacity function is treated as a weighted linear combination of a common set of basis functions. The set of weights is specific to each function and are therefore an alternative representation of the individual function. Similar to multivariate PCA, the basis functions each explain some percentage of the variance in the data. By treating those basis functions that explain only small amounts of variance as noise, we can achieve a concise representation of our data in terms of just the weights. The justification for applying fPCA is essentially the same as standard PCA; further details can be found in Ramsay and Silverman (1997).

To apply fPCA to capacity coefficient functions, we first calculated the empirical $K(t)$ by taking the natural log of the empirical cumulative distribution functions. Each of those functions were then interpolated using monotone cubic interpolation. The capacity coefficient for each condition was then calculated for each condition using those estimated K functions, then registered by aligning the median of each participants response time data in each condition. Functional principal components was then applied to the smoothed and registered functions, with the data weighted by the overall density function of the response time data and factor scores were computed based on a varimax rotation.

Results

Using a repeated measures ANOVA we found significant effects of condition $[F(4,18713) = 937.75, p < 2.2e-16]$ and whether the stimulus was a target for each of the string stimuli $[F(1,18713) = 10.75, p = 0.001]$, along with a significant interaction effect $[F(4,18713) = 57.73, p < 2.2e-16]$. Eight of the ten participants were faster on words and pseudowords than the other two conditions. Participant 6 was fastest on non-words while Participant 1 was fastest with words, but faster with non-words and upside-down non-words than pseudowords. Eight of the ten participants were slowest with Katakana, while Participant 7 was slowest with non-words and Participant 9 was slowest with upside-down words. While these results are interesting, this level of analysis does not account for the varied difficulty of processing each of the components. Hence, we turn to the capacity coefficient.

The results for the capacity analysis are shown in Figure 1. Bootstrap confidence intervals were used to determine significance, but are not included due to space limitations. Significance in comparisons against the UCIP model was determined by overlap of 99% confidence intervals with $C(t) = 1$.

In the word condition, nine participants had capacity coef-

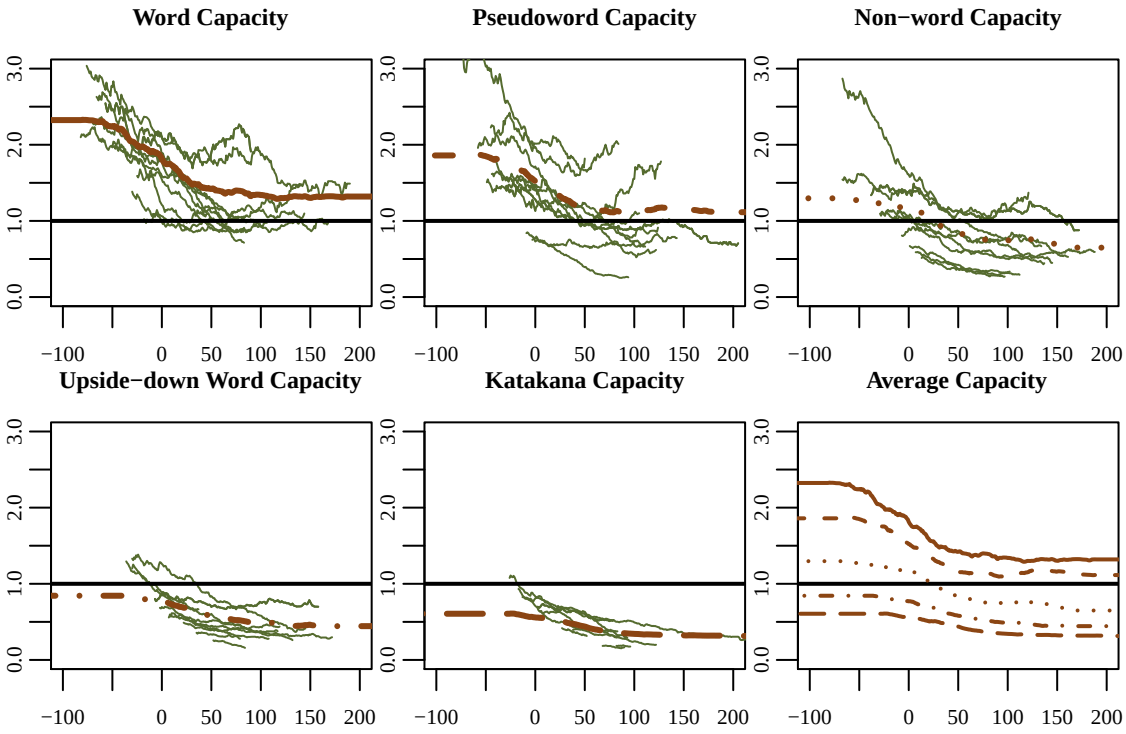


Figure 1: Capacity Coefficient values across time. Thin lines represent individual participant data and thick lines are the mean function across participants. The bottom right panel contains the mean function for each condition together.

ficient values significantly above one for at least some time (1-3, 5-10) and no participant had capacity values below 1 for any time. In the pseudoword condition, eight participants were super-capacity for some time (1, 2, 5-10), two of whom were limited capacity at later times (7, 8). The other two were limited capacity most of the time (3, 4). Three participants were super-capacity for some time in the non-word condition (1, 6, 8), one of whom was limited capacity for later times (8). The other seven participants were limited capacity for some time (2-5, 7, 9, 10), six of whom were limited capacity most of the time (3-5, 7, 9, 10). In the upside-down non-word condition, all participants were limited capacity for most times, with only three participants significantly super-capacity for early times (6-8). All participants were limited capacity for all times in the Katakana condition.

The bottom right panel of Figure 1 shows the capacity function averaged across participants for each condition. The average word capacity was the highest, followed by pseudoword, non-word, upside-down non-word, and lastly Katakana.

Figure 2 depicts the first principal component function of the capacity coefficients. This demonstrates that the feature that best distinguishes performance is a relatively large change in magnitude at early times, tapering off to a slight opposite change at later times. This first principal component describes 94% of the variance in the capacity functions. Furthermore, the condition was found to be a significant predictor of the factor score on this component in a repeated measures ANOVA $[F(4,36) = 18.56, p < 2e-8]$.

Discussion

Due to space limitations, we limit the majority of our discussion to the word and, to a lesser extent, the pseudoword results. We have demonstrated clear evidence of super-capacity processing of the word stimuli for nine of the ten participants. These participants are efficiently perceiving the whole word in comparison to individual letter perception. As mentioned earlier, evidence for the word superiority effect has been difficult to demonstrate with response times. These findings provide that evidence and thus agree with the majority of the word perception literature based on accuracy results. Based on comparisons across conditions, it is also clear that the word perception was more efficient than non-word, upside-down word, and strings of Katakana perception, findings that again match with the results reported for accuracy (e.g., McClelland & Rumelhart, 1981).

There is also evidence for a pseudoword superiority effect, another well established effect in the accuracy domain (McClelland & Johnston, 1977). Although the evidence was not as consistent as the word results, eight of the ten participants were super capacity for some time, with only two participants showing significantly limited capacity processing for most times.

The functional principal components analysis demonstrates that most of the differences in capacity across participants and types of stimuli occur early in the response time

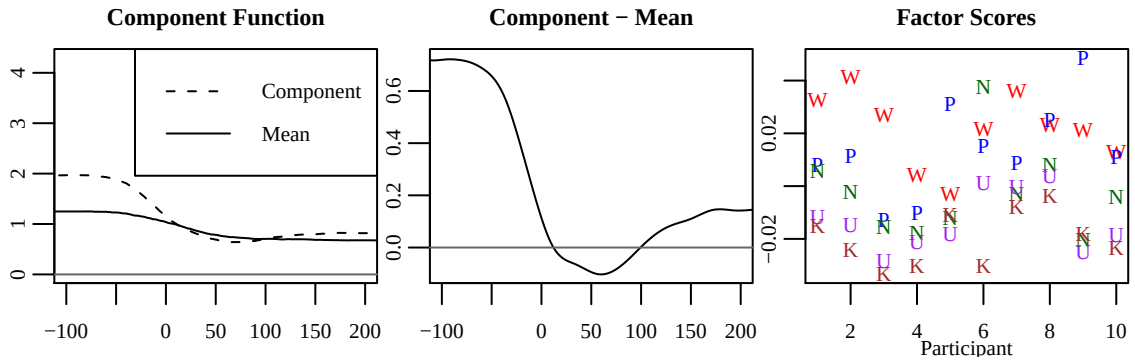


Figure 2: The first principal component of the capacity coefficient functions, which accounts for 94% of the variance. The first panel shows the component function with the mean capacity function for comparison. The second panel shows the difference between the component and the mean functions. The third panel depicts the factor scores on this component for each participant in each condition.

distribution. For all participants except Participant 6, the factor scores are higher for words and pseudowords than any of the other three conditions. This indicates that the word and pseudoword superiority effects are most pronounced in faster response times.

There are multiple plausible explanations for the capacity coefficient results demonstrating particularly efficient processing of words. At least one of the assumptions of the UCIP model must have been violated, so we examine each of those assumptions in turn. Each of these violations have been considered previously for modeling the accuracy based superiority effects.

One assumption that may have been violated is that of independence. If there is any type of facilitation between the letter processes, each letter would be processed faster within a word which would explain the capacity coefficient values above one. There could be many explanations of this facilitation. For example, word processing mechanisms may in fact take advantage of the considerable amount of co-occurrence between letters in English. As is often observed, there are only a fraction of possible four letter combinations used for words and it would be surprising if we did not take some advantage of this reduction in uncertainty. This correlation between letters is an important part of how connectionist models explain the word superiority effect (McClelland & Rumelhart, 1981; Plaut, McClelland, Seidenberg, & Patterson, 1996; Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001).

Another, related component of many visual word processing models is the phonological pathway. If a phoneme is activated as a possible interpretation of some letter combination, then it may in turn send positive feedback to those letters, speeding up their processing. Hence a phonological component of visual word processing could also lead to capacity coefficient values above one. Both the correlation between letters and the lack of a regular pronunciation of the non-words imply that these predictions are consistent with lack of evidence against the UCIP model of non-word processing. The phonological explanation is also supported by the evidence of a pseudoword superiority effect.

Another assumption of the UCIP model is that the letters are processed in parallel, with a separate detection of each letter. An alternative architecture that does predict capacity coefficient values above one is the coactive architecture. By pooling activation from each of the letters when processing a word, the word is processed much faster than if each letter is processed separately. A coactive architecture in this sense can be thought of as an extreme version of a facilitatory parallel model, in which all activation in each of the letters is shared. Many connectionist models of visual word perception assume a type of coactive architecture. In these models the activation accumulated in favor of a letter is immediately passed on to the word level. In this framework the type of parallel model assumed in the UCIP would not pass on any activation until the letter process is complete. There is some middle ground between these two extremes. One example is that of squelching suggested by Pelli et al. (2003). In this case, the activation from the letter process would only be passed on once it is above a certain threshold. It is important to note that these results are not necessarily inconsistent with serial processing, but for a serial model to predict capacity-values above one it would need to include facilitation and/or be super-capacity.

A coactive architecture could also lead to violations of the assumption of unlimited capacity, so that seemingly more resources are available to each component when more components are present. Capacity values above one imply that the participant had more than four times as many resources dedicated to the word task compared to the letter task, so that none of the individual letter processes has less. In this sense the advantage is similar to chunking; when groups of letters are recognized together, fewer resources are needed for each individual letter. Participants probably do not have truly unlimited resources to dedicate to the task, but having enough to act super-capacity with four letters is not so unreasonable.

There were clearly individual differences present in these

data, particularly in word and pseudoword processing capacity. This finding mirrors results reported in accuracy based studies (e.g., Reicher, 1969) and it will be an interesting extension of this work to compare the capacity measure to established measures of individual differences in reading.

Finally, we reiterate the importance of going beyond the simple ANOVA analysis of these data. Merely finding an ordering of the means in the string conditions says nothing about the relative processing efficiencies. For example, faster word processing than non-word processing could be due to the letters in "care" being relatively faster to process than the letters "rlkf". Workload capacity analysis, however, takes the processing of the components into account in estimating efficiency.

Summary

We have demonstrated response time based evidence for visual word perception as a particularly efficient process. This includes evidence that words are more efficiently perceived than predicted by the individual letter reading times, and evidence from comparing word perception efficiency to nonword stimuli. Based on the workload capacity analysis, there is also evidence for a pseudoword superiority effect in the response time domain although not as strong as for word superiority. The evidence we present negates models of word processing that assume parallel, independent processing of letters with separate decision thresholds on each channel. This deeper level of understanding of visual word perception required a shift from statistics based on comparing means toward a more theoretically rich, modeling-based approach.

Acknowledgments

This research was supported by NIH-NIMH MH 057717-07 and AFOSR FA9550-07-1-0078 awarded to JTT.

References

Blaha, L. M., Busey, T., & Townsend, J. T. (2009). An LDA approach to the neural correlates of configural learning. In N. A. Taatgetn & H. van Rijn (Eds.), *Proceedings of the 31st annual conference of the cognitive science society.* Austin, TX: Cognitive Science Society.

Cattell, J. M. (1886). The time it takes to see and name objects. *Mind*, *11*, 63–65.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review*, *108*, 204–256.

LaBerge, D., & Samuels, S. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology*, *6*, 293–323.

McClelland, J. L., & Johnston, J. C. (1977). The role of familiar units in perception of words and nonwords. *Perception and Psychophysics*, *22*, 249–261.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter percpetion: Part 1. an account of basic findings. *Psychological Review*, *88*, 375–407.

Pelli, D. G., Farell, B., & Moore, D. C. (2003). The remarkable inefficiency of word recognition. *Nature*, *423*, 752–756.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review*, *103*, 56–115.

Ramsay, J. O., & Silverman, B. W. (1997). *Functional data analysis*. New York: Springer-Verlag.

Rastle, K., Harrington, J., & Coltheart, M. (2002). 358,534 nonwords: The ARC nonword database. *Quarterly Journal of Experimental Psychology*, *55A*, 1339–1362.

Reicher, G. M. (1969). Perceptual recognition as a function of meaningfulness of stimulus material. *Journal of Experimental Psychology*, *81*, 274–280.

Townsend, J. T., & Nozawa, G. (1995). Spatio-temporal properties of elementary perception: An investigation of parallel, serial and coactive theories. *Journal of Mathematical Psychology*, *39*, 321–360.

Townsend, J. T., & Wenger, M. (2004). A theory of interactive parallel processing: New capacity measures and predictions for a response time inequality series. *Psychological Review*, *111*, 1003–1035.

The online processing of modal verbs: Parallel activation of competing mental models

Stephanie Huette (shuette@ucmerced.edu),
Teenie Matlock (tmatlock@ucmerced.edu),
and Michael J. Spivey (spivey@ucmerced.edu)
School of Social Sciences, Humanities & Arts, University of California, Merced, Merced, CA, 95344, USA

Abstract

A simple audio-visual two-alternative forced-choice task was conducted to examine processing differences between the modal verbs *should* and *must*. Unambiguous propositions were either agreed with or disagreed with, and participants' eye movements were monitored as they heard and read the sentence. Reaction times reveal no differences in processing. However, closer time course analyses revealed a divergence in fixations to the target for *should*. These results suggest two mental models are simultaneously activated, entailing both agreement and disagreement with the statement in question.

Keywords: Language understanding, semantics, modal auxiliaries, eyetracking

Introduction

Modal verbs are a central part of modality, the part of language that broadly concerns the factual status of affairs (Frawley, 1992). These verbs are often used to express how the world should be (*deontic* modality) or how the world must be because of some inference (*epistemic* modality) (Sweetser, 1990; Nuyts, 2006). The primary semantic function of modals is to deintensify the meaning of what is stated, by indicating some degree of uncertainty (Close, 1975). Humans are uniquely equipped to think and communicate states of affairs that are ideal, that have never been witnessed first-hand, or that can only exist if some past action were different. For instance, the epistemic use of the modal *must,* as in "You must be bored" indicates an inference (e.g., somebody is nodding off or appears inattentive), while its deontic use, as in "You must pay your taxes" indicates an obligation to perform a series of actions that are ideal according to customs and laws. This initial line of research focuses on the perception of deontic modality. Spoken language is processed incrementally (Marslen-Wilson, 1975), and thus the parallel processing and certainty differences of *should* and *must* may be accessible when temporally sensitive metrics are used.

Modals are of interest to developmental psychologists who investigate reasoning. In one study, children predicted the outcome of a scenario in which a character used a phrase such as "You can't go outside" directed toward a fictitious character (Byrnes & Duff, 1989). Statements including *might*, *have to*, *can*, and *can't*, were used. The results revealed that children ages 3-5 are able to combine negation (*can't*) and modal auxiliary intensity to infer future actions in a story, specifically performing more accurately with *have to,* and less accurately with *might*. Children often responded with the opposite of the correct prediction with a negated modal (i.e. "You can't go outside" was often interpreted as "Johnny went outside"), indicating some early competition in affirmative and negated reasoning. Even at an early age, children know how to make concrete, dichotomous decisions when modal verbs are used. (See also Bliss, 1988; Green, 1979; Hirst & Weil, 1982; Moore, Bryant & Furrow, 1989; Moore, Pure & Furrow, 1990.)

Adults also reason proficiently, in some cases facilitated by the use of modal verbs. In a Wason card selection task, participants made more errors with traditional if-then logic trials, when they were phrased as "if *p* then *q*" compared to those phrased with the deontic modal *must*, as in "if *p* then must *q"* (Cheng & Holyoak, 1985; see also Cosmides, 1989 and Girotto, Light, & Colbourn, 1988). These studies are offered as evidence for pragmatic constraints influencing decisions more than syntactic constraints. Thus, it is crucial to use natural stimuli one may hear in everyday conversation. This result also points to an amplification of certainty when must is used as compared with non-modal sentences.

These studies are informative because they provide insights into accuracy, but there remains a large gap between these and recent findings in decision-making and incremental, online speech perception. McKinstry, Dale & Spivey (2008) used computer-mouse tracking to study people's responses to phrases varying in degree of truth. They found that as uncertainty increases in a statement, deviations from a straight computer-mouse trajectory show up as a spatial attraction toward the competing response. For example, "You should brush your teeth everyday", rated as being completely true, showed a direct path to the "yes" response (with little curvature toward the "no" response), whereas the ambiguous phrase "Murder is sometimes justifiable" elicited substantially curved computer-mouse trajectories. Thus, making a decision about truth appears to be a process that unfolds over time, and is not made in terms of absolutes. Measurements of an ongoing response that begins while stimulus comprehension is still in progress can thus reveal subtleties in processing and brief considerations of alternative responses, which are not revealed in identification data (see Magnuson, 2005).

Some have referred to modal verb use as being indicative of degree of certainty, where deontic modal uses range from uncertain to certain (Close, 1975, p. 273). However, this hypothesis has not been tested perceptually, and it is unknown when or how this uncertainty is dealt with when a discrete response is required. Eyetracking can reveal probabilistic activations of two or more possible responses

unfolding over time, as eye fixations are closely time-locked to speech input. Studies that use this method have yielded much evidence for the idea that speech perception is incrementally processed (Allopenna, Magnuson & Tanenhaus, 1998; Altmann & Kamide, 1999; Spivey, Tanenhaus, Eberhard & Sedivy, 2002; Knoeferle, Crocker, Scheepers & Pickering, 2005), and that context is integrated continuously as a statement is unfolding (Sedivy, Tanenhaus, Chambers & Carlson, 1999). Thus, if modal verbs are simply contextual cues as to the intended strength of certainty, the process of coming to agreement or disagreement with a modal phrase should be quantifiable in terms of proportions of fixations over time.

Given that modal verbs presumably mitigate the truth of statements (Frawley, 1992), and that decisions about the truth appear to be probabilistically weighted according to degree of truth (McKinstry et al., 2008) it is reasonable to predict that different modal verbs will exert different forces on agreement dynamics. We predicted that *should* and *must* exert their influences differently, allowing for a temporary phase of considering an incorrect response. That is, the modal verb *should* may more readily allow the parallel consideration of both agreeable and disagreeable mental models of a given statement (or perceptual simulations of the possible states of affairs). If this uncertainty follows the Close (1975) scale, *should* will trigger more eye movements to the competing response than *must* because *should* is marking greater uncertainty.

The process of agreeing with a statement in an abstract sense may be made in terms of committing to reality, while simultaneously considering alternate, possible states of reality. We propose that predicates with modal verbs are processed as contingent statements, which are gradually resolved into one of the mental models, as they are being perceived.

The alternative hypothesis neglects interstitial stages of processing, and would be consistent with traditional linguistic definitions of modality. These grades of certainty hint at a probabilistic output, but make no predictions of competition being an intermediary mechanism in this decision-making process. Specifically, this would appear as a weakening or strengthening of looks to a target, with random looks to a competitor. This would be in line with certainty being a component, but the lack of competition would also predict reaction times to *must* statements would be faster than those to *should*. A lack of differences in looks to the competing response would also indicate a decision in terms of absolutes. In this case, the outcome could be modeled with traditional logical operations, which incorporate some degree of truth and/or certainty.

Methods

Participants. A total of 39 participants were run, 20 in the Should condition, and 19 in the Must condition. Each was right-handed or ambidextrous, native English speaking, and had normal or corrected to normal vision. Participants received course credit for participation. The study adhered to the university's IRB standards, and included a debriefing session.

Stimuli. Thirty-one sentences were created, once using the modal *should* and then replicated for *must*, yielding 62 sentences. All were written in the second person (directed at the reader), for instance, "You must/should brush your teeth everyday" where 15 of each set were unambiguously agreeable, and 16 were unambiguously disagreeable as in "You must/should eat from a dirty plate". All stimuli are available from the first author upon request. A male speaker recorded these sentences in a quiet room multiple times, and each recording was selected on the basis of being similar and least variant in pitch, loudness and timbre. Silences were cut off the beginning and end of the recording at zero crossings to avoid audible pops, so there was no silence at the beginning or end of each file. All files were amplitude normalized. Final wav files were sampled at 44kHz in stereo sound, to be presented binaurally over headphones.

Text versions of the stimuli were rated by 100 undergraduate students at UC Merced. Survey participants were excluded from eligibility to participate in the main experiment. Each rated the sentences on a Likert scale of 1 to 7 on agreement-disagreement. For Disagree statements, 7 was the most disagreement, and 1 was the least disagreement, and for Agreement statements 7 was most agreement, and 1, least. All stimuli were on the high end of the scale for *must* (Agree: M=6.1, SD=1.1; Disagree: M=6.1, SD=1.7) and *should* (Agree: M=6.1, SD=0.9; Disagree: M=5.7, SD=1.5). Differences between Agree and Disagree were insignificant for each condition ($p>.1$).

Procedure. A screen displayed Agree or Disagree response boxes on the right and left side of the screen, with stimulus text presented in the middle approximately 200 pixels below the choices. In a drift correction before each trial, participants were required to press the spacebar on a keyboard while looking at a marker on the screen. This marker also functioned as a fixation point, to control eye position at the beginning of each trial. The left and right response boxes were 250x250 pixels, and contained the word "Agree" or "Disagree," with their relative positions randomized across trials. Simultaneous with the onset of the written sentence, a matching auditory file was played. The spoken instructions that were given to the participants in advance of the experiment informed them about the kinds of sentences they would encounter, and asked them to respond with either the left or right arrow key to indicate which side of the screen contained their chosen Agree/Disagree response. Modal verb (*should*/*must*) was a between subjects variable to disguise the key manipulation of the study.

Participants wore a head mounted Eyelink II eyetracker, sampling at 500 Hz. In addition to eyemovements, the response, reaction time, and side of screen were collected.

Results

Responses were extremely accurate, with an average of 30 correct responses (SD: 2.6) and with 14 of the total 39

participants responding accurately to all stimuli. The Should and Must accuracy averages were similar, with Should at about 95% correct (*SD*: 1%) and Must at about 96% correct (*SD*: 5%). Incorrect responses are excluded from all analyses.

The final response occurred at an average of 2534ms (*SD*: 799ms). Reaction times by condition are presented with standard deviations in parentheses here in Table 1.

Table 1: Reaction times by condition

	Positive	Negative
Should	2503ms (542ms)	2454ms (565ms)
Must	2645ms (797ms)	2535ms (1178ms)

A 2x2 mixed ANOVA was conducted to explore differences, with Should versus Must as a between-subjects variable, and Positive versus Negative as a within-subjects variable. Should and Must did not differ as well as Positive and Negative showing no difference or any interaction (all tests: p>.1). To further explore if one condition may contain differences, separate paired-samples t-tests were conducted. Both conditions contained no significant difference between Positive and Negative statements (p>.1).

Eyetracking. One participant in the Should condition was excluded from all eyetracking analyses due to experimenter error. Fixations were coded in a binary fashion over time to 1 of 4 areas of the screen: target, competitor, text, or blank regions. Each port was extended 100 pixels in each direction to allow for error in the track. Graphs of the time course for each condition can be found in Figure 1.

Overall proportions of fixations were subjected to independent samples t-tests, to examine differences between looks to the target, competitor, text and blank regions of screen between the two conditions, for both Positive and Negative stimuli. In a comparison of Should and Must for Positive trials only, looks to target, text and blank regions of screen did not reliably differ (p>.1). Looks to the competitor reliably differed (t(36)=2.4, p=.02), with more fixations to the competitor for the Should condition (*M*: 0.097, *SD*: 0.044) than for Must (*M:* 0.064, *SD*: 0.065). A similar trend is seen for Negative trials, with fixations to the target and blank regions no different from one another (p>.1), fixations to the competitor greater in the Should condition (t(36)=2.9, p=.006), and fixations to text marginally significant with more looks to text for Must (t(36)=-1.8, p=.08). Thus, the prominence of the competitor is greater for *should* than for *must* regardless of response type, but not at the cost of looking to the target in particular.

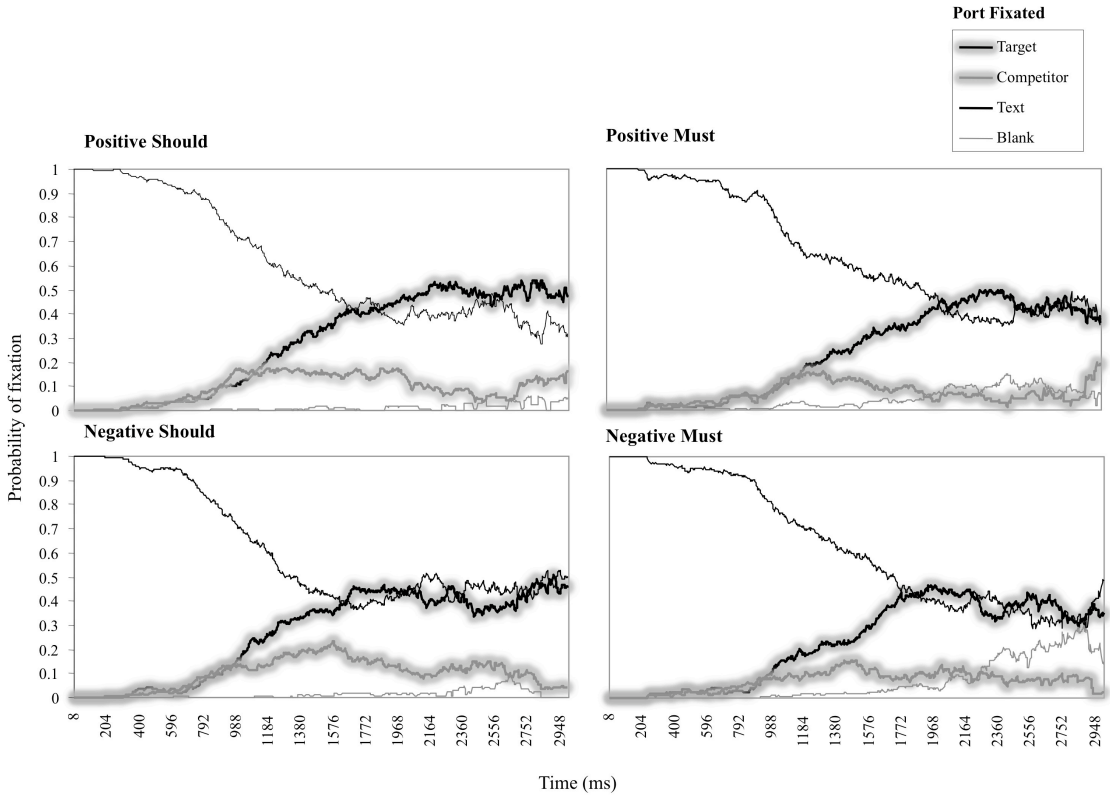


Figure 1: Timecourse by condition – Each panel represents one of four conditions, where Must and Should are between-subject conditions, and Negative (where the target is Disagreement) versus Positive (target Agreement) stimuli. Each line represents the probability of fixating on that particular region of the screen at that moment in time.

To test whether looks to the competitor in the Must condition were negligible or not, a one-sample t-test against a value of zero was performed. Positive Must was significant for the competitor against a value of zero ($t(18)=7.61$, $p<.001$) as was Negative Must ($t(18)=9.68$, $p<.001$). This analysis shows that there are still some looks to the competitor and thus, that processing statements with *must* is not likely to be an all or nothing process.

Because these tests examine positive and negative stimuli separately, it can be useful to look at the differences between Positive and Negative trial fixations over time. There may be a window of time during which there are more fixations to a certain region when Positive stimuli are heard, and another window during which a region is more prominent for a Negative stimulus. By pooling all time in previous analyses, there is a risk of averaging and washing out these differences. Graphically we can examine this by subtracting fixations in Negative trials from the Positive for each time step. Thus, the y-axis becomes a kind of valence scale, where 0 indicates no difference between looks for Positive and Negative stimuli. Positive y means a bias toward that particular port in Positive trials, and negative y is bias during Negative trials. Figure 2 shows that in the Should condition, between 1244ms and 2200ms there is a bias toward looking at the Negative Target, and a bias during this same time window toward looking at Positive Text. These differences were also mapped out for the Must condition, but were uninformative and will not be investigated further here.

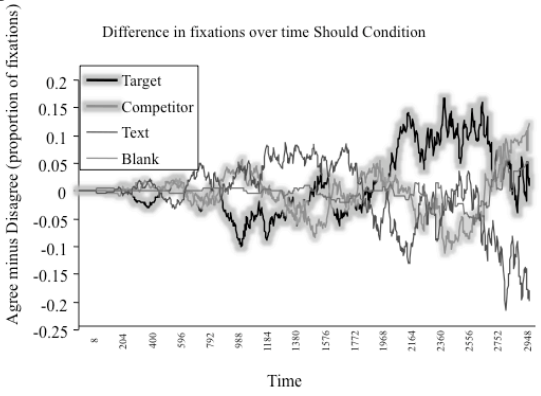


Figure 2: Difference in average fixations over time – This graph is constructed by subtracting average fixations on Disagree trials from Agree trials. Divergence from no difference was observed at 900ms through 1799ms for Text and Target areas of the screen.

Time was binned into Beginning (0-899ms), Middle (900-1799ms) and Final (1800ms-3000ms). These time bins were analyzed in a similar way to the previous fixation analyses, where average proportions of fixations to each port were output for each subject in each condition, and then compared across the Positive / Negative dimension in a paired samples t-test. As predicted, there were no differences in looks to any regions of the screen in the Beginning time window or the Final window ($p>.1$). Fixations to the Target in the Middle time window were greater for the Negative stimuli (Positive: $M=0.28$, $SD=0.11$; Negative: $M=0.322$, $SD=0.14$; $t(18)=2.17$, $p=.043$). Other ports revealed no differences ($p>.1$).

Discussion

For the Should condition, the increased looks to the competitor seemed to indicate that the competing response is more active throughout the trial than in the Must condition. A lack of differences in looks to the target where the competitor is prominent indicates that this increase is not competition, but rather an increase that comes at the cost of no particular region of the screen. Although there are very few looks to the competitor in the Must condition, a test against a value of 0 revealed that there are still some occasional looks here. However, because of the random assignment of side of response, we cannot rule out the possibility that this was due to scanning for the location of the correct response, instead of a brief consideration of responding in this way. Reaction times were uninformative, but further investigation of the middle of the trial revealed that the correct target Disagree is fixated on more, showing an earlier rise in looks to the target on Negative trials in the Should condition. For example, when the stimulus was "You should eat from a dirty plate", the participant fixated on the correct Disagree port earlier than with a Positive stimulus such as "You should study hard for exams".

Previous work has made bold claims for a domain-specific, innate, deontic-reasoning module (Cummins, 1996; see also Gigerenzer & Hug, 1992). Also, while not making direct comparisons between different modal verbs, Traxler, Sanford, Aked & Moxey (1997) indicate that one of their assumptions is that all modal verbs carry a possibility marker. As it has been shown here though, deontic reasoning takes place as the speech is being heard, as revealed by eye fixations. Because this decision is not solely driven by auditory information, but must rely on visual input and rapid sharing with areas that drive eye movements, it is unlikely that deontic-reasoning is domain-specific. Further, since the decision is being made as semantic and syntactic information are still coming in, it would be difficult to integrate a specific module for deontic reasoning without having continuous access to semantic and pragmatic constraints as this information arrives, which would be in line with constraint-based theories of sentence processing (e.g., Tanenhaus & Trueswell, 1995)

The semantic theory of possible worlds (Hintikka, 1975) fits with these results, where the set of possible worlds could be compared to output nodes in a dynamic neural network. Upon hearing a sentence with the word "should", the set of output nodes would have some activation for semantic features of the proposition, some true (partial activation above some threshold), and some false (below threshold activation). For example "You should eat from a dirty plate" would perhaps simulate a dirty plate, a clean plate (an alternative based on experience), motor plans of eating motions, etc. As time moves forward, although

"dirty plate" is explicitly stated, the activation for the dirty plate node would gradually fade while the clean plate node ramps up. Thus, although it is possible to eat from a dirty plate, and there may be some temporary fleeting time in which this is true in some possible world, we gradually resolve into a set of possible worlds or nodes based on previous experience in our world. While we do not propose any mechanism of how differences in certainty might be learned, it is possible to create a kind of possible worlds network with a simulation of simple visual, auditory, sensory feature nodes in combination with a simple recurrent word learning network (Anderson, Huette, Matlock, & Spivey, in press; Howell, Jankowicz & Becker, 2005).

These results reveal that agreement with obvious, unambiguous, everyday standards of conduct can be strongly affected by the choice of modal verb. Highly unambiguous sentence processing unfolds over time, and is susceptible to the influence of subtle variation in verbal modality. It is likely that these intermediate stages of processing are crucial to how one perceives, and makes decisions when these modal auxiliaries are used in everyday discourse, and an integration of compatible linguistic and psycholinguistic theories will be necessary for resolving further debates on the nature of reasoning where modal verbs are involved.

References

Allopenna, P.D., Magnuson, J.S., & Tanenhaus, M. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language, 38(4),* 419-439.

Altmann, G., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition, 73,* 247-264.

Anderson, S.E., Huette, S., Matlock, T., & Spivey, M.J.(in press). On the temporal dynamics of negated perceptual simulations. In F. Parrill, V. Tobin, & M. Turner (Eds.), *Meaning, form, & body*, (pp. 1-20). Stanford, CSLI.

Bliss, L.S. (1988). Modal usage by pre-school children. *Journal of Applied Developmental Psychology, 9,* 253-261.

Byrnes, J.P., & Duff, M.A. (1989). Young children's comprehension of modal expressions. *Cognitive Development, 4(4),* 369-387.

Cheng, P., & Holyoak, K.J. (1985). Pragmatic reasoning schemas. *Cognitive Psychology, 17,* 391-416.

Close, R.A. (1975). A reference grammar for students of English. Essex: Longman Group.

Cosmides, L. (1989). The logic of social exchange: Has natural selection shaped how humans reason? Studies with the Wason selection task. *Cognition, 31(3),* 187-276.

Cummins, D.D. (1996). Evidence for the innateness of deontic reasoning. *Mind & Language, 11(2),* 160-190.

Divers, J. (2002). Possible Worlds. London: Routledge.

Frawley, W. (1992). *Linguistic Semantics.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Gigerenzer, G., & Hug, K. (1992). Domain-specific reasoning: Social contracts, cheating and perspective change. *Cognition, 43,* 127-171.

Girotto, V., Light, P. H., & Colbourn, C. J. (1988). Pragmatic schemas and conditional reasoning in children. *Quarterly Journal of Experimental Psychology, 40A*, 342-357.

Green, M. (1979). The Developmental Relation between Cognitive Stage and the Comprehension of Speaker Uncertainty. *Child Development, 50(3)*, 666-674.

Halpern, J.Y., & Moses, Y. (1992). A guide to completeness and complexity for modal logics of knowledge and belief. *Artificial Intelligence, 54,* 319-379.

Hintikka, J. (1975). Impossible possible worlds vindicated. *Journal of Philosophical Logic, 4(3)*, 475-484.

Hirst, W., & Weil, J. (1982). Acquisition of epi- stemic and deontic meaning of modals. *Journal of Child Language, 9,* 659-666.

Howell, S. R., Jankowicz, D., & Becker, S. 2005. A Model of Grounded Language Acquisition: Sensorimotor Features Improve Lexical and Grammatical Learning. *Journal of Memory and Language, 53*, 258-276.

Knoeferle, P., Crocker, M.W., Scheepers, C., & Pickering, M.J. (2005). The influence of the immediate visual context on incremental thematic role-assignment: Evidence from eye-movements in depicted events. *Cognition, 95(1),* 95-127.

Magnuson, J. (2005). Moving hand reveals dynamics of thought. *Proceedings of the National Academy of Sciences of the USA, 102(29),* 9995-9996.

Marslen-Wilson, W.D. (1975). Sentence perception as an interactive parallel process. *Science, 189(4198),* 226-228.

McKinstry, C., Dale, R., & Spivey, M.J. (2008). Action dynamics reveal parallel competition in decision making. *Psychological Science, 19(1)*, 22-24.

Moore, C., Bryant, D., & Furrow, D. (1989). Mental Terms and the Development of Certainty. *Child Development, 60(1)*, 167-171.

Moore, C., Pure, K., & Furrow, D. (1990). Children's understanding of the modal expression of certainty and uncertainty and its relation to the development of a representational theory of mind. *Child Development, 61,* 722–730.

Nuyts, J. (2006). Modality: Overview and linguistic issues. In W. Frawley (Eds.), *The expression of modality,* (pp. 1-26). Berlin, Mouton de Gruyter.

Sedivy, J.E., Tanenhaus, M.K., Chambers, C.G. & Carlson, G.N. (1999). Achieving incremental interpretation through contextual representation: Evidence from the processing of adjectives. *Cognition, 71*, 109-147.

Spivey, M.J. Tanenhaus, M.K., Eberhard, K.M. & Sedivy, J.C. (2002). Eye movements and spoken language comprehension: Effects of visual context on syntactic ambiguity resolution. *Cognitive Psychology, 45*, 447-481.

Sweetser, E. (1990). *From etymology to pragmatics: Metaphorical and cultural aspects of semantic structure.* Cambridge: Cambridge University Press.

Tanenhaus, M. K. & Trueswell, J. C. (1995). Sentence comprehension. In J. Miller & P. Eimas (Eds.), *Handbook of Perception and Cognition.* (pp.217-262). San Diego: Acamdic Press.

Traxler, M.J., Sanford, A.J., Aked, J.P., & Moxey, L.M. (1997). Processing causal and diagnostic statements in discourse. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23(1),* 88-101.

INFLUENCE OF GRAMMATICAL GENDER ON DEDUCTIVE REASONING ABOUT SEX-SPECIFIC PROPERTIES OF ANIMALS

Mutsumi Imai (imai@sfc.keio.ac.jp)
Faculty of Environment and Information Studies, Keio University at Shonan-Fujisawa, Japan

Lennart Schalk (schalk@ifv.gess.ethz.ch)
Institute for Behavioral Sciences, ETH Zurich, Switzerland

Henrik Saalbach (saalbach@ifv.gess.ethz.ch)
Institute for Behavioral Sciences, ETH Zurich, Switzerland

Hiroyuki Okada (h.okada@eng.tamagawa.ac.jp)
Department of Engineering, Tamagawa University, Japan

Abstract

Grammatical gender is independent of biological sex for the majority of animal names (e.g., a male giraffe is grammatically treated as feminine). However, there is apparent semantic *motivation* for grammatical gender classes, especially in mapping human terms to gender classes. This research investigated whether this apparent motivation in mapping between grammar and biological sex affects deductive inference in German speakers. We identified two contexts in which speakers unconsciously over-generalize the grammar-semantics mapping to make inappropriate deductive inferences about sex-specific biological properties. They tended to erroneously accept deductions when the sex in the premise and the grammatical gender of the target animal agreed. The sex-gender agreement affected the inference even when the sex of the target was explicitly indicated (e.g., die$_{[FEM]}$ *männliche (male)* Giraffe). Experiment 2 further suggested that these effects occur only when the gender-marking article accompanied the noun. Implications of the results for lintuistic relativity is discussed.

Keywords: Add your choice of indexing terms or keywords; kindly use a semi-colon; between each term.

Introduction

Many languages of the world have a system of grammatical gender, where nouns are assigned to one of the limited number of gender classes (Corbett & Fraser, 2000). Unlike languages that mark gender only semantically (e.g., English), languages with grammatical gender assign gender to all nouns regardless of whether or not referents have a biological sex. The link between gender assignment and conceptual properties of non-human referents has widely been said to be arbitrary (Aikhenvald, 2000; Fox, 1990), as grammatical gender is not relevant to biological sex for a majority of words. For example, in German, the word *giraffe* is grammatically feminine and *elephant* is masculine, but it is not the case that all giraffes are female or that all elephants are male. Nonetheless, the feminine article *die* ($_{[FEM]}$) must be applied when one refers to a grammatically feminine noun and the feminine pronoun *sie* must be used as an anaphoric reference, whether the referent is biologically female or male (e.g., *die männliche* (male) *Giraffe*). Here, an interesting question is to what extent speakers are able to separate the biological sex of an animal from its grammatical gender. From the perspective of a speaker of language without grammatical gender, it appears confusing that one has to use the feminine article and the female pronoun even when the giraffe is actually male. Of course, speakers of a language with the grammatical gender system must know that grammatical gender does not directly reflect biological sex. However, are the speakers completely immune to the influence of grammatical gender when they draw inferences about the animal's sex-specific biological properties? It is possible that the few cases of semantic correspondence between grammatical gender and biological sex may have resulted in an overgeneralization during the process of language acquisition. In German, for example, salient female terms such as *woman*, *lady*, *mother* are grammatically feminine, while salient male terms such as *man*, *boy*, *father*, are grammatically masculine (*Natural Sex Principle*, cf. Zubin & Koepcke, 1986).

Thus, speakers may falsely generalize this *exceptional* mapping between gender class and biological sex to words for animated entities in general. This assumption is consistent with Vigliocco and colleagues' (Vigliocco, Vinson, Paganelli, & Dworzynski, 2005) *sex-and-gender-hypothesis*, which proposes that a conceptual influence of grammatical gender originates in speakers' first noticing the correspondence between grammatical classes and corresponding conceptual classes. In other words, acknowledging the link between biological sex and the grammatical gender class in the case of some salient human-specific terms leads speakers to develop a general anticipation that even non-human animals from the same grammatical gender class are more similar to one another than animals from different grammatical gender classes.

Most of the previous research has asked whether and to what extent grammatical gender influences speakers' concepts of entities in terms of typically feminine/masculine attributes assigned to those entities. Konishi (1993) looked at how Spanish and German speakers construe femininity or masculinity of non-animal objects by having them give gender-related ratings of various nouns on a potency scale

(e.g., weak vs. strong; tender vs. vigorous): German speakers rated *moon* (grammatically masculine in German and feminine in Spanish) to be higher on the scale of masculinity than *sun* (masculine in Spanish and feminine in German), while Spanish speakers showed the reverse pattern. Sera and colleagues (Sera, Berge, & del Castillo Pintado, 1994; Sera, Elieff, Forbes, Burch, Rodriguez, & Dubois, 2002) asked Spanish and French speakers to assign either a female or a male voice to artifact objects and reported that the judgments tended to agree with the grammatical gender of the objects (see also Boroditsky, Schmidt, & Phillips, 2003; Flaherty, 2001). However, to our knowledge, the question of whether speakers of a language with grammatical gender are at all influenced by grammatical gender when they make inferences about biological sex-specific properties of animals has not been addressed in the literature.

Importantly, Vigliocco et al. (2005) suggested that the relation between grammatical gender and speakers' concepts is weaker for languages with more than two gender classes, such as German. Using an odd-one-out categorization task, they in fact found an effect of grammatical gender on Italian speakers' construal of similarity among animals, but not on German speakers'. However, unlike judgments of abstract similarity among objects, inference about biological sex-specific properties is more directly linked to grammatical gender categories, and hence we might expect the influence of grammatical gender in German speakers in this case.

It is hard to imagine that German speakers are not aware of the motivated link between grammatical gender and biological sex, as human males are clearly mapped to the masculine gender and human females are mapped to the feminine gender. Yet, when thinking about animals at the level of generic species (dog, cat, giraffe, etc.), speakers have to separate grammatical gender and biological sex. Of course, adults speaking a language with grammatical gender must *consciously* understand that grammatical gender of basic-level animal names is independent of animals' biological sex. However, it may still be possible that their inference is still affected by the overgeneralization of the syntax-semantics mappings: For example, they may make a false deductive conclusion that grammatically feminine (or masculine) animals *in general* have a female (or male)-specific biological property.

Deductive reasoning plays a core role in human inference and learning, along with inductive reasoning (cf. Murphy, 2002). If grammatical gender affects deductive reasoning about biological properties even though people consciously understand that grammatical gender is independent of biological sex of animals, this will be taken as support for linguistic relativity.

Provided that such an effect is seen, however, it is important to be able to distinguish two possible mechanisms behind it. The effect may arise within the realm of syntactic processing but not at the conceptual representation of animal kinds. In other words, the effect may be seen only when a speaker processes the gender-marking article or pronoun. The alternative possibility is that the over-generalized syntax-semantic mapping penetrates into the conceptual level of generic-level animal kinds. If this is the case, the effect should be seen even when generic-level animal names are presented without the gender-marking article.

The present study

We tested German and Japanese speakers on deductive inferences about sex-specific animal properties. The Japanese speakers' performance served as a baseline because Japanese is a language without grammatical gender. We designed two experiments in such a way that we could identify at what level of processing the relation between grammatical gender and deductive reasoning is found, if it is found at all. In the first experiment, target words for deduction were presented in the singular form with their associated articles marking the gender class of each word. (In German, article + noun phrase can refer to a generic meaning.) In the second experiment, the target words were presented in plural form without any marking of gender class. Participants were asked to indicate whether the deductive conclusion would hold true or not; they were instructed to give a "No" response in cases in which the conclusion was logically indeterminable, in addition to the cases in which deduction would be clearly false.

Five conditions were set up within participants. The *Generic Animal Condition* was designed to test whether German speakers were more likely to draw a erroneous deductive conclusion when the sex specified for the biological property given in the premise and the grammatical gender class of the target animal's basic-level name were consistent (e.g., female – feminine) than when they were inconsistent (e.g., female – masculine). Here, the deductive conclusion is logically indeterminable, as the biological sex of the target animal is unknown, and thus, "No" is the correct answer. Nevertheless, German speakers may experience difficulties rejecting the deductive conclusion when the grammatical gender of the target animal agrees with the biological sex specified in the premise. In contrast, it should be easy for Japanese speakers to reject the deduction in this ambiguous case.

In order to test for possible baseline differences in deductive reasoning across the two language groups, we included *the Generic-Animal Control Condition*. Here, participants were to judge the correctness of the deductive conclusion about a property true for all animals regardless of their sex, while the targets were exactly the same as in the *Generic-Animal* condition.

The *Sex-specified Animal Condition* was set up to test whether grammatical gender affects deductive inference in German speakers even when the sex of the animal is explicitly specified in the conclusion. Here, unlike the Generic Animal condition, the target animal's sex was explicitly specified by the gender-specifying adjective and the specified sex and the grammatical gender of the target

animal was either consistent or inconsistent. Here, the deductive inference should of course be made based on the agreement between the sex in the premise and the target animal's sex indicated by the adjective. It is interesting to see if consistency between grammatical gender and sex affects German speakers' judgments in this obvious case.

The *Sex-specified Animal Control Condition* was included to rule out an alternative explanation for the potential gender effect in the *Sex-specified Animal Condition*. Provided that the expected effect was obtained, it may also have arisen from the difference in the difficulty in simply processing of the two types of (i.e., grammatical gender-sex specifying adjective matching and mismatching) noun phrases. To disambiguate the two possibilities, the conclusions in this condition were the same as those in the Sex-specified Animal Condition, but the property in the premise was not sex-specific. Finding the gender effect in German speakers in this control condition would indicate that the effect arises at the level of local phrase processing rather than during the deductive reasoning. In contrast, if there is no gender effect in this control condition, but the effect is found in the Sex-specified Animal condition, where the property in the premise is also sex-specific, this suggests that the grammatical gender affects deductive reasoning about a sex-specific property, even when the target animal's sex is explicitly given.

Finally, the *Artifact Condition* was included to examine whether German speakers' deductive reasoning about non-animate entities was affected by grammatical gender. The target object was an artifact whose grammatical gender was either consistent or inconsistent with the sex specified in the premise. The conclusion was logically determinable and should always be rejected. This condition allows us to see how pervasive the influence of grammatical gender on deductive inference about sex-specific biological properties: If the motivated sex-gender mapping is applied even in the realm of entities without sex, this would suggest that the influence of grammatical gender is overarching in German speakers.

Experiment 1

In this experiment, we tested whether there is a relation between grammatical gender and speakers' deductive reasoning about a sex-specific biological property when the grammatical gender of the target object was explicitly invoked by the gender-marking article.

Method

Participants

Twenty-one native German-speaking undergraduates from Zurich and 17 native Japanese-speaking undergraduates from Tokyo, both from a wide variety of majors, participated for payment.

Design and Materials

As described earlier, there were five within-subjects conditions: *Generic Animal*, *Generic Animal Control*, *Sex-specified Animal*, *Sex-specified Animal Control*, and *Artifact*. In each trail across the five conditions, the premise sentence containing a blank property X was shown, and followed by the target object. In the Generic Animal, Sex-specified Animal, and Artifact conditions, the premise stated that the property X was sex-specific. It said: "All and only male (or female) animal had X inside." In the two Control conditions, the premise statement was sex-general: "All and only animals had X inside." Prior to the experiment, the participants were told that X was an internal and important property.

In the *Generic Animal Condition*, 36 generic level animal names (half grammatically feminine, half masculine in German) that were commonly known to speakers of both languages, were used as targets. Each animal appeared once in the sex-gender consistent trials and once in the inconsistent trials, yielding a total of 72 trials in this condition. As described earlier, the correct response was "No" for all trials, as the deduction was not logically determinable. The same 36 animal names were used in the *Generic Animal Control Condition*, in which the property given in the premise sentence was general to all animals. Here, of course, the correct response was "Yes" for all trials.

In the *Sex-specified Animal Condition*, 18 animal names (half grammatically feminine, half masculine) that were not used in the Generic Animal Condition were presented twice, once in a consistent and once in an inconsistent trial. Here, the sex specified in the premise and the grammatical gender of the target animal always matched, but the specified sex and the grammatical gender of the target animal was either *consistent* ("*die*$_{[FEM]}$ *weibliche* (female) *Maus* (mouse)") or *inconsistent* ("*die männliche* (male) *Mous")* for the "all and only *female* animals have X inside" premise). The same targets were used for *the Sex-specified Animal Control Condition*, but here, the property in the premise was not sex-specific (e.g., "all and only animals has X inside").

In the *Artifact Condition*, the premise concerned a sex-specific animal property, as in the other two main conditions, but 28 artifact names (half grammatically feminine, half masculine) served as targets. All artifact names appeared once in a sex-gender consistent and once in an inconsistent trial. The "No" response was correct for all trials.

Altogether, , there were 208 trials including 90 trials with potential "Yes" responses and 118 trials with potential "No" responses.

Procedure

In each trial, a fixation cross appeared on the screen for one second. The premise statement was then shown for 1.5 seconds, followed by a blank screen for 0.5 seconds. For German participants, the name of the target object accompanied by the gender article was then presented until the participant made a response. For Japanese participants, the target object name was presented alone, without a classifier, as this was judged to be the most natural way of

presentation The participants were asked to indicate whether the deductive conclusion would hold true for the target by pressing a designated key for "Yes" or "No". After the response, the screen remained blank for 1.5 seconds and the next trial was then started. The presentation order of the 208 trials of all conditions was completely randomized within and across participants.

Results and Discussion

We report the results separately for each condition.

Generic Animal Condition Here, we only analyze the error responses (i.e., Yes responses, see Figure 1). Response times were not submitted to the analysis because of the high error rates in German speakers. As expected, there was a significant Language (German vs. Japanese) X Consistency (sex-gender consistent vs. inconsistent) interaction effect, $F_1(1,31)=9.1$, $F_2=(1,90)=98.8$, both $p<.01$). Paired t-tests were performed on subject (t_1) and item means (t_2) contrasting the performance in consistent and inconsistent trials across the different conditions. German speakers were more likely to erroneously accept a deductive conclusion when the sex in the premise and the grammatical gender of the target were consistent (53.4%) than when they were inconsistent (29.9%), $t_1(16) = 3.133$, $d = .626$, $p = .006$, $t_2(35) = 13.447$, $d = 2.898$, $p < .000$. No such difference was found in Japanese participants (17.0% vs. 17.2%), $t_1(15) = -.102$, $p = .920$, $t_2(35) = -.166$, $p = .869$. However, the performance in German speakers in the Control condition showed that they were in general no poorer in deductive reasoning than Japanese speakers (German:92.5%; Japanese:83.5%) , $t_1(31) = 1.821$, $p = .078$; $t_2(70) = 4.597$, $d = .969$, $p < .000$. These results suggest that the grammatical gender effect seen in the Generic Animal condition was not a reflection of generally poor deductive inference on the part of German speakers.

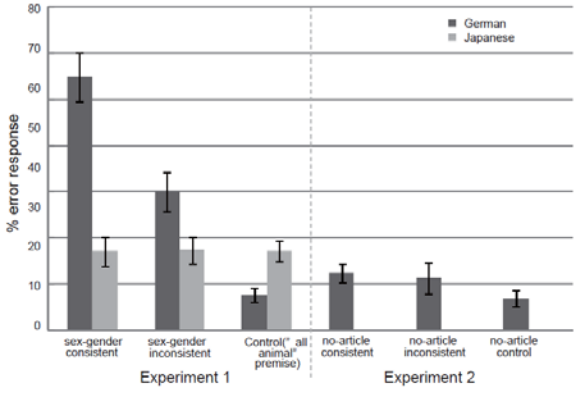


Figure 1. Percentages of error responses in the *Generic Animal Condition* (with sex specific premises) of Experiments 1 and 2 and in the *Generic Animal Control Condition* (with sex-general premises) in Experiment 1.

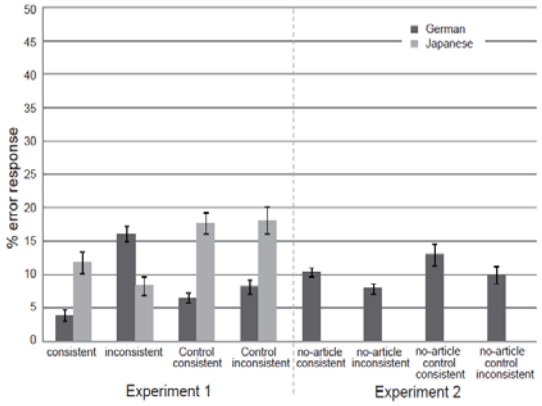


Figure 2. Percentages of error responses in the *Sex-specified Animal Condition* (with sex-specific premises) and *the Sex-specified Animal Control Condition* (with sex-general premises) in Experiments 1 and 2.

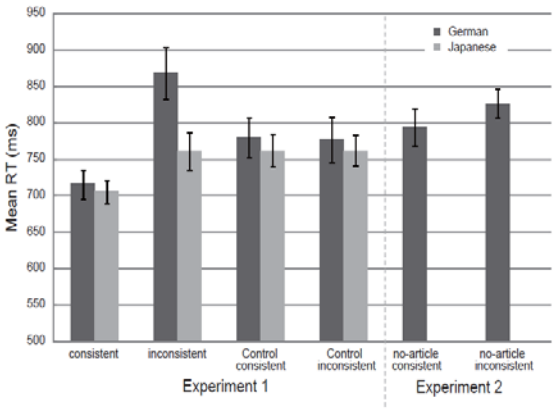


Figure 3. Response times (in milliseconds) for correct responses in the *Sex-specified Animal Condition* (with sex-specific premises) in Experiment 1 and the *Sex-specified Animal Control Condition* (with sex-general premises) in Experiments 1 and 2.

Sex-specified Animal Condition In this condition, both the error rates and response times were submitted to analyses. Again, a significant Language X Consistency interaction effect on the error rates was found, $F_1(1,31) = 8.5$, $F_2(1,34) = 8.9$, both $ps<.01$. Unlike the Generic Animal condition, the correct/error deduction was logically determinable according to the agreement or disagreement between the *sex* in the premise and the specified *sex* of the target animal. Here, the inconsistent trials, in which the sex specified by the adjective and grammatical gender in the target were inconsistent (e.g., *die männliche Giraffe)*, were expected to be more difficult than the consistent trials (*die weibliche Giraffe*). Indeed, German speakers were more likely to draw erroneous deductions in the inconsistent trials (16.1%) than in the consistent trials (3.9%), $t_1(16) = 2.917$, $d = .878$, p

= .010, $t_2(17) = 2.735$, $d = .766$, $p = .014$ (Figure 2). No such difference was found in Japanese participants (11.8% vs. 8.3 %), $t_1(15) = -1.274$, $p = .222$, $t_2(17) = -1.514$, $p = .148$. A similar pattern was found for response times. German speakers were slower in drawing deductive inferences in the inconsistent case (868ms) than in the consistent case (716ms), $t_1(16) = 3.442$, $d = .574$, $p = .003$, $t_2(17) = 4.986$, $d = 1.522$, $p < .000$, while no such difference was found in Japanese responses (706ms and 761ms), $t_1(15) = 1.342$, $p = .199$, $t_2(17) = 1.969$, $d = .559$, $p = .065$ (Figure 3). In the Control condition, there was no Language X Consistency effect on either accuracy, F_1 (1,31) = 0.3, F_2(1,34) = 1, or response times, F_1(1, 31) = 0.01 , F_2(1,34) = 0.6.

Thus, even when the biological sex of an animal was explicitly indicated, grammatical gender affected German speakers' inferences about sex-specific animal properties. The fact that German speakers' performance did not differ from that of Japanese speakers in the Control condition (where the premise was not sex-specific) indicates that the sex-gender consistency effect here emerged in the process of deductive reasoning rather than from mere disturbance of the local level processing of the target phrase due to gender-sex mismatch.

Artifact Condition In the *Artifact Condition*, no Language X Consistency effect was found, F_1 (1,31)=1.2, F_2(1,26)=0.9, both *p*s>.1. In neither language group did sex-gender consistent and inconsistent trials differ with respect to the error rates (German: 3.8% vs. 0.4%; Japanese: 1.3% vs. 0.4%) nor response times (German: 729ms vs. 727ms; Japanese: 601ms vs. 623ms). Thus, the influence of grammatical gender on sex-specific biological properties found in the animal domain did not extend to the artifact domain.

The results of Experiment 1 showed that German speakers were not immune to the motivated (but logically orthogonal) gender-sex mapping when they make deductive inferences about sex-specific properties of animals. When the biological sex specified in the premise agreed with grammatical gender of the target animal, they often made a false deduction that a sex-specific biological property holds for the target animal in general even though its biological sex was unspecified. German speakers experienced difficulty in rejecting the deductive conclusion even when the target animal's sex was explicitly indicated otherwise by a sex-specifying adjective, when the biological sex specified for the property and grammatical gender of the target animal agreed.

These results naturally lead to a question of whether the same effects are obtained when the target animal name is presented without the article. If German speakers' representation of animals per se is affected by grammatical gender, the same effects should be observed without explicit invocation of the article. Alternatively, the gender effects in Experiment 1 may vanish when the animal name is presented without the gender article. If so, this would indicate that the gender effect arises at the level of grammatical processing, but not at the level of the representation of animals. Experiment 2 was conducted to disambiguate these two possibilities.

Experiment 2

Method

Participants

Twenty-nine German-speaking undergraduates from Zurich participated in this study. None of them had participated in Experiment 1.

Design, Materials, and Procedure

The design, materials and procedure of Experiment 2 were identical to those in Experiment 1 with one exception: All target words were presented in plural form without articles marking grammatical gender. In the *Generic Animal Condition*, for example, the target "die $_{[FEM]}$ Maus (mouse)" was now presented as Mäuse (mice) and in the Sex-specified Animal Condition, "die männliche (male) Maus" was now presented as "männliche Mäuse".

Results

In stark contrast to Experiment 1, we found no significant difference between the gender-sex consistent and inconsistent trials in any of the conditions on the error rates or response times (for t_1 and t_2: all *p*s > .1; see Figures 1-3). When the performance of German speakers in this experiment was compared to that of Japanese speakers in Experiment 1, in no condition (including the Generic Animal and Sex-Specified Animal conditions) was there any Language X Consistency effect.

The results of Experiment 2 indicate that the grammatical gender effects found in Experiment 1 arise only when the speakers see the target animal name with the gender-marking article. This suggests that it was the *gender article* that affected German speakers' deductive reasoning about sex-specific animal properties; the effect did not arise because German speakers' representation of animals per se was changed by gender grammar.

General Discussion

Grammatical gender *in principle* is independent of biological sex, as grammatical gender is assigned to non-sexuated entities as well as to sexuated ones. This is even true for a majority of (basic-level) animal names. At the same time, there is apparent semantic *motivation* for grammatical gender classes, especially in mapping human terms to gender classes. This research investigated whether this mapping between grammar and biological sex is over-generalized in deductive inference--a core domain of human reasoning. We identified two contexts in which German speakers unconsciously over-generalize this grammar-semantics mapping to make erroneous deductive inferences.

First, German speakers tended to erroneously accept deductions when the sex specified in the premise and the grammatical gender of the basic-level name of the target animal agreed. Second, the sex-gender agreement affected the inference even when the sex of the target animal was explicitly indicated: German speakers experienced difficulty in rejecting the deduction when, for example, asked to judge whether a female-specific property would be true for "die$_{[FEM]}$ *männliche* Maus (male mouse)". Experiment 2 further suggests that these effects occur only when the gender-marking article was processed. Thus, German speakers seem to project biological sex onto gender-marking articles but not onto the conceptual representation of animals per se. Furthermore, this mapping does not go so far as to affect inferences when the targets are non-sexuated entities.

Researchers investigating the relation between the speakers' conceptual representation of objects and gender grammar have mostly approached the question in light of *whether* masculine or feminine images were projected on objects according to the grammatical gender of the name. This research examined the relation between gender grammar and cognitive processes more directly, asking how speakers handle the semantic motivation of gender classes on one hand and the fact that grammatical gender is independent of biological sex in animal terms on the other hand. The finding that German speakers could not help projecting biological sex on gender-marking articles (when they should not) can be taken as some evidence for linguistic relativity (Gentner & Goldin-Meadow, 2003, for an overview). On the other hand, our findings cannot be interpreted to be support for a strong version of linguistic relativity hypothesis, as the effect was not obtained without explicit invocation of the grammatical gender. Some researchers may argue that the gender effect here is only support for thinking for speaking (Slobin, 1996) but not for linguistic relativity per se, because the effect was obtained in a task using language (see also Vigliocco et al., 2005). Nevertheless, the influence of grammatical gender we found in this research should not be seen as trivial. For speakers of languages with grammatical gender, explicit gender marking by articles or pronouns is the *norm* rather than the exception in everyday discourse. If these speakers of languages unconsciously link the grammatical gender of an animal's name to its biological sex (even though the two are orthogonal), and if this link is strong enough to serve as a basis for inferences about sex-specific properties of animals, then we may conclude that grammatical gender has non-trivial cognitive consequences for these speakers, be it characterized as a "true" linguistic relativity effect or not. This research is important for the literature of language and thought in that it specifies how (i.e., the mechanism) and in what contexts gender grammar might affect cognitive processes rather than simply providing evidence for linguistic relativity (see also Imai & Saalbach, 2010)

Acknowledgement

This research was supported by Ministry of Education grant-in-aid for Scientific Research awarded to Imai and by DAAD Post-Doc Fellowship to Saalbach.

References

Aikhenvald, A. (2000). *Classifiers: A typology of noun categorization devices.* New York: Oxford University Press.

Boroditsky, L., Schmidt, L. A., & Phillips, W. (2003). Sex, syntax, and semantics. In D. Gentner, & S. Goldin-Meadow (Eds.), *Language in Mind: Advances in the Study of Language and Thought* (pp. 61-80). Cambridge, MA: MIT Press.

Corbett, G. G., & Fraser, M. N. (2000). Gender assignment: A typology and a model. In G. Senft (Ed.), *Systems of Nominal Classification* (pp. 293–325). Cambridge: Cambridge University Press.

Flaherty, M. (2001). How a language gender system creeps into perception. *Journal of Cross-Cultural Psychology*, 32(1), 18-31.

Fox, A. (1990). *The Structure of German*. New York: Oxford University Press.

Gentner, D., & Goldin-Meadow, S. (2003). *Language in mind: Advances in the Study of Language and Thought* (pp. 61-80). Cambridge, MA: MIT Press.

Imai, M., & Saalbach, H. (2010). Categories in mind and categories in language: Do classifier categories influence conceptual structures? In B. Malt (Ed.), *Words and the World*. New York: Oxford Press.

Konishi, T. (1993). The semantics of grammatical gender: A cross-cultural study. *Journal of Psycholinguistic Research*, 22(5), 519-34.

Murphy, G. L. (2002). *The Big Book of Concepts.* Cambridge, MA: MIT Press.

Sera, M. D., Elieff, C., Forbes, J., Burch, M. C., Rodriguez, W., & Dubois, D. P. (2002). When language affects cognition and when does it not: An analysis of grammatical gender and classification. *Journal of Experimental Psychology: General*, 131(3), 377-397.

Slobin, D. I. (1996). From "thought and language" to "thinking for speaking". In J. J. Gumperz, & S. C. Levinson (Eds.), *Rethinking Linguistic Relativity* (pp. 70-96). Cambridge: Cambridge University Press.

Vigliocco, G., Vinson, D. P., Indefrey, P., Levelt, W. J. M., & Hellwig, F. (2004). Role of grammatical gender and semantics in German word production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30(2), 483-497.

Vigliocco, G., Vinson, D. P., Paganelli, F., & Dworzynski, K. (2005). Grammatical gender effects on cognition: Implication for language learning and language use. *Journal of Experimental Psychology: General*, 134(4), 501-520.

Zubin, D., & Köpcke, K.-M. (1986). Gender and folk taxonomy: The indexical relation between grammatical gender and lexical categorization. In C. Craig (Ed.), *Noun classes and categorization* (pp. 139-180). Philadelphia: Benjamins

A Comparison of the Belief-Adjustment Model and the Quantum Inference Model as Explanations of Order Effects in Human Inference

Jennifer S. Trueblood (jstruebl@indiana.edu)
Cognitive Science Program, 819 Eigenmann, 1910 E. 10th St.
Bloomington, IN 47406 USA

Jerome R. Busemeyer (jbusemey@indiana.edu)
The Department of Psychological and Brain Sciences, 1101 E. 10th Street
Bloomington, IN 47405 USA

Abstract

One of the oldest and most reliable findings regarding human inference is that the order of evidence affects the final judgment. These order effects are non-Bayesian by nature and are difficult to explain by classical probability models. We use the empirical results of two jury decision-making experiments to compare two different models of human belief updating: the belief-adjustment model and the quantum inference model. We also provide evidence to suggest the belief-adjustment model has limited predictive power when accounting for tasks involving extreme evidence whereas the quantum inference model does not.

Keywords: inference; jury decision-making; order effects; recency effects; belief-adjustment model; quantum inference model

Introduction

Human inference provides a rich source of evidence for a non-Bayesian belief updating process. Consider a physician deciding whether a certain patient has an infection or not. The physician first examines the patient and takes a medical history. At this point, the physician has some degree of belief in the presence of the infection. Then the physician orders a laboratory test and revises those beliefs. Now, suppose the physician had proceeded by first administering the laboratory test followed by the physical examination. Would the physician ultimately have the same belief about the infection when the order of information is reversed? Bergus, Chapman, Levy, Ely, et Oppliger (1998) would argue that the order of information, physical examine followed by laboratory test versus laboratory test followed by physical examine, has a significant impact on the physician's final belief in the presence of the infection.

Order of information plays a crucial role in the process of updating beliefs across time (Hogarth & Einhorn, 1992). The presence of order effects makes a classical or Bayesian approach to inference difficult. Specifically, suppose a decision-maker must ascertain the probability that a certain hypothesis, H, is true after seeing two pieces of evidence, X and Y. Classical probability requires $P(X,Y|H) = P(Y,X|H)$, and by Bayes rule we must have $P(H|X,Y) = P(H|Y,X)$. Thus, a simple Bayesian model makes no distinction between different orders of information.

In this paper, we compare two possible explanations for order effects, the belief-adjustment model (Hogarth & Einhorn, 1992) and the quantum inference model (Busemeyer & Trueblood, 2009). The belief-adjustment model accounts for order effects by either adding or averaging evidence. The quantum inference model explains order effects by transforming a state vector with different sequences of operators for different orderings of information.[1] We first examine both models with data collected from a jury decision-making experiment conducted by McKenzie, Lee, et Chen (2002). Then we test both models using new data collected from two new experiments that extend the work of McKenzie et al.

A Jury Decision-Making Experiment

McKenzie et al. conducted two experiments to examine the effects of case order and strength on changes in subjects' confidence ratings. In this study, subjects were asked to read a criminal case concerning a burglarized warehouse and to rate their confidence in the defendant's guilt, G. In the first experiment, one group of participants read a strong prosecution, SP, followed by a weak defense, WD. The other group read the information in the reverse order, the weak defense followed by the strong prosecution. For the second experiment, the first condition was identical to the first condition in experiment 1. However, in the second condition, subjects read a weak prosecution, WP, followed by the weak defense. In both experiments, subjects provided confidence ratings as a number between 0 and 20 before reading either case, after reading the first case, and after reading the second case. A separate group of subjects rated the strength of the prosecution and defense and did not participate in the inference task. By averaging the data from condition one of experiment 1 with condition one of experiment 2 and converting the mean confidence ratings to probabilities, we have the results shown in Table 1.

One of the most interesting aspects of these results is that the weak defense increased confidence in guilt when preceded by the strong prosecution but decreased confidence in guilt when preceded by the weak prosecution. The interpretation of the defense as evidence for guilt when coupled with the strong prosecution and evidence for innocence when coupled with the weak prosecution resists explanation by the standard belief-adjustment model (McKenzie et al., 2002).

[1] We use quantum theory as a mathematical tool and do not attach the physical meaning associated with quantum physics. This type of approach is similar to the use of stochastic processes outside the domain of physics.

Table 1: Probability of Guilt from Experiments 1 and 2

After first case	After second case
Pr(G | SP) = 0.672	Pr(G | SP, WD) = 0.719
Pr(G | WD) = 0.51	Pr(G | WD, SP) = 0.75
Pr(G | WP) = 0.600	Pr(G | WP, WD) = 0.525

An extended version of this model, the minimum acceptable strength model (MAS), uses a variable reference point to model these results (McKenzie et al., 2002). As an alternative to the MAS model, the quantum model uses a series of transformations to explain the phenomena. Before we proceed with fitting the two models, we will outline the belief-adjustment model and the MAS model. We will also provide an intuitive description of the quantum model.[2]

The Belief-Adjustment Model

The belief-adjustment model assumes individuals update beliefs by a sequence of anchoring-and-adjustment processes (Hogarth & Einhorn, 1992). The algebraic description of the model is

$$C_k = C_{k-1} + w_k \cdot (s(x_k) - R) \tag{1}$$

where $0 \leq C_k \leq 1$ is the degree of belief in the defendant's guilt after reading case k, $s(x_k)$ is the strength of case k, R is a reference point, and $0 \leq w_k \leq 1$ is an adjustment weight for case k. Hogarth et Einhorn argue that evidence can be encoded either in an absolute manner or in relationship to the current belief in the hypothesis. If evidence is encoded in an absolute manner and there exists a positive/negative relationship between the evidence and hypothesis, $R = 0$ and $-1 \leq s(x_k) \leq 1$. However, if evidence is encoded in relationship to the current belief, $R = C_{k-1}$ and $0 \leq s(x_k) \leq 1$. Also, Hogarth et Einhorn assume that the adjustment weight w_k depends on the level of current belief and the sign of the difference $s(x_k) - R$. Specifically, if $s(x_k) \leq R$, then $w_k = C_{k-1}$. However, if $s(x_k) > R$, then $w_k = 1 - C_{k-1}$.

Using this information, we can rewrite the belief-adjustment model as either an adding model or an averaging model. The adding model results when information is encoded in an absolute manner and is given by

$$C_k = \begin{cases} C_{k-1} + C_{k-1} \cdot s(x_k), & \text{if } s(x_k) \leq 0 \\ C_{k-1} + (1 - C_{k-1}) \cdot s(x_k), & \text{if } s(x_k) > 0 \end{cases}$$

On the other hand, the averaging model results when information is encoded in relationship to the current belief and is given by

$$C_k = \begin{cases} C_{k-1} + C_{k-1} \cdot (s(x_k) - C_{k-1}), & \text{if } s(x_k) \leq C_{k-1} \\ C_{k-1} + (1 - C_{k-1}) \cdot (s(x_k) - C_{k-1}), & \text{if } s(x_k) > C_{k-1} \end{cases}$$

Rearranging the terms above shows that the current belief is an average of the prior belief and the strength of the new evidence weighted by the prior belief.

[2]Trueblood et Busemeyer (2010) contains a complete mathematical description of the quantum inference model.

The MAS model extends the belief-adjustment model by defining the reference point as a case's minimum acceptable strength (McKenzie et al., 2002). Thus, equation 1 becomes

$$C_k = C_{k-1} + w_k \cdot (s(x_k) - m_{k-1}) \tag{2}$$

where m_{k-1} is the minimum acceptable strength of the previous case and $-1 \leq s(x_k) \leq 1$. Neither the adding or averaging models can predict that a defense would increase confidence in guilt. However, it is possible to select a value for m_{k-1} such that the difference between the strength of the weak defense and m_{k-1} is positive. Therefore, confidence in guilt increases as a result of the weak defense. The downside to the MAS model is the increase in parameters. The adding and averaging models specify a parameter for each case, namely $s(x_k)$. However, the MAS model also needs a minimum acceptable strength parameter for each case; thereby, doubling the number of parameters needed in the original model.

The Quantum Inference Model

There are several reasons for considering a quantum approach to human judgments. First, judgment is not a simple read out from a pre-existing or recorded state, instead it is constructed from the current context and question. Thus, making a judgment changes the context which disturbs the cognitive system. This implies that changes in context produced by the first judgment influence the next judgment resulting in order effects. Therefore, human judgments do not obey the commutative rule of classic probability theory suggesting that classical probability theory is too limited to fully explain various aspects of human judgment and decision-making. Other such phenomena include violations of the sure thing axiom of decision-making (Tversky & Shafir, 1992) and violations of the conjunctive and disjunctive rules of classic probability theory (Gilovich, Griffin, & Kahneman, 2002).

We describe the quantum inference model in terms of the specific jury decision-making task outlined above; however, this model can be extended to any number of hypotheses and pieces of evidence (Busemeyer & Trueblood, 2009). The quantum inference model assumes that a decision-maker can view the two complementary hypotheses, guilty (h_1) and not guilt (h_2), from three different points of view. The first point of view is considered neutral (N) and is associated with the judgment made before either case is read. The second point of view is associated with the prosecution's case (P), and the third point of view is associated with the defense's case (D). The prosecution is assumed to present evidence for guilt (e_1), and the defense is assumed to present evidence for innocence (e_2). Considering all possible combinations of hypotheses and evidence, we have four patterns of the form $h_i \wedge e_j$. These four patterns or joint events define a four dimensional vector space. An individual's beliefs about these events are represented as a four dimensional state vector, ψ, situated within this four dimensional vector space. The three points of view are represented mathematically as three different bases for this vector space. Thus, there are three different vector repre-

sentations of ψ corresponding to the neutral basis, the prosecution basis, and the defense basis: $\psi_N = \omega$, $\psi_P = \alpha$, and $\psi_D = \beta$. The four dimensional unit column vectors ω, α, and β represent the probability amplitudes for the joint events, $h_i \wedge e_j$, with respect to the different bases, or points of view.[3]

A set of matrix operators act on ψ to transform an individual's beliefs in correspondence with changes of perspective. Specifically, the probability amplitudes for one point of view are transformed into the probability amplitudes for a different point of view by unitary transformations:

$$\alpha = U_{pn} \cdot \omega$$

$$\beta = U_{dn} \cdot \omega.$$

For example, suppose an individual makes a judgment after reading the prosecution and then again after reading the defense. First, the individual sees the prosecution present evidence (e_1) favoring the guilty hypothesis. We project $\psi_P = \alpha$ onto the subspace corresponding to the evidence:

$$\psi_P = \begin{bmatrix} \alpha_{h_1 \wedge e_1} \\ \alpha_{h_1 \wedge e_2} \\ \alpha_{h_2 \wedge e_1} \\ \alpha_{h_2 \wedge e_2} \end{bmatrix} \Rightarrow \begin{bmatrix} \alpha_{h_1 \wedge e_1} \\ 0 \\ \alpha_{h_2 \wedge e_1} \\ 0 \end{bmatrix}.$$

We then normalize this projection to ensure that the length of the new state vector, $(\psi_P \mid e_1)$, equals one. When the individual is questioned about the conditional probability of guilt given the prosecution, the revised state is projected onto the guilty subspace. With the presentation of the defense, the revised state vector, $(\psi_P \mid e_1)$, is transformed from its vector representation associated with the prosecution basis to its vector representation associated with the defense basis by $U_{dp} = U_{dn} \cdot U_{pn}^{\dagger}$.[4] Now, we project our revised state onto the e_2 subspace since the defense is assumed to present evidence for innocence. Again, we normalize the state vector and project it onto the guilty subspace to calculate the conditional probability of guilt given the prosecution followed by the defense. Order effects arise because the unitary transformations are non-commutative. Figure 1 provides a schematic for the different sequences of transformations used for the different case orderings.

The model parameters define the specific matrix operators, U_{pn} and U_{dn}, used to transform one representation of ψ to another. We define a parameter for each case. So, there is a parameter associated with the strong prosecution, weak defense, and weak prosecution. Thus, to model the data collected by McKenzie et al. the quantum model uses three parameters.

Fitting the Data

We fit both the MAS model and the quantum model to the six probabilities shown in Table 1. Both models capture the qualitative properties of the data. Namely, $Pr(G \mid SP) < Pr(G \mid SP,WD)$ and $Pr(G \mid WP) > Pr(G \mid WP,WD)$. The quantum model fit the data with three parameters with the sum of squared error (SSE) equal to 0.00056; whereas, the MAS model fit the data with four parameters with the SSE = 0.0022. The SSE for the models is very small since we are examining differences in probabilities. Of the four parameters in the MAS model, three were associated with the minimum acceptable strength. The fourth parameter was the gradient parameter of a logistic function used map the average independent strength ratings from participants into the interval $[-1,1]$. If we compare the SSE from the quantum model to the SSE from the MAS model, we see that the SSE for the quantum model is much smaller than the SSE for the MAS model:

$$\frac{SSE_{MAS}}{SSE_{qt}} = \frac{0.0022}{0.00056} = 3.93.$$

Furthermore, since the quantum model has less parameters and a smaller SSE, it will have a lower BIC value than the MAS model.

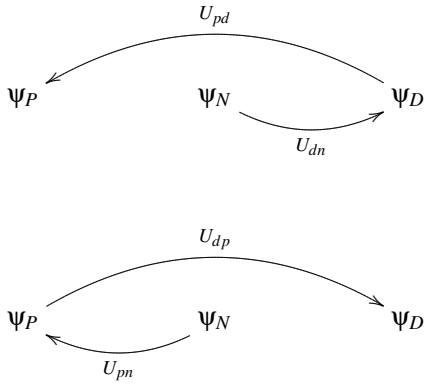


Figure 1: Transformations for different case orderings: defense followed by prosecution (top) and prosecution followed by defense (bottom).

Experiment 1: Extending McKenzie

McKenzie et al. did not examine all possible combinations of case strength and order. Assuming there are only two possible strengths, weak and strong, there are twelve total possible conditional probability judgments that can be made (see Table 2). Thus, we designed a new experiment to collected data for these twelve probabilities. Participants in this new study read eight different scenarios involving a defendant standing trial for either robbery, larceny, or burglary. Each participant was placed into one of eight conditions for each scenario. These eight conditions arise from the eight possible sequential judgments that can be made when taking into consideration order and case strength (e.g. weak prosecution followed by strong defense). Participants were placed in a different condition

[3] Probability amplitudes determine the belief about a particular event. Probabilities are calculated from probability amplitudes by taking the modulus of the amplitude and squaring.

[4] $U^{\dagger}$ is the conjugate transpose of U. For unitary matrices, $U^{\dagger}$ is also the inverse of U.

for each crime so they would experience all eight conditions by the end of the experiment. The participants reported the likelihood of the defendant's guilt before reading either case, after the first case, and after the second case.

Table 2: Conditional Probabilities for Jury Task

After first case	After second case	
Pr(G | WP)	Pr(G | WP, WD)	Pr(G | WP, SD)
Pr(G | SP)	Pr(G | SP, WD)	Pr(G | SP, SD)
Pr(G | WD)	Pr(G | WD, WP)	Pr(G | WD, SP)
Pr(G | SD)	Pr(G | SD, WP)	Pr(G | SD, SP)

Method

Participants in the study were 299 undergraduate students from Indiana University who received experimental credit for introductory psychology courses. For each scenario, there were approximately 38 participants in each each condition. All stimuli were presented on a computer and students entered their responses using the computer keyboard. For each scenario, participants were asked to imagine that they were jurors on the trial. They were also told that in each crime, the defendant was arrested after the police received an anonymous tip. One of the eight scenarios was directly patterned after the crime used by McKenzie et al. Likelihood of the defendant's guilt was reported on a continuous scale from 0 to 1 with 0 = certain not guilty, 0.5 = equally likely, and 1 = certain guilty.

Results

Eight of the 299 participants were excluded from the analyses because the majority of their initial ratings (before being presented with the prosecution or defense) were 0. These participants most likely assumed a literal interpretation of 'innocent until proven guilty'.

We first analyzed each scenario alone, and our analysis revealed a prevalence of recency effects. These effects arise when decision-makers place disproportionate importance on recent evidence (e.g. $Pr(G \mid SP,SD) < Pr(G \mid SD,SP)$). For each crime, there were four defense-prosecution pairs (SD v. SP, SD v. WP, WD v. SP, and WD v. WP) that could exhibit order effects. A two sample t-test showed the majority of pairs exhibited a significant recency effect ($p < 0.05$).

Since the scenarios were designed to be very similar, we reanalyzed the data by collapsing across all eight scenarios. A two sample t-test showed a significant recency effect for each of the four defense-prosecution pairs ($p < 0.001$).

Fitting the Data

The presence of recency effects in this new data set confirms earlier findings and provides the largest data set so far for comparing models that explain recency effects. Hogarth et Einhorn discovered that recency effects are prevalent in simple, step-by-step tasks with short series of evidence. Furthermore, there is evidence of recency effects in studies involving mock trials (Furnham, 1986 ; Walker, Thibaut, & Andreoli, 1972). Unlike the study conducted by McKenzie et al., none of the cases in this study caused a reversal in likelihood judgment when paired with opposing cases of different strengths. This might be due to the use of a standard numeric measure instead of a 21-point confidence scale. There is research showing that standard numeric measures can be insensitive to some judgment phenomena (Windschitl & Wells, 1996).

Since the data does not exhibit the effects found by McKenzie et al., we can fit the standard belief-adjustment model instead of the extended MAS model. We fit the averaging model, the adding model, and the quantum inference model to the mean likelihood of guilt for the eight different crimes as well as the averaged data. All three models use four parameters to fit the twelve data points associated with each crime. These parameters were fit by minimizing the sum of squared error between the data and model predictions. The four parameters used by the averaging and adding models arise from the four case strengths, $s(x_k)$, in equation 1. The four parameters for the quantum model arise from the matrix operators used to transform the state or belief vector. The minimized SSE for all three models are shown in Table 3. From this table, we see that both the adding and quantum models fit better than the averaging model. Also, the quantum model fits slightly better than the adding model in most cases.

Table 3: Model Fits

Crime	Averaging	Adding	Quantum
1	0.0719	**0.0112**	0.0132
2	0.0634	0.0083	**0.0056**
3	0.1185	0.0213	**0.0070**
4	0.0939	0.0156	**0.0127**
5	0.0913	**0.0091**	0.0109
6	0.0656	0.0130	**0.0113**
7	0.0913	0.0217	**0.0089**
8	0.0620	0.0164	**0.0023**
Average	0.0704	0.0059	**0.0058**

Figure 2 shows the model fits for the averaging model and quantum model for the strong defense-weak prosecution pair for the averaged data. From the figure, we see that the quantum model provides a much better fit. Fits for the remaining defense-prosecution pairs are similar.

Experiment 2: Extreme Evidence

To provide even more of a distinction between the quantum model and the belief-adjustment model, we conducted a second jury decision-making experiment involving extreme evidence. In this task, subjects read about an individual on trial for a crime in which the defense had an irrefutable argument. Specifically, the defense stated that the defendant was giving a public lecture when the crime was committed. The prosecution's argument was moderately strong: a witness claimed to

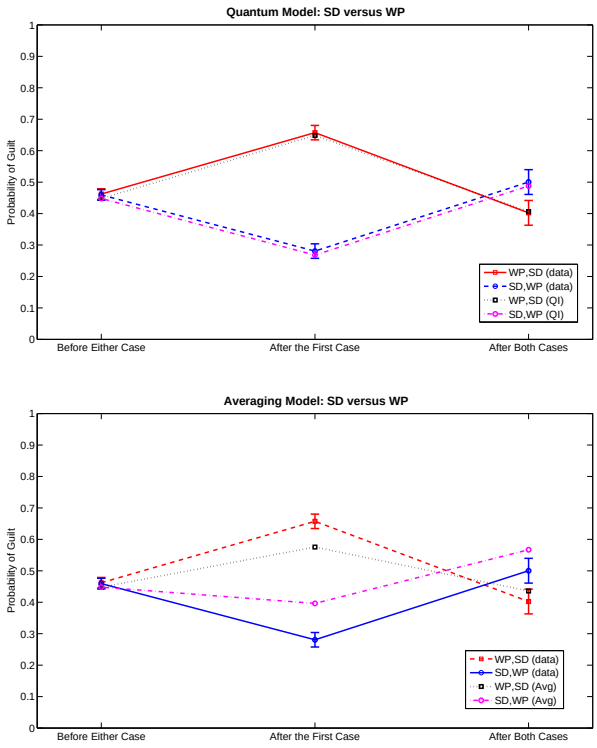


Figure 2: Averaging and quantum model fits to the mean likelihood of guilt for the strong defense-weak prosecution pair. Error bars on the data show the 95% confidence interval.

have seen the defendant near the scene of the crime. It seems reasonable to believe that the probability of guilt after hearing the defense will be near zero. Now, if the prosecution is presented after the defense, it is unlikely that the probability of guilt will increase by much. In terms of the belief-adjustment model, this places tight restrictions on the value of the prosecution strength parameter, $s(x_p)$. To see this, let's examine the adding version of the model:

$$C_p = C_d + (1 - C_d) \cdot s(x_p)$$

where C_p is the evaluation of the guilty hypothesis after hearing the prosecution's case and the irrefutable defense. We might assume the evaluation of the hypothesis after hearing just the defense, C_d, is near zero, say $C_d = \varepsilon_1$. Thus, $s(x_p)$ must also be a near zero, say $s(x_p) = \varepsilon_2$, in order for C_p to remain small:

$$C_p = \varepsilon_1 + (1 - \varepsilon_1) \cdot \varepsilon_2 = \varepsilon_1 + \varepsilon_2 - \varepsilon_1 \cdot \varepsilon_2 \approx 0.$$

Now, suppose the prosecution is presented before the defense. According to the adding model,

$$C_p = C_0 + (1 - C_0) \cdot s(x_p)$$

where C_0 is the evaluation of the guilty hypothesis before hearing either the prosecution or defense. We might assume that $C_0 \approx 0.5$. Thus, we have

$$C_p = 0.5 + 0.5 \cdot \varepsilon_2 \approx 0.5$$

showing the prosecution has little impact on the initial evaluation of the hypothesis. However, it seems unlikely that initial beliefs will be unaltered by the presentation of the prosecution. On the contrary, we might expect this prosecution to be very effective when no prior defense is presented. Essentially, the problem arises from the model's assumption that the strength of the prosecution, $s(x_p)$, is determined independently of other evidence.

This study used 164 undergraduate psychology students. Subjects were placed into one of two conditions corresponding to the two possible case orders: prosecution followed by defense or defense followed by prosecution. Similar to experiment 1, subjects entered responses on a computer and were told that the defendant was arrested after the police received an anonymous tip. Instead of providing the likelihood of the defendant's guilt, subjects were asked to rate their confidence in guilt on the same 21-point scale used by McKenzie et al. Like experiment 1, a significant recency effect was found ($p < 0.023$).

We converted the confidence ratings to probabilities and fit the quantum model and the adding model to the mean of these probabilities. We did not fit the averaging model since experiment 1 shows the adding model outperforms the averaging model. Figure 3 shows the model fits for the two models. Both the quantum model and the adding model use two parameters to fit the data. The SSE for the quantum model was 0.0002; whereas, the SSE for the adding model was 0.0158. By comparing the ratio of the SSE from the two models, we see that the quantum model provides a much better fit to the data:

$$\frac{SSE_{adding}}{SSE_{qt}} = \frac{0.0158}{0.0002} = 79.$$

The standard belief-adjustment model cannot capture dependences between the strength of the prosecution and the irrefutable defense. As a result, the model provides a poor fit to the data. Unlike the belief-adjustment model, the quantum model does not assume individuals combine evidence by simple arithmetic procedures such as adding or averaging. Instead, the quantum model supports the idea that evidence is viewed from different perspectives, and it is these different, or incompatible, points of view that allow the quantum approach to capture the effects of extreme evidence.

Conclusion

One might question the extent to which quantum probability models are rational. Like classic (Kolmogorov/Bayesian) probability theory, quantum theory is based on a coherent set of axioms. Then the question falls back on which set of axioms is most appropriate for an application. For example, models based on the axioms of quantum probability theory have been used to explain paradoxical phenomena arising in cognitive science such as violations of rational decision

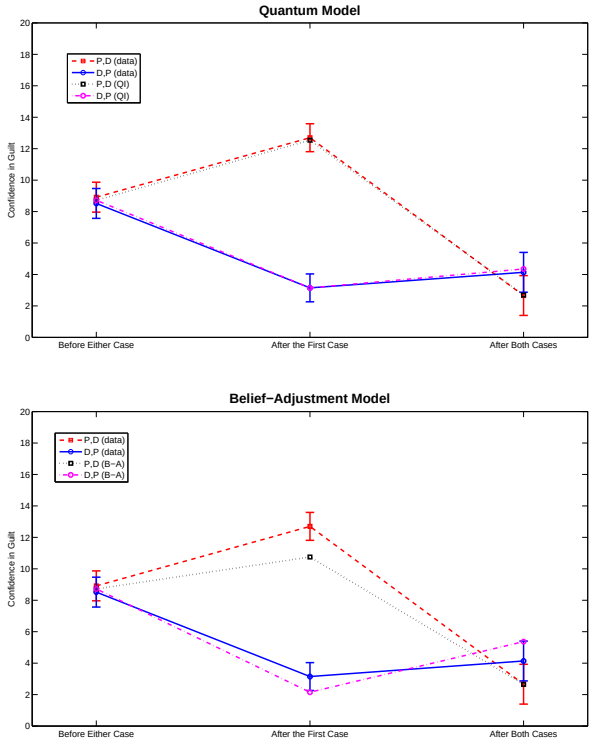


Figure 3: Adding and quantum model fits to the mean probability of guilt from experiment 2. Error bars on the data show the 95% confidence interval.

making principles (Pothos & Busemeyer, 2009), paradoxes of conceptual combination (Aerts, 2009), human judgments (Khrennikov, 2004), and perception (Atmanspacher, Filk, & Romer, 2004).

Here we provide evidence in support of a quantum probability explanation of order effects. Using data collected by McKenzie et al., we show that the quantum inference model out performs the minimum acceptable strength model. We also provide evidence that the quantum model performs as well or slightly better than the belief-adjustment model when fitting data from experiment 1. Finally, we describe some of the limitations of the belief-adjustment model in relationship to irrefutable evidence. We argue that the quantum inference model is not faced with these limitations and provides more reasonable predictions. In the future, we plan to continue empirically investigating the quantum inference model in the hope of developing a more coherent theory concerning human inference tasks.

Acknowledgments

This research was supported by the National Science Foundation/IGERT Training Program in the Dynamics of Brain-Body-Environment Systems and by the National Science Foundation under Grant No. 0817965.

Références

Aerts, D. (2009). Quantum structure in cognition. *Journal of Mathematical Psychology*, *53*, 314-348.

Atmanspacher, H., Filk, T., & Romer, H. (2004). Quantum zero features of bistable perception. *Biological Cybernetics*, *90*, 33-40.

Bergus, G. R., Chapman, G. B., Levy, B. T., Ely, J. W., & Oppliger, R. A. (1998). Clinical diagnosis and order of information. *Medical Decision Making*, *18*, 412–417.

Busemeyer, J. R., & Trueblood, J. (2009, March). Comparison of quantum and bayesian inference models. In *Quantum interaction: Third international symposium, qi 2009.* Saarbrucken, Germany : Springer.

Furnham, A. (1986). The robustness of the recency effect: Studies using legal evidence. *Journal of General Psychology*, *113*, 351–357.

Gilovich, T., Griffin, D., & Kahneman, D. (2002). *Heuristics and biases: the psychology of intuitive judgment*. Cambridge University Press.

Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology*, *24*, 1–55.

Khrennikov, A. Y. (2004). *Information dynamics in cognitive, psychological, social and anomalous phenomena*. Kluwer Academic.

McKenzie, C. R. M., Lee, S. M., & Chen, K. K. (2002). When negative evidence increases confidence: Change in belief after hearing two sides of a dispute. *Journal of Behavioral Decision Making*, *15*, 1–18.

Pothos, E. M., & Busemeyer, J. R. (2009). A quantum probability explanation for violations of 'rational' decision theory. *Proceedings of the Royal Society B*.

Trueblood, J. S., & Busemeyer, J. R. (2010). A quantum probability explanation for order effects on inference. *submitted to Cognitive Science*.

Tversky, A., & Shafir, E. (1992). The disjunction effect in choice under uncertainty. *Psychological Science*, *3*, 305-309.

Walker, L., Thibaut, J., & Andreoli, V. (1972). Order of presentation at trial. *Yale Law Journal*, *82*, 216–226.

Windschitl, P. D., & Wells, G. L. (1996). Measuring psychology uncertainty: Verbal versus numeric methods. *Journal of Experimental Psychology: Applied*, *2*, 343-364.

Information Relevance in Pseudodiagnostic Reasoning

Frédéric Vallée-Tourangeau and Gaëlle Villejoubert
Psychology Research Unit, Kingston University
Kingston-upon-Thames UNITED KINGDOM KT1 2EE
f.vallee-tourangeau/g.villejoubert@kingston.ac.uk

Abstract

When faced with two competing hypotheses, people sometime prefer to look at multiple sources of information in support of one hypothesis rather than to establish the diagnostic value of a single piece of information for the two hypotheses. This is termed pseudodiagnostic reasoning, and is understood to reflect a pervasive confirmation bias. Past research suggests that diagnostic reasoning may be more easily fostered when participants seek data to help in the selection of one of two competing courses of action as opposed to situations where they seek data to help inferring which of two competing hypotheses is true. In the experiment reported here, we provide the first empirical evidence demonstrating that the facilitating effect observed in action problems is driven by considerations of information relevance, reasoners' motivations and the numerical value of the first piece of information presented. The discussion of these findings focuses on implications for the ability to engage in diagnostic hypothesis-testing.

Keywords: decision making; utilities; pseudodiagnostic reasoning

Introduction

A sales manager advertised a new position for a sales assistant. After reviewing the curriculum vitae of the candidates, she selects two promising applicants, Ms. A. and Ms. B. The manager initially leans towards the first candidate, Ms. A., and discovers that she successfully completed 70% of her sales transactions in the last month in her previous position. The manager must now engage in inductive reasoning: She needs to collect more information in order to decide whether Ms. A or Ms. B. is the best candidate for the job. A long tradition of psychological research suggests that her search for information will be driven by a need for evidence confirming the hypothesis she is entertaining. Thus, if, at this point, the manager believes Ms. A. *is* the best candidate, she would naturally seek more information about Ms. A rather than checking Ms. B.'s sale performance. Yet, this strategy is potentially shortsighted: Ms. B. could well have outperformed Ms. A. on the sales front in her previous job, in which case, seeking more information about Ms. A. could be misguided and lead to the employment of a candidate with less potential. Without establishing the sales performance of Ms. B, the diagnostic value of Ms. A's sales performance is undetermined, and hence cannot judiciously inform the decision-making process.

More generally, the diagnosticity of a datum D for a given hypothesis X (H_X) is defined in terms of the ratio of the probability that D is observed given that H_X is true, $P(D|H_X)$, and the probability that D is observed given that an alternative hypothesis Y is true, $P(D|H_Y)$. Hence, diagnosticity can only be assessed from the perspective of *multiple* hypotheses. The likelihood ratio in Bayes's Theorem is the normative metric of the diagnosticity of information (Doherty, Mynatt, Tweney, & Schiavo, 1979).

People's Search for Information Is Not (Always) Driven by Considerations of Diagnosticity

Early research examining how people gather information in order to make inferences suggested they did not fully appreciate that diagnosticity is defined in terms relative to at least two hypotheses, not just one (Beyth-Marom & Fischhoff, 1983; Doherty et al. 1979; Kern & Doherty, 1981; cf. Trope & Bassok, 1982, 1983). For example, Beyth-Marom and Fischhoff (1983, Experiment 1) told participants that an individual possessed a distinguishing feature and asked them what information they deemed relevant to determine whether that individual was a member of Group A. Nearly 90% of their participants indicated that it was relevant to know P(D|group A), but of those 'only' 50% deemed important to know the probability that this information would also be observed given membership in a different group, or P(D|group B). Yet both probabilities must be examined in order to gauge the diagnosticity of the distinguishing feature.

Mynatt, Doherty and Dragan (1993) argued that people's hypothesis testing process is predominantly concerned with gathering evidence in favour of one hypothesis, rather than determining the diagnosticity of any given piece of information for multiple hypotheses. In one of their reasoning scenarios, participants were asked to determine which of two cars, X or Y, their "sister" purchased. They were told about *two features characterising this car*, its petrol consumption (D_1: "25 miles per gallon" –mpg) and its mechanical reliability (D_2:"no major mechanical problems in the first two years of ownership"). In addition, participants were given *an anchoring piece of information*: "65% of car Xs do over 25 mpg", or $P(D_1|H_X) = .65$. Participants were then asked to choose which of the following three pieces of information would help them

decide which type of car was owned by their sister (the participants did not see the information presented here in brackets).

1. The percentage of car Ys that do over 25 mpg. [Diagnostic, $P(D_1|H_Y)$].
2. The percentage of car Xs that have had no major mechanical problems for the first two years of ownership. [Pseudodiagnostic, $P(D_2|H_X)$].
3. The percentage of car Ys that have had no major mechanical problems for the first two years of ownership. [Switching, $P(D_2|H_Y)$].

The first choice would establish the diagnosticity of the petrol consumption data. If more than 65% of car Ys do over 25 mpg, then the sister's car is more likely to be a Y car. Otherwise, it is more likely to be an X car. The second choice would determine the mechanical reliability of car Xs. This choice leaves the reasoner with two pieces of information, and the diagnosticity of neither can be established. Learning that a high percentage of car Xs also featured good mechanical reliability could make one confident that the sister owned a car X, but this feeling of confidence would only be illusory: here again, car Xs could well be outperformed by car Ys. Hence, choosing to learn about $P(D_2|H_X)$ is considered a pseudodiagnostic choice. We term the third choice *switching* because the focus switches from the initial information (D_1) to the second piece of information (D_2) and from the initial hypothesis (H_X) to the alternative one (H_y).

In Experiment 1 of Mynatt et al. (1993), 60% of the participants chose to learn about the percentage of car Xs with good mechanical reliability in order to determine the identity of the car, while only 26% chose to know the percentage of car Ys that do 25 mpg. The majority of participants thus made what is considered a pseudodiagnostic choice, since opting to look at the mechanical reliability of car Xs cannot determine the diagnosticity of being informed that 65% of car Xs do over 25 mpg.

In contrast, there is evidence demonstrating that people *will*, under various circumstances, seek to know information that would establish the diagnosticity of the anchoring information. For example, if the anchoring information defines a relatively rare feature, for example 65% of cars of make X have a top speed of 165 miles per hour, participants are more likely to want to know the proportion of the alternative make of cars that reach that top speed (Feeney, Evans, & Venn, 2008, Experiment 1). The rare, and arguably more interesting feature (it's plausibly more interesting to be told about a top speed of 165 mph than that the car has an ashtray), thus invite participants to gauge its frequency given the alternative category, thereby encouraging diagnostic data selection. In addition, if the dimensions that define the target object (e.g., 'your sister's car') are couched in terms of an actor's motivation (e.g., you sister *wanted* a car with good petrol consumption) then anchoring information that appears to run counter to this motivation (e.g., car X does 20 miles per gallon) elicits *less* pseudodiagnostic reasoning. Likewise, a low value for the initial anchoring information (e.g., 35% of car Xs do over 25 mpg) encourages more diagnostic choices (Mynatt et al., 1993, Experiment 2). Thus, upon learning that $P(D_1|H_X)$ is relatively low, more participants are interested in $P(D_1|H_Y)$. Conversely, if the anchoring information given plausibly endorses the focal hypothesis, that is when $P(D_1|H_X)$ is high, participants appear less motivated to determine the diagnosticity of that information.

Information Relevance and Initial Values in Action Problems

Another important characteristic that seems to determine whether people will make diagnostic search choices is the goal of the task. Mynatt et al. (1993) distinguished between *inference* and *action* problems. The car example discussed above, they argued, represents an inference problem. The car has already been purchased and is owned by someone, the goal is to determine whether it is a car X or a car Y. In effect the problem is a categorization inference, and in principle the categorization can be true or false. In contrast, an action problem is one where hypotheses represent two courses of action. One might be better than the other, but the decision cannot in principle be evaluated in terms of whether one action is true and the other false. In a separate experimental condition, Mynatt et al. instructed participants to imagine *buying* a car, considering car X or car Y and told them they were "concerned about (…) petrol consumption and mechanical reliability" (p. 768). Participants were then given the same anchoring piece of information ("65% of car Xs do over 25 mpg") and were asked to choose one among the same three pieces of information in order to help them decide which car to buy (see options 1. through 3. above).

In that action problem, 52% chose the piece of information that could determine the petrol consumption of car Ys (the diagnostic choice) and 41% chose the piece of information that could determine the mechanical reliability of car Xs (a pseudodiagnostic choice). To explain the high proportion of diagnostic choices in action problems, Mynatt et al. propose that the choice among the three alternatives is determined by the datum which bears more utility for each individual participant: "Precisely how many subjects will select (the diagnostic choice) will depend on the content of a given problem and subjects' *idiosyncratic* utility function and decision strategies" (pp.765-766, emphasis added). On this account, those who consider petrol consumption to be more important than mechanical reliability would be motivated to establish the petrol consumption of car Ys and hence chose the diagnostic option. In contrast, those who are more concerned about mechanical reliability should seek information about car Xs' mechanical reliability, a pseudodiagnostic choice.

The authors, however, did not manipulate explicitly the perceived relevance of the two dimensions characterising each alternative (e.g., petrol consumption and mechanical reliability in the car scenario) nor did they seek to assess how relevant their participants believed these dimensions to be. Moreover, there is conflicting evidence showing that action problems may not necessarily promote diagnosticity. Maggi, Butera, Legrenzi and Mugny (1998, Experiment 1) asked participants to imagine having to choose between two cars or two political candidates. These authors found over 60% choices to be pseudodiagnostic. There was, however, an important methodological difference between their task and that of Mynatt et al. (1992): Maggi et al. (1998) presented participants with *four* possible pieces of information to choose from for each alternative (e.g., the car price, reliability, fuel consumption and performance). In addition, the authors found that people tended to be more diagnostic in their choices when the anchoring information concerned a characteristic they believe to be important (e.g., the price of a car or the competence of a political candidate). In light of these incongruous findings, one important issue to resolve would thus be to determine whether those who made a diagnostic choice by choosing to look up $P(D_1|H_Y)$ did so *because* they were more interested in D_1 than in any other piece of information D.

Another important difference between inference and action problems outlined by Mynatt et al. (1993) is the role of the initial $P(D_1|H_X)$ value of the anchoring information. According to the authors, in inference problems this initial value could be a cue to the truth value of H_X and, as such, dictate participants' information search. By contrast, the authors predicted and found that this initial value would not affect choices in action problems since, in those situations, information search would solely be determined by the perceived relevance of the anchoring information D_1. The authors, however, tested this prediction by comparing relatively narrow values, hovering modestly below and above the 50% mark (viz. 35% vs. 65%). It is therefore reasonable to assume that participants who believed, for example, petrol consumption was *the* most important attribute for a new car would always wonder if car Ys outperformed car Xs, even upon learning that 65% car Xs did over 25 mpg. However, this does not necessarily imply that participants' information search strategy will never be affected by the $P(D_1|H_X)$ value in action problems. It is not implausible to expect, for example, that participants may no longer search to establish a diagnosticity ratio when told that 95% of car Xs do over 25 mpg. In this case, the $P(D_1|H_X)$ value could be a cue to the superiority of H_X. Consequently, under such circumstances, participants might then be more interested in learning more about car Xs than in finding out whether car Ys outperform car Xs. Hence, when the $P(D_1|H_X)$ value is deemed satisfactory in action problems, we should expect more pseudodiagnostic choices.

The Present Study

In the experiment reported here, we examined the role of information relevance and initial values in determining diagnostic and pseudodiagnostic choices in problems structurally isomorphic to the one developed by Mynatt et al. (1993). The first aim of this experiment was to test the hypothesis that diagnostic choices in action problems occur because people believe D_1 is more relevant than D_2 in deciding whether to take action X or action Y. We manipulated the relative importance of D_1 and D_2 in two scenarios so that participants would care more about D_1 than D_2. We anticipated the higher perceived relevance of D_1 would lead to a higher proportion of diagnostic choices, since participants would seek to determine the probability of D_1 given the alternative course of action. Second, this experiment aimed to assess the degree to which people may revert to a pseudodiagnostic search for information when the initial value of the anchoring information $P(D_1|H_X)$ is deemed satisfactory. To do so, we manipulated the motivation underpinning participants' action. One scenario was designed to motivate people to find *the highest value* of $P(D_1|H)$. In this case, we anticipated that participants would never be satisfied by the initial value of $P(D_1|H_X)$ and hence we predicted that their search for information would not be affected by this initial value. The alternative scenario was designed to motivate people to find *a satisfactory value* of $P(D_1|H)$. In such a situation, we predicted that when the initial $P(D_1|H_X)$ presented could be deemed satisfactory, the rate of pseudodiagnostic choices would be greater.

Method

Participants

Participants were recruited by third-year psychology students at the University of Toulouse, France, as a course requirement. Each student made a list of several men and women who were older than 18 and not studying psychology, randomly drew one man and one woman from his or her list, and asked them to take part in a general survey which included the present study. Of the 1040 participants in the final sample (520 men, 520 women; mean age = 31.37, *SD* = 13.24), 11% had completed graduate school, 53% had an undergraduate education, 20% had graduated from high school only, and the remaining 16% had not graduated from high school. The sample included a large proportion of students (40%), but also working professionals (51%) and retired or unemployed individuals (8%). The survey was conducted in French.

Design and Procedure

The current experimental manipulation was embedded in a longer questionnaire. The experiment used a 2 × 2

between-subjects design. The independent variables were the implicit motivation of the decision-maker (maximizing vs. satisficing – Simon, 1955) and the numerical value of the anchoring piece of information (high or low). Participants were randomly allocated to one of the resulting four conditions. Their task was presented as follows: they were asked to imagine they were the director of a large zoo and that they had set up a programme aiming to promote reproduction in captivity of African elephants, a species at risk of extinction. Their calves, however, were facing a severe health issue. They were informed of the presence of a parasite whose eggs could lodge in the calves' aortic artery, causing strokes and killing the calves if left untreated, threatening the success of the reproduction programme. The experimental manipulation concerned the animals needed to be treated in order to rid the zoo of parasites. Half the participants were told the deadly parasites were infecting the calves directly and treatment was, therefore, to be administered to calves (satisficing scenario). The remaining half was told the deadly parasites were carried by roaming rats which were to be treated directly (maximizing scenario). In all cases, the zoo's chief veterinary suggested using one of two treatments to save the calves: treatment A or treatment B. Participants were then told about the mortality rate of calves (rats) treated with treatment A and that both treatments could also potentially cause infertility in calves (rats). The initial value of the anchoring piece of information was either high or low. Thus, half were told treatment A could cause the death of 80% calves (rats) whereas the remaining half were told it could cause the death of 20% calves (rats). Before making their choice, however, they were allowed to consult one additional piece of information among three alternatives: they could choose to consult the mortality rate of calves (rats) treated with treatment B (a diagnostic choice). They could also choose to learn more about treatment A and ask to consult the percentage of infertile calves (rats) among those treated with treatment A (a pseudodiagnostic choice). Finally, they could choose to learn about the rate of infertility observed in calves (rats) treated with treatment B (a switching choice). The order of the choices remained constant in all experimental conditions.

In both scenarios, we anticipated that people would be more concerned about mortality rates (D_1) than about infertility rates (D_2). Specifically, we anticipated people to place more value on the mortality rate of rats than on their infertility since rats made infertile would not eliminate their status as a contamination vector. Likewise, we anticipated people would be more concerned about avoiding the death of the endangered calves than about their potential infertility. A few pages later in the survey, all participants were asked to consider the calves (rats) task again and to rate the importance of avoiding killing the calves (rats) as well as the importance of avoiding making the calves (rats) sterile. Both ratings were recorded on an 8-point scale ranging from 1 (Absolutely not important) to 8 (Extremely important).

We expected that the greater relevance of the anchoring dimension (the mortality rate) induced by the scenarios would result in a large proportion of diagnostic choices. In the maximizing scenario, we predicted that participants would be motivated to find the best treatment to kill *all* the rats, and that consequently, short of a 100% mortality rate, they would be more interested in determining the mortality rate of the alternative treatment *regardless* of the mortality rate for treatment A. As a result, we expected high proportions of diagnostic choices in these scenarios when the mortality rate for treatment A was set at either .20 or .80. In the satisficing scenario, we anticipated that mortality would bear unacceptable consequences to a degree that varied with the rate associated with treatment A. We predicted that participants would deem it important to save as many calves as possible and that, consequently, the 80% chance of killing the host organism associated with treatment A would be deemed unacceptably high. In this situation, we predicted a strong preference for enquiring about the mortality rate associated with treatment B, the diagnostic option. With a lower mortality rate of .20, however, we predicted that some participants might deem it satisficingly low and hence be tempted to enquire about the infertility rate associated with treatment A, resulting in higher proportion of pseudodiagnostic choices in this condition.

Table 1: Mean and standard deviation of the importance ratings of preventing killing or preventing infertility in the host organism (rats or calves) in the conditions where the rats and elephant calves are treated as a function of the mortality rate for treatment A, .2 and .8.

		Motivation			
		Maximizing (rats)		Satisficing (calves)	
		Mortality Rate of Treatment A		Mortality Rate of Treatment A	
		.2	.8	.2	.8
Prevent Killing	*M*	3.37	3.59	7.19	7.18
	SD	2.30	2.48	1.17	1.32
Prevent Infertility	*M*	3.09	3.30	6.93	6.87
	SD	2.32	2.57	1.36	1.64

Results

Importance Ratings

Participants were asked to rate the importance of not killing the host animals or not making them infertile. These mean importance ratings are reported in Table 1. The importance of saving the calves was consistently rated higher than that of saving the rats. The same was true for the infertility ratings: Participants judged it more important to prevent infertility in the calves than in the rats. Most notably, as we had anticipated, participants deemed it more important to prevent mortality than infertility.

A 2 (goals: maximizing, satisficing) × 2 (mortality rate of treatment A: .2, .8) × 2 (rating type: killing, infertility) mixed analysis of variance (ANOVA) confirmed these observations. The main effect of goal was significant, $F(1, 1036) = 1255$, $p < .001$, $MSE = 5.69$, $\eta^2 = .55$, the main effect of the mortality rate was not significant, $F < 1$, and the main effect of rating type was significant, $F(1, 1036) = 20.1$, $p < .001$, $MSE = 2.11$, $\eta^2 = .02$. None of the interactions were significant, largest non reliable $F(1, 1036) = 1.46$.

Choice Preferences

Two participants did not make a choice and were discarded from subsequent analyses. Consistent with our predictions, the diagnostic option was by far the most frequently chosen in all experimental conditions with over 70% of participants opting for this type of information (see Fig. 1). In the maximizing scenarios (treating the rats), nearly 80% of the participants elected to examine the mortality rate associated with treatment B, and the remaining 20% of the participants were evenly split between the other two options (the infertility rate for treatment A or B). Moreover, this pattern of choice was identical whether the value of the mortality rate of treatment A was said to be 20% or 80%. In contrast, in the satisficing scenarios (treating the calves), while most participants still elected primarily to determine the mortality rate of treatment B (the diagnostic choice), the frequency of pseudodiagnostic choices was nearly twice as large when treatment A had a relatively low mortality rate (20%) compared to when it had a high mortality rate (80%). Approximately one fifth of the participants chose the irrelevant option, the infertility rate of treatment B in both versions of the satisficing scenario.

A number of χ^2 tests were conducted. The first determined that the choice frequencies in all four experimental conditions differed significantly, χ^2 (df = 6, $N = 1038$) = 35.2, $p < .001$. The proportion of diagnostic choices was significantly higher when the implicit goal was to maximize the number of rats killed, χ^2 (df = 1, $N = 869$) = 5.46, $p < .02$. Separate tests were then conducted within the maximizing (rats) and the satisficing (calves) scenarios, excluding the switching choice frequencies. Within the maximizing scenarios, the frequencies of diagnostic and pseudodiagnostic choices did not differ as a function of the mortality rate of treatment A, χ^2 (df = 1, $N = 463$) = .11, $p > .05$. In contrast, within the satisficing scenarios, the frequencies differed significantly as a function of the mortality rate of treatment A, χ^2 (df = 1, $N = 406$) = 9.22, $p < .005$.

Discussion

This experiment successfully manipulated the relative importance of the information dimensions available in a two-alternative action problem where participants were asked to choose which treatment they should use to save endangered calves whose life was threatened by a deadly parasite. Specifically, ratings of the importance of not killing the animals and not making them infertile confirmed that the mortality dimension was judged more important than the infertility dimension whether the animals were the calves themselves or rats hosting the parasite.

The examination of information search patterns in turn confirmed that when the anchoring dimension was perceived as being the most relevant, participants were strongly drawn to check the diagnostic option: nearly 71% of the 1038 responses collected were diagnostic. This high

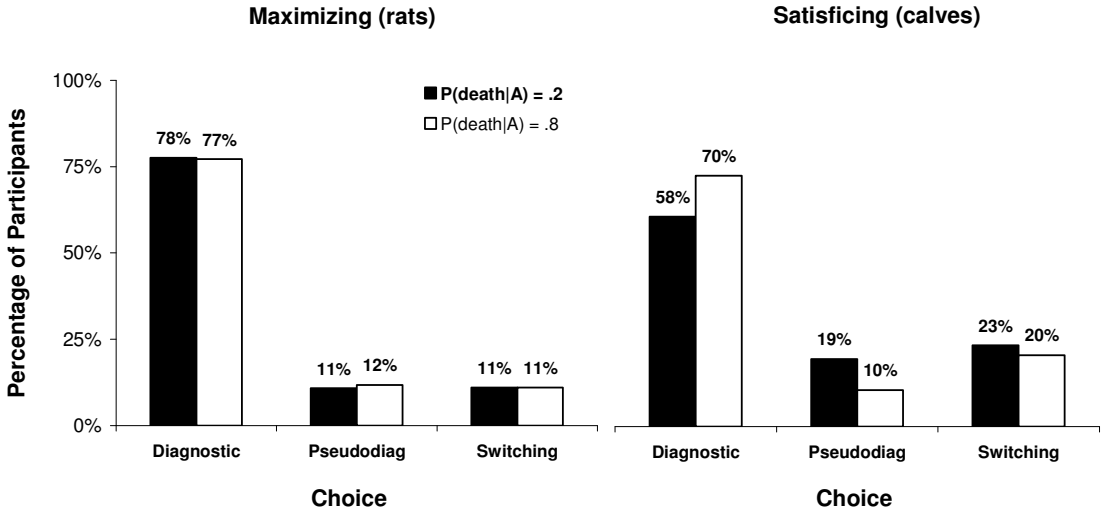


Figure 1: Percentage of participants making a diagnostic, pseudodiagnostic, or switching choice for the maximizing scenarios (involving the rats) and the satisficing scenarios (involveing the calves) for both values of P(death|treatment A).

degree of consensus is all the more impressive given the size and the variety of the sample from which it originated. This result supports Mynatt et al.'s (1993) initial, albeit untested, assumption that individuals' search for information is driven by considerations of utility in action problems.

We note the large inconsistency between the rate of diagnostic choice observed in our action problems (on average, more than 70% of our participants chose the diagnostic option) compared to those observed by Maggi et al. (1998) with a similar task (on average, less than 40% of their participants did so). These authors, however, had also found that people were more diagnostic when the anchoring dimension was one they judge to be important. Recall that an important methodological difference between their task and the original action problems used by Mynatt et al. (1993), as well as that used in the present study, was the number of dimensions participants could choose from. Whereas our participants and Mynatt et al.'s (1993) could only choose to look up the probability associated with two dimensions (D_1 and D_2) for each of two alternatives, Maggi et al.'s participants were presented with four such dimensions (D_1, D_2, D_3, and D_4). Moreover, the authors rotated the dimension defined as the anchoring dimension so that some participants would be first given information about D_1, others about D_2 and so on. Suppose that participants' search strategy is primarily driven by the importance of the dimension and suppose D_1 was the dimension they deemed most important. This means that whenever the anchoring dimension was *not* the most important dimension (3 times out of 4), participants would seek to learn more about D_1 for the current alternative and hence make a pseudodiagnostic choice. In other words, perhaps the reason why so many people made pseudodiagnostic choices in Maggi et al.'s (1998) task was because most of the time the anchoring dimension was not the dimension bearing the highest utility.

Finally, in line with our initial predictions, but contrary to Mynatt et al.'s (1993) conclusions, we were able to demonstrate that the numerical value of the anchoring dimension could affect people's search strategies when their motivation was to find a satisficing alternative. In such circumstances, a more satisficing value (viz., a relatively low mortality rate of endangered calves) resulted in almost twice as many pseudodiagnostic choices than a plainly unsatisfactory value (viz., a high mortality rate). This suggests that people will *also* engage in confirmatory search for information when they aim to choose between two courses of actions (and not only when they seek to make an inference, as the authors had previously concluded). These data therefore offer strong support for the hypothesis that the perceived relevance of the dimensions that define two courses of action governs the information search strategies adopted by reasoners. In addition, they establish that such strategies can also be modified depending on what the decision-maker is motivated to achieve, namely either identify a satisficing alternative or identify a utility maximizing alternative.

References

Beyth-Marom, R., & Fischhoff, B. (1983). Diagnosticity and pseudo-diagnosticity. *Journal of Personality and Social Psychology*, *45*, 1185-1195.

Doherty, M. E., Chadwick, R., Garavan, H., Barr, D., & Mynatt, C. R. (1996). On people's understanding of the diagnostic implications of probabilistic data. *Memory and Cognition*, *24*, 644-654.

Doherty, M. E., Mynatt, C. R., Tweney, R. D., & Schiavo, M. D. (1979). Pseudodiagnosticity. *Acta Psychologica*, *43*, 111-121.

Feeney, A., Evans, J., Venn, S. (2008). Rarity, pseudodiagnosticity and Bayesian reasoning. *Thinking & Reasoning, 14*, 209-230.

Kern, L., & Doherty, M. E. (1982). "Pseudodiagnosticity" in an idealized medical problem-solving environment. *Journal of Medical Education*, *57*, 100-104.

Maggi, J., Butera, F., Legrenzi, P., & Mugny, G. (1998). Relevance of information and social influence in the pseudodiagnosticity bias. *Swiss Journal of Psychology, 57*, 188-199.

Mynatt, C. R., Doherty, M. E., & Dragan, W. (1993). Information relevance, working memory, and the consideration of alternatives. *Quarterly Journal of Experimental Psychology*, *46A,* 759-778.

Simon, H. A. (1955). A behavioral model of rational choice. *Quarterly Journal of Economics, 69*, 99-118.

Trope, Y., Bassok, M. (1982). Confirmatory and diagnosing strategies in social information gathering. *Journal of Personality and Social Psychology*, *43*, 22-34.

Trope, Y., Bassok, M. (1983). Information-gathering strategies in hypothesis-testing. *Journal of Experimental Social Psychology, 19*, 560-576.

Accessing the Unsaid: The Role of Scalar Alternatives in Children's Pragmatic Inference

Neon Brooks (neonblue@uchicago.edu)
Department of Psychology, University of Chicago, 5848 South University Avenue, Rm. Green 510
Chicago, IL 60637 USA
Alan Bale (acbale@alcor.concordia.ca)
Department of Linguistics, Concordia University, H663, 1455 de Maisonnueve Blvd. W.
Montreal, Quebec CA
David Barner (barner@ucsd.edu)
Department of Psychology, University of California, San Diego, 5336 McGill Hall, 9500 Gilman Drive
La Jolla, CA 92093 USA

Abstract

When faced with a sentence like, "some of the toys are on the table", adults, but not preschoolers, compute a scalar implicature, taking the sentence to suggest that not all the toys are on the table. Although this difference is sometimes attributed to children's difficulties in processing and pragmatic understanding, this paper explores the hypothesis that children fail to compute scalar implicatures because they lack knowledge about relevant lexical alternatives to words like "some". Four-year-olds were shown pictures in which two objects fit a description and a third object did not, and were asked to judge the truth value of statements that relied on context-independent alternatives (e.g. *only some of the toys are on the table*) or contextual alternatives (e.g. *only the drum and the ball are on the table*). Children computed scalar implicatures only in the case of contextual alternatives, and only when the statements were grammatically strengthened, supporting the hypothesis that children's difficulties with scalar implicature result from a lack of knowledge of the relevant alternatives.

Keywords: scalar implicature; pragmatic reasoning; processing limitations; contextual alternatives

Introduction

As children acquire language, their task is complicated by the fact that speakers' intended meanings go beyond the literal meanings of their utterances. Word learning is not simply a process of mapping strings of words to speaker intentions. Instead, children must infer the core lexical meanings of words by distinguishing what is logically entailed from that which is merely implied. For example, given a dialogue like (1), John is likely to infer that Mary did not eat all of his cake.

(1) John: Did you eat my cake?
Mary: I ate some of it.

Although Mary's statement would be literally true if she had eaten the whole cake (eating *all* entails eating *some*), her utterance implies that she did not. This inference relies on the assumption that, if Mary had eaten the whole cake, and was communicating cooperatively, she would have uttered a more informative statement like "I ate *all* of it" (Grice, 1989).

The language acquisition literature is filled with examples of children learning words by making inferences about speaker intentions. A classic demonstration of this comes from experiments investigating mutual exclusivity. When a child is shown two objects, one of which has a known label (e.g., a car), they infer without difficulty that a novel label (e.g., *dax*) refers to the previously unlabeled object. Such an inference follows from the assumption that the speaker would not use two words to denote one kind of object (i.e., words exhibit *mutual exclusivity*, or *contrast*; Markman, 1989; Clark 1987). Children apply such strategies not only when learning nouns, but also when interpreting other classes of words, such as numerals. For example, when 2-year-olds who know the meaning of the word *one* (but no higher number) are shown two sets – e.g., 1 balloon and 5 balloons – they readily infer that *five* refers to the set of five objects (Wynn, 1992; Condry & Spelke, 2008).

Amidst such evidence, and further studies which find that children are sensitive to subtle intentional cues like eye gaze, speaker desires, etc. (Baldwin, 1993; Tomasello, 1992) children also exhibit striking failures in computing some simple inferences, including the inference in (1), which is a type of *scalar implicature*. Following Horn (1989), it is typically assumed that the quantifier *some* belongs to a larger class of terms called "scalar items". Scales are used to generate sets of alternative meanings, which are ordered according to their informativeness, and are implicitly contrasted during interpretation. In the case of *some*, the relevant scale includes other quantifiers – e.g., *a, some, many, most, all*. Examples of such scales are shown in (2):

(2) a. <*some, many, most, all*, etc.>
b. <*warm, hot, boiling*, etc.>
c. <*one, two, three*, etc.>

By most accounts, deriving a scalar implicature involves at least four steps, summarized in I - IV. First, the listener computes the basic, *literal*, meaning of the expression (Step I). Second, she considers the alternative sentences that might have been uttered (by substitution of scalar alternatives; Step II). Third, she restricts these alternatives by removing those that are less informative (Step III). Finally, she "strengthens" the interpretation of the sentence by negating the remaining alternatives – e.g., "I ate some (but not all) of the cake" (Step IV).

I. Compute basic meaning of a sentence S containing L, a scalar item.

II. Generate a set of alternatives (a_1, a_2, . . . , a_n) to S, called S_{alt}. These are all the sentences that can be generated by replacing L with its scalar alternatives.

III. Restrict the alternatives in S_{alt} by removing any alternative that is entailed by the original utterance S. Call this restricted set S*.

IV. Strengthen the basic meaning of S (containing L) with the negation of all of the members of S*.

A large number of studies have found that children fail to derive such implicatures. This has been shown for many scalar contrasts, including *might* vs. *must* (Noveck, 2001), *a* vs. *some* (Barner, Chow, & Yang, 2009), *some* vs. *all* (Huang & Snedeker, 2009; Papafragou & Musolino, 2003; Noveck, 2001), and *or* vs. *and* (Chierchia et al., 2001). For example, in a study by Papafragou and Musolino (2003), 5-year-old children were shown a scene including three horses, in which all three jumped over a log. When asked whether the sentence, "Some of the horses jumped over the log" was a good description of the event, most children said yes. Adults, in contrast, denied that this was a good description, since *all* of the horses jumped over the log. Adults, unlike children, computed a scalar implicature. Children do not always lack a so-called *strengthened* meaning. Papafragou and Musolino found that children provided adult-like responses when tested with numerals. Children denied that "Two of the horses jumped over the log" when three horses did. Thus, although children failed to have adult-like response with *some* and *all*, they interpret numerals with an *exact*-meaning just like the adult controls.

Previous studies have suggested factors that could affect children's derivation of implicatures, including limitations on working memory, limited understanding of context and meta-linguistic tasks, and the salience or availability of relevant scalar alternatives (see Chierchia et al., 2001; Papafragou & Tantalou, 2002; Pouscoulous et al, 2007; Reinhart, 2004). According to Papafragou and Musolino (2003), since each of these factors might limit children's computation of implicatures, and since children readily assign exact interpretations to numerals, children must not be using implicatures to derive exact meanings of numerals. Instead, by their view, the difference between quantifiers and numerals is due to the fact that numerals have lexically strengthened, exact meanings (see also Huang, Snedeker, & Spelke, under review).

Context clearly affects whether children (and adults) will compute implicatures (e.g., Papafragou & Musolino, 2002). It is also well established that working memory capacity grows over the course of development (e.g., Gathercole & Baddley, 1990). Nevertheless, the role of these factors in children's pragmatic difficulties has not been empirically established. First, although previous studies find that implicatures are more likely in some contexts than others (e.g., Papafragou & Musolino, 2002), the fact that strong contextual cues can push children towards one interpretation over another does not demonstrate that their difficulties are due to contextual misunderstanding. For example, strong contextual cues may compensate for difficulties that originate elsewhere in the process of deriving implicatures.

Second, there is currently no direct evidence that processing constraints are responsible for limiting children's implicatures. Studies that attribute their problems to processing limitations (Chierchia et al., 2001; Pouscoulous et al., 2007; Reinhart, 2004) do not actually assess working memory, nor do they demonstrate that individual differences in processing capacity predict differences in pragmatic abilities. For example, Chierchia et al. (2001) tested 3- to 6-year-old children's interpretation of *or*. Unlike adults, when children were told, "Every boy chose a skateboard or a bike," they accepted situations in which a boy chose both objects. Thus, they accepted the weak inclusive interpretation of *or*, when adults did not. However, when explicitly presented with a sentence containing *and* as an alternative, children strongly preferred it over a sentence containing *or*. This study shows that when children are presented with explicit scalar alternatives, they know when to use the stronger statement. However, it does not single out working memory as the source of children's difficulty. Instead, we suggest that it is also consistent with the idea that children lack knowledge of scales, and which words are activated as relevant alternatives during interpretation (Step II, see also Papafragou & Tantalou, 2004). An inability to generate relevant scale-mates could explain numerous failures in the literature, as well as the apparent discrepancy between children's difficulty with implicatures and their relatively sophisticated use of pragmatic cues elsewhere in language acquisition. Further, this account, as noted by Barner and Bachrach (2010), could explain children's ability to assign exact interpretations to numerals, which belong to an explicitly memorized set of alternatives – the count list.

Barner and Bachrach (2010) argued that young children routinely make inferences that are similar in structure to scalar implicatures when interpreting unknown numerals. As noted earlier, when a child who knows the meaning of *one* is shown two sets – e.g., one containing one balloon, and the other containing five – they systematically point to the larger set when asked to find *five balloons*. However, they do not do so when asked to find *blicket balloons*. According to Wynn (1992), "Since all the children knew that the word 'one' refers to a single item, then if they knew that, for example, the word 'five' refers to a numerosity, they should infer that it does not refer to a single item since they already have a word for the numerosity one." (p. 229).

This inference – that *five* refers to the larger set by virtue of *not* referring to *one* – requires all of the processing resources that an implicature would require, as well as several of the same steps. The child must generate a weak meaning for *five* (Step I), generate *one* as an alternative (Step II), and strengthen the interpretation of *five* by negating *one* (Step IV). The only missing component of implicature is that weaker items are not exhaustified by appeal to stronger ones (this would be impossible here, since stronger numeral words have not yet been acquired). Still,

once children acquire a meaning for *two*, they should be in a position to compute an implicature for *one*, meaning that even 2-year-olds could compute implicatures to derive exact meanings for numerals.

Children do not have difficulty accessing *one* as a relevant alternative to *five*. Also, once children have accessed *one* as an alternative, they appear capable of inferences not far from a full-fledged scalar implicature. These facts suggest that children's failure to compute implicatures for other scales may be due to a failure to generate relevant scalar alternatives. While children begin to explicitly memorize a count list well before they learn any numeral meanings (see Fuson, 1988), no child is taught to recite a list of quantifiers.

The present study tested the hypothesis that children's difficulty computing implicatures is caused by a failure to generate relevant alternatives. We asked whether children could strengthen their interpretation of utterances containing the quantifier *some* when used with the focus word *only*. In English, the algorithm for calculating scalar implicatures is grammatically mirrored by the semantics of *only*, a fact that allows us to isolate the role of access to alternatives in implicature. As with implicatures, *only* triggers the negation of alternative sentences. For example, consider the sentence in (5).

(5) I ate only some of the cake.

This sentence indicates that the speaker did not eat all of the cake, like Mary's statement in (1). The difference between the sentences in (1) and (5) is that in (5) the denial of the alternative "*I ate all of the cake*" is logically entailed by the sentence's core, literal meaning (it is not merely implied). Still, in order for this entailment relation to be realized, the listener must access *all* as a relevant alternative and negate it. Therefore, evidence that children comprehend *only* but fail to strengthen sentences containing *only some* would suggest that their difficulty is caused by a failure to access scalar alternatives.

To manipulate the accessibility of alternatives, we contrasted children's interpretation of *some*, whose scale members are specified in a context-independent way, with their interpretation of words that have contextually specified alternatives (for discussion, see Hirschberg, 1985). Previous studies find that young children are able to strengthen utterances that rely on contextual alternatives. For example, Goro, Minai & Crain (2006) found that children rejected sentences like "Only Bunny Rabbit will eat a carrot or a pepper" in contexts where another character ate a pepper.

We tested the hypothesis that children's difficulty is due to difficulty generating relevant alternatives by (1) manipulating the availability of alternatives by contrasting utterances that involve context-independent scales like <*some*/*all*> to utterances that draw on contextually specified sets of alternatives, and (2) forcing the exhaustification of utterances by including the focus element *only* in sentences. Critical trials in the experiment presented situations involving three things (e.g., three animals sleeping), and asked children to evaluate one of the questions in (6):

(6) a. Are some of the animals reading?
b. Are only some of the animals reading?
c. Are the cat and the dog reading?
d. Are only the dog and the cat reading?

If children's difficulty computing implicatures for context-independent scales is due to a failure to access alternatives, then they should accept statements like (6a) and (6b) regardless of whether *only* is present. They should fail to construct an alternative sentence containing *all*, and therefore be unable to strengthen either sentence. In contrast, children should have no difficulty strengthening a sentence like (6d), since the alternative contrast set is contextually specified and therefore readily available.

Method

Participants

Sixty 4-year-olds (M=53.94 months, age range: 48.7–59.8 months) participated in this experiment. Two additional children were excluded due to failure to complete the task.

Stimuli

Stimuli were twelve picture cards, each depicting a scene of three items. Four cards were used in a familiarization phase, and eight in the test phase. Familiarization cards depicted sets of animals with distinct characteristics, such as color or clothing. The test cards depicted four scenes (in 1 – 4).

(1) Cookie Monster holding fruit (an orange, an apple, and a banana)
(2) Animals sleeping (a dog, a cat, and a cow)
(3) Animals reading (a dog, a cat, and a rabbit)
(4) Toys on a table (a ball, a drum, and a train)

Two versions of each scene were created: one in which all three items shared a property (e.g., Cookie Monster is holding all three fruits), and one in which two of the three items shared the property (e.g., Cookie Monster is holding two fruits, and one is on the floor). An example is provided in Figure 1.

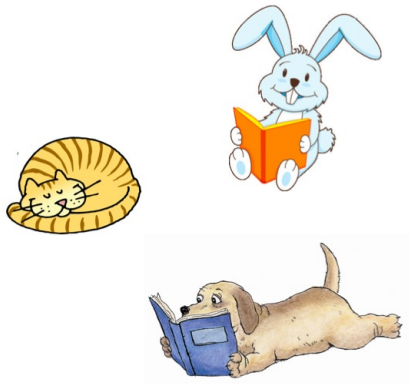


Figure 1: Example test stimulus card.

Procedure

Children were first shown the familiarization cards one at a time and asked to identify each animal ("What's this? That's right, it's a cow!"). If the child labeled an animal incorrectly, they were given the correct label and encouraged to repeat it ("That's a cow, can you say 'cow?").

Children were then asked a question about the scene (e.g., "Is the cow wearing a hat?" when the *fish* is wearing a hat). This exercise was designed to accustom children to answering both yes and no to questions. If a child answered any question incorrectly, the experimenter moved on to the next familiarization card, but returned to the problematic card after completing the remaining familiarization trials. If a child failed twice on any single familiarization trial, the experimenter ended the testing session.

At test, children were given nine trials using the test cards, presented in one of two counterbalanced orders. Again, children were asked to identify all of the items in the picture, and then to evaluate the truth-value of a statement.

Each child participated in one of four conditions. In Conditions 1 and 3 (*Context-Independent Alternatives)*, children were asked questions that required them to evaluate the meanings of the quantifiers *some* and *all*, e.g. *Is Cookie Monster holding some / all of the fruits?* In Conditions 2 and 4 (*Contextual Alternatives*), the individual animals, fruits, etc. were labeled separately, e.g., *Is Cookie Monster holding the banana, the apple and the orange?* The questions in each condition were identical, except that in Conditions 3 and 4 (*Grammatically Exhaustified* conditions) the word *only* was inserted– e.g., *Are only some of the animals reading?* or *Are only the dog and the rabbit reading?*

There were 9 questions in each condition. On "2-item, False" questions, children were shown a picture in which only two of the three items fit a description, and were asked whether the description was true for all the items (e.g., *Is Cookie Monster holding all of the fruits?* or *Is Cookie Monster holding the apple, the orange, and the banana?*). These were used as control trials, to be sure that children were attending to the task, and were presented identically in conditions 1 and 2 and conditions 3 and 4. On "2-item, True" questions, children saw pictures where two of the three items fit a description, and were asked whether the description was true for a subset of the items (e.g. *Are (only) some of the animals reading?* or *Are (only) the rabbit and the dog reading?*). Lastly, on "3-item, Test" questions, children were shown pictures in which all three items fit the description, and asked whether the description was true for a subset of the items (questions were identical in the 2-item, True and 3-item, Test trials).

Neither the word *only* nor the quantifier was emphasized by the experimenter's prosody.

Results

The use of the word *only* had a significant effect on how children interpreted sentences involving contextual alternatives, but had no effect on their interpretation of sentences involving context-independent alternatives (*some* and *all*). A 2x2x2 repeated measures ANOVA was conducted with Trial Type ("2-Item True" vs. "3-Item Test") as a within-subjects variable and Scale Type (context-independent vs. contextual) and Grammatical Exhaustification (*only* vs., *no-only*) as between-subjects variables. Two-Item False Trials were excluded from this analysis as children were expected to reject these sentences (results for these trials are described below).

Overall, children were significantly more likely to accept sentences on 2-Item True Trials (87.9%), than on 3-Item Test Trials (59.6%), $F(1,56)=37.05$, $p<.001$. They were also less likely, overall, to accept sentences with *only*, such as "only the drum and the ball are on the table" (84.8% of trials) than those that did not contain *only* (62.7% of trials, $F(1,56)=672.2$, $p<.001$). There was no main effect of Alternative Type ($p>0.05$). Crucially, there were two-way interactions between Alternative Type and Grammatical Exhaustification ($F(1,56)=13.74$, $p<.001$), Trial Type and Grammatical Exhaustification ($F(1,56)=15.08$, $p<.001$),Trial Type and Alternative Type ($F(1,56)=8.87$, $p<.01$), and a three-way interaction between Trial Type, Alternative Type, and Grammatical Exhaustification ($F(1,56)=13.28$, $p<.001$). These interactions were due to the fact that *only* had a significant effect on children's judgments only for contextual alternatives, and only on 3-item Test trials.

Figure 2 shows the percentage of children who said "yes" to questions in the contextual alternatives conditions. In contexts involving 2 items (e.g., Cookie Monster holding an apple and a banana), children correctly agreed to sentences like, "Is cookie monster holding the apple and the banana?" on 95.8% of trials, and correctly denied that he was holding "the apple, the banana, and the orange" on 80.5% of trials. As expected, adding the word *only* had no effect on either trial type ($ps>.05$). In contrast, on critical 3-item Test trials, children tested with contextual alternatives were highly sensitive to the presence of *only*. These children said "yes" when asked, "Is cookie monster holding the apple and the banana?" on 92.9% of trials, but rarely said "yes" when *only* was added: "Is cookie monster holding only the apple and the banana?" (14% of trials; $t(28)=8.98$, $p<.001$).

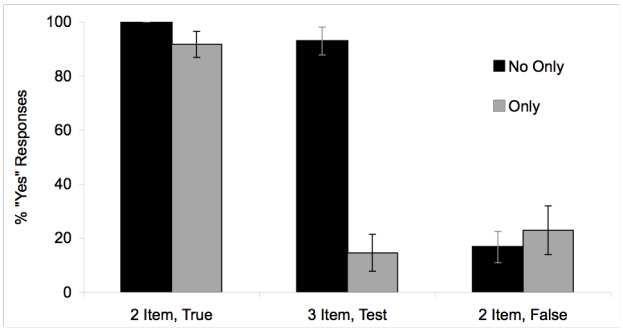


Figure 2: Percentage of children who said "yes" to sentences in contextual alternatives conditions.

When children were tested with the word *some* (Context-Independent Alternatives conditions), they also correctly said "yes" on 2-item true trials (80.0% of trials), and correctly said "no" 2-item false trials (87.2% of trials). The insertion of *only* again had no effect on children's responses for these trial types ($ps>0.05$). As in other studies, children did not strengthen utterances containing *some* in absence of *only*. There was no significant difference in children's response for 2-item True trials and 3-item Test trials

($t(14)=1.0$, $p>.3$). For example, children were equally likely to agree that *some* animals were reading when all three of them were, relative to when only two were. The insertion of *only* did not improve matters and had no effect on the 3-item test trials ($t(28)=.16$, $p>.8$). For example, when three animals were reading, children were equally likely to say "yes" when asked, "Are some of the animals reading" and "Are only some of the animals reading". Thus, whereas *only* had a huge impact on children's interpretation of utterances including contextual alternatives, it had no effect at all when children interpreted utterances containing the word *some*.

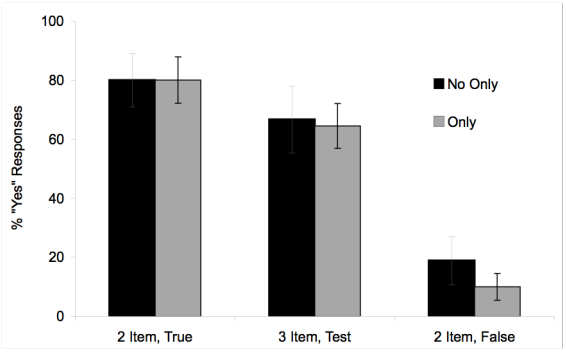


Figure 3. Percentage of children who said "yes" to questions in context-independent conditions.

Discussion

Children's ability to generate scalar alternatives places a significant constraint on their ability to compute scalar implicatures. In this study, children assigned strengthened interpretations to utterances when they included the focus element *only*, if alternatives were provided contextually. For context-independent scales (e.g., *some/all*) children failed to compute implicatures, even when *only* was added. Since *only* forces exhaustification grammatically (and did so for contextual alternatives), children's failure to derive strengthened readings for *some* must be attributed to a failure to generate relevant scalar alternatives – in this case the quantifier *all*.

These data also suggest, contrary to speculation in the literature (e.g., Chierchia et al, 2001; Pouscoulous et al., 2007) that children's difficulties are not purely attributable to processing constraints. Children were perfectly capable of deriving strengthened interpretations for utterances that involved contextual alternatives, but failed for identical sentences that involved *some*. This result is not predicted by theories that posit processing limitations, since the sentences in these conditions did not differ in grammatical complexity. They only differed with respect to the type of scale that they implicated. The possibility that there are differences in processing difficulty between these conditions cannot be completely ruled out. However, no previous study has provided direct evidence that children's failures are related to processing limits or working memory. Also, as they are presented, the previous accounts of processing limits are too vague to explain why they would affect Horn scales but not context dependent scales. Thus, we see no compelling reason to conclude that processing limits are at the root of children's difficulties on this task or with scalar implicature generally. Instead, we submit, children lack *knowledge* that is scale-specific – i.e., they lack the knowledge that *all* is a relevant alternative when interpreting *some*.

Our hypothesis – that children interpret "only" like adults but fail to compute scalar implicatures because they lack knowledge of specific scales – allows us to explain a much wider array of data than previous accounts, while explaining why children appear pragmatically sophisticated in some domains but not in others. As noted in the introduction, previous studies of children's number word acquisition find that children can make inferences that resemble scalar implicature from a very early age (Wynn, 1992; Condry & Spelke, 2008). These inferences – e.g., that *five* cannot refer to sets of one, because *one* does –involve processes similar to those needed for scalar implicature. (see Barner & Bachrach, 2010).

Children's ability to make such inferences for numerals and contextual scales, but not for scales like <*some, all*>, points to differences in scale-specific knowledge. In the case of contextual scales, no scale-specific learning is required since these scales are constructed on the fly in context. In the case of number words, children begin acquisition by learning numerals as an ordered list of alternatives. They acquire a partial count list *before* learning any individual numeral meanings (for review, see Carey, 2009). Thus, the first thing that children learn about the numeral *five* is that it is a member of the count list. In contrast, normal children never learn to recite a sequence of quantifiers like *some, many, most, all*, etc. This view of acquisition suggests, contrary to previous reports (e.g., Papafragou & Musolino, 2003; Huang, Snedeker, & Spelke, under review) that children may derive exact meanings of early numerals via scalar inference (by contrasting numerals with one another).

The idea that children's difficulties are scale-specific, rather than due to pragmatic immaturity, is also consistent with reports of pragmatic sophistication in other domains, such as noun learning (see Baldwin, 1993; Clark, 1987, 1988; Markman, 1989; Tomasello, 1992). For example, when shown a novel object next to an alternative with a known label – e.g., a shoe – children readily infer that a novel label like *blicket* must refer to the new object (Clark, 1987, 1988; Markman, 1989). Similarly, children infer that a novel color word, like *chromium*, must refer to a novel color, and not to known colors like *red* or *blue* (Carey & Bartlett, 1978). Children fail to respect mutual exclusivity if they believe the novel word is not at the same level of description as the label for the known object, or if they are told the word is from another language (Au & Glusman, 1990). In these cases, the known label is not considered a relevant alternative to the novel label. These simple inferences, though distinct from implicatures in many ways, nonetheless require both pragmatic understanding (including ascription of speaker intent), and the processing abilities needed to entertain and restrict possible alternatives. These abilities would be difficult to explain if children's

difficulties with scalar implicature were due to processing limits or a general insensitivity to pragmatics.

What must children learn about scales to use them for implicature? Clearly children must learn the meanings of scale mates, and how these meanings differ in informational strength in different contexts. At 4 years of age, children easily differentiate meanings like *some* and *all*, and are able to correctly choose stronger descriptions over weaker ones when provided with a forced choice (e.g., Chierchia et al., 2001). Children's difficulty, it seems, is in recognizing that, for communicative purposes, these scale mates are alternatives to one another – i.e., that using one implies that the others are not true. Thus, a failure to generate words as alternatives does not mean that children have difficulties with lexical retrieval. Rather, our claim is that even when children can retrieve *all* when interpreting *some*, they do not access it *as a relevant alternative* to *some*.

A remaining puzzle, and one that is not addressed by the current study, is how children eventually come to acquire such scales. Our results, and others from the literature, suggest that children are capable of strengthening utterances by appeal to alternatives, so long as these alternatives are contextually specified or memorized explicitly as a list. It is not clear children they come to associate scale mates, such as quantifiers, that they do not learn as a list. We suggest that the association of these lexical items may take place by trial and error learning – by hearing words used contrastively in context, or via explicit cancellations of implicature in the speech of adults. Future studies should explore the effects that such input has on children's pragmatic reasoning, and how experience with different scales affects their ability to compute implicatures.

References

Au, T. K., & Glusman, M. (1990). The principle of mutual exclusivity in word learning: To honor or not to honor? *Child Development, 61,* 1474-1490.

Baldwin, D. A. (1993). Infants' ability to consult the speaker for clues to word reference. *Journal of Child Language, 20,* 395–418.

Barner, D., Bachrach, A. (2010). Inference and exact numerical representation in early language development. *Cognitive Psychology, 60,* 40-62.

Barner, D., Chow, K., & Yang, S. (2009). Finding one's meaning: A test of the relation between quantifiers and integers in language development. *Cognitive Psychology, 58,* 195–219.

Carey, S. (2009). *The Origin of Concepts*. New York: Oxford University Press.

Chierchia, G., Crain, S., Guasti, M. T., Gualmini, A., & Meroni, L. (2001). The acquisition of disjunction: Evidence for a grammatical view of scalar implicatures. In A. H.-J. Do, L. Domingues, & A. Johansen (Eds.), *Proceedings of the 25th Boston University Conference on Language Development*). Somerville, MA: Cascadilla Press.

Clark, E. (1987) The *principle* of *contrast*: A constraint on language acquisition. In B. MacWhinney, Ed., *Mechanisms of language acquisition*

Condry, K. F., & Spelke, E. S. (2008). The development of language and abstract concepts: The case of natural number. *Journal of Experimental Psychology: General, 137(1),* 22-38.

Crain, S., Goro, T., & Minai, U. (2007). Hidden units in child language. In Shalley, A., & Khlentzos, D. (Eds.), *Mental states, Volume I: Evolution, Function, Nature* (pp. 275-294). Amsterdam: John Benjamins.

Fuson, K.C. (1988). *Children's Counting and Concepts of Number*. New York: Springer-Verlag.

Gathercole, S.E., & Baddeley, A.D. (1990). Phonological memory deficits in language disordered children: Is there a causal connection? *Journal of Memory and Language, 29(3),* 336-360.

Goro, T., Minai, U., & Crain, S.B. (2006). Bringing out the logic in child language. In Bateman, L., & Ussery, C. (Eds.), *Proceedings of the 35th Annual Meeting of the North East Linguistic Society* (pp. 245-256). Amherst, MA: GLSA Publications.

Grice, P. (1989). *Studies in the Way of Words*. Cambridge, MA: Harvard University Press.

Hirschberg, J. (1985). *A theory of scalar implicature.* Doctoral dissertation, University of Pennsylvania, Philadelphia, PA.

Horn, L. (1989). *A Natural History of Negation*. Chicago: University of Chicago Press.

Huang, Y., Snedeker, J. & Spelke, E. (under review). What exactly do numbers mean?

Huang, Y. & Snedeker, J. (2009). Online interpretation of scalar quantifiers: Insight into the semantics–pragmatics interface. *Cognitive Psychology, 58(3)*, 376-415.

Markman, E.M. (1989). *Categorization and naming in children*. Cambridge, MA: MIT Press.

Noveck, I. A. (2001). When children are more logical than adults: Experimental investigation of scalar implicatures. *Cognition, 78,* 165–188.

Papafragou, A., & Musolino, J. (2003). Scalar implicatures: Experiments at the semantics–pragmatics interface. *Cognition, 86,* 253–282.

Papafragou, A., & Tantalou, N. (2004). Children's computation of implicatures. *Language Acquisition, 12,* 71–82.

Pouscoulous, N., Noveck, I. A., Politzer, G., & Bastide, A. (2007). A developmental investigation of processing costs in implicature production. *Language Acquisition, 14,* 347–375.

Reinhart, T. (2004). The Processing Cost of Reference-Set Computation: Acquisition of Stress Shift and Focus. *Language Acquisition, 12*(2): 109-155.

Tomasello, M. (1992). The social bases of language development. *Social Development, 1,* 67–87.

Wynn, K. (1990). Children's understanding of counting. *Cognition, 36*, 155-193.

Wynn, K. (1992). Children's acquisition of number words and the counting system. *Cognitive Psychology, 24*, 220–251.

Developmental and computational perspectives on infant social cognition

Noah D. Goodman (ndg@mit.edu)
Chris L. Baker (clbaker@mit.edu)
Tomer D. Ullman (tomeru@mit.edu)
Joshua B. Tenenbaum (jbt@mit.edu)
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology

Kiley Hamlin (kiley.hamlin@yale.edu)
Karen Wynn (karen.wynn@yale.edu)
Paul Bloom (paul.bloom@yale.edu)
Department of Psychology
Yale University

Chris G. Lucas (clucas@berkeley.edu)
Thomas L. Griffiths (tom_griffiths@berkeley.edu)
Fei Xu (fei_xu@berkeley.edu)
Department of Psychology
University of California, Berkeley
Christine Fawcett (christine.fawcett@mpi.nl)
Max Planck Institute for Psycholinguistics

Tamar Kushnir (tk397@cornell.edu)
Department of Psychology
Cornell University
Henry Wellman (hmw@umich.edu)
Susan Gelman (gelman@umich.edu)
Department of Psychology
University of Michigan

Elizabeth Spelke (spelke@wjh.harvard.edu)
Department of Psychology
Harvard University

Keywords: Social cognition; Cognitive Development; Computational Modeling; Theory of Mind

Adults effortlessly and automatically infer complex patterns of goals, beliefs, and other mental states as the causes of others' actions. Yet before the last decade little was known about the developmental origins of these abilities in early infancy. Our understanding of infant social cognition has now improved dramatically: even preverbal infants appear to perceive goals, preferences (Kushnir, Xu, & Wellman, in press), and even beliefs from sparse observations of intentional agents' behavior. Furthermore, they use these inferences to predict others' behavior in novel contexts and to make social evaluations (Hamlin, Wynn, & Bloom, 2007).

Inspired by this work, computational modelers have in the last few years begun to formalize the knowledge and inference mechanisms underlying infants' social reasoning (Baker, Saxe, & Tenenbaum, 2009; Lucas, Griffiths, Xu, & Fawcett, 2009; Ullman et al., 2010). Many of these models share deep similarities, explaining social inference in terms of an intuitive understanding of how an agent chooses among actions. For instance, the principle of rational action, suggested in seminal work on infant social cognition (Gergely, Nádasdy, Csibra, & Biró, 1995), states that agents will select the best action to achieve their goals, given the constraints of their environment – or in a more sophisticated version, given their beliefs about the environment. This principle has been formalized using notions of planning and decision-making from economics and computer science. It underlies models that make accurate quantitative predictions of the social inferences of adults and young children in a variety of experimental tests.

The goal of this symposium will be to bring together developmental psychologists and computational modelers in a dialogue on the social inferences made by young infants, the mechanisms by which these inferences work and become more sophisticated in older children. The first talk of the symposium (Baker et. al) will briefly survey now-classic work on infants' understanding of goals and beliefs, and will introduce a general computational framework for modeling these social inferences based on intuitive principles of rational action. Next will be two pairs of developmental and computational talks, focusing on recent advances where there has been important exchange between empirical work and models. Kushnir, et al, and Lucas, et al, will describe work on understanding of others' preferences. Hamlin, et al, and Ullman, et al, will describe attribution of "prosocial" goals (such as helping). The symposium will conclude with a discussion led by Spelke, highlighting gaps in our understanding of infant social cognition, areas where more computational work is needed, and where computational ideas might suggest new areas for developmental experiments.

Close interaction and collaboration between developmentalists and computational modelers studying infant social cognition is a fairly recent trend, yet it has already proven fruitful, as the talks in this symposium hope to demonstrate. Previously, the research to be presented here has been discussed primarily at conferences on computational modeling (e.g., NIPS) or developmental psychology (e.g., the Cognitive Development Society), or in small workshops bringing together modelers and experimentalists. The Cognitive Science Conference would be an ideal venue for a broad symposium on this emerging, interdisciplinary subfield, due to its tradition of bringing together theorists and experimentalists from a broad array of disciplines. We expect the symposium will interest a wide audience and lead to new research directions and collaborations engaging different segments of the Cognitive Science audience.

Probabilistic models of belief-desire psychology

Baker, Goodman & Tenenbaum We propose a computational

framework for modeling how humans interpret intentional actions in terms of the mental states that cause behavior: chiefly, beliefs and desires. The framework represents a schema for intentional action using rational models of belief- and goal-based planning from economics and computer science known as partially observable Markov decision problems. Agents' beliefs and desires are inferred by inverting this model of rational planning using Bayesian inference, integrating the likelihood of the observed actions with the prior over mental states. This approach formalizes in precise probabilistic terms the essence of previous qualitative approaches to infant action understanding, (*e.g.* Gergely et al., 1995). We will present results showing that our models account for infants' and adults' social judgments from a body of experiments, from simple inferences about goals, to joint inferences of preferences and beliefs. We will also consider how a set of alternative, heuristic-based models compare to our account.

Young children use statistical sampling to infer the preferences of others

Kushnir, Wellman & Gelman Psychological scientists use statistical information to determine the workings of fellow humans. We argue so do young children. In a few years, children progress from viewing human actions as intentional and goal-directed to reasoning about the psychological causes underlying such actions. Here we show that preschoolers and 20-month-old infants can use statistical information – namely, a violation of random sampling – to infer that an agent is expressing a preference for one object over another. Children saw a person remove 5 items of one type from a container of objects. Preschoolers and infants only inferred a preference for that type of object when there was a mismatch between the sample and population. Mere outcome consistency, time spent with and positive attention toward the objects did not lead children to infer a preference. The findings provide an important demonstration of how statistical learning could underpin the rapid acquisition of early psychological knowledge.

A rational model of preference learning and choice prediction by children

Lucas, Griffiths, Xu & Fawcett We present a rational model of preference learning that explains the behavior of children in several recent experiments, as well as a developmental shift in which children come to understand that people have distinct preferences. We first show that a simple econometric model can account for young children's use of statistical information in inferring preferences and their ability to generalize others' preferences from one category to another. We then consider the question of how children begin to treat other individuals as having preferences that can differ from their own, showing that such a transition is consistent with Bayesian inference, given a model in which all people share preferences and one in which preference can vary as possibilities. Finally, we discuss novel predictions made by our model concerning preference understanding and the developmental shift.

The enemy of my enemy is my friend: Infants interpret social behaviors in context

Hamlin, Wynn & Bloom Recent research suggests that young infants prefer prosocial to antisocial individuals (Hamlin et al., 2007). While a preference for those who help others is certainly adaptive, there are potentially situations in which unhelpful behavior is more appropriate (*e.g.* punishing others for their wrongdoing) or more socially diagnostic (*e.g.* "The enemy of my enemy is my friend," Aronson & Cope, 1968). This talk examines whether infants always prefer those who are prosocial, in contexts in which antisocial behavior could be seen as punishment, or in which an individual's antisocial behavior may be an indication that he or she shares a negative opinion toward a disfavored other. Results suggest that even in the first year of life, infants evaluate behaviors not only in terms of their valence, but also in terms of certain qualities of their recipients.

Help or hinder: Models of social goal inference

Ullman, Baker, Goodman & Tenenbaum Everyday social interactions are heavily influenced by our snap judgments about others' goals. Even young infants can infer the goals of intentional agents from observing how they interact with objects and other agents in their environment: e.g., that one agent is 'helping' or 'hindering' another's attempt to get up a hill or open a box. We propose a model for how people can infer these social goals from actions, based on inverse planning in multiagent Markov decision problems. The model infers the goal most likely to be driving an agent's behavior by assuming the agent acts approximately rationally given environmental constraints and its model of other agents present. We also present behavioral evidence in support of this model over a simpler, perceptual cue-based alternative.

Discussion: Open challenges and future directions

Spelke The closing discussion will draw out gaps in our current understanding of infant social cognition, areas where more computational work is needed, and places where computational ideas might suggest new areas for developmental experiments.

References

Aronson, E., & Cope, V. (1968). My enemy's enemy is my friend. *Journal of Personality and Social Psychology*, *8*, 8–12.

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*, 329-349.

Gergely, G., Nádasdy, Z., Csibra, G., & Biró, S. (1995). Taking the intentional stance at 12 months of age. *Cognition*, *56*, 165–193.

Hamlin, J. K., Wynn, K., & Bloom, P. (2007). Social evaluation by preverbal infants. *Nature*, *450*, 557–560.

Kushnir, T., Xu, F., & Wellman, H. (in press). Young children use statistical sampling to infer the preferences of others. *Psychological Science*.

Lucas, C., Griffiths, T. L., Xu, F., & Fawcett, C. (2009). A rational model of preference learning and choice prediction by children. *Advances in Neural Information Processing Systems (NIPS) 21*.

Ullman, T., Baker, C., Macindoe, O., Evans, O., Goodman, N., & Tenenbaum, J. (2010). Help or hinder: Bayesian models of social goal inference. *Advances in Neural Information Processing Systems (NIPS) 22*.

Linking Meaning to Language: Linguistic Universals and Variation

Joshua K. Hartshorne (jharts@wjh.harvard.edu)
Department of Psychology, Harvard University
33 Kirkland St., Cambridge, MA 02138

Timothy J. O'Donnell (timo@wjh.harvard.edu)
Department of Psychology, Harvard University
33 Kirkland St., Cambridge, MA 02138

Yasutada Sudo (ysudo@mit.edu)
Department of Linguistics and Philosophy, Massachusetts Institute of Technology
77 Mass. Ave. 32-D808, Cambridge, MA 02139

Miki Uruwashi (mikiuruwashi@ruri.waseda.jp)
Graduate School of Human Sciences, Waseda University
33 Kirkland St., Cambridge, MA 02138

Jesse Snedeker (snedeker@wjh.harvard.edu)
Department of Psychology, Harvard University
33 Kirkland St., Cambridge, MA 02138

Abstract

To use natural language, speakers must map the participants in events or states in the world onto grammatical roles. There remains considerable disagreement about the nature of these so-called *linking rules* (Levin & Rappaport Hovav, 2005). In order to probe the nature of linking rules, we investigate verbs of psychological state, which demonstrate complex linking patterns both within and between languages. We find that the typical duration of the psychological state guides the application of linking rules to novel verbs in both English and Japanese, consistent with a universal constraint. Nonetheless, there are marked differences in the baseline preferences for the individual linking rules across the two languages. We discuss these findings both in terms of theories of exceptionless linking rules and accounts on which linking rules are governed by probabilistic biases as well as cross-linguistic variation.

Keywords: syntax; semantics; linking; UTAH; universal grammar; over-hypotheses.

The Linking Problem

To interpret *Mary broke the vase*, one must minimally identify the event described (*breaking*), the participants in that event (*Mary*, *vase*), and identify which participant played which role (*Mary* = breaker, not break-ee). This *linking problem* has received considerable attention both by theorists trying to correctly characterize the semantics-syntax links (see Levin & Rappaport Hovav, 2005, for review), and by developmental psychologists interested in how children discover these links (Bowerman, 1990; Pinker, 1984, 1989).

A key issue is identifying the right level of generalization for the *linking rules*. Many data points suggest linking rules are highly regular. Regularity is seen both within verbs and across verbs. Not just *Mary* but all breakers are the subject and not object of *break* (*John/the baby/the wind broke the vase/window/glass*). Similarly, in English the object of a transitive change-of-state verb is systematically the entity that changes state while the subject effects that change (*Mary broke/cleaned/opened the box*). These intuitions generalize to novel words. If interpretable, *The dax broke the blicket* must mean that the dax is the breaker and the blicket is broken. Adults and children prefer an interpretation on which *The bear pilked the horse* means the bear did something to the horse, not vice versa (Marantz, 1982; see also Pinker, 1989). Moreover, these patterns are sufficiently regular across languages to suggest that some (Pinker, 1984) or all (Baker, 1988) linking rules are innate.

However, there are numerous examples of apparent variation and exceptionality. An object moving from Mary's possession to John's can be described by *Mary gave/lent/sent the package to John* or *John received/took/obtained the package from Mary*. The same activity might be called *Mary chasing John* or *John fleeing Mary*. Many emotion verbs put the experiencer in subject position (*John feared/hated/loved Mary*), while others put the experiencer in object position (*Mary frightened/angered/delighted John*). Moreover, a relatively small number of languages appear to exhibit linking rules quite distinct from what is seen in languages like English (Dixon, 1994).

In the present study, we investigate linking rule regularity and variation within and across two unrelated languages with respect to one such problematic case: psych verbs.

Psych Verbs

Unlike change-of-state verbs, verbs of psychological state are highly variable in terms of their surface syntax. The experiencer of the mental state may appear as the verb's subject (experiencer-subject verbs: *Mary likes/hates/misses John*) or its object (experiencer-object verbs: *Mary surprises/confuses/angers John*). Both classes are seen in a wide variety of languages, though the subjects of experiencer-subject (ES) verbs can appear as dative subjects in languages that have such constructions (Levin, 1993). Interestingly, there appears to be some variation across languages in terms of which psychological states appear in which form: for instance, the apparent French equivalent to the ES psych verb *miss* is experiencer-object (EO; *manquer*; see also Croft, 1993).

Most authors have assumed there is no systematic semantic distinction between ES and EO verbs, and thus each verb must be marked for taking one linking rule or the other (e.g., Bowerman, 1990; Dowty, 1991; Jackendoff, 1990; Pinker, 1989). However, Pylkkanen (1999) finds that in Finnish, ES psych verbs describe individual-level predicates whereas EO psych verbs describe stage-level predicates.[1] Stage-level and individual-level predication differ in several ways; one relevant difference is that stage-level predicates can be narrowly bound temporally and physically (1), whereas individual-level states typically cannot be (2).

(1) a. John was sleepy yesterday in the kitchen.
b. John angered Mary yesterday in the kitchen.
(2) a. *John was tall yesterday in the kitchen.
b. *John hated Mary yesterday in the kitchen.

Thus, it may be that those psychological states which are deemed more likely to be bound in time and space are also more likely to be realized as EO verbs.

Interestingly, the psychological literature on emotional states typically distinguishes between *emotions* and *dispositions* (Ekman, 1999). The former are tied to specific physiological states and are brief in duration, whereas the latter are long-lived tendencies to feel or act in a particular way. Commonly-given examples of emotions are *surprise* and *anger*; frequent examples of dispositions are *love* and *hate*. Note that the former are EO verbs and the latter ES verbs.

Informal inspection of English psych verbs by the authors suggested that in fact ES verbs do typically describe dispositions thus defined while EO verbs typically describe emotions. This was further confirmed in an unpublished study in which naïve participants rated the states described by ES verbs as typically lasting longer than those described by EO verbs (Hartshorne, 2009).

In the present study, we investigate whether differences in the nature of the psychological state influence whether participants apply the EXPERIENCER→SUBJECT linking rule or the EXPERIENCER→OBJECT linking rule to novel psych verbs. We focus on the notion of duration: are long-lived psychological states (dispositions/individual-level states) more likely to be realized as ES verbs relative to short-lived psychological states (emotions/stage-level states)?

In order to investigate both linguistic universals and variation, we investigated the degree to which this proposed distinction guides generalization of linking rules in two historically unrelated and linguistically distinct languages: English and Japanese.

Experiment 1: English

Participants in Experiment 1 were introduced to novel transitive verbs describing psychological states for which there was no existing verb. To encourage participants to take the task seriously, the novel verbs were introduced as loanwords from Japanese. Half the verbs described long-lived psychological states; half described short-lived psychological states. For each verb, participants decided whether an ES structure or an EO structure was more likely to be "correct."

In English there is a preference for simple present tense verbs to be interpreted as generic statements (contrast *Bats frightened John* vs. *Bats frighten John*; see Carlson, 1988). As this may affect whether the novel psych verbs are seen to describe short-lived (event-like) or long-lived states, we tested separate groups of subjects using both simple present and past tenses (Experiments 1a and 1b). As ES verbs cannot be naturally used the progressive form (**John was fearing bats*), we used simple tense only.

Method

Participants

Forty native English-speakers participated in Experiment 1: twenty in 1a (18-60yo, M=25.3, SE=2.2) and twenty in 1b (18-39yo, M=23.1, SE=1.2). Participants, who were recruited outdoors on Harvard's campus, gave informed consent and were compensated with a small snack.

Materials

Sixteen Japanese nouns describing psychological states without clear English verbal equivalents were selected and turned into verbs, applying any necessary phonological accommodations. Eight were judged by the authors to be long-lived states (e.g. *tekitaishin: the feeling of rivalry; hankan: the feeling of being opposed to something or someone*) and eight to be short-lived states (e.g., *wabi: a sense of beauty of silence discovered in simplicity; tokimeki: the feeling of a heart beating because of encountering an attractive person or thing*). For each verb, an appropriate animate experiencer argument was chosen. The other argument was an inanimate *stimulus* of the emotion. Two sentences were constructed by placing the experiencer in either subject or object position (3). To further bias participants into conceiving of the long-lived states as long-lived states and short-lived states as short-lived states, the inanimate arguments for the former were themselves long-lived (e.g., *Harvard's basketball team; his company's*

[1] See also discussion of Pesetsky (1995) below.

policy) and the inanimate arguments for the latter were short-lived (*the unexpected rainbow; seeing the gorgeous necklace*). Four additional filler sentence pairs describing non-psychological events (*The ocean wave tsunamis the village* vs. *The village tsunamis the ocean wave*) were also constructed. Experiments 1a and 1b differed only in the tense of the verb: simple present in 1a and simple past in 1b.

Procedure

Participants were told that they would try to correctly use new Japanese loanwords. For each verb, they were given a definition and the two possible sentences. An example trial is shown below:

(3) Tekitaishin
The feeling of rivalry
a. Richard tekitaishins Harvard's basketball team.
b. Harvard's basketball team tekitaishins Richard.

They were asked to choose the sentence they thought most likely to be correct. Four test forms were constructed as follows: the order of verbs was pseudorandomized such that the same condition (emotion/disposition) did not occur more than twice in a row. We counter-balanced whether the ES sentence or EO sentence was displayed first within each condition. The second form was made by switching the order of the sentences for each verb. Forms 3 and 4 were made by reversing the order of the verbs in Forms 1 and 2.

Results and Discussion

As predicted, participants were more likely to choose the ES frame for long-lived verbs than for short-lived verbs, in both Experiment 1a (M=62.5%, SE=4.3% vs. M=32.5%, SE=4.2%) and 1b (M=58.7%, SE=5.5% vs. M=33.1%, SE=5.0%).[2] The main effect of short-lived/long-lived was significant ($F1(1,38)=60.8$, $p<.001$; $F2(1,14)=6.2$, $p=.03$),[3] and this effect did not interact with tense (Fs<1). Thus, semantics guides the preferences of native English-speakers for certain verbal syntactic forms. Interestingly, although the past tense is more amenable to the description of events, participants were not more likely to choose the object-experiencer frame when the verb was presented in the past tense (Fs<1), perhaps because the inanimate arguments used for the short-lived verbs strongly implied events (e.g., *Seeing the gorgeous necklace tokimekis Mary*).

Thus, the underlying semantics of the sentence (the verb and/or inanimate argument) biased participants to choose a particular syntactic frame: ES for short-lived states and EO for long-lived states. In Experiment 2, we test whether this distinction is cross-linguistically relevant by turning to Japanese, a language historically unrelated to English.

[2] Means and standard errors here and elsewhere calculated by subject.

[3] Items analyses consider a given verb in present or past tense to be the same verb. Treating them as separate items does not affect the pattern of results.

Experiment 2: Japanese

Japanese is widely considered to be a language isolate, and its grammar is distinguished from that of English in a number of important ways (Tsujimura, 2007). First, Japanese is a scrambling language, allowing considerable word-order variation, with the basic order being Subject-Object-Verb, while in English the word order is rigidly Subject-Verb-Object. Second, unlike in English, the grammatical roles of noun phrases are overtly marked by particle suffixes: the subject is generally marked by *-wa*, and the direct object is marked by *-o*. Third, in the verbal domain, Japanese is a highly agglutinative language in which a verbal stem must at least bear a tense suffix and also may appear with a number of other suffixes expressing various grammatical functions. One such verbal suffix that is relevant for our purposes is the causative suffix *(s)ase-*. For example, *aruk-ase-* is the causative form of the verbal stem *aruk-* 'walk', meaning 'to make somebody walk'. This suffix is productive and can combine with almost all verbal stems.

Interestingly, while English contains more morphologically simple EO verbs (220) than ES verbs (44; Levin, 1993), our survey of Japanese found only 5 morphologically simple EO verbs, with the vast majority (74) ES.[4] Additional, morphologically complex, EO verbs can be formed in Japanese by adding the causative *–(s)ase-* affix to a ES verb:

(4) a. Taro-wa koomori-o kowagat-ta.
Taro-TOP bat-ACC fear-PAST
Taro feared bats
b. Koomori-wa Taro-o kowagar-ase-ta.
bat-TOP Taro-ACC fear-CAUS-PAST
Bats frightened Taro.

As in Experiment 1, we tested verbs in both the present and past tense. However, since in Japanese ES verbs are unnatural in simple tenses (*John-wa Mary-o nikum-u; John-TOP Mary-ACC hate-PRES), we used the more natural progressive form (John-wa Mary-o nikun-dei-ru; John-TOP Mary-ACC hate-PROG-PRES; "John hates Mary") for both verb classes. Note that with certain stative verbs the progressive morphology does not force a progressive meaning (e.g. the previous example does not mean "John is hating Mary").

Method

Participants

Forty native Japanese-speakers participated in Experiment 2: twenty in 2a (20-35yo, M=22.3, SE=2.8) and twenty in 2b (19-65yo, M=31, SE=3.3). Participants, who were recruited in public spaces around Tokyo, gave informed consent and were compensated with a souvenir pencil.

[4] Throughout this paper we consider only transitive verbs that take direct objects (*John fears/frightens Sally*). Future research will investigate intranstive verbs that take oblique objects (*John cares about/matters to Sally*).

Materials and Procedure

Materials and procedure were modeled closely on Experiment 1. Participants were introduced to novel English-derived loanwords in Japanese (long-lived: reverence, greed, phobia, envy, credence, affection, loathing, pride; short-lived: déjà vu, anguish, grief, jolt, nostalgia, trepidation, glee, chagrin). Loan words in Japanese can be made using the semi-productive verbalizer *-r-* (e.g., *gugu-r-u*: 'to google') or the light verb *suru* (e.g., *enzyoi-suru*: 'to enjoy'). While the latter is more productive, it often carries an explicitly causative meaning, particularly when applied to states. Since our goal was to avoid explicit morphosyntactic markers of meaning (with any concomitant argument selection biases), we used the more neutral *–r-*.

Again, care was taken to ensure that the loanwords did not approximate any extant Japanese monomorphemic words (e.g. *hatred* was avoided, since Japanese already contains *nikum-u*, which means *to hate*). As in Experiment 1, long-lived psychological states were paired with long-lived inanimate arguments (e.g., *the mountain*; *the theory of evolution*) and short-lived psychological states with short-lived inanimate arguments (e.g., *news of her brother's accident; seeing the foreign town*). The four filler verbs were existing English-derived psych verbs.

Experiments 2a and 2b differed only in that the verbs were in the present-progressive in 2a and in the past-progressive in 2b. Two of the filler verbs in 2a were problematic and were replaced in 2b. An example trial for a short-lived verb from Experiment 2b are shown below:

guriifu (grief): deep sorrow (especially that caused by someone's death)
a. Tooru-wa aiken-no shi-o guriifu-t-tei-ru
Toru-TOP pet.dog-GEN death-ACC grief-V-PROG-PAST
Toru grieves the pet dog's death.
b. Aiken-no shi-wa Tooru-o guriifu-t-tei-ru
pet.dog-GEN death-TOP Toru-ACC grief-V-PROG-PAST
The pet dog's death grieves Toru.

Results and Discussion

Like English speakers, Japanese participants were more likely to select the ES interpretation for the long-lived verbs than for the short-lived verbs in both Experiments 2a (M=90.6%, SE=1.0% vs. 73.7%, M=0.9%) and 2b (M=73.1%, SE=0.9% vs. M=55.6%, SE=0.9%). The overall main effect of short-lived/long-lived was significant ($F1(1,38)=28.6$, $p<.001$; $F2(1,14)=16.8$, $p=.002$) and did not interact with tense ($Fs<1$). Unlike in English, there was a significant main effect of tense, with ES interpretations more likely in present tense than past ($F1(1,38)=6.3$, $p=.02$; $F2(1,14)=21.5$, $p<.001$).

These results suggest that linking rules in Japanese, as in English, are sensitive to the duration of the psychological state. Interestingly, however, Japanese participants were overall more likely than English speakers to choose the ES frame (M=72.5%, SE=3.5% vs. M=46.7%, SE=2.8%; $t1(78)=5.8$, $p<.001$; $t2(30)=3.5$, $p=.001$). This could show a broad preference for the EXPERIENCER→SUBJECT linking rule in Japanese. Alternatively or in addition, Japanese participants may have been sensitive to the fact that the novel verbs were all morphologically simple, and nearly all morphologically simple psych verbs in Japanese are ES (see above). EO verbs are typically formed with the addition of the causative affix *-(s)ase-*. We tested whether participants would be more likely to choose the EO form for *–(s)ase-* affixed verbs in Experiment 3.

Experiment 3: Causative Psych Verbs in Japanese

In Experiment 3, we tested whether Japanese participants would choose EO frames for *–(s)ase* affixed psych verbs.

Method

Participants

Twenty participants (19-34yo, M=22.5, SE=1.3), recruited in public spaces around Tokyo, gave informed consent and were compensated with a souvenir pencil.

Materials and Procedure

Materials and procedure were identical to Experiment 2b, except all verbs were causativized by the addition of the *-(s)ase-* affix and presented in the present progressive (guriifu-r-*ase*-tei-ru).

Results and Discussion

As in Experiment 2, Japanese participants were more likely to choose the ES interpretation for the long-lived verbs than the short-lived verbs (M=33.1%, SE=5.2% vs. M=21.2%, SE=3.5%; $t1(19)=2.41$, $p=.03$; $t2(14)=2.83$, $p=.01$). As predicted, participants were overall much less likely to choose the ES interpretation relative to Experiment 2a (M=27.2%, SE=3.7% vs. M=80.6%, SE=3.6%; $t1(38)=10.3$, $p<.001$; $t2(15)=20.5$, $p<.001$). Thus, the preference for the ES interpretation in Experiment 2 was not due to a global preference for EXPERIENCER→SUBJECT linking, but rather was specific to the verb form used (monomorphemic).

General Discussion

In order to discuss events and states, speakers must map the participants in the event or state onto grammatical roles. There remains considerable disagreement about the nature of these mappings or linking rules (Levin & Rappaport Hovav, 2005). Linking rules are typically defined in terms of features of the arguments such as agentivity or causativity (Dowty, 1991; Pesetsky, 1995; Pinker, 1984; 1989) or aspects of the predicate such as stativity and telicity (Hooper & Thompson, 1980). In this paper, we present evidence that in the case of psych verbs, linking rules are sensitive to duration of the psychological state: if the state is short-lived, the EXPERIENCER→OBJECT rule is more likely to apply; if the state is long-lived, the EXPERIENCER→SUBJECT rule

applies. This distinction appears in both English and Japanese, historically unrelated and grammatically distinct languages. Coupled with the fact that the this distinction may also characterize existing verbs in Finnish (Pylkannen, 1999) and Mandarin (Hartshorne, 2009), which are unrelated to each other or to English or Japanese, these results suggest this distinction *could* be universal across languages.

Causes, Stages and Emotions

The data in this paper demonstrate that the mapping from semantics to syntax for psych verbs is governed at least in part by the meaning of the verb. Although we discussed our manipulation in terms of the expected duration of the psychological state, that may not be the correct distinction.

Our experiments above were partly motivated by the distinction in the psychological literature on emotion between emotions and dispositions. Since one of the defining distinctions between emotions and dispositions is their duration, this distinction is fully confounded with our short-lived/long-lived distinction.

Similarly, we noted that Pylkkanen (1999) argues that Finnish ES verbs are individual-level predicates and Finnish EO verbs are stage-level predicates. Stage-level and individual-level predicates are usually defined in terms of the genericity of predicates—typically formalized as whether the predicate refers to a single event or quantifies over many events (Carlson, 1988). Genericity can be diagnosed by linguistic tests such as the permissibility of the progressive (see Pylkkanen, 1999). As noted above, at least one of the linguistic tests has apparent semantic consequences. One distinguishing factor of EO predicates is that they can be bounded by brief temporal durations, making the notion of *stage-level* similar to our notion of *short-duration*. Whether the two can be de-confounded is a question for future research.

Note that while it may be that *stage-level*, *short-lived* and *emotion* may simply be three ways of capturing the fundamental distinction that influences the semantic-syntactic mapping, the same may not be true for the other semantic distinction that has been suggested in the literature: Pesetsky (1995) presents linguistic analyses suggesting that EO verbs encode caused events, while ES verbs do not. Intuitively, brief states like emotions seem related to changes of state, which is a necessary component of *cause*, perhaps suggesting a way of integrating the notions.[5] Relatedly, Pylkkanen (1999) argues causally-affixed Finnish psych verbs either describe events or stage-level (rather than individual-level) states, providing another potential association. Nonetheless, the associations here are tenuous. Whether *cause* is a factor in the semantics-syntax linking rules for psych verbs – and, if so, whether it is a factor independent of the one(s) described above – remains a question for future research.

Universals

There have been several proposals suggesting that linking rules are universal, innate and exceptionless. Baker proposes his Uniformity of Theta Assignment Hypotheses (Baker, 1988), which posits a simple, exceptionless, many-to-one rules linking semantics roles (AGENT, EXPERIENCER) to syntactic position (SUBJECT, DIRECT OBJECT), at least at the level of deep structure. Pinker (1984) argues that linking rules may be innate. Such claims not only greatly simplify linguistic theory, but they also simplify the job of the language learner.

However, such theories have been challenged by apparent variation in the application of linking rules in some domains, such as psych verbs. The data presented here suggest a solution to this problem compatible with exceptionless linking rules: a rigid, innate linking rule that maps EXPERIENCER→SUBJECT for long-lived psychological states and EXPERIENCER→OBJECT for short-lived psychological states. Whether such rules apply beyond English and Japanese (and perhaps Finnish and Mandarin) remains an empirical question. This may suggest that other such cases of variation may similarly be resolved by closer inspection of the semantics (see also Pesetsky, 1995, for discussion). While this is an intriguing possibility, it is not the only possible conclusion (see below).

Variation

Despite the potentially universal sensitivity of linking rules to psychological state duration described above, Japanese and English speaking participants showed a striking difference in their baseline preference for the two argument mappings: Japanese participants were over 50% more likely than English-speakers to chose the ES form. At least three explanations for this cross-linguistic variation are possible.

First, although stimuli for the English and Japanese studies were constructed in an identical manner, the stimuli were not identical (the *different-stimuli hypothesis*). It may be that the semantics of the Japanese stimuli were biased in favor of the ES mapping; perhaps the short-lived verbs were less short-lived than those in the English study. Although such a possibility is difficult to rule out with certainty, the relative size of the effect limits the likelihood that poor stimulus selection explains the effect. Moreover, the discrepancy was highly consistent across stimuli: all but one of the short-lived English verbs in Experiment 1a had more EO attributions than *any* of the short-lived Japanese verbs in Experiment 2a. Similarly, all but one of the long-lived English verbs in 1a had more EO attributions than *any* of the long-lived Japanese verbs in 2a (the comparison for 1b and 2b is similar).

A second possibility is that linguistic differences between Japanese and English led the participants to construe the

[5] Consistent with this possibility, an additional experiment using novel Japanese psych verbs created with *–suru*, which typically gives rise to a causative interpretation, found that Japanese participants overwhelmingly chose the EO reading.

meanings of the novel verbs differently (the *different-construal hypothesis*). There are a number of reasons this might happen. For example, Pesetsky (1995) has argued that only EO verbs describe caused events. Japanese can mark verbs overtly as causal with the *–(s)ase* affix, and in fact there are only a handful of EO verbs lacking the causal affix. In Experiment 2, the verbs presented to the Japanese participants lacked the causal affix. These participants may then have made the inference that the verbs do not describe caused events, leading them to choose the ES reading. Since English does not explicitly mark verbs as causal or not, the English-speaking participants faced a more ambiguous inference problem.

Note that the *different-stimuli hypothesis* and the *different-construal hypothesis* are both consistent with rigid, exceptionless linking rules. The English and Japanese participants apply the linking rules in the same way; they simply disagree as to the meanings of the verbs. Another possible conclusion is that linking rules are constrained by universal biases but allow some cross-linguistic variation in their exact formulation (the *soft-universals hypothesis*). Imagine that based on the available cues Japanese and English speakers arrive at the same guess about the underlying semantics. They may still show different baseline preferences if argument mappings are probabilistic.

Our data provide evidence for a universal bias in argument mappings, however, they do not show that such mappings have to be either exceptionless or deterministic. Instead, semantics-to-syntax mappings for arguments could themselves be probabilistic and influenced by both soft universals and language-specific factors.

For example, as discussed above, unmarked psych verbs in Japanese are overwhelmingly ES while the opposite is true (to a lesser degree) in English (see above). Suppose that in addition to universal (and presumably innate) biases, mappings are also influenced by similarity to other verbs. In such a scenario, the baseline statistics of psych verbs in the two languages would predict the baseline difference in performance.

Models that allow for within-language, across-item generalizations of this form have a long history in both generative linguistics (where they often take the form of *parameter-setting* models) and non-generative approaches such as construction grammar. Recent work in computational modeling has shown how such systems can be expressed by hierarchical Bayesian models. These models encode the across-item generalizations as *overhypotheses*—hypotheses about hypotheses (see e.g. Perfors, et al., *in press*).

It remains for future work to determine whether cross-linguistic differences are better attributed to variation in how speakers of various languages construe situations, to probabilistic linking rules, or to some combination of both.

Acknowledgments

The authors wish to thank the members of SnedLab for discussion and suggestions. This material is based on work supported by a National Defense Science and Engineering Graduate Fellowship to JH and a grant from the National Science Foundation to JS (0623845).

References

Baker, M.C. (1988). *Incorporation: A Theory of Grammatical Function Changing*. Chicago, IL: University of Chicago Press.

Carlson, G. (1988). The semantic composition of English generic sentences. In G. Chierchia, B. Partee, & R. Turner (Eds.), *Property Theory, Type Theory, and Semantics*. Boston, MA: D. Reidel Publishing.

Croft, W. (1993). Case marking and the semantics of mental verbs, in J. Pustejovsky (Ed.), *Semantics and the Lexicon*. Dordrecht: Kluwer Academic.

Dixon, R. M. W. (1994). *Ergativity*. Cambridge, UK: Cambridge University Press.

Dowty, D. R. (1991). Thematic proto-roles and argument selection, *Language*, 67, 547-619.

Ekman, P. (1999). Basic emotions. In T. Dalgleish & M. Power (Eds.), *Handbook of Cognition and Emotion*. Sussex, U.K.: John Wiley & Sons, Ltd.

Hartshorne, J. K. (2009). The duration of psychological states. Unpublished manuscript.

Hooper, P. J. & Thompson, S. A. (1980). Transitivity in grammar and discourse, *Language*, 56, 251-95.

Levin, B. (1993). English Verb Classes and Alternations: A Preliminary Investigation. Chicago, IL: University of Chicago Press.

Levin, B., & Rappaport Hovav, M. (2005). Argument Realization. Research Surveys in Linguistics Series. Cambridge, UK: Cambridge University Press.

Marantz, A.P. (1982). On the acquisition of grammatical relations. *Linguistische Berichte: Linguistik als Kognitive Wissenschaft*, 80/82, 32-69.

Perfors, A., Tenenbaum, J.B. & Wonnacot, E. (in press) Variability, negative evidence, and the acquisition of argument constructions, *Journal of Child Language*.

Pesetsky, D. (1995). Zero Syntax: Experiencers and Cascades. Cambridge, MA: The MIT Press.

Pinker, S. (1984). Language Learnability and Language Development. Cambridge, MA: Harvard University Press.

Pinker, S. (1989). Learnability and Cognition. Cambridge, MA: The MIT Press.

Pylkkanen, L. (1999). On stativity and causation. In C. Tenny & J. Pustejovsky (Eds.), *Events and Grammatical Objects*. Stanford, CA: CSLI Publications.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Cambridge, MA: Harvard University Press.

Tsujimura, N. (1996). *Introduction to Japanese Linguistics*. Malden, MA:Wiley-Blackwell.

Comprehending Negated Sentences With Binary States and Locations

Sarah E. Anderson (sec57@cornell.edu)[1],
Stephanie Huette (shuette@ucmerced.edu)[2],
Teenie Matlock (tmatlock@ucmerced.edu)[2],
and Michael J. Spivey (spivey@ucmerced.edu)[2]
[1]Department of Psychology, Cornell University, Uris Hall, Ithaca, NY 14853 USA
[2]Department of Cognitive Science, University of California, Merced, Merced, CA 95344 USA

Abstract

Language theorists have argued that processing negated statements ("The eagle is not in the sky,") differs from affirmative propositions. However, evidence for these claims comes from studies that did not control for the possibility of numerous states (e.g., the eagle is perched on a branch or on the ground). Here, we explore whether constraining this number of possibilities provides more information about processing negation. In Experiment 1, the stimuli described binary states. For example, a coin can be either heads up or tails up; if it is not heads up it is necessarily tails up. In Experiment 2, preceding contexts constrained the number of possible locations of a negated proposition. The results, consistent with earlier evidence for negation's increased complexity, offer new data suggesting that perceptual simulation of negated proposition may be experimentally detected when the states or locations are sufficiently constrained, using binary states or contextual descriptions.

Keywords: Language Processing; Negation; Embodied Cognition; Perceptual Simulation

Negation is a fundamental part of everyday communication. Throughout the course of a typical day, people frequently have to report where things are *not* located, for instance, "Your keys are not on the table" or "My car is not in the garage." They must also describe events that are *not* happening, such as, "The Patriots are not playing tonight" or "Your car will not start." Despite the ubiquity of negated statements, surprisingly little is known about how they are processed and what their conceptual structure is.

One of the earliest and most reliable findings about negation is that people are slower to read negated sentences than they are to read affirmative sentences, due to their increased complexity (see Barres & Johnson, 2003; Carpenter & Just, 1975; Chase & Clark, 1972; Mayo, Schul, & Burnstein, 2004). Such findings have provided invaluable insights into sentence processing, but many important questions about processing negation remain. In particular, how do negated statements influence everyday cognition? Are negated sentences comprehended differently than affirmative sentences? The goal here is to further consider negation's influence on sentence comprehension.

Negation has been of interest to philosophers and language theorists for centuries (for review, see Gilbert, 1991), but only recently has its processing received close attention. Early cognitive work on processing suggested that negated statements about spatial relations are processed differently from similar affirmative statements. For instance, participants in Clark and Chase (1972) were presented with sentences followed by pictures and then asked whether the sentence was true or false of a corresponding picture. For example, a "true" trial might consist of the sentence, "The star is above the cross," followed by a picture with a star above a cross, accurately depicting the relationship expressed by the sentence. Participants responded true or false more quickly to a picture following the sentence, "The star is above the cross," than to the sentence, "The star is not above the cross." These differences suggested that affirmative and negated statements are processed differently, but the nature of this difference was unclear. One possibility is that the increased processing time associated with negation is the result of evaluating the core proposition and then applying a negation marker to this proposition.

More recent evidence also supports the hypothesis that negating or affirming a statement involves subtly different processes. McKinstry, Dale, and Spivey (2008) presented participants with questions of varying truth-values and asked them to answer "yes" or "no" using a mouse to click on a corresponding, visually presented box on the computer screen. A sentence's truth-value was defined as the proportion of participants who agreed the statement was true in an on-line survey. Therefore, the question, "Should you brush your teeth every day?" had a truth-value of 1.0, and the question, "Is murder sometimes justifiable?" had a truth value of .6. In addition to recording the end response and reaction time, the trajectory of the mouse as it moved across the computer screen to click on the appropriate answer was also recorded. These mouse-movement trajectories provide a continuous motor response that has been used to illustrate competition between alternatives in a number of cognitive tasks (Dale, Kehoe, & Spivey, 2007; Farmer, Anderson, & Spivey, 2007; Spivey, Grosjean, & Knoblich, 2005). McKinstry and colleagues' findings were consistent with those of Clark & Chase (1972) in that participants had more difficulty in evaluating a false statement than a corresponding true statement. Participants were also slower to respond negatively to a statement, and the "no" response trajectories showed more competition than the "yes" response trajectories. Similar effects also arise in research exploring the influence of negation on memory (Fiedler, Walther, Armbruster, Fay, & Naumann, 1996), supporting a general cognitive bias towards affirmative propositions.

Similarly, negated and affirmative sentences seem to be handled differently in language comprehension (e.g., Hasson & Glucksberg, 2006; Kaup, 2001; MacDonald & Just, 1989), and this may be due to differences in their corresponding

perceptual simulations. Recent experimental evidence suggests that understanding a single word embedded in a sentence is associated with the way people would actually perceive the noun they are asked to identify. Zwaan, Stanfield, and Yaxley (2002) asked participants to read sentences and to decide whether or not a subsequent picture was mentioned in the sentence they had just read. When a sentence such as, "The eagle was in the sky," was presented, participants were quicker to respond that a picture of an eagle with outstretched wings had been mentioned in the preceding sentence than when they saw a picture of an eagle with folded wings. These results support claims that participants construct an image to represent the sentences they read; this in turn makes that image more accessible, leading to faster subsequent responses to that image. Such images, constructed through the partial activation of the neurons used to actually perceive or interact with the objects, are called *perceptual simulations* (see Barsalou, 1999; Barsalou, Simmons, Barbey, & Wilson, 2003, for a more complete overview). When the test picture matched the image that had been mentally created, reaction times were faster than if the test picture did not match the state of the object described in the text. These data provide evidence that comprehending language may be grounded in perceptual representations.

Perceptual simulations seem able to explain the comprehension of affirmative sentences (Zwaan, Stanfield, and Yaxley, 2002), yet recent experimental results suggest that negated sentences are processed differently. Kaup, Yaxley, Madden, Zwaan, and Lüdtke, (2006b) presented participants with sentences like, "The eagle was not in the sky," and asked participants to judge whether a subsequently presented picture was mentioned in the sentence they had read. If participants perceptually simulate negation in the same way they simulate affirmative sentences, they should be quicker to respond "yes" to pictures that match the sentence (e.g., an eagle with its wings folded). The experimenters found that when participants read negated sentences, response times to a picture that matched the *affirmative* version of the proposition (eagle with wings spread) were faster than pictures that actually matched the negated sentence, suggesting that a perceptual simulation of the affirmative proposition was created in response to negated sentences. This suggested that negation is handled differently from other aspects of sentence processing, and specifically not through perceptual simulation.

Another possibility is that the experimental materials were not able to capture the perceptual simulation of the negation, similar to on-line research in verbal aspect. Madden and Zwaan (2003) examined potential differences produced by processing different verbal aspectual forms in narrative reading. In these experiments, participants were quicker to respond to pictures showing completed action after they had read a sentence containing a simple past verb than when they had read a sentence containing a past progressive verb, because simple past verbs emphasize the completion of a verb's action. Conversely, no such latency differences were found when participants read sentences containing past progressive verbs and then saw pictures of intermediate action. Like affirmative and negated sentences, these results suggest that perhaps one form of aspect is comprehended through perceptual simulations and that the other is comprehended via another mechanism.

However, the authors suggest that the past progressive's lack of effect was due to readers representing the ongoing action at different stages of completion. Simple past verbs place emphasis on the end state of the action, which typically corresponds to only a single state, while past progressive verbs place emphasis on the ongoing nature of the verb. After reading narratives containing past progressive verbs, participants may simulate a number of locations. In other words, past progressive aspect produces a diffuse number of possible stages of intermediate action that are un-captured by the task. Therefore, even though past progressive verbs, like simple past verbs, may be comprehended via perceptual simulations, the static images used in the task simply do not match the particular point in the action they are simulating.

Similarly, it may be that when participants read a negated sentence, they do create a perceptual simulation of the negation, but the pictures they are asked to respond to do not capture the simulation. Whereas an affirmative sentence identifies a particular state or location for the noun that is responded to more quickly when presented visually, the negated sentences do not. For example, when "the eagle is in the air," its wings are necessarily open to accommodate flying. However, when "the eagle is not in the air," there are many states it could be in other than sitting with it wings folded. Thus, when hearing a negated sentence of this sort, participants may appear to simulate the eagle in the air not because perceptual simulations are incapable of negation but instead because the alternative simulations are too numerous and too diffuse. If this is the case, then constraining the possible simulations to only two for a given object, one corresponding to the affirmative proposition and one to the negated proposition, should further inform this research.

Here, we wanted to explore this in sentence comprehension by constraining the possible states and locations of the event. In general, we hypothesized that binary states would allow us to capture the simulation of negation in sentences. Similarly, we hypothesized that creating contexts to limit the possible locations to only two options would allow us to observe and further extend results on processing negated sentences.

Experiment 1: Binary States

Evidence from recent research in sentence comprehension suggests that creating targets that themselves refer to binary states is promising for investigating the role of perceptual simulation in negation processing. Kaup, Ludtke, & Zwaan (2006a) investigated the way participants perceptually simulate sentences like those used in the earlier negation research (Kaup, et al., 2006b), but created materials that were binary, or had contradictory predicates. In Kaup, et al., (2006a), participants read binary sentences in the self-paced reading paradigm, and then after an SOA of either 750ms or

1500ms, they saw a picture of an object that they had to name as quickly as possible. At the 750ms SOA, responses to pictures that matched the non-negated state of affairs, even when the target sentence was negated, were significantly faster than pictures depicting the negated state of affairs. In other words, for the sentence "The door was not closed," reaction times to a picture of a closed door (matching the proposition of the sentence) were significantly faster than to pictures of an open door. At the 1500ms SOA, when the target sentence was negated, responses to pictures matching the negated state of affairs were significantly faster than pictures depicting the proposition.

Here, we wanted to further expand these findings. In Experiment 1, we used the picture verification task used in the earlier negation (Kaup, et al., 2006b) and perceptual simulation (Zwaan, et al., 2002) research. We anticipated that this method might be more robust and allow for a full statistical interaction to emerge from the data. Additionally, we used an intermediate SOA of 1000ms, in order to provide data on processing at this intermediate point.

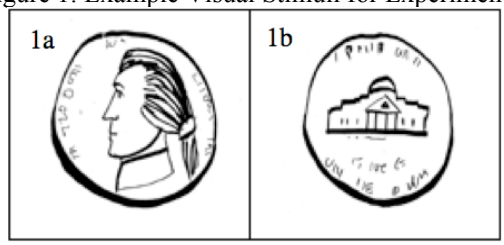
Method

Participants. A total of 32 Cornell University undergraduates participated in the experiment for extra course credit. All participants were right-handed, born in the United States, and native English speakers.

Materials. Sixteen target sentence frames were constructed. These frames were manipulated in order to produce two versions of each: a negated (The coin is not heads up) and an affirmative (The coin is heads up) version. The sentences were created such that they described a proposition that was true only in one way and untrue only in one way, thereby making the materials binary. Sixteen filler sentences, half of which were affirmative and half of which were negated, were also constructed. The target sentences contained binary state objects similar to the examples provided above, while the filler sentences did not contain binary objects.

Two pictures were created for each target frame: a picture matching the proposition of the sentence, and a picture matching the negation of that proposition. For example, for the sentence frame, "The coin is (not) heads up," the picture corresponding to the proposition would be a heads up coin (see Figure 1a) and the picture corresponding to the negation of the proposition would be a tails up coin (see Figure 1b). The correct response for either the negated or affirmative sentence and either picture is "yes," because the sentence is about a coin and the picture depicts a coin. Filler sentences also had corresponding pictures, although these pictures did not match anything in the sentence. All of the images were black and white ink drawings, created by a senior art major, with as much simplicity and as few lines as possible; this was done in order to make sure all pictures were as similar as possible. All the pictures were scanned into the computer in the same size to control for discrepancies between the objects.

Figure 1: Example Visual Stimuli for Experiment 1



1a) Picture Matches Proposition
1b) Picture Matches Negated Proposition

Procedure. Participants were seated at the computer and asked to make themselves comfortable. They read a page of instructions where they were informed that it was important for them to make decisions about the picture as quickly and accurately as possible. During the task, participants first read a sentence located in the center of the screen, pressing the spacebar when they had understood the sentence. A fixation cross appeared in the center of the screen for 1000ms and then a picture appeared. Participants indicated whether the pictured object had been mentioned in the previous sentence by pressing the f-key, covered with a "yes" sticker, or the j-key, covered with a "no" sticker. The correct response to all target trials was yes, and the correct response to all the filler trials was no. Half of the trials were followed by comprehension questions, in order to insure participants were paying attention. On these trials (half of the filler sentences, and half of the target sentences), a question regarding the sentence was presented next. Participants were asked to respond to the question as accurately as possible by clicking on the appropriate "yes" or "no" key. Participants were not given feedback on any of their responses. They were first given two practice trials before beginning the task (similar in construction to the filler items), and the task lasted approximately 10 minutes.

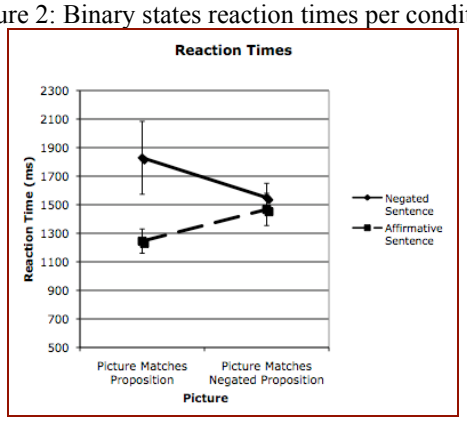
Results and Discussion

Incorrect trials (trials on which a participant responded "no," and the pictured object was not in the sentence they had just read) were removed before analyzing reaction time. Data from two participants were excluded from the analysis because they incorrectly answered one block of trials. In addition, the incorrect responses to 18 items from 9 different participants were discarded prior to the analysis.

There was a significant interaction of Sentence and Picture, $F(1,29)= 4.308$, $p<.047$. See Figure 2 below**.** Affirmative sentences that were followed by propositional pictures ("The coin was heads up" before a picture of a heads up coin) were responded to more quickly, $M = 1245.12$, $SD = 84.98$, than affirmative sentences that were followed by negated-propositional pictures ("The coin was heads up" before a picture of a tails up coin), $M = 1467.02$, $SD = 113.05$. These results are consistent with earlier perceptual simulation research (Zwaan, et al. 2002). Additionally, negated sentences followed by negated propositional pictures ("The coin is not heads up" followed by a picture of a tails up coin) were responded to more quickly, $M = 1550.17$, $SD =$

99.81, than were negated sentences followed by propositional pictures ("The coin is not heads up" followed by a picture of a heads up coin), M =1828.79, SD = 255.49. These results moderately extend the results obtained in Kaup et al. (2006a) by providing support that perceptual simulation does seem to operate for comprehending negated sentences whenever the experimental conditions are sufficiently constrained to capture it at an intermediate SOA.

Figure 2: Binary states reaction times per condition



In examining the main effects, the affirmative sentences, M = 1467.02, SD = 619.18, were responded to significantly faster, $t(29)$ = 2.08, $p<.05$, than negated sentences, M = 1245.12, SD = 465.47. These results are consistent with earlier research, like that of Clark and Chase (1972) and McKinstry et al. (2008), suggesting that there is a bias in favor of affirmative sentences over negated sentences. Also, there was no main effect of picture type, implying that the type of picture did not impact comprehension.

Finally, in examining accuracy, there was a main effect of picture type within the affirmative sentences, $F(1,29)$ = 6.916, $p<.014$. When the sentence was affirmative, participants were less accurate when responding to pictures that did not match. Accuracy did not differ for the two picture conditions in the negative sentence condition. This implies that the effects that have been described so far are not to the result of a speed accuracy trade off in the negated sentence condition.

Experiment 2: Binary Locations

Using the picture verification task of other perceptual simulation research (Kaup, et al., 2006b; Zwaan, et al., 2002), Experiment 1 converges with other negation research (Kaup, et al., 2006a): When the possible states of the items themselves are constrained to only two possibilities, perceptual simulations underlie the processing of these sentences at the intermediate 1000ms SOA. However, the materials that were used in Experiment 1 relied on binary *state* objects, leaving many questions regarding the processing of negation unanswered. Here, we sought to further extend these findings to investigate the role of context in the creation of binary *locations*. In Experiment 2, we again used the picture verification task was employed in the earlier negation (Kaup, et al., 2006b) and perceptual simulation (Zwaan, et al., 2002) research as well as the intermediate 1000ms SOA. Rather than relying on binary states, context sentences describing two possible locations for an item were created.

Method

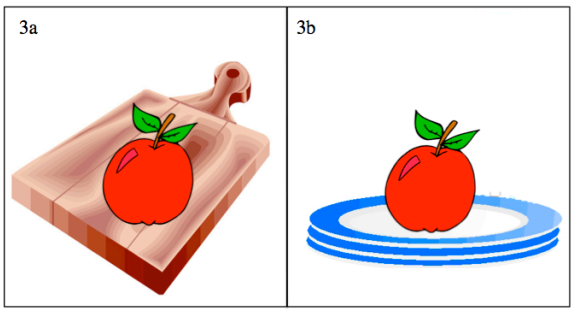
Participants. A total of 62 Cornell University undergraduates participated in the experiment for extra course credit. All participants were right-handed, born in the United States, and native English speakers.

Materials. Thirty-two target sentence frames were constructed. Each frame consisted of a context sentence, describing two possible locations for an item (i.e., The apple is either on the plate or on the cutting board), and a target sentence, identifying the location of the item. The context sentences were designed such that they limited the possible locations to only two; therefore, they could be true only in one way and untrue only in one way. The target sentences were manipulated in order to produce four versions of each: an affirmative version identifying the first location mentioned in the context sentence (The apple is on the plate); an affirmative version identifying the second location mentioned in the context sentence (The apple is on the cutting board); a negated version identifying the first location mentioned in the context sentence (The apple is not on the plate); and a negated version identifying the second location mentioned in the context sentence (The apple is not on the cutting board). Thirty-two filler sentences, one quarter of which corresponded to each of the four conditions described above, were also created.

Two pictures were created for each target frame: a picture matching the proposition of the sentence, and a picture matching the negation of that proposition. For example, for the target sentence frame, "The apple is (not) on the cutting board," the picture corresponding to the proposition would show an apple on a cutting board (see Figure 3a) and the picture corresponding to the negation of the proposition would show an apple on a plate (see Figure 3b). Pictures were also constructed for the filler items, although these pictures did not match anything in the sentence. All of the images were taken from clip art. The target item (i.e., the apple) was spliced into the different locations to maintain similarity in the pictures.

Eight lists were constructed such that each participant would respond to all of the conditions but only one version of each sentence frame and picture. Conditions were created as follows: 1) affirmative target, identifying first location of context, picture matching proposition; 2) affirmative target, identifying first location, picture matching negated

Figure 3: Example Visual Stimuli for Experiment 2



3a) Picture matches proposition for "The apple is (not) on the cutting board."
3b) Picture matches negated proposition for "The apple is (not) on the cutting board."
Pictures match final prepositional phrase of target regardless of the target's negation status.

proposition; 3) affirmative target, identifying the second location, picture matching proposition; 4) affirmative target, identifying the second location, picture matching negated proposition; 5) negated target, identifying the first location, picture matching proposition; 6) negated target, identifying first location, picture matching negated proposition; 7) negated target, identifying second location, picture matching proposition; and 8) negated target, identifying second location, picture matching negated proposition.

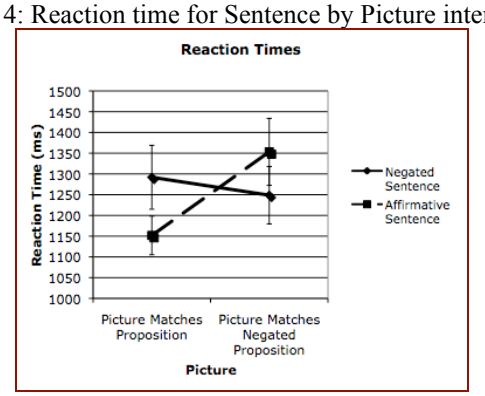
Procedure. The procedure of Experiment 2 was the same as that of Experiment 1, except it did not include comprehension questions and lasted 20 minutes.

Results and Discussion

Incorrect target trials (or trials on which a participant responded "no," the pictured object was not in the sentence they had just read) were removed before analyzing reaction time. Two hundred fifty-four trials, 12% of the data, were excluded from the reaction time analysis, resulting in the loss of fourteen participants. Each participant had at least one trial excluded from the reaction time analysis due to these criteria.

There was no significant three-way interaction of Sentence, Picture, and Context, $p > .28$. Also, the main effect of context was not significant, $p > .7$. Therefore, the order of the items in the context sentence did not significantly impact the results. However, there was a significant interaction of Sentence and Picture, $F(1,47) = 6.343$, $p < .015$. See Figure 4. Affirmative sentences ("The apple is either on the plate or on the cutting board. The apple is on the plate") followed by propositional pictures (a picture of the apple on the plate as in Figure 3b) were responded to faster, $M = 1155.497$, $SD = 413.419$, than affirmative sentences followed by negated propositional pictures (a picture of the apple on the cutting board as in Figure 3a), $M = 1365.495$, $SD = 701.461$. Similarly, negated sentences ("The apple is either on the plate or on the cutting board. The apple is not on the plate") followed by propositional pictures (a picture of the apple on the plate as in Figure 3b) were responded to slower, $M = 1335.8032$, $SD = 690.602$, than negated sentences followed by negated propositional pictures (a picture of the apple on the cutting board as in Figure 3a), $M = 1258.295$, $SD = 539.869$. These data further extend the results of Kaup et al. (2006a), providing support that perceptual simulations operate in comprehending negated sentences when contextual descriptions constrain possible locations to only two.

Figure 4: Reaction time for Sentence by Picture interaction



In examining the main effects, responses to negated sentences were not significantly slower than responses to affirmative sentences, $p > .8$. Also, there was a main effect of picture type, $F(1,61) = 7.67$, $p < .01$, such that pictures of the proposition were responded to faster, $M = 1221.829$, $SD = 563.79$, than pictures of the negated proposition, $M = 1300.929$, $SD = 622.402$. These results, combined with the percentage of incorrect responses, suggest that the complexity of the task may have caused problems or that some subjects employed strategies. For instance, it may have been possible to read the context sentence and respond to the pictures based on this information alone. Future work will further refine these preliminary data by using auditory files, rather than written text, in such a picture verification task. Such auditory presentation of the stimuli is less strategy prone than text presentation.

Finally, in examining accuracy, there was an interaction of negation and picture, $F(1,61) = 17.8$ $p < .001$. Participants responded more accurately when the target sentence matched the picture. Hence, participants responded more accurately to negated sentences when the picture matched the negated proposition than when the picture matched the proposition. Similarly, participants responded more accurately to affirmative sentences when the picture matched the proposition than when the picture matched the negated proposition. None of the other main effects were significant, p's $> .2$. This implies the effects described so far are not due to a speed accuracy trade off in the either sentence condition.

General Discussion

The materials of Experiment 1 were designed to reflect *binary state* propositions, such that affirmative and negated forms each referred to only one possibility. The interaction of picture and sentence types at the intermediate SOA of 1000 ms supports the hypothesis that appropriate perceptual simulations (of the negated proposition) may be used for

comprehending negated sentences. Further, Experiment 2 demonstrated that extrasentential context can constrain the possible perceptual simulations to reflect *binary locations*. Again, the interaction of picture and sentence types supports the hypothesis that perceptual simulations may be used in comprehending both affirmative and negated sentences. While the experiments here used the intermediate SOA of 1000 ms, future planned research, specifically in eye-tracking, will further investigate the time course of processing negated sentences.

The results reported here are promising, but future research is needed to further explore the mechanisms of negation. The exact mechanism underlying perceptual simulations in negation has not been thoroughly explored, and its explication is likely to require computational modeling. To this end, we have begun some preliminary explorations with a simple recurrent network (Elman, 1990). In addition to the word-prediction relation between 91 input word-nodes to 91 output word-nodes, this model includes 63 perceptual features on the output layer that are prominent properties of the relevant perceptual simulations (see also, Howell, Jankowicz, & Becker, 2005). Thus, combined with its learning to predict the next word in a sentence, this network also learns to activate the appropriate set of features for the perceptual simulation associated with the sentence (Anderson, Huette, Matlock, & Spivey, in press). In this network model, the only difference between an affirmative sentence and a negated sentence is that the input from the negated sentence has the word "not" immediately preceding its critical adjective (e.g., flying, not flying, heads-up, not heads-up). Thus, without a specialized logical operation of negation, this network can nonetheless reverse its perceptual simulation as a result of the presence or absence of the word "not" in the sentence. Current extensions of this model are exploring ways to include some temporal dynamics that may simulate the experimental results with different SOAs.

References

Anderson, S.E., Huette, S., Matlock, T., & Spivey, M.J.(in press). On the temporal dynamics of negated perceptual simulations. In F. Parrill, V. Tobin, & M. Turner (Eds.), *Meaning, form, & body*, (pp. 1-20). Stanford, CSLI.

Barres, P.E., & Johnson-Laird, P.N. (2003). On imagining what is true. *Thinking and Reasoning, 9*, 1-42.

Barsalou, L. (1999). Language comprehension: Archival memory or preparation for situated action? *Discourse Processes, 28*, 61-80.

Barsalou, L., Simmons, W., Barbey, A.,, & Wilson, C. (2003). Grounding conceptual knowledge in modality-specific systems. *Trends in Cognitive Sciences, 7*, 84-91.

Carpenter, P.A. & Just, M.A. (1975). Sentence comprehension: A psycholinguistic processing model of verification. *Psychological Review, 82*, 45-73.

Clark, H.H., & Chase, W.G. (1972). On the process of comparing sentences against pictures. *Cognitive Psychology, 3*, 472-517.

Dale, R., Kehoe, C., & Spivey, M. (2007). Graded motor responses in the time course of categorizing atypical exemplars. *Memory and Cognition, 35*, 15-28.

Elman, J. L. 1990. Finding structure in time. *Cognitive science*, *14*(2), 179-211.

Farmer, T.A., Anderson, S.E., & Spivey, M.J. (2007). Gradiency and visual context in syntactic garden-paths. *Journal of Memory & Language, 57*, 570-595.

Fiedler, K., Walther, E., Armbruster, T., Fay, D., & Naumann, U. (1996). Do you really know what you have seen? Intrusion errors and presuppositions effects on constructive memory. *Journal of Experimental Social Psychology, 32*, 484-511.

Gilbert, D. T. (1991). How mental systems believe. *American Psychologist, 46*, 107-119.

Hasson U., & Glucksberg, S. (2006). Does understanding negation entail affirmation? An examination of negated metaphors. *Journal of Pragmatics, 38*, 1015-1032.

Howell, S. R., Jankowicz, D., & Becker, S. 2005. A model of grounded language acquisition: Sensorimotor features improve lexical and grammatical learning. *Journal of Memory and Language*, *53*(2), 258-276.

Kaup, B. (2001). Negation and its impact on the accessibility of text information. *Memory & Cognition, 29*, 960-967.

Kaup, B., Lüdtke, J., & Zwaan, R.A. (2006a). Processing negated sentences with contradictory predicates: Is a door that is not open mentally closed? *Journal of Pragmatics, 38*, 1033-1050.

Kaup, B., Yaxley, R.H., Madden, C.J., Zwaan, R.A., & Lüdtke, J. (2006b). Experiential simulations of negated text information. *The Quarterly Journal of Experimental Psychology, 60*, 976-990.

MacDonald, M.C., & Just, M.A. (1989). Changes in activation levels with negation. *Journal of Experimental Psychology, 15*, 633-642.

Madden, C.J. & Zwaan, R.A. (2003). How does verb aspect constrain event representations? *Memory & Cognition, 31*, 663-672.

Mayo, R., Schul, Y., & Burnstein, E. (2004). "I am not guilty" vs "I am innocent": Successful negation may depend on the schema used for its encoding. *Journal of Experimental Social Psychology, 40*, 433-449.

McKinstry, C., Dale, R., & Spivey, M.J. (2008). Action dynamics reveal parallel competition in decision making. *Psychological Science, 19*, 22-24.

Spivey, M.J., Grosjean, M. & Knoblich, G. (2005). Continuous attraction toward phonological competitors. *Proceedings of the National Academy of Sciences, 102*, 10393-10398.

Zwaan, R.A., Stanfield, R.A., & Yaxley R.H. (2002). Language comprehenders mentally represent the shape of objects. *Psychological Science, 13*, 168-171.

On-line Interactions of Context and Grammatical Aspect

Sarah E. Anderson (sec57@cornell.edu)[a],
Teenie Matlock (tmatlock@ucmerced.edu)[b],
Michael J. Spivey (spivey@ucmerced.edu)[b]
[a] Department of Psychology, Cornell University, Uris Hall, Ithaca, NY 14853 USA
[b] Cognitive and Information Sciences, University of California, Merced, Merced, CA 95344 USA

Abstract

What role does *grammatical aspect* play in the time course of understanding motion events? Although processing differences between past progressive (*was walking*) and simple past (*walked*) aspect suggest differences in prominence of certain semantic properties, details about the temporal dynamics of aspect processing have been largely ignored. The current work uses mouse-tracking (Spivey, Grosjean, & Knoblich, 2005) to explore motor output in response to contextual descriptions and aspectual forms. Participants heard descriptions of terrain (difficult or easy) and motion events described with either the past progressive or simple past while placing a character into a scene to match this description. Overall, terrain descriptions modulated responses to past progressive more than to simple past in the region of the screen corresponding to the path. These results, which suggest that perceptual simulation plays a role in the interpretation of grammatical form, provide new insights into the understanding of event descriptions.

Keywords: Language Processing; Event Understanding; Mouse-Tracking; Embodied Cognition; Motion Verbs

The emerging consensus is that language influences thought (see Boroditsky, 2000; Lucy, 1992; Matlock, Ramscar, & Boroditsky, 2005), but the extent to which this generalizes is uncertain. *How* does language influence the way people think about everyday events? Can grammar influence the way events are conceptualized, and if so, how? Does hearing just "-ed" versus "-ing" on a motion verb influence listeners' cognitive processing and motor response, and if so, how? The goal here is to explore the influence of grammatical aspect on the conceptualization of motion events. The main question is how grammar and contextual descriptions differentially influence motor output as people understand language.

Many language theorists have observed that grammatical aspect provides temporal information about the internal structure of a verb, specifically providing information about the completion, duration, or repetition of the action (Comrie, 1976; Frawley, 1992). This temporal information, though subtle, can exert a substantial influence on the way a sentence is understood. Take, for example, the following sentences: "David ran to the university," and "David was running to the university." Both convey information about an event that occurred in the past, although they use different aspectual forms. The first sentence uses the simple past form of the verb "ran," emphasizing the completion of the action. In contrast, the second uses the past progressive form, emphasizing the ongoing nature of the action. Despite agreement that aspect provides such temporal "coloring" of a verb's information, potential processing differences between these aspectual forms have received little attention in psycholinguistic research.

More recently, however, aspect was explored in a series of offline studies that examined spatial outcome differences in response to simple past and past progressive verbs (Matlock, Fausey, Cargill, & Spivey, 2007). Participants read a sentence like "This morning David walked to the university" (simple past) or "This morning David was walking to the university" (past progressive), and looked at a schematic drawing that showed a path leading to the destination described in the sentence and ten unevenly spaced identical silhouette characters on the path (e.g., pedestrian with leg extended forward and arms bent as if in motion). Participants were instructed to "circle the man that the sentence is most naturally referring to." In brief, participants were more likely to circle a character in the middle region of the path with sentences containing past progressive verbs (*was walking*), and more likely to circle a character in the latter region of the path in response to sentences containing a simple past verb (*walked*). A similar pattern emerged in a subsequent experiment in which participants were asked to indicate where along the path an object had been dropped after reading simple past or past progressive sentences. These and other results suggest that when participants read simple past sentences, they focus on the end of the path, or the location of the completed action in the scene. In contrast, when participants read past progressive sentences, they focus on the middle section of the path, where the ongoing action would be taking place. These data suggest that different aspectual forms have consequences for thinking about motion events, but questions about the time course of processing remain.

On-line processing differences were initially addressed in a series of experiments by Madden and Zwaan (2003), in which the authors demonstrated different aspectual forms produce reaction time differences in narrative reading. Their participants were quicker to respond to pictures showing a completed action after they had read a sentence containing a simple past verb (e.g., The car sped through the intersection) versus a sentence containing a past progressive verb (e.g., The car was speeding through the intersection). However, no such latency differences were found when participants read sentences containing past progressive verbs and saw pictures of intermediate action. The authors suggest that the effect was missing in the past progressive condition because readers represented the ongoing action at different stages of completion. In other words, past progressive verbs could potentially correspond to any of a number of intermediate actions, and this could be un-captured by picture verification

and reaction time tasks. Therefore, their results suggest that different aspectual forms lead to processing differences in real time. (For other work on aspect and spatial representation, see Ferretti, Kutas, & McRae, 2007; Magliano & Schleich, 2000; Morrow, 1985).

Such reaction time data have revealed valuable insights into the processing of aspect. However, as suggested by the work of Madden and Zwaan (2003), they are somewhat limited when investigating diffuse representations. In addition to such offline and reaction time experiments, there is a great deal of information about real-time cognitive processing in the dynamics of the response. For example, evidence suggests that factors influencing latency to respond also influence later aspects of response dynamics meaning that the temporal dynamics of the motor movement that executes a response contain volumes of virtually untapped data. As a simple example, Abrams and Balota (1991) asked participants to perform a lexical decision task by making rapid limb movements in opposite directions to indicate whether a string of letters was a word or not. As expected, they found that the frequency of the word influenced reaction time, with high frequency words eliciting faster responses than low frequency words. Also, they found that word frequency influenced response kinematics after the response was initiated. Responses to high frequency words were executed with greater force than responses to low frequency words (Abrams & Balota, 1991). These findings suggest that word frequency not only influences the time required to recognize a word, but also influences response dynamics, implying that the response system is not slavishly executing a completed command regarding the categorical status of the word. This makes a compelling case for looking not only at reaction time differences, but also at variables of the motor movements themselves initiated in response to a stimulus.

To better understand the potential differences in the on-line processing of different aspectual forms, we have employed the methodology of *computer-mouse tracking*. Monitoring the streaming x- and y-coordinates of goal-directed mouse movements in response to spoken language is a useful indicator of underlying cognitive processes. In contrast to ballistic saccades, arm movements allow for a continuous, smooth motor output within a single trial to complement eye-tracking research. Spivey, Grosjean, and Knoblich (2005) demonstrated that these mouse movements can be used to index the continuous activation of lexical alternatives. By recording the x,y coordinates of the mouse as it moved with the goal-directed hand motion to click on the appropriate object, competition between the partially activate lexical representations was revealed in the shape and curvature of the hand-movement trajectories.

Further, some of our own data indicates that mouse-tracking is useful and informative for exploring research questions on the on-line processing of grammatical aspect (Anderson, Fausey, Matlock, & Spivey, 2008). In one experiment, participants listened to sentences like, "Tom jogged to the woods and then stretched when he got there," or "Tom was jogging to the woods and then stretched when he got there." While participants heard these sentences, they saw scenes consisting of a path curving upwards from left to right, and terminating at the destination described in the sentence. A character was located to the right of the beginning of the path and under the destination, separated from the scene by a black box framing the destination and path. Similar to our earlier offline results, participants dropped the character closer to the center of the path with past progressive verbs and closer to the destination with simple past verbs. Further, the two aspectual forms elicited significantly different movement durations: Participants moved the character into the scene for a longer duration of time with past progressive verbs than when they heard sentences containing simple past verbs. These drop location and movement duration results converge with and further inform earlier research, supporting that past progressive aspect focuses attention on the on-going nature of the action while simple past aspect focuses attention on the end state of that action, even during real time processing.

In the current experiment, we sought to extend these findings by investigating the way verbal aspect may interact with terrain descriptions. Research has shown that context descriptions interact with fictive motion verbs to produce both differences in patterns of eye movements and in reaction times (Matlock, 2004; Richardson & Matlock, 2007). However, the impact of such descriptions on grammatical aspect has not been explored. Here we use mouse-tracking methodology to investigate how different aspectual forms interact with similar context descriptions. Participants heard two sentences. The first provided a contextual description of the path and the second manipulated grammatical aspect. For example, on target trials participants heard a context sentence describing the path as either difficult (i.e., "The road to the university was rocky and bumpy") or easy ("i.e., "The road to the university was level and clear"), before a target sentence containing either a simple past verb (i.e., "David walked to the university where he sat in class") or a past progressive verb (i.e., "David was walking to the university where he sat in class"). While hearing these sentences, participants saw scenes containing a diagonal path that originated halfway up the screen and extended from the extreme left to the top and center of the screen (corresponding to the destination in the sentence). The orientation of the path was changed to this short, diagonal path from the long, curvy path of earlier research (Matlock, et al., 2007; Anderson, et al., 2008) to allow for more thorough and precise investigations of potential spatial and movement duration differences. A character was located to the right of the beginning of the path and under the destination. It was separated from the scene by a black box framing the destination and path.

We explored several hypotheses. If past progressive verbs sentences elicit more attention to the path, then the effect of context description was expected to be greater with past progressive verbs than when they contained simple past verbs. Specifically, we predicted that context would modulate movement durations and spatial attraction to the path more in the past progressive sentences than in the simple past verb sentences. Further, we wanted to explore the influence of the visual scene's path on movement durations.

The visual scene---with a path starting halfway up the screen---would enable us to examine if the trajectories produced in response to each aspectual form would reliably differ for the entire trajectory of the hand or only in the region of the screen corresponding to the path. If differences emerged across the entire trajectory, then the effect of grammatical aspect would appear to be more global, and to exert influence across the entire event description. However, if differences emerged only in the region of the screen corresponding to the path, then the effect of grammatical aspect would appear to be specific to the parts of the event it describes.

Method

Participants. A total of 64 undergraduates at Cornell University participated in the experiment for extra credit in psychology courses. All participants were right handed and native speakers of American English.

Materials. Twelve sentences were created from adapting the stimuli used in the offline studies of Matlock et al. (2007). As we hoped to elicit movements across the extent of the scene, from which we could extract differences in motor dynamics between the two conditions, a final clause that described an event at the destination was added, encouraging movement all the way to the destination. Similarly, two contexts for each stimulus were created. Hence, four versions of each of the 12 experimental items were created, as shown in (1) below: (1a) rough context description, simple past verb, (1b) rough context description, past progressive verb, (1c) smooth path description, simple past verb, (1d), smooth path description, past progressive verb.

1a) *The road to the university was rocky and bumpy*. / David walked to the university where he sat in class.
1b) *The road to the university was rocky and bumpy*. / David was walking to the university where he sat in class.
1c) *The road to the university was level and clear*. / David walked to the university where he sat in class.
1d) *The road to the university was level and clear*. / David was walking to the university where he sat in class.

Sentences were recorded using a Mac-based speech synthesizer program. Each of the 12 experimental items was spliced in order to produce both a past progressive and a simple past version, ensuring that the prosody of both of the targets were otherwise identical. Similarly, the context description was spliced onto the beginning of each of these target sentences. A pause of one second separated the offset of each context sentence from the onset of the target sentence. The experimental items were counterbalanced across four presentation lists. Each list contained three instances of each condition, so that all participants heard all twelve target sentences, but only heard one version of each.

Corresponding visual scenes were created for each target sentence pair. Each target visual scene consisted of a diagonal path starting halfway up and on the extreme left side of the screen. The path slanted to the right, terminating in the middle at the top of the screen. A character was located to the right of the beginning of the path and under the destination, separated from the scene by a black box framing the destination and path. See Figure 1. The depicted items in the scene were taken from clipart and edited in Adobe Photoshop. The only moveable item in the scene was the character, which subtended an average of 1.53 degrees of visual angle in width by 2.05 degrees in height. The destinations were an average of 11.22 degrees of visual angle in width by 4.09 degrees in height, and the path itself occupied a square of 8.42 degrees of visual angle in width by 6.11 degrees of visual angle in height. The character was located 14.25 degrees of visual angle from the destination. The stimuli were presented using Macromedia Director MX, and mouse movements were recorded at an average sampling rate of 40 Hz. The display resolution was set to 1024 x 768.

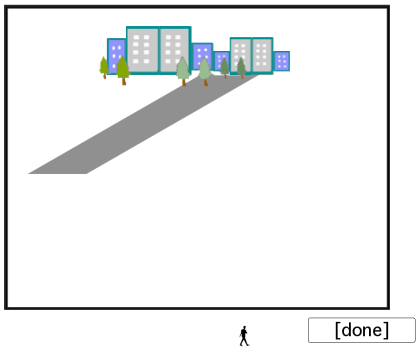


Figure 1: Example target visual scene accompanied by target sentences (1a, 1b, 1c, or 1d).

Additionally, to keep participants from developing strategies specific to the experimental sentences, 12 filler items were created. The fillers were of the same form as the target sentences: each contained a context description and either a past progressive or simple past verb. These filler trials varied from the target trials such that the context description provided no information about the path (i.e., "The weather in the valley was warm and humid") and such that they described no movement along the path (i.e., "Janet swam in the pool and then dried in the sun,"). These filler items were accompanied by 12 filler scenes, created using a short path beginning on the right side of the screen and slanting to the top, center of the screen. Besides the direction of the path, each filler scene was quite similar to the target scenes, for instance, character outside of a scene that contained the path and destination mentioned in the filler sentence.

Procedure. Participants were asked to make themselves comfortable in front of the computer, and allowed to adjust the mouse and mouse-pad to a location that suited them. First, participants read instructions to place the character in the scene to make the scene match the sentences they heard. Upon signaling to the experimenter that they understood the task, they were next presented with two practice trials (similar in form to the filler trials), followed by the experimental task. At the onset of each trail, participants were presented with the entire visual scene. After a 500 ms preview, the sound file began. After the participant had

moved the character (though not to any particular location), a "Done" button appeared in the bottom left corner of the screen. Participants clicked this button to move to the next trial. A blank screen with a button in the center labeled "Click here to go on" separated trials from each other. The entire experiment lasted approximately 20 minutes.

Results

Mouse movements were recorded during the grab-click, transferal, and drop-click of the character in the experimental trials. Prior to the analyses, the data were screened to remove extremely long trials. Movement durations 20 seconds or more were removed because they constituted an unusually long time for a mouse-movement. Using this criteria, only three trials(less than 0.4%) of trajectories, were excluded.

Drop Locations. Previous offline results revealed that participants chose a location closer to the middle of the path as the best representative of a sentence containing a past progressive verb, while selecting a location closer to the destination as the best representative of a sentence containing a simple past verb (Anderson, et al., 2008, Matlock, et al., 2007). By plotting the drop point (location along the path where each participant let go of the mouse to "drop" the character) in each of the four conditions, the current results demonstrate a similar trend. See Figure 2. There was not a significant interaction of terrain description and verb aspect (p's > .5). However, there was a main effect of verb aspect when comparing the average drop x-coordinate, F (1,62) = 8.462, p < 0.05, with the average drop x-coordinate being further left (closer to the path) when participants heard past progressive verbs (M = 476.71, SD = 68.81) than when they heard simple past verbs (M = 494.82, SD = 61.74). Similarly, there was a main effect of aspect when comparing the average drop y-coordinate, F (1,62) = 6.048, p < 0.05, with the average drop y-coordinate being lower (further from the destination) when participants heard past progressive verbs (M = 219.04, SD = 37.02) than when they heard simple past verbs (M = 210.65, SD = 41.01). This tendency to drop a character closer to the path in the past progressive condition, and close to the destination in the simple past condition, replicates previous evidence that the ongoing nature implied by a past progressive verb draws attention to the middle portion of the path, whereas there is a tendency to focus attention on the destination in response to simple past verbs.

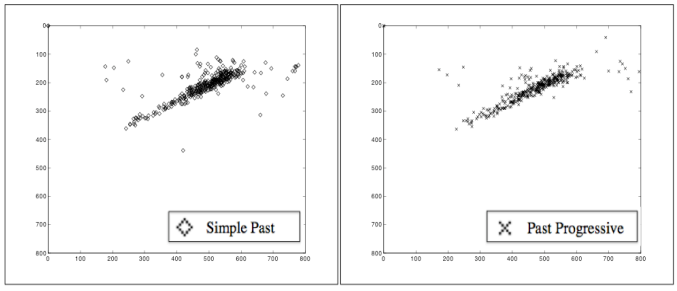


Figure 2: Drop locations in response to simple past verbs (left panel) and past progressive verbs (right panel).

Movement Durations. We began our investigation of online processing by looking at the temporal dynamics of the movement of the character. There was no significant interaction of context and aspect when comparing overall movement durations (i.e., the length of time from the initial grab of the character to the final drop of the character into the scene), p's> .2. There was a significant interaction of context and aspect on movement durations specifically *within in the region of the screen corresponding to the depicted path*, F(1, 63) = 4.6, p < .05. See Figure 3. In the region of the path, the average movement duration for simple past verbs was not substantially different when the context was first described as rough (M= 2448.33, SD=1848.88) or smooth (M=2478.72, SD= 1527.17). On the other hand, the average movement duration in the region of the path for the past progressive verbs was slower when the context was first described as rough (M = 2667.70, SD=1679.86) than when it was described as smooth (M=2121.88, SD=1240.13). Because simple past verbs focus attention on completed action, context descriptions do not significantly impact the movement dynamics. On the other hand, because past progressive verbs encourage attention to the ongoing-ness of the action, context descriptions of the location of that ongoing action do influence processing. These data extend previous research significantly, suggesting that aspect influences the real-time movement dynamics of the event being matched and that these dynamics are sensitive to visual information. Also, as predicted, the context descriptions modulate this on-line measure when aspect focuses attention to the ongoing action of the verb.

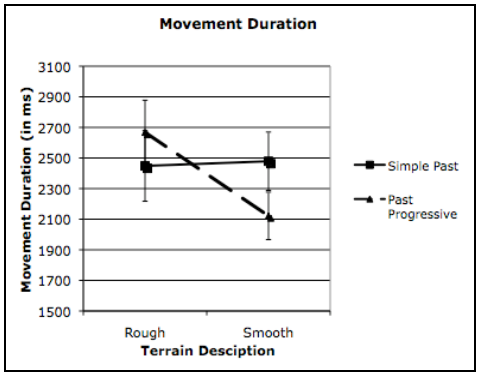


Figure 3: Movement duration differences in the region of the visual scene corresponding to the path.

Raw Time Analyses: To begin looking at the differences in spatial attraction to the visual scene's path across conditions, we first looked at average x- and y-coordinates within eight 500ms time-bins of the movement duration. There was no significant interaction between aspect and terrain, p's>.1, or main effect of either variable, p's>2. However, breaking the movement into time bins serves only as an approximation of actual attraction over raw time. These 500ms time-bins were not time locked to the sound files, and hence did not have a

fixed starting time. Because the offset of verb occurred late within the sound files and because many participants did not begin to move the character until after the end of the sound file (with an average 1400 ms lag between offset of verb and end of sentence), these data are not synchronized to a fixed point. Future work will address potential raw time spatial differences more fully.

Spatial Attraction. Figure 4 shows the average time-normalized trajectories in each of the four conditions. The mean simple past and past progressive trajectories at each of the 101 time-steps in the top panel of Figure 4 illustrate that in the rough terrain context, the average past progressive trajectory curved more toward the path than the average trajectory elicited by the simple past sentences, but only near the end of the trajectory. However, in the smooth terrain description, (Figure 2, bottom), there appears to be greater attraction toward the path across a greater portion of the trajectory for the past progressive verbs.

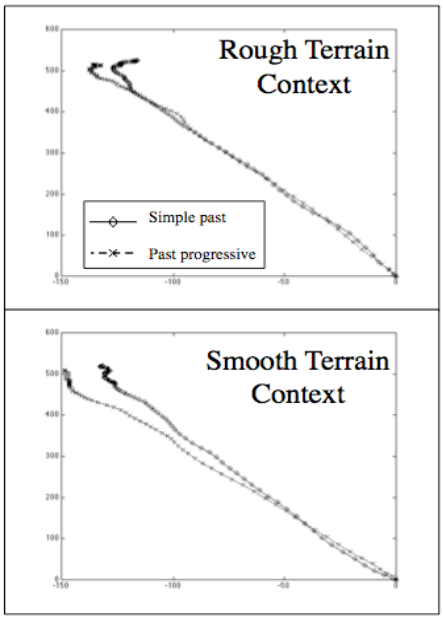


Figure 4: Average time-normalized simple past and past progressive trajectories in rough and smooth terrain contexts.

To determine whether the divergences observed across the simple past and past progressive sentence trajectories in the rough and smooth terrain descriptions were statistically reliable, we conducted a series of t-tests. These analyses were conducted separately on the x- and the y-coordinates at each of the 101 time-steps. In order to avoid the increased probability of a Type-1 error associated with multiple t-tests, and in keeping with Bootstrap simulations of such multiple t-tests on mouse trajectories (Dale, Kehoe, & Spivey, 2007), an observed divergence was not considered significant unless the coordinates between the simple past- and past progressive-sentence trajectories elicited p-values < .05 for at least eight consecutive time-steps.

In the rough context description condition, there was significant divergence of the past progressive x-coordinates away from the simple past x-coordinates and toward the path between time-steps 89 and 101, *p*'s <.05, and no significant divergence in the y-coordinates. This difference is commensurate with the observed differences in drop locations for past progressive and simple past verbs described earlier. Even though there was no significant interaction between aspect and context description on drop location, this significant divergence so late in the time-normalized trajectories may simply be an artifact of drop locations.

On the other hand, in the smooth context description, there were significant divergences of the past progressive x-coordinates away from the simple past x-coordinates towards the path between time steps 48 and 60, *p*'s <.05, and again between time steps 65 and 89, *p*'s < .05. There was also significant divergence of the average past progressive y-coordinates away from the average simple past y-coordinates and towards the path between time steps 89 and 101. Again, this divergence late in the trajectory may be an artifact of the drop locations in each condition.

While these results are encouraging, they are not as convincing as the path-movement duration results (Figure 3). It is curious that the spatial attraction differences were detected in the smooth context description but not as robustly in the rough context description. Perhaps the visual stimuli used to depict the path simply did not appear to afford difficult or uneven travel, and the incongruence in the linguistic description and the visual appearance of the path hindered the emergence of full spatial differences in this context description. Future work is slated to investigate this possibility.

General Discussion

The results reported here are consistent with previous research using mouse-tracking (Anderson, et al., 2008), narrative reading (Madden & Zwaan, 2003), and offline judgment tasks (Matlock, et al., 2007). They also provide new evidence that different grammatical forms influence the processing of event descriptions, with the simple past (e.g., *walked*) focusing attention on the end of the path and the location of the completed action, and past progressive (e.g., *was walking*) focusing attention to the "middle" of the event and the spatial region of that ongoing action. In addition to corroborating previous work on grammatical aspect, these data also reveal new insights about processing through the examination of continuous motor output in response to aspectual and contextual differences.

First, drop locations reliably differed between aspectual forms, with the past progressive condition eliciting drop locations closer to the path, and the simple past condition eliciting drop locations closer to the destination. These data are in line with earlier research, and were not significantly altered by terrain description. Contextual descriptions did interact significantly with verbal aspect in movement durations, specifically within the region of the screen depicting the path. Contextual descriptions did not significantly modulate simple past movement durations, because of the simple past's emphasis on the completed

action. However past progressive movement durations were significantly faster when preceded by an easy terrain description than when preceded by a rough terrain description. Because these differences emerge only in the region of the path in the visual scene, but not in the overall trajectory, these data suggest that grammatical aspect exerts processing influences specific to the parts of the event it describes.

Similarly, while the coarse measure of raw-time spatial attraction to the path did not reveal statistically significant results, there was a significant spatial divergence of the past progressive trajectory away from the simple past trajectory and toward the destination in both contextual descriptions. Divergences late in the trajectory may be a result of differences in drop location, but divergences across the trajectories after the smooth context description provide further evidence for processing differences between these two aspectual forms. More specifically, our results may suggest that differences in underlying perceptual simulations, resulting in these differences in the dynamics of the motor response, may account for observed processing differences.

The current research has notable implications for several areas of research. Although grammatical aspect has been considered to provide minimal semantic information by providing subtle temporal nuance, our results indicate that aspect can significantly influence on-line processing. This work also investigates grammatical aspect using a novel approach, allowing for the examination of more fine-grained temporal information, which complements the existing reaction time data. In addition, our results provide evidence to support cognitive linguists' claims regarding meaning as a conceptualization of linguistic descriptions, and the idea that aspect, like many domains of language, involves dynamic conceptualization (Langacker, 1987: Talmy, 2000).

More broadly, this work resonates with embodied cognition work on perceptual simulation and language understanding (Barsalou, 1999). It also dovetails with the methodological advances of Balota and Abrams (1995) by providing new evidence from the temporal dynamics of a response after the response has been initiated, and demonstrating that the motor system is not a robot-like automaton triggered by completed cognitive processes. Rather, motor processes are co-extensive with cognitive processes during perceptual/cognitive tasks (e.g., Balota & Abrams, 1995; Gold & Shadlen, 2000; Spivey et al., 2005; This work also comports with our understanding of how mental models and visual information are coordinated in motor output. Similarly to the way understanding of spatial events is created and observed through tracking eye movements (Richardson & Matlock, 2007; Spivey & Geng, 2001), this work demonstrates that event understanding takes place differently as a function of changes in context descriptions and grammatical aspect. Finally, the work explores a new way that language may influence thought.

References

Abrams, R.A. & Balota, D.A. (1991). Mental chronometry: Beyond reaction time. *Psychological Science, 2*, 153-157.

Anderson, S. E. . Matlock, T., Fausey, C., & Spivey, M.J. (2008). On the path to understanding on-line processing of grammatical aspect. *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 143-148), Mahwah, NJ: Lawrence Erlbaum Associates.

Barsalou, L. (1999). Language comprehension: Archival memory or preparation for situated action? *Discourse Processes, 28*, 61-80.

Balota, A. D. & Abrams, R. A. (1995). Mental chronometry: Beyond onset latencies in the lexical decision task. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 21*, 1289-1302.

Boroditsky, L. (2000). Metaphoric structuring: Understanding time through spatial metaphors. *Cognition, 75*, 1-28.

Comrie, B. (1976) *Aspect*. Cambridge: Cambridge University Press.

Dale, R., Kehoe, C., & Spivey, M. (2007). Graded motor responses in the time course of categorizing atypical exemplars. *Memory and Cognition, 35,* 15-28.

Ferretti, T. R., Kutas, M., McRae, K. (2007). Verb aspect and the activation of event knowledge in semantic memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 33*, 182-196.

Frawley, W. (1992). *Linguistic semantics.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Gold, J. I. & Shadlen, M. N. (2000). Representation of perceptual decision in oculomotor commands, Nature, 404, 390-394.

Langacker, R. W. (1987). *Foundations of cognitive grammar: Theoretical prerequisites*. Stanford, CA: Stanford University Press.

Lucy, J.A. (1992). Grammatical Categories and Cognition: A Case Study of the Linguistic Relativity Hypothesis. Cambridge: Cambridge University Press.

Madden, C.J. & Zwann, R.A. (2003). How does verb aspect constrain event representations? *Memory & Cognition, 31*, 663-672.

Magliano, J.P. & Schleich, M.C. (2000). Verb aspect and situation models. *Discourse Processes, 29*, 83-112.

Matlock, T. (2004). Fictive motion as cognitive simulation. *Memory & Cognition, 32*, 1389-1400.

Matlock, T., Ramscar, M., & Boroditsky, L. (2005). On the experiential link between spatial and temporal language. *Cognitive Science, 29,* 655-664.

Matlock, T., Fausey, C., Cargill, S., & Spivey, M. (2007, November). On the path toward understanding the dynamics of aspect descriptions in motion events. Paper presented at 48th Annual Meeting of the Psychonomic Society, Long Beach, California.

Morrow, D.G. (1985). Prominent characters and events organize narrative understanding. *Journal of Memory and Language, 24*, 304-319.

Richardson, D. C., & Matlock, T. (2007). The integration of figurative language and static depictions: An eye movement study of fictive motion. *Cognition, 102*, 129-138.

Spivey, M. & Geng, J. (2001). Oculomotor mechanisms activated by imagery and memory: Eye movements to absent objects. *Psychological Research, 65*, 235-241.

Spivey, M.J., Grosjean, M. & Knoblich, G. (2005). Continuous attraction toward phonological competitors. *Proceedings of the National Academy of Sciences, 102*, 10393-10398.

Talmy, L. (2000). *Toward a cognitive semantics*. Cambridge, MA: MIT Press.

Anaphora and Local Coherences

Lars Konieczny (lars@cognition.uni-freiburg.de)
Helmut Weldle (helmut@cognition.uni-freiburg.de)
Sascha Wolfer (sascha@cognition.uni-freiburg.de)
Daniel Müller (daniel@cognition.uni-freiburg.de)
Peter Baumann (peter@cognition.uni-freiburg.de)
Center for Cognitive Science, University of Freiburg, Friedrichstr. 50
D-79098 Freiburg i. Br., Germany

Abstract

We present two visual world studies indicating that local syntactic coherences interact with binding constraints (Chomsky, 1981) of both reflexives and pronouns. Gazes to depicted referents or events suggest that when sentences containing a local coherence with a pronoun or reflexive are presented, locally coherent antecedents become activated. Our results strengthen the assumption that local syntactic coherences are interpreted and extend the effect to online anaphora resolution and complementary binding constraints.

Keywords: anaphora, anaphora resolution, local syntactic coherences

Introduction

To arrive at a coherent interpretation of a sentence, we need to bind referring expressions to their correct referent. Binding theory (Chomsky, 1981) provides a syntax-driven structural account for the dependencies of co-reference within sentences. Principles based on c-command are assumed to constrain the possible co-referents of anaphoric expressions on a global level. Reflexives and pronouns have complementary structural binding domains, i.e. within sentence boundaries, the accessible antecedents for both anaphora types are mutually exclusive. In sentences like (1-a) and (1-b) determining the referent of the anaphoric expression *himself* or *him* is straightforward.[1]

(1) a. Ken$_i$ who likes John$_j$ saw himself$_{i/*j}$ in the mirror.
b. Ken$_j$ who likes John$_i$ saw him$_{i/*j}$ in the mirror.

Recent findings question strictly structure-driven accounts of anaphora resolution. Runner, Sussman, and Tanenhaus (2006), for instance, report violations of the binding domain complementarity assumption. They examined preferences for pronoun and reflexive binding in picture noun phrases in a series of *visual-world* studies. Fixation probabilities on depicted referents revealed violations of the binding theory assumption. They concluded that reflexives should rather be explained in terms of logophors, deposing reflexives beyond the scope of Binding theory explanations.

Converging evidence was found by Kaiser, Runner, Sussman, and Tanenhaus (2009), suggesting that the interpretation of reflexives is not only sensitive to structural but also semantic information. Moreover, they found differing degrees of sensitivity towards different sources of information for pronouns and reflexives.

Sturt (2003) on the other hand showed that the constraining principle for reflexives operates at the very earliest stages of processing. In eye-tracking-while-reading experiments, he found early effects of binding preferences. He concludes that the responsible binding principle is an early filter for the processing of referring expressions.

Interpretation of locally syntactic coherences

Local syntactic coherences (LSCs) have been shown to interfere with the global sentence interpretation. Tabor, Galantucci, and Richardson (2004) found increased reading times on the spill-over of *tossed* in sentences like *The coach chided the player tossed a frisbee by the opposing team*. Moreover, Konieczny, Müller, Baumann, Hachmann, and Wolfer (2009) have shown that LSCs temporarily affect the interpretation of globally unambiguous sentences.

Interestingly, there are locally coherent substrings in (1), leading to the opposite binding of the reflexive or the pronoun: In Sentence (1-a), *himself* is restricted t bound to *John* if only the local subparse *John saw himself* is taken into account. In the global parse however, *himself* is bound to *Ken*. In Sentence (1-b), *him* is bound to *Ken* or any other (unmentioned) referent if the LSC *John saw him* is interpreted.

It is still an open question though, whether or not LSCs can affect anaphora resolution, and if pronouns and reflexives are affected equally.

In the remainder of the paper, two experiments will be reported providing insight into the time-course of binding and its interaction with local syntactic coherences. We chose the *visual-world* paradigm (Cooper, 1974; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995) over reading times, as fixations on visual objects can indicate binding preferences in a much more direct way, without inferencing over processing difficulty. The results suggest that LSCs indeed have an effect on the binding of reflexives and pronouns in a way that strong constraints like *Principle A* are also applied to the local parse, temporarily overriding the globally correct binding.

Experiment 1: Depicted referents in the visual world

In the first experiment, the visual stimuli showed the depictions of three persons, two of which were depictions of the referents introduced in the spoken sentence. When the

[1] Subscripted indices mark coreference. A star * indicates that coreference with the indexed referent is not acceptable considering the global parse but suggested considering the local parse.

antecedents get reactivated when a pronoun or reflexive is processed, we should see inreased gaze proportions on the corresponding referent. Following this logic, the gaze pattern should indicate how local syntactic coherences affect anaphora resolution.

Materials and design

We tracked participant's gaze on depictions of three referents (cf. Figure 1) while they listened to sentences like (2). Materials were constructed according to a 2x2-design crossing the factors anaphor type (reflexive vs. pronoun) and presence of LSC (present vs. prevented), leading to four experimental conditions: reflexive with LSC (Sentence (2-a)), reflexive without LSC (Sentence (2-b)), pronoun with LSC (Sentence (2-c)), and pronoun without LSC (Sentence (2-d)). We prevented the LSC by inserting an adverb before the verb of the relative clause. Each participant was presented with 48 experimental sentences and an equal amount of sentence-picture pairs of comparable complexity. The task was to look at the pictures while listening to the sentences. Immediately after the sentence, participants had to click on the agent they considered to be most important in the scene. 25 participants took part in Experiment 1.

Auditory stimuli

We recorded 48 sentences with normal speech tempo. Locally coherent sequences were recorded separately – as main clauses (*Der Sohn kämmte sich im Wohnzimmer* ...) – and spliced into surrounding sentences (without the starting determiner of the LSC to prevent sentence-initial prosody). By doing so, we wanted to minimize prosodic cues induced by the relative clause boundary and thereby destroying the local coherence. To minimize prosodic differences between conditions, control conditions were produced by splicing the adverbs (*gründlich/thoroughly*) into the first two conditions. This method was necessary since earlier findings indicated strong sensitivity to prosodic cues (Konieczny et al., 2009). These effects are outside the scope of this study. The resulting experimental stimuli still sounded natural, as was established in a pre-test with native speakers who where naïve with respect to the research questions at hand.

(2) a. *Reflexive, LSC*
Während der Vater$_i$, den der Sohn$_j$ kämmte,
While the father$_i$, who the son$_j$ combed,
sich$_{i/*j}$ im Wohnzimmer anzog, ...
himself$_{i/*j}$ in the living room dressed, ...
While the father who the son combed dressed himself in the living room, ...

b. *Reflexive, no LSC*
Während der Vater$_i$, den der Sohn gründlich
While the father$_i$, who the son combed,
kämmte, sich$_i$ im Wohnzimmer anzog, ...
thoroughly himself$_i$ in the living room dressed, ...
While the father who the son combed thoroughly dressed himself in the living room, ...

c. *Pronoun, LSC*
Während der Vater$_i$, den der Sohn$_j$ kämmte,
While the father$_i$, who the son$_j$ combed,
ihn$_{j/*i}$ im Wohnzimmer anzog, ...
him$_{i/*j}$ in the living room dressed, ...
While the father who the son combed dressed him in the living room, ...

d. *Pronoun, no LSC*
Während der Vater, den der Sohn$_j$ gründlich
While the father, who the son$_j$ thoroughly
kämmt, ihn$_j$ im Wohnzimmer anzieht, ...
combs, him$_j$ in the living room dresses, ...
While the father who the son combed thoroughly dresses him in the living room, ...

Visual stimuli The visual stimuli of Experiment 1 consisted of three referents. The globally suggested agent, the agent contained in the LSC and a third, non-mentioned referent (see Figure 1).The positions of the referents was cross-balanced over all trials and participants.

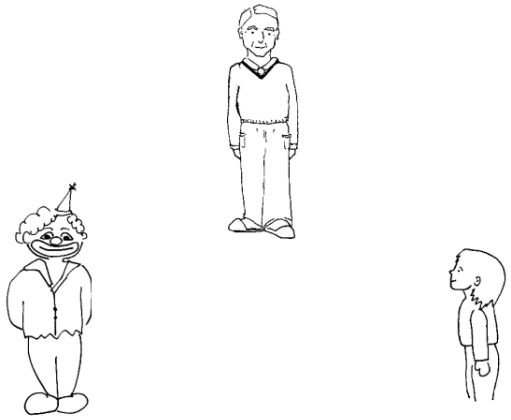


Figure 1: A picture of three referents from Experiment 1. The father (top), the son (right), and a clown (left).

Hypotheses

We expect the LSC to have an influence on fixation probabilities when the anaphor is heard or shortly after. If the LSC has an effect, we expect fixation probabilities on the referent denoted by the LSC to be increased, compared to sentences where no LSC is present. Therefore, when sentences like (2-a) are heard, fixation probabilities on the *son* should increase compared to sentences like (2-b) because the son is the only possible referent of the local interpretation of *sich/himself*. Accordingly when sentences like (2-c) are heard, we expect fixation probabilites on the *father* to increase compared to sentences like (2-d) because the father is one possible referent of the local interpretation of *ihn/him*. We expect the effect for the reflexive condition to be stronger than for the pronoun, because the reflexive

sich/himself corefers in its local interpretation with the agent of the LSC (the *son*) whereas the pronoun *ihn/him* corefers in its local interpretation with the first-mentioned agent (the *father*) but also with every other referent except the *son*.

Results

For reflexives (Figure 2), there are significantly more fixations on the referent denoted by the LSC, when the LSC was present than when it was absent. Fixation probabilities differ significantly in the range from 500 ms to about 1600 ms after the onset of the reflexive in the spoken stimulus. The referent denoted by the global meaning of the sentence is fixated the most after about 1200 ms after the synchronization point, indicating that the participants understood the spoken stimuli. The non-overlapping standard errors indicate significant differences of mean fixation probabilities in the different conditions. This was further validated by fitting a linear mixed effects model using the statistical software R (R Development Core Team, 2010) with the package lme4 (Bates, 2007). Analyses of fixation patterns for the relevant sections revealed significant differences, as tested with MCMC-sampling (all $ps < .05$). However, for the pronoun condition, we found no reliable difference in fixations on the referent denoted by the LSC when a LSC was present compared to when it was absent (Figure 3).

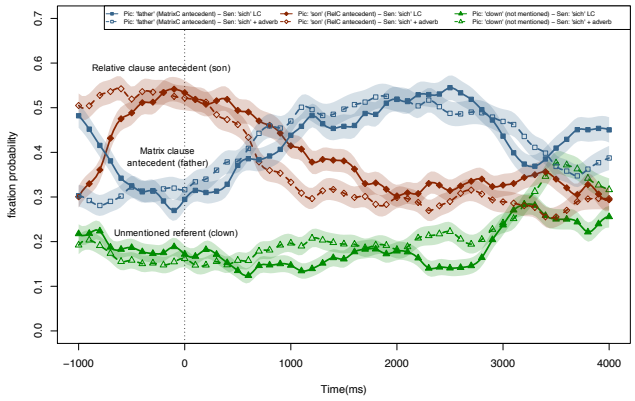


Figure 2: Results for *reflexives*: Proportion of fixations on the *globally correct referent* (blue lines), the *local referent* (red), and on an *unmentioned person* (green). Zero marks the onset of the reflexive in the spoken stimulus. There are significantly more fixations to the local referent (red lines), when a LSC was present (solid) than in the control condition (dashed), in the range of about 500 ms to 1600 ms.

Discussion

The results clearly indicate that binding can be disturbed by a local syntactic coherence, at least for reflexives. The lack of a local coherence effect for pronouns in this experiment might have several reasons: Due to their complementary binding domains, pronouns would have their antecedent outside the

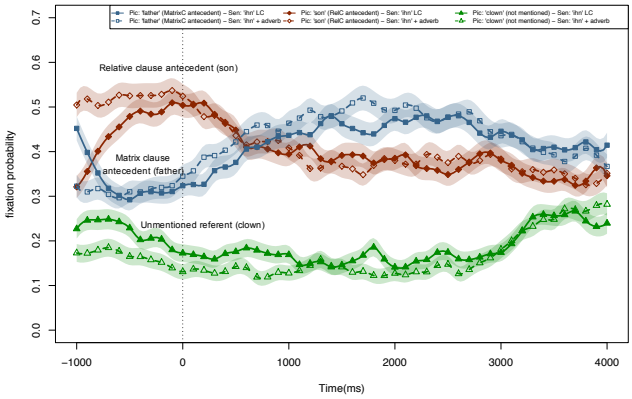


Figure 3: Results for *pronouns*: Proportion of fixations on the *globally correct referent* (blue lines), the *local referent* (red), and on an *unmentioned person* (green). Zero marks the onset of the pronoun in the spoken stimulus. There was no reliable difference in fixations to the local referent (red lines), when a LSC was present (solid) than in the control condition (dashed).

actual local coherence. Another possibility is that participants might have looked at both of the two actors (the *agent* and the *patient*) in transitive events, such that the gazes are not as informative for the pronoun cases as for the reflexive cases, where there is only one actor (the *agent*) in each event. These, and other potential problems of Experiment 1 were dealt with in Experiment 2.

Experiment 2: Depicted events in the visual world

In Experiment 2, we used depicted events instead of referents. By doing so, we were able to provide a unique depiction even for the transitive (i.e. pronoun) cases. We also used a different control condition (i.e. locally non-coherent condition), which is more effective than inserting an adverb before the verb, as in Experiment 1. We used particle verbs, such as *an-ziehen*, in the non-local conditions. In main clauses, the particle *an* would have to be seperated and placed at the end of the clause, as in: der Sohn *zieht* sich *an*.

Materials and design

Again, we used a 2x2-design crossing the type of anaphora (reflexive vs. pronoun) and the presence of the LSC (present vs. absent). Note, that the type of scene (depiction of transitive or reflexive actions) was always presented with the corresponding type of anaphor. So, reflexive scenes like those in Figure 4 were always presented with sentences like (3-a) and (3-b), transitive scenes (Figure 5) were presented with sentences like (3-c) and (3-d). The task was to click on (one of the) correct scenes after hearing the sentence. 36 participants took part in Experiment 2.

Auditory stimuli We recorded 24 sentences and used the double amount of sentence-picture pairs as fillers. Again, the local coherent substring was spliced into the surrounding sentence. To create the control conditions, we swapped the verb of the relative clause (*kämmte/combed*) with the second verb of the sentences (3-a) and (3-c) (*anzog/dressed*), which was always a particle verb. When a particle verb is placed inside the relative clause, the LSC is no longer valid (* *Der Sohn anzog sich/ihn im Wohnzimmer* ...).

(3) a. *Reflexive, LSC*
Während der Vater$_i$, den der Sohn$_j$ kämmte,
While the father$_i$, who the son$_j$ combed,
sich$_{i/*j}$ im Wohnzimmer anzog, ...
himself$_{i/*j}$ in the living room dressed, ...
While the father who the son combed dressed himself in the living room, ...

b. *Reflexive, no LSC*
Während der Vater$_i$, den der Sohn anzog, sich$_i$
While the father$_i$, who the son dressed, himself$_i$
im Wohnzimmer kämmte, ...
in the living room combed, ...
While the father who the son dressed combed himself in the living room, ...

c. *Pronoun, LSC*
Während der Vater$_i$, den der Sohn$_j$ kämmte,
While the father$_i$, who the son$_j$ combed,
ihn$_{j/*i}$ im Wohnzimmer anzog, ...
him$_{i/*j}$ in the living room dressed, ...
While the father who the son combed dressed him in the living room, ...

d. *Pronoun, no LSC*
Während der Vater, den der Sohn$_j$ anzog, ihn$_j$
While the father, who the son$_j$ dressed, him$_j$
im Wohnzimmer kämmte, ...
in the living room combed, ...
While the father who the son dressed combed him in the living room, ...

Visual stimuli We used scene depictions instead of referents. Sentences with reflexives were always presented with scenes depicting reflexive actions (Figure 4), whereas sentences with pronouns were always presented with scenes depicting transitive actions (Figure 5). This procedure lead to a total of 192 scenes. The positions of the scenes were cross-balanced over all trials and participants.

In the reflexive conditions (Sentences (3-a) and (3-b)) the global interpretation of Sentence (3-a) refers to the scene where *the father is dressing himself* (the lower left scene in Figure 4). The LSC is referring to *the son combing himself* (the upper left scene in Figure 4).

Because we generated the control conditions by swapping the verbs, the depicted target and control events were different. That is, for locally coherent sentences the target scene was *the son combing himself* (the upper left scene in Figure 4), whereas for the non-coherent controls, the corresponding scene was *the son dressing himself* (the upper right scene in Figure 4). Accordingly, the globally correct scene for Sentence (3-b) changes from *the father dressing himself* (the lower left scene in Figure 4) to *the father combing himself* (the lower right scene in Figure 4).

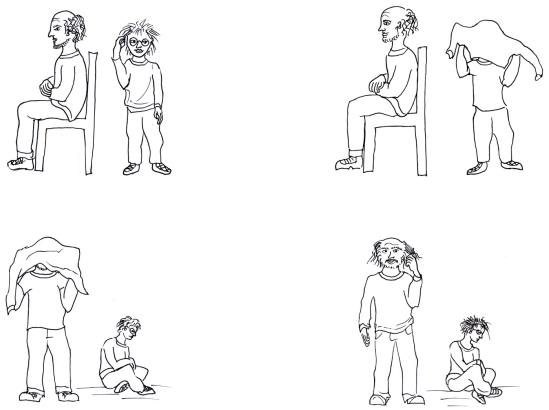


Figure 4: A picture of four scenes depicting reflexive actions from Experiment 2. The son combing himself (upper left), the son dressing himself (upper right), the father dressing himself (lower left), and the father combing himself (lower right).

In the pronoun condition with an LSC (Sentence (3-c)) the global sentence meaning refers to the scene where *the son is combing the father* (the upper left scene in Figure 5) as well as to the scene where *the father is dressing the son* (the lower right scene in Figure 5). The meaning of the LSC is also referring to the scene where *the son is combing the father*. Importantly, in the pronoun condition with an LSC, both global and local meanings are referring to the same scene.

Again, the control condition was generated by swapping the verbs, therefore the depicted target control events were also different. For locally coherent sentences the target scene was the *the son combing the father* (the upper left scene in Figure 5), whereas for the non-coherent controls, the corresponding scene was *the son dressing the father* (the upper right scene in Figure 5). Accordingly, the globally correct scene for Sentence (3-d) changes from *the father dressing the son* (the lower right scene in Figure 5) to *the father combing the son* (lower left scene in Figure 5).

Hypotheses

Reflexive condition We again expect a contrast in fixation probabilities between sentences with LSC (Sentence (3-a)) and sentences without LSC (Sentence (3-b), such that the picture depicting the event expressed by the LSC (*Der Sohn kämmt sich im Wohnzimmer*) is fixated more often when the LSC is present than when it is not (* *Der Sohn anzog sich im Wohnzimmer*). This effect should show up shortly after the offset of the reflexive. The other event, which is described by the main clause (*the father dressing* (LSC) *or combing* (no

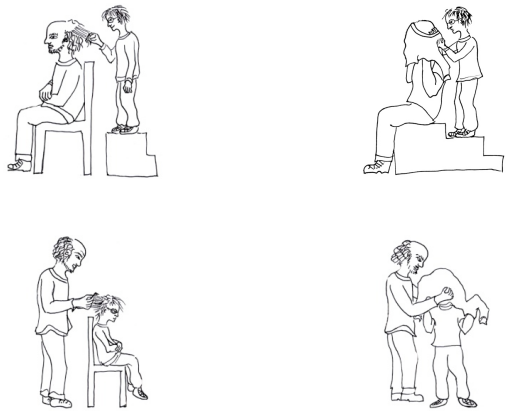


Figure 5: A picture of four scenes depicting transitive actions from Experiment 2. The son combing the father (upper left), the son dressing the father (upper right), the father combing the son (lower left) and the father dressing the son (lower right).

LSC) *himself*) should be fixated later as soon as the event is mentioned.

Pronoun condition In the pronoun condition both the global and the local meaning of the sentence refer to the same scene (*the son combing the father*), therefore we expect a "boost effect" on fixation probabilities on this scene, i.e. we expect fixation probabilities to be high in the control condition (Sentence (3-d)), but even higher in the local coherent condition (Sentence (3-c)). Furthermore, fixation probabilities on the event described in the main clause should rise as soon as the corresponding event is mentioned, i.e. in the LSC condition, the fixations on the main verb action depiction should be delayed until after the offset of the local coherence.

Results

For reflexives (Figure 6)[2], there are significantly more fixations to the scene denoted by the LSC when the LSC was present than when the LSC was absent. This effect ranges from about 800 ms to after 2000 ms from the onset of the reflexive. Of course, during the matrix clause at the end of the sentence, the highest proportion of fixations is on the scene denoted by the matrix clause (after about 2200 ms from the onset of the reflexive).

For pronouns (Figure 7), the meaning of the LSC coincides with the globally correct meaning of the relative clause, which are therefore depicted by the same scene. The significant difference between the LSC and the non-LSC condition demonstrates the expected "boost effect". This effect lasts from 200 ms to about 1000 ms after the onset of the pronoun. When the event denoted by the matrix clause is described, there are the most fixations on the corresponding scene (after about 2200 ms). Again, the indication of significant differences by non-overlapping standard errors were further validated by fitting a linear mixed effects model (all ps $< .05$).

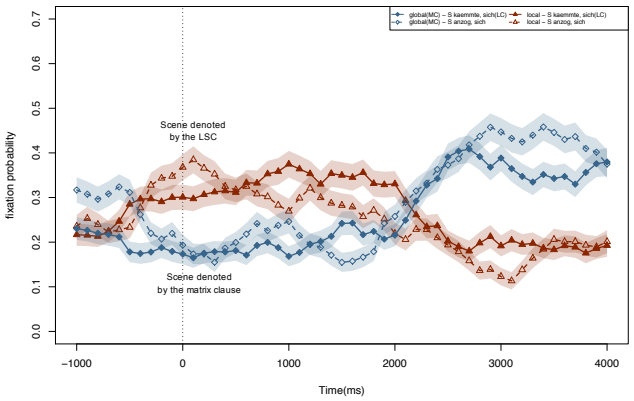


Figure 6: Results for *reflexives*: Proportion of fixations on the depicted event with the *globally correct binding* (blue lines), the *local binding* (red). Fixations on the other depictions are not plotted. Zero marks the onset of the reflexive in the spoken stimulus. There are significantly more fixations to the local binding depiction (red lines), when a LSC was present (solid) than in the control condition (dashed), in the range of about 800 ms to after 2000 ms.

Discussion

Experiment 2 replicates the results for reflexives in Experiment 1. Ignoring for a minute the period where the effect seems to be reversed, about 800 ms after the reflexive we see the local coherence effect, i.e. more looks to the locally coherent, or the control scene, respectively, when there is a local coherence in the speech input, than when there is none. Different from experiment 1, we also found a local coherence effect with pronouns. As discussed above, this might be due to the fact that the target picture in Experiment 2 is a single scene including the directionality of the action, whereas there are two depicted target actors involved in transitive actions in Experiment 1. Most notably, interpreting the pronoun within the local coherence overlaps with the meaning of the relative clause itself, so that the effect amounts to boosting the correct interpretation. The short inverse effect for reflexives starting even before the reflexive and lasting to about 200 ms after the onset of the reflexive might seem worrying at first glance. Note however that this effect is likely due to the fact that different target scenes where used for the target and the control condition (due to swapping the verbs). This difference could hence be attributed to differences in visual saliency between the two depictions. Moreover, the swapped verbs themselves might have added to the effect: particle verbs clearly morphologically indicate a clause boundary, whereas the verbs used for local coherences do not. Detection of a clause boundary might have triggered a short-lived attention-shift towards pic-

[2]For expository reasons only fixation probabilities on the locally and globally denoted scenes were plotted.

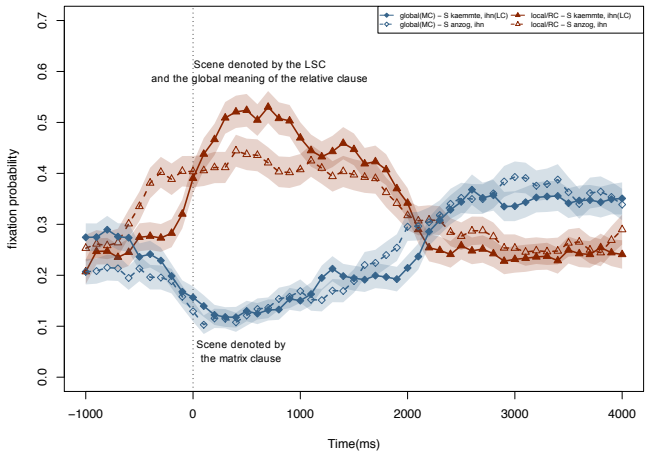


Figure 7: Results for *pronouns*: Proportion of fixations on the *globally correct scene described by the matrix clause* (blue lines) and the *scene described by the LSC and the global meaning of the relative clause* (red). Fixations on the other depictions are not plotted. Zero marks the onset of the pronoun in the spoken stimulus. There are significantly more fixations to the local binding depiction (red lines), when a LSC was present (solid) than in the control condition (dashed), in the range of about 200 ms to after 1000 ms.

tures depicting the verb's action. Note that the same holds for the pronoun cases, where we also see an early and short-lived advantage for the "local" picture in the non-local condition. The early appearance of the effect however, as in the reflexives, renders it unlikely that the effect is due to the local coherence itself.

Conclusions

Both experiments have shown that local syntactic coherences can influence the binding of pronouns and reflexives. The results suggest that LSCs open a short-lived window, during which binding constaints can work both locally and globally. With respect to the time course, shortly after an anaphoric expression is heard, potential referents or scenes corresponding to the binding of the anaphor are fixated. This argues for very early constraints (from both global and local interpretations) exerting their influence on co-reference assignment. This result is in line with Sturt (2003) who showed that *Principle A* operates at the very earliest stages of processing. However, we could also show that globally correct binding can be delayed when a local syntactic coherence interferes.

Furthermore, our findings suggest that the effects for pronouns are more fragile than those for reflexives, replicating similar findings by Kaiser et al. (2009) and Runner et al. (2006) who found that binding constraints for reflexives are harder than those for pronouns. Different to their findings, our results are not dependent on the specific semantic nature of the stimuli, as is the case with picture noun phrases.

On a larger scale, our results can be interpreted in two ways. They could be seen as an indicator that binding principles for pronouns and – especially – for reflexives are too restrictive because they only capture the by nature global structural characteristics of sentences. On the other hand our results speak for the validity of Binding Theory, because it is even applicable in non-global structures like the local syntactic coherences presented here.

What seems quite clear considering our results, is that LSCs are processed and interpreted in such a profound way that they exert a clear influence, even on the binding of anaphoric expressions.

Acknowledgments

We would like to thank Anne Karina Feldmeth, David Kühner and Minkus Teske for their detail work on the auditory and visual stimuli. We also want to thank Sarah Schwarzkopf for her voice and for recording the auditory stimuli. This work was supported by the German Research Foundation (DFG, grant KO 1932/3-1).

References

Bates, D. M. (2007). Linear mixed model implementation in lme4. *Manuscript, University of Wisconsin - Madison*.

Chomsky, N. (1981). *Lectures on government and binding*. Dordrecht: Foris.

Cooper, R. M. (1974). The control of eye fixation by the meaning of spoken language: A new methodology for the real time investigation of speech perception, memory and language processing. *Cognitive Psychology*, *6*, 84-107.

Kaiser, E., Runner, J. T., Sussman, R. S., & Tanenhaus, M. K. (2009). Structural and semantic constraints on the resolution of pronouns and reflexives. *Cognition*, *112*(1), 55-80.

Konieczny, L., Müller, D., Baumann, P., Hachmann, W., & Wolfer, S. A. (2009). Local syntactic coherence interpretation, and how prosody modulates it. In *Proceedings of the 22nd annual meeting of the cuny conference on human sentence processing*. Davis, CA.

R Development Core Team. (2010). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria.

Runner, J. T., Sussman, R. S., & Tanenhaus, M. K. (2006). Assigning referents to reflexives and pronouns in picture noun phrases: Experimental tests of binding theory. *Cognitive Science*, *30*, 1-49.

Sturt, P. (2003). The time-course of the application of binding constraints in reference resolution. *Journal of Memory and Language*, *48*, 542-562.

Tabor, W., Galantucci, B., & Richardson, D. (2004). Effects of merely local syntactic coherence on sentence processing. *Journal of Memory and Language*, *50*, 355-370.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*, 1632-1634.

Self-directed speech alters visual processing

Gary Lupyan (lupyan@sas.upenn.edu)
Institute for Research in Cognitive Science, Center for Cognitive Neuroscience
University of Pennsylvania
Philadelphia, PA 19104 USA

Daniel Swingley (swingley@psych.upenn.edu)
Department of Psychology, Institute for Research in Cognitive Science
University of Pennsylvania
Philadelphia, PA 19104 USA

Abstract

A major part of learning a language is learning connections between spoken words and their referents in the world. An open question concerns the consequence this learning has for cognition and perception. According to the label feedback hypothesis (Lupyan, 2007), processing a verbal label can change ongoing perceptual processing, e.g., actually hearing "chair" compared to simply thinking about a chair temporarily makes the visual system a better chair detector. Here, we test whether engaging in a non-communicative verbal act—speaking to oneself—also affects visual processing. Participants searched for common objects, sometimes being asked to speak the target's name aloud. Speaking facilitated search, but only when there was a strong association between the name and the visual target. Speaking appeared to hurt performance when there was even a slight discrepancy between the name and the target. Together these results speak to the power of words to evoke associated visual information.

Introduction

Learning a language involves, among other things, learning to map words onto categories of objects in the environment. In addition to learning that chairs are good for sitting on, one learns that this class of objects has the name "chair." Clearly, this learning is critical for linguistic communication. But beyond communication, what consequences does naming things—hearing and producing verbal labels—have on perception and nonverbal cognition?

On one account language is a "transparent medium through which thoughts flow" (H. Gleitman, Fridlund, & Reisberg, 2004, p. 363). Therefore, words are mapped onto concepts, but do not affect them (e.g., L. Gleitman & Papafragou, 2005; Gopnik, 2001). Thus, while word-learning is significantly constrained by nonverbal cognition, nonverbal cognition is not significantly influenced by learning or using words (Snedeker & L. Gleitman, 2004).

The alternative is that words are not simply mapped on to concepts, but actually change them, affecting nonverbal cognition, and even perception. The idea that words can affect the concepts to which they refer is not new: William James, for example, remarked on the power of labels to make distinctions more concrete (James, 1890, p. 333), and it has been argued that words stabilize abstract ideas in working memory and make them available for inspection (Clark, 1997; Clark & A Karmiloff-Smith, 1993; Dennett, 1994; Goldstein, 1948; Rumelhart, Smolensky, McClelland, & Hinton, 1986; Vygotsky, 1962). This is not to say that different languages necessarily place strong constraints on the speaker's *ability* to entertain certain concepts. Rather, it is a claim that language richly interacts with putatively nonlinguistic processes such as visual perception.

Insofar as performance on putatively nonverbal tasks draws on language, interfering with language should interfere with performance on those tasks (Goldstein, 1948). Indeed, verbal interference impairs certain types of categorization in a way strikingly similar to impairments observed in aphasic patients (Lupyan, 2009). Interfering with language can also affect perception. A number of studies have shown that interfering with language impairs categorical color perception (e.g., Gilbert, Regier, Kay, & Ivry, 2006; Roberson & Davidoff, 2000; Roberson, Pak, & Hanley, 2008; Winawer et al., 2007), suggesting that language actively modulates visual processing.

An additional way to study affects of language on perception is by attempting to *increase* rather than decrease its putative effect. A surprising finding is that when asked to find a certain visual item among distractors actually hearing its name immediately prior to performing the search—even when the label is entirely redundant—improves speed *and efficiency* of searching for the named object (or searching among the named objects). For example, when participants search for the numeral 2 among 5's (for hundreds of trials), actually hearing the word "two" (or hearing "ignore fives") immediately prior to doing the search, improves search RTs and reduces search slopes (Lupyan, 2007a, 2008a). Indeed, hearing an object name can even make an otherwise invisible object visible (Lupyan & Spivey, 2008; under review).

One way to understand such findings is in terms of an interactive activation framework (Rumelhart & McClelland, 1982; Spivey, 2008) in which recognition involves the combination of bottom-up perceptual information, with higher-level top-down (conceptual) information. As one learns a verbal label, it becomes associated with features that are most diagnostic (or typical) of the named category. With such associations in place, hearing the label provides top-down activation of visual properties associated with the label. In effect, the object name makes an object a "better" object by augmenting the idiosyncratic perceptual features of a given object with features typical to the named category (Lupyan, 2007b, 2008b).

Aims and Hypotheses

In the present work, we investigate whether non-communicative (self-directed) speech can affect visual processing in the context of a visual search task. Does producing the name of a pre-defined target object enable subjects to find it faster? Participants were asked to find an object among distractors while speaking its name or not. We predicted that actually speaking the object's name would facilitate visual search—even though such speaking can be seen to constitute a form of distraction. We also predicted that the effect of speaking would be largest for items most strongly associated with the label, and speaking might actually be detrimental when searching for objects having weaker associations with the label, e.g., objects judged as being less typical of their categories.

Experiment 1

The participants' task was to find and click on a target object among 35 distractors, positioned randomly in a 6×6 grid on a computer screen (Figure 1). For half the trials, participants were asked to speak the name of the target as they searched for it.

Participants

Twelve University of Pennsylvania undergraduates participated for course credit.

Materials

The targets and distractors were drawn from a set of 260 colored images of common objects (Rossion & Pourtois, 2004). For the targets, we selected 20 images having 100% picture-name agreement, as assessed by Rossion and Pourtois (2004) (airplane, banana, barn, butterfly, cake, carrot, elephant, giraffe, chicken, ladder, lamp, leaf, truck, motorcycle, mouse, mushroom, rabbit, tie, umbrella, windmill).



Figure 1: A sample search trial from Exp. 1

For a given trial, any of the 259 non-target images could serve as distractors. Rossion and Pourtois provide a number of measures for these pictures, which we included for item-analyses. Most relevant to the present work are: RT to name the picture, familiarity, subjective visual complexity, and imagery-concordance. The latter measure was derived by presenting participants with a picture name (e.g., butterfly), asking them to form a mental image of the object, and then, on seeing the actual picture, providing a rating of imagery agreement. For the lexical items themselves, we obtained log frequency from the British National Corpus, word length in phonemes and syllables, actual age-of-acquisition (AoA) norms (Morrison, Chappell, & Ellis, 1997), and several measures from the MRC Psycholinguistic Database (www.psych.rl.ac.uk/): imageability, concreteness, and word familiarity.

Procedure

Each trial began with a prompt informing the participant what object they would need to find. The prompt also informed them whether they should repeat the object's name as they searched for it, or not. For example, immediately prior to a no-speaking trial, participants saw a prompt such as "Please search for a butterfly. Do not say anything as you search for the target" For a speaking trial, the second sentence was replaced by "Keep repeating this word continuously into the microphone until you find the target." The speech/no-speech trials were intermixed, as were the target identities. Participants completed 320 trials: 20 targets × speech condition (speaking vs. not speaking) × 8 blocks. A block included all target × speech condition combinations. Participants used a computer mouse to click on the target object.

Results and Discussion

Participants showed excellent compliance with the instruction to speak the name of the target on the label trials and to remain silent on the no-speaking trials. We focus on accuracy and median RTs to find the target as the main dependent measures. Comparisons between conditions were made using a mixed-effects ANCOVA with speech condition as a fixed effect, subject as a random effect, and block as a covariate. For reasons described in Thomas et al., (2009), separate tests were run to assess fixed factor main effects and those of the covariate × factor interaction.

Accuracy was extremely high, M=98.8%, revealing that (1) subjects had no trouble remembering which item they were supposed to find, and (2) the word cues were sufficiently informative to locate the correct object. Despite this very high accuracy, saying the object's name during search resulted in significantly higher accuracy, M=99.2% than not repeating the name, M=98.4%, $F(1,11)=12.19$, $p=.005$. Participants' accuracy increased over the course of the experiment, $F(1,11)=10.90$, $p=.001$, but there was no reliable speech-condition × accuracy interaction, $F(1,11)=1.49$, $p>.2$.

The analysis of median RTs included correct responses only. Unsurprisingly, participants' speed improved over the course of the experiment, $F(1,11)=22.85$, $p<.0005$. There

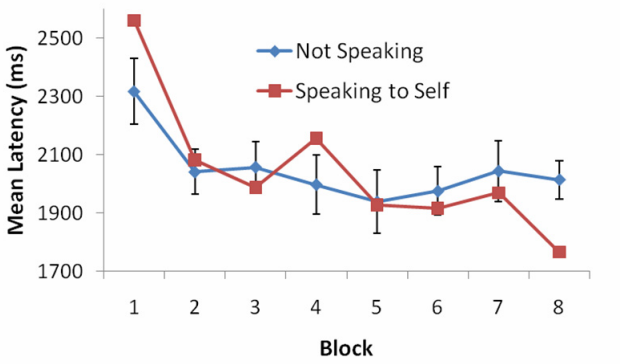


Figure 2: RTs in Exp. 1: Speaking significantly decreased RTs for the second half of the task. Error bars show ±1SE of the mean condition difference. Accuracy was significantly higher for the speaking condition throughout the task; see text.

was no main effect of the speech-condition on RTs, $F<1$, but there was a highly reliable speech-condition × block interaction, $F(1,11)=8.1$, $p=.004$. As shown in Figure 2, performance on the speech trials tended to be slower than on no-speech trials for the initial blocks, but this pattern reversed for the latter part of the experiment. Collapsing the last three blocks, participants were faster on speech trials than no-speech trials, $t(11)=2.91$, $p=.01$ (two-tailed). This finding suggests that although the target objects were very familiar, speaking the name decreased RTs only when participants had several opportunities to associate the picture name with the target picture, which presumably strengthened the picture-name association.

We next turn to the item analysis. A number of item factors predicted overall search performance. Search was faster, $r(18)=.55$, $p=.01$, and more accurate, $r(18)=-.54$, $p=.02$, for pictures that were visually simpler according to Rossion and Pourtois's (2004) norms. Search was faster, $r(18)=-.55$, $p=.01$, and slightly more accurate, $r(18)=.34$, $p=.15$ for pictures with higher imagery-concordance. Familiarity did not predict search times or accuracy. Lexical

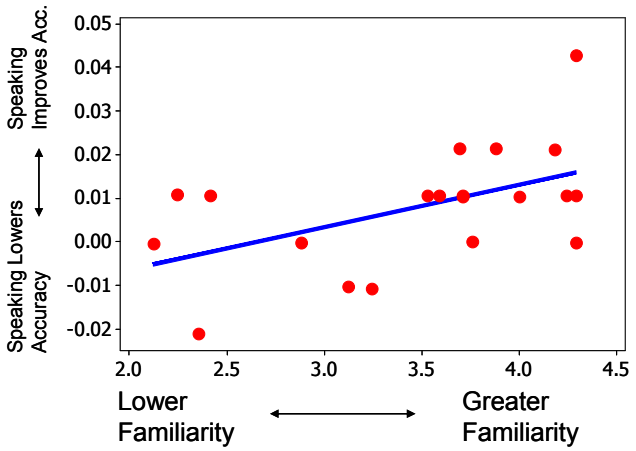


Figure 3: Relationship between item familiarity and effects of speaking on accuracy. Y-axis shows % correct when speaking - % correct when not speaking.

variables did not predict overall search performance, though there were marginal correlations of search times with word frequency, $r(18)=-.38$, $p=.10$, and of accuracy with age-of-acquisition (AoA) provided by adults, $r(18)=-40$, $p=.08$.

Finally, we examined which items were most affected by self-directed speech by subtracting performance on speech trials from performance on no-speech trials. Overall, speaking improved RTs most for the items which took, on average, the least time to find, $r(18)=-.57$, $p=.009$, and ones for which accuracy was, on average, the highest, $r(18)=.47$, $p=.037$. Recall that familiarity was not related to overall accuracy. However, separating accuracy into speech and no-speech trials revealed a very different pattern. Familiarity was unrelated to performance on no-speech trials, $p>.3$, but was highly correlated with performance on speaking trials, $r(18)=.55$, $p=.01$. The interaction was significant: speaking improved accuracy most for the more familiar items, $r(18)=.51$, $p=.02$ (Figure 3). Finally, RTs were improved marginally more for the items with the highest imagery-concordance, $r(18)=.39$, $p=.08$.

We also observed a relationship between AoA and self-directed speech. This relationship changed over the course of the experiment: for the first half of the task, AoA (both subjective and objective), correlated with the effect of speaking on search times, $r_{\text{objective AoA}}(28)=-.54$, $p=.02$, $r_{\text{subjective AoA}}=-.62$, $p=.003$: performance was impaired by saying words having higher AoA. By the second half of the task, these correlations disappeared entirely, rs<.1.

For interpretive ease, we performed a median split on the familiarity and imagery-concordance values. The label advantage ($RT_{\text{without-speaking}}$-$RT_{\text{speaking}}$) was larger for items having imagery-concordance scores above than below the median, $F(1,18)=6.32$, $p=.022$. Search items below the median were actually slowed by speaking, $t(10)=2.24$, $p=.049$ (two-tailed). The label advantage in accuracy trended in the same direction, being (marginally) larger for items with above-median familiarity ratings, $F(1,18)=4.19$, $p=.056$.

To summarize: speaking facilitated search for pictures judged in a separate norming study to be most familiar, and targets having the highest concordance between the actual image and the mental image formed by reading the name.

One way in which self-directed speech may help visual search is through verbal rehearsal: saying the name of the target might have helped participants remember what it was they were looking for. This account is not supported for two reasons. First, accuracy was extremely high, making it unlikely that difficulties in remembering the target played a significant role. Second, a memory-based account would predict that speech should help most for items that were most difficult to find. We found exactly the opposite pattern.

The item effects presented above place some constraints on the mechanisms by which labels affect visual search. One possibility is that saying the target name helps to find the target by activating and/or keeping active the visual features typical to that object (e.g., saying "cherry" makes it easier to attend to red things). Alternatively (or addition-

ally), repeating a label helps to reject distractors. If speaking facilitated search only by improving rejection of distractors, one would not predict correlations between the magnitude of the speaking advantage and properties of the target. The presence of these correlations supports the hypothesis that speaking the target's name facilitates deployment of attention to the target item over and above seeing the printed name of the target.

The present results can be viewed as an extension of findings showing that hearing a label, even when it is entirely redundant, facilitates visual search, and this facilitation is greatest for the stimuli most strongly associated with the label (Lupyan, 2007a, 2007b, 2008a). When visual quality of the item is reduced, or the item is made more ambiguous, hearing a label can impair performance (Lupyan, 2007b). Thus, compared to just being told what to find, speaking a target name—just like hearing it—affects visual search.

Experiment 2

The goal of Experiment 2 was to test whether self-directed speech affects performance on a more difficult and ecologically valid "virtual shopping" task in which participants search for supermarket products in a visually complex display and were required to find several instances of a category.

Participants

Twenty-two University of Pennsylvania undergraduates (14 women) participated for course credit.

Materials

We photographed products on supermarket shelves in the Philadelphia area and selected 30 to serve as targets, e.g, apples, Pop-Tarts, Raisin Bran, Tylenol, Jell-O. For each product, we obtained three pictures depicting instances of the product in various sizes and orientations. Some pictures depicted multiple instances of the product, e.g., a shelf containing multiple cartons of orange juice. See Figure 4 for some examples.

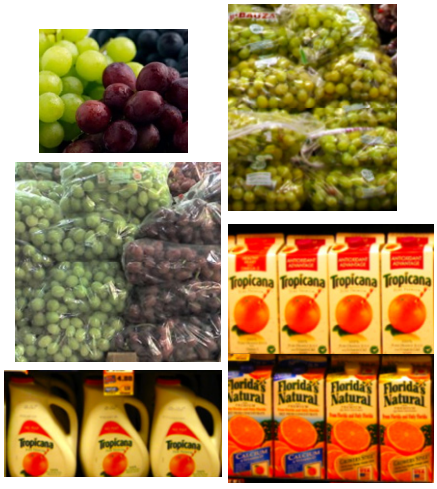


Figure 4: Samples of 2 search categories used in Exp. 2.

Procedure

As in Exp. 1, participants were instructed that they would need to search for various items while being asked to sometimes speak the items' names. Each trial included all three instances of the product and 13 distractors. Clicking on an object made it disappear, thus marking it as being selected. Once satisfied with their choices, participants clicked on a large "Done" button that signaled the end of the trial. To make the task more challenging, some of the distractors were categorically related to the target, e.g., whenever searching for "Diet Coke," some distractors were of other sodas, e.g., "Ginger Ale." There were a total of 240 trials (30 targets by × 8 blocks). Within each block, half the items were presented in a speech trial and half in a no-speech trial. Speech and no-speech trials alternated. Across the 8 blocks, each item was presented an equal number of times in speech trial and no-speech trials.

Prior to beginning the search task, participants rated each item on typicality ("How typical is this box of Cheerios relative to boxes of Cheerios in general?"), and visual quality ("How well does this picture depict a box of Cheerios?"). For each item category (i.e., all three images of Cheerios), participants rated its familiarity ("Overall, how familiar to you are the objects depicted in these pictures?") and visual similarity ("Considering only the visual appearance of these picture, how different are they from each other?"). In addition to providing us with item information, this task served to pre-expose participants to all the targets. We also obtained an imageability measure from a separate group of participants (N=28) who were shown the written product names, e.g., "Cheerios" and asked to rate how well they could visualize its appearance on a supermarket shelf.

Results and Discussion

The data were analyzed in the same way as in Exp. 1. Overall, participants were very accurate, averaging 1.5% false alarms and 97.7% hits (2.93 out of 3 targets). Overall performance (RTs, hits, and false alarms) correlated with all four item variables (visual similarity, visual quality, familiarity, and typicality). Correlation coefficients ranged from .35 to .65 (*p*s between .035 and <.0005). Items that were familiar, typical, of higher quality, and having least within-category similarity were found faster and with higher accuracy. Of course, the item variables were not all independent, e.g., familiar items and those of higher quality tended to be rated as more typical. The typicality and familiarity measures clustered together and were not independently predictive of performance (familiarity was the stronger of the two predictors). Within category visual similarity predicted performance independently of familiarity; multiple regression: $F(2,27)=9.15$, $p=.001$.

There was a reliable difference in hits between the two speech conditions: $M_{speech}=97.9\%$, $M_{no\text{-}speech}=99.1\%$, $F(1,21)=11.19$, $p=.003$. While speaking the product name, participants were more likely to miss one or more of the targets. As reported below, however, this effect was modulated strongly by the different targets in predictable ways. Speech-condition was not a reliable predictor of false-alarms, $F(1,21)<1$. There were no differences in total or per-click RTs between the speech and no-speech conditions, $F<1$. The speech-condition × block interaction was not reliable, $F<1$.

The item analyses in Exp. 1. suggested that effects of self-directed speech may be modulated by the relationship between the item and its name. Indeed, the cost in the hit rate incurred by speaking ($Hits_{no\text{-}speech}$-$Hits_{speech}$) was correlated with within-category similarity, $r(28)=-.34$, $p=.04$: the categories having the most dissimilar items incurred the highest cost when their names were repeated during search. The effect of self-directed speech ($RT_{no\text{-}speech}$-RT_{speech}) was also mediated by familiarity, $r(28)=-.51$, $p=.004$: labels tended to hurt performance for the less familiar items, but *improve*

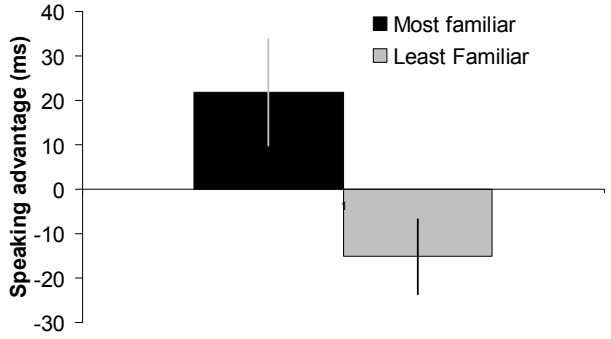


Figure 5: Speaking advantage (no-speech – speech trials) as a function of familiarity (median split). RTs were decreased by speaking the names of the more familiar items and increased by speaking the names of the least familiar items. Errors bars indicate 1±SE of the mean difference.

performance for the more familiar items (Figure 5). The label advantage also correlated positively with product imageability, $r(28)=.44$, $p=.01$. As an added confirmation of this finding, we divided the targets into those having characteristic colors (N=11), e.g., bananas, grapes, cheerios, raisin bran and those with weaker color associations, e.g., Jell-O, Pop-Tarts. The speaking advantage was greater for color-diagnostic items (for which speaking significantly improved RTs) than for non color-diagnostic items (for which speaking marginally increased RTs), $F(1,28)=7.35$, $p=.01$.

Exp. 2 revealed a striking gender difference in performance. Men had a significantly lower hit rate, $F(1,20)=5.02$, $p=.037$, and were significantly slower, $F(1,20)=6.37$, $p=.02$ to find the targets. The gender effect on RTs was substantial: men took on average 350 ms longer per trial. This effect was replicated in an item analysis, $F_2(1,29)=43.40$, $p<.0005$ (the *only* item on which men were faster than women was "Degree Deodorant"). There was a marginal gender × speech-condition interaction for hit rates, $F(1,20)=3.79$, $p=.066$: labels hurt performance slightly more for men than women. An examination of item ratings revealed that there were no gender differences in subjective ratings of familiarity, visual-quality, or visual-similarity, Fs<1, and only a marginal difference in typicality: women believed our items to be slightly more typical than did men, $F(1, 20)=2.66$, $p=.12$. In an effort to better understand the origin of this gender difference, we correlated the magnitude of the female advantage with various ratings of the stimuli. We observed a mildly reliable relationship between the magnitude of the female RT advantage and the measure of visual similarity: $r(26)=.38$, $p=.049$. The advantage was greatest for the most visually similar items (two items were excluded, as statistical outliers). There were no other reliable correlations.

Using a larger, more perceptually varied and true-to-life item set, the item analyses of Exp. 2 reinforced the conclusions of Exp. 1. As in Exp. 1, speaking aided search for the more familiar items. In contrast to Exp. 1, accuracy (hit rate) was actually decreased by speaking, though this decrease was limited to the items having low within-category similarity. This finding is consistent with the idea that speaking an object name activates a (proto)typical representation of the category. When the task requires finding items that diverge from this prototype (as when participants need to find visually heterogeneous items from the same category), speaking can impair performance.

General Discussion

Can language affect ongoing perceptual processing? A growing body of literature argues that it can. The present work is the first to examine effects of non-communicative (self-directed) speech on a visual task.

The findings show that speaking the name of the object that one is searching for improves search performance, provided that the object's name is strongly associated with the visual depiction of the object.

The present results are somewhat less reliable than those of hearing labels on visual search (Lupyan, 2007a, 2008a). Subsequent work has shown that the effects of speech on visual processing have a characteristic timecourse, peaking about 0.5-1.5 seconds after the presentation of the label, and declining afterwards (Lupyan & Spivey, 2010, under review). In the present studies we did not have precise control over the timing of the label. Recordings of participants' speech from the present work revealed a wide variability in the onset, speed, and duration of self-directed speech. Thus, more reliable effects may be obtained with finer control over speaking onset and rate.

Our results join work arguing for cognitive functions of self-directed speech. For example, even mild forms of articulatory suppression impair adults' ability to switch from one task to another (Baddeley, Chincotta, & Adlam, 2001; Emerson & Miyake, 2003; Miyake, Emerson, Padilla, & Ahn, 2004). The present results are consistent with Vygotsky's claim that the function of self-directed speech extends far beyond verbal rehearsal (Carlson, 1997; Vygotsky, 1962)—itself a learned strategy (Flavell, Beach, & Chinsky, 1966).[1]

The present work comprises a first step in understanding effects of self-directed speech on visual processing. One unanswered question is whether effects of speaking on visual search arise from the act of production itself, or from

[1] It is worth noting that these articulatory suppression effects on putatively nonverbal task-switching were compelling enough for Baddeley to concur with Vygotsky's claim (Baddeley et al., 2001, p. 655).

hearing one's speech. Although this distinction is of little practical importance—one almost always hears oneself speak—a full understanding of the mechanism by which speech and visual processing interact requires the two explanations to be teased apart. Despite these unknowns, the present results show that in the context of searching for a familiar object, knowing what an object is called is not the same as actually saying its name.

Acknowledgments

We thank Ali Shapiro, Joyce Shin, for their help with data collection and for assembling the stimulus materials.

References

Baddeley, A. D., Chincotta, D., & Adlam, A. (2001). Working memory and the control of action: Evidence from task switching. *Journal of Experimental Psychology: General*, *130*(4), 641-657. doi:doi:10.1037/0096-3445.130.4.641

Carlson, R. A. (1997). *Experienced Cognition* (1st ed.). Psychology Press.

Clark, A. (1997). *Being There: Putting brain, body, and world together again*. Cambridge, MA: MIT Press.

Clark, A., & Karmiloff-Smith, A. (1993). The Cognizer's Innards: A Psychological and Philosophical Perspective on the Development of Thought. *Mind & Language*, *8*(4), 487-519.

Dennett, D. (1994). The Role of Language in Intelligence. In *What is Intelligence? The Darwin College Lectures*. Cambridge University Press.

Emerson, M., & Miyake, A. (2003). The role of inner speech in task switching: A dual-task investigation. *Journal of Memory and Language*, *48*(1), 148-168.

Flavell, J. H., Beach, D. R., & Chinsky, J. M. (1966). Spontaneous Verbal Rehearsal in a Memory Task as a Function of Age. *Child Development*, *37*(2), 283-299.

Gilbert, A., Regier, T., Kay, P., & Ivry, R. (2006). Whorf hypothesis is supported in the right visual field but not the left. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(2), 489-494.

Gleitman, H., Fridlund, A., & Reisberg, D. (2004). *Psychology* (6th ed.). New York: Norton & Company.

Gleitman, L., & Papafragou, A. (2005). Language and thought. In *Cambridge Handbook of thinking and Reasoning* (pp. 633-661). Cambridge: Cambridge University Press.

Goldstein, K. (1948). *Language and language disturbances*. New York: Grune & Stratton.

Gopnik, A. (2001). Theories, language, and culture: Whorf without wincing. In *Language acquisition and conceptual development* (pp. 45-69). Cambridge, UK: Cambridge University Press.

James, W. (1890). *Principles of Psychology. Vol. 1*. New York: Holt.

Lupyan, G. (2007a). Reuniting categories, language, and perception. In D. McNamara & J. Trafton (Eds.), *Twenty-Ninth Annual Meeting of the Cognitive Science Society* (pp. 1247-1252). Austin, TX: Cognitive Science Society.

Lupyan, G. (2007b). *The Label Feedback Hypothesis: Linguistic Influences on Visual Processing*. PhD. Thesis. Carnegie Mellon University.

Lupyan, G. (2008a). The Conceptual Grouping effect: Categories matter (and named categories matter more). *Cognition*, *108*, 566-577.

Lupyan, G. (2008b). From chair to "chair:" A representational shift account of object labeling effects on memory. *Journal of Experimental Psychology: General*, *137*(2), 348-369.

Lupyan, G. (2009). Extracommunicative Functions of Language: Verbal Interference Causes Selective Categorization Impairments. *Psychonomic Bulletin & Review*, *16*(4), 711-718. doi:10.3758/PBR.16.4.711

Lupyan, G., & Spivey, M. (2008). Now You See It, Now You Don't: Verbal but not visual cues facilitate visual object detection. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 963-968). Austin, TX.

Lupyan, G., & Spivey, M. (2010). Redundant spoken labels facilitate perception of multiple items. *under review*.

Miyake, A., Emerson, M., Padilla, F., & Ahn, J. (2004). Inner speech as a retrieval aid for task goals: the effects of cue type and articulatory suppression in the random task cuing paradigm. *Acta Psychologica*, *115*(2-3), 123-142. doi:10.1016/j.actpsy.2003.12.004

Morrison, C. M., Chappell, T. D., & Ellis, A. W. (1997). Age of Acquisition Norms for a Large Set of Object Names and Their Relation to Adult Estimates and Other Variables. *The Quarterly Journal of Experimental Psychology A*, *50*, 528-559. doi:10.1080/027249897392017

Roberson, D., & Davidoff, J. (2000). The categorical perception of colors and facial expressions: The effect of verbal interference. *Memory & Cognition*, *28*(6), 977-986.

Roberson, D., Pak, H., & Hanley, J. R. (2008). Categorical perception of colour in the left and right visual field is verbally mediated: Evidence from Korean. *Cognition*, *107*(2), 752-762. doi:10.1016/j.cognition.2007.09.001

Rossion, B., & Pourtois, G. (2004). Revisiting Snodgrass and Vanderwart's object pictorial set: The role of surface detail in basic-level object recognition. *Perception*, *33*(2), 217-236.

Rumelhart, D., & McClelland, J. (1982). An Interactive Activation Model of Context Effects in Letter Perception .2. the Contextual Enhancement Effect and Some Tests and Extensions of the Model. *Psychological Review*, *89*(1), 60-94.

Rumelhart, D., Smolensky, D., McClelland, J., & Hinton, G. (1986). Parallel Distributed Processing Models of Schemata and Sequential Thought Processes. In *Parallel Distributed Processing Vol II* (pp. 7-57). Cambridge, MA: MIT Press.

Snedeker, J., & Gleitman, L. (2004). Why is it hard to label our concepts? In D. G. Hall & S. R. Waxman (Eds.), *Weaving a Lexicon* (illustrated edition., pp. 257-294). The MIT Press.

Spivey, M. (2008). *The Continuity of Mind*. Oxford University Press.

Thomas, M. S. C., Annaz, D., Ansari, D., Scerif, G., Jarrold, C., & Karmiloff-Smith, A. (2009). Using developmental trajectories to understand developmental disorders. *Journal of Speech, Language, and Hearing Research: JSLHR*, *52*(2), 336-358. doi:10.1044/1092-4388(2009/07-0144)

Vygotsky, L. (1962). *Thought and Language*. Cambridge, MA: MIT Press.

Winawer, J., Witthoft, N., Frank, M., Wu, L., Wade, A., & Boroditsky, L. (2007). Russian blues reveal effects of language on color discrimination. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(19), 7780-7785.

Is categorical perception really verbally mediated perception?

Andrew T. Hendrickson, George Kachergis ({athendri, gkacherg}@indiana.edu)
Department of Psychological and Brain Sciences, Indiana University
1101 East Tenth Street, Bloomington, IN 47405 USA

Todd M. Gureckis (todd.gureckis@nyu.edu)
Department of Psychology, 6 Washington Place
New York, NY 10003 USA

Robert L. Goldstone (rgoldsto@indiana.edu)
Department of Psychological and Brain Sciences, Indiana University
1101 East Tenth Street, Bloomington, IN 47405 USA

Abstract

Recent research has argued that categorization is strongly tied to language processing. For example, language (in the form of verbal category labels) has been shown to influence perceptual discriminations of color (Winawer et al., 2007). However, does this imply that categorical perception is essentially verbally mediated perception? The present study extends recent findings in our lab showing that categorical perception can occur even in the absence of overt labels. In particular, we evaluate the degree to which certain interference tasks (verbal, spatial) reduce the effect of learned categorical perception for complex visual stimuli (faces). Contrary to previous findings, our results show that a verbal interference task does not disrupt learned categorical perception effects for faces. Our results are interpreted in light of the ongoing debate about the role of language in categorization. In particular, we suggest that at least a sub-set of categorical perception effects may be effectively "language-free". **Keywords:** Perceptual Learning, Categorization, Concept Learning, Language.

Introduction

It is now well-known that the categories we know often influence the things that we perceive. For example, the phoneme categories in the native language of a listener dramatically influence their ability to perceive physical differences between two speech sounds. In particular, differences that span phonemic category boundaries are much more accurately discriminated than differences that fall within the same phonemic category (Liberman, Harris, Hoffman, & Griffith, 1957). This effect, known as Categorical Perception (CP), has been shown for many types of perceptual stimuli, and is known to be influenced by both innate and learned factors (e.g., Harnad, 1987; Goldstone, 1994; see Goldstone & Hendrickson, 2009 for a review).

Given the fact that CP effects are so ubiquitous, it is perhaps surprising that so little is known about how they arise. Theoretical analyses suggest that the very act of associating category labels with items can warp the representations of those items in the service of categorization. For example, Harnad, Hanson, & Lubin (1995) showed through neural network simulations that adding such a label, even without changing the representation space, changed the similarity of item representations in that space in a way consistent with CP effects. However, such simulations simply show how CP might *arise* without explaining the exact psychological factors that may contribute to it in humans.

On the other hand, recent work by Winawer et al. (2007) has argued that the change in representation that produces such a CP effect may be due to the inclusion of a "language-specific" component to the representation of an item in memory. In their study, Winawer and colleagues found that Russian speakers, who have unique words in their language for 'light blue' and 'dark blue,' show a standard CP effect: a higher accuracy for perceptual discriminations of blues that span the light-dark category boundary relative to blues within one category. English speakers, who only use one basic word for blue, did not show a similar CP effect for the same stimuli. Interestingly, the CP between-category advantage was eliminated for the Russian speakers when they were given a verbal interference task (repeating a string of digits) while performing the perceptual discriminations, though the CP effect was preserved if the interference task involved a spatial task (remembering a pattern) instead of a verbal task. From this, Winawer, et al. argue that linguistic processing not only influences the category learning processes, but has an online influence during perceptual discrimination as well (see also Lupyan, 2008).

Somewhat consistent with this viewpoint, learned CP effects are most often found in supervised learning tasks, where feedback about an item's correct category label drives learning to reduce classification error of category labels (Harnad, 1987; Goldstone & Hendrickson, 2009). However, Gureckis and Goldstone (2008) presented an interesting finding which would appear to challenge this view. In their study, a set of morph faces was created with varied along two arbitrary dimensions (Figure 1). Four "clusters" of items were created in the space by withholding a subset of the items from the training phase (the grey stimuli in Figure 1). Two of the clusters were assigned to category "A" and the other two clusters were assigned to category "B" by applying either a vertical or horizontal category boundary. Both before and after category learning participant's ability to make pair-wise discriminations

between items was measured. The results showed that discrimination of items within each small cluster was reduced following learning. In addition, discrimination of items across the category boundary was improved (a pattern consistent with the standard CP effect). They also found that discrimination performance was improved between clusters that belonged to the same category. These CP effects were largest in blocks in which performance on the categorization judgment task was highest, suggesting that learning drove both improvements in categorization performance as well as the changes in perceptual discrimination.

The improvement in perceptual discrimination *within* a category (and along the category-irrelevant dimension) would not be predicted if CP was only the result of verbal labeling processes since all of these items share the same label. Instead, it appears to suggest that a non-verbal learning mechanism is engaged during category learning that is sensitive to the internal structure of the categories (e.g., Love, Medin, & Gureckis, 2004). In this study we explore the hypothesis that the effect of this non-verbal learning is not impacted by verbal interference.

In our experiment, we taught participants to categorize the same set of morphed faces that have been previously shown to induce categorical perception effects in Goldstone (1994) and Gureckis & Goldstone (2008). Following the learning phase we had participants make perceptual discriminations between pairs of faces that span both the category and cluster boundaries while performing a set of spatial or verbal interference tasks. Similar to the approach adopted by Winawer, et al, our goal was to assess the impact that verbal interference has on CP of these stimuli relative to a spatial interference task. In light of these previous findings, we predict that verbal interference will disrupt the standard CP effect of improved discrimination across the category-relevant boundary by preventing online linguistic processing while a spatial interference task does not. In contrast, we predict that verbal interference would have little impact on the improved discrimination of items that belongs to different clusters within the same category (since such effects are unlikely to be driven by differences in verbal labeling). In line with previous work, we further predict that these effects will be most pronounced for perceptual discrimination judgments in blocks where categorization performance is most accurate. Our results replicate the effects of previous studies, but we found that the interference tasks had overall little effect on learned CP for our face stimuli.

An Experiment

Method

Participants 172 students at Indiana University participated in partial fulfillment of a course requirement and were assigned into one of two conditions based on which dimension (1 or 2 in Figure 1) was relevant for categorization (87 had dimension 2, and 85 had dimension 1). 61 participants were excluded who did not perform significantly above chance on either categorization or discrimination trials with no interference task (the threshold used for both tasks was 0.52, the upper threshold of a 95% confidence interval based on a binomial distribution centered at 0.5).

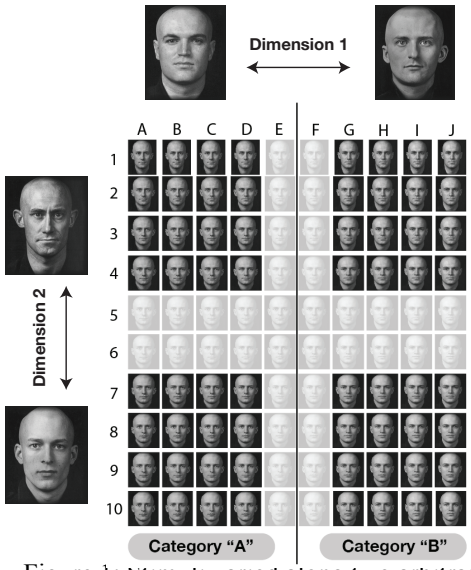


Figure 1: Stimuli varied along two arbitrary dimensions (1 and 2) forming a 10-by-10 grid of blended faces. The light grey stimuli were not included in category learning, introducing a source of within-category structure (two clusters of faces within each category). The vertical line between columns E and F shows an example category boundary used during category learning (the other category boundary was a horizontal line between rows 5 and 6).

Materials

The stimuli were morphs of bald male faces selected from Kayser (1997) using the blending technique outlined in Steyvers (1999). A stimulus space was constructed that varied along two arbitrary dimensions, each one formed by morphing between two anchor faces to create 10 faces per dimension that formed a continuum from one anchor face to the other (see Dimensions 1 and 2 in Figure 1). The two specific dimensions used in this study were selected because they were roughly equally salient and roughly orthogonal when subjected to a MDS analysis in preliminary work (Gureckis & Goldstone, 2008). A 10 by 10 matrix of stimuli faces were created by combining each face along dimension 1 with each face along dimension 2 to create a blended face that is the average of the two faces. Not all 100 faces in the 10 by 10 matrix were presented during categorization trials. In particular, a subset of faces was never presented (the light grey faces in Figure 1), creating two "clusters" of faces within each category.

Procedure

Categorization Task On each categorization trial, a single face was presented for 500 ms in the center of the display for study followed by a blank screen for 300 ms. Instructions were then presented directing participants to indicate if the correct category label for the item was 'P' or 'Q.' After a participant responded, the stimulus was again presented for 2000 ms along with feedback indicating whether the response was correct and the correct category of the stimulus.

Discrimination Task On each trial, a target stimulus was presented for 500 ms in the center of the display for study followed by a blank screen for 300 ms. For discrimination trials, immediately after the blank screen two stimuli were simultaneously presented: one stimulus that exactly matched the target stimulus and a foil stimulus. Participants were instructed to indicate by pressing one of two keys which stimulus matched the target. No feedback was provided after discrimination trials.

Sixteen stimuli were used as targets and foils in the discrimination task: the four corners of each of the four clusters shown in Figure 1 (e.g. stimuli A7, A10, D7 and D10 of the lower left-hand cluster). Each foil stimulus was two values away from the target stimulus on one of the two dimensions or two values away from the target stimulus on both dimensions (e.g. for stimulus A1 the set of foils was D1, A4, and D4; for stimulus D4 the set of foils was: A4 and G4 along Dimension 1, D1 and D7 along Dimension 2, as well as A1, G1, G7, and A7 along both dimensions).

Interference Tasks The verbal and spatial interference tasks involved participants memorizing a verbal string or a spatial pattern and recalling that information after a mini-block of categorization and discrimination trials. Verbal interference mini-blocks were preceded by the presentation of a string of 6 digits for 8 s followed by an interval of 3 s. Participants were instructed that they should memorize this string and would be tested on it later. At the end of the mini-block, memory for the studied string was probed by presenting the original string along with a foil stimulus (which had two randomly selected digits swapped). Participants simply indicated which string they recognized as the studied item by pressing one of two keys.

Spatial interference mini-blocks were preceded by the presentation of a 6 by 6 grid composed of half white squares and half black squares for 8 s followed by an interval of 3 s before the mini-block began. Participants were instructed to memorize this pattern and that they would be tested on it. At the end of the mini-block, recall of the pattern was tested by presenting the original pattern and a foil pattern that had the black-white state of one randomly selected square different than the original pattern.

A pilot study was done to control for the relative difficulty of the spatial and verbal interference tasks. The number of squares in the spatial interference task (36) and the length of the number of digits in the verbal interference task (8) were selected such that participants performed equally well at the discrimination task for the two interference tasks (0.72 vs. 0.73, spatial vs. verbal interference, $t(21) < 1$, $p = 0.58$). Participants in the pilot study were not exposed to any categorization trials.

The complexity of the verbal and spatial tasks differed from those used by Winawer et al. (2007). They used a verbal string of length 8 and a 4 by 4 grid for their spatial pattern. Using a pretest they found no significant differences in accuracy on the interference judgment for those two tasks. In our pilot study, we found a significant difference on discrimination performance (with no categorization training) between their two conditions (0.76 vs. 0.71, spatial vs. verbal, $t(21) = 2.26$, $p = 0.03$).

Phase 1: Mixed Categorization and Discrimination Phase one consisted of two blocks of 120 categorization learning and discrimination trials presented without interference tasks. This allowed participants to begin learning the correct categories before introducing interference tasks. Trials were randomly mixed such that for each mini-block of 15 consecutive trials, 8 trials were categorization and 7 were discrimination, randomly ordered and intermixed. Note that participants did not know the type of judgment they would have to make (categorization or discrimination) until after the stimulus disappeared. This manipulation increases the relevance of processing category-level information during discrimination. The first block of phase one discrimination trials was used as a baseline measurement of performance before learning.

Phase 2: Interference Tasks with Mixed Categorization and Discrimination Phase two consisted of 21 mini-blocks composed of eight categorization and eight discrimination trials presented in a random order. Of the 21 blocks, seven had a verbal interference task, seven had a spatial interference task, and seven had no interference task. The order of mini-blocks was randomized across participants.

Results

For all analyses presented below, responses faster than 150 ms (less than 2% of all responses) were excluded from analysis. Including these fast trials in the analyses does not change the significance of the results.

Interference Task Performance In phase two participants demonstrated above chance performance on the spatial interference task ($M = 0.89$, $SD = 0.08$, $t(110) = 31.5$, $p < 0.001$) and the verbal interference task ($M = 0.95$, $SD = 0.13$, $t(110) = 59.4$, $p < 0.001$). A paired-sample t-test found a significant difference in performance between accuracy on the two test types ($t(110) = 4.51$, $p < 0.001$). Participants were more accurate on the verbal interference task.

Categorization Performance In phase two participants demonstrated above chance categorization performance in the no interference condition ($M = 0.83$, $SD = 0.12$, $t(110) = 28.3$, $p < 0.001$), the verbal interference condition ($M=0.82$, $SD = 0.12$, $t(110) = 28.9$, $p < 0.001$), and the spatial interference condition ($M = 0.82$, $SD = 0.12$, $t(110) = 28.2$, $p < 0.001$). There was not a significant difference in

categorization performance across interference conditions ($F(2,220) = 0.1$, $Mse = 0.0003$, $p = 0.9$).

Discrimination Performance Discrimination trials were classified based on the relationship of the target and foil face stimuli and the category boundary. Trials were classified as *within-cluster* if both faces were contained in the same group, and therefore within the same category as well. If the faces were in different clusters but still in the same category, those trials were classified as *within-category*. All remaining trials contained faces that were in different categories and were classified as *between-category*.

Discrimination performance during phase two, containing interference tasks (blocks 3 and 4), was assessed as a change in performance relative to a baseline performance on discrimination trials. The average for each participant of all discrimination trials in the first block (all possible 56 discrimination pairs) of phase one was used as this baseline measure. Removing baseline performance minimizes the variance due to any initial differences in discrimination ability across individuals. In all the following analyses this change in discrimination performance was used as the dependent measure.

Discrimination Performance across all Interference Tasks A repeated-measures ANOVA with discrimination type (3 levels) as a within-subject variable found a significant main effect of discrimination type on change in discrimination performance ($F(2, 220) = 14.83$, $Mse = 0.06$, $p < 0.001$). Planned comparisons between discrimination types found significant differences between *between-category* and *within-category* conditions ($t(110) = 2.63$, $p = 0.010$), between *within-category* and *within-cluster* conditions ($t(110) = 2.61$, $p = 0.010$), and between *between-category* and *within-cluster* conditions ($t(110) = 5.98$, $p < 0.001$).

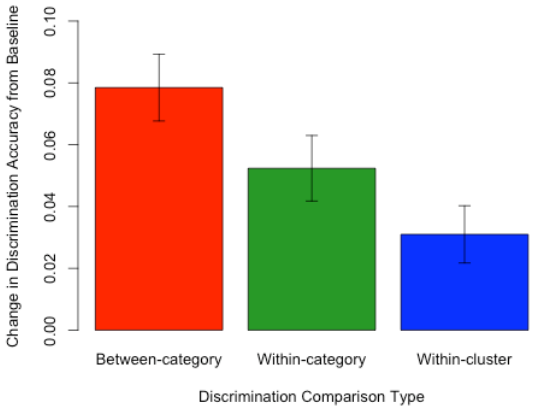


Figure 2: The change in discrimination performance relative to baseline averaged across interference condition. Participants show an increase in discrimination performance for all discrimination types relative to baseline (block 1), but a larger increase for judgments that cross category or cluster boundaries than are within-cluster. All error bars are standard errors.

Figure 2 shows these results support the predicted pattern of results and replicate the general pattern of results reported in Gureckis and Goldstone (2008). The learned CP effect was found: perceptual discriminations that span category boundaries showing the largest increase. Participants also learned the internal structure of categories, reflected in the significant difference between *within-category* and *within-cluster* perceptual discriminations, where discriminations that span within-category clusters had a larger increase. The main difference from Gureckis and Goldstone (2008) was that an increase in perceptual discrimination was found for all discrimination types (Gureckis and Goldstone (2008) found a non-significant decrease in the *within-cluster* condition).

Discrimination Performance within Interference Tasks A repeated-measures ANOVA with interference condition (3 levels: none, verbal, and spatial) and discrimination task (3 levels: as above) was performed with change in discrimination performance as the dependent measure. A main effect of discrimination type was found ($F(2,220) = 14.78$, $Mse = 0.19$, $p < 0.001$). Surprisingly, there was no main effect of interference task ($F(2,220) = 0.29$, $Mse = 0.003$, $p = 0.75$), nor a significant interaction between discrimination type and interference condition ($F(4,440) = 0.77$, $Mse = 0.008$, $p = 0.55$). Figure 3 shows this result.

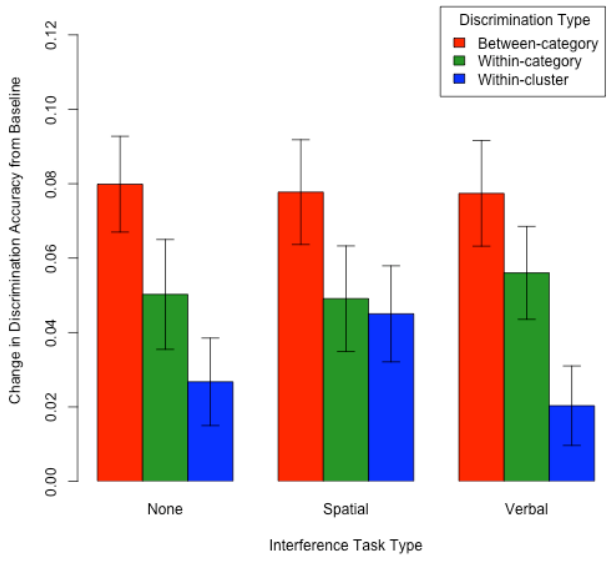


Figure 3: The effect of discrimination type and interference condition on change in discrimination performance relative to baseline (block 1). Participants show a consistent pattern in which *Between-category* improvement is greater than *Within-category* improvement, which is greater than *Within-cluster* improvement. There is no effect of interference condition or an interaction with discrimination type. All error bars are standard errors.

Within each interference condition the same pattern of results hold as across all conditions. *Between-category* discriminations increase more relative to baseline than *within-category*, which increases more than *within-cluster*. The difference in improvement between *between-category* and *within-category* is marginally significant for the no interference ($t(110) = 1.80$, $p = 0.075$) and the spatial interference conditions ($t(110) = 1.79$, $p = 0.077$), and not significant for the verbal interference condition ($t(110) = 1.34$, $p = 0.18$). The difference between *within-category* and *within-cluster* improvement is significant in the verbal interference condition ($t(110) = 2.50$, $p = 0.01$), marginally significant for the no interference condition ($t(110) = 1.81$, $p = 0.073$), and not significant in the spatial interference condition ($t(110) < 1$, $p = 0.75$). The difference in improvement between *between-category* and *within-cluster* is significant for all interference conditions (none ($t(110) = 3.75$, $p < 0.001$), spatial ($t(110) = 4.47$, $p < 0.001$), and verbal ($t(110) = 4.47$, $p < 0.001$).

Discrimination Performance grouped by Categorization Performance Following Gureckis and Goldstone (2008), an analysis was performed on the effect of discrimination task on discrimination performance within mini-blocks as a function of the accuracy of categorization trials within that mini-block. For each participant, mini-blocks selected from trials in phase two were grouped based on categorization accuracy within the mini-block into high categorization (75-100%, 322 mini-blocks among 107 subjects), medium categorization (50-75%, 312 mini-blocks among 79 subjects), and low categorization (0-50%, 124 mini-blocks among 22 subjects). Figure 4 shows these results.

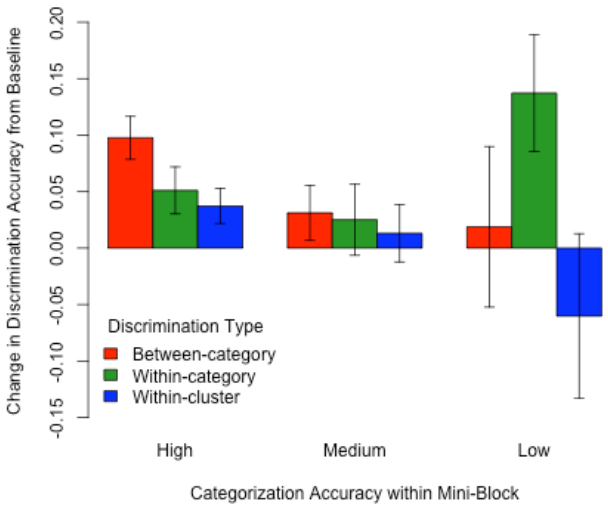


Figure 4: All bars are standard error bars but not all conditions had the same number of participants. Participants who did not have any low categorization accuracy mini-blocks did not contribute to the number of participants in the low categorization conditions. The small number of observations in the *within-category* low accuracy condition may have contributed to what appears to be a spuriously high increase in that condition.

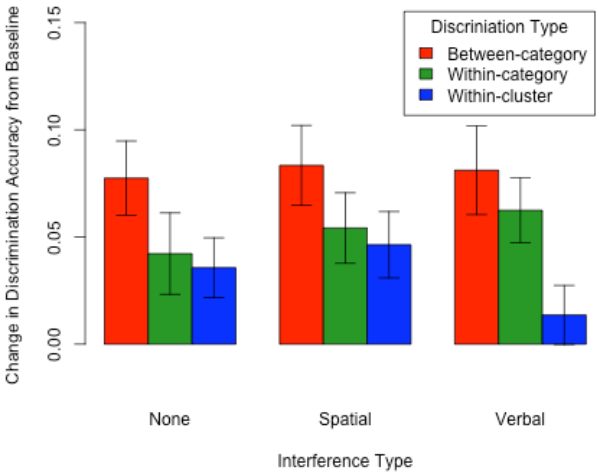


Figure 5: The effects of discrimination type and interference condition on change in discrimination performance relative to baseline for mini-blocks in which categorization accuracy was above 75%. Participants show a consistent pattern in which *Between-category* improvement is greater than *Within-category* improvement, which is greater than *Within-cluster* improvement. There is no effect of interference condition or an interaction with discrimination type. All error bars are standard errors.

The pattern of results in the high categorization accuracy set follows those of Gureckis and Goldstone (2008). The pattern among low categorization performance mini-blocks may be an artifact of having few participants at that level.

Looking specifically in the high categorization performance group (figure 5) where CP effects were predicted to be strongest and thus easiest to see an influence of interference condition, a repeated-measures ANOVA was performed with interference condition (3 levels) and discrimination type (3 levels) as within-subject factors. There was a significant main effect of discrimination type ($F(2, 214) = 8.56$, $Mse = 0.19$, $p < 0.001$) but not of interference condition ($F(2,214) = 0.4$, $Mse = 0.009$, $p = 0.66$) and no significant interaction between the two factors ($F(4,424) = 0.68$, $Mse = 0.02$, $p = 0.60$).

The high-categorization mini-block results echo our previous results (Figure 3) showing a strong effect of discrimination type but no influence or interaction with interference condition.

Discussion

Consistent with Gureckis and Goldstone (2008), we found strong evidence for learned categorical perception across the category boundary as well as learned sensitivity to the structure of information within the categories. This learning effect was strongest when averaged across all interference conditions, but the same pattern was exhibited in each interference condition: *between-category* discriminations improved the most, followed by *within-category* discriminations, and *within-cluster* discriminations

improved the least. This pattern was found within each interference condition with varying degrees of reliability. As predicted, it was also consistently found in mini-blocks that had high accuracy on categorization trials, more so than in blocks with low categorization accuracy.

Surprisingly, we did not find any indication that the interference tasks modulated the effects of increased discrimination. No main effect of interference condition was found in any discrimination types, across either category boundaries or within-category structures. This pattern was also found in mini-blocks with high categorization accuracy and in all discrimination types. The lack of interaction between interference condition and discrimination type is less startling than the lack of main effect of interference condition on overall discrimination performance because the difficulty of the spatial and verbal interference tasks was selected based on pilot data to have a relatively equal effect on perceptual discrimination tasks. The lack of main effect of interference condition is consistent with the results of Russian speakers in the Winawer et al. (2007) study (who only found an interaction between interference condition and the CP effect), though their interference tasks were pretested to equate for accuracy on the interference task itself. Winawer et al. also did not find an interaction between categorical perception and interference condition among the English speakers who did not show a main effect of categorical perception. This is not consistent with learners in our task who did show categorical perception, as well as sensitivity to inter-category structure, but did not show an interaction with interference condition.

This current work suggests that firmly entrenched verbal labels, such as color names (Winawer et al., 2007) or basic shapes (Lupyan, 2009), may be necessary to see verbal interference effects in perceptual discrimination. The incidentally learned information about the structure of categories that underlies the results found in Gureckis and Goldstone (2008) and replicated here may not have verbal labels attached that are influenced by an interference task. Instead, the preservation of this pattern across interference conditions is consistent with the non-verbally mediated account of CP that directs the focus of learning toward learning to weight perceptual dimensions rather than rely on verbal labels for categories. Clearly, the lack of effect of interference task does not justify strong claims about the nature of learned CP effects. However it does suggest that for non-automated categories verbal labels might not tell the whole story about what learning drives CP. Further work is needed to bridge the gap between our understanding of entrenched categories that do show verbal interference effects and newly-learned categories that might not, and how representations may change to incorporate more information about verbal labels.

References

Goldstone, R. (1994). Influence of categorization on perceptual discrimination. *Journal of Experimental Psychology: General, 123(2),* 178-200.

Goldstone, R. L., & Hendrickson, A. T. (2009). Categorical Perception. *Interdisciplinary Reviews: Cognitive Science*, *1,* 69-78.

Gureckis, T. M., & Goldstone, R. L. (2008). The effect of internal structure of categories on perception. *Proceedings of the Thirtieth Annual Conference of the Cognitive Science Society*, (pp. 1876-1881). Washington, D.C.: Cognitive Science Society.

Harnad, S. (Ed.). (1987). *Categorical perception: The groundwork of cognition.* New York: Cambridge University Press.

Harnad, S., Hanson, S., & Lubin, J. (1995). Learned categorical perception in neural nets: Implications for symbol grounding. *In V. Honavar & L. Uhr (Eds.), Symbol processors and connectionist network models in artificial intelligence and cognitive modeling: Steps toward principled integration* (p. 191-206). Boston: Academic Press.

Kayser, A. (1997). *Heads*. New York: Abbeville Press.

Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology, 54,* 358-368.

Logan, J., Lively, S., & Pisoni, D. (1991). Training Japanese listeners to identify English /r/ and /l/: A first report. *Journal of the Acoustical Society of America, 89,* 874-886.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A Network Model of Category Learning. *Psychological Review*, 111(2), 309-332.

Lupyan, G. (2008). The conceptual grouping effect: Categories matter (and named categories matter more). *Cognition*, 108, 566-577.

Lupyan, G. (2009). Extracommunicative functions of language: Verbal interference causes selective categorization impairments. *Psychonomic Bulletin & Review*, *16*, 711-718.

Steyvers, M. (1999). Morphing techniques for generating and manipulating face images. *Behavior Research Methods, Instruments, & Computers*, *31,* 359-369.

Winawer, J., Witthoft, N., Frank, M. C., Wu, L., Wade, A. R., & Boroditsky, L. (2007). Russian blues reveal effects of language on color discrimination. *Proceedings of the National Academy of Sciences, 104*, 7780-7785.

Visual Similarity Effects in Categorical Search

Robert G. Alexander[1] (rgalexander@notes.cc.sunysb.edu), Wei Zhang (weiz@microsoft.com)[2,3]
Gregory J. Zelinsky[1,2] (Gregory.Zelinsky@stonybrook.edu)
[1]Department of Psychology, Stony Brook University
[2]Department of Computer Science, Stony Brook University
[3]Microsoft Corporation

Abstract

The factors affecting search guidance to categorical targets are largely unknown. We asked how visual similarity relationships between random-category distractors and two target classes, teddy bears and butterflies, affects search guidance. Experiment 1 used a web-based task to collect visual similarity rankings between these target classes and random objects, from which we created search displays having either high-similarity distractors, low-similarity distractors, or "mixed" displays with high, medium, and low-similarity distractors. Subjects made faster manual responses and fixated fewer distractors on low-similarity displays compared to high. On mixed trials, first fixations were more frequent on high-similarity distractors (bear=49%; butterfly=58%) than low-similarity distractors (bear=9%; butterfly=12%). Experiment 2 used the same high/low/mixed conditions, but now these conditions were created using similarity estimates from a computer-vision model that ranked objects in terms of color, texture, and shape similarity. The same patterns were found, suggesting that categorical search is indeed guided by visual similarity.

Keywords: Visual search; eye movements; categorical guidance; visual similarity; object class detection

Introduction

You have probably had the experience of searching for your car in a parking lot and finding several other vehicles of the same color or model before finally finding your car. This is an example of visual similarity affecting search; the presence of these target-similar distractors made it harder to find the actual target of your search.

Such visual similarity effects have been extensively studied in the context of search, with the main finding from this effort being that search is slower when distractors are similar to the target (e.g., Duncan & Humphreys, 1989; Treisman, 1991). Models of search have also relied extensively on these visual similarity relationships (e.g., Pomplun, 2006; Treisman & Sato, 1990; Wolfe, 1994; Zelinsky, 2008). Despite their many differences, all of these models posit a very similar process for how similarity relationships are computed and used; the target and scene are represented by visual features (color, orientation, etc.), which are compared to generate a signal used to guide search to the target and to target-like distractors in a display. In general, the more similar an object is to the target, the more likely that object will be fixated.

All of these models, however, assume knowledge of the target's specific appearance in the creation of this guidance signal. This assumption is problematic, as it is often violated in the real world. Descriptions of search targets are often incomplete and lacking in visual detail; exact knowledge of a target's appearance is an artificial situation that typically exists only in the laboratory. Particularly interesting are cases in which a target is defined categorically, as from a text label or an instruction (i.e., no picture preview of the target). Given the high degree of variability inherent in most categories of common objects, search under these conditions would have few visual features of the target that could be confidently compared to a scene to generate a guidance signal. Indeed, a debate exists over whether categorical search is guided at all, with some labs finding that it is (Schmidt & Zelinsky, 2009; Yang & Zelinsky, 2009) and others suggesting that it is not (e.g., Castelhano et al., 2008; Wolfe et al., 2004).

The present study enters this debate on the existence of categorical guidance, focusing it on the relationship between target-distractor visual similarity and guidance to categorically-defined realistic targets. Guidance from a pictorial preview is known to decrease with increasing visual similarity between a target and distractors; does this same relationship hold for categorically-defined targets? Given that the representation of categorical targets is largely unknown, it may be the case that target descriptions are dominated by non-visual features, such as semantic or functional properties of the target category. If this is the case, guidance to the target may be weak or even nonexistent, potentially explaining the discrepant findings. To the extent that categorical search does use non-visual features, effects of target-distractor visual similarity would therefore not be expected. However, if target categories are represented visually, one might expect the same target-distractor similarity relationships demonstrated for target-specific search to extend to categorical search.

It is unclear how best to manipulate visual similarity in the context of categorical search. Traditional methods of manipulating target-distractor similarity by varying only a single target feature are clearly suboptimal, as realistic objects are composed of many features and it is impossible to know *a priori* which are the most important. This problem is compounded by the categorical nature of the task; the relevance of a particular target feature would almost certainly depend on the specific category of distractor to which it is compared. It is not even known how best to derive specific target features for such a comparison; should an average be obtained from many target exemplars or should features be extracted from a particular exemplar that is representative of the target class?

In light of the difficulties associated with directly manipulating the specific features underlying visual

similarity, we opted for a more holistic approach—to use ratings of visual similarity obtained from subjects. Specifically, we obtained ratings from Zhang et al. (2008), who used a web experiment to collect visual similarity estimates between random objects and categorical targets for the purpose of comparing these estimates to the behavior of a computational model of object class detection. Subjects were randomly assigned to either a butterfly target class or a teddy bear target class, and their task was to rate real-world objects (from the Hemera collection) to these target categories. They did this by rank-ordering groups of five objects; each trial showed five random objects, and the subjects' task was to give each a 1-5 ranking, where "1" indicated low target similarity and 5 indicated high target similarity (objects given the 2-4 rankings and objects with low inter-subject ranking agreement were considered medium similarity). There were 142 subjects, yielding a total of 71,000 butterfly and teddy bear similarity estimates for 2,000 different objects. Importantly, subjects were instructed to use only visual similarity and to disregard categorical or associative relationships between the objects and the target category when making their judgments. Consult Zhang et al. (2008) for additional details regarding this web-based collection of visual similarity estimates.

Using these estimates of visual similarity, Experiment 1 asked whether the visual similarity relationships known to affect search for specific targets also extends to categorical search. Previous arguments for the existence of categorical search guidance relied on evidence showing the preferential direction of initial saccades to targets (Schmidt & Zelinsky, 2009; Yang & Zelinsky, 2009). Although there is good reason to believe that these initial saccades are dominated by visual features, and occur too early in search to be influenced by semantic relationships between targets and distractors, it is still possible that the preferential fixation of categorical targets might have been influenced by non-visual factors. More compelling would be a demonstrated relationship between categorical guidance and a manipulation of target-distractor visual similarity; providing this evidence was the primary goal of this experiment.

We were also interested in determining whether explicit visual similarity judgments are predictive of effects of target-distractor visual similarity on categorical search. Search guidance is a largely implicit process, and as discussed can be expressed in even the first search saccade (Chen & Zelinsky, 2006); the task of assigning rankings to objects in a web experiment is comparatively slow and far more explicit. Do these two tasks use fundamentally different sources of information, or can visual similarity estimates obtained from explicit judgments be useful in describing guidance during search? Answering this question was a secondary goal of this experiment.

If categorical search is guided by target-distractor visual similarity, and if this relationship can be captured by explicit similarity judgments, we would expect a relatively high proportion of initial saccades to high-similarity distractors, and relatively few initial saccades to low-similarity distractors. However, if categorical guidance is mediated by non-visual factors, or if the visual similarity estimates obtained from an explicit task cannot be extended to search, we would expect no effect of our similarity manipulations on guidance or manual search efficiency.

Experiment 1

Method

Participants Twenty-four students from Stony Brook University participated in exchange for course credit. All subjects reported normal or corrected to normal vision.

Stimuli and Apparatus Targets and distractors were selected from the objects used by Zhang et al. (2008). The target categories were teddy bears, obtained from Cockrill (2001), and butterflies, obtained from the Hemera collection. The distractors were also Hemera objects. Each object was sized to subtend ~2.8º of visual angle.

Gaze position was recorded using an SR Research EyeLink® II eye tracking system. This eye tracker is video-based and has a sampling rate of 500 Hz and a spatial resolution of ~0.2º. Target present/absent search decisions were made using a GamePad controller connected to a USB port. Head position and viewing distance were fixed at 72 cm from the screen with a chin rest. Trials were displayed on a flat-screen monitor at a resolution of 1024 × 768 pixels (subtending 28º × 21º) and a refresh rate of 85 Hz.

Design and procedure Half of the subjects searched for a teddy bear target, the other half searched for a butterfly target. This search was categorical; subjects were not shown a specific bear or butterfly target preview prior to each search trial. Rather, subjects were told the target category at the start of the experiment. They were also shown examples of the target category, none of which were used as actual targets in the experimental trials.

Each trial began with the subject fixating a central dot and pressing a button on the controller to initiate the search display. The search display consisted of six evenly-spaced objects arranged on an imaginary circle with a radius of 300 pixels (8.4º) relative to the center of the screen. On target present trials (50%), one object was either a bear or a butterfly, depending on the condition, and the other five objects were randomly selected distractors. On target absent trials (50%), distractors were selected based on the similarity rankings from the Zhang et al. (2008) web task.

There were three target absent conditions: high-similarity trials (all distractors were similar to the target category), low-similarity trials (all distractors were dissimilar to the target category), and "mixed" trials, where two distracters were selected from the high-similarity category, two from the low-similarity category, and two from the medium similarity category (see Figure 1). The high and low similarity conditions were included to determine whether visual similarity affects search accuracy and manual reaction times (RTs). The mixed condition allowed us to

directly examine which distracters were preferentially fixated (i.e., search guidance) as a function of target-distractor similarity.

Target presence/absence and similarity condition were within-subjects variables, and both were randomly interleaved throughout the experiment. Subjects were asked to make their present/absent judgments as quickly as possible while maintaining accuracy. Accuracy feedback was provided following each response.

Results and Discussion

As only the target absent trials contained the similarity manipulation, analyses were restricted to these data.

Errors were less than 6% in all conditions, and were excluded from all subsequent analyses. This low false positive rate means that subjects were not confusing the high-similarity distractors for targets (e.g., a stuffed bunny distractor was not recognized as a teddy bear).

RTs were longest in the high-similarity condition and shortest in the low-similarity condition, with the mixed condition yielding intermediate RTs (Table 1). These differences were significant for both butterfly targets ($F(2,22) = 46.87$, $p < .001$) and for bear targets ($F(2,22) = 53.85$, $p < .001$). The number of distractors fixated during search also differed between the similarity conditions, and this again occurred for both butterfly ($F(2,22) = 30.41$, $p < .001$) and bear targets ($F(2,22) = 59.55$, $p < .001$). Distractors were fixated most frequently on the high-similarity trials (3.16±0.23 for bears; 2.50±0.36 for butterflies), followed by the medium-similarity trials (2.53±0.24 for bears; 1.83±0.31 for butterflies), and finally the low-similarity trials (1.51±0.23 for bears; 1.29±0.26 for butterflies); as distractor similarity to the target increased, so did the number of fixations on these distractors. All of these patterns are consistent with the suggestion that visual similarity rankings are predictive of search efficiency.

One of the most conservative measures of search guidance is the first fixated object—the object looked at first following search display onset. Consistent with the RT analyses we found that distractor similarity to the target determined which objects were fixated first on mixed condition trials (Figure 2A). High-similarity distractors were more often fixated first compared to medium-similarity distractors, which were more often fixated first compared to low-similarity distractors, and this pattern was found for both butterflies ($F(2,22) = 10.13$, $p < .01$) and for bears ($F(2,22) = 30.15$, $p < .001$).

Two conclusions follow from our data. First, categorical search guidance is affected by target-distractor visual similarity. As the visual similarity between a distractor and a target category increases, search efficiency decreases. This decreased efficiency is due to distractors becoming more distracting, as evidenced by an increase in the number of first fixations on the high similarity distractors. More generally, this finding adds to the growing body of evidence suggesting that categorical search is indeed guided (Schmidt & Zelinsky, 2009; Yang & Zelinsky, 2009), a question that had been the topic of debate (Castelhano et al., 2008, and Wolfe et al., 2004). Not only is categorical search guided, it is guided by matching visual features to a visual representation of the target category.



Figure 1: Objects from a typical mixed trial. (A) low-similarity, (B) medium-similarity, and (C) high-similarity distractors, as ranked to the teddy bear target category.

The second conclusion following from our data is that explicit visual similarity rankings from a web task are highly predictive of categorical search. Given the dramatic differences between these tasks, this finding is surprising. Judgments in the web task were highly deliberative. In piloting, a subject was observed agonizing over whether a wooden box or a backpack was visually more similar to a teddy bear. These highly explicit similarity judgments can be contrasted with the largely implicit visual similarity computations that drove search guidance. Whereas the web-based judgments could be measured in seconds, effects of similarity on search guidance appeared almost immediately, at least within the first 199 ms following search display onset (the average latency of initial saccades in this experiment). Our data suggest a common thread between these two processes. Regardless of whether a visual similarity relationship has to be completed in time for an initial eye movement, or the opportunity exists to deliberate on this relationship for an extended period, the same features seem to be represented and compared.

Table 1: Manual RTs by similarity condition, in seconds

	Experiment 1		Experiment 2	
	Butterfly	Bear	Butterfly	Bear
High	1.17 (.06)	1.48 (.14)	1.59 (.13)	1.24 (.15)
Medium	0.97 (.06)	1.15 (.11)	1.25 (.10)	1.07 (.15)
Low	0.82 (.05)	0.84 (.08)	0.92 (.09)	0.74 (.09)

Note. Values in parentheses indicate one standard error.

Experiment 2

Were subjects from Experiment 1 confining their similarity judgments to purely visual dimensions? The fact that this was the instructed task does not guarantee that non-visual factors were not creeping into the similarity judgments, raising the possibility that these factors, and not visual similarity, were responsible for the observed categorical guidance. Experiment 2 addressed this possibility.

It is unclear how best to separate visual from non-visual factors in estimates of similarity. Even when stimuli are oriented bars with no compelling semantic properties, semantic features might still influence perceptual decisions (Wolfe et al., 1992). The task of separating these factors

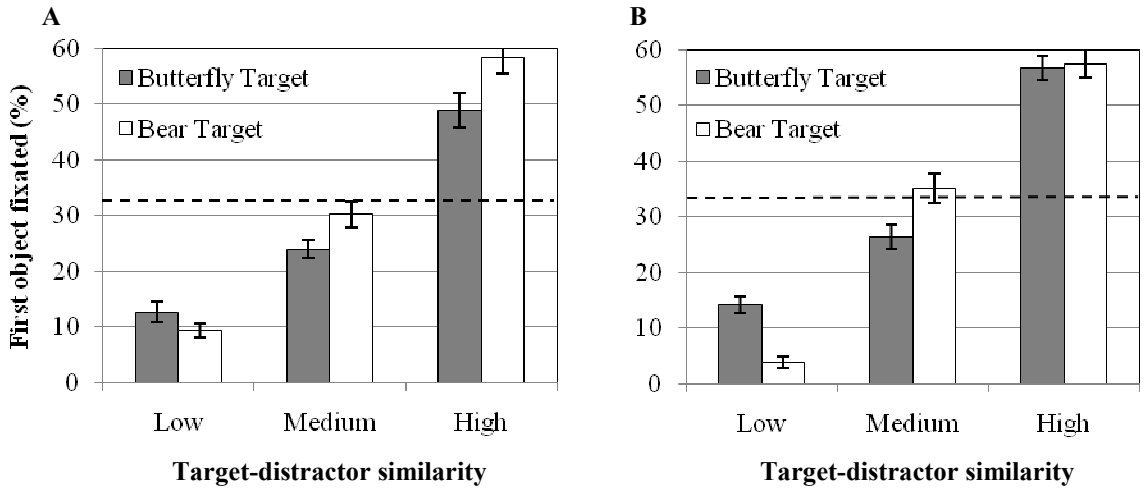


Figure 2: Percentage of mixed condition trials in which the first object fixated had a low, medium, or high target-distractor similarity for (A) Experiment 1 and (B) Experiment 2. Error bars show one standard error. Dashed lines indicate chance.

using purely behavioral methods is even more daunting in the present study, as our stimuli are realistic objects having an untold number of visual and semantic dimensions.

In Experiment 2 we take a different approach to this problem—turning to the computer vision literature to obtain similarity estimates. Recent years have seen considerable success in the development of automated methods for the detection of object categories in realistic scenes, a task with obvious relevance to categorical visual search. At the core of these methods is the computation of visual similarity relationships between visual images and features extracted from a target class. These similarity relationships are potentially useful for our current purpose, as they provide estimates of purely visual similarity between distractors and a categorically-defined target, free from any contamination by semantic properties. Whereas the similarity estimates used in Experiment 1 may have been based on some mix of visual and non-visual information, the similarity estimates obtained from a computer vision method are incontrovertibly exclusively visual.

To obtain these purely visual similarity estimates we used the computer vision method described in Zhang et al. (2008). We chose this method for two reasons. First, it works by having multiple visual features contribute flexibly to target classification (see also Zhang et al., 2005). Specifically, it combines state-of-the-art color histogram features (Swain & Ballard, 1991), texture features (the Scale Invariant Feature Transform, or SIFT; Lowe, 2004), and global shape context features (Belongie et al., 2002) with a well-studied machine learning technique (AdaBoost; Freund & Schapire, 1997) to create classifiers having features tailored for the detection of specific target categories. The advantage of this method over other automated object classification techniques is that similarity estimates can be based on the contribution of multiple features, not just one.

Our second reason for choosing the Zhang et al. (2008) model is that it has already been successfully applied to the identical target and distractor objects used in the present study. Specifically, it successfully classified the high-similarity and low-similarity objects from the above-described web task, regardless of whether the target category was a teddy bear or a butterfly. This makes the Zhang et al. model an obvious choice for our goal of collecting computer-vision-based similarity estimates; not only was this model able to learn classifiers to discriminate our target categories from random objects, these classifiers were also shown to be partially successful in capturing human visual similarity relationships between these random objects and the bear and butterfly target classes.[1]

To the extent that the Zhang et al. model is successful in capturing human visual similarity relationships, and to the extent that these similarity estimates extend to a search task (as we found in the previous experiment), then displays constructed of high-similarity or low-similarity distractors, as rated by the model, should produce the same patterns of guidance found in Experiment 1. Initial saccades should be preferentially guided to high-similarity distractors, and preferentially guided away from low-similarity distractors, with guidance to medium similarity distractors falling between these two levels. Replicating these patterns in the context of new search displays, assembled using the purely visual similarity estimates from a computer vision model, would offer converging evidence for our claim that visual similarity affects categorical search. Of course failing to replicate these patterns would weaken this claim, and would raise concerns that the evidence for guidance reported in

[1] Note that this agreement to human behavior does not mean that the features and learning method used by this model accurately describes how humans arrive at their visual similarity estimates. Making this correspondence is a goal to which we aspire, but one that we believe is still out of reach. However, this modest level of agreement does suggest that the model's multi-feature approach has the potential to generate visual similarity estimates having behavioral significance, which makes it relatively unique with respect to purely automated computational approaches.

Experiment 1 might have been due to semantic, associative, or other non-visual sources of information.

Method

Participants Twenty-four Stony Brook University students participated in exchange for course credit, none of whom participated in Experiment 1. All subjects reported normal or corrected to normal vision. Half searched for a teddy bear target, the other half searched for a butterfly target.

Stimuli and Apparatus Experiment 2 was conducted using the same equipment as in Experiment 1. The stimuli were also objects selected from the same image set, although the new selection criteria (described below) required the potential placement of these objects into different conditions. The search displays were therefore different, but were assembled from the same set of objects.

Design and procedure Experiments 1 and 2 had the same conditions and followed the same procedure, with the only difference being the distractor composition of target absent trials; distractors were now selected based on visual similarity estimates obtained from the Zhang et al. (2008) model rather than from similarity rankings from the web task. To derive these similarity estimates we again trained an AdaBoost-based classifier for each target class using color, shape, and texture features, then evaluated these same features for the distractors to compute target-distractor similarity. This resulted in the creation of two rank ordered lists, one indicating visual similarity to teddy bears and the other to butterflies. High-similarity trials for each target category were then constructed from distractors ranked in the top third of each list, and low-similarity trials were constructed from distractors ranked in the bottom third.
Mixed trials consisted of high-similarity distractors from the top third, low-similarity distractors from the bottom third and medium-similarity distractors from the middle third.

Results and Discussion

Errors were less than 3% in all conditions and were again excluded from subsequent analyses. These infrequent errors were likely just motor confusions rather than cases of confusing teddy bears or butterflies with random objects.

If categorical search is affected by the visual similarity between our target categories and random distractors, and if the Zhang et al. (2008) model is able to capture these relationships, then RTs should be the slowest on high-similarity trials, faster on mixed trials, and the fastest on low-similarity trials. These predictions were confirmed (Table 1). Search efficiency varied with target-distractor visual similarity for both teddy bears ($F(2,22) = 35.84$, $p < .001$) and butterflies ($F(2,22) = 60.95$, $p < .001$); post-hoc t-tests with Bonferroni correction showed slower RTs in the high-similarity condition relative to the mixed condition ($t(11) = 5.77$, $p < .01$ for teddy bears and $t(11) = 6.50$, $p < .01$ for butterflies) and faster RTs in the low-similarity condition relative to the mixed condition ($t(11) = 5.15$, $p < .01$ for teddy bears and $t(11) = 6.22$, $p < .01$ for butterflies).

Analysis of the number of distractors fixated during search revealed the same patterns. Fixated distractors varied with visual similarity for both butterfly targets ($F(2,22) = 74.55$, $p < .001$) and bear targets ($F(2,22) = 93.55$, $p < .001$). More distractors were fixated on high-similarity trials (2.42±0.20 for bears; 3.66±0.24 for butterflies) compared to either mixed trials (2.10±0.17 for bears; 2.88±0.23 for butterflies) or low-similarity trials (1.01±0.19 for bears; 1.94±0.24 for butterflies).

The availability of high-, medium-, and low-similarity distractors in mixed condition displays again enabled us to look for direct oculomotor evidence for categorical search guidance. Analyses of these trials showed a relationship between visual similarity and the probability of first fixation on an object ($F(2,22) = 19.42$, $p < .001$ for butterflies; $F(2,22) = 36.60$, $p < .001$ for bears – see Figure2B). Moreover, first fixations on high-similarity distractors were well above chance ($t(11) = 5.89$, $p < .01$ for bears; $t(11) = 10.01$, $p < .01$ for butterflies), and first fixations on low-similarity distractors were well below chance ($t(11) = 25.47$, $p < .01$ for bears; $t(11) = 8.32$ for butterflies), indicating that initial saccades were guided towards target-similar distractors and away from target-dissimilar distractors.

We also analyzed initial saccade latencies to see whether these patterns could be attributed to speed-accuracy tradeoffs, but none were found; initial saccade latencies did not reliably differ between the similarity conditions for either butterfly ($F(2,22) = 1.29$, $p = 0.30$) or bear targets ($F(2,22) = 0.76$, $p = 0.48$). The observed effects of visual similarity reflect actual changes in search guidance.

The conclusion from this experiment is clear. While the results of Experiment 1 could have been confounded by the unintentional inclusion of non-visual features in the behavioral similarity rankings, the same cannot be said for the similarity estimates used in Experiment 2. Even when estimates reflected purely visual features, target-distractor similarity still predicted categorical search performance. This strongly suggests that categorical guidance not only exists, but that it may operate in much the same way as search guidance from a pictorial target preview. The visual features used to represent a categorical target may be different and come from a different source (learned and recalled from memory rather than extracted from a target preview), but the underlying process of comparing these visual features to the search scene and using this signal to guide search may be the same. A goal of future work will be to determine what these categorical features are for a variety of real-world target classes. The present work constrains this goal by requiring that these features capture target-distractor visual similarity relationships.

Conclusions

Previous research had suggested that search is unguided to categorical targets (e.g., Castelhano et al., 2008; Wolfe et al., 2004). In light of recent evidence, this suggestion

should be revisited. Multiple studies have now shown guidance in the very first saccades made to categorical targets (Schmidt & Zelinsky, 2009; Yang & Zelinsky, 2009). The present work extends this finding to non-target objects from categories that are visually similar to the target class. Specifically, in the absence of a target our subjects preferentially directed their initial saccades *to* distractors that were target-similar, and *away* from distractors that were target-dissimilar (mixed condition; Figure 2). This pattern, when combined with the patterns of manual search efficiency found in the high-similarity and low-similarity distractor conditions (Table 1), provides strong converging evidence for categorical search guidance in our task. The fact that these results were obtained despite the highly non-obvious similarity relationships between random objects and teddy bears / butterflies, makes the clear expression of guidance reported here all the more striking.

We can also conclude that these effects of similarity on categorical search guidance are visual, and can be well described by explicit similarity estimates regardless of whether these estimates were obtained from behavioral rankings using a web task (Experiment 1) or generated by a computer vision model of object category detection (Experiment 2). This too is a striking finding. The lengthy deliberations that accompanied the behavioral judgments, and certainly the simplistic visual features underlying the model's estimates, might have easily resulted in no success whatsoever in predicting categorical search behavior. The fact that these radically different methods both successfully predicted patterns of search guidance is informative, suggesting that the computation of visual similarity is not only a core cognitive operation, but one that is remarkably stable across method. We speculate that visual similarity is computed early and automatically during perception, and once derived is used to mediate a variety of perceptual (e.g., search guidance) and cognitive (similarity judgments) behaviors. To the extent that this is true, it bodes well for the diversity of researchers in cognitive psychology, human-computer interaction, and vision science, all attempting to better understand human visual similarity relationships.

Acknowledgments

We thank Ryan Moore, Jonathan Ryan, and Arunesh Mittal for their help with data collection. This work was supported by NIH grant R01MH063748-07 to GJZ.

References

Belongie, S., Malik, J., & Puzicha, J. (2002, April). Shape matching and object recognition using shape contexts. *Pattern Analysis and Machine Intelligence, 24*(4), 509-522.

Castelhano, M. S., Pollatsek, A., & Cave, K.R. (2008). Typicality aids search for an unspecified target, but only in identification and not in attentional guidance. *Psychonomic Bulletin & Review, 15*(4), 795-801.

Chen, X., & Zelinsky, G. J. (2006). Real-world visual search is dominated by top-down guidance. *Vision Research, 46*, 4118-4133.

Cockrill, P. (2001). *The teddy bear encyclopedia.* New York: DK Publishing Inc.

Duncan, J., & Humphreys, G. (1989). Visual search and stimulus similarity. *Psychological Review, 96* (3), 433-458.

Freund, Y., & Schapire, R. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences, 55*(1), 119-139.

Lowe, D. (2004, November). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision, 60*(2), 91-110.

Pomplun, M. (2006). Saccadic selectivity in complex visual search displays. *Vision Research*, *46*, 1886-1900.

Schmidt, J., & Zelinsky, G. J. (2009). Search guidance is proportional to the categorical specificity of a target cue. *Quarterly Journal of Experimental Psychology*, *62 (10)*, 1904-1914.

Swain, M., & Ballard, D. (1991, November). Color indexing. *International Journal of Computer Vision, 7*(1), 11-32.

Treisman, A. M. (1991). Search, similarity, and integration of features between and within dimensions. *Journal of Experimental Psychology: Human Perception and Performance. 17* (3), 652-676.

Treisman, A. M., & Sato, S. (1990). Conjunction search revisited. *Journal of Experimental Psychology: Human Perception and Performance*, *16*, 459-478.

Wolfe, J. M. (1994). Guided search 2.0: A revised model of visual search. *Psychonomic Bulletin & Review 1*(2), 202-238.

Wolfe, J. M., Friedman-Hill, S., Stewart, M., & O'Connell, K. (1992). The role of categorization in visual search for orientation. *Journal of Experimental Psychology: Human Perception and Performance*, *18*, 34-49.

Wolfe, J. M., Horowitz, T. S., Kenner, N., Hyle, M., & Vasan, N. (2004). How fast can you change your mind? The speed of top-down guidance in visual search. *Vision Research,* 44, 1411-1426.

Yang, H., & Zelinsky, G. J. (2009). Visual search is guided to categorically-defined targets. *Vision Research*, *49*, 2095-2103.

Zelinsky, G. J. (2008). A theory of eye movements during target acquisition. *Psychological Review 115*(4), 787-835.

Zhang, W., Samaras, D., & Zelinsky, G. J. (2008). Classifying objects based on their visual similarity to target categories. *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 1856-1861).

Zhang, W., Yu, B., Zelinsky, G. J., & Samaras, D. (2005). Object class recognition using multiple layer boosting with heterogeneous features. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, *2*, 323-330.

Social Cues Support Learning about Objects from Statistics in Infancy

Rachel Wu (r.wu@bbk.ac.uk)
Centre for Brain and Cognitive Development, Department of Psychological Sciences, Birkbeck, University of London
Malet Street, London, WC1E 7HX, UK

Alison Gopnik (gopnik@berkeley.edu)
Institute of Human Development and Department of Psychology, University of California at Berkeley
Berkeley, CA 94720, USA

Daniel C. Richardson (dcr@eyethink.org)
Department of Cognitive, Perceptual and Brain sciences, University College London
Gower Street, London, WC1E 6BT, UK

Natasha Z. Kirkham (n.kirkham@bbk.ac.uk)
Centre for Brain and Cognitive Development, Department of Psychological Sciences, Birkbeck, University of London
Malet Street, London, WC1E 7HX, UK

Abstract

In laboratory experiments, infants can learn patterns of features that co-occur (e.g., Fiser & Aslin, 2002). This finding leaves two questions unanswered: What do infants do with the knowledge acquired from such statistical learning, and which patterns do infants attend to in the noisy and cluttered world outside of the laboratory? Here, we show that 9-month-old infants form expectations about co-occurring features remaining fused, an essential skill for object individuation and recognition (e.g., Goldstone, 2000; Schyns & Rodet, 1997). We also show that though social cues may temporarily detract attention away from learning events, they appear to stimulate infants to display learning better in complex situations than when infants learn on their own without attention cues. These findings suggest that infants can use feature co-occurrence to learn about objects and that social cues shape such foundational learning in a noisy environment during infancy.

Keywords: visual statistical learning; eye-tracking; cognitive development; social cues; eye gaze.

Introduction

Knowing *what* to learn is fundamental to all aspects of development. In particular, recognizing important features in a display and the relationships between them supports essential skills such as object recognition (Biederman, 1987), categorization (Mareschal, Quinn, & French, 2002; Rakison & Butterworth, 1998; Schyns & Rodet, 1997; Younger & Cohen, 1986), and word learning (Smith & Yu, 2008). Fiser and Aslin (2002) showed that 9-month-olds prefer to look at shapes that have co-occurred previously, rather than at shapes that did not co-occur. These findings suggest that infants are sensitive to statistical information about features in their visual environment, and support a growing literature showing that infants recognize such co-occurrence information within auditory and visual modalities (Kirkham, Slemmer, & Johnson, 2002; Kirkham, Slemmer, Richardson, & Johnson, 2007; Saffran, Aslin, & Newport, 1996). These findings raise two important questions. Do infants simply register these statistical patterns, or do they actually use them in order to make further predictions and inferences? In the literature on causal inference (e.g., Gopnik et al., 2004; Sobel & Kirkham, 2006), there is clear evidence that toddlers, and even infants, will go beyond the simple detection of statistical regularities among events and will use that information to make new predictions about what an object will do. Will infants similarly use the co-occurrence of features within an object to make new predictions about how that object will behave?

Moreover, in noisy natural environments infants are often presented with multiple co-occurrences. How do infants know which co-occurrences to attend to and learn from? Social cues may help infants select appropriate information. By the first few months of life, infants engage in joint attention (Butterworth, 2004), elicited by eye gaze, infant-directed speech, initial eye contact, head turn, and gestures (Carpenter, Nagell, & Tomasello, 1998; Senju & Csibra, 2008). Many investigators suggest that this attentional bias helps infants develop their social cognition and competence (e.g., understanding beliefs, desires, goals, and communicative intent, see Carpenter et al., 1998; Csibra & Gergely, 2006; Repacholi & Gopnik, 1997). However, these cues also can help shape basic cognitive development by helping infants *learn what to learn* in a noisy and exciting environment. While the impact of engaging in joint attention on social competence has been studied extensively (for review, see Carpenter et al., 1998), fewer studies investigate how following social cues can shape basic learning. Some studies have focused on word mapping (e.g., Houston-Price, Plunkett, & Duffy, 2006; Pruden, Hirsh-Pasek, Golinkoff, & Hennon, 2006) and learning linguistic structures (Goldstein & Schwade, 2008). Others have shown that such cues can lead to better recognition of an attended object (e.g., Striano, Chen, Cleveland, & Bradshaw, 2006) as well as to encoding an object's featural or spatial information (Yoon, Johnson, & Csibra, 2008). Recently, Wu

and Kirkham (in press) showed that social cues produce better association of audio-visual events than non-social cues (i.e., flashing squares). These findings suggest that following social cues could shape early basic learning. These conclusions, however, require stronger measures of learning than recognition of cued objects or simple matching of audio-visual events.

The present study tested whether infants develop expectations about object integrity based on the co-occurrences of features with and without social attention-directing cues. Binding co-occurring features is an essential skill for developing veridical representations of the visual world. Previous research shows that adults "chunk" co-occurring features into larger perceptual units (e.g., Orbán, Fiser, Aslin, & Lengyel, 2008). These larger units help to identify discrete objects (Baker, Olson, & Behrmann, 2004; Turk-Browne, Isola, Scholl, & Treat, 2008; for review, see Scholl, 2001). Co-occurrences can highlight the integral features of an object, the basis of visual categorization (Palmeri & Gauthier, 2004).

Methods

Experiment 1

We investigated infants' learning of shape feature co-occurrences and whether those co-occurrences shaped their expectations about the behavior of objects (No Cue condition). Events were presented in a simple spatial layout (Figure 2, top panel), maintaining the same spatial layout between familiarization and test trials.

Participants Eighteen 9-month-old infants (9 females, *M* = 9 months, 1 day, range: 8;14-9;23) participated in this experiment. Three infants were excluded from final analyses due to fussiness (i.e., completing only two out of four blocks). Thirteen infants completed all four blocks; the remaining 5 infants completed 3 blocks. Infants were recruited via local-area advertisements and given t-shirts or bibs to thank them for their participation.

Apparatus Infants' looks were monitored using a Tobii 1750 eye-tracker (www.tobii.com), and events were presented on a 17" monitor attached to the eye-tracking unit. Stimuli were displayed using Tobii's ClearView AVI presentation software, and sounds were played through stereo external speakers. An external video camera on top of the screen allowed the experimenter to determine whether the infant was looking at the display. The shape events were created using Macromedia Director MX 2004, and the movie clips were assembled using Final Cut Express HD 3 (Apple Inc, CA).

Stimuli During the familiarization trials, infants watched sequences of looming shapes. Each sequence contained three patterns (Patterns A, B, and C), and each pattern was composed of 3 differently colored component shapes (see Figure 1). Each infant saw one sequence (repeatedly) during familiarization with a total of 9 shape clusters in each sequence (3 clusters per pattern, 3 patterns per sequence). For each pattern, there was a pair of shapes that were always together and a third shape that varied for each cluster. Each cluster loomed from a minimum of 4.87° to a maximum of 9.72° for 2 seconds. When a cluster grew to its maximum size, it disappeared from the screen, and another cluster appeared. A single sequence appeared in one of two white frames arranged left and right in the lower half of a black screen. Infants viewed the looming pattern while the other frame remained empty.

During the test trials, infants were shown consistent and inconsistent splitting events. Consistent events showed an animation of a cluster breaking into two, with the variable shape moving apart from the other paired shapes. Inconsistent events showed shapes that had been paired together splitting up, where one stayed with the variable shape, and the other moved apart by itself. The same test events were seen by all infants. The events were labeled as consistent or inconsistent according to the pattern that each infant saw during familiarization (i.e., a test event that was a consistent split for Sequence A was inconsistent for Sequence B, and vice versa). Thus, differences in looking time to test events were due to the exposure during familiarization rather than basic perceptual preferences. Each splitting event loomed from a minimum of 4.87° to a maximum of 8.51° for 2 seconds. Then, either a variable or constant shape split off at either a 45°, 180°, or 270° angle (relative to the vertical depending on its position in the cluster) for another 2 seconds until it reached 9.72°. These test events appeared in the same frame as the familiarization events in the lower left or right frame. Consistent and inconsistent test events were shown sequentially in the frame.

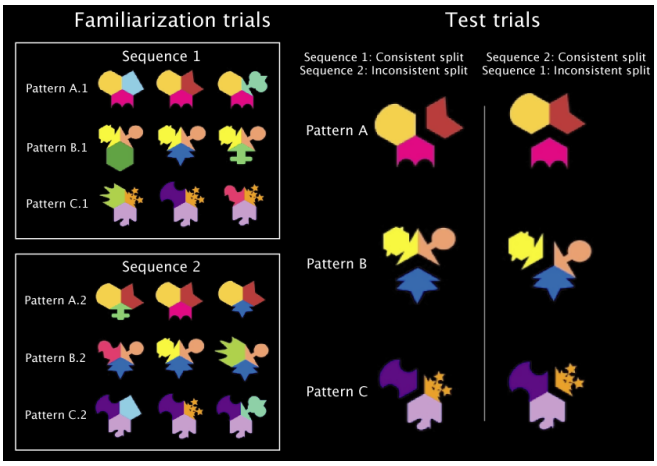


Figure 1: Shape cluster stimuli and patterns during familiarization and test trials.

Design and Procedure Infants sat in a car seat 50 cm from the eye-tracker monitor in a small, quiet room, while their caregivers sat out of their view. The caregivers were instructed to refrain from commenting on the movies or interacting with their infant. The experimenter used a 5-

point calibration with a Sesame Street clip on the infants' looks. For the experiment, infants were shown four blocks of trials. Each block consisted of 6 familiarization trials (i.e., two per pattern) that lasted 6 seconds per trial, and two test trials (consistent or inconsistent) that lasted 12 seconds per trial (3 splitting events per trial, one event per pattern). During each familiarization trial, the three shape clusters in a given pattern were presented sequentially. The training sequence (Sequence 1 or 2) and presentation side were counterbalanced across infants. For each infant, the clusters in the test trials were presented in four orders, which were counterbalanced with a Latin Square across all infants. Whether the first test trial displayed an inconsistent or consistent pattern also was counterbalanced across infants. Blocks 3 and 4 repeated all familiarization and test trials from Blocks 1 and 2. Attention getters (still kaleidoscopic circles or squares with either a "bling" or "boing" sound) spliced the familiarization and test trials. The 1-second clip looped until the infants returned their gaze to the screen for approximately 1500 ms.

Coding The Areas of Interest (AOIs) encompassed slightly larger areas than those of the frames to account for noisy infant saccades. For the familiarization trials, total looking time to the AOI containing the target event was calculated with Tobii's ClearView analysis software. For the test trials, a proportional looking time difference score was calculated by subtracting the percentage of looking to the inconsistent events (total looking time to the inconsistent divided by the total looking time to both test stimuli) from that of the consistent. A negative score reflected a preference for the inconsistent splits, and a positive for the consistent.

Results Infants looked at the familiarization sequence an average of 67.15 seconds, (*SE* = 4.28), 46.63% of the entire presentation time. For the test trials, a one-sample t-test revealed that there was a mean preference for the inconsistent splits overall, and that this preference was significantly higher than chance, $t(17) = -2.33$, $p = .03$, two-tailed, $M = -.11$, $SE = .05$.

Discussion Infants displayed an overall preference for the inconsistent split. They were sensitive to the internal statistics of the shapes within each cluster, and when the two co-occurring shapes separated, this attracted their attention. One interpretation is that the infants had represented the co-occurring pair of shapes as a single object, and noticed that object breaking apart. This preference gave us a baseline of learning that allowed us to investigate how learning under more difficult conditions can be influenced by social cues.

Experiment 2

Experiment 2 increased the difficulty of the test events and examined the effect of introducing a social cue during familiarization. At test, splitting events were presented simultaneously (rather than sequentially) in the lower left and right frames. In order to explore how a social cue might influence infants' learning of the sequences (Figure 2, middle panel), infants were either shown the familiarization sequence by itself (No Cue II condition) or with a face cue that turned down to the patterns (Social Cue condition).

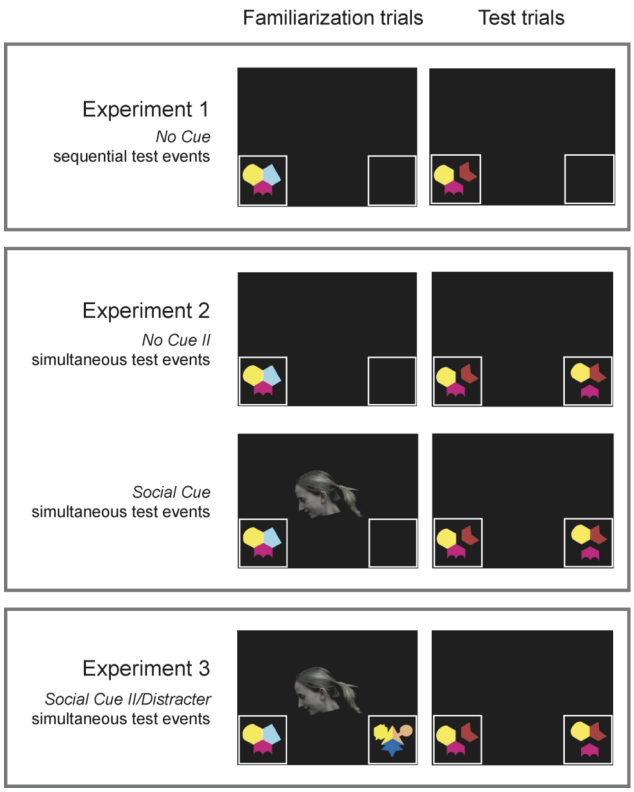


Figure 2. Familiarization and test trials for the four conditions in Experiments 1 to 3.

Stimuli Experiment 2 used the shape stimuli from Experiment 1. The familiarization trials in the No Cue II condition were identical to those from Experiment 1. For the Social Cue condition, before the onset of the shape pattern, a female face appeared in the center of the screen, looked forward, smiled, said "Hi baby, look at this!", and turned to look down at a frame. The pattern then appeared in that frame. The face stayed turned and smiling until the pattern finished (after displaying three clusters sequentially). The pattern disappeared, and the face turned back to the center and changed to a neutral expression as the trial ended. At the beginning of each familiarization trial, the face clip lasted for 5 seconds before the onset of the pattern.

In this experiment, we implemented a preferential looking paradigm by presenting the consistent and inconsistent splits simultaneously in both the right and left locations to increase the complexity of the display. Locations of the splitting events were counterbalanced across infants. To maintain consistency for presentation time and trial numbers, there were two test trials per block that showed the inconsistent and consistent splits simultaneously.

Coding A central AOI was added to the familiarization trials because of the addition of the central face cue. Since

the test trials now included simultaneous consistent and inconsistent splits, the difference score was calculated by subtracting the inconsistent proportional looking time from the consistent for each trial.

Results In addition to the analyses carried out in the previous experiment, we also ran a one-way ANOVA to investigate differences in total looking time during the familiarization trials between the two conditions. For the test trials, we ran a one-way ANOVA (in addition to analyses from Experiment 1) to investigate whether congruency between familiarization and test presentation side (whether infants had to switch sides to look at the inconsistent pattern) affected their preference for the inconsistent event. Infants in the No Cue II condition looked at the familiarization sequence an average of 68.39 seconds (*SE* = 5.30), 47.49% of the entire presentation time, similar to infants in Experiment 1. Infants in the Social Cue condition looked at the pattern for an average of 43.84 seconds, (*SE* = 5.53), 30.44% of the entire presentation time, because these infants split their attention between the face and target shapes. A one-way ANOVA revealed a significant effect of condition on total looking time to the target pattern during familiarization, $F(1, 32) = 10.25$, $p = .003$: Infants looked longer to the shapes in the No Cue II condition than in the Social Cue condition. For the test trials, a one-way ANOVA revealed an effect of condition (Social Cue or No Cue II) on the average difference scores, $F(1, 32) = 4.40$, $p = .04$. Thus, we divided the data set by condition to compare each set of average difference scores to chance: Infants in the Social Cue condition displayed a preference for the inconsistent split, $t(16) = -2.12$, $p = .05$, two-tailed, $M = -.11$, $SE = .05$, while infants in the No Cue II condition did not display a significant preference overall, $t(17) = 1.53$, $p = .15$, two-tailed, $M = .12$, $SE = .09$ (Figure 3).

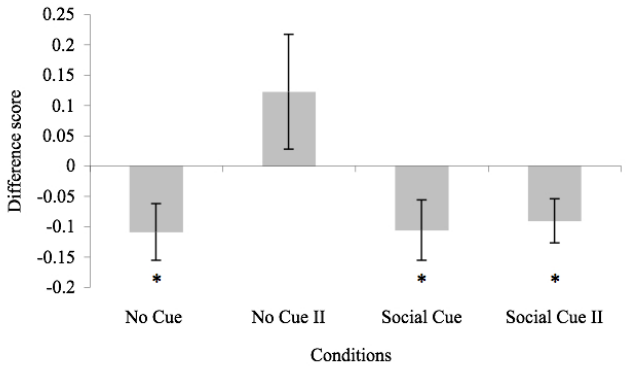


Figure 3. Difference scores across conditions (mean difference between proportional looking times for consistent minus inconsistent events during test). A negative value reflects a preference for the inconsistent splits. *p≤.05

A one-way ANOVA on switching sides revealed no significant effects of switching on the average difference score for both the No Cue condition, $F(1, 16) = .78$, $p = .39$, and the Social Cue condition, $F(1, 15) < .01$, $p > .90$.

Discussion With a noisier test layout, but without a social cue during familiarization, infants showed no significant preference for either the consistent or inconsistent splitting event. Infants who saw social cues during familiarization, however, exhibited a preference for the inconsistent split, similar to infants in Experiment 1. This preference discrepancy between the two conditions was not due to gross inattention to the target event. Infants in the No Cue II condition looked *longer* to the patterns during familiarization than infants in the Social Cue condition. Surprisingly, the lesser amount of attention that infants paid to the sequence in the Social Cue condition resulted in clear expectations about the objects.

Experiment 3

The learning challenge became more difficult again in this experiment, as we presented infants with two sequences simultaneously during familiarization (Social Cue II condition, Figure 2, bottom panel). This setup is a little bit closer to the situation an infant encounters in the real world, where there are many potential sources of statistical structure that could be learnt. Moreover, as is also often the case for the infant, there is a social cue present that could direct and shape learning.

Participants Eighteen 9-month-old infants (10 females, *M* = 9 months, 3 day, range: 8;24-9;22) participated in this experiment. Five infants completed 3 blocks, and 13 infants completed all four blocks.

Stimuli The stimuli, design, and procedure in this experiment were the same as those from the Social Cue condition in Experiment 2 except for the addition of a distracter sequence in the frame that was previously empty in Experiment 2. Infants who were directed to look at Sequence 1 had Sequence 2 as the distracter event, and vice versa. Therefore, the sequence that was the target for one infant was the distracter for another. The target and distracter patterns were displayed simultaneously during familiarization.

Coding With the introduction of a distracter during familiarization, we now measured the efficacy of social cues in directing the infants' attention to the target shape. A difference score for the familiarization trials was calculated in the same manner as for the test trials, except for the inclusion of the central AOI (face). During the test trials, the central AOI was disregarded while calculating the difference score, since there was no visual stimulus in the center during test.

Results Infants looked at the cued familiarization pattern an average of 39.14 seconds, (*SE* = 4.49), 27.18% of the entire presentation time and the non-cued pattern an average of

14.40 seconds, (*SE* = 1.65). Difference scores during familiarization indicated that 17 infants (94.44%) followed the cue and looked longer to the cued pattern than the non-cued pattern, $t(17) = 7.67$, $p < .01$, $M = .42$, $SE = .05$. For the test trials, a t-test on the difference score revealed a preference for the inconsistent splitting events compared to chance, $t(17) = -2.48$, $p = .02$, two-tailed, $M = -.09$, $SE = .04$. A one-way ANOVA on the effect of switching sides between familiarization and test trials revealed no significant effects of switching on the average difference score, $F(1, 16) = .55$, $p = .47$.

Discussion Infants in this experiment displayed a preference for the inconsistent split, similar to the preference in Experiment 1 (No Cue condition) and Experiment 2 (Social Cue condition). Interestingly, infants showed this preference despite a) being exposed to an equally salient pattern in the other frame during familiarization, and b) looking for a short amount of time to the target event compared to the other three conditions. These findings suggest that social cues elicit rapid learning in a noisy environment. Infants did not seem to process much of the distracter pattern, as they otherwise would have preferred the opposite event.

General Discussion

We have shown that 9-month-old infants use visual feature co-occurrences to form representations of object integrity, and that with multiple streams of visual information available in their environment, infants will use social cues to select the ones that they learn. First, these findings extend previous work showing that infants recognize visual feature co-occurrences (Fiser & Aslin, 2002) by demonstrating that infants consider co-occurring features as a larger perceptual unit that should remain fused. If this is the case, then infants may consider these larger perceptual units as integral to an object, as is the case with adults (see Scholl, 2001). Therefore, it is plausible that tracking co-occurrences in infants (as in adults) supports essential skills such as object recognition, categorization, and word learning. The fact that statistics are useful for infants in the visual domain echoes findings in the auditory domain and in studies of causal inference. For example, infants use speech segments that are statistically consistent as labels for objects (Graf Estes et al., 2007). Importantly, these findings support the idea that statistical learning is a powerful mechanism that is *useful* and that leads to inferences beyond detection of the statistical pattern itself.

Social attention cues shaped infants' learning about objects, the second finding in our experiments. This effect remained despite the distraction of the face, change in test spatial layout, and additional distracter patterns during familiarization. Again, this finding in the visual domain is similar to those in the auditory domain showing that infant-directed speech and visual face stimulus facilitates word segmentation (Sell & Kaschak, 2009; Thiessen, Hill, & Saffran, 2005). Importantly, this study grounds an emerging literature showing that social cues mediate infants' basic learning via three key aspects: 1) using a complex measure of learning, 2) comparing learning effects with and without social cues, and 3) investigating such learning with younger, prelinguistic infants. Together, these aspects allow for stronger evidence that social cues mediate learning from the first year.

One could ask whether the social character of the face drove the learning effect or whether a non-social cue would have been equally effective. In this regard, it is noteworthy that the face appeared equidistant from the target and distracter frame during familiarization. Hence, it was not merely the presence of the cue that facilitated learning but the fact that the cue was a face that turned towards one stimulus rather than another. Furthermore, our previous work showed that when cued by flashing squares, 8-month-old infants remember *where* they were cued to rather than *what* they were cued to (Wu & Kirkham, in press). An attention-directing non-social central cue is an interactive stimulus: If a stimulus interacts with the infant and then turns in one clear direction, the infant will follow the object's 'gaze' (e.g., Johnson, Slaughter & Carey, 1998). A recent study, however, showed that 18-month-olds do not map labels onto objects 'gazed' on by an interactive non-social stimulus, only those gazed on by human faces (O'Connell, Poulin-Dubois, Demke, & Guay, 2009), and this interactive stimulus might be argued to itself have social features. In future studies, we intend to find a suitable (and effective) non-social cue for this experimental paradigm. For now, we claim only that social cues facilitate visual statistical learning. We do not claim, however, that social cues are the *only* attention cues that aid learning, nor that they produce better learning than any other attention cue.

In conclusion, our findings suggest that co-occurrence information and social cues inform and direct learning in infancy. In particular, though social cues may temporarily detract attention away from certain learning events in the world, they appear to stimulate infants to display the learning better in complex situations than when infants learn on their own without attention cues. Investigating how infants interact with different cues in the environment (see Goldstein et al., in press) and the developmental trajectory of the use of such cues (e.g., Hollich, Hirsh-Pasek, & Golinkoff, 2000) would clarify the extent to which they shape cognitive development.

Acknowledgments

We thank Paul Quinn, Teodora Gliga, Denis Mareschal, Richard Aslin, and József Fiser for helpful comments on this work. This research was supported by a grant to NZK and RW from the University of London Central Research Fund and a grant to Mark Johnson from the UK Medical Research Council, G0701484.

References

Baker, C. I., Olson, C. R., & Behrmann, M. (2004). Role of attention and perceptual grouping in visual statistical learning. *Psychological Science, 15,* 460–466.

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review, 94,* 115-147.

Butterworth, G. (2004). Joint visual attention in infancy. In G. Bremner & A. Slater (Eds.), *Theories of Infant Development.* (pp. 317-354). Oxford: Wiley-Blackwell.

Carpenter, M., Nagell, K., & Tomasello, M. (1998). Social cognition, joint attention, and communicative competence from 9 to 15 months of age. *Monographs of the Society for Research in Child Development, 63* (4, Serial No. 255).

Csibra, G. & Gergely, G. (2006). Social learning and social cognition: The case for pedagogy. In Y. Munakata & M. H. Johnson (Eds.), *Processes of Change in Brain and Cognitive Development. Attention and Performance XXI* (pp. 249-274). Oxford: Oxford University Press.

Fiser, J., & Aslin, R. N. (2002). Statistical learning of new visual feature combinations by infants. *Proceedings in the National Academy of Sciences, 99,* 15822-15826.

Gopnik, A., Glymour, C., Sobel, D., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and bayes nets. *Psychological Review, 111*, 3-32.

Goldstein, M. H., & Schwade, J. A. (2008). Social feedback to infants' babbling facilitates rapid phonological learning. *Psychological Science, 19,* 515-523.

Goldstein, M. H., Waterfall, H., Lotem, A., Halpern, J., Schwade, J., Onnis, L., Edelman, S. (in press). General Cognitive Principles for Learning Structure in Time and Space. *Trends in Cognitive Sciences*.

Goldstone, R. L. (2000). Unitization during category learning. *Journal of Experimental Psychology: Human Perception and Performance, 26,* 86-112.

Graf Estes, K., Evans, J. L., Alibali, M. W., & Saffran, J. R. (2007). Can Infants Map Meaning to Newly Segmented Words?: Statistical Segmentation and Word Learning. Psychological Science, 18, 254-260.

Hollich, G., Hirsh-Pasek, K., & Golinkoff, R. M. (2000). Breaking the language barrier: An emergentist coalition model for the origins of word learning. *Monographs of the Society for Research in Child Development*, *65 (3, Serial No. 262)*.

Houston-Price, C., Plunkett, K., & Duffy, H. (2006). The use of social and salience cues in early word learning. *Journal of Experimental Child Psychology, 95,* 27-55.

Johnson, S. C., Slaughter, V., & Carey, S. (1998). Whose gaze will infant follow? The elicitation of gaze following in 12-months-olds. *Developmental Psychology, 1,* 233-238.

Kirkham, N. Z., Slemmer, J., & Johnson, S. P. (2002). Visual statistical learning in infancy: evidence for a domain general learning mechanism. *Cognition, 83,* B35-B42.

Kirkham, N. Z., Slemmer, J., Richardson, D. C, & Johnson, S. P. (2007). Location, location, location: development of spatiotemporal sequence learning in infancy. *Child Development, 78,* 1559-1571.

Mareschal, D., Quinn, P. C., & French, R. M. (2002). Asymmetric interference in 3- to 4-month-olds' sequential category learning. *Cognitive Science, 26,* 377-389.

O'Connell, L., Poulin-Dubois, D., Demke, T., & Guay, A. (2009). Can infants use a nonhuman agent's gaze direction to establish word–object relations? *Infancy*, *14*, 414–438.

Orbán G., Fiser, J., Aslin, R. N., & Lengyel, M. (2008). Bayesian learning of visual chunks by human observers. *Proceedings in the National Academy of Sciences, 105*, 2745–2750

Palmeri, T. J., & Gauthier, I. (2004). Visual object understanding. *Nature Reviews Neuroscience, 5,* 291-303.

Pruden, S. M., Hirsh-Pasek, K., Golinkoff, R. M., & Hennon, E. A. (2006). The birth of words: Ten-month-olds learn words through perceptual salience. *Child Development, 77*, 266-280.

Rakison, D. H., & Butterworth, G. (1998). Infants' use of object parts in early categorization. *Developmental Psychology*, *34,* 49-62.

Repacholi, B., & Gopnik, A. (1997). Early reasoning about desires: Evidence from 14- and 18-month-olds. *Developmental Psychology, 33,* 12-21.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science, 274,* 1926–1928.

Scholl, B. J. (2001). Objects and attention: The state of the art. *Cognition, 80,* 1-46.

Schyns, P. G., & Rodet, L. (1997). Categorization creates functional features. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23,* 681–696.

Senju, A., & Csibra, G. (2008). Gaze following in human infants depends on communicative signals. *Current Biology, 18,* 1-4.

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition, 106,* 1558-1568.

Sobel, D. M., & Kirkham, N. Z. (2006). Blickets and babies: The development of causal reasoning in toddlers and infants. *Developmental Psychology, 42*, 1103-1115.

Striano, T., Chen, X., Cleveland, A., & Bradshaw, S. (2006). Joint attention social cues influence infant learning. *European Journal of Developmental Psychology, 3,* 289-299.

Wu, R., & Kirkham, N. Z. (in press). No two cues are alike: Depth of learning is dependent on what orients attention. *Journal of Experimental Child Psychology*.

Yoon, J. M. D., Johnson, M. H., & Csibra, G. (2008). Communication-induced memory biases in preverbal infants. *Proceedings of the National Academy of Sciences, 105*, 13690 - 13695.

Younger, B. A., & Cohen, L. B. (1986). Developmental change in infants' perception of correlations among attributes. *Child Development, 57,* 803–815.

Electrophysiological Evidence for Multiple Representations of Number in the Human Brain

Frank Kanayet (kanayet.1@osu.edu)
Department of Psychology, 255 Psychology Building
Columbus, OH 43206 USA

John Opfer (opfer.7@osu.edu)
Department of Psychology, 245 Psychology Building
Columbus, OH 43206 USA

William Cunningham (cunningham.417@osu.edu)
Department of Psychology, 100G Lazenby Hall
Columbus, OH 43206 USA

Abstract

In adult human brains, the horizontal segment of the intraparietal sulcus plays a large role in representing numeric magnitude. In children and non-human primates, however, frontal cortex may play a larger role. We hypothesized that there is a link between observed developmental changes in locus of representation (frontal to parietal) and type of representation used (logarithmic to linear). Participants were presented with number lines and asked to judge accuracy of linear, logarithmic, or log-linear placements. Consistent with hypotheses, event-related potentials generally revealed greatest parietal N1 amplitudes for linear placements and greatest frontal P3 amplitude for logarithmic placements. Additionally, effects of linear placements on cortical activity were moderated by numerical magnitude: parietal N1 amplitudes decreased with magnitude, whereas frontal P3 amplitudes increased with magnitude. These results suggest adults possess logarithmic and linear representations of number, and when logarithmic representations were elicited; there was greater involvement of frontal cortex.

Keywords: Numerical cognition; representation; brain imaging; event related potentials.

Introduction

Whether a pollster evaluating the sampling process for an election poll, a parishioner telling the time by counting the tolls of a church bell, or a child figuring out how much candy she had received on this Halloween versus a previous one, mental representations of numerical magnitude are important for projecting the future, monitoring the present, and learning from the past. Moreover, this ability to code our experiences numerically must scale consistently regardless of the shape, size, sensory modality or context in which particular numeric magnitudes are presented.

Two prominent brain areas have been implicated in humans' and other animals' representation of numerical magnitudes: the *prefrontal cortex* and the *horizontal segment of the intraparietal Sulcus* (HIPS). Most studies have shown that HIPS plays a major role in numeric representation, with magnitude coded in this area as an abstract, notation-independent representation (Dehaene, Piazza, Pinel, Cohen, 2003; Dehaene, Dehaene-Lambertz, & Cohen, 1998; Libertus, Woldroff & Brannon, 2007). However, comparative and developmental studies have found HIPS playing a less prominent role. For example, single cell recordings in monkeys (Nieder & Merten, 2007 & Nieder & Miller, 2004) and *f*MRI studies in children (Ansari et al. 2005; Ansari & Dhital, 2006; Cantlon, Brannon, Carter & Pelphrey, 2006; Cantlon et al., 2009) have shown stronger effects of numerical magnitude on prefrontal cortex than HIPS. Similarly, ERP studies with infants, have also found that nonsymbolic numeric processing generates activity in a parieto-prefrontal network (Izard, Dehaene-Lambertz & Dehaene, 2008; Libertus, Pruitt, Woldorff & Brannon, 2009).

To explain this developmental trend, we propose that: (1) at any given age, the brain represents numeric magnitudes using both a logarithmically-compressed code and a linear code, with the probability of a number being processed by the linear code increasing with age and experience; and (2) logarithmic-coding is predominantly processed in frontal areas, whereas linear coding is predominantly processed in parietal areas. Here we test an implication of this account, namely that large magnitude (low-frequency) numerals are more likely than small magnitude (high-frequency) numerals to be represented in frontal cortex, whereas the reverse is true of parietal cortex.

Development of Numeric Representations

The origin of our developmental hypothesis stems from behavioral studies on development of numeric representations (Booth & Siegler, 2006; Opfer & Siegler, 2007; Opfer & Thompson, 2008; Siegler & Opfer, 2003; Thompson & Opfer, 2008). In these studies, children and adults were asked to estimate the position of numbers on a blank line with the end-points labeled "0" and "100", "0" and "1000", or "0" and "10000". This estimation task is particularly revealing about cognitive representations of numeric value because it transparently reflects the ratio characteristics of the number system. Overall, younger children's estimates typically follow Fechner's Law and

increase logarithmically with actual value, whereas older children's estimates increase linearly. At any given age, however, individual children use both logarithmic and linear representations of number, depending on numerical context. That is, for very large numeric contexts (e.g., on 0-1000 and 0-10000 number lines), children's estimates increase logarithmically; however, the same children will use linear representations when estimating the magnitudes of numbers for small numeric contexts (e.g., on 0-100 number lines). If our developmental hypothesis is correct, it should be possible to identify two different patterns in the brain that are consistent with the type of representation used, thereby providing neural correlates for the logarithmic-to-linear shift hypothesis.

Plausible candidates for these two different patterns of neural activation are provided by the developmental data showing a shift from prefrontal to parietal processing of numerical magnitude (Cantlon, et al., 2009; Rivera, Reiss, Eckert and Menon, 2005). More generally, evidence from perceptual learning has shown that complex conjunctive stimuli are processed by more posterior sites with gains in expertise, both within the visual cortex (Mukai et al., 2007) and between the prefrontal cortex and visual cortex (Eriksson, Larsson, Nyberg, 2008). As a result, information changes from being processed serially and with effort, to being processed in parallel and automatically. Possibly, the same is true for number representation, with the abstract representation that is needed for processing numeric magnitude regardless of shape, size, modality or context originally coming from the prefrontal cortex and gradually shifting to HIPS with gains in expertise.

Present Study

To test our hypothesis, we asked participants to judge whether a number had been accurately marked on a number line, and we evaluated the *Event-Related Potentials* (ERP) generated after participants saw number-line estimates that corresponded or not to a given numeral. By evaluating ERP components related to numeric estimation, we were able to test several predictions derived from our developmental hypothesis. Specifically, we were able to provide a novel test of whether subjects expected positions of numbers on number lines to increase linearly, logarithmically, both, or neither with numeric value, and we were able to test if the topography of those ERP components corresponded to our hypotheses.

Some ERP components can generate diagnostic data about representations of numerical magnitude, even before the subjects' response. Generally, targets that violate subjects' expectations elicit large P3 amplitudes (Donchin, 1981). Thus, numbers marked in non-linear positions would likely generate a higher P3 response than numbers marked in the linear position. Conversely, the N1 component is generally elicited when targets match the subject's orientation of attention (Luck, 2005; Folstein & Van Petten, 2008). Thus, numbers marked in the linear position would be expected to generate higher N1 responses than numbers marked in the non-linear position. Using this logic, ERP components are capable of early detection of both linear and non-linear representations of number. This provides an important test of our hypothesis because automatic, non-linear representations of number might occur in adults before they have time to provide formally correct, learned responses.

Method

Participants

Participants (N = 21, mean age = 20.5, 8 female) were recruited from an introductory psychology class and were awarded course credit for their participation in the experiment. Nineteen participants were right handed, and all had normal or corrected to normal vision.

Design and Procedure

Each problem presented a blank number line with a width of 255 pixels, labeled with '0' on the left end and '1000' on the right end. The numbers presented appeared on the top of the screen 192 pixels over the line (half point between the top of the screen and the number line). The numbers tested were 5, 78, 150, 606, 725 and 938. These numerosities were selected because they sample the whole length of the line and also maximize the discriminability between linear and logarithmic representations. All stimuli were presented in a dark and sound-attenuated room using DirecRT (Jarvis, 2006).

Participants were instructed to identify if the position of the hatch mark on a number line corresponded to the numeral presented by pressing one key if the position of the hatch mark were correct, and by pressing another if the position of the hatch mark were incorrect (keys were counterbalanced between participants).

At the beginning of each trial, the number line with the marked end points appeared and a fixation was placed where the target numerals were going to be shown for a period of 1 second. Next, the stimulus (i.e. the numeral) replaced the fixation for another 1-second interval. After this period, the hatch mark was placed either in the linear, logarithmic or log-linear position. Once the hatch mark was in place, participants had to decide if the mark was correctly placed and to press the appropriate key (no time limit was imposed on participants' responses). After the response, no feedback was provided and a 2000-ms intertrial stimulus interval (ISI) was used.

Participants were tested on three different sets of trials and the design of the study was all within subjects. Thus, on each block, participants encountered each of the six numerals compared to three possible hatch mark positions (linear, logarithmic, log-linear). The experiment consisted of 16 blocks and presentation of the trials was randomized within each block. This corresponds to a total of 288 trials (96 per trial type condition).

ERP Recording Procedure After attaining informed consent from participants, a NuAmps quick cap with 32 Ag/AgCl electrodes (Compumedics Neuroscan, El Paso, TX, USA) was placed on their heads to record their brain activity. Linked ears served as reference during recording. Before the beginning of the experiment, impedances were held below 40 kΩ[1]. The electroencephalogram (EEG) was amplified with an A/D conversion rate of 1000 and a gain of 250mV. Finally, a recording low-pass filter of 300Hz was used.

Before analysis of the data, the raw EEG data were processed offline using BESA (Version 5.2). Raw data were re-referenced to an average of all electrodes and a digital 0.1 to 30Hz bandpass filter was used. Also, artifact correction (Berg & Scherg, 1994) was used to reduce ocular artifacts and blinks. After artifact correction, an artifact rejection procedure (tailored to each individual) was conducted. After this process, 7 participants – who had less than 85% of the trials accepted – were removed from further analyses. ERP epochs (-200ms to 1000ms) were created for the three trial types (i.e. linear, logarithmic and log-linear), and for hatch mark number size (i.e. hatch marks that corresponded to small numbers and large numbers).

Results

Behavioral Results

Number comparison is typically characterized by effects of distance and size on speed and accuracy of judgments (Moyer & Landauer, 1967). We obtained similar results for judgments of number line placements. Consistent with distance effects, log-linear trials, which were closer to the correct (linear) placements than logarithmic ones, required more time to solve and resulted in the lowest accuracy rates. Consistent with size effects, judging the location of large numbers (i.e. larger than 500) on the number line required more time than judging the location of small numbers (i.e. smaller than 500), with accuracy also being lower for placement of large numbers compared to small numbers. Finally, there was evidence of interactive effects of size and trial type, with larger effects of trial type for small numbers (ω^2 = .54) than for large numbers (ω^2 = .32) on reaction times. This interaction is interesting because it suggests that representations of small numeric magnitudes are more strongly linear and non-logarithmic than representations of large numeric magnitudes, leading to less discriminability between trial types for the large magnitudes.

A potential problem with accuracy measures, such as those reported above is that they can fail to detect systematic response biases. To address this issue, we conducted d' analyses. Because performance of participants was near ceiling, hits and false alarm rates were corrected. Specifically, hit rates were constructed by the formula (hits + 1)/(total trials + 2), and false alarm rates were constructed by the formula (false alarms + 1)/(total trials + 2).

As predicted by the size effect, discriminability between the linear, and the logarithmic and log-linear trials declined with numeric size (see Figure 1). This result was confirmed by a one-way repeated measures ANOVA ($F(5,100)$ = 36.83, $p < .001$, $\omega^2 = 0.59$). An alternative explanation of this result is that it is due to the distance between the linear and logarithmic trials not being constant throughout the whole range of numbers. Thus, it is possible that the reason why discrimination decreases for the numbers 725 and 938 is because the distances between the linear and logarithmic trials decrease too. To test this alternative hypothesis, we performed a planned comparison between two numbers that differ in size but that have the same distance between the linear and logarithmic hatch mark positions (5 and 725). As predicted by the size effect, even though the distance between the linear and logarithmic trials is equal for these two numbers, discriminability was significantly smaller for 725 (d' = 2.25, SD = 0.78) than for 5(d' = 3.04, SD = 0.56; $p < .001$).

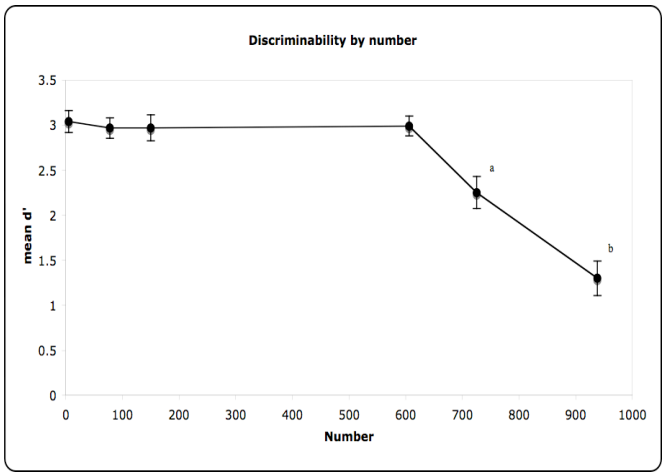


Figure 1: Mean d prime values for each numeral (+ SE). (a) and (b) represent significant differences at $p < .05$..

Electrophysiological Results

To understand the temporal characterization of the number line estimation task, average waveforms were computed for the three experimental trials (i.e. linear, logarithmic, log-linear). Additionally, these waveforms were averaged into four different electrode sites with the purpose of reducing experiment-wise error caused by computing multiple statistical comparisons. The frontal left (FL) electrode site was computed by averaging the electrodes FP1, F3, F7, FC3, and FC7. The frontal right (FR) electrode site was computed by averaging the electrodes FP2, F4, F8, FC4, FC8. The parietal left (PL) electrode site was computed by averaging the electrodes CP3, TP7, P3, P7. The parietal right (PR) electrode site was computed by averaging the electrodes CP4, TP8, P4, P8.

[1] Although the impedance threshold for accepting a participant was 40 kΩ, in reality most of the electrodes achieved impedances of 10 to 15 kΩ.

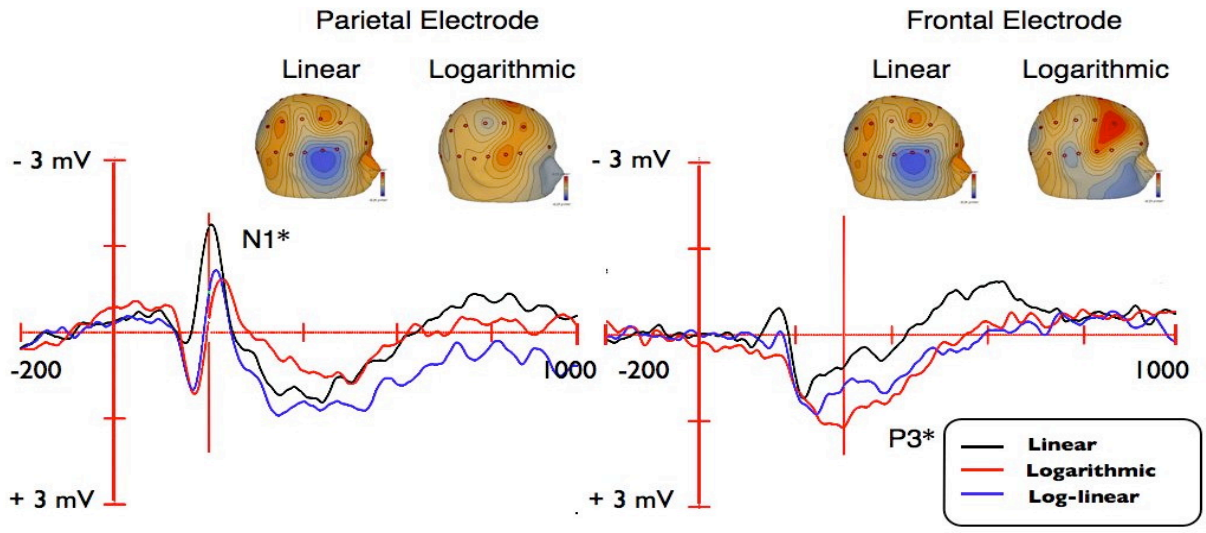


Figure 2: Top: Current source density topographies for linear and logarithmic trial types at 200 ms (left) and at 300 ms (right). Bottom: N1 (140-240 ms) and P3 (260-700 ms) ERP components for linear, logarithmic, and log-linear trials for parietal (left) and frontal (right) electrode sites.

Visual inspection of the waveforms is consistent with the main hypothesis from the study (see Figure 2). First, linear trials generated a greater N1 peak than both the logarithmic and log-linear trials, especially in parietal electrode sites. Moreover, at frontal electrode sites, the logarithmic trials generated a greater P3 peak than the log-linear trials, and in turn, the log-linear trials generated a greater P3 peak than the linear trials. These effects suggest that even before the behavioral response is effectuated, there is a strong recognition of the linear placements of numbers followed by a signal of surprise related to the logarithmic and log-linear placements of numbers.

To test these effects statistically, a 3-way (trial type: linear, logarithmic, log-linear x electrode site: FL, FR, PL, PR x component: N1, P3) repeated-measures ANOVA was conducted on the mean amplitudes calculated for the N1 and P3 time windows. All reported p-values are Greenhouse-Geisser corrected for violations of sphericity assumptions. Results indicated a significant component x electrode interaction, ($F(3,39) = 8.37$, $p < .001$, $\eta^2 = 0.39$). This effect is largely due to a larger N1 component in parietal sites compared to frontal sites. Furthermore, as expected, a trial type x electrode x component interaction was significant ($F(6,78) = 3.85$, $p = .033$, $\eta^2 = 0.23$). This interaction was due to different simple main effects of trial type at the N1 component for the PL ($F(2,26) = 8.07$, $p = .003$, $\omega^2 = 0.25$) and PR ($F(2,26) = 14.81$, $p < .001$, $\omega^2 = 0.40$) electrode sites versus simple main effects of trial type at the P3 component for the FR ($F(2,26) = 4.69$, $p = .048$, $\omega^2 = 0.16$) and PL ($F(2,26) = 18.25$, $p < .001$, $\omega^2 = 0.45$) electrode sites.

To explore this more closely, we computed average waveforms for the correct linear trials with hatch marks that corresponded to small numbers (i.e. 5, 78, 150) and to large numbers (i.e. 606, 725, 938). As can be seen in Figure 3, compared to small numbers, large numbers generated smaller N1 peaks at parietal electrode sites and larger P3 peaks at frontal electrode sites. This pattern of results indicates that small numbers were expected to appear in the linear position, whereas large numbers were not. Thus, even though participants made the correct response for both types of numbers, the brain shows evidence that large numbers and small numbers are processed differently.

To test these results statistically, we conducted a 3-way (Condition: small numbers, large numbers x Electrode: FL, FR, PL, PR x Component: N1, P3) repeated measures ANOVA. Results showed a significant electrode x component interaction ($F(3,39) = 9.14$, $p < .001$, $\eta^2 = 0.41$). This effect is largely due to a change in polarity from the N1 to the P3 components in parietal electrodes. Moreover, as expected there was a significant trial condition x electrode x component interaction ($F(3,39) = 8.05$, $p < .001$, $\eta^2 = 0.38$). Post-hoc analysis revealed that when hatch marks were positioned linearly, a greater N1 component at the PR electrode site ($F(1,13) = 5.67$, $p = .033$, $\omega^2 = 0.14$) was elicited by small numbers than by large numbers. Also, at the FL electrode site, linearly positioned hatch marks elicited a greater P3 component ($F(1,13) = 6.31$, $p = .026$, $\omega^2 = 0.16$) for large numbers than for small numbers.

DISCUSSION

We aimed to provide a temporal characterization of brain activity evoked by representations of numeric magnitudes. This characterization supported two conclusions: (1) the adult brain continues to represent numeric magnitudes using both a logarithmically-compressed code and a linear code, with the probability of a number being processed by the linear code decreasing with numeric magnitude (and thus frequency and prior experience); and (2) that logarithmic-coding is predominantly processed in frontal areas, whereas

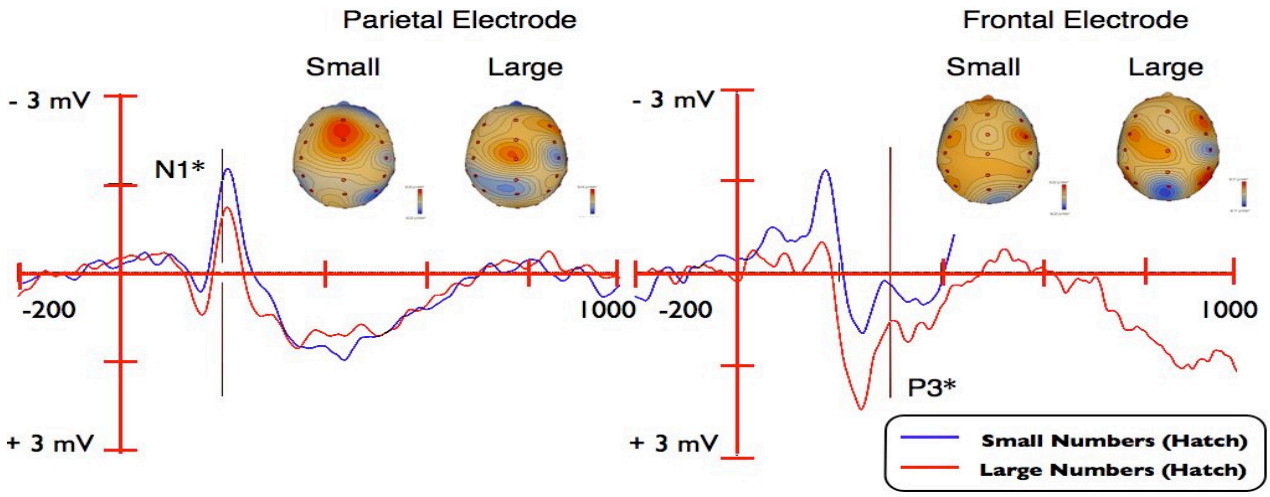


Figure 3: Top: Current source density topographies for correct linear trials divided into small and large numbers at 200 ms (left) and at 300 ms (right). Bottom: N1 (140-240 ms) and P3 (260-700 ms) ERP components for correct linear trials divided into small and large numbers for parietal (left) and frontal (right) electrode sites

linear coding is predominantly processed in parietal areas. These findings are important because they are consistent with our proposed explanation for a key finding in the developmental neuroscience of number representation. Namely, that although it has been found that HIPS is crucial for numeric processing (Dehaene, et al., 2003), studies that test children, have found a greater involvement of the prefrontal cortex (Ansari et al. 2005; Ansari & Dhital, 2006; Cantlon, et al., 2006; Cantlon, et al., 2009; Rivera, Reiss, Eckert and Menon, 2005).

Evidence supporting our first conclusion comes from several findings from this study. Behavioral results indicate that both linear and non-linear positions of numbers were judged as correct, with probability of non-linear positions being judged as correct increasing as numbers increased in size. Electrophysiological results were consistent with this behavioral finding. Small numbers shown in the linear position generated a greater N1 peak than did large numbers shown in the linear position. Similarly, large numbers shown in the linear position generated a greater P3 peak than did smaller numbers. Moreover, these electrophysiological findings held even when subjects' behavior correctly identified locations as linear. Thus, neither behavioral nor electrophysiological results are consistent with the idea that numbers are solely represented linearly or solely non-linearly.

Evidence supporting our second conclusion comes solely from electrophysiological data. As indicated by the N1 component, smaller numbers were more easily identified than large numbers, and this identification was predominantly found in parietal sites (Dehaene, 1996; Libertus et al., 2007). On the other hand, linear trials that corresponded to larger numbers (that are less entrenched) generated a greater surprise response (as indicated by the P3 component) in frontal electrode sites. Likewise, the results for the discrimination between linear and logarithmic conditions showed that the significant N1 component for linear trials was located in parietal electrodes, while the P3 component for logarithmic trials was located in frontal electrodes.

An alternative hypothesis that could explain the role of prefrontal cortex is that it could be signaling general attentional demands or processes of response selection that become more active for more difficult tasks. However, using habituation paradigms, Cantlon and her collaborators have found greater activity in the prefrontal cortex of children for numeric processing (Cantlon et al., 2006; Cantlon et al., 2009). Therefore, this finding rules out the response selection hypothesis because there was no response needed, and brings doubts about the attentional demands hypothesis because there should not be significant differences in attentional demands between the number and the control tasks used. Furthermore, in our analysis, large numbers have smaller discrepancies between the linear and logarithmic trials. Therefore, if this hypothesis were correct we would expect smaller P3 amplitudes for them. Instead, we found that large numbers elicited larger frontal P3 amplitudes.

In conclusion, this is the first study to provide neural evidence for competing representations of numerical magnitude. By using an ERP paradigm with a number line estimation task, we were able to investigate numeric processing both before participants had reached a final decision about magnitude and a behavior was executed. This paradigm led to the novel finding that not all numbers are represented as linearly positioned on the number line, despite the fact that participants' judgments are very linear at the behavioral level. In this way, our findings are consistent with a novel developmental proposal that can integrate apparently contradictory results regarding the neural representation of numeric magnitude.

References

Ansari, D., & Dhital, B. (2006). Age-related changes in the activation in the Intraparietal Sulcus during nonsymbolic magnitude processing: An event-related functional magnetic resonance imaging study. *Journal of Cognitive Neuroscience*, 18(11), 1820-1828.

Ansari, D., Garcia, N., Lucas, E., Hamon, K. & Dhital, B. (2005). Neural correlates of symbolic number processing in children and adults. *NeuroReport*, 16, 1769-1773.

Berg, P., & Scherg, M. A. (1994). A multiple source approach to the correction of eye artifacts. *Electroencephalography and Clinical Neurophysiology, 90,* 229-241.

Booth, J. & Siegler, R. (2006). Developmental and individual differences in pure numerical estimation. *Developmental Psychology, 42(1),* 189-201.

Cantlon, J. F., Brannon, E. M., Carter, E. J. & Pelphrey, K.A. (2006). Functional imaging of numerical processing in adults and 4-y-old children. *PLoS Biol 4(5)*: e 125. DOI: 10.1371/journal.pbio.0040125.

Cantlon, J. F., Libertus, M. E., Pinel, P., Dehaene, S., Brannon, E. M. & Pelphrey, K. A. (2009). The neural development of an abstract concept of number. *Journal of Cognitive Neuroscience, 21*(11), 2217-2229.

Dehaene, S. (1996). The Organization of brain activations in number comparison: Event-Related Potentials and the additive-factors method. *Journal of Cognitive Neuroscience, 8(1),* 47-68.

Dehaene, S., Dehaene-Lambertz, G. & Cohen, L. (1998). Abstract representations of numbers in the animal and human brain. *Trends in Neurosciences, 21,* 355-361.

Dehaene, S., & Mehler, J. (1992). Cross-linguistic regularities in the frequency of number words. *Cognition, 43(1),* 1-29.

Dehaene, S., Piazza, M., Pinel, P. & Cohen, L. (2003). Three parietal circuits for number processing. In J.I.D. Campbell, *Cognitive Neuropsychology, 20*, 487-506.

Donchin, E. (1981). Surprise…surprise? *Psychophyisiology, 18(5),* 493-513.

Eriksson, J., Larsson, A., & Nyberg, L. (2008). Item-specific training reduces prefrontal cortical involvement in perceptual awareness. *Journal of Cognitive Neuroscience, 20(10),* 1777-1787.

Folstein, J. R., Van Petten, C. (2008). Influence of cognitive control and mismatch on the N2 component of the ERP: A review. *Psychphysiology, 45,* 152-170.

Izard, V., Dehaene-Lambertz, G., & Dehaene, S. (2008). Distinct cerebral pathways for object identity and number in human infants. *PloS Biology, 6(2),* 275-285.

Jarvis, B. G. (2006). DirectRT (Version 2006.2.28) [Computer Software]. New York, NY: Empirisoft Corporation.

Libertus, M. E., Pruitt, L. B., Woldorff, M. G., & Brannon, E. M. (2009). Induced Alpha-band Oscillations Reflect Ratio-dependent Number Discrimination in the Infant Brain. *Journal of Cognitive Neuroscience, 21*,(12), 2398-2406.

Libertus, M. E., Woldorff, M. G. & Brannon, E. M. (2007). Electrophysiological evidence for notation independence in numerical processing. *Behavioral and Brain Functions, 3(1),* 1-15.

Luck, S. J. (2005). *An introduction to the event-related potential technique.* Cambridge, MA: MIT Press.

Moyer, R. S., & Landauer, T. K. (1967). Time required for judgments of numerical inequality. *Nature*, *215*, 1519-1520.

Mukai, I., Kim, D., Fukunaga, M., Japee, S., Marrett, S., & Ungerleider, L. (2007). Activations in visual and attention-related areas predict and correlate with the degree of perceptual learning. *Journal of Neuroscience, 27(42),* 11401-11411.

Nieder, A. & Merten, K. (2007). A Labeled-line code for small and large numerosities in the monkey Prefrontal cortex. *Journal of Neuroscience, 27(22),* 5986-5993.

Nieder, A. & Miller, E.K. (2004). A parieto-frontal network for visual numerical information in the monkey. *Porcedings of the National Academy of Scioences of the United States of America, 101*, 7457-7462.

Opfer, J. E. & Siegler, R. S. (2007). Representational change and children's numerical estimation. *Cognitive Psychology, 55*(3), 169-195.

Opfer, J. E., & Thompson, C. A. (2008). The trouble with transfer: Insights from microgenetic changes in the representation of numerical magnitude. *Child Development*, *79*, 790 -806.

Rivera, S. M., Reiss, A. L., Eckert, M. A., & Menon, V. (2005). Developmental changes in mental arithmetic: Evidence for increased functional specialization in the left inferior parietal cortex. *Cerebral Cortex, 15(11),* 1779-1790.

Siegler, R. S. & Opfer, J. E. (2003). The development of numerical estimation: Evidence for multiple representations of numerical quantity. *Psychological Science, 14*(3), 237-243.

Siegler, R.S., Thompson, C.A., & Opfer, J.E. (2009). The Logarithmic-to-linear shift: One learning sequence, many tasks, many time scales. *Mind, Brain, and Education, 3,*143-150.

Thompson, C. A., & Opfer, J. E. (2008). Costs and benefits of representational change: Effect of context on age and sex differences in magnitude estimation. *Journal of Experimental Child Psychology, 101*, 20 – 51.

The perception of number from long-term memory

Jiaying Zhao (jiayingz@princeton.edu)
Department of Psychology, Green Hall,
Princeton University, NJ 08540 USA

Nicholas B. Turk-Browne (ntb@princeton.edu)
Department of Psychology, Green Hall,
Princeton University, NJ 08540 USA

Abstract

The perception of numerosity is supported by two systems: an exact system for small quantities, and an approximate system for large quantities. Two properties arise from the combination of these two systems: the accuracy of numerosity judgments changes qualitatively above the capacity limit for exact representations, and the ability to discriminate two quantities depends on the numerical distance between the quantities and the relationship of this distance to the absolute magnitudes. These well-characterized aspects of number cognition have typically been studied in judgments of numerosity based on visual arrays. Across four experiments we demonstrate remarkably similar effects in numerosity judgments based on incidental long-term memory. These results suggest that similar mechanisms and constraints may operate when estimating numerosity from representations of external sensory input and internal representations derived from long-term memory.

Keywords: Numerosity judgments; perception; memory

Introduction

Perception is typically considered a set of processes for analyzing incoming sensory input. Some researchers have argued that the same perceptual and attentional mechanisms can be directed inward during prospection, memory retrieval, and memory search. To what extent are the mechanisms that underlie judgments on the basis of immediate perception similar or constrained in the same way as the mechanisms that underlie judgments derived from internal representations?

One way to answer this question is to examine the relation between numerical judgments on the basis of immediate external visual input and the numerical judgments based on internal representations from long-term memory. In the former case, several features of immediate numerical perception have been discovered. People are very accurate and fast at enumerating small quantities (6 or fewer), a process termed subitizing (Kaufman, Lord, Reese, & Volkman, 1949), while they are subject to capacity limitations with large quantities, a process termed approximation (Mandler & Shebo, 1982; Trick & Pylyshyn, 1994; Feigenson, Dehaene, & Spelke, 2004). Moreover, when discriminating between numerosities, error rates and response times are inversely related to the numerical distance between numbers (Moyer & Landauer, 1967; Dehaene, Dupoux, & Mehler, 1990). When distance is held constant, error rates and response times increase as the absolute sizes or magnitudes of the two numbers increase (Whalen, Gallistel, & Gelman, 1999; Barth, Kanwisher, & Spelke, 2003). Developmental research has also suggested core systems for representations of exact and approximate quantities (Feigenson et al., 2004; Wood & Spelke, 2005; Opfer & Siegler, 2007).

Numerical judgments from long-term memory have previously been examined in the context of event frequency (Hasher & Zacks, 1979; Hintzman & Block, 1971; Howell, 1973). For instance, according to the strength hypothesis proposed by Hintzman (1969), frequency judgments of an event are determined by the strength or the repetition of the memory trace representing the event. Another view is that frequency judgment is a direct readout of the number of stored traces of an event based on its time lag rather than the strength of a single trace (Hintzman & Block, 1971). More recently, Brown (1995, 1997, 2002) argues that judgments of event frequency depend on context memory in that people rely on enumeration when different contexts produce distinct memory traces.

Here we relate the two areas using tools from studies of immediate numerical perception to focus on how people make numerosity judgments from long-term memory. We investigate the extent to which properties and constraints of numerosity judgments on the basis of long-term memory mirror those of judgments based on immediate perception.

Experiment 1

The purpose of this experiment is to test whether unexpected numerosity judgments from long-term memory are accurate, and whether capacity limitations in subitizing and short-term memory for external visual input also apply for judgments based on retrieved internal representations.

Participants

Twenty students from Princeton University participated in exchange for partial course credit (13 female, mean age 19.2 yrs, SD = 1.1).

Materials

Stimuli were chosen from an image set containing 60 distinct object categories. To manipulate numerosity, 50 of these categories were pseudo-randomly assigned to a number between 1 and 10 such that each numerosity level was represented by 5 categories. One exemplar image was chosen from each of these categories, and was presented the corresponding number of times over the course of the first phase of the experiment. For example, if at numerosity level '3' the categories of *dog*, *bear*, *car*, *flower*, and *horse* were chosen, then one exemplar from each category would be presented 3

times throughout the first phase intermixed with images from other numerosity levels. The order of image presentation was randomized for each participant with the constraint that categories could not repeat back-to-back. In addition to the 275 images of interest ($[10+9+8+2+1] \times 5$), 20 additional images were selected randomly from the remaining 10 categories with half presented at the beginning and half at the end to control for primacy and recency effects.

Procedure

In the first phase of the experiment, the participant viewed each image and determined whether it corresponded to a *natural* or *artificial* category by pressing one of two keys. This cover task prevents an explicit strategy such as counting, and is orthogonal to the primary manipulation. On each trial, an image appeared on the screen for 2 seconds, followed by an interstimulus interval of one second. The full trial sequence of 295 images lasted about 20 minutes. The participant then completed an unrelated distractor task for 15 minutes.

In the second phase of the experiment, the participant again viewed single images of objects on the screen, and estimated how many times between 1 and 10 they had seen that image in the first part of the experiment by pressing a number key on the keyboard. The 50 exemplar images of interest were presented in a random order. The accuracy and response time for each image were recorded. Filler images were not presented. It is worth noting that participants often expressed surprise when receiving these instructions, and that in post-experiment debriefing no subject reported being aware that their memory for number would be tested in the second part. These responses suggest that any effects we observe reflect incidental encoding of number in long-term memory.

Results

We compared estimated numerosity from the second phase against the objective numerosity from the first phase. At every numerosity level we averaged across the five categories at that level for each participant, and then averaged these mean estimates across participants. These estimates were compared against the objective numerosity by computing differences within participant and averaging these differences across participants. Results are shown in Figure 1.

To quantify performance, estimated numerosities were modeled as a function of objective numerosities using linear regression. Since estimated and objective numerosities were bounded (from 1 to 10), perfect performance would result in a slope of 1 and an intercept of 0. In contrast, chance performance (i.e. guessing) would lead participants to randomly distribute their responses and would result in a slope of 0. If they randomly distributed their estimates across all response options, the expected intercept would be 5.5 ([10+1]/2). Thus, we can judge accuracy in estimating numerosity from incidental encoding on a continuum from perfect performance (slope = 1, intercept = 0) to chance performance (slope = 0, intercept = 5.5). The linear regression analysis was performed within each participant. The mean

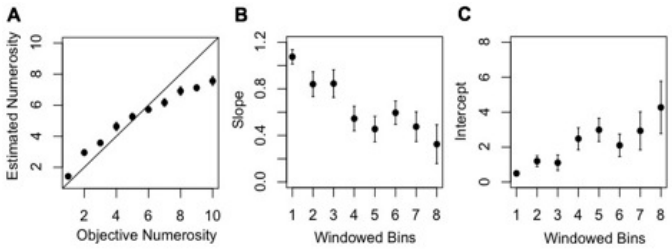


Figure 1: (A) Mean estimated numerosity plotted against the number of times each image was presented during the first phase (objective numerosity). (B) Mean slope of a linear model applied to the data in Figure 1A over windows of three numerosity levels (e.g. '1' reflects window from 1 to 3 on the x-axis of Figure 1A). (C) Mean intercept of a linear model applied over the same windows. Error bars reflect std. error.

slope across participants was 0.64 ($SD = 0.12$, median = 0.64) and the mean intercept was 1.59 ($SD = 0.71$, median = 1.53).

Prior research has indicated a capacity limitation in highly accurate numerosity judgments of about 4 objects (Mandler & Shebo, 1982; Trick & Pylyshyn, 1994). Thus, despite overall high performance, accuracy may be non-stationary across objective numerosity. In particular, the slope of linear regressions over smaller windows of objective numerosity may approach 0 (with a corresponding increase in intercept). Such a finding would support the existence of capacity limitation in numerosity judgments from long-term memory.

We thus ran a linear regression across all possible windows of 3 contiguous numerosity levels for each participant. That is, separate linear regressions were run on windows [1,3], [2,4]...[8,10]. For each window, the slope and the intercept values were averaged across participants (see Figure 1B). To quantify our results, one way repeated-measures ANOVAs were performed for slopes and intercepts. There were reliable main effects of numerosity on both measures (slope $F[7,145] = 9.4$, $p < .01$; intercept $F[7, 145] = 3.9$, $p < .01$). Post-hoc Tukey HSD tests revealed that the slope values for window [1,3] ($M = 1.08$, $SD = 0.28$) were reliably higher than the rest of the slope values, while the intercept values for the same window ($M = 0.50$, $SD = 0.65$) were reliably lower than the rest of the intercept values.

These results suggest that performance starts off near perfect, and declines steadily as a function of objective numerosity. From inspection of Figure 1B, there appears to be a marked drop in slope and increase in intercept after window [3,5], suggesting a capacity limitation around 4-5 repetitions. To quantify these intuitions, we imposed a mixture of perfect and chance performance on the data in Figure 1A. In particular, we tested a mixed linear model to identify at what point along the objective numerosity line at which participants' performance started to level off and to decline. In this mixed model, at a given point n on the numerosity line, $y = 1 \times x + 0$ for x in $[0, n]$, and $y = 0 \times x + 5.5$ for x in $[n + 1, 10]$. In other

words, we fit the perfect performance linear model to data up to numerosity n and a chance performance model to data from numerosity $n + 1$ to 10. It should be noted that at numerosity 0 the mixed model becomes a complete chance model and at numerosity 10 it becomes a complete perfect performance model. The average model fits across participants are shown in Figure 2A. SS*error* was minimized at n from 4 to 8.

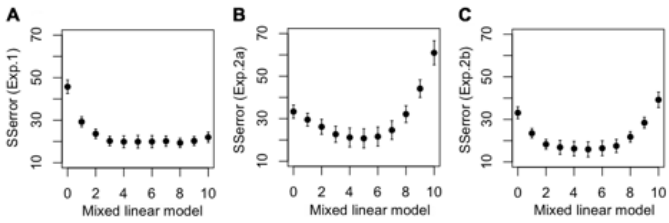


Figure 2: Estimated numerosities in (A) Exp. 1, (B) Exp. 2a, (C) Exp. 2b, tested against the following model: at a given point n on the x-axis, $y = 1 \times x + 0$ for x in $[0, n]$, and $y = 0 \times x + 5.5$ for x in $[n + 1, 10]$. Error bars reflect std. error.

Discussion

The results in Exp. 1 demonstrate that participants can make remarkably accurate numerosity judgments from long-term memory. Moreover, participants were unaware that their memory for number would be tested, and therefore this accuracy reflects incidental/automatic encoding of number in memory. Despite overall high accuracy, estimates of numerosity became less accurate when an image has been presented more than 5 times. To our knowledge, this provides the first demonstration of capacity limitations in judgments operating over internal representations, extending and replicating robust findings of similar limits in short-term memory for external visual input (Xu & Chun, 2006).

Experiment 2a

This putative capacity limit observed in Exp. 1 could reflect the fact that we repeated identical images many times. Such repetition could lead to habituation or reduced attention that would impair further encoding. To test this explanation, here we replicate Exp. 1, but present multiple exemplars of the same category once, rather than the same exemplar multiple times. This increased novelty may improve encoding and may facilitate retrieval.

Participants

Twenty students from Princeton University participated in exchange for partial course credit (14 female, mean age 19.7 yrs, $SD = 1.5$). None had served in the previous experiment.

Materials

The materials were identical to Exp. 1 with one important exception: instead of presenting the same exemplar image from each category n times, n distinct exemplars were randomly drawn from each category and presented only once. For example, if the category *dog* was assigned to the numerosity level '3', then images of three different dog breeds would each be presented once. This increased the novelty and variance within each category, possibly allowing for more accurate estimates about large numerosity levels.

Procedure

The procedure was identical to Exp. 1 except for one aspect of the second phase: category names (e.g., "dog") were used to elicit estimates of how many images of that category had been presented in the first phase. Names were used rather than images because there were several possible images to choose from for many of the categories. Thus, 50 category names were presented in a random order.

Results

Data were analyzed in the same manner as Exp. 1. Results are shown in Figure 3. To quantify performance, we modeled estimated numerosities as a function of objective numerosities using linear regression (as in Exp. 1). Surprisingly, the mean slope across participants was 0.33 ($SD = 0.16$, median = 0.34), reliably lower than the mean slope ($M = 0.64$) in Exp. 1 ($t[38] = 6.9$, $p < .01$). The mean intercept across participants was 2.72 ($SD = 1.18$, median = 2.58), reliably larger than the mean intercept ($M = 1.59$) in Exp. 1 ($t[38] = 3.7$, $p < .01$). Contrary to our predictions, these results suggest that performance was worse in Exp.t 2a vs. Exp. 1, i.e. farther from perfect performance, and closer to a chance uniform distribution.

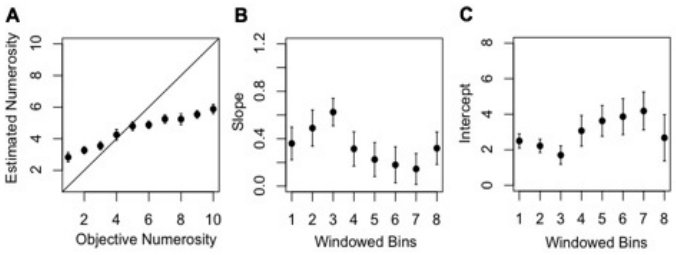


Figure 3: (A) Mean estimated numerosity plotted against the number of exemplars of each category from the first phase. (B) Mean slope of a linear model applied to the data in Figure 3A over windows of three numerosity levels. (C) Mean intercept of a linear model applied over the same windows. Error bars reflect std. error.

We again explored the presence of a capacity limit by computing the slopes and intercepts of linear functions over windows of objective numerosity. Despite the relatively poorer performance in this experiment, visual inspection of Figures 3B and 3C revealed a qualitative difference between windows [3,5] and [4,6]. One way repeated-measures ANOVAs revealed a main effect of numerosity on intercept values ($F[7, 145] = 5.6$, $p < .01$). The main effect of numerosity on slope values did not reach significance ($F[7, 145] = 1.5$, $p > .05$).

To further characterize a change in accuracy as a function of numerosity we also tested the same mixed linear models on the data in Figure 3A. The model fits are shown in Figure 2B. Again, across numerosities the data were best represented by a model in which performance was perfect up to 5 exemplars and plateaued for larger numbers.

Discussion

Contrary to our predictions, providing multiple exemplars for number estimation did not improve accuracy. In fact, performance was worse than in Exp. 1, where judgments were based on the number of repetitions of a single stimulus. This suggests that performance in Exp. 1 did not asymptote around 5 presentations because of habituation or diminished attention. The worse performance here could reflect poor encoding of images presented only once, or source confusion during retrieval in response to a category label. For example, "dog" may retrieve more than 3 exemplars, with reduced performance reflecting an inability to distinguish exemplars intruding from prior experience. Regardless, despite overall worse performance, participants nevertheless showed consistent capacity limitations to Exp. 1 of approximately 5 memories.

Experiment 2b

While the worse performance in Exp. 2a vs. Exp. 1 can be due to weaker encoding of number, it remains possible that potentially more accurate judgments were hampered by a less informative retrieval cue. To examine this possibility, here we replicate Exp. 1 with category labels during retrieval.

Participants

Twenty students from Princeton University participated in exchange for partial course credit (12 female, mean age 19.1 yrs, $SD = 1.3$). None had served in previous experiments.

Materials

The stimuli used here were the exactly same as those in Exp. 1. The exemplar image from each category was repeatedly presented depending on the numerosity value.

Procedure

The procedure was identical to Exp. 2a. Participants were cued by a category name (e.g. "dog") and estimated how many times they had seen an image from that category.

Results

Results are shown in Figure 4. We modeled estimated numerosities as a function of objective numerosities using linear regression. The mean slope across participants was 0.46 (SD = 0.15, median = 0.48), reliably higher than the mean slope in Exp. 2a (M = 0.33; t[38] = 2.6, $p < .01$) but reliably lower than that in Exp. 1 (M = 0.64; t[38] = 4.2, $p < .01$). The mean intercept across participants was 2.31 (SD = 0.82, median = 2.06), which was not statistically smaller than the mean intercept in Exp. 2a (M = 2.72; t[38] = 1.3, $p > .05$) but was reliably larger than that in Exp. 1 (M = 1.59; t[38] = 2.9, $p < .01$). Thus, after matching retrieval cues, numerosity judgments from multiple repetitions of the same exemplar remain more accurate than those from multiple exemplars of the same category.

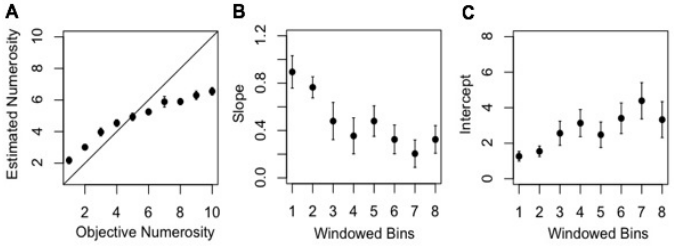


Figure 4: (A) Mean estimated numerosity plotted against the number of exemplars of each category from the first phase. (B) Mean slope of a linear model applied to the data in Figure 4A over windows of three numerosity levels. (D) Mean intercept of a linear model applied over the same windows. Error bars reflect std. error.

To explore possible capacity limitations, we computed slopes and intercepts for linear models over windows of three contiguous numerosities. One way repeated-measures ANOVAs revealed main effects of numerosity on both measures (slope F[7, 145] = 3.2, $p < .01$; intercept F[7, 145] = 4.6, $p < .01$). Post-hoc Tukey HSD tests revealed that the slope value for window [1,3] (M = 0.90, SD = 0.61) were reliably higher than the rest of the slope values, while the intercept for the same window (M = 1.26, SD = 1.20) were reliably lower than the rest of the intercepts.

Moreover, from Fig. 1, 3, and 4 the slope and intercept values for Exp. 2b appear to resemble those in Exp. 1 more than those in Exp. 2a. Collapsing across participants, the mean slope values were highly correlated with those in Exp. 1 (r = 0.84, $p < .01$), but not with those in Exp. 2a (r = 0.51, $p > .05$). Capacity limitations were examined for each numerosity level where performance plateaued (see Fig. 2C). As in previous experiments, the mixed model fit best at 5.

Discussion

This experiment reveals that judgments of numerosity are more accurate for multiple repetitions of the same exemplar than for single presentations of multiple exemplars of the same category. The category label did somewhat impair performance, but critically, cannot entirely explain the poor performance in Exp.2a. Across three experiments we observed evidence that unexpected judgments of numerosity from past experience are accurate and subject to capacity limitations.

Experiment 3

Capacity limitations are a signature property of perceptual number processing. To further build our case for a similarity between external and internal number estimation, we consider two additional classic effects in the numerosity literature: the *distance effect* and the *magnitude effect*. These

psychophysical effects are evident when two quantities must be discriminated. The distance effect refers to the relative ease with which participants can discriminate two quantities that are farther apart in number space (e.g., 2 vs. 3 compared to 2 vs. 4). The magnitude effect refers to the fact that a given numerical distance can become harder to discriminate at higher magnitudes (e.g., 2 vs. 3 compared to 8 vs. 9). We explore whether these psychophysical effects also occur when discriminating numerosities defined by long-term memory.

Participants

Twenty students from Princeton University participated in exchange for partial course credit (11 female, mean age 19.6 yrs, $SD = 1.6$). None had served in previous experiments.

Materials

The stimuli in Exp. 1 were used. The exemplar image of a category was repeatedly presented depending on the numerosity value.

Procedure

The first phase was identical to Exp. 1, such that all number encoding was incidental. In the second phase, participants judged which of two images they had seen more times during the first phase. Based on numerosity levels, we paired images so as to fully cover the space of possible distances and proportional distances (for the magnitude effect). At distance of 1 we paired an image that was presented n number of times with the one that was presented $n + 1$ number of times. Since numerosity levels range from 1 to 10, there were 9 pairs at the distance of 1 (e.g., 1 vs. 2, 2 vs. 3, etc.). The same pairing method was applied to the distances of 2, 3, 4, and 5, which resulted in 8, 7, 6, and 5 pairs, respectively. Thus, a total of 35 pairs were generated. For each pair, two images were presented side by side on the screen and participants judged which image appeared more times by pressing one of two buttons for left or right. The order of pairs was randomized for each participant and the position on the screen of the image with the larger numerosity was randomized on each trial.

Results

To assess distance effects, we pooled all of the pairs of each distance within participant and computed mean accuracy and RT. To assess magnitude effects, we conditioned every distance on the smaller number of the pair (e.g., 1 vs. 3: distance = 2/base = 1). Mean accuracy and RT were again computed within participant for each bin. Results are shown in Fig. 5.

As visible in Fig. 5A and 5B, accuracy increased while response time decreased as a function of distance (the opposite of a speed/accuracy tradeoff). Collapsing across participants, there was a strong correlation between accuracy and distance ($r = 0.98$, $p < .01$) but a weaker correlation between response time and distance (r = -0.61, $p > .05$). To test the reliability of these relationships, we correlated accuracy and response time with distance within each participant, transformed the resulting correlation coefficients to Z scores using Fisher's

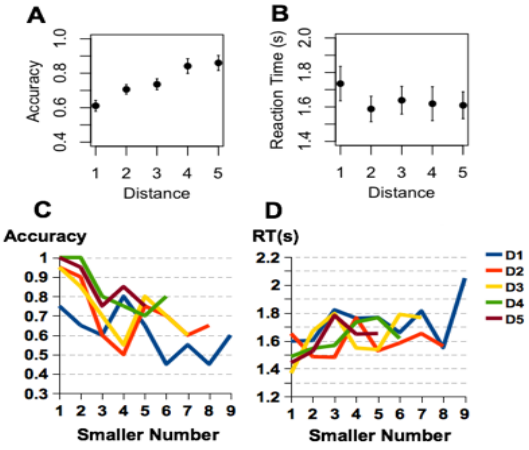


Figure 5: (A) Mean accuracy as a function of distance. (B) Mean response time as a function of distance. (C) Mean accuracy as a function of the smaller number in a pair and also the distance (e.g. D1 = distance of 1). (D) Mean response time as a function of the smaller number and also the distance. Error bars reflect std. error.

transform, and then compared these values against the null relationship of 0 using a one-sample t-test. Accuracy was positively correlated with distance (mean $z = 0.21$, $t(19) = 5.72$, $p < .01$), and response time was negatively correlated with distance (mean z = -0.08, $t(19) = 2.15$, $p < .05$). Thus, participants were able to discriminate between two numerosities in long-term memory, and performed better as a function of the absolute number difference.

To assess the magnitude effect, we performed repeated-measures ANOVAs on accuracy and response time as a function of the base and the distance of each pair. We could not include both the base and distance factors in a two-factor ANOVA because the design was not factorial (e.g., there were no distance = 5/base = 9 trials), and thus used separate one-way ANOVAs. There were main effects of base and distance on accuracy (base $F[8, 683] = 18.5$, $p < .01$; distance $F[4, 691] = 50.1$, $p < .01$). There were also main effects of base and distance on response time (base $F[8, 683] = 16.4$, $p < .01$; distance $F[4, 691] = 13.9$, $p < .01$). The robust effect of base demonstrates a magnitude effect in numerosity judgments from memory.

Discussion

When judging which exemplar appeared more times, participants were more accurate and faster when the distance was larger. Holding the distance constant, participants were also more accurate and faster when the base number of presentations was relatively small.

General Discussion

We have found that unexpected numerosity judgments based on long-term memory can be highly accurate, and that this accuracy is maintained for up to a small quantity of retrieved

memories. These findings are largely in agreement with studies of numerical judgments based on immediate visual perception. Moreover, judgments of numerosity were more accurate for multiple repetitions of the same exemplar than for single presentations of multiple exemplars. This rules out the possibility that apparent capacity limitations reflect habituation or reduced attention. The fact that multiple exemplars were in fact *worse* than single exemplars could relate to failures of source monitoring (Dougherty & Franco-Watkins, 2003; Johnson, Hashtroudi, & Lindsay, 1993): when cued by a category, participants may have been unable to screen out extra-experimental memories. Such failures may have been minimized when retrieval was cued by an image rather than a category label, but the effect persisted when retrieval cue was equated. When discriminating between two incidentally encoded numerosities, performance increased with distance but decreased as the magnitude or the absolute size of the numbers increased. These results were again in line with findings on numerosity comparison based on immediate visual perception (Whalen et al., 1999; Barth et al., 2003).

Since numerosity judgments based of long-term memory exhibited similar properties and constraints as compared to immediate perception, our findings are consistent with the existence of a common underlying numerosity mechanism for perception and memory. While our focus has been on drawing this analogy, there may also be important differences between the perception and memory of number information. For example, while the perception of number has been well-characterized by a hard capacity limit on exact judgments, memory for number may be better characterized by a more continuous logarithmic or power law function. Moreover, we do not yet know whether input representations retrieved from long-term memory are the same as those constructed during online perception (e.g., whether numerosity is estimated over a set of retrieved episodes, or directly read out from a symbolic or analog representation of quantity updated during encoding). These are important questions for future research, but our results nevertheless provide initial evidence for a striking symmetry between snap judgments of number from a single sensory stimulus, and delayed (surprise) judgments based purely on long-term memory. In sum, mechanisms that seem to exist in the service of sensory processing may have broader functional roles in cognition, operating similarly over input from internal or external sources.

References

Barth, H., Kanwisher, N., & Spelke, E. (2003). The construction of large number representations in adults. *Cognition*, *86*, 201–221.

Brown, N. R. (1995). Estimation strategies and the judgment of event frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 1539–1553.

Brown, N. R. (1997). Context memory and the selection of frequency estimation strategies. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*, 898–914.

Brown, N. R. (2002). Frequency processing and cognition. In P. Sedlmeier & T. Betsch (Eds.), (chap. Encoding, representing, and estimating event frequencies: A multiple strategy perspective). Oxford: Oxford University Press.

Dehaene, S., Dupoux, E., & Mehler, J. (1990). Is numerical comparison digital? analogical and symbolic effects in two-digit number comparison. *Journal of Experimental Psychology: Human Perception and Performance*, *16*, 625–641.

Dougherty, M. R. P., & Franco-Watkins, A. M. (2003). Reducing bias in frequency judgment by improving source monitoring. *Acta Psychologica*, *113*, 23–44.

Feigenson, L., Dehaene, S., & Spelke, E. (2004). Core systems of number. *Trends in cognitive sciences*, *8*, 307–314.

Hasher, L., & Zacks, R. T. (1979). Automatic and effortful processes in memory. *Journal of Epxerimental Psychology: General*, *108*, 356–388.

Hintzman, D. L. (1969). Apparent frequency as a function of frequency and the spacing of information. *Journal of Experimental Psychology*, *80*, 139–145.

Hintzman, D. L., & Block, R. A. (1971). Repetition and memory: Evidence for a multiple-trace hypothesis. *Journal of Experimental Psychology*, *88*, 297–306.

Howell, W. C. (1973). Representation of frequency in memory. *Psychological Bulletin*, *80*, 317–331.

Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source monitoring. *Psychological Bulletin*, *114*, 3–28.

Kaufman, E. L., Lord, M. W., Reese, T. W., & Volkman, J. (1949). The discrimination of visual number. *American Journal of Psychology*, *62*, 498–525.

Mandler, G., & Shebo, B. J. (1982). Subitizing: An analysis of its component processes. *Journal of Experimental Psychology: General*, *111*, 1–22.

Moyer, R. S., & Landauer, T. K. (1967). Time required for judgments of numerical inequality. *Nature*, *215*, 1519–1520.

Opfer, J. E., & Siegler, R. S. (2007). Representational change and children's numerical estimation. *Cognitive Psychology*, *55*, 169–195.

Trick, L. M., & Pylyshyn, Z. W. (1994). Why are small and large numbers enumerated differently? a limited-capacity preattentive stage in vision. *Psychological Review*, *101*, 80–102.

Whalen, J., Gallistel, C. R., & Gelman, R. (1999). Nonverbal counting in humans: The psychophysics of number representation. *Psychological Science*, *10*, 130–137.

Wood, J. N., & Spelke, E. S. (2005). Chronometric studies of numerical cognition in five-month-old infants. *Cognition*, *97*, 23–39.

Xu, Y., & Chun, M. M. (2006). Dissociable neural mechanisms supporting visual short-term memory for objects. *Nature*, *440*, 91–95.

Mapping number words to approximate magnitudes: associative learning or structure mapping?

Jessica Sullivan (jsulliva@ucsd.edu)
David Barner (barner@ucsd.edu)
Department of Psychology
University of California, San Diego

Abstract

How do we link number words to the magnitudes they represent? We investigated the roles of associative learning and structure mapping in linking the Approximate Number System to number words. Four tasks demonstrated that individuals have strong associative links between magnitudes and number words for relatively small sets, but have weak associative links for larger sets. These results point to multiple mechanisms for the mapping of number words to magnitudes.

Keywords: Language acquisition, approximate magnitudes, word learning, number words

Introduction

How does language represent human numerical knowledge? Are the referents of numerals determined primarily by inference and logical relations between words? Or do we identify the referents of words like *twelve* and *fifty-seven* by associating them, item-by-item, with sets in the world? Humans can represent the approximate numerical magnitude of a set nonverbally using the Approximate Number System (ANS), and previous research has shown that our system of number language is deeply linked to the ANS: adults recruit the ANS when estimating the cardinality of rapidly presented arrays, and judgments about verbally presented numerals show many of the same biases as nonverbal judgments about quantities (Barth, Kanwisher, & Spelke, 2003; Whalen, Gallistel, & Gelman, 1999, Piazza, Izard, Pinel, Le Bihan, & Dehaene, 2004; Holloway & Ansari, 2008; Duncan & McFarland, 1980). We know that mature mathematical thinkers can relate their verbal number system to the nonverbal ANS—but we don't know *how* these systems are mapped onto each other.

By some accounts, number words gain their numerical content in part via a mapping to the ANS (Verguts & Fias, 2006, Piazza, Pinel, Le Bihan, & Dehaene, 2007; Mundy & Gilmore, 2009). However, surprisingly little is known about the nature of this mapping, and descriptions of possible mechanisms are rare in the literature. It is therefore not well understood what roles associative and inferential processes play in relating number words to ANS representations. One possibility is that as humans accumulate experience with number words, they form item-specific associations between individual words (e.g., *ten*) and corresponding magnitudes (e.g., approximately 10 objects). This view, which we will call the associative mapping hypothesis (AM), predicts that the strength of mappings should vary according to how frequently words are used to refer to perceptually available quantities. Also, it predicts that mappings should be relatively independent of one another, such that a change in one mapping does not automatically impact another mapping. The AM hypothesis is supported by evidence from children's early mappings of small number words: children learn the associations between their first number words and the magnitudes they refer to one at a time, taking nearly two full years to learn the associations between the number words "one" through "four" and the magnitudes they denote (Wynn, 1992).

A problem with the AM hypothesis is that humans may get only limited experience with the denotations of some number words, and no experience at all with others. It seems implausible, for example, that experience with 1 million things would be required to support estimates for sets of this size. Instead, it appears that inferential abilities are required to give meaning to unfamiliar quantities, perhaps on the basis of more familiar amounts. One possibility, for example, is that associative learning about small quantities (e.g., 1 – 10) supports a structure mapping (SM): a linking of representations in one domain to those in another on the basis of their common structure (Izard & Dehaene, 2008; Thompson & Opfer, in press).

One signature of structure mapping is that when a subject's response for a given quantity is changed via feedback (i.e., calibrated), responses for other quantities should shift accordingly. Evidence for this comes from Izard & Dehaene (2008), who showed that mislabeling a visually presented set (e.g., calling a set of 30 dots "twenty-five") led participants to shift their estimates not only for the calibrated quantity, but for all quantities tested. However, this study provided only small amounts of miscalibration, resulting in tiny differences between conditions. For example, even for the most extremely miscalibrated trials, participants mapped the number word "thirty" onto arrays with an average set size between 31.5 and 33.7 (neither of which is perceptually discriminable from 30 for normal adults). As a result, although it appears that SM plays some role in estimation, it remains unclear how malleable estimation behaviors are, and thus what the relative roles of AM and SM are in the mapping of number words to the ANS.

The present study explored the nature of the relationship between the ANS and the count list, and the relative contributions of associative and structural mapping. We hypothesized that neither mechanism, in isolation, could explain how number words are mapped to the ANS.

Whereas AM is limited by the experiences that individuals have with particular magnitudes, in order for a SM to have content, it must be supported by reliable mappings between small number words and magnitudes. Without associative mappings for at least some number words, we submit, no structure mapping could take place. To our knowledge, no previous study has directly tested the contribution of these two mechanisms, and as a result, little is known about their relative contribution to number word mappings. To test our hypothesis, we conducted an experiment with four within-subjects measures that probed for evidence of associative and structural mappings at different numerosities.

In the Calibrated Estimation task, we measured participants' accuracy at labeling sets after they were provided misleading information about the range of set sizes to be presented. We predicted that quantities that have strong associative mappings should be less susceptible to calibration than those with weaker mappings. As in previous studies, we expected that participants' estimates of magnitudes would be influenced by the feedback they received. However, in the present task, we made two critical methodological changes. First, we provided only verbal calibration. While in past studies participants were shown an array and then told that it contained *x* dots (where *x* was either an accurate or inaccurate number word label), in the present study we did not mislabel arrays. Instead, we simply stated that "the largest set you will see is *x"*. In this way, we ensured that any influence of feedback was not because participants constructed new associative mappings, but was due purely to an inference about the structure mapping relation. A second difference was that we provided much more extreme calibration than in previous studies, in order to test the strength of associative mappings throughout the number line. We reasoned that misleading feedback should not influence estimation performance for any magnitude with a strong AM link to the number system, whereas structurally linked magnitudes should be quite susceptible to calibration.

An assumption in our analysis of the estimation task is that a participant's individual estimates act as inputs to a structure mapping, and that each act of estimation therefore constrains later estimates in an experiment. On this view, we predicted that a very similar task in which participants were asked to match a label to one of two visually presented sets would disrupt the structure mapping process. We reasoned that presenting two sets to participants would cause them to experience uncertainty, and thus prevent them from calibrating their mappings across trials. As such, we predicted that performance on this task should suffer for number words that have weak associative mappings to magnitudes. Where stronger AMs exist (e.g., for smaller numbers), performance should not suffer as much, since by our hypothesis the forced choice task relies on the associative strength between the number word and its corresponding magnitude representation. Although past studies have used a forced choice method to test mappings in young children (Lipton & Spelke, 2005; Gilmore & Mundy, 2009), these studies provided calibration before the study in the form of a familiarization phase. Our study, in contrast, sought to remove all forms of feedback, whether from the experimenter or from trial-to-trial self-calibration, in order to test the strength of associative mappings at different magnitudes.

We conducted two additional tasks as within-subject controls for the Calibrated Estimation and Number Matching tasks. The first was an Uncalibrated Estimation task, which served as a within-subject baseline for the Calibrated Estimation task. The second, a Numerical Discrimination task, used stimuli identical to those in the Number Matching task but asked subjects to judge which of the two sets on each trial was more numerous. This ensured that participants were able to discriminate the quantities used in the Number Matching task, and that any difficulties with the forced choice task were due to their number word mappings and not other stimulus properties.

For both the Number Matching task and the Calibrated Estimation task, we predicted that participants would exhibit strong associative mappings for some number words, resulting in smaller effects of calibration and higher levels of success on the forced choice task. In particular, we predicted that the strength of associative mappings would be strongest for the smallest number words, due to relatively greater experience with these words and their corresponding quantities. Corresponding to this, we also predicted that larger magnitudes would be more susceptible to miscalibration in estimation, and be more difficult to map to number words in the forced choice task.

Materials and Methods

Participants

Thirty adults from the UCSD community participated for course credit. One additional participant was excluded from analyses for failure to complete all tasks.

Procedure

Participants were seated approximately 40 cm from a 27" Mac OSX computer screen and completed 4 computerized tasks. Half of the participants completed the Number Matching task first, and half completed the Discrimination task first. All participants then completed first the Uncalibrated and then the Calibrated Estimation task.

Number Matching: This task evaluated participants' ability to match number words with one of two visually presented sets. Participants heard a number word, saw two arrays of dots flash sequentially on a computer screen, and judged which array best matched the word. Stimuli were sets of red dots on a black screen, and were presented for 400 ms. Trials compared sets that differed in numerical magnitude by either a 1:2 ratio or 3:4 ratio. Sets were matched for density on half of the trials and for total surface area on the other half, and comparisons ranged from small (4 vs. 8) to large (370 vs. 740). For 1/3 of the trials, the smaller of the two sets presented contained fewer than 30

items (Small Number Trials), for 1/3 it contained more than 30 and fewer than 110 (Medium Number Trails), and for the remaining 1/3, it more than 110 items (Large Number Trials).

Numerical Discrimination*:* This task served as a within-subjects control for the Number Matching task to ensure that participants could discriminate the quantities presented. The stimuli and procedure were identical to the Number Matching task, except that participants indicated which set contained more dots (instead of matching a word with a set).

Calibrated Estimation: This task tested the malleability of participants' numerical estimates. Participants saw sets of dots and estimated their numerosities, recording their estimates using the numeric keypad on a computer keyboard. Although the largest set that participants saw was 350 in all conditions, participants were told that the largest set they would see was either 75 (N=10), 375 (N=10), or 750 (N=10). Critically, this misleading feedback could not be used to alter associative mappings since, because unlike in previous calibration studies, the incorrect number word anchor was not paired with an array, and thus participants could not form new associative mappings between magnitudes and number word labels (e.g., Izard & Dehaene, 2008; Shuman, unpublished thesis). Instead, this feedback could only have influenced participants' notions about the structure and range of magnitudes.

Stimuli were sets of red dots on a black screen. Fifteen numerosities ranging from 8-350 were presented 36 times each. Each numerosity was matched for both density (15 trials) and total occupied area (15 trials) with each other numerosity presented, and non-numerical properties of the sets were otherwise varied for the remaining 6 trials. Participants received 270 of the possible 540 trials in the Calibrated condition, and 270 in the Uncalibrated condition. Trials were presented in random order.

Uncalibrated Estimation: This task served as a within-subjects control for the Calibrated Estimation task to provide a baseline of the participant's Uncalibrated estimates. Stimuli and instructions were identical to those in the Estimation task, except participants were given no information about the largest set they would see.

Results

Number Matching and Discrimination

If participants have associative mappings between individual number words and approximate magnitudes, then they should be able to use these mappings to guide the labeling of sets in the Number Matching task. For example, if the number word *twenty* is associatively mapped to a mental representation of 'about 20 things', then participants should never match the word *twenty* to an array that is discriminably different from 'about 20' (e.g., 40). In other words, for all magnitudes that have associative mappings, participants should perform equally well on the Discrimination and Number Matching tasks, because discriminably different magnitudes should be mapped to unique number words.

We first explored whether performance differed on the Number Matching task, as compared to the Discrimination task. Qualitatively, every participant performed worse on the Number Matching task than on the Discrimination task, indicating that matching a number word to the correct array is more difficult than judging which array is more numerous. A paired-samples t-test revealed that this trend reached significance (all $p<.05$) for 21/30 participants (binomial probability $p<.01$). This effect was consistent across the range of comparisons presented: participants were significantly less accurate on the Number Matching task than the Discrimination task for Small, Medium, and Large Number Trials (all $p<.01$).

Table 1: Mean accuracy on the Discrimination and Number Matching tasks by magnitude of smaller set

Set Size	Discrimination	Number Matching
Small (<30)	92%	85%
Medium	85%	63%
Large (>110)	82%	70%

To explore in greater detail whether magnitude influenced accuracy on these two tasks, we compared accuracy on the Number Matching task to accuracy on the Discrimination task for each magnitude presented. Interestingly, participants did *not* perform significantly worse on the Number Matching Task than the Discrimination Task for any of the comparisons containing magnitudes smaller than 15 (Dunnett's mean comparison, all $p>.05$)[1]. This suggests that participants may have associative mappings for relatively small number words. Consistent with this, accuracy on the Number Matching task was not constant for all magnitudes tested: a regression of accuracy onto set magnitude by ratio revealed an effect of ratio ($F(1,1954)=22.4$, $p<.01$) and an effect of set magnitude ($F(1,1954)=70.5$, $p<.01$), but no interaction ($F(1,1954)=1.4$, *ns*). This pattern of performance indicates that participants had greater difficult matching labels to larger sets relative to smaller ones, and is consistent with the hypothesis that small, but not large, magnitudes are associatively linked to number words.

To explore this trend further, we compared accuracy on the Number Matching task for Small, Medium, and Large Number Trials. Accuracy was significantly different as a function of set magnitude ($F(2,987)=54.3$, $p<.01$), and a post-hoc comparison of mean accuracy revealed that

[1] Accuracy also did not differ between the Number Matching and Discrimination tasks for the four largest comparisons presented. This effect appears to be driven by trials where the larger set was also the correct set: participants may have developed a simple response heuristic for these trials like "when I hear an unusually large number word, I select the larger of the two sets". This is unlikely to be evidence of associative mapping for large sets.

participants performed significantly better on Small Number Trials than either Medium or Large Number Trials (both $p<.01$), but that accuracy on the Medium and Large Number Trials did not differ significantly from each other (t=-.57, *ns*). Participants did not struggle to match number words to magnitudes when the words presented were relatively small, yet accuracy declined rapidly as a function of set magnitude, and remained low for the largest sets.

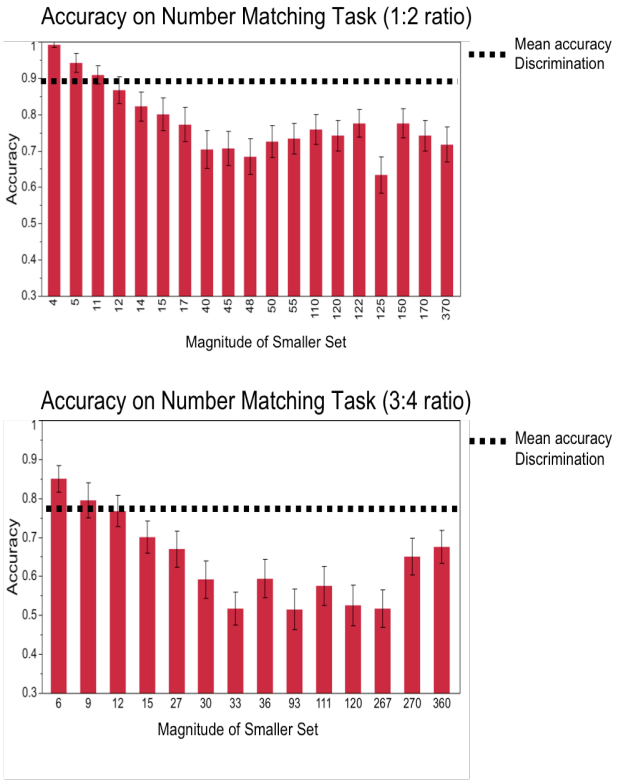


Figure 1: Accuracy on Number Matching task as a function of magnitude of the smaller set being compared

Estimation and Calibrated Estimation

Before completing analyses, we excluded all responses smaller than 2, and all responses more than a factor of 10 larger or smaller than the presented numerosity, as these were likely to be typos. Additionally, we removed outliers on a participant-by-participant basis by excluding all data points more than 3 SD from the mean of each participant's estimate of each presented magnitude (total exclusions: 313/15,450 data points)[2].

As a group, participants provided estimates that were related to the presented magnitude, and that were influenced by misleading feedback (miscalibration): a regression of estimates onto magnitude by calibration type revealed a significant relationship between set magnitude and estimate ($F(1,15129)=3912.5$, $p<.01$), a significant effect of Calibration ($F(1,15129)=605.3$, $p<.01$), and a significant interaction of Calibration and magnitude ($F(3,15129)=256.3$, $p<.01$). Participants were influenced by misleading feedback, and the influence of miscalibration differed as a function of magnitude.

To explore the influence of feedback at an individual level, we performed the identical regression on each individual's data. 24/30 participants showed an effect of calibration: 9/10 participants who were calibrated to 75, 8/10 who were calibrated to 375, and 7/10 who were calibrated to 750 (binomial $p<.01$ for each calibration type). Of the 24 participants who were influenced by calibration, 21 demonstrated a significant interaction of Calibration and magnitude of set (binomial $p<.01$), indicating that calibration influenced estimation patterns differently as a function of magnitude. Specifically, participants were less influenced by misleading feedback for smaller magnitudes, and were more influenced for larger magnitudes.

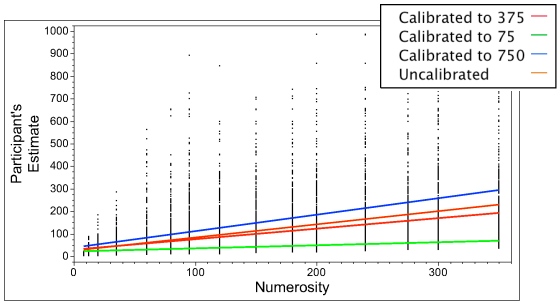


Figure 2: Estimation as a function of calibration

Next, we compared mean estimates for each presented numerosity in the Calibrated vs. Uncalibrated conditions. This isolated the magnitudes for which each participant[3] demonstrated the effects of miscalibration (Dunnett's mean, all $p<.05$). Overall, participants were influenced by calibration for relatively small magnitudes: one quarter of all participants provided different estimates for sets containing 8 items in the Calibrated vs. Uncalibrated conditions, and all but four participants were influenced by misleading feedback for at least one magnitude under 100. Participants were more resilient to misleading feedback for small magnitudes than large magnitudes, yet participants incorporated misleading feedback into the full range of estimates, and did not simply alter estimates for the largest sets presented.

Table 2: Smallest magnitude for which participants were influenced by calibration

Calibration	Mean	Median
75	35	61
375	36	42
750	20	45

[2] We also analyzed our data both excluding and including estimates of "75", "375", and "750" to ensure that any effect of calibration was not due to participants' repetition of the number word they had been miscalibrated to. There was no difference at either the group or individual level of analysis.

[3] Of the 24 who were influenced by calibration

Discussion

This study provides evidence for at least two mechanisms through which number words are mapped to approximate magnitudes. For the smallest numbers we tested, participants were accurate at matching arrays to number words, and they were resilient to misleading feedback. This suggests that, at least for relatively small and familiar magnitudes, adults may form associative mappings between number words and ANS representations. However, for larger magnitudes, adults struggled to correctly match number words with arrays, and were highly influenced by misleading feedback when estimating, making it unlikely that associative mappings play an important role in relating larger number words to magnitudes.

To our knowledge, this is the first study to demonstrate that adults do not directly link each number word in their count list to an ANS representation of approximately that magnitude. While past studies have shown that adults' estimates can be biased by misleading feedback (e.g., Izard & Dehaene, 2008; Shuman, unpublished thesis), in this previous research, the degree of miscalibration was minimal, and the sets that participants labeled as "thirty" before and after miscalibration were not discriminably different from each other using the ANS: this pattern of performance is consistent with either an associative mapping or structure mapping account. However, in the present study, participants failed to correctly match a number word to one of two discriminably different sets, and consistently provided different estimates of a set's magnitude when provided misleading information than when allowed to estimate without constraints. This suggests that adults do not possess associative mappings between large number words and magnitudes.

We posit a structural mapping hypothesis to account for the mappings adults make between large number words and magnitudes. By this hypothesis, adults recruit associatively mapped information about small numbers in order to map larger number words to magnitudes. We know that even large number words bear some relation to ANS representations of magnitudes, because adults' processing of verbal number words exhibit signatures typical of those found in perceptual judgments of numerosity using the ANS (e.g., Duncan & McFarland, 1980). Because of this, we suggest that structural mappings are constructed and supported by knowledge of associatively mapped magnitudes. This process may recruit more domain-general analogical or comparative mechanisms previously linked to the acquisition and extrapolation of spatial, numerical and categorical information during development (e.g., Gentner & Namy, 2006). As a result, structural links between number words and magnitudes may be based on analogy, proportional reasoning, or an understanding of the ordinality of both the verbal and nonverbal number systems. While each of these possible mechanisms for structure mapping is theoretically plausible and consistent with the current data, the present study cannot disambiguate between them. However, future research manipulating the availability and content of 'anchor' sets will allow us to construct a precise model of how small-number information is incorporated into structural mappings, by exploring how manipulating the availability of information about small quantities influences judgments about large quantities.

The present study also raises several developmental questions about the acquisition of number language. While much has been learned in recent years about the procedure of number word learning, little is known of the mechanisms that drive this learning. For example, we know that immediately after learning how counting represents sets, many children fail to map larger number words to larger sets in estimation tasks—however, by age 5, most children successfully provide larger estimates for larger magnitudes (Le Corre & Carey, 2008, Barth, Starr, & Sullivan, 2009). Clearly, 5-year olds have learned something about the count system that the 4-year olds have not—but what? The present study demonstrates that it is unlikely that these older children have improved at estimation solely because they have expanded their system of associative mappings between number words and magnitudes: even adults showed little to no evidence of any direct associative link between words like "one hundred" and sets of 100 things. Instead, children who are successful estimators must have learned something about the structural mapping of the count sequence onto magnitudes. What kind of structural relationships have these children learned?

Additionally, if children's structural mappings are constructed early in life, how much of the count sequence must be associatively mapped in order to support adult-like structural mappings? While the present study suggests that adults may have associative mappings for magnitudes as large as 20, we do not know which of these associative mappings are necessary to support a structure mapping. It may be the case that children need only to associatively map the smallest numbers (e.g., 1-10) in their count system in order to have enough information about the number system to develop a rich structural mapping between number words and magnitudes— by this theory, any additional associative mappings gained en route to adulthood (e.g., mappings between 10-20) are simply the result of additional experience with number words and magnitudes, and are not necessary to support structure mappings. However, it is also possible that children must have an adult-like system of associative mappings in order to support a structure mapping system. By this view, children would construct associative mappings between words and magnitudes for the numbers 1-20 *prior* to creating a structure mapping for larger number words. A continuation of the present line of research with children will help distinguish between these two possibilities, and in doing so, shed light on the nature of the inferences that children make about number words as they construct mappings between number language and magnitudes.

A better understanding of the roles of structural and associative mapping in the development of number knowledge may also help to illuminate other poorly

understood developmental phenomena in numerical cognition. For example, one signature of immature estimation ability is a tendency of young children to provide 'logarithmic' looking estimates: overestimating small numbers and underestimating large numbers. This tendency largely disappears in the 0-100 range by age 7, and in the 0-1000 range by age 9 (Booth & Siegler, 2006). What new information about the number system do these older children have? Several researchers have posited that the shift from immature-and-logarithmic to mature-and-linear patterns of estimation stems from a change in these children's underlying numerical representation (Siegler & Opfer, 2003; Booth & Siegler, 2006), or from an increased ability to reason about proportions (Barth & Paladino, 2010). However, the present study leads to an additional (though perhaps not mutually exclusive) hypothesis. Perhaps the shift towards more adult-like estimation can be best explained either by a realignment of structural mappings or to a refinement in the accuracy of the associative mappings that support these structure mappings. These two possible sources of the log-to-linear shift in estimation lead to distinct predictions of how both the younger and older children will reason about small and large numbers.

In conclusion, the present study failed to find evidence of associative links between most number words and magnitudes. Instead, the present study demonstrates that number word meaning is constructed through multiple mechanisms, and not necessarily through associations to real-world exemplars of their referents. By emphasizing the relative roles of associative and structure mappings, we hope to provide a new lens through which to view many of the developmental questions about number language acquisition, and in doing so, to open up new avenues for investigating the kinds of inferences we make about number words.

Acknowledgments

Thanks to Jennifer Audet, Amanda Chamberlain, David Huang, Michael Sim, and Tony Wang for help with data collection.

References

Barth, H., Kanwisher, N., & Spelke, E. (2003). The construction of large number representation in adults. *Cognition, 86*, 201-221.

Barth, H., Starr, A., & Sullivan, J. (2009). Children's mappings of large number words to numerosities. *Cognitive Development, 24*, 248-264.

Barth, H., & Paladino, A. (in press). The development of numerical estimation in children: evidence against a representational shift. *Developmental Science*.

Booth, J., & Siegler, R. (2006). Developmental and individual differences in pure numerical estimation. *Developmental Psychology*, *41*, 189-201.

Duncan, E. & McFarland, C. (1980). Isolating the effects of symbolic distance and semantic congruity in comparative judgments: an additive-factors analysis. *Memory and Cognition, 8,* 612-622.

Gentner, D., & Namy, L. (2006). Analogical Processes in Language Learning. *Current Directions in Psychological Science, 15,* 297-301.

Le Corre, M., & Carey, S. (2008). One, two, three, four, nothing more: An investigation of the conceptual sources of the verbal counting principles. *Cognition, 105*, 395-438.

Lipton, J., & Spelke, E. (2005). Preschool children's mapping of number words to nonsymbolic numerosities. *Child Development, 76*, 978-988.

Holloway, I., & Ansari, D. (2008). Mapping numerical magnitudes onto symbols: The numerical distance effect and individual differences in children's mathematics achievement. *Journal of Experimental Child Psychology, 103*, 17-29.

Izard, V., & Dehaene, S. (2008). Calibrating the mental number line. *Cognition, 106*, 1221-1247.

Mundy, E., & Gilmore, C. (2009). Children's mapping between symbolic and nonsymbolic representation of number. *Journal of Experimental Child Psychology, 103*, 490-502.

Piazza, M., Pinel, P., Le Bihan, D., & Dehaene, S. (2007). A magnitude code common to numerosities and number symbols in human intraparietal cortex. *Neuron, 53*, 293-305.

Shuman, M. (unpublished thesis). Computational characterization of numerosity perception and encoding.

Siegler, R., & Opfer, J. (2003). The development of numerical estimation: Evidence for multiple representations of numerical quantity. *Psychological Science, 14*, 237-243.

Thompson, C., & Opfer, J. (in press). How 15 hundred is like 15 cherries: Effect of progressive alignment on representational changes in numerical cognition. *Child Development*.

Verguts, T., & Fias, W. (2006). Priming reveals differential coding of symbolic and non-symbolic quantities. *Cognition, 105*, 380-394.

Whalen, J., Gallistel, C., & Gelman, R. (1999). Nonverbal counting in humans: the psychophysics of number representation. *Psychological Science, 10*, 130-137.

Wynn, K. (1992). Children's acquisition of number words and the counting system. *Cognitive Psychology, 42*, 220-251.

Analogue Magnitudes and Knower-Levels: Re-Visiting the Variability Argument

James Negen and Barbara W. Sarnecka ({jnegen,sarnecka}@uci.edu)
Department of Cognitive Sciences, 2201 Social & Behavioral Sciences Gateway Building
Irvine, CA, 92697-5100 USA

Abstract

What cognitive system(s) initially provide the numerical content that defines the cardinal number words for young children? Le Corre and Carey (2007) argued that the answer cannot be the analogue magnitude system. Here we re-examine the most powerful of their arguments, which concerned the system's signature *scalar variability* (the standard deviation of answers grows linearly with the mean). Using adult data, we explore a nuance of this signature: that while it is certainly true of the continuous, underlying activation in the brain, it may not always be true of the number-word responses that people produce. With this in mind, we re-examine the aforementioned variability argument; contrary to Le Corre and Carey, we conclude that young children's estimates of small set sizes (up to and including their number-knower-level) *do* show scalar variability.

Keywords: Cognitive Science, Psychology, Cognitive Development, Perception, Bayesian Modeling, Developmental Experimentation, Human Experimentation

How children acquire the meaning of number words is a question of great interest in cognitive development. In infancy, the child can create exact representations only of very small sets (up to 3 items) and can create only approximate representations of larger sets (for review, see Feigenson, Dehaene, & Spelke, 2004). It is from these foundations that the child must build an understanding of integers.

There are at least two families of theories about the role that the analogue magnitude number system plays in number-word learning. One says that (a) the numerical content of number words initially comes from the analogue magnitude system (i.e., the representations of numerosity are only approximate) and (b) number-word meanings are learned in the same way as any other words for antecedently available percepts (e.g., Dehaene, 1997; Gelman & Galistell, 1978; 2004). On this view, the child essentially understands the logic of numbers from the beginning, but must learn the *exactness of number words*. That is, the child must learn that "seven" means exactly 7 – not approximately 5 to 9.

The other family of theories contend that (a) the numerical content initially comes from enriched parallel individuation representations, (which are exact, but cannot go higher than 3 or 4) and (b) number-word meanings are learned one at a time, and in order (e.g., Carey, 2009; Le Corre & Carey, 2007; Sarnecka & Lee, 2009; Lee & Sarnecka 2010). On this view, the child figures out the principle of cardinality and the logic of the integer system by extrapolating from the first three or four exemplars—the words *one, two, three* and possibly *four*.

Our proposal takes something from both views. We consider it adequately demonstrated that children do learn the first few cardinal number-word-meanings one at a time, and in order (Carey, 2009; Le Corre, Van de Walle, Brannon & Carey, 2006; Lee & Sarnecka, 2010; Sarnecka & Lee, 2009; Wynn, 1990, 1992). Thus, in this paper we will use the knower-levels framework, wherein a *one-knower* is a child who just knows "one", a *two-knower* knows "one" and "two", etc., and a *CP-knower* understands how counting works and what it is used for. Though this view has historically been paired with the view that the numerical content of these words comes exclusively from enriched parallel-individuation representations, that linkage is not logically necessary.

This issue has been examined before by Le Corre and Carey (2007), who made a forceful argument against the analogue magnitude system being involved in number-word learning. The argument was based on variability signatures: Noise in analogue magnitude representations grows at the same rate as the number of objects being estimated. In other words, if you briefly show people 10 dots, and ask them how many are there, they will respond with variability σ. If you then show them 20 dots, they will respond with variability 2σ. More formally, the standard deviation of a person's responses, divided by the mean of their responses, will be a constant (the 'coefficeint of variance', hereafter COV). This *scalar variability* is a defining characteristic of the analogue magnitude system throughout the lifespan.

Le Corre and Carey (2007) argued that children's estimates of the number of items in a picture did not show scalar variability. Le Corre and Carey used a Give-N task (where children give a certain number of items to the experimenter) as well as a Cards task (where children quickly estimate the number of items on a card). Le Corre and Carey first identified children who were one-, two-, three-, or four-knowers as measured by the Give-N task. Then they analyzed the variability of children's estimates (on the Cards task) for sets of 1, 2, 3 and 4 items. The mean COV grew from 0 (for estimates of 1 item) to 0.4 (for estimates of 4 items), leading Le Corre and Carey to conclude that early on, the cardinal meanings of these words are not defined in terms of the analogue magnitude system.

However, there is a problem with this argument. Logically, scalar variability could be found in two places: in the underlying activations in the brain, or in the distribution of number-word responses. Most studies of the analogue magnitude system in adults treat these as the same. That is, they assume that number-word responses transparently

reveal the underlying activations. This is a reasonable assumption when the numbers involved are large. But with small numbers, the available words (i.e., *one*, *two*, *three* and *four*) may not be sufficient to express the variation in the underlying activations. Below, we describe these ideas in more detail, then support our argument with some new data from adults.

Scalar Variability of Activations versus Responses

Creating and using an analogue magnitude representation is a multi-step process. One place that scalar variability might be found is in the actual number-word responses that people produce. For example, if you show someone 10 and 20 dots in alternation, over and over again, and they say something like: 9, 22, 11, 18, … their responses could show scalar variability. That is, the standard deviation of the responses could grow linearly with the number of items to be estimated.

Another place you could look for scalar variability would be in a latent, real-valued variable – possibly the logarithm of activation in certain 'number neurons' – that describes the person's perception of how many dots are present. Thus, when shown 10 and 20 dots, the person may experience something like 8.82, 21.58, 11.21 and 18.44 units of activation. But of course, participants do not say that there are 11.21 dots on the screen – they just say "eleven"[1]. Scalar variability of the activations means that the standard deviation of this latent variable (as opposed to the responses) grows linearly with the number of dots shown.

For large numbers, these two ideas make almost exactly the same predictions. However, the predictions differ for smaller numbers. Imagine that a person says "four" when shown four dots, on a large number of trials. The variability in responses is zero, but the variability in activations could be much larger. The person might be experiencing anywhere from 3.5 to 4.5 units of activation on each trial. If we add the constraint that activation must be normally distributed, then the continuous standard deviation could be as large as 0.1 while the discrete standard deviation stays very near zero (out to several decimal places).

An analogy may be helpful: Imagine you are about to play a game with a friend. You will see a number of dots on a screen. Your friend has to say how many there are, but can't see the screen. You have to tell your friend how many dots there are, but you're not allowed to speak – only to draw a single line on a piece of paper. Your partner can then measure the line accurately with a ruler. Before the game starts, you agree on a simple code: 1 inch = 1 item. It's easy to see that the average amount of error in your line length will be proportional to the number of items.

Let's imagine that, when you see 2 items, all of your lines fall between 1.25 and 2.75 inches; you are accurate to ± .75 inches. Now cut that in half for when you see 1 item, and you are accurate to ± .375 inches. As such, your line lengths show scalar variability. But what about the answers your partner gives? There will be some variation at 2 (when she rounds 2.75 inches to 3, for example) but none at 1 (where all lines fall between .625 and 1.375 inches, and so are rounded to 1). The zero variation in her responses for trials of 1 cannot be half of the above-zero variation for trials of 2.

Thus, the length of your lines can show scalar variability while the number-word answers after measuring do not. This happens whenever the variation is small compared to the minimum distance between values after rounding. In more traditional statistical language, the variation is very fine-grained and the rounding is very coarse-grained, leading to the phenomenon of *heaping*: all of the data gets heaped onto one point.

Le Corre and Carey (2007) argued that one could expect to see scalar variability in the verbal responses all the way down to cards with 1 item. This claim was based on work by Cordes, Gelman, Galistell, and Whalen (2001). In that paper, the experimenters had instructed participants to tap a space bar a certain number of times in a steady rhythm, while saying "the" with each press to suppress counting. Cordes et al. found that participants' verbal responses showed scalar variability for numbers down to 2. However, they did not test on the number 1, and participants showed a relatively high amount of noise in their answers (a standard deviation of around .4 at minimum). As such, Cordes et al's data don't tell us where the variability was located: at such high levels of noise, the two sources of variability make much the same predictions.

The Present Study

In experiment 1, we examine the two sources of variability by extending the experiment by Cordes et al. (2001). Like them, we showed adult participants a numeral on a screen, and the participants' task was to tap the space bar that number of times. They had to do this in a steady rhythm while repeating "the" with each tap. We made one modification to reduce the amount of noise in the data: participants' responses were slowed to 2 Hz by having them tap along with a metronome.

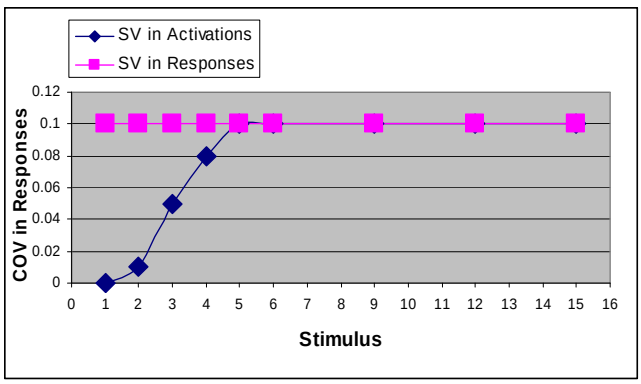


Figure 1: Different behavioral predictions for scalar variability in activations vs. responses.

[1] It may be more satisfying to imagine that this is the center of a confidence distribution. In other words, the participant would experience 11.21 units of activation, and be most confident in the integer 11 (which is closest), followed by 12, then 10, then 13, etc.

The two predictions from the two sources of scalar variability are illustrated in Figure 1. Conveniently, the pattern predicted by scalar variability in the responses is identical to the null hypothesis for a repeated-measures ANOVA.

In Experiment 1, we show that small-number estimation data from adults is consistent with the assumption of scalar variability in the activations. In experiment 2, we develop a simple model that incorporates number-knower-levels and the analogue magnitude system, and we test it using data from children's performance on the Give-N and Cards tasks. We find that the model provides a good fit to these child data, with no problems of variability.

Experiment 1

Methods

Participants Eleven undergraduates were recruited from the University of California, Irvine and successfully completed the task. Two additional participants were excluded because they could not keep a steady rhythm during training trials. Participants were each given a half point of extra credit in an introductory Psychology course.

Procedure On each trial, participants saw a numeral on the screen. They were asked to tap the space bar that many times, along with a 2Hz metronome, saying "the" with each tap to suppress counting. Participants were asked to avoid trying alternate strategies, like chunking, and to just keep tapping until they felt the number had been reached. The numerals used were 1, 2, 3, 4, 5, 6, 9, 12, and 15. For training, there were two trials with each number, and all participants declined the offer to go through training again. For data recording, there were 40 trials with the number 1, and 20 trials with each of the other numbers, for a total of 200 trials. Participants were allowed as many brief breaks as they wanted, though none took more than two. All participants finished the testing session in under 20 minutes.

Results and Discussion

All but two participants were able to keep a steady rhythm throughout the experiment (the two who couldn't were excluded). The mean coefficient of variance (COV) for each stimulus is shown in Figure 2. A repeated-measures ANOVA with the Greenhouse-Geisser correction suggests that the mean COV across stimuli is not constant, $F(2.8, 28.2) = 8.16$, $p = .001$. This discredits the idea of scalar variability in responses: if such variability were true of adult performance, then we should expect the mean COV in responses to be the same for all the numbers we tested.

On the other hand, these data are very consistent with the assumption of scalar variability of activations, with a correlation between prediction and observation of over $r = .8$. There appears to be a section where the rounding-off occurs (numbers 1 to 3), and a section with full observed variance (numbers 5 to 15), as predicted. It's a little surprising that there was so much variability in responses to 1, but this variability comes from only 2 of the participants, representing 3 errors out of 440 trials, and still appears to be lower than the section with full observed variance. It is clear that these data are better explained by scalar variability in the activations than in the responses.

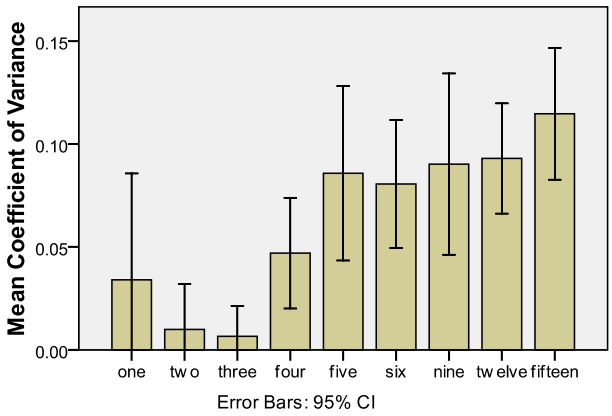


Figure 2: Observed COVs (standard deviation divided by mean) in the responses were lower on average for lower numbers of requested taps, consistent with scalar variability in the activations (but not in the responses).

The amount of variability at numbers 1 to 4 suggests that we were successful in repressing subitization, replicating Cordes et al., 2001. Also, subitizing is usually characterized by a hard limit at 4, but the lowest number where each participant made a mistake ranged from 1 to 6. In addition, if the effect was due to subitizing, we would be able to eliminate it by subtracting out all of the jumps from 4 to 5 (across the limit of what people can subitize). However, this is not the case, $F(7,4) = 7.546$, $p = .035$. Finally, all participants reported that they found it impossible to track individual taps during the task. (Tracking individual taps would show that participants were using the parallel individuation system). All of these facts point away from subitizing as an explanation.

Next, we test these predictions against a corpus of child behavioral data, to re-examine the question of whether the analogue magnitude number system may be involved in early number-word learning.

Experiment 2

We made a simple model of how children would solve the Cards task using the analogue magnitude system. In this task, children are shown a certain number of items on a card and are asked to guess the number of items. This is the same as Le Corre and Carey's (2007) Cards task, but without the time limit. (The purpose of the time limit is to prevent counting, but children who do not yet understand the cardinality principle do not use counting to answer anyway.)

This is how the model works: the child looks at the card once and forms an analogue magnitude representation of the number of items present. This representation could be measured in units of activation. The child then rounds off

this activation level to the nearest positive (non-zero) whole number. If the child knows this number word, she says it. If the child does not know the word for that number, she says a number from among the number words whose meanings she does not know. (For a discussion of how the child chooses *which* undefined number to guess, see Sarnecka & Lee, 2009; Lee & Sarnecka 2010.)

For example, suppose we show a two-knower 2 items on several trials. On the first, she experiences 2.3 units of activation, rounds it to 2, and says "two". On the second, she experiences .38 units, rounds it to 1, and says "one". On the third, she experiences 3.3 units, rounds to 3, and picks "six" from among the words that she does not know ("three" and up). These activation levels are being drawn from a normal distribution with a mean of 2.

The computations involved are based on the normal cumulative density function. Say that this child has a COV of α when shown a card with 2 items on it. The chances of saying "one" are $\Phi((1.5-2)/\alpha)$. The chances of saying "two" are $\Phi((2.5-2)/\alpha)-\Phi((1.5-2)/\alpha)$. The chances of saying "three" and up (beyond the knower-level) are proportional to $1-\Phi((2.5-2)/\alpha)$. The lower α is, the more often the child will be correct.

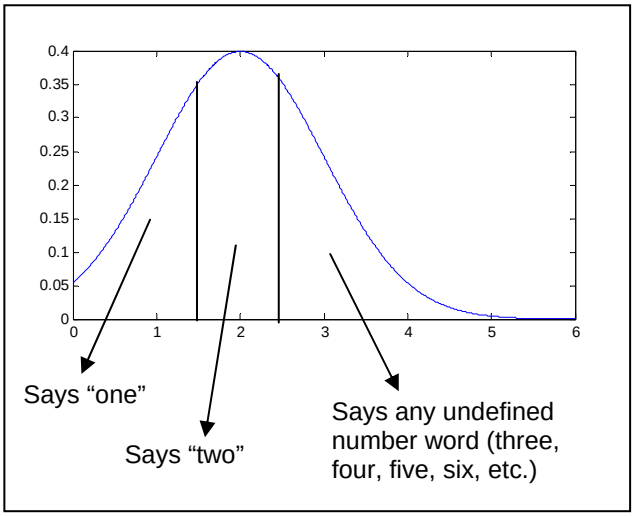


Figure 3: An example of the model: a two-knower, shown 2 items, draws an activation level from a normal distribution N(2,1) and responds accordingly.

Furthermore, we assume that for each stimulus γ (each number of items shown on the card, up to 4) and knower-level τ, a particular child's COV is drawn from a normal distribution with mean $\mu_{\gamma,\tau}$ and standard deviation σ. A variable δ describes the constant difference between each μ_γ and the next one up. So $\mu_{2,2} = \mu_{1,2} + \delta$, etc. We planned to use Bayesian inference on these latent variables: if $\delta \neq 0$, then this method of describing variation does not show scalar variability – i.e., a constant COV – and our model is wrong.

It may bother the reader that this model allows for the analogue magnitude system to output a negative number. However, there is no empirical argument against this– assuming the participant knows to round to the nearest positive whole number. A log-normal distribution may be substituted, which would limit the outputs to positive numbers, but this generally has little effect; having low outputs rounded to 1 accomplishes the same basic goal as making the distribution of outputs stay above zero. Since a normal distribution is symmetric, we assume that the variation to the higher side is sufficient to estimate total variation, e.g., the variation around 1 item is well-represented by the number of times the child says "two" or "three". Put another way, a two-knower has an equal probability of saying "two" as experiencing between -.5 and .5 units of activation, when shown 1 item. In any case, the probability of negative activation is often very small.

Methods

These data were taken from a longitudinal study testing children on both the Give-N task and the Cards task. The dataset thus offers a way of independently assessing the child's knower-level (using Give-N) and then estimating δ (using the Cards task).

Participants A total of 97 monolingual English-speaking children participated. Children were tested once every two weeks for twenty weeks, resulting in a total of 454 sessions. Because little is known about the week-to-week consistency of children's knower-levels and/or analogue magnitude acuity, each session was modeled as a new child. We included all children who were two-, three-, or four-knowers (as determined by Give-N), and who completed at least 15 trials of both the Give-N and the Cards task. This resulted in a total of 161 sessions (45 sessions with two-knowers, 51 with three-knowers, and 65 with four-knowers).

Give-N The purpose of this task was to determine what number-word meanings each child knew (i.e., to determine the child's number-knower-level.) The experimenter began the game by bringing out a stuffed animal (e.g., a lion), a plate, and 3 bowls, each containing 15 small identical rubber toys (e.g., toy bananas, approx. 3 cm long). The experimenter said to the child, "In this game, you're going to give something to the lion, like this [experimenter pantomimes putting an item on the plate and sliding it over to the lion]. I'm going to tell you what to give him." Instructions were of the form, "Can you give the lion TWO bananas?"

Trials were in pseudorandom order, always starting with a request for one item. There were a total of eighteen trials: six trials each asking for one, two, three, four, six, and eight items. Children were given generalized positive feedback after each trial (e.g., "Thank you!"), regardless of their responses.

The child's knower-level was estimated using Lee & Sarnecka's (2010) model. In this model, the child has a prior probability of giving each number of objects: small handfuls are very likely, as is giving the entire bucket. When the

child is asked for a specific number, the response probabilities are updated according to her knower-level. Every child is given a uniform prior chance to be each knower-level, and a posterior distribution is calculated based on the data. We sampled over this posterior distribution with 3 independent chains, 2,000 burn-in samples and 10,000 data collection samples. We then assigned each child's knower-level based on the expected value[2] from this posterior distribution.

Cards Task This was our comparison task, which provides a set of responses independent from Give-N, so that one can be used to determine knower-level and the other to model the analogue magnitude response. The stimuli were photographs of the same stimuli used in Give-N, on a white background. The difference is that children were asked to label the numerosity instead of generating it. For example, if a child was asked for three red trains in Give-N, she was later shown a picture of three red trains and asked the number on it. As a check that the child was paying attention, trials were thrown out if the child did not produce the correct object name (e.g., trains) along with the number word. Before the first trial, there was a check that the children knew the words for the objects being used, which did not seem to pose a problem for our participants.

Again there were eighteen trials: three trials each for one, two, three, four, six and eight items, in pseudorandom order. We asked children the question "What's on this card?" because questions that start with "How many …" are often interpreted as commands to count (Sarnecka & Carey, 2008). Children were given generalized positive feedback after every trial. Order of tasks was counterbalanced across sessions.

Results and Discussion

There were a total of 1,903 usable trials from the Cards Task. We decided to give the model a unit normal prior on effect size δ/σ, as a way to make the prior dimensionless and reasonably vague (Jeffreys, 1961; Rouder, Speckman, Sun, Morey & Iverson, 2009; Wetzels, Raaijmakers, Jakab, & Wagenmakers, in press). We then ran the model with MATLAB and WinBUGS, which uses a form of Gibbs sampling to describe the posterior distribution. We ran five independent chains, with 10,000 burn-in samples and 25,000 collection samples. The within-chain variability matched the between-chain variability, even though the chains were initialized with different effect sizes, which is a good indication of proper convergence.

At the point of the null hypothesis (δ = 0; i.e., the average COV is the same for every number of items shown), there was a prior density of .39. At the same point, there was a posterior density of 5.72, estimated with a normal kernel density method. By the Savage-Dickey Theorem (Dickey & Lientz, 1970), the Bayes factor is 5.72/.39 = 14.34, meaning that the data were 14.34 times more likely to be generated by the null hypothesis than the alternate. This is very strong evidence in support of the null hypothesis. In other words, Children *do* show scalar variability in at least one task that taps number-word knowledge – but it is scalar variability of activations, not of responses.

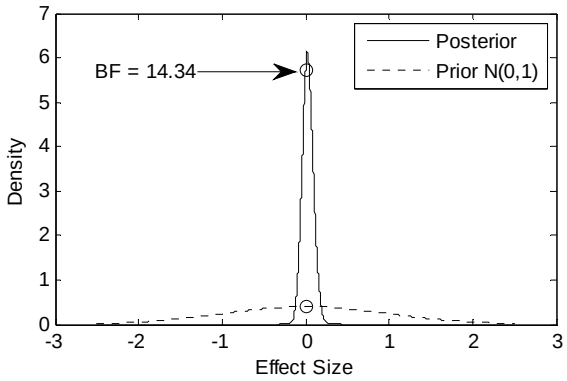


Figure 4: Prior and posterior distributions of δ/σ. There is a great deal of support for the null hypothesis (δ = 0; i.e., the mean COV for every stimulus is the same within each knower-level) in the child data.

General Discussion

In Experiment 1, we showed that scalar variability of adult small-number estimates may be located in the activations, but not in the responses. Our analysis suggests that scalar variability is not always present in the answers that people give when you ask them to estimate a number of items. Instead, scalar variability is present in the latent continuous variable that serves as input, and which gets rounded off in people's answers. In Experiment 2, we modeled how children respond when asked how many items are on a card. Our model included (a) the use of analogue magnitudes assuming scalar variability in the activations and (b) knower-levels, which are stages of number-word knowledge (e.g., Sarnecka & Lee, 2009). We showed that, contrary to previous reports, there is no problem fitting the well-known signature of scalar variability to real child data. In fact, we found strong evidence that children's responses *do* show scalar variability, with a Bayes factor of 14.34.

This reverses one of the most powerful arguments against the involvement of the analogue magnitude system in early number-word learning. Scalar variability is a very reliable feature of the analogue magnitude system. We have presented evidence that this signature is found in children's verbal responses to a small-number estimation task, even before the children have figured out the cardinal principle of counting. The statistical method we used has a very useful feature: it can quantify support for the null hypothesis, rather than simply rejecting or failing to reject it. Thus, we can report positive evidence that children's responses do show scalar variability. This suggests that children do imbue the first few number words (up to and including their

[2] Another reasonable method would be to take the mode instead of the expected value. This method has high agreement in this case, r = .992.

knower-level) with numerical content from the analogue-magnitude system.

Of course, evidence for the involvement of analogue magnitudes is not evidence against the involvement of enriched parallel individuation; it's possible that a parallel-individuation-based model could also be fit sensibly to these data. There are still many reasons to think that parallel individuation is involved. In particular, we do not dispute Le Corre and Carey's finding that new CP-knowers do not connect numbers above 5 to the analogue magnitude system (Le Corre & Carey, 2007; see Sarnecka & Lee, 2009 for a convergent finding). In addition, children seem to become three-knowers or four-knowers, but not five-knowers or six-knowers. The limit after 4 coincides with the set-size limit on the parallel individuation system (e.g., Feigenson & Carey, 2003).

Moving forward, we think tasks that tap more than just the child's number-word knowledge should be part of the debate. For example, there is new interest in the question of whether number-word knowledge is related to estimation acuity, after it was shown that acuity at age 14 retroactively predicts math grades back to first grade (Halberda, Mazzocco & Feigenson, 2008). In particular, the present model predicts that estimation acuity should correlate with the number of within-knower-level errors on the Cards task. This kind of argument could be very useful in determining which systems are important to number-word learning and how they contribute.

We hope that this line of work will lead to increasingly accurate descriptions of how children acquire integers, and will help to, resolve debates over the roles of various cognitive systems in number-word learning. The test case of number may then inform more general theories of how new representational resources are acquired.

Acknowledgments

This research was supported by NIH grant R03HD054654 to Barbara W. Sarnecka. We thank the children and familiews who participated in the study, as well as the teachers and administrators of the following Irvine, California preschools: Jenny Hart Early Education Center, Temple Bat Yahm Preschool, Turtle Rock Preschool, UCI Children's Center, UCI Early Childhood Education Center, UCI Infant and Toddler Center, and University Montessori.

References

Carey, S. (2009). *The Origin of Concepts*. Boston: MIT Press.

Cordes, S., Gelman, R., Gallistel, C. R., & Whalen, J. (2001). Variability signatures distinguish verbal and nonverbal counting for both large and small numbers. *Psychonomic Bulletin & Review, 8*(4), 698-707.

Dehaene, S. (1997). *The Number Sense*. New York: Oxford University Press.

Dickey, J. M., & Lientz, B. P. (1970). The weighted likelihood ratio, sharp hypotheses about chances, the order of a Markov chain. *The Annals of Mathematical Statistics, 42,* 204-223.

Feigenson, L., & Carey, S. (2003). Tracking individuals via object files: Evidence from infants' manual search. *Developmental Science, 6*(5), 568-584.

Feigenson, L., Dehaene, S. & Spelke, E. (2004). Core systems of number. *Trends in Cognitive Sciences, 8*(7), 307-314.

Gelman, R. & Gallistel, C. R. (1978). *The Child's Understanding of Number*. Oxford: Harvard University Press.

Gelman, R. & Gallistel, C. R. (2004) Language and the origin of numerical concepts. *Science, 306*(5695), 441-443.

Halberda, J., Mazzocco, M.M.M., & Feigenson, L. (2008). Individual differences in non-verbal number acuity correlate with maths achievement. *Nature, 455*(2), 665-669.

Jeffreys, H. (1961). *Theory of probability*. Oxford, UK: Oxford University Press.

Le Corre, M., Van de Walle, G., Brannon, E. M., Carey, S. (2006). Re-visiting the competence/performance debate in the acquisition of the counting principles. *Cognitive Psychology, 52*(2), 130-169.

Le Corre, M. & Carey, S. (2007). One, two, three, four, nothing more: An investigation of the conceptual sources of the verbal counting principles. *Cognition, 105,* 395-438.

Lee, M.D., & Sarnecka, B.W. (2010). A model of knower-level behavior in number-concept development. *Cognitive Science, 34,* 51-67

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t-tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*, 225-237.

Sarnecka, B. W. & Carey, S. (2008). How counting represents number: What children must learn and when they learn it. *Cognition*, *108,* 662-674.

Sarnecka, B. W. & Lee, M. D. (2009). Levels of Number Knowledge in Early Childhood. *Journal of Experimental Child Psychology, 103*(3), 325-337.

Wetzels, R., Raaijmakers, J. G. W., Jakab, E., & Wagenmakers, E.-J. (in press). How to quantify support for and against the null hypothesis: A fexible WinBUGS implementation of a default Bayesian t-test. *Psychonomic Bulletin & Review*.

Wynn, K. (1990). Children's understanding of counting. *Cognition, 36*(2), 155-193.

Wynn, K. (1992). Children's acquisition of number words and the counting system. *Cognitive Psychology, 24*(2), 220-251.

Integrating Reinforcement Learning with Models of Representation Learning

Matt Jones (mcj@colorado.edu) & Fabián Cañas (canas@colorado.edu)
University of Colorado, Department of Psychology & Neuroscience
Boulder, CO 80309 USA

Abstract

Reinforcement learning (RL) shows great promise as a model of learning in complex, dynamic tasks, for both humans and artificial systems. However, the effectiveness of RL models depends strongly on the choice of state representation, because this determines how knowledge is generalized among states. We introduce a framework for integrating psychological mechanisms of representation learning that allows RL to autonomously adapt its representation to suit its needs and thereby speed learning. One such model is formalized, based on learned selective attention among stimulus dimensions. The model significantly outperforms standard RL models and provides a good fit to human data.

Keywords: Reinforcement learning; attention; generalization

Introduction

Most challenging tasks people face are inherently dynamic and interactive. Choices affect not just immediate outcomes but also future events, and hence subsequent decisions that must be made. Normative and descriptive theories of learning in dynamic environments have advanced dramatically in recent years with the development of Reinforcement Learning (RL), a mathematical and computational theory drawing on machine learning, psychology, and neuroscience (e.g., Schultz, Dayan, & Montague, 1997; Sutton & Barto, 1998).

However, RL currently faces a fundamental challenge relating to the issue of knowledge representation. Dynamic tasks tend to be highly complex, with an enormous number of possible states (situations) that can arise. Therefore, efficient learning must rely on generalization from past states that are similar to the current one. Similarity, in turn, depends on how states are represented, including the features by which they are encoded and the relative attention allocated to those features (Medin, Goldstone, & Gentner, 1993). Thus representation is critical to the effectiveness of RL algorithms, because representation determines the pattern of generalization by which past experience is used to make new decisions.

Although there has been little research on representation in the context of RL, representation and representation learning have long been topics of psychological study. Empirical research in a number of domains, including perceptual learning, attention, categorization, object recognition, and analogy, has uncovered principles and mechanisms by which people learn to modify how they encode objects and situations in the service of learning, inference, and decision making. Here we describe a framework for a natural synthesis of these ideas with RL algorithms, which leads to models that learn representations for dynamic tasks. A specific model is presented that is based on principles of attention learning from the categorization literature (Kruschke, 1992; Nosofsky, 1986). Two sets of simulation studies are reported, which demonstrate both the power and the psychological validity of this approach.

Reinforcement Learning

RL comprises a family of algorithms for learning optimal action in dynamic environments. RL models characterize a task as a Markov Decision Process, in which the environment at any moment exists in one of a set of *states*, each associated with a set of actions available to the learner. The chosen action determines both the immediate reward received, if any, and the state of the environment on the next time step. This general framework encompasses most tasks of psychological interest (Sutton & Barto, 1998).

The key insight behind most RL algorithms is to learn a *value* for each possible state or action. This value represents the total future reward that can be expected starting from that point. Formally, given any state s and action a, the state-action value is defined as

$$Q(s,a) = E\left[\sum_{k\geq 0}\gamma^k r_{t+k} \middle| s_t = s, a_t = a\right], \quad (1)$$

where t is the current timestep; s, a, and r are the state, action, and received reward on each step; and $\gamma \in [0, 1]$ is a discount factor representing the relative value of immediate versus delayed rewards. This approach allows action selection to be based directly on the Q-values. Here we use a Luce-choice or Gibbs-sampling rule, with inverse-temperature parameter θ.

$$P(a_t = a) \propto e^{\theta \cdot Q(s_t,a)} \quad (2)$$

Once an action is selected, its value is updated according to the immediate reward and the values associated with the state that follows. One of the best-studied algorithms for learning action values, Q-learning (Watkins & Dayan, 1992), uses the update rule

$$\Delta Q(s_t,a_t) = \varepsilon_{\text{val}} \cdot \delta, \quad (3)$$

where $\varepsilon_{\text{val}} \in (0, 1]$ is a learning rate, and δ is the *temporal-difference (TD) error*, defined as

$$\delta = r_t + \gamma \cdot \max_a \{Q(s_{t+1},a)\} - Q(s_t,a_t). \quad (4)$$

The TD error represents the difference between the original action-value estimate, $Q(s_t, a_t)$, and a new estimate based on the immediate reward and ensuing state. The expression $\max_a\{Q(s_{t+1}, a)\}$ represents an estimate of the value of the new state, s_{t+1}, based on the best action that could be performed in that state.

The simplest implementations of *Q*-learning and other RL algorithms use a *tabular* (i.e., lookup-table) representation, in which a different set of action values is learned separately for every possible state that can occur. Tabular representations are impractical for most realistic tasks, because the number of states grows exponentially with the number of state variables. Therefore, most implementations of RL utilize some form of *generalization*, whereby knowledge about one state is extended to other, similar states. This approach dramatically speeds learning by reducing the amount of information that must be retained and updated, and by allowing the learner to draw on a richer set of past experiences when making each new decision.

Central to the success of generalization in all learning tasks (not just RL) is the choice of representation. In order for generalization to be effective, states (or stimuli in general) must be encoded so that stimuli that are perceived or treated as similar tend to be associated with similar outcomes or appropriate actions (Shepard, 1987). Such a representation facilitates generalization, and hence learning, because it leads the learner to draw on precisely those past experiences that are most relevant to the current situation.

Unfortunately, the choice of representation is a notoriously difficult problem, and the field of machine learning is far from having automated algorithms that discover useful representations for learning novel tasks. Successful applications of RL have instead tended to rely on hand-coded human knowledge for encoding states. For example, the state representation in Tesauro's (1995) celebrated backgammon program, TD-gammon, was based on complex features (configurations of pieces) suggested by expert human players. Likewise, psychological research in RL generally avoids the problem of representation by using small sets of stimuli with clearly defined features, so that the subject's representation can be confidently assumed by the modeler and is unlikely to change during the course of learning (e.g., Fu & Anderson, 2006). Arguably, representation is where the real challenge often lies, and therefore starting a model with a hand-coded representation, or using experimental stimuli with unambiguous features, presupposes the most difficult and interesting aspects of learning (Schyns, Goldstone, & Thibaut, 1998).

Selective Attention in Category Learning

One behavioral domain in which generalization and representation have been extensively studied is category learning. Much of the research on category learning has aimed to understand the internal representations that humans develop to facilitate classification of objects and inference of unobserved features. All of these models serve, in one way or another, to allow generalization of category knowledge from previously encountered to novel stimuli.

The most direct mechanism for generalization in categorization is embodied by exemplar models (Medin & Schaffer, 1978). In these models, the psychological evidence (E) in favor of classifying a stimulus (s) into a given category (c) is given by summing its similarity to all previously encountered exemplars (s'), weighting each exemplar by its association to c.

$$E(s,c) = \sum_{s'} sim(s,s') \cdot w(s',c) \qquad (5)$$

The property of exemplar models most relevant to the current investigation is the similarity function. Rather than being fixed, a large body of evidence indicates that similarity changes during the course of learning as a consequence of shifts of attention among the stimulus dimensions (e.g, Nosofsky, 1986). This flexibility is modeled by expressing similarity as a decreasing function of distance in psychological space, with each stimulus dimension, i, scaled by an attention weight, α_i (Nosofsky, 1986). Here we assume an exponential similarity-distance function, in accord with empirical evidence and normative Bayesian analysis (Shepard, 1987).

$$sim(s,s') = e^{-\sum_i \alpha_i |s_i - s'_i|} \qquad (6)$$

The effect of attention on similarity is to alter the pattern of generalization between stimuli so as to fall off more rapidly with differences along dimensions with greater attention weights. When stimuli differ only on unattended dimensions, their differences are unnoticed and hence generalization between them is strong. This adaptation of generalization leads to improved performance when attention is shifted to task-relevant dimensions, because the learner generalizes between stimuli that have common outcomes while discriminating between stimuli that are meaningfully different.

The influence of attention on generalization has extensive support, both theoretically (Medin et al., 1993) and empirically (Jones, Maddox, & Love, 2005; Nosofsky, 1986). An important question suggested by this research is how attention can be learned. One proposal is that attention weights are updated in response to prediction error (Kruschke, 1992). In a classification task, prediction error (δ) is simply the difference between the category evidence, $E(s, c)$, and the actual category membership given as feedback to the learner (e.g., +1 if $s \in c$ and -1 otherwise). The updating rule for attention is then based on gradient descent on this error, squared and summed over categories.

$$\Delta\alpha_i = -\varepsilon_{\text{att}} \cdot \frac{\partial}{\partial \alpha_i} \left(\frac{1}{2} \sum_c \delta_c^2 \right). \qquad (7)$$

This mechanism for attention learning has been implemented in ALCOVE, a highly successful model of human category learning (Kruschke, 1992). ALCOVE learns to shift attention to stimulus dimensions that are most relevant to predicting category membership and away from dimensions that are non-diagnostic. This leads to adaptation of generalization, which in turn speeds learning.

Attention Learning in RL

Because of the strong empirical support for attention learning in the categorization literature, we believe it is a potentially fruitful topic for study in the context of RL.

Selective attention may be especially relevant in this domain because most interesting RL tasks have complex state spaces of high dimensionality, and learning to distinguish relevant from irrelevant dimensions should be expected to greatly speed learning in such tasks.

The present investigation addresses two questions regarding the relationship between attention learning and RL. The first question is a psychological one, of whether attention learning as observed in supervised tasks such as categorization also operates in the dynamic tasks modeled by RL. This extrapolation is not trivial, because RL relies on TD error, which is an internally generated signal based in part on the learner's own value estimate of the ensuing state (see Eq. 4). It is an empirical question whether this internal signal can drive attention shifts and other forms of representation learning in the same way that external feedback does. A companion paper (Cañas & Jones, 2010) reports a behavioral experiment that supports an affirmative answer to this question, and the data from that experiment are modeled below.

The second question is a computational one, of whether the formal equations that describe RL and attention learning can be coherently integrated, and whether the resulting model will exhibit efficient learning. This normative question is important psychologically because computational power constitutes a significant motivation for expecting attention learning to play a role in human RL. If the two are computationally compatible, then the potential significance of RL is greatly increased, in that RL is capable of autonomously adapting the representations on which it operates.

Comparison of the equations describing Q-learning and attention learning reveals there is indeed a natural, highly complementary integration. The strength of Q-learning, and RL algorithms in general, is in the sophisticated updating signals they compute, which take into account both external reward and internal consistency of value estimates (Eq. 4). The updating itself is fairly trivial, consisting of adjusting the existing estimate by a proportion of the error (Eq. 3). Attention learning, and models of category learning more generally, have the opposite character. Their updating signals are fairly trivial (prediction error relative to external feedback), but the updates themselves are complex, driving adaptation of sophisticated internal representations. This complementary relationship suggests the solution of using the TD error signal from RL to drive representation learning, and in particular to update attention weights.

We refer to the model resulting from this integration as Q-ALCOVE. Q-ALCOVE estimates action values via similarity-based generalization among states, directly analogous to ALCOVE (Eq. 5).

$$Q(s,a) = \sum_{s'} sim(s,s') \cdot w(s',a) \tag{8}$$

The Q-values are used to generate action probabilities according to the response-selection rule used by both Q-learning and ALCOVE (Eq. 2). The w parameters, which act as pre-generalization action values, are updated analogously to both Q-learning (Eq. 3) and ALCOVE, using the same TD error signal as in Q-learning (Eq. 4).

$$\Delta w(s_t, a_t) = \varepsilon_{\text{val}} \cdot \delta \tag{9}$$

Similarity between states in Q-ALCOVE is defined identically to stimulus similarity in ALCOVE (Eq. 6), except that a normalization term is included that fixes the total similarity (i.e., the integral of the generalization gradient) to 1. We have found that attention learning in tasks requiring continuous prediction only functions well when normalization is included.

Learning of attention weights follows the same rule as in ALCOVE, except for the critical substitution of classification error with TD error. In addition, we only differentiate δ with respect to $Q(s_t, a_t)$ and not $Q(s_{t+1}, a)$, because the motivation behind Q-learning is to use $Q(s_{t+1}, \cdot)$ to adjust $Q(s_t, \cdot)$. Nevertheless, changing α also affects $Q(s_{t+1}, \cdot)$, and further analytical work is needed to understand the impacts of this fact on model behavior and predictions. The resulting rule for attention learning is

$$\Delta\alpha_i = \varepsilon_{\text{att}} \cdot \delta \cdot \frac{\partial}{\partial \alpha_i} Q(s_t, a_t) \tag{10}$$

The intuition behind attention learning in Q-ALCOVE is that, after feedback, the model adjusts attention weights to reduce generalization from states that contributed to error and to increase generalization from states that suggested more correct predictions. Over the course of experience, attention should shift to those dimensions that are most diagnostic of correct actions and their values.

Simulation Studies

Two sets of simulations were carried out to evaluate the behavior of Q-ALCOVE. The first set was based on Gridworld, a common benchmark task for RL models. These simulations aimed to test whether the attention-learning mechanism in Q-ALCOVE operates as predicted, to shift attention toward relevant dimensions and away from irrelevant dimensions. If so, a second question was whether selective attention leads to significant improvements in learning speed, and how such a potential advantage depends on the dimensionality of the task. The second set of simulations was based on a human behavioral experiment (Cañas & Jones, 2010) designed to test whether humans can learn selective attention using internal value estimates (i.e., TD error) as feedback, as proposed here. These simulations aimed to evaluate Q-ALCOVE's viability as a psychological model.

Directional Gridworld

Gridworld is a class of tasks with a long tradition as a benchmark for RL algorithms (e.g., Sutton & Barto, 1998). The states of a Gridworld task form a rectangular lattice of dimensionality D. We call the present task Directional Gridworld, because it was set up in such a way that one dimension was relevant and the others were irrelevant.

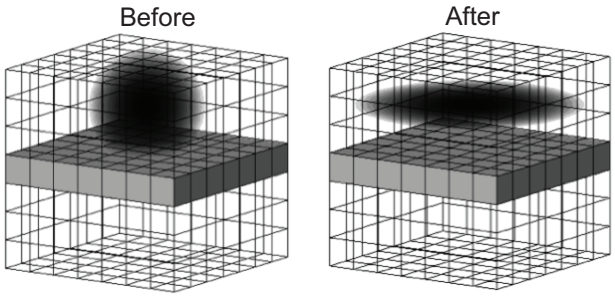


Figure 1. State space for 3-dimensional Directional Gridworld task. Grey states are goal states. Black cloud depicts Q-ALCOVE's generalization gradient, at the start of learning (left) and after 300 time steps (right).

Figure 1 illustrates the Directional Gridworld task for the case of $D = 3$ (the generalization gradients in the figure are discussed below). Each dimension has 7 levels, for a total of 7^D states. In each state, the learner has $2D$ available actions, corresponding to motion in either direction along any dimension. For simplicity, we assume that actions are deterministic and move the learner by 1 step in the chosen direction. Actions on the boundaries that would take the learner outside the space have no effect.

States are encoded as vectors corresponding to their values on the D dimensions. Other than this, the model has no prior knowledge of the topology of the environment or of the meanings or effects of actions. The spatial interpretation is only a convenient metaphor, and the task is not meant as a model of spatial navigation that might involve specialized psychological mechanisms. The stricter interpretation is as an abstract problem space (e.g., Newell & Simon, 1972).

The highlighted states (Fig. 1) spanning the center of the space are *goal states*. Whenever the learner reaches a goal state, a reward of 10 is provided. On the next step, the learner is taken to a random state maximally distant from the goal region. All actions that do not lead to a goal produce a reward of -1. The learner's task is to choose actions so as to maximize total temporally discounted reward (Eq. 1, with γ set to .5). Thus, optimal behavior consists of repeatedly moving in a straight line from the boundary to the nearest goal state.

For all values of the dimensionality D, the goal states form a hyperplane through the center of the space. The dimension perpendicular to the goal region is relevant to optimal action selection, as the learner needs to move in opposite directions depending on which side of the goal region it is on. All other dimensions are irrelevant. Indeed, it can easily be shown that the optimal Q-values for any state depend only on the state's position on the relevant dimension. Therefore, the most efficient generalization strategy for learning Q-values is to average over all states at each level of the relevant dimension but to learn separate values for each level. This strategy can be achieved by strong attention to the relevant dimension and zero attention to all other dimensions. A primary question was whether Q-ALCOVE would learn such an attention distribution.

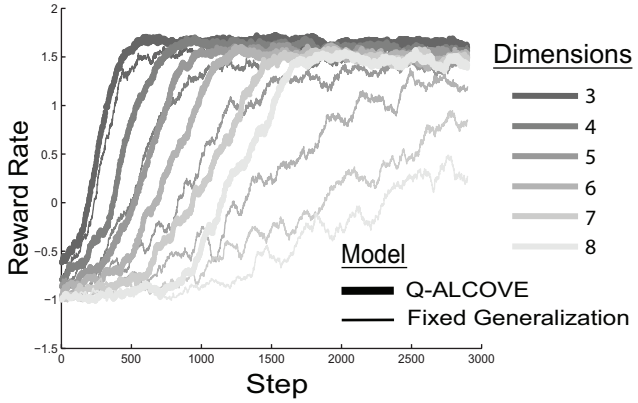


Figure 2. Learning curves in Directional Gridworld for Q-ALCOVE and version with fixed generalization.

Two models were simulated in addition to Q-ALCOVE. The first was tabular Q-learning, which learns actions values independently for all states. The second was a fixed-generalization model obtained from Q-ALCOVE by setting the attention-learning rate, ε_{att}, to 0. Q-ALCOVE was run using $\varepsilon_{att} = .01$. All models were run with value-learning rate $\varepsilon_{val} = 1$ and choice parameter $\theta = .5$. The models' value estimates (w or Q) were initialized at 0 at the start of each run. The initial values for attention weights were set to .4 for both Q-ALCOVE and the fixed-generalization model. This value was chosen so as to maximize performance of the fixed-generalization model on 3 dimensions.

Figure 2 shows average learning curves for Q-ALCOVE and the fixed-generalization model for Directional Gridworlds of 3 to 8 dimensions. Performance for tabular Q-learning was poor enough, especially at higher dimensionalities, that it is omitted. Each curve indicates reward rate, smoothed with a rectangular window of 100 time steps, and averaged over 4 separate runs of the model. As can be seen, Q-ALCOVE learns more quickly with attention learning than without, and the magnitude of this advantage grows rapidly with the number of dimensions. This result suggests that attention plays an indispensable role in natural tasks of much higher dimensionality.

Figures 1 and 3 illustrate how Q-ALCOVE's attention-learning mechanism facilitates learning, in the case of three dimensions. Figure 3 shows the attention weights for a single run, which increase for the relevant dimension and decrease toward 0 for the irrelevant dimensions. This shift of attention leads to the change in the generalization gradient depicted in Figure 1. The initial gradient (left) is spherical, reflecting the model's lack of knowledge of the dimensions' predictive validities. After 300 time steps (right), the gradient has been reshaped so that there is strong generalization between states as long as they match on the relevant dimension and very little generalization otherwise. Thus the model has learned the anisotropy of the task, which allows it essentially to estimate a common set of Q-values for all the states at each stratum (as an average over all the ws), while keeping the values for different strata separate.

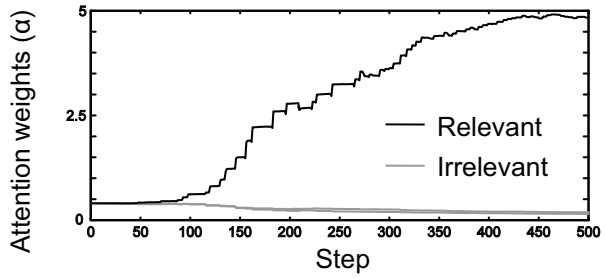


Figure 3. Dynamics of attention weights for one run of Q-ALCOVE in 3-dimensional Directional Gridworld.

The consequences of all three models' patterns of generalization are illustrated in Figure 4, which shows a two-dimensional slice through the center of the three-dimensional state space. Within each state, the four arrows indicate the model's estimated *Q*-values for the four actions within the plane. These values are from a single run of each model, after 300 time steps. Darker arrows indicate greater *Q*-values. The values for tabular Q-learning are irregular, reflecting the fact that they were learned separately for each state. In most states there has not been enough experience to obtain reliable estimates. The *Q*-values estimated by the fixed-generalization model are more accurate, because each draws on knowledge from neighboring states, so experience is used more efficiently. However, there is still irregularity among states at a given stratum (insufficient generalization across the irrelevant dimension) and too much smoothing (excess generalization) along the relevant dimension. Q-ALCOVE's estimated *Q*-values are much more accurate, allowing the model to select correct actions more reliably.

The Spores Task

Psychologically, the core assumption of Q-ALCOVE is that attention learning can be driven by internally generated TD-error signals, not just overt feedback. A behavioral experiment, reported by Cañas and Jones (2010), tested this hypothesis using a two-step task, in which Action 1 determined Stimulus 2, but reward was not received until after Action 2. The basic question was whether selective attention to the dimensions of Stimulus 1 could be learned, when the only immediate feedback was the identity of Stimulus 2.

Figure 5 illustrates the design of the task. Stimulus 1 (a cartoon mushroom spore) varied along two dimensions and was sampled from a circular set. This set was probabilistically divided into two regions, which had different consequences for the outcome of Action 1 (two options for how to grow the spore). The border between regions was oriented so that one dimension was more relevant than the other. The second step was designed so that the two possibilities for Stimulus 2 (two colors of mushrooms) each had a different optimal choice for Action 2 (selling the mushrooms to a troll or a goblin). Under these optimal actions, Stimulus 2a led to more reward than Stimulus 2b.

RL models in general predict subjects will learn internal values for Stimuli 2a and 2b (or their pairings with choices of Action 2), and these values will be used to generate internal feedback (TD error) for Action 1. This will in turn

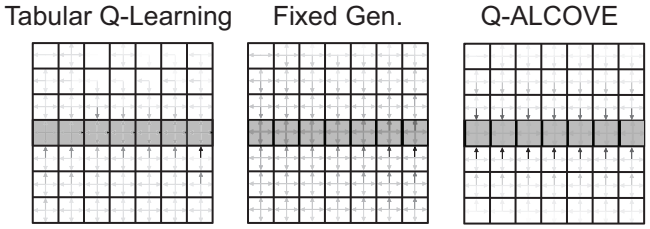


Figure 4. *Q*-values obtained from all three models after 300 steps in 3-dimensional Directional Gridworld. Shown is a 2-dimensional slice through the center of the state space. Arrows in each state correspond to the four actions within the plane. Darker arrows indicate greater *Q*-values.

allow subjects to learn to choose Action 1 so as to maximize the probability of obtaining Stimulus 2a (the more valuable mushroom). The key additional prediction of Q-ALCOVE is that TD error will also drive learning of attention to the more relevant dimension of Stimulus 1, to improve the effectiveness of generalization among stimuli.

Results revealed that subjects who learned the first step of the task also learned to selectively attend to the more relevant dimension (see Cañas & Jones, 2010, for details). Simulations of Q-ALCOVE corroborated this conclusion. Q-ALCOVE and the fixed-generalization version of the model were fit to the data of each subject using maximum likelihood. Aggregating over all 150 subjects, Q-ALCOVE fit reliably better, $\chi^2(150) = 1913.8$, $p \approx 0$. The difference between fits of the models was significant at the .05 level for 55 individual subjects. These results support the central hypothesis that attention learning, as embodied by Q-ALCOVE, was involved in learning the task.

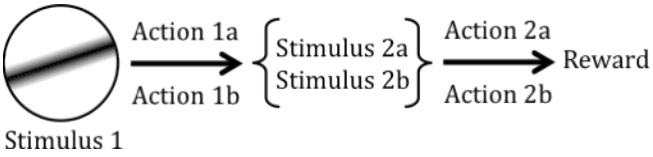


Figure 5. Structure of the Spores task.

Conclusions

Despite its computational power and neurological support, the basic principles behind RL are inherently limited by the representations it operates on. We argue here for a tight linkage between RL and mechanisms of representation learning established in other domains of psychology. Specifically, TD error, the engine behind nearly all RL models, can also drive updating of state representations. Representations thereby adapt so the pattern of generalization among states is tuned to the structure of the task, which in turn facilitates learning of optimal actions. RL's capacity to autonomously drive construction of representations that serve its needs greatly increases its power and flexibility, and hence its potential as a model of complex human learning.

The specific model proposed here draws on principles of selective attention from category learning and related domains (Nosofsky, 1986; Sutherland & Mackintosh, 1971). Shifting attention away from irrelevant dimensions allows

the learner to aggregate knowledge over states with similar outcomes, while attention toward relevant dimensions maintains discrimination of meaningful differences. Generalization in any learning task raises a bias-variance dilemma, in that more generalization reduces variance in parameter estimates but increases their bias. Selective generalization as modeled by attention learning is an elegant way of sidestepping this dilemma.

In a companion paper, we report empirical evidence supporting attention learning via TD error as a psychological mechanism (Cañas & Jones, 2010). Here we show how such a mechanism can be formalized in a mathematical model. Attention learning and RL in this model bootstrap off of each other, in that the internal value estimates generated by RL drive shifts of attention, and selective attention in turn improves RL's value estimates. This synergistic relationship, together with the elegance of the integration between the equations of Q-learning (Watkins & Dayan, 1992) and ALCOVE (Kruschke, 1992), suggests that RL and attention learning are similarly tightly coupled in the brain. The simulation studies reported here show that the unified model, Q-ALCOVE, is both computationally powerful and psychologically plausible.

Investigating attention is a useful first step because it acts to modify similarity directly, so that its effects on generalization are transparent. In further work, we plan to explore more complex psychologically supported mechanisms, such as stimulus-dependent attention (Aha & Goldstone, 1992), construction of new conjunctive features (Love, Medin, & Gureckis, 2004), and analogical mapping between structured stimulus representations (Markman & Gentner, 1993).

Psychological models that generalize knowledge based on pairwise similarity are closely related to kernel methods developed in statistics (Jäkel, Schölkopf, & Wichmann, 2007). Kernel methods add considerable flexibility to many learning algorithms, by allowing them to be recast from the raw stimulus space to a mathematical (Hilbert) space of functions (e.g., Cristianini & Shawe-Taylor, 2000). Q-ALCOVE can be viewed as a kernel method applied to RL. Viewed from the perspective of kernel methods, an important contribution of the present research is the proposal for adaptively modifying the kernel (i.e., generalization gradient) to improve learning. Learning the kernel has been a focus of recent research in machine learning (e.g., Micchelli & Pontil, 2007), but results thus far have been largely limited to existence theorems and global-search algorithms that seem psychologically implausible. Here we propose a simpler mechanism based on psychological principles. The mathematical results on kernel learning have been influential in guiding our design of well-behaved models and in inspiring more sophisticated mechanisms. Continuing to exploit this link to statistical and machine-learning techniques, while maintaining grounding in established psychological phenomena, seems promising for advancing the power and flexibility of psychological models.

References

Aha DW & Goldstone RL (1992). Concept learning and flexible weighting. *Proceedings of the 14th Annual Conference of the Cognitive Science Society*, 534-539.

Cañas F & Jones M (2010). Attention and reinforcement learning: Constructing representations from indirect feedback. *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*.

Cristianini N & Shawe-Taylor J (2000). *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press.

Fu W-T & Anderson JR (2006). From recurrent choice to skill learning: A reinforcement-learning model. *Journal of Experimental Psychology: General*, *135*, 184-206.

Jäkel F, Schölkopf B & Wichmann FA (2007). A tutorial on kernel methods for categorization. *Journal of Mathematical Psychology, 51*, 343-358.

Jones M, Maddox WT & Love BC (2005). Stimulus generalization in category learning. *27th Annual Meeting of the Cognitive Science Society*, 1066-1071.

Kruschke JK (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review, 99*, 22-44.

Love B, Medin D & Gureckis T (2004). SUSTAIN: A network model of category learning. *Psychological Review*, *111*, 309-332.

Markman AB & Gentner D (1993). Structural alignment during similarity comparisons. *Cognitive Psychology, 25*, 431-467.

Medin DL, Goldstone RL & Gentner D (1993). Respects for similarity. *Psychological Review, 100*, 254-278.

Medin DL & Schaffer MM (1978). Context theory of classification learning. *Psychological Review, 85,* 207-238.

Micchelli CA & Pontil M (2007). Feature space perspectives for learning the kernel. *Machine Learning, 66*, 297-319.

Newell A & Simon HA (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice Hall.

Nosofsky RM (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General, 115*, 39-57.

Schultz W, Dayan P & Montague P (1997, March). Neural substrate of prediction and reward. *Science, 275*, 1593-1599.

Schyns PG, Goldstone RL & Thibaut J-P (1998). Development of features in object concepts. *Behavioral and Brain Sciences, 21*, 1-54.

Shepard RN (1987). Toward a universal law of generalization for psychological science. *Science, 237*, 1317-1323.

Sutherland NS & Mackintosh NJ (1971). *Mechanisms of Animal Discrimination Learning*. NY: Academic Press.

Sutton R & Barto A (1998). *Reinforcement Learning: An Introduction*. The MIT Press.

Tesauro G (1995). Temporal difference learning and TD-gammon. *Communications of the ACM, 38*(3), 58-68.

Watkins CJCH & Dayan P (1992). Q-Learning. *Machine Learning*, *8*, 279-292.

Attention and Reinforcement Learning: Constructing Representations from Indirect Feedback

Fabián Cañas (canas@colorado.edu) & Matt Jones (mcj@colorado.edu)
University of Colorado, Department of Psychology & Neuroscience
Boulder, CO 80309 USA

Abstract

Reinforcement learning (RL) shows great promise as a theory of learning in complex, dynamic tasks. However, the learning performance of RL models depends strongly on how stimuli are represented, because this determines how knowledge is generalized among stimuli. We propose a mechanism by which RL autonomously constructs representations that suit its needs, using selective attention among stimulus dimensions to bootstrap off of internal value estimates and improve those same estimates, thereby speeding learning. Results of a behavioral experiment support this proposal, by showing people can learn selective attention for actions that do not lead directly to reward, through internally generated feedback. The results are cast in a larger framework for integrating RL with psychological mechanisms of representation learning.

Keywords: Reinforcement Learning; attention; generalization

Introduction

Humans have an incredible capacity to learn new and complex tasks in dynamic environments. In recent years, Reinforcement Learning (RL) has emerged as a theoretical framework that may explain how such powerful learning takes place (e.g., Sutton & Barto, 1998). Reinforcement learning draws on a synthesis of machine learning and neuroscience and offers a set of computational principles for describing learning of dynamic tasks. RL has led to major advances in the ability of machines to learn difficult tasks such as backgammon and autonomous helicopter flight (Tesauro, 1995; Bagnell & Schneider, 2001). RL has also received much interest in neuroscience, based on findings that phasic dopamine signals have similar properties to the internal feedback computed by RL algorithms (Schultz, Dayan, & Montague, 1997). This correspondence suggests that RL offers a useful model of biological learning.

Despite the promise of this framework, the learning performance of RL algorithms strongly depends on the representations on which they operate. RL works by learning which action to perform in each state of a task's environment. In realistically complex tasks with large state spaces, learning about every state individually is impossible, and instead the learner must generalize knowledge among states. Generalization is closely tied to similarity (Shepard, 1987), which in turn depends on how stimuli or situations are represented. Therefore the efficacy of generalization depends on how a task is internally represented. Most often in machine-learning applications, representations are pre-supplied by the modeler based on features that are carefully crafted to capture the most important aspects of the task being learned (e.g., Tesauro, 1995). In psychological contexts, stimuli are chosen so that the subject's representation is transparent, and consequently the question of how the representation is learned is neglected (Schyns, Goldstone, & Thibaut, 1998).

A great deal of psychological research in domains other than RL focuses on how people learn representations to facilitate learning, inference, and decision-making. The aim of our general research program is to explore how such mechanisms might interact with RL, and in particular how RL can build its own representations to bootstrap learning. In the present paper we focus on selective attention, building on models from the literature on category learning (Kruschke, 1992). In a companion paper (Jones & Cañas, 2010), we provide a formal framework for integrating representation learning with RL and implement a specific computational model based on selective attention. Here, we present a behavioral experiment that support the thesis that RL can drive representational learning. Our results show that the internally generated feedback signals at the core of RL can direct shifts of attention toward those stimulus dimensions that are most diagnostic of optimal action.

The remainder of this paper begins with background on RL and modeling of attention learning in categorization. We then outline our proposal for how RL and attention learning can bootstrap off of each other. We then report the results of a sequential decision-making experiment designed to test this specific proposal. Implications are discussed for the role of attention in more complex and temporally extended tasks, prescriptions for training in such tasks, and interactions between representation learning and declarative memory.

Reinforcement Learning

RL is a computational framework for learning dynamic tasks based on feedback from the environment. RL models represent a task as a set of environmental states together with a set of available actions in each state. The action selected at each step determines the immediate reward as well as the ensuing state. This general framework accommodates nearly any psychological task, from simple conditioning to elaborate planning (Sutton & Barto, 1998).

RL works by estimating values of states and actions, which reflect predictions of total future reward. From any given state, the action with the highest estimated value represents a best guess of the choice that will lead to the highest long-term reward. The key to learning value estimates, which lies at the heart of all RL models, is an internally generated feedback signal known as Temporal Difference (TD) error. TD error represents the discrepancy between the estimated value of an action prior to its execution and a new estimate based

on the immediate reward and the value of the ensuing state.

For the mathematically inclined, TD error is defined as

$$\delta = r_t + \gamma \cdot V(s_{t+1}) - Q(a_t, s_t).$$

Here, s_t represents the current state (at time t), a_t is the action selected, and $Q(a_t, s_t)$ is the estimated value of that action. The immediate reward received is denoted r_t, and $V(s_{t+1})$ is the estimated value of the ensuing state. The temporal discount parameter, γ, represents the degree to which the learner values immediate versus delayed rewards.

A critical question for all RL models concerns the relationship between value estimates (Q or V) for different states. The simplest approach is to learn values for all states independently, but for most realistic tasks with large state spaces this approach is unfeasible. Effective learning therefore requires generalization, or the use of knowledge about one stimulus or situation to make inferences or choose actions for other, similar stimuli. A number of methods have been proposed for implementing generalization in RL, and in all cases, the pattern of generalization depends strongly on the way in which states are represented. Representations relying on different features produce different patterns of similarity and hence different generalization. Learning will be most effective if generalization somehow respects the structure of the task, such that the learner pools knowledge about states with similar consequences but discriminates between states that are meaningfully different.

Representation

The various mechanisms for representation learning that have been identified in cognitive psychology all have potential application to RL as means for speeding learning through enhancing generalization. Our work thus far has focused on principles derived from research on category learning. Much of the literature on human category learning aims to understand how humans develop powerful internal representations that facilitate learning and inference of conceptual knowledge. The mechanisms that have been studied include selective attention (Kruschke, 1992; Nosofsky, 1986); feature discovery (Schyns et al., 1998), prototype formation (Smith & Minda, 1998); hybrid rule-exemplar representations (Erickson & Kruschke, 1998; Nosofsky, Palmeri, & McKinley, 1994); clustering representations that grow with task complexity (Anderson, 1991); mutable representations that evolve among exemplars, prototypes, and rules (Love & Jones, 2006; Love, Medin, & Gureckis, 2004); and conceptual networks based on causal knowledge (Murphy & Medin, 1985). In this paper, we examine the interaction of RL and selective attention.

Though attention has been studied under many guises in psychology, its implications for learning and generalization have been primarily explored in categorization and animal conditioning. In these literatures, attention has been proposed to act by reshaping generalization gradients (Sutherland & Mackintosh, 1971; Nosofsky, 1986). The generalization gradient is an empirical function that describes how strongly subjects generalize between stimuli as a function of how much those stimuli differ. This function is monotonically decreasing, but it decreases more rapidly along attended dimensions than unattended dimensions (Jones, Maddox, & Love, 2005). Thus subjects generalize less between stimuli when they are attending to the dimensions those stimuli differ on. An alternative view is that the generalization gradient is fixed and isotropic, but the perceptual scaling of individual stimulus dimensions is adjustable. Attention to a dimension serves to stretch the perceptual space so that stimuli differing on that dimension are less similar and thus produce less generalization (Nosofsky, 1986).

Theories of selective attention in category learning propose that people learn to reallocate their attention to improve performance. ALCOVE, a model of categorization with learnable selective attention, has an attention weight for each dimension that determines the degree of generalization along that dimension (Kruschke, 1992). The attention weights are learned by gradient descent on classification error, driven by external feedback. This process leads attention to shift to more predictive dimensions, which leads to less generalization along these dimensions and greater generalization along irrelevant dimensions. Selective attention can thus be thought of as a mechanism for representational learning, which facilitates future learning of the task by adapting generalization.

Incorporating Attention into RL

The previous two sections suggest a natural integration between RL and attention learning. RL's major focus is in updating value estimates by computing sophisticated feedback signals from temporal patterns of reward, but current RL models do not address how value estimates are represented. In contrast, theories of category learning focus on how representations are created that allow for effective generalization, but learning is driven by simple updating rules based on external feedback. We propose a natural unification, in which the feedback signals and updating rules from RL drive the representation-learning mechanisms identified in the categorization literature. This integration makes RL significantly more flexible and autonomous, and therefore possibly more aligned with biological learning.

The critical empirical question we explore operationalizes the idea that RL can adapt its own representation through learned selective attention. Specifically, we investigate whether attention learning can be driven by internally generated TD-error signals in the same way that has been observed with explicit external feedback (Nosofsky, 1986). In a companion paper (Jones & Cañas, 2010), we present a formal model that embodies this hypothesis, by synthesizing the learning mechanisms of ALCOVE (Kruschke, 1992) and Q-learning, a well-studied RL model (Watkins & Dayan, 1992). The formalism of the integrated model shows a tight and mathematically elegant synthesis of the two mechanisms, which we believe offers a strong candidate explanation of

how biological RL processes build their own representations. Here we present an experiment that tests that explanation, by assessing the human capacity for attentional learning via internal value and error signals as opposed to direct external feedback.

Experiment

The goal of the present experiment was to determine whether internal TD-error signals can drive attention learning in the absence of any immediate overt reward. The task consisted of a two-step decision process in which the action on the first step probabilistically determined the stimulus on the second step. Only after the second action did the subject receive feedback about reward.

The second stage of the task was a simple decision task with two possible stimuli and two possible actions. A different action was optimal (i.e., maximized reward) for each of these intermediate stimuli. Once this mapping was learned, one intermediate stimulus led to a higher reward than the other. RL predicts that once subjects learned the optimal actions on this second step, they would learn to assign differential values to the two intermediate stimuli. These values would in turn be used for computing a TD-error signal for actions in the first step, thereby allowing subjects to learn an action policy that maximizes the probability of obtaining the higher-valued intermediate stimulus.

The stimulus for the first choice varied on two continuous dimensions, one of which was more predictive of the outcome of the first action (i.e., the intermediate stimulus) and hence of which choice was optimal. The key question was whether learning the first action through TD error would also lead to learning of selective attention between stimulus dimensions, such that subjects would shift attention to the more relevant dimension. The stimulus set of the first step was designed so as to allow assessment of subjects' attentional allocation based on their patterns of errors, as described below.

Methods

150 undergraduate students from the University of Colorado, Boulder served as the experimental subjects in exchange for course credit.

Subjects were instructed they would pretend to be mushroom farmers. On each trial, they were presented with an image of a mushroom spore and asked to choose between two locations for growing the spore, Sun and Shade. This action determined the intermediate stimulus, a pair of blue or orange mushrooms. They were then given the option to sell the mushrooms to either a Troll or a Goblin, who paid them in gold coins. The structure of the task is outlined in Figure 1.

The stimulus in the first stage was a yellow spore shape, consisting of a circular center measuring 2.3 cm in diameter and radial spines arranged evenly around the center. The spines ranged from 8 mm to 260 mm in length and varied in number between 20 and 100. Spores were uniformly sampled from a circular region inscribed within this two-dimensional stimulus space. The spore was presented in the center of

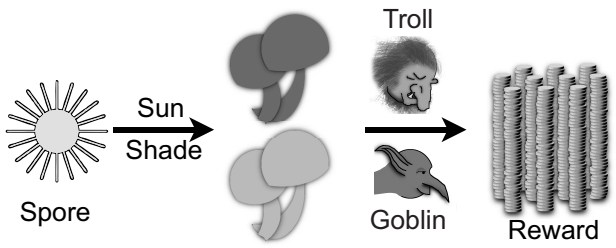


Figure 1: An overview of the task.

an LCD monitor over a black background. The subject selected an action by pressing either S (Sun) or C (Cave) on the keyboard. After this first response was given, the spore disappeared and a pair of cartoon mushrooms appeared in the center of the screen. The subject selected the second action by pressing T (Troll) or G (Goblin). The reward was then presented as stacks of gold coins with a numeric value underneath. The mushrooms and the chosen creature remained on the screen while the reward was displayed.

The transition after each step was animated, lasting 1200 ms between the first response and intermediate stimulus, and 970 ms between the intermediate stimulus and the reward. The reward remained on the screen for 800 ms. A blank screen separated the reward from the beginning of the next trial for 200 ms.

The reward structure for the second step was defined as shown in Table 1. Each mushroom color was associated with a different optimal action. Under these actions, one mushroom (henceforth referred to as the "good" mushroom) afforded a higher reward.

Table 1: Reward Structure of the Second Stage

	Creature Sold to	
Mushroom Color	Goblin	Troll
Blue	[200, 220]	[400 420]
Orange	[300, 320]	[100 120]

Note: Reward on each trial was sampled uniformly from the range shown.

The transition dynamics for the first step were defined as follows. For each action, the probability of one mushroom color versus the other was a logistic function of the dimension values of the spore, given by $p = 1/(1+exp(A(30L+10N)))$, where L and N represent the length and number of the spines, scaled to range from -1 to 1, and A represents the action on the first step, coded here as ± 1. The coefficients for L and N were counterbalanced between subjects, so that L was the more relevant dimension for half the subjects and N was more relevant for the other half. The effect of this design was to create an optimal decision bound, at an angle of 18.4° to one of the two axes, such that the action that maximized the probability of obtaining the good mushroom was determined by which side of the boundary each spore lay on.

Subjects were randomly assigned to Length-relevant and

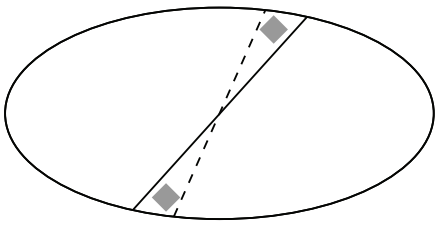


Figure 2: Predictions from selective attention in first step of task. Attention to the more relevant (horizontal) dimension leads to stretching of the stimulus space. Critical stimuli (grey) near ends of optimal decision bound (solid line) are predicted to lead to errors, producing rotation in best fit of linear classifier to subject's responses (dashed line).

Number-relevant conditions, which differed in which spore dimension was more predictive. The roles of the creatures, the colors of the mushrooms, and the labels for the first action were also counterbalanced between subjects. Each subject completed 240 trials (480 total decisions) in blocks of 40.

Predictions and Analysis

Our theory predicts subjects to shift attention to the more relevant spore dimension. Under the view of attention as a transformation of perceptual space, subjects' representations of the set of spores should become stretched along the more relevant dimension and compressed along the less relevant dimension, as shown in Figure 2. Consider the stimuli in the highlighted areas of the figure. Under the attention-altered representation, most of their neighbors lie on the opposite side of the optimal decision bound. Therefore, similarity-based generalization will lead to higher rates of suboptimal actions for these critical stimuli, as compared to matched stimuli on the other side of the optimal bound. The same prediction arises if one assumes subjects learn prototypes for spores associated to the two actions, because each critical stimulus is more similar to the opposite prototype (taken to be the centroid of the region on that side of the optimal bound). Therefore our predictions do not depend on an assumption of exemplar-based generalization.

To test this prediction, we used bivariate logistic regression to fit a linear classifier to each subject's responses. This classifier estimated a linear boundary in stimulus space that best divided the spores the subject chose to grow in the sun from those grown in the shade. To illustrate this analysis, Figure 3 shows the response distribution of a typical subject in the learning group (defined below). Open and closed circles represent stimuli for which the subject selected each of the two actions, the solid line represents the optimal bound, and the dashed line represents the output of the linear classifier. The prediction from selective attention, based on the analysis of expected errors described above, is that the boundary separating each subject's decisions will be rotated relative to the optimal boundary, as shown by the dashed line in Figure 2. Importantly, the estimation of a linear decision bound is a purely descriptive analysis that makes no commitment

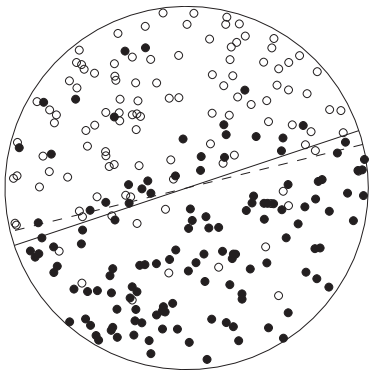


Figure 3: Distribution of responses on first step for a typical subject. Solid line shows optimal bound. Dashed line shows fit of linear classifier.

regarding psychological decision processes. In the companion modeling paper (Jones & Cañas, 2010), we fit a process model based on exemplar generalization, and it makes the same predictions.

In the absence of selective attention, the representation of the stimulus space would remain circular, and therefore by symmetry there should be no systematic bias in the subject's estimated decision bound. Therefore, testing for the predicted bias is a diagnostic way to determine whether our postulated attention-learning mechanism is operating.

Results

On average, subjects made the correct action on the second step of the task on 89.4% of trials. Figure 4 shows the distribution, across subjects, of the proportion of good mushrooms obtained following the first step. The histogram shows a clear bimodality, wherein many subjects performed at chance for the first step, but a significant number were able to learn effective actions.

As explained below, we only predict selective attention for subjects who learn the first stage of the task. Therefore we analyzed the responses of subjects who performed above 70% on the first stage. This cutoff was based on a visual inspection of Figure 4 to safely exclude subjects who were performing at chance. A total of 30 subjects performed at or above 70% on the first step of the task, 11 in the Length-relevant condition and 19 in the Number-relevant condition.

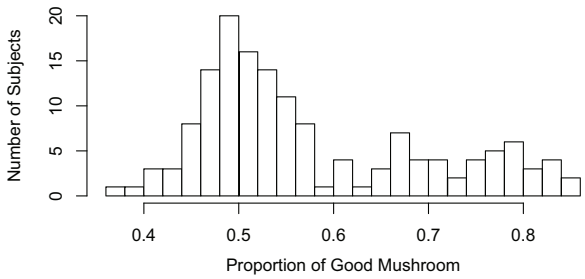


Figure 4: Distribution of performance on first step of task.

A linear classifier was fit to the first-step responses of each subject in the learning group. Figure 5 shows the orientations of the resulting decision bounds, indicated by dots on the circumference of the stimulus region. The mean orientation for each group is shown as a dashed line, and the optimal bound as a solid line. The Number-relevant condition is shown in black and the Length-relevant condition in grey. The mean orientation of the decision bound for subjects in the Length-relevant condition was 7.96° from the Number axis. This value was significantly different from the optimal bound ($18.4^\circ; t_{10} = -2.99, p = .014$) as well as from zero ($t_{10} = 2.29, p = .045$). The mean orientation for the Length-relevant condition was 7.33° from the Number axis. This too was significantly different from the optimal bound ($t_{18} = -3.25, p = .004$) and from zero ($t_{18} = 2.14, p = .046$).

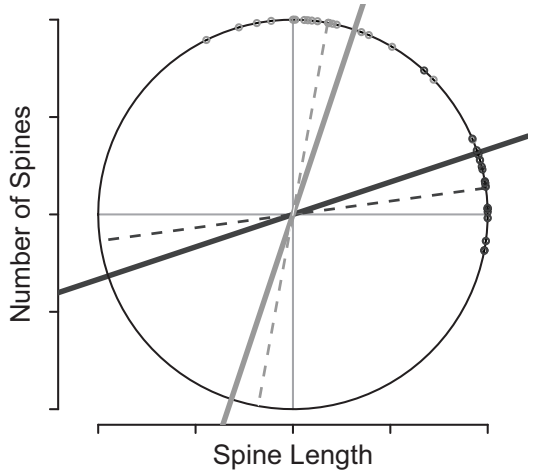


Figure 5: Orientations of empirical decision bounds for subjects in learning group. Small circles = subjects; dashed lines = means; heavy solid lines = optimal bounds; black = Number-relevant; grey = Length-relevant.

Discussion

The results of the decision-bound analysis confirm that subjects made more errors on the critical stimuli. This prediction follows directly from the assumption of selective attention to the more relevant dimension. Because actions on the first step only led to colored mushrooms and not nominal reward, our results support the proposal that attention learning can be driven by internal value estimates and error signals.

We only predicted selective attention for higher-performing subjects for three reasons. First, our theory only predicts attention to be learned once some amount of learning has taken place in associating stimuli to appropriate actions. Attention learning essentially works as a bootstrapping method operating by altering generalization and thus requires some amount of reliable knowledge to begin with in order for adaptation of generalization to have a useful effect. Second, because our theory predicts a bidirectional relationship between attention and value learning, those subjects who exhibit more selective attention should perform better on the task. Therefore, performance acts as a cue to indicate which subjects are more likely to exhibit a measurable effect. The third reason is purely methodological, in that the linear classifier requires a systematic set of responses in order to estimate a meaningful decision bound.

An alternative to our proposal of attention learning is that subjects simply disregarded one dimension of the stimulus entirely. This more strategic explanation is still consistent with our general theory of representation learning driven by RL, but the mechanism would be incompatible with continuous adjustment of attention weights. Regardless, the data rule out this explanation. The fact that the mean bound orientations were reliably different from zero (i.e., the less relevant axis) implies that subjects were sensitive to the less relevant dimension (they were just less sensitive to it than to the primary dimension). Another possibility is that some subjects disregarded one dimension and others disregarded the other, with most subjects in each condition disregarding the less relevant dimension. However, this explanation predicts a bimodal distribution of bound orientations at the subject level, which is clearly not present.

General Discussion

We have shown that humans can learn to shift attention in a dynamic task where reward is not given immediately following the decision that attention acts on. This finding tightly aligns with the internal TD-error signals that RL relies on, and it shows that direct external feedback is not required in order to learn selective attention.

At its core, RL uses predictions or knowledge about later states to build predictions and knowledge about prior states. Application of an RL model to our task predicts that after learning the second stage of the task, one mushroom becomes internally represented as more valuable than the other. This internal value in turn acts as a proxy reward that drives learning in the first stage of the task. Our findings support the proposal that this internal proxy reward signal is also capable of driving attention learning.

An alternative to the interpretation that our subjects are using RL-like internal values for the intermediate stimuli is the possibility of an explicit system that learns about both stages of the task simultaneously after the external reward at the end of each trial. Fu and Anderson (2008) found evidence for such a mechanism in a task structurally similar to ours. Explicit learning based on declarative memory is not, however, incompatible with RL. RL as we have discussed thus far, in its most simple form, only updates estimates about the most recent state. However, specific mechanisms, termed eligibility traces, have been explored within RL to maintain information across time steps to facilitate learning (Sutton & Barto, 1998). Eligibility traces permit simultaneous updating of multiple prior eligible states. Declarative memory may play an important role in encoding these eligibility traces, and therefore Fu and Anderson's results do not preclude an underlying RL

mechanism for learning several steps of a task at once.

Furthermore, declarative memory is unlikely to have played a role in the first step of the present experiment. First, in Fu and Anderson's design (2008), there was a direct correlation between the action in the first step and the eventual reward, which could support direct learning of the first action. In our design, only the conjunction of the spore and the action taken on it was directly related to the possible outcomes after the second step. Second, the spores were drawn from a rich set varying on two continuous dimensions, whereas the second stage of the task was very simple. Therefore subjects likely learned values for the intermediate mushrooms, which could then be used as feedback for the first action, well before the relatively weak correlation between spore-action pairs and final reward could be learned. Third, we have shown that subjects' decision bounds were consistently tilted away from unidimensional rules, indicating that subjects learned the first action using implicit information-integration processes not amenable to declarative memory (Ashby & Maddox, 2005). Though our current work does not completely preclude other learning mechanisms, we sought to isolate mechanisms directly related to RL and TD error, and our results show good support for such mechanisms.

Although not tested directly, the behavior of the subjects who did not learn the first stage sufficiently in our task fits well into the learning framework we propose. Before the differential value of the mushrooms is learned, the feedback to all actions of the first step is constant, which drives attention to generalize across the entire spore space. It is possible that by the time some subjects learned the optimal actions for the second step, they may have learned to entirely disattend any variability of the spores. This inattention is self-perpetuating and prevents future learning.

The potential for learned inattention in dynamic tasks has interesting theoretical and practical implications, because it could make aspects of a task far removed from overt reward difficult to learn. From this perspective, it is clear that an understanding of the mechanisms of attention learning could be beneficial in designing human training programs, such as backward chaining to train intermediate value representations before earlier stages are encountered.

The primary question we examined here was whether TD error, and therefore RL, can have an influence not just on learning values of stimuli within a fixed representation, but whether the representation itself can be altered. Shifts in attention alter the similarity structure of a stimulus space and therefore typify the sort of changes in representation we predict RL to effect. That humans exhibited changes in representation in the service of learning a new task involving fine discrimination of stimuli suggests a rich interplay of representation learning and RL.

References

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychol Rev*, *98*, 409–429.

Ashby, F. G., & Maddox, W. T. (2005). Human category learning. *Annu Rev Psychol*, *56*, 149-178.

Bagnell, J. A., & Schneider, J. G. (2001). Autonomous helicopter control using reinforcement learning policy search methods. *IEEE Int Conf Robo*, 1615-1620.

Erickson, M. A., & Kruschke, J. K. (1998). Rules and exemplars in category learning. *J Exp Psychol Gen*, *127*, 107-140.

Fu, W., & Anderson, J. R. (2008). Dual learning processes in interactive skill acquisition. *J Exp Psychol-Appl*, *14*, 179-191.

Jones, M., & Cañas, F. (2010). Integrating reinforcement learning with models of representation learning. *Proceedings of the* 32nd *Annual Conference of the Cognitive Science Society*.

Jones, M., Maddox, W. T., & Love, B. C. (2005). Stimulus generalization in category learning. *Proceedings of the* 27th *Annual Conference of the Cognitive Science Society*, 1066-1071.

Kruschke, J. K. (1992). Alcove: An exemplar-based connectionist model of category learning. *Psych Rev*, *99*, 22-44.

Love, B. C., & Jones, M. (2006). The emergence of multiple learning systems. *Proceedings of the* 28th *Annual Conference of the Cognitive Science Society*, 507-512.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). Sustain: A network model of category learning. *Psychol Rev*, *111*, 309–332.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychol Rev*, *92*, 289-316.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *J Exp Psychol Gen*, *115*, 39-57.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychol Rev*, *101*, 53–79.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*, 1593-1599.

Schyns, P. G., Goldstone, R. L., & Thibaut, J.-P. (1998). The development of features in object concepts. *Behav Brain Sci*, *21*, 1-54.

Shepard, R. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*, 1317–1323.

Smith, J., & Minda, J. (1998). Prototypes in the mist: The early epochs of category learning. *J Exp Psychol Learn*, *24*(6), 1411–1436.

Sutherland, N., & Mackintosh, N. (1971). *Mechanisms of Animal Discrimination Learning*. NY: Academic Press.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. The MIT Press.

Tesauro, G. (1995). Temporal difference learning and td-gammon. *Commun ACM*, *38*(3), 58-68.

Watkins, C. J. C. H., & Dayan, P. (1992). Q-learning. *Machine Learning*, *8*, 279–292.

Learning to Selectively Attend

Samuel J. Gershman (sjgershm@princeton.edu)
Jonathan D. Cohen (jdc@princeton.edu)
Yael Niv (yael@princeton.edu)
Department of Psychology and Neuroscience Institute, Princeton University
Princeton, NJ 08540 USA

Abstract

How is reinforcement learning possible in a high-dimensional world? Without making any assumptions about the structure of the state space, the amount of data required to effectively learn a value function grows exponentially with the state space's dimensionality. However, humans learn to solve high-dimensional problems much more rapidly than would be expected under this scenario. This suggests that humans employ inductive biases to guide (and accelerate) their learning. Here we propose one particular bias—sparsity—that ameliorates the computational challenges posed by high-dimensional state spaces, and present experimental evidence that humans can exploit sparsity information when it is available.

Keywords: reinforcement learning; attention; Bayes.

Introduction

Reinforcement learning (RL) in high-dimensional state spaces is a notoriously difficult problem in machine learning (Sutton & Barto, 1998), primarily because of the *curse of dimensionality*: The number of states grows exponentially with dimensionality (Bellman, 1957), and thus if one were naively to represent a separate value (expected reward) for each state, one would require astronomical amounts of data to effectively learn the value function (and thereby behave adaptively). Nonetheless, humans appear to learn rapidly from small amounts of data. Thus, while substantial evidence has accumulated that human behavior follows the predictions of RL models (Dayan & Niv, 2008), these models may fundamentally underestimate the learning capabilities of humans.

Following work in other areas of cognition (Braun, Mehring, & Wolpert, 2009; Kemp & Tenenbaum, 2009), we suggest that rapid learning arises from the exploitation of structured knowledge in the form of inductive biases. In particular, our proposal is that humans employ a sparsity bias: the assumption that only one (or a small number) of dimensions (input features) is relevant at any given time for predicting reward. For example, when you are at a stoplight, only the color of the light matters, not its shape, size, etc. In other domains (such as ordering food in a restaurant), you may know that dimensional relevance is sparse, but not which particular dimensions are relevant (does it matter which restaurant it is? which table I am sitting at? who the chef is? who the waiter is?); for this purpose, one requires a learning algorithm that can exploit sparsity. We formalize this idea in terms of rational statistical inference, and present new experimental evidence that human behavior is consistent with such a model.

Central to our analysis is the idea that selective attention is a direct consequence of Bayesian inference with a sparsity bias: Restricting attention to only a few dimensions is akin to the belief that only those dimensions are relevant for earning reward. This has the effect of reducing the space of possible value functions to a much smaller subspace.

While Bayesian probability theory stipulates the ideal learner, in general it may not be computationally tractable to perform the necessary calculations exactly (Kruschke, 2006; Daw & Courville, 2008). We therefore consider a "hybrid" model that combines the computational efficiency of model-free RL with the statistical efficiency of Bayesian inference. We compare the Bayesian and hybrid models to naive RL (no sparsity bias) and show that models that exploit structured knowledge better capture choice behavior in our experiment.

The Computational Problem

For concreteness, we consider one particular example of the general class of reinforcement learning problems for which the sparsity assumption holds. This example is meant to capture the abstract structure of many problems facing humans in the real world, where they must make choices between several multidimensional stimuli under conditions where most dimensions are unpredictive of reward. This example will also serve as a formal description of the task that we asked human subjects to perform, the results of which we report in a later section.

The subject plays N trials, and observes M stimuli simultaneously on each trial. The ith stimulus on trial n is denoted by a D-dimensional vector $\mathbf{x}_{ni}$, where each integer-valued component x_{nij} indicates the property of the jth stimulus dimension (for instance, [color = green, shape = triangle, texture = dots]). Each set of trials has a target dimension d (e.g., 'shape') and target property f on that dimension (e.g., 'circle'). The subject chooses a stimulus c_n on each trial and observes a binary reward r_n. The probability of reward given choice and target is

$$P(r_n = 1 | c_n, d, f, \mathbf{X}_n) = \begin{cases} \theta_1 & \text{if } x_{nc_nd} = f \\ \theta_0 & \text{otherwise,} \end{cases} \quad (1)$$

In other words, the subject receives a reward with probability θ_1 if the chosen stimulus posseses the target property on the target dimension, and with probability θ_0 otherwise.

Bayesian Model

Given uncertainty about the target dimension and property, a Bayesian model would use Bayes' rule to infer the posterior over the target dimensions and property and then calculate the

value of the stimulus by taking the expectation of reward with respect to the posterior:

$$V_n(c) = \sum_d \sum_f P(r_n = 1 | c_n = c, d, f, \mathbf{X}_n) P(d, f | \mathcal{D}_{n-1}, \mathbf{r}_{1:n-1}), \tag{2}$$

where $\mathcal{D}_{n-1} = \{\mathbf{X}_{1:n-1}, \mathbf{c}_{1:n-1}\}$. The intuition behind the Bayesian model is that the model weights the expected reward in each possible state of the world (i.e., target dimension and property) by the probability of the world being in that state given past observations. A key characteristic of this model is that a complete posterior distribution is maintained over states of the world, rather than a point estimate. The posterior distribution used by the Bayesian model is given by Bayes' rule:

$$P(d, f | \mathcal{D}_{n-1}, \mathbf{r}_{1:n-1}) \propto P(\mathbf{r}_{1:n-1} | \mathcal{D}_{n-1}, d, f) P(d, f), \tag{3}$$

where the prior is assumed to be uniform and the likelihood is given by:

$$P(\mathbf{r}_{1:n-1} | \mathcal{D}_{n-1}, d, f) = \prod_{t=1}^{n-1} P(r_t | c_t, d, f, \mathbf{X}_t). \tag{4}$$

Note that although this model describes the optimal learning rule, we shall assume that subjects are "weakly" rational in their decision rule (see the softmax choice function described below).

Reinforcement Learning Models

We now consider several alternative models based on RL. The intuition behind these models is that what ultimately matters for the choice value is the *expectation* under the posterior; so incrementally updating an estimate of this expectation from experience will eventually converge to the optimal choice values, even though these updates do not make optimal use of information on each trial. The various RL models differ principally in their construction of the value function.

Naive RL Model

The naive RL model represents a separate value for every possible stimulus-dimension-property configuration. Specifically, the choice value estimate is given by $V_n(c) = v_n(\mathbf{x}_{nc})$. This estimate is updated according to the learning rule:

$$v_{n+1}(\mathbf{x}_{nc_n}) = v_n(\mathbf{x}_{nc_n}) + \alpha \Delta_n, \tag{5}$$

where α is a learning rate and Δ_n is the *prediction error*:

$$\Delta_n = r_n - V_n(c). \tag{6}$$

Although the optimal solution is learnable by this model, its learning will be very slow, as we describe in the next section.

Function Approximation Models

One reason why the naive RL model may be ineffective in this task is that it lacks the ability to generalize across different configurations of features. Intuitively, if you knew the target dimension and property, then the value of a stimulus should be independent of the properties on the non-target dimension. However, the naive RL model yokes these together, such that learning operates on configurations of properties and hence fails to exploit this invariance. For example, the naive RL model learns a different value for green triangles with dots and for green triangles with waves, although the texture dimension may be completely incidental and not predictive of reward.

A more structured RL model that generalizes across configurations can be obtained by constructing the value function as a linear combination of D basis functions ϕ:

$$V_n(c) = \sum_{d=1}^{D} w_n(d, x_{ncd}) \phi_d, \tag{7}$$

where the weight matrix $\mathbf{W}_n$ determines how the basis functions are combined, with one weight for each dimension-property pair. This type of model is known as a *function approximation architecture* (Sutton & Barto, 1998). RL is used to update the weights according to:

$$w_{n+1}(d, x_{ncd}) = w_n(d, x_{ncd}) + \alpha \Delta_n \phi_d, \tag{8}$$

where the prediction error Δ_n is computed the same way as in the naive RL model (Eq. 6). This update can be understood as performing gradient ascent on the value function by optimizing the weight parameters (Williams, 1992).

We will consider a family of basis functions parameterized by η:

$$\phi_d = \frac{P_d^\eta}{\sum_j P_j^\eta}, \tag{9}$$

where $P_d = \sum_f P(d, f | \mathcal{D}_{n-1}, \mathbf{r}_{1:n-1})$. The basis function can be thought of as an "attentional focus" that encodes the subject's beliefs about what dimension is currently relevant. Thus, rather than maintaining the full posterior over target dimension and property (which may be quite computationally expensive), with the function approximation model the subject maintains the *marginal* posterior over target dimension (i.e., the probability of a dimension being the target, averaging over different target properties), which is then used to weight separate value functions, one for each dimension. When reward feedback is received, credit (or blame) is assigned to each value function in proportion to its posterior probability.[1] We refer to this model as the *hybrid* model, because it combines properties of RL and Bayesian inference.

Different settings of η lead to several special cases of interest:

[1] Note that the subject need not maintain and update the full posterior; any procedure that estimates the marginal posterior directly is consistent with this formulation.

- $\eta = 0$: uniform weighting of dimensions (diffuse focus).
- $\eta = 1$: exact posterior weighting of dimensions (optimal focus).
- $\eta \to \infty$: maximum a posterior (MAP) weighting (myopic focus).

Other intermediate scenarios are also possible. Thus, the value of η tells us how much information about the posterior distribution the subject is using to focus attention, with $\eta = 1$ being optimal focus and $\eta = 0$ completely ignoring information from the posterior and attending equally to all dimensions.[2] When η is larger than 1, the subject discards posterior uncertainty by focusing on the mode of the distribution, and is therefore overconfident in her beliefs about the relevant dimension.

One way to interpret the function approximation model is as a neural network in which the basis functions represent attentional units focusing on different sensory channels, and the weights represent synaptic connections between the attentional units and a reward prediction unit (Figure 1). The synaptic weights are updated using RL (Eq. 8). This interpretation resonates with ideas in computational neuroscience that view the dorsolateral prefrontal cortex as encoding attentional or task-related biases that interact with a striatal reward prediction system (Braver, Barch, & Cohen, 1999; Rougier, Noelle, Braver, Cohen, & O'Reilly, 2005; Todd, Niv, & Cohen, 2009). The prediction error Δ driving the weight updates is thought to be signaled by midbrain dopaminergic afferents to the striatum (Schultz, Dayan, & Montague, 1997)

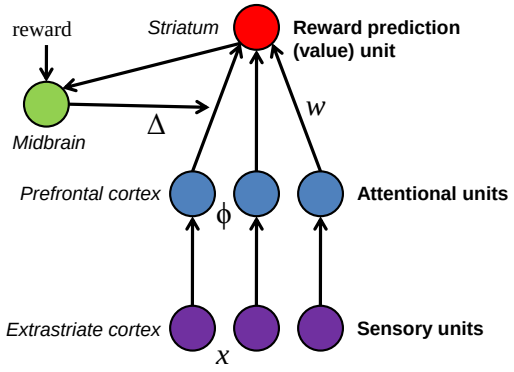


Figure 1: **Neural network interpretation of the hybrid model**.

Method

We now describe a behavioral experiment designed to quantitatively evaluate these models. Our experiment was inspired by the intra-dimensional/extra-dimensional set-shifting task (Dias, Robbins, & Roberts, 1996; Owen, Roberts, Polkey, Sahakian, & Robbins, 1991), in which subjects are asked to discriminate between visual stimuli on the basis of a particular (but unknown) dimension which they must learn from feedback, as well as the Wisconsin card-sorting task (Milner, 1963). We have adapted this task to a multi-armed bandit setting, such as has been used in previous studies of reinforcement learning (e.g., Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006; Schonberg, Daw, Joel, & O'Doherty, 2007).

Participants and Stimuli

Sixteen participants received 12 dollars reimbursement for performing 1800 trials of the task. The stimuli were triplets of stimuli varying along three dimensions: color (red, yellow, green), shape (circle, triangle, square), and texture (waves, dots, lattice). An example triplet is shown in Figure 2.



Figure 2: **Example experimental stimuli**.

Procedure

For each game, the target dimension and property were chosen randomly and with equal probability. On each trial, the subject was presented with a random triplet and asked to choose one of the stimuli. The stimuli on each trial were generated by a random permutation of the property assignments. After making the choice, the subject received feedback about whether or not her choice resulted in a reward. If the subject chose the stimulus with the target dimension/property pair, she received a reward with probability 0.75. Otherwise, reward was delivered with probability 0.25. The targets changed on each game (lasting 20-30 trials), and subjects were informed when a new game was beginning.

Choice Probabilities

To map from values to choices, we define a *policy* π_n that specifies the probability $\pi_n(c)$ of making choice c on trial n. Here we use the "softmax" policy defined by

$$\pi_n(c) = \frac{e^{\beta V_n(c)}}{\sum_a e^{\beta V_n(a)}}, \tag{10}$$

where β is an *inverse temperature* parameter that allows us to model stochasticity in the subject's responses.

Parameter Estimation and Model Comparison

We fit the parameters of each model with MAP estimation using gradient descent and calculated the Laplace approximation (Kass & Raftery, 1995) to the log marginal likelihood

[2] It is important to note that diffuse focus is not the same as the naive RL model. For all values of η, the function approximation model is still able to generalize across different configurations, unlike the naive RL model.

(evidence) for each model m according to:

$$E_m = \ln \int_{\omega_m} P(\omega_m)P(\mathbf{c}|\omega_m)d\omega_m$$
$$\approx \ln P(\hat{\omega}_m) + \ln P(\mathbf{c}|\hat{\omega}_m) + \frac{1}{2}G_m \ln 2\pi - \frac{1}{2}\ln|\mathbf{H}_m|, \quad (11)$$

where ω_m is the set of parameters for model m, $P(\mathbf{c}|\hat{\omega}_m) = \prod_{n=1}^{N} \pi_n(c_n|\hat{\omega}_m)$, $\hat{\omega}_m$ is the MAP estimate of the parameters, G_m is the number of parameters (length of ω_m), and $\mathbf{H}_m$ is the Hessian matrix of second derivatives of the negative log-posterior evaluated at the MAP estimate. We then calculated the log Bayes Factor relative to chance (where all choices are equiprobable) according to $E_m - N\ln(1/3)$. A larger Bayes Factor indicates greater support for a model. Note that the chance (null) model has no parameters. In addition to comparing models based on Bayes Factors, we also calculated predictive log-likelihood on a held-out game using a leave-one-out cross-validation procedure.

For all the models, we fit an inverse temperature β, placing on it a Gamma(2,2) prior. This served to ameliorate a well-known degeneracy in models with both a temperature and learning rate, such that these parameters tend to trade-off against each other (inverse temperature becoming very large and learning rate very small). For the RL models, we fit a learning rate α, placing on it a Beta(1.2,1.2) prior, which slightly biases the fits away from the parameter boundaries. We also allowed θ_1 and θ_0 to vary across subjects, since we only told subjects that the target would be rewarding more often than non-targets, placing on θ_1 a Beta(12,4) prior and on θ_0 a Beta(4,12) prior; these priors were chosen to have as their expected value the true—but unknown—values of θ_1 and θ_0. Finally we placed a Uniform(0,10) prior on η.

Results

Figure 3 presents the log Bayes Factors for each model, summed across subjects, along with the cross-validation resutls. Zero represents the null (chance) model in both cases. Clearly all the models do better than chance, but the naive RL model appears to perform substantially worse than the others. Overall, the hybrid model appears to best match behavior on this task. Figure 4 displays the distribution of log Bayes Factors for the Bayesian and hybrid models, showing that there are also individual differences in which model is favored for each subject.

Additional insight into these models can be gained by inspecting aggregate learning curves (the probability of choosing the optimal stimulus as a function of trials within a game). As shown in Figure 5, the naive RL model appears to consistently underestimate the speed of learning exhibited by subjects, whereas both the Bayesian and hybrid models hew closely to the empirical learning curve. One peculiarity of the learning curve is that subjects appear to learn faster than the Bayesian model. We believe that this is an artifact of the softmax choice probability function: the inverse temperature parameter appears to be too low early in a game and slightly too high later in a game. No single value of the inverse temperature would be able to capture this pattern.

	Log Bayes Factor	*Held-out log-likelihood*
Bayesian	5425	5620
Hybrid	**5892**	**6208**
Naive	3307	3312

Figure 3: **Model comparison results**. Highest scores are shown in bold.

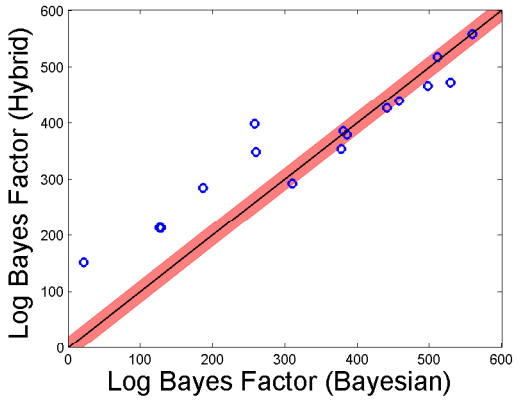


Figure 4: **Comparison of Log Bayes Factors for Bayesian and hybrid models**. Points above the diagonal are favored by the hybrid model. The red shaded region indicates the confidence interval outside of which one model is more likely than the other with $p < 0.05$.

Another question we can ask is whether subjects who behave more in accordance with the Bayesian model or hybrid model earn more reward overall. Figure 6 does indeed show this relationship (measured as the correlation between reward earned and the log Bayes Factors for the Bayesian model relative to the hybrid model), suggesting that subjects who more effectively exploit the structure of the task tend to perform better. The correlation is significant at $p < 0.01$. The hybrid model log Bayes Factor relative to the null model also correlates with total reward ($p < 0.02$).

Figure 7 shows the parameter estimates for η on a log-scale, demonstrating that subjects cluster around 0, corresponding to exact posterior weighting (optimal focus). This was supported by a sign test which failed to reject (p=0.45) the null hypothesis that $\ln(\eta)$ was drawn from a distribution with 0 median. Thus, within the family of possible basis functions, posterior attentional weighting best describes human behavior on this task.

Discussion

In this paper we have posed a problem that humans face in everyday life: how to learn value functions in high-dimensional state spaces. The crucial assumption that makes this possible

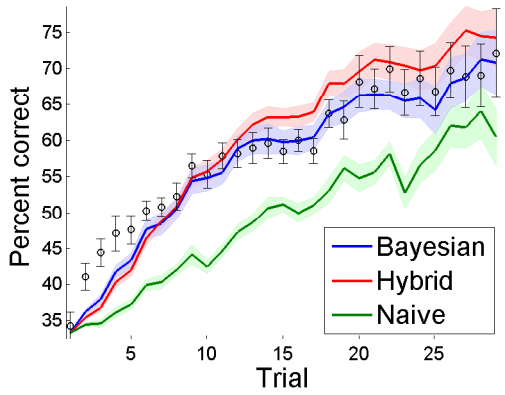


Figure 5: **Learning curves**. Probability of choosing the optimal stimulus as a function of trial within a game. The circles represent the empirical choice probability. Error bars are standard errors of the mean.

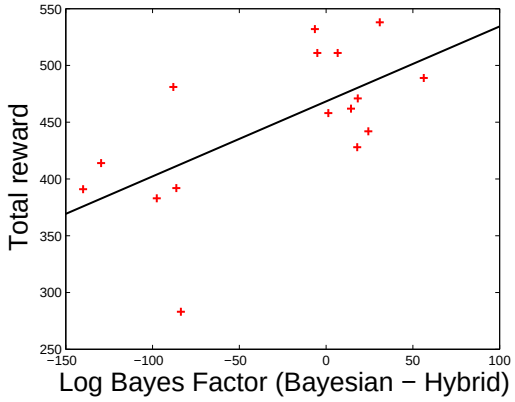


Figure 6: **Individual differences in earned reward**. On the x-axis is plotted the log Bayes Factors of the Bayesian model relative to the hybrid model, and on the y-axis is plotted the total reward earned. A least-squares line is superimposed on the scatter plot.

is that only one or a few dimensions is relevant at any given time. By employing this sparsity bias in the machinery of Bayesian inference, the effective dimensionality of the problem is reduced. This can be understood as a kind of selective attention that is learned through experience.

Our experimental results demonstrate that humans can exploit sparsity information when it is available. We compared a Bayesian model and a family of sophisticated RL algorithms against a naive RL model that ignores sparsity information and hence does not generalize across stimulus configurations, the key ingredient to efficient learning. Our computational analysis of behavior on this task suggests that humans combine reinforcement learning with Bayesian inference, rather than using a purely Bayesian strategy. This makes sense if the brain's learning algorithms are designed to deal with high-dimensional problems for which exact Bayesian inference is

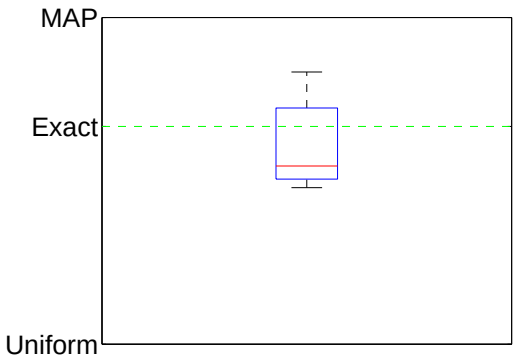


Figure 7: **Boxplot of $\ln(\eta)$ estimates**.

intractable. The hybrid model represents a compromise between the computational efficiency of model-free RL and the statistical efficiency of Bayesian inference.

The idea that selective attention can be framed as the outcome of Bayesian inference has been explored by several authors (Dayan, 2009; Rao, 2005; Yu, Dayan, & Cohen, 2009). Most relevant to our work is the competitive combination model of Dayan, Kakade, and Montague (2000), in which stimuli are assumed to vary in how reliably they predict reward. Dayan et al. (2000) showed that selective attention to particular stimuli falls naturally out of inference over the causal relationships between stimuli and reward in such a model. Our work is conceptually similar, with the main difference that we model inference over dimensions, rather than just stimuli. As emphasized by Dayan et al. (2000), the selectivity of attention in our model is based on proceses of statistical inference, rather than resource constraints. This point is particularly important to explaining how attention is *learned*; resource-limitation models, without further elaboration, do not speak to this issue.

The central role of selective attention has been extensively explored in the category learning literature, notably by Nosofsky (1986) and Kruschke (1992). The basic idea is that learned attentional weights amplify or attenuate specific stimulus dimensions in a way that facilitates category discrimination. Recently, Kruschke (2006) has attempted to connect these ideas to a form of approximate Bayesian inference he dubs "locally Bayesian learning" (LBL). Much as in our work, attention arises in LBL as a consequence of weighting different hypotheses about the currently relevant stimulus dimension in response to new evidence. In this sense, LBL is form of hybrid model; here we have attempted to identify a continuum through which Bayesian knowledge can influence RL, and fit this to data to identify where human learners fall on this continuum. At the same time, although our model shares some characteristics with categorization models such as LBL (see also Heller, Sanborn, & Chater, 2009), it is important to note that the problem it is designed to solve

is conceptually different: it does not receive feedback indicating the correct response (as in supervised category learning), but must instead learn from probabilistic reward feedback.

While our work was partly inspired by earlier neural network models (Braver et al., 1999; Rougier et al., 2005), our goal in this paper was to step away from implementational details and interrogate computational- and algorithmic-level concerns. Future work will need to examine more systematically how the algorithmic ideas presented here map onto neural mechanisms. We are currently investigating this question with functional magnetic resonance imaging.

In conclusion, the main theoretical and experimental contribution of this paper is showing that the human RL system is more sophisticated than previous computational models have given it credit for. This may not, after all, be that surprising; many years of machine learning research have shown that the naive assumptions of previous models simply do not scale up to high-dimensional real world problems. It remains to be seen what other hidden sophistications in the RL system await discovery.

Acknowledgments

We thank Michael Todd for invaluable discussion. SJG was supported by a Quantitative Computational Neuroscience training grant from the National Institute of Mental Health.

References

Bellman, R. (1957). *Dynamic Programming*. Princeton, NJ: Princeton University Press.

Braun, D., Mehring, C., & Wolpert, D. (2009). Structure learning in action. *Behavioural Brain Research*.

Braver, T., Barch, D., & Cohen, J. (1999). Cognition and control in schizophrenia: A computational model of dopamine and prefrontal function. *Biological Psychiatry*, *46*(3), 312–328.

Daw, N., & Courville, A. (2008). The pigeon as particle filter. *Advances in neural information processing systems*, *20*, 369–376.

Daw, N., O'Doherty, J., Dayan, P., Seymour, B., & Dolan, R. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, *441*(7095), 876.

Dayan, P. (2009). Load and attentional bayes. In D. Koller, D. Schuurmans, Y. Bengio, & L. Bottou (Eds.), *Advances in neural information processing systems 21* (pp. 369–376).

Dayan, P., Kakade, S., & Montague, P. (2000). Learning and selective attention. *Nature Neuroscience*, *3*, 1218–1223.

Dayan, P., & Niv, Y. (2008). Reinforcement learning: the good, the bad and the ugly. *Current opinion in neurobiology*, *18*(2), 185–196.

Dias, R., Robbins, T., & Roberts, A. (1996). Dissociation in prefrontal cortex of affective and attentional shifts.

Heller, K., Sanborn, A., & Chater, N. (2009). Hierarchical learning of dimensional biases in human categorization. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, & A. Culotta (Eds.), *Advances in neural information processing systems 22* (pp. 727–735).

Kass, R., & Raftery, A. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430).

Kemp, C., Perfors, A., & Tenenbaum, J. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science*, *10*(3), 307–321.

Kemp, C., & Tenenbaum, J. (2009). Structured statistical models of inductive reasoning. *Psychological review*, *116*(1), 20–58.

Kruschke, J. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*(1), 22–44.

Kruschke, J. (2006). Locally Bayesian learning with applications to retrospective revaluation and highlighting. *Psychological Review*, *113*(4), 677–698.

Milner, B. (1963). Effects of different brain lesions on card sorting: The role of the frontal lobes. *Archives of Neurology*, *9*(1), 90.

Nosofsky, R. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–57.

Owen, A., Roberts, A., Polkey, C., Sahakian, B., & Robbins, T. (1991). Extra-dimensional versus intra-dimensional set shifting performance following frontal lobe excisions, temporal lobe excisions or amygdalo-hippocampectomy in man. *Neuropsychologia*, *29*(10), 993–1006.

Rao, R. (2005). Bayesian inference and attentional modulation in the visual cortex. *Neuroreport*, *16*(16), 1843.

Rougier, N., Noelle, D., Braver, T., Cohen, J., & O'Reilly, R. (2005). Prefrontal cortex and flexible cognitive control: Rules without symbols. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(20), 7338.

Schonberg, T., Daw, N., Joel, D., & O'Doherty, J. (2007). Reinforcement learning signals in the human striatum distinguish learners from nonlearners during reward-based decision making. *Journal of Neuroscience*, *27*(47), 12860.

Schultz, W., Dayan, P., & Montague, P. (1997). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593.

Sutton, R., & Barto, A. (1998). *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press.

Todd, M., Niv, Y., & Cohen, J. (2009). Learning to use Working Memory in Partially Observable Environments through Dopaminergic Reinforcement. In *Neural information processing systems* (pp. 1689–1696).

Williams, R. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, *8*(3), 229–256.

Yu, A., Dayan, P., & Cohen, J. (2009). Dynamics of attentional selection under conflict: Toward a rational Bayesian account. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(3), 700–717.

Pavlovian conditioning from a foraging perspective

James J. Anderson (jjand@uw.edu)
School of Aquatic and Fishery Sciences, University of Washington
Seattle, WA 98195 USA

Chloe Bracis (cbracis@uw.edu)
Quantitative Ecology and Resource Management, University of Washington
Seattle, WA 98195 USA

R. Andrew Goodwin (Andy.Goodwin@us.army.mil)
Environmental Laboratory, U.S. Army Engineer Research & Development Center,
Portland, OR 97208 USA

Abstract

Principles in foraging and standard associative learning theories motivate a model for Pavlovian conditioning. The model tracks distal and proximal scales of expected reward probabilities plus the strength of signal-reward association. A combined reward probability is developed by combining the distal and proximal estimates through their uncertainties. Possible neural structure equivalents to the model variables are discussed. Model flexibility is demonstrated with data on the partial reinforcement extinction effect, a phenomenon difficult to explain with learning models.

Keywords: Mathematical model; Pavlovian conditioning; associative learning; foraging theory; partial reinforcement extinction effect, neural structures.

Foraging and Associative Learning

Pavlovian conditioning (PC), or associative learning, is one of the most well studied psychological processes and has an array of associated phenomena. The two main processes, acquisition of a behavior by pairing a signal and reward in trials and extinction of the behavior by removing the reward during the trials, can be explained by a number of models, the most common being the delta model where learning is guided by the error, i.e. delta, between the expected and received reward per trial (Rescorla & Wagner, 1972). However, the basic delta model cannot explain other widely known phenomena, including spontaneous recovery of behavior when signaling trials are conducted after extinguishing the behavior and the partial reinforcement extinction effect (PREE) where intermittent rewards during acquisition trials increase the resistance to extinguishing behavior and strengthen the response during spontaneous recovery trials. In particular, the PREE challenges associative models of PC, since lower reward expectations for partial compared to continuous reinforcement would appear in principle to produce faster extinction. One theory for PREE involve the rate of reinforcement, such the ratio of the current cumulative signal duration since the last reward to the average signal duration between rewards (Gallistel & Gibbon, 2000). However, this result is not supported experimentally (Haselgrove et al., 2004). An alternative verbal model proposed by Pearce et al. (1997) assumed that unrewarded trials during partial reinforcement schedules create a different context where the unrewarded trials signal rewarded trials. We develop a model that readily explains PC phenomenon in what we believe is a robust manner. Our model builds on PC and foraging theories and the neuroscience of memory and decision-making.

Animals in natural and laboratory environments meld their past and current experiences in making decisions; it is often assumed that the laws and processes in both environments are similar if not identical. In foraging, animals choose between staying in the current patch and moving to another. Deciding when to give up on a patch depends on the expected reward rate of the current patch relative to the expected reward rate on another patch discounted for the interpatch travel time (Charnov, 1976). In PC the response rate reflects expectations within the single environment as dependent on learning and unlearning of signal-reward associations (Bouton, 1993). Notably, in both frameworks, responses are based on comparisons of distal and proximal information. Distal information in PC is the memory of the previous signal-reward patterns, while in foraging distal information is the memory of the average habitat reward rate. Proximal information in PC reflects the most recent reward outcomes, while in foraging it reflects the expected foraging rate in the current patch. In both frameworks, expected reward for the next trial is computed by a delta rule, which is an exponentially weighted moving average (EWMA) that adds a percentage of the most recent outcome to a percentage of the previous expectation.

The two frameworks diverge in how non-reward events and extinction are treated. PC models commonly consider acquisition-extinction patterns in terms of distinct learning streams. A stream developed during the acquisition phase of the experiment characterizes signal-reward expectations, and a second stream developed during the extinction phase characterizes a signal-no reward expectation. Extinction learning inhibits the original acquisition learning (Bouton, 1993). However, when animals are retested after some interval of time the extinction learning is forgotten and spontaneous recovery of the original learning appears (Sissons & Miller, 2009). In foraging models, the learning streams do not inhibit each other nor are they forgotten.

Expected probabilities of rewards are tracked for both the current patch, i.e. the proximate patch, and the habitat, i.e. the distal patch. With parallel information streams, the animal does not need to distinguish whether information belongs in the acquisition or the extinction learning stream, an issue in PREE experiments where signals without rewards occur during the acquisition phase. Rather, the animal is constantly adapting to an always changing environment.

Patch foraging models naturally involve multiple temporal scales because information on the proximal patch is always more recent than information on distal patches. To capture these temporal differences, models have expressed distal and proximal learning with slow and fast learning rates respectively (Anderson, 2002; Moorter et al., 2009). Mixed learning rates are also used in the primary value learned value (PVLV) model (O'Reilly et al., 2007), that seeks a mapping to dopaminergic neuron dynamics during reinforcement learning.

Retaining reward probabilities across different temporal and spatial scales requires memory systems, and here neuroscience provides information on their nature. McClelland et al. (1995) postulated memories are created and stored in a two-stage process involving short- and long-term processes. First, events are stored via synaptic changes in the hippocampal system, a short-term memory (STM) which then supports reinstatement of recent memories into long-term memories (LTM) in the neocortex. The neocortical synapses change by a small amount on each reinstatement, which assures that learning, as a stochastic process, converges to the mean value of the statistical association of ensembles of experiences. The hippocampal system permits rapid learning of new items without disrupting the neocortex structure, and interleaves and integrates them into the neocortical system. In essence, the LTM is built-up incrementally from activation of STM. Furthermore, since extinction involves new learning, evidence suggests multiple memory systems may be applicable to the neural basis of extinction (Gabriele & Packard, 2006). We suggest the distal and proximal information streams which are contained in both PC and foraging models represent the STM and LTM system identified by neurological studies.

Forgetting is the other side of remembering and is important in PC models to explain spontaneous recovery. The idea being that the information stream acquired in the extinction phase is forgotten over time, which then removes the inhibition of the information streams acquired in the acquisition phase. This process is offered as an explanation for the stronger spontaneous recovery response that is observed with greater time between extinction and recovery tests and thus supports the view that learning in the extinction phase dissipates more rapidly than learning in the acquisition phase (Brooks & Bouton, 1993; Rescorla, 2004; Sissons & Miller, 2009).

Studies on forgetting provide valuable insight into its significance in associative learning. Recent memories are vulnerable to interference from other mental activity and Wixted (2005) suggested that forgetting is largely a function of nonspecific *retroactive interference* acting on memory traces that have not yet consolidated in the neocortex. Wang & Morris (2010) hypothesized that extinction trials involve reactivation of the acquisition-trial memories in the absence of further reinforcement. However, such interactions can be complex and two memories may mutually exclude each other or coexist depending on the timing of the signal during extinction (Perez-Cuesta & Maldonado, 2009).

Decision making is treated differently in PC and foraging models. In foraging models, the decision to leave a patch is depends on which patch has the higher reward probability (maximizing) or is selected probabilistically (matching) (Kacelnik, Krebs , & Ens, 1987). PC models do not have choice-based decision rules and express the response rate as a monotonic function of the reward expectation. However, if PC and foraging have the same basis, then PC models contain a hidden decision rule in which the animal chooses between proximal and distal information. However, decision rules in both PC and foraging models are incomplete because psychology, ecology, neuroscience, and machine learning research show that uncertainty in the reward assessment is an important factor in decision-making (Daw et al., 2005; Platt & Huettel, 2008).

The Model

We now develop a model for PC that has application to foraging models, draws on concepts from both modeling frameworks, and has some analogy to the neurology of decision-making. We model reward probability estimates for distal and proximal information streams, which correspond to the immediate patch and the surrounding habit in foraging models and to the short- and long-term estimates of rewards in PC models. We then combine the estimates with weightings based on their respective uncertainties. We also account separately for the process of learning that a signal can indicate a reward. Finally, we use the weighted expectation to model the animal's response rate in a trial.

Distal and Proximal Reward Estimates

For each trial we define the distal and proximal expected reward estimate with a modified delta model,

$$\hat{x}_{j,i} = m_j y_i \delta_{j,i-1} + \hat{x}_{j,i-1} \quad (1)$$

where j = 1, 2 indicates distal and proximal information streams, i designates a PC trial, m_j is the learning rate for stream j. For each stream the error between the expected reward probability and realized reward is

$$\delta_{j,i} = x_i - \hat{x}_{j,i} \quad (2)$$

where x_i is a reward/no-reward outcome (0,1) for trial i. The term, y_i, is a measure of the strength of the association of the signal-reward and is independent of reward probabilities. For convenience, we consider the distal and proximal information streams *unconscious* reward estimators because individually they are sub-process that must be combined to

affect the animal's response. We designate the combined estimator the *conscious* reward estimate.

Combined Estimate

The distal and proximal estimates of reward probability are combined into a single conscious reward estimate that the animal uses in making decisions:

$$\bar{x}_i = w_{1,i}\hat{x}_{1,i} + w_{2,i}\hat{x}_{2,i} \quad (3)$$

where the estimates are combined according to their respective weighting factors that depend on their associated uncertainties $\tilde{\delta}^2_{j,i}$. As we develop next, the uncertainties are in fact EWMAs of the variance in the distal and proximal estimators and so the estimates can be combined using a standard statistical weighting formula in which the weight for estimate *j* on trial *i* is

$$w_{j,i} = \left(1/\tilde{\delta}^2_{j,i}\right)/\left(1/\tilde{\delta}^2_{1,i} + 1/\tilde{\delta}^2_{2,i}\right). \quad (4)$$

It is noteworthy that this weighting scheme is not found in either PC or foraging models.

Temporal Discounting Uncertainty

The uncertainties used in weighting, $\tilde{\delta}^2_{j,i}$, are developed from the mean-squared errors of the distal and proximal reward estimates. Of relevance, the uncertainties depend on the time between trials as follows. First, compute unadjusted uncertainty estimates as EWMAs from errors defined by eq. (2):

$$\hat{\delta}^2_{j,i} = n(\delta^2_{j,i-1} - \tilde{\delta}^2_{j,i-1}) + \tilde{\delta}^2_{j,i-1}, \quad (5)$$

where *n* is the uncertainty learning rate. Next, adjust the uncertainties for the time interval $\Delta t = t_i - t_{i-1}$ between trials:

$$\tilde{\delta}^2_{j,i} = \hat{\delta}^2_{j,i} h_j^{\Delta t} \quad (6)$$

where h_j is a decay parameter that controls the rate at which the uncertainty in information stream *j* decays as time between trials increases. In this model, as the inter-trial time increases, we want to put more confidence on the distal (long-term) estimate and less on the proximal (short-term) estimate. The idea being that in a sequence of trials with uncertain outcomes, as time passes since the last trial we should trust the long-term estimate of reward probability more than the short-term estimate based only on the last few rewards. To insure this shift in confidence to the distal estimate, we decay the distal uncertainty but not the proximal uncertainty as time passes between trials: Mathematically this is achieved with $0 < h_1 < 1$ and $h_2 = 1$.

Signal-Reward Association

The term y_i in eq. (1) tracks the strength of the signal-reward association, which we assume is distinct from probability learning but also depends on the error of predictions. Learning requires repetition and reduction of errors in prediction, and we model these properties with a three step process. First, we track conscious error based on the difference between the trial outcome and the conscious expectation from eq. (3) giving

$$\delta'_i = x_i - \bar{x}_i. \quad (7)$$

Second, because errors are by nature random and one correct prediction, $\delta'_i = 0$, is insufficient to develop an association, we compute an average error with a EWMA:

$$\bar{\delta}^2_i = n\left(\delta'^2_{i-1} - \bar{\delta}^2_{i-1}\right) + \bar{\delta}^2_{i-1}. \quad (8)$$

where *n* is again the uncertainty learning rate. Third, to capture the repetitive and asymptotic nature of appetitive learning, we incrementally accumulate the uncertainties with a standard saturation function

$$y_i = \sum_{k=1}^{i-1} 1/\bar{\delta}^2_k \Bigg/ \left(g + \sum_{k=1}^{i-1} 1/\bar{\delta}^2_k\right) \quad (9)$$

where *g* is the halfway point in the learning process.

Response Rate

We relate the conscious reward expectation to the response rate with a matching function that asymptotically increases a response from a background level to a maximum and is defined with scale and shape parameters r_{max} and *r* as

$$R = r_{\max}\bar{x}_i / \left(\bar{x}_i + r(1-\bar{x}_i)\right). \quad (10)$$

Parameter Summary

The complete model combines elements of classical associative leaning and patch foraging. While several models contain multiple memory streams that track information over different time scales, the model presented here tracks the uncertainties in the estimates as information streams as well. The model contains 7 parameters (Table 1).

Table 1: Model parameters and values fitted to data.

Parameter	Fitted value	Meaning
m_1	0.055	Distal learning rate
m_2	0.248	Proximal learning rate
n	0.075	Uncertainty learning rate
h_1	0.126	Distal uncertainty decay rate
g	971	Association half-way constant
r_{max}	6.88	Response function scale parameter
r	0.13	Response function shape parameter

Comparison to Experiment

To demonstrate the flexibility and perspective the model provides, we fit it to a study of partial reinforcement extinction conducted by Haselgrove, et al. (2004). We selected this experiment because PREE is difficult for PC models to explain. In addition, the study covers an acquisition phase and two extinction phases, which demonstrate spontaneous recovery. Several models produce these basic patterns but not when one of the groups is trained with partial rewards.

In the experiment, rats divided into partial and continuous reinforcement groups received the same signal and number of rewards during an acquisition phase in which the reinforcement schedules differed. In the partial group, half of the trials were reinforced with two rewards, while in the continuous group one reward was given on every trial. Following the acquisition sessions, the rats received two sessions with unreinforced signals. In Figures 1-3, each point designates an entire session in the acquisition phase, while each point represents a block of two trials in the two extinction sessions following.

We fit the model to the data from both groups with a single set of parameters (Table 1) using the "mco" package in the R statistical programming language. This is a multi-criteria optimization algorithm based on a genetic algorithm (cran.r-project.org/web/packages/mco/mco.pdf).

The model fit the response patterns for the continuous and partial groups reasonably well. The mean responses in the acquisition phase developed in a similar manner for both groups, while in the extinction phase the continuous group response decayed more rapidly than the response in the partial group. Both groups exhibited spontaneous recovery in the final extinction session with the continuous group response again decaying faster than the partial group response (Figure 1).

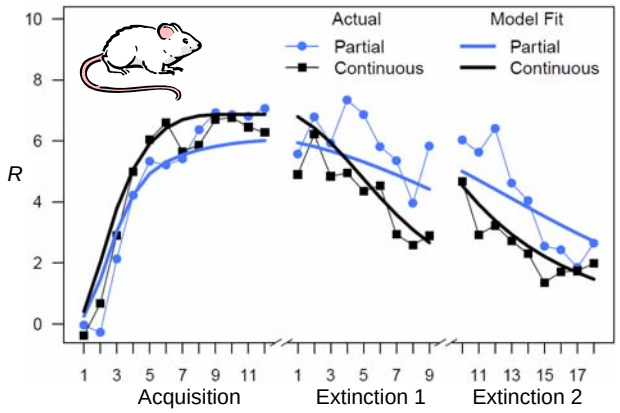


Figure 1: Haselgrove et al. (2004) data and model fit for partial and continuous reinforcement groups using parameter values in Table 1.

Discussion

The patterns of the underlying streams producing the fit to the Haselgrove et al. (2004) data for the continuous reinforcement group (Figure 2) and the partial reinforcement group (Figure 3) illustrate how a framework of multiple-scale estimators and uncertainties can account for seemingly complex patterns in PC studies. As in Figure 1, the first section consists of session averages for the acquisition sessions, and the next two sections each represent an extinction session in blocks of two trials.

Continuous Reinforcement Group In Fig. 2a the signal-reward association strength, *y*, rises over the acquisition phase to its full value and remains constant thereafter, implying that the animal has fully learned the association. The conscious reward probability also reaches its full value in the acquisition phase and then exponentially declines in the extinction phases. At the beginning of the second extinction phase, the expectation is higher than at the end of the first extinction phase, then the expectation again decays since the animal receives no rewards. This somewhat complex pattern of responses is generated by a unique weighting of relatively simple patterns in the distal and proximal estimators. The proximal estimator (Figure 2b), which is generated by a faster learning coefficient, rises quickly in the acquisition phase and then decays quickly in the first extinction phase and remains at zero throughout the second extinction phase. The distal estimator, being the slow learner, rises slowly in the acquisition phase and then decays slowly over the next two phases. The pattern in the weights (Figure 2c) that mix the two estimators produces the spontaneous recovery. Beginning in the acquisition phase, the weightings are equal. Because rewards are consistently received, the proximal estimator quickly adjusts and has less uncertainty than the distal estimator, giving the distal estimator the greatest weight in forming the conscious estimator in eq. (4). In the period between the acquisition and extinction phases, eq. (6) decays the distal uncertainty (trust the long-term estimate when information is old), so the two weights are nearly equal beginning the extinction phase. However, as signals are consistently unrewarded, the proximal estimator better represents the environment and its weight rises over the trials. The distal uncertainty decays again after the first extinction phase, and the pattern is repeated in the second extinction phase. At the beginning of the second extinction the proximal estimator, which predicts a reward, has a higher weight than the distal estimator, which predicts no reward, so the animal exhibits spontaneous recovery.

Partial Reinforcement Group In the acquisition phase, the patterns of conscious expectation and the signal-reward association (Figure 3a) are similar to the patterns in the continuous reinforcement group (Figure 2a), although the strengths are half the continuous reinforcement values. The patterns in the distal and proximal estimators are similar also (Figure 3b), and again the strengths are about half showing the accurate estimation of the 50% reward probability during acquisition. However, the experiments differ significantly in the weighing function patterns. These are reversed in the partial reinforcement group (Figure 3c) compared to that in the continuous reinforcement group (Figure 2c). This difference drives the differences in the response patterns (Figure 1). Again, at the beginning of the experiment, the distal and proximal uncertainties are equal, making for equal weights. However, both estimators

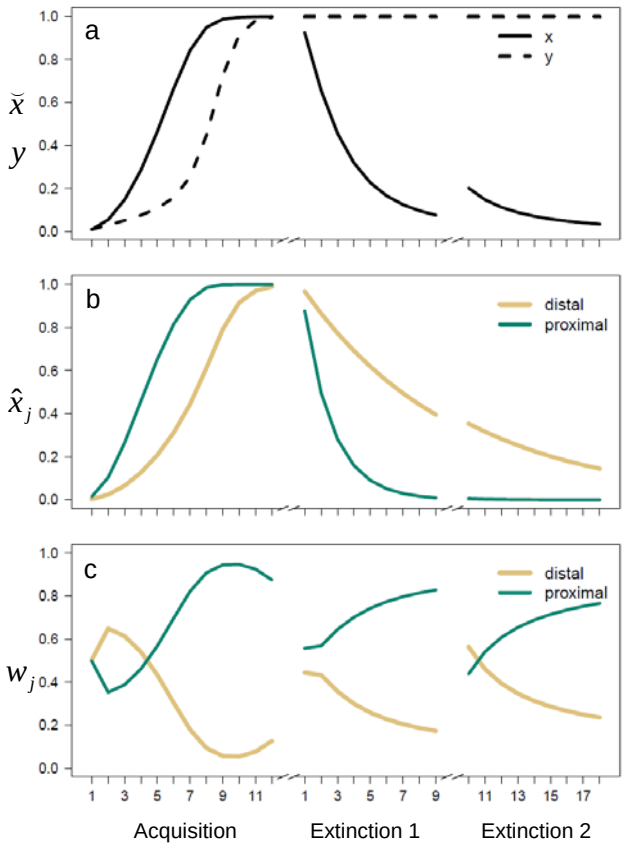


Figure 2: Changes in model variables for the continuous reinforcement group.

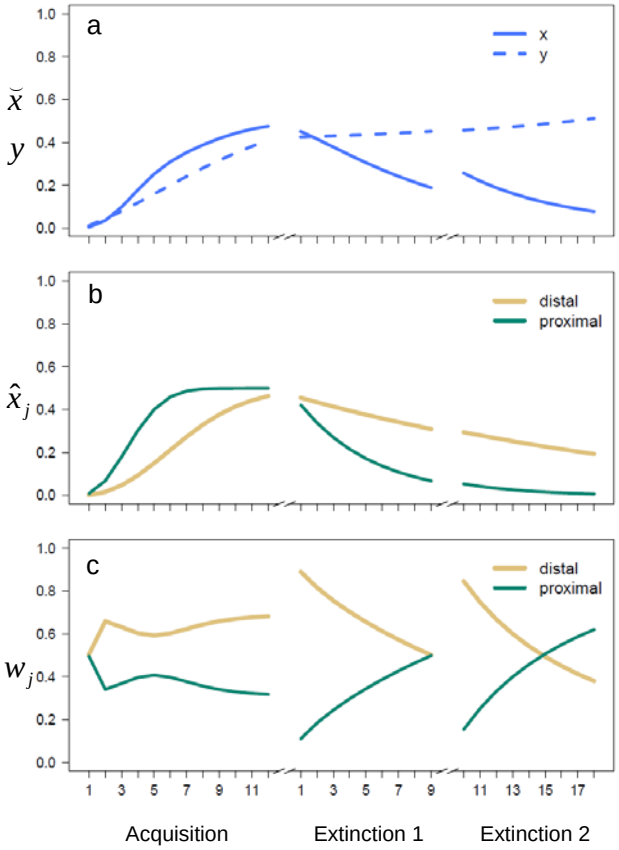


Figure 3: Changes in model variables for the partial reinforcement group.

have higher uncertainty with partial reinforcement, but the proximal estimator, which is strongly influenced by the previous trial, has higher uncertainty than the distal estimator, which integrates the reward expectation over multiple trials. The result is lower uncertainty for the distal estimator and thus greater weight in forming the conscious estimator. Between the acquisition and extinction phases, the distal uncertainty declines while the proximal uncertainty is fixed, so the distal estimator is dominant at the beginning of the first extinction phase. Over the phase the distal uncertainty increases while the proximal uncertainty decreases until they are equal at the end of the extinction. Therefore, at the end of the extinction phase, the animal has a higher response rate than in the continuous case, which is dominated by the proximal estimator. Between the first and second extinction phases, the distal estimator uncertainty again decays giving it more weight in the second extinction phase, resulting in a higher response and slower decline in response for the partial acquisition group.

Neurological Analogies

As our ultimate goal is to model the brain, not just observed behavior, we seek to identify possible equivalences between the model's elements and neural structures as has been encouraged by Rangel et al. (2008) and others. In a broad sense, we suggest that the distal and proximal information streams $\hat{x}_{1,i}, \hat{x}_{2,i}$ represent parallel memory systems that characterize reward probabilities estimated on different temporal scales. These terms might be candidates for STM-LTM systems involving the hippocampus and neocortex. However, the two streams are competitive and so they might be representative of competitive memory systems such as the hippocampus and basal ganglia (White & McDonald, 2002; Poldrack & Packard, 2003). In our model the signal-reward association y_i represents a separate memory stream that builds in a cumulative manner by summing the inverse of trial-by-trial uncertainties. This incremental building of memories is also a feature of the STM-LTM interaction of the hippocampus and neocortex (McClelland, McNaughton, & O'Reilly, 1995).

Uncertainty is specifically formulated in our model, and neural structures are clearly involved with uncertainty in decision-making. For example, Doya (2008) noted uncertainty has two flavors: one resulting from the environmental stochasticity (risk) and one from the limited knowledge of the decision-maker (ambiguity). Studies suggest that risk is represented in the striatum and precuneus while ambiguity is represented in the lateral obitofrontal cortex and amygdala (Platt & Huettel, 2008). Our model also has two flavors of uncertainty. The

uncertainty in the distal and proximal reward estimators $\hat{\delta}^2_{j,i}$ tracks variability in the environment that we suggest is akin to risk uncertainty. The uncertainty in signal-reward association $\bar{\delta}^2_i$ is a candidate for the decision-maker's ambiguity.

Final Thoughts

Under the assumption that animals in laboratory studies use behavioral strategies and neurological processes that evolved through natural selection, we reconsider Pavlovian conditioning in the context of animal foraging. From this perspective, animal behavior in a constrained environment has a hidden spatial component that leads us to consider the behavior in terms of distal (habitat) and proximal (local patch) information streams. In this framework the animal does not track distinct memory streams for acquisition and extinction phases, which we suggest is the experimenter's perspective. Instead of having to know when the experimenter ends one phase and starts another, the animal can view the environment as continuous yet random and simply track information measuring over two different time scales and weighting the estimates according to the trial-by-trial changes in their uncertainties.

Acknowledgments

This work was supported by the U.S. Army Engineer Research and Development Center. Permission was granted by the Chief of Engineers to publish this information.

References

Anderson, J. J. (2002). An agent-based event drive foraging model. *Natural Resource Modeling, 15*(1), 55-82.

Bouton, M. E. (1993). Context, time, and memory retrieval in the interference paradigms of Pavlovian learning. *Psychol Bull, 114*(1), 80-99.

Brooks, D. C., & Bouton, M. E. (1993). A retrieval cue for extinction attenuates spontaneous recovery. *J Exp Psychol Anim Behav Process, 19*(1), 77-89.

Charnov, E. L. (1976). Optimal foraging, the marginal value theorem. *Theor Popul Biol, 9*(2), 129-136.

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat Neurosci, 8*(12), 1704-1711.

Doya, K. (2008). Modulators of decision making. *Nat Neurosci, 11*(4), 410-416.

Gabriele, A., & Packard, M. G. (2006). Evidence of a role for multiple memory systems in behavioral extinction. *Neurobiology of Learning and Memory, 85*(3), 289-299.

Gallistel, C. R., & Gibbon, J. (2000). Time, rate, and conditioning. *Psychological Review*, 107, 289-344.

Haselgrove, M., Aydin, A., & Pearce, J. M. (2004). A partial reinforcement extinction effect despite equal rates of reinforcement during Pavlovian conditioning. *J Exp Psychol Anim Behav Process, 30*(3), 240-250.

Kacelnik, A., Krebs , J. R., & Ens, B. (1987). Foraging in a changing environment: an experiment with starlings (Sturnus vulgaris). In A. K. a. S. J. S. M.L. Commons (Ed.), *Quantitative Analysis of Behavior Foraging* (Vol. 6, pp. 63–87). Hillsdale: L. Erlbaum.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychol Rev, 102*(3), 419-457.

Moorter, B. V., Visscher, D., Benhamou, S., Börger, L., Boyce, M. S., & Gaillard, J.-M. (2009). Memory keeps you at home: a mechanistic model for home range emergence. *Oikos, 118*(5), 641-652.

O'Reilly, R. C., Frank, M. J., Hazy, T. E., & Watz, B. (2007). PVLV: The Primary Value and Learned Value Pavlovian Learning Algorithm. *Behavioral Neuroscience, 121*(1), 31-49.

Pearce, J. M., Redhead, E. S., & Aydin, A. (1997). Partial reinforcement in appetitive Pavlovian conditioning with rats. *Quarterly Journal of Experimental Psychology*, 50B, 273-294.

Platt, M. L., & Huettel, S. A. (2008). Risky business: the neuroeconomics of decision making under uncertainty. *Nat Neurosci, 11*(4), 398-403.

Poldrack, R. A., & Packard, M. G. (2003). Competition among multiple memory systems: converging evidence from animal and human brain studies. *Neuropsychologia, 41*(3), 245-251.

Rangel, A., Camerer, C., & Montague, P. R. (2008). A framework for studying the neurobiology of value-based decision making. *Nat Rev Neurosci, 9*(7), 545-556.

Rescorla, R. A. (2004). Spontaneous recovery. *Learn Mem, 11*(5), 501-509.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II* (pp. 64-99). New York: Appleton-Century-Crofts.

Sissons, H. T., & Miller, R. R. (2009). Spontaneous recovery of excitation and inhibition. *J Exp Psychol Anim Behav Process, 35*(3), 419-426.

Wang, S.-H., & Morris, R. G. M. (2010). Hippocampal-Neocortical Interactions in Memory Formation, Consolidation, and Reconsolidation. *Annual Review of Psychology, 61*(1), 49-79.

White, N. M., & McDonald, R. J. (2002). Multiple Parallel Memory Systems in the Brain of the Rat. *Neurobiology of Learning and Memory, 77*(2), 125-184.

Wixted, J. T. (2005). A Theory About Why We Forget What We Once Knew. *Current Directions in Psychological Science, 14*(1), 6-9.

A cognitive model of punishment

Francesca Giardini (francesca.giardini@istc.cnr.it)
Institute of Cognitive Sciences and Technologies, Via San Martino della Battaglia, 44
00185 Rome Italy

Giulia Andrighetto (giulia.andrighetto@istc.cnr.it)
Institute of Cognitive Sciences and Technologies, Via San Martino della Battaglia, 44
00185 Rome Italy

Rosaria Conte (rosaria.conte@istc.cnr.it)
Institute of Cognitive Sciences and Technologies, Via San Martino della Battaglia, 44
00185 Rome Italy

Abstract

People use sanctioning behaviours differently according to what they believe and want to achieve, according to the context and to the situation. We need to understand the motivations for different forms of punishment in order to explain why sanctions and incentives have different effects on human behaviour. Aim of this work is to propose a cognitive model of three distinct kinds of punishing behaviours, differentiated in terms of the defining cognitive patterns.

Keywords: Cognitive modeling; Punishment; Cooperation.

Introduction

Punishment is a core mechanism to enforce and support social order, to promote cooperation and to prompt group beneficial behaviours. Social scientists have long debated on the nature and the effects of this mechanism, but there are many questions still open, as for instance the relationship between counter-punishment and cooperation. There is a growing body of evidence that altruistic punishment plays a crucial role in enforcing cooperation and in promoting group welfare (Fehr & Gachter, 2000, 2002), but some recent experimental results raised the problem of antisocial punishment, that is sanctioning people who behave socially. Herrmann, Thoni, and Gachter (2008) compared results on punishment and cooperation collected in sixteen different participant pools around the world. They showed the emergence of antisocial punishment in repeated public goods experiments, and proposed that differences can be explained in terms of different societal background. Nikiforakis and Engelman (2008) used a public good game with multiple punishment stages aiming at investigating whether retaliatory behaviours would escalate into a feud. Interestingly, cooperation rates declined but feuds were avoided by participants.

Although a number of accounts (for some representative work see (Bowles & Gintis, 2004; Henrich & Boyd, 2001; Henrich et al., 2006) have stressed the relevance of punishment in human societies, they suffer the flaw that they consider punishment as a unique behaviour. In our view, punishing actually consists in a complex behavioral repertoire in which it is useful to disentangle at least revenge (social-status punishment), retaliation (strategic punishment) and sanction (normative punishment).

Treating punishment as a single behaviour without caring for its cognitive foundations could be misleading especially if one is interested in explaining cooperation and its maintenance in evolutionary terms. There is neither a single form of punishment nor a single motive to punish other people, and the question is: How can we distinguish between punishment aimed at making the individual internalize the norm and pure revenge? How do people choose between punishment and revenge?

Cognitive modelling allows us to disentangle apparently indistinguishable acts and to understand the motives and objectives that pave the way to distinct ways of punishing. Taking revenge is not the same as punishing a wrongodoer or sanctioning a deviant behaviour, and explaining these differences and the related motivations could effectively advance research on cooperation and prosocial behaviours under several respects.

The rest of this article is organized as follows. Firstly, we will introduce a general theory of cognitive social action, in order to provide some basic concepts. Secondly, revenge will be analyzed, focusing on the explicit mental representations behind it. Therefore we will turn our attention to punishment, showing what is inside the punisher's mind. Finally, sanction will be described. Future work and conclusions will follow.

The cognitive roots of social behaviour

In general, this work aims at unveiling the proximate mechanisms of enforcement behaviours, in order to understand the mental mechanisms underlying revenge, punishment and sanction.

Before these arguments are developed, some terminological issues need clarification. As stated elsewhere (Conte & Castelfranchi, 1995), an agent is a goal-governed system.

> By this, we mean an entity, not necessarily autonomous, that has the capacity to act upon the external world in order to reduce the discrepancy

between the world itself and some regulatory state that is somehow represented within the entity (p.1).

A cognitive agent is endowed with cognitive representations of the external world and of its internal states as well. Agents have *beliefs* about themselves and the world and they act on the basis of their *goals* to reduce the discrepancy between the world and what they want.

There are several ways to influence agents, but here we refer to *cognitive influencing* (Cialdini & Goldstein, 2004; Conte & Castelfranchi, 1995), a process by which a given entity, say *Ii*, acts on another entity, *mj*, in such a way that a given goal of *mj*' s be strengthened or generated anew. Notice that, since *mj* is an autonomous intelligent system, *Ii* must act on her beliefs in order to strengthen or generate new goals and modify her behaviours. We will address here the *goal-generation process*, as strengthening an existent goal is only a weaker case of cognitive influencing. To strengthen or generate a new goal, *mj* must acquire a new belief, say *Bjp* (*Ii* will harm *mj*, if she does not apply his will). This belief will activate a previous goal of *mj* 's, *Gjp* (avoid harm), and the interaction between *Bjp* and *Gjp* generates a new instrumental goal in *mj*'s, *Gjq* (adopt *Ii*'s will)[1] .

This is a social plan of action, which is based on a complex variant of the *theory of mind.* In the classic theory of mind (Leslie, 1991; Baron-Cohen, 1991; Dennett, 1987; Premack & Woodruff, 1978), others' mental states are harboured in one's mind, giving rise to *social beliefs*, namely beliefs about others' mental states (e.g. beliefs, intentions, desires, emotions, etc). In cognitive influencing, instead, the influencing entity has *social goals* as well, i.e. goals about others' mental states.

As we will see in the following sections, the presence and the type of cognitive influencing permits to discriminate between apparently similar enforcing mechanisms that are actually very different.

The three punishing strategies can be arranged on two axes: cognitive complexity and intentionality of deterrence. In this way, revenge easily appears to be the lowest in cognitive complexity and to pursue deterrence as an emergent and unintended self-reinforcing effect. The opposite is true for sanction (high cognitive complexity and intentional deterrence), whereas punishment occupies an intermediate position.

Revenge

Revenge appears to be a common human trait, widespread in human history and societies. According to the Merriam-Webster dictionary, vengeance is *punishment inflicted in retaliation for an injury or offense.* The retaliatory aspect is the main feature of revenge and is what makes this form of reaction differing from the two forms of punishment described below. Vengeance is also strongly characterized by the presence of emotional aspects that contribute to the common view of revenge as a not fully rational behaviour.

This "flavour of irrationality" could have contributed to the paucity of interest in revenge, compared to punishment, among scholars. While justifications for punishing and individual motives to punish have been widely investigated, research on retaliatory actions has considered them either as tribal and archaic forms of norm enforcement (Boehm, 1986) or as genetic predispositions evolutionary evolved to react to aggressions (Elster, 1990).

Amegashie and Runkel (2008) present a differential game model of revenge in conflicts. In their model, revenge has a positive value in economic terms; this means that, however destruction is costly, given what has been suffered in the past, the victim derives satisfaction and then utility from exacting revenge in the present. Similarly, deQuervain et al. (2004) used neuroscientific methodology to investigate how brain regions reacted to defection in an interaction game. According to Knutson (2004) their results show that punishing a defector activates brain regions related to the anticipation of a reward, even when punishment was costly, thus explaining human preference for punishing violators. Interestingly, Nikiforakis and Engelman (2008) reported data on revenge causing collaboration to decline in the lab but without boosting a chain of reciprocal vengeance.

Broadly speaking, the term 'revenge' refers to two diverse but connected phenomena. On one side, revenge is a social ritual that requires and prescribes specific behaviors to group members to repair an offense. The Kanun, a customary set of laws used mostly in northern Albania and Kosovo, disciplined people's reactions to murder (blood revenge or *gjakmarrje*) and other offenses (*hakmarrje*) according to the roles and degree of kinship of all the people involved. Shirking revenge or taking it without respecting what is stated in the Kanun lead to the same result: honour can not be restored and the whole family or clan is to blame. Shackleford (2005) considers "cultures of honor", in which revenge is the primary form of reaction to aggressions, likely to emerge and be maintained where the state is weak and can not prevent or punish theft.

It is worth noticing that in general retributive concepts of law and the creation of institutions are considered as advancements to replace vengeance and avoide blood feuds[2], but the Kanun itself was a social institutions aimed at preserving social order (*KLD*, 1989).

On the other side, revenge is an individual behaviour

[1] By means of the so called adoption rule (Conte & Castelfranchi, 1995), according to which an autonomous agent (adopter) will have another agent's (adoptee) goal as her own, if she, the adopter, comes to believe that the adoptee's achievement of this goal will increase the chances that the adopter will in turn achieve one of her previous goals.

[2] In this work we are not interested in analyzing the emergence and function of blood feud and we consider revenge in isolation

found both in human (Zaibert, 2006) and non-human primates (Jensen, Call, & Tomasello, 2007), reacting to personally harmful actions.

As Elster observes, revenge is "the attempt at some cost or risk to oneself, to impose suffering upon those who made one suffer, *because they have made one suffer* (emphasis added)".

In our view, revenge serves a terminal goal, that of making the aggressor suffer, and this excludes any other concerns. Usually, vengeance occurs in groups of equals, in which the offense is perceived also as an attempt to reduce an individual's prestige, to declass him or her family. Repaying the offense becomes a way to reaffirm one's status in front of both the aggressor and the social group and this behaviour is far from being extincted in present societies.

It is worth noticing that revenge may act as a deterrent from further aggressions, but this is an emergent function that can not be even represented in the avenger's mind. Revenge is not pursued to affect the likelihood that the wrongdoer will repeat the aggression in the future, inducing her to cooperate next time or deterring her from further aggressions. The avenger wants to repay the damage she suffered with an equal or greater offense, no matter how much risky or dangerous this retaliation is. In a sense, we can say that the avenger is a "backward looker" that revolves around the past and acts in the present to rebalance what happened, without any concerns for his future.

Into the avenger's mind

We claim that vengeance entails a specific configuration of goals and beliefs and that this configuration differs from those implied by terminal and instrumental punishment. This means that, although the punisher and the avenger could perform the same action, their aims and intentions were deeply different as well as the resulting state of the world.

In order to describe revenge, we need first to introduce its actors. There are at least three roles agents play in revenge. There is the avenger (A), the Target (T), and the Onlookers (O). The avenger's beliefs and goals involve both the target (T) and the onlookers (O), whose presence, as we shall see, is crucial.

Looking into the avenger's mind, we find a set of beliefs that are necessary to trigger the desire to take revenge[3]. The offended agent should, at least, believe that (1) the offense he received was intentional, (2) T was the main or the unique responsible and then liable for punishment, (3) there is a material and/or symbolic dimension to be restored in front of T and O.

The above set of belief should be paired with a set of goals, also necessary to trigger the retaliatory response. We identify at least three distinct goals: one referred to the material action, and the other two related to the influence the avenger wants to exert on the victim's and audience's representations. In fact, revenge is not motivated only by the desire of making the target suffering, but achieving this goal is pivotal to the objective of changing the target's and audience's beliefs about the avenger. What matters is what the others believe about the avenger and not what they are expected to do next time they are required to cooperate, as it is in punishment. In this case, cognitive influencing is aimed at modifying only the beliefs of the target and the onlookers, as depicted in Figure 2.

$$\boxed{(\mathrm{Gx}) \longrightarrow (\mathrm{By})}$$

Figure 1: Cognitive Influencing in Revenge

The avenger's action is driven by the following goals: first, the goal of imposing a suffering on the target; second, the goal of changing the target's beliefs, making her aware that the avenger does not passively accept the aggression and is able and willing to strike back at the aggressor (influencing the target). Finally, there is the goal of changing the beliefs of the onloookers (influencing the onlookers). In revenge the audience plays a crucial role because the damage suffered is not only material, but it usually has a strong symbolic component. Honour, for instance, is an intangible asset that can be threatened by the aggressor and that can be restored only if there is an audience in front of which the retaliatory action is performed and that recognizes that action as an attempt of restoring the initial situation.

This picture needs to be enriched by some additional considerations. First, the avenger can strike back at the aggressor's family or closer relatives, because they share some common traits. Posner (1980) views this issue the other way around: family obligation to retaliate is needed to make the threat of revenge work as a deterrent.

Another relevant issue is the cost-benefit analysis the avenger could carry out in order to choose the best conduct. According to Elster (1990), the retaliator does not calculate pros and cons of her action, but simply react to the offense. In our view, the avenger considers benefits and costs, but in her utility function there is an element that overrule any other consideration, that is the symbolic gain in terms of respect, honour, power, etc. the revenge allows to take.

Kant, I. (1952). The science of right (W. Hastie, Trans.). In R. Hutchins (Ed.), Great books of the Western world: Vol. 42. Kant (pp. 397 446).

[3]Here we are not concerned with the actual punishing behaviour chosen by the actor, but we are interested in investigating which behaviours he considers the most appropriate

Punishment

Punishment is a more controversial phenomenon, as showe by the two following definitions explaining the competing views on it:

> Punishment is the practice of imposing something unpleasant or aversive on a person or animal, usually in response to disobedient or morally wrong behavior (Stanford Encyclopedia of Philosophy, Punishment).

> [...] individuals (or groups) commonly respond to action likely to lower their fitness with behaviour that reduces the fitness of the instigator and discourages or prevents him or her from repeating the same action (Clutton-Brock & Parker, 1995).

According to the first view, punishment is meant to *righting* a wrong, while the second one stresses the influencing aim of punishment, that of discouraging or *preventing* an agent from repeating the same action.

The first one is a *retributive approach* to punishment: a person deserves a punishment that is *proportionate* to the moral wrong committed. Unlike revenge, punishment is proportionate to the offence. Immanuel Kant (Kant, 1952) argued that punishment can never be administered merely as a means for promoting another good and should be pronounced over all criminals proportionate to their internal wickedness (p. 397). Its justification lies in righting a wrong, not in achieving some future benefits. The punisher wants the victim to perceive punishment as a *natural consequence* of offence: the greater the offence, the greater the punishment. We can find such a view either in the lex talionis of early Roman law and in Old Testament and Koran.

In the second view, punishment is assigned a *deterrent effect*: it reduces the frequency and likelihood of future offences. This approach is referred to as utilitarian and is most often attributed to Jeremy Bentham (Bentham, 1962). Based on the rational choice model, deterrence theory works by modifying the *costs* and *benefits* allowed within the circumstances so that the criminal activity becomes an *unattractive* option[4].

According to these two views on punishment, we can say that the punisher is either a *backward-looker* and a *forward-looker*. The punisher aims to *repay* the damage she or someone else suffered with an offence *proportionate* to the one suffered, and to minimize the chance that the attacker will repeat the aggression in the future, thus *deterring* him from further hostility.

This enforcing mechanism, controlling modern societies, is not at all easy to distinguish from revenge, but we suggest that the punisher and the avenger are aimed at modifying the target and the audience's minds in different ways: unlike the latter, the punisher has the *explicit* goal to interrupt the chain of aggressions, with the further effect of preventing blood feuds and giving more stability to the social order.

Into the punisher' mind

Here follows a description of the punisher mental configuration - in terms of beliefs and goals. In order to trigger the punishing response, the offended agent should display the following beliefs (it is not necessary that he has all of them): (1) the damage/offense had a locus of responsibility then liable for punishment, (2) there is a material and/or symbolic damage to be refund and finally (3) the offense/damage will repeated in the future, so that punishment might be useful to avoid such a reiteration.

The above set of beliefs should be paired with a set of goals in order to trigger the punishing response. We identify the following set of distinct goals. More precisely, P aims at imposing an offence proportionate to the one suffered (*retributive goal*), and/or at establishing or mantaining a dominance hierarchy, and at deterring T (and possibly O) from further hostility (*deterrence goal*).

In order for the latter goal to be satisfied, P can employ different means, here we will focus on cognitive influencing. In order to achieve it, P has to act in such a way that the following belief is generated in T's an O's minds "P will harm me/will impose a cost to me, if I do not apply his will that the aggression will not be repeated in the future". This belief, *By*, will possibly activate a *previous* goal of T and O, *Gz* (avoid harm/avoid the costs of punishment), and the interaction between *By* and *Gz* will generate a new *instrumental* goal in T and O's minds, *Gy* (abstaing from repeating the aggression in the future). Social emotions - such as feeling of guilt - play a crucial role in achieving deterrence.

$$\boxed{\text{Gx } ((\text{By}) \longrightarrow (\text{Gy}))}$$

Figure 2: Cognitive Influencing in Punishment

Sanction

A particular case of punishment is that intended to deter future offences in observance no more of the punisher's will, but of a specific (social) *norm*. We refer to this case as (informal) sanction. In our view, a sanction is a particular case of cognitive influencing in which the sanctioner wants to modify the future action of T, making him form *two* beliefs at once: (i) a *normative belief* about the existence of a certain norm, and (ii) and the belief that T did violate that norm. Such a plan, which is incorporated to the act of sanctioning, is aimed at inducing the target to abstain from further offences not only

[4]It has to be said that deterrence can also be achieved through reinforcement learning, as suggested by behaviorism.

in order to avoid the sanction, but in order to *respect* the norm.

In our view, a norm - be it social, legal or moral - is a two-sided, internal (mental) and external (social), object, coming into existence only when it emerges, not only through the minds of the agents involved, but also *within* their minds (see (Conte & Castelfranchi, 2006; "On the Immergence of Norms: a Normative Agent Architecture", 2007). In other words, norms work as such only when agents recognize them and take decisions upon them as norms. Only when the normative, i.e. prescriptive, character of an input is recognized by the agent, that input gives rise to a normative behaviour of that agent. In order for the norm to be satisfied, it is not sufficient that the prescribed action is performed, but it is necessary to comply with the norm because of the *normative goal*, that is, the goal deriving from the recognition and subsequent adoption of the norm. Thus, for a norm-based behaviour to take place, a normative belief has to be generated into the minds of the norm addressees, and the corresponding normative goal has to be formed and pursued.

Unlike the punisher, the sanctioner aims at drawing the target's and the audience's attention on the existence and violation of the norm and on the fact that there is an high rate of surveillance. Our hypothesis is that sanctioning is characterized by a *signalling* function that has the aim of making explicit the casual link between violation and sanction: "you are being sanctioned because you violated that specific norm". Focusing T attention on the fact that the sanction is a consequence of a norm violation, possibly has the effect of encouraging the sanctionee to accept it as an *entitled* act, thus avoiding reiterated aggression (like in revenge) (see also (Bandura, 1991; Xiao & Hauser, 2009).

We also claim that sanction has the further effect, possibly aimed at by the sanctioner, to encourage the target to ground future decisions on *internal* evaluative criteria, established by the norm. This argument needs further elaboration, of course, and in order to test our hypothesis, we plan to conduct a series of laboratory experiments adopting a game-theoretical framework.

While imposing sanctions to them, we often request our children, pupils, etc. to observe the norm for its own sake. Isn't this behaviour irremediably paradoxical? However, it is far from an exception: it appears to be a *pedagogic* strategy rather frequent at least in Westernized societies. In sanction, the penalty is inflicted with the aim to favour a full autonomous compliance with the norm. How is this possible? A plausible explanation calls into question mechanisms of *norm internalization* (Durkheim, 1951; Scott, 1971; Gintis, 2004; Bicchieri, 2006; Bowles & Gintis, 2003). In particular, under conditions and by mechanisms that require specification (see also, ("On norm internalization", 2009), agents internalize external enforcement, converting it into self-enforcement, based on self-esteem and moral emotions, like the feeling of guilt.

Into the sanctioner's mind

In order to trigger the sanctioning response, the agent (S) should believe that (1) a norm has been violated. Regarding the motivations, there are at least two distinct goals that S aims to achieve. The first one is that of generating or reinforcing into the T's and O's minds a normative belief (NB) about the *existence* of a certain norm, and the belief that T did *violate* that norm. We will call this goal, a *pedagogic* goal. The second goal of S is that of making the norm be respected thus avoiding that the violation would happen again (*deterrence goal*) (Gx). In order for the latter goal to be satisfied, S has to act in such a way that the normative goal (I want to comply with the norm) will be activated. Once the normative goal has been activated, the agent will decide whether to adopt it or not. He can decide to obey a norm for instrumenental and terminal reasons. In the former case, the agent comply with the norm only to avoid punishment. In terminal norm adoption, agents decide to comply with the norm because "noms must be obeyed".

Such a plan, which is incorporated to the act of sanctioning, is aimed at inducing the target to abstain from further offences *not* only in order to avoid the sanction, *but* ideally in order to respect the norm. This kind of cognitive influence is the most complex, since it entails not only goals and beliefs but also the Normative Goal.

$$\boxed{\text{Gx } ((\text{NBy}) \longrightarrow (\text{NGy}))}$$

Figure 3: Cognitive Influencing in Sanctioning

To some extent the advantages of sanctions are easily identifiable: norm compliance is expected to be more robust than is the case when conducts are ruled only by external punishment: under ideal conditions agents abstain from violating because they want to respect the norm and not only in order to avoid punishment. Hence, sanctioned agents are expected to be more consistent and compliant than punished and endogenously motivated agents. A further consequence is that agents come to be much better at *defending* the norms: a consequence of the latter prediction is that sanction is decisive, if not indispensable, for *distributed* social control. A positive desired effect of sanction is an overall lowering of the costs associated to the social enforcement.

Conclusions and Future work

In this work we applied cognitive modelling to investigate the mental underpinnings of three different systems of norm enforcement: revenge, punishment and sanction. We argue that these are distinct behaviours people

choose in accordance with what they believe and want, thus entailing specific mental configurations. We also argue that without unraveling these cognitive bases, we can not fully explain complex phenomena like cooperation and altruistic punishment. Moreover, we claim that the transition from one to the other has been allowed by specific cognitive patterns, and suggesting that these mental mechanisms selected among given social structures, at the same time reinforcing and being reinforced by them. This preliminary model will be enriched by a simulation-based study of the different forms of enforcement.

References

(1989). New York: Gjonlekaj Publishing Company.

Amegashie, J. A., & Runkel, M. (2008). The desire for revenge and the dynamics of conflicts. 1–18.

Bandura, A. (1991). Social cognitive theory of moral thought and action. In W. M. Kurtines & J. L. Gewirtz (Eds.), *Handbook of moral behavior and development.* Hillsdale, NJ: Lawrence Erlbauml.

Baron-Cohen, S. (1991). Precursors to a theory of mind: Understanding attention in others. In A. Whiten (Ed.), *Natural theories of mind: Evolution, development and simulation of everyday mindreading.* Oxford: Basil Blackwell.

Bentham, J. (1962). Principles of penal law. In J. Bowring (Ed.), *The works of jeremy bentham.* New York: Russell and Russell.

Bicchieri, C. (2006). *The grammar of society: The nature and dynamics of social norms.* New York: Cambridge University Press.

Boehm, C. (1986). *Blood revenge: The enactment and management of conflict in montenegro and other tribal societies.* University of Pennsylvania Press.

Bowles, S., & Gintis, H. (2003). Origins of human cooperation. In P. Hammerstein (Ed.), *Genetic and cultural origins of cooperation.* Cambridge: MIT Press.

Bowles, S., & Gintis, H. (2004). The evolution of strong reciprocity: Cooperation in heterogeneous populations. *Theoretical Population Biology*, *61*, 17–28.

Cialdini, R., & Goldstein, N. (2004). Social influence: Compliance and conformity. *Annual Review of Psychology*, *55*, 591– 621.

Clutton-Brock, T. H., & Parker, G. A. (1995). Punishment in animal societies. *Nature*, *373*, 209–216.

Conte, R., & Castelfranchi, C. (1995). *Cognitive and social action.* London: University College of London Press.

Conte, R., & Castelfranchi, C. (2006). The mental path of norms. *Ratio Juris*, *19(4)*.

Dennett, D. C. (1987). Reprint of intentional systems in cognitive ethology: The panglossian paradigm defended. *Brain and Behavioral Sciences*, *6*, 343–390.

deQuervain, D., Fischbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A., et al. (2004, August). The neural basis of altruistic punishment. *Science*, *305*, 1254–1258.

Durkheim, E. (1951). *Suicide.* New York: The Free Press.

Elster, J. (1990, July). Norms of revenge. *Ethics*, *100*(4), 862–885.

Fehr, E., & Gachter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Review*, *90(4)*, 980–994.

Fehr, E., & Gachter, S. (2002). Altruistic punishment in humans. *Nature*, *415*, 137–140.

Gintis, H. (2004). The genetic side of gene-culture coevolution: internalization of norms and prosocial emotions. *Journal of Economic Behavior and Organization*, *53*, 57–67.

Henrich, J., & Boyd, R. (2001). Why people punish defectors. weak conformist transmission can stabilize costly enforcement of norms in cooperative dilemmas. *Journal of Theoretical Biology*, *208)*, 79–89.

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., et al. (2006). Costly punishment across human societies. *Science*, *312*(5781), 1767-1770.

Herrmann, B., Thoni, C., & Gachter, S. (2008). Antisocial punishment across societies. *Science*, *319 (5868)*, 1362–1367.

Jensen, K., Call, J., & Tomasello, M. (2007). Chimpanzees are vengeful but not spiteful. *Proceedings of the National Academy of Sciences USA*, *104*, 13046–13050.

Kant, I. (1952). The science of right (w. hastie, trans.). In R. Hutchins (Ed.), *Great books of the western world: Vol. 42. kant* (p. 397-446).

Knutson, B. (2004, August). Sweet revenge? *Science*, *305*, 1246–1247.

Leslie, A. M. (1991). Theory of mind impairment in autism. In A. Whiten (Ed.), *Natural theories of mind: Evolution, development and simulation of everyday mindreading.* Oxford: Basil Blackwell.

Nikiforakis, N., & Engelman, D. (2008). Feuds in the laboratory? a social dilemma experiment. *Research Paper University of Melbourne*, *1058*, 1–31.

On norm internalization. (2009). In *Proceedings of the 6th european social simulation association conference.*

On the immergence of norms: a normative agent architecture. (2007). In *Emergent agents and socialities: Social and organizational aspects of intelligence. papers from the aaai fall symposium.*

Posner, R. A. (1980, January). Retribution and related concepts of punishment. *The Journal of Legal Studies*, *9*(1), 71–92.

Premack, D. G., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Brain and Behavioral Sciences*, *1*, 515–526.

Scott, J. (1971). *Internalization of norms: A sociological*

theory of moral commitment. Englewoods Cliffs, N.J.: Prentice-Hall.

Shackleford, T. (2005). An evolutionary psychological perspective on cultures of honor. *Evolutionary psychology*, *3*, 381–391.

Xiao, E., & Hauser, D. (2009). Avoiding the sharp tongue: Anticipated written messages promote fair economic exchange. *Journal of Economic Psychology*, *30(3)*.

Zaibert, L. (2006). Punishment and revengel. *Law and Philosophy*, *25*, 81–118.

Culturally-Guided Beliefs about Opposing Categories and Their Consequences for Action: The Case of Cooperation and Competition

Josh Keller (jwkeller@ntu.edu.sg)
Division of Strategy, Management & Organisation, Nanyang Technological University
50 Nanyang Ave. S3-B2C-97, Singapore 639798

Jeffrey Loewenstein (jeffrey.loewenstein@mccombs.utexas.edu)
Department of Management, The University of Texas at Austin
1 University Station B6300, Austin, TX 78712 USA

Jin Yan (yanjin@zju.edu.cn)
School of Management, Zhejiang University
Gudun Road, Hangzhou, Zhejiang 310058 China

Abstract

We provide a new approach to how, why and with what results people think about opposing or paradoxical categories. Using a two-part laboratory study, we found differences in whether people in China and the US categorized "attempts to outperform others" as an instance of both competition and cooperation. We call membership in both categories in a paradox *integrative categorization*. We found that Chinese were more likely than in the US to engage in integrative categorization, and that the cultural difference was mediated by differences in lay dialecticism. Finally, we showed behavioral effects: integrative categorization predicted peoples' cooperative behavior after they experienced others' attempt to outperform them.

Keywords: Categories; paradox; coopetition; cooperation; competition; culture; dialecticism; China.

Introduction

Opposing categories, such as *past* and *future*, *nature* and *nurture*, or *habit* and *originality*, can be powerful organizers of cognition and action if they demarcate endpoints of important causal dimensions. Alternatively, they can distort cognition and action if they impose too simple a distinction on a complex space of possibilities. When faced with opposing categories, people could try to determine what the right answer is: what are the properties of the two opposing categories, are they endpoints of a single dimension and hence mutually exclusive, and does that dimension capture important causal forces in a domain of knowledge and activity. Alternatively, people might rely on general reasoning tendencies about how categories relate to one another and on the guidance of their cultural norms. Because most categories that people think about are complex and ambiguous in practice, a logical examination of the properties of an instance and whether those properties do or do not warrant category membership may not be possible. People may instead take predictable shortcuts in their reasoning about categories, with predictable consequences.

The specific case of opposing categories that we examine is the case of cooperation and competition. We study how people in China and the US understand these categories and the consequences for their behavior. Cooperation and competition are important categories. They are central to what it means to interact with others, be it in groups, organizations, industries or societies (e.g., Deutsch, 1949; Johnson & Johnson, 1989; Tyler & Blader, 2000).

Most research defines and operationalizes cooperation and competition as opposites (e.g., Bettenhausen & Murnighan, 1991). According to these views, individuals are either in a competitive situation or in a cooperative situation (Deutsch, 1949), either wanting to compete or wanting to cooperate (McClintock & Allison, 1989), or either acting competitively or acting cooperatively (Komorita & Parks, 1996). All of these views predict that the absence of cooperation indicates the presence of competition (and vice-versa) and treat the co-occurrence of both cooperation and competition as a contradiction.

Recently, an alternative theoretical perspective has emerged that conceptualizes cooperation and competition not as opposites, but distinct dimensions, which allows cooperation and competition to co-occur ("coopetition;" Brandenburger & Nalebuff, 1996; Tsai, 2002). For example, an individual might have a general disposition towards wanting to help others (a cooperative personality). At the same time, the individual might also like to be the most highly rewarded (a competitive personality; Xie, Chen, Yu, & Chang, 2006). All of these views predict that knowing about the presence or absence of cooperation will be uninformative regarding the presence or absence of competition.

There is a third logical possibility that no theory has yet defended but that is possible and empirically observable (Keller & Loewenstein, 2010). This is that cooperation and competition can at least partially overlap. At least in some cases, the presence of cooperation implies the presence of competition.

We provide a general framework for understanding how opposing or paradoxical categories can be related. We treat cooperation and competition as cultural categories (Atran, Medin, & Ross, 2005; Keller & Loewenstein, 2010; Sperber & Hirschfeld, 2004). Cultural categories are social conventions (Millikan, 2005) generated by cultural groups

for labeling and grouping sets of objects, practices, actors and other socially experienced examples (Douglas, 1986; Hannan, Pólos, & Carroll, 2007). Social conventions can also guide how people think about the relationship between categories. The words that are used to label cultural categories have semantic relationships (Lyons, 1977). If two words have antonymic semantic relationships (Jones, 2002; Murphy, 2003), this would imply a social convention that the named categories are in opposition.

The presence of these two kinds of social conventions—conventions about category membership and conventions about antonymy provides leeway for culture to shape which kind of convention has priority. If antonymy conventions dominate, then category membership conventions should conform, maintaining the distinction between categories by making category membership mutually exclusive. If category membership conventions dominate, then this allows paradoxes to be integrated. This is because an example that has features representative of two categories can be a member of both categories (Rosch, 1978; Smith & Medin, 1981). If those two categories happen to be antonyms, this example's dual categorization, which we call *integrative categorization*, represents a general account of how to integrate paradoxes. This is novel; discussions of paradoxes and paradoxical cognition (e.g., Miron-Spektor & Argote, 2008; Smith & Tushman, 2005) have claimed paradoxes can be integrated but not analyzed how in general this can be done.

For example, the words "work" and "play" are perceived as antonyms (Glynn, 1994). This is so regardless of the specific activities that constitute work or play, which might even have overlapping features (Jones, 2002). If an "engaging task" has features of both work and play, and work and play are antonyms, then engaging tasks establish a categorization paradox. Forcing engaging tasks to be categorized as either work or play would maintain the work-play distinction. Allowing engaging tasks to be categorized as both work and play (i.e., the integrative categorization of engaging tasks) would integrate the distinction.

We test whether people believe that cooperation and competition are antonymic cultural categories. Previous literature has found that antonymic patterns are often consistent across national cultures (Raybeck & Herrmann, 1996). Thus, our focus is on whether people engage in integrative categorization for cooperation and competition.

Testing whether people integrate the categories of cooperation and competition through overlapping category membership requires identifying an act with features of both categories. According to Tyler and Blader (2000), the key feature of a cooperative act is that it is an attempt to benefit the group. According to Johnson and Johnson (1989), the key feature of a competitive act is that it is an attempt to attain a higher relative position than others. Assuming these are accurate descriptions of conventional lay beliefs as well, then an individual's attempt to outperform others within a team or organization has features of both cooperation and competition. This act represents an attempt to gain a higher status and an increase in effort on group tasks. Therefore, it is possible to categorize an attempt to outperform others as an instance of both competition and cooperation. This act provides an opportunity for integrative categorization.

To be clear, integrative categorization does not require that all members of one category also be categorized as members of the other category. For example, attempts to undermine others are attempts to gain higher status (and hence representative of competition) by harming others, which is detrimental to group outcomes (and hence representative of non-cooperation). There is no need for integrative categorization to include attempts to undermine others as instances of both cooperation and competition.

Our account suggests that whether people categorize attempts to outperform others as an instance of both cooperation and competition is at least in part a function of social conventions. Social conventions can be generated at different social levels, yet for fundamental social categories like cooperation and competition, the social conventions are likely to be generated at the level of the society because the categories are used in many social contexts (Keller & Loewenstein, 2010).

Societies appear to differ in their approaches to paradoxical categories. Theories of paradoxes have pointed to East Asian philosophy, with its emphasis on holism, dynamism and a "middle-way," as fostering a societal level tendency towards integrating paradoxes (Chen, 2008; Eisenhardt, 1988). Integrating paradoxes is exemplified in the 阴阳(Yin-Yang) symbol found in the classic text 易经 (Yi Jing, Book of Changes; Wilhelm & Baynes, 1968) demonstrating that black and white are part of one whole. Integrating paradoxes is a prominent feature in Laozi's 道德经 (Dao de jing; Lao, 1997). In China, Japan, Korea and Vietnam, these texts have long been canonized (Schwartz, 1985), and the integration of paradoxes has long permeated stories, proverbs and other commonplace cultural artifacts within East Asian societies (Peng & Nisbett, 1999). As supporting evidence, cultural psychology research has found tendencies toward integration of paradoxes among lay people in East Asia, establishing societal-level lay theories on contradiction and change, or lay dialecticism (Norenzayan, Smith, Kim, & Nisbett, 2002; Spencer-Rodgers, Boucher, Mori, Wang, & Peng, 2009). Although dialecticism is present in Western philosophy (Walton 1990), its influence on lay people is less pervasive (Samson 2004), suggesting that societal-level cultural conventions that emphasize the integration of paradoxes are weaker in Western societies.

A heightened exposure to lay dialecticism encourages a tolerance of contradictions in people's general views of their self and their social relations (Spencer-Rodgers et al., 2009). Therefore, a tendency towards lay dialecticism could foster integrative categorization generally, and more specifically, could foster integrating the cultural categories of cooperation and competition (such as by categorizing attempts to outperform others as instances of both

cooperation and competition). So, people exhibiting a greater degree of lay dialecticism should be more likely to believe that even if cooperation and competition are generally opposites, it is possible that an act can be both cooperative and competitive because there are situations where contradictions can occur.

Taken together, the preceding discussion leads us to predict that national culture should influence people's predilection for lay dialecticism. Lay dialecticism, in turn, should influence people's tendency for integrative categorization—specifically, categorizing attempts to outperform as an instance of both cooperation and competition.

Integrative categorization should influence behavior. Categories serve as cognitive mediators between settings and actions (Keller & Loewenstein, 2010; Markman & Ross, 2003). Individuals use categories to interpret the type of setting they are in and the actions of others, and then use their interpretations to select appropriate responses (March, 1994; Smith, 1989). The interpretation and reaction to settings and prior actions is particularly important for cooperation, because cooperation requires reciprocity (Koster & Sanders, 2006). Reciprocity implies responding with an action of the same kind (Gouldner, 1960), that is, with a response drawn from the same category. Therefore, individuals' propensity to act cooperatively is contingent on whether they categorize the setting and others' actions as cooperative. If individuals categorize others' actions as non-cooperative, they are unlikely to respond with a cooperative act (Andersson & Pearson, 1999), even when the behavior does not have a material impact on the individual (Stanne et al., 1999). As a result, integrative categorization of attempts to outperform should increase people's likelihood of responding to attempts to outperform by cooperating.

Methods

Participants

Participants were 94 US undergraduates (62% female, mean age 20.3 years) and 100 Chinese undergraduates (65%, 21.2 years).

Part 1 procedure and materials

The study consisted of two parts, separated by 1-2 weeks. During the first part, participants completed computer-based questionnaire measures for lay dialecticism (from Spencer-Rodgers et al, 2009), integrative categorization (based on Keller & Loewenstein, 2010), antonymy (based on Herrmann & Conti, 1979), self-construal measures (as control variables) and demographics. All original materials were developed in English, translated into Chinese and back translated into English; tests of the back-translated versions showed comparable results.

The key new measure is the integrative categorization measure. Participants rated 25 behaviors three times; whether they indicated a strong or weak indicator of (1) cooperation, (2) commitment (as a foil), and (3) competition. Four of these 25 behaviors were key, because they represented attempts to outperform others. They were: 1) "A team member attempts to outperform other team members", 2) "A team member gauges others' performance and makes sure that the he or she is doing better that the others", 3) "A team member tries to get the quality of the his or her work to be better than the quality of others' work", and 4) "A team member tries to make sure that he or she isn't outdoing others in the team" (reverse-coded). These behaviors were consistently categorized by people in China and the US as indicating competition (α=.81, M=4.06, SD=.55). There was considerable variance as to whether these items indicated cooperation (α=.73), and hence we used their cooperation ratings as our measure of integrative categorization.

The remaining behaviors were mostly banal instances taken from prior research on lay beliefs about cooperation (Keller & Loewenstein, 2010) used as filler items so there would not be undue attention on attempts to outperform others. The exception was that we also included behaviors representing attempts to undermine others as a foil for attempts to outperform others. We found that people in both China and the US consistently rated attempts to undermine others as competitive (M=4.06) and non-cooperative (M=1.52). Thus, finding that some people's ratings indicate integrative categorization of attempts to outperform others should indicate their specific beliefs about attempts to outperform others rather than a general response bias.

Part 2 procedure and materials

Participants engaged in a group brainstorming task to facilitate the development of group entitativity (Campbell, 1958; Kramer, Kuo & Dailey, 1997). We assessed participants' ratings of how strongly they felt they were a group and part of a group as manipulation checks, and found that these ratings were generally high, and also that they did not account for the core findings we present later.

Participants next moved to a computer for a simulated group sales task. During the simulation, each participant managed a cart selling tea. Two simulated team members managed two other carts. During the simulation, participants had eight opportunities to share information with their teammates. The number of times they did so was our measure of cooperation.

Lastly, participants completed a post-task questionnaire. This included a manipulation check that showed that participants believed their teammates in the simulation were the people with whom they had completed the brainstorming task. The four participants who did not believe so were dropped from the analysis.

The participants in China and and the US engaged in one of two versions of the tea sales simulation. In the outperform condition, participants received messages from their teammates stating that they wanted the team to do well and that they wanted to perform the best. During the simulation, the teammates constantly checked on the others' performance (this act was made visible in the interface). In a

baseline condition, bland messages were sent and little signs of checking on the others' performance occurred. A post-task manipulation check showed that those in the outperform condition stated they experienced their teammates attempting to outperform them more so than those in the baseline condition, ($M_{outperform}$ = 5.45, SD= 1.15 $M_{baseline}$.= 2.83, SD=1.35, t(199)= 13.26, p<.001). Finally, we note that we used an unbalanced design, placing more participants in the outperform condition because at issue is whether there would be a difference in cooperation rates in the outperform condition. We expected (and found) no difference in the baseline condition.

Results

As shown in Table 1, we found China-US differences in lay dialecticism (t(143)=8.75, p<.001); integrative categorization (t(143)=10.50, p<.001); and cooperative behavior in the outperform condition (t(143)= 4.41, p<.001), but not the baseline condition. We also found China and US consistency in believing cooperation and competition to be antonyms.

Table 1: Descriptive Statistics.

	China (n=100) Mean (SD)	**US (n=94)** Mean (SD)
Lay Dialecticism	4.33 (0.49)	3.57 (0.58)
Integrative Categorization	3.70 (0.47)	2.68 (0.67)
Cooperative Behavior in Outperform Condition	5.12 (2.28)	3.54 (2.03)
Cooperative Behavior in Baseline Condition	4.95 (1.85)	4.95 (1.68)
Antonymy of "Cooperation" and "Competition"	1.84 (1.12)	1.88 (0.90)
Independent Self-Construal	4.49 (0.63)	5.31 (0.74)
Group-Collective Self-Construal	5.23 (0.82)	4.63 (0.99)
Perceived Task Difficulty	4.45 (1.56)	4.65 (1.61)

We used stepwise linear regression models to examine relations among variables just for those in the outperform condition. First, we found that lay dialecticism predicted integrative categorization (B=.48; SE=.11; p<.05). National culture also predicted integrative categorization (B=1.09; SE=.12; p<.05). To examine lay dialecticism as a mediator of the national culture effect, we ran a bootstrapped test of an indirect effect of national culture on integrative categorization through lay dialecticism (Preacher, Rucker, & Hayes, 2007). The mean indirect effect was 0.12 (95% CI: 0.01-0.22), p<.05, providing evidence of mediation. Therefore, the impact of national culture on integrative categorization can be at least partially attributed to differences in lay dialecticism.

Second, we found that lay dialecticism predicted cooperative behavior (B=1.09; SE=.31; p<.05). Integrative categorization also predicted cooperative behavior (B=.95; SE=.31; p<.05). A bootstrapped test of the indirect effect of lay dialecticism on cooperative behavior through integrative categorization found that the mean indirect effect was 0.13 (95% CI: 0.02-0.35), p<.05, providing evidence of mediation. Therefore, the impact of lay dialecticism on cooperative behavior can be at least partially attributed to differences in integrative categorization.

Analysis of control variables showed that the nationality to dialecticism to integrative categorization to cooperative behavior pathway was not explained away by alternative factors. For example, we found national differences in independent self-construal and group-collective self-construal, but the mediation analyses included these variables—as well as age, gender, and subjective ratings of task difficulty—as controls and still found the predicted patterns.

Discussion

We found US and Chinese consensus that cooperation and competition are antonyms, providing evidence of a coopetition paradox. We introduced the concept of integrative categorization as a specific means of integrating a paradox. We found cultural and individual differences in the integrative categorization of attempts to outperform as instances of competition and cooperation. We further found predictable consequences of integrative categorization on people's cooperative behaviors in a group simulation task. Therefore, we advance research on categories and on cooperation and competition.

We found societal-level differences between the US and China, suggesting that integrative categorization is culturally conditioned. The cultural differences were attributable to lay dialecticism differences. This implies that the national culture difference in integrative categorization was due to broad cultural belief systems about how to think about contradictions and change. The broad cultural tendencies towards lay dialecticism, by influencing integrative categorization, influenced people's reactions to others' behaviors. Therefore, the results suggest that culturally-influenced lay beliefs about paradoxes establish broad conditions that make particular behaviors more or less likely to occur. Specifically, lay dialecticism makes integrative categorization more likely, which in the case of cooperative and competitive behaviors, makes coopetition and the sustaining of cooperation within a group more likely to occur.

We note here that the data pattern described here has turned out to be robust. Subsequent research manipulating participants' social motivations (whether they are trying to maximize their own outcomes, group outcomes, or both) has shown that motivation effects are distinct from the dialecticism and integrative categorization effects that are our focus. The results are also robust when controlling for participants' performance on the simulation task.

Our results have implications for research on categories. There is growing interest in how categories are used (Markman & Ross, 2003), in complex categories (Gentner & Kurtz, 2005), and in how categories relate to each other (Goldstone, 1996; Loewenstein & Gentner, 2005). We contribute to these streams of category research by showing that people's decisions about category membership are not entirely a function of the features of the instance. Membership in one category can suppress the possibility of acknowledging membership in another category. Further, this suppression is a function of general beliefs about contradiction and change that are acquired through cultural experience exogenous to the immediate social context and the particular categories at hand. Thus, our study demonstrates that research on how people think about and use multiple categories is not only a matter of the features of exemplars, but also subject to broad and predictable cultural influence.

Our results also have implications for research on cooperation and competition. It is well established that cooperation can facilitate effective social outcomes (Campion, Medsker, & Higgs, 1993; Kogut & Zander, 1992). It is less well established but also supported that competition can increase individual effort towards collective goals, and thereby also generate effective social outcomes (Luo et al., 2006). Finally, it is also established that many social situations involve mixed motives (Komorita & Parks, 1996). The results from this study suggest that integrative categorization is important to making effective use of the positives of both cooperation and competition to advance social outcomes. People who engaged in integrative categorization were more likely to maintain cooperation and less likely to treat cooperation and competition as "trade-offs." Accordingly, people with beliefs that facilitate the integration of paradoxes may be more suitable for jobs with paradoxical situations, such as working in teams with mixed motive incentive structures. In teams with mixed motives, members with higher overall propensities for lay dialecticism and integrative categorization may perform better than teams whose members have low propensities or a mixture of propensities for lay dialecticism and integrative categorization. They might better take advantage of the positives aspects of both cooperation and competition. More broadly, the implication is that by examining categories central to social interaction, we can improve our ability to predict and provide prescriptions for obtaining positive social outcomes.

To conclude, how people think about and use specific categories can be influenced by broader cultural tendencies as to how to address oppositions and paradoxes. This is consequential; we showed that cultural tendencies to maintain separation between categories, rather than to seek out ways to integrate them, can lead to failures to support social opportunities for mutual gain.

Acknowledgments

We are grateful for the support of the McCombs School of Business, its Center for International Business Education and Research, and its Herb Kelleher Center for Entrepreneurship. We would also like to thank Caroline Bartel, George Huber, Martin Kilduff, Bradley Love and Kyle Lewis for their constructive feedback.

References

Atran, S., Medin, D. L., & Ross, N. O. 2005. The Cultural Mind: Environmental Decision Making and Cultural Modeling Within and Across Populations. *Psychological Review*, 112(4): 744-776.

Bettenhausen, K. L., & Murnighan, J. K. 1991. The Development of an Intragroup Norm and the Effects of Interpersonal and Structural Challenges. *Administrative Science Quarterly*, 36(1): 20-35.

Brandenburger, A., & Nalebuff, B. 1996. *Co-Opetition : A revolution mindset that combines competition and cooperation : the game theory strategy that's changing the game of business*. New York: Doubleday.

Campion, M. A., Medsker, G. J., & Higgs, A. C. 1993. Relations between work group characteristics and effectiveness: implications for designing effective work groups. *Personnel Psychology*, 46(4): 823-850.

Chen, M.-j. 2008. Reconceptualizing the Competition-Cooperation Relationship: A Transparadox Perspective. *Journal of Management Inquiry*, 17: 288.

Deutsch, M. 1949. A theory of cooperation and competition. *Human Relations*, 2: 199-231.

Eisenhardt, K. M. 1988. Agency- and institutional- theory explanations: The case of retail sales compensation. *Academy of Management Journal*, 31(3): 488-511.

Gentner, D., & Kurtz, K. (2005). Relational categories. In W. K. Ahn, R. L. Goldstone, B. C. Love, A. B. Markman & P. W. Wolff (Eds.), *Categorization inside and outside the lab* (pp. 151-175). Washington, DC: APA.

Glynn, M. A. 1994. Effects of work task cues and play task cues on information processing, judgment, and motivation. *Journal of Applied Psychology*, 79(1): 34-45.

Goldstone, R. L. (1996). Isolated and Interrelated Concepts. *Memory & Cognition*, 24, 608-628

Gouldner, A. W. 1960. The Norm of Reciprocity: A Preliminary Statement. *American Sociological Review*, 25(2): 161-178.

Hannan, M. T., Pólos, L., & Carroll, G. 2007. *Logics of organization theory: audiences, codes, and ecologies*: Princeton.

Herrmann, D., G. Conti, et al. 1979. "Comprehension of antonymy and the generality of categorization models." *Journal of Experimental Psychology: Human Learning and Memory*, 5(6), 585-597.

Johnson, D. W., & Johnson, R. 1989. *Cooperation and competition: Theory and research*. Edina, MN: Interaction book.

Jones, S. 2002. *Antonymy: a corpus-based perspective*. London: Routledge.

Keller, J., & Loewenstein, J. 2010. The cultural category of cooperation: a consensus model analysis of the united states and china. *Organization Science*, (in press).

Kogut, B., & Zander, U. 1992. Knowledge of the firm, combinative capabilities, and the replication of technology. *Organization Science*, 3(3): 383-397.

Komorita, S. S., & Parks, C. D. 1996. *Social Dilemmas*. Boulder, CO: Westview.

Koster, F., & Sanders, K. 2006. Organisational citizens or reciprocal relationships? An empirical comparison. *Personnel Review*, 35(5): 519-537.

Lao, T. 1997. *Tao Te Ching* (G.-F. Feng, Trans.): Vintage.

Loewenstein, J., & Gentner, D. (2005). Relational language and the development of relational mapping. *Cognitive Psychology*, 50, 315-363.

Luo, X., Slotegraaf, R. J., & Pan, X. 2006. Cross-Functional "Coopetition": The Simultaneous Role of Cooperation and Competition Within Firms. *Journal of Marketing*, 70(2): 67-80.

Lyons, J. 1977. *Semantics*. Cambridge: Cambridge University Press.

March, J. G. 1994. *A primer on decision making*. New York: Free Press.

Markman, A. B., & Ross, B. H. 2003. Category use and category learning. . *Psychological Bulletin*, 129(4): 592-613.

McClintock, C. G., & Allison, S. T. 1989. Social Value Orientation and Helping Behavior¹. *Journal of Applied Social Psychology*, 19(4): 353-362.

Millikan, R. G. 2005. *Language: a biological model* OUP.

Miron-Spektor, E., & Argote, L. 2008. The effect of paradoxical cogntion on individual and team innovation. *Academy of Management Proceedings*: 1-6.

Murphy, L. 2003. *Semantic Relations and the Lexicon*. Cambridge, UK: Cambridge University Press.

Norenzayan, A., Smith, E. E., Kim, B. J., & Nisbett, R. E. 2002. Cultural preferences for formal versus intuitive reasoning. *Cognitive Science: A Multidisciplinary Journal*, 26(5): 653 - 684.

Peng, K., & Nisbett, R. E. 1999. Culture, Dialectics, and Reasoning About Contradiction. *American Psychologist*, 54(9): 741.

Preacher, K. J., Rucker, D. D., & Hayes, A. F. 2007. Addressing Moderated Mediation Hypotheses: Theory, Methods, and Prescriptions. *Multivariate Behavioral Research*, 42(1): 185-227.

Raybeck, D., & Herrmann, D. 1996. Antonymy and semantic relations: the case for a linguistic universal. *Cross-cultural research*, 30(2): 154-183.

Rosch, E. 1978. Principles of Categorization. In B. Rosch, & E. Lloyd (Eds.), *Cognition and Categorization*: 27-47. Hillsdale, NJ: Erlbaum.

Samson, A. 2004. "Contradictions in counter-intuitive beliefs and naïve dialecticism", *Journal of Cognition and Culture*, 4(2): 373-390.

Schwartz, B. I. 1985. *The World of Thought in Ancient China.*: Harvard University Press.

Smith, E. E. 1989. Concepts and Induction. In M. I. Posner (Ed.), *Foundations of Cognitive Science*: 501-526. Cambridge, MA: MIT Press.

Smith, E. E., & Medin, D. L. 1981. *Categories and concepts.* . Cambridge, MA: Harvard University Press.

Smith, W. K., & Tushman, M. L. 2005. Managing Strategic Contradictions: A Top Management Model for Managing Innovation Streams. *Organization Science*, 16(5): 522-536.

Spencer-Rodgers, J., Boucher, H., Mori, S., Wang, L., & Peng, K. 2009. The Dialectical self-concept: Contradiction, Change, and Holism in East Asian cultures. *Personality and Social Psychology Bulletin*, 35: 29-44.

Sperber, D., & Hirschfeld, L. A. 2004. The cognitive foundations of cultural stability and diversity. *Trends in Cognitive Sciences*, 8(1): 40-46.

Stanne, M. B., Johnson, D. W., & Johnson, R. T. 1999. Does competition enhance or inhibit motor performance: A meta-analysis. *Psychological Bulletin*, 125(1): 133-154.

Tsai, W. 2002. Social Structure of "Coopetition" Within a Multiunit Organization: Coordination, Competition, and Intraorganizational Knowledge Sharing. *Organization Science*, 13(2): 179-190.

Tyler, T. R., & Blader, S. L. 2000. *Cooperation in groups : procedural justice, social identity, and behavioral engagement*. Philadelphia, PA: Psychology Press.

Walton, D. 1990. "What is Reasoning? What Is an Argument?" *Journal of Philosophy*, 87(8), 399-419.

Wilhelm, R., & Baynes, C. 1968. *The I ching; or, Book of changes*. London, UK: Routledge.

Xie, X. F., Chen, X. P., Yu, Y. Y., & Chang, S. Q. 2006. Cooperation, competition, and coopetition: Scale development and validation. *International Association of Chinese Management Research Conference Proceedings.*

Theories of God: Explanatory Coherence in a Non-Scientific Domain

Andrew Shtulman (shtulman@oxy.edu)
Department of Psychology, Occidental College
1600 Campus Rd., Los Angeles, CA 91106

Abstract

Public representations of God range from the highly anthropomorphic to the highly abstract, and the present study explored whether differences in the interpretation of those representations are correlated with differences in one's religious beliefs and religious practices more generally. American adults of varying ages and religious backgrounds completed a questionnaire that probed their beliefs about a wide range of religious matters, including prayer, ritual, worship, sin, cosmogenesis, anthropogenesis, angels, Satan, Heaven, and Hell. Participants were divided into two groups based on their propensity to anthropomorphize God in a property-attribution task, and their responses were analyzed for internal consistency. Overall, the two groups exhibited explanatorily coherent, yet qualitatively different, patterns of beliefs and practices – patterns interpreted contrastively as a "humanistic theology" and an "existential theology." These findings suggest that individuals' religious beliefs are organized in a theory-like manner despite their lack of direct perceptual support.

Keywords: Intuitive theories; religious cognition; conceptual representation; cultural transmission; explanation

Introduction

Belief in the existence of a divine being is prevalent both within and across cultures (Brown, 1991). This belief is particularly prevalent in the US, where the percent of individuals who report holding such a belief has hovered around 95% for the last six decades (Gallup, 2003). Theistic beliefs are of potential interest to cognitive developmentalists, as they present a challenge to standard, constructivist models of knowledge acquisition (e.g., Piaget, 1954; Gopnik & Meltzoff, 1997). Such models construe knowledge as a product of direct observation and exploration of the physical world, yet "knowledge" of God is rarely (if ever) acquired in this manner. Rather, individuals must learn about God from the art, literature, and discourse of their culture. How individuals make sense of such public representations is the topic of investigation in the present study.

The task of interpreting public representations of God is by no means trivial, for these representations range from the highly anthropomorphic (e.g., "heavenly father," "divine ruler," "intelligent designer") to the highly abstract (e.g., "first cause," "unmoved mover," "universal spirit"). As a group, they paint a picture of God that is neither consistent nor coherent. For instance, God is commonly said to listen to prayers, yet an omniscient being would already know the content of those prayers. Likewise, God is commonly said to have created man in his image, yet an omnipresent being presumably has no "image."

One way to resolve the tension between anthropomorphic and nonanthropomorphic representations of God is to treat the anthropomorphic representations as metaphors and the nonanthropomorphic representations as literal descriptions. Barrett & Keil (1996) investigated this possibility by comparing participants' self-professed beliefs about God to the kinds of beliefs revealed in a story-recall task. Although all participants claimed that God is omniscient and omnipresent when asked directly, many participants drew anthropomorphic inferences on the story-recall task that contradicted such claims. For instance, participants frequently mistook the statement "God was pleased by seeing the girl put the bird in its nest" for the statement "God was aware of the girl's deed and was pleased by it" in even though the former, but not the latter, implies that God must perceive an event in order to be aware of its occurrence. Likewise, participants frequently mistook the statement "When the woman awoke, God had already left" for the statement "When she woke, she saw no one" even though the former, but not the latter, implies that God occupies a discrete location in space.

These findings suggest that anthropomorphic descriptions of God do, in fact, influence the way individuals reason about God's actions and abilities, particularly in a narrative context. Still, not all the participants in Barrett and Keil's study anthropomorphized God to the same extent. Indeed, participants' accuracy in differentiating anthropomorphic descriptions of God from nonanthropomorphic descriptions ranged from 27% to 91%. Consistent with this finding, several other studies have documented significant differences in the anthropomorphization of God (e.g., Bassett & Williams, 2003; Shtulman, 2008; Trimeche, Vinsonneau, & Mullet, 2006), yet it is unclear how to interpret those differences in light of the commonly held view that what people say they believe about God is not necessarily true of what they actually believe (Boyer, 2003; Slone, 2004). One interpretation is that they are artifactual, reflecting nothing more than variation in participants' understanding of, or engagement with, the task at hand. Another (more interesting) interpretation is that they are symptomatic of variation in how to make sense of God's public representations as a whole, with anthropomorphic responses reflecting a fundamentally different interpretation of religious claims than nonanthropomorphic responses.

One reason to suspect the latter – i.e., that different God concepts are correlated with different patterns of religious belief – is that correlations of this nature have been documented in many other domains of knowledge. For instance, different concepts of matter are correlated with different beliefs about mass, weight, and density (Smith, Snir, &

Grosslight, 1992); different concepts of force are correlated with different beliefs about acceleration, momentum, and inertia (McCloskey, 1983); and different concepts of evolution are correlated with different beliefs about adaptation, speciation, and extinction (Shtulman, 2006). These correlations have been interpreted as evidence that our knowledge of natural kinds is organized in self-consistent, causal-explanatory networks (Carey, 1985; Keil, 1989; Murphy & Medin, 1985). Whether or not our knowledge of "supernatural kinds" is organized an a similar manner is an open question.

Previous research on religious cognition has not specifically looked for correspondences between God concepts and overall theologies. Instead, this research has focused either on explicating the content of God concepts apart from their associated beliefs (Bassett & Williams, 2003; Trimeche, Vinsonneau, & Mullet, 2006) or on comparing children's God concepts to those of adults (Barrett, Richert, & Driesenga, 2001; Gimenez-Dasi, Guerrero, & Harris, 2005). In contrast, the present study sought to determine whether variation in adults' God concepts tracks variation in their religious beliefs and religious practices more generally. Such a finding would imply not only that resolving the ambiguity inherent in God's public representations has different consequences for different individuals but also that religious beliefs, like scientific beliefs, are organized in a theory-like manner.

Method

Participants

Thirty-two American adults, ranging in age from 18 to 46, were recruited from the study pool of a large, urban university and were compensated for their participation either monetarily or with course credit in an introductory psychology class. Participation was restricted to individuals who believed in the existence of God, though participants varied widely in their particular religious affiliations: 34% self-identified as Protestant, 16% Catholic, 9% Unitarian, 6% Jewish, 6% Buddhist, 3% Muslim, and 25% claimed not to be affiliated with any particular religion.

Procedure

Each participant completed a six-part questionnaire that probed their beliefs about (1) God's appearance and occupation, (2) God's relationship to nature, (3) God's relationship to humankind, (4) supernatural beings associated with God (angels and Satan), (5) supernatural locations associated with God (Heaven and Hell), and (6) prayer, ritual, and worship. The particular questions on each topic are presented in combination with participants' responses in the Results section. Questions for which participants' responses exhibited little to no variation were omitted from these analyses for the sake of brevity.

Participants' religious beliefs were analyzed in relation to their God concepts as measured by a property-attribution task. In this task, participants were asked to decide whether God could or could not be attributed twelve human properties: "dreams," "sees," "talks," "thinks," "eats," "grows," "sleeps," "sneezes," "gets cold," "gets wet," "sits," and "stretches." The first four properties were intended to exemplify human psychological properties, the middle four human biological properties, and the last four human physical properties. The properties were arranged in alphabetical order, and the task itself was sandwiched between questions about God's occupation and questions about God's existence in the first part of the questionnaire.

Participants' responses to the open-ended questions were coded using the schema presented in Table 1. All responses were coded by two independent judges. Overall agreement between judges was 90%, and all disagreements were resolved through discussion.

Results

Beliefs about God

The first part of the questionnaire probed participants' beliefs about God's appearance and occupation. It also probed participants' beliefs about God's anthropomorphic properties, as assessed by the aforementioned property-attribution task. Replicating previous research (Shtulman, 2008), participants attributed more psychological properties to God (M = 2.8, SD = 1.1) than biological properties (M = 0.7, SD = 1.3) or physical properties (M = 0.7, SD = 1.1), and they varied widely in the total number of properties attributed (range = 0 to 12).

For the purposes of data analysis, participants were split into two groups: those who attributed zero to three human properties to God (n = 16) and those who attributed four to twelve human properties to God (n = 16). The first group were labeled "weak anthropomorphizers" and the second "strong anthropomorphizers." Note that the labels "strong" and "weak" denote relative, not absolute, amounts of anthropomorphism, for even the strong anthropomorphizers typically attributed fewer than half of the 12 properties to God. Still, 96% of the strong anthropomorphizers attributed at least one biological or physical property to God, whereas only 6% of the weak anthropomorphizers did. Thus, strong anthropomorphizers differed from weak anthropomorphizers not only in the *number* of properties attributed to God but also in the *type* of properties attributed.

Participants answered an additional five questions about God's nature and existence. In response to the question "What does God look like?," 56% of participants claimed that God has a definite physical appearance (e.g., "looks like a human being"), and 44% claimed that God's appearance is either unknown or unknowable (e.g., "in our limitation as humans we cannot conceive of what God looks like"). In response to the question "What does God do?", 69% claimed that God intervenes in human affairs (e.g., "he guides, chastises, advises, sacrifices, reminds and loves"), and 31% claimed that God's occupation is either unknown or unknowable (e.g., "God is omnipresent, so he does not 'do' anything in the conventional sense"). In response to the

Table 1: The percentage of weak anthropomorphizers (WA) and strong anthropomorphizers (SA) who professed each of the following beliefs and practices, and the strength of association (ϕ) between being a strong anthropomorphizer and professing each belief/practice ($df = 1$ for all statistical comparisons).

Topic	Professed belief/practice	WA	SA	ϕ
God	God has a physical appearance.	19	94	.76**
	God intervenes in human affairs.	50	88	.41*
	God answers prayers.	38	75	.38*
	God's existence is discernible from experience.	25	63	.38*
	God's existence is 100% certain.	13	63	.52**
Cosmogenesis	God created the universe as is.	25	63	.39*
	God created the universe via the Big Bang.	38	25	-.14
	God did not create the universe.	38	6	-.39*
Anthropogenesis	God created human beings as is.	6	56	.54**
	God created human beings via evolution.	50	13	-.41*
	God did not create human beings.	44	31	-.13
Problem of evil	God is not omnipotent and/or omnibenevolent.	50	25	-.26
	Suffering is part of the human condition.	31	0	-.43**
	God uses suffering to teach or to punish.	19	69	.50*
Problem of sin	God is not omniscient and/or judgmental.	69	25	-.44*
	God gave humans the freedom to disobey him.	25	69	.44*
Angels	Angels exist.	50	81	.33
	Angels have biological or physical properties.	6	56	.54**
	Angels have a physical appearance.	25	69	.44*
	Angels act as God's helpers.	19	56	.39*
Satan	Satan exists.	44	69	.25
	Satan has biological or physical properties.	13	63	.52**
	Satan has a physical appearance.	19	63	.45*
	Satan acts as God's enemy.	13	50	.41*
Heaven	Heaven exists.	50	81	.33
	Heaven occupies a discrete location in space.	13	38	.29
	Heaven has a physical appearance.	19	56	.39*
	Human activities continue in Heaven.	25	69	.44*
Hell	Hell exists.	38	81	.45*
	Hell occupies a discrete location in space.	13	38	.29
	Hell has a physical appearance.	19	56	.44*
	Human activities continue in Hell.	19	69	.50**
Prayer	Prays at least occasionally.	63	69	.07
	Prays once or more per day.	13	44	.35*
Worship	Attends religious services at least occasionally.	56	88	.35*
	Attends religious services once or more per week.	19	44	.27
	Belongs to an organized religion.	75	75	.00
	Belongs to a denomination of Christianity.	38	63	.25
Education	Acquired beliefs from a religious authority.	25	63	.38*
	Acquired beliefs from family.	38	31	-.07
	Acquired beliefs from scholarship, reflection.	38	6	-.38*

question "Does God answer prayers?," 56% claimed that he does, and 44% claimed that he does not. In response to the question "How do you know that God exists?," 44% provided an "experiential" justification (e.g., "I can feel him in my soul"), 41% provided an "intellectual" justification (e.g., "acknowledging a higher power feels like a good way to order the universe"), and 16% simply appealed to faith. Finally, in response to the question "How confident are you, on a scale from 1 (not confident) to 7 (100% confident), that God exists?," participants provided an average confidence rating of 5.2 ($SD = 2.0$), and a modal confidence rating of 7.

Displayed in Table 1 are the percentage of weak and strong anthropomorphizers who provided the five most common responses summarized above. Accompanying these percentages are a measure of the association between being a strong anthropomorphizer and providing each of response. As can be seen from this table, strong anthropomorphizers were significantly more likely than weak anthropomorphizers to claim that God (a) has a physical appearance, (b) intervenes in human affairs, and (c) answers

prayers. Moreover, strong anthropomorphizers were significantly more likely than weak anthropomorphizers to claim that they have experienced God's presence in their lives and are 100% certain that God exists. Participants' propensity to anthropomorphize God was thus correlated with their propensity to view God as a palpable (and pertinent) influence on everyday human affairs.

Beliefs about God's Relationship to Nature

The second part of the questionnaire probed participants' beliefs about God's role in the origin of the universe (cosmogenesis) and the origin of human beings (anthropogenesis). Participants' beliefs about cosmogenesis were elicited with the questions (1) "Do you believe that God created the universe?," (2) "Do you believe that the universe was created in the Big Bang?," and (3) "If you answered 'yes' to both questions, how do you resolve the apparent inconsistency between these two ideas?" Participants' beliefs about anthropogenesis were elicited with the questions (1) "Do you believe that God created human beings?," (2) "Do you believe that human beings evolved from other organisms?," and (3) "If you answered 'yes' to both questions, how do you resolve the apparent inconsistency between these two ideas?"

On the topic of cosmogenesis, 44% of participants claimed that the universe was created by God alone, 25% by the Big Bang alone, and 31% by both God and the Big Bang. Those who claimed that the universe was created by both God and the Big Bang justified their claim by appealing to some kind of dual process (e.g., "God set in motion the forces that created the Big Bang"). On the topic of anthropogenesis, 31% of participants claimed that human beings were created by God alone, 38% by evolution alone, and 31% by both God and evolution via some kind of dual process (e.g., "God created the organisms that ultimately evolved into humans").

The percentages of weak and strong anthropomorphizers who provided each of the above responses are displayed in Table 1. Strong anthropomorphizers were significantly more likely than weak anthropomorphizers to endorse a creationist explanation for both phenomena. Weak anthropomorphizers, on the other hand, were significantly more likely than strong anthropomorphizers to adopt a naturalistic explanation for cosmogenesis and a quasi-naturalistic explanation for anthropogenesis. Collapsing across "God only" explanations and "dual-process" explanations, strong anthropomorphizers were no more likely than weak anthropomorphizers to claim that God played at least *some* role in each process, indicating that the aforementioned differences are more nuanced than the difference between wholly accepting or wholly rejecting divine causation.

Beliefs about God's Relationship to Humankind

The third part of the questionnaire probed participants' beliefs about God's role in human suffering and human sin. These beliefs were elicited by asking participants to reason about two theological problems, traditionally known as the "problem of evil" and the "problem of sin" (see Plantinga, 1977). Reasoning about the first problem was elicited with the questions (1) "Do you believe that God is all powerful?," (2) "Do you believe that God is all good?," and (3) "If you answered 'yes' to both, why do you think God allows (or fails to prevent) human suffering?" Reasoning about the second problem was elicited with the questions (1) "Do you believe that God is all knowing?," (2) "Do you believe that God holds human beings responsible for their actions?," and (3) "If you answered 'yes' to both, why do you think God holds human beings responsible for actions he knows they will make?"

With regard to the problem of evil, 38% of participants denied that God is either omnipotent or omnibenevolent, 44% claimed that God is both omnipotent and omnibenevolent but that God uses suffering to teach or to punish (e.g., "God gave man free will and man chose to sin and suffering is a result of this sin"), 16% claimed that God is both omnipotent and omnibenevolent but that suffering is simply part of the human condition (e.g., "suffering, in various degrees, is part of the natural course of life"), and 3% plead ignorance. With regard to the problem of sin, 47% of participants denied that God is omniscient, judgmental, or both, 47% claimed that God is both omniscient and judgmental but that he also gave human beings the freedom to disobey him (e.g., "God gave man free will and hopes they will make the right choice, but sometimes they don't"), and 6% plead ignorance.

The percentages of weak and strong anthropomorphizers who provided each of the above responses are displayed in Table 1. Strong anthropomorphizers were significantly more likely than weak anthropomorphizers to claim that God uses suffering to teach or punish, and weak anthropomorphizers were significantly more likely than strong anthropomorph-izers to claim that suffering is part of the human condition. Moreover, strong anthropomorphizers were significantly more likely than weak anthropomorph-izers to claim that God gave humans the freedom to disobey him, and weak anthropomorphizers were significantly more likely than strong anthropomorphizers to deny that God is omniscient, judgmental, or both. In short, weak anthropomorphizers tended to treat suffering and sin as secular phenomena, not particularly linked to God, and strong anthropomorphizers tended to interpret sin as the defiance of divine law and suffering as the consequence of divine justice – beliefs reminiscent of those previously characterized as "belief in a just world" (Lerner, 1980).

Beliefs about Angels and Satan

The fourth part of the questionnaire probed participants' beliefs about two supernatural beings commonly associated with God: angels and Satan. Participants were first asked a series of property-attribution questions identical to those described earlier for God, and their responses were analyzed for the presence of biological and physical attributions. They were then asked whether they believed in the existence

of each being, and, if so, what they thought those beings looked like and how they thought those beings were related to God. Responses to the first question were coded for evidence that angels and Satan were believed to possess a physical appearance (e.g., "most angels have wings and are bright," "Satan looks like a ball of fire"), and responses to the second were coded for evidence that angels and Satan were believed to maintain a social relationship with God (e.g., "angels are God's servants," "Satan is God's enemy") rather than some type of existential relationship (e.g., "God is angels and angels are God," "Satan is a part of God because God is everything").

These responses are summarized in Table 1 as a function of participant group. Overall, strong anthropomorphizers were not significantly more likely than weak anthropomorphizers to believe in the existence of either angels or Satan, but they *were* significantly more likely to claim that these beings (a) possess the biological and physical properties of a human, (b) have a physical appearance, and (c) maintain a social relationship with God. In short, participants' propensity to anthropomorphize God was correlated with their propensity to anthropomorphize other members of their religious cosmology.

Beliefs about Heaven and Hell

The fifth part of the questionnaire probed participants' beliefs about two supernatural places associated with God: Heaven and Hell. Participants were asked whether they believed in the existence of each place, and, if so, where they thought those places were located, what they thought those places looked like, and what they thought the occupants of those places did. Responses to the first question were coded for evidence that Heaven and Hell were believed to occupy a discrete location in space (e.g., "Heaven is in the sky," "Hell is below the earth's crust"); responses to the second were coded for evidence that Heaven and Hell were believed to possess a physical appearance (e.g., "Heaven looks like a garden," "Hell looks like a prison"); and responses to the third were coded for evidence that the occupants of Heaven and Hell continue to engage in human activities (e.g., "singing, talking, dancing," "weeping and gnashing of teeth").

These responses are summarized in Table 1. Consistent with the belief that God possesses discrete physical properties, strong anthropomorphizers were significantly more likely than weak anthropomorphizers to claim that Heaven and Hell occupy discrete locations in space and that the occupants of Heaven and Hell engage in human activities. Strong anthropomorphizers were also more likely than weak anthropomorphizers to believe in the very existence of Heaven and Hell. In short, participants' propensity to anthropomorphize God was correlated with their propensity to accept, and to "spatialize," both places.

Religious Practices

The last part of the questionnaire contained questions about participants' religious practices and religious upbringing. Most participants (66%) claimed to engage in prayer at least occasionally, with 28% claiming to engage in prayer weekly. Likewise, most participants (72%) claimed to attend religious services at least occasionally, with 31% claiming to attend religious services weekly. In terms of affiliation, 50% claimed to belong to a Christian religion, 25% claimed to belong to a non-Christian religion, and 25% claimed to belong to no religious whatsoever. Finally, in response to the question "From whom did you acquire your current religious beliefs?," 44% claimed to have acquired their beliefs from a religious authority (e.g., a priest, a rabbi, "the church"), 34% claimed to have acquired their beliefs from their family, and 22% claimed to have acquired their beliefs from independent scholarship or personal reflection.

The percentage of weak and strong anthropomorphizers who claimed to engage in each of the aforementioned practices is displayed at the bottom of Table 1. Overall, strong anthropomorphizers were significantly more likely than weak anthropomorphizers to pray once or more per week, to attend religious services (at all), and to have acquired their beliefs from a religious authority. Weak anthropomorphizers, on the other hand, were significantly more likely to have acquired their beliefs from independent scholarship or personal reflection. The finding that participants' propensity to anthropomorphize God was correlated with their propensity to subscribe to the teachings of a religious authority is particularly interesting in light of the popular assumption that an anthropomorphic concept of God is not a "theologically correct" concept of God (e.g., Barrett & Keil, 1996; Boyer, 2003). Apparently, many strong anthropomorphizers would disagree.

Discussion

The present study explored the relationship between individuals' endorsement of an anthropomorphic God concept and their various beliefs and practices related to God. Overall, it was found that participants' propensity to anthropomorphize God was correlated with their propensity to (a) view God as a palpable and pertinent influence on human affairs, (b) adopt a creationist stance toward the origin of the universe and the origin of human beings, (c) adopt a "just world" view of human suffering and human sin, (d) anthropomorphize angels and Satan; (e) spatialize Heaven and Hell, and (f) engage in traditional religious activities, like prayer and worship.

Underlying these correlations were two qualitatively different, yet internally consistent, patterns of belief. One pattern, exhibited by strong anthropomorphizers, appeared to be structured around participants' understanding of human existence and human affairs. On this pattern, God is conceptualized as a divine ruler, angels are conceptualized as God's political allies, Satan is conceptualized as God's political opponent, Heaven and Hell are conceptualized as God's political territory, cosmogenesis and anthropogenesis are conceptualized as God's greatest achievements, and sin and suffering are conceptualized as God's primary spheres of influence. The other pattern of belief, exhibited by weak

anthropomorphizers, appeared to be structured around more abstract, and more limited, metaphysical commitments. On this pattern, God is conceptualized as an immaterial entity (rather than a physical object), angels and Satan are conceptualized as aspects of God (rather than independent agents), Heaven and Hell are conceptualized as states of being (rather than spatial locations), cosmogenesis and anthropogenesis are conceptualized as acts of nature (rather than acts of God), and sin and suffering are conceptualized as part of human nature (rather than part of a divine plan). Whereas the first pattern might best be described as a "humanistic theology," the second might best be described as an "existential theology."

The fact that these patterns of belief were associated with different religious practices suggests that different God concepts have different behavioral implications in addition to different cognitive implications. Presumably, the reason strong anthropomorphizers are more likely than weak anthropomorphizers to engage in prayer and worship is that only an anthropomorphic God would attend to, or care about, such activities. Of course, these correlations may be interpreted in the opposite manner – i.e., that individuals who engage in religious activities are more likely to hold a concept of God that is consistent with those activities. Tied to this concern is the broader concern that individual differences in God concepts may be due more to differences in religious education than to differences in the inferential relationship between one's God concept and one's God-related beliefs and practices.

There are at least three reasons to doubt that individuals inherent, rather than create, their personal theologies in the course of religious education. First, complete theologies are likely difficult to communicate given that God concepts are only one of many religious concepts open to multiple interpretations, as evidenced by participants' divergent interpretations of angels, Satan, Heaven, Hell, prayer, sin, and suffering. Second, the theologies documented in the present study were not specific to any one religion (see the section on religious affiliation in Table 1), implying that they are not the byproduct of a particular religious education (e.g., a Protestant education). Third, participants were unlikely to have pondered each and every issue broached by the questionnaire prior to participation, yet their responses to these questions were internally consistent nonetheless.

That said, future research could explore the development of personal theologies more directly. For instance, one could investigate the theologies of young children and chart how these theologies change over time, particularly their explanatory coherence and inferential scope. Alternatively, one could compare the theologies of different members of the same cultural unit, like the same church or the same family, to assess the dimensions along which personal theologies are most likely (and least likely) to differ. Such research would not only increase our understanding of religious cognition but would also increase our understanding of the interaction between culture and cognition more generally.

References

Barrett, J. L. & Keil, F. C. (1996). Conceptualizing a nonnatural entity: anthropomorphism in God concepts. *Cognitive Psychology, 31*, 219-247.

Barrett, J. L., Richert, R. A., & Driesenga, A. (2001). God's beliefs versus mother's: The development of nonhuman agent concepts. *Child Development, 72*, 50-65.

Bassett, J. F., & Williams, J. E. (2003). Protestants' images of self, God, and Satan as seen in adjective check list descriptors. *International Journal for the Psychology of Religion, 13*, 123-135.

Boyer, P. (2003). Religious thoughts and behaviors as byproducts of brain function. *Trends in Cognitive Sciences, 7*, 119-124.

Brown, D. E. (1991). *Human universals*. New York: McGraw-Hill.

Carey, S. (1985). *Conceptual change in childhood*. Cambridge, MA: The MIT Press.

Gallup, G. H. (2003). *Is religious faith collapsing in Great Britain?* Princeton, NJ: The Gallup Organization.

Gimenez-Dasi, M., Guerrero, S., & Harris, P. (2005). Intimations of omniscience and immortality in early childhood. *European Journal of Developmental Psychology, 2*, 285-297.

Gopnik, A., & Meltzoff, A. (1997). *Words, thoughts, and theories*. Cambridge: The MIT Press.

Keil, F. C. (1989). *Concepts, kinds, and cognitive development*. Cambridge: The MIT Press.

Lerner, M. (1980). *Belief in a just world: A fundamental delusion*. New York: Plenum.

McCloskey, M. (1983). Naïve theories of motion. In D. Gentner & A. Stevens (Eds.), *Mental models*, Hillsdale, NJ: Erlbaum.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review, 92*, 289-316.

Piaget, J. (1954). *The construction of reality in the child*. New York: Basic Books.

Plantinga, A. (1977). *God, freedom, and evil*. New York: Harper & Rowe.

Shtulman, A. (2006). Qualitative differences between naïve and scientific theories of evolution. *Cognitive Psychology, 52*, 170-194.

Shtulman, A. (2008). Variation in the anthropomorphization of supernatural beings and its implications for cognitive theories of religion, *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*, 1123-1138.

Slone, J. D. (2004). *Theological incorrectness*. Oxford, UK: Oxford University Press.

Smith, C., Snir, J., & Grosslight, L. (1992). Using conceptual models to facilitate conceptual change: The case of weight-density differentiation. *Cognition and Instruction, 9*, 221-283.

Trimeche, S., Vinsonneau, G., & Mullet, E. (2006). Individual differences in the theological concept of God. *International Journal for the Psychology of Religion, 16*, 83-100.

The Cultural Transmission of Explanations: Evidence that Teleological Explanations are Preferentially Remembered

Nicholas Z. Gwynne (nzg@berkeley.edu)
Tania Lombrozo (lombrozo@berkeley.edu)
Department of Psychology, 3101 Tolman Hall
Berkeley, CA 94703 USA

Abstract

Teleological explanations – explanations in terms of functions, purposes, or goals – are pervasive in religion and feature prominently in intuitive theories about the world, such as theory of mind and folk biology. Previous findings suggest that such explanations reflect a deep, explanatory preference. Here we explore the mechanisms underlying the prevalence and persistence of such explanations, following a method developed by Boyer and Ramble (2001) to examine which religious concepts are likely to survive processes of cultural transmission. Specifically, we test the prediction that novel teleological explanations are remembered better than mechanistic explanations, even when effects of an explanation's quality are taken into account. Two experiments support this prediction for artifact and biological trait explanations, but find the opposite pattern for explanations of non-living natural entities.

Keywords: explanation; teleological explanation; functional explanation; religion; memory; cultural transmission

Introduction

Why-questions are ubiquitous, ranging from those a child might ask to those of existential importance. While different questions solicit different answers, there seem to be systematic patterns in the properties of folk explanations (Lombrozo, 2006). Consider the difference between *mechanistic explanations*, which appeal to causal mechanisms, and *teleological explanations*, which appeal to functions and goals. The origin of human life can be explained mechanistically by appeal to evolution, but is often explained teleologically by appeal to some greater purpose.

In this paper we consider why folk explanations are so often teleological, and suggest that part of the answer lies in their mnemonic properties: teleological explanations are more likely to be remembered than mechanistic alternatives, and hence to survive processes of cultural transmission.

Teleology in folk explanations

Teleological explanations pervade intuitive theories. Folk psychological explanations, for example, often appeal to an agent's goals (e.g. Gopnik & Wellman, 1992; Wellman, 1992), while those in folk biology prominently feature biological functions (e.g. Atran, 1994). Teleological explanations even figure in early physics, as in Aristotle's appeals to teleological causation (Aristotle, Physics II).

Teleological explanations are particularly prominent in religion. Consider, for example, explanations for the origin of the universe. In the familiar Judeo-Christian creation story, the Old Testament God forms trees and animals for man to use (KJV, Genesis, 2:9, 2:18). Appeals to functions and goals likewise infuse the explanations for more mundane goods: we have wine because it "makes glad the heart of man," and bread because it "strengthens man's heart" (KJV, Psalm 104:14, 15).

What accounts for the prevalence of teleology in folk explanations? One possibility is that teleological explanations are common because they correspond to the structure of the world. This possibility is at best incomplete given that so many teleological folk explanations extend beyond those sanctioned by contemporary science (e.g. Kelemen & Rosset, 2009).

A second possibility is that teleological explanations are common because they are psychologically privileged, meaning that they are found more satisfying and generally preferred over alternatives. Evidence for this possibility comes from a growing literature on 'promiscuous teleology' demonstrating that young children prefer teleological explanations (e.g. "clouds are for raining"), and that this preference may persist into adulthood (Lombrozo, Kelemen, & Zaitchik, 2007; Kelemen and Rosset, 2009). Moreover, teleological and mechanistic explanations have unique consequences for categorization (Lombrozo, 2009). While some have suggested that teleological explanations are privileged in only some domains, such as folk biology (Atran, 1994, Keil, 1994), others suggest the preference is more widespread (Kelemen, 1999).

A third possibility is that teleological explanations are common because they are likely to survive processes of cultural transmission. Specifically, if teleological explanations are more likely to be reliably recalled, one should expect culturally transmitted beliefs such as religion to over-represent such explanations.

The possibility that mnemonic properties play a role in explaining the properties of religious explanations is particularly attractive in light of past research on cultural transmission. Within the domain of religion, Boyer and colleagues have successfully explored the role of memory and transmission in explaining the properties of religious concepts, such as demons and deities (e.g. Boyer, 2003). In these studies, participants read about

religious entities and were later asked to recall as many as possible. Boyer argued that if religious concepts have the properties they do because of cultural transmission, the concepts surviving this process should reflect the characteristics of concepts in the world's religions. Although the details have been disputed (e.g. Gonce et al. 2006; Norenzayan, Atran, Faulkner & Schaller, 2006; Tweney et al. 2006), Boyer's findings are broadly consistent with this proposal.

Beyond the domain of religion, research on iterated learning suggests that small biases in transmission can have large consequences over time (Kirby, 2001; Kalish, Griffiths, & Lewandowsky 2007). In the case of teleological explanations, a small bias in memory could have large consequences for the nature of folk explanations after several generations of transmission.

Our aim in this paper is to explore this third possibility: that teleology pervades folk explanations in part because teleological explanations are more likely to be remembered than mechanistic alternatives. We also explore the relationship between this hypothesis and the idea that teleological explanations are psychologically privileged – and hence deemed more satisfying – in all domains or in some domains.

How might memorability and satisfaction interact? It could be that teleological explanations are better remembered than mechanistic alternatives, and that this is *because* teleological explanations are judged more satisfying. Alternatively, memorability may influence satisfaction. Specifically, explanations that are more reliably encoded or recalled may lead to a greater sense of understanding, and hence be found more satisfying. A final possibility, and the one we favor, is that memorability and satisfaction have a common cause. If teleological and mechanistic explanations are supported by different kinds of representations or have a unique relationship to prior knowledge, both greater satisfaction and enhanced memory could result. By examining whether satisfaction mediates effects of explanation type on memorability we can begin to distinguish these alternatives.

In two experiments, participants read novel explanations that were either teleological or mechanistic and about biological traits, artifact properties, or (in Experiment 2) nonliving natural entities. Memory was then tested with recall and recognition tasks to examine the relationships between explanation type, explanation satisfaction, and memory.

Experiment 1

Experiment 1 examines whether teleological or mechanistic explanations are more reliably recalled and recognized. Additionally, it examines the relationship between explanations' memorability and their rated satisfaction, plausibility, detail, and unfamiliarity.

Explanation satisfaction and plausibility ratings were included to examine whether these factors mediate memorability. Ratings for detail and unfamiliarity were added because previous research suggests that explanations of moderate detail (Frazier et al, under review) and moderate unfamiliarity (Boyer & Ramble, 2001) are better remembered. By soliciting detail and unfamiliarity ratings we can examine whether memory differences, if found, result from differences in the detail or familiarity of the novel teleological and mechanistic explanations generated for the experiment. Having participants rate explanations along these four dimensions additionally provided a task to ensure that explanations were encoded prior to the memory tests.

Participants

One-hundred University of California students and community members (68% female, mean age = 22) participated in exchange for course credit or monetary compensation.

Materials and Procedure

The experiment involved twenty why-questions: ten regarding artifact properties and ten regarding biological traits. Each why-question had four possible answers of approximately equal length, two teleological and two mechanistic, for a total of 40 explanations of each type. Tables 1 and 2 provide sample why-questions and answers for each domain.

Table 1: Sample artifact trait explanations.

An istup is a kind of shovel with a compressible handle. Why do istups have compressible handles?
Teleological A: Because that way they can be used by aliens of various heights.
Teleological B: Because that way they can fit inside a regular toolbox.
Mechanistic A: Because the handle is made of distinct, interlocking segments.
Mechanistic B: Because the handle has hinges that allow it to fold.

These explanations were divided into four stimulus sets. In each set, half the questions were mechanistic and half teleological, with equal numbers across domains.

Data were collected via a computerized survey in a laboratory setting. Participants were instructed that they would "learn about the properties of plants, animals, and objects from alien planets and civilizations" and "receive explanations for those properties." Each participant was presented with explanations from a single stimulus set, and was asked to rate each explanation along four dimensions: satisfaction, plausibility, detail, and unfamiliarity. These scales

ranged from 1 (not at all satisfying, plausible, detailed, or unfamiliar) to 7 (very satisfying, very plausible, etc.).

Table 2: Sample biological trait explanations.

Bligs are a kind of animal with fur that is blue. Why do bligs have blue fur?
Teleological A: Because that way they can hide from predators in their environment, which contains blue rocks.
Teleological B: Because that way they can attract others of their species, who are drawn to blue.
Mechanistic A: Because of their mineral-rich diet, which contains compounds that are blue.
Mechanistic B: Because the surface of their planet contains fine, blue dust that sticks to their fur.

Following this encoding task, participants completed 24 distractor questions involving "Alien Math" which took 3 minutes to complete. Participants then completed a cued recall task and a recognition task. The cued recall task involved prompts such as the following:

> You previously saw the following: "Bligs are a kind of animal with fur that is blue. Why do bligs have blue fur?" What was the answer provided for this question? Please reproduce the answer you received as accurately as you can.

This prompt was repeated for each why question.

In the recognition task, the 20 why-questions were again repeated along with four candidate answers. The four answers included the one previously seen by that participant, as well as three additional answers for that why-question drawn from the three unpresented stimulus sets.

Participants were randomly assigned to a stimulus set. The question orders for encoding, recall, and recognition were randomized, as were the multiple-choice answers in the recognition task.

Results and Discussion

To analyze recall data, two independent coders categorized participant responses to cued recall questions as correct or incorrect/absent based on whether the explanation captured the gist of the mechanism or function. Coder agreement was 93%.

Recall accuracy was analyzed as a dependent measure in an ANOVA with explanation type and domain as within-subjects factors. This analysis revealed significant main effects of explanation type, $F(1,99)=25.5$, $p<.01$, and domain, $F(1,99)=7.90$, $p<.01$, with teleological explanation recalled more reliably than mechanistic explanations, and explanations for artifacts recalled more reliably than explanations for biological traits (see Fig. 1).

Recognition errors were analyzed as the dependent measure in an equivalent ANOVA, revealing a main effect of explanation type, $F(1,99)=5.05$,$p<.05$. Participants made an average of .83 errors (of 10) for teleological explanations, and 1.06 errors for mechanistic explanations.

These results suggest differential memory for teleological and mechanistic explanations, with teleological explanations remembered more reliably. To examine whether these effects stem from differences in the rated satisfaction, plausibility, detail, or unfamiliarity of teleological versus mechanistic explanations, the two analyses above were repeated as Linear Mixed Model analyses with satisfaction, plausibility, detail, and unfamiliarity as covariates. The main effect of explanation type on recall remained statistically significant, $F(1,396)=10.05$, $p<.01$; the effects of domain on recall and of explanation type on recognition did not. Notably, however, the explanations did differ in satisfaction, with higher ratings for the teleological explanations, $t(99)=5.71$, $p<.01$ (see Table 3).

Table 3: Mean satisfaction ratings in Experiment 1.

Biological Mechanistic	Biological Teleological	Artifact Mechanistic	Artifact Teleological
4.32	4.86	3.90	4.16

While these findings are consistent with several hypotheses, they suggest that differential recall is not a product of differential satisfaction. Instead, memorability and satisfaction may have a common cause, potentially stemming from the way teleological explanations relate to prior knowledge.

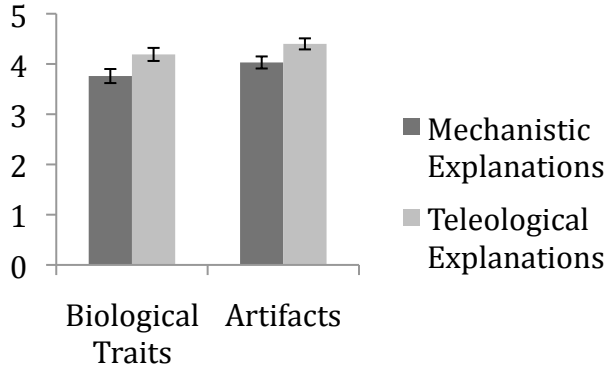


Figure 1: Recall accuracy in Experiment 1.

Experiment 2

Experiment 1 finds that teleological explanations are better remembered in two domains for which teleology is warranted. Are the mnemonic benefits of teleological explanation restricted to these domains, or do they extend more broadly? One aim of Experiment 2 is to

examine this question by extending the task to include nonliving natural entities (NNEs), such as lakes and mountains, which do not typically support teleological explanations.

Experiment 2 additionally aims to examine the nature of memory errors. To increase errors, Experiment 2 involves a larger number of explanations: 10 why-questions for each domain (artifacts, biological traits, NNEs), with one teleological and one mechanistic explanation for each why-question. Of particular interest are errors in explanation type, such as a mechanistic explanation that is "misremembered" as a teleological explanation, as systematic trends in error type could have implications for the cultural transmission of explanations. We target this issue in the recognition task by including two between-subjects conditions: one examining within-explanation type errors (e.g. teleological to teleological), and the other between-explanation-type errors (e.g. mechanistic to teleological).

Participants

Sixty University of California students and community members (72% female, mean age = 19 have participated for course credit or monetary compensation.

Materials and Procedure

The stimuli from Experiment 1 were augmented with 10 new why-questions involving the properties of Nonliving Natural Entities (NNEs), such as lightning and lakes. Four explanations for each question were generated as in Experiment 1 (see Table 4).

Table 4: Sample nonliving natural entity explanations.

A wernuct is a type of geyser that shoots very high into the air. Why do wernucts shoot high into the air?
Teleological A: Because that way giant reptiles can bathe under them.
Teleological B: Because that way the surrounding foliage can remain lush.
Mechanistic A: Because pressure builds up under ground and shoots water through cracks in the planet surface.
Mechanistic B: Because hot temperatures underground cause steam in the water, increasing its reach.

Explanations were subdivided into two stimulus sets, with each set containing a single teleological and mechanistic explanation for each why question.

The procedure mirrored Experiment 1, with the following modifications. In the explanation encoding task, participants saw each why-question twice: once presented with a teleological explanation, and once with a mechanistic explanation.

The distraction and recall tasks were identical to Experiment 1. Note that the recall prompt asked participants to report the answer to the question that they had previously seen; they were not explicitly prompted to provide two explanations.

There were two versions of the recognition task. In each version, participants received all 30 why-questions with two candidate responses: either one teleological and one mechanistic (between-type condition) or two of the same type (within-type condition), with the unseen explanations drawn from the unseen stimulus set.

Participants were randomly assigned to stimulus set and to recognition condition (between-type or within-type). Questions were presented randomly, but separated into two blocks, with order counterbalanced across participants, such that participants would see one answer for each why-question before seeing a second answer to any question. Answers for the recognition test were presented in random order.

Results and Discussion

Coding for the recall portion of the experiment was completed as in Experiment 1, with 88% agreement.

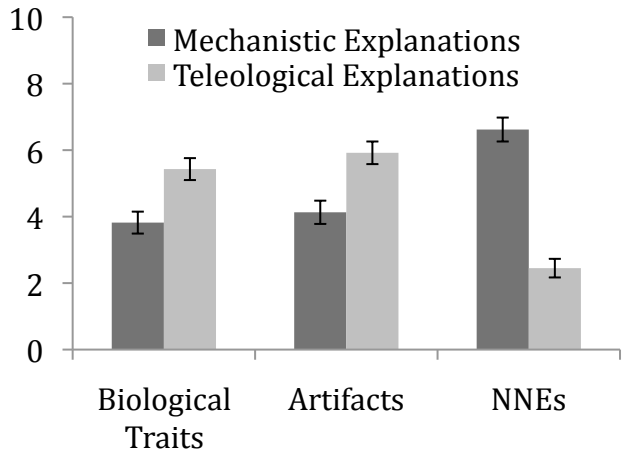


Figure 2: Recall accuracy in Experiment 2.

Recall accuracy was analyzed as a dependent measure in an ANOVA with explanation type and domain as within-subjects factors (see Fig. 2). This analysis revealed a significant effect of domain, $F(2,58)=13.4$, $p<.01$, as well as an interaction between explanation type and domain, $F(2,58)=70.5$, $p<.01$. Paired samples t-tests comparing recall for teleological and mechanistic explanations within each domain revealed significant differences, but in different directions. Teleological explanations were recalled more reliably than mechanistic explanations for biological traits, $t(59)=-3.91$, $p<.01$, and for artifacts, $t(59)=-4.39$, $p<.01$, replicating Experiment 1. In contrast, mechanistic explanations were recalled more reliably than teleological explanations for NNEs, $t(59)=8.94$, $p<.01$.

To determine whether these differences were driven by the satisfaction, plausibility, detail, or unfamiliarity of stimulus items, recall accuracy was analyzed separately for each domain in a Linear Mixed Model with explanation ratings as covariates. This analysis confirmed that teleological explanations were more reliably recalled than mechanistic explanations for biological traits, $F(1,118)=12.1$, $p<.01$, and artifacts, $F(1,118)=13.6$, $p<.01$, with the reverse pattern for NNEs, $F(1,118)=82.9$, $p<.01$. We did not replicate the satisfaction difference between mechanistic and teleological explanations found in Experiment 1.

Recognition errors were analyzed as the dependent measure in an ANOVA with explanation type and domain as within-subjects factors. This revealed a main effect of explanation type, $F(1,59)=63.9$,$p<.01$, with better recognition for teleological than mechanistic explanations (26.4 versus 22.4 of 30). There was also a main effect of domain, $F(2,58)=78.9$, $p < .01$, with better recognition for NNE (17.8/20) and artifact (17.7) explanations than for biological trait (13.3) explanations. Finally, there was an interaction between explanation type and domain, $F(2,58)=37.7$, $p<.01$, with teleological explanations recognized more reliably than mechanistic explanations for biological traits and for artifacts, but a non-significant trend in the reverse direction for NNEs (9.1 versus 8.8). These patterns of significance remained identical in a Linear Mixed Model with explanation ratings (satisfaction, etc) as covariates.

To examine the nature of memory errors, recognition accuracy was analyzed as the dependent measure in an ANOVA with recognition test (within-type versus between-type) as a between-subjects factor and explanation type as a within-subjects factor. This analysis confirmed the main effect of explanation type on recognition accuracy, $F(1,58)=75.3$, $p<.01$, with teleological explanations recognized more reliably than mechanistic explanations, and also revealed an interaction between explanation type and recognition test, $F(1,58)=11.5$, $p<.01$. Teleological explanations were correctly distinguished from teleological lures about as often as they were correctly distinguished from mechanistic lures (26.9 versus 25.8). In contrast, mechanistic explanations were more likely to be correctly distinguished from teleological lures than from mechanistic lures (23.4 versus 21.4). This suggests that mechanistic explanations were more "interchangeable" than teleological explanations.

It is worth noting that despite the sizeable advantage for mechanistic explanations of NNEs in recall, recognition performance was comparable for both teleological and mechanistic explanations. This suggests that recall performance may have reflected a preference to report warranted explanations in addition to or instead of differential effects of memory.

General Discussion

Experiments 1 and 2 suggest that when it comes to explaining the properties of biological traits or artifacts, teleological explanations are recalled more reliably than mechanistic alternatives. However, when it comes to explaining the properties of nonliving natural entities, mechanistic explanations are recalled more reliably than teleological alternatives. These effects persist when explanations' satisfaction, plausibility, detail, and unfamiliarity are taken into account.

Data for recognition were more variable, but yielded two suggestive results. First, mechanistic explanations were more likely to be "misrecognized" as other mechanistic explanations, suggesting that mechanistic explanations generated less distinctive memories. Second, despite the overwhelming advantage for mechanistic explanations of non-living natural entities in recall, recognition performance was comparable for teleological and mechanistic explanations, suggesting that even unwarranted teleological explanations were remembered remarkably well.

The current findings suggest that memory for teleological explanations is privileged in domains for which such explanations are typically warranted – namely biological traits and artifacts – but do not rule out the possibility that memory benefits extend more broadly. In follow-up work, we plan to explicitly prompt participants to report all provided explanations in recall, thereby helping to identify whether the benefit for mechanistic explanations of NNEs resulted from a genuine difference in memory or a preference to report explanations that are believed to be warranted.

While the memory differences we find are small – especially for recognition – models of cultural transmission suggest that small biases can quickly magnify over time (Kirby, 2001; Kalish, Griffiths, & Lewandowsky 2007). Thus a modest tendency to recall or report teleological explanations could result in a disproportionate representation of such explanations after only a few generations. When it comes to religious explanations, which are arguably less responsive to data than explanations in scientific or folk theories, biases in transmission may be responsible for systematic trends across the world's religions (see also Boyer, 2003).

We conclude by considering three important questions for future research. First, to what extent are differential effects of teleological and mechanistic explanations driven by the close relationship between teleology and intentional agency? Could it be that it is intentional explanations, not teleological explanations, that are preferentially remembered? While our effects extended to biological traits, it could be that participants construed the biological organism or natural selection as an intentional agent. Research suggests a close correspondence between teleology and intentional agency (e.g. Kelemen & DiYanni, 2005), but there is also

evidence that teleological explanations are reliably distinguished from intentional explanations (Lombrozo & Carey, 2006).

Second, to what extent are there individual differences in memory and transmission biases when it comes to explanations? In particular, what is the relationship between an individual's religious commitments and differential memory for explanation types? There is already some evidence that scientific training decreases teleological tendencies (Casler & Kelemen, 2008), while science training is inversely related with belief in god (Larson & Witham, 1997), particularly for top scientists (Larson & Witham, 1998).

Finally, and most critically, future research will need to explore the basis for differential memory for teleological and mechanistic explanations. Examining the contributions of explanation satisfaction, plausibility, detail, and unfamiliarity is a first step, as is considering a range of domains. But what is it about some explanations that make them more memorable, and hence more likely to survive processes of cultural transmission? We speculate that teleological explanations may be encoded differently from beliefs about causal mechanisms, and may have a more integrated relationship to prior knowledge. Of course, these speculations require further elaboration and empirical examination.

Acknowledgments

We thank the Concepts & Cognition Lab for feedback and help with data collection, and Kevin Uttich and Anna Rafferty for comments on earlier drafts. This research was supported by an Oxford Cognition, Religion, and Theology Grant awarded to the second author.

References

Aristotle, *Physics* (Vols. 1-8, Vol. 2).

Atran, S. (1994). Core domains versus scientific theories: Evidence from systematics and Itza-Maya folkbiology. *Mapping the mind: Domain specificity in cognition and culture*, 316ng th

Atran, S. (1998). Folk biology and the anthropology of science: Cognitive universals and cultural particulars. *Behavioral and Brain Sciences*, *21*(04), 547l and

Boyer, P. (2003). Religious thought and behaviour as by-products of brain function. *Trends in Cognitive Sciences*, *7*(3), 119n Cog

Boyer, P., & Ramble, C. (2001). Cognitive templates for religious concepts: cross-cultural evidence for recall of counter-intuitive representations. *Cognitive Science*, *25*(4), 535e Sci

Casler, K., & Kelemen, D. (2008). Developmental Continuity in Teleo-Functional Explanation: Reasoning about Nature among Romanian Romani Adults. *Journal of Cognition and Development*, *9*(3), 23.

Frazier, B.N, Wellman, H.M & Gelman, S.A. (Under review). Adults(UnderPreschool Childrenw) Preferences for Explanatory Detail.

Genesis. (n.d.). In *The Bible (KJV)* (pp. 2:9, 19).

Gonce, L. O., Upal, M. ,Afzal, D., Slone, J. & Tweney, R. D. (2006). Role of Context in the Recall of Counterintuitive Concepts. Jou*rnal of Cognition & Culture, 6,* 521-47.

Gopnik, A., & Wellman, H. M. (1992). Why the child childConcepts.uitive Coive Coounterin*Mind and Language*, *7*, 145and L

Kalish, M. L., Griffiths, T. L., & Lewandowsky, S. (2007). Iterated learning: Intergenerational knowledge transmission reveals inductive biases. *Psychonomic Bulletin and Review*, *14*, 288.

Keil, F. C. (1994). The birth and nurturance of concepts by domains: The origins of concepts of living things. *Mapping the mind: Domain specificity in cognition and culture* (p. 234. ng).

Kelemen, D. (1999). The scope of teleological thinking in preschool children. *Cognition*, *70*, 241tion

Kelemen, D., & DiYanni, C. (2005). Intuitions about origins: Purpose and intelligent design in children's reasoning about nature. *Journal of Cognition and Development*, *6*, 3rnal

Kelemen, D., & Rosset, E. (2009). The human function compunction: Teleological explanation in adults. *Cognition*, *111*, 138tion9

Kirby, S. (2001). Spontaneous evolution of linguistic structure-an iterated learning model of the emergence of regularity and irregularity. *IEEE Transactions on Evolutionary Computation*, *5*, 102Trans

Larson, E. J., & Witham, L. (1997). Scientists are still keeping the faith. *Nature*, *386*, 435-436. doi:10.1038/386435a0 0.Larson, E. J., & Witham, L. (1998). Leading scientists still reject God. *Nature*, *394*, 313.

Lombrozo, T. (2006). The structure and function of explanations. *Trends in Cognitive Sciences, 10*, 464-470.

Lombrozo, T. (2009). Explanation and categorization: How "why?" informs "what?" *Cognition, 110*, 248-253.

Lombrozo, T. & Carey, S. (2006). Functional explanation and the function of explanation. *Cognition, 99,* 167-204.

Lombrozo, T., Kelemen, D., & Zaitchik, D. (2007). Inferring Design. *Psychological Science*, *18,* 8,chologi

Norenzayan, A, Atran, S., Faulkner, J. & Schaller, M. (2006). Memory and Mystery: The Cultural Selection of Minimally Counterintuitive Narratives. *Cognitive Science 30,* 531-53.

Psalms. (n.d.). In *The Bible (KJV)* (pp. 104: 14, 15).

Tweney, Ryan D., M. Afzal Upal, Lauren O. Gonce, D. Jason Slone & Katie Edwards (2006). The creative structuring of counterintuitive worlds. *Journal of Cognition & Culture, 6,* 483-98.

Wellman, H. M. (1992). *The child's theory of mind*. MIT Press Cambridge, MA.

Semantic integration of novel word meanings after a single exposure in context

Arielle Borovsky (aborovsk@crl.ucsd.edu)
Center for Research in Language, 9500 Gilman Drive #0526
La Jolla, CA 92093-0515
Jeff Elman (elman@cogsci.ucsd.edu)
Department of Cognitive Science, 9500 Gilman Drive #0515
La Jolla, CA 92093-0515
Marta Kutas (kutas@cogsci.ucsd.edu)
Department of Cognitive Science, 9500 Gilman Drive #0515
La Jolla, CA 92093-0515

Abstract

We investigated the influence of sentence context on initial integration of novel word meanings into semantic memory. Adults read strongly or weakly constrained sentences ending in known and unknown (novel) words as electrical brain activity was recorded. Word knowledge was assessed via a lexical decision task where recently seen known and unknown word sentence endings served as primes for related, unrelated, and synonym/identical target words. N400 amplitudes to target words preceded by known word primes were reduced by prime relatedness. Critically, N400 amplitudes to targets preceded by novel words also varied with prime relatedness, but only if they initially appeared in highly constraining sentences. These results demonstrate that electrical brain activity accompanying one-shot contextual word learning is modulated by contextual constraint and reveals a rapid neural process that can integrate information about word meanings into the mental lexicon.

Keywords: word learning; N400; event-related brain potentials; language learning

Word learning is a lifelong process. However, research on adult first language word learning has largely been eclipsed by word learning in children and in adult bilinguals. Though these areas of study have yielded important insights into word learning, the mode by which adults learn words in their native language is likely to differ from that of children. For example, while children typically map words to novel or unnamed concepts (Markman & Wachtel, 1988), adults more often learn nuanced meanings for name-known concepts (e.g. jocund/happy). Furthermore, younger children often learn words in oral and ostensive contexts, whereas older children and adults acquire words largely via incidental learning in various language contexts, especially during reading (Jenkins, Stein, & Wysocki, 1984; Sternberg, 1987).

Word learning can be remarkably fast under the right conditions. A single exposure to a novel word can be sufficient for a learner to infer its probable meaning (Carey & Bartlett, 1978; Dollaghan, 1985). However, little is known about contextual influences on the representation of novel word meanings learned from a single exposure, how quickly this new information is integrated with the existing semantic system, or what the neural correlates of this rapid learning may be. The main goal of our research is to explore these issues by measuring modulation of event-related brain potentials (ERPs) during single trial word learning in sentence contexts of varying strength.

Background

Studies of adult first and second language learners provide evidence for rapid neural changes in young adults in association with word learning in both first (L1) and second (L2) languages (Borovsky, Elman, & Kutas, 2007; McLaughlin, Osterhout, & Kim, 2004; Mestres-Misse, Rodriguez-Fornells, & Münte, 2006; Perfetti, Wlotko, & Hart, 2005; Stein et al., 2006). For example, McLaughlin and colleagues (2004) compared brain responses in native French speakers to undergraduates learning French as a second language. They found that college language learner's brain responses during a semantic priming task using French words were indistinguishable from native speakers after only a few months of instruction. Their findings demonstrate not only that the brain may process word meanings acquired in childhood and adulthood similarly, but that lexical acquisition over extended training can be measured by modulations in neural activation.

L1 word learning studies have suggested that even faster neural changes due to word learning are possible (Perfetti, Wlotko & Hart, 2005; Mestres-Missé, Rodriguez-Fornells & Münte, 2007; Borovsky, Elman & Kutas, 2007). For example, Mestres-Missé and colleagues (2007) found that three presentations of a novel word in progressively constraining sentence contexts can significantly modulate the associated neural responses. We (Borovsky, Elman & Kutas, 2007) further examined the influence of contextual constraint on novel word usage after only a single presentation. Novel words were presented in a single highly or weakly constraining sentence context. Subsequently, participants were asked to differentiate between appropriate and inappropriate usages of these novel words as objects of particular verbs. Participants were able to incorporate significant information about the proper usage of novel words after a single exposure, but only when the novel words initially appeared in highly (and not in weakly) constraint contexts.

While comprehending a word's usage is an important aspect of vocabulary acquisition, knowledge of a word's relationship to other words is also vital. For example, part of our understanding of the words CAT, DOG and CHAIR, is that CAT and DOG have many overlapping similarities and features that neither shares with CHAIR. Research has indicated that adults can gain significant knowledge of these

relationships with a few exposures in sentences (Mestres-Misse et. al, 2006). In the present study, we ask whether even one exposure suffices to enable learners to incorporate the novel word into the semantic network that functionally connects words with related meanings, and how sentence contexts might influence this acquisition, if at all. More specifically, we use an event-related brain potential (ERP) component - the N400 - to index knowledge of word meaning via semantic priming when unknown words are initially presented in sentences that either strongly or weakly constrain their meaning.

The N400 is an ERP component that is a sensitive measure of semantic processing. It is a negative going wave with a centroparietal maximum that peaks approximately 400ms after the onset of any potentially meaningful stimulus. The N400 amplitude of word is reduced when it is (contextually) expected or when features associated with its meaning are easily integrated within the surrounding context (Kutas & Federmeier, 2000; Kutas & Hillyard, 1980) Additionally, the N400 to orthographically legal and pronounceable nonwords (pseudowords) is large (Ziegler, Besson, Jacobs, Nazir, & Carr, 1997); it is not present for true nonwords that do not have orthographically legal spellings, or are unpronounceable (Bentin, 1987). The N400 is modulated by lexical frequency and is larger for lower frequency words in lists (Smith & Halgren, 1987). N400 amplitude, thus, is associated with a word's meaningfulness in a given context, ranging from small in amplitude when a word is very easily integrated or understood, to large when a word's meaning is unknown. These findings suggest that N400 amplitude is likely to vary with the degree to which the meaning of a newly encountered word is appreciated – a prediction that has been borne out by recent research in L2 and L1 word learning (e.g., McLaughlin, Osterhout & Kim, 2004; Mestres-Misse et. al. 2007).

Target words preceded, or primed, by an identical or related word (for example doctor- NURSE, or doctor-DOCTOR) are associated not only with faster response times (e.g., Neely, 1991), but with reduced N400 amplitudes (Bentin, McCarthy, & Wood, 1985; Nobre & McCarthy, 1994), compared to target words preceded by words that are unrelated in meaning, or by nonwords (i.e. doctor-CHAIR, or doctor-FOOP). Such semantic priming effects have been interpreted as reflecting the functional organization of words in the brain (Collins & Loftus, 1975; Lucas, 2000).

In this study, we examine the impact of context on novel words' initial integration in semantic memory via semantic priming. Following an initial exposure in a strongly or weakly constraining sentence, we gauge successful word meaning acquisition by means of semantic priming in a lexical decision task. In this case, N400 amplitude modulation to a target word by a recently experienced prime word is taken as an index of semantic integration of the novel word's meaning into semantic memory. We use N400 amplitudes to gauge how contextual constraint influences acquisition of word meaning by contrasting how these novel words prime target words that are identical, related, or unrelated in (implied) meaning. We can also explore how context impacts the integration of novel word meaning into the mental lexicon by assessing the interaction between the priming effect and contextual constraint.

Methods

Participants:

24 college students (13 F) were given credit or paid $7/hr for their participation. Ages ranged between 18-30 (mean: 19.50). All participants were right-handed, native English speakers and had no significant exposure to another language at least before the age of 12. Participants reported no history of mental illness, learning disability, language impairment, drug abuse, or neurological trauma. All participants had normal hearing and normal (or corrected to normal) vision. An additional 11 participated but were not analyzed: 5 had excessive blinking or motion artifact, 1 due to equipment failure, and 5 reported a characteristic which disqualified them from analysis (4 had significant childhood second language exposure, 1 had non-normal vision.)

Materials:

Stimuli consisted of 132 sentence pairs selected from Federmeier and Kutas (1999), and 528 word pairs selected to correspond with 132 sentence final words. Both are described in detail below:

Sentences: 64 high constraint and 64 low constraint sentence pairs were selected from Federmeier and Kutas (1999). These pairs had previously been extensively normed to ensure adequate levels of cloze probability for high and low constraint sentences. Sentence pairs consisted of an initial sentence that set up an expectation of a meaning and item category, and a second sentence that was matched with sentence final words that were either plausible and expected known word sentence completions (Federmeier & Kutas, 1999), or unknown pseudowords, yielding 32 sentences in each of four main conditions: 1) High constraint / Known word ending, 2) High Constraint / Unknown word ending 3) Low constraint / Known word ending and 4) Low constraint / Unknown word ending (see Table 1a for examples). Sentences pairs were counterbalanced such that each appeared with a Known and Unknown ending equally across all versions, but not repeated within a subject. Known word target items consisted of words in 64 categories, and these categories were used as the basis for selecting semantically related and unrelated prime-target pairs, described below.

Word-Pairs: 528 word pairs were constructed that consisted of a prime followed by a target word presented one stimulus at a time. Since repetition is known to diminish N400 effects (Van Petten, Kutas, Kluender, Mitchiner, & McIsaac, 1991), and it is unclear if repetition and constraint might interact, we designed the priming task such that the N400 of interest was to the presentation of a target word that followed a prime that was either a Known or Unknown words from the sentence endings described above. The N400 effect of interest would thus be elicited to previously unseen real

word targets in three conditions: 1) Synonym/Identical (Syn/ID: rabbit-RABBIT), 2) Related (Rel: rabbit-MOUSE), and 3) Unrelated (Unrel: rabbit-RIBBON). Unrel and Rel word pairs were selected to be closely matched to other target conditions in word frequency ($F(2, 353)=1.09$, $p=0.34$) length ($F<1$)), syllables ($F<1$), and phonemes ($F<1$), as reported by the MRC psycholinguistic database (Wilson, 1988). Efforts were also made to match targets as closely as possible on Concreteness, Familiarity and Imageability ratings when they were available. Additionally, targets in each condition did not differ as a function of constraint in frequency [Syn/ID: $|t| <1$, Rel: $|t| <1$, Unrel: $t(130)=1.06$, $p=0.29$] length [Syn/ID: $t(130)=-1.45$, $p=0.15$, Rel: $|t| <1$, Unrel: $t(130)=-1.27$, $p=0.21$], # syllables [Syn/ID: $|t| <1$, Rel: $|t| <1$, Unrel: $|t| <1$], and #phonemes [Syn/ID: $t(130)=-1.36$, $p=0.18$, Rel: $|t| <1$, Unrel: $t(130)=-1.32$, $p=0.19$]. Highly associated word pairs were not included (like mouse-CHEESE), as confirmed via the Edinburgh Associative Thesaurus (Kiss, Armstrong, Milroy, & Piper, 1973). In cases involving Unknown word primes, Syn/ID, Rel and Unrel, was determined by its implied meaning from sentence context in which it had previously appeared.

An equal number of Nonword targets were also constructed so that the proportion of "Yes" and "No" lexical decision responses were equivalent. Nonwords were constructed using the ARC Nonword database (Rastle, Harrington, & Coltheart, 2002), and were selected to be pronounceable, and to contain between 4-7 letters.

In each version, each Known and Unknown prime was paired with two of three possible real word targets, and two nonword targets. The proportion of targets in each condition was: Nonwords=1/2, Syn/ID=1/6, Rel=1/6, Unrel=1/6. Known and Unknown prime was matched with the targets with equal frequency across versions. Table 1b includes examples of word pairs in the study.

Procedure:

Participants were tested in a soundproof, electrically-shielded chamber and were seated in a comfortable chair in front of a monitor. The experiment consisted of two interleaved tasks: sentence comprehension and priming.

In the sentence comprehension task, participants were instructed to read the sentence pairs for comprehension and to try to understand the sentences even when nonsense words appeared. The first sentence in each pair was presented in its entirety, and participants pressed a button to indicate that they were ready for the second. The second sentence was preceded by a series of crosses (500 ms duration with a stimulus-onset-asynchrony (SOA) varying randomly between 300-800 ms) to orient the participant toward the center of the screen. Sentences were presented one word at a time, each for 200 ms with a SOA of 500 ms. Participants were asked to minimize blinking and movement during sentences. The final target word appeared for 1400 ms.

In the priming task, participants were instructed to read every word that appeared on the screen and indicate with a button press if the target item (which always appeared in capital letters) was or was not a real word. Participants viewed two sets of prime/target pairs, and were given a 2500ms offset period to blink between pairs. Prime pair onsets were preceded by a set of fixation crosses that were randomly presented for 200-500ms. Immediately following the fixation cross, a prime word appeared for 200 ms, followed by an offset of 300ms, followed by the target word presentation for 200ms, and offset of 800ms. Participants provided a lexical decision response as soon as possible after each target word appeared in capital letters.

Table 1. Examples of sentences and word pairs

A) Context Sentence Pairs (Context Constraint / Word Type)	
High/ Known	Peter sat gaping at the centerfold. He asked his friend if he could borrow the MAGAZINE.
High/ Unknown	Peter sat gaping at the centerfold. He asked his friend if he could borrow the YERGE.
Low/ Known	The package was rectangular and heavy and suspiciously academic. Bianca was disappointed that her uncle was giving her a BOOK.
Low/ Unknown	The package was rectangular and heavy and suspiciously academic. Bianca was disappointed that her uncle was giving her a SHUS.

B) Word Pairs (prime – TARGET)	Syn/ID	Rel	Unrel
High/ Known	magazine-MAGAZINE	magazine - NOVEL	magazine-ACCIDENT
High/ Unknown	yerge – MAGAZINE	yerge – NOVEL	yerge-ACCIDENT
Low/ Known	book – BOOK	book – LETTER	book – ROAD
Low/ Unknown	shus – BOOK	shus – LETTER	shus – ROAD

Note: all word pairs were also paired with an equal number of pseudoword targets, not depicted in this table

The experiment consisted of 11 blocks of sentence/prime sets that were interleaved as follows. Participants read 12 sentence pairs, before completing the priming task consisting of 48 pairs, with primes being selected from the 12 immediately preceding sentence endings. Participants were given a break after each sentence/priming set.

In order to ensure that participants attended to the study sentences, participants were given a surprise old/new memory post-test containing 50 sentences that had appeared in the study, and 50 sentences that had not.

Electrophysiological recording:

Scalp potentials were continuously recorded from 26 geodesically arranged sites using an ElectroCap with tin electrodes and a left mastoid reference. Potentials were digitized at a sampling rate of 250 Hz and hardware bandpass filter of 0.1-100Hz with Grass Amplifiers. Impedances were kept below 5 kΩ.

Data analysis*:* Data were re-referenced offline to an average left and right mastoid. Trials contaminated by eye movements, blinks, excessive muscle activity, or amplifier blocking were rejected offline before averaging. ERPs were computed for epochs extending from 100 ms pre-stimulus onset to 920 ms post-stimulus onset. Averages of artifact-free ERP trials were computed for the target words in the four learning conditions (High/Known, High/Unknown, Low/Known, Low/Unknown) as well as to targets in all priming conditions (Syn/ID, Rel, and Unrel targets for each of the four main conditions High/Known, High/Unknown, Low/Known, Low/Unknown) after subtraction of the 100 ms pre-stimulus baseline

Table 2. Mean reaction times (ms) and mean percentage of correct responses for priming task.

	Real Word Primes		Novel Word Primes	
	Constraint		Constraint	
	High	Low	High	Low
% correct				
Syn/ID	99 (0.6)	99 (1.9)	97(6)	98(2.1)
Rel	97 (2.4)	93 (4.1)	94(4.3)	95(3.5)
Unrel	93 (6.8)	96 (3.2)	95(3.4)	94(3.8)
RT				
Syn/ID	512 (80)	488 (82)	543 (77)	553 (76)
Rel	568 (87)	561 (72)	567 (79)	570 (83)
Unrel	586 (79)	578 (75)	571 (75)	567 (79)

Note: Standard deviations are reported in parenthesis.

Results

Behavioral performance:

Participants made lexical decisions for words that were identical, related, or unrelated in meaning to a prime word. Mean accuracy and RTs are shown in Table 2. We did not statistically analyze accuracy since accuracy was near ceiling, with the lowest accuracy in any condition being 93%. For RT, A three factor repeated measures ANOVA on RT was carried out with factors of Word type (Unknown and Known) x Constraint (High and Low) x Prime relationship (Identical, Related and Unrelated). A main effect of Prime was found [$F(2, 46)=85.49$, $p< 0.0001$], with Tukey tests revealing that this effect was driven by faster responses to Identical targets than every other condition. No overall difference was found between Rel and Unrel conditions. There was also a main effect of Word Type [$F(1, 23)=11.94$, $p=0.002$] driven by faster responses to targets preceded by Known vs. Unknown words. An interaction of Prime x Type was also found [$F(2, 46)=29.2$, $p<0.0001$]. Follow-up Tukey tests revealed that this interaction was driven by targets that were preceded by Syn/ID Known words eliciting the fastest responses compared to other conditions. There were no other significant interactions. Although no significant three-way interaction was found, pairwise comparisons were conducted to examine the relationships between Syn/ID, Rel and Unrel meanings in each of the four prime conditions: Known/High, Unknown/High, Known/Low, and Unknown/Low. These analyses revealed that targets preceded by Known/High and Known/Low primes elicited faster RTs when preceded by a word identical in meaning, compared to a related or unrelated word. On the other hand, targets preceded by Unknown words did not elicit priming effects in any condition (all $p>0.05$).

ERP data: N400 amplitude

Context sentence endings*:* We analyzed ERP responses to sentence endings in four conditions: Known/High, Known/Low, Unknown/High and Unknown/Low. ERPs to sentence endings are shown in Figure 1. N400 mean amplitude was measured between 250-500ms post final word onset at four centro-parietal electrode sites (RMCe, LMCe, MiCe, MiPa) where N400 effects are typically largest. A two-factor repeated measures ANOVA with factors of Word Type (Known and Unknown) and Constraint (High and Low) revealed an effect of Word Type [$F(1,23)=28.85$, $p<.0001$] with Unknown word endings eliciting larger N400s than Known word endings. No other main or interaction effects were observed.

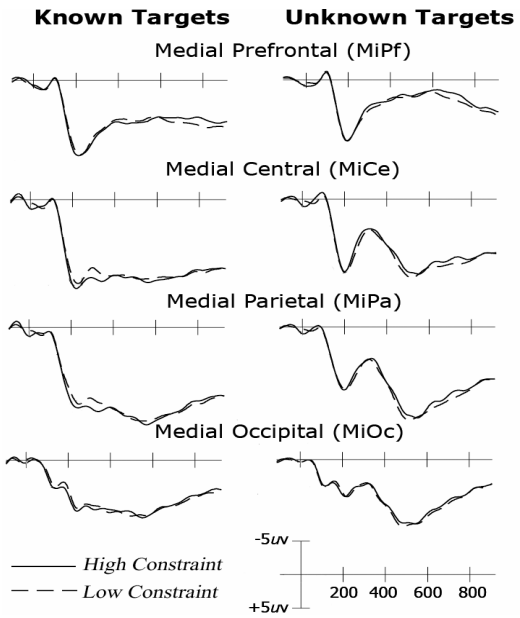


Figure 1. Grand average ERPs to known and unknown target words in context sentences at medial electrode sites.

Priming task: a shows ERPs to target words in the four main prime word conditions (Known/High, Known/ Low, Unknown/High, Unknown/Low). As can be seen from this figure, an effect of Target type is seen via modulation of the negative going peak from 250-500ms (N400) in all Prime conditions, except for Unknown/Low words. N400 mean amplitude was measured between 250-500ms post target word onset at four centro-parietal electrode sites (RMCe, LMCe, MiCe, MiPa) where N400 effects are typically largest (Figure 2b). A three-factor repeated measures ANOVA was conducted with factors of Prime-Type (Known or Unknown), Prime-Constraint (High or Low) and Target relationship (Sy/ID, Rel, Unrel), using Greenhouse-Geisser. univariate epsilon values

Table 3. F-values from pairwise ANOVAs comparing mean amplitude N400 to related, unrelated, and synonym/ID targets

		Syn/ID	Rel	Unrel
Known/High				
	Syn/ID	--	14.92**	30.22***
	Rel	14.92**	--	11.17**
	Unrel	30.22***	11.17**	--
Known/Low				
	Syn/ID	--	27.80***	23.69***
	Rel	27.80***	--	Ns
	Unrel	23.69***	ns	--
Unknown/High				
	Syn/ID	--	6.22*	32.24***
	Rel	6.22*	--	4.61*
	Unrel	32.24***	4.61*	--
Unknown/Low				
	Syn/ID	--	ns	Ns
	Rel	Ns	--	Ns
	Unrel	Ns	ns	--

* - $p < 0.05$, ** - $p < 0.01$, *** - $p < 0.0001$

This analysis revealed a significant effect of Word Type [$F(1,23)=5.4990$, $p=0.02$], with Unknown words eliciting larger N400 amplitudes than known words, and Target [$F(1.8922, 43.522)=32.439$, $p<0.0001$], with Syn/ID targets eliciting the smallest N400 amplitudes, but no main effect of Constraint [$F(1,23)<1$]. There was also an interaction of Constraint x Prime [$F(1,23)=6.29$, $p=0.02$]. No other interactions were significant. Preplanned pairwise repeated measures ANOVA comparisons were conducted to compare mean N400 amplitude between Rel, Unrel and Syn/ID targets in each of the four main Prime word conditions. The results of these comparisons are shown in Table 3. As seen from this table, significant priming effects were observed in all conditions, except for Unknown prime words that initially appeared in Low constraint contexts.

Discussion

This study explored the neural correlates of the rapid acquisition of recently experienced novel word meanings in adults' native language. Our goal was to understand the influence of sentential constraint on the integration of novel word meanings into the "mental lexicon" after only a single exposure. We measured behavioral and ERP responses in priming task to ask if the information that is rapidly integrated about novel word meanings includes information about a word's lexico-semantic relationships with other (known) words.

The behavioral (lexical decision) results did not reveal evidence of priming between novel words and related or synonymous targets. This result alone would suggest that no learning occurred regardless of sentential constraint. The electrophysiological results, however, support a different conclusion.

Known word primes produced N400 priming effects replicating a well-established result: smaller N400 amplitudes to target words preceded by identical or related words, relative to unrelated words. This was also the pattern for Unknown words (or perhaps more accurately, recently seen words) but only if it had initially appeared in a strongly constraining context. Semantic relatedness between an Unknown (novel word) prime and a real word target could only have been inferred from the sentence context in which that novel word previously appeared and apparently only strongly constraining contexts supported this inference.

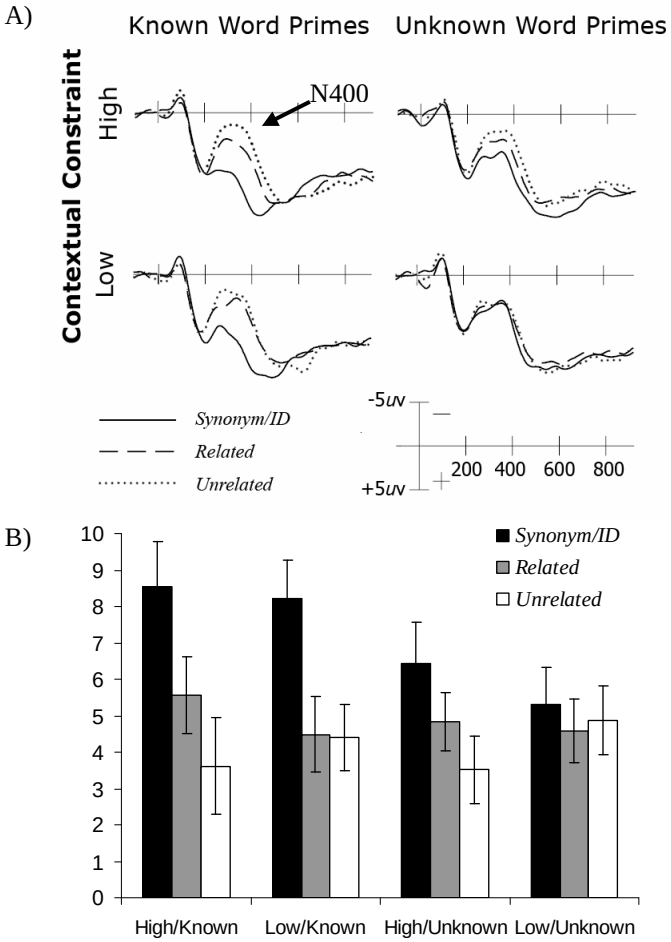


Figure 2. A) Grand average ERPs to target words in priming task at the vertex electrode (MiCE). B) N400 Mean amplitudes 250-500ms. Since the N400 is a negative going wave, larger N400 amplitudes are represented by smaller values on this figure.

Previous work has suggested that adults can integrate and organize information about word meanings after a few of weeks of second language instruction (McLaughlin et al., 2004; Stein et al., 2006), and even more rapidly in adult's first language, such as after only an hour of study of word definitions (Perfetti et al., 2005) or after three presentations in sentential context (Mestres-Misse et al., 2006). Our results extend these findings to show that in some cases a single exposure of a novel word in a strongly constraining sentence context is sufficient to convey significant information about its meaning to support semantic priming, and that there is a very fast neural process which enables the integration and retention of this information over at least a several minute delay. We add to a growing body of evidence that the rapidly acquired information about novel words includes

information not only about its usage in sentences but also about its meaning.

More generally, this paradigm suggests a novel method to examine the impact of sentential context and constraint on word processing. Further research will be necessary to extend these findings to other aspects of word meaning and knowledge, and to determine how long such information about a word's usage and meaning is retained and is effective.

Acknowledgements:

AB was supported by an NSF graduate fellowship and NIH training grant DC00041. This work was also funded by R01 MH60517 and R01 HD053136 to JE and RO1 AG08313 and R01 NICHD22614 to MK.

References

Bentin, S. (1987). Event-related potentials, semantic processes, and expectancy factors in word recognition. *Brain & Language, 31*(2), 308-327.

Bentin, S., McCarthy, G., & Wood, C. C. (1985). Event-related potentials associated with semantic priming. *Electroencephalography and Clinical Neurophysiology, 60*, 343-355.

Borovsky, A., Elman, J., & Kutas, M. (2007). *Getting the gist is not enough: An ERP investigation of contextual word learning.* Paper presented at the Proceedings of the 29th Annual Cognitive Science Society, Nashville, TN.

Carey, S., & Bartlett, E. (1978). Acquiring a single new word. *Papers and Reports on Child Language Development, 15*, 17-29.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review, 82*(6), 407-428.

Dollaghan, C. (1985). Child Meets Word: "Fast Mapping" in Preschool Children. *J Speech Hear Res, 28*(3), 449-454.

Federmeier, K. D, & Kutas, M (1999). A rose by any other name: Long-term memory structure and sentence processing. *Journal of Memory and Language, 41*, 469-95.

Jenkins, J. R., Stein, M. L., & Wysocki, K. (1984). Learning Vocabulary Through Reading. *American Educational Research Journal, 21*(4), 767-787.

Kiss, G. R., Armstrong, C., Milroy, R., & Piper, J. (1973). An associated thesaurus of English and its computer analysis. In A. J. Aitken, R. Bailey & N. Hamilton-Smith (Eds.), *The computer and literary studies*. Edinburgh: Edinburgh University Press.

Kutas, M., & Federmeier, K. D. (2000). Electrophysiology reveals semantic memory use in language comprehension. *Trends in Cognitive Sciences, 4*(12), 463-470.

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science, 207*(4427), 203-205.

Lucas, M. (2000). Semantic priming without association: A meta-analytic review. *Psychonomic Bulletin & Review, 7*(4), 618-630.

Markman, E. M., & Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology, 20*(2), 121-157.

McLaughlin, J., Osterhout, L., & Kim, A. (2004). Neural correlates of second-language word learning: minimal instruction produces rapid change. *Nature Neuroscience, 7*(7), 703-704.

Mestres-Misse, A., Rodriguez-Fornells, A., & Munte, T. F. (2006). Watching the Brain during Meaning Acquisition. *Cereb. Cortex*, bhl094.

Neely, J. H. (1991). Semantic priming effects in visual word recognition: A selective review of current findings and theories. In E. Derek Besner & E. Glyn W. Humphreys (Eds), *Basic processes in reading: Visual word recognition.* (pp. 264-336) Hillsdale, NJ, US.

Nobre, A. C., & McCarthy, G. (1994). Language-Related Erps - Scalp Distributions and Modulation by Word Type and Semantic Priming. *Journal of Cognitive Neuroscience, 6*(3), 233-255.

Perfetti, C. A., Wlotko, E. W., & Hart, L. A. (2005). Word learning and individual differences in word learning reflected in event-related potentials. *J Exp Psychol Learn Mem Cogn, 31*(6), 1281-1292.

Rastle, K., Harrington, J., & Coltheart, M. (2002). 358,534 nonwords: The

ARC Nonword Database. *Quarterly Journal of Experimental Psychology, 55*(A), 1339-1362.

Smith, M., & Halgren, E. (1987). Event-related potentials during lexical decision: effects of repetition, word frequency, pronounceability, concreteness. *Electroencephalography and Clinical Neurophysiology Supplement, 40*, 417-421.

Stein, M., Dierks, T., Brandeis, D., Wirth, M., Strik, W., & Koenig, T. (2006). Plasticity in the adult language system: A longitudinal electrophysiological study on second language learning. *Neuroimage, 33*(2), 774-783.

Sternberg, R. J. (1987). Most vocabulary is learned from context. . In M. G. McKeown & M. E. Curtis (Eds.), *The Nature of Vocabulary Acquisition* (pp. 89-106). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Van Petten, C., Kutas, M., Kluender, R., Mitchiner, M., & McIsaac, H. (1991). Fractionating the word repetition effect with event-related potentials. *Journal of Cognitive Neuroscience, 3*(2), 131-150.

Wilson, M. D. (1988). The MRC Psycholinguistic Database: Machine Readable Dictionary, Version 2. *Behavioural Research Methods, Instruments and Computers, 20*(1), 6-11.

Ziegler, J. C., Besson, M., Jacobs, A. M., Nazir, T. A., & Carr, T. H. (1997). Word, pseudoword, and nonword processing: A multitask comparison using event-related brain potentials. *Journal of Cognitive Neuroscience, 9*(6), 758-775

Fixation durations in first-pass reading reflect uncertainty about word identity

Nathaniel J. Smith
njsmith@cogsci.ucsd.edu
UC San Diego Department of Cognitive Science
9500 Gilman Drive #515, La Jolla, CA 92093-0515 USA

Roger Levy
rlevy@ling.ucsd.edu
UC San Diego Department of Linguistics
9500 Gilman Drive #108, La Jolla, CA 92093-0108 USA

Abstract

In reading, it is often assumed that words are recognized sufficiently quickly, accurately, and unambiguously that downstream processes may proceed with perfect information about word identity. For example, word predictability is believed to affect early reading time measures, yet a word's predictability cannot be calculated without knowledge of the word's identity. We argue that such information is not, in general, available to the language processing system, and that it proceeds with only probabilistic information about word identity. We predict therefore that what have been analyzed previously as predictability effects must instead be based on noisy estimates of word predictability that are influenced by the predictability of visually similar words (neighbors). We test this prediction by building a Bayesian model of visual word recognition, using it to compute the 'average neighborhood surprisal' of words in a corpus, and testing the ability of this novel measure to explain human reading time data.

Keywords: Psychology, Cognitive Science, Perception, Language Understanding, Bayesian Modeling, Neighborhood Effects, Visual Uncertainty, Reading

Performance in isolated word recognition tasks is often affected by the existence or properties of words that are not presented, but that are visually similar to the words which are presented. For instance, a word with many neighbors — especially high frequency neighbors — tends to produce a faster response in the lexical decision task, and a slower response in naming or reading tasks (Perea & Rosa, 2000). Norris (2006) has argued that these divergent results can be best explained by uncertainty in the processing system. That is, noise is an inevitable component of all biological computation, and if the processor receives only noisy information about a word's shape, then it must consider all similar looking words as candidates for identification. When there are many such candidates, identifying the single correct candidate (as in the naming task) becomes more difficult, because there are many incorrect distractors and only one correct target; resolving this difficulty requires the acquisition of more sensory information, which requires more time. In the lexical decision task, however, it is not necessary to determine *which* word is seen, only *whether* a word is seen, and therefore increasing the number of candidates only makes it easier to give a correct response (even if for the 'wrong' reason).

However, in reading — that is, processing connected language, rather than isolated words — another consideration arises. In a naming task, no response can be given until the word is identified, but when reading, the ultimate outcome is not the name of a single word, but an understanding of the text as a whole. Here, we ask: can the linguistic processing associated with a word proceed before that word is uniquely identified? And if so, what are the consequences for processing? It's possible that neighborhood effects are limited to some early, serial, word identification process, in which any uncertainty is resolved before higher-level linguistic processes begin. Alternatively, this uncertainty may be propagated through the linguistic processing system itself.

Most current models of high-level language processing fall into the former category; for instance, they take as input words, rather than probability distributions over words. However, there is some reason to suspect that the latter possibility is more plausible. Spoken language, in particular, is a very noisy signal, in which word identification is generally impossible without reference to high-level linguistic constraints. Furthermore, listeners are willing to revise their identification of perceptually ambiguous phonetic material in light of disambiguating material that follows within a short period (Connine, Blasko, & Hall, 1991). In reading, the availability of a stable visual record makes it possible in principle to acquire substantially more detailed perceptual information — but in practice the average fixation length in reading is 200 ms, comparable to the time required to plan a motor saccade. This suggests that the next saccade must be initiated almost as soon as the fixation begins, and that decisions about its timing — and thus the fixation time for the current word — must be made before the current word is fully processed. In addition, Levy, Bicknell, Slattery, and Rayner (2009) have recently used evidence from a reading task to argue that certain syntactic constructions associated with garden-path-like processing difficulty may arise from uncertainty about the identity of critical words earlier in the sentence. Therefore it seems plausible that the language processing system not only has the capacity to handle uncertain input, but that this ability is used in natural reading.

Here, we examine this question via the well-known effect of word predictability on reading time (predictable words are read more quickly, Ehrlich & Rayner, 1981). This is a useful tool, because *(i)* the effect is very early; it affects the duration of initial fixations on a word, in the 200–300 ms range, when we would most expect some uncertainty to remain, and *(ii)*

as word predictability depends on the fit between the present word and its context, it implicates higher-level linguistic processing in a way that word frequency, for instance, might not, and yet *(iii)* it cannot affect processing until the word is fully identified, because different words are differently predictable. All theories which invoke word predictability to explain early reading time measures therefore implicitly assume that word identification occurs early and fully.

We hypothesize that this effect does not arise from predictability *per se*, but from the processing system's 'best guess' at the word's predictability, given the uncertain information available to it. To test this hypothesis, we build a simple Bayesian model of visual word recognition, use it to estimate 'best guess' predictabilities on a corpus, and test whether this improves our ability to predict human reading-time measures.

Word recognition model

We begin with a standard Bayesian model of word recognition in sentence context, in which beliefs about the identity of the word on which the eyes are currently fixated are formed by integrating top-down prior expectations from language knowledge and context with bottom-up perceptual input:

$$P(\text{word}|\text{context, input}) = \frac{P(\text{word}|\text{context})P(\text{input}|\text{word, context})}{P(\text{input})} \quad (1)$$

The first term in the numerator, $P(\text{word}|\text{context})$, corresponds to top-down prior expectations and can be estimated from any of a variety of language-modeling techniques standard in computational linguistics (Manning & Schütze, 1999). The second term in the numerator, $P(\text{input}|\text{word, context})$, corresponds to bottom-up perceptual evidence and is the present focus: we are investigating the possibility that this evidence is imperfect and that this imperfection may be reflected in rapid eye-movement decisions in reading.

We introduce three simplifying assumptions to make our model of perceptual evidence more tractable. First, we assume conditional independence between input and context given word identity, which is natural since it is the word being identified rather than the preceding context that generates the relevant perceptual input. Second, we assume that readers are aware of how many letters exist in the word that they are looking at, and only their identity is in doubt. (A more detailed model would certainly relax this assumption, but we believe that the high visual salience of inter-word spaces makes it a reasonable initial approximation.) Third, we assume that the subjective evidence for a given letter depends only on the noisy input we receive describing that letter (and this noisy input, of course, depends in turn on the letter that is actually present in the world). In particular, we assume that our bottom-up perceptual evidence for each letter in a word is probabilistically independent of that for the other letters. Therefore, we can write the perceptual evidence for a word as simply the product of the evidence for each of the n letters which comprise it. If E is the complete perceptual input derived from a word and E_i is the component of that perceptual input arising from the i-th letter, then normative Bayesian inference for the word's identity looks as follows:

$$\begin{aligned} P(\text{letters}|\text{input}) &= \frac{P(E|\text{letters})P(\text{letters})}{P(E)} \\ &\propto P(E_1,\ldots,E_n|\text{letters})P(\text{letters}) \\ &= P(\text{letters})\prod_{i=1}^{n} P(E_i|\text{letter}_i) \end{aligned}$$

The term $P(\text{letters})$ is simply the prior probability of the word in question; the perceptual evidence for the word is represented by the term $\prod_{i=1}^{n} P(E_i|\text{letter}_i)$.

To estimate the perceptual evidence $P(E_i|\text{letter}_i)$ obtained from each position in the word, we made use of letter-confusion matrices derived from previous norming experiments with the lowercase English alphabet (Engel, Dougherty, & Jones, 1973; Geyer, 1977). In each of these experiments, participants were presented with isolated letters for durations brief enough to induce considerable identification error, and the frequency with some presented letter α was identified as some letter β was tabulated as $f_{\alpha\beta}$. Here $\alpha = \beta$ implies correct identification and $\alpha \neq \beta$ implies misidentification. Finally we used these frequency tables to obtain a matrix M, in which each entry $M_{\alpha\beta}$ denotes the estimated probability of identifying letter α as β. For example, M_{ti} is relatively high, presumably reflecting the visual similarity of the letters t and i, whereas M_{tn} is relatively low.

Since these norming studies used viewing conditions rather unlike those that occur in natural reading, we assume that the matrix entries $M_{\alpha\beta}$ specify only the *relative* perceptual evidence provided by each letter of the word, rather than the *absolute* evidence. We therefore introduce a single free parameter q which scales the matrix as a whole, so that for the i-th letter of a word in a sentence,

$$P(E_i = \alpha|\text{letter}_i = \beta) \propto (M_{\alpha\beta})^q. \quad (2)$$

This allows us to estimate the overall level of noise in the model when analyzing human reading-time data.[1] The parameter q can be interpreted as the overall quantity of information acquired by the reader and used to inform downstream decisions; each entry in the confusability matrix is raised to the power q, and then rows are renormalized. Thus, $q = 0$ creates a uniform posterior distribution over letters, or perfect ignorance, while in the limit as q goes to infinity, the matrix becomes diagonal — representing perfect in-

[1] Note that we are making a simplifying assumption by equating the perceptual evidence from the i-th letter with the letter actually in the word, rather than with noisy perceptual input generated from the actual letter, as is done in models such as (Norris, 2006). This simplifying assumption can be interpreted roughly as marginalizing over the perceptual input itself; see (Smith, Chan, & Levy, 2010) for discussion of the justification for and implications of this simplifying assumption.

formation about letter identity. Varying q between these extremes smoothly varies the overall accuracy of letter information available, while preserving relative differences in letter similarity and recognizability. Figure 1 depicts the resulting letter-confusion matrices for $q = 1$ and $q = 2$.

This idea of rescaling was also used in producing our perceptual confusion matrix M from the raw norming data. We assumed that the two experiments had different overall levels of perceptual noise, and we used maximum likelihood to find the single matrix M that — when rescaled for each experiment — best explained the data from both. However, simply averaging the two norming matrices would produce similar results.

In aggregate, these assumptions give us the following final estimate of the subjective probability that we are observing a particular word given both context and visual input:

$$P(\text{word}|\text{context, visual input}) \propto P(\text{word}|\text{context}) \prod_i P(\text{letter}_i|\text{visual input}). \quad (3)$$

Average neighborhood surprisal

Now that we have a model of the uncertainty affecting the language processing system, we can model its consequences for the predictability effect. Word predictability itself is well-described computationally by surprisal — the negative log-probability of a word in context (Hale, 2001; Levy, 2008). For clarity, in this paper we will refer to this as the *raw surprisal* (RS). We now define the *average neighborhood surprisal* (ANS) of a word in some context to be the average of the surprisal of every word that might occur in that context, weighted by that word's similarity to the visible word, $P(\text{word}|\text{context}, \text{visual input})$. More formally,

$$ANS(\text{word}_k|\text{context}) = \sum_i P(\text{word}_i|\text{context, word}_k) RS(\text{word}_i|\text{context}). \quad (4)$$

Our fundamental prediction is that ANS will better predict reading times than RS.

The intuition here is that the processing system would prefer to spend an amount of time on a word proportional to its RS, but since visual noise makes the RS unavailable, the ANS is the best available approximation. The visual system is accurate enough that in most cases $P(\text{word}_k|\text{context, word}_k)$, the subjective probability that one is looking at word k given that one is, in fact, looking at word k, will be close to one; therefore ANS will generally be close to the RS for any given word. However, if a word has visually similar neighbors with higher surprisals, then this will pull up the ANS, and the reader will spend more time on that word 'just in case' it turns out to be one of those high-surprisal neighbors that require more time to process. Contrariwise, if a word has visually similar neighbors with lower surprisals, then this will pull down the average, and our reader will hurry onward faster than they otherwise might. Note especially that in this model, a word with a dense neighborhood may be read either faster *or* slower than a word with a sparse neighborhood. It's not the size of your neighborhood that matters, it's who your neighbors are.

It should also be noted that other models of neighborhood effects generally predict that the presence of higher-frequency neighbors will produce an inhibitory effect on word identification, as these neighbors interfere with recognition of the true word (e.g., Perea & Rosa, 2000). Our prediction is nearly the opposite — that in reading, the presence of high probability neighbors should lead to shorter initial fixations (though it is possible that later, as more information about the word's true identity becomes available, the eyes may slow or regress in compensation).

Methods

We compared average neighborhood surprisal to raw surprisal as predictors of human reading times in the Dundee eye-movement corpus (Kennedy, Hill, & Pynte, 2003), which consists of all eye-movements made by 10 subjects while reading a collection of newspaper articles totaling approximately 50,000 words. Several previous studies have already demonstrated surprisal effects on reading times in the Dundee corpus (Demberg & Keller, 2008; Frank, 2009; Smith & Levy, 2008). We analyzed both first fixation times — defined as the duration of the first fixation to land on each fixated word in a text — and second fixation times, defined as the duration of the second fixation to land on each word that was fixated a second time. We eliminated all fixations on words that occurred at the beginning or end of a line, which preceded or followed punctuation, that did not occur in the BNC (i.e., unknown words), or that occurred in the BNC but in segmented form (e.g., the BNC codes *don't* as two words, *do* followed by *n't*). Finally, we eliminated any remaining words containing uppercase letters, since our confusion norms only cover the lowercase alphabet. This left 182,169 first fixations and 42,024 second fixations for further analysis.

To obtain conditional word probabilities for both raw surprisal estimates and noisy conditional word-probability estimates (Equation 3) we used a trigram language model trained on the 100 million word British National Corpus (BNC), using the SRI Language Modeling Toolkit (Stolcke, 2002); the trigram model was smoothed using modified Kneser-Ney (Kneser & Ney, 1995), a standard technique for broad-coverage language modeling. Average neighborhood surprisal was estimated for each fixated word by plugging in raw surprisal estimates to Equation (4), and repeating this process at each value of q required by the fitting process.

As Smith and Levy (2008) have previously demonstrated that the relationship between surprisal and first fixation times in this corpus is linear, we simply regressed fixation time on RS and ANS simultaneously, with frequency (estimated from the BNC) and word length as controls. The noise parameter q was fit simultaneously with the regression coefficients by maximum likelihood. Gamma distributed error was assumed,

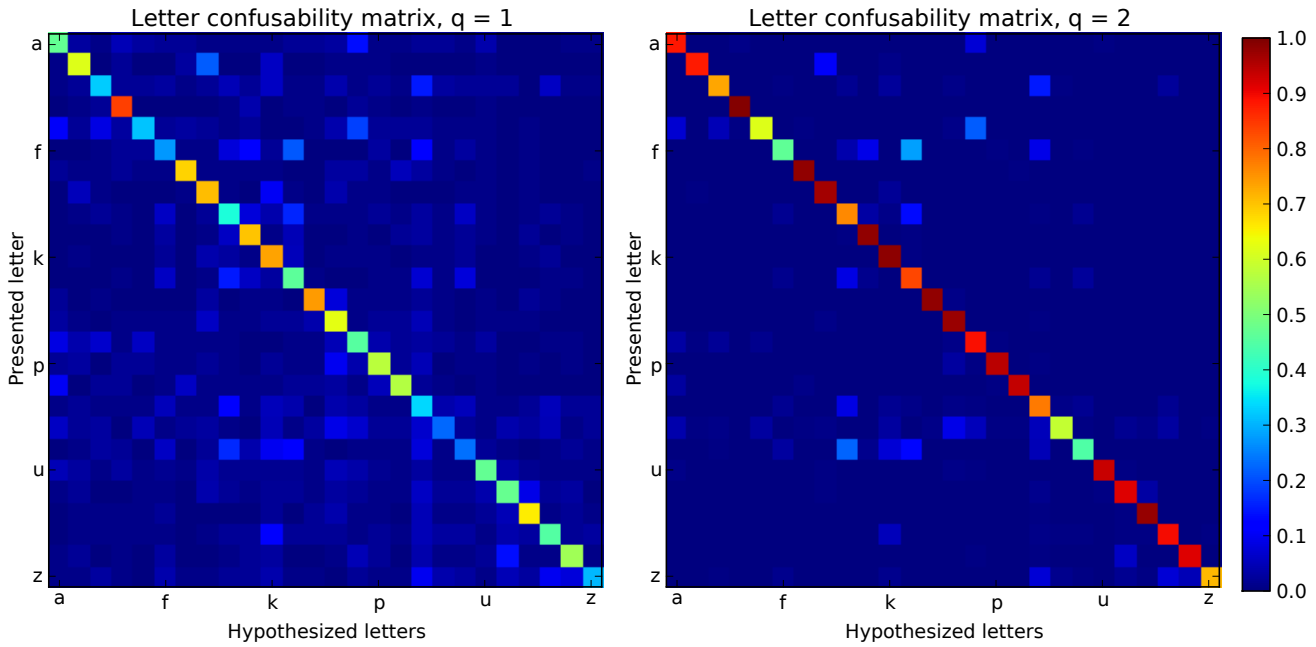


Figure 1: The letter confusability matrix, for different values of the scaling parameter q. For instance, presentation of the letter a to the noisy input system eventually gives rise to a particular posterior distribution over letters that is represented by the top row in each matrix. The diagonal represents the probability of veridical perception; we can see that the letter d is the least confusable in the lowercase English alphabet. As q increases (right), more information becomes available, causing the posterior distribution to cluster around the diagonal.

in order to properly account for the long right-ward tail in fixation durations.

Results

First fixations

The best fitting model had a moderate level of noise ($q = 1.306$), corresponding to a mean naming accuracy for individual letters of 66%. While this may seem low, most words contain enough letters that, combined with the constraint of linguistic context, this allows for substantial information about word identity. As a result, ANS and RS are highly correlated ($R^2 = 0.96$) — suggesting that while the models differ greatly in terms of the cognitive processes they postulate, they may be difficult to disentangle experimentally.

Even so, our data set turned out to be large enough for the regression model to give an unambiguous result: ANS better predicts human behavior than RS. That is, ANS is highly significant after controlling for RS ($t(182155) = 4.164, p \ll 0.001$), while RS has no significant effect after controlling for ANS ($t(182155) = 0.489, n.s.$). This result also remains after controlling for neighborhood size (N).

Second fixations

The same analysis on second fixations produces analogous results; ANS is highly significant ($t(42010) = 4.209, p \ll 0.001$), while RS is marginally significant in the wrong direction ($t(42010) = -1.847, p = 0.06$). More interesting, however, is examination of the q value; we predicted that a second fixation would provide more visual information about word identity, and thus result in a higher q. In fact, for second fixations, we found $q = 2.939$, suggesting that by the end of the second fixation, the eye movement control system has access to somewhat more than twice the information it has at the end of the first fixation.

Frequency prior

Equation (4) suggests that to compute the estimated, average neighborhood surprisal, the processing system must be able to, in some sense, compute the probability of *all possible* words in the current context, and sum over all of them, in time to affect the first fixation. This is a strong claim, and so to test it we calculated a simplified version of ANS in which we modified Equation 3 to replace the context-sensitive prior over words, $P(\text{word}|\text{context})$, with a simple, context-insensitive word frequency prior, $P(\text{word})$. This modified ANS was then added to our regression as an additional control. Our original context-sensitive ANS remained highly significant ($t(182154) = 4.413, p \ll 0.001$), suggesting that in the neighborhood effects we describe, the definition of 'neighborhood' is indeed sensitive to linguistic context.

Other determiners of reading time

While in this preliminary work we have focused on surprisal as a model reading time predictor, the essential argument applies to any word property which is believed to affect reading time, and one could define *average neighborhood X* for any interesting property *X* that was believed to affect reading time (or language processing behavior more generally). Generally, we would predict that to the extent the brain processes sensitive to property *X* must work from noisy representations of linguistic input, *average neighborhood X* would also be a better predictor of human behavior than *X* alone.

We have begun to examine this more general prediction, and in the process discovered a mystery. Using the above model to define average neighborhood word frequency, we find our regression against reading times gives just as unambiguous results as for surprisal — but the other way. That is, raw frequency is significant, and average neighborhood frequency is not. This suggests that whatever process produces word frequency effects in reading times appears to have exact information about the frequency (and therefore identity) of the word being processed, while the process which produces predictability effects has only noisy and imperfect information. Furthermore, this is true even on first fixations, so it cannot be a simple matter of the frequency effect arising later in the processing stream, when more information is available. (Evidence for frequency as a later effect than predictability would also, it seems safe to say, surprise most experts in the field.)

Discussion

Our fundamental prediction — that early predictability effects in reading are modulated by the predictability of visually similar (but unseen) words — was confirmed. Furthermore, the reduction of this effect on second fixations gives insight into the time course for resolution of uncertainty about word identity, and the failure of the word frequency prior to adequately explain the data argues for the ability of high-level linguistic constraint to quickly and robustly modulate the resolution of visual uncertainty. All our results — with the possible exception of the mysterious frequency non-effect — are compatible with a model of reading in which uncertainty about the input is propagated forward into the linguistic processing system itself.

Going forward, a major question is whether the noise we observe is truly visual noise, or whether it has another source. After all, biological computation necessarily involves noise and uncertainty at every level. When reading, for example, visual information must be gathered at the retina, transmitted and analyzed by the visual system, and converted to some higher level representation of word identity; then, this representation must be maintained in memory for semantic processing and integration. None of these processes can be perfectly veridical or reliable; all must introduce some amount of noise and uncertainty. Here, we built a specifically visual noise model, relying on a visual confusability matrix and a letter-based word representation, but presumably all models of word similarity/confusability are similar to the first order, and we did not compare against any other noise model; therefore, while our results suggest that average neighborhood surprisal drives reading time, it may be premature to conclude that the visual system is the source of uncertainty being averaged over.

In future work, we hope to make a sharper test of this part of the model in two ways. First, we can fit a different noise parameter q for letters at different degrees of eccentricity from visual fixation; if this reproduces the classic curve of acuity falling off with increasing eccentricity, then that would be stronger evidence that our noise arises from visual processing limitations. Second, looking the other direction, we plan to build a simple phonological/auditory noise model, and use it to estimate ANS for written words. If this model outperforms the visual noise model, then that would be strong evidence that the noise is in fact noise in some post-recoding internal representation. Finding auditory noise in a visual paradigm would be quite curious, but there is some precedent; for instance, it has been argued that the true determiner of neighborhood size for purposes of word naming effects is the number of words which are simultaneous visual and phonological neighbors (Adelman & Brown, 2007).

Finally, we hope that further investigation may shed light on the lack of a neighborhood effect on word frequency. One possibility is that further study of the noise, as described above, will provide a clue — perhaps visual information is highly accurate, the frequency effect is a relatively early and low-level effect acting on this low-level, accurate visual representation, and the predictability-sensitive process is working with a later representation more subject to internal noise. However, this remains mere speculation, and we welcome any suggestions on this matter. In another way, though, this dissociation of frequency and predictability is quite exciting, as it suggests a possible avenue for understanding the relationship between these highly similar linguistic properties. (Indeed, as they are inherently confounded in any study using isolated words stripped of context, and quite difficult to accurately measure and deconfound in more naturalistic stimuli, it has long been unclear whether they represented distinct effects at all.) This is, to our knowledge, the first study to find qualitatively different effects of each, and we hold high hopes that our current confusion may lead to a deeper future understanding.

Acknowledgments

We are grateful to Michael Tanenhaus for the initial suggestion of averaging surprisal over the visual neighborhood, and to Shane T. Mueller for maintaining the invaluable Letter Similarity Data Set Archive (`http://obereed.net/lettersim/`). This research was partially supported by NIH Training Grant T32-DC000041 to the Center for Research in Language at UC San Diego to NJS, and by NSF grant 0953870 to RL.

References

Adelman, J. S., & Brown, G. D. A. (2007). Phonographic neighbors, not orthographic neighbors, determine word naming latencies. *Psychonomic Bulletin & Review*, *14*, 455–459.

The British National Corpus, version 3 (BNC XML edition). (2007). (Distributed by Oxford University Computing Services on behalf of the BNC Consortium. URL: `http://www.natcorp.ox.ac.uk/`)

Connine, C. M., Blasko, D. G., & Hall, M. (1991). Effects of subsequent sentence context in auditory word recognition: Temporal and linguistic constraints. *Journal of Memory and Language*, *30*(2), 234–250.

Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, *109*(2), 193–210.

Ehrlich, S. F., & Rayner, K. (1981). Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning and Verbal Behavior*, *20*(6), 641–655.

Engel, G. R., Dougherty, W. C., & Jones, G. B. (1973). Correlation and letter recognition. *Canadian Journal of Psychology*, *27*(3), 317–326.

Frank, S. L. (2009). Surprisal-based comparison between a symbolic and a connectionist model of sentence processing. In *Proceedings of the 31st annual conference of the cognitive science society* (pp. 1139–1144).

Geyer, L. H. (1977). Recognition and confusion of the lowercase alphabet. *Perception and Psychophysics*, *22*, 487–490.

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of NAACL-2001* (pp. 159–166).

Kennedy, A., Hill, R., & Pynte, J. (2003). The Dundee corpus. In *Proceedings of the 12th European conference on eye movement.*

Kneser, R., & Ney, H. (1995). Improved backing-off for M-gram language modeling. In *Proc. ICASSP* (pp. 181–184).

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*, 1093–1582.

Levy, R., Bicknell, K., Slattery, T., & Rayner, K. (2009). Eye movement evidence that readers maintain and act on uncertainty about past linguistic input. *Proceedings of the National Academy of Sciences*, *106*(50), 21086–21090.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT Press.

Norris, D. (2006). The Bayesian reader: Explaining word recognition as an optimal Bayesian decision process. *Psychological Review*, *113*(2), 327–357.

Perea, M., & Rosa, E. (2000). The effects of orthographic neighborhood in reading and laboratory word identification tasks: A review. *Psicológica*, *21*(3), 327–340.

Smith, N. J., Chan, W.-H., & Levy, R. (2010). Is perceptual acuity asymmetric in isolated word recognition? evidence from an ideal-observer reverse-engineering approach. In *Proceedings of the 32nd annual meeting of the cognitive science society.*

Smith, N. J., & Levy, R. (2008). Optimal processing times in reading: a formal model and empirical investigation. In *Proceedings of the thirtieth annual conference of the Cognitive Science Society.*

Stolcke, A. (2002). SRILM — an extensible language modeling toolkit. In *Proc. intl. conf. on spoken language processing* (Vol. 2, pp. 901–904). Denver.

An fMRI Study of Strategic Reading Comprehension

Jarrod Moss (jarrod.moss@msstate.edu)
Department of Psychology, Mississippi State University,
Mississippi State, MS 39762 USA

Christian D. Schunn (schunn@pitt.edu)
Walter Schneider (wws@pitt.edu)
Learning Research and Development Center, University of Pittsburgh
Pittsburgh, PA 15260 USA

Danielle S. McNamara (dsmcnamara1@gmail.com)
Department of Psychology, University of Memphis
Memphis, TN 38152 USA

Kurt VanLehn (kurt.vanlehn@asu.edu)
School of Computing, Informatics and Decision Systems Engineering, Arizona State University
Tempe, AZ 85287 USA

Abstract

While there have been neuroimaging studies of text comprehension, little is known about the brain mechanisms underlying strategic learning from text. It was hypothesized that reading strategies would involve areas of the brain that are normally involved in reading comprehension along with areas that are involved in strategic control processes because the readers are intentionally using a complex learning strategy. The present study was designed to answer the question of what brain areas are active during performance of complex reading strategies. Activation was found in both executive control and comprehension areas, and furthermore, learning gains were found to be associated with activation in the anterior prefrontal cortex (aPFC).

Keywords: Reading Strategies; fMRI; Cognitive Control

Introduction

The importance and difficulty of comprehending expository text is obvious to anyone who has tried to learn about a new field of science by reading a textbook. The complexity of text comprehension and learning processes results in large individual differences in the strategies that students engage in to understand texts and what students extract from texts (e.g., Chi, Bassok, Lewis, Reimann, & Glaser, 1989; McNamara, 2004). While there have been neuroimaging studies of text comprehension, these studies have not examined the differences in brain activity associated with different reading strategies. Thus, understanding the neural correlates of different types of strategic reading comprehension should help us to better understand both the brain mechanisms underlying comprehension as well as the way in which these strategies affect comprehension.

There have been a number of neuroimaging studies that have investigated the brain areas involved in text comprehension (e.g., Xu, Kemeny, Park, Frattali, & Braun, 2005; Yarkoni, Speer, & Zacks, 2008). These studies show that a network of neural regions are used in text comprehension including inferior frontal and temporal areas associated with language comprehension and production as well as areas distributed throughout the temporal, parietal, and frontal cortices that appear to be associated with building coherent representations of texts.

When contrasting sentence-level processing with narrative-level processing, Xu et al. (2005) identified a network of areas including the hippocampus, caudate, thalamus, prefrontal cortex, precuneus, posterior cingulate, and angular gyrus. Hippocampal areas are likely associated with memory formation and retrieval. They hypothesized that the caudate, thalamus, and prefrontal cortex were involved in the sequencing of higher-level processes associated with reading comprehension. Medial prefrontal cortex, precuneus, and posterior cingulate were hypothesized to be involved with linking text content with global themes and other information in memory, and the angular gyrus was hypothesized to be involved in the mental scanning of spatial representations built from the text.

A number of the areas involved in discourse comprehension are also considered part of the brain's default network that is active when people are not engaged in an external task (Buckner, Andrews-Hanna, & Schacter, 2008). Some studies of discourse processing have noted this overlap between the default network and areas active during comprehension (e.g., Xu et al., 2005; Yarkoni, Speer, Balota, McAvoy, & Zacks, 2008). The default network has been associated with self-referential processing and the generation of coherent mental representations (Hassabis & Maguire, 2007). If the reader's goal is to form a coherent representation of the text, then these processes would be involved in all forms of comprehension including strategic reading comprehension.

Reading comprehension strategies improve readers' comprehension of text. Some readers use strategies naturally, and others benefit from being provided with strategy instruction. Self-explanation is one reading strategy

that has been shown to be effective at improving readers' comprehension when students are trained or prompted to use it (Chi et al., 1994; McNamara, 2004).

Because instructing readers to self-explain often benefits readers who are skilled self-explainers more than less skilled self-explainers (Chi et al., 1994), McNamara (2004) developed Self-Explanation Reading Training (SERT) in which students are provided with instruction and practice on using reading strategies while self-explaining texts. This approach combined the technique of self-explanation with reading strategies with demonstrated effectiveness. SERT includes five component reading strategies: comprehension monitoring, paraphrasing, elaboration, bridging, and prediction (McNamara, 2004). Comprehension monitoring is being aware of whether the text is being successfully understood while reading. Paraphrasing is putting the text into one's own words. The process of putting text into one's own words helps to activate relevant semantic knowledge in long-term memory and prepares the reader to make further inferences. Inferences are necessary in most text comprehension situations because most texts do not state all relevant pieces of information explicitly (Kintsch, 1998). Elaboration involves making inferences that aid in understanding the text by using knowledge from memory. Bridging involves making inferences across sentence boundaries to aid in understanding the text. Prediction is making predictions at the end of a sentence or paragraph about what information will be contained in the next section of the text. Collectively, these strategies help the reader to process challenging, unfamiliar material by scaffolding the comprehension process. The process of self-explaining externalizes the comprehension process and the reading strategies help the reader to understand the text (i.e., using paraphrasing and comprehension monitoring) and go beyond the text by generating inferences (i.e., using elaboration, bridging, and prediction).

Because self-explanation is a strategy that enhances existing comprehension processes, then it can be expected to involve areas of the brain that are normally involved in reading comprehension along with areas that are involved in strategic control processes. A network of brain areas have been shown to be active in a variety of tasks involving executive control (Chein & Schneider, 2005; Cole & Schneider, 2007). This control network includes dorso-lateral prefrontal cortex (DLPFC), anterior cingulate cortex (ACC), pre-supplementary motor area (pSMA), dPMC, anterior insular cortex (AIC), inferior frontal junction (IFJ), and posterior parietal cortex (PPC). These areas have been shown to be active in a variety of tasks involving executive control (Chein & Schneider, 2005; Cole & Schneider, 2007). Because learning strategies such as self-explanation are effortful and complex, we hypothesize that this executive control network will be active during self-explanation. We expect lower levels of activation for less complex learning strategies that do not involve as much effort and management of complex information, such as simple paraphrasing or rereading of information.

The present study contrasted these three learning strategies—rereading, paraphrasing, and self-explaining—differing in complexity and effectiveness. Rereading is commonly used as a reading strategy but has been found to be less effective than self-explanation and is often used as a control condition to evaluate the effectiveness of self-explanation training (Chi et al., 1994). Paraphrasing a text to put it into one's own words is another learning strategy that could be used to aid comprehension. It was predicted that more complex strategies would show more engagement of the executive control network as well as greater activation of areas that previous studies have associated with text comprehension. It is an open question whether strategy effectiveness is primarily a function of more engagement (as measured by activation of the executive control network) or primarily a function of specific text comprehension processes beyond the executive control components. In addition to examining activation associates with each learning strategy, this study will examine if there are areas that are associated with measurable learning gains.

Method

Participants

Twenty-two right-handed, native English speakers were recruited from the University of Pittsburgh and Carnegie Mellon University communities (14 female, M age = 20.7; SD = 2.4; range = 18-28). None of the participants were biology majors. One participant was excluded from analysis due to excessive head motion during the scanning session.

Materials and Design

Three biology texts that were matched on length were selected along with a set of 15 short-answer questions for each text. Text and question difficulty were equated using data from a pilot study in which students answered the questions before and after reading and self-explaining the texts. The texts were separated into 12 paragraphs, each containing 2-4 sentences. Each participant performed all three learning strategies: rereading, paraphrasing, and self-explaining. Each participant was instructed to use a given learning strategy to read all of a given text. The assignment of learning strategies to texts was counterbalanced across participants. The order in which participants performed the strategies was randomized. Each text was presented over three blocks consisting of four paragraphs each. Each block of paragraphs for each of the texts was presented before the next block of paragraphs for each text (e.g., Text1-Block1, Text2-Block1, Text3-Block1, Text1-Block2, …).

Procedure

This study took place over two sessions, separated by 2-5 days, with fMRI data collected only in the second session.

Session 1 During the first session, participants were given up to 30 minutes to complete a pretest including all of the questions for each of the three texts. Participants then

completed an iSTART session which provided instruction on how to self-explain using reading strategies.

iSTART provides high school students with instruction and practice on how to self-explain texts using the five SERT strategies described in the introduction. iSTART is described in greater detail by McNamara and colleagues (2004). iSTART training took approximately 90 minutes.

After iSTART training, the participants were provided with task practice in an MRI simulator. The MRI simulator was designed to closely simulate the physical conditions of the MRI scanner and included a magnetic tracking system to track and present feedback to the participant regarding head movement. Participants were presented with paragraphs from a practice text that was of a similar expository nature but contained different content than the texts in the experiment. Before each block of paragraphs, participants read instructions on the screen indicating the learning strategy they were to use for that block.

The title of the text was centered on the top of the screen with the paragraph appearing on the center of the screen. Along the bottom of the screen was a prompt reminding the participant of the current strategy. Participants were instructed to read the paragraph aloud once, and then to press a button on a response glove. Once they did so, the color of the paragraph's text changed from black to blue which served as a cue that they were to perform the given learning strategy aloud. The participants then reread, paraphrased, or self-explained the text and pressed a button.

The paraphrasing and self-explanation strategies were introduced within iSTART, and thus, participants were provided only brief instructions on how to either paraphrase or self-explain out loud each sentence in the text. In the paraphrase condition, participants were told to put each sentence in the paragraph into their own words without using any of the self-explanation strategies. In the self-explanation condition, participants were instructed to self-explain each paragraph using the reading strategies covered in iSTART. For the rereading condition, they were told to read and then reread each paragraph out loud until the computer indicated it was time to move to the next paragraph of text. A prompt, which flashed at the bottom of the screen, instructed the participant to stop rereading and move on to the next paragraph. The rereading condition was designed this way in order to roughly equate the amount of time spent rereading with the amount of time spent paraphrasing and self-explaining. The amount of time allotted for rereading was 45 seconds, which was determined from a pilot study in which participants applied the three strategies to the same texts.

Session 2 The second session occurred 2-5 days after the first session in order to reduce the chance that participants would read the passages with the pretest questions in mind. This session began with an iSTART practice session, which gave the participants additional practice self-explaining. fMRI data was collected for the remainder of the session. All tasks were presented using E-Prime (Psychology Software Tools, Inc., Pittsburgh, PA) on a Windows PC for task presentation and response collection. To verify strategy use within each condition, verbal responses were collected using an active noise canceling microphone system (Psychology Software Tools, Inc., Pittsburgh, PA), which almost entirely removed the scanner background noise.

A 30-second rest period was placed before and after each block of paragraphs. A fixation cross was presented in the middle of a white screen for the rest period. Participants were told to relax and to try not to think about anything during this time. The participants completed a total of 9 blocks with each block consisting of 4 paragraphs (3 blocks for each text/learning strategy pair). Following these 9 learning blocks, participants were presented with a posttest for each text. Although the posttest was collected in the scanner, we do not examine this fMRI data in this paper.

After the posttest, participants were presented with a line search task that served as a functional localizer to localize activity in control areas. The task involved detecting a target line orientation by monitoring lines of differing orientation in four locations on the screen. The lines in these four locations changed over time, and the participants were asked to press a button when one of the locations matched the target orientation. This task has been used in prior research on executive control (Cole & Schneider, 2007).

In order to increase statistical power in the learning comparison across learning strategy conditions while constraining the number of fMRI participants, a second group of 14 behavioral participants participated using the same learning strategy paradigm outside of the scanner.

Data Acquisition and Analyses

Structural and functional images were collected on a whole body Siemens Trio 3-T scanner at the University of Pittsburgh during a 2-hour scanning session. The functional runs were acquired as 39 oblique-axial slices parallel to the AC-PC plane using a T2*-weighted echo-planar imaging pulse sequence (TE = 25 ms, TR = 2000 ms, FOV = 21, thickness = 3.5 mm, flip angle = 76, in-plane = 3.28 mm^2).

The raw neuroimaging data were preprocessed and analyzed using the AFNI software package (Cox, 1996). All functional images were realigned to the first image of each run, which were aligned to the first run of each participant. The images were then transformed into Talairach space (Talairach & Tournoux, 1988). For visualization, statistical maps were mapped onto the cortical surface using Caret (Van Essen et al., 2001).

Analyses of the fMRI data used voxel-based statistical techniques. Unless otherwise specified, all results were corrected for multiple comparisons using family-wise error (FWE) cluster size thresholding. At the individual participant level, general linear models were fit to the data using a set of boxcar functions for the conditions of interest convolved with a standard hemodynamic response function. Each group-level analysis used a mixed effects model with participants treated as a random factor.

The line search task was used to define participant-specific regions of interest (ROIs) for the six bilateral areas

of the control network. Local peaks of activation corresponding to the anatomical location of the control net areas were used to identify each ROI. All statistically significant voxels within a sphere of radius 15 mm from the peak were included in the ROI.

Results

Behavioral Results

The proportion correct on the pretest and posttest were used to calculate a learning gain score, where gain = (posttest – pretest) / (1 – pretest). This gain score adjusts for the fact that questions already answered correctly on the pretest cannot be improved upon on the posttest (Cohen, Cohen, Aiken, & West, 1999). Due to technical difficulties, the recordings from a portion of two participants' posttests were not available to be scored. These missing scores corresponded to the paraphrase strategy for one participant and the self-explanation strategy for another.

The gain scores for the behavioral and imaging participants did not differ on any of the three conditions (for all comparisons, $p > .3$), so the data for these two groups were combined for the analyses of the effect of strategy on learning. Planned comparisons showed that rereading gain (M = .41, SD = .26) did not differ from paraphrasing (M = .42, SD = .22), $t < 1$. As expected, self-explanation led to greater learning (M = .51, SD = .19) than paraphrasing, $t(32) = 2.41$, $p = .02$, Cohen's d = 0.4, and rereading, $t(33) = 2.03$, $p = .05$, Cohen's d = 0.4. With a relatively short learning period for complex science materials and a short delay between learning and test, these moderately-sized condition differences in learning were as expected.

The verbal protocols were transcribed, and the self-explanation for each paragraph was coded for whether it contained each of the five techniques comprising self-explanation. Agreement between two independent coders was reliable, 89% agreement (Cohen's kappa = .66). The self-explanation coding was used to determine whether participants were performing the strategy that they had been instructed to perform. All participants in the imaging portion of the study performed the line search task well; d' was greater than 2 for all participants.

Imaging Results

In order to directly examine differences in activation between the different strategies, a voxel-wise ANOVA with strategy (reread, paraphrase, self-explain) as a within-participant factor was conducted followed by three planned contrasts (paraphrase – reread, self-explain – reread, and self-explain – paraphrase). Contrasts were done using the strategy participants had been instructed to perform as well as using the self-explanation coding process described above to determine the condition. If a participant did not use any self-explanation strategy other than paraphrasing during a self-explanation, then it was classified as being a paraphrase. This reclassification resulted in an average of 1.7 out of 12 self-explanations per participant being reclassified as paraphrases. The fMRI results were similar for both versions of this analysis so only the reclassified analysis is reported.

The areas more active for self-explanation compared to rereading are shown in Figure 1. The areas in the contrast between paraphrasing and rereading were a subset of these areas. Self-explanation and paraphrase both involve greater activation of the control network. These areas include DLPFC, IFJ, AIC, ACC/pSMA, PPC, and dPMC. The results are consistent with the notion that control activity increases with more complex learning strategies.

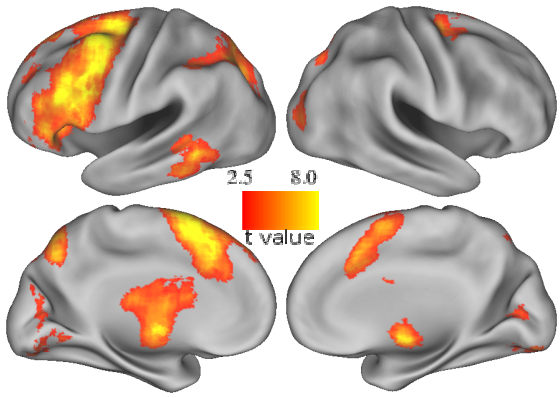


Figure 1. Areas more active while self-explaining than while rereading.

However, the contrast between the self-explanation and paraphrase conditions shows a different pattern of results as seen in Figure 2. These areas included posterior cingulate, precuneus, angular gyrus, middle temporal gyrus, and aPFC. Many of these areas are known to be part of the default network (Buckner et al., 2008). These results indicate that control areas do not account for the differential learning when using self-explanation and paraphrasing.

Analysis of areas that were active in the line search task indicated that all six areas of the control network were active, as expected. The amount of activation in control areas during performance of the learning strategies was examined by using the active voxels in a participant's line search task to identify ROIs for that participant. Average percent signal change was examined in these areas for each of the three learning strategies relative to the rest condition. For the average activation averaged across all ROIs, self-explanation and paraphrase both activated control areas more than reread, $F(1,20) = 8.94$, $p = .007$, $F(1,20) = 18.40$, $p < .001$, respectively. However, self-explanation and paraphrase did not differ in control area activation, $F(1,20) = 2.45$, $p = .13$. This analysis of control areas is consistent with the findings shown in Figures 1 and 2. Self-explanation and paraphrase do not differ in control activation.

The previous analysis examined areas that were active when participants were self-explaining. However, an alternative approach is to examine those times when it led to measurable learning. Thus, a separate analysis was conducted to examine whether there were brain regions with

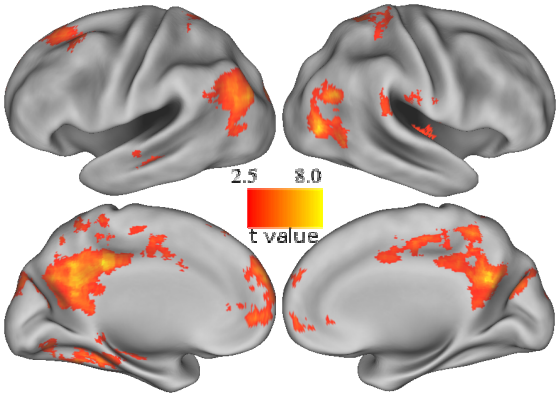


Figure 2. Areas more active while self-explaining than while paraphrasing.

activity associated during successful learning with self-explanation over and above that seen for self-explanation in general. This was achieved by creating an amplitude-modulated regressor in addition to the strategy regressor for the self-explanation runs. The amplitude of this regressor was based on the gain score for a particular slide. The gain score for each slide was calculated by first determining for each question on which slide the information to answer the question was presented. Some slides may have mapped to multiple questions. In this case, the average gain across all questions mapping to that slide was calculated. The regressor for the analysis was formed by convolving a boxcar function whose value was determined by the gain score with a hemodynamic reference function. This process was used to identify brain areas exhibiting a linear relation to gain scores (Buchel, Holmes, Rees, & Friston, 1998).

This learning analysis identified a set of bilateral pre-frontal areas that were positively associated with learning gain. These areas are shown in Figure 3. There were no areas negatively associated with learning gains. In addition to the areas which were active during self-explanation, these pre-frontal areas were more active during self-explanation trials during which material was learned well enough to be answered correctly on the posttest.

Discussion

The results presented here provide evidence that complex learning strategies engage executive control regions, semantic/comprehension regions, and bilateral aPFC. The behavioral learning results confirmed that the three learning strategies differed in effectiveness as hypothesized. Comparing the least complex strategy, rereading, with the next most complex strategy, paraphrasing, showed that predominantly areas known to be involved in executive control were more active for the more complex strategy. This is consistent with our initial hypothesis that more complex strategies would require more cognitive control.

However, the control network was not more active for self-explanation than it was for paraphrasing. The effectiveness of self-explanation was never expected to be

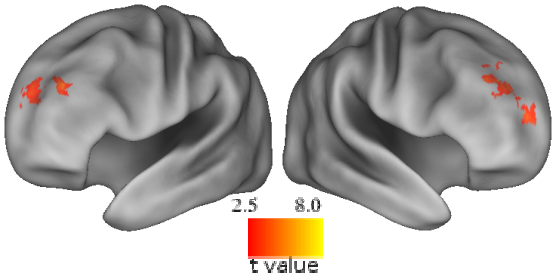


Figure 3. Areas linearly related to measurable learning gain during self-explanation.

solely due to the controlled effort involved, but it is interesting that the more effective learning strategy requires a similar amount of control activity as a less effective one.

The areas that were more active for self-explanation than the less effective strategies include areas associated with text comprehension, memory, and the default network. Areas previously shown to be associated with text comprehension that were more active during paraphrasing and self-explaining include L IFG, caudate, thalamus, PFC, bilateral precuneus, posterior cingulate, bilateral PPC, L parahippocampal gyrus, and L angular gyrus. Given that both paraphrasing and self-explanation usually lead to better comprehension than does rereading, it seems likely this network of areas are performing similar comprehension-related processing during performance of these reading strategies. In particular, the bilateral angular gyrus, right PPC, bilateral precuneus, bilateral posterior cingulate, left fusiform gyrus, and left parahippocampal gyrus were most active only in the self-explanation condition. Of these areas, PPC, left fusiform, and right precuneus have been previously been implicated in the construction and updating of situation models (Yarkoni et al., 2008). The angular gyrus, posterior cingulate, and precuneus have been associated with relating text to prior knowledge and the use and manipulation of mental models (Xu et al., 2005). The areas active in the MTG active in self-explanation are similar to areas that have been found when people draw inferences during text comprehension (Virtue, Haberman, Clancy, Parrish, & Jung Beeman, 2006). These are exactly the kinds of cognitive processes that a strategy such as self-explanation is supposed to engage to support deep comprehension of the text.

A number of the areas more active in the self-explanation condition than in the paraphrase condition are considered to be part of the default network that is active in the absence of goal-directed activity (e.g., Buckner et al., 2008). The areas of the default network typically include mid-orbital cortex, angular gyrus/inferior parietal, lateral temporal cortex, and the hippocampus. These areas were highly active during self-explanation. One hypothesis about the default network is that it is associated with an internal stimulus-independent mode of thought (Buckner et al., 2008). These stimulus-independent thoughts have been associated with lapses in attention and mind wandering (Christoff, Gordon, Smallwood, Smith, & Schooler, 2009), but this mode of

thought is also thought to have adaptive purposes including retrieval of episodic and semantic memory along with the generation of coherent and use of coherent mental representations (e.g., Hassabis & Maguire, 2007). The retrieval of prior knowledge and the generation of coherent representations during the use of reading strategies likely make use of these same brain areas.

The analysis of the areas active during self-explanation that were correlated with the amount learned mainly included bilateral aPFC. That is, in addition to the activity in executive control and default network areas associated with self-explanation, the aPFC was more active during self-explanation of paragraphs where measurable learning took place. The aPFC is active during performance of a number of higher-order tasks, but a recent theory of aPFC function refers to it as a router or gateway between modes of thought (Burgess, Dumontheil, & Gilbert, 2007). One mode of thought is one in which external representations drive thought, and the other mode is one in which internal representations drive thought. This gateway hypothesis might help to explain the correlation of the aPFC with learning in this study. The aPFC might be helping to coordinate the reading and processing of the text presented on the screen with the internal retrieval of memories and construction of situation models. It may also reflect the coordination of an explicit strategy with the internal thought processes associated with the default network. Self-explanation may be most effective when there is strategic processing of internal representations.

This initial exploration of the neural correlates of strategic reading comprehension has shown that a network of areas associated with executive control and the manipulation of internal representations and memories underlie the effectiveness of these strategies. Future work should explore the role of aPFC in reading strategies as well as whether these results will generalize to other texts.

Acknowledgments

This work was supported by The Defense Advanced Research Projects Agency (NBCH090053). The views, opinions, and/or findings contained in this article are those of the authors and should not be interpreted as representing the official views or policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the Department of Defense. Approved for Public Release, Distribution Unlimited.

References

Buchel, C., Holmes, A. P., Rees, G., & Friston, K. J. (1998). Characterizing stimulus-response functions using nonlinear regressors in parametric fMRI experiments. *NeuroImage*, 8(2), 140-148.

Buckner, R. L., Andrews-Hanna, J. R., & Schacter, D. L. (2008). The brain's default network: Anatomy, function, and relevance to disease. *Annals of the New York Academy of Sciences*, 1124, 1-38.

Burgess, P. W., Dumontheil, I., & Gilbert, S. J. (2007). The gateway hypothesis of rostral prefrontal cortex (area 10) function. *Trends in Cognitive Sciences*, 11(7), 290–298.

Chein, J. M., & Schneider, W. (2005). Neuroimaging studies of practice-related change: fMRI and meta-analytic evidence of a domain-general control network for learning. *Cognitive Brain Research*, 25(3), 607-623.

Chi, M. T. H., Deleeuw, N., Chiu, M. H., & Lavancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, 18(3), 439-477.

Christoff, K., Gordon, A. M., Smallwood, J., Smith, R., & Schooler, J. W. (2009). Experience sampling during fMRI reveals default network and executive system contributions to mind wandering. P*roceedings of the National Academy of Sciences*, 106(21), 8719-8724.

Cohen, P., Cohen, J., Aiken, L. S., & West, S. G. (1999). The problem of units and the circumstance for POMP. *Multivariate Behavioral Research*, 34(3), 315-346.

Cole, M. W., & Schneider, W. (2007). The cognitive control network: Integrated cortical regions with dissociable functions. *Neuroimage*, 37(1), 343-360.

Cox, R. W. (1996). AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research*, 29(3), 162-173.

Hassabis, D., & Maguire, E. A. (2007). Deconstructing episodic memory with construction. *Trends in Cognitive Sciences*, 11(7), 299-306.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge: Cambridge Univ Press.

McNamara, D. S. (2004). SERT: Self-explanation reading training. *Discourse Processes*, 38(1), 1-30.

McNamara, D. S., Levinstein, I. B., & Boonthum, C. (2004). iSTART: Interactive strategy training for active reading and thinking. *Behavior Research Methods Instruments & Computers*, 36(2), 222-233.

Van Essen, D. C., Drury, H. A., Dickson, J., Harwell, J., Hanlon, D., Anderson, C. H., (2001). An integrated software suite for surface-based analyses of cerebral cortex. *Journal of the American Medical Informatics Association*, 8, 443-459.

Virtue, S., Haberman, J., Clancy, Z., Parrish, T., & Jung Beeman, M. (2006). Neural activity of inferences during story comprehension. Brain Research, 1084(1), 104–114.

Xu, J., Kemeny, S., Park, G., Frattali, C., & Braun, A. (2005). Language in context: emergent features of word, sentence, and narrative comprehension. *NeuroImage,* 25(3), 1002–1015.

Yarkoni, T., Speer, N. K., Balota, D. A., McAvoy, M. P., & Zacks, J. M. (2008). Pictures of a thousand words: Investigating the neural mechanisms of reading with extremely rapid event-related fMRI. *NeuroImage*, 42(2), 973-987.

Yarkoni, T., Speer, N. K., & Zacks, J. M. (2008). Neural substrates of narrative comprehension and memory. *NeuroImage*, 41(4), 1408-1425.

Gricean Brainwaves:
Brain Responses to Pragmatic Violations in Dialogues

John C. J. Hoeks (j.c.j.hoeks@rug.nl)
Center for Language and Cognition, University of Groningen,
PO Box 716, 9700 AS Groningen, The Netherlands.

Petra Hendriks (p.hendriks@rug.nl)
Center for Language and Cognition, University of Groningen,
PO Box 716, 9700 AS Groningen, The Netherlands.

Gisela Redeker (g.redeker@rug.nl)
Center for Language and Cognition, University of Groningen,
PO Box 716, 9700 AS Groningen, The Netherlands.

Laurie A. Stowe (l.a.stowe@rug.nl)
Center for Language and Cognition, University of Groningen,
PO Box 716, 9700 AS Groningen, The Netherlands.

Abstract

During conversation, speakers and listeners act on certain basic assumptions, which enable them to communicate swiftly and seemingly effortlessly (Grice, 1975). The speaker, for instance, is supposed to say no more, but also no less than is necessary in a given conversational context (Maxim of Quantity). The present study looks at how language users react when this pragmatic assumption is violated. Participants were presented with written mini-dialogues while their ERPs (Event-Related brain Potentials) were measured. Dialogues in the violation condition, where the answer did not meet the quantity requirements, differed from control dialogues in three different time-windows, time-locked to the presentation of a critical word. Violating the Maxim of Quantity was signalled immediately and gave rise to effortful processing at different levels of representation.

Keywords: Psycholinguistics; Gricean Maxims, Implicature, Coordination, Pragmatics, Topic Structure, ERP.

Introduction

When taking part in a conversation, speaker and listener act upon specific assumptions about shared and private knowledge, and about the informativeness of the utterances that are exchanged. Grice (1975) formulated a framework in which these conversational assumptions are realized as four maxims:

1a. Quality: Be truthful
b. Quantity: Be as informative as required
c. Relation: Be relevant
d. Manner: Be clear

It is sometimes thought that the maxims are a kind of overly detailed puritan recipe for successful conversation. Indeed, Horn (2004) quotes a contemporary linguist exclaiming: "Would we want to have dinner with such a person, such an impeccably polite maxim observer?". A more fruitful approach, however, is to view these maxims as identifying a default set of assumptions - specifically the listeners' assumptions about the speaker - of which all participants in a communicative situation are aware (Horn, 2004). Grice's Maxim of Quantity, for example, describes how a listener expects the speaker to say no more, but also no less than necessary in a given conversational context. In the present experiment we will investigate what happens when the speaker does not comply to this conversational rule. Consider, for instance, the mini-dialogue in (2). There, the actions of two persons, John and Peter, are under discussion, and the answer provides all the information that is needed about these two protagonists, unlike dialogue (3).

2. Question: What did John and Peter do?
Answer: John kissed Annet and Peter kissed Hank.

3. Question: What did John and Peter do?
Answer: John kissed Annet and Peter on the cheek.

It is obvious that crucial information is missing, namely an answer to the partial question "What did Peter do?" By withholding this information, the speaker is violating the Maxim of Quantity.

There are different ways in which one can violate the Maxim of Quantity. For instance, someone can answer the question about how many children she has, with "two", when in fact she has three, or incorrectly say that the water is "not cold", while it is piping hot. These are called scalar implicatures, as they involve the computation of the intended meaning (i.e., what is implicated) from a *semantic hierarchy* or *scale* (e.g., cold - warm - hot). In another situation, a speaker wanting to *refer* to a specific object should refrain from giving too much or too little information describing it. For instance, Engelhardt, Bailey, & Ferreira (2006) present eye-tracking evidence suggesting that listeners are acutely sensitive to overdescription, even though they are not consciously aware of any processing problems.

The example that we are looking at in (3), however, takes place at a different level, and is closely related to the

pragmatic concept of 'topic-structure' (Hoeks, Vonk, & Schriefers, 2002). A *topic* can be loosely described as the entity about which the sentence imparts information (Lambrecht, 1994). The question in (3) introduces two entities in a way that makes them very likely topics of the answer, either as a unit ("They did X"), or in a construction with contrastive topics, in which each of the entities performs a separate action ("John did X, and Peter did Y"). Their results expectation of additional information due to the Maxim of Quantity clearly played a role in the resolution of the syntactic ambiguity seen in these sentences.

Until now there have only been very few investigations of how conversational assumptions impact on-line language processing. Most of these studies focus on scalar implicatures, which are instances of the class of generalized implicatures, that is, they can be computed without reference to the preceding context. In contrast, our study looks at the on-line processing of *particularized* implicatures, where the pragmatic interpretation of an utterance is crucially dependent on the preceding context.

Experiment

In this experiment, participants read short dialogues that appeared word-by-word in the middle of a computer screen. The sentences that are used in this experiment are all grammatically correct and semantically intact; they only differ in the extent to which the answer part of the dialogues is *pragmatically felicitous* with respect to the preceding question. During the reading of the mini-dialogues, brain activity of the participants was monitored by the continuous recording of ERPs (Event Related brain Potentials). Dialogues were structured such that at the final word of the answer sentence it became clear that the answer was pragmatically anomalous, as it violated the Maxim of Quantity (equals: give exactly as much information as required, no more and no less!).

Method

Participants The participants were 18 undergraduate students from the University of Groningen (6 male, 12 female, age-range 18-29, mean 20), who received payment or course credits for taking part in the experiment. All were right-handed native speakers of Dutch with normal, uncorrected vision.

Materials In this experiment we used sentences containing NP-coordinations that were based on materials taken from Hoeks (1999). For example, see sentence (4):

4. The mayor praised the councilor and the alderman exuberantly.

In the absence of a context, language users show a clear preference for structures where the conjunct *and* conjoins NPs, instead of sentences (as in e.g., "{the mayor praised the councilor} and {the alderman laughed}") (For English: Frazier, 1987; Frazier & Clifton, 1997; for Dutch: Hoeks, 1999; Hoeks et al., 2006). Using NP-coordinations in our experiment will thus avoid so-called 'garden-path' effects that occur when ambiguous utterances are ultimately resolved towards the non-preferred reading. These sentences were embedded in two kinds of dialogue: - in the neutral condition, the sentences were preceded by a 'neutral' question: "What happened?", which does not give rise to any specific expectation of the form or content of the answer (see, e.g., (5)); in the violation condition (see e.g., (6)) sentences were preceded by a question like "What did the mayor and the alderman do?", which requires a more specific answer pertaining to what both people actually did.

The adverb ("exuberantly") unambiguously indicates (at least in Dutch, the language used in this experiment) that the answer is a sentence with only one topic (i.e., "the mayor"), and not two, as would be expected from the question in the violation condition. Thus, the NP "the alderman" turns out not to be the expected topic, which constitutes a clear violation of the Maxim of Quantity.

5. *Neutral*:
Q: What happened?
A: The mayor praised the councilor and the alderman exuberantly.

6. *Violation*:
Q: What did the mayor and the alderman do?
A: The mayor praised the councilor and the alderman exuberantly.

Besides these two kinds of experimental dialogues - 40 in total, 20 per condition - where the answer sentence contained an NP-coordination, there were also 40 filler dialogues (half with a neutral and half with a two-topic question) in which the answer consisted of an S-coordinated sentence, so as to minimize the chance of participants developing processing strategies. In addition, there were 100 filler items from an unrelated experiment on relative clause processing; these will not be discussed further.

Design Experimental lists were created using a Latin Square, with equal numbers of items occurring in each condition on each list, and no list containing more than one version of a given item. The order in which experimental and filler items appeared was determined semi-randomly (i.e., allowing maximally three experimental items in consecutive order, but never two consecutive items in the same condition) and was the same for all lists. Each list was presented to an equal number of participants and each participant only saw one list.

Procedure Participants were tested in a dimly lit, sound-proof booth. They sat facing a computer screen at approximately 60 cm distance; a chin-rest was used to minimize movement artifacts. Participants were instructed to read each sentence for comprehension, and to respond to the occasional content question (35 in total, quasi-randomly distributed over the experiment) in order to answer "yes" or "no" by lifting the right or left index finger, respectively. Content questions were always followed by filler items, so that possible problems in answering the questions would not influence the processing of experimental items.

At the beginning of each trial, a fixation mark (an asterisk) appeared for 1 second. After that, the dialogue sentences were presented word-by-word in the centre of the screen. Each word remained on screen for 243 mSec (durations have to be a multiple of the screen refresh time), and was followed by a blank screen with a duration of 243 mSec. Between the question-part of the dialogue and the answer there was an interval of 729 mSec. At the end of an experimental item, the word "Knipper" (= "Blink") was shown for 3 seconds, giving participants the opportunity to blink; they were instructed to try and avoid blinking during the presentation of the sentence to avoid eye-movement and blink-related artifacts. After every 50 trials, the participant could take a short break. The experiment took about 105 min, including preparation.

EEG recording parameters The EEG activity was recorded by means of 20 tin electrodes mounted in an elastic cap (see Figure 1): FP1, FP2, FZA, F7, F3, FZ, F4, F8, T7, C3, CZ, C4, T8, P7, P3, PZ, P4, P8, O1, and O2. Bipolar horizontal EOG was recorded between electrodes at the outer left and right canthus. Bipolar vertical EOG was recorded for both eyes. Electrode impedances were kept below 5 kΩ. EEG and EOG signals were sampled at 1000 Hz, amplified (EEG: 0.2 mV/V; EOG: 0.5 mV/V; time constant: 10 sec.), and digitally low-pass filtered with a cut-off frequency of 30 Hz; effective sample frequency was 100 Hz.

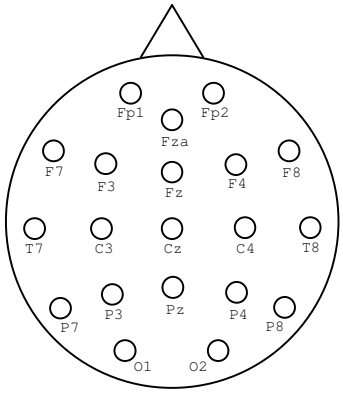


Figure 1: Electrode placement
(triangle indicates nose of participant)

Results

Data Analysis Participants read attentively, answering on average 85% (SD = 5.6) of the content questions correctly. Visual inspection of the ERP waveforms following the presentation of the critical adverb ('exuberantly') suggested three effects of the violation condition as compared to the neutral condition (see Figure 2): an early bipolar component in the ELAN time-window (180-320 mSec post-onset), which was followed by a positivity in the N400 time-window (350-550 mSec post-onset), and a late positivity in the early P600 time-window (550-750 mSec post-onset).

For each of those intervals, average ERPs were computed for each electrode site, each participant, and each condition separately. Prior to averaging, trials with ocular or amplifier-related artifacts were excluded from analysis.

The ambiguous NP in the answer sentence (e.g., "the alderman") was mentioned in the question of the Violation condition ("What did the mayor and the alderman do?") but not in the question of the Neutral condition ("What happened?"), which might have given rise to a reduction of the N400 due to repetition priming (Kutas et al., 2007) or other effects. To rule out the possibility that effects on the preceding word influenced the pattern of results at the critical adverb, a 100 mSec post-stimulus onset baseline was used instead of a pre-stimulus baseline, time-locked to the onset of the critical word (for a similar procedure, see Philips, Kazanina, & Abada, 2005; Mueller, 2008).

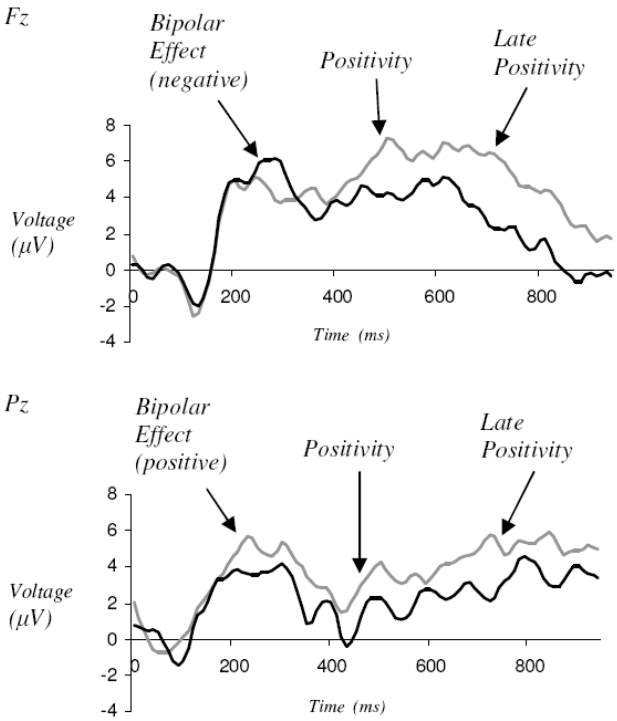


Figure 2. ERP-waveforms starting at the presentation of the *disambiguating adverb* for dialogues containing violations (**grey**) and neutral dialogues (**black**) on a frontal (Fz) and a posterior electrode (Pz).

For analysis purposes, three sets of electrodes were used: the prefrontal electrodes FP1, FZA, and FP2; the occipital electrodes O1 and O2; and the main set of electrodes. For each of these sets (and for each of the three relevant time-windows) average ERPs were statistically analyzed using Repeated-Measures ANOVA (Analysis of Variance) with Violation (violation vs. neutral) as a within-participant factor. In each of these ANOVAs topographical factors were also included: 1) for the prefrontal set this was the factor Laterality with 3 levels (i.e., left, midline, and right side of the scalp); 2) for the occipital set Laterality only had 2 levels (i.e., left and right); 3) for the main set of 15 electrodes, Laterality had 5 levels (far left, left, middle, right, far right),

and a second factor, Anteriority, had 3 levels (anterior, central, and posterior). Where appropriate, the Huynh-Feldt correction was applied to correct for violations of the statistical assumption of sphericity. We will report the corrected p-values with the original degrees of freedom. Because only effects involving the factor Violation tell us something about our pragmatic manipulation, other effects will not be reported.

Early Bipolar Effect (180-320 mSec post-onset: *ELAN time-window*)
In the analysis of the main set of electrodes, the interaction of Violation x Anteriority was significant ($F(2,30) = 5.34$; $p < .05$), but qualified by a significant three-way interaction of Violation x Anteriority x Laterality ($F(8,120) = 2.22$; $p < .05$).

Follow-up analyses showed significant and near to significant interactions between Violation x Anteriority for every level of Laterality, except for the electrodes on the far right (far left: $F(2,30) = 3.20$; $p = .07$; left: $F(2,30) = 7.36$; $p < .01$; middle: $F(2,30) = 8.72$; $p < .01$; right: $F(2,30) = 3.18$; $p = .07$; far right: $F < 1$). Each of these interactions was characterized by a frontal negativity (violation more negative than neutral), coupled with a posterior positivity (violation more positive than neutral), with central electrodes falling in between. Table 1 displays the size of the violation-effect as a function of Anteriority and Laterality. Analysis regarding the occipital electrodes produced a significant main effect of Violation ($F(1,15) = 5.35$; $p < .01$), where the violation condition was more positive than the neutral condition (a difference of 0.8 μV); there were no significant effects in the analysis of the prefrontal electrodes.

Table 1: Effect Sizes (*violation* minus *neutral*, in μV) for frontal, central, and posterior electrodes on every level of Laterality in the first time-window (180-320 mSec post-onset)

	Far Left	Left	Middle	Right	Far Right
Frontal	-0.8	-0.7	-0.9	-0.5	-0.4
Central	0.0	0.4	0.5	0.2	0.2
Posterior	0.9	1.1	1.2	0.6	0.1

Positivity (350-550 mSec post-onset: *N400 Time-Window*)
For the main set of electrodes we found a significant effect of Violation ($F(1,15) = 5.95$; $p < .05$), with a larger positivity for the violation condition as compared to the neutral condition (a difference of 1.3 μV). There was no interaction with topographical factors Anteriority and Laterality (all F-values < 1). In the analysis on the prefrontal electrodes there was also only a main effect of Violation (a difference of 1.8 μV; $F(1,15) = 7.61$; $p < .05$). There were no significant effects in the analysis of the occipital electrodes (all p-values > .19).

Late Positivity (550-750 mSec post-onset: *P600 Time-Window*)
The analysis on the main set of electrodes produced a significant main effect of Violation ($F(1,15) = 7.99$; $p < .05$), with a larger positivity for the violation condition versus the neutral condition (a difference of 1.9 μV). There was no interaction with Anteriority ($F < 1$); the interaction with Laterality was marginally significant (($F(4,60) = 2.22$; $p = .10$).
These effects were qualified by a significant three-way interaction of Violation x Anteriority x Laterality ($F(8,120) = 7.61$; $p < .05$). This interaction ensued from the effect of Violation (violation more positive than neutral) being quite pronounced on the left side of the scalp, and significantly less strong on the right (and even absent on far right electrodes). See Table 2 for the effect sizes on all electrodes contained in the main set. Analysis of the prefrontal electrodes showed a main effect of Violation where the violation condition was much more positive than the neutral (a difference of 2.8 μV; $F(1,15) = 11.65$; $p < .005$). At the occipital electrodes, the violation condition gave rise to a positivity on the left (O1: 0.5 μV), but to a slight negativity on the right (O2: -0.2 μV); this interaction was marginally significant ($F(1,15) = 3.64$; $p = .08$).

Table 2: Effect Sizes (*violation* minus *neutral*, in μV) for frontal, central, and posterior electrodes on every level of Laterality in the P600 time-window (550-750 mSec post-onset)

	Far Left	Left	Middle	Right	Far Right
Frontal	2.1	3.2	2.6	2.3	1.1
Central	2.3	2.1	1.2	1.4	1.3
Posterior	2.3	2.7	1.9	1.3	0.5

Discussion

Violating the Maxim of Quantity in these mini-dialogues had a very clear effect on ERPs, in three different time-windows.

The early frontal negativity seems to be related to the Early Left Anterior Negativity (ELAN) that has been found in response to word category violations (Friederici, 1995). The strong topic-structure expectation created by the question presumably translates into a strong syntactic expectation for an inflected verb to occur after the name of the second protagonist. If participants read an adverb instead of a verb, this may be detected very quickly. The positivity accompanying the anterior negativity may reflect the detection of the additional pragmatic violation.

After this early effect we found a broadly distributed positivity in the interval between 350 and 550 mSec after presentation of the critical word. This effect is highly reminiscent of a positivity reported by Bornkessel, Schlesewsky, and Friederici (2002). According to Bornkessel et al., this positivity reflected a form of thematic reanalysis that occurs when the thematic role that is initially assigned to a discourse entity turns out to be wrong. In the present experiment the ambiguous NP (e.g., 'the alderman') is expected to be an AGENT (the entity that performs an action) on the basis of the question, but turns out to be a PATIENT (the entity that undergoes an action), requiring thematic reanalysis.

Finally, there was a large positive effect for the Violation condition in the P600 time-window. A P600 is generally found as a response to syntactic violations (Hagoort, Brown, & Groothusen, 1993), syntactic dependencies (Kaan et al., 2000), but also to some kinds of semantic violations (e.g., Hoeks, Stowe, & Doedens, 2004). It is generally thought to reflect the effortful processing involved in syntactic integration, or syntactic reanalysis following an error somewhere in the utterance. This effortful processing is most likely motivated by the wish to create a coherent representation of the language input. The scalp distribution of the late positivity in the present experiment, however, is not centro-parietal, as in the typical case, but is shifted to the left, and especially large at frontal electrodes. Following Friederici et al. (2002) and Hagoort et al. (1999) we might assume that the more anteriorly oriented P600 reflects the difficulty of the revision process, whereas a posterior P600 effect might result from a general failure to compute. On a more speculative note, the late positivity that we find here may in part also reflect the computation of whether the speaker wants to impart something by not giving an adequate answer to a question. In Grice's terms there is an implicature: Answering a question about X and Y solely by relating what person X did, without reference to person Y, may be an indirect way of asserting for instance that what person Y did was in fact very insignificant. Ongoing research in our lab is in fact aimed at investigating under what circumstances people will compute implicatures of this kind.

Conclusion

If Grice is right, then all language users work from the default assumptions that their conversational partners are rational beings, who produce utterances that are true, clear, and relevant, and do not contain more, but certainly not less information than is required in the specific conversational setting in which they occur. And indeed, whenever a given utterance for instance violates the Maxim of Quantity, this will be detected within 200 mSec, leading to thematic and syntactic reanalysis - and possibly also the computation of an implicature - all of which are motivated by the desire to create a coherent representation of what the other person is saying.

Acknowledgements

We would like to thank Charlotte Wunderink and Ingeborg Prinsen for their help in running the experiment.

References

Bornkessel, I., Schlesewsky, M., & Friederici, A.D. (2002). Beyond syntax: language-related positivities reflect the revision of hierarchies. *NeuroReport, 13(3),* 361-364.

Engelhardt, P.E., Bailey, K.G.D, & Ferreira, F. (2006). Do speakers and listeners observe the maxim of quantity? *Journal of Memory and Language, 54,* 554–573.

Frazier, L. (1987). Syntactic processing: evidence from Dutch. *Natural Language and Linguistic Theory, 5,* 519-559.

Frazier, L., & Clifton, C.(1997). Construal: Overview, motivation, and some new evidence. *Journal of Psycholinguistic Research, 26,* 277-295.

Friederici, A.D. (1995). The time course of syntactic activation during language processing: A model based on neuro-psychological and neurophysiological data. *Brain and Language, 50,* 259–281.

Friederici, A.D., 2002. Towards a neural basis of auditory sentence processing. *Trends in Cognitive Sciences, 6,* 78–84.

Grice, H.P. (1975). Logic and Conversation. In P. Cole & J.L. Morgan (Eds.), *Syntax and Semantics,* Volume 3: Speech Acts (pp. 41–58). New York: Academic Press.

Hagoort, P., Brown, C.M., & Groothusen, J. (1993). The syntactic positive shift as an ERP measure of syntactic processing. *Language and Cognitive Processes, 8,* 439-484.

Hagoort, P., Brown, C.M., Osterhout, L. (1999). The neural architecture of syntactic processing. In: Brown, C.M., Hagoort, P. (Eds.), *The neurocognition of language.* Oxford University Press, Oxford.

Hoeks, J.C.J. (1999). *The processing of coordination: semantic and pragmatic constraints on ambiguity resolution.* Doctoral Dissertation, University of Nijmegen. Available: http://www.let.rug.nl/~hoeks/Hoeks1999_dis.pdf

Hoeks, J.C.J., Hendriks, P., Vonk, W., Brown, C.M., & Hagoort, P. (2006). Processing the NP- versus S-coordination ambiguity: thematic information does not completely eliminate processing difficulty. *The Quarterly Journal of Experimental Psychology, 59(9),* 1581-1599.

Hoeks, J.C.J., Stowe, L.A., & Doedens, L.G. (2004). Seeing words in context: the interaction of lexical and sentence level information during reading. *Cognitive Brain Research, 19,* 59-73.

Hoeks, J.C.J., Vonk, W., & Schriefers, H. (2002). Processing coordinated structures in context: the effect of topic-structure on ambiguity resolution. *Journal of Memory and Language, 46,* 99-119.

Horn, L.R. (2004). Implicature. In: L.R. Horn & G. Ward (Eds.), *The handbook of pragmatics* (pp. 3-28). Malden, MA: Blackwell.

Kaan, E., Harris, A., Gibson, E., & Holcomb, P.(2000). The P600 as an index of syntactic integration difficulty. *Language and Cognitive Processes, 15,* 159-201.

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: brain potentials reflect semantic incongruity. *Science, 207,* 203-205.

Kutas, M., Van Petten, C., & Kluender, R. (2007). Psycholinguistics electrified II (1994-2005). In: M. A. Gernsbacher & M. Traxler (Eds.), *Handbook of psycholinguistics* (2nd Edition). New York, NY: Elsevier.

Lambrecht, K. (1994). *Information structure and sentence form: Topic, focus, and the mental representation of discourse referents.* Cambridge, MA: University Press.

Mueller, J.L., (2009). The influence of lexical familiarity on ERP responses during sentence comprehension in language learners. *Second Language Research, 25,* 43-76.

Phillips, C., Kazanina, N., & Abada, S.H. (2005). ERP effects of the processing of syntactic long-distance dependencies. *Cognitive Brain Research, 22,* 407– 428.

Can grammar influence voting?

Caitlin M. Fausey (cmfausey@stanford.edu)
Department of Psychology, 450 Serra Mall, Building 420, Stanford University
Stanford, CA 94305 USA

Teenie Matlock (tmatlock@ucmerced.edu)
Cognitive and Information Sciences Program, University of California, Merced
Merced, CA 95344 USA

Abstract

The wording of political messages is known to affect voting behavior, including judgments about the electability of candidates. Yet the question remains whether voting behavior also depends on fine-grained grammatical details of political messages. Results from two studies suggest that the grammatical forms used in describing political candidates' past actions can affect attitudes about electability under certain conditions. The findings provide novel insights on how language can shape thought in the political realm.

Introduction

Millions of dollars are spent on campaign ads each year. Yet little is known about how linguistic details in these messages influence people's attitudes about political candidates and ultimately whether they are elected. Here we offer new results to show that altering grammatical information can lead to different opinions about electability.

We know that the linguistic content of political messages can influence attitudes about candidates running for office (e.g., Lau & Redlawsk, 2006). People base their voting decisions on criteria emphasized by news coverage (e.g., Druckman, 2004), and their votes can be biased by the editorial slant of the newspaper they read (e.g., Druckman & Parkin, 2005). People reject incumbent candidates if times are portrayed as bad (e.g., Quattrone & Tversky, 1988). They turn away from candidates or vote for no one at all if presented with an excess of negative language (e.g., Ansolabehere & Iyengar, 1995; Garramone, 1984). Their candidate preferences are more entrenched when opposition is emphasized (e.g., Bizer & Petty, 2005). They reject candidates who contradict their metaphorical conceptions of politics and government (e.g., Lakoff, 1996). What we do *not* know is how fine-grained linguistic details in political messages influence voters. Can grammatical information affect attitudes about candidates and whether they are electable, and if so, how?

In English and many other languages, information about the temporal organization of events is provided by aspectual markers that accompany verbs. For past events, imperfective aspectual markers (*was* VERB+*ing*) emphasize the ongoing nature of actions, and perfective aspectual markers (VERB+*ed*) emphasize the completion of actions (e.g., Comrie, 1976; Frawley, 1992; Madden & Zwaan, 2003; Magliano & Schleich, 2000). These grammatical markers can influence how people think about past events. In interpreting imperfective descriptions of past events, people take an internal perspective (e.g., Ferretti & Katz, 2010). In interpreting descriptions of motion events, for example, people tend to situate a moving character in the middle range of a trajectory toward a destination with imperfective information (Morrow, 1985; 1990). Also, details such as the individuals, objects and locations of the events are more accessible after imperfective event descriptions (e.g., Carreiras, Carriedo, Alonso, & Fernández, 1997; Ferretti, Kutas, & McRae, 2007; Madden & Therriault, 2009; Truitt & Zwaan, 1997).

In addition, when processing event descriptions people infer that more action occurs with imperfective descriptions than with perfective descriptions. For instance, people estimate that more houses were painted after reading "*John was painting houses last summer"* than after reading "*John painted houses last summer"* (Matlock, in press). People also remember past actions more easily, and are more likely to continue them in future behavior, after imperfective descriptions than perfective descriptions (e.g., Hart & Albarracín, 2009; Magliano & Schleich, 2000).

In the current work, we investigated the role of grammatical information in the interpretation of political messages, precisely, whether and how imperfective "*was* VERB + *ing*" and perfective "VERB + *ed*" would influence attitudes about the electability of political candidates. Would the imperfective form, which draws attention to details and the ongoing process of actions, lead to different attitudes about electability than would the perfective form? And might this effect be more pronounced for political messages that are negative versus positive, especially because negative information arouses emotions (e.g., Westen, 2007), captures attention (e.g., Baumeister, Bratslavsky, Finkenhauer, & Vohs, 2001; Rozin & Royzman, 2001), and affects perceptions of political candidates (e.g., Basil, Schooler, & Reeves, 1991; Lau, 1982; see Lau, Sigelman, & Brown Rovner, 2007, for a broad perspective)? Finally, can grammatical information ever influence inferences about actions themselves? Would, for instance, a phrase such as *was taking hush money* lead people to believe that more dollars were taken than a phrase such as *took hush money*? These questions are important because voters rely on information about the past to infer what politicians will do in future elected positions (e.g., Fiorina, 1981).

Two studies were designed to investigate these issues. In each, participants read about the past actions of a senator who was seeking re-election. Then they decided whether he would be re-elected. Next they provided a confidence rating for the decision. Last, participants provided a numeric estimate about the actions (e.g., amount of hush money in Study 1).

Study 1

Participants read a short passage about a fictitious politician who did (perfective) or was doing (imperfective) past actions that were either negative or positive. Based on previous research showing that an increment toward a negative pole may carry more weight in decision-making than "the same" increment toward a positive pole (e.g., Kahneman & Tversky, 1979), we hypothesized that grammatical form may more strongly influence people's judgments about negative past actions than about positive past actions. Further, people may pay closer attention to negative events than to positive events (e.g., Baumeister et al., 2001; Rozin & Royzman, 2001), perhaps heightening the effect of any particular linguistic construal of the past event. Thus, our main prediction was that the politician would be evaluated more negatively when negative past actions were described with imperfective rather than perfective grammatical markers.

Method

Participants. A total of 369 undergraduate students at the University of California, Merced, received partial course credit in an introductory cognitive science course or an introductory psychology course. Fifteen of the individuals provided illegible responses or did not finish the task, leaving 354 participants.

Materials, Design and Procedure. Participants completed a questionnaire that appeared on a single page in a booklet that contained a set of unrelated tasks. Participants had five days to complete and return the booklet, and were told not to discuss the task with others.

Participants first read a short description of a fictitious senator who was up for re-election. The senator did or was doing negative or positive actions (see Table 1 for the four description versions). For example, he *was taking hush money* or *took hush money*, and for positive actions, he *was collecting donations* or *collected donations.*

Then these participants answered two questions, "*Will this candidate be re-elected?*" (circled Yes or No) and "*How confident are you about your decision regarding re-election*" (used a seven point scale, ranging from "*Not at all confident*" (1) to "*Very confident*" (7)).

Next participants answered a question about the financial dealings of the senator, either "*Please estimate the total amount of hush money (in dollars)*" (in the negative valence condition) or "*Please estimate the total amount of donation money (in dollars)*" (in the positive valence condition). The senator was fictitious to avoid bias about actual political candidates.

Table 1: Stimuli for Study 1

	Grammatical form	
Action valence	**Perfective (verb+*ed*)**	**Imperfective (*was* verb+*ing*)**
Negative	Mark Johnson is a Senator in the United States Senate. He is up for re-election. He graduated from the University of Texas, Austin with a degree in political science. Mark's first term as a United States Senator is almost complete. Last year, Mark had an affair with his assistant and took hush money from a prominent constituent. (*N*=92)	Mark Johnson is a Senator in the United States Senate. He is up for re-election. He graduated from the University of Texas, Austin with a degree in political science. Mark's first term as a United States Senator is almost complete. Last year, Mark was having an affair with his assistant and was taking hush money from a prominent constituent. (*N*=96)
Positive	Mark Johnson is a Senator in the United States Senate. He is up for re-election. He graduated from the University of Texas, Austin with a degree in political science. Mark's first term as a United States Senator is almost complete. Last year, Mark rekindled his relationship with his wife and collected donation money for the American Cancer Society. (*N*=85)	Mark Johnson is a Senator in the United States Senate. He is up for re-election. He graduated from the University of Texas, Austin with a degree in political science. Mark's first term as a United States Senator is almost complete. Last year, Mark was rekindling his relationship with his wife and was collecting donation money for the American Cancer Society. (*N*=81)

Results

First, we examined the valence of past actions and electability. Not surprisingly, participants viewed the senator as more electable when his past actions were positive (80%) versus negative (22%), $\chi 2(1, N=354) = 119.94$, $p < .001$. Twenty-one percent of the participants did not conform to this pattern, and indicated that the candidate would be re-elected if he had done negative actions ($N = 41$), or not be re-elected if he had done positive actions ($N = 33$).

Second, we analyzed people's confidence about their electability decision. Electability decisions were weighted by confidence, resulting in a scale ranging from -7 (Strongly Confident "No" vote) to +7 (Strongly Confident "Yes" vote). Histograms of this weighted decision are shown in Figure 1a (negative actions) and Figure 1b (positive actions).

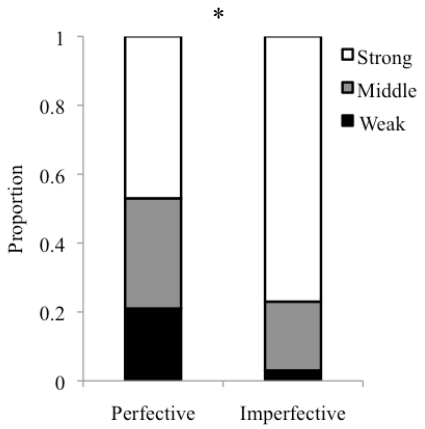


Figure 2: Grammatical aspect changes how people view a politician's negative actions. Voter confidence in deciding not to re-elect a politician. Proportion of sample is plotted on the y-axis.

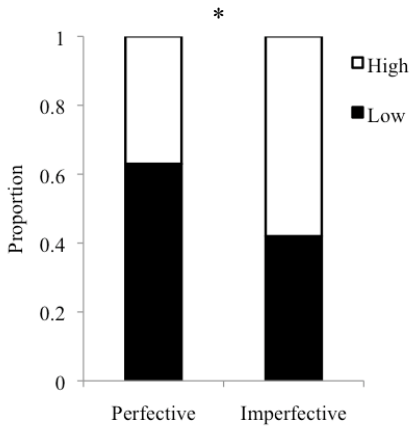


Figure 3: Grammatical aspect changes how people view a politician's negative actions. Median split judgments of hush money taken. Proportion of sample is plotted on the y-axis.

The subgroups that are evident in these data correspond to participants whose decision did, and did not, align with the action valence. This distinction may be analogous to the common distinction of correct versus incorrect responses in reaction time analyses. Only those responses that are clearly interpretable are submitted to further analyses. In this study, subsequent analyses were therefore restricted to those participants whose decision aligned with the action valence.

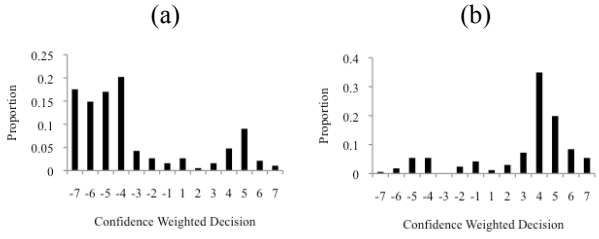


Figure 1: Confidence weighted electability decisions: (a) Negative Events, (b) Positive Events.

Data bearing on whether grammatical aspect influences electability were the confidence weighted scores for decisions that were consistent with the action valence. Because some of these data were skewed, and also showed some heteroskedasticity across conditions, we took a conservative approach and did a non-parametric analysis. Conclusions remain the same using parametric analyses.

Confidence ratings were divided into "Weak confidence" (rating of 3 or less extreme), "Middle confidence" (rating of 4), and "Strong confidence" (rating of 5 or more extreme) groups. As shown in Figure 2, participants' confidence about electability varied depending on the grammatical markers used to describe the senator's past actions. Participants were more strongly confident about their "no" decisions when the senator *was doing* negative actions (77%) than when he *did* negative actions (47%), $\chi^2(2, N = 147) = 18.27$, $p < .001$. They were about equally confident for their "yes" decisions when the senator *was doing* (45%) and when he *did* (39%) positive actions, $\chi^2(2, N = 133) = .65$, *n.s*.

Third, we analyzed estimates for money taken (hush money) or collected (donations) by the senator. Unsurprisingly, these distributions were highly skewed. We again took a conservative analysis approach, and conclusions remain the same using parametric analyses.

We divided responses into "Low" and "High" money groups based on the median estimate value of the respective decisions. The median estimate for hush money ($100,000) structured the two groups for negative financial actions, and the median estimate for donations ($50,000) structured the two groups for positive financial actions.

Grammatical form influenced the inferences that people made about money. Dollar estimates were higher when the senator *was taking* hush money (58% were above overall median) versus *took* hush money (37% were above overall median), $\chi^2(1, N = 147) = 6.74$, $p = .009$ (Figure 3). For positive actions, there was no difference (47% versus 53%, $\chi^2(1, N = 133) = .36$, *n.s.*).

Finally, using independent participants in a separate manipulation check, we confirmed that our "negative" and "positive" stories differed in valence. Forty-six participants who were among the English speakers who use Amazon's Mechanical Turk Service (mturk.amazon.com) read one story selected randomly from the four versions used in the main study (Negative perfective, Negative imperfective, Positive perfective, Positive imperfective). After reading the story, participants answered the question "*Please use the scale below to indicate what you think of the senator's actions*" using a 15-point scale ranging from "*Very Negative*" (1) to "*Very Positive*" (15).

As expected, participants judged the negative stories (M = 3.48, SE = .64) to be more negative than the positive stories (M = 11.91, SE = .52), $t(44) = 10.21$, $p < .001$. Further, grammatical aspect itself (perfective versus imperfective) did not influence participants' judgments of negativity, overall or within each kind of story (all p's > .18).

In sum, people were more confident in voting not to re-elect a senator who *was doing* negative actions than a senator who *did* negative actions. They also inferred that more negative action was involved when the past event was described using imperfective aspect compared to perfective aspect.

Study 2

In everyday life, politicians do good and bad things. Here we were interested in cases involving both a positive and negative outcome. In this study, the senator was responsible for an eminent domain policy with a negative *and* a positive outcome. All participants read about both outcomes, but some read about an imperfective negative outcome and a perfective positive outcome (*was removing homes and extended roads*) and others, about a perfective negative outcome and an imperfective positive outcome (*removed homes and was extending roads)* (see Table 2). We hypothesized that the overall eminent domain policy would be interpreted more negatively when the negative action was in the imperfective than when the negative action was in the perfective.

Method

Participants. A total of 127 members of the Stanford University community were paid to participate. Most were students. Data from participants whose age was greater than 3 SDs above the mean age (N = 5) and from individuals who returned incomplete surveys (N = 2) were excluded, leaving 120 participants.

Materials and Procedure. Participants read a passage about a fictitious senator who was seeking re-election and who had implemented an eminent domain policy with a negative and a positive outcome (home removal and road extension, respectively), and then answered the same questions as in Study 1. The task appeared on a single page in a booklet of unrelated materials. Participants had a week to complete the task.

Table 2: Stimuli for Study 2

Negative Imperfective	Mark Johnson is a Senator in the United States Senate. He is up for re-election. Last year, his district faced rush hour traffic problems. Under eminent domain Mark was removing homes and extended roads in his district. Traffic conditions improved. (N = 58)
Negative Perfective	Mark Johnson is a Senator in the United States Senate. He is up for re-election. Last year, his district faced rush hour traffic problems. Under eminent domain Mark removed homes and was extending roads in his district. Traffic conditions improved. (N = 62)

Results

As shown in Figure 4, participants who read about "*removing homes*" were more likely to respond that the candidate would *not* be re-elected (60%) than participants who read about "*removed homes*" (44%). The pattern was reliable, p = .049 (Fisher's exact test, one-tailed, was used given our directed prediction). Participants were about equally confident in their decisions in the two conditions. There were no reliable differences in estimates about the number of roads extended or homes removed. Thus, again, grammatical information influenced attitudes about electability. In this case, despite having read about both components of an eminent domain policy, participants were biased by the use of the imperfective: They judged a politician to be less electable when the negative outcome of his policy was highlighted using imperfective aspect compared to when it was described using perfective aspect.

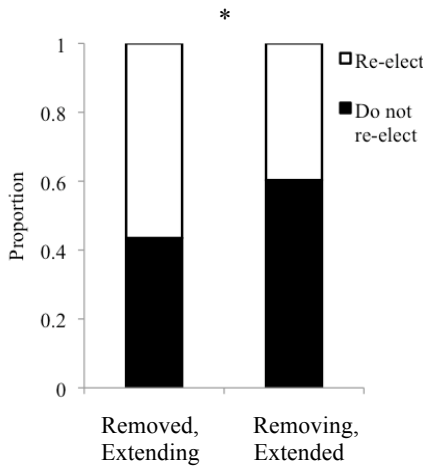


Figure 4: Grammatical aspect influences electability. Proportion of sample is plotted on the y-axis.

As in Study 1, we again queried independent participants ($N = 22$) about the valence of the senator's actions, using the same procedure and valence scale as Study 1. In this case, participants judged each version of the story to be about equally negative ("*removed*" $M = 6.18$, $SE = .50$; "*was removing*" $M = 8.00$, $SE = 1.21$, $t(13.32$(assuming unequal variances)) $= 1.39$, *n.s.*). It appears that the effect of grammatical aspect on electability may be somewhat insidious when reasoning is based on scenarios with mixed outcomes. When making explicit valence judgments, people see both the good and the bad, but grammatical aspect may implicitly color judgments about the political candidate himself.

General Discussion

Our studies suggest that grammar can influence electability. In Study 1, a change in the grammatical form of negative action descriptions resulted in a change in reasoning about a political candidate. People were more confident in their "no" vote, and provided higher dollar estimates for hush money when negative actions were described using imperfective than perfective. They were not sensitive to grammar when reasoning about a candidate's past positive actions. In Study 2, grammar again influenced electability, such that people reasoned about electability in line with whatever action was highlighted by imperfective aspect. Over 50% of people judged a candidate who "removed homes and *was* extend*ing* roads" to be electable while under 50% did so when the verb markers *–ed* and *–ing* were reversed.

Why did the imperfective form result in higher confidence ratings and larger money estimates than did the perfective form, for negative actions in particular? Several explanations are worth considering. First, people may pay more attention to negative events than to positive events (e.g., Baumeister et al., 2001; Rozin & Royzman, 2001), making any mental representation driven by a linguistic construal relatively more robust for negative events. Further, the contrast between two negative alternatives is often perceived to be larger than the contrast between two "equally spaced" positive alternatives (e.g., Kahneman & Tversky, 1979), and so any contrast due to grammatical form may have been amplified for negative events.

The effects of negative information and imperfective information on decision-making may be additive. The combination of negative information and imperfective information could have made for strong attitudes, including pronounced confidence about "no" votes. This is plausible given that negative information arouses emotions and captures attention (e.g., Baumeister et al., 2001; Rozin & Royzman, 2001; Westen, 2007) and the imperfective form widens scope (Frawley, 1992) and draws attention to details of actions (e.g., Carreiras et al., 1997; Ferretti, Kutas, & McRae, 2007; Madden & Therriault, 2009; Truitt & Zwaan, 1997). With heightened attention to negative details, it may be especially easy for voters to confidently reject a candidate.

Another possible explanation may simply be that people generally prefer to avoid losses when there are unknown outcomes (Kahneman & Tversky, 1979). More negative actions could be construed as risky, and lead to stronger confidence that a "no" vote was the right choice. In the same vein, the imperfective form may have prompted a sense of "ongoingness" of the politician's negative actions while the perfective form may have provided closure on negative actions. If a political candidate *did* negative events in the past, those actions could have been perceived as over and done with, and less likely to influence the future. With positive information, there are no risks or adverse consequences and thus no reason to have a strong opinion about a "yes" vote.

These mechanisms – heightened attention to negative details and risk aversion – may also operate when voters reason about mixed outcome scenarios as in Study 2. Here, the combination of imperfective and negative information ("*removing*") appeared to shift attention away from beneficial policy outcomes and lead to more decisions that the candidate would not be elected.

Further research on the fine-grained linguistic details of political messages must be conducted for a full understanding of how language influences everyday thought in the political realm. Our novel results are an initial attempt to detail these important effects of language, and suggest that under certain conditions grammatical information affects whether a political candidate is electable. Future research should examine a wider range of actions, including future actions and policy proposals, as well as other fine-grained grammatical features of political messages. Investigations of grammar using linguistic data from real political campaigns will also be informative.

Voters appear to be sensitive to fine-grained linguistic details when judging political candidates. When the past actions of a candidate were negative, descriptions using imperfective aspect damaged the candidate's electability more than descriptions using perfective aspect. Because 'scandals' involving political candidates are a hot topic in media coverage and campaign ads, insight into the power of the grammar used to communicate negative information will likely improve our understanding about how linguistic media shapes voting patterns. The current findings are consistent with previous psycholinguistic results and extend our understanding of the role of grammar in political decision-making.

Acknowledgments

Each author contributed equally to this research. We thank Michael Leon and Nilofaur Tahery for assistance with data collection, and the following individuals for sharing insights: Lera Boroditsky, Herbert Clark, Tom Hansford, Christopher Kello, Paul Maglio, Arthur Markman, Justin Matthews, Steve Nicholson, Daniel Oppenheimer, and Sally Rice.

References

Ansolabehere, A., & Iyengar, S. (1995). *Going negative: How political advertisements shrink and polarize the electorate*. New York, NY: Free Press.

Basil, M., Schooler, C., & Reeves, B.R. (1991). Positive and negative political advertising: Effectiveness of ads and perceptions of candidates. In F. Biocca (Ed.), *Television and political advertising, Volume 1: Psychological processes* (pp. 245-262). Hillsdale, NJ: Lawrence Erlbaum.

Baumeister, R.F., Bratslavsky, E., Finkenhauer, C., & Vohs, K.D. (2001). Bad is stronger than good. *Review of General Psychology, 5*, 323-370.

Bizer, G.Y., & Petty, R.E. (2005). How we conceptualize our attitudes matters: the effects of valence framing on the resistance of political attitudes. *Political Psychology, 26*, 553-568.

Carreiras, M., Carriedo, N., Alonso, M. A., & Fernández, A. (1997). The role of verb tense and verb aspect in the foregrounding of information during reading. *Memory & Cognition, 25*, 438-446.

Comrie, B. (1976). *Aspect.* Cambridge: Cambridge University Press.

Druckman, J.N. (2004). Priming the vote: Campaign effects in a U.S. Senate election. *Political Psychology, 25*, 577-594.

Druckman, J.N., & Parkin, M. (2005). The impact of media bias: How editorial slant affects voters. *The Journal of Politics, 67*, 1030 – 1049.

Ferretti, T.R., & Katz, A.N. (2010). Verb aspect and the retrieval of events from autobiographical memory. In A. Columbus (Ed.), *Advances in Psychology Research,* Nova Science.

Ferretti, T.R., Kutas, M., & McRae, K. (2007). Verb aspect and the activation of event knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*, 182-196.

Fiorina, M.P. (1981). *Retrospective Voting in American Elections.* New Haven, CT: Yale University Press.

Frawley, W. (1992). *Linguistic semantics.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Garramone, G. (1984). Voter response to negative political ads. *Journalism Quarterly*, *61*, 250-259.

Hart, W., & Albarracín, D. (2009). What I was doing vs. what I did: Verb aspect influences memory and future actions. *Psychological Science, 20(2)*, 238-244.

Kahneman, D., & Tversky, A. (1979). Prospect theory: an analysis of decision under risk. *Econometrica, 47*, 263-291.

Lakoff, G. (1996). *Moral politics: How liberals and conservatives think.* Chicago, IL: The University of Chicago Press.

Lau, R.R. (1982). Negativity in political perception. *Political Behavior, 4*, 353-377.

Lau, R.R., & Redlawsk, D.P. (2006). *How voters decide: Information processing during elections campaigns*. New York: Cambridge University Press.

Lau, R.R., Sigelman, L., & Brown Rovner, I. (2007). The effects of negative political campaigns: A meta-analytic reassessment. *The Journal of Politics, 69*, 1176-1209.

Madden, C.J., & Therriault, D.J. (2009). Verb aspect and perceptual simulations. *The Quarterly Journal of Experimental Psychology, 62(7)*, 1294-1302.

Madden, C.J., & Zwaan, R.A. (2003). How does verb aspect constrain event representations. *Memory & Cognition, 31*, 663-672.

Magliano, J.P., & Schleich, M.C. (2000). Verb aspect and situation models. *Discourse Processes, 29*, 83 – 112.

Matlock, T. (in press). The conceptual motivation of aspect. In G. Radden, P. Koch, & K. Panther (Eds.), *Motivation in lexicon, grammar, and discourse.* Amsterdam: John Benjamins.

Morrow, Daniel G. 1985. Prominent characters and events organize narrative understanding. *Journal of Memory and Language* 24: 304-319.

Morrow, Daniel G. 1990. Spatial models, prepositions, and verb aspect markers. *Discourse Processes* 13: 41-469.

Quattrone, G.A., & Tversky, A. (1988). Contrasting rational and psychological analyses of political choice. *The American Political Science Review, 82*, 719-736.

Rozin, P., & Royzman, E. (2001). Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review*, *5*, 296-320.

Truitt, T.P., & Zwaan, R.A. (1997, November). Verb aspect affects the generation of instrumental inferences. Paper presented at the 38th Annual Meeting of the Psychonomic Society, Philadelphia.

Westen, D. (2007). *The political brain: The role of emotion in deciding the fate of the nation.* New York, NY: Public Affairs.

How much for a transitive? Subtle linguistic cues influence blame and punishment

Caitlin M. Fausey (cmfausey@stanford.edu)
Lera Boroditsky (lera@stanford.edu)
Department of Psychology, 450 Serra Mall, Building 420, Stanford University
Stanford, CA 94305 USA

Abstract

When bad things happen, how do we decide who is to blame and how much they should be punished? In this paper we examined whether subtly different linguistic descriptions of accidents influence how much people blame and punish those involved. In three studies, participants judged how much people involved in particular accidents should be blamed and how much they should have to pay for resulting damage. The language used to describe the accidents differed subtly between conditions: either agentive (transitive) or non-agentive (intransitive) verb forms were used. Agentive descriptions led participants to attribute more blame and request higher financial penalties than non-agentive descriptions. Further, linguistic framing influenced judgments even when participants reasoned about a well-known event like the 'wardrobe malfunction' of Super Bowl 2004. Importantly, this effect of language held even when people were able to see the event for themselves on video. These results demonstrate that even when people have rich established knowledge and visual information about events, linguistic framing can shape event construal, with important real-world consequences. Subtle differences in linguistic descriptions can change the way people construe what happened and how they attribute blame and dole out punishment.

Introduction

When bad things happen, how do we decide who is to blame and how much they should be punished? Linguistic and contextual framing has been shown to affect people's reasoning in a variety of domains (e.g., Lee, Frederick, & Ariely, 2006; Levin, 1987; Levin & Gaeth, 1988; Loftus, Miller, & Burns, 1978; Loftus & Palmer, 1974; Shiv, Carmon, & Ariely 2005; Tversky & Kahneman, 1973; Tversky & Kahneman, 1981), including causal attribution (see Pickering & Majid, 2007, for a recent review). In this paper we build on this work by exploring the effects of linguistic framing in a domain of paramount real-world importance: blame and punishment.

Linguistic descriptions are of course ubiquitous in legal disputes. People linguistically frame incidents right from the very moment they occur and later in police reports, legal statements, court testimony and public discourse. Could the linguistic descriptions of an event influence how much we blame the people involved? Could language also influence how financially liable we think a person is for any resulting damage? Could linguistic framing shape construal even for well-known events (ones for which we already have rich knowledge and established mental representations) and even when we can witness the event with our own eyes?

The particular linguistic contrast of interest in this paper is between transitive agentive descriptions and intransitive non-agentive descriptions. A canonical agentive description (e.g., *Timberlake ripped the costume*) includes a person as the subject in a transitive expression describing a change of state (in this case, ripping). A canonical non-agentive description (e.g., *The costume ripped*) is intransitive and does not place the person as the subject for the change of state event.[1] Previous work has shown that people are sensitive to this distinction between agentive and non-agentive frames. For example, people are more likely to remember the agent of an event when primed with agentive language than with non-agentive language (e.g., Fausey & Boroditsky, 2010). The attributional consequences of these linguistic frames, however, are not well understood.

The linguistic contrast between agentive and non-agentive frames has the potential to have serious real-world consequences, especially in legal contexts. For example, in the 197,745 trials held between 1674 and 1913 at London's central criminal court (*Old Bailey Proceedings Online*, 2009), cases with the agentive phrase "*broke it*" in the court records resulted in a guilty verdict more often than cases with the non-agentive phrase "*it broke*" (76% and 70% guilty, respectively), with similar patterns for other consequential actions such as "*burned it*" versus "*it burned*" (77% and 57% guilty, respectively), $\chi^2(1, N = 2748) = 11.04$, $p < .05$. In the most serious of cases (when the charge was "killing"), the transitive/intransitive contrast as marked by different verbs also predicted verdicts. Saying "*killed*" resulted in more guilty verdicts than saying "*died*" (65% and 56% guilty, respectively), $\chi^2(1, N = 3814) = 21.34$, $p < .05$. These examples suggest that agentivity may be part of a suite of linguistic cues that are influential in legal reasoning.

In a correlational analysis like this, however, it is impossible to determine whether different linguistic forms actually caused a difference in verdicts. It could be that agentive descriptions indeed led the court more often to guilty verdicts. But it is also possible that people were simply more likely to use agentive language in cases where the defendant was actually more guilty. While the attributional consequences of transitivity have not been directly explored in the empirical literature, the question has been debated (and adjudicated!) in court. For example, in a case petitioning to change the title of a ballot measure (California's high-profile Proposition 8 in the 2008 election

[1] The distinction we draw here is different from active versus passive voice (e.g., Garvey, Caramazza, & Yates, 1976; Kassin & Lowe, 1979; White, 2003). Here we focus on transitivity and investigate not just the attributional consequences of transitivity (blame) but also the concrete real-world outcomes of these attributions (punishment).

titled "Eliminates right of same-sex couples to marry"), the judge rejected the petitioners' claim, ruling that "*There is nothing inherently argumentative or prejudicial about transitive verbs*" (*Jansson v. Bowen*, 2008). Few other questions in psycholinguistics have risen to a sufficient level of civic importance to be ruled on in high court.

With the high stakes of guilt, innocence and the legality of constitutional amendments on the line, it is important to empirically establish whether agentive and non-agentive frames indeed have any attributional consequences. In this paper we examine the effects of agentive and non-agentive linguistic frames on important real-world decisions about blame and punishment.

Study 1

In this study, participants read about an accidental restaurant fire that resulted in property damage. They then made judgments about the person involved in the accident. The survey was one of many unrelated surveys in a packet presented to participants.

Method

Participants. 236 students at Stanford University (96 male; mean age = 19.22 years) completed one survey in partial fulfillment of a course requirement. 116 read the agentive version of the story and 120 read the non-agentive version of the story.

Materials. Participants read either the agentive or the non-agentive account about an individual – Mrs. Smith – involved in a restaurant fire, and then answered two questions (Table 1). The two accounts contain all of the same content words (all of the same nouns, verbs and adjectives are used), involve the same individual and describe the same outcomes. The accounts differ only in the frames used to describe the accidental events (underlined in Table 1): transitive frames are used in the agentive account and intransitive frames in the non-agentive account.

Results and Discussion

Linguistic framing influenced both people's judgments of blame and financial liability. Participants who read the agentive account (M = 4.83, SE = .14) blamed Mrs. Smith more than did participants who read the non-agentive account (M = 4.01, SE = .15), $t(234) = 4.04$, $p < .001$, d = .53. Impressively, a subtle difference in language caused a big difference in dollars: people who got the agentive report ruled that Mrs. Smith should pay $247, or 36%, more in fines (M = $935.17, SE = $43.48) than participants who got the non-agentive report (M = $688.75, SE = $43.64), $t(234) = 3.99$, $p < .001$, d = .52.

In Study 1, linguistic framing influenced people's judgments of financial liability. One explanation for this result could be that Mrs. Smith was punished more harshly because she was also blamed more harshly. That is, the effect of language on financial liability might be indirect, such that language influences blame, which then determines punishment. Could language *directly* impact judgments of financial liability? This question is important because of the somewhat flexible sentencing process that occurs after guilt judgments in legal decision-making. A direct impact of language on sentencing would be an important applied result. Study 2 was designed to address this question.

Study 2

In Study 2, participants got an agentive or non-agentive accident description and also learned of a blame attribution generated by an independent review panel. This panel attributed either low, middle, or high blame to the person involved in the accident. After learning how blameworthy other people judged the person to be, participants determined the person's financial liability for the property damage. This paradigm allows us to target the independent role of language on financial liability sentences. People's decisions about financial liability may be guided by blameworthiness, language, or both.

Table 1: Studies 1 and 2 Reports and Questions

Agentive Report
Mrs. Smith and her friends were finishing a lovely dinner at their favorite restaurant. After they settled the bill, they decided to head to a nearby café for coffee and dessert. Mrs. Smith followed her friends and as she stood up, she flopped her napkin on the centerpiece candle. She had ignited the napkin! As Mrs. Smith reached to grab the napkin, she toppled the candle and ignited the whole tablecloth too! As she jumped back, she overturned the table and ignited the carpet, as well. Hearing her desperate cries, the restaurant staff hurried over and heroically managed to put the fire out before anyone got hurt.
Non-agentive Report
Mrs. Smith and her friends were finishing a lovely dinner at their favorite restaurant. After they settled the bill, they decided to head to a nearby café for coffee and dessert. Mrs. Smith followed her friends and as she stood up, her napkin flopped on the centerpiece candle. The napkin had ignited! As Mrs. Smith reached to grab the napkin, the candle toppled and the whole tablecloth ignited too! As she jumped back, the table overturned and the carpet ignited, as well. Hearing her desperate cries, the restaurant staff hurried over and heroically managed to put the fire out before anyone got hurt.
Questions for Study 1
Blame Mrs. Smith is discussing the damage with the restaurant. How much should she be blamed for the fire? (Likert scale from 1 to 7, anchored by "Not at all to blame" and "Completely to blame".) **Financial Liability** The restaurant's insurance policy does not cover minor fires. The restaurant has sought legal action to require Mrs. Smith to pay for the damage. Total costs to the restaurant were $1500. How much should Mrs. Smith be required to pay?
Question for Study 2
Financial Liability The restaurant's insurance policy does not cover minor fires and so the restaurant has sought legal action to require Mrs. Smith to pay for the damage. An independent review panel used their standard blame assessment scale in reviewing this case. On this scale, 0 means "not at all to blame" and 8 means "completely to blame". The panel gave Mrs. Smith a {1,4,7}. The total costs to the restaurant were $1500. How much should Mrs. Smith be required to pay?

Method

Participants. 179 students at Stanford University (59 male; mean age = 19.01 years) completed one survey in partial fulfillment of a course requirement. 91 read the agentive account of the restaurant fire accident (33 low-blame, 30 mid-blame, 28 high-blame) and 88 read the non-agentive account (33 low-blame, 28 mid-blame, 27 high-blame).

Materials. As in Study 1, participants read either the agentive or the non-agentive narrative and then answered the financial liability question shown in Table 1. Thus, participants in this study answered only the financial liability question, after learning that an independent panel judged the person to be either a "one" (low), a "four" (mid) or a "seven" (high) in terms of blame.

Results and Discussion

The level of blame assigned by the independent panel influenced participants' judgments of financial liability (Figure 1). Overall, people judged that Mrs. Smith should pay more in damages when the independent panel ruled her to be highly to blame (M = \$974.19, SE = \$61.97) than when the panel assigned her a middle level of blame (M = \$615.00, SE = \$56.27) than when she was ruled to be of low blame (M = \$425.63, SE = \$50.89).

Interestingly, language also influenced financial liability judgments. As in Study 1, a subtle change in language led to a substantial change in financial liability: Mrs. Smith was held responsible for \$153, or 26%, more in damages by people who got the agentive report (M = \$730.75, SE = \$49.57) than by those who got the non-agentive report (M = \$577.77, SE = \$52.35).

A 3 (Blame: Low, Mid, High) by 2 (Language: Agentive, Non-agentive) factorial ANOVA revealed reliable main effects of assigned blame level (F (2, 173) = 25.23, p < .001) and of language (F(1, 173) = 5.53, p = .02). Assigned blame level and language did not interact, F (2, 173) = 1.40, *n.s*.

Guilt and linguistic framing independently influenced how much someone was required to pay for accidental property damage. Increasing assigned blame led to greater financial liability and agentive framing led to greater financial liability than non-agentive framing. This finding replicates the result from Study 1. Further, sentencing itself appears to be susceptible to linguistic framing effects.

Results from the first two studies suggest that agentive and non-agentive language can shape how people attribute blame and financial liability to individuals involved in accidents. Of course, in these two studies the only information that reasoners had about the accident was linguistic. Were people inevitably swayed by language because it was the only thing that guided what they imagined about the event? Perhaps people who received differently phrased reports imagined substantially different scenarios of what happened? In many real-life situations, the information we have about an event is purely linguistic – in court arguments, insurance claims, news accounts. But in other situations we may also have visual evidence, either as eye-witnesses or on videotape. Would linguistic framing still have an effect even if people were able to see the event with their own eyes? Further, the restaurant fire described in Studies 1 and 2 was a novel event, one for which participants had no other previous information. Would people be so easily influenced by linguistic framing if they were reasoning about an event that they already knew something about, for which they already had a rich set of mental representations?

To address these questions, we capitalized on a widely known, much discussed, well-publicized and video-recorded event: the "wardrobe malfunction" of Super Bowl 2004 when a performance by Justin Timberlake and Janet Jackson ended with Janet Jackson's breast being exposed on national television. Post-experiment questioning confirmed that this is indeed a well-known event; nearly all of our participants (96.9%) had heard about it and many had also seen the video (67.9%) before the experiment. With prior knowledge, and current visual evidence, could linguistic framing still influence blame and punishment?

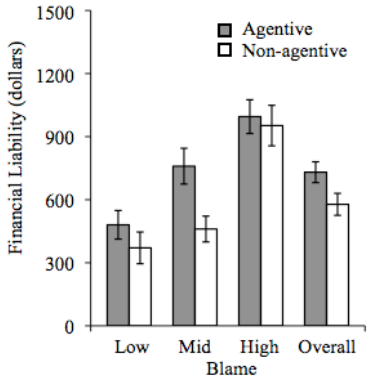


Figure 1: Independent contributions of guilt and linguistic framing to financial liability sentences (Study 2). Mean values are plotted on the y-axis, with whiskers representing +/- 1 SEM.

Study 3

In Study 3, participants reasoned about the wardrobe malfunction incident under one of three conditions: (a) they read about the incident, (b) they first read about the incident and then watched the video, or (c) they first watched the video and then read about it. In each condition, people read either an agentive or non-agentive account of the incident.

Method

Participants. 589 participants (188 male; mean age = 31.17 years) were paid for completing one survey online. Participants were recruited from the pool of English speakers who use Amazon's Mechanical Turk (https://www.mturk.com/mturk/welcome). 306 read the agentive account of the event (116 read-only; 88 read-then-watch; 102 watch-then-read) and 283 read the non-agentive account of the event (93 read-only; 106 read-then-watch; 84 watch-then-read).

Materials and Design. Participants read either the agentive or non-agentive account of the "wardrobe malfunction" incident (Table 2). In two conditions participants viewed a video of the final six seconds of the performance, which included the infamous malfunction.

After reading about the incident (and in two of the conditions also watching it on video), participants answered the questions shown in Table 2. The order of the three response options was randomized and the particular order presented to each participant was the same for the blame and financial liability judgments. Because Timberlake initiated movement right before the "wardrobe malfunction" and also because of his prominent apology to Super Bowl viewers (in which he coined the very phrase "wardrobe malfunction", Timberlake, 2004), our narratives focused on the actions of Timberlake. As a result, we expected that any effects of linguistic framing should be strongest for judging the guilt and financial liability of Timberlake. Also, because the FCC tried to fine CBS for broadcasting the incident, CBS was included among the possible targets for financial liability.

Results and Discussion

In brief, linguistic framing affected people's judgments of blame and financial liability in all conditions: language mattered whether it was presented before, after, or without video evidence. The main results of interest are shown in Figure 2.

Conclusions from these data are the same whether all three framing contexts are considered (as reported below) or whether only the two multimodal contexts are considered. Conclusions are also supported by nonparametric analyses.

Blame and financial liability attributions were analyzed using a 2 (Language: Agentive, Non-agentive) by 3 (Task context: Read-only, Read-then-watch, Watch-then-read) factorial ANOVA for each dependent measure. For clarity of presentation, we focus on effects of language here. Language and task context never interacted.

Blame. Linguistic framing influenced people's blame attributions (Figure 2a). Overall, people blamed Timberlake more after reading agentive language (M = 38.76%, SE = 1.59%) than after reading non-agentive language (M = 30.49%, SE = 1.43%), $F(1, 583) = 17.94$, $p < .001$. The effect of language was seen across the three conditions, with no interaction of the effect of language by condition, $F(2, 583) = .15$, *n.s.*

Language also affected attributions to chance. Overall, people attributed the outcome to chance more after reading non-agentive language (M = 42.87%, SE = 2.40%) than after reading agentive language (M = 33.92%, SE = 2.26%), $F(1, 583) = 8.99$, $p = .003$. Again this effect of language was seen across the three conditions, with no interaction of the effect of language by condition, $F(2, 583) = .20$, *n.s.*

Financial liability. The modal response for financial liability was $0 (57.2% of all data). This is likely because the sentence "*Eventually the fine was dismissed in court*" appeared in the liability question. Nevertheless, the linguistic framing of the event influenced people's judgments about financial liability. Overall, the proportion of people who gave any non-zero amount of financial liability to Timberlake depended on linguistic framing. 46.7% assigned a non-zero fine after reading agentive language, while only 38.5% did so after reading non-agentive language, $\chi^2(1, N = 589) = 4.05$, $p = .044$.

The amount of money for which Timberlake was held liable likewise depended on linguistic framing (Figure 2b). Participants who got the agentive report asked that Timberlake pay an extra $30,828.69, or 53%, more in fines than those who got the non-agentive report (*Agentive M* = $88,818.12, *SE* = $8,115.75; *Non-agentive M* = $57, 989.43, *SE* = $6,465.34), $F(1, 575) = 10.31$, $p = .001$.[2,3,4] Again there was no interaction of the effect of language by condition, $F(2, 575) = 1.22$, *n.s.*

Agentive and non-agentive linguistic framing did not affect people's attributions of blame or financial liability to Janet Jackson or CBS.

Table 2: Study 3 Reports and Questions

Agentive Report
Justin Timberlake and Janet Jackson performed during the 2004 Superbowl Half-time Show. Toward the end of the song, Timberlake followed Jackson across the stage and stood beside her. As they sang the last line, Timberlake reached across the front of Jackson's body. In this final dance move, he unfastened a snap and tore part of the bodice! He slid the cover right off Jackson's chest! This incident made for a lot of controversy.
Non-agentive Report
Justin Timberlake and Janet Jackson performed during the 2004 Superbowl Half-time Show. Toward the end of the song, Timberlake followed Jackson across the stage and stood beside her. As they sang the last line, Timberlake reached across the front of Jackson's body. In this final dance move, a snap unfastened and part of the bodice tore! The cover slid right off Jackson's chest! This incident made for a lot of controversy.
Questions
Blame. In your opinion, was someone to blame or was it just chance? Please allocate the percentage of blame. Be sure your numbers add up to 100%! (Response options: Justin Timberlake, Janet Jackson, Chance)
Financial Liability. The FCC (Federal Communications Commission) tried to fine CBS $550,000 for this incident. Eventually the fine was dismissed in court. How much do you think each of the parties below should have been fined for this incident? (Response options: Justin Timberlake, Janet Jackson, CBS)

[2] Eight participants whose financial liability responses exceeded $550,000 were excluded from this analysis.

[3] These conclusions are the same when analyses consider just those participants who assigned Timberlake a non-zero fine (N = 244). Among these participants, those who got the agentive report assigned more fines (M = $193,726.47 , SE = $12,893.53) than those who got the non-agentive report (M = $153,179.61, SE = $12,430.78), $t(242) = 2.22$, $p = .028$.

[4] These data show some heteroscedasticity, but our main conclusions remain the same after appropriate corrections. A t-test which does not assume equal variances confirms a reliable difference between the financial liability assigned by participants who got agentive versus non-agentive reports, $t(559.36) = 2.97$, $p = .003$.

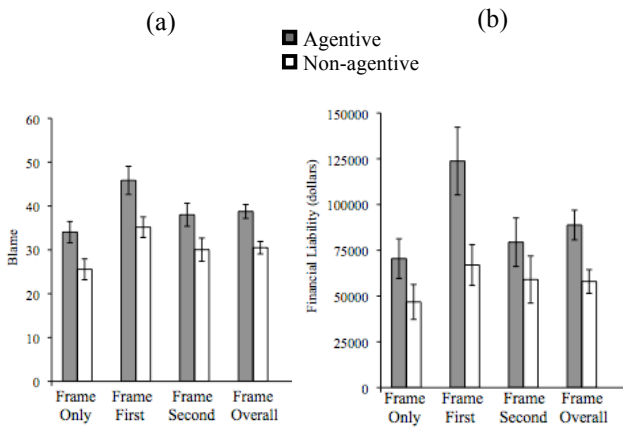


Figure 2: Language changes punishment of an observed individual (Study 3). (a) Blame attribution to Timberlake, (b) Financial liability to Timberlake. Mean values are plotted on the y-axis, with whiskers representing +/- 1 SEM.

In an additional set of analyses, all of the reported contrasts were conducted with an additional factor: whether or not the participant reported having seen the video of this incident prior to the experiment. This factor was not a reliable main effect nor did it interact with effects of linguistic framing in any of the analyses.

Linguistic framing influenced how much people punished an individual involved in an event, even when they witnessed the event with their own eyes, and even though the event was one our participants already knew about. Agentive language led to harsher punishment than non-agentive language. Replicating results from the first two studies, linguistic framing not only influenced attributions of blame but also financial liability. In the case of the wardrobe malfunction incident, an agentive report led people to think that Justin Timberlake owed more than $30,000 more (an extra 53%) in fines compared to a non-agentive report. In real-world contexts, visual evidence of accidents is rarely presented in the absence of linguistic framing. These results suggest that the form of this framing guides punishment.

General Discussion

In three studies, linguistic framing influenced participants' judgments about blame and punishment. Financial liability judgments in particular were strongly affected by linguistic framing: agentive descriptions led to 30-50% more in requested financial damages than non-agentive descriptions. Judgments of financial liability were affected by linguistic frame even when blame was held constant. This finding suggests that linguistic framing can have an influence not only on verdicts of guilt and innocence, but also on the sentencing process. Impressively, linguistic framing influenced reasoning even about an event that people knew a lot about, had seen before, and witnessed (again) right before judging the individual involved.

Previous inquiries into effects of language on attribution have examined the role of verbs, voice, and word order in guiding how people determine the cause of an event (e.g., Brown & Fish, 1983; Garvey, Caramazza, & Yates, 1976; Kasof & Lee, 1993; Kassin & Lowe, 1979; Pryor & Kriss, 1977; Schmid & Fiedler, 1988; Semin, Rubini, & Fiedler, 1995). Here, we provide the first report on the impact of transitivity on both people's attributions of blame and also on the real-world outcomes of these attributions (punishment). These studies extend previous research in several important ways. First, we probed people's decisions about a concrete form of punishment – financial liability, freely estimated in dollars – in addition to more abstract ratings of blame. Second, we examined effects of linguistic framing in the presence of previous knowledge as well as with current visual evidence – a condition that is absent from many previous attribution framing studies but present in many real-world reasoning contexts. Finally, we considered the transitive/intransitive alternation, a property of event description that both has important real-world consequences and differs interestingly across languages.

Previous work has shown that languages differ from one another in their preference for agentive versus non-agentive frames (e.g., Fausey & Boroditsky, 2010; Fausey, Long, & Boroditsky, 2009). The present findings raise the possibility that speakers of different languages may prescribe more or less severe punishment as a function of the frequency of particular grammatical frames in their language. While there have been many demonstrations showing the power of linguistic frames in shaping people's decisions, there has not been much contact between such findings and the literature investigating cross-linguistic differences in cognition. Establishing that linguistic framing has psychological consequences in a domain where languages naturally differ from one another opens the possibility for connecting these two rich bodies of knowledge.

In particular, as Sher and McKenzie (2006) have pointed out, the linguistic frames typically provided in framing studies often are not informationally equivalent. Each linguistic description is situated in a set of pragmatic norms within a language, and participants may be responding to the pragmatic cues implied by the choice of frame. The possibility of cross-linguistic comparisons offers an exciting extension to the framing literature: rather than having frames provided by an experimenter, in the cross-linguistic case, speakers of different languages may self-generate different frames for the same events because of the prevalent patterns in their respective languages (e.g., Maass, Karasawa, Politi, & Suga, 2006). In this way, cross-linguistic comparisons may allow us to investigate conceptual framing not just as a phenomenon in the communicative context (where participants may use pragmatic information to infer what the experimenter must mean by their choice of frame), but also in contexts where the participant naturally frames the event for themselves.

The linguistic (and cross-linguistic) framing of agentivity is of particular importance in court proceedings. Filipovic (2007) highlights a case from Northern California, in which a Spanish-speaking defendant's non-agentive (and

appropriate in Spanish) description of events ("*se me cayó*", roughly "*to me it happened that she fell*") was translated into English for the broader court into the agentive (and appropriate in English) "*I dropped her.*" Do these two descriptions mean the same thing? Or does this change in framing have serious attributional consequences? Our results raise the possibility that speakers of different languages may arrive at rather different conclusions regarding blame and punishment for the same events.

In three studies we find that agentive descriptions of events invite more blame and more severe punishment than do non-agentive descriptions. These results demonstrate that even when people have knowledge and visual information about events, linguistic framing can significantly shape how they construe and reason about what happened. In the case of agentive and non-agentive language, subtle differences in linguistic framing can have important real-world consequences. Deciding how much to blame an individual, and how much to hold them financially liable, appears to be broadly susceptible to linguistic framing.

Acknowledgments

We thank V. Vanchinathan and N. Heitz for help with data collection and entry. This research was supported by an NSF Graduate Research Fellowship to CMF and an NSF Career Award to LB.

References

Brown, R., & Fish, D. (1983). The psychological causality implicit in language. *Cognition, 14*, 237-273.

Fausey, C.M., & Boroditsky, L. (2010). *Whodunnit? Cross-linguistic differences in eyewitness memory*. Manuscript submitted for publication.

Fausey, C.M., Long, B.L., & Boroditsky, L. (2009). The role of language in eye-witness memory: Remembering who did it in English and Japanese. In N.A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, Austin, TX: Cognitive Science Society.

Filipovic, L. (2007). Language as a witness: Insights from cognitive linguistics. *The International Journal of Speech, Language and the Law, 14(2)*, 245-267.

Garvey, C., Caramazza, A., & Yates, J. (1976). Factors influencing assignments of pronouns and antecedents. *Cognition, 3*, 227-243.

Jansson v. Bowen, No. 34-2008-00017351 (Sacramento Superior Court August 7, 2008).

Kasof, L., & Lee, J.Y. (1993). Implicit causality as implicit salience. *Journal of Personality and Social Psychology, 65,* 877-891.

Kassin, S.M., & Lowe, C.A. (1979). On the use of single sentence descriptions of behavior in attribution research. *Social Behavior and Personality, 7(1),* 1-8.

Lee, L., Frederick, S., & Ariely, D. (2006). Try it, you'll like it: The influence of expectation, consumption, and revelation on preferences for beer. *Psychological Science, 17(12)*, 1054-1058.

Levin, I.P. (1987). Associative effects of information framing. *Bulletin of the Psychonomic Society, 25*, 85-86.

Levin, I.P., & Gaeth, G.J. (1988). How consumers are affected by the framing of attribute information before and after consuming the product. *Journal of Consumer Research, 15,* 374-378.

Loftus, E.F., Miller, D.G., & Burns, H.J. (1978). Semantic integration of verbal information into a visual memory. *Journal of Experimental Psychology: Human Learning and Memory, 4*, 19-31.

Loftus, E. F., & Palmer, J.C. (1974). Reconstruction of automobile destruction: An example of the interaction between language and memory. *Journal of Verbal Learning and Verbal Behavior, 13*, 585-589.

Maass, A., Karasawa, M., Politi, F., & Suga, S. (2006). Do verbs and adjectives play different roles in different cultures? A cross-linguistic analysis of person representation. *Journal of Personality and Social Psychology, 90*, 734-750.

Old Bailey Proceedings Online (2009). Retrieved November 3, 2009, from www.oldbaileyonline.org.

Pickering, M. J., & Majid, A. (2007). What are implicit causality and consequentiality? *Language and Cognitive Processes, 22(5),* 780-788.

Pryor, J.B., & Kriss, M. (1977). The cognitive dynamics of salience in the attribution process. *Journal of Personality and Social Psychology, 35(1),* 49-55.

Schmid, J., & Fiedler, K. (1998). The backbone of closing speeches: The impact of prosecution versus defense language on judicial attributions. *Journal of Applied Social Psychology, 28(13),* 1140-1172.

Semin, G.R., Rubini, M., & Fiedler, K. (1995). The answer is in the question: The effect of verb causality upon locus of explanation. *Personality and Social Psychology Bulletin, 21*, 834-842.

Sher, S., & McKenzie, C.R.M. (2006). Information leakage from logically equivalent frames. *Cognition, 101*, 467-494.

Shiv, B., Carmon, Z., & Ariely, D. (2005). Placebo Effects of Marketing Actions: Consumers May Get What They Pay For. *Journal of Marketing Research, 42(4),* 383-393.

Timberlake, J. (February 1 2004). "Statement From Justin Timberlake," PR Newswire.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology, 5*, 207-232.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211*, 453-458.

White, P.A. (2003). Effects of wording and stimulus format on the use of contingency information in causal judgment. *Memory & Cognition, 31(2),* 231-242.

Can mirror-reading reverse the flow of time?

Daniel Casasanto[1,2] **(daniel.casasanto@mpi.nl)**
Roberto Bottini[1,3] **(roberto.bottini@unibg.it)**
[1]Max Planck Institute for Psycholinguistics, Neurobiology of Language Group, Nijmegen, NL
[2]Donders Center for Brain, Cognition, and Behavior, Radboud University, Nijmegen, NL
[3]University of Bergamo, Department of Human Sciences, Bergamo, IT

Abstract

Across cultures, people conceptualize time as if it flows along a horizontal timeline, but the direction of this implicit timeline is culture-specific: in cultures with left-to-right orthography (e.g., English-speaking cultures) time appears to flow rightward, but in cultures with right-to-left orthography (e.g., Arabic-speaking cultures) time flows leftward. Can orthography influence implicit time representations independent of other cultural and linguistic factors? Native Dutch speakers performed a space-time congruity task with the instructions and stimuli written in either standard Dutch or mirror-reversed Dutch. Participants in the Standard Dutch condition were fastest to judge past-oriented phrases by pressing the left button and future-oriented phrases by pressing the right button. Participants in the Mirror-Reversed Dutch condition showed the opposite pattern of reaction times, consistent with results found previously in native Arabic and Hebrew speakers. These results demonstrate a causal role for writing direction in shaping implicit mental representations of time.

Keywords: Culture, Metaphor, Orthography, Space, Time

Introduction

Space and time are intertwined in the human mind, as they are in the physical world. The theory that people use spatial representations to think about time, first inspired by patterns in metaphorical language (Clark, 1973; Lakoff & Johnson, 1980), is now supported by numerous behavioral and neuroscientific experiments (e.g., Basso, et al., 1996; Boroditsky, 2000; Casasanto & Boroditsky, 2008; Torralbo et al., 2007; Weger & Pratt, 2008).

Yet, the way people use space to *talk* about time is not necessarily the same way they use space to *think* about it. In English and many other languages, metaphors suggest that time flows along the sagittal (front-back) axis: deadlines lie *ahead of us* or *behind us*; we can *look forward* to our golden years or *look back* on our salad days. Other languages also make use of the vertical axis to talk about time. In Mandarin Chinese, 'the up month' means a month earlier and 'the down month' a month later (Boroditsky, 2001). Yet, no known spoken language uses the lateral (left-right) axis to talk about time conventionally, and invented left-right metaphors for time sound nonsensical: Monday comes *before* Tuesday, not *to the left of Tuesday* (Cienki, 1998).

Despite the total absence of left-right metaphors in spoken language, there is strong evidence that people implicitly associate time with left-right space. Furthermore, the direction in which time flows along people's imaginary timeline varies systematically across cultures. In one study, Tversky, Kugelmass, & Winter (1991) asked children and adults to place stickers on a page to indicate where *breakfast* and *dinner* should appear relative to the *lunch* sticker, in the middle of the page. Whereas English speakers placed breakfast on the left and dinner on the right of lunch, Arabic speakers preferred the opposite arrangement. Fuhrman and Boroditsky (2007) showed a similar pattern in a reaction time (RT) task. English- and Hebrew-speaking participants judged whether the second of two pictures showed an earlier or later stage of an unfolding event. English speakers' judgments were fastest when *earlier* was mapped to the left button and *later* to the right, but Hebrew speakers showed the opposite pattern. Ouellet, et al. (in press) asked Spanish and Hebrew speakers to judge auditorily presented words referring to the past or future with either their left or right hand, and found a similar reversal of the lateral space-time mapping across groups.

These experimental data reflect patterns that can be found in spontaneous behavior, as well. When English speakers produce co-speech gestures they tend to use the lateral axis for time, much more often than the sagittal axis (Casasanto, 2009a; see also Boroditsky, 2008; Cienki, 1998; Cooperrider & Nunez, 2009). Earlier times are on the left and later times on the right of body-centered space. Preliminary data from our lab suggests that Spanish speakers' gestures follow a similar pattern, but Arabic speakers' spontaneous gestures show the reverse mapping (Romàn, Casasanto, Jasmin, & Santiago, in prep).

Across cultures, the direction in which time flows along the mental timeline varies predictably with the orthography of the dominant language: time flows rightward in cultures whose literate members use a left-to-right orthography and leftward in cultures that use a right-to-left orthography. Yet, despite this clear correlation, it is not known to what extent reading and writing direction is a *cause* or an *effect* of cross-cultural variation in implicit space-time mappings.

In principle, a culture's writing system could emerge with one directionality or another as a consequence of culture-specific conceptions of time -- not the other way around. This seems especially plausible for cultures where literacy (or mass-literacy) is a recent development. Alternatively, directionality in *both* orthography and in thought could arise due to cultural bootstrapping from material artifacts like calendars (whether a grid on a piece of paper, knots on a string, notches on a branch, etc.) or other devices for keeping track of time (e.g., a solar clock) or number (e.g., a horizontal abacus; Dehaene, 1999). Cultural practices tend to covary: groups who write from left to right

often spatialize time on calendars and numbers on graphs from left to right, as well. Based on the correlational data reviewed above, it is not possible to determine whether experience reading or writing plays any causal role in fixing the direction of implicit space-time mappings.

Here we performed an experimental intervention to determine whether experience with reading a left-to-right or right-to-left orthography is sufficient to determine the direction of people's implicit associations from space to time. Native Dutch speakers were assigned to perform one of two space-time congruity tasks. In one task (Experiment 1), participants saw past-oriented phrases (e.g. *a year earlier*) and future-oriented phrases (e.g. *a decade later*) appear on the screen one at a time, in standard Dutch orthography. As soon as each phrase appeared, they pressed a button (located on the left or right of a keyboard) to indicate the temporal reference of the phrase (past or future). Each participant performed two blocks: in one block the left-right key mapping required responses that were congruent with a left-to-right flow of time, and in the other responses were congruent with a right-to-left mapping. The order of blocks was counterbalanced across participants. We predicted that, on average, participants would show an RT advantage for responses consistent with standard Dutch orthography (left-to-right).

The other task (Experiment 2) was identical to the first, with one exception: all instructions and stimuli were presented in mirror-reversed text. Reading requires scanning the page in a particular direction: normally for Dutch speakers reading each line of a text requires moving the eyes gradually from the left to the right side of the page or the computer screen. As such, moving rightward in space is tightly coupled with 'moving' later in time. We reasoned that if the habit of reading from left-to-right contributes to an implicit left-to-right mapping of time in readers' minds, then practice reading in the opposite direction should weaken and eventually reverse this mapping.

Experiment 1: Standard Orthography

In Experiment 1, all instructions and stimuli were presented in standard Dutch orthography. We conducted Experiment 1 to validate the use of this space-time congruity paradigm in native Dutch speakers, and to provide a comparison group for the mirror-reading group.

Methods

Participants Native Dutch speakers (N=32) performed Experiment 1 in exchange for payment.

Stimuli Temporal phrases were constructed in Dutch, each with 3 words. The first word was an indefinite article, the second word a temporal interval, (tr., *second*, *moment*, *minute*, *hour*, *day*, *week*, *month*, *season*, *year*, *decade*, *century*, *millennium*), and the third word a temporal modifier (tr., *before*, *after*, *earlier*, *later*). The twelve temporal intervals were fully crossed with the four temporal modifiers to produce 48 temporal phrases (e.g., *a day before*; *a century after*; *a year earlier*; *a week later*). Half of the phrases referred to an earlier (past-oriented) interval of time (i.e., if the modifier was *earlier* or *before*), and the other half referred to a later (future-oriented) interval (i.e., if the modifier was *later* or *after*). Two of the modifiers were spatial terms used metaphorically (*before*, *after*), and the other two were purely temporal terms with similar meanings (*earlier*, *later*). Phrases were presented in the center of a Macintosh laptop screen (resolution=1024x768), in black 48-point Arial font, on a white background.

Apparatus Participants were seated at a desk. Two A4 Xerox paper boxes were stacked on the desk, and a laptop computer was secured on top of them, to raise the screen to approximately the participants' eye-level. A standard USB keyboard was mounted horizontally on the side of the upper box, with the keys facing the participant, at about shoulder level. The keyboard was covered with a sheet of black plastic with holes that exposed only the three keys needed for responses: the "A" key on the left, the "apostrophe" key on the right, and the "H" hey in the middle. The middle key was aligned with the center of the laptop screen, and the left and right keys were equidistant from it. The left key was covered with a blue sticker and the right key a red sticker, or vice versa, with the key colors counterbalanced across subjects.

Procedure The experiment consisted of two blocks. In each block, each of the 48 temporal phrases was presented once, for a total of 96 trials. Written instructions appeared on the screen before each block. In one of the blocks, participants were instructed that as soon as each phrase appeared, they should press the blue button if the phrase referred to an interval of time in the past (e.g., a week *earlier*) and the red button if it referred to an interval of time in the future (e.g., a week *later*). In the other block, the mapping between the red/blue keys and pastward/futureward phrases was reversed. To ensure that participants remembered the correct color-time mapping, after reading the instructions they were required to rehearse the correct color-time mapping aloud 5 times, before each block (e.g., "past=blue, past=blue, past=blue, etc.; future=red, future=red, future=red, etc.)

At the beginning of each trial the word 'ready' appeared in the center of the screen and remained there until the participant pressed the middle white button. 'Ready' was then replaced by a fixation cross. Participants were instructed to hold down the white button for as long as the fixation was shown. Its duration was varied randomly from 300-450 ms, in 50 ms increments, to make its duration unpredictable and discourage anticipatory movements. The fixation was then replaced by one of the 48 temporal phrases. Participants were instructed to press the colored button corresponding to the temporal reference of the phrase as quickly and accurately as possible. The phrase remained on the screen until the participant responded, at which time it was replaced by the 'ready' message to begin the next trial.

Participants pressed buttons with the index finger of the dominant (writing) hand. To ensure they would used the same hand for both rightward and leftward responses, participants were required to sit on their non-dominant hand.

The spatial direction of responses was never mentioned, but one colored button was on the right and the other on the left of the middle white button. Therefore, in one block pressing the correctly colored button called for a movement that was congruent with the space-time mapping encoded in standard Dutch orthography (e.g., pressing the blue button for a pastward phrase when the blue button was on the left); in the other block pressing the correctly colored button called for an incongruent movement (e.g., pressing the blue button for a futureward phrase when the blue button was on the left). The order of congruent-movement and incongruent-movement blocks was counterbalanced across participants. The space-time congruity effect was computed for each subject by comparing response times during Congruent and Incongruent responses (between-blocks, within-items). Testing lasted about 10 minutes.

Results and Discussion

Participants pressed the correct button on 96% of trials. Only accurate responses were analyzed. This resulted in the removal of 4% of the data. Responses greater than 5000 ms were also excluded, which resulted in the removal of 0.2% of the accurate trials trials.

A 2 X 2 ANOVA was conducted with Congruity of Movement Direction (Congruent with Time flowing leftward, Congruent with time flowing rightward) and Block (Block 1, Block 2) as within-subject and within-item factors. There was a highly significant main effect of Congruity ($F_1(1,15)=30.56$, $p=.0001$; $F_2(1,47)=119.38$, $p=.0001$). There was no main effect of Block ($F_1(1,15)=0.75$, *ns*; $F_2(1,47)=2.99$, *ns*). The Congruity X Block interaction was significant by items but not by subjects ($F_1(1,15)=1.41$, *ns*; $F_2(1,47)=6.27$, $p=.02$).

Congruity of Movement was then compared within each block (Block 1: $F_1(1,30)=9.62$, $p=.004$; $F_2(1,47)=116.31$, $p=.0001$; Block 2: $F_1(1,30)=3.64$, $p=.07$; $F_2(1,47)=32.55$, $p=.0001$). Mean RTs are shown in figure 1.

Overall, there was a strong effect of Congruity. Participants responded faster when the mapping between the color of the buttons and the temporal reference of the phrases required leftward movements for past-oriented phrases and rightward movements future-oriented phrases. This space-time congruity effect is similar to effects found previously in English and Spanish speakers (e.g., Torralbo, et al., 2006; Weger & Pratt, 2008). We are not aware of previous studies showing this effect in Dutch speakers, but given the correlation between writing direction and the direction of the space-time mappings across cultures, we had no reason to expect that Dutch speakers should perform differently from speakers of other languages that use a Roman alphabet.

For our present purposes, it is important that this paradigm produced a congruity effect in the same direction for both blocks. Having shown that this task provides clear evidence for the implicit space-time mapping typically found in left-to-right reading cultures, we can proceed to test effects of exposure to an orthography in which 'progress' along a spatio-temporal continuum proceeds in the opposite direction.

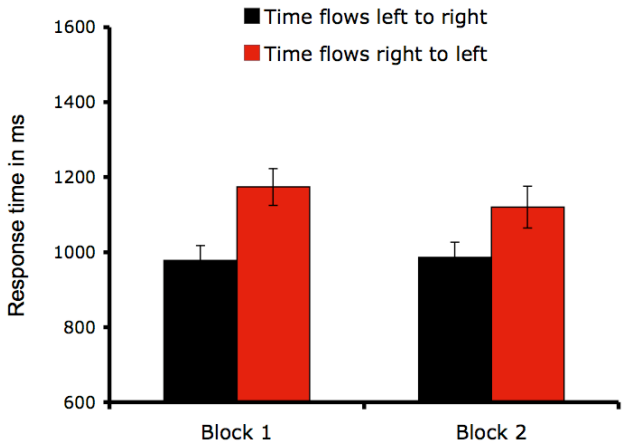


Figure 1. Results of Experiment 1. Error bars indicate s.e.m.

Experiment 2: Mirror-Reversed Orthography

To test for a causal role of orthography in the mental representation of temporal order, we replicated Experiment 1 in a new group of participants using stimuli and instructions presented in mirror-reversed font.

Methods

Participants A new sample of native Dutch speakers (N=32) performed Experiment 2 in exchange for payment.

Materials and Procedure

Materials and procedures were identical to Experiment 1, with one exception. All instructions and stimuli were presented mirror-reversed. Testing lasted about 15 minutes.

Results and Discussion

Accuracy

Participants pressed the correct button on 97% of trials. Only accurate responses were analyzed. This resulted in the removal of 3% of the data. Responses greater than 5000 ms were also excluded, which resulted in the removal of 4% of the accurate trials trials.

A 2 X 2 ANOVA was conducted with Congruity of Movement Direction (Congruent with Time flowing leftward, Congruent with time flowing rightward) and Block (Block 1, Block 2) as within-subject and within-item factors. There was no main effect of Congruity ($F_1(1,15)=.79$, *ns*; $F_2(1,47)= 2.29$, *ns*). There was a highly significant effect of Block ($F_1(1,15)= 66.37$, $p=.0001$; $F_2(1,47)= 321.81$, $p=.0001$), and crucially, a highly significant Congruity X Block interaction ($F_1(1,15)= 31.89$, $p=.0001$; $F_2(1,47)= 206.56$, $p=.0001$).

Congruity of Movement was then compared within each block (Block 1: $F_1(1,30)=5.00$, $p=.03$; $F_2(1,47)=98.36$, $p=.0001$; Block 2: $F_1(1,30)= 3.02$, $p.=.09$; $F_2(1,47)= 125.21$, $p=.0001$). Mean RTs are shown in figure 2.

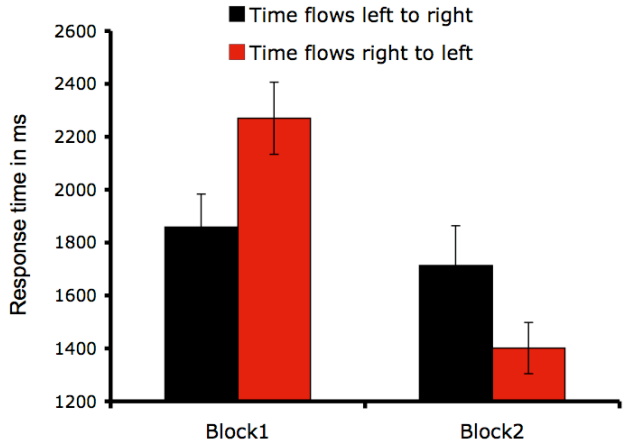


Figure 2. Results of Experiment 2. Error bars show s.e.m.

Finally, we compared the congruity effects found in Experiment 1 and Experiment 2 using a 2 X 2 X 2 ANOVA with Congruity and Block as within-subject/within-item factors and Orthography (Standard orthography, Mirror-reversed orthography) as a within-subject/within-item factor. Consistent with the prediction that orthography can influence mental representations of time, we find a highly significant 3-way interaction ($F_1(1,30)= 22.71$, $p=.0001$; $F_2(1,94)=125.38$, $p=.0001$). By subtracting the RTs during trials where movements were congruent with the leftward flow of time from RTs during trials where movements were congruent with the rightward flow of time (RT_rightward - RT_leftward), this 3-way interaction can be simplified, and conceptualized as a 2-way interaction of Block X Orthography (see figure 3).

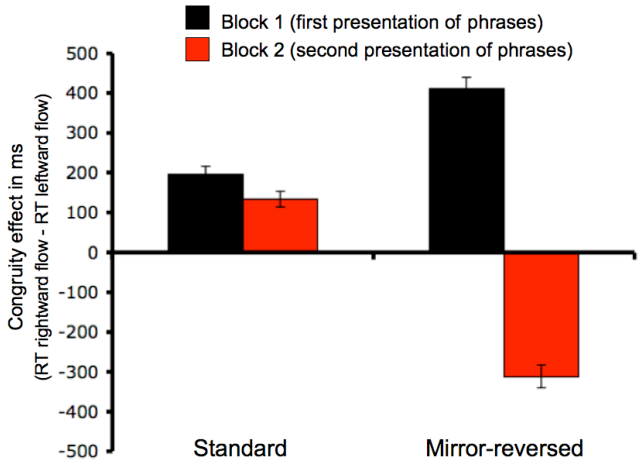


Figure 3. Congruity effects across blocks for Experiment 1 (left) and Experiment 2 (right). Error bars indicate s.e.m.

As is evident from figure 3, the absolute values (ABS) of the congruity effects in both blocks of Experiment 2 are greater than the ABS of the effects in Experiment 1. This was not expected, and although it is not relevant to our experimental hypothesis, it bears further investigation. On one possible explanation, congruity effects result from a failure of cognitive control; that is, they may result from participants' inability to ignore the irrelevant spatial dimension of their responses when judging the temporal reference of the stimuli. The cross-dimensional effect of space on time judgments may have been greater in Experiment 2 because cognitive control resources were taxed by reading backwards, contrary to habit.

Although the dominant space-time mapping in Dutch culture continued to influence RTs during the first block of Experiment 2, by the second block exposure to mirror-reversed writing was sufficient to reverse the congruity effect. Since this is the first experiment to test for a causal influence of writing direction on time representation, we did not have any *a priori* prediction about how much experience with reversed orthography would be needed to produce a significant change in the congruity effect, nor could we predict whether the congruity effect would be reversed or merely diminished. To support our hypothesis, it would have been sufficient to show a reduction in the left-to-right congruity effect from Block 1 to Block 2 that was greater in Experiment 2 than in Experiment 1. However, the fact that the congruity effect completely reversed here provides a particularly clear demonstration that even brief experience with one orthography or another can influence people's implicit spatial representations of time. (For compatible evidence of the flexibility of space-time metaphors in language and thought, see Boroditsky, 2000; 2001; Casasanto, 2008; Clark, 1973; Evans, 2004; Torralbo, et al., 2006).

General Discussion

It is now well established that people activate implicit associations between space and time when processing temporal language, and that the specifics of these associations vary systematically across cultures (Fuhrman & Boroditsky, 2007; Ouellet, et al. in press; Tverksy, Kugelmass, & Winter, 1991). Since time is not associated with left-right space in any known linguistic metaphors, it is unlikely that these culture-specific mappings are learned through experience with spoken language.[1] Here we tested whether orthography can play a causal role in fixing the direction in which time flows along the imaginary mental timeline. Experiment 1 showed that, when exposed to temporal phrases presented in standard left-to right orthography, Dutch speakers implicitly associated earlier time intervals with leftward movements and later time intervals with rightward movements, consistent with previous findings in members of other cultures that use the Roman writing system.

[1] Although spoken languages do not use the lateral axis for time, some signed languages do (see Emmorey, 2001).

However, when exposed to several minutes of mirror-reversed writing, Dutch participants began to show space-time congruity effects that revealed a reversal of their normally dominant implicit space-time mapping. By the second time they were judging each of the 48 temporal phrases (Block 2 of Experiment 2), participants were faster to make responses when key presses associated earlier events with *rightward* movements and later events with *leftward* movements -- a pattern observed previously in speakers of Hebrew, which is written from right to left. It appears that experience reading a right-to-left orthography (which requires the reader to 'progress' leftward across the screen with his/her eyes) is sufficient to reverse the flow of time in the reader's mind, at least transiently.

Although this rapid retraining of a space-time association stored in long-term memory may seem surprising, it is not unprecedented. In one study, Boroditsky (2001) found that horizontal spatial primes facilitated English speakers' judgments of temporal sentences (e.g., *April comes earlier than May*) more than vertical primes did, but found the opposite pattern in Mandarin speakers, consistent with the difference between these languages in the prevalence of horizontal and vertical metaphors for time. To test whether linguistic experience could affect these mappings, she trained a new group of English speakers to use Mandarin-like vertical spatial metaphors for time. After brief training, English speakers showed a pattern of priming similar to native Mandarin speakers.

In a test of a different set of space-time metaphors Casasanto (2008) and colleagues showed that when English and Greek speakers perform non-linguistic duration reproduction tasks, they show language-specific patterns of cross-dimensional interference from space. Whereas English speakers have a harder time screening out interference from (1-dimensional) spatial distance, Greek speakers have more difficulty screening out interference of (3-dimensional) volume. This pattern was predicted based on the relative prevalence and productivity of distance and volume metaphors for duration across languages (e.g., a *long* time (like a *long* rope); *a large amount of* time (like *a large amount* of water)). To find out whether using volume metaphors could cause the volume-interference found in Greeks, US English speakers were trained to use Greek-like volume metaphors for time. Results showed that after one brief (but concentrated) training session, English participants showed a pattern of cross-dimensional interference from volume in a low-level psychophysical task that was statistically indistinguishable from the pattern seen in native Greek speakers.

Time is not the only domain that appears to be mentally represented, in part, through spatial metaphors (which may or may not correspond to linguistic metaphors). Emotional valence is also spatialized on a left-right axis: whereas right-handers tend to associate the right hand and the right side of space with positive things and the left with bad, left-handers show the opposite set of implicit associations (Casasanto, 2009b). It was proposed that this mapping arises due to asymmetries in motor fluency: people like things on their dominant side better because they can interact with things on that side more easily. To test this proposal, Casasanto (2009c) asked right-handers to perform a 2-part training task. In the first part, they arranged dominoes according to a symmetrical pattern on a tabletop, standing them on end, moving both hands in synchrony. The challenge was that they were randomly assigned to wear a bulky ski glove one hand or the other while performing the task, which either enhanced their natural right-handedness or made them temporarily more skillful with their left hand.

After 12 minutes of this asymmetric motor experience, participants were taken to a different room by a different experimenter for some ostensibly unrelated questionnaire studies, one of which tested implicit associations between space and valence. This questionnaire was shown previously to produce distinctive patterns of judgments in right- and left-handers (Casasanto, 2009b). Participants whose training experience preserved their natural dominance showed the typical right-handers' pattern. But participants who had worn the skiglove on their right hand during training, becoming transiently left-handed, produced a pattern of responses that was indistinguishable from natural lefties'.

We are aware of one training study that manipulated writing direction in order to test a role for orthography in the spatial representation of gender and agency. Several studies suggest that males (seen as more agentive) tend to be represented to the left of females in the minds of people who speak left-to-right languages like English, but not for speakers of right-to-left languages like Arabic (Suitner, 2009). Yet, Suitner (2009) showed that this spatial bias can be nullified in speakers of Italian who are trained to perform a leftward writing exercise, reversing not only their habitual writing direction but also their habitual associations of gender, agency, and space.

How enduring are these training effects? Presumably, without further reinforcement of the new habits, participants who show rapid training effects will also revert to their long-term habits rapidly. Exactly how soon remains a question for further research. Depending on the goal of the training exercise, the durability of the behavioral change may matter more or less. In the present study the goal was to test the sufficiency of a proposed cause of cross-cultural differences. The total reversal of the congruity effect as a function of reading experience demonstrates that orthography can, indeed, influence the implicit spatial representation of time. This simple demonstration would serve its theoretical goal even if the effect were quickly reversed when participants resumed normal reading habits.

How best to characterize the learning mechanisms that afford this representational plasticity remains another open question. It may be fruitful to consider the changes participants undergo in Experiment 2 in terms of a hierarchical Bayesian model (Kemp et al., 2007). To sketch this suggestion briefly, people's associations between space and time could be characterized as intuitive hypotheses. Based on ordinary reading experience, Dutch speakers form

the hypothesis that by default events unfold from left to right. Yet after training, they appear to entertain the hypothesis that events unfold from right to left.

To explain how participants can switch from one hypothesis to a contradictory hypothesis (and presumably switch back) so quickly, it may help to posit that they also entertain a more enduring overhypothesis, of which both the 'Dutch-like' and 'Arabic-like' space-time associations are specific instances. The overhypothesis could be that time is associated with motion along a linear path. Such a belief would be well supported by observable correlations in the physical world: spatial succession is a reliable index of temporal succession.

Consistent with this proposal, we suggest that if orthography is responsible for determining the direction in which time flows along people's left-right mental timelines, this directional mapping likely builds upon a prior less-specific space-time association, which arises (either in developmental or evolutionary time) from space-time correlations that have no particular directionality: on any trajectory, it is the case that as a moving object travels farther, more time passes. The hierarchical model can help to explain how 'mental metaphors' linking space-time can be universal at one level of level of description but culture-specific at another.

Acknowledgments

This research was supported in part by the Max Planck Gesellschaft and by an NRSA fellowship (#F32MH072502) and by a grant from the Spanish Ministry of Education and Science (#SEJ2006-04732/PSIC, DGI) to DC.

References

Basso, G. et al. (1996). Time perception in a neglected space. *Neuroreport 7*, 2111 – 2114.

Boroditsky, L. (2000). Metaphoric structuring: understanding time through spatial metaphors. *Cognition*, 75(1), 1-28.

Boroditsky, L. (2001). Does language shape thought? English and Mandarin speakers' conceptions of time. Cognitive Psychology, 43(1), 1-22.

Casasanto, D. (2008). Who's afraid of the Big Bad Whorf? Cross-linguistic differences in temporal language and thought. *Language Learning, 58*(1), 63-79.

Casasanto, D. (2009a). When is a linguistic metaphor a conceptual metaphor? In V. Evans & S. Pourcel (Eds.), 127-145, *New Directions in Cognitive Linguistics*. Amsterdam: John Benjamins.

Casasanto, D. (2009b). Embodiment of Abstract Concepts: Good and bad in right- and left-handers. *Journal of Experimental Psychology: General. 138*(3), 351-367.

Casasanto, D. (2009c). Motor experience shapes abstract concepts. *Proceedings of the 50th Annual Meeting of the Psychonomic Society*. Boston, MA.

Casasanto, D. & Boroditsky, L. (2008). Time in the Mind: Using space to think about time. *Cognition, 106,* 579-593.

Cienki, Alan (1998). Metaphoric gestures and some of their relations to verbal metaphoric expressions. In Jean-Pierre Koening (Ed.), *Discourse and cognition: Bridging the gap* (pp. 189-204). Stanford: CSLI Publications.

Clark, H. H. (1973). Space, time, semantics and the child. In *Cognitive Development and the Acquisition of Language,* T. E. Moore (ed.), 27–63. New York: Academic Press.

Cooperrider, K and Nunez, R. (2009). Across time, across the body Transversal temporal gestures. *Gesture*, Vol. 9, pp. 181-206.

Dehaene, S., (1999). *The number sense*. Oxford University Press, Penguin press, New York, Cambridge (UK).

Emmorey, K. (2001). Space on hand: The exploitation of signing space to illustrate abstract thought. In M. Gattis (Ed), *Spatial schemas and abstract thought,* pp. 147 — 174, The MIT Press: Cambridge, MA.

Evans, V. (2004) *The structure of time: Language, meaning and temporal cognition*. Amsterdam: John Benjamins.

Fuhrman, O., & Boroditsky, L. (2007). Mental time-lines follow writing direction: Comparing English and Hebrew speakers. *Proceedings of 29th Annual Conference of the Cognitive Science Society*, Nashville, TN.

Kemp, C, Perfors, A. & Tenenbaum, J.B. (2007). Learning ovehypotheses with hierarchical Bayesian models. *Developmental Science*, 10 (3), 307-321.

Lakoff, G. and Johnson, M. (1980). *Metaphors We Live by*. Chicago: The Chicago University Press.

Ouellet, M., Santiago, J., Israeli, Z., & Gabay, S. (in press). Is the future the right time? *Experimental Psychology.*

Suitner, C. (2009). *Where to place social targets? Stereotyping and Spatial Agency Bias*. Doctoral dissertation, Department of Psychology, University of Padova.

Torralbo, A., Santiago, J., & Lupiáñez, J. (2006). Flexible conceptual projection of time onto spatial frames of reference. *Cognitive Science*, 30, 749–757.

Tversky, B., Kugelmass, S., & Winter, A. (1991). Cross-cultural and developmental trends in graphic productions. *Cognitive Psychology, 23*, 515-557.

Weger, U. W., & Pratt, J. (2008). Time flies like an arrow: Space-time compatibility effects suggest the use of a mental time-line. *Psychonomic Bulletin & Review*, 15, 426-430.

Implicit spatial length modulates time estimates, but not vice versa.

Roberto Bottini[1,2] **(roberto.bottini@unibg.it)**
Daniel Casasanto[2,3] **(daniel.casasanto@mpi.nl)**
[1]University of Bergamo, Department of Human Sciences, Bergamo, IT
[2]Donders Center for Brain, Cognition, and Behavior, Radboud University, Nijmegen, NL
[3]Max Planck Institute for Psycholinguistics, Neurobiology of Language Group, Nijmegen, NL

Abstract

How are space and time represented in the human mind? Here we evaluate two theoretical proposals, one suggesting a symmetric relationship between space and time (ATOM theory) and the other an asymmetric relationship (metaphor theory). In Experiment 1, Dutch-speaking participants saw 7-letter nouns that named concrete objects of various spatial lengths (*tr.* pencil, bench, footpath) and estimated how much time they remained on the screen. In Experiment 2, participants saw nouns naming temporal events of various durations (*tr.* blink, party, season) and estimated the words' spatial length. The implicit length encoded in object nouns modulated time estimates, but the implicit duration encoded in event nouns did not affect estimates of spatial length. Nouns that named short objects were judged to remain on the screen for a shorter time, and nouns that named longer objects to remain for a longer time. By contrast, variations in the duration of the event nouns' referents had no effect on judgments of the words' spatial length on the screen. This asymmetric pattern of cross-dimensional interference cannot be attributed to differences in the discriminability or perceptual salience of space and time in the stimuli. Results support metaphor theory and challenge ATOM.

Keywords: ATOM, Metaphor, Psychophysics, Space, Time

Introduction

Space and time are intimately related in the human mind, as they are in the physical world. But exactly *how* are these dimensions related? Here we evaluate two theoretical proposals, one suggesting a symmetric and the other an asymmetric relationship between space and time.

According to the first proposal, space and time are represented in the brain and mind by a common analog magnitude system, which also generates representations of number and quantity. This view, summarized in Walsh's ATOM (A Theory of Magnitude; 2003), is consistent with neurological data showing shared brain areas for processing space, time, and quantity (e.g., Basso, et al., 1996), and with many behavioral studies in animals and humans (e.g., Church & Meck, 1984; Fischer, 2003; Gallistel & Gellman, 2000; Cappelletti, et, al, 2009).

Implicit in ATOM is an assumption that these 'ATOMic' dimensions are symmetrically interrelated: not hierarchically related in the brain/mind. Accordingly, Walsh (2003) frames predictions in symmetrical terms, positing "overlapping brain regions" and "cross-domain, within-magnitude priming" between dimensions, without specifying any *directionality* to the priming (or interference) effects. Indeed, if space and time are both manifestations of the same general-purpose analog magnitude system, there may be no *a priori* reason to posit that one dimension should depend asymmetrically on another.

On an alternative proposal, space, time, and quantity are importantly related, but in a different way. According to theories of metaphorical mental representation (e.g., Lakoff & Johnson, 1999), representations of time, number, and quantity depend asymmetrically on representations of space. The claim that some domains are asymmetrically dependent on others, which is at the core of metaphor theory, was originally supported by patterns in metaphorical language. In English, it is nearly impossible to talk about domains like time without using words whose primary meaning is spatial (denotatively, developmentally, or historically (Clark, 1973)). Vacations can be *long* or *short*, meetings can be *moved forward* or *pushed back*, deadlines can loom *ahead* or lie *behind* us. Yet, it is far less common to use temporal words to talk about space (Lakoff & Johnson, 1999). This asymmetry in language has been echoed by behavioral findings in psycholinguistics (Boroditsky, 2000), cognitive development, (Casasanto, Fotakopoulou, & Boroditsky, in press), and psychophysics (Casasanto & Boroditsky, 2008).

In one set of studies, participants viewed lines of various spatial lengths that appeared on a screen for varying durations (Casasanto & Boroditsky, 2008). They were asked to estimate either the duration or the spatial length of each line, using mouse clicks. Participants were unable to ignore irrelevant spatial information when making judgments about duration, but not the converse. For stimuli of the same average duration, lines that extended shorter in space were judged to take a shorter time, and lines that extended longer in space were judged to take a longer time. By contrast, for stimuli of the same average spatial length, spatial estimation was not affected by the line's duration. This cross-dimensional asymmetry, predicted based on patterns in language, was shown here in non-linguistic psychophysical judgments. Five follow-up experiments varied the attentional, mnemonic, and perceptual demands of the stimuli, and all six experiments supported the same conclusion: mental representations of time depend on representations of space, more than vice versa.

This robust space-time asymmetry supports metaphor theory, but presents a challenge to ATOM. If space and time are both derived from (or are both manifestations of) a general-purpose magnitude metric, then why should representations of time depend on representations of space more than the other way around -- in adults and children, and in language and thought?

It might be possible to reconcile these results with ATOM by positing that in previous studies, space influenced time asymmetrically because space was either (a) the more discriminable dimension, or (b) the more perceptually salient dimension in the stimulus. Discriminability, in this context, refers to the resolution at which a dimension is sampled. Salience means the extent to which one dimension attracts attention relative to the other. Differences in discriminability and perceptual salience have been shown to modulate the strength or direction of cross-dimensional interference and priming effects across numerous studies (see Santiago, Román, & Ouellet, submitted, for review). In general, the dimension that is more discriminable or salient interferes with the dimension that is less discriminable or less salient. Can task-related differences in the relative discriminability or salience of stimulus dimensions account for the space-time asymmetries observed previously?

One set of studies reviewed above addressed these questions. Tests of cross-dimensional relationships often manipulate more levels of one dimension than of the other, creating an imbalance in discriminability (see Pansky & Algom, 1999). In the space-time experiments by Casasanto & Boroditsky (2008), however, there were 9 levels of each dimension fully crossed, to equate discriminability.

Differences in discriminability may correspond to differences in the accuracy, precision, or variability of judgments across domains. This complicates the interpretation of cross-dimensional interference effects. In the limit, if performance in one domain is perfect, there is no opportunity for variation in the other domain to influence it: the 'clean' domain can influence performance the 'messy' domain, but not vice versa. In Casasanto & Boroditsky's studies, however, within-domain performance was equivalent across space and time (see also Casasanto, Fotakopoulou, & Boroditsky, in press).

But is it possible that space was more salient than time in these studies? Following Garner (1976), Casasanto & Boroditsky (2008) asked participants to judge different dimensions of the same stimuli (e.g., the spatial or temporal extent of a line). Thus, people had the exact same perceptual input during space and time judgments. But this does not guarantee that the dimensions were equally perceptually salient: it is possible to *see* the spatial extent of a line, but not its duration. To address the concern that space may have been more salient than time, in one experiment each line was accompanied by a tone, which sounded for the duration that the line remained on the screen. Tones have temporal extent but no spatial extent. Thus, temporal information was available to the participant through two sensory channels, but spatial information through only one. Yet, increasing the salience of temporal information did not diminish the space-time asymmetry.

Still, on a skeptical interpretation, these previous studies did not definitively rule out cross-dimensional differences in perceptual salience. It is possible that space will *always* be more perceptually salient than time whenever perceptible spatial stimuli are used, since it is possible to perceive space, but arguably it is not possible to perceive time directly through the senses (Ornstein, 1969). The question remains, then, whether the space-time asymmetry would persist in psychophysical judgments if differences in the perceptual salience of space and time in the stimulus were eliminated.

In the present study, we eliminated differences in perceptual salience by eliminating perceptible variation in the critical dimension (space or time), altogether. We tested whether the implicit spatial information encoded in object nouns can influence estimates of time (in Experiment 1), and whether the temporal information encoded in event nouns can influence estimates of spatial length (in Experiment 2). Participants saw words presented one at a time and reproduced either the duration for which they remained on the screen or their spatial length, using mouse clicks as in Casasanto & Boroditsky (2008). In the duration estimation task (Experiment 1), the target words named objects of various spatial lengths (e.g., *pencil*, *clothesline*, *footpath*). All target words had the same number of letters in Dutch, and therefore the same physical length on the screen. In the spatial length estimation task (Experiment 2) the target words named events of various durations (e.g., *blink*, *party*, *season*). Again, all target words had the same number of letters, but they were presented with a varying number of spaces between letters (1-9 spaces), stretching them out to different spatial lengths on the screen.

Word meanings were irrelevant to the length and duration estimations. We expected, however, that participants would read the words while viewing them, and activate their meanings (voluntarily or involuntarily). Presumably, the meaning of an object noun typically includes a representation of the object's spatial form, and the meaning of an event noun a representation of the event's duration. If internally generated spatial and temporal representations cued by words are sufficient to modulate estimates of experienced duration and spatial length, then we should observe cross-dimensional interference. Following metaphor theory, we predicted that the cross-dimensional interference should be asymmetric, even in the absence of cross-dimensional differences in perceptual salience: spatial representations cued by object nouns should modulate estimates of their duration more than temporal representations cued by event nouns modulate estimates of their spatial extent on the screen.

Experiment 1: Does implicit spatial length modulate time estimates?

Experiment 1 tested whether the spatial length of a word's referent can modulate estimates of how much time the word remained on the screen.

Methods

Participants Native Dutch speakers (N=39) performed Experiment 1 in exchange for payment.

Materials Dutch nouns naming 9 concrete objects (Targets) and 9 abstract entities (Fillers) were presented on a computer monitor (resolution = 1024 x 768 pixels) for varying durations. The concrete nouns referred to objects whose characteristic spatial lengths ranged from short (normally measured in centimetres) to long (normally measured in kilometres). English equivalents of these nouns are listed here in order of increasing length: *cigarette*, *pencil*, *ruler*, *meter stick*, *bench, clothesline*, *footpath*, *lane*, *highway*. In Dutch, all 9 target nouns had 7 letters, and were presented on the screen in a fixed-width font (62-point Courier New). Therefore, the targets did not differ in their physical spatial lengths on the screen; rather, they differed in their implicit lengths (i.e., the typical spatial lengths of their referents).

The filler nouns referred to abstract entities that have no physical spatial length: *guess*, *idea*, *pride*, *opinion*, *envy*, *thought*, *philosophy*, *suspicion, dignity*. However, they varied in their number of letters in Dutch (from 3-11 letters) and therefore in their physical length on the screen (nine different lengths, varying from 50-450 pixels as measured from the left edge of the first letter to the right edge of the last letter). By contrast with the targets, the fillers did not differ in the implicit lengths of their referents; rather, they differed in their physical lengths on the screen.

Each target and filler word was presented 9 times throughout the experiment, for 9 different durations. Durations ranged from 1000 to 5000 ms in 500 ms increments. Fully crossing these 9 durations with the target words (which had 9 different implicit spatial lengths) produced 81 target trials. Likewise, fully crossing the 9 durations with the filler words (which had 9 different physical lengths on the screen) produced 81 filler trials. The 162 different trials were presented in random order, with fillers and targets intermixed. Words were presented in white letters on a black background in the center of the screen. Participants were tested individually and testing lasted about 30 minutes.

Procedure Participants viewed the 162 words, one word at a time, from a viewing distance of approximately 50 cm. Immediately after each word disappeared an "hourglass" icon appeared in the upper left corner of the monitor indicating that the subject should reproduce the amount of time the word remained on the screen. To estimate duration, subjects clicked the mouse once on the center of the hourglass, waited the appropriate amount of time, and clicked again in the same spot, thus indicating the beginning and end of the temporal interval. All responses were self-paced.

After the experiment there was a two-part debriefing. In the first part, the experimenter asked the participant "What do you think this experiment is about?" and "What do you think we were looking for?" to determine whether the participant was aware of any relationship between the implicit lengths of the target words and their durations. In the second part, participants saw each target word again, in random order, and verbally estimated the typical spatial length of the target words' referents (using an appropriate unit of measurement). These subjective length estimates were used in later analyses as predictors of subjective duration.

Results and Discussion

Four participants were removed from the analyses below: one for giving nonsensical answers in the debriefing, one for excessively poor time estimation performance according to the criterion used by Casasanto & Boroditsky (2008)[1], and two for guessing the that there was a connection between the meanings of the target words and time estimation.

For the remaining 35 participants, we first analyzed participants' duration estimates as a function of the actual duration of the stimuli. Overall, duration estimates for target words were highly accurate (mean effect of actual duration on estimated duration: $y=0.83x + 154.11$, $r^2=.99$, $df=7$, $p<.001$; fig 1a).

We then tested for effects of implicit length on duration estimation. Target words were rank-ordered according to the typical lengths of their referents (this *a priori* ranking was confirmed by participants' post-test length estimates). Non-parametric correlation showed that implicit spatial length affected estimates of duration ($y=3.77x + 2605.70$, $r_{s(\text{Spearman's rho})}=0.75$, $df=7$, $p<.001$; fig.1b).

Finally, we conducted a parametric analysis of the effect of implicit length on duration estimation. Participants' post-test ratings of the typical spatial length of each target word's referent were used as a predictor of their duration estimates. Ratings for each target item were averaged, and the average length estimates in meters were transformed by a base 10 logarithm. This analysis corroborated the non-parametric analysis, showing a highly significant effect of implicit spatial length on duration estimation ($y=5.60x + 2619.20$, $r^2=.57$, $df = 7$, $p< .001$).

Participants incorporated irrelevant spatial information into their temporal estimates. For stimuli of the same average duration, words with (spatially) shorter referents were judged to remain on the screen for a shorter time, and words with longer referents for a longer time. This was true even though the task did not require participants to process the words' meanings.

This result shows that perceptible spatial input is not necessary to modulate time estimates; rather, internally-generated spatial representations cued by words are sufficient. This outcome, *per se*, is equally consistent with metaphor theory and with ATOM. To distinguish between the theories, it is necessary to conduct a complementary experiment to determine whether implicit duration can affect estimates of spatial length, and whether cross-

[1] Participants were excluded if the slope of their within-domain duration or length estimates was less than 0.5 (see Casasanto & Boroditsky, 2008). This criterion, which resulted in the exclusion of only one participant overall, is unbiased with respect to the predicted cross-dimensional interference because length and duration are orthogonal in the designs of both experiments.

dimensional interference effects are as symmetric, as expected on ATOM (Effect of Space on Time ≈ Effect of Time on Space) or asymmetric, as predicted by metaphor theory (Effect of Space on Time > Effect of Time on Space).

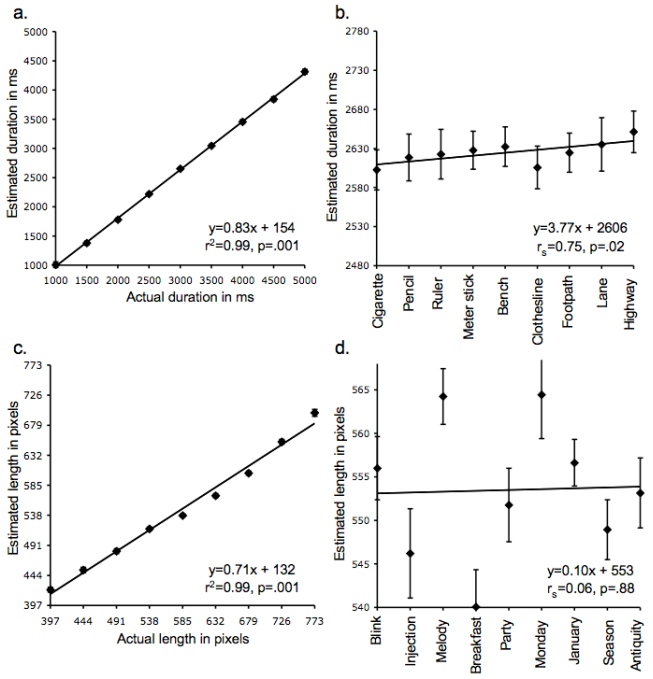


Figure 1. Results of Experiment 1 (top) and Experiment 2 (bottom). 1a. Within-domain effect of actual word duration on estimated duration. 1b. Cross-domain effect of words' implicit spatial length on estimated duration. 1c. Within-domain effect of actual word length on estimated spatial length. 1d. Cross-domain effect of words' implicit duration on estimated spatial length. The axes of the top and bottom plots (a-c, b-d) are proportional with respect to the total range of target values. Error bars show s.e.m.

Experiment 2: Does implicit duration modulate estimates of spatial length?

Experiment 2 tested whether the duration of a word's referent can modulate estimates of the word's spatial length as presented on the screen.

Methods

Participants Native Dutch speakers (N=35) performed Experiment 2 in exchange for payment.

Materials Dutch nouns naming 9 events (Targets) and nine concrete objects (Fillers) were presented on a computer monitor (resolution = 1024 x 768 pixels). The target nouns referred to events whose characteristic durations ranged from short (normally measured in seconds) to long (normally measured in years). English equivalents of these nouns are listed here in order of increasing duration: *blink*, *injection*, *melody*, *breakfast*, *party*, *Monday*, *January*, *Season*, *Antiquity*. All targets were presented for 3000ms. Therefore, the targets did not differ in the physical durations for which they remained on the screen; rather, they differed in their implicit durations (i.e., the typical durations of their referents).

The filler nouns referred to concrete objects that have no inherent duration: *doormat*, *ballast*, *portrait*, *detritus*, *crystal*, *device*, *case*, *sawdust, handle*. Each filler noun appeared for 9 different durations from 1000-5000ms, increasing in 500ms increments. By contrast with the targets, the fillers did not differ in the implicit durations of their referents; rather, they differed in the physical durations for which they remained on the screen.

In Dutch, all target and filler nouns had seven letters, and were presented on the screen in a fixed-width font (62-point Courier New). Each word was presented 9 times throughout the experiment, with a varying number of spaces in between the letters (1-9), to stretch the words out to 9 different spatial lengths on the screen. Due to the font selected, word lengths ranged from 397 to 773 pixels, in 47 pixels increments. Presenting each word at each of these 9 spatial lengths produced 81 filler trials and 81 target trials. For the fillers, spatial length was fully crossed with the physical duration for which they were presented. For the targets, spatial length was fully crossed with the implicit duration of their referents. The 162 different trials were presented in random order, with fillers and targets intermixed. Words were presented in white letters on a black background in the center of the screen. Participants were tested individually and testing lasted about 30 minutes.

Procedure Participants viewed the 162 words, one word at a time, from a viewing distance of approximately 50 cm. Immediately after each word disappeared an "X" appeared in the upper left corner of the monitor indicating that the subject should reproduce the spatial length that the word had occupied on the screen. To estimate length, subjects clicked the mouse once on the center of the X, moved the mouse to the right the appropriate distance, and clicked again, thus indicating the beginning and end of a spatial interval. All responses were self-paced.

After the experiment there was a two-part debriefing, as in Experiment 1. The first part was to determine whether the participant was aware of any relationship between the implicit durations of the target words and their spatial lengths. In the second part, participants saw each target word again, in random order, and verbally estimated the typical duration of the target words' referents (using an appropriate unit of measurement). These subjective duration estimates were used in later analyses as predictors of subjective spatial length.

Results and Discussion

One participant was removed from the analyses below for guessing that there was a connection between the meanings of the target words and spatial length estimation.

For the remaining 34 participants, we first analyzed participants' spatial length estimates as a function of the actual spatial length of the stimuli. Overall, length estimates

for target words were highly accurate (mean effect of actual length on estimated length: $y=0.71x + 132.44$, $r^2=.99$, df=7, p=.001; fig 1c).

We then tested for effects of implicit duration on spatial length estimation. Target words were rank-ordered according to the typical durations of their referents (this *a priori* ranking was confirmed by participants' post-test duration estimates). Non-parametric correlation showed that implicit duration did not affect estimates of spatial length ($y=0.10x + 553.00$, $r_{s(\text{Spearman's rho})}=0.06$, df = 7, *ns*; fig.1d).

Next, we conducted a parametric analysis using participants' post-test ratings of the typical duration of each target word's referent were used as a predictor of their length estimates. Ratings for each target item were averaged, and the average duration estimates in minutes were transformed by a base 10 logarithm. Again, there was no effect of implicit duration on spatial length estimation ($y= 0.04x + 553.39$, $r^2=.0003$, df=7, *ns*).

Finally, we compared the strength of the cross-dimensional interference effects across Experiments 1 and 2. The difference of correlations showed the predicted cross-dimensional asymmetry ($r_{\text{effect of spatial length on duration}}$-$r_{\text{effect of duration on spatial length}}=0.74$, z=1.66, p=0.05, one-tailed; see fig. 1b, 1d). This difference cannot be attributed to differences in within-domain performance ($r_{\text{effect of actual duration on estimated duration}}$-$r_{\text{effect of actual spatial length on estimated spatial length}}=0.00$, z=0.00, *ns*; see fig. 1a, 1c).

General Discussion

This study tested whether implicit spatial information encoded in concrete object nouns can influence estimates of time (in Experiment 1), and whether implicit temporal information encoded in event nouns can influence estimates of spatial length (in Experiment 2). When participants reproduced the duration for which an object noun remained on the screen, their estimates were influenced by the implicit length of the word's referent. Words that named shorter objects (e.g., *cigarette*, *pencil*) were judged to last a shorter time, and words that named longer objects (e.g., *bench*, *highway*) to last a longer time. By contrast, when participants reproduced the spatial length of an event noun, the duration of the word's referent did not influence judgments of spatial length.

This asymmetric pattern of cross-dimensional interference was predicted based on patterns in language: people talk about time in terms of space more than they talk about space in terms of time (Lakoff & Johnson, 1999). These data show that people incorporate spatial information into their temporal judgments even when they're not using any metaphorical language, and support the hypothesis that mental representations of time are asymmetrically dependent on representations of space: people use spatial length to think about duration, more than vice versa.

This space-time asymmetry cannot be attributed to differences in how well participants reproduced the actual durations and lengths of the stimuli, per se, since there was no significant difference between the effect of actual duration on estimated duration (fig. 1a) and the effect of actual length on estimated length (fig. 1c). Thus, differences in cross-dimensional interference were not due to differences in within-domain performance.

Furthermore, the space-time asymmetry cannot be attributed to differences in the perceptual salience of the interfering dimensions (i.e., space in Expt. 1, time in Expt. 2). In previous experiments, space could have influenced time asymmetrically because space is inherently more perceptually salient than time (which some scholars have argued can never be perceived directly (Ornstein, 1969)). But here there was no perceptible variation in the spatial component of duration-reproduction stimuli, and no perceptible variation in the temporal component of length-reproduction stimuli. Internally generated representations of spatial length, cued by words, were sufficient to modulate estimates of the words' physical duration. This was true even though the words' meanings were task-irrelevant.

Before discussing theoretical implications of these data further, it is important to consider whether the observed pattern could be due to unintended features of the stimulus words. For example, is it possible that duration estimates in Experiment 1 were influenced by implicit *speed* encoded in the concrete nouns, rather than implicit length? The three longest objects (*footpath*, *lane*, and *highway*) are all spatial paths. The speed of motion associated with these paths increases with their lengths (i.e., *footpath-walking*, *lane-slow driving*, *highway-fast driving*). The conflation of length and speed in these items was a consequence of restrictions on the stimuli: items had to increase in ordinal length unambiguously, and had to have 7 letters in Dutch.

If the effect of object length on duration estimates had been driven by these three items, this would be problematic. However, even a causal inspection of fig. 1b shows this was not the case. For the majority of the items there were no clear speed associations, and yet the effect of implicit length was found. For the first 5 items (*cigarette*, *pencil*, *ruler*, *meter stick*, *bench*), ordinal increases in implicit length corresponded to a monotonic increase in estimated duration. The predicted effect of length on duration was significant in these 5 items, alone ($y=6.84x + 2600$, $r_{s(\text{Spearman's rho})}=1.00$, p=.001). Thus, implicit speed was not responsible for the effect of implicit spatial length we report here (see Casasanto & Boroditsky, 2008, Expt. 6 for further evidence that spatial length affects duration estimates independent of speed).

On another skeptical possibility, could implicit *duration* encoded in object nouns have produced the observed effect on duration estimation? Looking at the longest and shortest items alone, this seems plausible. *Cigarette* could be associated with the time it takes to smoke a cigarette (a short time), and *highway* with the amount of time one typically drives on a highway (a longer time). Yet, looking at the full range of stimuli, this alternative explanation seems implausible. What durations are prepotently associated with *clothesline*, *pencil*, *ruler*, *bench*, or *meter stick*? Ordinal increases in spatial length predicted ordinal increases in

duration estimates for 7 out of the 8 ordinal pairs of stimuli (i.e., *cigarette* < *pencil*; *pencil* < *ruler*; *ruler* < *meter stick*; etc.) Pairwise differences in the typical spatial lengths of the words' referents are self-evident (and were confirmed by participants' post-test ratings), but for most of these word pairs, it seems unlikely that there are corresponding pairwise differences in durations associated with the words' referents.

Finally, although the space-time asymmetry cannot be due to differences in the *perceptual salience* of the interfering dimensions, could they be due to differences in *conceptual salience*? Could the spatial component of the object words' meanings be more salient than the temporal component of the event words' meanings? We cannot rule out this possibility definitively, but this seems unlikely to be the case. It is difficult to evaluate how salient spatial length is in the meaning of *bench* or *cigarette*, and to compare this with the salience of temporal duration in the meaning of *melody* or *party*. But a few of the stimuli are very strongly associated with a unit of space (*ruler*, *meter stick*) or a period of time (*Monday*, *January*, *season*, *Antiquity*). For these items, it is reasonable to assume that a spatial or temporal representation is the most salient aspect of the word's meaning. This was the case for only two of the object words (22% of targets) but for four of the event words (44% of targets). Therefore, overall, it seems likely that any asymmetry in conceptual salience favored the temporal meanings of the event words, thus working against the hypothesized space-time asymmetry.

These results suggest that the asymmetric dependence of time on space in psychophysical judgments is not an artifact of perceptual or conceptual asymmetries built into the stimuli. Rather, this performance asymmetry reflects a fundamental difference in the way people mentally represent space and time. Yet, this asymmetric relationship between space and time in the mind may, indeed, result from an asymmetry in how perceptible space and time are more broadly -- not in any particular experimental stimuli, but rather in the observable world, in general. Space and time are correlated in our everyday experiences (e.g., as objects travel farther more time passes), and tracking these correlations may be useful for anticipating changes in the physical environment. Correlation is a symmetrical relationship, but people may rely more heavily on the more perceptually available dimension (space), using it heuristically as an index of changes in the less perceptible dimension (time).

It appears that time and space are, in Garner's (1976) terminology, *asymmetrically separable* dimensions: it is possible to ignore irrelevant variation in time while judging space but not possible (or more difficult) to ignore irrelevant variation in space when judging time. At present, there is nothing in Walsh's (2003) ATOM proposal that can predict or explain the asymmetric separability of space and time. Yet, this cross-dimensional relationship is readily predicted by metaphor theory.

Importantly, space and time are predicted to be related *asymmetrically* but not *unidirectionally*. There is evidence that time can influence space in some paradigms (e.g., Miles, Nind, & Macrae, 2010), just as people can sometimes use temporal words to talk about space (e.g., "*I live two minutes from the station*" is a temporal metaphor for spatial distance). Simply showing that time can influence spatial judgments in some cases does not challenge the asymmetry we report here: to address the question of asymmetry, the cross-dimensional influences of time and space must be appropriately compared, controlling for salience and discriminability across dimensions, and for within-dimension performance.

We propose that Garner-like tests of dimensional separability will be critical for either modifying ATOM or deciding to abandon it in favor of a metaphorical theory of spatial, temporal, and numerical magnitude representation. In order to understand how space, time, and other prothetic dimensions are represented in the brain and mind, it is necessary to go beyond investigating *whether* these dimensions interact and determine *how* they interact.

References

Basso, G. et al. (1996). Time perception in a neglected space. *Neuroreport 7*, 2111 – 2114.

Boroditsky, L. (2000). Metaphoric structuring: understanding time through spatial metaphors. *Cognition*, 75(1), 1-28.

Cappelletti, M., Freeman, E.D. and Cipolotti, L. (2009). Dissociations and interactions between time, numerosity and space processing. *Neuropsychologia*, 47(13), 2732 – 2748.

Casasanto, D. & Boroditsky, L. (2008). Time in the Mind: Using space to think about time. *Cognition, 106,* 579-593.

Casasanto, D., Fotakopoulou, O, & Boroditsky, L. (2010). Space and Time in the Child's Mind: Evidence for a Cross-Dimensional Asymmetry. *Cognitive Science, 34,* 387-405.

Church, R.M. and Meck, W.H. (1984). The numerical attribute of stimuli. In *Animal Cognition* (Roitblat, H.L., Beaver, T.G. and Terrace, H.S., eds), pp. 445 – 464, Erlbaum.

Clark, H. H. (1973). Space, time, semantics and the child. In *Cognitive Development and the Acquisition of Language,* T. E. Moore (ed.), 27–63. New York: Academic Press.

Fischer, M. (2003). Cognitive representation of negative numbers. *Psych. Sci.*, 14(3), 278-282.

Gallistell, R.C. and Gellman, R. (2000). Non-verbal numerical cognition: from reals to integers. *Trends in Cog. Sci. 4*, 59 – 65.

Garner, W.R. (1976). Interaction of Stimulus Dimensions in Concept and Choice Processes. *Cog. Psych.*, 8, 98-123.

Lakoff, G., & Johnson, M. (1999). *Philosophy in the flesh: The embodied mind and its challenge to Western thought*. Chicago: University of Chicago Press.

Miles, L. K., Nind, L. K. and Macrae, C. N. (2010). Moving Through Time. *Psych. Sci.*, DOI: 10.1177/0956797609359333

Ornstein, R. (1969). *On the experience of time*. Hammondsworth: Penguin.

Pansky, A. & Algom, D. (1999). Stroop and Garner effects in comparative judgement of numerals: The role of attention. *JEP:HPP*, 25, 39-58.

Santiago, J., Román, A. and Ouellet, M. (submitted). Flexible foundations of abstract thought: A review and a theory.

Walsh, (2003). A theory of magnitude: common cortical metrics of time, space and quantity. *Trends in Cog. Sci., 7* (11), 483-488.

Emerging Insights from Eye-Movement Research on Category Learning

Bob Rehder (bob.rehder@nyu.edu)
Department of Psychology, New York University
6 Washington Place, New York, NY 10003

Mark R. Blair (mark_blair@sfu.ca)
Cognitive Science Program & Department of Psychology
Simon Fraser University, 8888 University Drive, Burnaby, BC, V5A 1S6, CANADA

Aaron B. Hoffman (aaron.hoffman@mail.utexas.edu)
Department of Psychology, University of Texas
1 University Station A8000, Austin, Texas 78712-0187

Marcus R. Watson (marcusw@psych.ubc.ca)
Department of Psychology, University of British Columbia
2136 West Mall, Vancouver, B.C., V6T 1Z4, Canada

Keywords: eye-tracking, category learning; categorization; attention; learning; optimization, Parkinson's disease, inference learning, error processing, working memory.

Introduction

Attempts to develop an accurate measure of eye movements are over a century old (e.g., Delabarre, 1898; Huey, 1898; as cited in Karatekin, 2007), and predate the earliest studies of categorization (Hull, 1920). Given the long history of both categorization and eye-tracking, it is surprising that eye-tracking has only recently been added to the categorization researcher's toolbox (Rehder & Hoffman, 2005a).

Selective attention is an important component of theories of categorization and eye-tracking provides a measure of what features of a stimulus participants have selected to attend. There are alternatives to eye-tracking, such as inferring attention allocation based on model fits or carefully designed transfer tasks. However, these methods lack the directness of eye-tracking and provide only a coarse measure of how attention shifts over the course of learning. Moreover, they provide no account of how attention is allocated early in learning and within a single categorization trial (including after feedback is presented). This more fine-grained data can not only clarify our understanding of key phenomena, it broadens the range of experimental questions that can be asked to understand how humans learn categories (Blair, Watson & Meier, 2009; Blair, Watson, Walshe & Maj, 2009; Hoffman & Rehder; 2009; Kim & Rehder, 2009; Rehder, Colner & Hoffman, 2009; Rehder & Hoffman, 2005a; Rehder & Hoffman, 2005b; Watson & Blair, 2008)

This symposium brings together four talks on eye-tracking and categorization. Each talk focuses on a different aspect of categorization and demonstrates how using eye-tracking can extend our knowledge. One recent trend in category learning is the use of alternative training procedures. The inference learning task is the most popular of these procedures and in the first talk Aaron Hoffman presents eye-tracking data illuminating the differences between inference learning and categorization. Bob Rehder then presents his recent work on understanding the learning difficulties associated with Parkinson's disease. Marcus Watson discusses work using eye-tracking to inform our understanding of the basic issue in category learning: error. Finally, Mark Blair discusses the relationship between working memory, attention and performance in a category learning tasks.

Inference versus classification learning

It has been proposed that whereas feature inference learning promotes learning a category's internal structure (e.g., typical features and feature correlations), classification promotes the learning of diagnostic information (Markman & Ross, 2003). We tracked learners' eye movements and found that inference learners fixated features that were unnecessary for inferring a missing feature—consistent with their acquiring the categories' internal structure. However, those fixations were limited to features that needed to be predicted on future trials. Inference learning appeared to induce both supervised and unsupervised learning of category-to-feature associations, rather than any general motivation to learn the internal structure of categories.

In a second study, we compared how inference and classification learning support learners' ability to draw *novel contrasts*—category distinctions that were not part of training. We found that classification learners were at a disadvantage at making novel contrasts. Eye movement data indicated that this *conceptual inflexibility* was due to (a) a narrow attention profile that fails to encode many category features and (b) learned inattention that inhibits the reallocation of attention to newly relevant information. Implications of these costs of supervised classification learning for views of conceptual structure will be discussed.

Using eye-movements to understand Parkinson's patients learning difficulties

Those with Parkinson's disease (PD) exhibit not only motor difficulties such as tremors, rigidity, and postural instability but also a variety of cognitive deficits, including deficits in procedural learning and in switching to new tasks ("set shifting"). Out central hypothesis is that deficits in selective attention are central to many of PD patients' learning difficulties. Moreover, assessing how attentional deficits in PD affect learning is critical to understanding how other learning mechanisms are affected by the disease. A probabilistic category learning paradigm known as the weather prediction task (WPT) has played a central role in theorizing about learning in PD patients. We report eye movement data from both PD patients and controls while performing the WPT and discuss implications our results have for current theories of category learning.

Over and under-estimating the importance of error-processing in categorization

The category label (i.e., the correct answer) has a central role in most models of categorization. It supplies the information necessary to improve both categorization and attentional performance. But despite its theoretical importance, there has been little direct investigation of how errors are processed.

In this presentation we first evaluate the necessity and sufficiency of errors for optimizing attention. Error-driven models predict large shifts of attention when errors are most common and the absence of shifts when learners are not making mistakes. We review data that shows the opposite result. We next use eye-tracking to assess how participants process stimuli and category labels while receiving feedback on their errors. Results show that temporal aspects of this process that are not captured in extant models are consequential for learning.

Working memory, attention and category learning

Categorization is a core cognitive task that involves accessing information, remembering relationships, focusing on relevant aspects of the stimuli, etc. While long-term memory and selective attention have long been employed by theories of categorization, working memory has had nothing much to do. This is especially surprising given that working memory is described by some researchers as executive attention, and its influence has been demonstrated to be very broad. Intuitively, working memory capacity might influence categorization performance in a variety of ways. High working memory span might be associated with faster learning or improved accuracy. It also might influence how participants attend to stimulus features.

This presentation will describe work aimed at demystifying the effects of working memory capacity on categorization performance, including on attentional optimization. Studies reveal that, depending on the task, working memory span is related to both attentional optimization and learning speed. Working memory span (measured by the symmetry span task) is compared to measures of attentional network efficiency (measured by the Attention Network Test), and to several other aspects of attentional learning and categorization data.

References

Blair, M.R., Watson, M. R., & Meier, K.M. (2009). Errors, efficiency, and the interplay between attention and category learning. *Cognition, 112,* 330-336.

Blair, M. R., Watson, M. R., Walshe, R.C., & Maj, F. (2009). Extremely selective attention: Eye-tracking studies on dynamic attentional allocation to stimulus features. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*, 1196-1206.

Delabarre, E. B. (1898). A method of recording eye-movements. *American Journal of Psychology*, 9, 572–574.

Huey, E. B. (1898). Preliminary experiments in the physiology and psychology of reading. *American Journal of Psychology*, 9, 575–586.

Hoffman, A.B. & Rehder, B. (2009). Attentional and representational flexibility of feature inference learning. In N. Taatgen, H. van Rijn, L. Schomaker, & J. Nerbonne (Eds.), Proceedings of the 31st Annual Conference of the Cognitive Science Society (pp. 1864-1869). Mahwah, NJ: Erlbaum.

Hull, C. (1920) Quantitative Aspects of the Evolution of Concepts. An Experimental Study. Psychological Monographs, 28.

Karatekin, C. (2007). Eye tracking studies of normative and atypical development. *Developmental Review*, 27, 283-348.

Kim, S. & Rehder, B. (2009). Knowledge effect the selective attention in category learning: An eyetracking study. In N. Taatgen, H. van Rijn, L. Schomaker, & J. Nerbonne (Eds.), Proceedings of the 31st Annual Conference of the Cognitive Science Society (pp. 230-235). Mahwah, NJ: Erlbaum.

Markman, A. B., & Ross, B. H. (2003). Category use and category learning. *Psychological Bulletin, 4*, 592-613.

Rehder, B., & Hoffman, A. B. (2005a). Eyetracking and selective attention in category learning. *Cognitive Psychology*, 51, 1-41.

Rehder, B., & Hoffman, A. B. (2005b). Thirty-something categorization results explained: Selective attention, eyetracking, and models of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31, 811-829.

Rehder, B., Colner, R.M., & Hoffman, A.B. (2009). Feature inference learning and eyetracking. Journal of Memory & Language, 60, 394-419

Watson, M. R., & Blair, M. R. (2008). Attentional allocation during feedback: Eyetracking adventures on the other side of the response. Proceedings of the 30th Annual Meeting of the Cognitive Science Society, 345-350.

Comparing Apples to Fruit:
Parent's Comparisons of Labels are Related to First and Second Label Learning

Chandra L. Brojde (chandrab@colorado.edu)
Department of Psychology and Neuroscience, CB345
Boulder, CO 80309 USA

Eliana Colunga (eliana.colunga@colorado.edu)
Department of Psychology and Neuroscience, CB345
Boulder, CO 80309 USA

Abstract

Young children often find it difficult to learn two labels for a single object. However, there is a great deal of variability across studies in children's bias to reject second labels. In this study, we investigated three possible factors affecting this variability including age, task, and parental input in a cross-sectional sample of children from 12- to 28-months-old. We show that children reject second labels differently depending on their age, task demands, and the amount and type of parental input. Importantly, there is also a correlation between the ways in which parent's use second labels and children's acceptance of first and second labels for objects. These results suggest that both previous experience and the task at hand determine children's learning of second labels.

Keywords: Word learning, language acquisition, vocabulary, parental input

Introduction

Sometime after their first birthday, children begin to add words to their vocabulary at an increasingly greater rate. These words almost exclusively share a one-to-one relationship with object categories. Learning more than one label for the same object, like "banana" and "fruit", can be difficult, especially for younger children (Liittschwager & Markman, 1994; Markman & Wachtel, 1988). The propensity to reject second labels can be useful when it comes to learning a new novel name. For example, when shown two objects, one familiar and one unfamiliar, and asked to hand the experimenter a "dax", children can correctly choose the unfamiliar object. But, this tendency can sometimes make it hard to learn certain kinds of words like adjectives (Hall, Waxman, & Hurwitz, 1993), part labels (Hansen & Markman, 2009; Saylor, Sabbagh, & Baldwin, 2002), proper names, (Gelman & Taylor, 1984) and labels at different levels of specificity (Au & Glusman, 1990).

Converging evidence from a variety of tasks supports the idea that children prefer a single label per object. There is also a great deal of variability from study to study in the degree to which children reject second labels. Context factors shown to influence this bias include: bilingual input (Au & Glusman, 1990; Davidson & Tell, 2005; Merriman & Kutlesic, 1993), pragmatic information (Bloom, 2000; Clark & Grossman, 1998; Diesendruck & Markson, 2001), and parts of speech and relationship between words (e.g. part versus a whole object or level of specificity) (Hall, Waxman, & Hurwitz, 1993; Saylor, Sabbagh, & Baldwin, 2002; Waxman & Senghas, 1992). All of these influences have in common that they depend on parental input. In this study we investigate the impact that the relationships between these different parental factors and how children learn second labels. Though attempts have been made to construct a unified explanation that includes all of these factors (Hollich et al., 2000), few studies have directly investigated the interaction between these input factors and the resulting impact on the learning of second labels.

In this study, we investigate the relationship parent input and the *process* of word learning. Specifically, we investigated the link between second label learning and the context in which second labels are learned. Both task differences and object properties may influence second label learning. We investigate both context variables in relation to parent's use of second labels in a naturalistic task.

Second Label Learning Tasks

In general, tasks used to measure second label learning can be separated into two groups (see Figure 1). They either 1) directly measure the child's ability to learn two labels for one object or they 2) require the child to infer by exclusion to which object a second label applies. This difference in task is often confounded with age such that older children do better than younger children when learning by exclusion (Markman, Wasow, & Hansen, 2003).

In direct learning, children are presented with a familiar object (e.g. a ball) and told that it is a "dax". They are then asked to identify the "dax" among one or more distractors. In this way, children are required to directly map the word

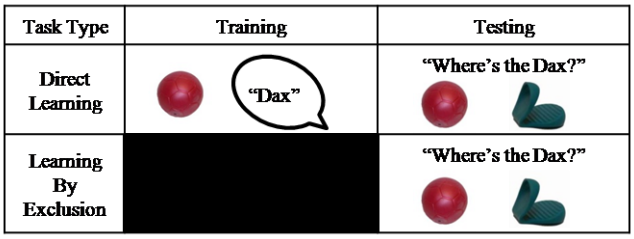


Figure 1. Examples of tasks used to test label learning.

"dax" to an object (Liittschwager & Markman, 1994; Mervis, Golinkoff, & Bertrand, 1994). Tasks requiring learning by exclusion, on the other hand, require that children infer the referent of a second label. For example, children may be shown two objects – one that they already have a name for and one that is unfamiliar. They are then simply asked to choose the "dax". Experimenters never directly label the unfamiliar object as "dax". Thus, children must infer that the novel word should refer to the unfamiliar object (Hollich et al., 2000; Markman, Wasow, & Hansen, 2003).

Parental Input

Several papers have also suggested that parental input influences second label learning. This is based on studies showing that parents differ in the amount and type of second labels they use for different age groups. This difference is related to vocabulary size (Callanan & Sabbagh, 2004; Masur, 1997). This conjecture is reasonable given that parental input effects language development in several ways (Girolametto, Weitzman, Wiigs, & Pearce, 1999; Hoff & Naigles, 2002; Huttenlocher, Vasilyeva, Cymerman, & Levine, 2002). The drawback to studies showing a difference in parent input is that they do not investigate the subsequent impact on children's biases. Furthermore, these studies have not investigated the interaction between amount and type of parental input and the type of task used to test children's second label learning. Different types of input may affect first and second label learning differently and may interact with task.

In this study, we investigate the amount and type of second labels that parents use and any to second label learning across age. Five age groups, from 12- to 28-months, were tested on their ability to learn first and second labels directly and by exclusion. These same children were also videotaped playing with one of their parents. This allowed us to not only determine the amount and type of second label use by parents but look at the interaction between parental input and task type.

Method

Participants

One-hundred and twenty child-parent dyads were recruited for the study, including 24 dyads per each of five age groups, included 12-, 16-, 20-, 24, and 28-month-olds. Equal numbers of males and females were included and were distributed approximately equally across age groups.

Materials

Parents completed a vocabulary checklist of words their child says using the MacArthur Communicative Development Inventories (MCDI) (Fenson et al., 1994). Total vocabulary size was determined using the number of items that parents indicated their child knew. In addition, parents and children completed two tasks twice, once each in two different sessions. The two tasks were always completed in the same order at each session. Tasks are described separately below.

Label Learning Task Children were taught four new labels (e.g. "lep") for four new objects, counterbalanced across two sessions. At one session they were taught two new labels for two familiar objects (i.e. a ball and a spoon). At the second session they were taught two different new labels for two unfamiliar objects (i.e. a rubber pot holder and a honey dipper). At each session, during training children saw eight objects in the following order: three objects that weren't labeled, one object that was labeled with a first new label, three more objects that weren't labeled, and a final object that was labeled with a second new label.

Children were then tested on six types of trials. The first two trial types were control trials: 1) known label trials where they were asked to pick an object they knew (e.g. doll) from two familiar objects and 2) no label trials where they were asked to "pick one" of two objects – one target and one non-target object. The remaining four trials tested 3) first labels (unfamiliar objects) directly and 4) by exclusion and tested 5) second labels (familiar objects) directly and 6) by exclusion (see Figure 2). The direct learning questions tested children's abilities to learn new words for objects where the new word was either a second label for a familiar object (i.e. ball) or a first label for an unfamiliar object (i.e. pot holder). Learning by exclusion trials were similar to the direct trials except that children were now asked to identify a "toma", a fifth new word that they had not heard in training, such that the unlabeled distractors from training now became the target objects.

Input Task This task consisted of a simple play session in which children and one of their parents (the primary caregiver when possible) played with four separate sets of toys for four minutes each. They played with two sets during one session and the other two sets at a second session, counterbalanced within and across sessions. The four sets included a sea animals set, a construction vehicles set, a fruit and vegetables set, and a kitchen utensils set. Each set consisted of 14 objects including 12 objects from the relevant category (roughly half familiar and half unfamiliar to the 20-month-olds according to MCDI percentages), one thematically related object, and one taxonomically related object. For example, the fruits and

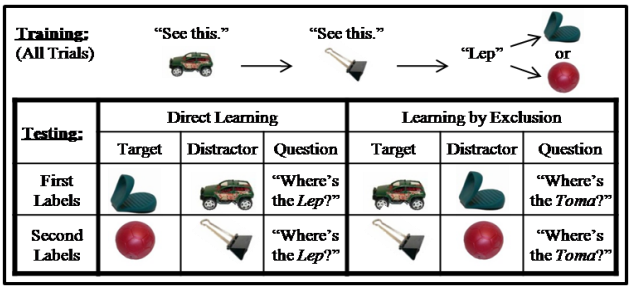


Figure 2. Testing trials design label learning task.

vegetables set included 12 related food toys (an orange, an apple, a banana, asparagus, a bean, an onion, a tomato, a slice of watermelon, a radish, an eggplant, a carrot, and a pepper), one thematic toy (a cutting board), and one taxonomic toy (an egg). Parents were told to sit on a large floor mat and play with their children as if they were at home. Audio and video was recorded.

Procedure

Parent/child dyads attended two sessions within two weeks of each other. Parents completed the MCDI vocabulary checklist on or close to the first session. At each session, the dyads completed the input task first and the label learning task second. Parents did not participate, but were present, for the label learning task.

Coding

Following data collection, all videos of child-parent dyads were reviewed. Coding of the videos consisted of two phases; one to identify instances of labeling and one to identify the types of relations used between first and second labels.

Label Identification First, all instances of the parent labeling an object were identified. Children's labeling instances were not considered. Specifically, the number of times that parents labeled the object was recorded separately for each different label used for each object. For example, a parent might label the orange object "orange" twice and "ball" three times. Four separate measures were calculated using this information including: 1) the proportion of objects that were labeled, 2) the proportion of labeled objects that were given two or more labels, 3) the number of times an object was labeled, and 4) the proportion of labeling instances that were applied to objects given two or more labels. Each measure was calculated separately for familiar and unfamiliar objects, giving us a total of eight input measures. Familiarity was determined separately for each child based on parent report.

Second Label Relations After all instances of second labeling had been identified, the videos were reviewed a second time and the relationship between each label pair was coded into one of eight categories: 1) no relation or labels separated in time (NR), 2) parent indicated that one label was "not" correct (NT), 3) parents stated that one label was not correct but that the object looked like another object (LK), 4) parent used one label as a proper name and one as a common name (PP), 5) parent used one label as a shortened version of the other (e.g. "crab" and "crabby") (SV), 6) parent stated that an object could be named using one label "or" another label (OR), 7) parent indicated that they didn't know which of two labels were correct (DK), and 8) parent stated that an object could be named using one label "and" another label (AND).

Results

Vocabulary

Both children's total vocabulary and their knowledge of the 56 items in the play sets was assessed using parental report. Overall, although children's total vocabulary scores did increase with age, $F(4,115)=92.94$, $p<.001$, their average vocabulary percentile rank did not, $F(4,115)=1.21$, $p=.31$. On average, children knew 23.65 (SD=12.42) of the 56 test objects. This average increased with age, $F(4,115)=25.24$, $p<.001$.

Label Learning Task

For each of the six trial types, the average number of times that each child chose the target object was recorded. The known and no label trials were analyzed separately from the four label learning trials.

Known and No Label Trials An analysis of the known label trials showed that overall children were able to correctly identify the known objects above chance, $t(113)=21.45$, $p<.001$, with older children doing better, $F(4,109)=11.94$, $p<.001$. In addition, all five age groups separately identified the target object greater than chance, all p's$<.05$. A similar analysis of the no label trials showed that, overall, children continued to choose the target object greater than chance, $t(119)=3.44$, $p<.01$. This did not vary by age, $F(4,115)=.80$, $p=.53$.

Label Learning Trials An initial 2 (learning type: direct or by exclusion) x 2 (label type: first or second labels) x 5 (age group) was conducted (see Table 1 for means and comparisons to chance). Results showed a main effect of age, $F(1,115)=25.52$,$p<.001$, such that older age groups learned labels more easily. A main effect of label type, $F(4,115)=16.05$,$p<.001$, showed that children learned first labels better than second labels overall. This interacted with learning type, $F(4,115)=4.81$,$p<.01$, such that this difference was greater when children had to learn labels by exclusion rather than by direct means. Finally, a significant 3-way interaction suggested that the greater difference between first and second label learning for exclusion than by direct

Table 1. Average percent of children correctly identifying the target object label learning task compared to chance.

	Direct Learning		Learning by Exclusion	
Age	1st Label	2nd Label	1st Label	2nd Label
12	.60† (.29)	.54 (.20)	.69* (.18)	.50 (.26)
16	.51 (.25)	.61* (.27)	.64* (.24)	.42† (.22)
20	.63* (.30)	.44 (.27)	.64** (.23)	.61* (.23)
24	.60 (.27)	.60* (.24)	.78** (.26)	.48 (.21)
28	.72** (.28)	.74** (.23)	.71** (.20)	.50 (.26)
All	.61** (.28)	.59** (.26)	.69** (.23)	.50 (.24

† p<.1, *p<.05, **p<.01

means was more pronounced for older than younger kids and somewhat reversed for 20-month-olds. No other main effects or interactions were significant.

Correlations Correlation analyses showed a significant correlation between learning a first label directly and age, r(120)=.22, p<.05, and a marginal correlation between learning a second label directly and age, r(120)=.16, p=.08. No correlations for leaning by exclusion were found.

Input Task

A series of one-way ANOVAs with age group as a between-subjects factor were conducted on each of the eight input measures. Results showed that all eight measures changed with age, all *p*'s<.05, with the exception of the proportion of unfamiliar objects labeled (See table 2 for means and SDs), *p*>*.05*. A series of correlations were also computed for each of the eight measures with age and vocabulary. Age was correlated with all eight measure, all *p*'s<.05, with the exception of the proportion of unfamiliar objects labeled, *p*>.05. Vocabulary was correlated with all measures except for the proportion of familiar and unfamiliar objects labeled, *p*>.05.

Generally speaking, the percent of familiar objects named, but not unfamiliar objects named, increased with age. The percent of both familiar and unfamiliar labeled objects that were given two or more labels also increased with age. The total number of labels used for familiar objects increased with age, whereas the total number of labels used for unfamiliar objects decreased. Finally, the percent of labeling instances that were second labels increased for both familiar and unfamiliar objects.

Factor Analysis on Relations between Labels A series of one-way ANOVAs with age were also conducted on the percentage of each of the eight label relations (NT, NR, etc…) used of the total relations used per participant. None of these types of relations changed with age, all *F*<.01, with the exception of the PP code, which decreased with age, F(4,115)=3.92, p<.01. Only one relationship type, NT, was correlated with age and vocabulary, r(120)=.19, *p*<.05 and r(120)=.18, *p*<.05, respectively.

Because it was likely that the seven codes in which parents provided relations for two or more labels (all but the NR code) were heavily interrelated, a factor analysis was conducted using PCA (principal components analysis) to look for relation types that loaded onto similar factors or components. The factor analysis passed several common criteria for use. First, with over 17 cases per factor, the factor analysis was reliable. Second, the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy was .53, above the cutoff of .5. Third, Bartlett's test of sphericity was significant, $\chi^2(21)=46.28$, $p<.01$. Finally, the diagonals of the anti-image matrix and the commonalities between the relations were all at or above .5.

The principal components analysis produced three components with Eigen values above 1.0 that were retained in the model. The first component explained 23.79% of the variance, the second 17.30%, and the third 14.77%, for a total variance explained of 55.85%. Four other components had Eigen values less than 1.0 and were excluded from the model. Rotation of the solution was utilized to facilitate interpretation of the three components. For this rotation, the verimax solution was used, though no difference in interpretation was obtained using an oblimin solution. None of the relation types were eliminated from the analysis as all seven had loadings or cross-loadings of .6 or higher on one of the three components (see Table 3).

Upon inspection, using values at or greater than .6, it was clear that the first component (henceforth called Contrast Relations) represented greater use of the NT and LK relations as opposed to the AND relation. A higher score on this component is consistent with input that rejects second labels while a lower score is consistent with acceptance of second labels. The DK and OR codes loaded onto the second component (called Ambivalent Relations) with a higher score indicating a greater use of relations that are ambivalent towards rejecting or accepting second labels. The PP and SV codes loaded onto the third component (called Elaborative Relations) with a higher score indicated a greater use of relations in which one label is an elaboration

Table 3. Factor loading for principal components analysis of multiple-labels relationships.

Code	Contrast	Ambivalent	Elaborative
NT	**.72***	.06	-.01
LK	**.60***	-.12	-.27
AND	**-.83***	-.11	-.08
OR	.09	**.77***	.08
DK	-.03	**.75***	-.07
SV	-.11	-.06	**.74***
PP	.02	.06	**.71***

Table 2: Means and standard deviations for input variables. Standard deviations are in parentheses.

	% of Objects Labeled		% of 2+ Object Labels		# of Labeling Instances		% of 2+ Total Labels	
Age	Familiar	Unfamiliar	Familiar	Unfamiliar	Familiar	Unfamiliar	Familiar	Unfamiliar
12	50.8 (28.1)	56.3 (15.9)	14.7 (14.5)	20.2 (11.7)	26.5 (28.6)	58.0 (33.4)	25.6 (27.6)	35.7 (18.8)
16	66.8 (20.9)	54.1 (23.7)	22.9 (21.6)	23.4 (16.2)	40.0 (24.6)	50.5 (31.1)	29.3 (22.7)	41.0 (21.2)
20	72.7 (14.8)	56.9 (13.2)	25.7 (13.2)	30.7 (13.0)	57.2 (28.7)	44.0 (20.4)	38.5 (13.4)	52.9 (18.9)
24	69.7 (19.3)	57.4 (22.3)	29.2 (13.8)	32.4 (19.6)	50.8 (22.6)	33.8 (16.2)	45.7 (18.3)	54.4 (19.1)
28	66.6 (14.0)	55.2 (19.6)	35.9 (15.2)	28.8 (17.8)	63.5 (22.1)	26.8 (16.9)	51.9 (14.2)	48.7 (24.7)

(e.g. longer version) of the other label. It should be noted that adding age and vocabulary to the model did not change the qualitative conclusions except that a fourth component reached an Eigen value above 1.0 on which age and vocabulary, but none of the relations, loaded. Further analyses showed that none of these three components were significantly different by age group, nor were they correlated with age or vocabulary, all *p*'s>.05.

Relationship between Tasks

In order to evaluate any relationship between the input task and language learning task, a series of correlations were computed between the four language learning measures, the eight measures of label use by parents in the input task, and the three components identified for the label type relations.

Label Use and Label Learning The number of times that parents labeled unfamiliar objects was negatively correlated with children's ability to learn first labels directly, r(120)= -.22, p<.05, whereas the ability to learn first labels by exclusion was positively correlated with the proportion of unfamiliar objects given two or more labels, r(120)=.19, *p*<.05.The ability to learn second labels directly was negatively correlated with the number of times that parents labeled unfamiliar objects, r(120)=-.22, *p*<.05. None of the input measures were related to second label learning by exclusion.

If the two second label learning measures are pooled together to get an overall measure of second label learning, there is a significant negative correlation with the number of labels used for unfamiliar objects, r(120)=-.16, *p*<.05, and a positive correlation with the proportion of *familiar* objects given two or more labels, r(120)=.16, *p*<.05.

Multiple-labels Relations and Label Learning Parents use of contrast relations, the first component, was positively correlated with the ability to learn second labels by exclusion, r(120)=.19, *p*<.05, such that the more likely parents were to state that one label was correct and one was not, the more likely children were to learn second labels by exclusion. The second component, ambivalent relations, was negatively correlated with direct learning of second labels, r(120)=-.18, *p*=.05), such that the more ambivalent relations that parents use, the less likely children were to learn second labels directly. Finally, the elaborative relations component was related to the learning of first labels. It was positively correlated with learning by exclusion, r(120)=.15, *p*=.10, and negatively correlated with direct learning, r(120)=-.25, *p*<.01.

Discussion

At the outset of this paper, we asked whether parent use of second labels was related to second label learning. Several interesting relationships between parents' use of second labels and children's learning of first and second labels were found. In particular, parents who labeled unfamiliar objects more had children who were *less* likely to learn first labels directly. This suggests that direct learning of first labels is hindered by parents labeling unfamiliar objects. On the other hand, parents who gave unfamiliar objects two or more labels, had children that learned first labels by exclusion more easily. This suggests that while labeling unfamiliar objects in general disrupts first label learning, if those same unfamiliar objects are given more than one label, it helps children make inferences about first labels. In addition, the more likely parents were to give two or more labels to familiar objects, the easier it was for children to learn second labels (either directly or by exclusion), providing some evidence for a link between amount of second label use by parents and second label learning in children.

Further support for this relationship was found when looking at the types of relations that parents used to connect multiple labels. Parents who use less elaborative relations have children who learn first labels easier when learning is direct, possibly because input is less muddled. However, more elaborative relations are associated with *better* learning *by exclusion*, possibly because they support more complex language relations. Additionally, children learned second labels directly when input relations were less ambivalent. This relationship seems, intuitively, to suggest that using ambivalent relations such as stating that you don't know which label is correct hinders second label learning by direct means. On the other hand, parents who made clear contrasts between the two labels, stating that one of the two labels was correct and the other not, had children who found it easier to learn second labels by exclusion. Though this may seem unintuitive at first glance, it can be explained by thinking of learning labels by exclusion as needing to clearly understand which object should *not* have a new label. Parental input that rejects one label in favor of another helps children do the same when they reject a new label for an already familiar object in favor of an unfamiliar object.

Overall, these results suggest that learning both first and second labels is related to the contrasts that parents make between labels. First labels are easier to learn directly when the input is simple and less ambivalent, but easier to learn by inference with complex input. Second labels are easier to learn either directly or by exclusion when input is heavy on clear, less ambivalent, contrasts between labels.

In addition to the relationship between input and second label learning, we were also able to characterize both the input and the process of second label learning separately. First, in regards to second label learning, children easily learned first labels regardless of whether learning was direct or by exclusion. However, they had a much more difficult time when learning second labels by exclusion than by direct means, and this difference was greater as children got older. In other words, children rejected second labels more as they got older, which is consistent with previous literature (Merriman, Bowman, & MacWhinney, 1989).

Second, we asked whether the amount and type of parental input in regards to second labels varies and whether this was related to age. Parents gave both familiar and unfamiliar objects a higher percentage of second labels as

children got older. In addition, a higher percentage of labels were used as a second label as children got older. More interesting, however, is the finding that, although the percent of second labeling changed with age, the types of relations that parents made between the two labels did not.

Together these results suggest that the manner in which parents label objects, not merely the amount, is related to children's processing of words. In particular, the types of contrasts that parents make can support or hurt children's word learning. However, it is not the case that providing children with more label contrasts will boost their word learning skills. Rather, whether elaborative contrasts with similar words or concrete contrasts of different words are better depends on the status of the word as a first or second label and the task demands.

In sum, not only does input relate to overall language variables such as vocabulary size, but it is also related to the way that children *process* language when presented with a new word. Several contextual influences, including previous experience, task and label type work together to determine children's responses at a given moment. More generally, these results suggest that parental input influences language biases in highly complex ways, something that should be carefully controlled and accounted for in future studies on linguistic biases.

Discussion

Au, T. K., & Glusman, M. (1990). The Principle of Mutual Exclusivity in Word Learning: To Honor or Not to Honor? *Child Development*, *61*(5), 1474 - 1490.

Bloom, P. (2000). *How children learn the meanings of words*. Cambridge, MA: MIT Press.

Callanan, M. A., & Sabbagh, M. A. (2004). Multiple labels for objects in conversations with young children: Parents' language and children's developing expectations about word meanings. *Developmental Psychology*, *40*, 746-763.

Clark, E., & Grossman, J. (1998). Pragmatic directions and children's word learning. *Journal of Child Language*, *25*(01), 1–18.

Davidson, D., & Tell, D. (2005). Monolingual and bilingual children's use of mutual exclusivity in the naming of whole objects. *Journal of experimental child psychology*, *92*(1), 25-45.

Diesendruck, G., & Markson, L. (2001). Children's avoidance of lexical overlap: A pragmatic account. *Developmental Psychology*, *37*(5), 630-641.

Fenson, L., Dale, P., Reznick, J., Bates, E., Thal, D., Pethick, S., et al. (1994). Variability in early communicative development. *Monographs of the society for research in child development*, *59*(5).

Gelman, S. A., & Taylor, M. (1984). How Two-Year-Old Children Interpret Proper and Common Names for Unfamiliar Objects. *Child Development*, *55*(4), 1535 - 1540.

Girolametto, L., Weitzman, E., Wiigs, M., & Pearce, P. S. (1999). The Relationship between Maternal Language Measures and Language Development in Toddlers with Expressive Vocabulary Delays. *American Journal of Speech-Language Pathology*, *8*(4), 364-374.

Hall, D. G., Waxman, S. R., & Hurwitz, W. M. (1993). How two- and four-year-old children interpret adjectives and count nouns. *Child Development*, *64*(6), 1651-1664.

Hansen, M. B., & Markman, E. M. (2009). Children's use of mutual exclusivity to learn labels for parts of objects. *Developmental Psychology*, *45*(2), 592-596.

Hoff, E., & Naigles, L. (2002). How Children Use Input to Acquire a Lexicon. *Child Development*, *73*(2), 418 - 433.

Hollich, G. J., Hirsh-Pasek, K., Golinkoff, R. M., Brand, R. J., Brown, E., Chung, H. L., et al. (2000). Breaking the Language Barrier: An Emergentist Coalition Model for the Origins of Word Learning. *Monographs of the Society for Research in Child Development*, *65*(3).

Huttenlocher, J., Vasilyeva, M., Cymerman, E., & Levine, S. (2002). Language input and child syntax. *Cognitive Psychology*, *45*(3), 337–374.

Liittschwager, J. C., & Markman, E. M. (1994). Sixteen- and 24-month-olds' use of mutual exclusivity as a default assumption in second-label learning. *Developmental Psychology*, *30*(6), 955-968.

Markman, E. M., & Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive psychology*, *20*(2), 121-157.

Markman, E., Wasow, J., & Hansen, M. (2003). Use of the mutual exclusivity assumption by young word learners. *Cognitive Psychology*, *47*(3), 241–275.

Masur, E. F. (1997). Maternal labelling of novel and familiar objects: implications for children's development of lexical constraints. *Journal of child language*, *24*, 427-439.

Merriman, W. E., & Kutlesic, V. (1993). Bilingual and monolingual children's use of two lexical acquisition heuristics. *Applied Psycholinguistics*, *14*, 229-249.

Merriman, W. E., Bowman, L. L., & MacWhinney, B. (1989). The Mutual Exclusivity Bias in Children's Word Learning. *Monographs of the Society for Research in Child Development*, *54*(3).

Mervis, C., Golinkoff, R., & Bertrand, J. (1994). Two-year-olds readily learn multiple labels for the same basic-level category. *Child Development*, *65*(4), 1163–1177.

Saylor, M., Sabbagh, M., & Baldwin, D. (2002). Children use whole-part juxtaposition as a pragmatic cue to word meaning. *Developmental psychology*, *38*(6), 993–1003.

Waxman, S., & Senghas, A. (1992). Relations among word meanings in early lexical development. *Developmental Psychology*, *28*(5), 862–873.

Thinking With Your Body: Modelling Spatial Biases in Categorization Using a Real Humanoid Robot

Anthony F. Morse (anthony.morse@plymouth.ac.uk)
Tony Belpaeme (tony.belpaeme@plymouth.ac.uk)
Angelo Cangelosi (angelo.cangelosi@plymouth.ac.uk)
Centre for Robotics and Neural Systems, University of Plymouth
Plymouth, Devon, PL4 8AA, UK

Linda B. Smith (smith4@indiana.edu)
Cognitive Development Lab, Indiana University, 1101 East Tenth Street
Bloomington, IN 47405-7007

Abstract

This paper presents a model of sensorimotor learning grounded in the sensory streams of a real humanoid robot (the iCub robot). The robot participates in a replication of two developmental psychology experiments, in which it is shown how spatial cues are sufficient for associating linguistic labels with objects. The robot, using auto-associated self-organizing maps connecting is perceptual input and motor control, produces similar performance and results to human participants. This model confirms the validity of a body centric account of the linking of words to objects as sufficient to account for the spatial biases in learning that these experiments expose.

Keywords: Developmental Robotics; Neural Networks; Sensorimotor; Learning; Spatial Bias; Category Learning.

Introduction

At the heart of all sensorimotor theories of cognition is the claim that perception is to a large degree based upon the use of sensorimotor knowledge in predicting the future sensory consequences of an action, either overtly executed or covertly simulated (Gallese & Lakoff, 2005; Morse, Lowe, & Ziemke, 2008; Noë, 2004, 2009; O'Regan & Noë, 2001). As such our perception of continuous contact with a rich visual world laid out in front of us is somewhat misleading, as sensory input is highly impoverished by comparison to perception; for example visual acuity is focused on an area the size of a thumb nail at arm's length. From a sensorimotor perspective, our perception of things outside the fovea is largely constructed from predictions of what you would see were you to look in this or that direction (Noë, 2004). Clearly such perception is supported by processing of the sparse input from the periphery of our visual field, and mechanisms drawing attention to movement, flashes, and other such changes, yet there remains a large disparity between sensory input and perception.

In taking a sensorimotor perspective, the recognition and categorization of objects in our perceptual field can be achieved through the identification of profiles of interaction unique to each object category. As an example we can perceive a plate as round, not because it projects a round image onto our retina, but rather because we can predict how our sensory contact will change as we move a little this way or a little that way. This rather sparse account supposes that such profiles can be constructed and recognized, leading to the recognition of objects in the world in terms of their Gibsonian affordances (Gibson, 1979). This construction of profiles of interaction is crucial to the ability of sensorimotor theories to account for high-level cognitive and mental phenomena such as perception, but is also the least detailed and most challenging aspect of these theories. Few sensorimotor theories do more than just suppose an ability to do this. Nevertheless such embodiment centric accounts of perception are supported by a large number of psychology experiments and neuroscientific evidence exposing various bodily biases in categorization (Richardson & Kirkham, 2004; Smith, 2005; Smith & Samuelson, 2010). For example, for Gallese and Lakoff (2005) the biological sensorimotor system is not merely foundational to our mental conceptual abilities but constitutes action and perception which are inseparably interwoven in those sensorimotor systems. In addition, the re-activation of visual and motor areas during imagined actions (Jeannerod, 1994; Kosslyn & Press, 1994) "shows that typical human cognitive activities such as visual and motor imagery, far from being of a disembodied, modality-free, and symbolic nature, make use of the activation of sensory-motor brain regions." (Gallese & Lakoff, 2005, p. 465). Similarly while paralysis and neuromuscular blockades do not disrupt conscious thought processes (Topulos, Lansing, & Banzett, 1994), the current activity of the motor cortex is highly influential on both perception and thought. Barsalou et al. (2003) highlight some of the ways in which body posture and action affect perception and cognition; for example, subjects rated cartoons differently when holding a pen between their lips than when holding it between their teeth. The latter triggered the same musculature as smiling, which made the subjects rate the cartoons as funnier, whereas holding the pen between the lips activated the same muscles as frowning and consequently had the opposite effect (Strack, Martin, & Stepper, 1988). Moreover, bodily postures influence the subjects' affective state; e.g., subjects in an upright position experience more pride than subjects in a slumped position. Further compatibility between bodily and cognitive states enhances performance. For instance, several motor

performance compatibility effects have been reported in experiments in which subjects responded faster to 'positive' words (e.g. 'love') than 'negative' words (e.g. 'hate') when asked to pull a lever towards them (Chen & Bargh, 1999).

In the remainder of this paper we describe a developmental robotics (Cangelosi & Riga 2006; Weng et al. 2002) model of a simple sensorimotor system grounded in the sensors and actions of iCub, a child-like humanoid robot. The robot then participates in a psychology experiment highlighting the role of body posture and spatial locations in learning the names of objects. Finally we compare the results of the robot experiments to the data from human child psychology experiments conducted by Smith and Samuelson (Smith & Samuelson, 2010).

The 'Modi' Experiment

In a series of experiments related to Piaget's famous A-not-B error (1963), and derived from experiments by Baldwin (1993), Linda Smith and Larissa Samuelson (Smith & Samuelson, 2010) repeatedly showed children between 18 and 24 months of age two different objects in turn, one consistently presented on the left, and the other consistently presented on the right. Following two presentations of each object, the child's attention is drawn to one of the now empty presentation locations and the linguistic label "modi" is presented. Finally the children are presented with both objects in a new location and asked; "can you find me the modi". Not surprisingly the majority (71%) of the children select the *spatially correlated* object despite the fact that the name was presented in the absence of either object. Varying the experiment to draw the child's attention to the left or right rather than to the specific location that the object, when saying "modi", resulted in a similar performance where 68% of the children selected the spatially linked object. The results of this experiment challenge the popular hypothesis that names are linked to the thing being attended to at the time the name is encountered.

In a follow up experiment, using the same basic procedure, one group of children were presented with only a single object labeled while in sight; a second group were repeatedly presented with a consistent spatial relationship until finally an object is labeled while in sight but in the spatial location where the other object was normally presented. In the control group, where a single object is presented and labeled, 80% correctly picked the labeled object over the previously unencountered object; in the second group (spatial competition) a majority of 60% selected the spatially linked object rather than the object that was actually being attended while labeled. In both experiments changes in posture from sitting to standing disrupted the children's ability to link the absent object to the name through space, while other visual or auditory distracters did not. This is strong evidence challenging the simple hypothesis that names are associated to the thing being attended at the time the name is heard, and strong evidence for the role of the body's momentary disposition in space playing a role in binding objects to names through the expected location of that object.

While several other variations of this experiment have been conducted with children, it is these two versions of the experiment that we have replicated with our robot model.

The Robot Experiments

The 'modi' experiments, though not conclusive, strongly suggest that body posture is central to the linking of linguistic and visual information, especially as large changes in posture such as from sitting to standing disrupt the effect reducing performance in the first experiment to chance levels. In our model this suggestion is taken quite literally, using body posture information as a 'hub' connecting information from other sensory streams in ongoing experience. Connecting information via a 'hub' allows for the spreading of activation via this hub to prime information in one modality from information in another. Furthermore using the body posture as a 'hub' also makes a strong connection to the sensorimotor literature reviewed in the introduction; as actions, here interpreted as changes in body posture, also have the ability to directly prime all the information associated with that new position and hence indicate what the agent would expect to see were it to overtly move to that posture. Such predictive abilities are the foundation of sensorimotor theories.

In this experiment we use the humanoid robotic platform iCub, an open source platform which has been recently developed as a benchmark platform for cognitive robotics experiments (Metta et al., 2008). It has 53 degrees of freedom, allowing experiments on visual, tactile and proprioceptive perception, manipulation and crawling. Initial iCub experiments were carried out in simulation through the open source iCub simulator (Tikhanoff et al. 2008), and then adapted and tested on the physical robot platform.

Grounding information in sensory streams

The information linked via the body-posture hub is the result of processing visual input from the iCub robots cameras, taking the average RGB color of the foveal area and using this as an input to a 2D self-organizing map (SOM) (Kohonen, 1998) described in Equation 1, Equation 2, and Equation 3 below. The SOM provides pattern recognition over the input space preserving input topology while capturing the variance of the data. The body-posture 'hub' similarly used the joint angles of the robot as input to another SOM. Though the iCub robot has 53 degrees of freedom, for simplicity in the experiments detailed herein only 2 degrees from the head (up/down and left/right), and 2 degrees from the eyes (up/down and left/right) were actually used, thus the body map of the iCub robot has 4 inputs, each being the angle of a single joint. Further experiments are underway using a more complex body posture map involving all the degrees of freedom of the iCub robot. Finally, auditory input is abstracted as a collection of explicitly represented 'words', each active only while

hearing that word. In the experiments herein these 'words' are artificially activated, though in related work we are using the open source CMU Sphinx library (http://cmusphinx.org/) to provide voice processing, achieving the same result from genuine auditory input.

Both the color map and the body posture map are initialized using random values in the appropriate sensory ranges with an increased probably of values in the extremes of each range until the SOM's have stabilized. Increasing the probability of extreme values ensures that the resulting stable map fully covers the range of possible input values, without this step mid range values would tend to pull in the extremities of the map resulting in poor coverage.

Equation 1: Initial activation of SOM units

$$A_j = \sqrt{\sum_{i=0}^{i=n} \left(v_i - w_{ij}\right)^2}$$

Where A_j is the resulting activity of each node in the map following a forward pass, v_i is an input, and w_{ij} is the weight between that input and the current node. The winning node is the node with the smallest value for A_i

Equation 2: Final activation of SOM units

$$y_i = e^{\left(\frac{-\beta_i}{2\sqrt{n}}\right)}$$

Where y_i is the final activation of the ith node in the map, β is the distance from node i to the winning unit, and n is the total number of nodes in the map. Note: units not within the neighborhood size are set to zero activation, the neighborhood size and learning rate are monotonically decreased and the map is taken to be stable when the neighborhood size is zero.

Equation 3: Weight changes

$$\Delta w_{ij} = \alpha\left(v_i - w_{ij}\right) y_i$$

Where w_{ij} is the weight between input j and unit i, and α is the learning rate.

The neural model forms the upper tier of a 2 layer subsumption architecture (Brooks, 1986) where the lower tier continuously scans whole images for connected regions of change between temporally contiguous images. The robot is directed to orient with fast eye saccades and slower head turns to position the largest region of change (above a threshold) in the centre of the image. This motion saliency mechanism operates independently from the neural model, generating a motion saliency image driving the motor system. This motion saliency image can be replaced with a color-filtered image to provoke orientation to regions of the image best matching the color primed by the neural model.

Using the model described we then replicated the experimental setup used by Smith and Samuelson (2010), linking the activity of the color map and the auditory words to the body map in real time using positive Hebbian connectivity following Equation 4 below.

Equation 4 Positive Hebbian learning

$$\Delta w_{ij} = \alpha x_i x_j$$

Where w_{ij} is the weight between node j and node i, α is the learning rate (0.01), x_i is the activity of the winning node in one map, and x_j is the winning node in the posture map.

These Hebbian associative connections were then only modified from the current active body posture node. Inhibitory competition between any simultaneously active nodes in the same map provides arbitration between multiple associated nodes resulting in dynamics similar to those expressed in Interactive Activation and Competition (IAC) models which have a long history of use in modeling psychological phenomena (Burton, Bruce, & Hancock, 1999; McClelland & Rumelhart, 1981; Morse, 2003).

As the maps are linked together in real time based on the experiences of the robot (see Figure 1), strong connections between objects typically encountered in particular spatial locations, and hence in similar body postures build up. Similarly, when the word 'modi' is heard, it is also associated with the active body posture node at that time. The relative infrequency of activity in the word nodes compared with continuous activity in the color map is not a problem as competition is between nodes within each map and not between the maps themselves. Finally at the end of the experiment, when the robot is asked to 'find the modi', activity in the 'modi' word node spreads to the associated posture and on to the color map node(s) associated with that posture. The result is to prime particular nodes in the color map, the primed color is then used to filter the whole input

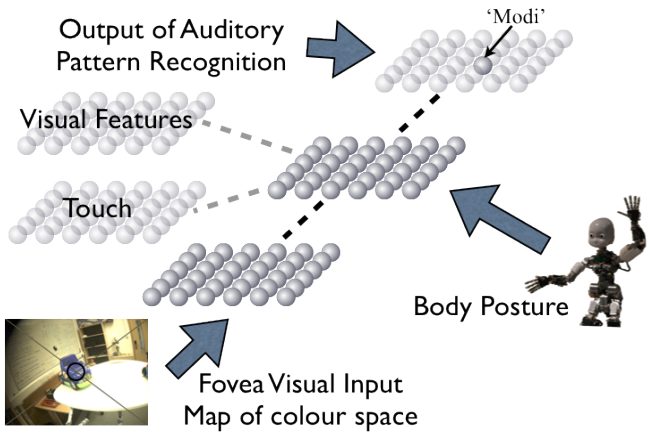


Figure 1: The general architecture of the model. SOMs are used to map the color space, the body posture, and the word space. These maps are then linked using Hebbian learning with the body posture map acting as a central 'hub'. The model can easily be extended to include other features such as visual and touch information in additional SOMs.

image and the robot adjusts its posture to center its vision on the region of the image most closely matching this color. This is achieved using the same mechanism that detects and moves to look at regions of change in the image, replacing the motion saliency image with a color-filtered image. Here the robot moves to look at the brightest region of the color-filtered image, circled in Figure 2 below.

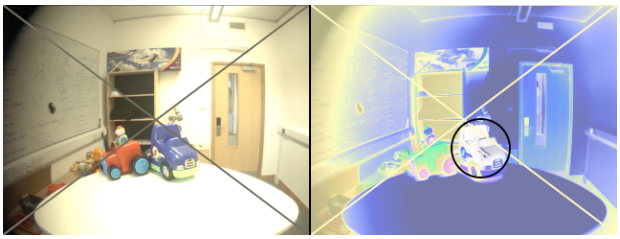


Figure 2 left: Image from the iCub robot's left camera. Right: the same image color-filtered with the primed blue color of the toy truck. The brightest area (circled) indicates the closest match to the primed color.

Given that the number of associations constructed will grow over time in the absence of negative Hebbian learning and in a changing environment, large changes in body posture are used to trigger a removal of these associative connections consistent with the eradication of spatial biases in the psychology experiment following changes from sitting to standing. Additionally, external confirmation that the correct object has been selected leads to more permanent connections being constructed either directly between word and color maps or via a second pattern recognition based 'hub'. As these mechanisms are superfluous to the experiments modeled herein their details have been omitted.

The model as described is then used to replicate each condition of the two psychology experiments described in the previous section as detailed below.

Experiment 1 No Switch Condition

1. Object A is presented to the robot's left – the robot then looks at object A,
2. Object B is presented to the robot's right – the robot then looks at object B,
3. Steps 1 and 2 are repeated,
4. The robot's attention is drawn to its left in the absence of objects A and B and the word 'modi' is spoken,
5. Steps 1 and 2 are repeated again,
6. Object A and object B are presented in a new location and the robot is asked 'where is the modi' – the robot then looks at one of the objects.

This experiment was repeated 18 times resetting the model between each run and starting with a different random seed thereby simulating 18 different individuals. The position of object A and object B (to the left and right) was swapped between each trial and the location that the robots attention was drawn to in step 4 was changed between the first 9 and the remaining trials thereby removing any bias favoring one object or one location over the other. The whole experiment was videoed and stills from steps 1, 2, 4 & 6 are shown in Figure 3. The results recorded which object was centered in the robots visual field following step 6.

Experiment 1 Switch Condition

In the switch condition the location of presentation of objects A and B was swapped for the first presentation only of each object (step 1). Subsequent presentations of each object in steps 2 and 5 remained consistent with the original locations in the no switch condition. Again the experiment was repeated, this time 20 times, with the same variations as used in the no switch condition and the results recoded which object if any is centered in the robots visual field following step 6.

Experiment 2 Labeling while in sight – Control Condition

Experiment 2 provides a variation on experiment 1 in which objects are labeled while in sight. In the control condition a single object is presented either to the left or to the right and labeled 'modi' while being attended, the object is then presented in a new location with a second object and the robot is asked to 'find the modi'.

Experiment 2 Labeling while in sight – Switch Condition

1. Object A is presented to the robots left – the robot then looks at object A
2. Object B is presented to the robots right – the robot then looks at object B
3. Steps 1 and 2 are repeated

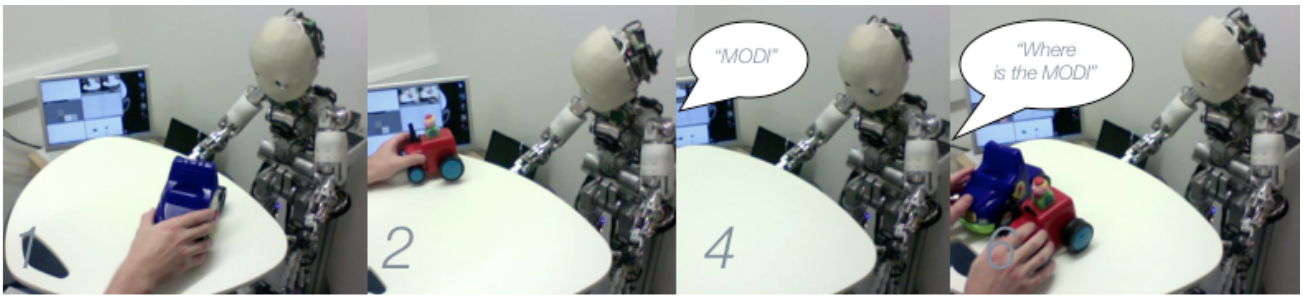


Figure 3: The experiment sequence with the iCub robot.

4. Steps 1 and 2 are repeated again
5. Steps 1 and 2 are repeated yet again
6. Object A is presented to the robots right (i.e. in the wrong location) and the word 'modi' is spoken
7. Steps 1 and 2 are repeated again
8. Object A and object B are presented in a new location and the robot is asked 'where is the modi' – the robot then looks at one of the objects

Experiment 2 was repeated 20 times in each condition with differently seeded networks where the identity of object A and object B was swapped on each consecutive trial and the locations (left and right) were reversed following 10 trials to remove any object or location specific bias.

This model represents preliminary work investigating spatial biases in object categorization. Further work developing and extending this model as a model of sensorimotor learning is currently underway.

Results

In each condition of each experiment, the results recorded which object, if any, was centered in the robots view following the final step of each experiment where the robot was asked to 'find the modi'. In the no-switch condition of experiment 1, 83% (15/18) of the trials resulted in the robot selecting the spatially linked object, while the remaining trials resulted in the robot selecting the non-spatially linked object. This is comparable to the reported result that 71% of children selected the spatially linked object in the human experiment in the same condition (Smith & Samuelson, 2010).

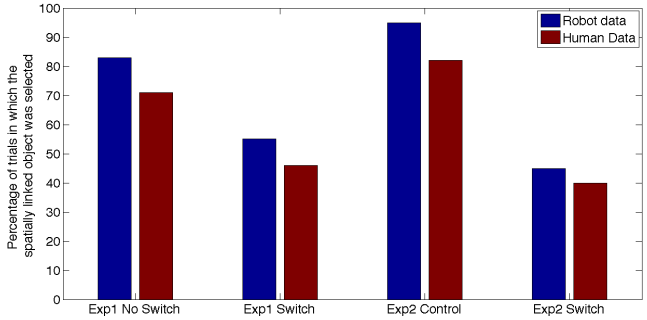


Figure 4: The percentage of spatially linked objects selected in each experimental condition for both robot data and for the human child data.

Reducing the consistency of the object-location correlation in the switch condition resulted in a significant reduction in the spatial priming effect with a close to chance performance of 55% (11/20) of the trials finishing with the spatially correlated object being centered in the view of the robot. The remaining 9 trials resulted in the other object being selected. In experiment 2 objects were labeled while being attended, the control group resulted in 95% (19/20) of the trials selecting the labeled object while in the switch condition only 45% (9/20) of the trials resulted in the labeled object being selected. The remaining trials all selected the other object. These results are compared to the reported human child data in Figure 4.

Discussion and Conclusion

The close match between the results from the robot experiments and the human child results reported by Smith and Samuelson (Smith & Samuelson, 2010) suggests that the hypothesis that body posture is central to early linking of names and object, and can account for the spatial biases exposed by these experiments. What is of relevance here is that the relations between the conditions of each experiment are consistent between the human and robot data, rather than the absolute values achieved. As can be seen from Figure 4 the robot data consistently produced a slightly stronger bias toward the spatially linked objects than the human data.

That the priming effect did not cause the robot to always select the spatially linked object in every variation of the experiments was due to a variety of factors including; noise in the input sensors, varying lighting and reflectance properties as objects are rotated slightly, inaccuracies in the orienting mechanism and so on. In combination these factors produced variations in which a node in the color map was activated as one particular object is being observed, this can lead to weak connections between several similar nodes rather than a single strong connection to one node. In the switch condition of experiment 1, this situation more frequently resulted in object B having a stronger connection to the body posture in which object A was more frequently observed, thus object B was more strongly primed and selected. In these cases increasing the consistency in which an object is seen in the labeled location promotes the strengthening of connections leading to that object being selected, as is seen in the no-switch condition of exp. 1.

It is anticipated that the inclusion of other visual features, though likely to be subject to similar variance, would increase the discrepancy between the data from this model and the human data. This would be due to activation spreading between maps, influencing the priming in much the same way a localist IAC model (Burton et al., 1999; McClelland & Rumelhart, 1981; Morse, 2003). Despite this the relative effects of the various conditions across each experiment should remain relatively consistent. We suggest that the close fit to human data could be misleading, as by comparison in the human case spatial priming would be in competition with far more complex factors influencing the saliency of the objects, factors we have not attempted to model here. Conversely such competition may in fact reduce the models tendency to over perform thereby more closely matching the human data.

As indicated in the introduction our model is consistent with the sensorimotor approach to understanding cognition as the model is able to predict the sensory input it would receive were it to move to different body-postures. This information is accessed simply by a spread of activation from primed body-posture nodes in the 'hub'. The model is also easily scaled up to include additional information presented in additional maps retaining the current IAC-like

architecture. Such models are also suitable for use in hierarchies providing a better fit to the underlying biology.

In conclusion our model accurately reproduces the human data from Smith and Samuelson's (2010) experiments, in an ongoing embodied human robot interaction. In fact, the close fit between our data and the reported human data is in part due to the difficulties and inaccuracies inherent in conducting experiments with complex real robots rather than simulations. In future work we are developing and demonstrating this architecture in a variety of related sensorimotor and psychological tasks involving object manipulations. The goal is close empirical studies of robots and children – in which robot models generate new predictions tested in children. Such joint studies should advance robotics, our understanding of human cognitive development, and the nature of embodied intelligence more generally.

Acknowledgements

This work has been supported by the EU FP7 ITALK project (no. 214668).

References

Baldwin, D.A. (1993) Early referential understanding: infants' ability to recognize referential acts for what the are. *Developmental Psychology, 29,* 832-43.

Barsalou, L. W., Niedenthal, P. M., Barbey, A. K., & Ruppert, J. A. (2003). Social embodiment. *Psychology of Learning and Motivation: Advances in Research and Theory, 43*, 43-92.

Brooks, R. A. (1986). A robust layered control system for a mobile robot. *IEEE Journal of Robotics and Automation, 2*(1), 14-23.

Burton, A. M., Bruce, V., & Hancock, P. J. B. (1999). From pixels to people: A model of familiar face recognition. *Cognitive Science, 23*(1), 1-31.

Cangelosi A., & Riga, T. (2006). An embodied model for sensorimotor grounding and grounding transfer: Experiments with epigenetic robots, Cognitive Science, 30(4), 673-689.

Chen, M., & Bargh, J. A. (1999). Consequences of automatic evaluation: Immediate behavioral predispositions to approach or avoid the stimulus. *Personality and Social Psychology Bulletin, 25*(2), 215.

Gallese, V., & Lakoff, G. (2005). The brain's concepts: The role of the sensory-motor system in reason and language. *Cognitive Neuropsychology, 22*, 455-479.

Gibson, J. J. (1979). *The Ecological Approach to Visual Perception*. Boston: Houghton Mifflin.

Jeannerod, M. (1994). The representing brain: Neural correlates of motor intention and imagery. *Behavioral and Brain Sciences, 17*(2), 187-201.

Kohonen, T. (1998). The self-organizing map. *Neurocomputing, 21*(1-3), 1-6.

Kosslyn, S. M., & Press, M. I. T. (1994). *Image and brain: The resolution of the imagery debate*: MIT Press.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. *Psychological Review, 88*, 375407.

Metta G., Sandini G., Vernon D., Natale L., & Nori F. (2008). The iCub humanoid robot: an open platform for research in embodied cognition. In R. Madhavan & E.R. Messina (Eds.), Proceedings of IEEE Workshop on Performance Metrics for Intelligent Systems Workshop (PerMIS'08). Washington, D.C.

Morse, A. F. (2003). *Autonomous Generation of Burton's IAC Cognitive Models*. Paper presented at the EuroCogSci03, The European Cognitive Science Conference.

Morse, A. F., Lowe, R., & Ziemke, T. (2008). *Towards an Enactive Cognitive Architecture.* Paper presented at the International Conference on Cognitive Systems, Karlsruhe, Germany.

Noë, A. (2004). *Action in Perception*. Cambridge, Mass: MIT Press.

Noë, A. (2009). *Out of our heads*. New York: Hill & Wang.

O'Regan, K., & Noë, A. (2001). A sensorimotor account of visual perception and consciousness. *Behavioral and Brain Sciences, 24*, 939-1011.

Piaget, J. (1963). *The origins of intelligence in children*. New York: Norton.

Richardson, D. C., & Kirkham, N. Z. (2004). Multimodal events and moving locations: Eye movements of adults and 6-month-olds reveal dynamic spatial indexing. *Journal of Experimental Psychology: General, 133*, 46-62.

Smith, L. B. (2005). Cognition as a dynamic system: Principles from embodiment. *Developmental Review, 25*(3-4), 278-298.

Smith, L. B., & Samuelson, L. (2010). Objects in Space and Mind: From Reaching to Words. In K. Mix, L. B. Smith & M. Gasser (Eds.), *Thinking Through Space: Spatial Foundations of Language and Cognition*. Oxford, UK.: Oxford University Press.

Strack, F., Martin, L. L., & Stepper, S. (1988). Inhibiting and facilitating conditions of the human smile: A nonobtrusive test of the facial feedback hypothesis. *Journal of Personality and Social Psych, 54*(5), 768-777.

Tikhanoff V, Cangelosi A., Fitzpatrick P., Metta G., Natale L., Nori F. (2008). An open-source simulator for cognitive robotics research: The prototype of the iCub humanoid robot simulator. In R. Madhavan & E.R. Messina (Rds.), *Proceedings of IEEE Workshop on Performance Metrics for Intelligent Systems Workshop (PerMIS'08)*. Washington, D.C.

Topulos, G. P., Lansing, R. W., & Banzett, R. B. (1994). The Experience of Complete Neuromuscular Blockade in Awake Humans. *Survey of Anesthesiology, 38*(03), 133.

Weng J., McClelland J., Pentland A., Sporns O., Stockman I., Sur M, Thelen E. (2001). Autonomous mental development by robots and animals. Science, 291, 599–600.

Effects of simultaneously presented visual information on adults' and infants' auditory statistical learning

Erik D. Thiessen
Department of Psychology
Carnegie Mellon University
5000 Forbes Ave, Pittsburgh PA 15213

Abstract

Infant and adult learners are able to identify word boundaries in fluent speech using statistical information. Similarly, learners are able to use statistical information to identify word-object associations. Successful language learning requires both feats. In this series of experiments, we presented adults and infants with audio-visual input from which it was possible to identify both word boundaries and word-object relations. Adult learners were able to identify both kinds of statistical relations from the same input. Moreover, their learning was actually facilitated by the presence of two simultaneously present relations. Eight-month-old infants, however, do not appear to benefit from the presence of regular relations between words and object. Adults, like 8-month-olds, did not benefit from regular audio-visual correspondences when they were tested with tones, rather than linguistic input. These differences in learning outcomes across age and input suggest that both developmental and stimulus-based constraints affect statistical learning.

Keywords: statistical learning, cross-modal stimuli, development of cross-modal integration

Introduction

Learners are able to identify many different kinds of statistical regularities from linguistic input, including phonological and syntactic patterns (e.g., Chambers, Onishi, & Fisher, 2003; Mintz, 2002; Thiessen & Saffran, 2003). Despite the power of statistical learning, though, there is little doubt that human learners are constrained. Learners do not identify all kinds of statistical patterns equally well (e.g., Newport & Aslin, 2004; Peperkamp, le Calvez, Nadal, & Dupoux, 2006; Redford, 2008; Saffran & Thiessen, 2003). However, most of the research into constraints on human learning has focused on how learners do when presented with a single learning task. This is insufficient for a complete understanding of statistical learning for two reasons. First, language frequently presents learners with multiple problems simultaneously. For example, when exposed to a novel word form in fluent speech, learners have the opportunity to both learn the word form, and to learn the referent of the word. Second, constraints on learning may be especially important when the input is complex enough to support multiple learning problems (e.g., Fiser & Aslin, 2002; Pinker, 1984).

Consider the interaction between the statistical information useful for segmenting words from fluent speech (e.g., Saffran, Aslin, & Newport, 1996), and identifying referents for words (e.g., Smith & Yu, 2008). Taken in isolation, both word segmentation (e.g., Thiessen, Hill, & Saffran, 2005; Toro, Sinnett, & Soto-Faraco, 2005) and referential learning are constrained (Golinkoff, Shuff-Bailey, Olguin, & Ruan, 1995; Landau, Smith, & Jones, 1988; Markman, 1990; Markman & Wachtel, 1988). It is also clear that these learning processes interact. Learners who are previously familiar with a word form map it more easily to a novel referent (e.g., Graf Estes, Evans, Alibali, and Saffran, 2007; Storkel et al., 2001). Conversely, children map familiar objects to novel labels more easily than unfamiliar objects (e.g., Hall, 1991). Because these learning tasks interact, different constraints may operate when learners are presented with both problems simultaneously. If the interaction between the problems is unconstrained, the additional complexity when they are presented together may hinder learning (Fiser & Aslin, 2002; Pinker, 1984). Alternatively, learning may be constrained in such a way that the learning occurs sequentially, with one problem privileged and learned first. It is even possible that learning, if appropriately constrained, could be facilitated by the simultaneous presentation of multiple regularities. This could be the case if learning of one regularity reinforces the other.

To explore these possibilities, it is critical to present learners with the opportunity to identify simultaneous regularities. This set of experiments did so by building on prior research demonstrating that learners benefit from the embedding of audio input in a visual context (e.g., Hollich, Newman & Jusczyk, 2005). Appropriate visual information helps learners determine whether speakers are producing one language or multiple languages (Soto-Farco et al., 2007). Similarly, the presence of a video improves adults' ability to identify word boundaries in fluent speech (Sell & Kaschak, under review). In all of these tasks, however, the auditory learning task is the only task, and vision facilitates that task. The current experiments differ by presenting learners with two problems simultaneously: word segmentation, and discovery of word-object relations. This better simulates the richness of language, where any single utterance may provide information about many different aspects of language (e.g., Saffran & Wilson, 2001).

Experiment 1

All learners in this experiment were presented with words embedded in fluent speech. As in previous statistical learning experiments (e.g., Saffran et al., 1996), these words could be segmented via use of transitional probabilities that were high within words, and low at word boundaries. A subset of the participants in this experiment (in the *no-video* condition) were presented solely with fluent speech. The only learning task this group faced was identifying word boundaries.

A second group of participants (in the *regular-video* condition) saw objects synchronized to the onset and offset of the words in the fluent speech. Each word in the fluent speech was consistently paired with a unique object. As such, this group of participants was presented with two potential statistical regularities to learn: word boundaries, and the relations between particular words and objects.

A third group of participants (in the *irregular-video* condition) also saw shapes synchronized to words in fluent speech, but these participants saw objects that were not consistently associated with the words. This condition serves as a control to make sure that performance in the regular-video condition is not affected by some aspect of the visual stimuli other than the regular relation between words and shapes.

Method

Participants

Participants were 60 undergraduates at Carnegie Mellon University. Twenty participants apiece were randomly assigned to one of three stimulus conditions: no-video, irregular-video, or regular-video.

Stimuli

Audio Stimuli

All participants were exposed to a stream of synthesized speech used in Saffran et al.'s (1996) experiments. This artificial language contained four words: *padoti*, *bidaku*, *tupiro*, and *golabu*. The transitional probabilities between syllables within a word were 1.0, and the transitional probabilities between syllables across word boundaries were .33. Two words (*bidaku* and *tupiro*) and two part-words (*tigola* and *bupado*) were used as test items. Unlike words, part-word test items contained a transition between syllables with low transitional probability.

Visual Stimuli

In the *no-video* condition, participants saw a static checkerboard image for the duration of their exposure to the synthesized speech.

Participants in the *regular-video* and *irregular-video* condition saw looming shapes synchronized with the word boundaries. Shapes appeared at the same instant the word began to play, and remained onscreen for the duration of the word. At the beginning of a word, each shape occupied roughly $1/16^{th}$ of the screen. Over the course of the presentation of the word, the shape increased in size until it filled the screen.

In the regular-video condition, each word was paired with a particular object (*padoti*: white cross; *bidaku*: green diamond; *tupiro*: purple heart; *golabu*: yellow hexagon). In the irregular-video condition, words and shapes co-occurred with no consistent pattern. *Procedure*

In all three conditions, participants sat in front of a portable DVD player with a 10'' screen wearing airline-pilot style headphones. Participants were simply informed that after watching the video, they would answer a series of questions about what they saw and heard.

Segmentation Test

There were 16 two alternative forced choice questions in the segmentation test. For each question, participants heard a word and a part word (in counterbalanced order), separated by one second of silence. They were asked to circle the item that sounded more like the speech they heard (for discussion of this procedure, see Saffran et al., 1997).

Word-Shape Correspondence test

After completion of the 16 segmentation test items, participants in only the regular-video condition were informed that they would now answer an additional series of 16 questions. These questions assessed whether participants learned that particular words corresponded to shapes. For each question, participants heard one of the four words from the synthesized speech. They then saw a sequence of four shapes on the screen, looming with the same animation as during the initial exposure. They were asked to circle which of the four shapes went with the word.

Results

A one-way ANOVA was performed on participants' scores on the word-segmentation test as a function of condition. There was a significant effect of condition, $F(2,57) = 5.4$, $p < .01$. Participants performed best in the regular-video condition ($M = 12.0$, $SE = 0.5$), and less well in the irregular-video ($M = 9.9$, $SE = 0.5$) and no-video condition ($M = 8.9$, $SE = 0.5$). Scores in all three condition differed from chance (all condition: binomial $p < .05$). To follow up the effect of condition indicated by the ANOVA, planned t-tests were performed. Here and elsewhere, all t-tests reported are two-tailed. There was no significant difference between participants' performance in the no-video and in the irregular-video condition: $t(38) = 1.1$, $p = .30$. However, participants in the regular condition scored significantly better than participants in either of the other two conditions (regular- vs. no-video: $t(38) = 3.4$, $p < .01$; regular- vs. irregular-video: $t(38) = 2.2$, $p < .05$).

Participants in the regular-video condition also learned word-object relations. On average, participants scored 8.8 (out of 16; chance = 4) correct on the correspondence test ($SE = 0.8$), which was significantly above chance, binomial $p < .01$. Further, as illustrated by Figure 2, the correlation between the two tests was positive, $r = .64$, and significant, $p < .01$. Higher scores on one test were associated with higher scores on the other. Results from the segmentation and correspondence test converge to indicate that the presence of regular word-object relations facilitated learning

Experiment 2

Prior experiments have demonstrated that infants are able to segment words from fluent speech via transitional probabilities (e.g., Saffran et al., 1996), and identify relations between words and shapes (e.g., Thiessen, 2007), but no experiments have assessed both simultaneously. Because infants are the primary learners of language, their performance is both theoretically and pragmatically important. For example, given the capacity limitations of infants, it is plausible to hypothesize that they would fail to integrate audio and visual information as effectively as adults. If so, they may not benefit from the audio-visual corresponded in the regular-video condition.

Method

Participants

Participants in this experiment were 60 infants between the ages of 7.5 and 9 months (M = 8.26). Infants were randomly assigned to one of three groups: no-video, regular-video, and irregular-video. In order to obtain data from 60 infants, it was necessary to test 66. The additional six infants were excluded for the following reasons: fussing or crying (3), parental interference (2), and experimenter error (1). According to parental report, all infants were full term, and free of ear infections at the time of testing.

Procedure

This experiment used a slightly modified version of the HPP, presenting the visual stimuli on a central monitor rather than from the side of the room. Preferential looking experiments with a central monitor are commonly and successfully used with infants (e.g., Fernald, 1985). Infant participants were seated on their parents' lap in a sound-isolated room, approximately one foot away from a 30'' monitor. There were two speakers adjacent to the monitor and a camera mounted above it. The parents wore noise-canceling headphones to eliminate bias. An experimenter outside the room watched the infant over a closed-circuit monitor to initiate test trials and code the direction of the infants' gaze.

There were two phases to this experiment: the segmentation phase, and the test phase. During the segmentation phase, infants heard the synthesized speech from speakers adjacent to the monitor, while the monitor displayed the visual stimuli appropriate to the infants' condition.

The test phase used the same two words and two part-words as the adult test. Each item was repeated 3 times, for a total of 12 trials. Before each trial, an attention-getter (a brightly colored Winnie the Pooh video, coupled with an excited exclamation) attracted infants' gaze to the monitor. Once the infant oriented to the monitor, the experimenter initiated the test trial. Each trial consisted of a repetition of a single word (or part-word), with a pause of 1 second between repetitions. For as long as infants' gazed at the monitor, the test item continued to repeat. When infants looked away from the monitor for two continuous seconds, the test trial ended.

Stimuli

The stimuli during the segmentation phase were presented for 50 seconds and were identical to the audio and video presentations used in Experiment 1. This exposure is half of the length in Experiment 1; pilot testing indicated that 100 seconds yielded an unacceptably high fuss-out rate. The test items were also identical to Experiment 1. During test phase, words and part-words were paired with an orange bar rotating like a propeller (it completed one revolution every three seconds). Pilot testing indicated that infants were far more likely to maintain their interest in the experiment if the monitor displayed a moving object rather than a static image. Both the color and the shape of the bar were novel with respect to the segmentation phase of the experiment, and the motion was unlike the looming animation infants saw during the segmentation phase.

Results

Infants in the no-video condition looked at word trials for 12.4 sec (SE = 1.0), and at part-word trials for 11.8 sec (SE = 0.9). This difference in looking trials between words and part-words was not significant, $t(19) = 1.1$, $p = .28$. Infants in the regular-video condition looked at word test trials for 9.7 sec (SE = 0.7), and to part-word test trials for 11.7 sec (SE = 0.8). This difference was significant, $t(19) = 3.7$, $p < .05$. Infants in the irregular-video condition showed the same pattern, looking at words (M = 9.9, SE = 1.2) less than part-words (M = 11.6, SE = 1.1). The difference in looking time to words and part-words was also significant for infants in this group: $t(19) = 2.5$, $p < .05$. Unlike infants in the no-video condition, infants in both the regular- and irregular-video condition listened longer to part-words than to words. This indicates that they had learned enough about the identity of words to distinguish them from part-words.

Infants appeared to perform better in the regular-video condition than in the no-video condition, as infants in the no-video condition failed to respond differentially to word and part-word trials. However, infants' performance in the regular- and irregular-video conditions was not significantly different, as indicated by a 2 (condition) x 2 (test item) ANOVA. As expected, since infants in both groups showed a preference for part-words, there was a main effect of test item: $F(1, 38) = 12.6$, $p < .01$. There was no main effect of condition: $F(1, 38) < 1$. There was also no interaction between test item and condition: $F(1, 38) < 1$. That is, infants' preference for part-words in the irregular-video condition was statistically equivalent to infants' preference in the regular-video condition. These analyses indicate that while infants may benefit from the presence of looming shapes synchronized with word boundaries (present in both video conditions), they do not gain an added advantage from the regular relations between words and shapes present in the regular-video condition. These results present two compelling questions, discussed separately below.

Why do infants fail to distinguish between words and part-words in the no-video condition, when they can do so in the regular- and irregular-video condition?

One possible explanation is that infants in the regular- and irregular-video conditions received some benefit not present in the no-video condition. The looming shapes may have facilitated learning by maintaining infants' attention (e.g., Frick & Richards, 2001; Thiessen, et al., 2005). Another possible benefit that infants may have received in both the regular- and irregular-video condition is the synchronization between the appearance of the shapes and word boundaries. For young infants, synchronization is one of the most important factors that enable identifying links between audio and visual events (e.g., Bahrick, Flom, & Lickliter, 2002; Gogate & Bahrick, 1998; Lewcowicz, 1986; 2003). Infants may have relied upon synchronization as a cue to word boundaries, a cue that was equally available in both the regular- and irregular-video conditions.

Why do infants, unlike adults, fail to benefit from the regular relations between words and shapes available in the regular-video condition?

The fact that infants' performance in the irregular-video condition is equivalent to their performance in the regular-video condition suggests that infants failed to detect the relations between words and shapes present in the regular-video condition. This suggestion is consistent with a variety of converging evidence indicating that infants at this age are relatively insensitive to relations between words and objects in the visual world. Eight-month-old infants have a small vocabulary (e.g., Fenson et al., 2002). In controlled word-learning experiments, infants typically fail to acquire names for novel objects until around a year of age (Werker, Cohen, Lloyd, Casasola, & Stager, 1998). If infants cannot detect the relation between words and objects in the regular-video condition, they cannot benefit from any facilitation that identifying the relation provides to adult learners.

Experiment 3

There are at least two (not mutually exclusive) factors that can explain why infants in Experiment 2 failed to benefit from the regular audio-visual pairing, unlike adults in Experiment 1. One is that adults' ability to take advantage of the regular-video condition is due to the fact that they are faster, more efficient information processors than infants (e.g., Pelphrey & Reznick, 2003). To detect a relation between words and shapes, learners must process the identity of the shape (and the word) in a brief time. There are several experimental results suggesting that young infants are less successful in processing multiple sources of information than older infants and adults (e.g., Stager & Werker, 1997). A second potential factor is that the difference between 8-month-olds and older learners is due to differences in their prior linguistic experience. Adults are well aware that one of the primary functions of words is to refer to features of the visual world such as shape. Eight-month-olds may not yet expect to discover relations between words and shapes (cf. Werker et al., 1998). Infants may fail to detect the regular relations in the input because they do not expect them.

Both factors converge to suggest that older infants should be more successful in identifying and benefiting from regular word-object associations. Thus, in Experiment 3, we presented 20-month-old infants with the same stimuli used in Experiment 2. These children have a year of additional word-learning experience, and more advanced cognitive processing abilities. Should infants of this age fail to benefit from the regular-video condition, it may suggest that the infant paradigm is simply insensitive to infants' abilities to benefit from word-object relations. However, should infants benefit from regular word-object relations in Experiment 3, it will indicate important developmental differences in infants' abilities to integrate audio and visual information in a statistical learning task.

Method

Participants

Participants were 45 infants between the ages of 19.5 and 20.5 months (M = 20.12). Infants were randomly assigned to one of three groups: no-video, regular-video, and irregular-video. To obtain data from 45 infants, it was necessary to test 63. The additional 18 infants were excluded for the following reasons: fussing or crying (16), test trial looking times averaging less than 3 seconds (1), or parental interference (1). According to parental report, all infants were full term, and free of ear infections at the time of testing.

Stimuli

The stimuli were identical to those in Experiment 2.

Procedure

The procedure was identical to that of Experiment 2.

Results

Infants in the no-video condition looked at word trials for 8.3 sec (SE = 0.6), and at part-word trials for 8.2 sec (SE = 0.5). This difference in looking trials between words and part-words was not significant, $t(14) < 1$. Like the younger infants in Experiment 2, 20-month-olds in the no-video condition failed to distinguish between words and part-words, showing no evidence of learning. At neither age should this be taken as evidence that infants are unable to learn from audio stimuli alone – prior experiments clearly demonstrate that infants are able to do so (e.g., Saffran et al., 1996). Infants' failure in the current experiments is due to the fact that the stimuli are presented much more briefly than in prior experiments. While infants can learn from stimuli presented for this duration, they may only do so for natural – as opposed to synthesized – speech (e.g., Thiessen et al., 2005).

Infants in the irregular-video condition also showed no significant preference, looking equivalently long at word trials (M = 9.6, SE = 0.7) and part-word trials (M = 9.1, SE = 0.6), $t(14) < 1$. Interestingly, unlike the 8-month-olds in Experiment 2, 20-month-olds did not appear to learn from the irregular-video condition. Note that 8-month-olds' looking times were much longer to both kinds of test trials (M = 11.1 sec) than that of the 20-month-olds. This may indicate that the testing situation was more interesting to 8-month-olds than 20-month-olds. Sustained attention to the input facilitates statistical learning (e.g., Toro et al., 2005). The 20-month-olds in the current experiment may simply have failed to attend to the stimuli long enough to learn.

Regardless of infants' performance in the other two conditions, the question that motivated this experiment was whether they are facilitated in learning from the regular-video stimuli. Infants in the regular-video condition looked at word test trials for 7.2 sec (SE = 0.6), and to part-word test trials for 8.5 sec (SE = 0.7). This difference was significant, $t(14) = 2.3$, $p < .05$. Only infants in the regular-video condition showed evidence of learning; no other group demonstrated the ability to distinguish between words and part-words. A series of planned 2 x 2 ANOVAs comparing looking times across conditions assessed this more rigorously. In none of the ANOVAs was there a significant main effect of condition, nor of test trial (all Fs < 1). Similarly, there was no interaction between condition and test trial when comparing participants in the no-video condition to participants in the irregular-video condition ($F < 1$).

Most importantly, though, there were significant interactions between condition and looking time when comparing participants in the regular-video condition to participants in both the no-video condition [$F(1, 28) = 3.2$, $p < .05$] and the irregular-video condition [$F(1, 28) = 3.4$, $p < .05$]. These interactions indicate that infants' preference in the regular-video condition was significantly different from their lack of preference in either of the other two conditions. This confirms that 20-month-olds, like adults, performed significantly better in the regular-video condition than either of the other two conditions. For children at this age, complexity can facilitate learning by providing multiple learnable regularities in the input. This suggests an important developmental difference between 8- and 20-months of age, with only 20-month-olds showing the ability to benefit from regular audio-video relations in a manner comparable to adults.

General Discussion

One of the reasons that language is such an effective communicative tool is that it allows speakers to express multiple pieces of information simultaneously. For example, a simple observation about the state of the world, such as "the Pirates won," is coupled with affective information that indicates how the speaker feels about that state of affairs. This means that language is rich in possible relations for learners to discover, both between aspects of the speech signal (such as words and pitch), and between speech and meaning. Indeed, infants are able to detect many of these possible relations (e.g., Fisher, Klinger, & Song, 2006; Saffran et al., 1996; Smith & Yu, 2008; Thiessen & Saffran, 2007). However, the need for constraints is necessary for learners presented with rich input, especially input in which multiple relations are present simultaneously.

The current results indicate that at least one of those constraints is a developmental constraint. The ability to integrate simultaneous audio and visual information in a statistical learning task develops during the first two years of life. Whereas adults benefit from the presence of regular word-object associations, 8-month-old infants do not. Critically, 20-month-olds, like adults, benefit from the regular relations between words and visual objects available in these stimuli. One possibility for this developmental difference relates to older children's vastly greater experience with word-object correspondences in language. Ongoing work with non-linguistic stimuli will assess this possibility. Though the current data do not differentiate between domain-specific maturational accounts (e.g., Waxman & Booth, 2000) and accounts that implicate more general processes, both kinds of accounts share an important commonality. On either account, young infants are not learning as much as adults are when presented with stimuli in which words and objects co-occur. This may actually be beneficial for young learners. By preferentially detecting only some of the available relations in the stimuli, infant learners may avoid a combinatorial explosion (e.g., Newport, 1990).

References

Bahrick, L.E., Flom, R., & Lickliter, R. (2002). Intersensory redundancy facilitates discrimination of tempo in 3-month-old infants. *Developmental Psychobiology, 41*, 352-363.

Chambers, K.E., Onishi, K.H., & Fisher, C. (2003). Infants learn phonotactic regularities from brief auditory experiences. *Cognition,* **87**, B69-B77.

Creel, S.C., Newport, E.L., & Aslin, R.N. (2004). Distant Melodies: Statistical learning of non-adjacent dependencies in tone sequences. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 30*, 1119-1130.

Elman, J.L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition, 48*, 71-99.

Fenson, F., Dale, P.S., Reznick, J.S., Thal, D., Bates, E., Hartung, J.P., Pethick, S., & Reilly, J.S. (2002). *MacArthur Communicative Development Inventories: User's Guide and Technical Manual.* Baltimore: Paul H. Brookes.

Fernald, A. (1985). Four-month-old infants prefer to listen to motherese. *Infant Behavior and Development, 8*, 181-195.

Fisher, C., Klingler, S.L., & Song, H. (2006). What does syntax say about space? 2-year-olds use sentence structure to learn new prepositions. *Cognition, 101*, B19-B29.

Frick, J.E., & Richards, J.E. (2001). Individual differences in infants' recognition of briefly presented visual stimuli. *Infancy, 2*, 331-352.

Gold, M.E., (1967). Language identification in the limit. *Information and Control, 10*, 447-474.

Golinkoff, R.M., Shuff-Bailey, M., Olguin, R., & Ruan, W. (1995). Young children extend novel words at the basic level: Evidence for the principle of categorical scope. *Developmental Psychology, 31*, 494-507.

Graf Estes, K.M., Evans, J.L., Alibali, M.W., & Saffran, J.R. (2007). Can infants map meaning to newly segmented words?: Statistical segmentation and word learning. *Psychological Science, 18*, 254-260.

Hall, G.D. (1991). Acquiring proper nouns for familiar and unfamiliar objects: Two-year-olds' word-learning biases. *Child Development, 62*, 1142-1154.

Hayes, J.R., & Clark, H.H. (1970). Experiments on the segmentation of an artificial speech analogue. In J. Hayes (Ed.) *Cognition and the Development of Language*, pp. 221-234. New York: John Wiley and Sons.

Hollich, G., Newman, R.S., & Jusczyk, P.W. (2005). Infants' use of synchronized visual information to separate streams of speech. *Child Development, 76*, 598-613.

Hunter, M. A., & Ames, E. W. (1988). A multifactor model of infant preferences for novel and familiar stimuli. *Advances in infancy research, 5*, 69-95.

Just, M.A., & Carpenter, P.A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review, 99*, 122-149.

Landau, B., Smith, L.B., & Jones, S.S. (1988). The importance of shape in early lexical learning. *Developmental Psychology, 28*, 273-286.

Lewkowicz, D.J. (1986). Developmental changes in infants' bisensory response to synchronous durations. *Infant Behavior and Development, 9*, 335-353.

Lewkowicz, D.J. (2003). Learning and discrimination of audiovisual events in human infants: The hierarchical relation between intersensory synchrony and rhythmic pattern cues. *Developmental Psychology, 39*, 795-804.

Marcus, G.F., Vijayan, S., Bandi Rao, S., & Vishton, P.M. (1999). Rule learning in 7-month-old infants. *Science, 283*, 77-80.

Markman, E.M. (1990). Constraints children place on word meanings. *Cognitive Science, 14*, 57-77.

Markman, E.M., & Wachtel, G.A. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology, 20*, 121-157.

Mattys, S.L., Jusczyk, P.W., Luce, P.A., & Morgan, J.L. (1999). Phonotactic and prosodic effects on word segmentation in infants. *Cognitive Psychology, 38*, 465-494.

Mehler, J., Peña, M., Nespor, M., & Bonatti, L. (2006). The "soul" of language does not use statistics: Reflections on vowels and consonants. *Cortex, 42*, 846-854.

Mintz, T.H. (2002). Category induction from distributional cues in an artificial language. *Memory and Cognition,* **30**, 678-686.

Neil, P.A., Chee-Ruiter, C., Scheier, C., Lewcowicz, D.J., & Shimojo, S. (2006). Development of multisensory spatial integration and perception in humans. *Developmental Science, 9*, 454-464.

Newport, E.L. (1990). Maturational constraints on language learning. *Cognitive Science, 14*, 11-28.

Pelphrey, K.A., & Reznick, J.S. (2003). Working memory in infancy. In R. Kail (Ed.), *Advances in Child Development and Behavior, 31*, pp. 173-227.

Pinker, S. (1984). *Language Learnability and Language Development*. Cambridge, MA: MIT Press.

Quine, W.V.O. (1964). *Word and Object*. Cambridge, MA: MIT Press.

Rohde, D.L., & Plaut, D.C. (1999). Language acquisition in the absence of explicit negative evidence: How important is starting small? *Cognition, 72*, 67-109.

Saffran, J.R. (2003). Statistical language learning: Mechanisms and constraints. *Current Directions in Psychological Science, 12*, 110-114.

Saffran, J.R., Aslin, R.N., & Newport, E.L. (1996). Statistical learning by 8-month-old infants. *Science, 274*, 1926-1928.

Saffran, J.R., Johnson, E.K., Aslin, R.N., & Newport, E.L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition, 70*, 27-52.

Saffran, J.R., & Thiessen, E.D. (2003). Pattern induction by infant language learners. *Developmental Psychology, 39*, 484-494.

Saffran, J.R., Newport, E.L., Aslin, R.N., Tunick, R.A., & Barrueco, S. (1997). Incidental language learning: Listening (and learning) out of the corner of your ear. *Psychological Science, 8*, 101-105.

Saffran, J.R., & Wilson, D.P. (2003). From syllables to syntax: Multilevel statistical learning by 12-month-old infants. *Infancy, 4*, 273-284.

Sell, A., & Kaschak, M.P. (under review). Does speech reading affect word segmentation?

Smith, L.B., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition, 106*, 1558-1568.

Soto-Faraco, S., Navarra, J., Weikum, W.M., Vouloumanos, A., Sebastian-Galles, N., & Werker, J.F. (2007). Discriminating languages by speech-reading. *Perception and Psychophysics, 69*, 218-231.

Stager, C.L., & Werker, J.F. (1997). Infants listen for more phonetic detail in speech perception than in word-learning tasks. *Nature, 388*, 381-382.

Thiessen, E.D. (2007). The effect of distributional information on children's use of phonemic contrasts. *Journal of Memory and Language, 56*, 16-34.

Thiessen, E.D., Hill, E.A., & Saffran, J.R. (2005). Infant-directed speech facilitates word segmentation. *Infancy, 7*, 53-71.

Thiessen, E.D., & Saffran, J.R. (2007). Learning to learn: Infants' acquisition of stress-based strategies for word segmentation. *Language Learning and Development, 3*, 73-100.

Voulomanos, A. (2007). Using probabilities to build a lexicon. Paper presented at the Calgary Workshop on Current Issues in Language Acquisition: Artificial and Statistical Language Learning, June 2007.

Waxman, S.R., & Booth, A.E. (2000). Principles that are invoked in the acquisition of words, but not facts. *Cognition, 77*, B33-B43.

Werker, J.F., Cohen, L.B., Lloyd, V.L., Casasola, M., & Stager, C.L. (1998). Acquisition of word-object associations by 14-month-old infants. *Developmental Psychology, 34*, 1289-1309.

Gesture in language: How sound symbolic words are processed in the brain

Mamiko Arata (arata@sfc.keio.ac.jp)
Graduate School of Media and Governance, Keio University at Shonan-Fujisawa, Japan

Mutsumi Imai (imai@sfc.keio.ac.jp)
Faculty of Environment and Information Studies, Keio University at Shonan-Fujisawa, Japan

Jiro Okuda (jokuda@cc.kyoto-su.ac.jp.ac.jp)
Faculty of computer Science and Engineering, KyotoSangyo University, Japan

Hiroyuki Okada (h.okada@eng.tamagawa.ac.jp)
Department of Engineering, Tamagawa University, Japan

Tetsuya Matsuda (tetsuya@lab.tamagawa.ac.jp)
Tamagawa University Brain Science Institute, Tamagawa University, Japan

Abstract

In traditional linguistics, it has been assumed that the sounds of words are not related to their semantic contents, and that meanings of words are not directly linked to sensory systems. Nevertheless, many languages have a word class in which the sound and meaning of words are systematically related. In this study, by using functional magnetic resonance imaging (fMRI), we scanned brain activity in adult Japanese-speakers while they were seeing locomotion videos together with sound symbolic mimetic words, non-sound symbolic adverbs or verbs. Mimetic words were neurally processed differently from non-sound symbolic adverbs and verbs: We identified extensive bi-hemispheric activations in the regions typically associated with nonverbal cognitive processes for mimetic words but not for non-symbolic verbs or adverbs. The results suggest that mimetic words, by their direct sound-meaning link, have dual neural status both as linguistic symbols and non-linguistic iconic symbols that are directly linked to sensory experience.

Keywords: sound symbolism, brain imaging, symbol grounding in language

Introduction

In the tradition of formal linguistics, language is regarded as an encapsulated system which is functionally separated from other cognitive functions. In this tradition, word meanings are assumed to be represented as a set of universal atomic semantic features that are amodal and not connected todirect sensory experiences. Here, sound symbolism, in which the sound and meaning of words are systematically related, is considered to be a marginal phenomenon in language. However, such a statement seems to be too strong when one looks beyond Indo-European languages. Many languages of the world have a large grammatically-defined word class in which sound symbolism is apparent. For example, in Japanese, mimetic words include not only onomatopoeias for animal sounds but also words referring to motion, tactile sensation and emotional states in which sound is not essential. Mimetic words constitute a large open class of words, and new words can be easily created.

These words are frequently used in everyday conversations and newspaper articles, as well as in various forms of verbal arts, from comic books to novels and poems. Japanese is by no means an exception among languages of the world. Many languages of the world have a similar grammatical class of words with clear sound symbolism (Hinton, Nichols, & Ohara,1994; Nuckrolls, 1999; Voetlz & Kilian-Hatz 2001), including most sub-Saharan African languages (Childs,1994) , and many of the South East Asian languages (called Diffloth, 1972; Watson, 2001; Enfield, 2005) and East Asian languages (Lee, 1992; Mok, 2001; Bodomo, 2006). Even in Indo-European languages that do not have a distinct grammatical class for sound symbolic words (e.g., English), linguists (e.g., Bloomfield, 1933/1984; Bolinger, 1950;Firth, 1935/1957) have pointed out that there is clear sound symbolism in some words (e.g., *squeeze*, *squirt*, *squint*, *bump*, *thump*, and *plump in English*).

Starting with Köhler (1929), there has been a body of empirical work, which demonstrates psychological reality of sound symbolism. Köhler found that, when presented with a curvy round shape and a spiky angular shape, one has the intuition that *baluma* is a better name for the former and *takete* is a better name for the latter (see also Ramachandran & Hubbard, 2001; Westbury, 2004). Sapir (1929) also demonstrated that English speakers associate novel words containing the vowel /i/ with smallness more frequently than words containing /a/.

More recently, empirical evidence for the role of sound symbolism in language processing and novel word learning has been accumulated. For example, sound-shape correlates facilitate category learning involving novel objects both in English-speaking children (Maurer, Pathman and Mondloch, 2006) and adults (Kovic, Plunkett, & Westermann, 2009; Nygaard, Cook & Namy, 2009). Imai and colleagues demonstrated that Japanese 25-month-olds and English

speaking adults who had no exposure to Japanese could detect the sound symbolism underlying novel mimetic words expressing human locomotion (Imai, Kita, Nagumo & Okada, 2008; Kita, Kantartzis & Imai, in press). They further demonstrated that Japanese- as well as English-reared children were greatly helped by sound symbolism in mimetic verbs when they needed to extend a novel verb.

These effects of sound symbolism are not harmonious with formal theories of linguistics. However, when considered from the neurological perspective (e.g., Maurer & Mondloc, 2005; Ramachandran & Hubbard, 2001), researchers may find sound symbolism much less problematic. However, theneural substrate of the phenomena of sound symbolism is still at a stage of speculation. For example, Ramachandran and Hubbard speculated that sound symbolism involves cross-domain mappings between sound contours and motor patterns in or close to Broca's area (possibly mediated by mirror neurons), and between hand gestures and articulatory gestures in the motor area. Also, if sound symbolism is considered as mimicry of the environment by sound, we might expect the activation in the area responsible for integration between sound and other sensory domains such as vision, motion, and touch (cf. Maurer & Mondoch, 2005) .

There are a few existing studies in the literature that examined neural representation of mimetic words (Hashimoto et al., 2006; Osaka, 2004). For example, Osaka (2004) compared mimetic words expressing pain and nonsense words. He identified the activation of anterior cingulate cortex (ACC) -the region known to be active when people experience pain-- when the pain mimetic words were processed.

Hashimoto et al. (2006) examined the pattern of neural activations when Japanese speakers processed mimetic words for animal sounds (e.g. *wan-wan*, bow-wow) as well as actual animal sounds (dog barking). These researchers showed that Japanese mimetic words for animal sounds (e.g. dog barking) elicited the bilateral activation in the superior temporal sulcus (STS) areas.

Importantly, Thierry et al (2003) demonstrated that there is a functional dissociation between the left and right STS: The left STS is mainly responsible for linguistic sound, whereas the right STS is used when environmental sound is processed. The bilateral STS activation may thus suggest that mimetic words expressing animal sounds have dual nature, being processed both as a linguistic sound (word) and environmental sound. Importantly, in this study, both mimetic words and actual animal sounds were presented auditorily. Hashimoto et al reasoned that mimetic words were processed as environmental sounds because mimetic words sounded like actual animal sound, in connection to Thierry et al.'s results,.and argued that the bilateral activation in the STS area reflected the prosodic property of the mimetic words.

These two studies suggest two important characteristics of mimetic words: (1) they are directly anchored to sensory experiences; (2) mimetic words have dual nature, being processed both as a linguistic sound(word) and environmental sound. However, they leave some important questions concerning the nature of sound symbolism unanswered.

First, it is difficult to determine whether the result by Osaka reflect the sound symbolism in the mimetic words per se, as recent neuro-imaging studies have shown that a word could activate the corresponding sensory area in the brain. For example, several studies revealed that verbs encoding face actions (e.g., *lick*), arm/hand actions (e.g., *pick*), and leg/foot actions (e.g., *kick*) differentially engage their corresponding sensory area in the primary motor and premotor regions (e.g., Hauk et al., 2004; Hauk & Pulvermüller, 2004). Thus, all or most words may be anchored to the sensory experience in some degree, whether or not they carry sound symbolism (Barsalou et al., 2003; Kemeler & Tranel., 2008). Still, it is possible that sound symbolic words, especially mimetic words, are tied to sensory experience more strongly and extensively than non- sound symbolic words due to the iconicity they carry.

Concerning the possibility for the cross-domain mappings between auditory and other sensory modalities in sound symbolic words, it is interesting to see whether or not the bilateral STS activations are also seen for sound symbolic words other than animal sound onomatopoeia. If the sound is strongly tied to the meaning in mimetic words, we might expect to see the bilateral STS activations not only for mimetic words expressing actual sounds but for those representing other sensory domains (e.g. motion) which do not directly involve environmental sound. Furthermore, we might expect to see the same activation pattern even when a mimetic word is presented orthographically instead of auditorily.

To uncover these questions, in this research, we compared the neural representation of mimetic words to that of non-sound symbolic verbs and adverbs, all of which express aspects of human locomotion. We scanned brain activities in Japanese speakers while they were presented with locomotion videos together with sound symbolic mimetic words, non-sound symbolic adverbs or verbs. Here, the words were presented orthographically, and the participants were asked to judge how the word semantically matched the locomotion.

As discussed above, mimetic words are expected to be more strongly tied to perceptual experience than non-sound symbolic verbs and adverbs because of their iconicity in the meaning. In fact, it is possible that mimetic words are processed as "gesture in language" by their sound symbolic nature (Ramachandran & Hubbard, 2001). If so, at a broad level, we may expect activations in the right as well as left hemisphere for the mimetic words, as is the case with non-linguistic gesture (e.g., Kita & Lausberg, 2008). When considering specific regions involved with processing mimetic words for locomotions, if they are in fact tied to sensory experiences more strongly than verbs and adverbs, we might expect stronger activations in the middle temporal (MT), motor, and premotor areas for mimetic words than

for verbs and adverbs. Furthermore, if the sound-meaning link is a part of the meaning for mimetic words but not for adverbs or verbs, stronger activation is expected in the STS and superior temporal gyrus (STG) in both hemisphers, as if linguistic sound and environmental sound are both processed (Hashimoto et al., 2006; Thierry et al., 2003).

Method

Participants

Sixteen native Japanese speakers who were either undergraduates or graduates students (mean age =23.7; age range = 22-25; 7 women) participated in the study.

All subjects were right-handed and had normal or corrected-to-normal vision and had no histories of neurological or psychiatric deseases. The data of five participants were excluded from analyses due to artifact components (e.g. head movements) and inadequate performance in the task. The rest of the data from eleven subjects (mean age = 23.4; age range = 22-25; 4women) were used for analyses. All participants gave a written informed consent for participation, and the study was approved by the local ethics committee.

Design and procedure

All participants went through the main experiment (comparing mimetic words, verbs and adverbs) first. After a break, they went through a control experiment, in which the same locomotion videos were presented without words. For the main experiment, we used 16 video clips showing different manners of locomotion, in which an agent moved from left to right on the screen. Each locomotion video was presented together with sound symbolic mimetic words, non-sound symbolic adverbs, or verbs. In half of the trials, the word and the locomotion semantically matched, while in the other half, they did not (e.g., the verb "aruiteiru" (to walk) was shown together with a skipping locomotion). At the end of each block, a fixation point was inserted for 10 seconds to separate the blocks. In each trial, the stimulus (a video clip with a written word) was represented on the screen for 5 seconds. During the stimuli presentation, the participants were instructed to think about the degree of match between the word and the locomotion, but they were asked not to make a response during this period. After the stimuli presentation, the fixation point appeared on the screen for 3 seconds, during which the participants were asked to respond on a scale from 1 to 5 by pressing an appropriate button. There were 4 blocks for each word class and each block consisted of 4 video-word pairs from the same word class. The order of blocks was rotated in the order of mimetic words, verb, and adverb.

All words were shown at the bottom of the video in *hiragana* (a type of orthography each of which represents a syllable). A block design was employed. In each trial, the participants were asked to judge the degree of matching on the scale of 1-5.

In the control experiment, in addition to the videos used for the main experiment, videos showing unnatural biological motions (which were created by morphing videos of natural biological motions) were also shown. The procedure of the control experiment parallels to that of the main experiment except that a word was not presented with the video. During the stimuli presentation, the participants were instructed to think whether the locomotion of the video clip was natural or unnatural as a human movement. After the stimuli presentation, the fixation point appeared on the screen for 3 seconds, during which the participants were asked to respond either 1 (natural) or 2 (unnatural) by pressing the an appropriate key. There were 8 blocks, each of which consisted of 4 video clips. The "natural" trials in which the videos used in the main experiments served as the baseline for visual recognition of the locomotion without verbal description (words).

Stimuli and stimuli validations

Three rating tests were carried out before the fMRI scanning to check whether words representing the three word classes (mimetic words, verbs, adverbs) do not differ in terms of imaginability, familiarity, and age of acquisisition (AOA). All participants were native speakers of Japanese, and none participated in the scanning experiment.

Including the words we used for fMRI scanning task, we prepared 120 words. Twenty-eight participants rated how imaginable each word was. Twenty-seven participants rated how familiar each word was, with a scale from 1 to 7. Finally, we asked other 22 participants to judge around what age they had learned the words to obtain AOA for each word. They were instructed to select the answer from 8 categories; infant period / pre-school age / lower-grade at elementary school/higher-grade at elementary school / junior high school / high school / university or college / do not know the meaning.

The results of the three rating tests indicated that there were significant differences among the three word classes with respect to imaginability (Mimetic words= 5.276 ; Verbs= 6.404 ; Adverbs= 5.616 ;$F(2,81)=3.11$, $p<0.05$) and familiarity (Mimetic words=5.423;Verbs= 6.511 ; Adverbs= 6.08 ; $F(2,78)=3.11$, $p<0.05$). However, importantly, there was no significant difference between mimetic words and adverbs ($t(1,54)=-1.192$ $p=0.119$). Also, the result of Friedman test indicated that there was no significant difference between the mimetic words and the verbs with respect to AOA (Mimetic words=1.523;Verbs=1.545; Adverbs=2.932 ,$p=0.763$), although adverbs were judged to have been acquired later than verbs and mimetic words.

Imaging parameters and analysis

Scanning of fMRI data was performed with a 1.5 Tesla MRI scanner. The fMRI images were obtained using multislice gradient-echo planner imaging (EPI) and were used to produce 20 contiguous, 6-mm thick axial slices covering the whole brain [echo time (TE), 50 ms; repetition time (TR), 2000 ms; flip angle, 90 degree; field of view (FOV), 192 mm; 64 × 64 matrix].

The fMRI data were analyzed using SPM2 software (SPM2; Welcome Department of Cognitive Neurology, UK). The EPI images for each time series were realigned with reference to the first image to correct for head motion. The anatomical images were co-registered with the mean functional images and normalized to the Montreal Neurological Institute (MNI) brain template. Functional data were then normalized using the same transformation parameters and were smoothed in the spatial domain (isotropic Gaussian kernel of 8 mm full width half maximum, FWMH).

Statistical analyses were based on general linear model and activations were modeled using a simple delayed box- car reference vector convolved with a hemodynamic response function (HRF). Low-frequency drifts were removed using a high-pass filter (Holmes *et al.,* 1997) and a first order autoregressive model (AR1) (Friston *et al.,* 2000) was applied for eliminating the temporal autocorrelation of the fMRI time-series data.

Results

Behavioral Results Inside the Scanner

The reaction times for making judgments about the degree of match between the word and the locomotion during the scanning were analyzed. The results indicated that the reaction times for the judgments did not differ across the three word classes, $F(2,20)=0.272$, $p>0.05$.

We also checked if the judged degree of match between the word and locomotion itself differed across the three word classes. No differene was found among mimetics, verb, and adverb conditions.

These results together with the results of the pre-scanning rating studies suggest that the three types of words did not differ in the task difficulty. Hence, if we see differences in the pattern of activations across the three word classes, it cannot be attributed to the task difficulty or processing difficulty of the words.

f-MRI Results

Activation pattern for each word class compared to the baseline

To identify the areas of activation due to processing each type of words, the gross activation for the mimetic, adverb, and verb condition[1] was subtracted by the activation obtained from the motion only (without words) control[1]. The usual left hemisphere dominance was observed for the adverbs and verbs. It is important to note that, for all word types including the verbs and adverbs, after removing the activation responsible for perceptual processing of locomotion video, the activation of the motor-premotor areas was identified. This result suggests that, whether the word is sound symbolic or not, words are anchored to direct perceptual/sensory experiences, in keeping with the embodiment view of concepts and word meanings (e.g., Barsalou et al., 2003, Kemerer, 2010) and in contrast to the formal view of language, which asserts that words are arbitrary symbols.

Interestingly, processing of the mimetic words elicited much broader and stronger activation of the brain than verbs and adverbs. In particular, extensive bi-hemispheric activations were observed when the mimetic word was processed together with the locomotion video, consistent with our hypothesis that mimetic words have dual natures, both as linguistic and non-linguistic gesture-like symbols (Figure 1). Note that the difference between the mimetic words and the other two types of words could not be attributed to familiarity or task difficulty, because the results of pre-ratings and the behavioral results inside the scanner found no difference across them.

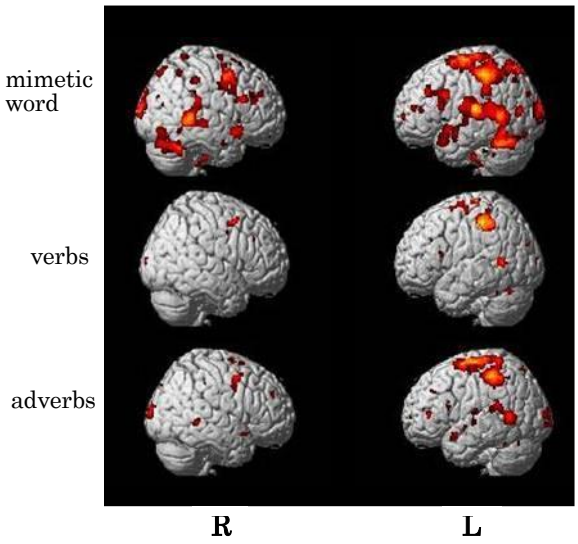


Figure1. Unique activations for each word class. The activation map is overlaid onto a rendered SPM normalized brain. Height threshold at $p<0.001$, uncorrected and 0 voxel extend threshold for one sample t-test were applied).

Unique areas of activation for each type of words

Next we examined the unique areas of activation for verbs, adverbs, and mimetic words. For this purpose, the activations observed for the target word class was subtracted by the other two word classes. For example, in order to see the unique regions for the mimetic words, the activations observed for the verbs and adverbs were subtracted from the activations elicited in the mimetic condition.

As is clearly seen in Figure 2, the verb and the adverbs showed virtually no unique regions left, when the activations for mimetics processing were subtracted. In contrast, as expected, activations of the bilateral STG/STS and the MT areas were shown as the specific regions for the mimetics. Also, the mimetics elicited stronger and broader activation in

[1] Here, we only used the blocks of the "natural" motion because we use only "natural" motions for the main experiment. The data from the unnatural motion blocks were discarded from the analysis.

the motor and pre-motor regions than the verbs and adverbs (Figure 2). This finding further supports the hypothesis that mimetic words are more strongly tied to sensory experience than non-sound symbolic verbs and adverbs, and that cross-domain integration between auditory and other (e.g., motor and motor perception) sensory domains are particularly important for processing of mimetic words.

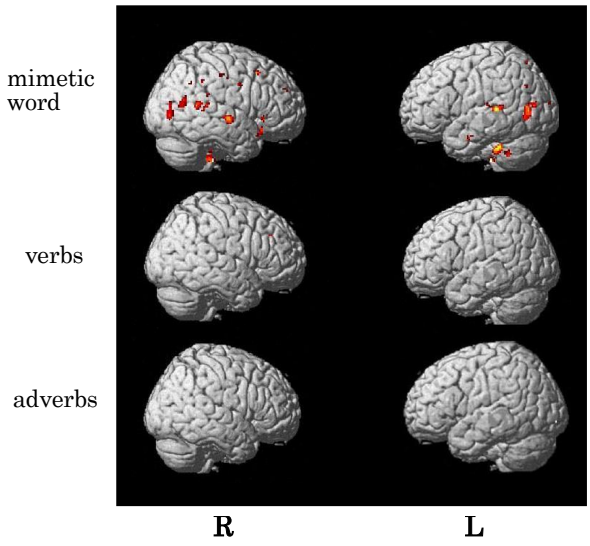


Figure2. The specific regions identified for each word class. The activation map is overlaid onto a rendered SPM normalized brain. Height threshold at p<0.001, uncorrected and 3 voxel extend threshold for one sample t-test were applied).

Correlation between the strength of activation and the degree of semantic match

We further examined if the pattern of brain activity varied as a function of the word-video match or mismatch. The strength of activation in the motor, premotor, and STS areas was correlated with the degree of word-locomotion match. This analysis revelaed that when the meaning of mimetic words matched the locomotion, the right motor area was activated more strong (r =.554, p <.05); when they did not match, the right pre-motor area was activated (r =.-555, p <.05) (Figure 3).

The shift of the areas of activation between the motor and pre-motor regions along with the change in the degree of the mimetic-locomotion match was intriguing: When the locomotion and the mimetic word semantically matched, the participants presumably mimicked the action easily in the brain; but if the mimetic word and locomotion did not match, the participants might have tried to model the action themselves, and as a result, the activation shifted to the primotor area. In contrast, no such correlation was found in the right STG/STS area, suggesting that the activation of the right STS region was related to the processing of mimetic words per se, independent of whether the word semantically matched the motion or not.

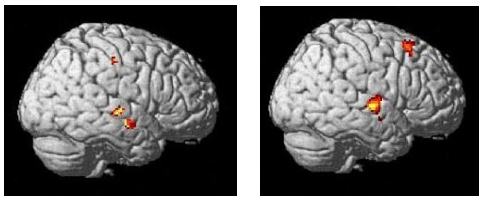


Figure3. The activations in the right hemisphere as a function of matching (Left) and mismatching (Right) mimetic words. The activation map is overlaid onto a rendered SPM normalized brain. Height threshold at p<0.001, uncorrected and 0 voxel extend threshold for ANOVA were applied).

Discussion and Conclusion

This research investigated the neural representation of mimetic words, verbs and adverbs in the domain of human locomotion using fMRI. In the traditional formal linguistics, sound symbolism has been considered as an unimportant aspect of language (e.g., Saussure, 1986; Sapir, 1921). However, recently, this view has been revisited in linguistics, psychology, language development, and neuroscience.

Researchers have demonstrated that certain phonological and prosodic properties are correlated with the meanings of words (e.g., voicedness are correlated with heaviness), and that people are able to detect this sound-meaning correlates from very early developmental stages (e.g., Maurer and Mondoch, 2005). It has also been suggested that sound symbolism may play a role in language development within a child (Imai et al., 2008; Kita et al., 2010) as well as evolution and origin of language (Ramachandran & Hubbard, 2005). In spite of the accumulating evidence for the presence of universal sensitivity to the sound-meaning correspondence, the neural mechanism behind it has been still at a stage of speculation.

This research was conducted to uncover the neural mechanism of sound symbolism by comparing the activation patterns for sound-symbolic mimetic words, (non-sound symbolic) verbs and adverbs in the domain of locomotion. The results largely support the hypothesis that mimetic words have dual natures, somewhat in between linguistic symbols and non-linguistic gesture, as not only the regions relevant to language processing but also those relevant to non-linguistic iconic gestures were activated. The stronger activation of the MT, motor and pre-motor areas also suggest that mimetic words invoke stronger attention to the motion and invites speakers to mentally simulate the action more strongly than regular, non-sound symbolic verbs.

Ramachandran and Hubbard (2005) speculate that processing of sound symbolic words involves cross-domain mappings in the brain between sound contours and motor patterns. The bilateral activation of the STS area found for the

mimetic word processing strongly indicates cross-domain mappings and integration between sound and motion, and provide support for their speculation. In future research, it is important to see if the pattern is replicated for mimetic words expressing other sensory domains (e.g., touch).

The results are also in great harmony with the embodiment view of language and cognition, demonstrating that words in general invoke activations of relevant sensory areas, consistent with previous neuro-imaging studies (e.g. Martin et al, 1995; Huak & Pulvermüller, 2004; Kemmerer& Tranel, 2008). However, they also suggest that the degree of embodiment depends on word types, with highly sound- symbolic words like mimetic words are most directly and strongly bound to sensory experience.

The issue of the origin of the sensitivity to sound symbolism is extremely interesting. Maurer and Mondoch speculate that sensitivity to sound symbolism is universally present prior to language learning, reasoning that it occurs as a bi-product of over-connectivity among different sensory areas in infants. In our lab, we are currently testing whether pre-semantic infants are sensitive to the sound-vision (shape) correlates and how this can be manifested in the brain. This may open the door to the big quest concerning the origin of language.

Acknowledgement

This research was supported by MEXT Ministry of Education, Culture, Sports, Science and Technology) Kakenhi grant (#15300088) to Imai and MEXT Global Center of Excellence (GCOE) program awarded to Tamagawa University.

References

Barsalou, LW., Simmons, W., Barbey, AK., & Wilson, CD. (2003). Grounding conceptual knowledge in modality-specific systems. *Trends in Cognitive Sciences, 7*, 84-91.

B. K. Bergen. (2007). Experimental methods for simulation semantics. In Methods in Cognitive Linguistics, Gonzalez-Marquez, M., Irene Mittelberg, S. Coulson and Michael J. Spivey (eds.), 277-301.

Bloomfield, L. (1984). *Language*. Chicago: University of Chicago Press. (Original Work published 1933).

Bolinger, D. (1950). Rime, assonance, and morpheme analysis. *Word, 6*, 117-136.

Hashimoto, T., Usui, N., Taira, M., Nose, I., Haji, T., & Kojima, S. (2006). The neural mechanism associated with the processing of onomatopoetic sounds. *Neuroimage, 31*, 1762-1770.

Hauk, O., Johnsrude, I., & Pulvermüller F. (2004). Somatotopic representation of action words in human motor and premotor cortex. *Neuron, 41*, 301-307.

Hauk, O., & Pulvermüller, F. (2004). Neurophysiological distinction of action words in the fronto-central cortex. *Human Brain Mapping, 21*, 191-201.

L. Hinton, J. Nichols, & J. Ohala. (eds, 1994). Sound Symbolism. Cambridge, UK: Cambridge University Press.

Imai, M., Kita, S., Nagumo, M., & Okada, H. (2008). Sound symbolism facilitates early verb learning. *Cognition, 109*, 54-65.

Kemmerer, D., & Tranel, D. (2008). Searching for the elusive neural substrates of body part terms: a neuropsychological study. *Cognitive neuropsychology, 25*, 601-29.

Kita, S., Kantartzis, K., & Imai, M. (in press). Children learn soundsymbolic words better: Evolutionary vestige of sound symbolic protolanguage. Proceedings of the 8th conference of *Evolution of Language*.

Kita, S., & Lausberg, H. (2008). Generation of co-speech gestures based on spatial imagery from the right- hemisphere: evidence from split-brain patients. *Cortex, 44*, 131-9.

Köhler, W. (1929). *Gestalt psychology*. New York: Liveright Publishing Corporation.

Kovic, Plunkett, & Westermann. (2010). The shape of words in the brain. *Cognition, 114*, 19-28.

Martin, A., Haxby, JV., Lalonde, FM., Wiggs, CL., & Ungerleider LG. (1995). Discrete cortical regions associated with knowledge of color and knowledge of action. *Science, 270*, 102-5.

D. Mauer, & C. J. Mondloch. (2005). Neonatal synesthesia: A re-evaluation. In L. C. Robertson & N. Sagiv(eds.), Oxford: Oxford University Press.

Maurer, D., Pathman, T., & Mondloch, C. J. (2006). The shape of boubas: Sound-shape correspondences in toddlers and adults. *Developmental Science, 9(3)*, 316-322.

Nuckolls, J. B. (1999). The case for sound symbolism. *Annual Review of Anthropology, 28*, 225-252.

Osaka, N., Osaka, M., Morishita, M., Kondo, H., & Fukuyama, H. (2004). A word expressing affective pain activates the anterior cingulate cortex in the human brain: an fMRI study, *Behav Brain Res, 153*, 123-7.

Ramachandran, V. S., & Hubbard, E. M. (2001). Synaesthesia - a window into perception, thought, and language. *Journal of Consciousness Studies, 8*, 3-34.

Sapir, E. (1929). A study in phonetic symbolism. *Journal of Experimental Psychology, 12*, 225-239.

de Saussure, F. (1983). *Course in general linguistics*. La Salle, IL: Open Court. (Original work published in 1916. Translated by R. Harris).

Thierry, G., Giraud, A.L. & Price, C. (2003). Hemispheric dissociation in access to the human semantic system. *Neuron, 38*, 499-506.

Word Order, Case Forms and Structural Priming in Czech Children's Comprehension

Filip Smolík
Institute of Psychology AS CR

Jií Lukavský
Institute of Psychology AS CR

Abstract: Two groups of Czech children, 3-year-olds (N=28) and 5-year-olds (N=26), participated in a preferential looking experiment testing their comprehension of simple transitive sentences and their susceptibility to structural priming. Four temporarily ambiguous target sentences were presented, two with the canonical SVO word order, two with OVS word order, which is possible but marked in Czech. Each target sentence was preceded by an unambiguous prime sentence with SVO or OVS word order. In 3-year-olds, the presence of OVS primes reduced the garden-path effect observed in the OVS sentences. In 5-year olds, there were no significant effects of structural priming. In unambiguous prime sentences, 5-year olds showed the same level of comprehension in both the canonical and non-canonical sentences. The results suggest that 3-year-olds represent the abstract relationship between agent and patient roles and word order. Results from prime sentences suggest that 5-year-olds can interpret sentences with non-canonical word order.

Modeling Implicit and Explicit Processes in Recursive Sequence Structure Learning

Jamie D. Alexandre (jdalexan@ucsd.edu)
Department of Cognitive Science, 9500 Gilman Drive
La Jolla, CA 92093 USA

Abstract

Recursive structure is viewed as a central property of human language, yet the mechanisms that underlie the acquisition and processing of this structure are subject to intense debate. The artificial grammar learning paradigm has shed light onto syntax acquisition, but has rarely been applied to the more complex, context-free grammars that are needed to represent recursive structure. We adapt the artificial grammar serial reaction time task to study the online acquisition of recursion, and compare human performance to the predictions made by a number of computational language models, chosen to reflect multiple levels and types of syntactic complexity (n-grams, hidden markov models, simple recurrent networks, and Bayesian-induced probabilistic context-free grammars). Evidence is found for a dissociation between explicit and implicit mechanisms of sequence processing, with the SRN more highly correlated with implicit performance, and the PCFG more correlated with explicit awareness of the sequential structure.

Keywords: artificial grammar learning; syntax; recursion; serial reaction time task; simple recurrent network; context-free grammars; implicit/explicit processes.

Introduction

The nature of linguistic structure, and the computational mechanisms by which humans comprehend it, have long been subject to heated debate. Recursion – the ability to hierarchically embed elements within instances of themselves – has been a central point of contention. Although the recursive structure of language was not a new idea at the time, Chomsky formalized the notion of syntactic recursion, touting it as *the* fundamental property that allows for human linguistic ability, a thesis he continues to popularize today (Chomsky, 1956; Hauser, Chomsky, & Fitch, 2002).

In the Chomskyan tradition, the human syntactic system implements a set of rules that allow for theoretically unbounded levels of recursive embedding ("competence"), but this system is then subject to processing constraints, such as working memory limitations, that explain our limited ability to process recursive structures beyond a few levels of embedding ("performance"). Other theorists, particularly from the connectionist camp, have attempted to explain the (limited) human ability to process recursive structure without hypothesizing unbounded competence, by modeling syntactic processing in systems that do not make use of rules or explicit representations (e.g. Elman, 1990; Pollack, 1990; Christiansen & Chater, 1999).

The artificial grammar learning paradigm (initiated by Reber, 1967) has been used to examine processes of syntactic acquisition, but this has been largely restricted to the class of regular grammars, which doesn't shed light onto the acquisition or processing of context-free or recursive structure. The goal of the present study is to obtain estimates of a subject's online string continuation expectancies while responding to sequences generated by a context-free grammar (palindromes), so that these may be compared with the predictions made by a variety of language models trained on the same input history as the subject. The traditional measures of successful acquisition in artificial grammar experiments – such as grammaticality judgments or recall error rates – are not able to provide the incremental (symbol-by-symbol) expectancy data that we require. We adapt a paradigm first employed by Cleeremans & McClelland (1991), known as a serial reaction time task, in which subjects respond to a sequences of stimuli (with a button mapped onto each stimulus class) by pressing the corresponding button as quickly as possible after perceiving stimulus onset. The resulting reaction times are then correlated with the probabilities generated by the competing computational models.

Surprisal

Surprisal, or self-information, is a notion from information theory that quantifies the amount of novel information that a particular event carries with it. An event's surprisal is defined as its negative log probability:

$$-\log(P(x \mid \text{context}))$$

The concept of surprisal has been used in psycholinguistics as a potential measure of incremental processing difficulty, and is thus expected to correlate with behavioral measures such as reading times in eye-tracking studies, and response times in self-paced reading studies (Hale, 2001; Levy, 2008).

The surprisal model requires that we adopt some measure of the probability of a word's occurrence given the preceding sentential context. Hale (2001) uses a probabilistic Earley parsing algorithm to generate incremental word probabilities, using the resulting surprisal values to explain the garden path effect. Levy (2008) uses a similar model to explain a wide-range of effects found in the psycholinguistic literature, such as predictability (e.g. effect of Cloze probability), locality effects (e.g. preference for local dependencies), competition/dynamical models (e.g. greater ease in highly constrained contexts), the tuning hypothesis (e.g. effect of structural frequency), and connectionist models (e.g. predictions made by an SRN). The case of the SRN is particularly interesting, because there are significant divergences between the predictions made by an SRN and a

PCFG-based surprisal model, particularly for constructions such as recursive center-embeddings, which PCFGs process flawlessly, and SRNs – much like humans – have difficulty processing beyond a few levels of embedding (Christiansen & Chater, 1999).

Frank (2009) tested a surprisal model against human eye-tracking data from the Dundee corpus, comparing PCFG-with SRN-generated probabilities, and found that the PCFG produced more accurate objective probabilities, but that the SRN produced probabilities that better matched the human data. He concludes from this, firstly, that subjective probabilities diverge from the objective probabilities, and secondly, that the SRN may in fact be a better model of human performance. Other surprisal studies have used n-gram statistics, such as a trigram model with Kneser-Ney smoothing (Smith & Levy, 2008), and also shown close correspondences with human eye-tracking data.

Language Models

A probabilistic language model is a distribution over the strings (sentences) in a language. The models considered in this paper also all support incremental prediction; that is, given a sentence prefix, they assign a distribution over the symbols that might come next.

To allow for comparison with the human data, each of the models is trained on the precise input that a subject has been exposed to at every point in the experiment (rather than training on a larger corpus, or simply using the probabilities assigned by the model that generated the stimuli). This allows us to observe how a subject's predictions change over the course of learning, to gain insight into the rate at which a system is acquired, as well as possible shifts in strategy, rather than simply comparing fully trained systems.

It is also important to note that none of the model parameters are fit to the human data; a model is trained to predict a sequence's continuation based on the set of sequences it has seen up to that point in the experiment, making use of the algorithmic and representational resources at its disposal, but agnostic to human performance.

The models were chosen from amongst those most commonly used within computational linguistics to model sequential structure, at various levels of complexity (some corresponding roughly to levels in the Chomsky hierarchy).

N-grams (bigrams/trigrams)

One of the simplest but most commonly used language models, n-grams calculate the probability of a symbol in terms of the frequency with which it occurs in its immediately preceding context. Here, we will consider bigrams (which take into account the preceding symbol of context) and trigrams (which take into account the 2 preceding symbols). The predictions made by the n-grams at every step were based on training on all preceding sequences (excluding the sequences that had not yet been seen).

Hidden Markov Model (HMM)

Whereas n-gram transition probabilities are defined between sets of adjacent words, the transitions in a hidden markov model (HMM) are defined over a set of "hidden" states, and these states, in turn, generate the individual words. The idea is that there is an underlying "hidden markov process" that we cannot access directly, and all we can observe is the final sequence of words that is produced by this underlying state sequence. Computationally, HMMs roughly correspond to regular languages at the bottom of the Chomsky hierarchy.

We use the standard Baum-Welch algorithm (Baum et al, 1970) to estimate the HMM's transition and emission matrices from the training corpus (the preceding sequences) for an HMM with 5 hidden states. The trained HMM is then used to compute the incremental posterior probabilities of each symbol given its preceding context. As always, the predictions only used the preceding sequences as a training corpus (so as to be comparable to the human data).

Simple Recurrent Network (SRN)

A simple recurrent network (SRN) is a standard three-layer feed-forward network, with the addition of a context layer that maintains a copy of the hidden layer's state from the previous timestep, and then allows the nodes in this context layer to feed back into the hidden layer during the next timestep, alongside the next input (Elman, 1990). The context layer in an SRN effectively implements time-tapped feedback loops from every node in the hidden layer back to each of the nodes in the hidden layer (delayed by one timestep). The addition of recurrent hidden layer connections allows an SRN to learn to use its hidden layer representations to maintain *task-relevant* contextual information over theoretically unbounded (though in most cases, rapidly decaying) distances.

The SRN used in this paper contained 9 input nodes (one for each symbol, plus a sequence boundary marker), 16 hidden nodes, and 9 output nodes. The network was trained using standard back-propagation, with a learning rate of 0.5 and no momentum, on a single pass through the sequences. Output activations at every timestep were converted into probabilities through the Luce choice rule (in effect, normalizing the network's output vector).

Probabilistic Context-Free Grammar (PCFG)

Context-free grammars (CFGs) have played a central role in linguistic theories of syntax ever since Chomsky (1956) proposed them as being necessary (and almost sufficient) to account for the types of recursive phrase structure observed in human language. A probabilistic context-free grammar adds probabilities to the production rules in a context-free grammar, allowing us to calculate a distribution over strings in the language.

Once we know the parameters of the grammar (see below), incremental predictions can be computed as follows (adapted from Jelinek & Lafferty, 1991):

1. The probability of a string is the sum of the probabilities of all its parse trees.
2. The probability of a string prefix is a sum over the probabilities of all possible completions of the prefix.
3. The probability that a particular symbol w_i will appear following the string prefix $w_1..w_{i-1}$ can be computed by dividing the probability of the prefix with that symbol appended, $P(w_1..w_i)$, by the probability of the prefix, $P(w_1..w_{i-1})$

Stolcke (1995) modified the Earley parsing algorithm to compute the above incremental probabilities efficiently, and we use an implementation by Levy (2008) in the present work.

Learning the parameters of a PCFG from an unparsed corpus is not a trivial task, however. Here, we use a Bayesian framework developed by Mark Johnson[1] that uses Gibbs sampling to learn the probabilities for a set of production rules, given a corpus of training sequences. All combinations of production rules with 8 states (in Chomsky Normal Form, e.g. A->BC) were included in set of candidate rules, and the sampler was given a prior of alpha=0.0001. The counts on the final sample grammar were normalized into probabilities. As with all the other models, the predictions made for every symbol were based on re-training after every sequence, using only on the sequences that occurred prior to that point in the experiment, so that the models have precisely the same information available to them at each timestep as the human subjects. This entire process was repeated 5 times, and the resulting sequences of probabilities were averaged together.

Experiment

Methods

Interface Care was taken in designing and constructing an interface device for the task, due to concerns about measurement noise. The button box (Figure 1) consists of 8 finger-sized push buttons arranged in a 2x4 array, with each button containing its own separately controllable LED for use as a response cue. The buttons and LEDs are interfaced to the PC via a USB-powered LabJack U3 DAQ device, which has very high sampling rates and low command-response latencies, allowing for RTs to be measured to millisecond accuracy.



Figure 1: Button box used in experiment.

[1] http://www.cog.brown.edu/~mj/Software.htm

Participants Eight subjects (mean age 20.5, all right-handed), drawn from the UCSD undergraduate subject pool, received 2 hours of course credit for their participation.

Stimuli Sequences were generated from the following grammar in Table 1.

Table 1: Context-free grammar used to generate stimuli.

Probability	Production Rule
0.193	S → T0 S T0
0.146	S → T1 S T1
0.112	S → T2 S T2
0.128	S → T3 S T3
0.077	S → T4 S T4
0.082	S → T5 S T5
0.159	S → T6 S T6
0.103	S → T7

This grammar generates palindromes, a particular type of "mirror recursion" in which the right-hand side of the sequence is a mirror image (flipped left-to-right) of the left-hand side. The 7th symbol serves as a consistent center marker, making the grammar deterministic. An example sequence would be "0 4 1 3 7 3 1 4 0".

Palindomes are the canonical example of context-free structures, and possibly the simplest type of grammar that is context-free and thus cannot be fully captured by finite state models such as an HMM, or by n-gram statistics.

An experimental session consisted of 16 blocks of 25 sequences each, with sequences ranging in length from 5 to 15. Each of the 8 subjects were presented with the same set of sequences, but with a different mapping of symbols to buttons, shuffled in a Latin-square design such that every symbol was mapped onto each of the 8 buttons for exactly one subject (to balance out any effects of button location or between-button distances).

Procedure Subjects were told that the purpose of the experiment was to study the "effects of practice on reaction times", and were told to "hit each button as quickly as possible when that button's light goes on". No mention was made regarding the structured nature of the stimuli; as far as the subjects were concerned, the sequences were entirely random.

Sequences were presented rapidly, with the next light in a sequence turning on 120ms after the previous button had been released. After the end of an individual sequence there was a 2 second pause before the next sequence began.

In between blocks, subjects were presented with a feedback screen indicating their performance on the block relative to their performance on earlier blocks (plotting their RT contour over time), and also relative to previous subjects, by means of a highscores list derived from earlier pilot testing. Subjects were given a chance to take a short break in between blocks.

After completing the experiment, subjects were interviewed about the strategies they had employed in the

task, the factors they thought affected their performance, and what sorts of patterns (if any) they had noticed in the sequences.

Results and Analysis

Reaction times longer than 1000ms (greater than ~4.2 std above mean) were excluded from analysis, to eliminate extreme outliers caused by events not related to the task (such as distractions, subject sneezing, etc). Only 0.2% of the trials were excluded by this criterion. In addition, the first trial of every sequence was excluded from correlation analyses, as earlier pilot testing using random sequences showed that mean reaction times for these sequence-initial trials were ~70ms slower than for the remainder of the sequence. Reaction times for error trials (when the incorrect button was pressed) were measured from when the light went on to when the correct button was pressed, ignoring the intervening erroneous button press. Subjects made an average of 65 errors each (1.7% of the trials), and these trials were not excluded from the analysis, but doing so has no noticeable effect.

The median reaction time for each trial is calculated across subjects, and then the resulting sequence of reaction times is correlated with the sequences of surprisal values (negative log probability) generated by each of the models. The experiment is divided up into four parts to visualize how the correlations change over the course of training. Standard correlation coefficients and 95% confidence intervals are plotted in Figure 2. Note that each of the models is significantly correlated with the human reaction time data throughout the experiment, though with no model clearly dominating (except perhaps a slight preference for the SRN).

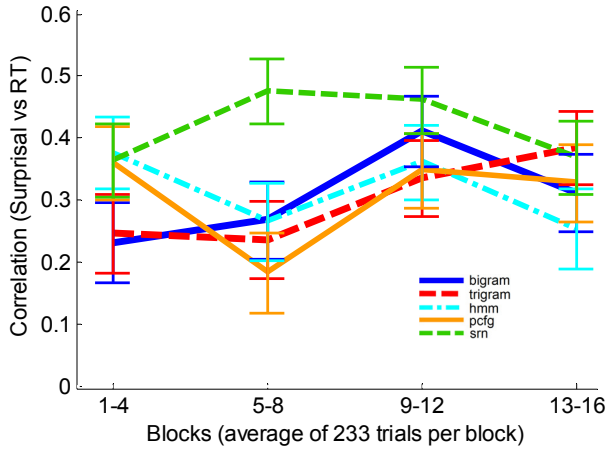


Figure 2: Correlations between models and human reaction times over the course of the experiment.

Several possible interpretations exist at this point. Since the models themselves are quite strongly inter-correlated, it is possible that the correlations for each of the models could be explained by a common shared component. In particular, each of the models is capable of representing n-gram statistics, so perhaps this could explain some portion of the correlation in the other models. To investigate this possibility, partial correlations between the human reaction times and the models are computed after regressing out the bigram and trigram statistics. The residual correlations are plotted in Figure 3.

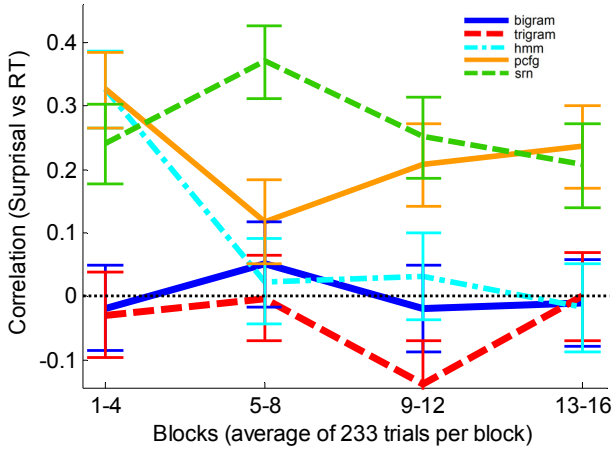


Figure 3: Partial correlations between model probabilities and reactions times, regressing out n-gram probabilities.

As is to be expected, the bigram and trigram correlations become insignificant. The HMM correlations are also eliminated after the first couple of blocks (at which point none of the models have learned very much), suggesting that the HMM was not explaining anything significant about the human behavior beyond n-gram statistics. Both the SRN and PCFG, however, maintain significant correlations throughout, suggesting that they are capturing more about the human reaction times than simply a sensitivity to n-gram statistics.

We might then wonder whether a common component is responsible for both the SRN and PCFG correlations, or if they are each accounting for distinct aspects of the human behavior. To test this, we regress out all models except for the model of interest, and see how much of the variance remains for that model to explain.

Regressing out all the models besides the PCFG reduces its correlations very slightly, but they remain highly over the course of a session, as can be seen in Figure 4 below.

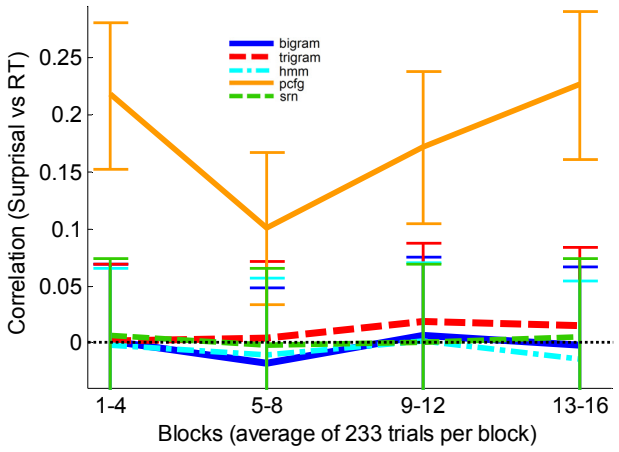


Figure 4: Partial correlations, regressing out all but PCFG.

Similarly, regressing out all models other than the SRN has very little effect on the SRN correlations, which remain strong throughout, despite declining somewhat towards the end (Figure 5).

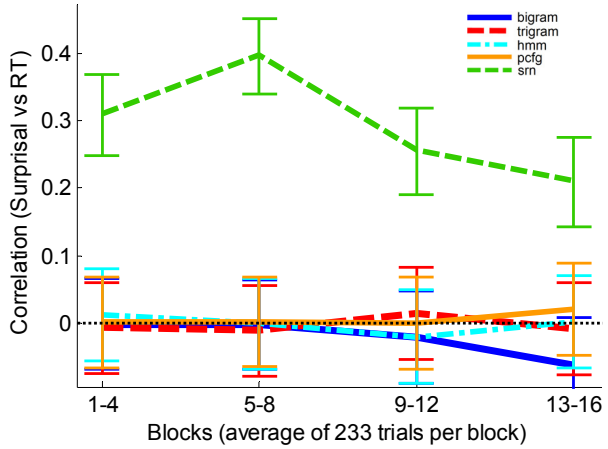


Figure 5: Partial correlations, regressing out all but SRN.

These results seem to suggest that multiple simultaneous processes are playing a role in human behavior on the task; on the one hand, an associative, incremental component captured by the SRN, and on the other hand, a more rule-based, recursive component exemplified by the PCFG. As SRN models have frequently been used to model implicit learning (e.g. Cleeremans, 1993; Misyak et al, 2009), whereas PCFGs are more often associated with explicit rule-based knowledge, we examined individual differences between subjects with regards to implicit and explicit learning, to see if this might help to explain this dissociation.

In the post-testing questionnaire, 3 of the 8 subjects identified some type of structure within the sequences; some referred to it as a "circular" or "mirror" pattern, and one also gave explicit palindromic examples. The 5 remaining subjects had not noticed any regularity to the sequences, even when probed further (2 of these "felt" like there might be some pattern, but could not articulate any details). We separated these two groups from one another and once again calculated partial correlations (regressing out n-grams and the hmm).

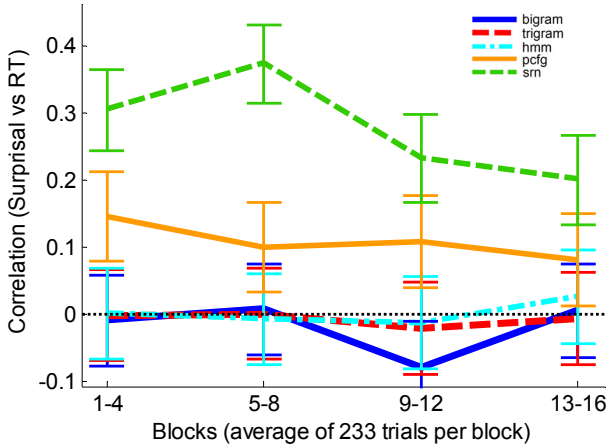


Figure 6: Subjects with no explicit awareness of structure; partial correlations, regressing out n-grams and hmm.

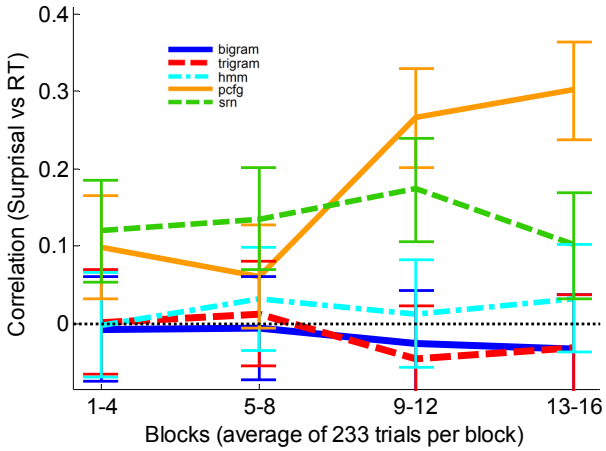


Figure 7: Subjects who were explicitly aware of structure; partial correlations, regressing out n-grams and hmm.

The subjects who were able to report explicit knowledge of aspects of the palindromic structure, by the end of the experiment, showed the strongest correlation with the PCFG (Figure 7), whereas the SRN correlated more strongly with the group that gained no explicit awareness of the structure (Figure 6), indicating that the variance explained by the SRN may reflect a more automatic, implicit processing of the sequential structure (as suggested, for example, by Cleeremans, 1993), whereas the acquisition of recursive, rule-like structures may involve more explicit, conscious processing. It was not possible to query subjects partway through the experiment about whether they had noticed any patterns without drawing their attention to the existence of structure, but the sudden divergence between the PCFG and SRN in Figure 7 lines up well with subjects' comments during the post-test interview that they had begun to notice the pattern somewhere in the "middle of the experiment".

It is also instructive to examine the pattern of reaction times over the course of an average sequence. As the sequences are of different lengths, position on the x-axis is represented as percentage of the way through a sequence (Figure 8).

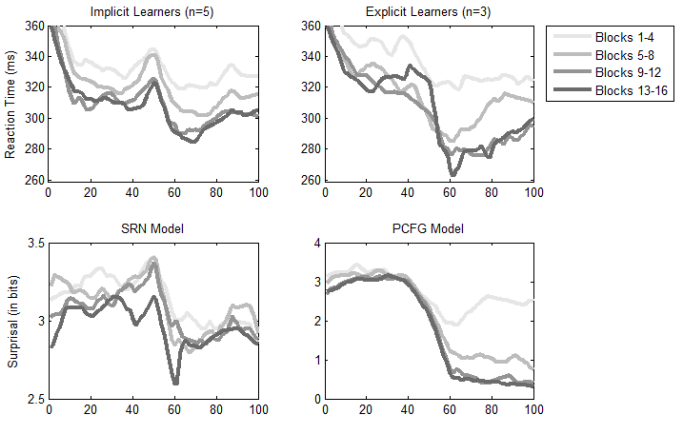


Figure 8: Comparison of RTs and model surprisal over the course of an average sequence (scaled by percentage).

There are several things to note in this reaction time data. Firstly, subjects seem to show a strong advantage in the second half of the sequence, which is consistent with the symbols in the second half being completely determined by the symbols in the first half (due to the palindromic nature of the sequences), and which is seen most strongly both in the PCFG and in the learners with explicit awareness of the structure. Secondly, this advantage is greater immediately following the center symbol and reaction time and then increases slightly as the sequence continues. This is consistent with the fact that later symbols in the second half involve longer-range dependencies, and thus may reflect working memory limitations. The reason for the peak seen halfway through the sequences in both the implicit learners and the SRN is at first unclear, but it is tempting to interpret it as reflecting the cognitive load involved in needing to flip around the first half of the sequence in order to predict the second half, although we might expect this to appear in the explicit rather than the implicit subjects.

Discussion

We attempted to shed light on the mechanisms underlying human processing of recursive structure, by extending the artificial grammar serial reaction time paradigm in two ways; firstly, by training subjects on more complex grammars than are typically used (context-free grammars); and secondly, by comparing performance not only to transitional n-gram probabilities and connectionist models, but also to a Bayesian-induced PCFG model, trained on the exact same set of sequences as the subjects. Evidence was found for a dissociation between implicit and explicit modes of processing, and these modes were seen to correlate most strongly with the predictions of the SRN and the PCFG, respectively.

It may also be fruitful to examine the effects of making subjects explicitly aware of the structure prior to beginning the task, as the results of the present study would suggest this would lead to greater correlation with the predictions of the PCFG. It would also be useful to provide a longer training period, to shed light on how these processes change over the course of more extensive exposure.

References

Baum, L. E., Petrie, T., Soules, G., and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *The Annals of Mathematical Statistics*, 41(1):164-171.

Chomsky, N. (1956). Three models for the description of language. *IEEE Transactions on Information Theory*, 2(3):113-124.

Christiansen, M. H. and Chater, N. (1999). Toward a connectionist model of recursion in human linguistic performance. *Cognitive Science*, 23(2):157-205.

Cleeremans, A. (1993). *Mechanisms of implicit learning: connectionist models of sequence processing*. MIT Press.

Cleeremans, A. and McClelland, J. L. (1991). Learning the structure of event sequences. *Journal of experimental psychology. General*, 120(3):235-253.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, 14(2):179-211.

Frank, S. (2009). Surprisal-based comparison between a symbolic and a connectionist model of sentence processing. *Proceedings of the 31st Annual Cognitive Science Society Conference* (pp. 1139-1144). Austin, TX: Cognitive Science Society.

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. *Proceedings of NAACL*, vol. 2: 159–166.

Hauser, M. D., Chomsky, N., and Fitch, W. T. (2002). The faculty of language: What is it, who has it, and how did it evolve? *Science*, 298(5598):1569-1579.

Jelinek, F. and Lafferty, J. D. (1991). Computation of the probability of initial substring generation by stochastic context-free grammars. *Computational Linguistics*, 17(3):315-323.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, 106:1126-1177.

Misyak, J.B., Christiansen, M.H. & Tomblin, J.B. (2009). Statistical learning of nonadjacencies predicts on-line processing of long-distance dependencies in natural language. *Proceedings of the 31st Annual Cognitive Science Society Conference* (pp. 177-182). Austin, TX: Cognitive Science Society.

Smith, N. and Levy, R. (2008). Optimal Processing Times in Reading: a Formal Model and Empirical Investigation. *Proceedings of the 30th Annual Meeting of the Cognitive Science Society* (oral presentation).

Pollack, J. B. (1990). Recursive distributed representations. *Artif. Intell.*, 46(1-2):77-105.

Reber, A. S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior*, 6(6):855-863.

Stolcke, A. (1995). An efficient probabilistic context-free parsing algorithm that computes prefix probabilities. *Computational Linguistics, MIT Press for the Association for Computational Linguistics*, 21.

The Impact of Starting Small on the Learnability of Recursion

Jun Lai (laij@fsw.leidenuniv.nl)
Cognitive Psychology Department, Leiden University, Pieter de la Court building,
P.O. Box 9555, 2300 RB Leiden, the Netherlands

Fenna H. Poletiek (poletiek@fsw.leidenuniv.nl)
Cognitive Psychology Department, Leiden University, Pieter de la Court building,
P.O. Box 9555, 2300 RB Leiden, the Netherlands

Abstract

Recursion is argued to be the crucial property distinguishing human and non-human primates language learning faculty (Hauser, Chomsky, & Fitch, 2002). Recently, 2 studies (Bahlmann & Friederici, 2006; de Vries, Monaghan, Knecht, & Zwitserlood, 2008), which investigated the learnability of a recursive artificial grammar of the type of A^nB^n, used the same material but reported divergent results. We propose that the organization of the linguistic environment crucially determines learnability of the recursive structure, and that this factor might offer some explanation to the incompatible findings. In a grammaticality judgment task using the same materials as in Bahlmann and Friederici (2006) and de Vries et al.'s (2008), we found significantly better performance when the training input was arranged in a starting small fashion, than when it was organized randomly.

Keywords: Starting small; Recursion; Artificial grammar learning; Statistical learning.

Introduction

Exploring the mechanism behind language learning has been the focus of an enormous body of research in linguistics, psychology and education. The question is how children can possibly acquire such an astonishing complex system so rapidly, while the linguistic environment input is noisy and limited. Sentences like *The rat the cat the dog chased killed ate the malt.* (Chomsky & Miller, 1963) with two recursive center embedding clauses are nearly unintelligible, even for native English speakers (Bach, Brown, & Marslen-Wilson, 1986; Hudson, 1996; Newmeyer, 1988; Vasishth, 2001), due to the associated elements in the sentence being distant from one another (e.g. "the rat" and "ate"). Moreover, recursion is a self-referential principle that can be applied an infinite number of times, producing sentences with numerous embeddings being cognitively very hard to process. Among all syntactical characteristics of natural language, recursion has therefore been argued to be the most fundamental and challenging to acquire (Hauser, Chomsky, & Fitch, 2002).

A recent experimental study (Fitch & Hauser, 2004) using an artificial language has reported that cotton-top tamarins could master the *finite state grammar* (FSG) with the $(AB)^n$ type, but not a higher-level recursive *phrase structure grammar* (PSG) with the A^nB^n type, which could be learned by human participants. Using a familiarization-discrimination paradigm, Fitch and Hauser (2004) first presented the animal participants two auditory sets of consecutive consonant-vowel nonsense syllables (e.g. *la, pa, ba*). Category A syllables were spoken by a female speaker, while Category B syllables by a male. The two sets were identical except for the underlying structure, as well as the pitch. The $(AB)^n$ set in FSG was formed by local transitions between A and B, while the A^nB^n sentences were made according to a center embedding recursive rule (see Figure 1). After this training phase, a discrimination task was performed by the tamarins using the familiarization paradigm. It showed that tamarins could detect the ungrammatical sequences from the grammatical ones in FSG, but not in PSG. Contrastively, humans demonstrated clear discrimination in judging grammaticality of both grammars. This study has raised a renewed interest concerning the inductive learnability of recursive structures, using *artificial grammar learning* (AGL) paradigm (Bahlmann & Friederici, 2006; Bahlmann, J., Schubotz, R.I., & Friederici, A.D., 2008; de Vries, Monaghan, Knecht, & Zwitserlood, 2008; Kersten & Earles, 2001; Perruchet & Rey, 2005). Nevertheless, a study (Gentner, Fenn, Margoliash, & Nusbaum, 2006) concerning song birds' capability of processing A^nB^n structure posed a challenge to this "uniquely human" claim.

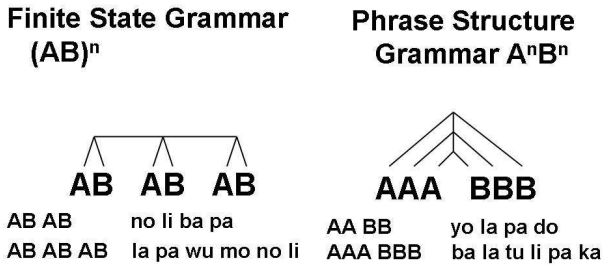


Figure 1. Structures of Finite State Grammar $(AB)^n$ and Phrase Structure Grammar A^nB^n used by Fitch and Hauser (2004). The phrase structure grammar is recursive, center-embedded, and generates long-distance dependencies.

Bahlmann and Friederici (2006, henceforth B&F) and Bahlmann et al. (2008) carried out an fMRI study to probe into the neural basis of processing center-embedding

structures in AGL. Significantly stronger activation in Broca's area, involved in natural language processing, was observed in processing of hierarchically recursive structure A^nB^n, than for the $(AB)^n$ grammar. By contrast, de Vries et al. (2008) replicated this study by B&F but reported no learning of center-embedding structures. De Vries et al. (2008) first trained all participants on the same stimuli as B&F, and required participants to judge the grammaticality of new items violating the center-embedding rule. However, participants were tested with different types of violations, namely: *scrambled* (e.g. $A_xA_yA_z\underline{\mathbf{B}_x\mathbf{B}_z\mathbf{B}_y}$)[1] sequences and *scrambled + repetition* sequences ($A_xA_yA_z\underline{\mathbf{B}_x\mathbf{B}_y\mathbf{B}_x}$). As they predicted, their participants could detect the scrambled + repetition violations, but not the scrambled ones. Therefore, de Vries et al. (2008) argued that successful performance in the study of B&F was due to alternative heuristics, such as counting or repetition-monitoring, instead of learning the abstract center-embedded principle. Indeed, B&F applied *replacement violations* (e.g. $A_xA_yA_zB_z\underline{\mathbf{A}_y}B_x$) and *concatenation violations* (e.g. $A_xA_yB_y\underline{\mathbf{B}_z}$) in their testing materials, which could possibly also be detected without any knowledge of the center-embedding rule, by merely counting the A's and the B's, or by simply detecting a B that was unrelated to any of the A's in a sequence. De Vries et al. (2008) concluded that surface features of A^nB^n sequences were learned by humans, such as repetition patterns and the match between the number of A's and the number of B's, but not the abstract recursive principle determining the long-distance dependencies between each A and each B in such a sequence. In sum, the learnability of center-embedded structures by mere exposure to input exemplars could not unambiguously be established in research using artificial materials, thus far. It seems still inconclusive to which extent AGL studies could help us understand the mechanism of learning recursion.

Here we propose that two fundamental properties of the training set might point at an alternative account of the inconclusive findings. One crucial property is *starting small*, which is the way learning input is ordered. The notion of starting small was first raised by Elman (1991, 1993). He trained a connectionist network to parse complex structures which contained embedded subordinates. The network succeeded in learning only if it was provided with a staged training input (starting small), but not after exposure to the entire random input as a whole. A number of empirical researches showed supporting evidence for this study (Cochran, McDonald, & Parault, 1999; Kareev, Lieberman, & Lev, 1997; Kersten & Earles, 2001), while some other findings yielded contradictory results (Rohde & Plaut, 1999). Possibly the diverging findings might be explained by the highly different methodologies, such as type of study (experimental designs versus simulation studies), stimulus set, input size, training and testing procedures, or the type of grammar used. An input 'growing' gradually, might be especially efficient for learning a complex recursive structure, when the input contains sequences with long distance dependencies, as in the study of B&F.

The second property is *frequency distribution* of the input. In natural language, simple phrases or sentences with zero-level-of-embedding (0-LoE) appear much more frequently than those with several levels of embeddings (Poletiek & Chater, 2006). In real life, this type of short and typical sentences with only adjacent-dependencies, is encountered much more often than more complex compound sentences with several sub-clauses. Sentences with simple structures occur frequently (Philips, 1973; Pine, 1994; Poletiek & Chater, 2006; Snow, 1972). We propose that the distribution of simple and complex sentences in the input set might play a role in rule induction. In our experiment, we presented the input stimuli of our artificial grammar in a distribution that reflected the unequal occurrence of simple and complex sentences in natural language.

To a large extent, both properties of the input we hypothesize to help learners, also occur in the natural linguistic environment of children. Compared to adult-directed speech, child-directed speech has shorter linguistic constituents, simpler structures, and mainly adjacent-dependencies (Pine, 1994). A large amount of repetitions of syntactically short utterances help children learn the basic structure of language. As children grow, child-directed speech develops into more mature speech types (Bellinger, 1980; Garnica, 1977) because more complex constructions are gradually introduced. Therefore, if we can demonstrate experimentally successful grammar learning with a growing environmental input and unequal frequencies for simple and complex exemplars, this might help understanding environmental factors involved in the mechanism of natural language learning.

In the present study, we tested whether participants could learn the hierarchical recursive rule when the learning set was organized 'starting small' rather than randomly, and when unique simple exemplars were repeated, whilst the complex ones were not. We predict that participants will show learning under these conditions.

Experiment

Method

Participants. Twenty-eight students (20 female), from Leiden University participated in the experiment for course credit or payment. All were native Dutch speakers. All had normal or corrected to normal vision.

Materials and design. The same stimuli were used as in B&F and de Vries et al. (2008). There were two sets of syllables, categorized by their vowels. Syllables in Category A contained vowels -e/-i, i.e. {be, bi, de, di, ge, gi}, whereas syllables in Category B contained vowels -o/-u, i.e. {po, pu, to, tu, ko, ku}.

[1] In the figure of Fitch and Hauser (2004), there were no indices for $(AB)^n$ or A^nB^n, because any A could be related with any B. Contrarily, in B&F, de Vries et al. (2008) and the current study, indices were used to indicate dependencies between specific A's and B's.

Each syllable in Category A was associated with its counterpart in Category B according to the onset consonants. For instance, any A_x could be related with any B_x. There were two possible syllables for A_x, i.e. "be" or "bi" and two for B_x, "po" and "pu". Therefore, the associated pairs were {be/bi-po/pu}, {de/di-to/tu} and {ge/gi-ko/ku}. Syllable strings were made out of two, four, and six paired-syllables following the hierarchical center-embedded rule A^nB^n. The resulting grammar G is schematically displayed in Figure 2. Frequencies of syllable occurrence were controlled for.

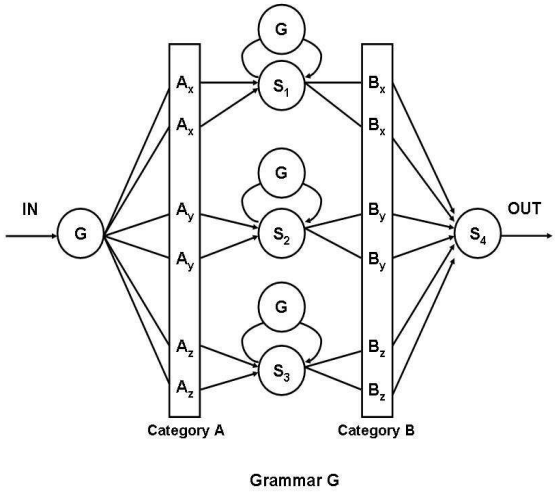


Figure 2. Grammar G, a recursive A^nB^n center-embedded structure. A_x={be, bi}; A_y={de, di}; A_z={ge, gi}; B_x={po, pu}; B_y={to, tu}; B_z={ko, ku}. Examples of strings generated by G are: bi pu (0-LoE), de ge ko tu (1-LoE), be di ge ku to po (2-LoE). "G" in the loops at states S_1, S_2 and S_3 refer to Grammar G, indicating that a center-embedded clause can legally be inserted at that state.

There were 12 blocks in total. Each block consisted of two phases, i.e. learning and testing. All learning and testing blocks together contained 144 strings respectively. Each learning phase was made of 12 syllable strings. After each learning phase, a testing phase followed with 12 novel syllable strings, of which six syllable strings were grammatical and six were ungrammatical.

Note that grammar G generates 12 unique 0-LoE items, 12^2 = 144 unique 1-LoE items, and 144 x 12 =1728 unique 2-LoE items. The 12 unique 0-LoE items were presented four times each (48 in total). Forty-eight 1-LoE items were sampled from the 144 possible ones and presented each once, without repetition. Finally, 48 2-LoE items were sampled from the 1728 unique exemplars of G, and not repeated. In this manner, the differential frequencies of repetitions of 'simple' vs. 'complex' exemplars of a grammar were represented in the input.

Participants were randomly assigned to one of the two experimental groups: the starting small (henceforth SS) group or the random group. All participants were exposed to the same items, i.e., syllable strings, generated by the grammar G in Figure 2. The learning items for the SS group were ordered by their levels of embedding (LoE). In the first four blocks of the SS group, only 0-LoE items were presented during learning. The following four blocks displayed 1-LoE items only. In the last four blocks, 2-LoE items were presented. In this manner, the learning phase was comprised of three consecutive stages, each of which contained four blocks. The ordering of syllable strings within one block was counterbalanced over participants. The random group would see exactly the same set of strings but in a random order. In the random group, each block and each stage contained an equal number of each LoE-category items.

Both groups were presented the same blocks of test items, in the same order. The grammatical test items were novel items with 0-, 1-, or 2-LoE. Ungrammatical items were made by mismatching syllables from Category A and their counterparts from Category B. To control for as many confounding surface cues as possible, the violations satisfied a number of demands. For two-syllable strings, violations appeared necessarily in the second position (e.g. $A_x\underline{\mathbf{B_y}}$); for four-syllable strings, violations appeared in the fourth position (e.g. $A_xA_yB_y\underline{\mathbf{B_z}}$, $A_xA_yB_y\underline{\mathbf{B_y}}$); and for six-syllable strings, violations appeared in the fifth or sixth position (e.g., $A_xA_yA_zB_z\underline{\mathbf{B_z}}B_x$, $A_xA_yA_zB_z\underline{\mathbf{B_x}}B_x$, $A_xA_yA_zB_zB_y\underline{\mathbf{B_z}}$, $A_xA_yA_zB_zB_y\underline{\mathbf{B_y}}$). In this way, no adjacent AB violations (illegal bigrams) were presented except for the two-syllable test items, in which violations were necessarily an illegal bigram, i.e. an illegal AB pair. Secondly, in contrast to B&F, no adjacent repetition of syllables appeared in the same sequence. All grammatical and ungrammatical test strings had an equal number of A's and B's. Hence, violations were not detectable by matching the number of A's to the number of B's. Thirdly, only one illegal pair was allowed in the same string to keep the global level of difficulty constant for each test item. As a result of these constrains, three types of violation were generated: first, violations of type $A_xA_yA_xB_xB_y\underline{\mathbf{B_y}}$ with A's and B's from the same subsets but not equally distributed; second, violations of type $A_yA_yB_y\underline{\mathbf{B_z}}$, or $A_xA_yA_yB_yB_y\underline{\mathbf{B_z}}$ with one B that could not be paired with any of the A's; third, violations of type $A_xA_yB_y\underline{\mathbf{B_y}}$, or $A_xA_yA_zB_zB_y\underline{\mathbf{B_y}}$, with one A missing a B from the same subset. Constructing the violations in this manner, violations detection by superficial heuristics could be largely excluded and categorization performance could be reasonably attributed to knowledge of the hierarchical structure

Procedure. At the beginning of every learning trial, a fixation cross appeared in the center of the screen for 500 ms. Then, each syllable was presented separately for 800 ms, with no interval in-between. Participants were instructed that there was a rule underlying the sequences that they had seen. After presentation of 12 syllable strings, the testing phase followed, in which the sequences appeared in the same fashion. When the last syllable of each test item had disappeared, participants had to press the keyboard buttons

indicating “YES” or “NO”. They were required to make a judgment whether the novel syllable string was grammatical or not, according to the rule underlying the sequences in the learning phase. After each judgment, appropriate feedback was given for 500 ms as B& F and de Vries et al. (2008) did. Approximately, the task took about 30 minutes.

Results and analysis

First, we estimated the mean proportion of “YES” responses to all test items. There was a small response bias favoring positive responses (M = .53, SE = .01, $p < .01$). Accordingly, d'-values were calculated and used as a measure for sensitivity to grammaticality of the responses, i.e. performance. We conducted an independent-samples t-test on mean d’-values for all test items, to compare performance between these two groups. Overall, the SS group (M = 1.51, SE = .36) highly outperformed the random group, M = .08, SE = .05, t (26) = 3.94, p = .001. Moreover, as indicated by a one-sample t-test comparing mean performance with chance level in both groups, only the SS group performed above chance, t (13) = 4.21, p = .001.

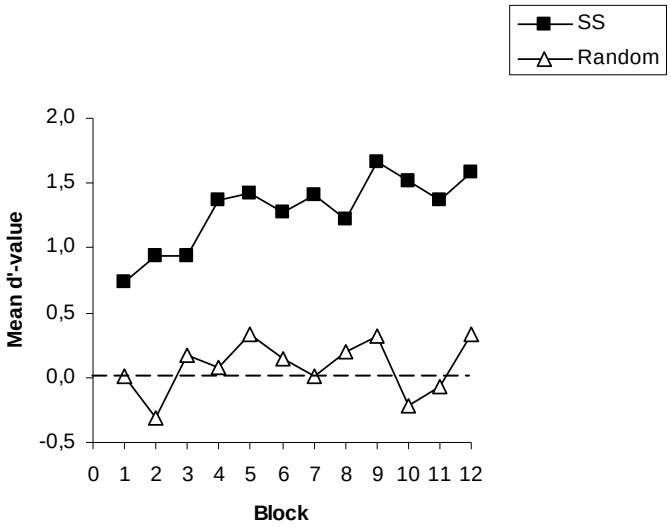


Figure 3. Experiment 1: Mean d’-values for all blocks in both conditions. Points represent mean d’-values per block. The dotted line represents chance level performance (d’= 0).

To evaluate the development over time, in both learning conditions, we compared performance on the first block (Block 1) with the last block (Block 12) for both groups. For the SS group, mean d’-values in Block 1 was M = .73 (SE = .30) and in Block 12, M = 1.59 (SE = .33). Performance had improved in the last block as compared to the first block as revealed by a t-test for means of paired samples, t (13) = 2.59, $p < .05$. In the random group, however, performance did not improve: in Block 1, M = .01 (SE = .21); in Block 12, M = .33, (SE = .29), t (13) = -.98, n.s.. Although in Block 1 the SS group performed slightly better than the random group in the same block, this difference was not significant, t (26) = 1.98, n.s.. However, in the last block, the SS group clearly outscored the random group, t (26) = 2.87, $p < .01$. In Figure 3, mean d’-values are displayed for all blocks in both conditions, showing learning in the SS group over time, but no learning for the random group.

To explore more in detail how the center-embedding recursive principle was learned, we looked into performance on test items with different LoEs. Performance on different types of test items (0-, 1-, and 2-LoE) was compared between conditions, at several stages of exposure. For this analysis, exposure was divided into three stages (Stage 1 consisted of Block 1-4, Stage 2 consisted of Block 5-8, and Stage 3 consisted of Block 9-12.). For the SS group, the stages of training reflected increasing LoE in the stimuli (Stage 1 comprised 0-LoE learning items only; Stage 2, 1-LoE items only; Stage 3, 2-LoE items only). In the random group, all LoEs were presented in the learning phases of every stage. To test the development of performance over time for test items with increasing LoEs, we carried out an ANOVA, with stage and LoE as within-subject factors, and condition as between-subject factor. The LoE × Stage × Condition interaction was significant, F (4, 104) = 2.94, $p < .05$, indicating that performance for different LoE test items developed differently in each learning condition.

Subsequently, an ANOVA was conducted with LoE as the within-subject factor and d’ performance as the dependent variable, for each group separately. For the SS group, a main effect of LoE was found, F (2, 26) = 10.86, $p < .001$. As can be seen in Figure 4, learning for test items with 0-LoE was quite high (M =1.89, SE =.39) and significantly better than learning for items with higher LoE in the SS group, M = 1.45, SE = .37, t (13) = 3.14, $p < .01$ and M = 1.29, SE = .33, t (13) = 4.19, p = .001 for 1-LoE and 2-LoE, respectively. This indicates that participants acquired fundamentally solid knowledge of the adjacent-dependencies of grammar G, under the SS learning condition. Violations of 0-LoE items were observed to be easier to detect than 1-LoE and 2-LoE ones because of their illegal adjacent-dependencies, i.e. bigrams. However, this advantage was only beneficial for the SS group, presented with all 0-LoE training items which clustered in the first stage of exposure. In the random group, participants did not perform differently for various LoE test items. No effect of LoE was found, F (2, 26) = 1.31, n.s. Chance level performance was observed in the random group for all types of test items.

Furthermore, our data revealed a main effect of stage in the SS group only: Performance on all types of test items improved along with exposure to increasing LoE items, F (2, 26) = 3.57, $p < .05$. The curves of 1-LoE and 2-LoE test items evolved equally (see Figure 4), suggesting that the center-embedding rule was learned and recognized equally well for items with one and two recursive loops. In contrast, no main effect of Stage was found for the random group, F (2, 26) = .87, n.s.: Performance was low at the beginning and did not increase significantly over time.

Finally, for the SS group, we compared participants' accuracy on all types of violations with an ANOVA, with Type of Violation as a within subjects factor, to test whether some surface characteristic of the test items (even after careful control for confounding surface cues) might have affected performance. No effect of Type of Violations on accuracy was found, $F(2, 26) = .151$, n.s.. This suggests that participants performed equally well over different types of violations, indicating knowledge of the hierarchical center-embedded structure learned in the SS procedure.

Hence, our findings indicate that center-embedded structures in an AGL could be learned through the SS procedure, but not in the random procedure, in accordance with our hypothesis. Moreover, an incremental exposure to the input in accordance with increasing applications of the recursive rule, correlated with a synchronic improvement in performance. Participants learned the center-embedding principle along with exposure to increasingly more complex exemplars. Robust knowledge of the 0-LoE exemplars could be shown in the SS group only, suggesting that this knowledge was a prerequisite for learning the embedding principle. Furthermore, the SS group did not judge less accurately test items with 2-LoE than items with 1-LoE, suggesting that the recursive rule was learned and recognized equally easily for 1- and 2-LoE strings.

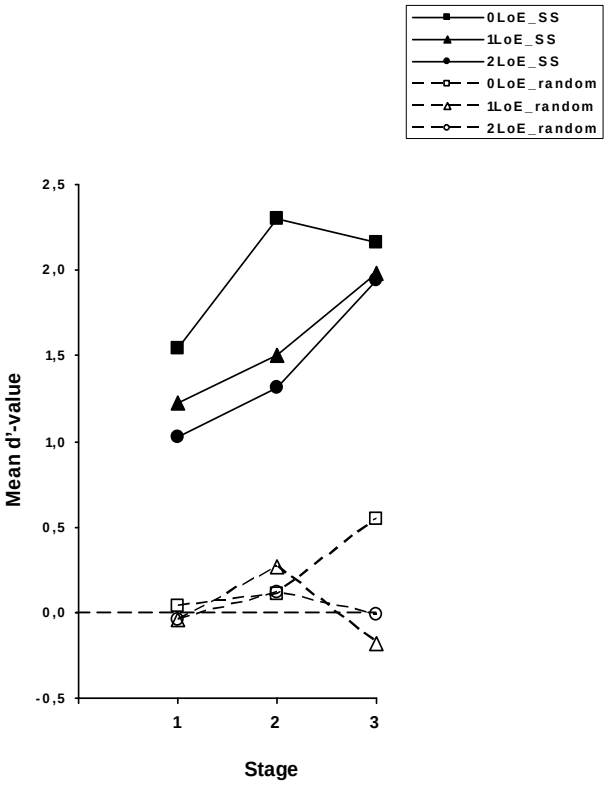


Figure 4. Experiment 1: Mean d'-values for 0-, 1-, and 2-LoE test items at different stages. Points represent mean d'-values of performance per stage. The dotted line represents chance level performance (d'= 0).

Discussion

We observed a 'starting small' effect highly facilitating learning a center-embedded recursive grammar. When participants were presented with a randomized input, there was no learning of the underlying hierarchical rule. Moreover, in our training materials as opposed to the materials presented in similar studies using the same unique training exemplars, simple stimuli were presented more frequently than complex ones, possibly contributing to the dramatic learning effect of the starting small ordering found in our study. In the AGL program, it is still under debate whether performance in learning reflects real knowledge of the abstract grammar, or local pattern learning, recognition of repetitions and other surface heuristics (Poletiek & Van Schijndel, 2009). In the present experimental set up, the violations inserted in the test materials were controlled as much as possible for surface cues that would make them easy to detect without knowledge of the structure. Though the use of cues can not be excluded definitely, our data make a strong case for the learnability of a center-embedded structure provided training with a staged input, and sufficient exposure to basic exemplars without embedded clauses.

Our training stimulus set may be regarded as a representation of the child's natural linguistic environment. The input contains not only a huge number of simple adjacent-dependencies (0-LoE items) produced by the grammar, but they were also presented repeatedly. From the complex items produced (1-, and 2-LoE items), a proportionally smaller sample was presented, and no repetitions occurred. This environment with both growing data and repetitions of basic patterns reflects, as we claim, the natural linguistic environment. In the SS group, due to an intensive training with only 0-LoE items, participants might become familiar with the most basic adjacent-dependencies, which might have provided them with a solid foundation for further induction of the recursive operation. Furthermore, the staged ordering helped participants gradually identify the recursive rule and the connections between long-distance dependencies. By contrast, previous studies failing to find recursion learning, trained participants with the whole corpus randomly presented as an entirety, and no 0-LoE items (de Vries et al., 2008). The two factors investigated here seem therefore to play a crucial role in learning complex recursive rules.

As Elman (1993) indicated humans' most amazing achievement in languages occurs in childhood. In this period, children are exposed to continuously repeated simple structures. Furthermore, the *less is more* proposal that the limited cognitive capacity of children is beneficial to language learning (Newport, 1988, 1990) is consistent with the starting small environmental factor found in our experiment.

In sum, the present study reveals crucial roles for staged input and for solid primary knowledge of the basically simple structures in learning a center-embedded recursive structure by induction. The picture raised is that preliminary

simple associative learning mechanisms such as adjacent-dependencies learning might prepare learners for subsequent processing of gradually encountered more complex and more distant dependencies. Our research suggests that the old puzzle of the inductive learnability of recursive structures might benefit from a shift of focus from the formal characteristics of the structure to the stimulus environment and how this environment is nicely shaped to fulfill the needs of the language learner.

References

Bach, E., Brown, C., & Marslen-Wilson, W. (1986). Crossed and nested dependencies in German and Dutch: A psycholinguistic study. *Language and Cognitive Processes*, 249-262.

Bahlmann, J., & Friederici, A.D. (2006). FMRI investigation of the processing of simple linear and embedded hierarchical structures: An artificial grammar task. *Journal of Cognitive Neuroscience*, Annual Meeting Supplement, p.126.

Bahlmann, J., Schubotz, R.I., & Friederici, A.D. (2008). Hierarchical artificial grammar processing engages Broca's area. *NeuroImage, 42*, 525-534.

Bellinger, D. (1980). Consistency in the pattern of change in mothers' speech: Some discriminant analyses. *Journal of Child Language, 7*, 469-487.

Chomsky, N., & G. Miller. (1963). Introduction to the Formal Analysis of Natural Languages. *Handbook of mathematical Psychology*, R. Luce et al., eds. New York: John Wiley.

Cochran, B.P., McDonald, J.L., & Parault, S.J. (1999). Too smart for their own good: The disadvantage of a superior processing capacity for adult language learners. *Journal of Memory and Language, 41*, 30-58.

De Vries, M.H., Monaghan, P., Knecht, S., & Zwitserlood, P. (2008). Syntactic structure and artificial grammar learning: The learnability of embedded hierarchical structures. *Cognition, 107*, 763-774.

Elman, J.L. (1991). Incremental learning, or the importance of starting small. *Technical Report, 9101*. Center for Research in Language, University of California at San Diego.

Elman, J.L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition, 48*, 71-99.

Fitch, W. T., & Hauser, M. D. (2004). Computational constraints on syntactic processing in a nonhuman primate. *Science, 303*, 377-380.

Garnica, O.K. (1977). Some prosodic and paralinguistic features of speech to young children. In C.E. Snow & C.A. Ferguson. *Talking to children: Language input and acquisition*, 63-68. Cambridge, England: Cambridge University Press.

Gentner, T. Q., Fenn, K. M., Margoliash, D., & Nusbaum, H. C. (2006). Recursive syntactic pattern learning by songbirds. *Nature, 440*, 1204-1207.

Hauser, M.D., Chomsky, N., & Fitch, W.T. (2002). The faculty of language: what is it, who has it, and how did it evolve? *Science, 298*, 1569-1579.

Hudson, R. (1996). The difficulty of (so-called) self-embedded structures. *UCL Working Papers in Linguistics 8*, 283-314.

Kareev, Y., Lieberman, I., & Lev, M. (1997). Through a narrow window: Sample size and the perception of correlation. *Journal of Experimental Psychology: General, 126*, 278-287.

Kersten, A.W., & Earles, J.L. (2001). Less really is more for adults learning a miniature artificial language. *Journal of Memory and Language, 44*, 25-273.

Newmeyer, F. (1988). Extensions and implications of linguistic theory: an overview. In F. Newmeyer, (ed.) *Linguistics: The Cambridge Survey 2. Linguistic Theory: Extensions and Implications*. Cambridge: Cambridge University Press. 1-14.

Newport, E.L. (1988). Constraints on learning and their role in language acquisition: Studies of the acquisition of American Sign Language. *Language Sciences, 10*, 147-172.

Newport, E.L. (1990). Maturational constraints on language learning. *Cognitive Science, 14*, 11-28.

Perruchet, P., & Rey, A. (2005). Does the mastery of center-embedded linguistic structures distinguish humans from non-human primates? *Psychonomic Bulletin and Review, 12 (2)*, 307-313.

Philips, J.R. (1973). Syntax and vocabulary of mothers' speech to young children: age and sex comparisons. *Child Development, 44*, 182-185.

Pine, J.M. (1994). The language of primary caregivers. In C. Gallaway & B.J. Richards (Eds.), *Input and interaction in language acquisition*. Cambridge, UK: Cambridge University Press.

Poletiek, F.H., & Chater, N. (2006). Grammar induction benefits from representative sampling. In R. Sun (Ed.), *Proceedings of the 28th Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum.

Poletiek, F.H., & Van Schijndel, T.J.P. (2009). Stimulus set size and statistical coverage of the grammar in artificial grammar learning. *Psychonomic Bulletin & Review, 16 (6)*, 1058-1064.

Rohde, D.L.T., & Plaut, D.C. (1999). Language acquisition in the absence of explicit negative evidence: How important is starting small? *Cognition, 72*, 67-109.

Snow, C.E. (1972). Mothers' speech to children learning language. *Child Development, 43*, 549-65.

Vasishth, S. (2001). An empirical evaluation of sentence processing models: Center embeddings in Hindi. In M. Daniels, D. Dowty, A. Feldman, & V. Metcalf (Eds.), *OSUWPL, Volume 56*, 159-181, Ohio State University.

Dissociating Sources of Knowledge in Artificial Grammar Learning

Michelle A. Hendricks (mpatte20@slu.edu), Christopher M. Conway (cconway6@slu.edu), and Ronald T. Kellogg (rkellogg@slu.edu)
Department of Psychology, Saint Louis University
St. Louis, MO 63108 USA

Abstract

Previous studies have suggested that individuals use both implicit and explicit, as well as rule and exemplar-based knowledge, to make grammaticality judgments in artificial grammar learning (AGL) tasks. Experiment 1 explored the importance of explicit mechanisms in the learning of exemplar and rule-based information by using a dual-task during AGL training. We utilized a balanced chunk strength grammar, assuring an equal proportion of explicit exemplar-based cues (i.e. chunks) between grammatical and non-grammatical test items. Experiment 2 explored the importance of perceptual cues by changing letters between AGL training and test, while still incorporating the dual-task design and balanced chunk strength grammar used in Experiment 1. Results indicated that participants with a working memory load learned the grammar in Experiment 1 just as well as the single-task no-load group, presumably by relying solely on implicit learning mechanisms. However, changing the letters from training to test resulted in no significant learning for dual-task participants in Experiment 2, suggesting that exemplar-based perceptual cues may the major contributor to implicit knowledge. Overall, the results suggest that implicit and explicit mechanisms for learning rule-based and exemplar-based information may both contribute to AGL via four independent, parallel routes, providing a new framework for understanding the complex dynamic of learning in AGL tasks.

Keywords: artificial grammar learning; implicit learning; working memory; dual-task

Introduction

There is widespread agreement that there exist two distinct forms of learning, explicit and implicit. Explicit learning refers to learning that happens actively, consciously, and with effort, such as the type of learning that occurs during much of formal education. Implicit learning, on the other hand, occurs passively, unconsciously, and without effort. Implicit learning is theorized to be involved in procedural motor activities such as riding a bike or typing, as well as in more complex phenomena such as social interaction and language learning (Reber, 1993).

Artificial grammar learning (AGL) has been a useful paradigm for the study of implicit learning. In the typical artificial grammar learning (AGL) paradigm, individuals are shown (or asked to memorize) letter strings that, unknown to them, conform to rules instantiated by an artificial grammar. Following presentation of the training exemplars, participants are able to reliably determine whether a newly presented letter string is grammatical according to the artificial grammar, without being able to explicitly verbalize the rules of the grammar. Originally, it was theorized that individuals rely on an implicit abstract rule-learning system during AGL tasks, with participants' failures to verbalize the rules as evidence that the rules were unconscious (Reber, 1989).

Additional support for implicit rule-based learning in AGL was provided by what are now referred to as "transfer" experiments. In an AGL transfer experiment, the surface features (e.g. letters) of the training exemplars are changed during the test phase, though the underlying grammar stays the same. Clearly, this would make grammaticality decisions based solely on item similarity difficult, if not impossible. Thus, the transfer manipulation is meant to increase reliance on (presumably implicit) rules divorced from the surface details of the exemplars. Impressively, results from multiple studies have indicated that individuals still successfully demonstrate above-chance classification performance, though the learning is often attenuated (Reber, 1989, Knowlton & Squire, 1996).

In addition to the transfer studies, multiple studies have shown that amnesic subjects, who putatively cannot rely on explicit forms of learning, demonstrate artificial grammar learning similarly to non-brain damaged controls (Knowlton, Ramus, and Squire, 1992; (Knowlton & Squire, 1996). The evidence from both the transfer and the amnesic studies suggest that AGL is mediated by implicit rule-learning mechanisms. Under this view, given that implicit learning is theorized to happen automatically and without effort, executive functions such as working memory (an explicit mechanism, by definition) should have a minimal impact on artificial grammar learning.

Although studies with amnesic patients strongly suggest that AGL can occur without explicit memory, research with non-brain damaged subjects suggests that under normal conditions, explicit processes are also recruited. For instance, test phase classification judgments have been found to be sensitive to the similarity between test and training items, specifically in terms of chunk strength (Chang & Knowlton, 2004; Knowlton & Squire, 1996). Chunks are bigrams and trigrams that are encountered frequently in an artificial grammar due to repetitions in the underlying structure. Studies have shown that individuals do retain some explicit information regarding the chunks of the training items (Dienes, Broadbent, & Berry, 1991; Dulany, Carlson, & Dewey, 1984), and that participants studying only training bigrams can classify the grammaticality of test items correctly at rates similar to controls (Perruchet & Pacteau, 1990). In addition, fMRI studies of AGL tasks have suggested some involvement of the medial temporal lobe (MTL; Fletcher, Buchel, Josephs, Friston, & Dolan, 1999; Opitz & Friederici, 2004). These findings suggest that individuals may rely on a combination of both implicit rule-based knowledge and explicit exemplar-based chunk knowledge to make grammaticality judgments (Vokey & Brooks, 1992; Knowlton & Squire, 1996).

However, although it was originally assumed that rule knowledge is implicit and exemplar-based knowledge is explicit (e.g. Reber 1989), the true picture appears to be much more complex. For instance, participants in a study by Dulany, Carlson, and Dewey (1984) were able to indicate which parts of letter strings were grammatical by crossing out ungrammatical portions, possibly suggesting some explicit knowledge of rules. Similarly, participants in another study demonstrated explicit knowledge of the grammar by being able to complete stems of letter strings to form grammatical strings (Dienes, Broadbent, & Berry, 1991).

Similarly, it appears that implicit learning can also be used to learn both types of information (rule-based and exemplar-based). For instance, Knowlton and Squire (1996) used a balanced chunk strength grammar to show that amnesic patients showed the same pattern of performance as controls, suggesting they were sensitive to both exemplar-based and rule-based information, despite not having explicit knowledge for either. Chang and Knowlton (2004) assessed the importance of low-level perceptual features in AGL performance. Using a balanced grammar, they conducted two experiments: one in which they used a concurrent articulatory suppression task during learning (designed to disrupt perceptual processing), and one where they changed the font and case of letters from acquisition to test. In both cases, participants exposed to the manipulation experienced a disruption in chunk sensitivity, suggesting that exemplar-based knowledge may be more implicit than commonly thought.

In summary, the existing evidence appears to suggest that depending on learning conditions, exemplar and rule-based knowledge may both be acquired implicitly or explicitly. We therefore hypothesized that there may exist at least four separate pathways to learning in AGL (see Figure 1). Exemplar information may be acquired explicitly through memory for chunks (Dienes et al. 1991), or implicitly via perceptual processing (Chang & Knowlton, 2004). Likewise, rule-based knowledge may be acquired via an implicit rule system (Reber, 1967) or via explicit knowledge of rules (Dulany et al. 1984).

The current study aimed to test this proposed four-pathway theory of AGL by attempting to behaviorally dissociate each source of knowledge available to participants. In each of two experiments, we attempted to neutralize one or more of the four hypothesized pathways to knowledge illustrated in Figure 1. In Experiment 1, we incorporated an explicit dual-task during AGL, designed to prevent participants from relying on either form of explicit learning during training (hypothesis generation and item memory), leaving available only implicit sources of knowledge (perceptual fluency and abstract rule-learning). If the four-pathway theory is correct, we should expect that even under this dual-task condition, participants will still demonstrate learning equivalent to single-task participants because they still have access to exemplar-based and rule-based information via implicit learning. In Experiment 2, we

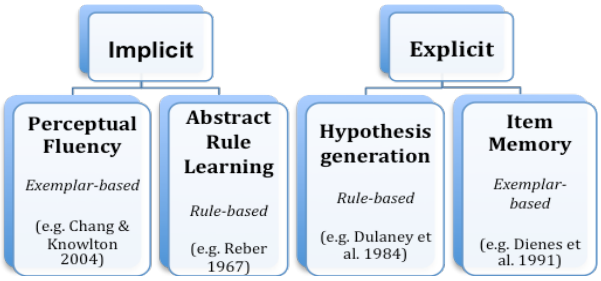


Figure 1: Hypothesized Pathways to Knowledge in Artificial Grammar Learning

furthermore neutralized the implicit perceptual fluency route to learning, leaving dual-task participants only with access to the hypothesized implicit rule-learning mechanism. Unlike Experiment 1, this manipulation is expected to drastically affect learning performance because only the (implicit) rule-based learning pathway is available. Finally, an additional aim of this study is to explore the relationship between individual differences in working memory ability and AGL performance.

Experiment 1: Dissociating Implicit from Explicit Learning

Experiment 1 was designed to address the question of whether learning in the AGL task can take place when explicit mechanisms, specifically working memory, are unavailable. To this end, half of the participants were engaged in a dual-task concurrently with the acquisition phase of the AGL task, designed to make explicit encoding of the stimuli during acquisition very difficult. The dual-task required participants to maintain a series of 6-digit strings in memory at the same time as they were exposed to the letter strings from the AGL task. For the AGL task, we used a balanced chunk strength design (Knowlton & Squire, 1996), which allows us to determine the relative contribution of learning processes to exemplar and rule-based knowledge. In a balanced chunk strength grammar, both grammatical and ungrammatical test items are balanced in terms of the chunks they have in common with the training items, thus ensuring that chunk learning alone cannot account for grammaticality performance. Since we have four categories of test items varying on two dimensions (chunk strength and grammatically), we are able to determine the impact of processing load from the dual-task on grammaticality and chunk strength separately.

We predicted that individuals with diminished explicit resources (i.e. via the concurrent working memory task during AGL acquisition) would still show learning (compared to a single-task control group) due to the availability of implicit mechanisms (perceptual fluency and abstract rule learning).

Finally, we also had each participant engage in an automated OSPAN task (Unsworth, Heitz, Schrock, & Engle, 2005) to measure their working memory abilities. This provided a way to assess the extent that working

memory ability correlates with AGL performance in the dual- vs. single-task groups. We predicted that OSPAN task performance would be associated with AGL performance for the single-task group only, because unlike the dual-task group, their explicit learning pathways are available.

Method

Participants

Participants were 45 undergraduate students (23 in the single-task condition, and 22 in the dual-task condition) who participated for course credit.

Materials

Automated OSPAN The Turner and Engle (1989) OSPAN task requires individuals to solve math problems while trying to remember a set of unrelated words, and is a common measure of working memory. We used an automated version of the OSPAN, designed by Unsworth, Heitz, Schrock, and Engle (2005).The automated OSPAN (AOSPAN) correlates well with other measures of working memory capacity, demonstrating both good internal consistency and test-retest reliability (Unsworth et al. 2005).

Artificial Grammar The artificial grammar used in this experiment is from Knowlton and Squire (1996), which has the advantage of being a balanced chunk strength design (see Figure 2). To determine chunk strength, Knowlton and Squire (1996) quantified the similarity between learning and test items by determining the number of trigrams and bigrams in a test string that corresponded to those appearing in the learning items. We used the same 23 training items and 32 test items as did Knowlton and Squire (1996). The test items are divided into four chunk-balanced categories of 8 items each: grammatical low chunk (G-LC), non-grammatical low chunk (NG-LC), grammatical high chunk (G-HC), and non-grammatical high chunk (NG-HC).

Procedure

Participants were assigned randomly to the dual-task or single-task condition, with all participants tested individually on a computer in a small, private room. All participants first completed the automated OPSAN task, followed by the AGL task. Participants in the dual-task condition completed a concurrent digit span task during the practice and acquisition phases, as described below.

Dual-Task Group After the automated OSPAN, the dual-task participants first received 3 blocks of practice trials to orient them to the task. Within each block, participants were presented with two or three sets of random letter strings consisting of the letters A, B, C, D, and E. For each string, participants were asked to type the letter string as shown in a space at the bottom of the screen; only after correctly typing the string were they allowed to proceed to the next trial. Participants were asked to use only one hand (their dominant hand) to type the strings. During these practice trials, the dual-task participants performed a concurrent working memory task. At the beginning of each practice block, participants were shown six random numbers

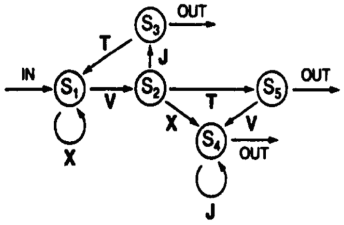


Figure 2: Balanced chunk strength grammar used in current study. From Knowlton and Squire (1996)

presented in the middle of the computer screen for 3000ms. Participants were instructed to maintain the number string in their memory while typing the letter strings as described above. At the end of each block, participants then were required to type the six digits from memory.

Following the practice blocks, participants next completed the acquisition phase, which was nearly identical to the practice phase except for the following differences. Within each block, participants were presented with eight blocks of two or three letter strings each, where each letter string corresponded to one of the 23 training items from the artificial grammar[1]. Each training string was presented only once. As with the practice phase, participants were required to type the string correctly before advancing to the next trial, as well as maintain a 6-digit number string in memory, with a different number string given each block.

During the testing phase, participants were informed that the letter strings shown previously conformed to very complex rules, and that they should use their gut feeling to determine whether the letter strings presented next also conformed to these same underlying rules. Participants were then presented with the 32 test strings, and asked to decide whether each was grammatical or not by pressing a corresponding key on the keyboard. Immediately following each grammaticality judgment, participants were asked to rate their confidence regarding the judgment they had just made on a scale of 1-4 with 1 being "I am sure" and 4 being "I am guessing".

Single-Task Group The single-task participants followed the exact same procedure as the dual-task participants, with the only difference being the nature of the concurrent task. The single-task participants saw a line of 6 asterisks instead of 6-digit number strings at the beginning of each AGL practice and acquisition block. They were not required to remember the asterisks during the trials; merely, at the end of each block, they saw each 6-digit number string and were asked to type it. In this way, the concurrent task did not tap working memory resources and thus serves as a good control to the dual-task group.

Results and Discussion

Main results are shown in Table 1. First we consider performance on the OSPAN and concurrent digit span tasks. As shown in the table, both groups performed comparably

[1] Training strings were randomized within blocks, and the blocks were presented randomly for every participant to account for any order effects.

on the OSPAN task, suggesting that the two groups possessed similar working memory abilities. The table also shows that for the concurrent digit span task, the dual-task participants correctly recalled all six digits at the end of each block 67% of the time (note that the single-task participants do not have a digit span score because they were not required to do the concurrent working memory digit span task). This score suggests that the dual-task had the desired effect of being challenging but not impossible to do. Furthermore, to act as a further control, a regression was conducted which indicated that the OSPAN score predicted 17% of the variance in digit span scores, which was marginally significant ($F(1, 20) = 4.13$, $p = .056$), implying that the effort expended on the dual-task was consistent with what would be expected given participants' working memory abilities. These results suggest that the dual-task had the desired effect of neutralizing or at the very least, attenuating, explicit processing resources for the dual-task group.

For the AGL task, Table 1 shows that both groups demonstrated learning as revealed by their test task performance being significantly greater than chance (single-task group, $t(22) = 5.30$ $p < .001$; dual-task group, $t(22) = 5.30$, $p < .001$). In fact, there were no significant differences between the single and dual-task participants on overall accuracy, the tendency to endorse items as grammatical, or classification confidence.

Even more strikingly, Figure 3 shows the test accuracy for each of the four categories of test items separately for each group. There were no differences between conditions on accuracy for each of the four categories. This indicates that both groups showed equivalent learning of the same two primary types of information present in the grammar (exemplar and rule-based information).

Interestingly, bivariate correlations indicated no correlation between accuracy and confidence judgments for either group. There was, however, significant positive correlation between the OSPAN score and accuracy in the single-task control condition ($r = 0.43$, $p < .05$), and a negative (but non-significant) correlation between the OSPAN score and accuracy in the dual-task condition ($r = -0.23$, ns). These results provide further support that our concurrent task did in fact neutralize working memory resources for the dual-task group; working memory positively contributed to control participants' ability to correctly classify the grammaticality of test items, while it did not contribute to dual-task participants' ability to correctly classify test items. This suggests that the single-group participants were successfully using working memory to learn the grammar, while the dual-task participants were relying on a separate pathway to learning, as indicated by the lack of correlation of OSPAN scores with accuracy in the dual-task.

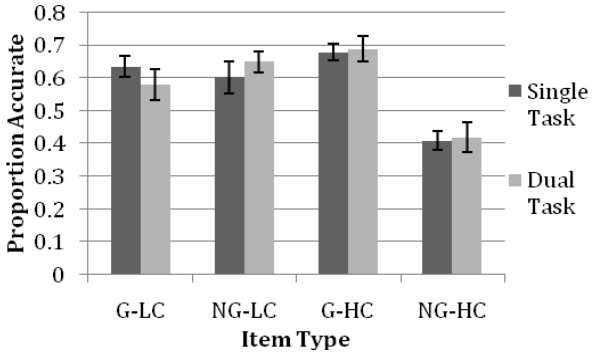


Figure 3: Proportion of Correct Grammaticality Judgments by Group and Item Type in Experiment 1.

In sum, the results from Experiment 1 indicate that dual-task participants exhibited equivalent performance on the AGL task, despite having limited explicit resources available due to the concurrent working memory task during encoding. Strikingly, dual-task participants showed a pattern of learning indistinguishable from controls, indicating that explicit information is not necessary for the acquisition of either exemplar or rule-based information. Our results are consistent with the finding that patients with bilateral hippocampal brain damage (who are unable to explicitly encode information) also showed normal learning on an AGL task (Knowlton & Squire, 1996). Thus, one way to conceptualize Experiment 1 is that it provides a way to behaviorally "simulate" hippocampal brain damage using a concurrent working memory task. By forcing participants to engage in the concurrent digit span task, it appears we successfully prevented participants from relying on the explicit pathways to learning (item memory and hypothesis generation as shown in Figure 1); however, even without full explicit resources available for the AGL task, participants still were able to learn both exemplar and rule-based information in a presumably implicit fashion, leading to performance that was identical to the single-task group.

Table 1: Mean (Standard Deviation) Proportion of Correct Responses, Proportion of Items Endorsed Grammatical, Confidence, and OSPAN Score.

	Experiment 1		**Experiment 2**		
	Single-task	**Dual-task**	**Single-task**	**Dual-task**	**Control**
Proportion Correct*	.58 (.08)	.58 (.08)	.59 (.06)	.53 (.11)	.49 (.11)
Proportion Grammatical	.58 (.10)	.55 (.12)	.54 (.15)	.49 (.22)	.46 (.15)
Confidence	2.97 (.44)	2.68 (.64)	2.45 (.55)	2.62 (.79)	1.92 (.61)
OSPAN Score	47.64 (15.90)	44.77 (15.83)	44.30 (14.91)	48.38 (12.08)	44.12 (18.04)
Digit Span Score	NA	.67 (.19)	NA	.68 (.21)	NA

*Experiment 2: Between single-task and dual-task: $t(49)=2.51$, $p<.05$

Experiment 2: Dissociating Implicit Rule-Based from Exemplar-Based Learning

In Experiment 1, we forced the dual-task participants to rely on implicit learning to learn both exemplar and rule-based information. The aim of Experiment 2 was to attempt to remove an additional pathway to learning, namely implicit perceptual fluency, a form of exemplar based knowledge (see Figure 2), leaving only the implicit rule-based system hypothesized by Reber (1967).

In order to remove the availability of perceptual exemplar-based cues, we incorporated the "transfer" methodology described earlier. Specifically, participants were required to do the test classification task on test strings that consisted of an entirely new letter set. With no perceptual similarity between the acquisition and test phases, dual-task participants can only rely on a more abstract form of knowledge gained via the implicit abstract rule-learning route.

We therefore predicted that single-task participants would show some learning even without exemplar-based cues, since explicit rule-based sources of information would still be available. For dual-task participants, however, only the hypothesized implicit rule-based information will be available. Therefore, dual-task participants should still be able to make correct grammaticality judgments, but they may lose the sensitivity to chunk strength due to lack of exemplar-based cues. Alternatively, dual-task participants may fail to learn the grammar entirely if exemplar-based cues are crucial to learning, as suggested by some accounts (Johansson, 2009; Vokey & Higham, 2005).

Method

Participants

Participants were 84 undergraduate students (26 in the single-task condition, 25 in the dual-task condition, and 32 in the control condition) who participated for course credit. A non-trained control condition was used to ensure that any learning that takes place was not due to the test materials themselves.

Materials& Procedure

The materials and procedure for the single and dual-task groups were identical to Experiment 1, with the exception that the test strings used letters F, Z, N, and C in place of X, T, V and J, respectively. The replacement letters were chosen to be perceptually dissimilar from the training letters, and vowels were avoided so that words could not be formed from strings. Care was also taken to ensure that the letters used for test strings did not result in common acronyms that may interfere with the expression of learning.

The control group completed the same procedure as the dual and single-task conditions, with the exception that they were not given the AGL training. During test, they were told that the letter strings they were about to see were created using a complex set of rules, and that they should use their gut feeling to decide if each string belonged to the rules or not.

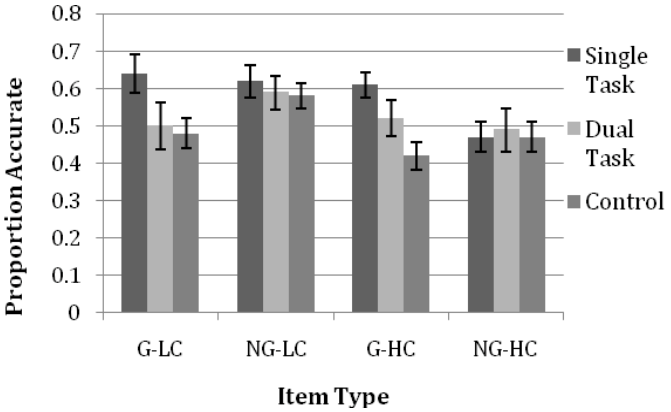


Figure 4: Proportion of Correct Grammaticality Judgments by Group and Item Type in Experiment 2

Results and Discussion

Again, we consider OSPAN and digit span scores first. As Table 1 shows, OSPAN results were equivalent between the two groups, suggesting that the groups' working memory abilities were evenly matched. In addition, performance on the dual-task (68%) was similar to Experiment 1.

As Table 1 also shows, accuracy on AGL test items was significantly greater than chance for the single-task participants only (59%, $t(25) = 6.86$, $p<.001$); dual-task test accuracy (53%, $t(22) = 1.08$, p = ns) and control accuracy (49%, $t(32) = -0.47$, p = ns) were not significantly greater than chance, indicating that only the single-task participants successfully learned the grammar. Further, single-task accuracy was significantly greater than dual-task accuracy ($t(49) = 2.51$, $p < .05$). As in Experiment 1, there were no significant differences between conditions on tendency to endorse grammaticality or classification confidence.

There were however significant differences between conditions on accuracy for the four categories of test items (See Figure 4). Though overall learning for dual-task participants was not significantly above chance, participants did show greater than chance accuracy for the NG-LC category of test items ($t(22) = 2.07$, $p < .05$). Nonetheless, control participants also demonstrated greater than chance accuracy on NG-LC items ($t(22) =$, $df = 32$, $p < .05$) suggesting that accurate performance on these items may reflect test item artifacts rather than implicit learning.

Bivariate correlations indicated no relationship between confidence, accuracy, and OSPAN scores for any group. This is in contrast to Experiment 1, in which there was a significant correlation between OSPAN scores and accuracy for single-task participants. It is unclear why this relationship would not persist in Experiment 2 given that access to explicit knowledge is presumably still available for single-task participants. It is possible that lack of perceptual information resulting from the transfer manipulation made explicit information regarding

exemplars more difficult to utilize during grammaticality judgments at test.

Experiment 2 demonstrates that without explicit learning mechanisms and perceptual features, no learning takes place. We hypothesized that using a combination of concurrent dual-task and transfer methodology, the only pathway to learning left to participants would be the hypothesized implicit abstract rule-learning route. If true, then our results suggest that the kind of implicit rule-based learning originally hypothesized by Reber (1967) does not occur, at least not for transfer tasks. Instead, it appears that explicit mechanisms may be the sole source of knowledge in AGL transfer experiments (Redington & Chater, 1996).

General Discussion

The goal of this study was to attempt to dissociate implicit from explicit learning in artificial grammar learning by selectively neutralizing one or more of the four pathways that we hypothesized are available to learners. In Experiment 1, a concurrent dual-task was used during AGL acquisition to diminish explicit forms of learning. Participants in the dual-task showed strikingly similar test classification performance to the single-task control group, suggesting that they relied on a different – and presumably implicit – set of learning mechanisms at training to demonstrate the same learning as the single-task group. In Experiment 2, we added an additional manipulation – changing the letter set used in the test phase – in order to remove exemplar-based information. Without three of the four hypothesized learning routes, dual-task participants showed patterns of performance similar to non-trained controls, indicating that little to no learning occurred. Therefore, our results bring into question the idea of a rule-based implicit learning system proposed by Reber (1967). Instead, our results are more consistent with recent proposals that implicit knowledge is acquired primarily through exemplar-based perceptual mechanisms (Chang & Knowlton, 2004; Vokey & Higham, 2005). Alternatively, if an implicit rule-learning mechanism does exist, it does not appear to be recruited during AGL transfer tasks.

These results are consistent with the existence of independent implicit and explicit learning mechanisms operating in parallel. Interestingly, access to both implicit and explicit learning systems (e.g. single-task in Experiment 1) does not substantially enhance learning relative to when only implicit learning is available (dual-task in Experiment 1). This suggests that these systems do not operate synergistically. Future work investigating the development of these hypothesized pathways to knowledge in young children, as well as neuroimaging studies to specifically isolate the underlying neural circuits, may prove fruitful. Furthermore, we anticipate that this framework may have ramifications for understanding the nature of certain cognitive and neuropsychological disorders, especially cases in which cognitive learning mechanisms may be disturbed, such as dyslexia or other language impairments.

References

Chang, G. Y., & Knowlton, B. J. (2004). Visual feature Learning in artificial grammar classification. *J Exp Psychol Learn Mem Cogn, 30*(3), 714-722. doi: 10.1037/0278-7393.30.3.7142004-13181-014 [pii]

Deines, Z., Broadbent, D., & Berry, D. (1991). Implicit and explicit knowledge bases in artificial grammar learning. *J Exp Psychol Learn Mem Cogn, 17(5)*, 875-887.

Dulany, D. E., Garlson, R. A., & Dewey, G. I. (1984). A case of syntactical learning and judgment: How conscious and how abstract? *J Exp Psychol Gen, 113*, 541-555.

Fletcher, P., Buchel, C., Josephs, O., Friston, K., & Dolan, R. (1999). Learning-related neuronal responses in prefrontal cortex studied with functional neuroimaging. *Cereb Cortex, 9(2),* 168-178.

Johansson, T. (2009). Strengthening the case for stimulus-specificity in artificial grammar learning: no evidence for abstract representations with extended exposure. *Exp Psychol, 56(3),* 188-197. doi: Y8071KR6Q47652GT [pii]10.1027/1618-3169.56.3.188

Knowlton, B.J., Ramus, S., J., & Squire, L. R. (1992). Intact artificial grammar learning in Amnesia: Dissociation of classification learning and explicit memory for specific instances. *Psychological Science, 3(3),* 172-179.

Knowlton, B. J., & Squire, L. R. (1996). Artificial grammar learning depends on implicit acquisition of both abstract and exemplar-specific information. J Exp *Psychol Learn Mem Cogn, 22(1),* 169-181.

Opitz, B., & Friederici, A. D. (2004). Brain correlates of language learning: the neuronal dissociation of rule-based versus similarity-based learning. *J Neurosci, 24(39),* 8436-8440. doi: 10.1523/JNEUROSCI.2220-04.2004 [doi]24/39/8436 [pii]

Perruchet, P., & Pacteau, C. (1990). Synthetic grammar learning: Implicit rule abstraction or explicit fragmentary knowledge? J Exp Psychol Gen 119, 264-275.

Reber, A.S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning andVerbal Behavior, 77*, 317-327.

Reber, A.S. (1989). Implicit learning and tacit knowledge. J Exp Psychol Gen, 118(3), 219-235.

Reber, A. S. (1993). Implicit learning and tacit knowledge: An essay on the cognitive unconscious. Oxford University Press: New York.

Turner and Engle, R. W. (1989) Is working memory capacity task dependent? Journal of Memory & Language,28, 127-154.

Reddington, M. & Chater, N. (1996). Transfer in artificial grammar learning: A reevaluation. *Journal of Experimental Psychology: General*, *125*, 123-28.

Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behav Res Methods, 37(3)*, 498-505.

Vokey, J. R., & Brooks, L. R. (1992). Salient of item knowledge in learning artificial grammars. J Exp Psychol *Learn Mem Cogn, 18(2),* 328-344.

Vokey, J. R., & Higham, P. A. (2005). Abstract analogies and positive transfer in artificial grammar learning. Can *J Exp Psychol, 59(1)*, 54-61.

Why Streaks Are Special: The Time of Patterns

Yanlong Sun (Yanlong.Sun@uth.tmc.edu)
Hongbin Wang (Hongbin.Wang@uth.tmc.edu)

University of Texas Health Science Center at Houston
7000 Fannin St. Suite 600
Houston, TX 77030 USA

Abstract

People seek for patterns and pay particular attention to streaks even when they are generated by a random process. The present paper examines statistics of pattern time in sequences generated by Bernoulli trials. We demonstrate that streak patterns possess some statistical properties that make them uniquely distinguishable from other patterns. Because of the uncontaminated continuity, streak patterns have the largest amount of self-overlap, resulting in the longest waiting time and the largest variance of interarrival times. We then discuss the psychological implications of pattern time such as in memory encoding and perception of randomness.

Keywords: Perception of randomness; representativeness; hot hand belief; gambler's fallacy; waiting time; patterns.

Introduction

When faced with temporal sequences of events, people often attempt to make sense out of apparent patterns even when they are completely random. Among the most perceptible patterns, "streaks" or "runs," defined as continuous series of the same outcomes, are notorious for they not only yield counterintuitive statistical properties but also inspire extensive investigations on the biases in human perception of randomness and probabilistic judgment and reasoning.

One well-known example is the hot hand belief. Many basketball fans believe that some players have the "hot hand" and tend to make successful shots in streaks. However, in a seminal study, Gilovich, Vallone, and Tversky (1985) find no significant statistical evidence to distinguish the actual shooting sequences from the sequences of Bernoulli trials. This finding has been controversial but withstood several critical attacks (for a comprehensive summary on the hot hand study, see, Bar-Eli, Avugos, & Raab, 2006). In explaining the hot hand belief, Gilovich et al. (1985) use the representativeness heuristic, which has also been used to explain the gambler's fallacy (Tversky & Kahneman, 1974). By such heuristic, people expect the essential characteristics of a chance process to be represented not only by the entire global sequence but also by local subsequences. For instance, when tossing a fair coin, a streak of four heads—which is quite likely in even relatively small samples—would appear to be non-representative.[1] Thus, in the gambler's fallacy, a tail is "due" to balance a streak of heads. In the hot hand belief, a streak of successful shots may lead people to reject the randomness of sequences and signal the existence of a hot hand. Several researchers have questioned the representativeness heuristic for its incompleteness in accounting for two opposite psychological dispositions, but their arguments are still based on the evidence that the hot hand belief is false (e.g., Ayton & Fischer, 2004; Burns, 2004). Together, the hot hand belief and the gambler's fallacy have been considered as two outright fallacies in people's perception of streak patterns, and this stance has a great impact on studies in other disciplines such as behavioral finance and economics (e.g., Camerer, Loewenstein, & Prelec, 2005; Gilovich, Griffin, & Kahneman, 2002; Rabin, 2002).

Moreover, studies on people's judgment and generation of random sequences show that people expect fewer and shorter streaks when observing sequences produced by an independent and identically distributed process (i.i.d.) and they tend to avoid long streaks when instructed to generate such sequences (e.g., Budescu, 1987; Falk & Konold, 1997; Nickerson, 2002; Olivola & Oppenheimer, 2008). Besides behavioral evidence, the salience of streak patterns is also indicated by the results from a functional magnetic resonance imaging (fMRI) study (Huettel, Mack, & McCarthy, 2002). In a "pattern violation task," participants were informed of the random order of the sequences. However, greater activation was found in prefrontal cortex (PFC) when participants observed violations of streak patterns (e.g., [AAAA] vs. [AAAB]) than violations of an alternating pattern (e.g., [ABABAB] vs. [ABABAA]) in a random binary sequence. In addition, the amplitude of fMRI hemodynamic responses (HDR) started increasing at lengths 2 for streak patterns (i.e., [AAB]) but only started increasing at lengths 6 and larger for alternating sequences (i.e., [ABABAA]). (Oskarsson, Van Boven, McClelland, & Hastie, 2009, provided a comprehensive review on judgments of random and nonrandom sequences of binary events.)

Given the unique role of streaks in people's perception and judgment of temporal sequences, an inevitable question is what is so special about streaks? To answer this question, we have to examine the statistics of patterns more carefully

[1] The probability of observing four heads in a row at least once in 20 tosses is 0.48.

since they are widely known for producing counterintuitive results (for the same reason, many people are surprised by the results of the runs test in the hot hand study).

It would seem too obvious to mention once again the unique composition of a streak: a streak is composed of an *uncontaminated* run of the same elements, which makes it exceptionally stand out from other non-streak patterns (such as alternation and symmetry) or any composition without an apparent order. While this property does not affect *how often* a streak occurs, it does affect *when* a streak *first* occurs. To exemplify, we compare two patterns HHH and THH (where H = heads and T = tails in tossing a fair coin). Governed by the independence and stationarity assumptions of Bernoulli trials, these two patterns have the same *probability of occurrence* in any three consecutive tosses (hence the fallacy in the gambler's fallacy and the hot hand belief). However when the coin is tossed repeatedly, the *probability of first occurrence*—the probability that a pattern first occurs at the nth toss, given that the pattern has not occurred before—can be different for different patterns (see Figure 1). For example, both patterns THH and HHH are equally likely to occur or not occur in the first three tosses. If THH has not occurred before, it will have a probability of 0.125 to first occur at the 4th toss. In contrast, if HHH has not occurred before, its probability of first occurrence at the 4th toss is only 0.0625, half of that for THH (for a method of calculating the probability of first occurrence, see Sun, Tweney, & Wang, 2010a). Overall, it will on average take 14 tosses to observe the first occurrence of HHH but only take 8 tosses to observe the first occurrence of THH. Moreover, if we monitor these two patterns simultaneously, it is more likely that we first encounter THH than we first encounter HHH (the odds are 7:1). In other words, it appears that the first occurrence of the streak pattern HHH has been "delayed."

The time it takes for the first occurrence of a pattern (measured by the number of trials) is a statistical property known as *waiting time*. Compared to the long history of studies on the gambler's fallacy (see, Ayton & Fischer, 2004), the development of waiting time and its related properties is fairly new (see, Gardner, 1988; Graham, Knuth, & Patashnik, 1994). Most recently, this development has received attention in psychological literature. Hahn and Warren (2009) argue that given people's limited exposure to the environment (i.e., the number of coin tosses is finite), the longer waiting time of streak patterns would have made them less likely to be observed, thus, "there is something right about the gambler's intuition that the longer the run, the more likely, by contrast, is a sequence with a final tails" (p. 458). Sun et al. (2010a) criticize Hahn and Warren's interpretation, and argue that it is the particular composition of patterns, rather than the length of the global sequence, that plays an essential role in both the statistics of waiting time and people's perception of randomness (also see Sun, Tweney, & Wang, 2010b).

Notwithstanding the debate above, the unique composition of a streak and its "delayed" first occurrence may provide a new prospective in the investigations on human perception of randomness. Particularly, different compositions of patterns may be directly related to memory encoding due to the limited working memory capacity (e.g., Falk & Konold, 1997; Olivola & Oppenheimer, 2008). For example, a streak of HHH can be easily memorized as "3Hs." In addition, different waiting times in effect indicate different variances in the distribution of pattern occurring times (Sun & Wang, 2010), and this fact may have direct consequence in people's intertemporal choices as it has been suggested that human brains are sensitive to time discounting (e.g., Ainslie & Monterosso, 2004; McClure, Ericson, Laibson, Loewenstein, & Cohen, 2007; McClure, Laibson, Loewenstein, & Cohen, 2004). In the following, we demonstrate some interesting properties in the statistics on the time of patterns and discuss the psychological implications.

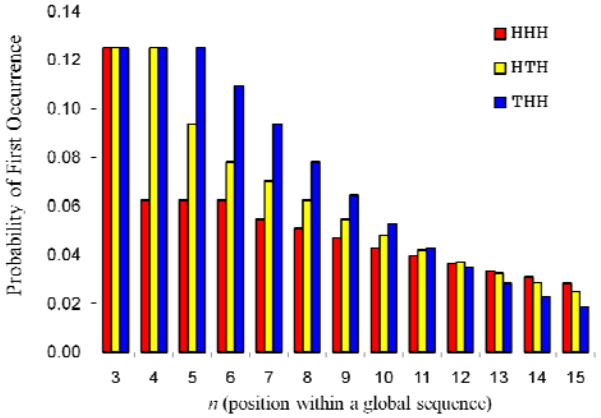


Figure 1: Probabilities of first occurrence for patterns HHH, HTH and THH when a fair coin is tossed repeatedly.

Mean Time and Waiting Time

The time of patterns has been studied by several different methods and different terminologies exist (e.g., Graham, et al., 1994; Li, 1980; Ross, 2007). To be consistent, here we clarify some basic concepts. In a process of coin tossing, the *interarrival time* (T) is the number of trials (tosses) between any two successive occurrences (arrivals) of the pattern, and the *first arrival time* (T^*) is the number of trials required to encounter the first occurrence of the pattern since the beginning of the process[2]. Then, *mean time* ($E[T]$) is the expected value of the interarrival time, and *waiting time* ($E[T^*]$) is the expected value of the first arrival time. We also distinguish the variance of interarrival time and the variance of the first arrival time by $\mathrm{Var}(T)$ and $\mathrm{Var}(T^*)$,

[2] T and T^* may have different distributions, so that the process of counting patterns is also called a general renewal process or a delayed renewal process (Ross, 2007).

respectively. To simplify the discussion, here we only discuss the case of a fair coin (i.e., $p_H = p_T = ½$) and focus on pattern length $r = 3$. Unless specified, the discussion in the following will extend to patterns for all $r \geq 3$. (A more general treatment can be found in Sun & Wang, 2010.)

Overlap and Waiting Time

We first note that when generated by an independent Bernoulli process, a pattern will have a mean time that is the inverse of its probability of occurrence. Thus, any pattern of the same length will have the same mean time. For example,

$$E[T_{HHH}] = E[T_{HTH}] = E[T_{THH}] = (1/2)^{-3} = 8.$$

However, waiting time can be different for different patterns. Compared to other patterns of the same length, streak patterns always have the longest waiting time. For example,

$$E[T^*_{HHH}] = 14,\ E[T^*_{HTH}] = 10,\ \text{and},\ E[T^*_{THH}] = 8.$$

Table 1 lists the mean and variance of interarrival time T and the first arrival time T^* for all possible patterns of length 3. Extra caution should be taken to properly explain these results. An example is given in Figure 2, which depicts pattern time in two different contexts where individual patterns are monitored either independently (panel A) or simultaneously (panels B and C). Note that the colored circles in Figure 2 highlight the position where individual patterns have occurred and they actually represent the values of an "indicator variable" for pattern occurrence. In addition, arrows represent the minimum interarrival time between successive occurrences of patterns—the "minimum succeeding distance" for a pattern to occur given a previous occurrence of either the same pattern or another pattern.

Table 1: Mean and variance of interarrival time T and the first arrival time T^* for patterns of length $r = 3$. Note that for non-overlapping patterns such as HHT, the two pairs of statistics are identical (shown in bold).

Patterns	$E[T]$	Var$[T]$	$E[T^*]$	Var$[T^*]$
HHT	**8**	**24**	**8**	**24**
HTT	**8**	**24**	**8**	**24**
THH	**8**	**24**	**8**	**24**
TTH	**8**	**24**	**8**	**24**
HTH	8	56	10	58
THT	8	56	10	58
HHH	8	120	14	142
TTT	8	120	14	142

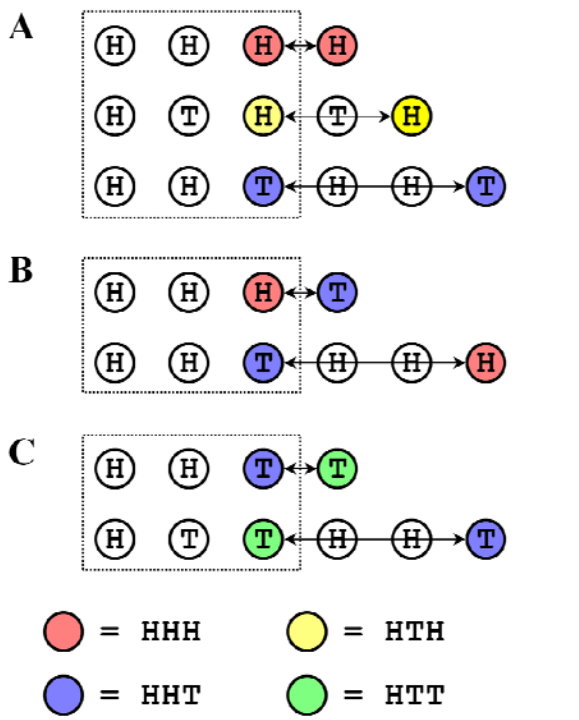


Figure 2: Visualization of pattern occurrences. Each circle represents the outcome of a single toss and the colored circle indicates one occurrence of the corresponding pattern. Arrows represent the "minimum succeeding distance" between successive occurrences of patterns, which also inversely indicate the levels of self-overlap (A) and inter-overlap (B and C).

Figure 2A illustrates the essence of waiting time as it is defined independently for each individual pattern. When the coin has been tossed exactly 3 times, the probability of occurrence for any pattern is the same, 1/8 (also see Figure 1)[3]. However, interesting phenomena will happen at the 4th toss (or, an observational window of size 3 starts moving from the beginning towards the end of the sequence one position a time). For example, if pattern HHH has occurred at $n = 3$, it can have an immediate reoccurrence at $n = 4$. In contrast, if pattern HTH has occurred at $n = 3$, its earliest next occurrence will have to be 2 tosses away at $n = 5$. More extremely, if we are monitoring pattern HHT and it has occurred at $n = 3$, then its earliest next occurrence will have to be 3 tosses away at $n = 6$. An intuitive explanation for this is that the reoccurrences of HHH can *self-overlap* with each other thus tend to be mostly clustered and the

[3] Alternatively, we can imagine that a coin is tossed repeatedly and a long sequence of heads and tails is generated. Then, an observational window of width $r = 3$ randomly lands on any position of the sequence and captures exactly 3 trials. Given the independence assumption of Bernoulli trials, the probability that the observational window will capture any pattern is the same 1/8, as if the process starts from scratch (i.e., the window lands at the beginning of the sequence).

reoccurrences of HHT cannot overlap thus tend to be mostly dispersed.

Figure 2A in effect illustrates all 3 possible levels of self-overlap for patterns of length 3, since reoccurrences of reciprocal patterns self-overlap in the same way (e.g., HHH and TTT, HTH and THT), and reoccurrences of HHT, HTT, THH, and TTH do not self-overlap. Comparing Figure 2A with Table 1, we can see that with other factors remaining constant, the self-overlapping property of a pattern completely determines the pattern's waiting time $E[T^*]$ and the variance of interarrival time $\mathrm{Var}(T)$. Since a streak pattern is an uncontaminated run of the same elements, by such unique composition, a streak pattern will have the largest amount of self-overlap and consequently, the longest waiting time.[4] As a comparison, non-streak patterns such as HTH or HHT only have partial self-overlap or no self-overlap at all so that they will have shorter waiting times. Moreover, waiting time grows approximately exponentially as the amount of self-overlap grows. As a consequence, the difference in waiting time between HHH and HTH is much greater than that between HTH and HHT.

The difference in waiting time can be viewed as one type of *precedence relationships* in which individual patterns are monitored independently and only the self-overlap within each pattern is considered. For example, suppose two players are betting on two patterns HHH and HHT, respectively, then each player tosses a coin of her own in isolation (i.e., the "solitaire game" in Sun, et al., 2010a). Because of the different waiting time, the player who bets on HHT will be more likely to get her desired pattern earlier than the player who bets on HHH.

Inter-overlap and Nontransitivity

The result above might give an impression that pattern HHT is always more likely to occur earlier than pattern HHH, thus the gambler's fallacy might actually have a valid statistical basis. However, such precedence relationship may not hold if two patterns are monitored simultaneously in the same sequence and both self-overlap and inter-overlap are involved (the "interplay game"). The fact is that although HHT is faster than HHH in the solitaire game, in the interplay game, HHT overlaps with the end of HHH (two positions) more than HHH overlaps with the end of HHT (none) (see Figure 2B). Overall, it can be calculated that in the interplay game, we are equally likely to first encounter HHH as to first observe HHT.

Figure 2C shows another comparison between HHT and HTT. Despite that these two patterns have the same waiting time of 8 tosses, because of the different amount of inter-overlap, the odds of HHT preceding HTT against HTT preceding HHT are 2:1. This indicates that the precedence relationship in the interplay game is *nontransitive*. That is, for pattern length $r \geq 3$, if one player first chooses any one of the patterns, the other player can always choose another pattern of the same length to ensure a better than even chance to win. In other words, the interplay game only favors the player who chooses later.

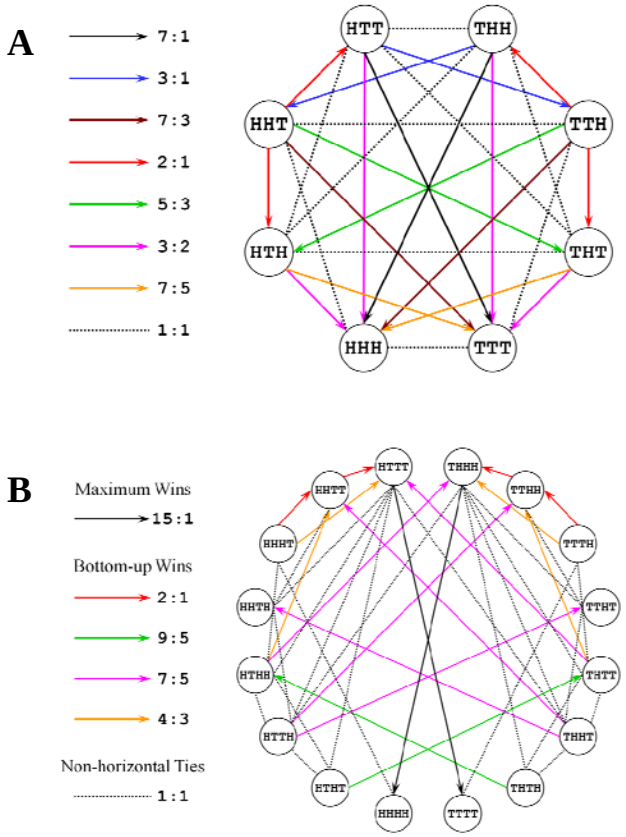


Figure 3: Pair-wise precedence in the interplay game and the corresponding odds. Arrows originate from the faster patterns and point to the slower patterns. A: pattern length $r = 3$. B: pattern length $r = 4$. Only the relationships in legends are connected between patterns and all other connections are either downward wins or horizontal ties. Note that in both A and B, no arrow originates from streak patterns.

Figures 3 shows the pair-wise precedence relationships in the interplay game for pattern length $r = 3$ and 4. Close examinations of Figure 3A confirm that the precedence relationship does not exactly follow the order of waiting time listed in Table 1. Particularly, a pattern with a shorter waiting time may not be necessarily encountered earlier than a pattern with a longer waiting time. Nevertheless, it appears that streaks are still the slowest patterns—at best, a streak pattern can tie with its "end-reversal" counterpart or its reciprocal streak (e.g., HHH vs. HHT, or, HHH vs. TTT), and it can never "beat" any other pattern. In other words, nontransitivity in the interplay game does not mean the

[4] It might seem counterintuitive that overlapped occurrences (hence faster reoccurrences) are associated with a longer waiting time. However, a reoccurrence of the pattern has to be based on a previous occurrence. Since a pattern of length $r \geq 2$ is more likely to have not occurred in the first r tosses than it has occurred, faster reoccurrences actually signify a delay in the waiting time.

equivalence (or indifference) between patterns in a circular fashion. Considering all possible pair-wise comparisons, streak patterns are still unique for their delayed first occurrences.[5] This fact holds for all pattern length $r \geq 2$. Figure 3B shows the pair-wise comparison in the interplay game when pattern length $r = 4$, in which streak patterns are still at the bottom of the game.

Discussion

We have examined several types of pattern time statistics in different contexts and demonstrated that streak patterns indeed possess some unique statistical properties. Here we discuss their psychological relevance and implications in the investigations on human perception of randomness.

First, the particular unbroken continuity of a streak leads to the maximum amount of self-overlap. As a consequence, successive occurrences of a streak tend to be clustered and such tendency would make it harder for human memory to keep an exact count of the actual number of occurrences. By contrast, successive occurrences of other patterns have to be either partially or completely separated (e.g., Figure 2A) and much more evenly distributed (indicated by the small variance of interarrival time in Table 1). For example, in Figure 2A, if the observational window is in size 4 instead of 3, two consecutive instances of 3Hs can be captured by one window. If the memory is encoded as the number of the windows containing the streak (at least once), two instances of 3Hs captured in the same window would have the same weight as one instance of 3Hs. Alternatively, two instances of 3Hs could be replaced by one instance of 4Hs. In either case, the remembered number of occurrences of 3Hs will be less than it actually is.

Moreover, compared with all other patterns, a streak is the slowest pattern to occur, determined by either self-overlap alone (solitaire) or a combination of self-overlap and inter-overlap with another pattern (interplay). In other words, as a random sequence unfolds over time, we are more likely to first encounter another pattern other than a streak. The only exception is the case in interplay where a streak can tie with its end-reversal counterpart or another streak (e.g., Figure 3). Even in this exception, a streak retains an inferior status because of the "minimum succeeding distance" (see Figure 2B). Although it is equally likely HHH preceding HHT as HHT preceding HHH, if HHH occurs first, HHT can immediately follow. If HHT occurs first, the next best shot for HHH has to be 3 tosses away. That is, the discrepancy in the minimum succeeding distance can obscure people's experience of HHH more than it does to HHT.

Together, although streak patterns have the same mean time as any other pattern, their longest waiting time and maximum clustering tendency can leave them underrepresented in people's experience thus make them appear rare or "non-representative" in recollection. Actually, a recent study by Olivola and Oppenheimer (2008) seems to confirm our speculation: when participants recalled the studied binary sequence, the lengths of streaks present in the original sequence were underestimated. Even more interestingly, Olivola and Oppenheimer found that when a streak was present *early* or late in a 25-event sequence, the overall sequence was judged as less likely to be random, compared to when the same streak occurred in the middle of the sequence. It appears that people may actually have an intuitive and accurate response to waiting time such that a streak is unlikely to occur early in sequences generated by a random process.

It should be noted that besides the delayed first occurrence, the particular composition of streaks can manifest itself in many other forms. One example is the probability of occurrence at least once and its complementary "probability of nonoccurrence," whose roles in affecting people's perception of event likelihood have been discussed (Hahn & Warren, 2009; Sun, et al., 2010a). Another example is the shear disparity in the variance of interarrival times between different patterns. When time is essential in predicting future events, different levels of variance may have direct consequences in people's risk preference (e.g., Lopes, 1996; Markowitz, 1991; Sun & Wang, 2010).

Last but not least, in the examples discussed throughout the paper, the sequences of coin tosses are generated by Bernoulli trials (hence inter-event independent and memoryless). However, the process of counting patterns, particularly streak patterns, are not exactly memoryless (this is implied by the unequal mean time and waiting time, see Table 1). As human memory plays essential roles in predicting and planning future events, studies on such process can be useful in order to untangle the interaction of human memory and perception of randomness. Among these different statistics of the similar nature, people can be more sensitive to one form of manifestation than to another or even completely indifferent. We may not be able to use these statistics to vindicate a certain type of bias or fallacy. Nevertheless, these statistics can aid us to better understand the task environment so that we may eventually be able to more precisely pinpoint the source of the error.

Acknowledgments

This work was partially supported by an AFOSR grant (FA9550-07-1-0181), an ONR grant (N00014-08-1-0042), and a Vivian Smith Foundation grant to HW. We would like to thank Ryan D. Tweney, Haiqing Wei, Xuebo Liu, and Franklin Tamborello for helpful discussions and comments.

[5] Guibas and Odlyzko (1981) and Graham et al. (1994) provide strategies to construct a "winning pattern" to beat a given pattern for pattern length $r \geq 3$, in which a streak can never be constructed as a winning pattern.

References

Ainslie, G., & Monterosso, J. (2004). A marketplace in the brain? *Science, 306*(5695), 421-423.

Ayton, P., & Fischer, I. (2004). The hot hand fallacy and the gambler's fallacy: two faces of subjective randomness? *Memory & Cognition, 32*(8), 1369-1378.

Bar-Eli, M., Avugos, S., & Raab, M. (2006). Twenty years of "hot hand" research: Review and critique. *Psychology of Sport & Exercise, 7*(6), 525-553.

Budescu, D. V. (1987). A Markov model for generation of random binary sequences. *Journal of Experimental Psychology: Human Perception and Performance, 13*(1), 25-39.

Burns, B. D. (2004). Heuristics as beliefs and as behaviors: The adaptiveness of the "hot hand". *Cognitive Psychology, 48*, 295-331.

Camerer, C., Loewenstein, G., & Prelec, D. (2005). Neuroeconomics: How neuroscience can inform economics. *Journal of Economic Literature, 43*(1), 9-64.

Falk, R., & Konold, C. (1997). Making sense of randomness: Implicit encoding as a basis for judgment. *Psychological Review, 104*(2), 301-318.

Gardner, M. (1988). *Time travel and other mathematical bewilderments*. New York: Freeman.

Gilovich, T., Griffin, D., & Kahneman, D. (Eds.). (2002). *Heuristics and biases: The psychology of inuitive judgment*. New York: Cambridge University Press.

Gilovich, T., Vallone, R., & Tversky, A. (1985). The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology, 17*, 295-314.

Graham, R. L., Knuth, D. E., & Patashnik, O. (1994). *Concrete mathematics*. Reading MA: Addison-Wesley.

Guibas, L. J., & Odlyzko, A. M. (1981). String overlaps, pattern matching, and nontransitive games. *Journal of Combinatorial Theory, series A, 30*, 183-208.

Hahn, U., & Warren, P. A. (2009). Perceptions of randomness: Why three heads are better than four. *Psychological Review, 116*(2), 454-461.

Huettel, S. A., Mack, P. B., & McCarthy, G. (2002). Perceiving patterns in random series: Dynamic processing of sequence in prefrontal cortex. *Nature Neuroscience, 5*(5), 485–490.

Li, S.-Y. R. (1980). A Martingale approach to the study of occurrence of sequence patterns in repeated experiments. *The Annals of Probability, 8*(6), 1171-1176.

Lopes, L. L. (1996). When time is of the essence: Averaging, aspiration, and the short run. *Organizational Behavior and Human Decision Processes, 65*(3), 179–189.

Markowitz, H. M. (1991). Foundations of portfolio theory. *Journal of Finance, 46*(2), 469-477.

McClure, S. M., Ericson, K. M., Laibson, D. I., Loewenstein, G., & Cohen, J. D. (2007). Time discounting for primary rewards. *Journal of Neuroscience, 27*(21), 5796-5804.

McClure, S. M., Laibson, D. I., Loewenstein, G., & Cohen, J. D. (2004). Separate neural systems value immediate and delayed monetary rewards. *Science, 306*(5695), 503-507.

Nickerson, R. S. (2002). The production and perception of randomness. *Psychological Review, 109*(2), 330-357.

Olivola, C. Y., & Oppenheimer, D. M. (2008). Randomness in retrospect: Exploring the interactions between memory and randomness cognition. *Psychonomic Bulletin & Review, 15*(5), 991-996.

Oskarsson, A. T., Van Boven, L., McClelland, G. H., & Hastie, R. (2009). What's next? Judging sequences of binary events. *Psychological Bulletin, 135*(2), 262-285.

Rabin, M. (2002). Inference by believers in the law of small numbers. *The Quarterly Journal of Economics, 117*(3), 775-816.

Ross, S. M. (2007). *Introduction of probability models* (9th ed.). San Diego, CA: Academic Press.

Sun, Y., Tweney, R. D., & Wang, H. (2010a). Occurrence and nonoccurrence of random sequences: Comment on Hahn and Warren (2009). *Psychological Review, 117*(2), 697-703.

Sun, Y., Tweney, R. D., & Wang, H. (2010b). Postscript: Untangling the gambler's fallacy. *Psychological Review, 117*(2), 704-705.

Sun, Y., & Wang, H. (2010). Gambler's fallacy, hot hand belief, and time of patterns. *Judgment and Decision Making, 5*(2), 124-132.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124–1131.

Social Learning and Cumulative Mutual Improvement in a Networked Group

Thomas N. Wisdom (tnwisdom@indiana.edu)
Department of Psychological and Brain Sciences, Indiana University
1101 East 10th St., Bloomington, Indiana 47405 USA

Robert L. Goldstone (rgoldsto@indiana.edu)
Department of Psychological and Brain Sciences, Indiana University
1101 East 10th St., Bloomington, Indiana 47405 USA

Abstract

We used a simple problem-solving game task to study imitation and innovation in groups of participants. Guesses were composed of multiple elements with linear and interactive effects on score, and score feedback was provided after each of a number of rounds. Participants were allowed to view and imitate the guesses of others during each round, and the score information accompanying others' guesses was either shown or hidden in two conditions. When scores were not visible, social learning was impeded; participants were less efficient in their searching of the problem space and achieved lower performance overall. When scores were visible, higher performance was observed, and results indicated a more equitable sharing of productive exploration among participants within groups as a result of selective imitation and cross-participant cumulative mutual innovations.

Keywords: Social learning; distributed cognition; innovation; imitation; problem solving; innovation diffusion.

Background

The act of learning about the world from others permeates human life. This is evident upon casual reflection about how people gather information and make choices about restaurants or movies, candidates for a job or political office, a new city to live in or a large household purchase, not to mention direct collaboration. Such "social learning" has been defined broadly as "the acquisition of behavior by observation or teaching from other conspecifics" (Boyd & Richerson, 2005). Social learning is a well-studied phenomenon in non-human animals, including foraging choices in starlings (Templeton & Giraldeau, 1996), food preferences in various rodent species (Galef & Giraldeau, 2001), and mate choices in black grouse (Höglund, Alatalo, Gibson & Lundberg, 1995). Humans' rare talent among animals for direct and flexible imitation has been called "no-trial learning" (Bandura, 1965), because it is even faster than the one-trial learning observed in animals with a strong built-in tendency to form certain associations (e.g. between the taste of a food and a subsequent stomach ache). This talent allows an imitator to add new behaviors to his or her repertoire without the costs of trial-and-error learning.

Social Learning Strategies

Tendencies toward individual and social learning depend on the availability and reliability of information in the environment, including other learners. Laland (2004) reviews strategies for *when* social learning is chosen, and *who* social learners choose to imitate. The first class of strategies (when to imitate) often uses the relative cost or uncertainty of asocial learning as criteria. For example, learning about predators on one's own can be very dangerous, so many animals have adapted to learn predator responses from others; in at least one instance this learning has occurred across species (Krause, 1993). The second kind of strategy (who to imitate) often relies on absolute or relative performance of candidate solutions (such as *copy the best* or *copy if better* strategies, respectively), or their relative popularity (such as the *copy the majority* strategy); each of these strategies has been shown in several species (Laland, 2004).

Consequences of Social Learning

Rogers (1988) performed simulations showing that in a temporally unstable environment, the extent to which imitation is beneficial depends on how recently the target of imitation has directly sampled the environment. Therefore, the addition of random social learners (information scroungers) to a population of asocial learners (information producers) does not improve the overall fitness of the population, because the costs of learning avoided by imitators will be offset by costs resulting from the increased use of outdated and inaccurate information. Boyd and Richerson (1995) and Kameda and Nakanishi (2003) confirmed and extended these results to show that when social learners can imitate selectively (e.g. imitating when individual exploration is relatively unreliable and thus more costly), the overall fitness of the population can increase, because both individual and social learning can become more accurate.

Of course, these models are greatly simplified in several ways, among them the assumption that social learners cannot discriminate between model solutions of varying quality without adopting them first. Even without this capability, the benefits for social learners (and thus average benefits for their group) in temporally stable environments are often assumed to be evident (Kameda & Nakanishi, 2002), but the mechanisms by which these benefits accrue are not necessarily clear. If social learning is essentially scrounging that only benefits imitators, then creating obstacles to social learning will only decrease the average performance of imitators. However, the results of previous experiments (Wisdom, Song & Goldstone, submitted) give

us reason to believe that imitators are often also explorers, and that social learning serves as a vital component of the creation of cumulative improvements. Thus impeding social learning is expected to lead to decreases in the performance of all participants.

Experiment Overview

The following experiment investigates both the causes and consequences of social learning. We employ a task in which participants in groups consisting of between one and nine persons are instructed to individually build solutions, which consist of multiple elements chosen from a larger set of elements over a series of rounds. These solutions are evaluated according to a score function that takes into account both individual element values and interactions between them. Groups of participants play simultaneously, and each can view the tentative solutions of all others. In one condition, participants may view fellow participants' scores alongside their solutions, and in another condition fellow participants' scores are invisible.

Predictions

We made the following predictions. When evaluative information about peer solutions was unavailable, participants would be unable to be sufficiently selective in imitation, and thus participants employing highly imitative strategies would have relatively lower scores. Imitation strategies in both conditions would be biased toward peers with solutions similar to the imitator's, and toward adopting solution elements that were more popular among peers, but these effects would be more pronounced in the invisible-scores condition in order to compensate for the lack of direct evaluative information. Mean scores would be lower for participants (including successful asocial learners) in the invisible-scores condition because they would be unable to easily take advantage of good solutions found by others through selective imitation and further improve upon them.

Methods

Participants were recruited from the Indiana University Psychological and Brain Sciences Department undergraduate subject pool, and were given course credit for taking part in the study. Participants populated each session by signing up at will for scheduled experiments with a maximum capacity of 9 persons. 234 individuals participated in the experiment, distributed across 65 sessions as shown in Table 1.

Table 1. Distribution of participants across group sizes.

Group size	1	2	3	4	5	6	7	8	9
# Sessions	16	8	11	11	5	2	3	3	2
# Participants	16	16	33	44	35	12	28	32	18

Task Details and Instructions

We implemented the experiment using custom software written in Java and Flash and run in a web browser (a version of the task can be run as "Creature League" at http://groups.psych.indiana.edu/). Each participant used a mouse to interact with the experimental game. A central game server recorded data and updated participant displays at the end of each round. In the game itself, participants attempted to maximize the scores earned by their chosen subsets ("teams") from a set ("league") of creature icons over 24 rounds. The display included an area for the participant's own current team, another area that could be toggled to show the participant's previous round team or their best-scoring team so far in the game (along with its associated score), a league area which showed all of the icons that were available for selection, and indications of the current round in the game and the amount of time remaining in the current round. If a session included more than one participant, each participant's display also showed the team and, in the visible-scores condition, the associated score for each other participant in the previous round. The ordering of other participants' teams in the display was not kept constant across conditions, to avoid imitation based on past behavior. Icons could be copied from any part of the display to a participant's current team by dragging and dropping them with the mouse, except for those already on the participant's current team. The current team could be replaced entirely by another team by using the score box above it as a "handle" to drag it to the current team area. A screen capture of the task interface is shown in Figure 1.

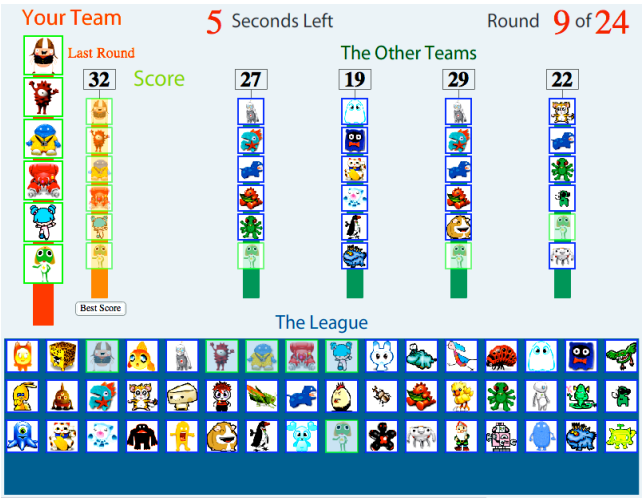


Figure 1: Example of experiment task display.

At the beginning of each session, players were given a hands-on demo of the game (including the various ways to move creatures to one's current team), and further informed about the mechanics of the game and what to expect in the remainder of the experiment session, including the following information. Each game consisted of 24 rounds, and each round was 10 seconds long. Score feedback was

given after each round: if the participant's score had improved from the previous round, the numerical score display counted up to the new score and turned green, and if it had worsened, the display counted down to the new score and turned red. At the end of each game, the display showed the player's final score, along with a table of the scores of each player in each round of the game, sorted by average score. The player's own score was highlighted to show their relative performance without placing competitive emphasis on it. Players were instructed to do their best to maximize their teams' scores over all 24 rounds. At the beginning of each game, each player's team was a random selection of creature icons from the league. Each group played 6 games; in 3 of the games, other participants' scores were visible, and in the other 3 they were not. These were called the visible-scores and invisible-scores conditions, respectively, and were played in random order in each session.

In each game, each icon was associated with a certain positive number of points, and several special pairs of icons were associated with separate score bonuses or penalties that captured interactions between icons. The score for a team was computed by summing the individual point values for each icon, and then adding or subtracting the value of any special pairs present. The pairs did not overlap, and the distribution was designed to be challenging: pairs which gave large positive bonuses were distributed among icons with small individual point values, and pairs which gave large negative penalties were generally found among icons with large individual point values. There was a greater number of positive interactions than negative ones, to give the score distribution a larger upper tail. For ease of comparison and analysis, all scores were normalized to the range [0,1] according to the minimum and maximum possible scores. The combinations of individual and pair values described above resulted in the probability distribution of scores among all possible teams shown in Fig. 2. Participants were not given explicit information about the maximum score, the score distribution, or the position of the interaction terms. The icons' display position and associations with the point distribution were shuffled randomly for each game, so that their appearance and placement in the display did not give clues as to their point values during the course of an experiment session.

Dependent Variables and Definitions

In each round, the following data were recorded for each player: the icons (*choices*) on the team at the end of the round, the *source* of each icon, and the resulting score. The *source* information indicated whether an icon was unchanged from the previous round (Retained), copied from the player's own best-scoring team so far (Retrieved), chosen from the league display (Innovated), or copied from another player's team (Imitated). When Imitation was chosen, the persistent identifier of the copied player was recorded to allow further analyses of imitation decisions. In the case of a player replacing the entire team with Imitated icons, only the choices that were not already present on the

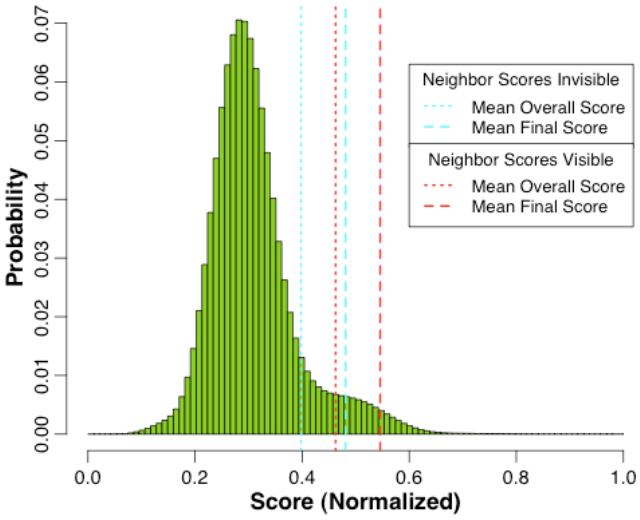


Figure 2: Distribution of scores for all possible teams.

team were counted as Imitated. Similar criteria applied to replacement of an entire team with Retrieved icons, or removing an icon and then returning it to the team via an Innovation choice. *Choice similarity* was defined as the proportion of icons that two teams have in common. An *improvement* was defined as an instance of a participant obtaining a score higher than all prior scores of all players within a particular game. Each participant's normalized *improvement share* was defined as their individually achieved proportion of the total improvements achieved by all participants in a condition, multiplied by the number of participants. A value of 1 indicated a "fair" share, e.g. a participant achieved one-third of the improvements in a three-person session. *Guess diversity* for a group in a particular round was defined as the proportion of icons in the league represented on one or more participants' teams in that round. This value was normalized by the average expected value of this proportion for each participant group size, generated by a Monte Carlo simulation.

Results

Differences in Performance

Grouped participants achieved mean overall (across all rounds) and final normalized scores of .398 and .481 respectively in the invisible-scores condition, and significantly higher scores (.463 and .546) in the visible-scores condition (see Figure 2). Isolated participants achieved mean overall and final scores of .356 and .395. Linear mixed-effects models revealed that score increased significantly with group size in the visible-scores condition ($F(1,63)=79.75$, $p<.0001$, $B=0.354$), as well as in the invisible-scores condition ($F(1,63)=14.94$, $p=.0003$, $B=0.129$), though the latter trend was not as strong. Of all grouped participants, 81.7% had higher mean scores in the visible-scores condition than in the invisible-scores condition (see Figure 3).

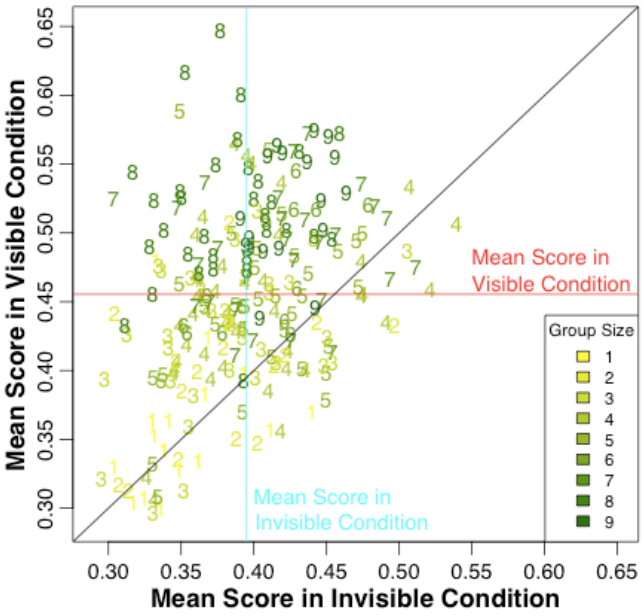


Figure 3: Scattergram of individuals' mean scores in each condition, labeled with their participant group size.

Linear mixed-effects models were used to examine trends across rounds for score and guess diversity, with a random effect of participant group. Analysis of score versus round showed a strong positive trend in the visible-scores condition ($F(1,1494)=295.96$, $p<.0001$, $B=.534$, mean increase=0.188), and a slightly shallower positive trend in the invisible-scores condition ($F(1,1494)=251.93$, $p<.0001$, $B=.615$, mean increase=0.145; see Figure 4). Guess diversity showed a similarly strong decrease across rounds in the visible-scores condition ($F(1,1126)=304.78$, $p<.0001$, $B=-.443$, mean change=-0.468), and a weaker decrease in the invisible-scores condition ($F(1,1126)=97.31$, $p<.0001$, $B=-0.453$, mean change=-0.271; see Figure 4).

Grouped participants achieved an average of 1.21 improvements per person in the visible-scores condition, and 0.95 in the invisible-scores condition. Isolated participants achieved an equivalent average of 2.44 improvements per person. Mean proportions of each choice source for improvement and non-improvement guesses in each condition are shown in Table 2. In both conditions, the proportion of Innovation choices was higher for guesses that yielded improvements relative to non-improvements (invisible-scores: $t(733.20)=-14.03$, $p<.0001$; visible-scores: $t(907.73)=-17.14$, $p<.0001$). In the invisible-scores condition, the proportion of Imitation choices was significantly lower for improvements than non-improvements ($t(916.77)=11.54$, $p<.0001$), while in the visible-scores condition, the proportion of Retention choices was significantly lower for improvements than non-improvements ($t(916.33)=9.34$, $p<.0001$). Overall there was significantly higher Retention in the visible-scores condition ($t(360)=-2.218$, $p=.027$, indicating that guesses changed more slowly.

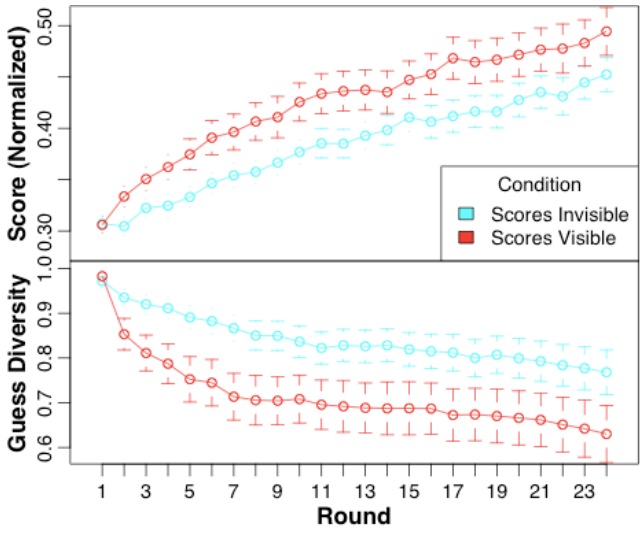


Figure 4: Change in score and guess diversity across rounds in each condition.

Analyses of relationships between mean individual score and mean individual choice source proportions showed a strong negative correlation in both conditions between score and prevalence of Innovation choices (invisible-scores: $F(1,196)=64.16$, $p<.0001$, $B=-0.497$; visible-scores: $F(1,196)=153.5$, $p<.0001$, $B=-0.663$) and a strong positive relationship between score and Retention (invisible-scores: $F(1,196)=15.27$, $p=.0001$, $B=0.269$; visible-scores: $F(1,196)=62.87$, $p<.0001$, $B=0.493$), while a strong positive relationship was shown for Imitation only in the visible-scores condition ($F(1,196)=9.70$, $p=.002$, $B=0.217$), and a strong positive relationship was shown for Retrieval only in the invisible-scores condition ($F(1,196)=14.28$, $p=.0002$, $B=0.261$).

Of all improvements in the invisible-scores condition, 14.5% resulted from guesses that included Imitation, versus 28.4% in the visible-scores condition. In a large majority (>70%) of those cases across both conditions, the focal player imitated at least one peer who had previously imitated the focal player. In other words, a player who was imitated by another player often later imitated the same player in the course of creating an improvement.

Table 2: Mean choice source proportions for (non-) improvement guesses in each condition. (Significant differences within a condition are in **boldface**, and significant differences between conditions are in *italics*.)

Condition	Improvement?	Imit.	Innov.	Retain	Retr.
Invisible Scores	No	**.100**	**.133**	*.712*	**.044**
	Yes	**.039**	**.216**	*.705*	**.035**
Visible Scores	No	.091	**.114**	***.763***	.022
	Yes	.082	**.194**	***.695***	.021

Normalized improvement share showed a relatively equitable distribution of improvements within groups in the visible-scores condition, with the distribution peaked near a "fair" share of 1. In the invisible-scores condition, however, the distribution had a strongly inequitable skew, with a modal share of zero (see Figure 5). A Kolmogorov-Smirnoff test of equality of distributions indicated that these distributions were significantly different (D=0.171, p=0.006). Mean overall score showed a strong positive correlation with improvement share in the invisible-scores condition (F(1,148)=34.94, p<.0001, B=0.329), but this relationship was not evident in the visible-scores condition.

Differences in Strategy

In the visible-scores condition, approximately 79% of imitation events were of the highest-scoring player, while in the invisible-scores condition, all players were approximately equally likely to be imitated with regard to score. A comparison between the mean choice similarity of participants' most recent guesses to those whom they imitated, and to those whom they did not imitate, revealed a slight but significant positive difference in the visible-scores condition: a similarity value of .563 for imitated and .524 for non-imitated guesses (t(5084.88)=-5.47, p<.0001). The opposite was true in the invisible-scores condition: .317 for imitated and .346 for non-imitated guesses (t(4041.53)=4.02, p<.0001). In other words, when scores were visible, imitation was biased toward similar guesses, and when scores were invisible, imitation was biased toward dissimilar guesses.

In order to measure the bias of participants to choose an icon according to its frequency in peers' teams, we tallied the number of players in the group whose teams included each icon in the previous round (N_{R-1}), as well as the

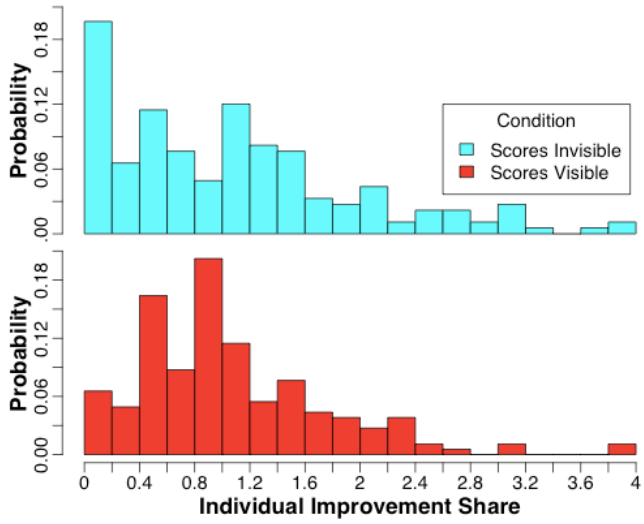


Figure 5: Histograms showing relatively equitable achievement of improvements within groups in the visible-scores condition, and an inequitable distribution in the invisible-scores condition.

number of the remaining players who added it to their team in the current round via Imitation. To convert these figures to normalized frequencies, the first number was divided by the participant group size (N), and the second number was divided by the number of participants who did not possess the icon in the previous round ($N - N_{R-1}$). If a participant had decided to imitate an icon at random from among all neighbors' teams, a certain chance correlation with choice frequency would be expected simply because more high-frequency icons are present. However, a linear mixed-effects analysis of imitation probability versus choice frequency showed a positive frequency bias that was significantly greater than chance in the visible-scores condition (F(1,604)=943.25, p<.0001, B=.741) and significantly below chance in the invisible-scores condition (F(1,604)=231.67, p<.0001, B=.470). This indicates that in the visible-scores condition, participants were biased toward imitating higher-frequency icons at a rate greater than expected by chance, but not in the invisible-scores condition.

Discussion

When scores were visible, participants were heavily biased toward imitating higher-performing peers (displaying the *copy the best* strategy discussed in Laland (2004)), and performance was correlated with the average amount of Imitation in a participant's choices. Participants also showed a bias toward imitating solution elements that were possessed by larger proportions of their fellow participants, similar to the *copy the majority* strategy. Another bias evident in the score-visible condition was toward imitating more similar guesses, which allowed the imitator to make use of social learning while keeping a solution partially compatible with previous solutions and existing knowledge of the problem space, a phenomenon explored in studies of innovation propagation (Rogers, 2003).

As expected, hiding other participants' score information strongly impeded social learning: when others' scores were not visible, the choice of whom to imitate was approximately random with respect to score, and performance was correlated with the average amount of Retrieved information on a participant's team, showing the incentive to focus on previously-acquired information rather than that of others. Surprisingly, participants in the score-invisible condition also seemed to be slightly biased against peer solutions that were similar to their own, as well as icons which were more popular among their peers, perhaps indicating a bias toward novelty, which would help explain the overall decrease in individual Retention in this condition.

However, participants in the invisible-scores condition still showed a slight bias toward imitating more popular icons, indicating that the lack of score information did not cause them to disregard the guesses of their fellow players entirely. Though it conflicts with the finding that imitation in this condition occurred without regard to score, this may explain some of the improvements using Imitation and the

positive relationship of score with participant group size in this condition. When players have relatively high incentives to explore for themselves rather than imitate, and yet have some solution elements in common, it is reasonable to conclude that those common solution elements may produce good scores. This is also consistent with many participants' self-reported strategies.

As seen in the increasing score and decreasing guess diversity trends across rounds, average performance increased via the convergence of group members on regions of the problem space that contained high-quality teams. This convergence combined with a small amount of individual exploration caused such regions to be explored more thoroughly and still better solutions to be found. However, in the invisible-scores condition, when imitation was not focused on a small group of better-performing neighbors (because performance information was not available), or similar guesses, this convergence happened much more slowly, search was more diffuse and less efficient, and lower performance resulted.

The significant correlation of improvement share with mean scores in the score-invisible conditions shows that individuals who were relatively more successful at individual exploration were rewarded with proportionately better overall scores compared to others, because their fellow players could not easily copy their improvements and achieve their scores. In the score-visible conditions this relationship disappeared, but mean scores increased significantly such that nearly all participants did better.

In other words, when social learning was unimpeded in the visible-scores condition, high and low individual achievers had approximately the same payoffs, but absolute payoffs were higher for both compared to the invisible-scores condition. This is because imitators were not merely scroungers; the substantial proportion of Imitation present in improvements shows that imitated guesses were often the basis for further cumulative innovations. The cumulative innovation hypothesis is supported by the fact that a large proportion of improvements which used Imitation involved mutual Imitation and improvement, in which solution elements were passed between players via copying and built into better solutions in the process. This enabled a more equitable sharing of the "labor" of producing improvements, and produced more improvements overall.

Gabriel Tarde, one of the founders of social psychology, considered innovation and imitation to be "the fundamental social acts" (Tarde 1903/1969). Cultural conventions can be thought of as a form of large-scale imitation of behaviors that evolve along with their associated populations, subject to accompanying adaptive pressures (Boyd & Richerson, 2005). Innovations are necessary to adapt to the challenges of changing environments, and when members of a group imitate them, adaptive solutions to problems can be effectively preserved within a culture.

The findings of this study point to new avenues for understanding how innovations are generated and spread, as well as how information, incentives and the dynamic behavioral interactions of individuals create higher-level consequences for the groups to which they belong.

Acknowledgements

The authors would like to thank Xianfeng Song, Zoran Rilak, and Todd Gureckis for their help in designing and programming the experiment, and Frances Kidwell and Bennis Pavisian for their help with running the experiment sessions. This work is funded by National Science Foundation REESE grant 0910218 and a National Science Foundation IGERT traineeship.

References

Bandura, A. (1965). Vicarious processes: a case of no-trial learning. In L. Berkowitz (Ed.) *Advances in Experimental Social Psychology, Vol. II*. New York: Academic Press.

Boyd, R., Richerson, P. J. (1995). Why Does Culture Increase Human Adaptability? *Ethology and Sociobiology, 16*, 125-143.

Boyd, R. & Richerson, P. J. (2005). The origin and evolution of cultures. New York: Oxford University Press.

Galef, B. G. Jr., & Giraldeau, L. A. (2001). Social influences on foraging in vertebrates: causal mechanisms and adaptive functions. *Animal Behaviour, 61*(1), 3-15.

Höglund, J., Alatalo, R. V., Gibson, R. M., & Lundberg, A. (1995). Mate-choice copying in black grouse. *Animal Behaviour, 49*(6), 1627-1633.

Kameda, T., & Nakanishi, D. (2002). Cost-benefit analysis of social/cultural learning in a non-stationary uncertain environment: An evolutionary simulation and an experiment with human subjects. *Evolution and Human Behavior, 23*, 373-393.

Kameda, T., & Nakanishi, D. (2003). Does social/cultural learning increase human adaptability? Rogers's question revisited. *Evolution and Human Behavior, 24*, 242-260.

Krause, J. (1993). Transmission of fright reaction between different species of fish. *Animal Behavior, 65*, 595-603.

Laland, K. N. (2004). Social learning strategies. *Learning & Behavior, 32*(1), 4-14.

Rogers, A. R. (1988). Does biology constrain culture? *American Anthropologist, 90*, 819-831.

Rogers, E. M. (2003). *Diffusion of Innovations* (5th ed.). New York: Free Press.

Tarde, G. (1969). *The Laws of Imitation.* (Elsie Clews Parsons, Trans.). Chicago: University of Chicago Press. (Original work published 1903.)

Templeton, J. J., & Giraldeau, L.-A. (1996). Vicarious sampling: the use of personal and public information by starlings foraging in a simple patchy environment. *Behavioral Ecology and Sociobiology*, *38*, 105-113.

Wisdom, T. N., Song, X., & Goldstone, R. L. (manuscript submitted for publication). The effects of peer information on problem-solving in a networked group.

Social Context Effects on the Impact of Category Labels

Rachel G. Stephens (rachel.stephens@adelaide.edu.au)
Amy Perfors (amy.perfors@adelaide.edu.au)
Daniel J. Navarro (daniel.navarro@adelaide.edu.au)
School of Psychology, University of Adelaide
Adelaide, SA 5005, Australia

Abstract

We explore whether social context affects how labels (relative to other features) affect category learning. We taught 104 participants four novel categories using a feature inference task. In a between-participants design, we manipulated: 1) the social context of the task (social context vs. on the computer); and 2) which dimension of the category members could be used to perfectly predict the target feature: the category label, a biased feature (which is salient and already associated with the target feature in the correct way) or a non-biased feature (which is less salient and not already associated with the target feature in any way). Learning curves were used to assess whether participants assumed that labels were uniquely helpful compared to other features. The results suggest that the extent to which labels are privileged depends on the context in which the category learning task is presented. When the task is social, people learn quickly regardless of whether a label or another feature is the most informative. When the task is not, both novel labels and biased features are more useful than non-biased features.

Keywords: categorization; feature inference; labels; features; social context.

Introduction

There is a Chinese proverb that says, "The beginning of wisdom is to call things by their right names." Category labels feel like an important and special part of our conceptual knowledge. We need labels to communicate about classes of objects, and labels (often unlike other features) are a property that all members of a category share. One might expect that labels help learners pick out the category members that have important similarities to each other, and that are different from members of other categories. However, although this has been the topic of study for decades, it remains unclear whether there is a psychological distinction between category labels and other types of features. It is also unknown (especially for adults) whether the effect of a label is affected by the social-referential context in which it is offered.

Work with children broadly supports the notion that labels have a privileged psychological status, although it is still debated what the source of that privilege is. Verbal labels appear to facilitate infant category learning (e.g., Balaban & Waxman, 1997), with shared names highlighting commonalties between objects (Waxman & Braun, 2005). Labels influence the number of categories formed by infants, overriding the categories that are suggested by perceptual similarity (Plunkett, Hu & Cohen, 2008). Additionally, when making decisions about whether a category feature can be generalized to a new object, preschool children rely more on category membership conveyed by a label than they do on perceptual similarity (Gelman & Markman, 1986; 1987). These experiments suggest that labels are special in some way, but these studies do not address whether a single salient feature possessed by all category members might produce the same effects.

The question is further complicated by the fact that, for children at least, the social and linguistic context influences the effects of category labels. Fulkerson and Haaf (2003) found that labels can help infants to form categories that are otherwise *not* formed when only a non-labeling sound or no sound is used in their place. However, for older infants (15 rather than 9 months) the source of the label matters: these infants formed categories when the labels were presented orally, but not when they were presented by a voice recorder. Consistent with this, Campbell and Namy (2003) found that infants learned object names only when the label was presented in a normal social-referential interaction. The names were learned when the label was verbalized by the experimenter and embedded in a familiar naming routine, such as, "Look at what you have! *Tillen*. That's what we call that one." Learning was unsuccessful when the label was emitted from a baby monitor and was not timed with the naming routine.

It is unclear whether we should expect similar effects of social context in adults, or whether any differences between adults and children are due to differences in the social context of label presentation. Unlike for children, most adult category learning experiments do not incorporate a social element: category labels are presented in written form on screen or on paper, or (at most) as a recorded sound. Grice's conversational maxims (Grice, 1975) suggest that labels presented in a social context should especially be presumed to be relevant and informative to the task at hand. Alternatively, since educated adults are well practiced at using labels, social context may not significantly change how labels are treated.

Labels do seem to play an important role in adult categorization. Categorical perception research suggests that learning that stimuli share a label can be sufficient to increase perceptions of similarity of the stimuli (e.g., Goldstone, Lippa & Shiffrin, 2001). In addition, Yamauchi and Markman (1998; 2000a) have begun to directly address the issue of whether category labels have a privileged status over other features for adults. They claimed that people employ different strategies to make feature inferences or classifications (i.e., infer labels), and that learning novel

categories through a classification task or a feature inference task will produce different category representations. However, the experimental tasks used were unfair as a test of a general distinction between labels and other features, because the category structures of the classification and inference tasks were not equivalent: labels were a diagnostic feature in the inference task, which seemed to drive the differences between conditions (see also Johansen & Kruschke, 2005). However, the studies could suggest that people expect labels to be useful in a feature inference task (see also Yamauchi, Love & Markman, 2002).

In a further series of experiments Yamauchi and Markman (2000b) showed adults a set of labeled exemplars of two categories and asked them to compare novel stimuli to the exemplars. Classifications of the novel stimuli were generally made according to the total number of features consistent with the appropriate category prototype, but feature inferences for novel stimuli were strongly influenced by the observed category label. As a result, when similarity and category membership were placed in opposition, participants were more likely to base their inferences on the label. This effect was decreased when the labels referred to a feature rather than to category membership, or when the label was replaced by a perceptual feature. These studies suggest that to the extent that labels convey category membership, they are privileged over other features. However, the experimental tasks used in these experiments were fairly unnatural, since participants did not have to learn the categories: they simply compared stimuli on a sheet in front of them.

This work has two aims. First, it contributes one of the first explorations in the adult literature focused on the question of whether social context has an impact on the status of labels. Second, it investigates whether labels have a privileged status over other category features for adults. Are people biased to assume that labels are uniquely helpful compared to other stimulus features when learning about novel categories? If so, they should assume that labels are important to pay attention to and therefore categories should be easier to learn when the labels are useful predictors. However, categories should be more difficult to learn when another feature is the more useful predictor (depending on the type of feature). Are these effects mitigated or amplified depending on the nature of the social context in which the labels are presented? Are labels assumed to be especially important in a social category learning context, involving communication with a knowledgeable human teacher?

Method

Participants learned about four novel categories during a feature inference task. Two between-participants experimental factors were manipulated to form a 3x2 design. The first factor was which aspect or dimension of the category members could be used to perfectly predict the target feature. This diagnostic feature dimension could be the CATEGORY LABEL, a BIASED FEATURE or a NON-BIASED FEATURE. The second factor was whether the category learning task was performed alone on a personal computer (PC), or in a more social context with the experimenter (SOCIAL).



Figure 1: Two sample images used in the category learning task. The image on the right includes the feedback of the hammer.

Participants

106 adults (either undergraduates at the University of Adelaide, or people recruited from the general community; 41 males) took part in the experiment. Ages ranged from 18 to 57 years. They received course credit or AU$10. One participant's data was removed from the SOCIAL, NON-BIASED FEATURE condition because the participant withdrew from the study before training was completed. An outlier was removed from the PC, LABELS condition for taking 26 blocks to complete the training task (this was more than 3 SDs above the mean for that condition). 16 to 18 people remained in each condition.

Materials

The category learning task was designed to be realistic and engaging, in order to encourage ecologically valid responses. Participants learned about four novel "alien people" categories, each of which contained four members. Images for the categories were created using *World of Warcraft*, an online computer game produced by Blizzard. Examples of the images are shown in Figure 1, and the category structure used across all conditions is shown in Table 1.

Participants were asked to predict the nature of a certain target feature: which item each alien wanted to buy (options were a timber axe, a dagger, a hammer or a staff). The four category members varied on five dimensions, which could each take one of four values. The five dimensions and their possible values were:

1) category labels, presented as community names: Goloth, Bragen, Lathor and Durgal
2) clothing: leather warrior-like garb, a robe, tradesperson-like overalls and "lumberjack" attire
3) hair style: long, cropped, bald and ponytail
4) skin color: red, cream, brown and blue-grey
5) facial hair: short square beard, long plaited beard, medium pointed beard and broad beard with upturned moustache.

Table 1: Category structure for the feature inference learning task, for all conditions. The diagnostic feature dimension perfectly predicts the target feature dimension, and demarcates the four categories. (F = feature)

Target features	Diagnostic features	F1	F2	F3	F4
Timber axe	1	1	3	2	1
	1	1	1	3	2
	1	2	1	1	3
	1	3	2	1	1
Dagger	2	2	4	3	2
	2	2	2	4	3
	2	3	2	2	4
	2	4	3	2	2
Hammer	3	3	1	4	3
	3	3	3	1	4
	3	4	3	3	1
	3	1	4	3	3
Staff	4	4	2	1	4
	4	4	4	2	1
	4	1	4	4	2
	4	2	1	4	4

As Table 1 demonstrates, four of the feature dimensions contributed to a family resemblance category structure, but one "best predictor" or diagnostic dimension perfectly predicted the target feature. Thus all conditions had rule-based categories: only a single dimension was needed to solve the categorization problem. One experimental factor was *which* dimension was the best predictor of the target feature. In the LABEL conditions, the community name was the diagnostic dimension. Thus participants could learn to perfectly predict which item an alien wanted to buy using only its community name. Alternatively, in the BIASED FEATURE conditions, the clothing was the diagnostic dimension. In these conditions, the clothing "value" corresponded to the target item one might expect based on prior background knowledge: the aliens wearing the robe wanted the staff, the aliens with the leather garb wanted the dagger, the aliens in the overalls wanted the hammer, and the aliens with the "lumberjack" attire wanted the timber axe. Finally, in the NON-BIASED FEATURE conditions, the facial hair was the diagnostic dimension, arbitrarily matched with the target items.

Procedure

Participants were randomly allocated to one of the six conditions. All participants were asked to imagine they were a space traveler who began working in a general store on another planet and needed to learn about the customers of the store. Participants were told that they needed to learn to predict which item each of 16 customers wanted to buy. They learned by trial and error, and the learning task continued until they made the correct prediction for all 16 customers. This criterion was chosen to encourage optimal performance: participants knew that the task would continue until no errors were made.

On each block the 16 trials were presented in random order. For each trial, participants were presented with an image of a customer with a label. In the SOCIAL conditions, the experimenter displayed a card with the image and verbally presented the label (e.g., "This is a Goloth"). In the PC conditions, the label was written in bold blue capital letters above the image on the screen. Participants were then asked to predict the target feature, either verbally to the experimenter in the SOCIAL conditions, or by clicking the appropriate button on the screen in the PC conditions. They were given immediate corrective feedback after each trial consisting of an image of the customer holding the correct item, with the community name written above the image and the name of the correct target item below the image. Participants also received additional feedback after each block of 16 stimuli about their total number of correct responses for that block.

Design

There were two factors of interest in this experiment. One factor was whether the target feature dimension was best predicted by: 1) the community LABEL; 2) the NON-BIASED feature (facial hair), which participants should not have expected to be useful *a priori*; or 3) the BIASED feature (clothing), which people should have had a prior bias to find useful for predicting what the creatures wanted to buy, since the sets of clothing each corresponded to the appropriate target item. This experimental factor tests whether people are biased to assume that labels are uniquely useful features, or whether they are similar to highly salient or useful features (like the BIASED feature). In each level of the factor, the diagnostic feature dimension plays an identical role in the category structure, allowing a fair test of the relative status of labels and features. If labels are not special, all else being equal, learning performance should be equivalent regardless of whether labels or other features are the diagnostic dimension. However, if labels have a special status, participants should be quicker to learn to predict the target feature when the LABEL is the diagnostic dimension. Learning should be slowest in the NON-BIASED FEATURE conditions, where labels are less useful and an unexpected feature *is* useful. Learning should be more rapid in the BIASED FEATURE conditions, where an unsurprising feature *is* useful, and it is relatively easy to remember which particular feature value (i.e., particular outfit) matches with each target item. Of critical importance, then, is whether learning in the LABEL conditions is closer to learning in the BIASED or in the NON-BIASED FEATURE conditions.

The other experimental factor was whether the labels were presented in a social context or not. In the PC conditions, participants worked on a computer; the labels and images were presented on the screen. In the SOCIAL conditions, participants learned by interacting with the experimenter, who presented the images on cards and verbally presented the category labels. This manipulation tests whether people assume labels to be particularly special when the context is more social and interactive.

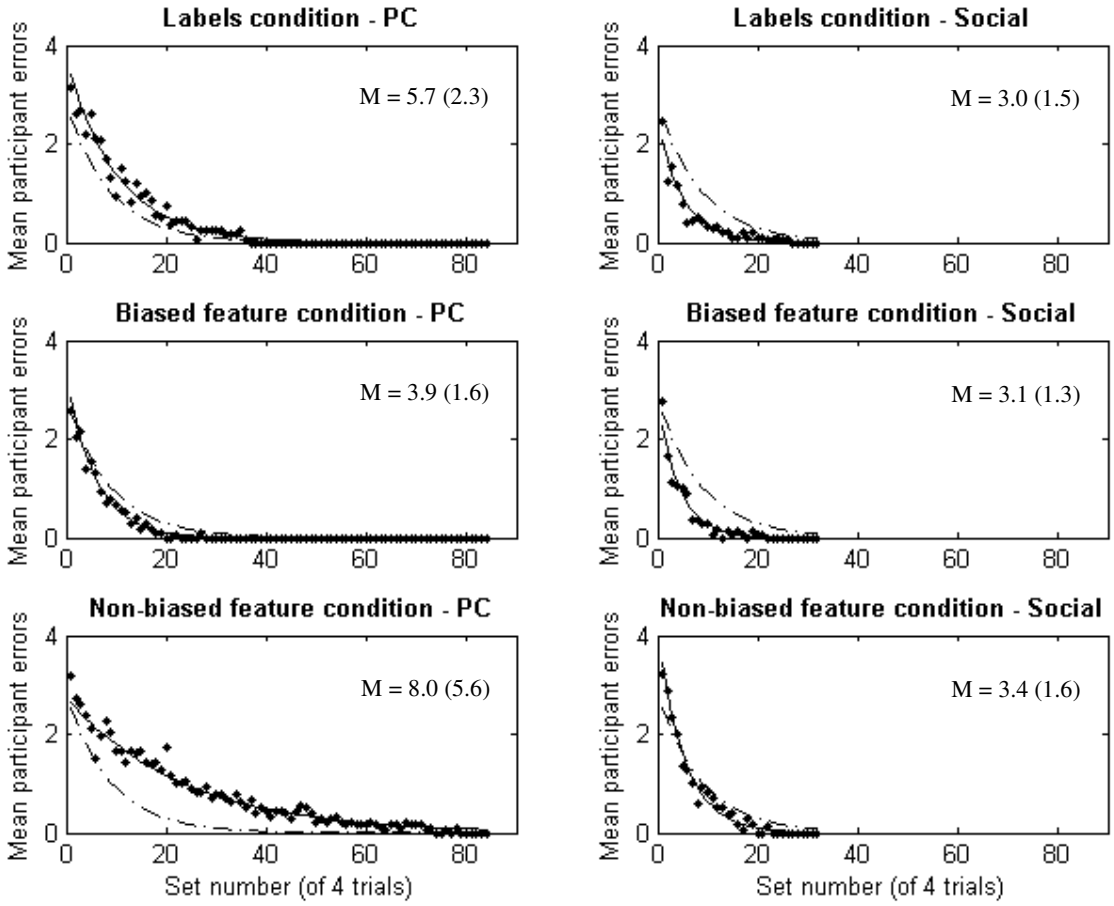


Figure 2: Learning curves averaged across sets of 4 trials, and averaged across participants in each condition. Two models using an exponential function are fit to the error data using BIC: a model with unique parameter values for each condition (solid line) is preferred over a model with the same two parameter values for all conditions (dashed line). M = the mean (SD) number of blocks taken to complete the learning task.

Results

Our results suggest that the extent to which labels are privileged depends on the context in which the task is presented. When the task is social, people learn quickly regardless of the nature of the diagnostic dimension. When it is not, labels are as useful as biased features.

Figure 2 shows the learning curves with error data averaged across sets of four trials, and across all participants in each condition (the black dots). Exponential functions of the form $Y = a \exp(-bX)$ (see Heathcote, Brown & Mewhort, 2000) were fit to the data of each condition. Model parameters were fit using maximum likelihood estimation, and Bayesian Information Criterion (BIC; Schwarz, 1978) was used for model selection. BIC is a measure that considers the data fit, but penalizes models for having excessive parameters (Myung & Pitt, 1997). The model that minimizes BIC should be preferred.

Two (of several[1]) simple models that were fit to the mean error data are also shown in Figure 2. Both models used the exponential function, but one model allowed unique parameter values for each condition, while the other model used the same two parameter values for all conditions. These two models were compared, using BIC to estimate the Bayes Factor (see Myung & Pitt, 1997), which provides the odds in favor of the model with the lower BIC score. The model that allowed unique parameter values was preferred according to this criterion (BIC for unique parameters model = 186.0 vs. BIC for same parameters model = 1499.1; according to the Bayes Factor approximation, a difference between BIC scores of such a large magnitude translates to extremely strong evidence in favor of the full model). This suggests that each condition

[1] Other models that were compared with the "unique parameter values" model to test for interaction effects were also found to be inferior according to BIC. (For instance, a model that allowed the PC, NON-BIASED FEATURE condition to have different parameter values to the other five conditions.)

had different learning curves – that is, that participants did not behave identically in each condition.

As Figure 2 demonstrates, the main effect was that learning was faster overall when the task occurred in a social context than when it was presented on a computer. Presenting the category learning task in a social context led to improved learning performance overall[2]. It seems that participants were much more engaged, and thus solved the task quite quickly across all three of the SOCIAL conditions.

There was a more pronounced difference between the LABEL, BIASED FEATURE and NON-BIASED FEATURE conditions when the stimuli were presented on the PC than when they were presented in a social context. That said, in both the SOCIAL and PC conditions, learning was slowest in the NON-BIASED FEATURE conditions, and fast in the BIASED FEATURE conditions, as expected. Of primary interest is to compare performance in the LABEL conditions with that of the other conditions. For both the SOCIAL and PC conditions, learning speed in the LABEL condition was closer to that of the BIASED FEATURE condition than to that of the NON-BIASED FEATURE condition. However, while on the PC, learning in the LABEL condition was slower than learning in the BIASED FEATURE condition. In contrast, in the SOCIAL conditions, the learning curves of LABEL and BAISED FEATURE conditions were very similar. This suggests a weak effect that labels were more privileged in the social context than on the computer. Nonetheless, in either context, diagnostic labels did not help category learning beyond help that could be given by a diagnostic biased feature.

Why was learning not *fastest* in the LABEL conditions? Let us consider the difference in the learning task between the LABEL and BIASED FEATURE conditions. To successfully complete the category learning task, participants in all conditions needed to: 1) notice that one particular feature dimension was diagnostic (e.g., the labels); and 2) learn the match between each particular diagnostic feature value and a target feature value (e.g., that "Bragens" wanted the dagger, and "Lathors" wanted the hammer). However, the BIASED FEATURE condition was easier: the diagnostic feature dimension (clothing) was not only salient and meaningful (and thus easy to notice), but each outfit also meaningfully corresponded to an item (e.g., the robe outfit matched with the staff). Thus, participants could essentially come to the task already knowing the correct answers. The LABEL condition was actually a more difficult task, because although the diagnostic feature dimension was perceptually salient, people needed to learn an arbitrary match between the novel names and the target features. It is interesting that despite the added difficulty, learning in the LABEL conditions was similar to that of the BIASED FEATURE conditions. However, the added difficulty might explain why learning in the LABEL conditions was not fastest.

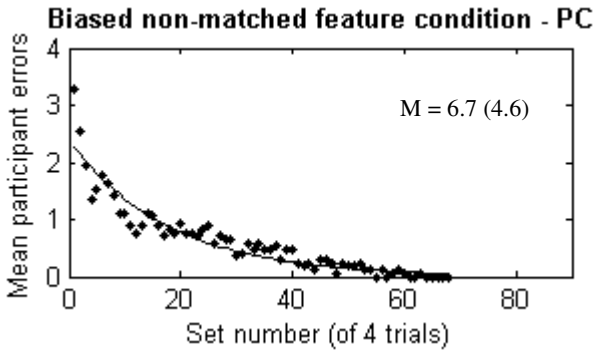


Figure 3: Learning curve averaged across sets of 4 trials, and averaged across participants in this condition. The exponential function is fit to the error data.

To determine whether this explains the observed similarity between the LABEL and BIASED FEATURE conditions, we ran another experimental condition on the PC with 17 new participants[3]. The category learning task was identical[4] to that of the PC, BIASED FEATURE condition, with clothing as the diagnostic feature dimension; however, the feature values no longer matched with the expected target feature values. Counterintuitively, the "lumberjack" wanted the dagger, the "warrior" wanted the hammer, the "tradesperson" wanted the staff, and the "wizard/priest" wanted the timber axe. This new condition still had the salient and expectedly meaningful clothes features as the diagnostic dimension, but the match between feature values was arbitrary. As Figure 3 shows, learning was slowed nearly to the same level as in the PC, NON-BIASED FEATURE condition. This suggests that the fast learning in the BIASED FEATURE condition was due to the pre-existing knowledge of the mapping between outfits and target items. Learning could be fast in the LABEL condition because the names were completely novel and "blank"; unlike the clothes in the new condition, the LABEL condition did not require any unlearning of associations between the diagnostic dimension and the target feature dimension.

Discussion

We set out to explore whether social context has an impact on the status of category labels for adults. We investigated whether people would pay special attention to labels presented in a social context, and hence learn quickly when these labels *are* most informative. While there was some suggestion that labels were more privileged in a social context than they were on a computer, the main result was

[2] Exponential functions for LABEL, BIASED FEATURE and NON-BIASED FEATURE conditions were $Y = 3.75\exp(-0.10X)$, $Y = 3.37\exp(-0.17X)$, and $Y = 2.76\exp(-0.04X)$ for the PC conditions, and $Y = 2.53\exp(-0.20X)$, $Y = 2.90\exp(-0.24X)$, and $Y = 4.14\exp(-0.19X)$, for the SOCIAL conditions, respectively. Note that the two parameters vary between conditions; larger a indicates more errors at the beginning of training and larger b indicates faster learning.

[3] Participants were undergraduates at the University of Adelaide, or recruited from the general community (9 males). Ages ranged from 17 to 38 years. Participants received AU$5.

[4] Feedback was slightly different in this condition, due to the availability of images: participants did not see an image of the correct target item being held by the alien creature.

that people can solve a category learning task much faster when they are in an engaging, social context, regardless of whether a label or another feature is actually more informative. We suspect that the participants were more motivated to do well in the presence of a human teacher and with a more enjoyable, interactive task. When the task is not social, novel labels are privileged over non-salient and arbitrarily-matched features, but are no more useful than biased features. Nonetheless, it is interesting that learning in the LABEL conditions was fast, despite participants having to learn an arbitrary match between each novel label and the target feature. Presumably, if the labels were *not* novel and were appropriately matched (e.g., "Wizard" and "Warrior"), the task would become trivially easy and participants would learn even faster.

The results of this study support an intermediate view between labels being "just another feature" and having a unique, privileged status. Novel labels can aid category learning better than arbitrary (or biased but arbitrarily matched) features can: labels are salient, and novel labels permit new associations with features to be learned without being hindered by knowledge about existing feature associations. However, context matters: if people are already fully engaged with the category learning task, there is less scope for labels to aid learning beyond other features. One caveat to this finding is that perhaps the influence of labels in the social context was somewhat hidden by a ceiling effect, since our rule-based category structure was a simple one, quickly solved by most participants. More challenging category structures may reveal a larger influence of labels within a social context.

Further work is required to determine whether adults *process* labels differently to other features, or simply weight them more heavily. Gliozzi, Mayor, Hu and Plunkett (2009) contrast an *unsupervised feature-based account* and a *supervised name-based account* of category formation. According to the latter account, objects given the same name belong to the same category, and so labels act as invitations to form categories and highlight commonalities between objects. The unsupervised feature-based account says that labels have the same status as other features. Labels may vary in salience, just like other features, but are handled with the same statistical inference processes as are other features. The model by Gliozzi et al. (2009) suggests that for infants, labels play a mundane but powerful role as simply additional features. Our study, and the experiments by Yamauchi and Markman (e.g., 1998; 2000a) are consistent with this view.

Finally, this experiment has implications for adult category learning studies that are not presented in an engaging, social context. We found that presenting a category learning task in the absence of a social context does not encourage optimal learning behavior. Category learning on a computer may not reflect category learning in the more engaging situations typically encountered in real life, so it is worth understanding category learning in more naturalistic, social contexts.

Acknowledgments

RGS was supported by an Australian Postgraduate Award and DJN was supported by an Australian Research Fellowship (ARC grant DP-0773794). Thank you to the reviewers for their helpful suggestions.

References

Balaban, M. T., & Waxman, S. R. (1997). Do words facilitate object categorization in 9-month-old infants? *Journal of Experimental Child Psychology*, *64*, 3–26.

Campbell, A. L. & Namy, L. L. (2003). The social-referential context in verbal and nonverbal symbol learning. *Child Development*, *74*, 549–563.

Fulkerson, A. L. & Haaf, R. A. (2003). The influence of labels, non-labeling sounds, and source of auditory input on 9- and 15-month-olds' object categorization. *Infancy*, *4*, 349–369.

Gelman, S. A., & Markman, E. M. (1986). Categories and induction in young children. *Cognition*, *23*, 183–209.

Gelman, S. A., & Markman, E. M. (1987). Young children's inductions from natural kinds: The role of categories and appearances. *Child Development*, *58*, 1532–1541.

Gliozzi, V., Mayor, J., Hu, J-F., & Plunkett, K. (2009). Labels as features (not names) for infant categorization: A neurocomputational approach. *Cognitive Science*, *33*, 709–738.

Goldstone, R. L., Lippa, Y., & Shiffrin, R. M. (2001). Altering object representations through category learning. *Cognition, 78*, 27–43.

Grice, H. P. (1975). Logic and conversation. In Cole, P. & Morgan, J. (Eds.) *Syntax and Semantics*, *3*. New York: Academic Press.

Heathcote, A. J., Brown, S. D., & Mewhort, D. (2000). The power law repealed: The case for an exponential law of practice. *Psychonomic Bulletin and Review*, *7*, 185–207.

Johansen, M. K., & Kruschke, J. K. (2005). Category representation for classification and feature inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 1433–1458.

Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin and Review, 4*, 79–95.

Plunkett, K., Hu, J-F., & Cohen, L. B. (2008). Labels can override perceptual categories in early infancy. *Cognition*, *106*, 665–681.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*, 461-464.

Waxman, S. R., & Braun, I. (2005). Consistent (but not variable) names as invitations to form object categories: New evidence from 12-month-old infants. *Cognition*, *95*, B59–B68.

Yamauchi, T., Love, B. C., & Markman, A. B. (2002). Learning nonlinearly separable categories by inference and classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 585–593.

Yamauchi, T., & Markman, A. B. (1998). Category learning by inference and classification. *Journal of Memory and Language*, *39*, 124–148.

Yamauchi, T., & Markman, A. B. (2000a). Learning categories composed of varying instances: The effect of classification, inference and structural alignment. *Memory & Cognition*, *28*, 64–78.

Yamauchi, T., & Markman, A. B. (2000b). Inference using categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 776–795.

Pedagogical Cues Influence Children's Inductive Inference and Exploratory Play

Lucas P. Butler (lpbutler@stanford.edu)
Department of Psychology, Stanford University
Building 420, 450 Serra Mall, Stanford, CA 94305

Ellen M. Markman (markman@stanford.edu)
Department of Psychology, Stanford University
Building 420, 450 Serra Mall, Stanford, CA 94305

Abstract

A hallmark of human cognition is the ability to learn from others—both via language and via non-linguistic cues. Children are sensitive to actions done for their benefit, treating pedagogical acts as conveying important information (Csibra & Gergely, 2009). The current research tapped children's exploration to investigate whether seeing a causal property either demonstrated pedagogically or produced accidentally influences children's expectations about that property's extension to other kind members. Experiment 1 found striking differences in 3- and 4-year-olds' exploration when a property was demonstrated intentionally rather than accidentally. Experiment 2 replicated this effect while also investigating possible influences of the emotional valence of causal events and the salience of property information. These experiments reveal that preschoolers use pedagogical cues to make inferences about generalizability and guide their exploration.

Keywords: Pedagogy; Exploratory Play; Inductive Inference; Causal Learning.

Introduction

One fundamental aspect of human cognition is our ability to learn from and teach others. Our abilities to read others' intentions and engage in collaborative learning may provide the necessary foundation for human culture, from law and government to industry and education (Gergely & Csibra, 2005; Tomasello, 1999). Children's understanding of intentions is inherent to many domains, including word learning (Baldwin, 1991, 1993a, 1993b) and imitation (Carpenter, Akhtar, & Tomasello, 1998; Meltzoff, 1995). Recent work has elaborated on the importance of explicit teaching and demonstration, which fundamentally rely on children's ability to read intentions. Tomasello and Carpenter (2007) argue that "instructed learning" is key to acquiring cultural knowledge, such as what we call objects and how we use them, and Csibra and Gergely (2006, 2009) suggest that humans have adapted a faculty for "natural pedagogy," enabling efficient social learning. On this account, children should treat pedagogical cues (e.g., eye gaze, pointing) as signaling that information is not only important, but that it is culturally agreed-upon and generalizable.

Indeed, pedagogical cues appear to influence processing of information even in infancy. For example, 8-month-olds expect eye gaze to be directed at referent objects when accompanied by pedagogical cues (Csibra & Volein, 2008). Further, pointing leads 9-month-olds to privilege a novel objects' identity over current location in memory (Yoon, Johnson, & Csibra, 2008). And when 14-month-olds see a person pedagogically convey affective information (e.g., disgust) about an object, they treat it as a stable property of the object (Gergely, Egyed, & Király, 2007), and expect others to react similarly towards it (Egyed, Király, Krekó, Kupán, & Gergely, 2007).

Thus, even infants treat pedagogy as communicating important information about novel objects. However, it is as yet unclear whether children take such information as generalizable to a kind, rather than merely to a particular object. Assessing whether novel information should be generalized to a kind is critical in category and concept formation, where children rely on others to impart often otherwise unknowable information (Gelman, 2009; Harris, 2002; Harris & Koenig, 2006). Such knowledge transmission is often linguistic, using language that refers to kinds and categories, and children make a variety of inductive inferences on the basis of kind-referring language. For example, children take labels as referring to kinds that share nonobvious properties, and generalize novel properties on the basis of shared labels—which signal shared category membership—rather than perceptual similarity (e.g., Booth & Waxman, 2002; Gelman & Coley, 1990; Gelman & Markman, 1986, 1987). Moreover, recent work has demonstrated that preschoolers expect novel objects that share a label to share a novel causal property, and selectively explore those objects more when that property fails to extend to additional kind members (Schulz, Standing, & Bonawitz, 2008). Children also understand that information conveyed in a generic statement (e.g., "dogs bark") has greater inductive potential than information conveyed non-generically (e.g., "this dog barks") (Cimpian & Markman, 2008; Gelman, Star, & Flukes, 2002; Hollander, Gelman, & Raman, 2009), and information conveyed generically becomes more central to their kind representations (Cimpian & Markman, 2009).

However, it is important to note that while linguistic cues such as kind labels are powerful in driving generalization, they are not always used pedagogically. One can use object or kind labels without having any intention of pedagogically conveying information, and certainly without intending such information to be taken as generalizable. Thus it might be

important for children to make use of non-linguistic cues when assessing the generalizability of novel information.

Indeed, Gergely and Csibra (2009) suggest that at the core of generic knowledge transmission is pedagogical intent—an intent to explicitly impart new information to a recipient—and that children are sensitive to whether or not information is communicated for the purpose of teaching them something important. Similar to a Gricean view of communication (cf. Clark, 1996; Sperber and Wilson, 1986), in which we expect speakers to be clear and informative, children may infer that when an adult intentionally communicates information for their benefit, it is because the adult intends to teach them something relevant and important, and thus children may use pedagogical cues to gauge generalizability. Given this, we hypothesize that, even given a shared label, children may make stronger inferences about the whether a property is generalizable when it is demonstrated pedagogically, treating it as more conceptually central and inferring that other kind members should share that property.

To test this, our methodology builds on prior research which has established that exploratory play is a window onto children's implicit inductive processes. Having learned that an exemplar of a kind has a causal property, young children, even infants, explore more upon encountering exemplars that share a kind label, but which lack that property (Baldwin, Markman, & Melartin, 1993; Schulz et al., 2008). In the current research, we tapped children's natural exploration to investigate whether, even given objects that share a kind label, they would form different expectations about generalizability depending on whether a novel property was demonstrated intentionally or produced accidentally. If so, then when a property is intentionally demonstrated for them, but fails to obtain for other kind members, children should explore more than when that same property is produced accidentally.

Experiment 1

In Experiment 1 we taught children a name for a novel object, and either intentionally demonstrated or accidentally produced a novel causal property (magnetically picking up paperclips). We then presented children with an identical set of exemplars with the same label but which lacked the property (they were not magnetic), and let them play.

Methods

Participants Thirty-two three-year-olds (16 girls; M = 42 months; range = 36-46 months) and 32 four-year-olds (16 girls; M = 54 months; range = 48-61 months) from a university preschool participated. Children came from predominantly middle- and upper-middle-class families, representing a variety of ethnic groups. Children were randomly assigned to condition, equating for gender and age.

Materials The novel objects were small wooden blocks. The active block had magnetic tape on one end, while the inert blocks had non-magnetic tape. All were covered with black tape, with green tape covering the magnetic/non-magnetic end.

Procedure All children were tested in a private room in their preschool by a trained experimenter. Children first learned a novel label (blicket) for the active block. When asked for the blicket, all children successfully selected it from 4 distracters on two trials, without error.

After learning the word, children did a short distracter task (making paper houses). This served two goals. First, it distanced the word-learning, which was necessarily pedagogical, from the demonstration. Otherwise, children may have remained in a pedagogical "mindset." Second, it provided a plausible excuse for placing a pile of paperclips on the table.

The experimenter then started to clean up the toys. He put away each the distracters saying, "Let's put this away. He then picked up the active block, and again said "Let's put this away," which served as an implicit invitation to attend to the blicket In the intentional condition, he said, "Look, watch this!" He deliberately placed the it on the paperclips, picked it up (with paperclips attached), and looked at it, saying "Hmmm" in a neutral tone. He then placed it next to the paperclips. Next, he placed 10 inert blocks on the table, saying, "here are some blickets." The accidental condition was *identical*, except that the experimenter appeared to "accidentally" drop the block on the paperclips as he was putting it away, exclaiming "Oops!" As in the intentional condition, he picked it up with paperclips attached, looked at it, said, "Hmmm," and placed it next to the paperclips.

The experimenter then told the child to "go ahead and play" while he left the table and sat facing away from the child for 60 seconds. Upon returning, the experimenter introduced a puppet and asked the child, "Can you tell Mr. Monkey about blickets?"

Results

None of the 3-year-olds explored the blickets in the accidental condition (leading to zero variance in that cell of the design), precluding parametric analyses. We analyzed 3- and 4-year-olds' responses separately, using non-parametric Mann-Whitney U and χ^2 tests.

4-year-olds Although there were no differences across conditions in whether or not 4-year-old children explored the blickets, they showed striking differences across conditions in the nature of that exploration, specifically the amount of time they spent exploring and the number of times they tried to elicit the property from the inert blickets. When 4-year-olds saw the property demonstrated intentionally, they spent more time trying to pick up paperclips with the blickets (M = 46.94 s, SD = 21.38) than when they saw it produced accidentally (M = 24.69 s, SD = 25.02), U = 66.0, N = 32, p = 0.019.

Four-year-olds also made more attempts to pick up paperclips with the blickets (M = 9.25, SD = 7.62) in the intentional condition than the accidental condition (M = 2.94, SD = 3.45), U = 61.5, N = 32, p = 0.011.

3-year-olds Three-year-olds showed an analogous effect of condition, which was even starker than for the 4-year-olds. In the accidental condition, zero out of 16 children explored at all, compared with 8 out of 16 in the intentional condition, $\chi^2(1, N = 32) = 10.67$, p = 0.001. Thus, despite lower overall levels of exploration, 3-year-olds were sensitive to how the property was produced, and this guided their inferences and exploration

Discussion

These results provide compelling evidence that children use pedagogical cues to guide their inductive inference and exploration. When 4-year-old children were deliberately shown a causal property of a novel object in a pedagogical manner, they explored more upon discovering that the property did not obtain for additional kind members, indicating that they expected the property to generalize. Furthermore, 3-year-olds explored only in the Intentional condition, suggesting a sensitivity to intentional demonstration even at a younger age.

Two additional factors beyond the pedagogical cues may have influenced children's exploration. First, to convey that it was accidental the experimenter said "Oops!" after producing the property in the accidental condition. But this may have also marked the property as negative, potentially inhibiting exploration. Additionally, the conditions may have produced slightly different evidence—more paperclips may have stuck to the block in the intentional condition, making the property potentially more salient. Experiment 2 explored the possible effect of these factors on children's exploration. We added an enthusiastic exclamation ("Wow!") in both conditions to mitigate any influence of negative affect, and also equated the number of paperclips picked up across conditions.

Experiment 2

The results of Experiment 1 make clear that, even when objects share a kind label, whether or not a property is demonstrated in an intentional, pedagogical manner has a powerful effect on children's inferences about the generalizability of that property and their exploration of novel kind members. Further, as mentioned above, there are other potentially interesting factors that could also be influencing children's exploration—specifically the inherent negativity of accidental events and the varying salience of the property information. If children are sensitive to the affective valence of causal events and attuned to the saliency of particular properties in making inferences and guiding exploration of novel kinds, then we might expect that equating these factors across conditions could dampen the effect of the manner of demonstration. However, if children's sensitivity to pedagogical cues is singularly important in guiding inference and exploration, equating for other facets of the event might have little impact on the effect of intentional demonstration.

Methods

Participants The participants were an additional 32 3-year-olds (16 girls; M = 41 months; range: 39-46 months) and 32 4-year-olds (16 girls; M = 52 months; range: 48-57 months), with comparable backgrounds to children in Experiment 1.

Procedure The procedure was identical to Experiment 1 with several modifications. First, while maintaining the manipulation of saying either "Look, watch this" or "Oops!" the experimenter also exclaimed, "Wow!" after producing the property in both conditions, rather that simply saying, "Hmm." This should mitigate any inhibitory effect that exclaiming "Oops!" in the accidental condition might have had on children's exploration. Second, we controlled for the number of paperclips picked up across conditions. The experimenter always picked up 2 paperclips in the intentional condition, while in the accidental condition the mean was 2.41 paperclips.

Results

Unlike Experiment 1, in which not one 3-year-old in the accidental condition explored, some 3-year-olds in both conditions of Experiment 2 did explore. However, violations of assumptions of normality and homoscedasticity precluded parametric comparisons across age groups. Instead, we used non-parametric ordinal logistic regressions (see Cimpian, 2009), with condition and age as predictors, to compare exploration across the two age groups and two conditions.

These analyses revealed a main effect of condition on children's exploration, with children in the intentional condition spending more time exploring (Wald χ^2 = 10.05, df = 1, p = 0.002) and making more attempts to elicit the property (Wald χ^2 = 18.29, df = 1, $p < 0.001$) than children in the accidental condition. The analyses also revealed a main effect of age, with 4-year-olds spending marginally more time exploring (Wald χ^2 = 3.21, df = 1, p = 0.073) and making significantly more attempts to elicit the property (Wald χ^2 = 6.82, p = 0.009) than 3-year-olds. To explore these effects further, we followed up these analyses by conducting Mann-Whitney U tests within each age group.

4-year-olds As in Experiment 1, 4-year-olds spent more time exploring in the intentional condition (M = 40.63 s, SD = 19.57) than in the accidental condition, (M = 20.75 s, SD = 22.27), U = 70.5, N = 32, p = 0.029.They also made more attempts to elicit the property in the intentional condition (M = 7.63, SD = 4.53) than in the accidental condition (M = 2.81, SD = 2.46), U = 44.5, N = 32, p = 0.001.

3-year-olds As in Experiment 1, significantly more 3-year-olds explored in the intentional condition (12 children; 75%) than in the accidental condition (5 children; 31%), $\chi^2(1, N = 32) = 6.15$, $p = 0.013$. Additionally, 3-year-olds spent more time exploring in the intentional condition ($M = 32.38$ s, $SD = 23.48$) than the accidental condition ($M = 11.56$ s, $SD = 21.09$), $U = 68$, $N = 32$, $p = 0.023$, and made more attempts to elicit the property in the intentional condition ($M = 5.63$, $SD = 6.29$) than the accidental condition ($M = 1.00$, $SD = 2.76$), $U = 59.5$, $N = 32$, $p = 0.008$. Thus, as with the older children, 3-year-olds used pedagogical cues to assess the generalizability of new information and guide their exploration.

Discussion

Even when controlling for the emotional valence of the event and the salience of the property, children showed different patterns of inductive inference and exploration on the basis of whether a property was demonstrated intentionally. Having seen a property intentionally demonstrated rather than produced accidentally, 3- and 4-year-olds showed increased exploration when that property failed to obtain for other kind members.

General Discussion

These experiments provide initial purchase on the question of how intentional demonstration influences children's inductive inferences. While previous research has documented an early sensitivity to pedagogy (Csibra & Volein, 2008; Egyed et al., 2007 Gergely et al., 2007; Yoon et al., 2008), the current work directly investigates the role of pedagogical cues in the process of theory-based categorization and concept formation in young children. As early as age 3, children take intentionally demonstrated information as more kind-relevant and generalizable than identical evidence produced accidentally.

Recent work has suggested that pedagogy might be a "double-edged sword," potentially dampening children's natural curiosity and constraining learning to only what is being taught (Bonawitz et al., 2009). However, our data indicate that children do not merely learn exactly what is taught (in our case, that a particular novel object is magnetic), but rather infer from pedagogical cues that this is an important and generalizable property of the novel kind. Upon encountering evidence conflicting with this inference, having seen the property demonstrated pedagogically increased curiosity and exploration. Thus, pedagogy may facilitate deeper learning of socially or culturally important information. Particularly to the extent that children are intuitively geared towards to learning not simply everything one can do with an object, but rather what we as a group or society use such artifacts for (Kelemen, 1999; Kelemen & Carey, 2007), selective use of pedagogical cues in this manner may be particularly important.

It is important to note that in the current research, we have not directly addressed the distinction between *pedagogical* as opposed to simply *intentional* action. In the current studies, the intentional condition was both intentional and pedagogical, while the accidental condition was neither. It is possible that simply seeing an artifact used in an intentional manner is enough to lead children to infer that other objects of the same kind can be used in the same way. However, children may remain particularly attuned to whether or not that action was done with *pedagogical* intent—that is, with the purpose of teaching them something new—or merely with the intent of carrying out a particular function. This is an important question, and one which we are addressing in further research.

Another open question what children are learning from, on one hand, information conveyed by the demonstration, and on the other hand, evidence produced by their own exploration. It is precisely this conflict between inferences about generalizability made on the basis of pedagogical cues and evidence that the property in fact fails to generalize which appears to drive continued exploration. But of course this conflict remains even after exploration, and how children resolve this conflict is as yet unclear.

More broadly, these results support the idea that, as generic language conveys information about the generalizability and conceptual importance of new information (Cimpian & Markman, 2009; Gelman et al., 2002; Hollander et al., 2009), so too does intentional, pedagogical action. When presented with the same novel causal property in a pedagogical manner rather than an accidental one, children make appear to make generic, kind-based inferences that drive their exploration. Furthermore, this obtains even when objects in both conditions share a label. Kind labels are known to license category-based inductive inferences, (e.g., Gelman & Markman, 1986), and having shared versus distinct kind labels does influence exploratory play (Schulz et al., 2008). Our research demonstrates that pedagogical cues play an important role above and beyond that of the kind label.

When facing inductive problems in generalization, children have many sources of information available to them, both non-social (e.g., observation, exploration, and prior knowledge) and social (e.g., labels, generic language, intentional and pedagogical cues). Children's ability to integrate sources of information—especially when they conflict—is an important skill. The current research suggests that this ability is developing during the preschool years, and that by as young as 3 children are particularly sensitive to intentionally communicated information as they form and test hypotheses about the world.

Acknowledgments

This research was supported by a NSF Graduate Research Fellowship to the first author. We are grateful to the teachers, staff, parents, and children at Bing Nursery School

and the Arboretum Child Care Center for their participation, to Hannah Jaycox and Cole Murphy-Hockett for their assistance with data collection and coding, to Andrei Cimpian for comments on a previous draft of this paper, and to Krishna Savani for advice on statistical analyses.

References

Baldwin, D. A. (1991). Infants' contribution to the achievement of joint reference. *Child Development*, *62*(5), 875-890.

Baldwin, D. A. (1993a). Early referential understanding: Infants' ability to recognize referential acts for what they are. *Developmental Psychology*, *29*(5), 832-843.

Baldwin, D. A. (1993b). Infants' ability to consult the speaker for clues to word reference. *Journal of Child Language*, *20*(2), 395-418.

Baldwin, D. A., Markman, E. M., & Melartin, R. L. (1993). Infants' ability to draw inferences about nonobvious object properties: Evidence from exploratory play. *Child Development*, *64*(3), 711-728.

Bonawitz, E., Shafto, P., Gweon, H., Chang, I., Katz, S., Schulz, L., et al. (2009). The double-edged sword of pedagogy: Modeling the effect of pedagogical contexts on preschoolers' exploratory play. *Proceedings of the Thirty-first Annual Conference of the Cognitive Science Society*.

Booth, A. E., & Waxman, S. R. (2002). Word learning is "smart": Evidence that conceptual information affects preschoolers' extension of novel words. *Cognition*, *84*(1), B11-B22.

Carpenter, M., Akhtar, N., & Tomasello, M. (1998). Fourteen- through 18-month-old infants differentially imitate intentional and accidental actions. *Infant Behavior and Development*, *21*(2), 315-330.

Cimpian, A., & Markman, E. M. (2008). Preschool children's use of cues to generic meaning. *Cognition*, *107*(1), 19-53.

Cimpian, A., & Markman, E. M. (2009). Information learned from generic language becomes central to children's biological concepts: Evidence from their open-ended explanations. *Cognition*, *113*(1), 14-25.

Clark, H. H. (1996). *Using Language*. New York: Cambridge University Press.

Csibra, G., & Gergely, G. (2006). Social learning and social cognition: The case for pedagogy. In Y. Munakata & M. H. Johnson, *Processes of Change in Brain and Cognitive Development* (Vol. XXI., pp. 249-274). New York: Oxford University Press.

Csibra, G., & Gergely, G. (2009). Natural pedagogy. *Trends in Cognitive Sciences*, *13*(4), 148-153.

Csibra, G., & Volein, Á. (2008). Infants can infer the presence of hidden objects from referential gaze information. *British Journal of Developmental Psychology*, *26*(1), 1-11.

Egyed, K., Király, I., Krekó, K., Kupán, K., & Gergely, G. (2007, March). Understanding object-referential attitude expressions in 18-month-olds: The interpretation switching function of ostensive-communicative cues. Poster presented at the Biennial Meeting of the SRCD, Boston.

Gelman, S. A. (2009). Learning from others: Children's construction of concepts. *Annual Review of Psychology*, *60*, 115-140.

Gelman, S. A., & Coley, J. D. (1990). The importance of knowing a dodo is a bird: Categories and inferences in 2-year-old children. *Developmental Psychology*, *26*(5), 796-804.

Gelman, S. A., & Markman, E. M. (1986). Categories and induction in young children. *Cognition*, *23*(3), 183-209.

Gelman, S. A., & Markman, E. M. (1987). Young children's inductions from natural kinds: The role of categories and appearances. *Child Development*, *58*(6), 1532-1541.

Gelman, S., Star, J., & Flukes, J. (2002). Children's use of generics in inductive inferences. *Journal of Cognition and Development*, *3*(2), 179-199.

Gergely, G., & Csibra, G. (2005). The social construction of the cultural mind: Imitative learning as a mechanism of human pedagogy. *Interaction Studies*, *6*(3), 463-481.

Gergely, G., Egyed, K., & Király, I. (2007). On pedagogy. *Developmental Science*, *10*(1), 139-146.

Harris, P. L. (2002). What do children learn from testimony? In P. Carruthers, S. Stich, & M. Siegal, *The cognitive basis of science* (pp. 316-334). New York: Cambridge University Press.

Harris, P. L., & Koenig, M. A. (2006). Trust in testimony: How children learn about science and religion. *Child Development*, *77*(3), 505-524.

Hollander, M. A., Gelman, S. A., & Raman, L. (2008). Generic language and judgements about category membership: Can generics highlight properties as central? *Language and Cognitive Processes*, *24*(4), 481-505.

Kelemen, D. (1999). Function, goals and intention: Children's teleological reasoning about objects. *Trends in Cognitive Sciences*, *3*(12), 461-468.

Kelemen, D., & Carey, S. (2007). The essence of artifacts: Developing the design stance. In E. Margolis & S. Laurence, *Creations of the mind: Theories of artifacts and their representation* (pp. 212-230). New York: Oxford University Press.

Meltzoff, A. N. (1995). Understanding the intentions of others: Re-enactment of intended acts by 18-month-old children. *Developmental Psychology*, *31*(5), 838-850.

Schulz, L. E., Standing, H. R., & Bonawitz, E. B. (2008). Word, thought, and deed: The role of object categories in children's inductive inferences and exploratory play. *Developmental Psychology*, *44*(5), 1266-1276.

Sperber, D., & Wilson, D. (1986). *Relevance: Communication and cognition*. Cambridge, MA: Harvard University Press.

Tomasello, M. (1999). *The cultural origins of human cognition*. Cambridge, MA: Harvard University Press.

Tomasello, M., & Carpenter, M. (2007). Shared intentionality. *Developmental Science*, *10*(1), 121-125.

Yoon, J. M., Johnson, M. H., & Csibra, G. (2008). Communication-induced memory biases in preverbal infants. *Proceedings of the National Academy of Sciences*, *USA*, *105*(36), 13690-13695.

The Facilitative Effect of Context on Second-Order Social Reasoning

Ben Meijering[1] (b.meijering@rug.nl), Leendert van Maanen[1], Hedderik van Rijn[2], & Rineke Verbrugge[1]
[1] Department of Artificial Intelligence, University of Groningen [2] Department of Psychology, University of Groningen

Abstract

This paper is about higher-order social reasoning such as "I think that you think that I think …". Previous research has shown that such reasoning seriously deteriorates in complex social interactions. It has been suggested that reasoning can be facilitated greatly if an abstract logical problem is embedded in a context. This has not yet been tested for higher-order social reasoning. We presented participants with strategic games that demand higher-order social reasoning. The games were embedded in the context of a marble game. Participants performed really well, that is, almost at ceiling. We argue that context has a facilitative effect on higher order-social reasoning.

Keywords: Theory of Mind; Social Cognition; Higher-order Social Reasoning; Strategic Game.

Social Reasoning

In many social situations we need to reason about one another. We do so to plan our actions and predict how our behavior might affect others. The ability to reason about another's knowledge, beliefs, desires and intentions is often referred to as Theory of Mind (Premack & Woodruff, 1978). It has been extensively investigated in children and seems to develop around the age of 4 years (Wimmer & Perner, 1983; but see Onishi & Baillargeon, 2005). Nevertheless, reasoning about others is very demanding, even for adults, which becomes apparent in more complex interactions. So far, empirical results have shown social reasoning to be far from optimal (Flobbe, Verbrugge, Hendriks, & Krämer, 2008; Hedden & Zhang, 2002). It has been suggested that (social) reasoning might be facilitated if it is embedded in a context (Wason & Shapiro, 1971). In the current study, we investigate whether social reasoning really is difficult and whether embedding it in a context can facilitate it.

When we ascribe a simple mental state to someone, we are applying first-order social reasoning. For example, imagine a social interaction between Ann, Bob and Carol. If Bob thinks "Ann knows that my birthday is tomorrow", he is applying first-order reasoning, which covers a great deal of social interaction.

However, first-order reasoning is not sufficient to cover more complex social situations. The interaction between Ann, Bob and Carol can easily demand reasoning of one order higher: If Carol thinks "Bob knows that Ann knows that his birthday is tomorrow", she is making a second-order attribution.

Bob's first-order attribution and Carol's second-order attribution are hierarchically structured: Bob applied first-order reasoning by attributing a mental state to Ann, and Carol applied second-order reasoning by attributing first-order reasoning to Bob. A third-order attribution involves the reader attributing second-order reasoning to Carol, and so forth.

The depth of reasoning in humans is constrained by cognitive resources (Verbrugge, 2009; Flobbe et al., 2008; Hedden & Zhang, 2002). As the order of reasoning increases, the demands on cognitive processing increase as well. Cognitive resources and processing speed seem to increase with age (Fry & Hale, 1996), and that increase could allow for the representation of increasingly more complex mental states. Findings from developmental studies support that idea. Where first-order social reasoning is acquired at the age of around 4 years (Wimmer & Perner, 1983), second-order social reasoning seems to develop some years later, at the age of around 6 to 8 years (Perner & Wimmer, 1985). However, 6- to 8-year-olds do not understand all kinds of mental states, and even adults cannot readily apply second-order reasoning in all kinds of contexts (Flobbe et al., 2008; Hedden & Zhang, 2002).

Paradigms to Test Social Reasoning

There are a few paradigms to test social cognition. Probably the most familiar paradigm is the False-Belief task (Wimmer & Perner, 1983), which has been adapted to test second-order social cognition (Perner & Wimmer, 1985). In a typical second-order False-Belief story, two characters, John and Mary, are independently informed about the transfer of an object, an ice-cream van, from one location to another. In the story, both John and Mary know where the van is, but John does not know that Mary also knows that the van has moved to a new location. Participants are told the story and asked where John thinks Mary will go for ice cream. To answer this question correctly, participants have to be able to represent the second-order false belief "John thinks that Mary thinks the van is still at the old location.". In Perner and Wimmer's (1985) study, some children of 6 to 7 years of age were able to make such second-order attributions, but only under optimal conditions; when the inference of second-order beliefs was prompted.

Apart from some concerns about the False-Belief task's aptness to test for the presence of a Theory of Mind (Bloom & German, 2000), Perner and Wimmer (1985) expressed concerns about the generality of their findings as participants were presented "rather pedestrian problem[s] of knowing where somebody has gone to look for something" (p. 469). They stressed that investigations into higher-order social reasoning will only achieve theoretical importance if a link with other domains can be established.

Various other language comprehension paradigms have been used to test social cognition (e.g., Van Rij, Van Rijn, & Hendriks, to appear; Hollebrandse, Hobbs, De Villiers, & Roeper, 2008; Hendriks & Spenader, 2006). Hollebrandse et al. (2008) presented discourse with multiple, recursive

embeddings. In their Experiment 1, no second-order reasoning was observed in children and adults. Hollebrandse et al.'s (2008) findings led them to conclude that "second-order theory of mind is a different milestone than first-order theory of mind." (p. 276).

The problem with the paradigms mentioned above is that they depend heavily on language skills (Apperly, Samson, Chiavarino, & Humphreys, 2004; Bloom & German, 2000), and cannot be adapted easily to investigate higher orders of reasoning. A paradigm that does not depend that much on language skills is that of strategic games (Verbrugge, 2009; Flobbe et al., 2008; Hedden & Zhang, 2002). In strategic games, players have to reason about one another, because a player's payoffs depend on what the other players do, and vice versa. Games are less prone to semantic idiosyncrasies and are as such easier to control. That allows games to be presented repeatedly in different variations to acquire a more accurate measure of second-order reasoning.

Strategic Games

Hedden and Zhang (2002) used a strategic game to study first- and second-order social reasoning. It is a sequential-move game, which is played on a 2-by-2 matrix (Figure 1). In each cell there are separate payoffs for Player 1 and Player 2, respectively. The goal is to attain the highest possible payoff. The players take turns; Player 1 begins. At each turn, a player has to decide whether to stay or to move to the next cell, as indicated in Figure 1. If a player decides to stay in a particular cell, the game ends and both players attain the respective payoffs in that cell. If a player decides to move, the turn passes to the other player.

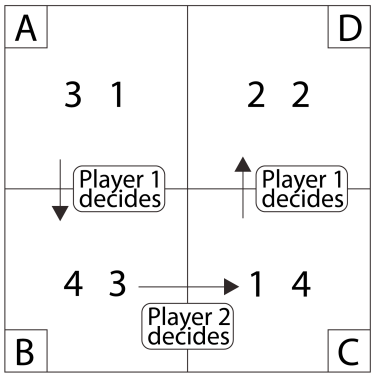


Figure 1: Schematic overview of a matrix game (Hedden & Zhang, 2002). The first number in each cell is Player 1's payoff, the second Player 2's payoff. The goal is to attain the highest possible payoff. Participants first had to predict what the other player would do at cell B before making a decision what to do at cell A. In this example, Player 1 would have to predict that Player 2 will stay, because Player 1 will move if given a choice at cell C, leading to a lower payoff for Player 2, namely 2 instead of 4. Consequently, the rational decision for Player 1 is to move to cell B.

Hedden and Zhang (2002) asked participants to (1) predict what the other player would do (stay or move) at cell B, and (2) decide whether to stay or to move at cell A. The first question provides a direct measure of what order of reasoning participants apply. To answer that question correctly, participants have to apply second-order reasoning; think about what the other player (at cell B) thinks that they think (at cell C). The second question measures whether the decisions that the participants make are based on the predictions that they have made. As a consequence of this procedure in which participants first have to predict what the other player will do, the application of second-order reasoning may not be completely spontaneous. Nevertheless, the proportion of games in which participants made second-order predictions is not that high, in the range of 60% – 70% at the end of the experiments 1 and 2, considering that by chance alone that proportion would be 50%, because there are just two predictions possible: either Player 2 stays or Player 2 moves.

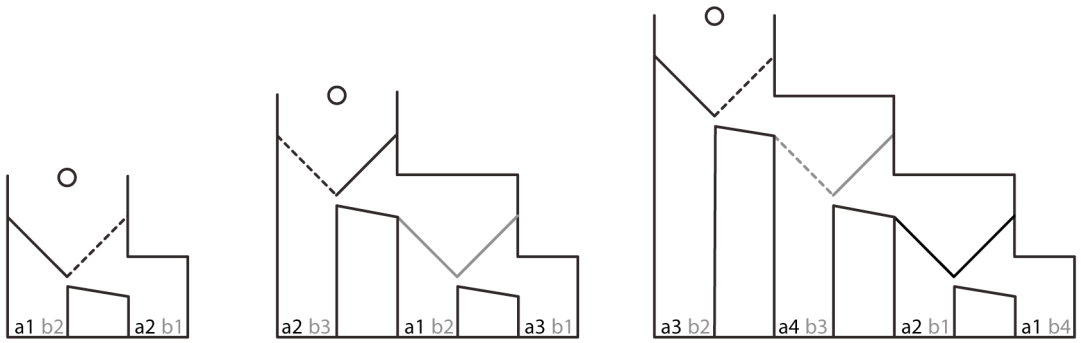
Poor performance could imply that second-order social reasoning is difficult or that participants had difficulties understanding Hedden and Zhang's (2002) matrix games. Participants could have had difficulties to comprehend the task, because the games are very abstract. The matrix games of Hedden and Zhang (2002) would be less abstract if embedded in a context. Higher-order social reasoning, which seems to be very demanding in these games, might benefit from a context embedding.

Context Effects

Some studies have investigated whether reasoning can be facilitated if a problem is presented in a (social) context. The Wason Selection Task (Wason & Shapiro, 1971) is an example of a task to investigate effects of context on reasoning. Wason and Shapiro (1971) presented a logical problem in an abstract form to one group of participants and in a "thematic" form (i.e., embedded in a social context) to another group of participants. Ten out of sixteen participants in the thematic group solved the problem, opposed to two out of sixteen participants in the abstract group. That finding implies a facilitative effect of context on reasoning.

However, there is another interpretation of Wason and Shapiro's (1971) manipulation, according to which the abstract and thematic forms are not logically equivalent (Stenning & Van Lambalgen, 2004; Manktelow & Over, 1991). If the logical problem does differ for these forms, Wason and Shapiro's findings do not support the argument that context has a facilitative effect on reasoning. To really appreciate facilitative effects of context on (higher-order social) reasoning, it is important that the context in which we embed the matrix games of Hedden and Zhang (2002) does not change their logical form. Then, improved performance can be attributed solely to context effects.

Not just any context will facilitate (higher-order social) reasoning. Flobbe et al. (2008) embedded Hedden and Zhang's matrix games in a context. Participants played games in which they, together with the computer, drive a car. The games are an adaptation of the Centipede game (Rosenthal, 1981), and are logically equivalent to Hedden and Zhang's (2002) matrix games. In second-order games,



(a) (b) (c)

Figure 2: A zeroth-order (a), first-order (b) and second-order (c) Marble Drop game. The participant's payoffs are represented by a1 – a4, the computer's by b1 – b4, both in increasing order of value. The goal is to let the white marble end up in the bin with the highest attainable payoff. The diagonal lines represent trapdoors. At the first set of trapdoors, the participant decides which of both trapdoors to remove, at the second set the computer decides, and at the third set the participant again decides. The dashed lines represent the trapdoors that both players should remove to attain the highest payoff they can get.

the road has three junctions, which correspond with the transitions, from one cell to another, in Hedden and Zhang's matrix games. At each junction either the participant or the computer decides to move ahead (i.e., continue the game) if there is a higher payoff to attain further in the game, or to turn right (i.e., end the game) if there is no higher payoff to be attained further in the game. The participant and the computer alternately take seat in the driver's position; the one in driver's seat makes the decision.

The performance of the adult participants in Flobbe et al.'s experiment was higher than in Hedden and Zhang's study: the mean proportion of games in which the adult participants gave a correct second-order prediction was more than 70%. It is important to note that this proportion is an average over Flobbe et al.'s entire experiment, whereas in Hedden and Zhang's experiments the proportion of games in which the participants applied second-order reasoning did not reach 60% – 70% until the end.

Flobbe et al.'s findings support the idea that context facilitates reasoning. Their findings also show, as did Hedden and Zhang's, that second-order reasoning is not impossible. However, performance was low, considering that the participants were explicitly asked to reason about their opponent. Both Flobbe et al. and Hedden and Zhang asked participants to first make a prediction before making a decision. This procedure is expected to scaffold second-order reasoning, as it prompts the participants to think about the other player (and what that player might think of them).

We expect that performance can be much further improved with a simpler context. In Flobbe et al.'s task, participants alternately change driver's seat with another player, which is not common practice in every day life. In the next section we will present a context that is more intuitive and will require less explanation.

We argue that second-order reasoning is not that difficult if it is embedded in an apt context, and that the facilitative effects of context render the scaffolding effects of making predictions (before making a decision) obsolete.

Experiment: Marbles and Second-order Reasoning

We present games, which we will call Marble Drop, in which the path of a white marble, which is about to drop, can be manipulated by removing trapdoors (Figure 2). Experience with world-physics allows players to see easily how the marble will run through a game. The interface of the game is very insightful; players can quickly see who can change the path of the marble, at what point in the game.

Marble Drop games are logically equivalent to Hedden and Zhang's (2002) matrix games and Flobbe et al.'s (2008) Centipede games (which we show with an informal proof in http://www.ai.rug.nl/~leendert/Equivalence.pdf). Marble Drop games only differ in appearance. The payoffs are color-graded marbles, which can easily be ranked according to preference, lighter marbles being less preferred than darker marbles. The ranking makes it possible to have payoff structures similar to those in matrix and Centipede games. The sets of trapdoors in Marble Drop games correspond with the transitions, from one cell to another, in Hedden and Zhang's (2002) matrix games.

We used color-graded marbles instead of numbers (of marbles) to minimize the usage of numeric strategies other than first- and second-order reasoning. We observed such alternative strategies in pilot studies in which we presented Flobbe et al.'s (2008) Centipede games with payoff numbers. Participants reported to use strategies such as maximizing the difference in both players' payoffs, maximizing the sum of both players' payoffs, and obstructing the other player.

Figure 2 depicts example games of Marble Drop. The goal is to let the white marble end up in the bin with the darkest color-graded marble. Note, for illustrative purposes, the color-graded marbles are replaced with codes: a1 – a4 represent the participants' color-graded marbles and b1 – b4 represent the computer's color-graded marbles (which are of another color); 1 – 4 being light to dark grades. (See http://www.ai.rug.nl/~meijering/MarbleDrop.html for the

original Marble Drop games.) The diagonal lines represent the trapdoors.

In the example game in Figure 2a, participants need to remove the right trapdoor to attain the darkest color-graded marble of their color (a2). The game in Figure 2a is a zeroth-order game, because there is no other player to reason about.

In first-order games (Figure 2b) participants need to reason about another player, the computer. The computer is programmed to let the white marble end up in the bin with the darkest color-graded marble of its target color, which is different from the participants' target color. Participants need to reason about the computer, because the computer's decision at the second set of trapdoors affects at what bin a participant can end up.

In the example game in Figure 2b, if given a choice at the second set of trapdoors, the computer will remove the left trapdoor, because its marble in the second bin (b2) is darker than its marble in the third bin (b1). Consequently, the participant's darkest marble in the third bin (a3) is unattainable. The participant should therefore remove the left trapdoor (of the first set of trapdoors), because the marble of their target color in the first bin (a2) is darker than the marble of their target color in the second bin (a1).

In a second-order game (Figure 2c) there is a third set of trapdoors at which the participants again decide what trapdoor to remove. They need to apply second-order reasoning, that is, reason about what the computer, at the second set of trapdoors, thinks that they, at the third set of trapdoors, think.

Method

Participants Twenty-two Psychology students participated in exchange for course credit. Two were excluded because of not adhering to the instructions.

Stimuli The colors of the marbles were taken from the HSV (hue, saturation and value) space. A sequential color palette was computed by varying saturation, for a given hue and value. This resulted in 4 grades (with saturation from 1 to .2) for each of the colors orange (hue = .1, value = 1) and blue (hue = .6, value = 1).

The payoff structures are constructed to be diagnostic of second-order reasoning. First- and second-order reasoning should yield opposite predictions and decisions in order to allow us to see at what order participants are reasoning.

All payoff structures in the experiment demand second-order reasoning. Consequently, payoff structures in which Player 1's first payoff is a marble with a color gradient of 1 or 4 are excluded. It is evident that in the former case participants should continue the game and in the latter case participants should end the game, whatever the other player does. The same holds for Player 2's second payoff, because at that bin (underneath the second set of trapdoors), Player 2 decides what to do.

Also, payoff structures in which Player 2's payoffs in bins 3 and 4 are lower or higher than the payoff in bin 2 are excluded. Player 2 does not need to consider Player 1's payoffs in these structures.

The payoff structures are doubly balanced for the number of left/right (trapdoor removal) predictions about Player 2 and decisions of Player 1.

Design & Procedure Before the experiment took place, participants were tested on colorblindness. They had to be able to distinguish the two colors blue and orange, and the 4 grades of each color. The experiment consisted of 3 blocks: a training block, an experimental manipulation block, and a test block.

The training block consists of zeroth-, first- and second-order Marble Drop games, respectively. In zeroth-order games, participants do not have to reason about another player. They have to find out in what bin the darkest color-graded marble of their target color is, and what trapdoor to remove to let the white marble end up in that bin. The target color is either blue or orange, which is counterbalanced between participants. If a participant's target color is blue, the computer's target color is orange, and vice versa. These games do not require social reasoning but are presented to familiarize the participants with the physics of the Marble Drop game. Participants are presented 4 zeroth-order games.

We assume that in first- and second-order games, participants reason about the decision of the computer at the second set of trapdoors. If a participant removes the left trapdoor of the first set of trapdoors, the white marble will drop into the first bin. If a participant removes the right trapdoor of the first set of trapdoors, the white marble will roll to the second set of trapdoors at which the computer decides what trapdoor to remove. If the computer removes the left trapdoor, the white marble will drop into the second bin. If the computer removes the right trapdoor, the white marble will drop into the third bin in first-order games, it will roll to the third set of trapdoors in second-order games. In the latter case, the turn passes to the participant. If the participant removes the left trapdoor, the white marble will drop into the third bin. If the participant removes the right trapdoor, the white marble will drop into the fourth bin. As soon as the white marble drops into a bin, participants are presented feedback ("correct!" or "incorrect!"). If they fail to let the white marble end up in the correct bin, a green arrow is depicted underneath the correct bin and participants are asked to explain verbally why that bin is the correct one. Participants are presented 8 first-order games and 8 second-order games.

In the experimental manipulation block, participants play second-order Marble Drop games. The participants are asked to decide what to do at the first set of trapdoors. They immediately receive feedback after making a decision. The experimental manipulation involves that one half of the participants is asked first to predict what the computer will do at the second set of trapdoors, before making a decision at the first set of trapdoors. This manipulation is included to investigate scaffolding effects of making predictions. In this block and the next, the games are not continued after the

participants have made a decision. The experimental manipulation block consists of 32 trials, all trials diagnostic of second-order social reasoning.

In the test block, the participants play second-order Marble Drop games. The participants that made a prediction before making a decision in the experimental manipulation block do not have to make predictions anymore. The test block has the same structure as the experimental manipulation block, except that none of the participants have to make predictions anymore.

The participants were randomly assigned to the group that makes a prediction and a decision in the experimental manipulation block and only a decision in the test block, the PD-D group, and the group that makes decisions in the experimental manipulation and the test block, the D-D group.

Results

To account for random effects of individual differences and payoff structures, we performed Linear Mixed-Effects (LME) analyses (Baayen, Davidson, & Bates, 2008). We first analyzed the proportion of games in which participants applied second-order reasoning (Figure 3). The analysis consists of a (logistic) LME with *block* (*experimental manipulation* and *test*) and *group* (*PD-D* and *D-D*) as fixed factors and *participants* and *payoff structures* as random factors.

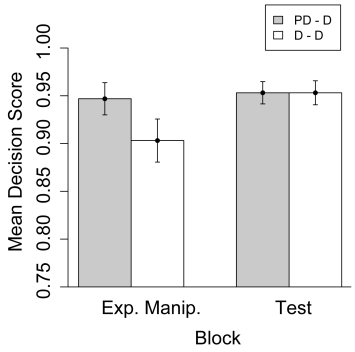


Figure 3: Mean proportion of games in which participants applied second-order reasoning, presented separately for the PD-D and D-D groups in the experimental manipulation block and the test block. The standard errors are depicted above and below the means.

The grand mean is 0.94. The factors *block* and *group* are significant: $\beta = .809$, $z = 2.348$, p = .009 and $\beta = 5.721$, $z = 1.844$, p = .033, respectively. The interaction *group* x *block* is also significant: $\beta = -1.024$, $z = -1.844$, p = .033.

Reaction Times The games in the experimental manipulation block are procedurally different for the PD-D and the D-D groups. We analyzed the reaction times of the games in the test block, because these are not procedurally different for the PD-D and the D-D groups.

After removing the trials in which participants unsuccessfully applied second-order reasoning, a LME analysis was performed, with *group* (*PD-D* and *D-D*) as a fixed factor and *participants* and *payoff structures* as random factors.

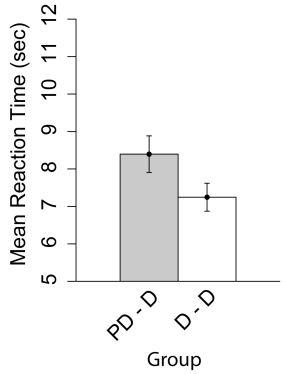


Figure 4: Mean reaction times for the PD-D and D-D groups. The standard errors are depicted above and below the means.

The grand mean is 7.82 seconds. The factor *group* is significant: $\beta = 1.523$, t = 3.09, p < .01. On average, the participants in the D-D group were faster to make a decision than the participants in the PD-D group (Figure 4).

Discussion

The participants performed really well in the Marble Drop games (Figure 3). The proportion of games in which they successfully applied second-order reasoning was very high, at 94% correct. That proportion is much higher than in Hedden and Zhang's (2002) matrix games (60% - 70%) and Flobbe et al.'s (2008) Centipede games (slightly above 70%). This finding supports the idea that a context can facilitate reasoning. It matters in what context reasoning is embedded. Flobbe et al.'s context facilitated higher-order reasoning, but not as strongly as in our experiment, which has a simpler context.

The interaction between *block* and *group* is significant. The performance of the participants that made a prediction before making a decision in the experimental manipulation block, the PD-D group, is almost at ceiling in both the experimental manipulation block and the test block (Figure 3). On the other hand, the performance of the participants that did not make a prediction in the experimental manipulation block, the D-D group, is not yet at ceiling in the experimental manipulation block, but reaches ceiling in the test block (Figure 3). This finding could imply that the D-D group lacked a scaffolding effect of making predictions in the beginning of the experiment. However, the D-D group, which was not explicitly asked to predict what the other player would do (before making a decision), probably did learn to make predictions during the experimental manipulation block. Eventually, there was no difference in performance anymore in the test block.

The main effect of *block* can be mainly attributed to the D-D group. The performance of the participants in the D-D group increases to ceiling in the test block, whereas the performance of the participants in the PD-D groups already

reaches ceiling in experimental manipulation block and remains stable (Figure 3).

The participants not only performed better, they also responded faster in Marble Drop games than in matrix games. Mean reaction times of second-order predictions were approximately 10 seconds in matrix games (Hedden & Zhang, 2002), whereas mean reaction times of second-order predictions took less than 7.5 (M = 7.1, SE = .57) seconds in Marble Drop games for the PD-D group in the experimental manipulation block. Mean reaction times of second-order decisions (based on second-order predictions) were approximately 3.5 second in matrix games, and less than 2.5 (M = 2.2, S = .77) seconds in Marble Drop games for the PD-D group in the experimental manipulation block.

Although these comparisons with Hedden and Zhang's (2002) results are informal, the differences are considerable. The better performance in Marble Drop games than in Hedden and Zhang's (2002) matrix games probably is not caused by a difference in our participants' speed-accuracy tradeoff. Our participants applied second-order reasoning more often and faster, which supports the idea that our context facilitated higher-order reasoning.

In the test block, on average, the participants in the D-D group were faster to make a decision than the participants in the PD-D group (Figure 4). In the test block, the behavior of the participants in the PD-D group could still have been constrained in a stepwise procedure of first making a prediction, then a decision. The participants in the D-D group were given more freedom in the experimental manipulation block to naturally interleave a prediction between the steps in their decision-making, which could have caused them to be faster than the participants in the PD-D group.

General Conclusion

Our findings seem to imply that embedding a logical problem in a context greatly facilitates (social) reasoning. Because of the facilitative effects of context embedding, second-order reasoning did not need to be scaffolded by explicitly asking participants to predict the behavior of other players. Second-order reasoning might still be difficult, but participants were able to apply it in Marble Drop games.

The question remains what strategies participants used to arrive at second-order decisions and predictions. We intend to investigate this with computational models (e.g., Van Maanen & Verbrugge, submitted). These models can help us to explore the cognitive mechanism involved in higher-order social cognition, and whether higher-order social cognition will generalize to more complex tasks.

References

Apperly, I. A., Samson, D., Chiavarino, C., & Humphreys, G. W. (2004). Frontal and temporo-parietal lobe contributions to theory of mind: Neuropsychological evidence from a false-belief task with reduced language and executive demands. *Journal of Cognitive Neuroscience, 16*(10), 1773-1784.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*(4), 390-412.

Bloom, P., & German, T. P. (2000). Two reasons to abandon the false belief task as a test of theory of mind. *Cognition, 77*(1), 25-31.

Flobbe, L., Verbrugge, R., Hendriks, P., & Krämer, I. (2008). *Children's application of theory of mind in reasoning and language. Journal of Logic, Language and Information, 17*(4), 417-442.

Fry, A. F., & Hale, S. (1996). Processing speed, working memory, and fluid intelligence: Evidence for a developmental cascade. *Psychological science, 7*(4), 237-241.

Hedden, T., & Zhang, J. (2002). What do you think I think you think?: Strategic reasoning in matrix games. *Cognition, 85*(1), 1-36.

Hendriks, P., & Spenader, J. (2006). When production precedes comprehension: An optimization approach to the acquisition of pronouns. *Language Acquisition, 13*(4), 319-348.

Manktelow, K. I., & Over, D.E. (1991). Social roles and utilities in reasoning with deontic conditionals, *Cognition, 39*(2), 85-105.

Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science, 308*(5719), 255.

Perner, J., & Wimmer, H. (1985). "John thinks that Mary thinks that..." Attribution of second-order beliefs by 5- to 10-year-old children. *Journal of Experimental Child Psychology, 39*(3), 437-471.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain sciences, 1*(4), 515-526.

Rosenthal, R. W. (1981). Games of perfect information, predatory pricing and the chain-store paradox. *Journal of Economic Theory, 25*(1), 92-100.

Stenning, K., & Van Lambalgen, M. (2004). A little logic goes a long way: basing experiment on semantic theory in the cognitive science of conditional reasoning. *Cognitive Science, 28*(4), 481-529.

Van Maanen, L., & Verbrugge, R. (submitted). A Computational Model of Second-Order Social Reasoning.

Van Rij, J., Van Rijn, H., & Hendriks, P. (to appear). Cognitive architectures and language acquisition: A case study in pronoun comprehension. *Journal of Child Language*.

Verbrugge, R. (2009). Logic and social cognition: The facts matter, and so do computational models. *Journal of Philosophical Logic*, 1-32.

Wason, P. C., & Shapiro, D. (1971). Natural and contrived experience in a reasoning problem. *The Quarterly Journal of Experimental Psychology, 23*(1), 63-71.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition, 13*(1), 103-128.

Phonological instability in young adult poor readers

James S. Magnuson (james.magnuson@uconn.edu)[1,2]
Anuenue Kukona (anuenue.kukona@uconn.edu) [1,2]
David Braze (braze@haskins.yale.edu) [1]
Clinton L. Johns (johns@haskins.yale.edu) [1]
Julie A. Van Dyke (jvandyke@haskins.yale.edu) [1]
Whitney Tabor (whitney.tabor@uconn.edu) [1,2]
W. Einar Mencl (einar@haskins.yale.edu) [1]
Kenneth R. Pugh (pugh@haskins.yale.edu) [1,2]
Donald P. Shankweiler (donald.shankweiler@uconn.edu) [1,2]
[1]Haskins Laboratories, 300 George St., New Haven, CT 06510
[2]Department of Psychology, University of Connecticut, Storrs, CT 06269-1020

Abstract

Phonology is held to play a central role in typical reading development (Shankweiler et al., 1979) and sensory or phonological deficits are often held to be a primary cause of reading disability (Snowling, 2008). However, little is known about the nature of phonology at the endpoint of atypical reading development -- that is, in adult poor readers. We examined the time course of (auditory) lexical activation, competition, and learning in a community sample with a high proportion of poor readers in two experiments. In Experiment 1, contrary to our expectations, we found that poor readers were *more* sensitive to subphonemic coarticulatory cues than better readers. In Experiment 2, we examined the time course of word learning along with the time course of phonological competition. Poor readers differed from better readers in the trajectory of learning, and also in phonological competition: typical readers exhibited strong competition between rhymes, but poor readers did not. Simulations with a computational model suggest that instability in phonological organization (simulated via reduced lateral inhibition) can explain differences in both studies in counter-intuitive ways, shedding new light on an old problem.

Keywords: phonology; reading; dyslexia; reading disability; spoken word recognition; computational modeling; visual world paradigm.

Introduction

A fundamental principle shared by nearly all theories of reading is that phonology plays a key role mediating the mapping from print to meaning (Harm & Seidenberg, 2004; Shankweiler et al., 1979; Snowling & Hulme, 2005; Ziegler & Goswami, 2005). This follows from repeated findings that impairments in reading are correlated with deficits in phonological abilities (Shankweiler et al., 1977; Snowling, 1981). While multiple hypotheses exist, linking the deficit to poor phonological quality (Joanisse, 1994) or low-level sensory impairments (e.g., Tallal, 1980), the precise nature of the phonological deficit in dyslexia and its causes remains a subject of intense debate.

Fairly little is known about the nature of phonological processing at the endpoint of atypical reading development, since studies of reading disability logically focus on developing samples. An exception is recent work by Szenkovits, Ramus, and colleagues (reviewed by Ramus & Szenkovits, 2008). They point out that deficits in phonological abilities in college-aged poor readers (self-reported "presumed dyslexics") are most readily detected in tasks with significant working memory demands (phonemic awareness tasks, or verbal short-term memory tasks) or under time pressure (as in rapid auditory naming). However, in tasks that do not impose such demands, poor readers are not strikingly different from typical readers (most notably, they report that poor readers in their sample exhibit phonological similarity effects similar to those exhibited by good readers, contra Shankweiler et al., 1977, who reported that poor readers fail to show such effects). Ramus and Sjenkovits suggest that the phonological deficit in dyslexia therefore may not be one of phonological representation, but rather one of phonological *access* -- and so manifests as difficulty in rapidly retrieving phonological forms into working memory. This new take on phonology in dyslexia has the potential to illuminate the nature and basis of the phonological deficit in new ways.

Techniques for examining the time course of on-line language processing provide the means to examine this hypothesis more closely. We report preliminary results of a project investigating the phonological abilities of adult poor readers. We use stimulus manipulations and time course measures that have been used to investigate lexical activation and competition at a fine timescale (Dahan, Magnuson, Tanenhaus, & Hogan, 2001) and lexical learning (Magnuson, Tanenhaus, Aslin, & Dahan, 2003) in typical adults.

Experiment 1

In Experiment 1, we sought a sensitive test of the fine-grained phonological processing of our sample, but in a task that minimizes cognitive demands. The study reported by Dahan, Magnuson, Tanenhaus and Hogan (2001) fits the bill. Dahan et al. investigated the impact of misleading coarticulation (subcategorical -- i.e., subphonemic -- mismatches). They achieved misleading coarticulation by cross-splicing recordings of words. For example, they took the initial consonant and vowel (CV) from "neck", cut as late as possible before the final stop consonant, and spliced it together with the final consonant of "net". This sounds like "net", but the vowel includes coarticulation consistent with /k/. They labeled this sort of item "W2W1" (word 2 spliced to word 1). They also had cases where the initial CV

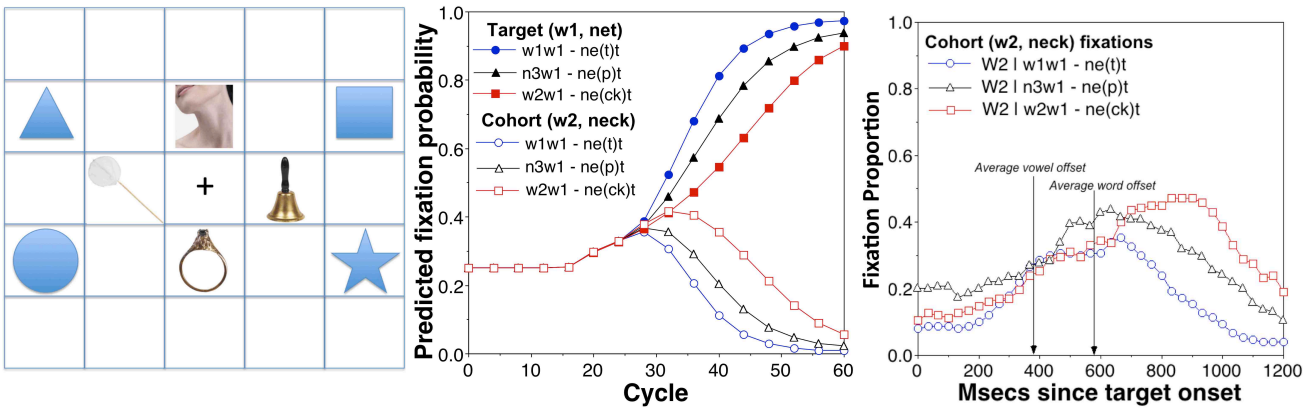


Figure 1. Left: Sample display. Center: TRACE predictions. Right: Competitor fixations over time from Dahan et al. (2001).

came from a nonword ("nep" + "net" → N3W1). Finally, they included cross-spliced items without misleading coarticulation by splicing together two recordings of a target word like "net" (W1W1).

Dahan et al. presented these items with displays like the one shown in Figure 1, using the Visual World Paradigm (VWP; Tanenhaus et al., 1995). Subjects heard instructions like "point to the net". Eye movements were recorded as subjects followed the spoken instructions.

The motivation for their study was the apparent deficiency in the TRACE model (McClelland & Elman, 1986) identified by Marslen-Wilson and Warren (1994) using these kinds of materials, in that lexical decision reaction times appeared inconsistent with the time course of activation in TRACE. However, the time course measure provided by the VWP (Figure 1, right) showed that the TRACE predictions (Figure 1, center) were remarkably accurate. Crucially, subjects fixated the competitor, "neck," most when there was misleading coarticulation consistent with that word (W2W1 condition), and least when the coarticulation was fully consistent with the target (W1W1). Fixation proportions were intermediate when misleading coarticulation did not map onto a word (N3W1). TRACE predicts the W1W1 and W2W1 patterns intuitively; the word with best bottom-up match is initially activated most strongly. The N3W1 results follow because neither net nor neck has an advantage as the nonword coarticulation is heard; thus, both reach a relatively high level of activation before the disambiguating final consonant.

Predictions What might we predict for our sample? If their linguistic difficulties arise from imprecise phonological representations (e.g., the phonological quality hypothesis of Joanisse, 2004) or slow-to-activate phonological representations (e.g., the generalized slowing hypothesis; Kail, 1994), we might expect them to be less affected by misleading coarticulation, and so show weaker competition effects. On the phonological access hypothesis (Ramus & Szenkovits, 2008), if the task minimizes cognitive demands, our sample ought to look no different from a typical sample.

Methods

Participants The participants were 56 college-aged adults (mean age = 21) recruited from community colleges and GED programs in the New Haven area. Previously, we have documented linguistic and other cognitive abilities in samples from this population (Braze et al., 2007), and demonstrated that the degree to which reading is subserved by common, supramodal brain areas also subserving speech is correlated with reading ability (Shankweiler et al., 2008). We examine this sample with a battery of 25 linguistic and other cognitive assessments. In this brief report, we only have room to mention that this population tends to lag in language and other cognitive domains, but a wide range of abilities is observed. Our goal is to conduct individual differences analyses. Given space constraints for the current report, though, we will compare the top 50% of readers in our sample with the bottom 50%. The most intuitive measure for conducting this median split is the standardized score from the Peabody Picture Vocabulary Test (which correlates closely with, e.g., a composite score derived from all subtests of the Woodcock-Johnson battery). The bottom 50% had standard scores ranging from 67 to 90, with a mean of 81. The top 50% had scores ranging from 91 to 137, with a mean of 104. The results we report do not differ if we remove, e.g., participants with low approximated IQ, and so the full sample is included.

Materials The auditory materials were those used by Dahan et al. (2001), and consisted of 15 word 1-word 2-nonword 3 triples (W1, W2, N3), such as *net, neck,* and *nep* (for the full set, see the Appendix B of Dahan et al.). The visual materials were similar to those used by Dahan et al., except that their line drawings were replaced with photographs.

Procedure The procedure was identical Dahan et al.'s. There were 3 lists, with 5 items assigned to each condition (W1W1 [consistent coarticulation], W2W1 [misleading cohort coarticulation], N3W1 [misleading nonword coarticulation]) in each list. Participants were randomly assigned to lists. On each trial, a fixation cross and four simple shapes appeared on the screen. When the participant clicked the cross, the trial began, and pictures of four objects appeared. A spoken instruction was presented over speakers, such as "point to the net; now click on it and put it below the circle." We tracked eye movements using an SR-Research Eyelink II head-mounted eye tracker, sampling at 250 hz. We tracked the probability of fixating each item

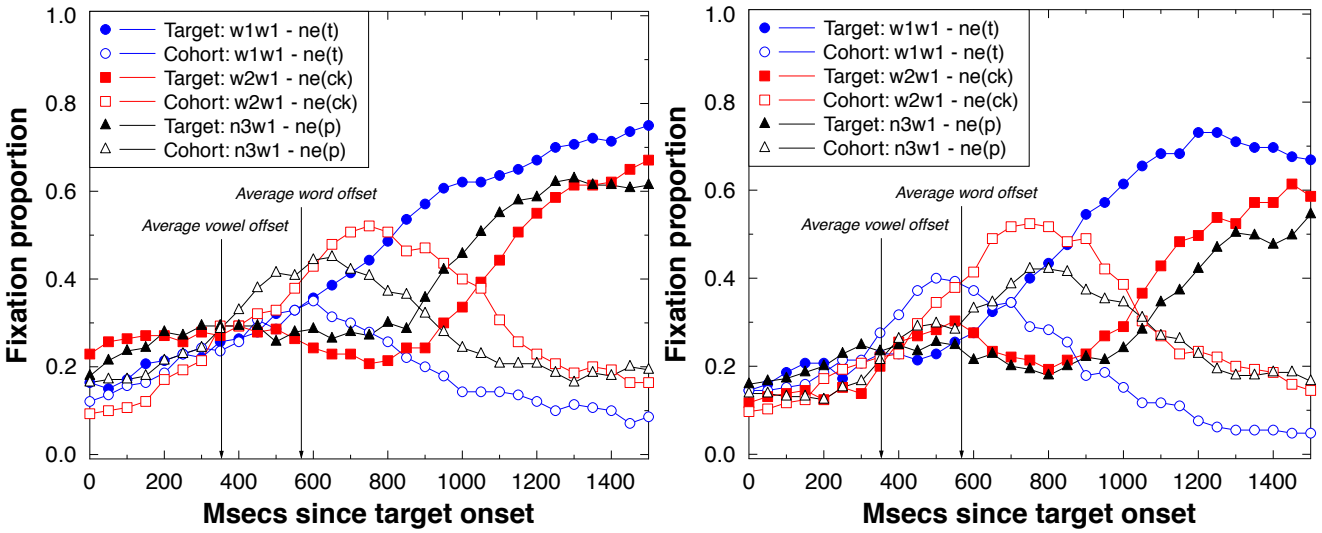


Figure 2: Subcategorical mismatch data for the top 50% (left) and bottom 50% (right) of readers from our community sample.

over time from the onset of the target word (e.g., *net*).

Results and discussion

Eye movements were parsed into saccades and fixations. Saccade time was attributed to the following fixation, since saccades are essentially ballistic; the initiation of a saccade is the earliest indicator of the choice to fixate the next gaze position. Eye tracking results are presented Figure 2. Qualitatively, the observed patterns for both halves of our sample resemble the (competitor) pattern in Figure 1. Notably, there is no apparent delay in the response to the bottom up signal in either half, when compared with the university sample in Figure 1. There are some differences between the two subsets in Figure 2 in the relative magnitude and timing of competitor proportion curves, but the most salient difference between the groups is in the target fixations in the mismatch conditions. The top 50% show the same ordering observed by Dahan et al.: W1W1 > N3W1 > W2W1. However, the pattern for the bottom 50% is W1W1 > W2W1 > N3W1. We explored this using a 2 (subset) x 2 (W2W1, N3W1) ANOVA on mean target fixation proportion in the window from 200 msecs after word onset (the expected average latency for a signal-driven saccade) to 1200 msecs (approximate target peak latency).

There was a main effect of Subset (top=0.40, bottom=0.32; $F(1,54)=6.8$, $p=.01$), but not Condition ($F < 1$), and a significant interaction ($F(1,54)=4.2$, $p<0.05$). This was due to a reliable effect of condition for the top subset (W2W1=.37, N3W1=.44; $F(1,54)=5.4$, p=.03), but not for the bottom (W2W1=.34, N3W1=.30; $F<1$).

Thus, there are several interesting patterns. There is no apparent delay in bottom up response. However, the later time course is different in both subsets compared to the sample of Dahan et al. (2001), and the subsets differ from each other. Most notably, it appears that lexical competition differs in the bottom subset. Target proportions for the mismatch conditions are depressed throughout the analysis window in comparison to the top subset, and the two mismatch conditions do not differ reliably in the amount of target interference they cause for the bottom subset.

Computational modeling To make sense of these patterns, we turned to the jTRACE re-implementation (Strauss, Harris, & Magnuson, 2007) of the TRACE model (McClelland & Elman, 1986) that includes several additional features (graphical user interface, plotting and scripting utilities). Starting with the default parameters used by Dahan et al. (2001) to obtain the simulations shown in the middle panel of Figure 1, we explored a wide range of changes to several parameters, one at a time. The goal was to determine whether any parameter could be changed to produce the observed changes in the bottom subset: increased competition effects without slowing initial lexical access. We tested a variety of parameters in TRACE (feedforward and feedback gain at various points, addition of input and "sensory" [model-internal] noise). *Lexical decay* was of particular interest, as the parameter McMurray et al. (2010) claim best fits individual differences in a lexical competition in a group of adolescents with a range of language and cognitive abilities; however, its influence is too weak and late. Two parameters could simulate the general trends: reducing *phonemic* or *lexical lateral inhibition* by approximately 50% from default levels. Reducing inhibition does not affect initial activation rates, but it allows larger competition effects because it delays the impact of late-arriving bottom-up disambiguation. In particular, it predicts larger cohort competition effects (note slight trends in this direction in the bottom subset) as well as less differentiation in target trajectories for the mismatch conditions.

Experiment 2

In Experiment 2, we continued our exploration of our sample's phonological abilities by examining lexical competition in the context of an artificial lexicon learning task (based on Magnuson, Tanenhaus, Aslin, & Dahan, 2003). This allowed us to simultaneously study



Figure 3: pictures of unusual animals used in the artificial lexicon study (Experiment 2).

phonological competition effects in word recognition (how strongly do "cohorts", like /pibo/ and /pibu/, compete? How strongly do rhymes, like /pibo/ and /dibo/, compete?) and word learning ability. Magnuson et al. (2003) were motivated in part by the goal of precisely controlling lexical characteristics such as phonological similarity, frequency, and neighborhood density. This approach has an added advantage for our sample. To the degree that our sample diverges from the performance of typical participants using real words, it is very difficult to determine the locus of the difference. There may be deep reasons, such as differential organization of processing mechanisms, or shallow ones, like simple differences in vocabulary size. An artificial lexicon paradigm allows us to put participants on maximally similar footing. While participants differ in linguistic and cognitive abilities, the items are equally unfamiliar to all.

Predictions Virtually any variant of the phonological deficit hypothesis might predict poor readers would perform worse in learning the artifical lexicon. With respect to the time course of cohort and rhyme competition, two precedents using familiar words in the visual world eye tracking paradigm suggest possible outcomes. Desroches, Joanisse, and Robertson (2006) examined cohort and rhyme competition in children with dyslexia. Unlike typically developing peers, they did not exhibit rhyme competition effects. In contrast, McMurray et al. (2010) reported that adolescents meeting criteria for SLI showed stronger cohort *and* rhyme effects, though only in the late time course. Thus, we might expect to see typical cohort effects but weak or absent rhyme effects (consistent with Desroches et al.) or late-enhanced cohort and rhyme effects (consistent with McMurray et al.).

Methods

Participants A subset of participants from Experiment 1 participated in Experiment 2: 14 individuals from the top 50% and 20 from the bottom 50%.

Materials 8 artificial words were constructed with one "cohort" (onset) competitor in the artificial lexicon and one rhyme. The words were /pibo, pibu, dibo, dibu, tupa, tupi, bupa, bupi/. The visual materials were pictures of 8 unusual animals (see Figure 3). Names were mapped randomly to pictures for each subject.

Procedure Each trial had identical structure. A fixation cross appeared in the center of the screen. When the participant clicked the cross, the trial began. Two pictures appeared, to the left and right of the cross. 500 ms later, an instruction was played, such as "find the pibo." At first, participants could only guess. If they clicked on the incorrect object, they heard "try again." When they clicked the correct object, they heard feedback, such as "that's right, that's the pibo!" The experiment consisted of 8 blocks of 24 trials. Each item appeared as the target 3 times per block, once each with its cohort, its rhyme, and an unrelated item. Thus, each block had 8 cohort, rhyme, and unrelated trials. There was no formal test; we measured behavior continuously over learning.

Results and discussion

Accuracy and response time Accuracy and response time (for accurate trials) are shown in Figure 4 for the two groups. We conducted ANOVAs with factors Type (Cohort, Rhyme, Unrelated) and Block for accuracy and RT. In the interest of space, we will only briefly summarize the results. The two subsets were both reliably more accurate for Unrelated than Rhyme trials, and more accurate in Rhyme than Cohort trials. In RT, the main effect of Type was not reliable for the top subset; in planned comparisons, none of the Types differed another. But for the bottom subset, Cohort trials were significantly slower than both Rhyme and Unrelated trials, which did not differ from each other. Thus, the bottom subset seemed to show less rhyme interference.

Fixation proportions over time are presented in Figure 5 by just showing target fixations (competitor fixations are essentially complementary) averaged over all correct trials (as the patterns did not change substantially with training). For qualitative comparison, results from a sample of 14 U. of CT (UConn) undergraduates are presented. Qualitatively, there is a very striking result. There are clear effects of both Cohort and Rhyme for the UConn sample. The Cohort effect is stronger and earlier, as with real words (Allopenna, Magnuson, & Tanenhaus, 1998; Desroches et al., 2006), while the Rhyme effect emerges later. Growth curve analysis (Mirman, Dixon, & Magnuson, 2008) revealed reliable intercept differences for the TD group (Unrelated > Rhyme > Cohort), analogous to differences in mean proportion over the analysis window. In contrast, the two community sample groups shows strong Cohort effects, but delayed Rhyme effects. The Rhyme condition differs reliably from the Unrelated condition for the top 50%, but not for the bottom 50%.

Our results are consistent with those of Desroches et al. (2006), who reported an absence of rhyme effects in children with dyslexia using a similar eye tracking paradigm with familiar, real words. They are partially consistent with the recent report of McMurray et al. (2010) that adolescents with SLI show larger but *later* competition effects than typically developing peers. We again turned to the model in order to explore possible bases for such a pattern.

Computational modeling As with Experiment 1, we used the jTRACE re-implementation (Strauss et al., 2007) of TRACE. Because TRACE is not a learning model (though see the Hebbian version of TRACE version developed by

Mirman, McClelland & Holt, 2006), we treated TRACE as a model of the stabilized system at the end of learning. Again, we changed one parameter at a time, looking for a change that would leave the magnitude and timing of the cohort effect intact while ideally wiping out the rhyme effect. We again tried several parameters. Lexical decay does not selectively affect rhyme effects. Reduced lexical lateral inhibition actually boosts rhyme effects. Only one parameter could generate the correct trends: a *reduction in lateral inhibition at the phoneme layer*. As it is reduced, rhyme effects are weakened and delayed, while leaving the cohort time course largely intact (though cohort effects are somewhat amplified). This counter-intuitive outcome follows from what happens to phonemes other than the initial phoneme of the target word. With inhibition reduced, similar phonemes get much more activated. Even though the phoneme inhibition parameter is lower, there is actually greater inhibitory flow at the phoneme level, putting rhymes that differ from the target in initial phoneme by more than a single feature at a disadvantage. Interestingly, lateral inhibition at the phoneme level was one of two parameters that could achieve the correct pattern to fit the bottom 50% subset behavior in Experiment 1.

Summary In Experiment 2, good and poor readers achieved similar accuracy in artificial lexicon learning. However, the time course of learning was substantially different, with poor readers exhibiting slower learning in early trials. Poor readers showed similar on-line onset (cohort) competition effects as better readers, but failed to exhibit a reliable effect of rhyme competition (instead showing a weak, delayed effect). This converges with a report that children with dyslexia did not exhibit rhyme effects in a similar study using real words (Desroches et al., 2006). In TRACE simulations, the only way to substantially reduce rhyme effects without inappropriately perturbing cohort (onset) effects was to reduce lateral inhibition at the phoneme level -- a parameter change that can also capture the poor reader differences in Experiment 1.

General Discussion

Adult poor readers continue to differ from good readers in phonological processing. Our poor readers showed greater interference effects from misleading coarticulation than better-reading peers in Experiment 1. Poor readers learned new words with a different trajectory than better readers in Experiment 2, and exhibited late, weak rhyme competition effects. The two primary patterns of differences -- enhanced competition due to misleading coarticulation and absence of rhyme effects -- can both be modeled in TRACE via reduced lateral inhibition at the phoneme level. The convergence on phoneme inhibition in the simulations of Experiments 1 and 2 increases our confidence that this parameter manipulation is capturing something important about phonological differences in poor readers. One next step will be to use the re-parameterized model to generate predictions for poor readers in new tasks.

We do not wish to imply that we believe that there are

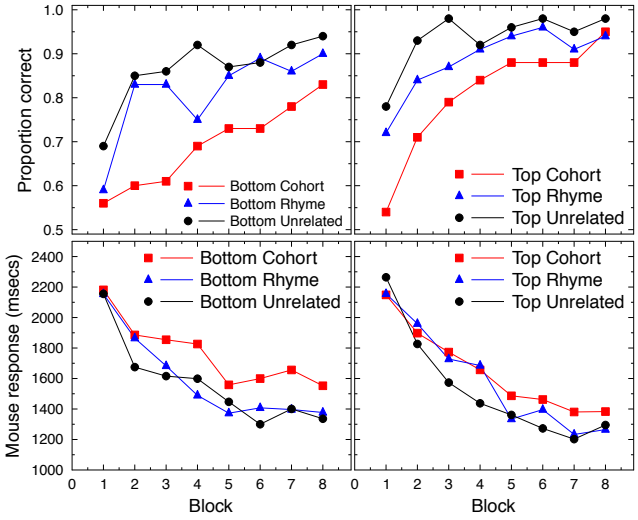


Figure 4: Accuracy (top) and RT for the bottom 50% of readers in our sample (left) and the top 50% (right) by training block.

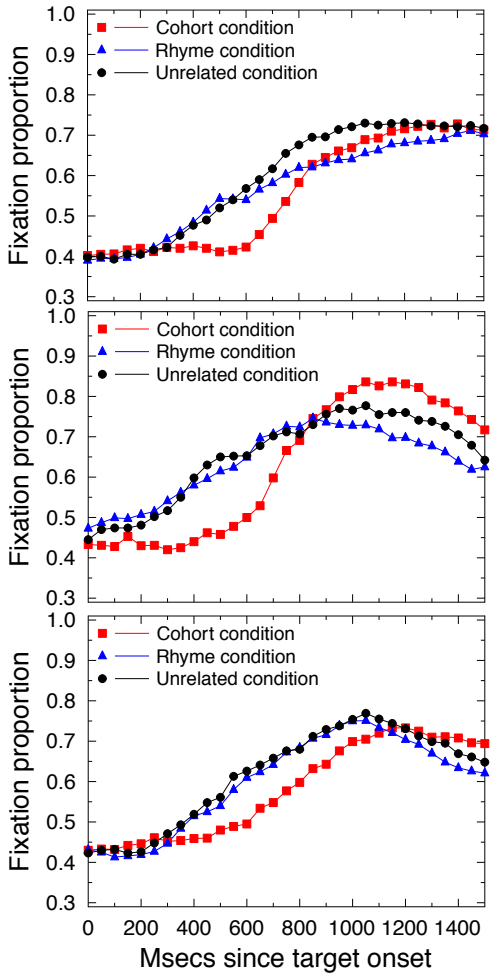


Figure 5: Target fixation proportions over time in Experiment 2, collapsed across block and only including correct trials, averaged over all 8 blocks. **Top**: typical university sample. **Middle**: to 50% of community sample readers. **Bottom**: bottom 50% of community sample readers. Patterns varied only slightly by block.

discrete representations of phonemes in the brain, let alone a discrete parameter controlling lateral inhibition. The ability of TRACE to simulate differences based on reduced phoneme inhibition instead points to the level of phonological organization in the dynamical system it is meant to simulate, i.e., the mechanisms underlying human word recognition. Thus, our simulations may identify the level of the system -- phonological organization -- that appears to be crucially different in poor readers.

Our results are potentially consistent with any form of the phonological deficit hypothesis, although they somewhat favor accounts that assume a typical level of phonetic resolution (given that poor and better readers showed similar timing in early lexical activation), and differences in the stability of phonological representations. In particular, our results may be compatible with the phonological access hypothesis (Ramus & Sjenkovits, 2008). However, our results also suggest differences in phonological access may be more subtle than suggested by Ramus and Sjenkovits, who emphasize working memory demands in conventional tasks that most clearly identify phonological deficits. That we observed differences in the time course of lexical activation, competition and learning in poor adult readers in minimally demanding, naturalistic tasks suggests that the locus of the phonological deficit may be a more low-level property of the system, even though this deficit may require difficult tasks or sensitive measures to be detected. We hope that our continuing exploration of individual differences in adult poor readers will illuminate this possibility further.

Acknowledgments

Supported by NIH grant HD-40353 to Haskins Laboratories.

References

Allopenna, P. D., Magnuson, J. S., and Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language, 38,* 419-439.

Braze, D., Tabor, W., Shankweiler, D. P., & Mencl, W. E. (2007). Speaking up for Vocabulary: Reading Skill Differences in Young Adults. *Journal of Learning Disabilities, 40.3*, 226-243.

Dahan, D., Magnuson, J. S., Tanenhaus, M. K., & Hogan, E. M. (2001). Subcategorical mismatches and the time course of lexical access: evidence for lexical competition. *Language and Cognitive Processes, 16*, 507-534.

Desroches, A.S, Joanisse, M.F. & Robertson, E.K. (2006). Specific phonological impairments in dyslexia revealed by eyetracking. *Cognition, 100,* B32-B42.

Harm, M. W., & Seidenberg, M. S. (2004). Computing the Meanings of Words in Reading: Cooperative Division of Labor Between Visual and Phonological Processes. *Psychological Review, 111*, 662-720.

Joanisse, M.F. (2004). Specific Language Impairments in Children: Phonology, Semantics and the English Past Tense. *Current Directions in Pysch. Science, 13,* 156-160.

Kail, R. (1994). A method for studying the generalized slowing hypothesis in children with specific language impairment. *J. Speech & Hearing Research, 37,* 418-421.

Magnuson, J.S., Tanenhaus, M.K., & Aslin, R.N. (2008). Immediate effects of form-class constraints on spoken word recognition. *Cognition, 108(3),* 866-873.

Magnuson, J. S., Tanenhaus, M. K., Aslin, R. N., and Dahan, D. (2003). The time course of spoken word recognition and learning: Studies with artificial lexicons. *J. Experimental Psychology:General, 132(2)*, 202-227.

Marslen-Wilson, W.D., & Warren, P. (1994). Levels of perceptual representation and process in lexical access: Words, phonemes, and features. *Psy. Rev., 101*, 653-675.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psych., 18*, 1-86.

McMurray, B, Samelson, V.M., Lee, S.H., & Tomblin J.B. (2010). Individual differences in online spoken word recognition: Implications for SLI. *Cog. Psych., 60,* 1-39.

Mirman, D., Dixon, J. A., & Magnuson, J. S. (2008). Statistical and computational models of the visual world paradigm: Growth curves and individual differences. *Journal of Memory & Language, 59(4),* 475-494.

Mirman, D., McClelland, J.L., & Holt, L.L. (2006). An interactive Hebbian account of lexically guided tuning of speech perception. *Psychonomic Bulletin & Review, 13*, 958-965.

Ramus, F., & Szenkovits, G. (2008). What phonological deficit? *Quarterly J. Experimental Psych., 61*, 129-141.

Shankweiler, D. P., Mencl, W. E., Braze, D., Tabor, W., Pugh, K. R., & Fulbright, R. K. (2008). Reading Differences and Brain: Cortical Integration of Speech and Print in Sentence Processing Varies with Reader Skill. *Developmental Neuropsychology, 33.6* 745-776.

Shankweiler, D., Liberman, I.Y., Mark, L.S., Fowler, C.A. & Fischer, F.W. (1979). The speech code and learning to read. *Journal of Experimental Psychology: Human Learning and Memory, 5,* 531-545.

Snowling, M.J., 1981. Phonemic deficits in developmental dyslexia. *Psychological Research, 43*, 219-234.

Snowling, M. J. & Hulme, C. (2005). *The Science of Reading: A Handbook*. Blackwell.

Strauss, T. J., Harris, H. D., & Magnuson, J. S. (2007). jTRACE : A reimplementation and extension of the TRACE model of speech perception and spoken word recognition. *Behavior Research Methods, 39,* 19-30.

Tallal, P. (1980). Auditory temporal perception, phonics, and reading disabilities in children. *Brain and Language, 9,* 182-198.

Tanenhaus, M. K., Spivey-Knowlton, M., Eberhard, K., & Sedivy, J. C. (1995). Integration of visual and linguistic information is spoken-language comprehension. *Science, 268*, 1632–1634.

Ziegler, J., & Goswami, U. (2005). Reading Acquisition, Developmental Dyslexia, and Skilled Reading Across Languages: A Psycholinguistic Grain Size Theory. *Psychological Bulletin, 131*, 3-29.

Phonological Encoding in Word Naming and Word Typing

Jenn-Yeu Chen (psyjyc@mail.ncku.edu.tw)
Institute of Cognitive Science, National Cheng Kung University
1 University Road, Tainan City, Taiwan 701

Cheng-Yi Li (adamli100@gmail.com)
Institute of Cognitive Science, National Cheng Kung University
1 University Road, Tainan City, Taiwan 701

Abstract

The process of phonological encoding was investigated in primed word naming and word typing with Chinese monosyllabic words. The target words shared or did not share the onset consonants with the prime words. The stimulus onset asynchrony (SOA) was 100 ms or 300 ms. Typing required the participants to enter the phonetic letters of the target word, which correspond roughly to the onset and the rhyme of the word's syllable. Regardless of SOAs, response times were shorter in the related condition than in the unrelated condition (an onset priming effect) for word typing, but were similar for word naming. The results suggest that naming and typing in Chinese may involve somewhat different phonological encoding processes even though both tasks require accessing the phonological codes. It is hypothesized that phonological encoding in Chinese is syllable driven in word naming, but is segment driven in word typing.

Keywords: Naming, Typing; Phonological Encoding; Word Production.

Introduction

The organization of a production system, natural or artifactual, must be constrained by the kind of outputs it is designed to produce. The production system for a car is organized differently than the production system for an airplane. The production systems for different kinds of cars (sedan vs. truck) are probably also organized differently. For natural languages, it has been shown recently that the word form encoding component of the word production system is organized differently for different languages such as Dutch and Chinese. In the present study, we show that the process of phonological encoding in word production is also somewhat different for naming and typing within the same language, Chinese, even though both tasks involve accessing phonological codes.

The phonological codes of a word may contain the syllables (e.g., /seg/ and /ment/ for the word 'segment'), the individual segments (e.g., /s, ɛ, g, m, ə, n, t/) and the prosodic features (the stress pattern ˈσσ) of the word. Theories of word production vary in whether the syllables are hypothesized as stored and retrieved units (Dell, 1986; Ferrand, Segui, & Grainger, 1996; Santiago, MacKay, Palma, & Rho, 2000), or whether they are assembled online during phonological encoding (Levelt, Meyer, & Roelofs, 1999). According to the model proposed by Levelt and colleagues (the LMR model), phonological encoding starts with retrieving the segmental contents and the wordshape frame of the word to be produced. The segments are then assigned to the slots in the wordshape frame sequentially from left to right according to the phonotactic principles of the language. The result of this segment-to-frame association (called syllabification) is phonological syllables, which are fed to the next stage of processing for phonetic encoding and articulation. In this model (illustrated in Figure 3), the syllables are assembled products of phonological encoding. The model was solely based on empirical evidence from Indo-European languages such as English, Dutch, and German.

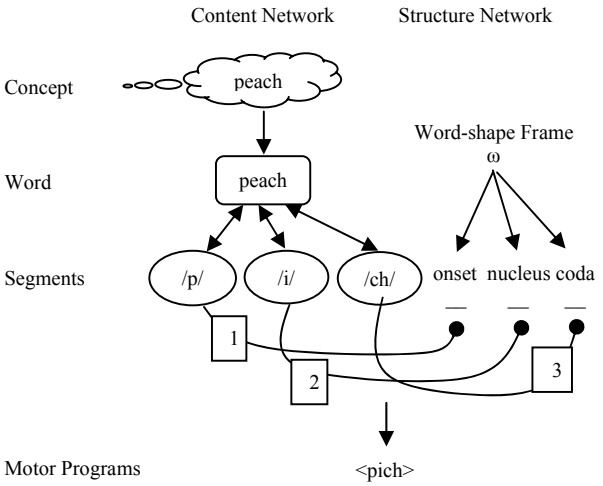


Figure 3: Production of an English CVC monosyllable for naming. Arrows signify activation. Button terminals signify assignment of contents to structures.

Assuming a similar architecture to the LMR model, Chen and colleagues (Chen, Dell, & Chen, 2002; O'Seaghdha, Chen & Chen, 2010) examined Mandarin Chinese recently but proposed that phonological encoding starts with retrieving the stored syllables of the word. The segmental contents and the syllable frame of each syllable are then retrieved for the same kind of segment-to-frame association process as in the LMR model. The difference between the LMR model (Figure 3) and the Chinese model (illustrated in Figure 4) can be characterized as the difference between segment-driven and syllable-driven processes. The difference, as we maintained previously, is due to the

different design characteristics of the phonological systems in the respective languages. The English and the Dutch phonology emphasize words and segments (large number of syllable types, syllable boundaries are often ambiguous, segments may be resyllabified in a different context, syllables carry stress and are not equally weighted), whereas the Chinese phonology emphasize syllables (clear syllable boundaries, syllabification prohibited, simple syllable structures, small number of syllable types).

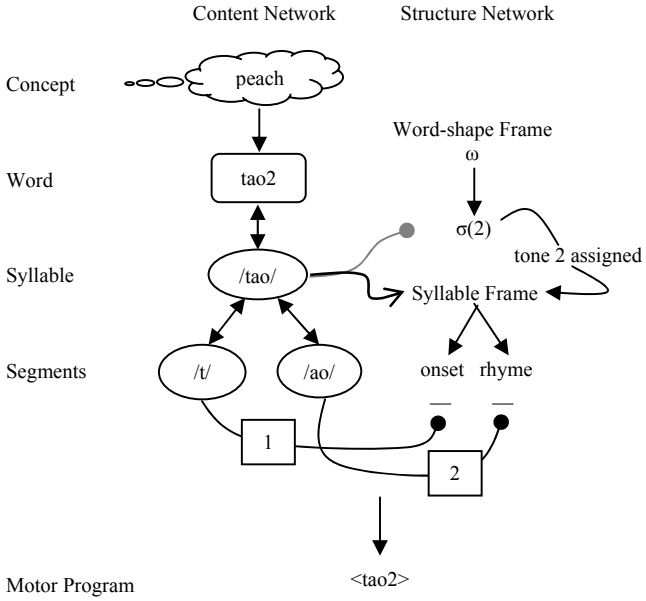


Figure 4: Production of a CV monosyllable in Mandarin for speaking. Arrows signify activation. Button terminals signify assignment of contents to structures. For simplicity, a rhyme is represented as a segment.

Typing is a production task which resembles speaking in many aspects except perhaps for the motor outputs. For speaking, the motor output involves moving several articulators simultaneously and sequentially in a highly coordinated fashion to produce syllable-sized gestures (MacNeilage, 1998). For typing, the motor output involves moving the fingers of both hands entirely discretely and sequentially (even though planning is done in parallel, Rumelhart & Norman, 1982; Salthouse, 1986). Both tasks, however, will need to access the phonological contents of a word, especially true in Chinese. The question we asked was whether the same kind of phonological encoding process operates in speaking and typing.

One hypothesis is that the same kind of phonological encoding process operates in speaking and typing. Although the motor outputs of typing are more discrete and sequential than those of speaking, the individual keystrokes may still be organized hierarchically into chunks of the word and the syllable sizes (Cooper, 1983). Accordingly, the entire process of producing a word would be identical in speaking and typing except that the specific motor muscles involved are different. Recent studies by Damian and colleagues (Zhang and Damian, in press; Shen and Damian, 2009) with English showed that writing accessed orthographic codes (graphemes) whereas speaking accessed phonological codes (phonemes or segments). However, writing involves a segment driven process just like speaking. In English, writing and typing are similar enough so that Damian and Shen's findings can be taken as the basis for the same-process hypothesis when speaking and typing are being compared. In Chinese, however, writing and speaking are distinctly different, and so are writing and typing (to be explained immediately). Therefore, hypothesizing about the phonological encoding processes in speaking and typing Chinese requires some explanation of the way Chinese characters are typed.

Chinese characters are logographs. Writing a Chinese character involves writing the strokes in a specific order and configuration. In contrast, the most commonly used methods of typing a Chinese character (*zhuyin* in Taiwan and *pinyin* in Mainland China) involve entering the phonetic letters of the word such as the onset consonant, the medial vowel, the rhyme, and the tone (for the *zhuyin* method), which bear no resemblance whatsoever with the strokes in writing. Nevertheless, what displays on the computer screen after phonetic typing is the orthographic form of the character. The phonological form is also shown, but only as an intermediate output before the typist hits the Enter key.

Given the way Chinese characters are typically typed (the phonetic typing method), it can be reasonably assumed that word typing might involve accessing the phonological codes of a word much like word speaking or naming. It can, then, be asked whether the same or different kinds of phonological encoding process underlie Chinese word naming and word typing.

Because previous studies have shown that speaking a word in Chinese is syllable driven, the same-process hypothesis predicts that typing a word in Chinese is also syllable driven. The contrasting hypothesis is that somewhat different kinds of phonological encoding process operate in speaking and typing. Because the individual keystrokes are organized discretely in typing, the process might emphasize the individual keystrokes, and, accordingly, the segments, more than the higher order units like the syllables. Support for the emphasis comes from analysis of typing errors, which, according to Norman and Rumelhart (1983), suggest that words are parsed into single-letter and two-letter units for execution. There may be two consequences of this emphasis. First, syllabification may not be necessary. Once the segmental contents of a word are retrieved, they are mapped to segment-sized motor programs for execution. This is different from speaking, where the initially retrieved segmental contents of a word are assembled back to syllables in order to be mapped to syllable-sized motor programs for execution. Second, the influence of higher-order units such as syllables and words may be weak because the end products are segments. In sum, the different-process hypothesis predicts that typing a word in

Chinese might be segment driven (illustrated in Figure 5), in contrast to the syllable driven process in speaking (Figure 4).

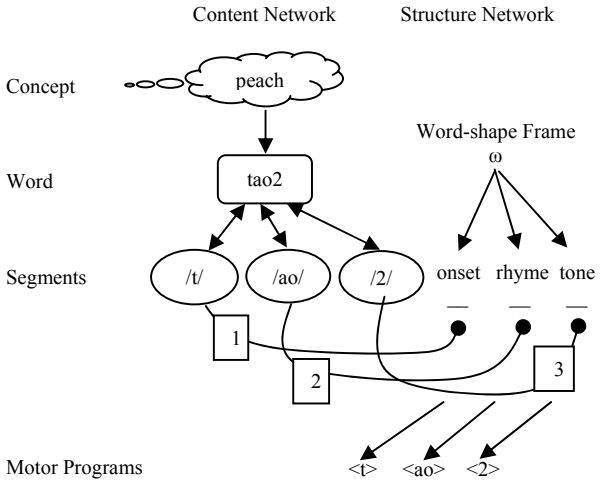


Figure 5: Production of a CV monosyllable in Mandarin for typing as modified from the speaking model of Figure 4 according to the prediction of the different-process hypothesis.

To test the hypotheses against each other, we employed a primed word naming task and a primed word typing task using Chinese monosyllabic words. Typing was performed with the zhuyin method. The target words shared or did not share the onset consonants with the prime words. We compared the onset priming effects (the difference in response times between the related and the unrelated conditions) between the two tasks. Because previous studies have observed no onset effect in Chinese speaking tasks (masked priming and implicit priming, Chen, Chen, & Dell, 2002; Chen, Lin, & Ferrand, 2003; O'Seaghdha, Chen, & Chen, 2010), the same-process hypothesis predicts no onset priming effects for either task, whereas the different-process hypothesis predicts no priming for the naming task but significant onset priming for the typing task.

In addition to manipulating phonological overlap, we also manipulated orthographic overlap such that the prime and the target words shared or did not share the first radical. With an unmasked priming procedure and stimulus onset asynchronies of varying lengths, previous studies showed that an orthographic overlap would produce positive (43 ms), negative (57 and 85 ms) or no (115 ms) priming in a word naming task (Perfetti & Tan 1998). According to one explanation, negative priming is due to form-related competition whereby the episodic memory trace of the prime is reactivated by the shared orthographic form in the target and competes with the target for phonological encoding (O'Seaghdha & Marin, 2000). If the level of competition is lexical, both the same-process hypothesis and the different-process hypothesis predict similar negative priming for naming and typing. If the level of competition is phonological, the same-process hypothesis predicts similar negative priming for the two tasks, while the different-process hypothesis predicts greater negative priming for naming than for typing (assuming syllable competition is greater than segment competition) or similar negative priming for naming and typing (assuming syllable competition is no greater than segment competition). Due to the uncertainties about the level of competition, the extent of competition, and the effect of SOA, the orthographic manipulation was included more for an explorative purpose than for testing the present hypotheses.

Method

Participants

Twenty-six native Mandarin Chinese speakers were recruited for the typing task and twenty-two for the naming task. They were students from National Cheng Kung University and the surrounding universities. The participants for the typing task were all habitual zhuyin typists with an average typing speed of 62.7 characters per min. All the participants had normal or corrected-to-normal vision and they were paid for participation.

Design and Materials

Thirty characters served as targets. Each was paired with four types of prime characters according to whether it shared the onset consonant or the first radical with the prime. An example is given in Table 1. The frequencies and the stroke numbers were matched among the four types of primes. There were a total of 120 pairs, which were randomly ordered for each participant. The experiment included one between-subjects factor (typing method) and three within-subjects factors (phonological relatedness, orthographic relatedness, and stimulus onset asynchrony), each with two levels. The SOA was either 100 or 300 ms. For each of the four types of prime-target pairs, half was presented with 100-ms SOA and the other half with 300 ms. The half which was presented with 100-ms SOA for half of the participants was presented with 300-ms SOA for the other half of the participants, and vice versa.

Table 1: An example of the prime-target pairs as a function of phonological and orthographic relatedness between the primes and the targets. The mean frequencies and the mean stroke numbers of the prime characters (standard deviations in parentheses) are also given.

		+Onset	-Onset
+Radical	Characters	梯- 桃	概-桃
	Pinyins	ti1-tao2	gai4-tao2
	Mean frequency	224 (34)	258 (119)
	Mean strokes	11.6 (2.8)	10.9 (3.6)
-Radical	Characters	泰-桃	棄-桃
	Pinyins	tai4-tao2	qi4-tao2
	Mean frequency	238(87)	276 (112)
	Mean strokes	11.9 (3.5)	12.4 (3.4)

Apparatus and Procedure

The experiment was programmed in Visual Basic for the typing task and in E-Prime for the naming task. Both were run on a personal computer (Intel® Core™2 Quad CPU, Q6600@2.40GHz) with a 20-inch LED screen (32bits, 1400x1050 pixels, 8-ms refresh rate), a standard keyboard, and a microphone. For the typing task, the experiment began with a familiarization phase, followed by a practice block of six trials and the experiment block of 120 trials. To ensure the participants knew the exact phonetic letters of each character used in the experiment, all of the characters (120 primes and 30 targets) were shown one at a time in a random order. The participants had to type in the correct phonetic letters of each character. If a mistake was made, the correct answer was offered. All of the incorrectly-answered characters were presented again at the end of the list. The procedure was repeated until no characters were incorrectly typed.

A trial for the practice and the experimental blocks consisted of a fixation cross appearing at the center of the computer screen for 1 sec. The prime character appeared in the Ximing font for 100 or 300 ms, followed immediately by the target character in the Biaokai font. Each character subtended about 2˚visual angle horizontally and vertically from a viewing distance of 50 cm. The participants were asked to type in the phonetic letters of the target character as quickly and accurately as possible. The response time was recorded and measured to the accuracy of millisecond from the onset of the target character to the first keystroke of the typing response. Response accuracy was also recorded.

All of the participants completed the typing task before coming back a month later for the naming task. For the naming task, exactly the same procedure was employed, except that the participants were asked to name the characters out loud. The response time was registered at the onset of the participants' vocal response.

Results

Typing

Errors were infrequent (less than 6%) and were not analyzed. Response times (RT) for the correct trials were analyzed using a linear mixed model (Statistical Analytic System, the PROC MIXED procedure) with subjects and items as random-effect variables and phonological relatedness, orthographic relatedness and SOA as fixed-effect variables. The mean RTs as a function of phonological relatedness, orthographic relatedness and SOA are plotted in Figure 1. The most notable effects in the figure are that of SOA and that of phonological relatedness. Response times were faster under 300-ms SOA than under 100-ms SOA: $F(1, 2895) = 62.4$, $p < .0001$. Response times were also faster when the prime and the target shared the onset consonant than when they did not (a positive onset priming effect of 30 ms): $F(1, 2895) = 35.3$, $p < .0001$. Response times were somewhat slower when the prime and the target shared the first radical than when they did not (a negative orthographic priming effect of 9.5 ms): $F(1, 2895) = 4.2$, $p < .05$. None of the interactions were significant, p's > .2.

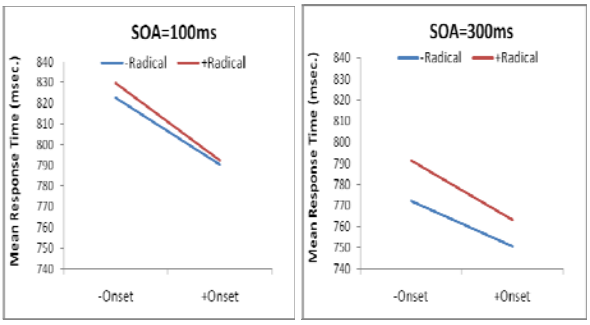


Figure 1: Mean RTs as a function of phonological relatedness, orthographic relatedness, and SOA from the typing task.

Naming

Errors were less than 2%. Response times were similarly analyzed as they were for typing. The mean RTs as a function of phonological relatedness, orthographic relatedness and SOA are plotted in Figure 2. The only significant effect in the figure is that of SOA. Response times were faster under 300-ms SOA than under 100-ms SOA: $F(1, 2516) = 323.9$, $p < .0001$. Response times were somewhat slower when the prime and the target shared the first radical than when they did not (a negative orthographic priming effect of 5.6 ms), but the effect fell short of the conventional level of significance, $p = .134$. None of the other effects were significant, p's ranging from . 454 to .932.

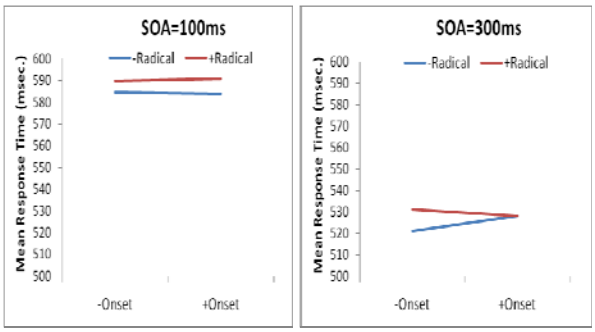


Figure 2: Mean RTs as a function of phonological relatedness, orthographic relatedness, and SOA from the naming task.

Combined Analysis

When the data from both tasks were included in the analysis with task as an additional fixed-effect variable, the results revealed significant main effects for all fixed-effect variables: $F(1, 5440) = 68.3$, $p < .0001$ for task, $F(1, 5440) = 259.3$, $p < .0001$ for SOA, $F(1, 5440) = 21.5$, $p < .0001$ for phonological relatedness, and $F(1, 5440) = 5.9$, $p < .02$

for orthographic relatedness. Importantly, the task x phonological relatedness interaction was significant: $F(1, 5440) = 23.5$, $p < .0001$, confirming the different findings in the separate analyses. The task x orthographic relatedness interaction was not significant, $F < 1$, confirming the similar findings in the separate analyses. The only remaining significant effect was that of task x SOA. All of the other effects were nonsignificant, p's ranging from .262 to .997.

Discussion

Single word naming and typing in Chinese both involve accessing the phonological codes of the word, but the motor outputs are different. Does word typing involve the same or different phonological encoding processes from word naming? Using an unmasked priming procedure and manipulating phonological and orthographic overlaps between the prime and the target characters, we observed significant positive onset (phonological) priming in the typing task, but no priming in the naming task. At the same time, we also observed significant and comparable negative radical (orthographic) priming in both tasks.

The orthographic priming effects were similar in the naming and the typing tasks. It is important to note that these effects did not vary with phonological relatedness. That is, whether the prime and the target shared the onset consonant did not affect the orthographic priming effect. This could suggest (1) lexical competition whereby the reactivated prime word competed with the target word for phonological encoding, or (2) phonological competition whereby the phonological contents of the prime and the target competed for selection. It is not possible for the present study to determine whether the observed orthographic priming was due to lexical or phonological competition, and whether phonological competition between syllables is equivalent in magnitude to that between segments. As a result, the similar orthographic priming effects are interesting, but uninformative for testing our hypotheses. The following discussion will focus on the different phonological priming effects.

The different phonological priming effects observed in word naming and word typing suggest that the two types of tasks likely involve somewhat different phonological encoding processes, but an alternative account needs to be considered first. The alternative account would argue that the segmental effect in word typing occurred at the stage of motor output. That is, knowing ahead of time the first segment of a word allowed the participants to prepare the motor act of typing that segment, or even to start typing before they had retrieved the response word. Either possibility is highly unlikely because the prime-target SOAs were too short (100 and 300 ms) to allow the processing of the prime to proceed to the motor stage in time to benefit the production of the target at that level. The finding that onset priming in word typing did not vary with SOAs also helps to rule out these possibilities.

The different phonological priming effects can be explained by postulating two different models for word naming and word typing. Figure 3 illustrates a production model for speaking a monosyllabic Chinese word proposed previously (O'Seaghdha, Chen, & Chen, 2010). The model is applicable to a word naming task if we focus on the production phase of naming and also ignore the concept level. In the model, syllables are retrieved as chunks. The segmental contents and the syllable frame are separately spelled out, followed by the sequential assignment of the individual segments to the categorized slots in the frame. The result of this phonological encoding process is a syllable-sized motor program for articulation. Figure 4 illustrates the same model modified for typing. In this model, tone is assumed to be one of the segmental contents of a syllable and is assigned last; there is no explicit syllable level; and the sequential assignment of the individual segments to their categorized slots leads to several segment-sized motor programs, rather than one syllable-sized motor program.

The segment-sized output characterizes an important feature of the typing model and distinguishes it from the naming model. In fact, it also contrasts with the naming model hypothesized for Germanic languages (Roelofs, 1997; Levelt, Meyer, & Roelofs, 1999). The different characteristics of the outputs for typing and naming serve to constrain the processing at earlier stages differently. Specifically, the syllable-sized output for naming prescribes that phonological encoding address the syllable, whereas the segment-sized output for typing prescribes that it address the segment in the planning process. The segment-addressing system gives rise to the onset priming effect observed in the typing task, whereas the syllable-addressing system produces no onset priming in the naming task.

To summarize, the results of the present study, as far as the onset priming effects go, support the hypothesis that somewhat different phonological encoding processes are involved in speaking and typing. The process is syllable-driven in speaking, but segment-driven in typing. And this is due to the different natures of the outputs the two tasks aim to produce.

In our previous work (Chen, Chen, & Dell, 2002; Chen, Lin, & Ferrand, 2003; O'Seaghdha, Chen & Chen, 2010), we emphasized and investigated cross-linguistic differences in the design characteristics of Chinese and English/Dutch (with respect to the phonological system) and how the processing mechanisms of phonological encoding differ accordingly. In the present study, we highlighted another important factor that might modulate the processing mechanism *within the same language*. The idea that the specific form of the output must in some way drive the form of the intermediate representations in a production system should surprise no one. The input of a production system can differ greatly from the output. Given that production is a process that translates a specific input to a specific output, the final form of the output (the goal state of the production system) must require that the intermediate representations approach that form, or else production would fail. This idea

is consistent with any system that is adaptive and goal-directed.

The hypothesis that the forms of the internal representations are constrained by the form of the output in a production system is consistent with previous models that postulate different modality-specific lexicons for speaking and writing, based on neuropsychological evidence (Ellis & Young, 1988; Caramazza, 1997). It is also consistent with the theoretical concept of embodiment in cognition (Lakoff & Johnson, 1999; Clark & Chalmers, 1998).

Many issues remain and further work awaits researchers. Convergent evidence is needed from other production tasks and procedures (e.g., word naming and picture naming with masked primes, the form preparation task). The difference between word typing in Chinese and word naming in English deserves investigations. Even though both involve segment-driven processes, they may be motivated differently. As mentioned earlier, the well-cited model of word production for English/Dutch assumes syllable-sized outputs. If the assumption is valid, the segment-driven process of phonological encoding must find a different motivation in the English/Dutch speaking system than that in the Chinese typing system. On the other hand, it could be that the assumption has been false. The hypothesis also bears a broader implication for understanding the cognitive system in general.

Finally, word typing (or typewriting) used to be a special skill of a small group of professionals. As a result, research on typewriting has been sparse. With the increasing popularity of computer word processing, typewriting has become a common skill of literacy like handwriting. As a production task, it is time for the production researchers to begin investigating this new technologically-driven skill of the digital age in order to understand its similarities and differences from the speaking task.

Acknowledgments

This work was supported by the NSC96-2413-H-006-004-MY2 grant.

References

Caramazza, A. (1997). How many levels of processing are there in lexical access? *Cognitive Neuropsychology*, *14*, 177-208.

Chen, J.-Y., Chen, T.-M., & Dell, G. S. (2002). Word-form encoding in Mandarin Chinese as assessed by the implicit priming task. *Journal of Memory and Language*, *46*, 751-781.

Chen, J.-Y., Lin, W.-C., & Ferrand, L. (2003). Masked priming of the syllable in Mandarin Chinese speech production. *Chinese Journal of Psychology*, *45*, 107-120.

Clark, A. and Chalmers, D. (1998). The extended mind. *Analysis*, 58, 7-19.

Cooper, W. E., (1983). Introduction. In W.E. Cooper (Ed.), *Cognitive aspects of skilled typewriting* (pp. 1-38). New York: Springer-Verlag.

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review*, *93*, 283-321.

Ellis, A. W., & Young, A. W. (1988). *Human cognitive neuropsychology*. Hove, UK: LEA.

Ferrand, L., Segui, J., & Grainger, J. (1996). Masked priming of word and picture naming: The role of syllabic units. *Journal of Memory and Language*, *35*, 708-723.

Lakoff, G., & Johnson, M. (1999). *Philosophy in the flesh: The embodied mind and its challenge to Western thought*. New York, NY: Basic Books.

Levelt, W. J. M, Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, *22*, 1–75.

MacNeilage, P. F. (1998). The frame/content theory of evolution of speech production. *Behavioral and Brain Sciences*, *21*, 499-546.

Norman, D. A., & Rumelhart, D. E. (1983). Studies of typing from the LNR Research Group. In W.E. Cooper (Ed.), *Cognitive aspects of skilled typewriting* (pp. 45-65). New York: Springer-Verlag.

O'Seaghdha, P. G., Chen, J.-Y., & Chen, T.-M. (2010). Proximate units in word production: phonological encoding begins with syllables in Mandarin Chinese but with segments in English. *Cognition*, *115*, 282-302.

O'Seaghdha, P. G. & Marin, J. W. (2000). Phonological competition and cooperation in form-related priming: Sequential and nonsequential processes in word production. *Journal of Experimental Psychology: Human Perception & Performance*, *26*, 57-73.

Perfetti, C. A., & Tan, L. H. (1998). The time-course of graphic, phonological, and semantic activation in Chinese character identification. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 1–18.

Roelofs, A. (1997). The WEAVER model of word-form encoding in speech production. *Cognition*, *64*, 249–284.

Rumelhart, D. E., & Norman, D. A. (1982). Simulating a skilled typist: A study of skilled cognitive-motor performance. *Cognitive Science*, *6*, 1-36.

Salthouse, T. A. (1986). Perceptual cognitive, and motoric aspects of transcription typing. *Psychological Bulletin*, *99*, 303-319.

Santiago, J., MacKay, D. G., Palma, A., & Rho, C. (2000). Sequential activation processes in producing words and syllables: Evidence from picture naming. *Language and Cognitive Processes*, *15*, 1–44.

Shen, X., & Damian, M. F. (2009). Role of phonology in handwritten word production. Poster presented at the 15th Annual Conference on Architectures and Mechanisms for Language Processing (AMLaP), 7-9 September, Barcelona.

Zhang, Q., & Damian, M. F. (in press). Impact of phonology on the generation of handwritten responses: Evdience from picture-word interference tasks. *Memory & Cognition*.

Visual and Task characteristics may explain hemispheric asymmetry in visual word recognition

Kloser Chee Fung Cheung (kloser@hku.hk)
Janet Hui-wen Hsiao (jhsiao@hku.hk)
Department of Psychology, University of Hong Kong
604 Knowles Building, Pokfulam Road, Hong Kong SAR

Abstract

Previous studies proposed that the left hemisphere (LH) lateralization in English word recognition is because of the LH superiority in language processing. Nevertheless, Chinese character recognition has been shown to be more bilateral or right hemisphere (RH) lateralized and thus is a counter example of this claim. Through computational modeling, here we show that at least two factors other than language lateralization may influence hemispheric asymmetry in visual word recognition: (1) Visual similarity among words, which can be influenced by the ratio between the alphabet size and the lexicon size and the visual similarity among letters: We show that the more similar the words are in the lexicon, the more high spatial frequency (HSF) information is required to distinguish them, and this leads to more LH lateralization (2) The requirement to decompose a word into letters in order to map them to corresponding phonemes in pronunciation: We show that letter identity mapping requires more HSF information than word identity mapping, and alphabetic reading requires more HSF information than logographic reading; this leads to more LH lateralization in alphabetic languages. These two visual and task characteristic factors alone may explain differences in lateralization between English word and Chinese character recognition, without assuming the influence from language lateralization.

Keywords: visual word recognition, hemispheric asymmetry, computational modeling

Introduction

Lateralization in visual word recognition

Words, which surround us ever since our childhood, have been extensively studied in the research on visual recognition. Previous studies have consistently shown a left hemisphere (LH) lateralization effect in visual word recognition in alphabetic languages such as English. A classical right visual field (RVF)/LH advantage in reading English words (or words in alphabetic languages) has been demonstrated first in tachistoscopic recognition (e.g., Bryden & Rainey, 1963) and consistently reported in other word recognition tasks, such as word naming (Brysbaert & d'Ydewalle, 1990) and lexical decision tasks (Faust, Babkoff, & Kravetz, 1995). Data from fMRI studies have shown a region inside the left fusiform area (Visual Word Form Area, VWFA) responding selectively to words (e.g., McCandliss, Cohen, & Dehaene, 2003). ERP studies also show that words elicit a larger N170 in the LH than strings of symbols (e.g., Maurer, Brandeis, & Dehaene, 2005). This RVF/LH advantage in visual word recognition in alphabetic languages has been argued to be because of the LH lateralization in language processing (e.g., Voyer, 1996).

Nevertheless, this claim has been challenged by at least one counter example, that is, the recognition of Chinese characters. In contrast to the RVF/LH advantage in the recognition of English words, the recognition of Chinese characters, a logographic writing system, has been shown to have a left visual field/right hemisphere (LVF/RH) advantage in orthographic processing, demonstrated in tachistoscopic recognition tasks (e.g., Tzeng et al., 1979; Cheng & Yang, 1989). In addition, Hsiao and Cottrell (2009) showed a left side bias effect in Chinese character perception in Chinese readers (experts), but not in non-Chinese readers (novices). This left side bias effect also suggests the RH involvement in Chinese character processing.

As for phonological processing in Chinese character recognition, Weekes and Zhang (1999) reported phonological priming effects on the recognition of phonetic compounds (i.e. characters with a phonetic radical that has information about character pronunciation) when the characters were presented in the RVF/LH but not in the LVF/RH; this effect was not observed in integrated characters (i.e. characters that do not have a phonetic radical; Weekes, Chen, & Lin, 1998). Thus, research on Chinese character recognition has exhibited a LVF/RH advantage for orthographic processing, and a RVF/LH advantage for phonological processing, especially for phonetic compounds. ERP and fMRI studies of Chinese character recognition have also shown a more bilateral or RH-lateralized activation in the visual system than those of English word recognition (e.g., Tan et al., 2000; Liu & Perfetti, 2003), which is consistent with the behavioral data.

The RH advantage in Chinese character recognition has been argued to reflect the RH superiority in handling holistic pattern recognition (Tzeng et al., 1979). Nevertheless, findings in later studies do not support this claim. For example, Cheng and Yang (1989) showed no laterality effect in the recognition of non-characters and pseudo-characters, suggesting that this RH advantage may be related to lexical knowledge of Chinese characters or learning experience. Also, in contrast to Tzeng et al.'s claim, Hsiao and Cottrell (2009) showed a reduced holistic processing effect in Chinese readers compared with non-Chinese readers. Thus, it remains unclear why Chinese character recognition and English word recognition involve different hemisphere lateralization.

Hemispheric processing model

In order to investigate why Chinese character and English word recognition involve different hemispheric

lateralization, here we adopt a computational approach, aiming to examine potential factors that may influence hemispheric asymmetry in visual word recognition, since computational modeling approaches enable us to have better control over variables.

Anatomical evidence shows that our visual field is initially split along the vertical midline, and the two visual hemifields are initially contralaterally projected to different hemispheres. In order to examine at which processing stage this split information converges, Hsiao, Shieh, and Cottrell (2008) conducted a hemispheric modeling study of face recognition, aiming to account for the left side bias effect in face perception. They proposed three models with different timing of convergence: early, intermediate and late convergence models (Figure 1). They showed that both the intermediate and late convergence models are able to account for the left side bias effect in face perception, whereas the early convergence model fails to show the effect.

Hsiao et al.'s hemispheric processing model (2008) incorporates several known observations about visual anatomy and neural computation: Gabor responses are used over the input images to simulate neural responses of cells in the early visual system (Lades et al., 1993); Principal Component Analysis (PCA), a biologically plausible linear compression technique (Sanger, 1989), is used to simulate possible information extraction processes beyond the early visual system. This PCA representation then is used as the input to a two-layer neural network (Figure 2).

In addition, the model implements a theory of hemispheric asymmetry in perception, Double Filtering by Frequency theory (DFF, Ivry & Robertson, 1998). The DFF theory argues that information coming into the brain goes through two frequency filtering stages: The first stage involves attentional selection of a task-relevant frequency range. At the second stage, the LH amplifies high frequency information, while the RH amplifies low frequency information. This differential frequency bias in the two hemispheres is implemented in the model by using two sigmoid weighting functions to assign different weights to the Gabor responses in the two hemispheres (Figure 2).

Here we apply Hsiao et al.'s hemispheric processing model (2008) to the modeling of visual word recognition, in order to examine whether visual and task characteristics alone are able to account for the differences in hemispheric lateralization in different languages, without assuming the influence of language processing being LH-lateralized. We introduce our hypothesis below.

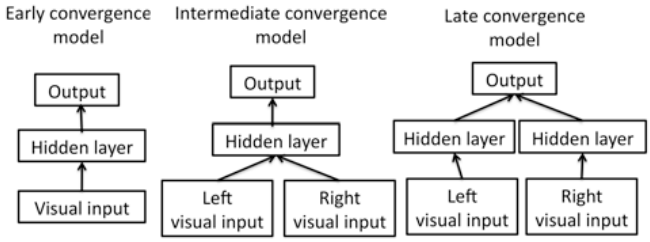


Figure 1: Hemispheric models with different timing of convergence (Hsiao et al., 2008)

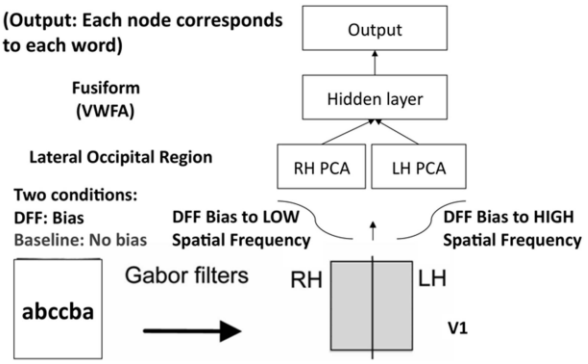


Figure 2: Hsiao et al.'s hemispheric processing model (2008)

Visual and task characteristics of a writing system

Here we test the hypothesis that differences in visual and task characteristics of a writing system alone are able to account for differences in hemispheric lateralization in visual word recognition in different languages. We hypothesize that at least two factors other than language lateralization may influence hemispheric lateralization in visual word recognition:

(1) Visual similarity among words in the lexicon:

The more similar the words look visually in the lexicon, the more high spatial frequency (HSF) information is required to recognize them; this leads to more LH lateralization. We hypothesize that at least two factors may influence visual similarity among words in the lexicon:

(i) Number of letters shared among words in the lexicon: The more letters are shared among words in the lexicon, the more similar the words look visually in the lexicon. This factor is influenced by the ratio between the alphabet size (i.e. the number of letters in the alphabet) and the lexicon size (i.e. the number of words in the lexicon); that is, given a fixed lexicon size, the smaller the alphabet size is, the more number of letters may be shared among the words in the lexicon, and thus the more similar the words look visually in the lexicon.

(ii) Similarity among letters in the alphabet: The more similar the letters in the alphabet look visually, the more similar the words look visually in the lexicon. This factor may be influenced by the number of letters in the alphabet; that is, given a fixed representational space for all possible letters, when we gradually increase the number of letters in the alphabet, it becomes more likely that some letters will look similar to each other (i.e. closer to each other in the space).

According to these two factors, we predict that with a fixed lexicon size, when we gradually increase the alphabet size, the model will first exhibit more and more low spatial frequency (LSF) reliance since the words will share fewer and fewer common letters (factor (i)); when the letters in the alphabet start to look visually similar to each other because of the alphabet size increase, the model will start to exhibit reduced LSF reliance (factor (ii)). In other words, we expect

that there will be an inverted-U-shaped curve in LSF reliance/RH lateralization in the model when we gradually increase the alphabet size given a fixed lexicon size.

(2) The requirement to decompose a word into letters in order to map them into corresponding phonemes in pronunciation

Maurer and McCandliss (2007) proposed the phonological mapping hypothesis to account for the difference in ERP N170 lateralization between faces and words: N170 has been found to be larger in the RH compared with the LH in face recognition, whereas in the recognition of English words, it has been found to be larger in the LH compared with the RH. They argued that given phonological processes are typically left-lateralized (e.g., Price et al., 1997; Rumsey et al., 1997), specialized processing of visual words in visual brain areas also becomes left-lateralized. Accordingly, the LH lateralization of N170 may be specifically related to the influence of grapheme-phoneme conversion established during learning to read. According to this hypothesis, this phonological modulation should be less pronounced in logographic scripts such Chinese (Maurer & McCandliss, 2007).

In contrast to the phonological mapping hypothesis, here we hypothesize that the LH lateralization in English word recognition is due to the requirement to decompose a word into letters, without assuming phonological processes being left-lateralized. We test this hypothesis through two simulations. In the first simulation, we contrast two mapping tasks using the same stimuli: word identity mapping and letter identity mapping. In the word identity mapping task, the model learns to distinguish different words, whereas in the letter identity mapping task, the model learns to identify the constituent letter in each letter position of an input word. We expect that the letter identity mapping task will require more HSF information (i.e. LH lateralization) compared with the word identity mapping task[1].

In the second simulation, instead of mapping word image input to either word or letter identities, we model visual word recognition more realistically by mapping them to pronunciations. We use an artificial lexicon with Korean-character-like pseudo-characters as the orthography. Two pronunciation conditions are created: in the alphabetic reading condition, each component (letter) of a character maps to a consonant or vowel in pronunciation systematically, whereas in the logographic reading condition, each character maps to a pronunciation randomly without a systematic relationship between its orthographic components (letters) and the phonemes in pronunciation. We expect that the alphabetic reading condition will require more HSF information (i.e. more LH lateralization) compared with the logographic reading condition.

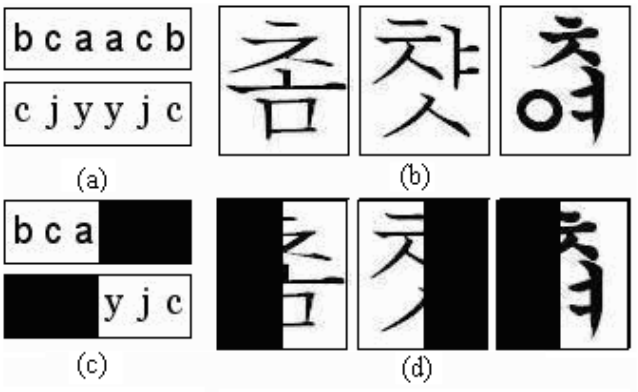


Figure 3: Images used in the current study: (a) palindrome English pseudo-words; (b) Korean pseudo-characters (from left to right, vertical structure, top heavy structure, and bottom heavy structure); (c) & (d) Left and right damaged images of the English pseudo-words and the Korean pseudo-characters

Modeling Method and Results

To test our hypotheses, we applied the intermediate convergence model proposed by Hsiao et al. (2008) to visual word recognition. In the model, the input word images were first filtered with a rigid grid of overlapping 2D Gabor filters (Daugman, 1985) to obtain Gabor responses. At each grid, we used Gabor filters of eight orientations and a fixed number of scales. The number of scales used depended on the task-relevant frequency range, which was determined according to the smaller dimension of the images; the highest frequency scale did not exceed the smaller dimension of the images (following Hsiao et al., 2008). In the current simulation, the dimensions of the two types of images used were 35 x 100 for the English pseudo-words and 70 x 80 for the Korean pseudo-characters (see Figure 3); thus the number of scales for English pseudo-word images was five ($2^5 = 32 < 35$, and $2^6 = 64 > 35$) and that for Korean pseudo-character images was six ($2^6 = 64 < 70$, and $2^7 = 128 > 70$). We applied the Gabor filters to a 5x18 grid of points on each English pseudo-word image, and to a 12x14 grid of points on each Korean pseudo-character image. So each English pseudo-word image was transformed into a vector of size 3600 (5x18 sample points x 8 orientation x 5 scales) while each Korean pseudo-character image was transformed into a vectors of size 8064 (12x14 sample points x 8 orientations x 6 scales).

After obtaining the Gabor magnitudes, two conditions were created: the baseline condition and the biased condition. In the baseline condition (the control condition), Gabor responses in different scales were given equal weights (i.e. no frequency bias), while in the biased condition, we implemented the second stage of the DFF theory by using a sigmoidal weighting function to bias the Gabor responses on the left half word (RH) to LSFs, and those on the right half word (LH) to HSFs (Figure 2). The perceptual representation of each of the left and right half

[1] Note that we reported some pilot data in Hsiao & Cottrell (2009b). Compared with Hsiao & Cottrell (2009b), here we have revised the hypotheses and modeling methods, and presented brand-new and more complete simulations.

words was compressed by PCA into a 50-element representation each (100 elements in total, following Hsiao et al., 2008)[2]. This PCA representation then was used as the input to a two layer neural network, as shown in Figure 2 (see Hsiao et al., 2008, for more simulation details).

We trained our neural network model to recognize the input images until the performance on the training set reached 100% accuracy. The training algorithm was gradient descent with an adaptive learning rate. To test hemispheric asymmetry effects, in contrast to the previous hemispheric models of face and word recognition (e.g., Hsiao et al., 2008, Hsiao & Cottrell, 2009b), here we did not use "chimeric images" (Figure 3(a) & (b)) as a way to give noise to one side of the stimulus in order to test the model's reliance on either the left or right half of the representation. A potential problem in using this kind of chimeric images for words is some letters may have a similar shape as their mirror images (such as 'o' and 'm' in the English alphabet), while others do not; thus these letters will give non-uniform noise distribution over the mirror-image sides of the chimeric words. Here we avoided this problem by using damaged images (Figure 3(c) & (d).) It was made by setting one half of the PCA representation to zero, so that when mapping these damaged images to their identities, only one of the visual hemifields was used for recognition. The left side bias effect thus was measured as the difference between the accuracy of recognizing a right-side-damaged word (carrying LSF/RH information only) as the original word and the accuracy of recognizing a left-side-damaged word (carrying HSF/LH information only) as the original word.

Visual similarity among words in the lexicon:

We first used images of six-letter English pseudo-words to examine how visual similarity among words in the lexicon influences lateralization in visual word recognition. To counterbalance the information carried in the two visual fields, we used palindrome pseudo-words as the stimuli (e.g., Figure 3(a)). We created artificial lexicons with an increasing alphabet size (a-c, a-e, a-g…), and trained the model to learn each lexicon 50 times. In each of the 50 simulations, 26 palindrome words were chosen randomly from all possible combinations of letters in the alphabet to form the artificial lexicon. In the model, each output node corresponded to a word identity.

In the first lexicon with letters from 'a' to 'c', there were 27 possible combinations: aaaaaa, aabbaa, aaccaa, abaaba, abbbba… The randomly chosen 26 words thus looked very similar to one another. When we increased the alphabet size to include 'a' to 'e', the number of combinations became 125, and the randomly chosen 26 words became more dissimilar visually to one another (i.e. the similarity among words decreased). In other words, the larger the alphabet size was, the lower the visual similarities among words in the lexicon were. Here we examined how the model's lateralization changed when we gradually increased the alphabet size.

In the datasets, we used 8 different fonts for each word, with 4 of them used as the training data, and the other 4 used as the testing data (counterbalanced across the simulations). Thus, in both the training and testing datasets, each word had 4 images of different fonts.

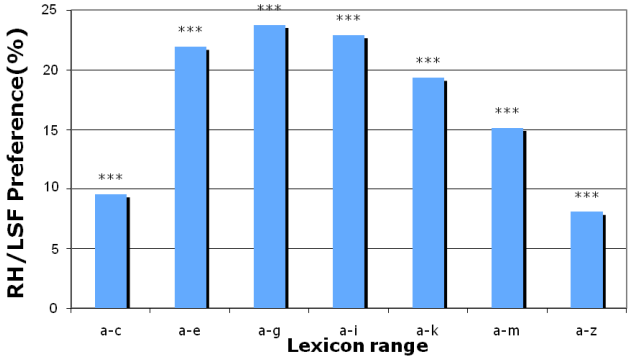


Figure 4: RH/LSF preference in the models trained with lexicons with different alphabet sizes in the word identity mapping task (*p<0.01; **p<0.001; ***p<<0.001).

The results are shown in Figure 4. The RH/LSF preference was defined as the difference in the left side bias effect between the biased condition and the baseline condition; it reflected how much the model preferred the RH/LSF-biased representation over the LH/HSF-biased representation compared with the control condition when no frequency bias was applied (Hsiao et al., 2008). As shown in Figure 4, when the alphabet size was small (e.g., 'a' to 'c'), the model had low RH/LSF preference. When we increased the alphabetic size, the RH/LSF preferences became stronger, and then decreased after the peak at around 'a-g' (i.e., an inverted-U shape in Figure 4).

Thus, the results showed that, when gradually increasing the alphabetic size of the lexicon, the visual similarity among words decreased, and the model relied more on LSFs to distinguish the words. But when the alphabetic size kept increasing, more and more letters with similar shapes were used in the alphabet (e.g., 'c' and 'o', 'b' and 'h', 'm' and 'n'), and the visual similarity among words in the lexicon increased; as the result, the model required more HSFs to distinguish the words.

The requirement to decompose a word into letters

When reading words in alphabetic languages, the readers have to decompose the visual input of a word into its constituent letters/graphemes and map them to the corresponding phonemes. This decomposition may require details of the word image and thus rely more on the HSF information. Here we examined lateralization effects in a letter identity mapping task using the English pseudo-words. Instead of learning to map word images to word identities, the model was trained to map a word image to its constituent letter identities. The output layer of the model

[2] In a separate simulation, we found that using 100 components each made the representation noisier and deteriorated the model's performance.

was divided into 3 parts corresponding to the first 3 letter positions in a word (the end 3 letters were the same as the first 3 since they were palindrome words). The number of nodes in each part was equal to the alphabetic size (see Figure 5).

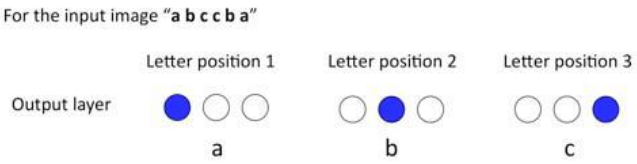


Figure 5: Output layers of the letter-position identity mapping task (Hsiao & Cottrell, 2009b).

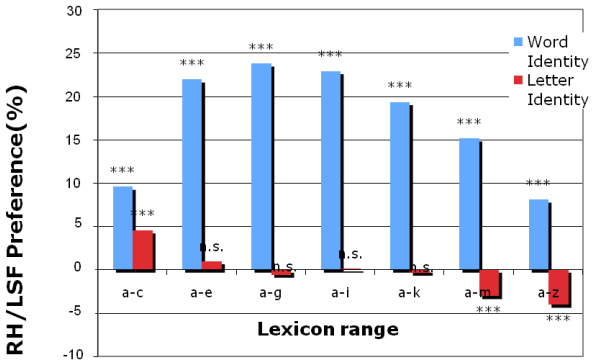


Figure 6: RH/LSF preference in the letter identity mapping task (in red) in the models trained with lexicons of different alphabet sizes, compared with the word identity mapping task (in blue; *p<0.01; **p<0.001; ***p<<0.001).

Figure 6 shows the results. The results showed that compared with the word identity mapping task, the letter identity mapping task required more LH/ HSF information. In addition, in the letter identity task, as the alphabet size increased, the model relied more on LH/HSF information.

In another simulation, we used artificial lexicons with Korean-character-like pseudo-characters to examine hemispheric asymmetry effects in recognizing square-shape characters, and more importantly, to examine hemispheric processing difference between logographic and alphabetic language reading. In this examination, we modeled visual word recognition more realistically by mapping each word input into its pronunciation with a consonant-vowel-consonant structure.

In the datasets, there were also 8 different fonts for each Korean-character-like pseudo-character. Each character consisted of three Korean-alphabet-like letters, arranging in three different structures: vertical, top-heavy, and bottom-heavy (Figure 3(b)). The frequency of each letter appearing in either side of the characters in the lexicon was balanced. In the alphabetic reading condition, each letter systematically mapped to either a vowel or a consonant in pronunciation, whereas in the logographic reading condition, each character mapped to a randomly assigned pronunciation without a systematic letter-phoneme mapping.

Figure 7 shows the results. As shown in the figure, the RH/LSF preference in the logographic reading condition was always stronger than that in the alphabetic reading condition. This result suggests logographic reading requires more LSF information compared with alphabetic reading, and is consistent with the visual word recognition literature showing a more RH lateralization in reading logographic languages such as Chinese compared with alphabetic languages such as English.

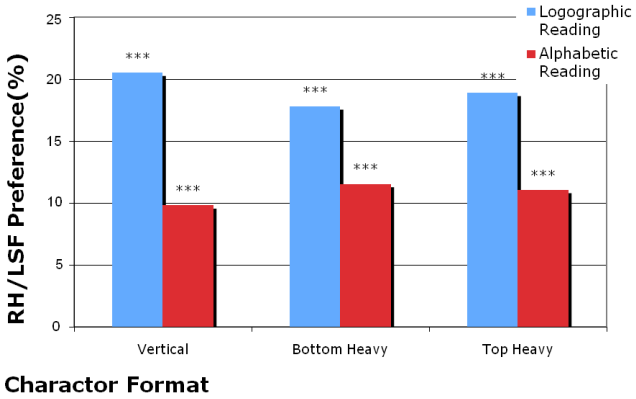


Figure 7: RH/LSF preference in the Korean pseudo-character reading task (*p<0.01; **p<0.001; ***p<<0.001).

Conclusion and Discussion

Visual word recognition in alphabetic languages such as English has been reported to be LH lateralized, and argued to be due to the LH lateralization of language processes. Nevertheless, a RH/LVF advantage has been reported in orthographic processing of Chinese character recognition. In this study, by applying the hemispheric processing model (Hsiao et al., 2008) to visual word recognition, we examined whether visual and task characteristics alone are able to account for differences in hemispheric lateralization in different languages without assuming the influence from language processing being LH-lateralized.

We first showed that visual similarity among words in the lexicon can influence lateralization in visual word recognition. We used artificial lexicons with the same number of words and word length, but with different alphabetic sizes, and trained the model to map word image input to their word identities. The results showed an inverted- U -pattern (Figure 4): When the alphabet size increases, the model initially relies more and more on the RH/LSF information, because words in the lexicon share fewer and fewer common letters and the visual similarity among words in the lexicon decreases. Nevertheless, with further increase of the alphabet size, the model's RH/LSF reliance starts to decrease, because of the increase of visual similarity among letters in the alphabet.

We then showed that the requirement to decompose a word in to its constituent letters can also influence lateralization in visual word recognition. We used the same artificial lexicons but trained the model to perform a letter-identity mapping task instead of the word identity mapping

task. The results showed that decomposition of words into letters requires more HSF information and thus results in more LH lateralization. In addition, we used Korean pseudo-characters to examine lateralization differences between logographical reading and alphabetic reading. The results showed that logographical reading requires more LSF information compared with alphabetic reading, and thus results in more RH-lateralization.

The two factors related to visual and task characteristics of a writing system we proposed here are able to account for the lateralization differences between English word and Chinese character recognition. Compared with Chinese, words in the English lexicon may look more similar to one other, because of the smaller alphabet size (only 26 letters) and a much larger lexicon size (more than 20,000 words). In contrast, Chinese has a smaller lexicon size (about 4500 characters for a native speaker), but a much larger "alphabet" (i.e., more than 1000 stroke patterns). In addition, English is an alphabetic language whereas Chinese is a logographic language. Chinese logographic reading may require more LSF information that leads to more RH-lateralization compared with English alphabetic reading, since logographic reading does not require a decomposition of words into letters in order to map them to corresponding phonemes.

In summary, here we show that visual and task characteristics of a writing system alone may account for lateralization differences in visual word recognition in different languages. Specifically, they are (1) visual similarity among words in the lexicon, and (2) the requirement to decompose a word into letters for performing grapheme-phoneme conversion during learning to read.

Acknowledgement

We are grateful to the HKU Seed Funding Program for Basic Research(project #10400471 to J.H. Hsiao) and the Research Grant Council of Hong Kong (project code: HKU 744509H to J.H. Hsiao).

References

Bryden, M. P. & Rainey, C. A. (1963). Left-right differences in tachistoscopic recognition. *Journal of Experimental Psychology*, 66, 568-571.

Brysbaert, M. & d'Ydewalle, G. (1990). Tachistoscopic presentation of verbal stimuli for assessing cerebral dominance: Reliability data and some practical recommendation. *Neuropsychologia*, 28, 443-455.

Cheng, C. M. & Yang, M. J. (1989). Lateralization in the visual perception of Chinese characters and words. *Brain and Language*, *36*, 669-689.

Daugman, J. G. (1985). Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two dimensional visual cortical filters. *Journal of the Optical Society of America A*, 2, 1160-1169.

Faust, M., Babkoff, H., & Kravetz, S. (1995). Linguistic process in the two cerebral hemispheres: Implications for modularity vs. interactionism. *Journal of Clinical and Experimental Neuropsychology*, 17, 171-192.

Hsiao, J. H. & Cottrell, G. W. (2009). Not all expertise is holistic, but it may be leftist: The case of Chinese character recognition. *Psychological Science*.

Hsiao, J. H. & Cottrell, G. W. (2009b). What is the cause of left hemisphere lateralization of English visual word recognition? Pre-existing language lateralization, or task characteristics? *Proceedings of the Thirty-First Annual Conference of the Cognitive Science Society*.

Hsiao, J. H., Shieh, D., & Cottrell, G. W. (2008). Convergence of the visual filed split: hemispheric modeling of face and object recognition. *Journal of Cognitive Neuroscience*, 20(12), 2298-2307.

Ivry, R. & Robertson, L. C. (1998). The Two Sides of Perception. Cambridge: MIT Press.

Lades, M., Vorbruggen, J. C., Buhmann, J., Lange, J., von der Malsburg, C., Wurtz, R. P., & Konen, W. (1993). Distortion invariant object recognition in the dynamic link architecture. *IEEE Transaction on Computers*, 42, 300-311.

Liu, Y. & Perfetti, C. A. (2003). The Time course of brain activity in reading English and Chinese: An ERP study of Chinese bilinguals. *Human Brain Mapping*, 18, 167-175.

Maurer U., Brandeis, D., & McCandliss, B. (2005). Fast, visual specialization for reading in English revealed by the topography of the N170 ERP response. *Behavioral & Brain Functions*, 1(1), 13.

Maurer U., & McCandliss, D. (2007). The Development of visual experticse for words: the contribution of electrophysiology. In E.L. Grigorenke & A. Naples (Eds.). *Single-Word Reading: Cognitive, behavioral and biological perspectives.* Mahwah, NJ: Lawrence Erlbaum Associates.

McCandliss, B. D., Cohen, L., & Dehaene, S. (2003). The visual word form area: expertise for reading in the fusiform gyrus. *Trends in Cognitive Sciences*, 7, 293-299.

Sanger, T. (1989). An optimality principle for unsupervised learning. In Touretzky, D. (ed) *Advances in Neural Information Processing Systems*, vol. 1, pp. 11-19, San Mateo: Morgan Kaufmann.

Tan, L. H., Spinks, J. A., Gao, J. H., Liu, H. L., Perfetti, C. A., Xiong, J., et al. (2000). Brain activation in the processing of Chinese characters and words: A functional MRI study. *Human Brain Mapping*, 10, 16-27.

Tzeng, O. J. L., Hung, D. L., Cotton, B., & Wang, S. Y. (1979). Visual lateralization effect in reading Chinese characters. *Nature* (London), 282, 499-501.

Voyer, D. (1996). On the magnitude of laterality effects and sex differences in functional literalities. *Laterality*, 1, 51-83.

Weekes, B. S. & Zhang, B. Y. (1999). Chinese character recognition in the left and right visual fields. *Brain & Cognition*, 40, 269-272.

Weekes, B. S., Chen, M. J., & Lin, Y. B., (1998). Orthographic Phonological and semantic priming of Chinese character recognition. *Reading and Writing*, 10 (3-5), 201-222.

Effects of Near and Distant Phonological Neighbors on Picture Naming

Daniel Mirman (mirmand@einstein.edu)
Moss Rehabilitation Research Institute
1200 W. Tabor Rd., Philadelphia, PA 19141, USA

Audrey K. Kittredge (akittre2@illinois.edu)
Gary S. Dell (gdell@cyrus.psych.uiuc.edu)
Beckman Institute, University of Illinois
405 N. Matthews Ave., Urbana, IL 61801, USA

Abstract

Many studies have examined the effects of co-activation of similar words ("neighbors") during processing, with some reporting facilitative effects and others reporting inhibitory effects. Attractor dynamics has provided a promising integrated account in which distant semantic neighbors (moderately similar words) tend to facilitate processing and near semantic neighbors (highly similar words) tend to inhibit processing. This framework was extended to phonological neighbor effects on the accuracy of word production. For aphasic patients (N=62) and speeded young controls (N=32), picture naming was more accurate for words with many distant phonological neighbors (words with matching onsets) and less accurate for words with a near phonological neighbor (homophones). In addition, the sizes of the facilitative and inhibitory effects were correlated, suggesting that the mechanisms responsible for both effects are functionally integrated. These results extend an attractor dynamics framework that predicts facilitative effects of distant neighbors and inhibitory effects of near neighbors.

Keywords: phonological neighbors; cohort density; homophones; neighborhood density; word production; attractor dynamics.

Introduction

Theories of language processing agree that similar words are co-activated during processing. Such co-activation provides a simple account of classic priming effects: processing *cat* partially activates *dog* (due to semantic similarity) and *can* (due to form similarity), facilitating responses to those words (Marslen-Wilson & Zwitserlood, 1989; Meyer & Schvaneveldt, 1971; Zwitserlood, 1996). Co-activation is also consistent with findings from studies using the visual-world paradigm: when instructed to click on a picture of a *cat*, participants are more likely to fixate images of a *dog* or a *can* (Allopenna, Magnuson, & Tanenhaus, 1998; Huettig & Altmann, 2005; Magnuson, Tanenhaus, Aslin, & Dahan, 2003; Mirman & Magnuson, 2009; Yee & Sedivy, 2006). Co-activation of similar words has also been used to account for global similarity effects: the number of similar words that are likely to be co-activated given a particular similarity metric, called "neighborhood density". In this report, we examine the effects of two kinds of phonological neighbors on word production in aphasic patients and speeded young controls.

The effects of neighborhood density on word processing are complex and poorly understood. Neighbors defined by form similarity (spelling or sound) have been found to facilitate printed word recognition (e.g., Sears, Hino, & Lupker, 1995) and spoken word production (e.g. Vitevitch, 2002). However, phonologically similar neighbors consistently produce inhibitory effects in many tasks involving spoken word recognition (e.g., Luce & Pisoni, 1998; Magnuson, Dixon, Tanenhaus, & Aslin, 2007). Neighbors defined by semantic similarity can also exert effects in both directions. Near neighbors (concepts with very similar meanings) inhibit word recognition and distant neighbors (concepts with moderately similar meanings) facilitate visual word recognition (Mirman & Magnuson, 2008). Mirman and Magnuson suggested that this contrast between the impact of near and distant neighbors on word processing may be a general property of word processing. For example, although orthographic neighbors (*salt* - *halt*) generally facilitate visual word recognition, transposed-letter neighbors (*salt* - *slat*) exert inhibitory effects (Andrews, 1996).

The attractor dynamics framework for cognition represents each concept as a stable state ("attractor basin" or simply "attractor") in a high-dimensional space of possible mental states (for an accessible introduction see Spivey, 2007). Word processing is a matter of traversing this space in order to reach the correct attractor. When the system has reached a stable state, it is deemed to have "settled" and the accuracy of the system's final state can be compared relative to the target attractor. Neighbors are other attractors and distance between attractors is determined by similarity. The critical insight from attractor dynamics is that different similarity relations between neighbors can exert different effects on the settling process (Mirman & Magnuson, 2008). Distant neighbors create a broader attractor basin, which facilitates settling to the correct attractor. In contrast, near neighbors are too few to substantially change the overall size of the attractor basin, but because of their high similarity (i.e., proximity) to the target, they function as conflicting subbasins, which slows the completion of the settling process.

An alternative to the attractor dynamics account would be to simply stipulate that neighbor effects are different in

different contexts or tasks. For example, Vitevitch and Luce (1998; 1999; see also Luce & Large, 2001) proposed that, in speech perception, sub-lexical neighbor effects are facilitative and lexical neighbor effects are inhibitory. However, there are three arguments against this view as a general account of neighborhood effects. First, the empirical data have been challenged (Lipinski & Gupta, 2005). Second, semantic neighbor effects appear to emerge at a single level of processing (i.e., semantics), thus, assigning different effects to different levels cannot account for the facilitative effects of distant semantic neighbors and the inhibitory effects of near semantic neighbors (Mirman & Magnuson, 2008). Third, it is unparsimonious to propose that neighbor interactions have fundamentally different properties at different levels of processing.

Attractor dynamics provide a parsimonious, integrated account in which neighbors can have different, context-dependent effects. However, the existing data have only examined the key attractor dynamics prediction in the domain of semantic neighborhoods. The present studies examine these same predictions in the domain of phonological neighbor effects on word production.

As noted above, previous studies have found facilitative effects of phonological neighbors on word production. Vitevitch (2002) found that healthy young controls produced more errors in an error-elicitation paradigm and were slower to name pictures for words with few phonological neighbors compared to words with many phonological neighbors. Similarly, aphasic patients produce more errors when naming pictures with low phonological neighborhood density names (Gordon, 2002; Kittredge, Dell, Verkuilen, & Schwartz, 2008).

Given the facilitative effect of phonological neighbors in picture naming tasks, one might expect that greater phonological similarity would strengthen this effect. In the extreme case, words with different meanings but identical phonological forms, that is, homophones (e.g. *can* [container] vs. *can* [able]) might be particularly easy. After all, the naming target's homophone is maximally phonologically similar to target's phonology. However, if both meanings are activated during an attempt to retrieve the name of one meaning of a homophone, those meanings may compete, consequently producing slower responses and higher error rates. Thus, there is reason to expect the opposite result. Indeed, this is the critical prediction from the attractor dynamics account of neighborhood effects.

There is an extensive experimental literature investigating homophony in word production, most of which is concerned with whether word frequency effects on picture naming arise from syntactic-semantic representations or phonological form representations (e.g. Caramazza, Costa, Miozzo, & Bi, 2001; Jescheniak & Levelt, 1994). Although there is no consensus among these studies, it is likely that production latencies for a homophone are influenced by the frequency of both its meaning and its form. More relevant to our analysis are findings that both meanings of a homophone are activated during word production. For example, priming the non-pictured homophone meaning affects response latency and accuracy in picture naming tasks (Cutting & Ferreira 1999; Ferreira & Griffin, 2003). Moreover, picture naming studies with aphasic patients have shown that training on one homophone meaning generalizes to the other meaning (Biedermann, Blanken, & Nickels, 2002; Biedermann & Nickels, 2008a; Biedermann & Nickels, 2008b).

These studies suggest that homophone production involves some degree of interaction between the target and its homophone mate. Given this, if homophones are viewed as very near phonological neighbors, the attractor dynamic approach of Mirman and Magnuson (2008) predicts that having a homophone should be associated with some kind of cost. Alternately, if the extreme similarity of the homophone just exaggerates the positive effect of having a similar neighbor, then the expectation is for a benefit. These conflicting predictions were tested by examining the accuracy of picture naming in aphasic patients and in speeded young controls.

Experiment

Methods

Participants. There were two sets of participants: aphasic patients and speeded young controls. The patients were 62 unselected aphasic patients recruited from the MRRI Cognitive Rehabilitation Research Registry (Schwartz, Brecher, White, & Klein, 2005) on the basis of chronic aphasia secondary to left cerebral vascular accident. They had a mean age of 58 (range 26–78), mean years of education of 14 (10–21), and most (over 90%) were at least 6 months post-onset. The patients were all premorbidly right-handed, had English as the primary language, adequate vision and hearing, and uniliateral left hemisphere damage (not restricted to subcortical areas). These patients included all aphasia subtypes and covered a wide range of performance (2%-97% correct naming). The young controls were 32 healthy college students with no known history of neurological, visual, or auditory impairments, who were recruited from the University of Illinois participant pool.

Materials. The 175-item Philadelphia Naming Test (PNT; Roach et al., 1996) was used to measure word production in picture naming. The black and white pictures represent objects from varied semantic categories and have high familiarity, name agreement, and image quality. Names range in length from 1 to 4 syllables and in frequency (normalized to occurrences per 1 million word tokens) from 1 to 100.

Our concern is with the effects of "near" and "distant" phonological neighbors on picture naming in order to test

the general attractor dynamics prediction that the effect of neighbors will depend on their impact on the attractor landscape. Distant phonological neighbors were defined as words that share onsets with the target word. These words are described as "cohorts" because they form the cohort of partially activated words during spoken word recognition (e.g., Allopenna et al., 1998; Magnuson et al., 2007; Marslen-Wilson & Zwitserlood, 1989). There are many possible phonological neighborhood measures, which are all strongly correlated with one another. The cohort density measure (the summed log frequency of the target word and all of its cohorts) was chosen because word onsets are particularly important for spoken word processing.

Lexical variables (phonological neighborhood, word frequency, etc.) were assessed using the American National Corpus (Ide & Suderman, 2004), a large-scale, representative corpus of American English containing over 3.2 million spoken word tokens. The words in the PNT were divided into "few" and "many" neighbor conditions based on the median cohort density (31.5) and a few words were eliminated to ensure that the conditions had an equal number of words and were matched in word frequency and length (the resulting conditions were composed of 85 words each). Table 1 shows that the two conditions were matched in word frequency and length and strongly different in cohort density as well as differing on other phonological neighborhood measures. For the purpose of this experiment, it was not necessary to isolate a particular measure of phonological neighborhood; rather, it was sufficient that words in the two conditions strongly differed in their number of phonologically similar words.

Table 1. Mean (standard deviations in parentheses) properties of stimuli for cohort density manipulation.

	Few neighbors	Many neighbors	t	p<
Phonological neighborhood measures				
Cohort Density	14.9 (8.7)	73.2 (36.0)	14.5	0.0001
Neighborhood Density	10.2 (8.8)	14.1 (11.3)	2.5	0.05
Number of Neighbors	11.8 (12.8)	16.2 (15.2)	2.0	0.05
Posit. Prob.	.211 (0.1)	.263 (0.1)	3.2	0.01
Transit. Prob.	.017 (0.02)	.023 (0.02)	2.3	0.05
Control Variables				
Num. Words	85	85	-	-
Log Frequency	1.07 (0.7)	1.16 (0.7)	0.94	0.35
Num. Letters	5.51 (1.9)	5.11 (1.9)	1.4	0.17
Num. Phonemes	4.33 (1.7)	4.35 (1.5)	0.09	0.93

Near phonological neighbors were defined as words with identical phonological forms and unrelated meanings, that is, homophones. The 175 words in the PNT include 14 homophones for which the pictured meaning is the dominant meaning (meaning dominance was assessed based on proportion of associated words in the USF free association norms (Nelson, McEvoy, & Schreiber, 2004): *M*=73.6%, *SD*=10.5, Range=50.4-86.7). Number of meanings (homophony) was assessed based on the number of distinct entries in the online Wordsmyth dictionary (http://new.wordsmyth.net/). For each of these homophones a control (unambiguous) word was selected from the PNT that was matched to the homophone on word frequency, length, and phonological neighborhood variables (see Table 2).

Table 2. Mean (standard deviations in parentheses) properties of stimuli for homophony manipulation.

	Homophones	Control Words	t	p<
Num. Meanings	2.21 (0.58)	1.0 (0)	-	-
Control Variables				
Num. Words	14	14	-	-
Cohort Density	50.6 (41.1)	46.7 (36.0)	0.87	0.40
Neighborhood Density	26.1 (14.7)	27.2 (15.2)	0.30	0.77
Number of Neighbors	22.4 (9.8)	22.00 (9.1)	0.22	0.83
Posit. Prob.	.202 (0.04)	.195 (0.06)	0.55	0.59
Transit. Prob.	.014 (0.01)	.013 (0.01)	0.61	0.55
Log Frequency	1.47 (0.80)	1.40 (0.50)	0.54	0.60
Num. Letters	4.07 (0.92)	4.00 (0.78)	1.00	0.34
Num. Phonemes	3.29 (0.61)	3.29 (0.61)	0.0	1.0

Procedure. The patients were tested using the standard PNT procedure (http://www.ncrrn.org/assessment/pnt; Roach et al., 1996; see also Dell et al., 1997; Schwartz et al., 2006): each picture was presented one at a time and the first complete (i.e. non-fragment) response produced within 20 s was scored. The young controls were tested using the tempo picture naming procedure (Hodgson & Lambon Ralph, 2008). This task provides a valuable source of converging data for comparison with the patient data because it has been shown to induce some characteristic aspects of aphasic picture naming errors. In the tempo picture naming task, participants heard a series of beeps set to a tempo (500 ms). On the fourth beep they were also presented with a picture (one of the PNT items), which they were to name and to time their response to coincide with the fifth beep.

Results

Cohort Density. The left panel of Figure 1 shows that picture naming accuracy was lower for low cohort density words than for high cohort density words (Patients: 66.7% vs. 70.4%, $t(61)=5.5$, $p<0.00001$; Speeded controls: 79.7% vs. 81.7%, $t(31)=2.37$, $p<0.05$). Patients also produced more nonword errors for low cohort density words than high cohort density words (8.24% vs. 6.62%, $t(61)=3.23$, $p<0.01$). Speeded controls produced very few nonword

errors (M=0.68%, SD=0.89%) and the numerical trend in the same direction as the patients (0.77% vs. 0.63%) was not significant ($t(31)$=0.49, p>0.6). There were no significant effects on any other error type. The cohort density finding is consistent with previous findings that words with many phonologically similar words are easier to produce (Gordon, 2002; Kittredge et al., 2008; Vitevitch, 2002).

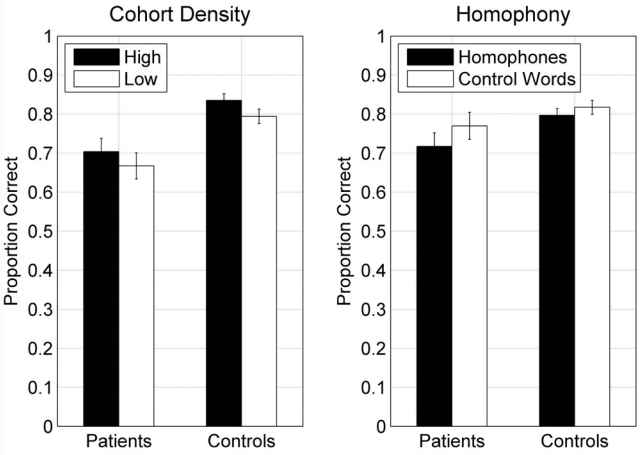


Figure 1. Picture naming accuracy for high and low cohort density words (left panel) and for homophones and control words (right panel). Error bars reflect 1SE.

Homophony. The right panel of Figure 1 shows that participants were more accurate for the control words than for the homophones (Patients: 77.0% vs 71.7%, $t(61)$=3.45, p<0.001; Speeded controls: 83.5% vs. 79.4%, $t(31)$=5.43, p<0.00001). This finding is consistent with previous results that indicate slowed processing due to competition between different meanings of homophones (e.g., Shatzman & Schiller, 2004; see also Ferreira & Griffin, 2003). The increased errors for homophones did not aggregate to a specific error type (i.e., no reliable differences for any error type).

Relation between effect sizes. We tested the correlation between cohort density and homophony effect sizes across participants to examine whether there is a possible relationship between them. Figure 2 shows each participant's homophony effect size (homophones – control) plotted against that participant's cohort density effect size (high – low). The effect sizes were reliably correlated for patients (r = -0.25, p<0.05) and for speeded controls (r = -0.76, p<0.0001).

One possible explanation for this effect size correlation is that there is simply an effect of overall accuracy. That is, participants who make more errors show bigger differences between any conditions. To test this hypothesis, we examined correlations between overall accuracy for the critical conditions and the effect size. For patients, neither correlation approached significance (homophony: r = 0.0032, p > 0.98; cohort density: r = 0.0937, p > 0.46). The same was true for controls (homophony: r = 0.1124, p > 0.54; cohort density: r = -0.2616, p > 0.14). Since it is not due to overall accuracy, the correlation between cohort density and homophony effect sizes suggests that the mechanisms involved in producing the benefit of similar-sounding words (cohort density effect) are closely tied to those involved in producing the cost of identical-sounding words (homophony effect).

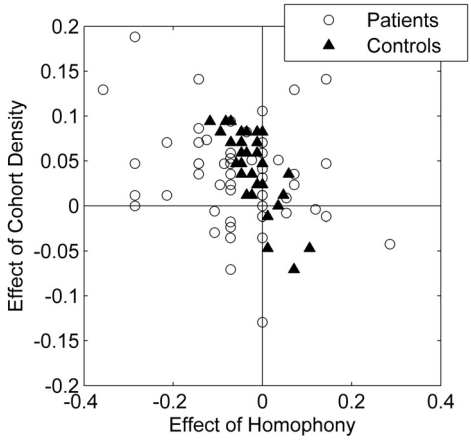


Figure 2. Relationship between homophony and cohort density effect sizes. Open circles correspond to patients, filled triangles correspond to speeded controls.

General Discussion

We examined the effects of phonological neighbors on picture naming in aphasic patients and speeded young controls. Two kinds of phonological neighbors were considered: similar-sounding words defined as words with matching onsets (i.e., cohorts) and identical-sounding words (i.e., homophones). These different phonological neighbor types capture the important distinction between distant and near neighbors. Mirman and Magnuson (2008) found that distant semantic neighbors facilitated word recognition and near semantic neighbors inhibited word recognition. Andrews (1996) found a similar contrast between the effects of (distant) orthographic neighbors and (near) transposed-letter neighbors on visual word recognition. Based on these results, we predicted facilitative effects of phonological neighbors and inhibitory effects of homophony.

The results were consistent with these predictions: both participant groups exhibited a facilitative effect of cohort density and an inhibitory effect of homophony. In addition, the effect sizes were correlated across participants; that is, participants who showed larger cohort density advantage effects also showed larger homophony disadvantage effects. This suggests that the mechanism or mechanisms that produce these effects are functionally integrated.

To account for the contrasting effects of near and distant semantic neighbors, Mirman and Magnuson (2008) proposed an account based on attractor dynamics. On this view, distant neighbors create a broader attractor basin,

which facilitates settling to the correct attractor. In contrast, near neighbors are too few to substantially change the overall size of the attractor basin, but because of their high similarity (i.e., proximity) to the target, they function as conflicting subbasins, which slows the completion of the settling process. These distinctions are shown schematically in Figure 3. Mirman and Magnuson confirmed this account using simulations of a computational model.

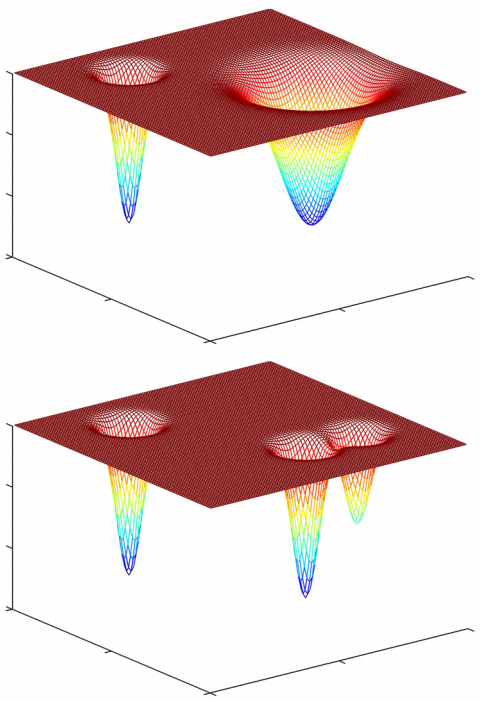


Figure 3. Top: Schematic diagram of narrow and broad attractor basins resulting from few and many distant neighbors, respectively. Bottom: Schematic diagram of a single attractor basin and an attractor with a subbasin formed by a near neighbor.

To extend this framework to word production it is helpful to consider picture naming as a process of settling to an attractor in a multidimensional space that combines semantic and phonological dimensions. For a given target word, cohort neighbors and homophone neighbors are equally semantically unrelated to the target (the cohort pair *can* – *cat* and the homophone pair *can* [container] – *can* [able] are equally semantically unrelated). On the phonological dimensions, the homophone neighbors have substantially more similarity to the target word than cohort neighbors do (i.e., complete phonological overlap vs. shared onsets). Therefore, a large number of cohort neighbors can increase the gradient and facilitate settling to the correct attractor. In contrast, a single homophone neighbor will not have a substantial impact on the gradient, but can form a competing subbasin, which can delay the settling process.

If the settling process is disrupted by damage or a time constraint, the system may fail to settle completely (no response) or may settle to an incorrect attractor (error). Since settling is facilitated by distant (cohort) neighbors and inhibited by near (homophone) neighbors, this account captures the observed pattern of facilitative effects of cohort density and inhibitory effects of homophony. The correlation between effect sizes could reflect the average sharpness of attractor basins in the landscape. In a landscape with relatively sharp attractor basins, distant neighbor attractors would have a relatively small impact on slope steepness and near neighbors would be less likely to act as a competing subbasin. Attractor dynamic models generally develop sharper attractors over the course of learning, so this individual difference variable could reflect language skill. Further research is required to test this hypothesis or other possible explanations of the correlation between effect sizes.

In sum, the present results demonstrate contrasting effects of near and distant phonological neighbors on picture naming and provide a new perspective on the mechanisms involved in word production. Furthermore, they contribute to the creation of a unified theory of neighborhood effects in lexical processing.

Acknowledgements

This research was supported by the Moss Rehabilitation Research Institute and National Institutes of Health grants DC000191 and HD44458. We thank Myrna Schwartz for her thoughtful suggestions.

References

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory & Language, 38*(4), 419-439.

Andrews, S. (1996). Lexical retrieval and selection processes: Effects of transposed-letter confusability. *Journal of Memory and Language, 35*(6), 775-800.

Biedermann, B., Blanken, G., & Nickels, L. (2002). The representation of homophones: Evidence from remediation. *Aphasiology, 16*(10-11), 1115-1136.

Biedermann, B., & Nickels, L. (2008a). The representation of homophones: More evidence from the remediation of anomia. *Cortex, 44*(3), 276-293.

Biedermann, B., & Nickels, L. (2008b). Homographic and heterographic homophones in speech production: Does orthography matter? *Cortex, 44*(6), 683-697.

Caramazza, A., Costa, A., Miozzo, M., & Bi, Y. (2001). The specific-word frequency effect: Implications for the representation of homophones in speech production. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*(6), 1430-1450.

Cutting, J. C., & Ferreira, V. S. (1999). Semantic and phonological information flow in the production lexicon.

Journal of Experimental Psychology: Learning, Memory, and Cognition, 25(2), 318-344.

Dell, G. S., Schwartz, M.F., Martin, N., Saffran, E.M., & Gagnon, D.A. (1997). Lexcial access in aphasic and nonaphasic speakers *Psychological Review, 104*(4), 801-838.

Ferreira, V. S., & Griffin, Z. M. (2003). Phonological influences on lexical (mis)selection. *Psychological Science, 14*(1), 86-90.

Gordon, J. K. (2002). Phonological neighborhood effects in aphasic speech errors: Spontaneous and structured contexts. *Brain and Language, 82*(2), 113-145.

Hodgson, C., & Lambon Ralph, M. A. (2008). Mimicking aphasic semantic errors in normal speech production: Evidence from a novel experimental paradigm. *Brain and Language, 104*, 89-101.

Huettig, F., & Altmann, G. T. M. (2005). Word meaning and the control of eye fixation: Semantic competitor effects and the visual world paradigm. *Cognition, 96*(1), B23-B32.

Ide, N., & Suderman, K. (2004). The american national corpus first release. In *Language resources and evaluation conference (lrec)* (pp. 1681-1684). Lisbon.

Jescheniak, J. D., & Levelt, W. J. M. (1994). Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 824-43.

Kittredge, A.K., Dell, G.S., Verkuilen, J., & Schwartz, M.F. (2008). Where is the effect of frequency in word production? Insights from aphasic picture-naming errors. *Cognitive Neuropsychology*, *25*(4), 463-492.

Lipinski, J., & Gupta, P. (2005). Does neighborhood density influence repetition latency for nonwords? Separating the effects of density and duration. *Journal of Memory and Language, 52*(2), 171-192.

Luce, P. A., & Large, N. R. (2001). Phonotactics, density, and entropy in spoken word recognition. *Language and Cognitive Processes, 16*(5/6), 565-581.

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing, 19*, 1-36.

Magnuson, J. S., Dixon, J. A., Tanenhaus, M. K., & Aslin, R. N. (2007). The dynamics of lexical competition during spoken word recognition. *Cognitive Science, 31*, 1-24.

Magnuson, J. S., Tanenhaus, M. K., Aslin, R. N., & Dahan, D. (2003). The time course of spoken word learning and recognition: Studies with artificial lexicons. *Journal of Experimental Psychology: General, 132*(2), 202-227.

Marslen-Wilson, W., & Zwitserlood, P. (1989). Accessing spoken words: The importance of word onsets. *Journal of Experimental Psychology: Human Perception and Performance, 15*(3), 576-585.

Meyer, D.E., & Schvaneveldt, R.W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology, 90,* 227-234.

Mirman, D., & Magnuson, J. S. (2008). Attractor dynamics and semantic neighborhood density: Processing is slowed by near neighbors and speeded by distant neighbors. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*(1), 65-79.

Mirman, D., & Magnuson, J. S. (2009). Dynamics of activation of semantically similar concepts during spoken word recognition. *Memory & Cognition, 37*(7), 1026-1039.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments & Computers, 36*(3), 402-407.

Roach, A., Schwartz, M. F., Martin, N., Grewal, R. S., & Brecher, A. (1996). The philadelphia naming test: Scoring and rationale. *Clinical Aphasiology, 24*, 121-133.

Schwartz, M. F., Dell, G. S., Martin, N., Gahl, S., & Sobel, P. (2006). A case-series test of the interactive two-step model of lexical access: Evidence from picture naming. *Journal of Memory and Language, 54*(2), 228-264.

Schwartz, M. F., Brecher, A. R., Whyte, J., & Klein, M. G. (2005). A patient registry for cognitive rehabilitation research: A strategy for balancing patients' privacy rights with researchers' need for access. *Archives of Physical Medicine and Rehabilitation, 86*(9), 1807-1814.

Sears, C. R., Hino, Y., & Lupker, S. J. (1995). Neighborhood size and neighborhood frequency effects in word recognition. *Journal of Experimental Psychology: Human Perception and Performance, 21*(4), 876-900.

Shatzman, K. B., & Schiller, N. O. (2004). The word frequency effect in picture naming: Contrasting two hypotheses using homonym pictures. *Brain and Language, 90*(1-3), 160-169.

Spivey, M. (2007). *The continuity of mind*. New York, NY, US: Oxford University Press.

Vitevitch, M. S. (2002). The influence of phonological similarity neighborhoods on speech production. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*(4), 735-747.

Vitevitch, M. S., & Luce, P. A. (1998). When words compete: Levels of processing in perception of spoken words. *Psychological Science, 9*(4), 325-329.

Vitevitch, M. S., & Luce, P. A. (1999). Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory & Language, 40*(3), 374-408.

Yee, E., & Sedivy, J. C. (2006). Eye movements to pictures reveal transient semantic activation during spoken word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*(1), 1-14.

Zwitserlood, P. (1996). Form priming. *Language and Cognitive Processes, 11*(6), 589-596.

How Does Anxiety Influence Analogical Mapping?

V. Feldman (vfeldman@nbu.bg), P. Hristova (phristova@cogs.nbu.bg), B. Kokinov (bkokinov@nbu.bg)
Central and East European Center for Cognitive Science, Department of Cognitive Science and Psychology,
New Bulgarian University, 21 Montevideo Street
Sofia 1618, Bulgaria

Abstract

This paper presents an experimental study of the influence that the anxiety state may have on analogical mapping. Contrary to the well-known study of Tohill & Holyoak (2000), where the anxiety state impeded the analogical mapping, in this study participants in the anxiety state were significantly more inclined to produce a relational choice which is structurally consistent with the target, even though this alternative was more superficially dissimilar to the target. This result was obtained in a match-to-sample paradigm. The implications for the theory of how anxiety influences analogy-making are discussed and it is argued in favor of a more detailed and specific approach to studying the influence of anxiety on each component mechanism of analogy-making.

Introduction

Imagine that you are in a stressful situation and you feel anxious. Will that make you more or less successful in making good analogies? Some researchers believe that this emotional state will impede the analogy-making process (Tohill & Holyoak, 2000), while others (Richert, Whitehouse, Stewart, 2005) argue that you will make more or better analogies and that is why some religious rituals are deliberately designed to increase your anxiety. This controversy has motivated our study.

The interplay between analogy and emotions has been studied from two opposite perspectives.

Thagard and Shelley (2001) have argued that analogy may influence emotions, since people may use analogies to convey emotions to others like in the famous "Saddam is like Hitler" example (Spellman & Holyoak, 1992). This theoretical account was empirically supported in a recent study using simple proportional analogies (Bliznashki & Kokinov, 2009) which demonstrated that the negative or positive attitude towards an item in one domain can be transferred to the corresponding item in the other domain via the analogy and that this transfer is bidirectional.

Several researchers explored the influence of emotions on the analogy-making process itself. Thus a series of studies was devoted to the influence of anxiety on analogy (Leon and Revelle, 1985; Keinan, 1987; Tohill and Holyoak, 2000). Why anxiety? The specific line of reasoning was that since it is well known that anxiety influences several cognitive processes, including working memory, one should expect also an influence on analogy. Thus Tohill and Holyoak (2000) provided evidence that state anxiety impedes the relational mapping and anxious participants prefer a more superficial attributive mapping. In their study anxiety was induced prior to the task by a serial subtraction task with a negative feedback. Participants were instructed to count aloud from 1000 backwards with a decrement of 13. One experimenter corrected participants' mistakes and another – urged participants to count faster. Moreover, participants in the anxiety group were informed that they would have to repeat this task at the end of the experiment, i.e. after the analogy-making task. The influence of anxiety on analogy-making was tested with a cross-mapping task, where participants were asked to indicate which object, presented on one of the pictures "goes with" the object, pointed to by the experimenter. The trick was that the object pointed to in the first picture could "go with" two different objects in the second picture for two different reasons, i.e. with the object which is similar in its physical appearance to the pointed object or with the object that participates in similar relations as the pointed one. Based on Eysenk's working memory restriction theory (Eysenk and Calvo, 1992) it was assumed that anxiety restricts working memory capacity which in turn impedes higher-order relational mapping needed for finding the relational mappings in the cross-mapping task used in this particular study. Correspondingly, anxious participants[1] indicated fewer relational mappings than non-anxious participants (Experiment 1) even in the presence of explicit instruction to find them (Experiment 2) (Tohill and Holyoak, 2000).

It was also shown that state anxiety impedes the range of generated analogies to a given base problem (Feldman and Kokinov, 2009). Anxious participants generated a significantly smaller amount of drastically different analogies, i.e. most of their analogies belonged to one domain, while non-anxious participants were more flexible and generated analogies belonging to two or three different domains. In addition, non-anxious participants produced analogies with remote domains, while anxious ones produced mainly close analogies. At the same time no difference was found between the quality of mapping and convincingness of the analogies produced. So, no direct evidence was produced neither in favor, nor against the hypothesis that anxiety impedes analogical mapping. It was only demonstrated that anxiety impedes analogical retrieval (in an analogy generation task).

On the other hand, it has been shown that people in negative mood are more likely to choose the relational match rather than the attribute matches compared to people in positive mood in a simple matching-to-sample task (Hristova, 2009). In this study both the relational and the attribute mappings were possible, but curiously people in negative mood prefer the former ones, i.e. they choose the

[1] state anxiety is used here, not trait anxiety.

relationally similar target as being more similar to the base stimulus. Hristova (2009) argues that the triples of figures used as stimuli in this experiment presuppose that the relational mappings were harder than the attribute ones, since they require an extra effort for encoding of relations, which were not explicitly drawn between figures. Hence participants in negative mood invested more effort while doing the task than participants in positive mood, consistent with the cognitive tuning hypothesis (Schwarz, 2002) that has inspired this experimental work. Since the core of analogy-making is exactly the mapping between relations, rather than attributes (Gentner, 1983) the straightforward inference from this work is that negative, rather than positive mood may enhance analogy-making by facilitating the encoding of relations.

In conclusion, it seems that there is controversial evidence for the role of emotions on analogical mapping. Tohill and Holyoak (2000) have found that state anxiety may change analogical mapping from relational toward attribute based one, while Hristova (2009) has found that negative mood, which can be considered as similar in valence to the state of anxiety (i.e. anxiety is a kind of negative emotional state) facilitates relational mapping compared to positive mood.

It could be that anxiety exerts a completely different influence on analogical mapping than negative mood: an interesting hypothesis that insists on fine grade distinction between the negative emotions themselves and therefore, between the cognitive mechanisms that these emotions may change. This hypothesis however, cannot fully explain the variety of results obtained in the field of analogy-making, since two experiments that manipulate state anxiety report different results with respect to analogical mapping: Tohill and Holyoak, (2000) demonstrated less relational mappings due to anxiety, while Feldman and Kokinov (2009) did not report any effect of anxiety on analogical mapping.

The present research aims to further explore the influence of anxiety on analogical mapping by exploiting the anxiety-inducing procedure used by Feldman and Kokinov (2009) and the analogical mapping task used by Hristova (2009). If anxiety impedes relational choices in this task rather than facilitate them, as shown under negative mood (Hristova, 2009), then it would be relatively safe to conclude that the diverse negative emotions (to be more specific, anxiety compared to negative mood) exert different effects on analogy-making. If the opposite trend is observed then the picture of influence of anxiety on analogical mapping is more complicated.

Experiment

Method

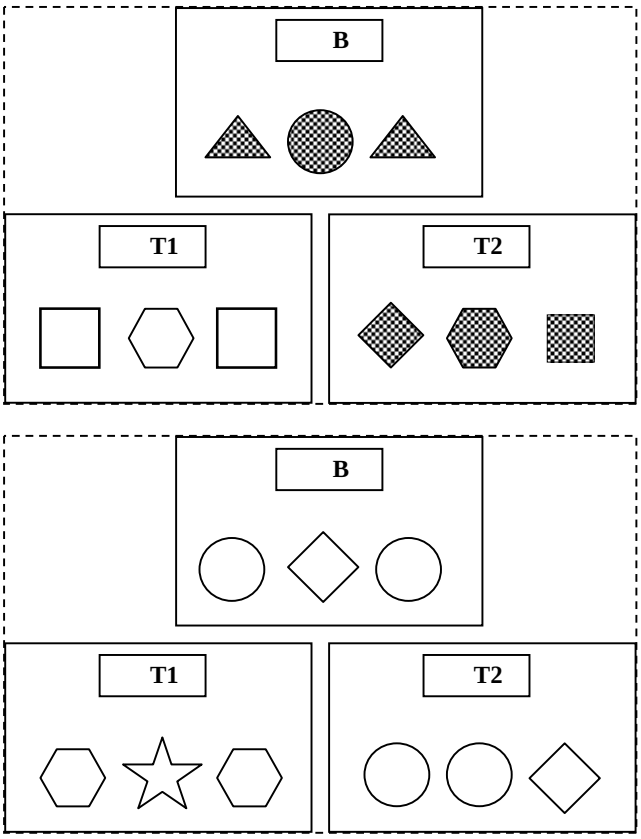
The main idea of this experiment is to test whether an induced state of anxiety will change the type of relational processing performed by the participants and in particular whether the proportion of relational choices will be higher or lower than in a non-anxiety state.

Design

This experiment has a between-subject design with one independent variable – the state of anxiety (an anxiety and a non-anxiety group), and one dependent variable the proportion of relational choices made. Two other variables were used for control purposes only – the state of anxiety as measured by a self-report on a scale and the response times.

Stimuli

22 stimuli were used in this experiment. Each of them was a match-to-sample-triple consisting of a base item B and two target items T1 and T2. The question that the participants had to answer was "whether T1 or T2 is more similar to B". The stimuli were prepared in such a way that one of the targets was sharing the same objects or the same color of the objects as the base, i.e. was superficially similar to the base, while the other one shared some spatial or transformational relations but consisted of different objects, i.e. was structurally similar. Both choices make perfect sense. Three groups of stimuli were used in the experiment and representatives of each group are presented in Figure 1. There is a forth group of stimuli which are only partially analogous (i.e. none of the two is a good match) since only some of the relations/attributes are shared (Figure 2). These stimuli were used by Hristova (2009) and are variations of the stimuli used by Medin, Goldstone, and Gentner (1990) and Sloutsky and Yarlas (submited). We have used them for replication purposes.



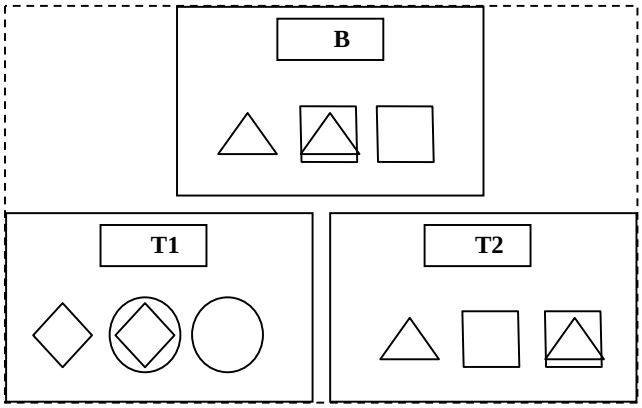


Figure 1. Three examples of items from the Match-to-Sample task, one example from each category of stimuli. In all three cases T1 is the relational choice, while T2 is the superficially similar one. Of course, in the experiment the order of T1 and T2 presented as relational/superficial choice has been contra balanced.

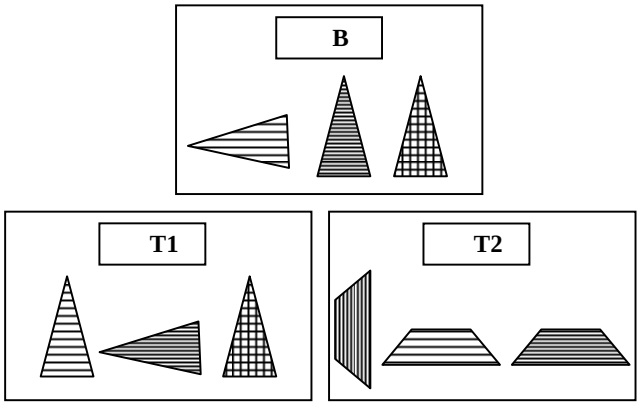


Figure 2. An additional type of examples used in the Match-to-Sample task for replication purposes. Neither T1, nor T2 makes perfect analogy to B, but T2 keeps the spatial relationships, while T1 keeps the relations between textures.

Procedure

The participants were tested individually by an experienced experimenter in a sound proof booth on a personal computer running e-Prime automated script.

Participants were enrolled in a matching-to-sample experiment for about 5 minutes. Their task was to judge whether "T1" or "T2" are more similar to the standard "B" by pushing the respective button on a BBOX: the left button "T1" and the right button for "T2". When participants gave their answer the next stimulus appeared on the screen. The presentation order of the stimuli was randomized across participants. A fixation cross was presented for 50 ms before each trial.

In the Experimental group the state anxiety was induced by a "public speech" procedure which was used successfully to induce state anxiety in a number of other studies (Graeff1, Parente, Del-Ben, Guimarães, 2003; Pertaub, Slater & Barker, 2002; Feldman & Kokinov, 2009). The participants in the Anxiety group when invited were instructed that at some point they will be interrupted and will be asked to make a presentation on a topic that they will not know in advance. The task will be to argue in favor of a specific claim. They will have to talk spontaneously and without interruption for 5 minutes. Their presentation will be video recorded and then later on their communication skills will be evaluated. In that moment the experimenter installed a camera in front of the participant, but no recording was initiated. They were asked meanwhile to participate in another experiment and they were given the match-to-sample task described above. The participants were never asked to make the public speech and were never recorded, however, they were constantly expecting that this was going to happen. At the end the participants were debriefed about how they were feeling and they also rated on a 5 point scale how nervous they were during the experiment.

Participants

38 participants (15 male and 23 female) took part in the experiment. All of them were students at the New Bulgarian University some in psychology and some in other programs. Their age varied from 17 to 37 years and the average was 22.95. The participants were randomly assigned in equal numbers to the two conditions, maintaining equal ratios between female and male participants in each group.

Results

First of all, our manipulation of anxiety seems to be successful. The two groups differed significantly on their self-evaluation of how nervous they had felt during the experiment on a 5 point scale ($t(36)=4.624$, $p<0.001$, $d=1.50$) – the Control group ($M=0.79$, $SD=1.134$) and the Anxiety group ($M=2.32$, $SD=0.885$).

The mean proportion of the relational choices was higher in the anxiety group (35%) than in the control group (24%) and this difference turned out to be significant tested with a t-test when the data were aggregated by item – $t(42)=5.695$, $p<0.001$, $d=0.31$ (Figure 3). At the same time importantly, RT did not differ significantly between the two experimental conditions: $t(42)=0.397$, $p=0.693$, (Figure 4). Thus, the influence of anxiety cannot be attributed to spending more time and more careful inspection of the task in the anxiety group.

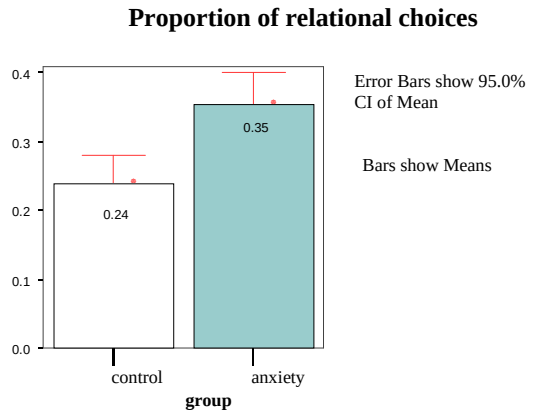


Figure 3. Mean proportion of relational choices per condition.

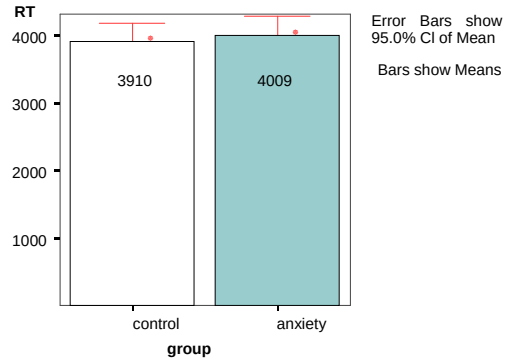


Figure 4. Mean RT per condition.

The same trend was observed for the six stimuli of Hristova (2009), included in the item pool of this experiment: the anxiety group made significantly more relational choices (38%) than the control group (25%) ($t(10)=0.424$, $p=0.016$, $d=1.78$), while the difference in the mean RT was again not significant ($t(10)=0.782$, $p=0.452$). This result is comparable to the one obtained by Hristova (2009) with the same task and stimuli, where people in negative mood also made significantly more relational choices than people in positive mood.

Discussion

The goal of this research was to clarify the role of anxiety for analogical mapping. Two conflicting findings were discussed at the beginning: anxiety may impede relational mapping (Tohill and Holyoak, 2000) or anxiety, as a kind of negative emotional state, may facilitate relational mapping (Hristova, 2009). The present research supports the latter prediction, i.e. anxiety facilitated relational mappings.

The question now is why we have obtained results opposite to the ones by Tohill and Holyoak (2000)? There are two important differences between the two studies: the procedure of anxiety inducement and the tasks of the participants. Each of them could potentially cause the difference.

With respect to the anxiety inducement procedure there are a number of differences. It could be that one of them is inducing stronger anxiety than the other. We cannot say this with certainty, since we have not used the same instrument for measuring the anxiety state of the participants, however, there are reasons to believe that the current procedure induces stronger anxiety since making a public speech and being recorded and then your communication skills being analyzed seems more stressful than counting backwards at high speed and being corrected. In addition, in the current procedure the participants were warned that they can be interrupted any time and asked to make the public speech, while the participants in the Tohill & Holyoak (2000) study knew that they will be counting again only after the analogy task is over. Of course, these are only speculations, it is also possible that the current procedure has produced much less anxiety than the Tohill & Holyoak (2000) study and the results are due to the classical Yerkes-Dodson law (1908) that describes the inverted U shaped relationship between arousal and performance: maybe we have found the optimal level of arousal for the matching-to-sample task used in our experiment, while Tohill and Holoak (2000) did not. In other words anxiety may both increase and decrease relational mappings depending on the degree of arousal. This explanation is unconvincing since the very same procedure has been applied by Feldman & Kokinov (2009) and it has significantly reduced the number of *different* analogies generated and their *scope*. Also additional analysis of the data shows that there is a trend: the higher the self-reported anxiety is, the more relational choices participants make, i.e. there is no point above which the relational choices have declined.

Alternatively, the difference might be due to the difference in the analogy tasks used in both experiments. This would be an interesting avenue for research since it would require task analysis and decomposition of the "analogy-making process" into simpler mechanisms and exploring the role of anxiety for each of these components. Thus, for example, in both tasks – the cross-mapping corresponding task used by Tohill & Holyoak (the subject has to point to the corresponding object of a hinted one) and the match-to-sample task used in the current study (the subject has to chose which of two alternative situations is more similar to the sample) – the participants have to encode certain relations and attributes of the objects and than build the two alternative mappings, and finally chose the better one. According to some models of analogy-making like ARCS (Holyoak & Thagard, 1989), AMBR (Kokinov, 1994, Kokinov & Petrov, 2001), CopyCat and TableTop (Hofstadter, 1995), LISA (Hummel & Holyoak, 1997) there are at least 3 subprocesses of analogy-making: perceiving (encoding) the relations, forming hypotheses of possible correspondences, and competition between them (constraint satisfaction), other models like SME (Gentner,

1983, Falkenheiner, Forbus, Gentner, 1989) offer alternative but analogous subprocesses. Anxiety may influence each of these subprocesses specifically. Since Tohill & Holyoak (2000) allow their subjects to observe the two pictures for 15 sec before the question was asked, this would mean that all relations are already encoded and possibly also most of the hypotheses are formed and after the query mostly the constraint satisfaction process continues. Thus the influence of anxiety would be mainly on the constraint-satisfaction outcomes. In our study participants used on average 4 sec for the whole task of encoding, hypotheses building, and constraint satisfaction. Therefore, it is reasonable to assume that they do not have the time for full relational encoding, building all possible hypothesis, etc. Most probably they encode only a few relations and form a few hypotheses and therefore the constrain-satisfaction process is quite straightforward. Thus most probably the anxiety state influences mostly the process of relational encoding in this case. According to the DUAL architecture and the AMBR model (Kokinov, 1994, Kokinov & Petrov, 2001) anxiety concentrates the activation over a smaller area of Long-Term Memory thus causing a smaller search space but faster processing within this space (Feldman & Kokinov, 2009). Thus maybe the anxiety state in our task causes a speeded search for relational encoding (especially given the restricted number of relations used in the stimuli) and hypotheses formation and that is how anxiety enhances relational choices.

Such a possibility is potentially and indirectly backed up by neuroscience approaches to anxiety and its influences on cognitive processes. Posner, Rueda, and Kanske (2007) distinguished 3 main attentional neural networks – *alerting network* (associated with the right frontal and parietal brain areas which contributes to the maintenance of the sensitivity level needed for perceiving and processing stimuli), *orienting network* (associated with the superior parietal lobe, frontal eye fields, and temporoparietal junction which contributes to the selection of information from among numerous sensory stimuli), and *executive control network* (associated with midline frontal areas, anterior cingulate gyrus, and lateral prefrontal cortex which contributes to the conflict resolution and voluntary action control) which could be somehow related to the three processes described above: encoding relations, building hypotheses, and constraint satisfaction. The encoding of relations would depend on the alerting network allowing bottom-up recognition of relations; the hypotheses formation – on the orienting network selecting potential correspondences; and the constraint satisfaction depending on the inhibitory capacity of the executive control. A recent study by Pacheco-Unguetti, Acosta, Callejas, Lupianez (2010) found that the anxiety state enhances the work of the alerting and orienting networks, while no significant effect was found on the executive network, while the trait anxiety has no effect on the alerting and orienting networks, but severely diminishes the executive control and its possibilities for inhibition. Thus "state anxiety is related to greater orienting and alerting effects, thus making participants more sensitive to bottom-up processing" (Pacheco-Unguetti et al., 2010). This might mean that in an anxiety state people are more rapidly encoding the relations which are otherwise difficult to be perceived and this could explain why the anxiety-induced subjects made more relational choices in our experiment. This hypothesis can be potentially backed up also by the study of Becker (2009) who found that in the presence of threatening stimuli people are faster in visual search also for non-threatening stimuli, i.e. faster encoding is performed. It is true that the search he has studied is for objects, not relations, but we plan an experimental study to test whether this speeded processing will also be extended to relations as we assume. At the same time the anxiety-induced subjects in the Tohill and Holyoak (2000) study had the necessary time to encode all relations in advance and therefore the effect could be due either to the limited capacity of working memory (Eysenck & Calvo, 1992) or to impoverished constraint satisfaction. Of course, all these are wild speculations and further studies are necessary to test these hypotheses.

The main conclusion from this study is that the influence of anxiety on analogical mapping is much more subtle and complicated than previously thought and that we need to study more carefully the influence of anxiety on each of the components of the analogy-making process before jumping to bold conclusions.

Acknowledgments

This research was supported financially by the EXREL project (NEST program, contract 43225) funded by the EC. We would like to specially thank Simona Dobrinova for helping us with the data collection process.

Reference

Becker, M. (2009). Panic Search: Fear Produces Efficient Visual Search for Nonthreatening Objects. *Psychological Science*, 20(4), 435-437.

Bliznashki, S., Kokinov, B. (2009). Analogical Transfer of Emotions. In: Kokinov, B., Holyoak, K., Gentner, D. (eds.). New Frontiers in Analogy Research. Sofia: NBU Press

Eysenck, M., Calvo M. *(*1992*).* Anxiety and performance: The processing efficiency theory. *Cognition & Emotion*, 6, 409-434

Falkenhainer, B., Forbus, K., Gentner, D. The Structure-Mapping Engine: Algorithm and Examples. *Artificial Intelligence*, 41(1): 1-63 (1989)

Feldman, V., Kokinov, B. (2009). Anxiety Restricts the Analogical Search in an Analogy Generation Task. In: Kokinov, B., Holyoak, K., Gentner, D. (eds.). New Frontiers in Analogy Research. Sofia: NBU Press

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. Cognitive Science, 7, 155-170.

Graeff1, F., Parente, A., Del-Ben, C., Guimarães, F., (2003). Pharmacology of human experimental anxiety. Brazilian Journal of Medical and Biological Research, 36: 421-432.

Hristova, P.(2009) People in Negative Mood May See Relations Where People in Positive Mood May Not In B. Kokinov, K. Holyoak & D. Getner (Eds.), New Frontiers in Analogy Research (pp. 204-210). Sofia, Bulgaria: NBU Series in Cognitive Science

Hofstadter, D. (1995). *Fluid concepts and creative analogies: computer models of the fundamental mechanisms of thought*, NY: Basic Books.

Holyoak, K., & Thagard, P. (1989). Analogical mapping by constraint satisfaction. *Cognitive Science*, *13*, 295-355.

Hummel, J., Holyoak, K. (1997). Distributed Representations of Structure: A Theory of Analogical Access and Mapping. *Psychological Review*, 104, 427-466.

Keinan, G. (1987). Decision making under stress: Scanning of alternatives under controllable and uncontrollable threats. Journal of Personality and Social Psychology, 52, 639–644.

Kokinov, B. (1994). A hybrid model of reasoning by analogy. In K. Holyoak & J. Barnden (Eds.), *Advances in connectionist and neural computation theory*: *Analogical connections*. Norwood, NJ: Ablex.

Kokinov, B., & Petrov, A. (2001). Integration of Memory and Reasoning in Analogy-Making: The AMBR Model. In D. Gentner, K. J. Holyoak, & B. Kokinov (Eds.), *The analogical mind: Perspectives from cognitive science*. Cambridge, MA: MIT Press.

Leon, M., Revelle, W. (1985). Effects of anxiety on analogical reasoning: A test of three theoretical models. Journal of Personality and Social Psychology. 49(5), 1302-1315.

Pacheco-Unguetti, A., Acosta, A., Callejas, A., Lupianez, J. (2010). Attention and Anxiety: Different Attentional Functioning Under State and Trait Anxiety. Psychological Science OnlineFirst published on January 22, 2010: http://pss.sagepub.com/content/early/2010/01/21/0956797609359624

Pertaub, D., Slater, M., Barker, B. (2002). An Experiment on Public Speaking Anxiety in Response to Three Different Types of Virtual Audience. *Presence-Teleoperators and Virtual Environments* 11(1), 68-78.

Posner, M., Rueda, M., Kanske, P. (2007). Probing the mechanisms of attention. In Cacioppo J.T., Tassinary J.G., Berntson G.G. (Eds.), *Handbook of psychophysiology* (3rd ed., pp. 410–432). Cambridge, England: Cambridge University Press.

Richert, R., Whitehouse, H., Stewart, E. (2005). Memory and analogical thinking in high arousal rituals. In: Whitehouse, McCauley, Eds., *Mind and religion: psychological and cognitive foundations of religiosity*. AltaMira Press: Walnut Creek, CA

Sloutsky, V. M., & Yarlas, A. S. (submited) Processing of information structure: Mental representations of elements and relations.

Spellman, B. & Holyoak, K. (1992). If Saddam is Hitler then who is George Bush? Analogical mapping between systems of social roles. *Journal of Personality and Social Psychology, 62*, 913-933.

Thagard, P., & Shelley, C. P. (2001). Emotional analogies and analogical inference. In D. Gentner, K. H. Holyoak, & B. K. Kokinov (Eds.), *The analogical mind: Perspectives from cognitive science* . Cambridge, MA: MIT Press. 335-362.

Tohill, J., Holyoak, K. (2000). 'The impact of anxiety on analogical reasoning', Thinking & Reasoning, 6:1, 27 – 40

Yerkes, R. M., & Dodson, J. D. (1908) The relation of strength of stimulus to rapidity of habit-formation. Journal of Comparative Neurology and Psychology, 18, 459-482

Impact of mood induction on temporal processing

Katrina S. Rodzon (krodzon@gmail.com)
Utah State University, Department of Psychology
2810 Old Main Hill, Logan UT, 84322

Kerry Jordan (Kerry.jordan@usu.edu)
Utah State University, Department of Psychology
2810 Old Main Hill, Logan UT, 84322

Abstract

The durations of negative events are overestimated when compared to the actual amount of time passed (Langer, et al, 1961; Meck, 1983). Similarly, emotionally valenced faces are temporally overestimated when compared to neutral ones (Droit-Volet, Bruno, & Niedenthal, 2004). In the current study, participants embodied emotion via mood induction prior to temporal estimation of neutrally valenced faces. Valenced mood induction led to overestimation of the duration of neutral faces. Results support the claim that embodiment of emotion can cause subjective temporal distortion.

Introduction

Several movies, advertisements, and societal icons portray the old adage, 'Time flies when you are having fun', but only recently did science begin investigating the legitimacy of this statement. Does subjective experience of duration really change when in emotionally valenced situations? Experiencing emotional events such as foot shocks (Meck, 1983) and approaching threatening stimuli (Langer, et al., 1961) results in overestimation of event duration, implying that subjective experience of time speeds up when participating in negative events. Simply put, negative situations produce the feeling of more time going by than has actually passed.

The current paper investigates the impact of emotions on temporal perception and the scalar theory of timing (Gibbon, 1977). Past research has found that humans overestimate the duration of emotionally valenced faces (Droit-Volet, Bruno, & Niedenthal, 2004), and the current study evaluates an embodied emotion premise for this subjective temporal bias. After mood induction, participants evaluated the duration of neutrally valenced faces. We hypothesized that participants in valenced moods would overestimate the duration of these neutral events, supporting the impact of embodied emotions on subjective temporal perception.

The Scalar Theory of Timing

A predominant theory of timing is the scalar timing theory (Gibbon, 1977) which comprises two fundamental properties: (1) the internal clock is, on average, accurate in estimating stimulus durations, and (2) the greater the mean duration of time has passed, the larger the variability of the internal clock's estimation. Three stages are outlined in scalar timing theory: (1) the clock stage, (2) the memory stage, and (3) the decision stage. During the clock stage, a pacemaker emits pulses that are stored in an accumulator, with more pulses representing longer durations. The accumulator is opened or closed with a mode switch, allowing specific events to be separated for temporal estimation. Once event timing is complete, the contents of the accumulator are stored for later use in the decision stage in working memory, where they are compared to previously experienced durations stored in long-term memory. The decision stage allows for appraisal of relative values in order to make an assessment of time.

In this scalar model of timing, an attentional system—which can allocate differential resources to incoming stimuli based on perceived importance—is added to the modal switch of the clock phase, helping explain erroneous time estimation. While scalar timing theory posits that the internal clock is generally accurate, over and under-estimation of time does occur. For example, previous research suggests

that attentional distraction can either delay the mode switch closing or prematurely open it, resulting in a net loss of pacemaker pulses. This loss results in an underestimation of time (Buhusi & Meck, 2006; Coull, et al., 2004; Lejeune, 1998; Macar, 2002; Meck & MacDonald, 2007).

Effects of Emotion on Attention and Timing

Emotional salience can significantly impact attentional priority, with highly emotional stimuli directing both conscious and unconscious attention away from neutral stimuli (Taylor & Fragopanagos, 2005). Emotional stimuli have been shown to: (a) be detected faster and more accurately than neutral stimuli, regardless of the number of distracters (Ohman, Flykt, & Esteves, 2001), (b) remain more detectable within an attentional blink paradigm, even persisting past the point at which neutral stimuli become minimally detected (Anderson & Phelps, 2001), and (c) capture automatic attention earlier than neutral stimuli when measured by event-related potentials (Carretie et al., 2004). Furthermore, affective priming's impact on emotional judgment (Murphy & Zajonc, 1993) and amygdala activation of backwards-masked emotional stimuli (Whalen, et al., 1998) also demonstrate how both detected and undetected emotional stimuli impact cognitive and neural processes involved with attention.

Recent findings have provided substantial evidence that emotions also impact temporal processing by causing overestimation of the duration of emotional: (a) events (Langer, et al., 1961; Meck, 1983; Stetson, Fiesta, & Eagleman, 2007), (b) faces (Droit-Volet, Brunot, & Niedenthal, 2004, Gil, Niedenthal, & Droit-Volet, 2007), and (c) other stimuli (Angrilli, et al., 1997; Noulhiane, et al., 2007). When experiencing stressful events, such as foot shocks (Meck, 1983), approaching threatening stimuli (Langer, et al., 1961), and forcing eye contact with an angry face (Schiff & Thayer, 1970), higher arousal level is hypothesized to increase the pacemaker's speed, thereby impacting the number of pulses acquired in the accumulator. In addition, a significant interaction between emotional valence and arousal has been found, with the duration of high arousal, negative stimuli being overestimated when compared to high arousal, positive stimuli (Angrilli, et al., 1997). The relationship is reversed when low arousal stimuli are presented, with the duration of negative stimuli being underestimated when compared to positive stimuli (Angrilli, et al., 1997). Finally, the durations of emotionally valenced faces (i.e. angry, happy, and sad) are significantly overestimated when compared to neutral faces in a duration bisection task (Droit-Volet, et al., 2004). These results are consistent with those of Schiff and Thayer (1970), who found that perceived duration of forced eye contact with an angry face was significantly longer than perceived duration of eye contact with a neutral face.

Together, these findings suggest that emotional stimuli, events, and faces impact the speed of the pacemaker invoked by scalar timing theory. Other research suggests that the impact of emotional faces on temporal processing may also involve embodied cognition of perceived emotion.

Embodiment of Emotion

Viewing emotional events, stimuli and faces similarly affect temporal processing, but a remaining question is whether the experience (embodiment) of emotions affects temporal estimates of neutrally valenced stimuli. Studies suggest that embodiment of emotions occurs when exposed to valenced faces (Chambon et al., 2008; Effron, et al., 2006). Embodiment of other groups' physiological behavior has been demonstrated by Bargh, et al (1996) with participants: (a) walking slower when exposed to elderly stereotype words in a word search, and (b) being more likely to interrupt when exposed to rude stereotype words. Similarly, Chambon, et al. (2008) hypothesized slowing down of the internal clock speed when exposed to elderly faces versus younger faces. Effron, et al. (2006) investigated if an embodied cognition approach could specifically explain the impact of emotional facial stimuli on temporal processing; indeed, inhibiting imitation of viewed facial expressions (by having participants hold a pen between their lips) eliminated any overestimation of the duration of valenced faces. These results suggest that imitation of facial

expression may influence timing processes, though additional evidence for the role of embodied emotion on temporal processing is needed.

The current study will thus further investigate the role of embodied cognition through evaluation of the implication set forth by Effron, et al. (2006). If perceived mood is embodied, and as such impacts temporal perception, similar effects should be seen when mood is induced and neutral stimuli are evaluated as when valenced stimuli are evaluated in a neutral mood. Induction of positive, negative and neutral moods was utilized to determine whether emotionally valenced mood leads to duration overestimation of neutral faces similar to the effect seen when timing valenced stimuli. If overestimation of neutral facial stimuli were found for those in a positive and negative mood, compared to participants in a neutral mood, results would indicate that subjective temporal distortion could occur via embodiment of emotion. The current study also evaluated the influence of emotion on the scalar timing theory, specifically the impact of emotional arousal and attention prioritization. Differences in point of subjective equality between moods revealed a bias shift, seen in previous literature, implicating an increase in pacemaker speed during emotionally arousing situations. Furthermore, attentional prioritization of any emotional stimuli used in previous studies was decreased through the use of neutral stimuli in this study, thereby allowing for the exclusive analysis of embodied emotional arousal on temporal perception.

Methods

Participants

Participants consisted of 32 undergraduates (males: n = 14, females: n = 18 in psychology classes at Utah State University (neutral mood: n = 12, positive mood: n = 7, negative mood, n = 13). Participants received course credit for participating.

Material: Apparatus and Stimuli

All participants were asked to complete a computer-based bisection task taking approximately 30 minutes. The experiment was run on a Dell Optiplex 755 computer with a 21 inch monitor in a dimly lit room. Participants sat approximately 45 cm from the display. The task stimuli were presented and data were recorded using E-prime, and participants made all responses using a keyboard. The stimulus presented for the practice trials was a white oval (9 x 10 cm) similar to that used by Droit-Volet, et al. (2004). One photo of a female face with a neutral expression, which had been coded using the Facial Action Coding System (Tracy, et.al., 2009), was used for the testing trials (44x .32 cm).

Procedure

Before the bisection task, a mood induction procedure was run in which each participant was presented with a series of either positive (n = 25), negative (n = 24), or neutral (n = 35) Velten statements (Velten, 1968) that progressed automatically on the computer screen over the course of 8 minutes. Participants were instructed to "read each and think about them as if you were experiencing them." This procedure has been used to induce both positive and negative moods in many previous studies (Jennings, et al., 2000; Sinclair, et al., 1994; Strickland, et al.1974).

Immediately following the mood induction, participants completed a temporal bisection task similar to the one used in Droit-Volet et al. (2004) with two trial phases: (1) practice, and (2) testing. Participants pressed the space bar to initiate each trial. In the practice phase, a white oval was presented for the longest (1600 ms) and shortest (400 ms) durations. Participants had to press the 'd' key if the duration was closer to 400 ms or the 'k' key if the duration was closer to 1600 ms. Each stimulus was presented 8 times, for a total of 16 trials. Accuracy feedback was given after each trial; positive feedback consisted of 'Correct!' displayed visually for 1500ms, while negative feedback consisted of 'Incorrect' displayed visually for 1500ms. Participants were then instructed to press the spacebar to begin the next trial. In the testing phase, participants were presented with a neutral face rather than a white oval as the stimulus to be timed, and feedback was eliminated. This face was presented for 18 trials each at each of 7 durations in random order, including the shortest and longest durations from the training phase and various

intermediate durations (400, 600, 800, 1000, 1200, 1400, and 1600 ms) for a total of 126 trials.

Results

For the training phase, participants in all mood conditions demonstrated accuracy on the bisection task prior to starting the test trials (neutral = 91.67%, positive = 98.43%, negative = 93.75%). For the testing phase, the mean proportion of long responses was calculated for each stimulus duration and separated by mood condition (see Figure 1).

To evaluate any significant differences between groups, a non-linear regression analysis was performed [model: Y=1/(1+[x/T50]^-E)] followed by a statistical comparison of the slopes (E) and subjective mid-point (T50) using a student's t-test. No significant differences between slope were found across groups. The following significant differences between subjective mid-point (T50)—the stimulus duration that the participant is equally like to categorize as 'short' or 'long'--were found: (a) those in a positive mood had a significantly lower T50 than those in a neutral mood (t (15) = -4.414; p <.01; positive: T50 = 893; neutral: T50 = 948.6), and (b) those in a negative mood had a significantly lower T 50 than those in a neutral mood (t (21) = -3.187; p <.01; negative: T50 = 904.8; neutral: T50 = 948.6). Thus, lower points of subjective equality were found in the positive and negative mood groups, as compared with the neutral mood group, supporting the premise that induction of valenced moods causes overestimation of the duration of neutral events.

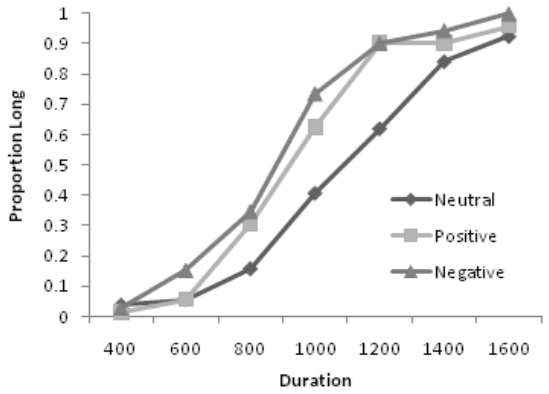


Figure 1. Proportion of 'long' responses for the duration of neutral facial stimuli for participants in neutral, positive, and negative moods.

As evidenced in Figure 1, this produces a leftward shift for the psychophysical functions of participants in valenced moods compared with neutral moods, and results in a bias to respond 'long'.

Discussion

The present findings replicate the effect of emotion on temporal perception found by Droit-Volet, et al. (2004) with emotionally valenced mood causing overestimation of neutral stimuli duration compared to duration estimation in neutral moods. The current use of mood induction, instead of emotional stimuli, supports the claims set forth by Effron et al. (2006) outlining the embodiment of perceived emotion significantly impacting temporal perception. The impact of experienced mood on temporal perception being identical to the impact of observed mood on temporal perception endorses the idea that embodying perceived emotion causes temporal bias when judging stimulus duration.

The current findings also speak to the influence of arousal on the scalar timing theory by illustrating a significant difference in point of subjective equality as well as no significant difference in sensitivity – slope- between groups. The use of neutrally valenced stimuli, as well as no slope differences between groups, indicates little if any impact of attentional demands on temporal perception. Furthermore, significant differences in point of subjective equality supports previous findings that arousal can impact pacemaker speed (Droit-Volet, et al., 2004; Gil & Droit-Volet, 2009). Positive and negative moods increase arousal levels, thereby causing faster pulse emission from the pacemaker and resulting in longer subjective judgment of time passed. Overall, in conjunction with previous research examining the effects of emotionally valenced stimuli on temporal perception, the current findings: (a) reveal the same impact via valenced mood induction on timing of neutral stimuli, and (b) suggest that

embodiment of emotions may distort temporal perception via increased arousal.

A body of research on depression and temporal perception suggests that the slowing of pacemaker speed in depressed individuals causes time to pass subjectively slower and underestimation of time (Blewett, 1992; Gil & Droit-Volet, 2009). There is a deceleration of general motor function in depression (Lemke, et al., 2000), however, manifested in reports of helplessness and resignation and not seen in non-depressed patients in a sad mood. This difference in motor function speed could account for the disparity in temporal perception in depressed and non-depressed patients. Regardless, differences in temporal perception between clinical populations with affective disorders (i.e. depression, bi-polar disorder) should be further addressed in future research.

Whether temporary valenced mood increases pacemaker speed or depression slows it, both support the idea that embodiment of emotions can drive temporal biases. When imitation of viewed facial expression is inhibited, for example, stimulus valence fails to impact temporal perception, suggesting that merely *perceiving* emotions in others is not sufficient to impact timing (Effron et al., 2006). The current finding that temporary mood induction produces the same effect on temporal perception as perceived mood further supports the claim that embodiment of emotions may play a mechanistic role in the influence of valence on timing.

References

Anderson, A.,K., & Phelps, E.,A. (2001). Lesions of the human amygdale impair enhanced perception of emotionally salient events. *Nature, 411*, 305-309.

Angrilli, A., Cherubini, P., Pavese, A., & Manfredini, S. (1997). The influence of affective factors in time perception. *Perceptuatl Psychophysiology, 59*, 972-982.

Bargh, J.A., Chen, M., & Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology, 71,* 230-244.

Blewett, A. E. (1992). Abnormal subjective time experience in depression, *The British Journal of Psychiatry, 161,* 195-200.

Buhusi, C.V., & Meck, W.H. (2006). Interval timing with gaps and distracters: Evaluation of the ambiguiting, switch, and time-sharing hypothesis. *Journal of Experimental Psychological Animal Behavior Processes, 32*, 329-338.

Caratie, L., Hinojosa, J.A., Martin-Loeches, M., Mercado, F., & Tapia, M. (2004). Automatic attention to emotional stimuli: Neural correlates. *Human Brain Mapping, 22*, 290-299.

Chambon, M., Droit-Volet, S., & Niedenthal, P.M. (2008). The effect of embodying the elderly on time perception. *Journal of Experimental Social Psychology, 44,* 672-678.

Coull, J. T., Vidal, F., Nazarian, B., & Macar, F. (2004). Functional anatomy of the attentional modulation of time estimation. *Science, 303*, 1506-1508.

Droit-Volet, S., Brunot, S., & Niedenthal, P.M. (2004). Perception of duration of emotional events. *Cognition and Emotion, 18*, 849-858.

Effron, D., Niedenthal, P.M., Gil, S., & Droit-Volet, S.(2006). Embodied temporal perception of emotion. *Emotion, 6*, 1-9.

Gibbon, J. (1977). Scalar expectancy theory of Weber's law in animal timing. *Psychological Review, 84*, 279-325.

Gil, S., & Droit-Volet, S. (2009). Time perception, depression and sadness. *Behavioral Processes, 80*, 169-176.

Gil, S., Niedenthal, P.M., Droit-Volet, S. (2007). Anger and time perception in children. *Emotion, 7*, 219-225.

Jennings, P.D., McGinnis, D., Lovejoy, S., & Stirling, J. (2000). Valence and arousal ratings for Velten mood induction statements. *Motivation and Emotion, 24,* 285-297.

Langer, J., Wapner, S., & Werner, H. (1961). The effect of danger upon the experience of time. *American Journal of Psychology, 74*, 94-97.

Lejeune, H. (1998) Switching or gating? The attentional challenge in cognitive models of psychological time. *Behavioral Processes, 44*, 127-145.

Lemke, M.R., Koethe, N.H., Schleidt, M. (2000). Segmentation of behavior and time structure of movements in depressed patients. *Psychopathology, 33,* 131-136.

Macar, F. (2002). Expectancy, controlled attention and automatic attention in prospective temporal judgements. *Acta Psychologica, 111*, 243-262.

Meck, W.H. (1983). Selective adjustment of speed of internal clock and memory processes. *Journal of Experimental Psychological Animal Behavioral Processes, 7*, 18-30.

Meck, W.H., & MacDonald, C.J. (2007). Amygdala inactivation reverses fear's ability to impair divided attention and make time stand still. *Behavioral Neuroscience, 121,* 707-720.

Murphy, S. T., & Zajonc, R. B. (1993). Affect, cognition, and awareness: Priming with optimal and suboptimal stimulus exposures. *Journal of Personality and Social Psychology, 64,* 723-739.

Noulhiane, M., Mella, N., Samson, S., Ragot, R., & Pouthas, V. (2007). How emotional auditory stimuli modulate time perception. *Emotion, 7*, 697-704.

Ohman,A., Flykt, A., & Esteves, F. (2001). Emotion drives attention: Detecting the snake in the grass. *Journal of Experimental Psychology: General, 130*, 46-478.

Shiff, W., & Thayer, S. (1970). Cognitive and affective factors in temporal behavior: Judgements of intrinsically and extrinsically motivated successful and unsuccessful performances. *Perceptual & Motor Skills, 30*, 885-902.

Sinclair, R.C., Mark, M. M., Enzle, M E., Borkovec, T. D., & Cumbleton, A.G. (1994). Toward a multiple-method view of mood induction: The appropriateness of a modified Velten mood induction technique and the problems of procedures with group assignment to conditions. *Basic & Applied Social Psychology, 4*, 389-408.

Stetson, C., Fiesta, M.P., & Eagleman, D.M. (2007). Does time really slow down during a frightening event? *PLoS ONE, 12*, 1-3.

Strickland B. R., Hale, W. D., & Anderson, L.K. (1974). Effect of induced mood states on activity and self reported affect. *Personality and Social Psychology Bulletin, 1,* 399-401.

Taylor, J.G., & Fragopanagos, N.F. (2005). The interaction of attention and emotion. *Neural Networks, 18* 353-369.

Tracy, J. L., Robins, R.W., & Schriber, R.A. (2009). Development of a FACS-verified set of basic and self-conscious emotion expressions. *Emotion, 9*, 554-559.

Velten, E. (1968). A laboratory task for induction of mood states. *Behavior Research & Therapy, 6,* 473-482.

Whalen, P. J., Rauch, S. L., Etcoff, N.L., McInerney, S.C., Lee, M. B., Jenike, M.A. (1998). Masked presentation of emotional facial expressions modulate amygdala activity without explicit knowledge. *The Journal of Neuroscience, 18,* 411-418.

Hot Cognitions in Coherence-Based Reasoning and Decision-Making

Stephen J. Read (read@usc.edu)
Department of Psychology, University of Southern California
Los Angeles, CA 90089

Dan Simon (dsimon@law.usc.edu)
Gould School of Law, University of Southern California
Los Angeles, CA 90089

Douglas Stenstrom (dstenstrom@email.com)
Department of Psychology, California State University, Los Angeles
Los Angeles, CA 90032

Abstract

The studies examine the role of *hot cognitions* alongside *cold cognitive appraisal* within the framework of *coherence-based reasoning.* In two simulated legal cases we find that emotions towards the suspect and motivation with respect to the outcome of the case are strongly correlated with the cognitive appraisal of the facts of the case, the judged credibility of the witnesses, and the overall judgment of the suspect's blame. Moreover, emotion and motivation partially mediate the effect of experimental manipulations on decisions.

Keywords: Decision-making; constraint satisfaction processes; coherence based reasoning; legal decision-making.

Introduction

Decision making in real-world situations characterized by complex patterns of facts often involves *coherence-based reasoning*; as decision makers consider a pattern of evidence and come to a conclusion, judgments about the facts of the case and the conclusion shift to become more coherent with each other (Holyoak & Simon, 1999; Simon, 2004; Simon et al., 2004a; Simon et al., 2004b). We sought to extend this research by investigating the role of *hot cognitions* in the *cold* cognitive appraisals involved in such judgments. We were particularly interested in whether and how emotions and motivation are implicated in conclusions about a suspect's guilt. Considerable research has recently examined the role of emotions in decision-making (Rick & Loewenstein, 2008). Particular attention has been directed at anger, which leads to systematic distortions in a variety of judgments. These distortions are especially problematic when the anger is aroused by a source that is unrelated to the person being judged. Observers aroused by such *incidental anger* are more likely to attribute blame to the person being judged, to perceive her conduct as intentional, to lower the required threshold of evidence, to neglect alternative explanations and mitigating circumstances (Lerner, Goldberg, & Tetlock, 1998; Goldberg, Lerner, & Tetlock, 1999; Quigley & Tedeschi, 1996), and to increase the desire for retaliation (Ferguson & Rule, 1983)

Social judgment has also been shown to be affected by motivation. As noted by Kunda (1990), reasoning processes under *directional goals* often lead to results that comport with those goals, whereas *accuracy goals* tend to lead to more objective conclusions (Balcetis & Dunning, 2006; Ditto & Lopez, 1992; Piercey, 2009).

In the current studies we sought to study the impact of *directional goals* by giving some subjects a specific role as either prosecutor (or investigator) and other subjects the role of defender. Taking on such *adversarial roles* may lead to biased information search and hypothesis testing. We also hypothesized that such *adversarial roles* may lead to negative emotions, such as anger.

Unfortunately the research demonstrating the effect of emotion and motivation on reasoning offers little insight into how these effects occur. How do emotion and motivation interact with the variables on which the judgments are supposed to be based: facts, preferences, values, etc.? One possibility is that emotion and motivation override these underlying variables. Another possibility is that emotion and motivation influence the underlying variables in the corresponding direction, which makes the corresponding judgments feel natural and obvious. The latter explanation is consistent with the Gestaltian notions that underlie *coherence-based reasoning*: the mental model of the task settles at a state of equilibrium at which all relevant elements—the underlying variables, conclusion, motivation, and emotion—all cohere with one another. Thus, we hypothesized that the constraint satisfaction processing that underlies *coherence-based reasoning* would engulf both the cold cognitions (as observed previously) and the hot cognitions. This prediction dates back to Heider's Balance Theory, in which liking for a person or an object was theorized to affect the overall balance of the structure (Heider, 1958). More recently, researchers have modeled hot and cold cognitions within the framework of constraint satisfaction processing (Nerb, 2007; Thagard, 2006).

Overview of Studies

In both studies, participants judged a quasi-criminal case concerning an allegation of academic misconduct by a university student. Participants were asked to imagine that they worked at a state university in the Office of Student

Disciplinary Affairs, which deals with allegations of academic misconduct. The Office investigates and adjudicates the allegations and, where appropriate, recommends disciplinary actions. The procedure consisted of an investigation followed by an adversarial-like disciplinary hearing, in which a University Representative prosecutes the case, and the student is defended by a Student Representative. The cases are ultimately decided by the university's Chief Judicial Officer.

The case involved an allegation that a student, Debbie Miller, cheated on a closed-book exam by copying from her notes. Participants received the case information, and were asked to make a variety of judgments about the incident. All participants received the same case information and instructions, except for assignment instructions, as described below. None of the manipulations provided any information pertaining to whether she cheated or not. In all conditions, participants were instructed to be "fair and objective."

The first study examines whether the objectivity of investigation is affected by directional motivations and emotions that are elicited by the adversarial nature of the process. Participants were asked to play the role of the investigator, and assigned to investigate the case for one of the two parties (the two adversarial assignments) or for both (the non-adversarial assignment).

The second study examines the effects of the intensity of adversarialism. Participants were asked to role-play a prosecutor-like role in a case of alleged academic misconduct. Half of the participants were given background information intended to induce low intensity (*non-partisan)* (*e.g.*, you feel that most of the time, the disciplinary process reaches correct decisions), while the other half were given information intended to induce high intensity (*partisan)* (*e.g.*, you believe that many of the students who were cleared by the disciplinary process did in fact cheat).

Study 1: Adversarial and Inquisitorial Investigations

This study tested whether and how investigations conducted in an adversarial framework might lead to different outcomes than investigations conducted in a non-adversarial mode. Participants were assigned to investigate the case for either one of the parties (two adversarial conditions) or for both parties (the non-adversarial condition). We predicted that relative to the non-adversarial assignment, the adversarial assignments would result in views of the case that would be tilted towards the respective assignments and that these views would be mediated by motives and emotions elicited by the role assignment.

Method

Participants. Participants were 296 individuals who completed the study via the Internet. The sample was 62% female, with an average age of 43.

Procedure. Participants went through a series of web pages containing the instructions, the case information, and the measures. They were informed that the assigned role of investigator entailed preparing the evidence to be submitted to the disciplinary hearing. All participants received the same case information and instructions, except for assignment instructions, as described below.

Assignment. Participants were randomly assigned to one of three conditions. The "university-assignment" condition was designed to simulate a police investigation. Instructions emphasized that the individual was performing the investigation on behalf of the University and their reports were central to the case. They were also told that someone would be fulfilling a similar function for the other side.

The "Debbie-assignment" condition was designed to simulate a private investigation for the defense. The instructions for this assignment were identical to the university assignment, but the sides were reversed.

The "Sole Investigator" condition was designed to mimic a non-adversarial investigation; participants were told that they were the sole investigator in the case. The instructions emphasized that they were the only investigator working on the case and that both sides would rely on their report.

All participants were exposed to the same case and instructed to be fair and objective. Participants performed the study alone, and there was no other investigator.

Case. The case was intricate and ambiguous. From the university files, participants learned that Debbie, a junior, was an "A" student, and was considered hardworking and ambitious. At high school, she was charged with cheating on an exam, but the file did not indicate whether she was disciplined or not. An interview with the examination room proctor revealed that Debbie sat against a wall, close to the back corner of the room. The proctor noticed that Debbie sat crouched over her papers, as if she was hiding something. At the end of the exam, she noticed also that Debbie stuck something into the pocket of her sweater, which later turned out to be a note with a summary of the course. Brad Loomis, a fellow student who sat behind Debbie, claimed to have seen her pull out the note from her sweater pocket and copy from it throughout the exam. The professor reported that Debbie was anxious about the exam, but did not believe that she cheated. He did mention that she was the only student to respond correctly to one of the questions. Debbie denied the allegations adamantly. She stated that as an A student, she had only to lose by cheating. She explained that she crouches when sitting for long periods of time because of a back injury she sustained while playing on the college volleyball team.

Dependent Variables. *1. Overall Judgments.* Participants estimated the likelihood that Debbie cheated on the exam (0-100%), how they would decide the case, how they expected the Chief Judicial Officer to decide the case, and which side their view supported.

2. Case facts and related beliefs. Participants evaluated 13 factual issues involved in the case, and 9 belief questions that corresponded to 9 of the factual questions (1 - 11 scale).

3. Judgments of Liking, Emotions, and Motivation. Participants indicated how much they liked Debbie (0-100). Next, they reported how much they felt three positive

emotions (sympathy, compassion, and sorrow) and three negative emotions (anger, scorn, disgust) towards Debbie. Another question gauged participants' motivation towards the outcome of the case by asking participants which side they wanted to see win the case. (all on a 1-11 scale)

4. Objectivity and Distrust. The questions measured participants' assessments of the objectivity of their own view of the case; the objectivity of the other investigator; how their own objectivity would be judged by the other investigator; and how the Chief Judicial Officer would assess their own objectivity and the other investigator's. (all on 1-11 scale).

Results

The prediction was that role assignment would influence participants' judgments of all aspects of the case.

1. Overall Judgments. The assignment had the predicted effects on overall judgments of the case. The estimates of the probability that Debbie cheated were 33%, 43%, and 53% for the Debbie-Assignment, Sole Investigator, and University-assignment conditions, respectively, $F(2, 292) = 12.75$, $p < .001$. A similar pattern was found in participants' judgments as to which side of the case was supported by their view: 3.5, 5.0, and 5.8, with higher numbers meaning more University support, ($F(2, 292) = 15.17$, $p < .001$. A chi-square analysis, Chi-square (2) = 6.99, $p < .05$, revealed that the assignment also influenced how participants would decide the case themselves (23%, 37%, and 40% would decide that Debbie cheated, respectively).

2. Case facts and related beliefs. First, consistent with prior research on *coherence-based reasoning* (Holyoak & Simon, 1999; Simon et al., 2001; Simon et al., 2004b), views of these items clustered around a coherent mental model of the case. The 13 fact items formed a reliable composite (alpha = .88). Participants developed globally coherent structures that tended to view the factual pattern as indicative either that Debbie cheated or that she did not. We found a similar clustering of the 9 beliefs that were related to the facts of the case (alpha = .60). This weaker alpha is understandable given that background knowledge is more stable than ad hoc judgments of specific events.

Second, the assignment influenced the facts and related beliefs as predicted, Facts $F(2, 292) = 15.87$, $p < .001$; Beliefs $F(2, 292) = 14.11$, $p < .001$. Those assigned to the university-condition were more prone to interpret the facts as incriminating Debbie (Fact M = 5.7, Belief M = 5.4), whereas those assigned to the Debbie condition interpreted them as least incriminating (Fact M = 4.4, Belief M = 4.5). The judgments in the Sole Investigator condition were in between (Fact M =5.2, Belief M = 5.0),

3. Judgments of Liking, Emotions, and Motivation. The assignment also influenced liking and emotional reactions to Debbie, as well as motivation with respect to the outcome. Participants in the university-condition were consistently the most negative toward Debbie, whereas those in the Debbie assignment condition were consistently most positive, with Sole Investigator in between: (Liking: 56 vs. 60 vs. 65; Negative emotions: 4.0 vs. 3.5 vs. 3.1; Positive emotions: 5.4 vs 6.0 vs. 6.8; Motivation to see University win: 5.7 vs. 4.6 vs. 3.7), all $ps < .05$.

4. Coherence: Correlations and Mediation

All the primary variables, whether cold (facts, likelihood, decision) or hot (liking, emotions, motivation), were strongly inter-correlated, rs = .57 -.76, $p < .01$, two tailed. These widespread correlations capture the essential core of the network that underlies constraint satisfaction processing.

Mediational analyses of the potential causal paths among the variables provided additional evidence to support the *coherence-based* mechanism. They were conducted with an SPSS macro by Preacher and Hayes (2004).

The first set of mediational analysis analyzed the relationship between the three primary variables—role assignment ("condition"), judgments of the case facts ("facts"), and the "likelihood" item ("likelihood that Debbie Miller did cheat on the exam"). Case facts were shown to be a significant mediator between assignment and likelihood, ($p < .001$). The assignment manipulation influenced the participant's perceptions of the case facts, which, in turn, influenced perceptions of guilt. A significant mediational effect was also observed in the reverse direction, with judgments of likelihood mediating the effect of assignment on the evaluations of the facts ($p < .001$). This is consistent with the bi-directional nature of *coherence-based reasoning,* in which all the elements in the network should mutually influence one other.

Another set of analyses examined whether participants' emotions and motivations mediated their "likelihood" judgments. Four Sobel tests were conducted, one for each mediator: facts, liking for Debbie, motivation (which side participant wanted to see win), and emotion. The effect of the assignment on the likelihood judgments was mediated significantly by each variable, all in the predicted directions. Similar meditation was observed when the "facts" were treated as the dependant variable.

To explore the relative strength of each mediator we conducted multiple mediational analysis. We included the four significant mediators (facts, liking, motivation, and emotion) simultaneously in the same analysis. The analysis revealed that two of the four remained significant, with the case "facts" being the strongest mediator ($z = 4.59$, $p < .001$), then "motivation" ($z = 4.00$, $p < .001$), while the emotion composite was marginal ($z = 1.79$, $p = .07$).

5. Perceived Objectivity – The Adversarial Mindset. The findings provide insight into the participants' metacognitive judgments. First, participants felt that their views of the case were equally objective in the adversarial conditions (7.9 and 8.0, on a 1 to 11 scale) as in the non-adversarial condition (7.9). They were unaware that the adversarial manipulation biased their judgments. Second, participants' in the two adversarial conditions had different views of their own and their adversary's objectivity. Participants deemed their adversary to be less objective, M=6.45, than they deemed themselves, M = 8.0, $t = 6.80$, $p < .001$. They also deemed him or her to be less trustful of themselves, M =

6.2, than they believed themselves to be, M = 8.0, t = 8.46, p < .001. Participants also believed that the other investigator's distrust was unwarranted, in that it was less credulous, M = 6.2, than the Chief Judicial Officer's evaluation of themselves, M = 7.2, t = 5.06, p < .001.

Discussion

The adversarial role strongly influenced people's perception of an ambiguous case. Relative to the non-adversarial assignment, adversarial role assignments skewed participants' views of the case in a self-serving manner. Participants in the condition that simulated police investigators were more likely to conclude that Debbie was culpable, whereas those simulating investigators for the defense were more prone to infer that she did nothing wrong. Most likely both conditions had a biasing influence on participants' judgments. Indeed, participants in the Sole Investigator condition viewed the case to be very close to the middle between the two adversarial conditions. The biasing impact of the adversarial assignment was manifested also by the arousal of mistrust towards their adversary.

Finally, the study provides the first experimental evidence of the interrelationship between hot and cold cognitions in *coherence-based reasoning.* More evidence for this relationship will be presented in Study 2.

Study 2: Partisanship and Coherence

Study 2 tested the effects of *strength* of *partisanship* on people's perceptions of a case and the role of motivation and emotion. We compared participants primed with a *non-partisan* manipulation with participants primed with a *partisan* one. We also examined whether the assignment would influence assessments of the trustworthiness of the witnesses. *Coherence-based reasoning* would lead to the prediction that judgments of the evidence would be positively related to judgments of the source's credibility.

We also sought to test *coherence shift* of beliefs. Study 2 introduced a pre-test instrument that tested participants' responses to the "belief" items, which were later included in the body of the study. This repeated-measures design enabled us to test within-subject shifts in the participants' responses to the belief items.

Method

Participants. The study used the same procedure as in Study 1. 163 individuals participated via the Internet. The sample was 48% female, with an average age of 46.

Procedure. We used the same case of Debbie Miller (with minor changes). The instructions described the adversarial hearing and the role of the University Representative ("University Rep"), which was substantively very similar to the role of a prosecutor, and role of the Student Advocate. All participants were assigned to the role of University Representative. After receiving the case, participants made a variety of judgments about it. All participants received the same case and instructions, except for information that was designed to manipulate the degree of partisanship.

Dependent Variables. Most of the variables were identical to those in Study 1. In addition, we measured participants' responses to the belief items on the pre-test and the judgments of the trustworthiness of the witnesses. To obtain a baseline measure for testing *coherence shifts*, participants received a pre-test questionnaire prior to the presentation of the case, containing questions probing their beliefs on a number of seemingly unrelated social issues. These questions were identical to the "belief" questions administered later on. Each of the belief items probed for a background belief that pertained to an ambiguous fact of the case (*e.g.*, "In general, people who have lower back pain tend to crouch when they sit for extended periods of time"). We predicted that responses to the belief items would shift from pre-test to post-test, ultimately cohering more strongly with the view of the case (see Simon, Snow, & Read, 2004).

Treatment. Participants were assigned to one of two conditions differing in their *partisanship*. Participants in the *non-partisanship* condition were told that for the most part they felt the process was fair. They were also provided with positive information about the Student Advocate assigned to represent Debbie Miller, Jim Cooper. He was said to be fair and professional and interested in the truth.

Participants in the *partisanship* condition were told that they had become frustrated by the number of students who had been cleared, despite being almost certainly guilty. They were upset about the impact of this on the University's reputation and the harm inflicted on students who did not cheat. Participants in this condition also received negative information about their adversary, Jim Cooper, being told that he was overzealous, strongly biased toward students, and responsible for many of the recent cases in which cheaters were cleared.

Results

The prediction was that participants in the *partisan condition* would be more inclined to believe that Debbie did cheat than participants in the *non-partisan condition* and that this would influence a range of different judgments.

1. Overall Judgments. The estimates of the likelihood that Debbie cheated were 40% in the *non-partisan* condition and 53% in the *partisan* condition, F (1, 161) = 6.93, p < .01. The assignment also influenced how participants would decide the case themselves, *non-partisan*: 33% Guilty vs. *partisan* : 49%, Chi-square (1) = 4.23, p = .04.

2. Case facts and related beliefs. Those in the *partisan* treatment perceived the case to be more consistent with the conclusion that Debbie cheated (Facts M = 5.6; Beliefs M = 5.4) than did participants in the *non-partisan* condition (Facts M = 4.8; Beliefs = 5.0), where higher numbers are more consistent with Debbie cheating, F (1, 160) = 9.65, p = .002 and F (1, 160) = 6.14, p = .014, respectively. The 13 "fact" items cohered to make a reliable composite (alpha = .86), as did the related "beliefs" (alpha = .61). Also partisanship affected the perceived trustworthiness of the witnesses. Both witnesses who testified that Debbie cheated were deemed more trustworthy by *partisan* participants

(Proctor: 6.8 v. 6.0; Brad Loomis: 6.0 v. 4.9), F (1, 160) = 4.96, p = .027 and F (1, 160) = 10.0, p = .001, respectively.

3. Judgments of Liking, Emotions, and Motivation. The partisanship manipulation also influenced participants' emotions and motivations. Compared with *non-partisan* participants, *partisan* participants liked Debbie less (M = 53 vs. 59), had stronger negative feelings (M = 2.7 vs. 2.1) and weaker positive feelings towards her (M = 3.3 vs. 4.0), and were more motivated to see the university prevail (M = 5.8 vs. 4.3), all differences p < .05.

4. Coherence Shifts of the Belief Items. Despite the initial ambiguity (as denoted by the non-significant differences at pre-test), by the time of the decision, the beliefs shifted to cohere with the decision and with one another, creating a strongly interconnected mental model (Holyoak & Simon, 1999; Simon et al, 2004a, Simon et al, 2004b). Figure 1 shows the coherence shifts in the belief items, plotting the data separately based on participants' response to the question: "if you were the Chief Judicial Officer, how would you decide the case"? (regardless of partisanship). A test of the interaction confirmed that these shifts were highly significant, F (2, 158) = 91.5, p = .000.

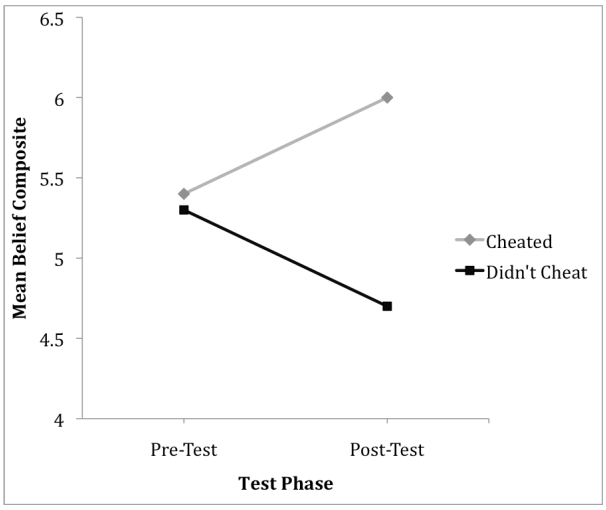


Figure 1. Coherence shifts in belief measures

5. Mediations and Structural Equation Modeling. We used SEM to perform simultaneous testing of the interrelationships among the study variables to identify which of the competing models best accounts for the relationships. The first analyses contained the four primary *cold cognitions*: partisanship assignment ("condition"), judgment of the case facts ("facts"), "likelihood" that Debbie Miller cheated on the exam, and the "decision" ("how would you decide the case?").

Two models (see Figure 2) show that partisanship predicts the primary variables. Model 1 shows that partisanship affected the judgment of the facts, which affected likelihood, which affected the decision. This is compatible with rational models of inference. Model 2 suggested that the inference chain may also run in reverse. These opposing models capture the bi-directionality of *coherence-based* reasoning; a central feature of the mutual influence in constraint satisfaction processes.

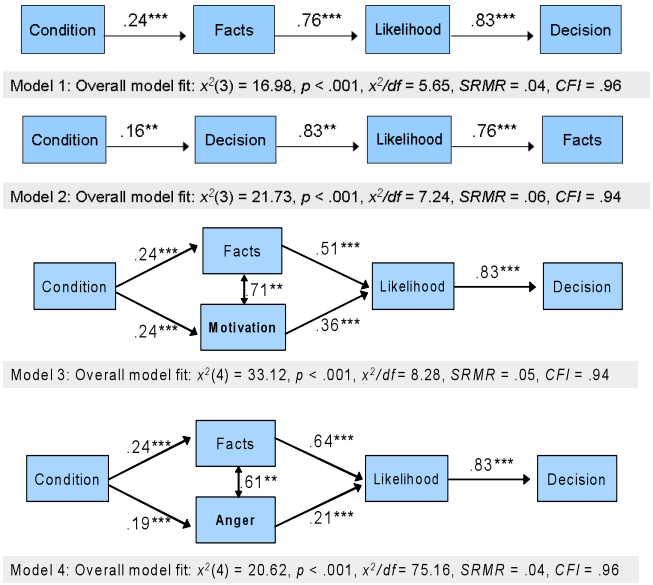


Figure 2: SEM models of hot and cold cognitions

Our primary question was whether hot cognitions are involved in the constraint satisfaction processes that drive the representation towards coherence. We first tested mediational relationships between the three hot cognitions (anger towards Debbie, motivation, and liking) and a central cold cognition: decision, (See Figure 3). A simultaneous mediational analysis between the condition and the decision revealed effects for "anger" and "motivation", but not for "liking". Mediation by hot and cold cognitions was also observed using SEM (see bottom of Figure 2), which found good fitting models for both "motivation" and "anger" as joint mediators, with "facts," of "likelihood", and decision.

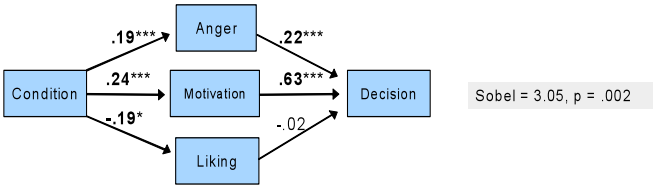


Figure 3: Multiple mediation model

Discussion

Judgments were influenced considerably by the intensity of partisanship. As in Study 1, the treatment assignment resulted in coherent mental models of the case, in which the wide range of variables involved in the judgment all cohered with the manipulated conclusion. Participants' assessments of the trustworthiness of the witnesses were also influenced by the assignment. Partisan participants were more likely to trust the witnesses who claimed to have seen Debbie cheat. Most important, we observed that hot cognitions mediated the effect of the assignment on the cold cognitions.

General Discussion

The studies show that the perception of a factually ambiguous case depends on the conditions under which the judgments are made. Study 1 simulated a police investigation and found that the perception of the case was strongly influenced by the participants' role assignment. Relative to the non-adversarial assignment, adversarial role assignments skewed participants' *hot* and *cold* cognitions in a manner that supported their assigned side. The non-adversarial assignment led to judgments close to the midpoint between the two adversarial conditions. The symmetry of the polarization supports the conclusion that adversarialism results in a distorted perception of the case. Participants in all conditions deemed their perception of the case to be equally objective, suggesting that the participants in the adversarial conditions were unaware of the influence of the assignment on their judgments. Study 2 simulated a prosecutorial view of the same case and found that both hot and cold cognitive judgments are influenced considerably by the intensity of the partisanship.

These studies provide further corroboration for the *coherence based reasoning* framework (Holyoak & Simon, 1999; Simon, 2004; Simon et al., 2004a; Simon et al., 2004b). We found again that participants' views of a complex task tend to cluster into large and coherent mental representations that encompass the overall judgments of the case as well as of the entire set of facts and related beliefs.

However, the most important contribution is the novel finding of the interrelationship between the hot and cold cognitive aspects of the task. While a great deal of research has observed the effect of emotion and motivation on cognitive processing (*e.g.*, Kunda 1990; Slovic, Finucane, Peters, & Macgregor, 2002; Zajonc, 1980), that research has not provided much insight into the mechanisms by which these effects occur. Mediational analyses and SEM revealed that emotion, motivation and to some degree also liking, mediated the effect of the assignment on the various cold cognitive judgments of the case, while similar mediations were observed in the reverse direction. While one ought to be cautious drawing causal conclusions from these data, these observations are strongly consistent with the Gestaltian features of high interconnectivity and bi-directional influence that characterize constraint satisfaction processing and *coherence-based reasoning.*

References

Balcetis, E., & Dunning, D. (2006). See what you want to see: Motivational influences on visual perception. *Journal of Personality and Social Psychology*, *91*, 612-625.

Ditto, P. H., & Lopez, D. F. (1992). Motivated skepticism: Use of differential decision criteria for preferred and nonpreferred conclusions. *Journal of Personality and Social Psychology, 63,* 568-584.

Ferguson, T. J., & Rule, B. G. (1983). An attributional perspective on anger and aggression. In R. G. Geen & E. I. Donnerstein (Eds.), *Aggression: Theoretical and empirical reviews,* Vol. 1. New York: Academic Press.

Goldberg, J. H., Lerner, J. S., & Tetlock, P. E. (1999). Rage and reason: The psychology of the intuitive prosecutor. *European Journal of Social Psychology*, *29*, 781-795.

Heider, F. (1958). *The psychology of interpersonal relations*. New York: Wiley.

Holyoak, K. J., & Simon, D. (1999). Bidirectional reasoning in decision making by constraint satisfaction. *Journal of Experimental Psychology: General, 128*, 3-31.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin, 108*, 480-498.

Lerner, J. S., Goldberg, J. H., & Tetlock, P. E. (1998). Sober second thought: The effects of accountability, anger, and authoritarianism on attributions of responsibility. *Personality and Social Psychology Bulletin*, *24*, 563-574.

Nerb, J. (2007). Exploring the dynamics of the appraisal-emotion relationship: A constraint satisfaction model of the appraisal process. *Cognition & Emotion*, *21*, 1382-1413.

Piercey, M. D. (2009). Motivated reasoning and verbal vs. numerical probability assessment: Evidence from an accounting context. *Organizational Behavior and Human Decision Processes*, *108*, 330-341.

Preacher, K. J. & Hayes, A. F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods, Instruments, & Computers, 36,* 717-731.

Quigley, B. M., & Tedeschi, J. T. (1996). Mediating effects of blame attributions on feelings of anger. *Personality and Social Psychology Bulletin*, *22*, 1280-1288.

Rick, S., & Loewenstein, G. (2008). The role of emotion in economic behavior. In Lewis, M. (Ed); Haviland-Jones, Jeannette M. (Ed); Barrett, Lisa Feldman (Ed), *Handbook of emotions (3rd ed.)*. New York, NY, US: Guilford Press.

Simon, D. (2004). A third view of the black box: Cognitive coherence in legal decision making. *University of Chicago Law Review, 71*, 511-586.

Simon, D., Krawczyk, D. C., & Holyoak, K. J. (2004a). Construction of preferences by constraint satisfaction. *Psychological Science, 15*, 331-336.

Simon, D., Snow, C., J., & Read, S. J. (2004b). The redux of cognitive consistency theories: Evidence judgments by constraint satisfaction. *Journal of Personality and Social Psychology, 86,* 814-837

Slovic, P., Finucane, M., Peters, E., & MacGregor, D. G. (2002). The affect heuristic. In Gilovich, Thomas (Ed); Griffin, Dale (Ed); Kahneman, Daniel (Ed), *Heuristics and biases: The psychology of intuitive judgment*. New York, NY: Cambridge University Press.

Thagard, P. (2006). *Hot thought: Mechanisms and applications of emotional cognition.* Cambridge, MA: MIT Press.

Zajonc, R. B. (1980). Feeling and thinking: Preferences need no inferences. *American Psychologist, 35*, 151-175.

Comparison-Induced Sequence Effects on Hedonic Evaluations

Jessica M. Choplin (jchoplin@depaul.edu) & Megan M. Lombardi (mlombar7@depaul.edu)
DePaul University Department of Psychology
2219 North Kenmore Avenue
Chicago, IL 60614-3504

Abstract

Hedonic evaluations and emotional reactions to experiences depend not only upon the conditions being experienced, but also upon the sequences in which conditions are experienced. The authors propose a comparison-induced distortion (CID) model of sequence effects on evaluation in which to-be-evaluated exemplars are verbally compared (Choplin & Hummel, 2002; Choplin, 2007) to the most similar, recent exemplars. Predictions of this model were tested and pit against Helson's (1964) adaptation-level theory, Parducci's (1995) range-frequency theory, and Haubensak's (1992) consistency model using a paradigm in which sequences periodically improved (i.e., improved for *n* trials, returned to the original state on a single trial, and improved for *n* trials again) or periodically deteriorated by small or large amounts. The results were consistent with the predictions of the proposed CID model of sequence effects and inconsistent with adaptation-level theory, range-frequency theory, and the consistency model.

Sequence Effects

Most theories of the causes of emotions posit that people's emotional reactions to the conditions they experience (e.g., prices, salaries, pain, tastes, wait times, and so forth) depend in part upon their hedonic evaluations of those conditions—how good or bad they judge those conditions to be (e.g., Kahneman, 1999; Lazarus, 1991; Tesser & Martin, 1996). While some researchers conceptualize these evaluations as measurements on a single good/bad dimension (Kahneman, 1999) and other researchers conceptualize these evaluations as separate measurements of how good conditions are and how bad conditions are (Cacioppo, Gardner, & Berntson, 1999; Watson & Tellegen, 1985), the general idea underlying this common notion is that these evaluations play a central role in the emotions people experience. If people judge conditions to be good, then their emotional reactions will generally be positive. That is, the evaluation that a condition is good will evoke emotions such as happiness, delight, and relief. If people judge conditions to be bad, then their emotional reactions will generally be negative. That is, the evaluation that a condition is bad will evoke emotions such as anger, frustration, and worry.

A challenge for research on emotion and hedonic evaluation arises from the fact that evaluations are not a pure function of the objective conditions being evaluated. Rather, hedonic evaluations depend upon the context in which conditions are experienced (see, for example, Parducci, 1995). Predicting people's emotional reactions, therefore, requires an understanding of the contextual factors that affect evaluations. The research reported here investigated one type of context effect on evaluation, namely, the effect of the sequence in which conditions are experienced. Paying $2.85 per gallon of gasoline, for example, might seem more reasonable if recent prices have been over $2.85 than if recent prices have been under $2.85. Waiting 5 minutes for a bus might seem more reasonable if one has lately been waiting more than 5 minutes than if one has been waiting less. The purpose of the research reported here was to develop and test an account of how the sequences in which conditions are experienced affect evaluations. We will propose a comparison-induced distortion (CID) account of sequence effects on evaluation and pit this account against Helson's (1964) adaptation-level theory, Parducci's (1995) range-frequency theory, and Haubensak's (1992) consistency model.

Comparison-Induced Distortions

The basic idea underlying the CID model of sequence effects is that people will verbally compare to-be-evaluated items to the most similar items they have recently encountered. The model selects previous exemplars to be compared to the to-be-evaluated item using two criteria: 1) giving more weight to exemplars that are more similar to the to-be-evaluated item and 2) giving more weight to more recent items (the 1-back item is weighted higher than the 2-back item, the 2-back item higher than the 3-back item, etc.). Although the proposal that these two criteria determine recall and comparison processes has not previously been applied to sequence effects on hedonic evaluations, it has been applied to other domains in which recall of previously presented exemplars affects judgment (see, for example, Nosofsky & Palmeri, 1997; Smith & Zarate, 1992).

After a comparison item is selected, we hypothesize that the to-be-evaluated item is verbally (often sub-vocally, not out loud) compared to the comparison item. CID theory predicts that verbally comparing items causes people to exaggerate small differences and under appreciate the size of large differences. The reason for this pattern of evaluation is that evaluations are biased toward the central tendency (i.e., mean or median) of values associated with a comparison word (Huttenlocher, Hedges, & Vevea, 2000).

For example, a longer wait for a bus could be 1 extra minute or 30 extra minutes, but when we consulted city bus schedules we found that the most common wait time between buses was 10 minutes. That is, there is a distribution of differences in wait times between busses where the central tendency of the distribution of "longer wait times" is around 10 extra minutes. Huttenlocher et al. (2000) demonstrated that judgments of values are typically

biased towards the central tendency of categories. If so, then a "longer wait time" of 5 extra minutes (i.e., less than the central tendency of the category of "longer wait times" which is 10 extra minutes) would be biased towards the evaluation of 10 extra minutes. That is, the difference would be exaggerated. Likewise, a "longer wait time" of 15 extra minutes (i.e., more than the central tendency of 10 extra minutes) would also be biased towards the evaluation of 10 extra minutes. This time, however, the bias would cause the size of the difference to be under appreciated.

The experiment described below was designed to test the predictions of this model. To do so, participants were asked to rate their aversion to several fictional wait times for a bus in the winter. Participants evaluated wait time sequences that improved or deteriorated in large or small increments, returned to a value near the original value, and then improved or deteriorated again. CID theory predicts that there will be an amount of change (small or large) by direction of change (improving or deteriorating) interaction effect. Specifically, wait times in the small-increment deteriorating sequence will be rated as worse than wait times within the small-increment improving sequence, because these small differences will be exaggerated. A little worse will seem like a lot worse and a little better will seem like a lot better. Conversely, wait times in the large-increment deteriorating sequence will be rated better than items in the large-increment improving sequence, because people will under appreciate the sizes of the differences. A lot worse will seem as if it is only a little worse and a lot better will seem as if it is only a little better.

Adaptation-Level Theory

Models of sequence effects on evaluation commonly start with Helson's (1964) proposal that evaluations are made relative the conditions to which people have adapted—that is, the conditions to which they have become accustomed, consider normal, and continue to expect (see Briesch, Krishnamurthi, Mazumdar, & Raj, 1997; Frederick & Loewenstein, 1999; Kalyanaram & Winer, 1995, for reviews). Helson (1964) modeled the conditions to which people adapted as the running average of all previously experienced conditions (see also Kalwani, Yim, Rinne, & Sugita, 1990; Rajendran & Tellis, 1994; Wedell, 1995). These running averages then serve as reference points against which all other conditions are evaluated. Models that appeal to this explanation of sequence effects on evaluation assume that sequence effects occur when the conditions people consider normal change as they experience more instances. If people experience additional favorable conditions, they will start to consider these favorable conditions to be normal. If people experience additional unfavorable conditions, they will start to consider unfavorable conditions normal. This change in what is considered normal, thereby, causes sequence effects wherein the same conditions might be evaluated as better (or worse) depending upon whether the previously experienced values were better or worse.

Contrary to CID theory, AL theory predicts that items in deteriorating sequences will always be judged worse than items in improving sequences regardless of the size of the difference between items. Since items are evaluated in comparison to the adaptation level—which is the average of all previous items—the same item (e.g., 36 minutes) will be evaluated differently based on the average of the items that precede it in the sequence. For a deteriorating sequence, the to-be-evaluated wait time will be worse than what people have gotten used to, the average that they consider normal. For an improving sequence, the to-be-evaluated wait times will be better than what they are used to and consider normal. This pattern would be true regardless of the sizes of the differences between wait times. That is, adaptation-level theory predicts that wait times will be evaluated as worse in deteriorating than in improving conditions.

Range-Frequency Theory

One of the most important models of hedonic evaluation is Parducci's (1995) range-frequency theory. Range-frequency theory is based on the idea that judgments are made based on a compromise between range and frequency principles. According to the range principle, individuals evaluate items relative to the smallest and largest values that they have previously encountered. The individual's evaluation is based on a calculation of the range value for the to-be-evaluated item, which is the proportion of the range at which the to-be-evaluated item is located relative to the smallest and largest values. The midpoint between the highest and lowest values would be 50% of the way to the largest value from the smallest; half way between the smallest value and the midpoint would be 25%; and half way between the midpoint and the largest value would be 75%. According to the frequency principle, individuals evaluate to-be-evaluated items by calculating their percentile rank among all of the items that they have seen. Individuals compromise between these two principles when making evaluations.

Unlike comparison-induced distortion theory, range-frequency theory predicts that there will be no effect of the amount of change between wait times as long as there are no changes in the range or frequency values of the to-be-evaluated wait times. The experiment described below controls for this issue by keeping range and frequency values constant across the amount of change manipulation. Furthermore, the frequency values (percentile ranks) of the to-be-evaluated wait times would be larger (worse) in the deteriorating sequence than in the improving sequence, because better previous exemplars would be included in the context of judgment. Like adaptation-level theory, then, range-frequency theory predicts that wait times will be evaluated as worse in deteriorating than in improving conditions.

Consistency Model

Similar to comparison-induced distortion theory, Haubensak's (1992) consistency model of evaluation relies on the basic assumption that recalled exemplars affect judgment. According to the consistency model, people strive for internal consistency in their responses since judgments are subjective in that the mapping between the real-world dimension and the category-rating dimension is

arbitrary. In attempting to maintain internal consistency in their responses, people often constrain their responses based on the first few exemplars they encounter. By making judgments about preceding stimuli in a sequence, for example, the person is confining subsequent judgments to a specific response scale. If the person evaluates the first few items in a sequence, doing so commits them to giving subsequent judgments that are consistent with the previous judgments. Evaluations that are consistent with those previous judgments can be calculated by linearly interpolating from previous evaluations.

To control for the effects of early judgments and the requirement that participants' evaluations be consistent with these judgments, an initial sequence was held constant across the direction of change manipulation for the current study. If participants linearly interpolate from the initial judgments they make, then evaluations should be the same for the deteriorating and improving conditions. To control for memory effects, a high and a low value were always present within the previous five trials (less than the seven represented in Haubensak's model) and these high and low values were constant across the direction of change conditions. Since this model predicts that the initial sequence will be the basis for wait time evaluations, this model predicts no effects of the amount of change between exemplars.

Experiment

To pit the predictions of the CID model of sequence effects against the predictions of the other models, we used a paradigm in which participants imagined that they had to wait for the bus in a rural town in northern Minnesota on each of 36 fictional winter days (manipulated within a single session). Wait times either periodically improved (i.e., times became successively shorter on each of *n* trials, returned to a value near the original state on a single trial, and then became successively shorter on each of *n* trials again) or periodically deteriorated (i.e., times became successively longer on each of *n* trials, returned to a value near the original state on a single trial, and then became successively longer on each of *n* trials again). Participants rated how aversive each wait time would be.

This sequence is effective in pitting the CID model against the other models. The CID model predicts that the size of the difference between consecutive exemplars will matter such that exemplars within periodically deteriorating series will be rated as worse than exemplars within periodically improving series when there is a small amount of change between consecutive exemplars. Additionally, exemplars within periodically deteriorating series will be rated as better than exemplars within periodically improving series when there is a large amount of difference between consecutive exemplars. Adaptation-level and range-frequency theories, by contrast, predict that exemplars within periodically deteriorating series will always be judged worse than exemplars within periodically improving series. The consistency model predicts no effects of the amount of change manipulation.

Method

Participants. An experimenter, who was blind to the hypotheses, approached individual prospective participants on a university campus or in the surrounding community. Two hundred and five people volunteered after being approached in this manner. Approximately half of the participants (n = 101) experienced wait times that changed (improved or deteriorated) by small amounts (i.e., 5 minutes) on each trial, excluding periodic large changes. Of these, 50 participants were in the periodically improving condition and 51 were in the periodically deteriorating condition. The other half of the participants (n = 104) experienced wait times that changed (improved or deteriorated) by large amounts (i.e., 15 minutes) on each trial, excluding periodic larger changes. Of these, 54 participants were in the periodically improving condition and 50 were in the periodically deteriorating condition.

Materials and Procedure. Participants imagined that they were spending 36 days in northern Minnesota during the middle of the winter and had to rely upon an erratic bus for transportation. The amount of time they spent waiting for the bus each day was presented aloud and participants rated how aversive that wait time for each day would be on a scale from 0 ("not bad") to 10 ("extremely bad").

CID hypothesized that participants would overreact to differences smaller than 10 minutes (i.e., 5 minutes) and under-react to differences larger than 10 minutes (i.e., 15 minutes), because the median of values from the category of "longer wait times" for busses was 10 minutes in local bus schedules for the campus community. Furthermore, the results of the experiment (presented shortly) suggest that participants did overreact to 5-minute differences and under-react to 15-minute differences. Sequences of presented values were constructed by dividing the middle 26 days of the experiment into two 13-day periods. Within each 13-day period, periodically improving and deteriorating sequences like those in Table 1 were presented. The order of the three series shown in Table 1 (i.e., Series A, B, and C) was fully counterbalanced to produce six counterbalanced groups for each of the four—2 (amount of change: 5 minutes or 15 minutes) x 2 (direction of change: improving or deteriorating)—conditions. The sequence of wait times in the second 13-day period was identical to the sequence in the first 13-day period.

To control for primacy effects and introduce participants to the range of values they would see prior to the sequence manipulation, a sequence of 5 days was inserted at the beginning of the experiment. The sequence on these 5 days was 22, 35, 50, 35, and 22 minutes respectively in the small-difference condition and 2, 35, 70, 35, 2 minutes respectively in the large-difference condition. To make peak and end values equivalent across periodically improving and deteriorating sequences before asking participants to make a retrospective evaluation (Kahneman, Frederickson, Schreiber, & Redelmeier, 1993; Redelmeier & Kahneman, 1996), a sequence of 5 days was added to the end of the

Table 1. Wait Time Sequences (all values are in minutes)

Initial Sequences:	
For Participants Experiencing Small Changes	22, 35, 50, 35, 22
For Participants Experiencing Large Changes	2, 35, 70, 35, 2

Manipulated Sequences:	Series A	Series B	Series C
Deteriorating in Small Increments	26, 31, 36, 41, 46	27, 32, 37, 42	29, 34, 39, 44
Improving in Small Increments	46, 41, 36, 31, 26	42, 37, 32, 27	44, 39, 34, 29
Deteriorating in Large Increments	6, 21, 36, 51, 66	11, 26, 41, 56	16, 31, 46, 61
Improving in Large Increments	66, 51, 36, 21, 6	56, 41, 26, 11	61, 46, 31, 16

Final Sequence:	
For All Participants	35, 35, 35, 35, 35

experiment. The wait time for each of these 5 days was 35 minutes.

After evaluating wait times for all 36 days, participants retrospectively evaluated all of the wait times they had seen in the experiment on a scale from -50 (not bad at all) to +50 (extremely bad).

Results

We first analyzed the ratings participants gave for each of the 36 days as they went through the experiment. To reduce variance caused by idiosyncratic reactions to wait times, participants' judgments during the initial 5-day sequence were used as a baseline. Each participant's judgments on trials 6 through 31 were divided by the average of her or his average judgment on days 1 and 5, days 2 and 4, and her or his judgment on day 3. The results are presented in Figure 1.

As shown in Figure 1, of the participants in the small change condition, those who experienced periodically deteriorating sequences rated wait times more aversive than did those who experienced periodically improving sequences for 11 of the 13 wait times. This proportion (.85) was significantly greater than .50, $\chi^2(1, N = 13) = 4.92$, $p < .05$. This effect was very weak, however. In fact, a 2 (direction of change: improving or deteriorating) x 2 (portion of sequence: days 6-18 and days 19-31) x 13(wait times) Mixed-Factors Analysis of Variance (ANOVA) on the evaluations of participants in the small-difference condition failed to find a difference due to the direction of change, $F(1,99) = 0.04$, $MSE = 28.61$, $p > .05$.

The participants in the large change condition showed a very different pattern of evaluations. Of these participants, those who experienced periodically improving sequences rated wait times more aversive than did those who experienced periodically deteriorating sequences for all 13 of the 13 wait times. This proportion (1.00) was significantly greater than .50, $\chi^2(1, N = 13) = 4.92$, $p < .01$. A 2 (direction of change: improving or deteriorating) x 2 (portion of sequence: days 6-18 and days 19-31) x 13 (wait times) Mixed-Factors ANOVA on the evaluations of participants in the large-difference condition also found a main effect of the direction of change, $F(1,102) = 8.92$, $MSE = 24.07$, $p < .05$.

The interaction between amount of change and direction of change was also significant as revealed by an omnibus 2 (amount of change: small or large) x 2 (direction of change: improving or deteriorating) x 2 (portion of sequence: days 6-18 and days 19-31) x 13 (wait times) Mixed-Factors ANOVA, $F(1,201) = 8.25$, $MSE = 26.3$, $p < .05$. This finding is consistent with the predictions of the CID model presented above and inconsistent with the predictions of adaptation-level theory (Helson, 1964) and range-frequency theory (Parducci, 1995) and not predicted by the consistency model (Haubensak, 1992). Post hoc least significant difference analyses found that participants who experienced periodically improving large differences rated wait times more aversive than did the other three groups. Evaluations on days 19-31 did not significantly differ from evaluations on days 6-18, $F(1,102) = 0.11$, $MSE = 2.32$, $p > .05$.

A 2 (amount of change: small or large) x 2 (direction of change: improving or deteriorating) Between-Subject ANOVA on participants' retrospective evaluations showed no main effect of the amount of change [$F(1,201) = 3.18$, $MSE = 621.1$, $p > .05$], no main effect of the direction of change [$F(1,201) = 0.46$, $MSE = 621.1$, $p > .05$], and no interaction between them [$F(1,201) = 0.87$, $MSE = 621.1$, $p > .05$] suggesting that the algorithms responsible for retrospective evaluations might be different from the algorithms responsible for online evaluations. This result also suggests that the finding that people prefer improving to deteriorating sequences (Hsee & Abelson, 1991; Hsee et al., 1991; Schifferstein & Frijters, 1992; Varey & Kahneman, 1992) might not generalize to online evaluations of periodically improving and deteriorating sequences such as the sequences investigated here.

Discussion

Theories of sequence effects on hedonic evaluation were assessed using a paradigm in which values periodically

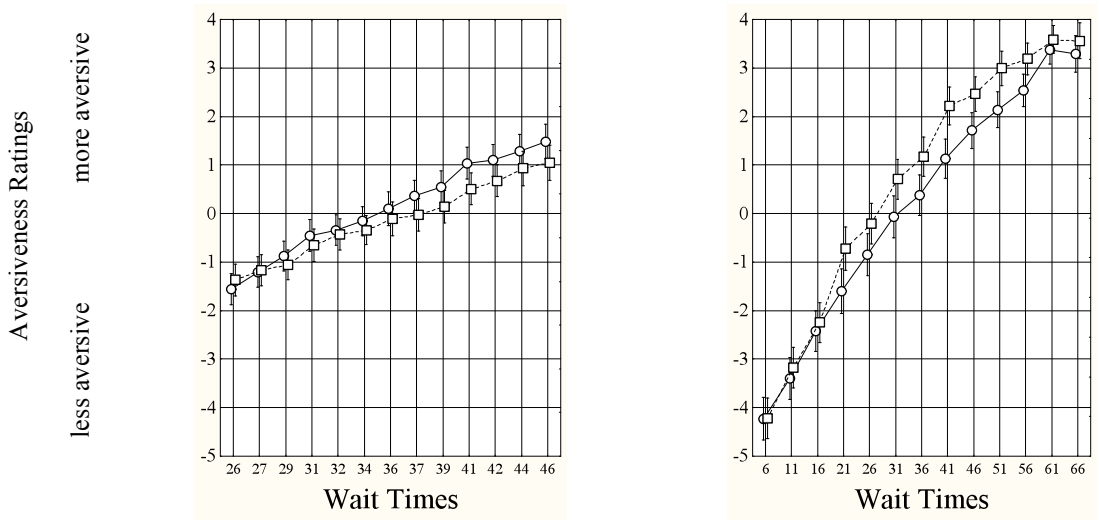


Figure 1. Results

improved or deteriorated by small or large amounts. When values changed by small amounts, participants evaluated periodically improving sequences more positively than periodically deteriorating sequences; but when values changed by large amounts, participants evaluated periodically deteriorating sequences more positively than periodically improving sequences.

Helson's (1964) adaptation-level theory cannot explain the finding that evaluations differed based on the size of the difference between exemplars since AL theory makes the prediction that items in deteriorating sequences will always be judged worse than items in improving sequences regardless of the size of the difference between items. Since wait times in the large change deteriorating sequence were preferred over wait times in the large change improving sequence it seems as though participants' evaluations were not based on a comparison of the to-be-evaluated wait time and the average of the preceding wait times, which would result in the opposite finding.

Since the range and frequency values of each wait time in the small change condition mapped onto the range and frequency values of each wait time in the large change condition, range-frequency theory (Niedrich, Sharma, & Wedell, 2001; Parducci, 1995) cannot explain the finding that the amount of change between items influenced wait time evaluations. Furthermore, range-frequency theory cannot explain the finding that wait times were evaluated as better in the large change deteriorating condition than in the large change improving condition; the frequency principle predicts that wait times will be worse in the deteriorating condition than in the improving condition since there would be more positive exemplars in the context of judgment. Based on the current findings, it seems as though range and frequency principles were not used to make evaluations of wait time since these principles would produce the opposite pattern of evaluations.

Haubensak's (1992) consistency model of judgment also cannot explain the finding that the amount of change between wait times influenced evaluations. The consistency model predicts that evaluations are based on linear interpolations from initial items in a sequence. Because linearly interpolating between the evaluations made during the initial sequence would have made evaluations in the periodically improving and deteriorating sequences identical, the finding that there were differences in the evaluated wait times in these sequences suggests that participants did not base their evaluations off of the initial sequence of wait times that was presented.

Of the four evaluation models presented in this paper, only the CID model proposed above was able to predict and explain the observed results. The finding that participants preferred periodically improving to periodically deteriorating sequences when wait times changed by small amounts is consistent with the CID prediction that differences will be exaggerated toward the central tendency of values that have been associated with a comparison word (Huttenlocher et al., 2000). Similarly, the finding that participants preferred periodically deteriorating to periodically improving sequences when wait times changed by large amounts is consistent with the CID prediction that differences will be under appreciated when the central tendency of values associated with a comparison word is smaller than the amount of change.

The results reported here have implications for how managers, price strategists, administrators, politicians, and other bearers of good and bad news ought to present news to others. Consistent with research on hedonic editing (Thaler & Johnson, 1990), if circumstances (e.g., prices, salaries, service quality, and so forth) are going to become better,

perhaps it is best to present the good news a little bit at a time. People will appreciate the good news; and they will be likely to exhibit positive emotions such as happiness, delight, and relief each time that good news is presented. If circumstances were going to become worse, however, perhaps it would be best to present all of the bad news at once. People might not realize how bad circumstances have actually gotten; and while they will likely exhibit negative emotions such as anger, frustration, and worry, the sum total of these negative emotions might be less than if the bad news were presented a little bit at a time. The CID model builds on this previous research by offering guidance on the size of the changes that will be underappreciated or exaggerated. Changes that are larger than the central tendency of the distribution of previously observed changes will be underappreciated; and changes that are smaller than this central tendency will be exaggerated.

References

Briesch, R. A., Krishnamurthi, L., Mazumdar, T., & Raj, S. P. (1997). A comparative analysis of reference price models. *Journal of Consumer Research, 24*, 202-214.

Cacioppo, J. T., Gardner, W. L., & Berntson, G. G. (1999). The affect system has parallel and integrative processing components: Form follows function. *Journal of Personality and Social Psychology, 76*, 839-855.

Choplin, J. M., & Hummel, J. E. (2002). Magnitude comparisons distort mental representations of magnitude. *Journal of Experimental Psychology: General, 131*(2), 270-286.

Choplin, J.M. (2007). Toward a comparison-induced distortion theory of judgment and decision making. In J.A. Elsworth (Ed.), *Psychology of decision making in education, behavior and high risk situations* (pp. 11-40). Hauppauge, NY: Nova Science.

Frederick, S., & Loewenstein, G. (1999). Hedonic Adaptation. In D. Kahneman, E. Diener & N. Schwarz (Eds.), *Well-being: The foundations of hedonic psychology* (pp. 302-329). New York: NY: Russell Sage.

Haubensak, G. (1992). The consistency model: A process model for absolute judgments. *Journal of Experimental Psychology: Human Perception and Performance, 18*, 303-309.

Helson, H. (1964). *Adaptation-level theory*. New York: Harper & Row.

Hsee, C. K., & Abelson, R. P. (1991). Velocity relation: Satisfaction as a function of the first derivative of outcome over time. *Journal of Personality and Social Psychology, 60*, 341-347.

Hsee, C. K., Abelson, R. P., & Salovey, P. (1991). The relative weighting of position and velocity in satisfaction. *Psychological Science, 2*, 263-266.

Kahneman, D. (1999). Objective happiness. In D. Kahneman, E. Diener & N. Schwarz (Eds.), *Well-being: The foundations of hedonic psychology*. New York, NY: Russell Sage.

Kahneman, D., Frederickson, B. L., Schreiber, C. A., & Redelmeier, D. A. (1993). When more pain is preferred to less: Adding a better end. *Psychological Science, 4*, 401-405.

Kalwani, M. U., Yim, C. K., Rinne, H. J., & Sugita, Y. (1990). A price expectations model of customer brand choice. *Journal of Marketing Research, 27*, 251-262.

Kalyanaram, G., & Winer, R. S. (1995). Empirical Generalizations from reference price research. *Marketing Science, 14*, G161-G169.

Lazarus, R. S. (1991). *Emotion and adaptation*. New York: Oxford University Press.

Niedrich, R. W., Sharma, S., & Wedell, D. H. (2001). Reference price and price perceptions: A comparison of alternative models. *Journal of Consumer Research, 28*(3), 339-354.

Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review, 104*, 266-300.

Parducci, A. (1995). *Happiness, pleasure and judgment: The contextual theory and its applications*. Mahwah, NJ: Lawrence Erlbaum.

Rajendran, K. N., & Tellis, G. J. (1994). Contextual and temporal components of reference price. *Journal of Marketing, 58*, 22-34.

Redelmeier, D. A., & Kahneman, D. (1996). Patients' memories of painful medical treatments: Real-time and retrospective evaluations of two minimally invasive procedures. *Pain, 66*, 3-8.

Schifferstein, H. N. J., & Frijters, J. E. R. (1992). Contextual and sequential effects on judgments of sweetness intensity. *Perception & Psychophysics, 52*, 243-255.

Smith, E. R., & Zarate, M. A. (1992). Exemplar-based model of social judgment. *Psychological Review, 99*, 3-21.

Tesser, A., & Martin, L. (1996). The psychology of evaluation. In E. T. Higgins & A. W. Kruglanski (Eds.), *Social psychology: Handbook of basic principles* (pp. 400-432). New York: Guilford.

Thaler, R.H. and Johnson, E.J. (1990). Gambling with the house money and trying to break even: The effects of prior outcomes on risky choice. *Management Science, 36*, 643-660.

Varey, C., & Kahneman, D. (1992). Experiences extended across time: Evaluation of moments and episodes. *Journal of Behavioral Decision Making, 5*, 169-185.

Watson, D., & Tellegen, A. (1985). Toward a consensual structure of mood. *Psychological Bulletin, 98*, 219-235.

Wedell, D. H. (1995). Contrast effects in paired comparisons: Evidence for both stimulus-based and response-based processes. *Journal of Experimental Psychology: Human Perception and Performance, 21*, 1158-1173.

The Effect of Cognitive Load and Meaning on Selective Attention

Rebecca A. Weast (rweast2787@gmail.com)
Department of Psychology, Franklin & Marshall College
Lancaster, PA 17604 USA

Nicole G. Neiman (nicole.neiman@fandm.edu)
Department of Psychology, Franklin & Marshall College
Lancaster, PA 17604 USA

Abstract

Nillie Lavie's Load Theory of selective attention suggests that the size of the cognitive load affects selective attention ability: the larger the cognitive load, the poorer the selective attention performance. Other authors have found that the relationship between distracting and relevant information can influence how well distractors are ignored. Our study hypothesized that a) larger cognitive load would (as previously shown) hinder reaction time on a selective attention task, b) that distractors (words) semantically related to the words being held in memory (as part of the cognitive load manipulation) would be more distracting than unrelated and neutral distracters. The findings instead showed that unrelated distractors were more distracting.

Lavie's Load Theory of selective attention suggests that the quantity of stimuli presented to a person determines how their selective attention system will function – whether they will be more or less distractible (Lavie, Hirst, de Fockert, & Viding, 2004). The Load Theory suggests that selective attention consists of two mechanisms: a passive, perceptual system, and an active mechanism of cognitive control. The perceptual system functions in line with the early selection model, where the number of stimuli modulates the effectiveness of attentional selection. When there is a low load on the perceptual system – as in a visual search task with few (1-3) items to search through – there is extra, unused perceptual capacity that involuntarily picks up other (irrelevant, distracting) environmental information, and the person is more likely to perceive distractors. In a high perceptual load condition, the opposite is true: the task uses up all attentional capacity, and extra environmental elements can't interfere. In this model, the second stage of selective attention is an *active*, cognitive process. In conditions of high cognitive load, where most of the person's cognitive capacity is consumed with a difficult working memory task, for example, the person has few cognitive resources available to resist distraction by irrelevant information. The person can better disregard distractors in a low-cognitive load condition. When the working memory (or other task) contains fewer items, there is more cognitive capacity available to focus on relevant information, while effectively weeding our perceived distractors. This cognitive system only comes into play in conditions of low perceptual load, when distractors have been perceived and need to be actively suppressed (Lavie, Hirst, de Fockert & Viding 2004; Lavie & Cox, 1997; Lavie, 1995).

Many studies have addressed the intricacies of this model, usually using simple, single letter or digit stimuli. Past studies used stimuli that are unrelated to each other and that carry as little semantic information as possible. Presumably, they do this to get at the attention issue in its purest form, with the simplest stimuli possible. Stimulus material used in past studies includes letters, numbers, colors, and simple black and white symbols. Few have examined the model in relation to the semantic, content of the stimuli. This raises the obvious questions: are distractors more distracting when they are meaningfully related to task-relevant information? Is this effect modulated by cognitive load?

Most of the work done to test Load Theory has focused on the perceptual mechanism of selective attention. Lavie and colleagues have investigated the effect of cross-modal distractor presentation in a decision task, and the possible relationship between the cognitive load resulting from task-switching between and within sensory modalities (Rees, Frith & Lavie, 2001; Brand-D'abrescia & Lavie 2008). Only Lavie's 2004 study has really addressed the effect of cognitive load on selective attention tasks. This makes sense; the question of involuntary attention grabbing by stimuli is a more direct way to study selective attention, and according to the Lavie model, the cognitive system plays only a supporting role in attention control. However, we wanted to investigate the effect of cognitive load further. The current study investigated the impact of the cognitive load on selective attention. More specifically, we examined what effect, if any, the semantic content of the information being held in working memory has on distractibility when

distractors are related to the information in working memory. Lavie et al.'s 2004 study used digits to compose their memory sets, the current study used words. By using words instead of digits, the memory set words could be related to each other, and could allow for distractors to also be related (or not) to the words in the memory set. By manipulating the meaningful relationships between memory set and distractor we hoped to have an effect on distractor interference.

One study has examined the effects of distractors with semantic meaning (words) when that meaning is either task congruent or non-congruent. Fabrice Parmentier (2008) found that task-relevant and task-irrelevant novel distractor words modulated performance of a decision task in a way that suggested that the words were semantically analyzed immediately following presentation. Specifically, when an auditory distractor word (either 'left' or 'right'), presented simultaneously with a target arrow, was incongruent with the direction of a target arrow, it took longer than with a congruent distractor word for participants to identify the direction of the arrow. This suggests that task relevant information can be more distracting than task irrelevant information. It should be noted, however, that this interference did not occur when the congruent and incongruent distractors were standard distractors (presented on every trial). The authors conducted two different manipulations of the neutral:word distractor ratio, and observed interference only when a neutral distractor was used (a sinusoidal tone) on 80% of trials, and congruent and incongruent distractors each appeared in 10% of trials. This indicates that interference occurred only when the meaningful distractors were novel as well. (The current study also examined the interference caused by semantically relevant and irrelevant distractors, but in a uni-modal design. Parmentier (2008) used a bimodal design, with auditory distractor words and a visual decision task.)

Lavie and colleagues have also manipulated the similarity of the distractor and the target in a visual search task. For example, in sections of her study outlining the two-part attention model, half of the distractors would be the same as the target (X and X), and half of them would be different (X target with N distractor) (Lavie et al., 2004). These studies have shown that task non-congruent distractors cause greater interference with attention control than congruent distractors. Still, these studies used only stimuli and distractors without semantic meaning.

Lavie and Forester (2008) noted that most selective attention studies (including most of Lavie's own) utilize the same or similar stimuli as the distractor item and target item (i.e. black and white letters, numbers, symbols, etc.). Therefore, the experimenters focused on the effects of truly-task irrelevant distractors, distractors that were completely unrelated to the target stimuli. Such studies are valuable because they aim to more closely simulate real-world distraction and selective attention in a controlled clinical setting. During a visual search task, participants had to identify whether an X or an N was present. During the task standard distractors (X or N) were presented in 80% of trials, related distractors (similar to target items) were presented in 10% of trials, and unrelated distractors (images of cartoon characters) were presented in the remaining 10%. They found that these novel distractors could create more disruption of performance than standard distractors, but only when participants had a longer period of time to identify the target. When a time pressure was added to the search task—they were given 500 ms to respond rather than no time limit—the extra interference of irrelevant distractors was eliminated in high perceptual load conditions, in agreement with the Load Theory's prediction.

Belke, et al. (2008) also investigated the effects of target-similar distractor items in a visual search task. Their experimental tasks presented participants with a target, presented as a single word, followed by an array of images. In some trials the array would contain the target, and some trials would contain an item semantically related to the target (target: "shirt," related item: an image of a pair of pants). Using an eye-tracker, they found that when the target was present, even in trials where the related item was also present the participant's first fixation would fall on that target significantly more frequently than non-target items. When the target was absent, however, the first fixation would fall on the related item significantly more than the other items in the array (Belke et al., 2008). These results suggest that, when primed with a target item, the participant is more likely to look at an item related to that prime than an item unrelated to that prime.

Finally, Belke et al. (2008) found that items semantically related to a target are more attention-grabbing than unrelated items. This, along with the broad finding that task-congruent distractors are more distracting than neutral distractors (Lavie et. Al. 2004; Parmentier, 2008; Lavie & Forester 2008), and Parmentier's findings that semantic information can be obtained from novel distracters, lead us to four hypotheses. First, we hypothesized that distractors with semantic content related to the semantic content of the memory set would cause more interference (more distraction, slower RTs) than either distractors unrelated to the memory set, or neutral distractors (which should have caused the

least interference). Second, we also expected to see an effect of load on distractibility compatible with the Load Theory: we expected to see greater distractibility (slower RTs) in the high cognitive load condition. Third, based on the Load Theory we hypothesized that there would be no interaction between cognitive load and distractor type; past studies produced no results that would suggest an interaction either way. Finally, we expected to see a higher rate of false positive identifications in the memory probe in the high load condition, as there should have been less cognitive capacity available to actively ignore distractor words.

Methods

Thirty-four Introduction to Psychology students from Franklin and Marshall College participated, in exchange for course credit. Participants volunteered for participation via sign up sheet. No demographic factors were recorded or controlled for.

Stimulus images were generated using Graphic Converter software, and sets of stimuli (trials) were constructed and ordered manually (although a random order was generated by computer to guide the organization of the trials). Stimulus sets were presented using Generic Psychology Lab software. All software was run on Mac OS9. Accuracy and mean reaction times for each participant were recorded by the software and analyzed using SPSS statistical software.

Each trial consisted of three phases that were presented serially at fixed time intervals: part 1, the memory set presentation, part 2, the selective attention task, and part 3, the memory probe. The memory set consisted of a small set of words the participants were asked to remember. The selective attention task was a simple decision task. Each attention task presentation consisted of a target letter (N or X) and a non-target (an O), one above the fixation point and one below. The target appeared randomly and equally in each location. The trials were also split evenly between the two target letters, each target letter appeared equally. One and only one of the two targets were present on every trial. Participants were asked to identify which of two target letters were present (X or N). During this time, a distractor presented simultaneously with the selective attention task, in the periphery of the screen aligned horizontally with the fixation point. The target and non-target stimuli (the N or X, and O) appeared, along with a flanker distractor, for 250ms. This display was followed immediately by a blank screen. Participants had from the offset of the selective attention screen (the onset of the blank screen) onward to make their response, either pressing the N key if the N was present, or the X key if the X was present. Finally, participants were shown a word, and asked if it was present in the original set.

This study had a 2x3 within subjects design: cognitive load x semantic content of distractors. Cognitive load was defined here (as in Lavie et. al. 2004) as the number of items presented in the memory set. The two conditions, high and low, were defined as memory sets containing 5 and 2 items respectively. In the low load condition, the memory set was presented for 2s and in the high load condition it appeared for 4s. These presentation times were meant to eliminate extra search/reading time from the low load condition, while still allowing enough time in the high load condition for the participant to read and process all memory set words. A similar method was used in Lavie (2004) when manipulating cognitive load. The two levels of cognitive load, as manipulated by memory set size, were presented in separate experimental blocks. Participants were assigned one of two groups at the start of the study, and group assignments alternated every-other participant. Presentation order was counterbalanced between groups.

In order to allow distractors to be semantically related (or not) to the words in the memory set, all memory sets consisted of either 2 or 5 words meaningfully related to each other. Our study's 3 distractor conditions were 1) a "neutral" distractor: a single symbol (#) without semantic meaning; 2) "related" distracters: words that are in the same semantic category as the words in the memory set; and 3) "unrelated" distracters: words that carry semantic meaning but are unrelated to the words in its trial's memory set. Words were semantically related based on broad categorization by meaning, or words were grouped under one broad category. For example, a high load memory set could consist of the words "apple, pear, grape, orange, cherry." A related distractor word would be "plum," and an unrelated distractor could be "truck."

The memory probe could have either been one of the words from the memory set (apple), or a word still related to the set, but not present (peach). While words were recycled between the two trial blocks, no words were presented more than once in the same trial block. Words were only used if they contained less than three syllables, and were easily recognizable.

In order to preserve the novelty of the meaningful distractors (which, according to Parmentier, 2008, was essential to the recognition of the distractors), a longer string of meaningless symbols (ex. #$%!?) was not used as the neutral distractor. Within each block of trials, 50% of distractors were neutral

distractors (#), 25% were related, and 25% were unrelated.

Reaction times on the decision task and accuracy rates in the memory probe were collected. A 2 x 3 within subjects ANOVA (cognitive load x relatedness of distractor) was run to analyze the possible effects of our variables as they pertain to the first three hypotheses, and appropriate post-hoc tests were run as necessary. Error rates were also calculated, and a 2x2 within subjects ANOVA (presence of the memory probe x cognitive load) was run to examine any patterns regarding false positives or false negatives (as discussed in our fourth hypothesis).

Results

The 2 x 3 ANOVA (cognitive load x relatedness of distractor) results indicated a significant main effect of relatedness, $F(2,62)=16.008$, $p<.001$, partial $\eta^2=.341$. However, there was no main effect of cognitive load, $F(1,31)=.018$, $p=.894$, and partial $\eta^2=.001$. The interaction between cognitive load and relatedness was statistically significant, $F(2,62)=.028$, $p=.028$, and partial $\eta^2=.109$. Fisher's LSD post hoc test was conducted to determine which groups of relatedness were significantly different in reaction times. Results revealed that all three groups differed significantly.

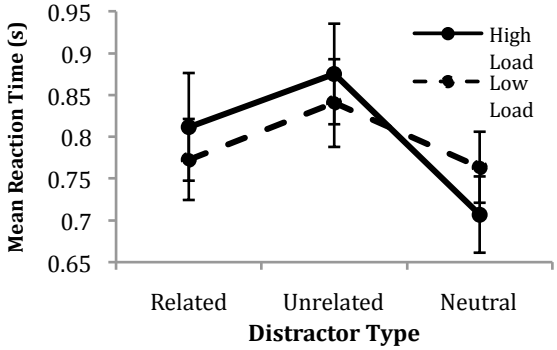


Figure 1: Mean reaction time as a function of cognitive load and distractor type.

A 2 x 2 ANOVA (presence of memory probe x cognitive load) indicated significant main effect of cognitive load, $F(1,31)=6.779$, $p=.014$, and partial $\eta^2=.179$, as well as a significant main effect of presence, $F(1,31)=30.998$, $p<.001$, and partial $\eta^2=..500$. There was no interaction between cognitive load and presence, $F(1,31)=.005$, $p=.943$, and partial $\eta^2<.001$.

Post-hoc t-tests were conducted to look at differences among related and neutral distractors in different cognitive loads. The results revealed a significant difference between mean reaction times of high related distractors and high neutral distractors, $t(31)=3.123$ and $p=.004$. There was not a significant difference between mean reaction times of low related distractors and low neutral distractors, $t(31)=.371$ and $p=.713$.

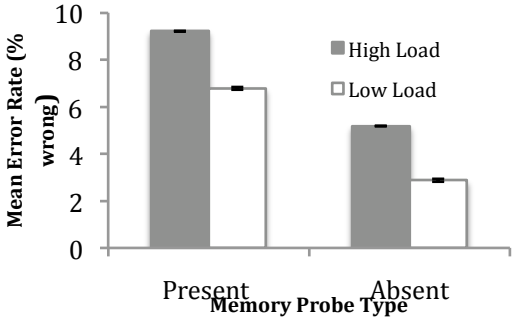


Figure 2: Mean error rate (%) as a function of cognitive load and memory probe condition.

Discussion

The first hypothesis was not supported. While a significant main effect of distractor type was found, the unrelated distractors caused more interference than related distractors. Neutral distractors caused the least interference, in line with the hypothesis. The background literature leading to this hypothesis was not cohesive, and at the time the hypotheses were written, the literature suggesting the greater interference capacity of related, rather than unrelated, distractor items was more compelling than the other findings available. Given the current findings, a second examination of the background information highlights several studies that do support the present results. As noted, Lavie et al. (2004) found that task non-congruent distractors were more distracting than task congruent distractors, Lavie & Forester (2008) found that novel, task irrelevant image distractors were more distracting than familiar or relevant distractors, and Parmentier (2008) found that task relevant items presented simultaneously with a task facilitated performance, while irrelevant items interfered.

These results might suggest a sort of priming effect. Perhaps the memory set primes some larger category or group meaning, and as a result the participant is not surprised by the presence of a related distractor; an unrelated distractor, then, is unexpected and more attention-grabbing. This is not completely satisfactory, however, because it might imply that related distractors would produce faster RTs than even neutral distractors would. This is only partially supported by our results. While in the high load condition, the neutral distractor caused significantly less interference than the related distractor, in the low load condition the two distractor categories produced mean RTs that were

virtually the same. If priming were involved, we would have expected to see related distractors produce significantly faster RTs than both unrelated and neutral distractors, or very similar RTs to neutral distractors. While these results have interesting implications in the broader discussion of priming effects and working memory, further discussion of the topic is beyond the scope of this study.

Although our results are not consistent with our original predictions, they still build on the previous literature in valuable ways. All previous work cited here related their distractors only to the items in their selective attention or perceptual load tasks; the related (or unrelated) information appeared all on the same screen. The current study sought to examine relatedness between the attention task and the cognitive load, to manipulate relatedness across tasks and attention mechanisms. Because of this the results were slightly unpredictable, but valuable nonetheless.

The second hypothesis was also not supported: no main effect of cognitive load was observed. This result was unexpected, as the effect of cognitive load has been observed in past work. A trend towards significance was observed—high load RTs were slower than low load RTs—but only in the 'related' and 'unrelated' condition. This suggests that, possibly, the cognitive load manipulations were not adequate representations of 'high' and 'low' cognitive load. One possible explanation of this involves the related nature of the stimuli. Perhaps the fact that all memory set items fell under one broad category or meaning provided a strategy for remembering them: maybe participants remembered the group rather than each individual word (consciously or not) in an effort to reduce the load on the cognitive system. The results regarding our fourth hypothesis provide more evidence to this effect.

Our fourth hypothesis was supported. We observed significantly more false positives in the high load condition than in the low load condition: participants were failing to accurately remember the words in the larger memory sets. After considering the related nature of the memory set words, this result suggests that participants noted the category of the words presented along with individual words. As all memory set words, present and absent, were related to their memory set (it would have been too obvious had the probe words been unrelated) false positives indicate that participants recognized the category membership/ semantic meaning of the probe word and responded accordingly.

The results concerning the second and fourth hypotheses point to semantic grouping as a characteristic that can mitigate the effects of cognitive load in both directions. In high load conditions, it appears that semantic grouping reduces the load that would otherwise be placed on a cognitive system by trying to retain five unrelated or meaningless items. Relating memory set provides a crutch, a strategy for the participant to make remembering easier. By relating the high load memory set items to each other, we may have created a pseudo-high cognitive load, not high enough to mimic previous results. The current manipulation of low load does differ from past studies'. In Lavie et al. (2004), the low cognitive load condition contained one item: a single letter. In Belke et al.'s (2008) manipulation of cognitive load a single number was used in the low load memory sets. To compare, the current study used two one- or two - syllable words. Setting aside the extra item in our low load condition, words have meaning, single letters do not. It is possible that the combination of these two factors brought the low load condition's difficulty closer to that of the high load condition. Adding the factor of meaningfully related memory sets, seems to have knocked the significance out of the effect of cognitive load.

Our third hypothesis was, again, not supported. A significant interaction occurred between cognitive load and distractor relatedness. The relationship between the two variables was very similar between the two conditions with meaningful distractors (the 'related' and 'unrelated' conditions)**.** The interaction appears between these conditions and the neutral distractor condition: neutral distractors were more distracting in the low load condition than in the high. This is not consistent with load theory, which would predict faster RTs in the low cognitive load condition all around. Most of the background studies cited here disregard their neutral data, there isn't much in the literature to compare our results to. The degree of the difference between RTs of contentful and non-contentful distracters—the fact that the trend is reversed in the neutral condition—may indicate a different mechanism is at work when the distractor and the memory set information are presented in the same form (all words, as opposed to words and symbols). This interpretation is not fully compatible with the past findings, however. Lavie and Forester (2008) found that when target stimuli were letters and distractors were images of cartoon characters, the images were still significantly distracting (when compared to letter distractors). This study, though, did not manipulate cognitive load; the novel distractors were not novel to cognitive load content, as in the current study, they were novel to perceptual load items. Perhaps, as the current results suggest, the interaction *between*

attention mechanisms is qualitatively different than that *within* mechanisms. This qualitative difference is interesting and warrants further study.

The size of the sample may have limited the study. Using additional participants was not feasible, however, and because of the within-subjects design was not essential. The software used to present stimuli, and record reaction times, responses and error rates also limited our ability to fully explore the data. A more sophisticated program like E-Prime could alleviate these technological issues.

The current study is incomplete in that it does not compare related memory sets to unrelated memory sets. Further exploration is necessary to determine whether the results seen here are truly attributable to the semantic relatedness of the memory set and the distractor, or if they are simply a product of using whole words as stimuli. Adding a third independent variable, that of memory set relatedness, would enhance the literature in this area.

More detailed data regarding error rates and the type of distractor task would shed further light on the issue of false memories: it is possible that the greater the interference of the distractor in the selective attention task, the higher the error rates would climb. If unrelated distractors are more distracting, it follows that error rates in the unrelated distractor condition would be significantly higher than those in the related or neutral conditions.

Because of technological limitations, it could be valuable to re-examine the interaction between distractor type and cognitive load. Validating the current results regarding neutral distractors could point to different mechanisms used in ignoring extraneous information.

Conclusion

A desire to better understand what interrupts or facilitates selective attention continues to drive research in cognitive psychology. While Lavie's Load Theory provides a valuable theoretical explanation of the phenomena, it is also vital that studies explore selective attention in real-world settings using real-world stimuli. The present results indicate that while the Load Theory provides a strong theoretical base, there are stimulus characteristics, like meaning and relationships between stimuli, that can alter the general pattern, but not without a cost. While it is possible that relating stimuli to one another reduces the cognitive capacity required to retain it, retention and recall suffer when such strategies are used. Knowing how attention and working memory are disrupted and aided could be particularly applicable in the field of education, and could be used to teach strategies for better retention and more effective methods of teaching and information presentation.

References

Belke, E., Humphreys, G., Watson, D., Meyer, A. and Telling, A. (2008). Top-down effects of semantic knowledge in visual search are modulated by cognitive but not perceptual load. *Perception & Psychophysics, 70* (8), 1444 – 1458.

Brand-D'abrescia, M. and Lavie, N. (2008). Task coordination between and within sensory modalities: effects on distraction. *Perception and Psychophysics, 70* (3), 508-515.

Forster, S. & Lavie, N. (2008). Failures to ignore entirely irrelevant distractors: The role of load. *Journal of Experimental Psychology: Applied, 14,* 73-83.

Lavie, N. (1995). Perceptual load as a necessary condition for selective attention. *Journal of Experimental Psychology: Human Perception & Performance, 21*, 451-468.

Lavie, N., and Cox, S. (1997). On the efficiency of visual selective attention: Efficient visual search leads to inefficient distractor rejection. Psychological Science, 8, 395 – 398.

Lavie, N., Hirst, A., de Fockert, J., and Viding, E. (2004). Load theory of selective attention and cognitive control. *Journal of Experimental Psychology: General, 133*(3), 339-354.

Parmentier, F. (2008). Towards a cognitive model of distraction by auditory novelty: The role of involuntary attention capture and semantic processing. *Cognition, 109*, 345-362.

Rees, G., Frith, C., and Lavie, N. (2001). Processing of irrelevant visual motion during performance of an auditory attention task. *Neuropsychologia, 39*(9), 937-949.

Wilson, D., MacLeod, C., & Muroi, M. (2008). Practice in visual search produces decreased capacity demands but increased distraction. *Perception and Psychophysics, 70*(6), 1130-113.

Is perceptual acuity asymmetric in isolated word recognition? Evidence from an ideal-observer reverse-engineering approach

Nathaniel J. Smith and Wen-Hsuan Chan
njsmith@cogsci.ucsd.edu, **whchan@ucsd.edu**
University of California at San Diego, Department of Cognitive Science
9500 Gilman Drive #515, La Jolla, CA 92093-0515 USA

Roger Levy
rlevy@ling.ucsd.edu
University of California at San Diego, Department of Linguistics
9500 Gilman Drive #108, La Jolla, CA 92093-0108 USA

Abstract

An asymmetrical optimal viewing position (OVP) effect in isolated word recognition has been well documented, such that recognition speed and accuracy are highest when the point of fixation within the word is slightly to the left of center. However, there remains disagreement as to the source of the asymmetry in the OVP effect. One leading explanation is that perceptual acuity in isolated word recognition is asymmetric, falling off more rapidly to the left than to the right. An alternative explanation is that of lexical constraint: perceptual acuity may be symmetric, but that the distributional statistics of the lexicon are such that the letters near the beginning of a word are on average of greater value in discriminating word identity than the letters near the end. On both these accounts, a left-of-center fixation point optimizes the efficient accrual of perceptual input from the word string, but for different reasons. These accounts have been difficult to tease apart experimentally due to the ubiquitous potential influence of lexical constraint. Here we take a novel approach, constructing an ideal-observer model of isolated word recognition which takes into account word frequency information and thus intrinsically accounts for the role of lexical constraint. Within this model, the shape of the perceptual acuity curve is governed by free parameters that can be estimated from purely behavioral response data from word recognition experiments. Fitting our model to the experimental data of Stevens & Grainger (2003), we find that the asymmetric version, in which perceptual acuity can differ to the left and to the right, fits human behavioral responses significantly better than symmetric versions in which the perceptual acuity curve is constrained to be the same to the left and to the right. Furthermore, in both parametric and nonparametric versions of the asymmetric model, perceptual acuity falls off more rapidly to the left than to the right. These results support the position that the perceptual acuity curve in isolated word recognition is indeed asymmetric.

Keywords: Psychology, Cognitive Science, Perception, Language Understanding, Decision Making, Bayesian modeling

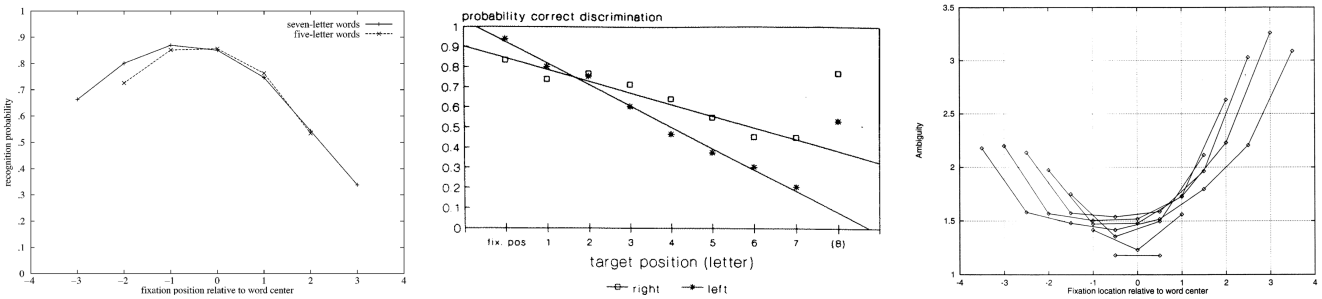
Introduction

Literate native speakers are exquisitely adapted to the visual and linguistic processing of written text in their language. The naturalistic task underlying most of this adaptation is reading (Rayner, 1998). In the study of eye movements in reading, one of the most striking examples of this adaptation in the last several decades has been discovery of *asymmetry of the perceptual span*: in languages which are written from left to right, readers are more sensitive to material to the right of the center of fixation on the page than they are to material on the left. In languages which are written from right to left, however, this sensitivity is reversed (Rayner, Well & Pollatsek, 1980). Since visual acuity per se is not itself asymmetric (as evidenced by experimental work on perception of non-linguistic visual inputs), the most intuitive interpretation of this finding is that, in ordinary progressive reading, because readers of languages written left to right have already seen what lies to the left of their eyes, they differentially attend to what lies to the right (and vice versa for languages written right to left).

However, the discovery by O'Regan, Lévy-Schoen, A., Pynte, J. & Brugaillére (1984) of the optimal viewing position (OVP) in *isolated* word recognition makes the picture more complex. The OVP in isolated word recognition can be succinctly described as follows: word recognition is fastest and most accurate when the initial fixation point of the eyes is slightly to the left of the center of the word (Figure 1a). The discovery of the OVP launched considerable discussion as to its nature and implications, since it cannot be obviously accounted for by an asymmetry in perceptual acuity that would be adaptive for the task.

At present, there are two leading explanations that have been proposed for the OVP in isolated word recognition. One is that the asymmetry of perceptual acuity to the left and to the right within reading may affect *all* processing of visual linguistic input. If acuity drops off more rapidly to the left of the fixation point than to the right, then the best strategy to recognize a word would be to fixate at a left-of-center location, maximizing average acuity across the word as a whole. Evidence supporting this position has been adduced by Nazir, O'Regan & Jacobs (1991) and Nazir, Heller & Sussman (1992), who demonstrated left-right acuity differences in tasks involving the detection of a target letter at a variable position within a masking letter string (e.g., *kkkkykk*). They found that the drop-off in performance was a monotonic function of visual eccentricity, and the left visual field showed steeper drop-off than the right visual field (Figure 1b).

The other leading explanation was put forth by Clark & O'Regan (1999), who argued that a better way to understand the contributions of these different mechanisms may lie in the distributional statistics of the written lexicon itself. They investigated the contributions of orthographic constraints, con-



(a) OVP effect (Stevens & Grainger, 2003)

(b) Asymmetric perceptual acuity in letter-in-mask recognition (Nazir et al., 2003)

(c) Asymmetry of lexical constraint (Clark & O'Regan, 1999)

Figure 1: The OVP effect in isolated word recognition and possible explanations for it.

structing a simple measure of residual lexical ambiguity that would hold assuming only that two letters near the fixation point and at the end of the word are known. For a range of word lengths in both French and English, this ambiguity measure is minimized just to the left of the word's center, capturing the OVP effect through lexical constraint without resorting to an asymmetric perceptual acuity curve (Figure 1c).

It is difficult to adjudicate between these two possible explanations through purely experimental means, because well-established effects on word recognition such as those of word frequency and neighborhood density make it is fairly clear that lexical constraint plays a ubiquitous role in the process but varies slightly for every word in the lexicon, making it difficult to design a word-recognition experiment that controls for lexical constraint while testing perceptual acuity. Conversely, our understanding of precisely how the perceptual acuity curve affects word recognition remains limited, making it difficult to hold it constant and test only lexical constraint as a possible source of OVP effects. In this paper, we take an alternative approach, constructing an ideal-observer model (Marr, 1982; Anderson, 1990) of isolated word recognition in which lexical constraints are assumed to be available. The perceptual acuity curve in this model is determined by a set of parameters which are free and can be fit to behavioral data using well-established techniques of statistical inference. Our ideal-observer model thus allows us to "reverse-engineer" the perceptual acuity curve active in isolated word recognition on the basis of a lexicon (with word frequencies) and a behavioral data set, and assess whether and how the reverse-engineered acuity curve may be asymmetric.

Data

The dataset used in our study is from Experiment Two of Stevens & Grainger (2003). In this experiment, words were presented for 50ms each at various positions relative to the center of the fixation. After presentation of each word, the participant was asked to type the presented word back into the computer.. The dataset contains 75 five-letter words and 105 seven-letter words and the human performance (correctness of the response) for each word from seventy subjects, for a total of 12,600 observations balanced across word identity and fixation position. We obtained a word frequency database from the Agence France Presse (AFP) French Corpus, which contains French journal articles from 1993-1996. Since the human dataset has both five-letter and seven-letter French words, we extracted frequency counts for all five-letter and seven-letter words in the corpus. The resulting lexicon contains 14,379 five-letter French words and 21,569 seven-letter French words.

Model

The Bayesian Reader

We use an ideal-observer model of isolated word recognition, the Bayesian Reader (Norris, 2006, 2009). Our version of the Bayesian Reader introduces several key assumptions regarding the nature of word recognition which determine the form of the probabilistic model:

- Word recognition is a Bayesian hypothesis test in which prior expectations regarding what word is likely to be presented are combined with perceptual evidence $\mathbf{d}$ to determine posterior beliefs about what word w is being seen:[1]

$$P(w|\mathbf{d}) = \frac{P(\mathbf{d}|w)P(w)}{P(\mathbf{d})};$$

- The prior probability of each word $P(w)$ is proportional to its corpus frequency of occurrence;
- Perceptual evidence consists of a sequence of independent identically distributed (i.i.d.) input samples $d^{(1)},\ldots,d^{(N)}$ drawn from a NOISE DISTRIBUTION $P(d|w)$, where samples accrue at a constant rate over time;
- If we denote the letters of a word w of length L as $w_1,\ldots,w_L$, then an input sample d can be decomposed into L samples $d_1,\ldots,d_L$, with d_i conditionally independent of d_j given w_i and w_j for all $i \neq j$, so that

[1] Tasks in which the decision to be made is something other than the identity of the word—e.g., a lexical decision about whether the input string is a word in the participant's language—can be formulated as Bayesian hypothesis tests among the possible choices (Norris, 2006; del Prado Martìn, 2008).

$$P(d|w) = \prod_{i=1}^{L} P(d_i|w_i).$$

We will refer to the term $P(d_i|w_i)$ as the noise distribution for letter w_i.

Thus far, this version of the Bayesian Reader is simpler and more general than that introduced by Norris (2006), who assumed (a) a specific representation of each sample d as a point in a high-dimensional space; (b) a multivariate Gaussian form for the noise distribution $P(d|w)$; and (c) a specific estimate of the noise variance used by the ideal observer in computing the posterior distribution over words given perceptual evidence.

Adaptation to modeling visual acuity curves

In order to adapt the Bayesian Reader to the task of estimating visual acuity curves, however, we need to introduce a dependence of the noise distribution for each letter on its physical positioning. In particular, we assume that the noise distribution for each letter is dependent on its eccentricity as measured in number of characters from the point of fixation, with negative values corresponding to left-of-fixation and positive values to right-of-fixation; and on its proximities v^L and v^R to the left and right edges of the word, also measured in characters. If we define these physical positioning characteristics of a letter w_i as $k_i \equiv \langle e_i, v_i^L, v_i^R \rangle$, then its position-contingent noise distribution can be denoted as $P(d_i|w_i,k_i)$.

For the empirical modeling studies presented here, we make the additional assumption that the value of

$$E_{d_i|w_i^*,k_i}\left[P(d_i|w_i,k_i)\right], \tag{1}$$

where $E_{d_i|w_i^*,k_i}$ denotes expectation under the conditional distribution $P(d_i|w_i^*,k_i)$ for the true letter w_i^* being presented, depends only on k and on whether $w_i = w_i^*$. This assumption can be interpreted as stating that every letter is equally confusable with all letters other than itself; the quantity in (1) for $w_i \neq w_i^*$ can be interpreted as the level of confusability of a letter as a function of its physical positioning. This assumption is not necessary within the overall framework, and indeed could be relaxed in order to incorporate letter confusability matrices (Engel, Dougherty & Jones, 1973; Geyer, 1977) into the model and even to learn them directly from behavioral word-recognition data. In the present studies, however, this assumption greatly simplifies and facilitates both the statistical learning problem and its computational implementation.

Learning visual acuity from word identification data

Recall that in their word-identification study, Stevens & Grainger (2003) presented experimental participants with five- and seven-letter words one at a time for a brief, fixed interval too short to permit refixation, with fixation position varying across trials. The behavioral response r in each trial was the participant's guess as to which word they saw. Our goal is to use these behavioral responses to learn the dependence of the visual acuity of a letter—as quantified by confusability in (1)—on its physical positioning.

We model the naming task using the assumptions outlined in the previous two sections. In general, the experimental participant must choose their response r through some possibly stochastic decision process based on their posterior beliefs $P(w|\mathbf{d},k)$ about what word they saw. We further assume that the participant makes their choice of response through probability matching, so that the probability of any response r given the word w^* actually being presented is given by its expected posterior probability:

$$P(r|w^*,k) = E_{\mathbf{d}|k,w^*}\left[P(r|\mathbf{d},k)\right] \tag{2}$$

where $E_{\mathbf{d}|k,w^*}$ represents the expectation marginalizing over possible perceptual input samples given the true word and its physical positioning. For notational simplicity, we omit the subscript on the expectation whenever it is clear from context.

Equation (2) can be rewritten using Bayes' rule as

$$P(r|w^*,k) = P(r)E\left[\frac{P(\mathbf{d}|r,k)}{P(\mathbf{d}|k)}\right]. \tag{3}$$

Equation (3) is the expectation of a ratio of random variables $E[Y/X]$, an expression which cannot in general be manipulated exactly. Using the method of propagation of error, however, a second-order approximation for the expectation of a ratio can be found (Rice, 1995):

$$E\left[\frac{Y}{X}\right] \approx \frac{E[Y]}{E[X]} + \frac{1}{E[X]^2}\left(\mathrm{Var}[X]\frac{E[Y]}{E[X]} - \mathrm{Cov}[X,Y]\right).$$

We now turn our attention to the rightmost part of this expression, the covariance between the numerator and the denominator—in Equation (3), these terms are $P(\mathbf{d}|r,k)$ and $P(\mathbf{d}|k)$ respectively. Insofar as any individual word plays only a small part in the calculation of the marginal probability $P(\mathbf{d}|k)$, we would expect the covariance of this marginal probability with $P(\mathbf{d}|r,k)$ to be small (with the important caveat that because words tend to look more like each other than like non-words, there will generally be some positive covariance, and its magnitude may depend on the orthographically typicality of w^* and r). Dropping the covariance from the above approximation allows us to approximate our posterior probability as

$$P(r|w^*,k) \approx P(r)\frac{E\left[P(\mathbf{d}|r,k)\right]}{E\left[P(\mathbf{d}|k)\right] + \frac{\mathrm{Var}[P(\mathbf{d}|k)]}{E[P(\mathbf{d}|k)]^3}}.$$

Ignoring the denominator (which is constant with respect to r) and decomposing the perceptual input $\mathbf{d}$ into its component independent samples at each time j and letter position i, we obtain

$$P(r|w^*,k) \propto P(r) \prod_{i,j} E_{d_i^{(j)}|k_i,w_i^*} \left[P(d_i^{(j)}|r_i,k_i) \right].$$

We take advantage of the identical distribution of the N samples to obtain the approximate unnormalized probability

$$P(r|w^*,k) \propto P(r) \prod_i \left(E_{d_i|k_i,w_i^*} \left[P(d_i|r_i,k_i) \right] \right)^N.$$

We are now ready to take advantage of our assumption from the previous section that each letter is equally confusable with all letters other than itself—that is, the value of each of the above terms $\left(E_{d_i|k_i,w_i^*} \left[P(d_i|r_i,k_i) \right] \right)^N$ depends only on k_i and on whether $w_i = w_i^*$. For each k_i, let us denote the value taken when $w_i = w_i^*$ as p_i, the value taken when $w_i \neq w_i^*$ as q_i, and the ratio $\frac{p_i}{q_i}$ as l_i. Substituting these terms in and dividing the entire expression by $q_1 \ldots q_L$ gives us our final approximate expression for the probability of the participant's response:

$$P(r|w^*,k) \propto P(r) \prod_{i:w_i=w_i^*} l_i \tag{4}$$

where l_i is dependent on the physical positioning of the letter in question. On the original assumption of the Bayesian Reader that the number of input samples N accumulates at a constant rate over time, the value $\log l_i$ can be interpreted as the average rate at which perceptual information accrues at position i.

Model parameterization and estimation

Within the context of our model, the goal of inferring a visual acuity curve from behavioral word-recognition data entails estimating these input-accrual rate parameters $\log l_i$. In the studies presented here, we assume as stated before that the $\log l_i$ for each letter is a function of three properties of its physical position: its eccentricity from the fixation point, its proximity from the left edge of the word, and its proximity from the right edge of the word. All measurements are made in characters. We consider two functional forms for the eccentricity parameters: a PARAMETRIC form in which the values of $\log l_i$ is assumed to follow a Gaussian curve centered at the fixation point with maximum value α and standard deviation σ; and a NONPARAMETRIC form in which each eccentricity has its own arbitrary parameter value. For each form, we consider a SYMMETRIC version in which the eccentricity parameters $\log l_i$ are determined by the absolute eccentricity, and an ASYMMETRIC version in which the parameters for negative and positive eccentricity values of the same magnitude can be different (in the parametric Gaussian case, the asymmetric model allows different standard deviations σ_L and σ_R to the left and the right of the fixation point). Additionally, all models include one left-edge and one right-edge "bonus" parameter, b_L and b_R, added to the eccentricity parameters to determine the input-accrual rate parameters $\log l_i$ for the first and last letters of a word respectively. That is, if the fixation position is on the f-th character of the word and the function $e(\cdot)$ maps eccentricities to values in $\log l_i$ space, we have

$$\begin{aligned}
\log l_1 &= e(1-f) + b_L && \text{(first letter)} \\
\log l_i &= e(i-f) && \text{(middle letters; } i \notin \{1, L\}\text{)} \\
\log l_L &= e(L-f) + b_R && \text{(last letter)}
\end{aligned}$$

In all cases, we fit the parameters of our models using maximum likelihood estimation. Fortunately, the gradient of our model is readily calculable and allows estimation using standard gradient-descent techniques.

Results

For each of our nonparametric ($-$PAR) and parametric (+PAR) models, we fit a symmetric (+SYM) and an asymmetric ($-$SIM) variety to the 12,600-observation dataset of Stevens & Grainger (2003, Experiment 2) estimation. This dataset contains presentations of both five-letter and seven-letter words, with every letter of each word serving as the fixation point for an equal number of trials. For each model, we used a single set of eccentricity and edge-bonus parameters parameters to cover all trials, giving us four parameters in the symmetric parametric case (one maximum acuity parameter, one standard deviation, and two edge bonuses), five in the asymmetric parametric case (two standard deviations instead of one), nine in the symmetric nonparametric case (seven eccentricity parameters and two edge bonuses), and fifteen in the asymmetric nonparametric case (thirteen eccentricity parameters instead of seven). These models are nested in the classical statistical sense as follows:

$$[-\text{PAR},-\text{SYM}] \prec \{[-\text{PAR},+\text{SYM}],[+\text{PAR},-\text{SYM}]\} \prec [+\text{PAR},+\text{SYM}]$$

Since we use maximum likelihood estimation with far more observations than parameters, we can use likelihood-ratio tests for pairwise comparisons of all models except between [$-$PAR,+SYM] and [+PAR,$-$SYM]. These tests indicate that asymmetric models explain participant response behavior far better than symmetric models in both parametric ($\chi^2(1) = 250, p \ll 0.001$) and non-parametric cases ($\chi^2(6) = 345.7800, p \ll 0.001$). Among asymmetric model variants, the nonparametric model explains participant response behavior significantly better than the Gaussian model ($\chi^2(10) = 564.7, p \ll 0.001$).

For the asymmetric nonparametric model, we estimated standard deviations for our parameter estimates using 100 bootstrap replicates. Figure 2 graphs the value of $\log l$ as a function of eccentricity, together with edge-bonus parameter

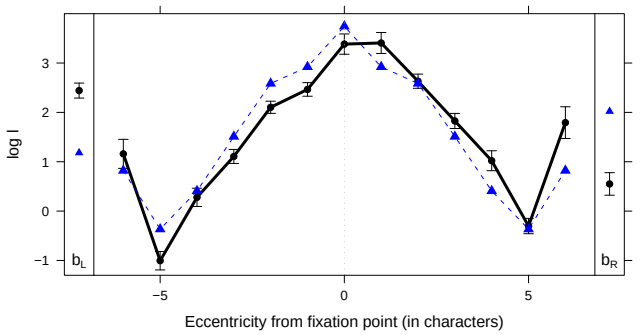


Figure 2: Symmetric (dashed blue lines) and asymmetric (solid black lines) non-parametric model parameter estimates. Error bars on the asymmetric model parameter estimates are standard deviations estimated from bootstrap replicates. b_L and b_R are "edge bonus" parameters added to the appropriate eccentricity parameters to obtain $\log l$ for the leftmost and rightmost letters in a word.

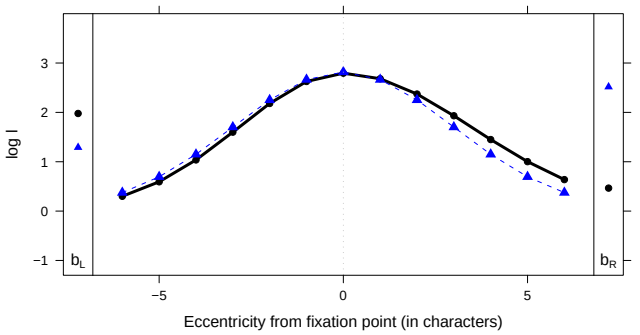


Figure 3: Symmetric (dashed blue lines) and asymmetric (solid black lines) parametric model results, assuming that eccentricity is piecewise Gaussian centered around the fixation point. b_L and b_R are "edge bonus" parameters added to the appropriate eccentricity parameters to obtain $\log l$ for the leftmost and rightmost letters in a word.

estimates (and error bars to indicate bootstrapped standard deviations for the asymmetric variant), for the two varieties of the nonparametric model. Figure 3 graphs these quantities for the two varieties of the parametric model. In both figures, eccentricity falls off to the left more rapidly in the asymmetric model than in the symmetric model, and to the right more rapidly in the symmetric model than in the asymmetric model.[2] This consistent pattern suggests that the data of Stevens & Grainger (2003) provide evidence for an asymmetry in the perceptual acuity curve in visual recognition of isolated words even when lexical constraint is taken explicitly into account.

Conclusion

The results of our modeling studies provide additional evidence for the idea that an asymmetric visual acuity curve contributes to the OVP effect documented in many studies on isolated word recognition (O'Regan et al., 1984; Vitu, O'Regan & Mittau, 1990; O'Regan & Jacobs, 1992; Stevens & Grainger, 2003). Even while explicitly accounting for the role of lexical constraint, we consistently found that the models which best accounted for the distribution of response accuracies found in Experiment 2 of Stevens & Grainger (2003) had an asymmetric perceptual acuity curve in which acuity dropped off more slowly as a function of visual eccentricity to the right than to the left. Although Stevens & Grainger (2003) also presented results of another experiment using the letter-within-mask identification task which called into question the generalizability of the findings of Nazir et al. (1991, 1992) regarding asymmetry of the perceptual acuity curve, the fact that an asymmetric perceptual acuity curve is required to provide the best account for their word-recognition data by an ideal observer with the knowledge of lexical statistics of the language calls into question the strong position staked out by Clark & O'Regan (1999) that the left-of-center position of the OVP may derive purely from lexical constraint. In our view, the most plausible theoretical position reconciling our modeling findings with empirical findings from the reading literature on the asymmetric perceptual span (Rayner et al., 1980) and with those of Stevens & Grainger (2003) on character-within-mask recognition is that isolated word recognition does indeed involve an asymmetry of perceptual acuity, but that it is a parasitic byproduct of cognitive adaptation to the naturalistic task of reading. The character-within-mask recognition task may be sufficiently unlike naturalistic reading that it does not trigger this asymmetric perceptual span. It is worth noting that the experiments of Nazir et al. (1991) and Nazir et al. (1992) always involved fixation on the leftmost or rightmost character of the word, and used a letter (instead of the hash mark # used by Stevens & Grainger) as the mask character. It is possible that the lower overall visual acuity due to fixation at the word's right or left edge together with the letter mask may have induced the language and perceptual systems to categorize input in this experiment more like that obtained in natural reading, inducing asymmetric perceptual acuity.

Although we believe our modeling approach represents an important step forward in resolving these issues, it is important to emphasize that several simplifying assumptions we introduced which led from the general ideal-observer formulation to the specific, highly tractable model of Equation (4) have the potential to significantly affect our modeling results and deserve more careful consideration in the future. The first

[2] In the nonparametric model, the most extreme eccentricities have oddly-behaving parameter estimates that indicate possible problems with model specification, perhaps because the edge bonus parameters are so often implicated in model predictions for these extreme eccentricities. We expect that fitting the nonparametric model to behavioral data involving presentation of words of a larger variety of lengths would be likely to reduce or eliminate this problem.

of these simplifying assumptions was equal confusability of letter pairs. Of course, it is well known that some letter pairs are more confusable than others (e.g., *o* is much more confusable with *e* than with *l* for native English speakers; Geyer, 1977). It is possible that the assumption of equal confusability could interact with the lexical statistics of French and the items chosen by Stevens & Grainger (2003) to create a confound in the explanation of our modeling results—for example, such a confound might arise if the letters near the end of their items were more highly confusable than the letters near the beginning of their items. This possibility can be explored in future work by relaxing the assumption of equal confusability and using established letter confusion matrices to scale our model parameters, or even to allow our model to learn confusability parameters directly from word-recognition behavioral data.

The second simplifying assumption deserving discussion is that participants's behavioral responses arose from probability matching. In other cognitive domains, this assumption seems to have reasonable theoretical and empirical support (Vulkan, 2000; Mozer, Pashler & Homaei, 2008; Vul & Pashler, 2008; Vul, Goodman, Griffiths & Tenenbaum, 2009). That being said, it is entirely possible that participants' responses may reflect maximization or some other similar decision process, and that most inter-trial response variability derives from the variation inherent in noisy perception. The consequences of this possibility may be explored in future work within our framework by explicit simulation of inter-trial noise instead of marginalizing over perceptual input as we have done here.

The final simplifying assumption is that of minimal covariance between the probability of noisy perceptual samples given the word under consideration and the marginal probability of those perceptual samples. As discussed earlier, this assumption is clearly wrong insofar as words in any given language tend to look more like each other than like nonwords; but the sheer number of words in the lexicon, combined with the considerable variability that does exist among wordforms, implies that this covariance should in general be rather small. It is also not obvious to us how this simplifying assumption might introduce a confound to our specific result of an asymmetric perceptual acuity curve in isolated word recognition. Nevertheless, there are two ways that this simplifying assumption could be relaxed in future work. First, explicit simulation of inter-trial noise would permit us to quantify the discrepancy between the simplified results we report here and the results which would obtain under the model more generally. Alternatively, we might try to quantify the covariance between $P(\mathbf{d}|w,k)$ and $P(\mathbf{d}|k)$ through explicit simulation, and then use these covariance estimates to adjust our expectation-based model directly. This latter alternative might have the added benefit of giving us more direct insight into the full range of top-down effects that are present in word recognition. Intuitively, this covariance should be larger for more "prototypically word-like" words, which should *decrease* the expected posterior belief for such words relative to less word-like words, a sort of second-order neighborhood-density effect. Further elucidation of all these issues awaits future research.

Acknowledgments

We are grateful to Michaël Stevens and Jonathan Grainger for graciously sharing their experimental data with us; and to Rebecca Colavin for her expertise in French. This research was partially supported by NIH Training Grant T32-DC000041 to the Center for Research in Language at UCSD to NJS, by a Taiwan Graduate Fellowship of government sponsorship for overseas study to WHC, and by NSF grant 0953870 to RL.

References

Anderson, J. R. (1990). *The Adaptive Character of Human Thought*. Lawrence Erlbaum.

Clark, J. J. & O'Regan, J. K. (1999). Word ambiguity and the optimal viewing position in reading. *Vision Research*, *39*, 843–857.

del Prado Martìn, F. M. (2008). A fully analytical model of the visual lexical decision task. In *Proceedings of the Cognitive Science Society*.

Engel, G. R., Dougherty, W. C., & Jones, G. B. (1973). Correlation and letter recognition. *Canadian Journal of Psychology*, *27*(3), 317–326.

Geyer, L. H. (1977). Recognition and confusion of the lowercase alphabet. *Perception & Psychophysics*, *22*, 487–490.

Marr, D. (1982). *Vision*. San Francisco: Freeman.

Mozer, M. C., Pashler, H., & Homaei, H. (2008). Optimal predictions in everyday cognition: The wisdom of individuals or crowds? *Cognitive Science*, *32*(7), 1133–1147.

Nazir, T. A., Heller, D., & Sussman, C. (1992). Letter visibility and word recognition: The optimal viewing position in printed words. *Perception & Psychophysics*, *52*(3), 315–328.

Nazir, T. A., O'Regan, J. K., & Jacobs, A. M. (1991). On words and their letters. *Bulletin of the Psychonomics Society*, *29*(2), 171–174.

Norris, D. (2006). The Bayesian Reader: Explaining word recognition as an optimal Bayesian decision process. *Psychological Review*, *113*(2), 327–357.

Norris, D. (2009). Putting it all together: A unified account of word recognition and reaction-time distributions. *Psychological Review*, *116*(1), 207–219.

O'Regan, J. K. & Jacobs, A. M. (1992). Optimal viewing position effect in word recognition: A challenge to current theory. *Journal of Experimental Psychology: Human Perception & Performance*, *18*(1), 185–197.

O'Regan, J. K., Lévy-Schoen, A., Pynte, J., & Brugaillére, B. (1984). Convenient fixation location within isolated words of different length and structure. *Journal of Experimental Psychology: Human Perception & Performance*, *10*, 250–257.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*(3), 372–422.

Rayner, K., Well, A. D., & Pollatsek, A. (1980). Asymmetry of the effective visual field in reading. *Perception & Psychophysics*, *27*(537–544).

Rice, J. A. (1995). *Mathematical Statistics and Data Analysis* (2 ed.). Duxbury Press.

Stevens, M. & Grainger, J. (2003). Letter visibility and the viewing position effect in visual word recognition. *Perception & Psychophysics*, *65*(1), 133–151.

Vitu, F., O'Regan, J. K., & Mittau, M. (1990). Optimal landing position in reading isolated words and continuous text. *pp*, *47*(6), 583–600.

Vul, E., Goodman, N. D., Griffiths, T. L., & Tenenbaum, J. B. (2009). One and done? Optimal decisions from very few samples. In *Proceedings of the Cognitive Science Society Conference*.

Vul, E. & Pashler, H. (2008). Measuring the crowd within: Probabilistic representations within individuals. *Psychological Science*, *19*(7), 645–647.

Vulkan, N. (2000). An economist's perspective on probability matching. *Journal of Economic Surveys*, *14*(1), 101–118.

Are grunters cheaters? The effects of grunting when judging the direction of a tennis shot

Scott Sinnett (ssinnett@hawaii.edu)
Department of Psychology, University of Hawaii at Manoa,
2530 Dole Street
Sakamaki C400
Honolulu, HI 96822
USA

Alan Kingstone (alan.kingstone@ubc.ca)
Department of Psychology, University of British Columbia,
2136 West Mall
Vancouver, BC V6T 1Z4
Canada

Abstract

There is a chorus of complaints that many professional tennis players who grunt when striking the ball gain an unfair advantage because the sound of the grunt distracts their opponent. However, scientific investigations of human attention and performance, specifically with regard to sound-vision interactions, would seem to predict that a grunting sound should help because it will draw attention to the visual event of a ball being struck. We tested the argument that a grunt has a negative impact by requiring participants to view videos of a professional hitting a ball to either side of a tennis court with or without a grunt. The task was to respond as quickly as possible to the ball's direction. Grunting interfered with performance making responses slower and less accurate. The competitive advantage afforded to the grunting player is potentially profound. The findings will be discussed in relation to current theory on multisensory integration.

Keywords: Attention; multisensory integration; distraction; tennis; action perception.

Introduction

Last year, for the first time, a Portuguese women's tennis player, Michelle Larcher de Brito, made it to the third round of the 2009 French Open. Unfortunately for Michelle she lost to Frenchwoman Aravane Rezai in a match where Michelle was heavily criticized for executing a loud and long grunt each time she hit the ball. The complaint is that Michelle, and many of the best players in tennis like her, such as Maria Sharapova (who grunts at over 100 decibels) and the Williams sisters, gain an unfair advantage by distracting their opponents with their grunts. Indeed, there is a growing chorus of critics who complain that many of the top-ranked professional women tennis players are cheating when they grunt. This complaint has been voiced not only by the media and fans, but also by the athletes themselves (Flatman, 2009; Navratilova, 2009). Indeed, further exemplifying the notion that grunting might distract an opponent, the governing body of the rules of tennis (International Tennis Federation, ITF) explicitly states (rule 26) that purposeful and excessive grunting is a hindrance and reason for a point penalty (International Tennis Federation, 2009).

Unfortunately, the scientific evidence to support these complaints and rules is less than compelling. While there is evidence that performance on a visually based task can be interfered with when a rare and unexpected distracting sound occurs, such as a phone ringing during an exam (Shelton, Elliott, Eaves, & Exner, 2009), a predominant complaint is that tennis players grunt too frequently (i.e., on every shot), so the grunts can hardly be unexpected. Furthermore, there is also evidence that when a sound and visual event occur at different moments and/or locations, attention may be drawn to the sound and away from the visual event (Morein-Zamir, Soto-Faraco, & Kingstone, 2003; Sekuler & Sekuler, 1997). However, that situation does not apply to tennis, as the sound of a grunt and the visual of a ball being struck share a common place in time and space. Accordingly, laboratory research indicates that when audio-visual events share a common origin, they are often integrated, thereby helping to focus attention on the visual event (Calvert & Thesen, 2004; Stein, London, Wilkinson, & Price, 1996). In fact, in some situations a certain degree of temporal and spatial disparity in the multimodal signals can actually be tolerated (Jones & Jarick, 2006). The science therefore suggests that a player who grunts while making a shot may help their opponent focus attention on the shot. This notion can be further bolstered by evidence that having a sudden, short sound can increase one's general level of alertness (Nickerson, 1973). Thus past research does not support the complaint that grunting puts an opponent at a competitive disadvantage.

Nevertheless, it would be the height of arrogance to dismiss the complaints from opponents and experts alike that the grunts have a negative impact. Indeed, past audio-visual studies have generally been limited to detecting flashes of light, and may not apply to the more complex situation of perceiving a tennis ball being struck. To take an initial step toward studying the effect of grunting in tennis we presented videos of a tennis player executing a forehand or backhand groundstroke to the left of right side of the court. Critically, half of the videos included a grunt whereas the other half did not. If the sound of the grunt is indeed distracting, longer response latencies and higher error rates would be expected when participants judged the direction of the tennis shot when a grunt was included.

Method

Participants

Thirty-three undergraduate students from the University of British Columbia participated in exchange for course credit. All reported normal hearing and normal or corrected-to-normal vision.

Materials

Participants sat approximately 60 cm from a computer screen in a dimly lit and sound attenuated testing room. The experiment was programmed and presented using DMDX software (http://www.u.arizona.edu/~jforster/dmdx.htm).

A total of 384 video clips were made of a professional tennis player hitting the ball (either forehand or backhand) to either the left or right of a video camera (Canon ZR10 digital video (DV) camera; 10x optical zoom, 200x digital zoom, image stabilizer, and 460K CCD pixel level) set up on the baseline of the court opposite the player. To be included as a video clip, the player had to hit the ball in a 2 X 2 meter target extending from the sideline and the baseline. The video clips were edited so as to include forehands hit crosscourt and down the line, and backhands hit crosscourt and down the line. There was a total of four clips for each shot type that were then edited such that each clip was played with or without a grunt and ended either at contact or 100 ms after contact. Each clip type (i.e., 32 total for each shot type, total of 128 video clips ranging in length from 1230 ms – 1666 ms) was repeated three times for a total of 384 trials. To mimic the sound of the grunt, while at the same time controlling for individual grunt types, white noise (500 ms; a very conservative and uniform grunt) was played for the last portion of the clips that included the 'grunt'.

Procedure

Participants were required to respond as quickly and accurately as possible indicating the direction of the shot in each video clip (3 blocks of 128 separated by breaks for rest). They were required to use the M key on a keyboard with their right hand if they thought that the shot was going to their right, and the X on a keyboard with their left hand if they thought that the shot was going to their left. Each trial began with a fixation cross (1250 ms), followed by the video. The experiment lasted approximately 25 minutes.

Results

Clips that ended at contact (Hard decision) were analyzed separately from clips ending 100 ms after contact (Easy decision). The data were analyzed for reaction time (RT) and accuracy. When the grunt was present and the video stopped at the time of contact, the participants were consistently 33 ms slower to respond to the direction of the ball (496 ms versus 463 ms; $t(32) = 3.7$, $p = .001$), and they made 4% more decision errors (39% vs. 35%; $t(32) = 2.7$, $p = .012$; see Figure).

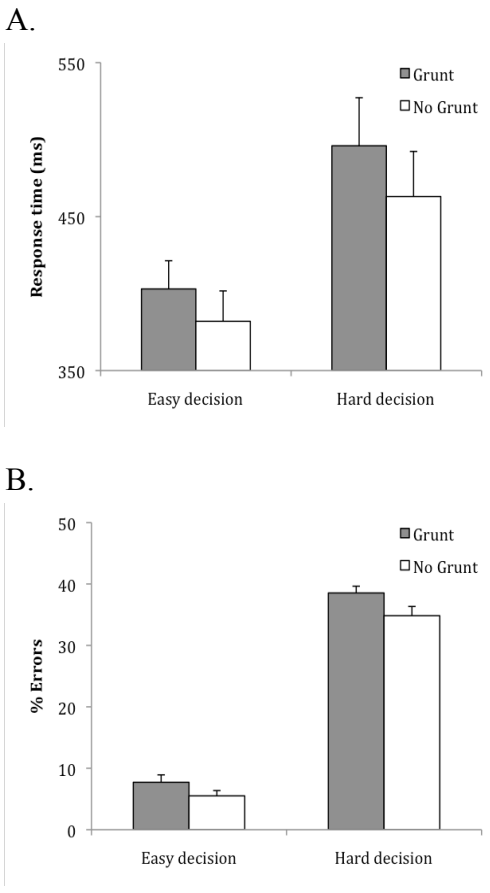


Figure: Dark grey bars represent when the grunt was present and clear bars when the grunt was absent for easy- and hard-shot decisions (A – response time in ms; B – total percentage of decision errors). All differences are significant.

When the video ended 100 ms after contact the exact same pattern was observed. If the grunt was present, participants

were 21 ms slower to respond to the direction of the ball (403 ms vs. 382 ms; $t(32) = 3.7$, $p = .001$), and they made 3% more errors (8% vs. 5%; $t(32) = 3.5$, $p = .001$). That a grunt had the same effect for hard and easy judgements was confirmed by analyses of variance of the overall RT and error data, which directly compared Grunt (Present vs. Absent) and Decision (Hard vs. Easy). The RT results revealed main effects of Grunt and Decision, reflecting the fact that participants were slower to respond when a grunt was present, $F(1,32) = 31.1$, $p < .001$, and the decision was hard, $F(1,32) = 21.8$, $p < .001$, but there was no interaction between grunt and decision, $F(1,32) = 1.74$, $p = .196$. Similarly, for response accuracy, there were more errors for grunts $F(1,32) = 16.0$, $p <.001$, and hard decisions, $F(1,32) = 525.8$, $p < .001$, but no interaction, $F < 1$.

Discussion

The findings are clear-cut. When a grunt occurs opponents are significantly slower (21-33ms) and make significantly more decision errors (3-4%) regarding the direction of the ball both for easy and hard decisions alike. Despite serve speeds now frequently exceeding 100mph (Miller, 2006), if a very conservative estimate that a professional tennis shot travels at 50mph during a rally, a 21-33ms response delay equates to a ball travelling two extra feet on every shot before an opponent can respond. This is a tremendous advantage given that rallies on average last five to seven seconds, with opponents executing generally four directional changes per point with approximately three strokes per rally (the precise values will of course vary with factors like game strategy and court surface; Fernandez, Mendez-Villanueva, & Pluim, 2006). Furthermore, based on data focusing exclusively on 481 matches played at Wimbledon from 1992-1995, an average of 6.4 points played per game can be calculated (Magnus, & Klaasen, 1999). Therefore, between the average number of points played per game and the average number of strokes per point, the additional 3-4% errors observed here could be equivalent to an opponent being wrong footed by a grunting-shot nearly once every game. Given that only four points are required to win a game, this is a definite advantage.

One can only speculate at present as to why a grunt affects the speed and accuracy of responding to a tennis shot. Because a tennis shot and grunt originate from the same location (i.e., the same person), contemporary evidence suggests that visual perception of the shot should be enhanced (see for example Calvert & Thesen, 2004; Sekuler & Sekuler, 1997). Yet we found the opposite. One possibility, suggested by past and present players, is that the sound of a ball making contact with a racket helps to indicate where a shot is going, and a grunt masks this crucial audio-visual integration. We are currently pursuing this issue by manipulating systematically the time of a grunt and the moment that a ball strikes a racket; benchmarking the data against past studies of audio-visual integration. An additional avenue for future research is to manipulate the sound of the grunt and the expertise of the observer. The latter idea is of particular interest, as it might be possible that tennis experts may attempt unique strategies to circumvent the negative impact of a grunt that we have demonstrated here. However, given the self-reports from the tennis players that an opponent's grunting interferes with their play, and our data showing that negative effects of grunting arise for both response latency and accuracy measures regardless of decision difficulty, it is likely that the negative effect of grunting persists for expert tennis players. Indeed, current research suggests that many multisensory phenomena are highly resistant to top-down processes (Driver, & Noesselt, 2008).

It is difficult to ascertain whether many of the most prolific grunters intentionally grunt to distract their opponent. There is little doubt, however, that they are cheating their opponents. Grunting not only decreases their opponent's ability to judge the direction of a shot, it also reduces the amount of time they have to respond to every shot. These consequences on faster tennis surfaces, such as the grass courts of Wimbledon, or the hard courts of the Australian and US Open, are likely to be profound.

References

Calvert, G.A., Thesen, T. (2004). Multisensory integration: methodological approaches and emerging principles in the human brain. *Journal of Physiology*, 98, 191-205.

Driver, J., & Noesselt, T. (2008). Multisensory interplay reveals crossmodal influences on 'sensory-specific' brain regions, neural responses, and judgments. *Neuron*, 57, 11-23.

Fernandez, J., Mendez-Villanueva, A., & Pluim, B.M. (2006). Intensity of tennis match play. *British Journal of Sports Medicine*, 40, 387-391.

Flatman, B. (2009, June 14). One more grunt and you're out: Wimbledon to crack down after complaints. Times Online. Retrieved from http://www.timesonline.co.uk/tol/sport/tennis/article6493899.ece

Jones, J.A., & Jarick, M. (2006). Multisensory integration of speech signals: The relationship between space and time. *Experimental Brain Research*, 174, 588-594.

International Tennis Federation (ITF). Rules of tennis 2009. Retrieved from http://www.itftennis.com/abouttheitf/rulesregs/rules.asp.

Magnus, J.R., & Klaasen, J.G.M. (1999). On the advantage of serving first in a tennis set: Four years at Wimbledon. *The Statistician*, 58, 247-256.

Miller. S. (2006). Modern rackets, balls, and surfaces. *British Journal of Sports Medicine*, 40, 401-405

Morein-Zamir, S., Soto-Faraco, S., & Kingstone, A. (2003). Auditory capture of vision: examining temporal ventriloquism. *Cognitive Brain Research*, 17, 154-63.

Navratilova, M. (2009, June 7). Martina Navratilova: The

grunting has to stop. The Times Online. Retrieved from http://www.timesonline.co.uk/tol/sport/tennis/article6446197.ece

Nickerson, R.S. (1973). Intersensory facilitation of reaction time: Energy summation or preparation enhancement. *Psychological Review*, 80,489–509.

Sekuler, R., Sekuler, A.B., & Lau, R. (1997). Sound alters visual motion perception. *Nature*, 385,308.

Shelton, J.T., Elliott, E.M., Eaves, S.D., & Exner, A.L. (2009). The distracting effects of a ringing cell phone: An investigation of the laboratory and the classroom setting. *Journal of Environmental Psychology*, 29, 513-521.

Stein, B.E., London, N., Wilkinson, L.K., & Price, D.D. (1996). Enhancement of perceived visual intensity by auditory stimuli: A psychophysical analysis. *Journal of Cognitive Neuroscience*, 8, 497-506.

Children's Inductive Inference with Synonymous Labels

Bryan J. Matlen (bmatlen@andrew.cmu.edu)
Department of Psychology, 5000 Forbes Avenue
Pittsburgh, PA 15213 USA

Anna V. Fisher (fisher49@andrew.cmu.edu)
Department of Psychology, 5000 Forbes Avenue
Pittsburgh, PA 15213 USA

Karrie E. Godwin (kegodwin@andrew.cmu.edu)
Department of Psychology, 5000 Forbes Avenue
Pittsburgh, PA 15213 USA

Abstract

Prior research indicates that young children can generalize object properties on the basis of category information when it is conveyed by identical labels or semantically similar labels at the same level of taxonomy (i.e., synonyms) (Gelman & Markman, 1986). However, in previous research semantic similarity was confounded with co-occurrence probability. Therefore, it is possible that synonym-based induction observed in prior research stemmed from children relying on cues other than the semantic similarity of labels. The present study investigated synonym-based induction with labels that do and do not co-occur in child-directed speech. Results indicated that adults made inferences on the basis of the semantic similarity of labels regardless of co-occurrence probability. In contrast, 4-year-old children generalized based on synonymous labels at above chance levels only when synonyms co-occurred in child-directed speech.

Keywords: Labels. Synonyms. Word Learning. Induction. Cognitive Development. Categories.

Introduction

Labels are pervasive in thought. Within the first six years of life, a child may learn up to 14,000 labels (Markman, 1990). It has been suggested that labels convey an object's category, thereby facilitating knowledge generalization (Gelman & Markman, 1986). For example, if a sheepdog has a certain property, would a bulldog be likely to share the same property? Though this question has no definitive answer, one might surmise that, both the sheepdog and the bulldog are likely to possess the property because both are referred to as 'dogs.'

It is well documented that adults rely on category information conveyed by labels to generalize from the known to the unknown, however, it remains contested when children begin to do so. Some research has suggested that children can rely on category information conveyed by labels as early as 13 months of age (Welder & Graham, 2001). Numerous studies have indicated that toddlers and preschool-age children view labels as communicating objects' kind, and that identical labels elicit category-based induction in young children as well as adults (Gelman 1988; Gelman & Coley, 1990; Gelman & Markman, 1986; Jaswal, 2004).

This interpretation has recently been challenged on the grounds that children may treat labels as perceptual attributes of objects rather than as category markers (Sloutsky & Fisher, 2004). Under this view, when two objects share a label, children engage not in *category-based* induction, but instead in *label-based* induction. In other words, children may rely on shared labels in the course of induction not because they understand that labels refer to categories, but because auditory information (including category labels) has a higher attentional weight than visual information early in development (Robinson & Sloutsky, 2004; Sloutsky & Napolitano, 2003). Indeed, there is ample evidence that auditory modality dominates the visual modality in infancy (Lewkowicz, 1994; Robinson & Sloutsky, 2004, 2007) and that these effects extend into early childhood (Napolitano & Sloutsky, 2004). Furthermore, there is evidence that a similarity-based account of early induction (which considers labels to be features of objects contributing to the overall perceived similarity) can readily account for children's reliance on identical labels in the course of property induction as well as categorization tasks (Sloutsky & Fisher, 2004; Sloutsky, Lo, & Fisher, 2001).

Both label-based and category-based accounts predict that children should rely on identical labels during the course of induction. One way to tease apart these two perspectives, is to convey category membership via non-identical semantically similar labels: If it is the case that children perceive labels as windows into categories, then children's generalizations based on semantically similar labels should be similar to their generalizations based on identical labels. If however, children are willing to generalize based on identical labels but not on semantically similar labels, then induction early in development can be label-based without necessarily being category-based.

There are two ways to convey semantic similarity using non-identical labels: by using hierarchically-related labels (e.g., *poodle-dog*) or semantically-similar labels at the same level of taxonomic hierarchy (e.g., *puppy-dog*). For the purpose of brevity, semantically similar labels at the same level of taxonomic hierarchy will be henceforth referred to as *synonyms*. It has been shown that the ability to base

inferences on familiar labels organized into taxonomic hierarchies does not mature until 7- to 8-years of age (Gelman & O'Reilly, 1988; Johnson, Scott, & Mervis, 1997). This finding could suggest that preschoolers' induction with identical labels is unlikely to be category-based. However, it is possible that children's difficulty using hierarchically-related labels stems from the lack of understanding of class inclusion relations, rather than from the lack of understanding that labels denote categories. Indeed, children have been shown to master class inclusion relations by 7- to 8-years of age (Klahr & Wallace, 1972) – the same age at which children can use hierarchically-related labels in the course of induction. The argument presented above suggests that preschool-age children should be successful in performing induction with synonyms, because these labels denote objects of similar kind at the same level of taxonomic hierarchy. At present, however, few studies have examined this possibility.

In a now classic study, Gelman and Markman (1986, Experiment 2) presented 4- to 5-year-old children with triads of pictures consisting of a target item and two test items: one test item looked similar to the target and the other belonged to the same category as the target. Category information was communicated by either identical or synonymous labels. Children were asked to generalize a property from one of the test items to the target. For example, children could be told that a 'rabbit eats bugs' whereas a 'squirrel eats grass', and asked whether the target item (referred to as a 'rabbit" in the Identical Labels condition and as a 'bunny' in the Synonyms condition) 'eats bugs like the rabbit' or 'eats grass like the squirrel.' Gelman and Markman found that children generalized properties to categorically similar items at above chance level in both labeling conditions. Notably, children's performance with synonyms was no different than their performance with identical labels (63% and 67% of category-based responses, respectively).

Gelman and Markman's (1986) study provided support to the notion that children utilize category information conveyed by linguistic labels. However, it has recently been suggested (Fisher, in press) that some label pairs in the Synonyms condition consisted of labels that were not only semantically similar, but also likely to co-occur as compound nouns in child-directed speech (e.g., *bunny-rabbit, puppy-dog*) according to the CHILDES database (MacWhinney, 2000). Co-occurrence of words in natural language has been argued to give rise to strong lexical associations (Brown & Berko, 1960; McKoon & Ratcliff, 1992); therefore in Gelman and Markman's (1986) study it is possible that when children were told that a 'bunny' had a particular property and were asked whether this property would be true of a 'rabbit' or a 'squirrel', children's responses were based not on the understanding that bunnies and rabbits are the same kind of animal, but on the fact that the word 'bunny' primed the word 'rabbit', whereas the word 'squirrel' did not.

A recent study by Fisher (2010) provides preliminary evidence to support this possibility. In this study participants were presented with a label extension task, in which they were taught a familiar label for a novel target object (e.g. "on a different planet, this one is called a *rock*"), and then asked which of the three test objects would likely be referred by a synonymous label (e.g., "which one do you think is called a *stone* on a different planet?"). The three test objects varied in perceptual similarity to the target: one test object looked similar, one looked less similar, and one looked dissimilar. The Co-occurring Synonyms condition included labels that co-occurred in child-directed speech (e.g., *bunny-rabbit*, *puppy-dog*, *kitty-cat*), whereas the Non-co-occurring Synonyms condition included labels that never co-occurred in child-directed speech in the CHILDES database (e.g., *rock-stone, couch-sofa, child-kid;* MacWhinney, 2000). Fisher found that adults and six-year-old children inferred that objects referred to by synonymous labels were likely to look similar, exhibiting a high proportion of choices of similar test items in both labeling conditions. In contrast, 4 year-old children were more likely to choose similar test items in the Co-occurring Synonyms condition than in the Non-co-occurring Synonyms condition. Moreover, young children's performance in the Non-co-occurring condition did not exceed chance.

The present study was designed to directly examine the possibility that label co-occurrence may play a role in inductive generalization. Four-year-old children and adults participated in a triad induction task; on half of the trials participants were asked to make inferences based on non co-occurring synonyms and on the other half of the trials participants made inferences based on co-occurring synonyms. An Identical Label condition was also included as a control condition

Method

Participants

Participants were 33 4-year-old children (M = 4.52 years, SD = .40 years, 18 females, 15 males) recruited from local preschools and 30 undergraduate students from a local university who received partial course credit.

Design

The experiment had a 2 (Label condition: Synonymous vs. Identical Labels) by 2 (Co-occurrence condition: Non-co-occurring vs. Co-occurring Labels) by 2 (Age: Preschoolers vs. Adults) mixed design. Labeling condition was a between-subject factor: participants were randomly assigned either to the Synonymous or Identical Labels condition. Co-occurrence probability of labels was a within-subject factor: every participant performed induction both with co-occurring and non-co-occurring labels.

Materials

Language materials consisted of nine label triads, with each triad comprised of a target item, a semantically related test

item and an unrelated test item. Related test items could be conveyed either by identical or by semantically similar labels (in the Identical and Synonymous Labels conditions, respectively). Unrelated items consisted of labels that a separate group of adult participants judged to be unrelated to the target items (see details below). To-be-generalized properties consisted of two-syllable blank predicates. A full list of linguistic stimuli is provided in Table 1.

Visual stimuli consisted of three sets of doors, with each set including three identical doors. Participants were told that objects were hiding behind each of the doors. This procedure was used to provide participants with conditions that were maximally favorable to relying on semantic information conveyed by labels as there was no perceptual conflict that participants had to resolve to perform category-based induction. Since visual stimuli were identical, category information conveyed by labels was the only basis for induction. Additionally, a set of 27 pictures was used for a Picture Identification task that all children completed after the experiment proper. The goal of this task was to ensure that children were familiar with all of the labels used in this study, and that children were willing to use semantically similar labels to refer to the *same* object (see the Procedure section below for details).

Label Selection

Assignment of label pairs to the Co-occurring and Non-co-occurring conditions was similar to the procedure used in Fisher (2010). Five different databases in the CHILDES corpus were analyzed (i.e., Bates, Brown, Gleason, HSLLD, and Wells). Children's ages ranged from 1 ½ to 9 years, and, across all databases a total of 2,264,722 words were included. To obtain normalized co-occurrence scores, the number of raw co-occurrences was divided by the sum of instances of each word occurring individually minus the number of times the two words co-occurred. For example, the word "kitty" occurred in the analyzed databases 847 times, the word "cat" occurred 2,319 times, and these words co-occurred 131 times. Using the normalization procedure the probability of the words "kitty" and "cat" co-occurring was calculated as 131 ÷[847+2,319–131] = .04.

Four co-occurring synonyms were selected based on their above-zero co-occurrence probability and their likelihood of being known to young children. Because all four co-occurring label-pairs referred to natural kinds, only non-co-occurring synonyms referring to natural kind objects were selected for this study. We did not use some of the non-co-occurring label pairs used by Gelman and Markman (1986) (e.g., *cobra-snake* and *rose-flower*) because these labels were hierarchically related, and thus unlikely to generate category-based induction in 4-year-old children (Gelman & O'Reilly, 1988; Johnson, Scott, & Mervis, 1997).

Overall, the average co-occurrence probability of synonyms was .033 in the Co-occurring condition and .000 in the Non-co-occurring condition, independent-samples $t(6) = 2.26$, $p = .03$. Unrelated test items were also labels that referred to natural kind objects. Unrelated test items were matched in syllable length to the related items for all triads except one[1].

A separate calibration study was conducted with an independent group of 22 adults to establish semantic similarity of labels within each triad. Adults were asked to rate semantic similarity of the Target items to the Related and Unrelated test items (e.g., *rock-stone*, *rock-cloud*, and *stone-cloud*) on a scale of 1 – 7, with 7 indicating that the labels could be used interchangeably, and 1 indicating that the labels had no overlap in meaning. Results of this calibration confirmed that targets and related test items (i.e. synonyms) were more semantically similar ($M = 6.3$) than targets and unrelated test items ($M = 2.8$), $t(14) = 11.43$, $p < .001$. There were no differences found when the analysis was separated by co-occurrence condition, $F(1, 15) < 1$, *ns*.

Table 1: List of stimuli and co-occurrence probabilities of semantically similar labels.

Target Items	Related Test Items	Unrelated Test Items	Blank Predicates	Co-Occ Prob
Rock	Stone	Cloud	Higa	.000
Dolphin	Whale	Seal	Omat	.000
Alligator	Crocodile	Hippo	Matlen	.000
Toad	Frog	Bird	Koski	.000
Mouse	Rat	Duck	Lignin	.000
Puppy	Dog	Cow	Erwin	.010
Kitty	Cat	Pig	Manchin	.040
Bunny	Rabbit	Squirrel	Creighan	.070
Pony	Horse	Fox	Troxel	.01

Procedure

Children were tested individually at their daycares in a quiet room or hallway. Adults were tested individually in a laboratory on campus. Visual stimuli were presented on a computer and labels were provided verbally by experimenters.

Labels used in the Synonyms condition are displayed in Table 1. The same set of labels was used in the Identical condition with the exception that the Target items and Related Test items were referred by identical labels (e.g., *rock-rock* for half of the participants and *stone-stone* for the other half of the participants). Half of the participants participated in the Co-occurring condition first, and the other half participated in the Non-co-occurring condition first. Within each co-occurrence condition trials were presented in one of two random orders. The *rock-stone-cloud* triad always appeared first as served as an instructional trial for all participants; the data from this trial were not included in the analyses reported below.

Participants were told that they would be playing a game about objects that were hiding behind doors (see Figure 1). The experimenter told participants what object was hiding

[1] Due to 4-year-old parlance, hippo was included as a lure for alligator-crocodile, despite it not matching the number of syllables of the related test item.

behind each door. The Target objects were always hidden behind the topmost door, and the location of the Related and Unrelated Test objects (to the left or to the right of the Target) was randomized across trials. The experimenter first introduced the Target item (e.g., *There is a rock hiding behind this door*) and then introduced the Test items in random order (e.g., *There is a cloud hiding behind this door*. *There is a stone hiding behind this door*). Then participants were told about the property of the Target item (e.g., *The rock behind this door has higa inside*) and asked to generalize this property to one of the Test items (e.g., *Do you think that the cloud behind this door or the stone behind this door also has higa inside?*).

Additionally, participants were asked to remember where each object was hiding. The memory check was included to ensure that possible differences in induction performance could not be attributed to children's better memory for co-occurring than for non-co-occurring synonyms. After the induction response was recorded, a memory check was performed: the experimenter asked the participant if (s)he remembered what was hiding behind each door, pointing to the doors in random order.

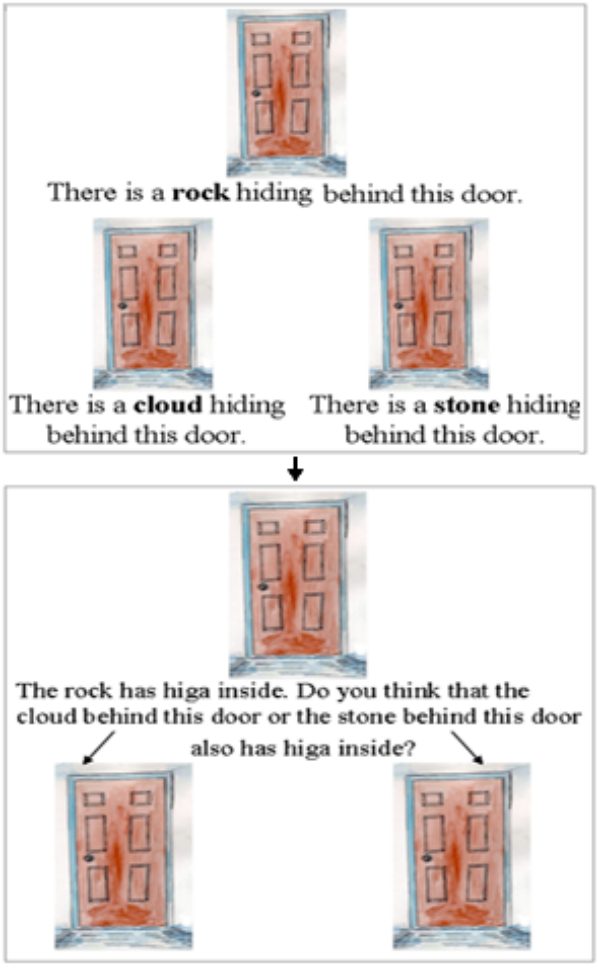


Figure 1: Schematic depiction of door task.

After the induction task children (but not adults) participated in a Picture Identification task similar to the Peabody Picture Vocabulary Test (Dunn & Dunn, 1997). The goal of this task was to confirm that children were familiar with each of the labels used in the induction task. In each trial children were presented with four different pictures, and asked to point to the target picture that was verbally indicated by the experimenter (e.g., "*can you find the rock?*"). Target items in the Picture Identification consisted of all Target, Related, and Unrelated labels that were included in the induction task. Importantly, knowledge of synonymous labels was always tested using *identical* pictures on separate trials (with location of the correct response counterbalanced across trials). There were 18 unique pictures of interest (for the two labels in eight experimental and one instructional trial in the induction task). Pictures testing knowledge of synonyms were presented twice and pictures testing knowledge of the unrelated items were presented once, resulting in a total of 27 trials in the picture identification task.

Results

Preliminary analyses revealed no effects of block order (all p's > .20). In the Picture Identification task, children's accuracy was .99 in each Label condition, indicating that 1) children were familiar with the words used in the experiment proper, and that 2) children could readily apply synonymous labels to the *same* objects.

Induction Accuracy

Proportions of category-based responses (i.e., choices of identical or synonymous labels) were analyzed in a 3-way mixed ANOVA, with Label condition and Age group as between-subject factors and Co-occurrence condition as a within-subject factor. The analysis revealed a significant effect of Age, $F(1, 58)=29.57$, $p<.001$; a significant interaction between Co-occurrence and Age $F(1, 59)=5.58$, $p<.05$; and a significant three-way interaction $F(1, 59)= 4.41$, $p<.05$. Follow-up analyses revealed no differences among conditions for adults (all p's >.63). Adults' category-based responding was above chance in all conditions (all p's < .001) (means in all conditions were ≥ .97, SD's ≤ .09).

Proportions of children's category-based responses are presented in Figure 2. For children there was a reliable difference in performance between the Non-Co-Occurring Synonyms and the Co-occurring Synonyms conditions, paired-sample $t(16) = 3.45$, $p <.005$ (M = .52 and .74, respectively). Within the Non-co-occurring condition, there was also a reliable difference between the Synonymous and Identical Label conditions, independent-sample $t(31) = 2.41$, $p <.05$ (M = .52 and .75, respectively). Furthermore, children's performance in the Non-co-occurring Synonyms condition did not exceed chance, one-sample $t(17) = .20$, *ns*., whereas performance in all other conditions was above chance (all one-sample t's > 2.54, p's <.05). There were no differences in children's performance with Identical Co-occurring and Identical Non-co-occurring labels, paired-samples $t(15) = .53$, *ns*, (M=.75 and .70, respectively).

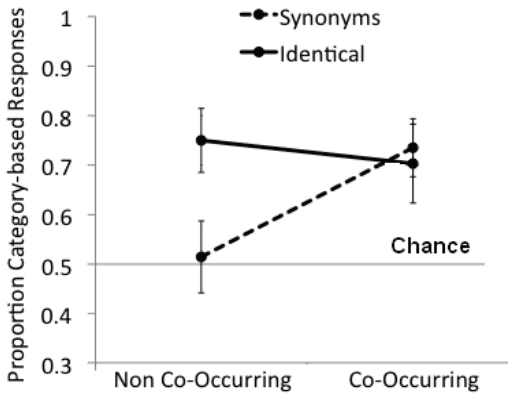


Figure 2. Proportion of category-based responses in children.

To investigate performance at an individual level, we classified participants into category-based and non-category-based responders. A category-based responder was defined as a participant who provided at least 3 of 4 category-based responses within each condition (see Figure 3). Individual response patterns mirrored the group data summarized above. In particular, all adult participants in all conditions were classified as category-based responders. In the Co-Occurring labels condition the majority of 4-year-olds were also classified as category-based responders: 11 out of 17 (65%) in the Identical labels condition and 10 out of 16 (63%) in the Synonyms condition (this association was not significant, Fisher's exact $p > .99$). Similarly, in the Non-Co-Occurring/Identical labels condition the majority of children were classified as category-based responders: 12 out of 16 (75%). However, in the Non-Co-Occurring/ Synonyms condition only 6 out of 17 children (35%) were classified as category-based responders. The association between condition and response type in the Non-Co-Occurring/Synonyms and Non-Co-Occurring/Identical labels condition was significant, Fisher's exact $p < .05$.

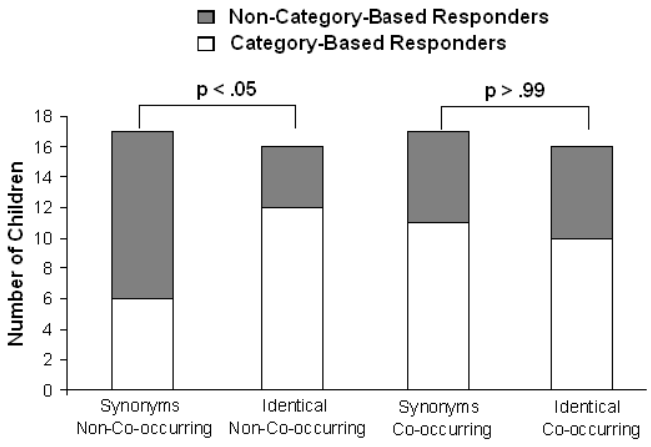


Figure 3. Number of children classified as category-based and non-category-based responders.

Memory Accuracy

Adults' overall memory scores were 99% in each label condition. Children's overall memory scores were 86% and 84% in the Synonymous and Identical Label conditions, respectively. Children's memory was well above chance level of 33% in both conditions, both one-sample p's $< .001$, indicating that children had little difficulty with the memory demands of the task. Most importantly, there was no difference in children's memory performance when separated by co-occurrence condition (86% and 85% correct in the Co-occurring and Non Co-occurring Synonyms conditions, respectively), paired-sample $t(32) < 1$, *ns*. A linear regression performed on children's memory scores and their induction performance revealed no significant relationship in the Synonyms condition, $r^2(134) = .055$, $p > .50$, or the Identical condition, $r^2(126) = .019$, $p > .95$.

Discussion

Contrary to the notion that the ability to perform synonym-based induction is well established by four years of age, the results of the present study suggest that this ability still undergoes development during the preschool years. In particular, 4-year-old children performed at chance in the triad induction task when semantically similar label-pairs did not co-occur in child-directed speech (e.g., *alligator-crocodile*). However, we observed a significant improvement in performance when children were presented with semantically similar labels that co-occurred in child-directed speech (e.g., *bunny-rabbit*) and with identical labels (e.g., *bunny-bunny* and *alligator-alligator*).

These findings are not easily explained by children's unfamiliarity with some of the words used in this research as our participants exhibited near ceiling accuracy on the picture identification task. Importantly, children readily applied different words with shared meaning (e.g., *alligator-crocodile*) to the *same* items in the picture identification task. Therefore, children clearly possessed the requisite knowledge to perform synonym-based induction. Yet, few 4-year-old children spontaneously relied on this knowledge in the induction task, unless the labels not only shared meaning but also co-occurred in child-directed speech.

Results reported in this paper suggest that poor understanding of class-inclusion relations is not the sole reason why preschool-age children fail to utilize taxonomic labels (e.g., *animal-cat*) in the course of induction tasks. The present findings add to the growing body of evidence suggesting that the understanding that labels refer to categories matures gradually between four and seven years of age (Fisher, 2010; Fisher & Sloutsky, 2005; Matlen & Fisher, 2008). In particular, consistent with the results reported in this paper, Matlen and Fisher (2008) found that only 15% of 4-year-old children spontaneously performed synonym-based induction with labels that did not co-occur in child-directed speech; this number increased to 51% of 5-year-olds. By 6 years of age the majority of children (86%) readily relied on synonymous labels to perform induction.

The present study is the first to demonstrate the effect of label co-occurrence on induction using a within-subject design. Therefore, this study provides direct evidence that results of earlier research on the development of synonym-based induction could stem from the fact that responses were averaged across items that were likely to result in above-chance performance (e.g., *bunny-rabbit, puppy-dog*) and items that were unlikely to result in above-chance performance (e.g., *rock-stone*, *cobra-snake*). It is conceivable that overall results aggregated over a bimodal distribution of responses could result in a mean proportion of synonym-based responses that exceeded chance level (i.e., 63%; Gelman & Markman, 1986; Experiment 2). Indeed, when children's responses in the Synonym condition of the present study were aggregated across both co-occurrence conditions, the average proportion of category-based responses was .63, above chance, paired-sample $t(16) = 2.17$, $p < .05$.

In sum, the results presented in this paper provide evidence that preschoolers' willingness to rely on semantically similar labels in the course of induction is influenced by co-occurrence probability of these labels in child-directed speech. This finding poses a challenge to the theoretical approach that assumes children's induction to be category-based from very early in development. At the same time, these results are consistent with the approach suggesting that the development of category-based induction follows a protracted developmental course.

Acknowledgments

We thank Malika Sinha, Sarah Shade, Janelle Higa, Tina Li, Stephanie Brown, and Kimberly Dye for help in data collection and the CHILDES analysis. We also thank the children, parents, teachers, and administrators for their participation. This research was supported by the Institute of Education Sciences, U.S. Department of Education through Grant R305B040063 to Carnegie Mellon University.

References

Brown, R. & Berko, J. (1960). Word association and the acquisition of grammar. *Child Development*, 31, 1-14.

Fisher, A. V. (2010). What's in the name? Or how rocks and stones are different from dogs and puppies. *Journal of Experimental Child Psychology*, 105, 198-212.

Fisher, A. V., & Sloutsky, V. M. (2005). When induction meets memory: Evidence for gradual transition from similarity-based to category-based induction. *Child Development*, 76, 583-597.

Gelman, S.A. & O'Reilly, A.W. (1988). Children's inductive inferences within superordinate categories: The role of language and category structure, *Child Development*, *59*, 876–887.

Gelman, S.A. & Coley, J.D. (1990). The importance of knowing a dodo is a bird: Categories and inferences in 2-year-old children. *Developmental Psychology, 26*, 796-804.

Gelman, S.A., & Markman, E. (1986). Categories and induction in young children. *Cognition*, *23*, 183-209.

Graham, S.A., Kilbreath, C.S., & Welder, A.N. (2004). Thirteen-month-olds rely on shared labels and shape similarity for inductive inferences. *Child Development*, 75, 409–427.

Jaswal, V.K. (2004). Don't believe everything you hear: Preschoolers' sensitivity to speaker intent in category induction. *Child Development*, 3, 279-300.

Klahr, D. & Wallace. Class inclusion processes. In S. Farnham-Diggory (Ed.), *Information processing in children*. New York: Academic Press.

Lewkowicz, D. J. (1994). The development of intercessory perception in human infants. In D. J. Lewkowicz & R. Lickliter (Eds.), *The development of perception: Comparative perspectives*. Hillsdale, NJ: Erlbaum.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk*. Third Edition. Mahwah, NJ: Lawrence Erlbaum Associates.

Matlen, B.J. & Fisher, A.V. (2008). Development of Synonym-Based Induction. Poster presented at the *XXX Annual Meeting of the Cognitive Science Society*, Washington, DC.

Markman, E. (1990). Constraints children place on word meanings. *Cognitive Science, 14*, 57-77.

McKoon, G. & Ratcliff, R. (1992). Spreading activation versus compound cue accounts of priming: Mediated priming revisited. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18, 1155 – 1172.

Napolitano, A. C., & Sloutsky, V. M. (2004). Is a picture worth a thousand words? Part II: The flexible nature of modality dominance in young children. *Child Development*, 75 (6), 1850-1870.

Robinson, C.W., & Sloutsky, V.M. (2004). Auditory Dominance and its change in the course of development. *Child Development*, *75*, 1387-1401.

Robinson, C. W., & Sloutsky, V. M. (2007). Linguistic labels and categorization in infancy: Do labels facilitate or hinder? *Infancy*, 11(3), 223-253.

Sloutsky, V.M., & Fisher, A.V. (2004). Induction and categorization in young children: A similarity-based model. *Journal of Experimental Psychology: General, 133*, 166-188.

Sloutsky, V. M., & Napolitano, A. (2003). Is a picture worth a thousand words? Preference for auditory modality in young children. *Child Development*, *74*, 822-833.

Sloutsky, V.M., Lo, Y.-F., & Fisher, A.V. (2001). How much does a shared name make things similar? Linguistic labels and the development of inductive inference. *Child Development*, *72*, 1695-1709.

Welder, A.N., & Graham, S.A. (2001). The influences of shape similarity and shared labels on infants' inductive inferences about nonobvious object properties. *Child Development*, *72*, 1653-1673.

Development of Relational Reasoning with Semantically Similar Labels

Sheela Ramesh (sheelar@andrew.cmu.edu)
Department of Psychology, Baker Hall 342c, 5000 Forbes Ave
Pittsburgh, PA 15213 USA

Bryan J. Matlen (bmatlen@andrew.cmu.edu)
Department of Psychology, Baker Hall 445b, 5000 Forbes Ave
Pittsburgh, PA 15213 USA

Anna V. Fisher (fisher49@andrew.cmu.edu)
Department of Psychology, Baker Hall 342c, 5000 Forbes Ave
Pittsburgh, PA 15213 USA

Abstract

The present study explored children's ability to utilize synonymous labels during relational reasoning. In Experiment 1, 4- to 5-year-old children and adults were presented with a base pair of related words (e.g., Castle:Rock) and then were given a partially completed target word-pair (Castle:?) that they could complete with a label that made the target word-pair relationally identical to the base word-pair (e.g., Stone). Additional response options included a label thematically related to the first word in the target pair (e.g., King) or an unrelated word (e.g., Milk). Results indicated that adults and 5-year-olds successfully completed the task, whereas 4-year-olds exhibited difficulty. In Experiment 2, 4-year-old children were presented with the same task, however relations were conveyed by identical rather than synonymous labels. Under these conditions, 4-year-old children exhibited no difficulty in either lure condition. These findings are discussed with regards to the theories of learning early in development.

Keywords: Synonyms; Language Acquisition; Word Learning; Relational Reasoning; Cognitive Development.

Introduction

Many objects in the world can be referred to by more than one label, a phenomenon called polyonomy. For example, one could accurately refer to a pet as *Fluffy*, *kitty*, *cat*, and *animal*. It is well-documented that in the beginning stages of language acquisition children struggle with this phenomenon; however by three years of age children are able to learn multiple labels in reference to the same object, both in the form of taxonomically-related labels (such as *cat-animal*) and semantically similar labels at the same level of taxonomic hierarchy (i.e. synonyms, such as *kitty-cat*) (Banigan & Mervis, 1988; Blewitt, 1994; Deák & Maratsos, 1998; Haryu & Imai, 1999; Johnson, Scott, & Mervis, 1997; Liitschwager & Markman, 1994; Mervis et al., 1994). Learning to refer to an object by more than one label may signify development of understanding that labels denote categories rather than individual objects. However, mature understanding of labels as category markers requires that one not only can use multiple labels in reference to the same object, but is also able to rely on related labels to perform a variety of reasoning tasks, such as categorization, inductive reasoning, and analogical reasoning.

Research investigating development of the ability to use hierarchically-related labels in reasoning tasks indicates that this ability does not mature until 7- to 8-years of age (Gelman & O'Reilly, 1988; Johnson, et. al., 1997). However, research into children's ability to use semantically similar labels at the same hierarchical level – or synonyms – in reasoning tasks has produced mixed results. In particular, Gelman and Markman (1986) observed that 4-year-old children can perform inductive reasoning tasks at above chance level with identical labels (e.g., generalizing a property from one rabbit to another rabbit, rather than from a squirrel to a rabbit) as well as semantically similar labels (e.g., generalizing a property from a bunny to a rabbit, rather than from a squirrel to a rabbit).

However, while Gelman and Markman's study provides valuable insight into children's reasoning with semantically similar labels, several factors warrant further investigation of this phenomenon. First, some stimuli used in this study included taxonomically-related labels (e.g. *rose-flower* and *cobra-snake*) rather than synonyms. Second, some of the semantically similar labels used in this study are likely to co-occur in the speech of children and their caregivers as compound noun-phrases (e.g., *bunny-rabbit* and *puppy-dog*) according to the CHILDES database (MacWhinney, 2000). For example, the word "bunny" occurred in CHILDES 803 times, the word "rabbit" occurred 579 times, and these words co-occurred 103 times. At the same time, other semantically similar labels used in the Gelman and Markman (1986) study (e.g., *rock-stone*) *never* co-occurred in the CHILDES database. It is possible that effects of semantic similarity of labels on inductive reasoning in 4-year-old children were amplified by co-occurrence probability of some of the label pairs used in this research. In support of this hypothesis, Matlen and Fisher (2008) found that 4-year-old children successfully relied on semantically similar labels in a property induction task only if these labels were likely to co-occur in child-directed speech, whereas children's performance with non-co-occurring semantically similar labels was not different from chance. Therefore, the extent to which young children can rely on semantically similar labels in the course of reasoning tasks remains unclear.

The goals of the research reported below were two-fold. The first goal was to explore to what extent young children are capable of using semantically similar labels that span the same level of taxonomic hierarchy as a basis for reasoning. Prior research on this topic has primarily been concerned with children's reasoning within property induction tasks (Gelman & Markman, 1986; Matlen & Fisher, 2008). To assess the robustness of children's reasoning with semantically similar labels, we employed a relational reasoning task where relations were conveyed by synonyms. In contrast to property induction tasks, relational reasoning tasks have typically been utilized in research aimed at assessing children's analogical thinking (see Goswami 1991 for review). These tasks tend to follow the format of A:B::C:?. For example, Goswami and Brown (1990) assessed 4- and 5-year-olds' ability to perform analogical reasoning tasks with familiar relations by presenting them with a base word-pair (e.g., Spider-Web) and an incomplete target word-pair (e.g., Bee-?). Children could complete the target word-pair with a relational choice (e.g., Hive), or with a word that did not preserve the relation specified in the base word-pair: a thematic lure (e.g., Honey). Goswami and Brown found that by four years of age children could correctly complete these analogies based on the relational choice, even in the presence of a thematic lure. It follows then that if young children possess the ability to reason with semantically similar labels, then they should be able to correctly complete relational reasoning tasks when semantically similar labels convey the relations.

The second goal of the present research was to examine children's understanding of linguistic labels as markers of category membership. It has been suggested that understanding of labels as category markers develops as early as two years of age (Gelman & Coley, 1991; Welder & Graham, 2001). Therefore, children realize that objects referred to by the same label – or by semantically similar labels – refer to objects of the same kind, and that objects of the same kind are likely to have many properties in common. However, it has recently been argued that understanding of labels as referents to categories may have a more protracted developmental course than previously believed (Sloutsky, Lo, & Fisher, 2001; Sloutsky & Fisher, 2004). In particular, children may rely on identical labels in reasoning tasks because labels are features contributing to the overall similarity of presented entities, and identical labels increase the *perceived* similarity of compared objects. At the same time, children may not rely on semantically similar labels in the course of reasoning tasks, unless there are factors other than shared meaning (such as co-occurrence probability) that promote label-based inference.

To achieve the goals outlined above, the present study utilized a modified analogical reasoning task of the type employed by Goswami and Brown (1990). Specifically, participants were presented with a base word-pair relation (e.g. Castle:Rock), and a partially completed target word-pair relation (Castle: ?). Participants could complete the target word-pair with a label that preserved the relationship specified in the base word-pair (i.e., a label semantically similar to the second term in the base word-pair, such as "Stone") or with a word that did not preserve the relation specified in the base word-pair: a word thematically related to the first term in the target word-pair (e.g., King; in the Thematic Lure condition) or an unrelated word (e.g., Milk; in the Unrelated Lure condition). Thus, this task followed an A:B::A:B' format (where B'-term was semantically similar to the B-term). Children's ability to perform the relational reasoning A:B::A:B' task using semantically similar labels was compared to their ability to perform this task using identical labels.

If children have acquired mature understanding that identical as well as semantically similar labels refer to objects of the same kind (Gelman & Markman, 1986; Gleman & Coley, 1991; Jaswal, 2004), then children should have little difficulty in completing the relational reasoning A:B::A:B' task using identical as well as semantically similar labels. However, if understanding that labels refer to categories has a protracted developmental course and is not yet complete by four years of age (Fisher, in press; Matlen & Fisher, 2008; Sloutsky & Fisher, 2004) then 4-year-old children, unlike adults, may have difficulty in completing relational reasoning tasks using semantically similar labels. At the same time, children should succeed in completing relational reasoning tasks using identical labels.

Experiment 1

Method

Participants

Participants were 35 four-year-old children (M = 4.43 years, SD = .30 years; 19 females and 16 males), 30 five-year-old children (M = 5.5 years, SD = .32 years; 12 females and 18 males), and 45 adults (M = 19.94 years, SD = 1.14 years; 26 females and 19 males).

Design

For the purpose of brevity, semantically similar labels will be referred to as "synonyms" henceforth. Experiment 1 had a two (Lure type: Thematic vs. Unrelated) by four (Age: 4-year-old, 5-year-olds, and adults) between-subjects design. Participants were randomly assigned to one of the two experimental conditions: the Thematic Lure or the Unrelated Lure condition.

Materials

Materials consisted of 12 picture sets presented on a computer screen, accompanied by 12 label sets provided by the experimenter (see Table 1 for the list of labels used in the experiment). Each picture set consisted of a series of four pictures: the first picture contained four doors in two rows of two. One by one, the first three doors disappeared to reveal objects hidden behind them (see Figure 1 for a schematic depiction of the task). As the doors disappeared to reveal hidden objects, the experimenter labeled each object. The objects behind the first and second doors had a

clear relationship, and the object behind the third door was identical in picture and label to the object behind the first door. The fourth door revealed no hidden object, but instead the experimenter provided two response options for the participant to guess the final object.

Table 1: Labels provided during task.

A-term	B-term	B'-term: Synonym Choice	Thematic Lure	Unrelated Lure
Bread	Jam	Jelly	Crumbs	Foot
Hand	Mitten	Glove	Foot	Ant
Castle	Stone	Rock	King	Milk
Cat	Couch	Sofa	Milk	Banana
Fly	Toad	Frog	Ant	Phone
Cheese	Rat	Mouse	Cracker	Dress
Apple	Belly	Tummy	Banana	Puppy
Duck	Lake	Pond	Feathers	TV
Vacuum	Carpet	Rug	Mop	Lion
Water	Ship	Boat	Fish	Cookie
Beach	Ocean	Sea	Sand	Chair
Car	Road	Street	Steering Wheel	Clock

In both the Unrelated Lure and the Thematic Lure conditions the relational choice was communicated by a label that was synonymous to the B-term of the base word-pair. For example, in the trial where the base word-pair consisted of the words "Castle:Rock", the relational choice consisted of the word "Stone". Note that half of the participants received the base word-pair of "Castle-Rock" with the word "Stone" being the relational response option, whereas the other half of the participants received the base word-pair of "Castle-Stone" with the word "Rock" being the relational choice (the B- and B'-terms alternated in this manner for all of the trials in this and other experiments described in this paper). To avoid the potential confound of co-occurrence probability influencing children's responses, only synonyms that never co-occurred in child-directed speech according to the CHILDES database (MacWhinney, 2000) were chosen for this study. Thus, common synonym pairs used in prior research, such as puppy-dog and bunny-rabbit, were not utilized. The outcome measure was the proportion of relational responses (i.e. choosing synonymous labels over thematic and unrelated lures) across the 11 experimental trials.

Calibration of Experimental Materials Experimental materials used in this research were calibrated in several separate studies to establish that (1) 4-year-old children were familiar with all of the labels used in this research and were willing to apply synonymous labels to the same object, (2) 4-year-old children were familiar and could identify the relationship specified by the A-term and the B-term labels, and (3) 4-year-olds children perceived labels chosen as thematic lures to be thematically related to the A-term labels.

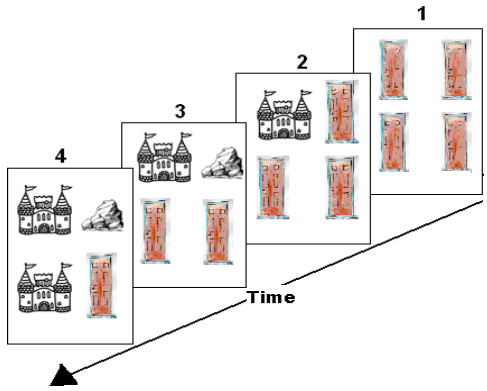


Figure 1: Schematic depiction of the A:B::A:B' task.

Label Calibration A group of four-year-old children, none of whom participated in the experiment proper (N = 12, M = 4.28 years) was presented with a picture-naming task analogous to the Peabody Picture Vocabulary Test (Dunn & Dunn, 1997). Participants were presented with a series of pictures on a computer screen, four pictures at a time, with one target picture for children to identify (location of all target pictures was counterbalanced across multiple presentation of the same picture) and three distracters. Children were asked to select the target picture according to the label spoken by the experimenter (e.g., "Can you point to the *rock*?"). Each of the 12 target pictures was presented to children in two separate picture sets, resulting in a total of 24 trials. The target pictures were identified by different synonymous labels during the first and second presentation of each target picture Children correctly identified pictures referred to by the B- and B'-terms used in Experiment 1 (see Table 1) with the overall accuracy of 97%. Therefore, the synonym labels used in Experiment 1 were familiar to four-year-old children. Importantly, four-year-old children were willing to accept the synonymous labels used in Experiment 1 as referents to the *same* objects.

A:B Relationship Calibration The same group of 4-year-old children who participated in the Label Calibration were presented with the relation familiarity check (the order of these tasks was counterbalanced across participants). Children were presented with a series of triads depicting objects that can be referred to by an A-term, an unrelated lure, and a B-term or a synonymous B'-term (see Table 1). Children were provided verbal labels for all three objects and asked to select the two objects that "go together." Half of the participants were asked about a B-term (e.g., a *rock*) and the other half about the synonymous B'-term (e.g., a *stone*). For instance, participants could be shown a triad consisting of pictures of a 'rock', a 'castle', and 'milk', and asked which objects go together. Additionally, children were asked to explain why the two pictures they had chosen "go together," and their explanations were recorded. All of the B- and B'- presented in Table 1 were judged to be identifiably related to the A-terms by 4-year-old children: children selected the B- or the B'-term over the unrelated lure and correctly specified the nature of the A:B

relationship with the average accuracy of 95% (all individual item *M*s were above 83%, above chance, all one-sample *t*s > 4.4, *p*s < .001).

Calibration of Thematic Lures A separate group of 4-year-old children (N = 12, M = 4.77 years) was presented with a series of triads depicting objects referred to by the A-term, thematic lures, and unrelated lures (see Table 1) and asked to select the two objects that "go together." For instance, participants could be shown a triad depicting a 'castle', a 'king', and 'milk', and asked which objects go together. Children selected the thematically related lure as the objects that "goes with the target" over the unrelated lure with the mean accuracy of 94% (all individual item *M*s were above 91%, above chance, all one-sample *t*s > 7.0, *p*s < .001).

Procedure

Children were tested individually in a quiet room in their preschools and adults were tested in a laboratory on campus. Participants were presented with pictures of four doors in two rows of two on a computer screen. The experimenter explained that there were objects hiding behind all of the doors, and that after showing the objects behind the first three doors, the participant would have to guess what was hiding behind the last door. The word-pair "Bread:Jam" served as a practice trial (see Table 1) and thus was always presented first. The order of the rest of the trials was randomized for each participant. When pictures of bread and jam were revealed during the practice trial, participants were told, "*bread and jam go together because jam goes on bread to make a sandwich.*" Participants were then asked to guess what object was hiding behind the fourth door and presented with two response options; participants were asked to choose the option that "*goes with the bread the same way that jam goes with the bread*". Upon completing the practice trial children were provided with corrective feedback. No feedback was provided after the experimental trials. At the conclusion of the practice trial children were told that they would keep playing the game, and that to solve the task they needed to think how the objects behind the first two doors go together.

Results

Proportions of relational responses in each condition are presented in Figure 2. A 2 x 2 Analysis of Variance was performed on the proportions of relational responses with experimental condition (Thematic Lure vs. Unrelated Lure) and age (4-year-olds, 5-year-olds, and adults) as between-subject factors. The results indicated the main effect of condition, F (1, 104) = 10.9, p = .001, and age, F (2, 104) = 50.7, p < .001. Performance was significantly higher in the Unrelated Lure condition (M = 89.8%) than in the Thematic Lure condition (M = 80.1%), F (1, 104) = 10.9, p = .001. These main effects were qualified by a condition by age interaction, F (2, 104) = 3.4, p < .05.

Post hoc Tukey tests indicated that performance increased significantly from 4- to 5- years of age (p < .001), and again from 5-years of age to adulthood (p < .05). Planned comparisons revealed that performance of adults was equivalent in the Thematic Lure and Unrelated Lure conditions (both means over 99%). However, 5-year-old children exhibited a higher level of performance in the Unrelated Lure condition compared to the Thematic Lure condition (97% and 83% of relational responses, respectively), independent samples t(28) = 2.98, p < .01. This difference in performance between the lure conditions was marginally significant in 4-year-old children (73% and 58% of relational responses, respectively), independent-samples t(33) = 1.9, p = .07.

Follow-up comparisons to chance indicated that in the Unrelated Lure condition participants in all age groups responded at a level above chance (chance = 50%), all one-sample *t*'s > 4.5, *p*'s < .001. However, in the Thematic Lure condition only 5-year-olds and adults responded at above chance level, both one-sample *t*'s > 7.6, *p*'s < .0001, whereas responses of 4-year-old children were not different from chance, one-sample t(16) =1.2, p = .25.

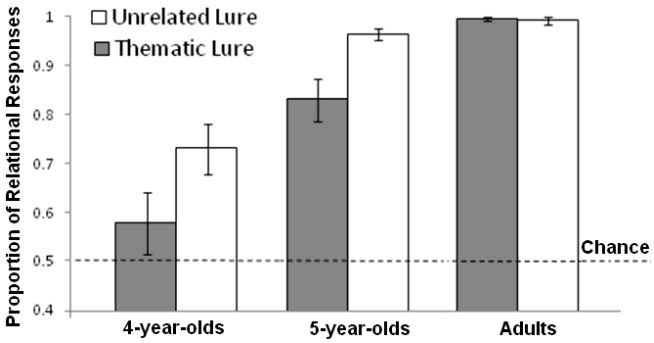


Figure 2: Proportions of relational responses in Experiment 1. Error bars represent standard errors of the mean.

To further understand the source of differences among conditions, we conducted analysis of the individual patterns of responses. Participants were judged to be relational responders if they selected the synonymous choice on at least 8 out of 11 trials (binomial p < .05). Proportion of relational responders in each condition is presented in Figure 3. All adult participants in both lure conditions were deemed to be relational responders. Among 5-year-old children, all participants in the Unrelated Lure condition were deemed to be relational responders, and 12 out of 15 (or 80%) participants in the Thematic Lure condition were classified as relational responders (marginally different from the pattern observed in adults, Fisher's exact p = .058). Among 4-year-old children, 10 out of 18 (56%) participants exhibited the relational pattern of responding in the Unrelated Lure condition, and only 5 out of 17 (29%) participants exhibited the relational pattern of responding in the Thematic Lure Condition. Proportion of relational responders in the 4-year-old group was significantly lower than that of participants in both older age groups in both experimental conditions, all Fisher's exact *p*s < .006).

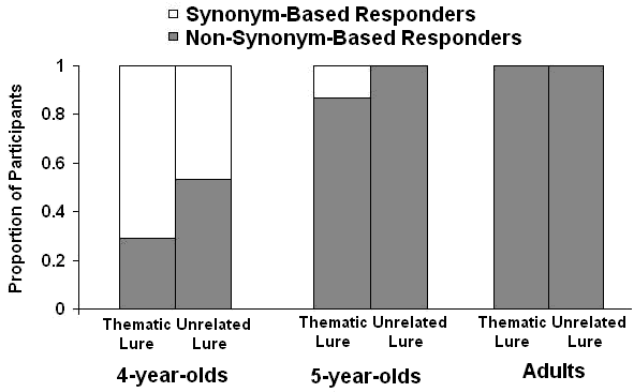


Figure 3: Individual patterns of responses in Experiment 1.

Overall, results of Experiment 1 suggest that the ability to utilize synonymous labels in the course of reasoning may follow a more protracted developmental course than it has been previously believed. Experiment 2 was designed to investigate 4-year-olds' performance in the A:B::A:B' task when identical labels were used. If by 4 years of age children realize that labels refer to kinds, then children's performance with identical labels should be similar to that with synonymous labels in Experiment 1 (Gelman & Markman, 1986). However, if 4-year-old children do not yet treat labels as category markers, their performance with identical labels may be superior to that with synonymous labels.

Experiment 2

Method

Participants

Participants in Experiment 2 were 27 four-year-old children (M = 4.48 years, SD = .27; years; 9 females and 18 males).

Design, Materials, and Procedure

Similar to Experiment 1, there were two between-subject conditions in Experiment 2: a Thematic Lure condition and an Unrelated Lure condition. Children were randomly assigned to one of the two experimental conditions. The order of trials was randomized for each participant.

Materials used in Experiment 2 were identical to those in Experiment 1, with the exception that relational choices were communicated by a label identical to the B-term, rather than by a synonymous label. The B- and B'-terms in Table 1 were counterbalanced across participants, such that half of the participants received the "Castle:Rock" base word-pair, whereas the other half of the participants received the "Castle:Stone" base word-pair.

Results

In both the Thematic Lure condition and the Unrelated Lure condition children averaged 94% of relational choices, above chance (chance = 50%), both one-sample ts > 16.7, ps < .0001). Analysis of the individual patterns of responses revealed that in both experimental conditions, 100% of participants successfully chose identical labels over unrelated as well as thematic lures.

Responses of 4-year-old children in Experiment 2 were compared to those in Experiment 1 in a 2 (Label condition: Identical vs. Synonymous labels) by 2 (Lure Type: Unrelated vs. Thematic Lures) ANOVA. This analysis revealed a main effect of the Label condition, $F(1, 58) = 34.17$, $p < .001$. The labeling condition by lure type interaction did not reach significance, $F(1, 58) = 2.55$, $p = .11$, possibly due to unequal variances in the Synonymous and Identical labels condition, Levene's test of equality of error variances: $F(3, 58) = 8.98$, $p < .0001$.

Overall, children performed significantly better when they could rely on identical labels rather than synonymous labels in both the Thematic Lure condition (94% and 73% of relational responses, respectively) and the Random Lure condition (94% and 58% of relational responses, respectively), both independent-sample ts > 3.2, ps < .005. This conclusion was supported by the analysis of the individual patterns of responses. In the Thematic Lure condition all 13 4-year-old children were classified as relational responders when children could rely on identical labels (Experiment 2), whereas only 5 out of 17 4-year-olds (29%) in the Thematic Lure condition exhibited this pattern of responding with synonymous labels (Experiment 1), Fisher's exact $p < .001$. Similarly, in the Unrelated Lure condition 13 out of 14 4-year-olds (93%) exhibited the relational pattern of responding with identical labels, whereas only 10 out of 18 4-year-olds (56%) exhibited this pattern Unrelated Lure condition with synonymous labels, Fisher's exact $p < .05$.

Discussion

The primary goal of the present study was to examine development of the ability to utilize semantic similarity of labels in a relational reasoning task. Results of the two experiments reported above point to several novel findings. First, adult participants successfully utilized semantic similarity of labels in the Semantic Completion (A:B::A:B') task and their performance was not affected by the type of lure. Importantly, all adult participants exhibited the same pattern of responding on this task. Second, 5-year-old children exhibited a decrease in performance in the Semantic Completion task in the presence of thematic lures compared to unrelated lures. At the same time, proportion of relational responses was well above chance level in 5-year-old children with both types of lures. Furthermore, the majority of 5-year-old children, similar to adults, were classified as relational responders in both lure type conditions. Third, 4-year-old children reliably made relational choices in the Semantic Completion task regardless of the type of lure when relational responses were communicated by identical labels; when relational choices were communicated by synonymous labels, 4-year-old children performed above chance only if correct responses were pitted against unrelated lures, but not thematic lures.

Finally, less than a third of 4-year-old children were classified as relational responders when response competition was strong (in the Thematic Lure condition) and only half of 4-year-olds reliably provided relational responses in the absence of response competition (in the Unrelated Lure condition).

Great care was taken to calibrate stimulus materials used in this research, therefore the reported results cannot be explained by children's unfamiliarity with the words or the relations used in the Semantic Completion and Semantic Substitution tasks. Instead, it appears that less than half of 4-year-old children spontaneously realize that synonymous labels refer to objects of the same kind and can use this knowledge in a reasoning task. This finding has important implications for different theoretical approaches to development of induction. In particular, successful performance on the Semantic Completion task with identical labels does not require that children realize that identical labels refer to objects of the same kind – children could have successfully performed the task based on simple matching of identical labels. However, successful performance on the task with synonymous labels (that do not co-occur in child-directed speech) can only be achieved if children understand that semantically similar labels refer to objects of similar kind. A large decrease in performance on the Semantic Completion task with synonymous labels compared to near ceiling performance with identical labels suggests that at four years of age many children do not yet treat labels as referents to object kind. Therefore, these results pose a challenge to the theoretical accounts of early induction that assume such understanding in very young children (Gelman & Coley, 1991; Welder & Graham, 2001) and suggest that children's understanding that labels refer to kinds continues to develop beyond toddlerhood.

Acknowledgments

We thank children, parents, and teachers for their participation in this research. This project was funded by a Small Undergraduate Research Grant at Carnegie Mellon University awarded to Sheela Ramesh.

References

Banigan, R. L. & Mervis, C. B. (1988). Role of adult input in young children's category evolution: An experimental study. *Journal of Child Language, 15,* 493-504.

Blewitt, P. (1994). Understanding categorical hierarchies: The earliest levels of skill. *Child Development, 65,* 1279-98.

Deák, G. O., & Maratsos, M. (1998). On having complex representations of things: Preschoolers use of multiple words for objects and people. *Developmental Psychology, 34,* 224-240.

Dunn, L. M., & Dunn, L. M. (1997). *Peabody Picture Vocabulary Test-Third Edition: Manual*. Circle Pines, MN: American Guidance Services.

Gelman, S. A. (1988). The development of induction within natural kind and artifact categories. *Cognitive Psychology, 20*, 65-95.

Gelman, S. A., & Markman, E. M. (1986). Categories and induction in young children. *Cognition, 23*, 189-209.

Gelman, S. A., & Coley, J. (1991). Language and categorization: The acquisition of natural kind terms. In S. A. Gelman, S. & J. P. Byrnes (Eds.), *Perspectives on language and thought: Interrelations in development* (pp. 146-196). New York: Cambridge University Press.

Gelman, S. A., & O'Reilly, A. W. (1988). Children's inductive inferences within superordinate categories: the role of language and category structure. *Child Development, 59*, 876-87.

Goswami, U. (1991). Analogical reasoning: What develops? A review of research and theory. *Child Development, 62*, 1-22.

Goswami, U. & Brown, A. L. (1990). Higher-order structure and relational reasoning: Contrasting analogical and thematic relations. *Cognition, 36*, 207-26.

Haryu, E. & Imai, M. (1999). Reorganizing the lexicon by learning a new word: Japanese children's interpretation of the meaning of a new word for a familiar artifact. *Child Development, 73*, 1378-91.

Jaswal, V. K. (2004). Don't believe everything you hear: Preschoolers' sensitivity to speaker intent in category induction. *Child Development, 75,* 1871-85.

Johnson, K. E., Scott, P. & Mervis, C. B. (1997). Development of children's understanding of basic-subordinate inclusion relations. *Developmental Psychology, 33,* 745-763.

Liittschwager, J. C., & Markman, E. M. (1994). Sixteen- and 24-month-old's use of mutual exclusivity as a default assumption in second-label learning. *Developmental Psychology, 30*, 955-968.

MacWhinney, B. (2000). The CHILDES Project: Tools for analyzing talk. Lawrence Erlbaum Associates.

Matlen, B. J., & Fisher, A. V. (2008). Development of synonym-based induction. Poster presented at the Cognitive Science Society, July 28th.

Mervis, C. B., Golinkoff, R. M. & Bertrand, J. (1994). Two-year-olds readily learn multiple labels for the same basic-level category. *Child Development, 65*, 1163-77.

Sloutsky, V. M., & Fisher, A. V. (2004). Induction and categorization in young children: A similarity-based model. Journal of Experimental Psychology: General, 133 (2), 166-188.

Sloutsky, V. M., Lo, Y., & Fisher, A. V. (2001). How much does a shared name make things similar? Linguistic labels, similarity, and the development of inductive inference. Child Development, 72, 1695-1709.

Welder, A. N., & Graham, S. A. (2001). The influences of shape similarity and shared labels on infants' inductive inferences about nonobvious object properties. *Child Development*, 72, 1653-1673.

Negative Transfer in Matchstick Arithmetic Insight Problems

Trina C. Kershaw (tkershaw@umassd.edu)
Department of Psychology, University of Massachusetts Dartmouth
285 Old Westport Road, North Dartmouth, MA 02747 USA

Jason L. G. Braasch (jason.braasch@univ-poitiers.fr)
Centre de Recherches sur la Cognition et l'Apprentissage, Université de Poitiers
5, rue Théodore Lefebvre, 86000 Poitiers, France

Christopher K. Flynn (christopherkflynn@gmail.com)
Department of Psychology, University of Massachusetts Dartmouth
285 Old Westport Road, North Dartmouth, MA 02747 USA

Abstract

The current experiment examined whether successful solution on one type of problem, indicating the relaxation of a constraint, had a negative impact on subsequent problems that did not involve the same constraints. One hundred and forty-five participants solved a series of matchstick arithmetic problems. In one group, participants were given three relatively simple "chunk decomposition" problems (CD). A second group solved one "operator decomposition" (OD) problem, involving more constraints, between the baseline CD problem and two later problems. The third group solved three OD problems, similarly placed. Results indicated that successful solution of an OD problem produced negative transfer to subsequent CD problems in the form of longer solution times. Participants who did not successfully solve OD problems did not slow down on subsequent problems; they displayed evidence of positive transfer. The findings were interpreted with reference to theories of constraint relaxation and its relationship to problem solving performance.

Keywords: mental set; insight problem solving; negative transfer

Procedures in Problem Solving

Whenever new problems are encountered in everyday life, our general approach is to apply procedures or solutions that produced successful outcomes in the past. For example, in the event that you are moving a couch to a new apartment, it is possible that the couch will not fit through a particular doorway or up a particular flight of stairs. Solutions that worked in the past were to unscrew the feet on the couch, or to try the other stairway/doorway into the apartment. If we try one of these solutions and it works again, we will likely bring them to bear when similar situations arise in the future.

Similarly, procedures or solutions that have worked in one context are often evoked and applied to another context. In the first author's most recent move, her bookcase was too large to fit up the front staircase, but could be brought up the back staircase with little trouble. Because the try-the-other-staircase procedure worked in a different situation, it may be that the procedure becomes generalized, thus making it more likely to be employed in a variety of situations.

However, known procedures do not always apply to new situations, and, in fact, may lead to situations of impasse. To use one final example, again from the first author's most recent move, her box spring would not fit up the front staircase. The back staircase was next attempted with no success. She and her movers tried rotating the box spring in multiple orientations on each staircase to no avail. Over an hour was spent attempting to apply a known procedure that was not leading to any progress. Eventually, a neighbor suggested using a chainsaw to split the box spring and fold it in half. The chainsaw procedure was used and the box spring entered the apartment. Thus, a successful procedure was applied, but only after lengthy misapplications of known, and previously useful, procedures.

The preceding everyday example of misapplying previously successful procedures to the moving of furniture is analogous to the sequence that occurs when solving insight problems. An individual's initial representation of an insight problem is often faulty because unhelpful prior knowledge and experiences are activated by the problem (Kershaw & Ohlsson, 2004; Knoblich, Ohlsson, Haider, & Rhenius, 1999; Ohlsson, 1992). The individual's initial problem-solving attempts are guided by this unsuitable knowledge. These initial attempts are usually unsuccessful and the individual then enters a period of impasse, in which no overt problem-solving behavior occurs. In order to exit the impasse, the individual must relax constraints (Knoblich et al., 1999; Ohlsson, 1992) or overcome mental sets caused by incorrect application of procedures (Luchins, 1942). The likelihood of relaxing constraints or breaking mental set depends on the number and strength of the constraints or procedures.

Constraint Relaxation and Breaking Mental Set

The difficulty of a particular insight problem is dependent upon several factors. For many insight problems, including famous examples such as the nine-dot problem, the necklace

.

problem, and the four tree problem, multiple types of constraints interact to make the achievement of solution difficult (Kershaw & Ohlsson, 2004). For example, Kershaw and Ohlsson identified perceptual (figural integrity and other Gestalt laws), knowledge (prior experiences and knowledge), and process (size and variability of search space) constraints that prevent solution of the nine-dot problem. Likewise, Flynn, Gordon, and Kershaw (2010) identified perceptual and knowledge constraints in the four tree problem.

Several researchers state that the difficulty of a particular problem can be found in the strength of constraints present in a particular problem. For example, Knoblich et al. (1999) identified three types of constraints in matchstick arithmetic problems: value, operator, and tautology. The value constraint, the weakest of the three, suggests that numerical values on one side of an equation cannot be changed without compensatory changes on the other. The operator constraint, which is described as having a moderate level of strength, signifies that arithmetic functions (operators) cannot be arbitrarily changed. The tautology constraint, which is the strongest of the three, signifies that arithmetic equations should follow a particular format in which a calculation is specified. That is, an arithmetic operation on one side of the equation should indicate a value on the other side of an equation, such as V + I = VI. While statements like II = II = II are valid, they are not common in arithmetic and therefore violate the tautology constraint.

Knoblich et al. (1999) also classified the difficulty of matchstick arithmetic insight problems by the strength of the chunks that had to be decomposed in order to solve the problem. People tend to view Roman numerals as perceptual chunks, but the strength of these particular chunks depends on the numeral or other element of the equation. Tight chunks, such as V and I, are composed of single units. Loose chunks, such as VII and III, are composed of other chunks. For example, VII is composed of three tight chunks, V, I, and I. Knoblich et al. also suggest there are intermediate chunks, such as operators like the plus sign (+) and the equal sign (=). Although these symbols are composed of other chunks, people are unlikely to have experience decomposing a + into its horizontal and vertical components, for example.

A different explanation of the difficulty of a particular problem is the success of the procedures applied to the problems that preceded it. In a classic demonstration of mental set, Luchins (1942) gave participants a series of water jug problems. The first five problems could all be solved successfully using a particular procedure, but the last five problems either could not be solved using the known procedure or could be solved using a simpler procedure. Luchins (1942) found that participants continued to apply the known procedure to the last five problems, and that over half of the participants were unable to solve problems for which the known procedure could not be applied. That is, participants experienced impasse on some problems and were unable to break impasse to reach solution.

Thus, the difficulty of particular insight problems may be due to the number and strength of constraints or procedures present. Likewise, the likelihood of relaxing these constraints or procedures should also be affected by number and strength. Knoblich et al.'s (1999) theory presupposes that relaxing one weak constraint will be much easier than relaxing multiple strong constraints. Researchers have implemented experimental interventions to increase the likelihood of constraint and procedure relaxation. For example, Kershaw and Ohlsson (2004) and Flynn et al. (2010) developed training procedures that targeted particular constraints, such as practicing non-dot turns for the nine-dot problem (Kershaw & Ohlsson) or comparing solved analogs of the four tree problem (Flynn et al.). Luchins and Luchins (1950) tried to prevent mental set by limiting the amount of liquid available, adding a fourth jar to the problems, and giving participants physical objects (actual jars and water) instead of using paper-and-pencil forms of the problems. Of these three manipulations, only adding a fourth jar was successful, because participants needed to figure out the amount that each jar could hold for each problem. Luchins and Luchins' other two manipulations did not work because participants were poor at keeping track of how much liquid they had used or they persisted in doing paper-and-pencil calculations prior to using the physical materials.

The Current Experiment

In the current experiment, we examine the connection between the strength of constraints and the effect of mental set by using matchstick arithmetic insight problems of two types. One type of problem we used required the decomposition of loose chunks. For example, to solve VI = VII + I, a participant needs to move a single matchstick (I) from VII to VI, thus making the solution of the problem VII = VI + I. Knoblich et al. (1999) states that these types of problems require the relaxation of the value constraint and the decomposition of loose chunks. For simplicity sake, we refer to these problems as chunk decomposition (CD) problems.

The second type of problem we used required the decomposition of the operator in the problem, in this case the plus sign (+). For example, to solve VII = VII + I, a participant needs to move the vertical matchstick from the + to the second VII, thus making the solution of the problem VII = VIII – I. Knoblich et al. (1999) note that these type of problems require the relaxation of the value and operator constraints as well as the decomposition of loose and intermediate chunks. We refer to these problems as operator decomposition (OD) problems. Although Knoblich et al. make a conceptual distinction between constraint relaxation and chunk decomposition mechanisms, we group both into the general category of constraints in the current work. Thus, CD problems contain two constraints and OD problems contain four. Because OD problems contain a greater number of constraints, as well as stronger constraints, they should be harder to solve than CD

.

problems as well as require longer solution times, predictions that are supported by Knoblich et al.'s findings.

In this experiment, all participants solve three CD problems. Participants differed in the number of OD problems that they received. A baseline group of participants did not receive any OD problems, a second group received one OD problem, and a third group received three OD problems. The groups that received the OD problem(s) solved one CD problem, the OD problem(s), followed by two additional CD problems, which functioned as transfer problems.

Our first research goal was to examine how the sequencing of constraint relaxation types affected solution time. In the group that did not receive any OD problems, we expected a general decrease in solution time across the CD problems because, as stated by Knoblich et al. (1999), once constraints are relaxed they will remain relaxed. In the groups that receive the OD problem(s), we explored the possibility that solving the OD problem(s) would make it more difficult to solve the subsequent CD problems. Knoblich et al. (1999, cf. Ohlsson, 1992) posit that constraint relaxation occurs through the natural spreading of activation after persistent failure is experienced via impasse. Because activation to memory nodes decays, it is possible that the CD solution space might become reconstrained after participants spend some time exploring the OD solution space. Therefore, successful solution of OD problems may make solving subsequent CD problems difficult because the constraint would need to be re-relaxed, thus leading to longer solution times for the CD problems received after the OD problem(s) relative to the CD problem received prior to the OD problem(s).

Our second research goal involved the amount of time that participants spent using the procedure needed to solve the OD problems. Thus, we manipulated the mental set that participants experienced due to the OD problems. Some participants only received one OD problem, while others received three. We expected that participants who received three OD problems would show longer solution times on subsequent CD problems than the participants who only received one OD problem relative to the CD problem solved prior to the OD problem(s). We will refer to these post-OD problems as return-to-chunk-decomposition problems and therefore they will be labeled RCD1 and RCD2.

Öllinger, Jones, and Knoblich (2008) explored similar questions using matchstick arithmetic problems. In Experiment 2 they found that solving a series of CD problems did not affect the solution rate for one OD problem (type CR1 in their experiment), although they did affect the solution rate for other constraint relaxation problem types. In Experiment 3 they found that solving a series of constraint relaxation problems negatively impacted CD problems, but the constraint relaxation problems were of a different type than the OD problems used in the current experiment. Thus, while Öllinger et al. (2008) explored similar questions to the current experiment, the current work builds on these findings in terms of providing solvers with different problem types and in varying the number of OD problems between participants.

Overall, we made the following predictions for the experiment:

1) Participants who receive OD problems will have slower solution times than participants who do not receive OD problems on RCD1 compared to the baseline CD problem (B).

2) Participants who receive three OD problems will have slower solution times than participants who receive one OD problem on RCD1 compared to B.

3) Participants who do not receive OD problems will show faster solution times from B to RCD1 and from RCD1 to RCD2. Participants who receive OD problems will not show this pattern.

Method

Participants

Participants were 145 introductory psychology students who received research credit for their participation. Sixty of the participants were from the University of Illinois at Chicago and 85 of the participants were from the University of Massachusetts Dartmouth. No demographic data were collected about the participants.

Materials

A series of matchstick arithmetic insight problems were developed for the study. The problems were of two types, chunk decomposition (CD) and operator decomposition (OD). Following the terminology of Knoblich et al. (1999), the CD problems required the decomposition of loose chunks, which are composite Roman numerals (such as IV, VII, etc.). In each problem, one matchstick is moved from one numeral to another. For example, the problem V = VI + I is solved by moving one matchstick from VI to V, thus making the answer VI = V + I (an acceptable alternate solution is V = IV + I). The CD problems and their solutions are in Table 1.

Table 1: CD problems and solutions.

Problem	Solution(s)
XI = XII + I	XII = XI + I
V = VI + I	VI = V + I, V = IV + I
VI = VII + I	VII = VI + I
VII = VIII + I	VIII = VI + I

The OD problems were akin to Knoblich et al.'s (1999) constraint relaxation (Type B) problems, and specifically required the relaxation of the operator constraint by decomposing the plus sign (+) into two matches and moving the vertical match elsewhere in the location, thus turning the operator into a minus sign (-).For example, the problem II = VIII + V is solved by moving the vertical matchstick from the + to the II, thus making the answer III = VIII – V (an

.

acceptable alternate solution is II = VIII – VI). The OD problems and their solutions are in Table 2.

Table 2: OD problems and solutions.

Problem	Solution(s)
VII = VII + I	VII = VIII - I
II = VIII + V	III = VIII – V, II = VIII – VI
V = VII + I	VI = VII – I, V = VII – II
I = V + III	II = V – III, I = IV – III

Procedure

Participants were run individually. After completing the consent process, participants were given a packet containing the experimental materials. Rules for solving matchstick arithmetic problems were provided on each problem page in the packet. The rules were:

A) Only one matchstick is to be moved.

B) A matchstick cannot be discarded; that is, it can only be moved from one position in the equation to another.

C) A slanted stick cannot be interpreted as a vertical matchstick.

D) The result must be a correct arithmetic equation.

In addition to these rules, participants were given a list of Roman numerals and their Arabic numeral equivalents (e.g., X = 10).

The first problem for all participants was XI = XII + I, which served as a practice problem. Participants were given five minutes to work on the problem and were instructed to alert the experimenter when they came up with a solution. If the participant correctly solved the problem, the experimenter summarized the participant's actions and stated that the solution was correct. The experimenter emphasized that one matchstick had been moved to create a correct equation. If the participant came up with an incorrect solution, the experimenter referred back to the rules to explain why the solution was incorrect. For example, the participant might be reminded that only one matchstick could be moved. If the participant did not solve the practice problem correctly within the time limit, the experimenter first checked to see if he/she had any questions and then gave him/her two more minutes to work on the problem. If, after this additional time, no solution was offered, then the experimenter explained how to move one matchstick to achieve the correct solution, XII = XI + I.

The second problem in the packet was the baseline chunk decomposition problem (B), V = VI + I. Participants had four minutes to work on this problem (and all subsequent problems). Participants wrote down their start time, worked on the problem, and wrote down the end time if they came up with a solution. The accuracy of the solution was checked by the experimenter. If the solution was incorrect, the experimenter used the rules to point out the inaccuracies of the solution.

The penultimate and final problems in the packet were also CD problems. As stated previously, the penultimate problem, VI = VII + I, will be referred to as the first return-to-chunk-decomposition problem (RCD1), and the last problem, VII = VIII + I, will be referred to as the second return-to-chunk-decomposition problem (RCD2). Participants received the same instructions and same amount of time to solve the B, RCD1, and RCD2 problems.

The problems in between B and the RCD1, RCD2 sequence differed by condition. One group of participants did not receive any OD problems. A second group of participants received one OD problem between B and the RCDs. A third group of participants received three OD problems between B and RCDs. On all OD problems, participants followed the same procedure as used for the CD problems by writing down their start and end times and checking their solutions with the experimenter.

After completing RCD2, participants filled out a problem familiarity survey, which asked participants if they had seen and solved any of the matchstick arithmetic problems prior to the experimental session. No participants had any familiarity with the matchstick arithmetic insight problems. At the end of the session, participants were debriefed and thanked for their participation.

Analysis

Participants were originally grouped by the number of OD problems they received. There were 46 participants who received no OD problems, 50 who received one OD problem, and 49 who received three OD problems. However, initial examination of the data revealed that not all participants in the OD conditions solved the OD problems. Therefore, participants were regrouped by the number of OD problems they solved. If participants did not solve the OD problems, we could not expect that they also relaxed this constraints associated with these problems. Thus, in the final analyses, there were 64 participants who solved no OD problems, 35 participants who solved one OD problem, and 46 participants who solved three OD problems.[1]

Time to solve the B, RCD1, and RCD2 problems was calculated by subtracting the end time from the start time for each problem. If a participant did not solve one of these problems, then his/her time data were not included. The number of non-solvers was low for each problem: one participant did not solve B, five participants did not solve RCD1, and three participants did not solve RCD2. The time-to-solve data were then screened for outliers, which were defined as time to solve values that were greater than three standard deviations above the mean. Rather than deleting data list-wise, data points were removed case-wise. Three time-to-solve values were removed from the B and RCD1 values, and four time-to-solve values were removed from the RCD2 values.

Three variables were computed for the planned comparisons between the solution times. First, a value was

[1] Removing participants who did not conform to their groups rather than regrouping participants led to the same pattern of results.

calculated for RCD1 – B, that is, the difference between the time needed to solve the baseline and first return to chunk decomposition problems. Next, a value was calculated for RCD1 – RCD2, that is, the difference between the time needed to solve the first and second return to chunk decomposition problems. Third, a value was calculated for B – RCD2, that is, the difference between the time needed to solve the baseline and second return to chunk decomposition problems.

Results

A one-way analysis of variance (ANOVA) compared participants on the difference between the time needed to solve the baseline chunk decomposition problem (B) and the time needed to solve the first return-to-chunk-decomposition problem (RCD1). The ANOVA was significant, F (2, 135) = 4.44, $p < .05$, $\eta^2 = .06$. Tukey post-hoc tests indicated that participants who solved three OD problems showed a significant increase in solution time from the B to the RCD1 problems (M = 19.26 seconds, SD = 43.44) compared to participants who did not solve any OD problems (M = -4.18 seconds, SD = 40.99), $p < .05$. Participants who did not solve any OD problems showed a decrease in time-to-solve between B and RCD1. There was also a marginal difference between participants who did not solve any OD problems and those who solved one OD problem (M = 15.41 seconds, SD = 45.74), $p = .09$. Importantly, there was no difference between participants who solved one OD problem and those who solved three.

A second analysis compared participants on the difference between the time needed to solve RCD1 and RCD2 (RCD1 – RCD2). A one-way ANOVA did not show any inter-group differences, F (2, 129) = .86, $p > .05$, $\eta^2 = .01$.

A third analysis compared participants on the difference between the time needed to solve B and RCD2 (B – RCD2). A one-way ANOVA showed an overall difference between the conditions, F (2, 131) = 4.64, $p < .05$, $\eta^2 = .07$. Tukey post-hoc tests indicated that participants who did not solve any OD problems needed significantly less time to solve RCD2 than to solve B (M = 15.14 seconds, SD = 23.08) compared to participants who solved three OD problems (M = 2.35 seconds, SD = 22.25), $p < .05$. There was also a marginal difference between participants who did not solve any OD problems and those who solved one OD problem (M = 3.69 seconds, SD = 24.30), $p = .07$. There was no difference between participants who solved one OD problem and those who solved three.

Discussion

The current work produced four main important findings. First, in the absence of successful OD performance, participants got progressively faster when solving CD problems. This finding suggests that relaxation of the value constraint made it easier to solve subsequent value constraint problems. In this sense, we found some evidence of positive transfer on problems that presumably required relaxation of the same constraint. This finding supported our third prediction, that participants who did not receive OD problems would show faster solution times from B to RCD1, while participants who received OD problems would not show this pattern. This finding also supports Knoblich et al.'s (1999) theory – once a constraint is relaxed, it will remain relaxed and affect subsequent performance on similar problems.

Second, the current results provide greater confidence that constraint relaxation can also negatively impact subsequent problem solving performance, particularly in the event that a different, more complex constraint was relaxed. Generally, successful solution of OD problems resulted in longer subsequent solution of CD problems compared to those who did not solve or were not presented with OD problems. Thus, the longer solution times indicate that there was at least some negative transfer associated with the relaxation of the operator constraint. Solution of an OD problem appeared to make it more difficult to solve the simple CD problems; this difficulty was absent for those who did not solve OD problems. This finding supports our first prediction, that participants who received OD problems would have slower solution times than participants who did not receive OD problems on RCD1 compared to B. Additionally, this finding replicates and extends the findings of Öllinger et al. (2008), who also found successful solution of constraint relaxation problems (of a different type) affected later problem solving performance.

Third, as stated previously, we varied the number of OD problems that were presented to participants and were solved in between the CD problems. Participants who solved one and three OD problems displayed similar indications of negative transfer on subsequent RCDs, as evidenced by longer solution times. This finding did not support our second prediction, in which we predicted relatively slower solution times for participants who received three OD problems than participants who received one OD problem. Our finding suggests that, indeed, after one successful solution of an OD problem, the operator constraint was relaxed. Moreover, there did not seem to be an additional slowing associated with solving multiple OD trials.

The final important point is that the negative transfer effects associated with the relaxation of the operator constraint were relatively lasting. That is to say that successful solution not only affected the immediate CD problem, but also the problem that followed. Although the current data does not provide an indication of how long-lasting this kind of negative transfer would be, there did seem to be a "downstreaming" effect into subsequent problem solving performance, beyond the problem situation that immediately followed the constraint relaxation.

Overall, the findings of this experiment point to the manner in which previously appropriate procedures can be persistently misapplied to a new situation. As demonstrated by Luchins (1942; Luchins & Luchins, 1950), mental set can hinder future problem solving. Mental set and other interference effects fit within the larger concept of negative

.

transfer, in which prior knowledge and experiences hinder learning in new situations that are similar to known situations. The type of negative transfer effects shown in this experiment are similar to those proposed by Singley and Anderson (1989): participants show behavioral slowing due to the misapplication of a procedure. However, the misapplication is an incorrect method, not a non-optimal method, and this misapplication of procedure lasts for more than one trial, thus lending some support to Woltz, Gardner, and Bell's (2000) theory of negative transfer.

Further work is needed to address the direction and duration of negative transfer effects. It is possible that completing a series of CD problems could lead to negative transfer on the OD problems. Likewise, it would be interesting to determine if increases in solution time on the RCD problems lasts more than two iterations. The negative transfer literature is divided on whether negative transfer effects are fleeting (e.g., Singley & Anderson, 1989) or lingering (e.g., Woltz et al., 2000). Additional studies extending the number of to-be-solved CD problems may inform on this issue.

Another future direction for this research would be to examine the processes that underlie the interaction between constraint relaxation mechanisms. Our findings, as well as the findings of Öllinger et al. (2008), show that relaxing some constraints hinders the relaxing of other constraints. One explanation for these findings is that successful solution of OD problems may make solving subsequent CD problems difficult because the constraint would need to be re-relaxed. Thus, there would be longer solution times for the CD problems received after the OD problems relative to the CD problem received prior to the OD problems. Alternatively, relaxing a stronger constraint, such as the operator constraint present in the OD problems, may cancel out a weaker constraint, such as the value constraint present in the CD problems. Although this a different explanation the same effect would be expected, in which solution times are longer for the CD problems received after the OD problem(s) than for the CD problem received prior to the OD problem(s). A third possibility is that the relaxing of multiple constraints opens up the problem space too much, thus leading to a difficulty in finding a correct solution path (c.f. Ohlsson, 1996; Ormerod, MacGregor, & Chronicle, 2002). This third explanation would also lead to the same pattern of results. Future research should address the mechanisms that underlie constraint relaxation interactions and, if possible, attempt to tease apart which of these three possibilities best explains negative transfer in problem solving performance.

Acknowledgments

We thank Valentina Pacheco and Justin Faria for their assistance with data collection. We also thank three anonymous CogSci reviewers for their comments and suggestions for future research.

References

Flynn, C.K., Gordon, L.T., & Kershaw, T.C. (2010). Multiple paths to transfer and constraint relaxation in insight problem solving. Unpublished manuscript.

Kershaw, T.C., & Ohlsson, S. (2004). Multiple causes of difficulty in insight: The case of the nine-dot problem. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*(1), 3-13.

Knoblich, G., Ohlsson, S., Haider, H., & Rhenius, D. (1999). Constraint relaxation and chunk decomposition in insight problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*(6), 1534-1555.

Luchins, A.S. (1942). Mechanization in problem solving: The effect of Einstellung. *Psychological Monographs, 54*, 1-95.

Luchins, A.S., & Luchins, E.H. (1950). New experimental attempts at preventing mechanization in problem solving. *Journal of General Psychology, 42*, 279-297.

Öllinger, M., Jones, G., & Knoblich, G. (2008). Investigating the effect of mental set on insight problem solving. *Experimental Psychology, 55*(4), 269-282.

Ohlsson, S. (1992). Information-processing explanations of insight and related phenomena. In M. Keane & K. Gilhooly (Eds.), *Advances in the psychology of thinking*. London: Harvester-Wheatsheaf.

Ohlsson, S. (1996). Learning from performance errors. *Psychological Review, 103*(2), 241-262.

Ormerod, T.C., MacGregor, J.N., & Chronicle, E.P. (2002). Dynamics and constraints in insight problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*(4), 791-799.

Singley, M.K., & Anderson, J.R. (1989). *The transfer of cognitive skill*. Cambridge, MA: Harvard University Press.

Woltz, D.J., Gardner, M.K., & Bell, B.G. (2000). Negative transfer errors in sequential skills: Strong-but-wrong sequence application. *Journal of Experimental Psychology, 26*(3), 601-625.

.

A critique of multi-voxel pattern analysis

Michael L. Anderson (michael.anderson@fandm.edu)
Department of Psychology, Franklin & Marshall College
Lancaster, PA 17604 USA

Tim Oates (oates@cs.umbc.edu)
Department of Computer Science, University of Maryland, Baltimore County
Baltimore, MD 21250 USA

Abstract

Multi-voxel pattern analysis (MVPA) is a popular analytical technique in neuroscience that involves identifying patterns in fMRI BOLD signal data that are predictive of task conditions. But the technique is also frequently used to make inferences about the regions of the brain that are most important to the tasks in question, and our analysis shows that this is a mistake. MVPA does not provide a reliable guide to what information is being used by the brain during cognitive tasks, nor where that information is. This is due in part to inherent run to run variability in the decision space generated by the classifier, but there are also several other issues, discussed here, that make inference from the characteristics of the learned models to relevant brain activity deeply problematic. These issues have significant implications both for many papers already published, and for how the field uses this technique in the future.

Keywords: neuroscience, machine learning, inference, philosophical issues.

Introduction

Multi-voxel pattern analysis (MVPA) is an increasingly popular analytical technique in neuroscience. MVPA involves searching through the Blood Oxygenation Level Dependent (BOLD) signal data produced in fMRI experiments to identify patterns that are highly predictive of task conditions. To illustrate, consider a simple experiment in which participants are asked to view pictures representing various object categories (e.g. faces, houses, chairs, shoes, etc.). One early MVPA study showed it was possible to determine, by looking only at BOLD data, which class of object an experimental participant was viewing when that data was collected (Haxby et al., 2001). The technique has since been used to predict the orientation of lines being viewed by a participant (Haynes & Rees, 2005), to differentiate between lying and truth-telling (Davatzikos et al., 2005), and to predict which action a participant was about to take (Haynes et al., 2007), among many other things (see Pereira, Mitchell & Botvinick, 2009; Norman et al., 2006; Haynes & Rees, 2006 for reviews of the technique and its applications).

This is indeed impressive, and we expect that MVPA will have many important experimental and diagnostic applications (Lao et al., 2004). It has become commonplace to make certain inferences about the way differences in BOLD signal patterns correspond to differences in mental states. For instance, by finding the set of voxels that are most predictive of a certain task outcome, studies have claimed to discover the "cognitive states associated with perception of tools and dwellings" (Shinkareva et al., 2008), "localizable task-specific representations of freely chosen intentions" (Haynes at al., 2007), and the regions of the brain that "contain information" (Preston et al., 2008) relevant to the cognitive or perceptual task under investigation.

To put it bluntly, however, such inferences are at best misleading and at worst entirely unwarranted. The issues dovetail with, but are distinct from, the more general concerns about the unreliability of "reverse inference" from neuroimaging data (Poldrack, 2006), and have significant implications both for how we ought to interpret some of the many papers already published, and for how the field applies this technique in the future.

Of course, not every MVPA study is governed by the logic that we will criticize here. For instance, Mitchell et al. (2008) take something like the opposite approach, and see if they can predict the pattern of brain activity that will be caused by listening to novel words. Here the point of the study is not to discover which brain regions are responsible for understanding; rather, they are testing the hypothesis that meanings of words are based on sets of "semantic features" that can be inferred from word co-occurrence in language corpora. McDuff, Frankel & Norman (2009) are likewise focused on hypothesis testing, in their case about the characteristics of targeted memory retrieval. We think that MVPA has a very promising future both as a diagnostic tool, and as a useful dependent variable—in part because the technique is sensitive to contingencies beyond classical single-voxel effects—but that for the reasons outlined in this paper it is a very poor tool for reliably localizing information or identifying cognitive states.

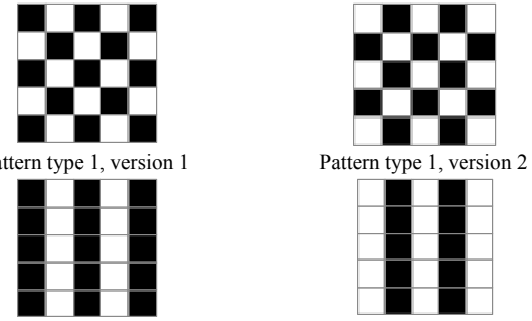
Information and the brain

There are three general ways in which information could inhere in the BOLD signal. First, the information could be non-local, that is, carried by irreducibly relational features of the signal like regional co-variance. We might expect this to occur when large-scale neural synchrony is the relevant aspect of brain activity (Varela, et al., 2001; Gross et al., 2004). Second, it could be local and distributed, that is, the information could be carried by the activity of individual voxels, and the information-carrying voxels could be spread throughout the brain. We might expect this for cognitive processes that require the cooperation of many different brain regions. Third, the information could be local and concentrated, that is, carried by individual voxels that are grouped together in one or a few clumps. This might happen when the work done by local neural circuits is most important to the cognitive task(s) in question. In this essay, we will consider the performance of MVPA in all three

situations, and discuss what can, and cannot, be inferred from features of the learned model in each case.

Local, distributed information

Consider the problem of differentiating between the following two patterns of hypothetical voxel-level activation data, each presented with two *versions* of the same pattern *type* (see Figure 1). In this simple example each of the 25 "voxels" can be in one of two states (active or inactive, if you like). Suppose "brain scans" like these had been observed during an experiment in which participants were asked to classify pictures as "living" or "non-living". If these judgments reliably corresponded to the two patterns, respectively, could we use MVPA to read the mind of the participants?



Pattern type 1, version 1 Pattern type 1, version 2

Pattern type 2, version 1 Pattern type 2, version 2

Figure 1: Simple patterns for use in MVPA test

Now, when the ratio of pattern versions within each pattern type is 1:1, every voxel is in both of its possible states in every task condition. That is: no voxel is by itself predictive of any cognitive state, and thus in this condition all information is non-local. In this condition *linear* MVPA cannot distinguish between these two patterns; it is blind to non-local information (Kamitani & Tong, 2005; Norman et al. 2006). For linear classifiers, since the evidence provided by each voxel is integrated separately, linear MVPA is successful only when there are individual voxels that are sensitive to the difference between classes. In general, a (binary) linear classifier over an input space of dimension n looks like this:

$$\text{prediction} = \text{sign}\left(b + \sum_{i=1}^{n} w_i * x_i\right)$$

where the ith weight is w_i and the ith component of the input vector (the list of numbers that describe the patterns to be classified) is x_i and the bias value is b. If the sum above is positive, the instance is classified one way; if it is negative, the instance is classified the other way.

However, manipulating the version ratio changes the situation from one in which no voxel is more informative than any other—a situation in which linear classifiers fail—to one in which there is indeed a set of voxels, scattered through the patterns, that are informative about class membership. That is to say, although there is still non-local information in the patterns—and it is arguable that the non-local co-variance structure is the crucial, relevant distinction between these patterns—the initial test situation is one in which there nevertheless is also relevant local information, distributed across many voxels.

For our analysis of the performance of MVPA with local, distributed information, we generated 20 sets of 80 "scans"—that is, 20 datasets, each containing 40 instances of each pattern type. Patterns were corrupted with 5% noise—a 5% chance for each voxel that it will be in a state inconsistent with the pattern. For each dataset, we used 40 of the 80 scans for training and 40 for test, and classified them using a Support Vector Machine. Because classification accuracy roughly tracks the relative proportion of pattern versions, our scans contained a 4:1 ratio of pattern versions within each type, and classification accuracy averaged 80%.

Thus, our hypothetical experiment would have produced a solid predictive success; we would be able to tell, 80% of the time, which task condition the participant was in just by looking at the fMRI data. But what, if anything, would we be permitted to conclude about the local neural conditions—representations, information content, activity, etc.—contributing to the differences in cognitive tasks (thinking about or judging the difference between living vs. non-living things)?

Although any of the input components could contribute to the prediction breaking one way or the other in an given case (and it needn't be the same components for each instance), in practice there can be a small number of voxels that contribute most to the classifier performance because they (literally) carry the most weight—that is, they have the highest values of w_i. In linear MVPA, this set of highly weighted voxels is considered the "most informative".

Figure 2 shows a map of the voxels that were most informative for distinguishing between pattern types 1 and 2 in dataset 1.

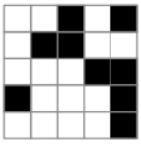


Figure 2: Most informative voxels for an MVPA classification

What is the proper interpretation of these results in the context of MVPA? These are the voxels that, had they been in a different state, would have been most likely to cause the classifier to place the pattern in the other class. But consider the following inference, an inference of similar structure to those being made in the MVPA literature: if the state of these voxels had been different in the right way—and note this picture provides no information about what the right way is—the brain would have been in the relevantly different state (or the participant would have been in the different cognitive state). This inference does not follow, because if covariance is the crucial cognitively relevant property of the activity here, then all the other voxels would also be different when the brain/participant is in the other state: they will be covarying with a different set of partners. And, even if covariance is not the crucial property—if the relevant information is the local information—it seems pretty clear that it isn't all or only the voxels in the "most

informative" set that would need to be in a different state to turn one pattern into the other.

Likewise, consider a similar inference (versions of which can also easily be found in the literature): the information contained in these voxels is the information crucial to the difference between the cognitive states under investigation (judging living vs. non-living things). This inference is also unwarranted, for similar reasons. For one strong possibility is that the relevant information is carried by the covariance structure of the patterns, and this non-local information is not contained in the set of "most informative" voxels. And even if the local information is what is relevant here, we can see from the results above that the set of most informative voxels does not consist of all or only the voxels carrying the relevant information.

The uncertainty of inferences about brain or cognitive states based on which voxels are most highly weighted is driven home even more strongly when one looks at the stability of the set of highly weighted voxels over multiple trials of the same task. Figure 3 shows the most highly weighted voxels from the first three datasets.

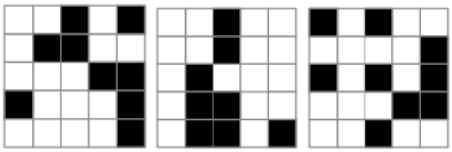


Figure 3: Most informative voxels for three different classification runs

Obviously, the highly weighted voxels vary from run to run. To get a better quantitative handle on the stability of the highly weighted voxel set, we counted the number of times each voxel was among the top 10 most highly weighted. Overall, every voxel was in this set at least twice, and none more than 12 times. 24 voxels were in the set between 6 and 12 times, and 22 between 6 and 10 times. The characteristics of the classification model can vary considerably, driven in part by noise in the training instances, but also by the fact that the algorithm needs only find *some* of the features that discriminate between *some* instances of the patterns *some* of the time. It is not guaranteed to find all of the relevant differentiating features, nor the best. The conclusion seems obvious, but is worth stating clearly: when any voxel can make it into the "most informative" set, and many voxels are more or less equally likely to end up there, this should make us a bit uneasy about their actual informativeness. If there is something stable to the cognitive states differentiating the task conditions, the set of most informative voxels is certainly not tracking it, nor can it therefore be a reliable indicator of the location of the cognitively relevant information.

In should be noted that cross-validation does not alleviate this issue. Cross-validation consists of a family of methods designed to prevent over-fitting of the model to what could be an unusually biased sample. Typically, it involves building multiple models based on multiple partitions of the sample, and averaging results over the range of different partitions (Pereira et al., 2009). For instance, K-fold cross-validation involves splitting the training data into K parts, and training K times on a rotating K-1 of the partitions. We performed 10-fold cross validation on our 20 training sets, and found similar variability in the set of most informative voxels in each fold. The mean number of inclusions among the top 10 most highly weighted voxels was 4 (SD 2.83). 23 of the voxels were among the 10 most highly weighted voxels in at least one fold.

Non-local information

So, that seems to be the situation when information is local and distributed. What about when the only information is non-local, that is, when the ratio of pattern versions is 1:1? It turns out it is possible to classify these patterns with 100% accuracy, applying MVPA using a support-vector machine with a polynomial kernel of degree two. Can we conclude anything in this case about the neural conditions—representations, information, activity, etc.—contributing to the differences in cognitive tasks?

One is of course tempted to simply dismiss the possibility. In our examples every voxel is in both of its possible states in every task condition. That is: no voxel is by itself predictive of any cognitive state, and thus no inference to the special status of activity in any voxel could possibly be supported by the predictive success of MVPA. Yet researchers do extract "most informative" voxel sets even when using polynomial kernels (e.g. Davatzikos et al., 2005), so it is wise to consider the matter more carefully.

In linear classifiers identifying most important voxels for the classifier is easy—the features in the decision space that have the highest weights are the most important, and these features have a 1:1 correspondence with components of the input vector, that is, with the voxel values fed into the classifier. But non-linear SVMs use "kernel functions", whereby a vector input is projected into a kernel-specific high-dimensional space, and the importance of each feature is determined in that space. The original space for a given vector x has one dimension for each component of the input vector, and the value of that feature—its position on dimension i—is just x_i. In contrast, a polynomial kernel of degree 2 (K_2) projects the input vector into a space having a dimension for each unique (unordered) pair of features in the vector, and the value of each of those features—its position on dimensions (i,j)—is $x_i * x_j$.

Thus, the K_2 classifier over an input space of dimension n looks more like this:

$$\text{prediction} = \text{sign}\left(b + \sum_{i,j=1}^{n} w_{i,j} * x_i * x_j\right)$$

In fact, we can get exactly this situation by manually projecting our input vectors into the polynomial space—turning them in this case from n-dimensional vectors into $n+((n^2-n)/2)$ dimensional vectors—and using linear SVMs to classify them. This procedure will produce the same decision surface as using K_2 in the original space, but will allow us to directly inspect the resulting weights to determine which features were most important.

However, given the nature of non-linear SVMs, relating features to individual components of the input vectors is inherently problematic. For note that what gets weighted in the decision function is the product of each pair of

components. So, if a given product turns out to be important to the classifier, shall we attribute this importance to just one of the components, or to both? Either decision seems likely to give misleading results. Nevertheless, for the sake of the discussion, let's adopt the simple rule that when a given feature is highly weighted, both components (voxels) will be counted as "informative". Given this, we can examine the frequency with which voxels are informative, and track the voxels that are frequently informative.

To test this procedure when using K_2, we generated 40, 10x10 versions of the standard patterns from Figure 1, 20 of each pattern type, with a 1:1 ratio of versions, and a noise level of 5%. We projected each of these patterns into the 5,050-dimensional feature space of K_2, and trained a linear SVM on the set. Then we found the top 500 highest weighted features, and projected these back onto the 10x10 pattern following the rule above. Now, it is perfectly legitimate to make the following inference from this procedure: the highly weighted voxels are the ones that, had they been in a different state, would have been most likely to cause the classifier to place the pattern in the other class. The trouble is, the weighting is often taken to tell us something about the relative importance of each voxel to the intrinsic difference between the patterns (and to the underlying cognitive states), and no such inference is warranted in this case.

First, there is a basic problem of interpretation given that the important features are in fact products of two voxels—so, every time a voxel is deemed informative, it has a partner with which it was important, and the set itself gives no information about the distribution of these partners. Second, it is clear in this case (because there is no local information) that the relevant information differentiating between the patterns is non-local, carried in the covariance structure of the pattern, and this information is not contained in the set of frequently informative voxels. Third, the most highly-weighted features are not those that contain the most information. As in the linear case, they are the features that contained sufficient information to drive the classifier on a given set of training examples. Fourth and finally, as should not be surprising, the set of informative features and informative voxels is highly unstable in this case, as well.

To explore the stability of the set of important features when using K_2, we generated 10x10 versions of the standard patterns above, creating 100 sets of 40 (20 of each pattern) with a noise level of 5%. We projected each of these patterns into the 5,050-dimensional feature space of K_2, and trained a linear SVM on each of the 100 sets. From each of these 100 sets, we extracted the top 500 most important features. Doing a pair-wise comparison of the most important features from each set revealed that, on average, only 101.08 (SD 16.94) of these features (20.21%) were common between each pair. Moreover, the common features varied from pair to pair. Doing a 5-wise comparison of the most important features sets reveals an average of just 0.81 (SD 1.09) of the features (0.16%) are shared across all five sets. Note that despite the instability of the "most informative" feature sets, classification accuracy in all cases was 100%.

Given the high degree of variability in the features considered most important, it seems certain that the set of frequently informative components (voxels) is likewise unstable. To confirm this, we generated 500 training sets of the 10x10 patterns, and, following the procedure above, found the top 500 most important features for each set. Then, we counted the number of times each individual component of the input vector was included in a pair that was in this important feature set. On average, each component was included in the set 10.00 times (SD 0.39). No component averaged fewer than 9 inclusions, or more than 11.00. Once again, if there is some stable difference between the cognitive states in the two task conditions, the set of most informative voxels is certainly not tracking it, nor can it therefore be a reliable indicator of the location of the cognitively relevant information.

Admittedly, this example was based on a very simple rule for mapping features in the multi-dimensional space to components of the original vector, and it is true that more sophisticated procedures for uncovering the most informative components have been developed (Davatzikos et al., 2005; Lao et al., 2004). But insofar as these techniques still depend in one way or another on identifying the most highly weighted features in a multi-dimensional space, and insofar as this set is not determinate for a given classification task, then the results of such analyses need to be interpreted with extreme caution.

Before moving on with the remainder of the analysis, it is worth pausing to summarize the findings. In the case where there is local information relevant to distinguishing patterns, linear MVPA does not reliably find it; and in the case where there is relevant non-local information, carried for instance by covariance patterns, linear MVPA cannot find it, and non-linear MVPA models can make it look as if they were using local information. More importantly, having discovered some features whose state matters most to the classification decision is not the same as having discovered the brain regions whose activity matters most (or even relatively more) to the participant (or her brain). Indeed, these two sorts of information need have no regular correspondence to one another; one need not track, be a reliable indicator of, or be otherwise instructive about the nature, scope or location of the other.

Local, concentrated information

How is this disconnect possible? Consider first an example from the MVPA literature meant to showcase the power of the technique. Haynes and Rees (2005) were able to use MVPA to correctly identify the orientation of visually-presented lines, even when the stimuli were presented briefly and masked so that the participant did not consciously perceive them. That is an intriguing result, and may tell us something interesting about the operation of V1 (the ROI they used to make the predictions). But note the broader implication for the method: since the participants cannot judge the orientation of the lines, they cannot be in whatever cognitive state gives the ability to judge the orientation of the lines. Thus, MVPA can be used to infer features of the task environment from characteristics of the

BOLD signal, without being a reliable indicator of the cognitive state of the participant.

Now consider extending the experiment in the following straightforward way: while the visual stimulus is being shown (and masked), experimenters play an auditory tone from which the participant could reliably infer the orientation of the line. If, as seems likely in this particular case, the most informative voxels for the pattern classifier would remain in V1, this outcome would provide a clear instance in which the information used by the participant and the information used by the classifier would not have the expected relation.

But is such an outcome really possible? In fact, this hypothetical example points in the direction of a well-known fact about the way classification algorithms perform. Numerous theoretical results and a tremendous amount of empirical evidence in machine learning demonstrate that there is no universally best learning algorithm (Wolpert, 1996). Every algorithm has a bias that is appropriate for some problems and inappropriate for others. This is true for the brain, and the same is true of kernels. There is no universally best kernel, and changing from one kernel to another can lead to large changes in the learned decision surface and thus to changes in what features in the data set seem to be important.

The relevance of this problem for MVPA is that a particular set of stimuli may elicit different patterns of activity, call them pattern A and pattern B, in different parts of the brain, and one kernel may be able to detect pattern A but not pattern B, whereas another kernel may be able to detect pattern B but not pattern A. Thus, when relating "most informative features" to "most important activity", the area of the brain implicated in the experiment will change depending on which kernel is used.

To make this concrete, consider two patterns with 20 binary features (f_1 - f_{20}) in which for every instance of the first (positive) pattern the following two conditions hold:

(a) Either $f_{19} = 1$ and $f_{20} = -1$, or $f_{19} = -1$ and $f_{20} = 1$
(b) The sum of the first 5 bits is less than or equal to zero

For every instance of the second (negative) pattern, the following two conditions hold:

(a) Either $f_{19} = 1$ and $f_{20} = 1$, or $f_{19} = -1$ and $f_{20} = -1$
(b) The sum of the first 5 bits is greater than zero

The values of the other bits are chosen uniformly at random from {−1, 1}. Condition (a) is the logical exclusive or (*XOR*) function on bits 19 and 20 and is easily learned by the polynomial kernel of degree two (the class label is $-\text{sign}(f_{19} * f_{20})$) but is impossible to learn with a linear kernel. Condition (b) is easily learned with a linear kernel (the class label is 1 if $f_1+f_2+f_3+f_4+f_5 \leq 0$ and is -1 otherwise), but is extremely difficult for the polynomial kernel of degree two because it has access to individual feature f_i only as $f_i * f_i$ which is 1 regardless of the value of f_i.

We created 100 datasets based on the above rules and trained an SVM with a linear kernel on both the original feature space and the feature space constructed for the polynomial kernel of degree two. In the latter space, the feature corresponding to $f_{19} * f_{20}$ had an average weight of 3.64. The remaining 209 features had average weights in the range (0.05, 0.10). In the former case, the average weights for features f_1 through f_5 were 1.92, 1.94, 1.94, 1.93, and 1.94. The remaining 15 features had average weights in the range (0.03, 0.10). Clearly, the choice of kernel can have a dramatic impact on which features are deemed important and, in the case of MVPA, which voxels are implicated in various cognitive tasks.

Thus, although much of this paper was spent detailing the worrying instability and potential deceptiveness of the most informative voxel set when information is non-local or distributed, the fact is that even if MVPA were perfectly reliable at the task of finding the most informative features in a data set, the inference from this to the brain activity most important determining the outcome in given task would remain fairly weak. This is because inference from most informative features to most important activity apparently relies on the unwarranted additional assumption that the pattern classification algorithm and the brain are classifying on a relevantly similar basis. While of course no one claims that the success of MVPA shows that the brain is implementing an identical classifier, the issue is that the hypothesis space is different for different classifiers, and so different information will be relevant to each. What is relevant in the brain, and what is relevant to classifying an image of the brain, need not bear much relation.

Conclusion

There are very many challenges to the task of reliably relating the features (of the BOLD signal) most important to classification success to the features (of brain activity) most important to cognitive states/outcomes. By way of summation, consider this general list of possible ways in which these features might fail to relate as expected.

(1) The highly informative elements of the pattern as discerned by MVPA are distributed in the brain in such a way that the brain is anatomically or functionally incapable of integrating the information. If people are nevertheless capable of making the relevant discrimination, it must have been on the basis of different information.

(2) There may well be classes of stimuli that differ in ways undetectable to subjects (under any presentation condition, conscious or otherwise), but which nevertheless create patterns in the BOLD signal allowing for successful classification by MVPA. Consider in this regard an experiment run by Hung et al. (2005). Macaques passively viewed picture stimuli in eight different categories while undergoing direct recording of neural activity using microelectrode arrays. Hung et al. were able to successfully classify the stimuli with a linear SVM taking the multi-unit activity as input. But here the macaques did not—indeed, in all likelihood could not—classify the stimuli, because they had not been trained to do so. In this case, the SVM might have been making distinctions that the (untrained) macaques were not.

(3) Stimuli may differ along more than one dimension, both of which lead to differences in the BOLD signal. MVPA classification could rely on patterns relating to one

dimension, while participants use information relating to the other. That is, even when there is information in the BOLD signal that is theoretically accessible by (or that is tracking information accessible by) the participant, this may not be the information that is being used by the participant.

(4) The MVPA classifier may be using a kernel that is significantly different from what is implemented in the brain. As we saw, classifiers with different kernels trained on the very same data will extract different features, and thus come to different decisions about which features (and which elements of the input vectors) are most important.

(5) Since there will always be a set of highly informative voxels produced by the MVPA classifier, the existence of such a set won't tell us whether the relevant information in the brain is local and concentrated, local and distributed, non-local, or some combination of these.

The discussion also raises a much more general issue. As we noted at the outset, MVPA offers an exciting new way to investigate the operation of the brain, by looking at the predictive value of (typically widely) distributed patterns of activity. The problematic inferences generally come in the attempt to reduce such patterns to local features of brain activity. But if the best predictor of cognitive states is not the location of an activated region, but rather the patterns of cooperation and coactivation between them—as the success of MVPA might be said to indicate, and as has been argued for independent reasons (Anderson, 2008; Sporns, et al., 2004; Uttal, 2001)—then perhaps it is time to pay more heed to the patterns than to the partners. We are just beginning to develop the tools to make such an investigation fruitful and rigorous—including not only MVPA but other forms of statistical pattern analysis, machine learning, graph theory, etc.—and it seems a shame instead to use these tools in the service of localization projects for which they are ultimately ill-suited. New tools often come with the opportunity to re-consider the strengths of theoretical perspectives and paradigms, and these are offering a chance to look beyond localization, to what other perspectives on brain organization might have to offer.

References

Anderson, M. (2008). Circuit sharing and the implementation of intelligent systems. *Connection Science*, 20(4): 239-51.

Davatzikos, C., Ruparel, K., Fan, Y., Shen, D., Acharyya, M., Loughead, J., Gur, R. & Langleben, D.D. (2005). Classifying spatial patterns of brain activity with machine learning methods: Application to lie detection. *NeuroImage*, 28(3): 663-68.

Gross, J., Schmitz, F., Schnitzler, I., Kessler, K., Shapiro, K., Hommel, B & , Schnitzler, A. (2004). Modulation of long-range neural synchrony reflects temporal limitations of visual attention in humans. *Proceedings of the National Academy of Sciences-USA*, 101(35): 13050-13055

Haxby, J.V., Gobbini, M. I., Furey, M.L., Ishai, A., Schouten, J.L. & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293: 2425-30.

Haynes, J-D. & Rees, G. (2006). Deconding mental states from brain activity in humans. *Nature Reviews Neuroscience*, 7:523-34.

Haynes, J.-D. & Rees, G. (2005). Predicting the orientation of invisible stimuli from activity in human primary visual cortex. *Nature Neuroscience*, 8(5): 686-91.

Haynes, J.-D., Sakai, K., Rees, G., Gilbert, S., Frith, C. & Passingham, R.E. (2007). Reading hidden intentions in the human brain. *Current Biology*, 17: 232-28.

Hung, C.P., Kreiman, G., Poggio, T. & DiCarlo, J.J. (2005). Fast readout of object identity from macaque inferior temporal cortex. *Science*, 310: 863-6.

Kamitani, Y. & Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nature Neuroscience*, 8(5): 679-85.

Kay, K.N., Naselaris, T., Prenger, R.J. & Gallant, J.L. (2008). Identifying natural images from human brain activity. *Nature*, 452: 352-6.

Lao, Z., Shen, D., Xue, Z., Karacali, B., Resnick, S. & Davatzikos, C. (2004). Morphological classification of brains via high-dimensional shape transformation and machine learning methods. *NeuroImage*, 21 (1): 46-57.

McDuff, S.G.R., Frankel, H.C. & Norman, K.A. (2009). Multivoxel pattern analysis reveals increased memory targeting and reduced use of retrieved details during single-agenda source monitoring. Journal of Neuroscience, 29(2):508-516.

Mitchell, T.M., Shinkareva, S.V., Carlson, A., Chang, K.-M., Malave, V.L., Mason, R.A. & Just, M.A. (2008). Predicting human brain activity associated with the meanings of nouns. *Science*, 320: 1191-5.

Norman, K.A., Polyn, S.M., Detre, G.J. & Haxby, J.V. (2006). Beyond mind-reading: Milti-voxel pattern analysis of fMRI data. *TRENDS in Cognitive Sciences*, 10(9): 424-30.

Pereira, F., Mitchell, T., Botvinick, M. M. (2009). Machine learning classifiers and fMRI: a tutorial overview. *NeuroImage*, 45: S199-S209.

Poldrack, R.A. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Sciences*, 2006; 10(2): 59-63.

Preston, T.J., Li, S., Kourti, Z. & Welchman, A.E. (2008). Multivoxel pattern selectivity for perceptually relevant binocular disparities in the human brain. *The Journal of Neuroscience*, 28(44): 11315-27.

Shinkareva, S.V., Mason, R.A., Malave, V.L., Wang, W., Mitchell, T.M. & Just, M.A. (2008). Using fMRI brain activation to identify cognitive states associated with perception of tools and dwellings. *PLOS One*, 3(1): e1394. doi:10.137/journal.pone.0001394

Sporns, O., Chialvo, D., Kaiser, M. & Hilgetag, C.C. (2004). Organization, development and function of complex brain networks. *Trends in Cognitive Sciences*, 8: 418-425.

Uttal, W. (2001). *The New Phrenology*. Cambridge: MIT Press.

Varela, F., Lachaux J.P., Rodriguez E. & Martinerie J. (2001). The brainweb: phase synchronization and large-scale integration. *Nature Rev. Neuroscience*, 2(4): 229-39.

Wolpert, D. (1996). The lack of a priori distinctions between learning algorithms. *Neural Computation*, 8(7): 1341-1390.

The Origins of Collective Overvaluation: Irrational exuberance emerges from simple, honest and rational individual behavior

Michael L. Anderson (michael.anderson@fandm.edu)
Department of Psychology, Franklin & Marshall College
Lancaster, PA 17604 USA

C. Athena Aktipis (aktipis@alumni.reed.edu)
Department of Ecology and Evolutionary Biology, University of Arizona
Tucson, AZ 85721 USA

Abstract

The generation of value bubbles is an inherently psychological and social process, where information sharing and individual decisions can affect representations of value. Bubbles occur in many domains, from the stock market, to the runway, to the laboratories of science. Here we seek to understand how psychological and social processes lead representations (i.e., expectations) of value to become divorced from the inherent value, using asset bubbles as an example. Using an agent-based model we explore whether a simple switching rule can generate irrational exuberance, and systematically explore how communication between decision makers influences the speed and intensity of overvaluation. We show that rational and simple individual level rules combined with honest information sharing are sufficient to generate the collective overvaluation characteristic of irrational exuberance. Further, our results demonstrate that simple noise in the exchange of value information leads to rapidly increasing expectations about value, even when no one is engaged in exaggerating their expectations for the assets they own.

Keywords: decision making; valuation; agent-based modeling; rationality; emergence.

Introduction

Chances are, your savings are invested in one or more kinds of assets—stocks, bonds, real estate, etc. Moreover, if you are an individual investor, or are planning on becoming one soon, you probably discuss the markets with various other investors, including friends, family, colleagues and investment professionals. You might also listen to one of the many market watch programs, or read the business section of your daily newspaper. In short, you are probably engaged in both soliciting and offering opinions on how various market sectors will perform in the future. Once in a while, this information will cause you to make a change in your portfolio. Imagine, for instance, that someone you trust shares with you their expectation for the performance of one of their investments. Imagine further that this expectation exceeds the expectation that you yourself have for your own investments. Surely there is some chance that you would sell (some of) your own portfolio, and invest in the asset with the higher expected return. Whether you would do this naturally depends on myriad other factors—your tolerance for risk, the perceived balance of your current investments, the liquidity of this new asset class, etc. But there remains some chance that you will make the switch. This is natural, and even—assuming that one of your financial goals is to maximize return consistent with other priorities—rational. But if we are right, this natural, rational behavior is sufficient to spark irrational exuberance.

Asset bubbles are among the most fascinating and puzzling phenomena in economic markets. Decision makers frequently drive up prices and demand to levels that seem completely divorced from the underlying value. Bubbles are common, and far from innocuous. Post-bubble market "corrections" have led to financial ruin for many, as occurred in the great depression and in the current real estate and financial crises. And there seem to be some important similarities between asset bubbles and other sorts of collective behavior, including clothing fashions, popular music trends and perhaps even the trajectory of science (with processes such as paper acceptances and grant funding being based on the expectations of reviewers about the future value of the work). Thus, bubbles are important to understand, to say the least. In the current paper, rather than seeking to understand these events through analyzing or modeling the complex historical and economic factors that lead to a specific instance of collective overvaluation, we have instead focused on formulating some simple and general individual rules that we hypothesize are sufficient to generate the phenomenon of irrational exuberance. We have isolated what we believe to be a key underlying cause of collective overvaluation / irrational exuberance across many contexts, and have constructed a simple model to explore whether it generates the predicted outcomes.

Here we model the genesis of collective overvaluation as a general phenomenon, using decision making about asset classes as an example. We aim to make this model as general as possible, making it potentially applicable to other domains.

Model Description

The model description offered below follows the standardized ODD protocol for describing individual and agent based models (Grimm and Railsback 2005; Grimm et al. 2006). This protocol for describing agent based models has been developed with input from modelers across the disciplines and is in wide use.

Purpose

A commonly observed behavior in markets of many kinds is continually increasing expectations about the future value of certain commodities/asset groups. Here we used agent based techniques to model a simple decision rule that we predict to be sufficient to generate both increasing expectations and overexploitation of certain assets (absorption of all individuals into a small number of asset groups). We also explore the impact of communication fidelity on the outcomes.

State variables and scales

In this model, time and space are both represented discretely. During each time period, all agents execute the commands described in the schedule. The simulation is constructed in a spatial environment for the purposes of visualizing interactions between asset groups.

Process overview and scheduling

This model proceeds in discrete time steps, and entities execute procedures according to the following ordering:

1. Individual A identifies random partner B to be recipient of information about asset value expectations.
2. Individual A communicates current expectation of value for A's current asset class to individual B with some fidelity
3. Individual B adopts expectation of individual A with some probability (opportunism) if A and B come from different groups, and A's expectation is higher than B's.
4. If B has adopted A's expectation then B switches to A's group.

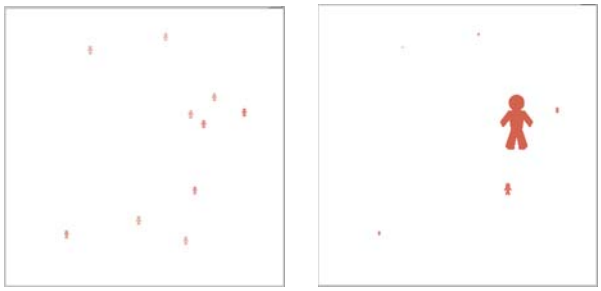


Figure 1: Two screen shots showing the initial conditions and the state of the simulation after 150 time steps under the default parameters (see Table 2). *Left:* The run begins with 10 groups of uniform size with an average expectation of 100. *Right:* After 150 time steps, there is one large group and the expectations of agents have increased to 131.5 (as indicated by the darker red shade of the agents).

Table 1. Overview of state variables associated with each type of entity in the simulation. Bold indicates manipulated independent variables and arrows indicate dependent variables.

Entity	State variable	Description
Global	• **Transmission fidelity**	Accuracy of communication of expectation. This is modeled by communicating to the partner not the agent's actual expectation, but an expectation taken randomly from a normal distribution with the transmission fidelity as its standard deviation and the agent's actual expectation as its average.
	• **Expectation distribution**	Initial variability (expressed as Standard Deviation) of expectations among individuals in the population
	• **Opportunism**	Probability of changing groups given a higher expectation communicated from partner
	➢ Number of groups	The number of groups (asset classes)
	• Number of agents	The number of individuals included in the model
	• Communication target	Binary, either random individual or individual in another group
	➢ Switches per step	The number of agents that change groups each step
	➢ SD switches per step	The standard deviation in the number of switches per step
	➢ Average expectation	The average expectation of all individuals regarding the future value of their investment
	➢ Change in expectation per step	The average change in the average expectation each step
	➢ SD change in expectation	The standard deviation of the change in expectation per step
	➢ Volatility	A combined measure of the SD switches per step and the SD change in expectation per step. This is an indication of the unpredictability of the changes from step to step.
Groups (Asset classes)	• Location	Coordinates of the group
	• Group size	Number of individuals in asset class
Agents	• Expectation	The future value the individual assigns to the current asset
	• Partner expectation	The information the individual has about their current partner's expectation in their asset class
	• ID number	The identification number of the individual
	• Partner ID number	The identification number of the current partner

Design Concepts

Emergence Irrational aggregate behavior emerged from individual-level rational decision making processes.

Prediction Agents did not have a complex function for predicting the future value of asset classes. They simply adopted information from partners if the information met the conditions described above.

Sensing Individuals have an initial expectation of the value of their asset class based on the expectation distribution. From this point forward, individuals' expectations change only from information transmission from other agents.

Interaction Individuals can transfer information about their expectation of the value of their asset class to partners (with some fidelity). Individuals can move to a new group (asset class), if the partners communicated expectation is higher than the current expectation.

Stochasticity Initial distribution of expectations is randomly distributed around the inherent value of a particular asset class. Opportunistic switching is implemented probabilistically and so has a stochastic element.

Collectives Agents were parts of groups (asset classes) and could transfer information to a 'partner' (from the same or other group). Partners were reset each time period and information transfers were unidirectional (i.e., A might transfer information to B, and B to C)

Observation Simulations were run for 2000 time steps or until only a single group remained. Each combination of independent variables (see Experiments, below) was run 10 times. The dependent variables were measured at the end of each run. Reported results are averages over 10 runs.

Initialization

Table 2 lists the variables associated with various entities in the simulation. All runs were initialized according to default parameters in the table.

Table 2. Initial and default values for all instance variables and independent variables (bold).

Entity	State variable	Initial/Default Value	Units
Global	• **Transmission fidelity**	Perfect (SD of 0)	
	• **Expectation distribution**	SD of 10	
	• **Opportunism**	5%	
	• Number of groups	10	count
	• Number of agents	1,000	count
Groups (Asset classes)	• Group size	100	count
	• Average expectation	100	Expected future value
Agents	• Expectation	Assigned from expectation distribution	
	• Partner expectation		
	• ID number		
	• Partner ID number		

Input

This model is designed as a general model of irrational exuberance and collective overvaluation. We did not initialize this model with real world data.

Experiments

We ran three simple and three complex experiments. In the three simple experiments, we used only a single independent variable, while in the three complex we used two, to look for interactions between the effects.

As noted above, all runs were initialized with 10 groups, each containing 100 agents, with an overall average expectation of 100. The three independent variables of interest were: initial expectation distribution, opportunism, and transmission fidelity.

Experiment 1, expectation

This experiment varied only the initial expectation distribution, setting it so the initial distribution of expectations had a standard deviation of 10, 20 and 30. Opportunism was fixed at 5%, and transmission fidelity was perfect.

Experiment 2, fidelity

This experiment varied only transmission fidelity, setting it at 0, 5 and 10. Recall that transmission fidelity is modeled by communicating to the partner not the agent's actual expectation, but an expectation taken randomly from a normal distribution with the transmission fidelity as its standard deviation and the agent's actual expectation as its average. Thus 0 equals perfect fidelity. Opportunism was fixed at 5% and the initial expectation distribution was fixed at 10.

Experiment 3, opportunism

This experiment varied only opportunism, setting it at 5%, 10%, and 15%. The initial expectation distribution was fixed at 10 and transmission fidelity was perfect.

Experiment 4, expectation x fidelity

This experiment varied both expectation distribution (10, 20, 30) and fidelity (0, 5, 10). Opportunism was fixed at 5%.

Experiment 5, fidelity x opportunism

This experiment varied both fidelity (0, 5, 10) and opportunism (5%, 10%, 15%). The initial expectation distribution was fixed at 10.

Experiment 6, expectation x opportunism

This experiment varied both expectation distribution (10, 20, 30) and opportunism (5%, 10%, 15%). Fidelity was perfect.

Dependent variables

The dependent variables measured in these experiments were:

A. The average expectation at the end of the run, representing the average agent expectation of the value of the asset class(es).
B. The number of groups remaining at the end of the run, representing the number of asset classes with investors
C. The number of switches per step, corresponding to the number of agents that switched groups each time step
D. The average change in expectation per step, corresponding to the change in expectation that occurs as agents switch and adopt the expectations of others
E. The volatility of the system, measured as the summed standard deviations of the number of moves per step and the average change in expectation per step.

Results

Descriptive statistics for experiment 1, expectation, are listed in Table 3. Increasing the distribution of expectations lead to a higher average expectation at the end of the run (ANOVA, $F(2, 27) = 112.45$, $p << .01$, see Figure 2) a larger change in expectation each time period (ANOVA, $F(2, 27) = 58.31$, $p << 0.01$), and higher overall volatility (ANOVA, $F(2, 27) = 34.34$, $p << 0.01$).

Table 3. Descriptive statistics for experiment 1, expectation.

Expectation distribution:	**10**	**20**	**30**
Groups at end	1.0 (0.0)	1.0 (0.0)	1.0(0.0)
Average expectation at end	131.55 (4.34)	165.99 (10.44)	199.95 (13.57)
Number moves per step	11.65 (2.25)	12.18 (0.92)	12.13 (1.31)
Δ-expectation per step	0.09 (0.02)	0.21 (0.04)	0.29 (.0.6)
Volatility	10.81 (0.52)	12.21 (0.47)	12.85 (0.68)

Descriptive statistics for experiment 2, fidelity, are listed in Table 4. Greater noise (low transmission fidelity) led to much higher average expectations at the end of the runs $F(2, 27) = 68.66$, $p << .01$ (see Figure 3); to more groups at the end of the simulation $F(2, 27) = 91.5$, $p << 0.01$; and to less overall volatility $F(2, 27) = 521.56$, $p << 0.01$.

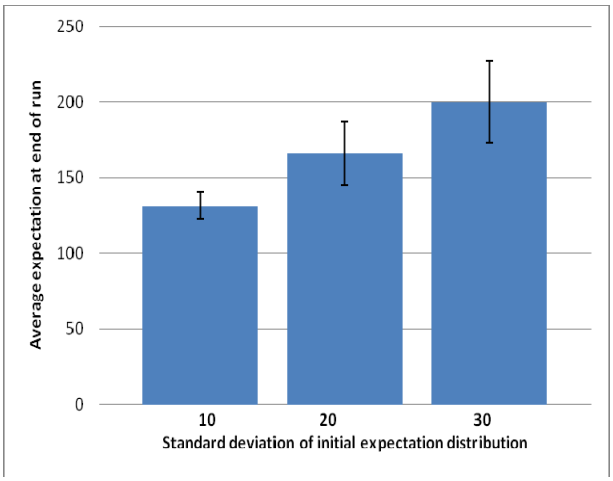


Figure 2: Increase in average expectation as a function of initial expectation distribution. Error bars represent ± 1 S.E.

Table 4. Descriptive statistics for experiment 2, fidelity.

Transmission fidelity:	**0**	**5**	**10**
Groups at end	1.0 (0.0)	3.3 (0.48)	2.7(0.48)
Average expectation at end	133.84 (2.86)	375.31 (107.04)	682.72 (146.98)
Number of moves per step	12.67 (1.23)	17.88 (0.68)	17.90 (1.30)
Δ-expectation per step	0.11 (0.02)	0.25 (0.01)	0.49 (0.02)
Volatility	10.94 (0.51)	5.68 (0.17)	6.57 (0.40)

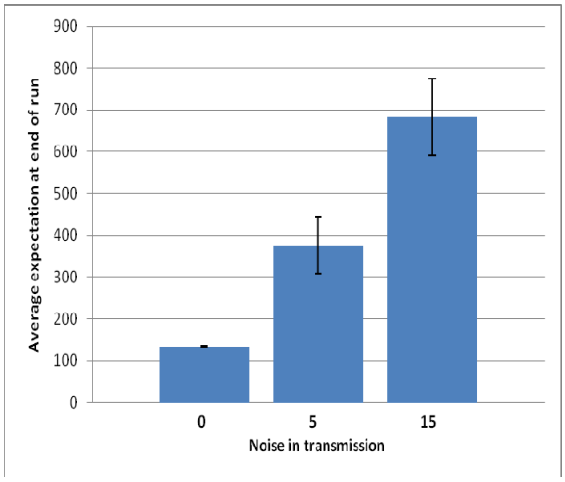


Figure 3: Increase in average expectation as a function of transmission fidelity. Error bars represent ± 1 S.E.

Note the increase in expectation is driven in part by the fact that with high noise, the number of groups never drops to one, as it always does when fidelity is perfect. Thus the

simulations when fidelity was > 0 lasted for all 2,000 steps, rather than stopping after around 300 steps, as is typical when fidelity is perfect. Even so, there was also a significant increase in the average change in expectation per step, indicating that the effect is not simply a matter of running the simulation for longer.

Descriptive statistics for experiment 3, opportunism, are listed in table 5. Greater opportunism increases the number of moves per step $F(2, 27) = 657.16$, $p << 0.01$; increases the amount by which expectations change each step $F(2, 27) = 657.16$, $p << 0.01$; and increases volatility $F(2, 27) = 1531.22$, $p << 0.01$. In addition, there was a decrease in the number of steps it took to achieve one group, and thus for the simulation to end $F(2, 27) = 260.41$, $p << 0.01$. That is, the more opportunistic the agents are, the faster the collective converges on a single asset. This explains why, despite a significant increase in the change in expectation each step, there was no main effect on average expectation at the end.

Table 5. Descriptive statistics for experiment 3, opportunism.

Opportunism:	**5%**	**10%**	**15%**
Groups at end	1.0 (0.0)	1.0 (0.0)	1.0(0.0)
Step when one group reached	298.80 (30.68)	162.30 (14.86)	98.70 (6.46)
Average expectation at end	133.45 (4.71)	132.21 (2.49)	133.76 (3.35)
Number of moves per step	12.71 (1.07)	25.33 (1.83)	38.73 (1.80)
Δ-expectation per step	0.11 (0.02)	0.20 (0.01)	0.34 (0.04)
Volatility	11.09 (0.41)	21.02 (1.22)	31.95 (0.69)

Interactions

The three complex experiments revealed the same main effects, which won't be repeated here. Instead we'll simply summarize some of the significant interactions.

Experiment 4, Expectation x Fidelity reveals a significant interaction between expectation distribution and fidelity on volatility $F(4,81) = 3.42$ $p = 0.012$. Whereas the general effect of fidelity on volatility is to decrease it when going from 0 to 5, and increase it slightly when going from 5 to 10, this latter effect disappears at higher levels of expectation distribution.

Experiment 5, Fidelity x Opportunism reveals an interaction between fidelity and opportunism on the number of moves per step $F(4,81) = 21.66$, $p << 0.01$; the change in expectation per step $F(4,81) = 341.86$, $p << 0.01$; and volatility $F(4,81) = 256.84$, $p << 0.01$. Both fidelity and opportunism increase the number of moves per step, and increase the change in expectation per step, and together the higher values increase the magnitude of the effect. As noted above, the change in fidelity tends to decrease volatility initially, then increase it slightly. These effects are greater as opportunism increases.

Experiment 6, Expectation x Opportunism reveals an interaction between expectation distribution and opportunism on the change in expectation per step $F(4,81) = 15.40$, $p < 0.01$ and on volatility $F(4,81) = 10.71$, $p << 0.01$. In each case the tendency of the independent variables to increase volatility and change in expectation per step is enhanced at higher levels of the other variable.

Discussion

On December 5, 1996, after nearly fifteen years of steady growth in the S&P 500 and Dow Jones Industrial Average (and just before the record-breaking bull market to follow), Federal Reserve chairman Alan Greenspan expressed his concern that the behavior of the stock market was characterized by "irrational exuberance". Whether he was right or not, it is certainly true that the price to earnings ratio had by then surpassed 27, a level that hadn't been seen since 1929, and was on its way to the record high of 47 it achieved in March of 2000. What leads to this sort of (apparent) disregard for underlying real value? There are several possible explanations. Some favor accounts based on individual irrationality—e.g. "animal spirits" like (over-) confidence and our tendency to be influenced by nominal amounts of money—that can be amplified under certain market and social conditions (Akerloff, 2005; Akerloff & Shiller, 2009). Others favor "herd behavior" models in which individuals allow their choices to be guided by other people's choices, on the (reasonable, but by no means certain) assumption that there is wisdom in crowds (Surowiecki, 2004). On these models, observations of early choices create an information cascade that causes late choosers to follow early ones, rather than following their own signal (Banerjee, 1992; Bikchandri, Hirshleifer & Welch, 1992). Finally, there is currently a great deal of discussion of the role of deception in the recent real-estate bubble (Bitner, 2008).

Here we consider the alternate possibility that irrational exuberance is driven by neither irrationality nor deception, nor requires individuals to ignore their own information and preferences, but instead emerges from simple, honest and rational individual-level behavior. To explore this possibility we created an agent-based model where agents have simple and seemingly rational individual-level rules for switching between asset classes and updating their representations of asset value based on information from others. Our results show that a simple rule—when another agent's expectation for the performance of their investment exceeds your own expectation for your own investment, consider switching investments—can generate collective behavior resembling irrational exuberance.[1] In particular,

[1] Although communication partners were chosen at random, agents adopted new expectations *only* when the partners represented different asset classes. Restricting communication to partners from other groups greatly speeds the dynamics outlined here, because members of smaller groups are bombarded with

we see rapidly increasing expectations for the value of commodities and the overexploitation of a single asset class. Further, our model shows that this collective overvaluation can occur even when there is no individual deception or bias in favor of exaggerating value when communicating to others about it. This suggests that surprisingly simple and rational individual level rules can generate some of the complex and irrational aggregate outcomes associated with market bubbles.

One especially interesting finding was the massive effect that transmission fidelity had on overvaluation. Here is a system in which increasing noise increases the rapidity and magnitude of overvaluation, and the interactions demonstrate that this effect can be magnified by other factors. Ironically, then, Alan Greenspan's infamous opacity could itself have been a contributor to the irrational exuberance he warned against. Although we do not explore this possibility explicitly here, it is clear that combining noise with even a few agents intent on deception would cause even greater overvaluation than we demonstrated in these experiments. This is perhaps part of the combination that led to the recent real-estate bubble.

This model has both specific implications for the phenomenon of market bubbles as well as general implications for the phenomenon of collective overvaluation across domains. Because this model simulates individual decision making processes (as is typical of agent based models) rather than simply aggregate dynamics, it is able to capture important effects of interactions among individuals (in terms of information sharing and switching). Models such as this can be used to improve our understanding of the psychological and social components of decision making behavior by allowing us to explore the generative sufficiency of individual rules as well as the sensitivity of the system to alterations in parameters such as those explored here (i.e., transmission fidelity, initial expectation, opportunism in switching). The model presented here demonstrates that representations/expectations of value can become dissociated from inherent value when individuals use simple and rational decision rules combined with well-intentioned communication. The emergence of increasing expectation from these simple and general decision making and communication processes may be the fundamental principle that gives rise to irrational exuberance, not just in the market place, but in any domain in which individuals switch from their current option when they hear about better opportunities elsewhere.

Thus, in addition to the potential relevance of this model for market phenomena, there are more general implications that can be drawn as well. The emergence of collective overvaluation from a simple switching rule could occur in a wide range of domains, making this model applicable to a wide range of phenomena. In fact, this model is sufficiently abstract that it can be applied to a variety of other situations in which individuals' assessments of value are based on social information. For example, clothing fashions, popular music, and even current trends in areas of scientific study might be subject to similar processes. These may be fruitful avenues for future research.

Future work will also explore market dynamics in greater depth and detail. For instance, we will explore the effect that broadcast information (e.g. announcements from the Fed, ratings agencies, etc) might have on the creation of asset bubbles. We will also allow for the dynamic creation of new asset classes, and allow agents to decide to temporarily opt out of the market. Finally, we will explore what can be done to reverse such overvaluation in a more controlled fashion than is typical in a market crash, or prevent high degrees of overvaluation from occurring in the first place.

messages from members of larger groups, thus increasing their likelihood of switching to the larger group.

Acknowledgments

This paper was prepared while M.L.A. was on junior faculty research leave from Franklin & Marshall College. He gratefully acknowledges the support. C.A.A. was supported by grants F32CA144331 and R01CA140657 from the National Cancer Institute.

References

Akerloff, G.A. & Shiller, R.J. (2009). *Animal spirits: How human psychology drives the economy, and why it matters for global capitalism*. Princeton, NJ: Princeton University Press.

Banerjee, A.V. (1992). A simple model of herd behavior. *The Quarterly Journal of Economics*, 107(3): 797-817.

Bikhchandani, S., Hirshleifer, D. & Welch, I. (1992). A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy*, 100 (5):992-1026.

Bitner, R. (2008). *Confessions of a subprime lender: An insider's tale of greed, fraud, and ignorance*. New York: Wiley.

Grimm, V. & Railsback S. F. (2005). *Individual-based modeling and ecology*. Princeton University Press, Princeton, NJ.

Grimm, V., Berger, U., Bastiansen, F., Eliassen, S., Ginot, V., Giske, J., Goss-Custard, J., Grand, T., Heinz, S. K., Huse, G., Huth, A., Jepsen, J. U., Jørgensen, C., Mooij, W. M., Müller, B., Pe'er, G., Piou, C., Railsback, S. F., Robbins, A. M., Robbins, M. M., Rossmanith, E., Rüger, N., Strand, E., Souissi, S., Stillman, R. A., Vabø, R., Visser, U. & DeAngelis. D. L. (2006). A standard protocol for describing individual-based and agent-based models. *Ecological Modelling*, 198 (1-2): 115-126.

Shiller, R.J. (2005). *Irrational exuberance*. New York: Broadway Books.

Surowiecki, J. (2004). *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies and nations*. New York: Doubleday.

A category theory explanation for systematicity

Steven Phillips (steve@ni.aist.go.jp)
National Institute of Advanced Industrial Science and Technology (AIST),
Tsukuba, Ibaraki 305-8568 JAPAN

William H. Wilson (billw@cse.unsw.edu.au)
School of Computer Science and Engineering, The University of New South Wales,
Sydney, New South Wales, 2052 AUSTRALIA

Keywords: systematicity; classicism; connectionism; compositionality; category theory; product; functor; adjunction

Abstract

Classical and Connectionist theories of cognitive architecture "explain" systematicity, whereby the capacity for some cognitive behaviors is intrinsically linked to the capacity for others, as a consequence of syntactically and functionally combinatorial representations, respectively. However, both theories depend on *ad hoc* assumptions to exclude specific architectures—grammars, or Connectionist networks—that do not account for systematicity. By analogy with the Ptolemaic (i.e., geocentric) theory of planetary motion, although either theory can be made to be consistent with the data, both nonetheless fail to explain it (Aizawa, 2003b). Category theory provides an alternative explanation based on the formal concept of adjunction, which consists of a pair of structure preserving maps, called functors. A functor generalizes the notion of a map between representational states to include a map between state transformations (processes). In a formal sense, systematicity is a necessary consequence of a "higher-order" theory of cognitive architecture, in contrast to the "first-order" theories derived from Classicism or Connectionism. Category theory offers a re-conceptualization for cognitive science, analogous to the one that Copernicus provided for astronomy, where representational states are no longer the center of the cognitive universe—replaced by the relationships between the maps that transform them.

Introduction

For more than two decades since Fodor and Pylyshyn's seminal paper on the foundations of a theory of cognitive architecture (Fodor & Pylyshyn, 1988), the problem of explaining systematicity remains unresolved (Aizawa, 2003b) despite numerous Classicist and Connectionist claims to the contrary (Fodor & McLaughlin, 1990; van Gelder, 1990; Smolensky, 1987).

The problem of systematicity for a theory of cognition is to explain why the capacity for some cognitive behaviours is intrinsically linked to some other cognitive capacities. The systematicity problem is actually three problems:

1. *Systematicity of representation*—why is it the case that the capacity to generate some representations (e.g., the representation `John loves Mary`) is intrinsically linked to the capacity to generate some other representations (e.g., the representation `Mary loves John`)?

2. *Systematicity of inference*—why is it the case that the capacity to make some inferences (e.g., that John is the lover in the proposition John loves Mary) is intrinsically linked to the capacity to make some other inferences (e.g., that Mary is the lover in the proposition Mary loves John)?

3. *Compositionality of representation*—why is it the case that the capacity for some semantic content (e.g., the thought that John loves Mary, however that thought may be represented) is intrinsically linked to the capacity for some other semantic context (e.g., the thought that Mary loves John, however that thought may also be represented)?

These problems are logically independent—one does not necessarily follow from another (Aizawa, 2003a), and so a theory is required it explain all three.

Classicists and Connectionists employ some form of combinatorial representations to explain systematicity. For Classicists, representations are combined in such a way that tokening of representations of complex entities entails tokening of representations of their constituent entities, so that the syntactic relationships between the constituent representations mirror the semantics ones—systematicity is a result of a combinatorial syntax and semantics (Fodor & Pylyshyn, 1988). For Connectionists, representations of complex entities are constructed more generally so that their tokening does not necessarily imply tokening constituent entity representations (van Gelder, 1990; Smolensky, 1987). We refer to the former as *classical compositionality*, and the latter as *functional compositionality*.

In general, a Classical or Connectionist architecture can demonstrate systematicity by having the "right" collection of grammatical rules, or functions such that one capacity is indivisibly linked to another. Suppose, for example, a Classical system with the following three production rules:

G1: `P` → `Agent loves Patient`
`Agent` → `John | Mary`
`Patient` → `John | Mary`.

The capacities to generate all four representations (i.e., `John loves John`, `John loves Mary`, etc.) are indivisibly linked, because the presence of all three, or absence of any one of those rules means the system is only capable of generating either all or none of those representations. In no case can the

system generate one without being able to generate the other. So, this Classical architecture has the systematicity of representation property with respect to this group of four propositions. Tensor products (Smolensky, 1990), or Godel numbers (van Gelder, 1990) are functionally compositional analogues to this explanation. Systematicity of inference follows from having additional processes that are sensitive to the structure of these representations. For Classical architectures, compositionality of representation also follows, because the semantic content of a complex representation is built up from the semantic contents of the constituents and their syntactic relationships (Aizawa, 2003a). Aizawa (2003a, 2003b) disputes whether a Connectionist architecture can also demonstrate compositionality of representation. Regardless, though, neither Classicism, nor Connectionism can derive theories that provide a full account of systematicity (Aizawa, 2003b).

A demonstration of systematicity is not an explanation for it. In particular, although grammar G1 has the systematicity of representation property, the following grammar:

G2: `P` → `John loves Patient` | `Agent loves Mary`
`Agent` → `John` | `Mary`
`Patient` → `John` | `Mary`

does not. This architecture cannot generate a representation of the proposition Mary loves John even though it can generate representations of both John and Mary as agents and patients, and the John loves Mary proposition. The essential problem for Classical theory—likewise Connectionist theory—is that syntactic compositionality by itself is not sufficient without some additional assumptions that admit grammars such as G1 that have the systematicity property, but exclude grammars such as G2 that do not. An explanation for systematicity in these cases turns on the nature of those additional, possibly *ad hoc* assumptions.

Ad hoc assumptions

Aizawa (2003b) presents an explanatory standard for systematicity and the problem of *ad hoc* assumptions by analogy with the Ptolemean (geocentric) versus Copernican (heliocentric) explanations for the motions of the planets (see Phillips, 2007, for a review). The geocentric explanation for planetary motion places the Earth at the center of the other planets' circular orbits. Although this theory can roughly predict planetary position, it fails to predict periods of apparent retrograde motion for the superior planets (i.e. Mars, Jupiter, etc.) across the night sky. To accommodate this data, the geocentric theory was augmented with the assumption that the other planets revolve around points that revolve around the Earth. This additional assumption is *ad hoc* in that it is unconnected with the rest of the theory and motivated only by the need to fit the data—the assumption could not be confirmed independently of confirming the theory. The heliocentric explanation, having all planets move around the Sun, eschews this *ad hoc* assumption. Retrograde motion falls out as a natural consequence of the positions of the Earth and other planets relative to the Sun. Tellingly, as more accurate data became available, the geocentric theory had to be further augmented with epicycles on epicycles to account for planetary motion; not so for the heliocentric theory.

The problem for Classical and Connectionist theories is that they cannot explain systematicity without recourse to their own *ad hoc* assumptions (Aizawa, 2003b). For Classicism, having a combinatorial syntax and semantics does not differentiate between grammars such as G1 and G2. For Connectionism, a common recourse to learning also does not work, whereby systematicity is acquired by adjusting network parameters (e.g., connection weights) to realize some behaviours—training set—while generalizing to others—test set. Learning also requires *ad hoc* assumptions, because even widely used learning models, such as feedforward (Rumelhart, Hinton, & Williams, 1986) and simple recurrent networks (Elman, 1990), fail to achieve systematicity (Marcus, 1998; Phillips, 2000) when construed as a degree of generalization (Hadley, 1994; Niklasson & Gelder, 1994). Hence, neither Classical nor Connectionist proposals satisfy the explanatory standard laid out by Aizawa, or Fodor and Pylyshyn for that matter.

Our category-theory based approach addresses the problem of *ad hoc* assumptions because the concept of an adjunction, which is central to our argument, ensures that the construct we seek (a) exists, and (b) is unique. That is to say, from this core assumption and category theory principles, the systematicity property necessarily follows for the particular cognitive domains of interest, because in each case the one and only collection of cognitive capacities derived from our theory is the systematic collection, without further restriction by additional (*ad hoc*) assumptions.

Basic category theory

Category theory is a theory of structure *par excellence* (see Awodey, 2006; Mac Lane, 2000, for introductions). It was developed out of a need to formalize commonalities between various mathematical structures (Eilenberg & Mac Lane, 1945), and has been used extensively in computer science for the analysis of computation (see, e.g., Pierce, 1991; Walters, 1991). Yet, applications to cognitive psychology have been almost non-existent (but, see Halford & Wilson, 1980; Phillips, Wilson, & Halford, 2009, for two examples). Our explanation of systematicity with respect to binary relational propositions is based on the concept of an *adjunction*. In this section, we provide definitions of this and other formal concepts that it depends.

Category

A *category* $\mathbf{C}$ consists of a class of objects $|\mathbf{C}| = (A, B, \dots)$; a set $\mathbf{C}(A,B)$ of morphisms (also called arrows, or maps) from A to B where each morphism $f : A \to B$ has A as its domain and B as its codomain, including the *identity* morphism $1_A : A \to A$ for each object A; and a composition operation, denoted "$\circ$", of morphisms $f : A \to B$ and $g : B \to C$, written $g \circ f : A \to C$ that satisfy the laws of:

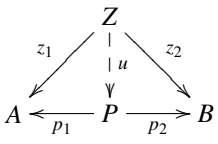
- *unity*, where $f \circ 1_A = f = 1_B \circ f$, for all $f : A \to B$; and
- *associativity*, where $h \circ (g \circ f) = (h \circ g) \circ f$, for all $f : A \to B$, $g : B \to C$ and $h : C \to D$.

The most familiar example of a category is **Set**, which has sets for objects and functions for morphisms, where the identity morphism 1_A is the identity function and the composition operation is the usual function composition operator "$\circ$".

A morphism $f : A \to B$ is an *isomorphism* if there exists a $g : B \to A$, such that $g \circ f = 1_A$ and $f \circ g = 1_B$. In this case, A is said to be isomorphic to B, written $A \cong B$.

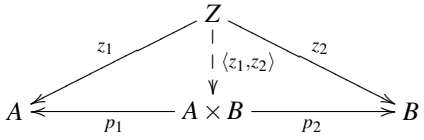
Product

A *product* of two objects A and B in a category $\mathbf{C}$ is an object P together with two morphisms $p_1 : P \to A$ and $p_2 : P \to B$, such that for any pair of morphisms $z_1 : Z \to A$ and $z_2 : Z \to B$, there is a unique morphism $u : Z \to P$, such that the following diagram commutes:



(1)

where a broken arrow indicates that there exists exactly one morphism making the diagram commute. That is, the compositions along any two paths with the same start object and the same finish object are the same. So, in this diagram, $z_1 = p_1 \circ u$ and $z_2 = p_2 \circ u$, where p_1 and p_2 are sometimes called projection morphisms. A product object P is *unique up to a unique isomorphism*. That is, for any other product object P' with morphisms $p'_1 : P' \to A$ and $p'_2 : P' \to B$ there is one and only one isomorphism between P and P' that makes a diagram like this one commute. Hence, P is not unique, only unique with respect to another product object via isomorphism. In **Set**, P is (up to isomorphism) the Cartesian product $A \times B$, $p_1 : A \times B \to A$, $p_2 : A \times B \to B$, where p_1 and p_2 are the projection maps to A and B, i.e., $p_1 : (a,b) \mapsto a$, and $p_2 : (a,b) \mapsto b$, and u is the function $\langle z_1, z_2 \rangle : Z \to A \times B$, sending x to tuple $(z_1(x), z_2(x))$, so that $p_1 \circ u = z_1$ and $p_2 \circ u = z_2$. (The $\mapsto$ arrow, often read as "maps to", indicates the action of a function on a domain element. Thus $f(a) = b$ is equivalent to $f : a \mapsto b$.) Since u is uniquely determined by z_1 and z_2, u is often written as $\langle z_1, z_2 \rangle$, and the diagram used in defining a product then becomes



(2)

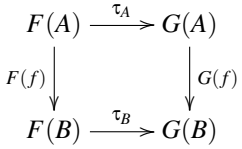
Functor

A functor $F : \mathbf{C} \to \mathbf{D}$ is a structure-preserving map between categories $\mathbf{C}$ and $\mathbf{D}$ that associates each object A in $\mathbf{C}$ to an object $F(A)$ in $\mathbf{D}$; and each map $f : A \to B$ in $\mathbf{C}$ to a map $F(f) : F(A) \to F(B)$ in $\mathbf{D}$, such that $F(1_A) = 1_{F(A)}$ for each object A in $\mathbf{C}$; and $F(g \circ_{\mathbf{C}} f) = F(g) \circ_{\mathbf{D}} F(f)$ for all maps $f : A \to B$ and $g : B \to C$ for which compositions $\circ_{\mathbf{C}}$ and $\circ_{\mathbf{D}}$ are defined in categories $\mathbf{C}$ and $\mathbf{D}$, respectively. The object and arrow components of a functor are sometimes explicitly distinguished as F_0 and F_1, respectively. Otherwise, the functor component is implicitly identified by its argument.

Functor composition and isomorphism are defined analogously to maps (above). That is, the composition of functors $F : \mathbf{C} \to \mathbf{D}$ and $G : \mathbf{D} \to \mathbf{E}$ is the functor $G \circ F : \mathbf{C} \to \mathbf{E}$, sending all objects A in $\mathbf{C}$ to objects $G \circ F(A)$ in $\mathbf{E}$; and maps $f : A \to B$ in $\mathbf{C}$ to maps $G \circ F(f) : G \circ F(A) \to G \circ F(B)$, such that identity and composition are respected. That is, $G \circ F(1_A) = 1_{G \circ F(A)}$; and $G \circ F(g \circ_{\mathbf{C}} f) = (G \circ F(g)) \circ_{\mathbf{E}} (G \circ F(f))$. A functor $F : \mathbf{C} \to \mathbf{D}$ is an *isomorphic functor*, if and only if there exists a functor $G : \mathbf{D} \to \mathbf{C}$ such that $G \circ F = 1_{\mathbf{C}}$ and $F \circ G = 1_{\mathbf{D}}$, where $1_{\mathbf{C}}$ and $1_{\mathbf{D}}$ are the identity functors sending objects and maps to themselves in the respective categories.

Natural transformation

A *natural transformation* $\tau : F \to G$ is a structure-preserving map from domain functor $F : \mathbf{C} \to \mathbf{D}$ to codomain functor $G : \mathbf{C} \to \mathbf{D}$ that consists of $\mathbf{D}-$maps τ_A for each object A in $\mathbf{C}$, such that $G(f) \circ \tau_A = \tau_B \circ F(f)$, as indicated by the following commutative diagram in the category $\mathbf{D}$:

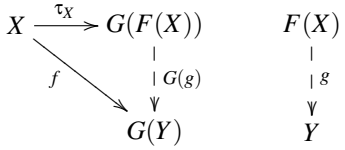


(3)

A natural transformation is a *natural isomorphism*, or *natural equivalence* if and only if each τ_A is an isomorphism. That is, for each $\tau_A : F(A) \to G(A)$ there exists a $\tau_A^{-1} : G(A) \to F(A)$ such that $\tau_A^{-1} \circ \tau_A = 1_{F(A)}$ and $\tau_A \circ \tau_A^{-1} = 1_{G(A)}$. Natural transformations also compose, and the composition of two natural transformations is also a natural transformation.

Adjunction

An *adjunction* consists of a pair of functors $F : \mathbf{C} \to \mathbf{D}$, $G : \mathbf{D} \to \mathbf{C}$ and a natural transformation $\tau : 1_{\mathbf{C}} \to (G \circ F)$, such that for every $\mathbf{C}-$object X and every $\mathbf{C}-$map $f : X \to G(Y)$ there exists a unique $\mathbf{D}-$map $g : F(X) \to Y$, such that the following diagram commutes:



(4)

where the functors are implicitly identified by (co)domain categories $\mathbf{C}$ (left subdiagram) and $\mathbf{D}$ (right subdiagram). The two functors are called an *adjoint pair*, (F, G), where F is the *left adjoint* of G, and G is the *right adjoint* of F; and natural transformation τ is called the *unit* of the adjunction.

Category theory explanation: Adjoint functors

We develop our adjoint functors explanation of systematicity in three movements. First, we show that a categorical product provides an account of systematicity of representation and systematicity of inference. However, a product of two objects may afford many isomorphic product objects that do not also account for compositionality of representation. Second, we show that the product functor provides the principled means for constructing only those products that also have the compositionality of representation property. However, there may be more than one product that has the compositionality property, but differs in semantic content by having different syntactic relationships between identical sets of constituents. So, a principled choice is needed to determine *the* product. Third, we show that the diagonal functor, which is left adjoint to the product functor, provides that principled choice. For concreteness, we refer to the category **Set**, but our explanation does not depend on this category.

First, suppose objects A (say, agents) and B (patients) are sets containing representations of John and Mary, denoted as $\{\mathtt{J},\mathtt{M}\}$. Although A and B are the same set in this example they may not be in the general case. Since our argument does not depend on equality, we maintain distinct names for generality, and for conceptual clarity. A categorical product of these two sets is the Cartesian product of A and B, which is the set of all pairwise combinations of elements from A and B, together with maps p_1 and p_2 for retrieving the first and second constituent in each case. That is, $A \times B = \{(\mathtt{J},\mathtt{J}),(\mathtt{J},\mathtt{M}),(\mathtt{M},\mathtt{J}),(\mathtt{M},\mathtt{M})\}$, $p_1 : (a,b) \mapsto a$, and $p_2 : (a,b) \mapsto b$. By definition, the Cartesian product, $A \times B$, generates all pairwise combinations of elements from A and B, therefore the Cartesian product has the systematicity of representation property. Moreover, by definition, the categorical product, $(A \times B, p_1, p_2)$, affords the retrieval of each constituent from each representation (otherwise it is not a product), therefore the categorical product also has the systematicity of inference property. In this case, Z from the categorical product definition takes the role of input, so inferring John as the lover from John loves Mary is just $z_1(\mathtt{JM}) = p_1 \circ u(\mathtt{JM})$, where JM is the input and u is the input-to-product object map, whose unique existence is guaranteed.

The Cartesian product, however, is not the only product object that satisfies the definition of a categorical product of A and B. An alternative product has $P = \{1,2,3,4\}$ as the product object, and $p'_1 : 1 \mapsto \mathtt{J}, 2 \mapsto \mathtt{J}, 3 \mapsto \mathtt{M}, 4 \mapsto \mathtt{M}$ and $p'_2 : 1 \mapsto \mathtt{J}, 2 \mapsto \mathtt{M}, 3 \mapsto \mathtt{J}, 4 \mapsto \mathtt{M}$ as the projections. However, this alternative does not have the compositionality of representation property: the semantic contents of these representations, whatever they may be, are not systematically related to each other, or the semantic content of John, or Mary. Hence, categorical products, in themselves, are not sufficient for an explanation of systematicity.

Second, for any category $\mathbf{C}$ that has products (i.e. every pair of objects in $\mathbf{C}$ has a product), one can define a product functor $\Pi : \mathbf{C} \times \mathbf{C} \to \mathbf{C}$, that is from the Cartesian product of categories, $\mathbf{C} \times \mathbf{C}$, itself a category, to $\mathbf{C}$, where $\Pi_0 : (A,B) \mapsto A \times B$, $\Pi_1 : (f,g) \mapsto f \times g$, as indicated by the following diagram:

$$\begin{array}{ccc} (A,B) & \xmapsto{\Pi_0} & A \times B \\ {\scriptstyle (f,g)}\downarrow & & \downarrow{\scriptstyle f \times g} \\ (C,D) & \xmapsto[\Pi_0]{} & C \times D \end{array} \tag{5}$$

omitting $\Pi_1 : (f,g) \mapsto f \times g$ for clarity. In this case, the semantic contents of these elements are systematically related to each other and their constituents John and Mary. This categorical construction is an instance of Classical compositionality, whereby the constituents $a_i \in A$, $b_j \in B$ are tokened wherever the compositions $(a_i, b_j) \in A \times B$ are tokened. As such, it has the compositionality of representation property.

Although the product functor explanation accounts for compositionality of representation, it introduces a new problem: $(B \times A, p'_2, p'_1)$, where $p'_2 : (b,a) \mapsto a$ and $p'_1 : (b,a) \mapsto b$ is also a valid product, but the semantic content of (a,b) is not the same as (b,a). That is because they have different order relationships between their constituents even though the corresponding constituents are identical. Thus, a principled choice is required to determine whether, for example, John loves Mary should map to (John, Mary), or (Mary, John). Otherwise, one can define an architecture that does not have the systematicity of inference property by employing both products to correctly infer John as the lover in John loves Mary via $(A \times B, p_1, p_2)$, yet incorrectly infer John as the lover in Mary loves John via $(B \times A, p'_2, p'_1)$, where position within the product triple identifies the relevant projection. The assumption that architectures employ only the first product is *ad hoc* just like the assumption that Classical architectures employ grammars such as G1, but not G2. So, a principled choice is needed to determine *the* product.

Third, and finally, the left adjoint to the product functor is the *diagonal* functor $\Delta : \mathbf{C} \to \mathbf{C} \times \mathbf{C}$, where $\Delta_0 : A \mapsto (A,A)$, $\Delta_1 : f \mapsto (f,f)$ as indicated by the following diagram:

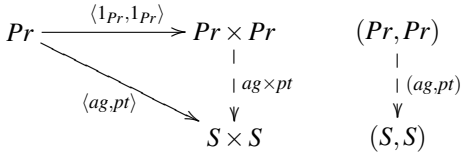
$$\begin{array}{ccc} A & \xmapsto{\Delta_0} & (A,A) \\ {\scriptstyle f}\downarrow & & \downarrow{\scriptstyle (f,f)} \\ B & \xmapsto[\Delta_0]{} & (B,B) \end{array} \tag{6}$$

The (diagonal, product) adjoint pair is indicated by the following commutative diagram:

$$\begin{array}{ccccc} C & \xrightarrow{\tau_C = \langle 1_C, 1_C \rangle} & C \times C & \qquad & (C,C) \\ & \searrow^{\langle s,t \rangle} & \downarrow{\scriptstyle s \times t} & & \downarrow{\scriptstyle (s,t)} \\ & & M \times N & & (M,N) \end{array} \tag{7}$$

(see Pierce, 1991, Example 2.4.6). In this manner, the John loves Mary family of cognitive capacities is specified by the

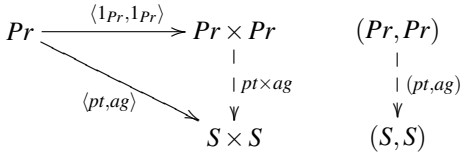
commutative diagram



(8)

where *ag* and *pt* are the *agent* and *patient* maps from the set of proposition inputs *Pr* into the set $S \supseteq A \cup B$ containing all the possible constituent representations. Given $\langle ag, pt \rangle$ as the morphism used by the architecture to map proposition inputs to their corresponding internal representations, then as mentioned (Introduction) the definition of an adjunction guarantees that $ag \times pt$ is unique with respect to making Diagram 8 commute. That is, $ag \times pt \circ \langle 1_{Pr}, 1_{Pr} \rangle(\texttt{JM}) = ag \times pt(\texttt{JM}, \texttt{JM}) = (\texttt{John}, \texttt{Mary}) = \langle ag, pt \rangle(\texttt{JM})$, where JM is the input for proposition John loves Mary. The alternative construction $pt \times ag$ is excluded because $pt \times ag \circ \langle 1_{Pr}, 1_{Pr} \rangle(\texttt{JM}) = pt \times ag(\texttt{JM}, \texttt{JM}) = (\texttt{Mary}, \texttt{John}) \neq (\texttt{John}, \texttt{Mary}) = \langle ag, pt \rangle(\texttt{JM})$. Having excluded $pt \times ag$ by the commutativity property of the adjunction, the only two remaining ways to map the other inputs (i.e., $\langle ag, pt \rangle$ and $ag \times pt \circ \langle 1_{Pr}, 1_{Pr} \rangle$) are equal. So, given that the architecture can represent John loves Mary as $(\texttt{John}, \texttt{Mary})$ via $\langle ag, pt \rangle$ and infer John as the lover via p_1 from the product $(A \times B, p_1, p_2)$, then necessarily it can represent Mary loves John and infer Mary as the lover using the same maps. That is, $p_1 \circ \langle ag, pt \rangle(\texttt{MJ}) = p_1(\texttt{Mary}, \texttt{John}) = \texttt{Mary}$, or $p_1 \circ ag \times pt \circ \langle 1_{Pr}, 1_{Pr} \rangle(\texttt{MJ}) = p_1 \circ ag \times pt(\texttt{MJ}, \texttt{MJ}) = p_1(\texttt{Mary}, \texttt{John}) = \texttt{Mary}$.

This explanation works regardless of whether proposition John loves Mary is represented as (John, Mary) via $\langle ag, pt \rangle$, or (Mary, John) via $\langle pt, ag \rangle$. In the latter case, the adjunction picks out the construction $pt \times ag$, because it is the one and only one that makes the following diagram commute:



(9)

$pt \times ag \circ \langle 1_{Pr}, 1_{Pr} \rangle(\texttt{JM}) = pt \times ag(\texttt{JM}, \texttt{JM}) = (\texttt{Mary}, \texttt{John}) = \langle pt, ag \rangle(\texttt{JM})$, but $ag \times pt \circ \langle 1_{Pr}, 1_{Pr} \rangle(\texttt{JM}) = ag \times pt(\texttt{JM}, \texttt{JM}) = (\texttt{John}, \texttt{Mary}) \neq (\texttt{Mary}, \texttt{John}) = \langle pt, ag \rangle(\texttt{JM})$. Given that the architecture can represent John loves Mary as $(\texttt{Mary}, \texttt{John})$ via $\langle pt, ag \rangle$ and infer John as the lover via p'_2 from the product $(B \times A, p'_2, p'_1)$, then necessarily it can do so for Mary loves John using the same maps. That is, $p'_2 \circ \langle pt, ag \rangle(\texttt{MJ}) = p'_2(\texttt{John}, \texttt{Mary}) = \texttt{Mary}$, or $p'_2 \circ pt \times ag \circ \langle 1_{Pr}, 1_{Pr} \rangle(\texttt{MJ}) = p'_2 \circ pt \times ag(\texttt{MJ}, \texttt{MJ}) = p'_2(\texttt{John}, \texttt{Mary}) = \texttt{Mary}$.

Importantly, the unit of the adjunction, $\langle 1_{Pr}, 1_{Pr} \rangle$, is not a *free parameter* of the explanation; it defines the adjunction. Also, there is no choice in representational format (i.e. left-right, or right-left constituent order)—the given capacity to represent a proposition fixes the same order for all the other propositions. Hence, systematicity is a necessary consequence of this adjoint pair without recourse to additional (*ad hoc*) assumptions, and so meets the explanatory standard set by Aizawa, and Fodor and Pylyshyn.

Explanatory levels: *n*-category theory

A generalization of category theory, called *n*-category theory (see Leinster, 2003) is used to formally contrast our category theory explanation against Classical and Connectionist approaches. Notice that the definitions of functor and natural transformation are very similar. In fact, they are morphisms at different levels of analysis. For *n*-category theory, a category such as **Set** is a 1-category, with 0-objects (i.e. sets) for objects and 1-morphisms (i.e. functions) for arrows. A functor is a morphism between categories. The category of categories, **Cat**, has categories for objects and functors for arrows. Thus, a functor is a 2-morphism between 1-objects (i.e. 1-categories) in a 2-category. A natural transformation is a morphism between functors. The functor category, **Fun**, has functors for objects and natural transformations for arrows. Thus, a natural transformation is a 3-morphism between 2-objects (i.e. functors) in a 3-category. (A 0-category is just a *discrete* category, where the only arrows are identities, which are 0-morphisms.) In this way, the order *n* of the category provides a formal notion of explanatory level.

Classical or Connectionist compositionality is essentially a lower levels attempt to account for systematicity. For the examples we used that level is perhaps best described in terms of a 1-category. Indeed, a context-free grammar defined by a graph is modeled as the *free* category on that graph containing sets of terminal and non-terminal symbols for objects and productions for morphisms (Walters, 1991). By contrast, our category theory explanation involves higher levels of analysis, specifically functors and natural transformations, which live in 2-categories and 3-categories, respectively. Of course, one can also develop higher-order grammars that take as input or return as output other grammars. Similarly, one can develop higher-order networks that take as input or return as output other networks. However, the problem is that neither Classical nor Connectionist compositionality delineates those (higher-order) grammars or networks that have the systematicity property from those that do not.

Discussion

In addition to explaining systematicity, our category theory approach has further implications. According to our explanation, systematicity with respect to binary relational propositions requires a category with products. Phillips et al. (2009) also provided a category theory account of the strikingly similar profiles of development for a suite of reasoning abilities that included *Transitive Inference* and *Class Inclusion*, among others—all abilities are acquired around the age of five years. The difference between the failures of younger children and the successes of older children (relative to age five) across all these reasoning tasks was explained as their capacity to compute (co)products. (A *coproduct* is related to a product

by arrow reversal—see, e.g., Pierce, 1991, for a formal definition.) Therefore, our explanation implies that systematicity is not a property of younger children's cognition. Some support for this implication is found on memory tasks that require binding the background context of memorized items (Lloyd, Doydum, & Newcombe, 2009), though further work is needed to test this implication directly.

Our explanation does not depend on **Set**, it only requires a category with products. For example, the categories **Top** of topological spaces and continuous mappings, and **Vec** of vector spaces and linear mappings (see, e.g., Awodey, 2006) could also be used. These possibilities imply that an explanation of systematicity does not depend on a particular (discrete symbolic, or continuous subsymbolic) representational format. Thus, a further benefit is that our approach opens the way for integration of other (sub/symbolic) levels of analysis.

For reasons of space, we have only sketched our category theory approach to systematicity. More detailed explanation and justification are given in Phillips and Wilson (in prep.), where we also address other examples of systematicity, such as multiple relations, and relational schemas. In our approach, we have not dealt with domains that are quasi-systematic, which appear to be particularly prevalent in language (see Johnson, 2004). For these cases, we would also need category theory-derived principled restrictions to products. *Pullbacks* (see Phillips, Wilson, & Halford, 2009, for an application to cognitive development) are one way to restrict product objects, in the same arrow-theoretic style.

From a category theory perspective, we now see why cognitive science lacked a satisfactory explanation for systematicity—cognitive scientists were working with lower-order theories in attempting to explain an essentially higher-order property. Category theory offers a re-conceptualization for cognitive science, analogous to the one that Copernicus provided for astronomy, where representational states are no longer the center of the cognitive universe—replaced by the relationships between the maps that transform them.

Acknowledgment. We thank the reviewers for extensive comments to help clarify the presentation of this work.

References

Aizawa, K. (2003a). Cognitive architecture: The structure of cognitive representations. In S. P. Stich & T. A. Warfield (Eds.), *The Blackwell guide to philosophy of mind* (pp. 172–189). Cambridge, MA: Blackwell.

Aizawa, K. (2003b). *The systematicity arguments*. New York: Kluwer Academic.

Awodey, S. (2006). *Category theory*. New York, NY: Oxford University Press.

Eilenberg, S., & Mac Lane, S. (1945). General theory of natural equivalences. *Transactions of the American Mathematical Society*, *58*, 231–294.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*, 179–211.

Fodor, J. A., & McLaughlin, B. P. (1990). Connectionism and the problem of systematicity: Why Smolensky's solution doesn't work. *Cognition*, *35*, 183–204.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*, 3–71.

Hadley, R. F. (1994). Systematicity in connectionist language learning. *Mind and Language*, *9*(3), 247–272.

Halford, G. S., & Wilson, W. H. (1980). A category theory approach to cognitive development. *Cognitive Psychology*, *12*, 356–411.

Johnson, K. (2004). On the systematicity of language and thought. *The Journal of Philosophy*, *101*(3), 111–139.

Leinster, T. (2003). *Higher operads, higher categories*. Cambridge: UK: Cambridge University Press.

Lloyd, M. E., Doydum, A. O., & Newcombe, N. S. (2009). Memory binding in early childhood: evidence for a retrieval deficit. *Child Development*, *80*(5), 1321–1328.

Mac Lane, S. (2000). *Categories for the working mathematician* (2nd ed.). New York, NY: Springer.

Marcus, G. F. (1998). Rethinking eliminative connectionism. *Cognitive Psychology*, *37*(3), 243–282.

Niklasson, L., & Gelder, T. van. (1994). Systematicity and connectionist language learning. *Mind and Language*, *9*(3), 288–302.

Phillips, S. (2000). Constituent similarity and systematicity: The limits of first-order connectionism. *Connection Science*, *12*(1), 1–19.

Phillips, S. (2007). Kenneth Aizawa, The systematicity arguments, Studies in brain and mind. *Minds and Machines*, *17*(3), 357–360.

Phillips, S., & Wilson, W. H. (in prep.). *Categorial compositionality: A category theory explanation for the systematicity of human cognition.*

Phillips, S., Wilson, W. H., & Halford, G. S. (2009). What do Transitive Inference and Class Inclusion have in common? Categorical (co)products and cognitive development. *PLoS Computational Biology*, *5*(12), e1000599.

Pierce, B. C. (1991). *Basic category theory for computer scientists*. Cambridge, UK: MIT Press.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagation of error. *Nature*, *323*, 533–536.

Smolensky, P. (1987). The constituent structure of connectionist mental states: A reply to Fodor and Pylyshyn. *Southern Journal of Philosophy*, *26*, 137–161.

Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, *46*(1-2), 159–216.

van Gelder, T. (1990). Compositionality: A connectionist variation on a classical theme. *Cognitive Science*, *14*, 355–384.

Walters, R. F. C. (1991). *Categories and computer science*. Cambridge, UK: Cambridge University Press.

The Effect of Word-internal Properties on Syntactic Categorization: A Computational Modeling Approach

Fatmeh Torabi Asr, Afsaneh Fazly, Zohreh Azimifar
Computer Sciences and Engineering
Shiraz University
Shiraz, Iran
{torabi,fazly,azimifar}@cse.shirazu.ac.ir

Abstract

We study the acquisition of abstract syntactic categories of words in children by using a computational model of categorization. Especially, we examine the effect of word-internal properties, such as morphological and phonological cues, on the identification of different categories, such as nouns, verbs, and determiners. To evaluate our model, we use it to determine the syntactic category of actual novel words selected from naturalistic child-directed utterances. We argue that such an evaluation is necessary for a better understanding of the effect of different cues (including word-internal properties and contextual cues) on category acquisition.

Keywords: Computational modeling, Syntactic category learning.

Introduction

Infants have a good understanding of the syntactic categories of words long before attending school. Psychological observations at different stages of child language development have shown the ability of children to recognize abstract (syntactic or semantic) categories, such as verb and noun, countable and uncountable (Brown, 1957; Gelman & Taylor, 1984; Samuelson & Smith, 1999). A variety of proposals exist in the psycholinguistics literature regarding the types of cues that are informative about such word categories, and the way children may use them to learn the categories. Computational modeling has often been used as a powerful tool to shed light on many aspects of language acquisition, including word categorization (Pearl, 2009). In this study, we draw on an existing categorization model in order to achieve a better understanding of the mechanisms and the information sources children use during the acquisition of syntactic categories, such as verbs and nouns.

Syntactic category learning in children has been suggested to be based on several information sources, such as word-external properties including distributional information about neighboring (co-occurring) words, as well as word-internal properties such as phonological and morphological cues (e.g., Brown, 1957; Gerken et al., 2005; Monaghan et al., 2007). Many of the computational studies on syntactic category acquisition focus on showing the relevance of the above properties to the acquisition of adult-like, linguistically-salient categories, such as verbs, nouns, and adjectives. For example, Mintz (2003), Monaghan et al. (2007) and Onnis and Christiansen (2008) present analyses of child-directed data to determine whether there are correspondences between particular syntactic categories and certain types of cues. Moreover, most of the existing computational models of child category learning lack cognitive plausibility in some respects: The categorization models proposed by Schütze (1993), Redington et al. (1998), and Clark (2003) incorporate a batch (non-incremental) clustering algorithm ; The connectionist model of Onnis and Christiansen (2008) is minimally supervised, assumes a fixed number of categories, and can only be used to study words in isolation.

A few studies have introduced cognitively-plausible models for syntactic category learning (Cartwright & Brent, 1997; Parisien et al., 2008; Alishahi & Chrupała, 2009). These incorporate fully-unsupervised incremental algorithms for clustering words as they appear in naturally-occurring utterances. However, these studies have focused solely on the role of context (co-occurring words) for inferring the syntactic category of a target word, and have overlooked the importance of other sources of information, such as phonology and morphology.

In our modeling of syntactic category acquisition, we address some of the aforementioned shortcomings. Specifically, we choose a simple incremental clustering algorithm (one proposed by Alishahi & Chrupała, 2009), which we further modify to increase simplicity. In addition, we examine the role of word-external information sources (namely, word co-occurrence), as well as that of word-internal sources (namely, phonology, and morphology) in order to better understand the interactions among these types of cues on the acquisition of syntactic categories. We use only very simple cues that are known to be accessible by children early in their language development. Finally, we propose and use a novel evaluation framework to examine the role of each type of information in the acquisition of syntactic categories.

Results of our experiments on naturally-occurring English child-directed utterances indicate that different cues are useful for the identification of different classes of words. In particular, we find that the identity of the word is essential to the identification of closed-class words. Open-class words, however, share similarities with respect to other types of cues, both word-external and word-internal. Nonetheless, even among these classes, different categories seem to be identified based on different properties: whereas verbs are better categorized with the help of morphological and phonological properties, co-occurrence information alone is reliable for categorizing nouns.

Algorithm 1: Incremental word clustering

initialize set of clusters $\mathcal{K} = \emptyset$
for every frame f **do**
  $C_M = argmax_{C \in \mathcal{K}} Sim(\mathrm{f}, C)$
  if $Sim(\mathrm{f}, C_M) \geq \theta$ **then**
    Add frame f to cluster C_M
  else
    Construct a new cluster for frame f
  end if
end for

(This algorithm is a modification of the one proposed by Alishahi & Chrupała, 2009).

Modeling the acquisition of syntactic categories

Our goal is to build a computational model of syntactic categorization that is cognitively plausible, i.e., we make as few assumptions as possible about the type of cues accessible to young children, and about the mechanisms children might use for categorization. We thus use an adaptation of a simple incremental algorithm proposed by Alishahi and Chrupała (2009), which forms categories simply by drawing on the similarity among words to be categorized. Here, we present an overview of our adaptation of the algorithm, and a description of three types of cues we use for categorization.

The categorization algorithm

The unsupervised clustering algorithm proposed by Alishahi and Chrupała (2009) works based on contextual similarities among words. The algorithm is incremental in that it processes words one by one, discarding each word after clustering. For each newly-observed *frame* (a target *head-word* along with its neighboring words from left and right), if the similarity to all of the already-shaped clusters is less than a predefined threshold, a new cluster is constructed. Otherwise, the word is assigned to the most similar cluster. We modify this algorithm in two ways: (i) the original algorithm of Alishahi and Chrupała includes a phase in which clusters are merged if they are sufficiently similar. To keep the algorithm simple, we removed this step; (ii) our frames are composed of three different types of *features* (five features in total besides the head-word content; see next subsection for details). We thus need to slightly modify the similarity score calculation in order to accommodate for more than one set of features. The similarity between a frame and a cluster (a group of frames) is calculated as in:

$$Sim(\mathrm{f}, C) = \sum_{i \in \mathcal{F}} \omega_i * Sim_i(\mathrm{f}, C) \in \mathcal{F} \qquad (1)$$

where f is a frame, C is a cluster, i is a feature, $\mathcal{F}$ is the set of all features, $Sim_i(\mathrm{f}, C)$ is the similarity of frame f to cluster C with respect to the ith feature, and ω_i determines the weight of the contribution of feature i in determining the overall similarity. Weights for all features need to sum to 1, i.e., $\sum_i \omega_i = 1$. The modified version of the algorithm is shown in Algorithm 1 .

Cues used for categorization

As previously mentioned, children are known to group words into syntactic categories by drawing on a number of different information sources. In our work, we include three different sources of information, and five types of cues (features) in total, as explained below:[1]

- Distributional information about word co-occurrences: This kind of information has been reported to be reliable and very important in syntactic categorization (Schütze, 1993; Redington et al., 1998; Mintz, 2003; Clark, 2000; Parisien et al., 2008; Alishahi & Chrupała, 2009). We take one word from each side of a target head-word as its co-occurrence features, because in many of the above studies words closer to a word have been shown to be more informative about its category. For example, considering sentences, such as "There is a *cat* in the basket", and "We need a *table* in our kitchen", "A cat is in the basket", and "A table is in the kitchen." provides a clue to the model to group *cat* and *table* together since they share similar co-occurrence features. In our framework, each co-occurring word is considered as an independent feature when determining similarity between a word (frame) and a cluster (as in many previous studies, and in contrast to representations such as "frequent frames" of Mintz, 2003). For example, even if the two tokens *cat* and *table* did not share the preposition *in*, they would still be considered as similar because of the preceding determiner *a* they have in common.

- Phonological information: Words belonging to the same syntactic category tend to have common phonological properties. For example, looking at child-directed utterances, (Monaghan et al., 2007) show that verbs and nouns are different with respect to several phonological features, including the number of syllables. The study of Monaghan et al. focuses on the relevance of syntactic categories and a large number of word-level, syllable-level, and phoneme-level phonological properties. We focus here on two of the simplest word-level phonological properties that we assume are readily accessible by young children, namely the length of a word in terms of number of syllables and phonemes (we use the number of letters to approximate the number of phonemes in a word).

- Morphological information: It has been shown that English affixes, such as *-ing* in verbs, can provide strong clues to the identification of syntactic categories, and that such information is abundant in child-directed speech (Onnis & Christiansen, 2008). Nonetheless, it is not clear whether we can assume that children have access to such accurate morphological knowledge about words and categories prior to syntactic category learning. Inspired by the work

[1] In this study, we do not consider one other important source of information for learning of syntactic categories, namely, semantic information about words. This type of information requires making assumptions about what *meaning* is and how children may represent it, and hence is outside the scope of this study.

of Onnis and Christiansen (2008), here we use the last phoneme (ending) of the words as an approximation of the morphological affixes.[2]

Overall, we include six different features (cues) in our categorization: two Cooc features, Head word, two Phon features, and one Morph feature. The Cooc cues are considered as properties external to the word (properties of the context the word appears in), whereas the rest are related to the word itself and hence are considered as word-internal cues. In our experiments, we examine the effect of each different type of cue on categorization, and also consider the role of word-internal cues versus external ones.

Experimental Setup

Corpus

We extract our input data (both for training and testing) from the Manchester corpus (Theakston et al., 2001), one of the English subsets in the CHILDES database (MacWhinney, 2000). The Manchester corpus contains conversations of parents/caregivers with 12 British children between the ages of 1;8 (years;months) and 3;0.[3] For training, we choose around 10000 child-directed utterances from the conversations of all 12 children, such that the chronological order of the utterances is maintained, and the utterances contain only words selected from a limited vocabulary of 500 words. When selecting the 500 words, we make sure that their distribution in the corpus matches a Zipfian distribution, so that our results are not biased towards words from certain frequency ranges. We limit the size of vocabulary because some feature values need to be determined manually. In addition, in one experimental task, we need access to actual novel words not previously seen in the training corpus, as opposed to made-up novel words used in many psychological experiments.

We use two different test corpora, one for each experimental task (as explained in the Evaluation subsection below). The first set of test data (used in the Word Category Prediction Task) is selected exactly as the training data, though from a non-overlapping portion of the original (Manchester) corpus. The other test data (used in the Novel Word Categorization Task) is selected such that the target words to be categorized are a novel word not in the vocabulary of 500 words. This second test set is similar to the training data in all other aspects. Each test corpus contains 2000 word usages (tokens to be categorized).

Feature Extraction

From each utterance (in the training or test data), we extract a number of frames to be clustered. As explained previously, each frame contains a head word (the target word to be categorized), as well as several other features (two Cooc, two Phon, and one Morph features). A sample frame is shown in Figure 1. The head word and the Cooc features can be directly extracted from the utterance. If any of the Cooc features are missing (i.e., the target word is the first or the last word of the utterance), that feature is set to "Unknown". For the two other types of features (Phon and Morph) we need to have access to a phonemic representation of words and other phonological features. We extract two of these features (the ending phoneme, and the number of syllables) from the MRC Psycholinguistic Database, a publicly available resource built for use in studies on child language (Wilson, 1988).[4] If a word is not found in MRC, we set the values of the above features manually. For the third feature, the number of phonemes in a word, we use the number of letters as an approximation.

Head:	*table*	Cooc:	*a*, *in*
Phon:	2, 5	Morph:	*l*

Figure 1: Sample frame extracted for the target word *table* from the utterance "We need a *table* in the kitchen".

Evaluation

To examine the contribution of different types of cues on syntactic categorization, we evaluate the effectiveness of clusters resulting from one or a combination of features in two tasks. Specifically, we train our model (on the training corpus) in three different conditions, that is, using one of the following feature combinations: Head+Cooc, Head+Cooc+Morph, Head+Cooc+Phon. We then determine the effectiveness of the resulting clusters in each condition by examining the performance of the model on inferring the category of a number of test words. Note that the model does not create any new clusters during the test phase, but assigns each word to one of the clusters formed in the training phase.

We evaluate our model using two experimental tasks: one is to predict the syntactic category of a word whose identity is known to the model/learner; the other one is to infer the syntactic category of a novel (previously-unseen) word. In the word category prediction task (Experiment 1) the Head of a frame is considered as a feature, whereas it is not included in the task of novel word categorization (Experiment 2). More details on each of these tasks is given in the following section.

Note that the resulting categories do not necessarily need to match the conventional adult-like categories put forth by linguists. Nonetheless, as a first-line evaluation, here we compare the categories learned by our model to a gold-standard set of syntactic categories. To measure test performance, we must compare the 'true' syntactic category of each test word (according to the gold-standard) to the label of its associated cluster. We thus need to label each cluster with a syntactic category. Words in the Manchester corpus are tagged with their parts of speech according to a fine-grained tag set. For

[2] We also included the first phoneme (beginning) of a word as also done by Onnis and Christiansen (2008). However, in our initial evaluations we found that the inclusion of this feature did not affect the results, and hence removed it from our set of features.

[3] Thanks to Chris Parisien for providing us with a preprocessed version of this corpus.

[4] http://www.psych.rl.ac.uk/

our evaluation, we use a coarse-grained version of this original tagging (also used by Parisien et al., 2008), including 11 tags, namely: Noun, Verb, Adjective, Adverb, Determiner, Negation, Infinitive, Auxiliary, Conjunction, Preposition, and Others. Each cluster is assigned the majority label among all its members. E.g., a cluster containing 30 nouns, 90 verbs, and 20 adjectives is labeled as Verb.

Test performance is measured using *Accuracy*: the proportion of test words assigned to their correct category. We also look into the accuracy for different groups of words, such as Verbs and Nouns, as well as open-class and closed-class words.

Model Parameters

Our model contains two sets of parameters: the weights ω_i used for measuring the similarity of a frame to a cluster (in Eqn 1), and a similarity threshold θ used for deciding whether to create a new cluster for a given frame. We set the weights ω_i uniformly, giving equal weights to all features. The value of θ affects the number of generated clusters: a low value increases the likelihood of grouping words, hence decreasing the total number of clusters. We set this parameter to different values for different experimental conditions (i.e., different combinations of features), so that we maintain the total number of clusters generated in each condition within a desired range.

We use two different ways of measuring $Sim_i(\mathrm{f}, C)$ in Eqn 1 depending on feature i. For categorical features (Head, Cooc, Morph) we use the cosine of the vectors (widely used for similar clustering algorithms). A vector representing a categorical feature such as Head is of the size of word types in the corpus. E.g., for a sample frame f this vector includes 0 in all elements except where the value of Head in that frame is presented. For numerical features (Phon) we use the Euclidean distance.

Experimental Results

Experiment 1: Word Category Prediction

Recall that to determine the effect of different types of cues (Head, Cooc, Phon, Morph) in the acquisition of syntactic categories, we train our model in three conditions (i.e., using three combinations of features, namely Head+Cooc, Head+Cooc+Phon, and Head+Cooc+Morph). In Experiment 1, we measure the accuracy of category prediction over a test data containing 2000 known words. Comparing the accuracy of the categorization model across these conditions is fair and meaningful only if the number of clusters are relatively close for all conditions. Generally, allowing a larger number of clusters makes the categorization more conservative (i.e., by forming too many small clusters each containing one or a few word types that are highly similar). Based on our observation, this implicitly affects the test accuracy. Hence, in the training phase for each of the three above-mentioned conditions, we use different values for the similarity threshold θ to obtain approximately similar number of final clusters (i.e., between 258–288).[5] This way we maintain one factor (number of clusters) constant, allowing us to focus on the effect of different features involved in categorization.

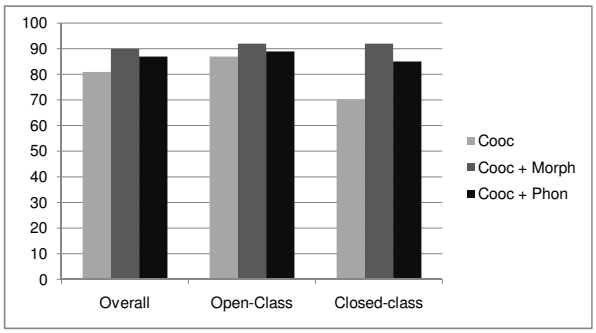


Figure 2: %Accuracy of known-word category prediction in three conditions; the total number of clusters constructed during training phase is in the range 258–288.

Results are presented in Figure 2. In each condition, we measure accuracy on all 2000 words (displayed in the figure as the Overall accuracy), as well as for open-class and for closed-class words separately. Since Head is used as a feature in all conditions, for the ease of exposition, the figure refers to the conditions as Cooc, Cooc+Phon, and Cooc+Morph.

Figure 2 shows that the overall categorization accuracy of the model is improved by adding morphological or phonological information, reinforcing that word-internal features are indeed informative about a word's syntactic category. The best performance is achieved by combining Cooc and Morph features, suggesting that our morphological feature might be more indicative of syntactic category than the phonological features.

Comparing the accuracy on open-class words and on Closed-class words, we can see that in two out of the three conditions (i.e., Cooc and Cooc+Phon), open-class words are better categorized in comparison to closed-class words. This is expected because it is more likely that the word co-occurrence information (which is the main source of information in all conditions) reveals the similarity among open-class (content) words more easily than for closed-class (function) words. As an example, we expect nouns to often appear after a small set of determiner types (e.g., *a, an, the*), whereas determiners may precede many different nouns, sharing fewer context features.

Previous studies have shown a strong effect for the Head feature in determining a word's syntactic category (e.g., Chang et al., 2006). It is thus reasonable to compare the over-

[5] We have performed experiments with different ranges of cluster numbers, and found that the general patterns in results are similar. As noted before, we prefer fewer clusters (fewer than our vocabulary size) to allow for generalization. Indeed, we observe that even with 258–288 clusters, the generalization of the model is reasonably good. Since more than 55% of the clusters contain three or more word types.

all performance of our model in the three conditions with that of a simple category learner that uses only the Head feature, which we refer to as the *lex-stat* learner following Chang et al. (2006). For the performance of our model and that of the lex-stat learner to be comparable, we must set the similarity threshold so that we end up with around 500 clusters for all conditions (since the lex-stat learner constructs a separate cluster for each word type in the vocabulary). Indeed, we find that the overall performance of lex-stat (92%; not shown in Figure 2) is better than for Cooc (89%), and is comparable to the other two categorization conditions, Cooc+Phon (92%), and Cooc+Morph (92%). This raises an important question: whether the positive effect we observe here for the addition of Phon and Morph features is a true effect. In other words, since both Phon and Morph features are word-internal, it is possible that their inclusion in categorization increases the contribution of the Head feature in calculating similarity, implicitly giving more weight to the Head feature.

Note that the lex-stat learner is a very conservative model with no generalization abilities (since each word type is in its own cluster). Such a model thus fails to properly categorize novel (previously unseen) words. In contrast to such a learner, children have the ability to categorize novel words (even meaningless artificial words made up for experimental purposes), by the help of the context, or based on their morphological properties (Brown, 1957). We thus argue that for a categorization model to reveal the true effect of features such as morphology or phonology, it should be able to generalize well on unseen words. In the second Experiment, we use our three categorization models to determine the category of novel words. We consider actual novel words in this task because we want to draw on word-internal features, e.g., phonological and morphological properties of words.

Experiment 2: Novel Word Categorization

In this task, we use our model (in the three conditions) to categorize 2000 novel words. In such cases, the Head feature is not informative (since test words have not been seen during training), and hence the model has to utilize other sources of information to determine the category of a word. Results are presented in Figure 3. Comparing performance on this Experiment with those on Experiment 1 (Figure 2) shows a substantial decrease in the overall categorization accuracy (note that here Head feature is taken out of consideration). We especially observe a significant drop in performance for closed-class words. This decrease in performance emphasizes the importance of the Head feature for word categorization, particularly in determining the category of closed-class words. This is again an expected result, given our discussion presented in the previous subsection about the weakness of co-occurrence features in categorizing closed-class words.

Comparing results for the conditions shown in Figure 3 reveals that, as in Experiment 1, the use of Morph features does not improve the overall accuracy of categorization. These results are in contrast to the findings of Onnis and Christiansen (2008), who claim that featuring words solely based on their

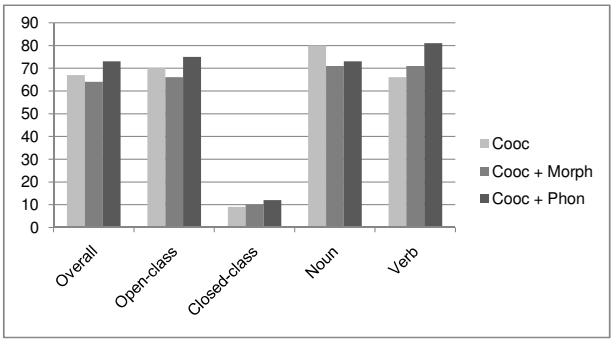


Figure 3: %Accuracy of novel word categorization in three conditions; the total number of clusters constructed during training phase is in the range 258–288.

(beginning and) ending phonemes results in good categorization. Their approach differs from ours in that they perform a batch processing over child-directed utterances, which allows their model to more easily learn the correspondences between a certain category, e.g., verbs, and endings shared by words from this category, such as *-ing* in *finishing*, *playing*, *reading*. Our model has to learn such correspondences incrementally, and hence is prone to making errors when calculating similarity between a word form such as "finishing" (a verb ending in the suffix *-ing*) and one such as "string" (a noun with a similar ending which is not a suffix but part of the word itself). Such errors in early stages may cause the algorithm to form incoherent clusters in later stages.

Figure 3 also includes the performance of our model (in all three conditions) separately shown for Nouns and Verbs. Although the use of Morph features does not help the overall categorization accuracy, it does seem to be particularly helpful in identifying Verbs. Interestingly, using Cooc features alone results in a better detection of novel nouns, whereas for verbs, other types of information (Morph and Phon) are helpful. Hence, even among open-class words, discovering different categories seems to rely on different types of information. This is supported by the observation that, typically, context words such as determiners mark the appearance of nouns; in contrast, verbs particularly share morphological and phonological properties. Related statistical analysis, such as that of (Monaghan et al., 2007; Clark, 2003) suggest such a complementary contribution of different cues; and moreover, some psychological studies implicitly take this into account when designing their experiments on children (Brown, 1957).

Conclusions

We have used an adaptation of a categorization algorithm proposed by Alishahi and Chrupała (2009) to model the acquisition of syntactic categories (e.g., verbs and nouns) in children, and to examine the effect of different types of cues on this task.

Our novel word categorization task provides a suitable

framework to evaluate the helpfulness of word-external (e.g., context) as well as word-internal features (e.g., morphological and phonological properties), independently from the identity of the word being categorized (head word). For example, our results indicate that categorizing closed-class words strongly relies on the head word. Specifically, these classes of words do not share intra-category similarities (neither contextual nor morpho/phonological similarities), and hence cannot be categorized well only by drawing on such properties. In contrast, open-class words can be successfully categorized based on a combination of word-internal and word-external properties, even without considering the head word.

In a more detailed investigation of the roles of word-external versus word-internal features, we find that verbs are better recognized when phonological and morphological properties are taken into account in addition to the context (co-occurring words). Note that we do not assume a full knowledge of morphology, but instead use word endingd as an approximation to word suffixes (as suggested by Onnis & Christiansen, 2008). Interestingly, for nouns, considering only the information about the co-occurring words results in a more accurate categorization. This finding is in contrast to that of Onnis and Christiansen (2008). We argue this difference to be due to the incremental nature of our model.

Evaluating the effect of different cues in word categorization models needs much care. Studies such as those of Parisien et al. (2008) and Alishahi and Chrupała (2009) have reported the capability of co-occurrence information in categorizing words. They include, however, the head word itself as part of their features used for categorization. These studies evaluated the performance of their models on various tasks, such as noun/verb disambiguation, and semantic feature prediction. But they did not provide a comparison between their models and a categorization model that only uses the head word. As shown in our experiments, it is possible to achieve a high accuracy on a task by using such a simple conservative model. The task of novel word categorization that we propose is appropriate for evaluating the ability of a set of categories generated by a model to make generalizations.

In this study, we have shown that different types of cues, e.g., contextual or word-internal properties, provide children with complementary information, each helping with the categorization of a particular group of words. However, our framework is general and can be extended to incorporate other similar features (e.g., other morphological or phonological cues), as well as information about the semantic properties of words.

References

Alishahi, A., & Chrupała, G. (2009). Lexical category acquisition as an incremental process. In *Proceedings of the CogSci-2009 workshop on psychocomputational models of human language acquisition.* Amsterdam.

Brown, R. W. (1957). Linguistic determinism and the part of speech. , *55(1)*.

Cartwright, T. A., & Brent, M. R. (1997). Syntactic categorization in early language acquisition: formalizing the role of distributional analysis. *Cognition*, *63*(2), 121–170.

Chang, F., Lieven, E., & Tomasell, M. (2006). Using child utterances to evaluate syntax acquisition algorithms. In *In Proceedings of the 28th Annual Conference of the Cognitive Science Society.*

Clark, A. (2000). Inducing syntactic categories by context distribution clustering. In *Proceedings of the CoNLL-2000 and LLL-2000* (pp. 91–94).

Clark, A. (2003). Combining distributional and morphological information for part of speech induction. In *Proceedings of the 10th annual meeting of the European Association for Computational Linguistics* (pp. 59–66).

Gelman, S., & Taylor, M. (1984). How two-year-old children interpret proper and common names for unfamiliar objects. *Child Development*, *55*(4), 1535–1540.

Gerken, L., Wilson, R., & Lewis, W. (2005). Infants can use distributional cues to form syntactic categories. *Journal of Child Language*, *32*(2), 249–268.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk* (3rd ed., Vol. 2: The Database). MahWah, NJ: Lawrence Erlbaum Associates.

Mintz, T. H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, *90*(1), 91–117.

Monaghan, P., Christiansen, M. H., & Chater, N. (2007). The phonological-distributional coherence hypothesis: Cross-linguistic evidence in language acquisition. *Cognitive Psychology*, *55*(4), 259–305.

Onnis, L., & Christiansen, M. H. (2008). Lexical categories at the edge of the word. *Cognitive Science*, *32*(1), 184–221.

Parisien, C., Fazly, A., & Stevenson, S. (2008). An incremental bayesian model for learning syntactic categories. In *Proceedings of the CoNLL-2008.*

Pearl, L. (2009). Using computational modeling in language acquisition research. *To appear in Experimental Methods in Language Acquisition Research*.

Redington, M., Chater, N., Finch, S., & Technology, T. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science*, *22*, 425–469.

Samuelson, L. K., & Smith, L. B. (1999). Early noun vocabularies: Do ontology, category structure and syntax correspond? *Cognition*, *73*(1), 1–33.

Schütze, H. (1993). Part of speech induction from scratch. In *Proceedings of the ACL-1993* (pp. 252–258).

Theakston, A. L., Lieven, E. V., Pine, J. M., & Rowland, C. F. (2001). The role of performance limitations in the acquisition of verb–argument structure: An alternative account. *Journal of Child Language*, *28*, 127–152.

Wilson, M. (1988). *MRC Psycholinguistic Database: Machine-usable Dictionary, version 2.00*.

Multiple-choice testing can improve the retention of nontested related information

Jeri L. Little (jerilittle@ucla.edu)
Department of Psychology, 1285 Franz Hall, Box 951563
Los Angeles, CA 90024 USA

Elizabeth Ligon Bjork (elbjork@psych.ucla.edu)
Department of Psychology, 1285 Franz Hall, Box 951563
Los Angeles, CA 90024 USA

Abstract

Taking an initial test leads to improved performance on later tests for those previously tested questions. Whether prior testing improves one's ability to answer related questions, however, is less clear, with some results showing impairment for related information, an effect called retrieval-induced forgetting (RIF; e.g., Anderson, Bjork, & Bjork, 1994). Two experiments investigated the use of initial multiple-choice tests on the retention of previously studied prose passages, specifically on the retention of related, but initially nontested information. In both experiments, an incorrect alternative on the initial test served as the correct answer to a related question on the final test. Results demonstrated that the retention of related information can, indeed, be facilitated by initial multiple-choice tests (Experiment 1) and that this benefit is dependant upon using competitive incorrect alternatives (Experiment 2). We discuss how and why our results differ from previous work (e.g., RIF) and address possible educational applications.

Keywords: memory; testing effects; prose passages; RIF

Introduction

Testing is ubiquitous in education. In most cases, teachers use tests to assess how much a student has learned. Similarly, when students self-test (e.g., with flashcards or practice tests), they typically do so in order to assess their current mastery of the to-be-learned materials. Testing, however, can have other benefits that extend beyond evaluation because retrieval modifies memory so as to improve future recall (see Bjork, 1975).

Multiple-Choice Tests in Educational Contexts

Nowhere is the implementation of testing more widespread than in educational contexts, and in such contexts, the use of multiple-choice (MC) tests is very popular. Some concerns exist regarding their use, however. One concern is that MC tests might provide less opportunity for learning than do cued-recall (e.g., short answer) or free-recall (e.g., essay) tests. Indeed, some studies have shown that although initial MC, cued-recall, and free-recall tests all lead to better retention of the tested information, as compared to nontested information, retention of tested information is better after cued-recall or free-recall tests (e.g., Gay, 1980; Kang, McDermott, & Roediger, 2007; McDaniel, Anderson, Morrisette, & Derbish, 2007).

Perhaps the increased difficulty of answering a recall question correctly (versus a similar MC question) accounts for this difference; that is, retrieval, but not necessarily the recognition and selection of a correct answer, modifies memory (e.g., Bjork, 1975; McDaniel & Masson, 1985). We argue, however, that answering an MC question need not be just a matter of recognizing the correct alternative. In a well-constructed MC test, the test-taker likely recognizes most or all of the alternatives from previous study, but must decide whether or not that alternative is an appropriate answer to the question at hand (Sax & Collet, 1968; Whitten & Leonard, 1980). Often processes of discrimination and memory search are utilized as one thinks not only of which alternative is correct and why, but also of which alternatives are incorrect and why. Certain MC tests could, therefore, invoke a type of processing comparable to that invoked by recall tests (Whitten & Leonard, 1980).

Related Information

Although previous testing is clearly beneficial for retention of identical information, it is less clear whether testing might also benefit the retention of related, but initially nontested information. For example, if one reads a chapter about several U.S. presidents and then answers questions about some of those presidents, will information about the other presidents be strengthened as well? This issue seems particularly germane to the educational context where instructors would rarely ask the same questions on both a quiz and a later exam. In addition, instructors often give practice tests to students, with the intention of providing them with an idea of what the later exam will be like, while not providing them with the actual questions.

On the basis of previous research examining the effects of initial testing on the later recall of related information, one might expect the retention of such information to be impaired. To illustrate, using a retrieval-practice paradigm, Anderson, Bjork, and Bjork (1994) found that—after an initial study phase—testing or giving retrieval practice to some items from a given category improved their later recall, but impaired the recall of other items in that category that were not themselves tested, as compared to the recall of items from another category, none of which were tested—a phenomenon now known as retrieval-induced forgetting. Thus, it seems possible that by giving initial tests or practice

questions, instructors could be inadvertently impairing their students' performance for related nontested questions that may appear on later exams.

Such retrieval-induced forgetting has been demonstrated for educational materials, including facts (Chan, 2009; Macrae & MacLeod, 1999); prose materials (Carroll, Cambell-Ratcliffe, Murnane, & Perfect, 2007); and even one's native language when words from a second language were practiced (Levy, McVeigh, Marful, & Anderson, 2007). Retrieval-induced forgetting has been argued to occur as the consequence of inhibitory processes needed to resolve competition among alternative responses to the same or similar cues (Anderson et al., 1994). Interestingly, it is argued that the processes that lead to forgetting are largely unconscious, as competitive alternatives need not be explicitly brought to mind for them to be suppressed. To the extent that related concepts are brought to mind, however, and can then be used to access the correct answer to a given question, related information might be facilitated.

Indeed, in recent work, Chan, McDermott, and Roediger (2006) developed question pairs such that answering one question on an initial test would encourage the spontaneous recall of information related to the second question that was then to be asked on the later test. Using these question pairs, Chan et al. found facilitation for related, but initially nontested information, although this result likely depended upon specific aspects of the procedure and materials in addition to the facilitative nature of the pairs (i.e., a 24-hr delay between the initial and final tests and integrated encoding of the to-be-learned information). In subsequent research, Chan (2009) demonstrated that although facilitation for these initially nontested, related items occurred at a 24-hr delay when the information had been learned in a prose context, forgetting occurred at a shorter delay when the information had been learned as an unordered series of facts. Importantly, in all of these studies that used short delays to final test, no facilitation was found for related information, even though the time spent on the initial test led to a greater amount of time-on-task—that is, time that the participant spent thinking about information from the tested passage.

In the present research, we tested whether MC tests might afford a benefit for related information that is not as easily afforded by cued-recall tests. Multiple-choice tests differ from free- and cued-recall tests in that they provide students with a set of related (and often competitive) concepts through which they can consciously search in selecting the correct answer, whereas cued-recall tests do not. For example, if given a cued-recall question about who served as the fourth president of the United States, although one may eventually recall the answer (i.e., *Madison*), in the process of doing so, other alternatives (e.g., *Adams*, *Jefferson*) may also become activated by the cue and compete for access and thus need to be suppressed in order to access *Madison*, according to inhibitory accounts of retrieval-induced forgetting. In contrast, if given an MC question with competitive alternatives provided (e.g., *Adams*, *Jefferson*), test-takers may be encouraged to consciously think about such competitors in selecting which president was the fourth (e.g., Adams and Jefferson held office prior to Madison, Jefferson was the third president, etc.), thereby strengthening information they spontaneously recall about these other presidents. Accordingly, MC tests (with competitive alternatives) might both reduce the possibility of retrieval-induced forgetting effects as well as encourage a type of spontaneous recall that later supports the enhanced recall of related, nontested information.

In Experiment 1, we explored this possibility by examining the effects of initial testing of some of the information presented in a prose passage on the later recall of related information using a variation of the retrieval-practice paradigm; specifically, we employed initial MC tests rather than cued-recall tests and then compared the recall of the previously tested items and related nontested items to that of control items from a passage not previously tested. We had two major questions in mind: (a) to what extent would the initial MC tests enhance the recall of previously tested information and (b) would the use of MC questions during initial testing enhance the recall of related information; that is, would the use of MC questions in the initial test allow related items to be facilitated instead of impaired—that is, escape retrieval-induced forgetting? In addition, we utilized a feedback manipulation to see whether being shown the correct answer after attempting to answer a question would affect later recall of both previously tested and related information. Although shown to improve recall of previously tested information, it is uncertain how feedback might affect recall of related information.

Experiment 1

Method

Participants A total of 112 students at the University of California, Los Angeles, participated for credit in an introductory psychology course.

Design We used a 2 (item type: previously tested, previously nontested related) x 2 (feedback: present, absent) within-subjects design plus an independent control group.

Materials Two passages were constructed, one about Saturn and one about Yellowstone National Park, and ten pairs of MC questions were created for each passage. The two questions in each pair were semantically related in that both questions tested the same topic (e.g., geysers) and had the same four alternatives (e.g., *Old Faithful*, *Steamboat Geyser*, *Castle Geyser*, and *Daisy Geyser*), but different correct answers (e.g., *What is the tallest geyser in Yellowstone National Park?* Answer: *Steamboat Geyser;* and, *What is the oldest geyser in Yellowstone National Park?* Answer: *Castle Geyser*). Questions were divided into two 10-item sets for a given passage, with the two questions from each pair randomly assigned to a different set.

Procedure All participants were given 10 min to read the first passage and were instructed to continue studying it if they finished early. Participants in the testing condition were then given an initial 10-item MC test (i.e., all items in one of the question sets for that passage) with questions presented one at a time on the computer. For a given test, all questions were either followed by feedback (feedback present) or not (feedback absent) after the participant provided an answer. Feedback entailed the entire question being re-presented, with the answer printed in red. Following study and test of the first passage, participants followed the same procedure for the second passage except that if feedback had been provided in the first MC test, then it was absent in the second test and vice versa.

Participants in the control condition received no tests; rather, they engaged in a non-verbal filler task (i.e., playing Tetris) following their study of each passage (for the same amount of time as would have been needed to take the test).

Finally, both tested and control participants received a final cued-recall test after a 5-min retention interval during which they played Tetris. Forty questions were presented one at a time on the computer screen; as cued-recall questions, they did not appear with any answer alternatives. For the tested condition, except for the absence of alternatives, half of the questions were identical to the MC questions (i.e., previously tested) and half were the nontested related items (i.e., the remaining questions from the two 10-item sets that had not appeared in the initial MC tests). Related questions were always tested before previously tested questions. For the control condition, all questions were previously nontested and served as a baseline. Topic (Passage) order, question set, and feedback (after Passage 1 or Passage 2) were counterbalanced.

Results and Discussion

Initial MC Test Performance Participants in the tested condition correctly answered an average of 70% (*SD* = 17%) of the questions on the initial MC tests.

Final Test Performance Final test performance is presented in Figure 1. As shown, we found evidence that taking an initial MC test improved the recall of both previously tested and previously nontested related information as compared to the control condition. [1]

Recall performance of participants in the tested condition was compared to the corresponding performance of participants in the nontested control condition via planned independent-samples *t* tests and, importantly, benefits were found for both types of questions. Specifically, these comparisons revealed that (a) previously tested questions given feedback (*M* = 65%, *SE* = 3%) and previously tested questions not given feedback (*M* = 51%, *SE* = 3%) were both answered correctly more often than the control questions (*M* = 29%, *SE* = 2%), $t(110) = 10.88$, $p < .001$, and $t(110) = 6.45$, $p < .001$, respectively; and (b) questions related to previously tested questions that had received feedback (*M* = 40%, *SE* = 3%) and questions related to previously tested questions that had not received feedback (*M* = 43%, *SE* = 3%) were both answered correctly more often than the control questions (*M* = 33%, *SE* = 2%), $t(110) = 2.10$, $p < .05$ and $t(110) = 3.10$, $p < .01$, respectively.

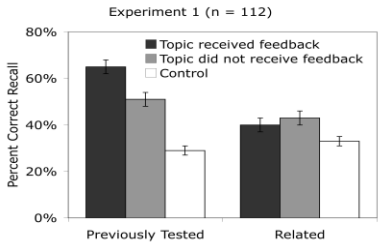


Figure 1: Correct recall percentages as a function of item and feedback type in Experiment 1. White bars show baseline recall for initially non-tested questions by control participants. Error bars represent +/- 1 SE.

To summarize, in Experiment 1, we found a generalized benefit of testing such that the answers to questions on a final cued-recall test were recalled more often when preceded by initial MC tests than when not. Most importantly, this benefit occurred even when the questions on the final cued-recall test were not identical, but only related to those on the initial MC tests, and even though answering such questions correctly on the final test involved recall of an answer that participants had needed to select against during the initial MC test. Thus, providing participants with practice on initial MC questions allowed related information not only to escape impairment but, indeed, to be enhanced. Although retrieval-induced forgetting is largely believed to occur as the result of the unconscious suppression of competitive alternatives, MC tests provide learners with the competitors and thus they can be consciously examined. For example, if students are given a set of alternatives that had all occurred in the required reading, as is the case for a MC question in our

[1] Overall, participants in the nontested control group correctly answered 31% (*SD* = 13%) of the questions on the final test, recalling marginally more answers in the first half of the test (*M* = 33%, *SE* = 2%) than in the second half (*M* = 29%, *SE* = 2%), $F(55) = 3.5$, $p = .07$, a finding consistent with previous accounts of output interference. Because of this marginal difference in performance for the first half and second half of the test, we compared recall for previously tested questions in the tested condition (which were always presented in the second half of the final test) with recall for questions in the control condition that were presented in the second half of the final test. Similarly, recall for nontested questions in the tested condition was compared with recall for the questions in the control condition that were presented in the first half of the test. This method of analysis provides a more conservative test of facilitation for related information.

experiment (e.g., *88 Earth days, 176 Earth days, 10 Earth hours,* and *30 Earth years* for the question: *How long does it take Saturn to revolve around the Sun?*), they could use these alternatives as a guide for searching their relevant knowledge set for the answer (e.g., 88 days and 176 days are wrong as they are related to Mercury; Saturn has a shorter day than Earth). Hence, even if students were unable to recall the answer to this particular question if asked in the format of a cued-recall question, if asked in the format of a MC question with possible alternatives provided, knowledge of related information presented in the passage might be utilized to reject incorrect alternatives; and, in this process, the student may spontaneously answer other related, but nontested questions. Indeed, we believe that such spontaneous retrievals may be the process by which the observed benefit for related but previously nontested items occurred in Experiment 1. For such a beneficial search process to be invoked, however, it would seem necessary that the incorrect choices be potential answers (i.e., competitive alternatives to the correct answer), thus requiring the student to select against them with the use of associated information from the passage. In contrast, without competitive alternatives, perhaps a benefit to related nontested information would not occur because the alternatives would not encourage this type of search strategy. We sought to explore this possibility in Experiment 2 by manipulating the competitiveness of the incorrect alternatives in the initial MC tests that followed the reading of prose passages.

Experiment 2

In Experiment 2, we tested whether the benefit of testing observed for related but previously nontested items in Experiment 1 arose from a type of search strategy engendered by the use of competitive alternatives in the initial MC tests, as described above. To do so, we manipulated the plausibility of the incorrect alternatives, hypothesizing that the more plausible the incorrect alternatives were as answers, the more competitive they would be and the more processing they would require in the attempt to reject them—processing that would likely involve retrieval of associated information from the passage and thus deeper processing of both the correct and the incorrect alternatives. Accordingly, we predicted that initial MC questions using more plausible incorrect alternatives would lead to a greater recall benefit for both previously tested information and previously nontested related information than would initial MC questions using less plausible incorrect alternatives.

Method

Participants A total of 28 students at the University of California, Los Angeles, participated for credit in an introductory psychology course.

Design We used a 2 (item type: previously tested, previously nontested related) X 2 (MC question type: competitive, non-competitive) within-subject design for the testing condition plus a control condition, with all participants serving in both conditions.

Materials Two passages were constructed, one about the Solar System and one about Ferrets, and eight question pairs were created for each passage. Related questions tested information from the same passage and the same type of information served as the correct answer for both questions (e.g., numbers, terms, proper names). To illustrate, the answers to two such questions (*How many inches long is an average ferret tail?* and *How many years ago were ferrets first domesticated, according to mitochondrial DNA evidence?*) were both numbers (i.e., *5* and *2500*, respectively).

To utilize a MC format for each of these questions, four incorrect alternatives were chosen from other information presented in the passage. Two incorrect alternatives were highly related to one question in the pair (and thus, plausible answers for it) and the other two alternatives were highly related to the other question (and thus, plausible answers for it). Thus, for a given pair, there were six alternatives (including the two correct answers). Because all of the alternatives for a given pair had the same type of answers (e.g., numbers), each of the six alternatives could be used in constructing two three-alternative MC questions for each question in these pairs. By manipulating which alternatives were used, we created a competitive and non-competitive version of each question. For example, in competitive versions, the incorrect alternatives were *7-10* and *20* for the first question and *1500* and *3500* for the second question. For the non-competitive versions, the incorrect alternatives were *1500* and *3500* for the first question and *7-10* and *20* for the second question.

Next, two new questions were constructed for each question-pair to serve as the nontested related questions on the final cued-recall test. As in Experiment 1, for these new questions, correct answers were previously incorrect alternatives on the MC questions. For example, although *7-10* was used as an incorrect alternative on the initial test, it was the correct answer to the question, *"How long do ferrets typically live?"* Similarly, *3500* was the correct answer to the question, "*According to archaeological evidence, how long ago were ferrets domesticated?"*

In summary, the six possible alternatives for each of the eight question-pairs were manipulated so as to make both of the three-alternative questions in each pair competitive or non-competitive for a given participant. On the initial MC test, all participants answered eight competitive questions and eight non-competitive questions. The final test included previously nontested questions for which previously incorrect alternatives (either competitive or non-competitive) were now the correct answers.

Procedure All participants were given 10 min to read the first of two passages and were instructed to continue studying the passage if they finished reading early. Next participants either took a test or engaged in a non-verbal filler task. When the passage was tested, participants were given an initial MC test with 16 questions (i.e., eight question pairs) for which they gave verbal responses. Questions appeared one at a time on a computer, and no feedback was given. When the passage served as the nontested control passage, participants played Tetris following its presentation for 3 min (the same time needed to take the test). If given a MC test after the first passage, then that participant engaged in the non-verbal filler task after the second passage and vice versa.

Finally, after a 4-min retention interval during which all participants played Tetris, a final 64-question cued-recall test was given. The 32 questions for the tested topic (previously tested and previously nontested related) and the 32 questions from the nontested control topic were presented on a computer screen, one-at-a-time, and participants gave a verbal response to each. Questions from the previously tested topic were always tested last. Topic (Passage) order, plausibility of alternatives, and testing (after Passage 1 or after Passage 2) were counterbalanced.

Results and Discussion

Initial MC Test Performance On the initial MC test, participants correctly answered more non-competitive questions (M = 86%, SE = 3%) than competitive questions (M = 66%, SE = 3%), $t(27) = 5.67$, $p < .001$, confirming that competitive alternatives make questions more difficult to answer correctly than do non-competitive alternatives.

Final Test Performance Final test performance is presented in Figure 2. For previously tested questions, correct answers to competitive questions (M = 37%, SE = 3%) were recalled marginally less often than were correct answers to non-competitive questions (M = 45%, SE = 4%), $t(27) = 1.76$, $p < .10$, a pattern consistent with the initial MC performance. For previously nontested questions from the same topic, however, the effect was in the opposite direction: correct answers that had previously been incorrect competitive alternatives (M = 47%, SE = 5%) were recalled more often than were correct answers that had previously been incorrect non-competitive alternatives (M = 36%, SE = 4%), $t(27) = 2.55$, $p < .05$, confirming our prediction that initial MC questions with competitive alternatives would lead to enhanced recall of related information as compared to initial MC questions with non-competitive alternatives.

When compared to control items (M = 27%, SE = 3%), answers to both previously tested competitive questions (M = 37%, SE = 3%) and previously tested non-competitive questions (M = 45%, SE = 4%) were facilitated, $t(27) = 3.10$, $p < .01$ and $t(27) = 4.54$, $p < .001$, respectively, demonstrating a testing effect. For previously nontested questions from the tested topic, those with answers that had previously been incorrect competitive alternatives (M = 47%, SE = 5%) were correctly answered more often than questions from the control passage (M = 36%, SE = 4%), $t(27) = 2.21$, $p < .05$, whereas those with answers that had been incorrect non-competitive alternatives (M = 36%, SE = 4%) were not, $t(27) = 0.1$, $p > .05$

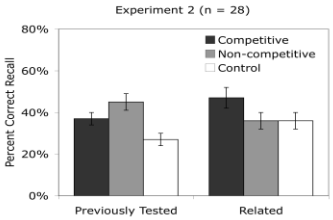


Figure 2: Correct recall percentages as a function of item type and competitiveness of MC alternatives on the initial MC test in Experiment 2. Error bars represent +/- 1 SE.

In Experiment 2, we manipulated the competitiveness of a given question by choosing incorrect alternatives that were either plausible or implausible answer choices to examine whether competitiveness of the alternatives was a critical factor in the facilitation of related information, and our results suggest this to be the case. Of concern, however, is whether the benefit we observed resulted from the increased processing of the incorrect alternatives as hypothesized, or simply occurred as an artifact of initial test performance. Because competitive questions were more difficult to answer than non-competitive ones, perhaps the benefit observed was merely a consequence of the participant being more likely—on the initial MC test—to select an incorrect competitive alternative than to select an incorrect non-competitive alternative, and then to recall that previously incorrect answer on the final test when given the related question (for which the answer might now be correct). For example, on the initial MC test, a participant might incorrectly choose *7-10* (instead of *5*) when given the question, "*How many inches long is an average ferret tail?"* If the participant then gives *7-10* as the correct answer for the question, "*How long do ferrets typically live?"* on the later test, one cannot be sure whether that participant is giving that answer believing it to be correct or giving that answer because it was chosen before and is now primed as an answer for all questions where it is plausible.

If such generalized strengthening of alternatives is the mechanism that leads to this effect, then participants should not demonstrate the pattern of results previously shown for related questions when recall is conditionalized upon answering the corresponding MC question correctly. A conditional analysis demonstrated, however, that marginally more answers to related questions were recalled correctly when those answers were previously incorrect competitive alternatives (M = 50%, SE = 4%) than when they were previously incorrect non-competitive alternatives (M = 41%, SE = 4%), $t(27) = 1.91$, $p = .07$. Thus, the possibility that

this effect is driven by cases in which a participant chooses the incorrect answer and then carries it to a new question (where it then happens to be correct) seems unwarranted.

General Discussion

The present results imply that taking an initial MC test not only improves one's ability to recall *that* information on a later cued-recall test, but also improve one's ability to recall nontested, but related information on a later test—provided that the initial test utilizes incorrect alternatives that are competitive. Furthermore, although an MC question is often easier to answer than a comparable question in a cued-recall format (i.e., same question, without the choices), to the extent that the question has competitive alternatives, that question may invoke processes that are similar to those involved in recall (e.g., memory search, retrieval checks), thus leading to comparable benefits to the tested information. Moreover, use of MC questions may provide a way to insure that access to related nontested information is not impaired on a later test.

Educators may be concerned that the initial test provides participants with additional time to think about the tested topic, whereas no such additional time is allocated to the nontested control topic. Although a valid concern, our findings need to be viewed in the context of previous work using the retrieval-practice paradigm in which additional time is not allotted for nontested control materials and in which nontested information from a tested topic is rarely facilitated and, in fact, is typically impaired (e.g., Macrae & MacLeod, 1999; Carroll et al., 2007). Indeed, with a similar procedure and educational prose materials, but with an initial cued-recall test, Chan (2009) did not find facilitation for related information, even when the questions on the initial test were created to be facilitative for questions on the final test. One might thus argue that our finding of facilitation occurred because our MC questions—unlike cued-recall questions—exposed participants to the future answers for related questions (in the form of incorrect alternatives), thus providing shallow priming that leads to facilitation on the later test. Against such an argument, however, are the findings of Experiment 2 where alternatives were exposed in both competitive and non-competitive conditions and yet facilitation only occurred when alternatives were competitive, thus ruling out an explanation in terms of priming. Instead, our findings are consistent with the explanation that competitive MC questions lead to enhanced performance for nontested related information, owing to the deeper processing of the incorrect alternatives that they engender, as compared to processing engendered by noncompetitive MC questions.

We believe that the present results have implications for both instructors and students. Instructors can create quizzes and study guides that improve retention for both initially tested information as well as related information that is not itself tested. Students can benefit from tests by thinking about all of the alternatives—not only why a given answer is correct, but also why other answer choices are wrong.

Acknowledgments

We thank Ashley Kees for her valuable contributions in both the conception and implementation of this work and Robert Bjork for his valuable insights. Grant 29192G from the McDonnell Foundation supported this research.

References

Anderson, M. C., Bjork, R. A., & Bjork, E. L. (1994). Remembering can cause forgetting: Retrieval dynamics in long-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 1063-1087.

Bjork, R. A. (1975). Retrieval as a memory modifier. In R. Solso (Ed.), *Information processing and cognition: The Loyola Symposium* Hillsdale, NJ: Lawrence Erlbaum Associates.

Carroll, M., Campbell-Ratcliffe, J., Murnane, H., & Perfect, T. (2007). Retrieval-induced forgetting in educational contexts: Monitoring, expertise, text integration, and test format. *European Journal of Cognitive Psychology*, *19*, 580-606.

Chan, J. C., McDermott, K. B., & Roediger, H. L. (2006). Retrieval-induced facilitation: Initially nontested material can benefit from prior testing of related material. *Journal of Experimental Psychology: General*, *135*, 553-571.

Gay, L. R. (1980). The comparable effects of MC versus short-answer tests on retention. *Journal of Educational Measurement*, *17*, 45-50.

Kang, S. H. K., McDermott, K. B., Roediger, H. L. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology*, *19*, 528-558.

Levy, B. J., McVeigh, N. D., Marful, A., & Anderson, M. C. (2007). Inhibiting your native language: The role of retrieval-induced forgetting during second language acquisition. *Psychological Science*, *18*, 19-34.

Macrae, C. N., & MacLeod, M. D. (1999). On recollections lost: When practice makes imperfect. *Journal of Personality and Social Psychology*, *77*, 463-473.

McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology*, *19*, 494-513.

McDaniel, M. A., & Masson, M. E. J. (1985). Altering memory representations through retrieval. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *11*, 371-385.

Sax, G., & Collet, L. S. (1968). An empirical comparison of the effects of recall and MC tests on student achievement. *Journal of Educational Measurement, 5*, 169-173.

Whitten, W. B., & Leonard, J. M. (1980). Learning from tests: Facilitation of delayed recall by initial recognition alternatives. *Journal of Experimental Psychology: Human Learning and Memory*, *6*, 127-134.

Holographic stimulus representation and judgement of grammaticality in an exemplar model: Combining item and serial-order information

Randall K. Jamieson (randy_jamieson@umanitoba.ca)
Department of Psychology, University of Manitoba
Winnipeg, MB, R3T 2N2, Canada

D. J. K. Mewhort (mewhortd@queensu.ca)
Department of Psychology, Queen's University
Kingston, ON, K7L 3N6, Canada

Abstract

We examine representation assumptions for learning in the artificial grammar task. Strings of letters can be represented by first building vectors to represent individual letters and then concatenating the letter vectors into a vector of larger dimensionality. Although such a representation works well in selected examples of artificial-grammar learning, it fails in examples that depend on left-to-right serial information. We show that recursive convolution solves the problem by combining item and serial-order information in a stimulus item into a distributed data structure. We import the representations into an established model of human memory. The new scheme succeeds not only in applications that were successful using concatenation but also in applications that depend on left-to-right serial organization.

Keywords: Artificial grammar learning; Holographic representation; Exemplar model

Introduction

In an artificial-grammar learning (AGL) classification task, participants study strings of symbols. Following study, the participants are told that the studied items were constructed according to the rules of an artificial grammar and are invited to sort novel rule-based (grammatical) exemplars from novel rule-violating (ungrammatical) ones. Even though the participants are unable to describe the rules, they can discriminate the two classes of stimuli.

Initial accounts proposed that the participants abstracted the grammar and used that knowledge to judge the status of the exemplars (e.g., Reber, 1967, 1993). Later investigators argued that the participants judged grammaticality without reference to the grammar. To support the latter position, investigators identified several sources of information that discriminate the two classes of test strings. Brooks (1978) suggested that whole-item similarity between training and test strings is used to infer grammaticality. Perruchet and Pacteau (1990) argued that bigram overlap is used to infer grammaticality. Vokey and Brooks (1992) identified edit distance as a predictor, and Brooks and Vokey (1991) argued that patterns of repetition within a string are used to infer grammaticality. Knowlton and Squire (1996) identified associative chunk strength (ACS), and Johnstone and Shanks (1999) identified chunk novelty. Finally, Jamieson and Mewhort (2009a, 2010) showed that global similarity predicts performance in the task. Regression analyses designed to sort the various predictors have confirmed a role for all of them (e.g., Johnstone & Shanks, 1999). Factorial designs that have pitted predictors against one another have been unable to identify a single dominant predictor (e.g., Kinder & Lotz, 2009; Vokey & Brooks, 1992).

We think that many of the predictors (e.g., ACS, bigram over, etc) point to a common underlying factor, namely left-to-right serial structure. If so, the problem is not to determine which predictor dominates but, rather, to decide how subjects encode material so that they have access to the left-to-right serial structure in the exemplars.

In this paper, we explore an encoding mechanism that folds several orders of left-to-right serial structure in a string into a coherent and distributed data structure (i.e., single letters, bi-grams, trigrams, and whole strings). To begin,we describe the representation scheme. After, we show that the new representations predict judgement of grammaticality when used in an established theory of retrieval (Jamieson & Mewhort, 2009a, 2010).

Holographic representation in memory

Many investigators have proposed that light holography provides a mathematical basis for memory representation (Borsellino & Poggio, 1973; Gabor, 1968; Khan, 1998; Longuet-Higgins, 1968; Poggio, 1973). Murdock's (1982, 1983, 1997) TODAM is probably the best-known use of the idea in experimental psychology. In TODAM, stimulus associations are formed using linear convolution and associations are unpacked using correlation (deconvolution).

More recently, Jones and Mewhort (2007) used recursive circular convolution (Plate, 1995) to develop a self-organizing model of semantic memory (BEAGLE). BEAGLE captures judgements of semantic typicality, categorization, priming, and syntax from word order. BEAGLE's ability to handle so many phenomena of semantic memory is in itself impressive. However, from our perspective, BEAGLE's strength is that it shows how holographic representation can account for complex decision behaviour without adding control structures (e.g., learning and the application of rules). BEAGLE's success suggests that holographic stimulus representation should be explored in related models of learning and memory. The present work adapts BEAGLE's representation scheme to represent strings in the artificial grammar classification task.

Recursive circular convolution

Circular convolution is a mathematical operation that forms an associative representation, **z**, for two input vectors, **x** and **y**,

$$z_i = \sum_{j=0}^{n-1} x_{j \bmod n} \times y_{(i-j) \bmod n} \quad , \qquad [1]$$

where *i* indexes the element in **z** and where *n* is the dimensionality of **z**, **x**, and **y**. Briefly, circular convolution forms the outer-product matrix—long used to represent associations in neural networks (e.g., Anderson, 1995)—and then collapses it into a vector (see Figure 1). Circular convolution is associative, commutative, and distributes over addition.

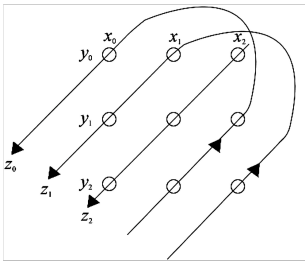


Figure 1. Collapsing an outer-product matrix with circular convolution, where **x** and **y** are the argument vectors and **z** represents the resulting compressed vector from collapsing the outer-product matrix. The values *i* and *j* represent the row and column indices, respectively, for an element in the outer-product matrix.

In the work that follows, we apply circular convolution recursively to encode a series, such as a sequence of letters. Consider the string *ABCD*. To represent *ABCD* as a series, first, generate a unique random vector for each of the individual letters in the string {**a**, **b**, **c**, **d**}. Next, apply circular convolution in a recursive fashion to bind the first letter to the second, the product of that binding to the third, and so on, until each letter has been folded into the representation. At this point, using **z** to represent the string *ABCD*, **z** = *ABCD* = (**a** * **b**) * **c**) * **d**, where * denotes circular convolution. No matter the length of the string, **z** has the same dimensionality as the input (i.e., letter) vectors.

Why holographic representation?

In previous studies of AGL, we represented exemplars by concatenating letters. For example, a string *ABCD* was represented by concatenating the vectors for *A, B, C*, and *D* to form a single vector **a**//**b**//**c**//**d**, where // denotes concatenation. The scheme captured a swath of data from the artificial grammar task and from serial reaction time tasks (see Jamieson & Mewhort, 2009a, 2009b, 2010). Nevertheless, concatenated representations are problematic.

In models using vector representation, it is traditional to compute the similarity between **x** and **y**, using a vector cosine. Thus, with concatenated strings, similarity is computed by comparing information in corresponding serial positions of two strings (i.e., element-for-element). Because of the serial-position constraint, a model using the concatenated representation scheme treats the strings *ABCD* and *CDAB* as if they shared no overlapping features—a judgement that is at odds with data. In contrast, a holographic representation scheme distributes information throughout the vector so that each part of it contains some information about the whole. Thus, in difference to the concatenation scheme, the cosine calculation compares all parts of **x** (i.e., *ABCD*) and **y** (i.e., *CDAB*) simultaneously and, thereby, acknowledges similarity between *ABCD* and *CDAB*. Because participants appreciate the similarity between *ABCD* and *CDAB*, the holographic scheme is preferred.

Critically, holographic stimulus representation finesses the problem of encoding serial structure. Importantly, it does so without requiring a change in the similarity calculation or other aspects of retrieval. This occurs because a representation of *ABCD* that is formed using recursive circular convolution superimposes overlapping orders of serial structure onto a single distributed structure. Because different orders of serial information about a string are superimposed in a single representation, a standard cosine of two vectors supports parallel comparison of multiple orders of serial structure. The question we pose, then, is if we import the holographic representations into an established model of retrieval, will the previously successful model still work; that is, can we still explain peoples' judgements in the artificial grammar task?

Minerva 2

Minerva 2 is an established model of retrieval (Hintzman, 1986, 1988). When a participant studies an item, an event is encoded to memory as a unique trace.

In Minerva 2, a stimulus is represented by a vector of *n* elements; each element takes values: +1 or -1. To represent stimuli in the artificial grammar task, we first, generate a unique random vector for each of the letters in the English alphabet and then apply recursive circular convolution to those vectors to represent a string of letters. Thus, a string *TXXV* is represented by a trace: ((**t** * **x**) * **x**) * **v**.

Memory is a matrix, **M**. Encoding an event involves copying its corresponding vector representation to a new row in the memory matrix. Encoding is sometimes imperfect. Imperfect encoding is implemented by setting some vector elements to zero (indicating that the element is indeterminate or unknown). A parameter, *L*, controls the probability with which an element is stored. As *L* increases, encoding quality improves.

All retrieval is cued. When a retrieval cue is presented, it activates each trace in memory in proportion to its similarity to the cue. The activated traces are aggregated into a

composite called the *echo*; the contribution of each trace to the echo is based on its activation.

The similarity of trace, i, to the probe, P, is computed using a vector cosine, i.e.,

$$S_i = \frac{\sum_{j=1}^{n} P_j \times M_{ij}}{\sqrt{\sum_{j=1}^{n} P_j^2} \sqrt{\sum_{j=1}^{n} M_{ij}^2}}, \qquad [2]$$

where P_j is the value of the jth feature in the probe, M_{ij} is the value of jth feature of the ith row in memory. Like the Pearson r, the similarity of the ith item to the probe, S_i, is scaled to the interval [-1, +1]. Similarity equals +1 when the trace is identical to the probe, 0 when the trace is orthogonal to the probe, and -1 when the trace is opposite to the probe.

The ith trace's activation, A_i, is the cube of its similarity to the probe, i.e.,

$$A_i = S_i^3. \qquad [3]$$

The activation function exaggerates differences in similarity between a probe and items in memory by attenuating activation of exemplars that are not highly similar to the probe. This allows traces most similar to the probe to dominate the information retrieved. Note that the exponent in the activation function preserves the sign of the argument, S_i.

The information retrieved by a probe is a vector, **c**, called the echo. The echo is computed by weighting each of the i = 1 ... m traces in memory by their activations and, then, summing all m traces into a single vector,

$$c_j = \sum_{i=1}^{m} A_i \times M_{ij}. \qquad [4]$$

The information in the echo is converted to decision variable called echo intensity, I, by computing the cosine similarity (see Equation 2) of the echo and probe. In the context of the artificial grammar task (i.e., classification), echo intensity is a proxy for judgement of grammaticality.

In the remainder of this paper we apply the model to data from the judgment of grammaticality task.

Evaluating the model

The judgement of grammaticality task was introduced by Reber (1967). In his experiment, participants memorized grammatical exemplars. After, they judged the grammatical status of novel test probes. Reber's subjects discriminated novel grammatical from novel ungrammatical test probes, but they could not articulate the rules of the grammar.

We have shown previously that Minerva 2 captures discrimination of grammatical from ungrammatical test probes in Reber's (1967) task, without reference to grammatical rules (Jamieson & Mewhort, 2009a, 2010). But we used concatenated stimulus vectors in that work. In the simulations that follow, we retest the model using the holographic rather than concatenated stimulus vectors.

To simulate Reber's (1967) task we began by representing his stimuli in our model.[1] First, we constructed a unique 100-element vector to represent each letter used to construct letter strings: {*T*, *V*, *P*, *X*, *S*}. Second, we generated a vector for each training and test string using recursive circular convolution. Third, we filled successive rows on the memory matrix with the training vectors. Fourth, we introduced moderate data-degradation to the items in memory, i.e., L = 0.7. Finally, we calculated the mean echo intensity for each of the 24 grammatical and 24 ungrammatical test strings.

The new model successfully discriminated grammatical from ungrammatical test items. The mean echo intensity for the 24 grammatical test strings was .57 (SD = .03); the corresponding value for the 24 ungrammatical test strings was .49 (SD = .02), $t(48) = 2.15$, $p < .05$.

In other simulations, we varied the integrity of data in memory (e.g., Jamieson, Holmes, & Mewhort, in press). As shown in Figure 2, the magnitude of the difference in mean echo intensity for grammatical and ungrammatical test strings (i.e., the model's discrimination of grammatical and ungrammatical items) grew as a function of L.

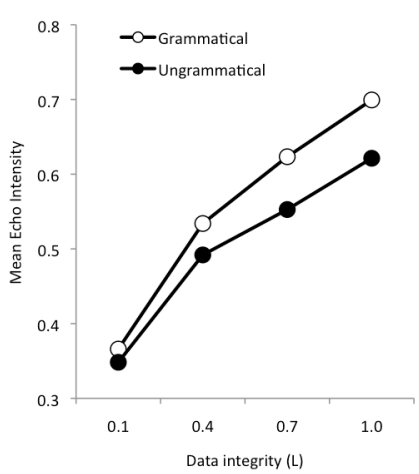


Figure 2. Mean echo intensity for grammatical and ungrammatical test strings as a function of data integrity in memory, L.

The simulation illustrates several points. First, it shows that the distributed stimulus representations generated using recursive circular convolution support discrimination of grammatical from ungrammatical test items. Second, because the model discriminated the two classes of stimuli without reference to grammatical rules, the simulation serves as an existence proof that grammatical strings can be

[1] Reber did not list the specific study and test items that he used in his original paper. He did, however, provide a list of representative strings from the same grammar in another source (Reber, 1993, p. 36). We took our strings from there.

discriminated from ungrammatical test strings without knowledge of the grammatical rules. Thirdly, the simulation shows that we can import holographic stimulus representations into Minerva 2 without a deleterious impact on the effects that the model captures using concatenated vectors (see Jamieson & Mewhort, 2009a, 2010).

Next, we test the new representation scheme by applying it to data collected by Kinder and Lotz (2009). Their data provide a more detailed challenge.

Kinder and Lotz (2009)

Kinder and Lotz (2009) engineered an artificial grammar to distinguish stimulus properties thought to predict judgements of grammaticality. They used the grammar to construct a list of 12 training items and 48 test items. The test items were of four different types. Type 1 and Type 2 items were ungrammatical; Type 3 and Type 4 items were grammatical. Type 1 test items violated both positional and sequential rules of the grammar; Type 2 items violated only sequential rules (i.e., the strings included at least one illegal bigram but all letters were in legal serial positions). Type 3 and Type 4 items obeyed positional and sequential rules of the grammar; but, Type 4 items had the additional property of being very similar to a specific training exemplar. Accordingly, if participants endorse Type 2 over Type 1 items, they must appreciate the positional dependencies of letters in the training set. If participants endorse Type 3 over Type 2 items, they must appreciate the difference between studied and unstudied chunks (i.e., bigrams and trigrams). If they endorse Type 4 over Type 3 items, they must appreciate whole-item similarity between training and test strings.

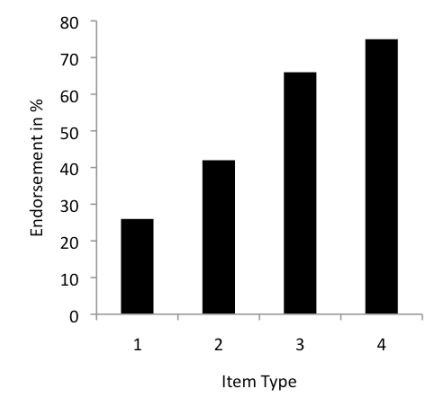


Figure 3. Empirical: Percentage of items endorsed as grammatical in Kinder and Lotz's (2009) Experiment 2.

Kinder and Lotz's (2009) results are reproduced in Figure 3. First, participants endorsed Type 2 over Type 1 items indicating they were sensitive to the positions of individual letters in the training strings. Second, participants' endorsed Type 3 over Type 2 items, indicating they were sensitive to test strings' inclusion/exclusion of studied and unstudied bigrams. Finally, participants endorsed Type 4 over Type 3 items, indicating they were sensitive to whole-item similarity between training and test strings.

The pattern of results demonstrates that judgement of grammaticality is influenced concurrently by the positions of single letters in a string, by knowledge of small chunks (i.e., knowledge of bigrams and trigrams), and by knowledge of larger chunks (i.e., whole training strings). To claim a model as a competent account of decision in the judgement of grammaticality task, the model must accommodate concurrent sensitivity to the three sources of information.

Simulation of Kinder and Lotz (2009; Exp 2)

Kinder and Lotz's (2009) data provide a principled challenge to test the idea that holographic stimulus representation allows multiple orders of serial-structure to exert a concurrent influence on judgements of grammaticality. Hence, we tested our model using Kinder and Lotz's (2009) materials.[2] The simulation was otherwise the same as before.

The results of the simulation are presented in Figure 4; the means were computed across 50 independent replications of the procedure. We treat mean echo intensity as a proxy for mean judgement of grammaticality.

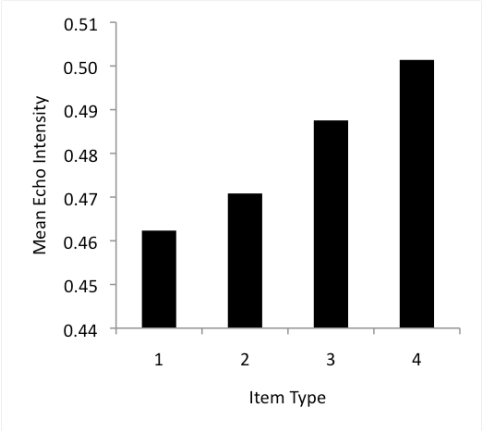


Figure 4. Simulation: Mean echo intensity for the four item types in Kinder and Lotz's (2009) Experiment 2.

As shown, the model reproduced the pattern of results from Kinder and Lotz's (2009) experiment. Firstly, mean echo intensity for Type 2 items was greater than for Type 1 items indicating that the model was sensitive to positional dependencies of individual letters in the training strings. Secondly, echo intensity for Type 3 items was greater than for Type 2 items indicating that the model was sensitive to bigram and trigram structure in the stimuli. Finally, echo intensity for Type 4 items was greater than for the Type 3

[2] A complete listing of Kinder and Lotz's (2009) materials is presented in their Appendix B. The simulations were identical for the two sets; a testament to their care at stimulus design.

items indicating that the model was sensitive to larger chunks of letters, possibly whole strings.

Importing a scheme for holographic stimulus representation into a Minerva 2-based account of retrieval allows the model to capture additional details of performance in the artificial grammar task. Minerva 2 now captures trends that previously required a very different kind of computational account (e.g., the SRN, see Elman, 1990).

General Discussion

Judgements of grammaticality reflect a concurrent consideration of discriminative cues (e.g., Johnstone & Shanks, 1999). To accommodate that fact, we developed a new kind of stimulus representation based on recursive circular convolution. The new representation folds information about several cues into a distributed data structure. More importantly, the holographic representation scheme supports parallel comparison of features in a string, unconstrained by serial position alignment. Using the holographic representations in a model of human memory captures judgement of grammaticality.

In previous work, Jamieson and Mewhort (2009a, 2010) demonstrated that judgment of grammaticality can be understood using Minerva 2—an exemplar model of memory. In that work, exemplars were represented by concatenating individual letter vectors. Judgement of grammaticality reflected a test probe's global similarity to the studied exemplars. The representation scheme worked because it preserved the spatial structure of the stimulus (i.e., letters from left-to-right). However, the account neglected to include information about left-right sequential properties of the exemplars—information that subjects notice during study. Because the model did not acknowledge sequential structure in stimuli, it incorrectly computed similarity between two exemplars based on bigram overlap; a factor measured by associative chunk strength.

The holographic stimulus representations developed here finesse the problem associated with the earlier scheme by folding information about serial-structure into the representation of a string. By using the holographic representations, the model now captures judgements that reflect serial structure (e.g., participants' appreciation of chunk overlap). Despite changes to the representation scheme in the model, we have not changed the model's account of retrieval and so we retain our previous conclusion: Judgement of grammaticality can be captured without an implicit rule-induction process that abstracts and applies grammatical information.

Kinder (2000; Kinder & Lotz, 2009) and others (e.g., Cleeremans, Servan-Schreiber, & McClelland, 1989) have argued for a Simple Recurrent Network (SRN) account of artificial grammar learning. The SRN accomplishes judgement of grammaticality by learning the sequential structure in a set of training sequences and then applying that knowledge to predict sequential regularities in test items. When the SRN can predict the sequential structure of a test string, it judges the test string as grammatical (see Reber, 2002, for an analysis of the approach; see Vokey & Higham, 2004, for model comparison of the SRN and a related instance-based model). Cleeremans et al. (1989) showed the SRN develops a veridical representation of the grammar used to generate the training strings. By contrast, our account treats judgement of grammaticality as an episodic memory task. At study, the model encodes information about individual exemplars, including serial structure. At test, the model judges a test strings' grammaticality by its global similarity to the exemplars in memory. The two classes of model (Minerva 2 and the SRN) offer very different explanations of the cognitive processes that underlie judgement of grammaticality. So, which approach is to be preferred? We think the answer should be based on the nature of the experimental problem.

In the training phase of a standard artificial grammar experiment, participants are asked to memorize exemplars. At test, they are given the problem of inferring the grammaticality of test probes. Of course, people *can* learn sequential structure in stimuli instructions. But they do not have to learn it: the task does not cue them to do so. In our view, although learning sequential structure in a set of exemplars provides a possible mechanism, for the judgement of grammaticality task, it implies compulsory learning of sequential regularities even though that action is neither implied by nor cued by the task. Unlike the SRN, Minerva 2 assumes people notice sequential characteristics of each exemplar, but they do not learn the regularities in the set of exemplars. Moreover, because our account treats judgement of grammaticality as a retrospective judgement, it is not necessary to justify or to describe prospective abstraction of structure in the training set.

In developing our holographic representation scheme, we have been careful to avoid altering our model's assumptions about retrieval. In both our original and our present accounts, we assumed a perceptual system loads memory with what the subjects notice about each of the studied exemplars. Judgment of grammaticality reflects the global similarity of a test probe to training items. The difference in our new account is that the new model assumes that subjects notice more about the order of the symbols than the old model assumed; a claim echoed in post-experimental interviews with our subjects. At a broader level, our solution honours an insight from Simon's (1969) parable of the ant. Simon noted that an ant's path on a beach may be complex and difficult to describe. But, the complexity of the path may be driven by complexity in the beach rather than complexity in the ant. Simon used the parable to goad theorists into considering explanations for a behaviour based on the complexity of the environment before assuming that the behaviour reflects complex psychological mechanisms. Here, we have followed Simon's advice. Peoples' behaviour in the artificial grammar task appears complex and difficult to describe. However, the complexity is in the materials, not in the subjects. Judgement of grammaticality reflects the storage and retrieval of studied exemplars.

Acknowledgments

R. K. Jamieson, Psychology, University of Manitoba; D. J. K. Mewhort, Psychology, Queen's University. The research was supported by grants to both authors from the Natural Sciences and Engineering Research Council of Canada. Mail correspondence to R. K. Jamieson, Department of Psychology, University of Manitoba, Winnipeg, MB, Canada, R3T 2N2. Electronic mail to: randy_jamieson@umanitoba.ca.

References

Anderson, J. A. (1995). *An introduction to neural networks*. Cambridge, MA: MIT Press.

Borsellino, A., & Poggio, T. (1973). Convolution and correlation algebra. *Kybernetik*, *122*, 113-122.

Brooks, L. R., & Vokey, J. R. (1991). Abstract analogies and abstracted grammars: Comments on Reber (1989) and Mathews et al. (1989). *Journal of Experimental Psychology: General*, *120*, 316-320.

Cleeremans, A., Servan-Schreiber, D., & McClelland, J. L. (1989). Finite state automata and simple recurrent networks. *Neural Computation*, *1*, 372-381.

Dienes, Z. (1993). Computational models of implicit learning. In D. C. Berry & Z. Dienes (Eds.), *Implicit learning: Theoretical and empirical issues* (pp. 81-112). Hove, UK: Lawrence Erlbaum Associates.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*, 179-211.

Gabor, D. (1968). Improved holographic model of temporal recall. *Nature*, *217*, 1288-1289.

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review*, *93*, 411-428.

Hintzman, D. L. (1988). Judgements of frequency and recognition memory in a multiple-trace memory model. *Psychological Review*, *95*, 528-551.

Jamieson, R. K., Holmes, S., & Mewhort, D. J. K. (2010). Global similarity predicts dissociation of classification and recognition: Evidence questioning the implicit/explicit learning distinction in amnesia. *Journal of Experimental Psychology: Learning Memory and Cognition*, in press.

Jamieson, R. K., & Mewhort, D. J. K. (2010). Applying an exemplar model to the artificial grammar task: String-completion and performance on individual items. *Quarterly Journal of Experimental Psychology*, *63*, 1014-1039 .

Jamieson, R. K., & Mewhort, D. J. K. (2009a). Applying an exemplar model to the artificial-grammar task: Inferring grammaticality from similarity. *Quarterly Journal of Experimental Psychology*, *62*, 550-575.

Jamieson, R. J., & Mewhort, D. J. K. (2009b) Applying an exemplar model to the serial reaction-time task: Anticipating from experience. *Quarterly Journal of Experimental Psychology*, *62*, 1757-1783.

Johnstone, T., & Shanks, D. R. (1999). Two mechanisms in implicit artificial grammar learning? Comment on Meulemans and van der Linden (1997). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 524-531.

Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information in a holographic lexicon. *Psychological Review*, *114*, 1-37.

Khan, J. I. (1998). Characteristics of multidimensional holographic associative memory in retrieval with dynamically localizable attention. *IEEE Transactions on Neural Networks*, *9*, 389-406.

Kinder, A. (2000). The knowledge acquired during artificial grammar learning: Testing the predictions of the two connectionist models. *Psychological Research*, *63*, 95-105.

Kinder, A., & Lotz, A. (2009). Connectionist models of artificial grammar learning: What type of knowledge is acquired? *Psychological Research*, *73*, 659-673.

Knowlton, B. J. & Squire, L. R. (1996). Artificial grammar learning depends on implicit acquisition of both abstract and exemplar-specific information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22,* 169-181.

Longuet-Higgins, H. C. (1968). Holographic model of temporal recall. *Nature*, *217*, 104.

Murdock, B. B. (1982). A theory for the storage and retrieval of item and associative information. *Psychological Review*, *89*, 609-626.

Murdock, B. B. (1983). A distributed model for serial-order information. *Psychological Review*, *90*, 316-338.

Murdock, B. B. (1997). Context and mediators in a theory of distributed associative memory (TODAM2). *Psychological Review*, *104*, 839-862.

Plate, T. A. (1995). Holographic reduced representations. *IEEE Transactions on Neural Networks*, *6*, 623-641.

Perruchet, P., & Pacteau, C. (1990). Synthetic grammar learning: Implicit rule abstraction or explicit fragmentary knowledge? *Journal of Experimental Psychology: General*, *119*, 264-275.

Poggio, T. (1973). On holographic models of memory. *Kybernetik*, *12*, 237-238.

Reber, A. S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning & Verbal Behavior*, *6*, 855-863.

Reber, P. J. (2002). Attempting to model dissociations of memory. Trends in Cognitive Sciences, 6, 192-194.

Reber, A. S. (1993). *Implicit learning and tacit knowledge: An essay on the cognitive unconscious*. New York: Oxford University Press.

Simon, H. A. (1969). The psychology of thinking: Embedding artifice in nature (Chapter 2). In *Sciences of the artificial (pp. 23-26)*. MIT Press.

Vokey, J. R., & Brooks, L. R. (1992). Salience of item knowledge in learning artificial grammars. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 1504-1510.

Vokey, J. R., & Higham, P. A. (2004). Opposition logic and neural network models in artificial grammar learning. *Consciousness and Cognition*, *13*, 565-578.

A Cross-Cultural Study of Change Blindness in Turkish and American Students

Treysi Terziyan (Treysi.Terziyan@FandM.Edu)
Franklin & Marshall College, 415 Harrisburg Ave.
Lancaster, PA 17603 USA

Joan Gilkey (JoanGilkey@Alumni.FandM.Edu)
Franklin & Marshall College, 415 Harrisburg Ave.
Lancaster, PA 17603 USA

Abstract

Change blindness is a phenomenon that occurs when a person fails to notice changes in their perceptual field. Previous studies have shown that East Asians are sensitive to both contextual and focal changes while Americans are sensitive to focal but not contextual changes (Masuda & Nisbett, 2006). This difference was attributed to the fact that Americans have analytical and East Asians have holistic perceptions. This study questions whether Turkish students' attention to changes in pictures is more like Americans or East Asians. Half of the study was conducted in Turkey and the other half in America. Participants looked at photographs that flickered back and forth from an original picture and an edited photograph. The photographs were Turkish, American, or Neutral. Half were complex, half were simple, and half the changes were made in the foreground and half in the background. We found that both Turkish and American students found the foreground changes a lot faster than the background changes. These results suggested that Turkish people's perception is analytical like Americans'.

Keywords: Change blindness; cross-cultural research

Introduction

One reason that there are continuity errors in movies, which go mostly unnoticed by the audience, is that resource limits prevent us from attending to every element of a visual scene. In one famous study (Simons & Chabris, 1999), 192 participants were shown a video of 6 people in two teams passing a ball. The participants were asked to count and report the number of passes occurring between players of the same team. While they were doing this task, one of two unexpected things happened in the video: either a woman with an umbrella or a woman in a gorilla suit walked by. Overall, only 54% of the participants reported seeing the unexpected event. This means that 46% of them did not "see" a very odd event, immediately obvious to anyone watching the video unburdened by other task demands.

To systematically investigate the interrelation between attention and visual awareness, Rensink et al. (1997) created the "flicker paradigm". In the flicker paradigm two versions of a picture are shown one after another repeatedly with a blank screen in between (for a review, see Simons & Rensink, 2005). The two versions of the picture are generally identical except for one small change. In a typical application of this paradigm, participants are asked to find the changes between the two versions. They almost always find all the changes if they are given enough time, but depending on the sort, size, and placement of the change, it can take several minutes or more. This failure to quickly notice changes in one's perceptual field is called change blindness.

Rensink, O'Regan and Clark (1997) ran a series of experiments investigating change blindness. In their first experiment they had their participants find the change in a regular flicker task. They found out that the participants took twice as long to find the changes in the background than it took them to find the changes in the foreground. In another experiment they gave the participants verbal cues where the change was. When the participants' attention was directed, they were significantly faster at detecting the changes. Moreover, there wasn't a difference between the time it took them to notice changes in the foreground and in the background. They concluded from these experiments that the "key factor" to notice a change is attention.

The facts that people need to pay attention to notice a change and that we naturally notice changes in the foreground, taken together, should mean that people pay more attention to the foreground. Masuda and Nisbett (2001) challenged this idea. They thought that since Westerners and East Asians have different attributions (Westerns attribute outcomes to individual factors, whereas East Asians to situational factors), they could have differences in perceptual orientations. They showed Japanese and American participants clips of underwater scenes and then asked to recount what they saw. Their results showed that Japanese participants stated significantly more information about the background than the Americans, whereas there wasn't a difference in their statements about the foreground. Moreover, Japanese participants referred to an object's relationship to the background twice as much as Americans.

Masuda and Nisbett (2006) extended these findings using the flicker paradigm. They hypothesized that Japanese would be more sensitive than Americans would be to changes made in the backgrounds of the pictures. Their results showed that Japanese participants were just as fast as Americans to find the changes in the foreground and were a lot faster than Americans to find the changes in the background. When they asked their participants to recall as many changes as possible in briefly shown flickering scenes, Americans remembered marginally more number of

changes in the foreground, whereas Japanese remembered significantly more number of changes in the background.

Nisbett and Miyamoto (2005) have attributed these differences in the attentional processes of Westerners and East Asians to Westerners' having analytical perceptions and East Asians' having holistic perceptions. Analytical Westerners attend to salient objects and their category memberships, whereas holistic East Asians attend to contexts and relationships. The authors think the reason behind this difference is the differences in social structures: the East Asian social world is interdependent, while the Western social world is individualistic.

While this proposal is compelling for two cultures that vary quite considerably in their social make-up, the predictions are less clear for cultures that may fall somewhere between classic "East" and "West" mentalities. In the present investigation, we seek to explore the attentional processes of just such a group: the Turkish. There have been very few studies done on Turkish people's perceptions. Hence, it is unclear whether Turkish perception is holistic or analytic. Turkey is located between the individualistic West and the collectivistic East. Moreover, although Turkish history is quite collectivistic, current trends and the growth of capitalism in Turkey suggest that people are becoming rapidly more individualistic (Çileli, 2000). Çileli administered surveys to hundreds of college-aged people in Ankara in 1989, in 1992, and in 1995. The participants scaled 36 values according to how they thought these values affected their lives. The values were divided into terminal and instrumental values; terminal values were about where one wanted their life to end up and instrumental values were about one's behaviors. The results in 1989 showed that most important values for the participants were self-respect, freedom, inner harmony, equality, independence, honesty, broad-mindedness and courage, whereas the least important values were having an exciting life, pleasure, national security, salvation, politeness, imagination, cleanliness and obedience. On the other hand, in 1992 and 1995 the results showed that most important values for the participants were inner harmony, happiness, mature love, exciting life, ambition, cheerfulness and capability, whereas the least important values were freedom, social recognition, comfortable life, true friendship, politeness, honesty, helpfulness and imagination. Çileli concluded from these results that Turkish people were becoming more hedonistic and competitive and hence more individualistically oriented. In short, Turkish people have a unique relationship with collectivistic and individualistic orientations, making them a particularly interesting test case for exploring how they perceive the world. The outcomes have implications for helping us better understand the social orientation of modern Turks and, furthermore, for providing additional support for proposals about how attention is involved in the visual perception of change.

In the current study we carried out a change blindness experiment with the flicker paradigm in which Turkish and American participants looked at some scenes which had changes in the foreground or background. Our experiment was intended to explore whether and how Turkish and American perceptions differ. We recorded how long it took for the participants to find the change. We expected Americans to be quicker at detecting changes in the foreground than the ones in the background and Turks to be quicker at detecting changes in the background than the ones in the foreground. In other words, we expected to find that Turkish people had holistic perspective because we assumed Turkish culture to be primarily collectivists since the trend of individualism was fairly new, whereas Turks have always been interdependent. We also looked into how the culture of photographs affected the participants' reaction times. We used photographs that were taken in the USA or in Turkey and some photographs were neutral, as in they could belong to either country. We thought that Americans would be quick at finding changes in American and neutral pictures but slower at finding changes in Turkish pictures and Turks would be quick at finding changes in Turkish and neutral pictures but slower at finding changes in American pictures.

Methods

Participants

Our participants included two groups: American and Turkish. The American participants consisted of 15 Franklin & Marshall College students who were only fluent in English. A sign-up sheet was posted so students could sign up independently to participate in our study. We stated the requirements (being American and being fluent only in English) in our sign up sheet and included questions about nationality and language proficiency in our demographic questionnaire. The Turkish participants were 15 college students in Bilkent University in Ankara, Turkey. Again, a sign-up sheet was posted for students to choose a time to participate in our study. A requirement on the sign up sheet to participate was that the participants should be Turkish and speak only Turkish fluently. As we did with the American participants, we double-checked this by including questions about nationality and language proficiency in our demographic questionnaire. The participants of both groups received class credit for participating in this experiment. Of all the participants we had to disregard 2 participants because of failure to follow directions and 3 participants because they were outliers (their reactions times were two or more standard deviations above or below the mean). The final number of participants was 12 American (8 women, 4 men, age range: 18-23, M = 19.8) and 13 Turkish students (10 women, 3 men, age range: 18-23, M = 20.8).

Materials

The materials consisted of an iBook, various photos, and two computer programs (Adobe Photoshop and a Change Blindness application created at Franklin and Marshall College). The pictures were edited to be the same size and the focal/contextual changes were made with Adobe

Photoshop. They became flickering movies using the Change Blindness application. There were three categories of pictures: American, Turkish and neutral. American pictures were scenes only found in the United States and not in Turkey. American participants would be familiar with these scenes; whereas, Turkish participants would be unaccustomed to them. These scenes were a Halloween party, a street from Los Angeles, an intersection in the Times Square, a baseball figure, a football game, a house with Christmas lights, a statue of Abraham Lincoln, and a Hollywood star. Turkish pictures were scenes that could only be found in Turkey and not in the states. These scenes were familiar for Turkish participants but not for Americans. These scenes were people doing halay (a traditional Turkish dance), a saz ekibi (an orchestra of classical Turkish instruments), the blue mosque, a fancy evil eye, a chestnut stand on the sidewalk, a women making gozleme (big Turkish crepes), kina gecesi (the pre wedding celebration where the women put henna on their hands), and the kiz kulesi (a very known building in the middle of Bosphorus). The neutral pictures were scenes that can be found in both countries. All participants would be accustomed to these scenes. These scenes were a family dinner, a girl with a birthday cake, three people in skiing outfits, a dorm room, a beach, girls eating dessert, a guy playing electric guitar and a dining hall.

A control variable for all the scenes was complexity because it has been suggested that complexity of a scene can prime the viewer to perceive holistically or analytically (Masuda & Nisbett, 2006). The photographs were all altered so they were of equal size. Each category has an equal number of simple and complex scenes. The simple scenes were scenes either with a straightforward focal point or with very few objects to focus on. The complex scenes had many objects and the subject of the scene isn't clear. In order to assess whether others think the photographs are simple or complex, we had non-participating students highlight the area of each photograph that they thought was the focal point. Simple scenes were defined has having only 1 area highlighted by all the people who did this pre-test. Complex scenes had 2 or more different areas of the picture highlighted. Half of all the pictures got a focal change (one that is in the area of the photograph that is the focus) and the other half received a contextual change (one that is made more in the background). We used the highlighted photographs to help us decide where the focus of each photograph was. In this way we were able to know where to make focal and contextual changes. The changes that were made to the photographs were taking an object away from the photograph. Half of the photographs shown to the participants started with the object in question and half started without it.

Each picture was made into a movie using the Change Blindness application with both versions going back and forth and a gray scene between them (for the flicker effect). The outline for a movie would be the original picture (560 msec), a gray scene (120 msec), the modified picture (560 msec) and a gray scene (120 msec). The movie played in a loop until the change was found. The Change Blindness application also recorded the reaction time of each participants' identification of the change in each picture. Fifteen files were created, each containing all 24 movies in different random orders. Based on their ID number, participants viewed one of the fifteen files.

Procedure

Participants signed up in posted time slots. They came to the study and sat at a table with the computer and a mouse in front of them. All participants were informed of what will be asked of them in the study. They signed an informed consent form indicating their willingness to participate in the experiment. They were assigned ID numbers that correspond with their data and demographics so confidentiality was preserved. Participants were given a demographics questionnaire including questions on age, gender, year in school, country of birth, and language proficiency.

As soon as they were ready, the experimenter started the Change Blindness application with the correct file of photographs for that participant. The participant was asked to click the provided mouse when they found the change. The computer program would then pause and record how long it took the participant to find the change in that flickering picture. The recorded time would be the reaction time of that participant to that picture. Then the experimenter asked the participant to show the change. All data in which the participant identified an incorrect change was disregarded. The participant was shown the each flickering picture (8 American, 8 Turkish and 8 neutral). After the participant was done, the experimenter gave them a copy of the informed consent, which included the experimenters' e-mail addresses. The American participants did this study in Franklin and Marshall College's Barshinger Life Sciences Building. The Turkish participants did this study in Bilkent University's Psychology building. This part was carried out in Turkish since the participants' most fluent language was Turkish. Therefore, the experimenter's script was in Turkish, as were all instructions, the demographics questionnaire, and the consent form.

The dependent variable was the reaction time. The independent variables were the place of the change, the category of the picture and the nationality of the participant. We analyzed the results with a 2 (Foreground/Background change) X 3 (Turkish/American/Neutral photograph) repeated measures ANOVA with nationality (Turkish/American) as a between-subjects factor. Moreover, we ran a 2 (Foreground/Background change) X 2 (Simple/Complex picture) repeated measures ANOVA with nationality (Turkish/American) as a between-subjects factor.

Results

We did not expect to find similar results for American and Turkish participants; however, our results showed that there were not significant differences between them. In other

words there was no main effect of nationality F(1, 23) = .449, p = .510. There was not an interaction between nationality and place of the change F(1, 23) = .134, p = .717; however, there was a main effect of the place of the change F(1, 23) = 10.3, p = .004. As can be seen in Figure 1 American participants found the foreground changes (M = 22.5, SD = 3.51) faster than they did the background changes (M = 31.1, SD = 3.47) and Turkish participants also found the foreground changes (M = 24.0, SD = 3.38) faster than they did the background changes (M = 34.7, SD = 3.33).

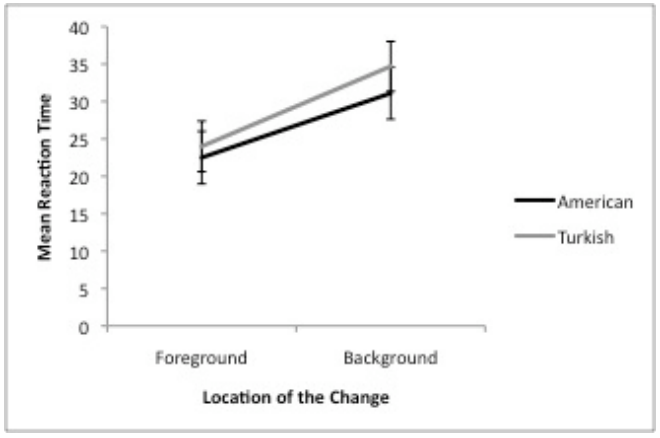


Figure 1: Mean reaction times of American and Turkish participants for finding the changes in the foreground and the background.

There was a main effect of the category of the picture F(2, 22) = 15.1, p < .001 ; however, there was not an interaction between nationality and category of picture F(2, 22) = 2.12, p = .143. That is to say Americans and Turks had similar reaction times within each category of picture. Americans were fastest in finding the changes in neutral pictures (M = 16.4, SD = 2.83) and took about the same time to find the changes in American pictures (M = 27.3, SD = 4.80) and Turkish pictures (M = 36.8, SD = 4.70) just like Turks were fastest in finding the changes in neutral pictures (M = 19.0, SD = 2.72) and took about the same time to find the changes in American pictures (M = 38.8, SD = 4.62) and Turkish pictures (M = 30.3, SD = 4.52). The mean times for finding the changes in the three categories of pictures for both nationalities are presented in Figure 2.

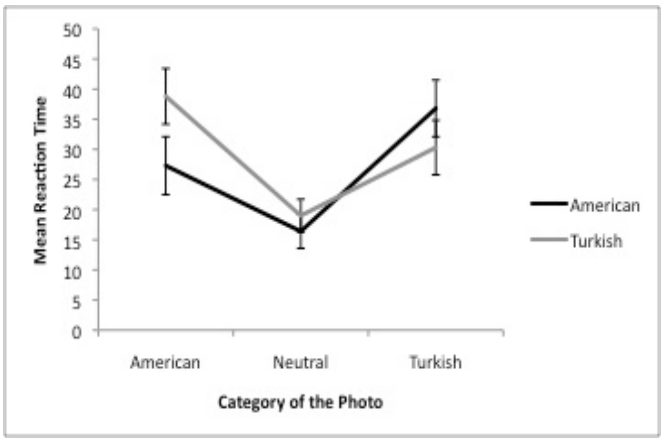
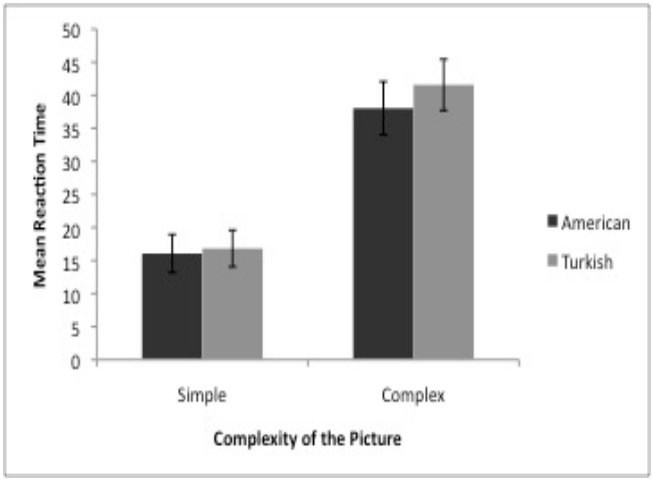


Figure 2: Mean reaction times of American and Turkish participants for finding the changes in the three picture categories

There was a main effect for complexity of the picture F(1, 23) = 61.6, p < .001 as shown in Figure 3. As one can see from Figure 3 the participants took significantly less amount of time to find the changes in the simple pictures than they did in the complex pictures. There was not an interaction between nationality and complexity F(1, 23) = .315, p = .580 which means that Americans' and Turks' reactions time were similar for both complexity conditions. There was not an interaction between complexity and location either F(1, 23) = 4.02, p = .057. The participants found the changes in the foreground fastest in both complexity conditions. There was not a three-way interaction between complexity, location and nationality F(1, 23) = .187, p = .669.



Firgure 3: Mean reaction times of American and Turkish participants for finding the changes in simple and complex pictures.

Discussion

As the past research suggested we found that Americans find the foreground changes faster than background changes. We also found that this pattern of results held for the Turkish participants. These results did not support our hypothesis that Turkish people have a holistic perspective. These results suggest that Turkish people have an analytical perspective.

This finding can perhaps be explained by the changes Turkey has been going through. Capitalism is growing in Turkey, which encourages people to be more individualistic and hence more analytic. Every day Turkish people are trying harder to be more like Westerners. American movies influence what is shown in Turkish movie theaters and on Turkish television. Traditional Turkish dances are being regarded as lame, whereas hip-hop and break dances are being regarded as cool. A lot of Turkish values are getting lost and Western ideas are becoming more popular. In a study a group of Turkish high school students were shown the video clip of Rammstein's "We are all living in America" and were asked about their thoughts (Pehlivan, 2007). The general consensus of the students was that the USA was more advanced than Turkey and also than various African or East Asian countries. In other words they associate that advancement with the West. One of the students even called Turkey "an orphan" since he reasoned the Western countries were so much better than Turkey. These current individualist attitudes might account for our results.

A reviewer has pointed out to us that our results also might be taken to mean that people from collectivistic cultures can have analytical perspectives, and that the relativistic viewpoint of Masuda and Nisbett (2006) should perhaps be questioned. We believe our reviewer's concern is a valid one since there has been some evidence found in the favor of it. For example, de Fockert et al. (2007) have found that people from a very traditional culture in South Africa had extremely analytical perceptions. We do not agree with this point of view for the following reason. Past research has suggested that Turkish youth are becoming more and more individualistic. We believe that is why our results suggest that Turkish people have analytical perspective. To confirm our hypothesis it might be useful to compare older Turkish people to younger ones in this paradigm. We expect that older adults might be more collectivistic and therefore show a different pattern in a change blindness study. A clear-cut difference between the patterns of older and younger Turkish people would indicate two things. Firstly, it would indicate that our results from the current study could be explained by the transition Turkish culture is going through. Secondly, it would indicate that the idea that holistic cultures might have analytical perspectives did not hold true in the case of Turkish people. The current study is only a preliminary to further research on Turkish people's perceptions.

In fact, there is much more room for research in this area, which has in general been under-explored. A comparison experiment between Turkish people and East Asians could be done in order to further investigate the possibility that Turks have an analytical perspective. A study like Masuda and Nisbett's (2001) would uncover whether Turkish people pay attention to background or the relationship between objects similar to East Asians. Another way of taking this research further would by doing a real-life change blindness experiment like the study in which random people on the campus of Vanderbilt University were asked to remember the color of a binder the experimenter was holding, and the word inside the binder (Varakin, Levin, & Collins, 2007). The participants were unaware of many changes in their environment like the font of the word in the experimenter's binder. The participants in this study were all Westerners. The results could have been different if the participants were East Asian and even if they were Turkish. Just because Turkish people were similar to the Westerners at noticing changes in flickering images on a laptop does not guarantee that they would be similar to Westerners at noticing changes in real life change blindness experiments.

In our experiment we tried to get rid of confounds by having only American and Turkish participants so that the groups would be more homogenous, unlike the study in which different cultures of East Asia were grouped all these different cultures (Chinese, Japanese and Korean) under one group (Masuda & Nisbett, 2005). Furthermore, we tried to control for differences in the photographs as much as possible. We used half simple and half complex pictures in each category of culture of the photograph (American, Turkish, and neutral). Our purpose in controlling complexity was that we did not want to end up with one of our categories of pictures that consisted of only complex pictures. A reviewer has suggested that American and Turkish participants might have been affected differently by the complexity of the picture. Our results did not support this. We found that American and Turkish participants had similar reaction times for both simple and complex pictures.

In our experiment were expecting the participants to be fastest in the pictures they were familiar with. There was a main effect of the category of the picture but it was not like how we expected it to be. All participants found the changes faster in the neutral pictures than the Turkish or American pictures. This could be because the changes in the neutral pictures might be slightly bigger than the changes in the other categories since we only approximate the size of the change. This difference could also be due to the subject of the pictures. Only half of the Turkish and the American pictures were about people but almost all of the neutral pictures were of people. The changes were not always made to the people but it could be that the participants were better at detecting changes in pictures of people. This unexpected effect emphasized that the pictures and the changes should be even more controlled. More control on the changes could be accomplished by having a constant change in all the pictures. For example, a cup would be appearing and disappearing in pictures. This cup could be in the

foreground or the background; hence, keep the size of change same no matter where it occurs.

The current study aimed to get some insight into Turkish people's perceptions. We ran a change blindness experiment on Turks and Americans. We were expecting to find that Turkish people have holistic perceptions like East Asians; however, our results suggest that they have analytical perceptions like Westerners. We believe that there needs to be more research done in this area to understand how exactly Turkish people perceive the world.

Acknowledgments

Prof. Michael Anderson served as scientific advisor for this project; we would like to thank him for is unwavering support and help. We would like to thank Bilkent Univeristy, especially Emre Ozgen for all their help. We would like to thank Prof. Krista Casler for her helpful comments on an earlier draft.

References

Çileli, M. (2000). Change in value orientations of Turkish youth from 1989 to 1995. *The Journal of Psychology, 134*(3), 297-305.

de Fockert, J., Davidoff, J., Fagot, J., Parron, C., & Goldstein, J. (2007). More accurate size contrast judgments in the Ebbinghaus Illusion by a remote culture. *Journal of Experimental Psychology: Human Perception and Performance, 33*(3), 738-742.

Masuda, T., & Nisbett R. E. (2001). Attending holistically versus analytically: Comparing the context sensitivity of Japanese and Americans. *Journal of Personality and Social Psychology, 81*(5), 922-934.

Masuda, T., & Nisbett R. E. (2006). Culture and change blindness. *Cognitive Science: A Multidisciplinary Journal, 30*(2), 381-399.

Nisbett, R. E., & Miyamoto, Y. (2005). The influence of culture: Holistic versus analytic perception. *Trends in Cognitive Sciences, 9*(10), 467-473.

Pehlivan, H. (2007). Are we all living in America?. *Elektronik Sosyal Bilimler Dergisi, 6*(22), 270-282. Retrieved April 26, 2008, from http://www.e-sosder.com/dergi/22270-282.pdf

Rensink, R. A., O'Regan, J. K., & Clark, J. J. (1997). To see or not to see: The need for attention to perceive changes in scenes. *Psychological Science, 8*(5), 368-373.

Simons, D. J., & Chabris, C. F. (1999). Gorillas in our midst: Sustained inattentional blindness for dynamic events. *Perception, 28*, 1059-1074.

Simons, D. J., & Rensink, R. A. (2005). Change blindness: Past, present, and future. *Trends in Cognitive Sciences, 9*(1), 16-20.

Varakin, D.A., Levin, D. T., & Collins, K. M. (2007). Comparison and representation failures both cause real-world change blindness. *Perception, 36*(5), 737-749.

Analogical Mapping Through Visual Abstraction

Jim Davies (jim@jimdavies.org)
Science of Imagination Laboratory
Institute of Cognitive Science; Carleton University; 1125 Colonel By Drive
Ottawa, Ontario, K1S 5B6 Canada

Patrick W. Yaner (patrick.yaner@gmail.com)
Artificial Intelligence Laboratory
College of Computing; Georgia Institute of Technology; 801 Atlantic Drive
Atlanta, GA 30332-0280 USA

Abstract

Analogical mapping theories tend to focus on matching identical symbols (either for objects or the relations between them). In the domain of visual representations we implemented a mapping system that uses separate domain knowledge (a shape-type superclass hierarchy) to re-represent analogs such that identicality can be found at different levels of abstraction. Such a scheme is useful where shape, and not the spatial layout of the analog images, is important to aligning visual objects.

Introduction

Mapping, a core part of analogy, is finding the alignments between the elements of two analogs. For example, in an analogy between a face and the front of a car, a mapping might include an alignment between the eyes and the windshield. Such an alignment might be based on the fact that the "perceptual" input to the car happens at the windshield for a car as do the eyes for a face.

Mapping is combinatorially complex, and this complexity is reduced by finding similarity between the elements to be mapped. In the example above, the alignment is justified by the *functional* similarity between the eyes and the windshield. In the discussion section we will describe different similarity measures for implemented mapping systems.

Our work focuses on mapping for purely *visual* analogs. That is, we are exploring different similarity measures that are appropriate for mapping the visual components of images. To return to the car/face example, rather than *functionally* aligning the eyes and the windshield, an agent that focused on visual similarity might align the eyes to the headlights because they both consist of two elements, are both round, and are horizontally oriented.

One way to find similarity between visual elements is by identification of identical symbols relating the elements. Most simply, if two elements are described in the representation as *square*, then the mapping agent can favor their alignment. Another way is to identify identical symbols that *relate* elements of the same image. For example, if in one image *element-x* is related to *element-y* with a symbol *is-above*, then a mapper interested in the structure of images might want to align x and y to f and g, if indeed f *is-above* g in the other analog—regardless of what shape x, y, f and g are. Structure Mapping Theory (Gentner, 1983) uses identicality of the symbols describing structure to find mappings between analogs.

Grouping

The Gestalt psychologists found that people perceptually grouped visual elements according to, among other aspects, shared orientation, color, and proximity. This provides psychological evidence for an explicit representation of visual element grouping. Broadly speaking there are two kinds of groups in our work: *aggregations* and *sets*. Aggregations are multiple visual elements that form one coherent shape (e.g. a square is an aggregate of four lines). Sets are groups of elements that are unconnected but similar in some way (e.g. nuts in a bowl).

Sets and aggregates appear at a certain level of abstraction, at which they can be aligned to each other as visual elements. An agent with a flexible representation can, however, zoom into these groups. There are two reasons an agent might want to do this: First, If the agent must decide which group to align to which other group, the conflict resolution might require an examination of the contents of that group. Second, it might be important to align group members. For example, imagine aligning two armies—at this level of abstraction armies can be moved and split apart, and it makes sense to have the armies' generals simply be members of the army sets. But if the agent needs to reason about the generals in particular, it could be important to know that the general in one set aligns to the general in the other. In analogical problem solving, for example, certain operations need to be applied to elements of analogs.

Different levels of representation are needed for different transformations applied to the analog (Davies, Goel, & Nersessian, 2003). Likewise with aggregates, mapping one box to another is the right level of abstraction for motion of the entire set, but opening one side of the box by moving one of its constituent lines requires a mapping at the component level.

For these reasons it is helpful for an agent to be

flexible in its representations such that it can reason at multiple levels of grouping abstraction.

Shape-type Superclass Hierarchy

The above similarity notions use the nature of the analogs as given in the representation. However similarity can also be found through the application of domain knowledge to the analogs. In the visual domain, this can take the form of a shape-type hierarchy (See Figure 2.) For example, a right triangle and an isosceles triangle are similar because they are both triangles. Even in cases where the term *triangle*, and its relation to *right-triangle* and *isosceles-triangle* are not explicit in the representation of the analogs, an agent can use the domain knowledge of a shape-type hierarchy to find element similarity. We will show that abstraction using this hierarchy is particularly useful (compared to structure-mapping) for analogs in which the spatial arrangement of the visual elements is less important than the shapes of the objects represented. The examples we explicate below are of this type. Using abstraction has been used in Minimal Ascension (Falkenhainer, 1988) and in cross-domain analogical learning (Klenk & Forbus, 2007).

Our theory is that aggregation and set abstraction are useful representations for mapping *visual* analogs, and that re-representation using a shape-type hierarchy can address some cases of ontological mismatch, where similar ideas cannot be identified as such because they are represented with different symbols.

Model

In this section we will describe our theoretical models for the three kinds of visual abstraction in more detail.

Our representational architecture consists of propositions, each of which connects two symbols with a relation. For example

```
(butterdish looks-like rectangle)
```

connects the *butterdish* symbol to *rectangle* with a relation that the agent uses to align symbols with the same shape type. This uses the Covlan visual language (Davies & Goel, 2007).

Set and Aggregation Hierarchies

Sets are explicit visual objects with no shape. They have links to their members, and the members likewise have back-pointers to the sets. Sets can contain other sets as members. Sets get aligned to other sets through some similarity measure based on the shapes of their members, but the members themselves will not be mapped unless the agent has a specific reason to do so. The cognitive justification for this is introspective: we do not appear, for example, to align each paper clip in one pile to each binder clip in another pile unless there is reason to do so, even though we might see the two sets as similar.

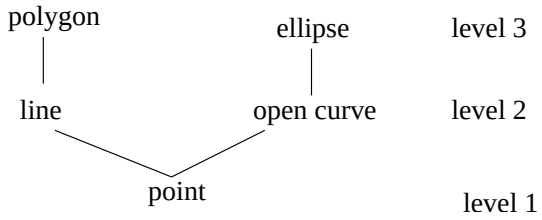


Figure 1: Component Aggregation Hierarchy

The same goes for aggregate objects. At the first pass of mapping, the aggregates are mapped to each other. In instances of conflict, the agent uses a measure of the shape similarity of the components to resolve it.

When there is a need to align the members or components, the entire mapping function can be recursively called on the sets or aggregates in question—that is, restart the mapping process as though the two aligned sets or aggregates were to the two images to be mapped.

The simplest kind of aggregates are visual elements aggregated together. Using knowledge of a shallow *aggregation hierarchy* (see figure 1) the agent can bring more domain knowledge to bear on the aggregate objects to align the components. When called upon to do align the components of simple visual elements, to resolve conflicts the most specific element in either aggregate object is decomposed into its aggregate parts. Then the identity-based mapping system tries again. This process repeats until all the align-able sub-elements are aligned.

Shape-Type Hierarchy

Each level of the shape-type hierarchy is associated with a level number. The higher the number, the less abstract the shape is. Abstraction is changing a shape to its more abstract form. For example, abstracting a *square* (level 8) means transforming it to a *rectangle* (level 7).

Process

Mapping is iterative. A *mapping* is a set of *maps*, which are alignments between a visual element in one analog to an element in the other.

1. Create maps of identical shape types, including aggregates and sets, ignoring the components of aggregates and the members of sets. If there is a conflict with mapping aggregates and sets, break them up into their constituents and see which are the most similar (a simple vector analysis of the contents).

2. If there are no more shapes to map in either the target or the source analog, recursively run this

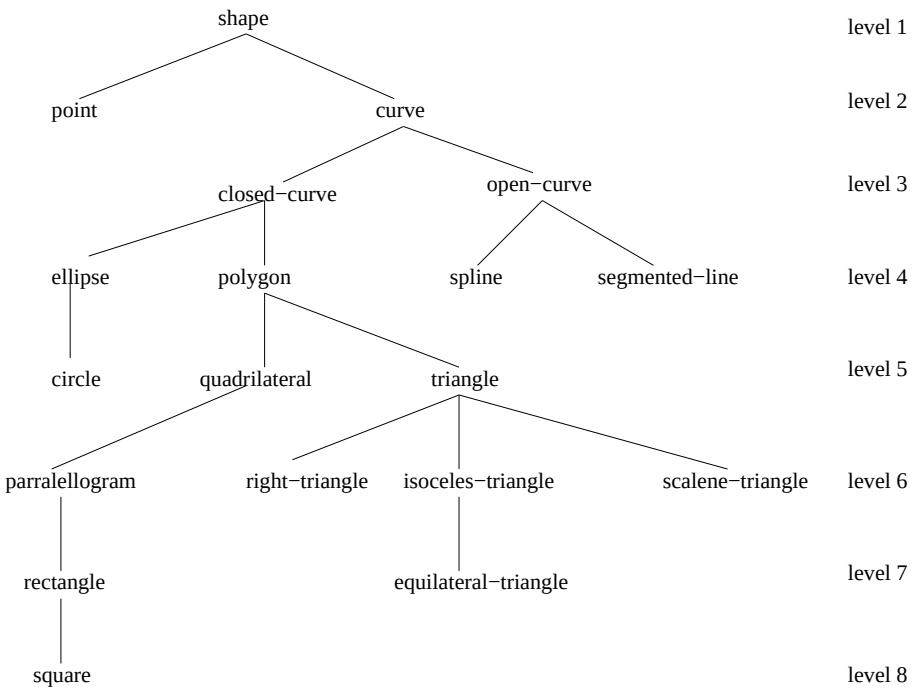


Figure 2: Shape Type Hierarchy

mapper on the members of sets and the components of aggregates, if any, then exit.

3. Abstract the highest numbered visual primitive in either analog one level. Go to step 1.

The visual elements need to be abstracted separately. If every element in the image is abstracted at once, matches will miss each other as they pass through levels of abstraction.

Implementation

We have some implemented the ideas above in a running computer program called *Thalassa*. Thalassa maps identical shapes and groups, and abstracts shapes with the shape-type hierarchy.

Thalassa has two basic components: a frame system and a "classical" problem solver. First, the "memory" of shape types and images with their elements and aggregate objects and such was built using a simple frame system. Relations of the sort `(building looks-like square)` lend themselves naturally to a frame-based representation, where the frame for `building` has a slot `looks-like` with the filler `square`. Likewise an image is a frame with a slot `contains-elements` whose filler is a list of elements (symbols naming visual element frames) in the system. Though the content of the frames in our actual implementation was sparse, the idea is that any extra information that might be useful to a larger problem-solving context could be added. For instance, surely people are aware of the location (qualitative or relative) within an image of a particular visual element, and the frame representation naturally allows one to add a slot `has-location` (or what have you). The only information actually used in the implementation was the `looks-like` slot for each visual element and the `has-size` slot (which took sizes like `small`, `medium`, and `big`).

Problem Solver

The second part of the implementation was the problem solver itself. Following the problem space hypothesis, and using the Classical Problem Solver (Forbus & DeKleer, 1993), we transformed the mapping problem into a search problem. One can think of one "state" of the search as the current set of maps—that is, the list of elements in each of the two analogs that have been mapped so far. An operator generating a new state in the search can do one of two things: map as many elements with the same shape as possible (generating a new state for each partial matching possible), or else pick one element to abstract.

The actual data structure representing a "state" in the search had the following elements:

Source The name of the source image frame

Target The name of the target image frame

Maps The list of maps gathered so far

Unmapped Source Elements A list of source elements that have not been mapped onto target

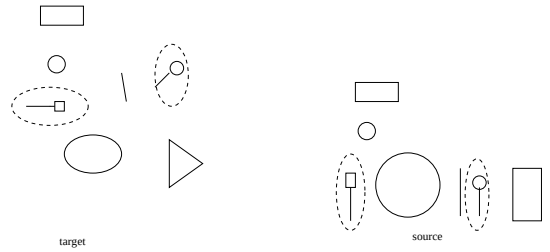


Figure 3: These two analogs represent a table before and after dinner. The fork and spoon are aggregate objects. Pictured are plates, butter dishes, forks, knives, spoons, and napkins.

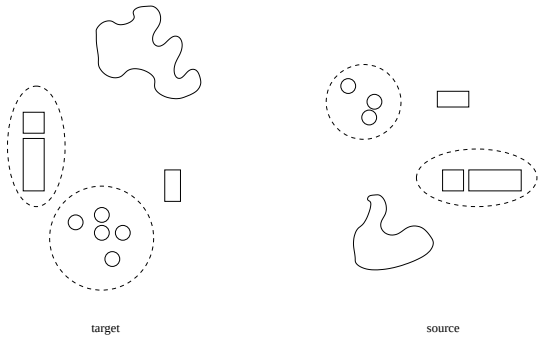


Figure 4: In the lot source there is a group representing a 18 wheeler (an aggregate of a square representing the cab and a rectangle representing the trailer). Also in the image is a set of garbage cans. Note there are a different number of cans in the analogs. The irregularly shaped object is a puddle, and the lone rectangle is a dumpster.

elements, initially set to all the source elements in the image.

Unmapped Target Elements A list of target elements that have not found source analogs, initially set to all the target elements in the image.

Needs Abstraction? A flag indicating that there are unmapped source and target elements that cannot be mapped without abstracting the shape types.

Abstractions A data structure that associates with each element (source and target) it's currently abstracted shape type. This is initially filled with the fillers from the `looks-like` slot, and as shapes are abstracted the contents slowly change.

The goal condition in this search is simply that either the `unmapped-source-elements` list or the `unmapped-target-elements` list becomes `nil`, in which case there are no more elements to match.

The next state operator `generate-new-mappings` simply checks the `needs-abstraction` flag, calling `abstract-one-element` if it is true, and `extend-mappings` otherwise. The `abstract-one-element` function is quite simple: it sorts the complete list of unmapped elements, choosing the one with the highest level arbitrarily (that is, if there are several, choosing the one at the front of the sorted list as an arbitrary choice) to abstract one level, if possible. It cannot abstract anything past `shape`, obviously (that being the top of the hierarchy), and so returns nothing if all of the shapes in the image are fully abstracted.

The `extend-mappings` function is much more complex. It has three parts: (1) generate all possible `maps-to` relations for each unmapped target element, separating them into whole mappings as it goes; (2) separate `maps-to` relations within each mapping that map onto the same source into separate whole mappings; and (3) generate a new state for each whole mapping.

The first part of this operation makes a list of all the elements that map to the given target (which is simply a list of all the unmapped sources that are identical under the `looks-like` relation and the current abstractions), and separates these into whole mappings, where one whole mapping has one `maps-to` relation for each target (there may be targets with no mapped sources, of course). It takes care to assemble all combinations when doing this.

The second part of the mapping looks within each mapping for two maps that map separate targets to the same source. If one is found, it is split into two mappings by removing one and then the other map from the mapping.

The third part simply takes each whole mapping, filters out those elements (sources and targets) which have been mapped from the unmapped elements lists, and generates a new state. A list of all new states is returned.

The search returns all mappings that it found. It's not clear to us that there is necessarily any cognitive plausibility in this decision, but for the sake of implementation we thought it best to have it return all mappings rather than choose one arbitrarily as the "best" mapping to return (one of the examples discussed below had several possible mappings).

One feature that has not been implemented in this version is recursing on aggregates and sets. Also, the system could often abstract in more than one way, and it's not clear that choosing one thing arbitrarily to abstract is the best decision; it seems wiser to have it abstract in all possible ways, generating new states (and hence new subtrees) in the search (space) for each one.

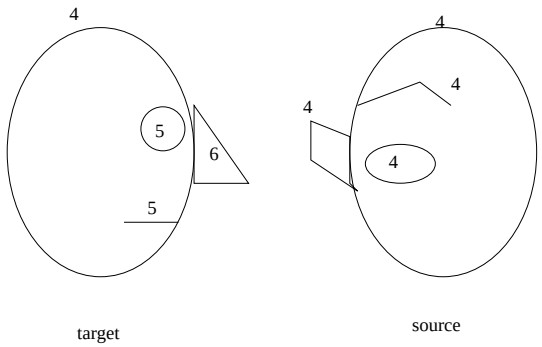


Figure 5: The faces example has similar faces reflected and rotated 180 degrees. Such transformations do not preserve many spatial relations, with certain exceptions such as containment and connections. The numbers represent the level of specificity of each shape in the shape-type hierarchy.

Test Examples

We ran this system on three test examples. The first, illustrated in Figure 3, was a table set up before and after a meal (not necessarily images from the same meal), where the elements are scattered about the table after the meal, and so structure would probably be uninformative. However, a fork looks like a for, a plate looks like a plate, and so on. In fact, the system found four mappings: the fork could map to the fork or the spoon (and vice versa), and the napkin could map to the napkin or the butter dish (and vice versa). Everything else mapped to the element of the same name, thus giving four mappings.

The second, illustrated in figure 4, was supposed to be an overhead view of a parking lot, with a dumpster and a group of trash cans and a truck aggregate (cab and trailer) and a puddle all in different positions. this had only one mapping.

The third example, illustrated in figure 5, mapped a face to another face reflected and inverted. Again, only one mapping was found.

Discussion

We are theorizing about solutions to the cognitive problem known as the ontological mismatch problem: When two ideas that should be thought of as similar are not because they are represented with different symbols. This problem manifests itself in mapping because the labels for objects and relations often do not match exactly. For example, mapping *orbits* with *revolves-around*.

Our solution uses shape-type abstraction, which is a content account of the visual domain. EMMA (Ramscar & Yarlett, 2003) uses a content account as well to solve ontological mismatching for analogy. The knowledge EMMA uses is the correlation of word proximity in text (Latent Semantic Analysis, LSA). Each word in LSA correlates with each other word. When trying to map, different words with a correlation above a certain threshold are considered equivalent, and can be mapped.

The Structure-Behavior-Function knowledge representation language (Goel et al., 1997) offers another means for resolving ontological mismatches. SBF language representations of systems include functional descriptions each device component.

Forbus et al. (1998, p.246) and Hummel and Holyoak (1997) both suggest that ontological mismatches can be resolved through re-representation using abstraction (e.g. *lift* and *push* can abstract to *move*. This idea also suggests a superclass hierarchy, but to our knowledge neither research group has implemented this.

Other implemented mappers rely on the identicality of symbols (either of the mapped concepts or the relations between them) and on a canonicalized representation to avoid ontological mismatches.

Ours is a content based account that uses re-representation using domain knowledge of visual objects.

Conclusion

In this paper we have described a method based on domain knowledge for analogical mapping of visual representations. Specifically, we focused on grouping and shape-abstraction. The one-to-one mapping constraint is maintained for our system, but sets and aggregates are treated as object to be mapped. Our ideas are implemented into a running computer program called Thalassa. Future versions of Thalassa will be able to recursively map set members and aggregate components.

We conjecture that these content-based mapping strategies will prove superior to structure-mapping in cases when the analogs share similarly-shaped components but where the spatial arrangement of the objects is disordered. A full, cognitively plausible model of visual mapping will need to include both structure and object mapping. Future research will test these claims through computational comparison with structure-based mapping engines on several examples.

References

Davies, J. & Goel, A. K. (2007). Transfer of Problem-Solving Strategy Using Covlan. Journal of Visual Languages and Computing: 18, 149–164.

Davies, J., Goel, A. K. & Nersessian, N. J. (2003). Visual Re-Representation in Creative Analogies. In A. Cardoso & J. Gero (Eds.) *The Third Workshop on Creative Systems.* International Joint Conference on Artificial Intelligence, 1–12.

Falkenhainer, B. (1988). *Learning from Physical Analogies.* Technical Report No. UIUCDCS-

R-88-1479, University of Illinios at Urbana-Champaign. (Ph.D. Thesis)

Forbus, K., & De Kleer, J. (1993). *Building Problem Solvers.* MIT Press.

Forbus, K., Gentner, D., Markman, A. B., & Ferguson, R. W. (1998) Analogy just looks like high level perception. Why a domain-general approach to analogical mapping is right. *Journal of Experimental and Theoretical Artificial Intelligence*, 10, 231–257.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7, pp 155-170.

Goel, A., Bhatta, S. & Stroulia, E. (1997) Kritik: An Early Case-Based Design System. In Maher, M. and Pu, P. (Eds.) *Issues and Applications of Case-Based Reasoning in Design*, Mahwah, NJ: Erlbaum, pages 87–132.

Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review*, 104, 427–466.

Klenk, M. & Forbus, K. (2007). Cross domain analogies for learning domain theories. In A. Schwering et al. (Eds.) *Analogies: Integrating Multiple Cognitive Abilities*, Volume 5-2007. Publication of the Institute of Cognitive Science, University of Osnabruck.

Ramscar, M. & Yarlett, D. (2003). Semantic grounding in models of analogy: an environmental approach. *Cognitive Science* 27:1. 41–72.

Output

```
CL-USER(33): (find-mappings 's-face-simage
't-face-simage) ; Fast loading
/net/hc283/yaner/work/current/7613/bps/proj/memory.fasl
(((S-FACE-HEAD MAPS-TO T-FACE-HEAD)
(S-FACE-MOUTH MAPS-TO T-FACE-MOUTH)
(S-FACE-EYE MAPS-TO T-FACE-EYE)
(S-FACE-NOSE MAPS-TO T-FACE-NOSE)))
CL-USER(34): (find-mappings 's-lot-simage
't-lot-simage) ; Fast loading
/net/hc283/yaner/work/current/7613/bps/proj/memory.fasl
(((S-PUDDLE MAPS-TO T-PUDDLE)
(S-TRUCK-AGGREGATE MAPS-TO
T-TRUCK-AGGREGATE) (S-DUMPSTER
MAPS-TO T-DUMPSTER) (S-CANS
MAPS-TO T-CANS))) CL-USER(35):
(find-mappings 's-table-simage
't-table-simage) ; Fast loading
/net/hc283/yaner/work/current/7613/bps/proj/memory.fasl
(((S-NAPKIN MAPS-TO T-BUTTERDISH) (S-GLASS
MAPS-TO T-GLASS) (S-FORK-AGGREGATE
MAPS-TO T-SPOON-AGGREGATE) (S-KNIFE
MAPS-TO T-KNIFE) (S-SPOON-AGGREGATE
MAPS-TO T-FORK-AGGREGATE) (S-PLATE
MAPS-TO T-PLATE) (S-BUTTERDISH MAPS-TO
T-NAPKIN)) ((S-NAPKIN MAPS-TO T-BUTTERDISH)
(S-GLASS MAPS-TO T-GLASS) (S-FORK-AGGREGATE
MAPS-TO T-FORK-AGGREGATE) (S-KNIFE MAPS-TO
T-KNIFE) (S-SPOON-AGGREGATE MAPS-TO
T-SPOON-AGGREGATE) (S-PLATE MAPS-TO
T-PLATE) (S-BUTTERDISH MAPS-TO T-NAPKIN))
((S-BUTTERDISH MAPS-TO T-BUTTERDISH)
(S-GLASS MAPS-TO T-GLASS) (S-FORK-AGGREGATE
MAPS-TO T-SPOON-AGGREGATE) (S-KNIFE
MAPS-TO T-KNIFE) (S-SPOON-AGGREGATE
MAPS-TO T-FORK-AGGREGATE) (S-PLATE MAPS-TO
T-PLATE) (S-NAPKIN MAPS-TO T-NAPKIN))
((S-BUTTERDISH MAPS-TO T-BUTTERDISH)
(S-GLASS MAPS-TO T-GLASS) (S-FORK-AGGREGATE
MAPS-TO T-FORK-AGGREGATE) (S-KNIFE MAPS-TO
T-KNIFE) (S-SPOON-AGGREGATE MAPS-TO
T-SPOON-AGGREGATE) (S-PLATE MAPS-TO
T-PLATE) (S-NAPKIN MAPS-TO T-NAPKIN)))
CL-USER(36): (dribble)
```

A Bottom-Up Parsing Model of Local Coherence Effects

Emily Morgan (emily@ling.ucsd.edu)
Department of Linguistics, 9500 Gilman Drive #108
La Jolla, CA 92093, USA

Frank Keller (keller@inf.ed.ac.uk)
Mark Steedman (steedman@inf.ed.ac.uk)
School of Informatics, 10 Crichton Street
Edinburgh EH8 9AB, UK

Abstract

Human sentence processing occurs incrementally. Most models of human processing rely on parsers that always build connected tree structures. But according to the theory of Good Enough parsing (Ferreira & Patson, 2007), humans parse sentences using small chunks of local information, not always forming a globally coherent parse. This difference is apparent in the study of local coherence effects (Tabor, Galantucci, & Richardson, 2004), wherein a locally plausible interpretation interferes with the correct global interpretation of a sentence. We present a model that accounts for these effects using a wide-coverage parser that captures the idea of Good Enough parsing. Using Combinatory Categorial Grammar, our parser works bottom-up, enforcing the use of local information only. We model the difficulty of processing a sentence in terms of the probability of a locally coherent reading relative to the probability of the globally coherent reading of the sentence. Our model successfully predicts psycholinguistic results.

Keywords: sentence processing; parsing complexity; local coherence; Good Enough parsing; Combinatory Categorial Grammar

Introduction

A major topic of inquiry in cognitive science is the process by which people produce and comprehend sentences. Human sentence processing is known to proceed incrementally: people construct syntactic and semantic interpretations gradually as a sentence unfolds, rather than waiting until after the whole sentence has been received. But although we know that syntactic information becomes available progressively while comprehending a sentence, it is still an open question to what extent decisions made early in the parsing process can constrain later decisions.

One phenomenon that can shed light on this question is local coherence effects. Local coherence effects arise when a sentence includes a substring with a plausible local interpretation that is incompatible with the global interpretation. (In other words, the interpretation is merely locally coherent, but not globally coherent.) A typical example (from Tabor, Galantucci, & Richardson, 2004) is:

(1) **A/R**: The coach smiled at the player tossed a frisbee.

A typical reader, seeing this sentence for the first time, will find it difficult to understand and will likely judge it to be ungrammatical. But this difficulty is unexpected in light of similar sentences:

(2) **U/R**: The coach smiled at the player thrown a frisbee.

(3) **A/U**: The coach smiled at the player who was tossed a frisbee.

(4) **U/U**: The coach smiled at the player who was thrown a frisbee.

These four sentences, all intended to be close paraphrases of one another, illustrate a puzzle: while the majority of readers reject (1), they accept (3) and (4), with mixed results for (2). These sentence differ on two dimensions: the past participle can be Ambiguous (such as *tossed*, which can be a past participle or a past tense form) or Unambiguous (such as *thrown*), and the relative clause can be Reduced (without *who was*) or Unreduced (with *who was*). Neither of these alternations generally changes the grammaticality of a sentence, so we would naively predict that if (4) is acceptable, then (1) is as well. Our challenge is to explain why this naive prediction is wrong. Intuitively, it seems that the local coherence of the substring *the player tossed a frisbee* in (1) as a plausible complete sentence is distracting from its globally correct interpretation as an object with a relative clause.

Tabor, Galantucci, and Richardson demonstrate the existence of local coherence effects as a psycholinguistic phenomenon in two different studies: in the first, they find increased reading times at the ambiguous past participle in (1). They present subjects with sentences from 20 sets of sentences like those seen above and measure reading times for each word using self-paced reading. In this methodology, longer reading times are taken to indicated increased processing difficulty. As expected based on previous studies (e.g. Ferreira & Clifton, 1986), they find substantially increased reading times for the Reduced cases as compared to the Unreduced cases, both on the past participle (e.g. *tossed*) and on the following word. Moreover, they find an unexpected interaction between Ambiguity and Reducedness: while the A/U reading times are not significantly different from the U/U reading times, the A/R reading times are substantially increased relative to the U/R reading times. This superadditive difficulty of the A/R condition is the signature of a local coherence effect.

In the second experiment, Tabor, Galantucci, and Richardson replicate the first using a grammaticality judgement task.

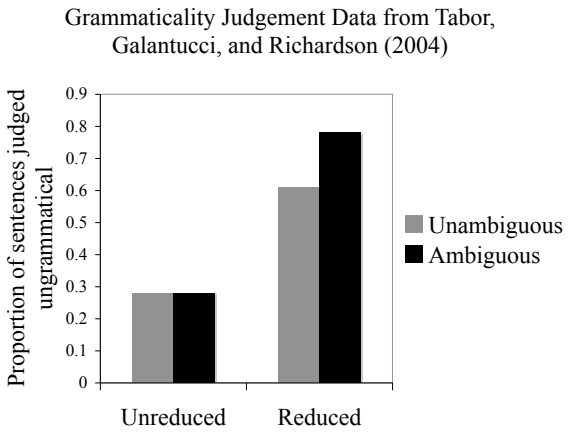


Figure 1: Grammaticality judgement data from Tabor, Galantucci, and Richardson (2004). The signature of a local coherence effect is the superadditive proportion of ungrammatical judgements in the Ambiguous/Reduced condition.

They find decreased acceptance of Reduced sentences as grammatical, with an interaction between Ambiguity and Reducedness such that A/R sentences are judged unacceptable superadditively often (see Figure 1). Once again, decreased acceptability judgements are taken to indicate processing difficulty.

Note that sentences in the A/R condition are not just standard garden path sentences. In a standard garden path sentence, the disambiguating information comes after the reader has already been led astray. In contrast, in sentences such as (1), the disambiguating information comes at the beginning of the sentence. Thus the reader in theory already knows that *tossed* cannot be a past tense form and must be a past participle. Yet despite that, these sentences cause processing difficulty.

A model of human sentence processing should be able to predict the difficulty of sentences with local coherence effects. However, most existing models cannot. In particular, most standard theories of parsing assume that that all accrued knowledge from the parsing process is taken into account at all times. Models following this assumption can straightforwardly account for standard garden paths because there is nothing inconsistent about initially misinterpreting a sentence before having access to the disambiguating information. But these models cannot take the same position in accounting for local coherence effects: when the disambiguating information has already been seen and *smiled* has already been recognized as the main verb of the sentence, they cannot entertain the inconsistent possibility that *tossed* is also a main verb. Computational implementations of wide-coverage parsers generally also make this assumption of global consistency (e.g. Roark, 2001; Sturt, Costa, Lombardo, & Frasconi, 2003; Demberg & Keller, 2008). For many applications, this assumption may be convenient. But for a parser to be credible as a model of human sentence processing, it must be able to predict these psycholinguistic effects, which requires relaxing this assumption.

An alternate theory of sentence processing is Ferreira and colleagues' *Good Enough* (GE) parsing. Ferreira and Patson (2007) describe GE parsing:

> People compute local interpretations that are sometimes inconsistent with the overall sentence structure, indicating that the comprehension system tries to construct interpretations over small numbers of adjacent words whenever possible and can be lazy about computing a more global structure and meaning.

The GE theory of parsing asserts that people do not construct full representations for sentences the majority of the time. Rather, they construct just enough to complete the task at hand, only constructing a further representation if necessary. Moreover, because people base their first-pass constructions on local information and generally construct only partial parse trees, these partial parses may contradict one another. A GE parsing account can thus easily account for local coherence effects. We will develop a computational model of why local coherence effects arise within the framework of GE parsing.

Previous Models of Local Coherence Effects

Two models have previously attempted to account for local coherence effects: Levy (2008) uses a noisy-channel model to argue that because there is uncertainty in linguistic input, the parse of a sentence should be modeled as a probability distribution over a set of candidate sentences (including the intended sentence and its near-neighbors). Given such a probability distribution, the effect of reading each word can be modeled and quantified in terms of a belief update. Levy predicts that a larger change in beliefs will correspond to greater processing difficulty and longer reading times. This in turn predicts local coherence effects because the rarer sentences provoke larger changes in belief.

Levy's model only considers fully connected and grammatical (partial) parses as candidates, thus it does not capture the intuition of GE parsing. An additional limitation of the model is that due to the computational load of calculating near-neighbors, it has only been implemented using a toy Probabilistic Context Free Grammar (PCFG), rather than a richer, wide-coverage language model.

The other previously existing model of local coherence effects comes from Bicknell and Levy (2009). They again model local coherence effects as arising from belief updates. Specifically, they model them as the consequences of an update from a bottom-up prior belief to a posterior belief that takes top-down information into account. They thus predict processing difficulty in the case of locally coherent substrings because the bottom-up statistics make strong predictions about the category of the substrings, which are then contradicted by top-down information.

This model begins to capture the idea of GE parsing by looking at substrings of different lengths. However, it has no way to integrate the information it receives from these different substring lengths because evaluating these substrings is post hoc, not an actual part of the parsing process. Additionally, like Levy's (2008) model, it has only been implemented using a toy PCFG.

Thus there is currently no general, wide-coverage model of human parsing that implements a GE parsing strategy. Computational models of local coherence effects have instead had to account for the phenomenon indirectly, either through a noisy channel model or by predicting the effects without actually simulating the parsing process, and have been confined to parsing with small toy grammars. We will develop a model to address these shortcomings.

A New Model of Local Coherence Effects

Our goal is to model the process by which local coherence effects emerge as the result of Good Enough parsing, within the context of a wide-coverage parser. In the example sentence *The coach smiled at the player tossed a frisbee*, our intuition is that processing difficulty arises from the locally coherent reading of *the player tossed a frisbee*, which distracts from the globally coherent reading. Our model will capture this intuition by using a strictly bottom-up parser to remove the top-down influence of non-local constraints.

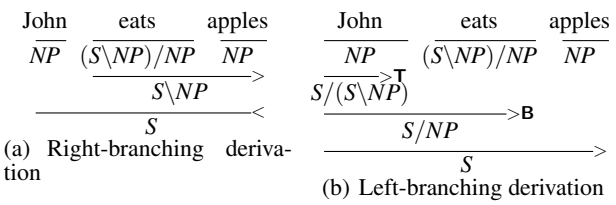
Strictly bottom-up parsing is frequently rejected as a plausible model for human parsing because, it is claimed, it does not allow for incremental interpretation. The standard argument says that a clause can only be interpreted when it is seen in full (i.e., at the end of a constituent). But in a strictly right-branching language, this means that nothing can be interpreted until the very end of the sentence because only then is any constituent completed.

To overcome this objection, our parser uses the Combinatory Categorial Grammar (CCG) formalism to represent linguistic structures. CCG was specifically designed to allow for incremental bottom-up parsing by using a more flexible notion of constituents.

Combinatory Categorial Grammar

Combinatory Categorial Grammar is a grammar formalism based on Categorial Grammar (CG). We base our description of it here on Steedman (2000).

CCG revolves around functional categories and rules for combining them. Categories can be either functions or arguments and are defined recursively: Base categories such as *S* and *NP* represent arguments. Functions are of the form α/β or $\alpha\backslash\beta$, where α and β are categories. To the right of the slash is the argument of the function, and to the left is its result. The direction of the slash indicates the directionality of composition: / means the argument is to the right and \ means the argument is to the left. An English verb phrase, for example, will have the category $S\backslash NP$, indicating that it combines with an NP on its left and results in a sentence. We also allow a finite set of features on our base categories, such as person, number, and gender on *NP*s. These are notated as e.g. $NP[3sf]$.



(a) Right-branching derivation

(b) Left-branching derivation

Figure 2: Right- and left-branching CCG derivations for the sentence *John eats apples*. $(S\backslash NP)/NP$ is the CCG category for a transitive verb. Without type-raising, *eats* can only combine with *apples*, yielding the typical right-branching derivation in (a). With type-raising, *John* can combine immediately with *eats*, yielding the left-branching derivation in (b).

A CCG derivation uses rules to combine categories. Pure CG relies on two rules, named $>$ and $<$, to combine categories:

(5) $X/Y \quad Y \rightarrow X \quad (>)$

(6) $Y \quad X\backslash Y \rightarrow X \quad (<)$

CCG introduces further combinatory rules that allow for more flexible notions of constituency than other grammar formalisms. In particular, it includes two lexical *type-raising* rules, named $>\mathbf{T}$ and $<\mathbf{T}$:

(7) $X \rightarrow T/(T\backslash X) \quad (>\mathbf{T})$

(8) $X \rightarrow T\backslash(T/X) \quad (<\mathbf{T})$

In these rules—which are here shown in the derivation, but in fact operate in the lexicon—T can be any lexical category taking X as argument. For instance, we could use $>\mathbf{T}$ to type-raise *NP* to $S/(S\backslash NP)$. Applying this rule limits the other categories the NP can combine with. Intuitively, we can think of the output of this rule as similar to an *NP* with nominative case-marking. It specifies not just that the word or phrase in question is a noun, but that it is a subject which must combine with a predicate.

These type raising rules allow us to parse a sentence incrementally by forming nontraditional constituents, leading to left-branching derivations (see Figure 2). CCG thus allows each new word of the input to be incorporated into the existing constituent structure as it is encountered, which makes incremental bottom-up parsing possible.

The Model

We take a bottom-up CCG parser as the basis of our model of human sentence processing. In order to predict processing difficulty caused by local coherence effects, we need a linking hypothesis to specify the relation between the parser output and psycholinguistic measures such as grammaticality judgements or reading times. Our linking hypothesis should embody the theory of Good Enough parsing, focusing on in-

terpretations of local substrings.

We adapt a model proposed by Jurafsky (1996) to predict garden path effects. To make graded predictions, rather than categorical distinctions, we will adopt a probabilistic framework, and consider the probabilities of various substrings of a sentence. In particular, we could consider either the inside probability $P(S \rightarrow \text{substring})$ (alternately written as $P(\text{substring} \mid S)$) or the inverse probability $P(S \mid \text{substring})$. We do not know of a computationally tractable way to calculate $P(S \mid \text{substring})$ from our parser. Calculating the inside probability, on the other hand, is a fundamental part of the parsing process. It is most parsimonious to base our model on the inside probabilities that are already being calculated.

Our intuition is that if an incorrect interpretation of a substring is highly plausible relative to the correct interpretation of the sentence, then it will cause processing difficulty. In a sentence such as *The coach smiled at the player tossed a frisbee*, the substring that we expect to cause difficulty is the locally coherent substring *the player tossed a frisbee*. We thus consider the ratio:

$$\frac{P(S \rightarrow \textit{the player tossed a frisbee})}{P(S \rightarrow \textit{The coach smiled at the player tossed a frisbee})}$$

In this case, the ratio will be high because *The player tossed a frisbee* is a relatively likely sentence. In the other three cases, the ratio will be low because none of the following are very plausible sentences:

(9) the player thrown a frisbee

(10) the player who was tossed a frisbee

(11) the player who was thrown a frisbee

Although in theory this ratio could be as low as 0, in practice this does not occur because there is generally some (low probability) way to parse each phrase as a sentence. We take this ratio as a measure of processing difficulty.

Implementation

We implement our model using a Combinatory Categorial Grammar parser based on the Cocke-Kasami-Younger (CKY) algorithm. This algorithm was originally developed for Context Free Grammars and uses dynamic programing to parse from the bottom up: given a sentence, it first calculates the probabilities of all ways to generate each word using a rule $X \rightarrow \textit{word}$. For each potential pair of categories X_1 and X_2 that could have generated adjacent words w_1 and w_2, it then calculates the probabilities of all ways to generate that pair using a rule $X_3 \rightarrow X_1X_2$. This allows us to calculate the inside probability $P(X_3 \rightarrow w_1w_2)$. Continuing iteratively, we can calculate the inside probabilities of all substrings of the sentence.

We used a modified version of the StatOpenCCG parser, developed by Christodoulopoulos (2008), which is itself an extension of the OpenCCG parser (White, 2008). StatOpenCCG implements a statistical version of the CKY algorithm which operates using a generative head-dependency model over CCG categories: From the parent (starting with a ROOT node), a head is generated with a certain probability. Then its sisters are generated with probability conditioned on the head category, the sister's direction from the head, and whether it is adjacent to the head. Although the number of CCG categories is theoretically infinite, our parser is constrained to only use categories that have appeared in the training data set. With this constraint, the runtime of the parser is bounded by $O(n^3)$. The parser has been trained on sections 1 through 22 of the CCGbank (Hockenmaier, 2003), a CCG version of the Penn treebank.

Our experiments use two different lexicons. The first lexicon is that taken from sections 1 through 22 of the CCGbank. However, this lexicon is too small to parse the majority of the sentences we wish to consider. To obtain a larger lexicon, we parsed six months of the New York Times (comprising approximately 50 million word tokens) taken from the Gigaword corpus (Graff, 2003). Sentences from the corpus were passed through the RASP tokenizer (Briscoe, Carroll, & Watson, 2006) and then parsed using the C&C CCG parser (Curran, Clark, & Bos, 2007). This state-of-the-art parser obtains labelled precision of 84.8% and labelled recall of 84.5% on section 23 of the CCGbank. It is extremely fast and provides the best parse accuracy from a CCG parser, making it convenient for obtaining large amounts of data to construct a larger lexicon. (However, it is not a cognitively plausible parser, as it relies on its supertagger and other cognitively implausible tricks to speed its parsing.) From this parsed sample, we extracted the lexicon for use in the StatOpenCCG parser (with the statistical parsing model over categories trained as before on CCGbank data). Although this lexicon of course contains quite a few errors, we verify that it nonetheless parses our test sentences correctly, placing the correct parses among the top results.

Experiments

We present two sets of experiments in which we test our model against the results from Tabor, Galantucci, and Richardson (2004). The first uses a small but high-quality lexicon to parse two test cases. The second uses a larger, error-ridden lexicon to parse a larger set of sentences. Recall that Tabor, Galantucci, and Richardson's (2004) study used 20 sets of sentences like those in (1)–(4).

Experiment 1: Test Cases using the CCGbank Lexicon

Because CCGbank is derived from a human-annotated treebank, the quality of the lexicon it yields is high. Nevertheless, it is small in comparison to human lexicons, and the passive relative constructions we are investigating are sparsely represented. In fact, the CCGbank lexicon contains only two words which are unambiguous ditransitive passive participles (i.e., $(S[\text{pss}]\backslash NP)/NP$ but not $(S[\text{dcl}]\backslash NP)/NP$—where [pss] indicates a past participle used in a passive construction, and [dcl] indicates a declarative sentence). These two words are

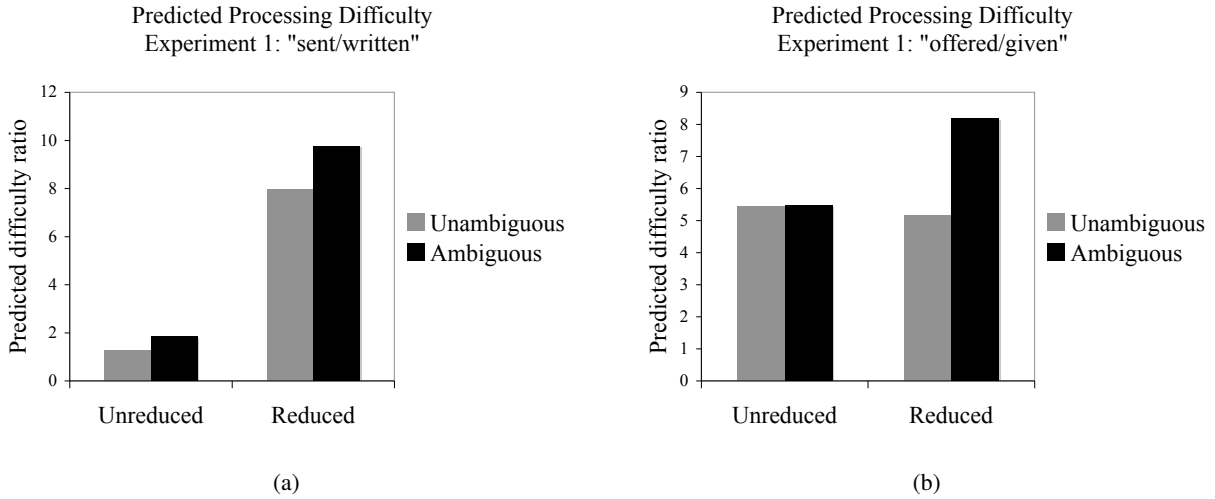


Figure 3: Results from Experiment 1, two test cases using the high-quality CCGbank lexicon. In both sets of sentences, the A/R case displays the correct pattern of superadditive difficulty.

written and *given*. Using these words, we construct two sentence sets, based on sentences used by Tabor, Galantucci, and Richardson:

(12) He questioned a congressman (who was) sent/written a letter.

(13) He addressed the woman (who was) offered/given a beer.

All words in these sentences are in the CCGbank lexicon. We parse them using our high-quality lexicon.

Results For these sentences, we obtain the predicted ratios:

$$\frac{\mathrm{P}(S \rightarrow \text{locally coherent substring})}{\mathrm{P}(S \rightarrow \text{whole sentence})}$$

Results are in Table 1 and Figure 3. We compare our results to the grammaticality judgements from Tabor, Galantucci, and Richardson (see Figure 1).

As we see in Figure 3(a), the set of sentences (12) displays the correct pattern of superadditive difficulty in the A/R case. While there is little difference in difficulty between the A/U and U/U conditions, there is a marked increase to the U/R condition, and a superadditive increase to the A/R condition. This mirrors the pattern seen in Tabor, Galantucci, and Richardson's grammaticality judgements.

We see the same superadditive pattern of difficulty in our results for the set of sentences (13), shown in Figure 3(b). Somewhat surprisingly, the U/R condition is in fact predicted to be marginally easier than the Unreduced sentences in this set. This may be because *given* is an extremely common word. Although it is unambiguous in that it cannot be a past tense, it is in fact a highly ambiguous word, with 18 entries in the CCGbank lexicon. For instance, it can serve as a preposition, as in *Given the weather, I will stay inside today*. Regardless of this slight puzzle, the A/R case displays the correct pattern of superadditive difficulty.

Table 1: Predicted difficulty ratios from all experiments, alongside grammaticality judgements from Tabor, Galantucci, and Richardson (2004).

Type	TG&R	Exp1: written	Exp1: given	Exp2
U/U	.28	1.27	5.45	5.74
A/U	.28	1.85	5.46	8.46
U/R	.61	7.96	5.16	11.60
A/R	.78	9.76	8.18	12.34

Experiment 2: Using the Gigaword Lexicon

Using the Gigaword lexicon, we are able to parse 13 out of the 20 sentences in the Tabor study. (Sentences were excluded only if their past participles were not present in the lexicon. All other vocabulary items are present.) We standardize all sentences to begin with a pronoun. Additionally, for the sake of parsing efficiency, we do not include the *by* phrases that give the agent of the sentence. We further shorten two sentence sets in ways that do not affect the target part of the sentence.

Results Results from Experiment 2 are shown in Table 1 and Figure 4. We compare our results to the grammaticality judgements from Tabor, Galantucci, and Richardson (see Figure 1). We find the correct trend of difficulties, with the A/R condition most difficult, followed by U/R, followed by the two Unreduced cases. We do not find the exact pattern of superadditive difficulty in the A/R case, due to the fact that the A/U case is in fact predicted to be much more difficult than the U/U case, in contrast to the grammaticality ratings. Because the Gigaword lexicon is very error-prone, it is difficult

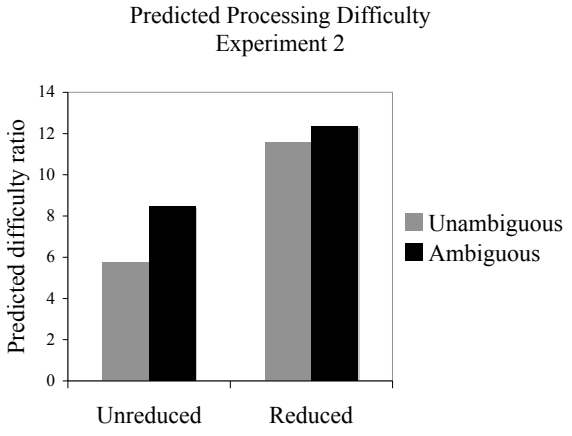


Figure 4: Experiment 2 results. We find the expected pattern of difficulty, but, due to the inflated predicted difficulty of the U/R case, do not see superadditive difficulty in the A/R case.

to draw any firm conclusions from this quirk in our results. However, we note that the A/R case is correctly predicted to be substantially more difficult than either of the Unreduced cases.

Conclusion

We have presented a model of local coherence effects using a wide-coverage bottom-up Combinatory Categorial Grammar parser. Our model can accurately predict which sentences humans will have difficulty in processing; specifically, it predicts the local coherence effects found by Tabor, Galantucci, and Richardson (2004). Our results support the psycholinguistic plausibility of CCG and the Good Enough theory of parsing by demonstrating that a parser that uses bottom-up local information can both perform well as a wide-coverage parser and predict specific psycholinguistic results.

Interestingly, the architecture of our version of the GE parser differs from Ferreira's original proposal. Ferreira (2003) proposes that GE parsing occurs via two separate strategies: one "algorithmic" and one "heuristic". In contrast, our parser does not include this separation: all analyses, both local and global, are produced by a uniform algorithm, and all are heuristically evaluated using the parsing model. This integration of strategies is a strength of our model, as it demonstrates how local coherence effects could emerge naturally as an inherent part of the parsing process.

In future work, we would like to make not just sentence-level predictions but word-by-word reading time predictions. Given that we have an entire parse chart, such predictions should be possible. We are currently choosing inside probabilities from two cells in the parse chart to compare, based on outside knowledge of where processing difficulty is likely to arise. We could do something similar for every cell in the chart, considering the inside probability of the substring it spans relative to the probability of the sentence as a whole. With word by word predictions, we could model reading time data as well as grammaticality judgement data. Such a model would be applicable to a wide range of psycholinguistic data beyond local coherence effects.

Acknowledgments

This work was supported by EU IST Cognitive Systems IP FP6-2004-IST-4-27657 "Paco-Plus".

References

Bicknell, K., & Levy, R. (2009). A model of local coherence effects in human sentence processing as consequences of updates from bottom-up prior to posterior beliefs. In *Proceedings of North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT) 2009 Conference* (pp. 665–673). Boulder, CO: Association for Computational Linguistics.

Briscoe, E., Carroll, J., & Watson, R. (2006). The second release of the RASP system. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions.* Sydney, Australia: Association for Computational Linguistics.

Christodoulopoulos, C. (2008). *Creating a natural logic inference system with Combinatory Categorial Grammar*. Master's thesis, University of Edinburgh.

Curran, J. R., Clark, S., & Bos, J. (2007). Linguistically motivated large-scale NLP with C&C and Boxer. In *Proceedings of the ACL 2007 Demonstrations Session (ACL-07 demo)* (pp. 29–32). Morristown, NJ: Association for Computational Linguistics.

Demberg, V., & Keller, F. (2008). A psycholinguistically motivated version of TAG. In *Proceedings of the 9th International Workshop on Tree Adjoining Grammars and Related Formalisms* (pp. 25–32). Tübingen.

Ferreira, F. (2003). The misinterpretation of noncanonical sentences. *Cognitive Psychology*, *47*, 164–203.

Ferreira, F., & Clifton, C., Jr. (1986). The independence of syntactic processing. *Journal of Memory and Language*, *25*, 348–368.

Ferreira, F., & Patson, N. D. (2007). The 'Good Enough' approach to language comprehension. *Language and Linguistics Compass*, *1*(1–2), 71–83.

Graff, D. (2003). *English Gigaword.* Linguistic Data Consortium, Philadelphia. (DVD)

Hockenmaier, J. (2003). *Data and models for statistical parsing with Combinatory Categorial Grammar*. Doctoral dissertation, University of Edinburgh.

Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science: A Multidisciplinary Journal*, *20*(2), 137–194.

Levy, R. (2008). A noisy-channel model of rational human sentence comprehension under uncertain input. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP).* Morristown, NJ: Association for Computational Linguistics.

Roark, B. (2001). Probabilistic top-down parsing and language modeling. *Computational Linguistics*, *29*(2), 249–276.

Steedman, M. (2000). *The syntactic process*. Cambridge, MA: The MIT Press.

Sturt, P., Costa, F., Lombardo, V., & Frasconi, P. (2003). Learning first-pass structural attachment preferences with dynamic grammars and recursive neural networks. *Cognition*, *88*, 133–169.

Tabor, W., Galantucci, B., & Richardson, D. (2004). Effects of merely local syntactic coherence on sentence processing. *Journal of Memory and Language*, *50*, 355–370.

White, M. (2008). *Open CCG: The OpenNLP CCG library.* (http://openccg.sourceforge.net/ [Online; accessed 27-July-2009])

Heuristics for Choosing Features to Represent Stimuli

Matthew D. Zeigenfuse (mzeigenf@uci.edu)
Michael D. Lee (mdlee@uci.edu)
Department of Cognitive Sciences, University of California, Irvine
Irvine, CA 92697 USA

Abstract

In this paper, we compare three heuristic methods for choosing which of a set of features to use to represent a domain of stimuli when we know the categories to which those stimuli belong. Our methods are based on three measures of category differentiation: cue validity, category validity, and their product, collocation. In a comparison of their ability to predict human similarity ratings in the Leuven Natural Concept Database, we find collocation to have the best performance, suggesting people use both cue and category validities in choosing which features to represent.

Keywords: Feature representation; basic-level categorization; similarity judgment.

Introduction

Of all the aspects of their world that could be represented, which do people actually choose? Imagine you are standing in front of a black dog named "Rover" with a small white patch of hair under its left eye. Which of its features do you choose to represent: its tail and four paws, its name, "Rover", and the spot under its eye? The last two of these may be useful for a representation of this particular dog, but are probably less useful to representing dogs as whole. Conversely, the first two may be useful for representing dogs, but are probably less useful for distinguishing Rover.

One method of learning about which aspects of a particular set of concepts people represent is the feature generation task (Rosch & Mervis, 1975). Often in this task people are asked generate a fixed number of features for each exemplar in a domain. In some cases, additional participants are asked to rate whether an exemplar has a feature for each combination of features and exemplars in a domain (Deyne et al., 2008). This leads to a large number of features describing each exemplar; however, not all of these features will be important to a person's representation.

Zeigenfuse and Lee (2008, 2010) provide a computational-level (Marr, 1982) approach to the problem. Similar to the theory of second-order isomorphism in perception (e.g. Shepard & Chipman, 1970), they argue that people represent those features that determine the similarity between objects and develop a model to infer which features are important using similarity judgments. Unfortunately, their method does not offer a psychological rationale for why one feature is important vis-à-vis an unimportant one, since it is more of a statistical solution than an account of feature importance.

This paper expands upon the computational approach of Zeigenfuse and Lee (2008, 2010) by exploring psychological theories of what makes a feature important. To this end, we propose heuristic methods for choosing important features based on how well a feature distinguishes categories from one another. We use these heuristics to begin answering the question of specifying what properties of a feature cause people to represent it.

Representation and Basic-Level Categories

Our heuristics are based on measures of category differentiation that have been proposed to explain basic-level categorization. Basic-level phenomenology refers to people's preference to categorize objects at a particular level in a category hierarchy, known as the basic level. Key finds are objects are categorized into basic-level categories more quickly than sub- or super-ordinate categories, basic level objects are named faster, objects are described preferentially with basic level names, more features are listed at the basic level than at the superordinate level, basic level names are learned before names at other levels, and basic level names tend to be shorter (Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976). These results suggest an intimate relationship between an object's basic-level category and its mental representation.

Category-Based Measures

Category Differentiation Given a feature representation, many theories of basic-level categorization score potential categorizations of the concepts in a domain through the information its categories give about the features of category members and vice-versa. Examples include, cue validity (Rosch et al., 1976), category validity, collocation (Jones, 1983), feature predictability (Corter & Gluck, 1992), category statistical density (Kloos & Sloutsky, 2006), and strategy length and internal practicability (SLIP: Gosselin & Schyns, 2001). Inverting this logic, given a set of categories, we can score features on their usefulness in providing information about which of the set of categories a concept belongs to, the information knowing a concepts category provides about whether it has the feature, or a mixture of the two.

Usefulness Measures The heuristics described here for choosing feature representations are based on three measures of feature usefulness. Suppose we have a domain of categories $\{c_1, \ldots, c_M\}$. Let $\boldsymbol{f}$ be an arbitrary feature. The first heuristic is *maximum cue validity*, which we define as $\max_{1 \leq j \leq M} p(c_j|\boldsymbol{f})$. The quantity $p(c_j|\boldsymbol{f})$ is known in the literature as the cue validity of feature $\boldsymbol{f}$ (implicitly, with respect to category c_j). Psychologically, it expresses how well having a feature predicts whether a stimulus belongs to a particular category.

We also look at *maximum category validity*, defined as $\max_{1 \leq j \leq M} p(\boldsymbol{f}|c_j)$. Here $p(\boldsymbol{f}|c_j)$ is often referred to as the category validity $\boldsymbol{f}$ (again, implicitly, with respect to category

c_j). It expresses how well belonging to a category predicts whether a stimulus has a particular feature.

Finally, we look at *maximum collocation*, $\max_{1\leq j\leq M} p(c_j|\boldsymbol{f})p(\boldsymbol{f}|c_j)$. The quantity $p(c_j|\boldsymbol{f})p(\boldsymbol{f}|c_j)$ is known as the collocation of feature $\boldsymbol{f}$ and category c_j. This measure has previously been applied by Jones (1983) in his feature possession score account of category basicness. Here it is applied as a measure that integrates both cue and category validity.

Alternative Measures

We supplement the usefulness heuristics by two additional heuristics, included as baselines. The first of these is based around a measure we term *feature prevalence*, defined to be the proportion of exemplars in a domain which possess a given feature. The purpose of this heuristic is to compare the usefulness heuristics to a simple heuristic using only base-rate information. The second is a "random" heuristic, which simply selects subsets of features at random. This heuristic is intended to illustrate how our usefulness heuristics compare to an arbitrarily chosen heuristic for selecting features.

The remainder of the paper compares the five heuristics using human similarity judgments. We procede as follows. First, we describe the data on which the heuristics will be compared, the Leuven Natural Concept Database (Deyne et al., 2008), a collection of normative data for semantic concepts. We then present the selection heuristics and how the representations chosen are used to generate similarity judgments. Next, we show the results of applying the heuristics to the Leuven database. We close by discussing what these results tell us about the features people choose to represent stimuli and the difference between natural and artificial kinds.

The Leuven Natural Concept Database

The Leuven Natural Concept Database (Deyne et al., 2008) contains normative data for semantic concepts falling into one of two domains, animals and artifacts. These data consist of typicality ratings, goodness ratings, goodness rank orders, generalization frequencies, exemplar associative strengths, category associative strengths, estimated ages of acquisition, word frequencies, familiarity ratings, imageability, and pairwise similarity ratings for concepts within a single category as well as exemplar-by-feature matrices and pairwise similarity ratings between a subset of the exemplars in a domain spread across its categories.

In our comparisons we make use of the exemplar-by-feature matrices and domain similarity ratings. The exemplar-by-feature matrices describe the exemplars of a domain in terms of a number of participant-generated features. For the animals domain, 129 exemplars, split among the categories birds, fish, insects, mammals, and reptiles, are described in terms of 765 features. For the artifacts domain, 166 exemplars, split among the categories clothing, kitchen utensils, musical instruments, tools, vehicles, and weapons, are described in terms of 1295 features. These features include both high frequency features such as "is a bird" and "is made of metal" and low frequency features such as "stands in the crib at Christmas" and "stored in the cellar".

Domain similarity judgments are pair-wise similarity judgments collected between exemplars in a set of consisting five exemplars from each of the categories in a domain. This results in sets of twenty-five exemplars for the animals domain and sets of thirty exemplars for the artifacts domain. Two distinct sets of exemplars were chosen for each domain, resulting four sets of domain similarity judgments.

Feature Selection Measures

Starting with a set of features that we wish to select a feature representation from (such as the 765 animal or 1295 artifact features in the Leuven sets), each heuristic chooses a feature representation using a two step process. First, the usefulness of each feature is computed under a particular usefulness measure. Then, we select those features whose usefulness is above a pre-defined threshold. For example, suppose we wish to use the collocation heuristic to choose among the seven features representing the exemplars of the three categories in Table 1. First, we would compute the maximum collocation over categories for each of the features (shown in the "Colloc." column of Table 1). Then, we would select all those features for which the maximum collocation over the categories was above our threshold. In this example, were the threshold one-half, we would select features 1, 2, and 3. The same procedure can be used with the benchmark importance measure of Zeigenfuse and Lee (2008, 2010) to select a representation.

The features selected by these heuristics to generate similarities according to a common features model (Shepard & Arabie, 1979). Suppose we have a set of features $\{\boldsymbol{f}_1,\ldots,\boldsymbol{f}_K\}$ from which we have selected a set of useful features indexed by $U\subseteq\{1,\ldots,K\}$. The common features model says that similarity between concepts i and j is

$$s_{ij} = c + \sum_{k\in U} w_k f_{ki} f_{kj}, \tag{1}$$

where c is the universal similarity and w_k is the salience of feature f_k.

The remainder of the section is devoted to discussing for the benchmark and other heuristics in greater detail. In the first subsection, we summarize the benchmark measure of importance. In the second, we provide a rationales for each of the three category-based usefulness measures. In the final subsection, we provide rationales for the two baseline heuristics.

Benchmark

The Zeigenfuse and Lee (2008, 2010) method for learning which of a set of features people use to represent stimuli is based upon latent variable selection. In this framework, those features that are included in a concept's representation are termed "important" features. For each feature, they define a variable z_k indicating whether feature $\boldsymbol{f}_k$ is used in similarity

	Category 1					Category 2			Category 3							Cue	Cat.	Colloc.
Feature 1	•	•	•	•	•											1	1	1
Feature 2	•	•	•	•	•	•										5/6	1	5/6
Feature 3	•	•	•	•												1	4/5	4/5
Feature 4			•													1	1/5	1/5
Feature 5	•	•	•	•	•	•		•			•	•	•	•		5/11	1	5/11
Feature 6	•	•	•	•	•				•	•	•	•	•	•	•	5/12	1	5/12
Feature 7							•		•			•				2/3	1/3	4/21

Table 1: Representative features illustrating behavior of the usefulness measures.

judgments. Then, the similarity between concepts i and j is then

$$s_{ij} = c + \sum_{k=1}^{K} z_k w_k f_{ki} f_{kj}. \quad (2)$$

To learn which features are included in the representation, Zeigenfuse and Lee (2008, 2010) develop a Bayesian model and sample from the marginal posterior over the z_k using Markov Chain Monte Carlo (MCMC). In this framework, a feature's importance is the marginal posterior probability the feature is represented. They found that a small number of important features are able to fit similarity almost as well as using all features.

Usefulness Measures

Different measures of usefulness correspond to different assumptions about what aspects of the environment lead a person to represent a particular feature. In the opening example, the small white spot under the dog's eye and its name, "Rover", may be useful for representing the family dog, but are probably not useful for representing dogs generally. This section outlines the psychological theories of feature importance embodied by each of the usefulness heuristics.

Maximum Cue Validity Maximum cue validity measures how concentrated a feature is in a single category. Formally, let r_k be the total number of objects with a particular feature f_k and let n_{jk} be the number of objects with the feature in category c_j. The cue validity of $\boldsymbol{f}_k$ is then $p(c_j|\boldsymbol{f}_k) = n_{jk}/r_k$ and the maximum cue validity is the maximum of n_{jk}/r_k taken over j.

As illustrated by example features Table 1, maximum cue validity is large when most of the exemplars possessing a feature belong to the same category (Features 1 – 4), though this need not be a large number of exemplars (Feature 4). To see why, note that maximum cue validity is large if and only if there exists a category for which n_{jk} is nearly r_k. Since $n_{lk} \leq r_k - n_{jk}$ for $l \neq j$, $r_k - n_{jk}$ must be small and few exemplars with f_k can belong to c_l.

Maximum Category Validity Category validity measures how diffuse a feature is within a particular category. As with maximum cue validity, let n_{jk} be the number of exemplars in category c_j with feature $\boldsymbol{f}_k$, and define a new quantity q_j to be the total number of exemplars belonging to c_j. Then, the category validity of $\boldsymbol{f}_k$ with respect to category c_j is $p(\boldsymbol{f}_k|c_j) = n_{jk}/q_j$ and the maximum category validity is the maximum of n_{jk}/q_j taken over j. Returning to Table 1, we see that features whose category validity is high (Features 1, 2, 5, and 6) are possessed by most of the exemplars in at least one category.

Maximum Collocation Maximum collocation is a measure of how simultaneous concentrated in and diffuse across a category a feature is. Using the terminology of the previous sections, the collocation of a feature $\boldsymbol{f}_k$ with respect to category c_j is $(n_{jk}/r_k)(n_{jk}/q_j)$. Maximum collocation is the maximum of this quantity taken over j.

Features with high collocation are possessed by most exemplars within a category and few outside it, as illustrated by the architypical Feature 1 in Table 1. Alternatively, Features 4 and 6 show why it is necessary for both of these to be true. Those features possessed by only a small fraction of exemplars within a single category will have high cue validity but low category validity (Feature 4). Those features possessed by most exemplars in more than one category will have high category validity but low cue validity (Feature 6).

Alternative Measures

The two baselines used here are intended to show both how well our usefulness heuristics performed against heuristics embodying contrasting assumptions. The first of these is based on the base rate of a feature across stimuli, which we refer to as feature prevalence. For feature $\boldsymbol{f}_k$, the prevalence is $p(\boldsymbol{f}_k) = r_k/K$, where r_k is as defined in the previous section. This shows that the ability of a feature to distinguish among categories does not affect its importance.

The random heuristic provides a different sort of foil for the usefulness heuristics. Many methods other than those included here could be imagined for selecting a sets of features. By selecting features at random, it allows us to compare the predictions of our heuristics to those an arbitrary method of choosing features.

Method Comparison

Here we describe a comparison of maximum cue validity, maximum category validity, and maximum collocation to each other as well as the benchmark and baselines using the Leuven Natural Concept Database (Deyne et al., 2008). In

the first section, we enumerate the procedure used to fit the domain similarity data. In the second, we present the results of this procedure for each of the heuristics.

Procedure

The fit procedures begins with the exemplar-by-feature matrices. Before applying any of the heuristics we filter out all features possessed by zero, one, or all of the 25 or 30 exemplars included in the domain similarity comparisons. Features possessed by one exemplar or fewer will not be used in any similarity comparisons, since $f_{ki} f_{kj} = 0$ for all distinct stimuli i and j. Features possessed by all exemplars will be used in every similarity comparison, so they can be included in the constant term c in Equation (2). Additionally, we find all groups of features possessed by exactly the same set of exemplars, and combine these into a single feature. Suppose $\mathbf{f}_k$ and $\mathbf{f}_l$ are features possessed by exactly the same set of exemplars. Then, $f_{ki} = f_{li}$ for all i and $w_k f_{ki} f_{kj} + w_l f_{li} f_{lj} = (w_k + w_l) f_{ki} f_{kj}$.

After pre-processing, for the benchmark and all of the heuristics except the random heuristic, we compute its corresponding measure using all of the exemplars in the domain, not just those included in the domain similarity judgments. The features are then sorted in order of decreasing value on these measures. Starting with only the top two features, we fit the common features model to the domain similarity judgments using non-negative least squares and compute the correlation between the fitted similarities and the actual similarities. We repeat this process with the top three features, the top four features, etc. To apply this procedure with the maximum collocation heuristic to the features in Table 1, we first compute the values in the collocation column. We then order the features in order of decreasing collocation, which in this case is 1, 2, 3, 5, 6, 4, 7. We first fit the model with features 1 and 2, then 1, 2, and 3, followed by 1, 2, 3, and 5, etc. Finally, for the random heuristic, we generated 100 random feature orders and apply this procedure to each of the orders.

Results

Figure 1 shows the correlation between observed and those fitted using the first x percent of features ordered by either cue validity, category validity, collocation, prevalence, or the benchmark. For example, on the collocation line (shown as a solid line) the correlation at a percentile rank of 20 percent is the correlation between the observed values and those fitted using the first 20 percent of features ordered by collocation. The smaller pane in the lower right-hand corner is a blowup of the lines in rectangular region extending from $0-20$ in percentile rank and from $0.6-1$ in correlation.

The gray shaded area shows 95% confidence intervals for the correlation between the values fitted using first x percent of features chosen by the random heuristic and the observed values. These orders give an estimate of how difficult the similarity data are to fit with a heuristic choosing x percent of the available features. A heuristic whose correlation is above the upper limit of the area fits better 95 percent of heuristics at that percentage of features. Alternatively, a heuristic whose correlation is below the lower limit of the area fits worse than 95 percent of heuristics at that percentage of features.

Regardless of data set, the orders produced by the Zeigenfuse and Lee (2008, 2010) measure is always able to fit the similarities in the top 5 percent of ordering, justifying its use a benchmark. The orders produced by feature prevalence nearly always perform worse than those generated by the other measures, often in the worst 5 percent of all orders. On the whole, cue validity, category validity, and collocation perform middling to well, rarely performing worse than feature prevalence.

For the animals data sets, cue validity outperforms category validity for small numbers of features (less than around 20 percent), category validity outperforms cue validity for larger numbers of features, and collocation is always commensurate to the best of these. For very small (less than around 10 percent) numbers of features, cue validity performs better than the benchmark; however, for larger numbers of features its performance is at best mediocre. After a slow start, category validity performs in the top 5 percent of orderings for larger numbers of features. Collocation always performs near the benchmark and is nearly always in the top 5 percent of orderings.

For the artifacts data sets, cue validity still performs better than category validity for very small (less than 10 percent) numbers of features, after which category validity performs better than cue validity. As with animals, collocation performs near or better than the best of these two measures. Category validity and collocation nearly always perform between the 5th and 95th quantiles of heuristics; however, for larger numbers of features (around 20 percent in the first set and around 40 percent in the second), cue validity performs in the bottom 5 percent of orderings.

Overall, these results suggest that both cue and category validity contain information about a feature's importance. Collocation always performs about the same as the best of cue and category validity, indicating that it tracks the best aspects of the two measures. This suggests that early on collocation is dominated by features with high cue validity, but later it is dominated by category validity.

Discussion

Cue and Category Validity

The major result of the previous section is that both cue and category validity seem to be important to choosing which of a set of features makes a good representation. Murphy (1982) suggests why this may be the case: cue validity cannot pick out basic-level categories because it can only increase for more inclusive categories. Consider the hierarchy of categories *animal*, *bird*, *duck*, in which *bird* is the basic-level category, and suppose we wish to compute the cue validity of the feature "has wings". Let r_{wings} be the number of things with wings and $n_{\text{ducks,wings}}$, $n_{\text{birds,wings}}$, and $n_{\text{animals,wings}}$ be the number of ducks, birds,

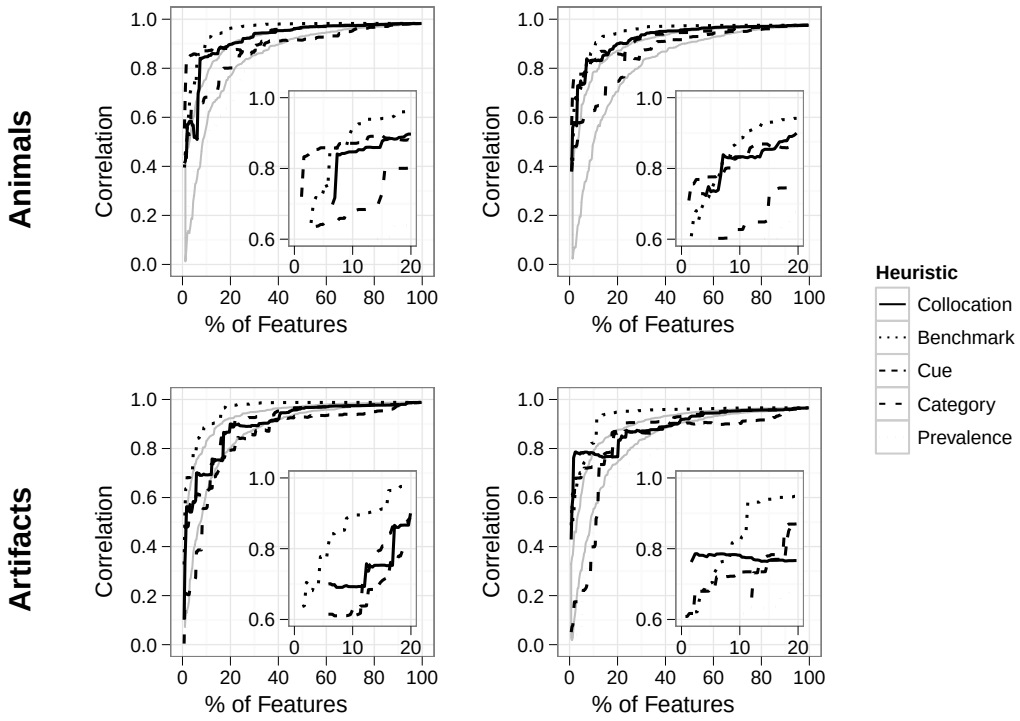


Figure 1: Model fit by the percent of features used for each of the four sets of domain similarities in the Leuven data set. The benchmark, three category-based heuristics, and feature prevalence baseline are shown as lines. In the legend, "collocation" corresponds to the maximum collocation heuristic, "benchmark" to the benchmark, "cue" to maximum cue validity, "category" to maximum category validity, and "prevalence" to feature prevalence. The gray area shows a 95% confidence interval for the fit of the random heuristic. The panels in the lower righthand corner of each of the plots enlarges the rectangular region from $0-20$ in percent of features and from $0.6-1$ in correlation in the main plots.

and animals with wings. Since ducks are birds and birds are animals, $n_{\text{ducks,wings}} \leq n_{\text{birds,wings}} \leq n_{\text{animals,wings}}$, so $n_{\text{ducks,wings}}/r_{\text{wings}} \leq n_{\text{birds,wings}}/r_{\text{wings}} \leq n_{\text{animals,wings}}/r_{\text{wings}}$. But the $n_{\cdot,\text{wings}}/r_{\text{wings}}$ is just the cue validity of "has wings", illustrating why, in settling on basic-level categories, people must be sensitive to more information than just cue validity. Since similarity is assumed to reflect representation, this should be reflected in measures used to select representations.

Along these lines, Tenenbaum and Griffiths (2001) offer a fuller explanation for why both cue and category validities should be important to choosing good representations. They argue that people generalize properties to novel instances only in the smallest set of instances consistent with known examples, a theory known as the "size principle", and further that similarity is the degree to which the consequences of being one object generalize to another. By this logic, choosing features on the basis of cue validity will lead to categories which are overly restrictive and choosing features on the basis of category validity will lead to categories which are overly broad. Appropriate generalization, then, requires taking both types of information into account. Thus, we would expect a heuristic that does this, like collocation, to choose better representations than heuristics that do not.

Natural Versus Artificial Kinds

A final point worth mentioning is the difference in performance of the heuristics on data sets containing natural kinds versus those containing artificial kinds. Numerous authors have suggested that natural and artificial kinds are represented in fundamentally different ways (e.g. Keil, 1989). Results of Zeigenfuse and Lee (2010) support this theory, finding the ratio between the probability two stimuli within the same category have a feature and the probability two arbitrarily chosen stimuli have a feature is larger for natural kinds than artificial ones.

Here we find a similar result: for animals data sets collocation nearly always performs in the top 5 percent of heuristics, whereas for artifacts data sets, collocation performs about as well as an arbitrary heuristic. In theory this difference could come from either differences in the types of features represented or the ability of the common features model to fit similarity judgments among exemplars of that domain. The latter seems unlikely, however, given that the benchmark performs well for all four data sets it seems a common features similarity model is able to fit the data well.

This, then, suggests that the difference in fits comes from differences in the types of features people choose to repre-

sent. Among animals, people prefer features that are closely tied to a particular basic category. Among artifacts, they seem to prefer a different strategy, representing features for multiple levels in a category hierarchy or selecting features using different criteria.

Extensions

A detailed explanation of this difference may requires extensions addressing one of both of these sources. The first of these begins from the recognition that the source of the apparent distinction between natural and artificial kinds may stem not from an actual difference but from an incorrect choice of selection heuristic. Thus, it makes sense to look at heuristics based on additional measures of category differentiation. The second supposes choosing just those features associated with basic-level category structure is not sufficient for selecting good feature representations.

Additional Heuristics In order to explore the first of these extensions, we could develop heuristics based on different measures, both those that have been proposed in the basic-level literature and outside it. Such measures could include the category likelihood ratio (Zeigenfuse & Lee, 2010), the mutual information between a category and a feature SLIP (Gosselin & Schyns, 2001). These last of these differs from the first two in that, in the first, each feature affects the quality of a categorization independent of all other included, whereas in the second two the effect of adding a new feature depends upon the features already included.

Category Hierarchies The second extension allows the method to deal with category hierarchies. The importance of structured representation in understanding human judgments of similarity has been illustrated by many authors (e.g. Markman & Gentner, 1993). Understanding how such structured representations influence those features represented is a crucial step towards bringing these models into contact with feature-based models such as Tversky's contrast model (Tversky, 1977). One potential method for acheiving this would be to compute the collocation, or other measure, at each level in a category hierarchy and to use a weighted combination of the collocations as the selection criterion.

Conclusion

In this paper, we have presented three heuristic methods for choosing a feature representation based on measures of category differentiation. We find these heuristics to fit human data better than heuristics that do not take this information into accounts, acheiving very good fits for natural kinds and above average fits for artificial kinds. Moreover, our results suggest both how concentrated in a particular category a feature is and how diffuse it is across exemplars in that category are important factors in whether a feature is represented as well as supporting a distinction between natural and artificial kinds. Though much still needs to be done, this work suggests people choose features in a systematic way and that these regularities can be uncovered by investigating the relationship between categories and features.

References

Corter, J. E., & Gluck, M. A. (1992). Examining basic categories: Feature predictability and information. *Pyschological Bulletin*, *111*, 291-303.

Deyne, S. D., Verheyen, S., Ameel, E., Vanpaemel, W., Dry, M., & Voorspoels, W. (2008). Exemplar by feature applicability matrices and other Dutch normative data for semantic concepts. *Behavior Research Methods*, *40*(4), 1030-1048.

Gosselin, F., & Schyns, P. G. (2001). Why do we SLIP to the basic level? Computational constraints and their implementation. *Psychological Review*, *108*(4), 735-758.

Jones, G. V. (1983). Identifying basic categories. *Psychological Bulletin*, *94*, 423-428.

Keil, F. (1989). *Concepts, kinds, and cognitive development*. Cambridge, MA: MIT Press.

Kloos, H., & Sloutsky, V. M. (2006). What's behind different kinds of kinds: Effects of statistical density on learning and representation of categories. *Journal of Experimental Psychology: General*, *137*(1), 52-72.

Markman, A. B., & Gentner, D. (1993). Structural alignment during similarity comparisons. *Cognitive Psychology*, *25*, 431-467.

Marr, D. (1982). *Vision: A computational approach*. San Francisco: Freeman & Company.

Murphy, G. L. (1982). Cue validity and levels of categorization. *Psychological Bulletin*, *91*(1), 174-177.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, *7*, 573-605.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*, 352-382.

Shepard, R. N., & Arabie, P. (1979). Additive clustering: Representations of similarities as combinations of discrete overlapping properties. *Psychological Review*, *86*(2), 87-123.

Shepard, R. N., & Chipman, S. (1970). Second-order isomorphism of internal representations: Shapes of states. *Cognitive Psychology*, *1*, 1-17.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, *24*(4), 629-640.

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*, 327-352.

Zeigenfuse, M. D., & Lee, M. D. (2008). Finding feature representations of stimuli: Combining feature generation and similarity judgment tasks. In V. Sloutsky, B. Love, & K. McRae (Eds.), *Proceedings of the 30th annual conference of the cognitive science society* (p. 1825-1830). Austin, TX: Cognitive Science Society.

Zeigenfuse, M. D., & Lee, M. D. (2010). Finding the features that represent stimuli. *Acta Psychologica*, *133*, 283-295.

When Two plus Two Does Not Equal Four: Event-Related Potential Responses to Semantically Incongruous Arithmetic Word Problems

Kristie J. Fisher (kjfisher@uw.edu)
University of Washington, Department of Psychology, Box 351525
Seattle, WA 98105 USA

Miriam Bassok (mbassok@uw.edu)
University of Washington, Department of Psychology, Box 351525
Seattle, WA 98105 USA

Lee Osterhout (losterho@uw.edu)
University of Washington, Department of Psychology, Box 351525
Seattle, WA 98105 USA

Abstract

Extensive research measuring event-related brain potentials (ERPs) shows that semantic incongruity is indexed by the N400 effect and syntactic/structural incongruity is indexed by the P600 effect. We used these indices to examine how people coordinate their semantic and arithmetic knowledge when they read simple addition and division word problem sentences (e.g., "Twelve roses plus three daisies equals fifteen"). Prior work in problem solving has shown that word-problem solutions are modulated by analogical alignment of semantic and arithmetic relations, such that people avoid or commit errors on misaligned problems (e.g., Aligned: "Twelve roses plus three daisies equals fifteen"; Misaligned: "Twelve cookies plus three jars equals fifteen"). Here, we found that such analogical alignments modulate the comprehension of word-problem sentences. Specifically, we found that analogically Misaligned semantic relations elicited a P600 effect. Furthermore, an N400 effect was elicited by the last number word of Misaligned problem sentences, even when it was a mathematically correct answer. These results show that analogical alignment between semantic and arithmetic relations can be indexed with the P600 effect and provide a foundation for future ERP work on analogical reasoning.

Keywords: ERP; analogy; mathematical cognition; N400 effect; P600 effect

Introduction

A common task facing the cognitive system is *conceptual integration* of individual items into a meaningful whole. For example, language comprehension requires conceptual integration of consecutive words into meaningful sentences. Similarly, comprehension of arithmetic problems requires conceptual integration of numbers and arithmetic operators into correct mathematical expressions. In this paper, we explore the conceptual integration of simple arithmetic word problems, which are unique in that they require conceptual integration of language and of mathematics.

Conceptual Integration & ERPs

The process of conceptual integration, and the conditions under which it can be disrupted, have been investigated in a variety of domains using event-related potential (ERP) methodology, which measures the electrical brain activity elicited by a particular stimulus. Work in this area has shown that two key aspects of conceptual integration, meaning and structure, are indexed by two distinct and highly reliable ERP components—the N400 and P600 components, respectively.

The N400 component is negative-going and peaks around 400ms after presentation of the stimulus. This component is highly sensitive to contextual semantic meaning. The magnitude of this component is larger for semantically incongruous compared to congruous items—a difference known as the *N400 effect*. The N400 effect was first documented in sentence processing. For example, the italicized word in the sentence, "The cat will *bake* the food" will elicit an N400 effect relative to, "The cat will *eat* the food" (Kutas & Hillyard, 1980). Subsequent work has shown that the N400 effect is elicited in response to conceptual incongruities in other domains. For example, incorrect answers to simple symbolic (e.g., "4 x 4 = *21*") and verbal (e.g., "Twelve plus three equals *sixteen*.") arithmetic problems elicit an N400 effect (e.g., Niedeggen & Rosler, 1999; Fisher, Bassok, & Osterhout, 2009). Thus, the N400 effect is generally accepted to be a domain-general index of semantic congruence.

The P600 component is positive and peaks at around 600ms after stimulus presentation. A P600 effect is elicited by violations of syntax within a sentence (e.g., "The cat will *eating* the food I leave on the porch."; Osterhout & Holcomb, 1992; Osterhout & Mobley, 1995) and by violations of structure, such as a wrong note played in a harmonic scale (Patel et al., 1998). Such violations of syntax or structure lead to larger P600 amplitudes, relative to control conditions (i.e., the P600 effect).

Furthermore, Osterhout and Mobley (1995) found that when there is any kind of violation within a sentence, syntactical or semantic, an N400 effect is also elicited by the ending word of the sentence, even though that word is perfectly correct. This last-item N400 effect is likely the result of the experimental paradigm typically used in

language research. Participants are typically asked to make binary judgments about the "acceptability" of the sentences they just saw (usually they are not instructed to look for any particular type of error). Thus, when participants reach the end of a sentence that contained a violation, the entire sentence must now be categorized as "unacceptable." The N400 effect to the final word in the sentence may be a result of this judgment processing.

Despite the relatively broad range of studies of conceptual integration, to our knowledge, no previous studies have used these ERP indices to examine a) how people integrate concepts that are presumably organized in distinct conceptual networks and b) the integration of concepts via analogy. Both of these characterize the integration process involved in the solution of mathematical word problems, whereby people are required to apply arithmetic operations in a way that fits the relations among objects in the "real world." This process involves analogical coordination of real world knowledge (e.g., *roses* and *daisies* are *flowers*) with one's knowledge of arithmetic properties (One can add 3 *roses* to 5 *daisies* to create a bouquet of 8 *flowers*). As we explain in the next section, people are highly systematic in the way they coordinate their semantic and arithmetic knowledge. The purpose of our study was to use ERP to examine such cross-network, analogical conceptual integration and, in particular, to test whether the same ERP components that index violations of meaning and structure in language also index violations of analogical alignment.

Mathematical Problem Solving in the "Real World"

Research by Bassok and her colleagues has shown that, when people reason about mathematical word problems, they tend to align structurally analogous semantic and arithmetic relations (Bassok, Chase, & Martin, 1998). Specifically, people align categorically related objects (e.g., cars and trucks) with the commutative addition operation and align functionally related objects (e.g., jars and cookies,) with the non-commutative division operation. Violating such *semantic alignment* (e.g., having to add jars to cookies or having to divide cars by trucks) severely impairs problem-solving performance (Bassok, Wu, & Olseth, 1995; Martin & Bassok, 2005), and even blocks retrieval of arithmetic facts from memory (Bassok, Pedigo, & Oskarsson, 2008).

In the present study we investigated how people conceptually integrate semantic and arithmetic relations while reading simple addition and division word problems presented in a sentence format (e.g., "Twelve roses plus three daisies equals fifteen."). We recorded ERPs as participants read these word-problem sentences. We analyzed the electrical waveforms elicited by the second object word in the sentence, which completed the semantic relation, and by the numerical mathematical answers (e.g., the two underlined words in, "Twelve roses plus three vases equals fifteen"). The semantic object relations were either analogically aligned (Aligned condition) or misaligned (Misaligned condition) with the arithmetic relation in the word problem, and the mathematical answers were either correct or incorrect (see Table 1 for example stimuli). After reading each word-problem sentence, participants were asked to make judgments as to whether or not the problem was "acceptable." As is standard practice in typical language research paradigms, we did not specify the criteria by which participants were to make their judgments.

Table 1: Example Stimuli

Object Alignment	Math Correct	Math Incorrect
Aligned Addition	Twelve limes plus three lemons equals fifteen.	Sixteen cars plus two trucks equals twenty.
Aligned Division	Fifteen roses divided by three bouquets equals five.	Six robins divided by two nests equals eight.
Misaligned Addition	Six questions plus three quizzes equals nine.	Eight cookies plus four jars equals two.
Misaligned Division	Eighteen skirts divided by two dresses equals nine.	Fifteen geese divided by three ducks equals six.

We had two main predictions. First, we expected that conceptual integration in Aligned word problems should be similar to conceptual integration in arithmetic problems, presented in sentence-form, which do not contain objects. Specifically, we expected that, in the Aligned condition, we would replicate the N400 effect elicited by mathematically incorrect answers to arithmetic problems (Fisher, Bassok, & Osterhout, 2009). Second, and most important, if conceptual integration via analogy is similar to conceptual integration in rule-governed sequence processing, then analogical misalignment of the semantic and arithmetic relations in the problem should elicit a P600 effect. In particular, we expected a P600 effect to occur at the second object word because that word completes a semantic relation that cannot be mapped onto the arithmetic relation in the problem, and thus constitutes a structural violation (e.g., Gentner, 1983). Furthermore, we expected that the mathematically correct answer (the final item of the word problem sentence) in Misaligned problems would elicit an N400 effect relative to correct answers in Aligned problems, replicating previous work by Osterhout and Mobley (1995).

Methods

Participants

The participants were 38 volunteer undergraduate students, graduate students, and staff from the University of

Washington (21 male, 17 female; M_{age} = 22.23 years, SD_{age} = 4.98 years) who were right-handed native English speakers. Participants were either given course extra credit or paid $30 for their participation.

Stimuli

The stimuli were simple word problem sentences that were composed of digit pairs and object word pairs that were either categorically related or functionally related. The digit and object pairs were selected based on pilot testing, as described below.

Arithmetic Problems The arithmetic problems were composed of two operands and satisfied a number of constraints established by cognitive arithmetic literature and required for our experimental manipulations. First, the two operands could be both added and divided to yield a whole-number answer (e.g., *12 + 3*; *12 / 3*). Second, we excluded tie problems (e.g., *2 + 2*) and problems containing a one, zero, or 10 as an operand, as evidence from prior work suggests that these types of problems are processed differently, and often more easily, than other simple arithmetic problems (Ashcraft, 1992; McCloskey, 1992). Third, we only selected problems that fell into the "small" category of division problems, defined as having a divisor lesser than 25, in order to avoid some of the issues of the problem-size effect[1] (see Zbrodoff & Logan, 2004, for a review). Finally, we controlled for answer parity (LeMaire & Reder, 1999).

Within these constraints, we created a set of 24 problems, 12 addition and 12 division, that were equivalent in difficulty. These problems were selected based on results of a pilot study (error rate and response time), in which 154 undergraduate students solved 48 addition and 48 division problems meeting the above criteria. To create an answer verification task, we constructed two different incorrect answers for each problem. The "Close" incorrect answer for both operations was derived by adding or subtracting the value one or two to or from the correct answer (e.g., *12 + 3 = 14*). The "Other" incorrect answers for addition were the correct answers to division problems with the same operands (e.g., *12 + 3 = 4*), and the "Other" incorrect answers for division were the correct answers to addition problems with the same operands (e.g., *12 / 3 = 15*)

Object Pairs We initially constructed a set of 163 word pairs that we considered to belong to one of the two semantic relations categories—categorical or functional. The set contained 83 possible categorical pairs and 80 possible functional pairs consisting of concrete, plural nouns (e.g., "cats, dogs"). From this set, we constructed rating surveys that were completed by 202 undergraduate students at the University of Washington as part of a class activity. Instructions asked students to rate, on a seven-point scale, either the extent to which the word pairs were *categorically* related or the extent to which they were *functionally* related. The average categorical and functional ratings in these two conditions were compared for each word pair using an independent t-test with an alpha level of .05. In order to be included in the final set, word pairs had to have significantly different categorical and functional ratings and an average rating of greater than 5 in one dimension and 4 or less in the other. Based on these ratings, we selected 48 categorical and 48 functional pairs. The word pairs in both relation conditions were equivalent in their average number of syllables and letters in each word.

Design

Operation (Addition vs. Division) was manipulated between participants ($N_{Addition}$ = 19; $N_{Division}$ = 19; participants were randomly assigned). Analogical alignment of the mathematical operation and the object sets (Aligned vs. Misaligned), and mathematical correctness of the problems (Correct vs. Close Incorrect vs. Other Incorrect) were manipulated within participants.

Verbal versions of the arithmetic problems (e.g., Twelve plus three" in place of "12 + 3") were created and were then combined with object pairs to create simple word problem sentences (e.g., "Twelve limes plus three lemons equals fifteen."). For the Addition problems, all of the Aligned stimuli were categorically related objects, and all of the Misaligned stimuli were functionally related objects; the reverse was true for the Division problems (see again Table 1). Thus, the same object sets were used for both operations, but for one operation the object sets were Aligned and for another they were Misaligned.

The experiment consisted of three blocks of trials. There were 96 trials in each block, for a total of 288 trials. Within each block, 50% of the trials were Aligned word problems, and 50% were Misaligned. Within each alignment type, 50% were mathematically Correct, 25% were Close Incorrect, and 25% were Other Incorrect. Trial order was pseudo-randomized within each of the three blocks. Each of the word pairs appeared once per block, and they were combined with different arithmetic problems each time.

Procedure

Participants were seated comfortably in front of a CRT monitor in an isolated room and fitted with electro-encephalography (EEG) recording equipment. Each trial consisted of a fixation point (500ms), and each item of the word-problem sentence was presented alone on a screen (450ms/350ms ISI). The final inter-stimulus interval before the appearance of the YES/NO response screen was 1,000 ms (total trial duration was 7.1 seconds). Participants were given a hand-held controller and were asked to respond *YES* (response hand counter-balanced) using one button if they thought the problem was completely "acceptable" and *NO*, using another button if the problem was "unacceptable" in

[1] Note, however, that "small" division problems translate into "large" addition problems. As described in this section, the stimuli selection pilot study was conducted primarily to ensure that the problems selected were of equivalent difficulty.

any way. They were told that the instructions were intentionally vague because the criteria by which they would judge the problems were at their discretion. Furthermore, the task did not include object labels, which are usually required in word problem solving. Participants were asked not to blink between the onset of the fixation point and the appearance of the response screen. They were permitted to blink and take a short break while the response screen was displayed. Response time was not recorded and responses triggered onset of the next trial. A break was given after each block. The entire experiment time, including set-up, was less than two hours.

Data Acquisition & Results

EEG recording

Continuous EEG was recorded from 19 tin electrodes attached to an elastic cap (Eletro-cap Inernational) in accordance with the extended 10-20 system. Vertical eye movements and blinks were monitored by two electrodes, one placed beneath the left eye and one placed to the right of the right eye. The 19 electrodes were referenced to an electrode placed over the left mastoid. Electrical signals were amplified, digitized at a rate of 250Hz, and bandpass filtered at 0.01-40Hz. Impedances at scalp and mastoid electrodes were held below 5 kΩ. Trials associated with blinking, excessive eye movement or amplifier blocking were removed prior to averaging (approximately 11% of all trials). Stimuli were displayed to participants on an 18" CRT monitor approximately three feet from the participants at eye-level with white font on a black background.

Behavioral Responses

Because participants were asked to make open-ended "acceptability" judgments, it is not surprising that there was variation in how they judged the Misaligned problems, particularly in the case where the problem was Misaligned but mathematically correct. These behavioral differences in acceptability judgments corresponded to differences in the magnitude of the overall ERP effects we report here. In this paper we do not discuss these individual differences, as they are not essential to our primary research question.

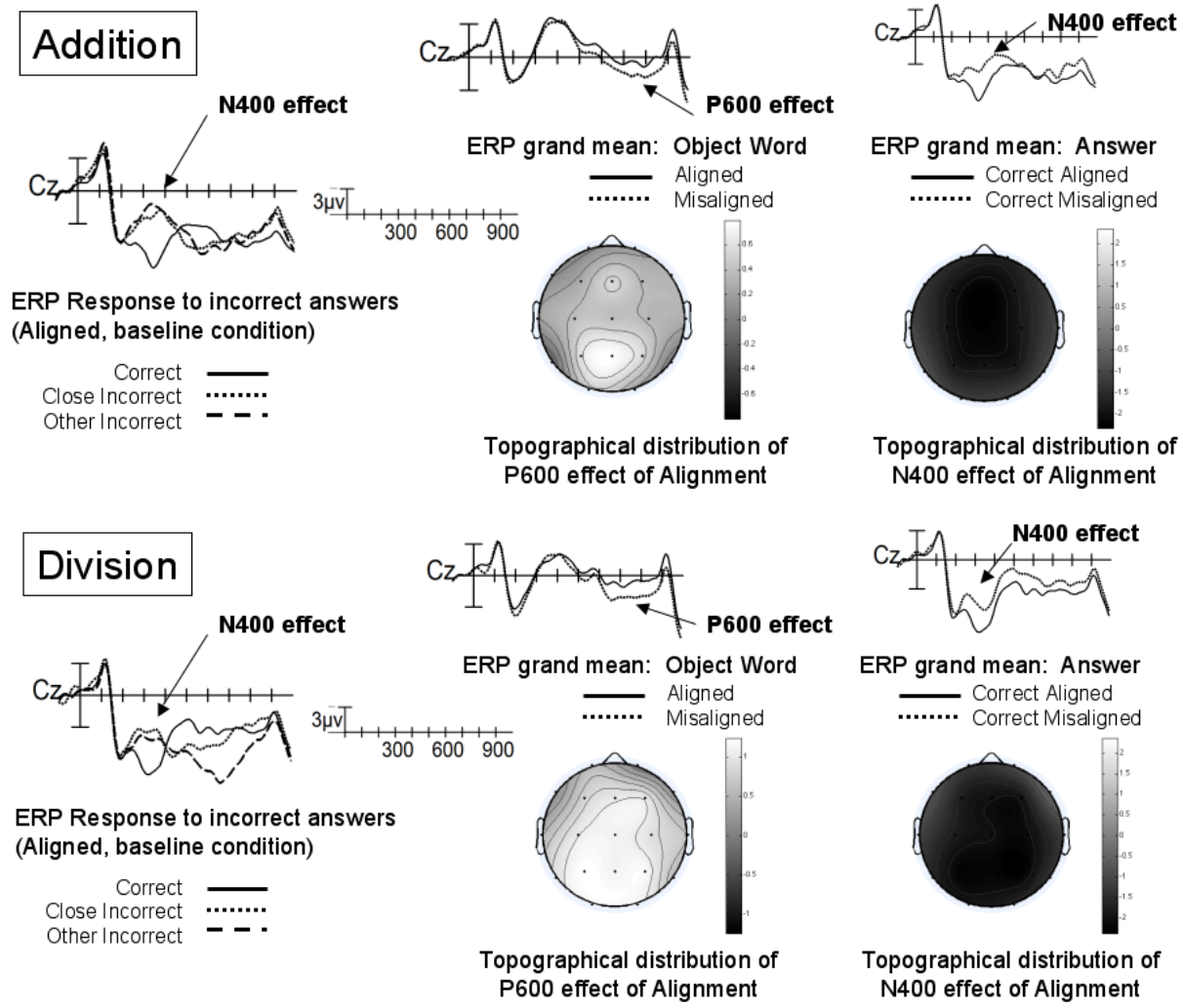


Figure 1: ERP responses to mathematically incorrect answers and semantic alignment.

ERP Responses

EEG amplitudes elicited by the second object word and by the last word (the mathematical answer to the problem) were averaged according to their respective Alignment and Answer conditions. Mean amplitudes were compared separately for the Addition and Division operations in the 250-450ms time window (N400 effect) and in the 500-700ms time window (P600 effect) following stimulus onset. For all analyses, separate ANOVAs[2] were conducted for midline (Fz, Cz, Pz), medial (Fp1, F3, C3, P3, O1, Fp2, F4, C4, P4, O2), and lateral (F7, T7, P7, F8, T8, P8) electrode sites, with electrode location and hemisphere included as factors in each ANOVA.

Incorrect Answers We first tested whether we replicated previous work with sentence-form arithmetic problems (Fisher, Bassok, & Osterhout, 2009), which found an N400 effect to incorrect numerical answers. We tested for this effect within the Aligned condition, which served as our baseline. Separately for Addition and Division, a 3-way (Answer Type - Correct, Close Incorrect, Other Incorrect) repeated measures ANOVA was conducted at each electrode grouping. Indeed, we found a main effect of answer types, such that mathematically incorrect answers elicited a significant N400 effect relative to correct answers for both Addition and Division word problems (see Figure 1) [Addition: $F_{\text{Midline}}(2,36) = 15.36$, $MSE = 11.99$, $p < .001$; $F_{\text{Medial}}(2,36) = 25.54$, $MSE = 17.71$, $p < .001$; $F_{\text{Lateral}}(2,36) = 22.29$, $MSE = 5.86$, $p < .001$; Division: = $F_{\text{Midline}}(2,36) = 6.08$, $MSE = 8.10$, $p < .01$; $F_{\text{Medial}}(2,36) = 10.37$, $MSE = 13.69$, $p < .001$; $F_{\text{Lateral}}(2,36) = 7.24$, $MSE = 6.23$, $p < .01$] Planned contrasts revealed no significant differences between the Close and Other incorrect answer types except at the lateral electrode sites for Addition[3].

Semantic Alignment When comparing ERP responses *between* semantic alignment conditions, we first compared the ERP waveforms elicited by the second object word in Aligned versus Misaligned problems (e.g., Twelve cars plus three *trucks* equals fifteen. vs. Twelve roses plus three *vases* equals fifteen.). Consistent with our predictions, we found that the second object word in the Misaligned condition elicited a P600 effect relative to the Aligned condition in both operations [Addition: $F_{\text{Midline}}(1, 18) = 5.28$, $MSE = 4.08$, $p = .03$; $F_{\text{Medial}}(1, 18) = 6.39$, $MSE = 6.66$, $p = .02$; $F_{\text{Lateral}}(1, 18) = 5.78$, $MSE = 1.68$, $p = .03$; Division: $F_{\text{Midline}}(1, 18) = 4.94$, $MSE = 4.25$, $p = .04$; $F_{\text{Medial}}(1,18) = 5.62$, $MSE = 7.16$, $p = .03$; $F_{\text{Lateral}}(1,18) = 3.20$, $MSE = 1.95$, $p = .09$]. As noted earlier, this effect occurred regardless of participants' behavioral response as to whether or not the problem was "acceptable."

Next, we compared ERP amplitudes elicited by the mathematically correct answer (the final item of the word problem sentence) between the Aligned and Misaligned conditions. Correct answers of Misaligned word problems elicited an N400 effect relative to the correct answers of Aligned word problems [Addition: $F_{\text{Midline}}(1,18) = 9.00$, $MSE = 13.69$, $p < .01$; $F_{\text{Medial}}(1,18) = 11.16$, $MSE = 23.81$, $p < .01$; $F_{\text{Lateral}}(1,18) = 8.62$, $MSE = 6.67$, $p < .01$; Division: $F_{\text{Midline}}(1,18) = 8.01$, $MSE = 15.67$, $p = .01$; $F_{\text{Medial}}(1,18) = 9.34$, $MSE = 32.98$, $p < .01$; $F_{\text{Lateral}}(1,18) = 7.26$, $MSE = 8.01$, $p = .02$].

This pattern of ERP results mirrors those found in studies of language processing (e.g., Osterhout & Mobley, 1995). That is, structural/syntactic violations within a sentence typically elicit a P600 effect, and the final word of sentences containing such violations elicits an N400 effect, even when those words contained no violations. In the case of our particular stimuli, the sentences were simple arithmetic word problems, and the structural violations were violations of analogical alignment between the semantic and arithmetic relations in the problem. The final words in the sentences were the mathematical answers to the word problems, and an N400 effect occurred for mathematically correct answers in the Misaligned, relative to the Aligned, condition.

Discussion

The goal of this study was to examine the conceptual integration process with respect to arithmetic word problems and how it compares to conceptual integration for sentences and other meaningful sequences. Arithmetic word problems are unique in that they combine elements of language and math and provide the opportunity for analogical alignment or misalignment between the semantic relations and the arithmetic relations in the problem (e.g., Bassok, Pedigo, & Oskarsson, 2008; Bassok, Wu, & Olseth, 1995).

Overall, our results provide evidence for the fluid integration of arithmetic and semantic knowledge during word problem processing. More broadly, our results suggest that the conceptual integration process does not change significantly when people must integrate concepts *across* two distinct knowledge networks. That is, the same ERP effects were elicited by violations of structure and meaning in word problems as are usually found in sentences and in arithmetic problems not containing objects. These results suggest that the N400 and P600 effects could be used as dependent measures in investigations of other situations wherein an individual has to integrate distinct types of knowledge, such as in reasoning problems that involve the applications of formal logic rules to object sets, or in song writing wherein one has to coordinate lyrics with a melody.

Moreover, in the word problems used in our study, semantic and arithmetic knowledge had to be coordinated via analogy. Thus, our results also demonstrate that the

[2] A Greenhouse- Geisser correction for sphericity violations was used when necessary

[3] Specific results for different answer conditions and the interactions between the semantic alignment variable, the answer type variable, and behavioral response pattern are not central to the research question addressed here and thus are not elaborated upon for the sake of brevity.

P600 effect can serve as an index of the integrity of analogical structure within arithmetic word problems just as it indexes syntactic integrity in sentences (e.g., Osterhout & Holcomb, 1992; Osterhout & Mobley, 1995) and structural integrity in other meaningful sequences (e.g., Patel et al., 1998). As such, our results provide a foundation for future ERP investigations of the cognitive processes related to analogical reasoning, using the P600 effect as an index of structure-mapping (Gentner, 1983). That is, when the structures of two relations cannot be mapped in an analogy task (e.g., Bird:Nest as Bear:Cave vs. Bear:Desert; Spellman, Holyoak, & Morrison, 2001), the size of the P600 effect could be used to discriminate the degree of relational structure mismatch. Such investigations could examine the analogical conceptual integration within one domain (e.g., animals and their habitats) or across two domains of conceptual knowledge (e.g., Bird:Nest as Car:Garage).

Interestingly, we found these ERP effects for violations of analogical alignment and mathematical correctness across all participants even though we observed distinctly different patterns of "acceptability" judgments and corresponding ERP effect magnitude within our sample. Though we are unable to elaborate on these differences here, initial analyses suggest that these patterns are consistent with prior work in mathematical reasoning suggesting that some people are better than others at coordinating their mathematical and "real world" knowledge when constructing and solving more complex mathematical expressions than the ones presented in this study (e.g., algebraic equations; Fisher & Bassok, 2009). Because this ability is arguably relevant to our simpler task, it is not surprising that there are individual differences among our sample of participants such that some were more sensitive than others to violations of semantic alignment in simple word problems, particularly because our sentences did not include labels as part of the solution.

Of course, further work is required to fully explore these individual difference patterns and elucidate the reason behind them. To expand on our current findings, we also plan to more thoroughly investigate the processes of analogical conceptual integration. Lastly, in the future we hope other researchers will continue to use ERP for investigations of conceptual integration in more complex, knowledge-diverse situations.

Acknowledgments

This research was partially funded by the University of Washington's Royalty Research Fund, with a grant awarded to Miriam Bassok and NIDCD Research Grant R01DC01947 awarded to Lee Osterhout. Thanks to the members of the Cognitive Neuroscience of Language Lab for help with data collection and theoretical insights and to Melody Sherry and Louis Wei for help with pilot data collection and analysis.

References

Ashcraft, M.H. (1992). Cognitive arithmetic: A review of data and theory. *Cognition*, *44*, 75-106.

Bassok, M., Chase, V., & Martin, S. (1998). Adding apples and oranges: Alignment of semantic and formal knowledge. *Cognitive Psychology, 35*, 99-134.

Bassok, M., Pedigo, S. F., & Oskarsson, A. (2008) Priming addition facts with semantic relations. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*, 343-352.

Bassok, M., Wu, L.-L., & Olseth, K. L. (1995). Judging a book by its cover: Interpretative effects of content on problem-solving transfer. *Memory & Cognition, 23,* 354-367.

Fisher, K.J., & Bassok, M. (2009) Analogical alignments in algebraic modeling. In B. Kokinov, K. Holyoak, & D. Gentner (Eds.), *Proceedings of for the 2nd International Conference on Analogy in Sofia, Bulgaria*, pp. 137-144.

Fisher, K. J., & Bassok, M., & Osterhout, L (2009). Conceptual integration is the same for digits and for words: It's the meaning, stupid! In N. A. Taatgen, H. Van Rijn, L. B. J. Schomaker, & J. Nerbonne (Eds.), *Proceedings of the 31st Annual conference of the Cognitive Science Society* (pp. 2142-2147).

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science, 7*, 155-170.

Kutas, M., & Hillyard, S. A. (1980). Reading Senseless Sentences: Brain Potentials Reflect Semantic Incongruity. *Science, 207*(4427), 203-205.

LeMaire, P. & Reder, L. (1999). What affects strategy selection in arithmetic? The example of parity and five effects on product verification. *Memory & Cognition*, *27*, 364-382.

Martin, S. & Bassok, M. (2005). Effects of semantic cues on mathematical modeling: Evidence from word problem solving and equation construction. *Memory & Cognition, 33,* 471-478.

McCloskey, M. (1992). Cognitive mechanisms in numerical processing: Evidence from acquired dyscalculia. *Cognition*, *44*, 107-157.

Niedeggen, M., & Rösler, F. (1999). N400 effects reflect activation spread during retrieval of arithmetic facts. *Psychological Science, 10*, 271-276.

Osterhout, L., & Holcomb, P. J. (1992). Event-related brain potentials elicited by syntactic anomaly. *Journal of Memory and Language, 31*, 785-806.

Osterhout, L., & Mobley, L. A. (1995). Event-related brain potentials elicited by failure to agree. *Journal of Memory and Language, 34*, 739-773.

Patel, A. D., Gibson, E., Ratner, J., Besson, M., & Holcomb, P. J. (1998). Processing syntactic relations in language and music: An event-related potential study. *Journal of Cognitive Neuroscience, 10*, 717-733.

Spellman, B. A., Holyoak, K. J., & Morrison, R. (2001). Analogical priming via semantic relations. *Memory & Cognition, 29*, 383-393.

Zbrodoff, N. J., & Logan, G. D.(2005).What everyone finds: The problem size effect. In J. I. D. Campbell (Ed.), Handbook of mathematical cognition (pp. 331-346). New York: Psychology Press.

Simplifying the Mapping from Referring Expression to Referent in a Conceptual Semantics of Reference

Jerry Ball (Jerry.Ball@mesa.afmc.af.mil)
Air Force Research Laboratory

Abstract

In Jackendoff's Conceptual Semantics, reference to objects, situations, places, directions, times, manners, and measures is supported, but reference is limited to instances of these conceptual categories. This paper proposes an extension of Jackendoff's referential types along an orthogonal dimension of reference which is cognitively motivated in suggesting the possibility of referring to types, prototypes and exemplars in addition to instances, as well as classes and collections of all referential types and vacuous instances and collections. The paper also introduces a bi-partite distinction between a situation model and the mental universe which helps to explain apparent non-referential uses of referring expressions. The primary motivation for expanding the ontology of referential types and distinguishing the situation model from the mental universe is to simplify the mapping from linguistic expressions to corresponding representations of referential meaning. The viability of this approach hinges on adoption of the mentalist semantics of Jackendoff. There is no direct reference to actual objects in the external world.

Keywords: referring expression; Conceptual Semantics

Introduction

In Jackendoff's *Conceptual Semantics* (Jackendoff, 1983, 1990, 2002, 2007), reference to places, directions, times, manners, and measures in addition to situations and objects is supported, but reference is limited to tokens or instances of these conceptual categories, adhering to the basic notion that reference is to individuals. This paper proposes an extension of Jackendoff's referential types along an orthogonal dimension of reference which is cognitively motivated in suggesting the possibility of referring to types, prototypes and exemplars in addition to instances. Reference to classes and collections of referential types and vacuous instances and collections is also considered.

The primary motivation for expanding the ontology of referential types is to simplify the mapping from referring expressions to corresponding representations of referential meaning. Hobbs (2003) pursues a similar strategy in arguing for logical representations that are as close to English as possible. Jackendoff's (1983, p. 13-14) *grammatical constraint* makes a related claim:

> …one should prefer a semantic theory that explains otherwise arbitrary generalizations about the syntax and the lexicon…a theory's deviations from efficient encoding must be vigorously justified, for *what appears to be an irregular relationship between syntax and semantics may turn out merely to be a bad theory of one or the other* (italics added)

Taking the grammatical constraint seriously, we assume that if a linguistic expression has the grammatical form of a referring expression, then it is a referring expression. For example, a nominal like "a man" which contains the referential marker "a", indicates that the expression can be used to refer. Unless there is a very strong reason to assume that any use of this referring expression is non-referential, it is assumed to refer. Further, the referential marker "a" indicates reference to a single referent as does the head noun "man" (i.e. both are grammatically singular). This expression cannot be used to refer to multiple individuals.

Where other approaches argue for the non-referential use of referring expressions or for a complicated mapping from referring expression to possible referents (see discussion below), it is argued instead that referring expressions may refer to something other than an individual, and that the notion of reference is complicated by a secondary relationship between the referents in a situation model and objects in the mental universe. By expanding the ontology of referential types to include types, prototypes and exemplars, and classes and collections of these, it is possible to retain a simplified mapping from referring expression to referent—one which is consistent with the grammatical features of the referring expression. By introducing a bi-partite relationship between a situation model and the mental universe it is possible to explain apparent non-referential uses of referring expressions. The viability of this approach hinges on adoption of the mentalist semantics of Jackendoff. Reference is to mental encodings of external experience and these encodings can provide alternative construals of reality. There is no direct reference to actual objects in the external world.

Theoretical Background

Ball (2007) presents a linguistic theory of the grammatical encoding of referential and relational meaning which is implemented in a computational cognitive model of language comprehension (Ball, Heiberg & Silber, 2007; Ball et al., 2010) within the ACT-R cognitive architecture (Anderson, 2007). The basic structure and function of nominals and clauses is bi-polar with a *specifier* functioning as the locus of the *referential pole* and a *head* functioning as the locus of the *relational pole*—where relational pole encompasses objects (noun, proper noun, pronoun) and relations (verb, adjective, preposition, adverb). If the head of the relational pole is a relation, one or more *complements* or *arguments* may be associated with the relation. *Modifiers* may surround the specifier and head and may be

preferentially attracted to one pole or the other. A specifier and head (or reference point, specifier and head) combine to form a *referring expression*. A determiner functioning as an *object specifier* combines with a head to form an *object referring expression* or *nominal* (ORE → Obj-Spec Obj-Head). A possessive nominal (e.g. "John's" in "John's book") or possessive pronoun (e.g. "his" in "his book") functioning as a combined reference point and specifier may also combine with a head to form an object referring expression (ORE → Ref-Pt/Obj-Spec Obj-Head). In this case the object referring expression contains two referring expressions: 1) the reference point, and 2) the referring expression as a whole.

Ball (2010) extends the theory of referential and relational meaning to a consideration of grammatical features like definiteness, number, animacy, gender and case in object referring expressions. These features provide important grammatical cues for determining the referents of object referring expressions.

The referring expressions in a text instantiate and refer to objects, situations, locations, etc. in a s*ituation model* which is a representation of the evolving meaning of the text. The term "situation model" originates in the research of van Dijk & Kintsch (1983). Originally a situation model was viewed as a collection of propositions extracted from a text and elaborated with additional propositions introduced by schemas activated by the text and resulting from inference processes operating over the text. However, situation models have evolved away from being purely propositional (or relational) representations towards encoding referential, spatial, imaginal and even motor aspects of meaning (cf. Zwann and Radvansky 1998). We view the situation model as the cognitive locus of Jackendoff's Conceptual Semantics. Jackendoff has adopted similar extensions in his recent work (Jackendoff, 2002, 2007).

A situation model is a mental scratchpad for maintaining information about the referents of the referring expressions in a text. However, referents can also be implicit in the text, inferred from background knowledge or encoded from the environment. The situation model is constructed in the context of a *mental universe*. The mental universe is the experience of the real world filtered through the perceptual and cognitive apparatus of an individual over the course of a lifetime. Like situation models, the mental universe may be full of counterfactual objects and situations. An individual may have a long history of experience of unicorns, both perceptual (e.g. from movies and picture books) and linguistic, despite the fact that unicorns only exist as figments of imagination in objective reality. The mental universe may also have well established and distinct referents for the morning star and the evening star, despite the fact that these referents map to the same planet in objective reality.

The combination of the mental universe and the situation model provide the basic sources for grounding the meaning of referring expressions. A referring expression may be bound to a referent in the situation model which may or may not be ground in the mental universe. If the referent is ground in the mental universe then the individual has personal experience of the referent. If the referent is not ground in the mental universe, then the individual has only limited information about the referent and it may appear that the referring expression is non-referential. But as Lyons (1977) notes, allowing referring expressions to be non-referential is problematic for co-reference. "Two expressions cannot have the same reference, if one of them is not a referring expression at all" (Ibid, 191). In "John's murderer, whoever he is…", "he" co-refers with "John's murderer". The attributive use of a referring expression like "John's murderer" is a type of reference which instantiates a referent into the situation model that is not grounded in the mental universe, but which supports co-reference.

The ontology of referential types presented in this paper follows from basic principles of *Cognitive Linguistics* (cf. Langacker, 1987; Lakoff, 1987) and *Cognitive Psychology* (Rosch, 1975; Collins and Quillian, 1969). There is extensive empirical evidence supporting the existence of conceptual categories corresponding to types, prototypes and exemplars. This paper takes the small step of suggesting that such conceptual categories can be referred to by linguistic expressions and explores the consequences.

The representation of referents in the situation model parallels the representation of referring expressions. Both are represented in ACT-R as chunks—i.e. frames with collections of slot-value pairs. Chunks are organized into an inheritance hierarchy which supports default inheritance and a distinction between chunk type and chunk instance. The value of a slot may be a chunk, supporting complex representations of structure needed for linguistic and Conceptual Semantic representation. With respect to object referring expressions which are the focus of this paper, a chunk representing an object referring expression is bound to a corresponding referent via a matching value in an index slot. Depending on the object referring expression, situation model and mental universe, the referent may be an instance, type, prototype, exemplar, class or collection.

An Expanded Ontology of Referential Types

First Order Predicate Calculus (FOPC) is typically grounded in a model theoretic semantics with an ontology limited to atomic individuals. The model consists of a domain and a set of individuals in that domain and nothing else. Typically these individuals are assumed to correspond to objects (or individuals) in the real world being modeled. In FOPC, a relation is modeled in terms of the set of individuals (for 1-ary relations or properties) or set of ordered sets of individuals (for n-ary relations, n>1) for which the relation is true. A relation with its arguments bound to individuals in the domain is either true or false of those individuals and it is said that the reference of the proposition is one of the values true or false.

Situation Semantics (Barwise and Perry, 1983) extends FOPC by allowing situations to be individuals. Not only are situations true or false of sets of individuals in the domain

being modeled, but they are themselves individuals in the domain. We may say that situations have "first-class" status in situation semantics, whereas they are a second-order (or derived) notion in standard FOPC.

Situation Semantics is a step in the right direction. Whereas it might make reasonable sense to suggest that a predicate like "dog" denotes the set (or class) of individuals that are dogs (although psychologically humans cannot quantify over such a large set), it makes little sense to suggest that the predicate "run" denotes the set of all individuals who run, or that "kick" denotes the set of ordered sets of kickers and kickees, as is typical in FOPC treatments with a set-theoretic model limited to individuals that are essentially objects of various types (and sets of such individuals). (It is this sleight of hand in FOPC that collapses the distinction between nouns and verbs, treating both as predicates corresponding to sets of individuals.) It is much more reasonable to suggest that "run" denotes the set of all running events and that "kick" denotes the set of all kicking events. And if "run" denotes a set of running events and "kick" a set of kicking events, then allowing "run" to be used in an expression that refers to an instance of a running event, and allowing "kick" to be used in an expression that refers to an instance of a kicking event, follows quite naturally and is cognitively plausible. However, Situation Semantics stops short. What is needed is a referential ontology which supports a mapping from the types of referring expressions which are linguistically attested to the types of referents which are cognitively motivated.

With an ontology of referential types limited to individuals and sets of individuals, it is often assumed that a referring expression like "a car" in an expression like "a car is a vehicle" quantifies over the set of all individuals for which the predicate "car" is true (i.e. the set or class of objects of type "car"). In FOPC, this can be represented as

$$\forall x \, (\text{car}(x) \rightarrow \text{vehicle}(x))$$

However, from a grammatical perspective, "a car" is clearly singular, and from a cognitive perspective, quantifying over all individuals is cognitively implausible. The need to quantify over all individuals in the FOPC representation of the linguistic expression stems from the limited ontology available in FOPC for representing the meaning of indefinite referring expressions. Only the universal and existential quantifiers—which fail to capture the full range of quantification in natural language—are available.

Similarly, one FOPC representation for the expression "every man owns a car" is given by

$$\forall x \, (\exists y \, (\text{man}(x) \text{ and } \text{car}(y) \rightarrow \text{own}(x,y)))$$

However, in English "every man" is grammatically singular, and a mapping to the universal quantifier is problematic. Johnson-Laird (1983) introduced mental models as a way of overcoming the limitations of quantification in FOPC (among other things). He suggests that the expression "a car" in the sentence "every man owns a car" maps to some representative subset of cars. This representative subset of cars corresponds to the representative subset of individuals referred to by "every man", plus a subset of cars that are not owned. He (1983, p. 421) represents this as

man → car
man → car
(car)

But if "every man" and "a car" are singular and not plural, then "every man" does not refer to multiple men and "a car" does not refer to multiple cars. Johnson-Laird's treatment is cognitively plausible, but inconsistent with the grammatical form of the referring expressions. From a perspective which assumes that the number feature of a referring expression corresponds closely to the number feature of the referent of the expression, there are several cognitively motivated referents for expressions like "every man" and "a car" which do not violate the singular status of the linguistic expressions:

- Type
- Prototype/Exemplar
- Indefinite/Definite Instance

"A car" may refer to a type of object, namely the type of object that is a car. "A car" may also refer to a prototype that represents what is common to most cars, or it may refer to an exemplar which is an instance that is a representative car. Further, "a car" may refer to an indefinite instance with the determiner "a" marking the indefinite status of the referent of "a car". Note that "indefinite instance" is used here as a referential type and not a type of referring expression. In all but a few cases, the type of the referring expression is an *indefinite, singular object referring expression* when grammatically marked by the determiner "a" and a singular head noun ("a few cases" being a notable exception where "a" combines with a plural head noun). Given the occurrence of the indefinite, singular determiner "a" and the singular noun "car" in this expression, "a car" cannot be used to refer to a definite instance of a car, or to a class or collection, but all the other referential types are potential referents of indefinite, singular object referring expressions. Likewise, "every man" may refer to a representative but indefinite, singular instance of a man as is suggested by the singular status of "every man".

Reference to Definite and Indefinite Instances. The determiner "the" marks reference to definite instances. Consider the definite object referring expression "the car". This definite expression indicates that there is already a referent in the situation model that is being referred to or that there is a salient "car" object in the mental universe that is being referred to and this object should be instantiated into the situation model. For a more complex example, consider:

A car is in the driveway. The car is red.

In the first sentence, the expression "a car" is indefinite and instantiates a new referent into the situation model—one that is not (known to be) ground in the mental universe. In the second sentence, the expression "the car" is definite and refers to the referent instantiated into the situation model by "a car". Note that this referent is ungrounded in the sense that it has not been identified with any object in the mental universe, although it could be (e.g. "Oh, it's your car"). It is the mental universe which ultimately grounds referents. In the first sentence, the expression "the driveway" is definite. In this case, the definiteness of "the driveway" indicates there is (or should be) a salient object in the mental universe that should be instantiated into the situation model. There are three primary types of definite reference: 1) reference to an existing referent in the situation model which is grounded in the mental universe, 2) reference to an existing referent in the situation model which is ungrounded in the mental universe, and 3) reference to a object in the mental universe which is not in the situation model, but is (or should be) salient. There are two primary types of indefinite reference: 1) reference to an object which is being introduced and should be instantiated into the situation model—this object is not known to correspond to any object in the mental universe, and 2) reference to a generic instance or type which exists in the mental universe and should be instantiated into the situation model.

Reference to Types. Type hierarchies are common in systems of knowledge representation and making types first class objects allows expressions like "a sedan is a (type of) car" or "a (type of) car I like is a sedan" to be represented as relating two types "a sedan" and "a car". "A sedan" and "a car" refer to *instances of a type*. The suggested reference to a type rather than a class of instances is based on the singular status of these referring expressions (i.e. "a sedan" vs. "all sedans"). A type is a reified class. From a referential perspective, the type is atomic with no subparts and singular reference is appropriate. An instance is added to the situation model which is grounded in a type in the mental universe. From a relational perspective, "is" establishes a relationship of equality between the two arguments "a sedan" and "a car". However, from a referential perspective, there are two basic possibilities: 1) both "a sedan" and "a car" may refer to types of objects which are equated, or 2) the occurrence of "a car" within the context of "is" suppresses the normal referential behavior of "a car" such that "is a car"—a *predicate nominal*—is treated as a non-referential expression which is ascribed to the subject "a sedan". The typical treatment of predicate nominals suggests that they are non-referential (cf. Jackendoff, 2002). In a sentence like "John is a fool", "is a fool" is treated as a predicate nominal that says something about the individual that "John" refers to and this sentence is often considered synonymous with "John is foolish". From the perspective of the grammatical constraint, there is a problem with this treatment. Grammatically, "a fool" has the form of an indefinite, singular object referring expression and all object referring expressions are capable of referring, regardless of context. In the case of a predicate nominal, the referent of the embedded object referring expression, if it is identified, is the same as the referent of the subject—they are co-referential. The assumption that "is a fool" is non-referential rests on the availability of a referring expression "John", the referent of which the predicate nominal "is a fool" is predicated. In the absence of a separate referring expression, it is unclear how to treat the predicate nominal. For example, in "I wonder who is a fool", if "who" is non-referential as Huddleston & Pullum (2002, p. 401) suggest, then what does "is a fool" get predicated of? An obvious suggestion is that "who" functions as an unbound variable (or variable bound via a lambda expression) which instantiates a referent whose grounding is yet to be determined, but which supports predication of "is a fool" and can be referred to subsequently as in the follow up "he better be careful". In fact, it may turn out that nobody is a fool since "wonder" is non-factive (i.e. doesn't entail the existence of its complement). Or it may be the case that the hearer can provide the grounding as in "It's John". In general, Huddleston & Pullum discuss a range of "non-referential" object referring expressions (they prefer to use the term NP) in which there is no object in the real world to which the expressions refer, overlooking the possibility of a more flexible notion of reference within a situation model embedded in a mental universe.

In Jackendoff (2002), types are treated as lacking an indexical feature. While this treatment is attractive in providing a simple distinction between types and tokens (i.e. tokens have an indexical feature, types don't), the lack of an indexical feature implies an inability to refer to types. Yet, Jackendoff acknowledges the existence of NPs which describe types. These NPs are necessarily non-referential. When an NP occurs as a predicate nominal and functions as a kind (or type) as in "a professor" in "John is a professor", this approach coheres. There is an object in the situation model to which the expression refers. But what happens when an NP describing a type occurs as the subject or object as in "A new kind of car is passing by" or "He wants a special kind of dog"? If the object referring expressions don't refer, then it is unclear how the situation model can represent the meaning of these expressions. At a minimum, Jackendoff needs to allow reference to generic instances and argue that apparent references to types are really generic instance references. However, since there is strong evidence that types exist as mental constructs (cf. Collins & Qullian, 1969), we see no good reason to preclude reference to them.

Reference to Generic Instances. The plural variant of the expression "a sedan is a car" is "sedans are cars". This variant suggests a representation based on a collection of generic instances rather than a type.

The generic instance category generalizes over prototypes and exemplars. It is difficult to distinguish reference to prototypes from reference to exemplars since they have much in common. A prototype may be viewed as a washed

out exemplar (some cognitive approaches treat prototype and exemplar as essentially synonymous). It is a washed out exemplar in that it is a generalization over the experience of particular instances of the type. In this respect, a prototype is more like a type than an instance, making the distinction between types and instances less clear cut than is typically assumed. The use of specific lexical items may help to make the distinction. Consider the sentence "the prototypical car is a sedan". If the expression "the prototypical car" actually picks out a prototype for a referent, and the expression "a sedan" picks out a type, then equating a prototype with a type has the effect of defining the prototype to be of a particular type.

Allen (1986) discusses the semantics of generic NPs noting that "there is no marking for the generic within NP morphology" and that generics have "to be inferred from context". Grammatically a singular object referring expression is either definite or indefinite. If the referent of the expression is a prototype or exemplar, then the reference is generic. In the expression "the sedan is a car" where there is no existing referent in the situation model for "the sedan" to refer to, "the sedan" presumably picks out a generic instance or type.

The motivation for distinguishing prototypes and exemplars is a cognitive one, although there is disagreement within the cognitive community as to whether or not both notions are needed. It may be sufficient to distinguish generic instances from types in the situation model without distinguishing prototypes and exemplars.

Reference to Classes, Collections and Masses. Classes, collections and masses complicate reference in interesting ways. Classes and types are two sides of the same coin. The type is atomic and has no subparts. However, the elements of a class are salient and a plural nominal is used to refer to classes as in "all men". Collections are also referred to by plural nominals as in "the men/all the men" where "the men/all the men" refers to some salient collection of men, and not to the entire class. In these expressions, the noun head "men" denotes the type, and the specifier and plural grammatical feature determine the nature of the referring expression (i.e. class or collection). Masses differ from classes and collections in that the elements of a mass are not salient. Singular nominals are used to refer to masses.

Mass and plural nouns, but not singular count nouns, may function as referring expressions without separate specification. In "rice is good for you", "rice" does not refer to any specific instance of rice and in "books are fun to read", "books" does not refer to any specific collection of books. Both expressions are indefinite. They refer to something non-specific: a type or generic instance for "rice" and a generic collection for "books". Reference to a specific mass or collection requires a definite determiner as in "the rice is ready" and "the books are fun to read".

The use of a plural nominal to refer to a class or collection suggests that the members of the class or collection are cognitively salient and may be separately represented. This opens up the possibility of either referring to the class or collection as a whole or referring to the elements of the class or collection. However, for cognitive reasons having to do with the limited capacity of humans to attend to multiple chunks of information (e.g. Miller, 1956), it is assumed that any linguistic expression may only introduce a small number of referents into a situation model (cf. Johnson-Laird, 1983). In the "sedans are cars" example, the instantiation of a sedan collection and two generic instances of a sedan, and a car collection and two generic instances of a car is the minimal number consistent with the plurality of the object referring expressions. Given these referents, it is possible to refer to the collections as a whole, and it is also possible to pair the members of one collection with the members of the other collection. These alternatives correspond to the *collective* and *distributive* readings discussed in Lyons (1977). Lyons presents the example "those books cost $5" which is ambiguous between a distributive—each book is $5—and collective—all the books are $5—reading. Distributive and collective readings involve inferential processes operating over collections and instances which are not part of the grammatically encoded meaning. However, addition of "each" to "those books cost $5 each" imposes a distributive reading.

We can now see that Johnson-Laird's representation of "every man owns a car" corresponds closely to a distributive reading (constrained to a small number of referents). We are also in a better position to consider the representation of "every man". Although expressions with "every" are singular, suggesting selection of an arbitrary instance of a collection, in "Everyone left. They went to eat.", subsequent references are plural. Further, "Everyone left. He went to eat" is infelicitous. There are two implications of these examples: 1) "every" instantiates or references a collection in the situation model, and 2) the arbitrary referent of "every" is not salient for subsequent reference. Even referring expressions with singular "a" as in "Everyone owns a car. They are indispensable." support subsequent plural reference, although in this case "Everyone owns a car. It is indispensable." is also felicitous. This may result from the flipping of the type/class coin. Subsequent singular reference is to the type (or generic instance), subsequent plural reference is to the class.

Reference to Vacuous Instances and Collections. The empty set is a useful notion in set theory. The null symbol (or empty list) is a useful symbol in the Lisp programming language. In both set theory and Lisp, these are actual objects that can be referred to and manipulated. The grammatical and lexical structure of English strongly suggests the possibility of referring to a corresponding empty or vacuous object whose existence is taken for granted. Yet Martinich (1985, p. 3) argues that the existence of nothing is an "absurd view" which rests on "a misunderstanding of how language works". However, not only does grammar suggest the existence of objects

corresponding to nothing, but it suggests that nothingness comes in lots of different types and collections. Consider

Nothing
No one, nobody
Nowhere, Never
No man, No dog
No men, No dogs

It is true that a logical representation for expressions like "no man" which requires quantifying over every individual in the model makes little practical sense

$$\forall x (\sim man(x))$$

but this is taken to be a problem for the logical representation of the meaning of negative expressions, rather than as a criticism of negative referring expressions in language. Allowing negative object referring expressions to refer to empty or vacuous objects and collections in the situation model which do not map to any objects or collections in the mental universe is perhaps the clearest demonstration of how to simplify the mapping from referring expression to referent, relative to other approaches.

Summary and Conclusions

This paper presents and supports an expanded ontology of referential types consistent with Jackendoff's Conceptual Semantics, basic principles of cognitive linguistics and empirical evidence from cognitive psychology. By expanding the ontology of referential types and introducing a distinction between situation model and mental universe, it is possible to simplify the mapping from referring expression to referent, relative to approaches with a more limited ontology and single semantic space.

We propose a bi-partite semantic space consisting of a situation model and mental universe that explains apparent non-referential uses of referring expressions, along with the existence of two partial orderings:

Universal (e.g., $\forall x$) >
Class (e.g., $\forall x$ (man(x)) or "all men") >
Collection (e.g. "some/the/all the men") >
Mass (e.g. "mankind") >
Instance (e.g. $\exists x$ (man(x)) or "a/the man") >
Null (e.g. "no man")

Type > Prototype > Exemplar > Token (Individual)

The partial orderings are motivated by the linguistic expression of referring expressions, cognitive theory and a computational interest in simplifying the mapping from referring expressions to corresponding objects and situations. The partial orderings are not definitive. They capture important aspects of the mapping from referring expressions to referents, but there are more dimensions of meaning involved in this mapping than these two orderings can accommodate.

References

Allen, K. (1986). *Linguistic Meaning*. London: Routledge & Kegan Paul.

Anderson, J. (2007). *How Can the Human Mind Occur in the Physical Universe?* NY: Oxford University Press.

Ball, J. (2007). A Bi-Polar Theory of Nominal and Clause Structure and Function. *Annual Review of Cognitive Linguistics*, 27-54. Amsterdam: John Benjamins.

Ball, J. (2010). Projecting Grammatical Features in Nominals: Cognitive Processing Theory & Computational Implementation. *Proceedings of the 19th Annual Conference on Behavior Representation in Modeling and Simulation*.

Ball, J., Heiberg, A. & Silber, R. (2007). Toward a Large-Scale Model of Language Comprehension in ACT-R 6. *Proceedings of the 8th International Conference on Cognitive Modeling*, 173-179. Edited by R. Lewis, T. Polk & J. Laird. NY: Psychology Press.

Ball, J., Freiman, M., Rodgers, S. & Myers, C. (2010). Toward a Functional Model of Human Language Processing. *Proceedings of the 32nd Conference of the Cognitive Science Society.*

Barwise, J. & J. Perry (1983). *Situations and Attitudes.*Cambridge, MA: The MIT Press.

Collins, A. & M. Quillian (1969). "Retrieval time from semantic memory." *Journal of Verbal Learning and Verbal Behavior*, 8, pp. 240-248.

Hobbs, J. R. (2003). Discourse and inference. Retrieved from http://www.isi.edu/~hobbs/disinf-tc.html

Huddleston, R. & Pullum, G. (2002). *The Cambridge Grammar of the English Language*. Cambridge, UK: Cambridge University Press.

Jackendoff, R. 1983. *Semantics and Cognition*. Cambridge, MA: The MIT Press,

Jackendoff, R. (1991). *Semantic Structures*. Cambridge, MA: The MIT Press.

Jackendoff, R. (2002). *Foundations of Language.* Oxford University Press, New York, NY.

Jackendoff, R. (2007). *Language, Consciousness, Culture, Essays on Mental Structure.* Cambridge, MA: The MIT Press.

Johnson-Laird, P. (1983). *Mental Models*. Cambridge, MA: Harvard University Press.

Kintsch, W. 1998. *Comprehension, a Paradigm for Cognition*. New York, NY: Cambridge University Press.

Langacker, R. (1987). Foundations of Cognitive Grammar, Volume 1, Theoretical Prerequisites. Stanford, CA: Stanford University Press.

Lakoff, G. 1987. *Women, Fire and Dangerous Things*. Chicago: The University of Chicago Press.

Lyons, J. (1977). *Semantics*, Volumes 1 & 2. Cambridge, England: Cambridge University Press

Martinich, A (ed) (1985). *The Philosophy of Language.* New York: Oxford University Press.

Miller, G. A. (1956). "The Magical Number Seven, Plus or Minus Two." *Psychological Review*, 63, pp. 81-94.

Rosch, E. (1975). "Cognitive Representations of Semantic Categories." *Journal of Experimental Psychology: General*, 104, pp. 192-233.

Van Dijk, T. and Kintsch, W. (1983). *Strategies of discourse comprehension*. New York: Academic Press.

Zwann, R., and Radvansky, G. (1998). Siutation models in language comprehension and memory. *Psychological Bulletin*, 123, 162-185.

Toward a Functional Model of Human Language Processing

Jerry Ball[1], Mary Freiman[2], Stuart Rodgers[3] & Christopher Myers[1]
Air Force Research Laboratory[1], L3 Communications[2], AGS TechNet[3]
Jerry.Ball@mesa.afmc.af.mil, Mary.Freiman@mesa.afmc.af.mil, Stu@agstechnet.com,
Christopher.Myers@mesa.afmc.af.mil

Abstract

This paper describes a computational cognitive model of human language processing under development in the ACT-R cognitive architecture. The paper begins with the context for the research, followed by a discussion of the primary theoretical and modeling commitments. The main theoretical commitment is to develop a language model which is at once functional and cognitively plausible. The paper continues with a description of the word recognition subcomponent of the language model which uses a perceptual span and ACT-R's spreading activation mechanism to activate and select the lexical unit that most closely matches the perceptual input. Next we present a description of the linguistic structure building component of the model which combines parallel, probabilistic processing with serial, pseudo-deterministic processing, including a non-monotonic context accommodation mechanism. A description of the mapping of linguistic representations into a situation model, follows. The paper concludes with a summary and conclusions.

Keywords: human language processing (HLP); functional; cognitively plausible; pseudo-deterministic.

Introduction

The capability to model the cognitive processes associated with language is a long sought-after goal of cognitive science. Computational cognitive process models help researchers to not only understand language processes in their own right, but to determine how they affect and interact with other cognitive processes (e.g., reasoning, decision-making, situation assessment, etc.). Scaled-up versions of these models also support the development of cognitive agents with communicative capabilities based on human linguistic processes (Ball et al., 2009; Douglass, Ball & Rodgers, 2009). In this paper we present a "snapshot" of a functional language comprehension model under development within the ACT-R architecture (Anderson, 2007). The model implements a referential and relational theory of human language processing (Ball, 2007; Ball, Heiberg & Silber, 2007) within ACT-R[1].

A key commitment of the language comprehension research is development of a model which is at once cognitively plausible and functional. We believe that adherence to well-established cognitive constraints will facilitate the development of functional models by pushing development in directions that are more likely to be successful. There are short-term costs associated with adherence to cognitive constraints; however, we have already realized longer-term benefits. For example, the integration of a word recognition capability with ACT-R's perceptual system and higher-level linguistic processing has facilitated the recognition and processing of multi-word expressions and multi-unit words in ways that are not available to systems with separate word tokenizing and part of speech tagging processes. Using an available tokenizer and part of speech tagger would have initially facilitated development, but the cognitive implausibility of using staged tokenizing and part of speech tagging led us to reject this approach. The benefits that we have realized as a result of this decision are described below.

Theoretical & Modeling Commitments

There is extensive psycholinguistic evidence that human language processing is incremental and interactive (Gibson & Pearlmutter, 1998; Altmann, 1998; Tanenhaus et al., 1995; Altmann & Steedman, 1988). Garden-path effects, although infrequent, strongly suggest that processing is essentially serial at the level of phrasal and clausal analysis (Bever, 1970). Lower level processes of word recognition suggest parallel, activation-based processing mechanisms (McClelland & Rumelhart, 1981; Paap et al., 1982). Summarizing the psycholinguistic evidence, Altmann & Mirkovic (2009, p. 605) claim "The view we are left with is a comprehension system that is 'maximally incremental'; it develops the fullest interpretation of a sentence fragment at each moment of the fragment's unfolding".

These cognitive constraints legislate against staged analysis models. All levels of analysis must at least be highly pipelined together, if not, in addition, allowing feedback from higher to lower levels. They also suggest the need for hybrid systems which incorporate a mixture of parallel and serial mechanisms, with lower levels of processing being primarily parallel, probabilistic and interactive, while higher levels of analysis are primarily serial, deterministic and incremental.

To adhere to and take advantage of these cognitive constraints, we have developed a *pseudo-deterministic* human language processing model—i.e. a model that presents the appearance and efficiency of serial, deterministic processing, but uses a non-monotonic context

[1] At the time of publication the model contained 6,395 declarative memory elements and 548 production rules which cover a broad range of grammatical constructions.

accommodation mechanism and relies on lower level parallel mechanisms to deal with the ambiguity that makes true deterministic processing impossible. This model makes use of the architectural mechanisms in ACT-R that are most compatible with incremental and interactive processing. For example, parallel, probabilistic processing taps into ACT-R's declarative memory (DM) and parallel spreading activation mechanism, with ACT-R's DM retrieval mechanism supporting probabilistic selection—without inhibition between competing alternatives as is typical of connectionist models (cf. Vosse & Kempen, 2000). Serial, incremental processing is based on ACT-R's procedural memory which is instantiated as a production system. ACT-R at once constrains the computational implementation and provides the basic mechanisms on which the model relies. Other than adding a collection of buffers to support language processing by retaining the partial products of retrieval and structure building, and improving the perceptual processing in ACT-R, the computational implementation does not add any language-specific mechanisms. In the following sections we discuss important subcomponents of the model, such as how the model recognizes words, builds linguistic representations, and maps linguistic representations to a situation representation.

Reading & Word Recognition

A functional language model must deal with the linguistic input as is. In an experiment involving human subjects communicating via text chat (cf. Ball, et al., 2009), we collected a text chat corpus that is riddled with variability in word forms—e.g., misspellings like "altitde", abbreviations like "alt.", and concatenations like "speedrestriction" and "speed=200-500". For competent readers, misspelled words activate the intended lexical items because they contain many of the same letters and trigrams (Perea & Lupker, 2003). Further, all the letters of a word can be transposed, yet still prime the intended word (Guerrera 2004). Key requirements of a functional language model are the ability to handle variability and misspellings in input forms, the ability to separate perceptually conjoined units (e.g. separating punctuation from words as in "He went.", but not "etc."); separating concatenated words, and the ability to recognize multi-word expressions (e.g. "speed up") and multi-unit words (e.g. "ACT-R", "a priori").

To satisfy these requirements, the model includes a word recognition subcomponent that uses ACT-R's spreading activation mechanism combined with a multi-word perceptual span to influence lexical item retrieval. It is assumed that word recognition involves mapping orthographic input directly into DM representations without recourse to phonetic processing (although a phonetic mapping is not precluded). The model does not treat each word as a sum of its parts, ignoring the complete form altogether. Rather, if the text input as a whole does not match, and thereby activate an item in the lexicon, the closest match can be retrieved based on the cues that do match, such as letters, word-length, and trigrams.

In the model's DM, word chunks have slots for letters, word-length, and trigrams. Multi-unit words and multi-word expressions have this information for all of the constituent units. Text input is distilled into this information by the model and put into buffers to spread activation to words in DM containing matching information. The activation mechanism allows the model to retrieve words from DM that are not an exact match to the input. Letters and trigrams in the text input increase the activation of word chunks containing those letters and trigrams in the mental lexicon. The most highly activated word chunk, which need not be an exact match to the input, is retrieved. These processes and encodings are based on the Interactive Activation model of word recognition (McClelland and Rumelhart 1981), with the addition of trigrams based on "letter triples" (Seidenberg and McClelland, 1989).

Besides breaking words into letters and trigrams, we modified the ACT-R architecture to better interpret multi-unit words and multi-word expressions. By default, ACT-R splits input text into perceptual units based on spaces and punctuation—even word internal punctuation, where "ACT-R" becomes "ACT" "-" "R"—and processes each perceptual unit separately. We replaced this behavior with a perceptual span that is based on human reading span data and a multi-level splitting of the input within the perceptual span into larger and smaller perceptual units which spread activation in parallel. We also added multi-word expression chunks and multi-unit lexical chunks to DM. The overall effect is a significant reduction in the number of DM retrievals per space and punctuation delimited input. Words with internal punctuation and multi-word expressions can now be retrieved as a single perceptual unit despite their internal structure (Freiman & Ball, submitted).

The new perceptual span is considerably larger than ACT-R's punctuation and space delimited span. There is a great deal of evidence that the perceptual span of adult readers is about 14-15 letters to the right of fixation (McConkie & Rayner, 1975; Rayner, 1986). We implemented a span of up to twelve letters, with the greatest amount of activation spreading from the first few letters of the span and decreasing toward the end of the span. Just as for adult readers, information to the right of fixation is obtained when the next word is predictable from the preceding text (see Rayner 1975; and Binder, Pollatsek, & Rayner, 1999).

Within the context of a functional language model—i.e. one that must interpret and act on the linguistic input, we are also attempting to model adult human reading rates (Freiman & Ball, submitted). Adult humans read at a phenomenal rate of 200-300 (space delimited) words per minute (Carver, 1973a; 1973b). The ACT-R architecture supports the timing of cognitive processes down to the msec level. The real-time it takes for a model to run can also be measured. Although we have not yet succeeded in achieving adult reading rates, we have improved the reading rate of the model significantly in both cognitive and real-time: 143 words per minute in ACT-R cognitive time (important for

cognitive plausibility); and 249 words per minute in real-time on a single-core, 2.1 GHz Windows Vista machine with 2 gigabytes of RAM (important for a functional model). Ultimately, we believe that achieving adult reading rates hinges on minimizing the amount of structure building and maximizing the average size of linguistic units which are retrieved. We are pursuing mechanisms and representations that will make this possible.

Building Linguistic Representations

The word recognition subcomponent typically delivers a lexical item categorized for part of speech to the higher level component that builds linguistic representations of referential and relational meaning. For example, consider the processing of "the pilot". The processing of "the" leads to its identification as a determiner via retrieval from DM. Selection of this lexical item is based on the probabilistic, context-sensitive mechanism discussed in the previous section. The subsequent processing of the determiner "the" leads to the projection or construction of a nominal construction. The processing of the word "pilot" in the context of the preceding word "the" and the projected nominal leads to retrieval of a DM chunk identifying "pilot" as a noun. The noun "pilot" is then integrated as the head of the nominal projected during the processing of "the".

Similar parallel, probabilistic mechanisms operate at the phrasal and clausal level, selecting between competing phrasal and clausal alternatives, and potentially interacting with lower level probabilistic mechanisms. As an example of this potential interaction, consider the processing of personal pronouns like "he" and "it". At the lexical level, these words are categorized as pronouns, but they are also closely associated with the nominal phrasal category since they typically function as the head of a complete nominal. Processing personal pronouns may involve their recognition as pronouns followed by projection of a nominal phrase from the pronoun, but it may also be that the perceptual form can directly lead to retrieval of a nominal phrase, without the intermediate step of identifying the word as a pronoun. The word recognition component, which prefers larger and higher level units, may deliver a pre-compiled nominal unit corresponding to the pronoun, rather than a lexical unit to the higher level construction process, blurring the distinction between lexical and phrasal units. The determiner "the" may behave similarly, resulting in direct retrieval of a nominal with an empty head, without the intermediate step of identifying "the" as a determiner.

The parallel, probabilistic mechanism which is capable of retrieving existing phrasal and clausal representations as well as lexical units, competes with a mechanism which builds novel representations. DM retrieval has priority over this alternative construction mechanism. However, lexical units are more likely to be available for retrieval than phrasal and clausal representations. Further, the parallel, probabilistic mechanism is not capable of building any structure—building structure is the function of the serial construction mechanism.

There are two basic ways of building structure: 1) integration of the current linguistic unit into an existing representation which contains an expectation for the linguistic unit (i.e. substitution), and 2) projection or construction of a novel representation coupled with integration of the current linguistic unit into the novel representation. For example, the processing of the word "pilots" recognized as a plural noun by the word recognition component can lead to projection of a nominal and integration of "pilots" as the head of the nominal. On the other hand, if "the" has already projected a nominal and set up the expectation for a head to occur, the processing of "pilots" can lead to its integration as the head of the nominal projected by "the".

The structure building mechanism is incremental in that it executes a sequence of productions that determine how to integrate the current linguistic unit into an existing representation and/or which kind of higher level linguistic unit to project. These productions execute one at a time within the ACT-R architecture which incorporates a serial bottleneck for production execution. Although supported by extensive empirical evidence, the serial production execution bottleneck is a characteristic of ACT-R that distinguishes it from other production system architectures which support parallel production execution.

The structure building mechanism uses all available information in deciding how to integrate the current linguistic input into the evolving representation. Although the parallel, probabilistic mechanism considers multiple alternatives in parallel, the output of this parallel mechanism is a single linguistic unit and the result of structure building is also a single representation. The structure building mechanism operates in a *pseudo-deterministic* manner. It is deterministic in that it builds a single representation which is assumed to be correct, but it relies on the parallel, probabilistic mechanism to provide the inputs to this structure building mechanism. In addition, structure building is subject to a mechanism of context accommodation capable of making modest adjustments to the evolving representation (Ball, 2010a). Although context accommodation does not involve backtracking or reanalysis, it is not, strictly speaking, deterministic, since it can modify an existing representation and is therefore non-monotonic. For example, in the processing of the expression "the altitude restriction", when the word "altitude" is processed, it can be integrated as the head of the nominal projected by "the". But when "restriction" is subsequently processed, the context accommodation mechanism can adjust the representation, shifting "altitude" into a modifying function so that "restriction" can function as the head. This context accommodation capability can apply iteratively as in the processing of "the pressure valve adjustment screw" where "screw" is the ultimate head of the nominal, but "pressure", "valve" and "adjustment" are all incrementally integrated as the head prior to the processing of "screw". Note that at the end of processing it appears that "pressure", "valve" and "adjustment" were treated as

modifiers all along, giving the appearance that these alternatives were carried along in parallel with their treatment as heads.

Context accommodation uses the full available context to make modest adjustments to the evolving representation or to construe the current input in a way that allows for its integration into the representation. As an example of construal, the verb "kick" is construed as an object and functions as the head of a nominal when it occurs in the context of "the", as in "the kick". Function overriding and function shifting are two additional mechanisms of context accommodation. We have already seen an example of function shifting (e.g. "the altitude restriction"). In the processing of "no altitude or airspeed restrictions", the conjoined head "altitude or airspeed" can override the initial treatment of "altitude" as the head of the nominal, with the subsequent shifting of "altitude and airspeed" into a modifying function during the processing of "restrictions". At a lower level, there are accommodation mechanisms for handling conflicts in the grammatical features associated with various lexical items. For example, the grammatical feature *definite* is associated with "the" and the grammatical feature *indefinite* is associated with "pilots". In "the pilots", the *definite* feature of "the" blocks the *indefinite* feature of "pilots" from projecting to the nominal. See Ball (2010b) for more details.

Context accommodation need not be computationally expensive—a single production may effect the accommodation, just as a single production may effect integration without accommodation. In this respect, context accommodation is not a reanalysis mechanism that disrupts normal processing—it is part and parcel of normal processing. Reanalysis mechanisms need only kick in when context accommodation fails and larger adjustment is needed. The mechanism of context accommodation is most closely related to the limited repair parsing of Lewis (1998). Context accommodation may be viewed as a very modest form of repair. According to Lewis (1998, p. 262) "The putative theoretical advantage of repair parsers depends in large part on finding simple candidate repair operations". The mechanism of context accommodation provides evidence for this theoretical advantage.

Overall, the highly interactive, parallel, probabilistic mechanism for selecting between competing alternatives combines with the incremental, serial construction and context accommodation mechanisms to provide an efficient, pseudo-deterministic language processing capability.

Mapping into the Situation Model

Although we borrow the term (cf. Zwann & Radvansky, 1998), we define *situation model* as a domain-specific mental representation of a set of objects, actions, events, and relationships related to a task, sufficient for reasoning about a set of actions within that task. The situation model is separate from the model's world knowledge but is related to and affected by world knowledge.

The situation model is implemented in three main subcomponents: the ACT-R module definition, a set of domain general production rules, and a set of domain specific production rules. The module is instantiated like other ACT-R modules (Anderson, 2007), and includes the module buffers and handlers for module requests and queries.

The main situation buffers are: sm-subject-context, sm-related-object-context, sm-sit-context, sm-action-context, sm-event-context, and sm-prior-attention. They are named and designed to reflect the semantics of the represented situations. The buffers will contain chunks representing the objects, actions, events, and relationships discussed or encountered in the task environment. The top level chunk types were based upon the Suggested Upper Merged Ontology (SUMO) (Niles and Pease, 2001) and are: Action, Attribute, Concept, Event, Object, Relation, and Situation. All entities represented in the situation model will be sub-typed from one of these top level chunk types. Because the situations being represented in our model may span multiple sentences, the contents of the sm-subject-context buffer will frequently not equate to the subject of an individually processed sentence. Rather, the contents of the sm-subject-context buffer should be thought of as the central topic or theme of the discourse at an individual moment. The situation chunk-type and its sub-types can be thought of as instances of *schemata* or structures for mental models of stereotypical situations (Alba, 1983). In our implementation, the situation chunk contains the relevant gist of the situation, where the "gist" can be thought of as an index to a specific category of situation.

It is the responsibility of the modeler to define any needed specific chunk subtypes. Because ACT-R's chunk inheritance mechanism does not permit inheritance from multiple supertypes, it is expected that there will be some redundancy in the definitions of the chunk subtype hierarchy. While this redundancy will create some inefficiency in the type hierarchy design, it should not preclude the modeling of necessary elements.

The domain general productions manage the relationships between elements within each individual situation. For instance, in a situation involving an uninhabited air vehicle altitude restriction for a reconnaissance waypoint, a situation chunk would contain a subject slot and a related object slot. The subject slot value would refer to the reconnaissance waypoint and the related object slot value would refer to the waypoint's altitude restriction. The domain general productions provide the mechanisms that manage the references between the situation elements.

The domain specific productions primarily consist of task knowledge and responses to the situations, events, actions, and objects that are learned from interacting with a specific task environment. It is the modeler's responsibility to define the needed domain specific productions. A central goal of current research is to discover regularities and useful abstractions within the domain specific production rules that can be generalized.

The situation model represents the domain specific objects and situations to which the linguistic representations refer. The linguistic comprehension system interfaces to the non-linguistic situation model via the identification of referring expressions in the linguistic input. For example, recognition of a nominal, or object referring expression, results in the mapping to a corresponding object in the situation model. There are two basic cases: 1) recognition of a definite object referring expression typically results in identification of an existing object in the situation model or surrounding context, and 2) recognition of an indefinite object referring expression typically results in the introduction of a new object into the situation model. Extensions to these basic cases are considered in Ball (2010c) which expands the ontology of referential types to include types, collections, exemplars, prototypes and even negative instances. The extended ontology has the important benefit of simplifying the mapping from referring expressions to situation model entities.

An object referring expression from the comprehension system is mapped to the situation model when the head of the object referring expression is identified. For example, if the input is "the altitude", then recognition of "altitude" as the head triggers the mapping to the situation model. Note that if the input is actually "the altitude restriction", an altitude object will still be mapped to at the processing of "altitude". At the processing of "restriction" an "altitude restriction" object will be mapped. Further, if a post-head modifier occurs as in "for Waypoint-A" in "the altitude restriction for Waypoint-A", the mapping may need to be modified following processing of the post-head modifier. The model does not currently attempt to map to an object on the basis of pre-head modifiers as in "the red…" although there is evidence that humans may do so in Visual World Paradigm tasks (Tanenhaus et al., 1995). It should be noted that object referring expressions contain ambiguous words, not word senses or abstract concepts. It is the mapping to objects in the situation model which disambiguates the words in the linguistic representation.

Other challenges include anaphora and co-reference resolution. We currently use grammatical features to constrain the possible co-referents of a pronoun (e.g. "it" is *inanimate* and *singular*). We plan to adhere to the constraints of binding theory with respect to binding pronouns and anaphors (Chomsky, 1981) and to adopt mechanisms of Centering Theory (Grosz, Joshi & Weinstein, 1995) in a more complete implementation. We are not proposing a general solution in our research program; however, we expect to implement an initial capability for co-reference resolution by relying on ACT-R's chunk merging feature. So long as the specific context for a chunk is the same for newly introduced references to previously referenced knowledge elements, some amount of the new references automatically merge with previously constructed chunks in DM. For a more general solution, existing approaches to co-reference resolution are being investigated for inclusion in our design.

Summary and Conclusions

This paper describes a model of human language processing which is intended to be both functional and cognitively plausible. It includes a linguistic structure building mechanism which combines a serial, deterministic processing mechanism with a non-monotonic mechanism of context accommodation, and a lower level parallel, probabilistic mechanism for selecting between competing alternatives. Overall, the model is pseudo-deterministic—it presents the appearance and efficiency of deterministic processing, and can handle much of the more mundane ambiguity evident in human language via the parallel, probabilistic and non-monotonic context accommodation mechanisms. The model adheres to well-established cognitive constraints on human language processing including incremental and interactive processing. This commitment led to the integration of a cognitively plausible word recognition subcomponent, rather than adopting an off-the-shelf tokenizer and part of speech tagger that lacked cognitive plausibility.

A key attribute of the language comprehension model is the capability to handle variability and mismatch at all levels of analysis from word recognition, through the generation of linguistic representations and the mapping into the situation model, to the determination of the conversational implicatures not literally described in the linguistic input (although the capability to handle conversational implicatures is not yet implemented). There is no level of analysis at which variability and mismatch can be ignored.

The language comprehension model is a key component of a larger synthetic teammate model which is capable of functioning as the pilot in a three-person simulation of an uninhabited air vehicle reconnaissance mission task (Ball, et. al, 2009). The main objective of the synthetic teammate project is to develop cognitive agents capable of being integrated into team training simulations while maintaining training efficacy. To achieve this goal, synthetic teammates must be capable of closely matching human behavior. To this end, we have developed and integrated models of several important cognitive capacities into a composite synthetic teammate model. In addition to language comprehension and situation modeling, these capacities include the ability to perform the UAV piloting task, and language generation and dialog modeling capabilities.

Although we do not report a direct comparison of model results to human data, Cassimatis, Bello & Langley (2009) argue that models of higher-level cognitive processes, such as language comprehension, may be better evaluated on model breadth, parsimony, and functionality. Ball (2008) provides similar arguments for a functional approach, but makes a stronger commitment to cognitive plausibility. The synthetic teammate is capable of receiving text communications from a teammate, reading the text, producing linguistic representations of the text, and mapping the representations into a situation model. Based

on the contents of the situation model, the synthetic teammate then interacts with its task environment, or responds to communications with its own text messages. We believe that this demonstrates the functionality and capability of the presented language comprehension model.

References

Alba, J. W., & Hasher, L. (1983). Is memory schematic? *Psychological Bulletin, 93*, 203-231.

Altmann, G. (1998). Ambiguity in sentence processing. *Trends in Cognitive Sciences, 2*(4), 146-152.

Altmann, G. & Mirkovic, J. (2009). Incrementality and prediction in human sentence processing. *Cognitive Science, 222*, 583-609.

Altmann, G., & Steedman, M. (1988). Interaction with context during human sentence processing. *Cognition, 30*, 191-238.

Anderson, J. (2007). *How Can the Human Mind Occur in the Physical Universe?* NY: Oxford University Press.

Anderson, J. R., Bothell, D., Byrne, M.D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review, 111*, 1036-1060.

Ball, J. (2007). A bi-polar theory of nominal and clause structure and function. *Annual Review of Cognitive Linguistics*, 5, 27-54.

Ball, J. (2008). A naturalistic, functional approach to modeling language comprehension. *Papers from the AAAI Fall 2008 Symposium, Naturally Inspired Artificial Intelligence*. Menlo Park, CA: AAAI Press.

Ball, J. (2010a). Context Accommodation in Human Language Processing. *Proceedings of the Natural Language Processing and Cognitive Science Workshop*. Lisbon: INSTICC Press.

Ball, J. (2010b). Projecting grammatical features in nominals: Cognitive Processing Theory and Computational Implementation. *Proceedings of the 19th Behavior Representation in Modeling and Simulation Conference*.

Ball, J. (2010c). Simplifying the mapping from referring expression to referent in a conceptual semantics of reference. *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society*.

Ball, J., Heiberg, A. & Silber, R. (2007). Toward a large-scale model of language comprehension in ACT-R 6. In R. Lewis, T. Polk & J. Laird (Eds.) *Proceedings of the 8th International Conference on Cognitive Modeling* (pp. 173-179). NY: Psychology Press.

Ball, J., Myers, C. W., Heiberg, A., Cooke, N. J., Matessa, M., & Freiman, M. (2009). The Synthetic Teammate Project. *Proceedings of the 18th Annual Conference on Behavior Representation in Modeling and Simulation*. Sundance, UT.

Bever, T. (1970). The cognitive basis for linguistic structures. In J. Hayes (Ed.), *Cognition and the development of language*, 279-362. NY: Wiley.

Binder, K., Pollatsek, A., & Rayner, K. (1999). Extraction of information to the left of the fixated word in reading. *Journal of Experimental Psychology: Human Perception and Performance, 25*, 1162-1172.

Carver, R. (1973). Understanding, information processing and learning from prose materials. *Journal of Educational Psychology, 64*, 76-84.

Carver, R. (1973). Effect of increasing the rate of speech presentation upon comprehension. *Journal of Educational Psychology, 65*, 118-126.

Cassimatis, N., Bello, P. & Langley, P. (2008). Ability, breadth, and parsimony in computational models of higher-order cognition. *Cognitive Science, 32*, 1304-1322.

Chomsky, N. (1981). *Lectures on Government and Binding*. Dordrecht, Holland: Foris.

Culicover, P. (2009). *Natural Language Syntax*. NY: Oxford University Press.

Douglass, S., Ball, J. & Rodgers, S. (2009). Large declarative memories in ACT-R. *Proceedings of the 9th International Conference on Cognitive Modeling 2009*, Manchester, UK.

Freiman, M., & Ball, J. (2008). Computational cognitive modeling of reading comprehension at the word level. *Proceedings of the 38th Western Conference on Linguistics*, 34-45. Davis, CA: University of California, Davis.

Freiman, M. & Ball, J. (submitted). Improving the reading rate of Double-R-Language.

Grosz, B., Joshi, A. & S. Weinstein (1995). Centering: A framework for modelling the local coherence of discourse. University of Pennsylvania: IRCS Technical Report Series.

Guerrera, C. (2004). Flexibility and constraint in lexical access: Explorations in transposed-letter priming. Unpublished dissertation, Department of Psychology, University of Arizona.

Gibson, E., & Pearlmutter, N. (1998). Constraints on sentence comprehension. *Trends in Cognitive Sciences, 2(*7), 262-268.

Lewis, R. L. (1998). Leaping off the garden path: Reanalysis and limited repair parsing. In J. D. Fodor, & F. Ferreira (Eds.), *Reanalysis in Sentence Processing*. Boston: Kluwer Academic.

McClelland, J., & Rumelhart, D. (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychological Review, 88*(5), 375-407.

McConkie, G. W., & Rayner, K. (1975). The span of the effective stimulus during a fixation in reading. *Perception & Psychophysics, 17*. 578-586.

Paap, K., Newsome, S., McDonald, J. & Schvaneveldt, R. (1982). An activation-verification model of letter and word recognition: the word-superiority effect. *Psychological Review, 89*, 573-594.

Perea, M., & Lupker, S. J. (2003). "Does jugde activate COURT? Transposed-letter similarity effects in masked associative priming". *Memory and Cognition, 31*, 829- 841.

Rayner, K. (1975). The perceptual span and peripheral cues in reading. *Cognitive Psychology, 7*, 65-81.

Rayner, K. (1986). Eye movements and the perceptual span in beginning and skilled readers. *Journal of Experimental Child Psychology, 41*, 211-236.

Seidenberg, Mark S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review, 96(4)*, 523-568.

Tanenhaus, M., Spivey-Knowlton, M., Eberhard, K., & Sedivy, J. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science, 268(5217)*,1632-1634.

Vosse, T. & Kempen, G. (2000). Syntactic structure assembly in human parsing: A computational model based on competitive inhibition and a lexicalist grammar. *Cognition, 75*, 105-143.

Zwann, R. & Radvansky, G. (1998). Situation models in language comprehension and memory. *Psychological Bulletin, 123(2)*, 162-185.

Linking Learning to Looking: Habituation and Association in Infant Statistical Language Learning

Daniel Yurovsky, Shohei Hidaka, Chen Yu, and Linda B. Smith
{dyurovsk, shhidaka, chenyu, smith4} @indiana.edu
Department of Psychological and Brain Science, and Cognitive Science Program
1101 East 10th Street Bloomington, IN 47405 USA

Abstract

Recent experiments have shown the importance of statistical learning in infant language acquisition. Computational models of such learning, however, often take the form of corpus analyses and are thus difficult to connect to empirical data. We report a cross-situational learning experiment which demonstrates robust individual differences in learning between infants. We then present a novel generative model of cross-situational learning combining two competing processes – *habituation* and *association*. The model's parameters are set to best reproduce each infant's individual looking behavior from trial-to-trial in training and testing. We then isolate each infant's word-referent learning function to explain the variance found in preferential looking tests.

Keywords: statistical learning; computational modeling; cognitive development; language acquisition

Introduction

Language acquisition should be hard but young children nonetheless move from babbling to complex sentences in a remarkably short time. One might thus expect the underlying language learning mechanism to also be complex, involving constraints and biases (Markman, 1990) and sophisticated inferences (Xu & Tenenbaum, 2007). However, even if the final mechanism is complex, it must begin with something simple – language learners develop. By understanding the tools available to very young learners, we may develop insight into how more complex mechanisms are created and how they might be understood as products of simpler mechanisms.

One candidate for a simple mechanism is the accumulation of associations between words and objects in a child's ambient environment (Hollich, Hirsh-Pasek, & Golinkoff, 2000, Smith, 2000). If the co-occurrence structure of the world is informative, such that words frequently occur with the objects they label, a child who can attend to this information could find a wedge into learning the more complicated structural aspects of her language (Landau, Smith, & Jones, 1988).

Recently, Smith and Yu (2008) have provided evidence of just such a sensitivity in 12 and 14-month-old infants. In the cross-situational learning paradigm, infants are exposed to a series of individually ambiguous learning trials containing multiple words and objects. While each trial contains several potential mappings, some of which are spurious, a child who can attend to the overall co-occurrence structure can unambiguously determine the correct mappings.

Attempts to understand the mechanism underlying this competence, however, have been aimed primarily at the abstract computational level. Computational models have taken the form of corpus analyses (Fazly, Alishahi, & Stevenson, in press, Frank, Goodman, & Tenenbaum, 2009, Yu, 2008) and thus have resisted direct comparison to empirical data. To understand the mechanisms available to budding language learners, however, models must account for and explain the *behavior* of young infants.

Because preferential looking is the primary measure of learning in studies of preverbal infants, it is this looking data that computational model must explain. Yu and Smith (in press) took a first step towards this goal. Using an associative model, they found that the *number* of words for which an individual infant showed preferential looking behavior was predictable from that infant's own eye movement in training. This might seem to fit an associative learning mechanism: one learns to associate words to the objects at which one is looking when one hears the words. However, Yu and Smith were unable to predict *which* word-referent mappings were learned. If associative learning is the relevant mechanism, something is still missing.

We propose to take two more steps towards understanding the mechanism supporting cross-situational learning. First, whereas Yu and Smith's model was descriptive – using patterns in training behavior to predict test behavior – we present a *generative* model of eye movements. That is, we construct a model which produces eye-movement behavior matching that of infants during training, and then show that the same model accounts for the test data. Second, we predict not only *how many* word-referent mappings each infant learned, but also *which ones*. This modeling is done at the *individual infant* level, allowing us to explain behavior as it unfolds trial-by-trial throughout training and testing.

To motivate our model, we first present results from a cross-situational learning experiment with 15-month-old infants. Analysis of preferential looking test results shows robust individual differences among infants, underscoring the importance of understanding cross-situational learning at a process level. We then construct a model that generates fixations through the competition of two well-known processes that organize infant behavior and learning – *association* (Smith, 2000) and *habituation* (Hunter & Ames, 1988. Model parameters are fit to best account for each individual infant's looking behavior over the course of the experiment, and then inferences about learning are drawn from these parameter fits.

Experiment

Method

Infants were exposed to a cross-situational word learning task (Smith & Yu, 2008; Yu & Smith, in press). Each child viewed a series of trials pairing two novel objects with one novel label. While the correspondence between words and objects on an individual trial was ambiguous, cross-trial co-occurrence statistics between words and objects indicated the correct pairings. After 60 training trials, preferential looking tests were used to determine whether infants had learned the correct pairings.

Participants. Twenty-five 15-month-old infants (14 females, *M* = 14 mos, 23 days, range: 13;22 to 16;4) composed the final sample. Twelve additional infants were excluded due to fussiness (*N*=11) or experimental error (*N*=1).

Stimuli. Six pseudoword labels were recorded by a female native English speaker in isolation and presented to infants over loudspeakers. Six novel two-dimensional objects, each a unique bright color, were presented to infants two at a time on a 47" by 60" white screen. All stimuli were constructed to be comparable to those used in previous cross-situational learning experiments (Smith & Yu, 2008, Yu & Smith, in press).

Procedure. Infants sat on their mother's laps 3.5 feet away from a large white projection screen. Direction of gaze was recorded by a Tobi X60 eye-tracker as well as a camera directed at the child's eyes. Parents were instructed to shut their eyes during the course of the experiment so as not to influence infant behavior.

Training consisted of 60 2-second long training slides. Each slide presented two objects, one on each side of the screen, and was accompanied by one of the recorded labels. A slide's label was presented 700ms after the objects' onsets. On each slide, one of the objects was the label's correct referent and one was a foil. This correspondence was uncorrelated with spatial location, but could be determined from cross-trial co-occurrence statistics: each label occurred 10 times with its correct referent and only 2 times with each of the other objects. Training trials were interspersed with presentations of Sesame Street characters intended to maintain infant attention. Total training lasted approximately 4 minutes.

Following training, infants were exposed to 6 testing trials, each 8 seconds long. Test trials began with approximately 1 second of silence, followed by six repetitions of a label – each separated by 1 second. Two objects were visible for the entire 8 seconds – the label's correct referent and a distractor object. Each of the 6 labels was tested once, and each object appeared equally often as a target and a distractor. Figure 1 illustrates the time course of training and testing with sample trials.

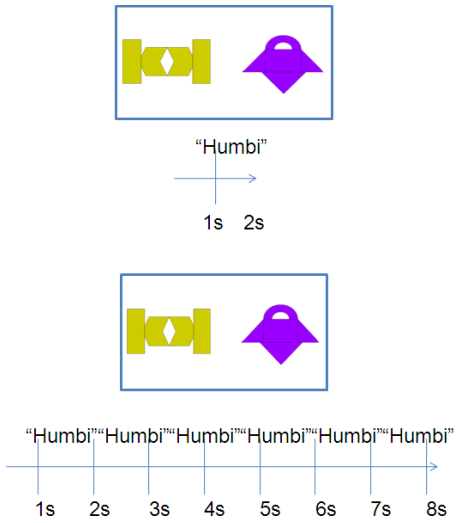


Figure 1: The time course of training (above) and testing (below) trials. Infants saw two objects and heard a label produced either once (training) or 6 times (test). The first 1 second window of each was silent; every subsequent window contained an auditory label.

Data. Gaze position was recorded via eye-tracker at a rate of 50Hz. Because of movement or looking away during the experiment, there were some discontinuities in automatic gaze recording. On average, 57.8% of each infant's gaze points were recorded. Naïve coders blind to the contents of each slide coded each of the remaining frames for direction of gaze (left, right, away/unknown). After hand-coding, 74.5% of all gaze points were mapped to a screen position where one of the objects appears.

Results and Discussion

Infants looking times to target and distractor objects on each of the 12 preferential looking test trials were submitted to a 2 (Target/Distractor) x 6 (Word) x 25 (Subject) mixed ANOVA. The analysis revealed no main effects, but showed a highly significant interaction between Target/Distractor and Subject (F = 3.66, p < .001, η^2 = .1). Individual infants thus showed reliably different looking patterns at test: some looked reliably longer at targets than distractors; others looked reliably longer at distractors than targets (Figure 2). This is consistent with previous work on slow vs. fast habituators (Cashon & Cohen, 2000, Schöner & Thelen, 2006).

Why should there be reliable individual differences? It is well known that the function that maps learning onto looking is *nonmonotonic* – it switches directions (Hunter & Ames, 1988 – Figure 3). This complicates the interpretation of looking behavior, with some investigators of word learning behavior suggesting that increased looking to the target indicates learning (e.g. Golinkoff, Hirsh-Pasek, Cauley, & Gordon, 1987) whereas others interpret increased looking to the distractor as evidence of learning via violation of expectation (e.g. Stager & Werker, 1997).

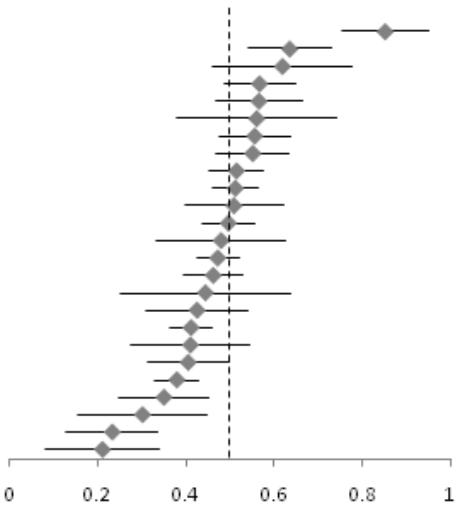


Figure 2: A plot of mean(std err) preferential looking to target for each infant. Values greater than .5 indicate an average preference for the target; those less than .5 indicate preference for the distractor.

The above analysis indicates that individual infants show reliable looking patterns when tested for their preference to look to or away from a label's referent. However, since individual infants show different patterns, it is unclear how to interpret their behavior. For which infants should we infer learning? In the following computational modeling effort, we propose to show that an unambiguous answer can be found through model selection. If we are explicit about the mechanisms which combine to generate looking behavior, we can ask if a learning mechanism is necessary to explain individual infants' looking behavior at test.

Computational Model

Throughout the experiment, infants were exposed to a series of slides presenting two objects along with an auditory label word. Infants responded to these stimuli - at any point in time - by fixating one of the two objects on the screen. Our goal was to derive a generative model for each infant that produced fixation patterns that best approximated his or her own generated fixations.

Because of the structure of the training and testing trials, we divided the time course of fixations into a series of 1 second bins (Figure 1). Proportion of looking to each of the two on-screen objects was calculated in each such window, and model was fit to this data.

Conceptually, the model is simple. Let us suppose that fixation patterns within a given window are generated by the combination of two processes: *habituation* to each of the objects on the screen, and *association* between each of the objects and the label being heard. Let us also suppose that each of these processes is a function of looking time to the input. However, because we do not know the true form of these functions (although see Schöner & Thelen, 2006), we *approximate* them with arbitrary degree polynomials. These polynomial approximations allow us to make inferences about the shape of the functions without making claims about their exact form.

We use each infant's individual training data to infer the *habituation* and *association* functions which best account for that infant's behavior. Because one cannot learn what one does not see, habituation and association are functions of gaze duration rather than occurrence frequency. We thus produce an explicit *linking function* from learning to looking at test, and this function is used to infer what each child learned from her looking behavior in training. Doing so allows us to move beyond preferential looking as a measure of learning, and to make deeper and more specific conclusions about the mechanisms supporting cross-situational learning in real time.

Data

In the experiment, infants were exposed to 60 training trials followed by six test trials. The label for each 2s trial was heard 700ms into the trial. Adding 367ms to the label's onset to account for processing time (Swingley & Aslin, 2000) results in two ~1s windows (Figure 1, top). In the first window, we assume that fixations are being driven by the objects (*habituation*) only, and in the second we assume that fixations are driven both by the objects (*habituation*) and the co-occurring word (*association*).

Test trials had a similar structure (Figure 1, bottom). Each began with a short period of silence, followed by the onset of a label which was then repeated 5 more times at 1 second intervals. We divide each testing trial into 7 1s windows: 1 in which fixations are driven only by the objects, and 6 in which fixations are driven by both objects and the label. The natural logarithm of the odds of looking at each of the on-screen objects was computed in each window, and these are the data to which the model was fit. Log odds is similar to proportion of looking, but has nicer mathematical properties for this particular analysis (see also, Barr, 2008). Any windows in which there was no fixation data for an infant were left out of that infant's dataset.

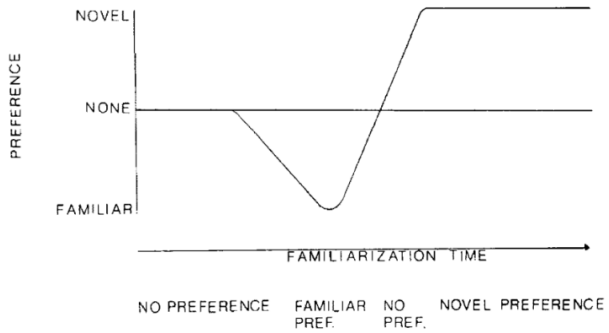


Figure 3: A schematic of the infant looking preference function reproduced from Hunter and Ames (1988). Because the function is *nonmonotonic* – direction of preference changes in opposite directions across time – looking time data resists straightforward interpretation.

Model Description

In a given window, each of the two objects had an activation level as described below. Odds of looking to each of the objects were computed using the ratio form of the Luce choice axiom (Luce, 1959). We additionally adjust the odds ratio in two ways.

Because saccades are controlled by a vision system subject to physical constraints, and because we model looks in 1 second windows, the current window depends on the location of the eye in the last window. For this reason, as an approximation, we modify $OddsLook_t(O)$ by a parameter p times the odds of looking in the previous window to the object in O's current location. Further, infants are known to display preferences for one side of the screen over another. For this reason, we build in a constant term b which models each infant's potential preference for the left or right side of the screen.

Thus, on trial t, if objects O_1 and O_2 are present,

$$OddsLook_t(O_1) = \frac{e^{Act_t(O_1)}}{e^{Act_t(O_2)}} \times \Big(b \cdot \big((Loc(O_1) = left)\big)\Big) \times \Big(p \cdot OddsLook_{t-1}(Loc(O_1))\Big)$$

Silent Window. In a window in which no label is present, activation is driven by an infant's *habituation* to each of the objects present. Habituation to an object was approximated by an arbitrary degree polynomial function *habit* evaluated on the cumulative looking time to that object so far in the experiment. Estimation of the parameters of this function for each infant will be described below.

$$Act_t(O) = habit(O)$$

Label Window. For windows in which a label was heard, we assume that activation is also driven by the association between each object and the label W. For these windows,

$$Act_t(O) = habit(O) + assoc(O, W)$$

Association and Habituation. Each infant's individual *habituation* and *association* functions were approximated by arbitrary degree polynomial functions. For each infant, all possible orders 0 to 2 were tried for each function, with the optimal parameters fit as described below. The final order of each function was chosen using AIC to be the most parsimonious fits for the infants looking behavior.

Formally, if t_o is cumulative looking time to an object, and $t_{o|w}$ is cumulative looking time to an object in the presence of a word,

$$habit(t_o) = \sum_{n=1}^{N_h} h_n \cdot {t_o}^n \qquad assoc(t_{o|w}) = \sum_{n=1}^{N_a} a_n \cdot {t_{o|w}}^n$$

Thus, one infant might have a quadratic *habituation* function (N_h = 2) and a linear *association* function (N_a = 1), while for another infant the best model may have been a linear *habituation* (N_h = 1) function and no *association* function (0 degree) at all.

Model Fitting

In order to determine the best approximation to each infant's individual learning functions, we constructed all 9 possible combinations of orders 0 to 2 for both *association* and *habituation* functions. The optimal parameters for each function were selected to best account for the infant's fixation data. Subsequently, model selection using AIC was performed for each infant by selecting from these models the one which also gave the best account of the individual infant's testing eye fixations without overfitting.

Results and Discussion

On average, the best generative model for each infant predicts a significant (r = .307, p < .001) proportion of the variance of looking. In comparison, a null model, which includes only a side bias (b) and inertia (p) term for each infant picks up a significantly small proportion of the variance (r_g = .307, r_n = .203, t = 2.68, p = .01).

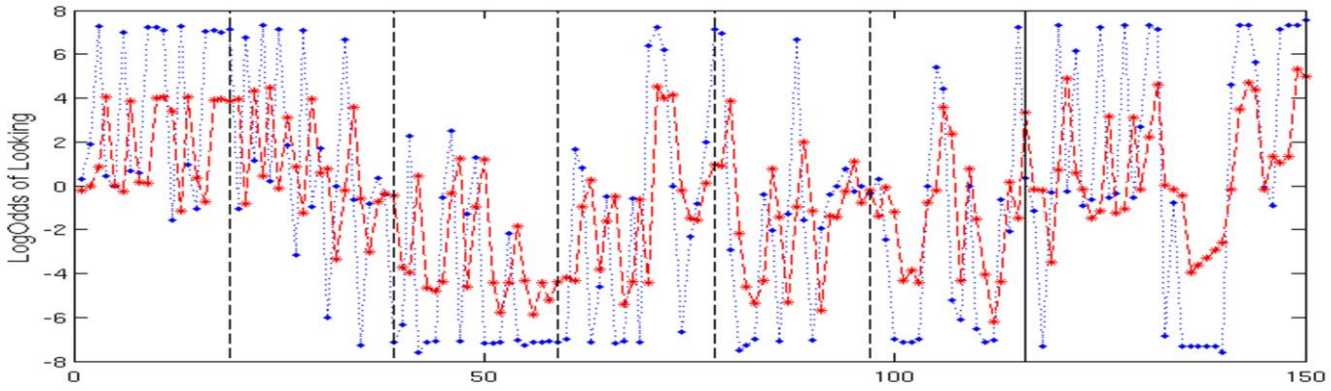


Figure 4: Log odds looking to the left side of the screen for one infant across both training and testing. Positive log odds indicate a preference for the left. Black dashed lines separate every 10 training trials and the black solid line indicates the start of testing. Infant behavior is the blue line with solid markers, model behavior is the red line with asterisk markers.

This indicates that *habituation* and *association* account for a significant proportion of each infant's looking behavior, both in training and testing. Further, functions which are appropriate for describing training can also describe test behavior. An example of the model's fit to one infant is shown in Figure 4 above.

Now that we have found the best model which accounts for each infant's looking behavior, we can determine which infants are likely to have learned word-referent mappings. Preferential looking behavior, while a good measure of learning at a group level, can be quite difficult to interpret at the individual level (Aslin, 2007; Houston-Price, Nakai, 2004; Hunter & Ames, 1988). There are several reasons for this. First, as mentioned above, the function which links learning to looking is *nonmonotonic*, and different infants learn at different rates. Hence whether preference for target or distractor should indicate learning in an individual infant is unclear. Second, as we have explicitly modeled, there are two principled reasons to move one's eyes in this task – in response to the objects on the screen (habituation), and in response to the relationship between objects and words (association). If we are interested in word-object mapping, then movement resulting from the first process adds noise to our measurement. Because both processes were modeled explicitly, however, we can probe association directly.

For each infant, model selection was used to determine which order polynomial best matched his or her association and habituation functions. If the optimal order of association for an infant was nonzero, then we can infer that the infant learned associations between words and objects. Thus, another way to measure whether an infant learned word-object associations is to ask about the order of that infant's association function. Of the 25 participants, 11 were best described as being driven by an associative process ($N_a > 0$). We can then look at what these association functions predict in the infant's test behavior.

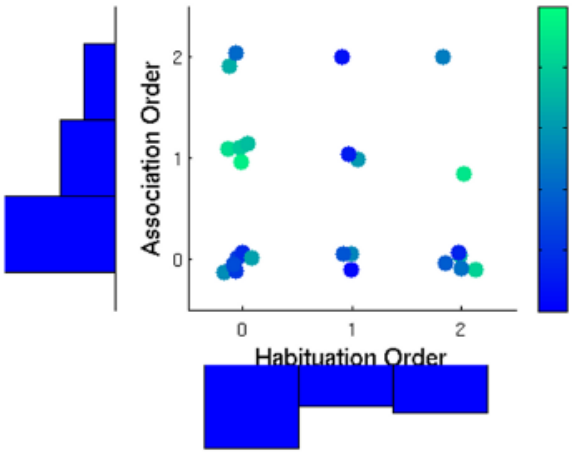


Figure 5: Distribution of association and habituation orders which best account for each infant. The scale on the right ranks infants from strongest preference for distractor (bottom) to strongest preference for target (top). Association order is correlated with strength of absolute preference.

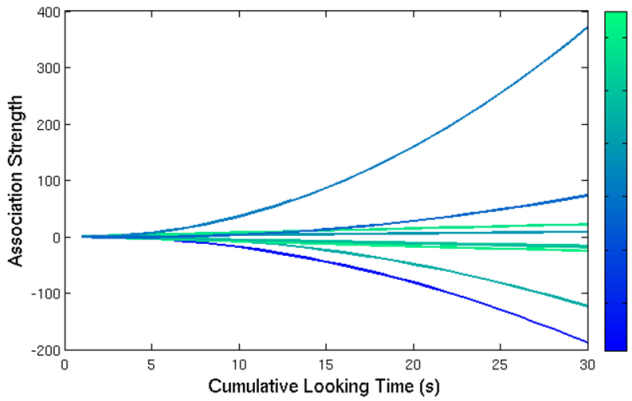


Figure 6: Theoretical association functions for each infant plotted over the course of 30 seconds of co-occurrence. The scale on the right ranks each infant by the strength of their preference for target or distractor. Throughout the entire 30 seconds, there is a significant correlation between the strength of the preference and the strength of the theoretical association function.

Figure 5 shows the distribution of association and habituation orders of the polynomial functions which best accounted for each individual infant's looking behavior. Points representing individual infants are color-coded by the strength of their preference for target(green) or distractor(blue). Analysis shows that the order of an infant's association function is strongly correlated with the strength of that infant's absolute mean preference in the 6 preferential looking trials (r = .55, p < .01). That is, the stronger an infant's preference at test (either for target or distractor), the higher the order of the association function that best described his or her data.

Second, because we have explicitly determined the polynomial function which best describes each infant's association learning, we can examine these functions in isolation. Figure 6 shows the association function for each infant plotted over 30 1 second exposures to a hypothetical word and object. We can compare the ordering of these functions – rank them in the order of their value after each window – and compare this to the mean preference for the target exhibited by each infant over the 6 test trials. The correlation between ordering and mean preference is significant at the .05 level over the entire course of the comparison, and peaks at four seconds (r_4 = .771, p < .001). This finding indicates that these theoretical learning functions are in deeply linked to preferential looking performance at test. The functions thus allow us to predict which infants will show *familiarity* preferences at test, and which infants will show *novelty preferences*. The two figures also reinforce the fundamental importance of understanding individual differences if we are to understand statistical learning. The 25 individual infants displayed the entire gamut of possible learning functions, and these functions fit sensibly to their looking performance at test.

General Discussion

Learning word-referent associations in cross-situational experiments, and in the world, must depend on moment-to-moment *behavior* of *individual* infants – what is looked at and when – and the co-occurrence of objects seen and words heard. Looking behavior, in turn, depends on previous experience in multiple ways and through multiple mechanisms. Two of these fundamental mechanisms are *habituation* and *association.* Repeated experience with an object increases the tendency to look away, but repeated experience with the object in a word's context increases the tendency to look towards the object in its presence.

The present analyses show what can be gained by attempting to understand the dynamic processes that underlie the *behaviors* used as indices of learning. Constructing trial-by-trial models of individual infants looking behavior in word-referent learning yields two major benefits. First, since looking behaviors themselves are commonly used as indices of learning, it allows us greater certainty in inferring learning in infants. Second, because we can track individual infants across the course of learning, it gives us a deeper theoretical understanding of how the mechanisms underlying this learning.

This work thus makes both specific and general contributions. First, the generative model of eye movements in cross-situational learning explains individual infant behavior in both training and testing. Second, we have delineated the interacting effects of two competing processes which produce infant eye fixations – habituation and association – and showed how they can be analyzed independently. Third, our experiment and model demonstrate the possibility of understanding cross-situational learning at the individual infant level, of making sense of the different ways in which different infants learn. Finally, we have demonstrated a novel methodology for analyzing infant learning tasks. In addition to the insight gained from preferential looking analysis, this work shows a promising role for model selection and the construction of explicit functions linking learning to looking.

Acknowledgments

This work was supported by a National Science Foundation Graduate Research Fellowship to the first author and National Institute of Health Grant R01HD056029. The authors would also like to thank Amara Stuehling and Melissa Elston for collecting much of the data. Finally, the authors would like to thank Mike Frank for thoughtful feedback on an earlier draft.

References

Aslin, R. N. (2007). What's in a look? *Developmental Science*, *10*, 48-53.

Barr, D. J. (2008). Analyzing 'visual world' eyetracking data using multilevel logistic regression. *Journal of Memory and Language, 59*, 457-474.

Cashon, C. H., & Cohen, L. B. (2000). Eight-month-old infants' perception of possible and impossible events. *Infancy*, *1*, 429–446.

Fazly, A., Alishahi, A., & Stevenson, S. (in press). A probabilistic computational model of cross-situational word learning. To appear in *Cognitive Science*.

Golinkoff, R., Hirsh-Pasek, K., Cauley, K., & Gordon, L. (1987). The eyes have it: Lexical and syntactic comprehension in a new paradigm. *Journal of Child Language*, *14*, 23–45.

Hollich, G., Hirsh-Pasek, K., & Golinkoff, R. (2000). Breaking the language barrier: An emergentist coalition model of word learning. *Monographs of the Society for Research in Child Development*, *65* (3, Serial No. 262).

Houston-Price, C. & Nakai, S. (2004). Distinguishing novelty and familiarity effects in infant preference procedures. *Infant and Child Development*, *13*, 341-348.

Hunter, M.A. & Ames, E.W. (1988). A multifactor model of infant preferences for novel and familiar stimuli. In L.P. Lipsitt (Ed.), *Advances in child development and behavior* (pp. 69–95). New York: Academic Press.

Landau, B., Smith, L. B., & Jones, S. S. (1988). The importance of shape in early lexical learning. *Cognitive Development*. *59*, 299-321.

Luce, R. D. (1959). *Individual choice behavior*. New York: Wiley.

Markman, E. M. (1990). Constraints children place on word meanings. *Cognitive Science*, *14*, 57-77.

Schöner, G. & Thelen, E. (2006). Using dynamic field theory to rethink infant habituation. *Psychological Review*, *113*, 273-299.

Smith, L. B. (2000). How to learn words: An associative crane. In R. Golinkoff & K. Hirsh-Pasek (Eds.), *Breaking the word learning barrier* (pp. 51 - 80). Oxford: Oxford University Press.

Smith, L. B. & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*, 333-338.

Stager, C. L., Werker, J. F. (1997). Infants listen for more phonetic detail in speech perception than in word learning tasks. *Nature*, *388*, 381-382.

Swingley, D. & Aslin, R. N. (2000). Spoken word recognition and lexical representation in very young children. *Cognition*, *76*, 147–166.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, *114*, 245–272.

Yu, C. & Smith, L.B. (in press). What you learn is what you see: Using eye movements to study infant cross-situational word learning. *Developmental Science*.

Simultaneous Cross-situational Learning of Category and Object Names

Tarun Gangwani, George Kachergis, and Chen Yu
{tgangwan, gkacherg, chenyu}@indiana.edu
Department of Psychological and Brain Science, and Cognitive Science Program
1101 East 10th Street Bloomington, IN 47405 USA

Abstract

Previous research shows that people can acquire an impressive number of word-referent pairs after viewing a series of ambiguous trials by accumulating co-occurrence statistics (e.g., Yu & Smith, 2007). The present study extends the cross-situational word learning paradigm, which has primarily been used to investigate the acquisition of 1-to-1 word-referent mappings, and shows that humans can concurrently acquire both 1-to-1 and 1-to-many mappings (i.e., a category relation), even when the many referents of a single word have no unifying perceptual features. Thus, humans demonstrate an impressive ability to simultaneously apprehend hierarchical regularities in their environment.

Keywords: statistical learning; cross-situational learning; category learning; mutual exclusivity; language acquisition

Introduction

In order to make sense of their world, human infants must learn relationships between words and referents in their environment. Infants simultaneously come into contact with many diverse, novel objects and equally diverse words that name them. Thus, there is much potential for acquiring erroneous word-referent mappings, given only a single situation. Despite this, both infants and adults have a remarkable ability to learn many novel word-referent associations quickly and accurately. Cross-situational word learning (CSWL) studies give us insight into how people are capable of learning multiple word-referent associations from individually ambiguous situations. Previous CSWL studies have shown that both infants and adults are able to learn simple 1-word to 1-referent mappings with astonishing speed (Kachergis, Yu, & Shiffrin, 2009; Smith & Yu, 2008; Klein, Yu, & Shiffrin 2008; Yu & Smith, 2007). In adult studies, participants are typically instructed to learn which words go with which referents, and are then presented with a few consistently co-occurring objects and spoken pseudowords on each of a series of training trials. On every trial, each pseudoword corresponds to a particular on-screen object, but the intended referent is never indicated. In a typical cross-situational training block, participants attempt to learn 18 word-referent pairs from 27 twelve-second trials consisting of four spoken words and four displayed objects (*i.e.* a 4x4 design). On average, participants in this condition managed to learn half of the 18 pairs by relying on cross-situational statistics (Yu & Smith, 2007). Further studies have shown that human learning often reflects statistics manipulated during training such as pair frequency, contextual diversity (the diversity of other pairs each pair appears with over time), and within-trial ambiguity (the number of co-occurring words and referents per trial) (Kachergis, *et al*., 2009).

However, simple 1-to-1 mappings are only a subset of the types of word-referent relations that exist in natural languages. 1-to-many mappings include referents that have one common label shared among them, such as a category or concept label. For example, both an apple and a banana may be labeled 'fruit.' Learners must learn to map both the superordinate label ('fruit') and each basic level name ('banana' 'apple') to the appropriate referent. Even in a learning paradigm like the 4x4 cross-situational learning condition discussed above, which is simpler than the real world, it is difficult to imagine that learners consider all 16 possible pairings, as might be necessary to learn higher-order relations. Constraints such as mutual exclusivity (ME) can drastically reduce the complexity of such ambiguous situations by limiting the possible pairings to a single word for each object (and vice-versa). Consider Markman and Wachtel's study (1988), in which a child was placed in front of a learned object (ball) and an unlearned object (gyroscope) and was prompted to retrieve the 'toma.' While 'toma' could be another name for the ball, the child moves to the unlearned object, exhibiting ME. However, despite its power to speed learning, the strict use of ME as a constraint in cross-situational learning would also make it impossible to learn non-1-to-1 mappings.

To determine whether learners use the ME constraint when learning names for previously unknown objects, Yurovsky and Yu (2008) presented learners with ME-violating mappings in the CSWL paradigm. An ME-violating mapping is a word (or object) that is consistently paired with more than one object (or word). Participants were trained on 12 words and 18 referents, where 6 words were paired with 12 referents (*i.e.*, 2 referents per word), known as *double* words, and the other 6 words were paired 1-to-1 with the remaining 6 referents, known as *single* words. Participants had to decide how to manage two names that co-occur with one referent in the same set of trials. The results showed that participants had equal performance in learning both single and double words if each double word's two referents were interleaved rather than temporally separated (*i.e.* one referent was shown in the first half only

and the other appeared only in the last half). Moreover, learners acquired more than half of both early and late pairings; thus, some must have violated ME.

In contrast, Ichinco, Frank, and Saxe (2009) presented participants with a study to demonstrate ME as a guide to learning word-to-referent associations. Participants were shown an additional referent (or word, in a different experiment) on each trial, alongside four previously-seen word-referent pairs. Both groups received training on a standard cross-situational task, which was followed by further training. In this training, a new stimulus (word or object, between groups) was added on each trial alongside four pairs from the early training. Rather than forming a 1-to-2 mapping with the additional object (or 2-to-1, for the extra word) on each trial, participants learned 1-to-2 (or 2-to-1) relations on average for only one item and consistently favored mutually exclusive mappings.

Thus, depending on how ME-violating word-referent mappings are added to the cross situational paradigm, learners vary their use of the mutual exclusivity constraint. In the Yurovsky & Yu study, additional referents are presented in the absence of old ones: when participants hear a word and see two referents consistently co-occurring with it, they may be more likely to violate ME and form a 1-to-2 mapping. In Ichinco, *et al*.'s study, all 1-to-1 mappings from the early stage occur simultaneously with the new mappings. Participants may have failed to learn the new mappings due to blocking, a known associative learning effect in which a previously learned pairing interferes with the acquisition of a new pairing involving old stimuli.

In the present study, in order to eliminate biases that participants may adopt as a result of training order, we provide participants with cross-situational training that is simultaneously consistent with both 1-to-1 (basic-level name to referent) and 1-to-many (superordinate-level name to multiple referents) relationships on every training trial. For example, in Experiment 1, on each 3x2 trial, two words are basic-level names for the visible referents, and a third will act as a superordinate-level identifier. These superordinate level labels hence refer to four referents, including two that are not on present on a given trial. Thus, participants are simultaneously faced with two labels for each referent, and one of these labels also applies to three other referents. In Experiment 2, we give learners a more complex learning scenario: 4x2 trials on which two labels map 1-to-1 to the objects and each of the other two labels refer to one of the present objects, and three unseen objects. In both experiments, participants must learn the unique name for a referent as well as a label it shares with three other objects, some of which are not present on a given trial.

One block in each experiment is composed of objects that share some unifying perceptual feature like a hook or arrow shape, somewhat like objects belonging to natural categories. We test for generalization using stimuli in which the objects share each category's identifying feature from training, but the objects have different textures and shapes than those from training.

Experiment 1

In Experiment 1, participants were merely instructed to learn which words go with which objects—with no mention of the potential to form 1-to-many relations—and were then given a sequence of cross-situational training trials, each consisting of three words and two referents. Unbeknownst to learners, two of the words on each trial map 1-to-1 to one of the visible referents, and the third word refers to both objects, and also will consistently appear with two other referents during training. Participants must determine which words specify a 1-to-1 reference to an object and which word specifies a 1-to-many reference to both objects on each trial. If participants assume ME, participants will either learn 1-to-1 mappings or 1-to-many mappings, but not both.

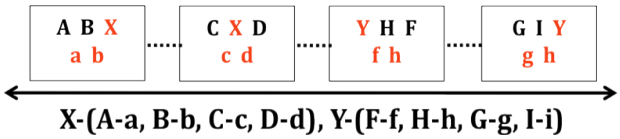


Figure 1: In Experiment 1, participants are trained on both 1-to-1 (e.g., A-a and B-b) and 1-to-many mappings (e.g., X-{a,b,c,d}) in the context of 3 words and 2 referents per trial. One word is the superordinate-level name that refers to both referents on each trial (shown in red).

In order to see if 1-to-many associations are facilitated by stimuli structure, subjects were trained on two different conditions (in three blocks in fixed order): **Block 1** was an *arbitrary category* condition, in which the objects had no obvious shared perceptual features but were consistently labeled by some other word. **Block 2** was a *natural category* condition, in which the objects in each category share a salient feature (e.g., a hook or arrow shape). **Block 3** was another *arbitrary category* condition (with different stimuli) to gauge attention shift after learning natural 1-to-many groupings. Given the salient features present in Block 2, performance in learning 1-to-many relationships will likely increase relative to Block 1, as participants' attention will be drawn to the 1-to-many relations due to the salient features acting as learning cues. Their performance on block 3 will indicate if this attentional shift is carried over from the natural category block.

Subjects

Participants were 33 undergraduates at Indiana University who received course credit for participating. None had participated in other cross-situational experiments.

Stimuli

Each training trial consisted of two objects shown on a computer screen and three pseudowords played sequentially. In each of the two arbitrary category conditions, the 12 referents were difficult-to-name, unrelated objects. For the

natural category condition, the 12 objects had one of three features protruding from the shape. The 45 computer-generated pseudowords are phonotactically-probable in English (e.g. "stigson"), and were spoken by a monotone, synthetic voice. 36 words are assigned to each referent, creating arbitrary word-object pairs which were randomly assigned to three sets of 12 1-to-1 mappings. One set of stimuli composed the natural category stimuli for the second block; the other sets composed of arbitrary strange objects for the first and third blocks. For the 1-to-many mappings, the remaining 9 pseudowords are assigned to three sets of four 1-to-1 mappings. Thus, in each block there are three groups (i.e., categories).

Words

Referents	X (12)				Y (12)				Z (12)			
	A	B	C	D	E	F	G	H	I	J	K	L
a	6	2	2	2	0	0	0	0	0	0	0	0
b	2	6	2	2	0	0	0	0	0	0	0	0
c	2	2	6	2	0	0	0	0	0	0	0	0
d	2	2	2	6	0	0	0	0	0	0	0	0
e	0	0	0	0	6	2	2	2	0	0	0	0
f	0	0	0	0	2	6	2	2	0	0	0	0
g	0	0	0	0	2	2	6	2	0	0	0	0
h	0	0	0	0	2	2	2	6	0	0	0	0
i	0	0	0	0	0	0	0	0	6	2	2	2
j	0	0	0	0	0	0	0	0	2	6	2	2
k	0	0	0	0	0	0	0	0	2	2	6	2
l	0	0	0	0	0	0	0	0	2	2	2	6

Figure 2: The accumulated stimulus co-occurrence matrix for each block in Experiment 1. Each word co-occurred with its intended referent 6 times (A-a, B-b, …) Note that each referent appeared twice with every other referent in its category, but never with referents from other categories. Each 1-to-many label appeared 6 times with each of its intended referents, and 12 times overall.

In the natural category condition, each of the three 1-to-many labels consistently maps to a salient feature present on the stimulus. An additional 12 pairs of testing stimuli were used for a generalization task, using the same category labels that correspond with the stimuli according to their feature.

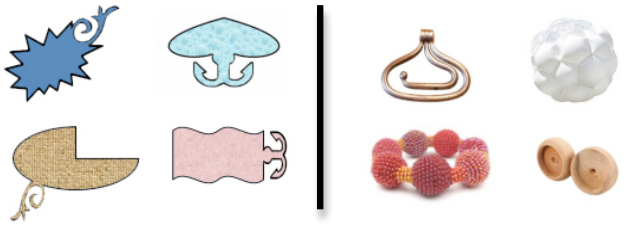


Figure 3: **Left**: In the natural category condition, objects with multiple types of textures and three different protruding shapes were used in training. **Right**: In the arbitrary category condition, objects had no apparent unifying feature.

Procedure

Participants were informed that they would experience a series of trials in which they would hear some words and see some objects. They were also told that their knowledge of which words belong with which objects would be tested at the end. Training for each condition consisted of 36 trials. Each training trial began with the appearance of two objects, which remained visible for the entire trial. After 2 s of initial silence, each word was heard (randomly ordered; 1 s of silence between each word) followed by 2 s of silence, for a total of 9 seconds per trial. After each training block, their knowledge was assessed using 12-alternative forced choice (12AFC) and 3AFC testing: on each test trial a single word was played—a 1-to-1 label or a 1-to-many label—and the participant was asked to choose the appropriate object from a display of all 12 objects (for 1-to-1 labels) or from 3 objects (for 1-to-many labels). For 3AFC testing, one representative from each category was used. The test slides for generalization were the same as the 1-to-many test slides except that the only previously-seen parts of the stimuli were the distinct, protruding shapes (e.g., a hook) that were seen in training to distinguish the different categories. Different stimuli were used in each block. Condition order was fixed.

Results & Discussion

Figure 4 shows the results across all three blocks for each pairing type. Unexpectedly, even in block 1 participants learned a significant number of 1-to-many mappings (M = .49, one-sided $t(32) = 4.95$, $p<.001$, chance=.33) and learned a significant proportion of 1-to-1 mappings (M = .52, one-sided $t(32) = 12.99$, $p<.001$).

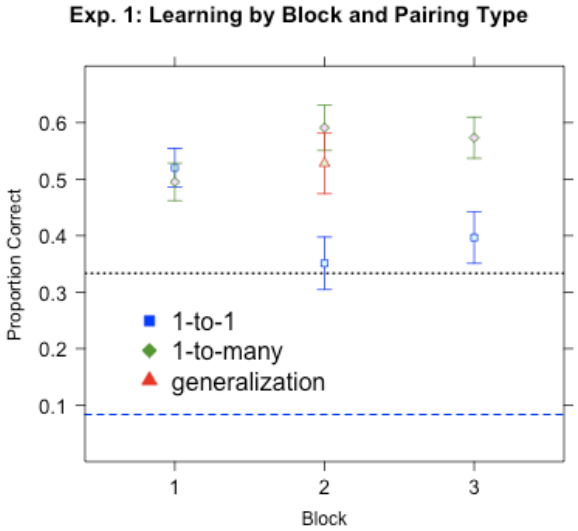


Figure 3: Mean performance for each experimental block by pairing type. Block 1 and 3 were arbitrary groupings and Block 2 was a category grouping; thus, generalization of category type was tested. Error bars show +/-SE. Blue dotted line indicates chance for 1-to-1 learning; black dotted line indicates chance for 1-to-many learning.

After the introduction of a unifying feature, learning of 1-to-1 pairings in block 2 decreased relative to block 1 (M = .35, paired $t(32) = 3.07$, $p<.01$). The perceptual similarity of category members in block 2 may have caused participants to focus on learning 1-to-many mappings, and thus drew attention away from 1-to-1 mappings. In addition, their ability to apply the superordinate name to new referents was reflected in their significantly above-chance (.33) performance on a generalization task (M = .53, one-sided $t(32) = 3.78$, $p<.001$).

Presented with a second arbitrary category condition in block 3, learning of 1-to-1 pairings was significantly lower compared to block 1 (paired $t(32) = 2.96$, $p<.01$), but performance on 1-to-many testing remained higher (M = .57, paired $t(32) = 6.69$, $p<.001$). That is, following the natural category condition in block 2, participants continued to focus on 1-to-many mappings, but still learned 1-to-1 mappings at a proportion over three times chance.

Overall, participants showed evidence of learning both 1-to-1 and 1-to-many mappings in every condition—even in the first condition, when they had no instructions telling them what type of relations would be present, and the referents belonging to each 1-to-many relation (i.e., an arbitrary category) had no unifying perceptual features. Moreover, we observed a shift in learning from block 1 to block 3: after the perceptually-similar category referents of block 2, participants learned more 1-to-many pairings in block 3 than block 1, and fewer 1-to-1 pairings. In Experiment 2, we investigate whether learners can still simultaneously acquire both 1-to-1 and 1-to-many mappings in a still more complex learning situation.

Experiment 2

Experiment 1 showed that humans can simultaneously learn superordinate and basic level names for referents. On each trial, there were two basic level names (1-to-1) and one superordinate level name (1-to-many). Thus, the mutual exclusivity constraint was relaxed and complex relations were formed, with two words referring to each object. After all three conditions in Experiment 1, participants still performed significantly above chance on 1-to-1 associations as well as on 1-to-many associations. However, an alternative learning scenario is an environment in which objects from different categories are learned simultaneously. For example, two referents such as an apple and a carrot could be presented. In this case, each referent has its own superordinate level name (fruit and vegetable, respectively). The learner would need to learn both the superordinate label and basic name label for each object while needing to assign each term to its appropriate referent. The potential for error is much greater because the learner is presented with a more ambiguous learning situation than in Experiment 1, where the superordinate label refers to both displayed referents. Experiment 2 thus presents learners with a four word and two referents (*i.e.* 4x2) on each trial, where two words are category labels referring to a single referent each, and two words are subordinate level names corresponding to one referent each (see Figure 5).

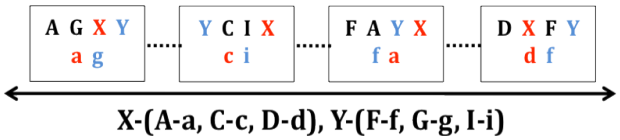


Figure 5: Participants are given 1-to-1 and 1-to-many mappings (e.g. A-a, C-c and X-{a,c,d}) in the context of 4 words and 2 referents per trial.

This extension of the cross-situational paradigm provides additional ambiguity beyond Experiment 1: presented with two more labels than referents on each trial, participants must now learn that the more frequent labels are superordinate, and apply not only to one of the objects on that trial, but also to three other objects seen on other trials. However, given the above-chance performance and particularly exceptional 1-to-many learning, participants may be able to tune themselves into the ambiguous superordinate label to referent pairings after the natural category condition in a manner similar to participants in Experiment 1.

Subjects

Participants were 24 undergraduates at Indiana University who received course credit for participating. None had participated in other cross-situational experiments, including the previous experiment.

Stimuli & Procedure

During training, two objects were shown on a computer screen with four spoken words played sequentially upon presentation of the objects, with time per word equal to that of Experiment 1. New sets of words and referents were used for this experiment. Training for each condition consisted of 36 trials, each lasting 12 s. due to the addition of a spoken category label. Immediately after training for each block, participants were tested for knowledge of the 1-to-1 relations using 12AFC and 1-to-many relations using 3AFC as in Experiment 1. Generalization was also tested for the natural category stimuli. Condition order was fixed.

Results & Discussion

Figure 6 shows results across all three blocks for each pairing type. In Block 1 with arbitrary category referents, participants learned only 1-to-1 names (M = .50; one-sided $t(23) = 7.76$, $p<.001$) while 1-to-many performance was at chance (M = .39, one-sided $t(23) = 1.76$, $p>.05$). Unlike in Experiment 1, block 2 did not see a performance shift. While performance was significant in learning 1-to-1 (M = .46; one-sided $t(23) = 6.45$, $p<.001$) associations, 1-to-many associations were still difficult to acquire, and were not learned significantly above chance (M = .42; one-sided $t(23)$ = 1.85, $p>.05$). Participants may have still not surmised that there was categorical structure involved due to the

confusion of four words per trial (including two category labels). Performance on the generalization task was also at chance, confirming that participants had not yet ascertained the presence and structure of the 1-to-many mappings.

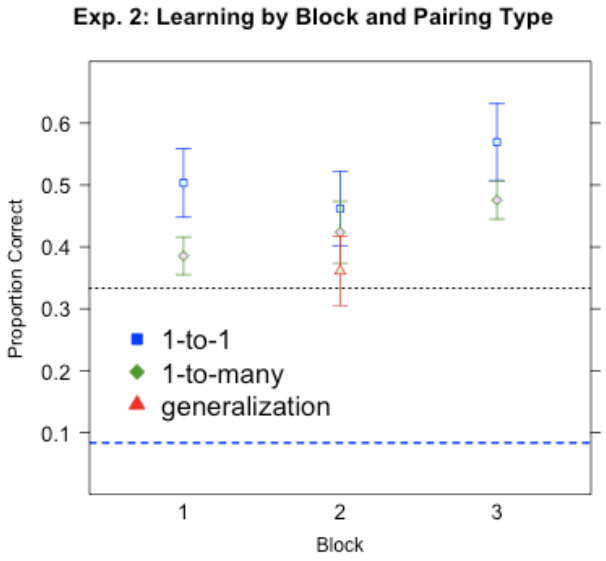


Figure 6: Mean performance by pairing type for each block. Error bars show +/-SE. Dotted lines indicate chance: blue for 1-to-1 pairings (.08); black for 1-to-many pairings (.33).

However, block 3 performance was significantly above chance for both 1-to-1 (M = .57; one-sided $t(23)$ = 7.98, $p<.001$) and 1-to-many (M = .46, one-sided $t(23)$, $p<.001$) associations. Thus, although the higher degree of ambiguity in Experiment 2 made participants take longer to catch on to the presence of multiple superordinate labels on each trial, in the final block they were able to learn these 1-to-many relationships in addition to the 1-to-1 relationships. In comparison to block 3 of Experiment 1, 1-to-many learning in Experiment 2 was significantly lower (Welch's t(55.0) = 2.08, $p<.05$), showing that the superordinate label structure (2 per trial) in Experiment 2 was indeed harder than the structure (1 superordinate label per trial) in Experiment 1.

However, even when participants were uncertain about the meaning of the superordinate labels in blocks 1 and 2, they learned a significant number of 1-to-1 mappings. In block 3 performance, not only did participants learn a significant number 1-to-many mappings, they also learned more 1-to-1 mappings than in the previous two blocks (block 2: paired $t(23)$ = 2.44, $p<.05$, block 1: paired $t(23)$ = 2.03, p=.05). The natural category condition once again provided a clue as to what learning strategy participants need to utilize. However, the significantly lower performance for block 1 in Experiment 2 as compared to Experiment 1 may also indicate interference due to confusion over the two extra labels.

In both experiments, it is important to note that since both 1-to-1 and 1-to-many word-referent mappings learned involving the same referents, each referent was thus part of a 2-to-1 word-referent mapping. Thus, it is possible to determine whether participants learned mappings that violate mutual exclusivity. In both experiments, participants were tested on each referent twice: for the 1-to-1 label (chance=1/12) and 1-to-many label (chance=1/3). Thus, learning that respects ME occurred when participants learn either 1-to-many or 1-to-1 mappings, but not both, and learning that violates ME occurred when participants learn both. As shown in Figure 7, across both experiments and in every block, the average participant learned a significant number of pairings that violate ME as they learned both 1-to-1 and 1-to-many mappings.

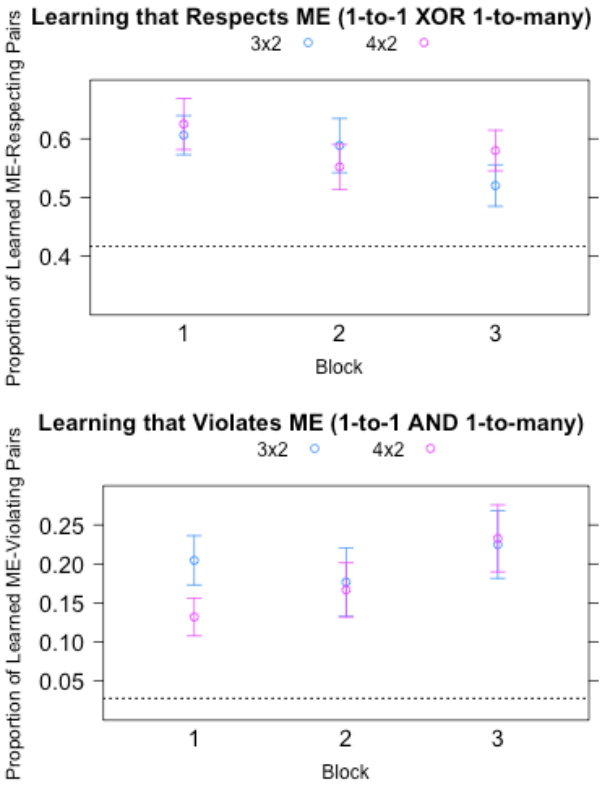


Figure 7: Comparison of proportion of learned ME violating vs. respecting pairs by block for each experiment. Chance (dotted line): Respects=1/3+1/12; Violates=1/3•1/12=.03

General Discussion

While the mutual exclusivity constraint can be a powerful tool for learning 1-to-1 mappings, the hierarchical structure of the real world—which is reflected in natural language—requires people to learn word-referent mappings that are not mutually exclusive. The present study demonstrates that learners learn both 1-to-1 and 1-to-many mappings from situations in which these regularities are simultaneously present.

By the end (block 3) of both experiments, performance for both 1-to-1 and 1-to-many testing was significantly above chance. Experiment 1 shows that participants on

average performed strongly on 1-to-many associations, particularly after the introduction of within-category perceptual similarity in block 2. This may be due to the natural stimuli serving as a primer for learning 1-to-many mappings in block 3. However, although there appears to be a trade-off in learning both types of relationships, participants nevertheless managed to learn both superordinate and basic level names in the first block of Experiment 1. Experiment 2 showed participants could not only learn superordinate and basic level names but can also handle an additional layer of ambiguity when the two referents on a trial belonged to two different superordinate categories. Consistent with Experiment 1, an increase in 1-to-many performance was seen after block 2 was observed in Experiment 2, but 1-to-many performance was overall lower than in Experiment 1. Correspondingly, generalization of superordinate labels to novel objects was also difficult for learners. The more complicated structure (four labels and two referents per trial, representing two categories) in Experiment 2 produces many more possible pairings per trial for a learner to consider. Naturalistic learning situations are even more complex, with multiple co-occurring words, events, and objects (Hart & Risley, 1995); Experiment 2 simulates a more natural scenario in which multiple referents with vague relationships to their superordinate labels are presented. This suggests that infant learning of higher order relations could be guided by creating more unambiguous learning scenarios in order to reduce the likelihood of attribution error.

Interestingly, participants were equally likely to know the superordinate level names (*e.g.,* fruit) regardless of their performance learning basic level names *(e.g.,* apple). Is this due to the mutual exclusivity constraint? In the 3x2 design of Experiment 1, participants were more likely to form a 1-to-many relationship if they do not know the superordinate level name than if they know both (P(Know Superordinate Name | Not Know Basic Name) = .31; P(Know Superordinate Name | Know Basic Name) = .19). The same relationship held in the 4x2 design (.25, .18 respectively). Therefore, participants seemed to form superordinate level relationships more easily rather than basic level relationships.

While the ME constraint may be useful in learning 1-to-1 relationships, the present study's experiments show that participants will focus on forming 1-to-many relationships rather than 1-to-1 relationships if the need to learn higher order relationships becomes apparent, which is often the case in category learning. The strong performance in 1-to-many learning independent of 1-to-1 performance may indicate that people are particularly tuned to learning complex relationships. Every day, we use categories as functional filters of our world to constrain the amount of information we must process at lower (basic) levels (Goldstone & Kersten, 2003). Furthermore, the addition of an exemplar to a category gives us more information about other novel candidate members of the category, allowing learners to generalize as demonstrated in Experiment 1. In future work, we hope to replicate our findings in infants as well as focus on what learning strategies are used by both infants and adults. We also expect that these findings will be useful in constraining formal models of cross-situational word learning.

Acknowledgements

This research was supported by National Institute of Health Grant R01HD056029.

References

Goldstone & Kersten (2003). Concepts and Categorization. In *Comprehensive handbook of psychology*. New Jersey: Wiley; pp. 599-621.

Hart, B. & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children*. Baltimore, MD: Brookes.

Ichinco, D., Frank, M., & Saxe, R. (2009). Cross-situational Word Learning Respects Mutual Exclusivity. In N. Taatgen, H. van Rijn, J. Nerbonne, & L. Schomaker (Eds.) *Proceedings of the 31st Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Kachergis, G., Yu, C., & Shiffrin, R. M. (2009). Frequency and Contextual Diversity Effects in Cross-Situational Word Learning. In N. Taatgen, H. van Rijn, J. Nerbonne, & L. Schomaker (Eds.) *Proceedings of the 31st Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Klein, K. A., Yu, C., & Shiffrin, R. M. (2008). Prior knowledge bootstraps cross-situational learning. *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, 1930-5. Austin, TX: Cognitive Science Society.

Smith, L. & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, 106, 1558-1568.

Yu, C. & Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, 18, 414-420.

Yurovsky, D., & Yu, C. (2008). Mutual exclusivity in cross-situational statistical learning. *Proceedings of the 30th Annual Meeting of the Cognitive Science Society*, 715–720. Austin, TX: Cognitive Science Society.

Extending Beyond Space

Brooke O. Breaux (brookebreaux@louisiana.edu)
Institute of Cognitive Science, University of Louisiana at Lafayette,
Lafayette, LA 70504 USA

Michele I. Feist (feist@louisiana.edu)
Institute of Cognitive Science, University of Louisiana at Lafayette,
Lafayette, LA 70504 USA

Abstract

Investigations into the semantics of the spatial and non-spatial uses of *in* and *on* have tended to assume that a type-level similarity exists between these two prepositions. However, their syntactic distributions, while overlapping, are not equal in scope (Navarro, 1998). In this paper, we ask whether these distributional differences might be related to semantic differences between the two terms. The preliminary evidence collected here suggests that *in* and *on* have slightly different levels of interpretability, even in their prepositional uses. Thus, both semantically and syntactically, the assumption of type-level similarity may need to be qualified.

Keywords: Semantics; prepositions; metaphor; language

Introduction

Investigations into the semantics of prepositions such as English *in* and *on* have tended to treat these lexical items as though they are different tokens of the same semantic and syntactic type. Such treatment seems to follow from the generative grammar tradition in which lexical category – rather than meaning – determines syntactic behavior. For example, Cook and Newson (2007) suggest "that arguments are interpreted in a particular way due to the structural positions they occupy" (p. 263). This assumption is also reflected in introductory linguistics and psycholinguistics text books, which state that words belonging to the same lexical category, or word class, are typically interchangeable syntactically (cf., Carroll, 2004; O'Grady, Archibald, Aronoff, & Rees-Miller, 2005). Together these suggest that different lexical items drawn from the same word class may interact with the rest of language in very similar ways.

Even in more cognitive views of language, we find evidence that prepositions are treated as a lexical category without indication that the individual differences between the prepositions will have important repercussions for the functions of the individual lexical items within the linguistic system. As a case in point, type-level equivalence has been assumed in examinations of the semantics of prepositions (e.g., Coventry & Garrod, 2004; Feist, 2000, 2008, in press; Feist & Gentner, 2003; Tyler & Evans, 2003; Vandeloise, in press). Much of this work focuses on the criteria that distinguish the meaning of one preposition from that of another, without discussion of the possibility that prepositions may differ in additional ways beyond their meanings. For example, while Tyler and Evans (2003) do acknowledge the importance of context in establishing the meaning of a lexical item and the fact that different prepositions will occur in different contexts, such contextual factors do not lead to different proposals regarding the nature of the meanings of individual prepositions.

However, evidence from corpus-based studies of prepositions challenges this assumption of distributional equivalence. For example, in his investigation into the semantic structure of English topological prepositions, Navarro (1998) found a differentiation between *in* and *on* based not only on their meanings but also on their syntactic distributions. While *on* tends to occur primarily in prepositional constructions, *in* is also quite prevalent within "a wide range of morphosyntactic usages that make it controversial to categorise it on behalf of a single syntactic construction" (Navarro, 1998, p. 273), including use as a full adverb, as an adverbial particle of a phrasal verb, and as a prefix for nouns, adjectives, and adverbs. This difference in syntactic distribution suggests that, despite their similarity as topological prepositions, *in* and *on* may behave quite differently within the language system as a whole.

Following up on these observations, we searched for uses of *in* and *on* in the more than 400 million word Corpus of Contemporary American English (COCA; www.americancorpus.org). Our first observation was that the frequencies of occurrence of *in* and *on* are highly unequal overall, with *in* (7,333,378 instances) appearing more than 2½ times more frequently than *on* (2,723,768 instances). Secondly, and more importantly, we examined the combinatorial possibilities for both *in* and *on* across a set of naturally occurring uses within a limited syntactic context (i.e., prepositional phrases containing the preposition immediately followed by a noun). Within the hundred most frequent collocations for each preposition, we observed an inequality in the distribution of uses, $\chi^2(1, N = 200) = 21.34$, $p = .0003$ (see Table 1).

Table 1: Noun types collocating with *in* and *on*

	Proper Nouns	Noun Phrases	Idioms	Concrete Nouns	Abstract Nouns
In	2	6	5	33	54
On	2	23	2	45	28

Taken together, these results suggest an imbalance between *in* and *on* that has yet to be thoroughly investigated.

Clearly, *in* and *on* have different meanings, which will result in the two prepositions collocating with different sets of nouns. However, these differences have not thus far led to a challenge to the assumption of type-level similarity based on their shared lexical category. As such, the differences in distribution and in combinatorial possibility that have been observed in corpus-based studies of *in* and *on* remain unexplained by the current state of thinking regarding their meanings.

There are two possible explanations for the observed differences between *in* and *on*. First, it may be that the differences are an artifact of the searches that yielded them, and that these differences would disappear given a large enough sample drawn from the corpus. In this case, the assumption of type-level similarity would remain intact, with the differences, which would be attributable to differences in meaning, limited to differences in the sets of nouns that collocate with each, but not to differences in the sizes of the sets or in the ranges of meaning types within the sets.

The second possibility is that *in* and *on* differ not only in meaning, but in *meaning potential*, with *in* able to collocate with a wider range of nouns than can *on*. In this case, the particular semantics of *in* and *on* will have a direct influence on their potential to combine with other lexical items, rather than that potential being determined by their belonging to the lexical class of prepositions, and the assumption of type-level similarity inherited from generative grammar will need to be abandoned.

In order to discriminate between these two explanations, we will seek evidence regarding the reality of the noted imbalance using a separate methodology. If the evidence gathered from an experimental investigation of the combinatorial possibilities of *in* and *on* fails to replicate the corpus evidence, this would support a type-level similarity-based explanation wherein the noted imbalance is an artifact of the corpus searches performed. If, on the other hand, the experimental data replicates the imbalance noted in the corpus, this would support the explanation that the range of combinatorial possibilities of a lexical item is not determined by its lexical class. Rather than having their influence limited to the specific referential situations within which prepositions are deemed appropriate, meaning differences may significantly determine prepositions' ranges of combinatorial possibilities.

In order to experimentally examine the combinatorial possibilities displayed by the prepositions *in* and *on*, we asked English speakers to interpret prepositional uses of *in* and *on* presented in the same novel syntactic and semantic contexts (i.e., the same novel sentence frames). If there is an imbalance in the combinatorial possibilities of these prepositions, then we should see different levels of interpretability for the two prepositions. To be clear, while we anticipate their different meanings to result in different interpretations of the sentences, if there are indeed differences in *interpretability* these should be evident in the rates at which participants attempt to provide interpretations for the novel sentences. Such an imbalance in interpretability between these lexical items when presented in identical sentence frames would suggest that the assumption of type-level similarity within lexical classes is unwarranted.

Experiment 1

Experiment 1 tested whether novel non-spatial uses of the preposition *in* would be more easily interpretable than matched non-spatial uses of the preposition *on*. If so, participants should make more attempts to interpret sentences containing *in* than sentences containing *on*.

Method

Participants A total of 82 UL Lafayette students participated in this experiment in exchange for course credit. One student, a native speaker of Vietnamese, was subsequently removed from further analysis; a second participant was removed for not following the task instructions. The 80 remaining participants were all native speakers of English. Of these, 39 took part in the *in* condition and 41 took part in the *on* condition.

Materials The stimuli consisted of forty sentences constructed from twenty sentence frames. Sentence frames were in the form *These Xs are Y*; each *Y* was a non-spatial prepositional phrase (i.e., *in* or *on* followed by an abstract noun), and each *X*, a concrete noun. Each sentence frame had both an *in* variant and an *on* variant (see Table 2).

In order to provide the prepositions with a neutral playing field, the sentence frames needed to constitute unfamiliar contexts for both prepositions under consideration. At the same time, we wanted the interpretability of each sentence as a whole to hinge on the interpretability of its prepositional phrase. Thus, in constructing our sentences, we (1) selected abstract nouns that would be considered unfamiliar prepositional objects for both *in* and *on* and (2) chose as the sentential subjects nouns which would be as stable in their meanings as possible.

To accomplish these goals, we searched for twenty abstract nouns that do not frequently occur as objects of either *in* or *on*. Francis & Kučera's (1982) rank list of lemmas was used to formulate a list of highly frequent nouns from which we could extract 100 that could potentially serve as abstract prepositional objects. Beginning with the most frequent lemma, one of us (B.B.) categorized each noun as either concrete or abstract. Nouns were categorized as concrete if they could refer to a concrete object, a concrete set of objects, a part of a concrete object, or the location of a concrete object; otherwise, they were labeled as abstract and set aside for potential use as a non-spatial prepositional object. Two-hundred and twenty-eight nouns had to be categorized in order to find 100 that fit the abstract criterion.

Table 2: The twenty sentence frames used to construct the experimental stimuli.

	Sentence frames
1	These houses are in/on system.
2	These rooms are in/on reason.
3	These cars are in/on idea.
4	These streets are in/on result.
5	These lights are in/on month.
6	These books are in/on hour.
7	These roads are in/on sense.
8	These tables are in/on moment.
9	These pictures are in/on voice.
10	These walls are in/on century.
11	These buildings are in/on situation.
12	These plants are in/on term.
13	These windows are in/on difference.
14	These floors are in/on statement.
15	These radios are in/on feeling.
16	These boats are in/on organization.
17	These parks are in/on basis.
18	These mountains are in/on event.
19	These blocks are in/on opportunity.
20	These apartments are in/on association.

We then compared our concreteness categorization with concreteness judgments collected from the MRC Psycholinguistic Database (Wilson, 1988; http://www.psy.uwa.edu.au/mrcdatabase/uwa_mrc.htm) for each of the 228 categorized nouns. Of the 201 queries that resulted in concreteness ratings (concrete, n = 117; abstract, n = 84), the mean concreteness rating for the nouns we labeled as concrete (M = 507.99) was significantly higher than for the nouns we labeled as abstract (M = 357.51; F (1,199) = 175.52, p < .0001).[1]

Next, we searched COCA for combinations of *in* and *on* with each abstract noun. The 20 abstract nouns chosen for the experiment were those for which (1) combinations with both *in* and *on* produced frequency totals lower than 100 and (2) the absolute differences between the frequencies of combinations with *in* and *on* was at a minimum. A post-hoc one-way ANOVA revealed that the average frequency of *in* combinations (M = 32.70) was not significantly different from the average frequency of *on* combinations (M = 23.75), F (1, 38) = 2.77, p = .1040.

Because we wanted the interpretability of our sentences to hinge on the prepositional phrase and, hence, the compatibility of the preposition and the abstract noun, we needed the other content words to be more stable in their meanings. Previous research has suggested that object terms may be more stable in their meanings than other terms (Feist & Cifuentes Férez, 2007; Gentner & Asmuth, 2008; Gentner & France, 1988). Thus, only count nouns that were considered by the experimenters to normally refer to inanimate concrete objects – especially when considered as a group of objects (i.e., when the noun is preceded by the adjective *these*) – were selected for use as sentential subjects.

Finally, to ensure that the *in* variants and the *on* variants of our resultant sentences were equally novel, a frequency search was conducted in COCA for each of the subject noun-prepositional phrase combinations (e.g., *house* together with *in system*). This search revealed that none of the final combinations appeared in the corpus.

Procedure Participants were randomly assigned to interpret either the *in* variants or the *on* variants. They were presented with all twenty sentences in their assigned condition in a randomized order on a computer screen. For each sentence, they were asked to either explain its meaning in the text box provided or, if they were unable to formulate a meaningful interpretation, to simply type *uninterpretable* in the text box instead of an interpretation.

Design We used a 2 (Preposition: *in* or *on*) x 20 (Sentence Frame) design with preposition as a between-subjects factor and sentence frame as a within-subjects factor.

Analysis and Results

Stimuli Check Before turning to our results, we ask whether the sentential subjects were less likely to shift in meaning within the context of the sentences than were the objects of the prepositions, as required by the design. To test this, we calculated for each sentence (e.g., *These houses are in/on system.*) the proportion of times the nouns used as sentential subjects (e.g., *houses*) and those used as prepositional objects (e.g., *system*) in the stimulus sentences were reproduced in the participants' interpretations. A one-way ANOVA revealed that sentential subjects were reproduced in interpretations significantly more often (M = .74, SD = .12) than their prepositional object counterparts (M = .29, SD = .17; F (1, 38) = 94.25, p < .0001), suggesting that any differences in the interpretability of the sentences would have more to do with interpretation of the prepositional phrases than with interpretation of the subjects within the wider context of the sentence, as was required by our experimental design.

Interpretability A repeated measures ANOVA on sentence interpretability revealed a significant main effect of sentence frame, F (19, 60) = 10.89, p < .0001, indicating that participants found some sentence frames to be more interpretable than others. Because the interpretability of the sentences was dependant on the interpretability of the prepositional phrases, this result suggests that the abstract nouns were not equally interpretable as objects of the prepositions.

Of greater relevance to the question of differences in combinatorial possibilities between the two prepositions, we observed a marginally significant sentence frame by

[1] Of the 27 noun queries that did not result in concreteness ratings, 11 were labeled as concrete and 16 as abstract.

preposition interaction, F (19, 60) = 1.72, p = .0574. Post-hoc t-tests comparing the proportion of participants willing to provide interpretations between conditions for each sentence frame, individually, revealed four significant differences in which interpretability was higher for participants in the *in* condition than for participants in the *on* condition and no significant differences in the opposite direction.

Although the ANOVA did not reveal a significant main effect of preposition, we did observe a trend in the predicted direction whereby participants who interpreted *in* sentences were more likely to provide interpretations (M = .58, SD = .19) than participants who interpreted *on* sentences (M = .51, SD = .28). Furthermore, we note that the lack of a significant difference between the two conditions may have been driven, in part, by the large variances in interpretability of the two groups. Therefore, we were interested in any broader patterns in the data that might be hidden within or beneath this high variability.

We turn first to the variances of interpretability for the two groups of participants. While a significant difference between the variances of interpretability for the *in* condition and the *on* condition would not be the original effect we were looking for, it would suggest an imbalance, or difference, between how the different groups responded to the prepositions in question. The data show that the interpretability of the *in* variant sentences resulted in lower standard deviations (SD = .19) than the *on* variant sentences (SD = .28). When interpretability was averaged across sentence frame, Levene's test for homogeneity of variances revealed that the mean interpretability of the *in* condition was significantly less variable than the mean interpretability of the *on* condition (F (1, 78) = 11.79, p = .0010). This difference in interpretation variability, while subtle, is suggestive of a difference between the two prepositions.

To see whether any broader patterns were underlying this high variability, we next categorized each of the participants as either high-percentage interpreters or low-percentage interpreters. Since overall interpretations were provided for 54.06% of the sentences, participants who provided interpretations for ten or fewer of the twenty sentences were considered low-percentage interpreters and participants who provided interpretations for more than ten sentences were considered high-percentage interpreters. In the *in* condition, 29 participants were categorized as high interpreters and 10 as low interpreters; in the *on* condition, 20 participants were categorized as high interpreters and 21 as low interpreters. This difference between conditions was significant, χ^2 (1, N = 80) = 5.60, p = .0179. Taken together, these results hint at an effect of preposition on interpretability.

Discussion

Although the data hint at an imbalance between the potential interpretability of the prepositions *in* and *on*, we did not find the main effect of preposition that we had originally predicted. The lack of a result is particularly curious because, in a separate attempt to create novel non-spatial uses of *in* and *on* that would be considered by participants to be nonsensical, we had the subjective experience that nonsense *on* metaphors were easier to construct than nonsense *in* metaphors. While this phenomenological experience was reflected in the trends from Experiment 1, the lack of a significant main effect of preposition suggests one of two possibilities. One possibility is that our phenomenological experience may simply be different in kind from the phenomenological experience of our participants. In fact, Sandra and Rice (1995) warn researchers against relying exclusively on their own linguistic intuitions since these might differ dramatically from the intuitions of the general population.

Another possibility is that our subjective experience was driven by the task at hand. It may be that attempting to gauge the interpretability of both prepositions within the same semantic and syntactic contexts is what highlights their differences in interpretability. This difference – between the task leading to our subjective experience and the experimental task performed by our participants – is not unlike the difference between a within-subjects experimental design and a between-subjects experimental design. Birnbaum (1999) argues that participants are exposed to different contexts depending on whether they are taking part in a within-subjects experiment or a between-subjects experiment, and it is this difference in context that could result in widely divergent results from the two kinds of experiments. For example, in the between-subjects design of Experiment 1, the context for each sentence was a set of sentences involving novel prepositional phrases built upon a single preposition. In contrast, the context of our subjective experience was the creation of novel prepositional phrases built upon both *in* and *on*, facilitating a comparison between them. This comparison is more like the everyday experience of using language, in which novel sentences are encountered in the context of similar structures built around a variety of lexical items. Similarly, a within-subjects design in which participants would be exposed to both *in* sentences and *on* sentences would allow for an implicit comparison of the two prepositions, akin to the range of contexts which speakers are exposed to in everyday language use. As a result of these differences in context, the lack of a between-subjects effect for preposition might reflect more about variation in the interpretability of the novel sentence frames than about similarity in the interpretability of novel *in* and *on* prepositional phrases.

Thus, Experiment 2 was designed to test whether the lack of a strong result in Experiment 1 was due to differences between the linguist and the language user or to differences between a task involving consideration of multiple prepositions and one involving consideration of a single preposition.

Experiment 2

Using a completely within-subjects design, Experiment 2 tested whether novel non-spatial uses of the preposition *in* would show higher interpretability than matched non-spatial

uses of the preposition *on*. If so, participants should make more attempts to interpret sentences containing *in* than sentences containing *on*.

Method

Participants A total of 20 University of Louisiana at Lafayette students participated in this experiment in exchange for course credit. Two students were removed from further analysis because they identified themselves as native speakers of Igbo and Arabic, respectively. The 18 remaining participants were native speakers of English.

Materials The materials were the same as those used in the first experiment.

Procedure The procedure was the same as in the first experiment, except that participants saw all 40 of the stimulus sentences.

Design We used a 2 (Preposition: *in* or *on*) x 20 (Sentence Frame) design. Both were treated as within-subjects factors.

Analysis and Results

Interpretability Unlike in the between-subjects design of Experiment 1, a two-way repeated measures ANOVA on the results of Experiment 2 revealed a main effect of preposition (F (1, 17) = 11.87, p = .0031), whereby participants were significantly more likely to attempt interpretations of *in* sentences (M = .64, SD = .48) than interpretations of *on* sentences (M = .54, SD = .50), as predicted.

In addition, as in Experiment 1, we observed a significant main effect of sentence frame (F (19, 323) = 3.45, p < .0001), confirming that the sentence frames were not all equally interpretable. Furthermore, as in Experiment 1, we observed a significant preposition by sentence frame interaction, F (19, 323) = 1.64, p = .0453 (see Figure 1). In support of our prediction, post-hoc contrasts revealed that for six sentence frames the *in* variant sentence was more interpretable than the *on* variant sentence, while for no sentence frame did participants find the *on* variant sentence to be more interpretable than the *in* variant sentence.

Discussion

In contrast to our own subjective experiences considering the interpretability of novel prepositional phrases headed by *in* and *on*, in Experiment 1 we failed to find a significant difference in the interpretability of sentences utilizing the preposition *in* and sentences utilizing *on*. The question we wanted to address in Experiment 2 was whether the difference between our experiences and the results of Experiment 1 were due to differences between the analyst and the language user (cf., Sandra & Rice, 1995) or due to differences between considering the interpretability of two prepositions and considering the interpretability of just one. In contrast to Experiment 1, in Experiment 2 we observed a difference in interpretability between *in* sentences and *on* sentences when participants were asked to interpret both kinds of sentence, suggesting that it was the task itself that masked the differences in interpretability in Experiment 1.

In line with the observed distributional differences from the corpus-based work (see Introduction), Experiment 2 revealed that *in* can more easily appear in novel combinations with other lexical items than *on* can. This difference in interpretability between *in* and *on* suggests that the two prepositions may be operating at slightly different semantic levels.

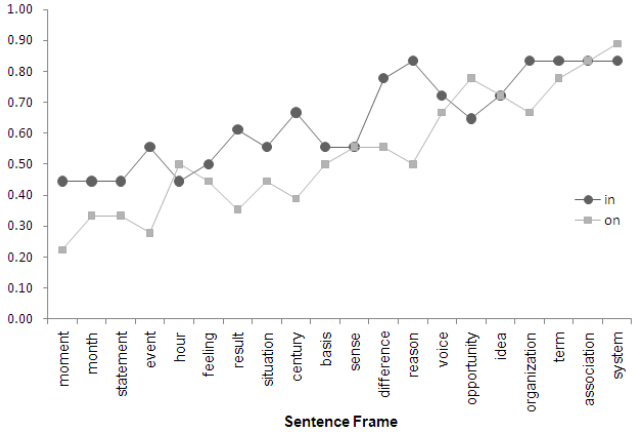


Figure 1: Proportion of participants providing interpretations for each sentence frame paired with each preposition. Each sentence frame is represented in the graph by its prepositional object.

General Discussion

Corpus-based studies of *in* and *on* have yielded observations of differences in morphosyntactic distribution (Navarro, 1998), overall frequency, and the range of non-spatial uses of the prepositions, calling into question the validity of the type-level similarity suggested by traditional treatments of prepositions in linguistics. In this study, we asked whether these differences correspond to differences in interpretability between the two prepositions, suggesting that the noted imbalance is in fact real and supporting the interpretation that the observed differences are due to a difference in meaning potential between *in* and *on*.

Across two studies, we found that *in* and *on* did evidence semantic differences in their combinatorial potentials. When participants were asked to interpret both novel *in* prepositional phrases and novel *on* prepositional phrases, we found that they were more likely to reject as uninterpretable sentences involving *on* phrases than sentences involving *in* ones, echoing the trend in interpretability found when participants were asked to interpret just one kind of sentence. In addition, we found that all sentences for which there was a significant difference in interpretability were more often interpreted in the *in* variant than in the *on* variant. In no case did we find the *on* variant to be more interpretable than the *in* variant in our novel contexts.

Taken together, this pattern of results suggests that, in addition to having different meanings, the prepositions *in* and *on* have different semantic combinatorial possibilities.

While this result is suggestive, further investigation is necessary to understand the strength and scope of the differences between *in* and *on*. For example, in balancing the frequency of co-occurrence of the abstract nouns and the two prepositions, we considered only the frequency of the collocations between the prepositions and the abstract nouns with no intervening lexical items, leaving aside co-occurrences at greater distances (e.g., *in a sense*, which is very high in frequency). However, our participants could potentially have used these phrases, if familiar, to interpret the novel sentences (e.g., *These roads are in a sense*.). Alternatively, participants may simply have been more likely to attempt an interpretation because of the high frequency of co-occurrence between the preposition and the noun at two-step (e.g., *in a sense*) and three-step positions (e.g., *in the traditional sense*). In order to gain a clearer understanding of the differences between *in* and *on*, we are planning a follow-up experiment in which these frequencies will also be balanced.

Conclusions

Taken together, the differences in distribution, frequency, and semantic combinatorial possibility argue against the assumption of a type-level similarity between *in* and *on*. In addition, the fact that all three types of data point toward *in* having a wider range of applicability than *on* suggests that these three phenomena may be linked.

Our results suggest that the overall combinatorial possibilities for *in* may be higher than those for *on*. In particular, this might result in a wider range of metaphorical extensions for *in* than for *on*. As a result, investigations into the semantics of non-spatial uses of these prepositions would benefit from taking into account the differences in meaning potential between these prepositions and the possibility that the structure of the extensions and their relations to spatial uses may similarly differ.

Acknowledgements

We would like to thank Marlene Burke for help with data collection, Carmen Comeaux for help recruiting participants, and the Language and Cognition Lab at UL Lafayette for comments and helpful discussion.

References

Birnbaum, M. (1999). How to show that 9 > 221: Collect judgments in a between-subjects design. *Psychological Methods*, *4*(3), 243-249.

Carroll, D.W. (2004). *Psychology of language* (4th ed.). United States: Thomson Wadsworth.

Cook, V., & Newson, M. (2007). *Chomsky's Universal Grammar: An introduction* (3rd ed.). United States: Blackwell Publishers.

Coventry, K.R. & Garrod, S.C. (2004). *Saying, seeing, and acting: The psychological semantics of spatial prepositions.* New York: Psychology Press.

Feist, M.I. (2000). On *in* and *on*: An investigation into the linguistic encoding of spatial scenes. (Doctoral dissertation, Northwestern University).

Feist, M.I. (2008). Space between languages. *Cognitive Science, 32* (7), 1177-1199.

Feist, M.I. (in press). Inside *in* and *on*: Typological and psycholinguistic perspectives. In V. Evans & P. Chilton (Eds.), *Language, cognition, and space*. London: Equinox.

Feist, M.I. & Cifuentes Férez, P. (2007). The object-relation continuum in language. *Proceedings of the Twenty-Ninth Annual Meeting of the Cognitive Science Society.*

Feist, M.I., & Gentner, D. (2003). Factors involved in the use of *in* and *on*. *Proceedings of the Twenty-Fifith Annual Meeting of the Cognitive Science Society.*

Francis, W.N., & Kučera, H. (1982). *Frequency analysis of English usage: Lexicon and grammar.* Boston: Houghton Mifflin.

Gentner, D., & Asmuth, J.A. (2008). Can relationality be distinguished from abstractness in noun mutability? *Proceedings of the Thirtieth Annual Meeting of the Cognitive Science Society.*

Gentner, D., & France, I. M. (1988). The verb mutability effect: Studies of the combinatorial semantics of nouns and verbs. In S. L. Small, G. W. Cottrell, & M.K. Tanenhaus (Eds.), *Lexical ambiguity resolution: Perspectives from psycholinguistics, neuropsychology and artificial intelligence* (pp. 343-382). San Mateo, CA: Kaufmann.

Navarro i Ferrando, I. (1998). A Cognitive Semantics analsyiss of the lexical units AT, ON, and IN in English. (Doctoral dissertation, University Jaume I, 1998) . Retrieved from http://www.tesisenxarxa.net/TESIS_UJI/ AVAILABLE/TDX-0804103-133233//navarro.pdf

O'Grady, W., Archibald, J., Aronoff, M., & Rees-Miller, J. (2005). *Contemporary linguistics: An introduction* (5th ed.). New York: Bedford/St. Martin's.

Sandra, D. & Rice, S. (1995). Network analysis of prepositional meaning: Mirroring whose mind—The linguist's or the language user's? *Cognitive Linguistics*, *6*, 89-130.

Tyler, A. & Evans, V. (2003). *The semantics of English prepositions: Spatial sciences, embodied meaning, and cognition.* New York: Cambridge University Press.

Vandeloise, C. (in press). Genesis of spatial terms. In V. Evans & P. Chilton (Eds.), *Language, cognition, and space*. London: Equinox.

Wilson, M.D. (1988). The MRC Psycholinguistic Database: Machine Readable Dictionary, Version 2. *Behavioural Research Methods, Instruments and Computers*, *20*(1), 6-11.

A computational model of cognitive interference without neural inhibitory mechanisms

Serge Thill (serge.thill@his.se)
Informatics Research Centre, University of Skövde
PO BOX 401, Skövde, Sweden

Robert Lowe (robert.lowe@his.se)
Informatics Research Centre, University of Skövde
PO BOX 401, Skövde, Sweden

Abstract

Interference between one cognitive behavior or sensory stimulus and subsequent behaviors is a commonly observed effect in the study of human cognition and Psychology. Traditional connectionist approaches explain this phenomenon by mutually inhibiting neural populations underlying those behaviors. Here, we present an alternative model, relying on a more detailed use of synaptic dynamics, in which populations of purely excitatory neurons can nonetheless interfere with each other, causing inhibition of activation for a varying amount of time. The fundamental, biologically motivated, mechanism in the model relies on current "spilling over" from an active neural population into another one, thereby depleting the latter population's synaptic resources. The principles underlying the model may find applications even in the design of problem-solving artificial neural networks.

Keywords: Neural modeling; Synaptic dynamics; Cognitive Interference.

Introduction

The effects on cognitive performance of *interference* in the process of associating temporally contiguous behaviors or events is a well studied phenomenon in the research disciplines of psychology and animal learning. Simply, it consists of the effects on working memory or memory recall of the presence of stimuli (or motor activations) that are non-critical to the learning of particular response/event associations. In the case of animal learning, it is best understood as entailing distractorstimuli introduced prior to (proactive) or after (retroactive) a task stimulus designed to be reliably predictive of another (e.g. rewarding) stimulus. In human learning, interference can manifest in learning deficits subsequent to pairing either context relevant (Oliveri et al., 2004) or incongruent (Buccino et al., 2005) motor actions and verbal descriptions. In every day human activities, the interference effect has implications for recall of important events, e.g. eye witness testimony (see Bouton, 2007).

Laboratory controlled studies of interference often utilize the delayed matching-to-sample (DMTS) paradigm whereby the subject is required to produce the desired behavioral response over a pre-determined delay period (or inter-stimulus-interval). In such cases, interference is a function of the strength of a 'distractor' stimulus and may induce forgetting (*cf.* Roberts & Grant, 1978), impaired learning (Revusky, 1971) or memory retrieval deficits (Gordon et al., 1981).

Some forms of associative learning may be more or less prone to the interference effect. Recent neuro-scientific evidence has uncovered that areas of motor and premotor cortex that become active during physical movement overlap with areas activated during the reading of the specific affected movement, e.g. hand, foot (Hauk et al., 2004). Buccino et al. (2005) for instance found an interference effect when human subjects, required to produce hand or foot responses to particular verb forms, produced physical movements apt to the action described in the particular sentence. Latency of response increasedin this case as compared to when a movement was required that was inapt to the particular action described (see Chersi et al., 2010, for a more detailed discussion).

Models exist that attempt to capture empirically demonstrated features of the interference phenomenon specified at the level of both connectionist and more neurobiologically motivated levels of abstraction. A seminal model of McGeoch (1932) proffered a connectionist account of interference whereby responses learned during a given time window would compete for retrieval by way of mutual inhibition. Essentially, this offered a classical account of 'distractor' stimuli inhibiting the influence of task-specific stimuli. The learned associative strengths of the responses determined the 'winner' which was, however, premised on the biological implausability of there being independence, as opposed to overlap, between the available responses.

Mensink & Raaijmakers (1988) provided a stochastic search model of retrieval that was able to describe behavioral data accounting for many of the effects of interference, e.g. proactive inhibition, retroactive inhibition, spontaneous recovery - where previously learned associations become behaviorally extinguished but, presumably still reside in memory.

More recently, neural models have been put forward to account for the ability of organisms to retain spatial information about stimuli over delay periods in the face of distracting (interfering) stimuli. Spencer et al. (2009) have described how the tuning of parameters of an interaction kernel on a dynamic neural field representing spatial working memory permits the development of activation peaks. These peaks are sustained through the use of tuned local excitation and global inhibition parameters on the kernel that afford more or less robustness to noise and distractor stimuli presented to the spatial field. Self-sustained activity can be achieved through bistable unit dynamics (*cf.* Amari, 1977) such that input or noise in-

duced supra-threshold individual unit activity may be maintained even following the withdrawal of the input. Neural field and bistable dynamics through the effective coupling of spatially mapped locally excited activation peaks in different fields provide mechanisms for coping with interference effects over delays between events of motor sequences to be associated.

The assumption in the above-mentioned models and theory is that interference (or distracting stimuli) induce inhibitory effects on the activity of applicable functional circuits or psycho-behavioral states whereas chaining of activations within populations of units entails excitatory activity. In dynamic field theory, for example, distracting stimuli induce elevated levels of global inhibitory activity serving to suppress existing continuous attractor states (i.e. activation peaks) potentially below threshold levels thus serving as a medium for forgetting.

Connectionist and population coding models seeking to enhance comprehension of the interference effect typically do not concern themselves with the biophysical details of the neuron units implied in the modeling approach, relying simply on 'point-to-point' synaptic transmission. However, considering that associations of activation may be somatotopically realized in the brain, i.e. via neighboring or overlapping populations of neurons (*e.g.* Chersi et al., 2010), and that current in a given population typically overlaps with or may otherwise 'spill over' into another population, it may be instructive to produce more detailed neural models taking into account these effects in order to better understand neural substrates of behavior.

A precedent for modelling the effects of a non-synaptic neuromodulatory process only recently thought to play a significant cognitive role exists. Nitric oxide (*NO*) gas is an inter-cellular signalling mechanism found in various structures of the brain. *NO* emissions affect neighbouring cells according to a slow diffusive dynamic different to standard point-to-point synaptic transmission. *NO* diffusion has been modelled (Philippides et al., 1998) and an analogue has been applied in the domain of cognitive robotics (Husbands et al., 1998). Recent evidence also suggests a functional role in homeostatic regulation of essential metabolic variables (e.g. Canabal et al. 2007).

The particular inter-cellular signalling mechanism we are concerned with here involves current that affects neighbouring regions of cells through non-standard synaptic transmission. A complete discussion of the different mechanisms that can cause current from one neural population to leak, or "spill over" into another population is beyond the scope of this paper. However, an interesting example of such a current spillover can for instance be observed when ionic neurotransmission at the synaptic cleft is not fully absorbed by the post-synaptic receptors of the receiving cell. Ions spill over the synaptic cleft and can thereby affect neighboring neurons, possibly of other populations leading to slow-rising increases in excitatory post-synaptic currents in the affected cells. Spillover has recently been recognized as a modulatory effect that may play a significant role in brain functioning, e.g. in the communication between the brain stem and cerebellum (Nishiyama & Linden, 2007), illustrating that neural communications do not necessarily rely solely on canonical synaptic transmission.

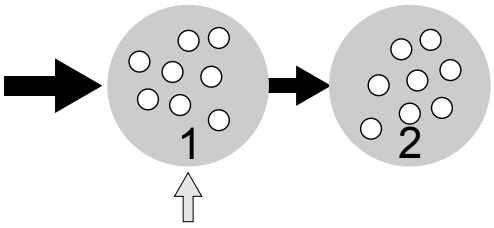


Figure 1: **Schematic of the neural model**. Two connected populations represent the neural substrate of a behavior. The behavior is triggered if the second population fires after triggering current arrives at the first one (large arrow). Weak spillover current, by itself insufficient to trigger the behavior, can also arrive at the first population (small arrow)

Here, we propose a neural model of the interference effect based primarily on synaptic dynamics. We model a sequence of two neural connected populations and show that, if spillover current from neural circuits external to the model reach the first population, activation of the second population may be prevented. Since we are mainly interested in the possible effects of the spillover current, we do not model or make assumptions on the precise underlying mechanisms. Nonetheless, we show that interference effects can be observed even though all currents are excitatory. Our model thus departs from the classically conceived models focusing on inhibitory inter-population inducement of interference. Our aim is to demonstrate that neural or neural network models of interference may be insufficient when focused solely on inter-population 'point-to-point' synaptic transmission effects. Accounting for biophysical dynamics when designing computational models or artificial neural networks may provide valuable insights to the fields of animal learning and psychology.

Methods

Neural and synaptic dynamics

We model the neural and synaptic dynamics following a standard model. The synaptic dynamics in particular take into account the fact that synaptic transmitters (or simply resources) are finite and both short term facilitation and depression can result from their dynamics (See Tsodyks et al., 1998, for a detailed discussion). Briefly, depression is caused by recognizing that synaptic resources may be "active" (in the synaptic cleft or at the post-synaptic receptors), "inactive" (returning to the pre-synaptic terminals and thus unavailable) or "recovered" (at the pre-synaptic terminals and available for release into the synaptic cleft on arrival of pre-synaptic current) and making the post-synaptic current dependent on the proportion

of active resources. The corresponding mean field equations are adapted from Tsodyks et al. (1998) with minor modifications to make the bounded nature of the resources explicit:

$$\frac{d\langle\rho\rangle}{dt} = \frac{1-\langle\rho\rangle}{\tau_{rec}} - \min\left(\langle\rho\rangle, \left\langle U_{SE}^{1}\right\rangle \langle\rho\rangle E(t)\right) \quad (1)$$

$$\frac{d\langle\alpha\rangle}{dt} = -\frac{\langle\alpha\rangle}{\tau_{in}} + \min\left(\langle\rho\rangle, \left\langle U_{SE}^{1}\right\rangle \langle\rho\rangle E(t)\right) \quad (2)$$

where ρ and α denote recovered and active resources respectively. Only recovered resources can generate post-synaptic current (by becoming active) and active resources affect the amplitude of post-synaptic current (Eqn. 5).The firing rate $E(t)$ is discussed further below. U_{SE}^{1} is a time-varying and firing-rate dependent parameter which models short term synaptic facilitation believed to be caused by residual calcium in the synaptic cleft. It is governed by the following equations:

$$\frac{d\left\langle U_{SE}^{-}\right\rangle}{dt} = \frac{\left\langle U_{SE}^{-}\right\rangle}{\tau_{facil}} + \min\left(1-\left\langle U_{SE}^{-}\right\rangle, U_{SE}\left(1-\left\langle U_{SE}^{-}\right\rangle\right)E(t)\right) \quad (3)$$

$$\left\langle U_{SE}^{1}\right\rangle = \left\langle U_{SE}^{-}\right\rangle (1-U_{SE}) + U_{SE} \quad (4)$$

Population dynamics

To model the effect one population of excitatory neurons may have on another, we also follow the model by Tsodyks et al. (1998). The mean firing rate of a given population r is thus dependent on the incoming current from other populations r' and external current I_r arriving directly at population r:

$$\tau_e \frac{dE_r}{dt} = -E_r + g\left(\sum_{r'} J_{rr'}\alpha_{r'} + I_r\right) \quad (5)$$

where $J_{rr'}$ denotes the absolute strength of the connections from r' to r multiplied by the average number of such connections and $\alpha_{r'}$ is given by Eqn. 2. It can be noted here that the original model is more complex since it also caters for inhibitory populations, but those aspects are not relevant to the present work. g, finally, is a transfer function, for which we use a standard sigmoid with a threshold:

$$g(x) = \max\left(0, \frac{2}{1+e^{(4-x)/3}} - 1\right) \quad (6)$$

Two or more populations governed by the above dynamics can then be seen to form the neural substrate of an observable behavior. In our model, the parameter choices are: $\tau_{rec} = 1000$ms, $\tau_{in} = 100$ms, $\tau_{facil} = 530$ms, $U_{SE} = 10^{-6}$ and $J = 4$. These parameters have been chosen to produce bell-shaped activation curves in the neural populations (rather than undesired firing patterns). They mostly (except where discussed below) affect the firing rates of the neural populations but the precise choices are not critical for illustrating the effect described in the present work.

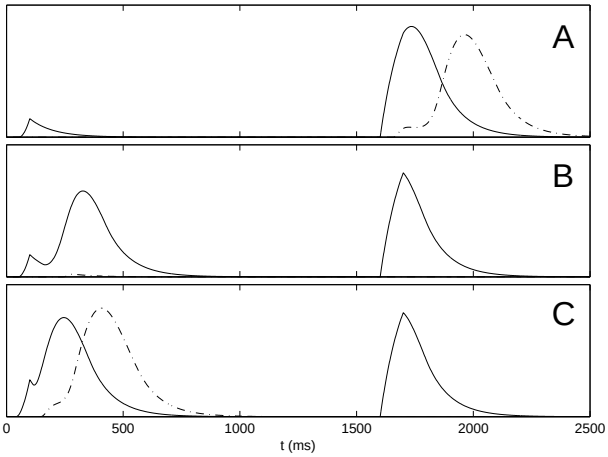


Figure 2: **The effects of weak, moderate and strong spillover current**. Solid (broken) line represents firing rate in first (second) population. Spillover current begins at t=0, behavior triggering at t=1600ms. (**A**) Spill-over current is insufficient to prevent activation of second population during behavior triggering. (**B**) Spill-over current causes significant but sub-threshold activation in the first population and prevents triggering of the second population later on. (**C**) Spill-over current is sufficient to prematurely trigger the behavior.

Results

We model two connected populations of neurons (Fig. 1) which are meant to represent the neural substrate (or part thereof) of an observable cognitive behavior. Such an arrangement is for instance thought to underlie action execution in the motor cortex (Chersi et al., 2006). The behavior is "triggered" if external current arriving at the first population is of sufficient amplitude to cause activation in the second population. In other words, a behavior is successfully triggered if the second population fires after the first one was stimulated (Fig. 2A, after 1500ms). We call *triggering current* any current that, in the absence of spillover current effects, is sufficient to trigger the behavior.

Conversely, we model spillover current as a type of external current arriving at the first population but of insufficient amplitude to cause the activation of the second population (Figs. 2A and B, the first 1000ms). For the present illustrative purposes, the spillover current is modeled as lasting 100ms and increasing linearly by a small amount I_{spill} during that time. After 100ms, the current dies away instantaneously. I_{spill} has a range of possible values, with the exact choice affecting overall behavior, which is explored below. It should be noted that the observation of the reported interference effect does not critically depend on this particular choice for modeling the spillover current. Of importance is merely the fact that supra-threshold activation is generated in the first population in some way.

To illustrate the effect spillover current can have (Fig. 2),

we first determine a sufficient triggering current for the behavior in a control case with no spillover current. We then measure the post-triggering firing rate of the second population in situations where the triggering current was preceded by a spillover current δt ms earlier. Any change in firing rate compared to the control case is of interest.

Interference without inhibition

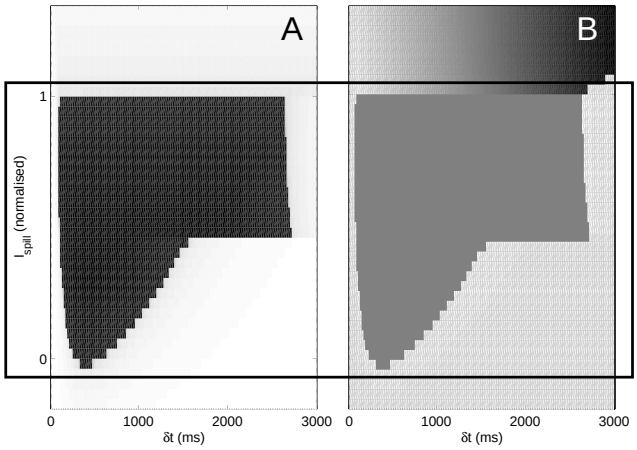


Figure 3: **Interference effect**. Y-axes indicate spill-over current strength I_{spill}, normalised so that values of interest fall between 0 and 1. Rectangle indicates this region of interest (bounds of I_{spill}). X-axes indicate values for δt. the waiting time between end of spillover and start of behavior-triggering current. Figures are grayscale ranging from black (0) to white (maximal values of the plotted parameters). (**A**) Firing rate of the second population determined by chosen values of I_{spill} and δt. Black region indicates no firing (and therefore interference). Other regions show firing rates all at similar, close to maximal levels. The interference effect thus either causes a strong suppression of firing rate or no significant effect at all. Further, the δt values for which the interference effect is observed depend on the value of I_{spill} (see text). (**B**) Time delay between peak of activation in first population and corresponding peak in second population. If the second peak was inhibited, this information does not exist (solid gray area). Region with $I_{spill} > 1$ shows premature activation (little to no time delay, dark colors) of second population due to excessively high values of I_{spill} (see Fig. 2C). Region with $I_{spill} < 0$ shows normal separation between peaks (see Fig. 2A). Region within rectangle ($0 \leq I_{spill} \leq 1$) shows separation similar to the normal case with $I_{spill} < 0$ but not to the premature activations observed when $I_{spill} > 1$. Thus, if both populations fire, I_{spill} does not significantly affect the timing between peaks in the region of interest (rectangle).

Since spillover current that is too low (Fig. 2A) or too high Fig. (2C) is not going to cause any interesting effects, we define lower and upper bounds of I_{spill} as follows: the spillover current should be strong enough to cause some measurable effect during an attempt at triggering the behavior but weak enough not to cause this triggering by itself (*e.g.* Fig. 2B). We define "measurable effect" simply as a difference in time-course and/or peak values in the firing rate of the second population, thus not excluding the possibility of a facilitation effect.

We find, however, that any spillover current sufficient to cause a measurable effect prevents activation of the second population (Fig. 3). The duration of this interference can vary and depends on the strength of the spillover current (Fig. 3A). For values near the lower boundary, the effect disappears if the behavior is triggered around 460ms or later after termination of the spillover current. Near the upper boundary, the interference window can last up to about 2800ms. For very small values of the spillover current, it is possible to avoid the interference effect if the behavior is triggered very shortly after the end of the spillover current (up to 340ms in the best case), since synaptic resources are depleting more slowly.

The maximal duration of the interference window is mostly affected by the choice of τ_{rec}. Interestingly, however, it is not reached monotonically. Rather, as can be seen in Fig. 3A, a threshold value for spillover current exists below which the interference effect disappears after a fraction of its maximal effect. Above the threshold, the interference effect lasts for its entire possible duration.

It would theoretically be possible for the spillover current to cause a delayed activation in the second population, rather than complete inhibition. This would be apparent if the time between the peak activation of both populations was a function of the strength of the spillover current. However, at least within the context of the work presented here, no such effect was found. Fig. 3B shows that, if the spillover current is within its bounds, it will either cause complete interference or, with a sufficient waiting period between spillover and behavior-triggering current, no effect at all. It should be noted however, that on a behavioral level, delays can still be observed. This would correspond to a control mechanism which re-triggers the behavior after noticing that the initial attempt was not succesful. Modeling these control mechanisms in detail is, however, beyond the scope of this work.

Fundamental cause

Since the behavior of the system described here is modulated only by synaptic dynamics, the cause for the observed interference effect is also found therein and illustrated in Fig. 4. Any activity within the first population will cause a reduction of recovered synaptic resources (as they become active). Since the amount of synaptic resources activated by incoming current is proportional to the recovered resources, fewer recovered resources mean smaller increase in current. If I_{spill} is very small, recovered resources do not deplete drastically during spillover current (Fig. 4A) and a following triggering current can have normal effects. If I_{spill} is larger, the recovered resources do deplete drastically but over a relatively long time-course (Fig. 4B). This slow depletion allows active resources to inactivate quickly enough to keep the proportion

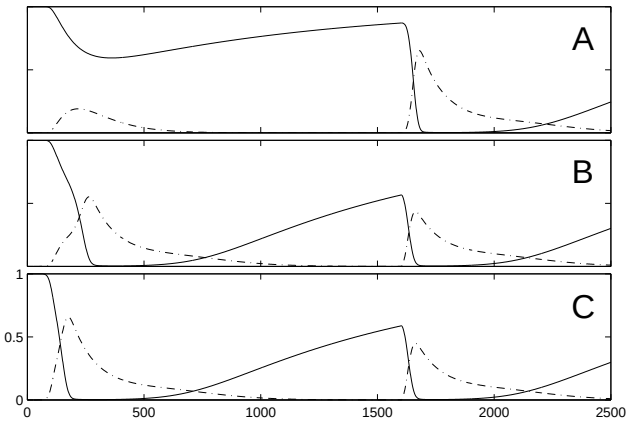


Figure 4: **Synaptic resources**. Solid (broken) line represents recovered (active) synaptic resources from the first population corresponding to the firing rates seen in Fig. 2. Spillover current begins at t=0, behavior triggering at t=1600ms. (**A**) Spill-over current causes a small decrease in recovered synaptic resources but the triggering current can activate sufficient amounts to cause firing in the second population. (**B**) Spill-over current causes complete but slow depletion of recovered resources. Not enough resources can recover and the fraction activated by the triggering current is insufficient to cause firing in the second population. (**C**) Spill-over current causes complete and fast depletion of recovered resources. Consequently, the proportion of active resources becomes sufficiently high to trigger the behavior prematurely.

of active resources below the necessary threshold for triggering the second population. At the same time, the depletion is significant enough that a later triggering current cannot activate a sufficient proportion of resources either - we observe interference. Finally, a very large value of I_{spill} works just like a triggering current: recovered resources activate quickly enough to push the proportion of active resources over the triggering threshold before it can decrease again due to inactivation.

Thus, the interference effect described here relies on a slow but significant depletion of synaptic resources. In theory, the effect of reduced available resources could be offset by the synaptic facilitation mechanism implemented here. However, since τ_{facil} is usually shorter than τ_{rec}, this is not observed in the present model.

Effects of parameter choices

Naturally, the exact values, most notably for the lower and upper boundaries of the spillover current, depend on the values chosen for the synaptic parameters in the model. The most important ones are the synaptic strength and the proportion of synaptic resources liberated. We do not address these effects in detail here but did find in a brief exploration that, as long as parameters are kept within ranges that allow a bell-shaped activation of both populations as seen in Fig. 2A after the 1500ms mark (as opposed to, e.g. self-sustaining, chaotic or oscillatory behavior), spillover current always appears to cause interference effect.

Discussion

The model presented in this paper departs from the more classical artificial neural network models in its use of more detailed biophysical dynamics. By taking into account the fact that synaptic resources are finite, we have been able to inhibit the execution of a behavior even though all currents within the model are excitatory. While our model merely provides an alternative account compared to those relying on inhibitory dynamics, it does not necessarily replace them. However, it does illustrate the power of more detailed biophysical dynamics in a model. There is therefore a necessity to move beyond simple point-to-point artificial neural networks if the purpose of such a network is to explain cognitive phenomena.

Although we do not provide an extensive parameter exploration here, the findings are rather robust. The parameters of the synaptic model affect the firing behavior of the populations more than the effect of the spillover current (the main exceptions to this are of course τ_{rec} and τ_{facil}). Likewise, we do not need to formulate any strong assumptions on the precise nature of the spillover current because the critical aspect is merely the activation generated within the first population. The effect is thus general but further work would be needed to explore the effects of different values for τ_{rec} and τ_{facil} respectively. For instance, one could discover values for which the spillover current causes both facilitation and interference (or only facilitation). However, it should be noted that this would mainly be interesting from a theoretical perspective, since typical short term facilitation time-courses tend to be faster than depletion ones (Tsodyks et al., 1998). In fact, related work (Chersi et al., 2010) which is concerned with modeling both interference and facilitation effects simultaneously has found that in such cases, neural dynamics including inhibitory currents may provide a better explanation.

Besides their role as explanatory tools for cognitive phenomena, neural networks also find applications as computational problem-solving tools. By illustrating the effects synaptic dynamics can have on the overall output of our model, we show that moving beyond the traditional connectionist models of nodes simply connected by a signed weight can be worth considering. While this will not extend the set of computations that a neural network can perform, it may simplify the topology or facilitate training. Such benefits have for instance been previously found in GasNets (Husbands et al., 1998). These networks have proven particularly amenable to efficient search of task solution space as cognitive robotics controllers situated according to spatial and temporal environmental constraints. This adaptive potential is tapped using a diffusive, non-purely point-to-point synaptic modulatory network. Exploration of the interaction of classically conceived synaptic transmission and less orthodox means of inter-cellular communication may provide scope to investi-

gate spatial and temporal interactions relevant to the study of cognitive phenomena particularly in an embodied context (*cf* Parisi, 2004). Again, these are possibilities that need to be explored further in future work.

Conclusions

We have presented a model that can explain temporal interference effects without relying on inhibitory dynamics in the underlying neural circuitry. Rather, the behavior is explained solely by synaptic dynamics which are modeled in a simple yet biologically plausible way. The contributions of this work are twofold: (1) We provide an alternative explanation for a range of interference effects which does not rely on explicit inhibitory dynamics. (2) We highlight the benefits of modeling synaptic and biophysical dynamics in more detail, both as a computational tool which may find applications even in artificial neural networks and as an explanatory mechanism as illustrated in the present paper.

Acknowledgments

This work was supported by the European Commission FP7 project ROSSI (www.rossiproject.eu), Grant agreement no. 216125,

References

Amari, S. (1977). Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological Cybernetics*, *27*, 77-87.

Bouton, M. E. (2007). *Learning and behavior: A contemporary synthesis*. Sinauer Associates, Inc. Publishers.

Buccino, G., Riggio, L., Melli, G., Binkofski, F., Gallese, V., & Rizzolatti, G. (2005). Listening to action related sentences modulates the activity of the motor system: A combined tms and behavioral study. *Cognitive Brain Research*, *24*, 355-63.

Chersi, F., Mukovskiy, A., Fogassi, L., Ferrari, P. F., & Erlhagen, W. (2006). A model of intention understanding based on learned chains of motor acts in the parietal lobe. In *Proceedings of the 15th annual computational neuroscience meeting.* Edinburgh, UK.

Chersi, F., Thill, S., Ziemke, T., & Borghi, A. M. (2010). Sentence processing: linking language to motor chains. *Frontiers in Neurorobotics*, doi:10.3389/fnbot.2010.00004.

Gordon, W. C., Taylor, J. R., & Mowrer, R. R. (1981). Enhancement of short-term retention in rats with pretest cues: Effects of the training-cueing interval and the specific cueing treatment. *American Journal of Psychology*, *94*, 309-322.

Hauk, O., Johnsrude, I., & Pulvermüller, F. (2004). Somatotopic representation of action words in human motor and premotor cortex. *Neuron*, *41*, 201-307.

Husbands, P., Smith, T., Jakobi, N., & O'Shea, M. (1998). Better living through chemistry: Evolving gasnets for robot control. *Connection Science*, *10*, 185-210.

McGeoch, J. A. (1932). Forgetting and the law of disuse. *Psychological Review*, *39*, 352-370.

Mensink, G. J.-M., & Raaijmakers, J. G. (1988). A model for interference and forgetting. *Psychological Review*, *95*, 434-455.

Nishiyama, H., & Linden, D. J. (2007). Pure spillover transmission between neurons. *Nature Neuroscience*, *10*, 675-677.

Oliveri, M., Finocchiaro, C., Shapiro, K., Gangitano, M., Caramazza, A., & Pascual-Leone, A. (2004). All talk and no action: A transcranial magnetic stimulation study of motor cortex activation during action word production. *Journal of Cognitive Neuroscience*, *16*, 374-381.

Parisi, D. (2004). Internal robotics. *Connection Science*, *16*, 325-338.

Philippides, A., Husbands, P., & O'Shea, M. (1998). Neural signaling - it's a gas! In L. N. M. Boden & T. Ziemke (Eds.), *Proceedings of the 8th international conference on artificial neural networks.* London: Springer-Verlag.

Revusky, S. H. (1971). Animal memory. In W. K. Honig & P. H. R. James (Eds.), (chap. The role of interference in association over a delay). New York: Academic Press.

Roberts, W. A., & Grant, D. S. (1978). An analysis of light-induced retroactive inhibition in pigeon short-term memory. *Journal of Experimental Psychology: Animal Behavior Processes*, *14*, 247-260.

Spencer, J. P., Perone, S., & Johnson, J. S. (2009). Toward a unified theory of development. In J. P. Spencer, M. S. C. Thomas, & J. L. McClelland (Eds.), (p. 86-118). Oxford.

Tsodyks, M., Pawelzik, K., & Markram, H. (1998). Neural networks with dynamic synapses. *Neural Computation*, *10*, 821-835.

Phonetic training makes word learning easier

Amy Perfors (amy.perfors@adelaide.edu.au)
Department of Psychology, University of Adelaide

David Dunbar (david.dunbar@student.adelaide.edu.au)
Department of Psychology, University of Adelaide

Abstract

Motivated by the idea that differences between adult and child language learners may stem in part from initially minor differences (such as in phonetic perception) that cascade throughout other aspects of language learning, we explored to what extent training adults on a novel phonetic contrast results in improved learning of words that incorporate that contrast. Results indicate that distributional training on a novel phonetic contrast improves word learning as well as the ability to discriminate a related contrast. We discuss implications for how adults' phonological abilities in affect other aspects of language learning, and also for understanding the effectiveness of different phonetic training regimes.
Keywords: language acquisition; phonetic learning; second language learning

Introduction

Children and adults differ both qualitatively and quantitatively in their ability to acquire a new language. Adults have difficulty with many aspects of language acquisition, from phonetic perception (Werker & Tees, 1984; Werker & Lalonde, 1988; Kuhl, 2004) to language processing (Clahsen & Felser, 2006) to certain aspects of syntax (e.g., Johnson & Newport, 1989; Birdsong, 2006). Scientists have proposed many theories to account for the difference between children and adults; these theories differ in both the degree and type of contribution made by pre-existing language-specific biases. Although nearly everyone agrees that (due to the inherent logical problem of induction posed by language learning) some bias must be necessary to explain successful language acquisition, explanations about the nature of the bias – and the difference between children and adults – vary considerably.

Some argue that there is a fundamental difference between first and second language acquisition: that acquisition in children is guided by an innate Universal Grammar and language-specific acquisition procedures, but that adult acquisition is directed by more domain-general learning mechanisms (e.g., Bley-Vroman, 1990). There are many other possibilities, however, since children and adults differ profoundly in their cognitive capabilities and typical linguistic input. Children have significantly poorer cognitive skills, including memory and processing speed; perhaps these differences aid children to learn language by enabling them to isolate and analyze components of a linguistic stimulus (Newport, 1988) or to over-regularize inconsistent input (Hudson Kam & Newport, 2005; Singleton & Newport, 2004). Another possibility is that learning a second language is made more difficult due to interference from the first language; indeed, the evidence that experience with a first language influences acquisition of a second is extensive (e.g., Mayberry, 1993; Iverson et al., 2003; Tan, 2003; Weber & Cutler, 2003; Hernandez, Li, & MacWhinney, 2005). This explanation overlaps considerably with the related point that adult brains are in many ways less plastic, and therefore less malleable in response to novel input (Elman et al., 1996; MacWhinney, 2005). Other explanations suggest that adults and children differ in their style of learning (Ullman, 2004) as well as the nature of the social support (Snow, 1999) and linguistic input (Fernald & Simon, 1984) they receive. Of course, many of these possibilities may be true simultaneously.

This work investigates yet another possibility – that small differences in children's abilities along one dimension or aspect of language can have cascading effects, resulting in larger differences in other aspects of language. These initial minor differences might be due to language-specific skills that naturally decay over time, or could be due to domain-general changes in the underlying cognitive abilities that subserve them. Key to this idea is the notion that, because language is such an intertwined, multi-dependent system, small differences in one aspect of language can be steadily amplified when it comes to the acquisition of other aspects. This idea is similar to the neo-constructivist view of Karmiloff-Smith (1998): both suggest that differences in eventual linguistic performance may derive from cascading effects that result from variation in more basic skills. That view focuses on abnormal development in children, however. Our work is motivated by an extension of this viewpoint: the notion that some of the well-attested differences between child and adult learners may result from the more minor, lower-level differences between adults and children. To investigate this, we begin by identifying aspects of language acquisition where one might expect to see cascading effects, and investigate whether performance in one improves performance in the other.

What minor difference between adults and children might have significant cascading effects onto other aspects of language? One possibility derives from children's well-attested superior phonological processing and perception abilities. Young infants can distinguish between phonemes in all natural languages, but lose that ability by the age of 10-12 months if they have not received sufficient linguistic input for a language containing that phoneme (Eimas, Siqueland, Jusczyk, & Vigorito, 1971; Werker & Tees, 1984; Kuhl, 2004). Adults who begin acquisition of a language later in life, even after decades of experience using the language, show phonological deficits in perception, production, and processing (e.g., Flege, 1995; Pallier, Colomé, & Sebastián-Gallés, 2001; Sebastián-Gallés & Soto-Faraco, 1999).

Moreover, it is quite difficult to train adults to learn a phonetic contrast that does not exist in their native language. Various training regimes exist; some rely on implicit learning of the phonemic categories based on distributional information (Maye & Gerken, 2001, 2002; Shea & Curtin, 2005; Hayes-Harb, 2007), while in others explicit feedback is given (Jamieson & Morosan, 1989; Bradlow, Akahane-Yamada, Pisoni, & Tohkura, 1999; McCandliss, Fiez, Protopapas, Conway, & McClelland, 2002). Although it is possible to train adults to discern non-native phonetic contrasts, the resulting phonetic representations are often fragile. For instance, when trained through implicit distributional learning, adults show little ability to generalize their knowledge to other non-native contrasts that differ along an analogous phonetic feature (Maye & Gerken, 2001), even though infants are able to do so (Maye, Weiss, & Aslin, 2008).

Why might difficulties in phoneme perception be responsible for adults' relatively poor performance on other aspects of language? It is well-known that adults have difficulty rapidly processing fluent speech in their second language (e.g., Guillelmon & Grosjean, 2001; Clahsen & Felser, 2006), which may be in part due to difficulty in perceiving and representing the phonemes that make up that speech. Difficulties in rapid processing could lead to difficulties in segmenting words and mapping those words onto their correct referents; difficulties in identifying words – particularly function words, which are generally shorter and more phonologically impoverished than content words – might result in more difficulty identifying the appropriate parse for sentences and therefore the correct underlying grammatical structure. Consistent with this, phonological working memory is correlated with second language skills in adults (e.g., Perani, 2005), and speech processing efficiency is related to other aspects of linguistic competence in children (Tsao, Liu, & Kuhl, 2004; Fernald, Perfors, & Marchman, 2006). Empirical evidence reveals that knowledge of lower-level aspects of language (such as phonological perception or statistical segmentation) can help in the acquisition of more complex linguistic phenomena (Werker & Yeung, 2005; Mirman, Magnuson, Graf Estes, & Dixon, 2008). Recent computational work suggests that word learning and phonetic category learning are more effective when occurring simultaneously (Feldman & Griffiths, 2009), and that knowledge of phoneme distributions may aid in speech segmentation and identification of lexical categories (Christiansen, Onnis, & Hockema, 2009). However, there is no work we are aware of that explores whether the ability to recognize a phonetic contrast assists adults in other areas of language.

The work here addresses that issue. We train adult learners to perceive a non-native phonetic contrast and then evaluate how this affects their ability to learn novel words containing the phonetic contrast in question. Our results are relevant not only to the possibility that deficits in phonetic skills may have cascading effects through other aspects of language; they are also relevant to the question of how generalizable adult phonetic learning is. As mentioned previously, existing work suggests that although adults can be trained to distinguish novel contrasts, this ability is fragile, and they have difficulty generalizing that contrast to analogous contrasts (Maye & Gerken, 2001). However, this work used synthesized stimuli not found in any natural language, and training included many filler items, so that there was effectively less than five minutes of exposure to the phonemes of interest. Would adults be able to generalize with more exposure or on a more naturally-produced contrast? In other training regimes adults show robust differences in both perception and production of a novel contrast (Lively, Logan, & Pisoni, 1993; Bradlow et al., 1999; McCandliss et al., 2002), but these regimes differed in many ways from Maye and Gerken (2001): they were significantly longer, used more natural stimuli, and involved explicit training with feedback, among other differences. Most importantly, most of these studies did not evaluate generalization to a novel but similar phonetic contrast. Among those that did, generalization to the novel contrast was successful, but the training paradigms involved giving explicit feedback rather than distributional training (e.g., McClaskey, Pisoni, & Carrell, 1983; Wang, Spence, Jongman, & Sereno, 1999). It is therefore unclear whether the limited generalizability observed in Maye and Gerken (2001) is due some inherent inability to generalize based on distributional information, or is due to other details in the training regime. In this work, we incorporate an implicit distributional training regime similar to that of Maye and Gerken (2001), but one of longer duration and with more natural stimuli. Do these changes in training result in improved generalizability, both in terms of novel but similar phonetic contrasts, but also in terms of the ability to use the new phonetic categories when learning new words?

Method

We trained 61 participants recruited from the student population[1] at the University of Adelaide on two tasks: a phonetic training task and a word-learning task. Participants[2] were randomly assigned to either a CONTROL or a TRAINED condition, which differed in terms of the nature of the phonetic training given.

Task 1: Phonetic learning

Training. The first task consisted of phonetic training based on distributional learning, similar to the task in Maye and Gerken (2001). Subjects in the TRAINED condition were trained on the unaspirated velar plosive voiced/voiceless contrast (/g/-/k/), which occurs in languages such as Hindi but not in English (both phonemes sound like a "g" to an English

[1] No participants were native speakers of a language with the phonetic contrast we sought to train. 52 were native English speakers. To ensure that native language was not a factor, we performed all analyses on the full population as well as the English speakers only. Results were identical, so we report the full population results.

[2] Of the original 61 subjects, 9 were excluded from the final analysis (5 due to technical difficulties, 1 for failure to follow instructions and 3 who performed at chance levels on the control task, indicating inattention). This left 25 participants in the CONTROL condition and 27 in the TRAINED condition.

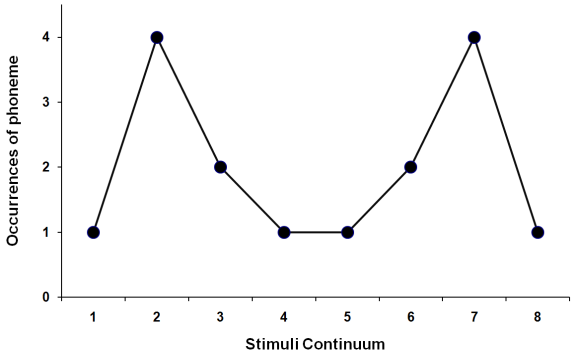


Figure 1: Distribution of stimuli used in phonetic training, defined along a continuum based on VOT. Tokens 2 and 7 occurred four times as often as tokens 1 and 8.

speaker). The /g/ and /k/ phonemes differ in terms of voice-onset time (VOT), such that /g/ contains a pre-voicing component while /k/ does not. It is therefore possible to gradually convert /g/ tokens into /k/ by successively removing parts of the pre-voicing component. Doing so yields a continuum of eight tokens from /g/ to /k/, separated by an average of 17ms in VOT from each other, and identical to each other except for the pre-voicing. As in Maye and Gerken (2001), we presented subjects with a bimodal distribution of these phonemes, as illustrated in Figure 1; thus, some tokens (e.g., 2 and 7) occurred four times as often as others (e.g., 1 and 8). Stimuli were recorded from a male native speaker of Hindi and edited using Praat phonetics software. Each of the phonemes occurred in one of three vowel contexts (/a/, /i/, and /u/).

In order to control for time spent listening to speech sounds across groups, subjects in the CONTROL condition also listened to a distribution of phonemes. However, they heard tokens from a phonemic contrast they could already recognize: the dental plosive aspirated/unaspirated voiced/voiceless contrast (/d/-/t^h/, which sound like "d" and "t" respectively to a native English speaker). As before, these phonemes were used to create a continuum of eight tokens extending from /d/ to /t^h/. Since these phonemes differ along aspiration as well as voicing, the tokens were created by removing voicing and then adding aspiration in continuous steps.

In both conditions, participants listened to a total of 912 tokens presented in random order and separated by 250 ms each, for a total of approximately 11 minutes of exposure to the sounds. During stimulus presentation the participants were told not to speak or read, but also that they need not consciously concentrate on the sounds. To alleviate boredom, they were allowed to doodle while listening.

Testing. Discrimination of the phonetic contrast was tested by presenting participants in both conditions with trials in which they heard three phonemes, two of which were identical. They were asked to press a button indicating whether the third phoneme they heard was the same as the first or the second (the distribution of correct answers was balanced across trials). There were three kinds of trials, defined by the nature of the phonemes tested. On *control* trials, the phonemes already existed in English (/d/ and /t^h/). On the *trained* trials, the phonemes were the ones that the TRAINED group had been trained on (/g/ and /k/). Finally, on the *untrained* trials, subjects were presented with a phonetic contrast that also does not exist in English and that is also defined by voice onset time, but differs in place of articulation – the unaspirated bilabial plosive voiced/voiceless contrast (/b/ and /p/, both of which sound like "b" to an English speaker).[3] The *untrained* trials enabled us to evaluate whether our subjects could generalize any learning to similar phonemes that differed on the same feature. There were 12 test trials for each contrast, totaling 36 testing trials in all; no feedback was given, and the order of all test trials was randomized.

Task 2: Word learning

Training. The word learning task was a standard task in which participants were presented with 12 different image types distributed over three stages of 36 trials each, making 108 trials in all. One each trial, an image was paired with a word, and the participants were instructed to try to learn the word-picture mapping. Words consisted of minimal pairs differing in initial position on each of the contrasts: *trained*: [g]ipur, [k]ipur, [g]anug, and [k]anug; *control*: [d]ipur, [t^h]ipur, [d]anug, and [t^h]anug; and *untrained*: [b]ipur, [p]ipur, [b]anug, and [p]anug. To ensure that the words differed only in the initial sound, words were created by splicing the same stem (-anug or -ipur) to the initial phonemes. The images corresponded to some of the earliest words spoken by children,[4] and were thus presumed to be highly familiar to all participants. The specific image-word pairing was randomized for each participant. The order of presentation of images was also random, with the constraint that each word-image pair was presented three times during each stage.

Testing. There were three testing sessions of 12 trials each, occurring after each stage. During each test trial, one of the 12 images was presented and participants heard two minimal pairs differing along the contrast in question (*trained*, *untrained*, or *control*). Thus, a participant might see a picture of a cat and hear [b]ipur followed by [p]ipur. Their task was to indicate whether the first or second word they heard was correct. No feedback was given.

Results

Task 1: Phonetic learning

Phonetic learning was evaluated by comparing performance on the phonetic test. As Figure 2 illustrates, participants in the TRAINED condition outperformed those in the CONTROL

[3]For *trained* and *control* trials, the exemplar tested on corresponded to token 1 and 8 from each continuum. Due to a coding error, the exemplar in the UNTRAINED trials corresponded to tokens 2 and 7 rather than 1 and 8. If anything, this is a more stringent test of generalization, but also means that it is more difficult to compare performance on the UNTRAINED trials to the other two. We discuss the implications of this in subsequent sections.

[4]They consisted of images of babies, balls, books, cats, chairs, birds, beds, cars, cookies, cups, dogs, and shoes.

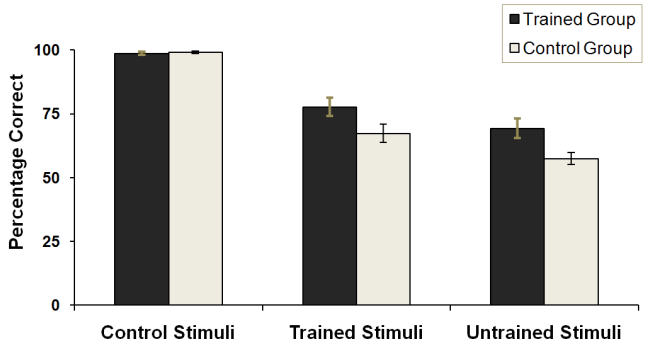


Figure 2: Phoneme discrimination test results. Participants who received distributional training outperformed participants in the CONTROL condition, but all participants performed above chance on all stimuli, suggesting that the test itself may have trained them. Error bars reflect standard error.

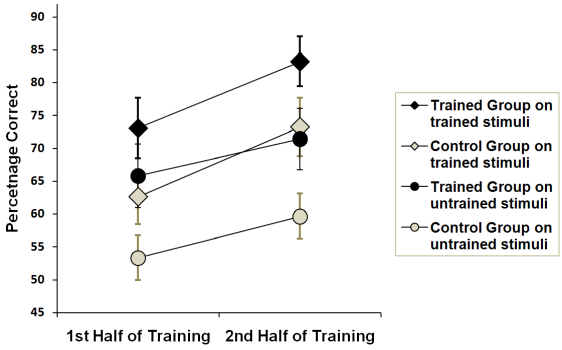


Figure 3: Were participants trained over the course of phonetic testing? Performance on the first half of testing is compared to performance on the second half. Both the TRAINED and CONTROL groups performed significantly better over the course of testing on the *trained* stimuli. While there was a positive trend, the difference in performance on the *untrained* stimuli for either group across the two halves of testing was not significant. The differential effects on *trained* and *untrained* stimuli is probably because the *trained* stimuli were easier to discriminate (tokens 1 and 8) than the *untrained* stimuli (tokens 2 and 7).

condition on both the *trained* and *untrained* stimuli.[5]

Interestingly, participants in both conditions performed above chance on the *trained* and *untrained* stimuli.[6] This suggests that the phonetic testing itself may have trained the participants in the CONTROL condition, which is not an unreasonable suggestion since it closely corresponds to the "prototype" training employed by Jamieson and Morosan (1989) or the "two-seven" condition of Hayes-Harb (2007). To evaluate to what extent such training occurred, we split scores on the phonetic test in half and compared performance on the first six test trials for each stimulus type with performance on the final six test trials of each. As Figure 3 indicates, both groups improved significantly over the course of the test.[7] There was no difference between the CONTROL group's performance in the final half of testing and the TRAINED group's performance in the first half: in other words, training during testing was so effective that it resulted in performance equivalent to having listened to distributional information for over 10 minutes.

It is also evident that performance on the *trained* stimuli was superior to performance on the *untrained* stimuli. This is true even for participants in the CONTROL condition, for whom there should have been no difference between the two types of stimuli (since they had heard neither before). This is probably an artifact of the coding error described earlier in which the *untrained* test stimuli consisted of tokens 2 and 7, rather than tokens 1 and 8 as for the *trained* stimuli. The *trained* stimuli were therefore probably both more effective at teaching participants the contrast, and also easier to differentiate (and hence get correct on the test). Consistent with the hypothesis that this was a training effect, analysis of the first trial of testing reveals that participants in the CONTROL condition performed equally, no better than chance, on both *trained* and *untrained* stimuli. In any case, the important finding – that subjects were able to generalize their phonetic learning to an untrained but related contrast – is unaffected by this detail.

Task 2: Word learning

Are participants able to generalize their phonetic discrimination abilities to a new task (word learning), as well as a new contrast? If the phonetic representations acquired are fragile enough, it is possible that they might not, since word learning incorporates many skills: hearing and identifying the phoneme in the context of an entire word; mapping that word onto an image; and doing so while simultaneously trying to learn other word-image mappings. If the task is difficult enough and the representation weak enough, one might expect that it would not transfer.

To answer this question we compared overall performance on the word-learning task, the results of which are shown in Figure 4. As one would expect, participants in both groups were able to identify the *control* words above chance. The TRAINED group performed above chance on the *trained* words, which began with the sound they were trained on; however, they performed at chance on the *untrained* words.[8] By contrast, the CONTROL group was unable to distinguish words beginning with any of the unfamiliar phonemes above chance. There was no difference in performance over the

[5]For *trained:* $t(50) = 2.11, p = 0.04$, *untrained:* $t(39) = 2.68, p = 0.011$, both two-tailed. Note that the degrees of freedom for the *untrained* trials were adjusted from 50 to 39; this was because Levene's test for equality of variance indicated unequal variance.

[6]TRAINED group on *trained* stimuli: $t(24), p < 0.001$; on *untrained* stimuli: $t(24) = 5.03, p < 0.001$; CONTROL group on *trained* stimuli: $t(26), p < 0.001$; on *untrained* stimuli: $t(26), p < 0.01$.

[7]Difference between the first and second half of the test trials for the CONTROL participants: $t(26) = 1.87, p = 0.036$; for the TRAINED participants: $t(24) = 2.12, p = 0.022$, both one-tailed.

[8]Differences from chance (50%) performance for the TRAINED group: on words with the *control* contrast: $t(24) = 8.118, p < 0.001$; on words with the *trained* contrast: $t(24) = 2.941, p = 0.007$; on words with the *untrained* contrast: $t(24) = 0.282, p = 0.781$. For the CONTROL group: on words with the *control* contrast: $t(26) = 7.710, p < 0.001$; on words with the *trained* contrast: $t(26) = -0.090, p = 0.929$; on words with the *untrained* contrast: $t(26) = 0.991, p = 0.331$. All tests are two-tailed.

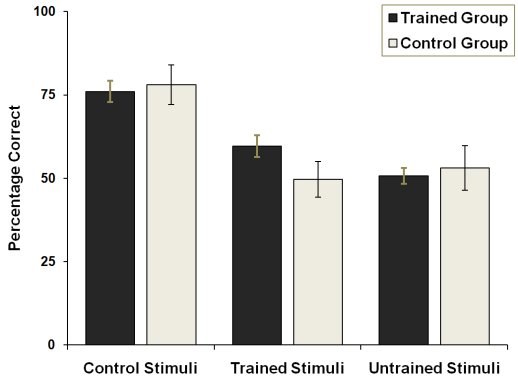


Figure 4: Word learning results. Participants in both groups were able to identify the correct words for the *control* stimuli above chance. The TRAINED group was above chance on words beginning with the sound they were trained on, but not on the related untrained sound. The CONTROL group, which was not trained on any phonemes, was unable to learn words beginning with both the *trained* and *untrained* sounds.

course of the word-learning task for any condition on any stimuli, suggesting that the task did not itself train phoneme discrimination.

Discussion

Motivated by the idea that differences between adult and child language learners may stem in part from initially minor differences that cascade throughout other aspects of language learning, we explored to what extent training adults on a previously unheard (novel) phonetic contrast results in improved learning of words that incorporate that contrast. Adults were assigned to either a TRAINED or CONTROL condition and trained distributionally, as in Maye and Gerken (2001). Both conditions were exposed to a bimodal distribution of phonetic sounds defined by voice onset time, but differed on whether the modes of the distribution mapped onto an existing phonetic contrast (the CONTROL condition: /d/ and /t^h/) or a novel contrast (the TRAINED condition: /g/ and /k/). We found that training on the phonetic contrast improved the learning of words beginning with that contrast, as well as the ability to discriminate a related contrast. These results have implications for how phonological abilities in adults affect other aspects of language learning, and for understanding how well distributional training enables phonetic generalization.

One interesting aspect of our findings is that as tasks became increasingly far removed from the original training, the ability to generalize diminished. The TRAINED group was able to generalize their phonetic learning to be able to discriminate a related but untrained contrast on a phonetic perception task, but when word learning was involved, they were only capable of learning words that began with the contrast they had been trained on. The CONTROL group was able to learn the *trained* and *untrained* contrast on the basis of the phonetic testing regime, but the resulting knowledge was more fragile than in the TRAINED group: their were unable to apply this ability to the problem of word learning. These results, in combination with the findings of other training studies (e.g., Bradlow et al., 1999; Maye & Gerken, 2001; McCandliss et al., 2002; Hayes-Harb, 2007), suggest that the ability to generalize phonetic learning (either to a related contrast or to another task) may depend strongly on the depth and nature of the training involved. It is possible that additional training would improve the ability to generalize even further. Relatedly, it is possible that our phoneme test did not measure phonetic category learning *per se*, and was more a measure of the raw ability to discriminate acoustically between two phonemes; if so, the limited generalization may have been due to the fact that our participants simply improved in their discrimination ability, but did not acquire phonetic categories in any reasonable sense (although the border between these two options is rather fuzzy). In general, the precise effect of training amount or type on generalization ability, and the nature of its dependence on the quantity and type of input, are matters for future study.

Our work was inspired in part by the idea that apparently major differences in language learning abilities may to some extent stem from smaller differences that have a cascading effect over time. While our findings are consistent with this notion, much work remains to be done to explore it more thoroughly, especially in the realm of adult language learning (research by Karmiloff-Smith and colleagues explores a similar idea in the area of language disorders). On one hand, it may appear unsurprising that being able to hear a phonetic contrast makes it easier to learn words that differ on that contrast. On the other hand, one might have expected phonological perception to have no effect on word learning: despite their poor perception, adults are arguably superior to children when it comes to acquiring vocabulary. Further work is essential, both for exploring whether linguistic abilities besides phonological perception affect other aspects of language, and for exploring whether phonological perception has effects on aspects of language besides word learning. This can include training studies like ours, as well as studies that evaluate how different aspects of language acquisition are affected by individual differences in adult phonetic perception (which are known to exist: see, e.g., McCandliss et al., 2002; Golestani & Zatorre, 2004; Perani, 2005; Golestani & Zatorre, 2009).

We conclude by noting an interesting puzzle: although the idea that deficiencies in one area of language acquisition can have cascading effects throughout other areas makes sense and is well-supported in the child acquisition literature (e.g., Tsao et al., 2004; Werker & Yeung, 2005; Fernald et al., 2006), so is the idea that jointly learning two aspects of language can improve performance in both (e.g., Feldman & Griffiths, 2009; Frank, Goodman, & Tenenbaum, 2009; Maurits, Perfors, & Navarro, 2009). However, the former implies that deficits in one area should propogate to another, while the latter implies that deficits in one area may be compensated for or overcome by skills or information from another. It is possible that both are true for different areas or in different ways, but as yet we know very little about the mechanisms or details

underlying either, so it is difficult to know for sure. As usual, further research is necessary.

References

Birdsong, D. (2006). Age and second language acquisition and processing: A selective overview. *Lang. Learning*, *56*(1), 9–49.

Bley-Vroman, R. (1990). The logical problem of foreign language learning. *Linguistic Analysis*, *20*, 3–49.

Bradlow, A., Akahane-Yamada, R., Pisoni, D., & Tohkura, Y. (1999). Training Japanese listeners to identify English /r/ and /l/: Long-term retention of learning in perception and production. *Perception & Psychophysics*, *61*(5), 977–985.

Christiansen, M., Onnis, L., & Hockema, S. (2009). The secret is in the sound: From unsegmented speech to lexical categories. *Developmental Science*, *12*(3), 388–395.

Clahsen, H., & Felser, C. (2006). How native-like is non-native language processing? *Trends in Cognitive Sciences*, *10*(12), 564–570.

Eimas, P., Siqueland, D., Jusczyk, P., & Vigorito, J. (1971). Speech perception in infants. *Science*, *171*, 303–306.

Elman, J., Bates, E., Johnson, M., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking innateness: A connectionist perspective on development*. Cambridge, MA: MIT Press.

Feldman, N., & Griffiths, T. (2009). Learning phonetic categories by learning a lexicon. In *31st Annual Conference of the Cognitive Science Society*.

Fernald, A., Perfors, A., & Marchman, V. (2006). Picking up speed in understanding: Speech processing efficiency and vocabulary growth across the 2nd year. *Dev. Psych.*, *42*(1), 98–116.

Fernald, A., & Simon, T. (1984). Expanded information contours in mothers' speech to newborns. *Dev. Psych.*, *20*, 104–113.

Flege, J. (1995). Second-language speech learning: Theory, findings, and problems. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 229–273). Timonium, MD: York Press.

Frank, M., Goodman, N., & Tenenbaum, J. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psych. Sci.*, *20*(5), 654–661.

Golestani, N., & Zatorre, R. (2004). Learning new sounds of speech: Reallocation of neural substrates. *Neuroimage*, *21*(2), 494–506.

Golestani, N., & Zatorre, R. (2009). Individual differences in the acquisition of second language phonology. *Brain & Language*, *109*, 55–67.

Guillelmon, D., & Grosjean, F. (2001). The gender marking effect in spoken word recognition: The case of bilinguals. *Memory & Cognition*, *29*, 503–511.

Hayes-Harb, R. (2007). Lexical and statistical evidence in the acquisition of second language phonemes. *Second Language Research*, *23*(1), 65–94.

Hernandez, A., Li, P., & MacWhinney, B. (2005). The emergence of competing modules in bilingualism. *Trends in Cognitive Sciences*, *9*(5), 219–224.

Hudson Kam, C., & Newport, E. (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language Learning and Development*, *1*(2), 151–195.

Iverson, P., Kuhl, P., Akahane-Yamada, R., Diesch, E., Tokura, Y., Kettermann, A., et al. (2003). A perceptual interference account of acquisition difficulties with non-native phonemes. *Cognition*, *87*, B47–B57.

Jamieson, D., & Morosan, D. (1989). Training new, nonnative speech contrasts: A comparison of the prototype and perceptual fading techniques. *Canadian Journal of Psychology*, *43*(1), 88–96.

Johnson, J., & Newport, E. (1989). Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology*, *21*, 60–99.

Karmiloff-Smith, A. (1998). Development itself is the key to understanding developmental disorders. *Trends in Cognitive Sciences*, *2*(10), 389–398.

Kuhl, P. (2004). Early language acquisition: Cracking the speech code. *Nature Reviews Neuroscience*, *5*, 831–843.

Lively, S., Logan, J., & Pisoni, D. (1993). Training Japanese listeners to identify English /r/ and /l/: II. the role of phonetic environment and talker variability in learning new perceptual categories. *Journal of the Acoustic Society of America*, *94*, 21242–1255.

MacWhinney, B. (2005). A unified model of language acquisition. In J. Kroll & A. De Groot (Eds.), *Handbook of bilingualism: Psycholinguistic approaches* (pp. 49–67). Oxford University Press.

Maurits, L., Perfors, A., & Navarro, D. (2009). Joint acquisition of word order and word reference. In *31st Annual Conference of the Cognitive Science Society*.

Mayberry, R. (1993). First-language acquisition after childhood differs from second-language acquisition: The case of American Sign Language. *Journal of Speech and Hearing Research*, *36*, 1258–1270.

Maye, J., & Gerken, L. (2001). Learning phonemes: How far can the input take us? In *25th annual conference of the BUCLD*.

Maye, J., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*, B101-B111.

Maye, J., Weiss, D., & Aslin, R. (2008). Statistical phonetic learning in infants: Facilitation and feature generalization. *Developmental Science*, *11*(1), 122–134.

McCandliss, B., Fiez, J., Protopapas, A., Conway, M., & McClelland, J. (2002). Success and failure in teaching the [r]-[l] contrast to Japanese adults: Tests of a Hebbian model of plasticity and stabilization in spoken language perception. *Cognitive, Affective, & Behavioral Neuroscience*, *2*(2), 89–108.

McClaskey, C., Pisoni, D., & Carrell, T. (1983). Transfer of training of a new linguistic contrast in voicing. *Perception and Psychophysics*(34), 323–330.

Mirman, D., Magnuson, J., Graf Estes, K., & Dixon, J. (2008). The link between statistical segmentation and word learning in adults. *Cognition*, *108*, 271–280.

Newport, E. (1988). Constraints on learning and their role in language acquisition: Studies of the acquisition of American Sign Language. *Language Sciences*, *10*, 147–172.

Pallier, C., Colomé, A., & Sebastián-Gallés, N. (2001). The influence of native-language phonology on lexical-access: Exemplar-based versus abstract lexical entries. *Psych. Sci.*, *12*, 445–449.

Perani, D. (2005). The neural basis of language talent in bilinguals. *Trends in Cognitive Science*, *9*(5), 211–213.

Sebastián-Gallés, N., & Soto-Faraco, S. (1999). Online processing of native and non-native phonemic contrasts in early bilinguals. *Cognition*, *72*, 111–123.

Shea, C., & Curtin, S. (2005). Learning allophones from the input. In *29th annual conference of the BUCLD*.

Singleton, J., & Newport, E. (2004). When learners surpass their models: The acquisition of American Sign Language from inconsistent input. *Cognitive Psychology*, *49*, 370–407.

Snow, C. (1999). Social perspectives on the emergence of language. In B. MacWhinney (Ed.), *The emergence of language* (pp. 257–276). Mahwah, NJ: Lawrence Erlbaum Associates.

Tan, L. (2003). Neural systems of second language reading are shaped by native language. *Human Brain Mapping*, *18*, 158–166.

Tsao, F., Liu, H., & Kuhl, P. (2004). Speech perception in infancy predicts language development in the second year of life: A longitudinal study. *Child Development*, *75*(4), 1067–1084.

Ullman, M. (2004). Contributions of memory circuits to language: The declarative/procedural model. *Cognition*, *92*, 231–270.

Wang, Y., Spence, M., Jongman, A., & Sereno, J. (1999). Training American listeners to perceive Mandarin tones. *Journal of the Acoustical Society of America*, *106*, 3649–3658.

Weber, A., & Cutler, A. (2003). Lexical competition in non-native spoken word recognition. *JML*, *50*, 1–25.

Werker, J., & Lalonde, C. (1988). Cross-language speech perception: Initial capabilities and developmental change. *Dev. Psych.*, *24*(5), 672–683.

Werker, J., & Tees, R. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant behavior and development*, *7*, 49–63.

Werker, J., & Yeung, H. (2005). Infant speech perception bootstraps word learning. *Trends in Cognitive Sciences*, *9*(11), 519–527.

The Influence of Within-Category Structure on Stimulus Similarity and Stimulus Generalization

James Close (james_close@eva.mpg.de)
Department of Developmental and Comparative Psychology, Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, 04103 Leipzig, Germany

Ulrike Hahn (hahnu@cardiff.ac.uk)
School of Psychology, Cardiff University, Tower Building, Park Place, Cardiff, CF10 3AT, UK

R. C. Honey (honey@cardiff.ac.uk)
School of Psychology, Cardiff University, Tower Building, Park Place, Cardiff, CF10 3AT, UK

Abstract

In Exp. 1, the authors report an influence of temporal contiguity in stimulus exposure on later judgments of similarity. Exposure to transformational information – that is, information that 'connects' two similar, but perceptually distinct stimuli – was found to have no influence on later judgments of similarity. In Exp. 2, exposure to transformational information was also found not to influence later property generalization; however, exposure to within-category structure that promoted a sense of 'surprise' (i.e., contained clear discontinuity) led to a reduction in later property generalization between two similar, but perceptually distinct stimuli. This latter effect was confirmed in Exp. 3 while ruling out any influence of temporal factors.

Keywords: Spontaneous categorization; within-category structure; similarity; generalization; transformational knowledge; temporal dynamics; perceptual learning; sensory preconditioning.

Introduction

Similarity and categorization are intimately intertwined: stimulus similarity is assumed to form the basis for many of our natural categories (Hampton, 2001), but categorization can also alter perceived similarity (Harnad, 1987). When taught that stimuli are members of the same category, participants will perceive these stimuli to be more similar than participants who are not (e.g., Livingston, Andrews, & Harnad, 1998). The reverse is also true: when taught that stimuli are members of contrasting categories, participants will perceive these stimuli to be more different (e.g., Goldstone, 1994). Moreover, given the lawful relationship that exists between similarity and stimulus generalization (Shepard, 1987), it is not surprising that many studies have shown that stimulus generalization is directly influenced by the 'classificatory status' of stimuli: when stimuli are 'classified together' (or acquire equivalence), increased levels of stimulus generalization are found between these stimuli. In contrast, when stimuli are 'classified apart' (or acquire distinctiveness), decreased levels of stimulus generalization are found between them (see Honey, Close, & Lin, 2010). In other words, categorization can warp psychological similarity space (Nosofsky, 1989).

While interesting, almost all studies to date that have indexed an influence of categorization on later judgments of similarity and stimulus generalization (commonly termed *categorical perception* (CP)) have employed supervised training procedures (but see Gureckis & Goldstone, 2008). Consequently, as Gureckis and Goldstone have noted, "it remains a somewhat opaque question if learned CP effects are restricted to cases where subjects make a differential response to each category or if other aspects of category organization, such as the similarity structure or distribution of items within a category, may also exert an influence on perception" (2008, p. 1876). This is important because although one may presume that the mechanisms of supervised categorization drive all classification, evidence has shown that this is unlikely to be the case (Pothos & Chater, 2002). To fully assess categorization's influence on later behavior, therefore, one needs to look to unsupervised categorization – that is, categorization that occurs in the absence of any external feedback.

Fundamentally, unsupervised categorization tasks afford an assessment of the principles that underlie categorization in an unconstrained manner, allowing greater insight into people's natural categorization biases (or preferences). However, much of the unsupervised categorization that occurs in the laboratory has been considered very different to that which occurs naturally (see Clapper & Bower, 1994; Love, 2002). Crucially, whereas any natural unsupervised category formation will have unlikely been the primary purpose of an interaction (meaning that any category formation is incidental), in laboratory investigations of unsupervised categorization, explicit instruction to categorize is generally given, meaning that any category formation is intentional (Love, 2002). Unlike the majority of laboratory-based unsupervised categorization, then, natural incidental categorization requires that a person first realize there is

some structure present, and then utilize this structure to guide their classifications[1]. The experiments presented in this paper, therefore, sought to assess how the similarity structure (i.e., the distribution of items) within a category influences incidental categorization, as indexed by the later perceived similarity of category items, and the level of generalization between category items.

What aspects of within-category structure might influence whether stimuli are incidentally classified together or apart? Zaki and Homa (1999) have proposed that the acquisition of an object concept will be facilitated by exposure to that object's successive changes (that is, exposure to transformational information). Based on this hypothesis, it seems plausible to suppose that transformational information should encourage the incidental 'classification together' of similar, but distinct stimuli (but see the categorical perception effects of Newell & Bülthoff, 2002). Another factor that might also encourage the 'classification together' of stimuli is temporally contiguous stimulus exposure (see, e.g., Bateson & Chantrey, 1972). Empirical investigation into the phenomenon of perceptual learning has shown that the temporal dynamics of stimulus exposure influence whether an increase or decrease in later perceived stimulus similarity (and stimulus generalization) is found (see Goldstone, 1998; Hall, 1991). More specifically, when two similar stimuli are exposed in close temporal contiguity, the perceived similarity of (and the level of generalization between) these stimuli should increase, relative to situations where no stimulus exposure is given and where stimulus exposure is not particularly temporally contiguous (see Bateson & Chantrey, 1972; Bennett & Mackintosh, 1999). Finally, a number of theories of spontaneous category learning link the formation of new categories (or clusters) to unexpected changes in stimulus structure. For example, Clapper and Bower (1994, 2002; see also SUSTAIN, Love, Medin, & Gureckis, 2004) propose that if a novel stimulus is perceived as sufficiently 'surprising' (sufficiently dissimilar) to previously stored stimulus encounters, then a new category (cluster) will be invented to accommodate that stimulus. Consequently, if a strong set of norms has been established about, for example, Category A membership (i.e., through a number of exposures to Category A exemplars), then it is more likely that a Category B exemplar will be accommodated in a newly invented category (cluster). Exposure to only a single Category A exemplar before exposure to a Category B exemplar, by contrast, will likely not lead to these stimuli being 'classified apart' (Clapper & Bower, 1994, 2002; Love et al., 2004).

In summary, much evidence has shown that categorization (using supervised training procedures) can alter the perceived similarity of stimuli, and concomitantly, the level of generalization between stimuli. While there is some preliminary evidence that similar alterations in perceived stimulus similarity can be found following unsupervised categorization (Gureckis & Goldstone, 2008), little research has directly assessed how the similarity structure (i.e., the distribution of items) within a category influences incidental categorization. Moreover, the discrimination based studies that have indexed an influence of categorization on stimulus similarity and stimulus generalization have typically employed designs in which participants engage in hundreds of experimental trials. However, it seems reasonable to suppose that people's sensitivity to category structure (if sufficiently obvious) should be immediate. This means that under certain conditions, incidental categorization should be a rapid process that can occur following only minimal stimulus exposure.

Experiment 1

In Exp. 1, we were interested in investigating those factors that should encourage incidental 'classification together' under conditions of minimal stimulus exposure. Specifically, we sought to test the hypotheses that transformational information and temporally contiguous stimulus exposure should encourage the 'classification together' of two similar, but distinct stimuli, as indexed by a later increase in their perceived similarity to one another.

Method

Participants 48 Cardiff University undergraduate students took part either for partial fulfillment of course credit or a small payment of £2, with 16 participants in each condition (see Table 1).

Table 1: The three conditions employed to assess within-category structure in Exp. 1.

Condition	Preexposure	Test
Baseline	A / - / - / - / - / F	A - F
Sys_trans	A / B / C / D / E / F	A - F
Contiguous	A / F	A - F

Stimuli The stimuli were individually rendered images taken with permission from Hahn, Close and Graf (2009). They were basic level objects from six biological categories (bird, fish, head, mushroom, starfish, turnip) and one artifact category (light bulb; see Figure 1). For every category, two objects formed the endpoints of each morph continuum (the 1% and 100% morph stimuli), from which 20%, 40%, 60% and 80% morph images were rendered (here, the 1%, 20%, 40%, 60%, 80% and 100% images are referred to as A, B, C, D, E and F, respectively). All morph images had a size of 256 × 256 pixels and were presented in gray scale on a 15-in.

[1] This contrasts with laboratory-based unsupervised categorization where the explicit instruction to categorize will likely promote a belief in participants that their task is to find some experimenter defined category structure.

computer monitor. Participants were seated approximately arms length from the monitor for the duration of the experiment.

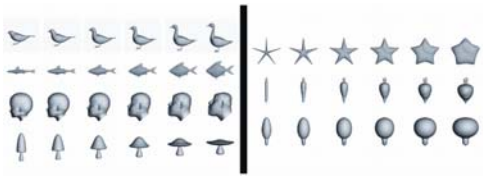


Figure 1. Illustration of the morph stimuli employed. The stimuli shown here are the 1%, 20%, 40%, 60%, 80%, and 100% morph images, respectively.

Design and Procedure Exposure condition was manipulated as a between-participants factor and participants in all conditions were exposed to the seven different object categories. On a given trial, participants were sequentially preexposed to a set of morph stimuli from one of the object categories. Within each of the three exposure conditions, half of participants received presentations of the morph stimuli in the order A to F, and half of participants received presentations of the morph stimuli in the order F to A. Each stimulus was presented for 3000 ms, and the temporal spacing between presentation of stimulus A and stimulus F was held constant in the Baseline condition and condition Sys_trans by introducing a fixation cross when no morph (object) stimulus was scheduled to be presented in the Baseline condition, relative to condition Sys_trans. Following stimulus preexposure, a 1000 ms inter-stimulus interval (blank screen) separated presentation of the test screen, on which was presented stimulus A and stimulus F. Within the subconditions created in each exposure condition following the previous counterbalancing operation, half of participants saw stimulus A surrounded by a red border on the test screen, and half of participants saw stimulus F surrounded by a red border on the test screen. Within each of the subconditions created by the previous counterbalancing operations, half of participants received presentations of stimulus A on the left-hand side of the test screen and presentations of stimulus F on the right-hand side of the test screen, and half of participants received the reverse. On the test screen, participants were simply asked to rate how similar they thought the object framed in red was to the object not framed in red, using a 1 (very dissimilar) to 9 (very similar) rating scale presented at the bottom of the test screen. Responses were made using the keys "1" through "9" on a standard keyboard. Following a response, a 1000 ms inter-trial interval (blank screen) separated participants' exposure to the next object category. Exposure to the seven object categories was random for all participants in each of the three exposure conditions.

Results

For the purpose of analyses, participant similarity judgments were averaged over the seven object categories. Figure 2 displays the results of interest: participants' overall mean similarity rating, split by preexposure condition. Inspection of this figure reveals that similarity ratings in condition Contiguous were higher than in the Baseline condition and condition Sys_trans. Similarity ratings in condition Sys_trans differed little from those in the Baseline condition. A one-way ANOVA revealed a significant effect of exposure condition, $F(2, 45) = 7.31$, $p < .003$, $\eta^2 = .25$. Tukey HSD post-hoc tests revealed that, overall, participants in the Contiguous condition reported significantly higher ratings of similarity than participants in the Baseline condition ($p < .05$) and Sys_trans condition ($p < .002$). Overall similarity ratings did not differ significantly between the Baseline and Sys_trans conditions ($p > .05$).

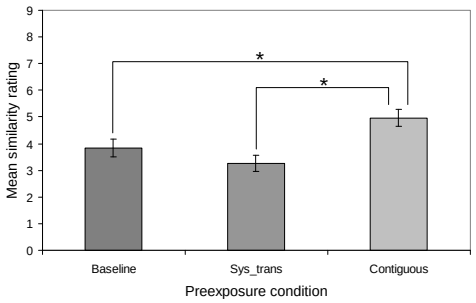


Figure 2. Results of Exp. 1: overall mean similarity rating, plotted by preexposure condition. Error bars indicate the standard error.

Discussion

In Exp. 1, the perceived similarity of stimuli A and F was influenced only by the temporal contiguity of preexposure to these stimuli. One interpretation of this result is that only the Contiguous condition was sufficient to encourage the 'classification together' of stimuli A and F. This 'classification together' can be conceptualized in a number of ways: one way of conceptualizing such is in terms of the formation of a blended representation of stimulus A and stimulus F (i.e., AF; see Hall, 1991). Alternatively, an account can be considered with respect to the assumption that temporally contiguous stimulus exposure provides the optimal conditions under which an excitatory association can form between two similar stimuli (Hall, 1991). Formation of such an A–F association would dictate that stimulus A will evoke a representation of stimulus F, creating a situation in which these stimuli will come to be perceived (somewhat) equivalently. Such *acquired equivalence* would lead to a concomitant increase in the perceived similarity of stimuli A and F (see Hall, 1991).

Interestingly, the results of Exp. 1 do not support the proposal of Zaki and Homa (1999). A number of possibilities exist for this failure: first, Zaki and Homa's (1999) proposal may simply be wrong. Second, the within-category similarity structure of condition

Sys_trans may have resulted in both associationistic and comparator processes operating (Honey, Bateson & Horn, 1994). If one assumes that the influence of these two processes was relatively balanced in condition Sys_trans, then this would have resulted in little change in the perceived similarity of stimuli A and F, relative to their baseline similarity.

In Exp. 2, we sought to further assess the influence of within-category structure using a property generalization task at test. Here, we were interested in investigating whether we could find evidence for 'classification apart' – driven by a surprise-driven category invention mechanism (Clapper & Bower, 1994, 2002) – under conditions of minimal stimulus exposure. To this end, we compared a skewed stimulus structure (condition Surprise) to the Baseline and Sys_trans conditions of Exp. 1, and a further scrambled transformational information condition (condition Scram_trans).

Experiment 2

In Exp. 2, we sought to assess the hypothesis that the Surprise condition – in which participants were exposed to a skewed stimulus structure – would lead to the 'classification apart' of stimuli A and F, as indexed by a later reduction in the level of property generalization between them. Such a finding would provide support for a surprise-driven category invention mechanism operating in incidental categorization (Clapper & Bower, 1994, 2002).

Method

Participants 64 Cardiff University students took part for partial fulfillment of course credit, with 16 participants in each condition (see Table 2).

Table 2: The four conditions employed to assess within-category structure in Exp. 2.

Condition	Preexposure	Conditioning	Test
Baseline	A / - / - / - / - / F	A+	F
Surprise	A / B / C / - / - / F	A+	F
Sys_trans	A / B / C / D / E / F	A+	F
Scram_trans	A / E / C / D / B / F	A+	F

Stimuli, Design and Procedure The same stimuli used in Exp. 1 were employed. As for Exp. 1, on a given trial, participants were sequentially preexposed to a set of morph stimuli from one of the object categories. Within each of the four exposure conditions, half of participants received presentations of the morph stimuli in the order A to F, and half of participants received presentations of the morph stimuli in the order F to A. Each stimulus was presented for 3000 ms, and the temporal spacing between presentation of stimulus A and stimulus F was held constant across conditions by introducing a fixation cross when no morph (object) stimulus was scheduled to be presented, relative to conditions Sys_trans and Scram_trans. Within the subconditions created by the previous counterbalancing operation applied in each preexposure condition, following a 1000 ms inter-stimulus interval (blank screen), half of participants were then presented with stimulus A, and half of participants were then presented with stimulus F. Situated above the stimulus was a sentence that informed participants about a particular property that the stimulus had: for example, "This person comes from a small, remote island in the Pacific Ocean". This information remained on the screen until the space bar was pressed, at which point participants were immediately presented with the test screen. On the test screen, participants were simply asked to rate on a scale from 1 (very unlikely) – 9 (very likely) how likely they thought it was that the stimulus now presented to them shared the property of the previously seen stimulus. If participants had previously been presented with stimulus A, then at test, they were presented with stimulus F, and if they had previously been presented with stimulus F, then at test, they were presented with stimulus A. The 1 – 9 rating scale was continuously presented beneath the test stimulus, and responses were made using the 1 – 9 keys on the top of a standard computer keyboard. A 1000 ms inter-trial interval (blank screen) separated participants' likelihood ratings and their exposure to the next object category. Exposure to the seven object categories was random for all participants in each of the four preexposure conditions.

Results

Again, for the purpose of analyses, participant similarity judgments were averaged over the seven object categories. Figure 3 shows the results of the generalization test: the overall mean likelihood ratings that the test stimulus shared the property of the previously seen stimulus, split by preexposure condition. Inspection of this figure reveals that participants in the Surprise condition reported lower mean likelihood ratings than participants in the other three preexposure conditions; likelihood ratings in the other three conditions were all very similar.

A one-way ANOVA[2] confirmed that there was an overall effect of preexposure condition, $F(3, 40.51) = 2.85$, $p < .05$, $\eta^2 = .12$. Dunnett T3 post-hoc tests (equal variances not assumed)[3] revealed that, overall, participants in the Surprise condition reported significantly lower mean likelihood ratings than participants in the Baseline condition ($p < .05$, $r = .35$). No other post-hoc comparisons were significant (all ps > .05).

[2] Due to a lack of homogeneity of variances between conditions, the Brown-Forsythe correction was applied.

[3] Tukey HSD post-hoc tests were not performed (as in Exp. 1) due to the lack of homogeneity of variances between conditions.

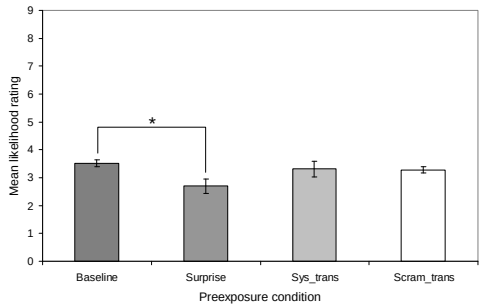


Figure 3. Results of Exp. 2: overall mean likelihood ratings, plotted by preexposure condition. Error bars indicate the standard error.

Discussion

The results of Exp. 2 are broadly consistent with the predictions of a surprise-driven category invention mechanism operating in incidental categorization (Clapper & Bower, 1994, 2002; also Love et al., 2004). This assumes that only the within-category similarity structure of the Surprise condition encouraged participants to invent an extra category (cluster) in which to accommodate the lone distinct stimulus, meaning that stimuli A and F were 'classified apart'. As a consequence of this, property generalization between A and F in the Surprise condition was reduced (Harnad, 1987).

In line with the results of Exp. 1, it is apparent that transformational information did not encourage the 'classification together' of stimuli A and F and a concomitant increase in the level of property generalization between A and F (condition Baseline vs condition Sys_trans). Moreover, there is no evidence to suggest that systematic transformational information influenced participants' response behavior differently to non-systematic transformational information (condition Sys_trans vs condition Scram_trans; cf. Zaki & Homa, 1999).

What aspect of the within-category structure of the Surprise condition encouraged the assumed 'classification apart' of stimuli A and F? Inspection of this structure reveals that not only does it have a similarity structure likely to engage a surprise-driven category invention mechanism, but also a distinct temporal structure. That is, while the three stimuli with the highest perceptual similarity were presented in a temporally contiguous manner, a temporal gap of six seconds separated presentation of the distinct stimulus from the highly similar stimuli. It is possible, therefore, that it was this temporal discontinuity, rather than the perceived perceptual discontinuity, that engendered the assumed invention of a new category (cluster) in which to accommodate the distinct stimulus.

Experiment 3

Exp. 3 replicated Exp. 2 with one exception: in order to determine if the temporal discontinuity contained within the Surprise condition of Exp. 2 was critical in producing the significant difference between the Baseline and Surprise conditions, the stimuli in condition Surprise_2 were preexposed with even temporal spacing.

Method

Participants 32 Cardiff University students took part for a small payment of £2, with 16 participants in each condition (see Table 3).

Table 3: The two conditions employed to assess within-category structure in Exp. 3.

Condition	Preexposure	Conditioning	Test
Baseline	A / - / - / - / - / F	A+	F
Surprise_2	A / B / C / F	A+	F

Stimuli, Design and Procedure The only difference to Exp. 2 was that during the preexposure phase of the Surprise_2 condition, presentations of the morph stimuli were separated by a 2000 ms long fixation cross. This maintained the equivalent temporal spacing between presentations of the object category endpoints (A and F) across the two conditions.

Results

Figure 4 shows the results of interest: the overall mean likelihood ratings split by preexposure condition. Inspection of Figure 4 shows that, overall, participants in the Surprise_2 condition reported significantly lower likelihood ratings than participants in the Baseline condition, $F(1, 30) = 6.14$, $p < .02$, $\eta^2 = .17$[4].

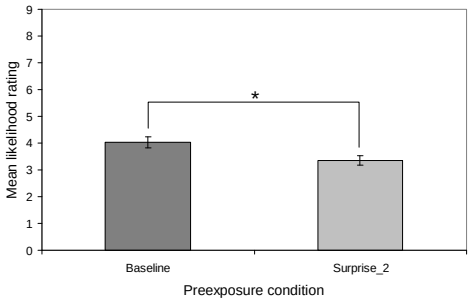


Figure 4. Results of Exp. 3: overall mean likelihood ratings, plotted by preexposure condition. Error bars indicate the standard error.

General Discussion

The method of these studies provides a fast and effective way of assessing the influence of within-category structure (i.e., the distributional properties of the stimuli)

[4] Due to a violation of normality in the data (Shapiro-Wilk test of normality, $p < .007$), we confirmed this result using the non-parametric Mann-Whitney U test, $U(16, 16) = 56.50$, $p < .008$, $r = .49$.

on people's incidental categorization behavior, as indexed by their later judgments of stimulus similarity and stimulus generalization. Indeed, one particularly notable feature of the designs of Experiments 1 – 3 is that participants only received a single presentation of each scheduled stimulus during preexposure.

Two main findings were made: First, transformational information did not encourage 'classification together', which would have resulted in a later increase in the perceived similarity of stimuli A and F (Exp. 1) and an increase in the level of property generalization between these stimuli (Exp. 2). Second, when perceptual discontinuity existed in the presented within-category structure, this resulted in a reduction in the level of later property generalization between stimuli A and F (Exp. 2 and Exp. 3). This result is consistent with the assumption of a surprise-driven category invention mechanism operating in human incidental categorization (Clapper & Bower, 1994, 2002; also Love et al., 2004), and supports previous work by Gureckis and Goldstone (2008). Importantly, the results of Exp. 2 demonstrate that this reduction in stimulus generalization was not simply a product of the amount of stimulus preexposure.

In conclusion, the present results support the idea that perceived discontinuity in the environment (be this temporal or perceptual) guides people's incidental categorization behavior, as indexed by their later judgments of stimulus similarity and stimulus generalization (Anderson, 1991; Rosch & Mervis, 1975). Finally, one of the particularly nice aspects of the design of Exps 2 and 3 is that it can be readily transposed to assessments of incidental categorization in nonhuman animals and prelinguistic children.

Acknowledgments

This work was supported by a Cardiff University, School of Psychology studentship to J. Close.

References

Anderson, J.R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*, 409-429.

Bateson, P.P., & Chantrey, D.F. (1972). Retardation of discrimination learning in monkeys and chicks previously exposed to both stimuli. *Nature, 237,* 173-174.

Bennett, C.H., & Mackintosh, N.J. (1999). Comparison and contrast as a mechanism of perceptual learning? *Quarterly Journal of Experimental Psychology, 52B,* 253-272.

Clapper, J.P, & Bower, G.H. (1994). Category invention in unsupervised learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 443-460.

Clapper, J.P., & Bower, G.H. (2002). Adaptive categorization in unsupervised learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28,* 908-923.

Goldstone, R.L. (1994). The role of similarity in categorization: Providing a groundwork. *Cognition, 52,* 125-157.

Goldstone, R.L. (1998). Perceptual Learning. *Annual Review of Psychology, 49,* 585-612.

Gureckis, T.M., & Goldstone, R.L. (2008). The effect of internal structure of categories on perception. *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 1876-1881). Washington, D.C.: Cognitive Science Society.

Hall, G. (1991). *Perceptual and associative learning*. New York: Oxford University Press.

Harnad, S. (1987). *Categorical Perception: The Groundwork of Cognition*. New York: Cambridge University Press.

Hahn, U., Close, J., & Graf, M. (2009). Transformation direction influences shape similarity judgments. *Psychological Science, 20,* 447-461.

Hampton, J.A. (2001). The Role of Similarity in Natural Categorization. In U. Hahn & M. Ramscar (Eds.), *Similarity and Categorization*. Oxford: Oxford University Press.

Honey, R.C., Bateson, P., & Horn, G. (1994). The role of stimulus comparison in perceptual learning: An investigation with the domestic chick. *Quarterly Journal of Experimental Psychology, 47B,* 83-103.

Honey, R.C., Close, J., & Lin, T.E. (2010). Acquired distinctiveness and equivalence: A synthesis. In C.J. Mitchell & M.E. Le Pelley (Eds.), *Attention and learning*. Oxford: Oxford University Press.

Livingston, K.R., Andrews, J.K., & Harnad, S. (1998). Categorical Perception Effects Induced by Category Learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 732-753.

Love, B.C. (2002). Comparing supervised and unsupervised category learning. *Psychonomic Bulletin & Review, 10,* 190-197.

Love, B.C., Medin, D.L., & Gureckis, T.M. (2004). SUSTAIN: A Network Model of Category Learning. *Psychological Review, 111,* 309-332.

Newell, F.N., & Bülthoff, H.H. (2002). Categorical perception of familiar objects. *Cognition, 85,* 113-143.

Nosofsky, R.M. (1989). Further tests of an exemplar-similarity approach to relating identification and categorization. *Perception & Psychophysics, 45,* 279-290.

Rosch, E., & Mervis, C.B. (1975). Family Resemblances: Studies of the Internal Structure of Categories. *Cognitive Psychology, 7,* 573-605.

Shepard, R.N. (1987). Toward a universal law of generalization for psychological science. *Science, 237,* 1317-1323.

Zaki, S.R., & Homa, D. Concepts and Transformational Knowledge. *Cognitive Psychology, 39,* 69-115.

Conservatism in Belief Revision and Participant Skepticism

Adam Corner (corneraj@cardiff.ac.uk)
School of Psychology, Cardiff University, Park Place, Cardiff, CF10 3AT, UK.

Adam J.L. Harris (a.j.l.harris@warwick.ac.uk)
Department of Psychology, University of Warwick, Coventry, CV4 7AL, UK.

Ulrike Hahn (hahnu@cardiff.ac.uk)
School of Psychology, Cardiff University, Park Place, Cardiff, CF10 3AT, UK

Abstract

Comparing the responses of participants in reasoning experiments to the normative standard of Bayes' Theorem has been a popular empirical approach for almost half a century. One longstanding finding is that people's belief revision is conservative with respect to the normative prescriptions of Bayes' Theorem, that is, beliefs are revised less than they should be. In this paper, we consider a novel explanation of conservatism, namely that participants do not perceive information provided to them in experiments as coming from a fully reliable source. From the Bayesian perspective, less reliable evidence should lead to more conservative belief revision. Thus, there may be less of discrepancy between normative predictions and behavioural data than previously assumed.

Keywords: Belief revision; Conservatism; Bayesian; Experimental Pragmatics.

Introduction

Bayes' Theorem provides a normative rule for updating beliefs in the light of new evidence, and therefore provides a valuable tool for studying human reasoning. In particular, participants' responses in experiments can be compared to normative predictions derived from Bayes' Theorem. There is a wealth of experimental data using the framework of Bayesian probability to study almost every aspect of human reasoning including judgement (Tversky & Kahneman, 1983), decision making (Edwards & Tversky, 1967), conditional reasoning (Evans & Over, 2004; Oaksford & Chater, 2003), category based induction (Kemp & Tenenbaum, 2009) and argumentation (Hahn & Oaksford, 2007).

Demonstrations of seemingly non-Bayesian reasoning behaviour abound, but the debate about whether people's reasoning behaviour can be considered normative has continued because deviations from supposedly rational standards have led to discussion about the standards themselves.

For example, Simon's notion of 'bounded rationality' (Simon, 1982) has led some researchers to focus on the adaptive value of cognitive strategies as the gold standard for rationality (Gigerenzer & Todd, 1999). Others (Hilton, 1995: Noveck & Sperber, 2004; Schwarz, 1996) have asked whether participants and experimenters share the same normative model – that is, are participants in reasoning experiments doing what experimenters *think* they are doing? These researchers propose that many of the most seemingly compelling demonstrations of irrationality may be attributable – at least in part – to the *pragmatics* of the experimental setting.

One question of fundamental importance in the debate about Bayesian rationality is whether or not people revise their beliefs in line with Bayesian predictions when they encounter new evidence. A consistent finding is that people are *conservative* relative to the predictions of Bayes' Theorem (Edwards, 1968; Fischoff & Beyth-Marom, 1983; Slovic & Lichtenstein, 1972). The provision of new evidence does not seem to have the impact on people's existing beliefs that Bayes' Theorem predicts it should.

In the following section we review some putative explanations for conservatism. We then propose that a consideration of the pragmatics of belief revision experiments suggests a novel explanation for conservatism: Participants do not treat the evidence they receive in belief revision experiments as fully reliable, and therefore do not 'maximally' revise their beliefs. Bayesian theory itself requires that less reliable evidence should lead to more conservative updating. Thus, conservatism in belief revision may reflect, at least in part, a normatively appropriate response to receiving evidence from a less than fully reliable source.

Conservatism

Conservatism in belief revision is a well-documented experimental finding. In a variety of different contexts, people have been shown to revise their beliefs more weakly than Bayes' Theorem predicts that they should when they encounter seemingly diagnostic evidence. In a typical conservatism experiment, participants are shown two 'bookbags', and told that they are filled with different distributions of red and blue 'chips' (Edwards, 1968; Peterson & Miller, 1964; Peterson, Schneider & Miller, 1965). For example, Bag A might contain 60% red chips and 40% blue chips, while Bag B contains 40% red chips and 60% blue chips. One of the bags is 'selected at random', and chips sequentially drawn from it (in reality, the distribution and the ordering of the chips is typically predetermined by the experimenter). Participants must judge which of the two bookbags the chips are being drawn from, using each new piece of evidence to update their existing beliefs.

Bayes' Theorem is a normative rule for updating beliefs based on new evidence:

$$P(H \mid E) = \frac{P(H)P(E \mid H)}{P(H)P(E \mid H) + P(\neg H)P(E \mid \neg H)} \quad \text{Eq. 1}$$

It allows calculation of posterior belief, $P(H|E)$, that is, one's belief in the hypothesis in light of the evidence received. The posterior is determined by one's prior degree of belief, $P(H)$, and the diagnosticity of the evidence received, that is, how much more likely it is that the evidence observed would have occurred if the hypothesis were true, $P(E|H)$, in this case the chips were drawn from Bag A, as opposed to if it were *false* (i.e., the chips were drawn from Bag B), $P(E|\neg H)$. In signal detection terms, these two quantities correspond to the *hit rate* and *false positive rate* associated with the evidence. The ratio between them, which captures the diagnosticity of this evidence is referred to as the likelihood ratio. The posterior degree of belief brought about increases as this likelihood ratio increases, as seen in Figure 1.

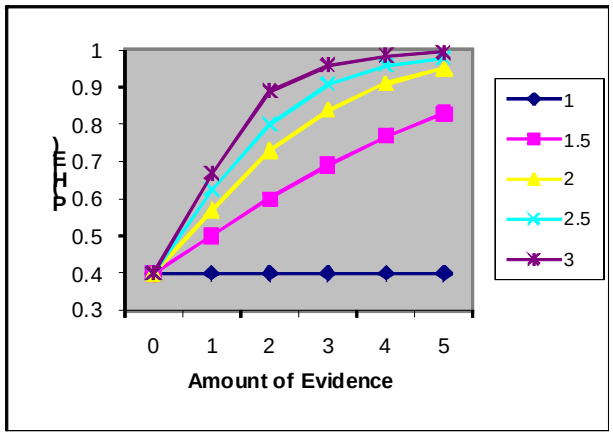


Figure 1: Impact of amount of evidence and source reliability (likelihood ratio) on posterior belief in a hypothesis. The figure plots posterior degrees of belief after receiving a unit of evidence of given diagnosticity, starting from a prior of .4. Each line represents a different likelihood ratio.

Each drawn chip represents a new piece of evidence, and thus provides information about which of the two 'hypotheses' is likely to be true (i.e., which of the two bookbags the sample is drawn from). As more evidence is obtained, participants should come to believe that one hypothesis is more likely to be true than the other. The dominant finding from the 'bookbag and poker chip' experiments is that this happens more slowly, and to a lesser extent than Bayes' Theorem predicts it should (Edwards, 1968; Peterson & Miller, 1964; Peterson, Schneider & Miller, 1965).

The finding that people tend to consistently underestimate the diagnostic impact of evidence unsurprisingly triggered a great deal of debate. Edwards (1968) suggested that people could either be mis-aggregating or misperceiving the true diagnostic value of evidence. Both of these explanations assume, however, that the 'true' value of the evidence is objectively known and available to both participant and experimenter (an assumption that we discuss in more detail below). By contrast, Slovic & Lichtenstein (1971) proposed a range of possible explanations for experimental conservatism in belief revision, including the idea that participants in reasoning experiments may anchor themselves to their initial beliefs and be unwilling to change them in the light of new evidence. This explanation does not assume that participant and experimenter necessarily assign the same weight to the evidence, but instead holds that people are too wedded to their initial assessments to properly incorporate new evidence.

In their review of the literature, Erev, Wallsten & Budescu (1994) conclude that while conservatism in belief revision is a fairly robust experimental finding, the locus of conservatism in participants' revisions of their opinions has never been definitively established. Mis-aggregation, misperception, and 'anchoring' are all explanations of conservatism that infer a normative fault in participants' responses – that is, participants' responses are viewed as non-Bayesian. But does conservatism in experimental demonstrations of belief revision really indicate a normative fault in participants' reasoning?

Edwards (1968) proposed a third explanation: Conservatism could simply be an experimental artefact. Edwards suggested that people become confused in experimental contexts that involve complex tasks, find it difficult to process all the explicit numerical information, and thus make performance errors. Slovic & Lichtenstein (1971; see also Erev et al., 1994) observed that people find the presentation (and production) of numerical probabilities difficult to deal with (as they do not typically come across explicit numerical probabilities in their daily lives). In addition, Slovic & Lichtenstein suggested that people are unwilling to use the extreme values of response scales, and that their responses therefore converge on central values. Similarly, Lopes (1985) suggested that non-Bayesian behaviour might be less likely to occur in situations where stimuli were more clearly 'marked' in support of or against a given hypothesis. Lopes (1987) succeeded in improving the match between participants' responses and normative predictions in a belief revision experiment by instructing them to separate their judgments into two steps. First participants labelled a piece of evidence as either favouring or countering a hypothesis, and then they made an estimate of how *much* it favoured one hypothesis of the other.

This second class of explanations locate the normative fault not with participants' responses, but with the nature of the experimental setting. Might conservatism in belief revision be more attributable to faulty assumptions on behalf of the experimenter than faulty reasoning on behalf of the participants?

The Pragmatics of Experiments

The normative construal of an experimental task can have wide-ranging implications (Hilton, 1995: Noveck & Sper-

ber, 2004; Schwarz, 1996). The key insight is that in order to be able to accurately understand behaviour in an experiment, it is vitally important to have a complete understanding of what the *participants* in the experiment think they are doing, in case it differs from what the *experimenters* think they are doing. Yet in many experiments the routine assumption is that participants' representation of the experimental task simply matches that of the experimenter.

Increasingly, some researchers have based their analyses of reasoning, judgement or decision making behaviour on the pragmatic, Gricean notion of *conversational implicature* – information that is not contained in the literal content of an utterance, but that can be implied from the context in which it is given (Grice, 1975). The notion of implicature is central to an understanding of the pragmatics of experiments: participants may infer more about the experiment than is contained in the literal content of the instructions. Similarly, experimenter and participant might have different ideas about what key task parameters are – such as the diagnosticity of the evidence in belief revision experiments.

Why might participants differ in their assessment of how diagnostic the evidence in belief revision experiments is? One possible explanation is that participants simply do not maximally trust the evidence they receive. In fact, several studies have investigated the idea that participants' trust in the context of experiments may be affected by participating in previous experiments – particularly if these experiments involved a deceptive manipulation.

Kelman (1967) proposed that the frequent use of deception in social psychological experiments was creating a new, suspicious breed of participant, who did not trust the experimenter and would be unlikely to react in a natural way. Christensen (1977) investigated the idea of the 'negative subject' empirically, and found that participants who were exposed to a prior experimental manipulation (not necessarily a deceptive manipulation) produced 'negative subject' responses, as demonstrated by a failure to exhibit verbal conditioning as effectively as subjects who had not received a prior manipulation. Similarly, Cook & Perrin (1971) found that experiencing deception caused a decrement in incidental learning in an immediately consecutive task – participants were more vigilant to the messages they were presented with, and therefore scrutinised them more carefully.

More recently, McKenzie, Wixted & Noelle (2004) observed that many demonstrations of supposedly irrational behaviour in the laboratory rely on the assumption that participants believe "key task parameters that are merely asserted by experimenters" (p947). McKenzie et al. then considered seeming rationality deficits in the context of changes in confidence judgments across yes-no and forced choice formats of the same cognitive task. Here previous empirical research has suggested that people's performance is suboptimal or irrational by comparison with the appropriate normative model. McKenzie et al. explicitly modelled participant skepticism toward aspects of the experimental materials. By including a 'believability' or 'confidence' parameter, the authors hoped to establish whether performance on such tasks was truly irrational (non-normative), or whether participants might actually be responding reasonably, given their understandable skepticism about task realism. Participant performance was found to be entirely in keeping with this modified normative model and hence rational.

The findings from McKenzie et al. (2004) suggest that the believability of experimental materials is likely to have a profound effect on experimental data. Noting that psychological experiments routinely involve systematic deception, the authors suggested that "maybe the only irrational thing to do in any experiment is to fully believe anything the experimenter tells you" (p.956).

This is a strong statement to make about the demands of the experimental setting. We do not wish to convey that participants in psychological experiments actively undermine experimental manipulations by seeking to discredit the information they receive. But the opposing assumption – that all information given to participants by experimenters is taken at face value – seems equally implausible. It seems possible that participants do not treat information they are given in experiments as deriving from a maximally reliable source.

Bayesian Updating & Source Reliability

In Bayesian terms, a reliable source will provide more diagnostic evidence; as a result, evidence from that source will lead to higher posterior degrees of belief than evidence from an unreliable source (Figure 1 above). In other words, a less reliable source leads to more conservative belief revision. If participants treat experimental evidence as obtaining from a somewhat unreliable source, their belief updating *should* be somewhat conservative in relation to a normative standard based on the assumption that the source is reliable.

There are two ways in which source reliability might be factored into a Bayesian model of a given task. The first is to consider source reliability as an endogenous variable; that is, inherent characteristics of the evidence and characteristics of the source providing that evidence are (implicitly) combined into a single, overall likelihood ratio (as in e.g., Birnbaum & Mellers, 1983; Birnbaum & Stegner, 1979; Corner & Hahn, 2009). The second possibility is to model source reliability exogenously as an explicit variable (as in e.g., Bovens & Hartmann, 2003; Hahn, Harris & Corner, 2009; Hahn & Oaksford, 2007; Pearl, 1988; Schum, 1981). This latter case involves a cascaded inference in a hierarchical model. Figure 2 shows a simple hierarchical model in which to capture an evidence report from a partially reliable source. This model captures explicitly the fact that what is received is a *report* of some evidence through a partially reliable source, not the evidence directly. In other words, it naturally captures cases of testimony where evidence of an event is based on witness description, not on first hand experience.

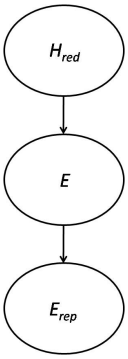


Figure 2: A hierarchical model in which the reliability of the reporting source is captured exogenously. Three levels are distinguished: the underlying hypothesis H, the evidence E, and the source's actual report of that evidence E_{rep}.

The likelihood ratio associated with such an evidence report, E_{rep}, is described by Eq. 2 (below):

$$\frac{P(E\mid H)[P(E_{rep}\mid E,H)-P(E_{rep}\mid \neg E,H)]+P(E_{rep}\mid \neg E,H)}{P(E\mid \neg H)[P(E_{rep}\mid E,\neg H)-P(E_{rep}\mid \neg E,\neg H)]+P(E_{rep}\mid \neg E,\neg H)}$$

Here, $P(E_{rep}|E,H)$ represents the probability of an evidence report, E_{rep}, to the effect that the evidence E obtains, given that both E and H (the hypothesis) are true, and so on (see also Schum, 1981). It can be seen that the evidential characteristics of the report vis à vis the hypothesis are a multiplicative combination of the diagnosticity of the evidence itself and the characteristics of the reporting source, that is, the source's own hit and false alarm rate regarding the true state of that evidence. If the witness is completely reliable and reports only the true state of the evidence, then Eq. 2 reduces simply to the standard relationship between evidence and hypothesis. Where the evidence is entirely deterministic and arises if and only if the hypothesis is true (i.e., $P(E|H)=1$, $P(E|\neg H)=1$), the hit and false positive rates of the witness completely determine the characteristics of the report. From this latter case, it can also be seen that partial reliability of the witness necessarily reduces the overall diagnosticity of the evidence received. How diagnostic the report can be, and hence what posterior degree of belief it can bring about is capped by the reliability of the witness (see also Hahn et al., 2009).

Simulating Bookbags and Pokerchips

How, then, can such a model be applied to the bookbag and pokerchip paradigm on which the vast majority of the evidence for conservatism is based?

We suggest that the conservative belief revision displayed in experimental settings may reflect rational responses to information from an information source that is less than perfectly reliable. Specifically, participants might not believe the asserted premise that the experimenter is drawing chips randomly from the bag. Such skepticism seems inherently sensible in light of the fact that draws in classic bookbag and poker chip tasks were frequently *not* random. Instead, the experimenter could determine the colour of the chip the to be drawn by a tactile cue. The 'random' laying of a hand on one poker chip, which is followed by a movement to another (experimenter desired) chip on the basis of a tactile cue could be construed as a mis-reporting of the nature of the initial, randomly chosen poker chip through the experimenter. Once the experiment is conceived of in this light, it is straightforward to model the effect of experimenter (un)reliability on belief revision, and we can show that such a model captures major effects demonstrated in the conservatism literature.

On this account, the participant is attempting to determine the truth of a hypothesis (e.g., that a bookbag contains predominantly red chips) on the basis of some evidence (the random drawing of a red or blue chip) that is reported by a source (the experimenter). The assumed characteristics of a single draw are represented by the model in Figure 2. H_{red} is the hypothesis in question, that is, whether the bag from which the chips are being drawn is a red bag. E represents the random drawing of a red chip; E_{rep} is the experimenter's report as to whether a red chip was randomly drawn - delivered in the form of the actual chip produced for the participant. This final piece of information is the only one at the participants' disposal in assessing the probability of H_{red}.

In these studies, prior degrees of belief are communicated to participants by explaining to them the number of bags of different composition and that this proportion should constitute their prior (see e.g., Phillips & Edwards, 1966). Consequently, under the assumption that the experimenter is a perfectly reliable source, who is merely exactly reporting the exact result of a random draw from the bag, participants posterior degree of belief should be determined completely by the diagnosticity of a given draw of red or blue. The diagnosticity of the chip drawn is fully determined by bag composition, that is, the relative proportion of red and blue chips within a bag. Because draws are independent, the overall diagnosticity of the evidence received across n trials thus far is a simple multiplicative function of the diagnosticity of a single draw.

To capture the fact that participants might (justifiedly) not consider the experimenter to be fully reliable, we likewise treat individual trials as independent, so that repeated draws correspond to repeated trials in the application of the model in Figure 2, which captures the believability of a single piece of testimony from one witness (Schum, 1981).

Arguably, this is not an *appropriate* model of what is going on in this task. All draws are coming from a single source and are ultimately neither random nor independent. However, the participant has no way of knowing what the purpose of the experiment is, and as a consequence, no way of knowing how the experimenter might be deviating from the model of independent random draws that the experimenter has explicitly set out. Consequently, the only model the participant arguably *can* establish if they are to engage in the task at all, is one of independent, random draws, in which experimenter distrust is captured simply through some additional, generic perturbation of those draws. This,

however, is readily captured through the repeated application of Eq. 2. Conceptually, the model of Figure 2 reflects, on the part of the participant, an inference to the chip that the experimenter would have drawn had he/she been drawing randomly from the bookbag. Once participants are assumed to treat the experimenter as a partially reliable source in this way, conservatism is unavoidable.

Unavoidable conservatism becomes apparent in the simulation of an idealized participant for a classic bookbag and pokerchip experiment. For these simulations, the prior probability of the bag containing predominantly red chips, $P(H_{red})$, was .5. In order to manipulate the diagnostic value of a single chip, the proportion of the predominant chips in any bag was either .6 or .7 (as in Phillips & Edwards, 1966, Experiment 1). To simulate belief revision on the basis of an imperfect information source, the sensitivity and specificity of the source, $P(E_{rep}|E)$ and $P(\neg E_{rep}|\neg E)$ were set to .6 (and thus the false positive rate $P(E_{rep}|\neg E)$ was .4). For the sake of simplicity, we only detail here the results of a simulation in which each of 10 draws from the bag (as reported by the experimenter) were red chips. The same general result, however, holds for sequences that also include some drawing of blue chips ($\neg E$). Belief revision occurs after each draw, with the prior probability of the hypothesis updated at each step.

Simulation of this model[1] produces not just basic conservatism, but also replicates the more specific findings of conservatism experiments. These are the findings that "conservatism increases as the diagnostic value of a single chip increases" and that "conservatism remains approximately constant as the diagnostic value of the sample increases" (Phillips & Edwards, 1966, p. 353). In other words, greater conservatism is observed for bags where the predominant color constitutes 70% of all chips than for those where it constitutes 60%. In order to facilitate comparison, we present results in terms of accuracy ratios as typical in conservatism research (as in Peterson & Miller, 1965; Peterson et al., 1964; Phillips & Edwards, 1966). The accuracy ratio is the ratio between participants inferred (and conservative) log likelihood ratio, and the 'true' log likelihood ratio corresponding to the task parameters as asserted by the experimenter. In our case, it is the ratio between the log likelihood ratio of the partially reliable and the fully reliable source. An accuracy ratio of less than 1 indicates conservatism (with smaller values indicating greater conservatism).

The results in Figure 3 clearly show that conservatism obtains regardless of bag composition, but that it is greater for the 70% bag, than for the 60% bag, in line with the experimental data of Phillips and Edwards. Finally, the accuracy ratios are constant across trials, in line with the experimental finding that conservatism remains approximately constant as the diagnostic value of the sample increases (Phillips & Edwards, 1966).

[1] Model simulations were created using the GeNIe modeling environment developed by the Decision Systems Laboratory of the University of Pittsburgh (http://dsl.sis.pitt.edu).

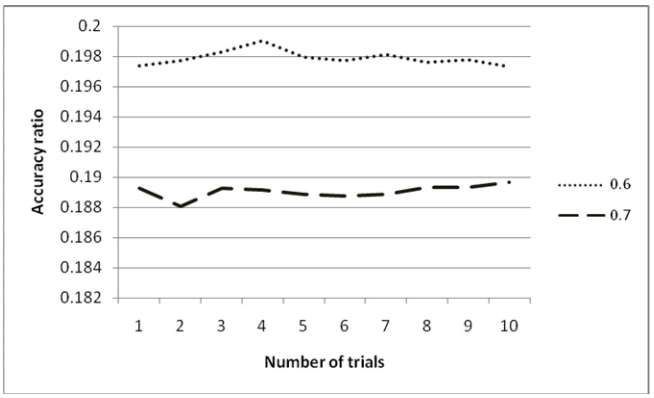


Figure 3: Accuracy ratios for a simulated participant who assumes that the experimenter is only partially reliable ($P(E|H) = .6$ and $P(\neg E|\neg H) = .6$). Different lines (.6 and .7) refer to bags of different composition (60% and 70% dominant chip color).

Finally, we note that there is nothing special about the specific values chosen here; these general relationships obtain across the range of meaningful values for source hit rate and false positive rate (i.e., wherever the hit rate exceeds the false positive rate).

General Discussion

In summary, the simple assumption that participants treat experimenters as partially reliable sources in classic conservatism studies generates, at least qualitatively, the main findings of such studies. It would be desirable in future work to not only model participant data exactly, but also to provide independent support for the source reliability account through experimental manipulation. For example, one might test whether conservatism vanishes if participants are allowed to make draws themselves, a methodological variant that has been found to reduce seeming base rate neglect (Gigerenzer, Hell & Blank, 1988).

In the meantime, these simulation results underscore why it cannot simply be assumed that participants take information presented to them by experimenters at face value. In the real world, most information sources are only partially reliable, and experimenters are no exception. Hence experimental demonstrations of conservatism do not necessarily indicate a gap between normative predictions and participants' responses – more conservative belief revision is the normatively appropriate response to less reliable evidence.

We are not suggesting that participants actively distrust or seek to undermine experimental materials. The tendency to treat experimental evidence as less than fully reliable is a mundane, default response to the experimental setting. Quite simply, participants know they are in an experiment, and do not necessarily (or automatically) assign as much weight to experimental evidence as they might in a non-laboratory situation. So, while participants in the classic 'bookbag and poker chip' experiments (Edwards, 1968) are unlikely to

have actively distrusted the experimenters, they are equally as unlikely to have treated the evidence as maximally reliable. Only when this possibility is either accurately modelled or empirically ruled out can the results of belief revision research fully be interpreted.

Acknowledgments

Adam Corner and Adam Harris were partly funded by ESRC postgraduate bursaries. Adam Harris was also funded by ESRC grants RES-000-22-3339 and RES-062-23-0952.

References

Birnbaum, M.H. & Stegner, S.E. (1979). Source credibility in social judgment: Bias, expertise and the judge's point of view. Journal of Personality and Social Psychology, 37, 48-74.

Birnbaum, M.H. & Mellers, B. (1983). Bayesian inference: Combining base rates with opinions of sources who vary in credibility. Journal of Personality and Social Psychology,45, 792-804.

Bovens, L. & Hartmann, S. (2003). *Bayesian Epistemology*. Oxford: Oxford University Press.

Christensen, L. (1977). The negative Subject: Myth, Reality, or a Prior Experimental Experience Effect? *Journal of Personality and Social Psychology. 35, 392–400.*

Cook, T.D. and Perrin, B.F. (1971). The Effects of Suspiciousness of Deception and the Perceived Legitimacy of Deception on Task Performance in an Attitude Change Experiment. *Journal of Personality. 39, 204–224.*

Corner, A. & Hahn, U. (2009). Evaluating Science Arguments: Evidence, Uncertainty & Argument Strength. *Journal of Experimental Psychology: Applied, 15, 199-212.*

Edwards, W. (1968). Conservatism in Human Information Processing. In B. Kleinmuntz (Ed.), *Formal Representation of Human Judgment (pp. 17-52)*. New York: Wiley.

Erev, I., Wallsten, T.S. & Budescu, D.V. (1994). Simultaneous over and under confidence: the role of error in judgement processes. *Psychological Review 101 (3) 519-527.*

Evans, J.St.B.T. & Over, D.E. (2004). *If*. Oxford: Oxford University Press.

Fischoff, B & Beyth-Marom, R. (1983). Hypothesis Evaluation from a Bayesian Perspective. *Psychological Review 90 (3) 239-260.*

Gigerenzer, G. & Todd, P.M. (1999). *Simple heuristics that make us smart*. Oxford: Oxford University Press.

Gigerenzer, G., Hell, W. & Blank, H. (1988). Presentation and content: The use of base rates as a continuous variable. *Journal of Experimental Psychology: Human Perception and Performance, 14*, 513-525.

Hahn, U., Harris, A.J.L., & Corner, A.J. (2009). Argument Content and Argument Source: An Exploration. *Informal Logic, 29*, 337-367.

Hahn, U. & Oaksford, M. (2007). The Rationality of Informal Argumentation: A Bayesian Approach to Reasoning Fallacies. *Psychological Review 114 (3)* 704-732.

Hilton, D.J. (1995). The Social Context of Reasoning: Conversational Inference and Rational Judgment. *Psychological Bulletin 118 (2)* 248-271.

Kelman, H.C. (1967). Human Use of Human Subjects: The Problem of Deception in Social Psychology. *Psychological Bulletin. 67, 1–11.*

Kemp, C. & Tenenbaum, J. B. (2009). Structured statistical models of inductive reasoning. *Psychological Review, 116(1), 20-58.*

Lopes, L. L. (1985). Averaging rules and adjustment processes in Bayesian inference. *Bulletin of the Psychonomic Society, 23*, 509-512.

Lopes, L. L. (1987). Procedural debiasing. *Acta Psychologica, 64*, 167-185.

McKenzie, C.R.M., Wixted, J.T. & Noelle, D.C. (2004). Explaining Purportedly Irrational Behavior by Modeling Skepticism in Task Parameters: An Example Examining Confidence in Forced-Choice Tasks. *Journal of Experimental Psychology: LMC 30 (5)* 947-959.

Noveck, I.A. & Sperber, D. (Eds). (2004). *Experimental Pragmatics*. New York: Palgrave Macmillan.

Oaksford, M. & Chater, N. (2003). Conditional Probability and the Cognitive Science of Conditional Reasoning. *Mind & Language 18 (4)*, 359–379.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Francisco: Morgan Kaufman

Peterson, C.R. & Miller, A.J. (1965). Sensitivity of subjective probability revision. *Journal of Experimental Psychology 70 (1) 117-121.*

Peterson, C.R., Schneider, R. & Miller, A.J. (1964). Sample size and the revision of subjective probabilities. *Journal of Experimental Psychology 69, 522-527.*

Phillips, L.D. & Edwards, W. (1966). Conservatism in a simple probability inference task. *Journal of Experimental Psychology, 72*, 346-354.

Schum, D.A. (1981). Sorting out the effects of witness sensitivity and response-criterion placement upon the inferential value of testimonial evidence. *Organizational Behavior and Human Performance,* 27, 153-196.

Schwarz, N. (1996). *Cognition & Communication: Judgemental Biases, Research Methods & The Logic of Conversation*. Hillsdale, NJ: Erlbaum.

Simon, H.A. (1982). *Models of Bounded Rationality*, Vols. 1, 2. Cambridge, MA: MIT Press.

Slovic, P. & Lichtenstein, S. (1971). Comparison of Bayesian and regression approaches to the study of information processing in judgement. *Organizational Behavior & Human Processes 6, 649-744.*

Tversky, A. & Kahneman, D. (1983). Extensional versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment. *Psychological Review 90 (4)* 293-215.

The Game Lies in the Eye of the Beholder: The Influence of Expertise on Watching Soccer

Michael Smuc (michael.smuc@donau-uni.ac.at)
Eva Mayr (eva.mayr@donau-uni.ac.at)
Florian Windhager (florian.windhager@donau-uni.ac.at)
Research Center KnowComm, Danube University Krems, Dr.-Karl-Dorrek-Str. 30
3500 Krems, Austria

Abstract

The influence of expertise on viewing soccer matches is already an area of extensive research focusing on training. However, free viewing of soccer matches did receive less attention. In an explorative eye-tracking study we compared the viewing behavior of novices, amateur players, and professional players watching soccer scenes freely. Overall, novices seem to view a soccer match quite similar to professional players, whereas amateurs engage in more visual work. The viewing behavior differs when watching soccer freely or with a task in mind – a result worth a second glance.

Keywords: Soccer, viewing behavior, eye tracking, expertise

Introduction

Watching soccer without the corresponding experience and domain knowledge is a real challenge. Without knowledge about standard situations and tactical behavior, an unskilled observer is restricted to following the ball's trajectory mainly. On the other extreme, a good commentator is able to take in the whole scene at once and comment on the events and possible next moves. But what is it that enables an experienced soccer viewer to direct his or her attention more strategically and to take in more relevant information in comparison to an inexperienced one?

To answer this question, we review existing research on eye-movements in sports and their relation to expertise. We present a study that compares the television viewing behavior of soccer laypersons, amateurs, and professional players.

Eye-Movements in Watching Television

In general, viewing television is a complex activity (Josephson & Holmes, 2006): A huge amount of information has to be processed at a speed, which cannot be controlled by the viewer. Kirkorian (2007) assumes that watching television is nearest to perceiving scenes (e.g., Henderson, 2007). Both convey complex visual stimuli, but instead of viewing only one scene, television includes a series of static frames.

To examine visual information processing, *eye tracking* technology provides a means to observe a viewer's point-of-gaze (e.g., Rayner, 1998). In the past, eye tracking focused mainly on scene perception and reading under laboratory conditions (Henderson, 2007; Rayner, 1998); only in the last years, applications in more everyday settings (e.g., Hayhoe & Ballard, 2005; Mayr, Knipfer, & Wessel, 2009) became possible with the emergence of more usable technology.

Central eye-movement measures are fixations and saccades. Saccades are shifts from one point of gaze to another; fixations indicate visual attention to that information (Rayner, 1998). In scene perception, top-down and bottom-up influences control where one looks (Henderson, 2007). Bottom-up influences are stimulus-driven, whereas top-down influences are viewer-driven.

Bottom-up influences are mainly based on the visual salience of the stimulus, i.e., color, saturation, and – which is especially important in television – movement (Mahapatra, Winkler, & Yen, 2008). Also, research on eye-movements during film watching shows that a high degree of the fixations is within the center of the screen (Goldstein, Woods, & Peli, 2007). An open question is whether this is due to a trend to fixate the center or due to movie making conventions placing the most relevant information in the center of the screen.

Top-down influences on the other hand are a viewer's knowledge about the stimulus, his or her domain knowledge, and his or her goals (Henderson, 2003). It was shown that expectations about camera angles, cuts and close-ups determine television viewing behavior (Kirkorian, 2007). These expectations are learned and, therefore, get stronger with viewing experience.

Another top-down influence is the viewer's domain knowledge. Chase and Simon (1973) showed that due to their higher knowledge on possible configurations experts in chess can easier create chunks of information. A similar mechanism can be assumed in soccer experts and was already shown to be influential (Ward & Williams, 2003).

A third top-down influence is the existence of specific goals. Only little research exists on humans watching television freely, i.e. without any task or instruction (see Goldstein et al., 2007, for an exception). However, Spanne (2006) showed that similar to viewing natural scenes (DeAngelus & Pelz, 2009) viewing behavior of movies differs according to the task at hand and in free viewing. But until now no research on free viewing behavior in soccer exists. Rather, most research asked players to anticipate the next move, recall the players' positions (e.g., Ward & Williams, 2003), or actively pass the ball (Helsen & Starkes, 1999).

As watching soccer for leisure purposes is a free viewing condition, it has to be questioned whether existing research

on the influence of expertise on watching soccer with a specific task holds under this condition as well.

Expertise in Soccer and Viewing Behavior

Research on eye-movements in sports focused on the sportsmen's performance and how it relates to perceptual processes mainly (see Memmert, 2009, for a review). The aim of such research was on the one hand to train the sportsmen's viewing behaviour and, thereby, to improve their performance; on the other hand, this research aimed at testing theories of expertise, perception, and attention under ecologically more valid conditions (Casanova, Oliveira, Williams, & Garganta, 2009).

In comparison to amateurs, professional soccer players can better use advance visual cues, they can better recall and recognize visual patterns, they engage in more effective search behaviour, and can better judge situational probabilities (Casanova et al., 2009). With respect to the viewing behaviour, experts have fewer fixations (Helsen & Starkes, 1999), but those last longer than the fixations of amateurs (Williams, 2000). It is assumed that during those longer fixations, experts take in information not only from central, but also from more peripheral areas (Casanova et al., 2009; Ghasemi, Momeni, Rezaee, & Gholami, 2009).

Informative visual cues in soccer are, next to the ball and the goal, the player's teammates and opponents, but also free spaces. Amateurs fixate on the more obvious informative areas only (players, ball), whereas professional players fixate on more sophisticated informative areas like possible free spaces as well (Casanova et al., 2009).

In dependence of the player's position, the number of players visible and the viewers´ tasks, different viewing patterns were observed (Poulter, Jackson, Wann, & Berry, 2005; Williams, Janelle, & Davids, 2003).

Pattern recognition is an important skill in watching games – especially in team sports, like soccer (Ward & Williams, 2003). Experts have a higher repertoire of patterns stored in their long-term memory and can more effectively retrieve appropriate patterns based on visual input (Casanova et al., 2009). Williams, Hodges, North, and Barton (2006) showed that the relation between players and the presence of key players are important features that facilitate pattern recall in soccer experts.

Research Questions

Based on the existing research on expertise in soccer, this study examines free viewing behavior while watching soccer without a concrete task. As prior studies compared only professional and amateur soccer players, we included a third less skilled group in our study: Novices, with little or no knowledge in soccer so far (like Poulter et al., 2005). In detail, we address the following research questions:

Do soccer laypersons, amateurs and professional players differ in their *soccer viewing behavior*? As reported in previous research (Casanova et al., 2009; Williams, 2000) we assume that professional players show less, but longer fixations than amateurs, and that they have better peripheral perception. No hypothesis for novices can be build upon the existing knowledge base.

Do professional soccer players pay more attention to *informative regions* than amateurs? Casanova and colleagues (2009) report that amateurs do focus on less informative regions like the ball and the players. We therefore assume that professional soccer players do fixate more informative regions than amateurs. As the informative content of some visual cues has to be acquired with soccer domain knowledge (e.g., free kick), we hypothesize that novices to soccer do fixate only the most obvious informative regions, i.e. the ball, the player in possession of the ball, and the goal.

Are experts better at *anticipating* the next pass? Ward and Williams (2003) found that professional soccer players are better in predicting the next pass in 11 to 11 simulations. We assume that this superior predictive performance also coincides with fixations on the according player and that amateurs and novices do have less fixations in this area prior to the pass.

Method

The study was conducted in November 2009. Professional soccer players' viewing behavior was recorded at the training camp of their Austrian first league soccer club Magna Wiener Neustadt. The viewing behavior of amateur soccer players and novices was recorded at the Austrian open research night at Danube University Krems.

Sample

The viewing behavior of 7 professional soccer players, 8 amateur soccer players, and 11 soccer novices was recorded. Three participants (1 amateur, 2 novices) with corneal irregularity and varifocals were excluded from further analyses, as there eye gaze data could not be recorded validly. An overall sample of 23 participants remained (see table 1).

The age distribution is similar in all three groups ($F_{2, 22} = 1.43$, $p > .05$). Though more female participants were soccer novices, this difference reached no significance ($\chi^2 = 5.35$, df = 2, $p > .05$).

Table 1: Descriptive statistics.

	professionals	amateurs	novices
N	7	7	9
age	30.9 (5.4)	39.9 (9.6)	33.4 (12.8)
male	100 %	86 %	55 %

Material

Some studies on soccer expertise used recordings from a single camera which takes in the whole soccer field instead of television reports (Vaeyens, Lenoir, Philippaerts, & Williams, 2007; Williams et al., 2006). As our study focuses on watching soccer on television, we used original soccer reports from different not well-known games. We chose

four scenes with an overall duration of 3'43 mins: Scene 1 consists of a cascade of successful passes. Scene 2 is a free kick sequence. Scene 3 deals with a questionable offside decision. Scene 4 shows a quick offense over the whole field.

Measures

Eye movements were recorded using an SMI iView X™ RED eye tracker at a temporal resolution of 60 Hz. It tracks the corneal reflection of the pupils and allows relatively free movement of the head when seated approximately 60 cm from the tracking device. As it allows eye tracking with glasses and contact lenses, a wide range of participants could be included.

Expertise was assessed with multiple questions: whether participants' had experience in actively playing soccer in a club or not, how often they watched soccer on television (never, seldom, several times a year, several times a month), and how they evaluate their own soccer knowledge in comparison to a famous soccer player on a rating scale.

Participants who actively played soccer in a club, watched soccer more frequently by trend ($t = -1.98$, $df = 16$, $p < .1$) and had higher knowledge ($t = 3.12$, $df = 16$, $p < .01$). Therefore, a differentiation based on experience in playing soccer seems to be a valid measure of expertise.

Procedure

Each participant was tested individually. After an explanation on the purpose of the study, the functionality of the eye tracking device was explained to the participants. The device was calibrated using a five-point-calibration. Then the participants were instructed to watch the scenes freely, as they would usually watch soccer.

Participants viewed the soccer scenes on the 17'' computer screen integrated in the eye tracking device. The experimenter was seated next to the participant with a control screen of the participant's gazes to intervene, if the gaze was lost by the eye tracking system (see figure 1).

After viewing the scenes, participants received some questions on demographic data and their soccer expertise.



Figure 1: Experiment setup at the training camp (left) and at the long research night (right).



Figure 2: Dynamic AOIs for two frames from the passing scene (top) and the free kick scene (bottom). Filled circles and ellipses denote the dynamic AOIs. The smaller bold rings represent the fixations of amateurs, novices (both red), and professional soccer players (blue, green).

Analyses

Eye tracking data were analyzed with BeGaze™ analysis software. We segmented the videos based on single scenes and extracted the fixations (number, duration) and saccades (number, amplitude).

To analyze the visual attention given to highly informative regions, the soccer scenes were coded in accordance to predefined Areas of Interest (AOIs) similar to Helsen and Starkes (1999), dependent on the scenes. In the following, two of four analyzed scenes are described in detail to exemplify the analysis procedure.

Scene 1 is shown from an overview perspective without close-ups. It consists of a cascade of successful passes for 33 seconds in the middle of the field. At least five players of the offending team and four players of the defending team can be seen. In this scene each player and the ball were coded as an AOI (see figure 2, top). To gain more information on peripheral perception of the ball's surrounding, we used 5 AOIs of different size with the ball at its center. The AOIs' radiuses grew approximately with the size of an average player. As a measure of anticipation, we coded the player who will be the recipient of the next pass as an AOI.

Scene 2 is a free kick sequence next to the goal (13 sec.), representing a typical standard situation in soccer. All players except for three of the offending team can be seen

(see figure 2, bottom). In this free kick situation, the viewers had approximately five seconds to follow the players' prearrangements before the free kick was carried out. The different roles were coded as individual AOIs; namely goalkeeper, referee, free kick shooter, the player next to the free kick shooter, and the bunch of players in front of the goal.

The ability to define dynamic AOIs allowed us to analyze eye movement parameters for each AOI automatically. The main measurements are the number of fixations per AOI, the glance time (where all saccades, including the entry and exit saccade, and all fixation times on an AOI are summed up), and the fixation duration in percent (fixation time in ms divided by the difference of end- and start-time).

Eye tracking data were exported and further analyzed in SPSS to examine differences between professional players, amateur players, and novices in soccer.

Results

Viewing Behavior

As a first indicator of viewing behavior we compared the fixations' number and average duration between the expertise groups across all scenes.

Overall and in contrast to our first assumption, the professional players did not make more fixations than the amateur players ($t = 0.65$, $df = 13$, $p > .05$). Still, as was assumed, their fixations endured longer ($t = -1.99$, $df = 13$, $p < .05$). Interestingly, novices' viewing behavior did not differ from professional players'. But they did make less fixation than the amateur players ($t = 2.37$, $df = 16$, $p < .05$).

Peripheral Perception

As stated earlier, the ball plays the decisive role in soccer and drives the game. Let us take a closer look at the ball and the space around it. As described in the analysis section, we analyzed areas of five different sizes around the ball. The AOI ball 1 equates an AOI with a radius of approximately the size of one player; ball 2 has the radius of two players and so forth.

While the smallest AOI around the ball showed no differences, amateurs differed from the professional players and novices for the bigger ones (see table 2). With increasing size of the AOI ball, professionals fixated this AOI less and had lower glance durations in this AOI. This is an indicator that they perceived the region surrounding the ball already with their central fixation in the nearer ball area; whereas amateurs had to fixate the outer areas as well to take in this information.

As for the overall viewing behavior, we found no remarkable differences between novices and professionals.

Seeking for Relevant Information

For the passing scene (scene 1), the AOIs of the offending and defending players were analyzed for a period of 15 seconds. This analysis offers some details about viewers' visual search patterns for relevant information.

Amateurs more often and longer fixated one player who had a rather longer ball possession time (*amateurs vs. professionals*: glance duration: $t = -2.86$, $df = 13$, $p < .05$; fixation count: $t = -2.98$, $df = 13$, $p < .01$; *novices vs. amateurs*: glance duration $t = -4.28$, $df = 16$, $p < .01$; fixation count. $t = -3.61$, $df = 16$, $p < .01$). One player who was an attractive alternative to pass to was fixated earlier by professionals than by novice viewers ($t = 4.58$, $df = 9$, $p < .01$). This attractive pass alternative had a defensive counterpart who covered him a bit later in the sequence. This defensive player turned out to be an interesting fixation object for professionals in contrast to amateurs (glance duration: $t = -2.63$, $df = 13$, $p < .05$; fixation count: $t = -2.32$, $df = 13$, $p < .05$) For professionals vs. novices a trend exists in the same direction.

As an indicator of anticipation, we also analyzed fixations to the player receiving the next pass prior to ball contact. No difference existed between participants of different expertise in the number of fixations and in their glance duration.

Knowledge-Driven Viewing

Scenes with an inactive ball provide more time for top-down, knowledge-driven processing of the scene. The beginning of scene 2, before the ball was shot, was therefore very interesting to analyze.

Table 2: Eye tracking performance matrix for AOIs of different sizes around the ball. The label of the group with higher values is plotted in case there is a trend. Asterisks denote significant differences.

higher for …		novices or amateurs	novices or profess.	amateurs or profess.
ball 1	glance dur.	-	-	-
	fix. count	-	-	-
	fix. time %	-	-	-
ball 2	glance dur.	amateurs	-	-
	fix. count	amateurs	-	-
	fix. time %	amateurs	-	-
ball 3	glance dur.	amateurs	-	amateurs
	fix. count	amateurs	-	amateurs*
	fix. time %	-	-	-
ball 4	glance dur.	amateurs*	-	amateurs
	fix. count	amateurs*	-	amateurs
	fix. time %	amateurs*	-	-
ball 5	glance dur.	amateurs*	-	amateurs
	fix. count	amateurs*	-	-
	fix. time %	amateurs	-	-

In the free kick sequence (scene 2), professional players fixated free regions (that is, regions without players) longer than amateurs ($t = -2.41$, $df = 13$, $p < .05$). Further, professionals more often fixated the player next to the free kick shooter ($t = -2.62$, $df = 13$, $p < .05$). Both, amateurs' glances (duration: $t = -2.31$, $df = 16$, $p < .05$) and professionals' (duration: $t = -2.21$, $df = 13$, $p < .05$), stayed longer in the area where most players were and where the ball will most likely be played to – in comparison to novices (see figure 2, bunch at the bottom). No significant differences were found between professionals, amateurs and novices in their viewing behavior on the goalkeeper, the free kick wall, the referee, and actions of single offensive or defensive players.

Discussion

Watching a soccer match freely is an everyday activity that is not connected to any task. Prior research on viewing behaviour during watching soccer was insofar restricted as participants were asked to answer questions, anticipate behaviour, or recall information (e.g., Ward & Williams, 2003). Human eye-movements in a free viewing condition of moving visual stimuli were recorded only seldom until now (Mayr et al., 2009; Spanne, 2006) and never for watching a soccer match. This study is the first to analyse the influence of expertise on viewing behaviour.

Still, some of the results which were gained under task conditions hold under free viewing as well: We observed longer fixations in professional than in amateur players like found in many other studies (Casanova et al., 2009, Williams, 2000). In addition, they exhibited higher peripheral perception skills like reported in prior research (Casanova et al., 2009; Ghasemi et al., 2009): Professional players fixate less often the wider area surrounding the ball compared to amateurs. Fewer fixations do not mean that professionals perceive less parts of the game but rather that they perceive more relevant visual cues with fewer fixations.

We assumed that professional soccer players pay more attention to informative regions. This top-down controlled viewing behaviour should increase with higher domain knowledge. Indeed, we ascertained that professional players fixated some informative regions (certain key players, free regions) which amateurs fixated only to a lesser extent. Other areas, like the bunch of players during the free-kick, were perceived by professional and amateur players to a similar extent. Due to our comparison with novices we could show that this perception is knowledge-driven as well.

In contrast to prior research, we observed some profound differences as well: Our assumption that professional players would anticipate the next pass visually is not supported by our participants' viewing behaviour. It remains to be studied whether this difference is due to the absence of an according task or due to the short duration of the analysed scene. Further analyses of longer sequences would be necessary to validate this finding.

In contrast to prior research (Casanova et al., 2009; Williams, 2000) we found no differences between the number of fixations by amateurs and professionals. As this visual indicator depends on the number of players displayed in a scene (see Vaeyens et al., 2007), a more differentiated analysis might reveal differences according to the proportion of the field displayed.

Prior research on *soccer expertise* compared only professionals with high and low performance (e.g., Vaeyens et al., 2007), professionals with amateurs (e.g., Ward & Williams, 2003), or people with high vs. low self-reported soccer knowledge (Dijksterhuis, Bos, van der Leij, & van Baarne, 2009). To our knowledge barely any research extended these boundaries of expertise so far to include also professional soccer observers (like referees, see Ghasemi et al., 2009, for an exception) or novices without any soccer knowledge (see Poulter et al., 2005, for an exception). Though the first gap remains to be filled, this study was able to shed some light on the viewing behaviour of novices:

In contrast to our assumption that novices would mainly focus on the ball (as the most obvious, and highly salient informative region) they watched similarly to professional soccer players. A possible explanation for this similarity could be that though they looked on the same region, they extracted different information.

Novices as well as professionals focused on the ball less time than the amateur players. This result raises the question, why amateurs do view a soccer match differently from professional players and novices? Maybe the amateurs were very motivated to compare their own gazes to those of professionals in comparison to the more carefree novices. They seemed to seek for as much information as possible, especially in regions of 8-10 meters around the ball. They also had a higher fixation dispersion than professionals and novices ($F_{2,21} = 3.34$, $p < .1$). Another explanation could be that the situation was not as goal-free as intended, because different learned viewing behaviors were activated: Professional soccer players frequently watch soccer matches to analyze their behavior for training purposes. Amateur soccer players in contrast watch the game not only to "read" it, but mainly to reach the soccer fan's "fever pitch".

Limitations

This study is limited by the artificial experimental setting of watching a match in front of a computer screen instead of a wide-screen television (cp. Josephson & Holmes, 2006). Even though nowadays soccer matches are often watched on youtube, a typical match-viewing situation is characterized by a stimulating, emotion-rich environment.

A second limitation of our results is the methodology used: Eye tracking methodology can only show the gaze focus, but not the focus of attention (e.g., Treisman, 2006). A triangulation with other methodologies would be necessary, but would restrict the free-viewing paradigm.

Further Research Questions

One of the main novelties in this study is the free viewing paradigm applied to watching soccer. It would be interesting – also in the sense of the limitations – to analyze free viewing in different environments (i.e., stadium, private TV, public viewing areas).

Qualitative analysis of the professional players' viewing behavior indicated differences between playing positions: Whereas goal keepers observed the behavior of the goal keepers to a higher extent, trainers scanned the soccer field more frequently. A more differentiated analysis of experts is needed (see also Casanova et al., 2009).

Further research should also compare passive sport experts, i.e. real viewing experts (e.g., referees – see Ghasemi et al., 2009) vs. couch potatoes, and active sport experts, i.e. professional vs. amateur players, in their viewing behavior.

Soccer is a male-biased sport – and so is research on it. With one exception (Poulter et al., 2005), no women were included in prior studies on soccer expertise. We would therefore like to encourage further research on female soccer players and their passive counterparts.

Acknowledgments

We are grateful for the willingness of the SC Magna Wiener Neustadt's players to participate in this study. Further thanks are given to the visitors at the "Lange Nacht der Forschung 09" (the Austrian open research night) for their interest and to ecoplus for supporting our travel expanses.

References

Casanova, F., Oliveira, J., Williams, M., & Garganta, J. (2009). Expertise and perceptual-cognitive performance in soccer: A review. *Revista Portuguesa de Ciências do Desporto*, *9*, 115-122.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology*, *4*, 55-81.

DeAngelus, M., & Pelz, J. (2009). Top-down control of eye movements: Yarbus revisited. *Vis Cognition*, *17*, 790-811.

Dijksterhuis, A., Bos, M. W., van der Leij, A., & van Baarne, R. B. (2009). Predicting soccer matches after unconscious and conscious thought as a function of expertise. *Psychological Science, 20*, 1381-1387.

Ghasemi, A., Momeni, M., Rezaee, M., & Gholami, A. (2009). The difference in visual skills between expert versus novice soccer referees. *Journal of Human Kinetics*, *22*, 15-20.

Goldstein, R. B., Woods, R. L., & Peli, E. (2007). Where people look when watching movies: Do all viewers look at the same place? *Computers in Biology & Medicine*, *37*, 957-964.

Hayhoe, M., & Ballard, D. (2005). Eye movements in natural behavior. *Trends in Cognitive Science*, *9*, 188-194.

Helsen, W. F., & Starkes, J. L. (1999). A multidimensional approach to skilled perception and performance in sport. *Applied Cognitive Psychology*, *13*, 1-27.

Henderson, J. M. (2007). Regarding scenes. *Current Directions in Psychological Science*, *16*, 219-222.

Josephson, S., & Holmes, M. E. (2006). Clutter or content? How on-screen enhancements affect how TV viewers scan and what they learn. In *Proceedings of the 2006 symposium on Eye tracking research &; applications* (pp. 155-162). San Diego, CA: ACM.

Kirkorian, H. L. (2007). *Age differences in eye movements during video viewing*. Dissertation, University of Massachusetts Amherst.

Mahapatra, D., Winkler, S., & Yen, S. C. (2008). Motion saliency outweighs other low-level features while watching videos. In *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, *6806*, 68060P1-68060P10.

Mayr, E., Knipfer, K., & Wessel, D. (2009). In-sights into mobile learning. An exploration of mobile eye tracking methodology for learning in museums. In G. Vavoula, N. Pachter, & A. Kukulska-Hulme (Eds.), *Researching mobile learning: Frameworks, methods, and research designs* (pp. 189-204). Oxford, UK: Peter Lang.

Memmert, D. (2009). Pay attention! A review of visual attentional expertise in sport. *International Review of Sport and Exercise Psychology*, *2*, 119-138.

Poulter, D. R., Jackson, R. C., Wann, J. P., & Berry, D. C. (2005). The effect of learning condition on perceptual anticipation, awareness, and visual search. *Human Movement Science*, *24*, 345-361.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*, 372-422.

Spanne, J. G. (2006). *Task impact on cognitive processing of narrative fiction film*. Master thesis, Lund University.

Treisman, A. (2006). How the deployment of attention determines what we see. *Visual Cognition*, 14, 411-443.

Vaeyens, R., Lenoir, M., Philippaerts, R. M., & Williams, A. M. (2007). Mechanisms underpinning successful decision making in skilled youth soccer players: An analysis of visual search behaviors. *Journal of Motor Behavior, 39*, 395-408.

Ward, P., & Williams, A. M. (2003). Perceptual and cognitive skill development in soccer: The multidimensional nature of expert performance. *Journal of Sport & Exercise Psychology*, *25*, 93-111.

Williams, A. M. (2000). Perceptual skill in team games: Research, theory, and practice. In *INSEP 2000* (p. 15-16).

Williams, A. M. (2002). Eye movement measurement systems and key visual search parameters. *Expertise in Elite Sports*, 43.

Williams, A. M., Hodges, N. J., North, J. S., & Barton, G. (2006). Perceiving patterns of play in dynamic sport tasks: Investigating the essential information underlying skilled performance. *Perception, 35*, 317-332.

Williams, A. M., Janelle, C. M., & Davids, K. (2003). Constraints on the search for visual information in sports. *Journal of Sport & Exercise Psychology, 2*, 301-318.

When Robot Gaze Helps Human Listeners: Attentional versus Intentional Account

Maria Staudte & Matthew W. Crocker
Department of Computational Linguistics
Saarland University
Saarbrücken, Germany
{masta, crocker}@coli.uni-saarland.de

Abstract

Previous research has shown that listeners exploit speaker gaze to objects in a shared scene to ground referring expressions, not only during human-human interaction, but also in human-robot interaction. This paper examines whether the benefits of such referential gaze cues are best explained by an attentional account, where gaze simply serves to direct the listeners visual attention to an object immediately prior to mention, or an intentional account, where speaker gaze is rather interpreted as revealing the referential intentions of the speaker. Two eye-tracking studies within a human-robot interaction setting are presented which suggest that close temporal synchronization of speaker gaze and utterance is not necessary to facilitate comprehension, while the order of gaze cues with respect to order of mentioned references is. We interpret this as evidence in favor of an intentional account.

Keywords: human-robot interaction; gaze; visual attention; referring expressions

Introduction

Gaze has been widely studied as an indicator for overt visual attention during language processing. Previous studies revealed that speakers look at entities shortly before mentioning them (Griffin & Bock, 2000; Meyer, Sleiderink, & Levelt, 1998), while listeners rapidly inspect objects as they are mentioned (Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). This shows that gaze during situated language production and comprehension is tightly coupled with the unfolding speech stream, reflecting both speakers' intentions and listeners' understanding on-line. In face-to-face communication, the speaker's gaze to objects in a shared scene provides the listener with a visual cue to the speaker's focus of (visual) attention (Emery, 2000; Flom, Lee, & Muir, 2007). By revealing a speaker's focus of visual attention, such gaze cues potentially offer the listener valuable information to ground and disambiguate referring expressions, to hypothesize about the speaker's communicative intentions and goals and, thus, to facilitate comprehension (Hanna & Brennan, 2007).

In human-robot interaction, robot gaze that was synchronized with speech in a human-like manner has been shown to be similarly useful for grounding and resolving spoken references (Staudte & Crocker, 2009b). Further evidence supported the hypothesis that the utility of robot gaze originates from people's inferences of referential intentions from gaze (Staudte & Crocker, 2009a). However, it remained an open question whether such human-like synchronization of gaze and spoken references is necessary for gaze to be beneficial.

Firstly, we hypothesize that people indeed infer referential intentions from robot gaze cues. And secondly, we hypothesize that this assignment of *intentional* states makes the utility of gaze relatively flexible with respect to temporal synchronization. That is, despite a substantial shift of gaze cues with respect to corresponding speech cues, the conveyed intentions of the speaker may still facilitate utterance comprehension. If, in contrast, gaze is only a purely visual cue that happens to directs listeners' *visual attention* to an object which is then mentioned, we hypothesize that close temporal synchronization would be necessary for any benefit of gaze. We present evidence from two experiments supporting an intentional account of processing and interpreting robot gaze.

Does robot gaze reflect referential intentions?

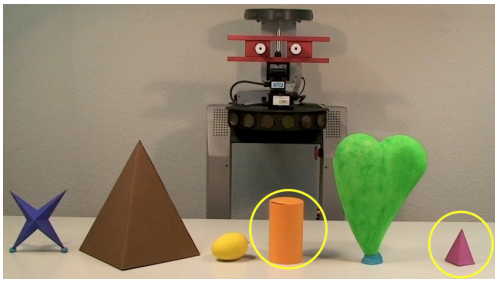
In previous experiments, Staudte and Crocker (2009b) showed that people follow and use robot gaze, similar to human gaze, and faster resolve referring expressions in the robot's utterance only when the gaze cue identified the actually mentioned object. Two different explanations of this result are conceivable. The influence of robot gaze could possibly be explained in terms of a purely "bottom-up" process: Robot gaze draws attention to one object (cf. Langton et al., 2000, for reflexive orienting in response to gaze cues) and the utterance subsequently draws attention to the same (congruent) or another object (incongruent). Thus, incongruent gaze elicits an *additional* shift of visual attention before utterance comprehension is completed. This additional shift could simply add to the total time needed to comprehend and respond, thus, accounting for an increase in response times (we will refer to this as the *Visual Account*). However, the effect of robot gaze could also be explained in terms of a (mis)match in expectations (elicited by robot gaze) and the actual utterance. Previous studies on the interpretation of human gaze have revealed that gaze is an extremely versatile cue which reflects attentional states as well as mental states such as goals, desires and intentions (Baron-Cohen, 1995). We therefore hypothesized that people's use of robot gaze may also be driven "top-down", by the belief that robot gaze also reflects attentional and intentional states and, thus, reveals what the robot intends to mention (*Intentional Account*). That is, participants may have thought that the robot attended to one object because it intended to mention it and, therefore, an incongruent reference would have led to a revision in referential expectations which slowed people. In a follow-up experiment, participants were asked to correct false robot utterances and were free to decide which objects they mentioned in their cor-

rection sentence (Staudte & Crocker, 2009a). Participants' responses suggested that their understanding about what the robot had originally intended to say was indeed affected by its gaze.

Another way to potentially distinguish between Intentional and Visual Account is to consider the relevance of temporal synchronization of gaze and speech cues. Recent findings from a study on the influence of *indirect*, human speaker gaze on utterance comprehension suggest that there is only limited flexibility in the requirement of synchronization of such a gaze cue with speech, while maintaining its utility for the listener (Kreysa, 2009). Kreysa (2009) found that gaze cues with a small shift with respect to their natural temporal co-occurrence still facilitated task completion whereas a greater shift (by more than 2 sec) was no more beneficial than random cues. Interestingly, the importance of synchronization between gaze and speech may illuminate the nature of gaze influence. On the Visual Account, the influence of gaze is attributed to the induced attention shift towards the right object at the right time such that changes in the temporal synchronization of gaze and speech should clearly affect the utility of gaze. Under the Intentional Account, in contrast, people would interpret robot gaze with respect to the robot's intentional states such that synchronization would not be critical. Understanding someone's (referential) intentions should persist and influence utterance comprehension as long as they seem relevant.

While Kreysa's results (2009) suggest that the effect of human gaze cues on utterance comprehension is flexible to some extent, gaze cues used in her studies were indirect and not necessarily qualitatively equal with the direct perception of speaker gaze. Depending on how people perceive speaker gaze compared to Kreysa's cursor, two different behaviors in response to substantially shifted robot gaze is possible: Robot gaze may be similar to a gaze cursor, a *visual cue* that may (reflexively) direct attention and, thus, is only helpful for processing referring expressions when it occurs within a short time window around the spoken reference. A substantial shift of gaze relative to speech would result in longer response times than the original synchronization. Alternatively, speakers' looks towards an object may be perceived as more *intentional* than a gaze cursor and as more robustly assigning relevance to the object in focus (similar to human gaze). Participants may persistently maintain and use this information when it seems relevant, leading to equal response time for shifted and synchronized gaze. Equally, non-congruent gaze cues may thus – even when shifted to precede the utterance – disrupt comprehension and cause slower response times.

We present results from two experiments which suggest that the utility of gaze is *not only* a matter of attention cueing to the right object at the right time. Rather people seem to interpret robot gaze as an indicator to the robot's referential intentions, leading to a persistent influence of gaze.



Original
Sync: <c>"The cylinder is taller than the <p> pink pyramid."
Prec: <c><p>"The cylinder is taller than the pink pyramid."

Reverse
Sync: <c>"The pink pyramid is shorter than <p> the cylinder."
Prec: <c><p>"The pink pyramid is shorter than the cylinder."

Figure 1: Sample scene from experiments, with original/reversed sentences and synchronized/preceding robot gaze (first at cylinder (<c>), then at pyramid (<p>)).

Experiment 1

In this study, we investigated whether referential robot gaze needs to be temporally synchronized with speech (in the way human gaze is synchronized) in order to be beneficial, or whether robot gaze conveys referential intentions that have a more persistent effect on utterance comprehension. Thus, we manipulated synchronization in two ways. While robot gaze was always directed to the mentioned objects, we manipulated the factor Order of Mention (sequence of mentioned objects crossed with sequence of 'gazed at' objects) which led to original (coherent) or reverse order of references (see Figure 1). The second factor, Synchronization, manipulated the temporal delay between gaze/visual references and corresponding linguistic references.

Method

Participants Thirty-two native speakers of German, mainly students enrolled at Saarland University, took part in this study (26 females). All participants reported normal or corrected-to-normal vision.

Materials We created 1920x1080 resolution video-clips showing a PeopleBot robot (kindly provided by the CogX-project, http://cogx.eu) onto which a pan-tilt unit was mounted. This pan-tilt unit carried a stereo camera which appeared as the head and eyes of the robot. The video-clips each showed a sequence of camera-movements consecutively towards the central object and then the peripherally located object. The utterance was a synthesized German sentence using the Mary TTS system (Schroeder & Trouvain, 2001).

In these videos we manipulated two factors: Order of Mention (original, reverse) and Synchronization (synchronized, preceding), so each item appeared in four conditions. The temporal delay between gaze and speech was roughly 5.3 seconds in the preceding condition and 1 second in the syn-

chronized condition. A sample scene is given in Figure 1 as well as examples for each type of sentence order and robot gaze synchronization. In condition original-synchronized, gaze and speech cues were coherent and synchronized in a human-like manner while the condition reverse-synchronized showed cues that were concurrent but reverse to each other.

Eight lists of stimuli were created, accounting for four experimental conditions and their counter-balanced versions. In addition to 24 items, 36 fillers were shown such that participants saw a total of 60 trials. The order of item trials was randomized for each participant individually and items were always separated by at least one filler.

Procedure An EyeLink II head-mounted eye-tracker monitored participants' eye movements on a 24-inch monitor. Before the experiment, participants received written instructions about the experiment procedure and task: They were asked to attend to the presented videos and judge whether or not the robot's statement in each was valid with respect to the scene. In order to provide a cover story for this task, participants were further told that the results were used as feedback in a machine learning procedure for the robot.

Analysis Videos were segmented into Interest Areas (IAs). That is, each video contained a region labeled "NP2 referent" which marked the object mentioned last in the robot utterance (i.e., before sentence validation was possible). Further, we recorded and analyzed participant fixations on this area. The speech stream was segmented into two Interest Periods (IPs). IP1 was defined as the 1000ms period ending at the onset of the second noun (in NP2). Importantly, it contained the robot's gaze towards the target object as well as verbal content preceding the target noun. IP2 was defined as the 700ms period beginning with noun onset in NP2. These IPs roughly segmented the sentences as follows: *"The cylinder is taller [than the pink]$_{IP1}$ [pyramid]$_{IP2}$"*. Defining IP1 and IP2 in this way made it possible to distinguish once again between gaze-mediated inspections in IP1 and utterance-mediated inspections in IP2. Trials that contained at least one *beginning* inspection towards an IA within an IP (coded as "1") were contrasted with trials that did not contain an inspection in the same slot ("0"). As a result, mean values represent inspection probabilities for a given IA and IP. For inferential analyses, we considered inspections on the NP2 referent as well as response time, recorded from NP2-noun onset to the moment of the button press. The analyses were carried out using mixed-effect models from the lme4 package in R and Chi-Square tests to asses the contribution of a predictor through model reduction (Baayen, Davidson, & Bates, 2008; Bates, 2005).

Results

Eye movements Mean inspection probabilities for the NP2 referent are depicted in Figure 2. Note, that the manipulation of Order of Mention coincided with a difference in location of the NP2 referent. That is, in original order, the NP2 referent is located in the periphery of the table (pink pyramid), while in reverse order it is located in the center of the scene (cylinder), as depicted in Figure 1. Results from inferential statistics on inspection data are reported in-text where necessary and are otherwise omitted due to space limitations.

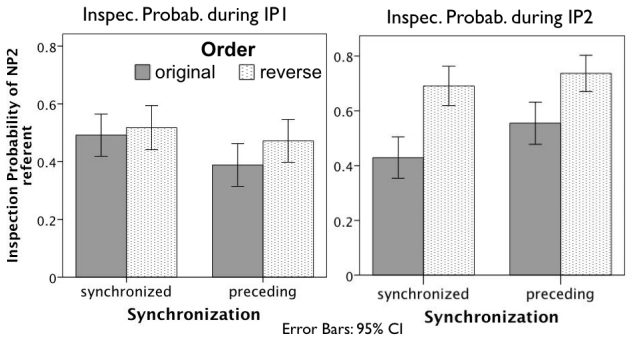


Figure 2: Inspection probability of NP2 referent in Exp1, for all conditions in IP1 (left graph) and IP2 (right graph).

In IP1, model reduction revealed a main effect of Synchronization on inspections on the NP2 referent ($\chi^2(1) = 4.03, p < .05$). People inspected the NP2 referent with lower probability when robot gaze preceded the utterance. Moreover, we did not observe a main effect for Order of Mention, i.e., people inspected the NP2 referent equally often irrespective of where this referent was located (centrally, peripherally) or whether the robot concurrently fixated this object. Interestingly, people were equally likely to inspect the NP2 referent in condition original-synchronized (when robot gaze identifies the actual NP2 referent) as in condition reverse-synchronized (when gaze does not). This result indicates that people may use the already mentioned reference (NP1) and the available visual cues to, at least *visually*, anticipate the NP2 referent even when cues were reversed.

In IP2, we observed a somewhat different inspection pattern. Order of Mention had a main effect on inspection probability ($\chi^2(1) = 35.67, p < .001$) such that participants inspected the (mentioned) NP2 referent significantly more often in the reverse condition than in the original (coherent) order condition. That is, when the robot fixated the peripheral object during IP1 and then mentioned the other, central object in IP2 (<central cylinder>"The pink pyramid is taller than <periph. pyramid> the cylinder.") participants were more likely to inspect the object mentioned in NP2 (centrally located cylinder) than in original order.

There are two possible explanations for these high probabilities of inspecting the NP2 referent in reverse order: Either participants inspected this central object more often because it was more salient due to its central location, predicting easy and quick reference resolution. Alternatively, the increased inspections on the NP2 referent in reverse condition reflect difficulty to resolve the reference as it includes conflicting information (gaze identified the pyramid while the mentioned noun referred to the cylinder). Thus, the response time results should reveal which of the two explanations is more likely.

Response Time Model reduction showed that Synchronization had no effect on response times. That is, participants were equally fast to determine the validity of the robot statement in synchronized and preceding conditions. Since no interaction between the two factors Synchronization and Order of Mention was observed, we excluded Synchronization as a predictor from our linear mixed-effects model. Model reduction further revealed a main effect of Order of Mention ($\chi^2(1) = 45.19, p < .001$, see also Figure 3 for averages).

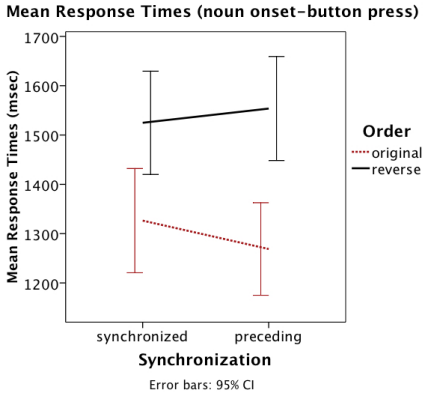


Figure 3: Avg. response times in all four conditions (Exp1).

The finding that Synchronization did not affect response time while Order of Mention did, cannot be explained by the Visual Account since both manipulations made robot gaze direct people's visual attention to relevant objects at non-synchronized points in time – and always prior to the last referring expression (NP2). Instead, the Intentional Account seems to provide more appropriate explanations for these results: The precise temporal synchronization is not crucial for people to interpret and use robot gaze as a cue to the robot's intentions. The inferred (referential) intentions, however, are expected to be executed in the *same order* as they were indicated by robot gaze. Thus, reversed order, even in the case of preceding gaze, slows people in utterance comprehension.

Discussion

By manipulating the order of references in the sentence, the location of the NP2 referent was effectively also manipulated. Since the center of the scene is the most salient area, this may have affected the effort needed to resolve a referring expression which identified the central object compared to one that identified an object in the periphery of the scene. Since in the reverse order condition, NP2 identifies the central object, this appears to benefit reference resolution given that the central object is most salient. However, response time results revealed that people were in fact *slower* in reverse order to judge sentence validity, compared to original order. This effect of Order of Mention suggests that reverse order was indeed more difficult to process than originally ordered cues, supporting the interpretation that people's increased inspections on the NP2 referent reflected *increased effort* to resolve the reference (due to conflicting information).

Experiment 2

In this experiment we further investigated whether the order of referring expressions in the robot utterance (accompanied only by neutral gaze) affects how fast people resolve these expressions and validate the utterance. Original and reverse sentence order were, thus, paired with neutral robot gaze and compared. This baseline condition showing neutral gaze allowed us not only to determine any effects of sentence order itself, but also to assess the actual benefit of original order versus a potentially disruptive effect of reversed gaze and speech cues. Since the temporal shift between synchronized and preceding condition did not affect people's responses, we did not include a preceding gaze condition again. We manipulated Order of Mention (*original*, *reverse*) and Synchronization (*synchronized*, *neutral*).

Synchronized robot gaze was again always directed first to the central object and then to the peripheral object. Using the sample sentence from the previous experiment "The orange cylinder is taller than the pink pyramid", the robot would first look at the cylinder and then to the peripherally located pyramid. The neutral gaze condition showed an initial glance down at the scene before the robot looked straight ahead and began to speak. We included an additional adjective for the central object (the "*orange* cylinder") in order to make sentences completely symmetric in both sentence orders. This symmetry also allowed us to change the onset for response time recordings from NP2-noun onset to NP2-adjective onset. Since the adjective already uniquely identifies the referent this most appropriately captures actual response time. Otherwise sentences and scenes were similar to the material used in Experiment 1.

Method

Participants & Procedure Thirty-two native speakers of German and mostly students at Saarland University took part in this study (21 females). All reported normal or corrected-to-normal vision. Task and Procedure were identical to Experiment 1.

Materials The manipulation of Order of Mention (original, reverse) and Synchronization (synchronized, neutral) resulted in four conditions. A set of 20 items was used as well as a set of 32 fillers which were evenly distributed across conditions. Participants therefore saw a total of 52 trials.

Analysis IAs used in this experiment were identical to those in Experiment 1. IP1 was again defined to begin 1,000ms prior to noun onset (in NP2). However, in this experiment IP1 did not stretch to noun onset but already ended with *adjective onset*. Thus, IP1 had no fixed duration but an average length of 600ms. This shortening of IP1 was done to incorporate the fact that the prenominal adjective already uniquely identified the referent. Consequently, IP2 was defined to stretch from adjective onset to 700ms after noun onset and had a mean duration of 1,100ms. Thus, sentences were segmented as follows: *"The cylinder is taller [than the]$_{IP1}$*

[pink pyramid]$_{IP2}$". Defining IP1 and IP2 in this way made it possible to distinguish once again between gaze-mediated inspections in IP1 (before the linguistic reference in NP2) and utterance-mediated inspections in IP2 (taking into account that the color adjective linguistically identifies the NP2 referent). Moreover, response time was defined to begin with NP2-*adjective onset* instead of the previously used noun onset for the same reason, that is, accounting for the nominal adjective as already identifying the final referent.

Results

Eye movements Mean probabilities for inspecting the NP2 referent are given in Figure 4. In IP1, both Order of Mention and Synchronization had main effects on inspection behavior (Synchronization: $\chi^2(1) = 5.83, p < .05$ and Order of Mention: $\chi^2(1) = 24.90, p < .001$). Participants generally inspected the NP2 referent more frequently when gaze was synchronized than when it was neutral. Moreover, model reduction revealed a significant interaction of the two predictors Order of Mention and Synchronization ($\chi^2(1) = 14.08, p < .001$). That is, the effect of Order of Mention varied depending on the Synchronization: Firstly, the neutral gaze condition reveals that Order of Mention by itself affected people's visual attention. In the reverse-neutral condition the NP2 referent was inspected significantly more often than in original-neutral. We argue that this effect is due to the NP2 referent being central and being additionally highlighted as the robot initially looked downwards. Secondly, the graph also reveals that the peripherally located object (NP2 referent in original order) was inspected more often when gaze was synchronized (original-synchronized) than when it was neutral (original-neutral), suggesting that a gaze cue in original (coherent) order helped people to visually anticipate the NP2 referent. In contrast, gaze cues in reverse order did not affect the inspections on the NP2 referent (central object) compared to reverse-neutral. Instead, the NP2 referent was rather frequently inspected in reverse order even when robot gaze was neutral throughout the utterance. This indicates that the central object was indeed more salient than the peripheral object.

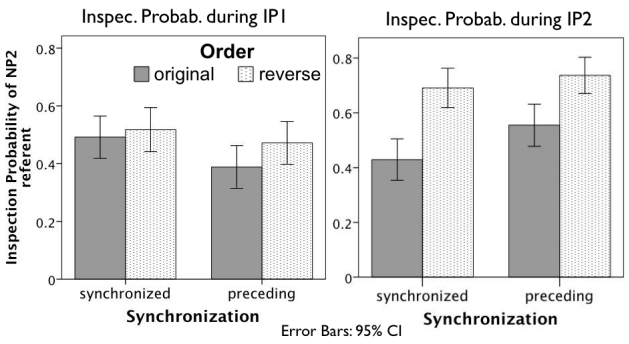


Figure 4: Inspection probability on NP2 referent in Exp2, for all conditions in IP1 (left graph) and IP2 (right graph).

In IP2, Order of Mention had a main effect on inspection probabilities ($\chi^2(1) = 51.99, p < .001$). That is, during NP2 noun mentioning, people inspected the NP2 referent more frequently in reverse order than in original order. As in Experiment 1, this suggests that people visually attended more closely to the mentioned object when the referring expression required more effort to be resolved.

Response Time Model reduction revealed a significant interaction of both predictors, Order of Mention and Synchronization ($\chi^2(1) = 16.85, p < .001$). Consequently, both predictors were included in the model fitted to response times. This model is specified by our dependent variable response time, the two predictors Order and Synchronization, and two random factors accounting for subject and item variation ($DV \sim Predictor1 \times Predictor2 + randomFactors$, see also Table 1). Both factors had a marginal main effect, however, the interaction is clearly more relevant for interpretation as is explained below. Firstly, pairwise comparisons reveal the following significant differences: Between reverse-neutral and reverse-synchronized ($p < .001$), reverse-synchronized and original-synchronized ($p < .05$), reverse-neutral and original-neutral ($p < .001$) and a marginally significant difference between original-synchronized and original-neutral ($p = .07$). These results suggest that order of references in a sentence indeed affected participant behavior, as already suggested by the inspection data in IP1. The response times in both neutral conditions show that people were significantly faster to validate the robot's utterance in the reverse-neutral condition than in original-neutral. This result is consistent with the findings of visual anticipation of the NP2 referent (for neutral gaze), i.e., when order was reversed people anticipated the NP2 referent, when order was original they hardly did. This suggests that reverse order of mention was generally *easier* to process than original order of mention. However, synchronization of gaze cues reversed this effect: Participants were significantly slower when gaze was synchronized and in reverse order (resulting in concurrent but conflicting referential cues) than when gaze was synchronized and in original order (concurrent and coherent order of cues).[1]

[1]The response time pattern in Experiment 2 was largely independent of the chosen onset. That is, results were qualitatively equal for starting recording at NP2-adjective onset or at NP2-noun onset.

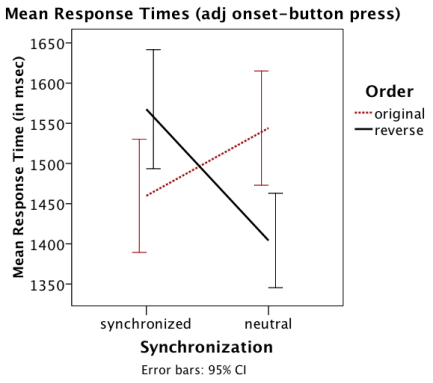


Figure 5: Avg. response times in all four conditions (Exp2).

Table 1: Model fitted to response time data. The last column shows p-Values calculated through Monte-Carlo-sampling.

Predictor	Coeff.	*SE*	t-value	pMCMC
(Intercept)	1475.79	55.19	26.741	$<.001$
Order-reverse	96.24	40.36	2.384	$<.05$
Synchr.-neutral	68.89	39.60	1.740	.075
reverse.:neutral	-230.67	55.94	-4.124	$<.001$

$Model: RT \sim Order of Mention \times Synchronization + (1|subject) + (1|item)$

Discussion and Conclusion

Results from Experiment 1 suggest that gaze cues are equally beneficial when preceding the spoken references as when they occur concurrently. However, the interpretation with regard to the influence of cue ordering was difficult as the manipulation of order was potentially confounded with the sentence order (i.e., referent location in the scene). This was addressed by adding a neutral gaze condition in Experiment 2 which revealed that reverse sentence order was *easier* to process than original sentence order. Thus, despite the advantage of reverse sentence order, synchronizing (reverse) robot gaze cues disrupted people whereas adding original (and coherently) ordered gaze cues to original sentence order significantly enhanced response time of this sentence order. The results for synchronized robot gaze may therefore be interpreted with respect to gaze and speech cue synchronization only: Synchronizing (reversed) gaze cue with reverse order of mention *increased* response times, while synchronizing (coherent) gaze cues with original order of mention *reduced* response times, when each is compared to its neutral gaze baseline.

The presented results thus suggest that large temporal shifts of robot gaze with respect to its 'natural' synchronization do not substantially affect the utility of the gaze cues whereas the order of the cues does. This contradicts the predictions derived from the Visual Account. The Intentional Account, in contrast, provides a plausible explanation for these results: The precise temporal synchronization is not critical since people interpret and use robot gaze as a cue to the robot's intentions. This may also explain why Kreysa (2009) found that substantial temporal shifts reduce the gaze cursor's utility while robot gaze and speech synchronization, in contrast, appears rather flexible. The order of cues, however, affects the the utility of robot gaze since the order of inspections reflects the speaker's intentions regarding order of mention. Thus, people seem to expect that the inferred referential intentions be realized in the corresponding order (Griffin & Bock, 2000). If this expectation is not met, gaze cues may even disrupt comprehension, as the comparison with neutral gaze suggests. Consequently, the presented evidence for a flexible use of robot gaze during utterance comprehension further supports the hypothesis that people assign attentional and intentional states to the robot. Thus, future research could use such a robot interaction setting to generally address questions such as the extent to which people infer intentions or other information from their partner's gaze. While it is not entirely clear whether robots and agents provide an unrestricted experimental test bed such that results generalize to human-human interaction, our results, among others (Breazeal, Kidd, Thomaz, Hoffman, & Berlin, 2005), suggest that people do establish basic joint attention also with robots (or artificial agents in general). However, this phenomenon is likely to depend on people's beliefs in the agent's competence or appearance, in particular, when signaling information processes and functionalities different from those of a human.

Acknowledgments

The research reported of in this paper was supported by the IRTG 715 "Language Technology and Cognitive Systems" funded by the German Research Foundation (DFG).

References

Baayen, R., Davidson, D., & Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390-412.

Baron-Cohen, S. (1995). *Mindblindness: An essay on autism and theory of mind*. MA: MIT Press/Bradford Books.

Bates, D. (2005). Fitting linear mixed models in R. *R News*, *5*, 27-30.

Breazeal, C., Kidd, C., Thomaz, A., Hoffman, G., & Berlin, M. (2005). Effects of nonverbal communication on efficiency and robustness in human-robot teamwork. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'05)* (p. 708-713).

Emery, N. (2000). The eyes have it: the neuroethology, function and evolution of social gaze. *Neuroscience and Biobehavioral Reviews*, *24*, 581-604.

Flom, R., Lee, K., & Muir, D. (Eds.). (2007). *Gaze-Following: Its Development and Significance*. Mahwah, NJ, US: Lawrence Erlbaum Associates.

Griffin, Z. M., & Bock, K. (2000). What the eyes say about speaking. *Psychological Science*, *11*, 274-279.

Hanna, J., & Brennan, S. (2007). Speakers' eye gaze disambiguates referring expressions early during face-to-face conversation. *Journal of Memory and Language*, *57*, 596-615.

Kreysa, H. (2009). *Coordinating speech-related eye movements between comprehension and production*. Unpublished doctoral dissertation, University of Edinburgh.

Langton, S. R., Watt, R. J., & Bruce, V. (2000). Do the eyes have it? Cues to the direction of social attention. *Trends in Cognitive Science*, *4*, 50-59.

Meyer, A., Sleiderink, A., & Levelt, W. (1998). Viewing and naming objects: Eye movements during noun phrase production. *Cognition*, *66*, B25-B33.

Schroeder, M., & Trouvain, J. (2001). The German Text-to-Speech Synthesis System MARY: A Tool for Research, Development and Teaching. In *4th isca workshop on speech synthesis.* Blair Atholl, Scotland.

Staudte, M., & Crocker, M. W. (2009a). The effect of robot gaze on processing robot utterances. In N. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31th Annual Conference of the Cognitive Science Society.* Amsterdam, Netherlands.

Staudte, M., & Crocker, M. W. (2009b). Visual Attention in Spoken Human-Robot Interaction. In *Proceedings of the 4th ACM/IEEE Conference on Human-Robot Interaction (HRI'09).* San Diego, USA.

Tanenhaus, M. K., Spivey-Knowlton, M., Eberhard, K., & Sedivy, J. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*, 1632-1634.

How Action Understanding can be Rational, Bayesian *and* Tractable

Mark Blokpoel[a,b], **Johan Kwisthout**[a], **Theo P. van der Weide**[a], **Iris van Rooij**[b]
(blokpoel@acm.org), (johank@science.ru.nl), (tvdw@cs.ru.nl), (i.vanrooij@donders.ru.nl)

[a]*Radboud University Nijmegen, Institute for Computing and Information Sciences*
[b]*Radboud University Nijmegen, Donders Institute for Brain, Cognition and Behaviour*

Abstract

An important aspect of human sociality is our ability to understand the actions of others as being goal-directed. Recently, the now classic rational approach to explaining this ability has been given a formal incarnation in the Bayesian Inverse Planning (BIP) model of Baker, Saxe, and Tenenbaum (2009). The BIP model enjoys considerable empirical support when tested on 'toy domains'. Yet, like many Bayesian models of cognition, it faces the charge of computational intractability: i.e., the computations that the model postulates may be too resource demanding for the model to be scalable to domains of real-world complexity. In this paper, we investigate ways in which the BIP model can possibly parry the charge. We will show that there are specific conditions under which the computations postulated by the model are tractable, despite the model being rational and Bayesian.

Keywords: *goal inference, inverse planning, computational complexity, intractability, NP-hard, fixed-parameter tractability*

Introduction

Imagine a mother and her son, sitting in the same room, when she hears his stomach rumble. She sees her son get up, walk to the kitchen and start searching for something. At first he finds a sour apple, which he discards in search of something else. Then the mother sees her son finding a delicious candy bar. When he starts to eat it she realizes her son is trying to still his hunger and at the same time wanting to eat something sweet. In this scenario, the son goes through a process of *planning*, choosing his actions to achieve his goals. The mother observes the actions of her son and based on her observations infers the goals she thinks her son is trying to achieve. This process is called *goal inference*.

In line with a long tradition of explaining the human ability to understand actions as goal-oriented (Dennett, 1987; Charniak & Goldman, 1991; Csibra, Gergely, Biró, Koós, & Brockbank, 1999; Cuijpers, Schie, Koppen, Erlhagen, & Bekkering, 2006), Baker, Saxe, and Tenenbaum (2009) have proposed that goal inference can be seen as a form of *inverse planning*, just as vision is a form of inverse graphics. Baker et al. go beyond existing psychological approaches by providing a precise formalization of 'inverse planning' in the form of a Bayesian inference model. We will refer to this model as the BIP model of goal inference (where BIP stands for Bayesian Inverse Planning). The BIP model has been tested in several experiments, and Baker et al. (2007, 2009) observed that it can account for the dynamics of goal inferences made by human participants in several different experimental settings.

According to the BIP model, observers assume that actors are 'rational' in the sense that they tend to adopt those actions that best achieve their goals. Given the assumption of rationality, and (probabilistic) knowledge of the world and how actions are affected by it, one can compute the probability that an agent performs an action given its goals, denoted

$$P(action \mid goal,\ environment) \tag{1}$$

When observing a given action, the probability in (1) can be inverted using Bayes' rule to compute the probability of a given goal:

$$P(goal \mid action,\ environment) \propto$$
$$\mathrm{P}(action \mid goal,\ environment)\, P(goal \mid environment) \tag{2}$$

Of all the possible goals that an observer can (or does) entertain, the goal that maximizes the probability in (2) best explains why the observed action was performed and is the goal that is inferred.[1]

Given that the BIP model belongs to the class of (rational) Bayesian inference models—and Bayesian inference is known to be intractable if no additional constraints are imposed (e.g. Chater, Tenenbaum, and Yuille (2006); see also Kwisthout (2009))—the question arises if the computations that it postulates can scale to situations of everyday complexity. As Gigerenzer and colleagues put it:

> The computations postulated by a model of cognition need to be tractable in the real world in which people live, not only in the small world of an experiment with only a few cues. This eliminates NP-hard models that lead to computational explosion, such as probabilistic inference using Bayesian belief networks ... including its approximations. (Gigerenzer, Hoffrage, and Goldstein (2008) p. 236)

Although we share the stance of Gigerenzer et al. (2008) towards intractable (NP-hard) models of cognition, we are not as pessimistic about the viability of Bayesian models. In our view, the key to understanding the computational feasibility of a Bayesian (or any cognitive) model lies in studying domain-specific constraints that hold in the model's domain of application (e.g., action understanding or vision) and investigating if and how such constraints may render the computations postulated by the model tractable for its domain,

[1] In other words, in the BIP model, goal inference is conceptualized as a form of probabilistic inference to the best explanation, a.k.a. *abduction* (e.g. Charniak and Shimony (1990)).

despite the intractability of those computations in general. In this paper we set out to perform such an investigation for the BIP model of goal inference.[2]

The remainder of this paper is organized as follows. We first introduce specific versions of the BIP model that Baker et al. (2007, 2009) formulated to account for their experimental data and observe that these versions are tractable but also too specific. We then propose a generalized model that breaks some implausible constraints in the original models. After this we introduce a method that we use to analyze the computational (in)tractability of the generalized model. We then give an overview of the (in)tractability results, and discuss their implications for Bayesian models of goal inference and for dealing with the intractability of Bayesian models in general.

Computational Models

Baker et al. (2009) propose three different versions of Bayesian Inverse Planning (M1, M2 and M3) to account for data gathered in several maze experiments. These two-dimensional maze experiments, based on earlier work (Gergely, Nádasdy, Csibra, & Biró, 1995; Schultz et al., 2003), were designed to assess subjects' inferences about the goals of a planning agent. Subjects were shown videos of agents moving in a maze, such as those in Fig. 1, and under different timing and information conditions had to infer the goal of the agent. In these experiments *changes in location* were considered *actions* and the *location* of the agent is considered its *state*. Specific locations (A, B and C) were possible *goals*.

All three models M1–3 can be seen as special cases of a more general BIP model, as depicted in Fig. 2, in which there is a goal structure template **G** that can encode different types of goal structures.[3] The simplest goal structure is present in M1 where the observer assumes that the agent has one single goal that does not change over time (Fig. 3(a)). In M2 the model allows the observer to infer the agent has a different goal at any given time (Fig. 3(b)). This models the ability of people to infer changes in an agent's goal over time. For instance, if someone is inspecting the contents of her fridge, you may infer she wishes to cook dinner, but when she closes the fridge, puts on her coat, and leaves the house, you may infer she is going to eat out. Finally, in M3 the goal structure encodes hierarchical goals (Fig. 3(c)), such that the observer can infer changes in the agent's sub-goals, which are subserving a common high-level goal. For instance, when you see someone gathering kitchen utensils, each individual gathering can be a sub-goal but the high-level goal is to cook dinner.

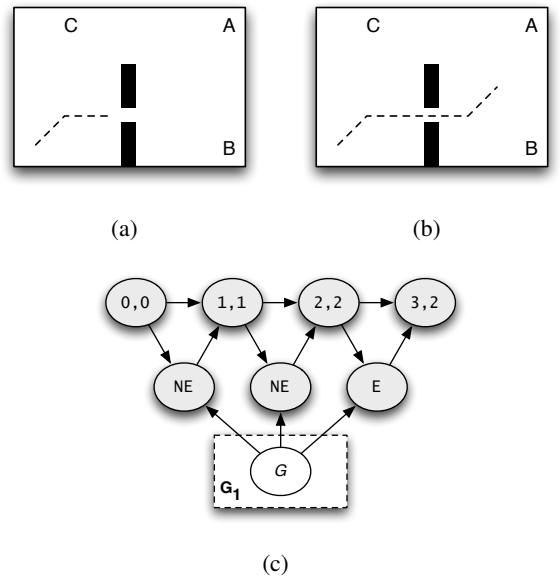


Figure 1: An illustration of the types of stimuli used in the maze experiments of Baker et al. (2009). Participants observe an agent (and the trail history as memory aid) move inside the maze, and are asked to judge which of the three possible goals (A, B or C) is most likely the agent's goal. Here (a) depicts an early judgement point where both human participants and the model infer B as most likely goal. (b) depicts a later judgment point where both human participants and the model infer A as most likely goal. (c) A possible BIP model (in this case M1) for the early judgement point.

Even though inference in Bayesian networks is hard in general, the BIP models proposed by Baker et al. are tractable.[4] This tractability is in some sense an artifact of the simplified experiments for which these models were designed. In the experiments an agent never has more than one (high-level) goal at any given time. This property does not seem to hold in general, however. Reconsider, for instance, the scenario in our opening paragraph. There the mother infers that the son wants to satisfy his hunger *and* he wants to eat something sweet. This type of goal inference where multiple goals are inferred at the same time cannot be modelled by M1, M2 or M3. To accomodate for this observation, we propose an extension called MULTIPLE GOALS BIP or MGBIP. Fig. 4

[2]The authors are well aware of common claims of approximability of Bayesian inferences, and that approximation is generally believed to provide a way to overcome the intractability of Bayesian models. In this paper, we will depart from this standard viewpoint for two reasons. First, the claims of approximability seem at worst incorrect and at best unfounded; for instance it is known that approximating the most probable explanation in a Bayesian network is itself also intractable (Abdelbar & Hedetniemi, 1998). Second, we believe that there are other, better ways of dealing with the intractability of cognitive models, viz., by identifying model constraints that render otherwise intractable models tractable (van Rooij, 2008).

[3]In the original BIP models (M1, M2 and M3) Baker et al. used additional parameters to model the effect of noise (β), the probability of changing a goal in M2 (γ) and the probability of having sub-goals in M3 (κ) to fit the model to the experimental data. As these parameters are assumed constants, they can be safely ignored for the purposes of our analyses.

[4]For the formal proof of these claims we refer the reader to the Supplementary materials available online at `http://tinyurl.com/suppl2010`

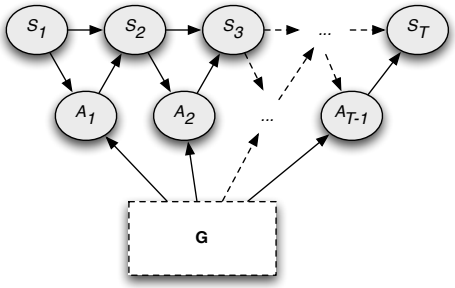


Figure 2: A graphical representation of the dynamic Bayesian network that describes the general form of BIP. Nodes represent variables, for example state node S_t is a variable and it can assume values corresponding to the state of the agent at time t. Arrows represent dependencies, for example the probability that a state S_{t+1} has a certain value depends on the previous state S_t and previous action A_t. States and actions are observed, i.e. the values of the states and action variables are given as input to the model. Given these observations the most probable combination of values for the goal variables in **G**. Finally, shaded nodes are observed and their values considered part of the input of the model. Examples of the possible contents of **G** are illustrated in Fig. 3

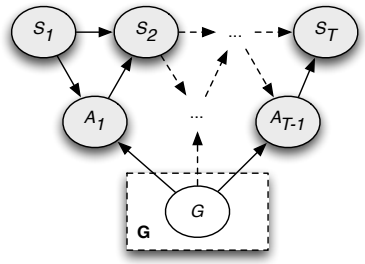
illustrates the dynamic Bayesian network of MGBIP.[5]

Because it is more general, MGBIP has wider range of applicability than M1–3. The introduced generality also comes at a cost: Whereas M1, M2 and M3 are tractable, MGBIP is intractable, in the sense that there are no tractable (more precisely: polynomial time) algorithms that can implement this model.[4] Even so, in real-world situations humans are often able to quickly infer an agent is pursuing multiple simultaneous goals. This suggests that, if MGBIP is to be psychologically plausible, we need to assume that some domain-specific constraints apply in those situations that render the goal inferences tractable under the MGBIP model (despite the model being intractable without such additional constraints). The next section describes how we set out to indentify such possible constraints.

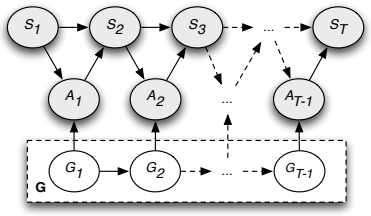
Identifying Sources of Intractability

In order to find constraints on the input domain of MGBIP that render the (restricted) model tractable, we adopt a method for identifying sources of intractability as described in (van Rooij, Evans, Müller, Gedge, & Wareham, 2008) (see also van Rooij and Wareham (2008)). The method works as follows.

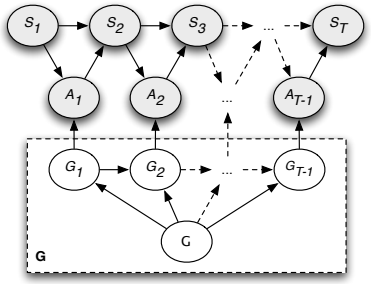
[5]Baker et al. (2009) also note, that the simplified models M1–3 unlikely suffice to model human action understanding in general and they argue that the models will need to be extended in various directions if they are to apply to real-world scenarios. Our extension can be seen as one such possible direction in which to extend the model. Other directions of extension are possible as well (see e.g. (Ullman, Baker, Macindoe, Goodman, & Tenenbaum, 2009).



(a) M1



(b) M2



(c) M3

Figure 3: Graphical representation of **G** for M1, M2 and M3. In M1 (a) goals are modeled by a single static goal. All actions are dependent on this goal. In M2 (b) goals can change over time. Actions at time t are dependent on goals at time t. In M3 (c) goals can consist of multiple subgoals. Actions at time t are dependent on subgoals at time t.

First, one identifies a set of model *parameters* $K = \{k_1, k_2, ..., k_m\}$ in the model M under study (for us, MGBIP). Then one tests if it is possible to solve M in a time that can grow excessively fast (more precisely: exponential or worse) as a function of the elements in K yet slowly (polynomial) in the size of the input.[6] If this is the case, then M is said to be *fixed-parameter (fp-) tractable* for parameter set K, and otherwise it is said to be *fp-intractable* for K.

Observe that if a parameter set K is found for which M is fp-tractable then the problem M can be solved quite efficiently, even for large inputs, provided only that the members

[6]More formally, this would be a time on the order of $f(k_1, k_2, ..., k_m)n^c$, where f is an arbitrary computable function, n is a measure of the overall input size, and c is a constant.

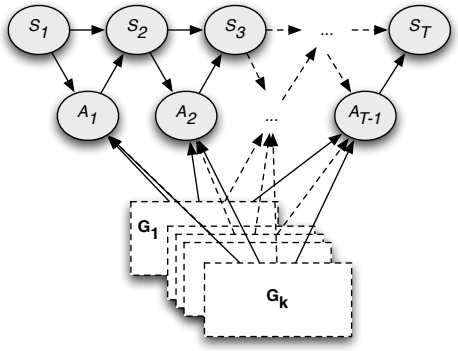


Figure 4: Graphical representation of the dynamic Bayesian network that describes MULTIPLE GOALS BIP (MGBIP).

of K are relatively small. In this sense the "unbounded" nature of K can be seen as a reason for the intractability of M. Therefore we call K a *source of intractability* of M.

The MGBIP model has several natural parameters, each of them a candidate source of intractability. In this paper we consider, five such parameters (see Table 1 for an overview and Fig. 5 for an illustration).

First consider parameters T, denoting the maximum number of observations the observer makes, and $1/T$, denoting the poverty of observations. Note that T is small if few observations are made, and $1/T$ is small if many observations are made. Based on intuition one might think, the less information we have, the harder it is to understand actions. This makes $1/T$ a candidate source of intractability. However as T grows, so does the size of the network and the necessary number of calculations and this also makes T a likely candidate source of intractability.

Second, parameter k is the maximum number of multiple goals that (the observer assumes) the agent can pursue. This parameter is also an excellent candidate source of intractability, because large k's introduce an exponential number of combinations of possible multiple goals leading to a combinatorial explosion.

Third, the parameter g is the maximum number of goal values per goal variable. As the number of possible values that a goal variable can take increases the necessary number of calculations, also g is a candidate source of intractability.

Finally, the parameter $1-p$ measures how far the most likely goal inference is from being completely certain (here p is the probability of the most likely explanation). If $1-p$ is small, this means that the most likely explanation is much more likely than any competitor explanation. If the value is large, it means that the most likely explanation has many competitor explanations of non-negligable probability (see e.g. Table 2). It seems intuitive that finding the most likely explanation is easier in the former case than in the latter case, and therefore also $1-p$ can be considered a candidate source of intractability.

Table 1: A list of parameters with short descriptions and their values based on the running example.

parameter	*description*	*value*
T	maximum observations	6
$1/T$	maximum observation poverty	$1/6$
k	maximum # multiple goals	2
g	maximum # goal values	3
$1-p$	distance from certainty	0.4

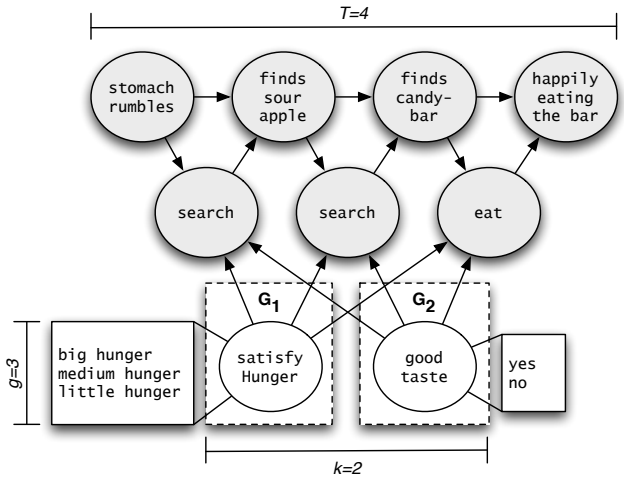


Figure 5: Illustration of the Bayesian network and different parameters of the MGBIP model applied to the "mother observes son"-example.

Table 2: Example probability distribution over the combinations of goal values. In this example $p = 0.6$ and $1-p = 0.4$.

satisfy hunger	desire sweet	P
big hunger	yes	0.05
medium hunger	yes	0.05
little hunger	yes	0.6
big hunger	no	0.3
medium hunger	no	0.0
little hunger	no	0.0

We have now reviewed five parameters that—on intuitive grounds—may be considered candidate sources of intractability in the MGBIP model. It is known, however, that human intuitions about what makes a computation tractable or intractable can be mistaken. Therefore it is necessary to verify such intuitions by means of mathematical proof.

Results

In this section we present our fp-(in)tractability results for the different parameters of the MGBIP model, and we explain how these results bear on the question 'which constraints render the MGBIP model tractable?'. Full details and proofs can

be found in the Supplementary materials.[7]

Result 1. MGBIP is fp-intractable for every subset of parameters $K \subseteq \{T, 1/T, g\}$.

Result 1 shows—contrary to the intuitions sketched in the previous section—that none of the parameters T, $1/T$ and g, nor any combination of them is a source of intractability for MGBIP. This means that even if we assume that one or more of these parameters is small for the domain of application, goal inference under the MGBIP model is still intractable.

Besides this negative result (Result 1), we also have two positive results (Results 2 and 3).

Result 2. MGBIP is fp-tractable for parameter $\{k\}$.

Result 2 confirms parameter k is a source of intractability. This means that goal inference is tractable under the MGBIP model provided only that we impose the constraint that (the observer assumes that) the agent can pursue only a handful of goals simultaneously. Importantly, this is true regardless the size of T, $1/T$, g or $1-p$. This is quite a powerful result, with great potential for explaining the speed of real-world goal inferences within the confines of a BIP model. After all, it seems to be a plausible constraint that real-world agents do not (typically) pursue a large number of goals in parallel at the same time (possibly also to keep their own planning tractable).

Result 3. MGBIP is fp-tractable for parameter $\{1-p\}$.

Finally, Result 3 confirms parameter $1-p$ is a source of intractability. This means that goal inference is tractable under the MGBIP model for those inputs where the most probable goal explanation is quite probable. Again, this is true regardless the size of T, $1/T$, g or k. Also, this result has potential for explaining the speed of real-world goal inferences within the confines of a BIP model, at least for certain situations—viz., those situations where the actions of the observed agents unambigously suggest a particlar combination of goals. For all we know, real world cases of speedy goal inference may very well match exactly these situations. Whether or not this is indeed the case is an empirical question which can be addressed by testing the speed of human goal inference for different degrees of goal ambiguity.

Discussion

We have analyzed the computational resource requirements of the Inverse Bayesian Planning (BIP) model of goal inference in order to study its viability as a model of inferences made by resource-bounded minds as our own. We generated several interesting theoretical findings. First, we observed that the three specialized models—M1, M2, and M3—that were developed by Baker et al. (2007, 2009) to account for their experimental data in maze experiments are in fact computationally tractable. This means that these specialized Bayesian models do not seem to make unrealistic assumptions about the computational powers of human minds/brains, even when operating on large networks of beliefs and observations. That being said, these models do seem to be theoretically problematic for a different reason: they are too specialized to count as models of goal inference in general.

The over-specialization of M1, M2 and M3 is revealed when pondering the assumptions that these models make about the agent and the observer. For instance, all three models assume that (the observer assumes that) the agent can pursue at most one goal at a time. In the real-world, however, people often can and do act in ways so as to try and achieve two or more goals at the same time, and observers can also often understand what these simultanous goals are from observing the actors behave in systematic ways. Recall, for example, the scenario from our Introduction where the son searches the kitchen for a candy bar. Under different circumstances, the mother may understand that her son has the goal to still his hunger (goal 1), to satisfy his craving for sweet (goal 2), to see how many bars are left (goal 3), to pretend that he did not hear his mom's request to clean up his room (goal 4), to bring back a candy bar for his mom (goal 5), etc., or any combination of these goals.

To accomodate the fact that real-world goal inference is not restricted to one goal at a time, we defined a more general BIP model—having M1, M2 and M3 as special cases—which we refer to as MULTIPLE GOALS BIP, or MGBIP for short. Complexity Analysis of the MGBIP model revealed that it *is* computationally intractable (i.e., NP-hard), meaning that this model, in all its generality, does indeed make unrealistic assumptions about the computational powers of human minds/brains. We took this negative theoretical result to mean that—if the BIP model is to account for human goal inference at all—it must be the case that in those situations where humans are able to infer multiple simulatenous goals quickly and effortlessly, specific constraints apply that render the inferences under the MGBIP model tractable.

To investigate which types of constraints could render the MGBIP model tractable, we used a methodology for identifying sources of intractability in NP-hard computational models (e.g. van Rooij and Wareham (2008)) and derived several theoretical results. For instance, we ruled out the possibility of explaining speedy real-world (multiple) goal inferences by an appeal to small values of T (modeling situations when goals can be inferred using only few observations) or an appeal to large values of T (modelling situations where a lot of information is available on which to base a goal inference). Similarly, we ruled out that the speed of such inferences could be explained by an appeal to a small number of values per goal node. Besides these negative theoretical results, we also had two important positive results. For one, we established that as long as the number of goals that can be simulatenously pursued, k, is not too large then goal inference is tractable under the MGBIP. Secondly we have shown that goal inference is tractable under the MGBIP model whenever the probability

[7] See `http://tinyurl.com/suppl2010`

of the most likely combination of simultaneous goals, p, is not too far from 1.

Whereas our negative theoretical results are useful to clarify that tractability is not a property that is trivially achieved (and often our intuitions about what constraints would render a model tractable can be wrong; cf. van Rooij et al. (2008)), our positive results show that a model of action understanding can nevertheless be rational, Bayesian, and tractable. Moreover, the nature of the constraints that need to be introduced to render the Bayesian Inverse Planning model of goal inference tractable yield new empirically testable predictions.

For instance, based on our results, we predict that human participants will be able to make quick and accurate goal inferences in the types of experimental set-ups such as used by Baker et al. (2007) (but see also Csibra et al. (1999)), but only if the number of simulatenous goals that the observed agents are pursuing is not too large, or the probability of the most likely combination of goals is not too small, or both. If both of these contraints were to be alleviated at the same time, we would predict that human performance on the goal inference task would deteriorate significantly. If our prediction were to be confirmed then this would provide corroborative support for the BIP model of goal inference, and validate that our theoretical results help explain the tractability of human goal inferences. If, on the other hand, the prediction were to be disconfirmed, then this would suggest that either the BIP model fails as an account of human goal inferences, or some constraint other than the ones we considered also suffices to render the BIP model tractable. The latter option may then be one that BIP modelers may be interested in pursuing further.

In closing, we remark that our approach can be seen as exemplary of a general strategy for dealing with intractability in Bayesian models, whether of action understanding or otherwise. Our approach reveals that—contrary to popular belief—Bayesian models *can* possibly scale to complex, real-world domains. To achieve this, Bayesian modelers need only identify constraints that apply in the real-world and suffice to render their models' computations tractable. By restricting Bayesian models in this way these models also become better testable: the constraints required to guarantee tractability of the models yield new predictions (specifically, about the speed of inferences) that can be used to perform more stringent tests of such models.

Acknowledgments

The 2nd author has been supported by the OCTOPUS project under the responsibility of the Embedded Systems Institute. The OCTOPUS project is partially supported by the Netherlands Ministry of Economic Affairs under the Embedded Systems Institute program.

The authors thank Vanessa Ferdinand and Max Hinne for comments on an earlier version of this paper.

References

Abdelbar, A. M., & Hedetniemi, S. M. (1998). Approximating MAPs for belief networks is NP-hard and other theorems. *Artificial Intelligence*, *102*, 21–38.

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*(113), 329–349.

Baker, C. L., Tenenbaum, J., & Saxe, R. (2007, Jan). Goal inference as inverse planning. *Proceedings of the 29th meeting of the Cognitive Science Society*.

Charniak, E., & Goldman, R. P. (1991). A probabilistic model of plan recognition. In *Association for the advancement of artificial intelligence* (pp. 160–165).

Charniak, E., & Shimony, S. E. (1990). Probabilistic semantics for cost based abduction. In *Aaai* (p. 106-111).

Chater, N., Tenenbaum, J. B., & Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences*, *10*(7), 287 - 291. (Special issue: Probabilistic models of cognition)

Csibra, G., Gergely, G., Biró, S., Koós, O., & Brockbank, M. (1999). Goal attribution without agency cues: the perception of 'pure reason' in infancy. *Cognition*, *72*, 237–267.

Cuijpers, R. H., Schie, H. T. van, Koppen, M., Erlhagen, W., & Bekkering, H. (2006). *Goals and means in action observation: A computational approach.*

Dennett, D. C. (1987). *The intentional stance*. Cambridge, MA: The MIT Press.

Gergely, G., Nádasdy, Z., Csibra, G., & Biró, S. (1995). Taking the intentional stance at 12 months of age. *Cognition*, *56*, 165–193.

Gigerenzer, G., Hoffrage, U., & Goldstein, D. G. (2008). Fast and frugal heuristics are plausible models of congition: Reply to Dougherty, Franco-Watkins, and Thomas (2008). *Psychological Review*, *115*, 230–239.

Kwisthout, J. H. P. (2009). *The computational complexity of probabilistic networks*. Unpublished doctoral dissertation, Faculty of Science, Utrecht University, The Netherlands.

Schultz, R. T., Grelotti, D. J., Klink, A., Kleinman, J., Gaag, C. van der, & Marois, R. (2003). The role of the fusiform face area in social cognition: Implications for the pathobiology of autism. *Philosophical Transactions of Royal Society of London, Series B: Biological Scences*(358(1430)), 415–427.

Ullman, T. D., Baker, C. L., Macindoe, O., Goodman, O. E. N. D., & Tenenbaum, J. B. (2009). Help or hinder: Bayesian models of social goal inference. *NIPS*.

van Rooij, I. (2008). The tractable cognition thesis. *Cognitive Science*, *32*, 939–984.

van Rooij, I., Evans, P., Müller, M., Gedge, J., & Wareham, T. (2008). Identifying sources of intractability in cognitive models: An illustration using analogical structure mapping. In K. M. B. C Love & V. M. S. (Eds.) (Eds.), *Proceedings of the 30th annual conference of the cognitive science society* (pp. 915–920). Austin, TX: Cognitive Science Society.

van Rooij, I., & Wareham, T. (2008). Parameterized complexity in cognitive modeling: Foundations, applications and opportunities. *Computer Journal*, *51*(3), 385–404.

Facilitating Low-Achieving Students' Diagram Use in Algebraic Story Problems

Julie L. Booth (julie.booth@temple.edu)
Department of Psychological Studies in Education, 1301 W. Cecil B. Moore Avenue
Philadelphia, PA 19122 USA

Kenneth R. Koedinger (koedinger@cmu.edu)
Human Computer Interaction Institute, Carnegie Mellon University
Pittsburgh, PA 15213 USA

Abstract

Recent research indicates that when solving algebraic story problems, adding a diagram is beneficial for seventh and eighth grade students, however, sixth graders—particularly low-achieving ones—do not benefit from the diagrams. In the present study, we further investigate the diagrammatic advantage in low-achieving pre-algebra students and examine whether and how picture algebra instruction improves diagram comprehension and use in the target population. Results replicate the lack of diagrammatic advantage in this population for two types of diagrams. Picture algebra instruction on mapping information in word problems to one type of diagrams yields improvement in both diagrammatic forms, but not story problems without diagrams; a diagrammatic advantage emerges following this instruction. Though low-achieving students may fail to use diagrammatic representations to their benefit when solving word problems, instruction on the use of one specific form may be sufficient to facilitate a more general diagrammatic advantage.

Keywords: Multiple representations; Algebraic problem solving; mathematics education

Introduction

Problem representation is a critical issue in education, as the way that information is conveyed to students can have a great impact on the degree to which they learn. (e.g., Cummins, Kintsch, Reusser, & Weimer, 1988). In the domain of mathematics, use of more grounded representations rather than more abstract ones (e.g., verbal descriptions of situations as opposed to equations), has been found to be useful for presenting simple algebra problems (Koedinger, Alibali, & Nathan, 2008); this practice may be especially useful for making problems concrete when students are early in their transition to algebraic thinking and are not yet capable of the abstract thinking necessary to comprehend equations (Koedinger & Nathan, 2004).

Another way that instructors often attempt to make problems or situations more concrete is to include external representations. External representations are an important part of mathematics education (Seeger, 1998), and are intended to increase understanding of mathematical concepts by allowing children to build relations between mathematical ideas (Hiebert & Carpenter, 1992). Pictorial representations, such as diagrams, charts, graphs, and tables, are often used in math classrooms because they are thought to be useful for helping students communicate and reason about mathematical concepts (Greeno & Hall, 1997), and the National Council of Teachers of Mathematics recommends that teachers include multiple forms of representations when teaching mathematical concepts (NCTM, 2000).

Indeed, evidence abounds on the cognitive benefits of external representations, including diagrams. For example, students learn better when a diagram is added to text than if they are studying the text alone (Mayer, 1989; 2009), likely because learners are able to build two mental representations of multimedia material, a verbal representation and a visual one, and build connections between them (Mayer, 2005). Diagrams may also be beneficial because the spatial organization and grouping of related components that are characteristic of diagrams better enables users to search, recognize relevant pieces in, and draw inferences about the represented information (Larkin & Simon, 1987). Diagrams may also be beneficial because they promote users to engage in self-explanation, which is in itself beneficial for learning (Ainsworth & Loizou, 2003).

Are diagrams universally helpful?

Despite the intention of these tools to help students succeed, the use of diagrams is not always beneficial. Larkin and Simon (1987) posited that diagrammatic representations of any sort are only useful if they are constructed in a way that groups information and facilitates inference in a better way than is possible with text. Further, even a well-constructed diagram will not be useful unless the user knows the computational processes that are necessary for taking advantage of them. Ainsworth (2006) also cautions that the usefulness of diagrams is influenced by characteristics of the user such as expertise in the content domain and familiarity with the structure and components of the representation, as well as characteristics of the diagram and interactions between the two.

Consistent with these assertions, recent research on using diagrams with algebraic story problems suggests that not all students benefit from the addition of diagrams. Booth & Koedinger (2007) found that older and higher-achieving middle school students do benefit from the diagrams as intended—they solve more diagram problems correctly than problems without a diagram. However, low-achieving students do not benefit from the diagrams; they perform better on story problems that do not have accompanying diagrams. In fact, the diagrams may actually hurt their performance—they perform just as poorly on the diagram problems as they do when solving the problems as symbolic equations.

For students that do experience a diagrammatic advantage, results suggested that the benefit comes from protecting those students from making common conceptual errors in interpreting the problem (Booth & Koedinger, 2007). For example, for a problem where students are given a sale price and asked to determine the original price if the buyer purchased it at 1/5 off, having a diagram showing the pieces of the equation makes it less likely that higher-achieving students solve the problem by multiplying the original price by 5, which is a common strategy for students solving the problem in story format. Of course, this benefit can only be realized when students are able to effectively use diagrams to understand the problem.

For lower-achieving students, there appear to be two barriers to successful diagram use. One is that they are less likely than higher-achieving peers to attempt diagram problems. This is perhaps unsurprising, as low-ability students generally perceive problems to be more difficult than high or average ability students, and are more likely to shut down and not attempt the problems as a result (Ericcson & Simon, 1980). The real or perceived need to attend to more than one representation at a time causes split attention demands on working memory (Chandler & Sweller, 1991), making the problem seem overwhelming, and these students' limited diagram comprehension skills preclude the realization that the problem could be solved by simply ignoring the diagram and working from the story alone.

The other barrier to success with diagrams was the failure to glean a correct conceptual understanding of the problem from looking at the diagram. Young and low-achieving students were *more* likely to make conceptual errors in problems that included diagrams than ones with stories alone. This is likely due to either misinterpretation of the diagram itself or, more crucially, failure to accurately map the story problem to the diagram. How can we help low-achieving students to better comprehend the diagrams?

Using instruction to improve diagram use

Research from the fields of cognitive development and mathematics education suggests that effective instruction on external representations is necessary for correct student use (Sowell, 1989; Fueyo & Bushell, 1998; Uttal, Scudder, & DeLoache, 1997). Brief instruction on a particular visual representation may not suffice (Rittle-Johnson & Koedinger, 2001), but more involved representation-specific instruction could consume a significant amount of precious classroom time, and may not transfer well to other representations.

An alternative to instruction on utilizing particular types of diagrams is having students construct diagrams to represent story problems themselves. Middle school students can use self-created representations to successfully solve algebraic story problems they wouldn't ordinarily be able to solve (Koedinger & Terao, 2002), and both constructing diagrams from scratch or filling in partially completed diagrams have yielded increases in student learning in a variety of domains (Lewis & Mayer, 1987; see Van Meter & Garner, 2005 for a review). Constructing diagrams has the potential to help students learn to coordinate and integrate text with visual representations (Ainsworth, 2006; Easterday, Aleven, & Scheines, 2007).

In the present study, we directly target low-achieving pre-algebra students to determine whether guided experience constructing one type of diagram from simple story problems facilitates broader use of diagrams for more complex problems. We also aim to investigate the mechanism underlying any resulting benefit by testing students on more than one type of diagram. If improvement is due to increased familiarity with the type of diagram used during instruction, benefits should manifest as increased willingness to attempt familiar-looking problems and improved performance on those items. However, if, as intended, the instruction provides students with the necessary tools for mapping between story problems and diagrams, benefits should be more likely to transfer to the other type of diagram as well.

Methods

Participants

Participating in this study were four classrooms of non-honors Pre-Algebra students (N = 73 eighth grade students; typically age 13) from a school in which only 8.5% of students reach the required state level of math proficiency. Eighty-nine percent of students at the participating school were economically disadvantaged; the ethnic breakdown of the school was approximately 95% Hispanic, 5% African-American, and < 1% Caucasian or other. Three additional students participated in the study, but were excluded because they were not given the correct version of the posttest.

All four classrooms used the Bridge to Algebra Cognitive Tutor curriculum. The Cognitive Tutor is a computer-based intelligent tutoring system which provides on-demand, step-specific help at any point in the problem-solving process and feedback on errors (Koedinger, Anderson, Hadley, & Mark, 1997).

Procedure

Prior to beginning the first Tutor unit in the curriculum (approximately three weeks after the beginning of the school year), participants completed a written pretest on which they were asked to solve algebraic story problems in three presentation formats: story alone, story with a vertical diagram, or story with a horizontal diagram. The test included six problem situations, each representing one of two underlying algebraic equations: 1) $ax + b = c$, and 2) $x + (x + a) + (x + b) = c$. Both equation types were represented in each of the three presentation formats on every test (See Figure 1 for examples of each presentation format for the second equation). There were three counterbalanced forms of the test, such that each problem

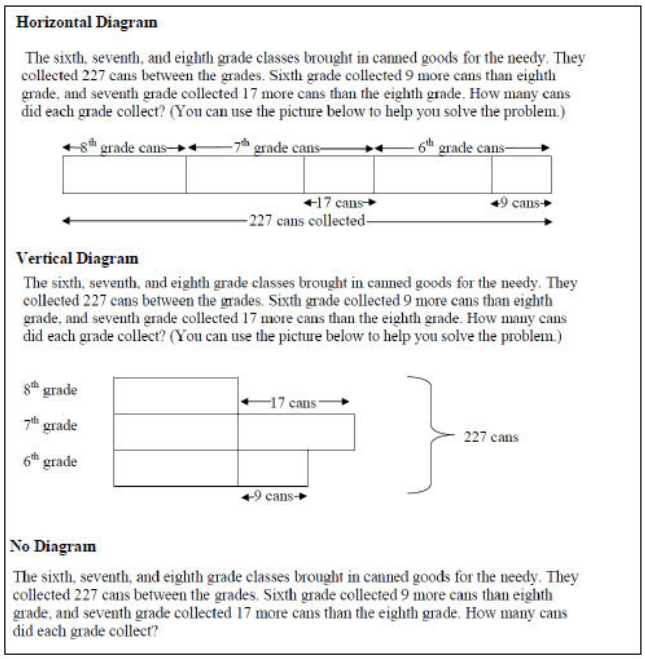


Figure 1: Sample problem in each presentation format

situation appeared in each of the three presentation formats on one of the three test versions. After completing the pretest, students began the interactive Tutor unit on Picture Algebra, in which they created, labeled, and used vertical diagrams to solve simple story problems using multiplication, addition, or subtraction The vertical training problems were simpler than those included in the test, in that they required the student to manipulate only two components compared with three or more components in test problems; thus, all test problems were transfer problems (see Figure 2 for examples of a problem in the Picture Algebra unit). Each student completed the unit at his or her own pace. As each student completed the unit, he or she was given a posttest by the classroom teacher; students were given the same version of the test that they had taken at pretest.

Results

Pretest and posttest scores for each of the three presentation types can be found in Figure 3. A 3 (presentation format: vertical diagram, horizontal diagram,

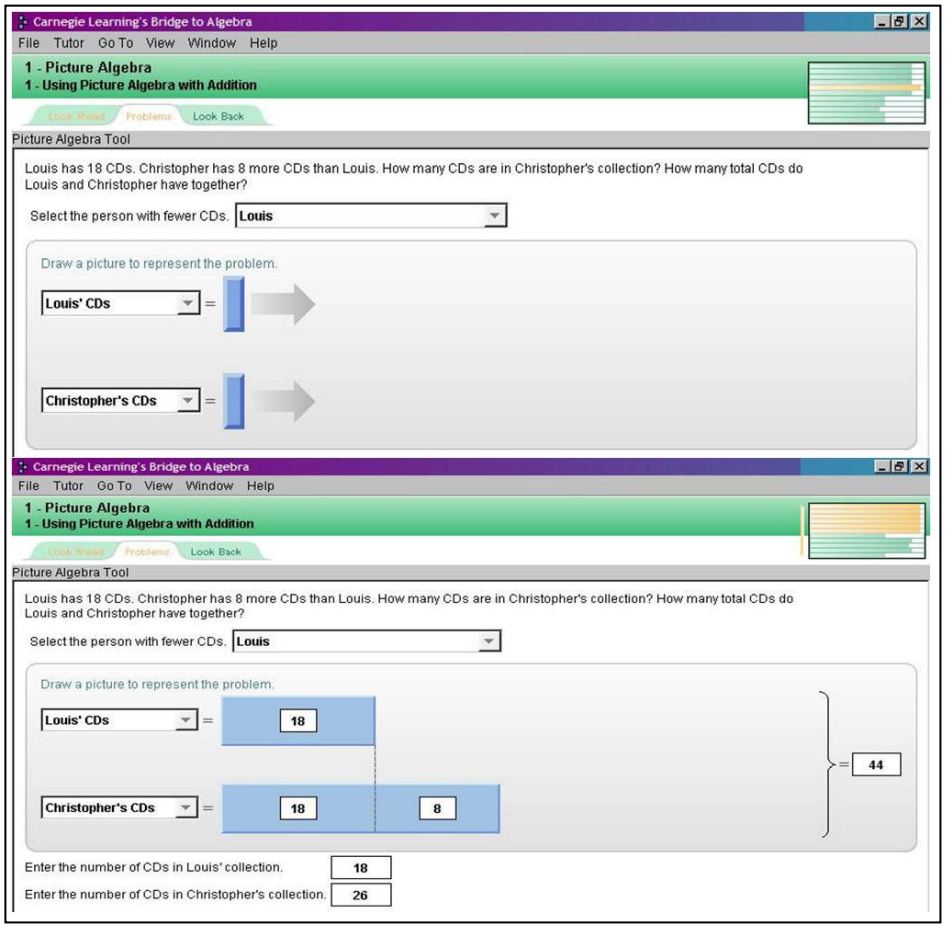


Figure 2: Screenshots from a Picture Algebra problem in the Bridge to Algebra Tutor. Students stretch the blocks out to represent the number of CDs owned by Louis and Christopher. Christopher's CDs are represented using the same sized box as for Louis and additional length to represent the 8 extra CDs.

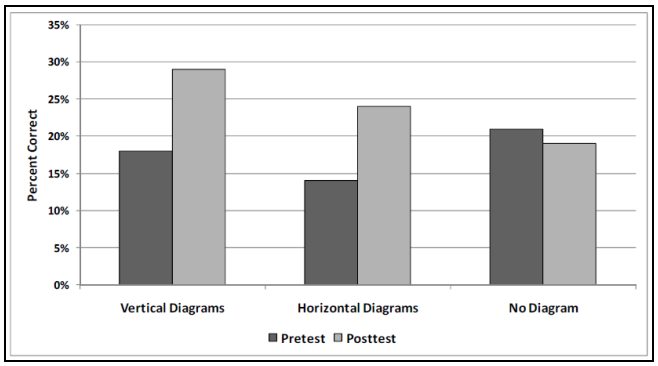


Figure 3: Percent correct at pretest and posttest for each presentation format.

no diagram) x 2 (test time: pretest vs. posttest) repeated measures ANOVA on the percent of problems answered correctly yielded a main effect of test time, $F(1, 72) = 5.10$, $p < .05$, $\eta_p^2 = .07$. There was no main effect of presentation format $F(2, 144) = 1.56$, *ns*. However, the interaction between presentation format and test time was significant, $F(2, 144) = 5.25$, $p < .01$, $\eta_p^2 = .07$. To interpret this interaction, we conducted follow-up repeated measures ANOVAs on presentation format, separately for pretest and posttest scores. No significant differences among presentation types were found at pretest $F(2, 144) = 1.91$, *ns*. In contrast, at posttest, a main effect of presentation format was found, $F(2, 144) = 3.95$, $p < .05$, $\eta_p^2 = .05$. Follow-up pairwise comparisons with Bonferroni correction indicated that students scored higher on problems with vertical diagrams than those with no diagrams ($p < .05$). No differences were found between scores on horizontal diagrams problems compared with either of the other presentation formats.

Students solved more vertical problems correctly on the posttest after receiving training on creating and using simpler versions of those diagrams ($t(72) = 2.83$, $p< .01$). Students also improved on horizontal diagrams problems after vertical diagram training ($t(72) = 2.80$, $p< .01$), but no improvement was found on problems that did not contain diagrams ($t(72) < 1$, *ns*). A repeated measures ANOVA on the amount of improvement shown yielded a main effect of presentation type, $F(2, 144) = 5.7$, $p < .05$, $\eta_p^2 = .07$. Follow-up pairwise comparisons with Bonferroni correction indicated that students improved more on problems with vertical diagrams (11%) and horizontal diagrams (10%) than those with no diagrams (both p's < .05). No difference was found between improvement on vertical and horizontal problems.

Error Analysis: The nature of the improvement

To investigate the source of this improvement in scores on the diagrams problems, we conducted a qualitative analysis of the types of errors made by students while solving each type of problem on the pretest and posttest. Four pretest and thirteen posttest problem attempts were found in which students drew diagrams to help them solve the no diagrams problems; these incidences were thus excluded from the subsequent response and error analysis.

One possible source of improvement was that exposure to the diagrams could have made students more comfortable with, and thus more likely to attempt, diagrams problems on the posttest, leading to a higher possible number of diagrams problems answered correctly. To examine this hypothesis, we coded whether students attempted to solve each problem or if they failed to respond to it. We then computed the percentage of each type of problem that was attempted at pretest and posttest. As can be seen in Figure 4, students attempted more problems of each of the three formats at posttest compared with the pretest. This suggests that a higher response rate for diagrams problems is not a viable explanation for the improvement.

A second, more plausible hypothesis was that students better understood the mapping between the diagrams and the stories as a result of instruction. Given that their experience with the Picture Algebra unit trained them to build components of diagrams to represent the information in a story problem, this hypothesis seemed plausible. This improved understanding should lead students to make fewer errors in which they demonstrate failure to make sense of the information in the problem. To test this, we coded student responses in terms of whether they were correct, contained an arithmetic error (e.g., adding 4 + 6 and getting 9), or contained a conceptual error—one that indicated a misunderstanding of the role of the numbers in the problem (e.g., for the problem pictured in Figure 1, solving the problem as if 7th graders collected 17 fewer cans than 8th graders, instead of 17 more cans). In previous work, adding diagrams to story problems was shown to prevent older, high-achieving students from making common conceptual errors when solving the problems (Booth & Koedinger, 2007), but younger and lower-achieving students did not receive this benefit. Results from the present study indicate that after instruction it appears that fewer horizontal and vertical diagrams problem attempts contained conceptual errors than did at pretest whereas no reduction was apparent for problems without diagrams (see Figure 5).

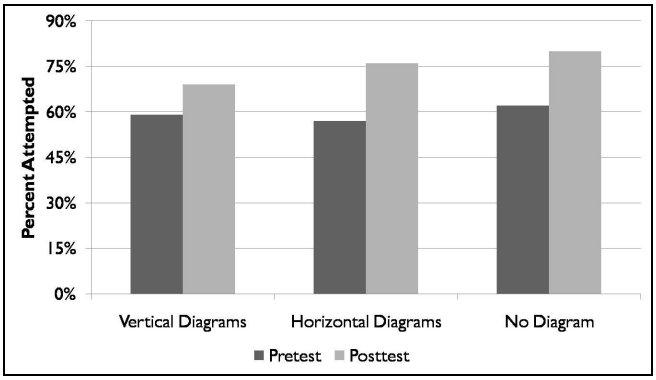


Figure 4: Percent of problems attempted at pretest and posttest for each presentation format.

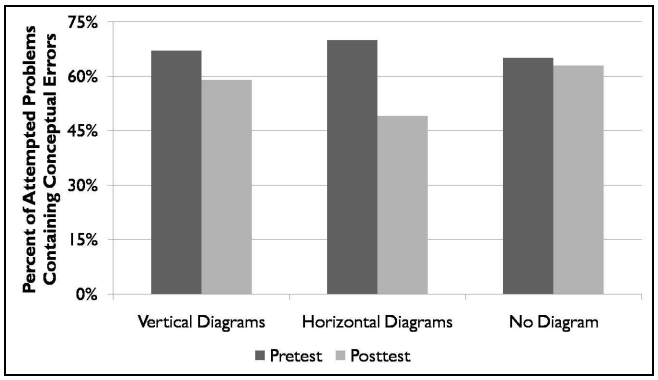


Figure 5: Percent of pretest and posttest problem attempts for each presentation format containing a conceptual error

Discussion

Results from the present study replicated the previous finding that diagrams are not inherently beneficial for solving algebraic word problems. At pretest, no diagrammatic advantage was found for low-achieving pre-algebra students. After instruction, however, the same students experienced a diagrammatic advantage, and there was evidence of transfer of instruction benefits to the non-instructed diagram format. Students made fewer conceptual errors with both types of diagrams after instruction, suggesting that increasing students' skill at mapping between story problems and supplemental diagrams can afford low-achieving students the same benefits as those enjoyed by their higher-achieving peers. Interestingly, no improvement was found for the no diagram condition, suggesting that students' general word problem solving abilities did not increase. Presumably, if students had drawn diagrams to help them solve those problems, as they did in their training, they would have had greater success.

One specific mechanism of the diagrammatic advantage is that it has been shown to increase the likelihood that students will achieve a conceptually sound understanding of a problem, and avoid common conceptually flawed solution paths (Booth & Koedinger, 2007). Consistent with this finding, results from the present study indicated that students reduced the number of conceptual errors made at posttest on transfer problems with diagrams compared to those without diagrams. The likely mechanism by which the diagrammatic advantage emerges is through increased experience coordinating information from two sources, which helps students learn to create appropriate links between the information (e.g., the components of the diagram with the corresponding components of the text). This general ability enables successful mapping between sources in new diagrammatic problems, yielding a sound representation of the overall problem, which leads to fewer critical conceptual mistakes in solution. The process of constructing the diagram facilitated this process by forcing students to make connections explicit; this is consistent with Van Meter & Garner's (2005) assertion that the benefit of diagram construction is that it necessitates integration between text and diagram. The Picture Algebra lesson provided practice opportunities with feedback for students to gain the general mapping ability, which they were then able to apply successfully to the test problems. Further research is needed to determine whether and how developing students naturally acquire this skill, whether through cognitive maturation (and perhaps the development of more formal reasoning skills), through certain types of experiences that become more prevalent as children age, or some combination thereof.

Results from this study suggest that, while low-achieving students have difficulty interpreting diagrams and using them to their benefit when solving problems, it may not be necessary for students to have specific instruction or experience with a given type of diagram in order to use it effectively. Rather, acquiring more general diagram-parsing skills that facilitate mapping between text and any sort of diagram may be more beneficial. The instruction presented in this study did not just increase comprehension of more complex vertical diagrams; it helped cultivate a broader skill which allowed them to also use complex horizontal diagrams to solve the problem. Future research should investigate the nature of broad diagram-parsing skills and determine how best to teach them to help all students benefit from diagrams and other external representations of instructional information in Algebra.

Acknowledgments

Funding for this research was provided by the National Science Foundation Grant Number SBE-0354420 to the Pittsburgh Science of Learning Center (PSLC, http://www.learnlab.org). Portions of this work were previously presented at the annual meeting of the American Educational Research Association in San Diego, CA. Thanks are also due to the Los Angeles Unified School District for allowing us to collect data in their classrooms.

References

Ainsworth, S. (2006). DeFT: A conceptual framework for considering learning with multiple representations. *Learning and Instruction, 16*(3), 183-198.

Ainsworth, S. E., & Loizou, A.T. (2003). The effects of self-explaining when learning with text or diagrams. *Cognitive Science, 27,* 669-681.

Booth, J.L., & Koedinger, K. (2007, March). Are diagrams always helpful tools? The effect of presentation format on students' solutions of algebra problems. Poster presented at the meeting of the Society for Research in Child Development in Boston, MA.

Chandler, P. & Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition and Instruction, 8,* 293-332.

Cummins, D. D., Kintsch, W., Reusser, K., & Weimer, R. (1988). The role of understanding in solving algebra word problems. *Cognitive Psychology, 20*, 405-438.

Easterday, M., Aleven, V., & Scheines, R. (2007). 'Tis better to construct than to receive? The effects of

diagramming tools on causal reasoning. In R. Luckin, K. Koedinger, & J. Greer (Eds.), *Proceedings of the 13th International Conference on Artificial Intelligence in Education* (pp. 93-100). Amsterdam, the Netherlands: IOS Press.

Ericsson, K. A. & Simon, H. A. (1980). Verbal reports as data. *Psychological Review, 87,* 215-251.

Fueyo, V., & Bushell, D. (1998). Using number-line procedures and peer tutoring to improve the mathematics computation of low-performing first graders. *Journal of Applied Behavioral Analysis, 31,* 417-430.

Greeno, J. G., & Hall, R. P. (1997, January). Practicing representation: Learning with and about representational forms. *Phi Delta Kappan, 78*, 361–367.

Hiebert, J. & Carpenter, T.P. (1992). Learning and teaching with understanding. In D.A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 65-97). New York: Macmillan.

Koedinger, K.R., Alibali, M. W., & Nathan, M.J. (2008). Trade-offs between grounded and abstract representations: Evidence from algebra problem solving. Cognitive *Science: A Multidisciplinary Journal, 32:2,* 366-397.

Koedinger, K.R., Anderson, J.R., Hadley, W.H., & Mark, M.A. (1997). Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education, 8*, 30-43.

Koedinger, K. R., & Nathan, M. J. (2004). The real story behind story problems: Effects of representation on quantitative reasoning. *Journal of the Learning Sciences, 13*, 129 -164.

Koedinger, K.R. & Terao, A. (2002). A cognitive task analysis of using pictures to support pre-algebraic reasoning. In C.D. Schunn & W. Gray (Eds.), *Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science Society*. Mawah, NJ: Lawrence Erlbaum Associates, Inc.

Larkin, J.H., & Simon, H.A. (1987). Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science, 11*, 65-99.

Lewis, A. B., & Mayer, R. E. (1987). Students' misconceptions of relational statements in arithmetic word problems. Journal of Educational Psychology, 79, 363-371.

Mayer, R. E. (1989). Systematic thinking fostered by illustrations in scientific text. *Journal of Educational Psychology, 81,* 240-246.

Mayer, R. (2005). Cognitive theory of multimedia learning. In R. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 31-48). Cambridge: Cambridge University Press.

Mayer, R. E. (2009). *Multimedia learning* (2nd ed). New York: Cambridge University Press.

National Council of Teachers of Mathematics (NCTM) (2000). *Principles and standards for school mathematics.* Reston, VA: Author.

Rittle-Johnson, B., & Koedinger, K. (2001). Using cognitive models to guide instructional design: The case of fraction division. In J. Moore & K. Stenning (Eds.), *Proceedings of the 23rd Annual Conference of the Cognitive Science Society,* (pp. 857-862). Mahwah, NJ: Erlbaum.

Seeger, F. (1998). Representations in the mathematics classroom: Reflections and constructions. In F. Seeger, U. Waschescio, & J. Voight (Eds.), *The culture of the mathematics classroom* (pp. 308-343). New York: Cambridge University Press.

Sowell, E.J. (1989). Effect of manipulative materials in mathematics instruction. *Journal for Research in Mathematics Education, 20,* 498-505.

Uttal, D.H., Scudder, K.V., & DeLoache, J.S. (1997). Manipulatives as symbols: A new perspective on the use of concrete objects to teach mathematics. *Journal of Applied Developmental Psychology, 18,* 37-54.

VanMeter, P., & Garner, J. (2005). The promise and practice of learner-generated drawing: Literature review and synthesis. *Educational Psychology Review, 17*(4), 285-325.

Why do four- year- olds show poor cross-modal transfer between haptic and vision?

Hilary Kalagher (hkalaghe@indiana.edu)
Department of Psychological and Brain Sciences, 1101 E 10th Street
Bloomington, IN 47405 USA

Susan S. Jones (jones1@indiana.edu)
Department of Psychological and Brain Sciences, 1101 E 10th Street
Bloomington, IN 47405 USA

Abstract

Four year olds have difficulty transferring information from the haptic to the visual modality. This difficulty may reflect qualitative differences in haptic and visual object representations or children's inability to obtain the same kinds of perceptual information in the two modalities. Twenty 4-year-olds explored novel objects either haptically or visually, then haptically chose a match from among three test objects that each matched the exemplar on one perceptual dimension. Children chose shape-based matches after visually exploring category exemplars. However, after haptic exemplar exploration, children were equally likely to pick a shape- or texture-based match. Analysis of children's hand movements during haptic exploration showed that certain movements reliably predicted shape-based matches. This finding suggests that children have difficulties in cross-modal transfer because their haptic exploration is not driven by a top-down perceptual focus as it is in adults.

Keywords: Haptic Perception; Cross-modal Transfer; Development

Introduction

The use of perceptual information obtained in one modality - for example, haptics – for use in a task in another modality - for example, a visual task - involves cross-modal information transfer. Cross-modal transfer is important because it allows for inter-sensory predictions. For example, being able to anticipate what an object will look like given that you have only touched it allows for efficient and quick interactions with the world. Adults appear to have no difficulty transferring novel information gathered in one perceptual modality for use in a second modality (e.g., Abravanel, 1971, 1973; Easton, Srinivas, & Greene, 1997; Reales & Ballesteros, 1999). However, there are reports that preschool-aged children have difficulty in cross-modal transfer. In particular, children perform poorly in object recognition tasks requiring the transfer of information from haptics to vision. Two explanations for these findings have been proposed. The first proposal is that there are qualitative differences between the representations that children form from visual and haptic experience, so that translation between the two modalities is hampered. The second proposal is simply that young children have poor haptic perception. The present study explores these two possibilities.

Qualitatively different representations

Bushnell and Baxt (1999) used real-world familiar and novel objects to test 5-year-olds in object recognition requiring either intra- or cross-modal use of haptic or visual information. The children did well in object recognition in intra-modal tasks with both familiar and novel objects, and in cross- modal tasks with familiar objects. Children's performance was markedly poorer, however, with novel objects. Bushnell and Baxt (1999) suggested that "hand-mages" – representations formed entirely from haptic exploration – differ importantly from visual images ("eye-mages"). They proposed that attention during haptic exploration might be focused on material-based properties (texture, mass, rigidity) whereas in vision attention is focused more on shape and color. These differences in perceptual focus would then presumably lead to qualitatively different representations and therefore poorer performance in cross-modal tasks as compared to intra-modal tasks.

Kalagher and Jones (2010) used a novel name extension tasks to test Bushnell and Baxt's (1999) hypothesis that representations formed through experiences in different modalities are qualitatively different. In the standard version of this task (c.f., Landau, Smith, & Jones, 1988), children are visually presented with a novel object (the exemplar) and told its novel name. Children are then visually presented with three test objects; each matches the exemplar on one perceptual dimension – color, texture, or shape - and differs from the other two test objects on the other perceptual dimensions. Children are asked to indicate the object that has the same novel name as the exemplar. Past experiments have shown that children by 2 years of age typically choose shape-based matches predominately over texture- or color-based matches (e.g., Smith, Jones, Landau, Gershkoff-Stowe, & Samuelson, 2002). This "shape-bias" is thought to reflect an acquired attentional bias. Kalagher and Jones (2010) tested children 2 ½ to 5 years of age and adults in a modified version of this task. Children in their experiment explored exemplar objects either haptically or visually. Test objects were then presented visually. All ages in that experiment chose a preponderance of same-shape matches after visual exemplar exploration. However, only 5-year-

olds and adults made shape-based matches after haptic exploration. Children younger than 5- years- old chose test objects at random. Importantly, when older children and adults did choose matches systematically after haptic exploration, they did not make texture matches as Bushnell and Baxt (1999) might predict. Instead, they chose shape-based object matches just as they did after visual exploration. This finding suggests that representations in the two perceptual modalities are not qualitatively different.

Haptic perception

The mature haptic perceptual abilities of adults have been studied extensively. Lederman and Klatzky (1987) studied the hand and finger movements of adults who were attempting to extract information about specific object properties using haptics alone. The researches identified a number of stereotyped manual movements that they called "exploratory procedures" (EPs). EPs are thought to be driven primarily by top-down processes such as task goals but also, to a lesser extent, by bottom-up sensory information. What then is known about young children's haptic perception? A number of studies have reported poor haptic perception in young children (e.g., Abravanel, 1972; Rose, Blank, & Bridger, 1972). For example, Milner and Bryan (1970) asked 5 to 7 year olds to make same/different judgments about object shape in both intra- and cross-modal conditions. The magnitude of the improvements made between 5 and 7 years of age were comparable in both the intra- and cross-modal conditions. The authors therefore concluded that the developmental change was due to gains in children's haptic abilities. However, other researchers have found that 5-year-old children's haptic perception is actually quite advanced, at least when they are asked to explore familiar objects (e.g., Bigelow, 1981). Kalagher and Jones (2010) found mature haptic exploratory behaviors in 5- year- olds but fewer such behaviors in younger children. Their analysis of children's hand movements during haptic exploration showed that certain movements reliably predicted subsequent shape- or texture-based matches: however, children younger than 5 produced these hand movements at very low frequencies.

In summary, young children's difficulties with haptic-to-vision information transfer appeared in a previous study to stem from their failure to execute mature hand movements rather than from qualitative differences of object representations in the two modalities. In the present study, we asked whether additional tests of children's haptic exploratory abilities would point to the same conclusion. More specifically, children in the present study participated in two new conditions: (1) a visual exemplar exploration – to – haptic recognition condition; and (2) a haptic exemplar exploration – to – haptic recognition condition. The visual exemplar exploration – to – haptic recognition condition, like the haptic exemplar exploration - to – visual recognition condition in the previous research, required children to transfer information across perceptual modalities. However, because children's attention during visual exploration is consistently biased towards object shape (e.g., Smith et al., 2002), we speculated that children might make more systematic choices using information from vision to make haptic object matches than they had made using information from haptic to make visual matches. In the former case, they would know what they were looking for: that is, their visual explorations would lead them to focus on object shape. A finding that children did not make systematic shape-based matches in this condition would be further evidence against the idea that representations in the two perceptual modalities are qualitatively different and do not translate.

The haptic exemplar exploration – to – haptic recognition condition eliminated the need to transfer perceptual information across modalities. Thus, this condition tested children's haptic perception only. A finding that children made systematic matches in this condition would indicate that their haptic perceptual abilities were good. Systematic texture matches would support Bushnell and Baxt's (1999) idea of "hand-mages". A failure to match objects systematically would suggest that neither shape nor texture had been perceived well enough during object exploration for subsequent use in object recognition in the haptic mode.

Methods

Twenty four- year- old children (range = 46.8 to 56.1 months; *Mean* = 51.65; 10 males) participated in the study. Participants reflected the local community in social class, ethnicity, and racial identity: almost all participants were from white, middle class families.

The stimuli consisted of 16 object sets, each with one exemplar, and three test objects. Exemplars and test objects were 3-dimensional, novel objects constructed from a variety of materials including wood, clay, and cloth. Sizes ranged from 7 to 17 cm. Colors, textures, shapes and masses were widely varied. Each of the test objects shared a different attribute –its color, texture, or shape – with the exemplar, and differed from the exemplar and the other two test objects on the other two dimensions (See Figure 1 for a sample stimulus set).

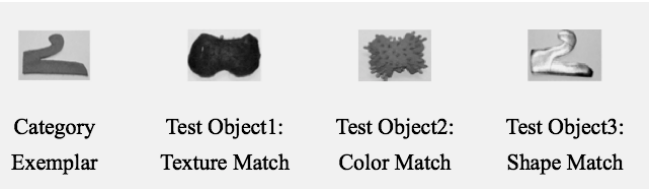


Figure 1: Sample stimulus set: 1 exemplar object, and 3 test objects, each matching the exemplar on 1 dimension – shape, texture, color – and differing from the exemplar and each other on the other 2 dimensions.

All participants completed two blocks of trials, each consisting of three warm-up trials and eight test trials. In one block of trials, children were limited to visual

exploration of the exemplars ("Visual Exemplar Exploration" condition). In the other block of trials, children were limited to haptic exploration of the exemplars ("Haptic Exemplar Exploration" condition). The order of conditions was counterbalanced across participants. The 16 stimulus sets were divided into two groups and the stimulus groups were counterbalanced within conditions. Thus, half of the participants saw Stimulus Group 1 in the Visual Exemplar Exploration condition, and the other half of the participants saw Stimulus Group 2 in that condition.

Each participant was seated at a table next to his or her parent and across from the Experimenter who explained that they were going to play a "matching game". The procedure began with three warm-up trials to ensure that participants understood the task. Warm-up trials differed between the two conditions. In each warm-up trial in the Visual Exemplar Exploration condition, participants were simply handed a familiar object and told its name (e.g., "Look, here is a spoon). After three seconds, the Experimenter retrieved the object. In each warm-up trial in the Haptic Exemplar Exploration condition, participants placed their hands and forearms inside a box; a piece of cloth was pulled over their arms to prevent participants from seeing inside the box. The Experimenter put a familiar object into the hands of the participant within the box, identified it by name, and asked the participant whether he or she could feel it (e.g., "This is a spoon. Can you feel the spoon?"). For the test trials in both conditions, participants had their hands and forearms inside the box and a piece of cloth draped over their arms to prevent them from seeing inside the box. On each trial, three test objects were placed inside the box (e.g., a cup, a comb, and a spoon) and the child was asked to pull out the test object with the same name as the exemplar (e.g., "Can you find me the spoon?).

Test trials followed warm-up trials immediately and were structured in the same way: participants were shown or handed the exemplar from one object set at a time and told its novel name (e.g., "This is a teeka") then asked to find a haptic match for the exemplar (e.g., "Can you find me another teeka?") from among the three test items inside the box. Children were given a sticker after each trial regardless of which choice they made. The experiment was digitally recorded, and records were later scored for the test objects – shape match, texture match, or color match – chosen on each trial.

The recordings were also coded for the children's hand movements while exploring test objects in the Visual Exemplar Exploration condition, and category exemplars and test objects in the Haptic Exemplar Exploration condition.

Results

Object Recognition To determine whether visual or haptic exemplar exploration led predominantly to shape or texture matches, the proportions of shape and texture matches in each condition were calculated. This resulted in 4 categories of scores: (1) visual exemplar exploration resulting in shape match (V->SH), (2) visual exemplar exploration resulting in texture match (V->TX), (3) haptic exemplar exploration resulting in shape match (H->SH), and (4) haptic exemplar exploration resulting in texture match (H->TX). The mean proportions of children's scores in each category can be found in Table 1.

Table 1. Means and Standard deviations for proportions of shape- and texture- based matching in the Visual Exemplar Exploration and the Haptic Exemplar Exploration conditions.

	Mean	SD
V-SH	0.54	0.27
V-TX	0.32	0.22
H-SH	0.42	0.14
H-TX	0.39	0.13

The proportions were first entered into a 2 Order (Visual Exemplar Exploration first or second) x 2 Gender (male/female) x 2 (Exemplar Exploratory Modality: Haptic/Visual) x 2 (Match Type: Shape/Texture) mixed analysis of variance. There were no between subjects main effects for either Gender ($F_{(1,16)} = 2.67$, $p = ns$) or Order ($F_{(1,16)} = .17$, $p = ns$). There was a significant main effect of Match Type ($F_{(1,16)} = 4.5$, $p < .05$) with more shape-based matches exceeding texture-based matches. We did not find a main effect of Exemplar Exploratory Modality ($F_{(1,16)} = .79$, $p = ns$). Figure 2 graphs the marginally significant Exploratory Modality by Match Type interaction and illustrates the fact that shape choices dominated choices after visual exploration ($F_{(1,16)} = 3.82$, $p = .06$).

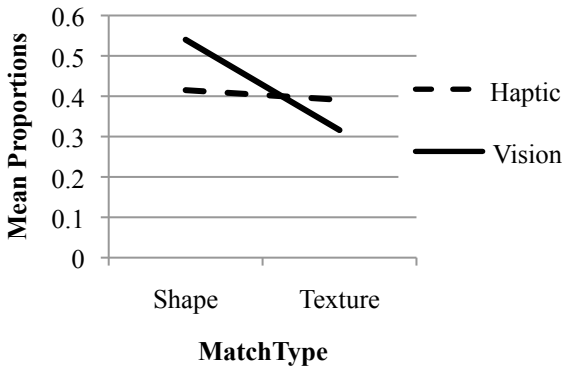


Figure 2. Match Type by Exploration Modality interaction

The proportions of shape and texture choices shown in Table 1 were also compared to chance (.33) using one-sample t-tests. Children chose same-shape matches at above chance levels in both conditions (H->SH: $t_{(19)} = 2.59$, $p < .05$; V->SH: $t_{(19)} = 3.46$, $p < .05$). Their texture-based matches were marginally above chance following haptic exemplar exploration of the exemplar ($t_{(19)} = 2.05$, $p =$

.054), but not following visual exploration ($t_{(19)} = -.34$, p = *ns*).

In sum: children showed the previously well-documented bias to preferentially attend to shape in object matching (e.g., Smith et al., 2002) in the Visual Exemplar Exploration condition. However, children in the Haptic Exemplar Exploration condition were equally likely to pick a shape or texture match. Thus, after visual exemplar exploration, children systematically chose shape-based matches suggesting that the representations they formed focused predominately on shape. After haptic exemplar exploration we do not see a similar systematic preference and therefore cannot claim that children's representations formed through haptic experiences are or are not qualitatively different from their representations formed through visual experiences.

We next examined children's hand movements during the test phase in the Visual Exemplar Exploration condition and during both the exemplar exploration and the test phase portions of the Haptic Exemplar Exploration condition.

Hand movements Initially, we attempted to use the taxonomy of exploratory hand movements developed by Lederman and Klatzky (1987) to code children's hand movements while exploring objects in both conditions. However, children in the present study did not produce these movements. Therefore, 5 categories of manual exploratory behavior identified by Kalagher & Jones (2010) in the same age group were used instead. The categories are: (1) "sequential finger movements" (rotating the object around only with fingertips), (2) "fingers palpating" (fingers palpating/ squeezing the object), (3) "static fingers" (fingers placed on object but not moving), (4) "hand grasping" (grasping the object with one hand), and (5) "hand press" (pressing the object between both hands with fingers outstretched).

We also coded instances of children's verbalization specifically recording shape-related verbalizations (e.g., "This feels like the letter 'Z'") and texture-related verbalizations (e.g., "This feels fuzzy"). However, such verbalizations were rare, occurring in fewer than 15% of trials, and were therefore not analyzed.

Hand movements were clearly visible in 108 exemplar exploration trials and 101 test trials in the Haptic Exemplar Exploration condition (H.E.E. and H.E.T.O., respectively) and in 97 test trials in the Visual Exemplar Exploration condition (V.E.E.). Figure 3 shows the frequencies of the 5 kinds of hand movements produced by children in each of these kinds of trials.

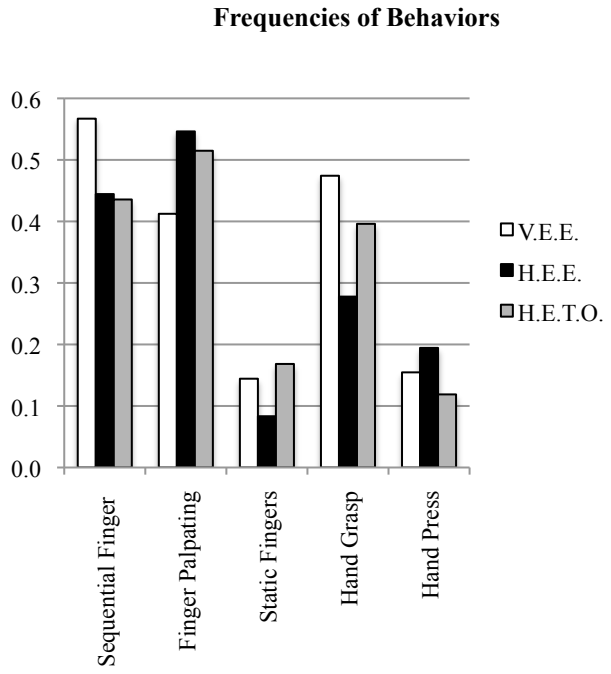


Figure 3. Frequencies of exploratory hand movement behaviors.

We then asked whether any of these five behaviors predicted whether children would make a shape-based, texture-based, or color-based (i.e., random) object match. We used multinomial logistic regression to address this question. Multinomial logistic regression is a generalization of the binomial regression and is useful when the dependent variable has more than two discrete choices. In a multinomial logistic regression model, the estimates for the parameter can be identified compared to a baseline category. For our analyses, we used a dependent variable SCORE (1= shape match, 2 = texture match, and 3 = color match), and independent predictor variables of hand movement patterns. A SCORE value of 3 was specified as the baseline category. The test then estimated the effects of the independent variables on choosing texture or shape matches over making a color-based match.

When the multinomial logistic regression test was carried out on the data the Likelihood Ratio (LR) chi-square test that at least one of the predictors' regression coefficients was not equal to zero yielded significant results, $\chi^2_{(2,14)} = 118.5$, $p < .0001$. This outcome indicated that particular hand movement behaviors affected subsequent matches (i.e., the participant's score). From the results of the multinomial logistic regression analysis, 2 main effects were significantly predictive of SCORE. The significant main effects were: *sequential finger movements* ($\chi^2_{(2,14)} = 48.22$, $p < .0001$), and *hand press* ($\chi^2_{(2,14)} = 24.4$, $p < .0001$). Further chi- square analyses showed that children's use of *sequential finger movements* was predictive of later shape

matches, $\chi^2_{(2,14)} = 73.34$, $p < .0001$, while the absence of the hand movement pattern *hand press* was also predictive of later shape matches, $\chi^2_{(2,14)} = 30.36$, $p < .0001$.

The fact that two specific hand movements (i.e., *sequential finger movements*, and *hand press*) during haptic exploration predicted children's subsequent choice of a same-shape match suggests that children were able to obtain shape information and use that information intra and inter-modally.

Item Analysis Individual items were classified into three categories by the extent to which each was matched predominately by shape, texture, or color. These categories and their criteria are as follows: (1) "dominant match" criterion: one feature (shape or texture) is matched more than twice as often as the second more frequently used feature; (2) "selective match" criterion: item is selectively matched but differently by different children (both shape and texture matches separately are chosen at least twice as often as color); (3) "random matches" – remainder. Figure 4 displays the results of applying these criteria to the objects in both conditions.

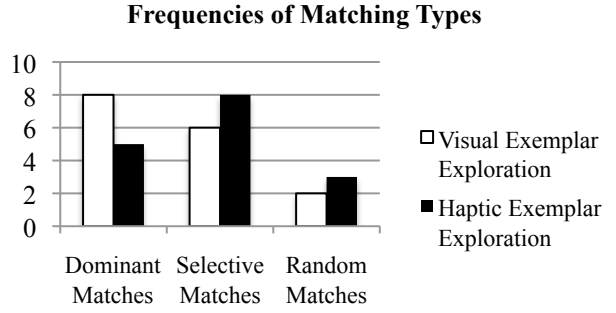


Figure 4. Frequencies of item analysis match types for the Visual Exemplar Exploration condition (white bar) and the Haptic Exemplar Exploration condition (black bar).

When the 16 category exemplars were explored visually, eight objects were subsequently matched consistently and by different children on one dominant perceptual dimension (shape-6 and texture-2). When those same objects were explored haptically, only 5 were matched on one dominant dimension (shape-3 and texture-2). A chi square analysis found no difference between conditions in the kinds of dominant matches (shape or texture) that children made ($\chi^2_{(1)} = .33$, $p = ns$).

Discussion

Our primary interest was in whether restricting exemplar exploration to either vision or haptics would have consequences for the kinds of test object that children chose in the haptic modality to match the exemplar objects. When children explored category exemplars visually, they were more likely to pick a same-shape match than if they had explored the exemplar haptically. This finding is consistent both with Bushnell and Baxt's (1999) suggestion that "eye-mages" represent different perceptual information than "hand-mages", and with the abundant evidence that young children's attentional focus on shape in visual object perception leads to a predominance of shape-based object matches (e.g., Smith, et al., 2002).

When category exemplars were explored haptically, children were equally likely to pick a shape- match or a texture- match. This finding is not consistent with Bushnell and Baxt's (1999) idea that "hand-mages" formed from haptic input predominantly represent object texture, mass, and rigidity. Instead, children in this study appeared to be matching each item on whatever perceptual information gained from exploration of each exemplar object was most salient to them. They do not appear to be using a top-down perceptual focus that would allow them to match objects systematically by either shape or texture.

Overall, we did not find compelling evidence of qualitative differences in the object representations from visual and haptic inputs. Instead, children's use of representations formed through haptic experience seemed to be affected by the most salient properties of the exemplar object (bottom- up), rather than by a particular perceptual focus such as the "shape bias" seen in vision (top-down).

The item analysis provides further support for this last point. When comparing children's consistency in making shape or texture matches in the Visual Exemplar Exploration condition to the children's consistency in making shape or texture matches in the Haptic Exemplar Exploration condition, we did not find a reliable difference in the number or kinds of *dominant matches* made. This result suggests that representations from input in the two modalities are not qualitatively different.

A secondary goal of this experiment was to further examine the status of young children's haptic abilities. Examination of haptic exploratory behavior indicated that when children executed *sequential finger movements* it was likely that they would make a subsequent shape-based match. This finding replicated the results Kalagher and Jones (2010). However, we found no parallel relation between particular hand movements during haptic exploration and children's later choices of texture matches. Again this finding suggests that children's haptic exploratory behavior is not guided by a particular perceptual focus or goal.

Interestingly, the present results show that 4- year -olds can haptically obtain shape information when they have a clear idea of what they are looking for as reflected in the predominance of shape matches in the Visual Exemplar Exploration condition. This predominance suggests that when children visually explored exemplar objects, they formed representations that contained and perhaps emphasized shape information. Guided by these representations, children's haptic abilities were good enough to obtain the shape information needed for a same-shape match. This finding is consistent with previous reports that

found that young children have good haptic perception of familiar objects (Bigelow, 1981; Bushnell & Baxt, 1999; Morrongiello, Humphrey, Timney, Choi, & Rocca, 1994).

In summary, 4 year olds' representations of novel categories experienced haptically do not appear to be focused on either texture or shape. However, the object representations constructed from haptic perceptual input appear to be good enough to support object matches on whichever perceptual dimension is most salient. Thus, the present results indicate that children younger than 5 years of age have functional haptic abilities.

Although haptic perceptual exploration did not appear to be bias towards one kind of perceptual information over another, visual experience of novel categories appeared in this study, as in many previous studies, to lead to the formation of representations focused of shape. A new finding in the present study is evidence that representations built from visual input can transfer, complete with their focus on shape, into the haptic mode. Specifically, the present findings of a predominance of same-shape object matches in the haptic modality given only visual experience of the exemplar object suggests that representations of that visual experience guided the haptic identification of a matching objet. Thus, it appears that the shape bias in visual object matching remained intact during the transfer of perceptual information about exemplar objects from the visual into the haptic realm.

References

Abravanel, E. (1971). Active Detection of Solid-Shape Information by Touch and Vision. *Perception & Psychophysics, 10*(5), 358-360.

Abravanel, E. (1972). Short-Term Memory for Shape Information Processed Intramodally and Intermodally at 3 Ages. *Perceptual and Motor Skills, 35*(2), 419-425.

Abravanel, E. (1973). Division of Labor between Hand and Eye When Perceiving Shape. *Neuropsychologia, 11*(7), 207-211.

Bigelow, A. E. (1981). Children's Tactile Identification of Miniaturized Common Objects. *Developmental Psychology, 17*(1), 111-114.

Bushnell, E. W., & Baxt, C. (1999). Children's haptic and cross-modal recognition with familiar and unfamiliar objects. *Journal of Experimental Psychology-Human Perception and Performance, 25*(6), 1867-1881.

Easton, R. D., Srinivas, K., & Greene, A. J. (1997). Do vision and haptics share common representations? Implicit and explicit memory within and between modalities. *Journal of Experimental Psychology-Learning Memory and Cognition, 23*(1), 153-163.

Landau, B., Smith, L. B., & Jones, S. S. (1988). The Importance of Shape in Early Lexical Learning. *Cognitive Development, 3*, 299-321.

Lederman, S., & Klatzky, R. (1987). Hand movements: A Window into Haptic Object Recognition. *Cognitive Psychology, 19*, 342-368.

Milner, A. D., & Bryant, P. E. (1970). Cross-Modal Matching by Young Children. *Journal of Comparative and Physiological Psychology, 71*(3), 453-458.

Morrongiello, B. A., Humphrey, G. K., Timney, B., Choi, J., & Rocca, P. T. (1994). Tactual Object Exploration and Recognition in Blind and Sighted Children. *Perception, 23*(7), 833-848.

Reales, J. M., & Ballesteros, S. (1999). Implicit and explicit memory for visual and haptic objects: Cross-modal priming depends on structural descriptions. *Journal of Experimental Psychology-Learning Memory and Cognition, 25*(3), 644-663.

Rose, S. A., Blank, M. S., & Bridger, W. H. (1972). Intermodal and Intramodal Retention of Visual and Tactual Information in Young Children. *Developmental Psychology, 6*(3), 482-486.

Smith, L. B., Jones, S. S., Landau, B., Gershkoff-Stowe, L., & Samuelson, L. (2002). Object Name Learning Provides On-The-Job Training for Attention. *Psychological Science, 13*(1), 13-19.

Collaborative Sensemaking in the Blogosphere

Richard Alterman (alterman@cs.brandeis.edu)
Johann Larusson (johann@cs.brandeis.edu)
Computer Science Department, Volen Center for Complex Systems
Brandeis University

Abstract

This paper presents a case study of a class of students co-blogging throughout the semester. The students collaboratively made sense of the course material. The class blogosphere became a repository of interpretations, reflections, opinions, monologues and dialogues about the course content. Over the course of the semester there was an aggregation of "sense made" that was "mined" by the students throughout the semester. The data shows that students leverage the contributions of other students when authoring their own posts and later when they write papers.

Keywords: Collaborative sensemaking; Online discourse communities; Co-blogging; Education; Case Study; Field study; Ethnography

Introduction

In a class it is not enough just to remember or retain the information that is presented, it must be "digested" or understood (Dewey, 1964: p. 249): "Of course intellectual learning includes the amassing and retention of information. But information is an undigested burden unless it is understood. It is *knowledge* only as its material is *comprehended*. And understanding, comprehension, means that the various parts of the information acquired are grasped in their relations to one another – a result that is attained only when acquisition is accompanied by constant reflection upon the meaning of what is studied. "

The in-class lecture and discussion provides an explanation of key concepts within the course content and a causal story about how the parts are connected. A student begins to learn the background knowledge, a foundation and framework for understanding the course material. The acquisition of this kind of background knowledge prepares the student for being able to produce causal explanations of key ideas, the relations between issues, and the connections between conclusions drawn from different evidences. Acquiring the background knowledge is a good part of what any course is about.

In itself, the in-class lecture and discussion is not enough to achieve a deeper understanding of the material. Other activities, including carefully reading the course texts, doing homework, and studying for exams, are exercises that help students "digest" material. Finding venues for students to cooperatively verbalize, explain, and discuss undoubtedly has positive educational value. However, finding a time and place for students to meet is a significant barrier for creating collaborative sensemaking opportunities.

This paper explores the value of online co-blogging as a discourse community that provides an arena for the students to work together and collaboratively make sense of the ideas and concepts taught in the class. The blogosphere is a play space for students to work at "understanding" the course material, even though the students work at different times and in different locations. Activity in the blogosphere is an opportunity to reflect, verbalize, get feedback, read alternate interpretations of the same material, and discuss: the students collectively make sense of the course material. There is a clear boundary between those items which are jointly made sense of in the blogosphere and those that are not.

This paper will present a case study of an interdisciplinary class on Internet & Society where the students co-blogged throughout the semester. The blogosphere provided an intersubjective space in which the students collaboratively worked at making sense of the lecture and course texts. In the blogosphere, the students created *common* and *background knowledge* (Lee, 2001). The data shows that the students drew on the sensemaking of other students that aggregated during the semester in support of their own individual sensemaking.

Co-Blogging

In a student co-blogging community, each student has a *blog*. The blog is composed of multiple *posts* written by the blog owner. Blog posts can summarize the key content of a text that was read for class, or develop an argument on some issue that was discussed during lecture. Students can read each other's blog posts and *comment* on them. A discussion emerges when a blog attracts a lot of commentary from other students. Blogging on the course material is a learning activity that invites reflection and self-explanation and improves learning. Reading and commenting on each other's blog posts provides students with other interpretations of the course material and the opportunity to discuss the content of the readings, which also helps learning.

In-class discussions have significant time constraints. Online, students can converse – and (co-)reflect – at their leisure, when they are prepared. Because co-blogging is a text-based community, literal quoting of the text is easier to do: the data shows that in many cases, students literally copied, or paraphrased, a small portion of an assigned reading in order to focus their blog post. Perhaps these kinds of activities occur in an in-class discussion, but since face-to-face discussions are serial there are fewer opportunities for students to present different quotations from the text or alternative viewpoints on the same quote.

The co-blogging community is social and student-owned (Oravec, 2002). Because each student has her own blog, she has full control over the content and can establish personal and intellectual ownership of her work (Fredig & Trammell,

2004). Because co-blogging is Web 2.0 technology, the "buy-in" for students is fairly cheap (Glogoff, 2005; Duffy 2008). Because co-blogging occurs outside the bounds of class time and it is an asynchronous learning activity that does not require the student be collocated, it expands the opportunities for co-reflection and fruitful discussion.

In contrast to discussion forums, in a co-blogging learning activity, students develop individual identities: each student has her own blog. In a discussion forum each discussion has a deep tree structure, and in the blogosphere, the range of discussion is broader with multiple viewpoints, and conversations, emerging. In a discussion forum, because of "the branching structure, the large proportion of messages that terminated branches, and the abstracted nature of student interaction demonstrate an overall incoherence in online discussion. ... Leads to poorly interrelated monologues." (Thomas, 2002). In the blogosphere discussions develop as smaller chunks of interaction. Where the comments of an individual student can be buried in an extended discussion in a discussion forum, in the blogosphere every student blog attracts a significant amount of attention (Larusson & Alterman, 2009).

When a student writes a blog post she has the opportunity to practice producing a narrative about the significant elements of the course material, making sense of the causal relations among the different elements of the course content (Williams & Jacobs, 2004). Co-blogging creates opportunities to exchange, explore, and present alternate viewpoints (Fredig & Trammell, 2004). It potentially exposes students to alternate ways of "seeing" and "constructing" what is significant and why (Oravec, 2002; Fredig & Trammell, 2004).

When a blog post attracts commentary, it serves to coordinate the students' work at aligning their views. In this manner, the students can work "through" (Bødker, 1990) a post or discussion together, working at different times in different places to reach a common understanding. Discussions on issues related to the course material naturally emerge, enabling students to collaboratively work through the arguments and trade-offs, weighing and comparing different explanations and justifications (Okada & Simon, 1997), which positively impacts learning (Andriessen, 2006).

The discussions that emerge among the students create a dimension of interactivity. Some students comment (interact) more frequently than others (Rafaeli & Sudweeks,1998). Each comment can be classified by its level of interactivity. Comments on posts can either elaborate or negotiate (Thomas, 2002). Comments can either be reactive, refer to a prior point in the emerging discussion, or they can be interactive, i.e. "recount the relatedness of earlier messages" (Beuchot & Bullen, 2005; Rafaeli & Sudweeks,1998).

Case Study

In the Internet & Society course taught in Fall 2008, 25 students collectively blogged throughout the semester. The course was an introductory course. Students in the class were from a variety of disciplines. There were 8 females and 17 males. All of the students were undergraduates. There were 3 science majors and 1 science minor in the class. There were 12 students majoring in the social sciences and 8 minoring in the social sciences. The remainder of the class was either in the humanities or fine arts.

The focus of the analysis presented in this paper is on the co-blogging work that the students did during the time the class read two of the books that were required reading. The students wrote a short paper on each of these books.

Methods

All of the students' online work was automatically recorded in a transcript, which enabled both quantitative and qualitative analyzes. The transcripts can be treated as an event log file and accessed using database queries. Additional tools enable a larger variety of alternate analysis methods, including discourse, conversation, or interaction analysis. One tool replays the transcripts just as if one was viewing a videotape showing the evolution of the blogosphere. Another tool makes it easy for an analyst to systematically annotate, and tag each of the posts and all of the discussions that emerged.

If a student used a newsletter to navigate to the blogosphere, it was possible to determine that the student read the newsletter and also which conversation or post was their destination. If a student's email client automatically viewed emails in a HTML format, it was possible to track whether a student opened a newsletter even if they did not navigate from the newsletter to the blogosphere. It was not possible to determine which parts of the newsletter were read.

At the end of the semester we distributed a survey, questions were on a 6-point Likert scale (from 1, not useful, to 6, very useful). The survey provided some data on the students' perception of the academic value of the learning exercise and the functionality and practicality of the collaborative technology. The survey also included open-ended questions.

Metrics

Lectures were presented using slides that summarized the key points of the presentation. At the beginning of each lecture, hard copies of the slides were handed out to support student note taking. PDF versions of the slides were downloadable from the class website.

The lecture slides were used as a basis for identifying the inputs to the blogosphere. For each set of slides, the instructor identified a set of key topics that were covered by the lecture. For each topic a tag was created that was organized into a taxonomy and treated as the potential input to the blogosphere. All posts and comments in the blogosphere were tagged using these topic/tags; this roughly identified the content of each contribution. When the students post on these topics they are reflecting on important course content. One way to measure the impact of a given post is to count the number of comments or reads that it accrued.

Procedure

At the beginning of the semester, an in-class tour and exercise introduced the students to the important features of the co-blogging environment. The students were required to blog at the pace of one post per lecture: there were two lectures per week. A typical post was 1 or 2 paragraphs in length. The students were also required to read and comment on other contributions to the blogosphere. The co-blogging work of each student counted for 35% of his or her grade. Students had the option to opt-out of the study. No student opted-out of the study.

During the semester the students read four books. The students wrote short papers on two of these books. The focus of the analysis presented in this paper is on the co-blogging work that the students did during the time the class read the two books for which they wrote papers.

The Co-Blogging Environment

The co-blogging environment has been developed over a number of years in several different courses following the *design-based research* methodology (Barab, 2006). It is implemented using the Wiki Design Platform (WDP), which is a wiki-based educational platform that supports a variety of collaborative learning activities (Larusson & Alterman, 2009).

In the co-blogging environment, each student has a blog. Each blog post shows a picture of the author, a title, and tag that relates the post to a lecture given in class. At the bottom of a post there is a list of people who read the post. Any threaded discussion that emerges is shown below the relevant post. As a student writes her blog, she can read another student's post on the same topic with a click of the mouse. At the "front entrance" to the blogosphere, there is a list of the ten most recent posts or comments on posts. Each item in the list displays the name of the author of the post or comment and a short excerpt from the contribution. Students can also access the blogs via a word cloud or by searching the content in the blogosphere using keyword(s) or tag(s).

Students receive daily email newsletters that summarize the online co-blogging activity of the class in the previous 24 hours. The newsletter lists the title, author, and first line of all the newly created blog posts, and a list of similar information for any new comment. Students can use the links on the newsletter to directly navigate to any post or comment on the blog site that is of interest.

In the blogosphere there are two ways to be a *primary participant*: author a blog or act as a discussant on another student's blog (Alterman & Larusson, 2009). *Secondary participation* occurs when a student reads either a post or a discussion that has emerged online. A *tertiary participant* reads a brief description of a recent post or a new comment on a post in a newsletter. The students can assume different participant roles at different times. A student can be the author of a post, a contributor to a conversation initiated by a post, a reader of a post or conversation, or an interested party who reads about the post or conversation in a daily newsletter. Secondary and tertiary participation are more peripheral kinds of participation.

Evaluation

Responses to the survey were positive. When the students were asked to rate the value of their online co-blogging work as a means of giving them first-hand experience with online collaborative learning, the average response was 5.6. In response to the question of whether the students felt the co-blogging community was useful, the average response was 5.3. When queried about the usefulness of the blogosphere for writing papers, the average response was 4.5. When asked as a yes/no question whether re-reading and reusing the blogging text helped the students write their papers, 67% answered in the affirmative.

There were a total of 155 blog posts, 113 comments, and 1010 reading events on the two books that are the focus of this study. There were 31 conversations of length 2, 15 of length 3, 7 of length 4, and 7 of length 5. The average conversation length was 2.85. The length of a conversation is defined as the number of contributions that were made to the discussion. For example, a post that receives one comment is a conversation of length 2.

There was no correlation between the number of tags on a given post and how often it was read; many of the best posts were thoughtful commentaries on a single topic. There was no correlation between the length of a conversation and the number of tags garnered in the conversation. There was, however, a strong positive correlation between the length of a conversation and the number of read events ($r(151) = .061, p < .01$).

Participating in the blogosphere

Students made two kinds of contributions to the blogosphere. As a *blogger*, each student produced an open journal, a monologue about the course content. As a *discussant* each student participated in a dialogue about the content of one or another post.

As a blogger, a student posted her reflections on some part of the course material. A blog post could refer to the text or quote the text; this occurred 75 times during the time the students co-blogged on the two books (roughly 48%). A post could refer to the lecture, an issue that was discussed in class, another blog, or to an outside article, site, or book (26 times; roughly 17% of the time for the two books). Frequently students included personal experiences or anecdotes as part of their post throughout the entire semester (73 times; roughly 14% of the time), and less frequently during the time they co-blogged on the two books (8 times; roughly 5%). Each of these were ways to initiate reflection.

Within the blogosphere, the monologues of the students were published and broadcast to the rest of the class, emerging in an open space, giving students exposure to multiple viewpoints and voices. Students viewed the same material

differently. Their different articulations complemented, regulated, or clashed with one another. All voices could "be heard". The ratio, the balance, of these voices potentially gave a student a textured view of the course material. By means of *perspective-taking* an intersubjective space emerged (Tomasello et al, 1993).

In addition to authoring posts, students acted as discussants on each other's posts. Much of the commentary was either an agreement with, or an expatiation of, another student's point (49 times during co-blogging on the two books; roughly, 43% of the comments). These sorts of confirmations moved the students towards creating a common understanding of a particular interpretation of some portion of a text or lecture. Sometimes a student posed a question or asked for a clarification in her blog or comment, which was answered later by the comment from another student (10 times; roughly, 9% of the comments on posts for the two books). Other responses were more discursive: students frequently disagreed, espousing different viewpoints on the same topic (52 times; 49% of the comments). Comments were linked to other posts (2 times; roughly 2%). Comments either referred to the initial post (102 times; roughly, 90%) or another student's comment on the post (14 times; roughly, 12%).

Intersubjective space

Participation in the creation and use of information in the blogosphere results in learning and the production of `common knowledge`. The students work together to "digest" the information that is presented during lecture or in the course texts.

The total number of additions to the blogosphere is a rough measure of the amount of information "digested" by the class while participating in the co-blogging exercise during the semester. One of the topics in the Internet & Society class was the advantages and disadvantages of "working home alone" as opposed to working in an office with your collaborators. Let x_1, x_2 ... represent the advantages and disadvantages of working home alone. Table 1 shows an idealized representative example sequence of events in the blogosphere that are ordered in time. At times t_1, t_2, t_3, and t_4 interpretations of content presented in the text or lecture are aggregated: x_1, x_2, x_3, and x_4 are added to the blogosphere.

Table 1: A sequence of events in the blogosphere.

Time	Event
t_1	Joe posts a blog on "working home alone", x_1.
t_2	Mary reads Joe's post x_1 and posts comment x_2 .
t_3	Mary posts a blog on "working home alone", x_3.
t_4	Joe reads Mary's comment on his post and replies. x_4.
t_5	Ed reads the conversation between Mary and Joe.
t_6	Ed reads Mary's post on "collocation".
t_7	Mary reads Joe's reply to her comment on x_1.

What each student learns, how much each student learns, and to what degree the students learn the same things, is all variable. The extent to which students converge on a set of agreed upon factors and arguments concerning some key concept is an open question. The degree to which the students share their beliefs is not clear either.

Common ground is defined in terms of a belief about some proposition p: p is a part of common ground for a set of actors if they all believe p and they believe that the other actors also believe p and that those other actors believe that they believe p and so on (Clark, 1996; Clark & Brennan, 1991). For the sequence shown in Table 1, at no point does it appear that Mary and Joe have attained *common ground* on x_1 (common ground: Clark & Brennan, 1991). At time t_4, Joe knows Mary read his post. At which point he may or may not believe that she understood his contribution. Suppose Joe believes Mary understood his contribution, he still does not know if Mary believes that he believes she understood his contribution. At time t_7, were Mary reads Joe's reply to her comment, even if Mary believes Joe believes she understood his contribution, Joe will not know that.

Lee (2001) makes a distinction between common, shared, and mutual knowledge. Each of these are distinguished by the *certainty of sharedness*. `Common knowledge` between two individuals is assumed to be held commonly by those individuals because that knowledge is considered to be general background knowledge within a community of which they are both a part. "`Shared knowledge`, on the other hand, is that information which has been established as shared as a result of interaction and discussion." `Mutual knowledge` requires an infinite regress of mutual belief, the certainty of sharedness is 100%. In the case of the sequence of events shown in Table 1, is common or shared or mutual knowledge established?

Lectures and student activity in the blogosphere are good venues for establishing common knowledge (background knowledge). Key points in an assigned reading or a lecture are likely to be `common knowledge` for the students; only likely because not all students read the assigned material or attend, or listen closely to, lectures. Sharing of knowledge within the blogosphere is asymmetric. When a student writes a post in the blogosphere and another student reads it, the second student believes she has `shared knowledge` with the first but not vise versa. So, for the sequence of events in Table 1, at time t_2, Mary believes she `shares knowledge` of x_1 with Joe, but Joe does not believe he `shares knowledge` of x_1 with Mary until time t_4. At time t_5, Ed may believe he `shares knowledge` of x_1, x_2, and x_4 with Joe and Mary, but neither share that with him. And so on.

In a face-to-face interaction, beliefs are grounded from a sequential interaction. In an online community, because all the students are not always together at the same time in the same place, `common` and `shared knowledge` emerges intermittently and non-uniformly; it is not clear that `mutual knowledge` ever emerges from the blogosphere alone. Many of the things the students learn/know as a result of their participation in the blogosphere are beliefs that may be held in common and shared but they are not mutually known.

During the co-blogging activity

Table 2 shows that on average 57% of the topics a student "considered" in the blogosphere were those the student wrote

about in one or another of her posts. On average, the other 43% of the topics that a given student "considered" occurred as a result of commenting or reading in the blogosphere. The variance is high for these numbers because there were a few students who were not very active at all.

Table 2: Learning from other students.

	Average	Median	Stdev
Blogging	57%	55%	22%
Commenting	12%	8%	15%
Reading	31%	28%	20%

These numbers do not reflect the fact that many of the students took advantage of a feature of the blogosphere environment that made it easy for a student writing a blog post to read other posts on the same topic. Over the entire semester, there were 13,408 reading events, 4,693 of them occurred while students were authoring blog posts (roughly 35%). Thus students were able to "mine" other interpretations of the same content even while they were authoring blogs.

While writing papers

Figure 1 shows how activity within the blogosphere exposed students to topics that were later included in one of the two papers they wrote.

1. The y-axis compares the number of topics/tags assigned to each student's posts and comments (primary participation) to that number for the same student's topics/tags in his or her paper. A positive number means that more of a student's paper was composed of topics they contributed initially in the blogosphere. A negative number means that a majority of the content in a student's paper did not originate in contributions to the blogosphere.
2. The x-axis computes a similar number for reads (secondary participation). A positive number means that more of a student's paper was composed of topics they read about in blogosphere prior to writing their paper. A negative number means that a majority of the content in a student's paper did not originate from reading in the blogosphere.

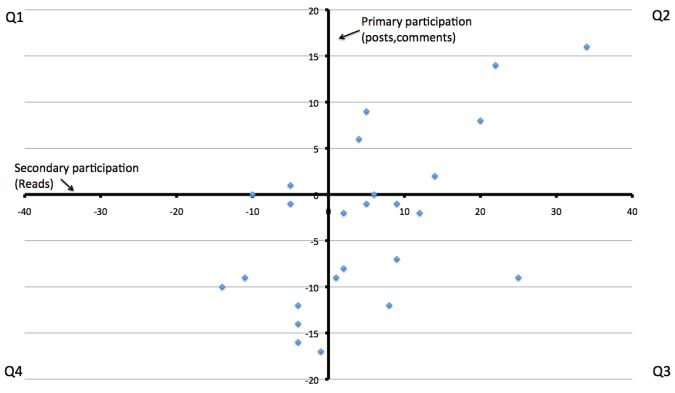


Figure 1: Influence on paper writing.

Consider each of the four quadrants of the graph starting in the upper left-hand quadrant:

Q1: Primary participation provided preparation for writing papers.

Q2: Primary and secondary participation provided preparation for writing papers.

Q3: Secondary participation provided preparation for writing papers.

Q4: Primary and secondary participation provided some help, but most of these papers were derived from work that was not influenced by a student's activity in the blogosphere.

For 16 of the 25 students, their work in the blogosphere provided background for the majority of the concepts that appeared in their two papers (their data is either positive on the x-axis or y-axis). The largest group of students (Q3) benefited most from the reading. The next largest group (Q2) benefited significantly from both primary and secondary participation in the blogosphere. These data confirm that students were "mining" the blogosphere to support their understanding of the material.

Figure 2 shows the correlations between the preparation for writing papers provided by reading, posting blogs, commenting, or doing all three. The trend line for all three activities combined is significant and positive ($r(23) = 0.485, p < .05$). The trend lines for reading ($r(23) = 0.402, p < .05$) and posting ($r(23) = 0.419, p < .05$) are also significant and positive. The trend line for commenting was not significant.

Discussion

Think of the blogosphere as a play space for students to work at "understanding" the course material. The blogosphere is an opportunity to reflect, verbalize, get feedback, read alternate interpretations of the same material, and discuss. The students are collectively making sense of the course material. The students leverage the aggregate online collaborative sensemaking throughout the semester.

The students produce multiple interpretations of the course material. The students reflect on the meanings of the assigned readings or a lecture given by the instructor in class. Frequently, posts include personal experiences or anecdotes. Comments on posts agreed with, or expatiated upon, another student's contribution; they also took contrasting views. The students are *many working minds* collaboratively making sense and creating `common` and `shared knowledge`; enabling many minds to work together is a significant outcome of Internet technology (Sunstein, 2006).

The blogosphere became a repository of interpretations, reflections, monologues and dialogues about the course content. At various points in the semester the students chose to *mine* the aggregated sensemaking. Because posts and discussions, once created, persist and can be re-considered at a later time, students can increase their common and shared knowledge throughout the semester.

On many occasions, as students composed their own posts, they first sampled another student's interpretation of the same lecture point or text. Right before a paper deadline, the students did heavy reading in the blogosphere in order to access

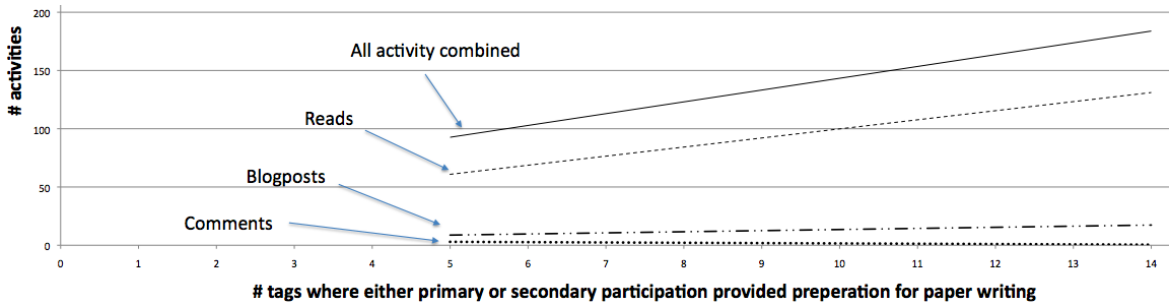


Figure 2: How different kinds of participation affect each student's preparation for writing a paper.

and review ideas, arguments, examples that were relevant to the paper they were writing, reproducing, in their own words, the content of relevant posts and discussions found in the blogosphere.

Concluding Remarks

During the semester, common and background knowledge is created by collaborative work in the blogosphere. The individual contribution of each student is an investment, the return on their investment is increased by the collective work of the class. The collective work of the class in the blogosphere produces multiple reflections on the course material. Students enrich their understanding by reading or commenting on the blogs and comments of other students.

The quantitative data from the case study shows that the students *mine* the blogosphere throughout the semester. When students write blog posts, 35% of the time they read other related contributions to the blogosphere first. On average 43% of the topics that a given student wrote about in an assigned paper was presaged by their participation in the blogosphere. The data also shows that for 16 out of the 25 students, the majority of the topics that appeared in their papers were first "played with" in the blogosphere as either a primary or secondary participation; the largest group of students benefited most from reading in the blogosphere. Finally, the data shows that there is a significant positive correlation between preparation for writing papers and a student's reading and posting activities in the blogosphere.

References

Alterman, R., & Larusson, J. (2009). Modeling Participation within a Community. In *Proceedings of the third annual conference of the cognitive science society* (pp. 1680–1685).

Andriessen, J. (2006). Arguing to learn. In R. Sawyer (Ed.), *The cambridge handbook of the learning sciences.* New York, NY: Cambridge University Press.

Barab, S. (2006). Design-based research: A methodological toolkit for the learning scientist. *The Cambridge handbook of the learning sciences*, *169*.

Beuchot, A., & Bullen, M. (2005). Interaction and interpersonality in online discussion forums. *Distance Education*, *26*(1), 67–87.

Bodker, S. (1990). *Through the interface: A human activity approach to user interface design*. Hillsdale, NJ: LEA.

Clark, H., & Brennan, S. (1991). Grounding in communication. *Perspectives on socially shared cognition*, 127–149.

Dewey, J. (1964). *John Dewey on education: Selected writings* (R. Archambault, Ed.). Modern Library.

Duffy, P. (2008). Engaging the YouTube Google-Eyed Generation: Strategies for Using Web 2.0 in Teaching and Learning. *The Electronic Journal e-Learning volume 6(2)*, 119-130.

Ferdig, R., & Trammell, K. (2004). Content Delivery in the'Blogosphere'. *The Journal (Technological Horizons In Education)*, *31*(7), 12–16.

Glogoff, S. (2005). Instructional blogging: Promoting interactivity, student-centered learning, and peer input. *Innovate. Journal of Online Education*, *1*(5).

Larusson, J., & Alterman, R. (2009). Wikis to support the collaborative part of collaborative learning. *International Journal of Computer-Supported Collaborative Learning*, *4*(4), 371–402.

Lee, B. (2001). Mutual knowledge, background knowledge and shared beliefs: Their roles in establishing common ground. *Journal of Pragmatics*, *33*(1), 21–44.

Okada, T., & Simon, H. (1997). Collaborative discovery in a scientific domain. *Cognitive Science: A Multidisciplinary Journal*, *21*(2), 109–146.

Oravec, J. (2002). Bookmarking the World: Weblog Applications in Education. *Journal of Adolescent & Adult Literacy*, *45*(7), 616–21.

Rafaeli, S., & Sudweeks, F. (1998). Interactivity on the Nets. *Network and netplay: Virtual groups on the Internet*, 173–90.

Sunstein, C. (2006). *Infotopia: How many minds produce knowledge*. Oxford University Press, USA.

Thomas, M. (2002). Learning within incoherent structures: the space of online discussion forums. *Journal of Computer Assisted Learning*, *18*(3), 351–366.

Tomasello, M., Kruger, A., & Ratner, H. (1993). Cultural learning. *Behavioral and brain sciences*, *16*, 495–495.

Williams, J., & Jacobs, J. (2004). Exploring the use of blogs as learning spaces in the higher education sector. *Australasian Journal of Educational Technology*, *20*(2), 232–247.

Inflectional Suffix Priming in Czech Verbs and Nouns

Filip Smolík (smolik@praha.psu.cas.cz)
Institute of Psychology, Academy of Sciences of the Czech Republic
Politických vězňů 7, Praha 1, CZ-110 00, Czech Republic

Abstract

Two experiments examined if processing of inflectional affixes is affected by morphological priming, and whether morphological decomposition applies to inflectional morphemes in visual word recognition. Target words with potentially ambiguous suffixes were preceded by primes that contained identical suffixes, homophonous suffixes with different function, or different suffixes. The results partially confirmed the observation that morphological decomposition initially ignores the affix meaning. With verb targets and short stimulus-onset asynchrony (SOA), homophonous suffixes had similar effects as identical suffixes. With noun targets, there was a tendency to respond faster after homophonous targets. With longer SOA in verb targets, the primes with identical suffix resulted in shorter responses than the primes with a homophonous suffix. Similar tendency was observed in some noun targets. The results confirm that it is possible to prime inflectional affixes, but that the mechanisms of morphological analysis may operate differently for different types of affixes.

Keywords: morphological priming, affix priming, word recognition, morphological decomposition, inflection

Introduction

Numerous studies suggested that morphologically complex words are decomposed to individual morphemes during visual word recognition. Most evidence for decomposition comes from priming studies, in which morphologically related words are presented in succession. Repeating the same morpheme in both the first word (prime) and the subsequent word (target) results in faster processing of the targets, as measured for instance by the lexical decision task.

The effects of morphological priming have been established first with words that overlapped in their root morphemes. Words consisting of roots and affixes have been shown to prime their roots (*friendly-friend*), as well other words derived from the same roots (*confession-confessor*, see e. g. Marslen-Wilson, Tyler, Waksler, & Older, 1994). This work established that word roots are accessed during the processing of morphologically complex words. If this is the case, functional morphemes should be accessed as well, and it should be possible to prime the access to these morphemes.

Priming of affixes proved more challenging than priming of word roots. Some studies found priming effects between words sharing prefixes, such as *dislike-disprove*. These effects were stronger than if there was mere orthographic overlap in the word initial segments, e. g. in *uncle-unhappy* (e. g. Chateau, Knudsen, & Jared, 2002; Giraudo & Graigner, 2003). While the findings with prefixes are quite robust, suffix priming has been more difficult to establish. (Marslen-Wilson, Ford, Older, & Zhou, 1996) found evidence for priming between auditory primes and visual targets that shared derivational suffixes (*darkness-toughness*). However, some research suggested that only prefixes could be primed, but not suffixes (Giraudo & Graigner, 2003).

Recently, Duñabeitia, Perea, and Carreiras (2008) were able to show affix priming in suffixed Spanish words. Their participants processed the suffixed words faster if they were preceded by words with the same suffixes. The effect was also present when the primes contained isolated suffixes only, or suffixes attached to strings of non-letter characters.

The literature thus indicates that affixes can be primed, even though there may be differences between prefixes and suffixes in the susceptibility to priming. However, all research sketched above worked with derivational affixes. It is not clear whether inflectional affixes are susceptible to morphological priming as well. Given that some languages have rich inflectional morphology and that many words in these languages appear with some inflection, the question about affix priming is highly relevant.

Early vs. late decomposition

The available evidence suggests that morphological decomposition of printed words proceeds by first removing all potential affixes and subsequently checks if this decomposition is the correct analysis. So, Rastle, Davis, and New (2004) showed that *brother* can prime *broth*, even though *brother* is not composed of the morphemes *broth+er*. Longtin, Segui, and Halle (2003) speak about *pseudo-derivation* in this context and show that pseudo-derived words may prime words that seem related to them. The meaning-blind early morphological decomposition may be responsible for the difficulties in detecting suffix priming. Duñabeitia et al. (2008) suggested that early decomposition is responsible for the lack of affix priming effects reported by Giraudo and Graigner (2003). Their study compared morphologically related primes (e. g. *fumet-MURET*) or orthographic control primes (*béret-MURET*). It is possible that the orthographic control primes were initially decomposed even though their final segment (*-et*) is not a true suffix. Because of this decomposition, Giraudo and Graigner (2003) did not detect any difference between these conditions. The evidence thus suggests that early stages of morphological decomposition ignore the meaning of affixes. If two homophonous affixes with different function are presented, they should initially have the same impact on the processing of subsequent words.

Current study

The present experiments explored whether the processing inflectional affixes in Czech nouns could be affected by morphological priming. Of particular interest was the issue of homophonous affixes and the process of their interpretation.

Participants saw suffixed target words. These targets were preceded with visual primes. In the two key conditions, the prime words ended in a suffix with the same phonetic form as the suffix in the target words. However, in one of these conditions, this suffix was fully identical to the target suffix, i. e. shared both its phonetic form and its function. In the other condition, the prime word contained a homophonous suffix with a different function.

The basic prediction was that the homophonous morphemes should have similar effects as the identical morphemes in masked priming with short stimulus-onset asynchrony. In unmasked priming, i. e. with longer SOA, suffix with identical function should result in stronger priming effects than the homophonous suffix that merely shares the form but not the function of the target suffix. Experiment 1 tested the prediction for short SOA, Experiment 2 for longer SOA. Each experiment involved two components, one with nouns and one with verbs as target words. The noun component of involved two additional conditions, the baseline, and a condition involving a prime suffix with different form but the same function as the target. The verb component only used primes in the two conditions with homophonous affixes.

Experiment 1

All target words in each component ended with potentially ambiguous suffix. In the noun targets, this was one of the Czech feminine nominative suffixes, *-a*. In the two key conditions, the primes also ended with *-a*. In the identical condition, the primes were feminine nominatives, in the homophone conditions, they were masculine accusatives/genitives. With regard to these two conditions, the prediction was that there would not be no difference between lexical decision times on target nouns. With 50 ms latency, the primes should be decomposed and suffixes identified by the time of target presentation, but the function of the suffixes should not be accessed yet. In order to test whether decomposition occurred at all, a condition with orthographically distinct nominative suffix primes was introduced in the noun component (no such controls were possible for the verbs). Reaction times in this allomorph condition should be slower because the search for the target suffix will not have started until the target is presented. The baseline condition served to establish the processing times for target target words with no primes.

The verb component focused narrowly on the comparison of the identical and homophone suffixes. No differences in the effect of these suffixes were predicted in Experiment 1.

Method

Stimuli The noun component contained 104 experimental items in four conditions summarized in Table 1. Four versions of the protocol were created so that each target word was presented in each condition to approximately the same number of participants. The verb component of the experiment presented 26 target words in two conditions: primes had either the same suffix, or a homophonous suffix. The verb component did not contain the baseline condition, nor the different-suffix condition. This was mainly because the number of possible stimuli was much smaller. Two versions of the verb component were created and presented to approximately equal number of participants, so that each target word was presented in each condition to equal number of participants. All the experimental conditions were constructed so that the prime and target words had approximately equal frequency, and that in each version of the protocol, the mean frequency of the primes and targets was approximately equal. The primes and targets always had the same number of letters. Primes were presented in uppercase, targets in lowercase letters.

Table 1: Sample stimuli from all conditions

	Condition	Prime	Target
Noun targets	baseline	XXXX	váha
	identical	LÍPA	váha
	homophone	SYNA	váha
	allomorph	VŮLE	váha
Verb targets	identical	BERETE	žijete
	homophone	KUŘETE	žijete

Besides the 140 experimental trials, 123 real word fillers and 270 nonword fillers were presented. The presentation was block-randomized so that trials from different conditions occurred with approximately equal probability during the whole experiment.

Pariticipants Thirty-nine students participated in the experiment as a part of their course requirement. All were native speakers of Czech.

Procedure The experiment was presented on a laptop computer using DMDX (Forster & Forster, 2003) as the presentation and response-collection software. Each trial started with a fixation cross presented for 500 ms. Then, the prime word was presented for 50 ms, followed by the subsequent presentation of the target word. The target word was shown until response was made or until 1500 ms from the onset. If no response was made within 1500 ms, the no response was recorded and the computer proceeded to the next trial.

Analysis The data were analyzed using linear mixed models with random effects for persons and items. This procedure replaces the separate ANOVA analyses by subjects and items (cf. Baayen, Davidson, & Bates, 2008). Post-hoc pairwise comparisons used the Tukey method as implemented in the `multcomp` package for R (R development core ceam, 2003).

Results

Results are summarized in Table 2. The initial analysis compared the reaction times in the experimental conditions to the baseline using planned contrasts. Compared to the baseline, reaction times were significantly longer in the nominative allomorph (*-e*) condition ($t = 3.55$, $p < 0.001$). In the identical (nominative *-a*) condition, the times were also slower and

Table 2: Top: baseline reaction times and the effects in experimental conditions. Bottom: pairwise comparisons of reaction times in experimental conditions. ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$

	SOA	
	50 ms	150 ms
Baseline	664	667
Nom. -a (identical)	*10	*13
Acc. -a (homophone)	2	***20
Nom -e (allomorph)	**17	*12

	50 ms	150 ms
Identical − homophone	8	7
Identical − allomorph	-7	-1
Homophone − allomorph	*15	-8

Table 3: Top: baseline reaction times and the effects in experimental conditions, separately for short and long nouns. Bottom: pairwise comparisons of reaction times in experimental conditions. **$p < 0.01$, *$p < 0.05$, †$p < 0.10$

	SOA			
	50 ms		150 ms	
Word length (letters)	4–6	7	4–6	7
Baseline	658	672	668	666
Nom. -a (identical)	*17	1	8	*19
Acc. -a (homophone)	6	-3	*17	**25
Nom -e (allomorph)	**19	*15	*18	4

	50 ms		150 ms	
Identical − homophone	11	4	-9	-6
Identical − allomorph	-2	-14	-10	15
Homophone − allomorph	-13	†-18	-1	*21

the difference was significant ($t = 2.01$, $p = 0.048$). There was no significant difference between the baseline and the incongruent homophonous (accusative *-a*) suffix condition ($t = 0.46$, $p = 0.67$). Post-hoc analysis using Tukey contrasts revealed a significant difference between the accusative homophonous condition and the nominative allomorph condition ($p = 0.01$). There was no significant difference between the congruent condition and the incongruent or allomorph condition.

The prediction for the experiment was that there should be no difference between the congruent and incongruent condition. This is in line with the results. However, both these conditions should be faster than the allomorph condition. This is only true about the incongruent condition. In order to examine this discrepancy, an analysis was performed that divided the items according to their length. It may be the case that the shorter prime words were processed to a larger extent that the longer primes. If there are any differences between shorter and longer words, the original prediction should be valid for the longer words, that were presumably processed to a lesser extent, In shorter words, differences between the two key conditions may surface because the prime suffix has been processed enough so that its function is being accessed.

Two analyses were performed separately for two groups. One group consisted of stimuli with 4-, 5- and 6-letter words (57 trials), the other group of stimuli with 7-letter words (47 trials). Results are summarized in Table 3. In longer words, the pattern of results seemed to fit the expectations: there was only a small difference between the identical and homophonous condition, but the allomorph condition appeared slower and was significantly slower than the baseline ($t = 2.17$,$p = 0.03$). However, the pairwise comparisons revealed only a marginally significant difference between the homophonous and allomorph condition ($p = 0.05$). There was no significant pairwise difference between the identical condition and the allomorph condition. In the group of shorter words, responses in the identical and the allomorph condition were slower than the baseline ($t = 2.58$,$p = 0.01$ for identical, $t = 2.94$,$p < 0.01$ for allomorph). In post-hoc analysis, there were no significant differences between conditions.

In trials with verb targets, there were only two conditions. Reaction times were 14 ms slower in the homophonous condition than in the identical condition. This difference approached statistical significance ($t = 1.71$,$p = 0.08$). Because the observed difference conflicted with the prediction, the analysis was repeated for shorter and longer words, with the expectation that the results for shorter words would fit the original prediction. In the 24 trials with 7- and 8-letter-long words, the responses were 8 ms slower in the homophonous condition, a nonsignificant difference ($t = 0.77$,$p = 0.44$). In the 12 6-letter trials, there was a significant 28 ms effect ($t = 2.06$,$p = 0.04$) with faster reactions in the identical condition.

Discussion

The predictions for Experiment 1 were only partially confirmed. No significant difference between the priming effects of homophonous suffixes was found, which was in line with the predictions. However, the reaction times in the identical condition were significantly slower than the baseline, while those in the homophone condition were very close to the baseline. The two conditions with the *-a* suffix in the primes may thus have differing priming effects, which was not expected. The allomorph condition resulted in the slowest reaction times, being significantly slower than both the baseline and the homophone condition.

The differences between the identical and allomorph condition might be due to differences in the progress of processing in prime words of different lengths. The subsequent analysis supported this view to a certain extent. The difference between the effects of identical and homophone primes was weaker in long words, which were presumably processed to a lesser extent. However, even here, the two critical conditions did not pattern in a completely identical manner and only the homophone condition showed a marginally significant advan-

Table 4: Top: overall reaction times and the condition effects in verbs. Bottom: reaction times and effects form Experiment 1, separately for long and short words. $^{**}p < 0.01$, $^{*}p < 0.05$, $\dagger\, p < 0.10$

	50 ms	150 ms
Verb 2pl. *-ete*	735	707
Genitive noun *-ete*	†14	**21

	50 ms	
Word length (letters)	6	7, 8
Verb 2pl. *-ete*	705	743
Genitive noun *-ete*	*28	8

tage over the allomorph condition. In longer words, there were no significant pairwise differences between the conditions, but the identical and allomorph conditions were significantly slower than the baseline, while the homophone condition was not.

The analysis of the verb component showed a marginally significant effect of condition, with homophonous targets showing a tendency to slower responses. The subsequent analyses for longer and shorter words suggested that the marginal effect could be attributed to short words, which showed significantly longer reactions in the homophone condition. Apparently, 50 ms is enough time for the word processing system to start accessing the function of a suffix at least in shorter words.

Some of the findings are surprising, especially the relative effects of the identical and homophone condition in nouns and verbs. The pattern in verbs was in line with intuition: if there is any difference between conditions, the homophone condition should be slower, since the suffix on the homophone primes only shares its form, but not its function with the target suffix. In shorter verbs, there was indeed a significant difference in this direction. However, the pattern in nouns appears to be opposite. There was a tendency in the identical condition towards slower reaction times than in the homophone condition. This was especially apparent in the group of shorter nouns, where the identical condition, but not the homophone condition, was significantly slower than the baseline. Possible reasons for this pattern are addressed in the general discussion below.

Experiment 2

Experiment 2 presented the same stimuli with longer SOA. Under these conditions, it was expected that the homophonous condition will elicit slower reaction times than the identical condition. If the function of the suffix is accessed within the chosen SOA (150 ms), the effect of the identical and allomorph suffix should be identical, or at least their difference should be smaller than in the homophonous condition.

Method

Design, procedure, participants Experiment 2 used the same design, stimuli and procedure as Experiment 1. The only difference was in the stimulus onset asynchrony. The primes were presented for 150 ms. Responses were collected from 38 students who volunteered or participated in exchange for course credit. None of the students participated in Experiment 1.

Results

The results are summarized in Tables 2, 3, and 4, along with the results from Experiment 1. In the noun component, the reaction times in all three experimental conditions were significantly slower than in the baseline condition (identical: $t = 2.57$, $p = 0.01$; homophone: $t = 3.93$, $p < 0.001$; allomorph: $t = 2.23$, $p = 0.03$). Pairwise post-hoc analysis revealed no significant differences between the individual levels. The direction of the differences was in line with the expectations, with the longest reaction times in the homophone condition, and the allomorph and identical condition eliciting similar responses. However, none of the pairwise differences between the experimental conditions were significant.

In order to examine the results more closely, the stimulus set was again split, and the groups of short and long words were analyzed. In the shorter words, there was a significant difference between the baseline and the homophone condition ($t = 2.41$, $p = 0.02$), as well as the allomorph condition ($t = 2.67$, $p = 0.01$). This would suggest an advantage for the stimuli primed with the identical suffix. However, post-hoc pairwise comparisons have not revealed any significant difference between the experimental conditions. Therefore, even though there seems to be an advantage for the identical condition, the prediction is not supported.

In the group of long words, the pattern of results is different. Compared to the baseline, the reaction times were significantly slower in the identical condition ($t = 2.42$, $p = 0.02$) and in the homophone condition ($t = 3.14$, $p < 0.01$). Pairwise comparisons revealed a significant difference between the allomorph and homophone condition, with homophone condition significantly slower than the allomorph condition ($z = 2.62$, $p = 0.04$).

In the verb targets, the reaction times in the homophone condition were significantly slower than in the identical condition ($t = 2.88$, $p < 0.01$). The results from the verb component are in line with the predictions.

Discussion

In Experiment 2, the predictions were again confirmed only partially. In the verb component, the homophone condition was slower than the identical condition, which is in line with the expectations. However, the expected differences in the more complex, noun component of the study have not materialized completely. Overall, there was a nonsignificant tendency for the reaction times to be longer in the homophone condition (20 ms effect against baseline) than in the identical or allomorph condition (13 and 12 ms effect, respectively).

This would be in line with the expectations. However, separate analyses for shorter and longer words complicated the picture. In the group of shorter words, more thorough processing of the primes is expected. The results should be in line with the predicted pattern. However, the homophone condition, predicted to be the slowest, has practically identical effects as the allomorph condition. These two conditions are significantly slower than the baseline. While this is not in line with the prediction, it is understandable under the assumption that the effects of orthography and function are about equally strong at 150 ms SOA. In the homophone condition, the response is inhibited by the difference in the morpheme function (accusative instead of a nominative marker). In the allomorph condition, the function is the same, but processing is inhibited by the difference in orthography. In any case, there was no significant pairwise difference between the identical condition and the two slower conditions, so the effects should be understood as a mere tendency.

In longer nouns, the pattern of results was more intriguing. Responses in the identical and homophone conditions were significantly slower than the baseline. Pairwise comparisons showed a significant advantage of the allomorph condition over the homophone condition. This appears to suggest that in these words, the function of the suffix is more influential than its orthographic form, since the condition with the different-function suffix is significantly slowed down. However, in such a case, the identical condition should be even faster than the allomorph condition. Another surprising aspect of the results is the fast response in the allomorph condition. The longer words are presumably processed to a lesser extent than the shorter words discussed above. Yet, the inhibiting effect of orthography against the baseline is present in the shorter, more completely processed words, and not in the longer words. This goes against the assumption that the orthographic form is accessed first and the function later. Moreover, it goes against other aspects of the present data. The allomorph condition was slower than the baseline both in Experiment 1, where the primes were presumably processed to a lesser extent, as well as in the short words in Experiment 2, where the processing of the primes progressed more than in the long words.

General discussion

The experiments examined the effects of morphological priming on word recognition. While the phenomenon has been well established with derivational morphemes, little research is available for inflectional morphemes. The results show that inflectional affixes can exert priming effects similar to those reported by Duñabeitia et al. (2008) and others for derivational affixes. However, the evidence is unequivocal only for the 2nd person plural verb suffixes at 150 ms SOA. For nominal suffixes, the result show a more complex pattern.

In verb targets in Experiment 2, the presentation of a noun prime with homophonous suffix inhibited word recognition compared to verb primes with identical suffix. This means that after 150 ms from the prime onset, the processing of the suffix moved beyond the level orthography, and words ending with homophonous suffixes inhibited the processing of target words. In Experiment 1, no such difference was predicted. It was expected that mere orthographic overlap between the prime and target suffix would initially affect the targets equally strongly as the repeated presentation of an identical suffix in the prime and target. However, it appears that in short words, the ending is recognized even within the 50 ms window, resulting in a morphological priming effect exceeding the effects of orthography.

The results from nouns require more detailed discussion. There was no significant difference between the two key conditions (identical and homophone) in Experiment 1, which was predicted. However, it was predicted that these two conditions would result in significantly faster responses than in the allomorph condition. This was true only for the homophone trials. Trials with identical suffix primes had longer reaction times than the homophone trials, and were not significantly different from the allomorph trials. This should not occur if the initial decomposition is blind to the function of the prime ending Moreover, the difference between the homophone and identical trials, though nonsignificant, goes in the unexpected direction and contradicts the findings from the verb component.

It is useful to summarize the results from the two key conditions based on the presumed amount of processing performed on the primes. On longer words with shorter SOA, i. e. after the least amount of processing, none of these conditions is faster than the baseline. In shorter words and short SOA, the identical condition is slower than the baseline. In longer words with 150 ms SOA, both identical and homophone conditions are slower than the baseline. Finally, with longer SOA and shorter words, only the homophone condition is slower than the baseline. This result is in line with the prediction that in the later stages of processing, the functional aspect of the affixes will play stronger role than their orthographic form. However, it is not clear why the primes with identical and homophone suffixes result in slower processing of long words in Experiment 2, and why identical suffixes inhibit processing of short word targets in Experiment 1.

One possible explanation is that morphological decomposition does not occur in frequent nominative forms. In this view, the processing system attempts morphological decomposition unless it can recognize the whole word form as a whole. If decomposition is attempted, the function morpheme is initially identified regardless of its function. If it is not attempted, there is nothing that would exert priming effect on the targets. If this view is correct, the accusative homophone primes in this experiment were decomposed. The *-a* suffix was initially not identified as accusative but activated all possible meanings, including nominative. For this reason, it facilitated the processing of the nominative target words. The nominative primes were not decomposed and thus could not exert the priming effects. This would explain the tendency

towards slower responses in the identical condition in Experiment 1, as well as the absence of the difference between identical and homophone primes in longer words in Experiment 2. In these longer words with 150 ms SOA, the homophone prefix presumably started to develop an inhibitory effect due to the functional difference between primes and targets. At the same time, the lack of priming due to the lack of nominative prime decomposition still inhibited processing after the identical primes.

This view may seem paradoxical. If nominatives are not decomposed, why should the decomposed *-a* affix from the accusative primes temporarily activate the nominative interpretation? The possibility must exist that a low-frequency word or a novel word will be analyzed as nominative. For this reason, separating the *-a* suffix activates the nominative interpretation, even though nominatives are not regularly decomposed. Another question that arises is why nominative targets should be facilitated if they are not decomposed. But it is not necessary to assume that facilitation of target processing operates on the suffix. The activation of nominative *-a* suffix on the prime may activate the whole corresponding declensional class of nouns (paradigm "žena"). There is independent evidence that declensional class of nouns is represented in an abstract manner (Bordag & Pechmann, 2009). This way, the target nouns could be primed even if not morphologically decomposed.

The reason why nominatives would not be decomposed lies in the fact that they function as the base and default form. Nominatives are considered the citation form of nouns, and they are the most frequent case form (Jelínek, Bečka, & Těšitelová, 1961). It has been proposed that only low-frequency words undergo morphological decomposition (Baayen, Dijkstra, & Schreuder, 1997; Baayen & Schreuder, 1999). Even though there is evidence that all suffixes, including pseudo-suffixes, are decomposed, the decomposition of nominatives may be slower than direct retrieval. In that case, nominatives would not be decomposed.

This proposed explanation might explain many aspects of the results reported here. Some aspects remain unexplained, especially the fast responses in the allomorph condition on long words in Experiment 2. In any case, the processes of morphological decomposition of inflectional suffixes deserve closer scrutiny. In particular, further research needs to test whether nominative words undergo morphological decomposition.

To summarize, findings from Experiment 1 suggest that the purely orthographic, function-blind stage of morphological decomposition may be over in less than 50 ms, at least in shorter words. At the same time, results from the noun targets in both experiments suggest the possibility that nominative forms do not undergo morphological decomposition.

Acknowledgments

The study was partially supported by the Czech Science Foundation award No. P407/10/2047 *Comprehension of grammar and lexicon in toddlers*.

References

Baayen, R. H., Davidson, D., & Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390–412.

Baayen, R. H., Dijkstra, T., & Schreuder, R. (1997). Singulars and plurals in dutch: Evidence for a parallel dual-route model. *Journal of Memory and Language*, *37*, 94–117.

Baayen, R. H., & Schreuder, R. (1999). War and peace: morphemes and full forms in a noninteractive activation parallel dual-route model. *Brain and Language*, *68*, 27–32.

Bordag, D., & Pechmann, T. (2009). Externality, internality, and (in)dispensability of grammatical features in speech production: evidence from Czech declension and conjugation. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *35*, 446–465.

Chateau, D., Knudsen, E. V., & Jared, D. (2002). Masked priming of prefixes and the influence of spelling-meaning consistency. *Brain and Language*, *81*(1-3), 587–600.

Duñabeitia, J. A., Perea, M., & Carreiras, M. (2008). Does darkness lead to happiness? Masked suffix priming effects. *Language and Cognitive Processes*, *23*, 1002–1020.

Forster, K. L., & Forster, J. C. (2003). DMDX: A Windows display program with millisecond accuracy. *Behavioral Research Methods, Instruments and Computers*, *35*, 116–124.

Giraudo, H., & Graigner, J. (2003). On the role of derivational affixes in recognizing complex words: Evidence from masked priming. In R. H. Baayen & R. Schreuder (Eds.), *Morphological structure in language processing* (pp. 209–232). Berlin: Mouton de Gruyter.

Jelínek, J., Bečka, J. V., & Těšitelová, M. (1961). *Frekvence slov, slovních druhů a tvarů v českém jazyce*. Praha, Czech Republic: SPN.

Longtin, C.-M., Segui, J., & Halle, P. A. (2003). Morphological priming without morphological relationship. *Language and Cognitive Processes*, *18*, 313-334.

Marslen-Wilson, W., Ford, M., Older, L., & Zhou, X. (1996). The combinatorial lexicon: priming derivational affixes. In G. Cotrell (Ed.), *Proceedings of the 18th Annual Conference of the Cognitive Science Society* (pp. 223–227). Mahwah, NJ: Erlbaum.

Marslen-Wilson, W., Tyler, L. K., Waksler, R., & Older, L. (1994). Morphology and meaning in the English mental lexicon. *Psychological Review*, *101*, 3-33.

R development core ceam. (2003). *R (programmable environment for statistical computing)*. Vienna. (Available from: http://www.r-project.org)

Rastle, K., Davis, M. H., & New, B. (2004). The broth in my brother's brothel: morpho-orthographic segmentation in visual word recognition. *Psychonomic Bulletin & Review*, *11*, 1090–1098.

Restructuring Causal Concepts

Eric G. Taylor (etaylor4@illinois.edu)
Department of Psychology, 603 East Daniel St.
Champaign, IL 61820 USA

Abstract

Typical studies of concept learning in adults address the learning of novel concepts, but much of learning involves the updating and restructuring of familiar conceptual domains. Research on conceptual change explores this issue directly but differs greatly from the formal approach of the adult learning studies. This paper bridges these two areas to advance our knowledge of the mechanisms underlying concept restructuring. The main idea behind this approach is that concepts are structured by causal-explanatory knowledge, and hence, models of causal induction may help to clarify the mechanisms of the restructuring process. A new learning paradigm is presented to study the learning *and revising* of causal networks. Results show that some behaviors indicative of conceptual change arise from basic causal learning mechanisms. Results also support models of causal induction that assume inhibition between competing causes.

Keywords: knowledge restructuring, conceptual change, belief revision, causal induction, concept learning.

Concept learning is an incremental process. We learn a concept for the first time only once, and often our initial understanding is flawed. The remainder of learning involves the updating, revising, and restructuring of previous conceptual knowledge. The critical implication—that most concept learning is actually the refinement of familiar concepts—runs counter to the traditional approach in the study of concept learning in adults, which has focused on the learning of entirely novel concepts (Murphy, 2002). Many open questions remain on the nature of concept restructuring.

The goal of this work is to better understand the basic mechanisms of concept restructuring by forging a connection between traditional work on concept learning and the literature on *conceptual change*. Although these two areas differ greatly (in everything from goals to dependent measures), this paper builds on recent work that highlights their commonalities.

Studies of conceptual change typically outline the process of knowledge restructuring in broad strokes: e.g., by showing that it often occurs abruptly (Kuhn, 1962), that people are highly resistant to giving up their prior beliefs (Chinn & Brewer, 1993), and that novice concepts appear to "differentiate" and "coalesce" over the course of development (Carey, 1985). To support these claims, authors have focused on specific real world domains and the shifts in knowledge therein, such as children's learning of biological concepts (Carey, 1985) and young adults' learning of physics (diSessa & Sherin, 1998).

These studies differ dramatically from the traditional research on concept learning in adults, despite great overlap in interests. The adult work has primarily used domain-general laboratory paradigms and formal models to assess the specific representations and processes underlying basic conceptual tasks like classification, inference, and category-based induction (Murphy, 2002).

A complete understanding of concept learning and restructuring requires explanations from both levels of analysis. This paper suggests that recent work developing the *theory view* of concept representation (Gopnik et al., 2004; Murphy & Medin, 1985; Wellman & Gelman, 1992) serves as a linkage between these levels. The theory view states that concepts are built upon networks of causal-explanatory knowledge. This knowledge affects performance in laboratory-based learning tasks (Murphy, 2002) and plays a role in the learning and development of real world concepts where conceptual change effects are typically demonstrated (Vosniadou, 2008). Assuming that concept learning amounts, in large part, to the learning of causal relations, then models of causal reasoning (Kim & Ahn, 2002; Rehder, 2003) provide the requisite theoretical tools for understanding the basic mechanisms of concept learning and potentially also conceptual change.

Few previous studies address this linkage to concept restructuring, however. Murphy's work (e.g., Kaplan & Murphy, 2000) has examined cases where prior causal knowledge is invoked when learning later concepts, but in these studies the prior concepts are not revised. Work on order effects in causal induction suggests that what is learned from the first half of a set of contingency data may be overwritten by later contingencies (Ahn & Marsh, 2006), but the initial learning (and hence, what is restructured) is not typically evaluated. A developmental study by Schulz, Bonawitz, and Griffiths (2007) showed that 4 to 5-year old children inferred causal relations from evidence that ran contrary to their prior beliefs. However, their evidence for belief revision, as measured by transfer performance, was mixed. This study is perhaps the strongest empirical evidence linking studies of causal induction to concept restructuring.

Other findings bearing directly on concept restructuring are less tied to the formal approach. Chinn & Brewer (1993) documented the many ways that people react to anomalous data, only one of which (the least common) was genuine concept revision. Chinn & Brewer (2001) also proposed a set of mental models for interpreting people's verbal evaluations of anomalous data and patterns of belief change,

but these were not formalized at the level specified in the causal induction models.

To directly address the linkage between concept learning research in the theory view tradition and studies of concept restructuring, I developed a task in which individuals would learn and then *revise* their hypothesized causal relations for a novel conceptual domain. The task was inspired by causal structure learning in real world domains, where one often develops a naïve, incorrect view of the underlying causal structure, and then with the accumulation of knowledge and evidence, restructures their original beliefs.

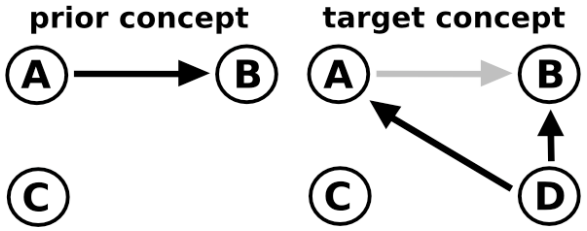


Figure 1: Diagrams of a hypothetical learner's causal representations en route to learning a common cause relation. The prior link remains in the target concept, though reduced, signifying a possible residual belief in that link.

The "common cause" scenario is one of many ways a learner may develop a prior, naïve concept and then need to restructure that concept based on new knowledge and evidence. See Figure 1 for an example. In this scenario, two variables—A and B—will appear correlated, and without further scrutiny, one may assume these variables share a direct causal relation. In fact, both A and B are caused by a third variable, the common cause. When the common cause becomes known, learners can track the relations it shares with variables A and B, and rule out the direct causal relation initially hypothesized.

This paper uses an empirical study based on the common cause scenario as a starting point to understanding the mechanisms underlying shifts in causal knowledge. Given that we currently know much about the initial learning of causal relations (i.e., the learning of the initial A causes B link), this study asks how that initial learning affects the process of concept restructuring. In particular, how does the belief in the prior concept affect later learning where one views contingency data in favor of the target explanation?

Consider the possible effects the prior concept may have on inferring the target structure. First, the prior concept may serve as an anchor, or bias, such that people show commitment to the A.B link (A.B means "A causes B") and later learning of alternative causes is more difficult. Previous work shows that prior beliefs are difficult to give up, especially when they figure centrally in other causal explanations (Chinn & Brewer, 1993).

Second, the acquisition of the prior belief may actually benefit later learning. In particular, evidence suggesting the lack of a correlation between other nodes in the system (between C&A and C&B) might draw resources away from those nodes and facilitate later search for the correct causal mechanism. This is especially true in Figure 1 since an alternative explanation for the A&B correlation is a mediating causal pathway, A.C.B. To the extent that one can rule out this "mediating cause" explanation, they might rule in the common cause explanation.

Third, both previous effects may occur. That is, learning the prior might increase one's belief in the A.B link, and independently, guide learners away from the wrong links and toward the right ones. If learners infer both the common cause and maintain a belief in the direct cause, they will have "over-explained" the occurrence of event B. Although previous work shows that people prefer simple explanations with fewer causal links (Lombrozo, 2007) and that competing causal hypotheses are considered in opposition (Lu et al., 2008), none have examined a case where learners are committed to a prior alternative conceptual structure, as is typically found in studies of conceptual change. In this case, people might over-explain to retain both possible causal pathways.

Experiment

The goal of the experiment was to determine how previously learned causal relations affect continued learning and concept revision. I created an experimental paradigm analogous to Figure 1. One group, the *change* condition, was verbally instructed on a prior structure with three nodes where A directly causes B, then in a second phase, was shown a fourth node (D) and had to infer the correct causal structure from contingency data. The control group, the *no-change* condition did not learn the prior structure and immediately attempted to infer the correct structure from contingency data with nodes A-D. The question is: How does the learning of the prior concept in the change condition affect the learning of the target concept, relative to that of the no-change condition?

Two dependent measures assessed learners' knowledge of the causal system. First, after the prior and target learning phases, participants rated the likelihood of each possible configuration of the system (e.g., A/~B/C for the prior phase, A/~B/~C/D for the target phase). These were used to infer participants "implicit" causal models of the system via model fitting, with the idea that some predictions offered above might not hold if participants were asked directly about their beliefs in the causal links (due to experimenter demands). Second, participants were asked at regular intervals during the target learning phase which of a set of possible links they believed were true. These judgments correspond to participants "explicit" beliefs about the causal system, similar to typical causal induction measures.

Method

Participants Forty-eight University of Illinois students participated in exchange for course credit.

Materials Participants learned about a fictitious ecosystem composed of four observable properties. Each property varied probabilistically during learning, taking one of two binary values (see Figure 2 for "on" values). The first property was the population size of a new fish biologists call "tespula": **above average** or **normal**. The second property was the color of a new type of algae called "plemocyn": **very green** or **normal**. The third property was the chemical composition of barium contained in the ecosystem's water: **crystallized** or **not crystallized**. The fourth property was the cloudiness of the water: **cloudy** or **not cloudy**. I refer to the first mentioned values as the "on" values.

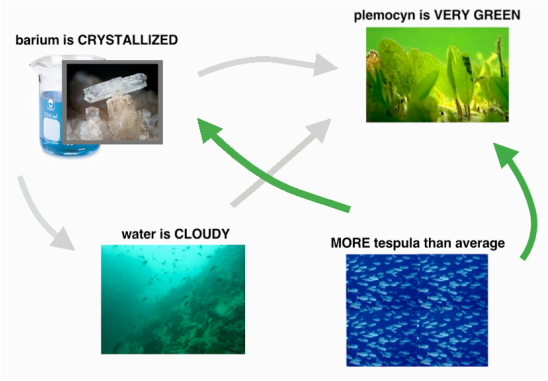


Figure 2. The causal structure of the ecosystem. Darkened links indicate that properties share a generative causal relation with causal power 0.85.

During covariation trials, the property values on each trial were determined by a causal system displayed in Figure 2. When the tespula population is more than average (base rate equal to 0.6), this will cause the barium to be crystallized with probability 0.85, and independently, the plemocyn to be green with probability 0.85. The water will be cloudy with probability 0.6. When the tespula population is average (depicted by a less colorful picture not shown in Figure 2), all other properties will be "on" with probability 0.6.

Covariation trials appeared like Figure 2, except that the property values varied probabilistically and all arrows appeared in grey. During the test phases, participants viewed all "on/off" combinations of the four properties and told to rate their likelihood (see *Procedure* section). In this phase, the arrows were completely absent.

Design Participants were divided into two groups: *change* and *no-change*, corresponding to those given a prior belief regarding the properties' causal relations and those who were not, respectively. Each group was subdivided into four counterbalance conditions, controlling for which properties were assigned to the roles in the causal system.

Procedure Prior to the experiment, participants read and signed a consent form. Participants then read instructions and completed all tasks on a computer.

Change condition: The instructions stated that the purpose of the task was to learn about a new oceanic ecosystem. Specifically, the task was to help a group of biologists to understand how the properties of the ecosystems cause one another. Three properties of that ecosystem were described—the top two properties from Figure 2 (A and B) and the bottom left property (C). The fourth property was absent during this phase. Participants were told that the biologists' current understanding was that property A causes property B (and told nothing else about C). They were also shown a picture with properties A-C and a green arrow connecting A to B. To ensure understanding, participants answered a multiple-choice question asking which properties were related and in what way. If they answered incorrectly, they repeated the instructions and re-took the question until they were correct.

Next, participants entered the *prior learning phase* where they viewed a sequence of 30 "snapshots" of the ecosystem. Each snapshot depicted a particular on/off configuration of properties A-C. Each snapshot appeared with a frequency proportional to its likelihood, which was determined using the probabilities given in the *Materials* section. To compute the probability of a particular snapshot, one computes the probability of each node taking its presented on/off value (conditional on the parent nodes) and then takes the product. Rehder (2003) describes this procedure building on Cheng's (1997) causal power theory, showing that the probability of node N being "on" is $1-(1-b_N)\prod(1-m_{CN})^{Con}$, where b_N is the probability of some unobserved background cause leading to the presence of node N, m_{CN} is the probability that node C generates the presence of N, and Con is an indicator variable equal to 1 when feature C is "on" and 0 otherwise. The snapshot frequencies were identical for all participants, but the order was random and different for each. Note that the causal system from Figure 2 creates a correlation between properties A and B, which supports the belief that A causes B when the status of property D is not visible.

After the 30 snapshots, participants entered the *prior likelihood rating phase* where they viewed each possible snapshot and were told to rate how likely the ecosystem is to look like the snapshot. They were also told, "when making the judgments, be sure to keep in mind the fact that the biologists think that [property A] causes [property B]." Ratings were given by moving a vertical bar up and down a scale, where the highest position indicated "VERY likely" and the lowest indicated "NOT likely."

Then, participants entered the *target learning phase*. They were told that the biologists discovered an important new aspect of the ecosystem, property D, and now they are wondering if their previous belief that A causes B was "wrong or perhaps missing something." They viewed a diagram similar to Figure 2 except with no links darkened, and were told their next task was to help the biologists figure out which of the shown potential causal relationships were true. Participants would learn which causes were true by viewing snapshots like those in the prior learning phase. The instructions also clarified that each property may occur

without being caused by another observed property (i.e., even if X causes Y, Y may appear in the absence of X) and that the links were not necessarily deterministic (e.g., if X causes Y, Y is simply more likely to appear in the presence of X). Finally, they were told that in addition to viewing the snapshots, they would sometimes be making predictions about which of the causes are true. Later during learning, the computer would give feedback about whether their hypotheses were close to or far from the true structure.

After every 10 snapshots participants were asked to guess which of the possible links were true. They were shown the picture in Figure 2 but with no links darkened, and told to click on the links to make their guess. Links darkened when selected. To assist with learning, participants were given indirect feedback regarding their link choices starting on their 4th hypothesis trial (after 40 snapshots)[1]. They were never told the status of any particular link choice (e.g., that the A.B link was right or wrong). Instead, they were told that the hypothesis was VERY GOOD, GOOD, WEAK, or VERY WEAK, indicating that 5, 4, [3 or 2], [1 or 0] links were correct, respectively. Participants were not told the correspondences between the feedback and number of accurate links. On the final hypothesis, participants were told, "This is your LAST PREDICTION. On the next trial, make your best guess as to what causes what."

Finally, in the *target likelihood rating phase*, participants again rated the likelihood of all possible snapshots of the ecosystem but this time with nodes A-D.

No-change condition: The no-change condition was identical to the change condition, but the prior learning phase and the prior likelihood ratings phase were excluded. The instructions immediately introduced participants to all four aspects of the ecosystem and the five possible links. Participants then began the target learning phase.

Results and Discussion

Hypotheses First, I present the results from the hypotheses participants made during the target learning phase. Each link was analyzed separately. Hierarchical logistic regression was used to evaluate the effects of condition and hypothesis trial on link choice. The "hierarchical" component refers to a random intercept term, which was used to model the between-participant variability in overall response tendency.

Results are plotted in Figure 3. To reduce inter-trial variability, I blocked the trials, except for the final trial: 1-4 (without feedback), 5-11 (feedback 1st half), 12-17 (feedback 2nd half), and 18 (the final trial). Main effects and interactions were assessed using Wald tests and likelihood ratio comparisons, but only Wald tests are reported. Likelihood tests led to similar interpretations.

The main effect of block on choosing the A.B link was significant, $\chi^2(1)=10.23$, $p<0.01$, suggesting that learning did occur, as participants selected this incorrect link less over time. The main effect of condition was marginally significant, $\chi^2(1)=3.44$, $p=0.06$, revealing an early and late bias in the change condition to select the prior link. The interaction was marginally significant, $\chi^2(1)=3.33$, $p=0.07$.

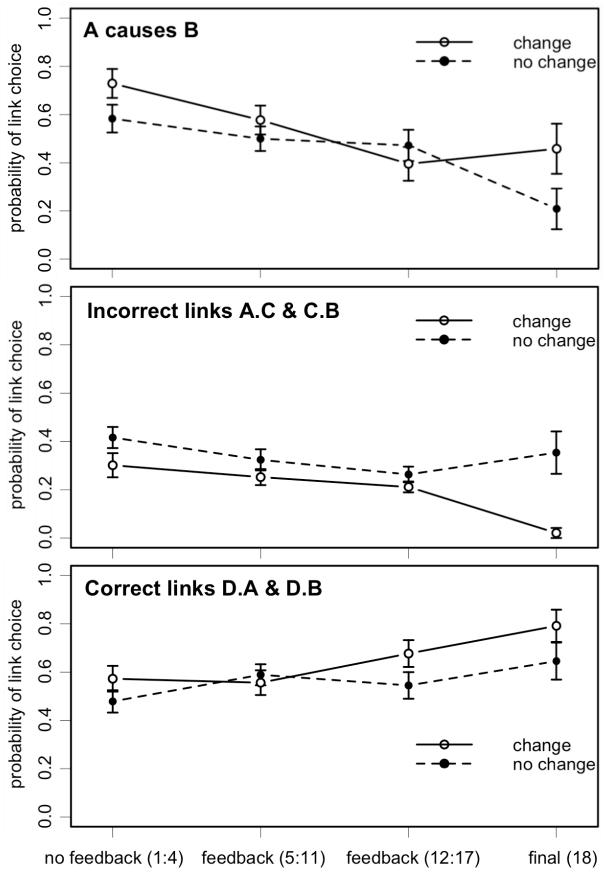


Figure 3. The probability of a participant including a link in their hypotheses during the target learning phase. Error bars are standard errors (binomial variance for final block).

Because the difference in conditions for the A.B link was non-monotonic over blocks, two separate regressions were fit to bocks 1-3 and blocks 3-4. The interaction between trial and condition was significant for blocks 1-3, $\chi^2(1)=9.36$, $p<0.01$, and for blocks 3-4, $\chi^2(1)=6.81$, $p<0.01$. Finally, the difference in conditions on just the final hypothesis was assessed using Fisher's exact test, which did not reach significance, $p>0.1$.

The interactions between trial and condition for the A.B link have two implications. First, although the change condition began selecting A.B more than the no-change condition, this difference went away by the third block as both conditions learned to not select A.B. Second, the difference in conditions increased from blocks 3 to 4.

[1] Feedback was added to improve learning based on the results of a pilot study and previous work showing poor learning for 3-4 node structures given only covariation data (e.g., Lagnado & Sloman, 2004; Steyvers et al., 2003). Feedback is natural in real world learning and is usually provided by confirming or disconfirming predictions made on the basis of hypothesized causal relations. The feedback in this task can be viewed as a proxy for the outcome of multiple such predictions.

Relative to the no-change condition, the change condition was more likely to retain a belief in the prior concept in their final judgment, despite both groups having chosen this link equally often during the final block of feedback.

The incorrect links A.C and C.B were analyzed together. The interaction between block and condition was not significant, $\chi^2(1)<1$. The main effect of block was significant, $\chi^2(1)=23.02$, $p<0.01$. The main effect of condition was also significant, $\chi^2(1)=4.49$, $p<0.05$, even when considering only the final hypothesis (Fisher's exact test, both ps<0.01). This advantage for the change condition is sensible; they are likely attributable to the extra learning in the change group during the prior learning phase. The scientists' tentative theory regarding the ecosystem implied no causal relation between node C and either A or B. Further, the 30 covariation trials suggested little correlation between these nodes, corroborating the scientists' view.

The correct links D.A and D.B were also analyzed together. The interaction between block and condition was not significant, $\chi^2(1)=2.63$, $p=0.10$. The main effect of block was significant, $\chi^2(1)=15.74$, $p<0.01$. The main effect of condition was not significant, $\chi^2(1)=1.34$, $p>0.10$, though there was a tendency for to change condition to choose these links more often.

Likelihoods judgments Likelihoods judgments were used to infer participants' latent causal representations via model fitting. Causal model theory (CMT; Rehder, 2003) and a version of causal support (Griffiths & Tenenbaum, 2005) were fit to each individual's data. Only the results from CMT are presented here, since they were very similar to the results from causal support.

Causal model theory fits were obtained via maximum likelihood estimation. Each fit yields an estimate of nine free parameters: the strength of each potential causal relation in Figure 2, plus an estimate of the probability that some unobserved background node causes each feature. The fitting routine worked by assuming that the participants' likelihood judgments were guesses about the relative frequency of the snapshots, should they be sampled again. Thus, 100 new snapshots were created with frequencies proportional to the normalized likelihood judgments of each participant. The MLE parameter values were those that maximized the likelihood of the snapshots.

The fits to CMT are presented in Table 1. Fitted background probabilities did not differ between the groups, but estimates of causal strength were different, and in the same direction as the differences present in the hypotheses data. First, the difference for link A.B was significant, $t(46)=2.39$, $p<0.05$, reinforcing the non-significant trend in the hypotheses data. This implies that the change condition represents the prior link stronger than the no-change condition, and this difference is robust for the more implicit measure, the likelihoods, where causal strength is not queried directly.

The conditions did not differ significantly in their representation of the incorrect links A.C and C.B, but the differences in the correct links were marginally significant: the change condition represented the D.A link more strongly, $t(46)=1.82$, $p=0.06$, as well as the D.B link, $t(46)=1.93$, $p=0.08$. In addition, when averaging the strength of the correct links, the difference in conditions was reliable, $t(46)=2.32$, $p<0.05$.

Table 1. Average causal strengths (standard deviations).

	No change	Change	*p*-values
Link A.B	0.06 (0.08)	0.14 (0.14)	0.02
Link A.C	0.10 (0.11)	0.06 (0.08)	0.12
Link C.B	0.07 (0.12)	0.05 (0.06)	0.50
Link D.A	0.20 (0.15)	0.28 (0.16)	0.08
Link D.B	0.25 (0.16)	0.35 (0.20)	0.06
Average of D links	0.22 (0.13)	0.31 (0.14)	0.02

The latter result is in line with a predictions stated earlier that the change group may benefit from the prior learning phase by observing the lack of a correlation between nodes A&C and between nodes C&B. Recall that links A.C and C.B constitute an alternative explanation of the A/B correlation; i.e., that A causes C causes B. Put simply, this set of links may be considered in opposition to the common cause links D.A and D.B in order to avoid over-explaining node B. If so, a reduced belief in the former may increase one's belief in the latter.

The idea that alternative causes compete or inhibit one another has empirical backing (Rehder & Milovanovic, 2007) and is made explicit in recent models of causal induction (Lu et al., 2008). In the current study, to evaluate the relation between choices of links involving the two explanations, I used a hierarchical linear regression with number of correct links chosen as the dependent variable and number of incorrect links as the predictor. The predictor variable was separated into two parts: the participant-level effect (the average number of A.C and C.B links chosen by a participant) and the within-participant effect (the number of links chosen on a given hypothesis minus the participant's average). These variables address different questions: the former asks whether participants who choose more incorrect links on average tend to choose more correct links; the latter asks whether on a given trial the number of incorrect links chosen affects the number of correct links chosen.

The effects of the two predictors were evaluated via model comparison. A model excluding the between-participant effect did not fit worse than a model including both effects, $\chi^2(1)=0.26$, $p>0.10$. However, a model excluding the within-participants effect did fit worse than the model with both effects, $\chi^2(1)=52.03$, $p<0.01$, suggesting that causal links involved in competing explanations inhibit one another on a trial-by-trial basis. To my knowledge this is the first evidence showing that competition occurs at the level of entire explanations (i.e., sets of causes), beyond simply individual causal relations.

Conclusion

The goal of this paper was to show that some aspects of concept restructuring might result from basic causal learning mechanisms, thus bridging the formal approach to concept learning with the conceptual change literature. In a novel learning task, participants first developed a prior conceptual belief and were then prompted to revise that concept through contingency learning. Results showed that the prior learning phase led participants to retain their original belief despite evidence against it but also led to enhanced learning of the target causal structure. That is, despite learning of the target, individuals retained the belief in the prior at the cost of over-explaining. Further evidence showed that when revising one's beliefs, alternative causal explanations are considered in opposition, building on the predictions of recent models for simpler causal structures.

Conceptual change surely involves many processes and representations, only some of which are the learning and revising of causal structures (and within that, only some of which are learning from contingency data; Ahn et al., 1995). For example, people also revise their taxonomic hierarchies (Thagard, 1992) and accrue domain-specific knowledge (Carey, 1985). In addition, full-blown conceptual change presumably requires the restructuring of numerous causal hypotheses and may result in emergent representations inherently unalike the prior beliefs. However, current models incorporate powerful learning mechanisms that are capable of such large-scale changes (Kemp & Tenenbaum, 2008). The hope is that improved cross-talk between formal, empirical, and developmental studies will help to build an integrated view of concept learning and conceptual change.

Acknowledgments

Thanks to Brian Ross, John Hummel, Jose Mestre, Woo-kyoung Ahn, Bill Brewer, Noah Goodman, Frank Keil, Tania Lombrozo, Bob Rehder, Pat Shafto, Dan Navarro, and three reviewers for their very helpful comments.

References

Ahn, W., Kalish, C. W., Medin, D. L., & Gelman, S. A. (1995). The role of covariation vs. mechanism information in causal attribution. *Cognition*, 54, 299-352.

Carey, S. (1985). Conceptual Change in Childhood. Cambridge, MA: Bradford Books, MIT Press.

Cheng, P.W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, 104, 367-405.

Chinn, C. A., & Brewer, W. F. (1993). The role of anomalous data in knowledge acquisition: A theoretical framework and implications for science instruction. *Review of Educational Research*, 63, 1-49.

Chinn, C. A., & Brewer, W. F. (2001). Models of data: A theory of how people evaluate data. *Cognition and Instruction*, 19, 323-393.

diSessa, A. A., & Sherin, B. L. (1998). What changes in conceptual change? *International Journal of Science Education*, 20(10), 1155-1191.

Gopnik, A., Glymour, C., Sobel, D., Schulz, L., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review*, 111, 1–31.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, 51, 354-384.

Fugelsang, J., Stein, C., Green, A., & Dunbar, K. (2004). Theory and data interactions of the scientific mind: Evidence from the molecular and the cognitive laboratory. *Canadian Journal of Experimental Psychology*, 58, 132-141.

Kaplan, A. S., & Murphy, G. L. (2000). Category learning with minimal prior knowledge. *JEPLMC*, 26, 829-846.

Kim, N. S., & Ahn, W. (2002). Clinical psychologists' theory-based representations of mental disorders predict their diagnostic reasoning and memory. *JEP:G*, 131(4), 451-476.

Kemp, C., & Tenenbaum, J. B. (2008). The discovery of structural form. *PNAS*. 105(31), 10687-10692.

Kuhn, T. S. (1962). The Structure of Scientific Revolutions, 1st. ed. Chicago: University of Chicago Press.

Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cognitive Psychology*, 55(3), 232-257.

Lu, H., Yuille, A. L., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2008). Bayesian generic priors for causal learning. *Psychological Review*, 115, 955-982.

Marsh, J. K., & Ahn, W. (2006). Order effects in contingency learning: The role of task complexity. *Memory & Cognition*, 34(3), 568-576.

Murphy, G. L. (2002). The big book of concepts. Cambridge, MA: MIT Press.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, 92, 289-316.

Lagnado, D. & Sloman, S.A. (2004). The advantage of timely intervention. *JEPLMC*, 30, 856-876.

Rehder, B., & Milovanovic, G. (2007). Bias toward sufficiency and completeness in causal explanations. Proceedings of the 29th Annual Conference of the Cognitive Science Society.

Rehder, B. (2003). A causal-model theory of conceptual representation and categorization. *JEPLMC*, 29, 1141-59.

Schulz, L.E., Bonawitz, E. B., & Griffiths, T. (2007). Can being scared cause tummy aches? Naive theories, ambiguous evidence and preschoolers' causal inferences. *Developmental Psychology*, 43(5), 1124-1139.

Steyvers, M., Tenenbaum, J., Wagenmakers, E.J., Blum, B. (2003). Inferring Causal Networks from Observations and Interventions. *Cognitive Science*, 27, 453-489.

Thagard, P. (1992). Conceptual Revolutions. Princeton University Press.

Vosniadou, S. (2008). International handbook of research on conceptual change. New York: Routledge.

Wellman, H. M., & Gelman, S. A. (1992). Cognitive development: Foundational theories of core domains. *Annual Review of Psychology*, 43, 337-375.

Analysis of the Variability of Three-Dimensional Spatial Relations in Visual Short-Term Memory

Carsten Winkelholz (carsten.winkelholz@fkie.fraunhofer.de)
FGAN, Neuenahrer Straße 20
53343 Wachtberg-Werthhoven, Germany

Michael Kleiber (michael.kleiber@fkie.fraunhofer.de)
FGAN, Neuenahrer Straße 20
53343 Wachtberg-Werthhoven, Germany

Christopher Marc Schlick (schlick@iaw.rwth-aachen.de)
RWTH Aachen University, Bergdriesch 27
D-52062 Aachen

Abstract

In a laboratory experiment, 13 participants reproduced from memory the position of a sphere relative to a second landmark sphere located on the viewing axis of the observer. The relative location of the second sphere varied both laterally and in depth. The stimuli were generated on a stereoscopic display. The paper focuses on the analysis of the structure of the noise in the reproduced object locations, this structure reflecting the mental representation of the stored spatial relations. The results showed that the spatial location of the landmark sphere affects the variability of the reproduced object locations. In particular, the variability in the frontoparallel plane increases with the length of the depth component of the spatial relation. This finding can be interpreted in two ways. First, spatial acuity in perception decreases, or second, participants encode sensory information by transforming it into a mental spherical coordinate system. Both interpretations are discussed.

Introduction

McNamara (2003) proposed that locations are memorized in egocentric and allocentric coordinate systems. Allocentric coordinate systems define locations with respect to objects in the environment. We believe that shifts of attention between several locations in space define the reference axes and planes of local allocentric coordinate systems within which the spatial relations are encoded. This assumption is consistent with the idea that locations are encoded by intrinsic frames of references (Mou & McNamara 2002; Schmidt 2004). These intrinsic reference frames would result naturally from salient landmarks of the scene that attract attention. The structure of the variability in the reproduced locations provides essential information about the nature of the allocentric reference systems and reveals the dimensions in which attributes of the location had been encoded. Only a few reports in the literature have provided a systematic investigation of the dispersion of locations recalled from memory. The most frequently cited work in this field is by Huttenlocher, Hedges, and Duncan S. (1993), who conducted an experiment in which the participants had to reproduce locations within a circle, the observed distribution of which was consistent with encoding the locations relative to the center of the circle in terms of the distance from the center and the polar angle. Furthermore, they found systematic distortions of the reproduced polar angles for locations near the virtual horizontal and vertical lines that divide the circle into quadrants. The participants misplaced the locations toward a central location in each quadrant. Huttenlocher et al. proposed a stochastic model based on hypothesized probability density functions for the recall of the locations from memory. Based on these findings Werner & Diedrichsen (2002) investigated the time course of the memory distortions for the location of a dot in relation to two horizontally aligned landmarks. These works and the work of McNamara (2003) complemented each another, if the recall of locations from memory is described by probability density functions according to the dimensions of the allocentric reference systems.

The aims of the experiment described in this paper are twofold. First to confirm basic parameters of the noise in the mental representation reported in the literature, which we have already used to model phenomena in memorizing object locations in graphical structures (Winkelholz & Schlick 2007a) and for symmetry detection (Winkelholz & Schlick 2007b). Second to gain insight into the structure of the probability distribution of basic three dimensional spatial relations reproduced from memory. Especially, we are interested if subjects encode the stimuli on the basis of values of the perceived attributes or if they transform the perceived attributes into a mental coordinate system.

Experiment

Within the experiment participants reproduced random virtual object locations on three predefined frontoparallel planes. If the object location is represented mentally by a distance and a solid angle relative to a landmark location, then the variability in the lateral coordinates of the

reproduced object locations should increase with their relative distance in depth from the landmark location. If the variability is independent of the reproduced location, the latter's mental representation might simply be its perceived projection on the screen and a relative distance in depth that is perceived by disparity and the visual angle of the circumference. In general, an increase in the variability of the lateral coordinates might be just the result of visual perception. When the visual system focuses on a location in three-dimensional space through convergence, only the points contained inside Panum's fusional area near the horopter are fused into a single image. Therefore, outside of Panum's fusional area oculomotoric sensor information will additionally be used by the visual system to determine the spatial relation. Accommodation should have no effect on spatial acuity, since the stereoscopic stimuli were generated synthetically on a display at a fixed distance from the observer.

Method

Participants

Thirteen volunteers (11 male, 2 female, average age 27), who were recruited from the staff of our institute, took part in the experiment. All participants had normal or corrected-to-normal vision.

Apparatus and Stimuli

The experimental task environment was generated on a Windows workstation equipped with a NVIDIA Quadro graphics card. The subjects used a spacemouse and the standard keyboard to provide input information. The spacemouse, a three-dimensional interaction device with six degrees of freedom, contains a controller cap that can be pushed, pulled and twisted in any direction. The subjects used the spacemouse to control the spatial movement of the object during the response stage. The stereoscopic images were rendered at 120 Hz on a 21″ CRT monitor and a resolution of 1280×1024 pixels. The images for the left and right eyes were separated by shutter glasses, which meant that the frame rate per eye was 60 Hz. The scene was rendered using antialiasing (16 times provided by the driver) to increase the visual spatial resolution and thereby enhance perception of the disparity The monitor screen was located 60 cm in front of the subject. The spheres were displayed using the user-centric projection method that is commonly employed in virtual environments such as caves and workbenches (Cruz-Neira, Sandin & DeFanti 1993). Points in object space are projected onto the screen according to the positions of the user's eyes. Each eye perceives the points on the surface of a virtual object from the correct solid angle as if the object was actually present. In other words, the disparity of the displayed objects on the screen and the viewing angle of the projected size of the spheres were the same as if real spheres had been placed at these coordinates.

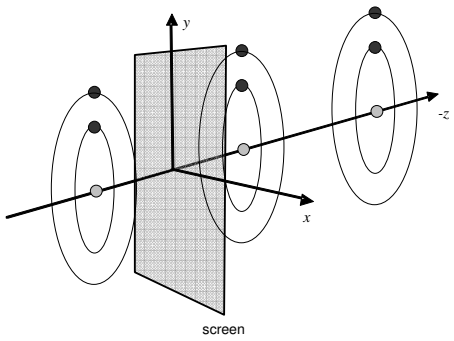


Fig. 1: Illustration of the experimental setup. The lower part of the figure shows two cross-sections of the display setup, the first from the side and the second from above.

To accomplish this, subjects were advised to sit in an appropriate position so that their head was within the range of the parameters used in the projection model. In the following, the stimulus parameters are reported in virtual coordinates according to this user-centric projection model. All spheres were displayed to appear on one of three virtual planes. The screen is defined to be at z = −60 cm[1]. The first plane was –1.5 cm [z = −58.5 cm] in front of the screen, and the second and third planes were located 1.5 cm [z = −61.5 cm] and 4.5 cm [z = −64.5 cm] behind the screen, respectively. Hence, the disparities shown on the display were –9.5′ for z = –58.5 cm, 9.1′ for z = –61.5 cm, and 26′ for z = –64.5 cm. The diameter of the spheres was 1 cm and the corresponding visual angles of the displayed size were 58.8′, 55.9′ and 53.3′, respectively. The landmark sphere was always displayed on the z axis. The sphere whose location had to be memorized was positioned at two distinct distances from the center axis. The radii of the circles were chosen so that the viewing angle of the distance to the center was constant across different virtual planes. The visual angle, $\alpha_{xy} = \tan^{-1}(\sqrt{x^2 + y^2} / z)$, of the lateral distance was 2.9° for the inner circle and 4.8° for the outer circle. The associated distances of the projected locations on the screen from the center were 3.0 cm for the inner circle and 5.0 cm for the outer circle. This procedure was used to ensure that effects in the *xy*-component of reproduced spatial relations did not result simply from different distances on the retina, which would make reasoning about the effects more difficult. On the other hand, this procedure makes it more difficult to analyze the effects in the displayed, virtual, three-dimensional space. However, variation in the radii of the test stimuli in virtual space was quite small and only resulted in additional noise that was identical for each factor level of Δz and therefore did not affect the main effects. In the virtual space, the range of the radii was [2.8 cm, 3.1 cm] for the inner circle and [4.9 cm, 5.4 cm] for the outer circle.

[1] The x, y, and z axes form a right-handed coordinate system.

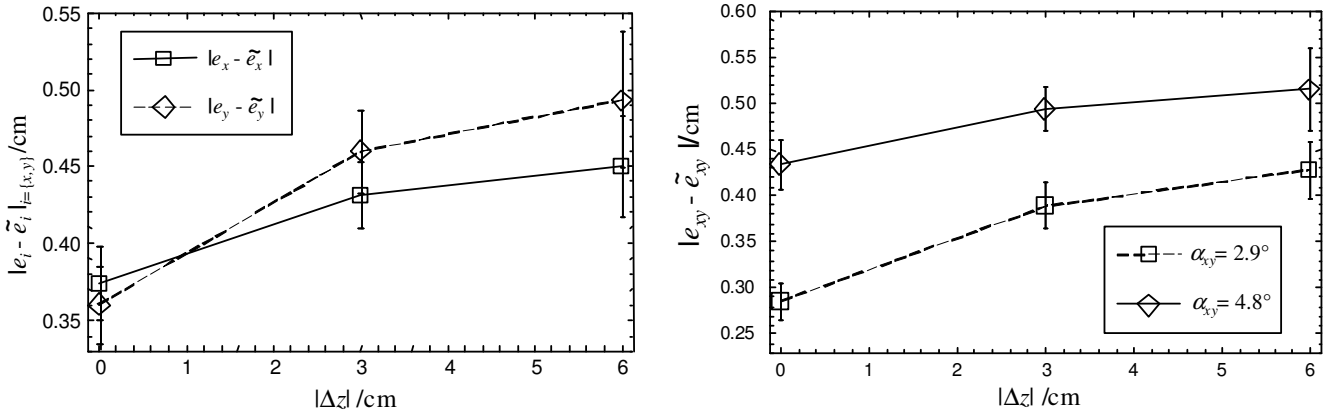


Fig. 2: (a) Absolute deviation of e_x and e_y as functions of |Δz|.
(b) Absolute deviation of e_{xy} as a function of |Δz| parameterized by the distance to the center axis.

Procedure

In each experiment the subject's task was to reproduce the location of one sphere relative to a second sphere. All participants performed training sessions to familiarize themselves with the stereoscopic information display and the spacemouse. Each experiment used a 3×3×2 within-participants design. The first factor was the virtual frontoparallel plane on which the landmark sphere was located. The second and third factors indicated the virtual frontoparallel plane and the eccentricity of the location that had to be memorized, respectively. The polar angle of the location on the circle in question was randomized by a uniform distribution. All object configurations were tested in a randomized order. At the beginning of each trial, both spheres were displayed for one second, followed by a blank screen shown for two seconds. Finally, the landmark sphere was displayed at its previous location and a second sphere was shown at the default location, the origin. This second sphere had to be moved to the memorized location using the spacemouse. When the subject was confident that the second sphere was located at its remembered location he/she confirmed the location by pressing the spacebar on the keyboard. After a blank screen had been displayed for a short time, a new trial containing new locations for the spheres followed. For movement of the sphere, the translation of the controller cap was modeled as a three-dimensional Cartesian vector. Because the controller cap can be moved along all dimensions simultaneously, this Cartesian vector can point in any direction and has no preferred movement along a particular axis. The sphere moved in the virtual display space in the direction of this vector with a speed proportional to the vector norm.

Dependent Variables

In the following, the triplet (x_0, y_0, z_0) represents the coordinates of the landmark location, (x, y, z) are the coordinates of the location that had to be memorized, and (x', y', z') are the coordinates of the location that was reproduced by a subject. In this study, the relative distances of the locations to the landmark location are of major interest: $\vec{v} = (x - x_0, y - y_0, z - z_0)^T$, $\vec{v}' = (x' - x_0, y' - y_0, z' - z_0)^T$ To test the hypothesis that relative depth is encoded independently of relative lateral location, we first investigated the response errors, $\vec{e} = \vec{v}' - \vec{v}$, in Cartesian coordinates. The reliability of the memorized location is reflected in the variability of the responses. By itself, the error vector reflects systematic distortions in the mental representation. Without a systematic component of distortion in the mental representation, the mean error equals zero. The variability of the errors is identical to the variability of the responses. We used the average absolute deviation to measure variability and the median to measure central tendency.

Results and Discussion

All trials on which the distance between the reproduced location and the correct location was larger than the distance between the correct location and the landmark location were considered as outliers. Since the exclusion of outliers resulted in empty cells for two of the participants, their data were excluded from further analysis. There were 6.9% outliers in the remaining group of 11 participants.

Cartesian coordinate system

For each factor level, the mean response error, $\vec{e}$, was determined. Using these means, the absolute deviations of each component, x and y, of the response error were calculated. The absolute deviation was analyzed using a repeated measures ANOVA with $|\Delta z| = |z - z_0|$ (0 cm, 3 cm, 6 cm), the visual angle of the lateral distance, α_{xy}, (2.9°, 4.8°) and component (horizontal (x), vertical (y)) as the within-subject factors. The ANOVA results showed that the absolute deviation varied systematically with |Δz| ($F(2,20) = 5.88$, $p < .01$, , $\eta_p^2 = .37$), its value being smaller for |Δz| = 0 cm (*Mean* = .36 cm, *SEM* = .04 cm) than for |Δz| = 3 cm (*Mean* = .44 cm, *SEM* = .03 cm) and |Δz| = 6 cm (*Mean* = .47 cm, *SEM* = .06 cm). No significant difference was found between |Δz| = 3 cm and |Δz| = 6 cm. The analysis revealed neither a main effect of the component type ($F(1,10) = .60$,

p = .46) nor an interaction effect of component type and $|\Delta z|$ ($F(2,20)$ = 1.00, p = .38). The absolute deviation of e_{xy}, parameterized with α_{xy}, is plotted in Fig. 2b. The absolute deviation varied systematically with α_{xy} ($F(1,10)$ = 19.5, p < .001, η_p^2 = .66), and the interaction of $|\Delta z|$ and α_{xy} was not significant (p > .5). The absolute deviation was smaller for α_{xy} = 2.9° (*Mean* = .37 cm, *SEM* = .04 cm) than for α_{xy} = 4.8° (*Mean* = .48 cm, *SEM* = .04 cm).

Spherical coordinate system

To analyze the variability of the responses using a spherical coordinate system, both the length and the zenith angle were calculated for all spatial relations that had been analyzed. Since it was assumed that the reference axis points in the same direction as the spatial relation, the zenith angle only varied from 0° to 90°. Based on the grouping of these two values, factor levels were defined for the zenith angle and the lengths of the tested spatial relation. The defined factor levels are shown in Fig. 3.

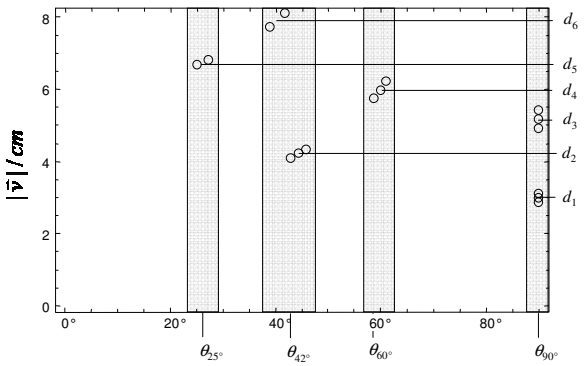


Fig. 3: Factor levels of the stimuli used for the analysis.

The factor levels with different lengths for a single zenith angle are of special interest. This is the case for $\theta_{42°} \approx 42°$ and $\theta_{90°} = 90°$. Therefore, if the response errors are examined in spherical coordinates, the absolute deviations of the angles should be identical for different lengths of the spatial relation. To verify this, the absolute deviations of zenith and azimuth angle for each response were calculated. A three-way repeated-measures ANOVA was conducted using Euclidean length and zenith angle of the tested spatial relation and the angular component of the reproduced spatial relation as within-subject factors.

There was no significant effect of the length of the tested spatial relation on the absolute response deviation ($F(1,10)$ = 1.52, p = .246). Therefore, the absolute deviation of the reproduced spherical angles was also calculated for $\theta_{25°}$ and $\theta_{60°}$. In Fig. 4a, the absolute deviations of the reproduced angles are plotted for all zenith angles under study. The absolute deviations of the reproduced zenith angles increased for smaller zenith angles of the tested spatial relation, whereas this dependence seemed to be weaker for the reproduced azimuth angle.

A two-way repeated-measures ANOVA showed significant effects for the angular component ($F(1,10)$ = 15.30, p < .005, η_p^2 = .54) and the zenith angle of the tested spatial relation ($F(3,30)$ = 6.46, p < .01, η_p^2 = .39). The interaction of these two factors was not significant ($F(3,30)$ = 2.13, p = .12, η_p^2 = .18). The increase in the absolute deviation of the reproduced azimuth angle was smaller ($\theta_{90°}$: *Mean* = 5.14°, *SEM* = .85; $\theta_{25°}$: *Mean* = 7.12°, *SEM* = .97) than the increase in absolute deviation of the reproduced zenith angle ($\theta_{90°}$: *Mean* = 9.15°, *SEM* = .79; $\theta_{25°}$: *Mean* = 15.53°, *SEM* = 1.96).The strong dependence of the absolute deviations in the reproduced zenith angles on the tested zenith angle contradicts the predictions of a pure spherical geometry for the mental representation. Therefore, as a next step the absolute deviation of the reproduced length of the spatial relation was analyzed. For each defined factor group, the tested lengths have a given absolute deviation, which need to be considered in the analysis. For a spherical geometry it must be expected that the absolute deviations of the reproduced lengths increase linearly. In contrast, the analysis showed a disordered picture for the reproduced lengths, the mean of the reproduced length being smaller for d_5 than for d_4 (Fig 4b). A one-way repeated-measures ANOVA revealed no significant difference for these two groups ($F(1,10)$ = 2.93, p = .12, η_p^2 = .23).

Therefore, the mean of the reproduced length for d_5 (*Mean* = 5.45 cm, *SEM* = .19 cm) was at least equal to or possibly smaller than that for d_4 (*Mean* = 5.73 cm, *SEM* = .13 cm). However, groups d_4 and d_5 also differed in zenith angle for the tested spatial relation. In d_4, the mean zenith angle was 42°, whereas for d_5 the mean zenith angle was 90°. For d_5, the spatial relation had no depth component, and the absolute deviations did not increase with the length of the tested spatial relation.

The absolute deviation for the tested length for d_3 was significantly smaller than that for d_2 ($F(1,10)$ = 13.6, p < .005, η_p^2 = .58). Again, both groups also differed in zenith angle (d_2: θ = 20°, d_3: θ = 42°), and consequently by the fraction of the depth component. These findings suggest an independent analysis of the depth and lateral components of the length of a spatial relation. Therefore, the data are grouped by $|\Delta z|$ and the length of the *xy*-component of the tested spatial relations. The absolute deviations increased with the length of the related length. A one-way repeated-measures ANOVA showed that this effect was significant for the *z*-component ($F(2,20)$ = 35.1, p < .001, η_p^2 = .78) and the *xy*-component ($F(1,10)$ = 20.3, p < .001, η_p^2 = .67). Since the absolute deviations from the given spatial relations also increased itself for the *xy*-component, an additional two-way repeated-measures ANOVA was performed on pooled data from the tested spatial relations and the reproduced spatial relations. This analysis, which included reproduced vs. original spatial relations as an additional factor, revealed a significant interaction between reproduced vs. original spatial relation and length ($F(1,10)$ = 6.59, p = .028, η_p^2 = .39). This interaction indicated that an additional increase in the absolute deviation results from the

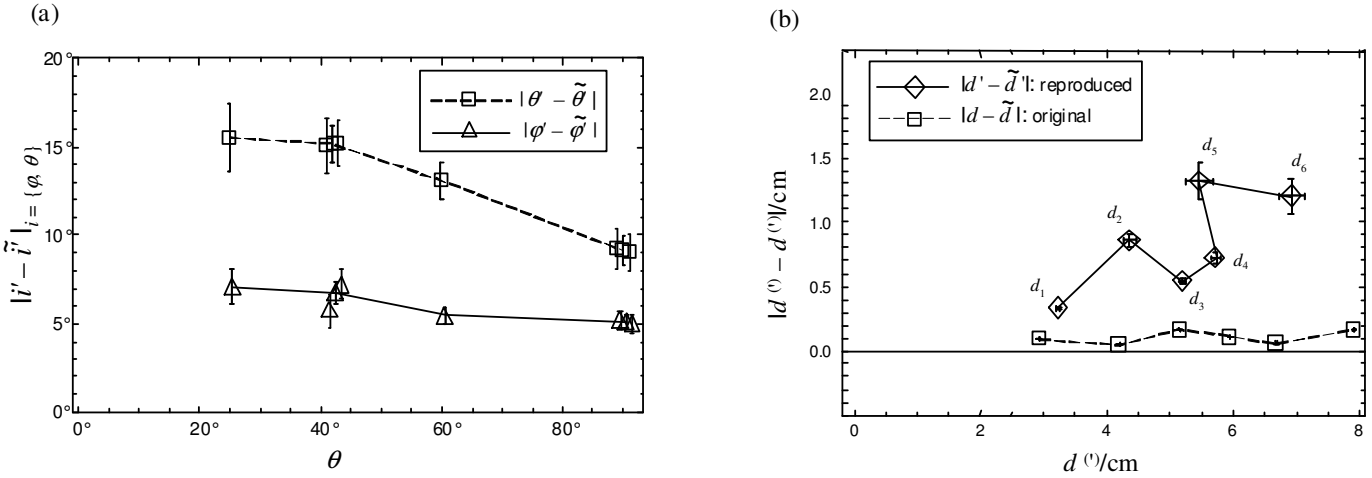


Fig. 4: (a) Absolute deviation of the reproduced spherical angular components as a function of the zenith angle of the tested spatial relation. The multiple measure points at $\theta_{42°}$ and $\theta_{90°}$ show the absolute deviations for the corresponding distance groups. (b) Mean and absolute deviation of the reproduced and tested Euclidean distances as a function of the corresponding mean.

mental representation. Notably, the absolute deviations of the *z*-component appeared to increase linearly with |Δz| for the tested spatial relation but not with the reproduced length. Furthermore, the z-component shrinks in memory. The strengths of the growth and shrinkage depended on |Δz|. A more detailed analysis is out of the scope of this paper

General Discussion

The results of the experiment showed that the variability of a location reproduced from visual spatial memory is influenced by the relative distance in depth to a landmark. With increasing distance in depth, not only did the variability of the reproduced depth component of the distance increase, but the variability of the reproduced lateral location also increased. The effect of landmarks on locations reproduced from memory generally indicates that participants include spatial relations between the location and the landmark in the encoded location. The structure of the variability of the reproduced locations provides insight into the mental representation. For an analysis a detailed model should describe the actual information processing steps that transform sensory information into a cognitive representation and then into a reproduction. Such a model can be greatly simplified if noise contained in the mental representation is much greater than the noise contained in visual sensory information. In this case, the noise from sensory information can be neglected. For two dimensional stimuli, visual acuity was much higher than the variability of the reproduced locations. For example, the visual acuity at an eccentricity of 5° is about 3″. Under the assumption that the landmark location can be assessed with a resolution of 1″, the lateral direction of a location relative to the landmark location should be discriminated by $2 \cdot \tan(5°/(4″/2)) \approx 0.08°$, which is much lower than the usually obtained variability of directions reproduced from memory. Similar arguments apply for the reproduced lateral distance to the landmark location. The lateral noise parameter σ_{φ} and $\sigma_{\alpha_{xy}}$ determined from the data can be compared to values reported in literature. Huttenlocher et al. (1991) reported $\sigma_{\varphi} = 10°$, which is somewhat higher than the value of $\sigma_{\varphi} = 6.3°$ found in our data. In contrast, we found $\sigma'_{f_\alpha} = 0.11$[2] for the standard deviation to reproduce radial distance, which is larger than $\sigma_{f_\alpha} = 0.025$, the value reported by Huttenlocher et al. (1991). This difference may be caused by the fact that participants in the experiments of Huttenlocher et al. had to estimate locations within a circle. To do so, initially they had to estimate the center of the surrounding circle as the landmark location, which adds noise to the direction, whereas the radial component could be estimated more efficiently by using more than one landmark located on the circumference of the circle. For three dimensional stimuli, the assessment of sensory acuity is much more complex than it is for two dimensional stimuli. There are several sources of sensory information that can be exploited by the visual system to deduce information about depth: disparity, accommodation, and vergence. To the best of our knowledge, the quantity representing the effect of an increase in disparity on lateral spatial resolution has not been described in the depth perception literature. In contrast, the dimensions of Panum's fusional area have been well studied (Kenneth & Ogle 1952). Additional studies have focused on the dependence of the stereo acuity on eccentricity (Rawlings & Shipley 1969) and the effect of object size on stereoscopic spatial depth acuity (Schlesinger & Yeshurun 1998) A decrease in spatial acuity in the lateral dimensions due to increasing disparity is to be expected, because double images are perceived outside Panum's fusional area. However, we believe that the additional noise from disparity is less than the increase in noise that was found in the data. Furthermore, because the stimuli had horizontal disparity, this noise should only affect the horizontal component of the

[2] To be compliant to Weber-Fechner-Law the standard deviation of reproduced eccentricity scales linear with the eccentricity of the actual memorized visual angle α_{xy} is given by: $\sigma_{\alpha_{xy}} = \sigma_{f_\alpha} \alpha_{xy}$

lateral location and not the noise in the vertical component, which surprisingly increased by similar amounts. Nevertheless, the analysis of the absolute deviation of the reproduced distances as a function of the distances examined in the experiment showed that the depth component of distance was crucial, since the variability was much greater in depth than it was in the lateral dimensions. This is consistent with the findings of Norman et al. (1996), who observed that participants are highly sensitive to small differences in the length of lines presented in the frontoparallel plane, while the sensitivity decreases by an order of magnitude when the line segments are presented at random orientations in depth. In case of a mental representation of the spatial relation in a spherical coordinate system a model should include this noise in the perception of depth, while the zenith angle is deduced from this noisy depth component. The dependence of noise in the depth component on eccentricity, where landmark location and the to-be-reproduced location are in the same frontoparallel plane ($\theta = 90°$), was similar to values reported in the literature. It is known that stereoscopic acuity is a decreasing function of eccentricity. Rawlings and Shipley (1969) reported a stereo acuity of 21″ at the point of focus and 155″ at an eccentricity of 4°. If 25% is assumed to be the threshold of the just-noticeable difference, an interpolation of the data reported in this paper will predict a stereo acuity of 221″ at an eccentricity of 4°. On the one hand, this finding does not deliver a new argument that spatial relations are mentally represented in a spherical coordinate system, since the additional noise might simply be the result of the subject's carelessness when adjusting the stimulus to the remembered location. Yet on the other hand, this finding does not contradict the argument that the noise contained in the mental representation results from noisy perception. A model assuming a mental representation in a spherical coordinate system would explain both effects—the increase in depth variability with eccentricity and the increase of lateral variability with relative distance in depth—using only one noise parameter for the zenith angle σ_θ, whereas a model considering independent dimensions for the depth and the lateral location needed two parameters: one noise parameter for the lateral projected distance in dependence on the depth component ($\sigma_{f_\alpha}(\Delta z)$), and a second noise parameter for the noise in the depth component in dependence on the eccentricity of the spatial relation ($\sigma_{\Delta z}(\alpha_{xy})$). In future research the mathematical modeling of human performance variability using probability density functions would clarify the underlying assumptions regarding dependencies between spatial attributes. The resulting parameterized models could be used to describe the recollection of locations from memory. The distortions at categorical boundaries emerged naturally at the boundaries of the probability density functions. Furthermore, the results of this study should be generalized. In the current experiment, the viewing axis was a natural choice for the polar axis of the spherical coordinate system, since there was only one landmark sphere present. If there are two landmark spheres, we suggest that the line connecting the two spheres serve as the polar axis of the mental representation.

Acknowledgment

This work was supported in part by the German Research Foundation DFG under the Cluster of Excellence "Integrative Production Technology for High-Wage Countries".

References

Cruz-Neira, C., Sandin, D.J., & DeFanti, A.T. (1993). Surround-screen projection-based virtual reality: The design and implementation of the CAVE. *Proceedings of SIGGRAPH*, 93, 135–142.

Huttenlocher, J., Hedges, L.V., & Duncan S. (1991). Categories and particulars: Prototype effects in estimating spatial location. *Psychological Review*, 98(3), 352–376.

Kenneth, N., & Ogle, K. N. (1952). Disparity limits of stereopsis, *Archives of Ophthalmology*, 48(1), 50–60.

McNamara, T. P. (2007). Commentary: The nature and development of spatial reference systems. In J. M. Plumert & J. Spencer (Eds.), The emerging spatial mind (pp. 104-113). London: Oxford University Press.

Mou, W., & McNamara, T.P. (2002). Intrinsic frames of reference in spatial memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28, 162–170.

Norman, J.F., Todd, J.T., Perotti, V.J., & Tittle, J.S. (1996). The visual perception of three-dimensional length. *Journal of Experimental Psychology: Human Perception and Performance*, 22(1), 173–186.

Rawlings, S.C., & Shipley, T. (1969). Stereoscopic acuity and horizontal angular distance from fixation. *Journal of the Optical Society of America*, 59(8), 991–993.

Schlesinger, B.Y., & Yeshurun, Y. (1998). Spatial size limits in stereoscopic vision. *Spatial Vision*, 11(2), 279–293.

Schmidt, T. (2004). Spatial distortions in visual short-term memory: Interplay of intrinsic and extrinsic reference systems. *Spatial Cognition & Computation*, 4(4), 313–336.

Werner, S., & Diedrichsen, J. (2002). The time course of spatial memory distortions. *Memory & Cognition*, 2002, 30(5), 718-730.

Winkelholz, C.; & Schlick, C. (2007a). Modeling human spatial memory within a symbolic architecture of cognition, In Barkowsky, Th., Knauff, M., Ligozat, G., Montello, D.R. (Eds.), *Spatial Cognition V: Reasoning, Action, Interaction*, (pp. 229-248), Berlin: Springer.

Winkelholz, C.; & Schlick, C.: (2007b) Bridging psychophysics and cognitive engineering in visual perception, *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, 07*, 2520-2527.

Spatial Factors in Social and Asocial Learning

Alexander Metz (alexander.metz@mail.mcgill.ca)
Department of Psychology, McGill University, 1205 Avenue Docteur Penfield
Montreal, QC H3A 1B1 Canada

Thomas R. Shultz (thomas.shultz@mcgill.ca)
Department of Psychology and School of Computer Science, McGill University, 1205 Avenue Docteur Penfield
Montreal, QC H3A 1B1 Canada

Abstract

Asocial learning is a mechanism by which innovations develop, and social learning is a mechanism by which innovations spread. Penetration of an innovative behavior through a population is measured by the proportion of the population that possesses the innovation. Via agent-based modeling, we examine innovation diffusion with agents learning and interacting in space. Simulations show that innovation spread systematically deviates from differential equations of the proportion of the population that has the innovation. Mediation analysis confirms that boundary surface length of groups having the innovation accounts for these spatial effects. Proportion of asocial innovative learners increases surface length which, in turn, increases social imitative learning.

Keywords: Social learning; asocial learning; imitation; innovation; spatial simulation; surface length; mediation analysis; agent-based modeling.

Introduction

Since Darwin's theory of evolution, researchers have sought to understand how organisms adapt to their environment to maximize their reproductive potential. In addition to biological evolution, some lasting adaptations manifest themselves through animal phenotypes with no genetic changes. Innovative behaviors allow relatively quick adaptation to rapidly changing environments, and can spread and persevere (Laland, Boyd, & Richerson, 1996; Reader & Laland, 2003). For our purposes, *innovation* refers to an adaptive behavior pattern with relative novelty.

Innovative behaviors can be acquired through either asocial or social learning. In asocial learning, an innovation is acquired through individual experience. In social learning, an innovation is acquired from a conspecific demonstrator (Heyes, 1994). Because individual discoveries are costly, they may occur in only a few key individuals through asocial learning and then diffuse through a population via social learning (Laland, Boyd, & Richerson, 1996). The dynamics of innovation diffusion and imitation are an important aspect of collective social cognition and behavior, and thus relevant to integrative cognitive science (Goldstone & Gureckis, 2009).

However, patterns of sequential spread in a population resembling those predicted by some models of social learning can result from asocial processes alone. It is not always clear which type of learning drives innovative acquisitions, so it is thus important to disentangle social from asocial learning. One useful method is diffusion curve analysis, or DCA (Reader, 2004; Franz & Nunn, 2009). Diffusion is the change in frequency of an innovative behavior in a population over time. In DCA, the shape of the curve is used to infer whether social or asocial learning is the mechanism of diffusion.

Our purpose is to simulate the spatial diffusion of innovation and compare the results to DCA predictions. Although there is a rich literature on learning in laboratory experiments, understanding of how social learning occurs in the wild is limited.

S-shaped logistic curves are predicted by DCA to characterize social learning. If the amount of social learning at a given time step is proportional to both the number of possible demonstrators and the number of possible learners, then it can be obtained as the product of the proportion of the population that knows the innovation and the proportion of the population that does not know the innovation (Laland, Boyd, & Richerson, 1996). This corresponds mathematically to the differential equation:

$$\Delta u = R_S u(1-u) \quad (1)$$

where R_S is a constant rate of social learning, and u is the variable proportion of the population with the innovation.

In a population of only innovators, assuming no social learning, the following differential equation applies:

$$\Delta u = R_I (1-u) \quad (2)$$

where R_I is the rate of innovation. As more innovators learn, the number of naïve innovators decreases in a decelerating curve (Franz & Nunn, 2009).

Social and asocial learning are not mutually exclusive. In an analysis of data from research by Hinde and Fisher (1949) on innovation spread in birds, Lefebvre (1995) concludes that milk-bottle-opening likely spread by some form of social learning from many unique points of origin. This is supported by evidence that some birds open bottles spontaneously without any prior experience with bottles or demonstrators (Sherry & Galef, 1984). Thus, asocial learning can occur alongside social learning, and Equations 1 and 2 can be summed to accommodate this:

$$\Delta u = R_S u(1-u) + R_I (1-u) \quad (3)$$

Equation 3, however, applies only to a population where every member is capable of being an innovator and a social learner. It may be more realistic to assume that only a certain proportion of the population is capable of either of

these things. No explanatory power is lost in making this assumption as these proportions can be set to 1, and the resulting model is only slightly less parsimonious. To accommodate this variation in ability, Equation 3 can be modified by multiplying the innovation and social learning parts of the equation by their corresponding proportions, *I* and *S*, respectively:

$$\Delta u = SR_S u(1-u) + IR_I(1-u) \qquad (4)$$

We refer to Equation 4 as the DCA equation. Based on the proportion of social learning compared to asocial learning, this differential equation generates a curve with either a logistic shape (greater social learning) or a decelerating shape (greater asocial learning). The DCA equation has been applied in various experimental contexts, including the diffusion of innovations in humans from peer and media influences (Lekvall & Wahlbin, 1973) and bystander effects in the diffusion of foraging techniques in pigeons (Laland, Boyd, & Richerson, 1996).

The DCA equation relies on one key variable: the proportion of the population that knows the innovation. This proportion thus serves as both the dependent and independent variable in the differential equation. Here we test the results of spatial simulations against the predictions of the DCA equation. The diffusion of innovation is in part a spatial process, a fact captured by the simulations, but not by the DCA equation. We answer several questions. What are the essential differences between asocial and social learning and how can these two types of learning be identified in wild populations? Does the DCA equation account for all aspects of these issues, or are other approaches required? Are these features realistic, or are they artifacts of abstract simulations?

To explore the spatial diffusion of an innovation, our simulations create a two-dimensional space containing agents. Depending on their genotype, agents can be innovators and/or social learners. Parameters of the simulation include the proportions of innovators and social learners, just as the DCA equation uses these factors as variables. Comparing the rate of learning in the simulation to the rate of learning predicted by the DCA equation could provide insight into any potential spatial factors affecting innovation diffusion.

Methods

The simulation is set on a torus, a 25 by 25 lattice in which each edge touches the opposite edge. Each of the 625 tiles contains one agent with on/off genes for innovation and social learning.

Agents with an activated innovation allele can spontaneously discover the innovation at a fixed innovation rate of .025. Agents with an activated social-learning allele can copy the innovation from their neighbors: for every adjacent neighbor that knows the innovation, a social learner's chance of learning the innovation increases by .25. The ten-fold difference between the success of social and asocial learning is based on an assumption of differential learning costs: if asocial learning has a greater cost and requires more resources than social learning, it should occur at a slower rate than social learning. The simulation experiment assigns genes to individual agents probabilistically depending on the proportion of social and asocial learners specified in simulation parameters. The simulation runs for 80 learning cycles, recording agent behavior, the times at which agents learn, and the neighbors from whom they learn if the learning is social.

The effect of number of innovators was investigated in simulations with the proportion of innovators ranging from .05 to 1.0, holding the proportion of social learners at 1.0. The effect of number of social learners was studied with simulations varying proportion of social learners ranging from 0 to 1.0, holding the proportion of innovators at 1.0.

Results

Figures 1-6 plot the change in the proportion of the population that knows the innovation over time, averaged across five runs. Figures 1 and 2 depict the results from varying the proportion of the population with the asocial learning allele when the whole population has the social learning allele. Figure 1 shows predictions of the DCA equation, and Figure 2 presents simulation results.

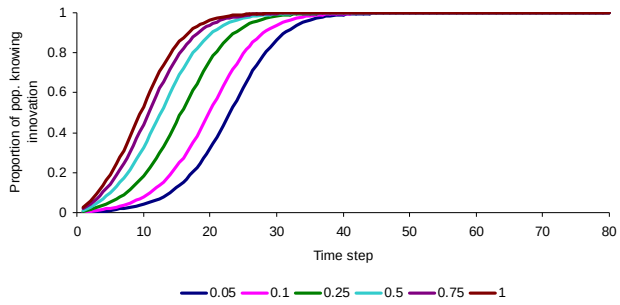


Figure 1: Diffusion curves predicted by DCA equation with asocial learning rate = .025, social learning rate = .25, proportion of social learners = 1, and the proportion of asocial learners varying from .05 to 1.

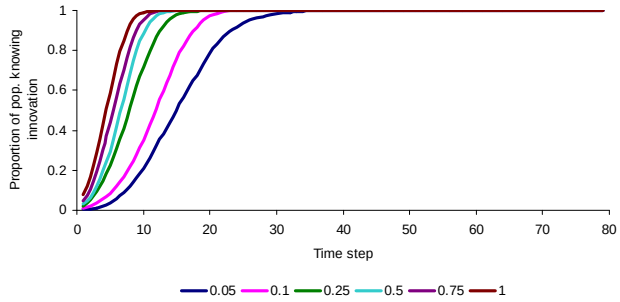


Figure 2: Simulations with asocial learning rate = .025, social learning rate = .25, proportion of social learners = 1, and proportion of asocial learners varying from .05 to 1.

These results reveal subtle but noticeable differences between the DCA equation and the simulations. For Figures 1 and 2, the whole population is capable of social learning;

what changes across curves is the proportion of the population capable of asocial learning. In Figure 1, the curves determined by the DCA equation appear more parallel than they do in the simulation results of Figure 2.

We can understand these differences by considering the DCA equation itself. This equation (4) has a social learning component (left half) and an asocial learning component (right half). Recall that the DCA equation's key variable is the proportion of the population that knows the innovation. At the beginning, the innovative behavior is introduced into the population by asocial learning, so the proportion of the population that can do asocial learning has a large effect as seen in Figure 1. Because this proportion of asocial learners is different in every curve, the curves differentiate quickly. However, as the proportion of the population that knows the innovation increases, the social learning component of the DCA equation has a greater effect. Because all of the curves in Figure 1 have the same social learning settings, with the proportion of social learners S set to 1 and the rate of social learning R_S set to .25, their learning rates are very similar after this original differentiation, causing the observed parallelism. Thus, the parallel nature of the equation-produced curves in Figure 1 is a direct consequence of using the proportion of the population that knows the innovation as the key independent variable.

The lack of parallelism in simulation curves can be quantified by examining the maximum learning slope for each curve, which represents the amount of learning when u, the proportion of the population that knows the innovation, equals .5. This is the point that maximizes the product $u(1-u)$ and thus also maximizes innovation spread according to the DCA equation. Figure 3 presents mean maximum slopes of diffusion curves as a function of the proportion of the population with the innovation allele.

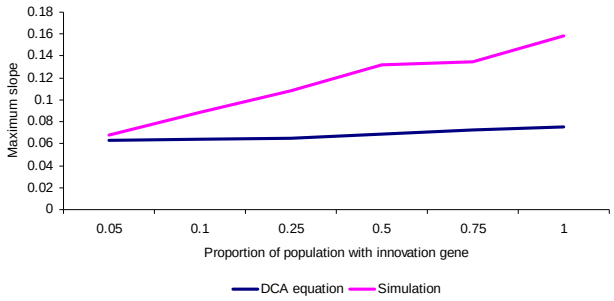


Figure 3: Maximum slope of curves (where u = .5) from the DCA equation and the simulations.

Figure 3 indicates that the maximum slope of each curve from the DCA equation is relatively stable across variation in number of innovators, consistent with a constant social learning component in the DCA equation. The corresponding simulations, however, do not follow this pattern; rather than being stable, the maximum slope increases with the proportion of innovators.

As Figure 6 indicates, there is no discrepancy between the asocial learning component of the DCA equation and asocial learning in simulations. Thus we can infer that this increase in maximum slope across number of innovators is due to social learning. This implies that increasing the proportion of the population with the asocial learning allele speeds innovation spread in the simulation, which is exactly what we see in Figures 1 and 2.

Analogously, Figures 4 and 5 depict results from adjusting the proportion of the population with the social learning allele when the whole population has the asocial learning allele. Figure 4 shows predictions of the DCA equation while Figure 5 presents simulation results. Again, the curves produced from the simulations have a greater maximum learning slope than the curves predicted by the DCA equation, and these discrepancies increase with the proportion of the population that is capable of social learning.

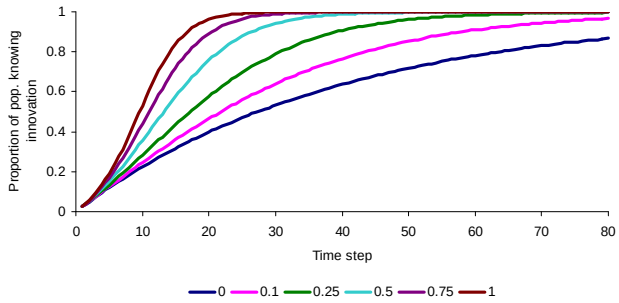


Figure 4: Diffusion curves predicted by DCA equation with asocial learning rate = .025, social learning rate = .25, and proportion of social learners varying from 0 to 1.

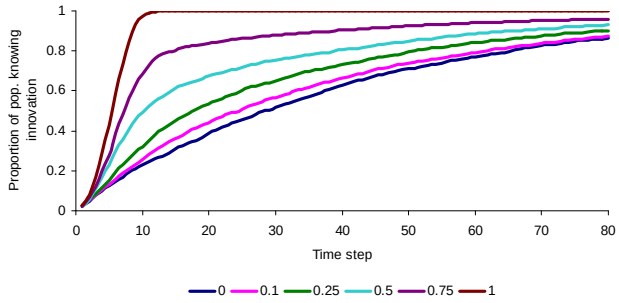


Figure 5: Simulations with asocial learning rate = .025, social learning rate = .25, proportion of social learners varying from 0 to 1, and proportion of asocial learners = 1.

With purely asocial learning (S = 0), the DCA equation closely tracks simulation results. The absolute differences between the equation and simulations averaged below .01 across all time steps. The lowest navy blue curves in Figures 4 and 5 are nearly identical. These two curves are re-plotted in Figure 6 to emphasize the overlap. This is the only simulation curve that the DCA equation successfully predicts. This predictive success makes sense because asocial learning in the simulation occurs as a random event based on a fixed probability, just as in the equation. Therefore, discrepancies between all other DCA and

simulation curves must result from social learning or possible interactions between social and asocial learning.

A possible cause of the increase in social learning as the proportion of innovators increases (Figure 3) is boundary surface length, the length of the perimeter surrounding groups of agents that know the innovation. These boundaries mark the area where naïve agents can learn the innovation. Thus, increasing this area should increase the speed of innovation spread.

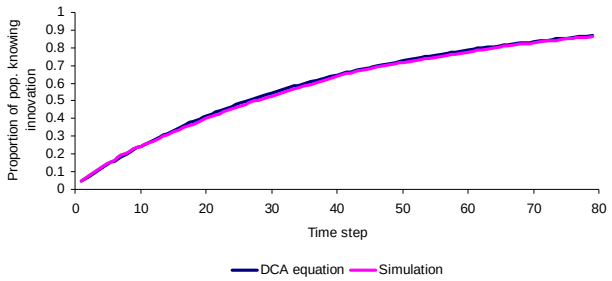


Figure 6: Diffusion curve predicted by the DCA equation compared to simulation results. Asocial learning rate = .025 and proportion of asocial learners = 1.

According to this analysis of the simulations, the spatial distribution of the agents that know the innovation affects social learning. Consider each innovator as a start point for an island of social learners. There will be more such islands when there are more initial innovators. More innovation islands generate more surface length and therefore more social learning. This suggests an interaction effect with the proportion of innovators in the population: when there are multiple initial innovators, there is a higher likelihood that more social learning will occur as a result of greater surface length. When there are fewer initial innovators, less social learning will occur as a result of less surface length.

Figure 7 shows two plots from simulations exemplifying this argument. These two tori present simulation outputs, each depicting the point where one-half the population possesses the innovation. In 7A, where the proportion of innovators = .05, there are two islands, resulting from a few early innovators. In 7B, where the proportion of innovators = 1, there are upwards of nine islands due to more innovators. Although the proportion of the population possessing the innovation is the same in both worlds, surface length is much greater for the simulation that was initialized with a higher proportion of innovators.

Thus, an explanation for the discrepancies between the predictions of the DCA equation and the simulation results is that asocial learning increases the number of start points for social learning, and therefore the emerging amount of surface length. Because surface length determines the amount of social learning that can take place, social learning and innovation spread increase substantially as surface length increases. Thus, increasing asocial learning increases social learning in the simulation (Figure 2) but not in the DCA equation (Figure 1; see Figure 3 for direct comparison). This explanation can be further validated by a mediation analysis (MacKinnon et al, 2007).

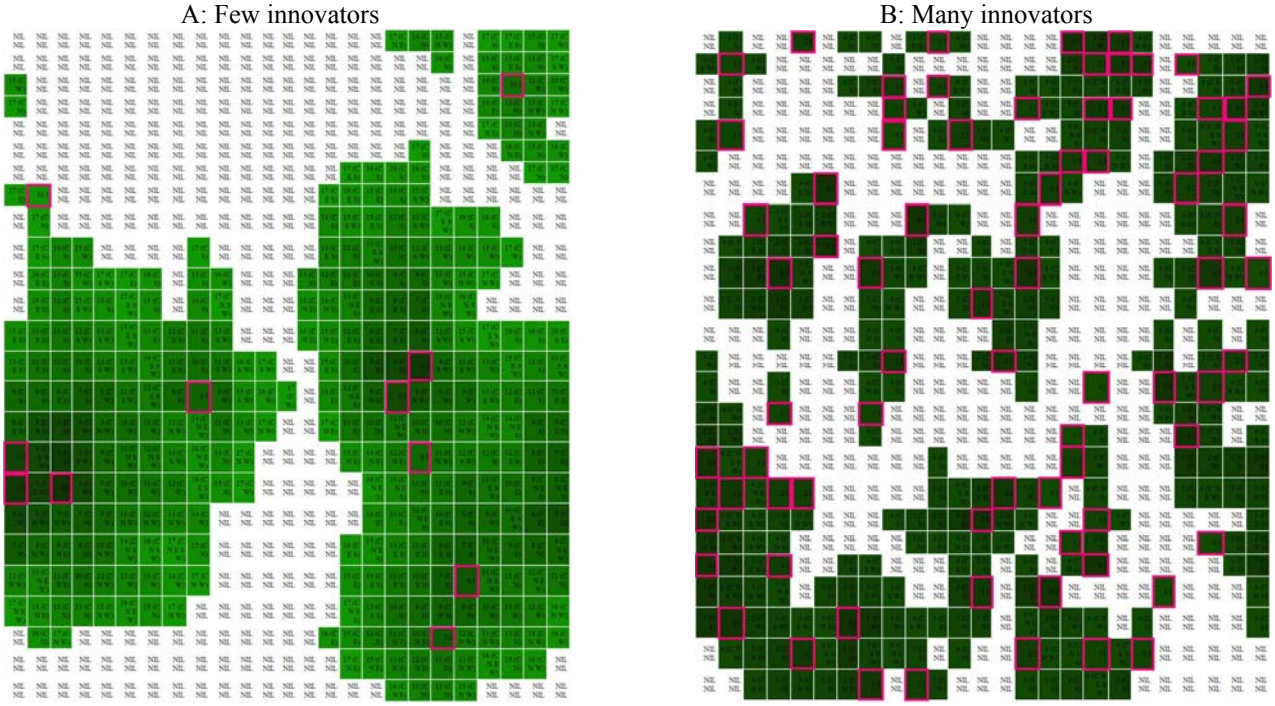


Figure 7: Two worlds with innovator proportions of .05 (A) and 1.0 (B). Time of acquisition is indicated by color saturation. Innovators are outlined in pink.

Mediation analysis is a type of linear regression that evaluates the relative effects of an independent variable (here, proportion of asocial learners) and a mediating variable (here, surface length) on a dependent variable (here, amount of social learning). The idea is that the independent variable affects the dependent variable, not only directly, but also indirectly via a mediating variable.

A mediation analysis of the simulation data across the six increasing proportions of asocial learners shows that 90.2 percent of the variance in the amount of social learning caused by variance in the proportion of asocial learners is mediated by surface length (total effect = 19.233 [$\beta = .945$], mediation effect = 17.351 [$\beta = .912 * .934 = .852$], $p < 0.0001$). As shown in Figure 8, the direct effect of the proportion of asocial learners on the amount of social learning becomes non-significant after controlling for the mediating variable of surface length, implying full mediation. This mediation analysis lends statistical support to the idea that surface length is the mechanism through which asocial learning causes social learning to speed up.

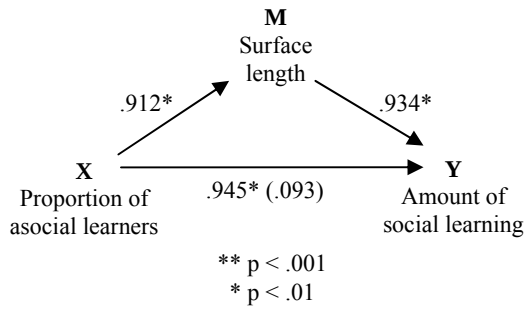


Figure 8: Standardized regression (beta) coefficients for meditational analysis. The path from X to Y falls to non-significance after controlling for the mediating variable of surface length (as indicated by the small coefficient in parentheses).

In summary, the simulations indicate that increasing social learning by adding more asocial learners increases surface length, and therefore increases the speed of social learning. This goes beyond the DCA equation which takes only the proportion of the population that knows the innovation as its independent variable. Also, speed of innovation spread is reduced as the number of agents with the social learning allele decreases.

Discussion

Our results show a difference between the DCA equation and the simulations, and this difference derives from the spatial factor of surface length. In the simulation, surface length is causally related to both social and asocial learning. Asocial learning increases surface length and surface length, in turn, increases social learning. The DCA equation, whose only independent variable is the proportion of the population that knows the innovation, does not capture this spatial factor. It is possible that the DCA equation could be improved on by a more sophisticated mathematical model that incorporates surface length.

A fundamental question is whether or not these results apply to the real world. After all, the results are a consequence of the design of the simulations. There is a high viscosity in the design, meaning that agents can only learn from their directly adjacent neighbors. This characteristic is presumably the cause of the spatial effect. If an agent could learn from any other randomly-selected agent, then the spatial arrangement of agents would have no bearing on the results. Therefore, the results are only applicable to real-world scenarios where social learning depends highly on spatial proximity. With tools like the telephone and internet, which allow social learning to take place across oceans, these results may not apply to diffusion of innovation for many human populations. This is not to say that diffusion in humans is random, but rather that these present simulations may be too constrained to model it. However, the current results do seem applicable to populations where social learning is heavily dependent on proximity, which would include a lot of human learning based on face-to-face interactions.

This consideration points to a distinction between *geographic* and *social-network* analysis. The simulations we present here are examples of geographic analysis, with agents learning from their immediate neighbors. Social networks can transcend spatial proximity by using communication technologies to cover great distances. This difference is not just one of viscosity but also of structural complexity, because social networks are often more complicated than geographic relationships.

Franz and Nunn (2009) developed a method of social network analysis called network-based diffusion analysis, or NBDA. NBDA uses the social network of a population and the times at which they learn innovations to probabilistically determine whether the learning mechanism is social or asocial. Their method of social network analysis seems promising, although it requires the researcher to determine the social network of a population. Such specification may not be feasible in excessively large populations. There are also cases where a geographic analysis may be more appropriate because some environments are in fact viscous (e.g., Lefebvre, 1995).

Also, Franz and Nunn's main interest was in detecting social or asocial learning when one such learning method was exclusively present. In contrast, our research used various, systematic combinations of these two learning mechanisms. Model sensitivity to such combinations of social and asocial learning is more interesting and important than detection of pure cases. Studying such combinations is critical to discovering interactions between social and asocial learning, as highlighted in our results.

A lattice structure permitting interaction only between immediately adjacent neighbors is actually a special case of a network that provides only those links (or edges). Thus, a generalization of our results would entail testing whether an analog of surface length would facilitate information

diffusion in networks of various topologies. Such an analogy might be the number of directed links between agents who possess, and agents who lack, an innovation. If such links indicate direction of causal influence, then it would be important to count the links from knowledgeable to naïve agents; if links indicate friendship choice, then count the links from naïve to knowledgeable agents, because agents are likely to be influenced by those they consider to be friends.

The original aim of this project was to look for ways to disentangle social learning from asocial learning through a spatial analysis of the diffusion of innovative behavior. The results suggest that a greater proportion of asocial learners results in more innovation islands and greater surface length. Although it may be difficult to determine surface length in wild populations, counting islands in a topographic analysis of observations of innovative behavior would seem feasible.

The spatial effect of surface length provides a mechanism to disentangle social and asocial learning that is not available in diffusion curve analysis. This kind of spatial analysis could become another valuable tool to measure and understand the differences between social and asocial learning. One next step is to apply the ideas developed from this simulation to real biological data. In doing so, we may be able to contribute new understanding of how adaptive innovations spread and how they interact with evolution. Another planned thrust is to study how evolution selects the best proportions of social and asocial learning alleles under different environmental conditions (Laland et al., 1996; Shultz, Hartshorn, & Hammond, 2008; Shultz, Hartshorn, & Kaznatcheev, 2009). In such research, faster learning cycles can be nested within slower evolutionary cycles.

Acknowledgments

This research is supported by a fellowship from the McGill Faculty of Science to AM, and operating grants to TRS from the Natural Sciences and Engineering Research Council of Canada and a Dean's Excellence Award from the McGill Faculty of Science. We are grateful to Simon Reader, Louis Lefebvre, Artem Kaznatcheev, Stuart Wright, and Charles Garfinkle for ideas and insights on our project.

References

Fisher, J., & Hinde, R. A. (1949). The opening of milk bottles by birds. *British Birds, 42,* 347-357.

Franz, M., & Nunn, C. L. (2009). Network-based diffusion analysis: A new method for detecting social learning. *Proceedings of the Royal Society B, 276* (1663), 1829-1836.

Goldstone, R. L., & Gureckis, T. M. (2009). Collective behavior. *Topics in Cognitive Science, 1* (3), 412-438.

Heyes, C. M. (1994). Social learning in animals: Categories and mechanisms. *Biological Reviews, 69,* 207-231.

Laland, K. N., Boyd, R., & Richerson, P. J. (1996). Developing a theory of animal social learning. In C. M. Heyes & B. G. Galef (Eds.), *Social learning in animals: The roots of culture.* San Diego: Academic Press.

Lefebvre, L. (1995). The opening of milk bottles by birds: Evidence for accelerating learning rates, but against the wave-of-advance model of cultural transmission. *Behavioural Processes, 34,* 43-54.

Lekvall, P., & Wahlbin, C. (1973). A study of some assumptions underlying innovation diffusion functions. *The Swedish Journal of Economics, 75* (4), 362-377.

MacKinnon, D. P., Fairchild, A. J., & Fritz, M. S. (2007). Mediation analysis. *Annual Review of Psychology, 58,* 593-614.

Reader, S. M. (2004). Distinguishing social and asocial learning using diffusion dynamics. *Learning and Behavior, 32* (1), 90-104.

Reader, S. M., & Laland, K. N. (Eds.). (2003). *Animal innovation*. Oxford: Oxford University Press.

Sherry, D. F., & Galef, B. G. Jr. (1984). Cultural transmission without imitation: milk bottle opening by birds. *Animal Behaviour, 32,* 937-938.

Shultz, T. R., Hartshorn, M., & Hammond, R. A. (2008). Stages in the evolution of ethnocentrism. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 1244-1249). Austin, TX: Cognitive Science Society.

Shultz, T. R., Hartshorn, M., & Kaznatcheev, A. (2009). Why is ethnocentrism more common than humanitarianism? In N. A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 2100-2105). Austin, TX: Cognitive Science Society.

Taking a Look (Literally!) at the Raven's Intelligence Test: Two Visual Solution Strategies

Maithilee Kunda (mkunda@gatech.edu)
Keith McGreggor (keith.mcgreggor@gatech.edu)
Ashok Goel (goel@cc.gatech.edu)
Design & Intelligence Laboratory, School of Interactive Computing, Georgia Institute of Technology
85 Fifth Street NW, Atlanta, GA 30332 USA

Abstract

The Raven's Progressive Matrices intelligence test is widely used as a measure of Spearman's general intelligence factor *g*. Although Raven's problems resemble geometric analogies, prior computational accounts of solving the test have been propositional. Studies of both typical and atypical human behavior suggest the possible existence of visual strategies; for example, neuroimaging data indicates that individuals with autism may preferentially recruit visual processing brain regions when solving the test. We present two different algorithms that use visual representations to solve Raven's problems. These algorithms yield performances on the Standard Progressive Matrices test at levels equivalent to typically developing 9.5- and 10.5- year-olds. We find that these algorithms perform most strongly on problems identified from factor-analytic human studies as requiring gestalt or visuospatial operations, and less so on problems requiring verbal reasoning. We discuss implications of this work for understanding the computational nature of Raven's and visual analogy in problem solving.

Keywords: Analogy; intelligence tests; knowledge representations; mental imagery; Raven's Progressive Matrices; visual reasoning.

Introduction

The Raven's Progressive Matrices tests (Raven, Raven, & Court, 1998) are a collection of standardized intelligence tests that consist of geometric analogy problems in which a matrix of geometric figures is presented with one entry missing, and the correct missing entry must be selected from a set of answer choices. Figure 1 shows an example of a 2x2 matrix problem that is similar to one in the Standard Progressive Matrices (SPM); other problems contain 3x3 matrices. The entire SPM consists of 60 problems divided into five sets of 12 problems each (sets A, B, C, D & E), roughly increasing in difficulty both within and across sets.

Although the Raven's tests are supposed to measure only eductive ability, or the ability to extract and understand information from a complex situation (Raven, Raven, & Court 1998), their high level of correlation with other multi-domain intelligence tests have given them a position of centrality in the space of psychometric measures (e.g. Snow, Kyllonen, & Marshalek 1984), and as a result, they are often used as tests of general intelligence in clinical, educational, vocational, and scientific settings.

Computational accounts of problem solving on the Raven's tests have, with the exception of Hunt (1974), assumed that visual inputs are translated into propositions, over which various kinds of reasoning then take place. In this paper, we provide evidence from two different methods that Raven's problems can be solved visually, without first converting problem inputs into propositional descriptions.

Existing Computational Accounts

Carpenter, Just, and Shell (1990) used a production system that took hand-coded symbolic descriptions of problems from the Advanced Progressive Matrices (APM) test and then selected an appropriate rule to solve each problem. The rules were generated by the authors from *a priori* inspection of the APM and were validated in experimental studies of subjects taking the test with verbal reporting protocols. Bringsjord and Schimanski (2003) used a theorem-prover to solve selected Raven's problems stated in first-order logic.

Lovett, Forbus, and Usher (2007) combined automated sketch understanding with the structure-mapping analogy technique to solve problems from the Standard Progressive Matrices (SPM) test. Their system took as inputs problem entries sketched in Powerpoint as segmented shape objects and then automatically translated these shapes into propositional descriptions. A two-stage structure-mapping process was then used to select the answer that most closely fulfilled inferred analogical relations from the matrix.

In contrast to these propositional approaches, Hunt (1974) proposed the existence of two qualitatively different strategies: "Gestalt," which used visual representations and perceptual operations like continuation and superposition,

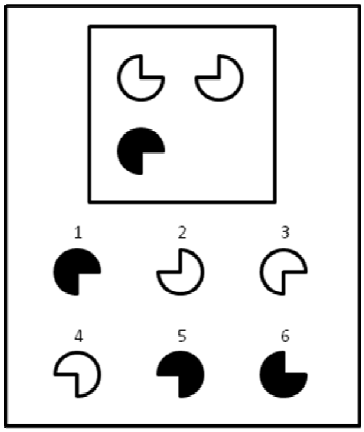


Figure 1: Example problem similar to one from the Standard Progressive Matrices (SPM) test.

and "Analytic," which used propositional representations and logical operations. The Analytic algorithm is similar to that of Carpenter, Just, and Shell (1990) in that it applied rules to lists of features representing each matrix entry. The Gestalt algorithm is similar to our methods in that it used visual operations over imagistic problem inputs, but it differs in that it operated on the entire problem matrix as a single image, whereas our methods treat each matrix entry as a separate image. While Hunt's algorithms provide an intuitively appealing account of solving Raven's problems, neither algorithm was actually implemented.

Behavioral Evidence for Multiple Strategies

Studies of human behavior suggest that qualitatively distinct problem solving strategies can be used to solve Raven's problems. Factor analyses of both the SPM (Lynn, Allik, & Irving, 2004; van der Ven & Ellis, 2000) and the APM (Dillon, Pohlmann, & Lohman, 1981; Mackintosh & Bennett, 2005; Vigneau & Bors, 2005) have identified multiple factors underlying these tests, which often divide test problems into two categories: those solvable using visuospatial or gestalt operations and those solvable using verbal reasoning. In support of this dichotomy, DeShon, Chan, and Weissbein (1995) found that simultaneously performing a verbal overshadowing protocol differentially impaired accuracy on about half of APM problems.

These studies of typically developing individuals have generally focused on within-individuals differences in solution strategies, i.e. a particular individual using different strategies on different portions of the test in a single sitting. Recent evidence from autism offers evidence of between-individuals strategy differences as well: individuals with autism do not show the same correlations between Raven's scores and other cognitive measures that are robustly demonstrated by typically developing individuals (Dawson, Soulières, Gernsbacher, & Mottron, 2007).

Even more striking are recent neuroimaging data that show increased brain activation in visual regions for individuals with autism solving the SPM than controls (Soulières et al., 2009). This study also found significant differences in reaction time as a function of problem type, with problems classified as "figural" or "analytic" based on previously published factor-analytic studies. The results from this study are highly suggestive of individuals with autism using a visual strategy that contrasts with the strategy used by controls. Evidence for a visual strategy preference in autism is found across several other cognitive task domains as well (Kunda & Goel, 2008).

Our approach

We hypothesize that Raven's problems can be solved computationally using purely visual representations. To test this hypothesis, we have developed two different algorithms that in this paper we will call the "affine" method and the "fractal" method. Both methods use image transformations to solve Raven's problems *without* converting the input images into any kinds of propositions. Below, we describe each of these algorithms, followed by an analysis of their performance on all 60 problems from the Raven's Standard Progressive Matrices (SPM) test.

Visual Methods for the Raven's Test

Similitude Transformations

At the core of each of our algorithms are image operations that fall under the category of affine transformations, and in particular similarity-preserving or "similitude" transforms. Similitude transforms can be represented as compositions of dilation (i.e. scaling), orthonormal transformation, and translation. Our implementations presently examine the identity transform, horizontal and vertical reflections, and 90°, 180°, and 270° orthonormal rotations, composed with various translations. The affine method restricts dilation to a value of one, i.e. no scaling, whereas the fractal method uses a short sequence of progressively smaller dilation values, i.e. its similitude transformations are contractive.

There is evidence that human visual processing can apply some of these types of transformations to mental images, or at least operations that are computationally isomorphic in some sense. In the theory of mental imagery proposed by Kosslyn, Thompson, and Ganis (2006), transformations of mental images include scanning (i.e. translation), zooming (i.e. scaling), and rotation, among others.

A Model of Similarity

Similarity lies at the core of both of our accounts of visual problem solving on the Raven's test. We calculate visual similarity using the ratio model (Tversky, 1977):

$$similarity(A,B) = \frac{f(A \cap B)}{f(A \cap B) + \alpha f(A-B) + \beta f(B-A)} \quad (1)$$

In this equation, *f* represents some function over features in each of the specified sets; for instance, *f* might simply be a count of features. The constants α and β are used as weights for the non-intersecting portions of the sets *A* and *B*. If α and β are both set to one, then this equation becomes:

$$similarity(A,B) = \frac{f(A \cap B)}{f(A \cup B)} \quad (2)$$

Equation (2) is used in both the affine and fractal methods, and it yields maximal similarity for sets in which A is equal to B. In contrast, if α is set to one and β is set to zero, it yields maximal similarity for sets in which A is a proper subset of B. If α is set to zero and β is set to one, then the opposite holds, and maximal similarity is found for sets in which B is a proper subset of A. These two variants are used in the affine method to capture notions of image composition, i.e. image addition and subtraction.

In the affine method, each feature is defined as a pixel, and intersection, union, and subtraction operations are defined as the product, maximum, and difference of RGB pixel values, respectively. The fractal method uses features derived from different combinations of elements from the fractal encoding (McGreggor, Kunda, & Goel, 2010).

The Affine Method

The affine method assumes that elements within a row or column in a Raven's problem matrix are related by similitude transformations. It tries to discover which similitude transformation best fits any of the complete rows or columns in the matrix, and then applies this transform to the last row/column to generate a guess for the answer. Then, it compares this guess to each of the answer choices, and chooses the answer that is most similar.

Each similitude transformation is represented as the combination of three image operations: a base transform, a translation, and a composition. Algorithm 1 shows how, for a pair of images A and B, these three components of the "best-fit" similitude transformation are found. Given a Raven's problem, then, the affine method seeks to discover the best-fit similitude transform over various combinations of the matrix entries. In particular, the algorithm assumes that certain analogical relationships exist based on the spatial arrangement of the entries. Similitude transforms are calculated for those combinations of entries that would yield an analogical mapping to solve for the missing entry. The specific base transforms and analogical relationships used by the affine algorithm are shown in Table 1, divided into those used for 2x2 and for 3x3 matrix problems.

Once the relationship and transformation are found that maximize similarity, the transformation is applied to the first entry or entries in the last row or column, as listed in Table 1. The resulting image represents the algorithm's best guess as to the missing entry. This image is compared to the answer choices, using Equation (2), and the best match is chosen as the final answer.

For each base transform T:
- Apply T to Image A.
- Find translation (tx, ty) which yields best match between T(A) and B, using Eq. (2).
- Find image composition operand X as follows:
  - Calculate similarity using Eq. (1) with:
    1) $\alpha = 1, \beta = 1$
    2) $\alpha = 1, \beta = 0$
    3) $\alpha = 0, \beta = 1$
  - Choose maximum similarity value.
    - If maximum is (1), then X = 0.
    - If maximum is (2), then X = B – A, and $\oplus$ refers to image addition.
    - If maximum is (3), then X = A – B, and $\oplus$ refers to image subtraction.

The best-fit similitude transformation can then be specified as:

$$[T_{max}+(tx, ty)](A) \oplus X = B$$

Algorithm 1. Affine method for calculating best-fit similitude transformation for a pair of images A and B. For three-element transforms, T is applied to images A and B, and the result is compared, as above, to image C.

Table 1: Base transforms and matrix relationships used by the affine algorithm.

	Transforms	2x2: A B / C ?	3x3: A B C / D E F / G H ?
Two-element transforms & relations	Identity Mirror Flip Rotate90 Rotate180 Rotate270	AB→C? AC→ B?	BC→H? DG→F? AC→G? AG→C? EF→H? EH→F? DF→G? BH→C? GH→H? CF→F?
Three-element transforms & relations	Union Intersection XOR	n/a	ABC→GH? DEF→GH? ADG→CF? BEH→CF?

For example, take the problem given in Figure 1. The similarity scores calculated for the various transforms and relationships are shown in Table 2. The best-fit similitude transformation is found to be a mirror (or reflection about the vertical axis) for the relationship AB, using an addition image composition (i.e. maximal similarity found using $\alpha = 1, \beta = 0$). Therefore, the answer image "?" is obtained using the analogous relationship of A:B :: C?. C is mirrored, translated by the (tx, ty) that was found in the search, and the composition operand of B – A (which in this case is mostly a blank image) is added on to the result. Finally, this "guess" image is compared to each of the six answer choices using Equation (2), and the best match is chosen as the final answer, which in this case is answer #5.

Table 2: Calculation of best-fit similitude transform and resulting answer guess for the problem shown in Figure 1.

Relation	Transform	**α = 1 β = 1**	**α = 1 β = 0**	**α = 0 β = 1**
AB	**Identity**	0.475	0.644	0.644
	Mirror	0.963	**0.981**	0.981
	Flip	0.337	0.504	0.504
	Rotate90	0.341	0.508	0.508
	Rotate180	0.453	0.624	0.624
	Rotate270	0.947	0.973	0.973
AC	**Identity**	0.256	0.764	0.277
	Mirror	0.252	0.759	0.274
	Flip	0.335	0.951	0.341
	Rotate90	0.331	0.941	0.338
	Rotate180	0.257	0.771	0.279
	Rotate270	0.250	0.752	0.273
	Generated guess:			

The Fractal Method

The fractal method proceeds in a manner which at once resembles and yet differs from the affine method. Like the affine method, the fractal method seeks to find a re-representation of the images within a Raven's problem as a set of similitude transformations. Unlike the affine method, the fractal method seeks these representations at a significantly finer partitioning of the images, and uses these representations (and more precisely, features derived from these representations) to determine similarity for each possible answer, simultaneously, across the bulk of relationships present in the problem.

The mathematical derivation for the process of fractal image representation expressly depends upon the notion of real world images, i.e. images that are two dimensional and continuous (Barnsley & Hurd, 1992). Two key observations are that all naturally occurring images we perceive appear to have similar, repeating patterns, and, no matter how closely we examine the real world, we find instances of similar structures and repeating patterns. These observations suggest that it is possible to describe the real world in terms other than those of shapes or traditional graphical elements —in particular, terms that capture the observed similarity and repetition alone. Computationally, determining the fractal representation of an image requires the use of the fractal encoding algorithm, which, given an image D, seeks to discover the set of transformations T that can transform any source image into D.

Decompose D into a set of N smaller images $\{d_1, d_2, d_3, \ldots, d_n\}$. These individual images are sets of points.

For each image d_i:

Examine the entire source image S for an equivalent image s_i such that a similitude transformation of s_i will result in d_i. This transformation will be a 3x3 matrix, as the points within s_i and d_i under consideration can be represented as the 3D vector <x, y, c> where c is the (grayscale) color of the 2D point <x,y>.

Collect all such transforms into a set of candidates C.

Select from C the transform which most minimally achieves its work, according to some predetermined, consistent metric.

Let T_i be the representation of the chosen affine transformation of s_i into d_i.

The set $T = \{T_1, T_2, T_3, \ldots, T_n\}$ is the fractal encoding of the image D.

Algorithm 1. Fractal encoding algorithm for determining the fractal representation of an image D.

This algorithm, shown in Algorithm 2, is considered "fractal" for two reasons: first, the transformations chosen are generally contractive, which leads to convergence, and second, the convergence of S into D can be shown to be the mathematical equivalent of considering D to be an attractor (Barnsley & Hurd, 1992).

Once fractal representations have been calculated for each pair of images in a Raven's problem, the metric shown in Equation (2) is used to calculate similarity between all of the pairwise relationships present in the matrix and those calculated with the given answer choices, using features derived from the fractal encodings. Whichever answer choice yields the most similar fractal representations across all pairwise relationships is chosen as the final answer. The fractal method is described in more detail in McGreggor, Kunda, and Goel (2010).

Results

We tested both the affine and fractal algorithms on all 60 problems from the Raven's Standard Progressive Matrices (SPM) test. To obtain visual inputs for the algorithms, we first scanned a paper copy of the SPM, aligned each page to lie squarely along horizontal and vertical axes, and then divided each problem into separate image files representing each of the matrix entries and answer choices. No further image processing was performed on these input images. As a result, these images were fairly noisy; they contained numerous misalignments and pixel-level artifacts from the scanning and subdividing processes.

Then, after answers for all 60 SPM problems were obtained from each algorithm, we scored each method according to standard protocols for the SPM. In particular, we looked at three different measures of performance:

1) The total score from the SPM summarizes the test-taker's overall level of performance.
2) This total score can be compared to national age-group norms to determine a percentile ranking.
3) A "consistency" measure is obtained by comparing performance on each of the five sets within the SPM, A through E, with the expected scores for each set given the same total score, which are obtained from normative data (Raven, Raven, & Court, 1998).

In addition, we conducted a separate analysis of results according to problem type, looking at accuracy as a function of three problems classifications: "gestalt continuation," "visuospatial," and "verbal-analytic," which we obtained from a published factor analytic study of the SPM (Lynn, Allik, & Irving, 2004).

Affine Results

The affine algorithm correctly solved 35 of the 60 problems on the SPM. For children in the U.S., this total score corresponds to the 75th percentile for 9-year-olds, the 50th percentile for 10½-year-olds, and the 25th percentile for 13-year-olds (Raven, Raven, & Court 1998).

The breakdown of this total score across sets is shown in Figure 2, along with the expected score composition for this

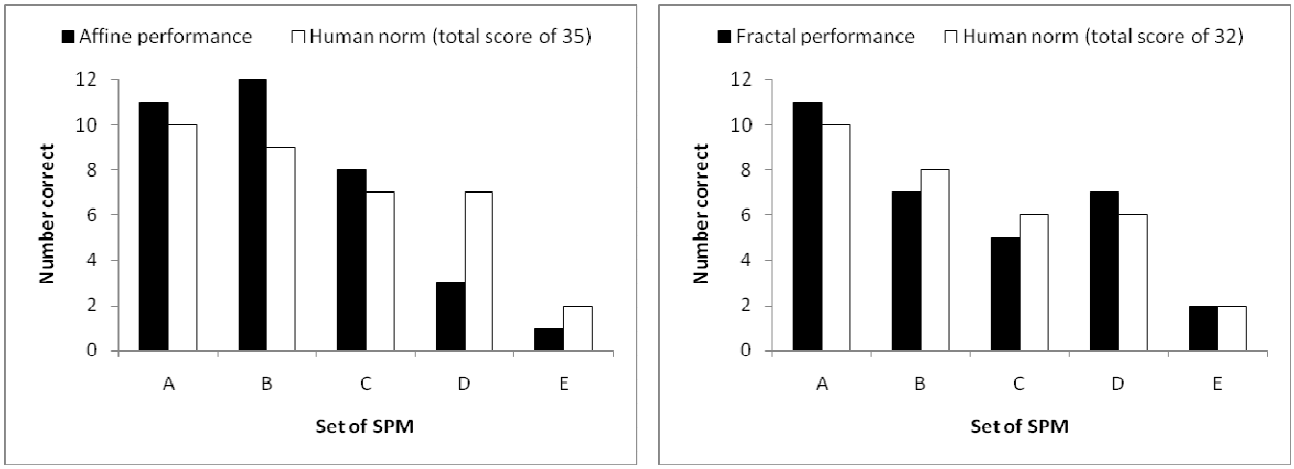


Figure 2: Breakdown of affine (left) and fractal (right) results across sets in the SPM. Also shown is the expected score breakdown for total scores of 35 and 32, from normative human data (Raven, Raven, & Court, 1998).

same total score. Scoring instructions for the SPM indicate that, if the score for any set deviates from the expected score for that set by more than two, the overall test results cannot necessarily be interpreted as a measure of general cognitive function (Raven, Raven, & Court, 1998). This check is intended to detect scores affected by a poor understanding of test instructions, random guessing strategies, or other departures from the intended test-taking framework. As shown in Figure 2, the affine scores deviate by more than ±2 from the expected scores on sets B and D. In particular, the affine algorithm does too well on Set B and not well enough on Set D to match typical human norms.

Fractal Results

The fractal algorithm correctly solved 32 of the 60 problems on the SPM. For children in the U.S., this total score corresponds to the 75th percentile for 8-year-olds, the 50th percentile for 9½-year-olds, and the 25th percentile for 11½-year-olds (Raven, Raven, & Court 1998).

The breakdown of this total score across sets is shown in Figure 2, along with the expected score composition for this same total score. The fractal scores fall within ±2 of the expected scores for each set, indicating that the fractal results are "consistent" with normative SPM scores.

Results by Problem Type

The final analysis we performed looked at the performance of both algorithms as a function of problem type on the SPM. Factor-analytic studies have often found evidence for multiple factors underlying problem solving on the SPM (e.g. van Der Ven & Ellis, 2000); we used the breakdown obtained by one such study to divide problems into those

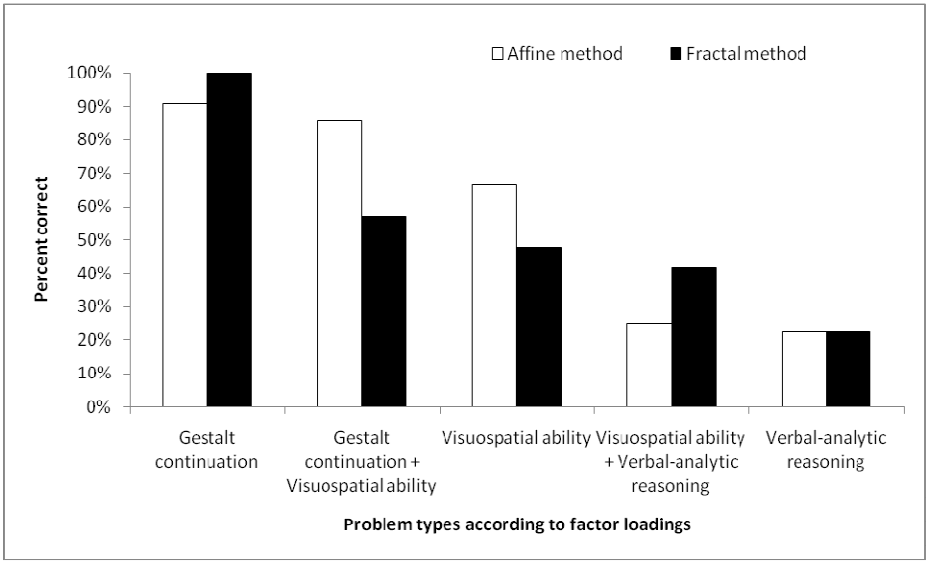


Figure 3: Breakdown of affine and fractal algorithm results on the SPM by problem type. Problem breakdowns were obtained from a factor-analytic study of human performance (Lynn, Allik, & Irving, 2004).

that loaded on "gestalt continuation," "visuospatial," or "verbal-analytic" factors (Lynn, Allik, & Irving, 2004).

Figure 3 shows the performance of both the affine and fractal algorithms on problems from the SPM which load on different combinations of these factors. Both the affine and fractal methods perform most strongly on gestalt problems, slightly less so visuospatial problems, and significantly less so on problems requiring verbal-analytic reasoning, though the relative difficulties of each of these problem types could represent a potential confound for these results.

Discussion

We have presented two different algorithms that use purely visual representations and transformations to solve more than half of the problems on the Raven's SPM test. Our results align strongly with evidence from typical human behavior suggesting that multiple cognitive factors underlie problem solving on the SPM, and in particular, that some of these factors appear based on visual operations. Whether these algorithms behave on the SPM similarly to individuals with autism, who may demonstrate a cognitive preference for solving the test visually, remains to be determined.

That purely visual methods can achieve such significant results on a standardized intelligence test is a little surprising to us, especially as the input images for both algorithms were taken "as is," from raw scans of a paper copy of the test. This robust level of performance calls attention to the visual processing substrate shared by the affine and fractal algorithms: similitude transforms as a mechanism for image manipulation, and the ratio model of similarity as a mechanism for image comparison. Of course, there are many other types of visual processing that may or may not be important for accounts of visual analogy, such as non-similitude shape transformations or image convolutions, which certainly bear further investigation.

While it has been shown (Davies, Yaner, & Goel, 2008) that visuospatial knowledge alone may be sufficient for addressing many analogy problems, the representations used in that work were still propositional. In contrast, the methods described here use only visual representations in the form of image similitude transformations. We believe the visual methods we have presented for solving the SPM can be generalized to visual analogy in other domains, such as other standardized tests (e.g. the Miller's Geometric Analogies test). We conjecture that these methods may provide insight into general visual recognition and recall.

Acknowledgments

This research has been supported by an NSF grant (IIS Award #0534266), "Multimodal Case-Based Reasoning in Modeling and Design," by ONR through an NDSEG fellowship, and by the NSF GRFP fellowship program.

References

Barnsley, M. F., & Hurd, L. P. (1992). *Fractal Image Compression*. Boston, MA: A.K. Peters.

Bringsjord, S., & Schimanski, B. (2003). What is artificial intelligence? Psychometric AI as an answer. *IJCAI*, 18, 887–893.

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: a theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological Review*, 97(3), 404-31.

Davies, J., Goel, A., & Yaner. P. (2008). Proteus: A theory of visual analogies in problem solving. *Knowledge-Based Systems*, 21(7), 636-654.

Dawson, M., Soulières, I., Gernsbacher, M. A., & Mottron, L. (2007). The level and nature of autistic intelligence. *Psychological Science*, 18(8), 657-662.

DeShon, R. P., Chan, D., & Weissbein, D. A. (1995). Verbal overshadowing effects on Raven's advanced progressive matrices: Evidence for multidimensional performance determinants. *Intelligence*, 21(2), 135-155.

Dillon, R. F., Pohlmann, J. T., & Lohman, D. F. (1981). A factor analysis of Raven's Advanced Progressive Matrices freed of difficulty factors. *Educational and Psychological Measurement*, 41, 1295–1302.

Hunt, E. (1974). Quote the raven? Nevermore! In L. W. Gregg (Ed.), *Knowledge and Cognition* (pp. 129–158). Hillsdale, NJ: Erlbaum.

Kunda, M., & Goel, A. K. (2008). Thinking in Pictures: A Fresh Look at Cognition in Autism. In *Proc. 30th Annual Conf. Cognitive Science Society* (pp. 321-326).

Lovett, A., Forbus, K., & Usher, J. (2007). Analogy with qualitative spatial representations can simulate solving Raven's Progressive Matrices. In *Proc. 29th Annual Conf. Cognitive Science Society* (pp. 449-454).

Lynn, R., Allik, J., & Irwing, P. (2004). Sex differences on three factors identified in Raven's Standard Progressive Matrices. *Intelligence*, 32(4), 411-424.

Mackintosh, N., & Bennett, E. (2005). What do Raven's Matrices measure? An analysis in terms of sex differences. *Intelligence*, 33(6), 663-674.

McGreggor, K., Kunda, M., & Goel, A. (2010). A fractal approach towards visual analogy. In *Proc. 1st International Conf. Computational Creativity*, Lisbon, Portugal, January, 2010.

Raven, J., Raven, J. C., & Court, J. H. (1998). *Manual for Raven's Progressive Matrices and Vocabulary Scales*. San Antonio, TX: Harcourt Assessment.

Soulières, I., Dawson, M., Samson, F., Barbeau, E. B., Sahyoun, C. P., Strangman, G. E., et al. (2009). Enhanced visual processing contributes to matrix reasoning in autism. *Human Brain Mapping*. 30(12), 4082-107

Tversky, A. (1977). Features of similarity. *Psychological Review*, 84(4), 327-352.

van der Ven, A. H. G. S., & Ellis, J. L. (2000). A Rasch analysis of Raven's standard progressive matrices. *Personality and Individual Differences*, 29(1), 45-64.

Vigneau, F., & Bors, D. A. (2008). The quest for item types based on information processing: An analysis of Raven's Advanced Progressive Matrices, with a consideration of gender differences. *Intelligence*, 36(6), 702-710.

The Dice are Cast:
The Role of Intended versus Actual Contributions in Responsibility Attribution

Tobias Gerstenberg (t.gerstenberg@ucl.ac.uk), **David A. Lagnado (d.lagnado@ucl.ac.uk)**
Department of Cognitive, Perceptual, and Brain Sciences
University College London, United Kingdom

Yaakov Kareev (kareev@vms.huji.ac.il)
Center for the Study of Rationality
The Hebrew University of Jerusalem, Israel

Abstract

How much are people's responsibility attributions affected by intended versus actual contributions in group contexts? A novel experimental-game paradigm dissociated intended from actual contributions: good intentions could result in bad outcomes and bad intentions in good ones. Participants acted as external judges and attributed responsibility to computer players engaging in a repeated game. On each round, three players formed a group and each chose to roll one of three dice that differed in terms of price and probability distribution. The team won if the sum exceeded a certain threshold. The results showed that both intended contribution, reflected in the choice of die, and actual contribution, reflected in the outcome of rolling the die, were determinants of participants' responsibility attributions. However, contrary to previous evidence (Cushman, Dreber, Wang, & Costa, 2009), more participants based their attributions on the intention rather than the outcome.

Keywords: responsibility; attribution; intentionality; outcome bias; experimental game.

Introduction

At the beginning of the movie "Naked Gun 2 1/2" the police officer Frank Drebin is honoured at the presidential dinner for his recent achievement of having eliminated his 1000th drug dealer. In response to this, Mr Drebin admits that he had run over the last two men with his car. Luckily, it turned out that they were wanted drug dealers. Cases of "moral luck" have drawn the attention of philosophers (Williams, 1981), legal scholars (Hart, 1985), and psychologists (Mitchell & Kalb, 1981). These situations are characterized by the fact that the outcome of an action influences its moral evaluation retrospectively, even if this outcome was to a large extent beyond the control of the agent. Mr Drebin, for example, receives praise for his reckless driving only because the men he ran over happened to be drug dealers: a circumstance which was clearly beyond his control.

That people are influenced by outcome knowledge is a well established psychological finding (Baron & Hershey, 1988; Fischhoff, 1975). Fischhoff (1975) showed that people are prone to a hindsight bias: knowledge about the real outcome influences the perceived likelihood of different possible outcomes. Furthermore, people appear to be unaware of the influence that outcome knowledge exerts on their judgments and are, hence, unable to control for its effect. Baron and Hershey (1988) showed that outcome knowledge influences how people evaluate decisions made under uncertainty. Even when participants had all information relevant to the decision, including the probabilities of each possible outcome, knowledge of the actual outcome nevertheless influenced their judgments of the competence of the decision-maker. Interestingly, when asked whether they *should* take the outcome into account, most participants answered in the negative.

Differential evaluations of identical decisions or actions are also reflected in the Law's differential treatment of negligence versus negligence that leads to harm, as well as cases of attempted murder versus murder. The latter cases share the fact that the person had the intention to kill; however, only in the case of murder did the intended event come about. Recently, experimental philosophers have put forward the claim that the folk notion of intention is deeply intertwined with the (moral) evaluation of the potential outcomes (Knobe, 2003). Whether a behaviour is thought to have been performed intentionally depends crucially on the outcome of that action. An identical action is judged by more participants as intentional when its outcome is blameworthy as opposed to praiseworthy.

Psychologists have also shown that intentions play a significant role when it comes to attributions of responsibility (Lagnado & Channon, 2008) and intentionality thus constitutes an important building block of psychological frameworks of responsibility attribution (Alicke, 2000; Shaver, 1985). Shaver's (1985) theory of blame assumes a linear process starting from considerations about causality, intentionality, foreseeability and potential justifications and leading to judgments of blame or praise. In contrast, Alicke's (2000) account acknowledges the possibility of that process being reversed. The valence of the outcome can trigger spontaneous moral or emotional evaluations which influence the perception of the antecedents of the outcome. This includes judgments about the causal impact of an agent, whether the action was performed intentionally as well as whether the agent should have foreseen the outcome.

The importance of the concept of intentionality has also been recognized by economists. Variations of classic economic games, like the ultimatum game, have been employed to investigate the effects of outcome versus intention on people's perception of fairness. In the ultimatum game (Güth, Schmittberger, & Schwarze, 1982), a first player is allocated a certain amount of money. She can then decide how much of that money to give to a second player, who can either accept or reject the offer. If he refuses, both players get nothing.

Two main findings with respect to the influence of intentions on the behaviour of the second player are worth mentioning. First, if the allocation of the first player is determined by a computer and hence cannot be ascribed an intention, the rejection rates for "unfair" offers are significantly lower (Falk, Fehr, & Fischbacher, 2008). Second, the rejection rates of unequal offers strongly depend on the allocator's set of possible alternatives (McCabe, Rigdon, & Smith, 2003). The acceptability of an action is hence evaluated with respect to the choice set and an unequal offer more readily accepted if the allocator could not have been kinder. In order to accommodate these findings, economists have moved from fairness theories that only considered outcomes (Fehr & Schmidt, 1999) to theories based on intentions (Dufwenberg & Kirchsteiger, 2004) and theories incorporating both intentions and outcomes (Falk et al., 2008).

As demonstrated by the moral luck example in "Naked Gun", there is another factor beyond intentions and outcomes that is relevant when it comes to considerations about fairness or attributions of responsibility: the control an agent has over the outcomes he brings about. Our environment is fundamentally noisy and, most of the time, we only have partial control over the effects of our actions. While it is true that the valence of intention and outcome are correlated in everyday life, this relationship is imperfect. Good intentions can sometimes lead to bad outcomes and bad intentions to good ones. For example, a careful driver might cause the death of a careless child. In order to understand the complex relationship between intentions, outcomes and control it is necessary to create experimental situations in which these factors can be dissociated.

In a recent study, Cushman et al. (2009) investigated the effects of intention versus outcome on perceived fairness in a two-player, allocater-responder game. Similar to the ultimatum game, the allocator proposed how a pot of \$10 should be shared. Allocations were either stingy (player 1: \$10, player 2: \$0), fair (\$5, \$5) or generous (\$0, \$10). The responder could punish or reward the allocation of player 1 by subtracting or adding up to \$9 to her account. Importantly, in one condition of the experiment, the allocator only had partial control over the outcome. She had to choose which one of three possible dice she wanted to roll. These dice differed in terms of the probability with which they led to stingy, fair or generous outcomes. Following a strategy format, responders had to indicate for each of the 9 possible combinations (e.g. generous die, stingy outcome) how much money they wanted to add or subtract from the allocator. The results revealed that participants were much more influenced by actual outcomes than by intentions. Responders tended to subtract money for selfish outcomes for all three dice, whereas they added money for fair and generous outcomes. The choice of die exerted only a small effect on this general pattern. Surprisingly, the results of a condition in which the allocator had perfect control were virtually identical. Hence, the study provides further support for the finding that people can be so sensitive to outcomes that they sometimes disregard the underlying intention that lead to that outcome. However, Cushman et al. (2009) admit that methodological limitations might have contributed to their findings. Importantly, since the responder is part of the game it is the outcome and not the intention that is the most relevant to him. In order to validate their findings, it is important to investigate how an independent judge would have decided.

The current experiment addressed this limitation. We created a setting in which an external observer evaluated the behaviour of agents participating in an experimental game. The following scenario helps to exemplify the main components of our experiment: Sarah is running for the position of student representative. Three friends are helping her campaign by distributing flyers. Tom puts in a lot of effort and distributes 100 flyers. John puts slightly less effort into the campaign and only distributes 50. Finally, Alex thinks that Tom's and John's contributions are probably already enough to win the campaign and he only distributes 30 flyers. As it turns out, 20 of Tom's, 20 of John's and 25 of the people who received their flyer from Alex voted for Sarah. As a result, Sarah won the election. Assuming that Sarah knows about both the number of distributed flyers and the votes she received, how much is she going to praise each of her three friends for their contribution to her win?

Two aspects of the outlined scenario are important with respect to the current study. First, it shows how intention and outcome can sometimes mismatch in situations over which agents exercise only partial control. Despite Tom's good intention and effort he contributed no more to the collective outcome than John and even less than Alex. Second, the scenario entails a component that is characteristic of social dilemmas (see e.g. Hardin, 1968). Each individual agent has to weigh the cost of the effortful process of distributing flyers with the potential gain of an election won. Alex's thought process indicates each person's motivation to free-ride on the effort of the others. Assuming the spoils of a victory are equally shared, the person who put in the least effort will have the highest net benefit.

The current study investigated the effects of intended and actual contributions on responsibility attribution in a group context in which agents had only partial control over their contributions. How well can intended contributions, actual contributions, or their combination explain participants' responsibility attributions?

Experiment

The aim of the experiment was to generate a situation in which intended versus actual contributions could dissociate. Participants took the perspective of an external observer and judged the behaviour of computer players engaging in an experimental game (see Figure 1). Hence, participants did not actively engage in the game themselves. In each round of the game, three computer players were randomly selected to form a group. Each player chose one of three available dice to roll. The dice differed in terms of their underlying probabil-

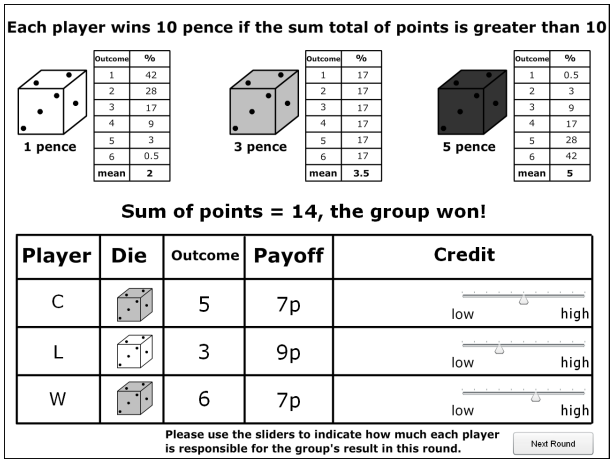


Figure 1: Screenshot of the game depicting a won round.

ity distributions (see top part of Figure 1). The white die had a higher probability of smaller values, the grey die was fair, and the black die was skewed towards higher values (in the experiment, the colours were bronze, silver and gold). The group of players won a round if the sum of their outcomes was greater than 10. If the group won a round, 30 pence were equally distributed between the players. If the group lost, no money was distributed. Importantly, the players had to pay different amounts for the dice before they rolled them. The white die cost 1 pence, the grey die 3 pence, and the black 5 pence. Each individual player's payoff was a function of the group's result, that is, whether they won or lost, and the money he had to pay for the die of his choice. The task of the participants as independent judges was to indicate how much they thought each player was responsible for the group's result in each round.

The computer players chose each of the dice with an equal probability. The chosen payoff function created a social dilemma. The overall probability of winning was 50%. The probabilities of winning given that a player had chosen the white, grey or black die were 33%, 50% and 68%, respectively. This led to an expected payoff of 2.3 pence per round for the white die ($33\% \times 9 - 67\% \times 1 = 2.3$). The expected payoffs for the grey and black die were 2 and 1.8 pence. Hence, there was an incentive for each player individually to choose the white die. However, if all of the players chose that die, the probability of the team winning was only 2%, and the expected payoff -0.8 pence.

Figure 2 shows the underlying structure of the experiment. The choice of die reflected the *intended* contributions of the players while the team's result was a function of the *actual* contributions. We predicted a main effect of intention: the same outcome of roll will elicit different responsibility attributions dependent on the choice of die. We also predicted a main effect of outcome: responsibility attributions for a given die will vary with the outcome of rolling this die. Finally, previous research suggested that outcomes will affect participants' responsibility ratings more strongly than intentions (Cushman et al., 2009).

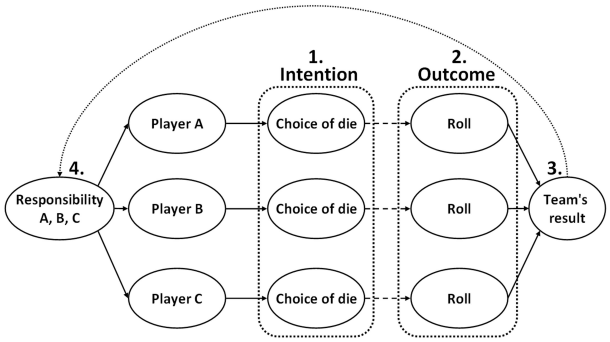


Figure 2: Underlying structure of the paradigm. Numbers 1. - 4. indicate the different components of each round.

Method

Participants and Materials 80 participants from the UCL subject pool participated for the chance of winning one of six amazon vouchers worth £150 in total. 55 participants were female and the mean age was 23.2 (5.94). With the second part of the experiment added at a later stage (see Procedure), 35 participants performed only the first part of the experiment, whereas the remaining 45 participants performed both. The study was conducted online and programmed with Adobe Flash.

Procedure Participants were informed that the experiment would take 20 minutes and that their task was to evaluate the behaviour of players engaged in an experimental game by attributing credit for wins and blame for losses. Participants read a description of the three dice which made it clear that they differed in terms of both probability distribution and price. A practice round served to familiarize participants with the structure of the game. After the practice round, participants had to answer questions to ensure that they had understood the rules of the game correctly. The game was then played for 20 rounds.

On each round, participants saw a table that showed for each player which die she had chosen, the outcome of having rolled that die and the amount won or lost in that round. In Figure 1, Player *C* chose the grey die and rolled a 5. Her payoff was 7p since she paid 3p for rolling the grey die and each player received 10p for winning this round. Players were indicated by capital letters which changed in each round. This was done to prevent participants from forming an overall impression about individual players. The header above the table showed the sum of points and changed its colour from green to red according to whether the round was won or lost. For each player, participants attributed blame for losses or credit for wins, by moving a slider ranging over a scale from 0-10. Its endpoints were labelled 'low' and 'high'. In line with the result of the round (loss/win), the label (blame/credit), color (red/green) and position of sliders (middle to left/ middle to right) of the last column changed.

45 of the 80 participants also completed a second stage of the experiment. Those participants were informed after the 20th round that they would see 14 novel situations that could

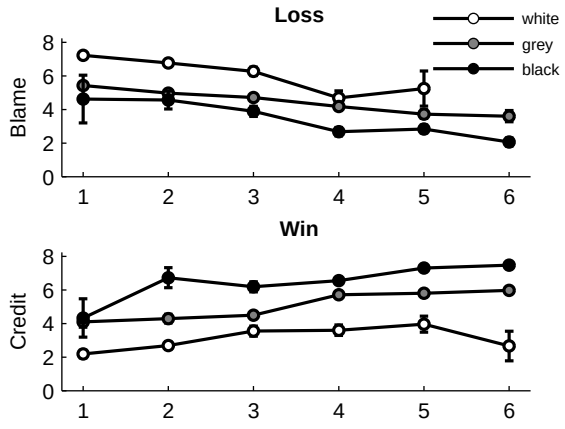


Figure 3: Mean responsibility ratings for each combination of die and outcome for both losses and wins. Lines represent the different dice and values on the x-axes indicate the outcome of rolling each die.

have occurred in the game and which were of special interest to the researchers. As explained below, the test cases were chosen so as to enable a fine assessment of the weight assigned to intentions and outcomes. The order of these test cases was randomized. Finally, participants were asked to indicate in a textbox whether they had focused on the choice of die, the outcome, or both.

Results

Mean Responsibility Ratings

In order to evaluate the effects of choice of die and outcome of roll for the first stage of the experiment, we ran separate 3 (Die) x 6 (Roll) ANOVAs for both wins and losses. Figure 3 shows the mean responsibility attributions as a function of choice of die and outcome of roll. For wins, there was a significant main effect of Die $F(2, 2472) = 87.10$, $p < .001$, $\eta^2 = .066$ and of Roll $F(5, 2472) = 9.53$, $p < .001$, $\eta^2 = .019$, as well as an interaction effect $F(10, 2472) = 1.91$, $p < .05$, $\eta^2 = .008$. For losses, there was a significant main effect of Die $F(2, 2327) = 31.62$, $p < .001$, $\eta^2 = .027$ and of Roll $F(5, 2327) = 15.31$, $p < .001$, $\eta^2 = .032$, but no interaction effect ($p > .05$).

To qualify these results, we ran linear contrasts on Die and Roll for both wins and losses. For wins, there was a significant positive linear trend of Die as well as for Roll. For losses, there was a significant negative linear trend of both Die and Roll (all p's $< .001$).

These analyses show that overall, both the choice of die and the outcome of its rolling influenced participants' responsibility ratings. However, the results cannot reveal how individual participants weighted these two factors. To find out, we conducted individual regression analyses, and report them below.

Regression Analysis

First, we ran the following three separate regression analyses based on the overall data (80 participants x 20 rounds x 3 ratings data points):

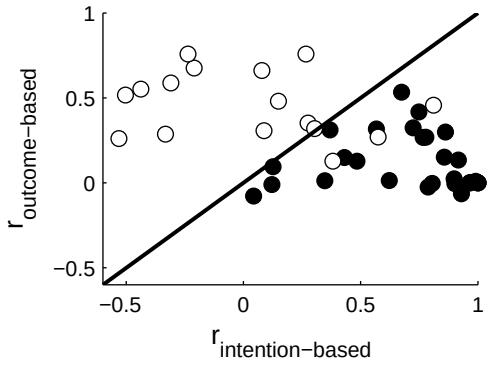


Figure 4: Scatterplot of correlations with outcome-based model and intention-based model. Black circles indicate participants classified as intention-based (N = 29), white circles indicate participants classified as outcome-based (N = 16).

$$\text{intention-based model: } responsibility = \beta_0 + \beta_{1\ die} \quad (1)$$

$$\text{outcome-based model: } responsibility = \beta_0 + \beta_{1\ roll} \quad (2)$$

$$\text{mixture model: } responsibility = \beta_0 + \beta_{1\ die} + \beta_{2\ roll} \quad (3)$$

All three regression models accounted for a significant amount of the variance in the data (see Table 1).

Evaluation of Test Cases

To break the results down even further, we ran the regression models for each individual participant. Based on the magnitude of the correlation with the intention-based regression model versus the outcome-based regression model, we grouped the 45 participants who completed the second stage of the experiment in two groups. We used this grouping to predict how participants would attribute responsibility for the chosen test cases (described below). Figure 4 shows how well the behaviour of the classified participants was predicted for the test cases.

The test cases were constructed to enable a fine analysis of the relative weights assigned by participants to intentions versus outcomes. It should be noted that in the first 20 rounds of the experiment the choice of die and outcome of roll were highly correlated due to the chosen probability distributions ($r = .68$, $p < .001$). In contrast, for the test cases the choice of die and outcome of roll were uncorrelated ($r = 0$). This shows that these test cases indeed created situations that could

Table 1: Results of overall regression analyses.

Model	R^2	F	$p <$	β	t	$p <$
intention-based	.268	1757.70	.001	0.518^a	41.93	.001
outcome-based	.219	1346.33	.001	0.468^b	36.69	.001
mixture	.303	1042.31	.001	0.370^a	24.02	.001
				0.238^b	15.48	.001

$a = \beta_{die}$, $b = \beta_{roll}$

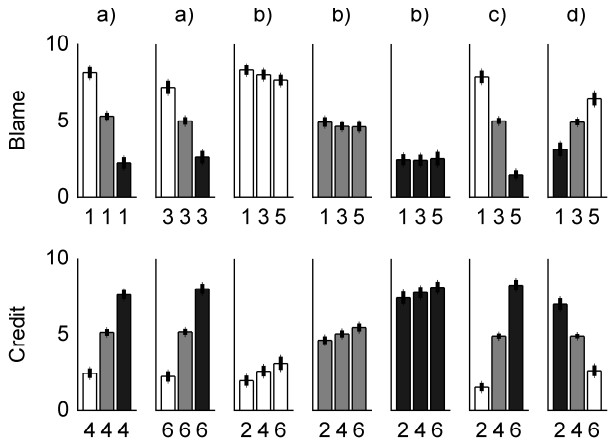


Figure 5: Mean responsibility ratings of *intention-based* participants for 14 test cases. The top row shows losses and the bottom row wins. The values on the x-axes indicate the outcome. The colours of the bars indicate the dice.

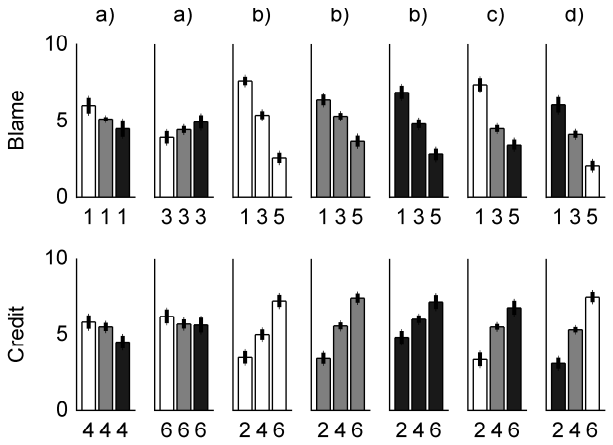


Figure 6: Mean responsibility ratings of *outcome-based* participants.

be used to distinguish between intention-based and outcome-based participants.

Figure 5 shows how the 29 intention-based participants attributed responsibility in the test cases. Figure 6 shows the responsibility attributions for the 16 participants who had been classified as outcome-based. The test cases can be categorized into 4 groups: a) different dice, same roll; b) same dice, different rolls; c) congruent; d) incongruent. 'Congruent' means that the quality of die and outcome of roll corresponded (i.e. the expensive die led to a high and the cheap die to a low outcome); 'incongruent' means that the quality of die and the outcome of roll mismatched.

Inspection of the graphs validates the original partition. First, in the congruent test cases ('c') - which serve as a manipulation check - both groups show the same pattern of attributions, with that for the intention group being steeper than that for the outcome group. For the intention test cases ('a'), the differences in attributions are large for participants in the intention group and small for participants in the outcome group. An opposite pattern of attributions is evident with the outcome test cases ('b'): there the intention group exhibits small differences and the outcome group exhibits large differences. Finally, and most interesting, the pattern of attributions reverses in the incongruent cases ('d'). Despite the fact that in these situations the expensive die led to the lowest outcome, the intention-based participants attribute the least blame to this player for the loss (Figure 5, top) and the most credit for the win (bottom). In contrast, the attributions of the outcome-based participants for these cases closely follows the number rolled, independent of the choice of die (Figure 6).

Discussion

The current study investigated the influence of intended versus actual contribution on the attribution of responsibility in a group context. We found that both intention and outcome exerted a significant influence on participants' attributions. Furthermore, we provided evidence that individuals differ in the extent to which they base their attributions on intentions or outcomes.

Our experimental procedure allowed us to dissociate intentions from outcomes and created a situation in which participants played the role of an external judge. We found that the majority of participants were better explained as having focused on intended rather than actual contributions. Methodologically, the current experiment shows that it is important to analyse the data on the level of individual participants. While on an aggregate level, it seems that participants are weighting both choice of die and outcome of roll to determine their responsibility attribution (see Figure 3), more careful analyses reveal that most participants actually tend to either focus on the intention or the outcome (see Figure 4).

At this point, we can only speculate about the factors driving these interindividual differences. Different interpretations of the notion of responsibility could have influenced participants' behaviour. Outcome-based participants might have endorsed a *causal* conception of responsibility. Accordingly, players that rolled high numbers were credited higher since their contributions caused the win. Intention-based participants, on the other hand, might have used a *moral* conception of responsibility. Hence, players were judged for their choice of die which reflected their underlying attitude towards the team. Alternatively, the results could reflect interindividual differences in the ability or motivation to mentalize. We would assume that people who find it hard to take another person's perspective are more likely to focus an the actual outcome rather than the underlying intention. We are planning to use a simplified version of the employed paradigm to test this hypothesis on a patient group. Finally, outcome-based participants' ratings might have been influenced by beliefs about the gambling-competence of players. On this view, rolling a high number with the cheap die reflects a special ability deserving credit. Some of the participants' written comments confirm the influence of such arguably non-normative considerations.

Why did we find a relatively stronger effect of intentions when previous studies postulated the existence of an outcome bias (Cushman et al., 2009)? Several differences between studies that draw their conclusions from economic games, such as the ultimatum game, and our study could potentially explain these divergent results. First of all, most of the studies in the economic literature were interested in exploring perceived fairness and not directly in responsibility attributions. Although we presume that these notions are tightly linked, it might be that considerations about fairness and responsibility can lead to different results. Second, the participants in those studies directly experienced the outcomes, while inferring the intentions of the other player was not incentivised. In our study, in contrast, participants acted as independent external judges. It is, hence, less likely that their attention was biased towards outcomes. In a future study, we aim to explore how the patterns of attribution change when participants actively take part in the game.

An important feature of the employed experimental paradigm is its potential to explore different combination functions between the individuals in the group. Gerstenberg and Lagnado (2010) have shown that the way in which individual contributions are translated into group outcomes significantly influences people's responsibility attributions. Accordingly, an identical individual contribution can lead to very different responsibility attributions as a function of the group context. While the current experiment used an additive combination function for the contributing players, we will investigate in future experiments how attributions change when the rule of the game reflects a minimum function (i.e., the group wins if no player rolls a 1) or a maximum function (i.e., the group wins if at least one player rolls a 6). Are participants more likely to focus on the actual rather than the intended contribution when the combination function is non-compensatory?

Finally, our paradigm can be used to explore how uncertainty affects responsibility attributions. In everyday life, we do not have direct access to other people's intentions. Rather, we try to infer the intention from a person's behaviour. Our paradigm allows us to model this situation. Instead of revealing all the information to the participant, we will only show the outcomes of rolling the dice but not which dice the players have chosen. We can then compare participant's ability to infer the underlying intentions from observed outcomes with an ideal Bayesian learner and evaluate in how far their attributions can be explained by their knowledge about the players.

In conclusion, the current study explored the influences of intentions versus outcomes on responsibility attributions in a group context. We found that a majority of our participants focussed on the intention rather than on the outcome. We introduced a novel experimental paradigm which is flexible enough to lend itself to the investigation of future questions that will help to disentangle the complex relationship between control, intention, outcome and responsibility attributions.

Acknowledgments

TG is the beneficiary of a doctoral grant from the AXA Research Fund. DL & TG were supported by the ESRC Centre for Economic Learning and Social Evolution; YK by the Israel Science Foundation 539/07.

References

Alicke, M. (2000). Culpable control and the psychology of blame. *Psychological Bulletin*, *126*, 556–574.

Baron, J., & Hershey, J. (1988). Outcome bias in decision evaluation. *Journal of Personality and Social Psychology*, *54*, 569–579.

Cushman, F., Dreber, A., Wang, Y., & Costa, J. (2009). Accidental outcomes guide punishment in a trembling hand game. *Plos One*, *5*, 1–7.

Dufwenberg, M., & Kirchsteiger, G. (2004). A theory of sequential reciprocity. *Games and Economic Behavior*, *47*, 268–298.

Falk, A., Fehr, E., & Fischbacher, U. (2008). Testing theories of fairness: Intentions matter. *Games and Economic Behavior*, *62*, 287–303.

Fehr, E., & Schmidt, K. (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics*, *114*, 817–868.

Fischhoff, B. (1975). Hindsight $\neq$ foresight: The effect of outcome knowledge on judgment under uncertainty. *Journal ol Experimental Psychology: Human Perception and Performance*, *1*, 288–299.

Gerstenberg, T., & Lagnado, D. (2010). Spreading the blame: The allocation of responsibility amongst multiple agents. *Cognition*, *115*, 166–171.

Güth, W., Schmittberger, R., & Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior & Organization*, *3*, 367–388.

Hardin, G. (1968). The tragedy of the commons. *Science*, *162*, 1243–1248.

Hart, H. (1985). *Punishment and Responsibility*. Wadsworth Publ. Co.

Knobe, J. (2003). Intentional action in folk psychology: An experimental investigation. *Philosophical Psychology*, *16*, 309–324.

Lagnado, D., & Channon, S. (2008). Judgments of cause and blame: The effects of intentionality and foreseeability. *Cognition*, *108*, 754–770.

McCabe, K., Rigdon, M., & Smith, V. (2003). Positive reciprocity and intentions in trust games. *Journal of Economic Behavior and Organization*, *52*, 267–275.

Mitchell, T., & Kalb, L. (1981). Effects of outcome knowledge and outcome valence on supervisors' evaluations. *Journal of Applied Psychology*, *66*, 604–612.

Shaver, K. (1985). *The Attribution of Blame: Causality, Responsibility, and Blameworthiness*. Springer-Verlag New York.

Williams, B. (1981). *Moral luck*. Cambridge University Press.

Learning Perceptual Aspects of Diagnosis in Medicine via Eye Movement Modeling Examples on Patient Video Cases

Halszka Jarodzka∗, Thomas Balslev†, Kenneth Holmqvist‡,
Marcus Nyström‡, Katharina Scheiter∗, Peter Gerjets∗, & Berit Eika○
∗Knowledge Media Research Center, Germany
†Viborg Hospital, Department of paediatrics, Denmark
‡Lund University, Sweden
○Aarhus University, Centre of Medical Education, Denmark

Abstract

Complex tasks with a visually rich component, like diagnosing seizures based on patient video cases, not only require the acquisition of conceptual but also of perceptual skills. Medical education has found that besides biomedical knowledge (knowledge of scientific facts) clinical knowledge (actual experience with patients) is crucial. One important aspect of clinical knowledge that medical education has hardly focused on, yet, are perceptual skills, like visually searching, detecting, and interpreting relevant features. Research on instructional design has shown that in a visually rich, but simple classification task perceptual skills could be conveyed by means of showing the eye movements of a didactically behaving expert. The current study applied this method to medical education in a complex task. This was done by example video cases, which were verbally explained by an expert. In addition the experimental groups saw a display of the expert's eye movements recorded, while he performed the task. Results show that blurring non-attended areas of the expert enhances diagnostic performance of epileptic seizures by medical students in contrast to displaying attended areas as a circle and to a control group without attention guidance. These findings show that attention guidance fosters learning of perceptual aspects of clinical knowledge, if implemented in a spotlight manner.

Keywords: example-based learning; eye tracking; expertise; attention; medical education

With progressing technical development, complex visualizations are increasingly in use for tasks, such as interpreting weather maps (Canham & Hegarty, 2010), classifying fish locomotion (Jarodzka, Scheiter, Gerjets, & Van Gog, 2010), driving (Underwood, Chapman, Brocklehurst, Underwood, & Crundall, 2003), or air traffic control (Helleberg & Wickens, 2003), to name just a few diverse examples. Dealing with such tasks requires not only knowledge about facts in this domain (i.e., conceptual knowledge), but also substantial visual search (i.e, perceptual skills). Sophisticated perceptual skills enable people in these professions to rapidly perceive the relevant out of the irrelevant and interpret it correctly. A large body of research has already shown that experts exceed novices in those skills (e.g., Antes & Kristjanson, 1991; Canham & Hegarty, 2010; Charness, Reingold, Pomplun, & Stampe, 2001; Jarodzka et al., 2010; Underwood et al., 2003; Van Gog, Paas, & Van Merriënboer, 2005; Vogt & Magnussen, 2007).

Perceptual Skills in Medical Education

The extensive presence of visually rich tasks and thus, the importance of sophisticated perceptual skills is especially true for the medical domain. Many medical imaging techniques developed only recent (like fMRI, CT, 3D displays). The task to diagnose medical images can also be seen as a visually complex task. In particular, since research could already show expertise differences on a perceptual level (e.g., Krupinski, 2005; Krupinski et al., 2006; Kundel, Nodine, Krupinski, & Mello-Thoms, 2008; Lesgold et al., 1988; Nodine, Kundel, Lauver, & Toto, 1996). However, not only diagnosing medical images is difficult from a perceptual perspective. Also the diagnosis in real-life situations of diseases that manifest in occasionally occurring behavioral patterns, like seizures, is difficult on a perceptual level (Balslev et al., in preparation). In the case of diagnosing seizures it is crucial to recognize the important features, which distinguish the seizure from normal behavior. Those features, however, might be short-term, subtle, time-sensitive, and not salient compared to other features.

Little children display many different movements. In rare cases some of these movements may be symptoms of diseases. In particular for small children that cannot be questioned it is important to carefully observe their movements for diagnosing certain diseases. The example we focus on are epileptic seizures. Epileptic seizures can be distinguished according to whether they involve one or both hemispheres of the brain ("International Classification of Mental and Behavioural Disorders (ICD-10)", 2006): if only one brain hemisphere is involved they are classified as partial seizures whereas if both hemispheres are involved they are classified as general seizures (here: spasms). Both seizure types have a normal behavior counterpart (i.e., differential diagnosis) with which they can easily be confused: epileptic seizures are easily confused with benign sleep myoclonus (BSM; Egger, Grossmann, & Auchterlonie, 2003), whereas spasms are easily confused with infantile masturbation (IM; Hansen & Balslev, 2009). A *general seizure* (spasm) is characterized by bilateral movements that can be spasmic or jerky, the face is affected, the infant briefly looses consciousness / awareness, and the movements are not stopped by touching the child. BSM is also characterized by bilateral, jerky movements, however, the face is not involved, the child is asleep, and the movements may rather worsen by touching the child. A *partial seizure* is characterized by lateral movements that can be spasmic or jerky, the face is affected, the infant looses briefly consciousness / awareness, and the movements are not stopped by touching the child. IM is also characterized by lateral, rather tension movements, however, the face is not in-

volved, the child is awake and conscious, and the movements stop by touching the child.

In order to convey diagnostic skills, medical education focused in its beginnings on the role of biomedical knowledge (Feltovich & Barrows, 1984; Kuipers & Kassirer, 1984; Lesgold et al., 1988). Biomedical knowledge is the knowledge contributing to the understanding of the functioning and dysfunctioning of the human body and gained during textbook or lecture study. It is composed of conceptual or factual knowledge. Thus, biomedical knowledge may be described as "inert knowledge". Knowledge is inert, if it is learnt in a formal setting and can be expressed by the student as facts without the ability to apply it in a real world situation (Whitehead, 1929). The focus on conveying inert knowledge in education has been extensively criticized by educational psychologists (e.g., Pozzi, Noss, & Hoyles, 1998).

In line with those findings, current research on medical education emphasizes that besides biomedical knowledge, also clinical knowledge is important (Boshuizen & Schmidt, 1992; Patel, Evans, & Groen, 1989a, 1989b; Patel & Kaufman, 1995; Schmidt & Boshuizen, 1992, 1993). Clinical knowledge is composed of manifestations, classifications, and treatments of diseases and it is gained during clinical praxis. One important aspect of clinical knowledge may be seen as being of perceptual nature (Manning, Gale, & Krupinski, 2005). As described above, novices have severe deficiencies on this level. However, there is little research so far, focusing on this aspect of clinical knowledge (Chen, Gale, & Evans, 2009; Jarodzka, Gog, Dorr, Scheiter, & Gerjets, in preparation; Litchfield, Ball, Donovan, Manning, & Crawford, 2008). Thus, the aim of the current study is to enhance the perceptual part of information-processing as part of medical expertise.

Conveying Perceptual Skills by Modeling Examples

An approach that has recently been developed to foster perceptual skills are eye movement modeling examples (Jarodzka et al., in preparation; Van Gog, Jarodzka, Scheiter, Gerjets, & Paas, 2009). To develop such modeling examples an expert model is recorded while performing a task. In addition, the model explains hers/his actions and hers/his eye movements are recorded. In a second step, those recordings are replayed to a student as an educational video. Those videos can be seen in the tradition of example-based learning and modeling. Example-based learning has been shown to be a powerful instructional method in early skill acquisition. Examples demonstrate a problem solution to students, either by providing them with a written, worked-out problem solution to study (i.e., worked examples; see, Atkinson, Derry, Renkl, & Wortham, 2000; Sweller, Van Merriënboer, & Paas, 1998) or by allowing them to observe an expert performing the task live or on video (i.e., modeling examples; Bandura, 1977; Collins, Brown, & Newman, 1989). For cognitive tasks, modeling examples require the model to verbalize his/her cognitive actions while performing the task (e.g., Wouters, Paas, & Merriënboer, 2008).

For cognitive tasks that require the processing of complex visual information, it is crucial that the student not only hears the expert's verbal explanations, but can also see the material the expert is looking at. However, this may not suffice. As Bandura (1977) noted, to learn from modeling examples students have to attend to the important features of the modeled behavior. The verbal explanations of the expert can only guide the students' attention to the important features of the material when students know exactly what the expert is referring to. However, the chance that they simultaneously attend to the same features is very small, as was shown by the eye tracking research described above. Thus, when learning from modeling examples that involve complex visual material, novices might need attention guidance to those task aspects that the expert is attending to. Otherwise, especially on dynamic tasks, they may miss important information relevant for understanding and learning from the example.

Although several studies exist on attention guidance via cueing instructional material (for a review: Koning, Tabbers, Rikers, & Paas, 2009), the decision concerning which information will be highlighted and when often remains arbitrary. In particular, since research has shown that experts cannot estimate the knowledge level of novices appropriately (Hinds, 1999). Thus, it is very unlikely that they would be able to estimate appropriately, where to place a cue for a novice. On the other hand, research has shown that choosing a cue based on eye movements of successful problem solvers, enhances the probability for correctly solving an insight problem (Grant & Spivey, 2003). The question remains, however, whether attention guidance based on eye movements, not only as a single cue but the entire perceptual process of the expert, cannot just influence insight problem performance on the task at hand, but also enhance learning. Learning refers to the resilient change in a person's knowledge about a task that enables him or her to independently perform that task after practice (Simon, 1983).

In an earlier study, we could show that students' attention can be guided directly by the eye movements of an expert and that this influences learning (Jarodzka et al., in preparation). We developed modeling examples for a classification task in which an expert's eye movements and verbal explanations were recorded while he was performing this task. Participants who studied examples in which the expert's eye movements were displayed either as a dot on the fixated area or by blurring non-fixated areas (spotlight display), closely followed the expert's gazes during example study. The spotlight display led to significant improvements in learning in terms of visual search during the test, and the dot display led to enhanced classification performance on the test. These findings showed that guiding students' attention can go beyond guiding thought, to guiding learning.

Still, two open questions remain. First, since none of the displays was optimal, the current study aimed at improving both types of display. The dot display partially occluded relevant problem features. Thus, instead of a solid dot, the ex-

pert's gazes are displayed as a circle. For the spotlight display, we decided to use a less intrusive blurring so that a holistic impression of the overall scene can be gained. This was done by compressing the video on non-attended areas. This procedure has shown to be well accepted by viewers (Nyström & Holmqvist, 2007). Second, although the classification task in the study described above was visually rich, the task in itself was simple. For this reason the benefit of this instructional approach might not have fully unfolded, because the task itself was too easy. Thus, the current study uses a medical diagnosis task based on video cases, which is not only a visually rich task, but is also composed of a complex underlying decision tree.

Research Question - Hypothesis

In line with prior research, we hypothesize that attention-guidance based on expert's eye movements will foster learning of perceptual skills not only in a simple classification task, but also in medical diagnosis. Since both ways of display were improved, we assume learning to enhance in terms of a better diagnostic performance.

Method

Participants and Design

Participants were 60 medical students in their final year of the University of Aarhus (age: M = 26.57 years, SD = 2.03; 41 female), who had no prior knowledge on the task and had normal or corrected-to normal vision. They had been randomly assigned to one of three conditions (n = 20 each): (1) control condition with no attentional guidance, (2) attentional guidance by a circle on fixated areas based on the model's eye movements (circle display), (3) attentional guidance by blurring non-attended areas and leaving fixated areas sharply displayed (spotlight display).

Apparatus and Materials

Eye tracking equipment The expert model's eye movements were recorded with a SMI High Speed eye tracking system with a temporal resolution of 240 Hz and the iView X 2.2 software. These eye tracking data were edited with BeGaze 2.3 software (www.smivision.com) and self-programmed MatLab algorithms. All video material was presented to the participants via Experiment Center 2.2. The questionnaire in the testing was presented via e-prime 2.0 software.

Modeling examples The modeling examples for the control group consisted of four digital videos (.avi format), sized 720 $\times$ 576 pixels and presented in fullscreen on a 1280 $\times$ 1024 pixels resolution (corresponding to 17.07 $\times$ 13.65 inches). Each video depicted a single infant (between 4 hours and 8 months old), whereby two infants deployed behavioral patterns corresponding to a focal seizure and two infants deployed different types of normal behavior (benign sleep myoclonus, infantile masturbation). The original sound was removed from the videos, because parents and clinical staff were talking, which would disturb the use of verbal explnations. The duration of the videos was between 71 and 103 seconds. All videos included a spoken description and diagnosis of the behavior by the expert model. The expert was a physician of pediatric neurology, with extensive experience in diagnosing epileptic seizures. Rather than using the expert's natural performance of these tasks as an example we decided to instruct the expert to behave didactically, that is, to explain to novice students what the relevant aspects of the behavioral pattern shown in each video are. Each recording was replayed to the expert so that he could self-evaluate the replay data based on a number of statements (e.g., for a novice student, the disease is explained in enough detail, in comprehensible terms, et cetera; cf. Jucks, Schulte-Löbbert, & Bromme, 2007), and if necessary, he could re-record it. This was done, because a prior study had shown that experts use knowledge-based shortcuts in verbal and eye tracking data due to automated processes as well as using many technical terms that a novice student would not understand (shortcuts in this domain, cf. Balslev et al., in preparation).

In the circle display condition, participants received the same examples as the control group but those additionally included the expert's eye movements. These were created using the manufacturer rendered "fixation scanpath display function". The saccadic definition was set at a peak velocity threshold of 40/s. The fixations were displayed as yellow circles with a line thickness of one pixel and a gaze trail for a temporal window of 1 sec. In the spotlight display condition, potentially distracting features in the unattended areas were filtered out. That is, the focus of the expert's attention (with a radius of 32 pixels) was visible as usual, whereas the areas surrounding it were blurred' by off-line foveation via video compression on non-attended areas (Nyström & Holmqvist, 2007). Figure 1 shows a screen shot from each of the three conditions.

Tests Participants were shown six new realistic videos for a mean duration of 31.00 seconds (SD = 18.35) of different children displaying different types of behavior (3 $\times$ seizures and 3 $\times$ normal behavior). The duration of the testing video depended on the duration of the seizure / normal behavior. Afterwards, their diagnositc performance was assessed by answering multiple-choice questions on those videos: (1) indicating from a list of body parts, which was moving, (2) indicating from a list the type of the movement, (3) indicating, whether the face was involved and whether or not this was important for the diagnosis, (4) indicating the level of consciousness of the child, (5) indicating, whether awaking the child would change the movement, if child was asleep, and (6) indicating, whether touching the child changes the movement.

Procedure

The recording ran in individual sessions of approximately 45 minutes each. At the beginning, participants filled in a questionnaire on their prior knowledge in this task and their de-

Table 1: Screenshots from the three conditions used in the study.

Control	Dot display	Spotlight display
		

mographic data. Then, they received a short introduction to the topic, stating very general information on seizures and the importance to distinguish them from normal behavior. Then, the learning phase started. Participants were told that they will subsequently receive videos of the to-be-learned disease, where an expert explains the according movements and behavioral pattern. Depending on the condition, they were told that they will additionally see where the expert's attention was attracted to on the video. Before watching the learning videos, participants received the age, gender, and a short problem description of the patient.

In the testing phase the testing videos were replayed once. Afterwards, each video disappeared, resulting in a blank screen. Then, the participants had to answer the multiple-choice questions in an arbitrary order. This procedure was repeated for six new patient video cases.

Data Analysis

Test performance The construction and scoring of the performance measure was derived from a task analysis and by the help of domain experts. Accordingly, to diagnose a focal seizure, the following guidelines should be applied: (1) correctly stating which part of the body is involved in the movement, (2) correctly stating how this part moves, (3) correctly stating, whether the face was involved and whether or not this is important for the diagnosis, (4) correctly indicating the child's level of consciousness, (5) correctly indicating, whether awaking the child would change the movement, if child was asleep, and (6) correctly indicating, whether touching the child would change the movements. Hence, participants could receive a maximum of six points per video (1 point for each category).

Results

For all statistical tests reported here, a significance level of .05 is used. Means and standard deviations for each condition are given in Table 2.

Table 2: Means (and SD) for Testing Performance for Diagnosing Seizures and Differential Diagnoses.

		Control group	Circle display group	Spotlight display group
Testing	Seizure diagnosis	3.25 (0.47)	3.38 (0.41)	3.90 (0.62)
	Differential diagnosis	3.78 (0.59)	3.48 (0.81)	3.57 (0.69)

Seizure Diagnosis

An ANOVA showed significant differences between conditions in performance on the multiple choice test, $F(1, 59) = 9.13$, $p < .01$, $\eta_p 2 = .24$. Bonferroni post-hoc tests indicated that the spotlight display condition outperformed the circle display condition ($p < .01$) and the control condition ($p < .01$), while the control condition and the circle condition did not differ significantly.

Differential Diagnosis

An ANOVA showed no significant differences between conditions, $F < 1$.

Discussion

This experiment showed that attention guidance based on displaying expert's eye movements in video-based modeling examples fostered learning in terms of improved diagnostic skills. Participants in the spotlight display group outperformed the control and the circle display group in diagnosing epileptic seizures. No differences were found in diagnosing normal behavior.

These findings leave us with two questions: (1) Why did both types of eye movement displays resulted in such differential effects and (2) why does this effect occur for diagnosing seizures, but not for diagnosing the differential diagnosis?

Considering the first question, the fact that the circle display did not enhance learning is surprising. Although, one study found a negative effect on learning by displaying the model's eye movements based on the manufacturers' gaze replay functions (Van Gog et al., 2009), three others found a positive effect of a comparable visualization (Chen et al., 2009; Litchfield et al., 2008; Jarodzka et al., in preparation). However, only the latter three studies used visually rich learning tasks. Adding information to a display in terms of eye movements, might only be a benefit for visually rich tasks. In this study, a visually rich task was also used, but what differed was the fact that the display was not a solid dot, as for the remaining three studies, but a fully translucent circle. Might it be that the total translucence in terms of a circle reversed the positive learning effect? This might be tested in future studies by investigating the students' ability or willingness to follow the circle during learning. Another possibility is that the circle display increased mental effort due to a noisy type of display? Van Gog et al. (2009) found a higher mental effort in their study for this display, which lead to detrimental effects on learning. This should be investigated in future research by assessing mental effort in the learning and the testing phase. In contrast, the spotlight display, which was rather an unintrusive guidance, might have enabled the students to infer the element behind the cue and thus, lead to a holistic impression of the behavioral pattern.

The second question, about the different effects for seizure and differential diagnosis, has two competing answers. First, it might be that detecting the symptoms of a seizure, strongly relies on interpreting the perceptual input. Whereas detecting the symptoms of types of normal behavior requires more conceptual knowledge. This type of knowledge, however, was not varied between the groups in the current study. Second, it might be that the lack of an effect for the differential diagnosis is due to the type of material used in this study. One type of normal behavior (benign sleep myoclonus) was quite easy to detect, because both children in training and in testing were asleep. Neither of the others was asleep. Thus, this cases might have been too easy. The other normal behavior (infantile masturbation) was trained with an older child, whereas the testing occurred with younger children. Thus, thus testing might have been too difficult. Both possible reasons should be investigated in future research with more cases.

In sum, the current study provided an interesting, first approach to train perceptual skills in medical diagnosis. Still many research questions remain. Not only the above mentioned issues should be further investigated, but also additional improvements of the training are conceivable. We trained the students in this case only for a short time and only within individual learning. Training students in longer sessions that allow them to re-view the cases and discuss them with peers might further improve the training. On the other hand, the training effects should be also investigated in a more direct manner via detailed analyses of potential influences on participants' eye movements.

References

Antes, J. R., & Kristjanson, A. F. (1991). Discriminating artists from nonartists b their eye-fixation patterns. *Perceptual and Motor Skills*, *73*, 893-894.

Atkinson, R. K., Derry, S. J., Renkl, A., & Wortham, D. (2000). Learning from examples: Instructional principles from the worked examples research. *Review of Educational Research*, *70*, 181-214.

Balslev, T., Jarodzka, H., Holmqvist, K., Grave, W. S. de, Muijtjens, A., Eika, B., et al. (in preparation). How do paediatricians do it? the influence of experience and training on attention and clinical reasoning.

Bandura, A. (1977). *Social learning theory.* Englewood Cliffs, NJ: Prentice-Hall.

Boshuizen, H., & Schmidt, H. (1992). On the role of biomedical knowledge in clinical reasoning by experts, intermediates, and novices. *Cognitive Science*, *16*, 153-184.

Canham, M., & Hegarty, M. (2010). Effects of knowledge and display design on comprehension of complex graphics. *Learning and Instruction*, *20*, 155-166.

Charness, N., Reingold, E. M., Pomplun, M., & Stampe, D. (2001). The perceptual aspect of skilled performance in chess: Evidence from eye movements. *Memory and Cognition*, *29*, 1146-1152.

Chen, Y., Gale, A. G., & Evans, A. (2009). The feasibility of vision-supported computer-based training in digital mammography. *Breast Cancer Research*, *11*, 10.

Collins, A. F., Brown, J. S., & Newman, S. (1989). Cognition and instruction: Issues and agendas. In L. Resnick (Ed.), (chap. Cognitive apprenticeship: Teaching the craft of reading, writing, and mathematics.). Mahwah, NJ: Erlbaum.

Egger, J., Grossmann, G., & Auchterlonie, I. A. (2003). Benign sleep myoclonus in infancy mistaken for epilepsy. *British Medical Journal*, *326*, 975-976.

Feltovich, P. J., & Barrows, H. S. (1984). Issues of generality in medical problem solving. In H. G. Schmidt & M. L. D. Volder (Eds.), *Tutorials in problem-based learning. new directions in training for the health professions.* (p. 128-142). Assen/Maastricht: Van Gorcum.

Grant, E. R., & Spivey, M. J. (2003). Eye movements and problem solving. guiding attention guides thought. *Psychological Science*, *14*.

Hansen, J. K., & Balslev, T. (2009). Hand activities in infantile masturbation: A video analysis of 13 cases. *European Journal of Paediatric Neurology*, *13*, 508-510.

Helleberg, J. R., & Wickens, C. D. (2003). Effects of data-link modality and display redundancy on pilot performance: An attentional perspective. *The International Journal of Aviation Psychology*, *13*, 189-210.

Hinds, P. I. (1999). The curse of expertise: The effects of expertise and debiasing methods on predictions of novice performance. *Journal of Experimental Psychology: Applied*, *5*, 205-221.

International classification of mental and behavioural disor-

ders (icd-10) [Computer software manual]. (2006). Genf: World Health Organisation (WHO).

Jarodzka, H., Gog, T. van, Dorr, M., Scheiter, K., & Gerjets, P. (in preparation). Guiding attentions guides thought, but what about learning? eye movements in modeling examples. *NN*.

Jarodzka, H., Scheiter, K., Gerjets, P., & Van Gog, T. (2010). In the eyes of the beholder: How experts and novices interpret dynamic stimuli. *Journal of Learning and Instruction*, *20*, 146-154.

Jucks, R., Schulte-Löbbert, P., & Bromme, R. (2007). Supporting experts' written knowledge communication through reflective prompts on the use of specialist concepts. *Journal of Psychology*, *215*, 237-247.

Koning, B. B. D., Tabbers, H. K., Rikers, R. M. J. P., & Paas, F. (2009). Towards a framework for attention cueing in instructional animations: Guidelines for research and design. *Educational Psychology Review*, *21*, 113-140.

Krupinski, E. A. (2005). Visual search of mammographic images: Influence of lesion subtlety1. *Academic Radiology*, *12*(8), 965 - 969.

Krupinski, E. A., Tillack, A. A., Richter, L., Henderson, J. T., Bhattacharyya, A. K., Scott, K. M., et al. (2006). Eye-movement study and human performance using telepathology virtual slides. implications for medical education and differences with experience. *Human Pathology*, *37*(12), 1543 - 1556.

Kuipers, B., & Kassirer, J. P. (1984). Causal reasoning in medicine: Analysis of a protocol. *Cognitive Science*, *8*, 363-385.

Kundel, H., Nodine, C., Krupinski, E., & Mello-Thoms, C. (2008). Using gaze-tracking data and mixture distribution analysis to support a holisitc model fort he detection of cancers on mammograms. *Academic Radiology*, *15*, 881-886.

Lesgold, A., Rubinson, H., Feltovich, P., Glaser, R., Klopfer, D., & Wang, Y. (1988). The nature of expertise. In M. T. H. Chi, R. Glaser, & M. Farr (Eds.), (p. 311-342). Hillsdale, NJ: Erlbaum.

Litchfield, D., Ball, L. J., Donovan, T., Manning, D. J., & Crawford, T. (2008). Learning from others: Effects of viewing another person's eye movements while searching for chest nodules. In B. Sahiner & D. J. Manning (Eds.), *Proceedings of spie medical imaging 2008: Image perception, observer performance, and technology assessment* (Vol. 9). San Diego, CA, USA: SPIE.

Manning, D. J., Gale, A., & Krupinski, E. A. (2005). Perception research in medical imaging. *The British Journal of Radiology*, *78*, 683-685.

Nodine, C. F., Kundel, H. L., Lauver, S. C., & Toto, L. C. (1996). Nature of expertise in searching mammograms for breast masses. *Academic Radiology*, *3*(12), 1000 - 1006.

Nyström, M., & Holmqvist, K. (2007). Deriving and evaluating eye-tracking controlled volumes of interest for variable-resolution video compression. *Journal of Electronic Imaging*, *16*(1).

Patel, V. L., Evans, D. A., & Groen, G. J. (1989a). Biomedical knowledge and medical reasoning. In D. A. Evans & V. L. Patel (Eds.), *Cognitive science in medicine: Biomedical modeling* (p. 53-112). Cambridge, MA: The MIT Press.

Patel, V. L., Evans, D. A., & Groen, G. J. (1989b). Reconciling basic science and clinical reasoning. *Teaching and Learning in Medicine*, *1*, 116-121.

Patel, V. L., & Kaufman, D. R. (1995). Clinical reasoning and biomedical knowledge: Implications for teaching. In J. Higss & M. Jones (Eds.), *Clinical reasoning in the health professions* (p. 117-128). Oxford, UK: Butterworth Heinemann.

Pozzi, S., Noss, R., & Hoyles, C. (1998). Tool in practice, mathematics in use. *Educational Studies in Mathematics*, *36*, 105-122.

Schmidt, H. G., & Boshuizen, H. P. A. (1992). Encapsulation of biomedical knowledge. In D. A. Evans & V. L. Patel (Eds.), *Advanced models of cognition for medical training and practice* (p. 265-282). New York, NY: Springer Verlag.

Schmidt, H. G., & Boshuizen, H. P. A. (1993). On acquiring expertise in medicine. *Educational Psychology Review*, *5*, 205-221.

Simon, H. A. (1983). Why should machines learn? In R. S. Michalski, J. G. Carbonell, & T. M. Mitchell (Eds.), *Machine learning: An artificial intelligence approach* (p. 25-38). Palo Alto, CA: Tioga.

Sweller, J., Van Merriënboer, J. J. G., & Paas, F. (1998). Cognitive architecture and instructional design. *Educational Psychological Review*, *10*, 251-296.

Underwood, G., Chapman, P., Brocklehurst, N., Underwood, J., & Crundall, D. (2003). Visual attention while driving: Sequences of eye fixations made by experienced and novice drivers. *Ergonomics*, *46*.

Van Gog, T., Jarodzka, H., Scheiter, K., Gerjets, P., & Paas, F. (2009). Attention guidance during example study via the model's eye movements. *Computers in Human Behavior*, *25*, 785-791.

Van Gog, T., Paas, F., & Van Merriënboer, J. J. G. (2005). Uncovering expertise-related differences in troubleshooting performance. combining eye movement and concurrent verbal protocol data. *Applied Cognitive Psychology*, *19*, 205-221.

Vogt, S., & Magnussen, S. (2007). Expertise in pictorial perception: Eye movement patterns and visual memory in artists and laymen. *Perception*, *36*, 91-100.

Whitehead, A. N. (1929). *The aims of education.* New York, NY: MacMillan.

Wouters, P., Paas, F., & Merriënboer, J. J. G. V. (2008). How to optimize learning from animated models: A review of guidelines based on cognitive load. *Review of Educational Research*, *78*, 645-675.

Structure Awareness in Action-Outcome Learning Eradicates the Detrimental Effect of Reinforcement Delays

W. James Grevile (GrevilleWJ2@Cardiff.ac.uk), Adam Cassar (AdCas43@Googlemail.com)
Mark Johansen (JohansenM@Cardiff.ac.uk), Marc J. Buehner (BuehnerM@Cardiff.ac.uk
School of Psychology, Cardiff University,
Tower Building, Park Place, Cardiff, CF10 3AT, Wales, UK

Abstract

Many studies of Action-Outcome Learning have demonstrated that reinforcement delays exert a detrimental influence on learning performance. Different theoretical perspectives offer varying explanations for this effect. A rational perspective suggests that as long as action-outcome pairings can be clearly recognized, delays should not interfere with the inductive process. Here we tested this idea by manipulating whether action-outcome contingencies were clearly identifiable as such by providing structural information in real time. In the absence of such information, we replicated the familiar detrimental effects of delay. Providing structural markers, and thus allowing easy identification of action-outcome pairings, eradicated this effect. Importantly, two additional experiments indicate that these results cannot be attributed to alternative explanations involving outcome salience or better awareness of timing. We conclude that when the environment allows Action-Outcome Learning to be conceptualized as a contingency learning task, learners are capable of covariation computation and immune to variations of response-outcome timing.

Keywords: Causality, Contiguity, Reinforcement, Structure, Computation

Introduction

The detrimental effect of a cause-effect delay on the learning of a causal relation is well established. However, the precise reason for this effect is still the subject of some debate. While it seems fairly intuitive that delayed causal relations might be more difficult to detect, and judged as weaker, compared to more immediate relations, this raises the question of how we ever manage to infer delayed causal relations of more than a few seconds. Yet we manage to do so routinely in day-to-day life. At the same time, laboratory experiments using basic stimuli have demonstrated that delayed causal relations of more than a few seconds could not be distinguished from non-contingent alternatives (Shanks, Pearson, & Dickinson, 1989). It therefore follows that in real-world causal induction, some other tangible source of information must be brought to bear that enables us to correctly identify delayed causal relations.

There have been a plethora of studies investigating the ability of humans to judge event contingencies (e.g. Shanks, 1987; Wasserman, Chatlosh, & Neunaber, 1983). A long-standing paradigm is the instrumental free-operant procedure (FOP), whereby participants evaluate the effectiveness of their responding (for instance pressing a key on a keyboard) in producing an outcome (such as a flash or a tone). These experiments are typically programmed with an invisible underlying trial structure, whereby the condition timeline is divided into several temporal segments. If a response is made during a particular segment, then an outcome will be scheduled to occur (with a certain probability) at the end of that segment. A key consideration that is often overlooked in such experimental designs is whether this trial structure is apparent. This may play a critical role in the mediation of empirical cues such as delay.

Several potential explanations for the effect of delay have been offered stemming from different theoretical motivations. Traditional associative accounts argue that causal induction is simply an extension of associative learning, and is as a consequence governed by the same principles as other forms of learning such as Pavlovian and instrumental conditioning. This perspective adopts the Humean assertion that temporal contiguity is necessary for learning to occur. Degradation of this contiguity leads to weaker increments of associative strength and thus universally attenuates learning.

Cognitive perspectives on causal learning, on the other hand, tend to focus on event contingencies. Most proponents of this view agree that the sensory input available to us, in the form of presence or absence of events, is computed to provide an assessment of the covariation between candidate causes and effects. In the simplest terms, the possible event combinations are as follows: Both cause and effect occur (c,e), the cause occurs without the effect (c~e), the effect occurs without the cause absent (~c,e), and neither cause nor effect occur (~c,~e). These event frequencies are often represented in a 2x2 contingency matrix, and form the basis for many different computational models of learning (see, e.g., Hammond & Paynter, 1983). Provided that this information can be discerned from the available evidence, contiguity is not essential.

The role of contiguity from this perspective is instead limited to determining whether or not events are classed as contingent. Longer intervals increase the likelihood of intervening events to occur between action and outcome, which compete for explanatory strength and place greater demands on processing and memory resources. Accordingly, where there is some temporal separation between cause and effect, the crucial decision revolves around deciding whether this constitutes a case of c,e, or separate cases of c,~e and ~c,e. The greater the delay, the more likely the latter becomes, and the effect will not be attributed to the cause. This is therefore known as the *attribution shift* hypothesis (Buehner & May, 2009).

Experiments by Buehner & May showed that by appealing to higher level knowledge, the detrimental effect of delay can be modulated (2002) and abolished completely (2004). Participants were presented with action-outcome learning tasks in different thematic scenarios. By manipulating the context using cover-stories, a delay between cause and effect was made to seem plausible by providing explicit information regarding the expected timeframe of the causal mechanism. In a scenario where participants evaluated the effectiveness of pressing a switch on the illumination of a lightbulb, one group of participants were told that the bulb was an ordinary bulb that should light up right away, while another group of participants was instructed that the bulb was an energy-saving bulb that lights up after a delay. For this latter group there was no decline in ratings with delay; delayed and immediate causal relations were judged as equally effective.

These findings were consistent with the knowledge-mediation hypothesis (Einhorn & Hogarth, 1986): reasoners have pre-existing ideas about specific mechanisms by which causes produce their effects, which in turn enables flexible interpretation of incoming evidence, including appraisal of delayed causal relations. However, a problem with this approach is circularity: if causal learning is governed by top-down assumptions regarding causal mechanisms, where does this knowledge come from in the first place?

Perhaps some causal knowledge is innate. Stimulus selectivity in rats (Garcia & Koelling, 1966) demonstrates that animals indeed have pre-existing conceptions about the types of stimuli that can elicit particular physiological reactions. It is therefore not unreasonable to suggest that animals (including humans) may likewise have some prior expectation about certain potential mechanisms, which may well include non-contiguous causal relations. Nevertheless, it seems appropriate to search for other means by which the connection between a proximal candidate cause and a distal effect may be bridged. Are there cues that can mitigate the impact of delay without recourse to knowledge of mechanism?

Our goal here was to create a paradigm by which the underlying trial structure could be made evident without appealing to prior knowledge, or manipulation of expectations using cover stories or thematic contexts. Instead, we aimed to convey this information using stimuli that are directly observable in the learning environment and thus demonstrate that empirical cues can be used to infer delayed causal relations without any prior cognitive bases. This was achieved by using a brief auditory tone to signal the end of each trial. This tone occurred regardless of whether an effect occurred or not, and if an effect was scheduled it occurred simultaneously with the tone. The tone thus marked the point at which an effect could potentially occur.

Our hypothesis represents a convergence of two traditionally opposing perspectives on causal learning. In accordance with associative learning theory, we predict a decline in causal ratings as delay increases and no additional information is provided. However, when the tone is introduced, providing markers that effectively reveal the delineation into trials, then contiguity becomes unimportant. The task will reduce to a simple contingency judgment, and we should see no delay-induced decline in ratings, as predicted by a computational account of causal learning.

Experiment 1

Method

Participants

33 undergraduate students from Cardiff University were recruited via an online participation panel. Participants included both males and females, with a modal age of 19 years. Either course credit or £3 payment was awarded for completion of the experiment. One participant failed to make any responses during two of the experimental conditions and thus was dropped from the analysis.

Design

The factors *trial length* (*2s*/*5s*) and *trial structure* (*apparent* vs. *not*) combined to produce four experimental conditions .Previous studies have found manipulation of trial length as an effective determinant of action-outcome delay, thus *2s* and *5s* were classed as *short delay* and *long delay* conditions respectively. For the *apparent* conditions, the end of each trial was signaled by an auditory tone, with the commencement of the next trial coinciding with tone offset. Meanwhile no additional cues were provided for the *not apparent* condition. Effectively, each trial ran seamlessly into the next, with no markers delineating one trial from the next (other than the occurrence of an effect). All participants experienced all four conditions, providing a 2x2 within-subjects design. The conditions were blocked such that the two *apparent* conditions were always presented one after the other, likewise for the two *not apparent* conditions. The order of which *apparent* or *not apparent* condition was presented first, or whether the *apparent* or *not apparent* block was presented first, was counterbalanced. At the end of each condition, participants were presented with the following question: "Please enter a rating from 100 to -100 to indicate the effect you think the button had on the triangle's behavior. 0 means it had no effect, +100 means it always made it light up, and -100 means it always prevented it from lighting up." The rating provided by participants constituted the dependent measure.

Apparatus, Materials & Procedure:

The experiment was conducted on an Apple "Mac Mini" computer running Microsoft Windows XP and Python 2.4.1, with a 17" LCD display, with standard headphones used to deliver the auditory stimulus. The stimuli consisted of an outline of an equilateral triangle and an image of a red circular button situated directly beneath it. Participants were free to click on this button with the mouse at any point. On doing so, the button stimulus 'depressed' for 500ms.

An effect constituted the triangle 'lighting up' (the transparent background became bright yellow and a 'glow' effect appeared around the triangle border) for 500ms. The

occurrence of the effect was determined probabilistically. If a response was made during the trial, P(e|c) was 0.7; if no response was made, P(e|~c) was 0.2. Only the first response in each trial altered the probability from 0.2 to 0.7, with subsequent responses having no influence.

For the apparent conditions, at the end of each trial, an auditory tone of 1000Hz was played for 500ms. This tone signaled the end of the trial, with the next trial beginning on termination of the tone. If an effect was scheduled, it occurred at this point of the trial to coincide precisely with the tone. For not apparent conditions, an equivalent 500ms was added to the end of each trial and the effect (if scheduled) occurred during this period. This ensured identical trial lengths and reinforcement delays between apparent and not apparent conditions. Each condition comprised 60 consecutive trials; total condition lengths were thus 150s and 330s for 2s and 5s conditions.

Participants were instructed to determine to what extent pressing the button caused or prevented the triangle from lighting up. Apparent conditions included the following additional instructions: "Each problem is divided into a series of trials. The end of each trial is marked by a beep. The triangle can only light up once per trial, and if it does so, it will light up at the end of the trial (i.e. to coincide with the beep)."

Results & Discussion

Causal Ratings

All analyses adopted a significance level of 0.05. One participant failed to make any responses during two of the experimental conditions and thus was dropped from the analysis altogether. One additional data point which was more than two standard deviations from the mean was also removed from the analysis for causal ratings. Figure 1 shows that ratings fell sharply in the *not apparent* conditions as trial length (and resultant action-outcome delay) was increased. However, a corresponding decline is not seen for the apparent conditions; there appears to be no difference between 2s and 5s. This suggests that the provision of trial structure information nullified the deleterious impact of delay.

A 2x2 within-subjects ANOVA corroborated these impressions, finding significant main effects of delay ($F_{(1,31)}$ = 7.276), trial structure ($F_{(1,31)}$ = 4.322), and a significant delay x structure interaction ($F_{(1,31)}$ = 4.719). This supports the original hypothesis. However we must exercise caution in the interpretation of these results. Because we employed a free-operant paradigm, it is possible that participants' response behavior differed between conditions, resulting in different objective response-outcome contingencies (cf. Buehner & May, 2003). If any such differences occurred were between the *apparent* and *not apparent* conditions, then manipulation of trial structure would be confounded with contingency and our results compromised. In addition, because participants were free to respond at any given time, there is no guarantee that increasing trial length will produce a concomitant increase in the action-outcome delay. A participant could respond at any point during the trial and therefore it is perfectly possible that contiguous cause-effect pairings will be experienced in both the 2s and 5s conditions. A closer inspection of the behavioral data is therefore warranted.

Behavioral Data

Response rate was calculated as the total number of responses, both reinforced and unreinforced, produced by participants across the entire duration of the condition and including all responses made during each trial. Mean action-outcome interval was calculated as the time between the first response in a given trial and the subsequent effect (if one occurred). If the response was unreinforced then this was not included in the calculation.

An analysis of behavioral data using 2x2 within-subjects ANOVAs on response rate and action-outcome interval revealed that, as expected, action-outcome intervals were significantly longer for trials of 5s length than for 2s ($F_{(1,32)}$ = 84.942) confirming that controlling trial length was effective in manipulating reinforcement delay. We also found an effect of trial length on response rate ($F_{(1,32)}$ = 28.437), which replicates earlier findings (e.g. Buehner & May, 2003). The important comparisons, however, were those involving trial structure. Specifically, if action-outcome intervals were significantly shorter for *apparent* than *not apparent* conditions, then our case for structural insight would be weakened by a mediation through experienced delay. Likewise, differences in response rate would entail different objective contingencies experienced across these conditions.

However, there was no significant main effect of trial structure on either response rate ($F_{(1,32)}$ = 0.814) or action-outcome interval ($F_{(1,32)}$ = 1.495); neither was there significant interaction between trial length and trial structure for response rate ($F_{(1,32)}$ = 0.026) or action-outcome interval ($F_{(1,32)}$ = 0.033). We can thus have confidence that our results concerning causal ratings are purely driven by structure information, and are not mediated by behavioral differences.

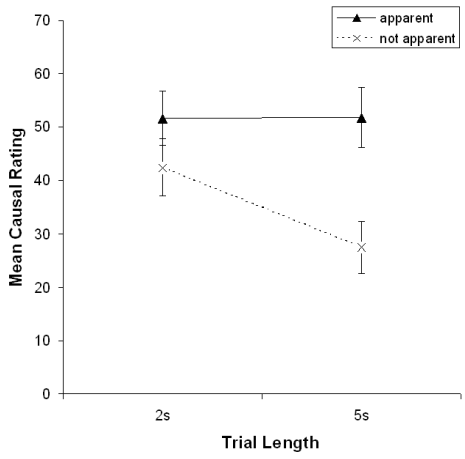


Figure 1: Mean causal ratings for Experiment 1. Error bars show standard error.

This finding suggests that causal learning in real time can, under certain conditions, be approached as a contingency learning task. When trial structure is apparent, contingency information can easily be discerned, and events accurately assigned to the cells of the contingency matrix. Under such circumstances, delays do not interfere with learning, as predicted by contingency-based or covariational models. Indeed in this case, judgments closely matched actual ΔP. Reinforcement delays thus are only detrimental to causal learning when they introduce ambiguity concerning response-outcome pairings.

It is important to note that our structural manipulation presented a tone simultaneously with the outcome. The tone therefore cannot act as a signal, bridging the temporal gap (Reed, 1992). There are however some other potential alternative explanations that must be ruled out.

Experiment 2A

Research in classical conditioning has demonstrated that increasing outcome salience increases the associative strength gained on each successive trial (e.g. Rescorla & Wagner, 1972). It could be argued that the tones marking the end of trials in the apparent conditions served to increase the saliency of the outcome, which coincided with them. If the causal learning process is subject to this property of associative learning, it might be responsible for alleviating the effect of delay. It could therefore be that our results are in fact driven by salience rather than structural insight.

To explore the effect of outcome salience, we modified the original paradigm such that under one set of conditions, outcome salience was increased, but without providing trial structure information. Accordingly, in one set of conditions, the triangle flash was accompanied by the same auditory tone used to provide structural markers in Experiment 1, adding to the salience of the outcome. The crucial distinction between this and the first experiment was that that here, the tone did not sound on occasions where there was no outcome, and thus did not convey trial structure information.

Method

Participants

32 participants, recruited as those in Experiment 1, completed the experiment to receive either £3 payment or course credit.

Design

Trial Length was either 2s or 5s as in the previous experiment, and Salience was either standard (no tone) or enhanced (tone present). Accordingly this gave four conditions which were presented in a blocked counterbalanced design as in the previous experiment.

Apparatus, materials & procedure

As before, except that in the *enhanced* conditions, the outcome was accompanied by the auditory tone, and participants received the following extra instructions: "When the triangle flashes, it will be accompanied by a tone." The *standard* conditions were identical to the *not apparent* conditions in the previous experiment. This and the following experiment were conducted in a small computer lab using Windows XP machines, and testing multiple participants at once. Partitions between machines and use of headphones ensured that each participant could focus exclusively on their own task.

Results & Discussion

Causal Ratings

Two data points which were more than two standard deviations from the mean were removed from the analysis. Figure 2 shows that causal ratings declined as trial length increased from 2s to 5s for both standard and enhanced conditions. There also appeared to be a slight positive influence of enhanced outcome salience. Most importantly there appeared to be little difference between 5s-salient and 5s-standard conditions, suggesting that increasing outcome salience alone cannot replicate the observed effect from Experiment 1.

A 2x2 within-subjects ANOVA found the expected significant main effect of delay ($F_{(1,30)} = 5.634$). There was no significant effect of salience ($F_{(1,30)} = 1.705$) nor was the salience x delay interaction significant ($F_{(1,30)} = 0.036$).

Behavioral Data

We found the expected main effects of delay on both response rate ($F_{(1,32)} = 33.512$) and action-outcome interval ($F_{(1,32)} = 355.372$). The effect of salience was non-significant on both response rate ($F_{(1,32)} = 0.199$) and action-outcome interval ($F_{(1,32)} = 1.361$). The interaction between salience and delay was non-significant for both response rate ($F_{(1,32)} = 1.779$) and action-outcome interval ($F_{(1,32)} = 1.643$).

The non-effect of increasing outcome salience was not wholly anticipated; the literature suggests that this manipulation might have enhanced learning (although such a trend, albeit non-significant, was seen). More importantly however, the decline from 2s and 5s remains for the salient condition, while there is no real difference between the 5s-

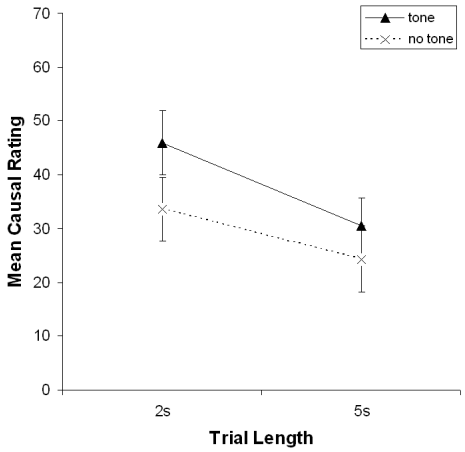


Figure 2: Mean causal ratings for Experiment 2A. Error bars show standard error.

salient and the 5s-standard conditions. We can thus rule out increased outcome salience as an alternative explanation for the effects observed in Experiment 1. We turn next to examine another potential confound, the presence of an auditory pulse.

Experiment 2B

In Experiment 1, the tone sounding at regular intervals (at the end of each trial) effectively produced a metronomic pulse that might have influenced participants' perception of time. In Experiment 1, the meter of this auditory pulse changed in line with trial length, such that there was either a relatively quick pulse occurring every 2s, or a slower pulse every 5s. Importantly, although trial length was different, there was one beat per trial in each case, so the marking of the passage of time was consistent for both conditions. Thus there was an imposed degree of perceptual similarity which could have accounted for the lack of difference between 2s and 5s when the tone was present.

To test this alterative explanation, we modified Experiment 1 in a fairly simple manner that would retain the auditory pulse, without necessarily providing information regarding trial structure. The tone was thus moved so that it did not occur at the end of each trial, thus demarcating one trial from the next, but rather occurred midway through each trial. Each tone was separated by the exact same interval, thus still providing a regular pulse, but was now no longer contiguous with the end (or beginning) of each trial, and therefore did not convey (useful) trial structure information.

Method

Participants

34 psychology students from Cardiff University received either £3 payment or course credit for participation.

Design

As for the previous experiments, four experimental conditions were produced by combining the factors *Trial Length* (2s or 5s) and *Pulse* (present/not present) and presented in a blocked counterbalanced design.

Apparatus, Materials & Procedure:

The apparatus, location and procedure was identical to the previous experiment, except that participants in the pulse conditions received the following extra instructions: "You will hear a tone sounding at regular intervals. This is a pulse to help you keep track of time."

Results & Discussion

Causal Ratings

Figure 3 suggests that the auditory pulse did not alleviate the effect of delay. Interestingly however, it does seem that the presence of the pulse did improve judgments of causality across the board; both for 2s and 5s, although it did not noticeably improve judgments at 5s relative to 2s. This general effect could be due to a slowing down of the internal pacemaker by the auditory pulse. Studies have provided evidence that human time perception is determined by a temporal oscillator, the frequency of which can be altered by interference from an imposed rhythm (Treisman, Faulkner, Naish & Brogan, 1990). Slowing the frequency means time seems to pass more quickly and the subjective duration of intervals is shortened. As a result, the perceived delay between cause and effect could have been decreased by the presence of the auditory pulse.

A within-subjects ANOVA found significant main effects of both pulse ($F_{(1,31)} = 4.413$) and delay ($F_{(1,31)} = 5.523$), but importantly no significant pulse x delay interaction ($F_{(1,31)} = 0.988$). These results suggest that the effect in Experiment 1 is not attributable to the presence of the auditory pulse alone, and is due to our manipulation of trial structure information. However, one has to be cautious in the interpretation of a null result. While the interaction indeed falls considerably short of significance, one can notice from Figure 3 that the slope from 2s to 5s for the *pulse* condition is less steep than that for the *no pulse* condition. One might therefore suggest that a more powerful experiment may also have elicited a significant interaction.

These slight concerns can be alleviated by an inspection of the behavioral data. The bisection of the trial by the tone had the potential to induce a change in the behavior of participants. Some significance may have been attached to the tone, for instance being perceived as marking the start of the trial, or a point at which they should respond. Participants may therefore only have responded at or after the tone, and by doing so, effectively cutting the trial in half, and significantly reducing the action-outcome delay. This would account for the increase of 5s relative to 2s – if trial length is indeed truncated in this fashion, it will have been shortened by approximately 2.5s compared to 1s.

Behavioral Data

An analysis of the behavioral data reflected these suspicions. While the main effect of pulse on response rate was non-significant ($F_{(1,34)} = 2.760$), there was indeed a significant effect of pulse on action-outcome interval ($F_{(1,34)} = 25.983$). Mean action-outcome intervals where no pulse

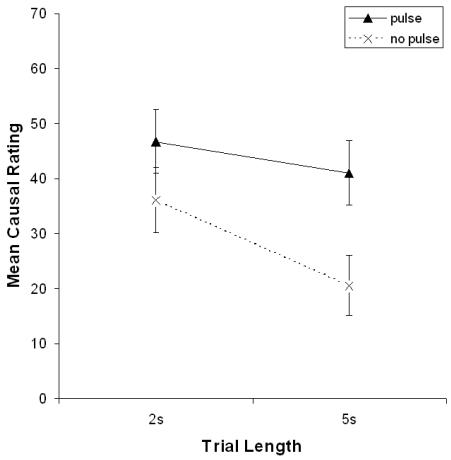


Figure 3: Mean causal ratings for Experiment 2B. Error bars show standard error.

was present were 1.38s and 3.46s at trial lengths of 2s and 5s respectively; with the inclusion of the pulse, these were shortened to 1.11s and 2.78s This can explain both the main effect of pulse, through the overall reduction in delay, and also the smaller decline in ratings from 2s to 5s (when the pulse was present) as there is a smaller discrepancy in delay. Consistent with the previous experiments, we also found the expected main effects of trial length on both response rate ($F_{(1,34)}$ = 13.819) and action-outcome interval ($F_{(1,34)}$ = 546.072). The interaction between pulse and trial length was significant for action-outcome interval ($F_{(1,34)}$ = 5.477) but not for response rate ($F_{(1,34)}$ = 1.054).

We can therefore be confident in our assessment that auditory pulse is not the determinant of the interaction observed in Experiment 1; when trial structure was present, 5s conditions received significantly higher ratings than when it was not, despite a lack of difference in actual action-outcome interval. If this effect were driven by the pulse, then in the present experiment, coupled with the behavioral shift, a significant interaction should have been even more likely, yet was not obtained.

General Discussion

This paper has demonstrated that by providing structural information in a real-time causal judgment task, the detrimental effect of temporal separation between action and outcome can be abolished. When cause and effect pairings are clearly delineated, the learning process appears to reduce to a simple contingency assessment which is unaffected by delay. Two follow-up studies ruled out potential alternative explanations for this effect, thus we can have confidence in the validity of the trial-structure manipulation.

These findings are consistent with a rational perspective on causal induction and could be regarded as a step towards overcoming the problem of circularity that hampers the causal mechanism view. It may well be that in the absence of clear structural information, other sources of knowledge (such as expectations based on previously acquired mechanistic beliefs) serve to divide the event stream into meaningful patterns of event co-occurrence. Importantly, we have shown that such beliefs are not necessary when structural information is apparent in the input, and furthermore, that such information serves to overcome the well-established detrimental effects of reinforcement delay.

References

Buehner, M. J., & May, J. (2002). Knowledge mediates the timeframe of covariation assessment in human causal induction. *Thinking and Reasoning, 8*(4), 269-295.

Buehner, M. J., & May, J. (2003). Rethinking temporal contiguity and the judgement of causality: effects of prior knowledge, experience, and reinforcement procedure. *Quarterly Journal of Experimental Psychology, 56A*(5), 865-890.

Buehner, M. J., & May, J. (2004). Abolishing the Effect of Reinforcement Delay on Human Causal Learning. *Quarterly Journal of Experimental Psychology, 57B*(2), 179-191.

Buehner, M.J. & May, J. (2009). Causal Induction from Continuous Event Streams: Evidence for Delay-Induced Attribution Shifts. *Journal of Problem Solving, 2(2)*.

Einhorn, H. J., & Hogarth, R. M. (1986). Judging probable cause. *Psychological Bulletin, 99*(1), 3-19.

Garcia, J., & Koelling, R. A. (1966). Relation of cue to consequence in avoidance learning. *Psychonomic Science, 4*(3), 123-124.

Reed, P. (1992). Effect of a Signaled Delay Between an Action and Outcome On Human Judgment of Causality. *Quarterly Journal of Experimental Psychology, 44B*(2), 81-100.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical Conditioning II: Current theory and research* (pp. 64-99). New York: Appleton-Century Crofts.

Shanks, D. R., Pearson, S. M., & Dickinson, A. (1989). Temporal Contiguity and the Judgment of Causality By Human Subjects. *Quarterly Journal of Experimental Psychology, 41B*(2), 139-159.

Treisman, M., Faulkner, A., Naish, P. L., & Brogan, D. (1990). The internal clock: evidence for a temporal oscillator underlying time perception with some estimates of its characteristic frequency. *Perception, 19*(6), 705-743.

Wasserman, E. A., Chatlosh, D. L., & Neunaber, D. J. (1983). Perception of Causal Relations in Humans. *Learning and Motivation, 14,* 406-432.

More than One Kind of Probability Matching: Evidence from a Dual-Task Paradigm

A. Ross Otto and Arthur B. Markman
(rotto@mail.utexas.edu, markman@psy.utexas.edu)
Department of Psychology, University of Texas, Austin, TX 78712 USA

Eric G. Taylor (etaylor4@illinois.edu)
Department of Psychology, University of Illinois, Champaign, IL 61820 USA

Abstract

Probability-matching is a well-documented suboptimal behavior that arises in simple prediction tasks. We identify two distinct, local choice strategies that both give rise to probability-matching behavior on a global level. Using a dual-task paradigm, we evaluate the hypothesis that these qualitatively different strategies exhibit different demands on individuals' central executive resources. We find that participants placed under a concurrent working memory are driven away from the one-trial-back strategy—utilized by participants without a working memory load—and towards a strategy that integrates a longer window of past outcomes into the current prediction. In other words, the demands of the concurrent task appeared to shift the prediction strategies used by decision-makers in our study.

Keywords: Decision-making; Prediction; Win-Stay-Lose-Shift; Working Memory; Dual Task; Heuristics

Introduction

One decision-making anomaly of great interest is the tendency for humans to match their responses to outcome probabilities in the prediction of binary outcomes. For example consider a laboratory task in which people need to repeatedly predict which of two outcomes (say Event A and Event B) will occur next. If Event A occurs at a base rate of p = .65, Event B occurs at a base rate of p = .35 and each outcome is conditionally independent of the last outcome, the optimal prediction strategy would be to always predict that Event A will occur next, which is called *maximizing*. However, a large body of empirical work suggests that people appear to predict events in proportion to their frequency of occurrence, known as *probability matching* (Estes, 1961; Vulkan, 2000). Under probability matching, a person would predict Event A 65% of the time and Event B 35% of the time. It is easy to see that this strategy produces an expected overall accuracy of 54.5% (calculated as .65 x .65 + .35 x .35), which is inferior to that produced by *maximizing*—which produces an expected overall prediction accuracy of 65%. In the present study, we examine strategies that be may underlying probability matching in random sequences of events.

The psychological mechanisms that give rise to probability matching behavior are unclear and are a matter of ongoing debate. One hypothesis posits that probability matching arises from the use of a suboptimal cognitive shortcut in which individuals allocates their responses according to an assessment of the observed outcome probabilities (e.g., Koehler & James, 2009). Under this strategy, termed *expectation matching* (EM), the decision-maker's responses are the result of integrating a moving window of past outcome information (Sugrue, Corrado, & Newsome, 2004). To generate a response, the individual stochastically and independently generates predictions in accordance with this historical assessment of outcome probabilities. Assuming a sufficiently long historical window, a decision-maker utilizing the EM strategy in the example above would stochastically allocate 65% of their predictions to Event A and 35% of their predictions to Event B.

Another proposal suggests that probability matching behavior seen at a more global level is the byproduct of a local decision process called *win-stay lose-shift* (WSLS: Herrnstein, Rachlin, & Laibson, 2000). Under WSLS, an individual persists with predicting one event, say Event A, until they make an incorrect prediction, at which point they shift responses and persist with predicting Event B until they are incorrect. While under certain task circumstances WSLS is an optimal choice strategy (Shimp, 1976), it is a suboptimal prediction strategy in the task outlined above. It can be shown that WSLS produces overall response rates (and hence, accuracy rates) equivalent to probability matching (Unturbe & Corominas, 2007). Further, there is evidence that people utilize WSLS in the simple binary prediction task described above (Gaissmaier & Schooler, 2008). Unlike the EM strategy, which involves integrating a comparatively long historical window of outcomes, WSLS requires that the decision-maker maintain a short-term memory for only the most recent response and outcome.

In the present study, we examined the cognitive demands imposed by the WSLS and EM strategies, with the idea that decision makers may utilize both strategies, but under different circumstances. While both strategies result in equivalent behavior at a global level—probability matching—they make different behavioral predictions at a local, trial-by-trial level. It is well documented that the working memory demands of a secondary task deplete mental resources that could otherwise be used to accomplish a primary task (Pashler, 1994). For example, Zeithamova and Maddox (2006) found that working memory load disrupts learning of explicit, rule-based categories and drives participants towards the use of an implicit, information-integration strategy. Here, we place decision-makers under a concurrent working memory load and find that they exhibit the same global tendency to probability match as decision-makers without a working memory

load. Using simple models, we demonstrate that different local strategies result in global probability matching. The distinction between these two matching strategies is theoretically significant because recent contributions to the probability matching literature (e.g., Gaissmaier & Schooler, 2008; Koehler & James, 2009) fail to find common ground on a) which strategies may give rise to probability matching behavior, and b) to what extent these strategies place demands on executive function.

Method

Participants One-hundred and sixty undergraduates at the University of Texas at Austin participated in this study, randomly assigned to one of two conditions: Dual-Task (DT) and Single-Task (ST). Participants were paid a small cash bonus of one cent per correct prediction.

Design and Procedure The experiment stimuli and instructions were displayed on 17-inch monitors. The participants were told that their goal was to predict repeatedly whether a red square would appear above a fixation cross or a green square below the fixation cross, using the up and down arrows respectively (see Figure 1 for a task screenshot). Like other studies (e.g., Koehler & James, 2009), the sequence of events was serially independent. The probability of the more common event was p = .65. The assignment of the high-probability event to the outcomes was counterbalanced across subjects. Subjects completed 10 practice trials in order to familiarize themselves with the response procedure, followed by 320 trials divided into 8 blocks of 40 trials each.

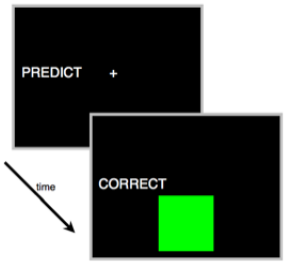


Figure 1: Example task screenshot of response and outcome for a correct prediction.

In order to accommodate the dual-task manipulation, the prediction task used a deadline procedure to ensure that a fixed amount of time elapsed each trial. At the start of each trial, the subject saw the word "PREDICT" and had two seconds to make a response. This response window lasted two seconds regardless of the timing of the response, and was followed by the actual outcome along with feedback indicating whether their prediction was correct ("CORRECT") or incorrect ("INCORRECT"). The outcome and feedback were displayed for one second, and was followed by a one second inter-trial interval. If a subject failed to respond within the response window, the message "TOO SLOW" was displayed along with the outcome. The timing of response windows and outcomes was the same for both the ST and DT conditions.

Blocks in the DT condition consisted of a secondary tone-counting task in addition to the prediction task. The design of the secondary task follows that of Foerde et al. (2007). Two types of tones, high-pitched (1000 Hz) and low-pitched (500 Hz) were played during each trial in the DT condition. Each three-second trial was divided into 12 intervals of 250 ms, with the tones occurring in intervals 3-10 (500-2,500 ms after trial onset). The number of tones presented each trial varied uniformly between 1 and 3 and occurred randomly within intervals 3-10. The pitch of each tone varied randomly, with the base rate of high tones varying uniformly from .3 to .7 each block. The subjects were instructed to maintain a running count of the number of high tones while ignoring the low-pitched tones. Note that the secondary task persisted during both the response window and the outcome. At the end of each 40-trial block, the subjects reported their running count using the keyboard and were instructed to restart their count at zero.

After subjects had completed 320 trials, they completed a questionnaire in which they were asked to provide estimates of the overall frequency of the red and green events. They were also given five prediction strategies to evaluate. These strategies included an expectation matching strategy ("Predict GREEN 65% of the time regardless of what happened during the last outcome"), a maximizing strategy, ("Always predict GREEN, regardless of what happened during the last outcome"), and a WSLS strategy ("Stick with predicting one outcome, and then change your prediction if you were incorrect on the last trial"). Subjects were instructed to rank these five strategies from 1 ("the best possible strategy") to 5 ("the worst possible strategy"), using each ranking only once.

Results

We removed data from 12 ST and 26 DT participants whose prediction behavior differed non-significantly from equiprobable responding (Binomial test at the p= .05 level of significance). We also removed the data of eleven participants who failed to respond before deadline more than 20 times during the experiment. One hundred and eleven participants (48 DT and 63 ST participants) remained in the analysis that follows.

Overall Prediction Performance Figure 2 depicts the subjects' accuracy, by condition, in predicting outcomes over the 320 trials. The dashed line depicts the level of accuracy expected under probability matching probability—that is, if participants allocated their 65% of their responses to the more frequent outcome. A 2 (task condition) x 2 (trial block) ANOVA revealed neither a significant main effect of task condition, $F(1,107) = .55$, $p = .46$, nor a significant interaction between condition and trial block, $F(1,107) = 0.27$, $p = .61$. There was a significant main effect of trial block, $F(1,107) = 25.51$, $p < .001$. Again, the lack of effect of task condition suggests that the dual task manipulation did not hinder subjects' overall accuracy, but rather, may have shifted the

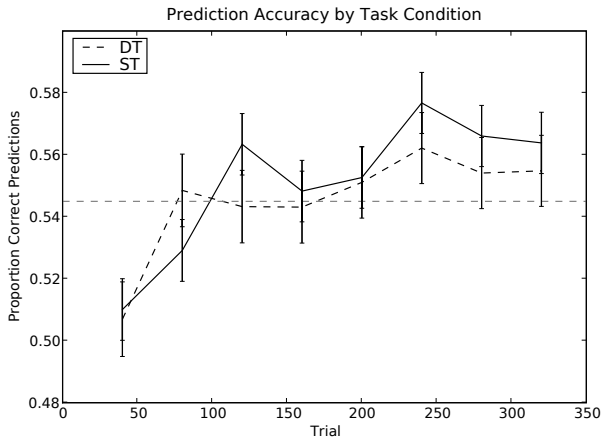


Figure 2: Left panel: mean prediction accuracy, by task condition and trial block. ST = Single-task condition, DT=dual-task condition. Error bars represent standard error of the mean.

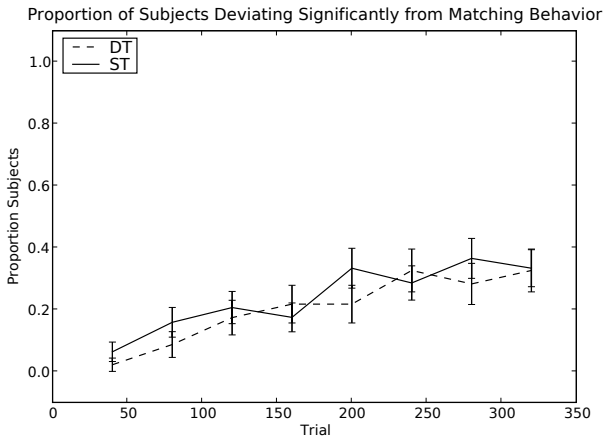


Figure 3: Proportion of Subjects Deviating Significantly from Matching (by Binomial test), by task condition and trial block. ST = Single-task condition, DT=dual-task condition. Error bars represent standard error of proportion.

prediction strategies that subjects employed.

Overall Deviation from Matching Recall that our main goal was to determine whether matching behavior results from different strategies across the ST and DT conditions. Before comparing strategy usage, we first determine that both groups were in fact predominantly matching—and to the same degree. Specifically, we determined whether the secondary task manipulation affected the degree to which subjects deviated significantly from matching behavior (that is, allocating 65% of one's responses to the more frequent event). For each of the 8 blocks, we calculated the proportion of subjects whose response allocations deviated significantly from a response allocation that matched the observed outcome frequency. The proportion of subjects in each condition, by block, that deviated significantly from probability matching behavior (under a Binomial test at the $p = .05$ level significance) are shown in Figure 3. We conducted a logistic regression with each subject's classification (deviating significantly or not) as the criterion and task condition and trial block as predictors, observing no significant coefficients for task condition (*Beta* = -.83, $p = .44$) or the interaction between task condition and trial block (*Beta* = .08, $p = .53$). Trial block did have a significant coefficient (*Beta* = .5, $p < .001$). The apparent null effect of task condition suggests that ST and DT subjects were engaging in prediction behavior that appears similar at a coarse level of analysis.

Exponentially-Weighted Averaging Model Analysis At least two distinct response strategies can manifest themselves as probability matching. Under WSLS, the decision-maker repeats the previous trial's response after a correct prediction and switches their response after an incorrect prediction. Thus responses under WSLS are determined by the outcome on the only the most recent trial. In contrast, EM requires that the decision-maker integrate a much longer window of previous outcomes, which in turn informs the decision-maker's response probabilities. By fitting a simple exponentially-weighted averaging model model to participants' responses, we identified the degree to which participants' predictions were dependent on recent outcomes. The probability P(t) of the decision-maker predicting the green event at time t is determined by:

$$P(t) = recency*outcome(t\text{-}1) + (1\text{-}recency)*P(t\text{-}1),$$

where *outcome(t-1)* is the outcome on the previous trial, *P(t-1)* is the model's estimate of the rate at which the green outcome occurs, and *recency* is a parameter that determines how much recent outcomes are weighted in updating *P(t)*. When the recency parameter is large, *P(t)* is based only on the most recent trial's outcome, and when the recency parameter is small, the model's predicated response on the next trial *P(t)* is based on a long window of previous outcomes. We fit this model to each participants' responses using maximum likelihood estimation, assuming separate parameter values across blocks. As shown in Figure 4, ST participants had larger estimated learning weights than DT participants, indicating that prediction strategies employed by ST participants were influenced more by recent outcomes. A 2 (task condition)x2 (trial block) ANOVA revealed a significant main effect of task condition, $F(1,107) = 4.13$, $p<.05$, a significant main effect of block, $F(1,107) = 21.38$, $p<0.001$, and a significant interaction between condition and trial block, $F(1,107) = 6.34$, $p<.05$. The effect of condition suggests that ST participants exhibited choice behavior characteristic of WSLS—dependence on only the most recent trials—while DT participants used a strategy characteristic of the EM strategy—involving integration of a long window of past outcomes.

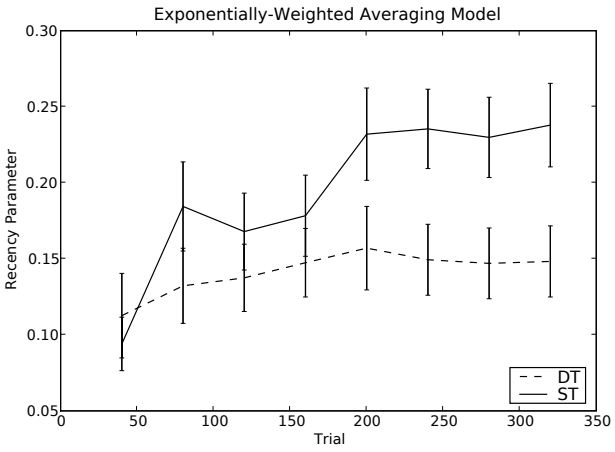


Figure 4: Average best-fitting recency parameter values for exponentially-weighted averaging model, by task condition and block. ST = Single-task condition, DT=dual-task condition. Error bars represent standard error of the mean.

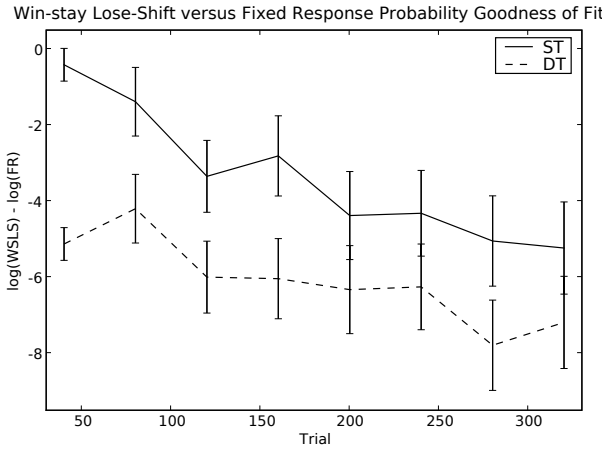


Figure 5: Comparison of model goodness-of-fit between WSLS and EM models. Average likelihood ratios using best-fitting parameter values for each block of each subject. Error bars represent standard error of the mean. ST = Single-task condition, DT=dual-task condition. Error bars represent standard error of the mean.

Models of the Two Prediction Strategies To more directly address usage of these strategies, we compared the relative goodness-of-fit of two models that instantiated the WSLS and EM strategies. To examine participants' WSLS usage, we fit a simple WSLS model to participants' choices, hypothesizing that ST participants would be better fit by this model than DT participants. This one-parameter model constrains the probability of a switching responses after an incorrect prediction (or a "loss") to the probability of persisting with the same response after a correct response (or a "win"). This model follows the WSLS implementation described by Steyvers, Lee, and Wagenmakers (2009). To examine usage of the EM strategy, we fit a simple stochastic response model, which we call the fixed response probability (FR) model, to participants' data. Under this model, a single parameter determines the base rate of predicting the green event. This model—which we use a proxy measure for EM strategy use—assumes that responses are determined stochastically and independently. One crucial difference between these two models is the dependence of the response on trial t to the outcome on trial t-1. We fit both models to each participants' choice data using maximum likelihood estimation allowing parameter values to vary across blocks.

We predicted that ST subjects would be better described by the WSLS model and that DT subjects would be better described by the FR model. Figure 5 depicts the relative goodness-of-fit (expressed as a log-likelihood ratio) between the two models, for each condition across the 8 blocks. Indeed, the likelihood ratios reveal that ST participants were better described by the WSLS model than the responses of DT participants, and conversely, DT participants were better described by the FR model—our proxy for the EM strategy. A 2 (task condition) x 2 (trial block) ANOVA revealed a significant main effect of task condition, $F(1,107) = 5.28$, $p<.05$, a main effect of trial block, $F(1,107) = 19.18$, $p<.001$, and no significant interaction between task condition and trial block, $F(1,107) = 1.14$, $p=.29$. The main effect of task condition suggests that the concurrent working memory load influenced the local prediction strategies utilized by decision-makers.

Offline Reported Event Probabilities We hypothesized that the secondary task would impair DT participants' ability to explicitly encode information about outcome frequencies. To test this, we calculated absolute deviations between participants' offline reported outcome probabilities and true empirical base rates. The average absolute deviations are shown in Figure 6. We found that DT participants' reported outcome probabilities deviated significantly more from observed base rates than DT participants, $t(107) = 2.82$, $p<.01$. Taken in conjunction with the similar overall accuracy profiles of the two groups, this result suggests that the two groups may have been using qualitatively different strategies to make predictions.

Strategy Self-Reports We assessed participants' offline endorsement of the strategies that were described in the questionnaire. To do this, we compared participants' relative preference for the WSLS over EM by their subtracting their ranking of the WSLS strategy from their ranking of the EM strategy, yielding a measure of endorsement of WSLS (note that this measure is equally informative about preference for EM). We found that ST participants' endorsement of WSLS significantly correlated with their overall WSLS model goodness-of-fit, $r(107) = .35$, $p < .01$, suggesting that ST participants had some explicit awareness of the strategies they employed. In contrast, DT participants' strategy endorsements did not significantly correlate with their average goodness-of-fit mea-

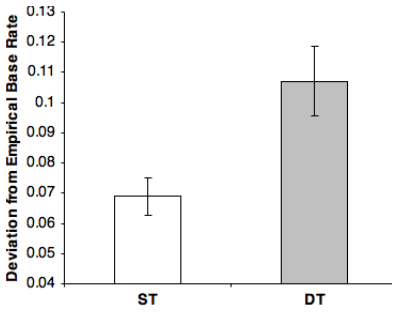


Figure 6: Mean absolute deviation from observed (empirical) base rate, by task condition. Error bars represent standard error of the mean.

sures for either model, suggesting the concurrent working memory load impaired decision-makers' ability to explicitly report the strategies they employed [WSLS model: $r(107) = .15$, $p=.28$, FR model: $r(107) = -.02$, $p = .82$].

Discussion

In this experiment, we investigated the effect of a concurrent working memory task on probability-matching behavior in a random, sequentially independent prediction task. To do so, we imposed a secondary working memory task on subjects, which was believed to deplete working memory resources that could have been used on the primary prediction task (Pashler, 1994). In the DT condition, subjects needed to both make responses in the prediction task and update their count of auditory tones, while in the ST condition, subjects needed only to make predictions. Although most subjects in both conditions demonstrated probability matching, subjects in the ST condition relied more on a WSLS strategy, which requires memory for the previous prediction and outcome. This finding suggests that while both ST and DT subjects appear to be using suboptimal strategies with similar base rates at a molar level, the two groups may actually be using different prediction strategies.

Our results are interesting in the context of previous dual-task studies of human learning. For example, Foerde et al., (2007) found that a concurrent working memory load during probabilistic classification learning impaired subjects' acquisition of explicit associations between perceptual cues and outcomes, although these subjects evidenced implicit learning of cue-outcome contingencies. Further, they were unable to flexibly apply knowledge about cues in an offline evaluation. Zeithamova and Maddox (2006) found that a concurrent working memory load disrupts learning of explicit, rule-based categories and instead drives subjects towards the use of an implicit, information integration strategy. Both of these studies point to the possibility that concurrent working memory load engenders the use of implicit learning systems. In our study, utilization of the EM strategy may be indicative of the operation of an implicit system.

Another possibility raised in the literature is that probability matching arises out of peoples' search for regularities in the event sequences (Gaissmaier & Schooler, 2008). Even when laboratory prediction tasks are probabilistic and outcomes sequences are conditionally independent, people may search for deterministic patterns in an attempt to achieve prediction accuracy above that of maximizing. Thus, if an individual believes that the event sequence contains structure, he or she will try to improve their accuracy by searching for patterns. Gaissmaier & Schooler's result suggests that that some individuals in the present study who appear to be probability matching—rather than maximizing—are more adept at detecting deterministic patterns when they are later introduced into the sequence of events.

One possibility in the present study is that subjects in the ST condition may have begun a search for deterministic patterns and abandoned the search given the very low likelihood of a pattern repeating itself in the random sequence, reverting later to a suboptimal WSLS strategy. Supporting evidence comes from the fact that over 60% of the ST condition's responses are consistent with WSLS and that this percentage increases over time. This hypothesis will be the subject of investigation in future studies.

References

Foerde, K., Poldrack, R. A., & Knowlton, B. J. (2007). Secondary-task effects on classification learning. *Memory & Cognition*, *35*(5), 864–874

Gaissmaier, W., & Schooler, L. J. (2008). The smart potential behind probability matching. *Cognition*, *109*(3), 416–422

Herrnstein, R. J., Rachlin, H., & Laibson, D. I. (2000). *The matching law*. Harvard University Press.

Koehler, D. J., & James, G. (2009). Probability matching in choice under uncertainty: Intuition versus deliberation. *Cognition*, *113*(1), 123–127

Pashler, H. (1994). Dual-task interference in simple tasks: Data and theory. *Psychological Bulletin*, *116*(2), 220–244

Shimp, C. P. (1976). Short-term memory in the pigeon: the previously reinforced response. *Journal of the Experimental Analysis of Behavior*, *26*(3), 487–493

Steyvers, M., Lee, M. D., & Wagenmakers, E. (2009). A bayesian analysis of human decision-making on bandit problems. *Journal of Mathematical Psychology*, *53*(3), 168–179

Sugrue, L. P., Corrado, G. S., & Newsome, W. T. (2004). Matching behavior and the representation of value in the parietal cortex. *Science*, *304*(5678), 1782–1787

Unturbe, J., & Corominas, J. (2007). Probability matching involves rule-generating ability: A neuropsychological mechanism dealing with probabilities. *Neuropsychology*, *21*(5), 621–630

Vulkan, N. (2000). An economist's perspective on probability matching. *Journal of Economic Surveys*, *14*(1), 101–118

Zeithamova, D., & Maddox, W. T. (2006). Dual-task interference in perceptual category learning. *Memory & Cognition*, *34*(2), 387–398

Probabilistic language acquisition: Theoretical, computational, and experimental analysis

Anne S. Hsu (ahsu@gatsby.ucl.ac.uk)
Division of Psychology and Language Sciences, 26 Bedford Way
London, WC1H 0AP

Nick Chater (n.chater@ucl.ac.uk)
Division of Psychology and Language Sciences, 26 Bedford Way
and Centre for Economic Learning and Social Evolution (ELSE)
London, WC1H 0AP

Abstract

There is much debate over the degree to which language learning is governed by innate language-specific biases, or acquired through cognition-general principles. Here we examine the probabilistic language acquisition hypothesis on three levels: We outline a theoretical result showing that probabilistic learning in the limit is possible for a very general class of languages. We then describe a practical computational framework, which can be used to quantify natural language learnability of a wide variety of linguistic constructions. Finally, we present an experiment which tests the learnability predictions for a variety of linguistic constructions, for which learnability has been much debated. We find that our results support the possibility that these linguistic constructions are acquired probabilistically from cognition-general principles.

Keywords: child language acquisition: Gold's theorem; poverty of the stimulus; probabilistic learning; simplicity principle; adult grammar judgments; natural language

Introduction

A central debate in cognitive science revolves around how children acquire their first language. A significant portion of this debate centers on how children learn complex linguistic structures, such as restrictions to general rules. An example restriction-rule can be seen in the contraction of '*going to*': '*I'm gonna leave'* is grammatical whereas '*I'm gonna the store'* is ungrammatical. Language communication requires the speaker to generalize from previously heard input. However, research shows children rarely receive feedback when they produce an over-general, ungrammatical sentence. Children also aren't explicitly told which generalizations are allowed and which are not (Bowerman, 1988). These observations evoke the question: how do children learn that certain overgeneralizations are ungrammatical without explicitly being told?

Traditionally, linguists have claimed that such learning is impossible without the aid of innate language-specific knowledge (Chomsky, 1975; Crain, 1991; Pinker, 1989; Theakston, 2004). However, recently, researchers have shown that statistical models are capable of learning restrictions to general rules from positive evidence only (Dowman, 2007; Foraker, Regier, Khetarpal, Perfors, & Tenenbaum, 2009; Grünwald, 1994; Perfors, Regier, & Tenenbaum, 2006; Regier & Gahl, 2004).

Here we examine language acquisition from a probabilistic perspective on a theoretical, computational and experimental level. We first revisit Gold's theorem and show that language identification *is* possible from a probabilistic perspective. Next we mention a recently proposed, general framework which can quantify learnability of constructions in natural language. This flexible framework allows for predictions to be made concerning the natural language learnability of a wide variety of linguistic rules. Finally, we experimentally test the learnability predictions obtained from this framework by comparing these predictions with adult grammaticality judgments for a wide range of linguistic constructions.

Gold revisited: probabilistic language acquisition with a simplicity prior

Inherent in a simplicity-based approach to language acquisition is the trade-off between simpler vs. more complex grammars: Simpler, over-general grammars are easier to learn. However, because they are less accurate descriptions of actual language statistics, they result in inefficient encoding of language input, i.e. the language is represented using longer code lengths. More complex grammars (which enumerate linguistic restrictions) are more difficult to learn, but they better describe the language and result in a more efficient encoding of the language, i.e., language can be represented using shorter code lengths. Under simplicity models, language learning can be viewed in analogy to investments in energy-efficient, money-saving appliances. By investing in a more complicated grammar, e.g. one which contains a restriction on a construction, the language speaker obtains encoding savings every time the construction occurs. This is analogous to investing in an expensive but efficient appliance that saves money with each use. A linguistic restriction is learned when the relevant linguistic context occurs often enough that the accumulated savings makes the more complicated grammar worthwhile. Because complex grammars become worth while as linguistic constructions appear more often,

simplicity models are able to learn restrictions based on positive evidence alone (See Figure 1).

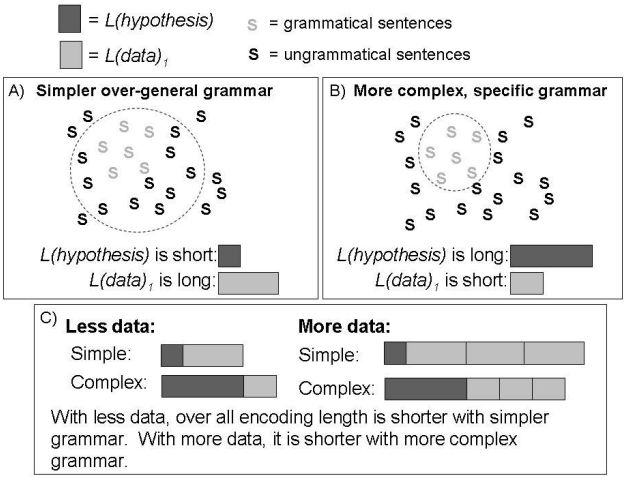


Figure 1: MDL simple grammar vs. efficient language encoding trade off. A) A simpler grammar is often over-general, i.e., allows for ungrammatical sentences as well as grammatical ones. Such an over-general grammar may be easy to describe (i.e., short grammar encoding length), but results in less efficient (longer) encoding of the language data. B) A more complex grammar may capture the language more accurately, i.e., allows only for grammatical sentences and doesn't allow for ungrammatical sentences. This more complex grammar may be more difficult to describe (i.e., longer grammar encoding length), but will provide a shorter encoding of language data. C) Initially, with limited language data, the shorter grammar yields a shorter coding length over-all, and is preferred under MDL. However, with more language input data, the savings accumulated from having a more efficient encoding of language data correctly favour the more complex grammar.

A central theoretical question is: given sufficient exposure to the language, can the learner recover a perfectly accurate description of that language? Gold (1967) famously showed that, under certain assumptions, this is not possible. However, a range of more positive results have since been derived, e.g., (J. A. Feldman et al 1969; Chater & Vitányi 2007). Here we show that under a simplicity-based probabilistic formulation, a new and strong positive result can be derived.

Suppose that the learner encounters sentences, s, which are independently sampled generated from a computable probability distribution, $C_P(s)$, which has Kolmogorov complexity $K(C_P)$. Here we will define learning a language as the process of identifying this distribution. $C_P(s)$ generates a corpus $S_n = s_1, s_2,...s_n,...$ which continues indefinitely. We assume that $C_P(s)$ allows all and only grammatical sentences in language L. That is, the probability of generating all sentences s, that are grammatical in L, is greater than zero, $C_P(s) > 0$; and conversely, if the probability of a sentence being generated is greater than zero, then it is grammatical according to L.

There is one additional mild constraint that we need to impose: that $C_P(s)$ has a finite entropy, i.e.,

$$H(C_p) \propto \sum_{j=1}^{\infty} C_p(s_j) \log\left(\frac{1}{C_p(s_j)}\right) < \infty$$

This is a modest constraint, because it follows from the assumption that the mean sentence length under distribution $C_P(s)$ is finite, which is clearly true for natural language.

The learning problem proceeds as follows: A learner is given an initial sample of the corpus S_n. The question then is: how should the learner assign probabilities to the various possible computable distributions C_Q that might have generated the corpus? This is equivalent to learning: $Pr(C_Q|S_n) \propto Pr(S_n|C_Q)Pr(C_Q)$

Also, we ask how these probabilities change as the corpus grows arbitrarily long, i.e., as n tends to infinity? In particular, can the learner identify the true probability distribution, C_P, in the limit?

Intriguingly, it turns out that this is possible – and indeed that an ideal learner (Chater & Vitányi 2007)) will 'converge' on the true probability distribution, C_P, with probability of measure 1, given a sufficiently large corpus. Suppose, for concreteness, that the learner "announces" its current most probable generating distribution each time a new sentence i arrives, based on the i sentences that he has received so far $S_i = \{s_1, s_2,..., s_i\}$. More formally, the following theorem holds: Consider any computable probability distribution C_P, from which samples, s_i, are drawn independently to generate a semi-infinite corpus S. Let m' be the number of initial items of S so that $S_{m'}$ is a "prefix" of S (i.e., a corpus consisting of the first m' items of s). With probability greater than $1-\varepsilon$, for any $\varepsilon > 0$, there is an m such that, under the simplicity principle, for all $m' \geq m$, the most probable C_Q, given $S_{m'}$ is the generating distribution C_P, i.e., $\text{argmax}(Pr(C_Q|S_{m'}))=C_P$.

Why is this true? A full proof is beyond the scope of this paper (see Chater & Hsu, in preparation); but the essence of the argument is the following. We know that almost all random samples from P will be incompressible (i.e., n sentences generated by the true generative model P will have no shorter description than the entropy $nH(P)$). This implies that, for *typical* data generated by P (which have summed probability arbitrarily close to 1), $K(P)+nH(P) \geq K(S_n) \geq nH(P)$. Now for each S_n, consider the set of probability distributions Q which satisfy this criterion: $K(Q)+nH(Q) \geq K(S_n) \geq nH(Q)$. For each n, there will be finitely many such Q; and, by our argument above, these will include the true distribution P. Now, for each n, the learner "announces" the simplest Q', i.e., the Q' such that for all Q, $K(Q) \geq K(Q')$. We know that P will always be in this set, by the argument above. However, there are only finitely many Q that are simpler than P. Once these simpler Q have been eliminated, then P will be the shortest element in the set, and will be announced indefinitely thereafter. We know that each of this finite set will be eliminated for

sufficiently large n, because the expected excess cost of encoding data generated by P with distribution Q is $n\mathrm{D}(Q||P)$, where $\mathrm{D}(Q||P) > 0$ unless $Q=P$; this excess cost tends to infinity as n tends to infinity. Hence, for some $n' > n$, after all probability distributions Q with shorter codes than P have been eliminated, P will be announced indefinitely.

Practical framework for quantifying learnability

The positive learnability results indicate that the probabilistic approach can be practically applied to the problem of language acquisition. Recently, researchers have used probabilistic models to show that many complex linguistic rules can be acquired by directly learning the probability distribution of grammatical sentence structures in language. These models learn this probability distribution under a cognition general prior for simplicity (Dowman, 2007; Foraker et al., 2009; Grünwald, 1994; Perfors et al., 2006; Regier & Gahl, 2004). Many of these studies used restricted language sets. In the context of natural language, a few studies have addressed specific linguistic cases such as anaphoric one (Foraker et al., 2009) and hierarchical phrase structure (Perfors et al., 2006).

Recently, a *general quantitative framework* has been proposed which can be used to assess the learnability of any given *specific linguistic restriction* in the context of real language, using positive evidence and language statistics alone (Hsu & Chater, 2010). This framework built upon previous probabilistic modeling approaches to develop a method that is generally applicable to any given construction in natural language. This new tool can be used to explicitly explore the learnability in a corpus relative to well-known information theoretic principles given a grammatical description. When using this framework to analyze learnability of a linguistic construction, there are two main assumptions: 1) The description of the grammatical rule for the construction to be learned. 2) The choice of corpus which approximates the learner's input. Given these two assumptions, the framework provides a method for evaluating whether a construction is present with adequate frequency to make it learnable from language statistics. The framework allows for comparison of different learnability results which arise from varying these two main assumptions. By making these assumptions explicit, a common forum is provided for quantifying and discussing language learnability.

Minimum Description Length hypothesis

Because this framework is detailed elsewhere (Hsu & Chater 2010), we will only provide a brief overview here. Learnability evaluations under a simplicity prior can be instantiated through the principle of minimum description length (MDL). MDL is a computational tool that can be used to quantify the information available in the input to an idealized statistical learner of language as well as of general cognitive domains (Jacob Feldman, 2000). When MDL is applied to language, grammars can be represented as a set of rules, such as that of a probabilistic context free grammar (PCFG) (Grünwald, 1994). An information-theoretic cost can then be assigned to encoding the grammar rules as well as to encoding the language under those rules.

Hsu & Chater (2010) used an instantiation known as 2-part MDL, which we will refer to as just MDL for brevity. In the context of language acquisition, the first part of MDL uses probabilistic grammatical rules to define a probability distribution over linguistic constructions, which combine to form sentences. Note that these probabilities are not necessarily the real probabilities of sentences in language, but the probabilities as specified under the current hypothesized grammar. The second part of MDL consists of the encoded representation of all the sentences that a child has heard so far. MDL selects the grammar that minimizes the *total* encoding length (measured in bits) of both the grammatical description and the encoded language length[1].

According to information theory, the most efficient encoding occurs when each data element is assigned a code of length equal to the smallest integer greater than or equal to $-\log_2(p_n)$ bits, where p_n is the probability of the nth element in the data. For our purposes, these elements are different grammar rules. The probabilities of these grammar rules are defined by the grammatical description in the first part of MDL. Because efficient encoding results from knowing the correct probabilities of occurrence, the more accurately the probabilities defined in the grammar match the actual probabilities in language, the more efficient this grammar will be.

Under MDL, the grammatical description is updated to be the most efficient one each time more data input is obtained. Savings occur because certain grammatical descriptions result in a more efficient (shorter) encoding of the language data. In general, more complex (i.e., more expensive) grammatical descriptions allow for more efficient encoding of the language data. Because savings accumulate as constructions appear more often, more complex grammars are learned (i.e., become worth investing in) when constructions occur often enough to accumulate a sufficient amount of savings. If there is little language data (i.e., a person has not been exposed to much language) a more efficient encoding of the language does not produce a big increase in savings. Thus, when there is less language data, it is better to make a cheaper investment in a simpler grammar as there is not as much savings to be made. When there is more language data, investment in a more costly, complicated grammar becomes worthwhile. This characteristic of MDL learning can explain the early overgeneralizations followed by retreat to the correct

[1] The MDL framework can also be expressed as a corresponding Bayesian model with a particular prior (Chater, 1996; MacKay, 2003; Vitányi & Li, 2000). Here, code length of the model (i.e., grammar) and code length of data under the model (i.e., the encoded language) in MDL correspond to prior probabilities and likelihood terms respectively in the Bayesian framework.

Table 1: Grammatical and ungrammatical sentences used in experiment.

Construction	*Grammatical usage*	*Ungrammatical usage*
is	She's as tall as he is.	She is as tall as he's.
arrive	The train arrived.	He arrived the train.
come	The train came.	I came the train.
donate	He donated some money to the charity.	He donated the charity some money.
fall	The ornament fell.	He fell the ornament.
disappear	The rabbit disappeared.	He disappeared the rabbit.
what is	What's it for?	What's it?
shout	I shouted the news to her.	I shouted her the news.
pour	I poured the pebbles into the tank.	I poured the tank with pebbles.
vanish	The rabbit vanished.	He vanished the rabbit.
whisper	I whispered the secret to her.	I whispered her the secret.
create	I created a sculpture for her.	I created her a sculpture.
who is	Who's it for?	Who's it?
going to	I'm gonna faint.	I'm gonna the store.
suggest	I suggested the idea to her.	I suggested her the idea.
that	Who do you think that she called?	Who do you think that called her?
want to	Which team do you wanna beat?	Which team do you wanna win?

grammar that has been observed in children's speech (Bowerman, 1988). The output of the framework described in Hsu & Chater (2010) results in an estimated number of occurrences needed for a specific linguistic rule to be learned and corpus analysis is then used to assess how many years on average are needed for the sufficient number of occurrences. The general applicability of this framework and its ability to produce clear learnability predictions allow us to take the crucial next step in addressing the language acquisition problem: experimentally assessing whether language might actually be probabilistically acquired.

Testing learnability predictions

Hsu & Chater (2010) used the above framework to assess language learnability of constructions, whose learnability have been commonly debated. These all involve restrictions on a general linguistic rule, which was described using PCFG's. Predictions for learnability in terms of years needed was made for constructions whose learnability have been commonly debated in the language acquisition field. These included restrictions on the following 17 constructions[2]: contractions of *want to, going to, is, what is and who is*; the optionality of *that* reduction; dative alternation for the verbs *donate, whisper, shout, suggest, create, pour*; transitivity for the verbs, *disappear, vanish, arrive, come, fall.* See Hsu & Chater (2010) for the explicit grammar descriptions of linguistic rules to be learned. The results showed a large spread in learnability. Some constructions appeared readily learnable within just a few years whereas other constructions required years that far outnumbered human life spans. Hsu & Chater (2010) compared predicted MDL learnablity with child grammar judgments of constructions for which there was data collected from previous experimental work (Ambridge, Pine, Rowland, & Young, 2008; Theakston, 2004). It was found that child grammar judgments for the constructions were more correlated with learnability than frequency counts (the entrenchment hypothesis (Theakston, 2004)).

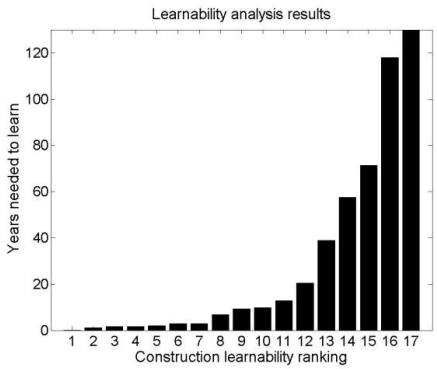


Figure 2: Estimated years required to learn construction. The constructions are sorted according to learnability: 1) *is* 2) *arrive* 3) *come* 4) *donate* 5) *fall* 6) *disappear* 7) *what is* 8) *shout* 9) *pour* 10) *vanish* 11) *whisper* 12) *create* 13) *who is* 14) *going to* 15) *suggest* 16) *that* 17) **want to.* *Predicted years for learning *want to* is 3,800years.

Here we propose that construction learnability should also correlate with adult grammaticality judgments: The more difficult a construction is to learn, the greater the difference

[2] Hsu & Chater (2010) also included analysis of two more linguistic rules concerning the necessary transitivity of the verbs *hit* and *strike*. Though these verbs are traditionally known to be transitive, in colloquial speech they have evolved to have a ambitransitive usage: e.g. *The storm hit. Lightening struck.* In COCA there are 3678 and 1961 intransitive occurrences of *hit* and *strike* respectively. Thus we did not assess rules regarding the intransitivity of these verbs in our experiment.

should be between judgments of the ungrammatical vs. grammatical uses of the construction.

Model Predictions

We conducted our learnability analysis using the full Corpus of Contemporary American English (COCA), which contains 385 million words (90% written, 10% spoken). We believe this is a reasonable representation of the distributional language information that native English language speakers receive. Learnability results using the British National Corpus were similar to that from COCA (Hsu & Chater, 2010). Figure 2 shows the estimated number years required to learn the 17 constructions. We quantified learnability as $log(1/N_{years})$, where N_{years} was the number of estimated years needed to learn a construction (Hsu & Chater, 2010).

Learnability vs. entrenchment To verify that our experimental results are not also trivially explained by a simpler hypothesis, we will also compare experimental results with the predictions of entrenchment theory. Entrenchment is the hypothesis that the likelihood of a child over-generalizing a construction is related to the construction's input occurrence frequency. There is some relation between learnability and entrenchment predictions because high construction occurrence frequencies do aid learnability. However, learnability differs from mere frequency counts because MDL also takes into account the complexity of the grammatical rule that governs the construction to be learned. Additionally, learnability is influenced by whether the restricted form would be commonly or uncommonly expected, if it were grammatically allowed. Here, we propose that under entrenchment hypothesis, the relative grammar judgment difference should be related to the construction's input occurrence frequency. (Frequencies estimated from COCA).

Experimental method

Participants 105 participants were recruited for an online grammar judgment study (age range: 16-75 years, mean=34 years). Results were included in the analysis only for participants who answered that they were native English speakers (97 out of 105 participants). The majority (74%) of our participants learned English in the United States. Other countries included the UK (14%), Canada (5%), Australia (4%). The rest learned English in either Ireland or New Zealand.

Procedure Participants were asked to rate the grammaticality of grammatical and ungrammatical sentences using the 17 constructions whose learnability were quantified above. These sentences (34 total) are shown in Table 1. Grammar judgments ranged from 1-5: 1) Sounds completely fine (Definitely grammatical) 2) Probably grammatical (Sounds mostly fine) 3) Sounds barely passable (Neutral) 4) Sounds kind of odd (probably ungrammatical) 5) Sounds extremely odd (Definitely ungrammatical).

Results

Results show a strong correlation between averaged relative grammaticality vs. log learnability as predicted by MDL, r=.35; p=.0045 (see Figure 3). Relative grammaticality for a given linguistic construction is the grammatical rating for the ungrammatical sentence subtracted by the rating for the grammatical sentence. Note that 4 is the maximum possible relative grammaticality because the lowest ungrammatical rating is 5 and the highest grammatical rating is 1. In contrast, there is no correlation between relative grammaticality and construction occurrence frequency, as would be predicted by entrenchment (see Figure 4).

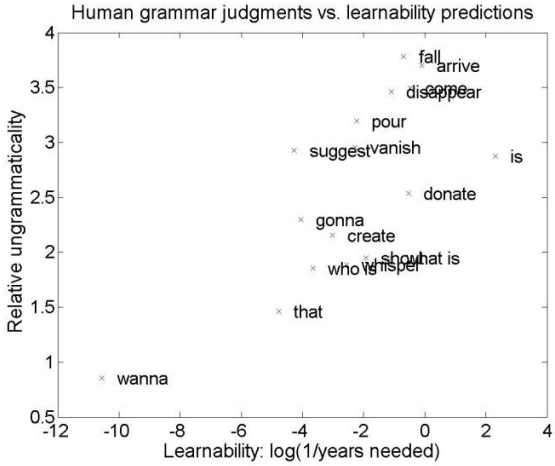


Figure 3: Human grammar judgments vs. learnability analysis. Learnability is log of the inverse of the number of estimated years needed to learn the construction. Correlation values: r=.35; p=.0045

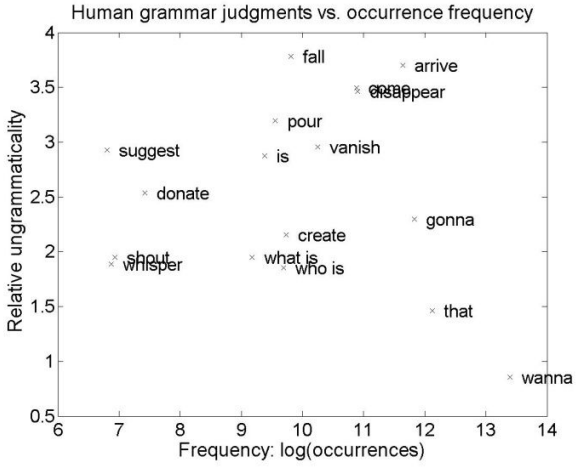


Figure 4: Human grammar judgments vs. log of occurrence frequency. Frequencies were estimated using Corpus of Contemporary American English.

Summary and Conclusions

This presented work helps evaluate how much of first language is probabilistically acquired from exposure. We show that, despite Gold's theorem, language is identifiable with a cognition general prior of simplicity under fairly general assumptions. We then describe a recently formulated framework which allows probabilistic learnability to be quantified in the context of natural language. This framework makes concrete predictions in terms of years needed to learn particular linguistic rules, given an assumed formulation of the rules to be learned and the corpus which represents a learner's language input.

There has now been a substantial body of work showing that probabilistic language learning *is theoretically and computationally possible.* The important next step in research on language acquisition is to assess whether probabilistic learning actually occurs in practice. Here we make the supposition that if language is probabilistically acquired, then there should be evidence of this in adult grammar judgments. There is a subtle leap of logic in this supposition. MDL learnability assumes that a grammar is learned in an absolute sense: once a grammar is chosen under MDL, that is the one used and there is no gradation of knowledge. However, here we are conjecturing that learnability should not only correlate with how long it takes for linguistic rule to be acquired, but also with how certain is one's knowledge of that rule. The more certain one is of a grammatical rule, the greater the difference should be one's acceptability rating of the ungrammatical form relative to the grammatical form. Experimental results show that predicted learnability correlates well with relative grammar judgments for the 17 constructions analyzed, chosen as controversial cases from the literature. Our experimental results support the possibility that many linguistic constructions that have been argued to be innately acquired may instead be acquired by probabilistic learning.

Our learnability predictions were calculated using a large corpus (COCA) to represent the distributional language input that native English speakers receive. This assumes that the distributional information estimated from this corpus is representative of that which influenced the language acquisition process in our adult participants. It also allows for the possibility that a speaker's certainty about different linguistic rules is updated through adulthood using probabilistic learning. If so, older adults might more certain in their grammar judgments, is a direction for future work.

References

Ambridge, B., Pine, J., Rowland, C., & Young, C. (2008). The effect of verb semantic class and verb frequency (entrenchment) on children's and adults' graded judgements of argument-structure overgeneralization errors. *Cognition, 106,* 87-129.

Bowerman, M. (1988). The 'No Negative Evidence' Problem: How do Children avoid constructing an overly general grammar? In J.Hawkins (Ed.), *Explaining Language Universals* (pp. 73-101). Oxford: Blackwell.

Chater, N. (1996). Reconciling simplicity and likelihood principles in perceptual organization. *Psychological Review, 103,* 566-581.

Chater, N. & Hsu, A. (in preparation). Language learning in the limit: theory and practice.

Chater, N. & Vitányi, P.M.B. (2007). `Ideal learning' of natural language: Positive results about learning from positive evidence, *Journal of Mathematical Psychology*, *51*,135-163.

Chomsky, N. (1975/1955). *The Logical Structure of Linguistic Theory*. London: Plenum Press.

Crain, S. (1991). Language Acquisition in the Absence of Experience. *Behavioral and Brain Sciences, 14,* 597-612.

Dowman, M. (2007). Minimum Description Length as a Solution to the Problem of Generalization in Syntactic Theory. *Machine Learning and Language, (in review)*.

Feldman, Jacob (2000). Minimization of boolean complexity in human concept learning. *Nature, 403,* 630-633.

Feldman, J.A., Gips, J., Horning, J. J., & Reder, S. (1969) Grammatical complexity and inference. Technical Report CS 125, Stanford University.

Foraker, S., Regier, T., Khetarpal, N., Perfors, A., & Tenenbaum, J. B. (2009). Indirect Evidence and the Poverty of the Stimulus: The Case of Anaphoric One. *Cognitive Science, 33,* 300.

Gold, E. M. (1967). Language identification in the limit. *Information and Control*, 10, 447-474.

Grünwald, P. (1994). A minimum description length approach to grammar inference. In S.Scheler, Wernter, & E. Rilof (Eds.), *Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language.* (pp. 203-216). Berlin: Springer Verlag.

Hsu, A. & Chater, N. (2010). The logical problem of language acquisition: A probabilistic perspective. *Cognitive Science, 2nd revision submitted*.

MacKay, D. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge: Cambridge University Press.

Perfors, A., Regier, T., & Tenenbaum, J. B. (2006). Poverty of the Stimulus? A rational approach. *Proceedings of the Twenty-Eighth Annual Conference of the Cognitive Science Society,* 663-668.

Pinker, S. (1989). *Learnability and Cognition: The acquisition of argument structure*. Cambridge, MA: MIT Press.

Regier, T. & Gahl, S. (2004). Learning the unlearnable: The role of missing evidence. *Cognition, 93,* 147-155.

Theakston, A. (2004). The role of entrenchment in children's and adults' performance on grammaticality judgment tasks. *Cognitive Development, 19,* 15-34.

Vitányi, P. & Li, M. (2000). Minimum Description Length Induction, Bayesianism, and Kolmogorov Complexity. *IEEE Transactions on Information Theory, IT, 46,* 446-464.

Learning concepts from sketches via analogical generalization and near-misses

Matthew D. McLure (mclure@u.northwestern.edu)
Scott E. Friedman (friedman@northwestern.edu)
Kenneth D. Forbus (forbus@northwestern.edu)
Qualitative Reasoning Group, Northwestern University, 2133 Sheridan Rd
Evanston, IL 60208 USA

Abstract

Modeling how concepts are learned from experience is an important challenge for cognitive science. In cognitive psychology, progressive alignment, i.e., comparing highly similar examples, has been shown to lead to rapid learning. In AI, providing very similar negative examples (*near-misses*) has been proposed as another way to accelerate learning. This paper describes a model of concept learning that combines these two ideas, using sketched input to automatically encode data and reduce tailorability. SAGE, which models analogical generalization, is used to implement progressive alignment. Near-miss analysis is modeled by using the Structure Mapping Engine to hypothesize classification criteria based on differences. This is performed both on labeled negative examples provided as input, and by using analogical retrieval to find near-miss examples when positive examples are provided. We use a corpus of sketches to show that the model can learn concepts based on sketches and that incorporating near-miss analysis improves learning.

Keywords: Concept learning; analogy; generalization.

Introduction

How concepts are learned from experience is a central question in cognitive science. It is well-known that some concepts can be viewed as analytic, having compact necessary and sufficient defining criteria (e.g., *grandparent* or *triangle)*, whereas others are based on similarity or typicality (e.g., *chair, bachelor*). Prior work has explored analogical generalization as an explanation for learning similarity-based categories. The SAGE model of analogical generalization, an evolutionary improvement over SEQL (Kuehne *et al* 2000a) has been used to model learning of perceptual stimuli (Kuehne *et al* 2000b), stories (Kuehne *et al* 2000a), spatial prepositions (Lockwood *et al* 2008) and causal models (Friedman & Forbus, 2008; Friedman & Forbus, 2009). SAGE's ability to construct probabilistic generalizations provides a model of typicality, i.e., high-probability relationships and attributes are more typical. SAGE has been used to model *progressive alignment* (Gentner *et al* 2007), where sequences of highly similar exemplars lead to more rapid learning (Kuehne *et al* 2000a). Progressive alignment alone may suffice to generate rule-like concepts (e.g., Gentner & Medina, 1998), but another possibility is to use negative examples to sharpen criteria for concepts. Winston (1970) proposed the idea of a *near-miss*, a labeled negative example that differs from the intended concept in only one way. A near miss exemplar should be highly alignable with some instances of a concept[1].

This paper describes a model of concept learning that combines analogical generalization and near-miss analysis to capture both similarity-based and analytic aspects of concepts. Its inputs are labeled positive or negative examples of concepts. It uses SAGE to construct generalizations for each concept, thus capturing similarity-based aspects of concepts (and typicality, via probability). When a positive example is provided, the corresponding concept is updated. When a negative example is provided, analogical retrieval is used to find the closest prior positive example or generalization, and analogical matching is used to construct and update hypotheses about inclusion and

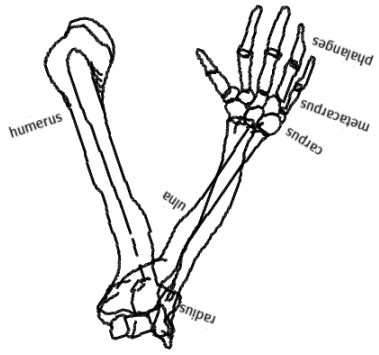


Figure 1: An example of the *skeletal arm* concept drawn in CogSketch.

exclusion criteria for that concept. Near-miss analysis is also attempted when a positive example is provided, using analogical retrieval over negative examples to look for a candidate near-miss. (Using analogical retrieval to find positive concepts and near-misses is a significant advance over Winston's model, which used hand-coded representations, a single abstract description for concepts and required a teacher to supply all negative examples.) To test the model, we use sketches to describe concepts, which are automatically encoded by a sketch understanding system. We show that the model can indeed learn concepts from sketches, and that including near-miss analysis improves learning. Our simulation is implemented using the Companions cognitive architecture (Forbus et al, 2009), which integrates analogical processing and sketching.

[1] For disjunctive concepts, some exemplars will not be similar.

The next section summarizes the simulations of analogical processing and sketch understanding that our model is built upon. We describe our model next, followed by a description of our experiments. We close with related and future work.

Simulation Components

Analogical processing

Our system uses three cognitive models as components to learn concepts and categorize examples. Similarity-based retrieval is used to find similar examples across conceptual boundaries. Analogical comparison is used to compare examples and generate classification hypotheses. Finally, analogical generalization is used to generalize examples. We use the Structure Mapping Engine (SME) (Falkenhainer et al, 1989) to model analogical matching, MAC/FAC (Forbus et al, 1995) to model retrieval, and SAGE (Keuhne et al, 2000) to model analogical generalization.

SME is based on Gentner's (1983) structure-mapping theory of analogy. Given two relational representations, a base and a target, SME computes *mappings* which represent how they can be aligned. A mapping consists of correspondences which describe "what goes with what" in the two representations and a numerical score indicating their degree of similarity. SME also computes *candidate inferences* from the base to the target and from the target to the base. Candidate inferences suggest possible relations that can be transferred across representations, using the correspondences in the mapping as support.

Given a probe case and case library, MAC/FAC efficiently retrieves a case from the case library that is similar to the probe. For scalability, its first stage estimates similarity via dot products on vectors automatically produced from the structured, relational representations used as cases. At most three descriptions are passed to the second stage, which uses SME to compare their full relational versions to the probe, in parallel, to find the best case, or up to three cases if they are very close to the best.

Our model uses SAGE for generalization. Each concept has its own *generalization context*, which SAGE uses to maintain a list of generalizations and ungeneralized examples. Given a new example, it is first compared against each generalization in the context, using SME. If the SME similarity score is over the *assimilation threshold*, the example is merged to update the generalization. Otherwise, the new example is compared with the ungeneralized examples in the context. Again, if the score is over threshold, the two examples are then combined to form a new generalization in the context. Otherwise, the example is added to the context's list of ungeneralized examples. Figure 2 depicts generalization contexts for concepts *Arch* and *Triangle*.

CogSketch

CogSketch[2] (Forbus et al, 2008) is an open-domain sketch understanding system. The ink that a user draws to represent an entity is called a *glyph*, which can be labeled with concepts from an OpenCyc[3]-derived knowledge base. For example, in the sketch shown in Figure 1, each bone is labeled a `Bone-BodyPart`, which is stored as an attribute for each of the individual entities.

CogSketch automatically computes qualitative spatial relations (e.g., `above`, `rightOf`, `touchesDirectly`) between glyphs. In the knowledge representation that is produced by CogSketch, these relations are automatically applied to the entities that the glyphs represent. CogSketch also computes candidate *visual/conceptual relations* (again, from the OpenCyc-derived knowledge base) for pairs of sketched entities based on the visual relationships that hold between them the conceptual labels they have been assigned, and the genre and pose of the sketch. For example, the fact that the glyphs depicting the carpus and metacarpus in Figure 1 touch suggests that the objects they depict might be touching or connected in some way. The list of candidate visual/conceptual relations for these objects is further constrained by the `Bone-BodyPart` concept labels they have been assigned, as well as the `Physical`

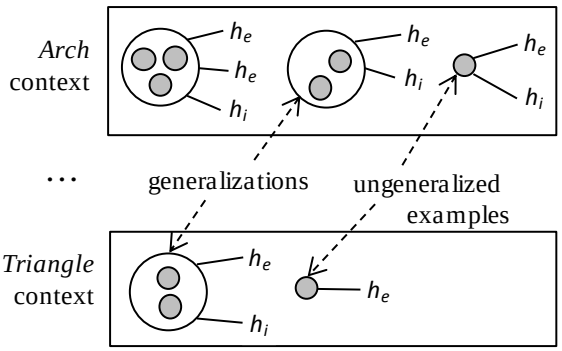


Figure 2: SAGE generalization contexts for *Arch* and *Triangle* concepts, with associated inclusion and exclusion hypotheses (h_i and h_e, respectively).

genre and `from-side` pose of the sketch. The user can browse the candidate relationships and select those which are accurate. In our input stimuli, correct visual/conceptual relationship candidates were always included.

CogSketch is based on the observation that people talk when they sketch, providing verbal labels for what they are drawing, and using language to express functional relationships (e.g. that two parts can rotate, or that one supports another) that the sketch alone cannot convey. The conceptual labels described above, which are applied by a simple menu system, model the effect of verbal labeling. The possible visual/conceptual relationships described above, which are computed automatically and are available

[2] http://www.qrg.northwestern.edu/software/cogsketch/
[3] www.opencyc.org

for the user to choose or not, model the effect of providing functional information via language. This makes the input process much closer to what happens in human-to-human sketching. The user draws ink, which CogSketch's visual system analyzes, producing visual and spatial relationships. The user-supplied conceptual labels plus the visual/spatial analysis enables CogSketch to automatically compute visual/conceptual relationship candidates, from which the user can select, if they choose. (In the experiments reported here, correct visual/conceptual relationships were always chosen, thereby providing some functional information about the concept.)

Similarity & near-miss concept learning

Our model takes as input a stream of labeled sketches. There are two kinds of labels: A positive label indicates that the example is an instance of a concept, e.g., an arch. A negative label indicates that, whatever it is, it is not an example of that concept (e.g. not an arch). Currently the model assumes that concepts are mutually exclusive. When the first positive example for a new concept is provided, a generalization context is created for that concept. Positive examples are added to the appropriate generalization context, invoking SAGE on it. MAC/FAC is used to find a negative example similar to the positive example. If a sufficiently similar exemplar from a different concept is found, near-miss analysis is invoked. Similarly, when a negative example is provided, MAC/FAC is used to retrieve the closest positive exemplar or generalization, which is then used for near-miss analysis.

When given an example to categorize, the model uses MAC/FAC to generate a reminding from each concept's context. The system tests the new example against the classification criteria for each concept. Of the concepts whose criteria are satisfied, the one with the most similar reminding is chosen as the category of the new example.

In explaining our model, we use as a running example learning the concept of an *arch*, which was first used by Winston (1970), who used hand-generated representations.

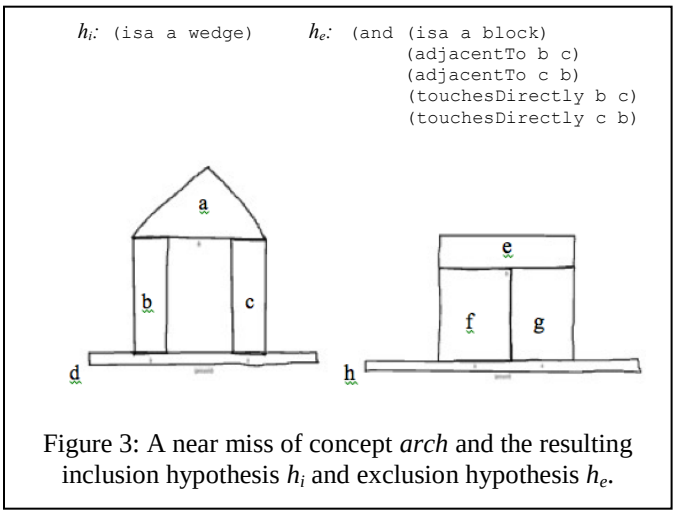


Figure 3: A near miss of concept *arch* and the resulting inclusion hypothesis h_i and exclusion hypothesis h_e.

Near-miss analysis. Winston argued for the importance of *near misses* in learning concepts. A near miss consists of a positive example e_1 (e.g. Figure 3, left) and a negative example e_2 (e.g. Figure 3, right) that differ only slightly.. In analogical reasoning terms, e_1 and e_2 are highly alignable, enabling a learner to conjecture that differences between them could be useful criteria for classification. Two kinds of hypotheses are computed to enhance concept discrimination. *Inclusion hypotheses* represent potential necessary conditions for something to be an instance of the concept. *Exclusion hypotheses* represent potential negative conditions that are sufficient to prevent something from being classified as an instance of that concept.

Near-miss analysis starts with a positive and a negative example. As noted above, one of these examples is a new learning example, while the other is a previous example retrieved via MAC/FAC. A similarity threshold of 0.75 is used for their comparison, to ensure high alignability.

Figure 3 shows a near miss that was processed by our simulation. The positive example is used as the base whereas the negative example is used as the target, and they are compared via SME. SME aligns a with e, b with f, c with g, and the grounds d with h. The conjunction of positive→negative candidate inferences in the mapping becomes a new inclusion hypothesis (Figure 3, h_i) designating criteria that might be necessary for concept membership. Similarly, the conjunction of all negative→positive candidate inferences is becomes a new exclusion hypothesis (Figure 3, h_e) designating criteria that might prevent concept membership. Here the attribute `(isa a wedge)` is the sole forward candidate inference, so it becomes the inclusion hypothesis h_i. Similarly, the `block` attribute, `touchesDirectly` relations, and `adjacentTo` relations comprise the conjunctive exclusion hypothesis h_e.

Inclusion and exclusion hypotheses are associated with the positive example in the near miss, as shown in Figure 2. Consequently, when MAC/FAC retrieves more than one near miss for a given positive example, the system computes more than one inclusion and exclusion hypothesis about the example, and must combine them. Inclusion hypotheses pertaining to the same example are combined via set union, since all necessary facts must hold for positive classification. Conversely, any exclusion hypothesis suffices to rule out that concept, so they are kept separate.

In Figure 3, the inclusion hypothesis h_i generated by the system erroneously asserts that all arches have wedges as their top. This error reflects one learning bias of the model, which is the immediate assumption that all differences detected in the near miss of a concept are important to the definition of the concept. Such errors can be removed during analogical generalization, which we discuss next.

Analogical generalization. During training, our learning system incrementally develops a disjunctive model of a concept through the observation of positive and negative examples. As positive examples are observed, they are added to a SAGE generalization context for the concept, where they are generalized with sufficiently similar

examples. When an example is generalized, resulting in new or larger generalizations (shown in Figure 2) the system revises the near-miss hypotheses associated with the generalization constituents.

Across generalizations, the near-miss hypotheses can be considered disjunctive hypotheses about the concept. For example, suspension bridges may be different enough from beam bridges that the classification hypotheses required of them differ. We can capture this distinction if suspension bridge examples and beam bridge examples form separate generalizations when added to the generalization context for the concept *bridge*. During classification, we may claim that an example is a bridge if it is similar enough to the *suspension bridge* generalization and satisfies the conditions for *suspension bridge*, or if it is similar enough to the *beam bridge* generalization and satisfies the conditions for *beam bridge*. The construction of disjunctive hypotheses based on similarity introduces another learning bias of the model, which assumes that similar examples of a concept are subject to the same rules for membership.

After an observed positive example is generalized with an existing generalization or ungeneralized example, their hypotheses are generalized. Figure 4 shows how a new example (top) and a previously ungeneralized example (middle) are merged into a new generalization with revised hypotheses (bottom).

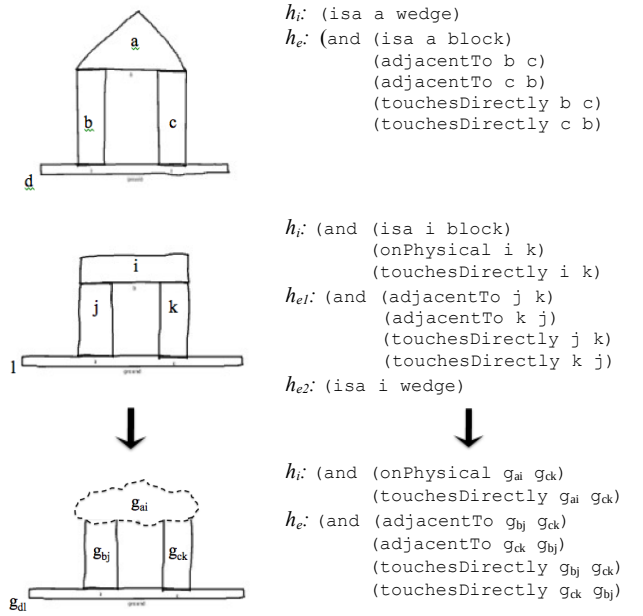


Figure 4: The generalization of two positive examples and their inclusion and exclusion hypotheses

The first step in generalizing inclusion hypotheses is mapping the hypotheses from their respective generalized examples to the newly created generalization. This involves replacing the names of entities with the names of corresponding entities in the generalization. Next, inclusion hypotheses are pruned by removing any assertions that do not hold on the new generalization. In Figure 4, the facts `(isa a wedge)` and `(isa i block)` are pruned from the inclusion hypotheses of the constituent examples because they are not true of the resulting generalization, i.e., the corresponding generalized entity g_{ai} is not known to be either `wedge` or `block`. After pruning, the facts of the two inclusion hypotheses are unioned to create a conjunctive hypothesis associated with the new generalization

Next, the system uses the generalization operation to identify and discard erroneous exclusion hypotheses. In Figure 4, exclusion hypothesis `(isa i wedge)` of the middle example is erroneous because it shares a generalization with the topmost example whose corresponding entity *a* is a `wedge`. Consequently, the exclusion hypothesis is discarded. Remaining exclusion hypotheses are mapped onto the resulting generalization. Finally, the system discards exclusion hypotheses of the resulting generalization that are more specific than other associated hypotheses (i.e., for every exclusion hypothesis composed of fact set *f*, any hypothesis of fact set *f'* such that $f \subseteq f'$ is eliminated). In Figure 4, hypothesis h_e of the topmost example is discarded for this reason.

Classification

Given a new testing example e_{new}, our model decides whether it is an instance of one of its learned concepts. The model decides this using similarity-based retrieval and by testing the hypotheses created during learning.

For each learned concept, the system uses MAC/FAC to retrieve the most similar generalization or ungeneralized example of the concept e_c from the concept's generalization context. The inclusion and exclusion hypotheses associated with e_c (as shown in Figure 2) are used as criteria for classifying e_{new}.

The inclusion and exclusion hypotheses associated with e_c are represented in terms of the entities in e_c, which typically do not exist in e_{new}. Consequently, structural alignment is used to perform the analogical equivalent of rule application. SME is used to find entity correspondences between e_c and e_{new}, and the entities of e_c are substituted with the corresponding entities in e_{new} in each hypothesis.

Testing the classification criteria is the final step in classification. If an inclusion hypothesis does not hold in e_{new}, or if an exclusion hypothesis does hold in e_{new}, it is not an instance of the concept. Otherwise, e_{new} is an instance of the concept. If e_{new} is a viable instance of multiple concepts, given the exclusion and inclusion criteria, the system chooses the concept whose MAC/FAC reminding similarity score was higher. Thus our model of concepts combines both rule-based and similarty-based aspects.

Experiment

We created a series of 44 sketches representing six concepts for learning and categorization, summarized in Table 1. The *false arches*, *false triangles*, and *false squares* sketches are all highly alignable with examples of their associated concept, but are not positive examples themselves.

Table 1: Sketched examples for simulation.

Arches:	8	Triangles:	4
False arches:	8	*False triangles:*	4
Bridges:	4	Squares	4
Skeletal arms:	4	*False squares:*	4
Skeletal legs:	4		

Our experiment follows a four-fold cross validation format covering all 44 sketches. The sketches were randomly assigned to four groups (folds) of 11 sketches each, with the constraint that all groups had the same distribution of sketches from the categories in Table 1 (two arches, two false arches, one bridge, one skeletal arm, etc). The system trained on three 11-example groups, for a total of 33 examples for learning. The remaining group of 11 examples is used for classification testing. We repeat this four times, so each group of 11 examples is used once for testing, resulting in 44 classifications total.

We tested our simulation under two conditions: The *full* condition uses the entire model, while in the *similarity-only* condition, near-miss analysis is turned off. In similarity-only, the system classifies a new example by using MAC/FAC to retrieve a similar representation from the concept context, and asserts concept membership if the normalized SME similarity score is above a threshold of 0.85. We expected that, based on prior experiments (Kuehne *et al* 2000b), similarity-only will learn quite well with only a handful of examples. However, we also expect that it will show false positives due to misleadingly similar negative examples, which near-miss analysis should prevent.

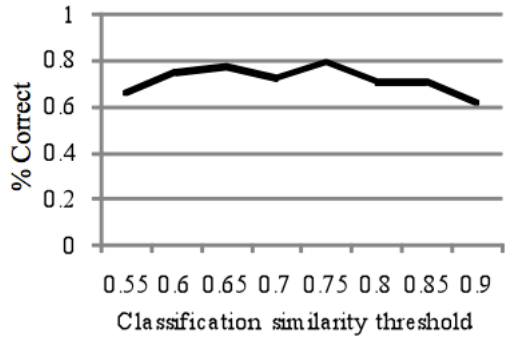


Figure 5: Effectiveness of using structural similarity alone for classification, as a function of similarity threshold.

In the similarity-only condition, 79% correct classification is achieved with a similarity threshold of 0.75 (Figure 5), well above chance ($p < 0.001$). Inspection of the results revealed that almost all of the 20% error can be attributed to false positives. One such false positive is the rightmost example in Figure 3, which shares considerable relational structure with other arches.

With near-miss analysis turned on, 86% correct classification was achieved, which is better than chance with $p < 0.001$. The number of false positives decreased from eight to two but the number of false negatives increased from one to four due to overly restrictive hypotheses. The rightmost example in Figure 3 was among the negative examples correctly classified. Just as with similarity-only, the model determined that this example was sufficiently similar to a generalization of the concept *arch*. However, it reported a failure to meet classification conditions due to a satisfied exclusion hypothesis,

```
(TheSet (adjacentTo f g)
        (touchesDirectly g f))
```

which expresses the justification "This is not an arch because f is adjacent to g and g touches f directly."

Discussion & Future Work

We have described a model that extends analogical generalization with near-miss analysis to learn concepts from sketches. We have generalized the notion of near-miss that Winston (1970) used in two important ways. First, Winston assumed that near-misses were always provided by a teacher. We have shown that near misses can also naturally arise from the process of similarity-based retrieval, thereby providing more self-direction in learning. Second, Winston's system had one description of the target concept it was learning, and hence did not capture the possibility of disjunctive concepts and finding the appropriate conceptual representation, which we do via a combination of SAGE and MAC/FAC. A version of the model without near-misses, using similarity alone, performs well over chance. However, similarity alone leads to a pattern of misclassification errors, which is partially corrected by near-miss analysis. The incorporation of classification criteria enables the model to make more expressive justifications for its classification decisions, as in the case of the negative example from Figure 3. We also believe that near-miss analysis will allow the model to more readily benefit from a larger training set, as hypotheses from new near-misses will add potentially valuable criteria to reduce false positives and hypothesis generalization will alleviate over-restrictiveness, which accounted for all but one of the false negatives. We expect the similarity-only classifier to gain less from additional training, since the examples it misclassifies are mostly negative examples that bear high relational similarity to positive examples. Thus near-miss analysis provides an important extension to similarity-based concept learning.

Our concept learning model learns several concepts simultaneously, with relatively few examples. It requires orders of magnitude fewer examples than existing connectionist models of concept learning (e.g., Krushke, 1992; Regier 1996; Elman 1999), and unlike such models, uses automatically encoded relational stimuli, to reduce tailorability. We believe our model makes more realistic demands, although it could be argued that our model learns too quickly. One reason that we see such rapid learning in simulation experiments is that our system, unlike people, has many fewer distracters. Everyday life does not always afford closely packed sequences of similar concept

instances, and human perception may contain more attributes and relations than CogSketch currently computes. However studies such Gentner *et al* (2009) suggest that people can learn spatial concepts quickly with highly alignable near-misses, which our model captures nicely.

Winston (1982, 1986) also explored learning rules from analogies, using simplified English inputs. His system generalized based on one example, rather than several, and produced logical quantified rules, while ours uses analogical matching to apply hypotheses to new examples. His if-then rules and censors are functionally similar to our inclusion and exclusion hypotheses, respectively.

There are several aspects of concept learning that our model does not currently capture. For example, our sketched input does not include causal relationships or goals (Lombrozo, 2009; Rehder & Kim, 2006). Based on prior work (Falkenhainer, 1987; Friedman & Forbus, 2009) we believe our model will handle such information if it is included in the initial encoding, since it basically adds relational structure that influences similarity judgments, and hence classification, in our model. Other factors, such as ontological structure (Medin & Smith, 1984) and centrality and mutability of properties (Sloman, Love, & Ahn, 1998) we believe can be handled by further exploiting the statistical information gathered via SAGE in cross-concept analyses. We plan to explore both of these issues in future work.

Acknowledgments

This work is supported by the Cognitive Science Program of the Office of Naval Research.

References

Elman, J. (1999). Generalization, rules, and neural networks: A simulation of Marcus et. al, (1999). Ms., University of California, San Diego.

Falkenhainer, B., Forbus, K. and Gentner, D. (1989). The Structure Mapping Engine: Algorithm and examples. *Artificial Intelligence*, 41, 1-63.

Forbus, K., Klenk, M., and Hinrichs, T. (2009). Companion Cognitive Systems: Design Goals and Lessons Learned So Far. *IEEE Intelligent Systems*, vol. 24, no. 4, pp. 36-46, July/August.

Forbus, K., Lovett, A., Lockwood, K., Wetzel, J., Matuk, C., Jee, B., and Usher, J. (2008). CogSketch. *Proceedings of AAAI 2008*.

Forbus, K., Gentner, D. and Law, K. (1995). MAC/FAC: A model of similarity-based retrieval. *Cognitive Science*, 19(2), 141-205.

Friedman, S. & Forbus, K. (2008). Learning Causal Models via Progressive Alignment & Qualitative Modeling: A Simulation. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society*.

Friedman, S. & Forbus, K. (2009). Learning Naïve Physics Models & Misconceptions. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society*.

Gentner, D. (1983). Structure-Mapping: A Theoretical Framework for Analogy. *Cognitive Science*, 7: 155-170.

Gentner, D., Levine, S., Dhillon, S. & Poltermann, A. (2009). Using structural alignment to facilitate learning of spatial concepts in an informal setting. In *Proceedings of the Second International Workshop on Analogy*, Sofia, Bulgaria, 2009.

Gentner, D., Loewenstein, J., & Hung, B. (2007). Comparison facilitates children's learning of names for parts. *Journal of Cognition and Development*, 8. 285-307.

Gentner, D. & Medina, J. (1998). Similarity and the development of rules. *Cognition* 65(2-3):263-97.

Kruschke, JK (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review 99*, 22-44.

Kuehne, S., Forbus, K., Gentner, D. and Quinn, B. (2000). SAGE: Category learning as progressive abstraction using structure mapping. *Proceedings of CogSci 2000*.

Kuehne, S., Gentner, D. and Forbus, K. (2000). Modeling infant learning via symbolic structural alignment. *Proceedings of CogSci 2000*.

Lockwood, K., Lovett, A., and Forbus, K. (2008). Automatic Classification of Containment and Support Spatial Relations in English and Dutch. In the *Proceedings of Spatial Cognition 2008*.

Lombrozo, T. (2009). Explanation and categorization: how "why?" informs "what?". *Cognition*, 110, 248-253.

Medin, D. and Smith, E. (1984). Concepts and concept formation. *Annual Reviews of Psychology*, 35, 113-138.

Regier, T. *The human semantic potential: Spatial language and constrained connectionism*, Cambridge Mass: MIT Press (1996).

Rehder, B. & Kim, S. (2006). How causal knowledge affects classification: A generative theory of categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32, 659-683.

Rips, L. J., & Handte, J. (1984). Classification without similarity. Unpublished manuscript, Univ. Chicago.

Sloman, S., Love, B., Ahn, W. K. (1998). Feature centrality and conceptual coherence. *Cognitive Science* 22(2). 189-228.

Winston, P.H. 1970. Learning structural descriptions by examples. Ph.D. thesis, MIT.

Winston, P.H. 1982. Learning new principles from precedents and exercises. *Artificial Intelligence* 23(12).

Winston, P.H. 1986. Learning by augmenting rules and accumulating censors. In Michalski, R., Carbonell, J. and Mitchell, T. (Eds.) *Machine Learning: An Artificial Intelligence Approach, Volume 2.* Pp. 45-62. Morgan-Kaufman.

Inferring Multitasking Breakpoints from Single-Task Data

Peter Bogunovich (pjb38@drexel.edu)
Drexel University, Department of Computer Science
Philadelphia, PA, USA

Dario D. Salvucci (salvucci@cs.drexel.edu)
Drexel University, Department of Computer Science
Philadelphia, PA, USA

Abstract

Recent research has shown that computer users placed in a deferrable multitasking situation generally postpone secondary-task interruptions until points of low mental workload in the primary task. Studies examining this phenomenon have relied on empirical data that explicitly show user switch points in the course of multitask performance. This paper addresses a related question: Can these same switch points, found empirically in a multitasking context, be inferred solely from single-task data? We investigate this question and propose an approach that analyzes a particular behavioral signature in single-task data—outliers in the distributions of time between task actions—to infer multitasking breakpoints. We evaluate this approach using behavioral data from a user-interface task, showing how the proposed method's inferences from single-task data match well to the real switch points observed during multitask performance.

Keywords: Multitasking; task analysis; data analysis.

Introduction

Multitasking is a concept that is familar to most computer users. It is not uncommon for a user to switch computing tasks every few minutes. In many cases switching is initiated by an interruption of the current task. For example, a notification of a newly received email may appear on the screen prompting a user to stop what he is doing and look at his email before continuing his previous task. Research has shown that interruptions can increase the overall time spent on a single task. One important source of this increase is the *resumption lag*, or time required to switch back to the task and resume after the interruption has been addressed (Trafton, Altmann, Brock, & Mintz, 2003; Monk, Boehm-Davis, Mason, & Trafton, 2004). Recently it has been shown that it is more beneficial to interrupt at certain points than at others (Adamczyk & Bailey, 2004; Bailey & Konstan, 2006; Cutrell, Czerwinski, & Horvitz, 2000). One particularly strong result states that the performance loss associated with interruption is reduced when interruptions occur at points of low mental workload (Iqbal & Bailey, 2005). This result has obvious importance when considering *forced interruptions* in which the user is required to address the interruption immediately before moving on with the primary task.

The relationship between mental workload and interruptibility has been strengthened in further studies of *deferrable interruptions* (Salvucci & Taatgen, 2010) in which a user is notified of a secondary task but the user can defer processing of this task until a later (presumably more comfortable) time. For example, it has been shown (Salvucci & Bogunovich, 2010) that in this situation users tend to defer switching tasks until a point where there is a drop in mental workload. As exemplified by these studies, a detailed analysis of when users switch tasks is critical to a deeper understanding of human multitasking behavior. A particular goal in this line of research involves the prediction of breakpoints, the points in a task sequence where the user can most conveniently switch tasks.

One approach to breakpoint prediction combines expert coding, feature detection and model prediction (Iqbal & Bailey, 2007). This approach begins by observing users in some natural multitasking environment. An expert manually examines user actions and identifies specific features which appear to signal breakpoints. A statistical model is then developed based on these features. Promising results have been obtained with his method, however it requires the human coders to identify the perceived breakpoints and features, and does not necessarily make use of the relationship between cognitive load and interruptibilty. A successful related approach that makes use of mental workload is to examine the typical execution structure of an action in advance and use this structure to estimate opportune breakpoints (Bailey, Adamczyk, Chang, & Chilson, 2006). This method still requires expert analysis and it may fail when variation in strategy is introduced.

There exists a well-known relationship between cognitive load and pupil dilation (Beatty, 1982). Researchers have made use of this link in another approach to breakpoint detection (Bailey & Iqbal, 2008). In this approach, pupil dilation data is recorded as users perform a task, and subtask boundaries, where there is an assumed drop in cognitive load, are estimated by changes in dilation. The result is a more general and more automatic estimation of good potential breakpoints that relies less on pre-computed models or experts. Despite these findings, it may not be possible to obtain pupil-dilation in practice for many tasks.

In this paper we attempt to infer multitasking breakpoints in a automatic, data-driven manner. In this respect our approach is most similar to (Bailey & Iqbal, 2008), but instead of relying on typically inaccessible equipment like eye-trackers, our goal is to come up with the good estimates using only data logs of system events generated by users performing a single primary task. Our analysis focuses on the distributions of elapsed time between recorded event pairs, using single-task data collected for a customer-support task

(Salvucci & Bogunovich, 2010). From our analysis of the recorded data, and particularly the estimation of observed outliers in distribution tails, we were able to infer breakpoints that closely mirror actual deferred user breakpoints as they arose in a multitasking context.

Task and Data

The task that we analyzed is taken from a recent experiment in which users performed a mail-based customer-support primary task while occasionally being interrupted by chat (instant-message) questions. The primary task simulated a typical customer-service scenario where a user receives email inquiries for the prices of a variety of products. The simulation was comprised of a simulated email program and a browser window used for looking up product prices, shown in Figure 1. Each email in the inbox contained a request for the price of a single product. Once the user read the email and became aware of the request, he or she had to look up the product in the browser to obtain the correct price. Each product consisted of a real manufacturer name and a fictitious model identifier (for example, "Canon H-44", or "Sony M-76"). To find the price of a product, the user had to first click on the proper manufacturer name from the top-level of the browser, and then click on the proper model identifier from a secondary browser level. The user could have at any time returned to the top level of the browser by clicking "home" button. Once the price of the product in question had been located, the user sent a reply email containing the requested information. The users were also asked to manually move the replied to emails to a "replied" bin by clicking and dragging.

In the multitasking setting a secondary chat task was introduced which simulated a typical instant messenger conversation. A chat window was included in which the users were occassionally asked questions about recent films by a simulated interlocutor. The users were notified of a new question by having the chat window flash, but it was up to the users to decide when to break from the primary mail task to address the questions once the notification was received.

It is important to note that in both the single mail task and dual mail and chat task situations, the simulation windows were arranged so that only the window that was currently being focused on could be seen. For example, while looking up a product price in the browser window, the name of the product given in the email window was obscured. This required the users to commit sub-task relevant information to memory.

For our analysis, we look specifically at single mail task data collected from six participants in this experiment. This data was collected in a sesssion where the chat simulation was not present. In particular, our goal is to analyze the single-task data, infer and estimate breakpoints from these data, and then evaluate our estimates by comparing the results to the also collected multitask data. The data recorded for the mail task (both single- and dual-task contexts) comprises a sequence of time-stamped events occurring in the task. Table 1 lists and describes these events. The full data recorded for a single event includes the event type, as given in Table 1, the time of the event, and any auxilary information about the event (for example, which character was typed, or which product link was clicked); we use only the event type and time information here.

mail-select:	Select (click on) an email from a list.
mail-move:	Move (drag) an email to the "Replied" bin.
browser-focus:	Change focus to browser window.
browser-home:	Press "browser home" button.
mfr-link:	Click on "product manufacturer" link.
model-link:	Click on "product model" link.
reply-button:	Press "Reply" button to open a new window to compose response.
reply-type:	Type characters in a response email.
reply-send:	Press the "Send" button to send response email.
reply-focus:	Change focus to an opened response window.

Table 1: User events in the mail customer-support task.

Analysis of Recorded Event Data

Starting with the recorded single-task data, we tried several theoretically-motivated approaches for analyzing the data and inferring multitasking breakpoints. In the following sections we discuss several of the approaches that we took. Motivations and limitations associated with each approach are given.

Frequency of Sequences

When relying solely on the frequencies of occurrence of given event sequences, perhaps the most naive hypothesis is that good locations for breakpoints are found between pairs of consecutive events that were observed infrequently. The motivation is that sequences which appear frequently consist of events that are strongly linked together, and thus switching tasks between the events is less desirable or at least less likely.

Problems with this hypothesis arise immediately, however, in noting that it is extremely unlikely or impossible for many pairs of events to occur consecutively. For instance, in the mail task, it is not possible for to observe the event "*model-link*" followed immediately by the event "*mfr-link*" due to the design of the task interface. Other pairs of consecutive events are unlikely due not to the design of the simulation, but simply because they make little sense for any user attempting to complete the mail goal. For example, the sequence "*mfr-link* → *browser-home*" is not useful in looking up a product price, since the price is not obtained until the "*model-link*" event. Any occurrences of "*mfr-link* → *browser-home*" are likely due to an error by the user and there is little reason to believe that this is a good place to switch tasks.

While it is clear that pairs of events with no or few occurrences do not necessarily represent good breakpoints, it still

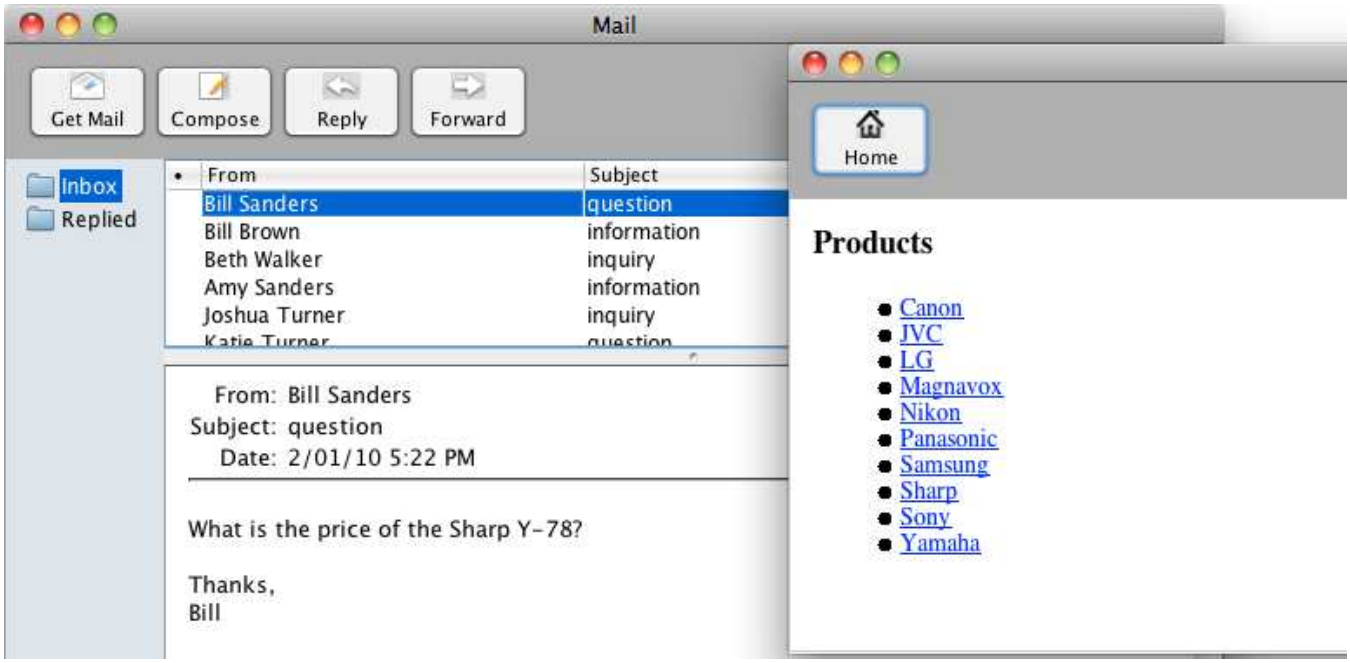


Figure 1: The customer service mail simulation.

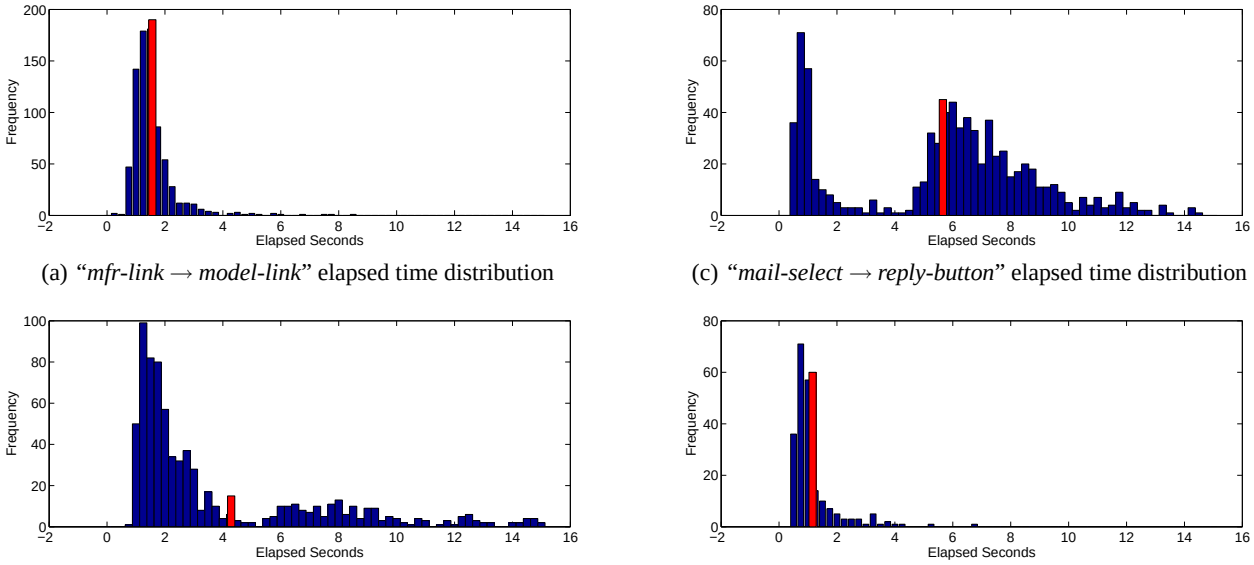
seems possible that pairs of consecutive events with high frequency represent strongly linked events and that task switching should not occur between them. This argument is supported when we compile a list of the most frequent consecutive event pairs and observe that one of the highest frequency consecutive event pairs is "*mfr-link* → *model-link*". It makes sense that we should link together these actions as they are the ordered steps required to look up a product's price. There should not be a drop in cognitive load after the "*mfr-link*" event since the model number is still required for the following "*model-link*" event and we should not expect task switching here. On the other hand, another high frequency consecutive event pair is "*reply-send* → *mail-select*". While these events appear to be strongly linked together, this pair actually does present a reasonable breakpoint. The "*reply-send*" event signals that a response email has been sent and a customer inquiry is completed. Handling a new customer inquiry is always marked by selecting a new mail from the list, or a "*mail-select*" event. It follows that the pair "*reply-send* → *mail-select*" is a task boundary and a drop in cognitive load should accompany it, making this a good breakpoint.

Mean Elapsed Time

A second attempt at identifying breakpoints involves considering the mean elapsed time between events. The hypothesis is similar to the frequency hypothesis: A low mean elapsed time between two events signals a strong link between them that should not be broken, while a large mean elapsed time between events indicates a weak link that may be broken when an interruption occurs.

For a given pair of events such as "*A*" and "*B*", it is not immediately clear how to construct the frequency distribution. We could look at all occurrences of "*A*" followed by a "*B*" any time thereafter, with the possibility of some events in between. This approach is appealing since it introduces some robustness to "noisy" user errors in the recorded events. We see some positive evidence supporting this choice in the distributions shown in Figure 2(a) and Figure 2(b). In both of these distributions, the mean of the histogram is indicated by a red (lighter) bar. The distribution shown in Figure 2(a) corresponds to the event pair "*mfr-link* → *model-link*", which as a sequence makes sense in a goal strategy and does not represent an expected boundary of cognitive subtasks. The mean of this distribution is about 1.56 seconds elapsed between the occurrence of the two events. The distribution shown in Figure 2(b) corresponds to the event pair "*model-link* → *reply-button*", which occurs when the user has completed the task of looking up the price of a product and is about to begin the process of responding to the inquiry. The mean of this distribution is 4.29 seconds of elapsed time between events. The larger mean found here supports the hypothesis, since this pair of events should straddle a subtask boundary and a drop in cognitive load should accompany it.

The idea of considering all occurrences of "*A*" followed some time later by "*B*" begins to break down, however, when we consider the distribution shown in Figure 2(c). This distribution corresponds to the elapsed time between the events "*mail-select*" and "*reply-button*". The mean elapsed time is 5.66 seconds, which seems to indicate that the events are not strongly linked. The problem with this assessment becomes clear when we take into consideration the variations in task strategies taken by different users. The consequence



(a) "*mfr-link* → *model-link*" elapsed time distribution

(c) "*mail-select* → *reply-button*" elapsed time distribution

(b) "*model-link* → *reply-button*" elapsed time distribution

(d) "*mail-select* → *reply-button*" distribution for one strategy

Figure 2: Distributions of elapsed time between pairs of events. The locations of the means are indicated by a red (lighter) bar.

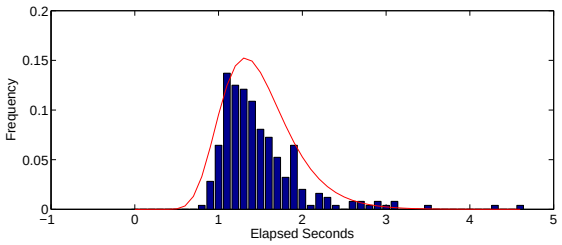
of this is that the sequence "*mail-select* → *reply-button*" is strongly linked together in several task strategies, but it is not found in all of them. This explains the two peaks seen in the histogram. The first peak (and the surrounding bins) correspond to the instances of the strategies which make use of the "*mail-select* → *reply-button*" sequence, while the second peak corresponds to the remaining strategies. Analyzing this sequence simply based on the mean of all of the possible occurrences does not provide a clear understanding of the data.

Addressing Multiple Strategies

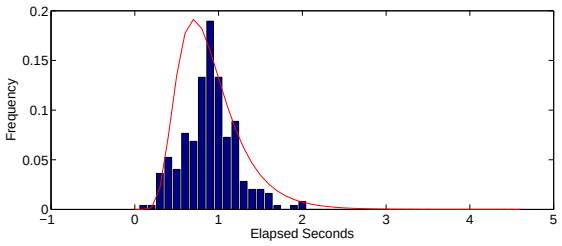
Regardless of the usefulness of the mean elapsed time in indicating the breakpoints, the observation concerning the multiple strategies needs to be addressed in any analysis of distributions. It seems that our distributions represent a classic example of a mixture distribution, which should lead us to consider a method such as expectation maximization (EM) (Moon, 1996) to fit a mixture model to the histogram. Once we've found a mixture model, we could then perform clustering to obtain only the instances of event sequences which should correspond to a single strategy. Another approach would be to use the T-Patterns method for identifying the critical interval (Magnusson, 2000) of elapsed time that we should consider acceptable for a given event pair. Both of these approaches present advantages and disadvantages for our data, and are likely to prove both useful and necessary in analyzing tasks containing variation in general.

We decided to use a much simpler approach to identifying the valid instances of a sequence. Based on the task that was assigned, we note that each task trial—the processing of a single email—must begin with a "*mail-select*" event to view the email. Furthermore, that once a new email has been selected, another "*mail-select*" event is very unlikely before this first email has been completely addressed. Following these assumptions, we can segment our raw event data stream into individual mail task instances by using each "*mail-select*" event as a boundary and consider unique sequences separately. This method is supported by Figure 2(d), where only the instances of the sequence "*mail-select* → *reply-button*" which are part of a strategy using those consecutive events are considered in the distribution. When compared to Figure 2(c) we now see a single a peak with a mean of 1.30 seconds versus two peaks and a mean of 5.66 seconds.

By considering instances of consecutive event pairs which are part of a particular observed task strategy, a lot of unexpected behavior in the elapsed time distributions is removed, but not enough to make the mean elapsed time a completely useful indicator of cognitive load or interruptibility. One reason for this lies in the simple nature of the data that was recorded. By comparing just the elapsed time between events "*A*" and "*B*", the analysis does not have at its disposal vital information about possible subtasks being performed. Consider once again the "*mfr-link* → *model-link*" sequence. Generally this sequence is observed when the user is looking up the price of a product for a customer inquiry. For one strategy which uses this sequence (actually all strategies must use this), we get a mean of 1.49 and a relatively large st. dev. of 0.59. Based on our hypothesis we should expect both a small mean and variance for such a strongly linked pair of events, but in fact we see a relatively large variance. This contradiction is explained when we consider that after a "*mfr-link*" event, a user completing this action is required to perform the relatively time-consuming task of reading through the list of model numbers to find the link for the model in question, before the "*model-link*" event can occur. A similar statement could be made about any event preceding the "*mfr-link*"



(a) "*model-link* → *reply-button*" distribution for one strategy



(b) "*reply-button* → *reply-type*" distribution for one strategy

Figure 3: Distributions for instances of pairs found in one strategy. The histograms are shown with a fitted log-normal distribution curve. Note that in 3(a) more of the mass falls in the far right end of its tail than in 3(b).

event. (More detailed data, such as eye-movement recordings, would further inform such an analysis—but again, such detailed data are not available in the general case.)

Tail Mass of Elapsed Time Distributions

Since basic statistics of our elapsed time distributions do provide an adequate signature with respect to multitasking breakpoints, we decided to take a closer look at the form of the distributions. When we compare the histogram distributions for different pairs of events, it becomes clear that certain histograms appear to have longer tails than others. To obtain a better picture of this, we could look at the amount of the histogram mass that falls several standard deviations to the right of the mean. We can also observe modeling the histogram with a normal distribution may not be the best choice, since there can be no negative elapsed times and typically the distributions exhibit an early peak followed by a right end tail. The log-normal distribution has these properties and we can easily find a maximum likelihood log-normal distribution to fit to our observations. Figure 3 shows two pair histograms that have been fitted with log-normal distributions. Figure 3(a) shows the distribution for the pair "*model-link* → *reply-button*", which corresponds to the boundary between the price lookup task and the email reply task and is a reasonable breakpoint. Figure 3(b) shows the distribution for the pair "*reply-button* → *reply-type*", which form consecutive events in the mail reply task and probably is not a good breakpoint. Notice that a significantly larger portion of the total observed mass in Figure 3(a) appears in the far right tail of the fitted distribution than does the mass in Figure 3(b). Another way to put it is that the "*model-link* → *reply-button*" distribution contains significantly more outliers than the "*reply-button* → *reply-type*" distribution.

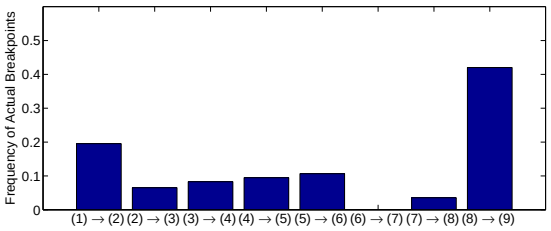
The hypothesis resulting from this analysis is that the amount of observed mass in the far end of the tails (outliers) of distributions of elapsed time between event pairs is a good indicator of the interruptibility between the events. We suspect that the underlying reason relates to people taking short mental breaks between these task steps: by resting for a short time (up to a few seconds) between actions, a person can mentally regroup for the next component of the task. It seems reasonable that such a mental regrouping would occur at higher-level task boundaries, or equivalently at places of low mental workload. Whatever the underlying reason, the tails of the distributions seem to serve as a good signature for multitasking breakpoints, as we detail in the next section.

To identify the outlier observations, we can simply fit the model to our observations and see how many observations fall n standard deviations to the right of the mean. Since we are specifically interested in outliers in the far right end of the tail, we should set n to be large, possibly $n = 3$ or 4. This simple method will certainly identify some outliers, but we can improve the method by performing it iteratively. In the iterative approach, we first fit the model, find the estimated std. dev., remove outliers n standard deviations from the mean from the distribution, and repeat. At each iteration the estimated mean will shift slightly to the left and we will consider more observations to be outliers. For large fixed n the estimates converge after a few iterations (i.e., no new outliers are found). At that point we have a good estimate of the percentage of the total observations which can be considered outliers.

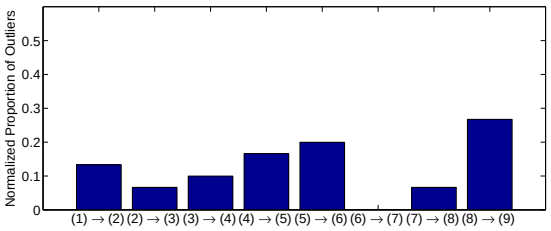
Results

To evaluate the outlier-based inference of multitasking breakpoints, we selected the events corresponding to the most frequently observed mail task strategy that we obtained from our data segmentation procedure. The complete sequence has the form: *mail-select*, *browser-focus*, *browser-home*, *mfr-link*, *model-link*, *reply-button*, *reply-type*, *reply-send*, *mail-select*. We calculated the outliers for each pair of consecutive events, and formed a normalized histogram of breakpoint likelihoods where the frequency of each bin is based on the number of outliers that were found. Our results were obtained using a log-normal distribution to fit our elapsed time distributions and a value of $n = 3.75$ standard deviations for the identifying outliers. Using the accompanying multitasking (mail and chat task) data, we also constructed a similar histogram of the actual deferred breakpoints that we taken by users while employing this strategy.

Both of the resulting histograms are shown in Figure 4. The inferred results match reasonably well to the observed breakpoints, $R = 0.83$. We obtained similar but not as good results using the normal distribution, and for several observed secondary strategy sequences.



(a) Observed Proportions of Breakpoints while Multitasking



(b) Proportions of Breakpoints Inferred from Single-Task Data

Figure 4: Comparison of actual breakpoints taken in (Salvucci & Bogunovich, 2010) with the outlier inferred breakpoints for the most frequent strategy: (1) *mail-select*, (2) *browser-focus*, (3) *browser-home*, (4) *mfr-link*, (5) *model-link*, (6) *reply-button*, (7) *reply-type*, (8) *reply-send*, (9) *mail-select*.

Discussion

To summarize, we found that the outliers (tails) of the distributions of time between task actions in a single task setting served as a good indicator of multitask breakpoints, were a secondary task to be introduced: The presence (or lack) of outliers in the tails of the distributions correlated well with people's tendency to switch away from a task between two given actions. These conclusions build on the results of (Bailey & Iqbal, 2008) which showed that users produce evidence of potential interruptibility in a single-task setting, but the proposed method was able to identify similar evidence using solely time and event data (rather than pupil-dilation or other data that may be more difficult to obtain). Our results suggest that when performing a task, users may occasionally take a short breaks (up to a few seconds) when a cognitive subtask is completed and before beginning a new subtask. Analysis based on this idea agrees well with multitask data from (Salvucci & Bogunovich, 2010) and hints at a strong relationship between distribution outliers and boundaries of cognitive subtasks.

Acknowledgments

This work was funded by ONR grant #N00014-09-1-0096.

References

Adamczyk, P. D., & Bailey, B. P. (2004). If not now, when?: the effects of interruption at different moments within task execution. In *Chi '04: Proceedings of the sigchi conference on human factors in computing systems* (pp. 271–278). New York, NY, USA: ACM.

Bailey, B. P., Adamczyk, P. D., Chang, T. Y., & Chilson, N. A. (2006). A framework for specifying and monitoring user tasks. *Computers in Human Behavior*, *22*(4), 709–732.

Bailey, B. P., & Iqbal, S. T. (2008). Understanding changes in mental workload during execution of goal-directed tasks and its application for interruption management. *ACM Trans. Comput.-Hum. Interact.*, *14*(4), 1–28.

Bailey, B. P., & Konstan, J. A. (2006). On the need for attention-aware systems: Measuring effects of interruption on task performance, error rate, and affective state. *Computers in Human Behavior*, *22*(4), 685–708.

Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the sturcture of processing resources. *Psychological Bulletin*, *91*(2), 276–292.

Cutrell, E. B., Czerwinski, M., & Horvitz, E. (2000). Effects of instant messaging interruptions on computing tasks. In *Chi '00: Chi '00 extended abstracts on human factors in computing systems* (pp. 99–100). New York, NY, USA: ACM.

Iqbal, S. T., & Bailey, B. P. (2005). Investigating the effectiveness of mental workload as a predictor of opportune moments for interruption. In *Chi '05: Chi '05 extended abstracts on human factors in computing systems* (pp. 1489–1492). New York, NY, USA: ACM.

Iqbal, S. T., & Bailey, B. P. (2007). Understanding and developing models for detecting and differentiating breakpoints during interactive tasks. In *Chi '07: Proceedings of the sigchi conference on human factors in computing systems* (pp. 697–706). New York, NY, USA: ACM.

Magnusson, M. (2000). Discovering hidden time patterns in behavior: T-patterns and their detection. *Behavior Research Methods, Instruments and Computers*, *32*, 93–110.

Monk, C. A., Boehm-Davis, D. A., Mason, G., & Trafton, J. G. (2004). Recovering From Interruptions: Implications for Driver Distraction Research. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *46*(4), 650–663.

Moon, T. (1996, Nov). The expectation-maximization algorithm. *Signal Processing Magazine, IEEE*, *13*(6), 47–60.

Salvucci, D. D., & Bogunovich, P. (2010). Monotasking and multitasking: the effects of mental workload on deferred task interruptions. In *Proc. CHI 2010*. ACM.

Salvucci, D. D., & Taatgen, N. A. (2010). *The multitasking mind*. New York: Oxford University Press.

Trafton, J. G., Altmann, E. M., Brock, D. P., & Mintz, F. E. (2003). Preparing to resume an interrupted task: effects of prospective goal encoding and retrospective rehearsal. *International Journal of Human-Computer Studies*, *58*(5), 583–603.

Comparing Human-Human to Human-Computer Tutorial Dialogue

Natalie B. Steinhauser (Natalie.Steinhauser@navy.mil) &
Gwendolyn E. Campbell (Gwendolyn.Campbell@navy.mil)
Naval Air Warfare Center Training Systems Division, Code 4.6.5.1
12350 Research Parkway, Orlando, FL 32826-3275

Katherine M. Harrison (Katherine.M.Harrison.ctr@navy.mil)
Kaegan Corporation
12000 Research Parkway, Orlando, FL 32826-2944

Leanne S. Taylor (Leanne.Taylor.ctr@navy.mil)
University of Central Florida
4000 Central Florida Blvd. Orlando, FL 32816

Myroslava O. Dzikovska (M.Dzikovska@ed.ac.uk) & Johanna D. Moore (J.Moore@ed.ac.uk)
Human Communication Research Centre, University of Edinburgh
2 Buccleuch Place, Edinburgh EH8 9LW, United Kingdom

Abstract

Intelligent Tutoring Systems are often modeled after human tutors; however, the effectiveness of this strategy is yet to be determined. Research on media interactions suggests that behaviors with humans are similar to those with computers. Intelligent Tutoring System studies have said the opposite. In this study we compared a human-human and a human-computer tutoring system in terms of metacognitive, social, and nonsense statements to dig deeper into these interactions. We discovered that the interactions were quite different between human-human and human-computer tutoring. With a human, participants expressed more positive metacognitive statements and social statements. When interacting with a computer tutor, students were more likely to make negative metacognitive statements and social statements. In addition, the interpretation of these results differed between the two corpora. In human-human tutoring, the more often a participant made positive metacognitive statements, the worse their learning gain. Their social dialogue had no impact on learning gain. In human-computer tutoring, the more negative and positive metacognitive statements and the more negative social statements they gave the worse their learning gain. It is clear from this study that students do not act the same with a human tutor as they do with a computer tutor. Therefore, designers of ITS systems should not just blindly model their systems after human tutors. The differences in human and computer interactions should also be considered.

Keywords: Human Computer Interaction (HCI), Intelligent Tutoring Systems (ITS), Metacognition, Social dialogue, Tutorial dialogue

Introduction

Over the years, Intelligent Tutoring Systems (ITSs) have become popular learning and teaching tools. Thus, their design is becoming more sophisticated. One approach to creating ITSs is to model them after a human tutor because human tutoring has been said to be the most effective form of teaching (Bloom, 1984). However, it has not yet been determined that this is a good strategy. Two unresolved questions are whether you will find the same kinds of dialogue when a student is interacting with a human and a computer tutor (ITS) and whether those types of dialogue can be interpreted in the same way with regards to the learning that is occurring.

Research on media interactions has stated that people interact socially and naturally with media (to include computers) as they do humans (Reeves & Nass, 1996). The researchers suggest that people follow rules of social relationships when interacting with media and that this occurs naturally and unconsciously. For example, media has been shown to induce emotions such as frustration and politeness.

Similarly, studies examining interactions with virtual humans have shown that people react in the same manner to these entities as they do with other humans (Zanbaka, Ulinski, Goolkasian, & Hodges, 2004; Pertaub, Slater, & Barker, 2002). While being observed by a crowd of virtual agents, people showed nervousness just as they did with a human audience. Women also show social inhibition effects with virtual agents like they do with humans.

In contrast, more recent research using ITSs has shown that students do not behave the same with computers as they do with humans, as evident in their dialogue acts. When students were conversing with a computer, but believed they were conversing with a human, they used more words and conversed longer than did students who were told they were talking to a computer (Schechtman & Horowitz, 2003). In addition, students provided more explanations and longer turns when they believed they were talking to a human versus a computer, even though they were talking to a computer in both cases (Rosé & Torrey, 2005).

Therefore, results as to how people respond to computers and computer entities, in comparison to humans, are mixed. While previous ITS studies have looked at the content based dialogue (dialogue relevant to the lesson material), we took

a broader perspective and considered other dialogue categories, such as metacognition, because they have also been shown to predict learning gain (Campbell et al., 2009). We examined and compared a human-human and a human-computer tutorial dialogue corpus. We categorized five types of dialogue found in these corpora. Most of the dialogue was related to the content of the lessons. The other four categories of dialogue that were present were management (discussing the flow of the lesson), metacognition (describing one's understanding), social (chit-chat and signs of frustration), and nonsense words (random sequences of letters). For this comparison, we will focus on metacognition, social dialogue, and nonsense words because these are the categories where research hasn't yet explored and, we believe, will also differ in regards to the interactions.

Method

To explore our research questions we conducted a human-human and a human-computer study. The two corpora were then analyzed and compared in terms of their dialogue.

Human-Human Tutoring Study

Data collection environment

A curriculum incorporating lessons on basic electricity and electronics was constructed. The curriculum covered topics including open and closed paths, voltage reading between components and positive and negative terminals, series and parallel configurations, and finding faults in a circuit with a multimeter. These basic concepts were taught in a computer-based learning environment within a single session lasting approximately four hours[1].

Figure 1 shows a screenshot of the learning environment that the participants interacted with during the study. The screen was divided into three sections. The top left-hand section displayed the core lesson material in slide form, including educational text, activities, and discussion questions. The participants were able to move through the lesson slides at their own pace. The top right-hand section provided participants with a circuit simulator, which allowed them to construct and manipulate circuits as a supplement to the material in the slides. The bottom section was the chat window where the participants and tutor conversed by typing.

The tutor and student were located in the same room, but were separated by a divider. The tutor had the ability to observe the student's learning environment and interact with the student through a computer screen and chat window. The tutor gave feedback, technical assistance, and/or encouragement that he or she considered appropriate. Participants directed their answers, comments, and/or questions to the tutor throughout the curriculum.

[1] Note that there was a second session, covering additional topics, but it will not be addressed further in this paper.

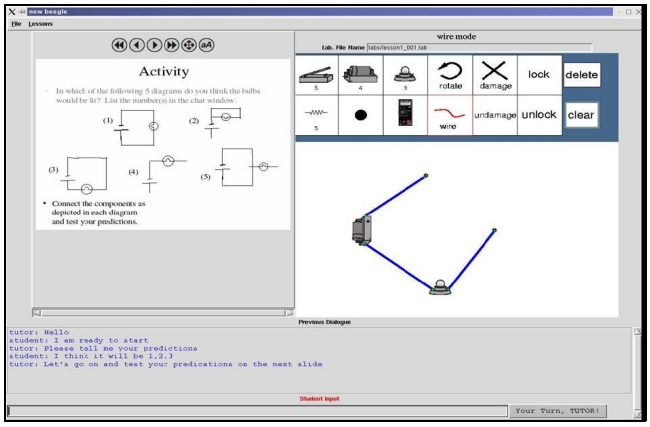


Figure 1. Participant screen for human-human tutoring.

Procedure

After completing informed consent paperwork, participants filled out a demographic questionnaire and took a pre-test consisting of 38 multiple choice questions. The participants were then introduced to their tutor and given a brief demonstration of how to operate the learning environment. The students spent the majority of the experimental session working through the lesson material and building circuits. At the conclusion of the experiment, participants completed a post-test, which included 21 multiple choice questions, and a reaction questionnaire. They were then debriefed and excused.

Corpus

The corpus of the human-human study was comprised of dialogues from each of the thirty participants distributed across three experienced tutors. The average age of the participants was 22.4 years (*SD* = 5.0) and exactly half of them were female. The corpus of this study includes 8,085 dialogue turns taken by the student and tutor and 56,133 tokens (words and punctuation).

Human-Computer Tutoring Study

Data collection environment

As much as possible, the same curriculum as the human-human study was used in the BEETLE II computer tutoring system (Dzikovska et al., 2010). Small changes were made to the curriculum so that the computer would be able to understand student responses (e.g., multi-part questions were simplified into single questions). The computer tutor (ITS) was created to implement the effective tutorial strategies used in our human-human corpus (e.g., hints). The ITS understood and responded to content (by providing feedback) and negative metacognitive statements (by giving a hint) made by a student, but not to the other types of dialogue (management, social, and nonsense). The responses and feedback given by the ITS was modeled after the human tutors from the previous corpus. The ITS used a friendly and encouraging tone similar to the human tutor. In

fact, in most cases, the ITS used identical phrasing for its comments to the student.

A screenshot of the learning environment is shown in Figure 2. The learning environment was similar to that of the human-human environment. The screen was divided into three sections. The upper left-hand section had the same function as the previous study; however the navigation buttons were slightly different. The right-hand section was the chat window where the participants and tutor interacted through typing. The lower-left section included the circuit simulator, which had the same purpose as the previous study, although the tools used to build circuits had a different display interface.

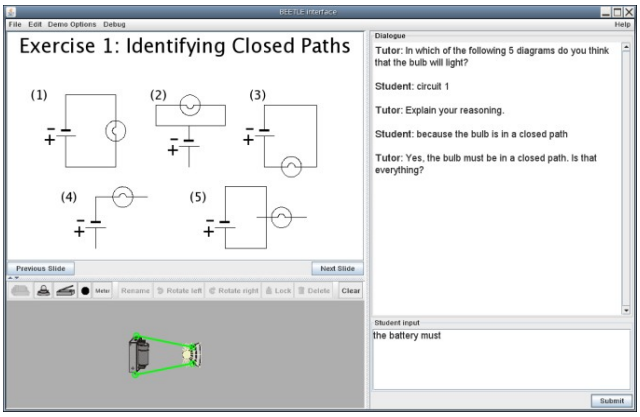


Figure 2. Participant screen for the BEETLE2 ITS.

Procedure

The procedure for the human-computer study was essentially the same as the human-human study with a few exceptions. The pre-test consisted of 22 multiple choice questions and the post-test consisted of 21 multiple choice questions. The human-computer pre-test had fewer questions because we removed questions associated with material from the second session of the human-human study, as mentioned earlier. In addition, instead of a reaction questionnaire at the conclusion of the study, participants were given a usability and satisfaction questionnaire.

Corpus

The human-computer corpus consists of dialogues from each of the forty-one participants in the study. The average age of the participants was 20.8 years (*SD* = 3.30) and there were almost twice as many females as males. The corpus includes an estimated 34,900 total dialogue turns taken by the student and tutor and an estimated 398,410 total tokens. There were many more dialogue turns and total tokens in the human-computer study because the computer asked the questions in this study (versus them being presented on the lesson slides in the previous study). In addition, more questions were presented in this study because, as stated earlier, multi-part questions were simplified into individual questions.

Coding

For the human-human data, two independent raters coded the student-tutor transcripts and were able to identify and distinguish between content, management, metacognitive, and social dialogue statements with perfect reliability (kappa = 1.00). In addition, raters were able to differentiate between positive and negative metacognitive statements made by the student with high inter-rater reliability (kappa = 0.99).

For the human-computer data, four independent raters coded the student-tutor transcripts and were able to identify and distinguish between content, management, metacognitive, social dialogue, and nonsensical statements with high reliability (kappa = 0.88). In addition, raters were able to differentiate between positive and negative metacognitive statements made by the student with high inter-rater reliability (kappa = 0.96).

A summary of the codes used in this study are presented in Table 1.

Content statements were described as comments including domain concepts that pertained to the lesson material. Answering a question fit into the content category (e.g., "The battery and the bulb in diagram 1", "1.5 volts", etc.).

Management consisted of dialogue that dealt with the flow of the lesson but does not contain information relevant to the lesson topics (e.g., "I give up", acknowledging the tutor's instructions to continue by saying, "OK", etc.).

Metacognitive statements were defined as statements that contained the student's feeling about his or her understanding, but did not include domain content. Metacognitive statements were further classified as positive or negative. Positive metacognitive statements were defined as statements that expressed understanding (e.g., "I get it", "I understand", etc.), whereas negative metacognitive statements expressed confusion (e.g., "I don't understand", "Give me a hint", etc.).

Social dialogue includes positive and negative statements. Positive social dialogue was defined as statements that included humor, rapport, chit-chat, and saving face. Examples included "Ha-ha", "Hi, how are you?", etc. Negative social statements included expressions of frustration, explicit refusals to cooperate, and even offensive statements. Examples included "Because I said so", "No", "You're stupid", expletives, etc.

Nonsense was classified as statements that were made up of random letters or numbers that are not content related (e.g., "ufghp", "3i9f", etc.). Nonsense did not occur in the human-human dialogue; therefore it was not coded in those transcripts.

Since we wish to look beyond just the content dialogue, we will focus on metacognition, social dialogue, and nonsense words in our results. Management was left out of the analyses because it was not very prevalent in the computer tutoring data and, when it was, it was ignored by the tutor. Also, it was not a relevant predictor of learning gain with the human tutor.

Table 1. Coding summary

Code	Definition	Example
Content	Statements including domain concepts that pertain to the lesson	"There is a battery and bulb in circuit 1." "1.5 volts."
Management	Dialogue that does not contain information relevant to the lesson material, but deals with the flow of the lesson	"I give up." "O.k." Acknowledging the tutor's instructions to continue
Metacognition	Statements containing the student's feelings about his or her understanding, but does not include domain concepts	Metacognitive statements can be positive or negative.
Positive	Statements that express understanding	"I get it." "I understand." "Oh, o.k."
Negative	Statements that express confusion	"I don't know." "I don't understand."
Social Dialogue	Dialogue that is not related to the content of the lessons and serves as motivation, encouragement, humor, frustration outlets, etc.	Social statements can be positive or negative.
Positive	Statements that include humor, rapport, chit-chat, or saving face	"Ha-ha" "Hi, how are you doing?"
Negative	Statements that include frustration, refusal to cooperate with the system, or offending the system	"Because I said so." "No." "You're stupid." Expletives
Nonsense	Random sequences of letters or numbers that do not pertain to the lesson material	"oidhf" "dsfafadgdfh"

Results

Learning Gain

Pre- and post-test scores were calculated in terms of percentage correct. A learning gain score was then calculated for each participant using the formula: (post-test score – pre-test score)/(1- pre-test score).

Metacognitive Statements

Students made metacognitive statements in both studies, regardless of whether the tutor was a human or a computer; however, the relative frequencies of positive and negative metacognitive statements depended upon the type of tutor. Specifically, students talking to a human tutor made significantly more positive metacognitive statements (M = 12.9, SD = 8.3) than negative metacognitive statements (M = 1.8, SD = 2.0), $t(28) = 7.16$, $p < 0.001$. Students talking to a computer tutor, on the other hand, made significantly more negative metacognitive statements (M = 3.8, SD = 5.5) than positive metacognitive statements (M = 0.2, SD = 0.5), $t(39) = -4.21$, $p < 0.001$.

The implications of the presence of metacognitive statements also varied depending upon the type of tutor. For students interacting with a human tutor, the amount of positive metacognitive dialogue, but not negative metacognitive dialogue, was significantly negatively correlated with learning gains; $r = -0.543$, $p = .002$ and $r = -0.210$, $p = 0.266$, respectively. However, for students interacting with the computer tutor, the frequency of both types of statements were negatively correlated with learning gains (positive statements: $r = -0.419$, $p = 0.006$; negative statements: $r = -0.537$, $p < .001$).

Social Statements

While students made social statements with both types of tutors, students interacting with a human tutor made exclusively positive social statements and students interacting with the computer tutor made exclusively negative social statements. On average, students interacting with a human tutor typed 37.5 positive social words to their tutor (SD = 52.3) and students interacting with the computer tutor typed 8.5 negative social words (SD = 20).

Interestingly, the amount of social dialogue with human tutors was unrelated to student learning gains, $r = -0.211$, $p = 0.262$, but the amount of social dialogue that the student produced when interacting with the computer tutor was negatively correlated with learning gains, $r = -0.372$, $p = .017$.

Nonsense

Finally, as mentioned earlier, students spontaneously exhibited a novel type of "utterance" when interacting with the computer tutor – nonsensical sequences of letters and/or numbers. On average, the students submitted nonsense to the computer tutor 1.7 times. There was quite a bit of variability across students in their likelihood of exhibiting this behavior, with a standard deviation of 5.1. This behavior was not statistically related to learning gains, $r =$

-0.073, $p = 0.651$. However, not surprisingly, the frequency of this behavior was significantly negatively correlated to the students' report of satisfaction with the computer tutor, $r = -0.33$, $p = 0.035$.

Discussion

As stated before, ITSs are often modeled after human tutors, but it is uncertain whether these interactions are similar and can be interpreted in the same manner. In fact, we found that students did not respond similarly to the computer tutor as they did with the human tutor. In both corpora, student's dialogue included metacognitive statements, but the nature of those statements was very different. With a human tutor the statements were mostly positive acknowledgements, whereas with the computer they were negative statements expressing confusion.

Social dialogue differed drastically as well. With a human the social dialogue was all positive and served the purpose of creating rapport. With the computer, the social dialogue was all negative and was concerned more with showing frustration with the system. Nonsense did not occur in the human corpus at all. This was a new category that occurred in the computer corpus only.

The human-human and human-computer dialogues also differ in their interpretations, specifically in the metacognition and social categories. In the human-human corpus, metacognition was a negative predictor of learning gain only when it consisted of positive statements. The more frequently students said things like "I get it" the worse they did. In the human-computer corpus, both types of metacognitive statements (positive and negative) were a bad sign, though they rarely gave positive metacognitive statements.

Social interactions also differ in their interpretation. With human-human tutoring, social dialogue was not related to learning gain, whereas in human-computer tutoring it was negatively correlated with learning gain. The social statements made in the ITS environment were all negative, reflecting the participant's frustration. Thus, expressing frustration through social dialogue was a good indicator that the student was struggling with the content.

These results indicate that interactions and interpretations may indeed be different between human-human and human-computer tutoring. They also suggest that perhaps human tutors are able to handle negative metacognitive statements like "I don't get it" more effectively than our computer tutor, since negative metacognition was not negatively correlated with learning gain in the human-human corpus.

Overall, it appears that politeness may be playing a role in human-human interactions, but is put aside in human-computer interactions. When conversing with another human, participants positively acknowledged what their tutor said and participated in rapport building with chit-chat. This seems to be driven by a need to be polite and courteous to the tutor, but wasn't a good indicator of what was really going on as far as learning was concerned. Based on the results, you may not be able to really trust a student who says "I understand" when they are interacting with a human because it is unclear if they really understand or if they are just being polite.

On the other hand, when interacting with a computer tutor, participants appear to be more honest in terms of their negative statements. If they show signs of confusion or frustration, they really seem to be indicating that they are struggling with the lessons. Such signs can be interpreted as more accurate indicators that additional remediation is needed. The rules of politeness are ignored and the true story seems to emerge.

From this study we found that students will not necessarily act the same with a computer tutor as they do with a human tutor. This suggests that designing an ITS to try to mimic a human tutor may not be the best strategy. The differences in interactions should also be considered. For example, positive social statements were not related to learning gains, so they do not necessarily need to be supported in an ITS; however, negative social and nonsense statements were negatively correlated with learning gains in the ITS and should be addressed. Perhaps additional help should be given or students should be offered a break when these forms of dialogue occur. All forms of metacognition impacted learning gain in the human-computer corpus, thus they should all be addressed in the ITS. Possibly giving additional remediation to students who make metacognitive statements could be helpful.

While modeling a human tutor may be a reasonable first step in the design of an ITS, the design cannot stop there. The ITS needs to be evaluated and tested with users to determine its effectiveness. Tweaks to the system should be made according to the ITS evaluation, like the ones suggested above, for each individual system and curriculum.

In this study we tried to model the human tutor as much as possible, but were limited by the current technological capabilities in computational natural language processing. Further advancements and improvements to the system's capabilities might yield different results. Additionally, these comparisons should be replicated in other domains and other curriculums to see how results compare. It would also be interesting to compare human-human and human-computer tutoring with spoken dialogue to see if the results would hold since tutoring is commonly done in spoken form.

Acknowledgments

We would like to thank our sponsors from the Office of Naval Research, Dr. Susan Chipman and Dr. Ray Perez, three former Research Associates who worked on this project, Leslie Butler, Lisa Durrance, and Cheryl Johnson, and two additional team members, Elaine Farrow and Charles Callaway for their contributions to this effort.

References

Bloom, B.S. (1984). The 2 Sigma problem: The search for methods of group interaction as effective as one-to-one tutoring. *Educational Researcher, 13*, 4-16.

Campbell, G.E., Steinhauser, N.B., Dzikovska, M.O., Moore, J.D., Callaway, C.B., & Farrow, E. (2009, July). Metacognitive awareness versus linguistic politeness:

Expressions of confusions in tutorial dialogues. Poster presented at the *31st Annual Conference of the Cognitive Science Society*. Amsterdam, Netherlands.

Dzikovska, M.O., Moore, J.D., Steinhauser, N.B., Campbell, G.E., Farrow, E., & Callaway, C.B. (2010). Beetle II: a system for tutoring and computational linguistic experimentation. *In Proceedings of ACL-2010 demo session.*

Pertaub, D., Slater, M., & Barker, C. (2002). An experiment on public speaking anxiety in response to three different types of virtual audiences. *Presence: Teleoperators and Virtual Environments*, *11(1),* 68-78.

Reeves, B. & Nass, C. (1996). *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. Cambridge, UK: Cambridge University Press.

Rosé, C.P. & Torrey, C. (2005). Interactivity and expectation: eliciting learning oriented behaviour with tutorial dialogue systems. *Proceedings of INTERACT* (pp. 323-336).

Schechtman, N. & Horowitz, L.M. (2003). Media inequality in conversation: how people behave differently when interacting with computers and people, In *Digital Sociability*, *5(1),* 281-288.

Zanbaka, C., Ulinski, A., Goolkasian, P., & Hodges, L.F. (2004). Effects of virtual human presence on task performance. *Paper presented at the International Conference on Artificial Reality & Telepresence*. Retrieved from http://www.vrsj.org/ic-at/papers/2004/S4-1.pdf .

Perceptually Grounded Word Meaning Acquisition: A Computational Model

Claudius Gläser (claudius.glaeser@honda-ri.de)
Honda Research Institute Europe
Carl-Legien-Strasse 30, 63073 Offenbach, Germany

Frank Joublin (frank.joublin@honda-ri.de)
Honda Research Institute Europe
Carl-Legien-Strasse 30, 63073 Offenbach, Germany

Abstract

We present a computational model for the incremental acquisition of word meanings. Inspired by Complementary Learning Systems theory the model comprises different components which are specifically tailored to satisfy the contradictory needs of (1) rapid memorization of word-scene associations and (2) statistical feature extraction to reveal word meanings. Both components are recurrently coupled to achieve a memory consolidation. This process reflects itself in a gradual transfer of the knowledge about a word's meaning into the extracted features. Thereby, the internal representation of a word becomes more efficient and robust. We present simulation results for a visual scene description task in which words describing the relations between objects have been trained. This includes relations in size, color, and position. The results demonstrate our model's capability to acquire word meanings from few training exemplars. We further show that the model correctly extracts word meaning-relevant features and therefore perceptually grounds the words.

Keywords: Word Learning; Computational Model; Complementary Learning Systems; Categorization

Introduction

When hearing a novel word, a language learner has to associate the word with its meaning. Establishing such word-meaning mappings is an inherently difficult task as the learner initially cannot know to what the word refers to. Quine (1960) illustrated this problem with the example of a stranger who hears a native saying "gavagai" after seeing a rabbit. How can the stranger determine the meaning of "gavagai"? It may refer to the rabbit, a part of the rabbit, its color, any fast moving animal, or even that a rabbit is tasty. This problem, usually referred to as *referential uncertainty*, cannot be solved from a single word-scene pairing. Rather the use of the word in different contexts enables the learner to extract its meaning. Nevertheless, children learn the meaning of words from few exposures to them. They rapidly construct hypotheses about word meanings, which may initially be linked to specific contexts in which the words occurred. Over time, however, children generalize among different observations, even though this may result in an overextension of a word's use (MacWhinney, 1998). This remarkable ability of children has been subject to many studies and resulted in numerous theories on early word learning.

In this paper we present a computational model for the incremental acquisition of word meanings which is inspired by the learning capabilities of children. More precisely, the system has been designed to rapidly build internal representations of words from few training samples. The thus acquired knowledge can be used to generalize to previously unseen scenes. Moreover, the framework is endowed with a learning mechanism that extracts features which are relevant to the core meaning of a word. This is done by exploiting the statistical evidence which resides from a word's use in different contexts. Our model tightly couples the rapid memorization of word-scene associations with the statistical feature extraction. This results in learning dynamics which resemble a gradual knowledge transfer and consolidation.

We will present experimental results which validate the model. Therefore, the model has been applied in a simulated visual scene description task where words for the relations between pairs of geometric objects have been trained. This includes relations in position, color, and size. The results from this experiment illustrate that our model rapidly acquires word meanings from few training exemplars and further extracts word meaning-relevant features.

The remainder of this paper is organized as follows. Next, we will review existing approaches for word meaning acquisition and relate our model to them. Afterwards, we will state contradictory needs that computational models have to satisfy. We proceed with the presentation of our computational model and subsequently show experimental results for it. Finally, we give a summary and outline our future work.

Related Work

Existing computational models address different levels of referential uncertainty. Firstly, there are approaches which consider the problem of how a learner establishes a mapping between words and a set of pre-defined meanings (e.g. Siskind, 1996; K. Smith, Smith, Blythe, & Vogt, 2006; Fontanari, Tikhanoff, Cangelosi, Ilin, & Perlovsky, 2009). In these models the first occurrence of a word typically induces multiple hypotheses about its meaning. These hypotheses become subsequently pruned either by incorporating learning constraints (Markman, 1990) or via *cross-situational learning* (L. Smith & Yu, 2008) - a technique making use of the statistical evidence across many individually ambiguous word-scene pairings. However, these models disregard the fact that learners can seldom rely on a set of pre-established concepts. Word meanings rather become flexibly constructed and shaped through language use (Boroditsky, 2001).

Therefore, a second group of models further asks how language use yields sensori-motor concepts to which words become associated (e.g. Steels & Kaplan, 2002; Skocaj et al., 2007; Kirstein, Wersing, Gross, & Körner, 2008; Wellens, Loetzsch, & Steels, 2008). In these models the learner observes the world through multiple (analog or discretized) input channels. The words finally serve as labels for categories, which become incrementally constructed on the multi-dimensional input space and gradually refined by concentrating on the most important input dimensions.

Lastly, there are models which aim at the acquisition of both phonological form and semantic form. Such models either build perceptual clusters in the acoustic space and the semantic space and subsequently associate them with each other (Yu & Ballard, 2003; Goerick et al., 2009) or clustering is directly carried out in the joint acoustic-semantic space (Roy & Pentland, 2002).

The model we present in this paper falls into the second group of methods, i.e. based on the observation of multiple word-scene pairs it acquires perceptual categories by which the words become grounded. To achieve realistic word meaning acquisition we further place several requirements on our model: (1) It should be capable of learning during online operation. Consequently, the model has to apply incremental learning techniques as training exemplars sequentially arise during a learner's interaction with its environment. (2) The model should further rapidly learn from few examples and afterwards apply the acquired knowledge to generalize to novel scenes. (3) However, to be efficient and robust the internally built categories should reflect the core structure underlying the word meanings. Thereby, we use the term core structure to refer to the essential aspects which define the meaning of a word. (4) Lastly, for systems with minimum predefined knowledge this core structure is usually hidden and thus cannot be directly accessed by concentrating on input dimensions which carry the meaning. The model rather has to extract word meaning-relevant feature dimensions in terms of a transformation from the input space.

The combination of these requirements is what distinguishes our model from existing approaches. Particularly the combination of rapid incremental learning with word meaning-relevant feature extraction has (to our best knowledge) not been realized previously. In (Skocaj et al., 2007; Wellens et al., 2008) and most notably (Kirstein et al., 2008) feature selection is applied, i.e. the learning focuses on the input dimensions which are considered to be relevant for representing the word meanings. By doing so the approaches inherently rely on the assumption that words can be grounded in a subset of the input dimensions. This in turn means that significant knowledge about the words to learn has to be put into the system by the designer. In contrast, our system generates new word meaning-relevant feature dimensions out of a set of basic input dimensions. We consider this ability to be crucial for life-long incremental learning systems for which the extent of words to be learned is unknown at design time.

Complementary Learning Systems Theory

The way how children acquire the meaning of new words is fascinating in multiple respects. When they hear a word for the first time they already get a glimpse on what it may mean. This ability may be facilitated by learning constraints or biases (Markman, 1990). It is anyway non-disputable that even the exposure to just a few uses of the word enables the child to generalize and apply the word in novel contexts. Even though generalization may occasionally result in errors (MacWhinney, 1998), over time children robustly identify the core meaning of a word.

A Computational Learning Dilemma

Modeling word meaning acquisition computationally, however, is difficult as contradictory needs have to be simultaneously satisfied. McClelland, McNaughton, and O'Reilly (1995) illustrated this fact on the example of artificial neural network models: On the one hand, the learning from few training samples requires a rapid or even one-shot memorization of the items which can be achieved by using high learning rates. This implies that localized representations, which keep the memory items separated from each other, have to be used. Otherwise, a neural network would suffer from *catastrophic forgetting* - the problem that the incorporation of new knowledge overwrites previously memorized items. On the other hand, the extraction of the core structure underlying a word meaning necessitates a statistical learning approach as knowledge has to be accumulated over many training exemplars. Such a learning can be achieved using low learning rates and overlapping representations. Artificial systems, which learn from few examples while they simultaneously extract statistical evidence, are consequently difficult to achieve.

A Solution to the Problem

Obviously, humans (and particularly children) successfully solve this learning task. Endowing artificial systems with mechanisms inspired by human learning may consequently lead a way to overcome the dilemma. *Complementary Learning Systems (CLS) Theory* (McClelland et al., 1995) suggests that the human brain makes use of separate but tightly coupled learning and memory devices which are specifically tailored to satisfy the contradictory needs. More precisely, it is proposed that new memories are first stored in the hippocampal system which is known to perform rapid learning while utilizing localized representations. The hippocampal system further allows the reactivation of recent memories during rest or sleep. This reactivation in turn enables neocortical areas to extract the core structure underlying different memories via *interleaved learning* - a technique where new items become gradually learned while learning is interleaved with the memorization of other items. Consequently, a gradual memory consolidation and transfer from the hippocampal system to neocortical sites can be observed. Furthermore, there is behavioral and neuroscientific evidence which is in accordance with a CLS theory for the lexical and semantic acquisition of novel words (Davis & Gaskell, 2009).

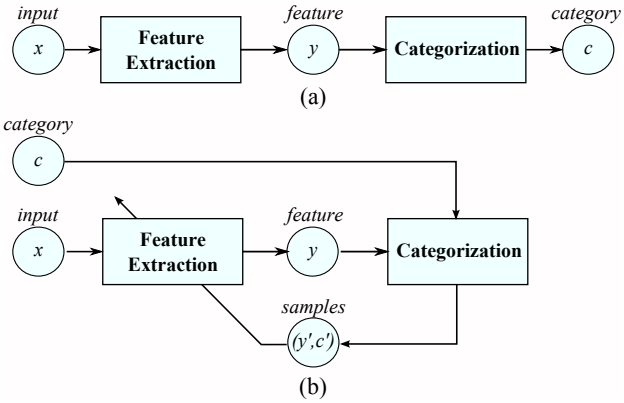


Figure 1: Architecture of the computational model: (a) Input samples x become transformed into feature patterns y which are subsequently categorized. (b) During learning the system components are recurrently coupled (see text for details).

Computational Model

In what follows we treat word learning as category learning. This is reasonable as a word refers to collections of entities which belong to the same category. Word meanings are consequently the conditions underlying category membership (Bloom, 2000). We restrict our description to the learning of one word. Multiple words can be learned straightforwardly by creating multiple instances of the system. As shown in Fig. 1, the framework consist of a feature extraction layer and a categorization layer which are recurrently coupled. The feature extraction transforms an input pattern x into a feature pattern y for which a category membership c is subsequently calculated. Here, c is a binary variable which signals whether the category's word label is appropriate for the description of the input pattern ($c=+1$) or not ($c=-1$).

Our model is largely inspired by CLS theory. Nevertheless, the model is not meant to provide a 1:1 mapping to certain brain areas. It rather resembles CLS theory from a functional perspective. For this reason, we will highlight functional correspondences of our model with different brain areas.

Feature Extraction

In the feature extraction layer word meaning-relevant features, which facilitate the subsequent categorization of a pattern, should become extracted. The learning consequently has to exploit the statistical evidence stemming from the observation of multiple word-scene pairings. Such a statistical feature extraction is obviously part of neocortical learning.

In (Hild, Erdogmus, Torkkola, & Principe, 2006) a learning technique called Maximizing Renyi's Mutual Information (MRMI) has been proposed. MRMI tries to maximize the information that the feature patterns carry about category memberships. Hence it is ideally suited to accomplish the learning task. We restrict learning to a linear feature extraction of form $y = R \cdot x$. We consequently aim at the identification of a transformation matrix R such that the mutual information $I(Y;C) = H(Y) - H(Y|C)$ between the feature patterns and

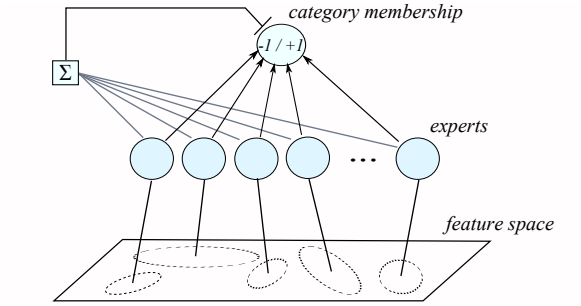


Figure 2: The architecture of an NGnet.

category labels becomes maximized. By relying on Renyi's quadratic entropy $H_2(Y)$ and its estimation using Parzen windows (Hild et al., 2006) the criterion to be maximized is

$$I(Y;C) = -\log \frac{1}{K} \sum_{k=1}^{K} G(y(k) - y(k-1), 2\sigma^2) + \sum_{j \in \{-1,+1\}} \left(\frac{K_j}{K} \log \frac{1}{K_j} \sum_{k=1}^{K_j} G(y_j(k) - y_j(k-1), 2\sigma^2) \right). \quad (1)$$

Here, $G(z, \sigma^2 I) = \exp(-\frac{1}{2} \frac{z^T z}{2\sigma^2})$ is a Gaussian kernel, $y_{+1}(k)$ and $y_{-1}(k)$ denote the k-th exemplars of feature patterns belonging to a category or not, K_{+1} and K_{-1} are the numbers of such patterns, and $K = K_{+1} + K_{-1}$. Since $y(k) = R \cdot x(k)$, we can estimate R via stochastic gradient ascent on $I(Y;C)$.

To de-correlate the feature dimensions and to perform dimensionality reduction we additionally apply Principal Component Analysis (PCA) on the extracted features. By assuming the inputs x to be white with zero mean and unit variance, the principal feature dimensions can be obtained via eigendecomposition of $R \cdot R^T$. Let Ψ be the matrix of eigenvectors whose cumulative energy content exceeds a threshold. Then we calculate feature patterns y according to

$$y = \Omega \cdot x = \Psi^T \cdot R \cdot x. \quad (2)$$

Categorization

To incrementally learn a category we use an adaptive Normalized Gaussian Network (NGnet) which we recently proposed (Gläser & Joublin, 2010). As shown in Fig. 2, the NGnet is composed of multiple locally operating experts, each of them being responsible for features stemming from its associated input region. The category membership $c \in \{-1, 1\}$ of a feature pattern y is calculated according to

$$c(y) = \text{sign} \left[\frac{1}{\sum_{j=1}^{M} \phi_j(y)} \cdot \sum_{i=1}^{M} \alpha_i \cdot \phi_i(y) \right]. \quad (3)$$

Here, M denotes the number of experts and α_i the weight of expert i to the output neuron. Furthermore, $\phi_i(y)$ is the response of the i-th expert to feature y which is described by a multivariate Gaussian of form

$$\phi_i(y) = \exp \left(-\frac{1}{2} \cdot (y - \mu_i)^T \Sigma_i^{-1} (y - \mu_i) \right), \quad (4)$$

where μ_i and Σ_i denote the center and covariance matrix of the Gaussian. The decision whether a feature pattern belongs to a category is finally obtained by application of the sign function to the continuously valued output. The network parameters are determined during online operation via Expectation-Maximization (EM) training as proposed in (Xu, 1998).

Since the NGnet statistically learns an internal category representation which associates inputs from different modalities, our categorization layer functionally resembles multimodal associative cortices, e.g. the perirhinal cortex. However, our adaptive NGnet is additionally endowed with mechanisms which allow a demand-driven allocation and removal of experts (Gläser & Joublin, 2010). This enables the network to perform a one-shot memorization of word-scene associations. Our categorization layer consequently also models the rapid initial learning as it is carried out by the hippocampus.

More precisely, our model accomplishes network growth and pruning as follows: (1) New word-scene associations become memorized based on the novelty or surprise of an input sample. Similarly, already memorized associations become (2) pruned if they became redundant, (3) split if the internal representation has to be refined, or (4) merged if they are sufficiently similar. For a detailed description of these mechanisms we refer to (Gläser & Joublin, 2010).

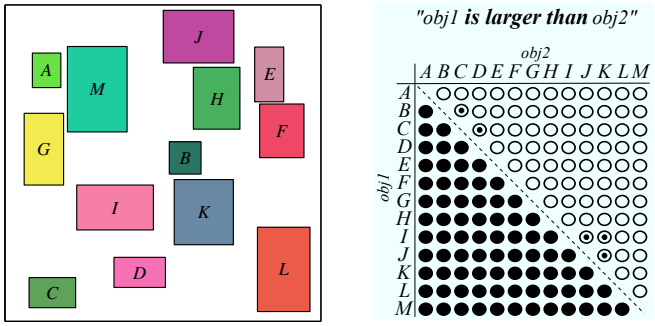
Coupling of the Components

Inspired by CLS theory we finally couple the slow statistical feature extraction and the rapid category learning. As shown by the pseudo-code in Alg. 1 the incremental learning mechanism consists of four steps which are carried out every time a new training exemplar is obtained.

Algorithm 1 Incremental Learning

```
Initialize the feature extraction to R = I, Ψ = I
Initialize an empty NGnet
for all training samples (x, c) do
    Update the NGnet with (y, c)
    Generate a set of samples (y', c') using the NGnet
    Train the feature extraction on the generated samples
    Adapt the NGnet to the changed feature space
end for
```

After updating the NGnet with a training sample, the internal representation of a category is used to reactivate memorized associations. This step resembles hippocampal dreaming. We consequently produce a set of samples (y', c') composed of feature patterns y' and associated category memberships c'. To do so, we first determine whether a local expert i represents category members ($c' = +1$) or non-members ($c' = -1$) and next randomly draw feature patterns y' from its Gaussian-shaped receptive field. Since the receptive field is described by its mean μ_i and covariance matrix Σ_i, a feature pattern y' can be generated by $y' = \mu_i + B \cdot z$, where $z \sim \mathcal{N}(0, I)$ is a random vector and B is obtained from the Cholesky decomposition $B \cdot B^T = \Sigma_i$.



(a) (b)

Figure 3: In (a) an example scene used in the visual description task is depicted. In (b) the output of the model after learning the meaning of *is larger than* is shown. Black circles correspond to category members, white circles to non-members, and dotted circles denote errors made by the system.

Afterwards, the generated samples are used to train the feature extraction. In other words, the feature extraction searches for commonalities among the reactivated patterns and tries to extract the condition which discriminates between members and non-members of the category. This learning process changes the feature space the categorization layer is operating on. For this reason, we finally adapt the NGnet to the changed feature space in an analytic way. Since we use a linear feature extraction, the change in the feature space can be expressed in terms of an affine transformation $\tilde{y} = A \cdot y$ with $A = \tilde{\Omega} \cdot \Omega^{-1}$. Here, Ω and $\tilde{\Omega}$ denote the feature extraction matrices before and after the learning. We consequently adjust a local expert's receptive field by calculating its new center $\tilde{\mu}_i$ according to $\tilde{\mu}_i = A \cdot \mu_i$ as well as its associated covariance matrix $\tilde{\Sigma}_i$ according to $\tilde{\Sigma}_i = A \cdot \Sigma_i \cdot A^T$.

Since these learning steps are carried out iteratively, knowledge about a category becomes consolidated as more training samples are processed. The knowledge, which has been first acquired in the categorization layer (via the memorization of word-scene associations), becomes gradually transfered into the extracted features. Due to the fact that the extracted features facilitate the categorization task, this knowledge transfer leads to a more robust categorization as well as a less complex NGnet needed to represent the category.

Experimental Results

We evaluated our computational model in a visual scene description task in which the meaning of words for the relations between objects has to be acquired. Thereby, a learner and a tutor observe a scene composed of geometric objects as the one shown in Fig. 3(a). The tutor selects two out of the objects and describes the relation between them, e.g. by saying

"K is larger than D."

Based on such exemplars of word use the learner has to incrementally build-up internal concepts which correspond to the words' meanings. The training of the model is illustrated in

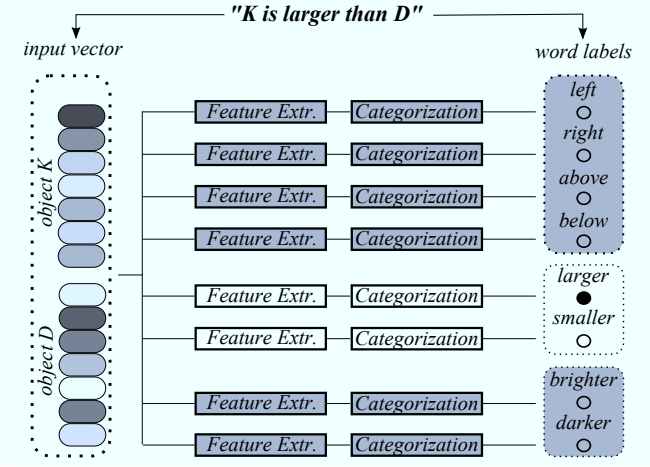


Figure 4: The training of the model in the visual scene description task is illustrated (see text for details).

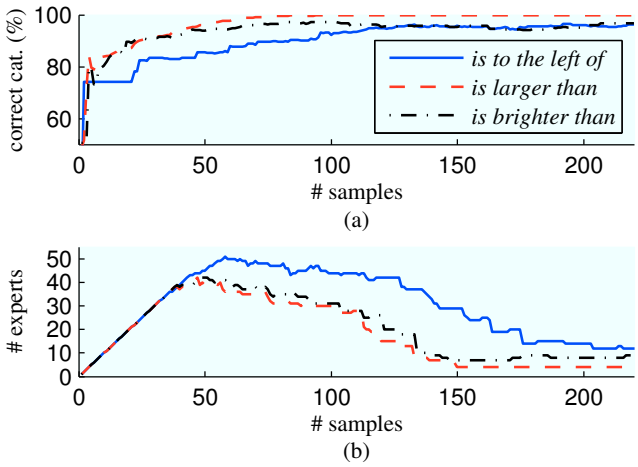


Figure 5: The evolution of (a) the correct categorization rate and (b) the complexity of the NGnet is shown for the learning of different words.

Fig. 4. For the present experiment we consider the learner to have sufficient syntactical knowledge to identify the objects of interest (e.g. *K* and *D*) as well as the word to be learned (e.g. *is larger than*). For computational purposes we further did not carry out the experiment in direct interaction with the system, but rather used simulated scenarios which provide a ground truth for performance evaluation.

Each of the objects in a scene is represented by its absolute position, its width and height, as well as its RGB color value. Consequently, tuples composed of a 14-dimensional perceptual vector (7 dimensions per object) as well as a word label served as training inputs to the system. Words for object relations concerning position (*is to the left of*, *is to the right of*, *is above*, *is below*), size (*is larger than*, *is smaller than*), and color (*is brighter than*, *is darker than*) has been trained. However, it is important to note that the system did not have prior knowledge about the relevance of input dimensions with respect to the meaning of the words. In contrast, important dimensions (e.g. the relative object positions) are even not present and have to be extracted by the system. For each of the words to learn we applied an adaptive NGnet as a binary categorization module and further extracted word meaning relevant features. To cope with missing negative training data we implemented the mutual exclusivity bias between words related to object positions, sizes, and colors, respectively. In other words, a positive training exemplar for *is larger than* has been additionally used as negative training sample for *is smaller than* (Regier, 1996).

The results of this experiment are shown in Fig. 5. In (a) we plot the system performance for the learning of individual words. The performance (correct categorization rate) has been determined on a set of scenes not included in the training data. In (b) we further plot the complexity of the individual classifiers for which the number of local experts comprising the NGnet is an indicator. To keep the plots readable, here we restrict ourselves to curves for the learning of *is to the left of*, *is larger than*, and *is brighter than*. The learning of the other words resulted in qualitatively similar curves.

From the plots we see that the system performance rapidly increases during the presentation of the first training exemplars and afterwards converges towards a near optimal level. In contrast, the complexities of the classifiers also increase at the beginning, but subsequently decrease and maintain a low level afterwards. The observed behavior of the model is in-line with CLS theory, insofar as it can be explained by two complementary learning processes which run at different timescales: (1) Initially, new knowledge is rapidly memorized. In our model this is accomplished by the on-demand allocation of local experts within the classifier. After a while, the experts adequately represent upcoming training samples such that they do not have to be memorized additionally. Consequently, the classifier complexity as well as the system performance increase at the beginning. (2) Afterwards, knowledge is gradually transferred. In our model the knowledge shifts into the iteratively extracted word meaning-relevant features. These features facilitate the classification task such that a less complex classifier can be applied. At the same time, however, the internal representation of a word meaning becomes more robust and, thus, further increases the system performance.

After training, an analysis of the extracted features revealed that the built categories solely rely on the meaning of the corresponding words. For example, for representing the meaning of *is larger than* the feature

$$(width_{obj_1} + height_{obj_1}) - (width_{obj_2} + height_{obj_2})$$

has been extracted which is an adequate linear approximation of the real decision criteria

$$(width_{obj_1} \cdot height_{obj_1}) - (width_{obj_2} \cdot height_{obj_2}) > 0.$$

Similarly, the relative horizontal and vertical positions have been extracted for the description of spatial relations. This shows that our framework is able to acquire the meaning of words and consequently grounds them.

Finally, the output of the classifiers can be used to describe a visual scene. For the scenario depicted in Fig. 3(a), we show the output of our framework concerning the judgment whether an object *is larger than* another object in Fig. 3(b). As can be seen, objects pairs are correctly categorized except for rare cases in which the object sizes are very similar.

Summary & Future Work

In this paper we presented a computational model for the incremental acquisition of word meanings. The novelty of the framework stems from its combined ability to (1) rapidly build categories which correspond to the learned words while (2) it simultaneously extracts features which underly the meaning of the words. We consider these abilities to be fundamental for life-long incremental learning systems which have to cope with minimal predefined task knowledge. To satisfy the contradictory needs of rapid learning from few examples as well as statistical feature extraction we modeled learning mechanisms which are known to be beneficial for humans. More precisely, our framework resembles CLS theory insofar as it uses separate but tightly coupled components which are specifically tailored to meet these criteria.

We evaluated our model in a visual scene description task, where words for the relations between objects have been taught. Our results demonstrate that the system acquires word meanings based on the observation of just a few word-scene pairings. It subsequently uses its knowledge to generalize to novel scenes. The results further showed that the system implements a memory consolidation process in which knowledge about a word's meaning gradually shifts from the rapidly learned category representation into the slowly extracted features. This consolidation process is beneficial as it abstracts the core meaning of a word and, hence, lets the internal representation of a word become more robust and efficient.

Part of our future work will be the extension of the model to incorporate a non-linear feature extraction. This would allow the system to extract more complex dependencies which may underly a word's meaning. Secondly, we will endow the model with a mechanism which detects the mutual exclusivity between words. This learning bias is currently pre-defined, but has to be autonomously applied by the system to enable a learning of an arbitrary set of words. Lastly, we will extend our teaching scenario to include social learning. Social learning enables an active learning by the system which is useful for testing hypotheses about a word's meaning.

References

Bloom, P. (2000). *How children learn the meaning of words*. MIT Press.

Boroditsky, L. (2001). Does language shape thought? Mandarin and English speakers' conceptions of time. *Cognitive Psychology*, *43*(1), 1–22.

Davis, M., & Gaskell, M. (2009). A complementary systems account of word learning: neural and behavioural evidence. *Phil. Trans. R. Soc. B*, *364*(1536), 3773–3800.

Fontanari, J., Tikhanoff, V., Cangelosi, A., Ilin, R., & Perlovsky, L. (2009). Cross-situational learning of object-word mapping using Neural Modeling Fields. *Neural Networks*, *22*(5-6), 579–585.

Gläser, C., & Joublin, F. (2010). An adaptive normalized gaussian network and its application to online category learning. In *Proc. IJCNN.*

Goerick, C., Schmuedderich, J., Bolder, B., Janssen, H., Gienger, M., Bendig, A., et al. (2009). Interactive online multimodal association for internal concept building in humanoids. In *Proc. IEEE-RAS Int. Conf. on Humanoids.*

Hild, K., Erdogmus, D., Torkkola, K., & Principe, J. (2006). Feature extraction using information-theoretic learning. *IEEE Trans. on Pattern Anal. and Machine Intell.*, *29*(9), 1385–1392.

Kirstein, S., Wersing, H., Gross, H. M., & Körner, E. (2008). An integrated system for incremental learning of multiple visual categories. In *Proc. ICONIP.*

MacWhinney, B. (1998). Models of the emergence of language. *Annu Rev Psychol*, *49*, 199–227.

Markman, E. M. (1990). Constraints children place on word meanings. *Cognitive Science*, *14*(1), 57–77.

McClelland, J., McNaughton, B., & O'Reilly, R. (1995). Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychol Rev*, *102*(3), 419–457.

Quine, W. V. O. (1960). *Word and object*. MIT Press.

Regier, T. (1996). *The human semantic potential: Spatial language and constrained connectionism*. MIT Press.

Roy, D., & Pentland, A. (2002). Learning words from sights and sounds: A computational model. *Cognitive Science*, *26*(1), 113-146.

Siskind, J. M. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, *61*(1-2), 39–91.

Skocaj, D., Berginc, G., Ridge, B., Vanek, O., Hutter, M., & Hawes, N. (2007). A system for continuous learning of visual concepts. In *Proc. ICVS.*

Smith, K., Smith, A. D. M., Blythe, R. A., & Vogt, P. (2006). Cross-situational learning: A mathematical approach. In *Symbol grounding and beyond.* Springer.

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*(3), 1558–1568.

Steels, L., & Kaplan, F. (2002). Aibos first words: The social learning of language and meaning. *Evolution of Communication*, *4*(1), 3–32.

Wellens, P., Loetzsch, M., & Steels, L. (2008). Flexible word meaning in embodied agents. *Connection Science*, *20*(2–3), 173–191.

Xu, L. (1998). RBF nets, mixture experts, and Bayesian Ying-Yang learning. *Neurocomputing*, *19*, 223–257.

Yu, C., & Ballard, D. H. (2003). *A computational model of embodied language learning* (Tech. Rep. No. TR791). University of Rochester.

A Motivationally Based Computational Interpretation of Social Anxiety Induced Stereotype Bias

Nicholas R. Wilson (wilson3@rpi.edu)
Ron Sun (rsun@rpi.edu)
Rensselaer Polytechnic Institute, Cognitive Science Department
110 Eighth Street, Troy, NY 12180 USA

Robert C. Mathews (psmath@lsu.edu)
Louisiana State University
210 Audubon Hall, Baton Rouge, LA 70803, USA

Abstract

Lambert et al. (2003) suggested that stereotyping could be thought of as automatic (implicit) responses that may become magnified in certain social settings through a loss of cognitive control. This type of explanation seems reasonable; however, to date, no attempts have been made to provide a more thorough, mechanistic (computational) explanation of the exact processes underlying the phenomenon. This paper proposes just such a detailed explanation using the CLARION cognitive architecture. Our CLARION-based theory takes into account motivational factors as well as the interaction between explicit and implicit processes and is used to provide a plausible interpretation of data from an identification task in Lambert et al. (2003).

Keywords: CLARION; cognitive architecture; cognitive modeling; motivation; social anxiety; stereotype.

Introduction

In line with existing studies of the effects of social anxiety on stereotype bias (e.g., Lambert et al., 2003; Payne, 2001), an explanation of such a phenomenon can be made within a computational framework, specifically the CLARION cognitive architecture (Sun, 2002, 2003). According to our interpretation, increases in anxiety related motivational drives (Sun, 2007, 2009) have a causal effect on the ability to make controlled (explicit) responses. The reduced capability can lead an individual to revert to a reliance on more "automatic" (implicit) systems.

In the remainder of this paper, we develop a motivationally based, mechanistic theory within the CLARION framework. This CLARION-based theory will then be used to simulate the Lambert et al. (2003) stereotype-inducing identification task and the simulation results will be matched to their human data. The next two sections will examine the task as well as the empirical findings from Lambert et al. (2003). The section following these will present the CLARION-based theory for capturing the phenomenon. The section after that will examine the simulation results and compare them to the human data. The final section will discuss how our theory relates broadly to the phenomenon of cognitive distracters and their impact on cognitive control.

Lambert et al.'s (2003) Experiment

Participants were instructed that they were to identify target objects being presented on a screen as belonging to either the "tool" category or the "gun" category. They were also told that the task required both speed and accuracy. Participants then completed a 48-trial "practice" phase allowing them to become familiar with the requirements of the experiment as well as the target objects (i.e., the tools and guns). After completing the practice trials, participants were told either that all of their responses would be kept confidential (i.e., they were in the private group) or that they would be asked to share and discuss their responses with the other participants in the testing room (i.e., they belonged to the anticipated public group).

For the test phase, an additional element was added to the identification task: the prime, a picture of a person's face, was presented briefly (for 200 ms.) before being replaced by the target object (which was presented for 100 ms.). Participants were given a total of 550 ms. to make a response (by pressing a button associated with the target's category).

Participants completed a total of three blocks of trials. In each block, each of the eight primes (4 black, 4 white) was randomly paired with each of the eight targets (4 tools, 4 guns) twice. This yielded 128 trials per block, and a total of 384 trials overall.

After completing the identification task, participants also completed a measure of social anxiety and the Dunton and Fazio (1997) Motivation to Control Prejudicial Reactions Scale. These scales attempted to measure the individual differences in social anxiety and motivations to control prejudicial reactions.

Experimental Results and Discussion

The results from Lambert et al. (2003) showed that participants tended to make stereotypic errors (i.e. misclassifying a tool as a gun when primed with a black face or a gun as a tool when primed with a white face) on tool trials regardless of context ($F = 20.03$, $p < .001$ for the anticipated public group, $F = 3.74$, $p = .058$ for the private group). In other words, when the results were collapsed over context, people who were presented with a tool were significantly more likely to mistake it for a gun when it was coupled with a black prime ($M = .24$) than a white prime ($M = .22$). In

addition, people who were presented with a gun were significantly less likely to mistake it for a tool if it was coupled with a black (M = .19) rather than a white (M = .21) face. This finding was evidenced by a significant Prime X Object interaction ($F = 22.13$, $p < .001$).

The results further indicated that people in anxiety-inducing situations (e.g. the anticipated public group) made significantly more stereotypic errors than those people who were not distracted by an anxiety-inducing context (e.g., the private group). This was confirmed by a significantly stronger Prime X Object interaction in the anticipated public condition compared to the private condition ($F = 20.03$, $p < .001$ vs. $F = 3.74$, $p = .058$, as mentioned before).

Further, the presence of the black prime had an enhanced effect on participants' responses than the white prime. In other words, on black prime trials, participants were significantly more inclined to make stereotypic errors ($F = 11.52$, $p < .001$ for the main effect of object). This tendency was not evidenced when primed with a white face ($p > .20$).

Based on the process dissociation procedure (Jacoby, 1991), it was found that participants in the private group had higher estimates of cognitive control (.60) than participants in the anticipated public group (.53). These numbers were essentially the same regardless of prime as confirmed by a Prime x Context ANOVA, which revealed a significant main effect for context ($F = 4.54$, $p < .05$), no significant effect of prime ($F = .67$, $p > .05$), and no evidence of a significant Prime x Context interaction ($F = .01$, $p > .05$).

Additionally, Lambert et al. (2003) hypothesized that accessibility bias (i.e., the likelihood of making a stereotyped response when control failed) was a separate (dissociated) process from cognitive control. The results on accessibility bias estimates showed that when participants were primed with a black face, estimates were significantly higher (.56) than when they were primed with a white face (≈ .50). To confirm this, a Prime x Context ANOVA was performed revealing a significant interaction ($F = 20.39$, $p < .001$). Beyond this, no other significant effects emerged from these analyses. Of particular importance, accessibility bias was not affected by manipulating context ($F < 1.00$, $p > .05$).

Lambert et al. (2003) also posited that accessibility bias estimates could be used to roughly capture individual variation in stereotypic associations about blacks (i.e., how strongly a person associates guns with this group). Taking into account that control is particularly low for high-anxiety participants in the anticipated public group, Lambert et al. (2003) predicted that, for the aforementioned group, a correlation exists between estimates of accessibility bias and performance. To test this, they constructed an overall index of stereotypic errors: Higher error indices indicated a greater propensity toward making stereotypic errors over counter-stereotypic errors when presented with a black prime.

A few important points resulted from that analysis. First, in the private group context, the relationship between accessibility bias and gun responses was moderate and about the same regardless of anxiety. However, the relationship was especially strong in the anticipated public group, but this was only among participants who were high in state anxiety. Those participants with higher accessibility bias scores and high anxiety made more stereotyped errors on black primed trials, whereas participants with lower accessibility bias scores made less stereotyped errors on those same trials.

Of additional pertinence to the present work is the effect that context had on reported levels of state anxiety. Recall that at the end of the experiment, participants completed a questionnaire aimed at measuring a person's reported level of anxiety. Analysis of the anxiety measure indicated that, consistent with expectations, participants reported significantly higher levels of (task-specific, i.e., state) anxiety in the anticipated public (M = 1.89) compared with the private condition (M = 1.32) [$F = 10.03$, $p < .01$].

A CLARION-based Theory

CLARION is a well-established cognitive architecture (Sun, 2002, 2003; Sun et al., 2005). It consists of a number of subsystems. The following three subsystems were used for simulating the task in Lambert et al. (2003): the action-centered subsystem (ACS), the motivational subsystem (MS), and the meta-cognitive subsystem (MCS). Each subsystem is divided into two levels of representation: the explicit (top) and implicit (bottom) levels (see Reber, 1989; Sun, 2002 for justifications).

One of the fundamental theoretical assumptions in CLARION is the distinction between implicit and explicit processing. What we term explicit processing is also known as ''controlled'' processing (Lambert et al., 2003). Explicit processes are often rule-based, require more time to obtain results, and sometimes require more than one step to reach a conclusion (Sun, 2002). Similarly, implicit processes are often referred to as ''automatic'' processes. Further, when researchers refer to ''a loss in cognitive control'', what they are referring to, in CLARION terms, is an inability to adequately rely on explicit processes over (or in addition to) implicit processes. A loss of cognitive control, therefore, is equivalent to using more implicit processes.

Moving now to the representations within the two levels, in the bottom level, CLARION takes note of the fact that the inaccessible nature of implicit knowledge is best captured by subsymbolic, distributed representations (such as in a backpropagation network). It has been extensively argued that the characteristics of distributed representations accord well with the relative inaccessibility of implicit knowledge (Sun, 2002). In contrast, explicit knowledge can be best captured in computational modeling by symbolic or localist representations (Sun, 2002; Sun et al., 2005), in which each unit is more easily interpretable and has a clearer conceptual meaning. This characteristic of symbolic or localist representations captures the characteristic of explicit knowledge being more accessible (Sun, 2002). Accessibility here refers to the direct and immediate availability of mental content for the major operations that are responsible for, or concomitant with, consciousness, such as introspection, forming higher-order thoughts, and verbal reporting, as well as meta-level control and manipulation.

The dichotomous difference in the representations of the two different types of knowledge led to a two-level architecture, whereby each level uses one kind of representation and captures one corresponding type of process (this paper focuses specifically on the interaction between implicit and explicit processing within the action-centered subsystem).

The Action-Centered Subsystem (ACS)

The Action-Centered Subsystem (ACS) consists of implicit processing (in the bottom level of the two-level structure, in the form of a backpropagation network) and explicit processing (in the top level, through explicit action rules; Sun, 2002). When both implicit and explicit knowledge is available in the ACS for determining appropriate actions, the two types of knowledge are ''integrated'', for example, through stochastic selection of one type or the other. For further details related to the ACS, see Sun (2002, 2003).

For our simulation, the ACS was responsible for generating responses to a set of featurized inputs (created based on the actual pictures from Payne, 2001, to make the inputs as accurate as possible; see table 1).

The bottom level of the ACS took the featurized descriptions of a prime and target as inputs and output the specification of whether the target item was a tool or gun. The backpropagation network had 25 input nodes (13 describing a person in 6 dimensions, 12 describing an object in 5 dimensions; see table 1), 10 hidden nodes, 2 output nodes (the classification of tool or gun), and the default parameter settings (Sun, 2003). Also, since this task required quick responding, it should be especially prone to noise. We captured this effect by setting the temperature (to .4) involved in stochastic selection of the output.

The bottom level was trained to focus on skin color, because it represents the stereotyping in its simplest form. According to Payne (2001), the primes were designed to filter the characteristics of the faces until race was the only distinguishing feature. We also chose to exclude specific target characteristics during training, because we felt that the link between race and guns was likely a connection between skin color and the concept of a gun (which is the output of the ACS), not any particular gun or tool feature.

Table 1. Featurized inputs as dimension/value pairs.

Primes (people)		**Targets (guns/tools)**	
Dim.	**Val.**	**Dim.**	**Val.**
Skin Color	Black, White, Gray	Handle Color	Black, White, Gray
Nose Shape	Thin, Wide	Shape	Bent, Straight
Nose Length	Short, Long	Handle Length	Long, Short
Eyebrow Shape	Thick, Thin	Head Length	Long, Short
Eye Size	Big, Small		
Sex	Male, Female	Head Color	Black, White, Gray

Furthermore, we posit that stereotype bias is developed slowly through subtle, cumulative experiences within a society. These biases have evolved from a fundamental need to easily "classify" other members of society for the purpose of ensuring survival. It has been argued that, in general, people have developed "classification" systems to provide help in making reasonable responses quickly to unexpected or unclear circumstances (Sun, 2002). People are not necessarily cognizant of these response mechanisms. In fact, research suggests that tasks requiring quick reactions are often performed implicitly (Reber, 1989; Sun, 2002; Sun et al., 2005). Taking these arguments together, we feel that it is reasonable to think of stereotyping as a form of "classification" that is often best explained as an implicit process.[1]

The bottom level was given 500,000 training trials presenting the black and white characteristic in such a fashion that was consistent with the accessibility bias estimates from Lambert et al. (2003). The accessibility bias estimate is the probability that a stereotyped response will be made if control fails, and in our simulation control failing means that only the bottom level of the ACS is used. Hence, it seemed appropriate to use this measure to help guide the training. On about 56% (plus or minus 3.5% for individual differences) of trials where a black face was presented to the network, it was coupled with a gun (on about 44% it was coupled with a tool). Tools and guns were paired at an equal rate (plus or minus 3.5%) when coupled with a white face.

The top level of the ACS learned appropriate response rules mapping inputs concerning specific tool/gun characteristics to the proper tool/gun classification output. The assumption is that these rules represent explicit knowledge learned during the 48 practice trials as well as prior experiences by the human participants.

The Motivational Subsystem (MS)

In addition to the ACS, the motivational subsystem (MS) is another major component in CLARION. The MS is responsible for motivational states (comprised of ''drives'' and ''goals''; Sun, 2007, 2009). In CLARION, drives are fundamental motivational forces behind decision-making (as well as other processes). Anxiety can be thought of as the biological/physiological consequence of heightened (avoidance-oriented) drive strengths (see the discussion of drives in Sun, 2009). Thus, in the simulation, an agent's drive strengths are set in the MS based on the experimental contexts (e.g., the existence of an anxiety-inducing situation).

Considering the specific aspects of this task, it was determined that a single drive, "honor" (i.e., obeying social norms and codes), best encapsulated the motivating factors involved with the contexts (groups). Based on an agent's context, its "honor" drive strength level was set in the MS.

The drive strength was obtained using a backpropagation network with 2 input nodes, 4 hidden nodes, 1 output node, and the default parameter settings (Sun, 2003). The first input specified the context (group) to which the agent be-

[1] Note that our interpretation is in line with the arguments made by Lambert et al. (2003).

longed. The second input represented the agent's predisposition toward anxiety in social settings. While more generalized drive-strength equations exist, for the purposes of this simulation, it was determined that a hyperbolic tangent function provided a reasonable approximation for translating "stimulus" (i.e. context) and "deficit" (i.e. the individual predisposition toward anxiety) into a drive strength.

Making the drive sensitive to both the context as well as the predisposition to anxiety is justified by analysis performed by Lambert et al (2003), which found the existence of a significant Context x Anxiety interaction using a hierarchical regression analysis.

Further analysis of the data of Lambert et al. showed that, among participants above the group median in state anxiety, there was a significant effect of context on estimates of control ($\beta = .25$, $p < .05$), reflecting lower control in the anticipated public context compared with the private context (Ms = .51 vs. .60, respectively). However, context had no significant effect on control for the participants reporting low levels of anxiety ($\beta = .08$, $p = .52$), reflecting the fact that control was relatively high and about equal across the anticipated public and private contexts (Ms = .57 vs. .60). This effect led to the two different values used for the parameter of the hyperbolic tangent curve for the drive strength in the MS. As a result of the two different parameter values, an agent's drive strength increased more rapidly and reached a higher level in the public group than in the private group. Figure 1 gives a graphical representation of the drive.

The Meta-Cognitive Subsystem (MCS)

Finally, in conjunction with the MS, the meta-cognitive subsystem (MCS) may be used for setting parameters in the ACS. The MCS performs a number of backend actions (including the setting of parameters for action selection, reasoning, and learning, etc.) based on drive states and so on (see Sun, 2007, 2009). In the simulation, (avoidance-oriented) drive strengths (levels of anxiety) from the MS are used as the basis by the MCS to determine the likelihood of making decisions in a more or less explicit (i.e., controlled) way by the ACS.

The MCS contains a module for determining the mode of action decision making (i.e., the proportion of implicit vs. explicit processing in the ACS). A backpropogation network with 1 input node, 4 hidden nodes, 1 output node, and the default parameter settings (Sun, 2003) was used. The network was used to produce outputs based on an inverted U curve (see Yerkes & Dobson, 1908) that mapped drive strengths (the input) to the probability of being explicit (i.e., using the top level of the ACS) during action decision making (see figure 2). The working hypothesis in this regard is that when anxiety is at a relatively low level, it has little (or possibly even a positive) effect on the ability to be controlled (explicit) in making action decisions. However, when anxiety reaches a certain higher level, it can begin impairing control, creating a need to revert to faster, more automatic, implicit processes (Sun, 2007, 2009; Wilson et al., 2009; Yerkes & Dobson, 1908)

Simulation Results

In exact correspondence with experiment 2 of Lambert et al. (2003), simulated agents were placed in either a simulated private group or a simulated anticipated public group. Like the human experiment, the test phase was run using 384 trials where each face/tool pairing was observed six times at intervals of 2 times per 128 trials. A total of 128 agents were used (as opposed to 127 human participants in Lambert et al., 2003) and 64 agents were placed into each group.

The results of the simulation were recorded as error rates for the four different possible pairings of prime and target. Consistent with the findings from Lambert et al. (2003), agents in the simulated private group made significantly fewer errors on gun trials than on tool trials when paired with a black prime (.174 vs. .224) [$F = 42.62$, $p < .001$]. Additionally, on trials containing a white prime, in the simulated private group, error rates on gun and tool trials were essentially the same (.202 vs. .199) [$F = .17$, $p > .05$]. In the simulated public group, when a black prime was paired with a gun, error rates were significantly lower than when paired with a tool (.214 vs. .27) [$F = 45.37$, $p < .001$]. Also, when a white prime was paired with either a gun or a tool, error rates were not significantly different (.244 vs. .238) [$F = .491$, $p > .05$] for the simulated public group. These findings were consistent with Lambert et al. (2003).

Further analysis of the simulation data revealed a signifi-

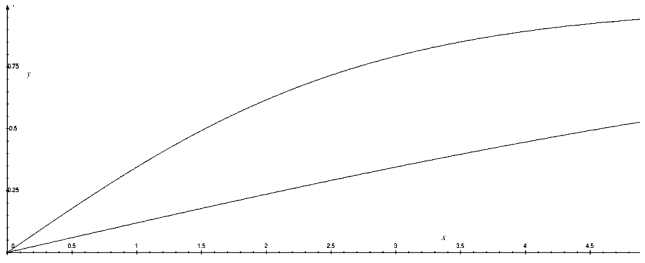


Figure 1. Graphical representation of the "honor" drive. The x-axis represents the predisposition toward anxiety ($0 \leq x \leq 5$); the y-axis represents drive strength ($0 \leq y \leq 1$). The bottom curve represents the private group [$y = tanh(.12x)$]; the top curve represents the anticipated public group [$y = tanh(.36x)$].

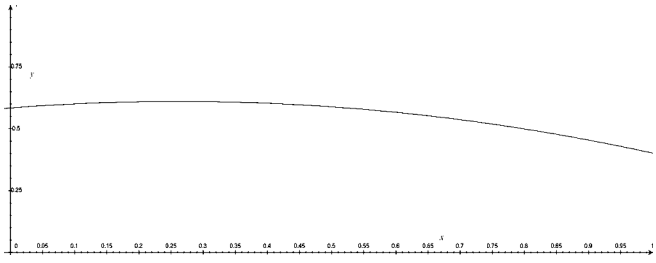


Figure 2. Inverted U-Curve. The x-axis represents the drive strength ($0 \leq x \leq 1$); the y-axis represents the level of cognitive control ($0 \leq y \leq 1$) [$y = -.38x^2 + .2x + .58$].

cant Prime X Object interaction ($F = 48.4$, $p < .001$). Collapsed over situational context, agents were significantly more likely to mistakenly identify a tool as a gun if they were primed with a black face ($M = .247$) than a white face ($M = .219$) [$F = 30.991$, $p < .001$]. Conversely, agents were significantly less likely to mistakenly identify a gun as a tool if they were primed with a black face ($M = .194$) than a white face ($M = .223$) [$F = 26.546$, $p < .001$]. Looking at it another way, agents showed a significantly stronger tendency toward mistaking a tool for a gun when primed with a black face, as opposed to mistaking a gun for a tool, when primed with a black face ($F = 88.42$, $p < .001$ for the main effect of object). When agents were primed with white faces, error rates did not vary significantly across object types ($F = .649$, $p > .05$). These findings were, again, consistent with Lambert et al. (2003).

Moreover, agents in the simulated public group made significantly more errors in general than agents in the simulated private group. This was confirmed statistically by comparing the mean error rates between the simulated public group ($M = .24$) and the simulated private group ($M = .20$) [$F = 56.64$, $p < .001$ for the main effect of context]. In a related statistic, the Object X Prime interaction was stronger in the simulated public group ($F = 28.01$, $p < .001$) compared with the simulated private group ($F = 22.26$, $p < .001$). Figure 3 graphically illustrates the above pattern of data and gives a comparison to Lambert et al. (2003).

Turning to analyses based on process dissociation, inferences into some of the mechanisms within CLARION can be made. First, the cognitive control estimate (Lambert et al., 2003) can be thought of as the probability that a person will be able to use their explicit processes (the top level of the ACS) when making a response (Sun et al., 2005). Second, the accessibility bias estimate (Lambert et al., 2003) can be thought of as the probability of making a gun response when cognitive control fails. According to our interpretation, a failure of control is tantamount to using implicit processing (see Sun, 2002; Sun et al., 2005).

Given this interpretation, there were two methods to report the cognitive control estimate from the simulation: by looking at the probability of using the top level of the ACS (as determined by the MCS), and by the process dissociation procedure (Jacoby, 1991; Lambert et al., 2003). Table 2 shows the MCS determined levels of cognitive control, the cognitive control estimates calculated using process dissociation, as well as the cognitive control estimates reported by Lambert et al. (2003). The cognitive control estimates from the simulation clearly correspond to Lambert et al.'s findings. A Prime X Context ANOVA on cognitive control

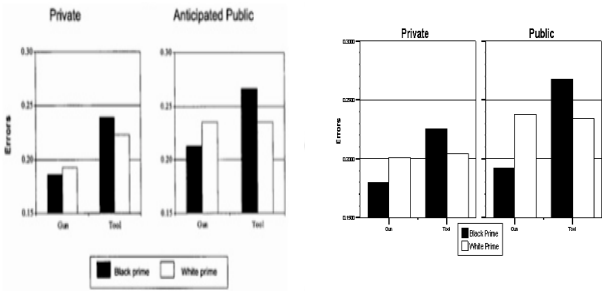


Figure 3. The graph on the left shows the human data from Lambert et al. (2003). The graph on the right is the simulation results.

estimates, calculated using the process dissociation equation (Lambert et. al., 2003), from the simulation data revealed the expected significant main effect for context ($F = 56.635$, $p < .001$), no significant effect for prime ($F = .861$, $p > .05$), and no significant Prime X Context interaction ($F = .683$, $p > .05$). This analysis provides support to the notion that cognitive control estimates are affected by context but not by prime.

Additionally, as per our interpretation, two methods for reporting accessibility bias estimates from simulation existed as well: process dissociation and actual levels of accessibility bias that were calculated by simply keeping track of the number of times the bottom level chose a gun classification when the bottom level was used. Table 3 shows the actual accessibility bias, the accessibility bias estimates calculated using process dissociation, as well as the accessibility bias estimates from Lambert et al. (2003). As expected, the accessibility bias estimates from the simulation, calculated using the process dissociation equation (Lambert et. al., 2003), were significantly higher for a black prime than a white prime and did not vary significantly by context. A Prime X Context ANOVA on accessibility bias estimates confirmed a significant main effect of prime ($F = 37.92$, $p < .001$), no significant effect of context ($F = .039$, $p > .05$), and no significant interaction ($F = .179$, $p > .05$).

Finally, a comparison between a standardized error index, which measured the agent's tendency toward making stereotypic vs. counter-stereotypic errors and the accessibility bias estimates, was calculated. Consistent with the findings from Lambert et al. (2003), the relationship between accessibility bias estimates and gun responses, as specified by the standardized error index, was moderate in the simulated private group, regardless of anxiety. However, this relationship became stronger in the simulated public group, but only when anxiety was high. A graphical representation of this analy-

Table 2. Cognitive control estimates.

Lambert et al. (2003)			**Simulation**			
Group	Black Prime	White Prime	Group	MCS	Black Prime	White Prime
Private	.61	.60	Private	.599	.602	.598
Public	.53	.53	Public	.528	.518	.517

Table 3. Accessibility Bias Estimates.

Lambert et al. (2003)			**Simulation**				
Group	Black Prime	White Prime	Group	ACS Black	ACS White	Black Prime	White Prime
Private	.56	.53	Private	.57	.508	.565	.504
Public	.56	.49	Public	.562	.505	.559	.506

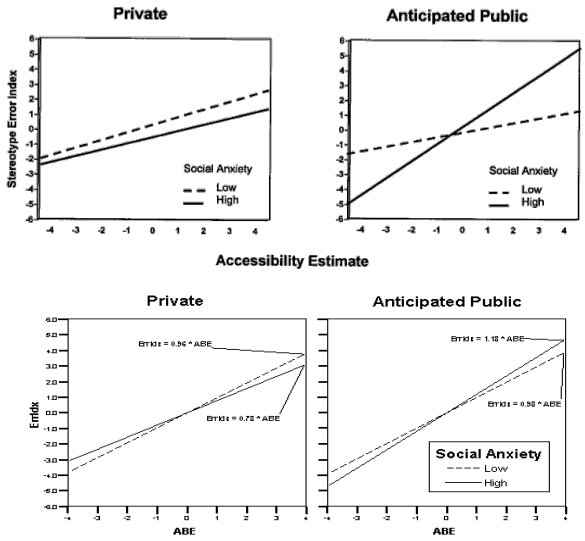


Figure 4. The top graph is the human data from Lambert et al. (2003). The bottom graph is the simulation results.

sis, along with a comparison to the findings from Lambert et al. (2003), can be seen in figure 4. Our finding of the correlation between accessibility bias estimates and error rates, as specified by the standardized error index, lend further support to the implicit nature of stereotyping. In addition, similar to the findings by Lambert et al. (2003), the connection between state anxiety and an agent's ability to make controlled (i.e., explicit) responses is characterized by the lack of a strong correlation between accessibility bias estimates and gun responses in both groups when agents were not highly effected by the anxiety-inducing cues. In other words, agents with lower levels of anxiety made more controlled responses and therefore had less chances of making stereotyped (implicit) responses.

General Discussion and Conclusion

Our CLARION-based theory appears to be capable of modeling the cognitive processes associated with the induction of stereotype biases in a social anxiety context, as illustrated by the successful simulation of Lambert et al. (2003). Moreover, our model captures the essence of the analysis of the empirical data by Lambert et al. (2003) (in a manor consistent with their interpretations).

Of related interest, our simulation supports the argument that stereotyping can be seen as mostly being an automatic (i.e., implicit) response that likely manifests itself as a result of a lessening in the ability to use more controlled (i.e., explicit) processes, as opposed to a strengthening of stereotyping habits (see Lambert et al., 2003 for further details related to this argument).

In conclusion, this article has laid out preliminary foundations that can later be applied to developing a more detailed theory of the mechanistic processes underlying the effects that anxiety and other cognitive distracters, in general, have on the control of cognition. Our theory suggests that the broader phenomenon (i.e., the effects that cognitive distracters have on performance in a variety of contexts) is explainable in a quantitative, process-based way. In this regard, CLARION provides a useful framework, which has been derived from our prior studies and simulations of human experimental data (e.g., Sun et al., 2005; Sun, 2002; Wilson et al., 2009). Our ability to explore such tasks in a more detailed, more unified fashion should be useful in better understanding the interaction between motivation, metacognition, and implicit and explicit performance.

Acknowledgments

This work is supported (in part) by the ARI contract W74V8H-05-K-0002 and the ONR grant N00014-08-1-0068. We would like to thank Sebastien Helie for his help.

References

Dunton, B. C., Fazio, R. H. (1997). An individual difference measure of motivation to control prejudiced reactions. *Personality & Social Psychology Bulletin*, 23, 316–326.

Jacoby, L.L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory & Language*, 30, 513-541.

Lambert, A., Payne, B., Jacoby, L., Shaffer, L., Chasteen, A., Khan, S. (2003). Stereotypes as dominant responses: On the "social facilitation" of prejudice in anticipated public contexts. *Journal of Personality & Social Psychology*, 84, 277-295.

Payne, K. (2001). Prejudice and perception: The role of automatic and controlled processes in misperceiving a weapon. *Journal of Personality & Social Psychology*, 81, 181-192.

Reber, A. (1989). Implicit learning and tacit knowledge. *Journal of Experimental Psychology: General*, 118, 219-235.

Sun, R. (2002). *Duality of the Mind*. Mahwah, NJ: Lawrence Erlbaum Associates.

Sun, R. (2003). *A Tutorial on CLARION 5.0.* Technical Report, Cognitive Science Department, Rensselaer Polytechnic Institute.

Sun, R. (2007). Motivation and metacognitive control of CLARION. In: W. Gray (ed.). *Modeling Integrated Cognitive Systems*. New York, NY: Oxford University Press.

Sun, R. (2009). Motivational representations within a computational cognitive architecture. *Cognitive Computation*, 1, 91-103.

Sun, R., Slusarz, P., Terry, C. (2005). The interaction of the explicit and the implicit in skill learning: A dual-process approach. *Psychological Review*, 112, 159-192.

Wilson, N., Sun, R., Mathews, R. (2009). A Motivationally-based Simulation of Performance Degradation Under Pressure. *Neural Networks*, 22, 502-508.

Yerkes, R.M., Dodson, J.D. (1908). The relation of strength of stimulus to rapidity of habit-formation. *Journal of Comparative Neurology & Psychology*, 18, 459-482.

Cultural Network Analysis: Mapping Cultural Theories of Mind

Winston R. Sieck (wsieck@ara.com)
Culture & Cognition Group, Applied Research Associates
1750 Commerce Center Blvd, Fairborn, OH 45324 USA

Louise J. Rasmussen (lrasmussen@ara.com)
Culture & Cognition Group, Applied Research Associates
1750 Commerce Center Blvd, Fairborn, OH 45324 USA

Abstract

People's ability to interact with members of other cultures is determined, in part, by their understanding of the folk psychological theories that explain behavior in those cultures. A comprehensive methodology is offered here for investigating such folk theories. It attempts to characterize the distribution of mental models within a cultural group. A network representation is used to depict the consensus elements (and level of consensus) in a cultural group's knowledge within a domain. The method is general with respect to knowledge domain, though the emphasis here is on folk theories of mind. The methodology is illustrated with two studies directed at Afghan explanations of an Afghan Mullah's decision making in a well-defined context.

Keywords: Cultural epidemiology, cultural modeling, folk psychology, mental models, decision making, mixture modeling.

Background

The epidemiological conception of culture has been gaining fairly wide acceptance among culture and cognition theorists. "Epidemiology" is used in the general sense of describing and explaining the distributions of any property within a population, and "cultural epidemiology" emphasizes explanations of the distribution of ideas within specific populations. According to Sperber (1996), culture is made up of contagious ideas, that is, the ideas that propogate effectively and durably in a population. One line of cultural research within this program seeks to directly explain how some ideas become widely distributed and long-lasting within a population (Berger & Heath, 2005). Other research has focused more on understanding the origins and distributions of complex concepts ("folk theories" or "mental models") that include systematic causal, explanatory beliefs (Atran, Medin, & Ross, 2005; D'Andrade, 1995; Hirschfeld & Gelman, 1994).

With respect to this latter area, cognitive psychology has long characterized the organization of knowledge in terms of networks of interconnected ideas, or mental models (Gentner & Stevens, 1983). Scientific investigations of mental models have spanned physical, biological, and psychological systems. Our research program focuses on understanding decision making within intercultural encounters. In particular, we have been investigating the relationship between how people represent the minds of culturally-different others and the decisions they make as regards those others (cf. D'Andrade, 1987). From this standpoint, mental models that pertain to psychological and social domains are especially useful.

Research in the psychological domain has sought to establish that people have theories about the workings of other people's minds and behavior, as well as to elucidate the abstract and general properties of theories of mind (Gopnik, & Wellman 1994). There is a general consensus among theory of mind researchers that basic awareness of mental constructs and their relationship with behavior is universal (Wellman, 1998). This, however, does not entail that higher level folk psychology interpretations of these basic theory of mind constructs are culturally universal. In fact, evidence suggests that they are not (Lillard, 1998). For example, in some cultures, "because one wants to" will be considered an important reason for any action. In contrast, within another culture, the most important reason for doing anything might be "because it is prescribed". In such a culture, actions follow a set of rules—often referred to as traditions or rituals. For instance, if upon encountering a Mursi woman from Ethiopia, it would be mistaken to think that she wears a lip plate because 'she wants to.' Now, that does not necessarily mean that she wears it unwillingly—"will" simply has little to do with it.

Recent research in cognitive science on the neural foundations of social learning also provides evidence to support the existence of differences in theory of mind constructs across cultures. For example, Leslie, Friedman, and German (2004) differentiate between the mechanism underlying theory of mind reasoning and the content of the reasoning process. Their research suggests that, along with very basic 'theory of mind' concepts, modular processes that promote attention to mental states and facilitate learning about them appear very early and develop rapidly. On the other hand, the heuristic processes that select appropriate contents for mental states develop over a longer timeframe and undergo several major changes. This timeframe in which the contents of mental states are selected suggests an important window of opportunity for the introduction of cultural variations in mental models of psychology. As children grow up in different cultures with different practices, different languages, and different external circumstances, they would correspondingly also generate different ideas about the mind to fit those experiences.

In summary, there are several reasons to believe that theory of mind functionality is supported by mental models of psychology that vary across cultures, in much the same

way that folk theories of biology differ between cultural groups. The direct investigation of such cultural theories of mind is useful for theoretical development, as it provides a comprehensive base of kinds of explanations, as well as correlations among explanatory elements. It also has considerable practical potential, primarily for the purposes of enhancing understanding of a crucial component of thought between culturally distinct groups. In order to support such investigations, we provide a method for directly eliciting, analyzing, and representing cultural models in any domain, and illustrate its use for mapping cultural theories of mind.

Cultural Network Analysis

We describe a comprehensive method for modeling culture as networks of ideas that are distributed among members of a population (Sieck, Rasmussen, & Smart, 2010). The method, Cultural Network Analysis (CNA), represents an interdisciplinary synthesis of techniques drawn from the fields of cognitive anthropology, cultural and cognitive psychology, and decision analysis. CNA is used to develop *cultural models* for groups and populations. The development of cultural models reflects a well-established practice in cognitive anthropology (D'Andrade, 2005; Quinn, 2005). Furthermore, CNA follows a similar overall pattern of research to other approaches for building cultural models, beginning with qualitative studies, followed by quantitative study and analysis. One refinement to the customary anthropological practice includes a common format for representing cultural models. Specifically, CNA cultural models are typically depicted as a network representation of the culturally-shared concepts, causal beliefs, and values that influence key decisions (Sieck, Rasmussen, & Smart, 2010). This and other refinements of the process are discussed more extensively by Sieck (2010). Here, we focus on providing a general overview of the process.

As mentioned, Cultural Network Analysis encompasses both qualitative, exploratory analysis, and quantitative, confirmatory analysis. The specific techniques used to achieve each step in the analysis depend on whether the cultural researcher is employing exploratory CNA or confirmatory CNA.

A primary goal of exploratory CNA is to develop an initial understanding of the concepts and characteristics that are culturally relevant within the domain. In exploratory CNA, concepts, causal beliefs, and values are extracted from interviews and other qualitative sources. Semi-structured interviews employ questions intended to elicit antecedents and consequences of concept states, as in the "explanatory models framework" sometimes used in cognitive anthropology (Garro, 2000). Questioning along these lines draws out a more comprehensive set of ideas than would typically be verbalized in standard think aloud procedures, and places particular emphasis on drawing out perceived causal relations. We have also combined this interview approach with "value focused thinking" from decision analysis to elicit values and objectives directly, along with the causal beliefs that link more fundamental values with the means intended to achieve them (Keeney, 1994; Rasmussen, Sieck, & Smart, 2009). Qualitative analysis and representation at this stage yield insights that can be captured in initial cultural models.

Influence diagrams have proven to be quite useful for representing mental models relevant to key judgments and decisions (Bostrom, Fischhoff, & Morgan, 1992). We further believe they are an important representation format for depicting cultural models, especially for showing both qualitative structure and numeric prevalence information. In an influence diagram, nodes are linked by arrows that represent local causal influences. That is, the value of the concept at the beginning of an arrow affects the value of the concept at the arrow's point. Fully-specified influence diagrams can also represent numerical quantities, but the basic structure is useful as well. Specifically, an influence diagram provides a relatively simple and useful representation of a cultural model of another's mind that includes key judgments and decisions of interest to the researcher, as well as the culture-specific concepts, values, and causal beliefs typically used to explain those decisions within a population.

Confirmatory CNA serves to test the structure of previously developed qualitative cultural models, as well as to elaborate the models with quantitative data on the prevalence of ideas in the population(s) of interest. In confirmatory CNA, specially-designed structured questionnaires are used to obtain systematic data that can be subjected to statistical analysis. Most questionnaires treat ideas as independent entities, and so do not provide any means for revealing their interrelated, network form. A few studies have attempted to capture first-order causal beliefs. We have begun developing questionnaires that permit the analysis of longer causal belief chains by starting with influence diagram representations of qualitative cultural models from exploratory CNA to provide a suitable reference.

Statistical models, such as cultural consensus theory and mixture models are employed in confirmatory CNA to assess the patterns of agreement from the "causal-belief" surveys, and derive statistics describing the distribution of concepts, causal beliefs, and values. Cultural consensus theory is a collection of formal statistical models that has been long used within cognitive anthropology to assess agreement in knowledge and beliefs among a set of respondents (Romney, Weller, & Batchelder, 1986). When a cultural consensus is found, it provides the consensual responses that indicate culturally shared knowledge and estimates of the strength of consensus for those responses.

Our group has increasingly been relying on mixture modeling as an alternative approach to cultural consensus theory, primarily as it permits direct segmentation of cultural groups based on clusters of consensus (Mueller & Veinott, 2008; Sieck & Mueller, 2009). Mixture models have been applied in many scientific fields, including

marketing, biology, medicine, and astronomy. A mixture model, or "finite mixture model," is given as a combination of different groups, each described by a distinct probability distribution. Mixture models sort through the data and group them into sets of relatively homogeneous cases or observations. In cultural modeling applications, the distinct segments resulting from the analysis represent *cultural groups*, i.e., groups defined by the similarity of their ideas.

Finally, influence diagram representations of the cultural models are constructed in confirmatory CNA that illustrate the statistical properties, as well as the qualitative structure elucidated in exploratory CNA. In the confirmatory CNA application, the influence diagram represents the "culturally correct" concepts, values, and causal linkages as determined by mixture modeling for each cultural group that was found. Furthermore, the numerical probability values in the diagram are populated with the prevalence of each idea, as measured by selection percentage, within a group. The result in this case is a summary of the full distribution of ideas, with probabilities indicating the consensus on any particular causal link (or node).

As discussed above, CNA provides an integrated collection of techniques and procedures that can be usefully employed to build cultural models in virtually any knowledge domain. Here, we illustrate their use for building cultural models that pertain to folk psychology. Specifically, the substance of our research primarily considers the general concept of corruption and its relationship to explanations of individual decisions in the context of Afghanistan.

Exploratory CNA: Afghan Expatriates

Method

Participants 14 Afghans living in the U.S. participated in the study. Most (80%) were men, aged between 20 and 34 (mean 26.7) years, and had resided in the U.S. for a time between 3 months and 9 years (mean 2.7 years). All spoke English in addition to Dari or Pashto**.**

Materials The interviews were structured around short scenarios based on real events. All involved Afghan actors who engaged in some behavior that puzzled Americans. Here, we focus on a scenario involving a Mullah who was helping a group of Afghans and Americans to distribute humanitarian assistance supplies in several villages. The scenario included that the Mullah was extremely helpful, especially in facilitating positive interactions with village elders. Everything was going very well. After finishing with the distribution, as everyone was packing up to leave, the American leader learns that the Mullah has kept a truckload of the supplies. The American who originally relayed this incident indicated a belief that the Mullah was operating out of a desire to increase his own wealth. As will be seen, the Afghan interpretations are somewhat different**.**

Procedure Each participant was interviewed individually using the same CNA interview guide. A primary and secondary interviewer were present for all interviews. The primary interviewer was responsible for covering the questions in the interview guide. The secondary interviewer took notes and asked additional questions of clarification. The interview guide consisted of open-ended questions to elicit participants' overall explanations of the situation, as well as their beliefs regarding the Mullah's intentions, objectives, fundamental values, and causal links between them. Questions also covered participants' beliefs about the Mullah's decision process, value conflicts, and anticipated reactions to possible mitigating actions.

Results and Discussion

Coding Two independent coders read through all of the transcripts and identified segments that contained concept - causal belief - value chains. Next, the two analysts coded each segment by identifying the antecedent, the consequence, and the direction of the relationship between the antecedent and consequence (increasing or decreasing) for each causal belief. Percent agreement for the coding was 95.2%.

Representation The concept - causal belief – value chain fragments were then integrated into a network diagram that synthesized participants' models of the Afghan Mullah's mind, as bounded by the scenario (See Figure 1). In the diagram, circles depict concepts, arrows represent causal beliefs, and values are indicated by color of the circles. One analyst constructed an initial draft of the overall diagram by first focusing on fragments related to key decision points and fundamental values within the scenario, and subsequently adding further detail. Three other analysts independently reviewed the resulting diagram against the original set of codes, and a final version determined by consensus.

Findings A rich structure emerged from the process, as illustrated in Figure 1. As can be seen in Figure 1, three main sections of peoples' beliefs about the Mullah's mental states are represented: 1) Mullah successfully acquiring supplies, 2) Mullah being caught taking supplies, and 3) Mullah's concept of theft. As shown, respondents believed that the Mullah's intentions for the supplies included using them within his own household, selling them, or distributing them among people within his village (this latter category can be further decomposed into distinct groups of people, as described in Study 2). Interestingly, the fundamental values projected onto the Mullah rarely reflected considerations of material gain for the sake of wealth alone. Instead, the Mullah's fundamental values guiding his decisions in this situation were believed to comprise considerations of status, respect, power, and honor. These considerations are examined further in Study 2. Being caught with the supplies was generally felt to negatively impact these four values, and so come at a significant personal cost to the Mullah. However, respondents believed the degree of impact would be minimized or exacerbated depending on how discreetly

the situation was resolved. With respect to the concept of theft, participants did not believe the Mullah would consider himself as stealing and also felt that he would likely view theft quite negatively. In order for the act of reserving supplies to be considered theft, the goods would need to come from local rather than foreign sources.

Confirmatory CNA: Afghans in Afghanistan

Method

Participants Participants were 405 men from 4 provinces (Balkh, Kandahar, Kabul, Herat) representing north, south, east, and west regions of Afghanistan. Approximately half were from rural villages (54%), and the remainder lived in urban areas. Ages ranged from 18 to 78 years old (mean = 32.6). The vast majority (96.8%) spoke either Pashto or Dari as their primary language**.**

Materials The Mullah section of the survey consisted of a series of brief vignettes describing possible intended actions, followed by closed-form questions about the objectives and consequences for those actions. An example vignette/question is, assuming the Mullah successfully acquires the supplies, "what will he most likely do with the supplies?"

1. Use in his immediate household;
2. Sell or trade them;
3. Give them to his close friends or supporters;
4. Give them to his extended family; and
5. Give them to needy people in his own village.

Again, the survey options presented were not created by the researchers, but instead derived from interviews with Afghans in the exploratory CNA study. The sequences ultimately led to six fundamental values that were also derived in the exploratory phase of the study: *status, honor/respect, wealth, power, safety, and family approval*. The CNA survey was translated into Dari and Pashto languages, and back-translated to ensure preservation of meaning (Brislin, 1970)**.**

Procedure The survey was administered to participants through structured face-to-face interviews. The interviews were conducted by trained Afghan interviewers who live in the same province where they collected data. Before the data collection began, supervisors from each provincial data collection center met with the principal investigator at the opinion research center headquarters in Kabul to discuss the study purpose, survey content, and data collection procedures. The supervisors then returned to their provincial centers and held similar sessions for the local Afghan interviewers. Local interviewers then collected data at randomly assigned sampling points in their province. Participants were interviewed individually by one interviewer. The interviewer documented participant responses to the survey using paper-and-pencil. The survey took approximately 1 hour to administer.

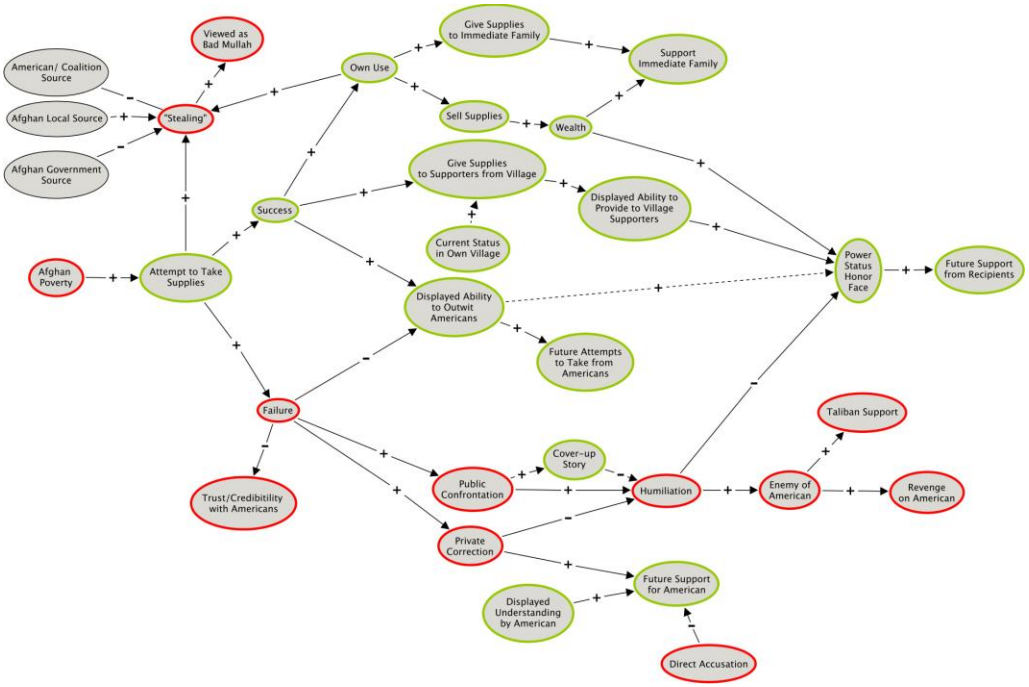


Figure 1. Qualitative cultural model of a Mullah's mind.

Results and Discussion

Model fitting was conducted using a statistical package called "FlexMix" (Leisch, 2004). FlexMix uses an iterative maximum likelihood procedure called the, "EM algorithm,"

for model estimation. A mixture of binomial distributions was fitted to the data set, after categorical variables were recoded in binary terms. The possible ("finite") number of resulting groups was allowed to vary between 1 and 7. The best fitting model was selected using the Bayes Information Criterion (BIC) statistic. The BIC statistic indicated that the best fit was achieved with 3 cultural groups (BIC = 73096.7) of roughly equivalent size (n_1=121, n_2=152, n_3=132).

Differences between the groups appeared to be fairly subtle. For brevity, a trimmed version of the cultural model is presented for Group 2, only (see Figure 2). As illustrated in Figure 2, participants in this group tended to believe the Mullah would use the supplies within his own household, though reasonable proportions felt he would either sell them, or distribute them among the needy in his village. Interestingly, the majority of possible motivations for Mullah actions center around fundamental values of status and respect. The possibility that the Mullah is simply seeking to increase his wealth appears to constitute a minority view among Afghans. This finding corroborates the initial results from the exploratory CNA study, and again differs from the original American interpretation that the Mullah was operating out of greed.

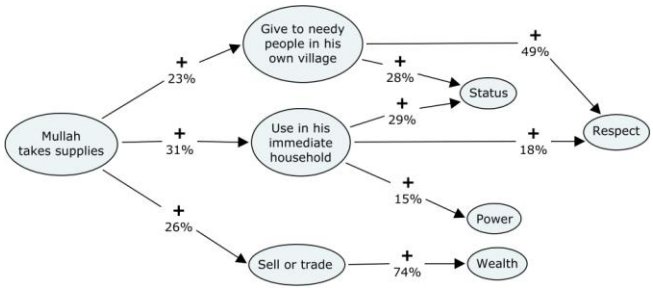


Figure 2. Quantitative cultural model of a Mullah's mind.

Discussion

We all have the ability to think and speculate about the behavior of objects, events, and other people. We do this naturally in a variety of domains. In the social domain, we are able to make guesses about other people's thoughts and therefore speculate about their intentions and their motives and use those guesses to generate plausible explanations for their behavior. Human interaction and communication relies heavily on our ability to anticipate each other's actions and questions. In fact, one could argue that the ability to predict and explain the behaviors of people around us in common terms is central to our ability to thrive in the local social environment. Methods to support the investigation of such explanations and predictions among localized populations are clearly warranted.

In this article, we described a method that can be used to study folk theories of psychology. The method, Cultural Network Analysis seeks to explicitly map the distribution of mental models within a cultural group. Specifically, the distribution of a cultural group's knowledge within a domain and situation is analyzed and displayed using a network representation of consensus elements. We also illustrated the use of the method to explicitly represent folk theories of mind in a cultural context. In particular, we used exploratory and confirmatory CNA, respectively, in two studies to tease out Afghan explanations of an Afghan Mullah's decision making in an ethically-charged scenario.

A core assumption of our program is that peoples' intuitive understandings of human psychology are fundamental to many more complex domains of interest in cultural research and applications (e.g., reading intentions, negotiating, and collaborating across cultures). Hence, cultural investigations of mental models of psychology using cultural network analysis, among other methods, will provide a useful starting point for addressing these more complex cultural domains.

Acknowledgments

This research was supported in part by the Human Social Cultural Behavioral (HSCB) modeling program, under CTTSO cooperative agreement W91CRB-09-C-0028. The authors thank collaborators Mansour Javidan, Rafik Ullah Kakar, Joyce Osland, Ben Simpkins, and Jennifer Smith, for assistance and suggestions at various stages of the research.

References

Atran, S., Medin, D. L., & Ross, N. O. (2005). The cultural mind: Environmental decision making and cultural modeling within and across populations. *Psychological Review, 112*(4), 744-776.

Berger, J. A., & Heath, C. (2005). Idea habitats: How the prevalence of environmental cues influences the success of ideas. *Cognitive Science, 29,* 195-221.

Bostrom, A., Fischhoff, B., & Morgan, M. G. (1992). Characterizing mental models of hazardous processes: A methodology and an application to radon. *Journal of Social Issues*, **48**, 85-100.

Brislin, R. W. (1970). Back-translation for cross-cultural research. *Journal of Cross-Cultural Psychology*, 1, 185-216.

D' Andrade, R. (1987). A folk model of the mind. In D. Holland and N. Quinn (Eds.), *Cultural Models in Language and Thought* (pp. 112-148). Cambridge: Cambridge University Press.

D'Andrade, R. (1995). *The Development of Cognitive Anthropology*. Cambridge: Cambridge University Press.

D'Andrade, R. (2005). Some methods for studying cultural cognitive structures. In Naomi Quinn (Ed.), Finding Culture in Talk (pp. 84-104). New York: Palgrave Macmillan.

Garro, L. C. (2000). Remembering what one knows and the construction of the past: A comparison of cultural consensus theory and cultural schema theory. *Ethos, 28*(3), 275-319.

Gentner, D., & Stevens, A. L. (Eds.). (1983). *Mental models*. Mahwah, NJ: Lawrence Erlbaum & Associates.

Gopnik, A., & Wellman, H. M. (1994). The theory theory. In L. Hirschfeld & S. Gelman (Eds.), *Mapping the Mind: Domain Specificity in Cognition and Culture* (pp. 257-293). New York: Cambridge University Press.

Hirschfeld, L., & Gelman, S. (Eds.). (1994). *Mapping the mind: Domain specificity in cognition and culture*. New York: Cambridge University.

Keeney, R. L. (1994). Creativity in decision making with value-focused thinking. *Sloan Management Review, 35*(4), 33-41.

Leisch, F. (2004). FlexMix: A general framework for finite mixture models and latent class regression in R. *Journal of Statistical Software, 11*(8), 1-18.

Leslie, A. M., Friedman, O., & German, T. P. (2004). Core mechanisms in 'theory of mind'. *Trends in Cognitive Sciences, 8*(12), 529-533.

Lillard, A. (1998). Ethnopsychologies: Cultural variations in theories of Mind. *Psychological Bulletin, 123*, 3-32.

Mueller, S. T., & Veinott, E. S. (2008). Cultural mixture modeling: Identifying cultural consensus (and disagreement) using finite mixture modeling. *Proceedings of the Cognitive Science Society*. Washington, DC.

Quinn, N. (2005). Finding Culture in Talk: A Collection of Methods. New York: Palgrave Macmillan.

Rasmussen, L. J., Sieck, W. R., & Smart, P. (2009). What is a good plan? Cultural variations in expert planners' concepts of plan quality. *Journal of Cognitive Engineering & Decision Making*, **3**, 228-249.

Romney, A. K., Weller, S. C., & Batchelder, W. H. (1986). Culture as consensus: a theory of culture and informant accuracy. *American Anthropologist, 88*, 313-338.

Sieck, W. R., & Mueller, S. T. (2009). Cultural variations in collaborative decision making: Driven by beliefs or social norms? In *Proceedings of the International Workshop on Intercultural Collaboration* (pp. 111-118). Palo Alto, CA.

Sieck, W. R. (2010). Cultural network analysis: Method and application. In D. Schmorrow & D. Nicholson (Eds.), *Advances in Cross-Cultural Decision Making*, CRC Press / Taylor & Francis, Ltd.

Sieck, W. R., Rasmussen, L. J., & Smart, P. (2010). Cultural network analysis: A cognitive approach to cultural modeling. In D. Verma (Ed.), *Network Science for Military Coalition Operations: Information Exchange and Interaction* (pp. 237-255). Hershey, PA: IGI Global.

Sperber, D. (1996). *Explaining culture: A naturalistic approach*. Malden, MA: Blackwell.

Some Attention Learning "Biases" in Adaptive Network Models of Categorization

Toshihiko Matsuka Chiba University
James E. Corter Teachers College, Columbia University
Arthur B. Markman The University of Texas

Abstract

In two simulation studies, we compare the attention learning predictions of three well-known adaptive network models of category learning: ALCOVE, RASHNL, and SUSTAIN. The simulation studies use novel stimulus structures designed to explore the effects of predictor diagnosticity and independence, and differentiate the models regarding their tendencies to learn simple rules versus exemplar-based representations for categories. An interesting phenomenon is described in which the models (especially SUSTAIN and RASHNL) learn to attend to a completely nondiagnostic constant dimension.

Keywords: category learning; selective attention; simulation.

Introduction

A key assumption of many computational models of categorization is that category learners do not merely form associations between instances and categories, but also learn how to allocate attention to each individual stimulus "dimension" (e.g., color). The present paper focuses on three such adaptive network models of classification learning: the ALCOVE model of Kruschke (1992); RASHNL (Johansen & Kruschke, 1999); and SUSTAIN (Love & Medin, 1998). These models are multilayer adaptive network models that accept as input a stimulus description (in the form of a set of input feature values), and produce as output category membership predictions that are based on the activation levels of a set of output nodes that correspond to the possible category responses. Over the course of training, these models learn both what dimensions to attend to, and how to correctly classify all the stimuli in the training set.

These three adaptive network models differ in several key aspects. ALCOVE and RASHNL are exemplar models, in the sense that each stimulus in the training set is allocated a node in the "hidden" or middle layer of the network. In contrast, SUSTAIN can form either exemplar-level or prototype-based representations. Prototypes are handled by using a reduced number of nodes in the hidden layer, corresponding to potential generalizations. SUSTAIN dynamically allocates new prototypes, allowing it to possibly use *multiple* prototype nodes for each category defined by the training feedback.

Exploring how these models adapt their attention weights is crucial to understanding their usefulness and validity by relating their learning accuracy predictions more directly to learning strategies. In previous studies (e.g., Matsuka & Corter, & Markman, 2002; Corter, Matsuka, & Markman, 2007), we found that all three models can account for human classification accuracy learning curves, but show distinct patterns in their "learning curves" for dimensional attention weights. In particular, ALCOVE and RASHNL seem to pay more attention to relatively independent predictors, while SUSTAIN shows the reverse pattern. The present Simulation 1 seeks to confirm this finding with a novel stimulus structure designed for this purpose, while Simulation 2 investigates an interesting phenomenon whereby the models sometimes learn to pay attention to a completely nondiagnostic feature. First, we briefly describe the models.

ALCOVE (Kruschke, 1992) is a multi-layer adaptive network model of categorization based on the Generalized Context Model (Nosofsky, 1986). The first layer of ALCOVE is a stimulus input layer. Each node in this layer represents the value of the presented stimulus on a single dimension. Importantly, each dimension has an attention strength (α_i) associated with it. Typically, attention strengths are initially equal across dimensions. However, the model learns to reallocate attention as learning proceeds, by adjusting these weights. The second layer in the network is the exemplar layer. Each node in this layer corresponds to an exemplar, described by its position in the multidimensional stimulus space. The activity of the exemplar nodes is fed forward to the third layer, the category layer, whose nodes correspond to the categories being learned. Separate learning rates are assumed for the association weights and attention strengths.

RASHNL (Kruschke and Johansen, 1999) is a modified and extended version of ALCOVE. The modifications introduced in RASHNL include: limited attention capacity; a capability for large and rapid shifts of attention; a gradually decreasing learning rate; and a parameter for salience of cues or features. RASHNL's architecture is similar to that of ALCOVE. However, each dimension has a dimensional salience parameter, the values of which are prespecified by the experimenter (i.e., not adjusted by learning). The dimensional attention strengths, α_i, are derived functions of separate underlying parameters, termed the "gains", which are adjusted by learning. An additional parameter P is incorporated, that can be set to vary between fixed attention capacity ($P = 1$) or unlimited attention capacity ($P = \infty$).

SUSTAIN (Love & Medin, 1998; Love, Medin & Gureckis, 2004), is comprised of two separate adaptive network components, a "supervised" network and an "unsupervised" one. The unsupervised network is a competitive network that clusters stimuli into prototypes. The term 'prototype' is used broadly, however, because an experimenter-defined category might be represented by one or many prototypes, and a prototype might represent only a single stimulus. This

flexibility also gives SUSTAIN the capability to form prototype-plus-exception representations or even exemplar-level representations. This clustering network is dynamic and incremental in its behavior, in the sense that new prototypes and/or exceptions are created when current prototypes are not predictive.

The "supervised" network is a feedforward network that classifies a stimulus based on similarity between the input pattern and the prototypes created by the unsupervised network. The activation of node *j* in the internal layer depends on several parameters: λ_i, which represents the "tuning" of the receptive field for a given dimension *i*, the distance between the centroid of prototype unit *j* and the input node on dimension *i*, and *r*, an overall attentional parameter that can be adjusted to create tighter or looser focus on highly tuned dimensions. The "tuning" (λ_i) parameters in SUSTAIN are the primary determinants of differences in attention among dimensions. When λ_i is large, difference between the input and the prototype node on dimension *i* are "stretched" or emphasized. At the output layer, SUSTAIN allows only the internal-layer unit with the highest post-transformed activation to determine output node activations, leading to "winner-take-all" learning.

Comparing the Models' Accounts of Attention Learning

We are interested in the attention learning behavior of these models. One clear difference between models is that RASHNL was designed with multiple attention learning iterations on each trial, in order to account for rapid shifts in attention that ALCOVE cannot predict. However, other differences among the models' assumed attention mechanisms have unknown implications. For example, it is not clear what follows from SUSTAIN's use of feedback from only the most-activated prototype to update the dimensional tuning parameters. Because of the complexity of these multilayer network models and their dynamic nonlinear performance, simulation studies are useful to establish the models' actual attention-learning behavior in complex learning tasks.

SIMULATION STUDIES

Simulation 1

Our previous findings (e.g., Corter et al., 2008; Matsuka et al. 2002) suggest that ALCOVE and RASHNL tend to incorporate dimensions that are relatively independent, even orthogonal, to the other predictors, compared to SUSTAIN. As an alternative (but related) hypothesis, it may be that relatively independent predictors are preferred by ALCOVE and RASHNL because such dimensions often are more useful for distinguishing exemplars, especially between categories. Simulation 1 explores this hypothesis by decoupling predictor diagnosticity (correlation with the criterion), predictor independence (inversely related to correlation with the other predictors), and "exemplar separation" (i.e., whether a predictor can be used in conjunction with other strong predictors in order to distinguish exemplars from different categories).

Method: Table 1 shows the category structure used in Simulation 1. In a typical classification learning task the classes (A and B) might be diseases, the exemplars patients, and the five "dimensions" might represent five types of test results or symptoms (each with two possible values). Correlations with the criterion are equal to .6 for Dimensions D1 and D2, to .2 for D3 and D4, and zero for D5. D3 and D4 differ in their configural validities, however: The variable subset (D1, D2, D3) gives a perfect R-square (RSQ) of 1.0 when these three dimensions are used in a linear model predicting the criterion, while the variable subset (D1, D2, D4) yields an RSQ of only .77. Addition of the orthogonal variable D5 alone does not increase the RSQ of the predictor set (D1, D2), which is equal to .60.

The dimensions also differ in their degree of independence from the other predictors. Dimension D3 is correlated .6 with D1 and with D2, while D4 is correlated -.2 with each of these two predictors. D5 has a zero correlation with all the other predictors and the criterion. However, the predictors D3-D5 are all comparable in one regard: each one can be used in conjunction with D1 and D2 to distinguish all category A exemplars from all category B exemplars. Thus, the simulation results for this structure should shed light on our hypothesis that this "exemplar separation" measure is key to predicting ALCOVE's and RASHNL's attention allocation behavior, by holding this factor constant across the "extra" dimensions D3-D5.

Table 1. Stimulus structure used in Simulation 1.

Class	D1	D2	D3	D4	D5
A	1	1	1	1	1
A	1	1	1	0	0
A	1	1	1	0	1
A	1	0	0	1	0
A	0	1	0	1	1
B	1	0	1	0	1
B	0	1	1	0	0
B	0	0	0	1	1
B	0	0	0	1	0
B	0	0	0	0	1

Using the three models, we simulated subjects (N=10,000) who were trained for 20 blocks on the stimulus structure shown in Table 1. For each individual subject parameter values were randomly selected from a uniform distribution within reasonable limits for each parameter. The main results recorded were the final-block attention weights for the five dimensions.

Results: Although we cannot identify any of the simulated subjects as being descriptively more plausible than others due to the lack of empirical data for this structure, we can assess the normative success of each simulated subject, by calculating their predicted final-block classification accuracy. Table 2 shows the mean final-block attention parameters for each model, by dimension. The table shows the final weights only for "successful" simulated learners, those achieving at least 80% correct classification accuracy

by the final block. The results do not differ if all simulated learners are included, however. All three models give highest attention weight to the two high-diagnostic dimensions D1 and D2. However, they differ widely in how they distribute attention to the three remaining dimensions. In particular, the results for ALCOVE show a surprising pattern, with nearly as much attention paid to D4 and D5 as to the two most diagnostic dimensions and with D3 weighted least, even though D3 has the highest configural validity (RSQ = 1.0) in conjunction with D1 and D2. Thus, this pattern of weights can be said to be non-optimal; it is a surprising result in that D5 is completely uncorrelated with the criterion. This ordering is consistent with the hypothesis that ALCOVE prefers relatively independent predictors, and cannot be ascribed to differences in "exemplar separability", because this latter factor is held constant for D3, D4 and D5.

Table 2. Simulation 1: Final block relative attention weights for dimensions for each model, for "successful" simulated learners, with number (N) of successful learners out of 10,000 total.

	N	D1	D2	D3	D4	D5
ALCOVE	8480	.248	.247	.098	.199	.209
RASHNL	7463	.274	.286	.183	.123	.135
SUSTAIN	6855	.240	.248	.230	.136	.148

RASHNL and SUSTAIN both predict normatively satisfactory patterns of attention weights in the sense that they give highest attention to D1 and D2, with D3 third highest. This set of predictors is the minimal sufficient set for perfect prediction, therefore these weights may be considered to be the monotonically "optimal" weights. However, both RASHNL and SUSTAIN weight D5 higher than D4. Again this is surprising, since D5 has zero correlations with the criterion (but also with the other predictors).

Discussion: In this simulation RASHNL and SUSTAIN yielded weights that are normatively justifiable by the customary criterion of "configural validity", by giving highest weighting to the three dimensions yielding a perfect multiple-R in predicting the criterion. However, they still gave nontrivial weights to the two remaining dimensions, D4 and D5. In this sense their attention allocation patterns cannot be described as optimal. Furthermore, most simulated learners gave attention to more than one of these "supplementary" dimensions, showing that the network models do not always learn minimal sufficient rules.

ALCOVE also gave highest weights to D1 and D2, but gave third highest weight to D5, a dimension that has a correlation of zero with the criterion and with all the other predictors. This pattern seems "irrational' by the usual criterion of configural validity. However, we note that it is reasonable from the standpoint of "exemplar separability": by this measure, the set (D1, D2, D5) is adequate for the classification task. ALCOVE also gives non-trivial weights to the remaining two dimensions, D3 and D4, again demonstrating that the network models do not tend to learn minimal representations across a broad range of parameter values. Finally, ALCOVE weights D5 higher than D4 and D4 higher than D3, an ordering that is consistent with the degree of independence of the three dimensions, while RASHNL weights D5 over D4 (but weights D3 highest, in line with its configural validity). This result supports the hypothesis that ALCOVE tends to give higher weight to more independent dimensions, even at the cost of finding a non-optimal solution. RASHNL and SUSTAIN both find the "optimal" configuration of dimensions (D1, D2, D3), and in fact exhibit the same ordering of weights (D1≈D2>D3>D5>D4). However, given that only SUSTAIN weights D3 nearly as high as the two diagnostic dimensions, the results are consistent with the hypothesis that this model "prefers" dimensions that are correlated with other important predictors, compared to the other models.

Simulation 2

Simulation 2 explores two issues. The first is the idea that SUSTAIN favors dimensions that are correlated with other predictors, at least relative to the other models. The second issue is the tendencies of the models to utilize exemplar versus simple rule based strategies when both strategies are sufficient for perfect performance.

Our previous simulations suggest that ALCOVE and RASHNL favor relatively independent predictors of the criterion. A form of independence that can arise with a very poor predictor of a criterion is the case of a constant predictor. A constant has a correlation of zero with the other predictors, and also with the criterion (very bad diagnosticity indeed). We explore whether ALCOVE and RASHNL have any attraction to this type or predictor.

There is reason to suspect that SUSTAIN may try to incorporate such a predictor. Although a constant dimension has zero correlation with other predictors, it will have maximal within-category consistency for any cluster. Thus, the inclusion of a constant dimension allows us to unconfound diagnosticity and between-predictor correlation from within-cluster consistency, possible aspects of the type of dimensions found to be attractive to SUSTAIN in previous simulations.

Inclusion of a constant dimension simulates important aspects of experimental stimuli that are usually ignored. The stimuli used in studies of category learning typically have many perceptually or conceptually salient aspects that are not coded or discussed by the experimenters, being treated as irrelevant because they are constant for all stimuli. For example, stimuli that are line drawings of bug-like creatures may differ in head shape, number of legs, and type of tail, aspects that are coded and manipulated by the experimenter to define the diagnostic input features to categorization models. But the line drawings all share certain basic characteristics that are constant across stimuli. Many models of similarity (e.g., Tversky, 1977; Markman & Gentner, 1993) assume that common features increase the similarity (and confusability) of stimuli. Thus, it seems interesting to use a simulation study to investigate what

predictions the three network models make for use of such constant or common-feature information.

Method: The category structure used for Simulation 2 is shown in Table 3. There are four exemplars of each category, A and B. Dimension D1 is a binary-valued variable, with values that are logically necessary-and-sufficient to identify each category. Dimension D2 is a constant dimension that has values of 1 for all exemplars in the population, regardless of category membership. Dimensions D3, D4, and D5 are binary-valued dimensions that together uniquely identify all eight exemplars. Note that this structure ensures that each network model not only has a relatively easy categorization strategy available (a unidimensional rule on D1), but can adopt a minimal attentional strategy that enables unique identification of all exemplars (attending to D3-D5).

Using the three models, we simulated subjects (N=100,000) who were trained for 20 blocks on the stimulus structure shown in Table 3. As in Simulation 1, for each individual subject parameter values were randomly selected from a uniform distribution within reasonable limits for each parameter. The main results recorded were the final-block attention weights for the five dimensions.

Table 3. Simulation 2: A simple two-category structure with one necessary-and-sufficient "category" dimension (D1), a constant dimension (D2), and three dimensions (D3-D5) that uniquely identify exemplars.

Class	D1	D2	D3	D4	D5
A	1	1	1	1	0
A	1	1	0	1	1
A	1	1	1	0	1
A	1	1	0	0	0
B	0	1	1	1	1
B	0	1	0	1	0
B	0	1	1	0	0
B	0	1	0	0	1

Results: Table 4 reports the mean pattern of relative attention in the final block for the successful classification learners, defined as those who had at least 80% classification accuracy in the final block.

Table 4. Mean final relative dimensional attention weights, by model, for the best-fitting simulated subjects of Simulation 2. Maximal mean attention weight for each model shown in bold type.

Model	D1	D2	D3	D4	D5
ALCOVE	**.338**	.097	.188	.188	.188
RASHNL	**.375**	.231	.132	.132	.132
SUSTAIN	**.389**	**.389**	.074	.075	.075

As can be seen in the table, learners simulated by ALCOVE gave the highest weight to D1, the dimension defining the simple rule. However, the total attention weight allocated by ALCOVE to the three dimensions uniquely identifying the exemplars (D3-D5) was greater than that given to the rule dimension D1, a pattern that could be interpreted as showing predominantly exemplar-based learning.[1] ALCOVE was relatively successful at ignoring the constant dimension D2, giving it about 1/4 the weight of the "rule" dimension D1. RASHNL showed a different pattern of final weights, giving the highest weight to the dimension (D1) defining the unidimensional category rule, an intermediate level to the constant dimension D2, and the least attention to the exemplar-identifying dimensions D3-D5. RASHNL's capability to emphasize D1, the rule dimension, is consistent with its capability to model simple rule-based strategies in other simulations we have conducted. It is somewhat surprising that this model cannot learn to ignore the constant dimension D2. SUSTAIN gave the least weight of any model to the "exemplar" dimensions D3-D5, and roughly as much weight as RASHNL to the perfectly diagnostic D1, but was the worst at ignoring D2, the constant dimension, giving it equal weight with D1.

Examination of the pattern of attention results across different regions of the parameter space for each model revealed that one key parameter affecting the results is the learning rate for association weights in the network. In order to display these results, we have created plots of the final pattern of attention weights for each model, separately for different ranges of the learning rate parameter.

Figure 1 presents the results for ALCOVE. The left panel plots the final attention weight for D2 (the constant dimension) versus that for D1 (the rule dimension). It can be seen that ALCOVE does not completely ignore D2 at any value of the learning rate, although D2 is consistently given lower weight than D1. The right panel plots the summed final attention weights for D3-D5, the "exemplar" dimensions, versus the weight for D1. These plots show a strong and consistent effect of the learning rate for associations. For higher values of this parameter (the upper row of the panel), the total attention weight given to the exemplar dimensions tends to exceed that for D1, meaning that exemplar learning predominates. For lower values of the learning rate (the bottom row) the dimension defining the unidimensional rule (D1) is weighted highly, sometimes even exclusively, meaning that a rule-based strategy is being used.

Figure 2 presents the corresponding plots for RASHNL. The left panel shows that RASHNL has trouble ignoring D2, the constant dimension, at any value of λ_w. However, D1 (the rule dimension) tends to receive more attention than D2 in the majority of solutions. The right panel shows that most simulated subjects pay more attention to D1, the rule dimension, than to the exemplar dimensions. This is especially true when the learning rate is very low (bottom row). However, the bottom row of the left panel

[1] Support for this interpretation is given by supplementary simulations we have conducted, in which various numbers of dimensions (1, 2 or 3) are used to uniquely code the exemplars. Across all of these simulations, the total final weight given to these "exemplar" dimensions is roughly constant, regardless of the number of dimensions involved.

underscores that for the low learning rates, considerable attention is also paid to D2, the constant dimension.

Figure 3 shows that SUSTAIN yields a very different pattern of results for this structure. For all values of λ_w SUSTAIN predicts that equal attention will be paid to D1 (the rule dimension) and D2 (the constant dimension). Also, the total amount of attention directed at D3-D5, the "exemplar" dimensions, is fairly stable across values of the learning rate, but there is more variability at the higher learning rates. Interestingly, the apparent constraint that the weights given to D1 and D2 be equal is so strong that any increase or decrease in the total weight given to D3-D5 trades off against the summed relative weight given to D1 and D2, creating a line of possible solutions with a slope of -2 in each plot of the right-hand panel.

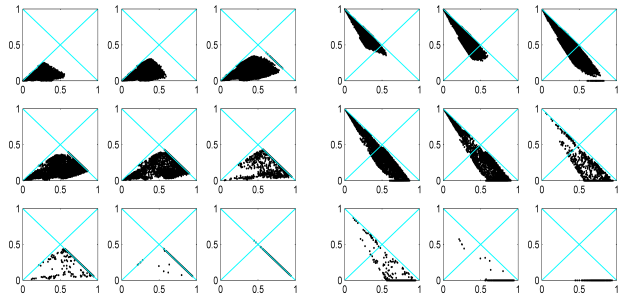


Figure 1. Simulation 2: Final relative attention weights for ALCOVE, separately for different values of λ_w, the learning rate for network association weights. Left panel: D2 (y-axis) versus D1 (x-axis) attention weights. Right panel: summed attention weights for D3, D4 & D5 (y-axis) versus D1 (x-axis) attention weights. In each panel, the nine plots summarize results for various ranges of the λ_w parameter. Top row: (>.8; .8-.4; .4-.2). Middle row: (.2-.15; .15-.10; .10-.05). Bottom row: (.05-.025; .025-.125; <.125).

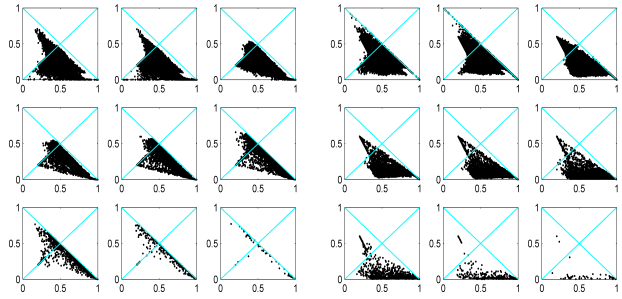


Figure 2. Simulation 2: Final relative attention weights for RASHNL

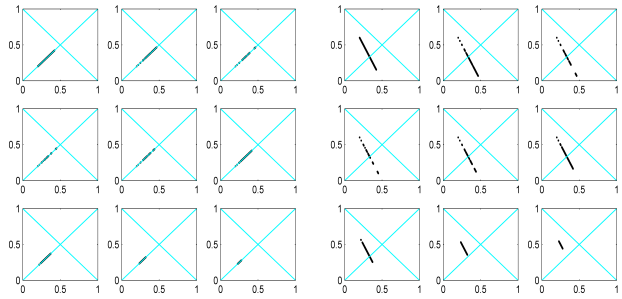


Figure 3. Simulation 2: Final relative attention weights for SUSTAIN

Discussion: The results of Simulation 2 are striking. First, both RASHNL and SUSTAIN pay considerable attention to a constant dimension (that has zero diagnosticity) under a wide range of parameter settings. In fact, RASHNL shows many solutions with relative weight exceeding 50% for D2 (with 5 dimensions). SUSTAIN invariably gives equal attention weight to D2 and D1, the unidimensional rule dimension. In this sense it is the least successful of the three models at ignoring D2. An explanation for this behavior of SUSTAIN is given below.

Second, the network models also differ in their tendencies to adopt the rule-based solution (using dimension D1) versus the exemplar-level representation (using D3-D5). For ALCOVE, successful learners tend to give high total attention weight to the "exemplar" dimensions D3-D5. These exemplar-based attention patterns occur often when the association learning rate is high, but rule-based attention patterns predominate when it is very low (Figure 1). For RASHNL, successful learners tend to weight the simple rule dimension (D1) more than the exemplar dimensions (D3-D5), and this tendency increases for low learning rates. Of the three models, SUSTAIN's successful learners give the least attention to the exemplar-identifying dimensions D3-D5. SUSTAIN pays somewhat more attention to these exemplar-identifying dimensions when the association learning rate is very low, the opposite pattern to that shown by ALCOVE and RASHNL.

Surprisingly, SUSTAIN gave the same amount of attention to D2 as to D1. Clearly, this tendency of SUSTAIN must arise from the structure and processing assumptions of the model. In fact, the reason that SUSTAIN finds D1 and D2 equally compelling is easy to identify, and stems from how SUSTAIN utilizes its reference points (i.e., clusters or prototypes) in learning. SUSTAIN utilizes only the single most activated cluster to determine an exemplar's classification and to guide learning. In this model, the update in attention strength for each dimension is inversely proportional to the distance from the most activated cluster's mean value and the value of the current input stimulus on that dimension (i.e., the smaller the dimensional distance, the more attention is increased for that dimension). For a constant dimension, any cluster and any input stimulus will have zero distance on that dimension, thus attention will be increased to the maximal degree possible on the constant dimension. In the present simulation, D1 is a perfect predictor with constant values *within* categories, thus any cluster that does not combine exemplars from across categories will also have zero distance on that dimension between the cluster centroid and the input stimulus, leading to an equivalent increase in attention strength to D2.

The critical aspect of the processing assumptions here is the winner-take-all nature of the utilization of the clusters, which means that the diagnosticity of a dimension relative to contrasting clusters has less effect. The net result in statistical terms is that the potential increase in attention to a dimension is a function of the similarity of the input

stimulus and the cluster centroid on that dimension. This places greater emphasis on within-category similarity and less on between-category differentiation, relative to the processing assumptions of ALCOVE and RASHNL. This line of analysis suggests that SUSTAIN will tend to select dimensions whose values have high category validity, P(f|c), over those with high cue validity, P(c|f), or with the best information gain (cf. Corter & Gluck, 1992).

Failing to ignore a dimension with zero diagnosticity seems like a major flaw of the three models, at least from a normative standpoint, because incorporating a constant dimension in a category's representation has cost without any obvious adaptive value. However, human data is needed to see if constant dimensions are indeed attended to and incorporated into a category's representation. It seems unlikely that in a category learning experiment human learners would waste time and effort memorizing or checking properties of a stimulus if those properties were seen to be useless for the task at hand.

On the other hand, it might be that such constant properties are learned implicitly, whether or not they are useful in a specific experimental task. An example might indicate why this is a reasonable possibility. A child learning the category *animal* might notice that all animals have mass. Is this fact incorporated into the child's representation? This certainly seems reasonable, though some normatively motivated theories of mental organization (e.g., Collins and Quillian, 1969) hold that the property of having mass should be stored at a superordinate level (say, under the category *object*) and merely inferred as needed in order to reason about animals and their properties.

Conclusions

The present analyses and simulation results show that the models examined here, ALCOVE, RASHNL, and SUSTAIN, incorporate differing attention learning mechanisms and processing assumptions that lead to distinct predictions regarding attention learning in the simulation studies reported here. The results from Simulation 1 supported the hypothesis that SUSTAIN tends to attend to dimensions that are correlated with other predictors, while the other models give relatively greater attention to more independent predictors, perhaps because they better support exemplar-level processing. Simulation 2 showed that the three models differ in their tendencies to use rule-based versus exemplar-based learning strategies. Another surprising result from Simulation 2 was that all three models incorporated a constant (i.e., completely nondiagnostic) dimension into their representations to some degree.

We believe that simulation studies on attention allocation in category learning are valuable for two reasons. First, they help us to better understand the behavior of complex computational models of category learning. Second, they can help to guide empirical work on attention by suggesting new hypotheses about human attention learning, hypotheses that can be verified using methods for assessing attention such as eye-tracking (e.g. Rehder & Hoffman, 2005) or information-board methods (Matsuka & Corter, 2008). These hypotheses may then be used to design empirical studies by suggesting stimulus structures and tasks that best differentiate predictions of the models.

References

Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior, 8(2),* 240-247.

Corter, J. E., & Gluck, M. A. (1992). Explaining basic categories: Feature predictability and information. *Psychological Bulletin, 111(2)*, 291-303.

Corter, J.E., Matsuka, T., & Markman, A. B. (2007). Attention allocation in learning an XOR classification task. Poster presented at the Second European Cognitive Science Conference (*EuroCogSci 2007*), Delphi, Greece, May 23-27.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review, 99,* 22-44.

Kruschke, J. K., & Johansen, M. K. (1999). A model of probabilistic category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 1083-1119.

Love, B. C., & Medin, D. L. (1998). SUSTAIN: A model of human category learning. *Proceeding of the Fifteenth National Conference on AI (AAAI-98),* 671-676.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of human category learning. *Psychological Review*, 111(2), 309-332.

Markman, A. B., & Gentner, D. (1993). Splitting the differences: A structural alignment view of similarity. *Journal of Memory and Language*, *32*, 517-535.

Matsuka, T., and Corter, J.E. (2008). Process tracing of attention allocation during category learning. *Quarterly Journal of Experimental Psychology, 61(7),* 1067-1097.

Matsuka, T., Corter, J. E., & Markman, A. B. (2002). Allocation of attention in neural network models of categorization. *Proceedings of the 24th Annual Meeting of the Cognitive Science Society*. Hillsdale, NJ: Lawrence Erlbaum.

Nosofsky, R. M. (1986). Attention, similarity and the identification–categorization relationship. *Journal of Experimental Psychology: General, 115*, 39-57.

Rehder, B., & Hoffman, A. B. (2005). Thirty-something categorization results explained: Selective attention, eyetracking, and models of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31(5),* 811-829.

Tversky, A. (1977). Features of similarity. *Psychological Review, 84,* 327-352.

Investigating Insight Using Compound Remote Associate Problems

Edward A. Cranford (eac53@msstate.edu)
Jarrod Moss (jarrod.moss@msstate.edu)
Department of Psychology, Mississippi State University
Mississippi State, MS 39762 USA

Abstract

Many new problems and paradigms have been developed to answer questions pertaining to insight problem solving. One problem type that may be useful to the study of insight is the Compound Remote Associate (CRA) problem, developed by Bowden & Jung-Beeman (2003). However, it is uncertain to what extent CRA problems are insight problems. We performed a protocol analysis of people solving CRA problems and found that CRA do exhibit some characteristics of insight. However, certain considerations should be taken into account. Particularly, problems solved when the solution is the first thing considered are often judged to be insight by participants, but these problems do not exhibit any characteristics of insight aside from the 'Aha' experience.

Keywords: Insight; Problem Solving; Restructuring; Impasse

Introduction

Problem solving enables us to discover solutions to problems. Sometimes, however, we reach an impasse, or road block, in the problem solving process. We may realize there is some flaw in our efforts, and the solution may seem unattainable, but the source of our error remains hidden. We may contemplate for a long period of time until, all of the sudden, the answer seems obvious. This phenomenon has been termed insight, and is loosely defined as achieving a solution without knowing where the solution came from.

Some phenomenological features are unique to insight. Insight solutions often appear from nowhere and solvers experience an affective response of suddenness and surprise (Aha! experience), sometimes resulting after an impasse; insight solutions are obtained through processes known as restructuring, whereby an incorrect representation of the problem is changed, leading to the access of an insightful, correct representation of the problem (e.g., Bowden, Jung-Beeman, Fleck, & Kounios, 2005; Ohlsson, 1992; Schooler, Fallshore, & Fiore, 1995). The key components of insight are often described as impasse, restructuring, and 'Aha!'.

Problems Used for Insight Research

Many types of problems have been used to study insight. Classic insight problems have been used extensively, and sometimes the reason is solely based on the fact that they have previously been used to study insight (Weisberg, 1995). With the emergence of advanced neuroimaging techniques (e.g., functional Magnetic Resonance Imaging [fMRI], Electroencephalography [EEG], etc.), and the great amount of time it takes to solve (if at all) complex, classic insight problems, new problems and paradigms have been used to investigate insight (Bowden et al., 2005). One specific problem type that has been used is the compound remote associate (CRA) problem (Bowden & Jung-Beeman, 2007). CRA problems involve finding the one word that can form a compound word or phrase with each of three different words. For example, if three words—tree, sauce, and big—are presented, the solution is apple. CRA problems are solved much quicker than classic insight problems. They can be solved by insight or by noninsight, search processes (i.e., generate-and-test or trial-and-error), and individual problems can be solved with insight regardless of learning effects over multiple trials (Bowden & Jung-Beeman, 2007). Given these differences between classic insight problems and CRA problems, the question is to what extent CRA problems are insight problems.

Are CRA problems Insight Problems?

Though differences have been found between insight-CRA problems and noninsight-CRA problems (Bowden & Jung-Beeman, 2007; Bowden et al., 2005; Jung-Beeman et al., 2004; Kounios et al. 2006), we need more empirical evidence that CRA problems can be used to study insight. Specifically, evidence is needed to show that CRA problems exhibit properties characteristic of insight, and that there are differences between solving a CRA problem with insight and solving one without insight beyond the Aha experience.

Bowden and Jung-Beeman (2003; 2007) and Bowden et al. (2005) claim CRA problems exhibit phenomenological features and components of insight found in classic insight problems and, therefore, should be used to study insight. The processing is often unreportable, the problems misdirect (or fail to direct) retrieval processes, and people experience the Aha!. These are reasons to assume that CRA problems can be used to study insight. However, because CRA problems are hybrid problems, rated insight or noninsight by the solver on a forced choice scale, and are such short and simple problems, there is concern about their use to study insight. A critical component not listed above for CRA problems is the process of restructuring, and it is unclear to what extent CRA problems exhibit restructuring prior to insight solutions.

We designed a study using concurrent verbal protocols of CRA problem solving to determine if CRA problems solved with insight exhibit more characteristics of insight, than CRA problems solved without insight. The characteristics of insight examined are impasse, restructuring, and verbal overshadowing. We expect higher rates of impasse and more restructuring processes in insight solutions than noninsight solutions. Concurrent verbalization of cognitive processing has been shown to inhibit solving a problem with

insight (Schooler, Ohlsson, & Brooks, 1993) resulting in a lower solution rate such that fewer problems are solved with insight than when problems are not verbalized. However, other research (i.e., Fleck & Weisberg, 2004) suggests verbalization does not necessarily inhibit solutions by insight and differences in verbalization instructions may differentially influence problem solving processes. In fact, the present research used an adaptation of the coding scheme used by Fleck and Weisberg as well as similar verbalization instructions, therefore, it is plausible that insight solutions may not show this verbal overshadowing effect. The results of this study should further inform the use of CRA problems in the study of insight.

Method

Participants

Participants were 31 undergraduates enrolled in a psychology course at Mississippi State University who received course credit for their participation. All participants were native English speakers.

Design

The design was a 2 (Verbal-Task: verbalization, nonverbalization) x 3 (Solution-Type: insight, noninsight, other) within-subject design. Solution-Type was measured by the subjective ratings given by the participant, rather than manipulated.

Materials

The task, described as word association problems to participants, consisted of a set of 60 CRA problems taken from a larger set of 144 normed CRA items (Bowden & Jung-Beeman, 2003). Problems were chosen based on information from a baseline study at Mississippi State University using all 144 problems. Problems with the highest solution rates that had been solved with insight, on average, half of the time were included in the set. An additional six problems from the set of 144 problems were used for practice trials. The 60 problems were presented in random order and randomly assigned to Verbal-Task condition for each participant. The problems were displayed on a 17-in. computer monitor and answers were given by typing on a keyboard. The task procedure was implemented using E-Prime software (Schneider, Eschman, & Zuccolotto, 2002). Concurrent verbal protocols were obtained via headset and recorded using E-Prime.

Procedure

Participants were run individually. After receiving informed consent, participants were instructed on the task and given descriptions of the differences for rating a problem with insight, noninsight, or other. These instructions were very similar adaptations taken from Jung-Beeman et al. (2004) and Kounios et al. (2006). The experimenter answered any questions the participants had about the tasks and the rating scale. Then the experimenter started the experiment on the computer, and the participants were instructed to read the directions presented for a second time via computer. After reading the general instructions, specific task instructions were given depending on the Verbal-Task condition for each participant. Participants were counterbalanced between Verbal-Task conditions so that half experienced the verbalization task first and the other half experienced the nonverbalization task first. For the verbalization condition, participants were instructed on how to verbalize their thoughts during problem-solving efforts based on the think-aloud instructions found in the appendix of Ericsson and Simon (1993). Participants who were asked to think aloud first were given instructions and training in how to verbalize their thoughts before the first task and told that they did not have to think aloud anymore right before the second task. If the first task was to solve problems without thinking aloud then participants were asked if they had any questions and allowed to continue. For these participants, think-aloud instructions were given after completing the nonverbalization task.

Participants were given three practice CRA problems before each Verbal-Task condition to make sure that they understood the difference in responses for rating a problem (insight, noninsight, and other), were verbalizing correctly, and understood the task requirements. The practice problems may have also helped to reduce some carryover effects from prior verbalization. Participants began the CRA task by pressing a button. The problem words were presented for a maximum of 30 seconds. Participants could give a solution at any time during the 30 second interval by typing their answer. If the given solution was incorrect they could continue work on the problem until time ran out. Upon correct solution within the time limit, participants were prompted to give a rating of whether they solved the problem via insight, noninsight, or other. The order of ratings was counterbalanced so that for half of the participants a rating of 1 was insight and a rating of 3 was noninsight and, for the other half, a rating of 1 was noninsight and a rating of 3 was insight. After a rating was given or solution time ran out, the next problem was presented. Thirty problems were presented and then the participant was asked to stop and notify the experimenter. The experimenter then gave the participant the appropriate instructions for the second Verbal-Task condition; after which, the participant continued to solve the next 30 problems while thinking aloud or keeping silent. Upon completion of the CRA task, participants were debriefed.

Results

Six subjects were dropped from all analyses (n = 25) due to outlier and zero data. One subject reported solutions by "other" much more often than any other subjects and five subjects reported solving problems only with insight, or only with noninsight, within at least one level of Verbal-Task.

Solution Rates and Times

Solution rates were calculated in each Verbal-Task condition for each Solution-Type as a percentage of problems solved (e.g. Insight percentage = number of insight solutions/number of total problems attempted). Response times were obtained for each solved problem to calculate the average time to solution for each Solution-Type and each Verbal-Task. Time started when the problem first appeared and ended when the final solution was entered. A verbal overshadowing effect is present if thinking aloud has a negative affect on response time and the number of problems solved (particularly with insight).

Overall, participants solved an average of 52.4% (*SD* = 11.10%) of problems. For solved problems, participants reported solution by insight 52.04% (*SD* = 15.27%) of the time, noninsight 37.14% (*SD* = 17.58%) of the time, and other 10.81% (*SD* = 11.16%) of the time. Average time to solution was 10.67 seconds (*SD* = 1.57). There were no effects of counterbalancing the orders of ratings or verbalization task conditions within subjects, so the data were collapsed across these levels of counterbalancing.

A 2 (Verbal-Task) X 3 (Solution-Type) within-subjects repeated measures analysis was performed to analyze the effects of Verbal-Task and Solution-Type (Insight, Noninsight, Other) on the percentage of problems solved (solution rates). More problems were solved in the Nonverbalization condition (M = 56.40%, *SD* = 11.37%) than in the Verbalization condition (M = 48.40%, *SD* = 14.52%), $F(1,24) = 8.64$, $p = .007$. Solution rates also differed for different Solution-Types, $F(2,48) = 30.69$, $p < .001$. Pairwise comparisons revealed that more problems were solved with Insight (M = .279, *SD* = .117) than Noninsight (M=.192, *SD* = .093), $t(24) = 2.58$, $p = .016$, and Noninsight than Other (M = .053, *SD* = .054), $t(24) = 5.353$, $p < .001$. Verbalization did not differentially affect the proportion of solutions across problem types as the Verbal-Task by Solution-Type interaction was not significant, $F(2,48) = .464$, $p = .631$ (Figure 1 shows a breakdown of the solution rates). Subsequent paired t-tests comparing the distributed proportions of correct insight, noninsight, and other solutions between levels of Verbal-Task also revealed no significant effects, all $t(24) < .70$, all $p > .50$, indicating that verbalization may affect the total number of problems solved but not the distribution of solution types.

For all following analyses only two levels of Solution-Type were used (Insight and Noninsight but not Other) because many subjects did not report "other" for any solved problems. A 2 (Verbal-Task) X 2 (Solution-Type) within-subjects repeated measure analysis was performed for solution response times. Response times were longer for Noninsight solutions (M = 12.9 seconds, *SD* = 3.29) than Insight solutions (M = 9.8 seconds, *SD* = 4.18), $F(1,24) = 5.74$, $p = .025$, and longer for Verbalization (M = 12.3 seconds, *SD* = 2.46) than Nonverbalization (M = 10.4 seconds, *SD* = 2.41), $F(1,24) = 10.21$, $p = .004$. As seen in the prior analysis, the interaction between Verbal-Task and Solution-Type was not significant, $F(1,24) = .553$, $p = .464$.

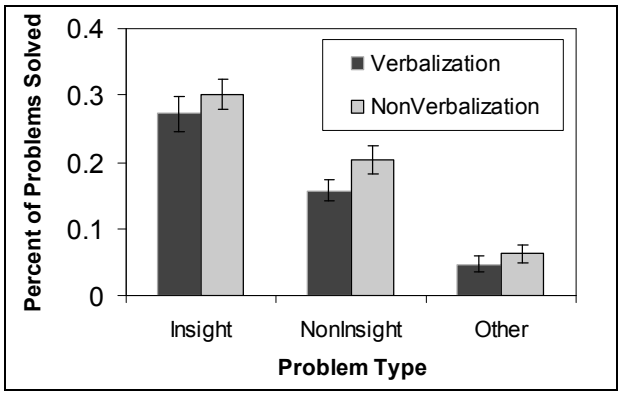


Figure 1: Solution rates of attempted problems. Mean percentage of attempted problems solved with insight, noninsight, or other while verbalizing and while not verbalizing. Error bars indicate one standard error.

Though it takes people longer to solve problems while thinking aloud, people will not likely solve any more problems if given more time. The results show that thinking aloud while attempting to solve CRA problems inhibits solution regardless of the solution method (insight, noninsight, or other) indicating an overarching verbal overshadowing effect. However, verbalization does not differentially affect insight and noninsight CRA solutions, which indicates that verbalization is not influencing the way in which the solution occurs but only if it occurs at all. Therefore, the verbal protocols may be used as data.

Verbal Protocol Analysis

The verbal protocols were coded for occurrences of impasses and restructuring elements. Impasse was calculated as the average number of impasses reached per correct solution and was coded into five different types. Restructuring was calculated as the average number of times restructuring occurred per correct solution and were coded into three different types. Because the amount of time a person spends on a problem affects the amount of impasses and restructuring that can possibly be obtained, each score for each problem was first divided by the amount of time it took to solve the problem before obtaining an average score for each participant. Details of each coding scheme are discussed below with each respective analysis.

When examining the verbal protocols we identified two distinct types of insight solutions. The first type of insight solution, termed "immediate-insight," occurs when the first candidate solution verbalized by the participant was the solution and this solution occurred within 15 s of the problem being presented. The second type of insight solution, termed "delayed-insight," are all other insight solutions not classified as "immediate-insight". A person may report the quickly solved problems as insight simply because they came to a solution so fast that it seemed sudden and surprising. However, it is unclear whether this should be called insight. Immediate insight solutions do not

Table 1: Within-subjects Repeated Measures Analysis for Each Level of Impasse (n=22).

	Means			
Analysis	Combined Insight	Immediate-insight	Delayed-insight	Noninsight
Rereading	0.0028 (0.0086)	0 (0)	0.0049 (0.0115)	0.0024 (0.0065)
Regenerating	0.0007 (0.0035)	0 (0)	0.0029 (0.0138)	0 (0)
Discontinuing	0.0006 (0.0020)	0 (0)	0.0028 (0.0099)	0.0038 (0.0131)
No New	0.0038 (0.0089)	0 (0) **	0.0089 (0.0173)	0.0058 (0.0083)
Frustration	0.0024 (0.0047)	0 (0) **	0.0135 (0.0297)	0.0071 (0.0157)
Total Impasse	0.0105 (0.0171)	0 (0) **	0.0332 (0.0375)*	0.0191 (0.0281)

Notes. Values represent mean number of occurrences per second. Standard deviations are in parentheses. Noninsight means remained the same for all analyses and are compared to each type of insight. Significant effects are represented by stars (*) in the "Insight" columns.
*$p < .10$. **$p < .05$.

exhibit any observable signs of impasse or restructuring in the verbal protocols and are the first solution candidate reported (clearly not insight as it is traditionally defined). Of solutions reported as insight, 77.73% (*SD* = 17.16%) were immediate-insight and 22.3% (*SD* = 17.16%) were delayed-insight. There were almost no immediate noninsight solutions. Therefore, in all subsequent analyses, noninsight solutions were not separated in order to simplify comparisons.

Three analyses for each coded variable were performed with only two levels of Solution-Type (Insight and Noninsight) and one level of Verbal-Task (verbalization). Three additional subjects were dropped from the analyses (*n*=22) because they did not report any delayed insight solutions. In an initial analysis we compared the combined delayed- and immediate- insight solutions ("combined insight") to noninsight solutions. A second analysis compared immediate-insight to noninsight solutions. A third analysis compared delayed-insight to noninsight solutions. The results of splitting insight into two categories reveal large differences between the effects seen in the combined insight versus noninsight analyses and the effects seen in the delayed-insight versus noninsight analyses.

Impasse Five types of Impasse were coded (Regenerating, Rereading, Discontinuing, No-New, and Frustration). Regenerating meant that a person generated the same solution candidate two or more times within a problem. Rereading meant that a person reread the problem words three or more times in succession without generating a solution candidate. Discontinuing meant that the person completely stopped solving the problem and in which no progress toward solution was being made. No-New meant that a person stopped generating new solution words for at least 15 seconds after onset of a problem or at least 10 seconds between candidates. Finally, Frustration meant that a person exhibited clear signs of emotional frustration and experienced real difficulty with the task or specific problem. The individual scores were summed to get a total impasse score. Two independent raters coded the data, and agreement between raters on the number of total impasses per solution was good (Pearson $r = .84$, Kendall's tau = .77).

Three within-subjects repeated measure analyses were performed on total impasse scores per Solution Type. In the first analysis, the amount of total impasse was no different for Noninsight (*M* = .191/second, *SD* = .028) than Combined Insight (*M* = .0105/second, *SD* = .017) problems, $F(1,21) = 1.687$, $p = .208$ (see Figure 2). In the second analysis, there was significantly less impasse for Immediate-insight solutions (*M* = 0, *SD* = 0) than Noninsight solutions, $F(1,21) = 10.11$, $p = .005$. When comparing Delayed-insight and Noninsight solutions, the amount of total impasse for Delayed-insight solutions (*M* = .033/second, *SD* = .038) is marginally greater than that of Noninsight solutions, $F(1,21) = 3.649$, $p = .070$ (see Figure 2). When immediate- and delayed- insight solutions are combined in the analysis there is a higher rate of impasse for noninsight solutions due to the effect of immediate-insight solutions, but when only delayed-insight solutions are included there is a lower rate of impasse for noninsight solutions (see Table 1).

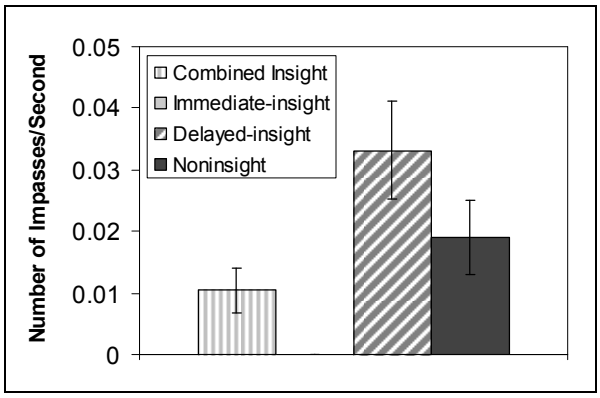


Figure 2: Mean rate of impasse for combined insight, immediate-insight, delayed-insight, and noninsight solutions. Immediate-insight does not show because it is zero. Error bars indicate one standard error.

Table 2: Within-subjects Repeated Measures Analysis for Each Level of Restructuring (n=22).

	Means			
Analysis	Combined Insight	Immediate-insight	Delayed-insight	Noninsight
Re-encoding	0.0174 (0.0199)	0 (0) **	0.0510 (0.0460)**	0.0296 (0.0332)
Elaboration	0.0039 (0.0096)	0 (0) *	0.0123 (0.0330)	0.0036 (0.0083)
Constraint Relaxation	0.0008 (0.0029)	0 (0)	0.0028 (0.0093)	0.0016 (0.0056)
Total Restructuring	0.0221 (0.0209)	0 (0) **	0.0660 (0.0497)**	0.0348 (0.0312)

Notes. Values represent mean number of occurrences per second. Standard deviations are in parentheses. Noninsight means remained the same for all analyses and are compared to each type of insight. Significant effects are represented by stars (*) in the "Insight" columns.
*$p < .10$. **$p < .05$.

Restructuring Restructuring was coded into three different types (Elaboration, Re-encoding, and Constraint Relaxation). Elaboration meant that a person switched to a different meaning of a problem word after trying and failing to find a solution with the first meaning of a word (i.e. star has multiple meanings: starlight or superstar). Re-encoding meant that a person switched to a different problem word to try and find a solution after failing with a previous word. Finally, Constraint Relaxation meant that a person revised the idea of the goal. The person switched the method for solving the problem or clearly stated that they needed to try something different to get a solution. Individual scores were summed to obtain a total restructuring score. Two independent raters coded the data, and agreement between raters on the number of total restructurings per problem was good (Pearson $r = .77$, Kendall's tau = .71).

Three within-subjects repeated measure analyses were performed on total restructuring scores per Solution Type. In the first analysis, the amount of total restructuring was no different for Noninsight solutions ($M = .0348$/second, $SD = .0312$) than Combined Insight solutions ($M = .0221$/second, $SD = .0209$), $F(1,21) = 2.251$, $p = .148$ (see Figure 3). In the second analysis, there is significantly less restructuring for Immediate-insight solutions ($M = 0$, $SD = 0$) than Noninsight solutions, $F(1,21) = 27.47$, $p < .0001$. Again, when comparing Delayed-insight and Noninsight solutions the amount of total restructuring is significantly greater for Delayed-insight solutions ($M = .066$/sec, $SD = .050$) than Noninsight solutions, $F(1,21) = 8.847$, $p = .007$ (see Figure 3). When immediate and delayed-insight solutions are combined in the analysis there is a higher rate of restructuring for noninsight solutions due to the effect of immediate-insight solutions, but when only delayed-insight solutions are included there is a lower rate of restructuring for noninsight solutions (see Table 2).

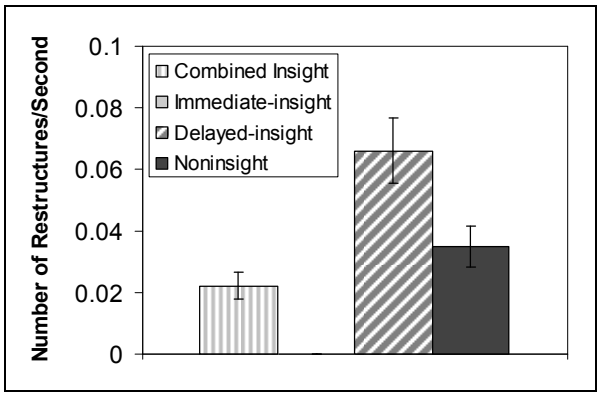


Figure 3: Mean rate of restructuring for combined insight, immediate-insight, delayed-insight, and noninsight solutions. Immediate-insight does not show because it is zero. Error bars indicate one standard error.

Discussion

The purpose of the experiment was to determine if CRA problems can reliably used as insight-like problems. If CRA problems are insight problems, then there should be higher restructuring scores and impasse scores for insight solutions compared to noninsight solutions. Also, according to Schooler et al. (2003), there should be a verbal overshadowing effect.

The results indicate at least some verbal overshadowing effect when solving CRA problems while thinking aloud. Verbalization hindered solution rates and response times for both insight and noninsight solutions. Short-term memory, working memory, and long-term memory retrieval are impacted by concurrent verbalization of processes that are not normally verbalized (Ericsson & Simon, 1993; Schooler et al., 1993). Verbalizing may hinder people from keeping track of where they are and where they are going (in working memory) as well as accessing seemingly distant concepts (in long-term memory). Therefore, both insight and noninsight CRA problem solving methods can be affected by concurrent verbalizations.

The analysis of verbal protocols provided a few unique findings. The results of the impasse and restructuring analyses comparing combined insight and noninsight solutions indicated that noninsight solutions had a slightly higher rate of impasse and restructuring. However, closer inspection of insight solutions revealed that many problems solved with insight were simply the first word that came to mind. The processes used here might not actually be that of insight. People may report insight simply because the answer was sudden. Or, there may have been insight, but that it occurred so quickly that participants were not able to verbalize much before solution. Using people's subjective 'Aha!' experience as a marker of insight might not be a

reliable indicator of insight by itself (or at least it indicates that different problem solving processes can lead to insight).

Further analyses revealed a significant difference in the methods used to solve the problems. The problems solved immediately with insight had no observable characteristics of insight (impasse and restructuring), while the delayed insight solutions had more of the characteristics of insight. When analyzed together insight solutions are derived faster and have significantly less rates of impasse and restructuring than noninsight solutions, but when immediate insight solutions are removed from the analysis the results differ. Delayed-insight solutions are derived in about the same time as noninsight solutions and have significantly more rates of impasse and restructuring than noninsight solutions. There may be two different methods for obtaining insight for CRA problems. Paraphrasing Newell (1973), people may perform a task using different methods and psychologists should take this into account when analyzing data. The effect of averaging over methods "conceals, rather than reveals" (p. 295) any true effect.

From the results, there is some concern that some prior results may have been clouded by averaging the data of the two distinct types of insight (immediate and delayed). For example, neuroimaging results using CRA problems have found more activity in the right anterior superior temporal gyrus, possibly indicating the sudden emergence of the correct solution, that may be facilitated by cognitive control activity prior to problem onset in the dorsal anterior cingulated cortex (Jung-Beeman et al., 2004; Kounios et al., 2006; Subramaniam et al., 2009). However, the results likely include many immediate-insight solutions in the data. There might be different, or additional, areas that are necessary for insight, which are not revealed in prior studies. The areas noted in these prior studies might actually be specific only to immediate-insight and not delayed-insight which often involve restructuring (a staple of the traditional insight definition).

The conclusion drawn is that separating the two types of insight solutions during analysis may reveal different results than prior studies. By pulling apart the two types of insight solutions the processes of insight can be further explored. For example, delayed-insight solutions reveal restructuring elements in the verbal protocols. Immediate-insight solutions do not show observable elements of restructuring. Therefore, the pattern of activation for immediate-insight solutions (and associated processes) may greatly differ from the pattern of activation for delayed-insight solutions. Comparing the solution types may reveal cortical areas necessary for restructuring while eliminating the activation of other common "insight areas." After all, the delayed type insight solutions seem to resemble real world insight more than the immediate type and fit better to the traditional definition of insight. In conclusion, CRA problems can, and should, be used to study insight. However, future work should differentiate immediate- and delayed- insight solutions.

Acknowledgments

We would like to thank Andrew Watkins, Willie Sullivan, Ariel Sibley, and Alesha Lindsey for their help in data collection. Special thanks go to Joshua Liddell for many hours helping to transcribe and code verbal protocols.

References

Bowden, E. M., & Jung-Beeman, M. (2003). Normative data for 144 compound remote associate problems. *Behavioral Research Methods, Instruments, and Computers, 35*, 634-639.

Bowden, E. M., & Jung-Beeman, M. (2007). Methods for investigating the neural components of insight. *Methods*, *42*(1), 87-99.

Bowden, E. M., Jung-Beeman, M., Fleck, J., & Kounios, J. (2005). New approaches to demystifying insight. *Trends in Cognitive Sciences*, *9*(7), 322-328.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data (rev. ed.)*. Cambridge, MA: The MIT Press.

Fleck, J. I., & Weisberg, R. W. (2004). The use of verbal protocols as data: An analysis of insight in the candle problem. *Memory & Cognition*, *32*(6), 990-1006.

Jung-Beeman, M., Bowden, E. M., Haberman, J., Frymiare, J. L., Arambel-Liu, S., Greenblatt, R., et al. (2004). Neural Activity When People Solve Verbal Problems with Insight. *PLoS Biology*, *2*(4), 500-510.

Kounios, J., Frymiare, J. L., Bowden, E. M., Fleck, J. I., Subramaniam, K., Parrish, T. B., et al. (2006). The Prepared Mind: Neural Activity Prior to Problem Presentation Predicts Subsequent Solution by Sudden Insight. *Psychological Science*, *17*(10), 882-890.

Newell, A. (1973). You can't play 20 questions with nature and win. In W. G. Chase (Ed.), *Visual Information Processing* (pp. 283-308).

Ohlsson, S. (1992). Information-processing explanations of insight and related phenomena. In M. T. Keane & K. J. Gilhooly (Eds.), Advances in the psychology of thinking (pp. 1-43). London: Harvester Wheatsheaf.

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). *E-Prime user's guide*. Pittsburgh: Psychology Software Tools Inc.

Schooler, J. W., Fallshore, M., & Fiore, S. M. (1995). Epilogue: Putting insight into perspective. *The nature of insight.*, 559-587.

Schooler, J. W., Ohlsson, S., & Brooks, K. (1993). Thoughts beyond words: When language overshadows insight. *Journal of Experimental Psychology: General*, *122*(2), 166-183.

Subramaniam, K., Kounios, J., Parrish, T. B., & Jung-Beeman, M. (2009). A brain mechanism for facilitation of insight by positive affect. *Journal of Cognitive Neuroscience*, *21*(3), 415-432.

Weisberg, R. W. (1995). Prolegomena to theories of insight in problem solving: A taxonomy of problems. *The nature of insight.*, 157-196.

Writing: The Process of Discovery

Veerle Baaijen (V.M.Baaijen@rug.nl)
Center for Language and Cognition Groningen, University of Groningen
Oude Kijk in 't Jatstraat 26, 9700 AS, Groningen, The Netherlands

David Galbraith (D.Galbraith@staffs.ac.uk)
Centre for Educational Psychology Research, Staffordshire University
College Road, Stoke-on-Trent, ST4 2DE, United Kingdom

Kees de Glopper (C.M.de.Glopper@rug.nl)
Center for Language and Cognition Groningen, University of Groningen
Oude Kijk in 't Jatstraat 26, 9700 AS, Groningen, The Netherlands

Abstract

This paper describes the results of a study investigating the process by which writers develop their understanding through writing. It argues that, contrary to problem-solving models of writing, the crucial ingredient is implicitly guided text production. Two groups of writers, varying in the extent to which their writing is assumed to be directed towards rhetorical goals, were asked to write either planned or non-planned texts. Key-stroke logs were collected, and changes in subjective understanding about the topic were measured. The results show that developments of understanding are strongly related to the extent to which writers modify their texts during writing, and this is highest in the conditions expected to promote implicitly guided text production. We conclude that these findings support a dual-process model of writing.

Keywords: Planning; knowledge change; writing processes; keystroke logging; text production.

Introduction

Writing is an ideal area in which to study the ebb and flow of thought. Although the end product is a fixed knowledge object which has to be comprehensible in the absence of the writer, the process by which it is produced is an extremely dynamic one, in which writers both have to work out what they think about a topic and how best to communicate this to their readers. For this reason, writing is typically characterized as a process of discovery. Bereiter and Scardamalia (1987), for example, characterize expert writing as a knowledge-transforming process, during which writers actively transform their thought in response to their evolving goals, and contrast this with the knowledge-telling process employed by novice writers, in which a fixed store of ideas in long term memory is translated directly into text. They claim that the knowledge-transforming model accounts for the "the peculiar value that many have claimed for writing as a way of developing one's understanding" (Bereiter & Scardamalia, 1987, p. 302). In this paper, we describe the results of an experiment investigating the conditions under which writers develop their understanding and how this is related to a simple indicator of one of the processes involved in writing.

According to Bereiter and Scardamalia's (1987) knowledge-transforming model, discovery through writing is a consequence of rhetorical problem solving. This claim has three important features. The first is an emphasis on the explicit thinking processes involved in the generation and evaluation of content rather than on the processes involved in translating thought into language. Second, and following from this emphasis, the crucial contrast between the knowledge telling and knowledge transforming processes is the goals toward which writing is directed. Thus, in the knowledge-telling model, the goal is to retrieve ideas stored in memory and translate these into text. By contrast, in the knowledge-transforming model, content retrieval and evaluation is mediated by the writer's communicative goals: expert writers develop an elaborate representation of their audience and the rhetorical situation and use this to guide the generation of content. This leads to the re-evaluation of existing content in long term memory and to the formulation of new content. Third, the extent to which writers are able to engage in this reflective evaluation of content depends on how they manage the interaction between high-level thinking processes and the formulation of content in text. Translating processes and higher level thinking processes are assumed to compete for limited cognitive resources, and hence it is assumed that writers will be less able to engage in rhetorical problem-solving the more they try to carry out text production at the same time as generating content. It is this conflict which is assumed to be responsible for the beneficial effects of outlining prior to writing. Kellogg (1988) has provided convincing evidence that outlining is associated with the production of better quality text, and that this is because it enables writers to clearly separate the reflective processes involved in generating, organizing and evaluating ideas from the processes involved in formulating these ideas in well-formed text.

Overall, the knowledge-transforming model and associated research on the benefits of outlining suggests that discovery through writing is a consequence of the strategic modification of content in order to satisfy rhetorical goals, and that this will be enhanced when the writer is able to focus on higher level thinking processes free from the demands of simultaneously formulating full text.

Recently, Galbraith (2009) has questioned this account on empirical grounds. In a series of experiments examining the conditions under which writers develop their understanding, Galbraith and his colleagues have found different patterns of development of understanding through writing than would be expected on the basis of the rhetorical problem-solving model. In brief summary, the essential pattern of their findings is as follows. First, although writers whose writing is assumed to be directed towards rhetorical goals (high self-monitors) do develop more new content after making notes than when they are required to write full text, as would be expected if discovery depended on the extent to which writing was directed towards rhetorical goals, this new content was not associated with increases in writers' subjective understanding of the topic. Second, there was also evidence that writers whose writing was assumed, not to be directed towards rhetorical goals, but rather to be implicitly organized (low self-monitors), developed new content after writing full text, without pre-planning, and that this was associated with developments of subjective understanding. (See Snyder and Gangestad, 1986, for a review of differences between low and high self-monitors.)

On the basis of these experiments, Galbraith has suggested an alternative dual-process account of discovery through writing. This proposes that the development of understanding in writing depends on an interaction between two different kinds of process. The first of these is an explicit planning process. This involves the retrieval of content from an explicit store of ideas and the goal-directed manipulation of these ideas in working memory designed to create a coherent knowledge object that satisfies rhetorical goals. This is essentially equivalent to the knowledge transforming model of Bereiter and Scardamalia, with the crucial difference that, by itself, this process only involves the reorganization of existing knowledge and is not associated with developments of understanding. The second is an implicit text production process. This operates on an implicit store of conceptual knowledge in semantic memory, which Galbraith defines as the writer's disposition towards the topic, and involves synthesizing content during text production. The key features of this process, for present purposes, are that it is engaged when writers have to formulate their thought in explicit propositions, and that, because the process is guided by the implicit organization of material in semantic memory, the sequence in which content is produced is unpredictable. Content is synthesized in the course of formulation rather than being directly retrieved from memory and translated into text. This process is assumed to lead to developments of understanding when the content it produces is different from the explicit content stored in episodic memory.

The model suggests that the implicit text production process will be at a maximum when writing is (i) dispositionally guided, i.e. for low self-monitors, and (ii) not outline planned, i.e. the order in which content is produced is governed by the implicit organization of content in semantic memory rather than by an explicit, pre-determined plan in working memory. The implicit text production process will be minimized when writing is (i) directed towards rhetorical goals, i.e. for high self-monitors, and (ii) controlled by an outline, i.e. when the sequence of text production is pre-determined. Furthermore, it suggests that, because changes in content can be induced by both explicit planning and implicit text production, but only implicit text production leads to the development of understanding, there will be no direct relationship between the overall amount of change in content and the development of understanding. Instead, the development of understanding will be directly linked to the extent that new content is produced by the implicit text production process.

This experiment set out to test these claims by using key-stroke logging to provide a direct measure of the extent to which content was modified during the course of text production, and examined how this varied depending on the conditions under which writing took place, and how it was related to developments in the writer's personal understanding of the topic. The present paper will report the results for a simple indicator of content modification during text production, which we will label as the *text modification index.* This corresponds to the total number of words recorded by key-stroke logging divided by the total number of words appearing in the final text. When writers transcribe their thoughts directly into text the index should be 1: all the words that are written down during text production will be retained in the final text. To the extent that the writer changes the way that they express their ideas during text production the index should increase: writers will produce more words during the process of text production than appear in the final text.

The design of the experiment was based on a previous experiment by Galbraith, Torrance and Hallam (2006) and manipulated two variables: self-monitoring and planning. Each group was asked either to make an outline before writing or to sum up their overall opinion of the topic prior to writing (a procedure we call synthetic planning, and which differs from outline planning in that it does not specify the order in which content should be produced during text production.). Our aim was to replicate the conditions of this earlier experiment with a view to assessing how the text modification index varied under these conditions. We expected that, if the dual-process model is correct, the text modification index should be at a maximum when low self-monitors produce synthetically planned texts, and that increases in subjective understanding should be associated with high levels of text modification, rather than with the overall amount of change in content produced in the different conditions.

Method

Participants

84 students from the faculty of Arts of the University of Groningen were recruited to participate in the experiment. They were all native Dutch speakers, average age 22.2 years

(*SD* = 3.8), and were pre-selected using Snyder's revised 18 item self-monitoring scale (Snyder & Gangestad, 1986). Participants could only take part if they were classified either as a high or a low self-monitor. They were classified as *high self-monitors* (HSM, *n* = 42) if they scored 11-18 on the scale and as *low self-monitors* (LSM, *n* = 42) if they scored 0 - 8 on the scale.

Design and procedure

High and low self-monitors were randomly allocated to the two planning conditions resulting in the following four experimental groups: (i) HSM outline planning, (ii) HSM synthetic planning, (iii) LSM outline planning and (iv) LSM synthetic planning.

Writing task In all four conditions, participants were asked to plan and write an article for the university newspaper discussing whether "our growing dependence on computers and the Internet is a good development or not". The writing task was divided into three phases.

In phase 1, participants were first given 10 minutes to list all the ideas they could think of relevant to the topic. It was stressed that each idea should be no longer than a sentence in length. They were then asked to rate how much they felt they knew about the topic on a 7-point scale.

In phase 2, participants were given 5 minutes to either write down a single sentence summing up their overall opinion (synthetic planning) or to construct a structured outline (outline planning). They were then given 30 minutes to write a well-structured article for the university newspaper. It was stressed that they had to produce a reasoned argument reflecting their own opinion about the matter. Participants were allowed to consult their written outlines. During writing, keystrokes were logged using Inputlog (Leijten & Van Waes, 2006).

In phase 3, immediately after writing, participants were asked again to rate how much they felt they knew about the topic. They were then given 10 minutes to again list all the ideas they could think of relevant to the topic. Finally, they were asked to compare the lists produced before and after writing, and to rate the extent that ideas on list 2 corresponded with ideas on list 1, using a 6-point scale ranging from 1=identical point to 6=no correspondence.

Measures

Subjective development of understanding The ratings of knowledge were used to assess subjective changes in understanding as a consequence of writing.

Development of ideas This was assessed using the procedure used in previous research. New ideas were defined as ideas in the second list that received ratings from 4 to 6 for their correspondence with ideas in the first list. Preserved ideas were defined as ideas in the second list that received ratings from 1 to 3 for their correspondence with ideas in list 1. The average length of these ideas was also calculated. These were assessed against baseline measures of the number and average length of ideas in the list produced before writing.

Text modification index In order to assess the process by which writing is carried out a text modification index was calculated. For the text modification index the total number of words recorded by Inputlog are divided by the number of words appearing in the final text.

Data screening Preliminary analysis of the data revealed 6 outliers (i.e. scores more than 3 SD's above or below the mean). Three participants had extremely low scores on the initial or post knowledge rating. One had an extremely high score on the mean length of ideas. Two had extremely high scores on the text modification index. These participants were removed from all analyses.

Results

Development of subjective understanding

A two-way (2*2) between subjects ANCOVA with self-monitoring and planning as factors and with prior knowledge as a covariate revealed a significant main effect of type of planning on subjective understanding after writing (F (1,73) = 4.61, p = .035, η^2 =.033). Figure 1 shows the mean ratings of knowledge before and after writing in each condition (with error bars showing standard errors).

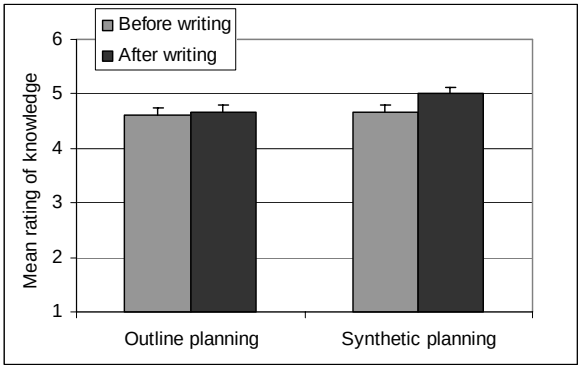


Figure 1: Development of subjective understanding as a function of type of planning.

Planned comparisons comparing mean knowledge ratings before and after writing in the synthetic and outline planned conditions showed that there was a significant increase in knowledge in the synthetic planned condition (t (39) = 3.34, p = .002) but no significant difference in the outline planned condition (t (37) = 0.47, p = .64).

Effects on idea change and relationships with developments of subjective understanding

To assess the relationship between changes in the content of the lists produced before and after writing and changes in subjective understanding, we converted the knowledge

ratings to a category variable representing the extent to which knowledge increased, decreased or remained the same. We then carried out a 3-way between subjects MANCOVA, with self-monitoring, type of planning and change in knowledge as independent variables; the number of new and preserved ideas, and the average length of these ideas, as dependent variables; and the number of ideas produced before writing and their average length as covariates. Using Pillai's trace, this showed a significant main effect of type of planning (V = .14, F (4, 60) = 2.51, p = .05) and a significant interaction between type of planning and knowledge change (V = .32, F (8, 122) = 2.85, p = .006). To describe these effects, we will consider them in two stages, starting with the main effect of type of planning and then considering the interaction between type of planning and change in knowledge.

Main effect of type of planning There were two important findings here. First, as can be seen in figure 2, the preserved ideas were significantly reduced in length in the outline planning condition but not in the synthetic planning condition (F (1, 66) = 5.80, p =.019, η^2 =.05). There was no equivalent effect for the new ideas.

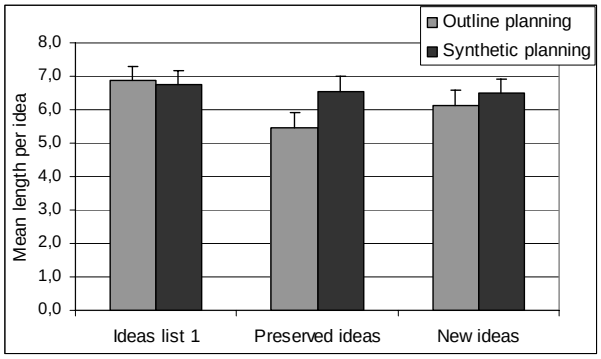


Figure 2: Words per idea for ideas in list 1, preserved ideas in list 2 and new ideas in list 2.

A possible explanation for the effect is that when an outline is constructed it is held in working memory to guide text production. In consequence, when writers refer to ideas in the outline, they label the idea held in memory in an abbreviated form. Although this effect may prove useful as a marker of the extent to which individuals within different conditions construct a mental outline during writing, it does not suggest a substantive effect of type of planning on the content of the lists produced after writing.

The second important finding here is a negative one. The follow-up analysis of the multivariate analysis revealed no apparent effect of either self-monitoring or type of planning on the number of new or preserved ideas produced after writing. Possible reasons for this will be considered in the discussion.

Interaction between type of planning and knowledge change To determine the source of this effect, we carried out simple effects analysis within the synthetic planning and outline planning conditions, using 1-way MANCOVAs, with change in knowledge as the independent variable, the four measures of the lists produced after writing as dependent variables, and the number of ideas in the initial list and their average length as covariates. This confirmed that there was no significant relationship between idea change and changes in subjective knowledge within the outline planning condition (V = .191, F (8, 54) =.71 p =.68). However, there was a highly significant effect of within the synthetic planning condition (V =.645, F (8, 60) = 3.57, p = .002). Univariate ANOVAs, followed by planned comparisons, on each of the dependent variables showed that there were significant effects for 3 of the variables.

First, there was a significant effect on new ideas (F (2, 34) = 6.25. p =.005, η^2=.25), with planned comparisons showing that participants whose knowledge remained the same produced more new ideas (M = 7.8, se = 0.68) than both participants whose knowledge decreased (M = 1.69, se = 2.34, p =.05) and participants whose knowledge increased (M = 4.85, se =0.91, p =.045). Although increased knowledge was associated with more new ideas than decreased knowledge, this difference was not significant (p = .65).

There was also a significant effect on the average length of new ideas (F (2, 33) = 5.94, p =.006, η^2=.14) with participants whose knowledge decreased producing longer new ideas than those whose knowledge remained the same (p =.008) and those whose knowledge increased (p =.04). Finally, there was a marginally significant effect on the average length of preserved ideas (F (2, 34) = 2.46, p =.10, η^2=.04), with a tendency for participants whose knowledge remained the same to produce preserved ideas shorter in length than those produced by participants whose knowledge either increased or decreased.

Taken together, these findings suggest, first, that decreases in knowledge occurred in this condition when writers were relatively unable to think of new ideas, and to express what ideas they could think of concisely. This implies that thinking of new content is generally necessary in order to produce satisfactory text. Second, increases in knowledge were associated with the production of fewer new ideas than when knowledge stayed the same. This contradicts previous research. A possible explanation for this is that new ideas were produced by different processes when knowledge remained the same than when it increased. On the assumption that the length of preserved ideas reflects the extent to which writing has been controlled by an outline (see above), then the marginally significant effect on the length of preserved ideas could indicate that texts where knowledge remained the same were relatively more outline planned than the texts where knowledge increased.

Relationship with processes

The preceding analysis revealed that, despite the significant difference between synthetic and outline planning in the extent to which writers reported increases in understanding,

there were no differences in idea change within the two planning conditions, and generally that there were no relationships between the amount of change in ideas and increased knowledge. According to the dual-process model, this is because new content is produced by two different processes -explicit rhetorical planning and implicitly guided text production- and only implicit text production leads to the development of understanding. To test these claims, we carried out a 3-way between subjects ANOVA on the text modification index, with self-monitoring, type of planning and knowledge change as dependent variables. This produced clear evidence to support these claims.

First, both self-monitoring and type of planning had a clear effect on the extent to which ideas were modified during text production. There was a significant main effect of type of planning (F (1, 66) =5.55, p= .02, η^2=.06), a close to significant main effect of self-monitoring (F (1, 66) = 4.53, p =.06, η^2=.03) and a significant interaction between self-monitoring and type of planning (F (1, 66) = 4.45, p =.04, η^2=.04). Figure 3 shows that low self-monitors produced higher levels on the text modification index than high self-monitors and this was reduced when text production was preceded by outline planning.

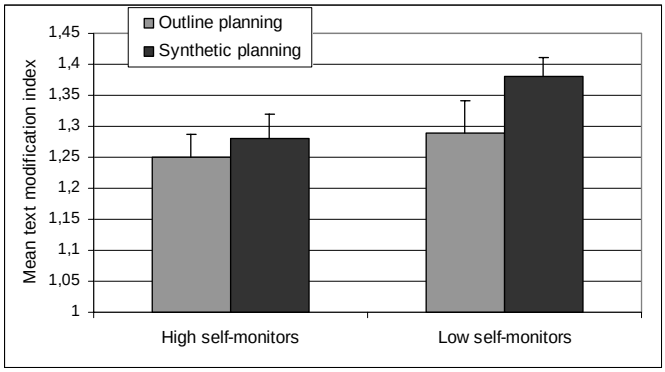


Figure 3: The text modification index as a function of type of planning and self-monitoring.

Second, there was a significant interaction between knowledge change and type of planning (F (2, 66) = 3.67, p =.03, η^2=.07). Analysis of simple effects, followed by planned comparisons of the differences between different types of knowledge change, revealed that there was a highly significant main effect of knowledge change within the synthetic planning condition (F (2, 34) = 5.59, p = .008, η^2=.22). As can be seen in figure 4, this was a consequence of the fact that increases in knowledge were associated with significantly higher levels of text modification than when knowledge remained the same (t (34) = 3.29, p = .007). Although decreased knowledge was also associated with lightly elevated levels of text modification, this was not significantly different from the other conditions.

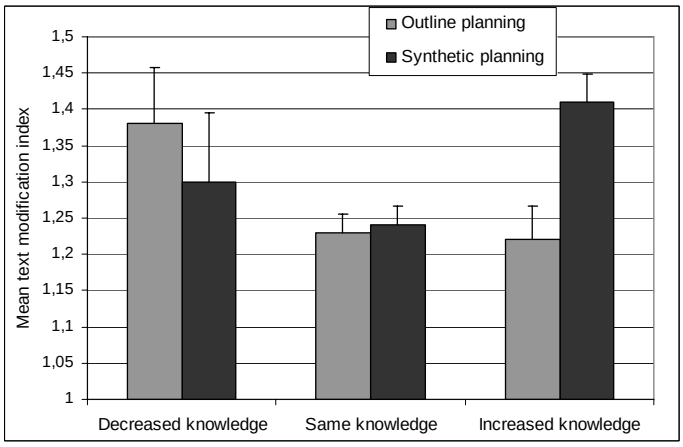


Figure 4: The text modification index as a function of type of planning and knowledge change.

Discussion

The dual-process model claims that new content is produced during writing by two different kinds of process: explicitly controlled planning to satisfy rhetorical goals and implicitly guided text production articulating the writer's developing understanding. This contrasts with the knowledge-transforming model in two key respects. First, it claims that, although explicitly controlled planning does lead to changes in content after writing, this is essentially a matter of retrieving already existing knowledge which is more appropriate to the rhetorical context than the ideas initially considered relevant to the topic, and hence that changes in content produced by explicit planning will not lead to developments in understanding. Second, it claims that implicitly guided text production is not simply a matter of translating the output of planning into words, but is an active knowledge-constituting process in its own right. Our results provide strong support for both claims.

First, there was clear evidence that content was produced by different processes in the outline planned and synthetically planned conditions. The outline planned condition involved significantly lower levels of text modification during writing than the synthetically planned condition. This is compatible with the claim that changes in content in this condition are a consequence of higher level thinking processes rather than of the modification of content in the formulation of the text itself. By contrast, text modification was at its highest in the condition – the low-self-monitors' synthetically planned texts – where the dual process-model assumes that text production is most implicitly guided, and where new content is assumed to be formulated in the text itself rather than through planning prior to text production.

Second, although both conditions led to a similar amount of change in ideas, as would be expected if both processes play an active role in developing content during writing, only the synthetic planning condition was associated with significant increases in subjective ratings of understanding.

This clearly supports the claim that explicit planning is less strongly associated with the development of understanding than implicitly guided text production is.

Third, there was no relationship between the amount of change in content in the different conditions and increases in subjective understanding. The dual-process model provides a straightforward explanation for this: increases in understanding depend on the extent to which new content is produced by implicitly guided text production. This explanation is strongly supported by the fact that synthetically planned writing involved significantly higher levels of text modification, and that it was precisely those writers within this condition whose understanding increased who produced the highest levels of text modification. The only exception to this extremely clear pattern was that the few writers who experienced decreases in knowledge in any of the conditions also appeared to engage in relatively high levels of text modification. The important feature of these writers, however, is that they also produced few new ideas. This leads to the general conclusion that increases in understanding occur when writers develop new ideas in the course of formulating the text itself. Understanding will remain the same when text production is either controlled to conform to a higher level plan (as in outline planning) or when the writer's knowledge prior to writing is sufficiently clear for text to be fluently produced. Understanding will decrease when text production does not lead to the formulation of coherent new content.

There is one aspect of these results which does not match previous research. Previous studies (see Galbraith, 2009) have found that low self-monitors typically produce more new ideas than high self-monitors under synthetic planning conditions, and that, under these conditions, the number of new ideas is positively correlated with increases in subjective understanding. The dual process model assumes that this is because high self-monitors typically impose more control on text production than low self-monitors, so reducing the extent to which ideas are formulated during text production. This was partially supported in the present experiment in that high self-monitors did engage in significantly less text modification than low self-monitors in the synthetic planning condition. However, there no was difference in the extent to which low and high self-monitors produced new ideas in this condition, and there was a negative rather than a positive relationship between the number of new ideas and increases in understanding.

We believe that this is a consequence of a difference in the constraints under which synthetic planning was carried out in this experiment. In previous research, the external constraints for the text under synthetic planning conditions have either been left unspecified or writers have been actively instructed to write down their thought free from rhetorical constraints. By contrast in this experiment, writers were instructed to produce a finished article for the university newspaper in the time available. According to the dual-process model, this should lead to an increase in the extent of explicit planning processes, and since these are prioritized by high self-monitors, a greater increase in the number of new ideas produced by high self-monitors compared to low self-monitors. Furthermore, since these new ideas are produced by explicit planning, which according to the dual-process model is not associated with changes in understanding, there will no longer be a straightforward relationship between the amount of new ideas and increases in subjective understanding, just as we found in this experiment. This explanation could be tested by comparing low and high self-monitors writing synthetically planned text, either with clear rhetorical constraints, as in the present experiment, or free from rhetorical constraints as in previous research.

Our general conclusion is that in order to explain how writers develop their understanding it is necessary to examine the processes by which their ideas are created rather than just assess the extent to which they have modified their beliefs. The simple index of text modification that we have used in this paper has shown clear distinctions between different kinds of knowledge change, which strongly supports the broad claim that the development of thought during writing depends on two different kinds of process. Further research is needed, using on-line measures such as key-stroke logging, to examine in detail how ideas are formulated during text production and how this results in developments of the writer's understanding.

Acknowledgements

We would like to thank Professor Kees de Bot for his helpful comments on earlier drafts of this paper.

References

Bereiter, C., & Scardamalia, M. (1987). *The psychology of written composition*. Hillsdale, NJ: Lawrence Erlbaum Associates Publishers.

Galbraith, D. (2009). Writing as discovery. *British Journal of Educational Psychology Monograph Series II 6 - Teaching and Learning Writing*, 5-26.

Galbraith, D., Torrance, M., & Hallam, J. (2006). Effects of writing on conceptual coherence. *Proceedings of the 28th Annual Conference of the Cognitive Science Society,* 1340–1345.

Kellogg, R.T. (1988). Attentional overload and writing performance: Effects of rough draft and outline strategies. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14 (2), 355-365.

Leijten, M., & Van Waes, L. (2006). Inputlog: New perspectives on the logging of online writing. In K.P.H. Sullivan & E. Lindgren (Eds.), *Computer Key-Stroke Logging and Writing: Methods and Applications*. Oxford: Elsevier.

Snyder, M., & Gangestad, S.W. (1986). On the nature of self-monitoring. *Journal of Personality and Social Psychology*, 51, 125-139.

Number Representations and their Development: A Connectionist Model of Number Comparison

Mark Rose Lewis (lewis505@umn.edu)
Department of Educational Psychology, 56 East River Road
Minneapolis, MN 55455

Sashank Varma (sashank@umn.edu)
Department of Educational Psychology, 56 East River Road
Minneapolis, MN 55455

Abstract

Building on prior work, the current study evaluated whether connectionist models can account for the distance and size effects in adults and the development of the distance effect in children. A family of models was constructed by orthogonally varying training environment (naturalistic versus non-naturalistic) and number representation (one-to-one versus magnitude). The ability of the models to account for the adult distance and size effects depended critically on a naturalistic training environment but was relatively independent of number representation. With respect to the developmental data, the naturalistic/one-to-one model provided a good account of response times and errors. The relation between the current models and prior models and avenues for future exploration are discussed.

Keywords: number comparison; distance effect; size effect; connectionism; models; development

Introduction

The nature of number representations is an enduring question in cognitive science. One clue to this representation is the *distance effect*: the time it takes to judge the greater (or lesser) of two numbers decreases with the distance between the numbers (Moyer & Landauer, 1967). For example, 1 vs. 9 is judged faster than 1 vs. 3. Another clue is the *size effect*: the time to judge the greater (or lesser) of two numbers that are a fixed distance apart increases with the absolute magnitude of the numbers (Parkman, 1971). For example, 7 vs. 9 is judged more slowly than 1 vs. 3. The distance and size effects conform to psychophysical laws (i.e., $RT = Klog(\frac{larger}{larger-smaller})$) and are therefore commonly interpreted as evidence that numbers are represented as analog representations, perhaps localized to the intra-parietal sulcus (Dehaene, Piazza, Pinel, & Cohen, 2003). Researchers have proposed various implementations of these analog representations. The classic ones are as points on a compressed mental number line (Dehaene & Mehler, 1992; Rule, 1969) and as points on a linear mental number line associated with increasing variability (e.g., Gallistel & Gelman, 2000). More recently, two connectionist models of number representation have appeared. Zorzi and Butterworth (1999) assumed magnitude representations whereby numbers are represented by banks of overlapping units. This model was able to account for the adult distance effect. By contrast, Verguts, Fias, and Steven (2005) assumed a coarse-coded representation, with each number corresponding primarily to one unit, but with graded activation of adjacent units. This model was able to account for the adult distance and size effects.

The purpose of the current study was to evaluate the ability of connectionist models to (1) account for the adult distance and size effects as a function of training environment and number representation and to (2) account for the development of the distance effect. In these regards, the reported simulations are the first of their kind.

With respect to training environment, some connectionist models (Zorzi & Butterworth, 1999) have employed a *non-naturalistic* training environment (i.e., every one-digit number appears with equal likelihood). However, corpus studies indicate that the frequency of a number falls off as a power function of its magnitude (Dehaene & Mehler, 1992), implying that one-digit numbers are non-uniformly distributed in a *naturalistic* environment. Some connectionist models have employed a naturalistic training environment (Verguts et al., 2005). We sampled comparisons (i.e., pairs of one-digit numbers) from these contrasting training environments to evaluate whether the distance and size effects were contingent upon naturalistic input.

With respect to number representation, we considered the magnitude representation implemented by the Zorzi and Butterworth (1999) model and a one-to-one variant of the coarse-coded representation implemented by the Verguts et al. (2005) model[1].

Finally, in the first study to model the development of the distance effect, we evaluated whether improvements in model performance throughout training parallel improvements in children's response times and error rates throughout development.

Method

We developed four connectionist models by orthogonally varying training environment (naturalistic versus non-naturalistic) and number representation (magnitude versus one-to-one). The models were implemented within a common connectionist architecture patterned after Verguts et al. (2005).

[1] Both of these codings represent exact numbers. We use the label "magnitude" to reflect the fact that the number of representation nodes activated in this coding corresponds to the number being compared.

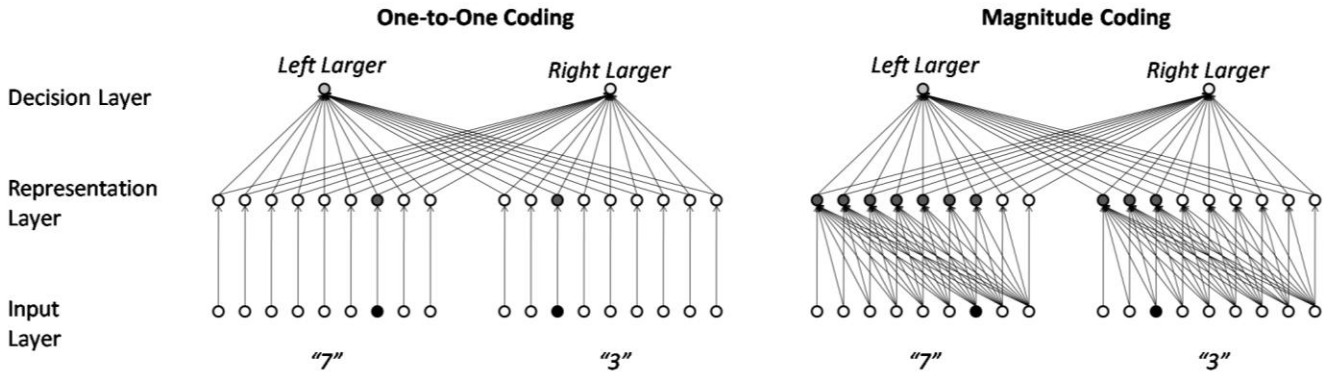


Figure 1: Schematic of models using one-to-one coding (left) and magnitude coding (right).

This architecture consisted of three layers of units (*input*, *representation*, and *decision* layers) (Figure 1). Each layer contained left and right fields. On each trial, the input units corresponding to the numbers being compared were clamped to an activation level of 1, and activation spread forward throughout the network. When a decision unit (*left larger* or *right larger*) reached an activation of 0.5 or greater, the model was considered to have made a decision

Architecture and Number Representation

Each model consisted of three layers of units. The input layer consisted of two fields of nine units each that corresponded to the numbers 1-9. The left field corresponded to the number presented on the left and the right field to the number on the right. Each number corresponded to one (and only one) unit in the input layer.

The representation layer consisted of two sets of nine units. The left field, *M*, represented the number presented on the left, and the right field, *N*, represented the number presented on the right. The left input field was connected to the left representation field and the right input field to the right representation field by connections with weights 0 or 1. The number representation scheme of the model determined the pattern of connection weights between the input and representation layers. For magnitude representations, the number of units activated in a representation field corresponded to the magnitude of the number presented (*e.g.*, if the number 5 was presented on the left, the 5 leftmost units of the left representation field would be activated). For one-to-one representations[2], one (and only one) unit in a representation corresponded to the number presented. The weights of the connections between the input and representation layers were held constant throughout learning to maintain the type of representation that the model *a priori* employed.

The decision layer consisted of two units representing *left larger* and *right larger* decisions. Units in the representation layer were fully connected with units in the decision layer. The initial weights of these connections were randomly sampled from a uniform distribution (0 to 1) and were adjusted during training by a supervised learning rule.

Model Dynamics

On each trial, the model compared two numbers, judging which was greater. (Following prior work, we did not model both greater and lesser judgments.) The left number was presented to the left input field by clamping the activation of the corresponding unit to 1, and the right number was presented similarly to the right input field. Activation spread from the input layer to the representation layer according to the equation[3]:

$$(1)\ \Delta r_{Mk}(t) = r_{Mi}(t-1) + \sum_{i=1}^{i=9} w_{Mkin_i} \lceil in_i(t) - \theta \rceil^+$$

Where $\Delta r_{Mk}(t)$ is the change in the activation of the kth representation unit in the left field (*M*), $in_i(t)$ is the activation of the ith input unit, w_{Mkin_i} is the weight of the connection between these two units, and θ is a firing threshold (set to .08 for these simulations). This equation results in the activation of representation units asymptotically approaching their maximum values.

Activation spread from the representation layer to the left-larger unit of the output layer according to the equation:

$$(2)\ \Delta o_{Left}(t) = o_{Left}(t-1) + \sum_{i=1}^{i=9} w_{Mi,Left} \lceil r_{Mi}(t) - \theta \rceil^+ + \sum_{i=1}^{i=9} w_{Ni,Left} \lceil r_{Ni}(t) - \theta \rceil^+$$

Where $\Delta o_{Left}(t)$ is the change in the activation of the left-larger unit, $r_{Mi}(t)$ is the activation of ith representation unit

[2] We employed one-to-one representations instead of coarse-coded representations (Verguts et al., 2005) to equate the architecture across models. Coarse-coding would have required adding additional units to the representation layer of models that employed magnitude representations, muddying the comparison of the models.

[3] All equations are for left fields. Equivalent equations governed model dynamics in the right fields.

in the left field, $w_{Mi,Left}$ is the weight of the connection between these two units, $r_{Ni}(t)$ is the activation of the ith representation unit in the right field, and θ is the firing threshold. This equation results in the activation of decision units asymptotically approaching their maximum values once the representation units have reached the firing threshold. A decision was considered made once activation in one of the decision units exceeds a threshold of 0.5.

Supervised Learning

During learning, weights between representation and decision units were adjusted according to the delta rule:

$$(3)\ \Delta w_{ik} = \epsilon(t_k - o_k)r_i$$

Where Δw_{ik} is the change in the weight between the ith representation unit and the kth decision unit, ϵ is a learning rate parameter, $(t_k - o_k)$ is the difference between the target decision unit activation t_k (1 for larger, 0 for smaller) and the actual decision unit activation o_k, and r_i is the activation of the ith representation unit. The delta rule apportions blame for incorrect decisions and adjusts weights accordingly. For this study, the learning rate parameter ϵ was set to 0.02. During learning, activation was allowed to settle prior to weight adjustment. Each model was trained for 30,000 trials, and weights were adjusted at the end of every trial.

Training Environment

Models were trained on one of two training environments. Naturalistic training environments were constructed by assuming, following Dehaene and Mehler (1992), that the frequency of a number in the environment is a decreasing function of its magnitude. Although Dehaene and Mehler favored a power function, Verguts et al. (2005) adopted a closely related exponential function. To facilitate the comparison of our results, we formed training comparisons by sampling pairs of numbers from an exponentially decreasing distribution (where the frequency of number *i* is $e^{-0.2i}$). The distribution of individual numbers and of comparisons (as a function of distance) is shown in Figure 2.

Non-naturalistic training environments were constructed by assuming that numbers are distributed uniformly in the environment. Training comparisons were formed by sampling from this distribution. The results are also shown in Figure 2.

It is interesting that naturalistic and non-naturalistic training environments result in strikingly similar distributions of comparisons as a function of distance. However, as we shall see, these environments have important differences as indicated by the ability of the resulting models to account for the adult distance and size effects.

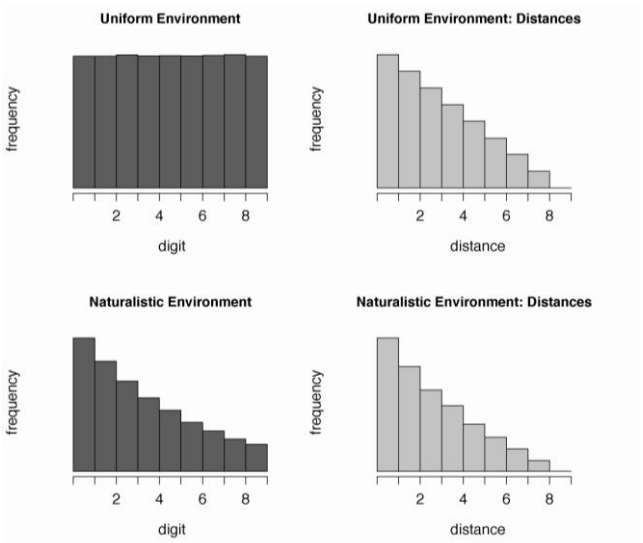


Figure 2: Training environments. Dark gray histograms give the distribution of numbers in the environment, light gray histograms the distribution of distances between the resulting comparisons (i.e., number pairs).

Training and Testing

Ten copies of each model (naturalistic versus non-naturalistic crossed with magnitude versus one-to-one) were created. Each copy was trained for 30,000 trials (following Verguts et al.) and tested with all possible pairs of numbers between 1 and 9 (excluding ties).

Results

Distance Effects

All four models produced distance effects (Figure 3).

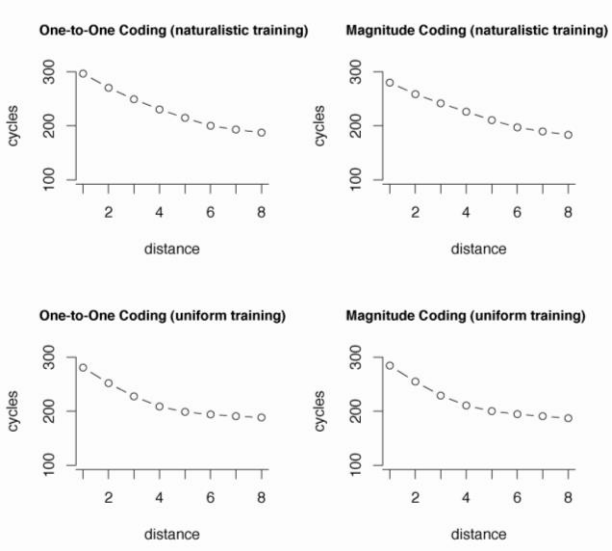


Figure 3: Distance effects for the four models.

Table 1: Model fits for the distance effect.

Representation	Training	R^2	p
Magnitude	Naturalistic	.73	< .001
Magnitude	Uniform	.43	< .001
One-to-One	Naturalistic	.78	< .001
One-to-One	Uniform	.46	< .001

To evaluate the fit of each model to human performance, we followed Zorzi and Butterworth (1999) in regressing human performance as captured by the equation:

$$\widehat{RT} = Klog\left(\frac{larger}{larger-smaller}\right) + C$$

against the number of cycles to make a decision. The results are shown in Table 1.

First, consider the question of training environment. The results indicate that models trained in naturalistic training environments provide better accounts of the distance effect than models trained in non-naturalistic environments. Although Figure 2 suggests that the difference between the training environments is negligible with respect to the amount of experience with different distances, the fit statistics indicate that differences between uniform and naturalistic environments are critical to the distance effect .

Next, consider the question of number representation. The results indicate that a model's ability to account for the distance effect is independent of whether it uses magnitude or one-to-one number representations. Additional work is necessary to determine how fundamentally different types of numerical coding can produce such similar results with respect to the distance effect.

Size Effect

The size effects produced by the four models are shown in Figure 4. There is a striking qualitative difference in the performance of models trained with naturalistic versus non-naturalistic training environments.[4] The former produce a generally positive linear relation between number size and judgment time, with the exception of distances 1-2. By contrast, the latter shows a size effect only for distances 5-8, and diverge considerably from a linear relation for distances 1-4. Additional modeling is necessary to determine what factors contribute to the failure of the uniformly-trained models to produce size effects for distances 1 and 2.

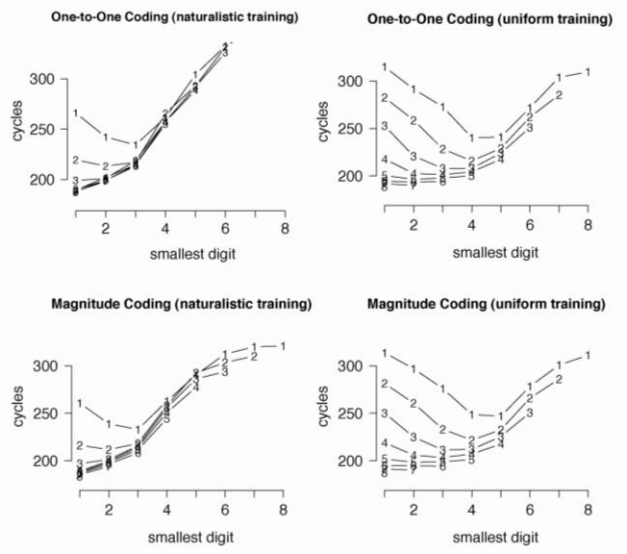


Figure 4: Size effects for the four models. Each line represents comparisons of the same distance.

By contrast, the ability of a model to account for the size effect appears to be relatively independent of whether it uses a magnitude or one-to-one number representation. As with the distance effect, additional work is necessary to determine how fundamentally different types of number representation can produce such similar results with respect to the size effect.

Development of the Distance Effect

We next turn to the development of the distance effect. The results thus far indicate that naturalistic training environments are critical for accounting for adult distance and size effects. Additionally, pilot simulations indicated that models that utilize magnitude number representations do not produce enough errors to account for that dimension of development. For these reasons, we focused our developmental efforts on the naturalistic/one-to-one model.

Sekuler and Mierkiewicz (1977) investigated the distance effect in kindergarten, first grade, fourth grade, seventh grade and adult subjects. Their results are shown in Figure 5. They reported that kindergarteners were significantly slower than other ages, first graders were significantly slower than all age groups except kindergarteners, and that the decision times of fourth graders, seventh graders and adults did not differ significantly. They also reported that the slope of the distance response curves was steeper for kindergarteners than other age groups.

Figure 6 presents the distance effect (averaged across 10 simulations) produced by the naturalistic/one-to-one model after 1200, 1600, 2000, 2400, and 2800 trials[5]. The model provides a nice qualitative account of the developmental data, showing distance effects at all time points as well as a steady decrease in response time.

[4] At the time of submission, we did not have access to empirical data on the size effect to quantify these models fits. We are working on gaining such access.

[5] These time points were chosen to align model-produced error rates with the error rates reported by Sekuler and Mierkiewicz.

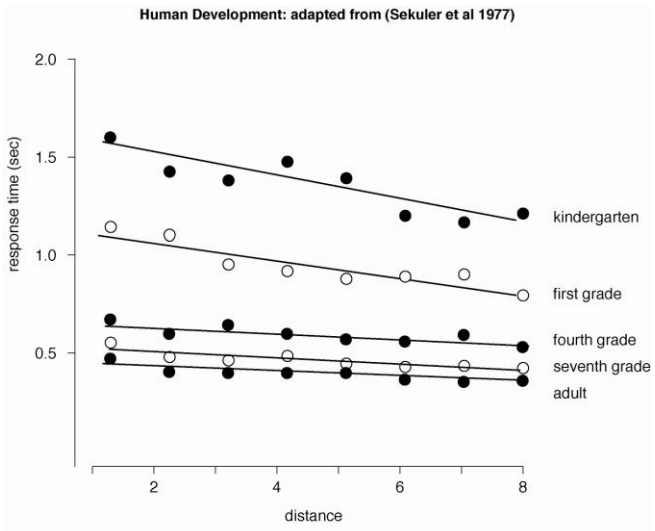


Figure 5: Development of distance effect from kindergarten to adulthood (adapted from Sekuler & Mierkiewicz, 1977).

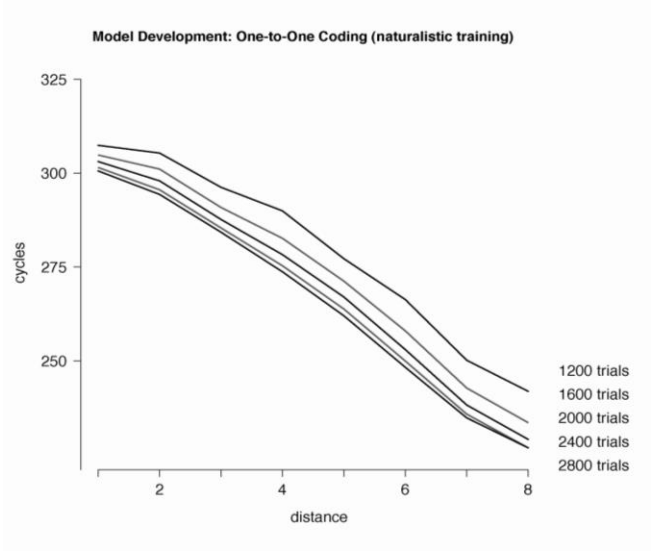


Figure 6: Development of distance effect for the naturalistic/one-to-one model from 1200 to 2800 trials)

However, the model fails to capture the interaction reported by Sekuler and Mierkiewicz : the slope of the 1200 trial line (corresponding to the kindergarten distance effect) is not qualitatively steeper than the slope of the 2800 trial line (corresponding to the adult distance effect).

We were unable to evaluate the quantitative fit of the model to the Sekuler and Mierkiewicz (1977) response time data because it is no longer available. However, Holloway and Ansari (2008) recently performed a similar experiment.[6] They had six, seven, and eight year old children make comparisons at distances 1-6. Their results are shown in Figure 7.

[6] We thank Daniel Ansari and Ian Holloway for sharing their data with us.

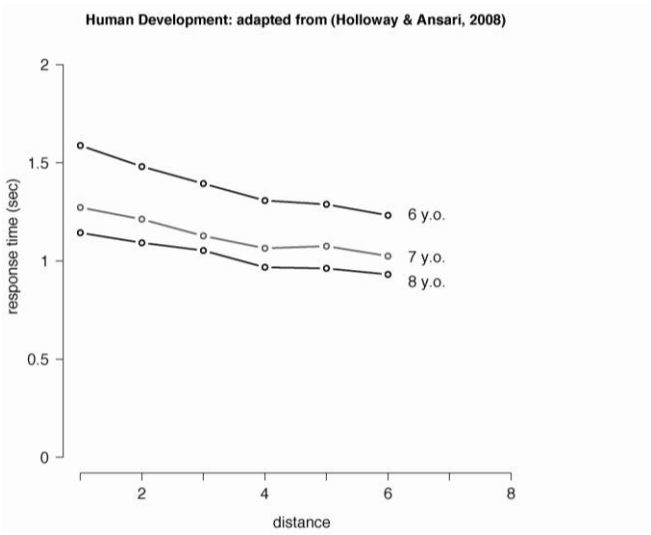


Figure 7: Distance effects at age 6, 7, and 8 years old (data from Holloway &Ansari, 2008).

We linearly regressed the performance of the model at 1200, 1600, and 2000 trials against their six, seven, and eight year old data, respectively. The model accounted for 44% of the variance in the data ($R^2 = .441,\ p = .003$).

Although we were unable to evaluate the quantitative fit of the model to Sekuler and Mierkiewicz's response time data because it was not available, we were able to evaluate the fit of the model to their error rate data because it was reported numerically in the original article. Table 2 presents their developmental error rate data and the error rates of our model. The model provides a good quantitative account of error rate as a function of age. The correlation between the model and the human data is 0.97 (p = .004).

Table 2: Error rates for the developmental simulations of the naturalistic/one-to-one model.

Human (Age)	Errors (%)	Model (Trials)	Errors (%)
Kindergarten	18.4	1200	18.3
First Grade	16.7	1600	15.1
Fourth Grade	11.8	2000	12.8
Seventh Grade	12.5	2400	12.1
Adult	7.9	2800	8.3

Discussion

The current study extends prior connectionist efforts to understand the distance and size effects. We systematically varied training environment and number representation and examined the effects on the adult distance and size effects. Models trained in naturalistic training environments, where the frequency of numbers falls off as a function of their absolute magnitude, provide better quantitative accounts of the distance effect and better qualitative accounts of the size effect. By contrast, the choice of number representation had little effect on these models' ability to account for the adult distance and size effects.

The current study is the first to address the development of the distance effect. The naturalistic/one-to-one model provided a good qualitative account of distance effects at different ages. It also provided a good account of decreasing error rates with development.

One limitation of the development simulation was that it did not account for the interaction observed by Sekuler and Mierkiewicz (1977), whereby the distance effect is most pronounced for kindergarteners and decreases throughout development. Further research is necessary to understand this limitation of the model.

Another limitation, one shared with the pioneering Zorzi and Butterworth (1999) model, is that the models considered here only perform the comparison task. By contrast, the Verguts et al. (2005) model also performs naming and parity judgment tasks and can thus be evaluated against a broader range of data. Future research is required to extend the range of the models considered here to new tasks.

Although the developmental model produced changes in error rates and comparison speed that parallel human data, further work is necessary to more completely model the development of number comparison. In particular, the model needs to account for the more pronounced distance effect of younger participants reported by Sekuler and Mierkiewicz. One reason our model may have failed to capture this trend is that we trained the model using distributions based on the occurrence of numbers in adult language. One approach to improving the developmental simulations may be to use training data that parallel the distributions of numbers in children's and child-directed speech.

Acknowledgments

Mark Lewis's work was supported by a U.S.D.E. Institute of Education Sciences (IES) training grant (Title: Interdisciplinary Education Sciences Training Program IES Award #R305C050059 University of Minnesota PRF# 473473).

References

Dehaene, S., & Mehler, J. (1992). Cross-linguistic regularities in the frequency of number words. *Cognition*, *43*(1), 1-29.

Gallistel, C. R., & Gelman, R. (2000). Non-verbal numerical cognition: From reals to integers. *Trends in Cognitive Sciences*, *4*(2), 59-65.

Holloway, I. D., & Ansari, D. (2008). Domain-specific and domain-general changes in children's development of number comparison. *Developmental Science*, *11*, 644-649.

Moyer, R. S., & Landauer, T. K. (1967). Time required for judgements of numerical inequality. *Nature (215),* 1519-120.

Parkman, J. M. (1971). Temporal aspects of digit and letter inequality judgments. *Journal of Experimental Psychology*, *91*, 191-205.

Rule, S. J. (1969). Equal discriminability scale of number. *Journal of Experimental Psychology*, *79*, 35-38

Sekuler, R., & Mierkiewicz, D. (1977). Children's judgments of numerical inequality. *Child Development*, 630-633.

Verguts, T., Fias, W., & Stevens, M. (2005). A model of exact small-number representation. *Psychonomic Bulletin & Review*, *12*(1), 66-80.

Zorzi, M., & Butterworth, B. (1999). A computational model of number comparison. In *Proceedings of the Twenty-First Annual Conference of the Cognitive Science Society* (pp. 772–777).

Causal stream location effects in preschoolers

David W. Buchanan (david_buchanan@brown.edu) and David M. Sobel (dave_sobel@brown.edu)
Department of Cognitive and Linguistic Sciences
Box 1978, Brown University, Providence, RI, 02912

Abstract

Based on the predictions of a computational model, we test preschoolers' ability to reason about *stream location effects*: reasoning that interventions that occur on a common part of a causal process should be more likely to affect multiple relations, than interventions which occur on independent parts of a causal process. In two experiments, we show that 3- and 4-year-olds both show a stream location effect. Children show this effect for both familiar and unfamiliar interventions.

Keywords: causal reasoning, cognitive development, models of causal reasoning.

Introduction

Even when we do not explicitly understand the details of how a causal system works, we often have strong intuitions about the effects of different interventions on that system. Consider a phenomenon that is a mystery to most adults – the way that a remote turns on a television. Even though most of us could not verbally describe the mechanism, we know that removing the batteries from the remote would make it fail. We also know that the television must be plugged in. We even know about correlations between relations: Say you have two remotes that both turn on your television – for instance, the one that came with your television, and a universal remote you bought to control all your devices. One day, both fail to turn on the television. You replace the batteries in remote A, and it now succeeds in turning on the television. You would not expect this intervention to change the efficacy of remote B – it would be odd if replacing the batteries in one remote made both effective. On the other hand, if you had noticed that the television was unplugged, and plugging it in had restored the efficacy of remote A, you would not be surprised if remote B started working as well. These inferences seem obvious, even to a person who knows nothing about how these devices operate. Interventions in one place in a causal system are expected to have wide-ranging effects, while interventions in other locations are not. Why is this?

To explain this intuition, we will use the metaphor of a *causal stream*. For instance, imagine we introduce pollution into a river that has several branches. The further upstream the pollution occurs, the more branches of the stream will be polluted. When we think of causation as flowing down a branching path, we can start to formalize these intuitions. Elsewhere (Buchanan, Tenenbaum, & Sobel, 2010) we have proposed a computational model that generates causal structures that have a branching, stream-like character. We call it the *causal edge replacement process*, or CERP. While the details are beyond the scope of this paper, we will outline its main implications.

CERP makes use of causal graphical models, a way of representing causal relations using graphs (Gopnik et al., 2004; Pearl, 2000; Spirtes, Glymour, & Scheines, 2001). Nodes represent events, and directed edges represent causal relations. Figure 1a shows an example, the simplest way of representing the common effect relation of the television (C) and the two remotes, A and B. There are other ways of representing this relation; CERP tends to generate graphs that are more complex, like Figure 1b. In this graph, we have enough detail to represent interventions on the mechanism that relates cause and effect. For instance, X is such an intervention, which disables both relations, preventing causation from flowing down the edge on which it falls. (The dashed edge indicates an inhibitory relation.) CERP implies that when causal relations share a node, they always share part of the path from cause to effect. In a common cause relation such as the television example, interventions (like X) that occur late in the causal stream (close to the common effect) are more likely to fall on this shared path, changing both relations. Interventions (like Y and Z) that occur early in the causal stream, are more likely to fall on the independent path, changing only one relation. We call this difference a *stream location effect*. Note that these implications are general and structural, and do not depend on the specific causal system involved.

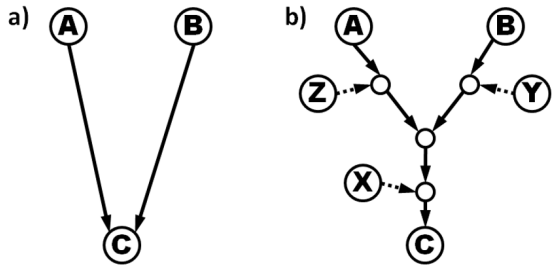


Figure 1: Examples of causal graphical models. Nodes represent events, and edges represent causal relations. Dashed edges indicate inhibitory relations. On the left, the simplest graph that captures a common effect relation: A and B both cause C. On the right, a graph generated by CERP, which allows us to make predictions about stream location effects. Intervening on X disables both relations, but intervening on Y or Z disables only one.

The stream location hypothesis is that human beings should expect stream location effects even about systems for which we have little or no knowledge of the causal mechanism involved (like the television remote, for most of us). While this hypothesis was inspired by CERP, it is not the only model which is consistent with these predictions. For instance, evidence that supports the stream location hypothesis

is not inconsistent with a general approach to causal graphical models. But only models (like CERP) that have a branching character directly and specifically predict stream location effects.

There is already some empirical evidence that suggests that stream location effects may exist in adults. Mayrhofer, Hagmayer, and Waldmann (2008) told participants a cover story involving mind-reading aliens: When the "cause" alien thought of food, he often caused the three "effect" aliens to think of food as well. The experimenters manipulated the number of other aliens that thought of food, and asked subjects to judge the probability that a given alien would also think of food, given that the cause alien was thinking of food. For instance, when the cause alien and two other effect aliens were thinking of food, participants judged it highly likely that the third alien was thinking of food. When the cause alien was thinking of food, but the two other effect aliens were not, subjects judged it less likely that the third effect alien was thinking of food.

This difference is known as a *nonindependence* effect, which CERP fits well in general[1]: Adults predict that collateral effects of a common cause should be correlated, even given their common cause. Crucially for the stream location hypothesis, the strength of nonindependence could be manipulated by changing the cover story. In the "sending" condition, participants were told that the cause alien sometimes had trouble concentrating; there was a strong nonindependence effect in this condition. In the "receive" condition, the effect aliens sometimes had trouble concentrating; there was a significantly weaker nonindependence effect in this condition. Mayrhofer et al. succeeded in showing that by changing the description of the mechanism, they could change the degree of nonindependence observed. We hypothesize that a stream location effect was responsible for this difference: the location of the described inhibitor (trouble concentrating) in the causal stream was different between conditions. Of course, we are only explaining their data in hindsight. The experiments in this paper present a more direct predictive test of the stream location hypothesis.

Because CERP makes such strong predictions about situations in which we have little or no knowledge, the best tests of the stream location hypothesis will be in children's causal reasoning. This is because children often have little specific causal knowledge about individual causal systems; we can see their reasoning as revealing the expected form of causation more than the expected content of causation. For instance, infants seem to initially expect that novel abstract objects need to make physical contact in order to interact causally (Leslie & Keeble, 1987). Among preschoolers, Bullock, Gelman, and Baillargeon (1982) found that even 3-year-olds expect that causes must precede their effects. They also found that 3-year-olds could reason appropriately about interventions on causal systems: They could recognize that some interventions would change a relation, whereas some would not. Buchanan and Sobel (submitted) showed that this ability depended on the specific causal system involved. For instance, 3-year-olds could not reason correctly about interventions on electrical connection, but they could reason correctly about interventions on batteries. On the other hand, 4-year-olds could reason appropriately about both connection and batteries. Because of these developmental differences, and numerous other studies on preschoolers' causal reasoning, we chose to test 3- and 4-year-olds in these experiments.

Our overarching hypothesis, which we test in two experiments, is that preschool-aged children will show stream location effects, expecting different changes to arise from interventions at different locations in the causal stream. Further, we predict that these differences will continue to hold regardless of the familiarity of the intervention involved, as long as that unfamiliar intervention appears to change the causal relation in the same way.

Experiment 1

In this experiment, we tested the hypothesis that children would reason differently about interventions to a causal system, depending on the location of the intervention in a causal stream. We presented children with a novel common effect relation in which both relations failed. Then we made a change, either early or late in the causal stream, which apparently enabled one of the relations. We asked children whether this change would enable the other relation as well. Our hypothesis was that in accordance with CERP, children would judge the late intervention as more likely to affect both relations than the early intervention.

Methods

Participants We tested 16 three-year-olds, (8 girls, mean age = 40.3 months, range = 36-45 months) and 16 four-year-olds (2 girls, mean age = 52.25 months, range = 48-59 months). Three additional children were tested, but were excluded due to experimenter error or equipment failure. About half the children were recruited from birth records, and the other half were recruited and tested either at a children's museum or at a local preschool. Children were randomly assigned to either the "early inhibitor" ($n = 16$) or "late inhibitor" ($n = 16$) condition. There were an equal number of 3- and 4-year-olds in each condition.

Materials Materials consisted of two sets of commercially available closet lights, modified for the experiment. In one set (the "cause lights") there were 8 lights, each 10 cm in diameter, with a large white button that illuminated only when actively depressed. These lights had a battery compartment on the underside that could hold two batteries; they required the presence of both batteries, inserted properly, in order to illuminate when pressed. It was possible to insert one battery backwards, in order to be able to show the presence of two batteries, without the light illuminating when pressed. The compartment also had a cover, which could be left on or off. The casing of each light was painted a different color, so that

[1]For details of this fit, see Buchanan et al. (2010)

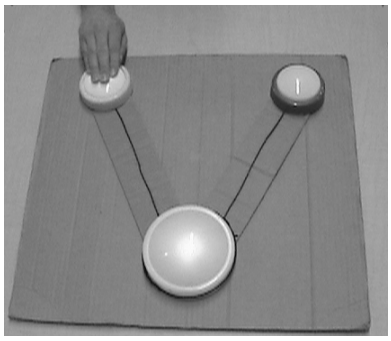


Figure 2: The push lights environment used in the present experiments, shown from the point of view of the child.

children could easily differentiate the lights.

Another set (the "effect" lights) consisted of four similar but larger lights, each about 14 cm in diameter. The effect lights were rendered distinguishable from one another by placing pipe cleaners of different colors around their casing. These lights each held four batteries. One of these lights was modified using radio-controlled car components, such that it illuminated only when a hidden remote was activated – depressing the light was not the actual cause of its illumination. It was possible to give adults and children the impression that they were causing the light to activate by depressing it, by activating the remote only when they pressed the light. The experimenter (a trained magician) practiced this effect until a convincing causal impression was achieved. The remote allowed the experimenter freedom to control which actions, if any, appeared to cause the effect light to illuminate. In post-tests, the experimenter was often able to use the remote to convince children that pressing their nose activated the effect light. All such children were subsequently debriefed, and allowed to play with the remote.

The lights were mounted on a piece of cardboard, together with wires that appeared to connect the lights. The cardboard was used so that the experimenter could easily retract the whole setup, controlling when and if children could intervene on the lights. The setup is shown in Figure 2. We refer to this setup as the "push lights environment."

Procedure The experimenter began by showing children all the lights to be used in the procedure, in order to establish that there were a large number of them. Then he showed them the push lights environment, arranged as shown in Figure 2: There were two small "cause" lights, connected to one large "effect" light using the wires. This began the training phase. The experimenter said: "I have some of these lights. When you push on them, they light up. See: [pushes on effect light, and it illuminates.] Here, you try." Children pushed the effect light, which illuminated. "Sometimes, when I push on these little lights, they'll make the big light go. Watch." He then pushed each cause light, both of which appeared to cause the large light to illuminate simultaneously. He then pointed to each of the cause lights and asked: "Does this one make the big light go?" Most children (26 out of 32) correctly answered "yes" to this question. The remaining children answered "yes" after one instance of corrective feedback. Excluding children who required feedback on this or any other training question did not change the statistical significance of the results we report.

The experimenter then removed the three initial lights, and arranged three visibly different lights in the same configuration. This began the first of three test phases. In the "late inhibitor" condition, the effect light in each test phase light was missing one battery. In the "early inhibitor" condition, the cause lights in each test phase were each missing one battery, and thus did not illuminate when pressed. The battery covers were left off so that the absence of batteries was visible, but only when the lights were flipped over. These new lights failed to activate the effect light. Children were asked about the efficacy of the relations. Most children (26 out of 32) correctly answered "no" to these questions on all three trials. Five responded correctly after one round of feedback, and one child required two rounds. Excluding these children did not change the significance of reported results. Note that at this point in the procedure, children had correctly answered "no" to two questions with feedback, and "yes" to two questions with feedback. Thus children were not coached on a strategy that would allow them to answer the test questions correctly.

At this point in the test phase, the experimenter made a modification to the causal system, which depended on the condition. In the "late inhibitor" condition, the experimenter flipped over the large light, exposing the fact that there was a battery missing. He said: "Look, this light has room for a battery, but there's no battery in there. Let's put a battery in." He then inserted a battery into the space, and replaced the light in its original position. In the "early inhibitor" condition, he instead flipped over and added a battery to one of the cause lights. Then he said "Let's try this light now." He pressed one cause light (side counterbalanced, and in the early inhibitor condition, always the effect light that had been intervened on), which made the effect light illuminate.[2] The experimenter then asked, pointing to this light: "Does this one make the big light go now?" All children answered "yes." He then pointed to the other light: "What about this one? Will this one make the big light go now?" Children's responses to this test question were recorded and analyzed. The experimenter repeated the test phase three times with three visibly different sets of lights, for a total of three answers from each participant. This meant that we collected three yes/no answers from each child, making 24 for each age group/ condition combination.

[2]In the "early inhibitor" condition, the cause light did not illuminate even when it was effective. Otherwise the illumination of the cause light would be diagnostic of its efficacy in causing the effect light to illuminate. That is, in the "early inhibitor" condition, when the experimenter pressed on an effective cause light, only the effect light illuminated.

Results

No effects were found for the age or gender of the children, or whether the question was initially asked about a light that was on the left or on the right. Results are shown in Table 1. In the "early inhibitor" condition, only 2 out of 24 responses from 3-year-olds, and 1 out of 24 responses from 4-year-olds were "yes." Both of these patterns were significantly below the proportion of "yes" responses predicted by chance, Binomial test, $p < 0.01$ in both cases. In the "late inhibitor" condition, all 3- and 4-year olds answered "yes" to every question, meaning that both age groups answered "yes" to 24 out of 24 questions. This was significantly above chance, Binomial test, $p < 0.01$ for both conditions.

Table 1: Mean number of "yes" answers in Experiment 1.

Age Group	Condition	"yes"/trials	Mean	SD
3-year-olds	Early($n = 8$)	2/24	0.25	0.46
	Late ($n = 8$)	24/24	3.00	0.00
4-year-olds	Early ($n = 8$)	1/24	0.12	0.35
	Late ($n = 8$)	24/24	3.00	0.00

Children of both ages were significantly more likely to answer "yes" in the late inhibitor than in the early inhibitor condition. For 3-year-olds, the average number of "yes" responses out of three was 0.25 in the early inhibitor condition and 3.00 in the late inhibitor condition. Among 4-year-olds, the means were 0.12 and 3.00, respectively. We ran a 2(age group) x 2(condition) ANOVA, which revealed a main effect of condition, $F = 746.05$, $p < 0.01$, partial $\eta^2 = 0.96$, but no main effect of age, $F = 0.37$, $p = 0.55$, partial $\eta^2 = 0.01$ or interaction, $F = 0.37$, $p = 0.55$, partial $\eta^2 = 0.01$. Because of the apparent difference in the variances, we supplemented this analysis using a Mann-Whitney U test: We found a significant difference between conditions, $U = 0.00$, $Z = 5.37$, $p < 0.01$, but not in the number of correct answers ("yes" in the late condition, and "no" in the early condition) between age groups, $U = 20.00$, $Z = 0.60$, $p = 0.78$.

Discussion

Children were significantly more likely to predict a change in both relations in the "early inhibitor" than in the "late inhibitor" condition. These results indicate that both 3- and 4-year-olds are sensitive to the location of an intervention in a causal stream. An open question is whether this is due primarily to structural inferences that apply to causal streams in general, or acquired knowledge about this specific intervention. Previous research (i.e. Buchanan & Sobel, submitted; Gottfried & Gelman, 2005) indicates that even 3-year-olds understand enough about batteries to make appropriate inferences about relevant and irrelevant modifications to a causal system when batteries are involved. To support stream location as a general structural principle, we needed to show a stream location effect for an unfamiliar intervention.

Experiment 2

The goal of this experiment was to show a stream location effect in a similar environment, but using an intervention that was not usually associated with a change in causal relations. In this experiment, instead of adding batteries, we added a battery cover. Since the presence of battery covers is not actually causally related to the efficacy of toys, children could not base their inferences on previous causal learning. We hypothesized that we would find the same effect in this experiment as in Experiment 1. We were agnostic as to whether children would be more uncertain in this experiment, generating a significantly more variable pattern of responses.

Participants As in Experiment 1, we tested 16 three-year-olds, (3 girls, mean age = 40.67 months, range= 36-46 months) and 16 four-year-olds (5 girls, mean age = 52.18 months, range = 48-57 months). One additional child was tested, but was excluded due to experimenter error. About half the children were recruited from birth records, and the other half were recruited at a children's museum or local preschool. Again, children were randomly assigned to either the "early inhibitor" ($n = 16$) or "late inhibitor" ($n = 16$) condition, with an equal number of 3- and 4-year-olds in each condition.

Methods Experiment 2 was identical to Experiment 1, using the same materials and procedure, except for two changes: First, all battery slots were filled, but as in Experiment 1 the battery covers were left off initially. Second, during the procedure, the experimenter did not add batteries; instead, he pointed out that each light was missing a cover, and added one. Thus, he said: "Look, this one does not have a cover. Let's put a cover on there." Just as in Experiment 1, in the early inhibitor condition, the cause lights did not illuminate. This was done in order to maintain similarity between experiments. Also as in Experiment 1, intervening on the light (the cause light in the late inhibitor condition, and the effect light in the early inhibitor condition) apparently changed the efficacy of one of the cause lights. Children were asked to verify this. Most children (22 out of 32) required no corrective feedback during this procedure. Six children required one round of feedback, two children required two rounds of feedback, one child required three rounds, and another four. Excluding all children who required any feedback does not change the statistical significance of the results reported below. In the test question (for which no feedback was provided), children were asked to predict the efficacy of the other cause light. To avoid negative effects on children's causal learning, all children were debriefed on the deception at the end of the procedure: They were allowed to play with the remote, and observed that replacing the covers did not in reality make the lights effective.

Results

Results are shown in Table 2. Again, no effects were found for gender, or the location of the intervened-on light. As in

Experiment 1, chance analyses showed that in all four condition/age group combinations, the proportion of "yes" responses was significantly different from what would be expected by chance, Binomial test, $p < 0.01$ in each case. In the "early inhibitor" condition, children were below chance, and in the "late inhibitor" condition, they were above chance.

Table 2: Number of "yes" answers in Experiment 2.

Age Group	Condition	"yes"/trials	Mean	SD
3-year-olds	Early ($n = 8$)	4/24	0.5/3	0.27
	Late ($n = 8$)	23/24	2.87/3	0.12
4-year-olds	Early ($n = 8$)	3/24	0.37/3	0.74
	Late ($n = 8$)	21/24	2.62/3	1.06

For 3-year-olds, the average number of "yes" responses out of three was 0.25 in the early inhibitor condition and 3.00 in the late inhibitor condition. Among 4-year-olds, the means were 0.12 and 3.00, respectively. As in Experiment 1, we ran a 2(age group) x 2(condition) ANOVA, which revealed a main effect of condition, $F = 72.05$, $p < 0.01$, partial $\eta^2 = 0.72$, but no main effect of age, $F = 0.47$, $p = 0.497$, partial $\eta^2 = 0.017$ or interaction, $F = 0.05$, $p = 0.82$, partial $\eta^2 = 0.002$. Because of the apparent difference in the variances, we supplemented this analysis using a Mann-Whitney U test: We found a significant difference between conditions, $U = 11.50$, $Z = 4.27$, $p < 0.01$, but not in the number of correct answers ("yes" in the late condition, and "no" in the early condition) between age groups, $U = 123.00$, $Z = 0.26$, $p = 0.87$.

Anecdotally, several 4-year-olds seemed surprised that merely changing the cover had changed the efficacy of the relation. Some initially responded "maybe" to the test question – they were asked to choose either a "yes" or "no" response. All children eventually responded appropriately to the test question.

Because of this phenomenon, we also tested for differences between the experiments: We gave each child a score based on the number of correct ("yes" in late inhibitor, and "no" in early inhibitor) responses they made. We then performed a t-test on the difference between scores in the two experiments. In Experiment 1, the mean score was 2.90, ($SD = 0.29$), and in Experiment 2, the mean score was 2.65, ($SD = 0.74$). This difference was only marginally statistically significant, $t = 1.76$, $df = 62$, $p = 0.08$. Because of the difference in the variances, we supplemented this analysis using a Mann-Whitney U test, which also failed to show a significant difference, $U = 443.50$, $Z = 1.48$, $p = 0.14$.

Discussion

Even in the case of an intervention that is not normally causally related to efficacy, 3- and 4-year-olds were able to reason appropriately about stream location. That is, when the unfamiliar intervention that resulted in a change in efficacy was early in the causal stream, children predicted that the other causal relation would be unaffected, but when the unfamiliar intervention was late in the causal stream, they predicted that both relations would be affected. The data are inconclusive about whether there is an effect of familiarity, possibly making children's responses more variable. Even if this effect exists, it is probably small, and manifestly not large enough to eliminate the stream location effect we observed.

General Discussion

Both experiments supported the stream location hypothesis: Children were significantly more likely to predict a change in both relations in the late intervention than in the early intervention condition. The results in both age groups indicate that children have a strong understanding of stream location, even as early as three years old. Furthermore, Experiment 2 showed that children would make these inferences even for an unfamiliar intervention. This suggests that stream location may reflect knowledge of the structure of causation in general, rather than just experience with a specific causal system. Further work is necessary to provide more support for this possibility. For instance, we may be able to find stream location effects when the intervention is not just unfamiliar but opposite to past associations – if batteries disable rather than enable the relation, for example.

CERP predicts and supports these findings. The model prescribes that early interventions on a common effect structure are likely to fall on the independent portion of the path from cause to effect, changing only one relation, whereas late interventions on a common effect structure are likely to fall on the shared portion of the path from both causes to the effect, changing both relations. While the data we present are consistent with the general causal graphical model framework – we have shown that children prefer Figure 1b over Figure 1a – only CERP explains *how* this preference is generated.

In Experiment 2, adding a cover appeared to change the efficacy of a the causal relation, a situation that would be counter to children's experiences (if any) with such causal systems in the real world. Why, then, did they not show a significantly different pattern of responding in Experiment 2? For instance, we might expect children to guess. The answer comes from noticing that the intervention was perfectly correlated with a change in efficacy: the light never worked until we added a cover. It seems that children required an explanation for this change, and the addition of the cover was the only explanation available. This is in line with previous research (i.e. Schulz & Sommerville, 2006) that shows that children are determinsists: they attribute such changes in efficacy to human interventions, rather than attributing them to randomness. In work currently underway, we are exploring the interaction between this type of determinism, and inferences about hidden interventions on a causal stream. CERP makes clear predictions here: for instance, if failures sometimes occur without an intervention, a given failure is less indicative of a changed causal relation. Thus, more variable relations should show weaker stream location effects.

The existence of stream location effects in preschoolers provides support for CERP as a model of causal reasoning. Although CERP arose from attempts to make quantitative fits to data on a different phenomenon with adults (namely, the nonindependence phenomenon mentioned above), it nonetheless predicted a novel, qualitative effect that could be detected in children. We see this as one of many examples (i.e. Sobel, Tenenbaum, & Gopnik, 2004; Thelen, Schoner, Scheier, & Smith, 2001) of a productive dialog between experiments and models in cognitive development and cognitive science.

Acknowledgments

This work was supported by NSF (DLS-0518161 to DMS). We would like to thank all of the parents and children who participated, and the Providence Children's Museum and Brown/Fox Point preschool for allowing us to recruit and test on-site. We also thank Karis Casagrande, Caroline Kleeman, Brianna Doherty, Katie Green, Rachel Shelley-Abrahamson, Kristen Swan, for help with data collection and analysis.

References

Buchanan, D., & Sobel, D. (submitted). Mechanism-based causal reasoning in young children.

Buchanan, D., Tenenbaum, J., & Sobel, D. (2010). Edge replacement and nonindependence in causation. In *Proceedings of the 32nd annual conference of the cognitive science society.*

Bullock, M., Gelman, R., & Baillargeon, R. (1982). The development of causal reasoning. In W. J. Friedman (Ed.), *The developmental psychology of time* (pp. 209–254). Academic Press.

Gopnik, A., Glymour, C., Sobel, D., Schulz, L., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review*, *111*, 3–32.

Gottfried, G., & Gelman, S. (2005). Developing domain-specific causal-explanatory frameworks: The role of insides and immanence. *Cognitive Development*, *20*(1), 137–158.

Leslie, A., & Keeble, S. (1987). Do six-month-old infants perceive causality. *Cognition*, *25*(3), 265–288.

Mayrhofer, R., Hagmayer, Y., & Waldmann, M. (2008). Violations of screening off: A Bayesian error attribution model of causal reasoning. *Unpublished presentation at Mathematical Psychology 2008*.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge University Press.

Schulz, L., & Sommerville, J. (2006). God does not play dice: Causal determinism and preschoolers' causal inferences. *Child development*, *77*(2), 427–442.

Sobel, D., Tenenbaum, J., & Gopnik, A. (2004). Children's causal inferences from indirect evidence: Backwards blocking and Bayesian reasoning in preschoolers. *Cognitive Science*, *28*(3), 303–333.

Spirtes, P., Glymour, C., & Scheines, R. (2001). *Causation, prediction, and search*. The MIT Press.

Thelen, E., Schoner, G., Scheier, C., & Smith, L. (2001). The dynamics of embodiment: A field theory of infant perseverative reaching. *Behavioral and brain sciences*, *24*(01), 1–34.

In What Sense is P(A|B) P(B) = P(A,B)? The Relationship between Distributional Format and Subjective Probability Estimates

Belinda Bruza, Matthew B. Welsh, Daniel J. Navarro & Stephen H. Begg
({belinda.bruza, matthew.welsh, daniel.navarro, steve.begg}@adelaide.edu.au)
The University of Adelaide, SA 5005, Australia

Abstract

The elicitation of uncertainty is a topic of interest in a range of disciplines. The conversion of expert beliefs into probability distributions can play a role in assisting key decisions in industry. However, elicitation methods can be prone to bias. In this paper we investigate the effect of changing the presentation of stimulus information and question format on elicited judgments of marginal, conditional and joint probabilities. Participants taught a probability distribution in one structure were expected to have difficulty assessing the distribution in another structure. While this pattern was not found, it turned out that training participants on the more difficult task (learning from a conditional structure) improved overall performance.

Keywords: decision making; cognitive biases; elicitation; probability learning

The "elicitation of uncertainty" is a general term that is often used to refer to methods for translating a set of implicit beliefs into an explicit probability distribution (Wolfson, 2001). The reason for using these methods is to allow researchers to incorporate subjective expert knowledge into a quantitative model that makes predictions about future events (Morgan & Keith, 1995). In view of this, good elicitation methods can play an important role in guiding decision making in a range of industries in which uncertain outcomes are central.

One of the main impediments to widespread use of elicitation techniques in applied settings is the inherent difficulty of the task. This difficulty is caused by the many well-known decision-making heuristics and biases, which can distort the estimates of the underlying beliefs. For instance, anchoring and adjustment, representativeness, availability, base rate neglect and overconfidence (see Tversky & Kahneman, 1974; Bar-Hillel, 1980; Lichtenstein, Fischoff, & Phillips, 1982) have all been found to influence the judgments people make in an elicitation context, in both lay and expert populations (see, e.g., Eddy, 1982; Welsh, Bratvold & Begg, 2005). Moreover, people often mistake conditional probabilities for joint probabilities (Pollasek et al., 1987) since these are easier to compute (Lewis & Keren, 1999), and often experience difficulties with characterizing the conditioning event (Bar-Hillel & Falk, 1982). People may confuse one conditional probability *P(A | B)* with another *P(B | A)*, or have difficulties interpreting instructions related to probability (Bar-Hillel, 1980; Fiedler et al., 2000).

Problem Representation

A consistent finding in the decision-making literature is that people are sensitive to the surface representation of a problem. For instance: options described in terms of gains are evaluated differently to the same options when described in terms of losses (Kahneman & Tversky, 1979); changing the surface form of the Tower of Hanoi problem can alter the difficulty of the task (Gunzelmann & Blessing, 2000); and statistical problems expressed in terms of frequencies seem to be easier than the same problems described in terms of probabilities (Gigerenzer & Hoffrage, 1995).

One interesting variation on the question of problem representation arises when people need to learn about and report on the joint distribution of two variables, *A* and *B*. Mathematically, we can describe the distribution to be learned and subsequently elicited in three formally equivalent ways, by noting that:

$$P(A, B) = P(A \mid B)\, P(B) = P(B \mid A)\, P(A) \qquad (1)$$

For the current purposes we refer to each of these three variations as a "problem format", and note that while all three formats describe to the same distribution over *A* and *B*, there is no guarantee that people will treat them as such. Indeed, in view of the known differences in how people estimate marginal probabilities, conditional probabilities and joint probabilities, we would expect to observe fairly substantial differences between formats.

In this paper we describe an experiment that examines (1) whether one format for the problem leads to superior learning and subsequent probability estimation in general, and (2) whether learning in one format makes it easier to report on questions framed in the same format. Should either of these two effects be observed, a natural method for improving elicitation in an applied context would be to alter the presentation format to be more suited to the expectations of the expert whose beliefs are to be elicited.

Method

Participants

Participants were 60 students (18 male) studying at the University of Adelaide, aged 18 to 37 years, and were paid $15 for their time.

Procedure

The experiment involved three learning tasks, and two testing conditions, and the measurement of several key covariates. All participants completed all three learning tasks, but were tested in only one of the two testing conditions (based on a random assignment to one of two groups). The basic procedure was as follows. Participants were individually tested in a quiet, well-lit room in front of a computer. Firstly, basic demographic data were collected. Participants then did a simple practice task to demonstrate how the interface works and to illustrate what they would be tested on. Participants then undertook all three learning-plus-elicitation tasks in a random order, with the covariate measurement tasks (APM & MHV; see later) used as filler tasks to help prevent order effects and learned probabilities from previous urn distributions affecting recall of later distributions. Participants were not allowed to use external resources (e.g., pen and paper, calculator) to aid calculations.

The learning tasks

The experiment involved showing participants 20 "candies" which could vary in color (red or blue) and shape (circle or triangle). The participants' task was to learn the distribution over colors and shapes. The experiment was conducted on computer, and the interface was designed so that the stimuli could be presented to participants in all three formats (i.e., *P(A, B)*, *P(A | B) P(B)* and *P(B | A) P(A)*). The cover story told participants that they had encountered a "vending machine" (which we refer to as the urn) filled with candies, which was varied slightly between conditions. Participants were shown the 20 candies one at a time: each candy appeared after the participant clicked on a "vend" button (see Figure 1). After viewing all candies, they were asked various elicitation questions (described later).

In the *wrapped* candy condition, participants were told that the candy was covered in a yellow wrapper. As a result, when they clicked on the "vend" button (see Figure 1) they would be able to see the shape of the candy but not its color. If they then clicked the "unwrap" button, the color would be revealed. Because of the sequential way in which the stimulus characteristics were revealed, the format in which "the world" presents the items is naturally described in terms *P(color | shape) P(shape)*.

In the *masked* candy condition, the distribution was also shown to people in a sequential fashion. However, the color of the candy was shown before the shape, so that participants would see items in a *P(shape | color) P(color)* format. The cover story in this case implied that the participants were initially viewing the candies through a small window, so they could see the color but not the shape. In this condition, the "unwrap" button was replaced by a "retrieve" button, which then revealed the shape.

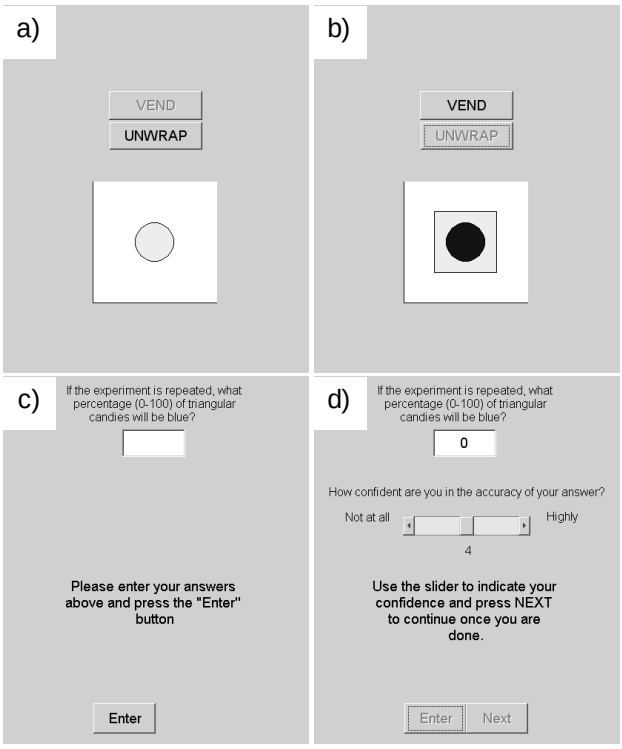


Figure 1: GUI of wrapped candy condition. Vended circular candy (a) unwrapped to reveal blue color (b). Percentage estimate requested (c) before confidence rating (d). All GUIs presented the same basic layout.

The *unveiled* candy condition was the simplest of the three, and presented the two features together as soon as the participants clicked on the "vend" button. As a consequence, participants observed the joint distribution *P(color, shape)* in a more direct fashion.

To allow for between-participant comparisons, the base rate for each type of candy was preset in all three conditions (see Table 1). The shape and color of each candy was randomly determined at each trial. After completing 20 trials, the elicitation questions were asked.

The elicitation questions

Participants answered 10 possible questions about the percentage of particular candies in a *future* urn distribution (two regarding marginal probabilities, four conditional probabilities, and four joint probabilities). The questions were asked in a random order. Participants in group 1 were asked to give estimates in terms of a "shape preceding color" structure. These estimates were therefore elicited in the same format in which the distribution of candies was learnt in the wrapped candy condition (e.g., *P*(*circle*), *P*(*red | circle*), *P*(*red, circle*) etc). Participants in group 2 were requested to give estimates in terms of a "color preceding shape" structure, hence estimates were elicited in the same format in which the distribution of candies was learnt in the masked candy condition (e.g., *P*(*red*), *P*(*circle | red*), *P*(*circle, red*) etc).

Thus, in order to produce estimates, participants in group 1, for example, needed to "flip" the probability distribution (using Bayes' theorem) that they learnt for candies in the masked candy condition (see Table 1). As shown in Figure 1, the elicited percentage was typed in an editable text box. Additionally, for every probability judgment that participants were asked to make, they were subsequently asked rate their confidence in their accuracy, using a horizontal scroll bar to enter a value that ranged from 1 (*not at all*) to 7 (*highly*). This process was repeated for each elicitation question. All GUI controls were sequentially locked and unlocked to prevent backtracking and to ensure that the participant answered questions in the prescribed order.

Covariate controls

Given that participants with higher cognitive functioning have been found to perform better on tasks involving conditional reasoning (Stanovich & West 1998) and to be less susceptible to overconfidence (Pallier et al., 2002), intelligence measures were included as controls. Bors and Stokes' (1998) short form of Raven, Court and Raven's (1988a) Advanced Progressive Matrices (APM) was used to measure fluid intelligence. Crystallized intelligence was measured using Senior Form 1 of the Mill Hill Vocabulary Scale (MHV) (Raven, Court, & Raven, 1988b). Finally, information regarding participants' TER (percentile Tertiary Entrance Rank derived from students' performance in the final year of secondary education in several Australian states) was collected.

Results

The accuracy of any given judgment was assessed in terms of the absolute error – the magnitude of the difference between the empirical probability experienced by the participant, and the participant's subjective estimate of that probability. Since the distribution of absolute errors was skewed to the right, a log transformation was performed on absolute error data points prior to model fitting (with the addition of 1 to each data point to prevent negative values).

Order, format and question type effects

It was hypothesized that participants taught a probability distribution in one conditional structure would have difficulty estimating probabilities in another conditional structure. Since group 1 participants were asked to answer questions consistent with the format learnt in the wrapped candy condition (i.e., a shape preceding color structure), they were expected to give estimates closer to the empirical rate than would group 2 participants. The same was expected for group 2 participants in the masked candy condition (i.e., a color preceding shape structure). Because a joint distribution was presented in the unveiled candy condition, question format was expected to have no effect on performance in either group.

Table 1: Base rates of candy color (red or blue) and shape (circle or triangle) and consistency of question format with presentation of candy features in each of the three conditions for group 1 and group 2. Since the unveiled candy condition contained a joint distribution, question format was neither consistent nor inconsistent.

	Condition		
	Wrapped	Masked	Unveiled
	Average base rate (%)		
Color			
Red	10	30	30
Blue	90	70	70
Shape			
Circle	30	90	30
Triangle	70	10	70
	Format consistent		
Group 1			
Shape, color	Yes	No	–
Group 2			
Color, shape	No	Yes	–

Examination of the relationship between questions of conditional probability and log absolute error in Figure 2a) showed what may be weak evidence for the predicted effect. That is, group 1 produced better conditional probability estimates in the wrapped candy condition, and group 2 produced better conditional probability estimates in the masked candy condition. There was also an effect of question type with the log absolute error score highest on questions of conditional probability (see Figure 2b).

Note that in the experimental phase there were four sets of questions that should sum to 100%: questions 1 and 2, which concerned marginal probabilities; 3 and 4; 5 and 6, which concerned conditional probabilities; and 7 to 10, which asked for joint probabilities. Errors within each set should therefore be positively correlated (e.g., if a participant estimated 50% of candies would be circular when the true value was 25%, the absolute error would be 25% and a similar absolute error score would thus be expected in their estimate of triangular candies). Moreover, there were participant-level correlations – some participants consistently had poorer or better performance than others. Linear mixed effects models were therefore fitted to further investigate the effect of condition (wrapped candy, masked candy and unveiled candy), group (1 or 2) and question type (marginal, conditional or joint) on absolute error while adjusting for interdependence of the data.

To adjust for the dependence in estimates within the same question set and within estimates from the same participant for a condition, random effects for participant and question set × condition × participant were added to the linear mixed effects models. Condition, group and question type were treated as fixed effects (predictor variables) in the model. The three-way interaction

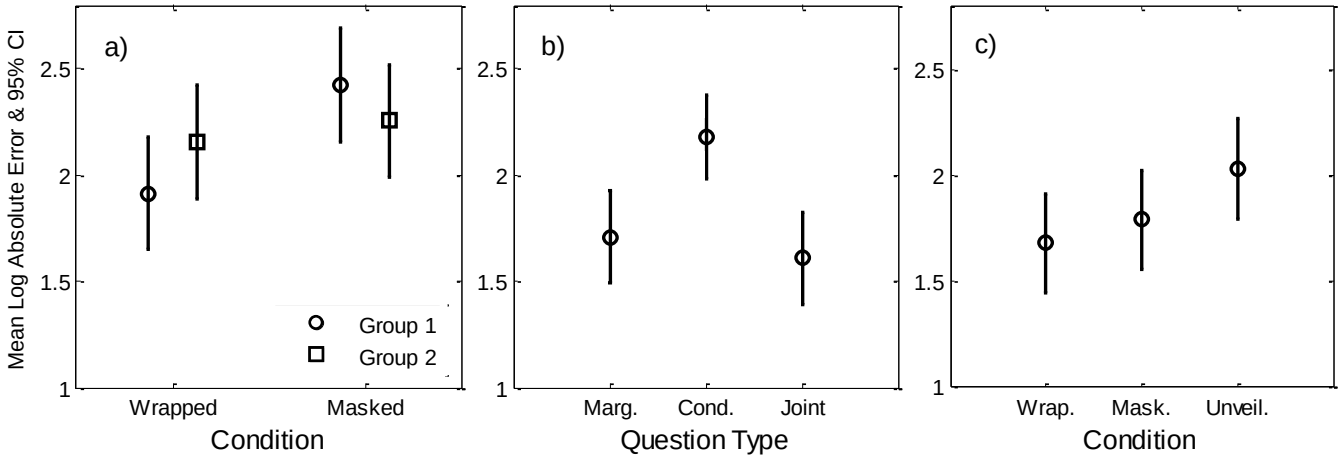


Figure 2: Mean log absolute error scores, with 95% confidence intervals, for (a) group 1 and 2 estimates of conditional probability in the wrapped and masked candy conditions; (b) combined estimates of marginal, conditional and joint probability; and (c) combined estimates in wrapped, masked and unveiled candy conditions. Group 1 $N = 30$, Group 2 $N = 30$. Sample size of estimates is $N = 240$ in each candy condition in (a); $N = 120$ for marginals, $N = 240$ for conditionals, $N = 240$ for joint in (b); and $N = 600$ for each candy condition in (c).

between these variables and all two-way interactions were examined. APM, MHV and TER scores were also included as fixed effects in the models to assess their influence on absolute error. Degrees of freedom were calculated using the containment method (see Littell et al., 1996).

There were no significant interactions so interaction effects were removed from the model. Significant main effects were found for question type $F(2, 1061) = 31.38$, $p < .001$; and condition (i.e., urn type), $F(2, 1061) = 4.75$, $p < .01$, as can be seen in Figures 2b) and 2c).

Bonferroni post-hoc tests indicated questions of conditional probability (adjusted $M = 2.18$, $SE = .10$) were associated with higher log absolute error relative to questions of marginal probability, (adjusted $M = 1.71$, $SE = .11$), $F(1, 1061) = 25.40$, $p < .001$; and questions of joint probability, (adjusted $M = 1.61$, $SE = .11$), $F(1, 1061) = 55.20$, $p < .001$.

The unveiled candy condition (adjusted $M = 2.03$, $SE = .12$) had a significantly higher log absolute error than the wrapped candy condition, (adjusted $M = 1.68$, $SE = .12$), $F(1, 1061) = 9.12$, $p < .01$ and masked candy condition, (adjusted $M = 1.79$, $SE = .12$), $F(1, 1061) = 4.12$, $p = .04$.

Intelligence and accuracy

Participants with higher APM, MHV and TER scores were expected to provide more accurate probability estimates and a significant main effect was found for APM, ($F(1, 1061) = 3.20$, $p = .04$). Looking at Table 2, it seems that MHV scores were also weakly related to accuracy on the estimation task, with 8 of 9 correlations in the predicted direction ($p = .002$ by a sign test), four of which were significant in their own right. TER scores, however, had no predictive power. Independent samples t-tests confirmed that there was no significant difference between groups on the covariates, specifically: the APM (group 1 $M = 10.47$, $SD = 2.16$; group 2 $M = 10.77$, $SD = 2.93$; $t(58) = -.45$, $p = .65$); and MHV (group 1 $M = 58.37$, $SD = 10.48$; group 2 $M = 56.40$, $SD = 10.53$; $t(58) = .73$, $p = .47$).

Table 2: Spearman correlations between MHV score, APM score, TER score and log absolute error broken down by question type.

Question type	Score	Condition		
		Wrapped	Masked	Unveiled
Marginal	MHV	–.08	–.11	.01
	APM	–.29**	–.23**	–.20*
	TER[a]	–.13	.02	.20
Conditional	MHV	–.09	–.11*	–.08
	APM	–.06	–.11	–.26**
	TER[a]	.12	.05	.02
Joint	MHV	–.16**	–.15**	–.12*
	APM	–.10	–.12*	–.19**
	TER[a]	–.06	.03	–.02

Note. $*p < .05$, $**p < .01$, one–tailed. $N = 60$, unless otherwise indicated. [a] $n = 41$. Sample size of estimates is $N = 120$ for marginals, $N = 240$ for conditionals, $N = 240$ for joint.

Confidence and accuracy

It was predicted that confidence ratings would decrease as absolute error scores increase. All correlations were significant and in the expected, negative direction (see Table 3).

Linear mixed effects models were also fitted to assess the relationship between confidence rating and absolute error. The relationship between confidence rating and log

absolute error was highly significant – with every one unit increase in confidence rating, log absolute error was expected to decrease by –.19 units. That is, an approximately 20% reduction in absolute error, $t(1559) = -7.31$, $p < .001$. No significant interaction effects were found but a significant main effect was found for condition, $F(2, 1061) = 14.69$, $p < .001$.

Table 3: Spearman correlations between confidence rating and log absolute error broken down by question type and condition.

	Condition		
Question type	Wrapped	Masked	Unveiled
Marginal	–.32**	–.22**	–.31**
Conditional	–.25**	–.20**	–.13*
Joint	–.27**	–.16**	–.19**

Note. $*p < .05$, $**p < .01$, one–tailed. $N = 60$. Sample size of estimates is $N = 120$ for marginals, $N = 240$ for conditionals, $N = 240$ for joint.

Bonferroni post-hoc tests indicated confidence ratings were significantly lower for the unveiled candy condition ($M = 3.92$, $SE = .20$) compared to the wrapped candy condition ($M = 4.57$, $SE = .20$), $F(1, 1061) = 28.41$, $p < .001$; and the masked candy condition ($M = 4.35$, $SE = .20$), $F(1, 1061) = 12.53$, $p = <.001$.

Discussion

In this study we found no significant evidence to suggest that performance on probability estimation tasks changes as a function of the order in which information is acquired. When items were presented in the *P(A | B) P(B)* format, there was no advantage to eliciting participants' knowledge in this same format, as compared to eliciting the knowledge in the *P(B | A) P(A)* format. However, we did find that participants who were shown the stimuli in the *P(A, B)* format actually had significantly higher error than participants taught in either of the other two formats, regardless of what type of question was asked. Given that joint probabilities are presumably easier to process than conditionals, one possibility is that this is a depth of processing effect (Craik & Lockhart, 1972). Recall that, when studying urns with a conditional structure, participants were presented with one characteristic of the candy at a time. This two stage learning process presumably required more attention, involvement and time spent to process each stimulus than the one stage learning process of the joint distribution. This may have contributed to the improvement in overall performance, precisely because the task is harder.

The expected effect of question type was also observed. Absolute error was smallest on questions related to marginal probabilities, and largest on questions related to conditional probability. This was observed regardless of question format or distribution format. These findings are consistent with previous research (see, e.g., Lewis & Keren, 1999), as is the relationship between accuracy and intelligence (see Stanovich & West, 1998). Finally, participants did seem to be aware of how accurate their performance was, since confidence and accuracy were related in a sensible fashion.

Future directions

Our finding that training on the more difficult task improves elicitation warrants further investigation. Future research could determine whether performance is improved by only the two stage learning process used here or by any training format that fosters increased depth of processing.

Limitations

Before concluding, it is worthwhile considering the limitations of this study. It should be noted, for example, that participants provided estimates for each urn distribution based on only 20 trials, which may not have been sufficient for them to form strong beliefs about the distribution. Increasing the number of trials to 100 might allow participants to get a better sense of the underlying distributions, while a larger sample size would enable a clearer understanding of the results; for example, clarifying whether the suggestive results seen in Figure 2a) actually reflect the hypothesized interaction between learnt distributional formats and probability estimates.

A secondary concern is the level of control over the empirically observed rates; although the "true" base rate for each urn was the same, random draws from the true distribution contain sampling error that results in participants observing slightly different empirical rates from each other, diluting control over the experiment. One solution to this would be to use a pseudo-random distribution with a fixed empirical rate, rather than the truly probabilistic approach taken here.

A third possibility is that the sequential presentation method did not have a strong effect because only one stimulus (the candy) was perceived. That is, the nature of the task may have undermined the experimental manipulation to some extent. A task in which *A* and *B* refer to distinct but causally related stimuli (instead of two features of a single object) might provide a better test of the hypothesis.

Conclusions

Although one of the main predicted effects did not appear, the overall results paint an intriguing picture of the potential impacts that training format can have on elicited probability estimates. For example, the fact that training people on the harder task improves estimates is interesting, and of potential applied value. The longer-term goal is thus to see how well these findings can be adapted to improve the elicitation of uncertainty in real world contexts.

Acknowledgments

MBW and SHB were supported by ExxonMobil and Santos' contributions to the CIBP at the Australian School of Petroleum. DJN was supported by an Australian Research Fellowship (ARC grant DP-0773794). We thank Thomas Sullivan for his assistance with the statistical models.

References

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica, 44*, 211-233.

Bar-Hillel, M., & Falk, R. (1982). Some teasers concerning conditional probabilities. *Cognition, 11*, 109-22.

Bors, D. A., & Stokes, T. L. (1998). Raven's Advanced Progressive Matrices: Norms for first-year university students and the development of a short form. *Educational and Psychological Measurement, 58*, 382-398.

Bratvold, R. B., Bickel, J. E., & Lohne, H. P. (2007) Value of information in the oil and gas industry: past, present and future. *Paper presented at the 2007 SPE Annual Technical Conference and Exhibition, 11-14, November, Anaheim, California.*

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11, 671-684.*

Eddy, D. M. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases.* Cambridge: Cambridge University Press.

Fiedler, K., Brinkmann, B., Betsch, T., & Wild, B. (2000). A sampling approach to biases in conditional probability judgments: Beyond base rate neglect and statistical format. *Journal of Experimental Psychology: General, 129*, 399-418.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review, 102*, 684-704.

Gunzelmann, G., & Blessing, S. B. (2000). Why are some problems easy? New insights into the Tower of Hanoi. In L. R. Gleitman and A. K. Joshi (Eds.), *Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society* (p. 1029). Mahwah: NJ: Lawrence Erlbaum Associates.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47*, 263-292.

Lewis, C., & Keren, G. (1999). On the difficulties underlying Bayesian reasoning: a comment on Gigerenzer and Hoffrage. *Psychological Review, 106*, 411–416.

Lichtenstein, S., Fischoff, B., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases*. Cambridge: Cambridge University Press.

Littell, R. C., Milliken, G. A., Stroup, W. W., & Wolfsinger, R. D. (1996). *SAS system for mixed models*. SAS Institute.

Morgan, M. G., & Henrion, M. (1990). *Uncertainty: A guide to dealing with uncertainty in quantitative risk and policy analysis*. Cambridge: Cambridge University Press.

Morgan, M. G., & Keith, D. W. (1995). Subjective judgments by climate experts. *Environmental Science and Technology, 29*(10), 468A-476A.

Pallier, G., Wilkinson, R., Danthiir, V., Kleitman, S., Knezevic, G., Stankov, L., & Roberts, R. D. (2002). The role of individual differences in the accuracy of confidence judgments. *The Journal of General Psychology, 129*, 257–299.

Pollatsek, A., Well, A. D., Konold, C., Hardiman, P., & Cobb, G. (1987). Understanding conditional probabilities. *Organizational Behavior and Human Decision Processes, 40*, 255–269.

Raven, J. C., Court, J. H., & Raven, J. (1988a). Manual for Raven's Progressive Matrices and Vocabulary Scales. London: H. K. Lewis.

Raven, J. C., Court, J. H., & Raven, J. (1988b). *The Mill Hill Vocabulary Scale Form 1 Senior*. Oxford: Oxford Psychologists Press Limited.

Stanovich, K. E., & West, R. F. (1998). Individual differences in framing and conjunction effects. *Thinking and Reasoning, 4*, 289-317.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124-1131.

Welsh, M. B., Bratvold, R. B., & Begg, S. H. (2005). Cognitive biases in the petroleum industry: Impact and remediation. *Paper presented at the Proceedings of the Society of Petroleum Engineers 81st Annual Technical Conference and Exhibition, Houston, Texas.*

Wolfson, L. J. (2001). Elicitation of probabilities and probability distributions. In E. Science (Ed.), *International encyclopedia of the social and behavioral sciences*. Elsevier Science.

Of Parrots and Parsimony: Reconsidering Morgan's Canon

Matthew Brian Welsh (matthew.welsh@adelaide.edu.au)
University of Adelaide, North Terrace
Adelaide, SA 5005 Australia

Abstract

Morgan's Canon is a specific restating of Occam's Razor that dictates that any description of animal behavior should never call upon higher order psychological processes if the behavior could, fairly, be explained in terms of lower processes. Herein, the Canon is discussed both historically and in light of current research into animal behavior. A reconsideration of the principle of parsimony, taking into account current states of knowledge, is also considered. In short, it is argued that Morgan's Canon, while a useful guideline, may have been over-enthusiastically applied in situations where the state of knowledge about a species would dictate that descriptions of its behavior in terms of higher order processes would be equally or more parsimonious. The potential benefits of reconsidering the Canon are then discussed.

Keywords: parsimony; animal behavior; comparative psychology; theory of mind; individual differences.

Morgan's Canon

In no case is an animal activity to be interpreted in terms of higher psychological processes, if it can be fairly interpreted in terms of processes which stand lower in the scale of psychological evolution and development. (Morgan, 1903).

Comparisons between animal and human behaviors have a long history, with scholars as far back as Aristotle (340BC/1952) arguing that 'reason' divides humans from the rest of the animal kingdom. This division, embedded in the Christian distinction between the creation and place of men and animals, was carried through the writings of such philosophers as Descartes (1640/1988) who placed the seat of reason in the soul; and little seems to have challenged this view until the publication of Darwin's Origin of Species (1876/1988).

The arguments presented by Darwin, regarding the common descent of all animals through natural selection acting on ancestor populations, broke down the clear-cut division between human and animal that had previously held sway in Western thought and promoted the idea that, across species, one should expect to see variation in traits – including such mental attributes as intelligence (Darwin, 1899/1965). Thus, while humans might be the species blessed with the greatest reasoning ability, one would expect that other species would have this to a greater or less extent – with the further understanding that closely related species would, likely, have similar levels of intelligence.

Romanes (1882), following this parsimonious line of reasoning, produced his book *Animal Intelligence* in which he described a great variety of animal behaviors (both collected by himself and sent to him by correspondents) in terms of the mental states and understanding required to produce them. The anecdotal nature of this work, however, provoked the responses of Morgan (1903) and Thorndike (1911), whose use of observational study of animals convinced them that many of the cases of 'intelligent' behavior reported by Romanes were, in fact, easily explained as the result of trial-and-error learning.

The reaction to Romanes' book and the subsequent research on conditioning by Pavlov (1927) led to a drastic change in approach to animal behavior research. Rather than considering the pre-existing knowledge of common lineage, researchers were, instead, motivated to explain behavior in the simplest, possible psychological terms. This was partly the result of a genuine belief in the equipotentiality principle (Pavlov, 1927) - which regarded all animals as largely equivalent in terms of their ability to learn through conditioning – but seems also to have resulted from a revision of people's interpretation of the principle of parsimony with a greater focus on the simplicity of the explanatory rules and less on the need for accord with prior knowledge.

Thus, for the greater part of the 20th century, Morgan's Canon has held sway - and been interpreted to mean that animal behaviors should be explained, wherever possible using simple, conditioning-based explanations as these were judged to be most parsimonious and, thus, best.

Occam's Razor and Parsimony

Parsimony in scientific research is often regarded in terms of Occam's Razor, which literally translates as "entities must not be multiplied beyond necessity" but is commonly understood to mean that the simplest hypothesis explaining an observation is the best (Kneale & Kneale, 1962). However, this simple restatement ignores the key phrase in the original: "beyond necessity". Thus, a more complete restatement would require that the best explanation be the simplest one that accords with our state of knowledge about the object or event in question.

The relevance of this to animal behavior research is that, when considering the most parsimonious explanation for an animal's behavior, we must take into account what we already know about that species, related species and even animals in general. Imagine, for example, if one were to see a small animal (of an unfamiliar species) moving along the ground and were interested in starting to explain its behavioral repertoire. Starting with the very broadest of behaviors, for example, we might ask whether the creature's appearance in this location is indicative of its environmental predilections and behaviors.

That is, is the simplest (most parsimonious) explanation for its presence that it is a terrestrial creature native to the area? The answer, in the absence of additional information should, clearly, be yes – this is the simplest explanation that explains the limited data we have. It does not require us to hypothesize about any alternative modes of movement beyond the observed, terrestrial movement nor does it require an additional explanation about why a non-native creature might be here.

If, however, while still unfamiliar with the species in question, you recognize that it is a type of bird this would, almost certainly, change the description judged most parsimonious. Given a general knowledge of birds, it would seem reasonable to decide, instead, that the most parsimonious explanation is that the creature is capable of flight and only currently on the ground – as the vast majority of birds are capable of flight. To take the example a step further, imagine that, in addition to recognizing the creature as a bird, you also recognize that it is, in fact, a type of penguin. This would cause another revision in the best explanation for its behavior (current and potential); in this case, concluding that it is, most probably, flightless and aquatic - as are all other penguin species.

Thus, knowledge about related species changes both the description of current behavior and *expected* behavioral repertoire of an animal; and, any attempt to find the *simplest* (most parsimonious) explanation for an animal's behavior must incorporate this knowledge.

Animal Cognition

Few people, of course, would disagree with the above examples and ethologists such as Tinbergen (1951) and Lorenz (2002/1949), despite their largely behaviorist view-points would, doubtless, start any observations of a new species with assumptions regarding its behavior based on the behavior of known, related species. The behaviors described by ethologists and those considered of greatest import by those comparative psychologists holding to Morgan's Canon, however, differ in significant ways. For the most part, ethologists deal with general types of instinctive behavior in the natural environment whereas comparative psychology concerns itself with animal cognition to gain insight into human cognition. That is, to what extent are animals capable of reason, learning and self-awareness and how can this knowledge be used to better understand human behavior?

As noted above, the behaviorist school of psychology (see, e.g., Skinner, 1938) applied Morgan's Canon uniformly and attempted to explain both human and animal behavior in terms of conditioned responses as the equipotentiality principle argued for all organisms learning in, essentially, the same fashion with differences only in the speed at which learning occurred.

The cognitive revolution, starting in the 1950s, however, convinced most psychologists that attempts to explain complex, human behaviors such as language use within a simple, reinforcement-learning paradigm was infeasible (see, e.g., Neisser, 1967). Perhaps the single greatest effect of this revolution was to move psychology away from regarding the mind as a black box about which nothing could be known beyond inputs (stimuli) and outputs (observed behaviors). Instead, it was recognized that: firstly, the mind cannot be a blank slate prior to learning because a blank slate will not react to inputs in any way (for a recent summary of the cognitive revolution, see Pinker, 2003); and, secondly, that observing the manner in which behaviors change as stimuli change allows us to meaningfully hypothesize about cognitive structures/processes.

This recognition of the need to understand an organism's cognitive processes or mind was not restricted to humans, however. Breland and Breland (1961) identified instinctive drift (the tendency for animals' trained behaviors to revert to the nearest equivalent instinctive behavior) and Garcia and Koelling (1966) exposed the difficulties of training animals when the conditioned and unconditioned stimuli did not 'match' (e.g., illness could be induced in rats by a flavor but not by a light or sound). That is, it was demonstrated that, in order to predict and understand experimental results, one needs to know not just the stimulus and resultant behavior but also the cognitive processes of the organism in question.

Despite such work, however, the shift from behaviorism to cognitive psychology stalled in animal research – no doubt partly because access to human cognitions is often as easy as asking someone what they are thinking while animal minds are much harder to read; but also, it seems, due to a continued belief that the most parsimonious explanation are those that posit the simplest possible processes without reference to 'human' cognitive processes (see, e.g., Wynne, 2007).

The question, though, how *should* our understanding of parsimony affect our beliefs regarding the best explanations for animal behaviors in terms of psychological processes? This is discussed in greater detail as regards two central areas of animal cognition that have provoked significant discussion: animal intelligence and theories of mind.

Animal Intelligence

Between Species Differences

Most people have very little difficulty in believing that certain types of animal are more intelligent than others. This seems to be one case where our understanding of the concept of common lineage has led us to conclude that animals more like us are likely to be more intelligent; and experimental work has offered some support for this. Work by Warren (1977), for example, comparing fish, chickens, mice and cats on a learning task returned the expected order of results – with the cats performing best, then the mice, the chickens and, finally, the fish – although only the cats performed significantly better than the other species.

The problem with such assessments, however, is clear. The very differences described by the Brelands (1961) and Garcia and Koelling (1966) make cross-species comparisons difficult as differences in instinctive behaviors mean that certain species learn particular tasks more easily, thus

making it difficult to determine whether any differences result from differences in "intelligence" or just differential degrees of match between a species and the task/apparatus being used.

Individual and Strain Differences

To avoid these problems, most researchers concentrate, instead, on within-species analyses as these should eliminate most differences in instinctive behavior and allow meaningful conclusions to be drawn. However, between research into human and animal intelligence lies a vast gulf - in the form of differential treatment of individual and group differences within a species.

In human research, individual differences is a major field of research, while group differences are very much a sideline – a result, Fraser (1995) argues, of the feeling that research into group differences in intelligence (in particular) is motivated by prejudice. By comparison, animal research is dominated by comparisons between strains of the same species – with tests of such attributes as spatial ability, memory and even reasoning using pigeons (Wilkie & Wilson, 1995), mice (Tang, et al., 1999) and rats (Anderson, 1992), respectively. These often include neuroanatomical studies to associate the cognitive differences with particular brain structures (the hippocampus, for example, is strongly linked to spatial learning by the above studies).

Individual differences in animals, by comparison, have been largely ignored or even dismissed – as by Warren (1977), who claimed that there was no evidence of individual animals performing above the level of their peers. This dismissal, however, seems to be driven, in part at least, by adherence to the narrow interpretation of Morgan's Canon described above. That is, individual differences in animal intelligence are not discussed because intelligence (which is largely understood in terms of studies of individual differences in humans) is regarded as a 'higher' order cognitive process and, therefore, inappropriate to apply to animal behavior.

This position, however, is at odds with both our everyday experience – those people who interact with animals on a regular basis such as animal trainers and researchers are adamant that certain, individual animals are smarter than others (see, e.g., Goodall, 1968; Kohler, 1925; Pepperberg, 1990) – and knowledge available to us from a variety of fields, including evolutionary theory and the strain differences studies mentioned above.

The first point, of course, relies on the same anecdotal evidence that led to the formulation of Morgan's Canon and runs the risk of the Clever Hans effect (Pfungst, 1911) where the trainer's own unconscious behavior is responsible for apparent differences in learning. As such, it must be treated with caution.

The second point, however, argues strongly for there being individual differences in animal "intelligence" – broadly defined here as any cognitive faculty affecting performance on a task. Specifically, according to the theory of evolution by natural selection, it is individual, genetic differences in traits that cause differential survival and (eventually) speciation (Darwin, 1876/1988). As such, if the argument is to be made that there are differences between the cognitive abilities of different species (for example, that humans have better reasoning abilities than other species) then these differences must have their origins in individual differences within the ancestral populations from which the compared species are descended (Griffin, 1976). Thus, in the ancestral species from which humans and chimpanzees are both descended, there must have been individuals with better reasoning abilities than their peers – otherwise these reasoning abilities could not be selected for and, thus, contribute to the evolution of differences between humans and chimpanzees.

Logically, this argument holds at every point of speciation where one believes there is a difference in cognitive abilities between current species. While this argument does not, in and of itself, make any statement regarding individual differences within *current* species, any attempt to argue that individual differences might, no longer, exist in species other than our own would seem so unlikely as to strain credibility. That is, the claim would have to be that: while, at every point in the past, individual differences in cognitive ability existed within a wide variety of species, now, for unexplained reasons, only one species has such individual differences.

In addition to the argument from parsimony proposed above, we also have evidence for individual differences in cognitive abilities in the form of our ability to selectively breed strains of a species for particular cognitive tasks such as maze-solving (Stewart, 1961); and the observation that strain differences are known to exist on a variety of tasks including those described above. Given the derivation of these strains from common, ancestor populations, it seems unavoidable to conclude that individual differences in the various cognitive abilities discussed do exist and that strain differences are just these writ large.

In addition to these logical arguments, there are also a number of studies (see, e.g., Anderson, 1992; Locurto & Scanlon, 1998; Welsh, 2002) that have shown individual differences in the performance of not just specific tasks but also the emergence of factor structures amongst various tasks reminiscent of the structure of human intelligence as described by Carroll (1993). Specifically, there is some evidence for attributes akin to human spatial intelligence and memory and learning (*Gv* and *Gy* in Carroll's model).

Given this, it seems reasonable to argue that, when attempting to explain animal behavior, appeals to differential levels of cognitive ability between individuals is not an 'unnecessary multiplication of entities' nor does it violate Morgans's Canon as, given the evidence for individual differences in various cognitive abilities, animal behavior cannot be *fairly* described without reference to such higher order cognitive constructs. In fact, any explanation for an animal's behavior that excludes this knowledge is likely to be overly simplistic rather than parsimonious.

Animal Theories of Mind

Another area of argument in which Morgan's Canon is frequently applied regards whether animals have a 'theory of mind'. That is, to what extent should animals be regarded as possessing minds in the way that humans do; are they self aware and aware of the minds of others (Premack & Woodruff, 1978)? A number of tests of this are commonly used and interpretations of experimental results are often hotly debated in terms of whether the behavior of the animals in questions indicates a theory of mind or can be explained via simple, stimulus-response relationships.

The goal, herein, is not to attempt to fully restate the debate; rather, key aspects of the debate will be considered along with findings relating to these and the interpretations will be discussed in terms of their parsimony in explaining not just the specific behavior at hand but also prior knowledge including phylogenetic relationships.

Attention

One of the preliminary tests for a theory of mind relates to whether an organism reacts to another organism's attention. That is, if one animal is looking in a particular direction, will the other animal look there as well. This is regarded as a test of an organism's theory of mind as it, theoretically at least, requires that the second organism be able to determine where the first creature is looking and what it could see from there.

For example, chimpanzees have been shown to understand point-of-view – that is, their behavior changes according to what an observing creature could see from its perspective (Hare, Call, Agnetta, & Tomasello, 2000). Further tests of this ability to understand attention have included observations of canine communication, where dogs' behaviors are affected by whether they can currently be seen by other dogs (Horowitz, 2009) or people (Call, Brauer, Kaminski, & Tomasello, 2003).

These tests of attention, however, are often criticized (in terms of their relevance to animal theories of mind) as their results can be explained in terms of selective rewards. That is, in environments when a human is directly facing them, a dog is more likely to have been punished for disobeying a command than when a human is facing away. Thus, differential learning could occur such that greater obedience is observed when the dog-human dyad is in certain spatial relations but not in others. This explanation requires only simple psychological processes to be hypothesized and, as a result, is often claimed to be a more parsimonious interpretation of animals' apparent ability to understand the attentional states of others.

Whether it is, in fact, a simpler explanation, though, is questionable. For example, the ability of the dog to distinguish between the situations when a second creature is and is not looking at it – as required by the stimulus-response explanation – requires the dog to have been in sufficient situations like this one to have learnt the difference between the various orientations of other creatures and their responses to various communication methods. That is, it pre-supposes a history of learning for which no evidence is presented.

Further, given that we know that one social mammal (humans) definitely has the ability to determine where another creature is attending (which assists with social communication and cooperative behaviors), should our starting assumption be that a species bred from another highly sociable mammal (wolves) and further selected for its ability to cooperate with humans does or does not have the same ability?

Imitation

Another central theme is theory of mind research is imitative behavior. That is, if an organism can observe another organism and then *imitate* the behavior, then this is argued to indicate its ability to understand the intentions of the first creature. Of course, there are provisos added to this simple description. The observer must be able to distinguish between accidental and deliberate behaviors and must also be able act in an intentional way – that is, the assumption must be that the organism's goal in imitating the behavior is to achieve the *outcome* that they observed the other creature achieving – rather than to simply mimic the action (Tomasello, Kruger, & Ratner, 1993).

The ever-present difficulties in designing animal experiments such that the animal is motivated to do as the experimenter intends make such analyses difficult with other species – to the extent that Zentall (2006) suggested that, given the number of social and non-social learning factors that need to be distinguished from imitation, inclusion of the recognition of intent might preclude any finding of imitation in non-verbal animals (including young humans).

Instead, Zentall (2006) proposes controlling for a list of pre-identified non-imitative learning behaviors and then, by a process of elimination, calling any learning that still occurs "imitative". Using this looser definition, there are a number of studies that compare how often organisms utilize a particular method to achieve a specific task – having seen conspecifics perform the task in one of the possible ways. Such studies, using budgerigars (Dawson & Foss, 1965), monkeys (Custance, Whiten, & Bard, 1999) and rats as subjects, show that an animal's preferred method of achieving specific aims varies according to how it has seen other animals perform the same task.

This has been demonstrated most clearly in chimpanzees (Buttelmann, Carpenter, Call, & Tomasello, 2007) who operated a device with their foot when an unencumbered human demonstrated its operation in this way but used their hands after seeing a human with his hands full operate the device with his foot. That is, they seem capable of differentiating between cases when the person could and could not use their hands and concluding that, when he could but didn't, there must have been a reason for this.

Once again, we are left with a question to answer: is it more likely, given the evidence we have seen from other species, that so useful a learning mechanism (bridging the

gap between instinctive and self-learnt behaviors, as Zentall, 2006, notes) is restricted to a single species or that imitative learning is likely to be a common ability of many social species?

False Belief

Perhaps the best known of the tests for theory of mind are those for false beliefs. That is, whether an organism can predict the actions of another organism based on the differences between their knowledge about a situation. The ability to understand false beliefs has proved very difficult to demonstrate in animals – in part, no doubt because of the required complexity of the task.

The classic design of such tests is to have an animal observe a conspecific observe a reward being hidden and then have the first animal observe the reward being moved while the second is not watching (see, e.g., Call & Tomasello, 1999; Hare, et al., 2000). The behavior of the first animal is then used to attempt to determine whether it realizes that the second animal's belief about the location of the reward is false.

The majority of attempts to test animals understanding of false beliefs, however, have failed. Chimpanzees and other great apes, generally regarded as the most likely of animals to share any particular trait with humans, have not shown an ability to distinguish between ignorance and false belief (Call & Tomasello, 2008). In fact, other than humans aged 5 and over, only dolphins have shown significant evidence of understanding false beliefs (Tschudin, 2006). Thus, false beliefs may mark a qualitative difference between human and (at least the majority of) animal minds. That said, chimpanzees are able to distinguish between another animal's true beliefs and ignorance, indicating some understanding of the complexities of other minds(Call & Tomasello, 2008).

Discussion

There has been a tendency, when considering the results of animal experiments to interpret parsimony as applying to each, new experiment as if it is independent of all other observations. That is, within each experiment, Morgan's Canon is applied and the researchers attempt to explain the results in the simplest psychological terms, without reference to our pre-existing stores of knowledge from previous experiments, related fields, similar organisms and so forth. It is like a physicist who, rather than attempting to create universal laws, attempts to explain the results of each, individual experiment in the simplest terms without reference to the known laws of physics.

Given the research and argument presented above, it seems difficult to conclude that restricting discussion of animal behaviors to 'lower' level psychological process (typically stimulus-reward learning) is an appropriate approach. While an explanation of any behavior can be attempted in stimulus-reward terms, the adequacy of said explanation must be considered. Where such an explanation has to posit the existence of a large number of unobserved learning trials in a variety of different contexts, and alternative explanations exist that accord with our knowledge about the behavior of other species and the relationships between them, a principled application of parsimony would seem to require a reconsideration of Morgan's Canon.

That is, while recognizing the potential dangers of anthropomorphism, it would seem that to adequately explain the findings from a variety of animal studies requires the use of higher-level psychological concepts such as intelligence and an understanding that animals are likely to have at least a limited theory of mind. In short, we need to consider animal behavior from a more cognitive view-point.

Future Research

An acceptance that animal behavior can meaningfully be discussed in similar, cognitive terms to that of humans opens up a range of research opportunities. For example, advancements in genetics and the mapping of the complete genomes of various species allows for the use of synteny homology (the fact that portions of one species genome have corresponding regions on other species genomes where large numbers of genes are found in the same order) would allow the use of analyses to investigate the genetic basis of cognition.

That is, those higher-level psychological processes that have clear equivalents between humans and animals could be isolated using animal genetic models, which have the advantage of large litter sizes and short inter-generational intervals, and then mapped to the human genome. This approach is, in fact, already underway in the medical sciences (see, e.g., Tang, et al., 1999) but its acceptance within psychology has been limited (for exceptions, see Anderson, 1992; Locurto & Scanlon, 1998; Welsh, 2002) with the result that those best suited to isolating and measuring the cognitive traits of animals have yet to start playing a major role.

Conclusions

Morgan's Canon has, over the past century been applied in a manner which, while seeming rigorous, has actually reduced the parsimony of explanations of animal behavior. Moving away from this too-broad application of the Canon, in addition to being necessary in order to develop the best and most parsimonious explanations of animal behavior, will allow animal research to join the cognitive revolution and allow comparative, cognitive research which will shed further light on human cognition.

Acknowledgments

I wish to thank Ted Nettelbeck and Nick Burns for their comments on an earlier version of this manuscript.

References

Anderson, B. (1992). Rat reasoning: a reliability and validity study. *Psychbiology, 20*(238-242).

Aristotle (340BC/1952). *The Works of Aristotle* (Vol. 2). Chicago: Encyclopedia Brittanica.

Breland, K., & Breland, M. (1961). The misbehavior of organisms. *American Psychologist, 16*, 681-684.

Buttelmann, D., Carpenter, M., Call, J., & Tomasello, M. (2007). Enculturated chimpanzees imitate rationally. *Developmental Science, 10*(4), F31-F38.

Call, J., Brauer, J., Kaminski, J., & Tomasello, M. (2003). Domestic dogs (*Canis familiaris*) are sensitive to the attentional state of humans. *Journal of Comparative Psychology, 117*(3), 257-263.

Call, J., & Tomasello, M. (1999). A nonverbal false belief task: the performance of children and great apes. *Child Development, 70*, 381-395.

Call, J., & Tomasello, M. (2008). Does the chimpanzee have a theory of mind? 30 years later. *Trends in Cognitive Science, 12*(5), 187-192.

Carroll, J. B. (1993). *Human cognitive abilities: a survey of factor-analytic studies*. Cambridge, UK: Cambridge University Press.

Custance, D. M., Whiten, A., & Bard, K. A. (1999). Social learning of artificial fruit task in capuchin monkeys (*Cebus apella*). *Journal of Comparative Psychology, 113*, 13-23.

Darwin, C. (1876/1988). *The Origin of Species*. New York: New York University Press.

Darwin, C. (1899/1965). *Expression of Emotion in Man and Animals*. Chicago: University of Chicago Press.

Dawson, B. V., & Foss, B. M. (1965). Observational learning in budgerigars. *Animal Behaviour, 13*, 470-474.

Descartes, R. (1640/1988). *Selected Philosophical Writings* (J. Cottingham, R. Stoothoff & D. Murdoch, Trans.). Cambridge, UK: Cambridge University Press.

Fraser, S. (1995). *The bell curve wars: race, intelligence and the future of America*. New York: Basic Books.

Garcia, J., & Koelling, R. A. (1966). Relation of cue to consequence in avoidance learning. *Psychonomic Science, 4*(3), 123-124.

Goodall, J. (1968). The behaviour of free living chimpanzees in the Gomba Stream Reserve Tanzania. *Animal Behaviour Mongraphs, 1*(161-311).

Griffin, D. R. (1976). *The question of animal awareness: Evolutionary continuity of mental experience.* New York: Rockefeller University Press.

Hare, B., Call, J., Agnetta, B., & Tomasello, M. (2000). Chimpanzees know what conspecifics do and do not see. *Animal Behaviour, 59*, 771-785.

Horowitz, A. (2009). Attention to attention in domestic dog (*Canis familiaris*). *Animal Cognition, 12*, 107-118.

Kneale, W., & Kneale, M. (1962). *The development of logic*. London: Oxford University Press.

Kohler, W. (1925). *The mentality of apes* (W. E. Winter, Trans.). Oxford: Harcourt Brace.

Locurto, C., & Scanlon, C. (1998). Individual differences and a spatial learning factor in two strains of mice (*Mus musculus*). *Journal of Comparative Psychology, 112*(4), 344-352.

Lorenz, K. (2002/1949). *King Solomon's Ring*. London: Routledge.

Morgan, C. L. (1903). *An introduction to comparative psychology* (2nd ed.). London: W. Scott.

Neisser, U. (1967). *Cognitive Psychology*. New York: Appleton-Century-Crofts.

Pavlov, I. (1927). *Conditioned reflexes*. Oxford, England: Oxford University Press.

Pepperberg, I. M. (1990). Conceptual ability of some non-primate species, with an emphasis on an African Grey parrot. In S. T. Parker & K. R. Gibson (Eds.), *"Language" and Intelligence in monkeys and apes*. New York: Cambridge University Press.

Pfungst, O. (1911). *Clever Hans (the horse of Mr von Osten): A contribution to experimental animal and human psychology.* (C. L. Rahn, Trans.). New York: Henry Holt.

Pinker, S. (2003). *The Blank Slate*. New York: Penguin.

Premack, D. G., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences, 1*, 515-526.

Romanes, G. J. (1882). *Animal Intelligence*. London: Kegan Paul.

Skinner, B. F. (1938). *The behavior of organisms*. Cambridge, MA: Copley Publishing Group.

Stewart, J. (1961). Some behaviour characteristics of maze-bright and maze-dull animals. *Canadian Journal of Psychology, 15*(75-80).

Tang, Y., Shimizu, E., Dube, G. R., Rampon, C., Kerchner, G. A., Zhuo, M., et al. (1999). Genetic enhancement of learning and memory in mice. *Nature, 401*(6748), 63-69.

Thorndike, E. L. (1911). *Animal Intelligence*. New York: Macmillan.

Tinbergen, N. (1951). *The study of instinct*. Oxford: Clarendon Press.

Tomasello, M., Kruger, A. C., & Ratner, H. H. (1993). Cultural learning. *Behavioral and Brain Sciences, 16*, 495-552.

Tschudin, A. (2006). Belief attribution tasks with dolphins: what social minds can reveal about animal intelligence. In S. Hurely & M. Nudds (Eds.), *Rational Animals*. Oxford, UK: Oxford University Press.

Warren, J. M. (1977). A phylogenetic approach to learning and intelligence. In A. Olivero (Ed.), *Genetics, environment and intelligence*. New York: North Holland.

Welsh, M. B. (2002). *Of Mice and Men: the structure and bases of murine cognitive abilities.* Unpublished Doctoral dissertation, University of Adelaide, Adelaide.

Wilkie, D. M., & Wilson, R. J. (1995). More evidence of robust spatial associative memory in the pigeon, Colombia livia. *Animal Learning and Behavior, 23*(1), 69-75.

Wynne, C. D. L. (2007). What are animals? Why anthropomorphism is still not a scientific approach to behavior. *Comprative Cognition and Behavior, 2*, 125-135.

Zentall, T. (2006). Imitation: definitions, evidence, and mechanisms. *Animal Cognition, 9*, 335-353.

Socially Induced Motor Plasticity Affects Language Comprehension

David Havas (dahavas@wisc.edu) & Julia Jenvey (julia.jenvey@gmail.com)
Department of Psychology, 1202 W. Johnson Street
Madison, WI 53706 USA

Hayley Shilling (hshilling@wisc.edu)
Department of Counseling Psychology, 244 Rust-Schreiner Halls
Madison, WI 53715 USA

Mitchell Nathan (mnathan@wisc.edu)
Department of Educational Psychology, 1025 W. Johnson Street
Madison, WI 53706 USA

Abstract

Language understanding is a socially coordinated activity, but the mechanisms of social coordination in language are poorly understood. Evidence from embodied cognition has shown that movement-induced fatigue of actions slows comprehension of language that refers to those actions. Research on the mirror neuron system suggests that action systems of the brain are also involved in social understanding of actions performed by another, empathy, and possibly language. Here, we show that simultaneous performance and observation of kinematically similar actions produced a fatigue-like effect in sentence judgment times relative to dissimilar control actions. The results suggest that the same action systems used in language processing are influenced by social actions.

Keywords: language comprehension, embodied cognition, social cognition, joint action, motor plasticity, mirror neuron system.

Introduction

Language is fundamentally a social activity in which individuals coordinate their actions (Clark, 1996; Grice, 1975). Conversation involves intricately timed verbal and non-verbal signals for clarifying, initiating, guiding, and ending dialog (Clark, & Wilkes-Gibbs, 1986; Garrod & Anderson, 1987). The mechanisms of language coordination are of current interest.

According to theories of dialog, conversation is successful to the extent that there is similarity of mental states between participants (Garrod & Anderson, 1987). For example, dialog requires that participants share a common ground, or similarity in mental states about referents (Clark, 1996; Clark & Wilkes-Gibbs, 1986). Interlocutors tend to show similarity across linguistic and non-linguistic levels, including word use (Garrod & Anderson, 1987), syntax (Branigan et al., 2000), semantics (Clark, & Wilkes-Gibbs, 1986) and movements (Chartrand & Bargh, 1999). It has been proposed that the same neural systems used for action imitation are also used in dialog (Garrod & Pickering, 2004; Pickering & Garrod, 2006b).

Recent evidence from embodied cognition has shown a close link between action and language. Glenberg and Kaschak (2002), for example, implicated the action system's influence on language comprehension. Participants in this study responded to a series of sentences depicting transfer, either away from (" Close the drawer") or toward the body (" Open the drawer"). After reading each sentence, participants deemed it sensible or nonsense by pushing "yes" or "no" buttons and in so doing, were required to move their hands either toward their bodies or away. The results demonstrated an action-sentence compatibility effect (ACE). Sentence judgment times were shorter when the sentence depicted movement compatible with the movement required to make a sensible response. The ACE effect has been demonstrated for sentences describing both concrete and abstract (or metaphorical) transfer (Glenberg, Sato, Cattaneo, Riggio, Palumbo, & Buccino, 2008).

More recent evidence has shown that language comprehension is influenced by prior fatugue of the motor system. Glenberg, Sato, & Cattaneo (2008) demonstrated that fatigue of specific actions influences comprehension of written language depicting those actions. To fatigue the neural motor systems involved in toward and away movement, participants were asked to engage in a repetitive action in a toward or away direction. Participants moved 600 beans individually from a container to a target. They were then asked to read a series of sentences describing transfer either toward ("Mark deals you the cards") or away from the body ("You deal Mark the cards"), and judge them as sensible or nonsense by pressing a button. Response times (the time to read the sentence and push the button) were longer when the sentence depicted motion congruous to the movement the participant had previously fatigued. When the sentence depicted an incongruous movement, response times were shorter. This finding was taken as evidence that repeated movement induced plasticity in motor areas recruited in language processing. In particular, the repetition of movement induced muscular fatigue, forcing action-controlling neurons in Broca's region to increase their output, but no longer target the specific action (Glenberg et al., 2008). Thus, participants' comprehension of written depictions of the compatible action (toward or away) was slowed, compromised by shared involvement in both action and language in Broca's region (Gallese, 2008).

Although these studies demonstrate a close link between action and language, they do not address language as a joint action. Research on the mirror neuron hypothesis (Rizzolatti, Craighero, & Fadiga, 2002) suggests that action systems of the brain, including Broca's region, are involved in social understanding of actions, emotions, and possibly language (Rizzolatti & Arbib, 1998).

Mirror neurons, first discovered in the premotor cortex of the macaque monkey, fire both during execution and observation of the same action (di Pellegrino, Fadiga, Fogassi, Gallese, & Rizzolatti, 1992). A comparable mirror neuron system (MNS) in humans may contribute to a wide range of behaviors, including action understanding (Iacoboni, 2005; Rizzolatti, Fogassi, & Gallese, 2001), empathy (Blakemore & Decety, 2001; Carr et al, 2003; Gallese, 2003; Gallese, Keysers, & Rizzolatti, 2004), and language understanding (Aziz-Zadeh, Wilson, Rizzolatti, & Iacoboni, 2006; Rizzolatti & Arbib, 1998).

Evidence supports the existence of a mirror-neuron-like mechanism in humans in which the observed actions of another are processed using the motor system of the observer. In one study, Fadiga, Fogassi, Pavesi, & Rizzolatti, (1995) used transcranial magnetic stimulation (TMS) to increase activation of the motor cortex responsible for grasping an object, while participants observed the same action, or just the object. The dependent variable was the motor evoked potentials (MEPs) in the muscle affected by the stimulated motor cortex. Muscle activation increased during action observation relative to the control conditions, showing that observed actions potentiate the execution of similar actions in the observer.

Other evidence shows that this mirror-like mechanism is specific to kinematically similar actions. Using fMRI, Calvo-Merino, Glaser, Grezes, Passingham, & Haggard (2005) showed that action observation produces the greatest increase in premotor and other cortical area activation for actions that the observers had been trained to perform themselves. And Stefan et al. (2005) used TMS to demonstrate the formation of a kinematically specific motor memory through action observations.

The role of the putative MNS in language has begun to be examined. Using fMRI, Aziz-Zadeh, Wilson, Rizzolatti, & Iacoboni (2006) asked participants to read sentences and observe actions involving either the foot, hand or mouth. They first located brain regions in each subject that were most active during observation of foot, hand, and mouth actions. Next, they compared activations in each region during the reading of sentences involving foot, hand, and mouth actions. They found congruence between areas active during observed actions and the activation levels during reading of sentences describing those actions. Brain regions responded most to sentences that involved the body part for which it was most active during action observation.

If the putative MNS in humans shares neural mechanisms with the action-based language system, then observing another agent repeatedly performing an action should elicit a generalizing response from action controllers similar to that observed by Glenberg, Sato, and Cattaneo (2008).

To test the influence of social actions in language comprehension, the present study adds to the beans task of Glenberg et al. (2008) and manipulates activation of action controllers through two kinds of simultaneous observed movements. In the Mirrored condition, the movement of both participants is kinematically identical; that is, both participants move beans in the same direction relative to their own bodies. In the Control condition, particpants' movements differ in the direction of movement; one participant moves the beans away from their body, while the other participant moves beans toward their body.

By definition of the MNS, both observation and execution of an action activate the same neural systems, and therefore simultaneous observation and execution of action in the Mirrored condition should elicit greater net activation of action controllers than in the Control condition. Based on the results of the Glenberg, Sato, & Cattaneo (2008) study, it is expected that greater activation of action controllers in the Mirrored condition will enhance the fatigue effect in sentence comprehension, relative to the control condition. That is, fatiguing movements toward or away from the body will slow the comprehension of sentences describing transfer toward or away from the body (respectively), and this effect will be enhanced in the Mirrored condition.

Participants

Participants were 80 Introductory Psychology undergraduate students at the University of Wisconsin—Madison (53 females, 27 males). Participants were native English speakers recruited using the UW psychology department's appointment scheduler and were offered course credit for their participation. They were paired randomly, irrespective of gender or handedness. All participants were treated in a manner consistent with the APA's "Ethical Principles of Psychologists and Code of Conduct" (American Psychological Association, 2002).

Design

The experiment consisted of a 2 (practice movement Mirrored or Control) x 2 (practice movement toward or away) x 2 (sentence movement toward or away) mixed design with repeated measures on the third independent variable. The first independent variable consisted of two levels—Mirrored or Control movement. In the Mirrored condition, both participants in a pair transferred beans from one container to another in the same direction, either toward or away from their bodies, while seated across from one another at a small table. In the Control condition, participants transferred the beans in opposite directions (see Figure 1 for a schematic illustration of these conditions). The second independent variable consisted of two levels - practice movement toward or away from the body. The third independent variable consisted of sentences describing transfer either toward the body ("Tony gives you the cup"), or away from the body ("Sarah passes the tray to you").

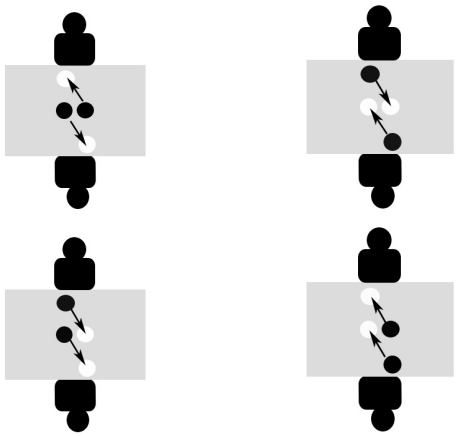


Figure 1: Schematic diagram of the Mirrored toward and away conditions (top row) and Control toward and away condition (bottom row).

In the reading task, the dependent variable was sentence judgment time; that is, the time between a sentence appearing on the screen and the participant pressing the "yes" button or "no" button to evaluate the sentence as nonsensical or sensible.

Materials

Experimenters used a protocol to guide the setup of materials, to assign participants to a condition, and to instruct participants in each phase of the experiment. For the first phase, the setup included four tupperware bowls on a card table, two for each participant. For each participant, one bowl contained three hundred beans, and the other bowl was empty but lidded with a hole in the top to serve as a target. All four containers were attached to the table by Velcro tabs.

For the second, reading comprehension phase of the experiment, short sentences were displayed one at a time on a computer monitor. Participants indicated that the sentence made sense or did not make sense using "yes" or "no" buttons located on the "3" and "8" keys on a keyboard.

For each participant, 280 sentences were shown in total. Half (140) of all sentences were sensible and half were nonsense ("You iron Linda the theory"). Of the 140 sensible sentences, 100 described transfer ("You give Angela a photo"), and the remaining 40 filler sentences did not ("Angela and you discuss the photo"). Following Glenberg, et al. (2008), half (50) the sensible sentences described transfer of a concrete object (" Tony gives you the cup"), and half described abstract transfer ("Liz tells you a story"). Also, half (50) the sensible sentences described transfer toward the body ("Meg hands you a paper" or "Liz tells you a story"), and half described transfer away from the body ("You hand Meg a paper" or "You give Chris advice"). Sentences were divided equally into two experiment halves. Example stimuli are provided in Table 1.

Table 1: Sample stimuli.

Sentence type	Example
Concrete transfer towards the body	Paul throws you the ball. Meg hands you a paper. Tony gives you a cup.
Abstract transfer towards the body	Chris gives you advice. Eric tells you a fact. Liz tells you a story.
Concrete transfer away from the body	You give Tony the cup. You throw Paul the ball. You hand Meg a paper.
Abstract transfer away from the body	You tell Liz a story. You give Chris advice. You tell Eric a fact.

Procedure

Participants were run in pairs. After participants signed consent forms, the experimenter read from a script, giving an overview of the experiment. First, the experimenter instructed each participant to go to one of two computer booths for practice in the language comprehension task. Practice consisted of instruction in the reading task, and six test trials. After finishing with practice, participants came out of the computer booths and were instructed to move 300 beans one at a time from the full container to the empty one using their right hands. Participants began the beans task at the same time. After both participants finished the bean transfer task, they returned to the reading booths to complete the first half of the sentence comprehension task. Participants typically finished the beans task within 1 minute of each other.

The experimenter then reversed the direction of movement for each participant by reversing the positions of the two containers. After both participants finished the first half of the reading task, they were instructed to return to the table and transfer the beans again, still using the right hand, to the empty container. When finished, the participants returned to the computer booths for the second half of the language task. When finished, participants were debriefed, thanked, and given course credit.

Results

Due to computer difficulties, experimenter error, or participant error, 7 participants were excluded from the analysis. We analyzed the data of the remaining 73 participants (48 females, 25 males; 70 right handed, 3 left handed). Because participants' accuracy in responding "yes" or "no" to the sensibility of each sentence reflects their reading comprehension, two participants with error rates higher than 10% were excluded from the analysis. Also excluded were trials containing erroneous responses or filler sentences. Only trials with raw sentence judgment times within two standard deviations of each participant's mean judgment time were used in the analysis.

We decided that the within subjects measure, away and toward movement, could produce carryover effects in the second half of the experiment. The analysis therefore includes only the first half of the experiment. A regression analysis adjusted judgment times to control for sentence length; these residual judgment times provide clearer data and the focus of our interpretation, but both raw and residual data were analyzed.

We predicted that participants in the congruent-toward practice direction condition would have higher (slower) judgment times on toward sentences than away and participants in the congruent-away practice direction condition would have higher judgment times on away sentences than toward sentences. Although we did not have specific predictions for participants in the incongruent condition, it was expected that less MNS stimulation and, by extension, less fatigue, would occur than in the congruent condition.

A three-way ANOVA was conducted separately for raw and residual judgment times. In the raw judgment times, there was a main effect of sentence direction on judgment times, F (1, 69) = 10.33, p = .002, showing longer judgment times for toward sentence (M=1718, SD=336) than for away sentences (M=1669, SD=305).

Contrary to the hypothesis, the difference between the Mirrored and Control conditions in raw judgment times only approached significance, F (1, 69) = 3.48, p= .066. An interaction between movement condition (Mirrored vs. Control) and practice direction (toward vs. away) also approached significance, F (1, 69) = 3.89, p =.053. None of these interactions were significant in residual judgment times.

Critically, the expected three-way interaction of sentence direction, practice direction, and movement condition was found in both raw [F (1, 69) = 4.60, p = .035] and residual judgment times [F (1,69) = 6.01, p = .017]. That is, the interaction of action and language depended on the movement condition. Mean residual judgment times for the three-way ANOVA are listed in Table 2.

Table 2: Mean residual judgment times.

	Practice toward	Practice away
Mirrored condition		
Toward sentences	41.0	10.3
Away sentences	-43.5	-8.0
Control condition		
Toward sentences	-19.8	31.6
Away sentences	18.7	-24.9

To decompose the three-way interaction, we ran a 2-way ANOVA for Mirrored and Control conditions separately. The 2-way interaction was not significant for the Mirrored condition in either raw or residual judgment times, but it was significant for the Control condition in both raw [F(1, 36)=7.86, p=.008] and residual judgment times [F(1,36)=6.88, p=.013].

To identify the source of the 2-way interaction in the Control condition, we conducted dependent-samples t-tests. There was a significant difference between raw judgment times for toward and away sentences after away practice [t(18)=3.420, p=.003], but not after toward practice [t(18)=.654, p=.522]. Similarly in residual judgment times, there was a significant difference between toward and away sentences after away practice [t(18)=2.575, p=.019], but not after toward practice [t(18)=1.235, p=.233].

We are aware that there are other ways to analyze the data that can take the nested design into consideration, and these alternatives are currently being explored.

Discussion

The aim of this study was to test one potential mechanism of social language coordination. Our results indicate an interaction between socially observed actions and language processing and support the hypothesis of Glenberg, Sato, and Cattaneo (2008) that action controllers in Broca's region are involved in comprehension of language describing concrete or abstract transfer. This study also adds to this hypothesis, suggesting that action controller output may increase during observation of others' similar actions. This finding implicates a mirror-neuron-like mechanism in mediating language comprehension and conversation.

As predicted, the pattern of results in the Mirrored condition indicates the fatigue of action controllers through simultaneous self-produced action and observation of action in the MNS. Participants in the Mirrored-toward practice direction, as expected, read toward sentences more slowly than away sentences. Participants in the Mirrored-away practice condition similarly demonstrated the expected pattern, judging away sentences more slowly than toward sentences.

In contrast, participants' judgment times in the control condition seem to reflect the opposite, or a facilitation effect. Participants in the Control-toward condition read toward sentences faster than away sentences and participants in the Control-away condition read away sentences faster than toward sentences. This finding is somewhat consistent with our prediction of a reduced fatigue effect in the control condition. We may attribute the discrepancy to an adjustment in procedure. Whereas participants in the Glenberg, et al. (2008) study transferred 600 beans in each condition, those in our study transferred only 300. Thus, the Control condition activates action controllers, although not to the point of fatigue. In this case, we would expect a pattern similar to an action-sentence compatibility effect (ACE) in which reading times are shorter when there is a match between the direction of motor response and the direction implied by the sentence. The fatigue effect found in the Mirrored condition would have resulted from the dual action and observation of movement, more closely approximating the experience of moving twice as many, or 600, beans.

The findings are generally consistent with several areas of research. The results support embodied theories of language

comprehension in which action systems of the brain play a role in processing of language about actions (Glenberg & Kaschak, 2002). In particular, we replicate the findings of Glenberg et al. (2008) in which action induced motor plasticity affected language processing. Here however, we extend the source of neural plasticity from action-induced fatigue of action controllers to socially induced fatigue of action controllers, in which the MNS is the hypothesized mechanism.

Our findings differ from those of Glenberg et al. (2008) by revealing a U-shaped effect of motor practice on the output of action controllers, with smaller amounts of practice leading to facilitation, and larger amounts of practice leading to fatigue.

Second, the findings support the existence of a MNS in humans in which the observed actions of another are processed using the motor system of the observer (Rizzolatti, & Craighero, 2004). Studies of the human MNS have shown that action observation potentiates the execution of kinematically similar actions in an observer (Calvo et al., 2005; Stefan et al., 2005). Similarly, it was recently found that concurrent observation of a similar action not only produces a kinematically specific motor memory in the observer, but also enhances the effect of training, relative to physical training alone (Stefan, Classen, Celnik, & Cohen, 2008). We found that observation of a kinematically similar action contributes to a fatigue-like effect associated with neural plasticity. To our knowledge, this is the first demonstration that action observation elicits practice-like effects in language comprehension.

Third, the results are consistent with the view of language as fundamentally a joint action (Clark, 1996) in which successful communication of meaning is achieved through alignment of mental states (Garrod & Pickering, 2004; Pickering & Garrod, 2006).

The results can also be compared with the literature on S-R compatility (e.g., the "Simon Effect"). Whereas that literature has shown that motor responses can reflect the "fatigue" of a spatial features of an irrelevant stimulus (e.g. Proctor & Lu, 1999), we show that such a fatigue effect can be modified by observation of another person doing a related movement.

This study suggests a mechanism by which alignment takes place, namely by the matching of motor states via the mirror neuron system. Interlocutors converge in terms of linguistic features, including grammatical structure (Bock; Branigan, 2000), word use, sematics (Clark & Wilkes-Gibbs, 1986), speech characteristics (Giles, H., Coupland, N., & Coupland, J., 1992), and phonetics (Pardo, J. S., 2006). But motor behavior also converges in social interaction (Chartrand & Bargh, 1999), particularly when there is a desire to create rapport (Lakin & Chartrand, 2003). Our results may shed light on the recent finding that physiological concordance correlates with client-therapist bond (Marci, Ham, Moran, & Orr, 2007). An interesting question is whether dyads in our study would report a greater sense of rapport in the Mirrored versus Control condition.

Recent theory suggests that the function of the MNS is for interpersonal coordination, rather than imitation of actions (Newman-Norlund, van Schie, van Zuijlen, & Bekkering, 2007), although the evidence for this view is equivocal (Kokal, Gazzola, & Keysers, 2009). Because our movement conditions differed only in terms of the similarity of movement rather than the coordination required by the task, our results support the view that the MNS is involved in imitation.

Nevertheless, understanding how the MNS interacts with brain mechanisms for interpersonal motor coordination is likely to shed light on how conversational alignment supports joint actions in communication.

Acknowledgments

We would like to acknowledge the invaluable research assistance of Katie Krol, and Arbor Otalora-Fadner, and Arthur Glenberg for his helpful comments and support.

References

Aziz-Zadeh, L., Wilson, S., Rizzolatti, G., & Iacoboni, M. (2006). Congruent embodied representations for visually presented actions and linguistic phrases describing actions. *Current Biology, 16*, 1818-1823.

Blakemore, S. J., & Decety, J. (2001). From the perception of action to the understanding of intention. *Nature Reviews Neuroscience, 2*, 561-567.

Branigan, H. P., Pickering, M. J., & Cleland, A. A. (2000). Syntactic co-ordination in dialogue. *Cognition, 75*, B13-25.

Calvo-Merino, B. Glaser, D. E., Grezes, J., Passingham, R. E., & Haggard, P. (2005). Action observation and acquired motor skills: an fMRI study with expert dancers. *Cerebral Cortex, 15*, 1243-1249.

Carr, L., Iacoboni, M., Dubeau, M.-C., Mazziotta, J. C., & Lenzi, G. L. (2003). Neural mechanisms of empathy in humans: A relay from neural systems for imitation to limbic areas. *Proceedings of the National Academy of Sciences, 100*, 5497-5502.

Chartrand, T. L., & Bargh, J. A. (1999). The chameleon effect: The perception-behavior link and social interaction. *Journal of Personality and Social Psychology, 76*, 893-910.

Clark, H. H. (1996). *Using Language*. Cambridge University Press.

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition, 22*, 1-39.

Di Pellegrino, G., Fadiga, L., Fogassi, L., Gallese, V., & Rizzolatti, G. (1992). Understanding motor events: a neurophysiological study. *Experimental Brain Research, 91*, 176-180.

Fadiga L., Fogassi L., Pavesi G., & Rizzolatti G. (1995). Motor facilitation during action observation: a magnetic stimulation study. *Journal of Neurophysiology, 73*, 2608-2611.

Garrod, S., & Anderson, A. (1987). Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition, 27*, 181-218.

Garrod, S., & Pickering, M. J. (2004). Why is conversation so easy? *Trends in Cognitive Sciences, 8*, 8-11.

Gallese, V. (2003). The roots of empathy: The shared manifold hypothesis and the neural basis of intersubjectivity. *Psychopathology, 36*, 171-180.

Gallese, V. (2008). Mirror neurons and the social nature of language: The neural exploitation hypothesis. *Social Neuroscience, 3*, 317-333.

Gallese, V., Keysers, C., & Rizzolatti, G. (2004). A unifying view of the basis of social cognition. *Trends in Cognitive Sciences, 8*, 396-403.

Giles, H., Coupland, N., Coupland, J. (1992). Accommodation theory: communication, context and consequences. In *Contexts of Accommodation* (Giles, H. et al., eds.), Cambridge University Press.

Glenberg, A., & Kaschak, M. (2002). Grounding language in action. *Psychonomic Bulletin & Review, 9*, 558-565.

Glenberg, A., Sato, M., Cattaneo, L. (2008). Use-induced motor plasticity affects the processing of abstract and concrete language. *Current Biology, 18,* R1-R2.

Glenberg, A.M., Sato, M., Cattaneo, L., Riggio, L., Palumbo, D., Buccino, G. (2008). Processing abstract language modulates motor system activity. *Quarterly Journal of Experimental Psychology, 61*, 905-919.

Grice, H. P. (1975). Logic and conversation. In P. Cole and J. L. Morgan (Eds.), *Syntax and semantics, Vol. 3: Speech acts*. New York: Seminar Press.

Iacoboni, M. (2005). *Understanding Others: Imitation, language, and empathy. Perspectives on imitation: From neuroscience to social science: Vol. 1: Mechanisms of imitation and imitation in animals*. Cambridge, MA US: MIT Press.

Kokal, I., Gazzola, V., & Keysers, C. (2009). Acting together in and beyond the mirror neuron system. *NeuroImage, 47*, 2046-2056.

Lakin, J. L., & Chartrand, T. L. (2003). Using non-conscious behavioral mimicry to create affiliation and rapport. *Psychological Science, 14*, 334-339.

Marci, C., Ham, J., Moran, E., & Orr, S. (2007). Physiologic Correlates of Perceived Therapist Empathy and Social-Emotional Process During Psychotherapy. Journal of Nervous and Mental Disease, 195, 103-111.

Newman-Norlund, R. D., van Schie, H. T., van Zuijlen, A. M., & Bekkering, H. (2007). The mirror neuron system is more active during complementary compared with imitative action. *Nature Neuroscience, 10*, 817-818.

Pardo, J. S. (2006). On phonetic convergence during conversational interaction. Journal of the Acoustical Society of America, 119, 2382-2393.

Pickering, M. J., & Garrod, S. (2006). Alignment as the basis for successful communication. *Research on Language and Computation, 4*, 203-228.

Proctor, R. W., & Lu, C. H. (1999). Processing irrelevant location information: Practice and transfer effects in choice-reaction tasks. *Memory & Cognition, 27*, 63-77.

Rizzolatti, G., & Arbib, M. (1998). Language within our grasp. *Trends in Neurosciences, 21*, 188-194.

Rizzolatti, G., & Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience, 27*, 169-192.

Rizzolatti, G., Fogassi, L., & Gallese, V. (2001). Neurophysiological mechanisms underlying the understanding and imitation of action. *Nature Reviews Neuroscience, 2,* 661-670.

Stefan, K., Cohen, L. G., Duque, J. Mazzocchio, R., Celnik, P. Sawaki, L., Ungerleider, L., & Classen, J. (2005). Formation of a motor memory by action observation. *The Journal of Neuroscience, 25*, 9339-9346.

Stefan, K., Classen, J. Celnik, P., & Cohen, L. G. (2008). Concurrent action observation modulates practice-induced motor memory formation. *European Journal of Neuroscience, 27*, 730-738.

Considering the Source: Preschoolers (and Adults) Use Talker Acoustics Predictively and Flexibly in On-Line Sentence Processing

Sarah C. Creel (creel@cogsci.ucsd.edu)
Department of Cognitive Science, University of California-San Diego
9500 Gilman Drive, La Jolla, CA 92093-0515 USA

Abstract

The identity of the person talking is likely to constrain the things that they talk about. Adults can use talker acoustics to make on-line predictions about upcoming spoken material (Van Berkum et al., 2008). However, this cue to meaning may take time to learn. Do preschoolers consider who is talking when they are comprehending spoken sentences? I explored this question in two eye-tracked picture selection experiments. Experiment 1 showed that children and adults use vocal cues to talker identity in predicting the color of upcoming referents in spoken sentences. Experiment 2 showed that children and adults flexibly use acoustic cues to talker for first-person requests ("I want the square") but reference to individuals for third-person requests ("Billy wants the square"). This suggests that children aged 3-5 years use who is talking to constrain the scope of reference in sentence processing, and know when this cue is likely to be useful.

Keywords: language development, talker identification, perspective-taking, spoken language processing

Introduction

No two people sound alike. Some research indicates that this poses a challenge for language processing (Mullennix, Pisoni, & Martin, 1989; Nusbaum & Morin, 1992). However, it may also provide additional, helpful information to the comprehender. That is, knowing who is talking can provide useful information in processing spoken language. For instance, adult listeners make different predictions about upcoming information in a sentence depending on who is speaking it (Van Berkum, Van Den Brink, Tesink, Kos, & Hagoort, 2008), suggesting they have particular semantic associations with certain voice characteristics (e.g., a child's voice vs. an adult's voice). Thus, acoustic differences among talkers potentially have rich semantic associations (Geiselman & Crawley, 1983). But how long does it take the developing language learner to form and use these associations in comprehending language?

Children are sensitive to familiar perceptual information about talkers from a very early age. For instance, they are better at generalizing words between talkers with a familiar accent than between talkers with an unfamiliar accent (Schmale & Seidl, 2009). This suggests that they are sensitive to the acoustic details in the speech signal. Less is known about how much semantic information children glean from talker acoustics. We do know that children have less positive affective responses to (Kinzler, Dupoux, & Spelke, 2007) and associate unfamiliar clothing, and dwellings with (Hirschfeld & Gelman, 1997) speakers who sound unfamiliar (they speak foreign languages). These studies suggest that children associate familiar-sounding speech with familiar objects and positive affect.

Beyond this, it is not clear whether children store more nuanced semantic information in relation to speech acoustics. This information might be somewhat difficult to learn for two reasons. First, children may be working to *ignore* talker-related acoustics to extract the attributes related to meaning (*dog* spoken by Mom still means the same thing as *dog* spoken by Dad, so why pay attention to irrelevant acoustic variation?). Second, knowing who is talking may only be useful what the person is referring to himself ("I really need a vacation") and not when talking about things irrelevant to himself ("It's raining outside"). That is, talker information may only be a reliable cue to meaning in a limited set of circumstances.

Use of other non-phonemic acoustic attributes in comprehension

Though talker information has not been explored as an influence on children's on-line sentence processing, recent studies on other non-phonemic acoustic cues—prosody and vocal affect—provide some hints about the potential of talker as a semantic information source during development. Children seem adept at processing prosodic information. Snedeker and Yuan (2008) showed that children were sensitive to a speaker's intonational phrase boundaries in their interpretations of prepositional-phrase attachment. Ito, Jincho, Minai, Yamane, and Mazuka (2009) and Bibyk, Ito, Wagner, and Speer (2009) found that children as young as 6 years use pitch accent to constrain upcoming referents to a set of items contrasting on the pitch-accented dimension. These studies suggest that children attend to non-phonemic sound patterns that cue differences in meaning.

Children seem to have more difficulty processing cues to vocal affect. Morton and Trehub (2001) found that when vocal affect conflicts with verbal content (e.g. hearing "I get to eat ice cream" in a sad voice, or "My dog got hit by a car" in a happy voice), children cannot ignore the verbal content when reporting the talker's affect (reporting the first sentence as sounding happy, and the second as sounding sad). Nonetheless, recent work by Berman, Graham, and Chambers (2009) using eye tracking, a more sensitive, implicit measure, suggests that children associate positive and negative vocal affect cues with positively- and negatively-valenced pictures (e.g. intact vs. broken dolls).

Children may be using these cues by making associations between sound properties and semantic attributes. For instance, pitch accent seems to semantically activate contrast sets. In the vocal emotion case, children might have associations between sad vocal cues and non-intact objects (Berman et al., 2009). This leaves open whether children are able to use non-phonemic acoustic information in the speech signal to make high-level inferences about the perspective of the talker.

In sum, children show some ability to glean semantic information from two non-phonemic acoustic information sources, prosody and vocal affect. Thus, one might expect that children would gain semantic information from non-phonemic acoustic cues to talker as well. However, it is not clear that children can go so far as to use it to invoke a particular talker's perspective.

The current study

To explore children's ability to exploit talker information in comprehending spoken language, I presented child and adult participants with an eye-tracked picture selection task (Cooper, 1974; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995) directed by two fictional child talkers, Anna and Billy. Each child professed a preferred color (pink vs. blue), and then asked for pictures on screen (e.g. "the square"), which were always their preferred colors. The question of interest was whether children would visually fixate the pictures in the talker's preferred color over the non-preferred color pictures based on which talker they hear.

I deliberately chose gender-stereotyped color preferences, reasoning that capitalizing on children's preexisting knowledge would minimize working memory demands that might mask sensitivity. I also queried the children's own color preferences, to determine whether they were able to predict color preferences (i.e., make looks to the talker's preferred-color pictures) when those preferences did not match their own.

In Experiment 1, I considered whether children (as well as adults) were able to use talker information early in the sentence as a cue to upcoming referent color. That is, are they able to infer what shape the talker might request, given that the talker is Anna, who prefers pink? In Experiment 2, I assessed children's flexibility in using talker information by making talker identity on its own a useless cue to referent color. Specifically, each child talker asked for a shape for herself half of the time, and for the other child the other half of the time.

Experiment 1

Method

Participants. Children (n = 24, ages 3-5 years) were recruited from local day-care and preschool facilities, and participated in the study at their day-care/preschool location. They were given a small toy as a thank-you gift. An additional two children were excluded due to high error rates (50% and 63%). Adults (n = 29) were recruited from the University of California San Diego human participant pool, and received course credit for participation.

Visual stimuli. Pink and blue squares, triangles, circles, and five-pointed stars were constructed in Microsoft PowerPoint and saved as 200 x 200 pixel .jpg files. Scenes of Anna with pink objects (a tutu, a bed, bunny slippers) and Billy with blue objects (a truck, a baseball cap, a watergun) were 1024 x 768 pixel .jpg files.

Auditory stimuli. Two native southern-Californian university students recorded requests for shapes, and descriptions of Anna's and Billy's favorite colors, in child-directed English. Recordings were made in a sound-attenuated chamber and saved to .wav files on a computer. Each utterance was edited for clarity, saved to its own sound file, and normalized to 70 dB. Target word (e.g. "square") onset was at 1003 ms after the sentence began, on average.

Procedure. Each experiment had four brief phases. During each phase, sound was presented over high-quality headphones as visual stimuli were presented on an LCD monitor. First, each talker appeared, surrounded by three pink (or blue) objects, and stated his/her preferred color. The talker named each colored object in turn. Children were then tested in their ability to distinguish the colors: on eight trials, they saw two of the same shape and heard Anna (Billy) ask "Where's the pink (blue) one?" Children did not proceed until they answered at least 7 of 8 trials in a row correctly. This verified that they could distinguish the two colors, and further reinforced each talker's preference. The two favorite-color trials were then presented again. Finally, there was a 32-trial test phase where Anna and Billy each requested objects (stars, squares, triangles, or circles). On each trial, children heard (for instance) Anna saying

(1) Can you help me find the square?

On every trial, two pictures were pink, and two were blue. Each talker requested squares, triangles, circles, and stars equally often. In this phase, neither talker used a color term, referring merely to the shapes themselves. Each shape+color combination occurred equally often in each screen position across trials. Each talker spoke on 50% of trials.

Adults clicked the desired picture with a computer mouse. Children pointed to their desired responses, which were then mouse-clicked by an experimenter. The measure of interest was whether participants, before knowing what shape was to be requested, would visually fixate pink things upon hearing Anna's voice and blue things upon hearing Billy's voice.

Equipment. The experiment was run in Matlab using PsychToolbox3 (Brainard, 1997; Pelli, 1997) and interfaced with the eye tracker using the Eyelink Toolbox (Cornelissen, Peters, & Palmer, 2002). Participants' eye movements were recorded by an Eyelink Remote eye

tracker (SR Research, Mississauga, ON) at 4-millisecond (ms) resolution. Offline, this was down-sampled to 50-ms resolution to enable easier processing.

Results

Figure 1 suggests that both children and adults were visually fixating pictures of the talker's preferred color well before the onset of the target word. To quantify this, I analyzed the data as follows. First, trials with erroneous responses (7% overall) were discarded. Then, a measure of color preference was calculated, which I will call the "color-look score." This was the proportion of looks to the non-target picture of the talker's **preferred color**, **minus** averaged looks to the two **nonpreferred-color** pictures. When this quantity was zero, listeners were not looking at pictures of either color more than the other. When it exceeded zero, listeners were looking more toward the talker's preferred color. (Negative values would imply looks to the talker's *nonpreferred* color, but this result did not occur in the current experiment.) Bear in mind that eye movements based on spoken material were most likely planned about 200 ms before they occurred, meaning that eye movements planned based on a signal at 1000 ms will show up around 1200 ms (Hallett, 1986).

An analysis of variance (ANOVA) was calculated on participants' color-look scores in three 400-millisecond (ms) time windows, with Time Window (200-600, 600-1000, 1000-1400) and Age (child, adult) as factors. The only significant factor was Time Window ($F(2,102) = 23.49$, $p < .0001$). Individual *t*-tests indicated that both children and adults had color-look scores greater than zero—that is, they were looking more to the talker's preferred color—by 200-600 ms (children: $t(23) = 2.27$, $p = 0.03$; adults: $t(28) = 3.21$, $p = 0.003$), which was also significant at 600-1000 ms ($t(23) = 4.49$, $p = 0.0002$; $t(28) = 5.64$, $p < .0001$) and 1000-1400 ms ($t(23) = 7.35$, $p < .0001$; $t(28) = 5.99$, $p < .0001$).Thus, both groups seem to be adept at utilizing talker information to decide whose preferences to invoke.

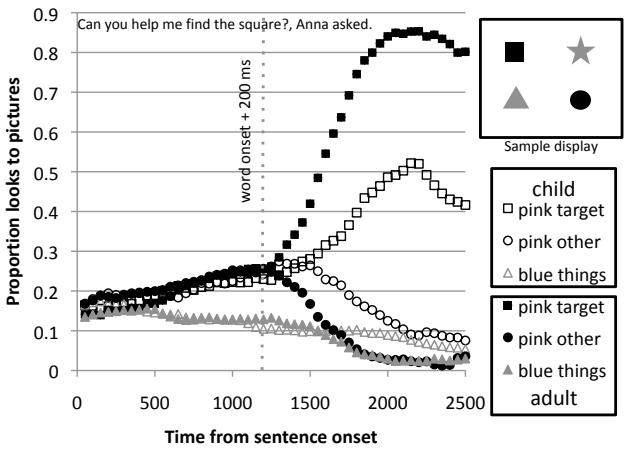


Figure 1: Adults' (solid) and children's (filled) looks to pictures in Experiment 1. Upper right inset: an example display where black=pink, gray=blue.

Note that children cannot be egocentrically fixating their own preferred color. If they were, then they should show no overall effect of the *talker's* preferred color: pink looks on pink trials and pink looks on blue trials should cancel each other out. A more subtle version of this egocentricity hypothesis is that children only fixate the talker's preferred color when it matches their own preferred color. This does not explain the results either: children whose preferred color matched neither talker ($n = 12$) still showed above-chance looks to the talker's preferred color at 600-1000 ms ($t(11) = 3.75$, $p = 0.003$) and 1000-1400 ms ($t(11) = 5.65$, $p = 0.0001$). This implies that children can use their knowledge of other individuals' color preferences, even when different from their own, to constrain the domain of reference.

Discussion

Both children and adults were able to use talker information early in the sentence to "predict" the color of the upcoming referent: they looked more at blue things when Billy began talking, and at pink things when Anna began talking. This verifies that, in a relatively simple situation, children use talker identity to constrain the referential domain of upcoming sentential material. Children showed looking effects equivalent to adults, suggesting that they are as able as adults to integrate talker information with verb information (Anna + want = pink, Billy + want = blue). This may depend on event knowledge that children have obtained through lifetime experience, or based on experimental conditions, but in either case, children are able to exercise this knowledge.

This experiment nicely demonstrates that children as well as adults are able to use talker characteristics to shape predictions of upcoming referents. One account of these data is that children and adults are using talker information to decide whose preferences to invoke to determine upcoming reference—they are constraining the domain of expected reference by talker. However, another explanation is that participants made a simple low-level audio-visual association between talker-related acoustic properties and color. That is, they associated the sound of a talker's voice with pinkness or blueness, rather than using talker acoustics to access a representation of the talker as an individual with a color preference. On this latter account, they might look at pink things even if Anna were to say "Let me out of this cage" because her voice is associated with pink things.

Related to this issue is whether children are aware of contexts where talker information is even useful in real-world language processing. In particular, talker identity in the real world may only be useful for prediction when the talker is talking about himself. When the talker is talking about someone else—for instance, if Billy said that Anna wanted to see a particular shape—it would be disadvantageous to activate colors associated with Billy's voice. This means that a smart listener would be able to use talker information in some (first person) situations, and ignore it in other (e.g. third person) situations. Presumably

adults do this readily, but it is unclear whether children do so.

Experiment 2

Experiment 2 explored whether children and adults were able to use talker information to activate characteristics (i.e., color preferences) of each individual. The experiment was introduced as before, but now in the test phase each talker asked for a shape either for herself or for the other talker, followed by "Can you show me/him/her where it is?":

(2) Anna: I want to see the square. Can you …
(3) Billy: Anna wants to see the square. Can you …
(4) Billy: I want to see the square. Can you …
(5) Anna: Billy wants to see the square. Can you…

If children are learning low-level auditory-visual associations between talkers and colors, they should fixate pink things for (2) and (5) and blue things for (3) and (4). However, if they are learning information about *individuals,* then they may use talker information only in first-person cases, and use reference to Anna or Billy in third-person cases, to determine whose preferences to invoke. If so, they should look to the *agent's* preferred color-pictures, looking at pink things in (2) and (3) and blue things in (4) and (5).

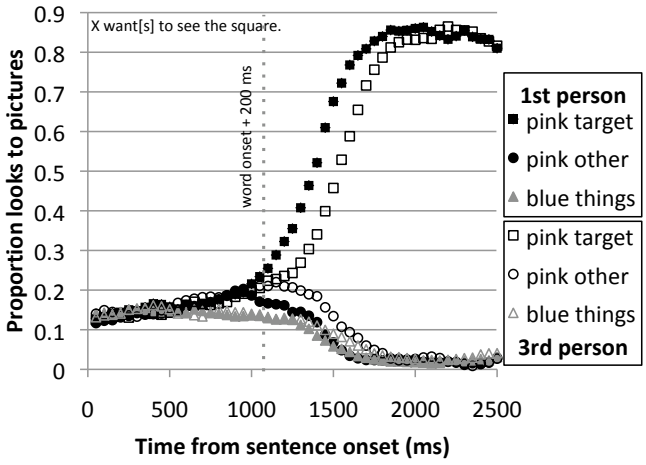


Figure 2: Adult fixations to targets and other pictures on 1st-person (circles) and 3rd-person (squares) trials.

Methods

Participants. Children (n = 33) and adults (n = 39) were recruited as in Experiment 1. Two more children with extremely high error rates (34% and 44%) were excluded.

Auditory stimuli. A new set of spoken instructions were recorded by the same individuals as in Experiment 1.

Procedure and Equipment. These matched Experiment 1 in all respects.

Results

Both adults (Figure 2) and children (Figure 3) seem to use talker information flexibly: when Anna is the agent of the sentence, they fixate pink things, regardless of whether Anna is the person talking. There were also somewhat later target fixations in the 3rd-person condition than in the 1st-person condition. While visually striking, this simply results from the 3rd-person sentences being slightly longer in duration than the 1st-person sentences (averaging 970 ms to word onset vs. 798 ms to word onset, respectively).

Error trials (5%) were discarded. Then, I conducted an ANOVA on color-look scores with Age (child, adult), Time Window (200-600, 600-1000, 1000-1400) and Person (1st person, 3rd person) as factors. This bore out the above observations. There was an interaction of Age x Time Window x Person ($F(2,140) = 5.18$, $p = 0.007$), so individual ANOVAs were conducted for each Age. For adults, only Time Window was significant ($F(2,76) = 10.3$, $p = 0.0001$), with color-look scores increasing over time. T-tests indicated that both 1st- and 3rd-person trials showed significant color looks at 600-1000 ms ($t(38) = 2.13$, $p = 0.04$; $t(38) = 2.73$, $p < 0.01$), and 1000-1400 ms ($t(38) = 2.25$, $p = 0.03$; $t(38) = 4.08$, $p = 0.0002$). For children, there was an effect of Time Window ($F(2,64) = 23.48$, $p < .0001$), with color-look scores increasing over time, and a Time Window x Person interaction ($F(2,64) = 3.36$, $p = 0.04$). T-tests comparing 1st-person and 3rd-person looks suggested nonsignificant differences in each time window (only 600-1000 ms approached significance, $t(32) = 1.82$, $p = 0.08$). Regardless, both 1st- and 3rd-person color-look scores were significant at 600-1000 ms ($t(32) = 2.22$, $p = 0.03$; $t(32) = 4.84$, $p < .0001$) and 1000-1400 ms ($t(32) = 8.11$, $p < .0001$; $t(32) = 4.78$, $p < .0001$). This suggests that children, as well as adults, used the talker's voice on 1st-person trials, but reference (the child's name) on 3rd-person trials, to determine whose color preferences to use in constraining the referential domain. As before, results held for children (n= 18) whose favorite colors were neither pink nor blue.

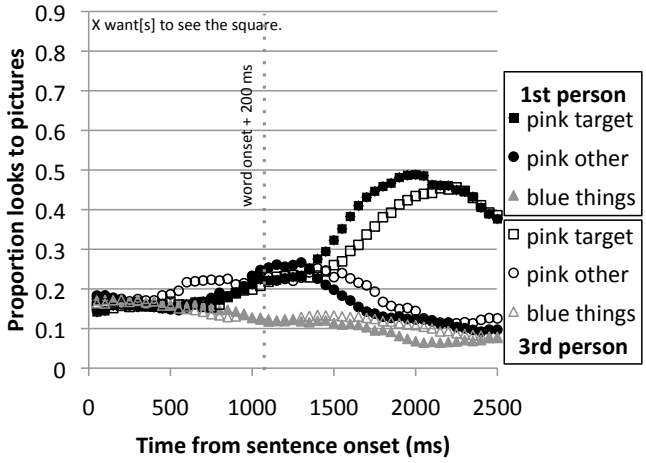


Figure 3: Child fixations to targets and other pictures on 1st-person (circles) and 3rd-person (squares) trials.

Discussion

Children and adults in Experiment 2 succeeded at predicting the agent's color preference. That is, they made more visual fixations to shapes of the *agent's* preferred color on both first-person ("I want") and third-person ("Anna/Billy wants") trials. This implies that they use talker acoustics not just as a low-level auditory-visual association, but as a source of information about a participant in an action. Thus, children as well as adults can use non-phonemic acoustic information to activate information about an individual, and then infer the likely referential domain for that individual.

General Discussion

Two experiments suggest that children are able to use their knowledge about particular talkers to constrain the domain of upcoming referents. In Experiment 1, listeners were instructed that Anna liked pink things, and Billy liked blue things. They then heard Anna and Billy request shapes of their preferred color. Both children and adults made more visual fixations to the shapes of the talker's preferred color than of the talker's nonpreferred color. This suggested that children were able to identify the talkers and use their individual preferences to constrain on-line interpretation of the request.

However, an equally good explanation was that children had associated female voice characteristics with pinkness, and male voice characteristics with blueness, a low-level auditory-visual cue correspondence rather than knowledge of an individual's preferences. Experiment 2 ruled out this explanation: listeners again heard Anna and Billy requesting shapes, but half the time, each talker requested a shape for the other talker. This meant that only on first-person trials ("I want") was talker a useful predictor, while on third-person ("Anna wants") trials, it was a misleading predictor. Impressively, children and adults were both able to use talker information on first-person trials, and proper nouns on third-person trials, to infer the identity of the sentential agent. That is, they always showed a visual fixation preference toward the *agent's* preferred-color shape, even when the agent was not the talker. This implies that, in a relatively simple task, children are able to use talker information selectively (only on first-person trials) to infer the identity—and thus the color preferences—of the agent.

Implications for development of language processing

This research adds to the existing literature on cue integration in spoken language processing. Specifically, this work demonstrates that, in addition to prosody and vocal-emotional cues, non-phonemic acoustic cues related to talker can also be used to constrain processing on-line fairly early in life. This suggests excellent facility on the part of children to use non-phonemic acoustic cues to talker identity to understand the situation described by a sentence. This work is similar to adult research by Van Berkum et al. (2008), in which listeners showed a larger semantic mismatch potential (N400) when the talker's identity and the action described were incongruous (e.g. a young child saying "I like to drink a glass of wine") than when they were congruous (an adult saying the same sentence). The current work suggests that preschool-aged children are similarly able to use talker acoustics to calculate likely (and unlikely) referents.

The current work, as well as Van Berkum's, fits nicely with a perspective on language processing (Kamide, Altmann, & Haywood, 2003; Bicknell, Elman, Hare, McRae, & Kutas, 2008) in which comprehenders use any available linguistic and nonlinguistic cues to construct event representations on-line. Acoustic information linked to talker identity is apparently useful in constructing event representations. Moreover, it is a robust enough cue that preschool-aged children can use it rapidly on-line (see Bates & MacWhinney, 1987; Snedeker & Trueswell, 2004 for further discussion of cue robustness and development).

Perhaps the most unique contribution of this study is the implication that children are using talker acoustics to infer *properties of individuals*, or at least of groups of individuals. That is, children are able to encode that Anna and Billy have particular color preferences, even when Anna and Billy have different preferences than the children themselves. As demonstrated in Experiment 2, this does not seem to be a simple auditory-visual association between Anna's voice (or female voices) and pink, and Billy's voice (or male voices) and blue, but an association with Anna and Billy as entities who have different preferences for color.

Remaining questions

One obvious question is how much of children's ability to use talker information in this task is subserved by children's long-term knowledge of gendered color preferences. A quick visual search of major toy retailers' products confirms strong tendencies for female toys to be pink (or purple), and for male toys to be blue (or a number of other colors, but not pink). Thus, children's use of talker information here could be due to a lengthy learning process through exposure to gender-stereotyped objects in their environments. On the other hand, children might readily associate idiosyncratic preferences with particular individuals. If so, then children should also be able to use learned, non-gender-stereotyped color preferences to constrain on-line language processing.

An experiment in progress addresses this question, using black and white as the preferred colors. Only one child (1.5%) in Experiments 1 and 2 reported black as his favorite color, and none reported white, suggesting that children have little experience or gender-preference information for black and white. Further, color preference is counterbalanced across talker gender. With 15 child participants so far, there are robust looks to talkers' preferred colors. This suggests that neither conformance to a gender-stereotypical color mapping nor long-term learning is necessary for children to be able to use talker information predictively. However, talker gender itself may still be an

important social anchor point for encoding talker preference.

Another question is how subtle children are in their appreciation of talker information. Are they as keen in their perceptions as adults? If not, how do they differ from adults? Direct comparisons may be limited somewhat by children's level of social knowledge relative to adults—adults may only seem more adept at using talker cues because they have more subtle knowledge of social variation.

Finally, it is unknown how semantic knowledge based on talker characteristics relates to talker-specific perceptual facilitation of word-forms (e.g. Goldinger, 1996; see also Creel, Aslin, & Tanenhaus, 2008). Does talker-specific perceptual information covary with semantic usefulness? Despite these remaining questions, though, the current research forms a solid basis for further explorations of children's sensitivity to talker as a cue to meaning.

References

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: evidence for continuous mapping models. *Journal of Memory and Language, 38,* 419–439.

Bates, E. A., & MacWhinney, B. (1987). Competition, variation and language learning. In B. MacWhinney (Ed.), *Mechanisms of language acquisition*. Hillsdale, NJ: Erlbaum.

Berman, J., Graham, S. A., & Chambers, C. (2009). Preschoolers' appreciation of vocal affect as a cue to a speaker's intentions. Paper presented at *Boston University Conference on Language Development 34,* Boston, MA.

Bibyk, S., Ito, K., Wagner, L., & Speer, S. (2009). Children can use contrastive pitch accent in on-line processing. Paper presented at *Boston University Conference on Language Development 34,* Boston, MA.

Bicknell, K., Elman, J. L., Hare, M., McRae, K., & Kutas, M. (2008). Online expectations for verbal arguments conditional on event knowledge. In B.C. Love, K. McRae, & V.M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 2220-2225). Austin, TX: Cognitive Science Society.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision, 10,* 433-436.

Cooper, R. M. (1974). The control of eye fixation by the meaning of spoken language. A new methodology for the real-time investigation of speech perception, memory, and language processing. *Cognitive Psychology, 6,* 84–107.

Cornelissen, F.W., Peters. E., & Palmer, J. (2002). The Eyelink Toolbox: Eye tracking with MATLAB and the Psychophysics Toolbox. *Behavior Research Methods, Instruments & Computers, 34,* 613-617.

Creel, S. C., Aslin, R. N., & Tanenhaus, M. K. (2008). Heeding the voice of experience: The role of talker variation in lexical access. *Cognition, 108,* 633 – 664.

Geiselman, R. E., & Crawley, J. M. (1983). Incidental processing of speaker characteristics: Voice as connotative information. *Journal of Verbal Learning & Verbal Behavior, 22,* 15–23.

Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*, 1166 – 1183.

Hirschfeld, L. A., & Gelman, S. A. (1997). What children think about the relationship between language variation and social difference. *Cognitive Development, 12,* 213 – 238.

Ito, K., Jincho, N., Yamane, N., Minai, U., & Mazuka, R. (2009). Use of emphatic pitch prominence for contrast resolution: An eye-tracking study with 6-year old and adult Japanese listeners. Paper presented at *Boston University Conference on Language Development 34,* Boston, MA.

Kamide, Y., Altmann, G. T. M., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language, 49,* 133-156.

Kinzler, K. D., Dupoux, E., & Spelke, E. S. (2007). The native language of social cognition. *Proceedings of the National Academy of Science, 104,* 12577 – 12580.

Morton, J. B., & Trehub, S. (2001). Children's understanding of emotion in speech. *Child Development, 72,* 834-843.

Mullennix, J.W., Pisoni, D.B., & Martin, C.S. (1989). Some effects of talker variability on spoken word recognition. *Journal of the Acoustical Society of America, 85*, 365-378.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10,* 437 – 442.

Schmale, R., & Seidl, A. (2009). Accommodating variability in voice and foreign accent: Flexibility of early word representations. *Developmental Science, 12,* 583-601.

Snedeker, J., & Trueswell, J. (2004). The developing constraints on parsing decisions: The role of lexical biases and referential scenes in child and adult sentence processing. *Cognitive Psychology, 49,* 238-299.

Snedeker, J., & Yuan, S. (2008). Effects of prosodic and lexical constraints on parsing in young children (and adults). *Journal of Memory and Language, 58,* 574-608.

Tanenhaus, M., Spivey-Knowlton, M., Eberhard, K., & Sedivy, J. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science, 268,* 1632-1634.

Van Berkum, J. J. A., van den Brink, D., Tesink, C. M. J. Y., Kos, M., & Hagoort, P. (2008). The neural integration of speaker and message. *Journal of Cognitive Neuroscience, 20,* 580 – 591.

Hebbian learning for deciding optimally among many alternatives (almost)

Patrick Simen (psimen@math.princeton.edu)
Princeton Neuroscience Institute, Princeton University, Green Hall, Washington Rd., Princeton, NJ 08544

Tyler McMillen (tmcmillen@fullerton.edu)
Department of Mathematics, California State University at Fullerton, Fullerton, CA 92834

Sam Behseta (sbehseta@fullerton.edu)
Department of Mathematics, California State University at Fullerton, Fullerton, CA 92834

Abstract

Reward-maximizing performance and neurally plausible mechanisms for achieving it have been completely characterized for a general class of two-alternative decision making tasks, and data suggest that humans can implement the optimal procedure. A greater number of alternatives complicates the analysis, but here too, analytical approximations to optimality that are physically and psychologically plausible have been analyzed. All of these analyses, however, leave critical open questions, two of which are the following: 1) How are near-optimal model parameterizations learned from experience? 2) How can sensory neurons' broad tuning curves be incorporated into the aforementioned optimal performance theory, which assumes decisions are based only on the most informative neurons? We present a possible answer to all of these questions in the form of an extremely simple, reward-modulated Hebbian learning rule for weight updates in a neural network that learns to approximate the multi-hypothesis sequential probability ratio test.

Keywords: Hebbian learning; diffusion model; neural network; multi-hypothesis sequential test; sequential probability ratio test; speed-accuracy tradeoff; response time

Introduction

We examine the problem of maximizing earnings from a sequence of N-alternative decisions about the identity of noisy stimuli, with $N > 2$. Our goal is to parameterize a simple neural circuit model whose behavior approximates optimal performance in such tasks, while simultaneously accounting for the fundamental role of tuning curves in the neural representation of sensory stimuli. Throughout, we take 'optimal' to mean *reward maximizing*, and we assume that correct decisions earn rewards for the decider.

As we show, simple principles of neural computation are sufficient to approximate this form of optimality quite closely in a class of N-choice tasks involving response-terminated stimuli: that is, stimuli that provide information continuously until the time (the response time) at which participants decide for themselves when to stop observing and make a response. This is somewhat surprising, given that a general decision policy that guarantees truly optimal performance cannot even be explicitly formulated for such tasks, as we discuss below.

N-choice, response-terminated decision tasks

We assume that participants earn rewards for correct responses, and earn less for errors (for simplicity, we assume errors earn nothing). In the simple tasks we consider, each stimulus type has a fixed prior probability within a block of trials, and the average signal-to-noise ratio of each stimulus is fixed. The duration, rather than the number of trials, is also held fixed, and the distribution of response-to-stimulus intervals (RSIs) that delay the onset of the next stimulus after a response is stationary. In this case, maximizing the rate of reward also maximizes the total reward.

Maximizing gains in this and a variety of similar tasks requires probabilistic inference. While the importance of a principled inference process is widely understood in psychology and neuroscience, the complexity of optimal decision policies in tasks with response-terminated stimuli (also known as 'free response' or 'response time' tasks) and $N > 2$ choices appears to be less well appreciated.

For 2-choice tasks of the type just described, reward-maximizing performance has been completely characterized (Bogacz et al., 2006): a sequential probability ratio test (SPRT) should be carried out in which the current likelihood ratio of the two hypotheses is multiplied by the probability of a given data sample under one hypothesis and divided by the probability of that data sample under the other hypothesis (equivalently, the logs of these probabilities can be added and subtracted, respectively — from now on, we will cast our discussion in terms of log-likelihoods). A response should be made when the resulting log-likelihood exceeds a fixed threshold (Wald & Wolfowitz, 1948). There exists an optimal starting point of the log-likelihood ratio (e.g., 0, for equally likely stimuli) and an optimal separation between the two response thresholds (one greater and one less than zero) that depends on the signal-to-noise ratio (SNR) and the RSI (Bogacz et al., 2006). Gold and Shadlen (2001) have demonstrated that for systems consisting of a neuron/anti-neuron pair, each of which is tuned for one of the two stimulus types in a 2-choice task, the log-likelihood ratio is approximately proportional simply to the difference between the activations of the two neurons, suggesting an extremely simple neural implementation of the SPRT.

In contrast, if the number of choices is greater than 2, the optimal policy for deciding based on accumulated information is nontrivial. In particular, a natural (but definitely suboptimal) approach to N-choice decision making is to compute the posterior probability of each of the N hypotheses, and then select whichever one first exceeds a fixed threshold. In fact, the best decision is made when the entire set of posterior probabilities meets conditions that are nontrivial func-

tions of the posterior values. A thought experiment may help make clear why this is true. Consider the case of a 3-choice task in which one choice has attained an 80% posterior probability of being correct, while the other posteriors are 10% and 10%. A fixed 80% threshold will therefore not distinguish this case from a case in which the posteriors are 80%, 19% and 1%. These two cases are quite different, however, and dealing optimally with them requires taking account of all posterior probabilities in a more sophisticated way. Because of this, and because of the inherent difficulty in defining truly optimal decision policies to apply to the posteriors, multi-hypothesis sequential probability ratio tests (MSPRTs) were designed to approximate optimality with a decision policy consisting of fixed thresholds applied to posteriors or likelihood ratios (Dragalin, Tartakovsky, & Veeravalli, 1999).

Tuning curves

Tuning curves are ubiquitous in neural responses to stimuli (Butts & Goldman, 2006). The relationship between tuning curve shape and decision making performance has intrigued researchers for several years (e.g., Pouget, Deneve, Ducom, & Latham, 1999). Naively, one may suppose that task participants improve their performance by sharpening the tuning curves of the neurons involved. However, wider tuning curves are in some cases more efficient in conveying information, and the most informative tuning curve shape depends strongly on the noise and correlations (Zhang & Sejnowski, 1999; Seriés, Latham, & Pouget, 2004). Moreover, in several tasks, participants may improve performance without significantly altering the shapes of the tuning curves in the neurons involved. For instance, in an angle discrimination task, monkeys are able to learn to discriminate between finer angles over time, while the tuning curves of primary sensory neurons are altered very little (Ghose, Yang, & Maunsell, 2002; Law & Gold, 2008). This suggests that improvements in performance take place in a learning process downstream from the receptor neurons.

In this paper we explore the ways in which a subject may improve performance in decision tasks, given tuning curve shapes in receptor neurons. We do not consider the alteration of receptor units' tuning curves, but rather how the information in tuning curves can be utilized more efficiently over the course of many trials.

The leaky competing accumulator model for decision making

We propose a three layer neural model for decision making (defined in Table 1, and depicted in Fig. 1). The first layer acts simply as a sensory amplifier; the next layer integrates the information from the first layer, but also exhibits competitive dynamics that gradually build a commitment to one course of action over the alternatives; the last layer triggers a discrete motor response when commitment to one response is sufficiently strong. For convenience, we refer to these three layers, respectively, as the MT, LIP and SC layers. These labels reflect the fact that our model exhibits known properties of neurons in the monkey middle temporal area (MT), the lateral intraparietal area (LIP) and the superior colliculus (SC) in decision making tasks requiring eye movements in response to visual motion stimuli (i.e., random dot kinematograms; Shadlen & Newsome, 2001). The architecture of this circuitry is expected to apply without major modification to other stimulus and response types, however.

Table 1: Three layer model with weight learning rule.

$S_i,\ i = 1, \ldots, n$	input signals (MT)
$dx_i = \left(-k x_i - m \sum_{j \neq i} x_j + S_i \right) dt + \ldots$ $c\, dB_i$	accumulators (LIP)
$z_i = H(y_i - \Theta), \quad y_i = \sum_{j=1}^{n} w_{ij} x_j$	decision units (SC)
$w_{ij}^{\text{new}} = (1-\alpha) w_{ij}^{\text{old}} + \alpha \Delta w_{ij}$ $\Delta w_{ij} = r z_i x_j$	LIP to SC weight-learning rule

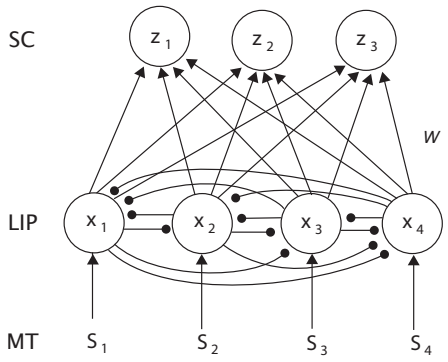


Figure 1: Neural network model with 4 accumulators and 3 alternatives. The weight matrix W denotes the weights of the connections between the x_i's and z_j's. Arrows represent excitatory connections; circles represent inhibitory connections.

We suppose that MT neurons have tuning curves that are preferentially sensitive to a single, given direction of visual motion, and that another layer is stimulated by the activity in this input layer. By virtue of their excitatory connections to LIP, model MT units' tuning curves and their feedforward connections to LIP in turn define tuning curves for LIP units. Questions of major importance in computational neuroscience are: Through what sort of learning process do these tuning curves arise? Can we define an optimal connection scheme that maximizes some function, such as the rate of reward earned by the model? We attempt to make progress on these questions while making the simplifying assumption that the brain circuits in question are approximately

linear systems (at least over a limited range of inputs), and that they employ simple learning schemes (such as Hebbian learning, or error-updating rules such as the Widrow-Hoff, Rescorla-Wagner or delta rule). Recent work (e.g. McMillen & Holmes, 2006; Bogacz & Gurney, 2007) that avoids discussion of tuning curves and learning shows that these assumptions allow simple neural network models to map precisely onto one or another form of MSPRT. We now demonstrate that a similar model that learns connections strengths and accounts for tuning curves does remarkably well at approaching optimal (reward maximizing) performance in decision making tasks with multiple alternatives. The model's layers are represented mathematically by S, x, and z.

Upon presentation of a stimulus, the model's MT layer presents a vector of signals to accumulators in the LIP layer. The signal presented to the ith unit in the LIP layer is referred to as S_i, representing the total weighted sum of MT signals to the ith accumulator. Each stimulus corresponds to a unique signal, so that the set of signals to the LIP layer may be represented as a vector indexed by μ:

$$S^{\mu} = \left(S_1^{\mu}, S_2^{\mu}, \ldots, S_n^{\mu}\right).$$

The task is to determine which of N possible signal vectors this represents. Notice that the number of vectors can be different from the number of signals, i.e. in general $n > N$.

Although it is not required, we will generally take the S^{μ} signals to be Gaussian:

$$S_i^{\mu} = a \exp\left[-\frac{(i - dir_{\mu})^2}{2\phi^2}\right],\ i = 1, \ldots, n. \qquad (1)$$

Here dir_{μ} is the peak of the signal, a is the height of the peak and ϕ is the width of the curve. As in McMillen and Behseta (2010), we interpret the S^{μ} in terms of approximately Gaussian MT tuning curves and weights from MT to LIP that preserve this Gaussian tuning in the LIP units. Notice that if $\phi = 0$, then

$$S_i^{\mu} = a\,\delta_{i,dir_{\mu}},$$

where $\delta_{i,j}$ is the Kronecker delta, so that the signal is concentrated in the channel dir_{μ}. But, if $\phi > 0$, the signal will have a spread around the peak. For MT units associated with the dot-motion task, tuning curves have been measured to have a width of about 40° (Law & Gold, 2008). The situation is illustrated in Fig. 2. Angles far apart have very little overlap in the signals, but when the angles are close the overlap is substantial. For a two-alternative task in which dots travel on average either up or down, the signals have very little overlap. Signals for alternatives corresponding to more similar motion directions have more overlap.

We model the LIP layer as a set of n leaky competing accumulators. The linearized model for their evolution is a stochastic differential equation (Usher & McClelland, 2001; Bogacz et al., 2006; McMillen & Holmes, 2006):

$$dx_i = \left(-kx_i - m\sum_{j \neq i} x_j + S_i\right) dt + c\,dB_i,\ \ i = 1, \cdots, n, \quad (2)$$

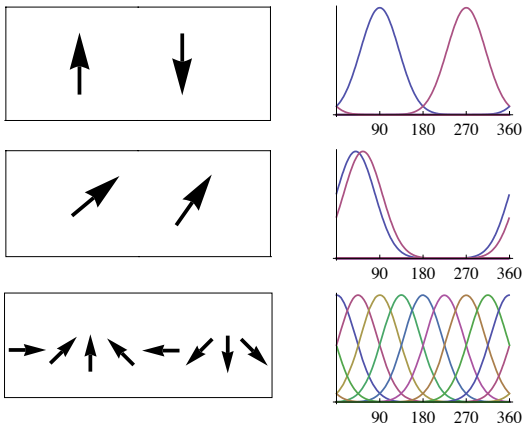


Figure 2: Possible directions of coordinated movement (left panels) and corresponding signal vectors (right panel).

where k is the decay rate, m is the mutual inhibition, and B_i is a Wiener process (integrated white noise) representing the noise in the signal and from other sources. The signal-to-noise ratio is the ratio of the magnitude of the largest signal to the variance of the noise, i.e. a/c. We can thus model changes in the direction coherence by changing this ratio. The effect of decay and inhibition is to concentrate the values of the accumulators onto the signal vectors. Thus, moderate values of w and k tend to increase the accuracy. Best results are achieved when decay and inhibition are balanced, i.e. $w = k$ (McMillen & Holmes, 2006). For simplicity, and to be concrete, throughout the rest of this paper we will present results for $k = w = 0.5$, $a = 2$ and $c = 1$. Results are qualitatively insensitive to these choices.

The output from accumulator j feeds into the ith unit of SC with weight w_{ij}. SC units apply step functions H with thresholds Θ to their inputs. A response is made when SC unit j transitions from 0 to 1 (i.e., when $y_j = \sum_{j=1}^{n} w_{ij} x_j > \Theta$).

The results in this paper are generally applicable, but to be precise we consider a motion direction task with 36 accumulators and interpret these as representing increments of 10°. If the direction $j \cdot 10°$ is presented, the signal vector takes the shape S_i^{μ} as in (1), with $dir_{\mu} = j$. For concreteness we consider four possible directions of motion: $30°, 60°, 140°, 220°$. Thus, if say, the direction of coordinated movement is 60°, the signal vector has a peak at the sixth accumulator. The four possibilities are represented by the four possible signal vectors with peaks at accumulators 3, 6, 14 and 22. In this paper we only consider the case when all the possibilities are equally likely, in which case the appropriate initial condition for the accumulators is $x_i(0) = 0$.

McMillen and Behseta (2010) showed that the optimal weights w_{ij} in the above are achieved when the weights mimic the shape of the possible incoming signal vectors. That is to say, a threshold crossing test best approximates an MSPRT when $w_{ij} = S_j^{\mu_i}$. The magnitude of the weights

are not important in terms of optimality, as the magnitude may be incorporated into the thresholds. The performance of the threshold crossing tests is illustrated in Fig. 3. Here we consider a test with 36 accumulators and the four alternatives as described above. In Fig. 3 we plot the mean response time (MRT) for a fixed value of the error proportion (ER). For each value of the spread we compute the threshold such that ER $= 0.1$, and find the corresponding MRT. Each panel demonstrates an important fact, as we elucidate below.

In the left panel of Fig. 3 we take the signal vectors to be as (1), and allow ϕ to vary. Thus, $\phi = 0$ corresponds to the case when the signal is concentrated in a single channel. Positive values of ϕ correspond to signals that are spread about a peak. In these computations, the weights are as desired for MSPRT approximation, i.e. $w_{ij} \propto S_j^{\mu_i}$. This panel thus shows the minimal MRT that can be achieved by a threshold crossing test for an ER of 0.1. We see that there is an advantage to a moderate spread in the signals if this information can be utilized by the decision mechanism. In fact, the optimal spread is near $\phi = 3$. It is interesting to note that this corresponds to a width in the shape of the signal vectors of about 30°, while the width of tuning curves in MT associated with the direction task as measured in Law and Gold (2008) are approximately 40°.

In the right panel we fix the spread in the signal vectors at $\phi = 4$, and compute MRT for various spreads in the weights. In order to get an idea of how the spread in the shape of the weights affects performance when the signal shape is fixed, in these simulations we suppose that the weights also have a Gaussian shape:

$$w_{ij} = w_0 \exp\left[-\frac{(i-j)^2}{2\phi_W^2}\right],\ j = 1,\ldots,n,$$

where w_0 is a normalizing factor chosen so that $\sum_{j=1}^n w_{ij}^2 = 1$ (this normalization step is not in fact required). The spread ϕ_W controls how the values of the accumulators are weighted before making a decision. In the case $\phi_W = 0$, we have $y_i = x_i$, so that the accumulator values are not weighted. When $\phi_W = \infty$, each y_i is the same, i.e. the sum of all accumulators. The right panel of Fig. 3 shows that MRT is minimized when $\phi_W = \phi$. That is, the optimal weights occur when the width of the weight shape is the same as that in the signal vector.

To reiterate, a moderate spread in the signals is a significant advantage, but only if the LIP-to-SC weights can be tuned to take on the same shape as the possible signal vectors defined by MT activity. In the following section we consider how the weights may be modified over the course of trials.

An algorithm for learning the LIP to SC weights

We propose a simple Hebbian weight learning algorithm for the weights w_{ij}. The learning algorithm is a modification of a classical Widrow-Hoff rule (see, e.g., Hertz, Krogh, & Palmer, 1991). In rules of this type, the connection strength being modified acts as a filter that tracks an input signal. At

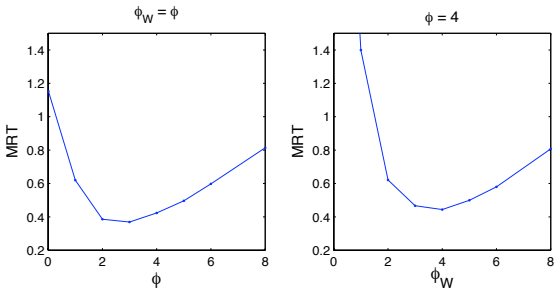


Figure 3: Effects of signal spread and weight shape. Left panel: Simulated MRT vs. spread in the signal vectors, where the weights have the same shape. Right panel: MRT vs. spread in shape of weights with signal vector fixed with $\phi = 4$. In all cases the threshold is such that ER $= 0.1$.

any point, its value is an approximately exponentially decaying, weighted average of past input values. High frequency changes in this signal (representing noise) are filtered out by the algorithm, producing little change in the updated weight. In contrast, low frequency signal changes (representing, hopefully, the uncorrupted input signal) produce significant changes in the weight. If the signal is constant and noise is absent, the weight will converge approximately exponentially on the value of the signal. If what is being tracked is a signal that depends on the product of activations in a sending unit and a receiving unit, then this rule is simply a Hebbian update rule with a decay term for forgetting old co-activation levels — a useful feature in a noisy neural system.

After each trial the subject responds with a choice among alternatives, say i. At this time the weights to the output unit z_i corresponding to the choice made are updated, according to whether a reward is received or not. Then, if the choice corresponding to z_i is chosen, the weights are updated by the rule

$$w_{ij}^{\text{new}} = (1-\alpha) w_{ij}^{\text{old}} + \alpha \Delta w_{ij}\,, \quad (3)$$

$$\Delta w_{ij} = r z_i x_j\,, \quad (4)$$

where r is the magnitude of the reward, and α is the learning rate. Notice that only the weights to the unit corresponding to the choice made are updated, and this is the sense in which the rule is Hebbian. For simplicity, we assume here that a reward is either earned or not, so that r is either 1 or 0 depending on whether a correct decision is made.

Thus, after each trial, if a correct decision is made the weights to the correct output unit are increased in proportion to the values of the accumulators $\mathbf{x}$. There is no need to estimate the probability of making a correct decision or an expected value of the reward, as for example in reinforcement learning methods, since only the values of the units are used in the update rule. With this rule the weights track the shape of the vectors being passed from the LIP layer. The weights thus tend to oscillate around the means of the accumulator

values, $\langle x_j(t) \rangle$.

The accumulator values on average take on the shape of the signal vector from the MT layer. This can be proved analytically, but here we show only simulation results. The update rule (3-4) thus causes the weights to track values whose means take on the shape of the MT-to-LIP signal vectors. Therefore the weights tend, on average, to mimic the shape of the signal vector, with oscillations about this shape that depend on the learning rate.

Results of simulations

Figs. 4 and 5 show results of simulations using the update rule (3 - 4). The weights are initially chosen randomly, with a peak added at w_{ii}. Fig. 4 shows how the weights evolve over time, and how this affects the performance of the subject. The reward rate continually increases on average, and the ER continually decreases. The bottom panels show the weights to SC corresponding to $i = 14$, or to angle 140°. The weights for the other alternatives behave similarly. Simulations in which the weights are chosen differently show similar improvements in performance and similar matching of the weight profiles to the signal vector shapes. Cases in which the weights are all chosen randomly show a more dramatic improvement in reward rate (RR) since then the accuracy will initially be very low. Fig. 4 shows that even when the weight has a peak at the right position, a dramatic improvement occurs: for example, the RR more than doubles and the RT and ER both decrease over time.

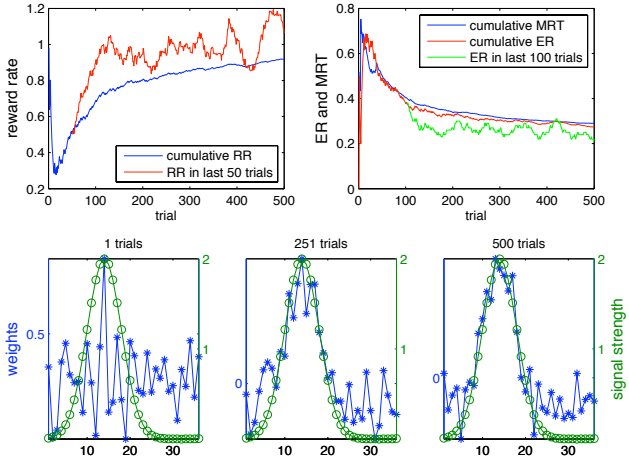


Figure 4: Effects of weight learning rule. The threshold is fixed at $z = 1$. There are four alternatives (3, 6, 14, 22), and the learning rate is $\alpha = .05$. In the bottom panel the signal strength is plotted on the right axis (circles), and the weights are shown on the left axis (stars). The inter-trial delay used in the calculation of reward rate (RR) here is 500 msec.

Figure 4 shows one block of 500 trials. In order to see how the weight update rule behaves on average, we carried out the same simulation for a number of blocks and averaged

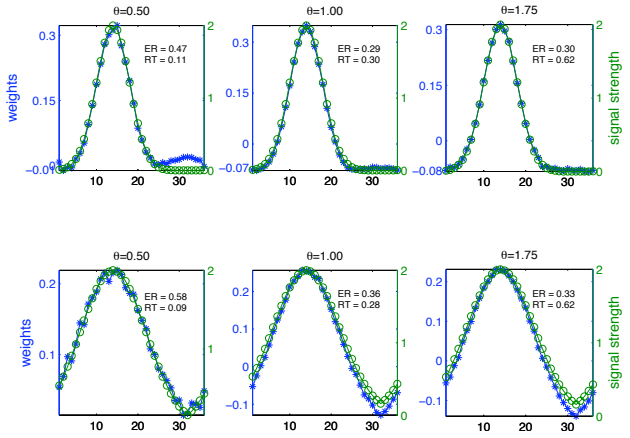


Figure 5: Averaged weights over 150 blocks of 500 trials. In the top row $\phi = 4$; in the bottom row $\phi = 8$.

the weights over each block, and then took the average over 150 blocks of trials. Fig. 5 shows the averaged weights for different values of the threshold, as well as different values of the spread in the signals. We see that on average, the weight profile shape is very close to the signal shape. Also indicated in these figures are the ERs and MRTs for these blocks of trials. Notice that in the lower left panel, the ER = .58 is not much smaller than would be achieved by random guessing (.75). In this case the threshold is very small, as is the corresponding MRT of .09. In this situation it will take the weights much longer to learn the shape of the signal vectors, since most of the time the decision will be incorrect. This is why the weights appear more erratic in this frame than in the others. However, even in this case, the average values of the weights take the same shape as the signal vector. Similar comments apply, *mutatis mutandis*, to the upper left panel.

Generally, the model is insensitive to changes in the parameters a, c, k, m, in the sense that the weights tend on average toward the optimal weight shape mimicking the shape of the signal vectors. If the learning rate α is made smaller, the weights take longer to track to the shape of the signals, but there is less variation around these mean values.

Fig. 6 shows the dynamics of evidence accumulation within trials, demonstrating that Gaussian bumps of activation arise on the LIP layer (preserving the Gaussian input signal profiles, and therefore producing Gaussian LIP-to-SC weights through Hebbian learning).

Discussion

The simple rule (3-4) works remarkably well at learning the shapes of the signal vectors from MT to LIP. This leads to a dramatic improvement in performance, and occurs without any direct connection to the MT layer. The three layer model incorporates integration of information, a rule for making the decision, as well as a simple algorithm for learning to optimize reward rates by learning the shapes of the vectors of

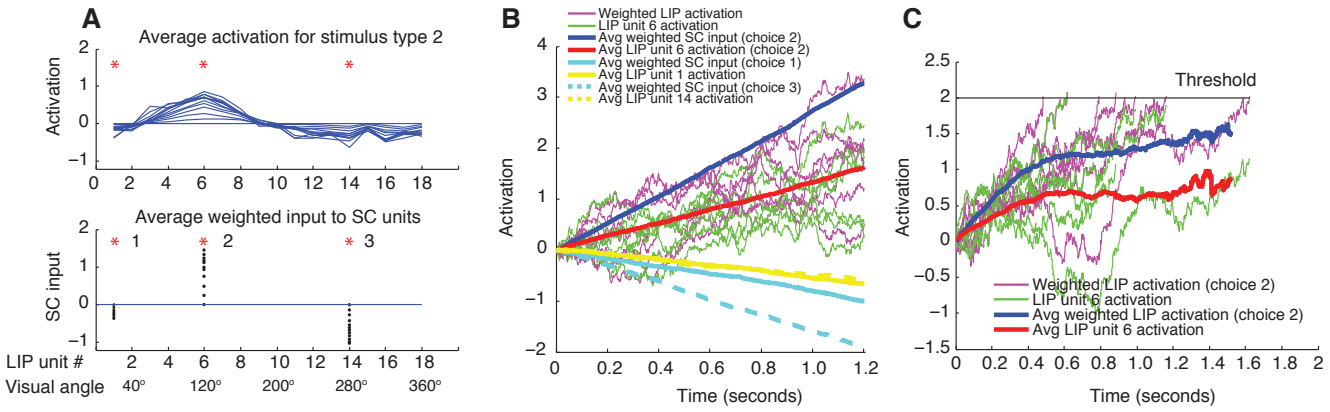


Figure 6: Panel A, top, shows LIP unit activations at several time points within a decision. Activations are averaged over many instances of stimulus type 2, which produces maximal activation in LIP unit 6 (visual angle 120°; here we arbitrarily quantized visual angle into 18 levels). Panel A, bottom, shows the weighted values of these activations feeding into each of 3 SC units. Panel B shows the average state of evidence accumulation for choice 2 (input to SC unit 2; blue) and average LIP unit 6 activity (red) within fixed-viewing time trials, without thresholds applied to the evidence (the interrogation protocol). Panel C shows the average state of weighted evidence accumulation for choice 2 (blue) and average unit 6 activity (red) within free response trials (black line indicates threshold). The weighted sum produces a higher signal-to-noise ratio, and therefore better performance than evidence from unit 6 alone. Red and blue traces fall off over time because the average is based on fewer and fewer trials as time progresses (more and more decisions have already taken place by the end of the plot).

neural signals coming from an input layer. These features are essential elements of a complete decision-theoretic model.

Acknowledgments

P. Simen was supported by a postdoctoral training fellowship from the National Institute of Mental Health (MH080524).

References

Bogacz, R., Brown, E., Moehlis, J., Hu, P., Holmes, P., & Cohen, J. (2006). The physics of optimal decision making: A formal analysis of performance in two-alternative forced choice tasks. *Psych. Rev.*, *113*(4), 700-765.

Bogacz, R., & Gurney, K. (2007). The basal ganglia and cortex implement optimal decision making between alternative actions. *Neural Comput.*, *19*(2), 442–477.

Butts, D. A., & Goldman, M. S. (2006). Tuning curves, neuronal variability, and sensory coding. *PLoS Biol*, *4*(4), e92.

Dragalin, V., Tartakovsky, A., & Veeravalli, V. (1999). Multihypothesis sequential probability ratio tests, part I: Asymptotic optimality. *IEEE Trans. Inform. Theory*, *45*, 2448-2461.

Ghose, G., Yang, T., & Maunsell, J. (2002). Physiological correlates of perceptual learning in monkey V1 and V2. *J. Neurophysiol.*, *87*, 1867-1888.

Gold, J., & Shadlen, M. (2001). Neural computations that underlie decisions about sensory stimuli. *Trends in Cognitive Sciences*, *5*, 10-16.

Hertz, J., Krogh, A., & Palmer, R. (1991). *Introduction to the theory of neural computation*. Redwood City, CA: Addison-Wesley Publishing Company Advanced Book Program.

Law, C.-T., & Gold, J. (2008). Neural correlates of perceptual learning in a sensory-motor, but not a sensory cortical area. *Nat. Neurosci.*, *11*, 505-513.

McMillen, T., & Behseta, S. (2010). On the effects of signal acuity in a multi-alternative model of decision making. *Neural Comput.*, *22*(2), 539-580.

McMillen, T., & Holmes, P. (2006). The dynamics of choice among multiple alternatives. *J. Math. Psych.*, *50*(1), 30-57.

Pouget, A., Deneve, S., Ducom, J.-C., & Latham, P. (1999). Narrow vs. wide tuning curves: what's best for a population code? *Neural Comput.*, *11*, 85-90.

Seriés, P., Latham, P., & Pouget, A. (2004). Tuning curve sharpening for orientation selectivity: coding efficiency and the impact of correlations. *Nat. Neurosci.*, *7*(10), 1129-1135.

Shadlen, M., & Newsome, W. (2001). Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey. *J. Neurophysiol.*, *86*, 1916-1936.

Usher, M., & McClelland, J. (2001). On the time course of perceptual choice: The leaky competing accumulator model. *Psych. Rev.*, *108*, 550-592.

Wald, A., & Wolfowitz, J. (1948). Optimal character of the sequential probability ratio test. *Ann. Math. Statist.*, *19*, 326-339.

Zhang, K., & Sejnowski, T. (1999). Neural tuning: to sharpen or broaden? *Neural Comput.*, *11*, 75-84.

Contributions of Prosodic and Distributional Features of Caregivers' Speech in Early Word Learning

Soroush Vosoughi[1], **Brandon C. Roy**[1], **Michael C. Frank**[2] **and Deb Roy**[1]
{soroush, bcroy, dkroy}@media.mit.edu, The Media Laboratory[1]
mcfrank@mit.edu, Department of Brain and Cognitive Sciences[2]
Massachusetts Institute of Technology

Abstract

How do characteristics of caregiver speech contribute to a child's early word learning? We explore the relationship between a single child's vocabulary growth and the distributional and prosodic characteristics of the speech he hears using data collected for the Human Speechome Project, an ecologically valid corpus collected from the home of a family with a young child. We measured F0, intensity, phoneme duration, usage frequency, recurrence, and MLU for caregivers' production of each word that the child learned during the period of recording. When all variables are considered, we obtain a model of word acquisition as a function of caregiver input speech. Coefficient estimates in the model help to illuminate which factors are relevant to learning classes of words. In addition, words that deviate from the model's prediction are of interest as they may suggest important social, contextual and other cues relevant to word learning.

Keywords: language acquisition; word learning; corpus data; prosody

Introduction

How does the linguistic environment contribute to children's early word learning? We address this question by making an in-depth study of a single child's vocabulary growth and the relationship of this growth to prosodic and distributional features of the naturally occurring caregiver speech that the child is exposed to. Studying this relationship has the potential to illuminate not only the role of environmental factors in word learning, but also the child's underlying learning mechanisms.

Children's linguistic environments plays a crucial role in determining what they learn, but the precise relationship between what children hear (their input) and what they learn is still unknown. Much of the debate about the role of the linguistic environment has centered around whether the particular properties of child-directed speech (CDS) are useful for the acquisition of syntax. On the one hand, Snow (1986) emphasized the importance of CDS for conveying communicative intent and its consequent importance to development. However, the work of Newport, Gleitman, and Gleitman (1977) challenged the assumption that CDS is a simplified teaching language that facilitates the acquisition of specific syntactic constructions. More recent work has focused on broader patterns of development, documenting a correlation between grammatical and lexical developmental trajectories (Bates & Goodman, 1999).

Stronger evidence for the contributions of CDS to language development have been found in the realm of lexical acquisition. For example, Huttenlocher, Haight, Bryk, Seltzer, and Lyons (1991) found a positive correlation between the quantity of CDS and a child's vocabulary size and rate of growth. Increased frequency of use of particular words in CDS has also been tied to earlier acquisition of those words by the child (Huttenlocher et al., 1991; Goodman, Dale, & Li, 2008; Roy, Frank, & Roy, 2009). Frequency is not the only factor that affects acquisition, however. The production of a word in isolation is also a consistent predictor of lexical development (Brent & Siskind, 2001). Finally, prosodic factors in caregiver speech also likely play a role in acquisition: Echols and Newport (1992) found that children were much more likely to produce and recognize syllables that were stressed in caregivers' speech.

While previous studies of the relationship between CDS and children's vocabulary acquisition have largely focused on examining a small section of the input to a range of children, here we take a different approach. We make a very detailed study of this relationship in a very large, dense, longitudinal dataset collected in an ecologically valid setting. This dataset was collected as part of the Human Speechome Project (Roy et al., 2006). At present, the Speechome Corpus consists of time aligned orthographic transcripts as well as a complete audio and video record of all data collected. Therefore, our analysis is not limited to factors like frequency (which can be computed from transcripts alone): instead we are able to include additional prosodic variables that can only be computed from aligned audio and transcripts.

Our goal in this current analysis is to predict the child's age of acquisition (AoA) for individual words on the basis of information from CDS. AoA is usually categorized as the age of *receptive* and *productive* acquisition. Receptive acquisition is typically determined by the caregiver via diary studies or checklists, and is consequently relatively difficult to assess with high accuracy for a large sample of words. Age of productive acquisition is more easily measured from transcripts, although there are complications here as well, since early productive word forms often differ from the corresponding adult word form. However, we are able to overcome this limitation to a greater extent than previous studies, because of the density of our data and the accessibility of caregivers for help in the transcription process.

The plan of our paper is as follows. We begin with an overview of the Human Speechome Project. We then review the regression framework we used for the prediction of vocabulary acquisition and describe in detail the predictors we included in this framework. We report both simple correlations

between individual predictors and age of word acquisition as well as the results of a series of regression models. We end by considering the implications of our work for future research in language acquisition.

The Human Speechome Project

The Human Speechome Project (HSP) (Roy et al., 2006) was launched in 2005 to study early language development through analysis of audio and video recordings of the first two to three years of one child's life. The house of one author's (DR) family was outfitted with fourteen microphones and eleven omnidirectional cameras at the time of birth of their first child. Audio was recorded from ceiling mounted boundary layer microphones at 16 bit resolution with a sampling rate of 48 KHz. Due to the unique acoustic properties of boundary layer microphones, most speech throughout the house including very quiet speech was captured with sufficient clarity to enable reliable transcription. Video was also recorded to capture non-linguistic context using high resolution fisheye lens video cameras that provide a bird's-eye view of people, objects, and activity throughout the home.

The Speechome project captures one child's development in tremendous depth. While this aspect of the project limits conclusions about general aspects of language development, the dense sampling strategy affords many advantages over other corpora (eg. (Lieven, Salomo, & Tomasello, 2009)). First, the Speechome corpus is higher in density than other reported corpus, capturing an estimated 70% of the child's wakeful experiences during the recording period. Second, since data were collected without specific theoretical assumptions or hypotheses, they can be reanalyzed in multiple ways from different theoretical perspectives. Finally, since high resolution video was also collected the role of non-linguistic context can also be studied (though in the current study we restrict our analysis to aspects of speech input).

The current study builds on our first analysis of the Speechome data (Roy et al., 2009). In that study, we focused on the child's 9-24 month age range and explored several aspects of word learning, examining variables such as the child's vocabulary growth, increase in mean length of utterance (MLU) as well as properties of caregiver speech such as caregiver MLU over time. Due to the high density of data, with several days per week fully transcribed over the course of this 9–24 month period, a surprising picture emerged of the tuned relationship between the child's development and caregiver speech. Congruent with other reports, we found that words used more frequently in caregiver speech tend to be learned earlier by the child, with a much stronger effect when words are grouped by class (Huttenlocher et al., 1991; Goodman et al., 2008).

Methods

The Speechome Audio Corpus

The dataset collected for the Human Speechome Project comprises more than 120,000 hours of audio and 90,000 hours of video. Most analysis depends on annotated data, however, so an effective annotation methodology is critical to the project's success. We have developed a semi-automated speech transcription system called BlitzScribe that facilitates fast and accurate speech transcription (Roy & Roy, 2009). Automatic speech detection and segmentation algorithms identify speech segments, presenting them to a human transcriber in a simple user interface. This focuses human effort on the speech and leads to a smoother transcription process. We have obtained an approximately five-fold performance gain at comparable accuracy to other tools.

Speaker identification algorithms are then applied to the transcribed audio segments, selecting from one of the four primary speakers (mother, father, nanny, and child) and producing a classification confidence score. Speaker annotation tools allow a human to review low confidence segments and make corrections as necessary. Since identifying CDS currently requires significant human effort, we operationalized the definition to refer to caregiver speech when the child is awake and close enough to hear. We refer to this as "child available speech" (CAS).

Our current study focuses on the child's 9–24 month age range, and the corresponding subset of the corpus contains 4260 hours of 14-track audio, of which and estimated 1150 hours contain speech. Of the 488 days in this time range, recordings were made 444 of the days with a mean of 9.6 hours recorded per day. The current results are based on 72 fully transcribed days containing an average of 23,055 words per day of combined CAS and child speech, totaling 1.66 million words. We estimate that the fully transcribed 9-24 month corpus will contain 12 million words. Our long term goal is to fully annotate all speech in the corpus with transcriptions, speaker identity, and prosodic features.

Three limitations of the speech annotation process required us to filter the 1.66 million words of transcripts and only use a subset of the transcripts for the current analyses. First, roughly 700,000 words belong to utterances marked by human transcribers as containing more than one speaker. In other words, about 40% of pause separated spoken utterances contain abutting or overlapping speech of two or more people, reflecting the realities of "speech in the wild." Since our objective here is to examine interaction of CAS and child speech, and since we cannot currently distinguish the sources of this type of speech, we removed these utterances. Second, to reduce errors due to automatic speaker identification, we sorted utterances based on a confidence metric produced by the speaker identification algorithm and removed approximately the bottom 50% of utterances. Third, about 15% of the remaining utterances were deemed by human transcribers to be of insufficient clarity to transcribe reliably. After removing those utterances, we obtained the 399,141 word corpus used for all analyses in this paper.

Outcome and Predictor Variables

The goal of our study was to use measurements of the prosodic and distributional characteristics of CAS to predict

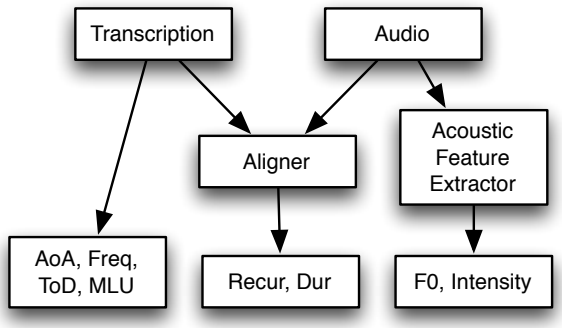


Figure 1: Schematic of the processing pipeline for outcome and predictor variables.

AoA for the child's early vocabulary. We use linear regression to provide a computational framework for this goal. We therefore used age of acquisition as our outcome variable and extracted six predictor variables to quantify aspects of CAS. Figure 1 shows the pipeline used to extract these predictor variables from our speech and transcription files. Below we give our operational definition for age of acquisition and for each of the six predictor variables we used in our analysis. All variables are computed using the sample up to the AoA for a particular word.

Age of Acquisition We defined the AoA for a particular word as the first time in our transcripts that the child produced a word. Using this definition, the first word was acquired at nine months of age with an observed productive vocabulary of 517 words by 24 months (though the actual productive vocabulary might be considerably larger when transcription is completed). In order to ensure reliable estimates for all predictors, we excluded those words from the child's vocabulary for which there were fewer than six caregiver utterances. This resulted in the exclusion of 56 of the child's 517 words, leaving 461 total words included in the current analysis.

Frequency Frequency measures the count of word tokens in CAS up to the time of acquisition of the word divided by the period of time over which the count is made. Thus, this measure captures the average frequency over time of a word being used in CAS.

Recurrence Distinct from frequency, recurrence measures the repetition of a particular word in caregiver speech within a short window of time. The window size parameter was set by searching all possible window sizes from 1 to 600 seconds. For each window size, we performed a univariate correlation analysis to calculate the correlation between recurrence at that window size and AoA. We then selected the window size which produced the largest correlation (51 seconds).

MLU The MLU predictor measures the mean utterance length of caregiver speech containing a particular word. In order to be consistent with the direction of correlation for other variables (a negative correlation with the AoA) we use $1/\text{MLU}$ as the predictor.

Duration The duration predictor is a standardized measure of word duration for each word. We first extracted duration for all vowel tokens in the corpus. We next converted these to normalized units for each vowel separately (via z-score), and then measured the mean standardized vowel duration for the tokens of a particular word type. For example, a high score on this measure for the word "dog" would reflect that the vowel that occurred in tokens of "dog" was often long relative to comparable vowel sounds that appeared in other words. We grouped similar vowels by converting transcripts to phonemes via the CMU pronunciation dictionary.

Fundamental frequency The fundamental frequency predictor is the measure of a word's change in fundamental frequency (F0) relative to the utterance in which it occurred. We first extracted the F0 contour for each utterance in the corpus using the PRAAT system (Boersma & Weenink, 2009). We then calculated the change in F0 as a sum of two terms shown in the equation below. The first term captures the change in F0 for the word relative to the utterance in which it's embedded. $\overline{\text{F0}}_{\text{w}}$ is the mean F0 value of the word, and $\overline{\text{F0}}_{\text{utt}}$ is the mean F0 of the whole utterance. The second term captures the maximum change in F0 within the word. $t_{\max}$ and $t_{\min}$ are the times at which the max and min F0 values occur within the word. α_0 and α_1 are constants set using the same optimization technique described in the recurrence section.

$$\alpha_0 * \left|\overline{\text{F0}}_{\text{w}} - \overline{\text{F0}}_{\text{utt}}\right| + \alpha_1 * \left|\frac{\max(\text{F0}_{\text{w}}) - \min(\text{F0}_{\text{w}})}{t_{\max} - t_{\min}}\right|$$

Intensity Relative word intensity was calculated in the same manner as F0 using the intensity contour in place of the F0 contour. The intensity contour was extracted using the PRAAT system.

Results

Correlation analysis

Correlations between AoA and the six variables we coded in caregiver speech are shown in Figure 2. All correlations were negative and highly significant (all p-values less than .001) though their magnitude varied. Correlations with recurrence and intensity were largest, while correlation with F0 was smallest.

Replicating results in Roy et al. (2009), the correlation with frequency was -.23. This figure is slightly lower than the -.29 reported in the earlier paper. There are two differences in analysis that account for the different result. First, a small subset of words were excluded from this analysis due to data sparsity. Second, frequency data are estimated only up to the time the child first produces the word. This second difference leads to a potentially interesting conclusion. If the distribution of word frequencies is stationary with respect to time, then correlations should go up as more data are included for

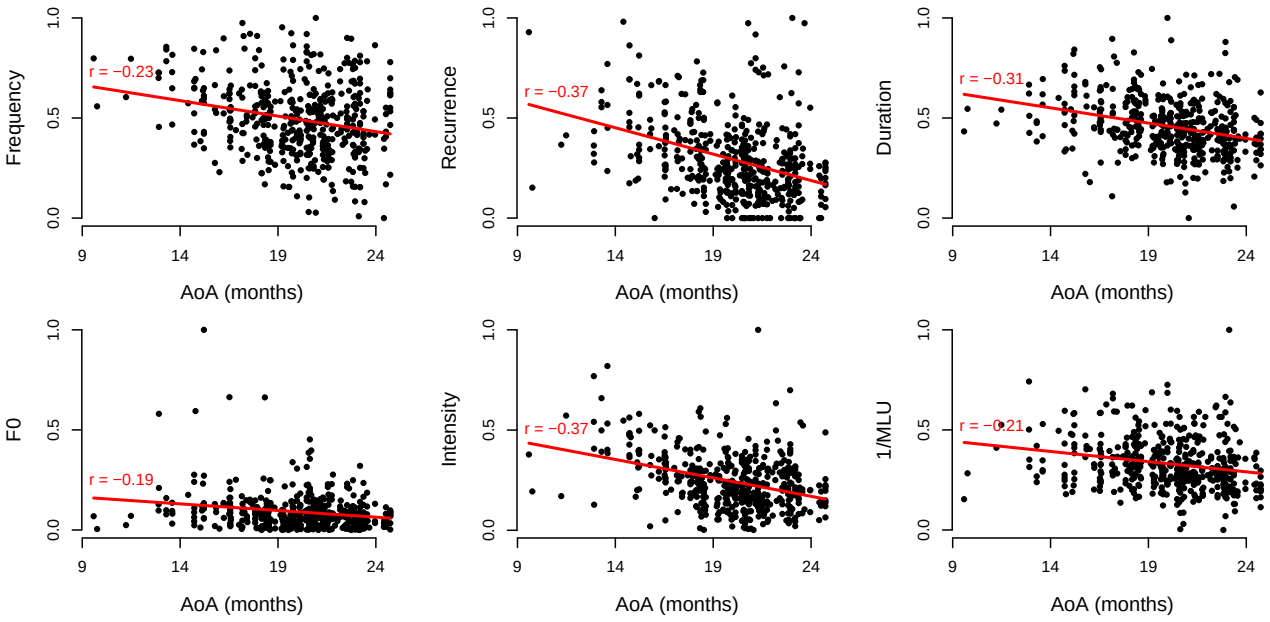


Figure 2: Each subplot shows the univariate correlation between AoA and a particular predictor. Each point is a single word, while lines show best linear fit.

Table 1: Correlation coefficients (Pearson's r) between all predictor variables. Note: $' = p < .1$, $* = p < .05$, and $** = p < .001$.

	Recur	Dur.	F0	Int.	1/MLU
Frequency	.36**	-.05	.19**	.35**	-.22**
Recurrence		.25**	.20**	.22**	.10*
Duration			.12*	.22**	.33**
F0				.10*	-.15*
Intensity					.02

each word. In contrast, if caregivers tune the frequency distribution of words to an estimate of the child's knowledge, correlations should go down as more data are included. Because we observed (slightly) larger correlations with frequency for the earlier dataset, this provides some evidence against caregiver tuning of word frequencies.

Correlations between predictor values are shown in Table 1. The largest correlations were between frequency and recurrence, frequency and intensity, and inverse MLU and duration. The correlation between frequency and recurrence is easily interpreted: the more times a word appears, the more likely it is to recur within a small window. On the other hand, correlations between prosodic variables like frequency and intensity or duration and inverse MLU are less clear. For example, perhaps words are more likely to have longer duration vowels when they are being accented in a shorter sentence.

Regression analysis

We next constructed a regression model which attempted to predict AoA as a function of a linear combination of predictor values. The part of speech (POS) was included as an additional predictor. We created POS tags by first identifying the MacArthur-Bates Communicative Development Inventory category (Fenson, Marchman, Thal, Dale, & Reznick, 2007) for each word that appeared in the CDI and generalizing these labels to words that did not appear in the CDI lists. To avoid sparsity, we next consolidated these categories into five broad POS categories: adjectives, nouns, verbs, closed-class words, and other. The inclusion of POS as a predictor significantly increased model fit ($F(4) = 107.37$, $p < .001$).

Coefficient estimates for each predictor are shown in Figure 3. All predictors were significant at the level of $p < .05$. The full model had $r^2 = .32$, suggesting that it captured a substantial amount of variance in age of acquisition.

The largest coefficients in the model were for intensity and inverse MLU. For example, there was a four-month predicted difference between the words with the lowest inverse MLU ("actual," "rake," "pot," and "office") and the words with the highest inverse MLU ("hi," "silver," "hmm," and "sarah"). Effects of POS were significant and easily interpretable. We used nouns as the base contrast level; thus, coefficients can be interpreted as extra months of predicted time prior to acquiring a word of a non-noun POS. Closed-class words and verbs were predicted to take almost two months longer to acquire on average, while adjectives and other words were predicted to take on average less than a month longer.

Assessing model fit

Residuals from the basic linear model were normally distributed. Figure 4 shows the relation between predicted age of acquisition (via the full predictive model including part of

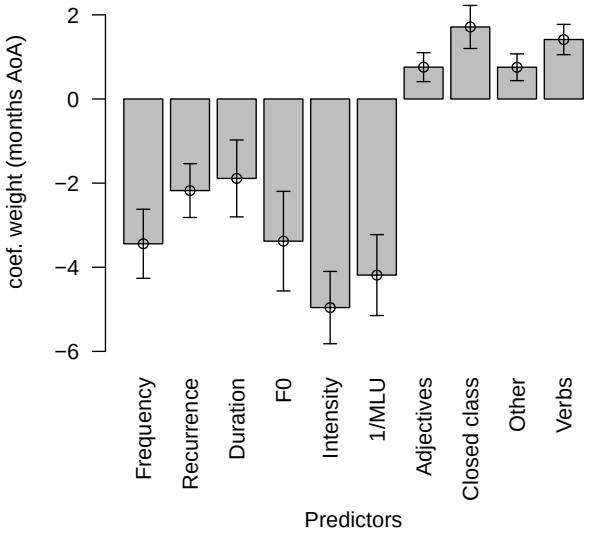


Figure 3: Coefficient estimates for the full linear model including all six predictors (and part of speech as a separate categorical predictor). Nouns are taken as the base level for part of speech and thus no coefficient is fit for them. Error bars show coefficient standard errors. For reasons of scale, intercept is not shown.

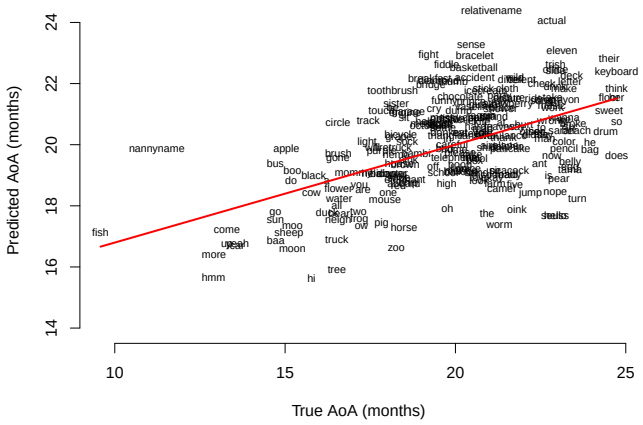


Figure 4: Predicted AoA vs. true AoA. To to avoid overplotting, only half of the 461 words are shown. The red line shows the line of best fit.

speech) and the age of acquisition of words by the child. One useful aspect of plotting the data in this way is that it makes clear which words were outliers in our model (words whose predicted age of acquisition is very different than their actual age of acquisition). Identifying outliers can help us understand other factors involved in age of acquisition.

For example, words like "dad" and "nannyname" (proper names have been replaced for privacy reasons) are learned far earlier than predicted by the model (above the line of best fit), due to their social salience. Simple and concrete nouns like "apple" and "bus" are also learned earlier than predicted, perhaps due to the ease of individuating them from the environment. In contrast, the child's own name is spoken later than predicted (20 months as opposed to 18), presumably not because it is not known but because children say their own name far less than their parents do. Future work will use these errors of prediction as a starting point for understanding contextual factors influencing word learning.

Interactions and more complex models

Our first linear model had two limitations. First, we found that there was significant variation in the effects of the six predictors depending on what POS a word belonged to. Second, we did not include any interaction terms. We followed up in two ways. First, in order to investigate differences in predictor values between word classes we built separate linear models for each POS. Second, we used stepwise regression to investigate interactions in our larger model.

Table 2 shows coefficient estimates for five linear models, each one for a different group of words. None (including the "all" model) include a predictor for POS. Coefficient estimates varied considerably across models, suggesting that different factors are most important for the acquisition of different kinds of words. For example, frequency, intensity, and inverse MLU were most important for nouns, suggesting that hearing a noun often in short sentences where it is prosodically stressed leads to earlier acquisition. In contrast, adjective AoA was best predicted by intensity, duration, and inverse MLU, congruent with reports that children make use of prosodic cues in identifying and learning adjectives (Thorpe & Fernald, 2006). Finally, both verbs and closed-class words were best predicted by recurrence, supporting the idea that the meanings of these words may be difficult to decode from context; hence frequent repetition within a particular context would be likely to help (Gleitman, 1990).

Table 2: Coefficient estimates for linear models including data from adjectives, nouns, closed-class words, verbs, and all data. Note: $' = p < .1$, $* = p < .05$, and $** = p < .001$.

	Adj.	Closed	Nouns	Verbs	All
Icept	27.66**	25.03**	25.00**	25.93**	25.57**
Freq	0.38	6.73	-5.84**	-0.89	-1.53*
Recur	-2.36	-12.02*	-1.53′	-7.47**	-2.85**
Dur	-5.22*	1.81	0.09	-2.74	-2.66*
F0	-7.43	-6.42	-2.28′	0.54	-3.42*
Int.	-8.60*	-12.16	-4.66**	-1.56	-4.78**
1/MLU	-5.70*	-9.37	-3.71*	-5.26	-3.89**

We next constructed a model that included every pairwise interaction between each of the six predictors and between the predictors and POS. We then used stepwise regression to remove predictors that did not increase model fit. Stepwise regression prunes predictors using AIC, a measure which balances increases in likelihood with complexity. This model increased r^2 to .44, and added a large number of interaction

terms. We report only the general outlines of results in this model as they confirm intuitions from other analyses.

While frequency had an overall *positive* coefficient value in this model, all four interactions were negative, indicating that there was considerable shared information between frequency and other predictors. Recurrence and intensity also interacted significantly, suggesting that when words were spoken repeatedly with high intensity (possibly because they were a topic of discourse over a period of time) they were acquired at earlier ages. Finally, both duration and intensity interacted with POS, with significant coefficients for closed-class words. As seen in Table 2, longer closed-class words are acquired slightly later (probably because longer closed-class words are less frequent). In addition, higher intensity closed-class words are acquired considerably earlier, probably because one major challenge in function word acquisition is understanding their prosodic structure (Demuth & McCullough, 2008).

Discussion and Future Work

Our study quantified six variables describing the prosodic and distributional characteristics of words in child-available caregiver speech: frequency, recurrence, mean length of utterance, duration, fundamental frequency, and intensity. We found that each of these variables helped to predict the age at which the child acquired words. There were considerable differences in the predictive power of each variable across different parts of speech, however. For example, frequency and intensity mattered most for nouns, while recurrence in a small window of time seemed to matter more for verbs and closed-class words. These results complement previous smaller-scale, cross-sectional investigations and provide a variety of new directions for potential experimental manipulations.

Our current model only takes into account variables in caregiver speech, omitting the visual and social context of word learning. One of the benefits of the Speechome Corpus is that this information is available through rich video recordings. Computer vision algorithms and new video annotation interfaces are being developed to incorporate this aspect of the corpus into future investigations. In addition, our current investigation has been limited to the child's lexical development; our plan is that future work will extend the current analysis to grammatical development.

Finally, the analysis and findings presented in this paper assume a linear input-output model between child and caregivers: the caregivers produce input to the child, who then learns. In other words, our current model treats the child as the only agent whose behavior can change. Beyond a first approximation, however, this assumption is inconsistent with our own previous findings (Roy et al., 2009). Our ongoing work continues to investigate the mutual influences between caregivers and child and to measure the degree of adaptation in this dynamic social system.

References

Bates, E., & Goodman, J. (1999). On the Emergence of Grammar from the Lexicon. In B. MacWhinney (Ed.), *The emergence of language* (pp. 29–79). Mahwah, NJ: Lawrence Erlbaum Associates.

Boersma, P., & Weenink, D. (2009). *Praat: doing phonetics by computer (version 5.1.01).* http://www.praat.org/.

Brent, M., & Siskind, J. (2001). The role of exposure to isolated words in early vocabulary development. *Cognition*, *81*, 33–44.

Demuth, K., & McCullough, E. (2008). The prosodic (re) organization of children's early english articles. *Journal of Child Language*, *36*.

Echols, C., & Newport, E. (1992). The role of stress and position in determining first words. *Language acquisition*, *2*.

Fenson, L., Marchman, V. A., Thal, D., Dale, P., & Reznick, J. S. (2007). *MacArthur-Bates Communicative Development Inventories: User's Guide and Technical Manual*. Paul H. Brookes Publishing Co.

Gleitman, L. (1990). The structural sources of verb meanings. *Language acquisition*, *1*, 3–55.

Goodman, J., Dale, P., & Li, P. (2008). Does frequency count? Parental input and the acquisition of vocabulary. *Journal of Child Language*, *35*, 515–531.

Huttenlocher, J., Haight, W., Bryk, A., Seltzer, M., & Lyons, T. (1991). Early vocabulary growth: Relation to language input and gender. *Developmental Psychology*, *27*.

Lieven, E., Salomo, D., & Tomasello, M. (2009). Two-year-old children's production of multiword utterances: A usage-based analysis. *Cognitive Linguistics*, *20*.

Newport, E., Gleitman, H., & Gleitman, L. (1977). Mother, I'd rather do it myself: Some effects and non-effects of maternal speech style. In C. E. Snow & C. A. Ferguson (Eds.), *Talking to Children: Language Input and Acquisition* (pp. 109–149). Cambridge University Press.

Roy, B. C., Frank, M. C., & Roy, D. (2009). Exploring word learning in a high-density longitudinal corpus. In *Proceedings of the 31st Annual Cognitive Science Conference.*

Roy, B. C., & Roy, D. (2009). Fast transcription of unstructured audio recordings. In *Proceedings of Interspeech.* Brighton, England.

Roy, D., Patel, R., DeCamp, P., Kubat, R., Fleischman, M., Roy, B., et al. (2006). The Human Speechome Project. In *Proceedings of the 28th Annual Cognitive Science Conference* (pp. 2059–2064). Mahwah, NJ: Lawrence Earlbaum.

Snow, C. E. (1986). Conversations with children. In P. Fletcher & M. Garman (Eds.), *Language acquisition: Studies in first language development* (pp. 69–89). Cambridge, UK: Cambridge University Press.

Thorpe, K., & Fernald, A. (2006). Knowing what a novel word is not: Two-year-olds 'listen through' ambiguous adjectives in fluent speech. *Cognition*, *100*.

Concreteness and Relational Matching in Preschoolers

Jennifer A. Kaminski (kaminski.16@osu.edu)
Center for Cognitive Science, Ohio State University
208A Ohio Stadium East, 1961 Tuttle Park Place, Columbus, OH 43210, USA

Vladimir M. Sloutsky (sloutsky.1@osu.edu)
Center for Cognitive Science, Ohio State University
208D Ohio Stadium East, 1961 Tuttle Park Place, Columbus, OH 43210, USA

Abstract

This study investigated the effect of concreteness on preschool children's ability to recognize simple relations. Participants, age 3.0 to 5.0 years, were asked to make one-shot relational matches from a base to a target display. Two types of questions were posed: Generic in which the base display contained simple geometric shapes and Concrete in which the base display contained colorful familiar objects. Two between-subjects conditions varied the order in which the Concrete and Generic questions were asked. The results reveal relational matching on Concrete questions was significantly higher when preceded by Generic questions than when answered first, suggesting children transferred relational knowledge acquired through the Generic questions to answer the Concrete questions. However, there was no improvement on Generic questions when preceded by Concrete questions. These are novel findings suggesting that young children can better acquire and subsequently transfer relational knowledge from a generic format than from a concrete, perceptually rich format.

Keywords: Cognitive Science; Psychology; Transfer; Relations, Structure Recognition.

Introduction

The ability to recognize common relations across different situations is not always easy, but tends to improve through the course of development. Most researchers agree that some form of a relational shift occurs in development (e.g. Gentner, 1988; Gentner & Ratterman, 1991, see also Goswami, 1991); young children are more likely to attend to object-level similarities between systems or displays and overlook relations. Later in development, people become more likely to attend to relational similarities. For example, when given a simple metaphor such as a plant stem is like a straw, children's interpretation is often based on superficial attributes, such as both are thin and straight. Adults tend to interpret such metaphors through deeper relations; in this case, both can carry water (Gentner, 1988).

One category of theoretical accounts of relational development is that the relational shift is knowledge-driven (Brown, 1989, Brown & Kane, 1988; Gentner, 1988, Gentner & Ratterman, 1991, Vosniadou, 1989). By such accounts, domain-specific knowledge is the primary predictor of ability to attend to relations. In support of this position, there is considerable evidence that while young children may fail to reason analogically (i.e. based on relational structure) in many instances, they can reason analogically in contexts that are familiar to them (see Gentner, Ratterman, Markman, & Kotovsky, 1995 for discussion). For example, Gentner (1977a, 1977b) found that when 4-year-old children were shown a picture of a tree and asked, "If a tree had a knee, where would it be?", they interpreted the relational correspondence and responded as accurately as adults. Similarly, preschool children, aged 3 to 5 years, successfully transferred problem-solving strategies from contexts involving simple, familiar relations such as mimicry and camouflage (Brown & Kane, 1988). Additionally, 4-year-olds applied relational reasoning on tasks involving known relations, such as cutting and melting (Goswami & Brown, 1989). Taken together, there is ample evidence of successful relational reasoning by young children when the relations are known to them.

Yet, even in the context of simple relations and familiar objects, attention to relations can be diverted by interference of surface similarities across the base and target domains. For example, preschool children, age 3 and 4 years, were tested on their ability to make relational matches involving the relation of monotonic increase or decrease of three items (Gentner & Rattermann, 1991; see Gentner, et al., 1995 for summary). In the task, the experimenter and the participant each had sets of three items arranged in monotonically increasing or decreasing order. The child was asked to close his/ her eyes while the experimenter hid stickers under one object in each set. The stickers were always placed under items in the same relational roles across sets. When the child opened his/her eyes, the experimenter showed the child an object with sticker in the experimenter's set and asked the child to find the sticker was in the child's set. This study had a 2 x 2 design: literal similarity or cross-mapping by stimuli type. In the literal similarity condition, the correct item matched the target on both object appearance and relational location. In the cross-mapping condition, the correct item differed in appearance and matched the target only on relational location. Also, in the cross-mapped condition, an incorrect relational choice matched the target object in appearance. Hence, children could make either relational matches or appearance matches. The stimuli type varied the perceptual richness of the objects: either sparse,

such as clay pots or blue boxes, or rich, such as colorful toys or silk flowers.

It was found that children were more likely to choose relational responses in the literal condition than in the cross-mapped condition. Four-year-olds were very accurate on matching literal similarity for both rich and sparse material. However, 3-year-olds had difficulty with the sparse stimuli. In the cross-mapped condition, 3- and 4-year-olds generally matched on object appearance rather than on relational role. Furthermore, performance was much worse for perceptually rich objects than for perceptually sparse objects, suggesting that the richer objects were more likely to divert attention from relations than the more sparse objects.

Adults are also susceptible to interference from cross-mapped elements involved in complex relational tasks (Ross, 1987, 1989). When attempting to transfer mathematical solution strategies from one example problem to another, college students tend to align structure based on similarity of elements, placing similar elements in the same relational roles. This leads to incorrect solutions if the similar elements do not actually hold analogous roles.

The ability to perceive common relational structure underlies not only simple analogies, but also higher-order cognitive processes such as the acquisition and transfer of mathematical knowledge. This is because mathematical concepts are defined, not by surface features, but by their relational structure. Therefore, relational knowledge can potentially be transferred between situations that appear very different on the surface but have the same underlying structure. For example, the same probability principles can be applied to problems involving the number of ways computers can be assigned to offices or the number of ways toppings can be applied to pizza (e.g. Ross, 1987). Therefore, the study of factors that promote the recognition of common relations has importance to both the study of general cognition as well as practical importance for the potential improvement of acquisition of abstract concepts such as mathematical concepts.

One way of facilitating recognition of common relations is through explicit comparison (e.g. Catrambone & Holyoak, 1989; Loewenstein, Thompson, & Gentner, 1999; Gentner, Loewenstein, Thompson, 2003). Learners are more likely to recognize common relational structure between two instances when they explicitly compare them than when they encounter them sequentially. The process of comparison requires alignment that highlights common structure. Comparison appears to promote the formation of a schema which can in turn allow for successful transfer of relational knowledge to novel analogous situations (Gentner, et al., 2003; Gick & Holyoak, 1983).

Another factor that has been shown to affect the detection of common relations is the concreteness of the learning material. Concreteness of a given instantiation of an abstract concept can be construed as the amount of information communicated to an individual by that particular instantiation. By this interpretation concreteness can be in the form of perceptual richness or contextual richness including prior knowledge. In contrast to concrete instantiations, generic instantiations communicate little extraneous information. Concrete, perceptually rich objects and contexts can hinder performance on relational tasks in comparison to more abstract generic instantiations of the same concepts. This pattern is suggested by the performance of preschoolers on the relational matching task involving monotonic increase and decrease mentioned above (Gentner & Rattermann, 1991; see Gentner, et al., 1995). Children were more likely to make relational responses in the face of conflicting object matches when the task was conducted with perceptually sparse material than with perceptually rich material.

Other evidence for the hindering effects of concreteness is found from studies investigating the development of children's symbol use (DeLoache, 1995a, 1995b, 1997, 2000). Successful symbol use requires the detection of common relations. For example, to effectively use a map as a symbol for a real location, one must recognize the common relations between entities on the map and their real-world analogs. Young children have difficulty using concrete, perceptually rich objects as symbols. In a series of studies, 2½ to 3-year-old children were shown a 3-dimensional scaled model of a real room and told that a stuffed animal was hidden in the actual room. The experimenter then placed a miniature toy in the model telling the children that the location of the miniature toy in the model corresponded to the location of the actual toy in the real room. The children were then asked to retrieve the real toy. Only 16% of the children were able to make errorless retrieval of the actual toy. The children were then asked to retrieve the miniature toy. The accuracy of the miniature toy retrieval was 88% implying that the poor performance on the retrieval of the actual toy was not due to inability to remember the location, but an inability to realize that the model symbolically represented the actual room. In subsequent studies, the salience of the model was decreased by putting it behind a glass window. Under this condition, more than half of the participants accurately retrieved the toy. Similarly, when children were shown the location in a picture and not a 3-dimensional model, 80% of participants ably retrieved the real toy.

By 3 years of age, most children are successful in such a task. However, when the 3- year-old study participants were encouraged to play with the model first only 44% of them successfully retrieved the toy, compared to 78% of 3-year-olds who retrieved the object with no opportunity to play. The physical interaction with the model made it more difficult for the children to treat it as a symbol. In sum, decreasing the concreteness of the objects increased the ease of their symbolic use.

The hindering effects of concreteness demonstrated in the mentioned studies were found in the context of similarities between the base and target situations. In the Gentner and Ratterman study, object similarity was directly pitted against relational similarity. In the DeLoache et al. studies, there was some alignable similarity across base and target as

the model was intended to represent the real room. Little is known about the effects of concreteness on children's relational reasoning in absence of either relationally alignable or cross-mapped similarities between base and target.

There is some recent evidence for an advantage of generic material over more concrete material for children's relational reasoning in the absence of overt interdomain similarities. Kindergarteners were more likely to acquire the concept of proportion and correctly match displays of different objects based on proportion when training instantiated proportions using generic shapes than when proportions were instantiated with colorful, concrete objects (Kaminski & Sloutsky, 2009).

There is also evidence that concreteness can hinder the ability of adults to detect common relations (Kaminski, Sloutsky, & Heckler, 2006). Undergraduate students were less able, or often unable, to transfer complex relational knowledge to novel analogous when knowledge was acquired in a concrete format than when knowledge was acquired in a more generic format (Goldstone & Sakamoto, 2003; Goldstone & Son, 2005; Kaminski, Sloutsky, & Heckler, 2008; Sloutsky, Kaminski, & Heckler, 2005).

Taken together, prior research shows that adults are better able to recognize learned relations in novel contexts when they have initially acquired those relations through a more abstract, generic instantiation than through a more concrete, contextualized one. Generic instantiations also have advantages over concrete instantiations for children's ability to acquire novel relations such as proportion. It is unclear whether this advantage will hold for young children's ability to recognize simpler relations such as monotonic increase in the absence of surface similarities. It is possible that without competition of element similarity, children's attention can be focused on the underlying relation. At the same time, relations are less observable than elements and perhaps added perceptual richness of the elements will itself detract from relations.

The goal of the present study was to examine the effect of concreteness of elements on young children's ability to detect common relations. We considered the relations of monotonic increase, monotonic decrease, and symmetry involving three elements. These relations should be easier for children to recognize than proportion (Kaminski & Sloutsky, 2009) because they are built on the simple and familiar relation of "bigger than". Like previous research, we asked 3- and 4-year-old children to make one-shot relational matches across displays. This task prompts participants to compare two displays instantiating the same relation, therefore it allowed us to see whether or not generic instantiations can provide an advantage for recognition of relations beyond the comparison process alone.

Experiment

Method

Participants Participants were 100 preschool children from middle-class, suburban preschools and child care centers in the Columbus, Ohio area (51 girls and 49 boys). Participants' ages ranged from 3.0 to 5.0 years (M = 3.72 years, SD = .47 years).

Materials and Design Participants were shown two displays presented side by side involving a common relation. The task was to choose an item in the right display that was in the same relational role as an indicated item in the left display. Each display involved three objects. The relations considered in this experiment were monotonic increase, monotonic decrease, and symmetry. There were a total of 18 test questions (six increase, six decrease, six symmetry); half were Generic questions and half were Concrete. Generic questions presented simple colored, geometric shapes (circles, triangles, rectangle, or non-rectangluar parallelograms) in the base display. Concrete questions presented colorful perceptually rich objects (dogs, bugs, little girls, shoes, piggy banks, frogs, cats, jack-o-lanterns, and slices of cake) in the base display. The target display for all questions involved colorful perceptually rich objects (ducks, cats, fish, crayons, birds, flowers, ice cream, rocking horses, and ginger bread houses). Each of the target objects were used twice, once for a generic question and once for an analogous concrete question (see Figure 1). The color of the shapes for a generic question was the same as the predominant color of the perceptually rich objects in the analogous concrete question.

Participants were randomly assigned to one of two conditions (Generic-then-Concrete or Concrete-then-Generic). In the Generic-then-Concrete condition, participants were presented with the Generic questions first and then presented with the Concrete questions. The Concrete-then-Generic condition presented the Concrete questions first followed by the Generic questions. Prior to the test questions, participants were shown one example which illustrated the relation of bigger than. The base display of this example showed a bigger boy and a smaller boy in the Concrete-then-Generic condition and a bigger triangle and a smaller triangle in the Generic-then-Concrete condition. For both conditions, the target display showed a bigger teddy bear and a smaller teddy bear.

For the example and all test questions, the elements were identical except in size within each display. For example, in the questions shown in Figure 1, the triangles, dogs, and fish are identical except in size. There were no variations between elements in any other surface features.

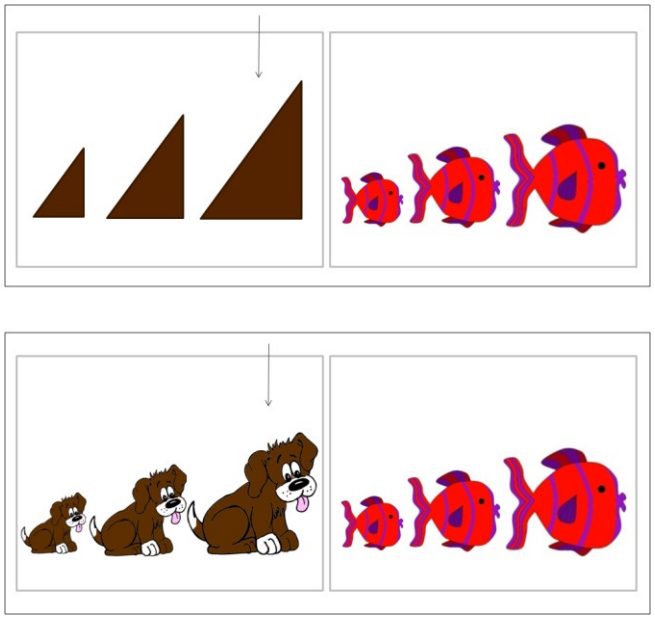


Figure 1: Example of a Generic question (upper) and its analogous Concrete question (lower).

Procedure Participants were asked to play a matching game with the experimenter. All questions were presented on the computer.

The experimenter told the child that he/she would see two pictures and showed the child the example of "bigger than". The experimenter explained that one picture had a certain pattern in it and the same patter was in the other picture, but it looked different. The following is the script:

> "See, in the top picture, there are a bigger boy and a smaller boy. This is the bigger boy, and this is the smaller boy (*the experimenter pointed to each as described)*. Now in the bottom picture, there is a bigger bear and a smaller bear *(the experimenter pointed to each as described)*. See, the same pattern happens in both, but it looks different. Now, in this game, first you have to figure out what the pattern is that happens in both pictures. Okay? Then I am going to point to one thing in one picture, and your job is to tell me what is in the same part of the pattern in the other picture."
> "So here, we have a bigger boy and a smaller boy, and a bigger bear and a smaller bear. Now I am going to point to the smaller boy *(the experimenter pointed)*. Which one is like the smaller boy in the bottom picture according to the pattern? Which one is in the *same part* of the pattern in the bottom picture?"

The Generic condition presented an analogous script which replaced the word "boy" with the word "triangle". The experimenter gave corrective feedback to this example.

Test questions first presented the base display alone on the left side of the computer screen and participants were told to "look for the pattern between things in this picture". The next slide showed the original base display and a new target display on the right side of the screen. In addition an arrow appeared over one of the objects in the base (left) display. The experimenter asked the child, "According to the pattern, what in this picture (*the experimenter gestured to the right picture*) is like this?" *(the experimenter pointed to the object with the arrow)*. The experimenter recorded the child's response on a paper. Then feedback which explicitly stated the relation was given after both correct and incorrect responses. For example, the feedback to the Concrete question that appears in Figure 1 was: "Right *or* No, actually… because in this picture (*the experimenter pointed to the left picture*), these dogs are getting bigger and bigger (*the experimenter gestured*) and I pointed to the biggest one. And in this picture (*the experimenter pointed to the right picture*), these fish are also getting bigger and bigger and this is the biggest one. So, you should point to this one (*the experimenter pointed*).

Results and Discussion

In both the Generic-then-Concrete and Concrete-then-Generic conditions, children were successful at relational matching on both the Generic and Concrete questions. Mean test scores are presented in Figure 2. Scores were above a chance score of 33% (3 out of 9 correct), one-sample t-tests, ts > 8.1, ps < 0.001. However, there was a significant difference in performance as a function of the order in which participants received the Generic and Concrete questions. Test scores were submitted to a two-way analysis of variance with order of the test question type as a between-subjects factor, age as a covariate, and test question type as a repeated measure. The analysis indicated a significant order x question type interaction, $F(1,91) = 11.09$, $p < .001$, $\eta_p^2 = .11$. There were no differences in scores on the Generic questions between the Generic-then-Concrete condition and the Concrete-then-Generic condition, independent samples t-test, $t(92) = .079$, $p > .93$. At the same time, there were differences in scores on the Concrete questions, participants in the Generic-then-Concrete condition scored significantly higher than participants in the Concrete-then-Generic condition, independent samples t-test, $t(92) = 3.16$, $p < .003$. These findings suggest that children who first answered the Generic questions acquired knowledge of the relevant relations that they were able to transfer to the Concrete questions. The reverse was not the case, answering the Concrete questions first did not improve scores on the Generic test. Therefore, experience answering the Generic questions offered an advantage for subsequent transfer that answering the Concrete questions did not.

Additionally, there were improvements with age in test scores on both question types, ANCOVA $F(1,91) = 21.66$, $p < .001$, $\eta_p^2 = .19$. Figures 3 and 4 present accuracy for the Concrete and Generic questions respectively split across the participant age range. Figure 3 illustrates that the differences in accuracy on Concrete questions is present across the age range. Therefore, while development leads to better

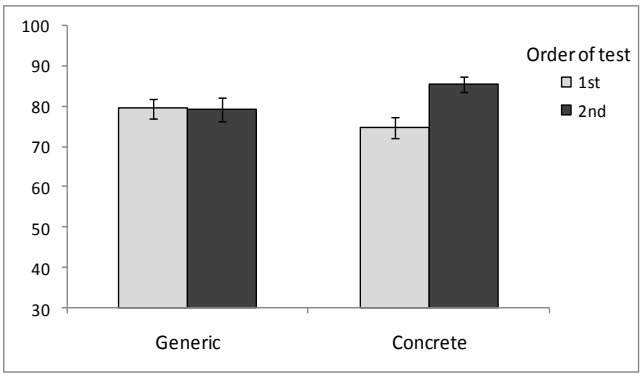


Figure 2: Mean test scores (% correct) by order of test. Error bars represent standard error of mean. Chance score is 33%.

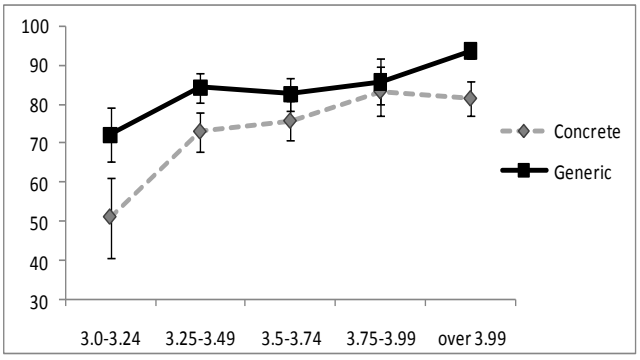


Figure 3: Mean test scores (% correct) on Concrete questions by age of participant in years. Error bars represent standard error of mean.

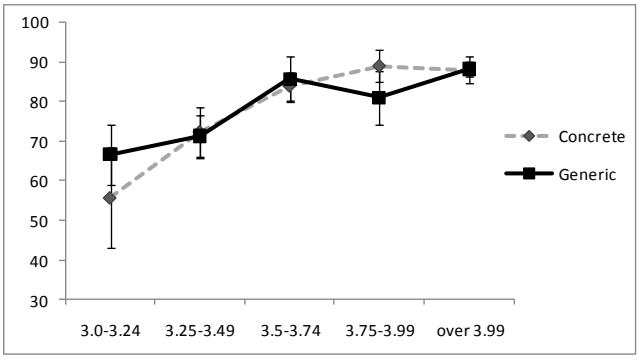


Figure 4: Mean test scores (% correct) on Generic questions by age of participant in year. Error bars represent standard error of mean.

recognition of relations, there is a consistent transfer advantage when first answering the Generic questions.

General Discussion

Previous research has demonstrated the difficulty young children have attending to common relations across displays particularly in the face of competing surface similarities (e.g. Gentner & Ratterman, 1991; Gentner et al, 1995, Richland, Morrison, & Holyoak, 2006). Explicit comparisons, as well as the use of relational language, have been shown to increase relational reasoning (e.g. Gelman, Raman, & Gentner, 2009). There is also evidence that learning a generic instantiation of an abstract concept can facilitate subsequent relational transfer for adults (Goldstone & Sakamoto, 2003; Goldstone & Son, 2005; Kaminski, Sloutsky, & Heckler, 2008; Sloutsky, Kaminski, & Heckler, 2005). However, little research has considered what types of learning instantiations might help promote young children's relational reasoning in the absence of competing surface similarity.

The present study considered preschool children's ability to recognize the relations of symmetry and monotonic increase and decrease. Preschool children were asked to make one-shot mappings across displays of three items. This task encourages participants to make comparisons between instantiations of the same relations. Participants were given generic questions in which relations were mapped from displays of generic shapes to displays of colorful, concrete items. They also answered concrete question in which the mapping was between two displays of different colorful, concrete objects. The results found that when participants first answered the generic questions they scored markedly higher on the subsequent concrete questions than when the concrete questions were answered first. This suggests that by answering the generic questions, participants acquired solid knowledge of the relations which they ably transferred to the concrete questions. At the same time, there were no differences in scores on the generic questions as a function of when they were answered. Therefore, answering the concrete questions provided no benefit for subsequent transfer of relations.

In order to successfully recognize common relations in two different instantiations, the learner must focus attention on the relations between the objects and not directly on the objects themselves. Perceptually rich, concrete objects communicate much more information than perceptually sparse objects. Consider how much more information is communicated by the dogs versus the triangles in the base displays of Figure 1. This abundance of extraneous information may divert the learner's attention from relevant relations making it difficult to recognize these relations. In addition, the present findings suggest that when acquiring relations in the presence of extraneous concrete information, learners may form a weaker representation of the relational knowledge that can hinder future transfer.

Simple generic objects likely have less potential to capture attention, allowing more attentional resources to be focused on relevant relations. Therefore, instantiating relations with generic elements may provide an advantage for later transfer even for very young children.

While it is well accepted that the process of comparison can facilitate abstraction of relations and transfer, these findings suggest that comparisons between some types of instantiations may be more beneficial than comparison between other types of instantiations. Furthermore, this

advantage may not be detectable in immediate performance, but in later tasks involving the same relations.

Acknowledgments

This research is supported by a grant from the Institute of Educational Sciences of the U.S. Department of Education (#R305B070407.) to V. M. Sloutsky and J. A. Kaminski.

References

Brown, A. L. (1989). Analogical learning and transfer: What develops? In S. Vosniadou & A. Ortony (Eds.), *Similarity and Analogical Reasoning* (pp. 369-412). Cambridge, England: Cambridge University Press.

Brown, A. L., & Kane, M. J. (1988). Preschool children can learn to transfer: Learning to learn and learning by example. *Cognitive Psychology*, 20, 493-523.

Catrambone, R., & Holyoak, K. J. (1989). Overcoming contextual limitations on problem-solving transfer. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*, 1147–1156.

DeLoache, J. S. (1995a). Symbolic functioning in very young children: Understanding of pictures and models. *Child Development, 62,* 736-752.

DeLoache, J. S. (1995b). Early understanding and use of symbols: The model model. *Current Directions in Psychological Science, 4,* 109-113.

DeLoache, J. S. (1997). Manipulatives as symbols: A new perspective on the use of concrete objects to teach mathematics. *Journal of Applied Developmental Psychology, 18*, 37-54.

DeLoache, J. S. (2000). Dual representation and young children's use of scale models. *Child Development, 71*, \329-338.

Gelman, S. A., Raman, L., & Gentner, D. (2009). Effects of language and similarity on comparison processing. *Language Learning and Development, 5*, 147-171.

Gentner, D. (1977a). Children's performance on a spatial analogies task. *Child Development*, 48, 1034-1039.

Gentner, D. (1977b). If a tree had a knee, where would it be? Children's performance on simple spatial metaphors. *Papers and Reports on Child Language Development*, 13, 157-164.

Gentner, D. (1988). Metaphor as structure mapping: The relational shift. *Child Development, 59*, 47-59.

Gentner, D., Loewenstein, J., & Thompson, L. (2003). Learning and transfer: A general role for analogical encoding. *Journal of Educational Psychology 95* (2) 393-408.

Gentner, D., & Rattermann, M. J. (1991). Language and the career of similarity. In S. A. Gelman & J. P. Byrnes (Eds.), Perspectives on language and thought: Interrelations in development (pp. 225-277). London: Cambridge University Press.

Gentner, D., Ratterman, M. J., Markman, A., & Kotovsky, L. (1995). Two forces in the development of relational similarity. In T. J. Simon & G. S. Halford (Eds.), *Developing cognitive competence* (pp. 263-314). Hillsdale, NJ: Erlbaum.

Goldstone, R. L. & Sakamoto, Y. (2003). The transfer of abstract principles governing complex adaptive systems. *Cognitive Psychology, 46(4)*, 414-466.

Goldstone, R. L., & Son, J. Y. (2005). The transfer of scientific principles using concrete and idealized simulations. *The Journal of the Learning Sciences, 14*, 69-110.

Goswami, U. (1991). Analogical reasoning: What develops? A review of research and theory. *Child Development, 62*, 1-22.

Goswami, U., & Brown, A. L. (1989). Melting chocolate and melting snowmen: Analogical reasoning and causal relations. Cognition, 35, 69-95.

Kaminski, J. A., Sloutsky, V. M., & Heckler, A. F. (2006). Effects of concreteness on representation: An explanation for differential transfer. In R. Sun & N. Miyake (Eds.), *Proceedings of the XXVIII Annual Conference of the Cognitive Science Society.*

Kaminski, J. A., Sloutsky, V. M., & Heckler, A. F. (2008). The advantage of abstract examples in learning math, *Science, 320*, 454-455.

Kaminski, J. A., & Sloutsky, V. M, Heckler, A. F. (in press). The devil is in the superficial details: Why generic instantiations promote portable mathematical knowledge. *Child Development Perspectives.*

Kaminski, J. A., & Sloutsky, V. M. (2009). The effect of concreteness on children's ability to detect common proportion. In N. Taatgen & H. van Rijn (Eds.), *Proceedings of the XXXI Annual Conference of the Cognitive Science Society.*

Loewenstein, J., Thompson, L., & Gentner, D. (1999). Analogical encoding facilitates knowledge transfer in negotiation. *Psychonomic Bulletin & Review, 6*, 586-597.

Richland, L. E., Morrison, R. G., & Holyoak, K. J. (2006). Children's development of analogical reasoning: Insights from scene analogy problems. *Journal of Experimental Child Psychology, 94*, 249-273.

Ross, B. H. (1987). This is like that: The use of earlier problems and the separation of similarity effects. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *13*, 629-639.

Ross, B. H. (1989). Distinguishing types of superficial similarities: Different effects on the access and use of earlier problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*, 456-468.

Sloutsky, V. M., Kaminski, J. A., & Heckler, A. F. (2005). The advantage of simple symbols for learning and transfer. *Psychonomic Bulletin & Review*, *12*, 508-513.

Vosniadou, S. (1989). Analogical reasoning as a mechanism in knowledge acquisition: A developmental perspective. In S. Vosniadou & A. Ortony (Eds.), Similarity and analogical reasoning (pp. 413-437). New York: Cambridge University Press.

How does the presence of a label affect attention to other features?

Amy Perfors (amy.perfors@adelaide.edu.au)
School of Psychology, University of Adelaide

Daniel J. Navarro (daniel.navarro@adelaide.edu.au)
School of Psychology, University of Adelaide

Abstract

Are labels cues to category membership or simply highly salient features? This question is difficult to answer definitively because of the challenge in identifying empirical predictions that would be distinct in each case: either way, one would expect labels to be highly interesting, easy to process, and preferentially used as the basis of generalization. Here we suggest that one difference should be in how the label directs (or fails to direct) attention to the other, less-salient features of the object. We perform a categorization experiment with complex objects containing many low-salience features, and find that labels affect attention to the other features in the same way that highly salient features such as color or sounds do (and unlike an explicit cue to category membership). This results in a diminished ability to use the less-salient features of the categories to generalize appropriately.
Keywords: labels; features; categorization; generalization

Introduction

Shakespeare once famously asked "What's in a name?" Over the past few decades, psychologists have studied the scientific version of this question: what is the role of labels in category learning? How do labels affect categorization: the categories people form, the inductions they license, and the generalizations they make? What assumptions about labels do people bring to the tasks of word and category learning? These questions have been of special interest in the study of language acquisition, because understanding the assumptions children bring to the problem of word learning is key to understanding their linguistic development.

Much evidence demonstrates that children assume that labels are special in some way. Infants familiarized to items from a novel category will treat it like a category if they hear a label attached to the items, but not if they hear a non-linguistic sound (Balaban & Waxman, 1997; Fulkerson & Waxman, 2007) or hear nothing at all (Waxman & Markow, 1995; Waxman & Braun, 2005). Moreover, infants use labels but not sounds for individuation (Xu, 2002) and as a basis for inductive inference (Gelman & Markman, 1987; Davidson & Gelman, 1990; Graham, Kilbreath, & Welder, 2004).

Why do labels have this special status? Although infants appear uniquely interested in speech (Vouloumanos & Werker, 2004), they are equally capable of learning mappings involving non-linguistic sounds as words (Roberts & Jacob, 1991; Woodward & Hoyne, 1999). This suggests that the "specialness" of labels is not solely due to increased attention or interest in speech in general (although it may be related to the fact that the input is auditory; see Robinson & Sloutsky, 2004, 2006). Furthermore, when labels are inconsistent with apparent category structure or similarity, infants and children are much more reluctant to form categories based on them (Davidson & Gelman, 1990; Waxman & Braun, 2005; Plunkett, Hu, & Cohen, 2008); this may suggest that words are important because they tend to pick out useful categories. Perhaps children make the assumption that labels map cleanly onto category structure because labels are referential: younger infants will categorize using symbolic forms other than words (e.g., gestures or pictograms) if they are used in a referential context (Namy, 2001; Campbell & Namy, 2003), and older infants will use labels to pick out global categories only if they are presented in person by an experimenter rather than a recording(Fulkerson & Haaf, 2003). Another possibility is that infants assume that words identify useful categories because they statistically tend to do so (Samuelson & Smith, 1999), and infants' statistical learning mechanisms are well-attuned for picking this sort of pattern up (Smith, Jones, & Landau, 1996).

As this discussion illustrates, there is some disagreement about how and why labels are special. It may be that labels are special because they are linguistic – referential and used for communication – and infants realize this (Balaban & Waxman, 1997; Namy, 2001; Xu, 2002; Fulkerson & Waxman, 2007). Alternatively, it may be that infants have learned to pay special attention to words because they are statistically likely to be useful indicators of category structure (Smith et al., 1996). The special status of labels may also be perceptual in origin: perhaps labels play a unique role in category formation because of their auditory properties (Robinson & Sloutsky, 2004, 2006, 2007).

This debate parallels a similar, but not identical, discussion in the adult literature – one focused on whether labels act as category indicators or just a highly salient feature. On one hand, labels certainly do appear to hold a privileged psychological status in some ways. When objects share a label, this is sufficient to increase their similarity (e.g., Goldstone, Lippa, & Shiffrin, 2001), and people often make inductive inferences based on an object's label rather than its features or overall similarity(e.g., Yamauchi & Markman, 2000; Johansen & Kruschke, 2005). On the other hand, formal models of categorization have often been remarkably successful at matching human performance simply by treating labels as another – possibly highly salient – feature of the stimulus (e.g., Anderson, 1991; Gliozzi, Mayor, Hu, & Plunkett, 2009).

One of the difficulties inherent in resolving this debate is that it is hard to identify characteristics that an indicator of category membership would have but a very salient feature would not. For instance, one might suggest that the differ-

ence might be that if something is an indicator of category membership, it should be used to pick out categories even when it seems to be inconsistent with the observed similarity or category structure. There is evidence that this is the case for labels when they are mildly inconsistent (Yamauchi & Markman, 2000), but not when they grow too inconsistent (Davidson & Gelman, 1990; Waxman & Braun, 2005). But does this mean that words are strong markers of category membership or salient object features? The problem is that the results make sense under either theory. On one hand, if labels are especially salient features then one would expect them to be followed even if other (less salient) features seemed to pick out a different category structure; on the other hand, if labels are treated as markers to category membership without being features themselves, they could still be such strong markers that they are nearly impossible to override.

More generally, both highly salient features and cues to category membership should share many other characteristics: easy to represent, quick to process, and preferentially used as a basis for generalization. What, then, is the difference between them? To address this question, it helps to consider the two possibilities individually.

- *What are the cognitive effects of a salient feature?* Much work suggests that salient features share two important characteristics. One is that they tend to be the features that people examine first when making choices (e.g., Tversky, 1972; Gigerenzer & Goldstein, 1996). The other is that if the feature is predictive and useful, it will become even more salient over the course of learning (Kruschke, 1992, 2003). As a consequence, if a feature is initially quite salient and later turns out to be predictive of category membership, even more attention will be devoted to it, and the attention devoted to the *other* features will decrease commensurately, particularly if they themselves are not salient or are difficult to process.

- *What are the cognitive effects of a cue to category membership?* Less research bears directly on this question, but we can begin by considering the case of something that is unequivocally a cue to category membership and also unequivocally *not* a feature: explicit instruction. Imagine telling someone that objects from category *A* were sorted into one box and objects from category *B* were sorted into another. Those boxes (along with the instructions) would be cues to category membership, but not features of the objects. How would this affect processing of the objects? Not surprisingly, providing this kind of structure in the visual presentation of stimuli tends to improve learning by calling attention to the relevant features and minimizing the processing load imposed on the learner (e.g., Bruner, Goodnow, & Austin, 1956, ch. 4). As a result the effect on attention is expected to be in the opposite direction: those object features that are less salient, will be processed much more than they otherwise might.

Do the cognitive effects of labelling look more like those of features, or of cues to category membership? We address this question by presenting participants with a simple categorization task involving objects with numerous non-salient and difficult-to-process features paired with a *category indicator* of some sort. In one condition, the category indicator is intended to be a strong cue to category membership: the objects are explicitly categorized by being sorted into boxes. In two other conditions, the category indicator is a highly salient non-linguistic feature (a color or a non-linguistic sound). In two final conditions, the category indicator is a label (either written or oral). After sorting the objects, participants are asked how they would classify new objects for which the category indicator is unknown. Importantly, because the category indicator is unknown and the other features so complex and low-salience, performance on the generalization task reflects how much people have attended to those features. If the category indicator acts like a cue to category membership by calling attention *to* the less-salient features, generalization on the basis of them should be improved when given the indicator; however, if the category indicator acts more like a salient feature by directing attention *away from* the less-salient features, generalization should be poor. Our results suggest that labels behave much like other extremely salient features in the way that they focus attention away from other features of an object.

Method

92 adult participants were recruited from the University of Adelaide and surrounding community and were paid $5 for their participation in the half-hour experiment. Two participants were excluded due to failure to understand the task, leaving 18 people in each of five possible conditions. Each participant saw a series of trials in which they were asked to sort novel objects into categories. They were then asked two generalization questions about how they would categorize additional objects without category indicators. Each of the objects has eight features, four of which vary coherently according to the category structure, and four of which are random. In half of the trials (the NO INDICATOR trials), participants were asked to sort these objects into clusters. In the other half (the INDICATOR trials) the task was the same except that the objects were also each associated with a category indicator, the nature of which varied by condition.

Items. Each item consists of a square with four symbolic characters (one in each quadrant) surrounded by circles (also containing symbolic characters) at each corner; we refer to each location as one of the eight low-salience *features* of the objects, and the particular character in that location as its *feature value*. Each feature can take on a value corresponding to one of ten specific characters, and there is no overlap of possible character sets (feature values) from feature to feature. For each participant and trial, features were generated independently, according to the following pattern: four features are randomly selected to be *dispersed*, meaning that they do not respect category structure because they are uniformly se-

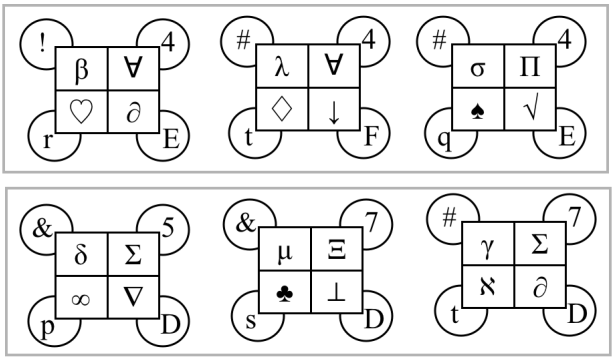


Figure 1: Example INDICATOR trial in the BOXES condition (for visual clarity, we show 6 objects rather than 8 or 16). In this condition, objects are presented already sorted into boxes corresponding to two categories. Here the four coherent features are the two upper circles, the upper right square, and the lower right circle. These features have a 75% coherence level: each of the four coherent features has 25% probability of being "flipped" from the value appropriate to its category.

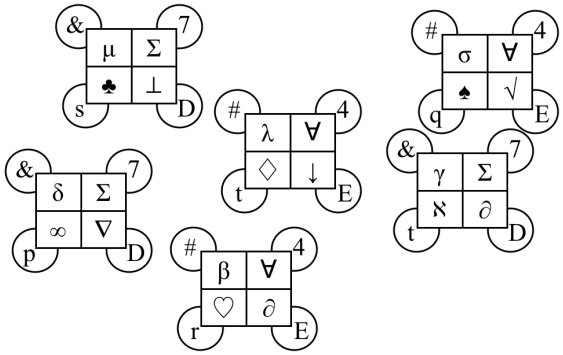


Figure 2: Example NO INDICATOR trial, which participants in all conditions were exposed to. In this sort of trial, people are told to sort the objects in whatever way appears sensible, and are not told in advance how many categories there are or what features are important or useful. In this trial the coherence level is 100%: each of the four coherent features (which are the same as in Figure 1) follow the category structure precisely.

lected from the possible set of values for that feature. The other four are *coherent*, meaning that they correspond to the underlying category structure: feature *A* corresponds to category structure if all members of category *X* share a the same feature value for *A* (say, all of them have a δ in the upper left corner of the square). We systematically varied the coherence[1] level of the four *coherent* features so that half of the trials involved items with a coherence level of 75%, and half involved a coherence of 100%. This mimics real-world categories, which have a probabilistic, graded structure.[2] It is possible to identify the correct categories on the basis of the coherent features, as people have succeeded in doing in other studies (Perfors & Tenenbaum, 2009). However, because these features are numerous, of low salience, and representationally complex, it can be difficult.

Sample objects as they appeared in the experiment are shown in Figures 1 and 2.

Trial structure. Each participant saw eight NO INDICATOR and eight INDICATOR trials. In order to ensure that participants were not relying on external knowledge about how many categories the correct sorting contained, trials varied in the number of items (8 or 16) and the number of categories (2 or 4). Since items varied also in coherence, this resulted in the following factorial design: 2 (INDICATOR or NO INDICATOR) x 2 (coherence level of 75% or 100%) x 2 (containing 8 or 16 items total) x 2 (categories made of 2 or 4 items). This resulted in 16 trials per participant. Due to a coding error, trials with 8 items and 2 categories were not properly counterbalanced according to category indicator, so all analyses excluded these trials and therefore consisted of 12 trials per participant. Figure 1 shows the sort of situation a participant might see on an INDICATOR trial in the BOXES condition, while Figure 2 shows a typical NO INDICATOR trial.

Conditions. The five conditions are defined by the nature of the category indicator involved in the INDICATOR trials. In the BOXES condition, participants saw the objects already pre-sorted into boxes; this is intended as an explicit cue to category membership, and was described to participants as such. In two of the other conditions, the objects in the INDICATOR trials were associated with a label. In the WRITTEN LABEL condition, participants were told that the label would be written above the object. To evaluate whether it mattered if the label was presented visually or orally, in the ORAL LABEL condition, the label was presented out loud (over headphones) whenever the participant clicked on the object. Since participants had to click on all objects in order to sort them, they ended up hearing the labels for every object at some point. The label conditions were compared to two conditions in which the category indicator was simply a highly salient feature. In the COLOR condition, objects were colored (unlike the objects in the NO INDICATOR trials and other conditions, which were always white). And in the SOUND condition, objects were associated with non-linguistic sounds (distinct buzzes, beeps, and tone sequences without semantic associations). As in the ORAL LABEL condition, these sounds were heard through headphones whenever the participant clicked on the object.

Procedure. Each trial consisted of two phases. The first was the "sorting" phase, in which participants were presented with all of the objects in the trial randomly scattered on the computer screen and asked to sort them in categories. (The exception is the INDICATOR trials in the BOXES condition, in which the objects appeared already sorted with square "boxes" drawn around each of the categories, as depicted in Figure 1). During the sorting phase, participants were allowed unlimited time in which to move the objects around on the screen by clicking and dragging them into clusters. They then drew boxes around the objects to indicate cate-

[1] A coherence of c means that a feature value has a $(100-c)\%$ chance of being randomly generated rather than following category structure.

[2] There were no interaction effects between coherence and any of the results of interest here, so all analyses combine coherence levels.

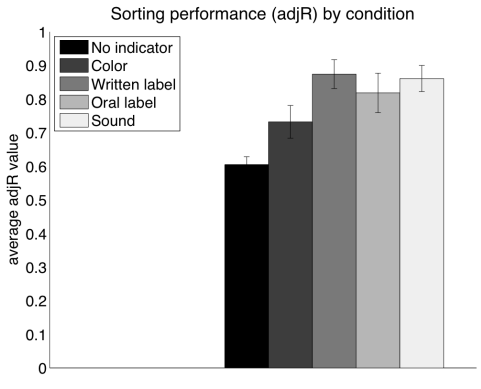


Figure 3: Performance in the sorting task. Subjects in the ORAL LABEL, WRITTEN LABEL, and SOUND conditions used the category indicators to sort at close to optimal levels. When there was no category indicator, people were able to use the less-salient features to sort, but were significantly worse than when there was one.

gories. People were told ahead of time that not all trials would have the same number of items or categories, and they should just sort in whatever way seemed sensible.

After the sorting task was completed, the items remained on the screen and participants were presented with two generalization questions in random order. In *first-order* generalization, participants were shown one of the items they had sorted (without category indicator) and asked which of two novel items would go in the same category as that one. The correct answer had the four coherent features in common with the first, and the incorrect answer had the four other features in common. The *second-order* generalization trials were identical, except that the item shown to the participants had specific feature values that had not been seen before: a person could only answer correctly if they realized that the coherent features (rather than specific values) were what mattered for category organization. As in Perfors and Tenenbaum (2009), our participants performed identically in the first- and second-order generalization, so all analyses collapse them together into one variable, *gen*.

Results

There are two natural questions to ask. First, does the nature of the category indicator affect people's sorting behavior? Second, does it affect how people pay attention to the other, less-salient features of the objects? We can address the second question by examining generalization performance in each condition, since our generalization tasks do not include the category indicator and therefore necessarily rely on the other features. The answer to the first is important for knowing how to interpret the answer to the second: for instance, if generalization performance is poorer because people cannot figure out the correct categories, that does not tell us anything about how people are attending to the less-salient features *given* those categories. We therefore begin with addressing how sorting performance depends on the nature of the category indicator.

Sorting performance

Sorting performance is evaluated using a standard measure for evaluating the similarity between two clusterings of items known as the adjusted Rand Index (*adjR*) of Hubert and Arabie (1985). In this case, we use *adjR* to measure the similarity between the correct category clustering and the category assignments made by the participants. An *adjR* of 1 indicates that the clusters are identical, while 0 is the score one would expect from two random clusterings; scores below 0 indicate that the clusters match less than one would expect by chance.

Figure 3 indicates that category indicator has a strong effect on sorting performance.[3] Participants in the ORAL LABEL, WRITTEN LABEL, and SOUND conditions sorted nearly optimally, which suggests that they used the category indicators to create their categories (since sorting according to category indicator is optimal sorting). Participants on the NO INDICATOR trials were able to use the less-salient features to sort at an above-chance level, but performed worse than when given a category indicator. Finally, people in the COLOR condition sorted halfway in-between, suggesting that color was a more salient feature than the symbolic characters, but not as salient as labels or sounds.

Generalization

Based on sorting performance it appears that participants generally created sensible categories. Were they able to form generalizations about category membership based on the less-salient features? We test this, as explained earlier, by presenting participants with additional items and asking how they would categorize a novel item they had not seen before. Figure 4 demonstrates that generalization in the BOXES condition was generally superior to generalization in the other conditions, all of which were similar to each other.[4] Since generalization depends on what the participant notices about the less-salient features *other* than the category indicator, this suggests that in the BOXES condition people were paying more attention to those features than in any of the other conditions.

These two results, taken together, drive the main conclusion of this paper: labels appear to act more like highly salient features than overt category indicators (boxes). Labels, like highly salient features, support accurate sorting, but are associated with poorer levels of generalization to new items. We have suggested that the reason for this may be because the labels and salient features are directing attention away from the non-salient features during the sorting task; this impairs generalization because attention to the non-salient features is

[3] A one-way Anova on *adjR* by condition was significant: $F(4,158) = 9.77, p = 4.34e^{-7}$. Post-hoc comparisons using the Tukey-Kramer test indicated that the mean *adjR* in the NO INDICATOR condition was significantly different than mean *adjR* in the ORAL LABEL, WRITTEN LABEL, and SOUND conditions.

[4] A one-way Anova on generalization by condition was significant: $F(5,176) = 2.91, p = 0.0149$. Post-hoc comparisons using the Tukey-Kramer test indicated that the generalization in the BOXES condition was significantly different from the NO INDICATOR and ORAL LABEL conditions, and nearly significantly different from the other three.

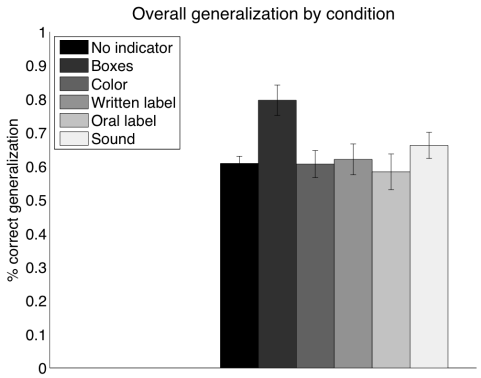


Figure 4: Generalization on the basis of the non-salient features in the BOXES condition was superior to generalization in the other conditions, suggesting that participants in the other conditions did not attend as much to the less-salient features when generalizing.

necessary for accurate generalization of novel items (which are not associated with a label or highly salient feature). This would explain why generalization in those conditions is lower than generalization in the BOXES condition.

However, one minor yet confusing aspect of these results remains: if the salient features are truly directing attention away from the non-salient features, why is generalization performance not poorer on the INDICATOR trials than the NO INDICATOR trials, at least in all conditions other than BOXES? After all, it might be assumed that people are *less* able to use the non-salient features when they have the distracting, highly-salient features around, especially since those features do a very good job at picking out the category members.

Relating sorting and generalization

We address this question by realizing that two factors drive generalization performance, which depends ultimately on knowing which of the less-salient features pick out which categories. It therefore requires not only being able to attend to and identify the less-salient features, but also knowing what the correct categories are. On the INDICATOR trials in the WRITTEN LABEL, ORAL LABEL, SOUND, and COLOR conditions, participants may be less able to attend to the non-salient features, but be better at identifying the categories in the first place. These factors may therefore be cancelling each other out, resulting in generalization that is very similar to the NO INDICATOR conditions.

This possibility yields a testable prediction, namely that in the NO INDICATOR trials sorting performance should be positively correlated with generalization, but in the INDICATOR trials it should be more irrelevant.[5] We would not expect it to be *entirely* irrelevant since, after all, one must be able to identify the categories in order to generalize correctly. However, the converse is not necessarily true: identifying the categories in the INDICATOR conditions would not imply that one should be able to generalize correctly, since generalization requires attention to the less-salient features but categorization does not. We test this by calculating the correlation between sorting accuracy (*adjR*) and generalization (*gen*) for both the INDICATOR and NO INDICATOR trials. Although both are significant, the size of the effect on the INDICATOR trials is markedly weaker.[6] While not conclusive, this is consistent with our interpretation of the results: sorting is less predictive of generalization in the INDICATOR trials because sorting does not depend on the less-salient features in those trials, unlike in the NO INDICATOR situation.

Discussion

This research is motivated by the question of whether labels are cues to category membership or simply highly salient features. The question is difficult to answer in part because it is hard to predict what would be empirically different in each case: no matter what, one would expect labels to be highly interesting, easy to process, and preferentially used as the basis of generalization (but also to be ignorable if they were inconsistent with category structure). We suggest that one difference between cues to category membership and highly salient features is their effect on the processing of the other, less salient features of the objects: highly salient features should direct attention away from the less salient ones, while cues to category membership should direct attention toward them. We tested this by presenting participants with a sorting task involving objects with many complex, low-salience features, and then posing generalization questions that required attention to the less-salient features to answer correctly. Our main results, shown in Figures 3 and 4, suggest that labels act more like highly salient features than they act like boxes (an explicit external cue to category membership).

One might object that this result is not very surprising. After all, stimuli in the BOXES condition may be easier to process since they have one fewer feature – the cue to category membership is the box and the visual organization of the objects, not any features inherent to them. However, in a very real sense this is precisely our point: if something is acting as a cue to category membership, it *should* improve performance by reducing the load required to process the actual features of the objects. Labels, whether oral or written, did not do that in our study.

An important subtlety lies in how we define salience. In what way are the labels in our study really "highly salient"? All of them except for the written label are perceptually noticeable; is this what we mean? The written label was actually fairly small relative to the size of the entire object, so why do people treat it as highly salient? In answer, we note the importance of distinguishing *perceptual* salience from what we might call *conceptual* salience. A feature is perceptually salient because our basic perceptual mechanisms automatically notice and process it preferentially or more easily; this

[5] Note that when we refer to sorting in the INDICATOR trials, we are excluding the BOXES condition, since participants do not actually have to sort anything – the items are already placed into boxes. All of these analysis, therefore, excluded the BOXES condition from the INDICATOR trials.

[6] Spearman's: INDICATOR: $\rho = 0.192, p = 0.007$; NO INDICATOR: $\rho = 0.499, p < 0.0001$.

might be true of speech input (Vouloumanos & Werker, 2004) or auditory input in early childhood (Robinson & Sloutsky, 2004). By contrast, a feature may be *conceptually* salient if we have learned to attend to it preferentially for more abstract conceptual reasons – perhaps because it has proven useful in the past, or if it is easier to process because we have practiced processing it for many years. If the written labels are highly salient, this is probably the sense in which they are. The distinction between the two types of salience gets somewhat blurry at the edges, since many features may be both perceptually and cognitively salient, or change in salience over time. The important point, however, is that for our purposes something is salient if it invites preferential attention or is easier to process; that may be because of perceptual factors, learned conceptual factors, or some mixture of both, and we do not address that question in this work.

One limitation of our study is the fact that it was presented entirely on a computer using bizarre objects with many representationally complex features. The complexity of the features was intentional since we wanted to maximize our chances of creating a situation in which low attention to the features had a measurable effect on generalization; however, it is possible that, due to the unnaturalness of the situation, people adopted a strategy unlike that which they use in the real world. It is also possible that labels, since they are normally referential and communicative, might have a different effect when presented in a communicative, social context rather than on a computer. There is evidence that for children, labelling by a person results in different behavior to labelling by a recorder (Fulkerson & Haaf, 2003), and that non-labels can behave more like labels when presented in a referential context (Campbell & Namy, 2003). However, it is unclear how (or if) these findings will generalize to children, to more naturalistic stimuli, or to different contexts; future work is necessary.

Acknowledgments

We thank Jia Ong for his invaluable help recruiting and running the experimental participants. DJN was supported by an Australian Research Fellowship (ARC grant DP0773794).

References

Anderson, J. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*(3), 409–429.

Balaban, M., & Waxman, S. (1997). Do words facilitate object categorization in 9-month-old infants? *Journal of Experimental Child Psychology*, *64*, 3–26.

Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956). *A study of thinking*. New York, NY: Wiley.

Campbell, A., & Namy, L. (2003). The role of social-referential context in verbal and non-verbal symbol learning. *Child Development*, *74*(2), 549–563.

Davidson, N., & Gelman, S. (1990). Inductions from novel categories: The role of language and conceptual structure. *Cognitive Development*, *5*, 151–176.

Fulkerson, A., & Haaf, R. (2003). The influence of labels, nonlabeling sounds, and source of auditory input on 9- and 15-month-olds' object categorization. *Infancy*, *4*(3), 349–369.

Fulkerson, A., & Waxman, S. (2007). Words (but not tones) facilitate object categorization: Evidence from 6- and 12-month-olds. *Cognition*, *105*(1), 218–228.

Gelman, S., & Markman, E. (1987). Young children's inductions from natural kinds: The role of categories and appearances. *Child Development*, *58*, 1532–1541.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, *103*, 650-669.

Gliozzi, V., Mayor, J., Hu, J.-F., & Plunkett, K. (2009). Labels as features (not names) for infant categorization: A neurocomputational approach. *Cognitive Science*, *33*, 709–738.

Goldstone, R., Lippa, Y., & Shiffrin, R. (2001). Altering object representations through category learning. *Cognition*, *78*, 27–43.

Graham, S., Kilbreath, C., & Welder, A. (2004). Thirteen-month-olds rely on shared labels and shape similarity for inductive inferences. *Child Development*, *75*(2), 409–427.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*, 193–218.

Johansen, M., & Kruschke, J. (2005). Category representation for classification and feature inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(6), 1433–1458.

Kruschke, J. K. (1992). Alcove: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*(1), 22-44.

Kruschke, J. K. (2003). Attention in learning. *Current Directions in Psychological Science*, *12*, 171-175.

Namy, L. (2001). What's in a name when it isn't a word? 17-month-olds' mapping of nonverbal symbols to object categories. *Infancy*, *2*(1), 73–86.

Perfors, A., & Tenenbaum, J. (2009). Learning to learn categories. In *31st Annual Conference of the Cognitive Science Society.*

Plunkett, K., Hu, J.-F., & Cohen, L. (2008). Labels can override perceptual categories in early infancy. *Cognition*, *106*, 665–681.

Roberts, K., & Jacob, M. (1991). Linguistic versus attentional influences on nonlinguistic categorization. *Cognitive Development*, *6*, 355–375.

Robinson, C., & Sloutsky, V. (2004). Auditory dominance and its change in the course of development. *Child Development*, *75*(5), 1387–1401.

Robinson, C., & Sloutsky, V. (2006). Auditory overshadowing and categorization: When decreased visual processing facilitates categorization. In *28th Annual Conference of the Cognitive Science Society.*

Robinson, C., & Sloutsky, V. (2007). Linguistic labels and categorization in infancy: Do labels facilitate or hinder? *Infancy*, *11*(3), 233–253.

Samuelson, L., & Smith, L. (1999). Early noun vocabularies: Do ontology, category structure, and syntax correspond? *Cognition*, *73*, 1–33.

Smith, L., Jones, S., & Landau, B. (1996). Naming in young children: A dumb attentional mechanism? *Cognition*, *60*, 143–171.

Tversky, A. (1972). Elimination by aspects: A theory of choice. *Psychological Review*, *79*, 281-299.

Vouloumanos, A., & Werker, J. (2004). Tuned to the signal: The privileged status of speech for young infants. *Developmental Science*, *7*(3), 270–276.

Waxman, S., & Braun, I. (2005). Consistent (but not variable) names as invitations to form object categories: New evidence from 12-month-old infants. *Cognition*, *95*, B59–B68.

Waxman, S., & Markow, D. (1995). Words as invitations to form categories: Evidence from 12- to 13-month-old infants. *Cognitive Psychology*, *29*, 257–302.

Woodward, A., & Hoyne, K. (1999). Infants' learning about words and sounds in relation to objects. *Child Development*, *70*(1), 65–77.

Xu, F. (2002). The role of language in acquiring object kind concepts in infancy. *Cognition*, *85*, 223–250.

Yamauchi, T., & Markman, A. (2000). Inference using categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(3), 776–795.

Wisdom of the Crowds in Minimum Spanning Tree Problems

Sheng Kung Michael Yi (skyi@uci.edu)
Mark Steyvers (steyver@uci.edu)
Michael D. Lee (mdlee@uci.edu)
Department of Cognitive Sciences,
University of California, Irvine
Irvine, CA 92697, USA

Matthew Dry (matthew.dry@adelaide.edu.au)
Discipline of Pharmacology
University of Adelaide
Adelaide, SA 5005, Australia

Abstract

The 'wisdom of the crowds' effect describes the finding that combining responses across a number of individuals in a group leads to aggregate performance that is as good as or better than the performance of the best individuals in the group. Here, we look at the wisdom of the crowds effect in the Minimum Spanning Tree Problem (MSTP). The MSTP is an optimization problem where observers must connect a set of nodes into a network with the shortest path length possible. A method is developed that creates aggregate solutions based only on the nodes connected in individuals' solutions, without access to spatial information about the nodes. Across the three problems analyzed, the solutions produced by the aggregation method perform better than even the best individual, leading to a strong wisdom of the crowds effect. We show this effect can be observed even with sample sizes as small as 6 individuals.

Keywords: Wisdom of the Crowds; Minimum Spanning Tree Problem; Decision Making; Problem Solving

Introduction

When a problem is posed to a group of individuals, a variety of answers or solutions may be returned. If the accuracy of the individual solutions is unknown, it would be useful to have the ability to extract the collective wisdom contained in the collection of individual responses by aggregating their solutions. The idea that an aggregate solution will perform better than the majority of individuals in the group is referred to as the 'wisdom of the crowds' effect (Surowiecki, 2004). Unlike most research in the topic of distributed cognition and collective intelligence (see Goldstone & Gureckis, 2009 for an overview), where individuals are able to interact in some fashion, individuals in a wisdom of the crowds environment tend to operate independently of one another. Despite this independence and the fact that group members may have widely varying levels of proficiency, aggregation can be found to be effectual in a number of scenarios.

The wisdom of the crowds effect has traditionally been demonstrated for simple questions for which there is a single answer. For example, Galton (1907) asked a large number of individuals to estimate the weight of an ox. He found that the median estimate for the weight of the ox was within 1% of the ox's actual weight. Similarly, Surowiecki (2004) reports that, when polled, the modal answer given by the audience in the US version of the game show "Who Wants To Be A Millionaire" for multiple choice questions is correct more than 90% of the time.

Recently, the wisdom of the crowds idea has also been applied to more complex problems. Steyvers, Lee, Miller, and Hemmer (2009) demonstrated the wisdom of the crowds effect for ordering problems, such as ordering a list of ten states from east to west, ordering the first ten amendments to the U.S. Constitution, or remembering the order of U.S. Presidents. For ordering data, simply taking the mode of individual answers can be problematic because, in many cases, all of the individual orderings are unique. Instead, Steyvers et al. (2009) developed several Bayesian aggregation models that looked at the underlying consistencies in the individuals' orderings to produce an aggregated solution.

A wisdom of the crowds effect has also been observed recently by Yi, Steyvers, Lee, and Dry (submitted), for a difficult combinational optimization problem known as the Traveling Salesman Problem (TSP: see Applegate, Bixby, Chvátal, & Cook, 2006 for a review). In the TSP, the goal is to connect a set of nodes to make the shortest path possible, with the constraints that each node can be visited only once, and the path must end at the same node as it started. The aggregation method developed by Yi et al. (submitted) did not require any spatial information about the locations of the nodes. Instead, the method took advantage of the knowledge of which nodes are connected in individual solutions and selected a solution that maximized the agreement across individuals as to the sequence of nodes visited.

Generating a wisdom of the crowds effect for TSP problems in this way provides an example of a potentially powerful and general approach to aggregating individual knowledge and abilities. The key feature is that all of the aggregation is based on the observed ordering of individuals and their patterns of agreement. No representation was needed of the complex multidimensional TSP stimuli, nor were evaluation measures for individual performance used. For these reasons, the results of Yi et al. (submitted) suggest

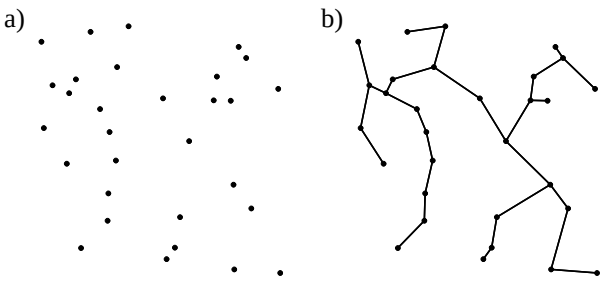


Figure 1: An example MSTP solution (a) and its optimal solution (b).

an approach to finding the wisdom of the crowd in challenging real-world situations where the problem space is too large or complicated to represent formally, and there is no clear way to quantify the merits of proposed solutions.

Of course, however, it may be that the TSP result is simply a special or isolated case. Accordingly, in this paper, we explore the possibility of a wisdom of the crowds effect for another complex problem solving task, known as the Minimum Spanning Tree Problem (MSTP). First, as for TSPs, we develop an aggregation method that is based on easily observed features of individual solutions. Then, we apply the method to previously collected data for several MSTPs. We observe a strong wisdom of the crowds effect, in which the aggregate solution is closer to optimal than any individual solution. Finally, we examine how many individual solutions are needed for good aggregation, and discuss how our approach could be extended, modified, and applied to more general problems.

Minimum Spanning Tree Problems

In MSTPs, participants are required to find the shortest possible network that links together a set of nodes in some spatial configuration. An example stimulus and optimal solution for an MSTP is shown in Figure 1. In contrast to the TSP, there is no constraint on the paths that can be formed. Each node can be connected to multiple nodes. The optimal solution is an open, branching path system or tree, in which nodes can be linked to one or more other stimulus nodes.

Finding the optimal solution to MSTPs has an obvious real-world engineering application in regards to finding the minimal length network of cables or pipes needed to join discrete geographical locations (e.g., Borůvka, 1926). However, MSTPs are also of interest from a psychological perspective, providing insight into human decision-making, individual differences in cognitive abilities, and visuo-perceptual organization (e.g., Burns, Lee & Vickers, 2006; Vickers, Mayo, Heiman, Lee & Hughes, 2004). Specifically, the MSTP belongs to a class of difficult visual optimization problems such as the TSP and the Generalized Steiner Tree Problem (GSTP). Despite the apparent difficulty (and in some cases intractability) of these optimization problems, human observers are often able to find optimal or close-to-optimal solutions in a time frame that increases as a linear function of problem size (e.g., Dry, Lee, Vickers & Hughes, 2006; Graham, Joshi, & Pizlo, 2000).

An important finding from the literature on human solutions to MTSPs is that there are meaningful individual differences (e.g., Burns et al., 2006). As Surowiecki (2004) and others have emphasized, a precondition for the wisdom of the crowds effect is that there is variation between individuals. Intuitively, the hope is that some individuals complete some parts of an MSTP optimally or near-optimally, while other individuals complete different parts well. In this scenario, the aggregation of the individual solutions could potentially improve on both.

Dataset

The data were taken from Burns et al (2006). In that study, as part of a larger battery of optimization tasks and cognitive abilities tests, 101 participants completed 3 MSTPs, with 30, 60 and 90 nodes. The problems were comprised of black nodes on a uniform white background and were presented on color computer monitors.

The participants generated spanning trees by pointing and clicking with the mouse cursor, and were allowed to add or remove links as they saw fit. They were instructed to connect the nodes by making a system of links, using as many links as they felt necessary, under the condition that the resulting system had the minimum overall possible length. The participants worked without time limits and were asked to be as accurate as possible. The results of the empirical solutions are displayed in Figure 2, expressed as the percentage above optimal solution length (PAO = 100*[empirical length/optimal length - 1]). Participants provided solutions that were on average around 6% longer than the optimal solution. Importantly however, there were significant individual differences with some individuals providing solutions that were much closer to the optimal solution. Despite the large number of participant solutions available, there was no case in any problem where any participant's solution exactly matched that of another participant.

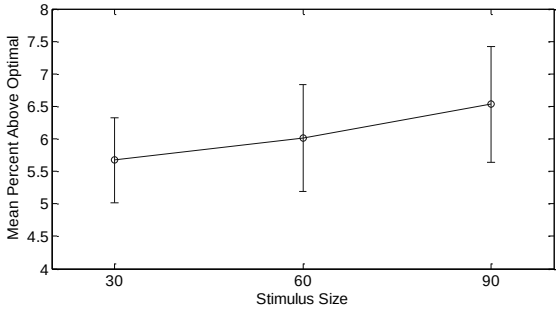


Figure 2. Mean empirical PAO for MSTP with 30, 60 and 90 nodes; error bars indicate standard deviation of individual performance.

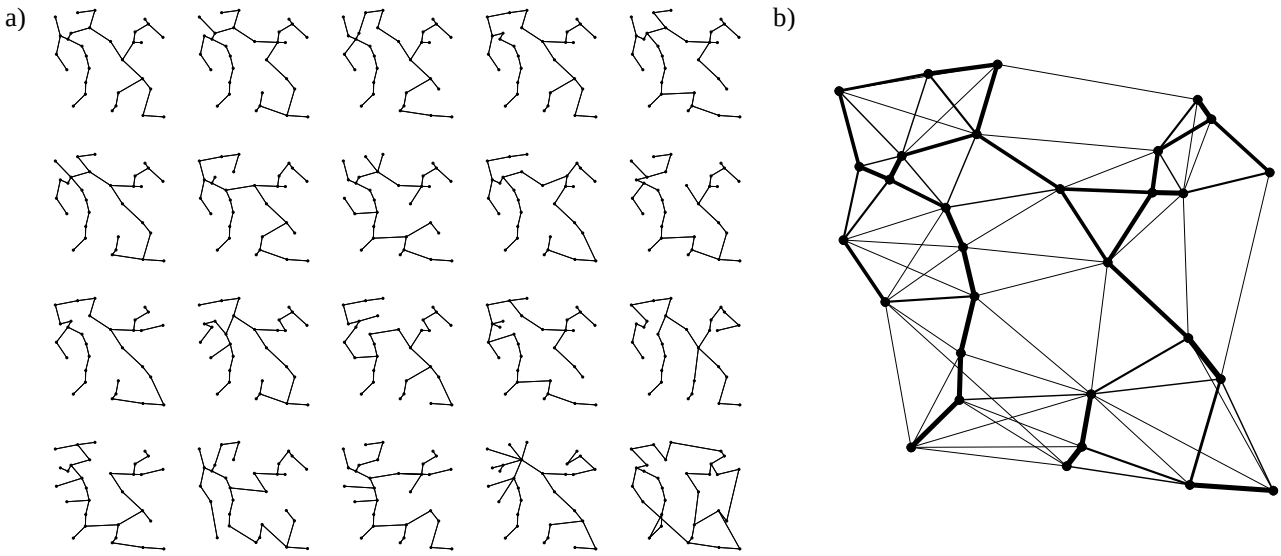


Figure 3. a) Representative subject solutions for the 30-node MSTP, the best subject solution in the upper left with decreasing performance across rows and the worst subject in the lower right. b) Visualization of agreement matrix on problem nodes. Vertices selected by at least one subject are drawn; thicker lines indicate higher agreement.

Aggregation Method

The data for the aggregation method were restricted to the information of which nodes each participant connected in their solutions. In particular, the method was not given any spatial information about the node locations, and so relied solely on the information contained in the participant solutions to create a proposed network. The aggregation method operates under the assumption that vertices between nodes that are better for inclusion in a MSTP solution tend to be selected by more participants. An aggregate solution that maximizes the degree of agreement with participant solutions can therefore be expected to have good performance.

In order to obtain an aggregate solution, we first arranged the solutions of all individuals in an $n \times n$ agreement matrix, where n is the number of nodes in the problem. Every cell a_{ij} in the matrix records the number of participants that connected nodes i and j in their solutions. A visualization of the agreement matrix is depicted in Figure 3b. We then derived a cost matrix of the same size with cell values $c_{ij} = k - a_{ij}$, where k was the total number of participants; connections with higher agreement would thus have lower costs. This cost matrix is then used as the input to a standard MSTP algorithm to obtain a proposal solution for the aggregate.

The MSTP can be solved optimally in polynomial time through the use of simple greedy algorithms such as Prim's algorithm (Jarník, 1930; Prim, 1957). In Prim's algorithm, a starting node is randomly selected from all nodes. At each step in the algorithm, the vertex with the smallest cost that connects an unconnected node to the already-connected nodes (or starting node, in the first step) is added to the network, until all nodes are connected. Despite the fact that the algorithm is greedy in nature, it is always guaranteed to output the minimum spanning tree depending on the cost metric being used. When the vertex costs are equal to the distances between nodes, Prim's algorithm is guaranteed to produce a spanning tree with the shortest total length. In our research, the vertex costs upon which Prim's algorithm is

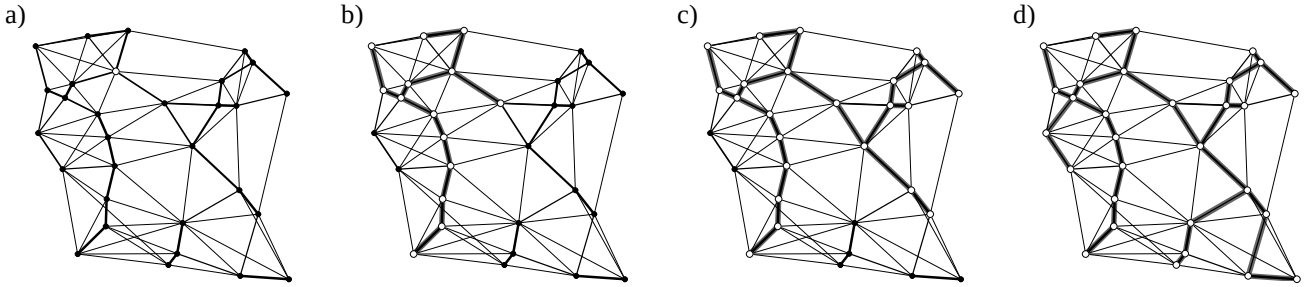


Figure 4. Example demonstration of Prim's algorithm on the 30-node MSTP. A random node is selected, shown in white (a.). At each step of the algorithm, vertices with the smallest cost (i.e., highest agreement) that connect an unconnected node (black) to those already connected (white) are added to the network until all nodes are connected (b-d.)

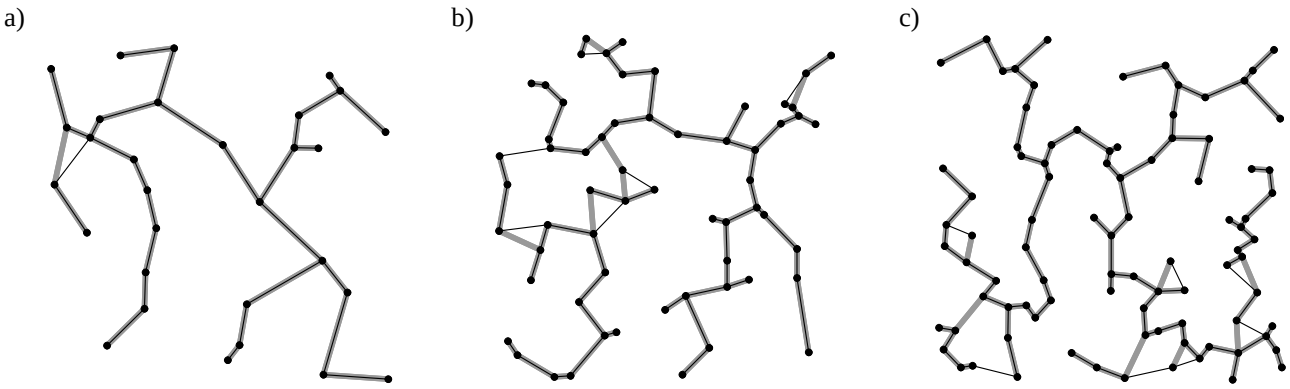


Figure 5. Solution paths for the aggregate method (thin black) and the optimal minimum spanning tree (thick gray) for the a) 30-node, b) 60-node, and c) 90-node MSTPs.

applied are set using the cost matrix based on subject agreement above. The algorithm will still produce a network with minimum total cost, but in this case, the network represents the spanning tree that has the highest agreement with the participant solutions. It is this solution that is selected by the aggregation method. A demonstration of the algorithm is shown in Figure 4.

Optimality of Prim's algorithm can be verified by considering the necessary conditions for a minimum spanning tree. For a solitary node, it is necessary for it to connect to its nearest neighbor using the vertex with the lowest cost. If a spanning tree is created without using such a vertex, and that node is connected to the others via some other vertex, it does not change the connectedness of the network by deleting that other vertex and instead connecting to the nearest neighbor, but it does reduce the total path length. This makes the first step of Prim's algorithm, connecting a random node to its nearest neighbor, a sensible action. The logic can be followed by induction to the sub-networks drawn by Prim's algorithm by treating each sub-network as if it were a single node, thus showing optimality. In cases where multiple potential vertices with the same cost may be selected for addition to the spanning tree, then any of the candidates may be chosen without affecting the solution's optimality.

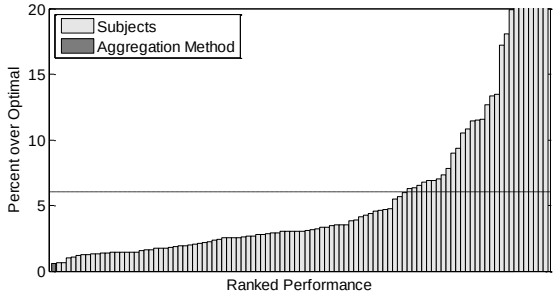


Figure 6: Ranked performance of subjects and the aggregation method averaged over all problems. Dashed horizontal line indicates mean subject performance.

Results

Figure 5 shows the optimal minimum spanning trees in thick gray lines and solutions selected by the aggregation method in thin black lines while participant and aggregate solution performance is provided in Table 1. Additional performance statistics are noted for the aggregate solutions: the amount of agreement the aggregate solutions had with subject solutions and a count of the number of participants whose performance is better than, worse than, or same as the aggregate. Subject agreement values were calculated as the proportion of subject vertices coinciding with vertices present in the aggregate solution; these can be obtained by noting the value of the aggregate path as measured on the agreement matrix, then dividing by $(n\text{-}1)k$, the number of vertices multiplied by the number of subjects. The aggregate

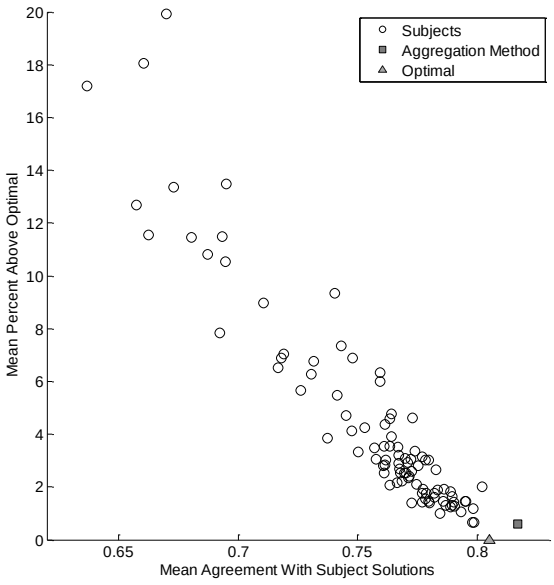


Figure 7: Performance averaged over all problems against mean solution agreement with subject solutions.

Table 1: Subject and Aggregate Method Performance on MSTPs (% network length over optimal)

Problem	subject performance		aggregate method performance				
	subj. best	subj. mean	path length	subj. agreement	# subj. better	# subj. same	# subj. worse
30 nodes	+0.000%	+5.672%	+0.059%	.7856	1	0	100
60 nodes	+0.037%	+6.010%	+1.410%	.8263	21	0	80
90 nodes	+0.235%	+6.533%	+0.310%	.8392	1	0	100
Overall	+0.644%	+6.072%	+0.593%	.8171	0	0	101

method solutions perform quite well, beating the average participant by a large margin. In the 30- and 90-node problems, the performance of the aggregate is bested only by a single participant out of the full set of 101. The aggregate performs relatively worse in the 60-node problem, but still better than most individuals. When performance is averaged over all problems, the aggregate performs better than any individual (Figure 6). Interestingly, the proportion of vertex agreement with participants increased with problem size, and solutions selected by the aggregate did not completely match any single individual on any problem. Figure 7 contains a plot of solution performance against the proportion of agreement with participant solutions averaged equally over all problems for all subjects, the optimal solution, and the aggregate solution. There is a clear correlation between individual performance and the amount of agreement their solutions had with other participants ($r = -.9602$). The optimal solution also has a high rate of coincidence with participant solutions, more than any individual.

Performance of the aggregation method under smaller sample sizes was also investigated. For each sample taken, subjects were selected randomly from the full dataset and aggregate solutions were created for all problems, their performances compared to the subjects in the sample that generated them. In cases where Prim's algorithm encountered a choice between vertices of the same cost, one was chosen at random to create the proposal solution. Solution performance for selected sample sizes is noted in Figure 8, averaged over 1000 random draws at each sample size. We find that for samples of as small as size 6, the aggregate is able to obtain performance that is, on average, significantly better than the mean subject and close to that of the best subject in the sample. Averaged over all problems, the aggregate was outperformed by about one participant at all sample sizes investigated. In certain cases for individual problems, the aggregate solution outperformed all participants in the sample; this was much more common for the 30-node and 90-node problems than the 60-node problem.

Conclusions

We have demonstrated a strong wisdom of the crowds effect for the MSTP using a simple aggregation method on participant solutions. The aggregation method was reliant only on the knowledge of which nodes were connected by each participant, requiring no information regarding the spatial characteristics of the problems themselves. In addition, the simple greedy algorithm used to generate solutions required no input parameters to run. The aggregation method solutions generally have performances ranking among the best participants on individual problems, and perform better than any individual when averaged over all problems. Even when the number of available participants was reduced down to as low as 6, the aggregation method was still able to extract enough information to propose solutions that produced performances significantly better than the mean subject and exceeding most or all participants in the sample.

While performance of the aggregation method is quite good, there are potential areas for expanding on the method. It was noted that there was a clear correlation between a

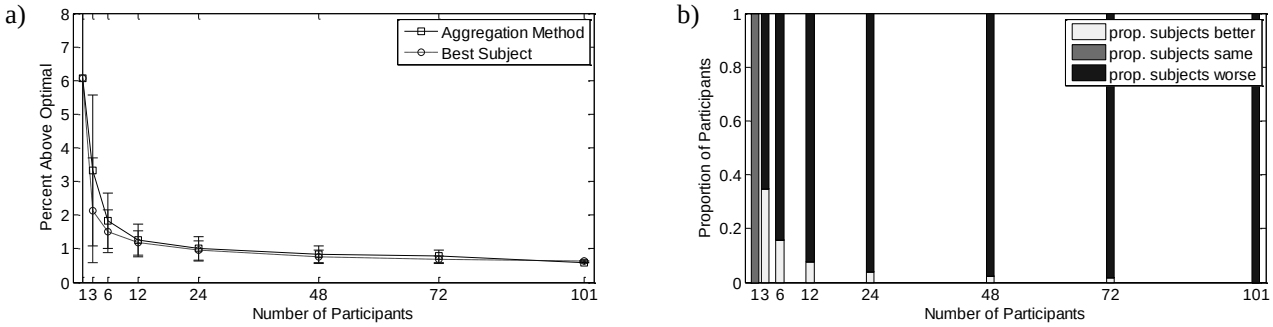


Figure 8. Performance of the aggregate method for selected sample sizes, taken across problems. a) Mean PAO for aggregate and best subject in each sample, error bars indicate standard deviation of individual samples. b) Proportion of subjects with better, same, or worse performance than the aggregate.

participant's performance and the amount of agreement they had with other participants. It may be useful if it were possible to identify 'experts' in the data and weight their responses over that of others. This approach of amplifying expertise may be most useful for when sample sizes are small. Due to the fact that there are so few participant solutions to draw from, there may be many networks that can potentially be chosen by the algorithm that share the same agreement with participant solutions, but carry very different performances in terms of actual distance. If participants can be weighted differently, then there will be less ambiguity. However, with the complexity of the problem structure, it is a difficult problem to create a formal system in which this can be done.

More generally, the results presented here, when coupled with those presented by Yi et al. (submitted) for the TSP, suggests that it may be possible to achieve wisdom of the crowds effects for complicated and only partly defined problems. While the MSTP does have a simple solution algorithm, and the TSP has good approximate solution algorithms for small numbers of nodes, our results show that near-optimal performance can be obtained from simple properties of the sub-optimal sets of solutions produced by a group of people.

In other words, our results show that there is an alternative route to solving these problems, not based on complicated algorithms, detailed stimulus information, and precise performance metrics. Instead, we have shown that the orders people produce can be combined to achieve near-optimality. Of course, for TSPs and MSTPs, there is not much reason to go to the effort of collecting human solutions when good algorithms are available. But our approach will continue to apply for different sorts of difficult problems where, for example, the stimuli or problem space is hard to represent in a formal way. This representational burden is borne by the individual providing solutions, and there is no need for any formal attempt to characterize the problem space. Even more intriguingly, our approach will apply in situations, such as some types of aesthetic judgment, where people agree on what constitutes a good answer once it is produced, but cannot define exactly what metric they are using. Since our aggregation approach just uses the patterns of relationships between individual judgments, and does not need a performance measure, it is equally applicable to these poorly defined problems.

We are currently investigating the use of the wisdom of the crowds approach described in this paper to the "wisdom of the crowds within", the idea that one can aggregate over multiple judgments from a single individual to obtain performance better than the individual judgments alone (Vul & Pashler, 2008). By applying transformations to MSTPs, we can easily test an individual on multiple repetitions of the same problem while minimizing bias from their responses on previous trials. We are also looking at applying the aggregation approach to a less-well defined aesthetic judgment task. Participants were asked in Dry, Navarro, Preiss, and Lee (2009) to connect point stimuli based off of constellations into perceived structures. It is possible that a structure created by aggregating over individuals is perceived as more aesthetically pleasing than individual patterns. The application of our approach to aggregation to these sorts of challenging problems seems a promising direction for further wisdom of the crowds research.

References

Applegate, D. L., Bixby, R. E., Chvátal, V., & Cook, W. J. (2006). *The Traveling salesman problem: A computational study*. Princeton, NJ: Princeton University Press.

Borůvka, O (1926). O jistém problému minimálním. *Práce Moravské Přírodovědecké Společnosti*, 3, 37–58.

Burns, N. R., Lee, M. D., & Vickers, D (2006). Are Individual Differences in Performance on Perceptual and Cognitive Optimization Problems Determined by General Intelligence? *Journal of Problem Solving*, *1*(1), 5-19.

Dry, M. J., Lee, M. D., Vickers, D., & Hughes, P. (2006). Human Performance on visually presented traveling salesperson problems with varying numbers of nodes. *Journal of Problem Solving*, 1, 20-32.

Dry, M.J., Navarro, D.J., Preiss, K., Lee, M.D. (2009) The Perceptual Organization of Point Constellations. In N. Taatgen, H. van Rijn, J. Nerbonne, & L. Shonmaker (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, 1151-1156. Austin, TX: Cognitive Science Society.

Galton, F. (1907). Vox Populi. *Nature*, *75*, 450-451.

Goldstone, R. L., Gureckis, T. M. (2009) Collective Behavior. *Topics in Cognitive Science*, *1*, 412-438.

Graham, S. M., Joshi, A., & Pizlo, Z. (2000). The Traveling Salesman Problem: A hierarchical model. *Memory & Cognition*, *28*(7), 1191-1204.

Jarník,V (1930). O jistém problému minimálním. *Práce Moravské Přírodovědecké Společnosti*, 6, 57-63.

Prim, R. C. (1957). Shortest connection networks and some generalizations. *Bell System Technical Journal*, 36, 1389-1401.

Steyvers, M., Lee, M.D., Miller, B., & Hemmer, P. (2009). The Wisdom of Crowds in the Recollection of Order Information. In J. Lafferty, C. Williams (Eds.), *Advances in Neural Information Processing Systems*, *23*. MIT Press.

Surowiecki, J. (2004). *The Wisdom of Crowds*. New York, NY: W. W. Norton & Company, Inc.

Vickers, D., Mayo, T., Heitmann, M, Lee, M. D., & Hughes, P. (2004). Intelligence and individual differences in performance on three types of visually presented optimization problems. *Personality and Individual Differences*, *36*, 1059-1071.

Vul, E. & Pashler, H. (2008). Measuring the Crowd Within: Probabilistic Representations Within Individuals. *Psychological Science*, *19*(7), 645-647.

Yi, S. K. M., Steyvers, M., Lee, M. D., & Dry, M. J. (submitted). Wisdom of the Crowds in Traveling Salesman Problems.

The Effect of Labels on Visual Attention: An Eye Tracking Study

Catherine A. Best (best.140@osu.edu)
Center for Cognitive Science, The Ohio State University
208G Ohio Stadium East, 1961 Tuttle Park Place, Columbus, OH 43210, USA

Christopher W. Robinson (robinson.777@osu.edu)
Center for Cognitive Science, The Ohio State University
208F Ohio Stadium East, 1961 Tuttle Park Place, Columbus, OH 43210, USA

Vladimir M. Sloutsky (sloutsky.1@osu.edu)
Center for Cognitive Science, The Ohio State University
208D Ohio Stadium East, 1961 Tuttle Park Place, Columbus, OH 43210, USA

Abstract

The effects of language on categorization are well documented; however, underlying mechanisms are under debate. According to one account, words facilitate categorization by highlighting commonalities among labeled objects. Although there is some behavioral evidence consistent with this claim, research remains limited for whether labels can direct infants' attention to corresponding visual features. In the current study, adults and infants were presented with 10 different exemplars that were either associated with 10 different labels, the same label, or presented in silence. An eye tracker recorded visual fixations to common and unique features throughout familiarization. Experiments 1 and 2 provide evidence that unique labels can direct infants' and adults' attention to unique features (compared to a silent condition); however, the effect of hearing the same label associated with different objects was less robust in both age groups.

Keywords: Attention; Language; Categorization

Introduction

Beginning at birth, infants must learn to make sense of the world, and the ability to form categories is an important part of this learning. Although very young infants can quickly learn visual categories (Bomba & Siqueland, 1983; Eimas & Quinn, 1994), there is some evidence that words and other types of sounds influence this process. For example, young infants are often better at learning visual categories when category members are associated with the same word than when the same visual stimuli are paired with a nonlinguistic sound (Balaban & Waxman, 1997; Fulkerson & Waxman, 2007; Robinson & Sloutsky, 2007). Exposure to words may also help infants individuate objects. Research demonstrates that infants who hear two different words (but not two sounds) expect two objects to be hidden by an occluder (Xu, 2002). Labels also influence what category structure infants learn. For instance, while looking at the same visual images, infants who heard one word associated with all exemplars learned one category; whereas, infants who heard two words learned two categories (Plunkett, Hu & Cohen, 2008). Finally, although words and sounds often have different effects on categorization and individuation, only a few studies have directly compared infants' performance in label and sound conditions to a silent baseline. These comparisons illustrate that compared to a silent condition, words and sounds can interfere with categorization of visual input, often with greater interference from sounds than from words (Robinson & Sloutsky, 2007; 2008).

To account for the effect of labels on category learning, several mechanisms have been put forth. First, Waxman and colleagues argue that infants understand the conceptual importance of words and that words (but not sounds) facilitate categorization by highlighting the commonalities among labeled entities (Fulkerson & Waxman, 2007; Waxman, 2003). Given the findings reported by Plunkett *et al*. (2008) and Xu (2002), it is also possible that unique words may also facilitate the formation of multiple categories by highlighting unique features among labeled entities. In contrast, Sloutsky and colleagues argue that infants and young children have difficulty processing multimodal information, with words and sounds often attenuating visual processing (Robinson & Sloutsky, 2004; Sloutsky & Napolitano, 2003). Differential effects of words and sounds stem from sounds interfering with visual tasks more than words (as opposed to words facilitating categorization above a silent control). Thus, according to Waxman and colleagues, hearing common and unique words should increase attention to common and unique features in the early stages of development. In contrast, according to Sloutsky and colleagues, early in development words should have no facilitative effect above a silent condition and may even interfere with visual processing.

The aim of the current set of studies was to explore *how* words might affect visual attention by utilizing eye-tracking technology. Measuring eye movements during experimental tasks provides an online measure of attention. By tracking the gaze of infants and adults during a simple familiarization task, we can investigate whether patterns of visual attention during learning differ with respect to varying language cues.

Overview of Current Studies

To investigate the effect of labels on visual attention, gaze data were collected from both infants and adults while viewing novel stimuli paired with novel labels. Half the features on each stimulus were shared among the

sequentially presented stimuli (i.e., common features); whereas, half of the features were unique. If participants inferred identical labels indicated that images were members of the same object category, it was predicted that participants who heard the same label associated with different images would accumulate more looking to common features than participants in the silent condition. Similarly, if participants inferred different labels indicated that images were members of different object categories, it was predicted that participants who heard different labels associated with different images would accumulate more looking to unique features than participants in the silent condition. Experiment 1 compared adults' attention to common and unique features across familiarization when labels were consistent, varying, or when images were presented in silence. Experiment 2 tested infants with the same three sets of stimuli as presented to adults.

Experiment 1

Method

Participants Thirty-six adults (20 men, 16 women), ranging in age from 18 to 21 years (M = 18.58, SD = 0.79) were tested, with 12 adults per condition. Adults were recruited from an Introductory Psychology class. Participants provided written consent upon arrival to the laboratory. All adults reported normal or corrected-to-normal vision and normal hearing prior to recruitment.

Apparatus A non-invasive Tobii T60 eye tracker measured eye gaze by computing the pupil-corneal reflection at a sampling rate of 60 Hz (i.e., 60 gaze data points collected per second for each eye). The eye-tracking device, which is integrated into the base of a high-resolution 17-inch computer monitor, was located on a table inside a darkened testing booth, enclosed by curtains. A trained experimenter monitored the experiment on a 19-inch Dell OptiPlex 755 computer located outside of the testing booth. A Sony Network camera was located inside the testing booth to the side of the eye tracker displaying a live feed view of the participant that an experimenter monitored on a 9-inch black and white Sony SSM-930/930 CE television. Two Dell computer speakers were positioned behind a curtain and out of view on either side of the eye tracker.

Stimuli Stimuli included 12 audio-video interleave (AVI) files. Each AVI file combined a static bitmap image with an auditory speech component. The visual images consisted of four uniquely-shaped parts. Two parts were common across all stimuli and two parts were unique across all stimuli. See Figure 1 for example stimuli. The common parts were the same color and shape; whereas, the unique parts varied in color and shape. The auditory input consisted of one-syllable novel labels spoken by a female adult (e.g., *dax, bim*, *fep*, *gid*, *jup*, *meb*, *pof*, *raz*, *sop*, and *zot*). All labels were spoken within the context of a simple command (*e.g.,* "Look at the dax."). Speech was recorded using Cool Edit 2000. Each sound file was saved as an audio compression manager waveform at 44.10 kHz, 16 Bit, in stereo. Audio files were then imported into Macromedia Flash, paired with corresponding bitmap images, and exported as Windows AVI files. All AVI files were 6000 ms in duration. The image lasted for the total duration of 6000 ms. The audio occurred with the onset of the image and lasted 2000 ms in duration. The remaining 4000 ms consisted of silence. The stimuli for the silent condition were identical to the label conditions; except, the speech was removed entirely.

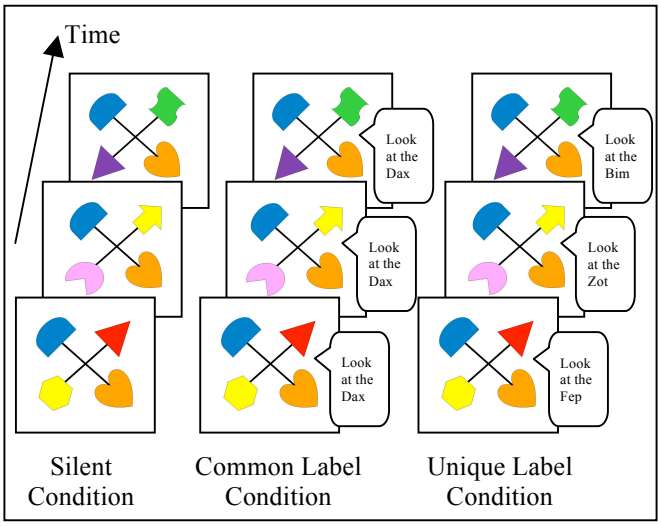


Figure 1: Example stimuli

Design The experiment utilized a between-subjects design. Participants were randomly assigned to one of three experimental conditions (*i.e.*, common label, unique label, or silent). The common condition consisted of 10 different novel images, each paired with the same novel label. The unique condition consisted of 10 different novel images, each paired with 10 different novel labels. The silent condition consisted of 10 different novel images, each presented in silence. The visual input was the same for all conditions and was presented in a random sequence. Only the auditory input differed across conditions.

Procedure Participants sat centered in front of the eye tracker within an approximate viewing distance of 60 cm. Prior to the experiment, participants completed a 5-point calibration sequence lasting less than one minute. The calibration points consisted of a moving red dot appearing in different locations on the screen. The experiment commenced after successful calibration. Participants were asked to pay close attention to the images because they would be asked about them later. All participants were familiarized to 10 stimuli presented one at a time for 6000 ms. Each stimulus was presented subtending an approximate horizontal visual angle of 11° and an approximate vertical visual angle of 11°. A black screen was presented for 1000ms between trials. After training, participants were

tested with four paired preference trials, each trial displaying one old image and one new image presented in silence. Adults were asked to select the new image from the old image. Test stimuli were the same size as familiarization stimuli. Each test trial remained visible until adults made a decision. All gaze data were recorded by the computer using Tobii Studio gaze analysis software.

Results and Discussion

All participants correctly identified the novel stimuli on every test trial; therefore, no one was excluded from the current study. Primary analyses presented below focused on adults' attention to common and unique features during familiarization.

Unfiltered gaze data were exported from the computer using Tobii Studio gaze analysis software. A point of gaze was recorded if a participant made a fixation to pre-determined areas of interests (AOIs). Four AOIs were defined as a rectangle surrounding the four parts of each stimulus image. Gaze data were combined for the two common features and for the two unique features to obtain a measure of looking to unique or common features per refresh rate. These data were used to calculate unique and common feature preference scores based on the proportion of looking time to either unique or common features as compared to the total time looking to all features combined.

Effect of Unique Labels To determine if unique labels pushed adults' attention to unique features, we compared preference for unique features in the unique label condition to preference for unique features in the silent condition.

Gaze data were analyzed using a moving average of participants' attention across time to smooth out temporary fluctuations within a given trial (*i.e.*, 3 trials were averaged per time point such that time point 1 averaged trials 1 to 3, time point 2 averaged trials 2 to 4, and so on). A repeated-measures analysis of variance (ANOVA) was conducted on mean preference for unique features with Condition (unique label vs. silent) as a between-subjects factor and Time Point (1 vs. 2 vs. 3 vs. 4 vs. 5 vs. 6 vs. 7 vs. 8) as a within-subjects factor. Results revealed a significant main effect of Time Point, $F(7, 154) = 6.56$, $p < .001$, a main effect of Condition, $F(1, 22) = 4.44$, $p < .05$, and a significant Time Point X Condition interaction, $F(7, 154) = 3.74$, $p < .01$. Overall, mean preference scores for unique features were significantly greater in the unique condition ($M = .66$) compared to the silent condition ($M = .53$). Specifically, as shown in Figure 2, post-hoc comparisons revealed that mean preference scores for unique features were significantly greater in the unique condition compared to the silent condition at time points 5, 6, and 7, ts > 2.16, ps < .05. These results support the idea that unique labels facilitate attention to unique features.

To obtain a better understanding of the dynamics of attention, unique preference scores were averaged across trials and plotted as a function of time. As can be seen in Figure 3, preference for the unique features in the unique label condition was consistent across the entire 6000 ms trial duration. This attention pattern in adults corroborates evidence for unique labels facilitating attention to unique features.

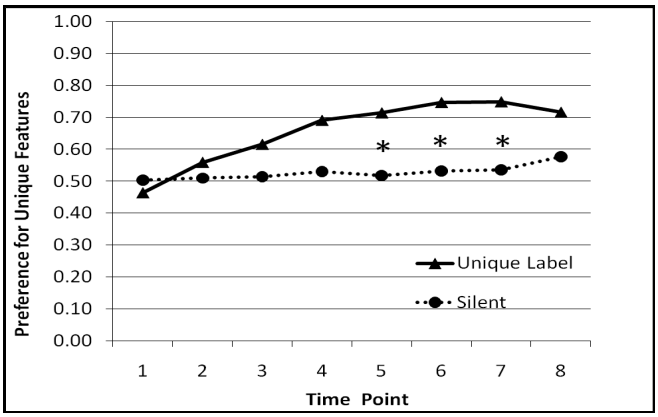


Figure 2: Adults' mean preference for unique features by time point. (Note: Time points represent moving averages).

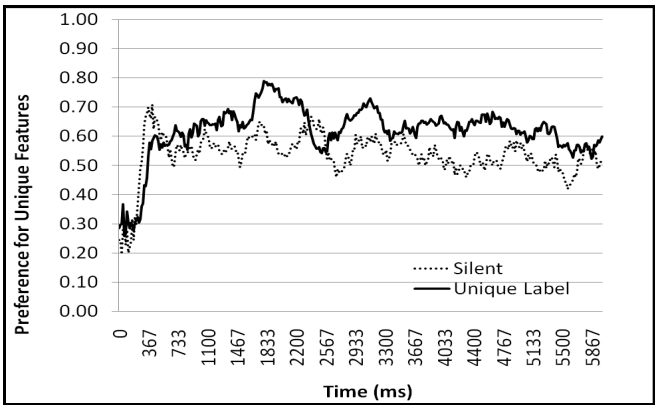


Figure 3: Adults' mean preference for unique features over time.

Effect of Common Labels To determine if common labels pushed adults' attention to common features, we compared preference for common features in the common label condition to preference for common features in the silent condition, using the same sample of adults that was previously compared to the unique label condition.

As in the unique label condition, gaze data were analyzed using a moving average. A repeated-measures ANOVA was conducted on mean preference for common features with Condition (common label vs. silent) as a between-subjects factor and Time Point (1 vs. 2 vs. 3 vs. 4 vs. 5 vs. 6 vs. 7 vs. 8) as a within-subjects factor. Results revealed a significant main effect of Time Point, $F(7, 154) = 2.20$, $p < .05$. Preference for common features attenuated over time for both the common label and silent conditions. However, as shown in Figure 4, mean preference scores for common features were not significantly different between conditions at any point in the course of familiarization.

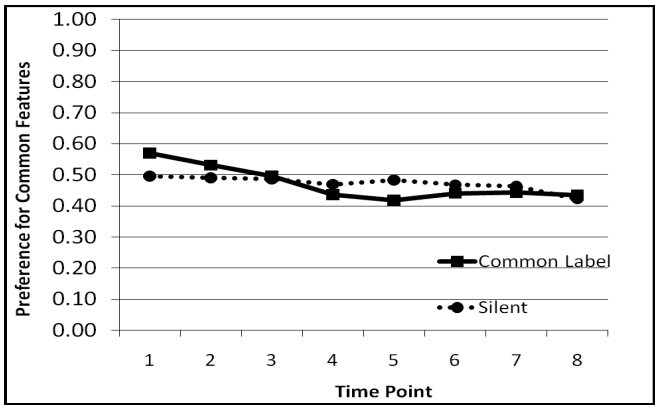


Figure 4: Adults' mean preference for common features by time point. (Note: Time points represent moving averages).

Furthermore, preference for common features was analogous for the entire 6000 ms trial duration in the common label and silent conditions when preferences scores were averaged across trials and plotted as a function of time (see Figure 5). Therefore, the pattern of attention over time did not suggest that common labels directed adults' attention to common features.

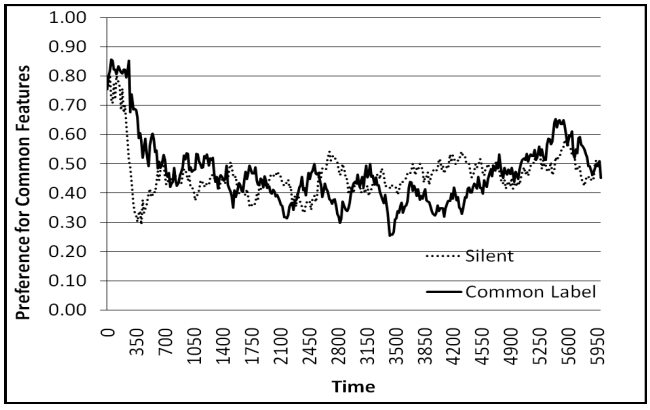


Figure 5: Adults' mean preference for common features over time.

Summary Experiment 1 found a robust effect of unique labels directing attention to unique features and no significant effect of common labels directing attention to common features. Adults presented with varying labels (*i.e.*, unique) compared to silence disproportionately distributed their attention to objects' unique versus common features. In contrast, adults presented with consistent labels (*i.e.*, common) compared to silence did not disproportionately distribute attention to objects' common versus unique features. The purpose of Experiment 2 was to investigate how labels affect visual attention in infancy. Do unique and common labels direct infants' attention to correlated visual features?

Experiment 2

Method

Participants Thirty-six infants, (19 boys and 17 girls), ranging in age from 16 to 24 months (M = 19 months, 9 days; SD = 3 months, 21 days) were tested, with 12 infants per condition. Ten additional infants were excluded from analyses due to fussiness. Infants were recruited from local birth records. Parents provided written consent upon arrival to the laboratory. All infants were healthy and developing typically.

Materials and Design The apparatus, stimuli, and design were identical to Experiment 1.

Procedure Infants sat on a caregiver's lap and were positioned in front of the eye tracker within an approximate viewing distance of 60 cm. The procedure was identical to Experiment 1 with three exceptions. First, during calibration, rather than a shrinking red dot, infants saw a dynamic kitten image appearing on the screen with a corresponding "bounce" sound. Second, unlike adults, infants were not provided with instructions. Third, a dynamic bouncing ball was presented as an attention-grabbing fixation between trials.

Results and Discussion

Infants in all three conditions (*i.e.*, unique label, common label, and silent) demonstrated a mean novelty preference based on the average looking time to new versus old objects across all four test trials, ts > 2.55, ps < .05. Mean novelty preference scores did not differ among the three conditions. Primary analyses presented below focused on infants' attention to common and unique features during familiarization.

Effect of Unique Labels As in Experiment 1, unfiltered gaze data were exported and combined into looking to common features and looking to unique features. To determine if unique labels directed infants' attention to unique features, we compared preference for unique features in the unique label condition to preference for unique features in the silent condition. As with adults' data, infants' gaze data were analyzed using a moving average of participants' attention across time to smooth out temporary fluctuations within a given trial (*i.e.*, 3 trials were averaged per time point such that time point 1 averaged trials 1 to 3, time point 2 averaged trials 2 to 4, and so on). A repeated-measures ANOVA was conducted on mean preference for unique features with Condition (unique label vs. silent) as a between-subjects factor and Time Point (1 vs. 2 vs. 3 vs. 4 vs. 5 vs. 6 vs. 7 vs. 8) as a within-subjects factor. Results revealed a significant main effect of Time Point, $F(7, 147) = 3.96$, $p < .01$. Although the effect of Condition did not reach significance, as shown in Figure 6, independent t-tests revealed that mean unique preference scores were

significantly greater in the unique label condition compared to the silent condition at time point 2, $t(22) = 2.03$, $p = .05$.

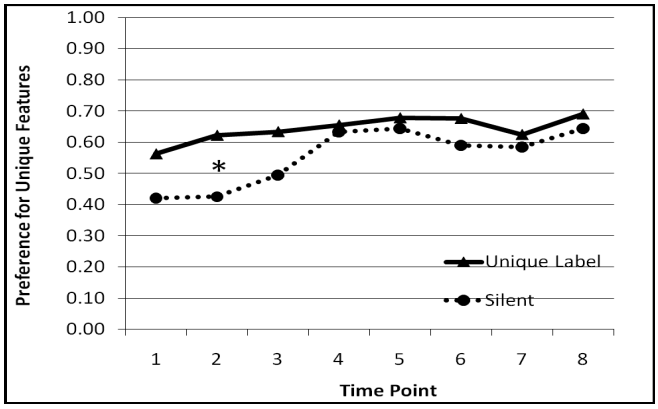


Figure 6: Infants' mean preference for unique features by time point. (Note: Time points represent moving averages).

To obtain a better understanding of the dynamics of attention, unique preference scores were averaged across trials and plotted as a function of time. As can be seen in Figure 7, preference for the unique features was greater in the unique label condition than the silent condition within 1000 ms to 4000 ms. Although, the effect of unique labels was less pronounced in infants than adults, these data provide some evidence for unique labels facilitating infants' attention to unique features.

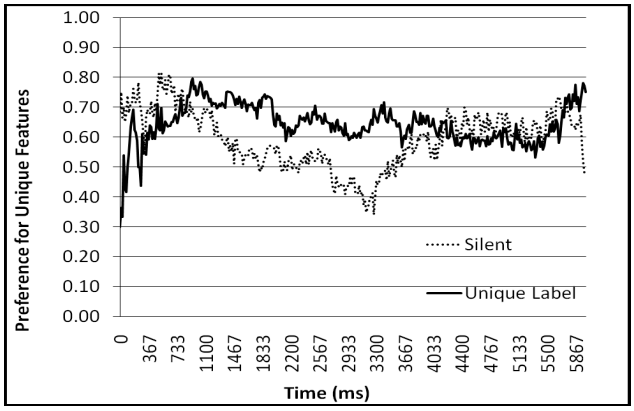


Figure 7: Infants' mean preference for unique features over time.

Effect of Common Labels To determine if common labels directed infants' attention to common features, we compared preference for common features in the common label condition to preference for common features in the silent condition, using the same sample of infants that was previously compared to the unique label condition. As in the unique label condition, gaze data were analyzed using a moving average. A repeated-measures ANOVA was conducted on mean preference for common features with Condition (common label vs. silent) as a between-subjects factor and Time Point (1 vs. 2 vs. 3 vs. 4 vs. 5 vs. 6 vs. 7 vs. 8) as a within-subjects factor. Results revealed a significant main effect of Time Point, $F(7, 147) = 9.06$, $p < .001$. Like adults, infants' preference for common features attenuated over time for both the common label and silent conditions. However, as shown in Figure 8, mean preference scores for common features were not significantly different between conditions at any point in the course of familiarization.

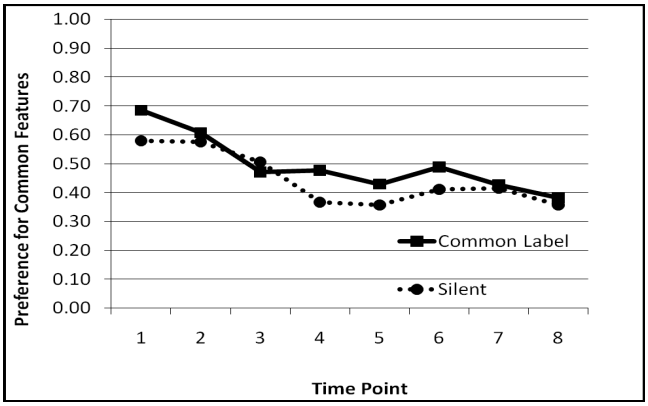


Figure 8: Infants' mean preference for common features by time point. (Note: Time points represent moving averages).

Common preference scores were averaged across trials and plotted as a function of time (see Figure 9). Although results from the ANOVA and *t*-tests revealed no differences between conditions, preference for the common features in the common label condition exceeded the silent condition for the first and last 1000 ms of the trials. Although not illustrated by adults, this pattern of results revealed that if common labels directed infants' attention to common features, the effects were subtle.

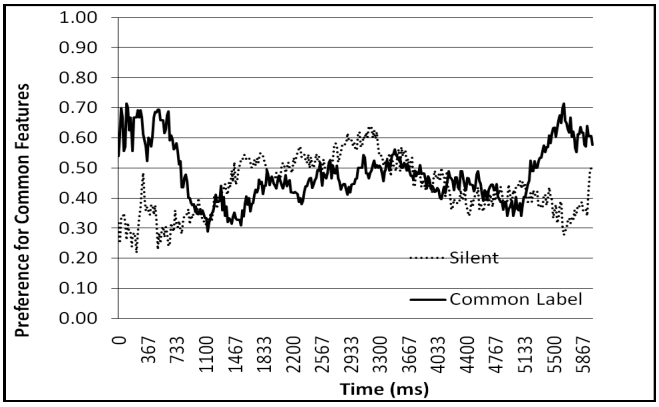


Figure 9: Infants' mean preference for unique features over time.

Summary Experiment 2 found comparable, yet less pronounced results as Experiment 1 with regard to unique labels affecting visual attention. Infants presented with

varying labels (*i.e.*, unique) compared to silence disproportionately distributed their attention to objects' unique versus common features. Effects of common words were less robust, and if they directed infants' attention to common features, these effects are subtle.

Conclusion

The current study reveals several important findings. First, adults, and to a lesser extent, infants, who heard unique labels accumulated more looking to unique features compared to the silent condition. Second, for adults, this effect was robust across familiarization and occurred throughout the entire trial. Third, there was no clear evidence of common labels directing attention to common features for adults or for infants.

Many studies have examined how different types of auditory input affect categorization, as assessed by increased looking to novel categories in a subsequent testing phase (Balaban & Waxman, 1997; Fulkerson & Waxman, 2007; Plunkett, Hu & Cohen, 2008; Robinson & Sloutsky, 2007). The current study, in conjunction with research by Althaus and Mareschal (2010), are the first studies we are aware of that have directly tested the hypothesis that words draw attention to category relevant features for infants. Directly testing this hypothesis (*i.e*., as opposed to inferring it from infants' looking to the novel category at test) is crucial for understanding possible mechanisms underlying effects of words on category learning.

The findings of the current study are partially consistent with both proposed mechanisms. First, in support of the claim that words direct attention to category relevant information (e.g., Waxman, 2003), there was clear evidence for adults, and to a lesser extent, infants, that unique labels highlight unique features. However, there was little support for the claim that common words highlight commonalties, which may have stemmed from a general tendency to increase looking to novel features across familiarization.

Support for the claim that auditory information can attenuate visual processing (e.g., Robinson & Sloutsky, 2007) is also mixed. Support for this claim primarily comes from the finding that infants in the label conditions did not show better discrimination at test, and there was no robust facilitation across familiarization. However, this account assumes that differential effects of words and sounds stem from sounds attenuating visual processing more than words. A direct test of this account would require a non-linguistic sound condition. At the same time, there was little evidence that words slowed down visual processing. However, studies showing that words interfered with visual processing tested 8- and 12-month-old infants (Robinson & Sloutsky, 2007; Sloutsky & Robinson, 2008), which is younger than the infants tested in the current study.

The current study raises an interesting question. Why was the effect of common labels weaker than the effect of unique labels? One possibility is that adults were told to pay attention to the pictures because they were going to be asked about them later. These instructions, in combination with habituation to the common features may have biased attention to unique features. Future research will need to systematically manipulate the category structure by changing the proportion of common to unique features. It is possible that effects of words may interact with the structure of the to-be-learned category. It will also be important to test categorization abilities to connect performance at test to training data, allowing for a better examination of individual differences in category learning.

Acknowledgments

This research is supported by grants from NSF (BCS-0720135), NIH (R01HD056105), and the US Department of Education (R305H050125 and R305B070407) to V. M. Sloutsky and from NIH (RO3HD055527) to C. W. Robinson.

References

Althaus, N., & Mareschal, D. (2010). Speech facilitates categorization but only novel labels direct infants' attention to commonalities. Poster presented at the biennial meeting of International Conference on Infant Studies, Baltimore, MD.

Balaban, M.T., & Waxman, S.R. (1997). Do words facilitate object categorization in 9-month-old infants? *Journal of Experimental Child Psychology, 64*, 3 – 26.

Bomba P.C., & Siqueland E.R. (1983) The nature and structure of infant form categories. *Journal of Experimental Child Psychology, 35*, 294– 328.

Eimas, P., & Quinn, P. (1994). Studies on the formation of perceptually based basic-level categories in young infants. Child Development, 65, 903-917.

Fulkerson, A.L., & Waxman, S.R. (2007). Words (but not tones) facilitate object categorization: evidence from 6- and 12-month-olds. *Cognition, 105*, 218–228.

Plunkett, K., Hu, J.F., & Cohen, L.B. (2008). Labels can over-ride perceptual categories in early infancy. C*ognition, 106, 665– 681.*

Robinson, C.W., & Sloutsky, V.M. (2004). Auditory Dominance and its change in the course of development. *Child Development, 75 (5), 1387-1401.*

Robinson, C.W., & Sloutsky, V.M. (2007). Linguistic labels and categorization in infancy: do labels facilitate or hinder? *Infancy, 11*, 233–253.

Sloutsky, V.M., & Napolitano, A.C. (2003). Is a picture worth a thousand words? Preference for auditory modality in young children. *Child Development, 74*, 822–833.

Sloutsky, V.M., & Robinson, C.W. (2008). The role of words and sounds in visual processing: from overshadowing to attentional tuning. *Cognitive Science, 32*, 342–365.

Waxman, S. R. (2003). Links between object categorization and naming: Origins and emergence in human infants. In D. H. Rakison & L. M. Oakes (Eds.), *Early category and concept development: Making sense of the blooming, buzzing confusion* (pp. 213–241). London: Oxford University Press.

Xu, F. (2002). The role of language in acquiring object kind concepts in infancy. *Cognition, 85*, 223–250.

Mental Representations of Diagrams, Views about Diagrams, and Problem Solving

Emmanuel Manalo (emmanuel.manalo@aoni.waseda.jp)
Center for English Language Education in Science and Engineering (CELESE)
Waseda University, 3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555, Japan

Yuri Uesaka (y.uesaka@nm.hum.titech.ac.jp)
Department of Human System Science, Graduate School of Decision Science and Technology
Tokyo Institute of Technology, Japan

Yoshio Yano (yano@kyokyo-u.ac.jp)
Department of Psychology, Kyoto University of Education
Kyoto 612-8522, Japan

Abstract

This study investigated people's mental representations of diagrams and whether these related to views about diagrams and problem solving performance. The participants were 93 undergraduate students who were asked to complete a questionnaire which included free writing on the topic of diagrams, and problem solving. Analysis of the statements and ideas that the students wrote revealed four categories through which diagrams may be mentally represented: uses/purposes, exemplars, personal opinions, and structure. Personal opinions responses were found to negatively correlate with views about the usefulness of diagrams, and with experiences and confidence in using diagrams. In contrast, responses about the uses/purposes of diagrams positively correlated with confidence in using diagrams. Evidence was also found suggesting that, among students studying math, greater knowledge about the uses/purposes of diagrams facilitated better problem solving performance.

Keywords: Mental representation of diagrams; problem solving; articulation; free writing.

Introduction

Diagrams have many different fields of application (see, e.g., Blackwell & Engelhardt's, 2002, list of academic fields that they identified as having research interest in diagrams), and their use is generally considered as efficacious. In problem solving, for example, Larkin and Simon (1987) explained how diagrammatic representations have distinct advantages over sentential representations because the ways in which diagrams index information can more effectively support useful and efficient computational processes. Hembree (1992) also found that, among the instruction methods he examined in a meta-analysis, training in diagram drawing provided the largest performance improvement in problem solving.

Despite the many reported positive attributes of diagram use, there are numerous problems that have been identified in relation to that use. For example, prior knowledge about diagrams appears necessary for their effective use (see, e.g., Grawemeyer & Cox, 2008; Larkin & Simon, 1987), and student have generally been found to lack spontaneity in using diagrams (see, e.g., Dufour-Janvier, Bednarz, & Belanger, 1987; Uesaka, Manalo, & Ichikawa, 2007). In essence, these suggest that many individuals probably fail to benefit from diagram use: if they lack sufficient knowledge about how to effectively use them, and – even if they did know how to use diagrams – if they nevertheless neglect to make use of them.

Most of the published research on diagrams have focused on their effects and functions (e.g., Ainsworth & Th Loizou, 2003; Cheng, 2002, 2004; Mayer, 2003), with very few studies that have investigated possible ways of understanding and addressing the problems associated with users noted above. The few studies that have considered issues concerning users of diagrams include Uesaka et al. (2007) which found that lack of confidence and perceptions of difficulty in diagram use, and viewing diagrams more as a strategy that teachers use (rather than a strategy that they themselves can use), were deterrents to students' spontaneous use of diagrams. Uesaka et al.'s findings indicate that how individuals view diagrams influence their use of diagrams – suggesting that understanding the ways in which diagrams can be mentally represented could be key to addressing issues/problems about their use.

There is not a lot in the research literature, however, that deals with how people mentally represent diagrams. Numerous studies have considered mental processes relating to graphical representation: for example, Stern, Aprea, and Ebner (2003) examined the effect of "active" versus "passive" graphical representation (i.e., passive encounter with, as opposed to active construction of, linear graphs) on processing transfer from one subject content area to another. However, such studies have not directly addressed the question of how people might structurally represent diagrams in their minds (e.g., as images and/or propositions, in terms of their functions and/or specific examples?).

Blackwell and Engelhardt (2002) proposed a meta-taxonomy that can be used to analyze and compare existing taxonomic systems of diagrams. Their meta-taxonomy was aimed at facilitating the study of diagrammatic representations, such as assessing the relevance of different representations to specific research questions. One of the taxonomic dimensions they proposed was "cognitive" and, although they did not elaborate on this dimension in any detail, their suggestion of focusing on the nature of the

representation and the ways in which people might differ appears appropriate in any attempt to understand the user perspective in diagrams use.

Perhaps the closest attempt at finding out how people mentally represent diagrams was carried out by Cox and Grawemeyer (2003). They used a card sorting task to assess their participants' semantic knowledge about a wide range of diagrams (that they referred to as "external representations" or "ERs"), and found through cluster analysis 9 major categories of ERs. Furthermore they found that participants differed in the categories they created according to how well they scored on ER reasoning tasks: the group of participants who scored well tended to create fewer categories that were based on semantic distinctions between ERs, while the group who scored lower created more categories that tended to focus more on superficial aspects of ERs (e.g., what the ERs looked like).

Cox and Grawemeyer's (2003) findings revealed some important points about how people mentally deal with diagrams: for example, that those who had (presumably) greater knowledge about diagrams were able to perceive meaning-based commonalities between diagrams that may not look alike, while those who had (again, presumably) less knowledge about diagrams may have had to rely on feature-based processes which in turn may have been based on their recollections of diagrams they had previously encountered. Essentially, this suggests that with greater knowledge about diagrams, a person can perceive meaningful relationships between different forms, and categorize accordingly.

It is questionable, however, whether the categories that Cox and Grawemeyer (2003) identified based on their participants' responses reflect the categories that people normally posses as their mental representations of diagrams (i.e., in normal circumstances, not in response to requests to sort/group diagrams). Would people naturally use such categories in mentally organizing and representing what they know about diagrams? The present authors believe otherwise as the categories that people come up with in response to item sorting tasks would inevitably be influenced by their efforts at incorporating and making sense of items that they either did not know about or had not previously considered as part of the subject in consideration. The categories would also reflect the absence of items they may know about but had not been presented. In other words, the kinds of items presented in the task would unavoidably bias the kinds of categories that are produced.

Furthermore, in the Cox and Grawemeyer (2003) study, the participants' ability to find meaning (e.g., the semantic distinctions) in the task given does not necessarily mean that those meanings previously guided their mental representation of diagrams: the task itself could have facilitated the development of their insight about those semantic distinctions between different ERs. This therefore means that, despite the valuable contributions of the Cox and Grawemeyer study, the question of how people normally represent diagrams in their minds remains largely unanswered.

One method that has been used to gain insights into people's cognitive structures about target "objects" is articulation (see, e.g., Scott, 1966): through the descriptions that participants provide of the target object, the structural properties of their cognition can be inferred (i.e., through the definitions, categorizations, connections, and elaborations they articulate). As information provided with the target object can be restricted, potential biases that can inadvertently be communicated to the participant can be reduced. Steps to reduce the potential for such biases are important when attempting to understand how people normally represent certain concepts in their minds.

Free writing, which can be defined as "a procedure in which students are asked to write down whatever they think of and to keep writing without worrying about quality of ideas" (Hayes & Flower, 1986, p. 1106), is one technique that has been used to facilitate participants' articulation of their beliefs and ideas about target concepts. It has, for example, been used to tap into and understand students' knowledge about, and associations with, basic scientific concepts they were learning (Curtis & Millar, 1988); and to understand the specific content of marital ideals among newly married couples (Knobloch-Fedders & Knudson, 2009).

In the present study, free writing was utilized as a method to facilitate participants' articulation of their thoughts and ideas about diagrams – the aim being to explore and try to understand how people might normally represent diagrams in their minds. A further aim of the study was to find out if such mental representations of diagrams are related to participants' (i) views about the importance of diagrams in teaching and learning situations, (ii) self-assessments of experience and confidence in using diagrams, and (iii) competence in using diagrams in problem solving.

Method

Participants

The participants were 93 undergraduate first-year (freshmen) students in a university of education in Japan (i.e., they were studying to become teachers) who voluntarily participated in completing the questionnaire used in this study. Their mean age was 19.0 years (SD = .53 year); 50 were females, and 43 were males.

Materials and Procedure

The questionnaire administered to students was written in Japanese and comprised of three parts. In part 1, after briefly being informed about "free writing", participants were first asked to practice free writing for 1 minute on the topic of "friendship". Following this, they were given 3 minutes to free write on the topic of "diagrams".

In part 2, participants were asked their opinions about diagrams. First, they were asked to indicate on a 5-point Likert-type scale how important they considered diagrams in teaching and learning. They were then asked to briefly write reasons for their response (however, because of space

constraints, analyses concerning the participants' responses to this question have not been included in this paper). Next they were asked to indicate, again on a 5-point Likert-type scale, how much they usually used diagrams, and how confident they felt in using diagrams.

In part 3, participants were asked to solve three problems. Three minutes were allowed for each of the problems, and participants were explicitly asked to try to construct and use diagrams in their attempts to solve them. The first problem required the comparison of heights, and a pictorial depiction of the heights indicated would have been helpful towards solving it. The second problem required working out the circumference following the arrangement of similar-sized pieces of paper; for this, the construction of a table would have been helpful. The third problem concerned applicant placement at a training and employment agency, and for this problem a decision flow chart would have been helpful. (Again, due to space constraints, analyses relating to appropriateness and quality of the diagrams that participants produced have not been included in this paper.)

Results

Categories that Emerged from the Free Writing Task

The participants' responses to the free writing task about diagrams were analyzed initially by looking through these responses and identifying themes or categories of ideas that they conveyed. Understandably, because it was a free writing task and participants were asked to write continuously for the 3-minute duration irrespective of the relevance of the ideas that came to their minds, a large proportion of what they produced appeared unrelated to the topic of diagrams (e.g., single word statements like "compass" and "PC", phrases like "book that has a catchphrase of 'easy to understand'", and sentences like "It keeps appropriate distance."). Apart from these unrelated statements, however, the participants' responses appeared to fall into four broad categories: statements or ideas concerning (a) the uses of diagrams, (b) specific examples of diagrams, (c) personal opinions about diagrams, and (d) the structure of diagrams.

These categories were therefore used to sort and tabulate the participants' responses. The responses were sorted in terms of single, complete 'units of ideas': these could be single words that conveyed a complete idea and appeared intended by the participant to be so (e.g., by being separated from other ideas spatially or through the use of punctuations), complete phrases, sentences, and – in a few cases – diagrams that participants drew.

Table 1 shows the five categories (including the "unrelated" category), the number of ideas or statements that participants wrote in these categories, and the number and percentage of participants who wrote ideas or statements that belonged in these categories.

Under the category of "Uses or purposes" were included statements or ideas that pertained to general or specific uses, purposes, or functions of diagrams. Examples of general statements/ideas of this kind were: "it can help to summarize" and "promotes understanding". Examples of more specific references to uses, purposes, or functions included: "diagrams are used to represent problems more concretely, as a result people can visualize better and find a hint for finding the solution more easily" and "although mathematics is separate from daily life, many people can reach common understanding by using diagrams".

Table 1: Responses to the free writing task.

Category	No. of ideas/ statements	No./percentage of participants
Uses or purposes	178	69 (74%)
Specific examples	163	52 (56%)
Personal opinions	80	46 (49%)
Structure	17	12 (13%)
Unrelated	470	81 (87%)

The category of "Specific examples" was used when participants simply listed, described, or drew specific kinds or forms of diagrams. Examples included: "graph", "bar chart", "pie chart", and "table".

Included in the "Personal opinions" category were participants' ideas or statements that pertained or related to experiences they have had with diagrams. Examples of statements and ideas placed in this category were: "troublesome … complicated", "many are difficult to understand", and "I get irritated when I cannot draw them well". Almost all were negative.

Under the "Structure" category were included participants' references to the general or specific ways in which diagrams structure, organize, or present data/information. Examples of the general ways participants mentioned included: "a diagram is a visual representation", and "a different approach from one that uses language". An example of a more specific reference to the way in which diagrams structure data/information included: "something represented line-by-line".

To check the reliability of coding the participants' responses under these categories, another person independently carried out coding on 10% of the participants' responses (10% being the minimum acceptable subsample size recommended by Lombard, Snyder-Duch, & Campanella Bracken, 2008, for such purposes). The inter-rater agreement (Cohen's kappa coefficient) was found to be .63, which was considered as being substantially concordant.

Relationship of Categories to Views About Diagrams

As previously noted, in part 2 of the questionnaire, Question 1 ("Importance") asked participants how important they considered diagrams in teaching and learning situations, Question 2 ("Experience") asked how much experience they had had in using diagrams, and Question 3 ("Confidence")

asked how confident they felt in using diagrams. The participants were asked to respond on 5-point Likert-type scales where 1 was most negative (e.g., not important) and 5 was most positive (e.g., very important).

For the "Importance" question, the mean response was 4.38 (*SD* = .59), indicating that the participants generally viewed diagrams as being very important in teaching and learning situations. For the "Experience" question, the mean response was 3.44 (*SD* = 1.04), indicating that most of the participants had experiences of occasionally using diagrams. And for the "Confidence" question, the mean response was 2.59 (*SD* = .84), indicating that the participants were generally tending toward being doubtful about their ability to use diagrams well.

To find out whether there were any possible relationships between (i) the categories of statements and ideas that participants produced in free writing about diagrams and (ii) their views about diagrams as gauged in part 2 of the questionnaire, correlations between these were examined.

The correlations between the participants' use or otherwise of the categories, and their responses to the Likert-type scales used in part 2 of the questionnaire, are shown in Table 2.

Table 2: Correlations between participants' use of the categories and their opinions about diagrams.

Category Used	Importance	Experience	Confid.
Uses or purposes	.043	.109	.212*
Specific examples	.127	.107	.020
Personal opinions	– .195	– .193	– .160
Structure	– .138	.022	– .004
Unrelated	– .081	.009	– .112

* $p < .05$

The significant correlation found here suggests that participants who wrote statements/ideas about the uses of diagrams also indicated greater confidence in being able to use diagrams.

The correlations between the number of ideas/statements that participants wrote under each of the categories, and their responses to the items in part 2 of the questionnaire, are shown in Table 3.

Table 3: Correlations between number of ideas in each of the categories and opinions about diagrams.

Category	Importance	Experience	Confid.
Uses or purposes	.003	.011	.229*
Specific examples	.118	.095	.021
Personal opinions	– .233*	– .255*	– .212*
Structure	– .086	.068	.050
Unrelated	.051	.064	– .009

* $p < .05$

The significant correlation here between "Uses or purposes" and "Confidence" suggests that participants who wrote more ideas/statements about the uses of diagrams also possessed greater confidence in their ability to use diagrams. In contrast, the significant negative correlations between "Personal opinions" and "Importance", Experience", and "Confidence" suggest that those who wrote more about their personal opinions about diagrams tended to have lesser appreciation of the value of diagrams in teaching and learning situations, and less experience and lower confidence in diagrams use.

Relationship of Categories to Problem Solving Performance

The mean score for the three problems administered to participants in part 3 of the questionnaire was 2.32 (SD = .80). Seventy-three percent correctly solved Problem 1 (heights comparison), 76% correctly solved Problem 2 (paper arrangement circumference), and 83% correctly solved Problem 3 (employment agency placement). In general therefore, the problems appeared quite easy for most of the participants to solve and ceiling effects may have been encountered.

No significant and/or meaningful correlations were found between (i) the categories of statements and ideas that participants produced in free writing about diagrams (both whether or not they used the categories, and the number of statements they made in each of the categories) and (ii) the scores they obtained in their attempts to solve the problems given in part 3 of the questionnaire.

However, when participants were differentiated on the basis of their "math involvement" (i.e., whether or not they were in a math course, or were seeking a math teacher's license), a significant correlation was found between math involvement and problem solving performance ($r = .28$, $p < .05$). This suggested the possibility that the relationship between math involvement and problem solving performance was mediated by participants' knowledge about the uses/purposes of diagrams; thus, a mediation effect analysis (Baron & Kenny, 1986) was undertaken. This revealed that when regression analysis on problem solving performance was carried out only with the math involvement variable, the standardized coefficient was significant ($\beta = .208$, $p < .05$). However, when the same analysis was done with math involvement and participants' use (or otherwise) of the uses/purposes category as independent variables, the standardized coefficient of math involvement diminished and became non-significant ($\beta = .198$, n.s.). This finding suggests that the better problem solving performance of participants with math involvement was likely due to their greater knowledge about the uses/purposes of diagrams.

Discussion

How Do People Mentally Represent Diagrams?

Through the statements and ideas that participants in the present study articulated via the free writing task, it can be inferred that they viewed or mentally represented diagrams

in terms of their uses, specific examples of them, personal opinions/experiences of them, and their structure. Almost three-quarters of the participants (74%) wrote something about the uses, purposes or functions of diagrams. This high proportion is probably understandable considering that diagrams are tools or strategies that can serve particular purposes: thus the majority of people are likely to mentally represent them in terms of, or in relation to, those purposes they are aware of.

Just over half of the participants (56%) also provided specific examples of diagrams, either on their own or in relation to other categories of statements/ideas they articulated (i.e., uses, personal opinions, structure). This suggests that many people also mentally represent diagrams in terms of, or in relation to, specific kinds of diagrams they know – or exemplars.

Almost half of the participants (49%) wrote statements and ideas that appeared to fall into a category of being about their personal experiences and opinions about diagrams. Again, it probably makes sense that many people would do this when one considers that people make sense of the world around them through their personal experiences and the resulting opinions that they form. Thus, where tools/strategies are concerned, these are likely to be represented in terms of notions like "helpful" or "difficult to use" depending on their experiences of using these.

Finally, a small proportion of participants (13%) referred to the structure of diagrams, suggesting that such structure formed at least part of their mental representation of diagrams. However it should be noted here that the statements/ideas that participants wrote in relation to structure were fairly general and superficial – mostly just expressing that diagrams represent information in visual or pictorial format. They did not refer to more complex and specific structural qualities/portrayals of diagrams like arrays, sequences, notations, and so on.

Only the four categories of uses/purposes, specific examples, personal opinions, and structure were identified in the written data collected in the present study. However, it is possible that other groups of participants would evidence other categories (different from the four identified here) depending on their knowledge about and experiences in the use of diagrams. It would be important to investigate this in future research.

In the present study, only the participants' responses in terms of their use of the categories identified, and the number of distinct statements/ideas they wrote that belonged to each of those categories, were coded, analyzed, and reported. However, there are a number of other dimensions of the data that, at the time of writing this paper, the authors had not yet examined. These dimensions include the 'quality' of the statements and ideas that participants articulated: for example, participants wrote both fairly superficial as well as more meaningful uses/purposes of diagrams that were not differentiated. Also, most of the statements that participants wrote relating to their personal opinions about diagrams were "negative"; very few could be considered "positive". Another potentially important dimension is the connectedness of the statements and ideas – both within and between categories. Although outside the scope of the present paper, it would clearly be useful to examine in more detail the possible effects or relationships that may stem from these other dimensions of the data.

Contributions to Cognitive Research

As noted in the introduction to this paper, no prior research had looked into how people might naturally represent diagrams in their minds. The present findings suggest that such mental representations involve categories of uses or purposes of diagrams for the majority of people. Furthermore, approximately 50% of people would have mental representations that incorporate their personal opinions about diagrams and/or specific examples or exemplars of diagrams. A small minority may also have mental representations relating to the structure of diagrams.

Although at first glance the mental representations suggested by these findings may appear completely different from those identified by Cox and Grawemeyer (2003) through their card sorting task, there are possible connections and congruence between these representations. Firstly, the 9 categories identified by Cox and Grawemeyer pertained to the structure of diagrams – both semantic and superficial. Although only a small proportion, some of the participants in the present study did articulate structure-related statements and ideas about diagrams. Categorizing diagrams according to their structure may be a natural response in a task like the one used in the Cox and Grawemeyer study (where structure may appear as the most salient feature of the stimuli presented). However, diagrammatic structure may also be a natural way of mentally representing diagrams for some people: perhaps for those with limited knowledge/experience about diagrams, superficial structures may be the only salient basis for mental representation. Likewise, for those who have greater than average knowledge/experience about diagrams, the semantic structures of diagrams may in fact be a natural way of mentally representing and organizing diagrams.

Secondly, in the same way that the Cox and Grawemeyer (2003) study identified semantic and superficial distinctions in participants' responses according to their possession of greater or lesser knowledge about diagrams, it is possible that the same semantic-superficial dimension underpins the participants' responses across the different categories identified in the present study. Thus, for example, the participants' responses in the personal opinions category may well differentiate those with greater from those with lesser knowledge and skills about diagrams according to whether the opinions expressed are superficial in nature (e.g., basic references to ease or difficulty) or more meaningful (e.g., pertaining to what they have learnt about themselves or about diagrams). One possibility is that the mental representation of diagrams lies along two dimensions – one dimension being the kinds of categories identified in the present study, and the other being

meaningfulness-superficiality. Future research will need to examine this, and whether other strategies/tools may also be represented mentally in a similar manner.

Contributions to Educational Research

There is evidence in the findings of the present study to suggest that mental representations of diagrams could influence students' views about diagrams as well as their problem solving performance. That responses in the personal opinions category negatively correlated with participants' views about the usefulness of diagrams, and their experiences and confidence in using diagrams, is understandable in light of the fact that the majority of statements/ideas written in the personal opinions category were negative. Many of the participants' more positive personal opinions about diagrams were probably expressed as statements/ideas about their uses – thus falling into the uses/purposes category instead.

The finding that responses in the uses/purposes category not only correlated with confidence in using diagrams, but also appeared to mediate the problem solving performance of those studying math, is likely due to two simple explanations (cf. Uesaka et al., 2007). First, greater knowledge about the uses/purposes of a strategy/tool should promote greater confidence in the use of that strategy/tool. Second, greater knowledge about the uses/purposes of diagrams should enable more appropriate use of them in problem solving situations, which in turn should assist toward better problem solving performance. Further research into the mechanisms of these relationships would be helpful toward the development of their applications in classroom instruction.

Acknowledgments

The authors would like to thank Muneyuki Mizutani and Takahiro Komatsu for their assistance in the collection of data for this study.

References

Ainsworth, S., & Th Loizou, A. (2003). The effects of self-explaining when learning with text or diagrams. *Cognitive Science, 27,* 669–681.

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic and statistical considerations. *Journal of Personality and Social Psychology, 51,* 1173–1182.

Blackwell, A., & Engelhardt, Y. (2002). A meta-taxonomy for diagram research. In M. Anderson, B. Meyer, & P. Olivier (Eds.), *Diagrammatic representation and reasoning* (pp. 47–64). London: Springer-Verlag.

Cheng, P. C. H. (2002). Electrifying diagrams for learning: principles for complex representational systems. *Cognitive Science, 26,* 685–736.

Cheng, P. C. H. (2004). Why diagrams are (sometimes) six times easier than words: Benefit beyond locational indexing. In A. Blackwell, K. Marriott, & A. Shimojima (Eds.), *Diagrammatic representation and inference, third international conference, diagrams 2004, LNAI 2980* (pp. 242–254). Heidelberg: Springer.

Cox, R., & Grawemeyer, B. (2003). The mental organisation of external representations. *Proceedings of the European Cognitive Science Conference (EuroCogSci - joint Cognitive Science Society and German Cognitive Science Society conference), Osnabrück, September, 2003.* Available from: http://www.cs.bath.ac.uk/~bg230/Cox&GrawemeyerEuroCogsci03.pdf

Curtis, S., & Millar, R. (1988). Language and conceptual understanding in science: A comparison of English and Asian language speaking children. *Research in Science & Technological Education, 6,* 61–77.

Dufour-Janvier, B., Bednarz, N., & Belanger, M. (1987). Pedagogical considerations concerning the problem of representation. In C. Janvier (Ed.), *Problems of representation in the teaching and learning of mathematics* (pp. 110-120). Hillsdale, NJ: Erlbaum.

Grawemeyer, B., & Cox, R. (2008). The effects of users' background diagram knowledge and task characteristics upon information display selection. In G. Stapleton, J. Howse, & J. Lee (Eds.), *Diagrams 2008 (Lecture Notes in Artificial Intelligence 5223)* (pp. 321–334). Berlin Heidelberg, Germany: Springer-Verlag.

Hayes, J. R., & Flower, L. S. (1986). Writing research and the writer. *American Psychologist, 41,* 1106–1113.

Hembree, R. (1992). Experiments and relational studies in problem-solving: A meta-analysis. *Journal for Research in Mathematics Education, 23,* 242–273.

Knobloch-Fedders, L. M., & Knudson, R. M. (2009). Marital ideals of the newly-married: A longitudinal analysis. *Journal of Social and Personal Relationships, 26,* 249–271.

Larkin, J. H., & Simon, H. A. (1987). Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science, 11,* 65–99.

Lombard, M., Snyder-Duch, J., & Campanella Bracken, C. (2008). Practical resources for assessing and reporting intercoder reliability in content analysis research projects. Retrieved May 13, 2010, from: http://astro.temple.edu/~lombard/reliability/

Mayer, R. E. (2003). The promise of multimedia learning: Using the same instructional design methods across different media. *Learning and Instruction, 13,* 125–139.

Scott, W. A. (1966). Brief report: Measures of cognitive structure. *Multivariate Behavioral Research, 1,* 391–395.

Stern, E., Aprea, C., & Ebner, H. G. (2003). Improving cross-content transfer in text processing by means of active graphical representation. *Learning and Instruction, 13,* 191–203.

Uesaka, Y., Manalo, E., & Ichikawa, S. (2007). What kinds of perceptions and daily learning behaviors promote students' use of diagrams in mathematics problem solving? *Learning and Instruction, 17,* 322–335.

Metacognitive Judgments of Improvement are Uncorrelated with Learning Rate

Corinne L. Townsend (ctownsend@ucmerced.edu)
Evan Heit (eheit@ucmerced.edu)
Department of Social and Cognitive Sciences, 5200 North Lake Road
Merced, CA 95343 USA

Abstract

Being able to assess one's own learning rate is essential for optimal learning. Can students accurately assess their learning rate, and is the timing of judgments of improvement important? In this experiment, students were to estimate their learning rate on each trial, either before the trial, or immediately after. If students typically make these judgments before embarking on further study, accuracy might be greater in the predictive judgment condition. No evidence was found that students could accurately judge improvement, in either condition. Implications for models of self regulated learning are discussed in light of these findings.

Keywords: metacognition; self regulated learning; metamemory.

Introduction

Judgments of improvement are metacognitive judgments regarding one's speed of learning. These can be thought of as a student's estimation of how quickly he or she is acquiring more knowledge, or put into practical terms, how useful a given amount of study time is likely to be. These judgments are crucial, as the ability to estimate one's learning rate will affect how well students are able to allocate their time optimally during self regulated learning, which then in turn will influence academic achievement. One such example of how these judgments might inform study time allocation is in the proximal learning model (Kornell & Metcalfe, 2006; Metcalfe, 2002; Metcalfe & Kornell, 2005). In this model, it is proposed that decisions seek to maximize the rate of return per time studied, and those regarding when to switch topics or stop studying may rely on judgments of improvement. This way, students can avoid working in vain while not making progress, and instead move on to more fruitful pursuits. The proximal learning model contrasts with an earlier account, the discrepancy reduction model (Thiede & Dunlosky, 1999), which assumed that students focused on the most difficult items first, and stopped when material reached a satisfactorily high level of learning, thus depending on JOL level to determine stopping times. Additionally, Son and Sethi (2006, in press) have derived mathematically that the most optimal behavior is usually to focus on the items with the highest current rate of return, consistent with the proximal learning model. There is some evidence to support this account, which is sometimes referred to as the shift-to-easier-materials effect; this is the finding that under time pressure, students prioritize by studying the easiest (high rate of return) items first, before moving on to more difficult material (Metcalfe & Kornell, 2003; Dunlosky & Thiede, 2004; Kornell & Metcalfe, 2006).

However, there is not yet evidence to support the idea of using improvement rates to inform decisions, and current research has not shown that students have the ability to make judgments of improvement accurately in any sense. In our previous work (Townsend & Heit, 2010), participants estimated their amount of improvement after completing each study trial, in a repeated series of study trials for a set of verbal materials. Students' judgments of improvement (or JOIs) were not significantly correlated with actual improvement rates, and in some cases were even negatively correlated. The negative correlation occurred when judgments of learning and judgments of improvements were made using different rating scales, which prevented participants from attempting to infer their JOIs from their judgments of learning. Work by Kornell and Bjork (2009) has also shown that students have difficulties estimating how much they will learn during one or more study trials, dramatically underestimating the usefulness of study. They referred to this type of judgment as a prediction of learning, but the concept is the same. Thus, there is reason to be concerned that students are not able to make the metacognitive judgments that would lead to optimal learning.

Students' post-study JOIs showed an interesting shift from underconfidence to overconfidence over the course of learning (Townsend & Heit, 2010), but predictive JOIs that estimate the fruitfulness of further study may or may not show the same pattern. It is important to assess predictions of future learning (predictive, pre-study trial JOI) rather than just a postdictive assessment of learning during a study trial, as decisions regarding study time allocation may depend more on how much is expected to gain from further study, rather than how much was gained from recent study. This experiment was designed to compare the two conditions to evaluate how (or whether) timing affects JOIs. For comparison, we also collected judgments of learning (JOLs, which are predictions of recall test performance) from an additional group of participants, to compare the relative accuracy of JOIs and JOLs. For example, whereas it may be too difficult for students to judge their level of improvement, they still may be able to judge their level of learning in absolute terms.

Experiment

In this experiment, we compared two different rating scales (percentage vs. absolute number of words), as well as different types of improvement judgments. One might expect that judgments in terms of number of words learned would be easier and more successful, due to their simplicity as well as their close nature to other judgments of optimal foraging (Gigerenzer & Hoffrage, 1995). Judgment types were either postdictive (made after a study trial) or predictive, occurring before the next study trial, i.e. "if you were to study this list for another minute, how much do you think you would improve?" "Answer: I think I would learn another ___% of the material". Predictive JOIs may be more informative than postdictive JOIs for study decisions, and if students do make predictive JOIs (and not postdictive) they should have better accuracy for this kind of judgment. It may be more likely that students would make predictive JOIs, especially if they are determining whether or not further study would be worthwhile. Type of judgment (Predictive JOI, Postdictive JOI, or JOL) and type of scale (percent or number of words) were both manipulated between subjects.

Method

Participants. 171 students from the subject pool at the University of California, Merced, volunteered to participate for class credit. The number of participants in each condition was as follows: 32 making prospective, percent scale JOIs, 31 making prospective, numerical JOIs, 34 making postdictive percent JOIs, 30 making postdictive numerical JOIs, 23 making percent scale JOLs, and 21 making numerical JOLs.

Materials. A list of 50 Swahili – English word pairs was constructed from the Nelson and Dunlosky (1994) norms. These stimuli have been used in much previous metacognitive research. The list of word pairs was constructed to include a range of difficulty.

Design and Procedure. The experiment consisted of six trials, with each trial consisting of a study phase, judgment phase, and test phase. All manipulations were between subjects. The design was 3 judgment types (predictive JOI, postdictive JOI, or JOL) by 2 scales (absolute number or percentage), so each subject only experienced one judgment type and one scale type for a total of 6 different conditions.
For the prospective JOI conditions, each trial consisted of judgment – study – test (with the exception of the first trial, which did not include a judgment). Judgments were solicited with the question "if you were to study this list for another minute, how much do you think you would improve? Answer: I think I would learn another ___[% or words] of the material."

For the postdictive JOI conditions, each trial consisted of study – judgment – test (with the first trial not including a judgment). These judgments were made after the question "Compared to the previous trial, what percent more of the list will you be able to recall? Answer: I will recall another ___ % of the list" OR "Compared to the previous trial, how many more words of the list will you be able to recall? Answer: I will recall another ___ words of the list".
The JOL conditions consisted of study – judgment – test. Participants were asked "What percent of the list will you be able to recall? Answer: I will recall __ % of the list" OR "How many words of the list will you be able to recall? Answer: I will recall ___ words of the list".

Scoring. Responses on the test trial were marked correct if they matched the target word. No points were deducted for misspellings. Percentage judgments were converted to number of words for the purpose of analysis.

Results

Preliminary analyses revealed that some participants were not successful in learning Swahili-English word pairs. On this basis, 37 participants were removed from analyses due to either not entering any judgments, responding with the same judgment on each trial, not learning more than 5 words after all 6 trials, or technical errors. There were a total of 25 participants in the predictive JOI – percent judgment condition, 23 in the predictive JOI – numerical judgment condition, 25 in the postdictive JOI - percent rating and 25 in the postdictive numerical rating condition. Finally, 15 participants gave percentage JOL judgments, and 20 gave numerical JOL judgments.

Judgments of Learning. Judgments of learning were compared to recall performance, and significant correlations were found for both percentage (ρ = .61, *min* = -.58, *max* = 1.0, *SD* =.56, t (15) = 4.38, $p < .001$) and number rating conditions (ρ = .42, *min* = -.88, *max* = 1.0, *SD* =.68, t (18) = 2.67, $p < .015$). There was no significant difference between the two conditions, t (33) =.36, p = .72.

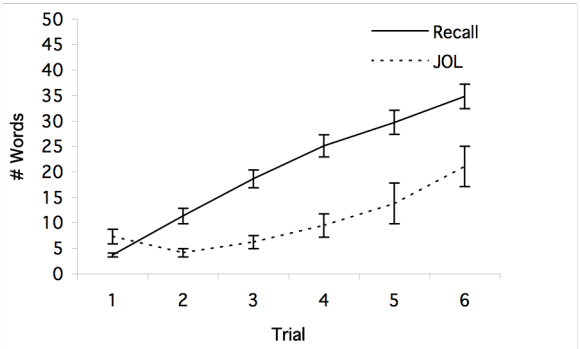


Figure 1: Mean JOL values and recall per trial, percent scale converted to number of words.

Confidence Bias. Relative accuracy of JOLs is not particularly informative, since it is reasonable to assume that participants understand that performance generally increases with each trial. For this reason, we examined absolute accuracy of these judgments as well. Absolute accuracy (in terms of bias) was assessed for JOLs by computing the difference between JOLs and actual recall. For percentage judgments, the percentage was converted to number of words. Biases were also analyzed to see if they differed for judgment type. There was a trend toward more underconfidence for percentage judgments, F (1, 32) = 3.86, MSE = 111.22, p = .058, $\eta 2$ = .108. There was a significant effect of trial, F (5, 160) = 61.33, MSE = 44.76, p < .001, $\eta 2$ = .657. Similarly to previous work that included both JOLs and JOIs (Townsend & Heit, 2010), there appeared to be increasing underconfidence with practice, but with a small upturn on the last trials, as seen in Figures 1 (percentage scale judgments) and 2 (numerical scale judgments).

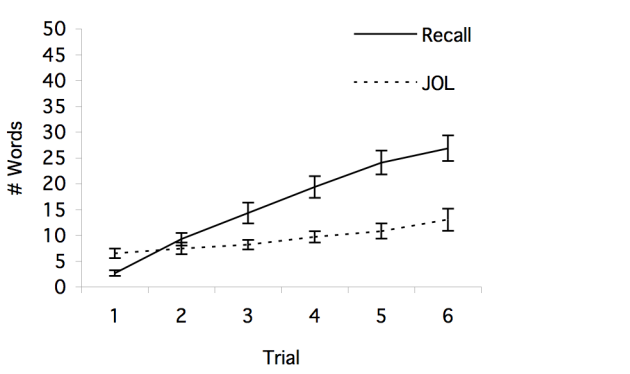


Figure 2. Mean JOL values and recall per trial, numerical scale

Judgments of Improvement. Judgments of improvement were compared with actual improvement, with no significant correlation found for either judgment type or either scale type. For predictive JOIs, neither percentage (average ρ = .11, min = -.89, max = .95, SD = .50) nor numerical judgments (average ρ = .06, min = -.98, max = 1.0, SD = .52) were significantly different from zero; for postdictive JOIs, percentage (average ρ = .05, min = -.89, max = .95, SD = .51) and numerical (average ρ = .04, min = -.89, max = .89, SD = .52) judgments were also non-significant.

Changes in JOLs are a possible basis of judgments of improvement. In this experiment, JOIs and JOLs were made between subjects to avoid influencing participants towards inferring JOIs this way. A between subjects repeated measures analysis of vaiance comparing mean JOIs and mean JOL difference scores by trial suggests that participants may not have been covertly making JOLs and using them to infer JOIs; F (1,125) = 13.302, p < .001, $\eta 2$ = .096.

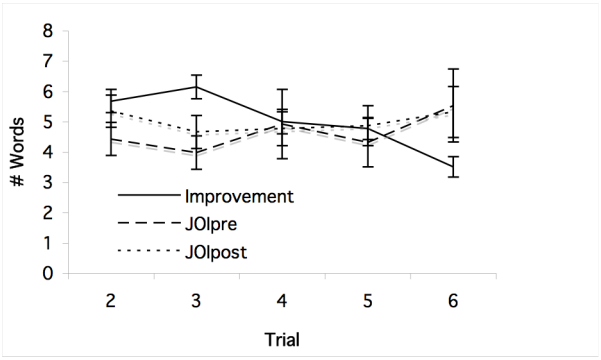


Figure 3. Average JOIs and improvement values per trial, by judgment time.

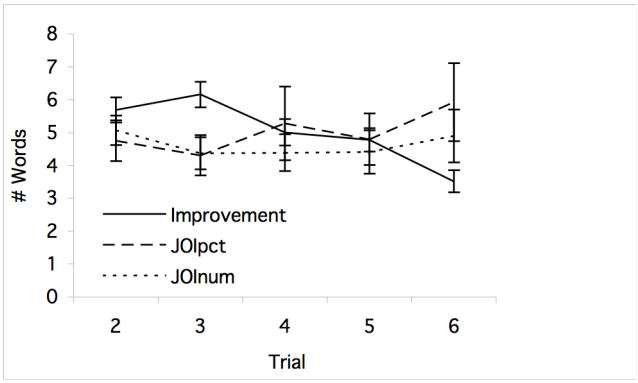


Figure 4. Average JOIs and improvement values per trial, by scale type.

JOI Bias. Absolute accuracy for JOIs was examined, and no significant differences in bias were found for judgment type or scale type, though there was a significant effect of trial, F (3.05, 259.51) = 9.13, MSE = 25.34, p < .001, $\eta 2$ = .097. Percentage judgments were converted into number of words for the purpose of comparison. There appeared to be increasing confidence with trial, as illustrated in Figures 3 and 4 which corroborates with the results from previous work (Townsend & Heit, 2010) which found that JOIs increased with trial, and in that case, were correlated with JOLs. Average total bias across participants was -.2872, min = -7.10, max = 20.0, SD = 4.67.

The low values for JOI biases may lead one to conclude that judged improvement was very close to actual improvement, despite the low correlations. This would be an erroneous conclusion, however, because an examination of the absolute accuracy (Schraw, 2009) of JOIs (average squared deviations between JOIs and improvement) shows a large discrepancy. The average value of absolute accuracy across participants was 45.68, min = -78.8, max = 897.0, SD = 115.21. No significant differences in absolute accuracy were found for judgment time, t (96) = -.262, p = .091, or for judgment type, t (96) = -.45, p = .66.

Discussion

In this experiment, we failed to find a significant correlation between JOIs and actual improvement. The type of scale (percentage or number of words) did not make a difference for judgment accuracy, nor did the time of judgment; predictive JOIs were no more accurate than postdictive JOIs. In comparison, JOLs made before and after a test have been found to differ (Hacker, Bol, Horgan, & Rakow, 2000). One possible reason for why JOL values differ between pre and post test is that students may routinely make JOLs, assessing how well they are likely to do on exams, and then make post test judgments of performance, e.g. "I think I aced the exam!", and there are more cues with which to base posttest JOLs on, as compared to pretest JOLs (e.g. once they've taken the exam, they know what the actual questions were, how quickly the answers came to mind, etc). In contrast, JOIs may be a judgment that is not made very often, without as many informative cues, and is a judgment on which students don't generally get feedback; JOLs do get feedback over time, as students are given grades on assignments and exams (and this feedback may also help savvy students to learn what cues are more informative). To get feedback on a JOI, it would be necessary to test oneself before and after a study session, and then calculate how much more information was known compared to pre-study. This is a cumbersome and unlikely task for a student to perform; more likely, students will rely upon subjective feelings, like how much more fluent the information seems, how answers may seem to come to mind faster, and perhaps even reduced feelings of anxiety about exams—and without feedback, students cannot learn whether or not these feelings are actually informative.

Other research that has looked at JOI predictions also found judgments to be uncorrelated with actual learning. In Kornell and Bjork (2009), they found a large degree of underconfidence in predictions of learning. Participants in their experiment made their predictions on the first study trial, so their results might not predict how learners will feel about the fruitfulness of study if asked beyond that point. For example, if asked initially about how much they will learn in four study trials, they may be incredibly unoptimistic, but if the students were to be asked after two study trials, they may have different predictions, perhaps based on their subjective experience of the task becoming easier. Kornell and Bjork (2009) showed that JOIs were inaccurate, observing that students were incredibly underconfident when it came to predicting future learning beyond one study trial, but they only experienced one trial at the time of judgment, and did not yet have the experience of repeating study (which is the very thing they are asked about). In our experiment, students made their judgments on each study trial, and a different pattern emerged: a shift from underconfidence in early trials, which is consistent with their results, to overconfidence in later trials. Unfortunately it would seem that experience with the task does not improve JOI accuracy at all, but rather shows a more interesting pattern of inaccuracy.

The inaccuracy of these judgments of improvements has significant implications for models of study time allocation that rely on them; specifically, it is highly unlikely that student behavior would approximate optimality by the use of JOIs. The inability to accurately assess the speed at which one is learning means that learners could not accurately make JOIs to reliably know if further study would be made in vain, and when time would be better spent on a different item or task, leading to much wasted time. Even worse, if students do make JOIs and base decisions on them, they may make bad decisions. Students may give up early in the process of learning (as JOIs are underconfident in the beginning of study), and instead work on better-learned material, on which they persist longer than they should due to overconfidence in the later periods of study. This would lead to very inefficient studying, and could have disastrous results- yet many students do manage to achieve reasonable performance in their courses, so this cannot be the whole story. It may be the case that stopping and switching is based not on explicit JOIs but is done implicitly; Reder and Schunn (1996) suggest that much of metacognitive monitoring and control may actually be implicit. Supporting this somewhat, Payne, Duggan, & Neth (2007), found that in a task switching situation where people performed two different tasks (scrabble and word search), they were sensitive to rate of rewards, and able to spend more time in the easier task. This possibility of implicit control will be investigated in future research that more closely resembles a learning situation, rather than tasks in which participants have such obvious successes and failures.

Whether they are informed by explicit JOIs or implicit control, decisions may also be based on other factors: subjective feelings such as frustration and fatigue, idiosyncratic rules (e.g. study for X amount of time, or until I fall asleep, or all day before the exam), JOLs, or on the results of self-testing. It would also be adaptive if students do not simply stop studying low JOI material, because no learning would take place, and that is not always a viable option. In the cases where the item has a low JOI and a low JOL (meaning that the item is not well learned, and is not being learned very quickly), the ideal behavior would be to change strategies, seek other sources of learning, or the guidance of the instructor.

We also leave open the possibility that students could be taught better study habits, and to make more accurate JOIs. In the framework of Stanovich (2009), otherwise intelligent students may act suboptimally because they lack the "mindware" that allows them to reflect on their own level of learning, simulate the possible consequences of further studying, and override their default study strategies. We are hopeful that at least some of these abilities are teachable. Future research will examine these possibilities.

References

Dunlosky, J., & Thiede, K. W. (2004). Causes and constraints of the shift-to-easier-materials effect in the control of study. *Memory and Cognition, 32*(5), 779-788.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review, 102*, 684-704.

Hacker, D. J., Bol, L., Horgan, D. D., & Rakow, E. A. (2000). Test prediction and performance in a classroom context. *Journal of Educational Psychology, 92,* 160–170.

Kornell, N., & Bjork, R. A. (2009). A stability bias in human memory: Overestimating remembering and underestimating learning. *Journal of Experimental Psychology: General, 138*(4), 449-468.

Kornell, N., & Metcalfe, J. (2006). Study efficacy and the region of proximal learning framework. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*(3), 609-622.

Metcalfe, J. (2002). Is study time allocated selectively to a region of proximal learning? *Journal of Experimental Psychology: General, 131*(3), 349-363.

Metcalfe, J., & Kornell, N. (2003). The dynamics of learning and allocation of study time to a region of proximal learning. *Journal of Experimental Psychology: General, 132*(4), 530-542.

Metcalfe, J., & Kornell, N. (2005). A region of proximal learning model of study time allocation. *Journal of Memory and Language, 52*(4), 463-477.

Nelson, T. O., & Dunlosky, J. (1994). Norms of paired-associate recall during multitrial learning of Swahili-English translation equivalents. *Memory, 2*, 325-335.

Payne, S. J., Duggan, G. B., & Neth, H. (2007). Discretionary task interleaving: Heuristics for time allocation in cognitive foraging. *Journal of Experimental Psychology: General, 136*(3), 370-388.

Reder, L., & Schunn, C. D. (1996). Metacognition does not imply awareness: Strategy choice is goverened by implicit learning and memory. In L. M. Reder (Ed.), *Implicit memory and metacognition (pp. 45-78)*. Mahwah, NJ: Erlbaum.

Schraw, G. (2009). A conceptual analysis of five measures of metacognitive monitoring. *Metacognition and Learning, 4*(1), 33-45.

Son, L. K., & Sethi, R. (2006). Metacognitive control and optimal learning. *Cognitive Science, 30*(4), 759-774.

Son, L. K., & Sethi, R. (in press). Adaptive learning and the allocation of time. *Adaptive Behavior.*

Stanovich, K.E. (2009). *What intelligence tests miss: The psychology of rational thought.* New Haven: Yale.

Thiede, K. W., & Dunlosky, J. (1999*)*. Toward a general model of self-regulated study: An analysis of selection of items for study and self-paced study time*. Journal of Experimental Psychology: Learning, Memory, and Cognition, 25, 1024–1037.*

Townsend, C. L. & Heit, E. (2010). *Judgments of Learning and Improvement.* Manuscript submitted for publication.

Learning to Explore the World through its Statistics: Infants' Visual Search in the A-not-B task

Hanna Popick (hmuenke@stanford.edu)
Department of Psychology, Stanford University
450 Serra Mall, Jordan Hall
Stanford, CA 94305, USA

Michael Ramscar (ramscar@stanford.edu)
Department of Psychology, Stanford University
450 Serra Mall, Jordan Hall
Stanford, CA 94305, USA

Natasha Kirkham (Natasha.Kirkham@gmail.com)
Department of Psychological Sciences, Birkbeck, University of London
London, Malet Street, WC1E7HX, England

Abstract

In searching for hidden objects, infants younger than 12 months frequently commit "A-not-B errors," in which they successfully search for an object in one location (A) and then fail to search for it when it is conspicuously hidden in a new location (B). Why do they fail to make the switch and perseverate at the first location? Although these errors have often been attributed to cognitive and conceptual limitations, we suggest that the answer is far more basic: in order to search successfully, children must first learn to do so. In what follows, we present an error-driven learning account of "A-not-B" search which suggests that failing to make the switch is an essential part of learning the appropriate searching cues and contextual search strategies. We elaborate the findings of an eye-tracking experiment with 9 month-olds that behaviorally confirms the predictions of our learning model.

Keywords: error-driven learning; search; A-not-B; feedback

It is Monday morning. You haven't seen your car keys since Friday evening. Where do you look for them? In an ideal world, you will go straight to where you last saw them. As an adult, you will have learned from previous experience that keys are not (usually) assigned random locations, and that what best predicts a key's location is the conjunction of a given spot and that spot being the most recent place the keys were seen. In a less than ideal world, however, you may not remember where you last saw the keys; other memories might compete with your specific memories from Friday. Indeed, when you aren't precisely sure where you last saw the keys, you may check the hook where you usually keep them first, because you know that searching at a location where they keys are seen frequently can (on other days) be a successful search strategy.

This story illustrates the task a child faces in learning to find things in the world. A child must learn that some things are most likely found at the location they were *last* seen, while other things are most likely found they are most *often* seen (while it makes sense to look for keys in the last place you saw them, if you haven't seen your cat for a while, you would be best off looking in the most frequent place the cat is found), and that a successful search will involve weighing these considerations against what the child remembers about the last and most frequent locations of an object. From this perspective, "perseverative errors," in which a child searches for an object in a most frequent location rather than a most recent, can be seen as a misapplication of what in other circumstances might be a logical strategy.

Piaget (1954) first described what are often called "A-not-B errors" in infants, namely that 8- to 12-month infants will generally search successfully for an object in an initial location (A) and then fail to search for it when it is conspicuously hidden in a new location (B) instead, continuing to search at location A. Subsequent studies have confirmed that in actively searching for hidden objects, infants robustly commit this prototypical A-not-B error, ignoring the last location of an object when they reach for it (see Marcovitch & Zelazo, 1999 for a meta-analyses). In seeking explanations for this reaching behavior, accounts have tended to assume that infants' errors stem from problems associated with implementing a correct search, such as limited working memory and inhibitory control, or from weak memory traces for the object and hiding location (e.g. Baillargeon, Graber, Devos, & Black, 1990; Diamond, 1988; Diamond, Cruttenden, & Neiderman, 1994; Munakata, 1997; Thelen, Schöner, Scheier, & Smith, 2001).

In what follows, we take a slightly different approach to thinking about the prevalence of the perseverative searching behavior. Rather than assuming that a child "knows" how to search, all other things (object concepts, memory, etc.) being equal, we consider what might be expected when children are *learning how* to search. As noted above, in

learning to find objects, children have to figure out which strategy is appropriate in a particular context.

We consider the question of how children learn to discriminate between possible object retrieval strategies in a given context within the framework provided by formal *learning theories* (e.g. Rescorla & Wagner, 1972), which view learning as a process of acquiring information about the relationships between salient events (*outcomes*) in the environment, and the cues that allow those outcomes to be predicted. From this perspective, children's learning to search is a process of trial and error, each iteration of which strengthens or weakens cues depending on how well they predict an outcome (termed *error-driven learning*). While from an adult perspective, children's "perseverative" search may be erroneous, we suggest that from an infant perspective, their behavior is rational in following the often accurate cue of where something has been found most frequently. In the approach of learning theories, A-not-B "errors" are an inevitable, and logical, step along the path to adult search expertise, as infants go through the process of learning which situational cues best predict an outcome.

Mastering Search

In considering perseveration as part of a logical learning strategy, it is interesting to note that it is also evident when infants learn other (novel) relationships between cues and outcomes, and not just in hiding events. Aguiar & Baillargeon (2000), showed that 7-month-old infants perseverated in pulling a towel that had previously had a toy on it, even when the toy was now visibly on a different towel, and Smith, Thelen, Titzer, & McLin (1999) found that directing infants' attention to lids in an A-not-B pattern drew reaching behavior similar to when objects were hidden. These examples suggest that the perseverative response has more to do with the process of learning where to direct an action than particular properties of the objects or hiding events themselves. Munakata (1997) found that perseveration was reduced if the experimenter waved lids at A, but then hid a toy at B, suggesting that infants learn about particular outcomes at particular locations, such that if a *different* outcome is observed at a new location, it is less influenced by prior evidence.

Thus, it appears that A-not-B errors are not only due to prior motor habit (given that infants can switch when a new object is hidden at the new location), or by infants having trouble conceptualizing objects or not being interested in them (given that they search successfully at the first location). Infants appear to "understand" the task, so *why* do they fail to search correctly after a switch in location?

As we noted above, successfully searching for an object involves weighing a number of cues to its likely location: its last location, its usual location, the independent mobility of the object, etc. If a child doesn't know how to weigh those items correctly in search, then in the early stages of learning within a particular context, cues learned when an object is hidden at location A and then reappears at A may suggest to the infant that location A is the most *likely* location a hidden object will reappear from. When the object is first hidden at location B, the infant will have no experience of objects reappearing at B, and given that the situational cues provided by an object being at A and an object hidden at B overlap, the infant's best guess ought to be that the object will reappear at A. Given only this information, a child thus *ought* to continue to search at location A when the object has first been hidden at location B.

Over time, if we assume that the child in the A-not-B task is capable of learning (e.g., Rescorla & Wagner, 1972), the error resulting from incorrect searches at location A during trials when the object is hidden and retrieved from location B will *weaken* the cues that continue to predict that an object will be at location A. At the same time, cues supporting the prediction that location B is the correct location will *strengthen*. Eventually, the cues predicting that the object will reappear at location B will have more support than those predicting that it will reappear at A, and the infant will slowly come expect the object at B.

Moreover, since the cues supporting the general "search at the most frequent location" response will generate error over time as hiding locations are switched, while conjunctive cues that support searching at the specific place an object was last seen will continue to be accurate, this process will gradually result in conjunctive cues (that favor the most recent location) over cues favoring search at the most frequent location. Thus, the infant will gradually learn to weigh search strategies within a context from the evidence of their success and failure, and will then come to resemble the adult strategies described above.

Further, although we have talked so far about children 'learning to search,' the evidence is that children do not initially appear to learn abstract, generalized "search." While 9-month-old infants succeed in Aguiar & Baillargeon's (2000) towel pulling task, they still fail the standard A-not-B task (Piaget, 1954), even though the tasks are structurally similar. Rather than learning abstract "search," it appears that children may instead learn search and retrieval strategies within particular contexts.

To formally illustrate how children might learn the appropriate search strategy in the A-not-B task, we simulated this process using the Rescorla-Wagner (1972) learning model. In the model the change in associative strength between a cue C_i and a relevant environmental event E_j given a learning trial n is defined to be:

$$\Delta V_{ij}^{n} = \alpha_i \beta_j (\lambda_j - V_{total})$$

This rule specifies how the associative strength (V) between individual cues (C_i) and an event (E_j) changes as a result of discrete exposure trials, where n indexes the current trial (V_{total} is the sum of the associative strengths between all CSs present on the current trial and US_j). The individual saliency of cues can be denoted by a parameter α_i (where C_i $0 \leq \alpha_i \leq 1$), the rate at which cues are learned with respect to an

event is determined by a learning rate parameter β_j (where $0 \leq \beta_j \leq 1$), and the maximum amount of associative strength that an event E_j can support is denoted by value λ_j, such that the *amount learned* by the set of cues on a given trial is the value of $\lambda_j - V_{total}$, modulated by β and α.

In simulating children learning about the two locations (A and B) of the A-not-B task, we assume that on each trial infants will search at the location that they most strongly associate with the object's reappearance, and that the outcome of each search, either finding the object or not, will be incorporated into task learning and affect later search behavior. To reflect the fact that infants will most likely spend longer looking towards the location at which they expect objects to reappear (keeping in mind that this may not match the location in which the object was hidden), the saliency of the unattended location was set lower than that rate for the attended location.[1] This allowed the model to reflect the likelihood that infants would learn less quickly about location B when they were still primarily attending to A. This meant that, as a consequence of their attending to A, infants were initially slow to learn about hiding events at B.

To initially examine the learning problem facing an infant, we simulated the learning of cues that represented each location, and the strength of association between those cues and the object. Figure 1 shows these associative strengths changing across the trials (first at location A, then at location B) of a typical A-not-B task. Initially the association for the object at location A increases as objects are being hidden there. The association with location A then decreases slowly when hiding events begin at location B, because the error associated with unsuccessful search at location A reduces the value of cues predicting location A. For cues associated with location only, the model predicts that if hiding were to be later switched back to location A, learning to search at A would again proceed slowly, given that B is now strongly associated with the object's location.

However, unlike this model, which predicts that an infant will simply search at the location most associated with an object, children *do* eventually succeed on the A-not-B task, switching back and forth with ease, suggesting they need to learn about the conjunction between a location and the most recent hiding event. To reflect this, we added conjunctive cues to the model. The changing associative strength of the conjunctive cues, along with those of the simple location cues discussed previously, is shown in Figure 1. During the initial hiding events at A, there is nothing to distinguish between the simple and conjunctive cues, because each predicts the correct outcome at location A. However, after the switch in locations, the associative strengths of the simple cues are weakened because they fail to predict the object location as well as the conjunctive cues.

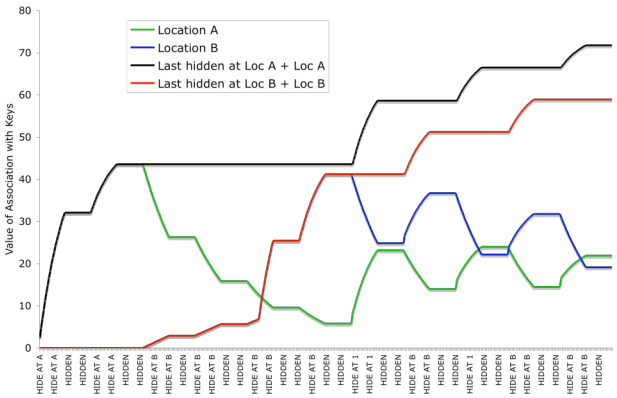


Figure 1: A Rescorla-Wagner model of cue competition between two cues representing location only, and two conjunctive cues representing last known location. The model shows associations with the hidden object across two trials at A, then four trials at B, then hiding events alternating between the two locations.

Learning and Visual search: Experiment 1

We suggested above that A-not-B errors are the result of a particular kind of search in a particular context, and that the relationships (and similarities) between the particular cues observed and the particular outcomes expected from those cues is a very important component in shaping children's early search behaviors. To examine our account of learning to search, we conducted a study of the visual search behavior of infants in an A-not-B paradigm using eye-tracking. Though the standard A-not-B task generally involves infants *reaching* for hidden physical objects, if our hypothesis about the need for infants to learn appropriate search strategies is correct, we would expect the same pattern of behavior to be apparent across modalities (see also Diamond, 1990; Hofstadter & Reznick, 1996; Bell & Adams, 1999). In addition, by measuring eye gaze across the duration of a specified search period in which both hiding locations were visible but neither was cued, we sought to acquire a continuous measure of children's attention to—and association between—each of the two regions and the objects shown to have been hidden there.

Methods

Participants 32 9-month-old infants successfully completed our testing procedure (range 8 months 17 days to 9 months 17 days, median 9 months 7 days; 16 males). An additional 13 infants failed to complete the experiment due to fussiness. Participants were recruited from a volunteer

[1] α_i=0.05 for attended stimuli and 0.075 for unattended stimuli; the other parameter values for the simulations reported were: λ = 100% when the keys are visible at a given location or 0% when they are not visible at a given location and β_j=0.5.

pool, which reflects the properties of the community surrounding Stanford University.

Stimuli Stimuli were movies of colorful keys, accompanied by music. The keys were familiarized in center screen, and were then shown moving across the screen and disappearing into a bucket on one side of the screen. An identical bucket was present on the other side of the screen.

Following the disappearance of the keys, a pinwheel distracter appeared in the center of the screen for three seconds, then disappeared. For the following four seconds only the buckets were visible, while the music that accompanied the keys was played to encourage searching. After the four-second search period, the keys reappeared from the same bucket into which they had disappeared, before moving back towards center screen. The pinwheel animation then reappeared in center screen and remained until infants' attended to it, at which point the next trial began.

Procedure and design Participants sat on a caregiver's lap during testing, facing a 152cm projection screen, which was approximately 180cm from them. An Applied Science Laboratories (ASL) Model 504 corneal reflection eye tracking system collected eye movement data as infants were shown the stimulus displays. A computer script translated the gaze coordinates recorded by the system into gaze durations to regions of interest (ROI) defined around each of the hiding wells during the 4-second search period after each hiding event.

Infants were shown the key-hiding sequence six times: the keys were hidden twice in the bucket on one side of the screen, and then four times in the bucket on the other side of the screen, mimicking the sequence of a typical A-not-B task. Side of initial presentation was counterbalanced across participants.

Results and Discussion Looking-time data are presented as a difference in milliseconds between the amount of looking to the two ROIs, with positive values reflecting greater looking (bias) towards the A-side, and negative values reflecting greater looking (bias) towards the B side, calculated for each participant on each trial.

For data-analysis, the six trials were grouped into three pairs; the first two trials, in which the keys were shown hidden in location A, are labeled 'A trials'; the two subsequent trials, immediately following the switch to the new hiding location B are labeled 'early B trials'; and the last two trials in location B are labeled 'later B trials.' Because not all of the infant participants provided clean data from all six trials, pairing the trial data in this way allowed some missing cells to be filled in. Data was averaged for both trials if available, but if only one trial of the pair had clean data, then this trial was used.

Although the display shown to the infants was intended to mimic the manual A-not-B search task in presentation, the presentation could not be infant-controlled in the same way that manual studies can. In a manual search task, the toys can continue to be hidden at location A until the infant has reached a success criterion for searching at that location, ensuring that the infant has been attending to, and learning about, the hiding events; however, in the current visual search task, we were unable to employ such a criterion accurately in real time. Therefore, the presentation of hiding events continued without any performance-based contingency. Because our visual search task did *not* require success at location A prior to the switch in hiding to location B, we expected that there might be individual differences among the participants in extent of learning about location A, and that this difference might affect later search behavior. Consistent with the design of our model, we would not expect children to learn about hiding events that occurred at a location to which they were not attending.

We considered two possible measures of how much children learned about hiding events at location A. One that is predicted by the model is that infants who attend more to the actual hiding event, as the keys move towards the location in which is will be hidden, learn more about it. However, infants are capable of deploying covert attention, such that gaze does not necessarily imply attention (e.g. Johnson, M.H., Posner, M. & Rothbart, M.K., 1994). Therefore, we decided to use infants' looking behavior during the search period of hiding events at location A as evidence for how much they learned about those events. Following a hiding event at location A, an infant who looks a lot at the ROI for location A, and little at the ROI at location B, demonstrates a greater expectation for the object to be at (and reappear from) location A – the infant has learned something about what to expect about hiding events in the context! As might be expected, there is a correlation in the trials at location A between how much infants attend to the object as it moves towards a hiding location, and how much they look at location A during the search period (r=.456, p=.005), but in terms of making predictions about later search expectations, the actual extent of learning demonstrated during trials at location A seems more directly relevant as a measure, which is why we have chosen to focus on that.

Analysis confirmed that infants in the study varied in how much they looked towards location A during the search period of A trials, with 17 infants who (accurately) looked more at location A and another group of 15 infants who looked more at location B during that search period. An ANOVA comparing the searching patterns across the study between these two groups of infants (who searched differently during A trials) revealed the anticipated difference in the patterns of infants' looking across the study, $F(2,90)=34.597$, $p<.001$. Accordingly, the children were separated for remaining analyses: an 'attenders' group

of children who looked more to A during the initial search trials, and a 'non-attenders' group who looked more to B during the initial search trials even though the keys were hidden at location A.

A further omnibus ANOVA, including attending status as a variable, revealed an overall 'side' x 'time' interaction, $F(1,92)=2.622$, $p=.022$, and a 'side' x 'attending status' interaction, $F(1,92)=5.435$, $p<.001$ (Figure 2). These results revealed an overall change in where the infants were looking during the search period across trials, and that this change was driven by the attenders, who searched first at location A and then changed their locus of search over time, as more hiding events occurred at location B. Unsurprisingly, the non-attenders did not change their searching behavior throughout the study.

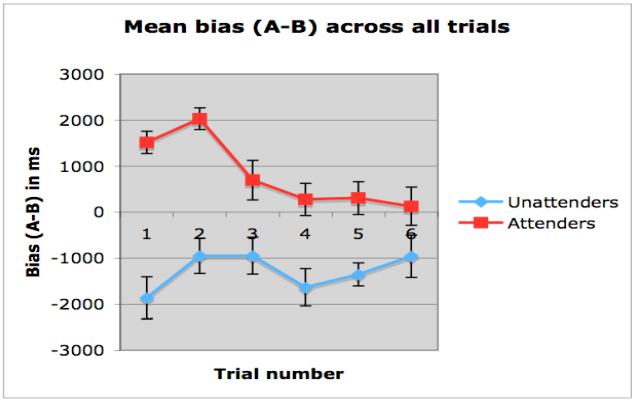


Figure 2: A plot of the difference in looking time to locations A and B across the visual search A-not-B task, (with the first two trials at location A and the remaining at location B) for each of the two groups (those who searched more at location A during the first two trials, attenders, and those who did not, non-attenders.

An analysis of the visual search of the attenders revealed a change in looking bias across the trials, with decreasing looking to location A, $F(1,49)=14.057$, $p<.001$. Despite this trend however, there was still a main effect of side in the study, $F(1,49)=29.468$, $p<.001$, with significantly more looking to A (M=2038 ms) than B (M=1192 ms), $t(1,50)=4.611$, $p<.001$, even there are twice as many hiding events at B than at A over the course of the experiment. This overall greater searching at location A within the attending group is noteworthy because it demonstrates the same perseverative trend seen in the typical A-not-B task with manual search. Along with the overall perseverative trend, however, the data are also consistent with the learning model presented. Specifically, children's searching at the formerly correct location gradually lessens as the cues that predict that location are weakened following hiding events in a new location, and the rate of learning to search the alternative location is also slowed until attention shifts away from the initial location. Individual differences among the attenders elucidate this learning, with a regression showing that the extent of the searching bias A events "predicted" the extent of bias on early B-trials, $p=.018$, a relationship that was not significant for non-attenders.

The non-attenders, who did not learn about location A or the hiding events that occurred there, were not expected to behave in the same way as the attenders. While these non-attenders showed a main effect of side, $F(1,42)=10.979$, $p=.002$, this resulted from more overall looking to location B, $t(1,43)=7.282$, $p<.001$. More distinctly from the attenders, the non-attending infants did not change their looking bias over the course of the trials in different locations, $F(1,42)=.378$, ns. Since the non-attenders failed to notice the hiding events at A, there was no reason that they should later begin to search there. The fact that the non-attenders do not change their bias over time suggests that changes in search do not result simply from regression to the mean (a possible concern, because groups were split based on early search behavior), but reflect different patterns of learning over time in the two groups.

General Discussion

Children who initially learned about an object hidden at one location continued to search visually at that location even after the object was hidden elsewhere, but then showed a gradual shift in their search behavior away from the initial location and towards the new location. This pattern of data is consistent with the idea that learning is a function of experience and the expectations that experience produces (Rescorla & Wagner, 1972; Ramscar, Yarlett, Dye, Denny, Thorpe, in press; Ramscar & Dye, 2009) and suggests that when infants initially learn that objects will appear from A, they will "perseverate" in that response before gradually learning to predict the objects' appearance at B. The correlation between the attenders' bias during A trials and the early B trials, but *not* the later B trials, also supports the idea that the initial bias towards A must be unlearned, and that this will happen only as more hiding / appearance events are shown at location B (see also Diedrich, Thelen, Smith, & Corbetta, 2000). This gradual change in looking preference over time supports our hypothesis that search is something children have to learn, and that success or failure at different kinds of search is, to a degree, a matter of contextual experience.

Our results further suggest that infants given the same exposure to a particular hiding location may actually learn differently about it, in part because of the degree to which they attend to training. This variability in attending might not be evident in a reaching paradigm. For example, the results of a recent A-not-B study by Topal, Gergely, Miklosi, Erohegyi & Csibra (2008) are consistent with the idea that the degree to which children attend to hiding events at location A will impact the degree to which they perseverate in search to that location rather than a new hiding location. In their study, Topal et al. found that infants

who were directed to hiding events with the highest level of engagement (both words and gestures) later showed the greatest perseveration to location A, while a group who saw only gestures was more likely to switch successfully to searching at location B. Although the groups had similar rates of searching at location A, that does not necessarily imply equal attention to or learning about hiding events in that location, because any attention drawn to location A should make a reach there more likely than to another location, given the forced-choice single measure outcome. A more continuous measure of attending or learning during trials at location A, such as eye tracking, could have confirmed whether this later difference in perseveration was because attention was actually increased to location A when the experimenter verbally engaged the infants, thereby increasing their learning about that location, and therefore increasing the time it took them to unlearn their response to location A, therefore leading to the observed greater perseveration to location A after the change in hiding location.

While there is much to explain with regards to the development of children's ability to search—and not least how the learning of conjunctive cues over extended trials might impact performance on the A-not-B task—we believe that there is insight to be gained from seeing infants' behavior in the A-not-B task in terms of *learning* to search, and the patterns of behavior that accompany such learning, rather than as a failure to search correctly. Not only does this approach offer an answer to the often puzzling search behavior of children, but we believe that the combination of eye-tracking and computational modeling methods used in the current study provide a useful formal framework for addressing these questions.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant Nos. 0547775 and 0624345 to Michael Ramscar. We are grateful to Melody Dye for insightful comments and discussions. We also thank the participants, and Research Assistants.

References

Aguiar, A., & Baillargeon, R. (2000). Perseveration and problem solving in infancy. In H. Reese (Ed.), *Advances in child development and behavior* (Vol. 27, pp. 135-180). New York: Academic Press.

Baillargeon, R., Graber, M., Devos, J., & Black, J. (1990). Why do young infants fail to search for hidden objects? *Cognition, 36*(3), 255-284.

Bell, M. A., & Adams, S. E. (1999). Comparable performance on looking and reaching versions of the A-not-B task at 8 months of age. *Infant Behavior and Development, 22*(2), 221–235.

Diamond, A. (1988). Abilities and neural mechanisms Underlying AB Performance. *Child Development, 59*(2), 523-527.

Diamond, A. (1990). The development and neural bases of memory functions as indexed by the AB and delayed response tasks in human infants and infant monkeys. In A. Diamond (Ed.), *The development and neural bases of higher cognitive functions* (pp. 267-309). New York: New York Academy of Sciences Press.

Diamond, A., Cruttenden, L., & Neiderman, D. (1994). AB with multiple wells: 1. Why are multiple wells sometimes easier than two wells? 2. Memory or memory+ inhibition. *Developmental Psychology, 30*(2), 192–205.

Diedrich, F. J., Thelen, E., Smith, L. B., & Corbetta, D. (2000). Motor memory is a factor in infant perseverative errors. *Developmental Science, 3*(4), 479-494.

Hofstadter, M., & Reznick, J. S. (1996). Response modality affects human infant delayed-response Pperformance. *Child Development, 67*(2), 646-658.

Johnson, M. H., Posner, M. I., Rothbart, M. K. (1994). Facilitation of saccades toward a covertly attended location in early infancy. *Psychological Science, 5,* 90-93.

Marcovitch, S., & Zelazo, P. D. (1999). The A-Not-B error: Results from a logistic meta-analysis. *Child Development, 70*(6), 1297-1313.

Munakata, Y. (1997). Perseverative reaching in infancy: The roles of hidden toys and motor history in the AB task. *Infant Behavior and Development, 20*(3), 405–416.

Piaget, J. (1954). *The Construction of Reality in the Child* (M. Cook, Trans.). New York: Basic Books, Inc.

Ramscar, M., & Dye, M. (2009) Expectation and negative evidence in language learning: the curious absence of mouses in adult speech. *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, Amsterdam, Netherlands.

Ramscar, M., Yarlett, D., Dye, M., Denny, K. & Thorpe, K. (in press). The Feature-Label-Order Effect in Symbolic Learning. *Cognitive Science.*

Rescorla, R. A., & Wagner, A. R. (1972). A Theory of Pavlovian Conditioning: Variations in the Effectiveness of Reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical Conditioning II: Current Research and Theory* (pp. 64-99). New York: Appleton-Century-Crofts.

Smith, L. B., Thelen, E., Titzer, R., & McLin, D. (1999). Knowing in the context of acting: the task dynamics of the A-not-B error. *Psychol Rev, 106*(2), 235-260.

Thelen, E., Schöner, G., Scheier, C., & Smith, L. B. (2001). The dynamics of embodiment: A field theory of infant perseverative reaching. *Behavioral and Brain Sciences, 24*(01), 1-34.

Topal, J., Gergely, G., Miklosi, A., Erohegyi, A., & Csibra, G. (2008). Infants' perseverative search errors are induced by pragmatic misinterpretation. *Science, 321*, 1831-1834.

The Impact of Collective Opinion on Online Judgment

Yasuaki Sakamoto (ysakamot@stevens.edu)
Center for Decision Technologies
Howe School of Technology Management, Stevens Institute of Technology
Hoboken, NJ 07030 USA

Abstract

Social media are part of our everyday lives. These technologies allow people to share their opinions with others. Here I examine whether the opinions posted online actually change people's perception of the world or they simply serve as anchors when people post their own opinions. Participants rated the interestingness of given stories. In one condition, the stories were presented with invented average ratings of others that matched the rating task. In another condition, the assumed opinions of others mismatched the rating task. Only in the task match condition, people used the opinions of others when rating the stories. The results suggest that the other's opinions are used as anchors when making judgments and do not influence people's perception as much as one may expect. The current work provides insights into cognitive mechanisms underlying collective behavior in online environments as well as a lesson for users and designers of social media websites.

Keywords: Collective opinion; online judgment; social influence; anchoring and adjustment.

Introduction

Many websites allow users to contribute content. Examples include the product reviews on Amazon, the user ratings on eBay, and the votes for stories on Digg. Although these social media websites are used to share information with others (Glushko, Maglio, Matlock, & Barsalou, 2008), little is known about how people process the opinions of others in online environments. In the current work, I examine whether the opinions posted online actually change the way people perceive the world or they simply serve as anchors when people post their own opinions.

Previous offline studies have suggested that people have a strong motivation to compare their opinions with others (Festinger, 1954). People often adopt the decisions of others (e.g., Cialdini & Goldstein, 2004; Deutsch, & Gerard, 1995; Gureckis & Goldstone, 2006) due to their desire to make correct responses under uncertainty (Sherif, 1935) or their desire to be like others (Asch, 1951; 1956).

More recent work has shown that knowing other's decisions also influences people's decisions online. Salganik, Dodds, and Watts (2006) found in an online market study that whereas good music was always downloaded by many and bad music was always unpopular, the popularities of the pieces in between varied depending on whether or not the number of downloads the pieces had was publicly available. Sakamoto, Sadlon, and Nickerson (2008) showed that only a computational model that assumed that users copied other users' decisions could account for the popularity of stories in an online community. Sakamoto, Ma, and Nickerson (2009) further showed that participants in their online experiments switched their preferences for stories when the assumed numbers of previous supporters were flipped.

These previous studies clearly show that the opinions of others influence decisions. Nevertheless, it is unclear whether the opinions of others change people's mental representations. For instance, when people become aware that many others like a story and decide that they also like the story, (1) are they simply using the opinions of others as anchors for making their response or (2) do the opinions of others actually change their perception of the story? Relevant to this question, Berns et al. (2005) found changes in the activation of the visual regions of the brain when participants conformed to the majority's decisions, suggesting that the decisions of others might actually influence people's perception of what they saw. On the other hand, the anchoring and adjustment heuristic often observed in decision making (Tversky & Kahneman, 1974) has been proposed as a process consumers use to weight information from others when evaluating products (Wooten & Reed, 1998).

To tease apart the two accounts, the current experiment examined how people behave online using materials from real environments. Participants from an online community (Amazon's Mechanical Turk) were asked to rate the interestingness of stories obtained from another online community (Digg). The assumed opinions of others associated with the stories were manipulated. In the task match condition, the opinions of others took the form of previous average ratings, which matched the rating task the participants completed (see Figure 1). In the task mismatch condition, the opinions of others took the form of the number of previous users who found the story interesting, which mismatched the rating task (see Figure 2). The two conditions differed only in the information about collective opinion associated with the stories.

If people use the opinions of others as anchors to make their own judgments, then the participants in only the task match condition will be influenced by the collective opinion. According to this account, when the format of the other's opinions and the format of the task mismatch, people will not be able to use the previous opinions as anchors to complete the task. This account predicts that whereas the stories associated with higher pervious ratings will be rated higher in the task match condition, there will be no influence of collective opinion in the task mismatch condition.

1. **Browse the iTunes Store Without Installing iTunes Software** **(previous rating: 2)**
simples.in — Even if you are not buying movies or songs from the iTunes store, it's still a perfect place to learn about new audio books and free podcast shows that have just been released onto the web. Other than that, iPod Touch and iPhone customers use the iTunes store to download games and apps for their devices.

Extremely Boring	1	2	3	4	5	Extremely Interesting

2. **Vacation Ideas** **(previous rating: 4)**
mpakvngwl.org — Great Vacation Ideas: As a "Professional Vacationer," I often times am blown away by how easily we tourists are taken advantage of. The prices at some places are astounding, and down right despicable. That's why I want to write some fun vacation ideas.

Extremely Boring	1	2	3	4	5	Extremely Interesting

3. **Curing Cancer with Baking Soda!** **(previous rating: 2)**
healingcancernaturally.com — An Italian oncologist, Dr. Tullio Simoncini, has devised a simple, very inexpensive and apparently frequently effective cancer treatment centered around the use of sodium bicarbonate, taken orally or by infusion.

Extremely Boring	1	2	3	4	5	Extremely Interesting

4. **Netbooks Become The New Notebooks!** **(previous rating: 4)**
blogblek.co.cc — Despite the rough economy, certain segments of the computer market are faring well. One of these is the netbook (also known as the ultra-portable or mininotebook), which continues to carve a comfortable niche in the PC market by providing an ideal mix of power and portability. HP's latest entries into this segment should further cement the netbook.

Extremely Boring	1	2	3	4	5	Extremely Interesting

Figure 1: Four news stories presented to the task match condition are shown. The information about the collective opinion matches the rating task. The first and third stories have lower previous ratings than the second and fourth stories.

1. **Browse the iTunes Store Without Installing iTunes Software** **(82 people)**
simples.in — Even if you are not buying movies or songs from the iTunes store, it's still a perfect place to learn about new audio books and free podcast shows that have just been released onto the web. Other than that, iPod Touch and iPhone customers use the iTunes store to download games and apps for their devices.

Extremely Boring	1	2	3	4	5	Extremely Interesting

2. **Vacation Ideas** **(2377 people)**
mpakvngwl.org — Great Vacation Ideas: As a "Professional Vacationer," I often times am blown away by how easily we tourists are taken advantage of. The prices at some places are astounding, and down right despicable. That's why I want to write some fun vacation ideas.

Extremely Boring	1	2	3	4	5	Extremely Interesting

3. **Curing Cancer with Baking Soda!** **(85 people)**
healingcancernaturally.com — An Italian oncologist, Dr. Tullio Simoncini, has devised a simple, very inexpensive and apparently frequently effective cancer treatment centered around the use of sodium bicarbonate, taken orally or by infusion.

Extremely Boring	1	2	3	4	5	Extremely Interesting

4. **Netbooks Become The New Notebooks!** **(2412 people)**
blogblek.co.cc — Despite the rough economy, certain segments of the computer market are faring well. One of these is the netbook (also known as the ultra-portable or mininotebook), which continues to carve a comfortable niche in the PC market by providing an ideal mix of power and portability. HP's latest entries into this segment should further cement the netbook.

Extremely Boring	1	2	3	4	5	Extremely Interesting

Figure 2: Four news stories presented to the task mismatch condition are shown. The information about the collective opinion mismatches the rating task. The first and third stories have fewer supporters than the second and fourth stories.

In contrast, if the opinions of others actually change the participants' representations, both the task match condition and the task mismatch condition should show influence of collective opinion. This is because according to this account, knowing the other's opinions lead to actual changes in people's perception, and such changes will transfer across tasks that differ on the surface. In this way, the current work will provide new insights into cognitive mechanisms underlying collective behavior in online environments as well as a lesson for users and designers of websites.

Method

Participants

Two hundred and seven (109 females and 98 males, $M = 31$ years old, $SD = 11$ years) members of Amazon's Mechanical Turk community (www.mturk.com) completed the experiment. They earned 10 cents for participation.

Materials

Six news stories were selected randomly from an online community (www.digg.com) with the constraints that they (1) were not about exceptional events, (2) were not promoted to the front page, (3) were submitted to the community on the same date, and (4) had between 3 and 5 supporters in the community. These constraints were used to minimize the possibility that participants were already familiar with and had strong opinions about the stories.

Design and Procedure

Table 1 summarizes the manipulation of collective opinion in the current work. There were three conditions: task match, task mismatch, and control. The same six stories were used in the three conditions. To measure any differences in interests among groups, no information about the opinions of others was provided for the same two stories in all groups (Story X and Story Y in Table 1). Collective opinion was manipulated for the remaining four stories (Story 1 – 4 in Table 1). No information about the opinions of others was given in the control group.

Condition		Story X	Story Y	Story 1 (A)	Story 2 (B)	Story 3 (A)	Story 4 (B)
Task match	Low-high	–	–	2	4	2	4
	High-low	–	–	4	2	4	2
Task mismatch	Low-high	–	–	82	2377	85	2412
	High-low	–	–	2377	82	2412	85
Control		–	–	–	–	–	–

Table 1: Manipulation of collective opinion is summarized. Each value represents the average interestingness rating of the story in the task match condition, and the number of people who found the story interesting in the task mismatch condition. A dash indicates that no information about collective opinion was provided. Story X and Story Y were used to measure if the three groups differed in their interests. Stories 1 and 3 are grouped as Story A because they are the same type within each condition. For the same reason, Stories 2 and 4 are grouped as Story B.

In the task match condition, the four stories were associated with the invented average interestingness ratings from previous raters. This information about collective opinion matched the interesting rating task the participants completed. Although the stories selected had similar number of supporters in Digg, suggesting that they are similar in popularity, some stories might be inherently more interesting than others for some participants. To cancel out any effect due to the difference in stories, the assumed information about collective opinion was counterbalanced. In the low-high group in the task match condition, the previous ratings were 2, 4, 2, 4 for the first, second, third, and fourth stories, respectively, as shown in Table 1. In the high-low group in the task match condition, the previous ratings were 4, 2, 4, 2 for the first, second, third, and fourth stories respectively, the flip of those in the low-high group.

In the task mismatch condition, the four stories were associated with the invented number of people who found the stories interesting, which mismatched the rating task. As shown in Table 1, for the low-high group in the task mismatch condition, the number of people who found the stories interesting were 82, 2377, 85, 2412 for the first, second, third, and fourth stories, respectively. In the high-low group in the task mismatch condition, the number of people who found the stories interesting were 2377, 82, 2412, 85 for the first, second, third, and fourth stories, respectively. The first and third stories are grouped as Story A because they are the same type within each condition. For the same reason, the second and fourth stories are grouped as Story B.

The results from previous studies (e.g., Sakamoto et. al, 2009) suggest that in the task match condition, whereas the low-high group will provide lower ratings on the first and third stories (Story A) than on the second and fourth stories (Story B), the high-low group will provide higher ratings on Story A than on Story B. Thus, there will be an interaction between Group (low-high vs. high-low) and Story (A vs. B) in the task match condition. Whether this interaction will result in the task mismatch condition is unclear. If knowing the collective opinions of others can indeed change people's internal representations, then there should be an interaction in the task mismatch condition. This is because if knowing how many people think a story is interesting indeed changes one's perception of the story, this change should influence her responses when she rates the interestingness of the story. Failure to find an interaction in the task mismatch condition suggests that knowing the opinions of others merely anchors people's responses in a task that is compatible with the format of the opinions. The control group's responses will provide the baselines for the interestingness of the stories.

Participants completed the experiment online. They were instructed to read six brief news stories and rate the interestingness of the stories using a 5-point scale. The instruction for the task match condition informed the participants that the previous ratings indicated the average ratings of previous readers. The instruction for the task mismatch condition informed the participants that the number of people indicated how many readers found the stories interesting previously. The instruction for the control group contained no information about the opinions of others. The first two stories were the ones used to measure whether the groups differ in their interests and had no information about the other's opinions in all groups. Then, the participants rated the four stories. Finally they completed the questions asking demographic information.

Results

All participants were included in the analyses. The groups did not differ in their ratings for the first two stories ($F < 1$), suggesting that the groups did not differ in their interests. I first present the results from analyzing the task match condition. One interest is whether the current work replicates previous findings (Sakamoto et al., 2009) that one can influence people's judgment by manipulating the information about the opinions of others. A 3 by 2 Analysis of Variance (ANOVA) was performed on the task match condition's interestingness ratings, with Group (task match low-high vs. task match high-low vs. control) and Story (A vs. B) as independent variables. The main effect of Group approached significance, $F(2, 123) = 2.79$, $p = .07$. As shown in Figure 3, the control group had overall higher ratings than the other groups. Perhaps people tend to use higher end of scales in these tasks, and the invented previous ratings shifted their responses down. Alternatively,

all stories might have been quite interesting to the participants, and the information about the previous ratings could only lower their ratings. There was no significant main effect of Story, $F < 1$. As predicted, there was a significant interaction, $F(2, 123) = 5.75$, $p = .004$. As can be seen in Figure 3, whether Story A or B was rated higher depended on the type of the Group. As predicted, whereas the low-high group rated Story B higher than Story A, the high-low group rated the rated Story A higher than Story B, $F(1, 83) = 5.96$, $p = .017$. The participants conformed to the opinions of others.

Further analyses revealed that whereas the high-low group rated Story A significantly more interesting than Story B, $t(41) = 2.03$, $p = .049$, the difference in the low-high group's ratings on Story A and Story B did not reach significance $t(42) = 1.40$, $p = .17$. Unexpectedly, the control group rated Story B significantly more interesting than Story A, $t(40) = 2.73$, $p = .009$. The information about the opinions of others in the low-high group was consistent with the participants' ratings without social influence. Thus the previous ratings might have provided no new information to the participants in the low-high group.

It is surprising that the high-low group in the task match condition shows the opposite pattern from the natural ratings shown by the control group. The participants in the high-low group were willing to rate Story B less interesting than Story A consistent with the invented collective opinions, even though the invented collective opinions went against the true collective opinions suggested by the control group's ratings. This suggests that either the participants did not have strong opinions about these stories, or the desire to conform was so strong that they gave untruthful ratings. The results from the task match condition showed that the participants conformed to the given information about the opinions of others.

Our main interest is whether a similar pattern of results can be found in the task mismatch condition. Figure 4 shows that the pattern of results for the task mismatch condition was rather flat. A 3 by 2 ANOVA was conducted on the task mismatch condition's interestingness ratings, with Group (task mismatch low-high vs. task mismatch high-low vs. control) and Story (A vs. B) as independent variables. There was no significant main effect of Group, $F < 1$. There was a significant main effect of Story, $F(1, 119) = 6.95$, $p = .009$. Figure 4 shows that collapsing across Group, Story B has overall a higher rating than Story A. However, the effect of Story was mostly due to the control group. Both the low-high group and the high-low group did not differ significantly in their ratings of Story A and Story B, $t(40) = 1.41$, $p = .17$, and $t < 1$, respectively. As can be seen in Figure 4, there was no significant interaction, $F < 1$. The similar pattern of ratings in the three groups in the task mismatch condition indicates that knowing how many people found the stories interesting had little influence on the participants ratings.

Discussion

In the current study, the participants rated the interestingness of stories with varying information about the opinions of others. In one condition, the information about other's opinions matched the rating task they were asked to complete. In another condition, the information about the other's opinions mismatched the format of the rating task. In this way, I examined whether the opinions of others could actually change the way people perceived the world, or they simply served as anchors when people made their own judgments. The results were consistent with the latter hypothesis. Only when the format of the information about the opinions of others matched the format of the rating task, the opinions of others had significant influence on the participants' judgments.

One might say that perhaps knowing the number of previous readers who found the stories interesting simply do not influence people's judgment. However, previous work showed that information about the number of people who liked a story had significant influence on which of two stories the participant liked better (Sakamoto et al., 2009). Thus, I predict that knowing the number of previous people who have found the stories interesting will influence people's responses on a binary decision task that asks which of the two stories they find more interesting. I further predict that information about the previous ratings will have no significant influence on such binary tasks because the information and the task mismatch in this case. Data from these two groups are being collected right now.

The current work provides insights into cognitive mechanisms underlying collective behavior. Anchoring and adjustment (Tversky & Kahneman, 1974) may be one cognitive mechanism that drives preferential attachment, (Barabási & Albert, 1999), which characterizes the rich-get-richer effect observed in real social networks. Further, the current findings extend existing theories of social influence by suggesting that social influence is not as internal as one may think.

The current results may also have connections to other cognitive theories. The finding that the information about other's opinions must match the task to have influence might be related to the idea of transfer appropriate processing from the memory literature (Morris, Bransford, & Franks, 1977). In transfer appropriate processing, performance is improved not only by the depth of processing but also by the extent that the format of initial encoding of information matches the format of later retrieval. Analogously, the current work suggests a kind of transfer appropriate processing in online social influence: the information about other's opinions needs to match the online judgment task.

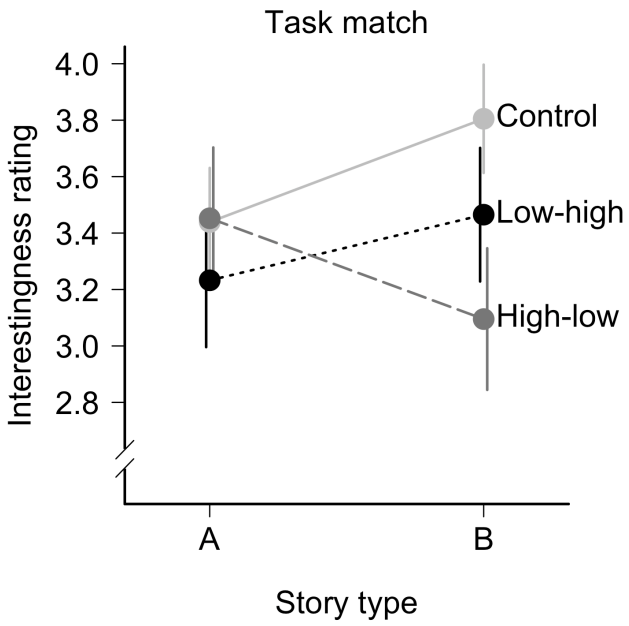


Figure 3: The low-high group, the high-low group, and the control group's interestingness ratings in the task match condition are shown for Story A and Story B. Error bars represent the 95% confidence intervals (Loftus & Masson, 1994). The low-high group thought Story A was rated less interesting than Story B. The high-low group thought the opposite. The control group had no information about the previous ratings.

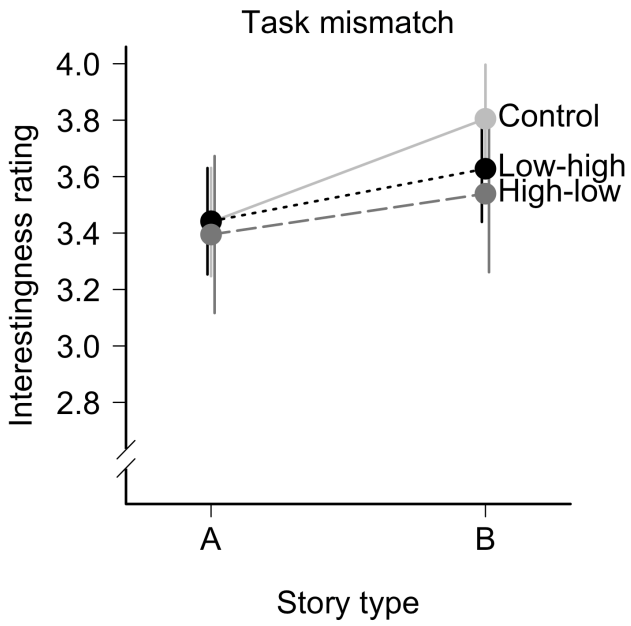


Figure 4: The low-high group, the high-low group, and the control group's interestingness ratings in the task mismatch condition are shown for Story A and Story B. Error bars represent the 95% confidence intervals (Loftus & Masson, 1994). The low-high group thought fewer people found Story A interesting than Story B. The high-low group thought the opposite. The control group had no information about the previous ratings.

The idea of alignability in analogy also has a bearing on the current findings. Two things are alignable when they have common dimensions. For example, knowing that the previous rating of a story is 4 and the task of rating the interestingness of a story using a 5-point scale are alignable. On the other hand, knowing the number of people who found the story interesting and the task of rating the interestingness of a story using a 5-point scale are nonalignable. Work in analogy has shown that alignability plays major roles in memory retrieval (Markman & Gentner, 1997) and preference formation (Zhang & Markman, 1998). The current work suggests that alignability also plays an important role in people's use of other's opinions.

Related to the idea of transfer appropriate processing and alignability, previous studies in social influence have shown that only other's decisions that are relevant have influences on decisions (Cialdini, 1998; Cason & Mui, 1998). Although the social information in both the task match condition and the task mismatch condition was relevant to the interestingness task, perhaps the participants did not regard the number of previous readers who found the stories interesting as appropriate information for the rating task.

The present study also provides useful information for users and designers of websites. Many people use online stores, such as eBay and Amazon, which provide information about collective opinion in the forms of reviews and ratings, and by listing the top selling items or the number of items available in stock. Knowing the responses produced by others can bias people's sampling of information (e.g., Lewandowsky et al., 2009; Stasser & Titus, 1985). The current results show that not all outputs by others influence people's behavior in the same fashion. The information about other's opinions needs to align with the response task to have significant influence on people's response. This is a note for the designers of social media websites, whether they want to encourage or minimize online social influence, as well as for marketers who want to take advantage of social media.

The users of social media websites may also find the current findings useful. Often times, users put too much attention to the opinions of others and too little attention to the actual content of the item (Sakamoto et al., 2009). By doing so they may be creating a trend for an item whose content is not so great. Knowing the present findings, users may be able to focus more on the content and less on what others think.

Perhaps people's desire to attend to the other's opinions survived for a good reason. Observing and imitating others

can allow people to try out solutions that they would not have considered otherwise (Bandura, 1965). Solutions selected by many people are often useful. Collective opinion of a community can be more informative than the opinions of a few experts (Surowiecki, 2004). Learning from the previous outputs of others is considered as a process for the creation of innovative solutions (Kraatz, 1998), the evolution of language (Smith et al., 2003), and the development of culture (Dennett, 1995).

In conclusion, we are surrounded by the opinions of others in online environments. Although online decisions are usually made privately and anonymously, online users influence and are influenced by their opinions as in offline environments. We need to know more about how people process collective opinion in online environments.

References

Asch, S. E. (1951). Effects of group pressure upon the modification and distortion of judgment. In H. Guetzkow (Ed.) *Groups, leadership and men*. Pittsburgh, PA: Carnegie Press, pp.177–190.

Asch, S. E. (1956). Studies of independence and conformity: A minority of one against a unanimous majority. *Psychological Monographs*, *70* (Whole no. 416).

Bandura, A. (1965). Behavioral modification through modeling procedures. In L. Krasner & L. P. Ulmann (Eds.), *Research in behavior modification: New development and implications* (pp. 310–340). New York: Rinehart and Winston.

Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, *286*, 509–512.

Berns G. S., Chappelow J., Zink C. F., Pagnoni G., Martin-Skurski M. E., Richards J. (2005). Neurobiological correlates of social conformity and independence during mental rotation. *Biological Psychiatry*, *58*, 245–253.

Cason, T., & Mui V.-L. (1998). Social influence in the sequential dictator game. *Journal of Mathematical Psychology*, *42*, 248–265.

Cialdini, R. B. (1998). Influence: The Psychology of Persuasion. Perennial Currents.

Cialdini, R. B., & Goldstein, N. J. (2004). Social influence: Compliance and conformity. *Annual Review of Psychology*, *55*, 591–621.

Dennett, D. C. (1995). *Darwin's dangerous idea*. New York: Touchstone.

Deutsch, M., & Gerard, H. B. (1995). A study of normative and informational social influences upon individual judgment. *Journal of Abnormal Social Psychology*, *51*, 629–636.

Festinger, L. (1954). A theory of social comparison processes. *Human Relations*, *7*, 117–140.

Glushko, R. J., Maglio, P. P., Matlock, T., and Barsalou, L. W. (2008). Categorization in the wild. *Trends in Cognitive Sciences*, *12*, 129–135.

Gureckis, T. M., & Goldstone, R. L. (2006). Thinking in groups. In S. Harnad & I. Dror (Eds.), *Distributed cognition: Special issue of pragmatics & cognition*, *14* (pp. 293–311). Amsterdam, The Netherlands: John Benjamins.

Kraatz, M. S. (1998). Learning by association? Interorganizational networks and adaptation to environmental change. *Academy of Management Journal*, *41*, 621–643.

Lewandowsky, S., Griffiths, M. L., and Kalish, M. L. (2009). The wisdom of individuals: Exploring people's knowledge about everyday events using iterated learning. *Cognitive Science*, *33*, 969–998.

Loftus, G. R., & Masson, M. E. J. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review*, *1*, 476–490.

Markman, A. B., & Gentner, D. (1997). The effects of alignability on memory. *Psychological Science*, *8*, 363–367.

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior*, *16*, 519-533.

Sakamoto, Y., Sadlon, E., & Nickerson, J. V. (2008). Bellwethers and the emergence of trends in online communities. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society*.

Sakamoto, Y., Ma, J., & Nickerson, J. V. (2009). 2377 people like this article: The influence of others' decisions on yours. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society*.

Salganik, M. J., Dodds, P. S., and Watts, D. J. (2006). Experimental study of inequality and unpredictability in an artificial cultural market. *Science*, *311*, 854-856.

Sherif, M. (1935). A study of some social factors in perception. *Archives of Psychology*, *27*, 1–60.

Smith, K., Kirby, S., and Brighton, H. (2003). Iterated learning: A framework for the emergence of language. *Artificial Life*, *9*, 371–386.

Stasser, G., & Titus, W. (1985). Pooling of unshared information in group decision making: Biased information sampling during discussion. *Journal of Personality and Social Psychology*, *48*, 1467–1478.

Surowiecki, J. (2004). *The wisdom of crowds: Why the many are smarter and how collective wisdom shapes business, economies, societies, and nations*. New York: Random House.

Tversky, A. & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*, 1124-1130.

Wooten, D. B., & Reed, A., II. (1998). Informational influence and the ambiguity of product experience: Order effects on the weighting of evidence. *Journal of Consumer Psychology*, *7*, 79–99

Zhang, S., & Markman, A. B. (1998). Overcoming the early entrant advantage via differentiation: The role of alignable and nonalignable differences. Journal of Marketing Research, 35, 413–426.

Enhancing Acquisition of Intuition versus Planning in Problem Solving

Dawn Chen (sdchen@ucla.edu)
Keith J. Holyoak (holyoak@lifesci.ucla.edu)
Department of Psychology
University of California, Los Angeles
Los Angeles, CA 90095 USA

Abstract

The acquisition of intuition, which guides problem solving by pruning unpromising strategies, is essential to the development of expertise in any domain. Problem-solving intuition may be viewed as analogous to search heuristics in artificial intelligence. One prediction inspired by this analogy is that practicing on subproblems and relaxed problems (versions of a problem with fewer constraints on the goal state and on the possible moves, respectively) may enhance the development of intuition for the full problem. Using the *n*-puzzle, we found that practice on relaxed problems did promote intuition compared to practice on the full problem, but impaired performance on solving the full problem. More detailed analyses suggest that practice on relaxed problems may discourage planning and encourage reliance on intuition. Planning is slower but more likely to produce optimal solutions if given enough time, whereas relying on intuition is faster but may lead to suboptimal solutions.

Keywords: Problem solving; intuition; planning; learning; heuristic search; *n*-puzzle.

Introduction

When encountering a problem in an unfamiliar domain for the first time, the novice may feel lost among what seems to be an indefinitely large number of possible actions that seem about equally promising, and end up pursuing some arbitrary path that leads nowhere. But after solving some number of problems from the same domain, the solver will eventually learn to consider only a limited number of approaches, those that are likely to prove effective. In commonsense terms, the learner has acquired *intuition* about the problem domain: an implicit sense of what to do in various types of situations that arise during problem solving (Gobet & Philippe, 2009). How is such intuition acquired through practice?

The standard account of general problem solving is Newell and Simon's (1972) proposal that the problem solver performs search within a problem space. A problem space can be visualized as a graph or tree in which the nodes represent possible states in the problem and each edge represents a legal move transforming one state into another. The legal moves in a problem are defined by its *operators*, or possible types of actions. The problem solver can search the problem space by starting at the node representing the initial state of the problem and moving to adjacent nodes by applying operators, until one of the nodes representing a goal state is reached. The solution to the problem is the successful path that the solver took through the problem-space graph.

Importantly, the problem solver may search the problem space not only by physically manipulating the external representation of the problem state (*external search*), but also by mentally transforming an internal representation (*internal search* or *planning*). During internal search, the problem solver need not always move from the current state to an adjacent node.

For most realistic problems, the problem-space tree is enormous, so that it is terribly inefficient even for a computer to solve the problem by using brute-force search algorithms that traverse the entire tree until a goal state is found. Heuristic search algorithms, on the other hand, are much more efficient because they use domain-specific knowledge to prune branches of the tree that never lead to the goal state or do not do so in an optimal way (i.e., in the minimum number of moves). A search heuristic may guide search by estimating the *distance* (minimum number of moves required) from any state to the goal so that, for example, a search algorithm can always choose to explore next the state that is closest to the goal (i.e., the greedy best-first search algorithm). This form of a search heuristic, commonly used in artificial intelligence, is called a *heuristic function*.

In many ways, the formal concept of a search heuristic is closely related to the commonsense concept of intuition in problem solving. Search heuristics prune branches in the problem-space tree that are unlikely to lead to the goal efficiently, just as problem-solving intuition focuses attention on just those paths that are likely to lead to a solution quickly. Search heuristics are usually fast to compute, but may lead to suboptimal solutions. Similarly, intuitive judgments arise quickly, but are fallible and may result in diminished accuracy or optimality compared to a solution strategy based on systematic analysis or careful planning. Furthermore, just as search heuristics rely on domain-specific knowledge, problem-solving intuition is restricted to a particular domain and is acquired only through multiple experiences with solving problems in that domain. Nonetheless, certain search heuristics are more general than others and apply to several domains with overlapping structure, just as the intuition gained from solving problems in one domain may apply to a related domain (see Hatano & Inagaki, 1986, for a discussion of routine vs. adaptive expertise). Finally, and most importantly for the present study, heuristic functions yield estimates analogous to the intuitive sense of closeness to the goal available to experienced problem solvers. The task we use to assess intuition will be based on subjective judgments of distance to the goal state.

The analogy between problem-solving intuition and search heuristics provides insights into how it might be possible to facilitate the acquisition of intuition in human problem solving. AI researchers have discovered that the solution lengths of *subproblems* and *relaxed problems* often provide good heuristic functions for the original problem (Prieditis, 1993). A subproblem removes one or more constraints on the goal state from the original problem, whereas a relaxed problem removes one or more constraints on the legal moves (i.e., it adds one or more operators). Thus, an instance of the original problem can be solved in fewer moves when translated into a corresponding subproblem or relaxed problem.

Applying the results from AI to the domain of human problem solving, solving subproblems and relaxed problems may facilitate the acquisition of intuition for the original problem. Therefore, learners who practice solving subproblems or relaxed problems may acquire better intuition for the original problem than those who receive the same amount of practice on only instances of the original problem. At the same time, planning may seem less necessary when solving subproblems and relaxed problems. Thus, the kind of learning experience that fosters development of intuition the most may also have a detrimental impact on planning. We will elaborate on these points in discussing our experimental findings.

Method

Participants

Seventy-two undergraduates from the University of California, Los Angeles participated for course credit. Participants were randomly assigned to either the control condition (n = 24), the subproblem condition (n = 24), or the relaxed problem condition (n = 24).

Materials

The *n*-puzzle Participants solved a computer version of the *n*-puzzle, which is illustrated in Figure 1. The *n*-puzzle consists of a square bounded space containing a smaller empty square and *n* initially misplaced square tiles numbered 1 to *n*. A legal move consists of sliding any tile into the empty square, and the goal state contains all the tiles in ascending order.

4	1	3
	2	5
7	8	6

initial state

1	2	3
4	5	6
7	8	

goal state

Figure 1: An 8-puzzle with a 5-step solution: Move 4 down, 1 left, 2 up, 5 left, and 6 up.

Subproblems and Relaxed Problems In the subproblems for the *n*-puzzle, participants were required to move only some of the tiles into their correct places. In the relaxed problems, participants could swap some of the tiles with adjacent tiles, in addition to sliding any tile into the empty square. These *swappable* tiles were displayed in a lighter color than the non-swappable tiles. Defined in this way, a subproblem that removes k goal constraints requires moving tiles 1 through $n - k$ into their correct places, and a relaxed problem that removes k move constraints contains one empty square and k tiles that can be swapped with neighboring tiles.

Generation of Puzzles All puzzles were generated randomly. The optimal A* search algorithm was used to ensure that each puzzle had the desired minimum solution length.

Procedure

All instructions and stimuli were presented on a computer, and participants responded using a mouse. In each condition, the participant was first given instructions on how to solve the type of puzzles (full, subproblem, or relaxed problem) in that condition. The participant then attempted to solve an initial 8-puzzle of the appropriate type, solvable in a minimum of three moves. An experimenter ensured that the participant understood the instructions and could solve the initial puzzle. In the subproblem condition, the initial puzzle required tiles 1-4 to be moved into place. In the relaxed problem condition, tiles 5-8 were swappable. That is, the number of constraints removed, k, was four for the initial puzzle in both the subproblem and relaxed problem conditions. After solving the initial puzzle, the participant took part in a training phase, a test phase, and finally an intuition assessment phase.

Training Phase The participant was told that more puzzles would now be given for practice, with a time limit of one minute and 30 seconds for each. The participant was told to solve each puzzle in as few moves as possible, and that there would be a penalty for every extra move made. These instructions were designed to discourage external search (the usual strategy for solving *n*-puzzles) and encourage internal search, which has been shown to enhance learning (O'Hara & Payne, 1998).

The participant then attempted to solve a sequence of 12 8-puzzles. In all conditions, the minimum solution lengths (a measure of difficulty) of the puzzles increased from 4 to 10 (i.e., the puzzles in the experimental conditions were not subproblem or relaxed versions of those in the control condition). In the experimental conditions, k also decreased from four to zero across the puzzles. During the presentation of each puzzle, the minimum solution length and the number of moves the participant had made so far were shown above the puzzle. After the participant solved each puzzle or the time limit expired for that puzzle, a dialog box informed the participant which event had occurred, the number of extra moves the participant made (if the puzzle was solved), and in the subproblem condition, the tiles to

slide into place for the next puzzle. The participant could then take a break and click on a button to start the next puzzle when ready.

Test Phase After all 12 puzzles in the training phase had been presented, participants were told that there would now be a test, with the same instructions as for the practice puzzles. In the subproblem condition, participants were told to slide all tiles into place. Participants in all conditions then attempted to solve the same sequence of 12 full *n*-puzzles. The first six were 8-puzzles and the last six were 15-puzzles, and all puzzles could be solved in 12 moves. After each puzzle had been solved or had timed out, the next puzzle was presented without any feedback or time to rest. During both the training and test phases, the computer recorded for each puzzle whether it was solved, the solution time, the moves the participant made, the initial latency (the amount of time the participant took to make the first move), and the inter-move latencies (the time to make each subsequent move).

Intuition Assessment Phase After the test phase, participants made a series of 40 *pairwise distance comparisons*. In each comparison, they were presented with two different puzzle states and had to click on the one that they believed was closer to the goal within a short time limit. No feedback was given. The short time limit was designed to elicit a quick, intuitive judgment and prevent participants from solving the puzzles mentally and then counting the number of moves used. Because experts in a domain often have an intuitive sense of how close they are to solving a problem, and heuristic functions estimate the distance of any given state to the goal, this distance comparisons task serves to assess participants' intuition on the *n*-puzzle.

The first 20 pairs to be compared were 8-puzzles, with 10 seconds each, and the last 20 pairs were 15-puzzles, with 12 seconds each. The true distances of the puzzles ranged from 1 to 28, and the ratio of the shorter distance to the longer distance in each pair was between .2 and .91. For each comparison, which puzzle was chosen and the time taken to make that choice were recorded.

Results and Discussion

Dissociation of Performance on Solving Puzzles and Comparing Distances

The mean percentage of full *n*-puzzles solved during the test phase in each condition is shown in Figure 2. The relaxed problem group solved a significantly lower percentage of puzzles during the test phase (M = 57.99, SD = 23.25) than the control group (M = 69.79, SD = 14.08), $F(1, 69) = 5.18$, $p = .026$, and also the subproblem group (M = 68.75, SD = 15.20), $F(1, 69) = 4.30$, $p = .042$. The latter two groups did not differ reliably.

However, as shown in Figure 3, the relaxed problem group correctly solved the most problems on the distance comparisons task, which assesses intuition. The percentage

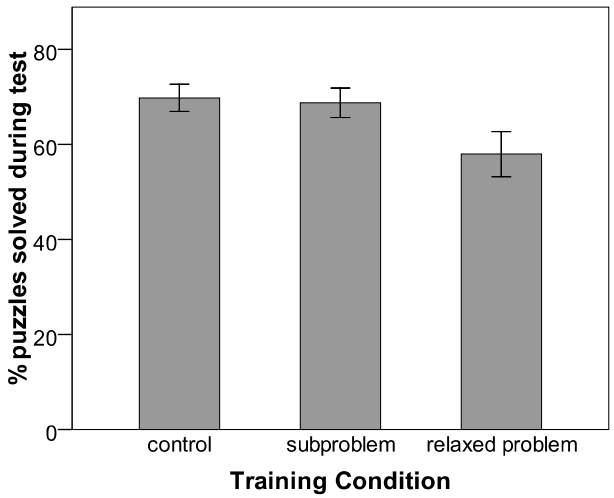


Figure 2: Mean percentage of *n*-puzzles solved by participants in each training condition during the test phase. Error bars in all data figures represent 1 standard error of the mean.

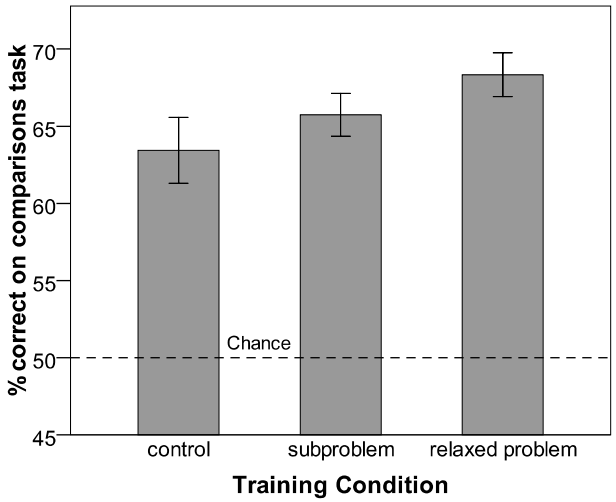


Figure 3: Mean percentage of comparisons solved correctly on the distance comparisons task in each condition.

of comparisons correct was significantly higher for the relaxed problem group (M = 68.33, SD = 6.94) than for the control group (M = 63.44, SD = 10.47), $F(1, 69) = 4.22$, $p = .044$. Performance of the subproblem group on the comparisons task fell between that of the other two groups, but did not differ significantly from either.

To further investigate the difference in performance on the distance comparisons task, we divided the pairwise distance comparisons into an "easy" set and a "hard" set based on the overall performance of the participants on each comparison. For each comparison problem, we calculated the proportion q of participants (over all three conditions) who solved that problem correctly. We then calculated the median value of q over all comparisons. A comparison that

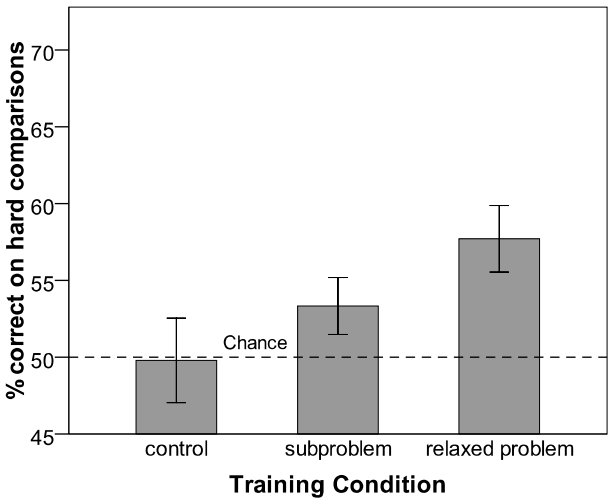


Figure 4: Mean percentage of hard comparisons solved correctly on the distance comparisons task in each condition.

had a *q*-value higher/lower than the median was assigned to the "easy"/"hard" set. All groups performed about the same on the easy comparisons, but as Figure 4 shows, the relaxed problem group performed the best on the hard comparisons. In particular, the relaxed problem group correctly solved a significantly higher percentage of the hard comparisons (M = 57.71, SD = 10.63) than the control group did (M = 49.79, SD = 13.47), $F(1, 69) = 6.00$, $p = .017$. Thus, the relaxed problem group performed very well on the intuition task, especially the harder problems, compared to the control group.

How could participants in the relaxed problem group have apparently acquired such good intuition on the full *n*-puzzle, and yet perform relatively poorly in actually solving it? A possible explanation is that because planning (internal search) is harder and seemingly less necessary when solving the relaxed problems, participants in the relaxed problem group learned to plan less and rely more on their intuition during the training phase. Thus, even though their intuition became more developed (as evidenced by their performance on the distance comparisons task), their decreased use of planning caused them to perform poorly on solving the puzzles in the test phase. Participants in the control group, on the other hand, learned to rely more on planning and less on their intuition during the training phase, because they were trying to minimize the number of moves they made and it was easier for them to plan. Increased planning led them to perform better on the test puzzles, but their intuition was less developed. We will now present evidence to support each of these claims.

The Relaxed Problem Training Condition Discourages Planning

Planning is Harder on Relaxed Problems This is true for two reasons. First, internally visualizing the move of swapping two tiles in the relaxed problem imposes a greater working memory load, because the participant must now keep track of the new locations of both tiles, rather than just one tile in the sliding move. Manipulating an internal representation of the puzzle state to reflect a swapping move might take longer as well. Second, the introduction of additional legal moves in the relaxed problem also makes planning harder because participants have to consider more moves at each state (that is, the *branching factor* is higher). In order to plan, participants must also remember more information about which paths they have already mentally explored to some depth and have determined to be unpromising.

The hypothesis that the swapping move consumes more working memory is supported by the finding that the average length of unbroken sequences of backtracking moves during the training phase was significantly lower in the relaxed problem group (M = 1.34, SD = .36) than in the control group (M = 1.88, SD = 1.11), $F(1, 61) = 4.66$, p = .035, and also the subproblem group (M = 2.04, SD = .82), $F(1, 61) = 8.93$, $p = .004$. In contrast, no reliable differences among conditions were observed in the test phase. Backtracking for a number of moves requires remembering all those previous moves, and participants solving relaxed problems may have backtracked for fewer moves because they could not remember as many past moves, since storing a single move requires more working memory capacity on average.

Planning Seems Unnecessary on Relaxed Problems Because relaxed problems have a higher branching factor, the problem-space graphs for relaxed problems are more connected and so there are more ways to reach the goal state. Thus, it may seem unnecessary to plan one's moves before executing them, since no matter how far away one wanders from the goal, there is always some way to get back onto the right track. In other words, local minima do not exist in the problem space, so a greedy (hill-climbing) search algorithm that always chooses the state with the shortest estimated distance to the goal to explore next cannot become trapped, and is thus sufficient. Accordingly, participants in the relaxed problem group probably learned to use a greedy search algorithm, which does not look ahead and thus requires little effort. Moreover, a greedy search algorithm relies heavily on the heuristic function, so its use would foster development of intuition for participants in this condition.

One piece of evidence that participants in the relaxed problem group planned less than those in the other conditions is that they made extra moves more often during the training phase. The percentage of solved puzzles in the training phase that were solved with extra moves was significantly higher in the relaxed problem group (M = 49.99, SD = 19.76) than in the control group (M = 20.92, SD = 13.67), $F(1, 69) = 34.53$, $p < .001$, and also the subproblem group (M = 25.97, SD = 17.43), $F(1, 69)$ = 23.58, $p < .001$. Furthermore, the relaxed problem group had significantly higher average solution times during the

training phase (M = 35.23s, SD = 9.07s) than did the control group (M = 27.68s, SD = 8.29s), $F(1, 69) = 9.86$, $p = .002$, and also the subproblem group (M = 24.64s, SD = 7.53s), $F(1, 69) = 19.41$, $p < .001$. Participants in the relaxed problem condition may have found planning harder and thus took longer on average to plan a single move (when they did plan); in addition, their longer, less optimal solutions took more time to execute. These differences indicate that the relaxed problem participants did not or could not plan as far ahead as did the participants in the other conditions, and tended to meander around the problem space for a while before reaching the goal.

The average initial latency on a puzzle, or the average amount of time a participant spent thinking before making the first move on a puzzle, is a clear indicator of how much a participant plans voluntarily. (While the average inter-move latency is also an indicator of planning, higher inter-move latencies could also indicate that the participant was stuck in the middle of solving a puzzle and was forced to think carefully about what to do next.) The average initial latency was not significantly lower for the relaxed problem group during the training phase, as might be expected if these participants were planning fewer moves ahead; however, the lack of a difference could reflect the offsetting effect of planning each move being harder for the relaxed problems and thus taking longer. During the test phase, when all participants were solving the full n-puzzles, the average initial latency was indeed significantly lower for the relaxed problem group (M = 10.37s, SD = 4.46s) than for the control group (M = 14.75s, SD = 6.02s), $F(1, 69) = 7.33$, $p = .009$, indicating that the relaxed problem group continued to plan fewer moves ahead during the test phase.

Increased Planning is Associated with Better Puzzle-Solving Performance

Not surprisingly, increased planning is associated with better puzzle-solving performance. The average initial latency was not correlated with the number of puzzles solved during the training or test phase, perhaps because some participants tended to get stuck at the very beginning and could not solve many puzzles, or were just too slow in general to solve many puzzles. However, average initial latency was negatively correlated with performance measures such as the average number of extra moves made on solved puzzles [$r(70) = –.37$, $p = .002$ for the training phase and $r(70) = –.46$, $p < .001$ for the test phase], and the percentage of backtracking moves [$r(70) = –.26$, $p = .026$ for the training phase and $r(70) = –.31$, $p = .007$ for the test phase]; and positively correlated with the percentage of moves that decreased the true distance of the problem state to the goal [$r(70) = .33$, $p = .005$ for the training phase and $r(70) = .47$, $p < .001$ for the test phase]. These results indicate that the more the participant planned before making the first move, the better the moves the participant made later on.

Recall that on relaxed problems, which do not have many local minima, a greedy search algorithm is sufficient. However, greedy search may get stuck in local minima on the full n-puzzle, for which the problem-space graph is not as well-connected. Accordingly, if participants in the relaxed problem group did indeed use a greedy search algorithm, they would perform poorly during the test phase. The control group, on the other hand, may have learned to use a more effective search algorithm involving greater look-ahead. Such a search algorithm could achieve an acceptable level of performance with a relatively poor heuristic function. Thus, participants in the control condition would not acquire intuition during the training phase to the degree that those in the relaxed problem group did.

Planning and Intuition are Dissociated

For every participant, we calculated a composite score on the intuition task by summing the values of $1 – q$ for all comparison problems that the participant solved correctly. Recall that for each comparison, q is the proportion of all participants who solved that comparison correctly. Thus, $1 – q$ is the estimated probability of choosing the incorrect response on a given comparison, an empirical measure of its difficulty. Therefore, the composite score on the intuition task gives greater weight to more difficult problems.

We calculated correlations between the composite intuition score and measures of planning for each training condition separately to test whether planning and intuition are dissociated within each group. The following correlations appeared for measures of planning during the training phase: The composite intuition score for the control group was negatively correlated with the average initial latency, $r(22) = –.41$, $p = .047$, as well as the average inter-move latency, $r(22) = –.47$, $p = .021$. For the subproblem group, the composite intuition score had a negative correlation with the average inter-move latency, $r(22) = –.50$, $p = .013$, and a near-significant positive correlation with the percentage of puzzles that were solved with extra moves, $r(22) = .40$, $p = .055$. Finally, for the relaxed problem group, there was a weak negative correlation between the composite intuition score and the percentage of moves that decreased the true distance of the problem state to the goal, $r(22) = –.35$, $p = .098$.

During the test phase, the composite intuition score for the control group had a near-significant negative correlation with the average initial latency, $r(22) = –.39$, $p = .061$, as well as a slight positive correlation with the average number of extra moves, $r(22) = .36$, $p = .082$.

These findings indicate that participants in our study mainly took one of two approaches to solving the puzzles and comparison problems. One was a more analytic or algorithmic approach based on planning, and the other was a more holistic or heuristic approach based on intuition. While the first approach was more effective for solving the full n-puzzles, the second approach was more effective on the task requiring speeded comparison of distances to the goal state. The control training condition encouraged the more analytic problem-solving style, and participants in this

condition developed a more effective search algorithm. In contrast, the relaxed problem training condition encouraged the more intuitive problem-solving style, and participants in this condition developed a more accurate heuristic function.

Conclusions

The present study demonstrates a dissociation between two core mechanisms on which expertise in problem solving depends: internal search (planning) and use of a heuristic function to evaluate locally available moves (intuition). Training on problems with fewer possible moves at each choice point (full *n*-puzzles) encouraged a more analytic problem-solving style, whereas training on relaxed versions of the same problem type that allow more possible moves encouraged a more intuitive problem-solving style. In the present study, the analytic style led to better performance on actually solving the full *n*-puzzles, but the more intuitive style led to better performance on a task requiring fast evaluations of how close a problem is to being solved.

Our results should not be construed as evidence that the development of analytical thinking and intuition are mutually exclusive. In fact, true experts in solving problems in complex domains such as chess (Chase & Simon, 1973; Gobet & Charness, 2006) appear to rely heavily on both intuition and planning, with the relative importance of intuition increasing when performance is time-constrained (Gobet & Simon, 1996). The time frame of the present study was far shorter than the years required to develop true expertise (Ericcson, 1996). Even by the end of the experiment, our participants remained novices on the *n*-puzzle. An expert solver of the *n*-puzzle would no doubt plan ahead more as well as make better intuitive judgments relative to a novice. The ability to quickly evaluate problem states should allow the problem solver to plan more moves ahead, just as heuristic functions reduce the branching factor and thus allow the search algorithm to search to a greater depth within the same amount of time. In fact, Charness (1981) found that skilled chess players search more deeply than novice players do, indicating that good intuition aids planning in problem solving.

What our findings do indicate is that these two basic approaches to problem solving may not be acquired in lock-step fashion, and to some extent constitute competing problem-solving strategies. Moreover, the two different approaches may be maximally effective for different types of problems. The systematic, analytic approach is slower and places a greater burden on working memory, but is more likely to lead to optimal solutions, and thus may be preferable for problems that can be solved slowly with the assistance of external aids to memory. In contrast, the holistic, intuitive approach is faster and less dependent on working memory, and hence will often be preferable when the problem must be solved under severe time constraints.

One example of this dichotomy is battlefield versus hospital triage. In the hospital, medical personnel may take a more analytic approach, carefully considering the consequences of each possible action. On the battlefield, by contrast, the need for decisions may be so urgent that the only possible approach is to rely on intuition or "gut feelings." An important direction for future research will be to determine whether the present findings using the toy example of the *n*-puzzle in fact generalize to real-world problem solving (cf. Gobet & Philippe, 2008).

Acknowledgments

This research was funded by a University Fellowship and a Chancellor's Prize from the Graduate Division at the University of California, Los Angeles, and by ONR grant N000140810186.

References

Charness, N. (1981). Search in chess: Age and skill differences. *Journal of Experimental Psychology: Human Perception and Performance*, *7*, 467–476.

Chase, W. G., & Simon, H. A. (1973). The mind's eye in chess. In W. G. Chase (Ed.), *Visual information processing*. New York: Academic Press.

Ericsson, K. A. (1996). The acquisition of expert performance: An introduction to some of the issues. In K. A. Ericsson (Ed.), *The road to excellence: The acquisition of expert performance in the arts and sciences, sports, and games*. Mahwah, NJ: Erlbaum.

Gobet, F., & Charness, N. (2006). Expertise in chess. In K. A. Ericsson, N. Charness, P. J. Feltovich, & R. R. Hoffman (Eds.), *The Cambridge handbook of expertise and expert performance*. New York: Cambridge University Press.

Gobet, F., & Philippe, C. (2008). Towards an alternative to Benner's theory of expert intuition in nursing: A discussion paper. *International Journal of Nursing Studies*, *45*,129-139.

Gobet, F., & Philippe, C. (2009). Expertise and intuition: A tale of three theories. *Minds and Machines*, *19*, 151-180.

Gobet, F., & Simon, H. A. (1996). The roles of recognition processes and look-ahead search in time-constrained expert problem solving: Evidence from grandmaster-level chess. *Psychological Science*, *7*, 52-55.

Hatano, G., & Inagaki, K. (1986). Two courses of expertise. In H. Stevenson, H. Azuma, & K. Hakuta (Eds.), *Child development and education in Japan*. San Francisco: Freeman.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

O'Hara, K. P., & Payne, S. J. (1998). The effects of operator implementation cost on planfulness of problem solving and learning. *Cognitive Psychology*, *35*, 34-70.

Prieditis, A. E. (1993). Machine discovery of effective admissible heuristics. *Machine Learning*, *12*, 117-141.

Infants Expect Others to Help One Another Achieve a Goal

Woo-yeol Lee (wlee@yonsei.ac.kr)
Department of Psychology, Yonsei University, 134 Shinchon-dong, Seodaemun-gu, Seoul, Korea

Eun Young Kim (majilake@gmail.com)
Department of Psychology, Yonsei University, 134 Shinchon-dong, Seodaemun-gu, Seoul, Korea

Jeong-ae Won (bloom1015@naver.com)
Department of Psychology, Yonsei University, 134 Shinchon-dong, Seodaemun-gu, Seoul, Korea

Yoonha Lee (una3263@hotmail.com)
Department of Psychology, Yonsei University, 134 Shinchon-dong, Seodaemun-gu, Seoul, Korea

Yoon Kim (alsey07@naver.com)
Department of Psychology, Yonsei University, 134 Shinchon-dong, Seodaemun-gu, Seoul, Korea

Abstract

What makes people help each other? To explore the origin of human altruism, we tested whether 16-month-old infants have an expectation of helping behavior when they observe an interaction between others. Infants watched videos in which one (capable) agent had achieved a goal while the other (incapable) one could not. In a subsequent situation, the capable agent either helped the incapable agent achieve the goal (helping event), or ignored the incapable agent and achieved the goal alone (ignoring event). Infants looked longer at the ignoring event, suggesting that they expected helping behavior rather than ignoring behavior. The results are discussed in terms of infants' understanding of the connection between goals and altruistic behaviors.

Keywords: altruism, helping behavior, violation of expectation paradigm, goal understanding, infants

Introduction

In everyday life, we often help others not necessarily expecting rewards in return. We willingly donate money to charities when we hear news about people on the other side of the earth suffering from hunger and distress due to a tragic natural disaster. We hear about doctors and rescue teams rushing into places of catastrophe to save others' lives. These behaviors cannot be explained from an economic perspective because expending resources without profits could be viewed as irrational. What makes people benevolent toward others? The origin of human altruism has been a major interest of philosophers for a long time. Recently, developmental psychologists have begun to examine infants in order to discover the development of human altruism.

Recent research with toddlers and infants demonstrates that they take some actions to help others under certain circumstances. In Warneken and Tomasello (2006), for example, when 18-month-old children observed that an adult accidentally dropped a marker pen, they picked up the pen and brought it to the adult. Infants do such behaviors spontaneously without external rewards. Another study showed that children's motivation to help others was in fact *decreased* by material rewards (Warneken & Tomasello, 2008). Meanwhile, it is difficult for younger children to give instrumental aid to others through actions because they have yet to master control of their bodies. Nevertheless, there is some evidence that even 12-month-old infants give relevant information to others by using pointing actions (Liskowski, Carpenter, Striano, & Tomasello, 2006; Liskowski, Carpenter, & Tomasello, 2008).

Infants also discriminate helping behaviors from hindering behaviors (Hamlin, Wynn, & Bloom, 2007; Kuhlmeier, Wynn, & Bloom, 2003). In studies by Kuhlmeier and her colleagues, 12-month-old infants watched a series of computer-animated videos including a social interaction between geometric shapes. In the videos, an agent (e.g., a triangle) helped a circle climb up a hill, whereas another agent (e.g., a square) hindered the circle from climbing the hill. In the following test trials, infants observed scenes in which the circle approached one of the two agents. The looking time of the infants was longer when the circle approached the helper than hinderer. This result indicates that infants are able to make a distinction between a helper and a hinderer. In addition, infants themselves show preference for agents who have helped others over agents who have not (Hamlin et al., 2007).

In summary, infants often show and prefer helping behaviors and distinguish helpers from hinderers. The present study further investigated infants' expectation of others' helping behaviors. More specifically, it asked: Do infants anticipate someone would help another when that other is in trouble or need? For instance, consider the following situation. A person sees another person repeatedly fall down while hiking. We may expect the first person to offer some help to the second person. If the first person simply passes by the second, we may be surprised.

The current research examined what 16-month-old infants expect of an agent when they watch a similar

situation. We employed the violation of expectation paradigm using computer-animated videos as stimuli (see Figure 1). The violation of expectation paradigm measures infants' looking time patterns to evaluate their reasoning about an event, where infants show longer looking times for surprising or unexpected scenes (e.g., Gergely, Nadasdy, Csibra, & Biro, 1995; Onishi & Baillargeon, 2005; Song, Baillargeon, & Fisher, 2005; Woodward, 1998). The infants were randomly assigned to either the experimental or control condition.

In the experimental condition, infants first received familiarization trials in which they watched videos about two agents, a square and a circle. The videos showed that the square was able to achieve the goal of climbing a tall hill whereas the circle was not. During test trials, the infants watched two events. In the helping event, the square helped the circle achieve the goal of climbing the hill by pushing the circle to the top of the hill. In the ignoring event, the square did not help the circle; it simply passed by the circle as if completely ignoring the circle striving to climb the hill. If infants expect the square to help the circle, they should look longer at the ignoring event than at the helping event because their expectation would be violated in the ignoring event.

To rule out the possibility that infants would look longer at the ignoring event than at the helping event simply because the agents' movements are more interesting or perceptually salient in the ignoring event, another group of infants were tested in the control condition. The control condition was identical to the experimental condition except that the circle did not show an intention to climb the tall hill during the familiarization trials. Instead, it simply moved around aimlessly. If infants reason that the circle does not have the goal of climbing the tall hill, and thus that the square does not have to help the circle achieve the goal, infants should look for equal amounts of time at the helping and ignoring events. However, if the ignoring event is simply more interesting than the helping event, infants in both the control and experimental conditions should look longer at the ignoring event than at the helping event.

Experiment

Participants

A total of 31 infants initially participated in the study. However, 7 infants were excluded from the data analyses because of parental interference (1), distraction (1), experimental error (2), and fussiness (3). So, 24 16-month-old infants (12 boys, 12 girls, M = 16;12, range 15;8 – 17;22) were kept for data analyses. They were randomly assigned to the experimental condition or the control condition.

Materials and procedure

Figure 1 shows examples of the stimulus videos. In the videos, a red circle and a yellow square-like geometric shapes climbed small and tall hills. The shapes had some personifying features, i.e., eyes and a nose.

In the experimental condition, the infants received 4 trials during the familiarization phase. In the first two trials, only the square was in the video and infants watched it climb the two hills successfully.

At the beginning of the third and fourth familiarization trials, the square was on top of the tall hill and the circle was at the bottom left corner of the scene. The circle approached the small hill and successfully climbed it. It then tried, but failed, to climb the tall one—it moved up the tall hill until it reached the middle, slid down, and ended up stuck between the two hills. It attempted to climb the tall hill twice more, but continued to fail. The square watched all of these attempts from the top of the tall hill.

In the following test phase, infants received 2 test trials comprising the helping and ignoring events. At the beginning of each trial, infants saw a static scene in which the square was now at the bottom left corner of the scene and the circle was stuck between the two hills. In the helping event, the square pushed the circle up the tall hill and they successfully reached the top together. In the ignoring event, by contrast, the square simply passed by behind the circle and climbed up the tall hill alone, as if ignoring the circle.

In the control condition, the infants watched videos that were identical to those in the experimental condition, with the exception of the movement of the circle in the third and fourth familiarization trials. At the beginning, the circle was at the bottom middle of screen, between the hills. The circle rolled only half up the tall hill, and then reverted to the valley. After that, it moved to the left corner of the scene over the small hill and returned to the original place. The circle stopped at the valley between the two hills. Thus, the circle did not show the intent to climb the tall hill.

The duration of each video was 6 seconds, and these videos were played repeatedly until the end of each trial. Each trial ended if the infants looked away from the monitor for 2 consecutive seconds after watching at least 6 cumulative seconds, or if they looked at the videos for 60 cumulative seconds.

Half of the infants in each condition saw the helping event first, and half saw the ignoring event first. Infants sat on a parent's lap, approximately 45 cm away from a 20-inch computer monitor. The parents were asked to close their eyes and remain silent during the experiment.

Two observers monitored each infant's looking behavior through peepholes in cloth-covered frames on either side of the apparatus. The primary observer's responses determined the end of each trial. Interobserver agreement averaged 93% per trial per infant.

Results

The infants' looking times during the familiarization and test trials were analyzed. Preliminary analyses did not reveal any effect of gender or order of test events

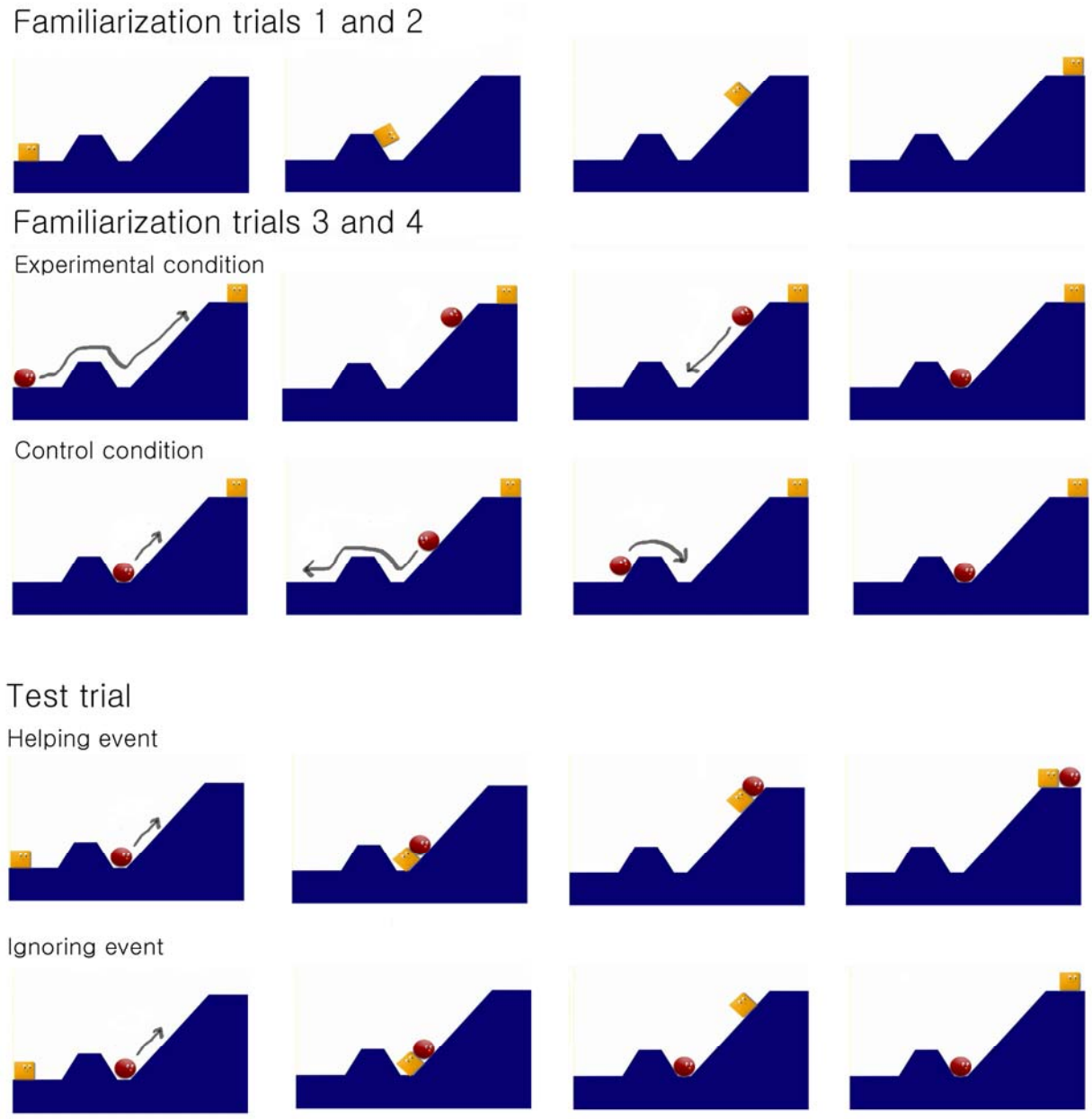


Figure 1: Video stimuli used in the experiment.

(helping event first or ignoring event first) for the looking times during these trials, *F*s(1, 16) < 3.34, *p*s > .086. Therefore, these factors were collapsed in further analyses.

During the 4 familiarization trials, the mean looking time of the infants was 23.9 seconds (*SD* = 10.4) in the experimental condition and 23.1 seconds (*SD* = 11.1) in the control condition. A single-factor analysis of variance (ANOVA) with condition (experimental or control) as a between-participants factor demonstrated no main effect of condition, $F(1, 22) < 1$, indicating that the infants in the two conditions did not significantly differ in their mean looking times during the familiarization trials.

The infants' looking times during the test trials were analyzed with a 2 X 2 ANOVA with condition (experimental or control) as a between-participants factor and event (helping or ignoring) as a within-participants factor (see Figure 2). The results revealed no significant main effect of condition or event, *F*s(1, 22) < 1. However, the interaction between condition and event was significant, $F(1, 22) = 5.26$, $p < .05$. A planned comparison indicated that the infants in the experimental condition looked reliably longer at the ignoring event (*M* = 34.0 seconds, *SD* = 19.6) than at the helping event (*M* = 23.2 seconds, *SD* = 15.7), $F(1, 22) = 4.77$, $p < .05$, whereas those in the control condition did not show a difference in looking times between the events (ignoring event, *M* = 24.3 seconds, *SD* = 18.1; helping event, *M* = 30.2 seconds, *SD* = 19.8), $F(1, 22) = 1.22$, $p > .2$.

A non-parametric Wilcoxon signed-ranks test revealed the same pattern as above. In the experimental condition, 11 of the 13 infants looked longer at the ignoring event than at the helping event ($Z = 2.13$, $p < .05$), whereas in the control condition, 4 of 11 infants looked longer at the ignoring event than at the helping event and one of them looked equally at both events, $Z = .66$, $p > .5$.

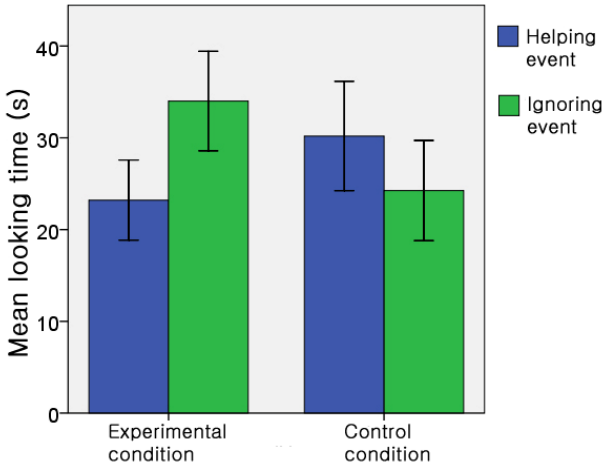


Figure 2: Mean looking times of the infants during the test trials. Error bars denote standard error.

Discussion

This study demonstrates that 16-month-old infants expect an agent to be helpful when the agent sees another in need of aid. In the experimental condition, infants looked reliably longer at the ignoring event than at the helping event. This result indicates that the infants expected the competent agent to help the less competent agent achieve the less competent agent's goal.

The infants in the control condition, by contrast, looked for comparable durations at the helping and ignoring events. The results of the control condition confirmed that the results of the experimental condition are not simply due to infants' perceptual preference for movement of the agents in the ignoring event. Note that infants' understanding of basic physics such as *solidity* and *continuity* emerges very early in life (Spelke, 1994). Therefore, the ignoring event could have been more interesting simply because it seems to defy a law of physics, i.e., that solid objects cannot "pass through" one another. However, this possibility was not the case because the infants in the control condition did not show the difference in their looking times between the events. The only difference between two conditions was the motion of the circle during the third and fourth familiarization trials. The circle showed an intention to climb the hill only in the experimental condition. Thus, the infants could have expected that the square would help the circle in the experimental condition, but not in the control condition. The square pushing the circle to the top of the tall hill hence could have been viewed as helping the circle achieve the goal in the experimental condition. In contrast, the same motion in the control condition could not have been viewed as helpful because climbing the hill was not the circle's demonstrated goal.

The present study thus supports and extends previous studies that investigated infants' understanding and showing of helping behaviors. According to previous findings, infants show spontaneous helping behavior (Warneken & Tomasello, 2006), distinguish helpers from hinderers (Kuhlmeier, Wynn, & Bloom 2003), and prefer helpers to hinderers (Hamlin, Wynn, & Bloom, 2007). In addition, our findings suggest that infants expect an agent to willingly help, rather than neglect, others. In our study, infants expected to see helping behavior even though (1) they did not observe interactions between the agents before the test trials, and (2) they were not informed about the characteristics of the agents beyond the agents' competence to achieve the goal.

Furthermore, our findings extend previous findings that infants of this age can attribute goals to nonhuman agents. Previous research has found that infants are able to notice the goal of a nonhuman agent when several cues to animacy are provided (Biro & Leslie, 2007; Luo & Baillargeon, 2005). In Biro and Leslie (2007), for instance, 9-month-old infants can reason what an object's goal is when it moves freely, as though its movements are being directed by its free will. In our experiment, agents' actions through self-propelled movements and personifying features such as eyes and a nose may have helped the infants detect goals of the agents.

Our results also suggest that infants can infer an agent's goals or intentions even when it fails to accomplish the goal. That is, infants in the experimental condition did not see the circle reach the top of the hill during the familiarization trials, but they were able to infer the goal of the circle. The findings are consistent with previous evidence that infants can infer an agent's goal when observing others' failed actions (Bradone & Wellman, 2009; Hamlin, Newman, & Wynn, 2009).

What do the current findings suggest about the developmental origin of human altruism? Where does the expectation about others' helpful actions come from? On the one hand, the propensity to expect helping behavior could be acquired from interaction with others. Attachment with parents in infancy may especially influence the development of their social models. A recent study suggests that 12- to 16-month-old infants have different expectations of others' behavior in a social context depending on the infants' experiences with their mothers (Johnson, Dweck, & Chen, 2007). On the other hand, the possibility exists that the expectation of helping behavior is an innate tendency since 16-month-old infants are not old enough to have had extensive social interactions in groups. In either case, our findings suggest that the expectation of altruistic behavior emerges in a very early period of human life. To further investigate the root of this altruistic mechanism, future studies can examine the relationship between these results and social factors such as parenting styles, daycare systems, or presence of siblings.

Acknowledgment

This work was supported by the Students' Association of the Graduate School of Yonsei University and by the Brain Korea 21 Project in 2010.

References

Biro, S., & Leslie, A. M. (2007). Infants' perception of goaldirected actions: Development through cues-based bootstrapping. *Development Science, 10,* 379–398.

Brandon, A. C., & Wellman, H. M. (2009). You can't always get what you want: Infants understand failed goal-directed actions. *Psychological Science, 20,* 85-91.

Gergely, G., Nadasdy, Z., Csibra, G., & Biro, S. (1995). Taking the intentional stance at 12 months of age. *Cognition, 56,* 165–193.

Hamlin, J. K., Newman, G, E., & Wynn, K. (2009). Eight-month-old infants infer unfulfilled goals, despite ambiguous physical evidence. *Infancy, 14,* 579-590.

Hamlin, J. K., Wynn, K, & Bloom, P. (2007). Social evaluation by preverbal infants. *Nature, 450,* 557-559.

Johnson, C., Dweck, C. S., & Chen, F. S. (2007). Evidence for infants' internal working models of attachment. *Psychological Science, 18,* 501-502.

Kuhlmeier, V., Wynn, K., & Bloom, P. (2003). Attribution of dispositional states by 12-month-olds. *Psychological Science, 14,* 402–408.

Liszkowski, U., Carpenter, M., & Tomasello, M. (2008). Twelve-month-olds communicate helpfully and appropriately for knowledgeable and ignorant partners. *Cognition, 108,* 732-739.

Liszkowski, U., Carpenter, M., Striano, T., & Tomasello, M. (2006). Twelve and 18-month-olds point to provide information. *Journal of Cognition and Development, 7,* 173–187.

Luo, Y., & Baillargion, R. (2005). Can a self-propelled box have a goal? Psychological reasoning in 5-month-old Infants. *Psychological Science, 16,* 601-608.

Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science, 308,* 255–258.

Song, H., Baillargeon, R., & Fisher, C. (2005). Can infants attribute to an agent a disposition to perform a particular action? *Cognition, 98,* B45–B55.

Spelke, E. (1994). Initial knowledge: six suggestions. *Cognition, 40,* 431-445.

Warneken, F., & Tomasello, M. (2006) Altruistic helping in human infants and young chimpanzees. *Science, 311,* 1301–1303.

Warneken, F., & Tomasello, M. (2008) Extrinsic rewards undermine altruistic tendencies in 20-month-olds. *Developmental Psychology, 44,* 1785–1788.

Woodward, A. L. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition, 69,* 1–34.

Priming Effects on Event Types Classification: Effects of Word and Picture Stimuli

Alessandra Zarcone (zarconaa@ims.uni-stuttgart.de)
Institut für Maschinelle Sprachverarbeitung, Azenbergstr. 12
70174 Stuttgart, Germany

Alessandro Lenci (alessandro.lenci@ling.unipi.it)
Dipartimento di Linguistica "T. Bolelli", Via Santa Maria 36
56126 Pisa, Italy

Abstract

Event types (ET) have been widely addressed in linguistic literature, but few studies have dealt with the questions of how they are represented, retrieved and processed in the mental lexicon. We report two experiments in which ET categories were found to give rise to semantic priming effects, both with word and picture stimuli. These effects are argued to provide empirical correlates for ET categories in the mental lexicon not only at the lexical level but also at a deeper conceptual level.

Keywords: Semantic priming; event types; verb processing; verb semantics; psycholinguistics.

Introduction

Event Types

Event types (ET) are an important component of the "event structure template" (Kemmerer & Gonzales-Castillo, 2010) of the verb, and play crucial role in the temporal constitution of the sentence. We refer here to Vendler's (1967) standard classification of predicates into *states* (STA), *activities* (ACT), *accomplishments* (ACC) and *achievements* (ACH)[1]. These categories can be further cross-classified with respect to the features of dynamicity (DYN), durativity (DUR) and resultativity (RES)[2] (see Table 1).

Table 1: Features of Vendler's event types

ET	[dyn]	[dur]	[res]
STA	−	+	−
ACT	+	+	−
ACC	+	+	+
ACH	+	−	+

In particular, we focused on ACHs and ACTs, because they contrast with respect to DUR and RES: ACH $[-dur, +res]$ (e.g., *land*, *die*); ACT $[+dur, -res]$ (e.g. *sing*, *walk*).

[1] **STA** denote properties and situations experienced by the subject as being static (e.g. *to know*, *to be tall*); **ACT** denote non-resultative activities (e.g. *to sing*, *to walk*); **ACC** denote activities with a clear goal or outcome (e.g. *to write a book*, *to walk to the fence*); **ACH** denote a change of state (e.g. *to stumble*, *to die*).

[2] *DYN* distinguishes among stative events and dynamic events (e.g. *to live*, *to know*, vs. *to run*, *to stumble*). *DUR* events are events perceived as lasting over time (e.g. *to knit*, *to stir*), *non-DUR* events are perceived as punctual (e.g. *to fall*, *to die*). *RES* events entail the existence of a clear outcome or resulting state that has to be reached for the event to be considered completed (e.g. *to land*, *to write a book*, vs. *to fly*, *to talk*).

Empirical Correlates of Event Types

Event types (ET) have been widely addressed in linguistic literature, but few studies have dealt with the questions of how they are represented, retrieved and processed in the mental lexicon. Noteworhy exceptions are: Gennari and Poeppel (2002, 2003); Finocchiaro and Miceli (2002); Heyde-Zybatow (2004); Bott (2008); Bonnotte (2008).

In particular, Bonnotte (2008) shows semantic priming effects of ET for ACTs and ACHs in French, reporting differences between processing of durativity and resultativity: "facilitation was shown on the former with similar and opposite priming, whereas it was shown on the latter only with similar priming".

Goal of the work

The main goal of the work was the investigation of ETs in the mental lexicon, their representation and retrieval. We report two experiments based on the semantic priming paradigm (see McNamara, 2005, for a review), aimed at providing empirical correlates for ET categories.

Our starting point was the study in Bonnotte (2008), which we replicated for Italian with some crucial design innovations (Experiment 1). This experiment was conducted at a lexical level, using word stimuli. A second experiment (Experiment 2) introduced picture primes, in order to compare lexical semantic priming with non-linguistic priming, with the aim of delving into a deeper conceptual level than word stimuli.

ET categories were found to give rise to semantic priming effects, both with word and picture stimuli, but with a different pattern of results than in Bonnotte (2008): crucial differences were found at the ET level, and not between processing of durativity and resultativity. Priming effects were registered not only at the lexical level but also at a deeper conceptual level.

Experiment 1

Experiment 1 replicated the study conducted for French by Bonnotte (2008). As in Bonnotte (2008), Experiment 1 was designed to explore semantic priming effects of ET categories in Italian.

Nevertheless, two main differences were introduced. First of all, prime-target pairs and ACH-ACT sets were checked and tagged with respect to their semantic class, in order to rule out influences of the semantic class and to isolate effects of features pertaining to ETs, i.e. DUR and RES. Semantic

Table 2: Examples of prime-target pairs in Experiment 1

	target ACH	target ACT
neutral prime	XXX - sparare XXX - *to shoot*	XXX - dormire XXX - *to sleep*
opposite prime	ballare - sparare *to dance - to shoot*	entrare - dormire *to enter - to sleep*
similar prime	entrare - sparare *to enter - to shoot*	ballare - dormire *to dance - to sleep*

classes correspond to WordNet topnodes for verbs (Fellbaum, 1998). The prime and target of each test pair never belong to the same semantic class. Semantic classes were also used as a source of variance in the inferential statistic model. As a further difference with Bonnotte (2008) a slightly longer stimulus onset asyncrony (SOA) was used (300ms), in order to avoid spillover effects with longer stimuli.

Method

Participants 48 native Italian speakers from the University of Pisa and the Scuola Normale Superiore in Pisa volunteered to participate in the experiment and were paid for their participation. All had normal or corrected-to-normal vision.

Materials Two groups of 18 intransitive ACT Italian verbs and 18 intransitive ACH Italian verbs were pair-wise balanced for variables known to affect processing costs, such as length, frequency, syntactic frame frequency, ET polysemy; they were used as targets in the priming experiment.

The average length was 8 characters for ACTs (SD = 1.5) and 8 for ACHs (SD = 1.5) and did not differ significantly between the two groups (Kruskal-Wallis: df $= 1, \chi_2 = 0, p = 1$). Mean frequency (estimated from ColFis, Laudanna et al., 1995) was 129.5 occurrences per 3 million words for ACTs (SD = 165.5), and 88 for ACHs (SD = 173.5) and did not differ significantly between the two groups (Kruskal-Wallis: df $= 1, \chi_2 = 1.683, p = 0.2$). Syntactic frame frequencies were estimated from Repubblica corpus (Lenci et al., 2010): all verbs were intransitive and strongly monoargumental; ET polysemy was assessed with pre-test 1.

Each target appeared in one of three prime contexts: after a neutral prime (a string of Xs), after a similar prime (a verb of the same ET), after an opposite prime (a verb of opposite ET). As prime verbs we used different verbs than the target verbs. See examples in Table 2. Each prime-target pair was assigned to one of three lists so that an equal number of pairs per each condition appeared on each list, so that exactly one version of each target appeared on each list and so that each participant saw not more than one version of each target.

Pre-test 1 Italian lacks morphological clues for ET, and verbs tend to be ambiguous with respect to their ET category. Our experiments required non-ambiguous verbs, to be assessed with an inter-annotators pre-test inspired by the one in Bonnotte (2008).

Pre-test 1 was carried out to check our annotation of the verbs according to their ET. Materials for pre-test 1 were 136 predicates (114 transitive VPs - verb + object - and 22 intransitive verbs). Both transitive and intransitive verbs showing all four of Vendler's (1967) ET categories were used, both to have less constrained answers and to have a broader stimuli set for further experiments.

20 native Italian-speaking students performed the test in a web-based format. Per each event, subjects were asked to choose one of four pictures, one representative of each ET:

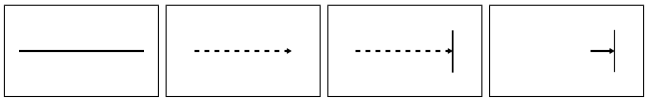


Figure 1: Pictures used in pre-test 1: the long continuous line depicts a state that lasts in time, the long dashed arrow depicts a process that develops over a certain period of time, the long dashed arrow ending with a vertical dash depicts a process that develops over a certain period of time and leads to a result, the short arrow ending with a vertical dash depicts an event that causes a change of state.

Results showed a mean accuracy of .61, inter-subject observed agreement of .5, inter-subject expected agreement of .25 and a kappa mean value of .33. Kappa was .46 on intransitive ACHs and .34 on intransitive ACTs. Agreement values were above chance and significantly good, since the subjects were naive to linguistics and ET classification. 3 ACHs and 3 ACTs showing low agreement (< 0.19) were ruled out for future experiments.

Procedure Participants were instructed to read the prime and the target and perform a semantic decision task. Half of the subjects were assigned a durativity decision task, the other half were assigned a resultativity decision task. Within the DUR task, subjects were asked:

Does the target denote a process
lasting over a period of time?

Within the RES task, subjects were asked:

Does the target denote an event with a clear outcome?

The semantic decision task directly references the manipulated variables, but nevertheless it was preferred over a more neutral lexical decision task for a better comparison with previous results and procedures in Bonnotte (2008). Task choice was later supported by the good accuracy results achieved.

Prime-target stimuli were presented on a screen in white upper-case letters on a black background with an SOA of 300ms. The target was deleted after the response. Participants answered by pressing one of two buttons on a button box, which recorded the decision latencies (DL) to one tenth of ms accuracy. DL were recorded as the time between the target onset and the response. Participants were given a detailed description of the experimental trials and were trained during a special simulation session (9 practice trials) before beginning the experiment.

Table 3: Experiment 1 - Mixed Effect Model: $log(dl) \sim prime + et + task + (1|subj) + (1|verb) + (1|sem_cl)$

	Estimate	MCMCmean	HPD95lower	HPD95upper	pMCMC	$Pr(>\|t\|)$	
(Intercept)	9.49	9.66	−12.78	30.79	0.16	0.00	
primeopp	−0.09	−0.09	−0.14	−0.04	0	0.00	***
primesim	−0.05	−0.05	−0.10	−0.01	0.02	0.02	*
etACT	−0.10	−0.11	−0.21	0.01	0.06	0.04	*
taskris	0.09	0.09	0.00	0.18	0.06	0.12	

Table 4: Experiment 1 - Separate analyses: $log(dl) \sim prime + (1|subj) + (1|verb) + (1|sem_cl)$

	Estimate	MCMCmean	HPD95lower	HPD95upper	pMCMC	$Pr(>\|t\|)$	
DUR, ACH targets							
(Intercept)	9.48	9.48	9.34	9.62	0	0	
opp	−0.1	−0.1	−0.18	−0.02	0.02	0.02	*
sim	−0.03	−0.03	−0.11	0.05	0.47	0.45	
DUR, ACT targets							
(Intercept)	9.4	9.4	9.23	9.56	0	0	
opp	−0.06	−0.06	−0.15	0.02	0.13	0.12	
sim	−0.11	−0.11	−0.20	−0.03	0.01	0.01	**
RES, ACH targets							
(Intercept)	9.61	9.6	9.45	9.77	0	0	
opp	−0.15	−0.15	−0.26	−0.04	0.01	0.01	**
sim	−0.06	−0.06	−0.16	0.06	0.32	0.29	
RES, ACT targets							
(Intercept)	9.45	9.45	9.32	9.58	0	0	
opp	−0.07	−0.07	−0.17	0.03	0.16	0.14	
sim	−0.02	−0.02	−0.12	0.08	0.71	0.66	

Design Experiment 1 had a 2x3 within-subjects design (2-levels factor being the ET of the target, and 3-levels factor being the type of prime context) with one between-subjects factor (DUR task, RES task).

Results

The neutral prime level was used as a baseline to evaluate the effect of opposite and similar prime on decision latencies (DL): both primes show smaller mean DL, suggesting a general facilitation effect (see Figure 2, more detailed information on Table 5). A mixed effect model (see Table 3) of DLs[3] showed that the difference between the neutral prime and the opposite prime was highly significant, and the difference between the neutral prime and the similar prime was significant; furthermore, it yielded a significant effect of the target's ET.

General accuracy was .86 (.89 for DUR, .82 for RES). A logistic regression analysis performed on errors did not yield any effect of the priming context or of any other factor.

Separate analyses Four separate analyses were conducted, one for each ET (ACH and ACT) within each task (DUR and RES), using four smaller-scale mixed effect models (see Table 4). A significant difference between the neutral prime level and the opposite prime level was found on ACH targets for both DUR and RES tasks. A significant difference between the neutral prime level and the similar prime level was found on ACT targets in the DUR task.

[3]Fixed effects were prime, task, ET of the target; random effects were subject, target verb, semantic class of the target.

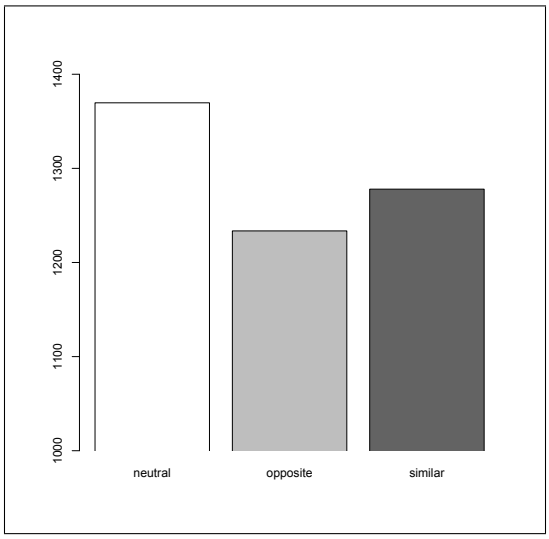


Figure 2: RT means for the different priming contexts in Experiment 1.

In the DUR task, ACTs are able to prime both ACT and ACH targets; on the other hand, in the same task ACHs never prime either ET. This might depend on the fact that ACT are positively marked with the feature of DUR, which is relevant for this task. However, in the RES task, only ACTs have a significant priming effect on ACH targets, suggesting that priming occurs in this case only when the target is positively marked with the feature of RES, activated in this task. Here

Table 5: RT means (in ms) and standard deviations in Experiment 1 and Experiment 2

	Experiment 1								Experiment 2							
	DUR				RES				DUR				RES			
	ACH		ACC		ACH		ACC		ACH		ACC		ACH		ACC	
	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd
neu	1370	403	1291	501	1558	539	1309	484	1226	409	1076	332	1305	426	1316	360
opp	1242	382	1186	456	1335	428	1227	465	1238	421	1128	372	1289	380	1297	327
sim	1312	370	1100	305	1472	530	1296	434	1302	472	1102	325	1422	474	1358	380

it is more the contrast between the [−res] of ACTs and the [+res] of ACHs to produce a priming effect.

Experiment 2

Experiment 1 showed significant priming effects of ET at the lexical level. The aim of Experiment 2 was to delve to a deeper conceptual level, in order to assess if ETs are "pure linguistic" categorizations or if they rather apply also to non linguistic input. With this purpose, a key modification was applied to Experiment 1: picture primes were used instead of word primes.

Method

Participants 42 native Italian speakers from the University of Pisa and the Scuola Normale Superiore in Pisa volunteered to participate in the experiment and were paid for their participation. All had normal or corrected-to-normal vision.

Materials Targets from Experiment 1 were also used as targets for Experiment 2. Each target appeared in one of three prime contexts: after a neutral prime (a pattern of Xs), after a similar prime (a picture depicting an event of the same ET), after an opposite prime (a picture depicting an event of opposite ET). Picture primes were selected from the IPNP database (Bates et al., 2000, see examples in Figure 3) and their association with ET categories was assessed through pre-test 2. Each prime-target pair was assigned to one of three lists so that an equal number of pairs per each condition appeared on each list, so that exactly one version of each target appeared on each list and so that each participant saw not more than one version of each target.

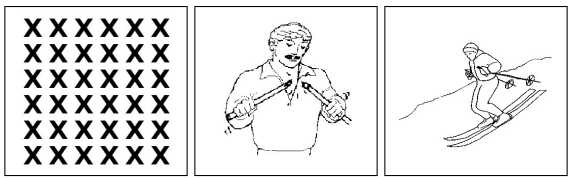


Figure 3: Picture primes: neutral prime, ACH picture (*to break*), ACT picture (*to ski*).

Pre-test 2 Non-ambiguous verb stimuli were selected through pre-test 1; a similar pre-test was conducted to select picture stimuli for Experiment 2. Materials for pre-test were 87 pictures from the IPNP database. Again, all four of Vendler's ET categories were used as possible answers. 20 native Italian-speaking students performed the test in a web-based format. Procedure was the same as in pre-test 1.

Results showed an inter-subject observed agreement of .42, inter-subject expected agreement of .26 and a kappa mean value of .21. 12 ACH pictures and 12 ACT pictures showing best agreement were chosen as primes for Experiment 2. Kappa mean value for chosen pictures was .42 (.41 for ACHs, .43 for ACTs).

Procedure Procedure was the same as in Experiment 1, with one difference: picture stimuli required longer times to be processed, and so SOA was set to a higher value (700 ms); SOA was assessed by asking 10 more participants from the same pool as in Experiment 1 and 2 to name the pictures. 700 ms was estimated as the shortest presentation time to allow the participants to identify the picture.

Design Design was the same as in Experiment 1.

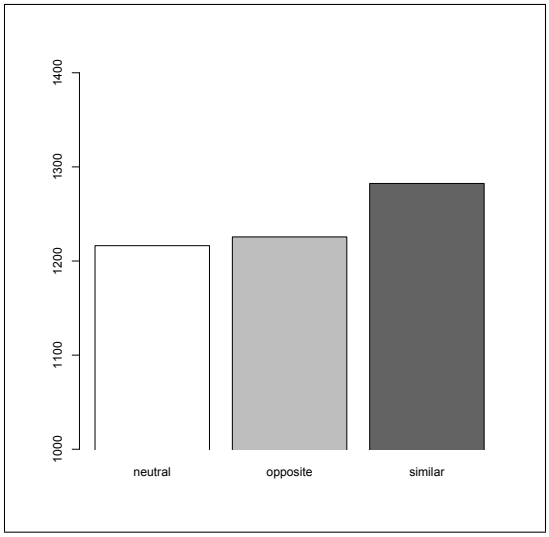


Figure 4: RT means for the different priming contexts in Experiment 2.

Results

In contrast with Experiment 1, in Experiment 2 picture primes showed longer mean DL than the neutral prime, suggesting a general inhibitory effect (see Figure 4, more detailed information on Table 5). A mixed effect model (see Table 6) of

Table 6: Experiment 2 - Mixed Effect Model: $log(dl) \sim prime + et + task + featval + (1|subj) + (1|verb) + (1|sem_cl)$

	Estimate	MCMCmean	HPD95lower	HPD95upper	pMCMC	$Pr(>\|t\|)$	
(Intercept)	9.4	9.4	9.31	9.49	0	0	
primeopp	0.01	0.01	−0.02	0.03	0.68	0.69	
primesim	0.05	0.05	0.03	0.08	0.00	0.00	***
etACT	−0.08	−0.08	−0.14	−0.02	0.01	0.01	**
taskris	0.14	0.14	0.05	0.22	0.00	0.02	*
featval+	−0.05	−0.05	−0.08	−0.03	0.00	0.00	***

Table 7: Experiment 2 - Separate analyses: $log(dl) \sim prime + (1|subj) + (1|verb) + (1|sem_cl)$

	Estimate	MCMCmean	HPD95lower	HPD95upper	pMCMC	$Pr(>\|t\|)$	
DUR, ACH targets							
(Intercept)	9.38	9.37	9.24	9.53	0	0	
opp	0.02	0.01	−0.04	0.07	0.59	0.55	
sim	0.08	0.08	0.02	0.13	0.00	0.00	**
DUR, ACT targets							
(Intercept)	9.26	9.26	9.18	9.33	0	0	
opp	0.03	0.03	−0.02	0.08	0.22	0.21	
sim	0.02	0.02	−0.02	0.07	0.36	0.34	
RES, ACH targets							
(Intercept)	9.52	9.51	9.35	9.68	0	0	
opp	−0.02	−0.02	−0.08	0.03	0.52	0.48	
sim	0.07	0.07	0.01	0.12	0.01	0.01	*
RES, ACT targets							
(Intercept)	9.46	9.46	9.39	9.53	0	0	
opp	−0.01	−0.01	−0.06	0.04	0.71	0.7	
sim	0.03	0.03	−0.02	0.09	0.21	0.19	

DLs[4] showed that the difference between the neutral prime and the similar prime was the only one to reach significancy; furthermore, it yielded a significant effect of the target's ET, of the task and of the featural value ([+/−dur], [+/−res]) of the target.

General accuracy was .92 (.94 for DUR, .90 for RES). A logistic regression analysis performed on errors did not yield any effect of the priming context or of any other factor.

Separate analyses Four separate analyses were conducted, one for each ET (ACH and ACT) within each task (DUR and RES), using four smaller-scale mixed effect models (see Table 7). A significant difference between the neutral prime level and the similar prime level was found on ACH targets for both DUR and RES tasks.

A striking difference with respect to Experiment 1 is the absence of priming effects with ACTs, both as target or prime. This fact might be due to the inherently "static" character of picture stimuli, which makes the [+dur] of ACTs less salient.

General Discussion and Conclusions

In line with both our expectations and the study in Bonnotte (2008), our experiments yielded significant priming effects of ET, thus providing evidence to the idea that ETs are indeed relevant for the mental lexicon. This conclusion is further supported by the crucial innovation we introduced in the experiments, i.e. controlling the semantic class of prime and target verbs. The priming effects can thus be related to the more abstract event structure shared by verbs that greatly differ for other dimensions of their meaning.

In addition to this, two different modalities were explored and contrasted: word primes and picture primes. Using picture stimuli is a first but significant attempt to place the study of ETs within a broader frame of study of event meaning in cognition. The Embodied Cognition Framework (Evans & Green, 2006; Haggard et al., 2007; Barsalou, 2008) suggests that semantic representations are not purely amodal, but rather grounded in our sensorimotor perception, and it has been suggested that processing a verb might involve "covertly recapitulating" the event it refers to (Kemmerer & Gonzales-Castillo, 2010).

The effect of facilitation given by the word primes is not surprising; the negative priming which we report for the picture primes is usually explained with a combination of inhibition (an effort of selective attention to avoid a previous stimulus) and memory retrieval (see Tipper, 2001, for a review). Picture primes seem to act at a deeper level than word primes, and it is crucial that the negative priming is found in the similar prime condition: similar primes seem to be more difficult for subjects to ignore.

Moreover, the pattern of results offered by this study suggests a different explanation of such priming effects than the one offered by Bonnotte (2008) (see Table 8). Bonnotte (2008) suggests a crucial difference between processing of

[4]Fixed effects were prime, task, ET of the target, featural value ([+/−dur], [+/−res]) of the target; random effects were subject, target verb, semantic class of the target.

Table 8: Comparison with results in Bonnotte (2008).

	DUR		RES	
	ACH	**ACT**	**ACH**	**ACT**
Bonnotte 2008	–	sim and opp	sim	sim
Exp 1	opp	sim	opp	–
Exp 2	sim	–	sim	–

durativity and resultativity: "facilitation was shown on the former with similar and opposite priming, whereas it was shown on the latter only with similar priming". Nevertheless, the pattern emerging from our experiments does not show great differences between processing of durativity and resultativity, but rather suggests a difference at the level of ET categories, which seem to differ with respect to their behavior in both tasks. Differences in priming effects across ET categories can be ascribed to different lexical encodings of their ET features: the [+dur] and [−res] of ACTs is more ductile and subject to contextual adaption, whereas ACHs are more "inherently" [−dur] [+res]. Moreover, ACTs do not seem to be affected by priming with picture stimuli, which might be problematic in conveying the [+dur] [−res] nature of ACTs. In the near future a comparison will be carried out with a similar study of Russian (Batiukova et al., 2010).

The use of videos was also contemplated for this study, but pictures were preferred for a first exploration of the visual modality because the IPNP database provided a convenient standard of stimuli and because picture primes allowed for shorter SOAs. This would not have been the case for video stimuli. Nevertheless, videos have been used in the investigation of event representations (e.g. Gennari et al., 2002) and, since they could provide a better depiction of both DUR and RES, this modality would definitely be of some interest for further work, in order to more thoroughly investigate ET representations in the mental lexicon.

ET categories were found to give rise to semantic priming effects, at both word and picture levels, which, albeit with crucial cross-modal differences, provide empirical correlates for ET categories in the mental lexicon and suggest that ETs are not only a linguistic phenomenon, but relate with our way of conceptualizing events in the world.

Acknowledgments The experiments reported were funded by the Laboratorio di Linguistica of the Scuola Normale Superiore in Pisa. The authors thank Pier Marco Bertinetto, Valentina Bambini, Berry Claus and Pirita Pyykkönen for helpful discussions.

References

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology*, *59*, 617–645.

Bates, E., Federmeier, K., Herron, D., Iyer, G., Jacobsen, T., Pechmann, T., et al. (2000). Introducing the CRL International Picture-Naming Project (CRL-IPNP). *Center for Research in Language Newsletter*, *12*.

Batiukova, O., Bertinetto, P. M., Lenci, A., & Zarcone, A. (2010). Semantic priming study of Russian aspect and resultativity. In *Proceedings of The Russian Verb, formal and contrastive approaches to aspect, tense and mood in Russian.* St. Petersburg.

Bonnotte, I. (2008). The role of semantic features in verb processing. *Journal of Psycholinguistic Research*, *37*, 199–217.

Bott, O. (2008). Doing it again and again may be difficult but it depends on what you are doing. In *Proceedings of the 27th West Coast Conference on Formal Linguistics.* Somerville, MA: Cascadilla Proceedings Project.

Evans, V., & Green, M. (2006). *Cognitive linguistics*. Mahwah: Lawrence Erlbaum.

Fellbaum, C. (Ed.). (1998). *Wordnet. language, speech and communication*. Cambridge, MA: The MIT Press.

Finocchiaro, C., & Miceli, G. (2002). Verb actionality in aphasia: data from two aphasic subjects. *Folia Linguistica*, *36*, 335–357.

Gennari, S., & Poeppel, D. (2002). Events versus states: Empirical correlates of lexical classes. In *Proceedings of the 24th Annual Conference of the Cognitive Science Society* (pp. 351–356). Mahwah, NJ: Lawrence Erlbaum Associates.

Gennari, S., & Poeppel, D. (2003). Processing correlates of lexical semantic complexity. *Cognition*, *89*(1), B27–41.

Gennari, S., Sloman, S. A., Malt, B. C., & Fitch, W. T. (2002). Motion events in language and cogniton. *Cognition*, *83*, 49–79.

Haggard, P., Rossetti, Y., & Kawato, M. (Eds.). (2007). *Sensorimotor foundations of higher cognition*. Oxford, UK: Oxford University Press.

Heyde-Zybatow, T. (2004). *Achievements: two experimental studies and one semantic analysis.* (Talk delivered to Sinn und Bedeutung 9, Nijmegen, 1-3 November 2004.)

Kemmerer, D., & Gonzales-Castillo, J. (2010). The two-level theory of verb meaning: an approach to integrating the semantics of action with the mirror neuron system. *Brain and Language*, *112*, 54–76.

Laudanna, A., Thornton, A. M., Brown, G., Burani, C., & Marconi, L. (1995). Un corpus dell'Italiano scritto contemporaneo dalla parte del ricevente. In S. Bolasco, L. Lebart, & A. Salem (Eds.), *III giornate internazionali di analisi statistica dei dati testuali* (Vol. 1).

Lenci, A., Johnson, M., & Lapesa, G. (2010). Building an italian framenet through semi-automatic corpus analysis. In *Proceedings of LREC 2010.* La Valletta, Malta: ELDA.

McNamara, T. (2005). *Semantic priming - perspectives from memory and word recognition*. New York: Psychology Press.

Tipper, S. P. (2001). Does negative priming reflect inhibitory mechanisms? a review and integration of conflicting views. *The Quarterly Journal of Experimental Psychology*, *54A*(2), 321–343.

Vendler, Z. (1967). *Linguistics in philosophy*. Ithaca, NY: Cornell University Press.

Motor Effects in Rating Lines' Length Using a Dichotomous Scale

Lyuben D. Laskin (laskin@mail.bg)
Department of Cognitive Science and Psychology,
New Bulgarian University, 21 Montevideo Street
Sofia 1618, Bulgaria

Abstract

The aim of this study is to demonstrate how the execution of particular task-specific motor movements can influence subjects' ratings of simple stimuli. Sixty-four participants in one control and two experimental groups rated lines of 36 different lengths. Lines appeared on a computer screen and subjects gave their ratings using a standard keyboard. In the experimental groups trials did not change automatically, but subjects had to press a specific button (called the "trial change button"), which was next to one of the response buttons. It was hypothesized that this manipulation would lead to assimilation of the ratings toward the category whose button was next to the trial change button. The results confirmed this hypothesis. Possible explanations of the results are discussed.

Keywords: context effects; scale ratings; grounded cognition; motor actions.

Introduction

According to traditional views in psychology and cognitive science, the role of sensory and motor processes in cognition is only peripheral. Our sensory organs receive information from the outside world and that information is transduced into amodal symbols which represent knowledge. High-level cognition (language, memory, decision making, problem solving, etc.) consists of the interaction of these symbols with each other, the product of which is either the activation of other amodal symbols, or their transduction into motor commands.

Researchers from the field of grounded cognition (Barsalou, 2008) assign a very different role to our sensory and motor systems. According to that view, the brain does not explicitly represent amodal symbols, but rather high-level cognition emerges from the interaction between the brain, the body, and the environment. This can also be stated by saying that high-level cognition is grounded in sensory and motor representations, not amodal, abstract symbols.

An ample amount of empirical results supports the views of grounded cognition. Evidence shows that haptic, visual, auditory sensations, proprioception, as well as execution of motor actions, all influence higher-level cognitive processes, like memory, language processing, visual and motor imagery, and so on (for a review, see Barsalou, 2008).

The aim of this study is to demonstrate how motor actions required for the execution of a particular cognitive task can affect high-level cognitive processes. More specifically, we are going to try to show this by demonstrating how motor actions necessary to perform a scale rating task can influence the ratings.

There already exists a field in psychology which deals with the so-called context effects in scale ratings. There is bountiful experimental literature demonstrating changes in subjects' ratings, influenced by factors like the range of the stimuli, their distribution, the sequence of their presentation, and so on.

Some of the studies demonstrate how context can systematically[1] change the ratings of stimuli evaluated only by one dimension. Examples include judgments of square sizes (Parducci & Perrett, 1971; Sarris & Parducci, 1978), weights (Parducci & Marshall, 1962; Sherif, Taub, & Hovland, 1958), and the length of lines (Kokinov, Hristova, & Petkov, 2004; Petrov & Anderson, 2005).

Other studies demonstrate contextual effects in the ratings of more complex stimuli (stimuli that must be evaluated based on more than one dimension). For example, Cooke & Mellers (1998) asked participants to rate flats' attractiveness based on their rent, number of rooms, and distance from campus. Mellers (1982) demonstrates such effects in equity judgments, and Wedell, Parducci, & Geiselman (1987) show contextual effects in ratings of the attractiveness of female faces.

There are a number of influential theories which try to explain such experimental results. One of the first theories in the field is the **adaptation-level theory** (Helson, 1964). According to that theory, the stimuli a person has rated leave a general impression with which all other stimuli are compared while being assessed. Another powerful theory in this field is the **range-frequency theory (**Parducci, 1965, 1968, 1974). It claims that a stimulus' rating is a compromise between the range and frequency principles. The former refers simply to the lower and higher end of the stimulus material (e.g., the smallest and the biggest square, if the task is to judge the size of different squares). The latter principle is concerned with the distribution of the stimuli (e.g., uniform, positively or negatively skewed, etc). Discussing in detail these and other theories in the field of contextual effects in scale ratings is beyond the scope of this paper.

To our knowledge, there are no studies showing changes in people's ratings of stimuli caused by "peripheral" factors like the specific motor actions executed during the process of rating itself. Furthermore, none of the theories presented

[1] For a change to be considered systematic, it has to be in one particular direction. When the change is in the direction of the context (e.g., when there are more big squares than small squares in the stimulus material and an average square receives a higher rating than normal), the effect is called *assimilation*, whereas when the change is in the opposite direction of the context, the effect is called *contrast* or *compensation*.

above predict any such effects. As was mentioned earlier, the aim of the current study is to demonstrate how task-specific motor actions can influence subjects' performance in the task. Next, we will review some of the literature concerned with how executing particular motor actions affects some cognitive processes which don't seem to be directly related to the motor actions.

Motor Effects in High-Level Cognitive Processes

In this section, we will briefly present empirical results showing how different motor actions can influence high-level cognitive processes. The most commonly used experimental paradigms are described below.

Cacioppo, Priester, & Berntson (1993) use the isometric arm flexion and extension paradigm to trigger the so-called approach and avoidance systems. The authors argue that when these systems are activated, stimuli that people interact with are perceived as more positive or more negative, respectively. In a series of experiments they ask subjects to either place their palms at the bottom of a table and to lift slightly (flexion condition) or to place their palms at the top of the table and to push slightly (extension condition), while at the same time observing Chinese ideographs which they later rate as pleasant or unpleasant. The results show that ideographs observed during arm flexion are later rated as more pleasant, whereas those observed during extension are rated as less pleasant.

Another commonly used method is inducing a smile or a frown by asking subjects to hold a pen or a pencil with their teeth or with their lips, respectively. The muscles activated during these actions are also active during one of the above facial expressions. Using this manipulation, Strack, Martin, & Stepper (1988) showed that holding a pen between the teeth or between the lips leads to evaluating cartoons as funnier or less funny, respectively.

Head nodding or shaking are meaningful gestures in most cultures. They convey agreement/disagreement with or approval/disapproval of someone else's behavior, a witnessed event, etc. Wells & Petty (1980) used the association between the type of head movement and the created mental set toward the currently active concepts to show that making vertical head movements while listening to a message leads to higher agreement with the message, whereas making horizontal movements leads to lower agreement.

For a review of other experiments showing motor influence on high-level cognitive processes, see Briñol & Petty (2008).

Possible Explanations of These Findings

The explanation that most papers provide for the obtained results is related to the existing associations between a motor action and a cognitive response (e.g., nodding associated with agreement). These associations are created during a person's life and influenced by their culture. But how are they created?

Zwaan & Madden (2005) provide one possible mechanism by which such associations can be established. According to their **interconnected experiential traces** theory, all mental representations are experiential, that is, created during some form of interaction with the outside world. They define two types of representations: **referent** and **linguistic**. The former are multimodal memory traces laid down during interaction with the environment. The latter representations are laid down during receiving or producing linguistic information (e.g., talking, listening, writing, etc.). A very important feature of these representations is that they can be interconnected (associated).

The authors propose **co-occurrence** as a possible mechanism for establishing these associations. When two events occur simultaneously or in succession, the neural assemblies which represent those events establish stronger connections with each other (Hebb, 1949). For example, the visual image of a falling glass of water is likely to be associated with the sound of breaking glass. This happens because in a person's lifetime, the experience of a glass falling on the ground from a certain height has almost always been followed by a specific sound (that of breaking glass). Thus, that person develops anticipation for that sound after seeing a falling glass.

Experiment

Likert scales are often used in pilot studies or even as dependent measures in experiments. Researchers exploring contextual effects in ratings have showed that these measures can sometimes be affected by factors other than those being investigated by the particular study (see studies reported in the introduction). However, they have emphasized on "cognitive" factors and have not studied any possible influence of sensory or motor processes on subjects' ratings. The current experiment's goal is to make the first step in filling this gap by demonstrating changes in subjects' ratings influenced by the specific hand movements they make while rating lines of different lengths.

One common feature of the experiments demonstrating motor effects in high-level cognitive processes reviewed in the previous section is that they all exploit associations between different types of representations that have already been formed throughout participants' lives. The current experiment will try a different approach by attempting to *create* new short-term associations between particular motor movements and conceptual categories.

Having in mind the interconnected experiential traces theory of Zwaan & Madden (2005), it can be hypothesized that if the activation of a particular category is repeatedly coupled with the execution of a motor action, a temporary association between the respective category and motor action might be created. After that, the execution of the motor action alone may be sufficient to activate the category with which it was associated.

Hypotheses

In the current experiment, participants' task was to rate lines' lengths using a dichotomous scale (a line could be rated as "short" or "long"). If the motor actions (hand movements) required for giving one of the two responses are different in nature this can lead to the creation of a new associative connection between them (i.e., between one of the two categories and the respective hand movement). If, then, one of the motor actions is activated, it should also activate the associated category.

When one of the two categories is more active than the other, this can *increase* the probability of the line being currently rated to receive that particular rating (e.g., a middle-sized line may be rated as "long" if that category's base-level activation is higher that usual). This hypothesis was tested using the procedure described below.

Method

Participants 64 New Bulgarian University (26 males, 38 females) undergraduate students volunteered for this study.

Stimuli The stimulus material consisted of 36 lines of different lengths appearing in the middle of a computer screen. The shortest line was 38 pixels and the longest line was 668 pixels, with an increment of 18 pixels. Lines were 2 pixels thick.

Apparatus Lines were presented on a 17'' TFT monitor with a resolution of 1024 x 768 pixels. The responses were obtained using a standard computer keyboard. The experimental script was written with the E-Prime 1.1 software.

Design and procedure The experiment was conducted in small rooms with each participant being tested individually. Subjects sat in front of a computer and the instructions were presented to them in written form across the screen, as well as explained to them by the experimenter. In short, the instructions said that subjects would take part in a study concerned with people's judgment of length and that their task would be to rate different lines presented on the screen as "short" or "long" using the two specified buttons on the keyboard.

The experiment employed a between-subject design with one control group and two experimental groups (see Table 1). In the control group the procedure was the following: after they heard the instruction subjects went through a training session to be familiarized with the experiment. The training session was the same as the experimental session, but only 10 (out of 36) lines were presented in random order. In the experimental session all 36 lines were presented 2 times each, resulting in a total of 72 trials. Figure 1 shows what a single trial looked like.

The procedure in the experimental groups was identical to that of the control group, except for the transition between trials. In the experimental groups, subjects had to press an additional button at the end of each trial (called the "trial change button") in order to see the next trial (Figure 2). The trial change button was positioned either next to the "short" button or next to the "long" button (see Table 1).

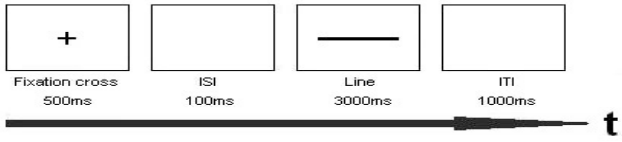


Figure 1: Every trial in the control group started with a fixation cross for 500 ms, followed by a 100 ms inter-stimulus interval, followed by a 3000 ms exposure of the line to be rated, and a 1000 ms inter-trial interval.

Subjects were asked to use only their right index finger to give their responses. Between every two trials they had to put their finger in a neutral position between the response buttons (the black rectangle in Figure 2)[2]. The sequence of actions in every trial (after the line's appearance on the screen) was: press one of the response buttons – press the trial change button – return to neutral position.

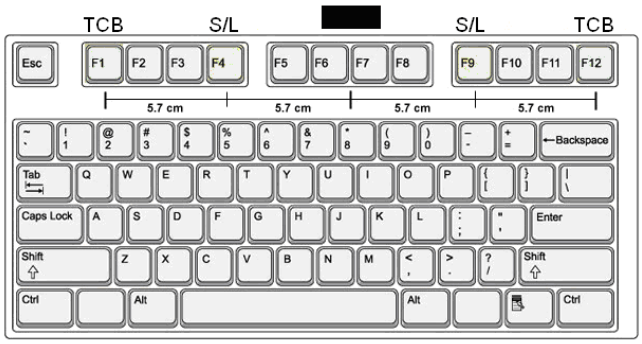


Figure 2: The F4 and F9 buttons were used as response buttons (for responding "short" or "long"), and the F1 and F12 buttons were used as trial change buttons. Both response buttons and trial change buttons were counter-balanced across conditions.

The dependent measure in this study was the response to each line ("long" vs. "short").

Table 1: The position of the trial change button with respect to the response buttons in the three groups.

	Ex. Group 1	**Ex. Group 2**	**Control Group**
TCB next to	"Long" button	"Short" button	No TCB

After the end of the experiment, subjects were debriefed, thanked, and dismissed.

[2] All other keyboard button functions were disabled, so pressing other buttons accidentally did not affect the experimental procedure. Thus, subjects were instructed to rest their wrists on the keyboard without worrying about accidentally pressing buttons other than those which were part of the procedure.

According to the main assumption in this study, the different types of movements should be associated with one of the two categories. That is, the categories "long" and "short" should be associated with a hand movement to the left or a hand movement to the right (depending on the experimental condition). Since in the two experimental groups the position of the trial change button also requires a hand movement either to the left or to the right immediately before the presentation of the next line to be rated), that movement should activate the respective category more than its rival category and the probability that each line is rated with that category should increase.

Results and Discussion

The expected results following this manipulation were that there is going to be an assimilation of the responses toward the position of the trial change button. That is, if the trial change button is next to the "long" response button, the probability that an arbitrary line is rated as "long" should be higher than in the control group, and if the trial change button is next to the "short" response button, the probability should be lower. The actual results confirmed these expectations (Figure 3).

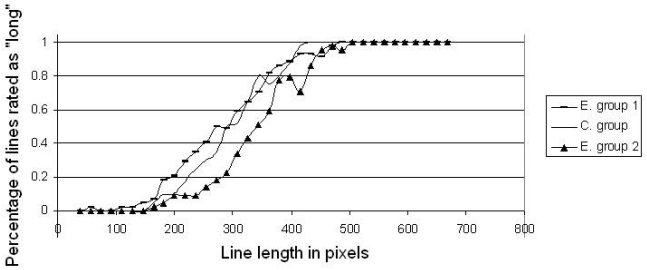


Figure 3: The probability for every line to be rated as "long" in the three conditions. As was expected, in experimental group 1 more lines were rated as "long" and in experimental group 2 more lines were rated as "short", in comparison with the control group.

All individual responses (the number of individual responses was *number of subjects * number of trials per subject*) were divided in three groups (the two experimental groups and the control group). "Long" responses were coded as "1", and "short" responses were coded as "0". A chi-square analysis was performed in order to test if the results in the three groups differed significantly, χ^2 (2) = 26.54, $p < 0.01$. The standard residuals[3] are given in Table 2. As can be seen, the two experimental groups were the main contributors for the significance of the results.

Since this analysis had too many individual measures, for a higher certainty in the significance of the results a repeated-measures ANOVA was also performed after aggregating the data for individual lines (that is, one individual measure stood for the percentage of a particular line rated as "long" in one of the three conditions), F (1, 35) = 68.11, $p < 0.001$. Three individual t-tests were performed to compare the three groups. The analyses revealed significant results between experimental group 2 and the control group, t (35) = 4.38, $p < 0.001$, ES = 0.73, also between experimental groups 1 and 2, t (35) = 4.98, $p < 0.001$, ES = 0.83, and marginally[4] significant results between experimental group 1 and the control group, t (35) = -2.3, p = 0.027, ES = 0.4.

Table 2: The standard residuals for the chi-square analysis.

	"Short"	"Long"
Control Group	-,9	,8
Ex. Group 1	-2,2	1,9
Ex. Group 2	3,1	-2,7

These results show that the presence of a trial change button always affects subjects' ratings. However, it is also evident that there is an asymmetry in the difference between the two experimental groups and the control group. That is, when the trial change button is next to the "short" response button the effect is stronger than when it is next to the "long" response button. This and other questions are discussed in the next section.

General Discussion

The results of this study showed that "non-cognitive" factors can also affect "cognitive" processes like judgment and categorization under certain circumstances. We think these results contribute to both the field of embodiment and grounded cognition, as well as to the field of context effects in scale ratings. The two main findings are: (1) task-specific motor actions can potentially affect subjects' ratings in a scale rating task, and (2) temporary associations between referent and/or linguistic representations can be established even for a short period of time.

Of course, there still remain a lot of open questions. In relation to the first finding, one thing that needs to be explored empirically is whether the same results can be obtained with a larger scale (e.g., a 7-point Likert scale). One might argue that the task in the current study was not scaling at all, but rather simple categorization.

[3] Standard residuals are used to determine which cells contribute most for the rejection of the null hypothesis in a chi-square analysis. Absolute values equal to or greater than 2 are considered statistically significant.

[4] Due to the increasing probability of making a type I error when performing more than 1 t-test on overlapping statistical data, the acceptable level of significance was not 0.05, but was set to $\alpha = 1 - \sqrt[3]{(1 - 0.05)} = 0.017$. For that reason p = 0.027 is considered a marginally significant result.

Another open question regarding the first finding is concerned with the observed asymmetry between the two experimental groups. All performed statistical analyses showed that the assimilation is stronger when the trial change button was next to the "short" response button than when it was next to the "long" response button. A possible explanation for this result can be found in the linguistic notion of **markedness** (Andrews, 1990). This term was coined by the Russian linguist Nikolai Trubetzkoy. Even though he used it to explain some phonological phenomena, other authors later extended the notion to other linguistic fields, including semantics. An unmarked form of a concept is a basic and natural form, whereas a marked form is one that is derived from the unmarked form. For example, lioness is the marked form of lion, since lion can refer to both male and female lions, whereas lioness only refers to female lions. Since, as was mentioned earlier, the task that subjects received in this experiment can be considered categorization, some markedness effects can also be observed. When talking about the size of a line, it is more natural to think about its "length", rather than its "shortness". This suggests that "long" is the unmarked category, and "short" is the marked category[5]. That might be the reason why the experimental manipulation was weaker for the experimental group in which the trial change button was next to the "long" response button. Subjects are simply more confident in responding "long" than in responding "short". However, this clearly is a *post-hoc* explanation and needs further confirmation.

It has also been brought to our attention that the results could be explained by assuming that subjects press the button closer to the TCB in order to save time and effort and not because of the activated referent or linguistic representations. This is a valid point and needs to be addressed in future studies.

Regarding the second finding in this study, the open questions are concerned with the exact mechanisms underlying these associations. Zwaan & Madden (2005) propose a sound theory, but it is not specified in enough detail.

Returning to the current study, one interesting question is related to the exact representations that are associated. Throughout this paper, it was assumed that the semantic category ("long" or "short") is associated with the particular type of movement (hand movement to the left or to the right). A second possibility is that it is the visual image of a button that is associated with the respective category. In that case, every time subjects have to press the trial change button, their attention is directed toward the respective response button too, and that activates its category. The results of the current experiment are unable to disambiguate between these possibilities.

Future Studies

It is clear that there are a lot of open questions that need to be investigated empirically. In this section, we will propose two experiments that might clarify some of them.

The first one is a natural extension to the current study. Namely, can the same results be obtained if there are more than two responses, that is, if a larger scale is used? The procedure in that study is going to be the same, but there are going to be more than two response buttons (e.g., 7 buttons for a 7-point scale) and again the trial change button is going to be placed at one side of the scale. If our hypothesis is correct, the same assimilation effect should be observed.

The second proposed experiment is aimed at answering the question of whether or not the obtained results are simply due to the fact that subjects' attention is being directed towards a particular response button every time they press the trial change button (see the discussion in the previous section) or if the results are due to a time/effort saving incentive. The proposed procedure is the following.

There are going to be three experimental sessions. In the first session, subjects will have the same rating task as in the current study (i.e., rating lines' lengths). However, instead of "long" and "short", the available response categories are going to be "big" and "small". If the assumptions made in this paper are correct, during this session these categories should be temporarily associated with the hand movements required for giving these responses.

In the second session, subjects will have a task whose goal will be to make them press one of the two response buttons more frequently than the other (e.g., a circle will appear on the left or on the right side of the screen and the subjects' task will be to press the respective button on the keyboard; the circle will appear more frequently in the right or in the left, depending on subjects' experimental condition). Again, if the assumptions made in this paper are correct, this should make the movement that has been executed more frequently more active than the other movement, and that would make the associated category more active as well.

In the third experimental sessions, subjects will have a task identical to that of the first session, but the stimuli will be different (e.g., rating squares, instead of lines), but again using the same categories for responses ("big" and "small"). If one of the two categories is more active than the other (because of the manipulation in the second experimental session) this should lead to a higher probability of responding with that category. Since in no part of this procedure is there any trial change button, the "attention" and time/effort saving explanations of the results can safely be ruled out.

Both proposed experiments are going to be performed in the near future.

[5] Results from the control group support this hypothesis. About 60% of the lines were rated as long, and only 40% as short, $\chi^2(1) = 45.93$, $p < 0.001$ (this difference would not be expected if subjects have no bias toward either category). It seems that subjects find it more natural to call a middle length line "long", rather than "short".

Conclusions

This study showed effects of motor actions on the process of rating lines as "long" or "short". The results are considered to contribute to both the field of grounded cognition and the field of context effects in scale ratings (or even to fields like psychophysics, if the effects of the methodology of measuring subjects' perceptions for different stimuli are to be taken seriously).

There are still many open questions which must be further explored. Despite all uncertainties however, these results show that it is quite likely that sensory and motor processes can be significant factors in the process of scaling (a finding that is not predicted by the main theories explaining contextual effects in scaling).

Acknowledgments

I would like to thank Armina Janyan, Boicho Kokinov, Encho Gerganov, Georgi Petkov, Ivan Vankov, and Stefan Mateev for the fruitful discussions. I would also like to thank my thesis advisor Penka Hristova for her advice, support, and encouragement without which my first publication would not have been the same. And last but not least, I want to thank Meryl Varadinov for helping me in fixing many small but significant details in this paper.

References

Andrews, E., (1990). Markedness theory: the union of asymmetry and semiosis in language. *Duke University Press.*

Barsalou, L.W. (2008). Grounded cognition. *Annual Review of Psychology*, 59, 617-645.

Briñol, P., & Petty, R. E. (2008). Embodied persuasion. In G. R. Semin & E. R. Smith (Eds.), *Embodied grounding: Social, cognitive, affective, and neuroscientific approaches.* Cambridge University Press.

Cacioppo, J. T., Priester, J. R., & Berntson, G. G. (1993). Rudimentary determinants of attitudes: II. Arm flexion and extension have differential effects on attitudes. *Journal of Personality and Social Psychology*, 65, 5-17.

Cooke, A. D. J., Mellers, B. A. (1998). Multiattribute judgment: attribute spacing influences single attributes. *Journal of Experimental Psychology: Human Perception and Performance*, 24 (April), 496-504.

Hebb, D. O. (1949). *The* Organization of Behavior: A Neuropsychological Theory. *New York:Wiley.*

Helson, H. (1964). *Adaptation level theory: an experimental and systematic approach to behavior*. New York: Harper & Row.

Kokinov, B., Hristova, P., Petkov, G. (2004). Does Irrelevant Information Play a Role in Judgment? *In Proceedings of the 26th Annual Conference of the Cognitive Science Society,* Erlbaum, Hillsdale, NJ., pp. 720-725.

Mellers, B. A. (1982). Equity judgment: a revision of Aristotelian views. *Journal of Experimental Psychology: General*, Vol. 111, No. 2, 242-270.

Parducci, A. (1965). Category judgment: a range-frequency model. *Psychological Review*, Vol. 72, No. 6, 407-418.

Parducci, A. (1968). The relativism of absolute judgments. *Scientific American*, Vol. 219 (6), 84-90.

Parducci, A. (1974). Contextual effects: a range-frequency analysis. *Handbook of Perception*, Vol. 2, NY: Academic Press, 127-141.

Parducci, A., Marshall, L. (1962). Assimilation vs. contrast in the anchoring of perceptual judgments of weight. *Journal of Experimental Psychology*, Vol. 63, No. 5, 426-437.

Parducci, A., Perrett, L. F. (1971). Category Rating Scales: Effects of relative spacing and frequency of stimulus values. *Journal of Experimental Psychology*, Vol. 89, 427-453.

Petrov, A., Anderson, J. (2005). The dynamics of scaling: a memory-based model of category rating and absolute identification. *Psychological Review*, 112 (2), 383-416.

Sarris, V., Parducci, A. (1978). Multiple Anchoring of Category Rating Scales. *Perception and Psychophysics*, Vol. 24 (1), 35-39.

Sherif, M., Taub, D., and Hovland, C. (1958). Assimilation and contrast effects of anchoring stimuli on judgments. *Journal of Experimental Psychology*, Vol. 55, No. 2, 150-155.

Strack, F., Martin, L., & Stepper, S. (1988). Inhibiting and facilitating conditions of the human smile: A nonobtrusive test of the facial feedback hypothesis. *Journal of Personality and Social Psychology*, 54, 768–777.

Wedell, D. H., Parducci, A., and Geiselman, R. E. (1987). A formal analysis of ratings of physical attractiveness: successive contrast and simultaneous assimilation. *Journal of Experimental Social Psychology*, 23 (May), 230-249.

Wells, G. L., & Petty, R. E. (1980). The effects of overt head movement son persuasion: Compatibility and incompatibility of responses. *Basic and Applied Social Psychology*, 1, 219–230.

Zwaan, R.A. & Madden, C.J. (2005). Embodied sentence comprehension. In Pecher, D. & Zwaan, R.A. (Eds.) Grounding cognition: *The role of perception and action in memory, language, and thinking*. Cambridge University Press, Cambridge, UK, pp 224-245.

The Role of Event Knowledge in Comprehending Synesthetic Metaphors

Tetsuaki Nakamura (tatnakac@edu.hc.uec.ac.jp)
Department of Human Communication, The University of Electro-Communications
1-5-1, Chofugaoka, Chofushi, Tokyo 182-8585, Japan

Maki Sakamoto (sakamoto@hc.uec.ac.jp)
Department of Informatics, The University of Electro-Communications
1-5-1, Chofugaoka, Chofushi, Tokyo 182-8585, Japan

Akira Utsumi (utsumi@se.uec.ac.jp)
Department of Informatics, The University of Electro-Communications
1-5-1, Chofugaoka, Chofushi, Tokyo 182-8585, Japan

Abstract

A synesthetic metaphor (e.g., "*sweet touch*") is a metaphor that results from a combination of a modifier and a head, where they express different perceptual qualities. Most of the existing studies examine how the acceptability of synesthetic metaphors can be explained by the pairing of adjective modifier's and head noun's modalities. However, little attention has been paid to how people comprehend synesthetic metaphors. This paper explores how people comprehend Japanese synesthetic metaphors. In our psychological experiment we collected 10388 words associated with 62 synesthetic metaphors and classified them into the following four kinds of features: common (features listed for the metaphor, the vehicle and the topic), vehicle-shared (features listed for both the metaphor and the vehicle, but not listed for the topic), topic-shared (features listed for both the metaphor and the topic, but not listed for the vehicle), and emergent (features listed for the metaphor, but not listed for either the vehicle or the topic). The result showed that there were significantly more emergent features than the other kinds of features in the comprehension of synesthetic metaphors. This result suggests that we do not so directly comprehend synesthetic metaphors based on salient features of the vehicle or the topic. In this paper we focus on event knowledge which is assumed to play a crucial role in comprehending synesthetic metaphors. We analyzed how many words associated with synesthetic metaphors could be classified into those based on event knowledge. The results showed that there were significantly more words based on event knowledge than those which could not be classified as words based on event knowledge. This result suggests that event knowledge play an important role in comprehending synesthetic metaphors.

Keywords: synesthetic metaphors; Japanese language; event knowledge; words association; emergent features.

Introduction

Synesthetic metaphors such as "*sweet touch*" or "*sweet voice*" are one kind of adjective metaphor, in which an adjective denoting the perception of some sense modality modifies a noun's modality. Metaphor studies in the domain of cognitive science have paid little or no attention to adjective metaphors. Many existing studies have paid much attention to nominal metaphors such as "*My job is a jail*" (e.g., Bowdle & Gentner, 2005; Glucksberg, 2001; Jones & Estes, 2006; Utsumi, 2007) and predicative metaphors such as "*He shot down all of my arguments*" (e.g., Lakoff & Johnson, 1980; Martin, 1992).

Many studies focusing on synesthetic metaphors, including Werning et al. (2006), have examined how the acceptability of synesthetic metaphors can be explained by the pairing of adjective modifier's and head noun's modalities. Ullmann (1951), in a very early study on synesthetic metaphors, proposes a certain hierarchy of lower and higher perceptual modalities. He claims that qualities of lower (e.g., tactile) senses should preferentially occur in the source domain (i.e., adjective), while qualities of higher (e.g., optic) senses should be preferred in the target domain (i.e., noun). After Ullmann, Williams (1976) makes a more differentiated claim of directionality, in which a similar order of sense modalities is proposed. Werning et al. (2006) explores the factors that enhance the cognitive accessibility of synesthetic metaphors for the German language. Very few studies, however, have attempted to explore how people comprehend synesthetic metaphors.

Utsumi & Sakamoto (2007a) is one of the few studies to have explored how people comprehend synesthetic metaphors. They proposed a two-stage categorization theory and argued that the comprehension process of adjective metaphors including synesthetic metaphors could be explained as a two-stage categorization process. The intuitive idea behind two-stage categorization is that correspondences between the properties literally expressed by the adjective and the properties to be mapped onto the noun would be indirect, mediated by an intermediate category. In the case of "*red voice*", for example, the adjective "*red*" first evokes an intermediate category "*red things*," to which "*blood*," "*fire*," "*passion*," "*apple*" and "*danger*" typically belong. Then exemplars relevant to the noun "*voice*" are selected and they evoke a final abstract category of property like "*scary*," "*screaming*" and "*dangerous*." However, they did not mention the relationship between the intermediate category and the noun and the detailed process in which certain exemplars are selected as those relevant to the noun was left unexplored.

In this study we focus on experience-based event knowledge to explain how people comprehend synesthetic metaphors.

Event knowledge has been recognized to be important for metaphor comprehension process by many scholars. For instance, Lakoff & Johnson (1980) argue that metaphors like HAPPY IS UP as in "*She is in high spirits*" and ANGER IS HEAT as in "*boil with anger*" are grounded in correlations in our experience. The HAPPY IS UP metaphor is grounded in the experience that a person in a positive emotional has an erect posture, and the ANGER IS HEAT metaphor is grounded in the experience that the angry person feels hot.

As for synesthetic metaphors, Taylor (2003) argues that they cannot be reduced to correlations. He argues that synesthesic metaphors are based on perceived similarity across different domains. Unlike Taylor (2003), Sakamoto & Utsumi (2008) point out that there are a number of synesthetic metaphors which seem to be based on correlations in experience. For example, a metaphor "*sweet smell*" ("*amai nioi*" in Japanese) is based on correlations in experience. "*Sweet smell*" is the smell you feel when you eat something sweet. A metaphor "*delicious autumn*" ("*oishii aki*") is also based on correlations in experience because you can eat lots of delicious meals in autumn (especially in Japan). However, Sakamoto & Utsumi (2008) did not verify their argument based on psychological experiment.

To sum up, we propose the following comprehension process: an intermediate category is evoked by the adjective to which various things belong. Then exemplars correlated in experience with the noun are selected as those mapped onto the noun and they evoke a final abstract category of property. The experience-based event knowledge plays an important role in the process of relating the intermediate category evoked by the adjective to the concept expressed by the noun.

Experiment

Participants

Participants were recruited through Macromill, Inc., an organization that maintains a panel of more than 533579 people who have agreed to participate in web-based online survey research. 3266 Japanese males and females, aged 20-78, agreed to participate in our experiment.

Materials

Materials used for our experiment (i.e., 62 Japanese synesthetic metaphors) were made by combining 24 Japanese adjectives with 5 Japanese nouns. The adjectives were "*light*" ("*karui*" in Japanese), "*hot*"(in temperature) ("*atsui*"), "*cold*" ("*tsumetai*"), "*hard*" ("*katai*"), "*soft*" ("*yawarakai*"), "*tasty*" ("*oishii*"), "*sweet*" ("*amai*"), "*sour*" ("*suppai*"), "*bitter*" ("*shibui*"), "*hot*"(in taste) ("*karai*"), "*fragrant*" ("*koubashii*"), "*smelly*" ("*namagusai*"), "*sweet-smelling*" ("*kaguwashii*"), "*stinking*"(1) ("*kusai*"), "*stinking*"(2) ("*kinakusai*"), "*red*" ("*akai*"), "*blue*" ("*aoi*"), "*yellow*" ("*kiiroi*"), "*white*" ("*shiroi*"), "*black*" ("*kuroi*"), "*quiet*" ("*shizukana*"), "*noisy*"(1) ("*urusai*"), "*noisy*"(2) ("*yakamashii*"), "*noisy*"(3) ("*sawagashii*"). The nouns were "*color*" ("*iro*"), "*touch*" ("*tezawari*"), "*voice*" ("*koe*"), "*taste*" ("*aji*"), "*smell*" ("*nioi*") .

Procedure

3266 participants were classified into 20 groups. 3-8 linguistic expressions were assigned to each group. The linguistic expressions assigned to one group were randomly assigned to each participant in that group (e.g., linguistic expressions assigned to group 1 were randomly assigned to each participant belonging to group 1).

Participants of group 1-4 were each assigned 7-8 adjectives or nouns, and the remaining 16 groups were assigned 3-4 metaphorical expressions per participant. They were asked to list 3 words associated with each linguistic expression.

Japanese is written with a mixture of hiragana, katakana, and kanji. Hiragana, katakana, and kanji of the same concept (e.g., rose can be written as "ばら," "バラ," or "薔薇") were regarded as the same feature. This feature combination procedure was completed by three judges. Features regarded as the same by at least two judges were unified into one expression, and we got 8594 features. After this combination procedure, all features listed by at most 1 participant were dropped. The following analyses were based upon these amended feature lists.

Analysis 1

According to Becker (1997), when a person interprets a novel metaphor such as "*A child is a sponge*," that interpretation has the potential to contain information from four logically possible sources. The first is a feature which is salient only for the vehicle (i.e., "*sponge*"). Thus, this feature appears in the interpretation for the vehicle and the metaphor. She refers to such a feature as a "vehicle-shared feature." The second is a feature which is salient only for the topic (i.e., "*child*"). Thus, this feature appears in the interpretation for the topic and the metaphor. She refers to such a feature as a "topic-shared feature." The third is a feature which is salient for both the vehicle and the topic. Thus, this feature appears in the interpretation for the vehicle, the topic and the metaphor. She refers to such a feature as a "common feature." The fourth is a feature which is not salient either for the vehicle or for the topic. Thus, this feature appears in the interpretation only for the metaphor. She refers to such a feature as an "emergent feature."

Becker (1997) conducted a psychological experiment for "*A is a B*" metaphors. Participants were divided into two groups. One group of participants listed features of metaphors. The other group of participants listed features of the topic or the vehicle presented alone. Features from metaphor interpretations were compared with features listed for vehicle interpretations and topic interpretations in order to identify the four kinds of features: common, vehicle-

shared, topic-shared, or emergent. These features are shown in Table 1.

Table 1: Features.

features	detail
common	features listed for the metaphor, the vehicle and the topic
vehicle-shared	features listed for both the metaphor and the vehicle, but not listed for the topic
topic-shared	features listed for both the metaphor and the topic, but not listed for the vehicle
emergent	features listed for the metaphor, but not listed for either the vehicle or the topic

The result of her experiment showed that metaphor interpretations contained larger numbers of vehicle-shared and emergent features than either common or topic-shared features. In Particular, there were significantly more vehicle-shared features than the other kinds of features. Furthermore, she found that altering a metaphor's vehicle produced greater changes in emergent content than did altering the topic and suggested that emergent features were influenced primarily by salient features of the vehicle.

In Analysis 1 we compare what Becker (1997) says for the comprehension of nominal metaphors with the comprehension of synesthetic metaphors.

Features listed by participants were classified into one of the four kinds as in Table 1. For each metaphor, the frequency of each of the four kinds was counted. Features were counted both as types (i.e., counted only once no matter how often the feature was listed) and as tokens (i.e., counted as often as the feature was listed). The result was 1198 types and 10388 tokens.

The mean value of common, vehicle-shared, topic-shared, and emergent features are presented in Figure 1 (type counts) and Figure 2 (token counts).

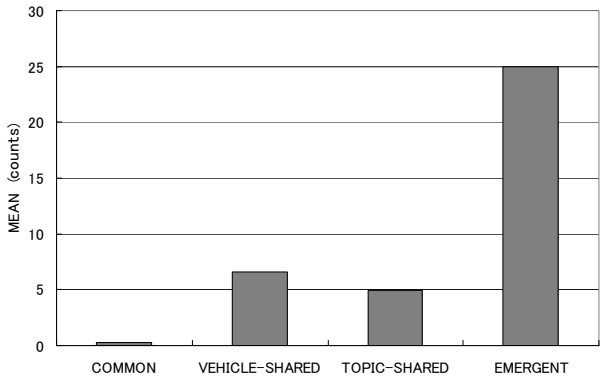


Figure 1: The mean value of the four kinds of features (type counts).

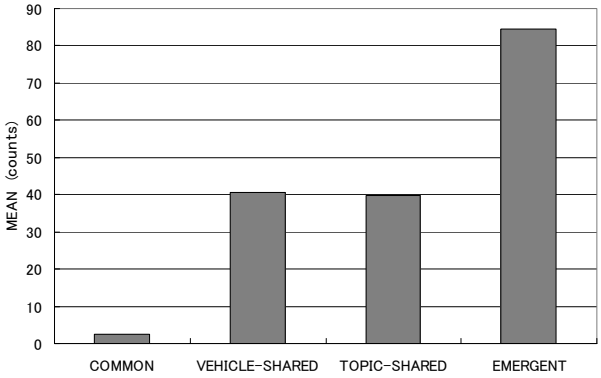


Figure 2: The mean value of the four kinds of features (token counts).

As can be seen from the two figures, regardless of whether one counts features as types or as tokens, participants produced more emergent features than the other kinds of features.

Analyses of variance (ANOVAs) among the four kinds (common, vehicle-shared, topic-shared, emergent) were conducted for both type and token counts. The type count analysis revealed a significant feature type main effect, $F(3, 183) = 456.82$, $p < .001$. Post hoc analyses (Ryan's method) to explore the interaction revealed that significantly more emergent features were produced than the other kinds of features ($p < .05$) and significantly less common features were produced than the other kinds of features ($p < .05$). The token count analysis also produced a significant main effect, $F(3, 183) = 74.79$, $p < .001$. Post hoc analyses (Ryan's method) to explore the interaction revealed that significantly more emergent features were produced than the other kinds of features ($p < .05$) and significantly less common features were produced than the other kinds of features ($p < .05$). In the type count analysis significantly more vehicle-shared features were produced than topic-shared features ($p < .05$), but in the token count analysis this difference was not significant.

These results are different from the results of Becker (1997) which analyzed nominal metaphors. According to Becker (1997), in the interpretation of nominal metaphors there were significantly more vehicle-shared features than the other kinds of features, and nominal metaphors were influenced primarily by salient features of the vehicle. Our results show that in the interpretation of synesthetic metaphors there were significantly more emergent features than the other kinds of features. Thus, our results suggest that we do not so directly comprehend synesthetic metaphors based on salient features of the vehicle or the topic.

Analysis 2

If, as shown in Analysis 1, synesthetic metaphors were not so directly comprehended by salient features of the vehicle

or the topic, where do the emergent features come from? We address this question based on the assumption that the influence of salient features of the vehicle or the topic in the comprehension process of synesthetic metaphors is indirect, mediated by experience-based event knowledge. As we described in the introduction, Sakamoto & Utsumi (2008) suggest that synesthetic metaphors such as "*sweet smell*" ("*amai nioi*" in Japanese) is based on correlations in experience. Thus, in Analysis 2 we explore whether features listed for synesthetic metaphors could be explained by experience-based event knowledge.

Considering experience-based event knowledge and the fact that the vehicle and the topic of a synesthetic metaphor are an adjective and a noun, respectively, we can elaborate the claim of Sakamoto & Utsumi (2008) as follows:

[Hypothesis]
Synesthetic metaphors are interpreted based on event knowledge in which we typically perceive a property denoted by the vehicle (i.e., adjective) and an object denoted by the topic (i.e., noun) simultaneously.

According to this hypothesis, words associated with synesthetic metaphors reflect the process shown in Figure 3; we understand the metaphorical expression as "an object of perception readily evoked by an event in which an entity characterized by the adjective figures prominently." Then we evoke a concrete event in which we typically perceive a property denoted by the adjective and an object denoted by the noun simultaneously. Therefore, words associated with the synesthetic metaphor reflect the evoked concrete event. That is, words associated with the synesthetic metaphor are either feature 1 (hereafter, F1), feature 2 (F2) or feature 3 (F3) in Figure 3.

a synesthetic metaphor
↓
an object of perception readily evoked by an event in which an entity characterized by the adjective figures prominently
↓
a concrete event in which we typically perceive a property denoted by the adjective and an object denoted by the noun simultaneously
↓
feature 1: an entity with a property denoted by the adjective
feature 2: an object denoted by the noun
feature 3: other salient entities or objects in the event

Figure 3: comprehension process of synesthetic metaphors

For example, in the comprehension of "*red taste*" ("*akai aji*" in Japanese), as shown in Figure 4, an event in which we eat chili peppers is evoked as an event in which we perceive "*red*" and "*taste*" simultaneously. This comprehension process is verified when features such as "*chili peppers* (F1)," "*hot* (F2)" and "*sweat* (F3)" are listed for "*red taste*" in the experiment.

"*red taste*"
↓
taste in events including red things
↓
an event in which we eat chili peppers
↓
feature 1: "*chili peppers*"
feature 2: "*hot*"
feature 3: "*sweat*"

Figure 4: comprehension process of "*red taste*"

If this hypothesis is valid, the ratio of features corresponding to either F1, F2 or F3 against all the features collected in the experiment will be very high. Therefore, in Analysis 2, we explore the ratio of features corresponding to either F1, F2 or F3 against all the features collected in the experiment. This exploration is based on the following procedure.

Step 1: Labeling features

This step is a preparation for Step 2 and Step 3. Step 2 and Step 3 are procedures for identifying features corresponding to either F1 or F2.

In this step, features satisfying either of the condition shown in Table 2 are labeled either X or Y. This labeling procedure is conducted based on majority decision of three judges. Hereafter, WXs denote features labeled X and WYs denote features labeled Y. For example, features such as "*chili peppers*" or "*tomato*" listed for "*red taste*" are WXs, and features such as "*hot*" listed for "*red taste*" are WYs.

Notice that at this step we cannot yet determine whether WXs and WYs correspond to either F1 or F2.

Table 2: Labels and Conditions.

label	condition
X	an entity with a property denoted by the adjective
Y	an object denoted by the noun

Step 2: Identifying F1

One situation (hereafter, S1) in which we perceive properties denoted by the adjective (hereafter, PA) and objects denoted by the noun (hereafter, ON) simultaneously is one situation in which there is an entity satisfying both a PA and an ON. In S1, if an entity satisfying both a PA and an ON is a WX, the expression "the ON of a WX" is natural. For example, since "*chili peppers*" for "*red taste*" is a WX, "the ON of a WX" is "*the taste of chili peppers*." This expression is natural.

In this step, therefore, the three judges mentioned in Step 1 consider whether "the ON of a WX" is natural for each synesthetic metaphor. If two or more judges find this

expression natural, the WX is regarded as F1. So, "*chili peppers*" for "*red taste*" is regarded as F1.

Step 3: Identifying F1 and F2

If an entity satisfying both a PA and an ON is evoked as shown in Step2, a possibility, in which our participants will very likely list not only WXs but also a concrete ON of a WX as a feature, is very high. Features corresponding to concrete ONs of WXs are WYs.

Another situation in which we perceive a PA and an ON simultaneously (hereafter, S2) is one situation in which an entity with a PA is different from an object of a category ON. In S2, if an entity with a PA is listed as a feature, the entity is a WX. In S2, if an object of a category ON is listed as a feature, the object is a WY.

If a synesthetic metaphor evokes S1 or S2 for our participants, that will indicate that WXs are strongly connected with WYs. This in turn will make it easy for the three judges mentioned in Step 1 to imagine a concrete event based on the closely associated WXs and WYs.

In this step, the three judges combine all WXs with all WYs for each synesthetic metaphor. If two or more judges can easily imagine concrete events, a WX and a WY comprising the combination are regarded as F1 and F2 respectively.

For example, the features "*chili peppers*" and "*hot*" are listed for "*red taste*." "*Chili peppers*" and "*hot*" are a WX and a WY respectively. The judges make a pair "*chili peppers*, *hot*." Then, they consider whether they can easily imagine an event on the basis of the pair. Since they can imagine an event easily (e.g., eating chili peppers), "*chili peppers*" and "*hot*" are regarded as F1 and F2 respectively.

Step 4: Identifying F3

In this step, we identify F3. Features unlabeled in Step 1 may correspond to F3. If unlabeled features correspond to F3, concrete events are most likely to have already been evoked. Thus, there is a strong possibility that F1 and F2 are included in the features listed by our participants.

In this step, based on this line of reasoning and the rationale presented in Step 3, we conduct the following procedure; the three judges mentioned in Step 1 combine features regarded as either F1 or F2 in Step 2 and Step 3 with unlabeled features. If two or more judges can imagine concrete events easily, the unlabeled feature included in the combination is regarded as F3.

For example, "*chili peppers*", "*hot*" and "*sweat*" are listed for "*red taste*." "*Chili peppers*" and "*hot*" are F1 and F2, respectively, in Step 3. Thus, the judges combine "*sweat*" with the pair "*chili peppers*, *hot*." Since the judges can imagine an event easily (e.g., eating chili peppers), "*sweat*" is regarded as F3.

Result of Analysis 2

All the features regarded as either F1, F2 or F3 are those based on experience-based event knowledge. Table 3 shows the total number of token counts and the mean value of token counts when the features are classified into either those based on event knowledge or those not based on event knowledge. The proportion of the token counts classified as the features based on event knowledge was significantly higher than those which could not be classified as the features on event knowledge, $\chi 2$ (1, N = 10388) = 804.01, p < .01. Furthermore, the T-test using the mean value of token counts revealed that there were significantly more features based on event knowledge than those which could not be classified as features based on event knowledge, $t(61)$ = 3.28, p < .01.

This result shows that synesthetic metaphors tend to be understood based on event knowledge.

Table 3: Classification Result.

	event knowledge	not event knowledge
total	6639 (63.91%)	3749 (36.09%)
mean	107.08	60.47

General Discussion

Indication for the theory of metaphor

Analysis 1 showed that in the interpretation of synesthetic metaphors there were significantly more emergent features than the other kinds of features. The result of Analysis 1 suggests that we do not so directly comprehend synesthetic metaphors based on salient features of the vehicle or the topic.

Utsumi & Sakamoto (2007a, 2007b) proposed a two-stage categorization theory. In the two-stage categorization theory, correspondences between the properties literally expressed by the adjective and the properties to be mapped onto the noun would be indirect, mediated by an intermediate category. Utsumi & Sakamoto (2007a, 2007b) tested their argument by means of computer simulation in which the meanings of adjective metaphors including synesthetic metaphors are computed in a multidimensional semantic space. In the simulation, three theories for adjective metaphor comprehension, i.e., two-stage categorization theory, categorization theory (Glucksberg, 2001; Glucksberg & Keysar, 1990) and comparison theory (Bowdle & Gentner, 2005), were compared in terms of how well they mimic human interpretation of adjective metaphors. The simulation result was that the two-stage categorization theory is a more plausible theory of adjective metaphors than the other kinds of theory.

As for the fact that we do not so directly comprehend synesthetic metaphors based on salient features of the vehicle or the topic, the result of Analysis 1 is consistent with the two-stage categorization theory. Thus, our results support the arguments by Utsumi & Sakamoto (2007a, 2007b). Furthermore, while Utsumi & Sakamoto (2007a, 2007b) left unsolved the detailed process in which certain exemplars are selected as those relevant to the noun, the result of Analysis 2 showed that experience-based event knowledge played an important role in that process.

Importance of Event Knowledge

We showed that experience-based event knowledge play an important role in the comprehension process of synesthetic metaphors. How we use knowledge to interpret new experiences is an important topic in cognitive science. Since 1970's many studies have been conducted based on the concepts of "frame" (Minsky, 1975), "schema" (Rumelhart, 1980) and "script" (Schank & Abelson, 1977). Recent studies such as Bicknell & Rohde (2009) also argue the important role of real-world event knowledge in processing linguistic expressions. Our study showed that experience-based event knowledge also play an important role in the comprehension process of metaphorical expressions.

Conclusion

This paper explored how people comprehend synesthetic metaphors, to which previous studies had paid little attention. The results of psychological experiments showed that there were significantly more emergent features than the other kinds of features. This suggests that we do not so directly comprehend synesthetic metaphors based on salient features of the vehicle or the topic. We argued that experience-based event knowledge played an important role in the comprehension process of synesthetic metaphors.

In our future work we are planning to confirm this finding by different psychological experiments. Since the tendency of synesthetic metaphors to evoke negative images was pointed out by Sakamoto and Utsumi (2009), it would also be interesting for further work to investigate how those negative images were evoked in the comprehension process of synesthetic metaphors using experience-based event knowledge.

Acknowledgments

This research was supported by Grant-in-Aid for Scientific Research (C) (No. 20520350) from Japan Society for the Promotion of Science. The authors wish to thank Yoshiki Nishimura and the anonymous reviewers for their very helpful suggestions.

References

Becker, A. H. (1997). Emergent and common features influence metaphor interpretation. *Metaphor and Symbol*, 12(4), 243–259.

Bicknell, K., & Rohde, H. (2009). Dynamic integration of pragmatic expectations and real-world event knowledge in syntactic ambiguity resolution. In *Proceedings of the 31st annual meeting of the cognitive science society* (pp. 1216–1221).

Bowdle, B. F., & Gentner, D. (2005). The career of metaphor. *Psychological Review*, 112(1), 193–216.

Glucksberg, S. (2001). *Understanding figurative language: From metaphors to idioms*. Oxford: Oxford University Press.

Glucksberg, S., & Keysar, B. (1990). Understanding metaphorical comparisons: Beyond similarity. *Psychological Review*, 97(1), 3–18.

Jones, L. L., & Estes, Z. (2006). Roosters, robins, and alarm clocks: Aptness and conventionality in metaphor comprehension. *Journal of Memory and Language*, 55, 18–32.

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. Chicago: The University of Chicago Press.

Martin, J. H. (1992). Computer understanding of conventional metaphoric language. *Cognitive Science*, 16, 233–270.

Minsky, M. (1975). A framework for representing knowledge. In P. H. Winston (Ed.), *The psychology of computer vision* (pp. 211–277). New York: McGraw-Hill.

Rumelhart, D. E. (1980). Schemata: The building blocks of cognition. In R. J. Spiro, B. C. Bruce, & W. F. Brewer (Eds.), *Theoretical issues in reading comprehension: Perspectives from cognitive psychology, linguistics, artificial intelligence, and education* (pp. 33–58). Philadelphia, PA: Lawrence Erlbaum Associates.

Sakamoto, M., & Utsumi, A. (2008). Semantic diversity revealed by a comparison between two types of adjective metaphors: Correlation vs. resemblance. In *Proceedings of the 6th international conference of cognitive science* (pp.309–393).

Sakamoto, M., & Utsumi, A. (2009). Cognitive effects of synesthetic metaphors evoked by the semantic interaction. In *Proceeding of the 31st annual meeting of the cognitive science society* (pp. 1593–1598).

Schank, R. C., & Abelson, R. P. (1977). *Scripts, plans, goals and understanding*. Philadelphia, PA: Lawrence Erlbaum Associates.

Taylor, J. R. (2003). *Linguistic categorization*. Oxford: Oxford University Press.

Ullmann, S. (1951). *The principles of semantics*. Oxford: Blackwell.

Utsumi, A. (2007). Interpretive diversity explains metaphor-simile distinction. *Metaphor and Symbol*, 22(4), 291-312.

Utsumi, A., & Sakamoto, M. (2007a). Computational evidence for two-stage categorization as a process of adjective metaphor comprehension. In *Proceedings of the 2nd european cognitive science conference* (p. 77-82).

Utsumi, A., & Sakamoto, M. (2007b). Predicative metaphors are understood as two-stage categorization: Computational evidence by latent semantic analysis. In *Proceedings of the 29th annual meeting of the cognitive science society* (pp.1587–1592).

Werning, M., Fleischhauer, J., & Beşeoğlu, H. (2006). The cognitive accessibility of synaesthetic metaphors. In *Proceedings of the 28th annual conference of the cognitive science society* (pp. 2365–2370).

Williams, J. M. (1976). Synesthetic adjectives: A possible law of semantic change. *Language*, 52(2), 461–478.

Simple Pointing to Objects may Facilitate Remembering

Georgi Petkov (gpetkov@cogs.nbu.bg), Prolet Nikolova (bonbonisezoni@gmail.com)
Central and East European Center for Cognitive Science,
Department of Cognitive Science and Psychology,
New Bulgarian University, 21 Montevideo Street
Sofia 1618, Bulgaria

Abstract

Two experiments demonstrate the impact of the self-performed actions during the encoding phase on the amount of the learned information. People memorized more items if they had touched the stimuli during learning. The experiments differ from many of the classical studies testing embodiment of human memory in two main respects:
First, the performed actions are completely unrelated to the essence of the learned stimuli, thus the results can not be explained by pure association-based facilitation. Second, the actions are performed during the encoding phase only, thus the results maybe directly linked to the nature of the encoded representations. The possible mechanisms that may underlie the observed influence are discussed shortly.

Keywords: embodiment; memory encoding; representations

Introduction – embodied cognition

The classical AI theory of manipulations of abstract symbols had been reconsidered during the last decades. Very influential in the field of the so-called embodied view of cognition are the books of Lakoff and Johnson (Lakoff & Johnson, 1980, 1999). They argue for a massive cross-domain interrelation of various structures that map each other. Furthermore, the authors claim that maybe the whole cognition is ground in the body. The growing theory of embodied cognition rejects in its extreme version even the very idea of representations: "What you get underlying our representations of the world - the kinds of things we formulate, for instance, in declarative sentences - is not further representations but rather a certain grasp of the world that we have as agents in it" (Taylor, 1987, p. 432).

The relation between the sensory-motor and the conceptual system was widely explored from different perspectives. O'Regan & Noë, (2001) focus on the dynamic interrelation of all cognitive systems – the perceptual one, the action one, and in turn the conceptual one. The fundamental base for the view that perceptual signals lie at the core of the conceptual representations is enriched by the works of Barsalou (1999) and Brooks (1987). According to the Barsalou's perceptual symbol system, the representations of all concepts, even the abstract ones, are based on associations with huge number of perceptual and motor neural signals. A wonderful theoretical analysis of the philosophical view of embodiment had been made by Anderson (2003), as well by Shepard (1984), Glenberg (1997), and Varela et al. (1991).

At the same time, many empirical studies gave additional support for the idea of embodiment. Sinai et al. (1998) and Proffitt et al. (2003) demonstrated how physical difficulties may change abstract judgments of people. For example, people judge a certain distance as longer if they stay with a heavy rucksack on their back. Myung et al. (2006) performed an experiment to show that recognition of the action, for example, typing on typewriter, is facilitated by the context of a piano just because of the common typical finger movements performed on both objects. In addition, not only the performed actions influence perceptions, but also the perceived objects influence directly some motor commands. Thus, in a series of experiments Tucker and Ellis (Tucker & Ellis, 1998; Ellis & Tucker, 2000), as well as Richardson et al. (2001) demonstrated this opposite effect – the perceived objects automatically and immediately activate certain action responses.

Evidence supporting the embodiment view on cognition can be found also in brain imaging researches (Hauk, Johnsrude & Pulvermüller, 2004, Damasio, 1999). Even the mirror neurons (Rizzolatti & Craighero, 2004) often are speculatively related to this view.

Together with the interdependency between perceptions and actions, the relationship between language and constraints of the body is explored widely by the scientists. For example, people prefer to say that a given umbrella is above the man's head if the umbrella protects the man from the rain even if the real position of the umbrella is at 45 degrees according to the head (Coventry et al., 2001). The claim that language is grounded in our bodies and actions is supported by a huge number of empirical evidences (Glenberg & Kaschak, 2002; Pecher et al., 2003; Solomon & Barsalou, 2001; Spivey et al., 2000; Stanfield& Zwaan, 2001; Zwaan et al.,2002). Catrambone et al. (2006) demonstrated that an irrelevant touch of the objects may facilitate relatively abstract analogy-making. Maybe babies first ground the meaning of verbs that are closely related to the body parts (Tardif & Wellman, 2000).

Actions – Massively associated with the representations or essence of the memorized knowledge

However, most of the empirical studies can not answer whether there are pure symbolic representations of objects in our mind that are massively associated with the action and perceptual representations or indeed the body actions and perceptions are the very essence of the memory traces. Many classic theories in the field of memory and learning assume that learning can be improved if the target

information is processed from different modalities. For example, The Levels of Processing Theory (Craik & Lockhart, 1972), which is deeply based on representational view of memory, is grounded exactly on the interaction between memory and vision (Craik & Lockhart, 1972), memory and hearing (Fletcher et al., 1998; Srinivas et al., 1997), memory and touching (Srinivas et al., 1997), and memory and smell (Schab, 1991). Dual-coding theory (Paivio, 1986) also assumes that the visual and auditory information are processed separately but nevertheless claims for a deep relation between them (Anderson & Bower, 1973). Thus, many empirical data that support embodied view on memory are actually arguments for the relationship between perceptions, actions, and conceptual system but do not contradict in any way to a possible existence of pure symbolic representations of the concepts.

In most of the experiments that test embodied effects on human behavior, the performed actions are closely associated to the respective test items. For example, recognition of a typewriter is faster in the context of a piano (Myung et al., 2006). However, this does not mean that the concrete movements of fingers are inseparable part of the representation of the typewriter. Instead, maybe there are huge number of associative links between the symbolic representation of the concept 'typewriter' and many concrete situations in which a typewriter has been used. Furthermore, maybe the representation of these concrete situations is linked (again associatively) to the representation of the concrete finger movements.

Thus, we decided to conduct an experiment in which the manipulated action is not associatively linked to the essence of the tested items in any way. More concretely, we decided to test whether simple pointing to a colour sample may improve actor's memory of this sample.

In addition, we wanted to ensure that the effect of action should not be manifested during the test phase. With other words, we attempted to avoid the possible explanations that actions and movements may influence the process of retrieval. Thus, we ask participants to perform or not certain actions during the phase of memorizing only. Then people from both acting and not-acting groups were tested in the same way – by asking them to write on a sheet of paper what do they remember.

Experiment 1: The role of the own action

Method

Design

One-factorial between-subject design was used. People from both the control and the experimental group were asked to memorize the colours of twelve small rectangles, placed at different positions on the screen. People from the experimental group were asked to open the rectangles themselves by touching different parts of the touch-screen. When any of the rectangles on the screen was touched a colour appears on its place. Participants from the control group observed the same procedure of opening the colours without touching the screen. The dependent variable was the number of correctly recognized colours on the respective positions.

Stimuli

Six different colours, each of them used twice, were randomly placed on a 4x3 table. The exact positions of each colour, as well the predefined order for their exposure, are shown on table 1. Initially all colours are "closed", i.e. the rectangles were gray.

Table 1: The order of opening and the colour of each rectangle.

6. red	3. blue	8. green	11. orange
10. yellow	12. black	1. red	7. blue
9. green	4. orange	5. yellow	2. black

Procedure

A 4x3 table with 12 gray rectangles was placed at the middle of a touch-screen monitor. When a fixation cross appeared within a certain rectangle, participants from the experimental group touched it and the rectangle changed its colour from gray to one of the six target colours (i.e. red, green. yellow, orange, black and blue). The duration time for the colour presentation was fixed to 1500ms than the rectangle became gray again. One second later the fixation cross appeared on a different location and the procedure was repeated until all twelve rectangles were seen.

Participants from the control group did not touch the screen. Two seconds after the fixation cross the rectangle changed its colour alone for 1500ms and then turned into gray again. Thus, the control participants were only permitted to observe the same procedure as participants in the experimental group but were not actively involved in it.

The order of presentation of the rectangles, as well the position of the colours, was randomly assigned at the beginning of the experiment and was the same for all participants.

The memory test for both groups was performed five minutes later, at which time the participants looked a short movie. Each participant received a sheet of paper with a 4x3 empty table graphed on it. Then he/she was asked to fill the table with the colour labels that he/she can memorized. For each participant the number of positions filled with correct colour labels was counted.

Participants

53 students from New Bulgarian University (26 in the experimental and 25 in the control group) took part as volunteers in the experiment. The range of their age was from 19 to 32 years; 24 of them were males and 29 were females, randomly assigned to both groups.

Results

Nobody had memorized correctly eleven or twelve colours; everybody had memorized at least one; one person had memorized correctly only one colour.

The mean number of correctly recognized colours for the control group was 4.33, st. dev. 1.94. People from the experimental group had recognized correctly mean 5.96, st. dev. 2.60 (see figure 1). The difference was significant: $t(51) = -2.59$, $p = 0.012$; the size effect (Cohen's *d*) was 0.725.

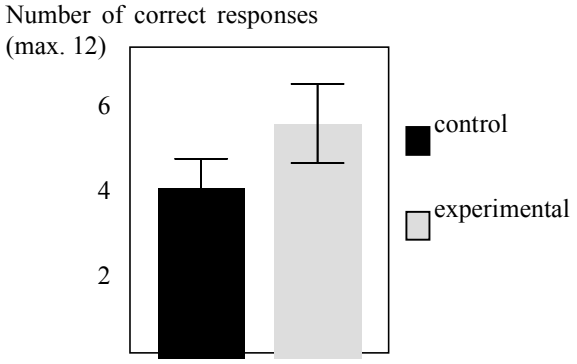


Figure 1. The results from the first experiment - mean number of positions, filled with correct colours during the memory test for both groups. The error bars show 95% CI for the means.

Discussion

The results from the experiment demonstrated that people memorize better when they use their own hand for touching the stimuli during learning. These results differ from most of the empirical data supporting the idea for embodiment of our memory traces because the memorized items were completely unrelated to the specific hand action and because the action was performed during the encoding phase only, but not during the test phase. Thus, the results are in favor of the hypothesis that concrete situated actions, performed by people, are important part and key factor of the representation of relatively abstract and in some sense purely symbolic items.

However, two alternative explanations of the experimental results may arise: First, maybe memories of people from the experimental group are richer because a representation of a movement is added to the representation of each colour position. Thus, because the overall amount of information for the experimental group is larger, maybe the respective memory traces are more accessible. Second, it could be that participants from the control group were less motivated and less involved in the task.

Thus, a second experiment had been performed. The amount of the information that maybe encoded was controlled. In addition, the experiment was performed in an ecological environment by a trained experimenter who tried to keep the attention of people from both groups.

Experiment 2: Control of the amount of information

This experiment differs from the first one in three aspects: First, a manually made cardboard was used instead of a computer touch-screen. Second, the experimenter opened and closed the covers of the coloured rectangles for the control group. This ensured that for both groups somebody's movements can be encoded. Third, the experimenter was trained on several things: to keep the motivation and attention of the participants; to know the exact order of opening the covers; and to keep the time for exposition of the colours as equal as possible.

Method

Design and stimuli

The design of the second experiment followed exactly the respective one from the first experiment. However, the stimuli used differed significantly. A 4x3 cardboard was manually modeled and each of the twelve rectangles was differently coloured. The pattern of the colours followed exactly the respective pattern from the first experiment (see table 1). Twelve gray covers that could be opened were stuck in one side of the rectangles.

Procedure

The experimenter touched one of the covers till the respective participant attended it. Then in the control group she opened the cover for about one and a half second and then closed it. Participants from the experimental group were instructed to open the cover that was pointed from the experimenter themselves and after one and a half second the they closed the cover. The order of presentation of the stimuli was the same as in the first experiment (see table 1).

After the presentation of the twelve stimuli, all participants watched a five minutes movie on a portable computer screen. After that all of them received a graphed sheet of paper and were asked to fill the positions with the colours they remember. Thus, the overall procedure was the same as in the first experiment.

Participants

40 persons (20 women and 20 men) took part as volunteers in the experiment. The range of their age was from 18 to 35 years. All they were randomly assigned to one of the two groups.

Results

Everybody had memorized at least one colour position; two persons had memorized correctly just one; one of the participants had filled correctly all twelve colours.

The mean number of correctly recognized colours for the control group was 4.50, st. dev. 2.37. People from the

experimental group had recognized correctly mean 6.45, st. dev. 2.69 (see figure 2). The difference was significant: t(38) = -2.43, p = 0.020; the size effect (Cohen's *d*) was 0.77.

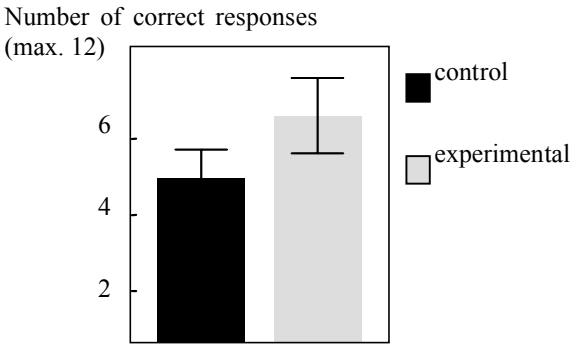


Figure 2. The results from the second experiment - mean number of positions, filled with correct colours during the memory test for both groups. The error bars show 95% CI for the means.

Discussion

During the second experiment the time of colour exposition was not controlled but the amount of the exposed information was equalized for the two groups, i.e. someone's hand opened and closed the covers. At the same time, in the first experiment, any possible influences from the exposure time or from the behavior of the experimenter were eliminated. Nevertheless, similar pattern of results was observed in both experiments. Moreover, although during the ecological experiment people memorized a bit more in both groups, the size effect was almost the same in the two experiments. Thus, the effect of the authentic actions on the amount of the memorized information seems to be stable enough.

Models that can account to the experimental results

Often the phenomena of the embodiment cognition are related to the constructive processes of cognition. Thus, the models of constructive memory like the CHARM model (Metcalfe, 1990), the TODAM2 model (Murdock, 1995), the Trace synthesis model (Nystrom, McClelland, 1992) and the Complementary Learning Systems (CLS) model (McClelland, McNaughton, & O'Reilly, 1995) can account to many of the empirical data that support the embodied view. All these models are based on massively interconnected associative networks. Thus, they can explain the influence of the performed actions to the perceptions (for example, the contest of a piano may facilitate the recognition of a typewriter). They can explain as well the opposite relationship – the perceived objects may automatically activate motor commands.

However, the results from the two experiments can not be explained satisfactory from this type of models. The main reason is that the associative links can be excluded as a possible reason for the effect, because people did not make any movements during the recognition phase.

A second possible explanation of the experimental results can arise from the encoding of the information that comes from proprioception. As it was mention during the discussion of the first experiment, maybe people from the experimental group have richer representations, because their own pointing is an additional portion of information. This was one of the reasons for conducting the second experiment, in which somebody's hand movement can be encoded in both groups. However, the information that comes from the proprioception still is available for the people from the first group only. Unfortunately, there are not any memory models that take into account this type of information.

Ballard and colleagues (Rao, Ballard, 1995, Ballard et al., 1997) propose their model, based on the idea for deictic pointers. According to the authors, eye-fixations and the attention serve for creation and manipulation of pointers to the objects in the environment. The pointers, instead of the representation of the objects can be encoded in the memory. If necessary, it is easy to use these pointers for finding the objects and to encode from them the necessary information. From one hand, the deictic codes model proposes a way for a drastic decreasing of the necessary calculations for performing tasks in a 3-D environment. From other hand, they answer to the question why people are limited for the amount of information that can process simultaneously.

The paradigm of the proposed experiments seems very close to the deictic codes view. Maybe pointing to the objects with a hand is an additional source for creating such deictic pointers. Thus, it seems natural why people from the experimental group have better memories – they can just use more deictic pointers.

Unfortunately, the model that Ballard and colleagues propose is still not enriched with mechanisms for creation, manipulation, and retrieval of memory traces from the long term memory. Thus, the relation between the deictic codes view of the embodiment and the results from the two experiments seems promising but still speculative.

Finally, maybe pointing to the objects has additional social value that in turn can influence memory. (Nathan, in press) proposes various examples how gestures may enrich the listener' understanding as well as the speaker's one during a conversation. Thus, people from the experimental group point to the colour samples and they point not only for themselves but to the experimenter too. Maybe this is the reason for their better memory.

Conclusion

The theory of embodiment of memory and cognition opposes in its extreme versions the classical representational based view. Many empirical data support the close relationships and interdependencies between our conceptual system and the sensory-motor inputs and outputs. However, it is still an open question whether actions and perceptual signals lie at the very core of the memory traces or it is just a massive associative interconnection between the separate conceptual and sensory-motor systems.

The two related experiments, presented here, tried to highlight more this question. People memorized better colour samples if they touched them instead of just observed them. The same effect of the action of touching has been observed both during the controlled laboratory experiment and during the more ecological second experiment.

It is demonstrated that the simple touching influences what people had learned even if the respective touching is not related in any way to the essence of the information, required to be learned. Thus, the results are in support of the hypothesis that the movements that we perform maybe are a substantial part of the representation of the things we learn during these movements. This hypothesis is supported additionally by the fact that the observed effect is caused by what is actually encoded, not by any influences of the actions during the recall phase.

The experiments, however, do not highlight any possible mechanisms that may underlie the observed effect. It is not clear whether the concrete touching influences the attention, the way of encoding, both, or something else.

Nevertheless, the observation that simple touching can influence the rate of memorizing of relatively abstract items seems promising for further investigations.

Acknowledgments

This material is based on research sponsored by the ANALOGY project (NEST program, contract 29088) funded by the EC and the Air Force Research Laboratory, under proposal number 103061**.** The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Air Force Research Laboratory or the U.S. Government or the European Commission.

References

Anderson, M. (2003). Embodied Cognition: A field guide. *Artificial Intelligence 149*, 91–130

Anderson, J. & Bower, G. (1973). Human associative memory. Washington, DC: Winston.

Ballard, D., Hayhoe, M., Pook, P., Rao, R. (1997). Deictic codes for the embodinment of cognition. *Behavioral & Brain Sciences 20*, 723–767.

Barsalou, L. (1999). Perceptual symbol systems. *Behavioral & Brain Sciences 22*, 577–660.

Brooks, R. (1987). Intelligence without representation. *Arti.cial Intelligence 47*, 139–159.

Catrambone, R.,Craig, D., & Nersessian, N. (2006). The role of perceptually represented structure in analogical problem solving. Memory & Cognition 34 (5), 1126-1132

Coventry, K. R., Prat-Sala, M., & Richards, L. (2001). The interplay between geometry and function in the comprehension of 'over', 'under', 'above' and 'below'. *Journal of Memory and Language*, 44, 376-398.

Craik, F., Lockhart, R. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning & Verbal Behavior* 11 (6), 671-84.

Damasio, A. (1999). The Feeling of What Happens: Body and Emotion in the Making of Consciousness. *M.D.; New York, Harcourt Brace & Company.*

Ellis, R. & Tucker, M., (2000). Micro-affordance: The potentiation of components of action by seen objects. *British Journal of Psychology 91(4)*, 451-471.

Fletcher, P., Shallice, T., Dolan, R. (1998). The functional roles of prefrontal cortex in episodic memory. I. Encoding. *Brain* 121 (7), 1239–1248.

Glenberg, A. & Kaschak, M. (2002). Grounding language in action. *Psychonomic Bulletin & Review 9*, 558–565.

Glenberg, A. (1997). What memory is for? *Behavioral and Brain Sciences 20*, 1–55.

Hauk, O., Johnsrude, I., & Pulvermüller, F. (2004). Somatotopic representation of action words in the motor and premotor cortex. *Neuron, 41*, 301-307.

Jong-yoon Myung, Blumsteinp S., Sedivy, J. (2006). Playing on the typewriter, typing on the piano: manipulation knowledge of objects. *Cognition 98 (2006)*. 223–24.

Lakoff, G., Johnson, M. (1999). Philosophy in the Flesh: The Embodied Mind and Its Challenge to Western Thought, Basic Books, New York.

Lakoff, G & Johnson, M. (1980). Metaphors we live by. Chicago: University of Chicago Press.

McClelland, J.L., McNaughton, B.L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hyppocampus and neocortex: Insights from the successes and failures of connectionists models of learning and memory. *Psychological Review*, 102, 419-457.

Metcalfe, J. (1990). Composite holographic associative recall model (CHARM) and blended memories in eyewitness testimony. *Journal of Experimental Psychology: General, 119,* 145-160.

Murdock, B. B. Jr. (1995). Developing TODAM: three models for serial-order information. *Memory & Cognition, 23,* 631-645.

Nathan, M. (in press). An embodied cognition perspective on symbols, grounding, and instructional gesture. In

DeVega, M., Glenberg, A, M. & Graesser, A. C. (Eds.) Symbols, Embodiment and Meaning: A Debate (pp. 375-396). Oxford, England: Oxford University Press.

Nystrom, L., McClelland, J. (1992). Trace synthesis in cued recall. *Journal of Memory and Language, 31*, 591-614.

O'Regan, J., Noë, A. (2001). A sensorimotor account of vision and visual consciousness, *Behavioral Brain Sci.* 24 (5).

Paivio, A (1986). Mental representations: a dual coding approach. Oxford. England: Oxford University Press.

Pecher, D., Zeelenberg, R.,& Barsalou, L. (2003).Verifying conceptual properties in different modalities produces switching costs. *Psychological Science 14*, 119–124.

Proffitt, D., Stefanucci, J.,Banton, T.,Epstein, W. (2003). The role of effort in perceiving distance. *Psychological Science*, Vol. 14, No. 2. (March 2003), pp. 106-112.

Rao, R.P.N., Ballard, D.H. (1995). Object indexing using an iconic sparse distributed memory. In Proceedings of the 5th International Conference on Computer Vision, Cambridge, Massachusetts, USA, p. 24-31.

Richardson, D., Spivey, M., Cheung, J. (2001). Motor Representations In Memory And Mental Models: The Embodied Zork. *In: Proceedings of the Twenty-third Annual Meeting of the Cognitive Science Society*, (pp.867-872), Erlbaum: Mawhah, NJ

Rizzolatti, G., Craighero, L. (2004). The mirror-neuron system, *Annual Review of Neuroscience* 27, 169–192.

Schab, F. (1991). Odor memory: taking stock.. *Psychological Bulletin* 109 (2), 242–51.

Shepard, R. (1984). Ecological Constraints on Internal Representation: Resonant Kinematics of Perceiving, Imagining, Thinking, and Dreaming. *Psychological Review*, Vol. 91, 4.

Sinai, M., Ooi, T.,He, Z. (1998). Terrain influences the accurate judgment of distance. *Nature*, Vol. 395, 6701. pp. 497-500.

Solomon, K., & Barsalou, L. (2001). Representing properties locally. *Cognitive Psychology 43*, 129–169.

Spivey, M., Tyler, M., Richardson, D., & Young, E. (2000). Eye movements during comprehension of spoken scene descriptions. *Proceedings of the 22nd Annual Conference of the Cognitive Science Society* (pp. 487–492). Mahwah, NJ: Erlbaum.

Srinivas, K., Greene, A., Easton, R. (1997). Visual and tactile memory for 2-D patterns: Effects of changes in size and lef-right orientation. *Psychonomic Bulletin & Review* 4 (4), 535–540.

Stanfield, R., & Zwaan, R. (2001). The effect of implied orientation derived from verbal context on picture recognition. *Psychological Science 12*, 153–156.

Tardif, T. & Wellman, H.M. (2000). Acquisition of mental state language in Mandarin- and Cantonese-speaking children. *Developmental Psychology, 36*, 25-43.

Taylor, C. (1987). Overcoming epistemology, in: Baynes, et al. (Eds.), After Philosophy: End or Transformation? MIT Press, Cambridge, MA.

Tucker, M., & Ellis, R. (1998) On the relations between seen objects and components of potential actions. *Journal of Experiemental Psychology: Human Perception and Performance, 24(3),* 830-846.

Varela, F., Thompson, E., Rosh, E. (1991). The Embodied Mind, MIT Press, Cambridge, NA

Zwaan, R., Stanfield, R., & Yaxley, R. (2002). Language comprehenders mentally represent the shapes of objects. *Psychological Science 13*, 168–171.

A Bayesian Antidote Against Strategy Sprawl

Benjamin Scheibehenne (benjamin.scheibehenne@unibas.ch)
University of Basel, Missionsstrasse 62a
4055 Basel, Switzerland
&
Jörg Rieskamp (joerg.rieskamp@unibas.ch)
University of Basel, Missionsstrasse 62a
4055 Basel, Switzerland

Abstract

Many theories in cognitive science assume that people possess a repertoire of strategies or a "toolbox" from which they choose depending on the situation. This approach suffers from the problem that the number of assumed strategies is often not constrained and may be extended post-hoc to improve the fit to the data. This makes it difficult to rigorously test and compare strategy repertoire models. To prevent this "strategy sprawl", a criterion is necessary to decide how many strategies a toolbox should include. Here, Bayesian statistics provide a powerful tool to evaluate toolboxes of different sizes based on their marginal likelihoods. The present work illustrates how such a Bayesian approach can be implemented and demonstrates its applicability by means of parameter recovery studies. Our approach also makes the novel contribution of showing how Bayesian statistics allow testing the strategy repertoire theory against alternative decision theories.

Keywords: Strategy repertoire theories, Bayes factor, model selection, simulation, Bugs.

The problem of strategy sprawl

A common assumption within many research areas in cognitive science is that people possess a repertoire of cognitive strategies to solve the problems they face. For example, people use different strategies for making consumer decisions (Payne, Bettman, and Johnson, 1993), for organizational memory tasks (Coyle, Read, Gaultney, & Bjorklund, 1998), for estimations of frequencies (Brown, 1995), for categorization problems (Patalano, Smith, Jonides, & Koeppe, 2001), for the development of mathematical skills (Siegler, 1991), or for inference problems (Gigerenzer, Todd, and the ABC Research Group, 1999). The strategy repertoire approach provides a fruitful way to explain intra- and inter-individual differences in cognitive processes. This approach has also been described by the metaphor of an adaptive toolbox according to which individual decision makers select between different cognitive strategies to solve specific tasks just as a craftsman selects tools from a toolbox.

Despite its undisputed success in explaining a wide range of human behavior, the idea of a toolbox raises the question of how many different strategies the mental toolbox should contain in the first place. A larger number of possible strategies will always lead to a better description of the data but not necessarily to greater insight. For example, by assuming a specific tool for each possible task, the toolbox should provide a good description of observed behavior due to its great flexibility. Along the same lines, Dougherty, Thomas, and Franco-Watkins (2008) criticized that in a situation in which no strategy out of a set is able to describe a person's choices, an unconstrained toolbox could be enlarged by a new strategy to describe the data. On the other hand, if the toolbox is restricted to only a few or to a single strategy it would lose its ability to describe different cognitive processes.

In the following, we outline a possible solution to the question of how many tools a toolbox should contain based on a Bayesian approach. Having a criterion for determining how many strategies to include keeps strategy sprawl at bay and is also a necessary pre-condition for rigorously comparing different toolbox models with competing cognitive theories that do not assume different strategies but rely on the idea of an "all-purpose" process (Newell, 2005).

Example of a cognitive toolbox

As an illustrative example of a cognitive toolbox, imagine a situation in which a person tries to determine which of two used cars is a better deal. To make this decision, a person could use different pieces of information (i.e., cues) such as mileage, number of previous owners, or accident history of the cars. In such a situation, each single cue provides a hint of which car might be better, but none of the cues provide an indisputable prediction, because it could be that a car with many previous owners still turns out to be superior overall. In other words, the cues are probabilistically related to the criterion, so that even an object with positive cue values for all cues could sometimes be inferior compared to an object with negative cue values. Probabilistic inferences can be complicated because it is not always clear which information is relevant and how and whether the different pieces of information should be combined.

To make probabilistic inferences, such as choosing the better of two cars, people may choose from a variety of cognitive strategies, that is, from their adaptive toolbox (Gigerenzer, Todd & the ABC Research Group, 1999). For instance, when choosing between two options, people could use a simple non-compensatory decision strategy called take the best (TTB) that only focuses on the single most important or valid cue that discriminates between the two

options. If the most valid cue discriminates between the alternatives, TTB chooses the object with the positive cue value. Only in cases where the most valid cue does not discriminate does TTB then consider the second most valid cue, and so on. An alternative strategy in the toolbox could be a weighted additive (WADD) rule. This model adds up all available cue values weighted by their respective validities and then selects the alternative with the largest score. The WADD rule is compensatory because a highly valid cue may be compensated for by a number of other cues that point to the opposite choice.

This example illustrates that different decision strategies can be applied to make a choice between two options that are described by several attributes. Here, proponents of a toolbox approach could argue that people use either TTB or WADD depending on the decision situation and the characteristics of the decision maker.

Preventing strategy sprawl

When examining how people solve an inference problem researchers aim to identify the model that best describes the cognitive processes, that is, the one that most likely generated the observed data. Under the assumption that people have a repertoire of strategies, the goal is to identify the strategy that a person has selected. Possible strategies are not known a priori. Therefore, a researcher may add more and more strategies to the toolbox to increase chances that one strategy provides a good fit to the data. From a model comparison perspective, a given toolbox becomes more flexible and complex when more strategies are added. Accordingly, it is not surprising when it provides a better description of the observed data. Therefore, the question of how many strategies to include in the toolbox essentially becomes a trade-off between the complexity of the toolbox and its fit in describing observed data. Adding another strategy is only justified if it increases the fit substantially. Bayesian techniques offer a valuable theoretical framework to make this trade-off and to identify a toolbox that fits the data well. More precisely, the probability of a specific dataset given a specific toolbox model (referred to as the evidence or marginal likelihood of that model) can be used as a criterion of how many tools to include.

Bayes' theorem

In a Bayesian framework, the marginal likelihood $p(D)$ is a measure of how well a given model M describes the observed data D across all possible parameter values of that model (Kass & Raftery, 1995; Shiffrin, Lee, Kim, & Wagenmakers, 2008):

$$p(D) = \int p(D|M) \times p(M) \mid \mathrm{d}M \qquad (1)$$

where $p(D|M)$ is the likelihood of the observed data given the model and $p(M)$ is the prior probability of the model. The evidence provides a viable metric to compare different models against each other. However, to eventually apply this criterion the possible toolbox models and the strategies within each toolbox need to be specified in a Bayesian framework. In particular, this requires the specification of prior distributions and likelihood functions.

In the following, we lay out the necessary specifications in detail. To illustrate the basic principle behind this approach, we start with the simple example of comparing two toolboxes consisting of only one strategy each, before proceeding to the more complex case of comparing toolboxes of different sizes.

Model specification for a simple toolbox

As a first step, we compare two toolboxes that only consist of one single strategy each. For illustrative purposes, we assume that the first toolbox consists of TTB, described above. In its basic form, TTB is a deterministic strategy that assumes people make no errors. This is a rather unrealistic assumption because if someone uses TTB but occasionally makes an error, strictly speaking the resulting choice data would contraindicate the application of the strategy. Therefore we allow for the possibility of inconsistent choices due to application errors or "unsystematic noise". We extended the deterministic model with a simple error theory, so that a parameter α_{TTB} indicates the probability that a decision maker will chose the alternative that was not predicted by TTB in a pairwise choice. In the following, we refer to this probabilistic version of TTB as TTB_α. Other deterministic choice strategies such as WADD can be extended by similar error terms, in an analogous manner.

Specifying the prior distribution

In the case of a toolbox that only consists of TTB_α as a single strategy, the only free model parameter is the application error α'_{TTB}. A reasonable prior on α'_{TTB} may be to assume an average application error of about 10%. Of course, other values are also possible. In any case, the application error will probably vary depending on the situation and the type of experiment. Therefore, a moderate degree of uncertainty concerning the prior distribution seems justified. As α'_{TTB} may fall within a range from 0 (no implementation error) to 1 (100% implementation error), we choose the prior to be beta distributed. For illustrative purposes, we set the rate and shape parameters of this beta distribution to 1 and 9, resulting in a mean of 0.1 and a standard deviation of 0.09.

$$\alpha'_{TTB} \sim \mathrm{beta}(1, 9) \qquad (2)$$

Specifying the likelihood function

Next, a likelihood function needs to be defined that indicates the probability of the observed data given the model. The predictions of a deterministic choice model like TTB are readily available as long as the attributes of options in the experiment are known. In this case, the likelihood is just a function of the implementation error α. If a single choice between a pair of options is in line with the deterministic predictions of the model for that pair, then likelihood of that choice equals $1 - \alpha$, otherwise, the

likelihood equals α. Hence, the likelihood of a series of N choices in an experiment is the product of likelihood values for all pairwise choices:

$$p(D|M) = \Pi\,[d_n \times (1-\alpha) + (1-d_n) \times \alpha] \quad (3)$$

where d_n is 1 if the decision for the nth pair of options is in line with the deterministic prediction of the model for that pair and 0 otherwise.

For illustrative purposes, suppose a participant in an experiment made z pairwise choices that were inconsistent with TTB's deterministic prediction and $N - z$ choices that were consistent with it. Following Equation 3, the likelihood of this data can be expressed as a Bernoulli function:

$$p(D|\mathrm{TTB}_\alpha) = \alpha'^{(z)}_{\mathrm{TTB}} \times (1-\alpha'_{\mathrm{TTB}})^{(N-z)} \quad (4)$$

Deriving $p(D)$ for a single strategy

Once the prior and the likelihood function of a model are specified, the evidence for the observed data $p(D)$ can be estimated. In the present case of a single strategy with one free parameter α, a closed-form solution exists:

$$p(D) = B(a+z,\, b+N-z) \,/\, B(a,b) \quad (5)$$

where B is the beta function, a and b are the rate and shape parameter of the beta distribution that defines the prior, and z quantifies how many out of a total of N choices are in line with the prediction of the deterministic model.

Specification for WADD

The outlined approach for TTB_α can be conveyed to other deterministic strategies like WADD. Like TTB, WADD can be extended with a similar beta-distributed error term leading to WADD_α. The prior distribution and likelihood function for WADD_α can be specified analogously to TTB_α with the only difference being different deterministic predictions of WADD.

Comparison between two simple toolboxes

Now that two toolboxes are specified, they can be compared with regard to a given set of data. This case is analogous to a model selection task in which an individual decision maker can be classified as a TTB_α or WADD_α user.

To lay out the approach in a concrete way, we assume a hypothetical decision experiment in which a single participant made 40 choices among pairs of options described on a number of attributes, as in the used-car example outlined in the introduction. We further assume that the options were chosen to differentiate TTB from WADD, such that both strategies would make opposing predictions. In this example, a decision maker chooses option *A* 30 times and *B* 10 times. If we set the prior as beta(1,9) for both models, we can calculate the respective marginal likelihoods according to Equation 2. For TTB_α, this yields $p(D) = 2.8 \times 10^{-11}$ for TTB_α as compared to $p(D) = 2.5 \times 10^{-14}$ for WADD_α. The ratio of the marginal likelihoods, also known as the Bayes factor (Kass & Raftery, 1995), is 1118:1 in favor of TTB_α. Therefore, Bayes' rule clearly indicates that the decision maker should be classified as a TTB_α user.

Next, we outline how this procedure can be extended to toolbox comparisons that include more than one decision strategy.

Specifying toolboxes with more than one strategy

The concept of a cognitive toolbox relies on the idea that decision makers have several decision strategies available to them. To account for this assumption, we extend the Bayesian approach to toolboxes that contain more than one strategy. Again, precise model specifications are required so that Bayesian techniques can be applied.

Model specification

For illustrative purposes, we assume toolbox $\mathrm{TB}_{\mathrm{TTB,WADD}}$ contains two strategies, TTB_α and WADD_α. We further assume that an individual decision maker who uses this toolbox will choose according to TTB_α with probability β and according to WADD_α with the complementary probability (1 − β). Thus, $\mathrm{TB}_{\mathrm{TTB,WADD}}$ has three free parameters: The implementation error for TTB in the toolbox (α_{TTB}), the implementation error for WADD in the toolbox (α_{WADD}), and the probability of selecting TTB_α (β). The likelihood function of this toolbox is simply a function of the likelihood for each single strategy weighted by the probability of selecting it:

$$p(D|\mathrm{TB}_{\mathrm{TTB,WADD}}) = \beta \times p(D|\mathrm{TTB}_\alpha) + (1-\beta) \times p(D|\mathrm{WADD}_\alpha) \quad (6)$$

Next, a prior distribution for each of the three parameters needs to be specified. Without any prior knowledge about the probability of selecting TTB_α over WADD_α all possible values between 0 and 1 seem equally likely a priori. Accordingly, we assume that the prior on β is uniformly distributed:

$$\beta \sim \mathrm{uniform}(\min = 0,\, \max = 1) \quad (7)$$

Likewise, in this example we do not make any a priori assumptions about the probability of particular implementation errors.Thus, we assume priors for α_{TTB} and α_{WADD} are uniformly distributed. Based on these specifications, the marginal likelihood of $\mathrm{TB}_{\mathrm{TTB,WADD}}$ can be estimated by integrating out all three parameters in the model analogous to Equation 1:

$$p(D) = \iiint p(D|\,\alpha_{\mathrm{TTB}}, \alpha_{\mathrm{WADD},}, ß) \times p(\alpha_{\mathrm{TTB}}, \alpha_{\mathrm{WADD},}, ß) \;|\; d\alpha_{\mathrm{TTB}}, d\alpha_{\mathrm{WADD},}, dß \quad (8)$$

While this approach is conceptually similar to the case with only one free parameter, it becomes more elaborate to

estimate the integral of Equation 1 as the number of free parameters increases.

MCMC methods to estimate the evidence

Fortunately, a closed-form mathematical solution, is not mandatory to estimate marginal likelihood because all that is needed is a representative sample from the integral of Equation 1 that is large enough to draw reliable conclusions on the shape of the distribution. Such a sample may be obtained by means of Monte Carlo Markov Chain (MCMC) methods for which statistical packages are readily available (Gilks, Richardson, & Spiegelhalter, 1996). For the present analysis, we utilized utilized the *OpenBugs* software implemented in the *BRugs* package, version 0.51, that can be integrated into the statistics software *R*.

Similar to the mathematical solution outlined above, the implementation in *BRugs* requires the specification of prior distributions and likelihood functions. Provided it is properly implemented, the software returns the evidence as well as a representative sample of the full posterior distribution across all parameters.

Comparison between a small and a large toolbox

To illustrate the principle of comparing toolboxes that differ in the number of cognitive strategies they contain, we will compare a simple toolbox TB_{TTB} that only consists of TTB_{α} as a single strategy to a more complex toolbox $TB_{TTB,WADD}$ that contains both TTB_{α} and $WADD_{\alpha}$. Thus, TB_{TTB} is nested within $TB_{TTB,WADD}$.

Transdimensional prior

Instead of calculating the evidence for both toolboxes separately, we directly compared the two models by means of a transdimensional prior θ. This prior acts like a model indicator, controlling which of two models most likely generated the observed data. Thus, θ immediately informs us which of the two toolboxes best describes the choices of an individual decision maker. The parameter θ is Bernoulli-distributed with a prior that assigns equal probabilities to both models (Han & Carlin, 2001; Shiffrin, et al., 2008). Like any other estimated parameter in the model, θ is updated during the course of MCMC simulation. The Bayes factor (*BF*) is simply the odds ratio of this probability, that is, $BF = \theta/(1 - \theta)$. Figure 1 graphically depicts the model's implementation in *OpenBugs*. The graph follows the notation used by Lee and Wagenmakers (2009) and Shiffrin et al. (2008).

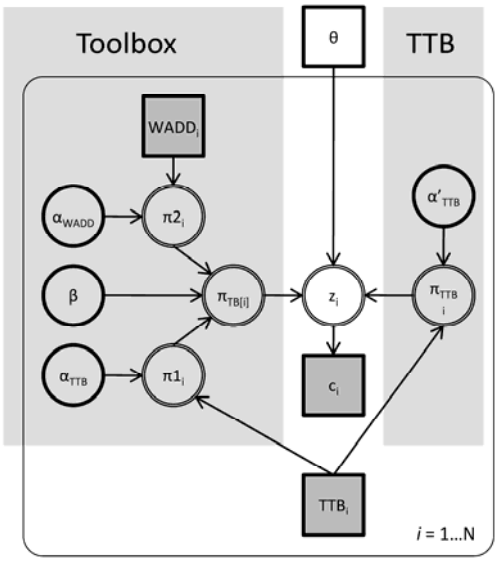


Figure 1: Graphical representation of the model comparison. TTB_i and $WADD_i$ depict the deterministic predictions for each choice *i* out of *N* choices. The notation *c* indicates the actual choice (*A* or *B*) and π depicts the probability of choosing *A* over *B* as predicted by the particular model.

Model recovery

To test this approach, we set up a model recovery simulation based on 1,000 pairs of options described on 30 attributes to ensure that the results would not be influenced by an extreme constellation of binary attribute values that were randomly drawn from a Bernoulli distribution with $p = .5$. The importance weights for each attribute were set to increase linearly from the least to the most important attribute. Next, we simulated hypothetical decision makers who repeatedly chose among the pairs of options according to either TTB_{α} or $WADD_{\alpha}$ (i.e., ß was set to either 1 or 0). The implementation error α for both decision strategies varied from 0 to 0.5 in steps of 0.1 across decision makers. For each value of α in the simulation, we estimated θ by sampling from three separate MCMC chains in OpenBugs that were run for 2,000 steps each with a thinning of 10.

For $TB_{TTB,WADD}$ the priors were set to uniform distributions ranging from 0 to 0.5 for α_{TTB} and α_{WADD} and from 0 to 1 for β. For TB_{TTB}, the prior on α'_{TTB} was set to a uniform distribution ranging from 0 to 0.5.

Predictions

If a comparison based on Bayesian evidence is a feasible way to solve the problem of strategy sprawl, the method should assign more evidence to the model that generated the data. Thus, if the data was generated by choices according to TTB_{α}, the evidence for a simple toolbox TB_{TTB} should be higher than that of a toolbox $TB_{TTB,WADD}$ even though the latter one contains TTB_{α} as a special case.

Results

The samples from the three estimated chains in OpenBugs provided representative samples as indicated by a visual inspection of the trace plot, the autocorrelation and the Gelman−Rubin statistic. The results clearly indicate that the marginal likelihood for a smaller toolbox can indeed be higher than that for a larger toolbox. Figure 2 plots the actual implementation error α_{TTB} for choices based on TTB_α (Figure 2a) and the actual implementation error α_{WADD} for choices based on $WADD_\alpha$ (Figure 2b) against the estimated θ. Here, θ indicates the probability of the more complex $TB_{TTB,WADD}$ over TB_{TTB}. As can be seen from Figure 2a, θ increases as α_{TTB} increases. This indicates that a decision maker who uses TTB_α with a small implementation error is better described by TTB_α as compared to $TB_{TTB,WADD}$. Likewise, a decision maker who chooses according to $WADD_\alpha$ is clearly better described by $TB_{TTB,WADD}$ even if the implementation error α_{WADD} is large. This relationship seems plausible because the larger toolbox contains $WADD_\alpha$ whereas the smaller toolbox does not.

For an α of 0.5, which indicates random choice, Bayes' rule tells us to favor TTB_α, because in the case of very noisy data a simpler model is favored over a more complex one. Together, the results show that a small toolbox with only one strategy should be preferred over a more complex one provided that the small toolbox contains the appropriate tool to describe the initial decision-making process.

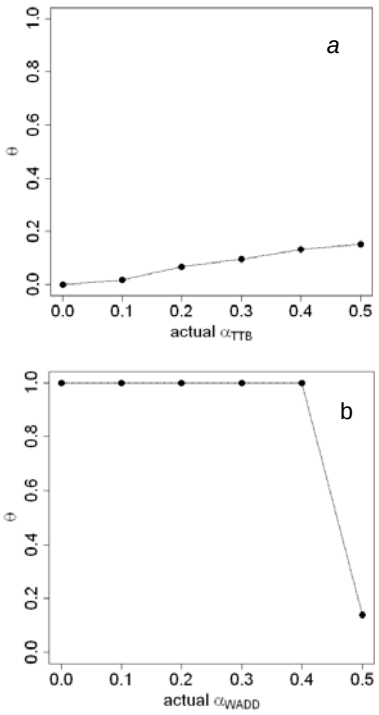


Figure 2: Plot of θ against the implementation error α_{TTB} (Figure 2a) and α_{WADD} (Figure 2b) used in the simulation. θ indicates the evidence in favor of $TB_{TTB,WADD}$ relative to TB_{TTB}.

Parameter recovery

The results so far show that Bayesian methods can be fruitfully applied to solve the question of how many strategies a toolbox should contain. However, we have implicitly assumed that the estimation methods accurately estimate the free parameters of the choice models within the toolboxes. To test if this condition really holds, the actual parameters used in the choice simulation can be compared with their respective posterior estimates. Figure 3a shows the marginal of the posterior distribution for α_{TTB} in the toolbox plotted against the actual parameters used in the simulation. As can be inferred from Figure 3, the posterior distributions of α_{TTB} match the actual values of α_{TTB} used in the simulation quite well. The parameter recovery for the α_{TTB} parameter of the simple TTB_α model appears similar.

Figure 3b shows the estimated β values plotted against α_{TTB} parameters used in the simulation. For low values of α_{TTB}, estimated β values are high. As β indicates the probability of using TTB_α over $WADD_\alpha$ within the toolbox, this relationship also seems plausible.

Figure 3 also shows the actual values for α_{WADD} from the choice simulation plotted against the posterior of α_{WADD} (Figure 3c) and β (Figure 3d) in the toolbox. Again, the estimated values match up with the actual values, indicating that the parameters were recovered across the whole parameter space. For high values of α_{WADD} the Bayesian model slightly underestimates the implementation error. Presumably this is the case because the prior distribution constrains the parameter space between 0 and 0.5. If the prior distribution is extended to range from 0 to 1, the estimated parameters match more closely.

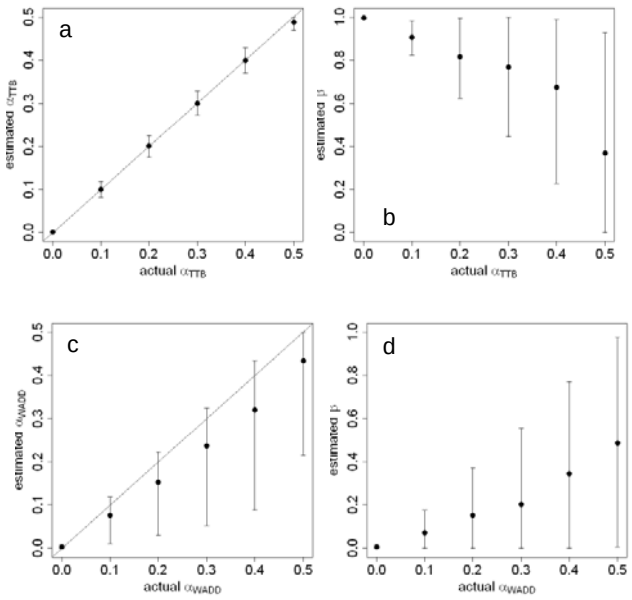


Figure 3: Parameter recovery $TB_{TTB,WADD}$ for choices according to TTB_α (Figure 3, a & b) and $WADD_\alpha$ (Figure 3, c & d). Error bars indicate the 95% highest probability density region of the posterior distribution.

Discussion

Due to higher degrees of freedom, a cognitive toolbox that includes many strategies will always provide a better fit to the data, but it will not necessarily provide the highest evidence or marginal likelihood. The results of our analyses indicate that the marginal likelihood within a Bayesian framework can be fruitfully used to determine the number of tools to include in a toolbox. Thus, Bayesian statistics are well suited to prevent strategy sprawl.

As outlined above, the marginal likelihood can only be estimated with regard to a specific set of choice data and to precisely defined cognitive strategies. Thus, the number of tools to include may vary depending on cognitive strategies included in the toolbox. Yet, within these boundaries, the approach indicates that a small toolbox may be preferred over a large toolbox if the small toolbox contains a tool that describes the data well—even if the small toolbox is nested within the larger one.

The reason why the marginal likelihood provides a common comparison metric is because it implicitly accounts for differences in model complexity. This happens because the prior probability of each possible combination of parameters decreases with an increase in parameters. This carries over to the marginal likelihood that weights the likelihood of the data by the probability of each combination of possible parameter values. Thus, even though the likelihood of the data can be expected to increase with more free parameters, this increase is counteracted by a lower prior probability for each possible combination of parameters.

A comparison of the prior distributions of TB_{TTB} and $TB_{TTB,WADD}$ illustrates this mechanism. The prior probability of TB_{TTB} follows a beta-distribution around a single parameter α. On the other hand, the prior probability of the parameters in $TB_{TTB,WADD}$ embrace a total of three parameters. As prior distributions are probability distributions, they must integrate to 1. With more parameters in the model, the probability of each specific combination of parameter values should decrease because parameter space is more spread out.

Limitations and future research

Here we demonstrated that the Bayesian approach provides a powerful statistical tool for comparing and evaluating cognitive toolboxes that contain rather few strategies. In principle, the same approach can also be used for more complex scenarios. The only constraint of this methodology lies in the potential difficulties of implementing efficient MCMC sampling for vastly more complex models. As long as these computational challenges are met, the approach is not constrained to comparing toolboxes but can also be extended to compare different toolboxes against alternative cognitive models that do not conform to the notion of a repertoire of strategies.

Acknowledgments

We would like to thank Eric-Jan Wagenmakers for helping us implement the Bugs code.

References

Brown, N. R. (1995). Estimation strategies and the judgment of event frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*, 1539–1553.

Coyle, T. R., Read, L. E., Gaultney, J. F., & Bjorklund, D. F. (1998). Giftedness and variability in strategic processing on a multitrial memory task: Evidence for stability in gifted cognition. *Learning & Individual Differences, 10*, 273–290.

Dougherty, M. R., Thomas, R., Franco-Watkins, A. M. (2008). Postscript: Vague Heuristics Revisited. *Psychological Review,115*, 211-213.

Gigerenzer, G., Todd, P. M., & the ABC Research Group. (1999). *Simple Heuristics That Make Us Smart*. New York: Oxford University Press.

Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (1996) *Markov Chain Monte Carlo in Practice*. Chapman and Hall, London.

Han, C. & Carlin, P. (2001). Markov Chain Monte Carlo methods for computing Bayes factors: a comparative review. *Journal of the American Statistical Association, 96*, 1122-1132.

Kass, R. E. & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*, 773-795.

Lee, M. D. & Wagenmakers, E.J. (2009). *A Course in Bayesian Graphical Modeling for Cognitive Science*. Unpublished lecture notes. http://users.fmg.uva.nl/ewagenmakers/BayesCourse/BayesBook.pdf.

Newell, B. R. (2005). Re-visions of rationality. *TRENDS in Cognitive Sciences, 9*, 11-15.

Payne, J. W., Bettman, J. R., & Johnson E. J. (1993). *The Adaptive Decision Maker*. Cambridge: Cambridge University Press

Patalano, A. L., Smith, E. E., Jonides, J., & Koeppe, R. A. (2001). PET evidence for multiple strategies of categorization. *Cognitive, Affective & Behavioral Neuroscience, 1*, 360–370.

Shiffrin, R. M., Lee, M. D., Kim, W., & Wagenmakers, E.-J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive Science, 32*, 1248-1284.

Siegler, R. S. (1991). Strategy choice and strategy discovery. Learning and Instruction, Volume 1, Issue 1, 1991, 89-102.

Meaning Representation in Natural Language Categorization

Trevor Fountain (t.fountain@sms.ed.ac.uk) and
Mirella Lapata (mlap@inf.ed.ac.uk)
School of Informatics, University of Edinburgh
10 Crichton Street, Edinburgh EH8 9AB, UK

Abstract

A large number of formal models of categorization have been proposed in recent years. Many of these are tested on artificial categories or perceptual stimuli. In this paper we focus on categorization models for *natural language concepts* and specifically address the question of how these may be represented. Many psychological theories of semantic cognition assume that concepts are defined by features which are commonly elicited from humans. Norming studies yield detailed knowledge about meaning representations, however they are small-scale (features are obtained for a few hundred words), and admittedly of limited use for a general model of natural language categorization. As an alternative we investigate whether category meanings may be represented *quantitatively* in terms of simple co-occurrence statistics extracted from large text collections. Experimental comparisons of feature-based categorization models against models based on data-driven representations indicate that the latter represent a viable alternative to the feature norms typically used.

Introduction

Considerable psychological research has shown that people reason about novel objects they encounter by identifying the category to which these objects belong and extrapolating from their past experiences with other members of that category. This task of *categorization*, or grouping objects into meaningful categories, is a classic problem in the field of cognitive science, one with a history of study dating back to Aristotle. This is hardly surprising, as the ability to reason about categories is central to a multitude of other tasks, including perception, learning, and the use of language.

Numerous theories exist as to how humans categorize objects. These theories themselves tend to belong to one of three schools of thought. In the *classical* (or Aristotelian) view categories are defined by a list of "necessary and sufficient" features. For example, the defining features for the concept BACHELOR might be *male*, *single*, and *adult*. Unfortunately, this approach is unable to account for most ordinary usage of categories, as many real-world objects have a somewhat fuzzy definition and don't fit neatly into well-defined categories (Smith and Medin, 1981).

Prototype theory (Rosch, 1973) presents an alternative formulation of this idea, in which categories are defined by an idealized prototypical member possessing the features which are critical to the category. Objects are deemed to be members of the category if they exhibit enough of these features; for example, the characteristic features of FRUIT might include *contains seeds*, *grows above ground*, and *is edible*. Roughly speaking, prototype theory differs from the classical theory in that members of the category are not required to possess *all* of the features specified in the prototype.

Although prototype theory provides a superior and workable alternative to the classical theory it has been challenged by the *exemplar* approach (Medin and Schaffer, 1978). In this view, categories are defined not by a single representation but rather by a list of previously encountered members. Instead of maintaining a single prototype for FRUIT that lists the features typical of fruits, an exemplar model simply stores those instances of fruit to which it has been exposed (e.g., apples, oranges, pears). A new object is grouped into the category if it is sufficiently similar to one or more of the FRUIT instances stored in memory.

In the past much experimental work has tested the predictions of prototype- and exemplar-based theories in laboratory studies involving categorization and category learning. These experiments tend to use perceptual stimuli and artificial categories (e.g., strings of digit sequences such as 100000 or 0111111). Analogously, much modeling work has focused on the questions of how categories and stimuli can be represented (Griffiths et al., 2007a; Sanborn et al., 2006) and how best to formalize similarity. The latter plays an important role in both prototype and exemplar models as correct generalization to new objects depends on identifying previously encountered items correctly.

In this paper we focus on the less studied problem of categorization of *natural language concepts*. In contrast to the numerous studies using perceptual stimuli or artificial categories, there is surprisingly little work on how natural language categories are learned or used by adult speakers. A few notable exceptions are Heit and Barsalou (1996) who attempt to experimentally test an exemplar model within the context of natural language concepts, Storms et al. (2000) who evaluate the differences in performance between exemplar and prototype models on a number of natural categorization tasks, and Voorspoels et al. (2008) who model typicality ratings for natural language concepts. A common assumption underlying this work is that the meaning of the concepts involved in categorization can be represented by a set of features (also referred to as properties or attributes).

Indeed, featural representations have played a central role in psychological theories of semantic cognition and knowledge organization and many studies have been conducted to elicit detailed knowledge of features. In a typical procedure, participants are given a series of object names and for each object they are asked to name all the properties they can think of that are characteristic of the object. Although feature norms are often interpreted as a useful proxy of the structure of semantic representations, a number of difficulties arise

when working with such data (e.g., Sloman and Ripps 1998; Zeigenfusse and Lee 2009). For example, the number and types of attributes generated can vary substantially as a function of the amount of time devoted to each object. There are many degrees of freedom in the way that responses are coded and analyzed. It is not entirely clear how people generate features and whether all of these are important for representing concepts. Finally, multiple subjects are required to create a representation for each word, which limits elicitation studies to a small number of words and consequently the scope of any computational model based on these feature norms.

Even when the stimuli in question are of an abstract or linguistic nature, the features elicited are assumed to be representative of the underlying referents. As an alternative we propose to model the categorization of linguistic stimuli according to their distribution in corpora. Words whose referents exhibit differing features likely occur in correspondingly different contexts; our question is whether these differences in usage can provide a substitute for featural representations.

The idea that words with similar meaning tend to be distributed similarly across contexts is certainly not a novel one. *Semantic space* models, among which Latent Semantic Analysis (LSA, Landauer and Dumais 1997) is perhaps known best, operationalize this idea by capturing word meaning *quantitatively* in terms of simple co-occurrence statistics (between words and paragraphs or documents). More recently, *topic models* (Griffiths et al., 2007b) have arisen as a more structured representation of word meaning. In contrast to more standard semantic space models where word senses are conflated into a single representation, topic models assume that words observed in a corpus manifest some latent structure — word meaning is a probability distribution over a set of topics (corresponding to coarse-grained senses). Each topic is a probability distribution over words whose content is reflected in the words to which it assigns high probability.

In this work we investigate whether semantic representation models based on the statistical analysis of large text collections can provide a viable alternative to feature norms for natural language categorization. Specifically, we compare categorization models that represent concepts by features against LSA, Latent Dirichlet Allocation (LDA, Griffiths et al. 2007b; Blei et al. 2003), a well-known topic model, and a semantic space that takes syntactic information into account (Padó and Lapata, 2007). These semantic representations are used as input to two well-established categorization models, namely Nosofsky's (1988) generalized context model (GCM) and a prototype model derived from the GCM. We evaluate the performance of these models on three adult categorization tasks — category naming, typicality rating, and exemplar generation — which have been previously modeled using exclusively feature norms (Storms et al., 2000). Our results indicate that LSA-based meaning representations outperform more sophisticated alternatives across the board, whilst lagging behind feature norms only by a small margin.

Meaning Representation

In this section we briefly describe the feature norms used in our experiments. These were based on an existing general purpose database (McRae et al., 2005) which we augmented in several ways to suit our categorization tasks. We also describe three corpus-based models of meaning representation, highlight their differences, and motivate their selection.

Feature Norms

As mentioned earlier, many behavioral experiments have been conducted to elicit semantic feature norms across languages. One of the largest samples for English has been collected by McRae et al. (2005). Their norms consist of 541 basic-level concepts (e.g., DOG and CHAIR) with features collected in multiple studies over several years. For each concept several annotators were asked to produce a number of relevant features (e.g., *barks*, *has-four-legs*, and *used-for-sitting*). The production frequency of a feature given a particular concept can be viewed as a form of weighting indicating the feature's importance for that concept. A spatial representation of word meaning can be extracted from the norms by constructing a matrix in which each row represents a word and each column a feature for that word. Cells in the matrix correspond to the frequency with which a feature was produced in the context of a given word. An example of such a space is shown in Table 2 (a) (the numbers correspond to production frequencies, e.g., 12 participants thought *has-legs* is a feature of TABLE).

Unfortunately, McRae et al.'s (2005) norms do not include any explicit relational information. Because we are interested in using the norms in a model of categorization it was necessary for us to augment the concepts with category labels (e.g., 'dog' is an ANIMAL) and typicality ratings (e.g., 'dog' is a typical ANIMAL whereas 'Snoopy' isn't). We collected this information using Amazon Mechanical Turk[1], an online labor marketplace which has been used in a wide variety of elicitation studies and has been shown to be an inexpensive, fast, and (reasonably) reliable source of non-expert annotation for simple tasks (Snow et al., 2008).

We obtained category labels as follows. We presented each participant with twenty unrelated, randomly selected concepts from McRae et al.'s (2005) data set and asked them to label each with the category to which it best belonged. Responses were in the form of free text, i.e., participants were asked to key in a label rather than select one from a list. Each concept was labeled by ten participants; concepts were then grouped according to the resulting categories. Because annotations collected from Mechanical Turk can be noisy we then discarded those categories containing fewer than five unique concepts, leaving 41 categories for 541 exemplars. These category labels are listed in Table 1. To fully integrate them into the norms it was necessary to collect semantic features for them. To do this, we replicated the norming study of McRae et al. (2005), again using Mechanical Turk. Participants were presented with a single concept (drawn from the set of category labels collected in our previous study) and asked to generate ten relevant features. Instructions and examples were taken from McRae et al. (2005). For each category label we collected features from 30 participants, resulting in a large number of features per item. These features were then mapped into the features already present in the norms; as in McRae et al. (2005) this mapping was performed manually.[2]

[1] http://www.mturk.com

[2] The extended database can be downloaded from http://homepages.inf.ed.ac.uk/s0897549/data/.

INSTRUMENT	keyboard	FURNITURE	chair	HOUSING	apartment	DEVICE	stereo
REPTILE	rattlesnake	CONTAINER	bin	VEHICLE	bike	TRANSPORTATION	van
CLOTHING	jeans	STRUCTURE	building	VEGETABLE	carrot	FOOD	bread
HARDWARE	drill	APPLIANCE	stove	BIRD	seagull	GARMENT	coat
HOUSE	cottage	PLANT	vine	TOOLS	hammer	FISH	trout
EQUIPMENT	football	UTENSIL	ladle	THING	doll	ENCLOSURE	fence
TOY	surfboard	KITCHEN	dish	RODENT	rat	INSECT	grasshopper
BUG	beetle	HOME	house	FRUIT	grapefruit	SPORTS	helmet
MAMMAL	horse	OBJECT	door	ACCESSORIES	necklace	COOKWARE	pan
STORAGE	cabinet	BUILDING	apartment	ANIMAL	cat	WEAPON	bazooka

Table 1: Category labels with most typical exemplars produced by participants in category naming and typicality rating study.

This augmented dataset could be used as-is to evaluate a model of categorization on either a category naming or an exemplar generation task (we describe these tasks in detail in the following section). We further wished to use typicality rating as an additional means for evaluation (Voorspoels et al., 2008). We therefore elicited typicality ratings again via Mechanical Turk. Participants were presented with a single category (e.g., FRUIT) along with twenty randomly selected exemplars belonging to the category (e.g., 'cherry', 'apple', and 'tomato') and asked to rate the typicality of each exemplar among members of the category. Typicality ratings for each exemplar-category pair were collected from 20 participants and an overall rating for each exemplar was computed by taking their mean. The highest rated exemplar for each category is shown in Table 1.

We assessed the quality of the data obtained from Mechanical Turk by calculating their *reliability*, namely the likelihood of a similarly-composed group of participants presented with the same task under the same circumstances producing identical results. We split the collected typicality ratings randomly into two halves and computed the correlation between them; this correlation was averaged across three random splits. These correlations were adjusted by applying the Spearman-Brown prediction formula (Storms et al., 2000; Voorspoels et al., 2008). The reliability of the ratings averaged over 41 concepts was 0.64 with a standard deviation of 0.03. The minimum reliability was 0.52 (INSTRUMENT); the maximum was 0.75 (FURNITURE). Reliability on the category naming task was computed similarly, with an average of 0.72, a maximum of 0.91 (INSTRUMENT), and a minimum of 0.13 (STRUCTURE). These reliability figures may seem low compared with Storms et al. (2000) who perform a similar study. However, note that they conduct a smaller scale experiment; they only focus on eight common natural language concepts (whereas we include 41), and 12 exemplars for each concept (our exemplars are 541).

Data-driven Approaches

In addition to feature norms, we obtained semantic representations for categories and exemplars from natural language corpora. We compared three computational models: Latent Semantic Analysis (LSA; Landauer and Dumais 1997), Latent Dirichlet Allocation (LDA; Griffiths et al. 2007b; Blei et al. 2003), and Dependency Vectors (DV; Padó and Lapata 2007). LSA has historically been a popular method of extracting meaning from corpora, and has been successful at explaining a wide range of behavioral data — examples include lexical priming, deep dyslexia, text comprehension, synonym selection, and human similarity judgments (see Landauer and Dumais 1997 and the references therein). LSA provides a simple procedure for constructing spatial representations of word meanings. The same is true for dependency vectors where co-occurrence statistics are computed between words attested in specific syntactic relations (e.g., *object-of*, *subject-of*). The assumption here is that syntactic information provides a linguistically informed context, and therefore a closer reflection of lexical meaning. LDA, in contrast, imposes a probabilistic model onto those distributional statistics, under the assumption that hidden topic variables drive the process that generates words. Both spatial and topic models represent the meanings of words in terms of an n-dimensional series of values, but whereas semantic spaces treat those values as defining a vector with spatial properties, topic models treat them as a probability distribution.

Latent Semantic Analysis To create a meaning representation for words LSA constructs a word-document co-occurrence matrix from a large collection of documents. Each row in the matrix represents a word, each column a document, and each entry the frequency with which the word appeared within that document. Because this matrix tends to be quite large it is often transformed via a singular value decomposition (Berry et al., 1995) into three component matrices: a matrix of word vectors, a matrix of document vectors, and a diagonal matrix containing singular values. Remultiplying these matrices together using only the initial portions of each (corresponding to the use of a lower dimensional spatial representation) produces a tractable approximation to the original matrix. This dimensionality reduction can be thought of as a means of inferring latent structure in distributional data whilst simultaneously making sparse matrices more informative. The resulting lower-dimensional vectors can then be used to represent the meaning of their corresponding words; example representations in LSA space are shown in Table 2 (b) (vector components represent tf-idf scores).

Dependency Vectors Analogously to LSA, the dependency vectors model constructs a co-occurrence matrix in which each row represents a single word; unlike LSA, the columns of the matrix correspond to other words in whose syntactic context the target word appears. These dimensions may be either the context word alone (e.g., *walks*) or the context word paired with the dependency relation in which it occurs (e.g., *subj-of-walks*). Many variants of syntactically aware semantic space models have been proposed in the literature. We adopt the framework of Padó and Lapata (2007) where a semantic space is constructed over dependency paths, namely

sequences of dependency edges extracted from the dependency parse of a sentence. Three parameters specify the semantic space: (a) the *content selection function* determines which paths contribute towards the representation (e.g., paths of length 1), (b) the *path value function* assigns weights to paths (e.g., it can be used to discount longer paths, or give more weight to paths containing subjects and objects as opposed to determiners or modifiers.), and (c) the *basis mapping function* creates the dimensions of the semantic space by mapping paths that end in the same word to the same dimension. A simple dependency space in shown in Table 2 (c) (vector components represent co-occurrence frequencies).

Latent Dirichlet Allocation Unlike LSA and DV, LDA is a probabilistic model of text generation. Each document is modeled as a distribution over K topics, which are themselves characterized as distribution over words. The individual words in a document are generated by repeatedly sampling a topic according to the topic distribution and then sampling a single word from the chosen topic. Under this framework the problem of meaning representation is expressed as one of statistical inference: give some data — words in a corpus, for instance — infer the latent structure from which it was generated. Word meaning in LDA is represented as a probability distribution over a set of latent topics. In other words, the meaning of a word is a vector whose dimensions correspond to topics and values to the probability of the word given these topics; the likelihood of seeing a word summed over all possible topics is always one. Example representations of words in LDA space appear in Table 2 (d) (vector components are topic-word distributions).

Implementation All three models of word meaning were trained on the British National Corpus. For the LSA model we used the implementation provided in the Infomap toolkit[3], with words represented as vectors in a 100-dimensional space; for the DV model we used the implementation[4] of Padó and Lapata (2007) with dependency paths up to length 3 and a length-based path value function that assigns each path a value inversely proportional to its length, thus giving more weight to shorter paths corresponding to more direct relationships. We obtained dependency information from the output of MINIPAR, a broad coverage dependency parser (Lin, 2001). Infrequent dependencies attested less than 500,000 times in the BNC were discarded. The LDA model used the implementation[5] of Phan et al. (2008) with 100 topics. Inference in this model is based on a Gibbs sampler which we ran for 2,000 iterations. Additionally, LDA has two hyperparameters α and β which were set to 0.5 and 0.1, respectively.

(a) Feature Norms

	has_4_legs	*used_for_eating*	*is_a_pet*	...
TABLE	12	9	0	...
DOG	14	0	15	...

(b) LSA

	Document 1	Document 2	Document 3	...
TABLE	0.02	0.98	-0.12	...
DOG	0.73	-0.02	0.01	...

(c) DV

	subj-of-walk	*subj-of-eat*	*obj-of-clean*	...
TABLE	0	3	28	...
DOG	36	48	19	...

(d) LDA

	Topic 1	Topic 2	Topic 3	...
TABLE	0.02	0.73	0.04	...
DOG	0.32	0.01	0.02	...

Table 2: Semantic representations for 'table' and 'dog' using feature norms, Latent Semantic Analysis (LSA), Dependency Vectors (DV), and Latent Dirichlet Allocation (LDA).

Categorization

Models

The semantic representations described above served as the input to two categorization models, representative of the exemplar-based and prototype-based approaches. In the generalized context model (GCM, Nosofsky 1988; Medin and Schaffer 1978) categories are represented by a list of stored exemplars and inclusion of an unknown item in a category is determined by the net similarity between the item and each of the category's exemplars. Specifically, the similarity $\eta_{w,j}$ of a novel item w to the category c is calculated by summing its similarity to all stored items i belonging to c:

$$\eta_{w,c} = \sum_{i \in c} \eta_{w,i} \tag{1}$$

To calculate the inter-item similarity $\eta_{w,i}$ we compute the cosine of the angle between the vectors representing w and i:

$$\eta_{w,i} = cos(\theta) = \frac{v_w \cdot v_i}{||v_w||\ ||v_i||} \tag{2}$$

Following Vanpaemel et al. (2005), we can modify Equation (1) into a prototype model by replacing the list of stored exemplars with a single 'prototypical' exemplar c_j:

$$\eta_{w,c} = \eta_{w,c_j} \tag{3}$$

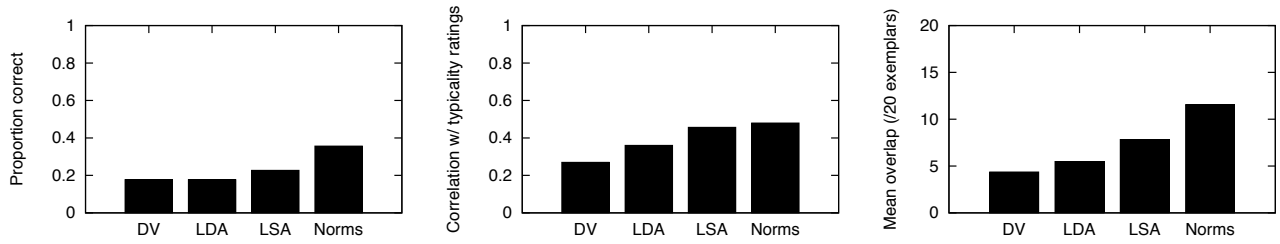
For the category prototype c_j we use the representation of the category label, e.g., the prototype for the category FRUIT is the semantic representation of the word 'fruit'. The similarity between an item and a category thus reduces to the cosine distance between the item and prototype representations.

Tasks

We evaluated the performance of our models on three categorization tasks introduced in Storms et al. (2000): category naming, typicality rating, and exemplar generation.

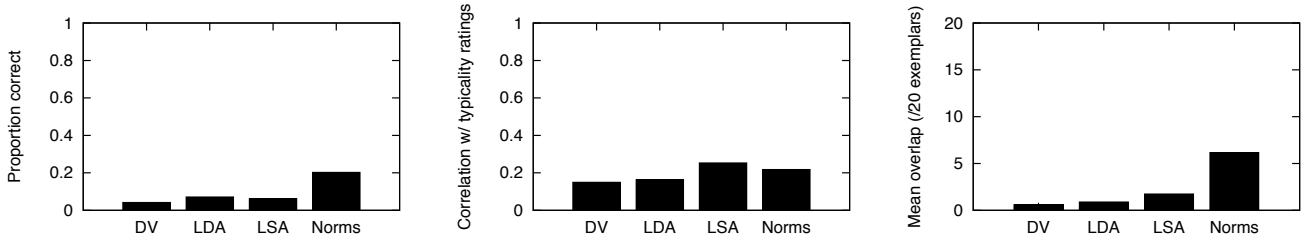
In *category naming* the model is presented with a previously unencountered word and must predict the most appropriate category to which it belongs, e.g., the exemplar 'apple' would be most correctly identified as a member of the category FRUIT, or (with lesser likelihood) FOOD or TREE. In the exemplar model (see (1)), we measure the similarity $\eta_{w,c}$

[3] http://infomap.stanford.edu/
[4] http://www.nlpado.de/~sebastian/dv.html
[5] http://gibbslda.sourceforge.net/



(a) Category Naming (b) Typicality Rating (c) Exemplar Generation

Figure 1: Performance of exemplar model using feature norms and data-driven meaning representations.



(a) Category Naming (b) Typicality Rating (c) Exemplar Generation

Figure 2: Performance of prototype model using feature norms and data-driven meaning representations.

of the novel word against all previously encountered exemplars and select the category with the highest *net* similarity between its exemplars and the word in question; for the prototype model (see (3)) this is the category with the highest similarity between the word and the category's label. Performance on the category naming task was determined in a leave-one-out fashion: a single exemplar was removed from the training examples and then categorized. This was repeated for each exemplar in the training set. The latter consisted of 41 subject-produced category labels each with an average of 30 exemplars.

In a *typicality rating task* the model is presented with both an exemplar and label of the category to which it belongs, and must predict the degree to which it is common amongst members of that category. For the category FOOD, for example, 'pizza' or 'bread' would be considered highly typical exemplars, while 'lutefisk' or 'black pudding' would likely be considered much more atypical. The predicted typicality rating for a word and a category is simply the similarity between the two. In the exemplar model this is the sum similarity between the word and each of the category's exemplars; in the prototype model this is the similarity between the category's label and the word. Performance on the typicality rating task was evaluated by computing the correlation between the models' predicted typicality ratings and the average value predicted by the participants of our rating study. The dataset included typicality ratings for 1,228 exemplar-category pairs.

In an *exemplar generation* task the model is given a category label and must generate exemplars typical of the category, e.g., for FOOD we might generate 'pizza', 'bread', 'chicken', etc. Given a category the model selects from the exemplars known to belong those that are most typical; typicality is again approximated by word-category similarities as determined by the model-specific $\eta_{w,c}$. We evaluate performance on the exemplar generation task by computing the average overlap (across categories) between the exemplars generated by the model and those ranked as most typical of the category by our participants.

Results

Figure 1 summarizes our results with the exemplar model and four meaning representations: McRae et al.'s (2005) feature norms (Norms), Latent Semantic Analysis (LSA), Latent Dirichlet Allocation (LDA), and Dependency Vectors (DV). Results are shown for category naming (Figure 1(a)) typicality rating (Figure 1(b)) and exemplar generation (Figure 1(c)). We examined performance differences between models using a χ^2 test (category naming and exemplar generation) and Fisher's r-to-z transformation (to compare correlation coefficients for the typicality rating task).

On category naming the exemplar model performs significantly better with the feature norms than when using any of the three corpus-derived representations ($p < 0.01$); however, LSA performs significantly better ($p < 0.05$) than DV or LDA. On typicality rating there is no significant difference between the feature norms and LSA. The norms are significantly better ($p < 0.01$) than either DV or LDA, while LSA surpasses both of the other two corpus-derived representations ($p < 0.01$). Additionally, LDA performs significantly better than DV ($p < 0.05$). On the exemplar generation task the feature norms are significantly better ($p < 0.01$) than any of the corpus-based representations; similarly, LSA performs significantly better than LDA or DV ($p < 0.01$), while LDA again outperforms the dependency space ($p < 0.05$).

Our results with the prototype model are shown in Figure 2 and broadly follow a similar pattern. On category naming the feature norms outperform any of the corpus-based representations ($p < 0.01$), LSA is significantly better than LDA which

in turn is better than DV ($p < 0.05$). On typicality rating there is no significant difference between the feature norms and LSA; the difference between LSA and either of the other two representations is significant ($p < 0.01$). On the exemplar generation task feature norms significantly outperform all other representations ($p < 0.01$); LSA is significantly better ($p < 0.01$) than LDA or DV.

Discussion

In this work we have quantitatively evaluated feature norms and alternative corpus-based meaning representations on three natural language categorization tasks. Perhaps unsurprisingly our results indicate that feature norms are more accurate representations when compared to corpus-based models. As feature norms rely on explicit human judgment, they are able to capture the dimensions of meaning that are psychologically salient. Corpus-based models on the other hand learn in an unsupervised fashion and require no human involvement or external sources of knowledge.

Overall we find LSA to be a reasonable approximation of feature norms, superior to both LDA and the syntactically more aware dependency vectors. This result is consistent across models (exemplar vs. prototype) and tasks. Importantly, the LSA model is language-independent and capable of extracting representations for an arbitrary number of words. By contrast, feature norms tend to cover a few hundred words and involve several subjects over months or years. Albeit in most cases better than our models, feature norms themselves yield relatively low performance on all three tasks we attempted using either an exemplar or prototype model (see Figures 1 and 2). We believe the reasons for this are twofold. Firstly, McRae et al.'s 2005 norms were not created with categorization in mind, we may obtain better predictions with some form of feature weighting (see Storms et al. 2000). Secondly, the tasks seem hard even for humans as corroborated by our reliability ratings.

The differences in performance between LSA, LDA, and DV can be explained by differences between the notion of similarity implicit in each. Closely related words in LDA appear in the same *topics*, which are often corpus-specific and difficult to interpret; words belonging to different categories may be deemed similar yet be semantically unrelated. By contrast, the poor performance of the DV model is somewhat disappointing. Our experiments used a large number of dependency relations; it is possible that a more focused semantic space with a few target relations may be more appropriate.

Finally, our simulation studies reveal that an exemplar model is a better predictor of categorization performance than a prototype one. This result is in agreement with previous studies (Voorspoels et al., 2008; Storms et al., 2000) showing that exemplar models perform consistently better across a broad range of natural language concepts from different semantic domains. This finding is also in line with studies involving artificial stimuli (e.g., Nosofsky 1992).

Directions for future work are two-fold. Firstly, we wish to explore alternative meaning representations more suited to the categorization task. A potential candidate is the feature-topic model (Steyvers, 2009; Andrews et al., 2009), in which documents are represented by a mixture of learned topics in addition to predefined topics derived from feature norms. Secondly, we expect that developing specialized models for natural language categorization that are tailored to data-driven meaning representations would improve performance.

References

Andrews, M., Vigliocco, G., and Vinson, D. (2009). Integrating experiential and distributional data to learn semantic representations. *Psychological Review*, 116(3):463–498.

Berry, M., Dumais, S., and O'Brien, G. (1995). Using linear algebra for intelligent information retrieval. *SIAM review*, 37(4):573–595.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Griffiths, T. L., Canini, K. R., Sanborn, A. N., and Navarro, D. J. (2007a). Unifying rational models of categorization via the heriarchical dirichlet process. *Proceedings of the Twenty-Ninth Annual Conference of the Cognitive Science Society*.

Griffiths, T. L., Tenenbaum, J. B., and Steyvers, M. (2007b). Topics in semantic representation. *Psychological Review*, 114:2007.

Heit, E. and Barsalou, L. (1996). The instantiation principle in natural language categories. *Memory*, (4):413–451.

Landauer, T. and Dumais, S. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211–240.

Lin, D. (2001). LaTaT: Language and text analysis tools. In *Proceedings of the 1st Human Language Technology Conference*, pages 222–227, San Francisco, CA.

McRae, K., Cree, G. S., Seidenberg, M. S., and McNorgan, C. (2005). Semantic feature production norms for a large set of living and non-living things. *Behavioral Research Methods Instruments & Computers*, 37(4):547–559.

Medin, D. L. and Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, 85(3):207–238.

Nosofsky, R. M. (1988). Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14:700–708.

Nosofsky, R. M. (1992). Exemplars, prototypes, and similarity rules. In Healy, A. F., Josslyn, S. M., and Shiffrin, R. M., editors, *From Learning Theory to Connectionist Theory: Essays in Honor of William K. Estes*, volume 1, pages 149–167. Hillsdale, NJ: Erlbaum.

Padó, S. and Lapata, M. (2007). Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199.

Phan, X.-H., Nguyen, L.-M., and Horiguchi, S. (2008). Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In *Proceedings of The 17th International World Wide Web Conference (WWW 2008)*, pages 91–100.

Rosch, E. (1973). Natural categories. *Cognitive Psychology*, pages 328–350.

Sanborn, A. N., Griffiths, T. L., and Navarro, D. J. (2006). A more rational model of categorization. *In Proceedings of the Twenty-Eighth Annual Conference of the Cognitive Science Society*.

Sloman, S. A. and Ripps, L. J. (1998). Similarity as an explanatory construct. *Cognition*, (65):87–101.

Smith, E. and Medin, D. (1981). *Categories and Concepts*. Harvard University Press.

Snow, R., OĆonnor, B., Jurafsky, D., and Ng, A. (2008). Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. *Proceedings of EMNLP 2008*.

Steyvers, M. (2009). Combining feature norms and text data with topic models. *Acta Psychologica*. (in press).

Storms, G., Boeck, P. D., and Ruts, W. (2000). Prototype and exemplar-based information in natural language categories. *Journal of Memory and Language*, 42:51–73.

Vanpaemel, W., Storms, G., and Ons, B. (2005). A varying abstraction model for categorization. In *Proceedings of the 27th Annual Conference of the Cognitive Science Society*.

Voorspoels, W., Vanpaemel, W., and Storms, G. (2008). Exemplars and prototypes in natural language concepts: A typicality-based evaluation. *Psychonomic Bulletin & Review*, 15(3):630–637.

Zeigenfusse, M. D. and Lee, M. D. (2009). Finding the features that represent stimuli. *Acta Psychological*. (in press).

Temporal Chunk Signal Reflecting Five Hierarchical Levels in Writing Sentences

Erlijn van Genuchten (e.genuchten@iwm-kmrc.de)
Knowledge Media Research Center, Konrad-Adenauer-Str. 40
72072 Tübingen, Germany

Peter C-H. Cheng (p.c.h.cheng@sussex.ac.uk)
School of Informatics, University of Sussex
Brighton, BN1 9QJ, UK

Abstract

Previous research on the temporal chunk signal has focused on the use of pauses in behaviour to probe chunk structures in working memory. On the basis of some of these studies, a hierarchical process model has been proposed, which consist of four hierarchical levels describing different kind of pauses. In this model, the lowest level consists of pauses between strokes within letters. On higher levels, there are pauses between letters, words, and phrases. Each level is associated with a larger amount of processing when retrieving these chunks from memory. The main aim of the present study is to test whether the temporal chunk signal can distinguish a fifth level, the sentence level. A secondary goal is to replicate the findings which were used to construct the hierarchical process model in a manner that overcomes some of the limitations of the earlier experiments.

Keywords: Temporal chunk signal, graphical protocol analysis, writing, working memory, sentences.

Introduction

Chunks have a fundamental role in information processing in the human cognitive architecture. Chunks are individual pieces of information grouped into larger units that increase our information retention (Caroll, 2004). It is widely accepted in Cognitive Science that the hierarchical storage and processing of chunks in working memory provides a fruitful basis for explaining a substantial range of the behavioural phenomena, such as recall from long-term and working memory and expert performance. This acceptance comes in part from various established methods that have been developed to infer the particular structure of chunks possessed by individuals from their behaviours. Such methods have made an important contribution to developing accounts of cognition in complex tasks. One method is computational modelling. Another method, on which this research focuses, provides information about the structure of chunks in memory by measuring pause lengths that occur in verbal or motor actions, the inter response latency. Chase and Simon's (1973) work on chess expertise and Reitman's (1976) work on Go experts are classic examples of the uses of pause lengths during the recall of a board containing chess pieces or Go discs. There are many other studies that have also exploited pause lengths in actions in order to define chunks. In one of the earliest studies using this approach, McLean and Greg (1967) studied the chunking of arbitrary letter sequences. Later on, Buschke (1976) examined the gradual acquisition of chunks comprising clusters of every day words originally presented in unstructured lists. Egan and Schwartz (1979) showed how electronics experts chunked components of electrical circuits in terms of their functioning. In all these studies, the duration of a pause preceding an action that generated an element of the domain (in these studies letters, words, and components respectively) is taken as an indication of whether the element is within a putative chunk or at the boundary between chunks. The term *temporal chunk signal*, TCS, is used in order to refer to the basic phenomenon that underpins the use of pause lengths to probe chunk structures. Typically, the TCS is often used in a binary fashion, which includes setting some threshold (e.g., 500 ms) as a criterion upon which to classify successively produced elements as intra-chunk if the pause length is less than the threshold, or as inter-chunk if the pause length is greater than the threshold.

The present experiment is a continuation of our studies on the nature and application of the TCS that is manifest in the process of writing and drawing, or more general graphical production. We call our general approach to using the TCS to study chunk related behaviour in writing and drawing tasks, *graphical protocol analysis*, GPA. A standard graphics tablet is used to record pen strokes. Pause lengths are computed by finding the difference in time between the lift of the pen from the tablet at the end of one stroke and the time at which the pen touches the tablet at the beginning of the stoke of interest: $pause_{item} = time_{pen\text{-}down\text{-}current\text{-}item} - time_{pen\text{-}up\text{-}previous\text{-}item}$. In our previous experiments, tasks with known hierarchical structures have been used, such as 'to be or not to be', so that each pause could be coded as intra-chunk or inter-chunk. For sentences and language-like stimuli identified pauses have included: intra-chunks pauses between strokes within a letter (e.g., second stroke of a 't', level 0 or L0); inter-chunk pauses between letters within a word (e.g., between 't' and 'o', L1); and inter-chunk pauses between words within a phrase (e.g., between 'to' and 'be', L2).

Our previous experiments have shown that the TCS is a richer source of information about chunk structure than just a binary signal. We consider that TCS within GPA has potential to be used as general technique for the study of various cognitive phenomena. In the domain of copying mathematical formulae, the TCS was able to distinguish

participants who had four different levels of expertise in mathematics (Cheng & Rojas-Anaya, 2007). The TCS has also been used to distinguish between children with and without dyslexia (van Genuchten et al., submitted). Cheng, McFadzean, and Copeland (2001) have shown that the TCS reflects three distinct levels of processing when drawing geometric figures. That experiment showed the TCS to be present when drawings are made with pen on paper or with a mouse on a computer screen. Obaidellah and Cheng (2009) used the TCS to reveal the role of perceptual chunks and spatial schemes in different modes of drawing complex abstract diagrams. In Cheng & Rojas-Anaya (2005) participants wrote number sequences that had been memorized with different chunk structures. The TCS showed the existence of three levels corresponding to: pauses between strokes within a digit (e.g., second stroke in '⺡', L0, ≈90 ms); digit level chunks (e.g., between '1' and '2', L1, ≈280 ms); and digit group level chunks (e.g., between '111' and '222', L2, ≈440 ms). In Cheng & Rojas-Anaya (2006) the TCS again showed the existence of the same three levels of pauses when writing familiar and jumbled sentences (≈90, ≈270, ≈400 ms respectively). Finally, moving beyond three hierarchical levels, Cheng & Rojas-Anaya (2008) devised an artificial sentences copy task with four hierarchical levels (e.g., 'ITH* ITH* ITH*, ITH* ITH* ITH*') and found the same pattern of stroke, letter, word and phrase level pauses (≈90, ≈250, ≈440 and ≈600 ms respectively). On the basis of these studies, a hierarchical process model has been proposed (see Figure 1) to explain these patterns in terms of the depth first serial processing. In this model, the hierarchical structure of chunks corresponds to the amount of processing associated with different branch lengths of the hierarchy, with longer branches indicating longer pause lengths.

The main aim of the present experiment is to test whether the TCS can distinguish more than four hierarchical levels, by adding a fifth level of pauses between sentences, and thereby extend the previous findings. Is it simply the case that this fifth level in the chunk hierarchy will merely result in an additional amount of processing and a corresponding increment of pause duration? If so, will the increase in magnitude of the pause length be linear as is the case between the other levels? A secondary goal of the experiment is to replicate the previous findings in a manner that overcomes some of the limitations of the earlier experiments. In particular, the found significant effects existed at the level of individuals using pairwise comparisons of the pauses between levels, but typically involved relatively small numbers of participants. Hence, a subsidiary aim of the present experiment is to test whether

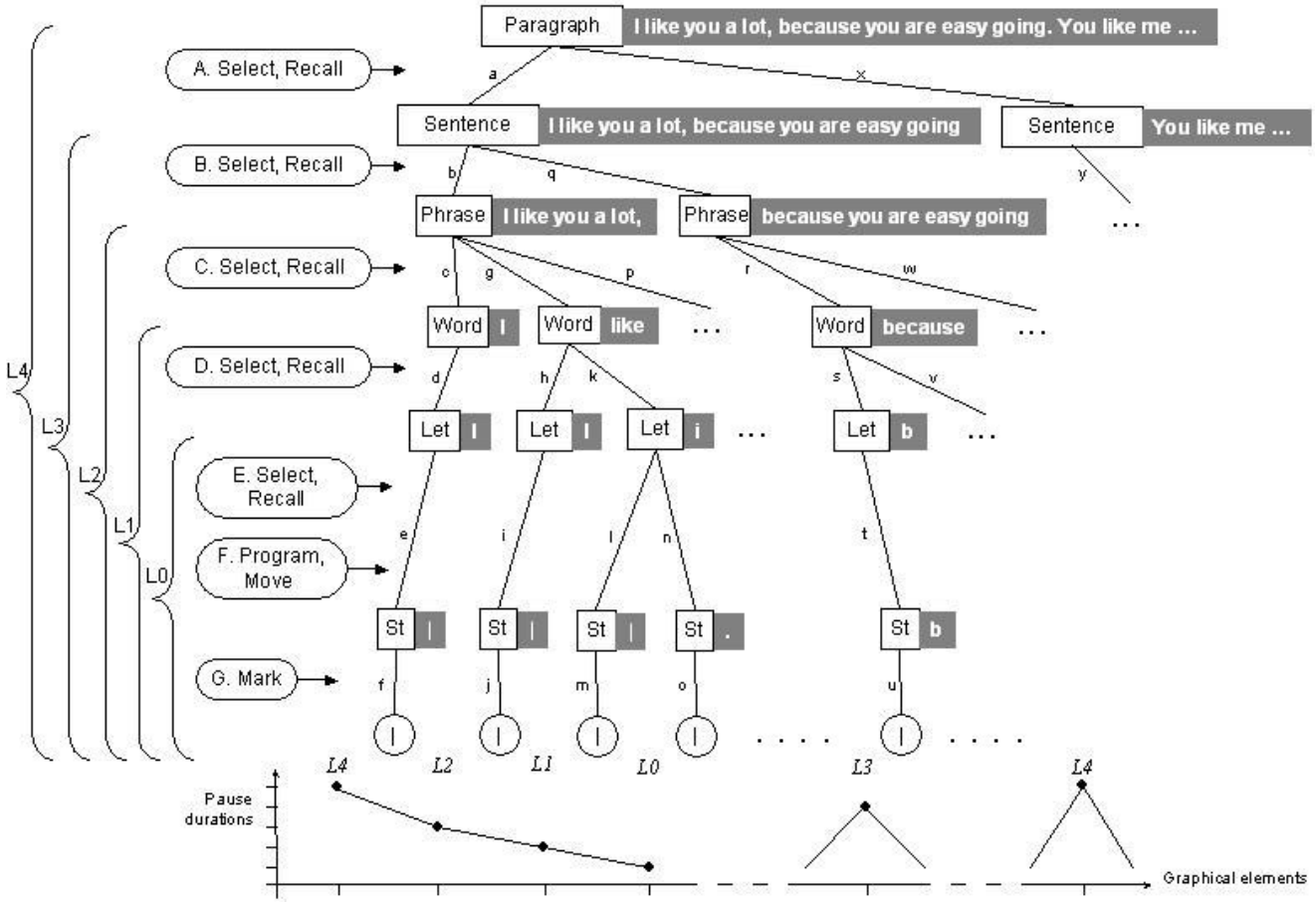


Figure 1: Hierarchical process model including the sentence level, reflecting relations between chunk structure, processing steps, and pause levels (adapted from Cheng & Rojas-Anaya, 2008).

the differences between the hierarchically levels is a robust effect by using multilevel analysis to simultaneously compare all the levels from the data of a large number of participants.

Method

Participants

Participants in this study were 32 adults, 19 female and 13 male, between 18 and 33 years old, working or studying at a large university in the UK. The participants (M = 22.99 years, SD = 3.98 years) were all native English speakers.

Measures and Materials

In order to answer the research question, pause lengths between sentences, phrases, words, letters, and strokes in paragraphs were compared. These five measures were calculated for each stimulus. The eight English sentence stimuli were specially written to obtain these five hierarchical levels. Each stimulus comprised of three or four sentences (L4), which were made up of two or three phrases (L3), which in turn were comprised of between 4 and 8 words (L2), which contained letters (L1) that may have required more than one stroke (L0) to write. This hierarchical structure was emphasized by including punctuation marks (periods between sentence and commas between phrases). Example stimuli are:

'You just signed up for a trip, from your favourite society, because you like visiting different places. You paid with some money, which you got from your mum, because you did shopping for her. You have never been to Holland, so you would like to visit Amsterdam, and have a great time.' (3 sentences, 3 phrases)

'We like swimming, in the pool next door. You like to cycle, to towns far away. They like to play football, on the top of the hill. As they play all day, they should eat enough.' (4 sentences, 2 phrases)

The median per level was calculated in order to reduce the influence of outliers, which could only occur in the direction of longer pause lengths, and which would consequently severely distort the mean, rendering it unsuitable as a measure in this study (Stavig & Gibbons, 1977).

All sentences were written on a piece of paper attached to a graphics tablet containing horizontal rows of rectangles. One letter had to be written in each rectangle (width: 6 mm, height: 8 mm), so that participants were encouraged to lift their pen from the paper and to put it down again for the next letter, and allowing the distance between each letter to be approximately equal (see Figure 2). The equal distance between rectangles rules out the possibility that differences in hand movements account for the different pause lengths. Therefore, pauses between the last letter of a line and the first of the next were ignored, because of increased hand movement.

Design and Procedure

The administering of the test had a duration of 45 to 75 minutes per participant. A quiet room was used to minimize disturbing background noises. The session began with an acclimatization period which allowed the participants to become familiar with writing on a tablet by having them write their names on the tablet. The actual experiment did not start until the participant was considered to have followed all instructions. Participants were asked not to write any punctuation in order to make sure that increased pause lengths were not due to writing an extra symbol. The task itself consisted of remembering and writing down the eight visually presented target stimuli.

All stimuli were presented in turn in random order. After presenting a stimulus, participants were allowed to apply any strategy and take as long as needed to rehearse the stimulus. When participants finished rehearsing, the experimenter tested recall accuracy by asking participants to recite the stimulus sentences without errors twice. Once this was accomplished, participants were allowed to start writing. A hash (#) had to be written at the beginning of each sentence to ensure that the writing process was well underway before the first letter was generated (Cheng & Rojas-Anaya, 2006).

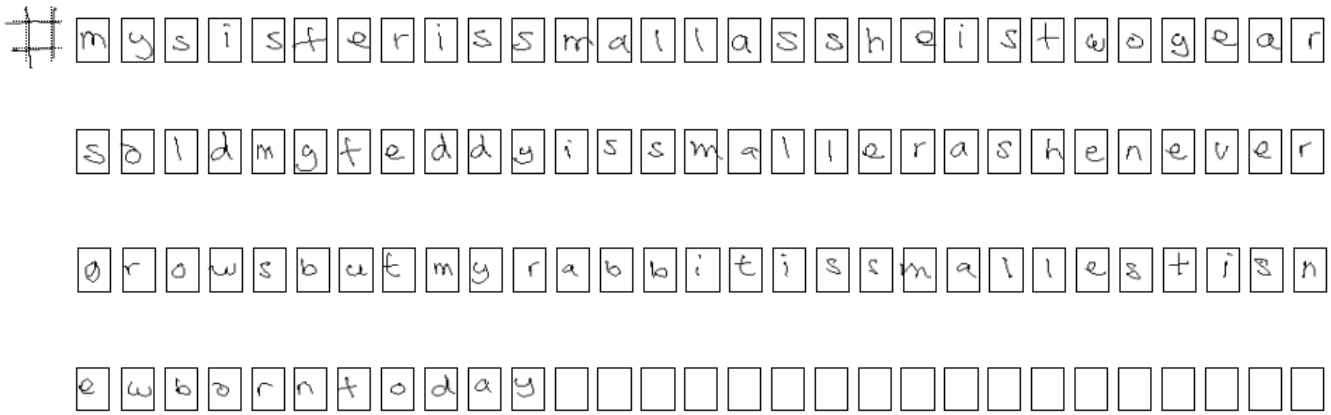


Figure 2: Example part of a written stimulus in equally spaced rectangles.

Table 1: Parameter estimates for multilevel models.

	Intercept only model		Model with predictors [a]		Robust standard errors [b]	
	Par.	SE	Par.	SE	Par.	SE
Fixed effects						
Intercept	488	29.8				
Predictors:						
L0 (pauses between strokes)			90*	37.9	90*	3.1
L1 (pauses between letters)			273*	37.9	273*	13.6
L2 (pauses between words)			374*	37.9	374*	19.8
L3 (pauses between phrases)			567*	37.9	567*	41.3
L4 (pauses between sentences)			1134*	37.9	1134*	96.0
Random effects						
Predictor level variance	306818	7105.5	175177	7012.7	175177	7012.7
Participant level variance	20726	12282.6	24016	7101.0	24016	7101.0
Deviance	19846		19146		19146	

Notes: [a] The constant has been left out of the model with predictors, because a complete set of dummy variables was used.
[b] Robust standard errors were used, because the assumption of linearity of residuals was not met.
* $p < .001$.

Data and Analysis

A specially written program, TRACE, was used to record the writing actions and to extract the pen positions, times of points and pause lengths (Cheng & Rojas-Anaya, 2003). All files generated by TRACE were analysed using a specialized computer programme (Pause Level Extraction Tool, PLET, van Genuchten, 2009). Automatic detection of letters and automatic calculation of pause lengths had to be applied, because of the large amount of data (at least 1200 measures per participant). This involved identifying the horizontal position of strokes making up a letter within a rectangle and the horizontal separation between these strokes between rectangles. The written input was specified for each stimulus, so that errors (e.g., wrongly spelled or omitted words) were taken into account. Only in those cases where no automatic detection of letters was possible (e.g., when strokes of subsequent letters were written relatively close together), manual calculation of the pause lengths was applied.

To test whether there were differences between pause lengths of different pause types, a multilevel analysis was performed. On the lowest level, data of pause lengths of each stimulus for each pause type were used as predictors and were measured within participants. This data was gathered on the highest level, the participant. It is expected that the correlation between measures is higher within a participant than between participants. By performing a multilevel analysis, the dependency between data measures within individuals is controlled. Dummy variables were created for each of the five pause types.

Results

The comparison between the model-fit of the one-level and two-level intercept only models, indicates that there is significant variance at the participant level ($\chi^2(1) = 41.78$, $p < .01$). This means a two-level multilevel analysis is appropriate. Parameter estimates for the intercept only model and model with predictors are presented in Table 1.

A difference between the different pause types was expected, which was confirmed by the model with predictors. However, plots of standardized residuals against normal scores indicated that the assumption of linearity of residuals was not met. Therefore, robust standard errors were calculated using the Sandwich method (Hox, 2002). The resulting model indicates that a distinction can be made between pause lengths on the basis of pause type. Specifically, the regression coefficients show that pause lengths between sentences are longest, and that pause lengths become successively shorter when considering pauses between phrases, words, letters, and strokes (sentences: $B = 1134$, $SE = 96.0$, $p < .001$; phrases: $B = 567$, $SE = 41.3$, $p < .001$; words: $B = 374$, $SE = 19.8$, $p < .001$; letters: $B = 273$, $SE = 13.6$, $p < .001$; strokes: $B = 90$, $SE = 3.1$, $p < .001$). This means that pause lengths can be very well predicted when it is known which type of pause is concerned.

Discussion

One aim of the present experiment was to replicate the findings of previous experiments concerning the temporal chunk signal, TCS, using a more rigorous methodology. These earlier studies showed that the TCS reflects a hierarchical chunk structure as increasing durations of pause lengths between written elements. In this research, differences between every pause level within this structure were also found to be significantly different in a single multilevel statistical test. Although the outcomes of the previous experiments had to be carefully qualified, it does

Table 2: Pauses (ms) for various stimulus levels over different stimulus types (rounded to 10 ms).

Experiment	Stimuli	L0	L1	L2	L3	L4
Cheng & Rojas-Anaya (2005)	Number sequences	90	280	440		
Cheng & Rojas-Anaya (2006)	Familiar and jumbled phrases	90	270	400		
Cheng & Rojas-Anaya (2008)	Artificial sentence	90	250	440	600	
Present	Natural language paragraphs	90	270	370	570	1130

appear that the effects found are genuine, because of the consistency with the present experiment.

Regarding the primary aim of this experiment of adding a fifth level to the structure, the results show that when a rehearsed stimulus that possesses five hierarchical levels is written, the TCS, which is based on the pauses between written elements, reflects the ordering of the levels. The stroke level pause lengths are the shortest and the duration increases for each successive increment of level, through letter, word, phrase and sentence level. The increase of the pause with the addition of the fifth sentence level is consistent with the proposal that in graphical production of well rehearsed stimuli each successive chunk level requires specific processing to deal with the particular information associated with that level (Cheng & Rojas-Anaya, 2008).

The direct comparison of the approximate absolute values of the pauses associated with each level for this and the previous experiments reveal some interesting patterns (see Table 2). The three prior experiments noted in this table involved graphical production using the same experimental task methodology: specifically, the writing of sequences from memory after rehearsal with one character in one rectangle. The experiments differ in the important respect that each used a different type of stimulus, as indicated in Table 2. The similarity of the absolute values of the pauses over each level across the different experiments is noteworthy, because it suggests that the same underlying processes are responsible for the pattern of pauses irrespective of the nature of the stimulus. The differences between the pauses on successive levels range between 100 and 200 ms, with mean values of L1-L0=178, L2-L1=145 and L3-L2=180 ms. Taking Newell's (1990) estimate of the time scale for elementary deliberated operations as circa 100 ms, this suggest that there is at least one additional operation occurring when preparing to graphically produce an element that is one level higher in the hierarchy. One such operation will be a process to select the next chunk at a particular level. At the beginning of a new phrase this will involve selection of a phrase, a word, a letter and a stroke. The increase in time suggests that this selection occurs serially and is therefore consistent with the predictions of the hierarchical process model (Cheng & Rojas-Anaya, 2008).

The increase in pause length up to the sentence level from the phrase level is more than three times greater than the increase between any of the other levels. As this is new data from just one experiment, some caution must be taken with its interpretation. It seems to suggest that additional operations that occur at this level do not occur at the levels below. The additional time may be an indication that working memory is fully loaded when complex stimuli comprising multiple sentences with several sub phrases are being processed. Furthermore, the additional time may indicate that retrieval from long-term memory is required as the complete stimuli cannot all be held in working memory despite the rehearsal. As there are approximately ten times as many letters in each of the present stimuli as there were in the stimuli of the previous experiments and as the number of chunks is larger than Miller's magical number 7 ± 2 (Miller, 1956), this is a likely interpretation. In future research, verbal working memory measures, such as the digit span task (Wechsler, 1985) and the listening span task (Daneman & Carpenter, 1980), could be used to gain insight into how pause lengths are related to working memory capacity and the possible involvement of long-term memory. However, it should be noted that as there were 19 pauses on the sentence level at the most, the actual value might be less robust than for the other levels, because outliers have a larger influence with a small number of measurement points. For comparison, there were 33 phrase level, 238 word level, and about 1148 letter level pauses (the number of strokes depended on whether cursive of block letters were used).

Another possible interpretation of this large increase in pause length is that, in addition to retrieving a sentence, inhibiting processes take place to suppress the inclination of writing punctuation. A possibility to overcome this problem is to require participants to write punctuation in a separate rectangle. However, in this case, it is unclear whether the pause between the last letter of for example, the sentence and the period (full stop) or the pause between the period and the first letter of the next sentence, should be taken as an indication of pause length. An alternative for future experiments would be to require participants to write punctuation marks in the same rectangle right after the last letter of the sentence or phrase.

In summary, other processes than selection and retrieval processes might also underpin this pattern of pause lengths. Therefore, empirical and modelling studies are conducted to unravel which processes contribute to the increase in pause length accompanied with each level.

Irrespective of the precise explanation for the increase in duration between each hierarchical level, the results of this experiment reconfirms the claim that there is a temporal signal. This signal may be associated with chunking processes and is a source of high resolution information concerning participants' task performance. With appropriately designed tasks, the TCS could provide valuable evidence to probe the relations among the sub-processes that underpin cognitive phenomena.

A logical next step is to investigate whether a sixth level, the paragraph level, can be added to the hierarchical process model. However, as the demands on working and long-term memory will increase even more, such an experiment has to be designed carefully in order to be feasible.

Acknowledgements

We would like to thank Keith Smith of the University of Sussex for his assistance with the recruitment of participants for this study. We would also like to thank Cora Maas of Utrecht University for her assistance in performing the multilevel analysis.

References

Buschke, H. (1976). Learning is organized by chunking. *Journal of Verbal Learning and Verbal Behavior, 15,* 313-324.

Carroll, D. W. (2004). *Psychology of language.* Belmont, KY: Thomson Wadsworth.

Chase, W., & Simon, H. A. (1973). The mind eye's in chess. In W. Chase (Ed.), *Visual information processing*. New York, N.Y.: Academic press.

Cheng, P. C-H., McFadzean, J., & Copeland, L. (2001). Drawing out the temporal structure of induced perceptual chunks. In J. D. Moore & K. Stenning (Eds.), *Proceedings of the Twenty Third Annual Conference of the Cognitive Science Society* (pp. 200-205). Mahwah, New Jersey: Lawrence Erlbaum.

Cheng, P. C-H., & Rojas-Anaya, H. (2005). Writing out a temporal signal of chunks: Patterns of pauses reflect the induced structure of written number sequences. In B. G. Bara, L. Barsalou & M. Bucciarelli (Eds.), *Proceedings of the 27th Annual Conference of the Cognitive Science Society* (pp. 424-429). Mahwah, NJ: Lawrence Erlbaum.

Cheng, P. C-H., & Rojas-Anaya, H. (2006). A temporal signal reveals chunk structure in the writing of word phrases. In R. Sun, & Miyake (Eds.), *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (pp. 160-165). Mahwah, NJ: Lawrence Erlbaum Associates.

Cheng, P. C-H., & Rojas-Anaya, H. (2007). Measuring Mathematical Formula Writing Competence: An Application of Graphical Protocol Analysis. In D. S. McNamara & J. G. Trafton (Eds.), *Proceedings of the Twenty Ninth Annual Conference of the Cognitive Science Society* (pp. 869-874). Austin, TX: Cognitive Science Society.

Cheng, P. C-H., & Rojas-Anaya, H. (2008). A Graphical Chunk Production Model: Evaluation Using Graphical Protocol Analysis with Artificial Sentences. In B. C. Love, K. McRae & V. M. Sloutsky (Eds.), *Proceedings of the Thirtieth Annual Conference of the Cognitive Science Society* (pp. 1972-1977). Austin, TX: Cognitive Science Society.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Learning and Verbal Behavior, 19,* 450-466.

Egan, D. E., & Schwartz, B. J. (1979). Chunking in recall of symbolic drawings. *Memory and Cognition, 7,* 149-158.

Hox, J. (2002). *Multilevel analysis: Techniques and applications.* London: Lawrence Erlbaum Associates.

McLean, R. S., & Gregg, L. W. (1967). Effects of induced chunking on temporal aspects of serial recitation. *Journal of Experimental Psychology, 74,* 455-459.

Miller, G. A. (1956). The magical number seven plus or minus two: Some limits on our capacity for information processing. *Psychological Review, 63,* 81-97.

Newell, A. (1990). *Unified theories of cognition*. Cambridge, MA: Harvard University Press.

Obaidellah, U. H., & Cheng, P. C-H. (2009). Graphical production of complex abstract diagrams: drawing out chunks and schemas. In N. Taatgen & H. v. Rijn (Eds.), *Proceedings of the Thirty-first Annual Conference of the Cognitive Science Society* (pp. 2843-2848). Austin, TX: Cognitive Science Society.

Reitman, J. S. (1976). Skilled perception in Go: Deducing memory structures from inter-response times. *Cognitive Psychology, 8,* 336-356.

Stavig, G. B., & Gibbons, J. D. (1977). Comparing the mean and the median as measures of centrality. *International Statistical Review / Revue Internationale de Statistique*, *45,* 63-70.

van Genuchten, E. (2009). Pause Length Extraction Tool (PLET), Computer Software, University of Sussex. Brighton, UK.

van Genuchten, E., Cheng, P. C-H., Leseman, P. P. M., & Moerland, J. (submitted). Detection of a Working Memory Deficit in Dyslexia: Children Writing from Memory. *Dyslexia.*

Wechsler, D. (1958). *The measurement and appraisal of adult intelligence*. Baltimore: Williams & Wilkins.

Emotion in Good Luck and Bad Luck: Predictions from Simplicity Theory

Jean-Louis Dessalles (dessalles@telecom-paristech.fr)
Telecom PARISTECH, 46 rue Barrault,
F-75013 Paris, France
&
ILCAA, Tokyo University of Foreign Studies,
Fuchu-shi, 183-8534, Tokyo, Japan

Abstract

The feeling of good or bad luck occurs whenever there is an emotion contrast between an event and an easily accessible counterfactual alternative. This study suggests that *cognitive simplicity* plays a key role in the human ability to experience good and bad luck after the occurrence of an event.

Keywords: Kolmogorov complexity; simplicity; emotion; luck; probability; unexpectedness.

Good Luck and Bad Luck

Situations spontaneously associated with good luck or bad luck are an important source of emotion. They are frequent in daily life: missing (or catching) the train by five seconds, forgetting one's cell phone the very day one is late for an important appointment, finding a banknote on the ground, etc. They are heavily used in popular fiction, precisely to arouse emotion: the gun gets jammed just at the right (or bad) time, the heroine defuses the bomb just before it explodes, etc. Regarding oneself or someone else as lucky or unlucky on specific occasions may induce gratitude or guilt, and for those who downplay the role of chance, intense feelings of good or bad luck may strengthen supernatural beliefs (Teigen & Jensen, *in press*). Reasoning about good luck and bad luck may also significantly influence rational judgment (Roese, 1997; Wohl & Enzle, 2003).

The feeling of having good or bad luck is a clear-cut phenomenon. Different individuals have consistent views of which situations can be regarded as bad or good luck (what the present study will confirm). This ability therefore gives rise to a well-posed problem, worth investigating. Previous studies have identified various parameters that control the feeling of luck. These include physical or temporal closeness (Kahneman & Varey, 1990; Teigen, 1996; Roese, 1997; Pritchard & Smith, 2004), deviation from norms and expectations (Kahneman & Miller, 1986), mutability of causes (Kahneman & Miller, 1986; Byrne 2002, 2007) and controllability (Roese, 1997).

Many authors have acknowledged the prime importance of counterfactuals in any situation that generates a strong feeling of good or bad luck. Individuals systematically go through thoughts such as "If only…" or "I almost…" when regarding situations as (un)lucky. The theoretical treatment of counterfactuals in general, and in emotional situations in particular, remains however complex, as a multitude of determining factors seem to be involved.

The purpose of the present study is to propose a new perspective on the phenomenon, imported from two other scientific domains. One is the study of narrative relevance (Dessalles 2008a). Spontaneous conversations are replete with stories about (un)lucky episodes, and the laws of interestingness seem to apply to them. The other import is the mathematical notion of complexity, which is involved in several important cognitive phenomena (Chater 1999; Chater & Vitányi, 2001).

After mentioning existing attempts to capture the good/bad luck phenomenon formally, I will briefly present the Simplicity Theory and its first predictions concerning our problem. I will then present a study that seems to corroborate those predictions. Then, I will consider situations in which individuals adopt causal thinking. The results and the scope of the theory will be discussed in a last section.

Formal accounts of luck

Various determining factors have been identified that control the intensity of luck. One of them is the low probability of the (un)lucky event s. According to Rescher (1995:211), the intensity of luck is given by $L = E\,(1–p)$, where E measures the difference that the occurrence of s makes for the interests at stake, and p is its probability. This formula has two major drawbacks. First, contrary to intuition, it does not distinguish moderately unlikely outcomes from highly unlikely ones, as both would provide emotion roughly equal to Δu. Second, as pointed out by Teigen (2005), it fails to capture the crucial presence of a counterfactual. As shown by Teigen in various studies, the amplitude of (un)luck is controlled by the 'distance' to an alternative outcome that would have provided an emotional contrast. Teigen (2005) represents these effects through the formula: $L = \Delta u / D$, where Δu is the difference in 'utility' between the counterfactual s_2 and the actual situation s_1, whereas D represents the 'distance' between s_1 and s_2.

This formula makes predictions that are much closer to observation, and thus represents a significant progress in comparison with Rescher's initial proposal. It has, however, its limitations. First, the influence of low probability, as identified by Rescher, is lost. The problem is illustrated in figure 1, where the feeling of unluck after missing the winning sector (in color) in a wheel of fortune game is stronger in (b) than in (a). Second, the notion of 'utility', imported from economics, does not account for situations of pure surprise ('I almost got six on all dice'). Third, the

notion of distance is not properly defined. Sitting next to a lottery winner doesn't make you feel unlucky; you might however feel unlucky to have played her winning numbers, but a week to soon. Lastly, Teigen's formula fails to capture one property of counterfactual s_2 that contributes to (un)luck, namely its simplicity. In figure 1(c), the winning sectors (in color) of the wheel of fortune occupy the same area as in (a) and the distance to the landing site is the same in both cases. Judgment of bad luck is, however, stronger in (a) than in (c). This phenomenon, due to the greater complexity of the counterfactual in (c), is not predicted by Teigen's formula.



(a) (b) (c)

Figure 1: Three examples of near miss

We will propose an alternative account, based on Simplicity Theory. It can be formulated in an informal way:

(Un)lucky events are situations that occurred despite of simple, easily accessible alternatives.

Simplicity Theory

Simplicity Theory (ST) (formerly called 'Complexity Drop Theory') has been developed to predict how people select events worth to tell. It has applications in the study of spontaneous conversations, of narratives, of news, and in the definition of subjective probability (Dessalles, 2006; 2008a). ST's main principle can be stated:

Interesting situations are those which are 'too' simple.

ST uses the notion of cognitive complexity, which is a slightly modified version of the mathematical notion know as Kolmogorov complexity.

The complexity C(s) of a situation s *is the size of the ideal minimal description of* s *that is available to the observer.*

(the last restriction is crucial for the notion to be useful in cognitive science). The concept is much less trivial than it seems at first sight, and has given rise to a growing literature since its definition in the years 1960.

ST uses two notions of complexity. The second notion is *generation* complexity.

$C_w(s)$ is the minimal size of the parameters to be set for the 'world' w *to generate situation* s.

To compute $C_w(s)$ of a lottery draw, for instance, one adds up the descriptions of all drawn numbers, as the 'world' (in this case, the lottery machine) had to 'choose' them independently. Note that the notion refers, not to any objective world, but to the observer's perception of the world. ST's central notion is unexpectedness, noted $U(s)$.

$$U(s) = C_w(s) - C(s) \quad (1)$$

A situation is unexpected if it is 'too' simple, *i.e.* simpler to describe than to generate. In the lottery example, a 'remarkable' lottery draw such as 22-23-24-25-26-27 is unexpected, since C is much smaller than C_w. It only requires to instantiate 22 and to mention that it is a continuous series. C thus spares five instantiations by comparison with C_w. Hence a strong feeling of unexpectedness if such a draw actually occurs (Dessalles 2006). This definition of unexpectedness accounts for various cognitive abilities, such as the perception of coincidences (Dessalles, 2008b) and of interestingness (Dessalles, 2008a; Dimulescu & Dessalles, 2009) (see details on www.simplicitytheory.org). It is consistent with the observation that 'contrast' (what we call unexpectedness) is more relevant than (standard) probability to explain surprise (Teigen & Keren, 2003).

Complexity is usually linked to probability p_0 thanks to the following formula $p_0=2^{-C_{w_0}}$, where w_0 is a blank world (Solomonoff, 1978). This formula is, however, unsatisfactory, as it assigns a virtually zero probability to most situations of daily life, as they depend on a huge quantity of parameters. If we replace the blank world w_0 by the observer's model w of the actual 'world', we get $p_w = 2^{-C_w}$, which corresponds to the usual definition of 'objective' probability. In a lottery, for instance, p_w is the same for all draws. ST (Dessalles 2006) defines *subjective* probability p by subtracting cognitive complexity C from C_w. We get:

$$p = 2^{-U} \quad (2)$$

Hence the statement about unexpected events being 'too' simple. In ST's framework, the concept of probability is a derived notion and should be replaced by the notion of unexpectedness to account for many aspects of cognition.

To account for good luck and bad luck, we must say how emotion is related to simplicity (Dessalles, 2008a). Let's call $E(s)$ the (always positive) intensity of the emotional experience caused by situation s.

$$E(s) = E_h(s) + U(s) \quad (3)$$

$E_h(s)$ is the hypothetical emotional intensity attached to the occurrence of s. It corresponds to a not unexpected experience (when $U = 0$). In many cases, $E_h(s) = V(s)$, where V is a utility function. Events that were complex for the world to produce (C_w large) arouse more intense emotion when they occur, as they are more unexpected. Using (2), (3) can be rewritten: $e(s) = e_h(s)/p(s)$, where e_h and e stand for non-logarithmic emotions. The cognitive complexity $C(s)$ decreases $E(s)$ in (3). It acts like an emotional 'tax' paid for considering the event.[1]

[1] In (2), U must remain positive. In (3), U may be negative, but E must be positive. These constraints can be used to define the *relevance* of events (Dessalles, 2008a).

If s is not an event, but an anticipated situation, the expected emotion can be expressed using utility function V:

$$E_h(s) = V(s) - U(s) \quad (4)$$

The perspective of a situation that is complex for the world to produce (C_w large) arouses less emotion. In the non-logarithmic domain, equation (4) reads $e_h(s) = v(s) \times p(s)$.

In causal reasoning, we suppose that expected emotion propagates through causal links (Dessalles 2008). If a known emotional situation s is believed to result from situation s', then $E_h(s) = E_h(s')$. Using conditional complexity, we may write:

$$U(s) = U(s') + C_w(s|s') \quad (5)$$

By adding $E_h(s)$ to both sides, we get:

$$E(s') = E(s) - C_w(s|s') \quad (6)$$

ST's Predictions

In the absence of any precise counterfactual, as when one's house is struck by lightning, (3) provides a definition of luck, in line with (Rescher 1995):

$$L_1 = E_h(s_1) + U(s_1) \quad (7)$$

To assess the expected emotion $E_h(s_1)$ in such case, individuals may recall a known situation s of lightning on a house (or imagine it), and consider $E_h(s_1) = E_h(s) = L_1(s) - U(s)$.

In wheel of fortune situations, the expected emotional intensity $E_h(s_+)$ of winning corresponds to landing on a winning site s_+. The colored segment in figure 2 represents the winning sector in a linear version of the wheel of fortune. The complexity of landing on s_+ is $C_w(s_+) = \log_2 l_0$. This is the number of bits required by the 'world' to choose a landing position. According to (4), the maximum value of $E_h(s_+)$ is obtained for typical, *i.e.* maximally complex s_+: $C(s_+) = \log_2 l_2$. This is the number of bits required to discriminate among all winning positions. We get:

$$\max E_h(s_+) = V(s_+) - \log_2 l_0/l_2$$

This corresponds to the classical expected utility in the non-logarithmic domain.

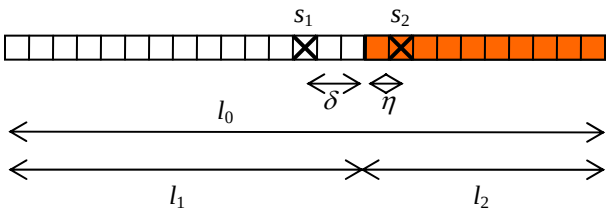


Figure 2: Discrete bounded near miss

When playing with a wheel of fortune, individuals acknowledge that the probability of landing in various sectors of the roulette is constant, but that landing close to a winning sector involves more intense bad luck (Teigen, 1996). Let us considered a linear version of the problem (figure 2).

After the draw, possibly for (self-)narrative purposes, individuals pick the situation s that maximizes emotional intensity $E(s)$. It may be the actual situation s_1, as in (7), or a counterfactual one s_2. Individuals are supposed to opt for the computation that gives the more intense emotion. In the counterfactual case, s_1 is seen as an intermediary step toward s_2. (3) and (6) give a new value for $E(s_1)$: $E_c(s_1) = E_h(s_2) + U(s_2) - C_{wc}(s_2|s_1)$. Luck is measured by the emotional gap between both emotions for s_1:

$$L_2 = E_h(s_2|s_1) + U(s_2) - C_{wc}(s_2|s_1) \quad (8)$$

Conditional $E_h(s_2|s_1)$ means that the expected emotional intensity is assessed using the actual emotional intensity of s_1 as baseline. The counterfactual nature of s_2 requires the introduction of a fictitious world wc that is able to keep a memory of s_1 to generate s_2. The term $C_{wc}(s_2|s_1)$ is the minimal price to pay for the 'If…'. It represents the size of the minimal parameter modifications that the observer can imagine for the 'world' to have generated s_2 instead of s_1.

In the case of figure 2, $E_h(s_2|s_1) = V(s_+)$, and $C_{wc}(s_2|s_1) = 1 + \log_2(\delta + \eta)$, which is the amount in bits needed to indicate the (non zero) targeting shift to the right toward s_2. On the other hand, $C_w(s_2) = \log_2 l_0$ and $C(s_2) = 1 + \log_2(1+\eta)$ (one bit to choose the left edge of the winning region, plus the representation in bits of the (possibly null) shift to reach s_2). We get: $L_2 = V(s_+) + \log_2 l_0 - \log_2(\delta+\eta)(1+\eta) - 2$. Taking $\eta = 0$ to maximize the intensity of unluck:

$$L_2 = V(s_+) + \log_2 l_0/\delta - 2 \quad (9)$$

The experience of bad luck in this near miss experience is an increasing function of the missed opportunity $V(s_+)$ and of the number l_0 of possibilities, and a decreasing function of the miss δ.

If the counterfactual is assessed against the expected emotion, here $\max E_h(s_+)$, instead of s_1, we get:

$$L_3 = V(s_+) + \log_2 l_2/\delta - 2 \quad (10)$$

This model accounts for the fact that when s_2 is more complex, as in figure 1(c), the intensity of (un)luck is smaller. We have $C(s_2) = \log_2 k + 1 + \log_2(1+\eta)$, where k is the number of winning regions. The intensity of luck is thus diminished by $\log_2 k$.

The extension to the continuous case is straightforward (figure 3). We suppose that the space is bounded to the left but not to the right. If we call α the landing precision, then $C_w(s_2) = \log_2(l_0/\alpha)$, as we need that number of bits to decide where to stop.[2] As previously, $C_{wc}(s_2|s_1) = \log_2(\delta+\eta)/\alpha$, and $C(s_2) = \log_2(1+\eta/\alpha)$. After taking the best choice $\eta = 0$, we get:

$$L_2 = V(s_+) + \log_2 l_0/\delta - 1 \quad (11)$$

[2] This supposes that there is a way to delimit numbers in the algorithm.

(the one-bit difference with (9) comes from the fact that the winning region has only one edge). Equation (11) accounts for emotions described by the expression: "fall short of the goal".

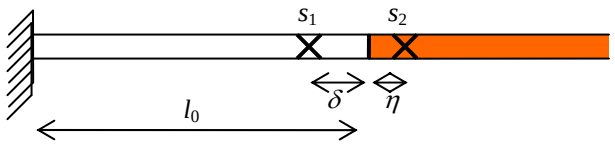


Figure 3: Continuous unbounded near miss

Equations (9) and (11) define the intensity of luck, but not only. They predict what the counterfactual situation s_2 will be (what many models of counterfactual thinking omit to do). Individuals pick the alternative s_2 that realizes the best compromise between high emotion $E(s_2)$ and low counterfactual complexity $C_{wc}(s_2|s_1)$.

Nine Stories

The following experiment was conducted to validate the predictions of the model. We tested 61 participants who accepted to pass the test on a Web site (www.dessalles.fr/expe/histoires). All contacted individuals had a high level of academic education, though in domains different from psychology or language sciences (mainly students in engineering). Nine short stories were presented to them (see Table 1). Each involved two or three choices. Instructions invited participants to choose options that made emotion maximal. Some choices irrelevant to the present study (such as the age of the victim in story S9) have not been exploited. Answers given after less than 20 sec. of reading were automatically discarded (median answering time per story was 90 sec.), which leaves us with a minimum of 56 answers per story. Presentation order (stories and options) was randomized.

Table 1: Abridged translation of the stories
(originals on www.dessalles.fr/expe/histoires).

S1- René is a railway worker. He works at the border, at a place where signals must me manually transmitted between the two networks. There is single-track line at [**9** (71*) / **23** (21) / **15** (7)] km from René's post. That day, René forgot to send the signal as a train crossed the border. He eventually did, but [**ten** (59*) / **fifty** (21) / **thirty** (20)] seconds before that, another train had entered the single-track line. The collision killed one of the two drivers.

S2- Lucas was heading for the metro station. At [**30** (71*) / **100** (20) / **800** (9)] m from the station, he stopped to lace up his shoe. As he arrived on the platform, the doors of the train closed in front of him. He had to wait [**25** (89*) / **15** (9) / **6** (2)] minutes for the next train.

S3- Michèle has been playing lotto every week for [**6** (84*) / **4** (11) / **2** (5)] years. On December [**19** (70*) / **3** (18) / **12** (12)], she told [**two** (60*) / **four** (32) / **three** (9)] friends of hers that she would stop playing. They persuaded her to bet for the special Christmas draw, on December 26. She did and won 62 000 Euros.

S4- Jacques was badly injured at his workplace by a defective machine on November 7. The defect had been previously notified and the machine was planned to be repaired on [**Novembe**r 8 (75*) / **November 17** (12) / **December 18** (12)].

S5- Florence works in a biology lab. Her two-[**year** (84*) / **month** (11) / **week** (5)] experiment on cell cultures was ruined by a student who knocked over a shelf. This broke [**all boxes containing** (35) / **a bottle of formalin that fell on** (45) / **the automatic device nourishing** (20)] the cell cultures. Florence was furious. She discovered that the student was the son of [**her neighbor** (67*) / **her former PE teacher** (15) / **the piano teacher of her sister** (18)].

S6- A young writer is admitted to Magalie's emergency department at the hospital. Her condition deteriorates. [**8** (66*) / **4** (21) / **6** (14)] infectious agents may explain the illness. Magalie sends samples to the lab and tests are conducted in parallel. It takes [**seven** (79*) / **three** (16) / **five** (5)] hours to get the result and the patient is saved at the last minute. Magalie remembers that she saw the name of the virus in [**the media, as well-know singer recently died of it** (52*) / **the record of another patient** (28) / **a specialized journal** (21)].

S7- For [**four months** (76*) / **two months** (21) / **two weeks** (3)] I was thinking of changing my cell phone. I eventually went to SFR Thursday at 1pm. I had to pay part of it because I was lacking 1000 points. [**Thursday** (74*) / **Friday** (21) / **Tuesday** (5)] evening, I received an offer: "change your phone, SFR offers you [**1500** (55) / **4000** (38) / **500** (7*)] points".

S8- Ms Tsuda's daughter had invited [**two friends** (71*) / **all girls in her class** (17) / **four friends** (12)] to her house. One of them was late. She had left her own house long ago. Ms Tsuda walked toward the girl's house and arrived at a level crossing, located at [**200** (55*) / **500** (24) / **900** (21)] m from Ms Tsuda's house. There was indeed an accident involving a young girl. It turned out that the invited girl was not involved and was late because of a detour caused by the accident.

S9- Helen, retired teacher, fainted as she was walking in the woods. She was found by [**a retired couple** (49) / **a colleague teacher** (26) / **a member of her bridge club** (25)] who called the rescue team. Helen would not have survived if she had reached the hospital [**half an hour** (77*) / **one hour** (16) / **one hour and a half** (7)] later.

Note: Choices irrelevant to the present study are not shown here. Numbers in parentheses indicate percentages. Asterisk indicates significance ($p < 0.001$). Underlined numbers indicate model predictions.

As shown in Table 1, most results were significant and 19 of the 21 majority choices are congruent with the model's predictions.

Analysis

Some results are commented now in the light of the theory.

Emotions: The intensity of the actual event, $E(s_1)$, was tested in story S2 (Lucas's waiting time), and in story S5 (duration of Florence's lost experiment). Unsurprisingly, majority choices make $E(s_1)$ maximal. In story S7, the third choice (number of points offered) influences $E_h(s_2)$: option "500",

which would lead to a smaller value of $E_h(s_2)$, was discarded by participants.

Counterfactual simplicity: In story S8, counterfactual s_2 corresponds to the invited girl (G) being involved in the accident ('it could have been her'). Both majority choices in S8 tend to make s_2 simpler, in agreement with equation (8). Participants clearly preferred that the invited girl (G) be one among 2 (71%) instead of one among 5 (17%) or 30 (12%), thus making the minimal description of G smaller by $\log_2 n - 1$ in comparison with $n = 5$ and $n = 30$. Similarly, by choosing the closest location (200m (55%)) instead of 500m or 900m for the counterfactual accident, they saved bits on $C(s_2)$ ($\log_2(500/200)$ and $\log_2(900/200)$).

Duration before near miss: In story S7, participants judged important that the hero hesitated four months (76%) instead of two months or two weeks before buying her/his telephone. We are in a case of unbounded near miss, and as predicted by equation (11), participants preferred the largest value for L. The same phenomenon explains the strong preference for the fact that Michèle has been playing for 6 years (84%) in story S3 (in this case, the winning 'sector' is s_1 and it is reached, but the computation is identical).

Proximity in near miss: Equation (11) predicts that emotion is maximum when one ends up close to the border between 'winning' and 'loosing' sectors (δ small). Several stories represent near miss situations. In S1, the train accident would have been prevented if the signal had been sent k×10sec before (k = 1 preferred (59%)); In S4, the worker would not have been badly injured if the accident had occurred k days later (k = 1 preferred (75%)); in S7, the cost would have been saved if the purchase had been made k days later (k = 1 preferred (74%)); in S9, Helen would have died if her admission had been delayed by k×30min (k = 1 preferred (77%)).

Causal Thinking in Good or Bad Luck

When confronted with events they perceive as (un)lucky, people tend to construct causal explanations for why these events happened (Pritchard & Smith, 2004). Causal thinking may produce counterfactuals by negating causes of the actual event, but also by enabling conditions for the counterfactual (Byrne, 2007). In what follows, we show how causal thinking can be accounted for within the ST framework.

Suppose that a cause s_3 can be found to explain s_1. If we use (5) together with (7), we get:

$$L_1 = E_h(s_1) + U(s_3) + C_w(s_1|s_3) \qquad (12)$$

This relation shows that unexpected causes ($U(s_3)$ large) and materially complex causal links will tend to increase the feeling of (un)luck in the non-counterfactual case.

If s_4 is a counterfactual alternative to s_3 that would have led to s_2, we can compute L_2 from s_3. Using (8):

$$L_2 = E_h(s_2|s_3) + U(s_2) - C_{wc}(s_2|s_3)$$

We may decompose $C_{wc}(s_2|s_3)$:

$$L_2 = E_h(s_2|s_3) + U(s_2) - C_w(s_2|s_4) - C_{wc}(s_4|s_3) \qquad (13)$$

The term $C_{wc}(s_4|s_3)$ measures the mutability of s_3 (Byrne, 2007). Equation (13) can be used to find a cause that people will be likely to select as mutable. Let us check these predictions against the experimental results.

Cause simplicity: Relation (12) predicts that simple causes ($C(s_3)$ small) will augment emotion since they are more unexpected. This is verified in story S5, where participants preferred the student responsible for the damage to be a neighbor's son (67%) instead of more complex individuals. In story S6, they preferred the virus to have been mentioned in the media (52%), rather than in a medical journal or a medical record where it would have been more complex to discriminate. Story S9 was also designed to test causal simplicity. We expected participants to reject option 'a retired couple', as these individuals would be more complex to discriminate than in the two other options ('a colleague teacher' and 'a member of her bridge club'). However, participants did not show the expected preference (49% *vs.* 26%+25%).

Causal link complexity: Relation (12) predicts that materially complex causal links ($C_w(s_1|s_3)$ large) are more unexpected and thus will augment emotion. Story S6 has been designed to check this point. Participants did prefer Magalie's eventual success to go through a seven hour (79%) test to decide between 8 (66%) infectious agents, rather than easier alternatives.

Causal link simplicity: Relation (13) conversely predicts that in counterfactual thinking, simple causal links will be preferred ($C_w(s_2|s_4)$ small). In story S1, participants chose the shortest distance between the railway worker's faulty action and its effect (71%); in story S2, they preferred Lucas to lace up his shoe close to the station (71%). In both cases, the material simplicity of the causal link diminishes the counterfactual complexity from the cause ('if he had sent the signal…', 'if Lucas had not paused to lace his shoe…') to the counterfactual effect. We failed to show the same effect in story S5, where we expected participants to chose the simpler causal mechanism ('broke all the boxes') instead of more complex ones ('broke a bottle of formalin'; 'broke the nourishing device'). The probable reason is that a simple causal link is preferable if one adopts Florence's counterfactual thinking, whereas a complex causal link is preferable if we only consider the newsworthiness of the story, what some participants seem to have done despite the instructions.

Discussion

The strong point of this study was to show the relevance of the notion of complexity in the study of the perception of luck. Many judgments about (un)lucky situations are not explained by variations of probability (even perceived probability) (Teigen, 1996). However, they vary in a

systematic way according to variations in complexity. We tried to connect people's attitude toward luck with predictions from Simplicity Theory, with some positive results.

Another positive aspect of the study is to highlight several intervening factors that have gone unnoticed in previous studies, such as the simplicity of the counterfactual situation (story S8), the fact that proximity is measured on a relative scale (stories S3, S7), or the simplicity of causes (stories S5, S6). The model also provides quantitative laws, e.g. for the wheel of fortune near miss.

We had two negative results in the experiment (story 5, choice 2 and story 9, choice 1). Note, however, that both consist in qualitative choices, which are more prone to complex interpretations by participants. The failure in S5 is likely to result from the bad design of the story; the failure in story 9 remains a mystery (perhaps the association due to word 'retired' being used twice is sufficient in rapid readers to make the rescuers seem simple).

In its current state, this theory of luck is not as simple as it should be. There are still some conceptual connections to be done that will make the link between equations and the processing of emotional intensities more transparent. The present account is meant as an attempt to depart from mere lists of factors and to outline an integrated model of the human ability to perceive luck in events.

The research, initiated in the recent years, on the cognitive role of descriptive complexity has already produced valuable results. The model presented in this paper is meant as a contribution to this enterprise. The sensitivity to complexity differences, which is central to ST, seems to be a general law, which applies across modalities and at all levels of abstraction. Its importance in the processing of some emotions that are involved in decision processes, such as the feeling of being (un)lucky after the occurrence of an event (Loomes & Sugden 1982), should encourage further investigation in this domain.

References

Byrne, R. M. J. (2002). Mental models and counterfactual thoughts about what might have been. *Trends in cognitive sciences*, *6* (10), 426-431.

Byrne, R. M. J. (2007). Précis of the Rational Imagination: How people create alternatives to reality. *Behavioral and Brain Sciences*, *30* (5/6), 439-480.

Chater, N. (1999). The search for simplicity: A fundamental cognitive principle? *The Quaterly Journal of Experimental Psychology*, *52* (A), 273-302.

Chater, N. & Vitányi, P. (2003). Simplicity: a unifying principle in cognitive science? *Trends in cognitive sciences*, *7* (1), 19-22.

Dessalles, J-L. (2006). A structural model of intuitive probability. In D. Fum, F. Del Missier & A. Stocco (Eds.), *Proceedings of the seventh International Conference on Cognitive Modeling*, 86-91. Trieste, IT: Edizioni Goliardiche. www.dessalles.fr/papiers/pap.cogni/Dessalles_06020601.pdf

Dessalles, J-L. (2008a). *La pertinence et ses origines cognitives - Nouvelles théories.* Paris: Hermes-Science Publications. http://pertinence.dessalles.fr

Dessalles, J-L. (2008b). Coincidences and the encounter problem: A formal account. In B. C. Love, K. McRae & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, 2134-2139. Austin, TX: Cognitive Science Society. www.dessalles.fr/papiers/pap.conv/Dessalles_08020201.pdf

Dimulescu, A. & Dessalles, J-L. (2009). Understanding narrative interest: Some evidence on the role of unexpectedness. In N. A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, 1734-1739. Amsterdam, NL: Cognitive Science Society. http://141.14.165.6/CogSci09/papers/367/paper367.pdf

Kahneman, D. & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review*, *93* (2), 136-153.

Kahneman, D. & Varey, C. A. (1990). Propensities and counterfactuals: The loser that almost won. *Journal of Personality and Social Psychology*, *59* (6), 1101-1110.

Loomes, G. & Sugden, R. (1982). Regret theory - An alternative theory of rational choice under uncertainty. *The Economic Journal*, *92*, 805-824.

Pritchard, D. & Smith, M. (2004). The psychology and philosophy of luck. *New ideas in psychology*, .

Rescher, N. (1995). *Luck: The brilliant randomness of everyday life.* New York: Farrar, Straus, and Giroux.

Roese, N. J. (1997). Counterfactual thinking. *Psychological Bulletin*, *121* (1), 133-148.

Solomonoff, R. J. (1978). Complexity-based induction systems: Comparisons and convergence theorems. *IEEE transactions on Information Theory*, *24* (4), 422-432. http://world.std.com/~rjs/solo1.pdf

Teigen, K. H. (1996). Luck: The art of a near miss. *Scandinavian journal of psychology*, *37* (2), 156-171.

Teigen, K. H. & Jensen, T. K. (2010). Unlucky victims or lucky survivors: Spontaneous counterfactual thinking by families exposed to the tsunami disaster. *European psychologist*, , in press.

Teigen, K. H. & Keren, G. (2003). Surprises: low probabilities or high contrasts?. *Cognition*, *87*, 55-71.

Teigen, K. H. (2005). When a small difference makes a big difference - Counterfactual thinking and luck. In D. R. Mandel, D. J. Hilton & P. Catellani (Eds.), *The psychology of counterfactual thinking*, 129-146. Oxon, UK: Routledge.

Wohl, M. J. A. & Enzle, M. E. (2003). The effects of near wins and near losses on self-perceived personal luck and subsequent gambling behavior. *Journal of experimental social psychology*, *39*, 184-191.

Scan Patterns on Visual Scenes predict Sentence Production

Moreno I. Coco (M.I.Coco@sms.ed.ac.uk) and
Frank Keller (keller@inf.ed.ac.uk)
School of Informatics, University of Edinburgh
10 Crichton Street, Edinburgh EH8 9AB, UK

Abstract

A range of cognitive modalities are involved in everyday tasks, which raises the question to which extend these modalities are coordinated. In this paper, we focus on two particular aspects of this coordination: linguistic structure and visual attention during sentence production, based on the hypothesis that similar scan patterns are associated with similar sentences. We tested this hypothesis using a dataset from an eye-tracking experiment in which participants had to describe photo-realistic scenes. We paired each sentence produced with the corresponding scan pattern, and computed a range of similarity measures for both modalities. Correlation and mixed model analyses confirmed that trials involving similar scan patterns also involve similar sentences productions. This was true for all pairs of linguistic and scan pattern similarity measures we investigated. The result holds both before and during sentence production, and for within-scene and between-scene analyses.

Keywords: scan patterns; sentence production; eye-tracking; sentence similarity.

Introduction

Everyday tasks demand the coordination of a range of cognitive modalities. If the task is to make tea, for example, then motor actions (e.g., arm-hand movement) and visual attention (e.g., looking at the pot) have to be coordinated (Land, 2006). This implies that if two different persons perform the same task, they will do so in a similar way. It follows that the sequence of scan patterns, i.e., eye fixations across spatial locations in temporal order (Noton and Stark, 1971) as well as the sequence of motor actions, will be similar across participants (Land, 2006).

In this paper, we investigate whether a similar evidence of cross-modal coordination can be found when vision and language have to be coordinated. In particular, we focus on the similarity between scan patterns and linguistic structures in a language generation task.

In the visual cognition literature, similarity of scan patterns has not received much attention, mainly because of the high variability across participants (Henderson, 2003). There are some results, however, that point toward a range of visual factors that can trigger similarity. Often, these factors are related to the task (Castelhano et al., 2009) and to the degree of cross-modal interactivity required to perform it.

In tasks with a low level of interactivity, i.e. free viewing, visual attention is mainly guided by **exogenous** factors like *saliency* (Itti and Koch, 2000): a measure of visual prominence based on low-level features (color, intensity and orientation). A free viewing task does not require visual attention to interact with knowledge-based (top-down) information. The low interactivity of free viewing means the visual responses are driven by exogenous visual mechanisms while minimizing the need for cross-modal coordination.

Different patterns of visual attention emerge in other visual tasks, such as *memorization* or *imagery* (Humphrey and Underwood, 2008), where participants are asked to memorize images in preparation for a recall phase. In the recall phase, despite the absence of the original stimuli (preventing bottom-up effects), scan patterns on a blank screen were more similar across participants within the same scene than across different ones (Humphrey and Underwood, 2008). In this case, the task requires an **endogenous** control of visual attention through top-down knowledge, i.e., scene layout, contextual information, and even semantic relations between objects (Hwang et al., 2009). Thus, exogenous bottom-up effects are overridden by endogenous contextual guidance effects.

These results, consistent with similar findings from visual search studies (Yang and Zelinsky, 2009), suggest that in tasks requiring endogenous control, categorical and semantic information is activated. It seems plausible that this endogenous access to categorical information is activated during daily actions (Land, 2006); e.g., categorical knowledge about the tea pot (i.e., it has a handle to grasp) is necessary to allow cross-modal coordination between visual attention and motor action.

It is important to notice that this information has a direct link with language processing. Categorical information, in fact, is typically expressed verbally, and drives linguistic tasks such as scene description. It seems likely that the same mechanism, based on categorical information, which allows coordination between motor-action and visual attention might also underlie the coordination between language processing and visual attention.

Previous research has looked at the interaction between vision and language principally using the visual world paradigm (VWP, Tanenhaus et al. 1995), an eye-tracking paradigm which has demonstrated clear links between the processing of certain linguistic constructions and the access to visual contextual information (Knoeferle and Crocker, 2006). Research in this field suggests a tightly coupled relation between vision and language, but previous works has largely focused on specific psycholinguistic phenomena (e.g., attachment ambiguity), rather than uncovering the shared mechanisms by which this interaction takes place. We explain this coupled relation assuming a categorical interface which coordinates the cross-modal, visual and linguistic, interaction.

In this paper, we investigate the extent to which visual and



Cue - Animate: "Man" **Cue - Inanimate: "Suitcase"**

Figure 1: Example of scene and cues used as stimuli for the description task

language processing are synchronized when participants perform a task viz., scene description in a visual context, which requires endogenous interaction between linguistic and visual processing. Our main hypothesis is that scan patterns and sentences are correlated, i.e., if two trials involve similar scan patterns, then the sentences produced in these two trials will also be similar.

Experimental Setting

In this section, we discuss how the data was collected and processed, and explain how we computed the measures of scan pattern and linguistic similarity.

Data Collection and Pre-processing

In an eye-tracking language production experiment (Coco and Keller, 2010), we asked participants to describe photo-realistic indoor scenes after being prompted with cue words which referred to visual objects in the scenes. The cue words were either animate or inanimate (e.g., *man* or *suitcase*) and were ambiguous with respect with the scene (see Figure 1 for an example trial). Participants' eye-movements were recorded using an Eyelink II eye-tracker with at sampling rate of 500 Hz on a 21" screen (1024 x 768 pixel resolution), while the speech of the participants was recorded with a lapel microphone. We collected a total of 576 sentences produced for 24 scenes which were drawn from six different scenarios (e.g., bedroom, entrance). The sentences were manually transcribed and paired with the scan patterns that participants followed when generating the sentences. We removed two pairs because the sentences were missing.

The data varies across participants and scenes both in terms of the complexity of the sentences (i.e., *one man waits for another man to fill out the registration form for a hotel* vs. *the man is checking in* for Figure 1) and in the length of the scan patterns produced both in preparation for production (min = 800 ms; max = 10205 ms) and during production (min = 2052 ms; max = 18361 ms). Both types of variability have to be taken into account when developing metrics for sentence and scan pattern similarity.

Similarity Measures

Before quantifying the association between scan patterns and sentence productions, we measure similarity within each modality. We defined two similarity (or equivalent, dissimilarity) measures for both modalities. Applying more than one measure makes it less likely that our results will be an artifact of the type of measure used.

Sentence Measures We define two similarity measures on sentences: Feature Dissimilarity (FD) and semantic similarity computed using Latent Semantic Analysis (LSA). We pre-process the sentences produced by the participants using an automatic part of speech (POS) tagger (Toutanova and Manning, 2000), whose reported accuracy is 96.8% on the Penn Treebank. The POS tags make it easy to extract relevant information from a sentence.

For FD measure, we represent each sentence as a vector, each element of which represents a feature of the sentence. We include semantic and syntactic features, as well as contextual information derived from the scenario a scene belongs to. (In the result section, we also report correlation coefficients obtained when excluding the contextual features.)

Syntactic features include (1) the length of the utterance, which is a general indicator of complexity while also reflecting the total number of visual referents, and (2) the presence of coordination, which reflects disambiguation strategies. As **semantic features** we include (1) the verb frame and (2) semantic similarity between verbs. The verb frame encodes the arguments the verb can take, obtained from WordNet (e.g., transitive or intransitive); semantic similarity is computed using Jiang and Conrath's (JC) synset path-distance (Budanitsky and Hirst, 2006). This distance measure is based on the number of nodes from one verb to another in the WordNet database. We calculate pairwise JC distance on all pairs of unique verbs in our corpus of sentence productions; we then apply hierarchical clustering to group together similar verbs. Cluster membership is the feature value included in the FD vector.

The **contextual features** include (1) the animacy of the cue word, useful to discriminate between different descriptive routines and, (2) the scenario in which the sentence was produced (e.g., bathroom, entrance). Notice that the contextual features are not scene specific; each scenario is represented by four different scenes.

After converting each sentence into a vector of features, we calculate FD between all pairs of sentences by applying Gower distance (Gower, 1971), which can be used when the data is both numerical and categorical.

LSA measures the similarity between words based on the co-occurrence of content words within a collection of docu-

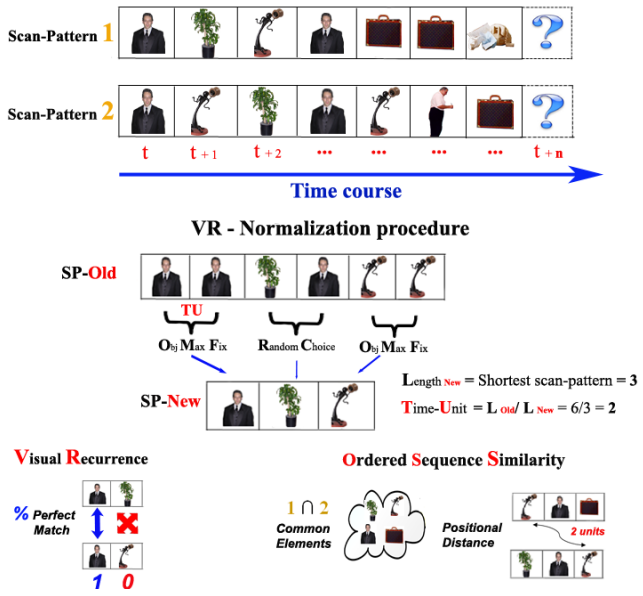


Figure 2: Example of how scan patterns are represented and normalized (for VR only); and how measures of scan pattern similarity are computed

ments (in our case the British National Corpus). It indicates how likely two words are to occur in the same document. Different from Hwang et al. (2009) where LSA is calculated between individual words, we implemented a version of LSA generalized to compute the similarity of sentences (Mitchell and Lapata, 2009). We compute an LSA vector for each content word in the sentence (context window of size five; low frequency words are removed) and then combine these vectors using addition to obtain a sentence vector (an alternative discussed by Mitchell and Lapata 2009 would be vector multiplication). Similarity between sentence vectors is measured using cosine distance.

Scan Pattern Measures We use two measures to compute the similarity between scan patterns: Visual Recurrence (VR) and Ordered Sequence Similarity (OSS, Gomez and Valls 2009).

We consider scan patterns as temporally ordered sequences of fixated target objects. Each trial is therefore encoded as a sequence of discrete time points, each annotated with the object fixated at that time, encoded numerically (see Figure 2). VR is a percentage measure of scan pattern similarity that counts the frequency of looks to the same objects during the same time points between two scan patterns relative to its total length. For example, in Figure 2, we have two matches on a total of seven time points, i.e., 25.87% agreement between the scan patterns.

VR can only compare scan patterns equal in length. We therefore normalize each scan pattern (SP_{old}) by mapping it onto a normalized time course of fixed length (SP_{new}). The length of SP_{new} is set on the basis of the shortest eye-movement sequence found across all participants. For each SP_{old}, we obtain the number of time-points corresponding to a time unit of SP_{new} by simply dividing the length of SP_{old} with the length of SP_{new}. Over the SP_{old} time-points, we look for the object which has received the highest number of looks and map it into the corresponding time-unit of SP_{new}. The final result is a normalized scan-pattern of fixed length containing the objects most likely to be fixated.

The second method used to compare scan patterns is Ordered Sequence Similarity (note that despite its name, OSS is in fact a dissimilarity measure). Its main advantage is that it can be used with sequences of different lengths, and has shown to be more effective than established measures such as edit distance (Gomez and Valls, 2009). OSS is based on two aspects of sequential data: the elements the sequence is composed of, and their positions. When comparing two sequences, it takes into account the number of common elements and their relative order. The first step is to find target objects that occur in both scan patterns. For example in Figure 2, four objects are shared by the two scan patterns (man, plant, statue, suitcase). For each common element, we calculate the distance between the two sequences, e.g., statue of scan pattern 1 is two units distant from statue in scan pattern 2. Distances are then summed and normalized on the basis of sequence lengths (for details refer to Gomez and Valls 2009).

All four measures of similarity are computed pairwise, i.e., every trial (sentence and scan pattern) is paired with every other trial. This resulted in a total of 164,164 pairs for each region of analysis, i.e., Before and During production.

Analysis

To analyze the correspondence between sentences and scan patterns, we divide the data into two regions: *Before* speech onset, and *During* production. The *Before* region provides evidence about the process of utterance planning and visual information retrieval, whereas *During* is informative about linguistic encoding and the utilization of visual information during this process. We perform two types of analysis: descriptive and inferential.

In the descriptive analysis, we investigate the data at two levels: (1) globally, i.e., by performing comparisons between all pairs of trials in the full data set, and (2) locally, i.e., by comparing only the trials that pertain to a given scene (24 in total). These two forms of analysis make it possible to test whether the correspondence between sentences and scan patterns is scene specific. For comparison, we also report a baseline correlation (Humphrey and Underwood, 2008) that is obtained by pairing sentences and scan patterns randomly (rather than pairing the scan patterns with the sentences they belong to).

We quantify the strength of the correspondence between similarity measures by computing Spearman's ρ for all pairs of measures. We do not report coefficients for the baselines, as they are not significant across all combined measures: $\rho \approx 0.002; p > 0.1$. For the correlation analysis, we also con-

sider a variant of the Feature Dissimilarity measure for which we remove the contextual features (FD-C). This makes it possible to investigate the contribution of scenario and animacy of the cue word to the correspondence between scan pattern and sentence similarity.

The distinction we made between global and local similarity has implications for the nature of correspondence. A correlation found globally (across all scenes) would imply that scan patterns are partially independent from the precise layout of the scene, i.e., position of the objects, etc., as these factors varied across scenes, but rather dependent on the categorical structure shared, i.e. the visual referents common across scenes. A correlation found at the local level would be consistent with well-known scene-based effects, both bottom-up and top-down, which guide visual attention (Itti and Koch, 2000; Humphrey and Underwood, 2008).

In the inferential analysis, we apply linear mixed effects modeling (LME) (Baayen et al., 2008) using the R-package lme4. We use scan pattern similarity as the dependent variable (fitting a separate model for OSS and VR) and sentence similarity (FD and LSA) as predictors. The region of analysis (before or after) is also included as a factor. As random effects, we included participants and trials.[1] All fixed factors were centered to reduce collinearity. The models are built following a forward step-wise procedure. We start with an empty model, then we add the random effects. Once all random effects have been evaluated, we proceed by adding the predictors. The parameters are added one at time, and ordered by their log-likelihood improvement of model fit: the best parameter goes first. Every time we add a new parameter to the model (fixed or random), we compare its log-likelihood against the previous model. We retain the additional predictor if log-likelihood fit improves significantly ($p < 0.05$). The final model is therefore the one that maximizes fit with the minimal number of predictors.

Results and Discussions

Figure 3 plots the linguistic similarity measures LSA and FD against the scan pattern similarity measure OSS[2], computed globally, i.e. across all scenes. We bin the data on the x-axis and include 95% confidence intervals. The plots also include the random baseline.

For both linguistic measures, we observe a clear trend between sentence and scan pattern: when LSA similarity increases, scan pattern dissimilarity decreases; when feature dissimilarity (FD) increases, OSS also increases. This effect is observed both Before and During region, but not in the random baseline.

We also observe a difference in the intercept between the Before and During region. In the Before region, there is less dissimilarity between scan-patterns overall. This could indicate a higher degree of coordination between the two modalities during sentence planning, compared to sentence encoding. During planning, visual attention integrates the categorical information of the scene with the linguistic referents selected as arguments of the sentence. When production starts, detailed information is sourced from the visual processor to drive encoding, thus triggering more specialized routines of visual information retrieval.

[1]Similarity is calculated pairwise. Thus, we need to include as random variables two participants and two trials for each pair.

[2]For reason of space, VR is shown only in the LME results.

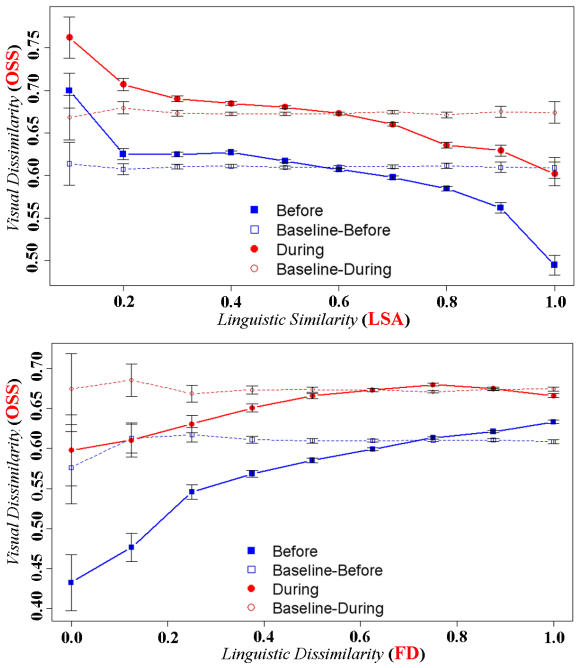


Figure 3: Correlation between LSA similarity, Feature dissimilarity (FD) and Ordered Sequence Similarity (OSS)

Figure 4 plots local similarity values, i.e., values computed separately for each scene (OSS against LSA)[3]. Generally, the

[3]Again, for space limitation, we can show only one pair of combined measures, OSS/LSA. However, we observe a similar trend for all the other pairs.

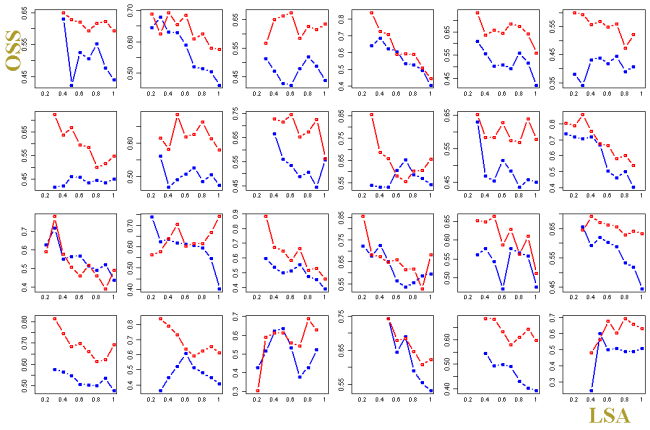


Figure 4: Scan pattern dissimilarity (OSS) as a function of the Linguistic Similarity (LSA) across all 24 scenes

Table 1: Correlations (Spearman ρ) between the different similarity measures. *Before* and *During* aggregated

Measures	VR	OSS	FD	LSA
OSS	-0.63^{***}			
FD	-0.07^{***}	0.15^{***}		
LSA	0.15^{***}	-0.10^{***}	-0.06^{***}	
FD-C	-0.02^{**}	0.01^{*}	0.86^{***}	-0.10^{***}

Table 2: Minimum and Maximum correlations (Spearman ρ) across different scenes between the different similarity measures

Measures		VR	OSS	FD
OSS	min	−0.10		
	max	−0.56		
FD	min	−0.01	−0.02	
	max	−0.55	0.44	
LSA	min	0.01	−0.001	−0.52
	max	0.33	−0.30	−0.79

trend previously observed at the global level is confirmed, both for the *Before* and the *During* region, though there is some variation in the degree of association between scan patterns and linguistic similarity across scenes.

Table 1 lists the correlation coefficients for all pairs of similarity measure. There are weak but significant correlations across all measures. In particular, both VR and OSS are significantly correlated with both FD and LSA in the direction expected, i.e., positively in case of dissimilarity and negatively in the case of similarity. Between the two scan pattern measures (OSS and VR), we observe a strong correlation, whereas the association between the two linguistic measures (FD and LSA) is weak. We also observe that FD-C, the measure obtained by removing contextual information from FD is highly correlated with FD, but the removal of contextual information weakens the correlation with the scan pattern measures. On the other hand, FD-C is somewhat more strongly correlated with LSA than FD is. It seems that the contextual information, even if at the level of the scenario, prominently contribute to the prediction of scan pattern similarity.

In Table 2, we show the minimum and maximum values of the correlation coefficients for similarity measures observed locally, i.e. computed trials aggregated by scene. As expected from the plots in Figure 4, correlation coefficients vary across scenes for all pairs of measures. The context of the individual scenes modulates the correspondence between scan patterns and linguistic productions. Compared to the global coefficients, the most noticeable difference is a strengthening of the correlation between the two linguistic measures FD and LSA. It seems that in a scene context, the semantic information expressed by LSA more directly matches the information in FD, which also includes verb semantics and scenario information.

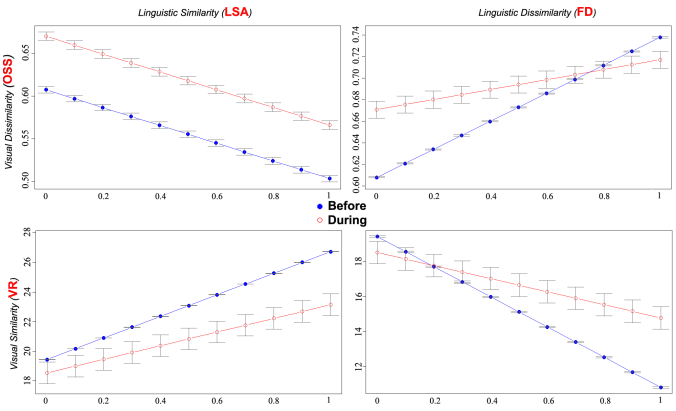


Figure 5: Predicted values of the linear mixed effects model: linguistic similarity predicted by scan pattern similarity

Table 3: LME coefficients. The dependent measures are: *OSS* and *VR*. The predictors are: *Region* (contrast coding: Before = −0.5; During = 0.5) and the Linguistic Measures (LM) FD or LSA. Each column shows which linguistic/scan pattern similarity measure is compared

Predictor	FD/OSS	FD /VR	LSA/OSS	LSA/VR
Intercept	0.0879^{***}	18.95^{***}	0.639^{***}	18.97^{***}
Region	0.062^{***}	-0.907^{***}	0.062^{***}	-0.906^{***}
LM	0.087^{***}	-6.151^{***}	-0.104^{***}	5.953^{***}
LM:Region	-0.083^{***}	4.866^{***}	n.sig.	-2.687^{***}

Turning now to the inferential analysis, Figure 5 plots LME predicted values calculated globally for all pairs of measures. The models closely follow the empirical patterns in Figure 3. Table 3 lists the coefficients of the mixed models; we find a significant main effect of scan pattern similarity for both FD and LSA, for both the OSS and the VR model. Moreover, we observe a main effect of region across all combined measure: for the *Before* region, sentence similarity is more strongly related to scan pattern similarity, compared to the *During* region.

Furthermore, we observe an interaction of region of analysis and linguistic similarity: for *Before* region, the similarity between sentence and scan pattern has a steeper change, compared to *During*. In linguistically driven visual planning, we retrieve the referents going to be encoded. Thus, if two sentences are going to be very different, the set of referents chosen during visual planning is also going to be very different. During encoding instead, the visual system is already sourcing detailed information in a sentence specific way, thus the magnitude of change is smaller compared to planning.

General Discussion

A range of cognitive modalities are involved in everyday tasks, which raises the questions to which extend these modalities are coordinated. In this paper, we focused on two particular aspects of this coordination: linguistic structure and visual attention during sentence production. Our main hy-

pothesis was that similarity of scan patterns predict the similarity of sentences.

We tested this hypothesis using a dataset from an eye-tracking experiment in which participants had to describe photo-realistic scenes. We paired each sentence produced with the corresponding scan pattern, and computed similarity measures for both modalities. We used Visual Recurrence and Ordered Sequence Similarity to compare scan patterns, while for sentences we used a semantic similarity measure based on LSA and a feature dissimilarity measure that combines syntactic, semantics, and contextual information.

Both descriptive and inferential analysis confirmed our hypothesis: if two trials involve similar scan patterns, then the sentences produced in these two trials are also similar. This was true for all pairs of linguistic and scan pattern similarity measures. Furthermore, we subjected the data to a global analysis (i.e., we computed similarity across different scenes) and a local analysis (i.e., we only compared scan patterns and sentences within the same scene). Significant correlations were found in both cases, which suggests that the correspondence between sentences and scan patterns cannot be explained as a simple mapping between individual scene content and the objects mentioned in the corresponding sentence. This conclusion is confirmed at the level of individual scenes, where the variability observed suggests the presence of different visual and linguistic factors modulating the strength of the correspondence.

An important point emerged during our analysis regarding the role of contextual information in predicting similarity. When contextual features were removed from the linguistic measure, the strength of the correlation was reduced (but was still significant). Even though our contextual features were not scene specific, but rather pertained to more general scenarios, they were still helpful in predicting scan patterns.

Within the broader context of cognition, our results indicate that in tasks demanding the interaction of vision and language, where endogenous control plays an essential role, they synchronize processing through coordination over a shared categorical interface.

Ongoing work is currently investigating the sequential and temporal aspects of the correspondence using alignment techniques borrowed from bio-informatics. Preliminary results show that the inclusion of temporal information together with a more stringent analysis of sequential data increase correlation between sentences and scan patterns.

Finally, in future work we plan to investigate a range of linguistic features separately, thus enabling us to establish which aspects of scan patterns predict syntactic, semantic, or contextual aspects of sentence production.

References

Baayen, R., Davidson, D., and Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59:390–412.

Budanitsky, A. and Hirst, G. (2006). Evaluating wordnet-based measures of semantic distance. *Computational Linguistics*, 32(1):13–47.

Castelhano, M., Mack, M., and Henderson, J. (2009). Viewing task influences eye movement control during active scene perception. *Journal of Vision*, 9(3)(6):1–15.

Coco, M., I. and Keller, F. (2010). Sentence production in naturalistic scene with referential ambiguity. In *R. Catrambone & S. Ohlsson (Eds.), Proceedings of the 32th Annual Conference of the Cognitive Science Society, Portland*.

Gomez, C. and Valls, A. (2009). A similarity measure for sequences of categorical data based on the ordering of common elements. *Lecture Notes in Computer Science*, 5285/2009:134–145.

Gower, J. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, 27:623–637.

Henderson, J., M. (2003). Human gaze control during real-world scene perception. *Trends in Cognitive Sciences*, 7:498–504.

Humphrey, K. and Underwood, G. (2008). Fixation sequences in imagery and in recognition during the processing of pictures of real-world scenes. *Journal of Eye Movement Research*, 2(2):1–15.

Hwang, A., Wang, H., and Pomplun, M. (2009). Semantic guidance of eye movements during real-world scene inspection. In *N.A. Taatgen & H. van Rijn (Eds.), Proceedings of the 31th Annual Conference of the Cognitive Science Society, Amsterdam*.

Itti, L. and Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40(1):1489–1506.

Knoeferle, P. and Crocker, M. (2006). The coordinated interplay of scene, utterance and world knowledge. *Cognitive Science*, 30:481–529.

Land, M. (2006). Eye movements and the control of actions in everyday life. *Progress in Retinal and Eye Research*, 25:296–324.

Mitchell, J. and Lapata, M. (2009). Language models based on semantic composition. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 430– 439.

Noton, D. and Stark, L. (1971). Eye movements and visual perception. *Scientific American*, 224(1):34–43.

Tanenhaus, M., Spivey-Knowlton, M., Eberhard, K., and Sedivy, J. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, (268):632–634.

Toutanova, K. and Manning, C. (2000). Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000)*.

Yang, H. and Zelinsky, G. (2009). Visual search is guided to categorically-defined targets. *Vision Research*, 49:2095–2103.

Modulation of motor-meaning congruity effects for valenced words

Geoffrey Broookshire[1] (geoff.brookshire@mpi.nl)
Richard Ivry[2] (ivry@berkeley.edu)
Daniel Casasanto[1,3] (daniel.casasanto@mpi.nl)

[1]Max Planck Institute for Psycholinguistics, Neurobiology of Language Group, Nijmegen, NL
[2]University of California at Berkeley, Department of Psychology, Berkeley CA, USA
[3]Donders Center for Brain, Cognition, and Behavior, Nijmegen, NL

Abstract

We investigated the extent to which emotionally valenced words automatically cue spatio-motor representations. Participants made speeded button presses, moving their hand upward or downward while viewing words with positive or negative valence. Only the color of the words was relevant to the response; on target trials, there was no requirement to read the words or process their meaning. In Experiment 1, upward responses were faster for positive words, and downward for negative words. This effect was extinguished, however, when words were repeated. In Experiment 2, participants performed the same primary task with the addition of distractor trials. Distractors either oriented attention toward the words' meaning or toward their color. Congruity effects were increased with orientation to meaning, but eliminated with orientation to color. When people read words with emotional valence, vertical spatio-motor representations are activated highly automatically, but this automaticity is modulated by repetition and by attentional orientation to the words' form or meaning.

Keywords: Automaticity, Metaphor, Motion, Space, Valence

Introduction

Do some abstract concepts depend, in part, on mental representations of physical space? According to theories of metaphorical mental representation, linguistic metaphors like 'a *rising* price', 'a *sliding* scale, or 'a *long* engagement' suggest that many of our abstract ideas are grounded in representations of *motion* and *space*. These are, in turn, grounded directly in perceptuo-motor experiences (e.g., Clark, 1973; Lakoff & Johnson, 1999; Talmy, 1988). Although initial arguments for metaphor theory were based on descriptive linguistic data, psychological experiments provide evidence for important links between spatio-motor representations and mental representations in more abstract domains like *power* (Schubert, 2005), *happiness* (Meier & Robinson, 2004), *time* (Boroditsky, 2000), *number* (Dehaene et al., 1993), and *similarity* (Casasanto, 2008). Yet researchers are just beginning to specify what roles spatial representations may play in abstract thought.

Debates about metaphorical representation have focused on two theoretical possibilities outlined by Murphy (1996), which were impossible to distinguish based on observational linguistic data, alone. On the Strong View, representations in metaphorical source domains (e.g., space) are necessary for conceptualizing target domains (e.g., time). According to Lakoff and Johnson (1999), activating source-target mappings is obligatory: without them, "abstract thought is virtually impossible." On the Weak View, however, source domain representations make an optional contribution to people's understanding of target domains. Boroditsky (2000) tested whether spatial representations are necessary for understanding temporal language, and concluded that "spatial schemas are *useful*, but *not necessary*" (italics added).

Framing experiments in terms of the necessity of source domain representations for understanding target domains (and for understanding target-domain language in particular) helped to transform a question that was long the province of linguists and philosophers into a question that is tractable using the psychologist's toolkit. Yet continuing to test a Strong-Weak dichotomy seems unlikely to lead to further new discoveries.

On nearly any theory of metaphor, source domain representations are hypothesized to be *part* of a more complex mental representation or word meaning: on the Strong View, a necessary part. The idea that there are *necessary parts* (i.e., features) of concepts or word meanings, however, is difficult to maintain. Wittgenstein (1953) famously exploded the notion that even a simple, relatively concrete word like *game* has any features that are necessarily present in all of its instantiations. It seems unlikely that more abstract words like *value* or *justice*, whose meanings are notoriously fluid, would have any necessary parts. This suggests the necessity question should be reframed in terms of functionality: What causes source domain representations to be activated, and what functional roles do they play in understanding target domains?

Psychologists have also raised a related question about metaphor (e.g., Meier & Robinson, 2004; Meier, et al., 2007): Are source domains activated *automatically* when people understand target domains? Automaticity is of interest because it is taken as evidence against the possibility that source-domain representations are only activated strategically (perhaps consciously) when people need to communicate about abstract ideas, or in response to task demands (Meier, et al., 2007). Curiously, however, automaticity has been treated as binary; source domains either *are* or *are not* activated automatically. Yet for most aspects of concepts and word meanings, it seems unlikely that activation is fully automatic − not in the same sense that people automatically perceive the lines in the Müller-Lyer illusion to be of different lengths. As classic studies of 'semantic flexibility' suggest, context can modulate the

activation of even those aspects of a word's meaning that might seem to be indispensable (e.g., Barclay, et al., 1974). Notions of automaticity that are well-suited for characterizing aspects of perceptual and motor processes may not be appropriate for characterizing aspects of meaning: meaning is not a reflex.

Traditional notions of necessity and automaticity must be tailored to fit questions about metaphor (and about meaning, more broadly). Rather than asking whether source domains are necessary for understanding target domains, it may be more fruitful to ask 'what functional roles do source-domain representations play in understanding target domains?' Rather than investigating *whether* source domain representations are activated automatically, it may be useful to ask '*to what extent* is their activation automatic, and *under what conditions* is their activation increased or diminished?' We take up these latter questions of automaticity here, assuming automaticity to be a continuum.

Emotional valence is an abstract domain that people often talk about using metaphors from space and motion: when people are optimistic they're *looking up*, and when they're sad they're *feeling down*; hopes can *rise*; morale can *drop*; spirits can *soar* or *plummet.* Behavioral studies suggest these linguistic metaphors correspond to mental metaphors: non-linguistic associative mappings from representations of motion or space to the representations of emotional valence. Stroop-like experiments show these mappings are activated when people process language with positive or negative valence, even when they're not using any linguistic metaphors.

In one study (Meier & Robinson, 2004), participants were faster to judge words like *polite* and *rude* as having positive or negative valence when positive words were presented at the top and negative words at the bottom of a computer screen (Experiment 1). Furthermore, judging words to be positive directed attention to the top of the computer screen, and judging them to be negative directed attention to the bottom (Experiment 2). Yet based on these experiments it would be premature to conclude that space-valence associations are 'automatic'. For one thing, the spatial variation from trial to trial was highly salient in Meier & Robinsons' experiments (in fact, impossible to ignore), and for another, participants made explicit judgments about the valence of the words. Thus, the tasks strongly focused attention on both the source and target domains.

To address these concerns, Casasanto (2008) adapted a spatial interference task of Zwaan & Yaxley's (2003) for use with valenced words. Participants saw pairs of words, one above and the other below fixation, and made speeded synonym-antonym judgments. Target word pairs were antonyms, one with positive and the other with negative valence. Participants were fastest to classify the pairs as antonyms when the positive word appeared above the negative (e.g., *wealthy* above *poor*). In a second experiment, participants were faster to make lexical decisions on positive-valence words (e.g., *brave*, *ethical*) when they were presented above non-word distractors, and on negative-valence words (e.g., *failure*, *hate*) when presented below non-word distractors. This was true even though neither the spatial position of the words, nor their valence, nor any other part of their meaning was relevant to the task.

In a third experiment, Casasanto (2008) presented positive and negative words in the center of a screen, in either red or blue letters. On the right and left of the screen there were three large boxes. The top box was red and the bottom box was blue (or vice versa). The middle box was white, and was filled with marbles. Participants were instructed that as soon as each word appeared, they should move one marble with each hand into the box corresponding to the color of the word's font, as quickly as possible. They moved marbles fastest when the direction of movement was congruent with the spatial schema suggested by the word's valence. This was true even though movements were cued only by the words' colors: not only was their meaning irrelevant, the tasks did not even require participants to process the words *as words.*

These Stroop-like congruity effects suggest that spatial representations are activated with a considerable degree of automaticity when people read valenced words. The goal of the present study was to test the limits of this automaticity. In Experiment 1, we tested whether repeating stimuli modulated the magnitude of the space-valence congruity effect. Casasanto's (2008) marble-moving task was adapted for use with button presses, to automate response coding. Stimuli were presented twice, in successive blocks, and reaction times were compared across blocks. In Experiment 2, we tested whether attentional orientation influenced the magnitude of space-valence congruity effects. We used a Task Set Inertia manipulation (Allport & Wylie, 2000). Distractor trials oriented attention during the target trials toward either semantic or perceptual aspects of the target words.

Experiment 1: Does repetition modulate motor-meaning congruity effects?

Experiment 1 tested whether motor-meaning congruity effects observed in previous studies would be modulated by repetition of the same stimulus words.

Methods

Participants Native English-speaking UC Berkeley students (N=20) participated in exchange for course credit or payment.

Materials

Two lists of 48 English words were created, one with positive and the other with negative valence (e.g., *wealthy*, *poor*, *virtuous*, *evil*, *joy*, *disgust*, etc.), totaling 96 stimuli. The words were nouns and adjectives that have no literal spatial meaning, but which subjects in a previous norming study spatialized consistent with their metaphorical associations (e.g., placing *wealthy* above *poor*; *virtuous* above *evil*, etc.) Positive and negative words did not differ

in frequency (p=0.70), number of syllables (p=0.60), or number of letters (p=0.12), by two-tailed t-tests.

Stimuli were presented on a CRT monitor with a refresh rate of 60 Hz. A standard QWERTY keyboard was mounted vertically directly underneath the monitor, and participants responded using three of the keys: top (the A key), bottom (the apostrophe key), and middle (the H key). The top and bottom keys were colored green and purple, and the assignment of colors to keys was counterbalanced across participants. The middle key was always colored white.

Procedure All 96 words were presented one at a time in random order in block 1, and again in a new random order in block 2. Half of the words were in green letters and half in purple letters. The assignment of colors to words was the same for both blocks within-subjects, and counterbalanced between subjects.

Participants began each trial by holding down the middle (white) key with the pointer or middle finger of the dominant hand. A fixation cross appeared for 1000ms-1500ms on a rectangular distribution (to prevent anticipatory releases of the middle key). When the fixation disappeared, a word appeared in the center of the screen for 2000 ms in lowercase, bold 28-point Arial font (purple or green), on a black background. Participants were instructed to release the white key and press the key matching the color of the text as quickly as possible. Only the color of the word was relevant to the response: the word's meaning was irrelevant, and the direction of the response was incidental. But because the purple and green keys were positioned vertically, one above the other, each key press required the participant to make either an upward or a downward movement. After pressing the colored key, participants returned their finger to the white key. Pressing the white key initiated the next trial.

The color of the words was orthogonal to their valence. Therefore, for half of the trials the direction of the correct response was congruent with the valence of the word (e.g., if the word *joy* appeared in green and the green key was on top), and for the other half of the trials direction and valence were incongruent (e.g., if the word *joy* appeared in purple and the purple key was on bottom).

Participants received warning messages, displayed for 2500 ms, if they released the middle key too early (less than 200 ms after word onset) or too late (more than 1000 ms after word onset). Participants performed 16 practice trials prior to the first block. Halfway through each block, they were given a rest, and chose when to continue.

Results and Discussion

Accuracy

Participants pressed the correct button for over 99% of trials. Accuracy did not differ as a function of congruity or block (t-values<1).

Reaction Times

We collected two reaction times: Release Time (measured from the onset of the word to the release of the middle white key), and Press Time (measured from the onset of the word to the press of the colored key). From these we computed Travel Time (Press Time - Release Time). Trials for which Press Time was more than two standard deviations from the participant's mean were excluded from further analysis (143 out of 3840 trials, 3.7%).

Release Times Mean Release Times are given in fig 1a-b. Omnibus 2 × 2 × 2 ANOVAs showed a 3-way interaction of Direction (upward, downward), Valence (positive, negative), and Presentation (first, second), both by subjects ($F_1(1,19)=5.95$, $p=.03$) and by items ($F_2(1,94)=5.83$, $p=.02$). The predicted motor-meaning congruity effect would be indicated by a 2-way interaction of Direction × Valence. There were no significant 2-way interactions in the data from both presentations, combined (all F's<1), so separate 2-way ANOVAs were conducted to test for this effect within each block.

Presentation 1 showed the predicted Direction × Valence interaction ($F_1(1,19)=4.67$, $p=.04$; $F_2(1,94)=3.26$, $p=.07$). Presentation 2 showed a slight trend in the opposite direction, but the Direction × Valence interaction did not approach significance ($F_1(1,19)=1.60$, *ns*; $F_2(1,94)<1$, *ns*).

Press Times Mean Press Times are given in Figure 1c-d. Omnibus 2 × 2 × 2 ANOVAs showed a 3-way interaction of Direction (upward, downward), Valence (positive, negative), and Presentation (first, second), by subjects and by items ($F_1(1,19)=9.17$, $p=.007$; $F_2(1,94)=3.72$, $p=.06$).

Presentation 1 considered alone showed the predicted Direction × Valence interaction ($F_1(1,19)=4.43$, $p=.05$; $F_2(1,94)=3.32$, $p=.07$). Presentation 2 showed a slight trend in the opposite direction, but the Direction × Valence interaction did not approach significance ($F_1(1,19)=2.84$, *ns*; $F_2(1,94)<1$, *ns*).

Overall, there was a strong main effect of direction for Press Times ($F_1(1,19)=131.62$, $p=.0001$; $F_2(1,94)=764.76$, $p=.0001$), which was not present for Release Times. This effect appears to be an artifact of kinematic differences between top and bottom key presses, which used different muscle groups due to the positioning of the keyboard. This main effect is not relevant to the predicted motor-meaning congruity effect.

Travel Times Neither the omnibus 3-way ANOVAs nor the separate 2-way ANOVAs testing relationships between Direction and Valence in Presentation 1 and Presentation 2 showed any interactions that approached significance. This suggests that congruity effects arise during action planning rather than action execution.

In summary, we found the predicted Direction × Valence interaction only during the first presentation of the stimulus words. This motor-meaning congruity effect was absent when words were presented a second time (in Block 2). To test the effect of repetition directly, we compared the magnitude of the congruity effect (incongruent trials -

congruent trials) across blocks, both for Release Times ($t_1(19)=2.46$, $p=.02$; $t_2(95)=2.37$, $p=.02$) and Press Times ($t_1(19)=3.02$, $p=.007$; $t_2(95)=1.95$, $p=.05$). Repetition significantly reduced the effect of congruity between movement direction and valence.

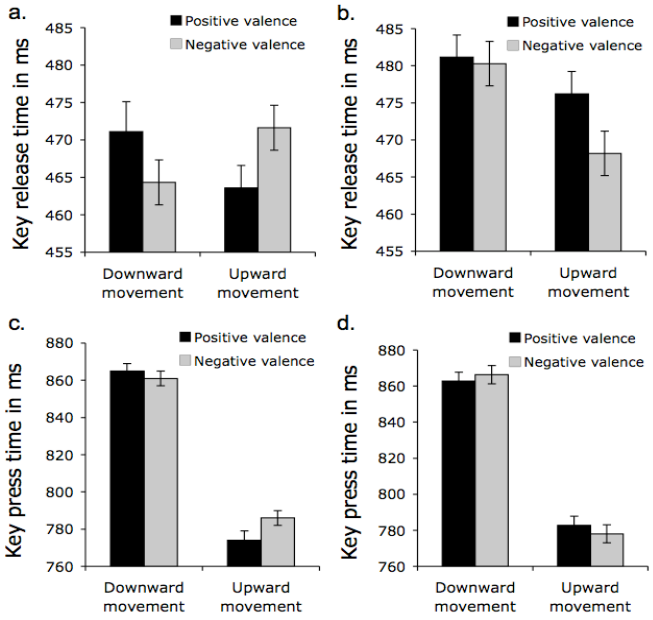


Figure 1. Results of Experiment 1. Top: RT measured from the release of the middle key for Presentation 1 (1a) and Presentation 2 (1b). Bottom: RT measured from the press of the colored key for Presentation 1 (1c) and Presentation 2 (1d). Error bars indicate s.e.m.

Experiment 2: Does attentional orientation modulate motor-meaning congruity effects?

What accounts for the disappearance of the congruity effect when words are repeated? On one possibility, participants may have become so efficient at performing the task that there was no opportunity to detect any interference from irrelevant dimensions of the stimuli: a ceiling effect. Yet an increase in efficiency should result in an overall decrease in reaction times from Presentation 1 to Presentation 2. Since we found no main effect of Presentation, this explanation is not well supported.

Alternatively, it may be that with practice, participants are better able to attend to the relevant dimension of the stimuli (their color) as opposed to irrelevant dimensions (their valence, and more generally their meaning). To test this explanation, for Experiment 2 we adapted Allport & Wylie's (2000) Task Set Inertia paradigm. Target trials were the same as in Experiment 1, but distractor trials were added. For one group of participants, the distractor trials oriented attention toward the meanings of the target words. For the other group, distractors oriented attention toward the target words' colors. We compared reaction times across groups to determine whether attentional orientation modulates the magnitude of space-valence congruity effects.

Methods

Participants Native English-speaking UC Berkeley students (N=48) participated for course credit or payment.

Materials and Procedure

The experimental apparatus for Experiment 2 was the same used in Experiment 1. The primary task was identical to Presentation 1 of Experiment 1, except that 48 distractor trials were added, randomly intermixed with the 96 target trials, for a total of 144 trials. Participants were assigned to perform one of the two versions of the task, one with distractors designed to orient attention to the Meaning of target words, and the other to the Color of target words. Responses to these distractors were not recorded.

Stimuli in the Meaning Orientation condition were 24 concrete nouns, half referring to animate and half to inanimate objects. Whereas target words were shown in purple or green letters, distractors were in white letters. Participants performed a go/no-go animacy judgment, releasing and then re-pressing the middle white button to indicate the distractor word named something animate. In the Color Orientation condition, a 2×2 grid of grey squares appeared. On half of the trials the grid was empty, and on the other half an unsaturated red "X" appeared in one of the squares, balanced across the 4 positions. Participants performed a go/no-go X-detection judgment, re-pressing the middle white button to indicate that a red X was present.

Only one block of trials was performed, and brief rests were provided twice, after the first 48 trials and then after the next 96 trials.

Initially, 16 participants were assigned to each of the distractor conditions. Upon preliminary analyses, the predicted congruity effect was present in the Meaning Orientation condition but not in the Color Orientation condition. Sixteen new participants were added to the Color Orientation condition, to ensure that the absence of a congruity effect was not due to lack of statistical power. Since results for the second cohort did not differ from results in the first, data from both cohorts were combined for the analyses reported here.

Results and Discussion

Accuracy

Participants correctly pressed the button corresponding to the color of the word for 100% of target trials. Performance on distractor trials was not analyzed.

Reaction Times

Omnibus 2 × 2 × 2 ANOVAs showed no significant 3-way interaction of Direction (upward, downward), Valence (positive, negative), and Distractor Type (Meaning, Color). The Press Time data showed the predicted 2-way interaction of Direction and Valence in the Meaning Orientation condition ($F_1(1,15)=6.12$, $p=.03$; $F_2(1,94)=4.23$, $p=.04$), but not in the Color Orientation condition ($F_1(1,31)=.11$, *ns*; $F_2(1,94)=.55$, *ns*). A slight trend toward the same Direction × Valence interaction in the Meaning Orientation condition

was found for Release Times ($F1(1,15)=1.61$, $p=.22$; $F2(1,94)=1.57$, $p=.21$) and Travel Times ($F1(1,15)=4.81$, $p=.05$; $F2(1,94)=.82$, $p=.37$). The absence of a significant effect on Release Times was unexpected, given the results of Expt. 1. This may have been the result of noise introduced into the early phase of target responses when participants were required to task-switch following distractor trials.

To test the predicted effect of attentional orientation on Press Times directly, we compared the magnitude of the congruity effect (incongruent trials - congruent trials) across conditions. According to a Wilcoxon signed rank test, the congruity effect was greater in the Meaning Orientation condition (15.1 ms) than in the Color Orientation condition (1.7 ms; difference of means=13.4 ms, W=176, $p=.04$, one-tailed). Orienting attention toward Meaning or toward Color during distractor trials modulated the size of the motor-meaning congruity effect observed during target trials.

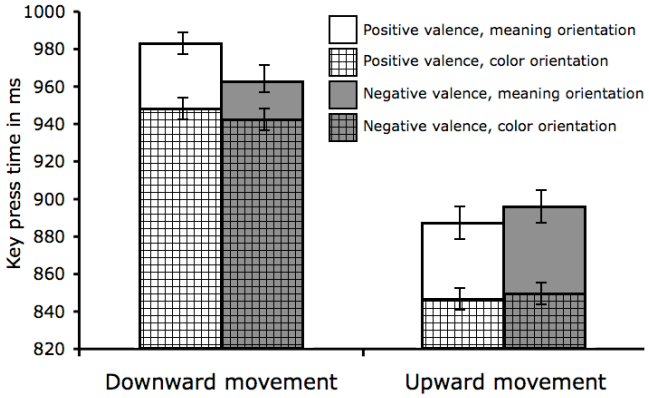


Figure 2. Results of Experiment 2. Space-valence congruity effects were found for target trials when distractors oriented attention to word meaning but not to word color. Error bars indicate s.e.m.

General Discussion

In two experiments, we show effects of congruity between the valence of a word and the spatial direction of the response it cued. In both experiments participants responded only to the color of the target words, pressing the button that matched in color. The spatial directions of the responses were task-irrelevant, as were the meanings of the words. Still, participants responded fastest when the direction of the response and the valence of the word were in agreement: upward movements for positive-valence words, and downward for negative-valence words. The presence of space-valence congruity effects even during shallow, incidental processing of both space and valence suggests that the spatial component of the words' meanings was activated with a high degree of automaticity.

Both experiments also illustrate that automaticity has its limits. In Experiment 1, the motor-meaning congruity effect was found only during the first presentation of the stimuli, but not upon their repetition. Since there was no overall reduction in response times between Presentation 1 and Presentation 2, the extinction of the congruity effect does not appear to be a ceiling effect.

Experiment 2 tested an alternative explanation for the effect of repetition: perhaps with practice, participants became more adept at focusing on the task-relevant dimension of the stimuli (their color) rather than the task-irrelevant dimension (their meaning). Consistent with this proposal, when distractor trials oriented participants to the meaning of the target words, a strong congruity effect was found. By contrast, when distractor trials oriented participants to the color of the target words the congruity effect disappeared.

It is possible to interpret both the repetition effect (in Expt. 1) and the Task Set Inertia effect (in Expt. 2) as effects of attention. During the initial presentation of the words in Expt. 1 and in the Meaning Orientation condition of Expt. 2, participants failed to fully disregard the task-irrelevant meanings of the target words, one component of which is a spatial (or spatio-motor) representation with a certain direction. During the second presentation in Expt. 1 and the Color Orientation condition of Expt. 2, participants more successfully attended to the target words' colors. In Expt. 1, this was because the participants became better at restricting attention to the task-relevant dimension of the stimuli, as a result of practice. In Expt. 2, this was because of attentional 'inertia' from the colored-letter-detection distractor task.

Although this standard interpretation may be valid, there is a potential alternative that does not rely on the construct of attention ("psychology's Weapon of Mass Explanation", according to Vincent Walsh (2003). Implicit in the attentional account is an assumption that reading a word activates *its meaning.* On standard psycholinguistic theories, *the meaning* of a word is retrieved from the mental lexicon, much the way a definition can be looked up in a dictionary. Then attention determines how strongly the word's meaning is activated, and which aspects of the meaning are highlighted.

On alternative accounts of the mental lexicon, however (e.g., Elman, 2004), words don't have meanings; rather, words are cues to activate stored information. The particular constellation of information that gets activated in any instance depends both on the cue, *per se*, and on the context in which the cue is encountered. As a consequence, a word's meaning is unlikely to ever be the same over successive experiences (see James, 1892/2001). 'Meaning', then, is nothing more (or less) than the effect that the word-in-context has on the representations formed in the mind of its reader (or hearer).

On this dynamic view of word meaning, our stimulus words cued the activation of spatio-motor representations in some contexts more than in others. The results of the first block of Expt. 1 suggest that the target words typically cue upward or downward spatio-motor representations such that these representations were activated even though they are irrelevant to the task at hand. But the same words serve as weaker cues for activating such task-irrelevant

representations in contexts where the participant's experience (either with the preceding block of target trials or with the intermixed distractor trials) has adjusted the cue validity of the words' color relative to validity of other pieces of information associated with the words, such as their valence.

Ordinarily, for the words we used as stimuli, valence has high cue validity and the color of the ink has low cue validity: reading that someone is *a hero* is normally a valid cue that the reader should construe the referent positively, regardless of the color *hero* is printed in. But the typical cue validity of words' color and valence is reversed in our tasks, because of the tasks' goals. Seeing a word in green letters is a valid cue that the item should be construed as a member of the category of "up-words" (or "down-words"), regardless of the word's valence or other aspects of its meaning. The weights that participants assign to Color and Meaning as cues, it seems, can be adjusted by the experience of doing the primary task repeatedly, or by the addition of distractor trials that require either color processing or meaning to be processed exclusively.

The present data may be equally consistent with the first proposed account (that words have meanings and attention determines which parts of their meanings get activated) and with the second (that words are cues, and the same cues activate different sets of information depending on the contexts in which they are encountered). Arguably, the second view is preferable on grounds of parsimony: the appearance and disappearance of space-valence congruity effects can be explained based on contextual modulation of retrieval cue weights, alone, rather than on retrieval dynamics *and* the intervention of attention. Distinguishing these accounts definitively will require further experiments.

Conclusions

Some versions of metaphor theory propose that source domain representations are activated automatically when people process words or concepts in target domains (Lakoff & Johnson, 1999). Experimental results have been interpreted as evidence for this automaticity (e.g., Meier & Robinson, 2004). Here we show that, indeed, spatio-motor representations are activated with a surprising degree of automaticity when people read words with positive or negative emotional valence. Space-valence congruity effects are found even when both space and valence are processed shallowly and incidentally.

The present results make clear that automaticity has its limits. The magnitude of space-valence congruity effects was modulated both by repetition of the valenced words and by a Task Set Inertia manipulation (Allport & Wylie, 2000). Spatio-motor representations may be activated by default when people read valenced words, but their activation is also context-dependent. These results are consistent with dynamic views of mental metaphor and of meaning construction, more broadly (Elman, 2004; Evans, 2009; Feldman, 2006).

Acknowledgments

Research was supported in part by a Haas Fellowship to GB and by an NRSA Fellowship #F32MH072502 and a grant from the Spanish Ministry of Education and Science (#SEJ2006-04732/PSIC, DGI) to DC.

References

Allport, A. & Wylie, G. (2000). 'Task-switching', stimulus-response bindings, and negative priming. In S. Monsell & J. S. Driver (Eds.), *Control of cognitive processes: Attention and Performance XVIII*. Cambridge: MIT press.

Barclay, J. R., Bransford, J. D., Franks, J. J., McCarrell, N. S., & Nitsch, K. (1974). Comprehension and semantic flexibility. *Journal of Verbal Learning and Verbal Behavior, 13*, 471–481.

Boroditsky, L. (2000). Metaphoric structuring: understanding time through spatial metaphors. *Cognition*, 75(1), 1-28.

Casasanto, D. (2008). Universal processes generate body-specific representations. *Proceedings of the 30th Annual Conference of the Cognitive Science Society*. Washington, D.C.

Casasanto, D. (2009). Embodiment of abstract concepts: good and bad in right- and left-handers. *Journal of Experimental Psychology: General, 138,* 351-67.

Clark, H. H. (1973). Space, time, semantics and the child. In T. E. Moore (Ed.), *Cognitive Development and the Acquisition of Language*. New York: Academic Press.

Dehaene, S., Bossini, S., & Giraux, P. (1993). The mental representation of parity and number magnitude. *Journal of Experimental Psychology: General, 122,* 371–396.

Elman, L. J. (2004). An alternative view of the mental lexicon. *Trends in Cognitive Sciences, 8*(7), 301-306.

Evans, V. (2009). *How words mean: Lexical concepts, cognitive models and meaning construction.* Oxford, Oxford University Press.

Feldman, J. (2006). *From molecules to metaphor: A neural theory of language.* Cambridge: MIT Press.

James, W. (1892/2001). *Psychology (Briefer Course)*. New York: Dover Publications.

Lakoff, G., & Johnson, M. (1980). The metaphorical structure of the human conceptual system. *Cognitive Science, 4,* 195–208.

Lakoff, G., & Johnson, M. (1999). *Philosophy in the flesh: The embodied mind and its challenge to Western thought*. Chicago: University of Chicago Press.

Meier, B. P. & Robinson, M. D. (2004) Why the sunny side is up: Associations between affect and vertical position. *Psychological Science,* 15, 243-247.

Meier, B. P., Robinson, M. D., Crawford, L. E., & Ahlvers, W. J. (2007). When 'light' and 'dark' thoughts become light and dark responses: Affect biases brightness judgments. *Emotion, 7,* 366-376.

Murphy, G. (1996). On metaphoric representation. *Cognition, 60,* 173–204.

Schubert, T. (2005). Your highness: Vertical positions as perceptual symbols of power. *Journal of Personality and Social Psychology, 89,* 1–21.

Talmy, L. (1988). Force dynamics in language and cognition. *Cognitive Science, 12,* 49–100.

Walsh, V. (2003) Time: the back-door of perception. *TRENDS in Cognitive Sciences, 7,* 335 – 338.

Wittgenstein, L. (1953/2001). *Philosophical Investigations*. Oxford: Blackwell Publishing.

Zwaan, R. A. & Yaxley, R. H. (2003). Spatial iconicity affects semantic relatedness judgments. *Psychonomic Bulletin & Review, 10,* 954–958.

Does micro-variability make models more complex? A comparison between diffusive and linear evidence accumulation

Chris Donkin and Richard M. Shiffrin
Department of Psychological and Brain Sciences, Indiana University
1101 E. 10th St, Bloomington IN 47405 USA

Scott Brown and Andrew Heathcote
School of Psychology, The University of Newcastle,
University Drive, Callaghan NSW 2308 Australia

Abstract

Most theories of how decisions are made assume that the accumulation of evidence from the environment is a noisy process. Recently, models have been proposed which do not have this micro-variability, and as a result are simple in the sense of being analytically tractable. We use a global model analysis method called landscaping to show that in terms of flexibility, simply removing micro-variability does not necessarily make a model more simple. Our landscaping also highlights an experimental design which might be helpful in discriminating between different response models.

Keywords: response time models; complexity; landscaping

A wide range of experimental psychology tasks involve a decision between two alternatives. Which alternative is chosen and the time taken to make that choice has been the subject of intense investigation. The most successful theories for the decision process usually come from a class of evidence accumulation (or sequential sampling) models. Evidence accumulation models assume that participants collect information from the environment to use as evidence as to which potential response is correct. Evidence is accumulated until there is enough to indicate that one of the responses should be given. This response is then made and the time taken for evidence accumulation makes up the decision time component of observed reaction time (RT). Though there are many models which follow this basic framework, the particular assumptions about evidence accumulation that each model makes varies considerably.

Historically, the collection of evidence from the environment has been modeled as a stochastic process (e.g. Ratcliff & Tuerlinckx, 2002; Usher & McClelland, 2001), such that how much evidence there is for a response varies randomly from moment-to-moment. For example, in a random walk process, the amount of evidence accrued between any two moments in time is a sample from a normal distribution.

A small number of recently proposed models, however, have demonstrated that it is not necessary to explicitly model the micro-variability in evidence accumulation (e.g. Reddi & Carpenter, 2000; Reeves, Santhi, & Decaro, 2005). Brown and Heathcote's (2008) Linear Ballistic Accumulator (LBA) model assumes that while a decision is being made, evidence accumulates at a fixed linear rate. Despite this lack of micro-variability the model provides a full account of benchmark choice and response time phenomena.

Brown and Heathcote (2008) proposed the LBA as a simple model of choice and RT because it makes few, and relatively basic, assumptions about how evidence accumulation occurs. Here we investigate a slightly different question, whether or not the LBA, with its lack of micro-variability, is a functionally simpler model. More generally, we aim to examine whether the addition of micro-variability necessarily increases the complexity of a model. Since Occam's Razor says that we should prefer the simplest and complete description of data, and models both with and without micro-variability have been shown to account for empirical data, our investigation may shed light on whether decision models need to necessarily assume micro-variability. We will use a technique called landscaping (Navarro, Pitt, & Myung, 2004) to assess complexity. First, however, we provide an overview of the diffusion and LBA models.

Overview of Models

The Diffusion Model

Consider a recognition memory task in which participants have been asked whether or not a stimulus currently presented was either previously studied, "old", or not studied, "new". A diffusion model account of this choice assumes that participants sample information continuously from the stimulus. Each sample of information counts as evidence for one of the two responses and is used to update an evidence counter, shown by the irregular line in the right panel of Figure 1. Total evidence begins at some starting point and evidence that favors an "old" response decreases the evidence counter and evidence for a "new" response increases the counter. Evidence accumulation continues until the counter reaches one of the response boundaries, the horizontal lines in Figure 1. The choice made depends upon which boundary was reached, the top barrier for "new" and the bottom barrier for "old". The observed RT is the time taken for accumulation plus a non-decision time component made up of things such as encoding time and the time taken to make a motor response.

A key feature of the diffusion model is its micro-variability, such that the amount of evidence accumulated varies from moment-to-moment according to a normal distribution whose mean we call the *drift rate*. On top of this within-trial variability, there are typically three forms of between-trial variability added to the diffusion model. Drift rate and start point

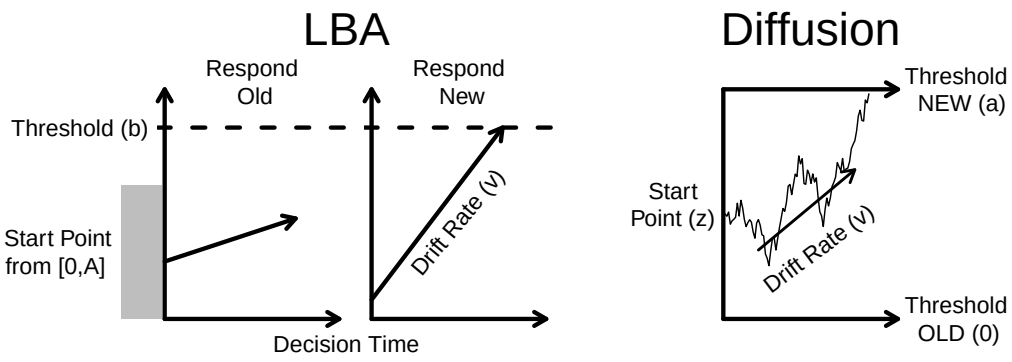


Figure 1: Overview of the diffusion and LBA models (left and right panel, respectively)

are generally assumed to vary from trial-to-trial according to a normal and uniform distribution, respectively. Finally, Ratcliff and Tuerlinckx (2002) included between-trial variability in non-decision time in the form of a uniform distribution.

The LBA Model

In the LBA there are separate accumulators gathering evidence for each of the "new" and "old" responses. As indicated by the straight lines in the left panel of Figure 1, these accumulators accrue evidence linearly and without micro-variability. Accumulation begins at some start point and continues until evidence in one accumulator reaches a response boundary. The accumulator which reaches the boundary first selects its associated response and predicted RT is accumulation time plus non-decision time. As in the diffusion model, the LBA also features between-trial variability. Like the diffusion model, drift rate and start point are assumed to vary between-trials according to normal and uniform distributions, respectively. Unlike the diffusion model, the LBA typically does not require between-trial variability in non-decision time to fit empirical data.

The Complexity of the Models

The LBA was considered by Brown and Heathcote (2008) as a relatively simple model because of its simpler assumptions about variability, and hence fewer parameters. However, recent work has demonstrated that the complexity of a model is not determined simply by the number of parameters in a model, but by how the parameters of the model interact within the model architecture to produce different patterns of predictions – also known as the functional form complexity of a model (e.g. Myung, 2000; Shiffrin, Lee, Wagenmakers, & Kim, 2008). Functional form complexity differs among models when they are able to produce differing ranges of predictions, even when they share the same number of parameters. In this way an overly complex model can provide an excellent fit to data, but not because the model gives a good account of the underlying process, but simply because of the model's flexibility. In particular, a more complex model can "overfit" the data by explaining the noise specific to a particular sample, as well as the structure due to the underlying processes. Because only the structure re-occurs in new data, overfitting limits the model's ability in terms of prediction.

There are many techniques for analyzing the complexity of a model (see Shiffrin et al., 2008 for a review). We will focus on one particular method proposed by Navarro et al. (2004) called landscaping. This method is highly related to parametric bootstrap methods proposed by Wagenmakers, Ratcliff, Gomez, and Iverson (2004). Landscaping, as a means of determining model complexity, is based on the idea that a more flexible model will be better able to mimic the predictions of an alternative model. Landscaping is used to compare the relative flexibility of any two models, and for our purpose these will be models with and without micro-variability (a diffusion and an LBA model, respectively). Note that landscaping tells us about a specific form of local, relative flexibility, rather than the model's general flexibility. In particular, landscaping tells us about how flexible one model is relative to another model, specifically for the regions of the parameter space in which we observe real data. In what follows we will refer exclusively to this local flexibility.

Landscaping

To do landscaping we generate data from one model, say model A, and fit these data with both models, i.e. model A and the alternative model, say model B. We then repeat the process with model B as the data-generating model. How well model B can fit the data generated by model A, and vice versa, gives insight into the relative flexibilities of both models. We will focus on two measures of model flexibility, the first is the difference between how well model B fits model A's data compared to model A, and the second is how often model B can better fit model A's data. The first measure tells us how flexible model B is compared to model A, i.e. if model B gives better fits to data from model A than vice versa, then model B is more flexible. The second measure tells us how distinguishable, or confusable, the two models are, i.e. how

often we expect to have model B fit data better than model A when model A is actually the true model.

In all of our landscaping analyses we simulated 3200 data sets from each model. For each data set a random sample of parameters was chosen from uniform distributions whose ranges were determined by previously observed parameters estimated from real data. In particular, Matzke and Wagenmakers (2009) identified the range of parameter values previously estimated across all previous applications of the diffusion model to data. Donkin, Brown, Heathcote, and Wagenmakers (2009) used these values to identify a range of parameters values for the LBA which spanned the same range of data space. Note that parameters are sampled from each distribution independently and so may not reflect the correlations between parameters in real data.

Table 1: Range of parameter values used to generate data sets. Parameters not previously defined are as follows: T_{er} is non-decision time in both models, s and η represent between-trial standard deviation in drift rate in their respective models, and s_z and s_t represent the ranges of between-trial variability in start point and non-decision time in the diffusion model, respectively.

Model		$b-A$	A	T_{er}	s	v	
LBA	Min	0	.15	.1	.15	.5	
	Max	.5	.45	.4	.35	1	
		a	T_{er}	η	s_z	s_t	v
Diffusion	Min	.06	.3	.01	.01	.01	.01
	Max	.25	.6	.25	.08	.3	.5

Landscaping is known to depend on the design of the data simulated. Here we selected two commonly used designs, one in which only the difficulty of the task was manipulated, and one in which both difficulty and response caution were manipulated. To simulate a difficulty manipulation we used three conditions (easy, medium and hard) across which only the drift rate parameter of the model could change. In practice this meant that the distribution of drift rates shown in Table 1 was divided evenly into three smaller distributions, with the ease of the task increasing with drift rate. To simulate a caution manipulation we used the same procedure to create two conditions (speed emphasis and accuracy emphasis) across which only the response boundary parameter could change, i.e. boundary parameter distributions were divided in two and two values were sampled.

Micro-variability

In these first set of analyses we aim to investigate whether the micro-variability of the diffusion model makes it more flexible than a model without micro-variability, the LBA model. The models, however, differ in more ways than just micro-variability. In an attempt to make the models more similar, and hence make the effect of micro-variability more salient, we use a slightly simplified version of the standard diffusion model (cf. Ratcliff & Tuerlinckx, 2002) in which there is no between-trial variability in non-decision time. The models now share the same assumptions about between-trial variability – it is in both drift rate and start point of accumulation (but see the General Discussion for talk of other key differences between the models).

Difficulty Manipulation To create our landscape we first simulated data from both the LBA and the diffusion model. The data were simulated with all parameters except for drift rate fixed across three difficulty conditions, with 200 observations simulated per condition. We used 200 observations per condition because this amount is standard in applications of choice RT models. Both models used seven parameters for both simulating and fitting data – the diffusion model: a, T_{er}, η, s_z, v_{easy}, v_{medium}, v_{hard}, and the LBA: b, T_{er}, s, A, v_e, v_m and v_h. The simulated data were summarized using five quantiles (.1, .3, .5, .7 and .9) and both models were fit using quantile maximum probability estimation (Heathcote, Brown, & Mewhort, 2002) as the objective function and simplex as a search algorithm.

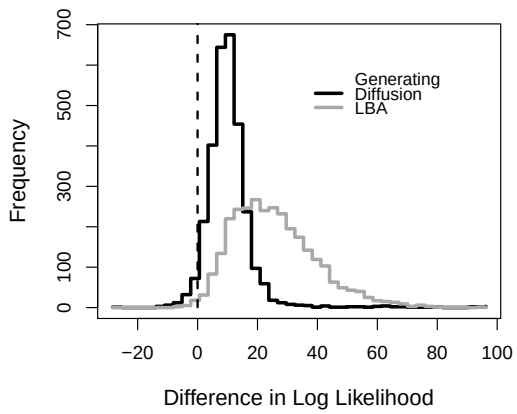


Figure 2: Difference in log-likelihood values between the data-generating model and the alternative model. The black and gray lines represent the diffusion and LBA as the generative models, respectively. The dotted line represents the point at which the data-generating and alternative models give equal quality fits, negative values indicate cases in which the alternative model fits better than the generative model. In this plot the simulated data come from a difficulty manipulation and the models used make the same assumptions about between-trial variability.

Figure 2 shows the difference in quality of fit between the generating and alternative model when the diffusion was the generating model (black histogram) and when the LBA was the generating model (gray histogram). Positive values indicate that the data-generating model fits better than the alternative model, and negative differences indicates that the alternative model is fitting the generating model's data better than the generating model itself. Two things are apparent from

the figure: the gray histogram is generally more positive than the black histogram, and neither histogram has much mass in the region of negative differences. The first observation tells us that when the LBA was the generating model the diffusion tended to fit worse than how well the LBA fit when the diffusion was the generating model. In other words, the diffusion model appears to be less flexible than the LBA model in terms of how closely it can resemble the other model's data. The second observation tells us that neither model is very capable of better fitting the other model's data – the LBA fit data generated from a diffusion model better than the generating model in only 3.2% of the 3200 data sets, and the diffusion model better fit data generated from an LBA only 0.8% of the time.

Visual inspection of the fits suggested that the predictions of both models matched the simulated data closely, regardless of which model the data had come from. Indeed, in all but the most extreme cases, the models appeared to be mimicking each other closely. This suggests that the differences in log-likelihood we observe in Figure 2, and in all other figures, are not simply due to the models occupying completely separate data spaces, but reflect differences in the ability of one model to better fit the other model's data (i.e., what we define as model flexibility).

Caution and Difficulty Manipulations To create the landscape for a design in which both caution and difficulty were manipulated we simulated data in which all parameters except for drift rate were fixed across the three difficulty conditions and all parameters except for response boundary were fixed across the two caution conditions. Fits were as in the previous landscape except that each model now had eight parameters – the diffusion: a_{speed}, $a_{accuracy}$, T_{er}, η, s_z, v_e, v_m, v_h, and the LBA: b_s, b_a, T_{er}, s, A, v_e, v_m and v_h. Landscapes were created using both 200 observations per condition (as in the previous landscape), as well as 100 observations per condition (since twice as many conditions meant that total sample size was twice that of the previous landscape). Sample size had little effect on the pattern of results, but the smaller sample size did lead to slightly more confusion between the models. We present, therefore, the results of the landscape using the smaller sample size (i.e. where total sample size was equated across landscapes).

A quick look at Figure 3 suggests that the current landscape is similar to the one where only difficulty was manipulated. Closer inspection, however, reveals two differences: Firstly, the histograms in Figure 3 show a larger mean and variance than those in Figure 2, and secondly, the histograms show even less mass below zero. The first observation suggests that when both caution and difficulty are manipulated that both models are not as good at accounting for the alternative model's data. Note, however, that the relative position of the black and gray histograms continue to suggest that the diffusion model has less flexibility than the LBA. The second observation implies that the models are even more distinguishable when both caution and difficulty manipulations

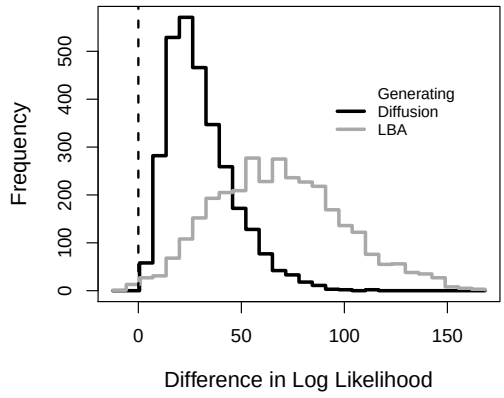


Figure 3: Difference in log-likelihood values between the data-generating model and the alternative model. The data come from a caution and difficulty manipulation, and the models make the same assumptions about between-trial variability.

are made – in 3200 data sets, the LBA never better fit data generated from a diffusion model, while the diffusion model better fit data from an LBA only 0.4% of the time.

Discussion Our first measure of flexibility, the relative shapes and positions of the histograms in our figures, suggest that the LBA is capable of getting better fits of data generated from a diffusion model than vice versa. We take this to mean that the LBA model is more flexible than our simplified version of the diffusion model (i.e. one without non-decision time variability). Since the models were equated on assumptions about between-trial variability, we also take this result as evidence against the idea that the micro-variability in the diffusion model makes the model more flexible than the model without micro-variability, the LBA. Indeed, there may be evidence to suggest the opposite – that micro-variability reduces the functional form complexity of a model. We do not mean our results as conclusive evidence of such a result, however, particularly because micro-variability is not the only difference between the LBA and diffusion models. We direct the reader to our General Discussion for suggestions of how the effects of micro-variability could be more investigated more specifically.

Our second measure of flexibility, how often the alternative model can better fit data from the generating model, gives a less clear result. This is largely because both models seem relatively incapable of better capturing the other model's data, at least for the sample size we use. When we repeated our landscaping analysis with a greatly reduced sample size (just 20 observations per condition) we observed an interesting result, consistent with our first measure of flexibility – the LBA better fit diffusion data in almost one in ten samples, while the diffusion still only better fit LBA data in less than one in two hundred samples. The results reported in Figure 3, however,

suggest that the two models are distinguishable based on fit alone for the types of sample sizes typically used. In other words, in the unlikely case that one of the two models was truly responsible for empirical data, then our results suggest that the alternative model would rarely be mistakenly chosen as the best fitting model, provided at least 100 observations were recorded per condition. However, this result is not very useful since we do not believe that a diffusion model without between-trial variability is appropriate. We now repeat our landscaping using a diffusion model *with* between-trial variability in non-decision time, paying particular focus as to whether or not the models remain distinguishable.

Comparing the LBA and the Full Diffusion

The method for creating the following two landscapes was the same as for the previous two landscapes, however, between-trial variability in non-decision time was assumed for the diffusion model (but not the LBA).

Difficulty Manipulation Figure 4 suggests that a full diffusion model may be slightly more flexible than the LBA when only difficulty is manipulated. In particular, though largely overlapping, the grey histogram looks like a slightly left-shifted version of the black histogram, suggesting that the difference between quality of fit for the data-generating and alternative models was smaller when the LBA generated the data. In other words, the diffusion model was slightly better able to fit LBA data than vice versa. When we look at just the cases in which the alternative model fits better than the data-generating model we see that the same pattern continues, the LBA model better fits data simulated from a diffusion model in 6% of simulated data sets, while the diffusion model better fits LBA data 10% of the time.

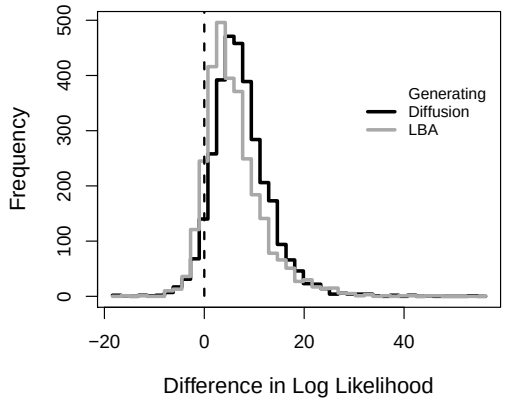


Figure 4: Difference in log-likelihood values between the data-generating model and the alternative model. The data come from a difficulty manipulation, and the diffusion model makes the additional assumption that non-decision time has between-trial variability.

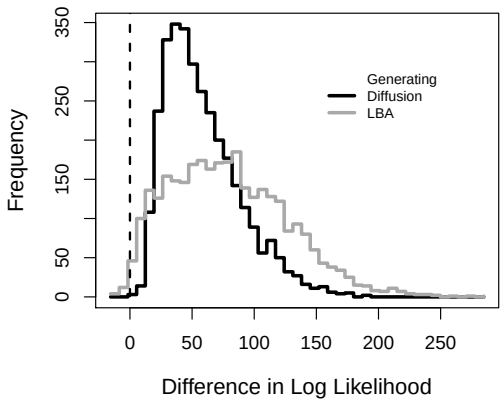


Figure 5: Difference in log-likelihood values between the data-generating model and the alternative model. The data come from a caution and difficulty manipulation, and the diffusion model makes the additional assumption that non-decision time has between-trial variability.

Caution and Difficulty Manipulations The landscape in which both caution and difficulty were manipulated was created using 200 simulated data points in each of the six conditions. From Figure 5 it is not clear which of the full diffusion model or the LBA is more flexible. In particular, the grey histogram has more mass than the black histogram at both very small and very large positive values, suggesting that the diffusion model fit LBA data both very well and very poorly. In terms of how often the alternative model fit better than the data-generating model, when the diffusion model was the generative model then the LBA never fit better, while the diffusion model fit LBA data better in only 0.8% of the simulated data sets.

Discussion The first two landscapes we created suggested that the LBA and the diffusion models were distinguishable, such that each model was relatively incapable of better fitting the other model's data. These second pair of landscapes looked at whether these results extended to the full diffusion model (with between-trial variability in non-decision time). The first landscape we created suggested that this might not be the case. When data came from a design in which only difficulty, i.e. drift rate, varied then both models displayed some reasonable mimicry, such that the LBA looked more like a diffusion model in 6% of the simulated data sets and the diffusion model looked more like an LBA 10% of the time. These proportions are not overly large, but they do suggest that if one of the models actually was the true model, that we would observe the alternative model fitting data better for about one in ten to twenty participants.

The results of our fourth and final landscape suggest that the models become highly distinguishable when both difficulty and caution are manipulated. Indeed, the results suggest that if one of the two models were the true model then the al-

ternative model would be mistaken as the best fitting model for fewer than one in a hundred participants. The difference in distinguishability between these two final landscapes is remarkable, however it is possible that the difference occurs because there are twice as much data under the design with both caution and difficulty manipulations. Equating total sample size using a simplified diffusion model, however, had little effect on distinguishability – doubling sample size meant that the largest confusion occurred 0.4% of the time instead of 0.2%. We expect, therefore, that it is something about the design rather than sample size which causes such a large change in distinguishability. Consistent with this idea, Donkin et al. (2009) showed that the boundary parameters of the LBA and diffusion model do not have a similar effect on model predictions. These results further cement the idea that the key to distinguishing between these two models may lie in the differential effect of manipulating the response boundary parameter in each of the models.

General Discussion

We compared the flexibility of the LBA model, which contains no micro-variability in evidence accumulation, with a simplified version of the diffusion model, which does contain micro-variability. Our results suggest that micro-variability does not necessarily make a model more flexible than one without micro-variability. We can not, however, confidently conclude that micro-variability does not increase flexibility at all. This is because the LBA and the diffusion model, even a simplified version without between-trial variability in non-decision time, do not have identical frameworks. In particular, the LBA has multiple, independent, accumulators while the diffusion has a single accumulator, which implies that evidence for one response is perfectly negatively correlated with evidence for the alternative response. Without further investigation, we can only confidently conclude that a ballistic multiple-accumulator framework gives the LBA more flexibility than a stochastic single-accumulator framework gives the diffusion. Further investigation into the effects of micro-variability might directly compare a multiple accumulator framework with and without micro-variability (e.g. the LBA compared to a simplified version of Usher and McClelland's, 2001, model). Such a study will be more difficult than that carried out here because analytic expressions do not exist for Usher and McClelland's model.

Our final two landscapes compared the relative flexibility of the LBA model and the full Ratcliff diffusion model. Most impressive here was the increase in distinguishability which arose out of the inclusion of a caution manipulation. This increase is quite remarkable, when only difficulty was manipulated the models show the largest overlap of any landscape we analysed, but when caution is added there are almost no cases in which a model fit data better than the data-generating model. The distinction between models may arise because the effect of changing caution (i.e., boundary separation parameters) is different for the two models (Donkin et al., 2009). In particular, micro-variability in processing means that some responses will terminate quickly regardless of the position of response boundaries. Without this micro-variability, however, increasing caution will slow down even the fastest responses. This means that changing caution in the diffusion model effects the speed of the fastest responses much less than in the LBA. Regardless of the cause, our results suggest that whichever model can better account for a combined caution and difficulty manipulation is probably closer to the true model, and unlikely to be due to model mimicry.

References

Brown, S., & Heathcote, A. J. (2008). The simplest complete model of choice reaction time: Linear ballistic accumulation. *Cognitive Psychology*, *57*, 153-178.

Donkin, C., Brown, S., Heathcote, A., & Wagenmakers, E. J. (2009). Diffusion versus linear ballistic accumulation: Different models for response time, same conclusions about psychological mechanisms? *Manuscript submitted for publication*.

Heathcote, A., Brown, S. D., & Mewhort, D. J. K. (2002). Quantile maximum likelihood estimation of response time distributions. *Psychonomic Bulletin & Review*, *9*, 394–401.

Matzke, D., & Wagenmakers, E. J. (2009). Psychological interpretation of exgaussian and shifted wald parameters: A diffusion model analysis. *Psychonomic Bulletin & Review*, *16*, 798–817.

Myung, I. J. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology*, *44*, 190–204.

Navarro, D. J., Pitt, M. A., & Myung, I. J. (2004). Assessing the distinguishability of models and the informativeness of data. *Cognitive Psychology*, *49*, 47-84.

Ratcliff, R., & Tuerlinckx, F. (2002). Estimating parameters of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability. *Psychonomic Bulletin & Review*, *9*, 438–481.

Reddi, B. A. J., & Carpenter, R. H. S. (2000). The influence of urgency on decision time. *Nature Neuroscience*, *3*, 827–830.

Reeves, A., Santhi, N., & Decaro, S. (2005). A random–ray model for speed and accuracy in perceptual experiments. *Spatial Vision*, *18*, 73–83.

Shiffrin, R. M., Lee, M. D., Wagenmakers, E.-J., & Kim, W. J. (2008). A survey of model evaluation approaches with a focus on hierarchical bayesian methods. *Cognitive Science*, *32*, 1248–1284.

Usher, M., & McClelland, J. L. (2001). On the time course of perceptual choice: The leaky competing accumulator model. *Psychological Review*, *108*, 550–592.

Wagenmakers, E.-J., Ratcliff, R., Gomez, P., & Iverson, G. J. (2004). Assessing model mimicry using the parametric bootstrap. *Journal of Mathematical Psychology*, *48*, 28–50.

Self-Organization, Embodied Cognition and the Bounded Rationality Concept

Maria Luísa Bissoto (malubissoto@yahoo.com)
Centro Universitário Salesiano de São Paulo (UNISAL) Americana, SP, Brazil
Faculdade Salesiana Dom Bosco, Piracicaba, SP, Brazil
Post-Graduate Program
"*Apprendere*" - Cognition, Education and Learning Research Group
Prudente de Moraes, 1341, Piracicaba, SP, Brazil

Abstract

This research has aimed to debate some decision making theoretical principles, in particular the bounded rationality concept. The conceptions of cognition subjacent to this concept, its main limitations for understanding the human decision process and the possible contributions from two other theoretical perspectives- the self-organization process and the embodied cognition concept have been methodologically analyzed. The concluding comment claims that: a. other forms to conceive the human cognition are still necessary to better understand the cognitive basis of human making decision process; b. the Self-Organizing and Embodied Cognition Theories, as understood here, might constitute themselves as relevant contributions to the development and reflection concerning the making decision process and the Bounded Rationality concept.

Keywords: decision making; bounded rationality; self-organization; embodied cognition.

Introduction

Electing perspectives for acting and making decisions are critical aspects of the human life and theoretical explanations about these topics could be retraced to the Greek Antiquity. A common characteristic to several of these explanations is concerning the conception that the decision making is essentially a logical-rational process, in which utility principles are prevalent. This conception has became historically stronger and influenced by four sources: the Illuminist thought from XVII century, the probabilistic mathematical theories elaborated from XVIII century, the Utilitarian philosophical-economic theories from XVIII century, and the Neoclassic Economic theories, developed from XIX century on (Buchanan & O'Connell, 1996; Taleb, 2007; Glimcher, Camerer, Fehr & Poldrack, 2009). Since the middle of XX century the approaches that conceive the decision making process under the *expected utility* premises - like those that have been derived and influenced by F. Ramsey, von Neumann e Morgestern (1944), L. Savage (1954) e R. Jeffrey (1956) works – have obtained a wide and important diffusion. These theories have constituted an important scientific advance for understanding the human decision process, presenting for the first time questions about the decision maker preferences, the capacity to order among several variables and rationally choice the *best decisional option*. They have thus created conditions to think more deeply and systematically about the role of the decision maker subjectivity to the decision process. However since the last years from XX century the Expected Utility Theory has been the target of a strong criticism, mainly concerning: a restrictive conception of the human cognition, understanding it basically as an "algorithmic machine", b. the idealization of the decision maker as a super-rational agent, omniscient and omnipotent within the decision process and c. the concept of information on which this theory relays, quantitative and syntactic, making the relation between the informational input and the value attribution to the decisional variables a paradox[1]. These critics have stimulated other theorizations characterized by reconsidering two principal questions: the optimized reason principle and the Maxim Expected Utility principle.

The Bounded Rationality Concept: cognitive presuppositions

H. Simon when contesting the conception of rational/optimized decision considered its substitution for the concept of satisfactory decision; founded on three main issues: a) the human beings are cognitive and perceptually restricted, never being able to fully apprehend the environmental complexity, b) these restrictions impact the decision making process, generating a "cost", compelling the decision maker to find alternative actions that satisfy the decision requirements at other levels ("satisficing" principle) and c) the difficulties and restrictions found in making a decision disclose and clarify its significance, making the process to find satisfactory alternatives adjustable to the decision maker limitations and to the environmental parameters.

These statements sustain the bounded rationality concept developed by H. Simon and formalized in *A Behavioral Model of Rational Choice*, in Models of Man, 1957. In Simon's definition this is the term used "to designed rational choice that takes into account the cognitive limitations of the decision maker- limitations of both knowledge and computational capacity" (1997, p. 291).

[1] See Kahneman & Tversky, 1979; Simon, 1989, Juarrero, 1999; Gigerenzer & Selten, 2001.

The bounded rationality concept is widely used at present and has been employed in several areas "to relax" the strong rational view (decision optimized paradigm) concerning the Classical Decision Theories. Despite that, the comprehension about the *nature* of human cognition upholds the same principles in both approaches, mainly: a. the appeal to the symbolic-normative reasoning, b. the ontological and epistemological gap between subject/object and c. a tight relation among the conception of rationality, truth criteria and a necessary coherence (consistency) between the decision maker and their purposes.

In an article from 1993 A. Vera and H. Simon explained the conception of cognition underlying to the bounded rationality concept. The reasoning process is understood as a sequential and symbolic (internal) computational processing of information1, recursively operating within this basis: input - processing of information - output (reply/behavior). Symbols, on this perspective, are patterns: "(…) when we say that symbols are patterns, we mean that pairs of them can be compared (by one of the system's processes) and pronounced alike or different, and that the system can behave differently, depending on this same/different decision."(Vera&Simon,1993,p.03).

The information/symbols processing is thus conceived: a) inputs are received from the exterior environment as patterns of sensorial stimulations and codified by perceptual processes in symbols; b) these symbols are indexed and stored in the long term memory; c) the elicitation of the meaning denoted by the symbols is made by another symbol, used as input, to get access to a referring object stored in the memory, to affect it or to be affected by it.

It is observed that this conception is, essentially, the same used for computational mind theories[2]; that have not been satisfactorily successful to explain the plasticity and flexibility of the human cognitive – and decision making- process. Simon and Vera (1993) have added contributions from the Behavioral Psychology to this conception of computational mind, relaxing thus some of the hardest cognitivist arguments. This Behaviorist basis can be especially observed when the authors argued for a semantical basis in their conception of computational cognition[3]: the patterns received/perceived for a system are already abstracted, represented and stored with an aggregated meaning. This meaning, not being universal, which would be opposite to the concept of bounded rationality, would follow the material and cultural surroundings in which the (symbolic) system is inserted. In such a way it would –circularly- justify, for example, the different attributions of meaning that exist among/within distinct cultures.

[2] See also Argyris, 1973; Walczack, 1998 e Patokorpi, 2008.

[3] See Rastier, 1996; Floridi, 2004.

Simon had aimed during his academic life to understand *how* human cognition and decision making process really *work*. But does the juxtaposition of elements of two theories (computational mind theory and Behavior Psychology) both of that understanding the cognition and the behavior in terms of causal or functional relations, based in input/output or stimulus/response, appropriately elucidate the human cognitive processes? Or the making decision process? Does the human cognition really "operate" on a computational-representational model? Could the information, at least in the scope of the living beings, still continue to be narrowly understood as a "data flow"?

The concept of bounded rationality, while circumscribing the rationality and decision process limits, situating them as context-time dependents, modified the general comprehension about the human decision-making process. But this "new" focus does not seem yet satisfactory to *really* understand the human cognition; at least not under the behaviorist-rationalist perspectives that remain underlying to it. For a more "realistic" theoretical approach about the decision and cognitive human processes other ontological and epistemological bases are fundamental. We discuss that these bases must be searched in a conception of embodied cognition, which conceives the cognition as a vital self-organizing process.

Cognition: an embodied self-organizing process

The self-organization concept is intended as concerning to the natural process of trends ordering observed in complex systems, both artificial and natural (Debrun, Gonzales & Pessoa, Jr., 1996; Haken, 2000; Piers, Muller & Brent, 2007). It was a term "coined in the 1940s to label processes in which systems become more highly organized over time, without being ordered by outside agents or by external programs" (Shalizi et. al., 2004). A concept strongly attached to this is that of emergence, here understood as the appearance (materialization) of qualities not yet observed in a system from its self-organizing interaction and that cannot be understood by the analysis (on an individual basis) of the relations or elements of the system[4].

Within the scope of this paper both concepts the one about self-organizing process and the one about emergence are relevant for representing by which the organization of a system modifies itself; reaching other levels of complexity. This complexity alteration that enables a system to diversify its surrounding coupling is the definition here conceived to the embodied cognition concept, or "vital cognition". In the core of this conception (and following Hutchins, 1995, Clark, 1997, Zunda, 1999, Wheeler & Clark, 2007, Calvo & Gomila, 2008), the cognitive process is qualified by some undissociated attributes: it's situated, social and

[4] See Bissoto, 2007, 2008; Halley & Winkler, 2007.

distributed. It's observed that within this conception it is not necessary to dichotomously disembody the affective/rational attributes from the human cognition. As embodied beings situated and embedded in a physical circumstantiality, whose comprehension is semantically and socially constructed, reason and emotion/affection are imbricated, mutually influencing themselves, being not possible to disentangle one from the other.

Some theoretical approaches between the bounded rationality and the embodied cognition concepts are possible. The (behavioral) premise of the first concept implies in an interactive relation system/environment and in a situated and embodied action - in the sense that there is a material "body", natural or artificial, acting in a determined time/space. However there is a fundamental difference between both the proposals concerning to *how* the system-environmental interaction occurs.

The bounded rationality concept, when epistemologically considered, can be described as a meaning-sign appropriation one. The system is always acting to apprehend the reality "really" existent in the exterior world, generating diachronically a response/a behavior resulting of the symbolic decoding. When understanding the embodied cognition as a self-organizing movement of a system the main epistemological assumption is that the interrelation system/surrounding is an *interpretative* one. Although it does not discard the assumption of an existent materiality that sustains, displays and sets parameters for the embedding system/environment, there is not in the embodied cognition concept the comprehension that this materiality contains any meaning that could *objectively* be abstracted by a system. According to these considerations signs - and information - are not entities that "carry" an aggregated meaning. They are rather material elements that arise modifications: they provoke the formation of an interpretation, implying in a systemic attribution of value and in changing perspectives for the interacting system. There is not, in this optics, incomplete or badly-structuralized information: everything that can be perceived/selected and meant/interpreted as relevant, from the vital dynamics of each system, comes to be meaningful; guiding the action of this system in the space-time of its surroundings[5].

The making decision process

Decision, within the embodied cognition perspective as here understood, is the *choice* executed by a system concerning to its adaptative efforts. This adaptative process is, by the way, understood as integration, as relational adjustment to the changing enviroment, rather than forced behavior of adequateness. It is a process to make the world meaningful[6] and must address the dynamical self-organizing system "health" and not just looking for a satisfactory or excellent platform of stability. A "good" decision is one that prepares the system to get energy and informational resources, which will lead it to other (richer, in the system's optics) possible organizing openings.

Socially, the decisor's choices still within the embodied cognition perspective are understood as enlargement/disclosure of other interactive horizons which will impulse the enaction of new meaning attributions and therefore the institutions self-organizing vitality; rather than the statement of decisions that aims narrowly a prompt and short-termed efficiency of certain functions of those institutions.

In this scope the bounded rationality concept might be understood considering the perception and action limitations inherent to a determined system in the circumscription of the possibilities for the embedding of this system resulting of order parameters: those boundaries that once emerging from the system/surrounding coupling work as attractors, "forming"/enacting a decision pattern. The analysis of these order parameters by the system itself (or by an observer) can "materialise" for this system an *interpretative* understanding of this dynamics, causing it to be less "evanescent" and allowing the system to disclose which will be the next organizational parameters to be configured, making a dialogue with its trajectory possible.

Understanding the decision making processes under this embodied-self-organizing cognitive perspective is relevant for considering: a. that the "utilities" or preferences of the system are dependents from the historical of interactions system/surroundings already constituted for a system, b. how this system has been successful to perceive other organizing horizons and c. to make possible to think the decision making process as a decentralized one, systemic and surrounding distributed and not circumscribed to the logic-cerebral rationality. The "not-rational", "irrationals" or the "bad" decisions are not conceived as errors but as tied within the organizing perspectives of a system, any evaluation of that could just be thought *a posteriori*, and for the observer's perspective.

The actions assumed for a system (a decision maker) within the incessant informational and energetic flow that this access, are conflicting. Any decision attends to the certain states of the system, ignoring others. There is a momentary "pacification" of the system although a state of "unsated" is always latent, which pressures the system to levels of criticality. From these levels the system could organize itself in new relational situations. But it can be "crystallized" or also "paralyzed" when not obtaining informational and energetic resources that allow it to foment or to choose other organizing routes; which would stimulate the emergence of other systemic configurations. Furthermore a system could reach so high disorder levels

[5] See C.S.Peirce, 1972.

[6] See von Uexkhüll, s/d; Skarda & Freeman, 1987; Johnson, 1999; van Djik et. al., 2008.

(from inherent reasons, like a disease, or from surroundings reasons, like a catastrophic event) that its decision acts would attend just minimal organizational requirements, and its organizational continuity would become impracticable.

Concluding Comments

The decision making theories, including those that follow the bounded rationality principles, have traditionally supported the idea of a decisor agent that controls a data flow – received from the outside- by rational/intellective cognitive process, obtaining therefore a better decisional management. Nevertheless, other decision making process conceptions are possible, mainly when one considers that the decisor system/environmental relation involves more than the creation of syntactic mental models - or, still, as defended by the bounded rationality principles, a (weakly) semantic model.

Kunreuther & Meyer's research (2001) show us information relative to a survey about complex making decisions, which are referents to important aspects of our daily life like health, family, security and financial decisions. The authors analyzed the contrast between how this kind of decision should be made and how they are made (Kunreuther & Meyer, 2001, p. 05). The concluding comments claim that the human complex decision making are characterized for not adequately considering the available information about the probability of an event to occur, fail when differentiating these probabilities, in terms of relevance, attitudes showed for thoughts like "these things will not happen with me"; are strongly influenced by normative social rules, the social *status quo*, the present situation- "I'd better not think about this ever", emotions and affects, failing to learning from other decisional situations.

These remarks also allow us defend the assumption that the research for others ways to comprehend the human making decision process is still a strong scientific requirement. As has been discussed in this paper the embodied concept and the self-organization theory are serious theoretical alternatives for another understanding of the human cognitive and decision making process. The decision making process, as understood here, is closely cohesive within the system/surroundings embedding, in the scope of an embodied and self-organizing cognition.

Some theoretical decision-making perspectives have been incorporating both of the concepts here approached. The Embodied Cognition Theory has been employed in the Consumer Decision-Making Research (Malter, 1996), in the studies on Cognitive-Decision Making (Stewart, 2006) and Neuroeconomics (Hardy-Vallée, 2007) and the Self-Organizing Theory has been thought in the analysis of the collective making-decision (Johnson et. al., 1998) and on the descentralization in (economical) social networks (Roy, Nair & Venema, 2009). Despite of these new branches about the decision making process, the main focus of the Decision-Making field still "targets" the human cognition rational aspects.

In the perspective of this paper further researches could analyze if under this theoretical bases we could increment the human decision process a. debating alternative ways to theorize what is information and its role in the decisional process; b. understanding the cognitive process as "using" the information interpretatively, when this "usage" is nestled in a systemic organization and not relying just on rational capacities and c. widening the interpretative universe of a system, searching to better comprehend *how* the social and distributed cognitive attributes would favor to this system "disclose" organizing alternatives.

References

Argyris, C. (1973). Some Limits of Rational Man Organizational Theory. *Public Administration Review*, Vol. 33, No. 3, 253-267.

Bissoto, M.L. (2007). Auto-organização, cognição corporificada e os princípios da racionalidade limitada. *Ciências & Cognição*; ano 04, v. 11, 80-90.

_____. (2008). Das (im)possibilidades da Relação Auto-organização e Informação: uma perspectiva de análise. *Auto-Organização Estudos Interdisciplinares*, col. CLE, vol. 52, 33-58.

Buchanam, L. & O'Connell, A. (1996). A Brief History of Decision Making. *Harvard Business Review*, Boston, v. 74, n. 06, 61-78.

Calvo, P. & Gomila, T. Handbook of Cognitive Science: an Embodied Approach. Amsterdan: Elsevier.

Clark, A. (1997). *Being There:* Putting Brain, Body and World Together Again. Cambridge MA: MIT Press.

Debrun, M; Gonzales, M.E.Q.; Pessoa Jr., O. (1996) *Auto-organização*: estudos interdisciplinares em filosofia, ciências naturais e humanas, e artes, Campinas/SP: Unicamp, Centro de Lógica, Epistemologia e História da Ciência, v.18.

Floridi, L. (2004). Open Problems in the Philosophy of Information. *Metaphilosophy*, 35, 4, 554-582.

Gershenson, C. & Heylighen, F. (2003). When Can We Call a System Self-Organizing? *7th European Conference on Advances in Artificial Life* (pp. 606–614). Dortmund, Germany: Springer.

Gigerenzer, G.; Selten, R. **(**2001). *Bounded Rationality.* Cambridge/MA: MIT Press.

Glimcher, P.; Camerer, C.; Fehr, E.; Poldrak, R. (2009). *Neuroeconomics:* Decicion Making and the Brain. Elsevier.

Halley, J. D., & Winkler, D. A. (2008). Classification of emergence and its relation to self-organization. *Complexity*, 13(5), 10–15.

Haken, H. (2000). *Information and self-organization:* A Macroscopic Approach to Complex Systems. Berlin: Springer-Verlag.

Hardy-Vallée, B. (Ed.). (2007). *Cognitive Decision-Making:* Empirical and Foundational Issues. Newcastle, U.K.: Cambridge Scholars Publishing.

Hutchins, E. (1995). *Cognition in the Wild*. Cambridge/MA: MIT Press.

Johnson, M.L. (1999). Embodiment Reason. In G. Weiss & H. F. Haber (eds) *Perspectives on Embodiment:* The Intersections of Nature and Culture. UK: Routledge, 1999.

Johnson, N., Rasmussen, S., Joslyn, C., Rocha, L., Smith, S. & Kantor, M. (1998). "Symbiotic intelligence: self organizing knowledge on distributed networks, driven by human interaction." In: *6th International Conference on Artificial Life*. C. Adami, et al. (Eds.). MIT Press.

Juarrero, A. (1999). *Dynamics in Action:* Intentional Behavior as a Complex System. Cambridge: MITpress.

Kahneman, D.; Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk, *Econometrica*, 47, 263-291.

Kunda, Z. (1999). *Social Cognition:* Making Sense of People. Cambridge/MA: MIT Press.

Kunreuther, H., R. Meyer, R. Zeckhauser, P. Slovic, B. Schwartz, C. Schade, M. F. Luce, S. Lippman, D. Krantz, B. Kahn and R. Hogarth. (2002). High Stakes Decision Making: Normative, Descriptive and Prescriptive Considerations. *Marketing Letters*, 13(3): 259-278.

Malter, A. (1996). An Introdution to Embodied Cognition: Implications for Consumer Research, *Advances in Consumer Research*, v. 23, Provo: Association for Consumer Research, 272-276.

Newell, A. *Human Problem Solving*. USA: Prentice Hall, 1972.

Patokorpi, E. (2008). Simon's paradox: Bounded rationality and the computer metaphor of the mind. *Human Systems Management*, IOS Press, v. 27, n 04, 285-294.

Peirce Edition Project. (1992). *The Essential Peirce***.** (vol.1). Indianapolis, Indiana, USA. (original manuscript 1867-1893).

Piers, C., Muller, J., Brent, J. Self-Organizing Complexity in Psychological Systems. (2007). *Psychological Issues*, 67. Lahan MD: Jason Aronson.

Rastier, F. (1996). Problématiques du signe et du texte. *Intellectica*, 2, 23, pp. 11-52.

Roy, D., Nair, S., Venema, H. (2009). Enabling Self-Organization and Social Networking. In Swanson, D. & Badhal, S. *Creating Adaptive Policies:* A Guide for Policy-Making in an Uncertain World. Canada: The International Development Research Centre.

Shalizi, C. R., Shalizi, K. L., Haslinger, R. (2004). Quantifying self-organization with optimal predictors, *Physical Review Letters*, vol. 93, no. 11, pp. 1–4.

Simon, H. A. (1989). *A Razão nas Coisas Humanas*. Trad. M.G. Segurado. Lisboa: Gradiva Publicações.

Simon, H. (1997). *Models of Bounded Rationality***.** vol. 3. Cambridge/MA: MIT Press.

Skarda, C.A., Freeman W. J. (1987). How brains make chaos in order to make sense of the world. *Behavioral and Brain Sciences*, 10, 161-195.

Stewart, T. (2007). Embodied Decisions: Models of Decision Making Within a Larger Cognitive Framework. In Hardy-Vallèe, B. *Cognitive Decision Making:* Empirical and Foundational Issues. Newcastle, U.K.: Cambridge Scholars Publishing.

Taleb, N. *The Black Swan*: The Impact of the Highly Improbable. NY: Random House, 2007.

van Dijk, J.; Kerkhofs, R.; van Rooij, I., Haselager, P. (2008). Can There Be Such a Thing as Embodied Embedded Cognitive Neuroscience? *Theory Psychology*, 18, 297-318.

Vera, A.; Simon, H. (1993). Situated action: a Symbolic Interpretation. *Cognitive Science***,** 17, 7-48.

von Uexkhüll, J. *Dos animais e dos homens*. Lisboa: Livros do Brasil, s/d.

Walczack, S. (1998). Neural network models for a resource allocation problem, *IEEE Transactions on Systems, Man, and Cybernetics*, Part B 28(2), 276-284.

Wheeler, M. & Clark, A. (2008). Culture, embodiment and genes: unravelling the triple helix. *Philosophical Transactions of the Royal Society B* 363(1509): 3563-3575.

The Role of Inhibition in Theory of Mind Performance: Evidence for a Non-Modular View of Theory of Mind

Lindsey Frederixon Byom (ljfrederixon@wisc.edu)
Department of Communicative Disorders, 1975 Willow Drive
Madison, WI 53706 USA

Margarita Kaushanskaya (kaushanskaya@wisc.edu)
Department of Communicative Disorders, 1975 Willow Drive
Madison, WI 53706 USA

Abstract

As the modularity of Theory of Mind continues to be debated, the present study sought to investigate the relationship between inhibitory control and performance on a linguistic Theory of Mind (ToM) task. Performance on ToM tasks that relied on inhibitory control was contrasted with performance on ToM tasks that did not rely on inhibitory control. In addition, a range of executive function tasks were administered to all participants. It was hypothesized that if Theory of Mind shares resources with the executive process of inhibition, performance on the ToM task would diminish when inhibition demands were high. Results indicated that performance on the ToM task was significantly lower when participants were required to inhibit superficial discrepancies in the Theory of Mind stories. Moreover, performance on the ToM task correlated with the ability to resist non-linguistic interference. These findings challenge the modular views of Theory of Mind, and suggest that Theory of Mind and executive functions may rely on common cognitive resources.

Keywords: Theory of Mind; inhibition; executive function; modularity.

Introduction

Human beings, as the most social of primates, rely heavily on complex social knowledge or social cognition (Adolphs, 1999). Social cognition has been described by Adolphs (1999) as "the processes that subserve behavior in response to other individuals of the same species…especially those higher cognitive processes subserving the extremely diverse and flexible social behaviors that are seen in primates." Social cognition, as a high-order cognitive process, relies on several cognitive functions for appropriate interaction including goal-directed planning, emotional control and recognition, arousal, vigilance, and memory integration, (Adolphs, 2009). One aspect of social cognition, Theory of Mind (ToM), allows humans to understand that others have mental states and to use reason about these mental states in order to predict the behavior of others (Fletcher, et al., 1995; Frith & Frith, 1999). While the importance of the Theory of Mind for successful social functioning is not debated, there is controversy surrounding the degree to which the development and use of Theory of Mind relies on domain-general cognitive processes. A modular approach views Theory of Mind as a specific and independent cognitive module (Frith & Frith, 1999). A domain-general approach sees Theory of Mind as a skill that relies on executive function (EF) – a collection of complex cognitive processes that includes inhibitory control (or the ability to resist interference), updating (or working memory), and task switching (e.g., Miyake et al., 2000; Smith & Jonides, 1999; Zelazo, Craik, & Booth, 2004).

The modular view is supported by clinical data, with certain clinical impairments, such as autism or schizophrenia, characterized by marked difficulty with Theory of Mind tasks in the face of relatively spared intellect (e.g., Frith & Frith, 1999; Happé, 1994). The modularity of Theory of Mind has also been evaluated in individuals with traumatic brain injury (TBI), a population commonly found to demonstrate impairments in Theory of Mind (Bibby & McDonald, 2005; Channon, Pellijeff, & Rule, 2005; Turkstra, Dixon, & Baker, 2004). For instance, Bibby and McDonald (2005) found that individuals with TBI exhibited impairments in Theory of Mind tasks and that these difficulties could not be accounted for by inference abilities or language skills. As further support for the modular view, functional neuroimaging techniques have identified specific brain regions that are selectively active during Theory of Mind tasks (e.g., Adolphs, 2009; Amodio & Frith, 2006; Fletcher et al., 1995).

In contrast to domain-specific views of Theory of Mind, it appears that other cognitive domains, most notably the executive functions (Leslie, German, & Polizzi, 2005; McKinnon & Moscovitch, 2007), may influence Theory of Mind performance. For example, McKinnon and Moscovitch (2007) used a dual-task experiment to examine whether Theory of Mind and working memory relied on the same resources. Dual-task experiments assume that when two cognitive processes compete for shared resources, performance will be diminished for one or both tasks. McKinnon and Moscovitch (2007) found that adults' performance on a Theory of Mind task was significantly worse when participants were required to simultaneously perform a working memory task. This finding suggested that both the working memory and Theory of Mind tasks rely on common cognitive resources. In another dual-task study, Theory of Mind performance was shown to decrease in both older and younger adults when the executive function demand of the task was increased by requiring participants to also reason about approach or avoidance

beliefs (German & Hehman, 2006). Similarly, Carlson, Moses, & Claxton (2004) found that children's performance on inhibitory tasks was significantly related to their performance on Theory of Mind tasks.

The current study sought to evaluate the relationship between adults' performance on a linguistic Theory of Mind Task and inhibitory control – a component of the executive function. Previous work on the relationship between ToM and executive function either focused on very young children (e.g., Leslie et al., 2004), used dual-task methodology (e.g., McKinnon & Moscovitch, 2007), or used a purely correlational approach (e.g., Carlson & Moses, 2001). In the current study, the relationship between inhibitory control and ToM was examined in two ways. First, we manipulated the executive function demands of the ToM task itself. Second, we administered a battery of Executive Function tasks to all the participants, being particularly careful to index inhibitory control in both the linguistic and the non-linguistic domain.

The ToM task designed for the current study presented participants with pairs of short stories that either matched or mismatched in the ToM Structure (see Table 1 for examples of stories). Stories described human behavior that required understanding of people's intentions and beliefs (including engaging in white lie, using sarcasm, etc.). Stories that matched in ToM structure described situations that shared the underlying intent (e.g., both were stories about white lies). Stories that mismatched in ToM Structure described situations that diverged in the underlying intent (e.g., one story was about a while lie and another story was about sarcasm). Participants were asked to make same/different judgments on pairs of stories that either matched or mismatched in ToM structure based on the underlying intentions of the story characters. The key manipulation involved the Surface Structure of the stories. Half of the stories matched in superficial contextual elements (characters had the same names, actions took place in the same location, etc.), while half of the stories mismatched in superficial contextual elements (characters had different names, the actions took place in different locations, etc.). This design yielded four conditions: Stories that matched in both the ToM Structure and Surface Structure; stories that matched in ToM Structure but differed in Surface Structure; stories that matched in Surface Structure but differed in ToM Structure; and stories that mismatched in both the ToM and the Surface Structures. The logic was that making a "same" decision on stories that matched in ToM but that mismatched in Surface Structure would require inhibition of attention to superficial discrepancies. Similarly, making a "different" decision on stories that mismatched in ToM but that matched in Surface Structure would require inhibition of attention to superficial similarities. Conversely, performance on stories that either both matched or both mismatched in ToM and Surface Structure would not require inhibitory control. We hypothesized that if performance on ToM tasks relies on executive function, then participants should be less accurate and slower making judgments of similarity on the conflicting stories than on the non-conflicting stories, since performance on conflicting stories would require inhibitory control. If, on the other hand, performance on ToM tasks relies on domain-specific mechanisms that are separable from executive function mechanisms, then participants should show similar performance on conflicting and non-conflicting stories.

In addition to embedding inhibitory-control manipulation within the ToM task itself, we also examined the relationship between executive function and Theory of Mind by administering a range of executive function tasks to the participants. We hypothesized that if ToM relies on executive function, then adults' performance on the ToM task would correlate with performance on executive function measures. Since executive function is a complex construct that subsumes a number of dimensions, finding that performance on the ToM task correlates with some executive function measures, but not others would be informative with regards to the specific executive function mechanisms that may underlie performance on Theory of Mind Tasks in adulthood. Because the ToM task designed for the current study was linguistic in nature, we were especially interested in examining the relationship between ToM performance and performance on linguistic vs. non-linguistic executive function tasks.

In summary, the goal of the present study was to examine the modularity of Theory of Mind in adults by testing the relationship between ToM and inhibitory control. We theorized that if performance on ToM tasks requires inhibitory control, then participants should perform less well on ToM tasks that place increased demands on the inhibitory control mechanism. We also theorized that if performance on the ToM tasks is related to executive function, then measures of executive function should correlate with ToM performance.

Methods and Procedures

Participants

Twenty-two participants were recruited for this study. Participants ranged in age from 18.9 to 22.8 years, and all were native speakers of English. Each participant scored within the normal range on English receptive vocabulary as measured by the Peabody Picture Vocabulary Test-III (Dunn & Dunn, 1997) (*Mean* = 108.57; *SD* = 7.00), and on reading ability as measured by the Reading Fluency subtest of the WJ III Tests of Achievement, (Woodcock, McGrew, & Mather, 2001a) (*Mean* =110.41 *SD* = 8.28). Participants also scored in the normal range on the non-verbal intelligence measure (Visual Matrixes subtest of the Kaufman Brief Intelligence Test, Second Edition, K-BIT2; Kaufman & Kaufman, 2004) (*Mean* = 101.76, *SD* = 11.87).

Materials and Procedure

Each participant was tested in one two-hour session. Theory of Mind tasks, executive function tasks, language ability tasks, and a non-verbal IQ test were administered in random order.

Table 1: Examples of Stimuli in Four ToM Conditions

	ToM Match		ToM Mismatch	
	Story A	Story B	Story A	Story B
Surface Match	Ann and her husband left their home for work on a gloomy, rainy day. Ann said, "What a bright cheery day."	Ann and her husband left their home for work on a gloomy, rainy day. Ann said, "It's a good thing I packed my sunglasses today."	Dan attempted to cook dinner for his sister's birthday, but burnt everything to a crisp. His sister's friend Kristy said, "You're quite the chef, Dan."	Dan attempted to cook dinner for his sister's birthday. His sister's friend Kristy said, "Your dinner was great, I just wasn't very hungry tonight."
Surface Mismatch	Jacob's history professor assigned six chapters of reading for the following day's class. On the way out of class, Jacob said to his friends, "We'll have plenty of free time tonight, huh?"	Joan had to stay late at work for the next week while her boss was out of town. Her coworker John said, "Aren't you the lucky one this week."	Ben and Ryan were walking to class when Ben said, "Did you see John's shoes at track practice today? They were awful." Just then Ryan turned around and said, "Oh hi John, I didn't see that you were behind us."	Karen really wanted to try out a new café in town but didn't want to go alone. Karen said to her best friend, Joanne, "I'll probably try out that new café, but I suppose I'll have to go alone since no one will go with me."

Theory of Mind Task The ToM task in this study evaluated participants' ability to identify Theory of Mind inferences in the face of varying inhibitory control demands. The ToM task presented participants with 40 pairs of short stories to be read silently from a computer screen (see Table 1). Stories ranged from two to four sentences in length and were constructed using vocabulary and syntax at the sixth grade level. Five story types, each requiring Theory of Mind for accurate interpretation were included in the task. Story types included white lie, deception, faux pas, sarcasm, and persuasion. Participants were not informed of the story types. Executive function demands were manipulated through variation in the Surface Structure or context of each story. In the low executive function conditions, the story context and the ToM inference were both similar or both different across stories. In the high executive function conditions, the stories either shared the ToM inference, but differed with regards to story structure or shared story structure, but differed with regards to the ToM inference. Participants first completed two practice trials, each of which was followed by an explanation of the correct response. The order of presentation of story pairs was randomized. Participants first saw a screen with only a black vertical line bisecting the screen at the midline, and then were presented with one story (Story A) on the left side of the screen. After reading the story, participants pressed the space bar and Story A disappeared and Story B was presented on the right side of the screen. After reading story B, the participant again pressed the space bar and both stories appeared on the screen, separated by the vertical black line. This procedure was implemented in order to minimize the effect of reading times and of working memory demands on ToM performance. While both stories were available on the screen for review, the participant chose whether the stories required the same inference, (e.g., both stories included a faux pas situation) or if they required different inferences, (e.g., one story included faux pas and one demonstrated sarcasm). Participants indicated their decision by pressing the forward slash key if the inferences were the same across stories, or the "z" key if the inferences were different across stories. Both accuracy and reaction times were recorded. Reaction time measurements began as both stories were presented simultaneously and ended as soon as a decision key was pressed. Participants were instructed to respond to each stimulus as quickly as possible while maintaining response accuracy.

Executive Function Tasks Tasks measuring distinct EF components were administered to each participant. *Linguistic inhibitory control* was measured via the Color-Word Interference Task (a version of the Stroop task, Stroop, 1935), where participants were asked to name ink colors and inhibit the more automatic processing of print (Delis, et al., 2001). *Non-linguistic inhibitory control* was measured via the Simon task (Simon & Rudell, 1967), where participants were presented with either red or green circles in the center, or on either the left or the right side of the computer screen and were required to press a left key when they saw a green circle and a right key when they saw the red circle. On incongruent trials, the location of the colored circle conflicted with the response key, and participants had to inhibit the automatic spatially-based response. *Complex problem solving and planning* were measured via The Towers Task (Delis, et al., 2001), where participants were presented with disks of different sizes and

three rods, and were required to achieve the target arrangement of disks on rods in as few moves as possible. *Working Memory* was measured using the Numbers Reversed subtest of the Woodcock Johnson III Tests of Cognitive Abilities, (Woodcock, McGrew, & Mather, 2001b), where participants heard increasingly-long sequences of digits, and were required to repeat each sequence in the reverse order.

Results

Accuracy and Reaction Time data were analyzed using 2 x 2 Repeated-Measures ANOVAs, with ToM Structure (matching vs. mismatching) and Surface Structure (matching vs. mismatching) as within-subjects independent variables. A-priori follow-up paired-samples comparisons were conducted to examine (1) whether surface mismatch impaired participants' ability to identify similar ToM structure in the stories, and (2) whether surface match impaired participants' ability to differentiate distinct ToM structures in the stories. Finally, correlation analyses were conducted to examine the relationship between ToM performance and Executive Function measures. Here, we were especially interested in comparing the relationships between ToM performance and language-based EF tasks and the relationship between ToM performance and non-linguistic EF tasks.

For accuracy, a 2 x 2 ANOVA with ToM Structure and Surface Structure as within-subjects Independent Variables yielded a marginally-significant interaction between the two independent variables, F (1, 20) = 3.23, p = 0.09, η_p^2 = 0.14. The interaction of ToM and Surface Structure variables indicates that performance on the ToM task was influenced by superficial contextual information, and suggests that the inhibitory-control demands mediated ToM performance. For RTs, a similar analysis yielded a main effect of ToM structure, F (1, 19) = 8.10, p = 0.01, η_p^2 = 0.30, indicating that conditions in which ToM matched required shorter response time than those in which ToM differed.

A-priori pair-wise t-tests were conducted to determine whether accuracies and reaction times differed across high and low inhibitory control conditions. For accuracy, a significant difference was observed between performance on stories where both ToM and Surface Structure matched and stories where ToM matched but Surface Structure mismatched, t (20) = 2.03, p = 0.06. However, there were no differences in performance on stories where ToM and Surface Structure both matched and stories where ToM mismatched but Surface Structure matched, t (20) = 0.83, p = 0.42. Figure 1 displays average accuracy for each condition.

For RTs, a significant difference was observed between performance on the condition in which both ToM and Surface Structure matched, and the condition in which ToM Structure differed but Surface Structure Matched, t (19) = -2.21, p = .04. Figure 2 displays average RTs for each condition.

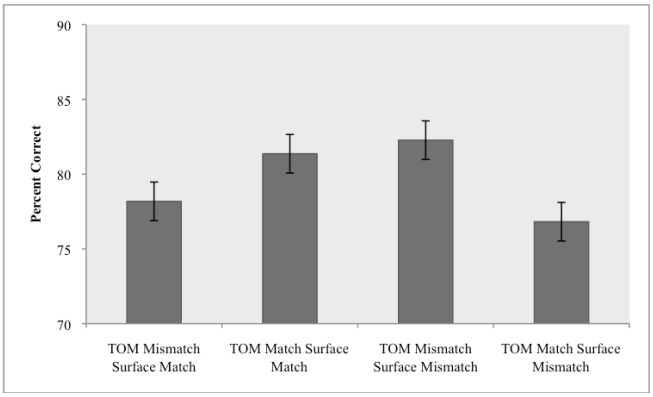


Figure 1: Performance Accuracy on ToM Task

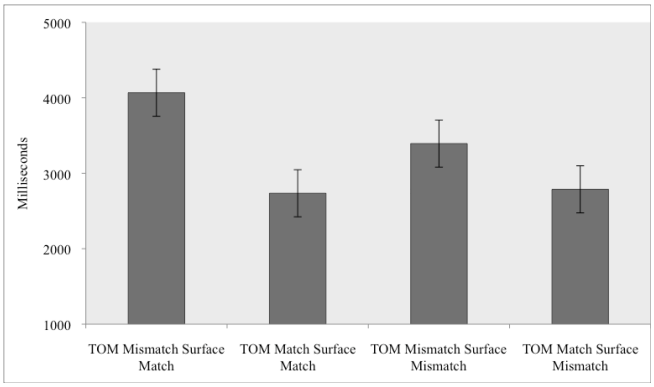


Figure 2: Performance RTs on ToM Task

Pearson Correlation analyses were conducted to examine the relationship between participants' accuracy and RTs on the ToM task and their performance on the executive function tasks.

Performance on the Digits Reversed task did not correlate with any of the performance measures. This indicates that the ToM task in the current study did not tax working memory capacity. Performance on the Stroop task was positively correlated with accuracy on the ToM task where both ToM and Surface Structure differed (r = 0.45). The finding that only one condition of the ToM task correlated with Stroop performance was unexpected, especially because this ToM condition did not require inhibition. It may be that the lack of association was due to the fact that the Stroop task demanded inhibition of an irrelevant perceptual dimension (conflicting color word) while the ToM task required inhibition of an irrelevant response dimension (response based on surface structure). It may also be that the lack of association was due to differences in response modality across the two tasks, with the Stroop task requiring vocal responses, and the ToM task requiring button-press responses.

Unlike the Stroop findings, performance on the Simon task was associated with ToM performance. To measure non-linguistic inhibitory control, a difference score was calculated where participants' RTs on the incongruent Simon trials (requiring inhibition) were subtracted from the

neutral Simon trials. Small difference scores indexed successful conflict resolution, and thus, superior inhibitory control. This measure of conflict resolution was correlated with ToM performance in each of the four conditions, and only one analysis yielded a significant correlation. Namely, successful conflict resolution on the Simon task was associated with higher accuracy on the ToM task in a condition where two stories shared the underlying ToM structure but diverged in Surface Structure (r = -0.51, p = 0.02). This finding suggests that non-linguistic inhibitory control was associated with ToM performance only in a condition where participants had to select the "match" response and inhibit the "mismatch" response based on non-overlapping superficial structural characteristics.

Interestingly, accuracy scores on the Towers Task were inversely correlated with accuracy on the ToM condition in which both ToM and Surface Structure matched (r=-0.55), and the total time taken to complete the Towers Task was inversely correlated with RTs for all conditions in the ToM task (correlation coefficients ranged from -0.51 to -0.58). It is difficult to interpret these correlations since it is unclear what cognitive abilities the Towers Task indexes. In our data, performance on the Tower task did not correlate with any other executive-function measure, indicating that the skill(s) it was indexing may not have been related to executive function. It is possible that these inverse relationships between performance on the Tower task and performance on the ToM task are due to the different modalities tapped by each task: visuospatial in the Towers and linguistic in the TOM task.

Discussion

Questions regarding the degree to which higher-order cognitive tasks rely on domain-general processes permeate every aspect of cognitive science. The goal of the current study was to inform the debate surrounding the modularity of Theory of Mind by examining the relationship between performance on the Theory of Mind task and inhibitory control. The results indicated that performance on the linguistic ToM task was associated with inhibitory control function. This conclusion was supported by three main findings.

First, higher accuracy was observed on the condition with similar ToM and Surface Structure compared to the condition with similar ToM Structure but differing Surface Structure. Because the condition with divergent Surface Structure required more inhibitory control than the matching condition, we interpret this pattern of results to suggest that ToM and inhibitory control draw on common cognitive resources.

Second, participants were significantly quicker to respond to trials in which both ToM and Surface Structure were similar than when the ToM Structure differed, but the Surface Structure was similar. This finding suggests that the inhibitory-control demands imposed by the incongruent ToM and Surface Structure resulted in prolonged response times.

Finally, performance on an executive function measure as assessed by the Simon task correlated with ToM performance, particularly for the condition where participants had to detect matching ToM across two structurally-distinct stories. This finding suggests that performance on the linguistic ToM task (especially one that involved inhibition) was associated with performance on the non-linguistic inhibitory-control task.

While this study included a small sample size and all participants scored very high on measures of receptive English vocabulary, the findings of a link between ToM and executive function support the non-modular view of the Theory of Mind (e.g., Carlson, et al., 2004; McKinnon & Moscovitch, 2007). It appears that Theory of Mind performance in adulthood may in fact draw on the same complex cognitive processes as inhibitory control. However, the findings are also consistent with the view of Theory of Mind proposed by Leslie, Friedman, & German (2004). Leslie et al. (2004) argued that Theory of Mind is comprised of an innate, modular 'Theory of Mind mechanism' that generates alternate interpretations of social situations, and an executive selection process that chooses one interpretation from those suggested by the Theory of Mind mechanism. According to this view, the selection process is inhibitory in nature. This theory has been tested previously using a false belief task, in which the participant must correctly identify that a character in the task has a belief that is different from the actual state of reality (Leslie, et al., 2004; Leslie, et al., 2005). In the case of a false-belief task, Leslie (2004) argued that the Theory of Mind mechanism generates several possible beliefs with the reality of the situation being the default selection. In a false-belief task, however, because the character's belief is false, the selection process must inhibit the default interpretation in favor of a belief that is different than the reality of the situation.

When considering the findings of the present study, it could be argued that in the condition in which the ToM and surface structure are incongruent, the irrelevant surface structure information must be inhibited in favor of the deeper ToM structure. Therefore, the present data may in fact support the view of Theory of Mind that construes performance on ToM tasks as a process that consists of mechanisms specific to the Theory of Mind, and domain-general inhibitory control mechanisms.

Whatever the interpretation of the findings, it is intriguing that only one ToM condition was taxing for the participants – the condition with similar ToM and differing surface structure. The opposite condition, in which the ToM structures differed, but the surface structure matched did not seem to incur higher inhibitory demands. Perhaps suppressing a "no" response requires more inhibitory control than suppressing a "yes" response, although it is unclear why this may be so. The degree to which different ToM tasks require inhibitory control is therefore a crucial area of further research.

This study provides evidence that performance on Theory of Mind tasks may rely on domain-general inhibitory control mechanisms, and more broadly provides insight into the non-modularity of processes associated with high-order cognition. It is possible to increase inhibitory-control demands of the ToM task by pitting similarities in the underlying intentional structure of the stories (ToM) against superficial similarities in the linguistic structure of the stories. Requiring participants to make decisions about ToM similarities while ignoring structural differences imposes inhibition demands on performance. Crucially, performance on the linguistic ToM task correlated most highly with a measure of non-linguistic inhibitory control, pointing to an association between ToM and executive function in particular, and linguistic and non-linguistic performance in general. This pattern is in line with non-modular views of Theory of Mind, that construe performance on social cognition tasks as drawing on the same basic cognitive mechanisms that underlie performance on complex planning tasks, i.e., executive function.

Acknowledgments

The authors thank Dr. Lyn Turkstra for helpful discussions of this work. This research was made possible by the training fellowship to Lindsey Frederixon Byom from the University of Wisconsin-Madison Department of Communicative Disorders T32 Interdisciplinary Research Training Grant.

References

Adolphs, R. (1999). Social cognition and the human brain. *Trends in Cognitive Sciences, 3*(12), 469-479.

Adolphs, R. (2009). The social brain: Neural basis of social knowledge. *Annual Review of Psychology, 60*, 693-716.

Amodio, D. M., & Frith, C. D. (2006). Meeting of minds: The medial frontal cortex and social cognition. *Nature Reviews Neuroscience, 7*(4), 268-277.

Bibby, H., & McDonald, S. (2005). Theory of Mind after traumatic brain injury. *NEUROPSYCHOLOGIA, 43*(1), 99-114.

Carlson, S. & Moses, L. (2001). Individual differences in inhibitory control and children's Theory of Mind. *Child Development*, 72(4), 1032-1053.

Carlson, S.M. Moses, L.J., & Claxton, L.J. (2004). Individual differences in executive functioning and Theory of Mind: An investigation of inhibitory control and planning ability. *Journal of Experimental Child Psychology, 87*(4), 299-319.

Channon, S., Pellijeff, A., & Rule, A. (2005). Social cognition after head injury: Sarcasm and Theory of Mind. *Brain and Language, 93*(2), 123-134.

Delis, D. C., Kaplan, E., & Kramer, J. H. (2001). *D-KEFS: Delis Kaplan Executive Function System* San Antonio, TX: The Psychological Corporation.

Dunn, L. M., & Dunn, L. M. (1997). *PPVT-III: Peabody picture vocabulary test* (3rd ed.). Circle Pines, MN: American Guidance Service.

Fletcher, P. C., HappÈ, F., Frith, U., Baker, S. C., Dolan, R. J., Frackowiak, R. S. J., et al. (1995). Other minds in the brain: a functional imaging study of "Theory of Mind" in story comprehension. *Cognition, 57*(2), 109-128.

Frith, C., &, & Frith, U. (1999). Interacting minds -- a biological basis. *Science, 286*(5445), 1692-1695.

German, T.P., & Hehman, J.A. (2006). Representational and executive selection resources in 'Theory of Mind': Evidence from compromised belief-desire reasoning in old age. *Cognition, 101*(1), 129-152.

Happe, F., G.E. (1994). An Advanced Test of Theory of Mind: Understanding of Story Characters' Thoughts and Feelings by Able Autistic, Mentally Handicapped, and Normal Children and Adults. *Journal of Autism and Developmental Disorders, 24*(2), 129-154.

Kaufman, A. S., & Kaufman, N. L. (2004). *K-BIT2: Kaufman Brief Intelligence Test* (Second ed.). Minneapolis, MN: NCS Pearson, Inc.

Leslie, A. M., Friedman, O., & German, T. P. (2004). Core mechanisms in 'Theory of Mind'. *Trends in Cognitive Sciences, 8*(12), 529-533.

Leslie, A. M., German, T. P., & Polizzi, P. (2005). Belief-desire reasoning as a process of selection. *Cognitive Psychology, 50*(1), 45-85.

McKinnon, M. C., & Moscovitch, M. (2007). Domain-general contributions to social reasoning: Theory of Mind and deontic reasoning re-explored. *Cognition, 102*(2), 179-218.

Miyake, A., Friedman, N.P., Emerson, M.J., Witzki, A., & Howerter, A. (2000). The unity and diversity of executive function and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology, 41*, 49-100.

Simon, J. R. & Rudell, A. P. (1967). Auditory S-R compatibility: the effect of an irrelevant cue on information processing. *Journal of Applied Psychology, 51*, 300-304.

Smith, E.E., & Jonides, J. (1999). Storage and executive processes in the frontal lobe. *Science, 283*, 1657-1666.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18*, 643–662.

Turkstra, L. S., Dixon, T. M., & Baker, K. K. (2004). Theory of Mind and social beliefs in adolescents with traumatic brain injury. *NeuroRehabilitation, 19*(3), 245-256.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001a). *Woodcock-Johnson III Tests of Achievement*. Itasca, IL: Riverside Publishing.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001b). *Woodcock-Johnson Tests of Cognitive Abilities*. Itasca, IL: Riverside Publishing

Zelazo, P. D., Craik, F. I. M., & Booth, L. (2004). Executive function across the life span. *Acta Psychologica, 115*(2-3), 167-183.

An ACT-R List Learning Representation for Training Prediction

Michael Matessa (mmatessa@alionscience.com)
Alion Science and Technology
6404 Cooper Street
Felton, CA 95018 USA

Abstract

This paper presents a representation of training based on an ACT-R model of list learning. The benefit of the list model representation for making training predictions can be seen in the accurate a priori predictions of trials to mastery given the number of task steps. The benefit of using accurate step times can be seen in the even more accurate post-hoc model results.

Keywords: Training; prediction; list length; ACT-R.

Introduction

Numerous studies have documented operational and training problems with the modern autoflight systems, in particular the flight management system (FMS) and its pilot interface, the control display unit (CDU). During the last few years, more attention has been given to the limitations of current autoflight training methods. Many studies have concluded that current training programs are inadequate in both depth and breadth of coverage of FMS functions (Air Transport Association, 1999; BASI, 1998; FAA Human Factors Team, 1996).

Matessa and Polson (2006) proposed that the inadequacies of the programs are due to airline training practices that encourage pilots to master FMS programming tasks by memorizing lists of actions, one list for each task. Treating FMS programming skills as lists of actions can interfere with acquisition of robust and flexible skills. This hypothesis of the negative consequence of list-based representation was validated by Taatgen, Huss, and Anderson (2008), who show poorer performance for list-based representation compared to a stimulus-based representation.

This paper extends the table-based training time predictions of Matessa and Polson (2006) by presenting a computational model that represents procedure training as list learning. The model is meant to describe training programs where to-be-learned procedures are formally trained, and trainees must demonstrate mastery before they can go on to more advanced, on-the-job training. Airline transition training programs are examples of this paradigm. The model takes as input the number of steps in a procedure and the time per step, and it generates estimates of the training time required to master the procedure. Predictions of the model are compared to human data and show the benefit of the number-of-steps and step-time parameters.

Model

Novice pilots lack an organizing schema for memorizing lists of actions and so the actions are effectively represented as nonsense syllables (Matessa & Polson, 2006). Therefore, the list model does not represent the actual information to be learned, but instead as an engineering approximation represents the training as learning a list of random digits. The model is motivated by the table-based list model of Matessa and Polson (2006), but is implemented in the ACT-R cognitive architecture (Anderson, 2007).

Table-Based List Model

The following description from Matessa and Polson (2006) shows how procedure learning can be represented as list learning, and a table-based prediction of training time can be created based on procedure length. A representation of a task must encode both item (actions and parameters) and order information. Anderson, Bothell, Lebiere, and Matessa (1998) assumed that item and order information is encoded in a hierarchical retrieval structure incorporated in their ACT-R model of serial list learning shown in Figure 1. The order information is encoded in a hierarchically organized collection of chunks. The terminal nodes of this retrieval structure represent the item information. The model assumes that pilots transitioning to their first FMS-equipped aircraft master a cockpit procedure by memorizing a serial list of declarative representations of individual actions or summaries of subsequences of actions. It is assumed that each of these attempts to learn the list is analogous to a test-study trial in a serial recall experiment.

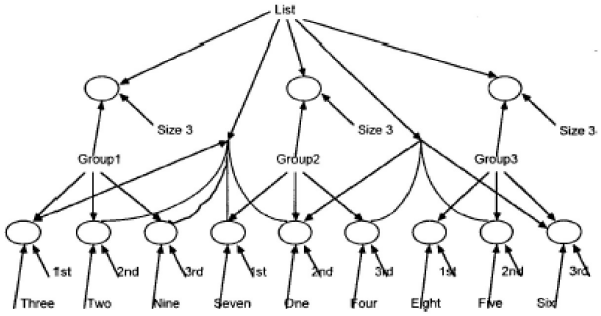


Figure 1: The List Model representation for a list of nine digits (from Anderson et al., 1998).

An interpretive process uses the list to perform the procedure. This process incorporates the knowledge necessary to understand each step description and to execute actions necessary to perform each step. Thus, an item such as "Press the LEGS key" would generate the actions required to locate the Legs key on the CDU keyboard and press it. A parameter such as a waypoint identifier would be represented in working memory as a sequence of letters. The interpretative process would generate the keystrokes necessary to enter the identifier into the scratch pad.

The list actions representation is a consequence of pilots' decisions to treat the task of mastering FMS procedures as learning serial lists of actions. The retrieval structure shown in Figure 1 is generated by processes that adults use to memorize any arbitrary serial list of items. It is assumed that a novice representation of a FMS procedure with nine actions would be represented by replacing the terminal-node chunks with chunks representing individual actions in the procedure. The retrieval structure only encodes order information and supports access to the chunks representing individual actions. The groupings of the actions imposed by this structure have no relationship to the underlying task structure. Because these retrieval structures are unique to each task, they block transfer of training.

The following figure is a possible list describing an FMS procedure for the Boeing 777 for responding to the following hold clearance that would be generated by a pilot with limited glass-cockpit experience.

"NASA 1: Hold west of Haden on the 270 degree radial. Right turns. 10 mile legs. Expect further clearance at 2130 z."

1. Press HOLD Function/Mode Key.
2. Press LSK 6L, if a holding pattern is in the route.
3. Line select waypoint identifier for Haden to scratchpad.
4. Press LKS 6L.
5. Enter the quadrant and the radial, W/270.
6. Press LSK 2L.
7. Enter the turn direction into the scratchpad, R.
8. Press LSK 3L.
9. Enter the leg distance into the scratchpad, 10.
10. Press LSK 5L.
11. Enter expect further clearance time, 2130.
12. Press LSK 3R.
13. Verify the resulting holding pattern on the ND.
14. Press EXECUTE.

Figure 2: A possible novice representation of a FMS procedure for responding to a Hold clearance.

This probably looks like a list of nonsense syllables to you, as it does to novice pilots. Pilots do not receive an explicit instruction on how to encode FMS procedures in memory early in training and lack organizing schemas that would help in memorizing instructions. Catrambone (1995) has shown that novices tend to describe problem solutions in terms of actions used to solve the problem. In the case of FMS programming skills, this process leads to long lists that are very difficult to memorize.

The list shown in Figure 2 has undesirable properties and would be difficult to memorize. It is long—14 items—and it is organized as a linear sequence of actions that cannot be directly stored in memory (Anderson, et al., 1998). Some kind of idiosyncratic organization would have to be imposed on it to break it up into sublists before it could be successfully memorized. Furthermore, the representation of the procedure for programming a hold shown in Figure 2 is specific to a particular clearance. It would be relatively easy to generalize this representation to clearances with identical parameters but with different values. However, generalizing this procedure to cover the entry of any hold clearance requires numerous nontrivial inferences.

The Savings Paradigm The list model assumes that learning a FMS procedure is analogous to memorizing serial lists of nonsense syllables for a pilot with limited FMS experience. Training times can be estimated using results of an experimental paradigm initially developed by Ebbinghaus (1888/1913, Chapter 8). On the first day of the experiment, participants learn a serial list of items to a criterion of mastery of one perfect recitation of the list. Performance is measured as the number of trials to mastery. Participants return to the laboratory 24 hours later and relearn the list to the same criterion of mastery. Training stops on the first day that participants perform perfectly on the first presentation of the list after a 24-hour retention interval.

Table-based Prediction Matessa and Polson (2006) developed a table that presents the number of retentions on each successive day and the number of days of training required to be able recall a list perfectly after 24 hours. The numbers in the table were derived by synthesizing the results of several experiments from the list-learning literature starting with the data from Ebbinghaus (1885/1913, Chapter 8). The numbers are extrapolations generated by fitting power functions to Ebbinghaus's results and then adjusting them to account for the fact that he used a very rapid presentation rate.

Training time is estimated by calculating the amount of time it would take to administer N repetitions of a procedure of length L during one session in a fixed-base or full-motion simulator. The model's description of the training processes has three time parameters: session setup time (SST), repetition setup time (RST), and step time (ST). SST is the time required to set up a simulator to begin training a specific procedure. RST is the time required to set up the simulator for the next repetition, and ST is the time required for a trainee to perform a step and receive feedback from the instructor if necessary. These values are then summed over days to generate a training- time prediction for a given procedure.

The time devoted to training a procedure on one day = SST + N*RST + N*L*ST.

The values for N, the number of repetitions on a day, are taken from the table. Values for SST and RST were set to 120 seconds, and ST was set to 5 seconds. Current fixed-based and full-motion simulators were found to be ill-suited to this kind of training; they are designed to simulate the execution of complete missions.

Numerous studies have shown that PC-based, part-task simulators can be used successfully to train skills such as performing FMS procedures (e.g., Salas, Bowers, and Prince, 1998; Salas, Bowers, and Rhodenizer, 1998; and Polson, Irving, and Irving, 1994). The lesson planners incorporated into commercially developed simulators can be programmed to deliver the necessary repetitions while minimizing the SST and RST (Aerosim Technologies, www.aerosim.com; Tricom Technologies, www.tricom-tech.com/products.htm; CAE, www.Cae.com; and Wicat, www.wicat.com). Use of such a trainer was modeled by reducing the values of SST and RST to 5 seconds.

ACT-R List Model

This paper presents a computational list model developed in the ACT-R cognitive architecture (Anderson, 2007). ACT-R includes a subsymbolic level of representation where facts have an activation attribute which influences their probability of retrieval and the time it takes to retrieve them. The activation A_i of a chunk i is computed from two components – the base-level and a context component. The base-level activation B_i reflects the recency and frequency of practice of the chunk. The equation describing learning of base-level activation for a chunk i is

$$B_i = \ln(\sum_{j=1}^{n} t_j^{-d})$$

where n is the number of presentations for chunk i, t_j is the time since the jth presentation, and d is the decay parameter. The equation for the activation A_i of a chunk i including context is defined as:

$$A_i = B_i + \sum_j W_j S_{ji}$$

where the base-level activation B_i reflects the recency and frequency of practice of the chunk as described above. The elements j in the sum are the chunks which are in the slots of the chunk in module k. W_{kj} is the amount of activation from source j in module k. The strength of association, S_{ji}, between two chunks is 0 if chunk j is not in a slot of chunk i or is not itself chunk j. Otherwise it is set using the following equation:

$$S_{ji} = S - \ln(m)$$

Built into this equation is the prediction of a fan effect (Anderson, 1974) in that the more things associated to j the less likely any of them will be, on average, in the presence of j. That is, if there are m elements associated to j their average probability will be 1/m.

The current model is an ACT-R 6.0 model based on the ACT-R 4.0 list learning model developed by Anderson et al. (1998) and can account for phenomena such as length and serial position effects. Figure 3 plots the probability of correctly recalling a digit in position as a function of serial position in input. There is considerable variation in recall of items both as a function of list length and input position. These variations are predicted by the model as a reflection of the changes in activations of the elements being retrieved. These activations increase with rehearsal (base-level activation), decrease with time (base-level activation), and decrease with list length (associative activation). As the list is longer, there will be greater interference because there will be more associations from the list element and less associative activation to any member of the list.

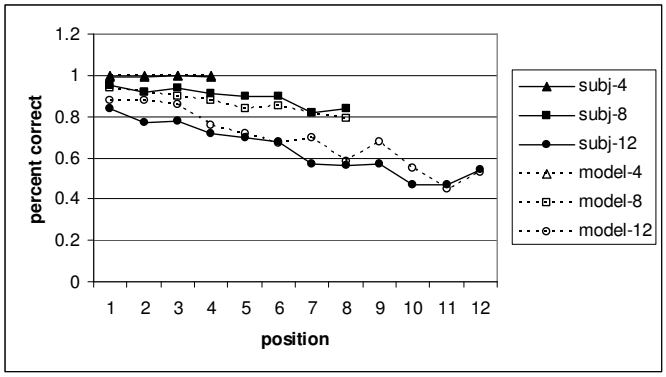


Figure 3: List model showing length and serial position effects.

In order to approximate training, the current model differs from the Anderson et al. (1998) model by not implementing its rehearsal strategy. In this way, presentation rate represents task step time (ST). As a consequence, longer presentation rates produce poorer performance, in contrast to findings from studies that allow rehearsal.

The model also uses the Pavlik and Anderson (2005) version of memory decay that accounts for spacing effects. They developed an equation in which decay for the ith presentation, d_i, is a function of the activation at the time it occurs instead of at the lag. This implies that higher activation at the time of a trial will result in the gains from that trial decaying more quickly. On the other hand, if activation is low, decay will proceed more slowly.
Specifically, they propose

$$d_i(m_{i-1}) = ce^{m_{i-1}} + a$$

to specify how the decay rate, d_i, is calculated for the ith presentation of an item as a function of the activation m_{i-1} at the time the presentation occurred, with

$$m_n(t_{1\ldots n}) = ln\left(\sum_{i=1}^{n} t_i^{-d_i}\right)$$

showing how the activation m_n after n presentations depends on the decay rates, d_is, for the past trials.

These equations result in a steady decrease in the long-run retention benefit for additional presentations in a sequence of closely spaced presentations. As spacing gets wider in such a sequence, activation has time to decrease between presentations; decay is then lower for new presentations, and long-run effects do not decrease as much.

The model is run inside code that simulates the savings paradigm in order to determine trials to mastery. The model uses the same parameters as Anderson et al. (1998) except that the rate of presentation (representing step time) and repetition setup time are both set to 5 seconds, as in Matessa and Polson (2006). The activation retrieval threshold is set to -0.85 in order to match the predictions of the trials to mastery table found in Matessa and Polson (2006).

Experiment

In order to gather data for an experimental interface, Boeing conducted experiments with a PC-based, part-task simulator to compare the new interface to the current 777 interface (Prada, Mumaw, Boehm-Davis, & Boorman, 2007). Results from these experiments can be compared with model predictions to show the usefulness of the list modeling approach.

Boeing Pilot Performance

Boeing gathered performance data on flight tasks in a medium-fidelity, setting to get feedback on proposed interface improvements and to generate performance data comparing the 777 design to the proposed design (Prada et al., 2007). Two desktop computer simulations of the 777 and proposed automatic flight control panels and associated displays were created. The simulations provided appropriate feedback, including mode changes, as controls were manipulated. However, the aircraft remained frozen in time and space until advanced by the experimenter. Participants controlled the simulation using a standard two-button mouse. For this paper, only data from the 777 interface is considered.

Participants The participants consisted of twelve FMC-naïve subjects who were male Boeing employees. All were general aviation pilots with instrument rating. Six had commercial certification and four were not instrument current. They had no previous exposure to the 777 FMC.

Procedure Twenty training tasks were selected to capture tasks that are difficult on each interface and to provide a representative set of functions. In the training tasks, for each action (click) on the interface, the location and time were collected. Also collected were overall task time, number of steps correct, and trials to mastery.

Results The number of steps in the tasks ranged from two steps to thirteen steps. For this paper, tasks are grouped into those with an average of two, four, seven, and thirteen steps. Trials to mastery increased with the number of steps in the task (Figure 4).

Model Performance

The original list model of Anderson et al. (1998) made predictions for lists with three items up to twelve items. The current model retains this range, and so, for analysis, tasks with two steps are compared to lists with three items and tasks with thirteen steps are compared to lists with twelve items (four steps are compared directly, as are seven).

Results Model runs with the step time of 5 seconds used by Matessa and Polson (2006) show trials to mastery increasing with the number of steps in the task. The difference in trials to mastery between the model and subjects averaged 1.5 trials (Figure 4, model-pre).

A post-hoc analysis used the actual average step time from subjects as input to the model. For tasks with an average of two, four, seven, and thirteen steps, the average step time was 15.2, 8.1, 8.0, and 6.5 seconds, respectively. The difference in trials to mastery between this model run and subjects averaged 0.8 trials (Figure 4, model-post).

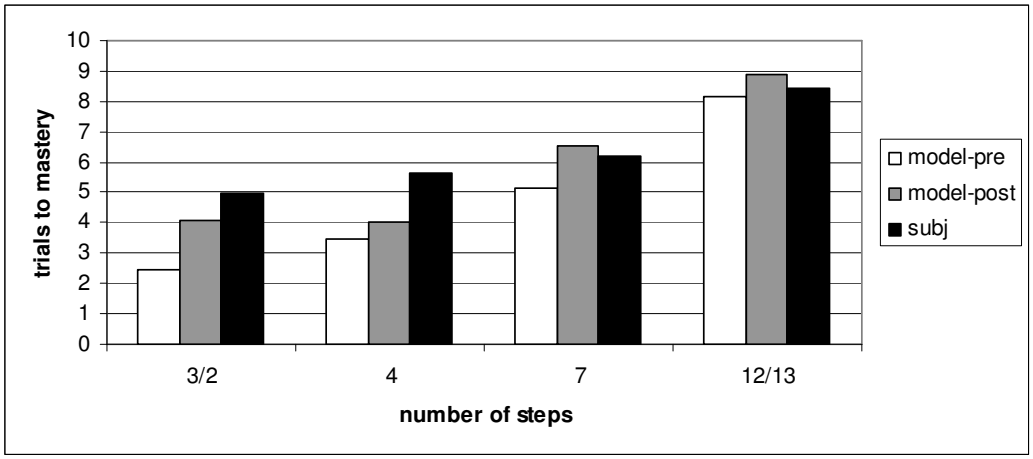


Figure 4: Trials to mastery for model and subjects.

Conclusions

The benefit of the list model representation for making training predictions can be seen in the accurate *a priori* predictions of trials to mastery given the number of task steps. The benefit of using accurate step times can be seen in the even more accurate post-hoc model results.

Ideally, the list model would be given an accurate estimate of step times without seeing the data ahead of time. To this end, the list model is currently being integrated with CogTool (John, Prevas, Salvucci, & Koedinger, 2004). CogTool takes as input a demonstration of an interface task and returns a zero-parameter prediction of task performance time based on ACT-R primitives. With this information, the number of steps in the task and average step time can be fed into the list model in order to make training predictions. A number of open issues remain, such as the level of abstraction of a "step". Does a step to push a button include the visual search for that button, or is that a separate step? More empirical work is needed to determine in what situations the list model representation can be useful in training prediction.

Acknowledgments

Funding for this work was provided by the National Aeronautics and Space Administration.

References

Air Transport Association. (1999). *Performance of Standard Navigation Tasks by FMS-Generation Aircraft* (Third report by the Human Factors Committee Automation Subcommittee).

Anderson, J. R. (2007) *How Can the Human Mind Occur in the Physical Universe?* Oxford University Press.

Anderson, J. R. (1974). Retrieval of propositional information from long-term memory. *Cognitive Psychology*, 5, 451-474.

Anderson, J. R., Bothell, D., Lebiere, C., & Matessa, M. (1998). An Integrated Theory of List Memory. *Journal of Memory and Language*, vol. 38, 1998, pp. 341–380.

BASI (1999) Advanced Technology Safety Survey Report. Flight Safety Digest Special Issue. Flight Safety Foundation, June-Aug 1999, pages 137-216.

Catrambone, R. (1995) Aiding subgoal learning: Effects on transfer. *Journal of Educational Psychology*, 87, 5-17.

Ebbinghaus, H. (1885/1913). Memory: A contribution to experimental psychology (Henry A. Ruger & Clara E. Bussenius, Trans.). New York: Teachers College, Columbia University.

Federal Aviation Administration (FAA) Human Factors Team. 1996. Report on the Interfaces between Flightcrews and Modern Flight Deck Systems (June 18, 1996). Washington: U.S. Department of Transportation, Federal Aviation Administration.

John, B., Prevas, K., Salvucci, D., & Koedinger, K. (2004) Predictive Human Performance Modeling Made Easy. *Proceedings of CHI, 2004* (Vienna, Austria, April 24-29, 2004) ACM, New York.

Matessa, M., & Polson, P. (2006). List Models of Procedure Learning. *Proceedings of the International Conference on Human-Computer Interaction in Aeronautics* (pp. 116-121), San Francisco, CA.

Pavlik, P. I., Jr., & Anderson, J. R. (2005). Practice and forgetting effects on vocabulary memory: An activation-based model of the spacing effect. *Cognitive Science*, 29, 559-586.

Polson, P. G., Irving, S., & Irving, J. E. (1994). Final report: Applications of formal methods of human computer interaction to training and the use of the control and display unit. Washington, DC: System Technology Division, ARD 200, Department of Transportation, FAA.

Prada, L. Ricardo; Mumaw, R J.; Boehm-Davis, D. A.; Boorman, D.J. (2007) Testing Boeing's Flight Deck of the Future: A Comparison Between Current and Prototype Autoflight Panels. *Human Factors and Ergonomics Society Annual Meeting Proceedings*, Aerospace Systems, pp. 55-58(4).

Salas, E., Bowers, C.A., and Prince, C. eds. (1998). Special Issue: Simulation and Training in Aviation. *International Journal of Aviation Psychology*, 8(3).

Salas, E., Bowers, C. A., & Rhodenizer, L. (1998). It is not how much you have but how you use it: Toward a rational use of simulation to support aviation training. *International Journal of Aviation Psychology*, 8(3), 197-208.

Taatgen, N.A., Huss, D. & Anderson, J.R. (2008). The Acquisition of Robust and Flexible Cognitive Skills. *Journal of Experimental Psychology: General*, 137(3), 548-565.

Eye Movements During Mental Imagery are Not Reenactments of Perception

Roger Johansson (Roger.Johansson@ling.lu.se)
Department of Cognitive Science, Lund University
Kungshuset, Lundagård, 222 22, Lund, Sweden

Jana Holsanova (Jana.Holsanova@lucs.lu.se)
Department of Cognitive Science, Lund University
Kungshuset, Lundagård, 222 22, Lund, Sweden

Kenneth Holmqvist (Kenneth.Holmqvist@humlab.lu.se)
Humanities Laboratory, Lund University
Box 201, 221 00, Lund, Sweden

Abstract

In this study eye movements were recorded for participants under three different conditions. All three conditions consisted of a perception phase and a mental imagery phase. The imagery phase was similar for all conditions: i.e., participants looked freely at a blank white screen. The perception phase was different for each condition. In the control condition, participants looked freely at a complex picture. In the first experimental condition, they looked at another complex picture but maintained fixation at the center of the picture. In the second experimental condition, they maintained central fixation while listening to a verbal scene description. The results revealed that despite central fixation during the perception phase under the two experimental conditions, participants' eye movements were spread out during the imagery phase, reflecting the spatial positions and directions within the picture or scene. These results contradict the theory that eye movements during mental imagery are re-enactments of perception.

Keywords: Eye movements, mental imagery, spatial cognition, visual attention, scene description.

Introduction

Since the late Nineties, several eye-tracking studies have reported that spontaneous eye movements occur with mental imagery and that these eye movements closely reflect the content and spatial relations from an original picture or scene (e.g., Brandt & Stark, 1997; Holsanova, Hedberg & Nilsson, 1998; Laeng & Teodorescu, 2002; Gbadamosi & Zangemeister, 2001; Altmann, 2004; Johansson, Holsanova & Holmqvist, 2006; Humphrey & Underwood, 2008). A similar effect has been found for spatial relations and scenes that are verbally described (e.g., Demerais & Cohen, 1998; Spivey, Tyler, Richardson, & Young, 2000; Spivey & Geng, 2001; Johansson et al, 2006). It has further been shown that this effect is equally strong irrespective of whether the original elicitation was visual or verbal (Johansson et al., 2006). Additionally, an eye movement effect of this kind has also been found during problem-solving tasks (e.g., Yoon & Narayanan, 2004; Freksa & Bertel, 2007) as well as with visual motor imagery (Heremans, Helsen & Feys, 2007; Gueugneau, Crognier & Papaxanthis, 2008). From this large body of research, it appears that eye movements play an important role in visual imagery and in the construction of mental models. However, what role these eye movements have, and why they appear, are issues of debate (cf., Johansson et al., 2006; Ferreira, Apel, & Henderson, 2009; Richardson, Altmann, Spivey & Hoover, 2009).

Hebb (1968) suggested a *functional* role for eye movements during mental imagery, and proposed that they are necessary to assemble and organize "part images" into a whole visualized image. This functional view has gained strong support from a study by Laeng and Teodorescu (2002). In their study participants inspected visual stimuli of two kinds: 6×6 grid patterns with 5 black filled cells or a small fish in various locations on the screen. One group was instructed to maintain fixation onto the screen's center and another group was free to inspect the stimuli. In a subsequent imagery phase, both groups were instructed to 'build a visual image of the figure' they had just seen and were then allowed to move their eyes freely while looking at a blank screen. The results revealed that those who maintained their gaze centrally in the perception phase did the same, spontaneously, during the imagery phase, while those who inspected the original stimuli freely had eye movements during the imagery phase which, to a high degree, resembled those in the perception phase. Laeng and Teodorescu (2002) argued that this implied eye movements are stored along with a visual representation of the scene, and are used as spatial indexes to properly arrange the parts of a mental image. They concluded that eye movements during mental imagery are *re-enactments* of perception and have a necessary and functional role in "constructing" the mental image. However, the question can be raised whether the instruction to 'build a visual image', in combination with the relatively simple stimuli, might necessarily lead to spatial scanning of the mental image.

As discussed in Johansson et al. (2006), the task and the complexity of the stimuli are important when the scene is recalled during mental imagery. For instance, it is possible that the mental image is only covertly scanned or is not scanned at all. Thomas and Lleras (2009) have shown that shifts in covert attention can produce identical results in a

problem-solving task to overt eye movements. It is however less likely that shifts in covert attention, or lack of scanning altogether, would be sufficient when recalling scenes that are rich in detail and contain many objects: i.e., visualizing highly complex scenes would increase the cognitive load such that more internal operations would be needed to construct the parts of the image and then tie them together and place them into a context.

The purpose of the present study is to investigate whether Laeng and Teodorescus' (2002) 'central gaze effect' occurs even for visual scenes of high complexity. To ensure that spatial scanning is actually employed, the experimental design and method from Johansson et al. (2006) was used. In this method the imagery task is to orally describe the scene from memory, which introduces a great need for spatial scanning. Additionally, by including two types of stimuli – visual scenes and verbal descriptions – we can investigate mental imagery for scenes that have never been seen in the first place.

Experiment

The experiment consisted of three conditions: a *control* condition, a *fixed-picture* condition and a *fixed-verbal* condition. All three conditions consisted of a *perception* phase and a *mental imagery* phase. The imagery phase was similar for all conditions: i.e., participants looked freely at a blank white screen. The perception phase was different for each condition. In the control condition, participants looked freely at a complex picture. In the fixed-picture condition, they looked at another complex picture but were instructed to maintain fixation at the center of the picture. In the fixed-verbal condition, they were instructed to maintain central fixation while listening to a verbal description of a scene.

Participants

Twenty students at the University of Lund – ten females and ten males – participated in the experiment. All subjects reported either normal vision or vision corrected to normal (i.e., with contact lenses or glasses). All participants were native Swedish speakers. The mean age of the participants was 21.4 years (SD = 1.9).

Apparatus and stimuli

The participants were seated in front of a computer screen at a distance of 600-700 mm. (The distance varied slightly because of the subjects' freedom to move their head and body.) The eye tracker used was the SMI iView RED250, which has a sampling frequency of 250 Hz and a precision of 0.02°. The data was recorded with the iView X 2.4 software. The eye-tracking data was analyzed with BeGaze 2.4 and in-house MatLab programs.

The visual stimulus in the experiment was presented using Experiment Center 2.4 on a 480 mm. x 300 mm. computer screen with a resolution of 1680 × 1050 pixels. The auditory stimulus was a pre-recorded description (one minute and 38 seconds long). An English translation of the scene description follows:

"In the center, right in front of me, I see a large, green spruce. At the top of the spruce, a bird is sitting. To the left of the spruce – to the far left – there is a yellow house with a black tin roof and with white corners. The house has a chimney on which a bird is sitting. To the right of the large spruce – to the far right – there is a tree as high as the spruce. The leaves of the tree are colored yellow and red. Above the tree a bird is flying. Between the spruce and the tree is a man in blue overalls, raking leaves. Below the spruce, the house, the tree and the man, i.e. in front of them there is a long red fence, which runs all the way from left to right. At the left end, a bike is leaning against the fence. Just to the right of the bike is a yellow mailbox. On top of the mailbox, a cat is sleeping. Below the fence, i.e. in front of and along the fence, a road leads from the left side to the right side. On the road, to the right of the mailbox and the bike, a black-haired girl is bouncing a ball. To the right of the girl, a boy in a red cap is sitting and watching her. To the far right along the road, a lady in a big red hat is walking with some books under her arm. Just to the left of her, on the road, a bird is eating a worm."

Procedure

Participants were told that the experiment concerned pupil dilation in relation to mental workload. It was explained that we would be filming their eyes, but nothing was said about us recording their eye movements. They were asked to keep their eyes wide open so that we could film their pupils, and to look directly ahead so that our equipment could accurately measure their pupil dilation. The eye tracker was calibrated using a five-point calibration procedure with validation. (This is the default setting in Experiment Center 2.4). Participants' eye movements were recorded during both the perception and imagery phase under all three conditions.

In the *control* condition, a picture was shown for thirty seconds. Then the screen went blank, and participants were asked to describe the picture in their own words. They were told explicitly to keep their eyes wide open and to look directly ahead so that the equipment could record their pupil dilation. Figure 1 shows the schematics of the control condition.

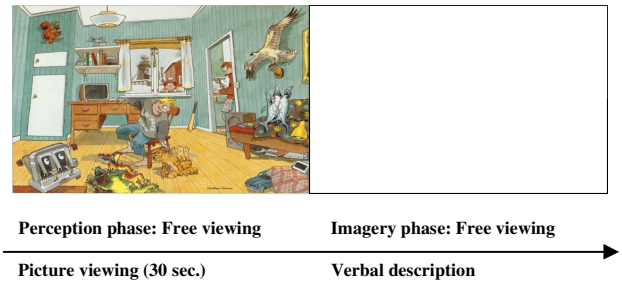


Figure 1: Control condition

In the *fixed-picture* condition, participants were instructed to maintain fixation on a cross in the center of the screen until

it disappeared. The cross was first shown for five seconds, after which a picture appeared behind it for an additional thirty seconds. Then the screen went blank, and participants were asked to describe the picture in their own words. They were told explicitly to keep their eyes wide open and to look directly ahead so that the equipment could record their pupil dilation. Figure 2 shows the schematics of the fixed-picture condition.

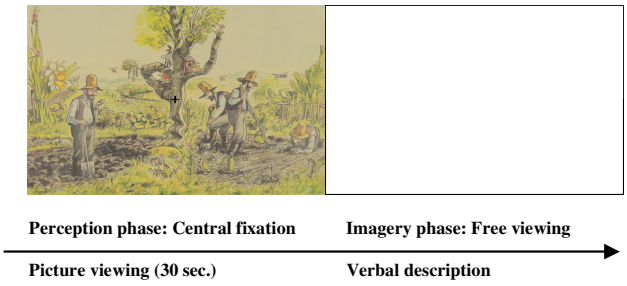


Figure 2: Fixed-picture condition

In the *fixed-verbal* condition, participants were likewise instructed to maintain fixation on a cross in the center of the screen until it disappeared. The cross appeared in an otherwise blank screen while a pre-recorded scene description was played from speakers in front of the participants for 1:38 minutes. Then the cross disappeared. Participants were asked to retell the scene. They were told that they could retell it in their own words and did not have to follow the same order. Participants were told explicitly to keep their eyes wide open and to look directly ahead so that the equipment could record their pupil dilation. Figure 3 shows the schematics of the fixed-verbal condition.

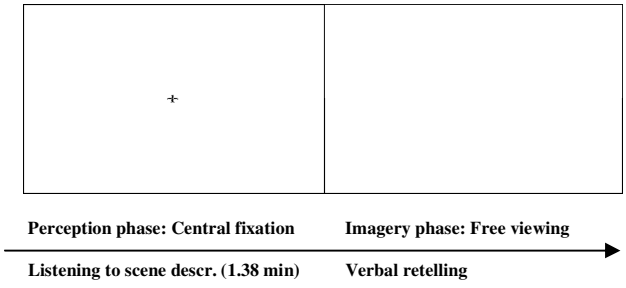


Figure 3: Fixed-verbal condition

Afterwards, to assess whether any of the participants had seen through the nature of the experiment, we asked what they thought the true objective of the experiment was.

Analysis

If Laeng and Teodorescus' (2002) conclusion – that eye movements during imagery functionally reenact those of perception – is supported then participants' eye movements should remain centrally fixated during the imagery phase for the fixed-picture condition and the fixed-verbal condition: i.e., their eye movements should have similar spatial dispersion as during the perception phase and therefore not correspond to directions and positions from the imagined scene. To test this, we chose to analyze eye movements in two regards. First, the overall spatial dispersion of the eye movements was considered. However, spatial dispersion does not give any information about how eye movements correspond to directions and positions in a mental image. Also, it is common that participants "shrink" their mental image and only look at a limited part of the screen during imagery (Gbadamosi & Zangemeister, 2001; Johansson et al., 2006). Therefore, as a second step, a method combining eye movement data and verbal data (cf., Holsanova, 2008) was used.

To analyze the overall spatial dispersion of the eye-tracking data, a modified version of the *coverage measure* proposed by Wooding (2002) was calculated for each phase (perception/mental imagery) and condition. An "attention map" was created by centering a Gaussian function (σ = 0.1W, W = 1680 pixels) at each fixation point and then superimposing all the other functions. The volume under the attention map, after being normalized to unit height, was then used to estimate the spatial dispersion of the eye-tracking data. Within-subject ANOVAs were done to analyze the spatial dispersion between the perception and imagery phases in each condition, as well as between conditions for the imagery phase.

To analyze whether eye movements corresponded to directions and positions from the verbal descriptions and retellings, the method developed and described in Johansson et al. (2006) were used. Since it is possible that participants can make use of either the whole screen or only a part of it in imagining the scene, one cannot simply take physical coordinates on the computer screen as one's areas of interest. Instead, this method uses the relative position of an eye movement compared to each participant's individual gaze pattern over the entire description or retelling. Eye movements are then scored as correct or incorrect according to either *global correspondence* or *local correspondence* coding. The spatial criteria for an eye movement to be considered correct in global correspondence coding is defined as when an eye movement shifts from one object to another it must finish in a position that is spatially correct relative to the participant's gaze pattern over the entire description or retelling. The spatial criteria for local correspondence is defined as when an eye movement shifts from one object to another during the description or the retelling it must move in the correct direction (up, down, left or right). The minimum threshold for the saccadic amplitude to be considered an actual movement from one object to another was set at 35 pixels (10 mm on the screen). In addition to these spatial criteria, we used the temporal criteria from Johansson et al. (2006), where an eye movement from one position to another must appear within five seconds before or after an object is mentioned.

The key difference between global and local correspondence is that global correspondence requires

fixations to take place at the categorically correct *spatial position* relative to the whole gaze pattern, whereas local correspondence only requires that the eyes move in the correct *direction* between two consecutively mentioned objects. Eye movements are considered incorrect when neither the local correspondence nor the global correspondence criteria are met: e.g., when the eyes move with amplitudes below the 35-pixel threshold or in the wrong direction.

As a consequence of applying these spatial criteria a binomial distribution in the data is obtained: the spatial relations are either correct or incorrect. The possibility that a participant would move his or her eyes to the correct position by chance was then defined. For global correspondence coding, both the direction and the distance of the movement must be correct. Many movements are possible. In this study a conservative estimate was chosen, whereby the eyes could move in at least four directions (up, down, left, and right) to at least two locations (full and half distance). In addition to these eight possibilities, the eye might stand still (or move with an amplitude below the 35-pixel threshold). For global correspondence, the probability that the eyes moved to the correct position at the correct time by chance is thus definitely less than one in nine (11%). For local correspondence coding, which requires only correct direction, the corresponding probability is one in five (20%). The data could then be analyzed using a Wilcoxon signed-rank test for significance between the total number of correct eye movements and the expected number of correct movements made by chance.

Finally, to compare the proportion of correct eye movements in global and local correspondence coding between the three conditions a within-subjects ANOVA was used.

Results and discussion

None of the participants saw through to the true objective of the experiment and data from all participants could be included in the results.

The comparison of spatial dispersion between the perception and imagery phases revealed a significantly larger spatial dispersion in the imagery phase under the fixed-picture condition ($F(1,19) = 29.429$, $p < 0.001$) and the fixed-verbal condition ($F(1,19) = 32.934$, $p < 0.001$). The results for the control condition were the opposite: i.e., spatial dispersion was significantly larger in the perception phase ($F(1,19) = 114.553$, $p < 0.001$). The comparison of spatial dispersion in the imagery phase between conditions revealed a significant main effect ($F(2,38) = 8.175$, $p = 0.002$). Bonferroni *post-hoc* tests revealed that spatial dispersion was significantly larger for the control condition than for either the fixed-picture condition ($p = 0.01$) or the fixed-verbal condition ($p = 0.02$). No significant difference was found between the fixed-picture and the fixed-verbal condition.

The average proportion for all participants of correct eye movements by local and global correspondence coding under each condition is presented in Table 1. Consistent with the results from Johansson et al. (2006) the control condition generated a high proportion of correct eye movements, both by local and global correspondence. However, also the central gaze conditions generated a high degree of correct eye movements in the local correspondence coding as well as a certain degree of correct eye movements in the global correspondence coding. Except for the eye movements in global correspondence coding in the fixed-picture-condition the results were significantly above chance ($p < 0.001$).

The comparison of correct eye movements by global and local correspondence coding in the imagery phase between the three conditions revealed a significant main effect for global correspondence coding ($F(2,38) = 5.544$, $p = 0.008$) but not for local correspondence coding. Bonferroni post-hoc tests revealed that there were significantly more correct eye movements (for global correspondence coding) under the control condition than under the fixed-picture condition ($p = 0.03$). No significant difference was found between the other conditions.

Table 1: Percentages of objects with correct eye movements in the imagery phase for all three conditions by both local and global correspondence coding.

	Control	Fixed-Picture	Fixed-Verbal
Global	55.8 %	26.5 %	34.5 %
Local	81.6 %	73.6 %	60.0 %

These results reveal that spatial dispersion of the eye movements was significantly larger in the imagery phase than in the perception phase under the two central gaze conditions, and that there was a significant degree of correct eye movements under the two central gaze conditions; especially for local correspondence coding. However, the results also showed that spatial dispersion was smaller in the imagery phase under the two central gaze conditions than under the control condition.

Figures 4-6 show scanpaths in both the perception and imagery phase for one and the same participant. Figure 4 shows that in the control condition this participant used a lot of the computer screen during imagery and her eye movements had a large spatial dispersion. Positions and directions for the eye movements corresponded to a high degree with described elements of the picture. Figure 5 shows that in the fixed-picture condition this participant had a large number of fixations in the center of the screen during the mental phase but also executed eye movements away from the center, resulting in a larger spatial dispersion than during the perception phase, and eye movements that spatially corresponded to what was described. For example, the eye movements to the far left were executed when the flowers to the far left of the picture were described. It is clear that during the perception phase, the participant looked at the central cross the entire time and never shifted to the flowers. Figure 6 shows that in the fixed-verbal condition

this participant executed a lot of eye movements across a large extent of the screen during the imagery phase, resulting in a larger spatial dispersion than during the perception phase, and eye movements that spatially corresponded to the described scene. For example, the eye movements to the left in this figure were executed when the house, the bike and the mailbox were mentioned, and the eye movements to the far right were executed when the second tree and the lady on the road were mentioned.

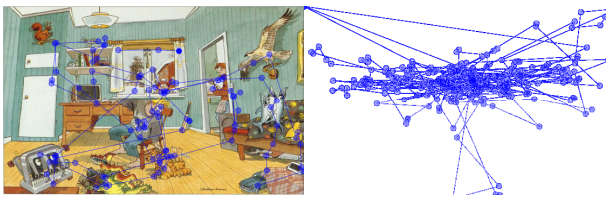


Figure 4: Control condition
(left: perception phase, right: imagery phase)

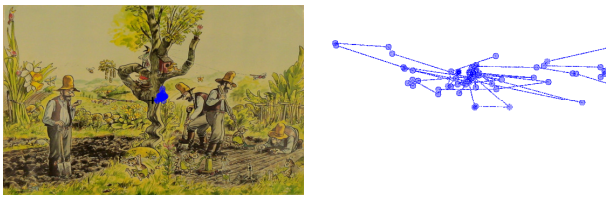


Figure 5: Fixed-picture condition
(left: perception phase, right: imagery phase)

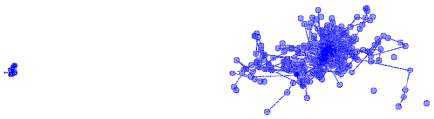


Figure 6: Fixed-verbal condition
(left: perception phase, right: imagery phase)

General discussion

The results show that despite maintaining central fixation during visual perception of either a complex picture or a verbal scene description, eye movements spread out and to a high degree correspond to spatial positions and directions during mental imagery of the picture or scene. These results contradict Laeng and Teodorescus' (2002) conclusion that eye movements during visual imagery reenact those of perception of the same scene. Nevertheless, it was also revealed that eye movements were less spread out during imagery under the two central gaze conditions than under the control condition and the proportion of correct eye movements was by global correspondence coding significantly lower (and not significantly above chance) for the fixed-picture condition than for the control condition. Therefore, it seems that central gazing in the perception phase to some degree did affect eye movements during imagery. We do, however, propose that this is an effect of the limitation of not being able to move the eyes during perception rather than a support for Laeng and Teodorescus' (2002) functional view. For example, under the fixed-picture condition most of the picture was only seen peripherally and participants were not able to describe as many objects (mean: 4.1) as in the control condition (mean: 7.6) and the description focused to a high degree on picture elements that were in focus during perception (the tree and the bird's nest). For the fixed-verbal condition we propose a similar explanation. Since the participants could not move their eyes when they listened to the scene description it was harder for them to form a mental image of the scene and less objects and spatial relations among them were remembered when the scene was recalled.

If eye movements during imagery are not reenactments of perception would this mean that they do not have a functional and necessary role for the construction of mental images? There has been a vibrant debate recently whether 'looking at nothing' can facilitate memory retrieval of visual scenes and what role internal depictive image representations have in this process (Ferreira, Apel, & Henderson, 2008; Richardson, Altmann, Spivey, & Hoover, 2009). Nevertheless, in this debate, eye movements to regions of a blank screen are interpreted in relation to a previous perception phase: i.e., again eye movements during mental imagery were seen as reenactments of perception. We propose that this is the wrong approach. Johansson et al. (2006) showed that participants who listened to a scene description while looking at a blank screen spontaneously performed eye movements that closely corresponded to spatial positions and directions from their visualizations of the scene. In this case, there was no previous perception phase that the eye movements could be reenacting. Another big problem for the 'reenactment approach' is that eye movements during imagery are idiosyncratic. For example, participants frequently "shrink" their mental image, and only look at a limited part of the screen when visualizing a previously seen picture that covered the entire screen (Gbadamosi & Zangemeister, 2001; Johansson et al., 2006). The results from the current study together with these previous findings strongly show that the phenomenon of eye movements during mental imagery is more complex than a mere reenactment of a perceptual phase. Therefore, to conclude that eye movements are necessary and functionally connected with the construction of a mental image is too strong of an assumption. A better approach might be to see them as a support that can relieve working memory load during imagery. If this is right, they become more likely to appear when a difficult imagery task is performed. This could explain why the results in this paper differ from those of Laeng and Teodorescu (2002). It is a much harder task to visualize and verbally describe a complex picture or scene description than to 'build an image' of the much simpler stimuli used in their study. Another possible interpretation comes from various versions of simulation theory (e.g.

Hesslow, 2002; Thomas, 1999), where eye movements during imagery do not have a direct and necessary link to eye movements from a perception phase. For example, the perceptual activity theory (Thomas, 1999) states that imagery is the reenactment of a perceptual behavior that would be appropriate for exploring the imagined scene as if it were actually present. Eye movements would therefore be likely to appear independently of how they were executed in perception.

Nevertheless, to explain the complex interplay between eye movements and mental imagery fully, further studies need to be performed: e.g., to investigate whether memory retrieval is enhanced by eye movements to blank areas of a screen and how individual differences in spatial cognition and working memory capacity are related to these movements.

Summary

This study showed that despite maintaining central fixation, either while looking at a complex picture or listening to a scene description, participants' eye movements spread out and did correspond to directions and positions during mental imagery of the picture or the scene. Laeng and Teodorescus' (2002) conclusion that eye movements during imagery reenact those of perception was therefore not supported.

Acknowledgements

This research was funded by the Swedish Research Council (Vetenskapsrådet 2007–2637) and by a Fellowship at Hanse Wissenschaftskolleg Delmenhorst. Special thanks to Marcus Nyström for programming the MatLab algorithms.

References

Altmann, G. T. M. (2004). Language-mediated eye movements in the absence of a visual world: the 'blank screen paradigm'. *Cognition*, 93 (2): B79-B87.

Brandt, S. A., & Stark, L. W. (1997). Spontaneous eye movements during visual imagery reflect the content of the visual scene. *Journal of Cognitive Neuroscience, 9,* 27–38.

Demarais, A., & Cohen, B.H. (1998). Evidence for image-scanning eye movements during transitive inference. *Biological Psychology*, *49*, 229-247.

Ferreira, F. Apel, A., & Henderson, J. M. (2008). Taking a new look at looking at nothing. *Trends in Cognitive Science*, 12(11), 405-410.

Freksa, C., & Bertel, S. (2007). Eye movements and smart technology. *Computers in Biology and Medicine,* 37, 983 – 988.

Gbadamosi, J., & Zangemeister, W. H. (2001). Visual imagery in hemianopic patients, *Journal of Cognitive Neuroscience.* 13 (7), 45–56.

Gueugneau, N., Crognier, L., & Papaxanthis, C. (2008). The influence of eye movements on the temporal features of executed and imagined arm movements. *Brain Research*, 1187, 95–102.

Hebb, D. O. (1968). Concerning imagery. *Psychological Review*, 75, 466-477.

Heremans, E., Helsen, W. F., & Feys, P. (2008). The eyes as a mirror of our thoughts: quantification of motor imagery through eye movement registration, *Behavioural Brain Research,* 187(2), 351–360.

Hesslow, G. (2002). Conscious Thought as Simulation of Behaviour and Perception. *Trends in Cognitive Science*, 6, pp. 242-247.

Holsanova, J., Hedberg, B., & Nilsson, N. (1998). Visual and Verbal Focus Patterns when Describing Pictures. In Becker, Deubel & Mergner (Eds.), *Current Oculomotor Research: Physiological and Psychological Aspects*. Plenum: New York, London, Moscow.

Holsanova, J. (2008). *Discourse, vision, and cognition*. Human Cognitive Processes 23. John Benjamins Publishing Company: Amsterdam/Philadelphia.

Humphrey, K., & Underwood, G. (2008). Fixation sequences in imagery and in recognition during the processing of pictures of real-world scenes. *Journal of Eye Movement Research*, 2(2):3, 1-15.

Johansson, R., Holsanova, J., & Holmqvist, K. (2006). Pictures and spoken descriptions elicit similar eye movements during mental imagery, both in light and in complete darkness. *Cognitive Science,* 30:6, 1053-1079.

Laeng, B., & Teodorescu, D.-S. (2002). Eye scanpaths during visual imagery reenact those of perception of the same visual scene. *Cognitive Science*, 26, 207-231.

Richardson, D. C., Altmann, G. T. M., Spivey, M. J., & Hoover, M. A. (2009). Much ado about eye movements to nothing: a response to Ferreira *et al.,*: Taking a new look at looking at nothing, *Trends in Cognitive Science,* 13(6), 235-236.

Spivey, M., & Geng, J. (2001). Oculomotor mechanisms activated by imagery and memory: eye movements to absent objects, *Psychological Research*, 65, 235-241.

Spivey, M., Tyler, M., Richardson, D., & Young, E. (2000). Eye movements during comprehension of spoken scene descriptions. *Proceedings of the 22nd Annual Conference of the Cognitive Science Society* (pp. 487-492). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Thomas, N. J. T. (1999). Are Theories of Imagery Theories of Imagination? An Active Perception Approach to Conscious Mental Content. *Cognitive Science,* vol. 23 (2).

Thomas, L. E., & Lleras, A. (2009). Covert shifts of attention function as an implicit aid to insight. *Cognition*, 111, 168-174.

Wooding, D. S. (2002). Fixation maps: quantifying eye-movement traces. *Proceedings of the 2002 symposium on Eye tracking research & applications*, New Orleans, Louisiana.

Yoon, D., & Narayanan, N. H. (2004). Mental imagery in problem solving: An eye tracking study. *Proceedings of the Third ACM Symposium on Eye Tracking Research & Applications*, Association for Computing Machinery, ACM Press pp. 77-83.

The Role of Vagueness in the Numerical Translation of Verbal Probabilities: A Fuzzy Approach

Franziska Bocklisch[1] (franziska.bocklisch@psychologie.tu-chemnitz.de)
Steffen F. Bocklisch[1] (steffen.bocklisch@etit.tu-chemnitz.de)
Martin R. K. Baumann[2] (martin.baumann@dlr.de)
Agnes Scholz[1] (agnes.scholz@psychologie.tu-chemnitz.de)
Josef F. Krems[1] (josef.krems@psychologie.tu-chemnitz.de)
[1] Wilhelm-Raabe-Str. 43, Chemnitz University of Technology, Germany
[2] Lilienthalplatz 7, German Aerospace Center Braunschweig, Germany

Abstract

The paper describes a general two-step procedure for the numerical translation of linguistic terms using parametric fuzzy potential membership functions. In an empirical study 121 participants estimated numerical values that correspond to 13 verbal probability expressions. Among the estimates are the most typical numerical equivalent and the minimal and maximal values that just correspond to the given linguistic terms. These values serve as foundation for the proposed fuzzy approach. Positions and shapes of the resulting membership functions suggest that the verbal probability expressions are not distributed equidistantly along the probability scale and vary considerably in symmetry, vagueness and overlap. The role of vagueness for further investigations in reasoning and decision making is discussed and relations to knowledge representation and working memory are highlighted.

Keywords: verbal probability expressions; vagueness; fuzzy potential membership functions; knowledge representation; diagnostic reasoning; working memory

Introduction

Since the 1960s up to the present time researchers of different scientific areas have sustained an interest in studying the relationship between verbal and numerical probability expressions (Lichtenstein & Newman, 1967; Teigen & Brun, 2003; Smits & Hoorens, 2005). Among these are cognitive psychologists that inquire about the influence of uncertainty expressions on basic cognitive processes such as reasoning and decision making (Windschitl & Wells, 1996) as well as engineers, computer scientists and others that focus on the characterization (Zadeh, 1978, 2002) or on the treatment of uncertainty in applications such as medical decision support systems (Boegl, Adlassnig, Hayashi, Rothenfluh & Leitich, 2004). This broad interdisciplinary interest may be motivated by the essential role language plays in our daily life. Verbal probability terms, such as *probably* or *thinkable* are very widely used to express uncertainty about the occurrence of future events or about the degree of belief in hypotheses. For example, a typical statement that illustrates the use of linguistic terms in the conversation of stock market traders could be: "It is *very unlikely* that there will be a significant increase in the price of oil in the next month vice future.". Several studies consistently show that people prefer words over numbers to express uncertainty (e.g. Wallsten, Budescu, Zwick & Kemp, 1993). This preference may be explained by the possibility of saying something about two different kinds of subjective uncertainty by using only one word. First, the stochastic uncertainty about the occurrence of an event (e.g. the probability of an increase of the oil price) and second, the vagueness of the event (e.g. what is meant by "a significant increase").

The understanding of these two kinds of uncertainty, their relations to each other and the way in which they influence human reasoning and decision making is crucial for any application that aims to support decision makers for example in medicine, business, risk management, marketing or politics. In our view, in order to contribute to the understanding of uncertainty, it is essential to first uncover the underlying relationship between word meaning and mathematical concepts such as subjective probability or fuzzy membership. Therefore, we propose a general two-step procedure for the numerical translation of verbal probability expressions based on (1) empirical estimates modelled by (2) fuzzy membership functions (Zadeh, 1965, Bocklisch & Bitterlich, 1994).

The paper is structured as follows: first, we compare verbal and numerical probability expressions and discuss existing translation approaches. Second, we present our proposal that goes beyond other methodical issues and the results of an empirical investigation. Thereafter, the results are discussed and conclusions (e.g. for the construction of verbal probability scales for questionnaires) are highlighted. Further, potentialities of the fuzzy pattern classification method for reasoning and decision processes are pointed out.

Verbal and Numerical Probabilities

There is broad agreement concerning the different features of verbal and numerical expressions (see Teigen & Brun, 2003 for an overview). Numerical probabilities are commonly described as precise, unambiguous and especially useful for calculations. Additionally, the quality of numerical expressions can be evaluated and compared to predictions of normative models such as Bayes nets. Currently many researchers in the area of cognitive

psychology utilize subjective probabilities for the modelling of human reasoning (e.g. Bayes nets in inductive learning and reasoning (Tenenbaum, Griffiths & Kemp, 2006)). This approach is very fruitful and the obtained results contribute highly to the understanding of psychological processes but, at the same time, it focuses only on the probability dimension of uncertainty. Generally, vagueness is another facet of people's subjective uncertainty and should not be neglected. The effects of vagueness, such as exemplarily described by Kuhn and Budescu (1996) for hazard risk decisions, have received much less research attention in psychology. Although it is investigated more in engineering and other domains, where the practical significance is clearly observable from its prevalence in real-world decisions, vagueness is also crucial for psychological approaches. Zadeh (1965) proposed the fuzzy framework for the handling of vagueness and pointed out that probability theory and fuzzy approaches are complementary rather than competitive (Zadeh, 1995). Hence, it is possible to combine probability and fuzzy accounts and the advantages of bridging the gaps have been discussed recently (Singpurwalla & Booker, 2004).

In contrast to numerical probabilities, probability words are vague, with ambiguous meaning. They cannot be easily used for calculations and their meaning is often only clarified by means of a context (such as domain, speakers' prior knowledge and experience, reference point or prior probabilities and base rates of events). Nevertheless, most people in most everyday situations use words rather than numbers when describing their own uncertainty. Words are perceived as more natural, easier to understand and communicate and they are useful in situations when uncertainty can not at all be verbalized exactly. Numerical and verbal expressions are closely associated and refer to the underlying concept of probability and there is evidence that people can use numbers and words interchangeably (Jaffe-Katz, Budescu & Wallsten, 1989). But, at the same time, words and numbers do not mean exactly the same thing.

Furthermore, it can be assumed from various experiments that the use of numbers versus words affects human reasoning processes under certain circumstances. Windschitl and Wells (1996) show that numeric measures of uncertainty tend to sway people toward rule-based, deliberate thinking, whereas verbal expressions tend to elicit more associative and intuitive reasoning. These findings are of particular importance for reasoning situations that create conflicts between logical reasoning and intuitive beliefs (e.g. the belief-bias effect (Evans, 2003)).

In belief updating processes, such as customers product evaluation, there is evidence for the influence of information format (verbal vs. numerical) on order effects. Shen and Hue (2007) report that numerical information lead to order effects whereas verbal expressions do not. It can be assumed that the utilization of numerical vs. verbal expression formats result in different cognitive processes that in turn have different consequences for decisions.

Translating Words Into Numbers

In order to investigate the impact of verbal versus numerical probability expressions on order effects, decision making and the communication of uncertainty methods have to be developed for the "translation" of verbal into numerical expressions. There are already a number of translation studies that utilized different estimation and translation procedures. Among these are empirical approaches using direct estimation techniques for instance on a scale from 0 to 100 (Beyth-Marom, 1982) or pair comparison methods (Wallsten, Budescu, Rapoport, Zwick & Forsyth, 1986) as well as expert consultations for example to create knowledge bases for decision support systems (Boegl et al., 2004). A summary and discussion of different estimation approaches, that map verbal probabilities onto a numerical probability scale, is provided by Teigen and Brun (2003).

Recurrent findings in the studies using empirical estimations are that the mean estimates of the verbal probability expressions are reasonably similar supporting the idea that words are translatable. At the same time, there is a large variability between individuals indicating inconsistency in word understanding which may lead to communication problems. Although there are different views on whether verbal probability expressions are quantifiable or not (Teigen & Brun, 2003), we agree with Budescu et al. (2003). They propose to treat probability words as fuzzy sets and use fuzzy membership functions (MFs) over the probability scale to represent their vague meanings. They elicited judgments of membership by using a multiple stimuli estimation method in which probability values (0, 0.1, …, 0.9, 1) are presented simultaneously with a verbal probability expression. Their results show, that the peak value and skew of the MF describing a probability expression depends on the words meaning. Therefore, they conclude that properties of the MF can predict for example the directionality (positive vs. negative verbal expressions, such as probable vs. improbable) of probability words.

Objective of the Paper

This paper has the goal to present a general two-step procedure for the numerical translation of linguistic terms. It is composed of (1) a direct empirical estimation method that yields numerical data participants assigned to presented words and (2) a fuzzy approach for the analysis of the data resulting in parametric membership functions (MFs) of the potential type (Bocklisch & Bitterlich, 1994). We outline this method for verbal probability expressions (e.g. *possible*) but the proposed procedure can also be applied for other linguistic terms such as expressions of frequency (e.g. *occasionally*), strength (e.g. *strong*) or others and is therefore of potential interest for many research areas and applications. Furthermore, our method goes beyond existing approaches for two reasons: first, the presented direct estimation method is frugal, efficient and easy to use to yield data from human decision makers. Therefore, it is suitable for research purposes and especially for applications where expert knowledge is crucial but also rare

or expensive. Second, the proposed parametric MFs of the potential type bring along advantages compared to other MFs (Zadeh, 1965; Budescu et al., 2003). For instance, they are able to account for asymmetric probability terms and are defined continuously over the numerical probability scale. Hence, linguistic terms can be modelled very realistically. In addition, the MFs can be implemented directly in applications (e.g. in fuzzy decision support systems) and the fuzzy pattern classification approach has potentials for psychological research (see Future Prospects at the end of this paper).

In contrast to Boegl et al. (2004) we do not expect that the MFs of the probability words are distributed equidistantly along the numerical probability scale and just like Budescu et al. (2003) we predict the functions to be asymmetric in shape.

Two-Step Translation Procedure

In this section we present the details of the two-step translation procedure for the numerical translation of verbal probability expressions. At first, the estimation technique and the method we used in the empirical study is outlined. Thereafter, the fuzzy analysis and the MFs are specified.

Empirical Investigation

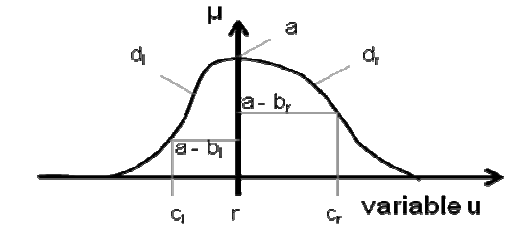
Participants. 121 participants (19 males) took part in the study mainly for exchange of credits. The majority were undergraduate students of the Universities of Chemnitz, Göttingen and Zurich with an average age of 23 years (*SD*=4.6).

Materials and Procedure. Participants read a short contextual story from the area of medical decision making and were requested to take over the perspective of a physician. Then they assigned three numerical values to each of 13 exemplars of probability words (see translated words in Table 1, the original material was presented in German language) that were chosen from previous studies (e.g. Budescu et al., 2003). Among the three numerical values that had to be estimated were: (1) the one that represents the given probability word best and the (2) minimal and (3) maximal values that just correspond. The estimations can be interpreted according to the semantic meaning of the words: the first value characterizes the most typical numerical equivalent for the word, whereas the other values indicate the lower and upper border of the verbal probability expression. Participants were instructed to give their estimates in the frequency format (e.g. "In how many of 100 cases a certain diagnosis is correct if it is for instance *improbable*?"). This frequency format of estimation was proved to be better than for instance the estimation of percentages (Gigerenzer & Hoffrage, 1998). Participants used a PDF online questionnaire to provide their estimates.

Fuzzy Analysis

Fuzzy Membership Functions. Membership functions are truth value functions. The membership value (μ) represents the value of truth that an object belongs to a specific class (e.g. that the numerical probability value 0.25 belongs to the word *doubtful*). For the analysis of the empirical data provided by the 121 participants a parametric membership function of the potential type (Bocklisch & Bitterlich, 1994; Hempel & Bocklisch, 2009) was used.

This function (see Figure 1) is based on a set of eight parameters: r marks the position of the mean value, a is representing the maximum value of the membership function. Regarding a class structure, a expresses the "weight" of the class in the given structure (we use a fixed a=1 in this investigation). The parameters b_l and b_r assign left and right-sided membership values at the borders of the function. Hence, they represent the border memberships whereas c_l and c_r characterize the left and right-sided expansions of the class and therefore mark the range of the class (in a crisp sense). The parameters d_l and d_r specify the continuous decline of the membership function starting from the class centre, being denoted as representative of a class. They determine the shape of the function and hence the fuzziness of the class.



$$\mu(u) = \begin{cases} \dfrac{a}{1 + (\frac{1}{bl} - 1)(\frac{r-u}{cl})^{dl}} & u \leq r \\ \dfrac{a}{1 + (\frac{1}{br} - 1)(\frac{u-r}{cr})^{dr}} & u \geq r \end{cases}$$

Figure 1: Parameters of the membership function (for r=0)

A continuous range of membership functions, varying from a high degree of fuzziness to crisp, is available. This function type allows considering asymmetry in fuzzy classes by individual parameters for the left and right hand branches of the function. As we expect the MFs for the probability expressions to be asymmetric, this feature is especially important for the present study.

Results

In this paragraph we present the results of the statistical and fuzzy analysis of the present study. The descriptive statistics were calculated with the help of SPSS software. For the fuzzy analysis and the modelling of the MFs a software package (Fuzzy Toolbox, 2008) was used.

Descriptive Statistics

Table 1 shows the descriptive statistics for the empirical estimates of the most typical values that correspond to the

presented words. The minimal and maximal estimates, that indicate the borders of the semantic meaning of the linguistic terms, were necessary for modelling the MFs.

Results show that the probability words are distributed all over the numerical probability scale with varying distances. The standard deviation and kurtosis show a systematic pattern: probability words near to the borders of the numerical probability scale (e.g. *impossible* and *certain*) have small standard deviations but high values of kurtosis. And probability words in the middle (e.g. *thinkable* and *possible*) offer a larger spread but smaller kurtosis values. Also systematic differences exist for the skew indicating that probability expressions with means smaller than P=0.5 are skewed to the right whereas words with means higher than P=0.5 are asymmetric to the left. These findings are consistent with the results reported by Budescu et al. (2003).

Table 1. *Descriptive statistics for the estimates (most typical values)*

probability words	Mean	SD	Skew	Kurtosis
Impossible	1.44	3.01	3.25	13.39
very improbable	5.53	5.48	1.71	2.72
quite improbable	9.99	7.94	1.42	2.2
Improbable	11.68	9.03	1.43	1.82
hardly probable	17.01	11.05	1.15	1.02
sparsely probable	18.57	12.19	1.12	.89
Doubtful	21.34	13.61	.72	.32
Thinkable	49.33	20.24	.35	.1
Possible	51.49	21.6	.54	.53
Probable	67.68	12.49	-.01	-.85
quite probable	75.07	12.89	-1.01	1.02
very probable	83.95	9.08	-1.02	1.2
Certain	96.28	6.45	-2.87	9.99

Fuzzy Analysis

Figure 2 shows the MFs for the 13 verbal probability expressions. The representative values (r) indicating the highest memberships are identical to the reported means in Table 1.

Obviously, the functions differ considerably in shape, symmetry, overlap and vagueness. Functions at the borders (e.g. *impossible*) are narrower than those in the middle (e.g. *thinkable*) which is consistent with the observed standard deviations and kurtosis values. Most functions are asymmetric and are not distributed equidistantly along the probability scale. From the functions' positions, three clusters arise, that may be described by (1) *low* (MFs 1-7), (2) *medium* (MFs 8 and 9) and (3) *high* (MFs 10 - 13) probability ranges. The 13 MFs overlap in large parts and especially when they belong to the same cluster.

To test whether the probability expressions are distinct or not, participants' estimates were reclassified. Table 2 shows the results of the reclassification.

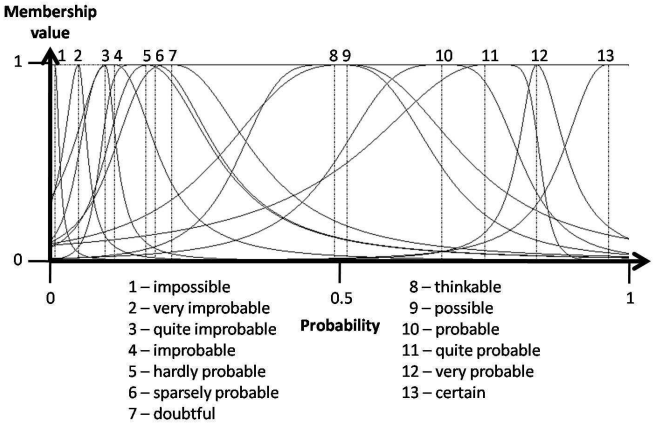


Figure 2: Membership functions of the 13 verbal probability expressions

The second column of the table presents percentages of the corresponding estimation data that was reclassified correctly. According to these results, some of the probability words are unambiguous and the reclassification was very successful (e.g. *certain*; 93.5% reclassified correctly). Others are inconclusive and almost no estimation data point that was used to describe the MF was reclassified correctly (e.g. *improbable*; 2.5 % classified correctly). Instead, the data was classified as belonging to the neighboring functions.

Table 2. *Percentages correct reclassification*

probability words	Scale (13)	Scale (5)
impossible	80.0	95.0
very improbable	33.1	
quite improbable	24.8	
improbable	2.5	
hardly probable	15.1	
sparsely probable	2.5	
doubtful	42.4	77.1
thinkable	41.2	61.3
possible	6.6	
probable	44.2	72.5
quite probable	33.9	
very probable	18.4	
certain	93.5	93.5

For a verbal probability scale that could be employed in psychological research or application, a scale with 13 probability words would not be useful because the words are too indifferent according to their meanings. But if a few words with small overlaps are selected, it is possible to create a scale that differentiates very well (see reclassification rate computed by the Fuzzy Toolbox Software in column three of Table 2). Figure 3 shows an example scale with five probability words described by their MFs.

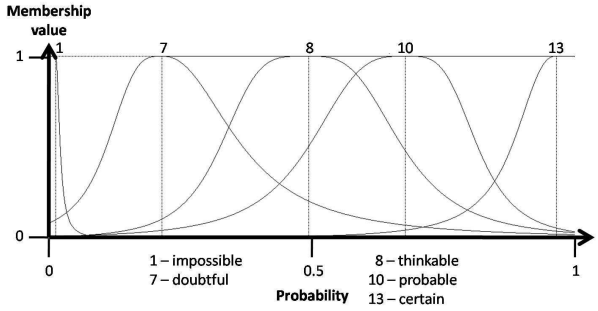


Figure 3: Membership functions of 5 selected verbal probability expressions

Discussion

This paper aims to present a two-step procedure for the numerical translation of linguistic terms that goes beyond existing approaches. First of all, the estimation of three numerical values for each linguistic term (the most typical, minimal and maximal corresponding values) is very frugal and data can be gained very efficiently, whereas most alternative procedures are more costly (Budescu et al., 2003). The resulting estimation data can be analyzed using the proposed parametric MFs of the potential type. Results show, that the functions are able to model the data in a very efficient way, creating averaged membership functions that describe the linguistic terms continuously over the numerical probability scale. Because of the eight parameters, the functions take into account asymmetry, which was indeed found in the empirical data. Parametric MFs with fewer parameters would model the data without considering asymmetry and would therefore be less accurate and suitable for the reported data. Another advantage of the proposed function type is that the parameters can be interpreted in terms of content on a semantic meta level and illustrate the vague meaning of probability words very realistically.

Large overlaps of the functions (see Figure 2) indicate that the words are very similar in their meanings. Despite the imprecision of natural language, the MFs allow identifying words that are more distinct in their meaning than others. Just as Dhami and Wallsten (2005) we also found five probability expressions (see Figure 3) that are sufficiently distinct. This is especially useful for the creation of verbal probability scales for purposes of research and application that should include unambiguous words when possible.

Finally, the presented translation procedure serves as foundation for future investigations concerning the influence of contexts on word understanding. This influence can then be quantified by changes in the parameters defining the MFs. As these parameters can be semantically interpreted the influence of context on the interpretation of the expressions can be investigated in detailed way. As Wallsten and Budescu (1990) claimed, it is a promising instrument to uncover the various communication roles that probability phrases serve. For instance, it is likely that some of the ambiguous probability words are clarified by the context in which they are used and therefore will become less vague which can be observed in the MFs.

Future Prospects

Finally, we will present a short outlook that highlights the potentials of the fuzzy approach for further psychological research in the area of diagnostic reasoning and decision making.

An advantage of the proposed MFs and the underlying fuzzy pattern classification method (Bocklisch & Bitterlich, 1994) is that the functions serve for the representation and combination of various kinds of vague knowledge (e.g. fuzzy degrees of symptom intensity such as "*high* fever" or "*low* blood pressure") in a multidimensional way. For example, a physician considering the likelihood that a patient has a certain disease presumably takes into account the intensity of two (or more) present symptoms in combination prior to stating the diagnosis. Figure 4 exemplifies the content of a possible mental model in a simplified manner: three fuzzy classes (diseases A, B and C) resulting out of the multivariate combination of two features (intensities of the symptoms 1 and 2) that are described by fuzzy potential membership functions.

Furthermore, it is possible to integrate both vague and crisp information (such as precise predictions of probabilistic models) in this framework.

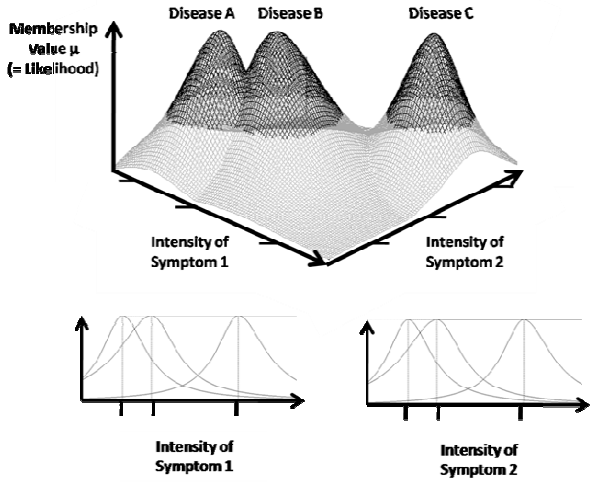


Figure 4: Representation of medical knowledge using fuzzy pattern classification method

The distance of the classes as well as their overlap can be interpreted in terms of similarity (disease classes A and B are near to each other and therefore cause similar symptom intensities, whereas disease C is apart and less similar to the other diseases). Furthermore, shapes and positions of the classes provide information about the discriminability of items in working memory which in turn affects reasoning performance. According to Oberauer, Süß, Wilhelm and Wittman (2003), the coordination function of working memory (WM) allows the integration of information (such as symptoms in a diagnostic reasoning process). Therefore,

WM provides simultaneous access to independently varying elements (such as symptoms and diseases) by placing them in a common coordinate system. The coordinate system has limited capacity to hold information and keep them separated from each other. Hence, it is likely that the precision or vagueness of the information elements (as it is described by the MFs) is an important variable influencing diagnostic reasoning processes and decision making performance. Moreover, it seems possible to predict to which extent relevant and irrelevant diagnostic hypotheses will interfere during the reasoning process (Dougherty & Sprenger, 2006) from the fuzzy knowledge representation. For example, it is plausible to assume that irrelevant diagnostic hypotheses that show a strong overlap with the relevant ones interfere more than irrelevant hypotheses that show less overlap. And the overlap can be quantified with this fuzzy approach. This is currently the object of further investigation.

References

Beyth-Marom, R. (1982). How Probable is Probable? A Numerical Translation of Verbal Probability Expressions. *Journal of Forecasting, 1*, 257-269.

Bocklisch, S.F. (2008). *Handbook Fuzzy Toolbox. GWT-TUDmbH*, Department of Electrical Engineering, Systems Theory, Chemnitz University of Technology, Germany

Bocklisch, S.F. & Bitterlich, N. (1994). Fuzzy pattern classification – methodology and application. In Kruse, R., Gebhardt, J., & Palm, R. (Eds.) *Fuzzy Systems in Computer Science*. Vieweg.

Boegl, K., Adlassnig, K.-P., Hayashi, Y., Rothenfluh, T.E., & Leitich, H. (2004). Knowledge acquisition in the fuzzy knowledge representation framework of a medical consultation system. *Artificial Intelligence in Medicine, 30*, 1-26.

Budescu, D.V., Karelitz, T.M., & Wallsten, T.S. (2003). Predicting the Directionality of Probability Words from Their Membership Functions. *Journal of Behavioral Decision Making, 16*, 159-180.

Dhami, M.K. & Wallsten, T.S. (2005). Interpersonal comparison of subjective probabilities: Towards translating linguistic probabilities. *Memory & cognition, 33(6)*, 1057-1068.

Dougherty, M.R. & Sprenger, A. (2006). The Influence of Improper Sets of Information on Judgment: How Irrelevant Information Can Bias Judged Probability. *Journal of Experimental Psychology: General, 135(2)*, 262-281.

Evans, J.S.B.T. (2003). In two minds: dual-process accounts of reasoning. *Trends in Cognitive Sciences, 7 (10)*, 454-459.

Gigerenzer, G. & Hoffrage, U. (1998). Using Natural Frequencies to Improve Diagnostic Inferences. *Academic Medicine, 73(5)*, 538-540.

Hempel, A.-J. & Bocklisch, S.F. (2009). Parametric Fuzzy Modelling for Complex Data-Inherent Structures. In *Proceedings of the Joint 2009 International Fuzzy Systems Association World Congress and 2009 European Society of Fuzzy Logic and Technology Conference (IFSA-EUSFLAT 2009)*, 885-890.

Jaffe-Katz, A., Budescu, D.V., & Wallsten, T.S. (1989). Timed magnitude comparisons of numerical and nonnumerical expressions of uncertainty. *Memory & Cognition, 17*, 249-264.

Kuhn, K.M. & Budescu, D.V. (1996). The Relative Importance of Probabilities, Outcomes, and Vagueness in Hazard Risk Decisions. *Organizational Behavior and Human Decision Processes, 68(3)*, 301-317.

Lichtenstein, S. & Newman, J.R. (1967). Empirical scaling of common verbal phrases associated with numerical probabilities. *Psychonomic Science, 9*, 563-564.

Oberauer, K., Süß, H.-M., Wilhelm, O., & Wittman, W.W. (2003). The multiple faces of working memory: Storage, processing, supervision, and coordination. *Intelligence 31*, 167-193.

Shen, Y.-C. & Hue, C.-W. (2007). The role of information presentation formats in belief-updating. *International Journal of Psychology, 42(3)*, 189-199.

Singpurwalla, N.D., & Booker, J.M. (2004). Membership Functions and Probability Measures of Fuzzy Sets. *Journal of the American Statistical Association, 99 (467)*, 867-877.

Smits, T. & Hoorens, V. (2005). How Probable is Probably? It Depends on Whom You're Talking About. *Journal of Behavioral Decision Making, 18*, 83-96.

Teigen, K.H. & Brun, W. (2003). Verbal Expressions of Uncertainty and Probability. In D. Hardman (Edt.): Thinking: *Psychological Perspectives on Reasoning, Judgment and Decision Making*. Chapter 7, Wiley and Sons.

Tenenbaum, J.B., Griffiths, T.L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences, 10 (7)*, 309-318.

Wallsten, T.S., Budescu, D.V. (1990). Comment. *Statistical Science 5(1)*, 23-26.

Wallsten, T.S., Budescu, D.V., Rapoport, A., Zwick, R., & Forsyth, B. (1986). Measuring the Vague Meanings of Probability Terms. *Journal of Experimental Psychology: General, 115*, 348-365.

Wallsten, T.S., Budescu, D.V., Zwick, R., & Kemp, S.M. (1993). Preferences and reasons for communicating probabilistic information in numerical or verbal terms. *Bullet of the Psychonomic Society, 31*, 135-138.

Windschitl, P.D., & Wells, G.L. (1996). Measuring psychological uncertainty: Verbal versus numeric methods. *Journal of Experimental Psychology: Applied, 2*, 343-364.

Zadeh, L.A. (1978). Fuzzy Sets as a basis for a theory of possibility. *Fuzzy Sets and Systems, 1*, 3-28.

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control 8*, 338-353.

Zadeh, L.A. (1995). Discussion: Probability Theory and Fuzzy Logic Are Complementary Rather Than Competitive. *Technometrics, 37*, 271-276.

Selective attention and development of categorization: An eye tracking study

Xin Yao (yao.64@osu.edu)
Center for Cognitive Science
The Ohio State University
209C Ohio Stadium East, 1961 Tuttle Park Place
Columbus, OH 43210, USA

Vladimir M. Sloutsky (sloutsky.1@osu.edu)
Center for Cognitive Science
The Ohio State University
208C Ohio Stadium East, 1961 Tuttle Park Place
Columbus, OH 43210, USA

Abstract

Some researchers argue that categorization in early development is knowledge-based rather than perceptually based. This approach requires young children to be able to attend to unobservable properties instead of perceptual features, which are usually more salient. However, potential immaturity of selective attention makes this possibility questionable. Current study tested both young children and adults with a match-to-sample task in which perceptual features were in conflict with the matching rule. Both behavioral and eye tracking data were collected. Eye-tracking results suggested that young children (3- and 4-year-olds) could not inhibit attention to the perceptual features, although behaviorally, 4-year-olds could. These findings are discussed with respect to theoretical accounts of category learning in early development.

Keywords: Cognitive Development, Categorization, Attention, Psychology, Human Experimentation.

Introduction

The ability to learn categories is a critical component of human cognition and this ability is present early in development (e.g., see Eimas & Quinn, 1994; Madole & Oakes, 1999, for reviews). However the mechanisms underlying category learning remain highly contested. Some researchers argue that early categories are perceptually-based, whereas other argue that even early in development, unobservable conceptual properties (such as animacy) play an important role in infants and young children's category learning and category use (see Rakison & Poulin-Dubois, 2001; Sloutsky, in press, for reviews). According to the latter view, early categorization (some have argued that as early as at 7 months of age) is based on features that are not given directly in the input. However, to be able to do so, infants and young children have to be able to selectively attend to these unobservable properties. This problem is particularly evident when salient perceptual features are in conflict with less salient, often unobservable, "conceptual" features. For example Gelman & Markman (1986) presented 4-year-olds with an inductive inference task. The task was structured as a match to sample triad, such that one of the items belonged to the same kind as the target (but was dissimilar) and another looked similarly (but belonged to a different kind. The authors argued that the unobservable conceptual feature (i.e., taxonomic kind) would override the salient observable features (e.g., appearance similarity). In this case, in addition to the ability to attend selectively to less salient input, young children should also have the ability to inhibit more salient (yet irrelevant) choice option. Given the critical immaturities in the executive function early in development (see Rueda, Fan, McCandliss, Halparin, Gruber, Lercari, & Posner, 2004, Davidson et al., 2006, for reviews), such selectivity seems questionable.

Current research addresses this issue by presenting participants with a simple match-to-sample task and examining their eye movement in the course of the task. This task is substantially simpler than the match-to-sample task used by Gelman and Markman (1986). First, in the current task, participants were explicitly told which aspect of the stimuli they should focus on. And second, instead of pitting appearance versus unobservable properties, we pitted more salient features against less salient ones. Our reasoning was as follows. If participants focus on unobservable information in a more difficult induction task, they should have no difficulty focusing on less salient information in this highly simplified task.

The task includes a target and two test items. There are three within-subjects conditions. In the Supportive condition, the test item that shares the matching rule with the target is also similar to the target. In the Neutral condition, both items are equally similar to the target, with one test item sharing the matching rule. And finally, in the Conflict condition, one test item shares the matching rule, whereas the other one looks similar to the target. Therefore, the latter condition required participants to reject a salient appearance-based item in favor of less salient rule-based item. In sum, the task requires the ability to attend selectively that is critical for many category learning and inductive inference tasks. Given that the task is exceedingly simple, participants' failure in the conflict condition might be particularly informative. If they cannot resolve the conflict in this simple task, it is reasonable to ask: how could they resolve a conflict in more difficult and demanding categorization and induction tasks?

Experiment 1

Method

Participants Sixteen adults (6 women and 10 men, *M* = 20.1 years, *SD* = 2.7 years) participated in this experiment. Adults were undergraduate students from The Ohio State University participating in the experiment for course credit. The experiment used a within subject design and each subjects took all the three conditions in the experiment: Supportive, Neutral, and Conflict conditions. All the participants were tested in a quiet room on campus.

Stimuli consisted of triads of artificial creatures, which were irrelevant components of the task. Each triad also included three rows of circles (referred to as cookies that creatures eat). Examples of stimulus triads are presented in Figures 1-3. These cookies were the critical features that participants were instructed to focus on in the current matching task. To make the irrelevant features perceptually more salient, creatures were bigger and colorful, while the critical features were smaller and shared the same color. The only difference for the critical features was different patterns on the cookies. Two had wave lines on them while the remaining one had diagonal lines. The irrelevant features were drawn from two categories. One category consisted of objects with hands and feet, and the other consisted of bug-like objects with wings and tails. In each triad, the bottom object was the target item, and the two top ones were test items. Half of target and test items were selected from one category and half were selected from the other category. The top two sets of cookies were always different with only one matching the target set. At the same time, irrelevant items varied across the conditions. In the Supportive condition, the "matching distracter" (i.e., the one that had the same kind of cookies as the target) came from the same category as the target. So the one that looked more similar to the target item also shared the matching rule with the target item. Therefore, the perceptually irrelevant information was consistent with and supportive of the critical features. In the Conflict condition, the "matching distracter" came from the opposite category than the target distracter. So the perceptual information was in conflict with the matching rule. Finally, in the Neutral condition, both test distracters and the target distracter came from the same category. As a result, the matching rule was neither supported, nor in conflict. The right and left sides of the stimuli were counterbalanced.

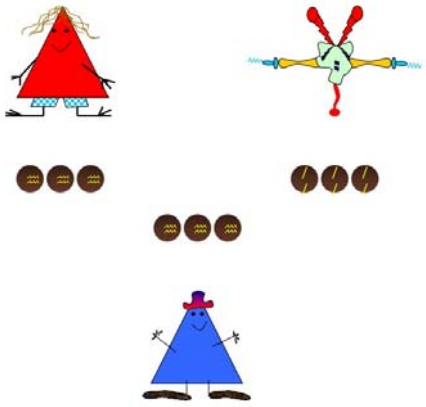


Figure 1: An example of the stimuli in the supportive condition

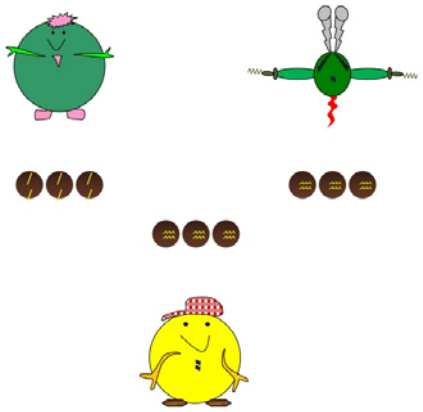


Figure 2: An example in the conflict condition

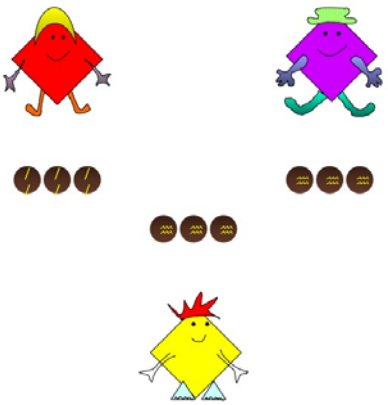


Figure 3: An example in the neutral condition

The locations of the cookies and the creatures were fixed for each trial. The distracters were subtended at visual angles equaling to 6.2° horizontally and 5.2° vertically. The cookies were subtended at visual angles equaling to 4.2° horizontally and 1.6° vertically. The distance between the creatures and the cookies were 2.6° vertically.

Procedure Eprime 2.0 was used for controlling the experiment and Tobii T60 with the sampling rate of 60 Hz was used for collecting eye tracking data.

Before the task, the eye tracker was calibrated to each participant. Participants were told that in this matching task they should choose one of the objects on the top to match the object at the bottom by matching the cookies. They were also instructed to make the choice as quickly as possible. If it was the left creature that matched, they should press "1", and press "4" if it was the right one. They were given the following instructions: *This is a matching game. The game is to decide which one on the top goes with the one at the bottom. To win the game, you need to choose the one likes the same as cookies as the one as the bottom.*

Prior to testing, participants had three warm-up trials at first, one for each condition. Feedback was provided for the three warm-up trials. During the test phase, participants were given 30 trials, with 10 Supportive, 10 Neutral, and 10 Conflict trials. The trials were mixed and pseudo-randomly assigned into 3 blocks, with 10 trials in each block. The order of the three blocks and the order of the trials within each blocks were randomized. Each trial was preceded by a fixation point at the center of the screen. The duration of the fixation varied between 300 ms to 800 ms. No feedback was provided during the test phase.

Eye tracking Dependent Variables A stream of eye fixations corresponding to their x-y locations on the screen were collected by the eye tracking software for each subject. Six areas of interest (AOIs) for fixations were defined: three circular areas encompassing the creatures and three rectangular areas encompassing the cookies displayed on the screen. All fixations outside the AOIs were discarded.

Results and Discussion

Behavioral Data The average of accuracy across the three conditions was 97% (SE = 2.1%) and exceeded chance level, one-sample t compared to 50%, t (15) = 22.94 p = .01. No difference was found between different conditions, F (2, 30) = .92, p = .41.

Eye Tracking Data The primary analyses focused on the proportion of the eye fixation on the critical features, which were the kinds of cookies in this study. The proportion was calculated by total fixations on the triads of cookies divided by the sum of fixations on the triads of cookies and the fixations on the triads of creatures. The absence of a preference would result in comparable looking across the areas of interest. Before 200 ms, all the eye fixations were at the center of the screen which indicated that participants did focus on the fixation stimulus and did not exhibit eye movements during that period. The time window for eye tracking analysis was two standard deviations above the mean reaction time (M = 1013.8 ms, SD = 480.7). Therefore, the time window for eye tracking analyses was between 200 ms and 2000ms. The proportions of looking at the critical feature in the Conflict condition across time are presented in Figure 5. The overall proportion of looking at the critical features, i.e., the cookies, was 84.4% (SE= 5%). No difference was found across the three conditions, F (2, 30) = 765, p = 474. Perhaps not surprisingly, these findings indicate that adults had little difficulty focusing on the critical features and ignoring more salient distracters. As a result, participants exhibited near ceiling accuracy in all three conditions. The importance of these data is that they represent a necessary point of comparison for children's data. Experiment 2 focused on performance of 3- and 4-year-old children.

Experiment 2

Method

Participants Young children were recruited from the suburbs of Columbus, Ohio. There are 15 4 year olds (9 girls and 6 boys, M = 50.5 months, SD = 2.5 months) and 15 3 year olds (8 girls and 7 boys M = 41.8 months, SD = 3.4 months). All participants were tested in a lab on campus.

Procedure The procedure for young children was almost identical to that for adults, except for the following differences. First, a female experimenter presented the task to the participants, controlled the pace of the experiment, and pressed the key based on children's verbal response during the experiment. And second, the instructions "Choose the one that likes the same kind of cookies as the one at the bottom" were repeated before each trial.

Results and Discussion

Behavioral Data Accuracy data are presented in figure 4. For 3-year-olds, difference was found in accuracy across the three conditions, F (2, 28) = 6.81, p < .01. Specifically, accuracy in the conflict condition did not exceed chance, one-sample t compared to 50%, t (14) = 1.56, p = .14, two-tailed. However, the accuracy was above chance in the neutral and supportive condition, ts (14) > 5.78, ps < .01. For 4-year-olds, accuracy for all the three conditions exceeded chance, one-sample t compared to 50%, ts (14) >3.67, ps < .01, one-tailed. Difference was also found across the three conditions, F (2, 28) = 3.7, p = .037. In particular, participants were less accurate in the conflict condition than the other two conditions.

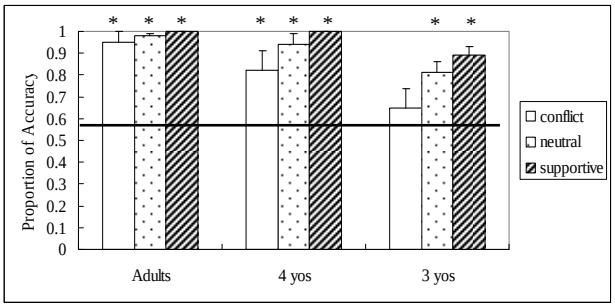


Figure 4 Behavioral data of different age groups
Note: * -- Above chance, p < .05.

Eye Tracking Data The time stream between 1000ms and 3000ms was used for analysis. Before 1000ms, eye fixations did not reliably move from the center of the screen to the areas of interest. Data were analyzing by averaging across trials and individuals. The proportions of looking at the critical feature in the Conflict condition by sampling rate (16 ms) and age are presented in Figure 6. Main effect of age was found in proportion of looking at the critical feature, F (1, 238) = 408.219, p < .01. 4 year olds showed more fixations on the critical features than 3 year olds. Difference between conditions was found, F (2, 476) = 109.4, p < .01. There was an age by condition interaction, F (2, 476) = 14.94, p < .01, with larger age difference in looking at the critical feature found in Conflict condition. Therefore, 4-year-olds were not only more accurate in the conflict condition, but also were more likely to look at the critical feature in the Conflict condition. At the same time, the proportion of looking at critical features by 3- and 4-year-olds was consistently below 50%.

Individual patterns of responses were also analyzed. We were particularly interested whether individuals who were more likely to look at the critical features in the Conflict condition also exhibited greater accuracy. For 3-year-olds, a significant correlation was observed between the accuracy and the overall proportion of looking at the critical features in the Conflict condition (r = .574, p = .03). This indicated that accurate participants were more likely to pay attention to the critical features. However, there was no significant correlation in 4-year-olds, r = .29, p = .29. This is probably because there was very little variability in the accuracy of 4-year-olds.

To further examine the connection between looking and response accuracy, we split the children into two groups according to their accuracy in the Conflict condition. Those with accuracy above .5 were assigned to the high accuracy group, and those with accuracy below or equal to .5 were assigned to the low accuracy group. Difference in overall proportion of looking was found in conflict condition between these two groups. Those high accuracy children were more likely to focus on the critical features (M = 36.1%, SE = 5%) than those low accuracy children (M = 16.5%, SE = 6%), t (28) = 2.26, p = .016, one-tailed.

Further analysis was carried out for examing the online learning during the task. If there was any learning or strategy optimization happening during the task, we should expect the difference in looking across trials. The participants should show more looking to the critical featuers during the later part of the task than during the earlier part. To test this, data was divided into the earlier 5 and later 5 trials of each condition. Comparison between these two half of the task were made for each condition and each age group. However, no difference was found, t (14) < .05, p > .16, one-tailed. Therefore, in the absence of feedback given to participants, there was little evidence of on-line learning to allocate attention to critical features.

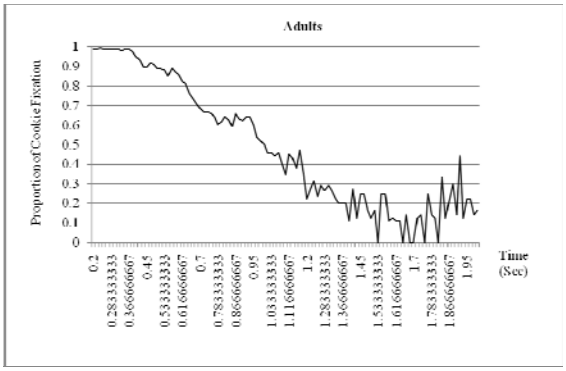


Figure 5: Adults' eye tracking data in Conflict condition

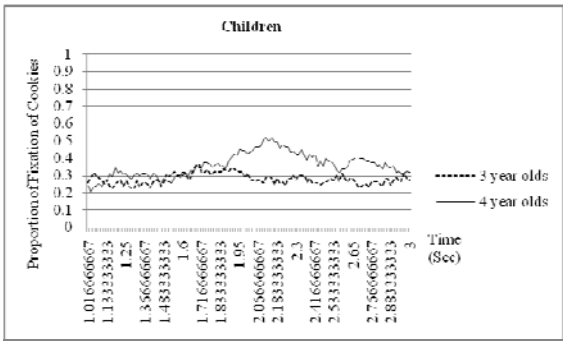


Figure 6: Childrens' eye tracking data in Conflict condition

General Discussion

The results point to several important findings. First, under comparable experimental conditions, adult participants and 4-year-olds were less likely to be distracted by the appearance of the stimuli, which were the irrelevant features in this study. When the critical features conflicted with the irrelevant features, their behavioral performance was still above chance, although performance of 4-year-olds (but not of adults) decreased in the Conflict condition. However 3-year-olds could not ignore the irrelevant features and their performance was at chance in the Conflict condition.

Second, performance in the Conflict condition was associated with the proportion of looking to the critical feature. This proportion in adults was greater than in 4-year-olds, and in 4-year-olds greater than in 3-year-olds. In addition in 3-year-olds, correlation was found between the proportion of looking to the critical features and the accuracy on the task in the Conflict condition. Moreover, when children were divided into the groups by their performance in conflict condition, difference was observed in their looking pattern. Thus, the proportion of looking to the target in the Conflict condition was a predictor of performance on the task in this condition.

Third, for 4 year olds, the pattern of their behavior data looked more like adults data, while their pattern of eye tracking data was closer than that of 3 year olds. During the task, most adults' fixations were focusing on the critical features, while children spent more looking on the irrelevant features. Proportions of looking at the critical features in 4-year-olds were above of those in 3-year-olds, but were remarkably lower than that of adults and never excelled that of looking at the irrelevant features. This indicated that even though 4-year-olds exhibited high accuracy in the Conflict condition, they could not inhibit looking at the irrelevant features. Unlike adults, 4-year-olds' performance was not optimized and their choice between critical features and irrelevant features was not as efficient as adults. This suggested that children at this age were more likely to be attracted to the salient perceptual features instead of the critical but less salient one. Therefore, it is likely that if task demands were increased, 4-year-olds' performance in the Conflict condition would decrease as well.

Fourth, there was no evidence for the learning during the task. Participants did not look more to the critical features later in the task. This indicated that participants used the same strategy throughout the task and the trend that young children could not inhibit looking at more salient perceptual features was robust.

These findings indicate that young children have difficulty attending to less salient but critical task features, while ignoring more salient, but irrelevant features. Even in the very simple task used in the current research with warm up trials and instructions repeated on every trial, 3-year-olds failed in the Conflict condition, whereas 4-year-olds exhibited significant performance decrease. These findings present interesting challenges to the knowledge-based assumption that young children (and even infants) are capable of learning and using categories by spontaneously focusing on unobservable features, while ignoring salient observable features.

At the same, the study also raises a number of important questions for future research. One of them is how the low proportion of looking to critical features explained the high accuracy for 4-year-olds and whether the pattern will change for more difficult tasks. We have preliminary evidence addressing this issue. In an ongoing study, young children were presented with a more challenging induction task. While the stimuli and the procedure are the same as in the current talk, participants are asked a more difficult questions. They are informed about an unobservable property of the creature at the bottom and asked which at the top had the same property. For instance, on one trial, experimenter pointed to the creature at bottom, told children that "this one has thick blood", and asked them "Which one on the top do you think also has thick blood". The instructions that those like the same kind of cookies go together

in the matching task were changed into that those like the same kind of cookies have the same thing inside. Similar to the current task, this rule of induction was also repeated every time before each trial. Compared to the matching task, the induction task was more challenging to young children as there was more information they needed to keep track during the task. As a result, the working memory demand was higher and so was the executive function demand. Considering the results of the matching task presented here (i.e., 4-year-olds spent most of the time looking at the irrelevant features), we expected that accuracy of 4-year-olds will drop in the Conflict condition. The results support this prediction: 4-year-olds exhibited low accuracy in the Conflict condition, and it did not exceed accuracy of 3-year-olds in the current study.

Another issue that has to be addressed in future research is related to the online strategy learning and whether children could move from a less efficient learning strategy to a more efficient one during the task. For instance, whether the time pressure and the feedback will help children pay less attention to the irrelevant features.

Finally, an investigation of whether training on selective attention would accelerate children's category learning in general would provide some insight into the development of this ability and also the interaction between the development of executive function and generalization ability.

In summary, many studies have examined how young children learn new categories. The current study provided evidence indicating that young children have difficulty inhibiting attention to irrelevant information. This evidence provides challenges to the knowledge-base approach assuming the ability of infants and young children to focus on less salient aspects of the input, while ignoring more salient.

Acknowledgments

This research has been supported by grants from the grants from the NSF (BCS-0720135), from the Institute of Education Sciences, U.S. Department of Education (R305B070407), and from NIH (R01HD056105) to Vladimir M. Sloutsky.

References

Davidson, M. C., Amso, D, Anderson, L. C., & Diamondd, A. (2006). Development of cognitive control and executive functions from 4 to 13 years: Evidence from manipulations of memory, inhibition, and task switching. *Neuropsychologia, 44*, 2037–2078.

Eimas, P. D., & Quinn, C. (1994). Studies on the formation of perceptually based basic-level categories in young infants. *Child Development, 65(3)*, 903-917.

Gelman, S. A., & Markman, E. (1986). Categories and induction in young children. *Cognition, 23*, 183 – 209.

Madole, K. L., & Oakes, M. (1999). Making sense of infant categorization: Stable processes and changing representations. *Developmental Review, 19(2)*, 263-296.

Rakison, D. H., & Poulin-Dubois, D. (2001). Developmental origin of the animate-inanimate distinction. *Psychological Bulletin, 127*, 209-228.

Rueda, M., Fan, J., McCandliss, B. D., Halparin, J., Gruber, D., Lercari, L., & Posner, M. I. (2004). Development of attentional networks in childhood. *Neuropsychologia, 42*, 1029- 1040.

Sloutsky, V. M. (In Press). From perceptual categories to concepts: What develops? *Cognitive Science.*

Eye-Movements of Dyslexic Children Reading in Regular Orthography: Exploring Word Frequency and Length Effects

Evgenia Hristova (ehristova@cogs.nbu.bg)

Alexander Gerganov (agerganov@cogs.nbu.bg)

Ekaterina Todorova (e.todorova@nbu.bg)

Severina Georgieva (severina.georgieva@cogs.nbu.bg)

Central and East European Center for Cognitive Science, Department of Cognitive Science and Psychology
New Bulgarian University
21 Montevideo St., 1618 Sofia, Bulgaria

Abstract

Eye-movements represent a great interest in studying the specificity of the reading difficulties that individuals with developmental dyslexia have. In the present study dyslexic children were pair-matched with control children in a sentence reading task. The children read sentences in Bulgarian – a Cyrillic alphabet language with regular orthography. Target nouns with controlled frequency and length were embedded in the sentences. Eye movements revealed highly significant group differences in the gaze time and the total fixation times, word frequency and word length effects as well as interaction for both frequency and length with the group factor. These results, especially the frequency effect found in the dyslexic children, are discussed in the context of previous studies.

Keywords: development dyslexia; eye movements; reading.

Introduction

Developmental dyslexia is described as a condition found in children as young as 6-7 years that impairs their reading skills, while their IQ, reasoning and communication abilities are intact. Still, there is large variability in both the symptoms that dyslexic children demonstrate and in the experimental findings that give support to several theories explaining the underlying causes for dyslexia (see Vellutino, Fletcher, Snowling & Scanlon, 2004 for a review).

Usually, dyslexic children are given non-verbal, phonological or single word reading tasks, which aim to distinguish between different theories. While the rationale behind these experiments is very sound, classical reading experiments are also of great interest. For many years eye movements during reading provide insight into psycholinguistic research. During reading, dyslexic readers exhibit more and longer fixations and a higher percentage of regressions than normal readers. It is still a matter of debate, whether these divergent eye movement patterns of dyslexic readers reflect an underlying problem in word processing or whether they are – as the proponents of the oculomotor deficit hypothesis claim (e.g. Pavlidis, 1981) – associated with deficient visual performance that is causal for dyslexia.

It is a well-documented (and undisputed fact) that eye movements of dyslexic readers differ from those of normal readers. During reading, dyslexic readers exhibit more and longer fixations, shorter saccades and a higher percentage of regressions than normal readers (for review, see Rayner, 1998).

Hutzler, Kronbichler, Jacobs and Wimmer (2006) used a string processing task that imposes the same requirements as reading to visual perception (letter identification) and oculomotor control (moving the eyes in the same pattern as during reading). The task is different from reading as it does not require linguistic or language processing of the visual information beyond letter identification. In the study above the authors found no differences between the eye-movements of dyslexic and normal readers and concluded that differences in eye-movements during reading are not the cause for the impaired performance.

Hyona & Olson (1995) also tested the hypothesis that the specificity of eye-movements of dyslexic readers is the cause for their reading difficulties. They found word length and word frequency effects on eye-movement characteristics of dyslexic readers. The conclusion they made was that the eye-movement patterns of dyslexic readers are affected by the properties of the linguistic material encountered during reading and therefore eye-movement patterns of dyslexic readers are reflection of the difficulties these readers have during linguistic processing (and not vice versa).

Still, there are few studies of text- or sentence-level reading with dyslexic children. An eye-movement study on reading German text passage found word length effects for both dyslexic and normal readers as well as interaction between the groups (Hutzler & Wimmer, 2004). The words taken from the text passage, however, could not be controlled for possibly confounding factors like predictability and frequency. In a similar task of reading short text passages in Italian, De Luca, Di Pace, Judica, Spinelli and Zoccolotti (1999) found once again strong length effects but much smaller frequency main effect that was marginally significant and did not interact with the group factor. Finally, Hyona & Olson (1995) compared a group of dyslexic children with younger ones and found highly significant word frequency and word length effects

for both groups in a somewhat similar task – reading aloud of English texts. Although there was no main effect of the group factor, an interaction between length and group was still observed (but only in the subject means).

These experiments show an interesting pattern of results. Dyslexic children seem to show strong length effects and weaker frequency effects in text reading but the differences between normal readers and the dyslexic ones resembles the difference between experienced and average readers or in the children case – of younger, less trained in reading children (Olson, Conners, & Rack, 1991). School practices show that children diagnosed with Dyslexia tend to resent reading and as a result of their reading difficulties, they are less exposed to written text than normal children. Whatever the underlying reason for the various symptoms may be, it is clear that reading practice plays some important role in the later reading behavior of dyslexic children and adults. Indeed most theories predict the length effects which can be explained by difficulties in grapheme-phoneme decoding, oculomotor control, attention. The word frequency effect, however, is closely related to reading experience. It could be argued, that for languages with irregular orthography the grapheme-phoneme decoding could be more problematic for less frequent words than for languages with regular orthography – an explanation suggested by De Luca et al. (1999) for their results that showed much stronger length effects than frequency effects for Italian dyslexic readers when compared to the Hyona & Olson study (1995) on English readers (English is a language with irregular orthography, while Italian – with regular).

Clearly, a further investigation of word frequency and word length effect in reading is necessary in order to explore these inconclusive results.

Experiment

This experiment aims to study word length and word frequency effects in Bulgarian language (a Cyrillic language with regular orthography). Target nouns were embedded in sentences that were controlled for the preceding context (neutral) among other possibly confounding factors, thus providing much more reliable results than words selected from text passages.

Stimuli and design

Before conducting the study there was a preparatory phase. As a first step, we collected a large corpus of children texts in Bulgarian. The corpus contains children books, fairy tales, etc. representative for the age groups studied. It consists of 931 320 words in total, among them 58 605 unique.

From this corpus we selected *short* (5 letters) and *long* (8 letters) concrete nouns (animals, objects, flowers, etc.) that were either *high*- or *low-frequency*. To calculate word frequencies we first computed the raw frequency (number of occurrences per million words) and then we performed a logarithmic transformation. After this we chose 16 short words (5 letters) and 16 long words (8 letters) that have similar low frequency. We also chose 16 short words (5 letters) and 16 long words (8 letters) that have similar high frequency. Summary of the frequencies of the words chosen is presented in Table 1.

Table 1: Summary of the words used in the study. Frequency was assessed as normalized number of occurrences in a 1 million words corpus of texts that are usually read by children (fairy tales, novels, etc.).

		Frequency per million (ln)		
		min	Max	average
5-letter words	low frequency	0	2.26	1.27
	high frequency	3.29	4.6	3.85
8-letter words	low frequency	0	2.01	1.32
	high frequency	2.96	5.97	3.9

In this way, we were able to vary both word length and frequency in a 2x2 design with factors: word length (short vs. long words) and word frequency (high vs. low frequency).

Each of the 64 target words were embedded in a sentence with neutral preceding context. The target word was never the first word in the sentence. The sentences were with content appropriate for children. Example sentences are as follows (the target words are in bold):

- 5-letters, high-frequency: ‘Подробна **карта** на океаните е нужна на всеки пират’. (A detailed **map** of the oceans is a necessity for every pirate).
- 5-letters, low-frequency: ‘Добрият **бобър** живееше край омагьосаната река’. (The good **beaver** lived near the enchanted river).
- 8-letters, high-frequency: ‘Хитрото **чудовище** пресрещаше пътниците и им задаваше гатанки’. (The clever **monster** stopped passengers and gave them riddles).
- 8-letters, low-frequency: ‘Червеният **карамфил** беше във високата ваза на земята’. (The red **carnation** was in the tall vase on the ground).

The sentences were counterbalance in two lists, so that each participant saw 32 sentences (8 sentences from each condition).

Procedure and apparatus

Sentences appeared one by one on a screen and were read silently. The task of the children was to read each sentence and to understand it. After reading the sentence, the participant had to press the space bar on a standard computer keyboard. The sentence stayed on the screen until the space bar was pressed and then it disappeared. In order to assure careful reading, control questions appeared after

some of the sentences (the questions were related to the content of the sentence). The questions required a 'yes' or 'no' answer. After reading the question, the participant had to press one of two keys on the keyboard marked with labels 'YES' and 'NO'. There was a fixation cross between the sentences and the participants were instructed to look at it when it appeared.

Each participant had to read 32 sentences, which were presented in a pseudo-randomized order. In the beginning there were 8 practice trials. Data from the practice trials were not included in the analysis. The practice trials were intended to provide an opportunity for the participants to get used to the task.

Eye-movement data were recorded with a Tobii 1750 remote eye-tracker and ClearView 2.7.1 software. The eye-tracker looks like a computer screen with in-built cameras and sensors. That allowed for comfortable and completely unobtrusive recording of eye-movements. Each participant was seated at a distance of approximately 55 cm from the screen. The sentences were presented in black letters on white background. The sentences were presented in Tahoma font (a sans-serif typeface). The size of the letters was chosen to space 3 letters per degree of visual angle. The screen was an integrated 17' TFT monitor set to its native resolution (1280 x 1024).

The equipment recorded gaze coordinates on the screen every 20 ms. ClearView algorithms were used to compute fixation duration and location from these raw data (the fixation analysis filter was set to 40 pixels fixation radius and 100ms minimal fixation duration). ClearView was also used to control stimulus presentation and to collect participants' answers.

Participants

Seven dyslexic children and seven children with normal reading skills were matched (in pairs) on age and nonverbal IQ. Full matching data are presented in Table 2. Children with attention disorders were excluded from the sample. All participants had normal or corrected to normal vision.

Data analysis and results

Participants performed well on the control questions – all reported participants had above 80% correct answers (see Table 2 for individual scores).

One of the items (8-letters, high-frequency word) was excluded from the analyses due to typo in the stimulus material.

First-pass durations (gaze duration) and total times were selected as dependent measures that reflect well both word frequency and length effects in reading (Rayner, 1998). **First-pass duration** is calculated as the sum of all fixation durations beginning with the first fixation in a region (the target word) until the reader's gaze leaves the region, left or right. **Total time** is calculated as the sum of all fixation durations in a region (the target word), regardless of their order.

The eye-movement data were analyzed using two separate analyses of variance (ANOVA): using subjects (F1) and items (F2) as cases.

Table 2: Participants in the study. Each dyslexic child is matched with the child in the row below. The column IQ represents the raw score on a non-verbal Raven test with 36 questions (a point is granted per correct answer). The column "Correct Answers" gives the percentage of correct answers for the comprehension questions during the reading task.

Group	Age (months)	Gender	IQ (raw score)	Correct Answers (%)
Dyslexia	103	Male	29	95
Norm	106	Female	33	100
Dyslexia	120	Male	33	90
Norm	120	Male	32	95
Dyslexia	123	Female	32	100
Norm	125	Female	35	100
Dyslexia	128	Male	32	100
Norm	128	Male	32	100
Dyslexia	130	Male	33	90
Norm	131	Male	35	100
Dyslexia	136	Female	34	95
Norm	142	Female	30	85
Dyslexia	141	Female	20	80
Norm	144	Female	27	90

First-pass duration

First-pass durations (Table 3) were analyzed as a function of word length, word frequency, and group (dyslexic or normal readers).

Comparison between dyslexia group and control group

The subjects analysis (repeated-measures ANOVA) on first-pass duration was performed with two within-subjects factors: word length (short and long) and word frequency (low and high), and group (dyslexic or normal readers) as a between-subject factor. The item analysis on first-pass duration was performed with word length and word frequency as between-item factors, and group as within-item factor.

The main effect of group (dyslexic vs. normal readers) on first-pass duration is significant: $F1(1, 12) = 20.28$, $p \leq 0.001$; $F2(1, 59) = 150.9$, $p < 0.001$. In general, dyslexic readers showed much longer first-pass durations (means were 2044 ms for the dyslexic readers vs. 467 ms for the normal readers).

Table 3: Mean first-pass duration (in ms) as a function of word length and word frequency in dyslexia and control groups.

	Word length	Word frequency		
		high	low	M
Dyslexia group	short	1418	2068	1743
	long	2072	2620	2346
	M	1745	2344	2044
Control group	short	405	421	413
	long	441	600	520
	M	423	511	467

The main effect of word length (short vs. long) on the first-pass duration was significant in the items analysis ($F2(1, 59) = 6.02$, $p < 0.05$) and marginally significant in the subjects analysis ($F1(1, 12) = 4.44$, $p = 0.057$). Longer (8-letters) words lead to longer first-pass durations compared to the short (5-letter) words. Length by group interaction did not reach statistical significance ($F1(1, 12) = 2.16$, $p = 0.17$; $F2(1, 59) = 2.65$, $p = 0.11$). Long words (8-letters) received longer first-pass durations both in the dyslexia and in the control group (see Figure 1).

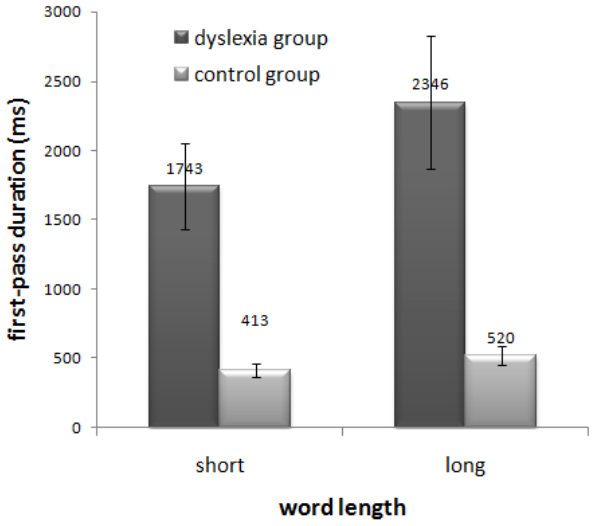


Figure 1: Average first-pass duration (in ms) as a function of word length in dyslexia and control groups. Error bars represent standard error of the mean.

The main effect of word frequency (high vs. low) on first-pass duration was significant in the subjects analysis and in the items analysis: $F1(1, 12) = 10.1$, $p < 0.01$; $F2(1, 59) = 8$, $p < 0.01$. Low-frequency words lead to longer first-pass durations compared with the high-frequency words. Frequency by group interaction was also statistically significant in both subjects and item analysis: $F1(1, 12) = 5.6$, $p < 0.05$; $F2(1, 59) = 5.86$, $p < 0.05$ (see Figure 2). Additional tests on simple effects in items analysis reveal that word frequency effect is significant in dyslexia group ($p < 0.05$) and not significant in the control group ($p = 0.13$). The interaction reflects the fact that low frequency words (compared to high-frequency words) lead to greater increase in first-pass duration only in the dyslexia group.

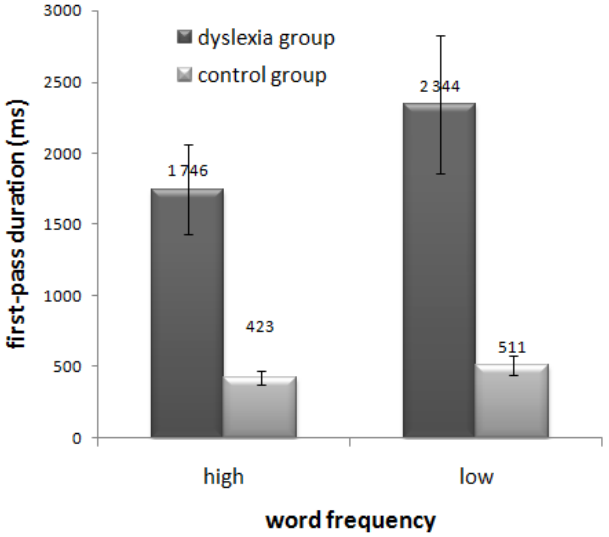


Figure 2: Average first-pass duration (in ms) as a function of word frequency in dyslexia and control groups. Error bars represent standard error of the mean.

Summary of the results for first-pass duration The comparison between dyslexia group and control group demonstrates that *dyslexic children have much longer first-pass duration* in general. Their first-pass durations are approximately 4-5 times longer than for the control group. There was also main effect of word frequency. However, the frequency by group interaction and the additional analysis revealed that the increase in first-pass duration for low-frequency words is present only for the dyslexic group. Main effect of word length on first-pass duration is also found: long words receive longer first-pass durations both in the dyslexia and in the control groups.

Dyslexic children show longer first-pass durations for the long words compared to the short words (word length effect) and for low-frequency words compared to high-frequency words (word frequency effect). So, it seems that eye-movements of dyslexic children are affected by such lexical factors as word length and word frequency.

First-pass durations for dyslexic readers seem to be affected to a greater extend by word-frequency, unlike in some of the previous studies.

Total time

Total times (see Table 4 for means) were analyzed as a function of word length, word frequency, and group (dyslexic or normal readers).

Table 4. Mean total time duration (in ms) as a function of word length and word frequency in dyslexia and control groups.

	Word length	Word frequency		
		High	low	M
Dyslexia group	short	1928	2783	2355
	long	3063	3671	2346
	M	2495	3227	2861
Control group	short	545	556	551
	long	620	747	683
	M	583	652	617

Comparison between dyslexia group and control group The subjects ANOVA on total time was performed with two within-subjects factors: word length (short and long) and word frequency (low and high), and group (dyslexic or normal readers) as a between-subject factor. The item analysis on total time was performed with word length and word frequency as between-item factors, and group as within-item factor.

The main effect of group (dyslexic vs. normal readers) is significant in the subjects and in the items analysis: $F1(1, 12) = 28.4$, $p < 0.001$; $F2(1, 59) = 154.3$, $p < 0.001$. In general dyslexic readers showed longer total time gaze durations compared to the normal readers (means are 2861 ms for the dyslexic readers vs. 617 ms for the normal readers).

The main effect of length on total time was significant in both subjects analysis and items analysis ($F1(1, 12) = 5.52$, $p < 0.05$; $F2(1, 59) = 6.66$, $p < 0.05$). Longer words (8-letters) led to longer total time gaze durations compared to the shorter (5-letters) words. Length by group interaction did not reach statistical significance in both analyses ($F1(1, 12) = 3.25$, $p = 0.096$; $F2(1, 59) = 3.31$, $p = 0.074$). Additional tests on simple effects reveal that word length effect is significant in dyslexia group ($p < 0.05$) and in the control group ($p < 0.01$). Reading for long words had longer total time durations both in the dyslexia and in the control group (see Figure 3).

The main effect of frequency was significant in both analyses: $F1(1, 12) = 6.47$, $p < 0.05$; $F2(1, 59) = 6.68$, $p < 0.05$. Low-frequency words led to longer total time durations compared with the high-frequency words. Frequency by group interaction was statistically significant in the items analysis ($F2(1, 59) = 5.71$, $p < 0.05$) and marginally significant in the subjects analysis ($F1(1, 12) = 4.44$, $p = 0.057$) The interaction between word frequency and group is presented in Figure 4. Additional tests on simple effects revealed that word frequency effect is significant in the dyslexia group ($p < 0.05$) and not significant in the control group ($p = 0.43$). The interaction reflects the fact that low frequency words lead to greater increase in total time duration only in the dyslexia group.

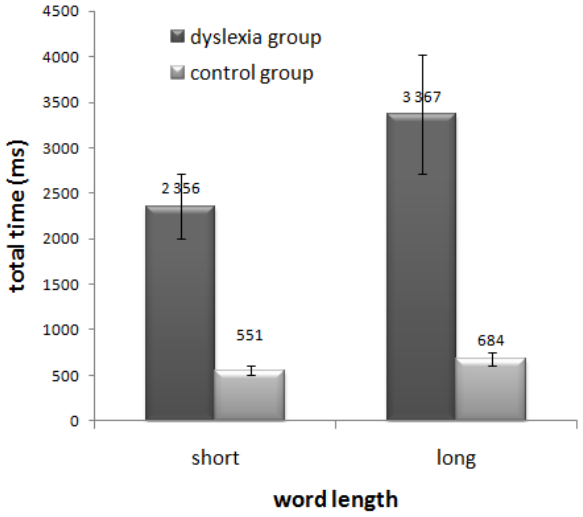


Figure 3: Average total time duration (in ms) as a function of word length in dyslexia and control groups. Error bars represent standard error of the mean.

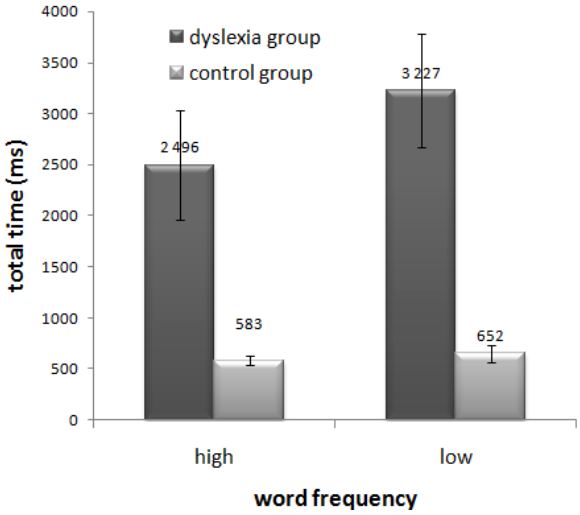


Figure 4: Mean total time (in ms) as a function of word frequency in dyslexia and control groups. Error bars represent standard error of the mean.

Summary of the results for total time The comparison between dyslexia and control groups showed that dyslexic children had generally longer total time for viewing the target words than the controls. There was main effect of frequency. However, the frequency by group interaction and the additional analysis revealed that the increase in total time for low-frequency words is present only for the

dyslexic group. Long words receive longer total time durations both in the dyslexia and in the control groups.

The dyslexic children show word length and word frequency effects in eye-movements.

General Discussion

The overall gaze duration and total time for the target words were much longer than the reported by Hutzler & Wimmer (2004) and Hyona & Olson (1995). This could be explained by the age of the dyslexic children – in this study they were between second and fourth grade, while in the above-mentioned studies, the dyslexic children were about 7th grade (about 14 years old). As many dyslexic children tend to develop different strategies with age which help them overcome their reading problems, thus we reason that studying eye-movements of younger dyslexic children give us the possibility to study the specificity of the their reading difficulties without the confounding effect of such strategies.

Another possible explanation for the results lies in the silent reading for comprehension task. The children were highly motivated to reply accurately, which can be seen by the very high number of correct answers (see Table 2). We argue that this task is more natural than reading aloud, which sometimes can be done without any comprehension.

Using sentences with embedded target words allowed better controlling for confounding factors and successfully varying word length and word frequency as independent factors.

The main effect of word length replicated most of the previous findings. The interaction between length and group (well-established in previous research) failed to reach significance probably due to the small number of participants. Word frequency, however, showed somewhat different pattern than former studies. Word frequency effects were very weak for Italian dyslexic readers (De Luca et al. 1999) and did not interact with the group factor in neither De Luca at al. (1999), nor Hyona & Olson (1995). This discrepancy once again can be explained by both the age of the participants and the task - the later, however, seems more probable, since frequency effects reflect not only lexical access but also some comprehension and integration processes that take important part in reading for comprehension unlike reading aloud for example. These results contradict previous findings that claim there is no frequency effects in dyslexic children in regular orthography, or that these effects are much weaker than the length effects. The explanation that is suggested is that the frequency effects stem from the irregular orthography. Our data show that this is not the case and that frequency effects are well manifested in the dyslexic population even in a language with regular orthography.

Conclusion

The results from the current study show that young dyslexic children have extremely slow, but otherwise normal reading patterns that are governed not only by word length but also by frequency – an effect that usually marks good reading skills. The interaction between frequency and group implies that there is some higher-level processing impairment that inhibits the recognition of rare words or that the children simply do not have the same vocabulary range as the controls.

Acknowledgments

We would like to thank all participants in the current study. We also like to thank 151st School, Sofia, for their support during the study in granting us access to participants in the control group. And finally, we would like to thank Veronika Penkova, Kalina Seksenova, Stanislava Borisova, and Yoana Vergilova for their help in data collection.

References

De Luca, M., Di Pace, E., Judica, A., Spinelli, D., & Zoccolotti, P. (1999). Eye movement patterns in linguistic and non-linguistic tasks in developmental surface dyslexia. *Neuropsychologia*, *37*, 1407–20.

Hyona, J. & Olson, R. K. (1995). Eye fixation patterns among dyslexic and normal readers: effects of word length and word frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21* (6), 1430-40.

Hutzler, F., & Wimmer, H. (2004). Eye movements of dyslexic children when reading in a regular orthography. *Brain and Language*, *89*, 235-242.

Hutzler, F., Kronbichler, M., Jacobs, A.M., & Wimmer, H. (2006). Perhaps correlational but not causal: No effect of dyslexic readers' magnocellular system on their eye movements during reading. *Neuropsychologia*, *44*, 637-648.

Olson, R., Conners, F., & Rack, J. (1991). Eye movements in dyslexics and normal readers. In Stein, J. (ed.). *Vision and visual dyslexia*. Houndmills: Macmillan Press.

Pavlidis, G. (1981). Do eye movements hold the key to dyslexia? *Neuropsychologia*, *19*, 57-64.

Ramus, F., Rosen, S., Dakin, S. C., Day, B. L., Castellote, J. M., White, S., & Frith, U. (2003). Theories of developmental dyslexia: Insights from a multiple case study of dyslexic adults. *Brain*, *126*, 841-865.

Rayner, *K. (*1998*)*, Eye Movements in Reading and Information Processing: *20* Years of Research, *Psychological Bulletin*, *124*, 372-422

Vellutino, F. R., Fletcher, J. M., Snowling, M. J., & Scanlon, D. M. (2004). Specific reading disability (dyslexia): what have we learned in the past four decades? *Journal of Child Psychiatry*, *45*, 2-40.

Mathematical reasoning with higher-order anti-unifcation

Markus Guhe, Alison Pease, Alan Smaill
(m.guhe|a.pease|a.smaill@ed.ac.uk)
University of Edinburgh, School of Informatics, Informatics Forum, 10 Crichton Street, Edinburgh EH8 9AB, Scotland

Martin Schmidt, Helmar Gust, Kai-Uwe Kühnberger, Ulf Krumnack
(martisch|hgust|kkuehnbe|krumnack@uni-osnabrueck.de)
University of Osnabrück, Institute of Cognitive Science, Albrechtstr. 28, 49076 Osnabrück, Germany

Abstract

We show how *heuristic-driven theory projection* (HDTP, a method based on higher-order anti-unification) can be used to model analogical reasoning in mathematics. More precisely, HDTP provides the framework for a model of the inductive analogy-making process involved in establishing the fundamental concepts of arithmetic. This process is a crucial component for being able to generalise from the concrete experiences that humans have due to their embodied and embedded nature. Such generalisations are a cornerstone of the ability to create an abstract domain like arithmetic. In addition to generalisations, HDTP can also transfer concepts from one domain into another, which is, for example, needed to introduce the concept ZERO into arithmetic. The approach presented here is closely related to the theories of *Information Flow* and *Institutions*. The latter in particular provides a compelling way to integrate *concept blending* into the HDTP approach.

Keywords: mathematical cognition; mathematical reasoning; analogy; anti-unification; concept blending

Mathematical reasoning as a cognitive process

Although mathematics is usually presented in terms of axioms, concise proofs, theorems and so on, the actual cognitive process of mathematical reasoning is very different. For example, when a mathematician changes a definition this affects the proofs that use it, but such changes are not discussed in mathematical papers. Additionally, mathematics, at least partly, does not consist of discovering eternal, Platonic ideals but in creating mathematical concepts. For example, Lakatos's (1976) account of the history of Euler's conjecture illuminates how the concept POLYHEDRON can differ and how its definition depends on the current circumstances and needs of the mathematician. Put differently, if the Platonic ideal POLYHEDRON does exist, it is not clear how it can be identified by mathematical means – what cognitive processes mathematicians can use to find the correct definition. Thus, mathematical concepts are not necessarily the same as the ideals.

Lakoff and Núñez (2000) describe how our embodied, situated experience is the basis on which abstract mathematical concepts are developed by a process of metaphorical abstraction and transfer. In chapter 3, they describe how basic arithmetic is created from four everyday experiences, which are the source domains of the metaphors. In this way, arithmetic is grounded in situated cognition. To motivate that these four domains in particular are source domains, Lakoff and Núñez analyse linguistic expressions used in the target domain, arithmetic, which they trace back to these four domains. For example, we use the terms *add* and *take away* in arithmetic. Lakoff and Núñez argue that these terms were originally used for talking about collections of objects, such as physically placing an object into a container, e.g. *adding an onion to the soup*, or physically removing a substance or an object from a container, e.g. *take a book out of the box*.

Analogical reasoning is a central component of the process transforming knowledge of this kind into mathematical concepts. For present purposes we assume that metaphor and analogy are essentially the same cognitive process (Gentner, Bowdle, Wolff, & Boronat, 2001), and we have demonstrated how structure mapping (Gentner, 1983; Gentner & Markman, 1997) – a basic method to compute analogical relations – can account for the overall cognitive process (Guhe, Pease, & Smaill, 2009).

In this paper, we describe a formal cognitive model of this process. This has a twofold motivation: firstly, we want to specify the cognitive processes that mathematicians use, to better understand how mathematical discovery works; secondly, we want to use the model to improve automated theorem provers by incorporating cognitive mechanisms. In Guhe, Smaill, and Pease (2009a, 2009b) we presented formal representations of the four grounding metaphors (the 4Gs) and suggested how *Information Flow* theory (Barwise & Seligman, 1997) may be used to model the analogies involved. The 4Gs are: (1) arithmetic is object collection, (2) arithmetic is object construction, (3) measuring stick and (4) arithmetic is motion along a path.

Here, we present a proof-of-concept of how performing anti-unification (Plotkin, 1970) on such representations can account for aspects of the analogical reasoning involved in the 4Gs. This inductive kind of reasoning provides us with a procedural version of the otherwise static Information Flow models and enables us to computationally determine the relationships between the domains (classifications in the case of Information Flow). More precisely, we will use *Heuristic-Driven Theory Projection* (HDTP; Schwering, Krumnack, Kühnberger, & Gust, 2009), a general framework for making analogies. HDTP provides us with the means to generalise over two of Lakoff and Núñez's domains to establish a basis for arithmetic as well as the means to generalise over one of the domains as source domain and arithmetic as the target domain to add concepts to arithmetic that are only present in one of the grounding domains. We will also outline how this conception of mathematical reasoning is linked to Goguen's (2006) notion of *concept blending* (which is based on notions by Gärdenfors, 2000 and Fauconnier & Turner, 2002), a further cognitive process for creating mathematical concepts.

Metaphors for arithmetic

Arithmetic is object collection

The arithmetic is object collection metaphor (Table 1) is based on the notion that the repeated manipulation of (small, countable, physical) collections of objects lets us notice certain regularities. For example, we can determine which one of two collections is bigger by aligning the objects in the two collections one-to-one, and the collection that has at least one unpaired object left over is the bigger collection. (Smaller and equal are, of course, determined correspondingly.) This corresponds to the (abstract) arithmetic notion GREATER.

By comparing collections of objects in this way we can also group such collections into groups of collections of equal size, i.e. where after the aligning procedure no object is left unpaired. Each of these groups corresponds to a number in arithmetic.

There are two things to note about this basic metaphor. Firstly, it does not produce a concept of ZERO, because the empty collection is a collection that does not exist physically. (Even calling one object a *collection with one object* is an abstraction of the term *collection*.) Lakoff and Núñez (2000, p. 64) propose that an *entity-creating metaphor* is required to create a concept that is not part of the basic metaphor (like ZERO). This corresponds well with the fact that, historically, ZERO was a rather late invention. Secondly, the subtraction operation requires that a smaller collection be taken from a bigger one, because physically, negative objects do not exist.[1]

Table 1: Arithmetic is object collection metaphor (Lakoff & Núñez, 2000, p. 55)

object collection	arithmetic
collections of objects of the same size	numbers
size of collection	number
bigger	greater
smaller	less
smallest collection	the unit (one)
putting collections together	addition
taking a smaller collection from a larger collection	subtraction

Arithmetic is object construction

The arithmetic is object construction metaphor (Table 2) runs along the same lines, except that it is not based on collections of objects, but on objects that are constructed from smaller objects. In this way, fractions are added to arithmetic, although they are not part of the basic metaphor. Consider, for example, an object that is constructed out of seven smaller objects. If now a smaller object that consists of three of the seven overall objects is removed from the original object, the two resulting objects have a size of $\frac{3}{7}$ and $\frac{4}{7}$ of the original.

[1] One is reminded of the old joke: If on one floor 5 people leave an elevator containing 3 people, 2 people have to enter the elevator on the next floor in order for it to be empty.

Table 2: Arithmetic is object construction metaphor (Lakoff & Núñez, 2000, pp. 65–66)

object construction	arithmetic
objects	numbers
smallest whole object	the unit (one)
size of object	size of number
bigger	greater
smaller	less
constructed object	result of arith. operation
whole object	a whole number
putting objects together to form larger objects	addition
taking smaller objects from larger objects to form other objects	subtraction

Arithmetic is motion along a path

The *motion along a path* metaphor (Table 3) adds concepts to arithmetic that we experience by moving along straight paths. Numbers are point locations on paths. Addition and subtraction correspond to a movements from point one point on the path to another point on the path. An important new concept that is added to arithmetic by this metaphor is ZERO, which is based on the concept of a path's origin and which provides a direction for the movements along paths, namely towards the origin or away from it.

Table 3: Arithmetic is motion along a path metaphor (Lakoff & Núñez, 2000, p. 72)

motion along a path	arithmetic
acts of moving along the path	arith. operations
a point location on the path	result of an operation; number
origin; beginning of the path	zero
unit location, a point location distinct from the origin	one
further from the origin than	greater
closer to the origin than	less
moving away from the origin a distance	addition
moving toward the origin a distance	subtraction

Heuristic-Driven Theory Projection

Overview

This section provides a short overview of the basic ideas of heuristic-driven theory projection (HDTP), a formal framework to model analogical mapping and reasoning. A more detailed description can be found in Schwering et al. (2009).

HDTP establishes analogies between two domains, the source and the target, by detecting common structures. In the mapping phase, source and target are compared for structural commonalities and a generalised description is created, which

subsumes the matching parts of both domains. In the transfer phase, unmatched knowledge in one domain can be mapped to the other to establish new hypotheses.

HDTP is a formal framework that computes analogical relations and inferences for domains represented in first-order logic. Both, source and target domain, are given by axiomatisations, i.e. finite sets of first-order formulae. The basic idea is to associate pairs of formulae from the domains in a systematic way. HDTP uses anti-unification (Plotkin, 1970) to identify common patterns in formulae. In anti-unification, two formulae are compared and the least general generalisation that subsumes both formulae is identified.

Figure 1 provides some examples of anti-unification of terms. Terms are generalised to an anti-instance where different constants or function symbols are replaced by a variable. In (i), first-order anti-unification is sufficient. However, the terms in (ii) differ in the function symbols, i.e. first-order anti-unification fails to detect structural commonalities. Here, higher-order anti-unification generalises function symbols to a variable and retains the structural commonality. It is even possible to generalise terms in which common parts are embedded structurally in a different way, as shown in (iii).[2] Substitutions accompanying the generalised terms are created, which can be used to reconstruct the original terms.

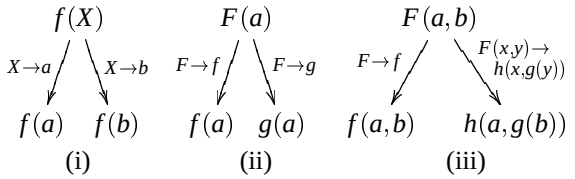
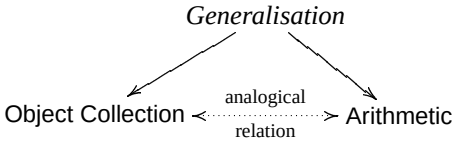


Figure 1: Anti-unification of terms

HDTP extends this classical anti-unification of terms to formulae and logical theories by iteratively picking pairs of formulae to be generalised from the domains. This process is driven by heuristics. Coherent mappings are preferred, i.e. mappings in which substitutions can be reused. The generalised theory together with its substitutions specifies the analogical relation between source and target. Additional information about the source domain, i.e. formulae with no correspondence in the target domain, can be transferred by replacing symbols using the established substitutions.

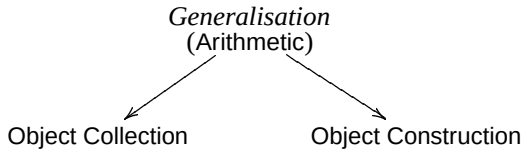
Modelling the arithmetical metaphors

HDTP provides two different ways in which Lakoff and Núñez's (2000) grounded domains (Object Collection, Object Construction etc.) can be related to the abstract domain of Arithmetic. Following Lakoff and Núñez, the grounded domains constitute the source, while Arithmetic is the target domain. To establish an analogical relation between Object Collection and Arithmetic, HDTP can construct a generalisation of these domains:

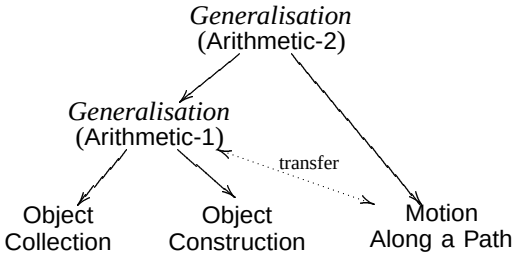


In this model, both domains are already given. The analogy explains abstract concepts like numbers by linking them to familiar entities from the grounded domains. Thus, the generalisation provides a description of the commonalities of the grounded and the abstract domains.

However, from the cognitive perspective, Arithmetic does not initially exist – it has to be created by an act of abstraction as well. This idea can be modelled by analogically relating two grounded domains, e.g. Object Collection and Object Construction. Arithmetic then emerges as a generalisation of these domains.



In our view, a combination of both approaches is needed to model the cognitive bootstrapping process. By generalising over two grounded domains, an abstract domain is established, which serves as a 'proto-domain' of Arithmetic, i.e. a domain that already contains some arithmetical concepts. This is then refined subsequently, by relating it analogically to other grounded domains, removing peculiarities of the two original domains and/or adding new concepts by analogical transfer.



It should be noted that in pursuing this approach the results may vary depending on which grounded domains are chosen for generalisation and on the order in which other grounded domains are added for refinement. This is due to the heuristics that HDTP applies when building up the generalisation. The more similar the grounded domains are, the richer the generalisation will be, while dissimilar domains give coarser results. Nevertheless, we expect that this effect can be compensated by further mapping the initial generalisation to other domains. A detailed examination of this will be a focus of our future work.

Formalisation of domains

We demonstrate the feasibility of the outlined approach by applying it to simple formalisations of Lakoff and Núñez's

[2] HDTP uses a restricted form of higher-order substitutions, that allows to expand terms by introducing arguments and nested structures as described in Krumnack, Schwering, Gust, and Kühnberger (2007).

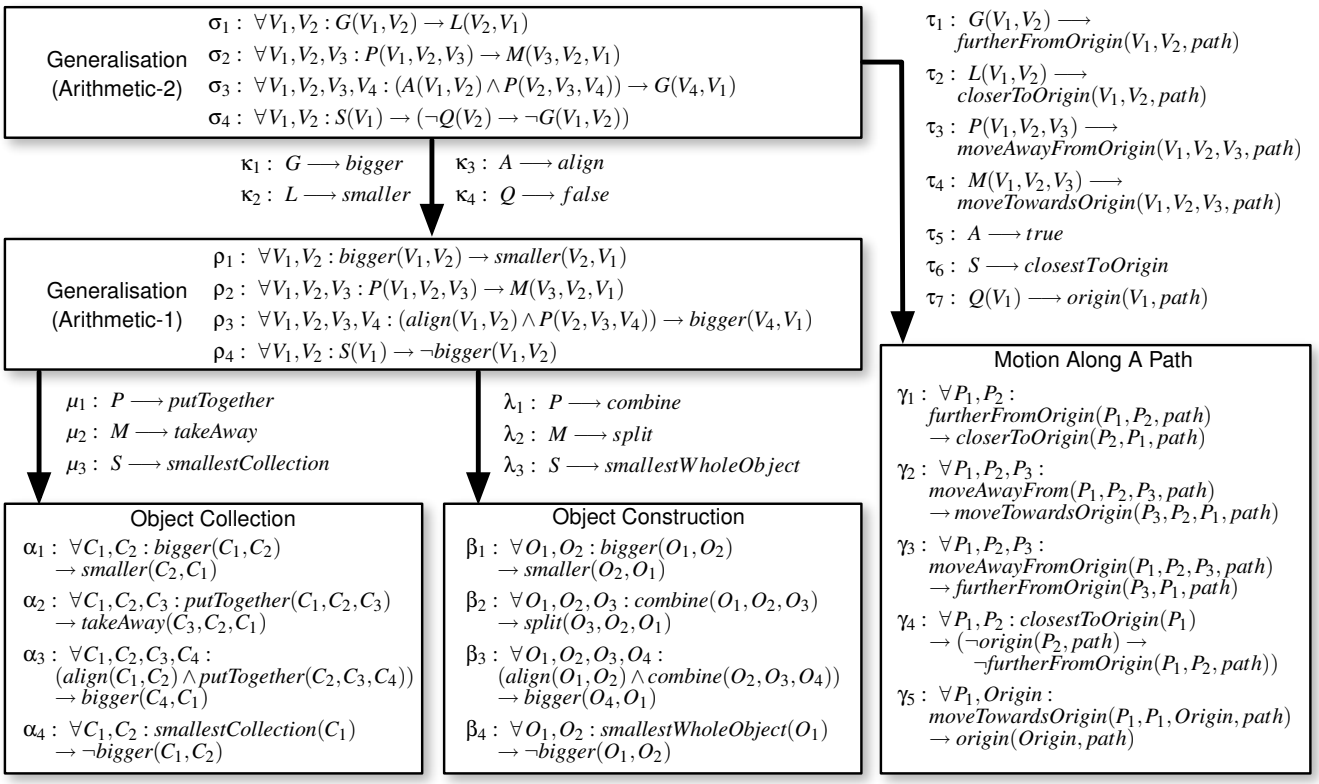


Figure 2: Developing Arithmetic from Object Collection, Object Construction and Motion Along a Path

grounding metaphors. The original descriptions of the domains are only given informally, but we tried to stay as closely to original as possible. One possible axiomatisation in HDTP of the Object Collection domain is given in Table 4. Such a formalisation specifies the vocabulary that is used in the form of *sorts*, *entities* and *predicates* and then provides *facts* and *laws* to describe the structure of the domain. For example, axiom α_3 states that if two collections C_1 and C_2 can be aligned, i.e. all their objects can be paired up, and C_4 is created by putting C_2 and C_3 together, then C_4 will be bigger than C_1. Note that further formulae need to be added to get a complete axiomatisation, but such formulae can easily be introduced into the system as long as some elementary consistency constraints are satisfied. While adding more formulae to this formalisation might strengthen the support for a specific alignment, it does not necessarily introduce new mappings of concepts to other domains. An example for this is α_6, which states the transitivity of *bigger*. This formula embeds *bigger* further in the structure of the domain but does not introduce new concepts. *putTogether*, *takeAway*, *bigger* and *smaller* are considered core concepts of the Object Collection domain. In what follows, we will restrict our axiomatisations to such simple versions in which just the cores of the domains are represented and connected to each other. Furthermore, we will omit technical details as well as the specification of sorts and signatures for a more concise presentation.

Generalising two domains

We tested various alternative formalisations, which all resulted in HDTP being able to establish appropriate analogies. Here we present axiomatisations of the grounded domains that are compact and consistent and that import integral parts of the domains. Furthermore, we demonstrate how the transfer of knowledge from one domain to another one works, because this is a hallmark of 'interesting' analogies.

In a first step, we generalise the domains of Object Collection and Object Construction. (We only use the basic version of the Object Construction domain here, which largely resembles Object Collection. This version is not sufficient to introduce the concept of fractions.) The axiomatisation of the two domains can be found in the two boxes in the bottom left of figure 2. The grounded domains are restricted to the operations that in arithmetic correspond to *greater*, *less*, *addition* and *subtraction*. The axioms α_i and β_i (for $i \in \{1,\ldots,4\}$) correspond to each other and are generalised in the obvious way by introducing individual variables and variables for operations. For example, the predicate *putTogether* of the Object Collection domain is identified with *combine* of Object Construction and generalised to a variable P. The substitutions μ_1 and λ_1 can be used to reconstruct the original expressions. Note that aligning corresponding clauses in formalisations is only done for the convenience of the reader; HDTP does not rely on such an ordering to find the best possible analogical mapping.

Table 4: Formalisation of the Object Collection domain

Sorts
coll
Entities
$singleton : coll$
Predicates
$smallestCollection : coll$
$bigger : coll \times coll$
$smaller : coll \times coll$
$equal : coll \times coll$
$putTogether : coll \times coll \times coll$
$takeAway : coll \times coll \times coll$
Laws
$\alpha_1 : \forall C_1 : coll, C_2 : coll : bigger(C_1, C_2) \rightarrow smaller(C_2, C_1)$
$\alpha_2 : \forall C_1 : coll, C_2 : coll, C_3 : coll : putTogether(C_1, C_2, C_3) \rightarrow takeAway(C_3, C_2, C_1)$
$\alpha_3 : \forall C_1 : coll, C_2 : coll, C_3 : coll, C_4 : coll : align(C_1, C_2) \wedge putTogether(C_2, C_3, C_4) \rightarrow bigger(C_4, C_1)$
$\alpha_4 : \forall C_1 : coll, C_2 : coll : smallestCollection(C_1) \rightarrow not(bigger(C_1, C_2))$
$\alpha_5 : \forall C_1 : coll, C_2 : coll : equal(C_1, C_2) \rightarrow (\neg bigger(C_1, C_2) \wedge \neg smaller(C_1, C_2))$
$\alpha_6 : \forall C_1 : coll, C_2 : coll, C_3 : coll : (bigger(C_1, C_2) \wedge bigger(C_2, C_3)) \rightarrow bigger(C_1, C_3)$
...
Facts
$\alpha_7 : smallestCollection(singleton)$
...

Refining the generalisation

The formulae computed above by generalising Object Collection and Object Construction serve as a first formalisation of elementary arithmetic (labelled Arithmetic-1 in figure 2). The variables introduced by anti-unification are regarded as entities and predicates of this new domain. Because the grounded domains chosen were very similar, and in particular because the grounded domains neither have the concept EMPTY COLLECTION nor EMPTY CONSTRUCTION, the system computes only a subtheory of arithmetic that lacks a neutral element with respect to the operation P (representing ADDITION). A second step of creating analogical mappings is needed to transfer the concept ZERO from a differently structured domain into our Arithmetic-1. This is achieved by the second generalisation between the formalisation of Motion Along a Path and Arithmetic-1 resulting in Arithmetic-2 depicted in figure 2.

The formalisation we chose for Motion Along a Path is different from the other domains in that the predicates take an extra argument, *path*, to indicate the path along which the motion occurs. As a consequence, higher-order anti-unification is applied which leads to the slightly more complex substitutions τ_1 to τ_7 and κ_1 to κ_4. As before, these substitutions can be used to reconstruct the source and target domains from the generalisation. A further point to note is that γ_4 contains an additional condition in comparison to ρ_4. This mismatch is handled by introducing a generalised predicate Q, which is mapped to *false* by κ_4 and therefore can be neglected in Arithmetic-1. However, this dummy predicate is used as a hint for refinement. It indicates that an elaborated version of ρ_4 might be used to describe Arithmetic-1, namely

$$\rho_4' : \forall V_1, V_2 : S(V_1) \rightarrow (\neg Q(V_2) \rightarrow \neg bigger(V_1, V_2))$$

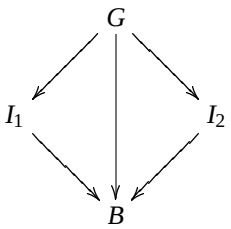
which mainly states that if V_1 is the smallest number, then either V_2 is ZERO or V_1 is not bigger than V_2. This new predicate Q can also be used for the transfer of additional formulae, e.g. γ_5 can be introduced into Arithmetic-1 resulting in

$$\rho_5 : \forall V_1, O : M(V_1, V_1, O) \rightarrow Q(O)$$

basically saying V_1 minus V_1 equals ZERO. Thereby, the basic ideas on ZERO are incorporated into Arithmetic-1 by refinement and transfer from the Motion Along a Path domain.

Goguen's notion of concept blending

Another important means to create new mathematical concepts is *concept blending*, in particular in the form presented by Goguen (2006). His figure 3, reproduced below, gives an overview. Each node in this graph corresponds to a *conceptual space* in the sense of Gärdenfors (2000), which, roughly speaking, is a subset of the system's knowledge. The arrows preserve the inferential structure from space to space, and the diagram commutes. Goguen does not discuss examples from arithmetic, but how from the concepts HOUSE and BOAT the concepts HOUSEBOAT and BOATHOUSE are created by concept blending. The G space contains generic elements, such as PERSON or OBJECT; the I spaces represent more specific conceptual spaces, in his example I_1 represents that a HOUSE is on LAND and that a PERSON LIVES in it and I_2 that a PASSENGER RIDES in a BOAT and that the BOAT is on WATER. Concept blending takes these conceptual spaces and maps them into another space (the B space) in a way that, for example, that the BOAT is mapped onto the PERSON LIVING in a HOUSE, resulting in the concept of a house in which the boat 'lives' – a BOATHOUSE.



Fauconnier and Turner (2002, pp. 242–245) discuss blends in arithmetic. Their presentation can be formulated in the form suggested by Goguen (they are a major influence on Goguens conception in the first place), thus giving an extension to the work described in this paper. For example, Lakoff and Núñez's extended version of the motion along a path metaphor supports an analogue of the rational numbers. Taking this as I_2 and object collection as I_1, a generalisation G can be found as above, which ignores the division operation of I_2. Forming the blend B then allows the extra operation to be incorporated into a conceptualisation which respects the generalisation. The blend can be seen as an updated view of I_1:

> Once we have the blend, and reify it, we can adopt the view that the previous conception of number was 'miss-

ing' several numbers that were 'there' but not yet 'discovered'. (Fauconnier & Turner, 2002, p. 244)

Conclusions and future work

We examined to which extent the cognitive processes underlying mathematical thinking can be made formally precise and algorithmically operationalised. For this purpose we took the mathematical metaphors of Lakoff and Núñez (2000) and used the analogy engine HDTP to compute generalisations from the basic source domains of arithmetic based on higher-order anti-unification.

For this, we used formalisations similar to the ones in our earlier approaches using Information Flow theory and created a first generalisation that contained the fundamental concepts of arithmetic. We extended the first generalisation produced by HDTP by incorporating a transfer of concepts, which added new concepts to the 'growing' domain of arithmetic (in our case the idea of a neutral element). Thus, anti-unification cannot only serve to find abstractions of two source domains but also to transfer concepts.

We also briefly described, how Goguen's concept blending is a direct extension of the HDTP approach. A paper detailing the role of concept blending for arithmetic and a treatment within the HDTP framework is currently submitted.

The demonstrated generalisation examples are still quite simple. Enriching the domains to get more interesting transfer candidates is therefore the next step. This notion of 'interestingness' is central to a comprehensive treatment of mathematical discovery, because there is an unlimited number of possible theorems and theories, but only a fraction of these are deemed interesting and useful enough for mathematicians to consider. For automated theorem provers, this is a hard problem; one on which we expect the heuristic nature of the HDTP engine will shed more light.

The grounded domains as we used them here are already generalisations of concrete situations, e.g. for the object collection domain the person/system must already have abstracted over concrete instances of the acts of putting collections together and realised that this is a general law holding in this domain. HDTP should be suited to create these abstractions as well. A more pressing and fundamental case seems to be, however, to create an abstract, generalised number concept that extends beyond the subitising range, i.e. those cardinalities (ranging from one to three or four) for which humans don't need to count but immediately perceive the number of objects and which seem to be innate.

Some other directions in which to extend our work are: (1) How are the results influenced by the order of generalisation? (2) How can the object construction domain be extended such that fractions (rational numbers) can be introduced into the domain of arithmetic? (This is Lakoff and Núñez's fraction extension of the basic metaphor.) (3) How can our approach be extended to include Lakoff and Núñez's *linking metaphors*, which are used for creating more abstract mathematical concepts.

Acknowledgments

The research reported here was supported by EPSRC grant EP/F035594/1 for the *Wheelbarrow* project.

References

Barwise, J., & Seligman, J. (1997). *Information flow: The logic of distributed systems*. Cambridge: Cambridge University Press.

Fauconnier, G., & Turner, M. (2002). *The way we think: Conceptual blending and the mind's hidden complexities*. New York: Basic Books.

Gärdenfors, P. (2000). *Conceptual spaces: The geometry of thought*. Cambridge, MA: MIT Press.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science, 7*(2), 155–170.

Gentner, D., Bowdle, B. F., Wolff, P., & Boronat, C. (2001). Metaphor is like analogy. In D. Gentner, K. J. Holyoak, & B. N. Kokinov (Eds.), *The analogical mind: Perspectives from cognitive science* (p. 199-253). Cambridge, MA: MIT Press.

Gentner, D., & Markman, A. B. (1997). Structure mapping in analogy and similarity. *American Psychologist, 52*(1), 45–56.

Goguen, J. (2006). Mathematical models of cognitive space and time. In D. Andler, Y. Ogawa, M. Okada, & S.Watanabe (Eds.), *Reasoning and Cognition: Proceedings of the Interdisciplinary Conference on Reasoning and Cognition* (pp. 125–128). Keio University Press.

Guhe, M., Pease, A., & Smaill, A. (2009). A cognitive model of discovering commutativity. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Guhe, M., Smaill, A., & Pease, A. (2009a). A formal cognitive model of mathematical metaphors. In B. Mertsching, M. Hund, & Z. Aziz (Eds.), *KI 2009: Advances in Artificial Intelligence* (pp. 323–330). Berlin: Springer.

Guhe, M., Smaill, A., & Pease, A. (2009b). Using Information Flow for modelling mathematical metaphors. In *Proceedings of the 9th International Conference on Cognitive Modeling.*

Krumnack, U., Schwering, A., Gust, H., & Kühnberger, K.-U. (2007). Restricted higher-order anti-unification for analogy making. In *AI 2007: Advances in Artificial Intelligence* (pp. 273–282). Berlin: Springer.

Lakatos, I. (1976). *Proofs and refutations: The logic of mathematical discovery*. Cambridge: Cambridge University Press.

Lakoff, G., & Núñez, R. E. (2000). *Where mathematics comes from: How the embodied mind brings mathematics into being*. New York: Basic Books.

Plotkin, G. D. (1970). A note on inductive generalization. *Machine Intelligence, 5*, 153–163.

Schwering, A., Krumnack, U., Kühnberger, K.-U., & Gust, H. (2009). Syntactic principles of Heuristic-Driven Theory Projection. *Journal of Cognitive Systems Research, 10*(3), 251–269.

Head and Hand Movements in the Orchestration of Dialogue

Stuart A. Battersby (stuart@dcs.qmul.ac.uk)
Queen Mary, University Of London
Interaction, Media & Communication Group
School Of Electronic Engineering & Computer Science
London, E1 4NS

Patrick G. T. Healey (ph@dcs.qmul.ac.uk)
Queen Mary, University Of London
Interaction, Media & Communication Group
School Of Electronic Engineering & Computer Science
London, E1 4NS

Abstract

Gaze and head orientation are considered to be the most important non-verbal cues people use to help manage the flow of conversation. However, if there are more than two participants, gaze and head orientation become problematic. People can only look at a single participant at a time. When speakers concurrently engage with more than one participant, they often make use of both head and hand orientation. We show two contrasts with existing findings. Firstly, people do not automatically look where the speaker is looking. Secondly, we demonstrate that hand movements are more important for the interaction than head movements. Specifically, changes in speaker hand orientation prompt quicker and more frequent responses from recipients than changes in head orientation.

Keywords: Dialogue; Non-verbal interaction; Multi-party; Gesture; Gaze; Simultaneous engagement;

Introduction

Consider the following situation: Ann, Bob and Claire are discussing a film that Bob and Claire went to see the previous night. Ann asks "Was it good"? Claire responds by saying "I really enjoyed it" while Bob simultaneously pantomimes a yawn. More than one person's responses are potentially relevant to the interpretation of the answer. Moreover, the orientation of each participant to those responses is also relevant. For example, it matters whether Bob is looking at Ann or Claire as he pantomimes a yawn and it also matters whether Claire is aware that he is looking at her when he yawns.

Putting puzzles about mutual-knowledge to one side, this example highlights the intuition that in multi-party interactions participants often face the challenge of simultaneously monitoring the responses of several people to each contribution (see Goodwin (1979)). People can also design their contributions in ways that directly convey how different participants stand in different relationships to what is said. In a variation of the example above, Claire might look at Ann and say "I really enjoyed it but Bob was bored" while simultaneously pointing toward Bob as she says his name (see Healey and Battersby (2009), for documented examples of this kind).

In the literature on non-verbal communication, a significant body of evidence has accumulated that shows gestures have managerial functions within dialogue (see Bavelas, Coates, and Johnson (2002) and Jokinen and Vanhasalo (2009)). However, eye gaze and, by association, head-orientation are normally singled out as the most important cues to the current orientation of participants in interaction (see, for example, Argyle (1975)). Kendon (1990) uses the term 'Face Address System' to make the claim that speakers use their gaze to identify the intended recipient of their utterance and Streeck (1993) observed that it is the speaker's gaze that addressees follow, potentially to the speaker's gesture. Langton, Watt, and Bruce (2000) reflect upon the claims about gaze and although they agree on its importance, suggest that gaze cues should be considered along with cues from the head orientation and hands.

Gullberg (2003) provides a quantitative estimate of the relative importance of a speaker's face and hands by measuring the eye-gaze patterns of addressees. Her live condition consisted of two people one of which had watched a cartoon. This person then retold the cartoon in narrative form to an addressee who had been configured with eye tracking equipment. The gaze patterns of this addressee were recorded. Only 7% of the speaker's gestures were fixated by the addressee. 96% of the time the addressee looked at the speaker's face; only 0.5% of the time was spent on their gestures with the remaining time spent looking at other objects in the room. Whilst this data points to a marked difference in the relative importance of the head and the hands, the interactional situation is different to open multi-party conversation.

Coordinating Multi-Party Interactions

Although eye gaze is an effective cue to focus of attention in dyadic (two-person) interactions it has more limited value in multi-party interactions. We can only look at one person at a time and we can only monitor the gaze of one person at a time. As Loomis, Kelly, Pusch, Bailenson, and Beall (2008) have shown, direction of eye gaze is difficult to estimate in the physical arrangements typical of conversation. In small group conversations people are only able to judge another's eye gaze direction with a maximum 4° retinal eccentricity whereas other people's head orientation can be judged effectively up to a 90° retinal eccentricity. This leads to the pre-

diction that, in multi-party conversations, auxiliary cues such as head and hand orientation should therefore be much more important to the conduct of the interaction.

Healey and Battersby (2009) describe how in three-way task-oriented dialogues speakers frequently use combinations of head and hand orientation to enable simultaneous engagement with two other participants. These moments of simultaneous engagement occurred on average once every 25 seconds. However, it is unclear what the consequences of the events are for the other participants in the interaction. Specifically, do these head and hand movements have any demonstrable impact on the responses of the other participants?

This paper addresses the question of whether changes in a speaker's orientation reliably prompt changes in the behaviour of the other participants. It also compares the relative impact of head and hand movements on other participants.

Method

Materials

All data was gathered in the Augmented Human Interaction (AHI) lab at Queen Mary. This lab houses a Vicon optical motion capture system consisting of an array of 12 infra-red cameras which track reflective markers attached to the clothing of participants. Each participant wears an upper body motion capture suit and a baseball cap with reflective markers attached. The motion capture system records the precise 3D coordinates of each marker at a rate of 60 frames per second (see Battersby, Lavelle, Healey, and McCabe (2008) for more details). Video cameras are placed above and to either side of the participants and are time synchronised with the motion capture system. Audio is recorded on the video cameras. Motion capture data from each interaction is time synchronised with the video data. A custom piece of software reads the motion capture data and integrates it with hand-annotation data from ELAN.

Participants

Participants were recruited from undergraduate and masters courses at Queen Mary and either received pay or modules credits and pay for their time. 33 participants (19 female and 14 male) aged between 18 and 30 took part. Each group consisted of three people meaning that the data presented are from 11 triads.

Task Description

Six tuition tasks were developed that consisted of a description of either a short Java program or a description of a system of Government. They were designed to involve an abstract hierarchy with no direct visual analogue. All material was text based with no graphical descriptions.

Procedure

Each group completed three rounds, based on either three Java or three Government tasks. On the first round one member of the triad is randomly assigned to a 'learner' role and the other two participants are assigned 'instructor' roles. These roles are then rotated on subsequent rounds so that each participant is as a learner once and an instructor twice. The instructors are asked to collaborate to teach the leaner the structure described in the task description. The learner is removed from the group to another room whilst the instructors are given the descriptions of the task for next round. Once the instructors signal that they understand the task, the descriptions are returned to the experimenter and the learner rejoins the group. All three participants are seated on pre-positioned stools in the AHI lab (see Fig. 1).

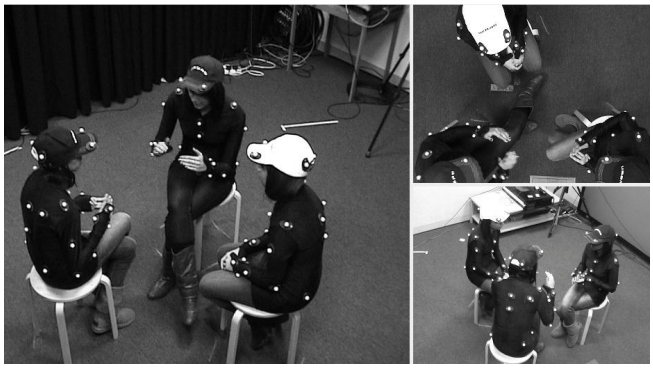


Figure 1: Three participants wearing upper body motion capture suits.

There was no time limit on the tuition stage and no restrictions on the interaction other than they were not allowed to use pens or paper. The participants notified the experimenter when they finished each round of tuition. To motivate the participants to adequately teach and learn the material a post completion test (comprising of a drag and drop arrangement of the classes for the computer program, or some multiple choice questions for the government material) was given after each round. Tasks were systematically rotated across groups and the order of the printed sheets of paper was randomised before each round.

Hand Coding of Target Events

All interactions were recorded on video, with cameras above and either side of the group, using synchronised video recording. ELAN was used to hand code these videos. The recordings were coded for all instances of simultaneous engagement in which a speaker who is making a gesture visibly changes the orientation of their hands or head with respect to the other participants. For example, by turning their hand from one participant to another or changing their head orientation. These changes were coded as:

- Head Moves: Here the head orientation changes but the gesture remains stationary
- Hand Moves: Here the gesture orientation changes, but the head orientation remains stationary
- Both Move: Here both the gesture and the head orientation shift

Motion Data Analysis

Taking the hand coded target events for the speaker as the starting point, the motion capture data was used to provide quantitative measures of recipients' responses to target events.

Assigning Recipient Role The motion data was used to provide an operational definition of recipient role. Recipients are either primary or secondary recipients. This role is judged by the head orientation of the speaker. We project a vector from the middle of the forehead for each speaker. The orientation of this vector is compared to a centre line between the two recipients. The primary recipient is defined as the recipient who is on the same side of the centre line as the speaker's current head orientation (see Figure 2).

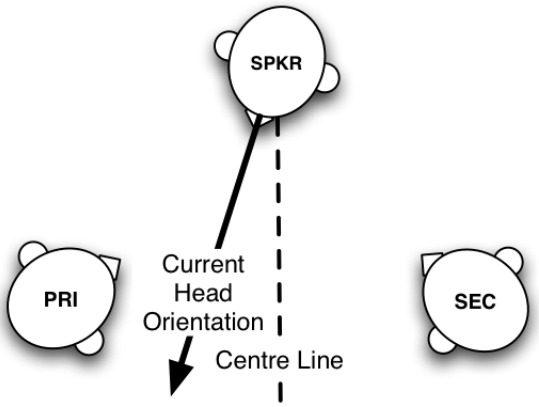


Figure 2: Defining primary and secondary recipients

Indexing Head Orientation Responses It is impossible to be sure exactly which head movements correspond to changes of orientation by the recipients. Instead, we set a criterion for counting movements as changes of orientation based on a vector projected from each recipient's head as it was for the speaker. A change in orientation is thus defined as a shift of head orientation that crosses the centre line between the speaker and the other recipient (see Figure 3).

Indexing Nod Responses A second index of responses, 'nods', was also generated from the motion capture data. As for changes in head orientation it is impossible to be sure when a head movement really constitutes a nod or is simply a shift in position or unintentional motion of some kind. As for

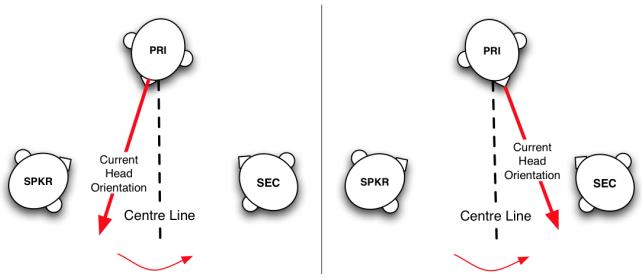


Figure 3: Indexing head orientation responses

changes in head orientation we set a criterion level of movement of a single frontal head marker in the vertical axis (see Figure 4 for some sample movement). Only movement with a frequency of between 2Hz and 8Hz is used. This removes some of the effects of gross body sways (below 2Hz) and very minor body shakes or fluctuations in data from the cameras (above 8Hz). Movements with an amplitude greater than 5cm are removed as these could likely be a result of shifts in position. The resulting signal, which is smoothed using a window size equivalent to 0.5 seconds, is used to represent periods of head movement that approximate nodding.

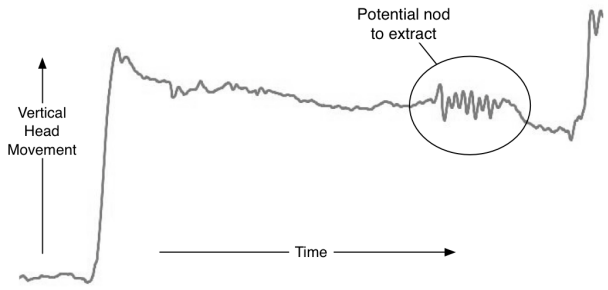


Figure 4: Raw head movement motion capture data. An area of potential nodding is circled.

In order to analyse frequency of responses to each simultaneous engagement event by the speaker we create a 5 second window after the event and score, for each recipient, whether a head re-orientation and whether a head nod occurs in that window. In order to provide a measure of response latency we record the first change of head orientation or nod that occurs after the target event and before another target event occurs.

Baseline Response Rates In order to interpret the measures of responses to the target events, it is important to know what the baseline likelihood of a recipient nodding or changing orientation is. To provide this a control comparison sample was created by randomly selecting points where someone was speaking but not producing a target event. Recipient

responses after these control points were then analysed in exactly the same way as for the target events.

Results

The total time for all the dialogues was 2 hours and 54 minutes, each task took on average 5 minutes and 16 seconds. A total of 287 target, simultaneous engagement events involving a change in orientation of the speaker were identified.

Inter-rater Agreement

In order to check the reliability of the hand coding of event types by the 1st author, a random sample of 25 events taken from experimental and control data was independently coded for event type by a second coder. The inter-rater reliability was good with Kappa = 0.78,($p < 0.001$). The number of each type of target event is show in Table 1.

Table 1: Number of Changes in Speaker Orientation

Event Class	Count
Head Moves	170
Both Move	86
Hand Moves	31

Recipient Responses

In analysing responses to changes of speaker orientation we distinguish the task role of the recipients (learner or instructor) and their recipient status at the time of the event (primary or secondary). In addition we code whether each recipient is oriented toward the speaker or the other recipient at the time the simultaneous engagement event, i.e. the change in speaker orientation, begins. These judgements are made using the motion capture data.

Recipient Orientation

As Table 2 shows, at the point when the speaker initiates a change of orientation, the primary recipient is more likely to be looking at the speaker than the secondary recipient. The secondary recipient, by contrast, is equally likely to be looking at the other participants ($\chi^2 = 16.9$, $p < 0.001$)).

Table 2: Initial Head Orientations

Recipient Role	Oriented To Speaker	Oriented To Other
Primary Recipient	68.0%	32.0%
Secondary Recipient	50.2%	49.8%

Response Frequency

In contrast to recipient orientation (and our preliminary findings (Healey & Battersby, 2009)), there was no difference between the response rates of the primary and secondary recipients. Both were equally likely to respond.

Combining the responses of the two recipients together we can compare the overall frequency of response to a target coordination event with the baseline response rate. For changes in head orientation the recipients' baseline response rate is 41.3% and their response rate to target events is 48.6%; a small but reliable difference ($\chi^2 = 5.75$, $p < 0.05$).

Table 3 illustrates the differences in response rate for each type of event.

Table 3: Response rates by type of event, measured by recipient reorientations

Event Class	Response Rate	Baseline Response Rate	Sig
Head Moves	43.1%	41.3%	Not Significant
Both Move	56.2%	41.3%	$\chi^2 = 10.26$, $p < 0.01$
Hand Moves	63.0%	41.3%	$\chi^2 = 8.14$, $p < 0.01$

For target events in which only the head changes orientation there is no significant increase in response rate (measured by a shift in recipient head orientation) relative to the baseline rate. However, for targets events that involve changes to both gesture and head orientation we see a significant difference of 14.9% between the baseline and the target event. Where only the gesture changes orientation there is a 21.7% increase in response rate.

A slightly different pattern is evident in the head nodding response measure. Combining target events, recipients respond 72.4% of the time compared to a background response rate of 66.0% ($\chi^2 = 5.08$, $p < 0.05$). The breakdown by type is shown in Table 4.

Table 4: Response rates by type of event, measured by recipient nodding

Event Class	Response Rate	Baseline Response Rate	Sig
Head Moves	70.0%	66.0%	Not Significant
Both Move	73.6%	66.0%	Not Significant
Hand Moves	87.0%	66.0%	$\chi^2 = 8.51$, $p < 0.01$

In order to provide a direct comparison of the recipient's relative sensitivity to changes in the speaker's head and hand orientation responses to 'Head Moves' events and 'Hand Moves' events can be compared. This shows a significant difference between the groups using the values for both head

re-orientations and head nods as a measure of response shown above ($\chi^2 = 6.43, p < 0.02$ and $\chi^2 = 5.75, p < 0.02$ respectively).

Response Latency

The time elapsed between a target event until the first response (nod or change of head orientation) for each recipient was analysed in a Mixed Model linear analysis with Recipients and Task as random factors and 'Condition' (Target Event vs. Baseline) and Task Role (Learner vs Instructor) as within subjects factors. This showed a reliable main effect of Condition ($F_{(1,1089)}$ = 14.88, p = 0.00) but no main effect of Task Role ($F_{(1,1088)}$ = 1.29, p = 0.25) and no Task Role $\times$ Condition interaction ($F_{(1,1078)}$ = 0.39, p = 0.53).

As Table 5 shows, recipients' responses to target events were approximately 1 second faster than the baseline responses.

Table 5: Marginal Means for Recipient Response Times

Condition	Marginal Mean	Standard Error
Target Event	2.4 seconds	0.37
Baseline Event	3.4 seconds	0.35

Discussion

The results show two important contrasts with existing findings on non-verbal cues and the co-ordination of interaction. First, in the dialogues reported here people do not automatically look where the speaker is looking. In fact, in the cases where the speaker only changes their head orientation there is no reliable shift in recipient's head-orientation. The second key finding is that changes of hand orientation are significantly more likely to invoke a response from the recipients than changes in head orientation; the opposite of what would be predicted on the basis of previous work.

The results also show that recipients are demonstrably responsive to the target events, but with a pattern of responses that is different to that typically described in the literature. This provides support for the claim that they are distinctive and significant interactional events. Although it is difficult to generalise beyond the particular task we have used, it seems likely that these moments of simultaneous engagement are a response to the demands of co-ordinating a conversation with multiple participants. As Healey and Battersby (2009) note, they are also distinguished by relying on physical co-presence in mutually shared space as a specific resource for interaction. For example, they cannot be deployed in point-to-point video communication.

Our analysis suggests that recipient role (primary or secondary) can manifest itself non-verbally. Whilst hand movements are more marked than head movements in initiating recipient responses, we see differing patterns of recipient head orientation through the dialogue. The primary recipient is more likely to be looking at the speaker than they are to be looking at the secondary recipient before a simultaneous engagement event occurs. Secondary recipients do not share this pattern though, and are equally likely to be looking at either party. It is interesting that this distinction between roles is not found when measuring responses, perhaps suggesting that the target events unify the recipients' behaviour.

The clear difference between our data and that of previous findings is the introduction of the third person. It would be intuitive, and logical, to understand the conflicting results with the statement that multi-party dialogue is simply different to dyadic dialogue. Whilst this is true, there is also the possibility that multi-party dialogue only allows us to see fully the underlying process that is present in *all* dialogue; dyadic interaction simply masks them.

Conclusion

We examined a corpus of multi-party dialogues comprising of video and motion capture data. Moments where the speaker simultaneously engaged both recipients were coded for. These events were broken down by changes in the speaker's orientation of their head, their gesture or both and the significance of these changes for the recipients was examined. These changes in speaker orientation were shown to hold interactional significance. In contrast to existing findings in the literature, movements of the hands elicited a higher and faster response rate than movements of the head.

References

Argyle, M. (1975). *Bodily Communication*. Bristol: Methuen & Co. Ltd.

Battersby, S. A., Lavelle, M., Healey, P. G. T., & McCabe, R. (2008, May). Analysing Interaction: A comparison of 2D and 3D techniques. In *Conference on multimodal corpora.* Marrakech.

Bavelas, J. B., Coates, L., & Johnson, T. (2002). Listener Responses as a Collaborative Process: The Role of Gaze. *Journal of Communication*, *52*, 566–580.

Goodwin, C. (1979). The interactive construction of a sentence in natural conversation. In G. Psathas (Ed.), *Everyday language: Studies in ethnomethodology* (pp. 97–121). Irvington Publishers.

Gullberg, M. (2003). Eye movements and gesture in human face-to-face interaction. In J. Hyönä, R. Radach, & H. Deubel (Eds.), *The mind's eye: Cognitive and applied aspects of eye movements* (pp. 685–703). Oxford: Elsevier.

Healey, P. G. T., & Battersby, S. A. (2009). The Interactional Geometry of a Three-way Conversation. In *Proceedings of the 31st annual conference of the cognitive science society* (pp. 785–790). Amsterdam.

Jokinen, K., & Vanhasalo, M. (2009). Stand-up Gestures Annotation for Communication Management. In *Nodalida*

2009 workshop multimodal communication: from human behaviour to computational models (pp. 15–20).

Kendon, A. (1990). *Conducting Interaction: patterns of behavior in focused encounters*. University of Cambridge.

Langton, S. R. H., Watt, R. J., & Bruce, V. (2000). Do the eyes have it? Cues to the direction of social attention. *Trends in Cognitive Sciences*, *4*(2), 50–59.

Loomis, J. M., Kelly, J. W., Pusch, M., Bailenson, J. N., & Beall, A. C. (2008). Psychophysics of perceiving eye and head direction with peripheral vision: Implications for the dynamics of eye gaze behaviour. *Perception*, *37*, 1443–1457.

Streeck, J. (1993). Gesture as Communication I: Its Coordination with Gaze and Speech. *Communication Monographs*, *60*(275-299).

Tracking Lexical and Syntactic Alignment in Conversation

Christine Howes, Patrick G. T. Healey and Matthew Purver
{chrizba,ph,mpurver}@dcs.qmul.ac.uk
Queen Mary University of London
Interaction, Media and Communication Group
Mile End Road, London E1 4NS

Abstract

As much empirical work attests, people have a reliable tendency to match their conversational partner's body movements, speech style, and patterns of language use – amongst other things. A specific version of this tendency, *Structural priming*, which occurs when prior exposure to a particular linguistic structure facilitates one's subsequent processing of the same structure, has gained widespread acceptance. Pickering and Garrod (2004) propose that *cross*-person structural priming is a basic mechanism of conversational coordination – part of an automatic, resource-free *alignment* mechanism that is the basis for all successful human interaction. We present evidence to the contrary from two analyses of a corpus of ordinary conversation. The first suggests that the level of structural (syntactic) matching is no different from chance, and the second that the observed statistical correlation between prime form and target form may be entirely associated with repetition of lexical form.

Keywords: structural priming; alignment

Introduction

The apparent tendency for speakers to repeat their own or others syntactic or structural choices in conversation – a phenomenon referred to as *structural* or *syntactic* alignment – has been a subject of particular scrutiny (see Pickering and Ferreira (2008) for an overview).

The evidence for such alignment in dialogue comes from two main sources: experimental studies of task-oriented dialogue (e.g. (Branigan, Pickering, & Cleland, 2000)), and corpus studies that track frequency of use of these same constructions in language use outside of the laboratory setting (e.g. (Gries, 2005)).

In the basic experimental set-up of Branigan and colleagues, there are two participants, one of whom is a confederate of the experimenter. The participants describe picture cards to each other, the critical items of which require the use of ditransitive verbs in their descriptions. In English, there are two syntactic structures which can be used; one a double object structure ("The thief giving the nurse the banana"), and the other using a preposition ("The thief giving the banana **to** the nurse"). The confederate uses a scripted description of the ditransitive prime sentences, thus manipulating which type naive subjects are exposed to. Participants are more likely to use the type of structure that they have just used or been exposed to. This has been found to hold across comprehension and production (Branigan, Pickering, Stewart, & McLean, 2000; Bock, Dell, Chang, & Onishi, 2007), from main clauses to relative clauses (Branigan, Pickering, McLean, & Stewart, 2006) and even across languages in bilingual speakers (Hartsuiker, Pickering, & Veltkamp, 2004). Different factors found to increase the *strength* of syntactic alignment include the distance between the prime and the target, participant role (Branigan, Pickering, McLean, & Cleland, 2007) and, importantly for the interactive alignment model (see below), reuse of the same or semantically related lexical items (Branigan, Pickering, & Cleland, 2000).

In a corpus study using the International Corpus of English (ICE-GB), Gries (2005) looked at the same syntactic alternation. His data show that there is a tendency to reuse the form of a ditransitive verb most recently encountered (double object or prepositional), in line with the experimental results. Similar results have been found to hold with different constructions such as particle placement of phrasal verbs (Gries, 2005), future markers ("will" versus "going to") and comparatives ("cleverer" versus "more clever than") (Szmrecsanyi, 2005).

Pickering and Garrod (2004) argue, in their *Interactive Alignment* model, that alignment is the basis for successful communication; "successful dialogue occurs when interlocutors construct similar situation models to each other" (Pickering & Garrod, 2006, p206). In order to do this, interlocutors *align* on situation models; however, as this alignment is not usually negotiated explicitly, it is hypothesised to arise automatically from local alignment, via resource-free *priming* mechanisms.[1] Alignment at local levels, including lexical (repetition of words) and the syntactic alignment discussed above, "percolates", leading to alignment at other levels.

From priming to alignment?

There are three problems with using these studies to support the claim that cross-speaker structural priming is ubiquitous in conversation. First, automatic priming predicts an increase in matching of *all* structures across turns, but this claim has not been directly tested. For practical reasons, experimental studies have focussed on situations in which specific syntactic alternatives can be used to describe the same situation. Similarly, corpus studies have tended to track the frequency of use of spe-

[1] Note that the observed effects are alignment effects; priming mechanisms are their hypothesised cause, leading to two distinct questions – does such alignment occur; and if it does, is it caused by priming?

cific constructions across participants and time, rather than addressing whether or not people tend to match one-another in general (e.g. Gries (2005)). One exception is Reitter, Moore, and Keller (2006), who examined general syntactic similarity, but results were unclear. Reitter et al. (2006), used two corpora, one task-specific (Map Task) and one more general (Switchboard), and saw a large difference: while *same*-person priming was found in both datasets, *cross*-person priming was found only in the task-specific dialogues.[2]

The second problem is that the data used in these studies is not adequately representative of ordinary dialogue. As Pickering and Garrod (2004, p187) say:

> The interactive alignment model was primarily developed to account for tightly coupled processing of the sort that occurs in face-to-face spontaneous dyadic conversation between equals with short contributions. We propose that in such conversation, interlocutors are most likely to respond to each other's contributions in a way that is least affected by anything apart from the need to align.

However, in the experiments, the confederate is scripted, and the naive participants were told that if they didn't understand they "could say "Please repeat," but nothing else" (Branigan et al., 2007, p175). And while corpora can provide more spontaneous data, Gries (2005)'s corpus is biased towards written and spoken *monologue*, and a significant proportion of the dialogues it samples involve specialised institutional settings, e.g. legal cross-examinations and broadcast interviews.

The third problem is that these studies have not used a control condition. As a result the chance level of structural matching is unknown and effects such as conversational genre cannot be discounted (cf. Tannen (2007)).

In order to address two of these issues,[3] we conducted an experiment which tested the degree of match of dative alternation structures in a corpus of naturally occurring dialogue data. We compared this measure to control conditions for the same genuine conversational data manipulated to create 'dialogues' from turns actually occurring in different conversations (see below).

Experiment 1

Method

The corpus used here is the Diachronic Corpus of Present-Day Spoken English (DCPSE). This consists of 885,436 words together with a full set of parse trees that have been hand-checked by linguists. It includes several distinct genres of dialogue. We consider the two-person portions of the three largest samples: Face-to-Face Formal (90,000 words), Face-to-Face Informal (403,000 words) and Telephone Conversations (47,000 words). This gives us 127 dialogues with an average of 45.24 turns per person (per dialogue), which ought to provide us with the data most likely to exhibit alignment phenomena (see above).

Creating control dialogues In order to discount the potential biasing effect of conversational structure (e.g. recurrent patterns of turn-taking, topic shifts, openings and closings) on syntactic similarity, a control condition that captures how similar two people's conversational turns would be by chance is needed. For each 'real' dialogue in each genre in the corpus, we therefore create two types of 'fake' control dialogue. For the first, the *random-speaker* control, one speaker's turns are kept and interleaved with the turns of another speaker from a different dialogue (matching dialogues by genre, matching by length as closely as possible, and discarding any 'unmatched' turns). This 'fake' dialogue thus maintains turn order for each speaker; but consists of the turns of two speakers who did not, in fact, interact. For the second 'fake' dialogue, the *random-sentence* control, a new dialogue of the same length is created by randomly choosing sentences, each time allowing a new choice of dialogue and speaker (but always matching dialogue genre). This 'fake' dialogue thus maintains neither turn order nor speaker identity (see table 1 for comparisons).

Table 1: Real and control dialogues comparison

GENUINE DIALOGUE:
A: Are you going to go to all of the phonology lectures
B: I think I ought to do that
A: Yes. I think you had. Yeah
B: I mean I don't know how much I'll take in
A: I think I'll go to most of them. But I won't go to all of pragmatics the day before
RANDOM-SPEAKER CONTROL DIALOGUE:
A: Are you going to go to all of the phonology lectures
C: Well uh ask one of the stallholders down Chapel Street. They'll all know
A: Yes. I think you had. Yeah
C: Uhm I was down there the other day and I got some excellent salmon
A: I think I'll go to most of them. But I won't go to all of pragmatics the day before
RANDOM-SENTENCE CONTROL DIALOGUE:
A: Are you going to go to all of the phonology lectures
D: Uhm one of the few. Oh George was impossible
E: Just normal water
F: Yes. What do they call it
G: Oh dear. It does not bode very well

Creating these control dialogues allows us to compare the syntactic similarity observed in the real data with the similarity that would be observed by chance. By choosing a suitable similarity metric, we can express the

[2] In fact, the opposite appeared to hold in the general corpus – participants seemed to *avoid* repeating each others' syntactic structure.

[3] The first issue we address in additional work; see e.g. Healey, Howes, and Purver (2010); Healey, Purver, and Howes (2010).

average similarity observed between turns or between speakers; and examine the difference between the real and control corpora. Choosing a general syntactic similarity metric (which takes all observed structural rules into account) would allow us to compare with Reitter et al. (2006); see e.g. Healey, Purver, and Howes (2010). In this paper, we only consider the specific ditransitive alternation discussed above, allowing comparison with Branigan, Pickering, and Cleland (2000) and Gries (2005).

Metric and predictions Considering only a single syntactic phenomenon gives us essentially a binary metric: a target sentence scores 1 if it reuses the form of the most recent prime sentence, and 0 otherwise. More concretely, each sentence is given a score of 1 only if:

1. it uses one possible form of the phenomenon in question: a double-object or prepositional-object construction; and
2. the most recent prior sentence in the same dialogue which exhibits the same phenomenon also uses the same form.

and 0 otherwise. Summing sentence scores and normalising by the number of sentences gives us the score for each individual in a dialogue. These scores can then be compared between the real and control corpora.

We test three key predictions:

1. Priming: Sentences in real conversations should display reliably more turn-by-turn structural matching than would occur by chance.
2. Person: Structural matching should be observed both between sentences produced by the same participant, and between those produced by different participants.
3. Genre: Relatively restricted registers should promote a higher level of cross-speaker structural matching than less restrictive registers.

Results

Two different analyses were carried out: the first compares real levels of matching against the control dialogues as outlined above, and the second compares the level of (real) same-person matching against (real) other-person matching.

In order to test predictions on Priming (1) and Genre (3) the average turn-by-turn syntactic similarity scores for each dialogue participant[4] in each Genre were analysed in a mixed analysis of variance with Dialogue Type (Real × Control) as a within subjects factor and Genre (Face-to-Face Formal × Face-to-Face Informal × Telephone Conversations) as a between subjects factor.

[4]Shown as N in tables 2 and 3. As we were only looking at 2-person dialogues this equates to 127 dialogues overall.

For overall similarity (this measure includes both same-person and other-person matching), the analysis showed no reliable difference between the Real and Control (i.e. 'fake') dialogues (random-sentence control: $F_{(1,251)} = 1.067, p = 0.30$, random-speaker control: $F_{(1,251)} = 0.11, p = 0.92$),[5] no significant main effect of Genre (random-sentence control: $F_{(2,251)} = 1.279, p = 0.28$, random-speaker control: $F_{(2,251)} = 1.881, p = 0.16$) and no interaction between Dialogue Type and Genre (random-sentence control: $F_{(2,251)} = 0.213, p = 0.81$, random-speaker control: $F_{(2,251)} = 0.809, p = 0.45$). The absolute levels of syntactic matching of the dative alternation were not reliably different from chance (see Table 2). There were also no significant results when comparing only cross-person similarity with its control condition.

Comparing same-person versus cross-person similarity using a mixed analysis of variance with Speaker (Same × Other) as a within subjects factor and Genre as a between subjects factor showed a reliable difference between the Same and Other person ($F_{(1,251)} = 4.124, p = 0.043$), no significant main effect of Genre ($F_{(2,251)} = 1.058, p = 0.35$) and no interaction between Dialogue Type and Genre ($F_{(2,251)} = 0.499, p = 0.61$) (see Table 3). This means that there is reliably more matching to one's own prior utterances than to another person's.

Discussion

These results seem to show that, at least for the dative alternation construction, in contrast to hypothesis (1), sentences in the DCPSE do not show reliably more structural matching than would occur by chance. In regards to (2), the overall level of same-person matching was higher than that of other-person matching (in line with experimental findings that production-production priming is higher than comprehension-production). However due to the control conditions used, it is not possible to ascertain whether the same person matching on its own is reliably higher than chance (though recall that both other person matching and overall levels of matching were not). As for hypothesis (3), although it appears from tables 2 and 3 that there is greater matching in the more restricted registers as predicted, pairwise comparisons did not show any significant effects. This could be due to the relatively small values, and limited number of cases, and further work is necessary to see if this is a genuine effect.

As the observed power values were in some cases as low as 0.2, we cannot reject the null hypothesis outright. Power calculations suggest that we require four times more data in order to be able to do so, and to this end we are currently conducting analyses on the

[5]For completeness we report exact probabilities but throughout adopt a criterion probability level of < 0.05 for accepting or rejecting the null hypothesis.

Table 2: Mean Dative Alternation Similarities

Dialogue Type	N	Real Similarity	(s.d.)	Random-Sentence	(s.d.)	Random-Speaker	(s.d.)
Face-to-Face Formal	60	0.017	(0.016)	0.013	(0.016)	0.018	(0.019)
Face-to-Face Informal	94	0.013	(0.015)	0.012	(0.014)	0.014	(0.016)
Telephone Conversation	100	0.012	(0.025)	0.012	(0.018)	0.011	(0.022)
Overall Mean	254	0.014	(0.019)	0.012	(0.016)	0.014	(0.019)

Table 3: Mean Dative Alternation Similarities

Dialogue Type	N	Same Person	(s.d.)	Other Person	(s.d.)
Face-to-Face Formal	60	0.010	(0.014)	0.007	(0.010)
Face-to-Face Informal	94	0.008	(0.012)	0.005	(0.007)
Telephone Conversation	100	0.007	(0.012)	0.006	(0.019)
Overall Mean	254	0.008	(0.012)	0.006	(0.014)

British National Corpus (BNC), which includes 2884 2-person conversations (Healey, Purver, & Howes, 2010). Another alternative to increase power would be to treat each occurrence of either form of the dative alternation as a separate datapoint, as Gries (2005) did, rather than taking an overall value per person per conversation. Experiment 2 reports such an approach.

Experiment 2

These results suggest that there is little or no priming above chance for the dative alternation in ordinary dyadic conversation. Prima facie, this is inconsistent with the evidence from Branigan, Pickering, and Cleland's experiments and also Gries' corpus study on the same constructions.

Other than the power issues discussed above, these differences could be due to differences in the data used. Whilst our natural conversational data is obviously different from the task specific experiemntal data, it is also different to the corpus data used by Gries, in one important respect. Although the DCPSE corpus overlaps with the ICE-GB corpus used by Gries, the data in the DCPSE is all spoken, while the ICE-GB contains a mixture of written and spoken data.[6] Additionally, our experiment 1 used only dyadic (two-person) dialogues, as this makes creation of the control corpora more straightforward.[7]

Method

A further study was therefore carried out, following Gries' methodology but using the DCPSE, to attempt to replicate his positive results. We once again restricted the analysis to the three largest genres, but this time included all conversations in those genres (i.e. we did not restrict this to dyadic conversation as in experiment 1, but still discounted e.g. broadcast interviews, legal cross-examinations and spontaneous commentaries, which would also have been included in Gries' data; see table 4). Following Gries, prime-target pairs in the DCPSE were coded for the variables shown in table 5, using the DCPSE's ICECUP tool to detect particular forms based on fuzzy tree fragments (Nelson, Wallis, & Aarts, 2002).

Table 4: Comparison of corpus data used

	Spoken	Written	Total prime/ target pairs
Gries (2005)	600,000	400,000	3003
This paper	540,000	N/A	1438

Results

The general result, as for Gries, is the significant effect between CPrime and CTarget ($\chi^2_{(1)} = 10.573, p = 0.001$), as shown in table 6. We observe priming for both the ditransitive and prepositional dative forms: observed target frequencies of each are greater than expected frequencies when following a prime of the same form, and lower than expected when following a prime of the other form.

The variables in table 5 were entered into a General Linear Model (GLM) analysis with CTarget as the dependent variable and CPrime, VFormID, VLemmaID, SpeakerID as independent variables and Distance as a covariate. Like Gries (2005), we found a main effect of CPrime ($F_{(1,1425)} = 76.364$, $p = 0.000$) as expected given the general result above, and indicating that the form of the prime strongly predicts the constructional choice of the target, and an interaction effect of CPrime

[6] Note, however, that (Gries, 2005) did not find any significant effect of Medium.

[7] Although one might expect that priming would be stronger in the canonical two-person case – see (Pickering & Garrod, 2004).

Table 5: Variables

Variable	Description
CPrime	the form of the prime (ditransitive v prepositional dative)
CTarget	the form of the target (ditransitive v prepositional dative)
CID	yes if CPrime and CTarget are the same form, no otherwise
Distance	the number of parsing units between prime and target
VFormID	yes if the verb and its form were identical in prime and target, no otherwise
VLemmaID	yes if the verb lemma was identical in prime and target, no otherwise
SpeakerID	yes if the speaker of prime and target was the same person, no otherwise

× VLemmaID, ($F_{(1,1425)} = 28.969$, $p = 0.000$) indicating that when the verb lemma is identical across prime and target, the effect of priming is stronger. We did not find an effect of CPrime × SpeakerID, as Gries did, however, this could be due to the different corpora used, as written material would inevitably only include cases where the producer of prime and target are the same (note also that the effect he found was a marginal one).

Following Gries, a second analysis using CID as a dependent variable was carried out. There was a significant main effect of CPrime ($F_{(1,1425)} = 4.935$, $p = 0.026$), the direction of which suggests that there is more likely to be an identical target following a ditransitive prime than a prime in the form of the prepositional dative. There was also a significant main effect of VLemmaID ($F_{(1,1425)} = 27.255$, $p = 0.000$), such that the target is more likely to have the same form as the prime if the verb lemma used is the same. Like Gries, we did not find an effect of distance when it was entered into the model linearly, but when transformed to a logarithmic scale, it had a significant effect on CID ($F_{(1,1425)} = 4.540$, $p = 0.033$). Adding Genre to the model did not reveal any additional effects to those outlined above.

Table 6: Observed v expected frequencies

CPrime:	CTarget: Ditran		CTarget: Prep		Total
Ditran	527	(497.1)	319	(348.9)	846
Prep	318	(347.9)	274	(244.1)	592
Total	845		593		1438

These results suggest that whilst there are genuine alignment effects being observed, due to the large effect of VLemmaID we cannot rule out the possibility that they are lexically specified, or collocational, rather than specifically syntactic or structural. To test this possibility, two post-hoc analyses were carried out. When the prime-target pairs which have an identical lemma are removed from the analysis, there is no longer any effect of CPrime on CTarget ($F_{(1,1211)} = 0.563$, $p = 0.45$), and there are also no other significant effects. See also table 7 ($\chi^2_{(1)} = 0.454, p = 0.50$). Conversely, looking just at those with an identical lemma we get a large effect of CPrime on CTarget ($F_{(1,1211)} = 171.358$, $p = 0.000$), as is obvious from table 8 ($\chi^2_{(1)} = 105.6, p = 0.000$). Note that these findings do not, in fact, contradict Gries (2005), as his major finding was that individual verbs differ in their sensitivity to priming effects, a finding that is supported by the evidence that the variation in our data can be accounted for by those cases in which the lemma is identical between prime and target.

Table 7: Observed v expected frequencies of prime-target pairs where LemmaID = no

CPrime:	CTarget: Ditran		CTarget: Prep		Total
Ditran	370	(375.8)	304	(298.2)	674
Prep	308	(302.2)	234	(239.8)	542
Total	678		538		1216

Table 8: Observed v expected frequencies of prime-target pairs where LemmaID = yes

CPrime:	CTarget: Ditran		CTarget: Prep		Total
Ditran	157	(129.4)	15	(42.6)	172
Prep	10	(37.6)	40	(12.4)	50
Total	167		55		222

Conclusions

The results show that, in ordinary dyadic conversation, there is no unequivocal evidence of syntactic priming effects for the specific constructions that have been the focus of previous experimental and corpus work. The results presented here show that individual people do tend to repeat the same structure. However, they are no more likely to converge on the same version of each structure with their conversational partners than would be expected by chance. In addition, the overall likelihood of a match in syntactic structure across turns appears to be accounted for by the repetition of specific words.

Our results seem to be inconsistent with previous findings, however, as already noted, there may be several reasons for this disparity. Firstly, laboratory based experiments on dialogue are always subject to concerns about

ecological validity and it's possible that the restricted, task-oriented, exchanges used in previous studies do not generalise well to the more open-ended dialogue samples in the corpus data. Note though, that the present results do replicate the strong effects of lexical choice on syntactic similarity reported by Branigan, Pickering, and Cleland (2000). Another point of contrast between the current study and previous work are the specific characteristics of the corpus we use. Our data only includes exchanges in ordinary dialogue (and is further restricted in experiment 1 to dyadic exchanges). We specifically exclude spoken monologue, institutionally specialised contexts such as tutorials and broadcast interviews and one-sided interactional activities such as story-telling. Note however, that in doing so we focus on just those cases where Pickering and Garrod (2004) predict that priming should be strongest.

Our data are also compatible with studies on *lexical* alignment – reuse of previously encountered *words*. Despite well documented experimental evidence of lexical alignment (Brennan & Clark, 1996), there are also questions as to how this scales up to genuine conversation – a study of relative lexical overlap in conditions allowing or prohibiting verbal feedback (Hadelich, Branigan, Pickering, & Crocker, 2004) found that in the conditions which were more akin to genuine dialogue (where verbal feedback was permitted), there was less relative lexical overlap.

Additionally, our experiment 2 is in fact an extension of Gries' (2005) work, and completely compatible with it, though it does suggest a shift of focus. While a statistical correlation between prime form and target form is observable, this may be almost entirely associated with repetition of lexical form, rather than reuse of syntactic structure *per se*.

While there is insufficient data in the DCPSE corpus to definitively prove that structural priming effects are absent in ordinary conversation, these results indicate that the strength and ubiquity of structural priming (see e.g. Pickering and Ferreira (2008)) may have been overstated.

Acknowledgements

The research presented here was carried out as part of the *Dynamics of Conversational Dialogue* project, funded by the UK ESRC (RES-062-23-0962).

References

Bock, K., Dell, G., Chang, F., & Onishi, K. (2007). Persistent structural priming from language comprehension to language production. *Cognition*, *104*(3), 437–458.

Branigan, H., Pickering, M., & Cleland, A. (2000). Syntactic co-ordination in dialogue. *Cognition*, *75*, 13–25.

Branigan, H., Pickering, M., McLean, J., & Cleland, A. (2007). Syntactic alignment and participant role in dialogue. *Cognition*, *104*(2), 163–197.

Branigan, H., Pickering, M., McLean, J., & Stewart, A. (2006). The role of local and global syntactic structure in language production: Evidence from syntactic priming. *Language and cognitive processes*, *21*(7-8), 974–1010.

Branigan, H., Pickering, M., Stewart, A., & McLean, J. (2000). Syntactic priming in spoken production: Linguistic and temporal interference. *Memory and Cognition*, *28*(8), 1297–1302.

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *22*, 482-1493.

Gries, S. (2005). Syntactic Priming: A Corpus-based Approach. *Psycholinguistic Research*, *34*(4), 365–399.

Hadelich, K., Branigan, H., Pickering, M., & Crocker, M. (2004). Alignment in dialogue: Effects of visual versus verbal-feedback. In *Proceedings of the 8th workshop on the Semantics and Pragmatics of Dialogue.*

Hartsuiker, R., Pickering, M., & Veltkamp, E. (2004). Is syntax separate or shared between languages? Cross-linguistic syntactic priming in Spanish-English bilinguals. *Psychological Science*, *15*, 409–414.

Healey, P. G. T., Howes, C., & Purver, M. (2010). Does structural priming occur in ordinary conversation? In *Proceedings of Linguistic Evidence 2010.* Tübingen.

Healey, P. G. T., Purver, M., & Howes, C. (2010). Structural divergence in dialogue. In *Proceedings of 20th annual meeting of the Society for Text & Discourse.*

Nelson, G., Wallis, S., & Aarts, B. (2002). *Exploring natural language: Working with the British component of the International Corpus of English.* Amsterdam: John Benjamins.

Pickering, M., & Ferreira, V. (2008). Structural priming: A critical review. *Psychological Bulletin*, *134*(3), 427–459.

Pickering, M., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, *27*, 169–226.

Pickering, M., & Garrod, S. (2006). Alignment as the basis for successful communication. *Research on Language and Computation*, *4*, 203–228.

Reitter, D., Moore, J., & Keller, F. (2006). Priming of syntactic rules in task-oriented dialogue and spontaneous conversation. In *Proceedings of the 28th conference of the Cognitive Science Society.*

Szmrecsanyi, B. (2005). Language users as creatures of habit: A corpus-based analysis of persistence in spoken English. *Corpus Linguistics and Linguistic Theory*, *1*(1), 113–150.

Tannen, D. (2007). *Talking voices: Repetition, dialogue and imagery in conversational discourse.* Cambridge: Cambridge University Press. (Second Edition)

Investigating phonotactics, lexical analogy, and sound symbolism using xenolinguistics: A novel word-picture matching paradigm

Vsevolod Kapatsinski (vkapatsi@uoregon.edu)
Department of Linguistics, 1290 University of Oregon,
Eugene, OR 97401 USA

Lamia H. Johnston (lhj@uoregon.edu)
Department of Linguistics, 1290 University of Oregon,
Eugene, OR 97401 USA

Abstract

All human languages have restrictions on sound sequences, called *phonotactic constraints*. Knowledge of phonotactic constraints is typically tested using pseudoword rating tasks, e.g., an English speaker might be asked to rate acceptability or wordlikeness of the phonotactically illegal /bnɪk/ and the phonotactically legal /blɪk/. We introduce a new method of testing knowledge of phonotactic constraints. Instead of asking subjects to rate pseudowords, we ask them to assign pseudowords to pictures of novel objects. The set of available pseudowords is larger than the set of pictures and includes both legal and illegal pseudowords. We find legal pseudowords to be less likely to be left unassigned to pictures than illegal pseudowords. Thus, the listeners show knowledge of the phonotactics of English. We suggest that the present method has important advantages over rating tasks: it is a more direct measurement of the influence of phonotactics on the lexicon, and it allows the experimenter to detect influences of sound symbolism and lexical analogy and separate them from the influence of phonotactics.

Keywords: phonology; phonotactics; sound symbolism; analogy; acceptability

Introduction

The grammars of all languages contain restrictions on possible sound sequences, called *phonotactic constraints*. For instance, despite /bnɪk/ and /blɪk/ not being actual English words, /blɪk/ obeys the phonotactic constraints of English but /bnɪk/ does not because there are no word-initial stop+nasal sequences in English. Native English speakers would also rate /bnɪk/ as being less acceptable than /blɪk/, showing that they have knowledge of the phonotactic constraints of their language (Chomsky & Halle 1965).

The phonotactic constraints are thought to place restrictions on the way the lexicon of the language can develop in the future, such that newly coined or adopted words are likely to also obey the phonotactics of the language. If a word does not obey the phonotactics of a language into which it is borrowed, it often changes to fit the phonotactics. One way this change can happen is through misperception (Ohala 1981). Berent et al. (2007), Dupoux et al. (1999), and Pitt (1998) have documented that phonotactically illegal sequences are often perceived as similar legal sequences, e.g., English listeners often perceive natural productions of /bnɪk/ by speakers of Russian, for whom the /bn/ cluster is phonotactically legal, as having a vowel between /b/ and /n/. Thus a word like /bnɪk/ is likely to be misperceived by English speakers as /bənɪk/ and borrowed into English as /bənɪk/.

An additional, and much more controversial, way in which phonotactic constraints can influence the development of a language is by militating against the adoption or retention of phonotactically illegal words. Thus, phonotactically illegal words may be less likely to be borrowed and retained in the language than phonotactically legal words. An intriguing piece of evidence for this influence of phonotactics is provided by Berg (1998:230-233) who examines the probability of Old English words surviving into Modern English depending on the phonotactics of the initial cluster in Modern English. He finds that 803/968 (83%) words containing a phonotactically legal cluster (/kr/, or /sn/) have survived, compared to 555/774 (72%) for words containing now illegal clusters (/kn/, /gn/, and /wr/, $\chi^2(1)=31.1$, $p<.001$). He argues that "a word may pass out of the system because of phonological problems" (Berg 1998:231), suggesting that phonotactic constraints may not only force illegal words to change but also force illegal words out. A plausible mechanism for this effect is suggested by Martin (2007), who provides simulation data from neural networks showing that, as long as sublexical-to-lexical feedback is assumed, words that are phonotactically suboptimal are less likely to be selected for production than more well-formed competitors.

Knowledge of phonotactics is typically tested using rating tasks (for recent representative examples, see Bailey & Hahn 2001, Coleman & Pierrehumbert 1997, Frisch et al. 2000, in press, Shademan 2005, Treiman et al. 2000). involving a metalinguistic judgment of 'acceptability', 'grammaticality', 'goodness', 'wordlikeness' etc. However, judgment tasks offer at best an indirect way to gauge the hypothesized effect of phonotactics on lexical selection. One goal of the present paper is to develop a more direct method for examining the potential influence of knowledge

of phonotactics on lexical choice experimentally (Berg 1998, Martin 2007).

Phonotactic constraints are not the only influence on lexical selection. Two other potential factors are sound symbolism (e.g., Sapir 1929, Ultan 1978 vs. Diffloth 1994) and lexical analogy (e.g., Bailey & Hahn 2001, Shademan 2005). A word containing a consonant cluster that is never observed in English may nonetheless be selected (and receive high ratings in a judgment task) if it is sufficiently phonologically similar to an existing English word. In addition, words that contain sounds that iconically represent some aspects of their referents may be especially likely to enter the lexicon. In the present study,we focus on size symbolism, where high vowels like [i] symbolize small creatures while low vowels like [a] symbolize large ones (Sapir 1929, Ultan 1978).

Methods

40 native English speakers were recruited from the Psychology/Linguistics human subjects pool and participated for course credit. All reported being native English speakers. Each subject was presented with a Microsoft Powerpoint file containing instruction slides followed by experimental slides.

The instructions asked the subject to imagine oneself in the distant future, arriving on an unknown planet (called Terra Enigmatica) and discovering the remains of an Earth colony that was established by speakers of both English and Wilkipaengo (the language name was invented, so as to avoid the influence of knowledge regarding non-English phonotactics). The rest of the story, shown in (3), explained the importance of matching names to creatures and stressed that the lists 'inadvertently' included non-English names that should not be assigned to creatures.

(3) It appears that the colony was established by speakers of both English and Wilkipaengo. Before disappearing, the colonists recorded an archive of messages.

Listening to the English, you notice some unfamiliar words. The words appear to be names for creatures common to Terra Enigmatica. According to the recordings, some creatures are benign while others are extremely dangerous and may be responsible for wiping out the entire colony!

Now you need to match the creatures you've encountered to the names given to them by the English-speaking colonists.

You are not interested in the Wilkipaengo names that seem to have somehow crept into your lists.

The backstory was designed to avoid the speakers treating the nonsense words as loanwords from another language, since languages often have more tolerance of phonotactic violations in borrowings than in the native vocabulary (e.g., McCauley 1968, Pierrehumbert 2006, Schutze 2005). We also wanted to avoid asking speakers to 'name' the creatures believing that such an instruction would unleash the subjects' creativity and perhaps lead them to choose the strangest-sounding words to match the strangeness of the novel creatures (although see Martin 2007 for corpus data showing that even names of characters of role-playing games produced (largely) by English speakers tend to obey the phonotactics of English). Thus, the backstory is designed to suggest to the speakers that the words to be assigned to creatures should be ordinary English words that speakers of English would be using in speech. In Schutze's (2005) terms, we are after the "dictionary scenario" where the word is assumed to be unknown to the subject but to be a regular English word that could be found in a big enough dictionary of the right variety of the language. An important goal for future work is to determine the extent to which subjects' behavior in the task is influenced by instructions.

The experimental slides, which followed the instruction slides, are exemplified by Figure 1.



Figure 1: An experimental slide containing draggable and playable sound files and creature animations.

When a subject came to an experimental slide, s/he clicked on 'Play animations', which played all creature animations simultaneously. The animations were made using Electronic Arts' Spore™ and featured movement and animal sounds. After playing the animations, the subject would double click on the sound files of pseudowords on the left and drag the desired sound files onto the creatures they name using the computer mouse. This procedure avoids presenting subjects with orthography (see Clopper & Pisoni 2007 for a related free classification paradigm for acoustic stimuli). The subjects could listen to the sound files as much as they wanted to and could also replay creature animations if desired. They were instructed to make sure that they listened to all the words on a slide before proceeding to the next one.

There were six experimental slides, each containing six animated creatures and twelve sound files of pseudowords. Six of the pseudowords on each slide began with a consonant cluster that is phonotactically illegal in English

while six began with either a single consonant or a legal consonant cluster. Consonant cluster legality was fully crossed with vowel identity such that half of the words with legal clusters contained one vowel, and half another vowel. The vowels contained in words differed across slides, with two slides featuring [i] and [a], two featuring [u] and [æ], and two featuring [oʊ] and [eɪ]. The order in which vowel pairs were presented was counterbalanced: half the subjects were exposed to each of the slide sequences in (4).

(4) i/a → u/æ → eɪ/oʊ → u/æ → i/a → eɪ/oʊ
eɪ/oʊ → u/æ → i/a → eɪ/oʊ → u/æ → i/a

There were two matched sets of pseudowords such that for each phonotactically illegal pseudoword there was a legal pseudoword that differed from the illegal counterpart only in the onset. All pseudowords had a (C)CVC structure. The legal and illegal counterparts were never presented to the same subject. Rather, they appeared in the same positions on the same slides but for different subjects. This was done to avoid presenting minimal pairs differing only in the (legality of) the onset and thus perhaps drawing abnormal degree of attention to phonotactics. Half of the subjects assigned to each vowel sequence order received each pseudoword set. The mappings between legal and illegal clusters are shown in (5) with numbers of word pairs exemplifying a mapping in parentheses.

(5) bd/bl (9), bn/bl (3), bn/br (1), bw/kw (1), bz/sp (3), bz/sk (1), bz/bl (1), dg/dw (3), dg/dr (5), fn/fl (3), fn/fr (1), gd/gl (3), gd/gr (3), kp/kw (6), ks/sk (3), lb/bl (4), lb/w (3), sr/fl (1), nd/dr (1), nd/pl (1), pn/pl (2), pt/pr (2), pw/pl (2), sr/sw (3), sr/tr (3), tk/tw (2), tn/tw (2)

Results and Discussion

The effect of phonotactic legality is shown in Figure 2. Phonotactically legal words were significantly more likely to be assigned to creatures than the corresponding phonotactically illegal words (by items, $t(71)=8.05$, $p<1/10^{11}$; by subjects, $t(39) = 5.57$, $p<1/10^{5}$).

The legal/illegal pairs in which the illegal pseudoword was (unexpecdtedly) used less often than the legal one are drVC/dgVC (n=4), dwæʃ/dgæʃ, fneɪk/freɪk, kwum/kpum, and twis/tnis.

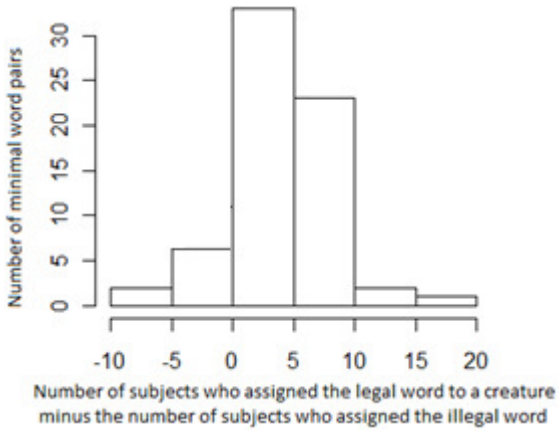


Figure 2: The effect of phonotactic legality on a word's frequency of being assigned to any creature (maximum possible difference = 20; pairs with no difference in popularity between legal and illegal words (n=4) not shown).

It is important to distinguish between underuse of legal clusters, which could then be argued to have been perceived as illegal by the subjects, and overuse of illegal clusters. Figures 3 and 4 show that in the present case we are dealing primarily with underuse of the legal clusters [dr] and [Cw] (all words beginning with these clusters are shown as darkened blocks in Figure 3) rather than overuse of the corresponding illegal words.

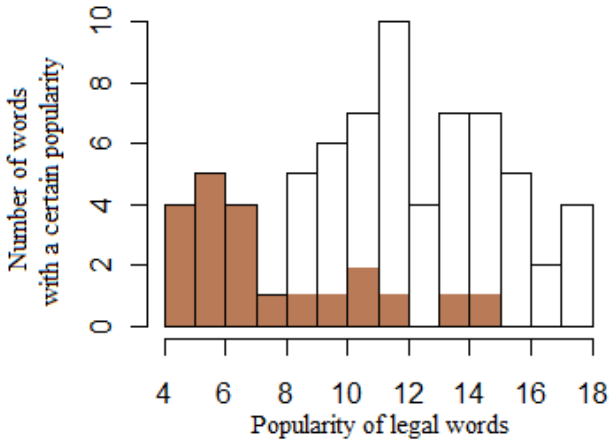


Figure 3: The distribution of popularities of legal words with the legal words beginning with /dr/ or /Cw/ shown darkened

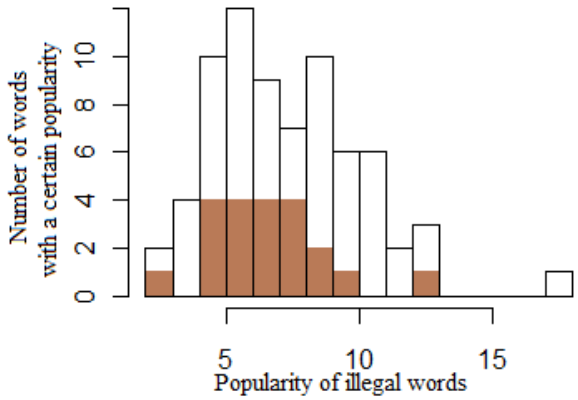


Figure 4: The distribution of popularities of illegal words. Darkened blocks represent illegal words that are minimal pairs for the legal words in Figure 3 (differing in onset cluster)

The underuse of /dr/ onsets may be due to the speaker's strong affrication of /d/ in these clusters, possibly resulting in the cluster being perceived as the phonotactically illegal cluster /dʒr/ by listeners who produce less affrication of /d/ in /dr/ (cf. Ohala 1981). The lack of preference for Cw over illegal clusters may be due to the legal clusters having a very low type frequency in English, which makes these clusters, though legal, marginal (for effects of type frequency on acceptability ratings, see Bailey & Hahn 2001, Coleman & Pierrehumbert 1997, Frisch et al. 2000, in press, Treiman et al. 2000).

Finally, the strong preference for /fneɪk/ over /freɪk/ (the former is used by 8 more subjects than the latter and is the most popular illegal word in the present study: the clear outlier in Figure 4) is likely to be an effect of lexical analogy to the word 'snake'. To assess possible effects of lexical analogy and sound symbolism, we tested whether some words might be preferentially paired with certain creatures by cross-tabulating sound files and the creatures they are paired with and looking for cells with values that are significantly higher than expected under the null hypothesis. We tested three different null hypotheses: 1) subjects are randomly pairing words with creatures within a slide (which produces a 1/12 change of assigning a word to a creature), 2) subjects randomly pair phonotactically legal words with creatures within a slide, and 3) for each slide, subjects choose a set of words to assign to creatures, and then randomly match the words within the set with creatures on the slide. With any of the three null hypotheses, there were three words that were paired with particular pictures more often than would be expected if the null hypothesis were true. The words were /fneɪk/, /blun/, and /blut/ (assigned to their preferred creatures 43%, 42%, and 37% of the time they were assigned to *any* creatures; p=.0005, p=.0003, p=.0006 respectively according to the binomial test with null hypothesis 3; the Bonferroni-adjusted critical p value is .05/72=.0007). The preferred creature-word pairings are shown in Figure 5. The likely explanation for these preferred assignments is lexical analogy to the words 'snake' [sneɪk], 'bloom' [blum], and 'blue' [blu] respectively: the creatures in question are the only snake-like, bloom-like, and blue creatures on their slides.

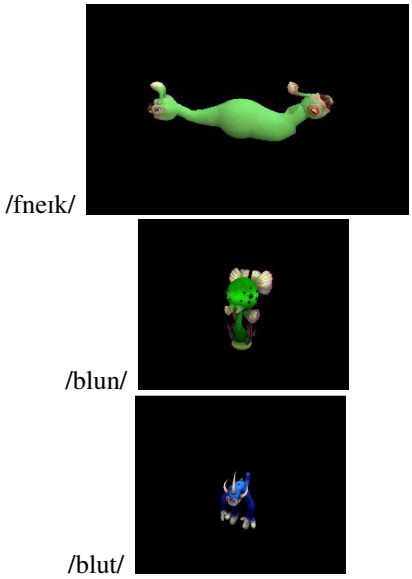


Figure 5: Non-random word-creature pairings.

Schutze (2005) objects that the "dictionary scenario" (exemplified by our backstory) is inappropriate for use in nonce probe tests of grammatical knowledge because of being particularly subject to effects of lexical analogy. The present findings confirm the presence of lexical analogy effects in the scenario. However, we do not believe this invalidates the use of the "dictionary scenario" in the present paradigm even if one believes in grammar as a cognitive module that is separate from the lexicon (Schutze 2005). Unlike in rating tasks, lexical analogy effects can be detected (and factored out) in the present task by searching for non-random picture-word co-occurrences. In order to examine possible differences between rating tasks and word-picture matching, we have conducted a wordlikeness rating task where "1" meant "not at all like English words" and "5" meant "very much like typical English words". The same pseudowords were used but no pictures were presented. We observed that [fneɪk] received the highest ratings out of all phonotactically illegal pseudowords. Given the results of the picture-matching task, we would argue that this result is due to lexical analogy to the word /sneɪk/. We would not have been able to infer this based on the rating data alone, leaving the effect unexplained.

The use of pictures in the present experiment may discourage the use of phonological analogy to existing words that are phonologically similar to the experimental pseudowords but not semantically similar to any of the pictures of the slide, e.g., the pseudoword /glog/ could be

rated highly wordlike on analogy with /grog/ or /log/ but the existence of /grog/ and /log/ might not lead the subjects to assign /glog/ to a creature because /grog/ and /log/ are not names for animals (or features of animals). This hypothesis remains to be tested.

Both rating tasks and the present paradigm are limited by the fact that phonotactically illegal sound sequences are often misperceived as phonetically similar legal sequences (Berent et al. 2007, Dupoux et al. 1999, Pitt 1998). Furthermore, as Berent et al. (2007) show, phonotactically illegal sequences are not equal in how likely they are to be misperceived. In particular, typologically marked onsets with falling sonority like /lg/ are more likely to be misperceived by English speakers than onsets with flat sonority like /bd/, which are less likely to be misperceived than clusters with rising sonority like [bn] or [pw]. While we might have expected that English listeners would judge words beginning with /lg/ to be particularly unnatural and would be unlikely to assign them to objects, the finding that such clusters are most likely to be misperceived as legal sound sequences (e.g., /ləg/) throws a wrench into this expectation. Thus, it is a priori unclear whether illegal clusters strongly violating sonority sequencing should be assigned to creatures more often or less often than illegal clusters that do not violate sonority sequencing (as much).

The breakdown of onsets by sonority is shown in Figure 6. Assuming that [s] is extrasyllabic, the optimality of the sonority sequence in the onset rises from left to right.

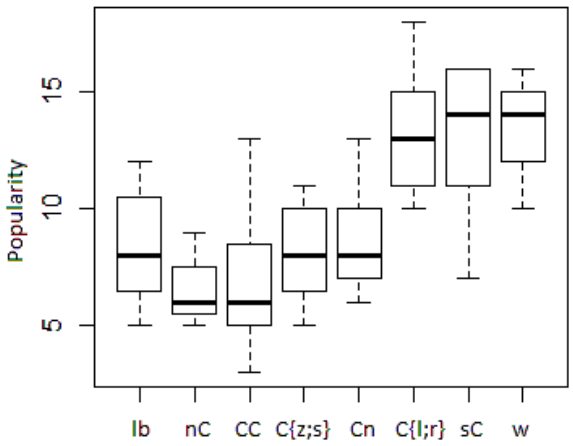


Figure 6: Frequency of being assigned to any creature as a function of sonority (C="obstruent"). This figure does not include /dr/ clusters.

There is a statistically significant difference between C{z;s;n} and C{l;r} (W=349, p=.00001). However, there is only a trend for {n;C}C clusters to be used less than C{z;s;n} clusters (W=137, p=.034, which would not reach $p_{critical}$ with the Bonferroni correction), and /lb/ clusters are assigned to creatures numerically more often than clusters that should be more acceptable according to sonority sequencing. The effect of sonority on acceptability of illegal clusters is thus ambiguous and requires perception data for interpretation.

In future work, it appears important to supplement data from picture-word matching with data on how the stimuli are perceived by the same subjects. We expect that subjects who often misperceive an illegal cluster as a related legal sequence should be more likely to assign words containing the cluster to pictures of novel objects. Nonetheless, the presence of the effect of phonotactic legality in the present data as well as in rating studies of phonotactics (Bailey & Hahn 2001, Coleman & Pierrehumbert 1997, Frisch et al. 2000, in press, Treiman et al. 2000) shows that the perceptual mechanism of repairing phonotactically illegal sequences does not succeed in repairing the sequence 100% of the time, leaving room for speakers to choose between borrowing or retaining phonotactically legal and illegal pseudowords, thus repairing phonotactic violations on the lexical level (Berg 1998). The imperfection of perceptual repair is what allows rating studies as well as the present method to assess knowledge of phonotactics.

Following the completion of all experimental slides, we asked subjects to review all creature animations and rate the creatures' size and cuteness. Subjective and objective (height, width, area, thickness) measures of the size of a creature, the height of F0 in the creatures' vocalizations, and ratings of creature cuteness did not correlate with the presence or absence of any segments or segment features in the words subjects assigned to the creature (all p>.1). Thus, size sound symbolism did not seem to play an important role in this experiment. We hypothesized that this may be due to the presence of many dimensions other than size in the visual stimuli. Figure 7 presents the results of an ongoing follow-up study. Thus far 7 subjects have been asked to name ten (5 big, 5 small) monochromatic 2-dimensional creature pictures using 20 words (half phonotactically illegal, half containing [i] or [u], half containing [a] or [au]). As Figure 7 shows, words with high vowels tended to be assigned to small creatures while words with low vowels tended to be assigned to large creatures ($\chi^2(1)=8.21$, p=.004). Thus, size sound symbolism effects may be observed in the present task when size is a salient dimension of variation for the presented objects.

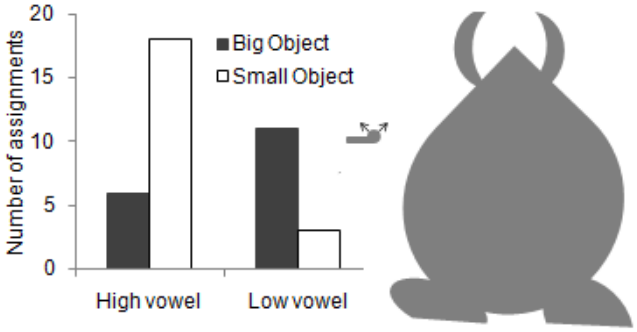


Figure 7: An effect of size sound symbolism with simpler creatures.

Conclusion

Asking subjects to match a set of pseudowords with a smaller set of novel objects provides a new way to assess the subjects' knowledge of phonotactics. This method provides important advantages over the traditional method of assessing knowledge of phonotactics (acceptability or wordlikeness ratings). First, the proposed method is a much more direct way of assessing the influence of phonotactics on lexical selection (found to operate in historical data by Berg 1998 and Martin 2007). Second, the method facilitates separating out and investigating the effects of lexical analogy and may restrict the occurrence of lexical analogy to words that are semantically related to the pictures, although analogies based on such words may be more likely in the present task than in rating. The method may also be profitably used to examine the effect of sound symbolism and how it competes with phonotactics.

This task does share some shortcomings with rating tasks. First, it requires somewhat accurate perception of illegal clusters. Given the evidence that phonotactically illegal sequences are often misperceived as similar legal sequences (e.g., Berent et al. 2007, Dupoux et al. 1999, Pitt 1998), the present task should ideally be followed by an assessment of the same subjects' perception of the stimuli. This might be accomplished using discrimination, transcription or identification tasks, or testing for the presence/absence of identity priming between the similar-sounding legal and illegal sound sequences (Berent et al. 2007, Dupoux et al. 1999, Pitt 1998). Second, the present instructions still require subjects to explicitly judge whether or not the presented words could be words of English. Future work should investigate the importance of this instruction.

Finally, the principal disadvantage of the present task compared to rating is that subjects perform the task much more slowly than a comparable rating task (the subjects in the word-picture matching version of the present task took on average 15 minutes to go through the 72 words, while a rating task using the same words took only 3 minutes). A possible way to reduce the time demands is to present fewer words and pictures per slide, thus simplifying the decision. The principal potential disadvantage of such a move is a reduction in the possibilities for detecting effects of lexical analogy due to an even more restricted set of referents to be assigned to the words.

References

Bailey, T., & Hahn, U. (2001). Determinants of wordlikeness: Phonotactics or lexical neighborhoods? *Journal of Memory & Language*, *44*, 568–91.

Berent, I., Steriade, D., Lennertz, T., & Vaknin V. (2007). What we know about what we have never heard: Evidence from perceptual illusions. *Cognition*, *104*, 591–630.

Berg, T. (1998). *Linguistic structure and change: An explanation from language processing*. Oxford: Oxford University Press.

Chomsky, N.., & Halle, M. (1965). Some controversial questions in phonological theory. *Journal of Linguistics*, *1*, 97–138.

Clopper, C., & Pisoni, D. B. (2007). Free classification of regional dialects of American English. *Journal of Phonetics, 35*, 421-38.

Coleman, J. S., & Pierrehumbert, J. (1997). Stochastic phonological grammars and acceptability. In *Third Meeting of the ACL Special Interest Group in Computational Phonology* (pp. 49–56). Somerset, NJ: Association for Computational Linguistics.

Diffloth, G. (1994). i: big, a: small. In L. Hinton, J. Nichols, & J. J. Ohala, eds. *Sound symbolism*. Cambridge: Cambridge University Press.

Dupoux, E., Kakehi, K., Hirose, Y., Pallier, C., & Mehler, J. (1999). Epenthetic vowels in Japanese: A perceptual illusion? *Journal of Experimental Psychology: Human Perception & Performance, 25*, 1568-78.

Frisch, S. A., Brea-Spahn, M. R., & Orellana, C. I. (In press). Metalinguistic judgments of phonotactics by bilinguals. *Laboratory Phonology*, *11*.

Frisch, S. A., Large, N. R., & Pisoni, D. B. (2000). Perception of wordlikeness: Effects of segment probability and length on the processing of nonwords. *Journal of Memory & Language*, *42*, 481-96.

Martin, A. (2007). *The evolving lexicon*. PhD Dissertation, UCLA.

McCauley, J. (1968). *The Phonological Component of a Grammar of Japanese*. The Hague: Mouton.

Ohala, J. J. (1981). The listener as a source of sound change. *Chicago Linguistic Society*, *17-2*, 178-203.

Pierrehumbert, J. (2006). The statistical basis of an unnatural alternation. *Laboratory Phonology*, *8*, 81-107.

Pitt, M. A. (1998). Phonological processes and the perception of phonotactically illegal consonant clusters. *Perception & Psychophysics*, *60*, 941–51.

Sapir, E. (1929). A study in experimental symbolism. *Journal of Experimental Psychology, 12,* 225-39.

Schutze, C. T. (2005). Thinking about what we are asking speakers to do. In S. Kepser & M. Reis, eds. *Linguistic evidence: Empirical, theoretical, and computational perspectives*. Berlin: Mouton de Gruyter.

Shademan, S. (2005). Is phonotactic knowledge grammatical knowledge? *Proceedings of the West Coast Conference on Formal Linguistics, 25*, 371-9.

Treiman, R., Kessler, B., Knewasser, S., Tincoff, R. & Bowman, M. (2000). English speakers' sensitivity to phonotactic patterns. *Laboratory Phonology*, *5*, 269-82.

Ultan, R. (1978). Size sound symbolism. In J. H. Greenberg, C. R. Ferguson, & E. A. Moravcsik, eds. *Universals of human language, vol. 2*. Stanford: Stanford University Press.

Computer-based Learning of Neuroanatomy: A Longitudinal Study of Learning, Transfer, and Retention

Julia H. Chariker (julia.chariker@louisville.edu)
Farah Naaz (farah.naaz@louisville.edu)
John R. Pani (jrpani@louisville.edu)
Department of Psychological and Brain Sciences, University of Louisville
Louisville, KY 40292 USA

Abstract

Using interactive computer-based methods of instruction, this research examined the contribution of whole (3D) anatomical knowledge to learning sectional anatomy. Participants either learned sectional anatomy alone or learned whole anatomy prior to learning sectional anatomy. Sectional anatomy was explored either with perceptually continuous navigation or discretely, as in the use of an anatomical atlas. Learning occurred over repeated cycles of study, test, and feedback, and continued to a high performance criterion. After learning, transfer of knowledge to interpreting biomedical images and long-term retention were tested. Whole anatomy was learned quickly and transferred well to the learning of sectional anatomy: initial accuracy was higher, learning of sectional anatomy was completed more rapidly, and there was less error over the entire course of learning. Knowledge of whole anatomy benefited the long-term retention of sectional anatomy at 2-3 weeks. Learners demonstrated high levels of transfer to the interpretation of biomedical images.

Keywords: learning; transfer; computer; anatomy.

Introduction

In medicine and many areas of science, anatomy education serves as a vital foundation for high level knowledge and skill. Unfortunately, anatomy is challenging to learn. Large volumes of material must be learned in relatively short periods of time. Anatomical structures often have irregular and indistinct shapes. They have little variation in color and texture, and they are related to each other in complex three-dimensional arrangements. Moreover, a comprehensive education in anatomy extends to include a thorough knowledge of sectional anatomy, which is necessary for diagnostic imaging, microscopy, and dissection.

Sectional anatomy is particularly challenging to learn. A spatial transformation occurs when a two-dimensional section is taken from a three-dimensional object. The two and three-dimensional structures may look very different from each other. In addition, multiple mappings are possible between these representations of anatomy. One-to-many mappings occur because anatomy can be sectioned at different depths and orientations, resulting in significant variation in the presentation of structures across a series of sections. Many-to-one mappings occur because differently shaped structures can appear similar in a sectional image.

The challenges in learning sectional anatomy might be reduced by facilitating cognitive organization of the mass of information in the sections (consider Bower, Clark, Lesgold, & Winzenz, 1969). Given that anatomical sections are derived from whole anatomy, helping students develop a thorough understanding of the shapes and relationships of whole structures prior to learning sectional anatomy would seem an ideal way to help students organize the information in the sections. The benefit of organization for learning and memory has been established for verbal materials, but it is not clear what effect organization has in domains where spatial reasoning is required.

Knowledge of whole anatomy may also serve as a mental model that supports reasoning about sectional anatomy. Reasoning has been found to play a large role in the successful interpretation of histological sections viewed under the microscope (e.g., Pani, Chariker, & Fell, 2005).

A second approach to helping students organize information in sectional anatomy may be in the presentation of sectional anatomy itself. Serial presentation of the sections would be expected at a minimum, but additional support may be found by providing smooth, seamless navigation through the sections. Work in *anorthoscopic perception* and *kinetic completion* suggests that with this approach, learners may see the series of sections as a unified whole. On the other hand, continuous presentation of sectional anatomy can be considered a form of animation, and there has been mixed success in using animation in instruction (e.g., Hegarty, 2005; Tversky, Morrison, & Betrancourt, 2002).

In the current study, we explored both approaches to organizing sectional anatomy. Half of the participants learned whole anatomy before learning sectional anatomy (*transfer* groups), while the other half learned only sectional anatomy (*sections alone* groups). Within each of these groups, half of the participants learned sectional anatomy using a continuous presentation, and half learned with a discrete presentation -- analogous to turning the pages of an anatomical atlas.

Participants learned neuroanatomy in interactive computer-based environments. This approach holds potential for helping learners build rich mental representations of anatomy. For example, a computer-based model of 3D anatomy can be rotated to allow exploration of anatomy from any angle. It can be virtually dissected, restored to its original state, and then dissected again.

The instructional programs were designed to promote efficient learning through a method that we call *adaptive exploration*. With graphical models and exploratory tools

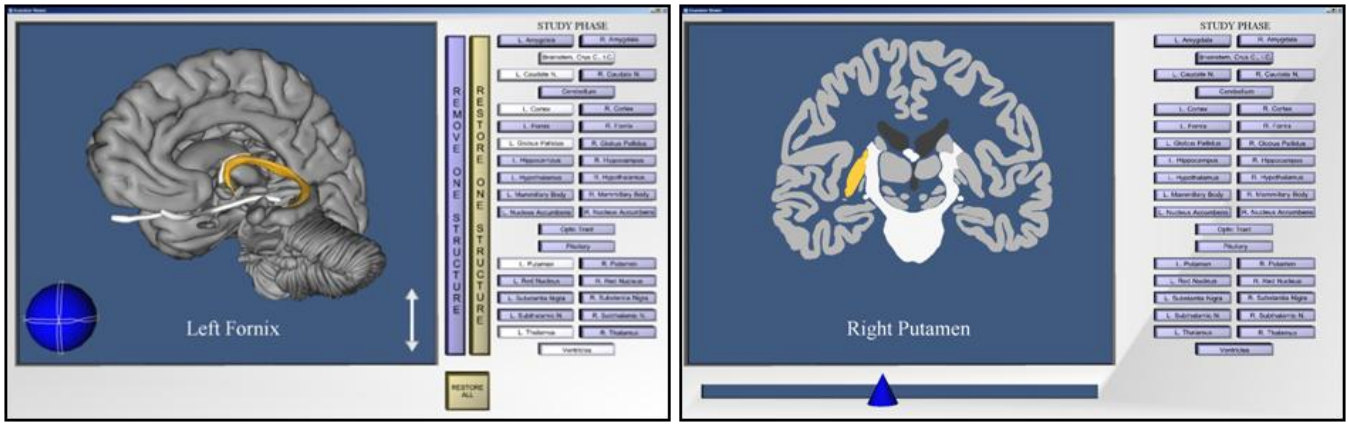


Figure 1: Screenshots of the anatomical model and the interface in the study phase of the whole anatomy learning program (left) and the sectional anatomy learning program (right).

available, learning was measured over multiple trials of study, test, and feedback until a high performance criterion was reached. In testing and feedback, participants learned the nature of the test to be mastered and were continually updated on progress in learning. This information allowed learners to adaptively adjust exploration of anatomy during study. Additionally, this approach to learning conforms to what appears to be best practices in regard to optimizing long-term retention through repeated testing (e.g., Karpicke & Roediger, 2008).

All participants learned 19 neuroanatomical structures across three standard views of anatomy: coronal, sagittal, and axial. After learning was completed, we measured the degree to which participants could transfer anatomical knowledge to interpreting biomedical images. Retention of anatomical knowledge was measured 2-3 weeks after learning was completed.

Method

Participants

Seventy-two undergraduate students at the University of Louisville were recruited for the study through advertisements placed around campus. All were at least 18 years of age. Only those respondents who reported minimal knowledge of neuroanatomy were enrolled. Participants were paid $8.00 per hour for their participation.

Each participant was administered the Space Relations subtest of the Differential Aptitude Tests, a test of spatial ability, prior to beginning the study (DAT-SR; Bennett, Seashore, & Wesman, 1989). The mean and the distribution of scores were balanced across the four learning groups.

Materials

A three-dimensional (3D) computer graphical model of the human brain was created for this research (see Figure 1). Digital images of neuroanatomical cryosections in the Visible Human project (Vers. 2.0) of the National Library of Medicine were used as source material for the model (Ratiu, Hillen, Glaser, & Jenkins, 2003). The brain model is composed of 19 structures, including the cerebral cortex, ventricles, cerebellum, brainstem, amygdala, caudate nucleus, fornix, globus pallidus, hypothalamus, hippocampus, mammillary bodies, nucleus accumbens, optic tract, pituitary, putamen, red nucleus, substantia nigra, subthalamic nucleus, and thalamus. The structures were colored in dark gray (ventricles), medium gray, and white to approximate the basic appearance of light and dark structures in typical biomedical images of the brain.

Three relatively dense sets of serial sections were created from the brain model. There were 60 coronal sections, 50 sagittal sections, and 46 axial sections. All sections were taken at equal intervals.

MRI images were used to test transfer of knowledge. The images were made available from the SPL-PNL Brain Atlas (Kikinis et al., 1996). The images are typical gray scale T1 images of structures in the head and neck. The images were slightly brightened and contrast enhanced and presented at a screen resolution of 895 x 895 pixels. Visible Human images also were used to test transfer. These images were from the Visible Human 2.0 dataset. The images were high resolution color images of structures in the head and neck.

Computer programs for learning neuroanatomy were created using the C++ programming language and the Open Inventor library for interactive graphics. There was a common format for all of the learning programs. The differences between the programs were modifications related to the type of anatomy presented and the different presentations of sectional anatomy.

In all of the learning programs, a participant completed two learning trials -- one block of trials -- before a single run of the program terminated. Participants were presented with the same form and view of anatomy (e.g., sectional anatomy, coronal view) throughout the two trials in a block.

Each learning trial was composed of three phases: study, test, and feedback. Throughout each phase, tools were available that functioned specifically for either whole or sectional anatomy. In the study phase, participants had three minutes to freely explore the brain. On selecting a structure, its name appeared on the screen. In the test phase, the

participant's task was to identify the anatomical structures in the model. Testing was self-paced. In the feedback phase, the participant saw the same orientation of the brain, and used the same tools and procedures, as in the study phase. In addition, structures were color coded to provide participants with information about their performance on the test.

In the study and feedback phases for whole anatomy learning, a rotation tool allowed participants to smoothly rotate the model 360 degrees forward and backward or right and left. A zoom tool allowed participants to move the model closer or further from view. Buttons were available that allowed participants to remove or restore structures.

In the test phase of a trial, model rotation was constrained to a total range of 90 degrees of motion -- 45 degrees in any direction from the initial viewpoint. This ensured that a participant's performance on the test was specific to the viewpoint being learned in that trial.

Two programs were created for learning sectional anatomy, one for the continuous and one for the discrete form of navigation. In the study phase of a trial, both programs presented a set of anatomical sections in serial order in a single viewing plane. There was a slider at the bottom of the screen, and the two learning programs differed in the way the slider functioned. In the continuous program, moving the slider resulted in continuous movement back and forth through the series of sections. A section of the brain was always visible, and the transition between sections comprised a type of animation. In addition, a highlighted structure remained highlighted in each section in which it appeared.

In the discrete presentation program, movement between sections was perceptually discontinuous. When participants moved the slider, the brain became invisible. The number of the corresponding section in the series appeared prominently at the bottom of the viewing area. On stopping at a numbered section, a 0.75 second delay occurred before the appropriate section of the brain appeared. When participants moved to a new section, highlighting was removed.

The test phase of a learning trial was the same in the two programs. Participants were given a series of test sections, presented one at a time. In each section, one or more structures were indicated with a red arrow, and the participant's task was to correctly label those structures. Although all 19 structures were tested in each trial, the section of a structure that was tested varied across trials.

During the feedback phase of the trial, participants used the slider to find each of the test sections in the series. A message reading "Test Section" appeared prominently on the screen when a test section was accessed by the slider. The tested structures in each test section were identified with the same red arrows that appeared in the test.

Three computer programs were created to test transfer of knowledge to the interpretation of biomedical images. In the first test, Uncued Recognition, participants were presented with a set of 9 images, one at a time, and asked to identify all of the structures they thought they recognized in each image. Participants identified structures by indicating the location of a structure with the mouse (leaving a red dot on the image) and then selecting the name of the structure from a list on the interface. The images alternated through coronal, sagittal, and axial views, in that order.

The remaining two test programs provided cues to the presence of structures in the images. In the Submit Structure test, the name of a single structure was presented at the bottom of each image, and participants selected the appropriate structure in the image. In the Submit Name test, a single structure was designated by a red arrow in each image, and participants selected its name from a list on the interface. Each test was comprised of three subtests, one for each view of anatomy.

A sectional anatomy test and a whole anatomy test were created for testing long-term retention. For participants who had only seen sectional anatomy, the test of whole anatomy was a test of transfer rather than retention. These tests were the same as tests given during learning and were created for all three views of anatomy.

Apparatus

Participants sat individually at computer workstations with large high resolution LCD screens (24 inch, 1200 x 1952 pixels). Participants were tested alone in small quiet rooms with the doors closed.

Design and Procedure

The core experimental design was a 2 X 2 between-groups factorial: anatomy course (transfer vs. sections alone) by sectional anatomy presentation (continuous vs. discrete).

Prior to beginning any of the learning or testing programs in the study, participants were trained on all aspects of the task using instructional software developed for this purpose.

During the learning portion of the study, performance in identifying 19 neuroanatomical structures was measured over multiple blocks of trials. Percent correct was calculated for each trial, and mean percent correct was calculated for each block of two trials. Participants continued learning anatomy until they reached a minimum of 89.5 percent accuracy (17 of 19 structures) in each of three consecutive learning blocks—all three views of anatomy. Across blocks of learning trials and throughout testing, the order in which view was presented was standardized at coronal, followed by sagittal, and then axial.

Immediately after learning was completed, participants were given the three tests of transfer to biomedical images in the order Uncued Recognition, Submit Structure, and Submit Name. For each test, participants were tested with each image type (MRI and Visible Human) in all three views of anatomy. The two image types were counterbalanced across participants.

Two to three weeks after learning was completed, participants were given the test of long-term retention for sectional anatomy followed by the test of long-term retention/transfer for whole anatomy. Tests were given for all three views of anatomy.

Results

Learning

Learning Trajectories Multilevel modeling was used for statistical analysis of performance in learning (Raudenbush & Bryk, 2002). Binomial models were appropriate for these data. Variables tested for inclusion in the multilevel model included learning block, anatomy course (AC), sectional anatomy presentation (SAP), and spatial ability (DAT-SR). Spatial ability was a significant factor in each of the models of learning but will not be discussed in this paper. Details of model parameters are available from the authors.

To establish the relative efficiency of learning whole anatomy and sectional anatomy, the transfer group's performance in whole anatomy was compared to the sections alone group's performance in sectional anatomy. Participants learning whole anatomy had substantially higher performance in the first block of trials and learned at a faster rate than participants learning sectional anatomy (see Figure 2). Mean percent correct identification in block one was 54 percent for whole anatomy and 36 percent for sectional anatomy, $t(69) = 5.780$, $p < .001$. Both groups improved in performance over successive blocks, $t(68) = 15.746$, $p < .001$; however, the increase in performance was much greater for participants learning whole anatomy: AC, $t(68) = 7.359$, $p < .001$.

There were no effects on the efficiency of learning sectional anatomy due to the type of sectional anatomy presentation in any of the analyses of learning. This variable was not retained in the multilevel models and will not be discussed further in the presentation of results on learning.

In a second analysis, transfer of learning from whole to sectional anatomy was measured by comparing performance in sectional anatomy for the transfer and sections alone groups. Participants in the transfer groups performed significantly better in the first block of sectional anatomy learning than participants in the sections alone groups (see Figure 2). Mean percent correct identification was 73 percent in the transfer groups and 36 percent in the sections alone groups, $t(69) = 13.522$, $p < .001$. Although both groups improved over time, the transfer groups continued their learning at a slower rate than the sections alone groups: AC, $t(70) = -3.321$, $p = .002$.

In a third analysis, differences between conditions were further explored by comparing performance in sectional anatomy for the transfer and sections alone groups after relating performance to the total time spent learning neuroanatomy. For the transfer groups, learning blocks were numbered to reflect the time participants spent learning both whole and sectional anatomy. Nearly two thirds of the participants in the Transfer groups (21 of 36) completed whole anatomy learning in 4 blocks and transferred to sectional anatomy in block 5. Therefore, performance in sectional anatomy learning was compared beginning at block 5. Modeled performance in Block 5 was 71 percent for the Transfer groups and 81 percent for the Sections alone groups, AC, $t(69) = -3.030$, $p = .004$. The 10 percent difference is equivalent to 2 of the 19 structures on the test.

Learning Time to Achieve Criterion Performance In each learning trial, time was constrained to 3 minutes for study and 3 minutes for feedback. Therefore, we considered the number of blocks required to reach the performance criterion as one measure of learning efficiency. An ANCOVA was performed to compare the number of blocks of trials required to complete learning for whole and

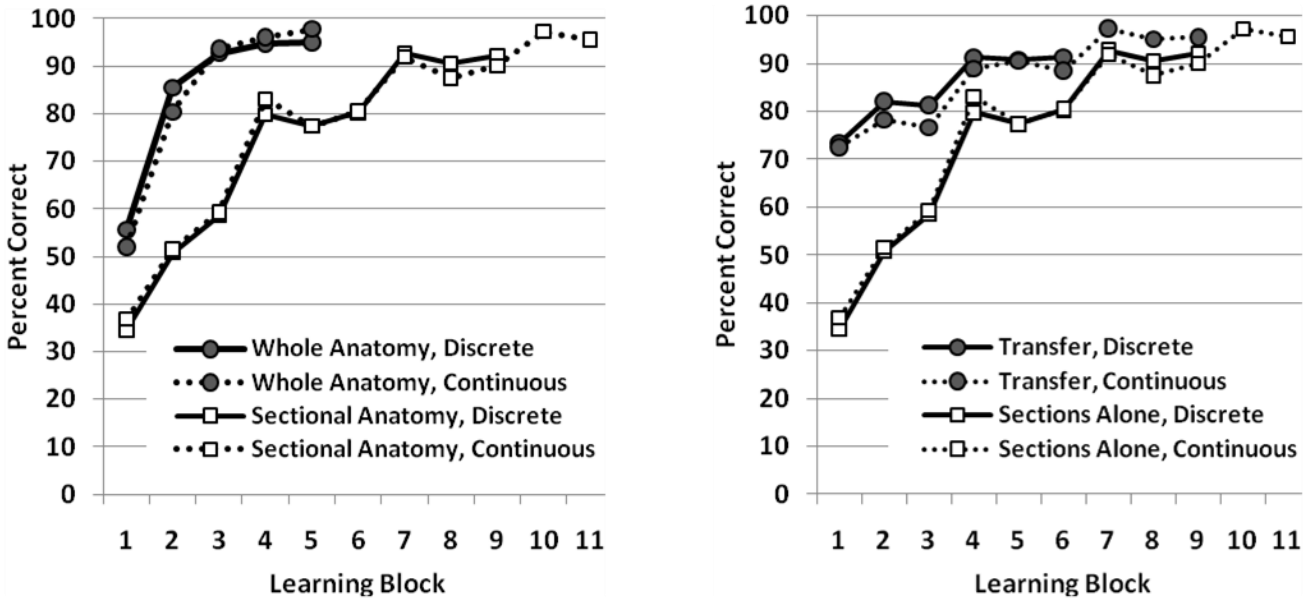


Figure 2: A comparison of performance in whole anatomy and sectional anatomy (left) and a comparison of performance in sectional anatomy beginning at block 1 (right).

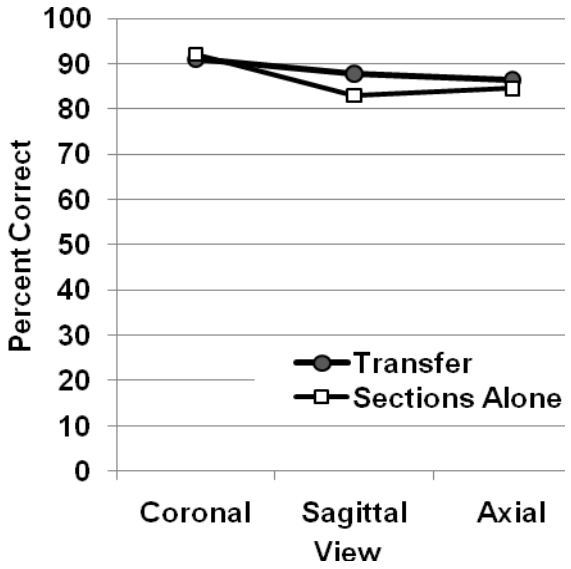


Figure 3: Sectional anatomy retention.

sectional anatomy. Spatial ability was correlated with the number of blocks required (r = -.236, p = .046) and was entered as a covariate, $F(1,67)$ = 7.785, p = .007. Participants learned whole anatomy in significantly fewer blocks (M = 5.2) than participants learned sectional anatomy (M = 10.7), $F(1, 67)$ = 57.555, p < .001.

A second ANCOVA compared the transfer and the sections alone groups on the number of trial blocks required to reach criterion in sectional anatomy learning. Again, spatial ability was correlated with the number of blocks to reach criterion (r = -.298, p = .011) and was included as a covariate, $F(1,67)$ = 7.678, p = .007. Participants in the transfer groups completed sectional anatomy in 2.5 fewer blocks (M = 8.2) than participants in the sections alone groups (M = 10.7), $F(1, 67)$ = 7.282, p = .009.

A third ANCOVA was performed to look for differences between the groups in the number of blocks of trials necessary to complete all learning in neuroanatomy. Spatial ability was correlated with the number of blocks to reach criterion (r = -.344, p = .003) and was included as a covariate, $F(1,67)$ = 10.129, p = .002. Participants in the transfer groups completed whole anatomy and sectional anatomy in 2.7 more blocks than participants in the sections alone groups completed sectional anatomy (transfer, M = 13.4; sections alone, M = 10.7), $F(1, 67)$ = 6.021, p = .017.

Total Error in Learning Neuroanatomy Over the entire course of learning, participants in the transfer groups made fewer errors (M = 77) in learning neuroanatomy than participants in the sections alone groups (M = 100), $F(1, 67)$ = 3.870, p = .053. This occurred even though the transfer groups were required to complete two presentations of anatomy and took 2.7 more blocks to do so. Spatial ability was a significant covariate in the analysis of total error, $F(1, 67)$ = 13.995, p < .001.

Testing

Long-Term Retention and Transfer MANCOVA was used to analyze retention of sectional anatomy and retention/transfer of whole anatomy. DAT-SR was included as a covariate.

Retention of sectional anatomy remained high two to three weeks after learning, with several participants reaching 100% accuracy in the first test (see Figure 3). There was an interaction of AC with view, Wilks' Lambda (Λ) = .898, $F(2, 63)$ = 3.570, p = .034. The transfer groups were more accurate than the sections alone groups for retention of the sagittal view of sectional anatomy (transfer M = 87.8, sections alone M = 83.1), $t(57)$ = -2.675, p = .03 (Bonferroni). No differences between the groups occurred for retention of the coronal and axial views.

In the analysis of retention/transfer for whole anatomy, participants in the transfer groups were more accurate than participants in the sections alone groups in identifying whole anatomy, $F(1, 64)$ = 15.306, p < .001. Participants in the transfer groups tested at 97% mean accuracy in identifying whole brain structures. Although participants in the sections alone groups had never seen whole anatomy, they reached an overall mean accuracy of 89.5%. This meets the numerical criterion used for successful learning. Given this high rate of transfer, it is important to consider that there was a relatively substantial effort required to achieve this performance. All tests in this experiment were self-paced. In an analysis of test duration, participants in the sections alone groups took substantially more time than the transfer groups to complete the three tests for whole anatomy (M = 14.5 minutes vs. 8.8 minutes, a difference of nearly 6 minutes), $F(1, 61)$ = 54.331, p < .001. This

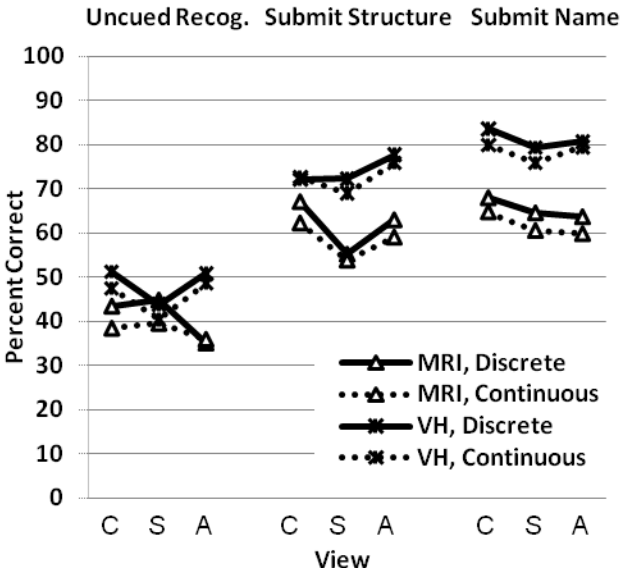


Figure 4: Transfer to MRI and Visible Human images for the discrete and the continuous sectional anatomy presentation groups.

suggests that participants who had received sectional learning alone were not recalling a representation of whole anatomy but were inferring it.

Transfer to Biomedical Images In scoring Uncued Recognition and Submit Structure, correct answers were decided ahead of time, and images were created with the structure boundaries drawn on them. During scoring, the experimenters were blind to the participants' identities and experimental conditions. MANOVA was used to analyze performance on each test.

Transfer performance was quite high, particularly for the two cued tests (Submit Structure and Submit Name; see Figure 4). Within each test, performance varied widely among individuals, with some participants performing extremely well. In Submit Structure and Submit Name, the best performing participants were above 90% accuracy.

There were no differences in transfer due to learning group. Performance was higher for Visible Human than for MRI images in all three tests: Uncued Recognition, VH M = 47%, MRI M = 40%, Λ = .440, $F(1, 65) = 82.659$, $p < .001$; Submit Structure, VH M = 72%, MRI M = 58%, Λ = .704, $F(1, 64) = 26.913$, $p < .001$; Submit Name, VH M = 80%, MRI M = 64%, Λ = .378, $F(1, 64) = 105.282$, $p < .001$.

In two of the three transfer tests, Uncued Recognition and Submit Name, there was a main effect of sectional anatomy presentation: Uncued Recognition (continuous M = 41.7, discrete M = 44.9), $F(1, 65) = 3.962$, $p = .051$; Submit Name (continuous M = 70.1, discrete M = 73.4), $F(1, 64) = 4.835$, $p = .032$. Participants who learned with a discrete presentation were more accurate in identifying structures than participants who learned with a continuous presentation.

Discussion

Knowledge of whole anatomy served as an effective basis for learning sectional anatomy. Whole anatomy was learned quickly—in half of the time of sectional anatomy. Knowledge of whole anatomy transferred well to learning sectional anatomy. Accuracy in block 1 of sectional anatomy was twice as high for the transfer groups, and learning of sectional anatomy was completed more quickly. There was less error over the entire course of learning for participants learning both representations of anatomy.

Knowledge of whole anatomy benefited long term retention of sectional anatomy. Because the participants who learned whole anatomy required *fewer* trials with sectional anatomy, this advantage for retention is inconsistent with the well-known test effect. In the test effect, a *greater* number of tests of knowledge during learning leads to an advantage for long-term retention. However, tests administered during learning and at retention are identical to each other. For the present research, such a test effect would show better long-term retention for the *sections alone* groups.

On the other hand, the groups that learned both whole and sectional anatomy did require more total trials to learn. Thus, the improvement in long-term retention is potentially due to a type of test effect, one that we have not seen described elsewhere. In this case, additional testing of whole anatomy is contributing to the long-term retention of sectional anatomy, a further instance of transfer of learning.

The transfer of knowledge to the interpretation of biomedical images served as a gold-standard test of the present methods of computer-based learning of neuroanatomy. The high levels of transfer obtained, along with the high levels of long-term retention, strongly encourage the use of these methods in neuroanatomy instruction.

Acknowledgments

This research was supported by a grant from the National Library of Medicine, National Institutes of Health to J. Pani (Grant 1 R01 LM008323; Histological Reasoning: Visual Cognition in Microanatomy).

References

Bennett, G.K., Seashore, H. G., & Wesman, A. G. (1989). *Differential Aptitude Tests for Personnel and Career Assessment: Space Relations.* San Antonio, TX: The Psychological Corporation, Harcourt Brace Jovanovich.

Bower, G. H., Clark, M. C., Lesgold, A. M., & Winzenz, D. (1969). Hierarchical retrieval schemes in recall of categorized word lists. *Journal of Verbal Learning and Verbal Behavior, 8*, 323-343.

Hegarty, M. (2005). Multimedia learning about physical systems. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning*. New York, NY: Cambridge University Press.

Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science*, *319*, 966-968.

Kikinis, R., Shenton, M. E., Iosifescu, D. V., McCarley, R. W., Saiviroonporn, P., Hokama, H. H., … Jolesz, F. A. (1996). A digital brain atlas for surgical planning, model driven segmentation and teaching. *IEEE Transactions on Visualization and Computer Graphics, 2*, 232-241.

Pani, J. R., Chariker, J. H., & Fell, R. D. (2005). Visual cognition in microscopy. *Cogsci 2005: Proceedings of the XXVII Annual Conference of the Cognitive Science Society, 27*, 1702-1707.

Ratiu, P., Hillen, B., Glaser, J., & Jenkins, D. P. (2003). Visible Human 2.0: The next generation. In J.D.Westwood, H. M. Hoffman, G. T. Mogel, R. Phillips, R. A. Robb, & D. Stredney (Eds.), *Medicine Meets Virtual Reality 11 -- NextMed: Health Horizen*. Amsterdam, The Netherlands: IOS Press.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods.* Thousand Oaks, CA: Sage Publications.

Tversky, B. Morrison, J. B., Betrancourt, M. (2002). Animation: Can it facilitate? *International Journal of Human Computer Studies, 47*, 247-262.

Introspection and Mindreading as Mental Simulation

Paul Bello (paul.bello@navy.mil)
Office of Naval Research, 875 N. Randolph St.
Arlington, VA 22203 USA

Marcello Guarini (mguarini@uwindsor.ca)
University of Windsor, Dept. of Philosophy, 401 Sunset.
Windsor, Ontario Canada N9B 3P4

Abstract

We present a sketch of a computational account of the relationship between certain aspects of introspection with aspects of third-person ascription of mental states (mindreading). The theory we propose is developed in large part as a reaction to what we perceive to be a lack of precision in the literature and a lack of experimental techniques to properly inform the debate on the relationship between 1st and 3rd-person ascription. We first discuss the set of phenomenology associated with self-ascriptions and other-ascriptions before briefly mentioning patterns of deficits associated with each. We sketch the very beginnings of a theory of mindreading in both the 1st and 3rd person within a computational cognitive architecture having mental simulation as one of its core operations. The theory we develop provides computationally-grounded explanations that are compatible with both clinical data and the phenomenology of 1st-person attribution.

Keywords: Mental Simulation; Cognitive Architecture; Metacognition; Mindreading; Philosophy of Mind.

Introspection and Mindreading

The ability to predict and explain behavior, both self- and other-generated, is a defining feature of human intelligence and a crucial phenomenon to be accounted for at the process-level; especially for those of us interested in computational theories of cognitive architecture. One of the major constituents of this ability takes the form of being able to ascribe mental states in service of behavior prediction and/or explanation. We will refer to mental state ascription more colloquially as "mindreading." Typically, mindreading is mentioned as being related to predicting and explaining the behavior of others, but what of our ability to report on our own mental lives? This ability is generally termed introspection, and one important scientific task will be to clarify its relationship (or lack thereof) to mindreading.

After presenting some of the generally agreed-upon phenomenological features of introspection, we briefly summarize the theoretical options for the mindreading-introspection relationship and some of their immediate entailments. Finally, we present our own account of their relationship in terms of a computational cognitive architecture capable of both 1st and 3rd-person ascription via mental simulation.

Introspection: Phenomenology

Characterizing the nature of introspection has been one of the most active areas of epistemology and the philosophy of psychology. This being the case, many distinctions have been made in the process, as definitions of what it is to introspect become ever-more specialized. While some of these distinctions have arisen from a priori philosophical analysis, the advent of novel experimental procedures and the further development of neuroscience have added a substantial amount of data on introspection that is providing constraints on what our theories of self-ascription look like.

Even with its many distinctions, there seem to be a few phenomenological features that all parties agree to be related to, if not constitutive of introspection (Schwitzgebel 2010). While there is a minority who believe that either we have no mental states like beliefs to introspect or that self-attributions are only unconscious, automatic processes of self-interpretation (Carruthers 2009); the majority of others agree that humans have a window on their mental lives. Most philosophical work in the area has been dedicated to clarifying the role, function, and features of introspection.

Following the discussion in (Schwitzgebel 2010), what mostly seems to be agreed upon is that:

1. Introspection is about the mental/internal, and thus not about the non-mental/external.
2. Introspective judgments are accompanied by a strong sense of certainty, even stronger than judgments about other forms of sense data.
3. Introspective judgments are relatively direct in the sense that they occur directly without needing to be inferred from other supporting data, supporting a distinction between detecting versus reasoning about one's mental states.
4. Introspection occurs in the "specious present," comprised of a very short time period just before and just after the introspective act.
5. While effortful and non-automatic, introspective judgments about one's own mental life seem easier to produce and less prone to subjective feelings of uncertainty than judgments about the mental lives of others.

Whatever sort of theory we intend to develop ought to at least coarsely capture these features and preferably provide

explanations for them in terms of computational mechanism.

Psychological and Clinical Data

In the case of mindreading, it's been long established that those on the autism spectrum have deficits associated with mindreading; especially in regard to appreciating the false beliefs of others when trying to predict or explain their behavior (Baron-Cohen 1995). The same subjects have trouble engaging in spontaneous pretence, both self-directed and with other children. Of course, a small percentage of those on the autism spectrum are high-functioning enough to pass typical tests of false belief understanding, and more advanced tests that probe second-order false belief understanding. Results as to performance of autistic subjects on introspective tasks have been somewhat mixed. Some data suggest that autistics are capable of self-report and robustly utilize self-ascriptions of beliefs, intentions, desires and the like (Nichols & Stich 2003) to describe how they feel at randomly cued intervals. On a more contrarian note, the number of subjects in these experiments are small (N less than 5) and consisted of extremely high-functioning patients, blunting some of the force of such a charitable interpretation. Other experimental results with autistic populations suggest serious deficits with introspective judgments as well as mindreading.

Those diagnosed with schizophrenia provide a second set of clinical data on both mindreading and introspection. Recently, large scale studies conducted by (Sprong 2007, Corcoran 2001) have suggested deficits in mindreading across different categories of schizophrenia. Schizophrenia has long been thought of as a characteristic deficit in introspection and self-monitoring, with delusions resulting from an inability to properly identify stimuli as being generated internally by the operations of the mind (e.g. inner speech, volitional imagery) or externally by other sources (Frith & Done 1988).

A third set of individuals consists of those with severe brain damage or those who have for some reason, required a commissurotomy, or severing of the main bundle of neural fibers connecting the right and left hemispheres of the brain. It has been reported that this subject pool demonstrates that the left hemisphere of the brain generates unconscious, automatic self-interpretations of the form we mentioned earlier (Gazzaniga 1967). Finally we have numerous psychological studies purporting to show healthy subjects having only the most tenuous grip on their inner lives. Perhaps most famous are the early studies of Nisbett and Wilson demonstrating subjects' lack of insight into the processes whereby they arrive at a decision (Nisbett & Wilson 1977). In this case, the subject falls prey to a particular form of automatically induced bias, but is asked for an explanation for why they chose as they did. It's unclear to us and apparently to Nisbett and Wilson as attested in their later writings (Wilson, 2002) that these results challenge the notion of introspection as traditionally conceived.

Prior Work

As we've mentioned, introspection and mindreading have been perennial topics in the philosophy of mind, and have now become important areas of study for psychologists and neuroscientists. While it isn't feasible to even topically review the prior work in the area, two sets of items are worth mention. The first of these concerns the lack of consensus on how to perform experiments to test claims about introspection, and subsequently how to interpret the results. Many of the studies performed have subject pools with $N < 5$, and rely on hermeneutical analyses of written reports by these subjects to draw conclusions (Hurlburt & Heavey 2006). The second claim, which relates in a way to the first, is that while purporting to explain the variety of phenomena we've mentioned so far, contemporary theories of introspection (Carruthers 2009, Nichols & Stich 2003) provide little more than box-and-arrow diagrams and verbal argumentation to support their favored position. Much of the verbal argumentation is aimed toward giving a convincing interpretation for the so-called data on introspection, which itself seems to defy consistent analysis, even by co-authors (Hurlburt & Schwiztgebel 2007)! Many of these theories endorse one form or another of the so-called theory-theory, simulation theory, or modular theory of mindreading. While space doesn't allow for detailed descriptions of the commitments made by each of the preceding options, we think it to be generally the case that each provides a set of constraints as to how computations underlying both introspection and mindreading might be made. In very broad strokes, theory-theory is committed to the existence of a body of theoretical knowledge about how beliefs, desires and other mental states stand in causal relation to one another to enable the prediction and explanation of behavior. Various strains of theory-theory have been proposed to underwrite both mindreading and introspection (Gopnik 1993). One way that theory-theory can be applied is inside a cognitive module, which is somewhat isolated from central cognition, and houses specific representational and processing resources dedicated solely to mindreading and introspection. Modules are generally thought to implement specific computational constraints on the variety and complexity of information allowed in and out of them, but different theorists have different takes on what these constraints are (Carruthers 2009, Leslie & Thaiss 1992). Finally, simulation theorists propose that we use our own mental states and inferential resources to construct mental simulations of ourselves-as-the-target, where the target is an agent whose behavior is to be predicted or explained (Goldman 2006). Current theorists have used these frameworks to define their particular notions of mindreading and introspection. Along with interpretation of clinical and other data, constraints generated by theory-theory and its' alternatives have led researchers to draw conclusions about whether or not these two abilities are served by different or identical computational mechanisms.

Imprecision

What seems so curious to us is why these theorists choose to commit to any of the frameworks we just mentioned in the last section. In essence, both simulation theory and modular theories of mindreading were developed as reactions to what are perceived implausibilities associated with theory-theory. For example, questions remain about what the contents of such a theory would be and how inference is performed efficiently using them. Classical questions from the artificial intelligence perspective regarding computation over such theories in dynamic environments (e.g. the frame problem, the relevance problem and their cousins) have never been addressed by the leading proponents of theory-theory. In addition theory-theory seems to commit to theories about the mental states of others, but also theories about how mental states are manipulated by inference procedures. Having detailed theories of the inferential tendencies of others seems to be a bit of an intellectual stretch for many. Similar questions about the structure and constraints that modules impose plague supporters of modular ideas about mindreading and introspection. The imprecision we describe poses not only a problem for a theory-laden interpretation process, but also for off-line simulation theorists (Goldman 2006) and some simulation-theory hybrids (Nichols & Stich 2003). In these cases, the mindreader selects a number of "pretend" beliefs, desires, and other relevant mental states and inserts them into their own practical decision-making system, taking the result "off-line;" meaning, any actions inferred in light of these pretend states are not actually sent to the motor system for execution as they would normally be for non-pretend inputs. While at least one of us (PB) is sympathetic to simulation, it isn't clear on any account of simulation how the pretend inputs are selected for simulation in the first place. All of these concerns serve to illustrate a more general point about theories of mindreading. In general, those who propose conceptual models for mindreading do so with an eye to philosophical issues or to empirical data without regard to how computations performed by these models might take place.

We feel that computational implementation provides at least a coarse guide to how feasible one option might be over another. Most computational models have been of the false belief task (Wimmer & Perner 1983). Examples from (Goodman et al. 2006), (Bello et al. 2007) and (Berthiaume 2008) almost completely cover the space, which is somewhat disappointing, given the many hundreds of false belief studies and associated variants that have been conducted since Wimmer and Perner's original experiment. While space doesn't allow for a detailed discussion, we now turn toward sketching an implementation of mindreading and introspection in a computational cognitive architecture that captures some of the general phenomenology and is sensitive to the constraints imposed by psychological and clinical studies.

Cognitive Architecture

Descriptions of the Polyscheme cognitive architecture in which we have conducted our modeling efforts can be found in (Cassimatis et al. 2009). A detailed account of the architecture and how coordination is achieved between its various elements can be found therein. For the sake of exposition, we only describe architectural features that are central to our account of the mindreading-introspection relationship.

Cognitive Architecture: Specification

Polyscheme is comprised of a number of *processing elements* (PE's) that communicate with one another via a *focus of attention* (FoA). Each PE maintains its own proprietary memory, data structures, algorithms for elaborating propositions, and internal knowledge representation that maps onto propositional form. Every PE is wrapped in an interface that allows two-way communication with the FoA through a propositional language. Choices of what PE's to include in the architectural specification are made through appeal to evolutionary, cognitive developmental, neuroscientific, and computational constraints. The PE's that serve our purposes in explaining mindreading are represented in figure 1 and include rule matching, categorization, gaze detection, difference detection, identity hypothesis generation/evaluation, temporal and spatial reasoners, and a perceptual buffer.

Strings of the form $P(x_0, \ldots, x_n, t, w)$ are called *propositions*. Simply stated, *P* is a relation (i.e. *Loves*, *Hates*, *Color*, *MotherOf*) over the set of objects x_i during the temporal interval *t* in a world *w*, which bears a truth value. We designate "E" as the temporal interval containing all other temporal intervals. A proposition's truth-value is a tuple <F, A> consisting of the positive evidence for (F) and negative evidence against (A) the proposition and a scalar valence. Evidence takes on one of the following values: F, $A \in \{C, L, l, m, n\}$ representing *certainly*, *very likely*, *likely*, *maybe*, and *unknown*.

Cognitive Architecture: Mindreading

Propositions in Polyscheme have truth-values in mentally simulated worlds. Polyscheme's "beliefs" that are derived from perceptual data or via inference exist as propositions that are true in "R" or the real world; however the architecture is also capable of entertaining counterfactual, past, future-hypothetical, and other forms of simulated worlds. Polyscheme's "beliefs" about the real world are propositions with "R" in the final argument slot. What we're really interested in is how Polyscheme is able to identify and reason about the beliefs of other agents, including reflection on its own beliefs. In past work, we have shown how 3rd-person ascription is reducible to a substrate of domain-general representational primitives and processing elements including mental simulation of counterfactual worlds, reasoning about identity, categories, and by applying conditional rules (Bello et al. 2007). While

this surely sounds like quite a lot of mechanism, all of these abilities seem to be roughly in place by two years of age in typical human children, and none of them implies any commitment to innate modules or core theories. We do take mental simulation to be a critical operation for the ascription

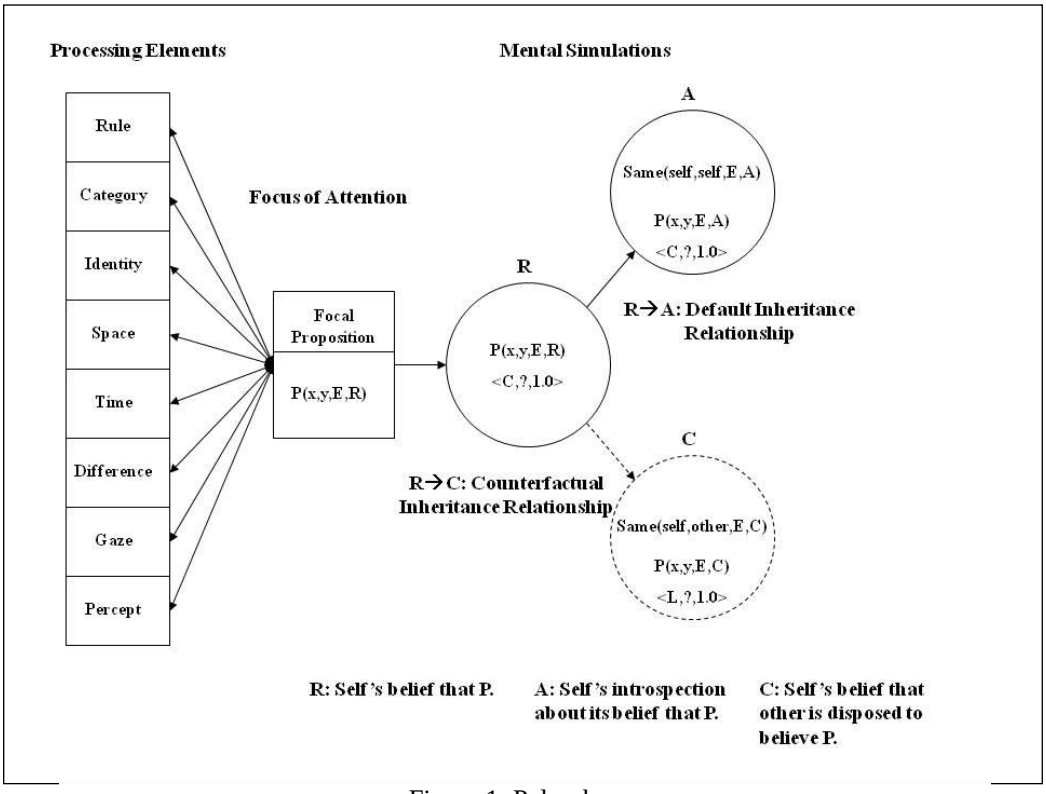


Figure 1: Polyscheme

of beliefs, which according to our theory proceeds in the following way:

1. Categorize other entity as an agent using category PE.
2. Construct counterfactual world C where Same(self, other, E, C) is true.
3. Detect differences between self and other using identity PE
4. Apply an override for each difference detected using conditional rule PE, forcing self-related propositions to resemble other-related propositions.
5. Proceed with inference and predict behavior appropriately.

The conditional rule PE implements a general-purpose rule that roughly looks like the following:

Holds(?P, self, ?t , ?w) ^ -Holds(?P, other, ?t, ?w) ^ Same(self other, E, ?w) ➔-Holds(?P, self, ?t, ?w)

Actual implementation of this rule is somewhat more complex, but incidental to our discussion. It suffices to say that mismatches between self and other-related propositions are detected as exceptions in simulated worlds C where Same(self,other,E,C) is true. An immediate concern is how such a rule fails to immediately generate a contradiction, since Holds(?P, self, ?t, ?w) is true, and –Holds(?P, self, ?t, ?w) is inferred as a consequent. Recall that propositions in Polyscheme have truth-values that are more differentiated than bivalent true or false. Also recall that Polyscheme's beliefs are propositions indexed to "R," the real world. Worlds in Polyscheme are related to one another via a process of *inheritance*. Inheritance relates a child world to a parent world, and operates in the following way: if during the course of inference, Polyscheme is asked to focus on a proposition P in a child world, it will check to see if P has a truth value in that world. If it doesn't, Polyscheme will look at the child's parent world to see if P has a truth value there. If it does, the truth value for P in the child world will be assigned the same value it has in the parent world. The inheritance procedure is visually depicted in figure 1 above. The inheritance procedure captures the idea that if we are to imagine a world in which some proposition like "pegasus exists" is true, other unrelated things we know about, such as "New York is north of DC" are vacuously true in our imagined world by virtue of the fact that they inherit truth values for these propositions from "R," the real world.

The rule we've given that performs an override looks like it might generate a contradiction. Polyscheme's world-simulation PE detects that Same(self,other, E, C) is a counterfactual claim, and when inheriting truth-values from

the parent world "R" for propositions in the counterfactual child-world C, they inherit into C as only being very likely true or very likely false, rather than the certainly true or certainly false values they would be assigned if the counterfactual status of Same(self,other, E, C) was never detected. Since Holds(?P, self, ?t, C), etc. would inherit into C with less-than-certain truth values, Polyscheme can continue to infer in C without running into the danger of contradiction.

Inheritance, Overrides and Mindreading

How do inheritance and overrides in simulation relate to one another, and to both mindreading and introspection? We will differentiate between introspection of currently-held beliefs and 3rd-person ascription by appealing to different inheritance relationships with "R" that define them. Specifically, we are interested in the difference between *alternate* worlds and *counterfactual* worlds. We qualify what we mean by alternate world in the following fashion: an alternate world is such that no proposition in it is the truth-functional negation of a proposition in its parent world. For purposes of our discussion, "R" will always be the parent world of whatever simulations we are considering, whether they are alternate worlds or counterfactual worlds. This is in contrast to counterfactual worlds, which we've already explained, and which contain propositions that are truth-functional negations of propositions in their parent worlds. The difference between these two modes of simulation is illustrated in figure 1. When introspecting on currently-held beliefs, Polyscheme entertains an alternate world in which it is the same as itself. It does so by inheriting from its parent world "R" using an inheritance relationship called I_{aw}. We call this the "default" inheritance relationship since it perfectly preserves truth-values for propositions between parent and children worlds. In contrast, the counterfactual inheritance relationship, called I_{cw}, weakens the truth values for propositions inherited from a parent world R into a child world C, allowing counterfactual reasoning to proceed without immediately inferring a contradiction.

When introspecting, an alternate world A is considered in which Same(self, self, E, A) is true. According to the definition of strict identity, there are no differences between self and self, and thus nothing to override in such a world. However, when simulating oneself in the past or in the future, we might simulate a counterfactual world where Same(self, self_at_now-2, E, C) or a world where Same(self, self_at_now+10, E, C), and so on. Since these past or future versions of oneself might be importantly different from the standpoint of mental states, we note differences between these versions of ourselves and our current self, perform appropriate overrides, and make subsequent predictions or develop explanations. In this way, some sorts of introspective judgments work exactly the same way as 3rd-person ascription of mental states, while not committing us to the idea that introspection and mindreading are somehow identical and served by exactly the same set of cognitive operations (Carruthers 2009).

Accounting for the Data

Our theory satisfies a number of the conditions discussed in our introduction. Firstly, it should be clear that since we are simulating a world where we are ourselves, introspection about current mental states is clearly not aimed at perceptual features or external objects. The objects under consideration are propositions inherited from Polyscheme's set of beliefs. This satisfies #1, the *mentality condition.* Since we differentiate simulating alternate worlds in which currently-held mental states are considered, versus counterfactual worlds in which either simulate ourselves as another agent entirely, or simulate ourselves in the past or future, there is a temporal constraint put on what we consider to be introspection proper. Simulation of past and future-selves certainly would count as self-knowledge, but there are acknowledged differences between self-knowledge broadly speaking, and introspection proper. This satisfies #3, or the *temporal locality condition.* Inheritance is not an inferential operation in the sense of having an associated logical operator with an associated semantics. Inheritance floats and attenuates the truth values of propositions from parent worlds to their children when required. In this way, truth of a proposition in a simulated world is arrived at non-inferentially, satisfying #3, the *directness condition.* Introspective judgments made in alternate worlds do not require any overrides relative to their counterparts arrived at counterfactually. If we associate some degree of effort or cognitive cost to performing an override of any sort, judgments about currently held beliefs will be guaranteed to seem at least as easy and likely much easier than judgments made about the mental lives of others, or of ourselves in the distant past or future. This satisfies the #5, the *ease condition.* Finally, properties of the two different inheritance relationships produce propositions in child worlds with different truth values. Inheriting from R into an alternate world produces propositions in the alternate world that have exactly the same truth value that they do in R. This contrasts to the relationship between propositions in R, and how they inherit into counterfactual worlds with slightly weakened truth values. This suggests that introspectively considered propositions are more certain than their non-introspective counterparts, satisfying #2, the *certainty condition.*

As for the clinical and psychological data, it's difficult to speculate on how any existing model correctly accounts for disorders of mindreading and introspection. But speaking purely speculatively, some of the psychological data on confabulation (e.g. the Nisbett and Wilson results) can be attributed to the mechanisms in Polyscheme which produced its base set of beliefs in R. Since there is no requirement to have introspective access to the workings of these mechanisms, Polyscheme would merely take any propositional content generated by these mechanisms, and ascribe them to itself in an alternate world. In this way,

Polyscheme has introspective access to the propositional content, without necessarily having access to the means by which it is acquired. In the case of autism, much has been said about cognitive deficits associated with autistic patients. Some of these deficits include the inability to follow and understand the targets of other agents gaze, thus eliminating a major source of evidence for understanding what other people currently believe. Other deficits have been hypothesized to include an inability to separate self versus other-centric representations, marked deficits in engaging in pretence and other forms of counterfactual simulation, and general lack of global coherence in cortical processing, all of which are critical elements of our story about mindreading and introspection. Similar deficits in schizophrenic subjects might be addressed by lesioning or confusing our inheritance and world-simulation mechanisms, which detect whether or not we're mindreading self or other-related targets. Of course, these are wild speculations, and we haven't produced any implementation. We only mention them to provide a prima facie story about how much deficits might be reproduced in a computational cognitive architecture.

Summary

We have given the rudiments of an account of the relationship between mindreading and introspection in an existing computational cognitive architecture using a single simulative mechanism, but having separate conditions of operation for each. We discussed our model's capacity to capture some of the defining features of introspection that have yet to be accounted for by competing models, providing a new way to generate and test hypotheses regarding the relationship between mindreading and introspection. While space hasn't permitted the inclusion of detailed computational models and associated model traces, these can be found for an example of 3rd-person ascription (the false belief task) and 1st-person ascription (the smarties task) on the first author's website: http://www.pbello.com/mindreading.html produced in a deprecated version of Polyscheme.

References

Schwitzgebel, E., (2010). "Introspection", The Stanford Encyclopedia of Philosophy, Edward N. Zalta (ed.).

Carruthers, P. (2009). How we know our own minds: the relationship between mindreading and metacognition. *Behavioral and Brain Sciences, 32*, 121-138.

Baron-Cohen, S, (1995). *Mindblindness: an essay on autism and theory of mind*. Cambridge MA: MIT Press.

Nichols, S. & Stich, S. (2003). *Mindreading: an integrated account of pretence, self -awareness, and understanding of other minds*, USA: Oxford University Press.

Sprong, M., Schothorst, P., Vos, E., Hox, J. & van Engeland, H. (2007). Theory of mind in schizophrenia: meta-analysis. British Journal of Psychiatry, 191(1), pp 5-13.

Corcoran, R (2001). Theory of Mind in Schizophrenia. In: D. Penn and P. Corrigan (eds.) Social Cognition in Schizophrenia. American Psychiatric Association, Washington DC.

Frith, C. & Done, C. (1988). Towards a neuropsychology of schizophrenia. *British Journal of. Psychiatry 153*: 437–43.

Gazzaniga, M.S. (1967). The split-brain in man. *Scientific American 217*, 24-29.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84*, 231-259.

Wilson, Timothy (2002). *Strangers to Ourselves: Discovering the Adaptive Unconscious*. Cambridge: Belknap Press.

Hurlburt, R. & Heavey, C. (2006). *Exploring inner experience*, Amsterdam: John Benjamins.

Hurlburt, R., & Schwitzgebel, E. (2007). *Describing inner experience? Proponent meets skeptic*, Cambridge, MA: MIT

Gopnik, A. (1993). How we know our minds: The illusion of first-person knowledge of intentionality, *Behavioral and Brain Sciences*, 16: 1–14.

Leslie, A.M., & Thaiss, L. (1992). Domain specificity in conceptual development: Neuropsychological evidence from autism. Cognition, 43, 225–251.

Goldman , A. (2006). *Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading*. USA: Oxford University Press.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in children's understanding of deception. *Cognition, 13*, 103–128.

Goodman, N. D., Bonawitz, E. B., Baker, C. L., Mansinghka, V. K, Gopnik, A., Wellman, H., Schulz, L. and Tenenbaum, J. B. (2006). Intuitive theories of mind: a rational approach to false belief. *In Proceedings of the Twenty-Eighth Annual Conference of the Cognitive Science Society*.

Bello, P. Bignoli, P. & Cassimatis, N. (2007). Attention and Association Explain the Emergence of Reasoning About False Belief in Young Children. *In Proceedings of the 8th International Conference on Cognitive Modeling*, Hillsdale, NJ: Lawrence Erlbaum Associates.

Berthiaume, V., Onishi, K. H., & Shultz, T. R. (2008) A computational developmental model of the implicit false belief task. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, 825-830. Austin, TX: Cognitive Science Society.

Cassimatis, N., Bignoli, P., Bugajska, M., Dugas, S., Kurup, U. Murugesan, A. & Bello, P (2010). An Architecture for Adaptive Algorithmic Hybrids. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*.

Illusions of consistency in quantified assertions

Niklas Kunze[1] (niklas.kunze@uni-konstanz.de)
Sangeet Khemlani[2] (khemlani@princeton.edu)
Max Lotstein[2] (lotstein@princeton.edu)
P.N. Johnson-Laird[2] (phil@princeton.edu)
[1]Department of Psychology, University of Konstanz, 78457 Konstanz, Germany
[2]Department of Psychology, Princeton University, Princeton, NJ 08540, USA

Abstract

The mental model theory of reasoning postulates that individuals establish the consistency of a set of assertions by constructing a mental model in which all the assertions hold. Mental models represent what is true but not what is false, and this principle of 'truth' predicts that certain assertions should yield systematic errors. We report an experiment in which participants evaluated the consistency of assertions based on quantifiers and sentential connectives, e.g., *All of the artists are barbers or else all of the barbers are artists; Some of the artists are not barbers*. The results showed that participants judged consistent assertions to be inconsistent, and vice versa, much more often for the predicted assertions than for control problems, which should be unaffected by the failure to represent what is false. These results provide a litmus test for mental models, because no current alternative theories of reasoning predict them.

Keywords: deductive reasoning; mental models; consistency; illusions.

Introduction

Are humans inherently rational? Without any training, they are able to make valid deductions. Consider the following problem:

Carol invested in capital securities or else she invested in municipal bonds.
She did not invest in municipal bonds.
Therefore, she invested in capital securities.

You do not need to know anything about securities or bonds to tell that the inference is valid. The ability to make valid inferences is a cornerstone of rationality, and as such, many theories argue that humans make use of formal rules akin to those in logic (Braine & O'Brien, 1998; Rips, 1994). According to those theories, humans make mistakes because they misapply the rules. Likewise, theories based on a probabilistic calculus believe that humans are rational, and that cognitive scientists use the wrong criteria to assess rationality, because everyday reasoning is probabilistic (Oaksford & Chater, 1998, 2007). Our own alternative theory is that human reasoning is based on *mental models*, or iconic representations of possibilities (Johnson-Laird, 2006; Johnson-Laird & Byrne, 1991). Mental models represent what is true and not what is false, and this constraint can lead to inaccurate models of assertions. Thus, human reasoning is fallible in practice, and individuals should succumb to systematic errors in judgment and inference. These errors are at present a unique prediction of the model theory, and so they serve as a litmus test for mental models. In the present paper we examine fallacious judgments of consistency for assertions combining quantifiers such as 'all' and connectives such as 'or else'.

Mental models and illusions

A foundational assumption of the model theory is the *principle of truth*: the mental models of a set of assertions represent only those possibilities consistent with the truth of assertions. The principle applies to assertions as a whole as well as to clauses within them. For example, an exclusive disjunction *A or else not B* yields the following mental models, where each horizontal line represents a model of a possibility, and '¬' is used to denote negation:

A
¬ B

The models do not represent what is false according to the disjunction, such as the case in which *A* is false and *B* is true. And, for those possibilities that make the exclusive disjunction true, such as the case in which *A* is true, the falsity of the corresponding possibility, in this case the negation of ¬*B*, hence *B*, is not represented in the models. This means that a *literal* (a proposition such as ¬*B* that contains no sentential connective) is represented in a model only if it is true in a possibility. Thus, the first of the models above represents the possibility that A is true, but it does not represent the fact that the literal ¬*B* is false in the possibility. Mental models do not represent what is false, whether it is an affirmative or negative literal, but in certain cases, such as when an inferential task is easy, individuals can construct *fully explicit models*. They represent both what is true and what is false in each possibility and therefore yield the correct representation of the assertion. The fully explicit models of *A or else not B* are as follows:

A B
¬ A ¬ B

where the affirmative *B* in the first model represents the falsity of the negation, ¬*B*, and the negative ¬*A* in the

second model represents the falsity of affirmative, *A*. As these models show, the disjunction is equivalent to the biconditional: *A if and only if B*. However, very few people grasp the equivalence, because it is evident only to those who envisage what is false and construct fully explicit models.

The principle of truth may seem like a useful compromise to reduce the load on working memory, but it has an unexpected consequence: it predicts *illusions*, i.e., judgments and inferences that are compelling but erroneous (Johnson-Laird & Savary, 1999). Illusions can occur when individuals assess conclusions, make inferences, or evaluate the consistency of a set of assertions. For example, consider this problem:

> There is a pin and/or a bolt on the table, or else a bolt and a nail on the table.
> There is a bolt and a nail on the table.
> Is it possible that both assertions could be true at the same time?

Reasoners in one study overwhelmingly responded 'yes' (Johnson-Laird, Legrenzi, Girotto, & Legrenzi, 2000), but the response is a fallacy predicted by the principle of truth. It is a fallacy because 'or else' in the first premise is an exclusive disjunction, i.e., if one clause of the disjunction is true, the other must be false. Hence, the truth of the second premise, *there is a bolt and a nail on the table*, implies that both clauses in the first premise are true. And that contravenes the meaning of 'or else', which means that the two assertions cannot be true at the same time. However, reasoners do not grasp this inconsistency and instead incorrectly judge the two assertions to be consistent.

Previous studies have corroborated the existence of illusory *deductions* from disjunctive and biconditional premises (Johnson-Laird & Savary, 1999). They have also corroborated them in singly quantified premises (Yang & Johnson-Laird, 2000). But, no study has examined the interaction between connectives and quantifiers. Our aim was accordingly to test whether illusions also occurred in a new domain: the evaluation of the *consistency* of assertions that depend on both quantifiers and connectives (as in Problem 1 below).

Illusions with quantified assertions

Our experiment examined two sorts of assertions that should yield illusions: exclusive disjunctions of quantified assertions, such as *All the A are B or else some of the B are A,* and biconditionals of quantified assertions, such as *All of the A are B if and only if All of the B are A*. Half of the problems used in the study were those that the principle of truth predicts should yield illusory judgments of consistency, and the other half were those that the principle of truth predicts should yield correct responses. Here is an example of an illusory problem based on an exclusive disjunction:

> *1. Illusion (disjunction)*
> All of the artists are barbers or else some of the barbers are artists.
> All of the barbers are artists.
> Is it possible for both statements to be true at the same time?

The disjunction yields two mental models that represent the two clauses (*All of the artists are barbers* and *Some of the barbers are artists*). Each model contains a set of individuals, where each line represents an individual and denotes the individual's properties. We lay out the two models as follows:

1.		2.	
[artist]	barber	barber	artist
[artist]	barber	barber	artist
		barber	

These models represent each set by a small but arbitrary number of individuals (two or three in the present models), and the square brackets denote that a set has been represented exhaustively (cf. the notion of 'distribution' in logic). One consequence of these exhaustively represented properties is that they cannot be added to new individuals in the model (see, e.g., Johnson-Laird, 2006). So, you cannot add instances of artists that are not barbers to the first model.

Consider the first mental model, which represents the first clause of the disjunction, *All of the artists are barbers*. The second assertion in the problem, *All of the barbers are artists*, is true in this model. The model theory predicts that individuals judge a set of assertions to be consistent if all the assertions hold in at least one mental model. Hence, people should judge that the two assertions are consistent. However, this judgment is flawed. The principle of truth predicts that the mental models represent the truth of each clause in an exclusive disjunction, but not the concurrent falsity of the other clause. Suppose that the first clause in the disjunction, *All of the artists are barbers*, is true. Hence, the second clause must be false, i.e., none of the barbers is an artist. This case is inconsistent with the first clause of the disjunction, and so it is impossible. Now suppose that the second clause of the disjunction is true, i.e., some of the barbers are artists. In this case, it must be false that all of the artists are barbers, i.e., at least some of them are not barbers. So, we have the conjunction of at least some of the barbers are artists and at least some of the artists are not barbers. There is accordingly just one fully explicit model of the disjunction:

artist	barber
artist	barber
	barber
artist	¬barber

The second assertion in the problem, *all the barbers are artists*, is accordingly inconsistent with this model, and the

correct evaluation of the two assertions is that they are inconsistent.

The same compound assertion used in (1) yields a control problem, as in (1'):

1'. Control (disjunction)
All of the artists are barbers or else some of the barbers are artists.
None of the barbers is an artist.
Is it possible for both statements to be true at the same time?

The second assertion, *none of the barbers is an artist*, is inconsistent with the mental models above, and so the theory predicts that individuals should respond that the two assertions are inconsistent. In this case, they will be correct, because the second assertion is also inconsistent with the fully explicit model above.

An example of an illusory problem based on a biconditional assertion is:

2. Illusion (biconditional)
All of the artists are barbers if and only if all of the barbers are artists.
None of the artists is a barber.
Is it possible for both statements to be true at the same time?

Biconditional assertions are true whenever both of its clauses are true or else when they are both false. In the case of a biconditional, the principle of truth predicts that a mental model of a biconditional will represent the possibility in which both clauses are true, but not the possibility in which both clauses are false. Hence, the biconditional assertion yields the following mental model in which the two sets of individuals are co-extensive:

[artist] [barber]
[artist] [barber]

According to the principle of truth, individuals should respond that the second assertion, *None of the artists is a barber*, is inconsistent with the first assertion, because the second assertion does not hold in the mental model above of the biconditional assertion. Mental models fail to represent the possibility in which both clauses of the biconditional are false, i.e., at least some of the artists are not barbers and at least some of the barbers are not artists. The fully explicit models would include both the model above and represent such a possibility, e.g.:

¬artist [barber]
[artist] ¬barber

This model *is* consistent with the second assertion in the problem, and so the correct response is that the two assertions are consistent.

The same compound assertion can also yield a control problem:

2'. Control (biconditional)
All of the artists are barbers if and only if all of the barbers are artists.
Some of the artists are barbers.
Is it possible for both statements to be true at the same time?

The second assertion is consistent with the mental model, which is a correct possibility, and so individuals should respond correctly that the two assertions are consistent.

Method

Participants and procedure. 28 participants were recruited through an online platform hosted by Amazon.com. None of the participants had received any training in logic. Participants were told to take as much time as they needed to answer the questions and were asked to answer as accurately as possible.

Design and materials. Participants acted as their own controls and evaluated 18 sets of assertions (see Appendix), and each set contained one compound quantified assertion (e.g., *All the artists are beekeepers if and only if some of the beekeepers are not artists*) and one simple assertion. There were four sorts of problem: illusions of consistency (C/I), where 'C' denotes the predicted response of consistent, and 'I' denotes the correct response of inconsistent, their controls (I/I), illusions of inconsistency (I/C), and their controls (C/C). 12 of the problems were based on disjunctions, and 6 of them were based on biconditionals, for which it is impossible to have illusions of inconsistency. For each set of assertions, participants pressed one of two buttons on the screen (labeled 'yes' and 'no') to respond to the question 'Is it possible for both statements to be true at the same time?' The contents of the assertions concerned occupations (e.g., artists, beekeepers, and chemists). Each participant received the problems in a different random order. The corresponding mental models and fully explicit models are given in the Appendix.

Results

Table 1 provides the overall percentages of correct responses for the six sorts of problem. The data strongly support the predictions of the model theory.

Table 1: The percentages of correct responses for illusory and control problems in the different conditions

	Illusions	Controls
Disjunctions		
Consistent problems	36%	85%
Inconsistent problems	7%	75%
Biconditionals		
Consistent problems	43%	80%

Overall, the illusions (29% correct) were reliably harder than the control problems (80% correct; Wilcoxon test, $z = 4.49$, $p < 0.0001$), and 24 out of the 28 participants did worse on illusions than controls (Binomial test, $p < .00001$). There was no reliable difference in performance between disjunctive and biconditional consistent problems (Wilcoxon test, $z = 0.12$, $p > 0.9$). And participants succumbed to illusory inferences in both disjunctive and biconditional assertions. As the Table shows, however, a reliable interaction occurred: the illusions of consistency were more compelling than the illusions of inconsistency (Wilcoxon test, $z = 3.24$, $p < 0.002$). The control problems demonstrated that participants interpreted the sentential connectives correctly. The illusions of inconsistency likewise rule out the possibility that individuals had alternative interpretations of the connectives.

One might be tempted to argue that the results could be explained if individuals interpreted the exclusive disjunctions as inclusive ones. Yet this cannot be the case, because inclusive disjunctions merely add possibilities, and so they do not change the consistency of the assertions in the problems. We conclude that illusions of consistency in quantified assertions are a robust phenomenon.

General Discussion

Our results show that robust illusions of consistency, and of inconsistency, occur with compound assertions that consist of quantified clauses. Participants were more likely to succumb to illusions of consistency – they judged that assertions were consistent when in fact they were inconsistent. One contributory factor may have been that individuals need to show that no model exists in which the assertions hold in order to establish inconsistency. In contrast, to establish consistency, they only need to construct one model in which the assertions hold. That is, inconsistency calls for a more exhaustive search than consistency.

In general, illusions serve as a litmus test for mental models, because no other current alternative theory can predict or explain the results. The results of the present study support the principle of truth., They also show that judgments of inconsistency are more difficult than judgments of consistency (Johnson-Laird, Girotto, & Legrenzi, 2004). Theories based on formal rules of inference (Braine & O'Brien, 1998; Rips, 1994) cannot account for the illusions, because these theories rely on valid rules of inference. If such theories incorporated invalid rules to explain illusions, they would predict many inferences that individuals never make. Invalid inference rules are a recipe for irrationality, and could render theories of deduction unstable. Moreover, performance can be enhanced when reasoners are given remedial instructions (Khemlani & Johnson-Laird, 2009; Yang & Johnson-Laird, 2000).

Theories based on the probability calculus cannot readily account for performance in the task of evaluating consistency either. Chater and Oaksford (1999, 2007) assign probabilistic meanings to quantified clauses. For instance, *All the A are B* is interpreted as meaning $p(B|A) = 1$, and *Some of the A are not B* means that $p(B|A) < 1$. But, how does one assess consistency? If it is simply a matter of consistent conditional probabilities, then a problem based on these assertions:

All of the A are B or else all the B are A.
Some of the A are not B.

should be judged as consistent, because both assertions can have a probability > 0. But, the model theory predicts that these assertions should be evaluated (erroneously) as inconsistent, and indeed 61% of our participants corroborated this prediction. At the very least, the probabilistic theory needs to add some additional machinery to cope with inconsistency. A conditional probability of the form:

$p(B \& C \& D \mid A)$

has the value of zero in case the conjunction of B, C, and D, is inconsistent, even if the individual conditional probabilities $p(B \mid A)$, $p(C \mid A)$, $p(D \mid A)$ all have non-zero values.

In sum, reasoners in our study made systematic errors in reasoning about the consistency of disjunctions and biconditionals of singly quantified assertions. They tended to err on problems that called for them to take into account possibilities that rendered the assertions false, and thus corroborated the model theory's principle of truth.

Acknowledgments

This research was supported by a National Science Foundation Graduate Research Fellowship to the second author, and by National Science Foundation Grant No. SES 0844851 to the third author to study deductive and probabilistic reasoning. We thank Olivia Kang, Cathy Haught, Matt Johnson, Sam Glucksberg, Adele Goldberg, and Laura Suttle for their helpful ideas and assistance.

References

Braine, M.D.S., & O'Brien, D.P., Eds. (1998). *Mental logic*. Lawrence Erlbaum Associates, Mahwah, NJ.

Chater, N., & Oaksford, M. (1999). The probability heuristics model of syllogistic reasoning. *Cognitive Psychology, 38.*

Evans, J.St.B.T., Newstead, S.E., & Byrne, R.M.J. (1993). *Human reasoning: The psychology of deduction*. Hillsdale, NJ: Erlbaum.

Johnson-Laird, P.N. (2006). *How we reason*. Oxford University Press.

Johson-Laird, P.N., & Byrne, R.M.J. (1991). *Deduction*. Psychology Press.

Johnson-Laird, P. N., Girotto, V., & Legrenzi, P. (2004). Reasoning from inconsistency to consistency. *Psychological Review, 111,* 640-661.

Johnson-Laird, P.N., & Savary, F. (1999). Illusory inferences: a novel class of erroneous deductions. *Cognition*, 71, 191-229.

Khemlani, S., & Johnson-Laird, P.N. (2009). Disjunctive illusory inferences and how to eliminate them. Memory & Cognition, 37, 615-623.

Oaksford, M., & Chater, N. (2007). *Bayesian rationality*. Oxford University Press.

Oaksford, M., & Chater, N. (1998). *Rationality in an uncertain world*. Hove, UK: Psychology Press.

Rips, L.J. (1994). *The psychology of proof*. Cambridge, MA: MIT Press,.

Rips, L.J. (1997). Goals for a Theory of Deduction: Reply to Johnson-Laird. *Minds and Machines*, 7, 409–424.

Yang, Y., Johnson-Laird, P.N. (2000). How to eliminate illusions in quantified reasoning. *Memory & Cognition*, 28, 1050-1059.

Appendix

The problems in the experiment in an abbreviated form, their mental models and their fully explicit models

Forms of Premises and Questions	Mental Models	Fully Explicit Models	% Correct
1. All of the A are B or else some of the A are B	[A] B [A] B A B A B	A B A ¬B ¬A B	
Some of the A are not B		Illusion of consistency	11
None of the A is a B		Control "no" response	82
2. All of the A are B or else some of the B are A.	[A] B [A] B B A B A	A B A ¬B ¬A B	
All of the B are A		Illusion of consistency	0
None of the A is a B		Control "no" response	71
3. All of the A are B or else all of the B are A.	[A] B [A] B [B] A [B] A	[A] B [B] A [A] B [B] A ¬A B ¬B A	
Some of the A are not B		Illusion of inconsistency	39
Some of the A are B		Control "yes" response	68
4. Some A are not B or else some B are not A.	A (A) [B] [B] B (B) [A] [A]	[A] B [B] A [A] B [B] A ¬A B ¬B A	
All of the B are A		Illusion of inconsistency	32
Some of the A are B		Control "yes" response	96
5. None of the A is a B or else some of the A are not B	[A] [B] A (A) [B] [B]	A [B] A ¬B	
None of the A is a B		Illusion of consistency	11
All of the A are B		Control "no" response	71
All of the B are A		Illusion of inconsistency	36
Some of the B are A		Control "yes" response	89
6. None of the A is a B if and only if some of the B are not A	[A] [B]	A [B] A	
All B are A		Illusion of inconsistency	39
Some A are not B.		Control "yes" response	71
7. All of the A are B if and only if all of the B are A	[A] [B] [A] [B]	[A] [B]	
None of the A is a B		Illusion of inconsistency	50
Some of the A are B		Control "yes" response	71
8. Some A are not B if and only if some B are not A.	A ¬B ¬A B A B	[A] [B]	
All of the A are B		Illusion of inconsistency	39
Some of the A are B		Control "yes" response	96

Can similarity-based models of induction handle negative evidence?

Daniel Heussen (Daniel.Heussen@psy.kuleuven.be)

Wouter Voorspoels (Wouter.Voorspoels@psy.kuleuven.be)

Gert Storms (Gert.Storms@psy.kuleuven.be)
Department of Psychology, Tiensestraat 102, 3000 Leuven, Belgium

Abstract

Even if we don't like it, we often face counterexamples to the inferences we have made or would like to make. With the exception of the SimProb model (Blok, Medin & Osherson, 2007), models of inductions to date have predominantly focused on the relevance of positive evidence to the inference process. Here we provide data from single and double premise arguments in a category-based property induction task using positive and negative evidence. A simple similarity model, the Similarity-Coverage model (Osherson et al., 1990) and the SimProb model are tested on negative and mixed evidence arguments.

Keywords: Induction; Negative evidence; Similarity

The relevance of negative evidence

Ever since Hume, induction has been an area of immense research efforts in philosophy (e.g., Goodman, 1955; Hempel, 1966; Lipton, 2004), psychology (e.g., Blok, Medin, & Osherson, 2007; Heit, 2000; Osherson et al., 1990; Rehder, 2009; Rips, 1975; Sloman, 1993) and cognitive science (e.g., Kemp & Tenenbaum, 2009) in general. Among the prominent questions studied have been: What is the logical basis for induction? What role does prior (semantic) knowledge play in inductive reasoning? Why are some kinds of fact more easily projectable than others? And how should we model inductive inference? Despite these extensive efforts little is known on the influence of negative evidence in induction.

Negative evidence, however, is ubiquitous in everyday reasoning. In some circumstance, evidence may go against our established views. Your favorite restaurant serves you a bad meal, your friend, that is always late, shows up on time and your oh so reliable car won't start. In other instances, you might be making a new inference with both positive and negative evidence present. You check out a new restaurant and receive a great starter and desert but a burned steak and overcooked vegetables. Negative evidence in category-based property induction is defined here as evidence from an instance of the conclusion category that does not possess the to-be-projected property. In other words the evidence constitutes a clear counterexample of something possessing the to-be-projected property. The questions we would like to address here are: How does negative evidence affect our generalizations? What determines the relevance of negative evidence? How do we combine evidence to reach a conclusion?

In research on induction involving positive evidence, Rips (1975) found that the similarity of the evidence to the conclusion influences its relevance. People are more willing to generalize the attribution of a property from a robin to a sparrow than from an eagle to a sparrow because robins and sparrows are more similar. Models of induction involving positive evidence have tried to capture this intuition. The similarity coverage model for instance uses the maximum similarity between premises and conclusion as one component to their model (Osherson et al., 1990). Similarly Sloman's (1993) feature model uses the overall match in the number of features between the premises and the conclusion as a determinant of argument strength. The SimProb model (Blok, Medin, & Osherson, 2007) turns similarity between premises and conclusion into probabilities and uses those to determine argument strength.

The question we are addressing here is whether similarity also determines the relevance of negative evidence. If similarity functions in the same way for positive and negative evidence in determining whether a piece of evidence is considered to be relevant to the conclusion, then existing models of induction based on similarity should be able to handle arguments involving negative evidence. To our knowledge, the SimProb model (Blok, Medin, & Osherson, 2007) is the only model explicitly designed to handle negative evidence. Other models require some adaptation to handle the intuition that the belief in a proposition should decrease with the encounter of negative evidence.

A second question of importance when modeling induction is how to combine the evidence. One approach might be to simply add to argument strength for positive evidence and subtract for negative evidence. Alternatively as the SimCov (Osherson et al., 1990) and the SimProb model (Blok, Medin, & Osherson, 2007) suggest, one could assign the greatest importance to one premise by virtue of its similarity to the conclusion for instance and adjust the resulting argument strength in accordance with the remaining evidence. Furthermore the manner in which the second premise exerts its influence can be implemented in different ways. The SimProb model suggests a weighting by similarity to the first premise. The SimCov model uses the relative positions of the premise categories in a conceptual similarity representation to determine the influence of additional premises. These are only a few examples of the various possibilities to combine data, but they highlight the complexity of the issue.

The aim here is to test whether similarity based models of induction are able to handle negative evidence in a category-based property induction task. We present data from an induction task involving single and double premise arguments with positive and negative evidence and fit three models. In the next section we'll describe in more detail the three models used.

Similarity-based models of induction

We evaluated three models, each relying essentially on similarity to predict the strength of an argument. The models differ in how information is combined in arguments with two or more premises and in the implementation of negative evidence premises. The first model is a simple similarity based model (Sim). The second model is the similarity-coverage model (SimCov) as proposed by Osherson et al. (1990). In the present study, we adapted the model to account for negative evidence. The third model is the similarity-probability model (SimProb; Blok, Medin, & Osherson, 2007).

The Sim model

In this model the strength of the argument is directly related to the similarity of the conclusion category and the premise category (or categories). Formally, the argument strength S_c of an argument with conclusion c and a set of premises then is:

$$S_c = \sum_{p=1}^{n} e_p sim_{cp}$$

where sim_{cp} is the similarity between the conclusion category and the category of premise p and e_p indicates whether the premise is positive or negative (respectively $e_p=1$ or $e_p=-1$). Note that in this expression similarities are combined in a very straightforward manner, summing them (or subtracting, depending on whether it's a positive or a negative premise) across the number of premises.

The SimCov model

In the SimCov model, the strength of an argument depends on two components. A similarity component captures the similarity between premise and conclusion categories, and thus the relevance of the premise. The coverage component captures the idea of how much of the nearest superordinate category containing both premise and conclusion categories is covered by the premise(s). We modified the model to account for negative evidence by making the similarity of a premise and a conclusion category negative when the premise is negative.

Formally, the argument strength according to the SimCov model is a weighted sum of the similarity and the coverage component:

$$S_c = \alpha \times similarity_c + (1-\alpha) \times coverage_p$$

where α is a free parameter determining the relative weight of each component. The similarity component represents the similarity between premise and conclusion category. In case of multiple premises, the similarity component is equal to the premise category that is most similar to the conclusion category. As in the previous model, when the most similar premise category is in a negative premise, the similarity is negative.

The coverage component is calculated as follows:

$$coverage_p = \frac{1}{N}\sum_{i=1}^{N} \max\left(\mathrm{sim}_{p_1 i}, \mathrm{sim}_{p_2 i}, \ldots, \mathrm{sim}_{p_n i}\right)$$

where i is an element of a relevant comparison set and N is the size of that set. The comparison set consists of known members of the nearest superordinate category containing both premise and conclusion categories. The coverage term implements the diversity principle (Carey, 1985). In a double positive premise argument, the more diverse the two premise categories are, the larger the coverage term will be – the more the nearest superordinate category is "covered" by the premise categories. Again, when the most similar premise category is in a negative premise, the similarity is negative in the expression.

The SimProb model

In the simprob model, inductive reasoning is considered as a conditional probability judgment. Given a certain prior belief about something, the evidence considered will update this prior belief. Formally, the belief update elicited by the premise a is given by:

$$P(c|a) = P(c)^{\alpha}$$

with

$$\alpha = \left(\frac{1 - sim_{ca}}{1 + sim_{ca}}\right)^{1-P(a)}$$

When there are two premises, the most relevant premise a (the premise that would influence the prior belief the most) is combined with the lesser relevant premise in the following way:

$$P(c|a,b) = P(c|a) + \begin{bmatrix} (1 - P(c|a)) \times \\ (1 - sim_{ab}) \times \\ (P(c|b) - P(c)) \end{bmatrix}$$

There are elegant symmetrical expressions to implement negative evidence (see Blok et al., 2007, for details). The basic idea is that the probability of a negative premise is 1 minus the probability of the same but positive premise, and that similarity between two premises will raise the posterior probability of the conclusion instead of decreasing it.

The SimProb model makes use of prior beliefs regarding the premises and conclusion. In the present study, we use blank properties. Following Blok et al., (2007) in their handling of blank properties, we use a uniform and low prior probability (fixed at .2) for all premises and conclusions.

An obvious parallel between the three models is that they all rely heavily on similarity to account for argument strength. There are differences however, in how similarity is used and – for arguments with multiple premises – how premise information is combined. The Sim model simply adds and subtracts similarities in the multiple premise case. SimCov picks the most relevant premise based on similarity and discards the similarity of the other premise. SimProb picks the most relevant premise, updates the conclusion probability and then modifies the resulting probability according to the less dominant premises.

Present research

The primary goal of this study was to see whether models that use similarity as a determinant of relevance of the evidence are able to handle negative evidence. To that end, we first established what influence negative evidence has on argument strength. We then tested a simple similarity model (the Sim Model), that only takes similarity into account, the SimCov model (Osherson et al., 1990) that also considers the coverage of the conclusion category and the SimProb model (Blok, Medin, & Osherson, 2007), that was specifically designed to be able to handle negative evidence.

The models are evaluated on data from a standard category-based property induction task using properties that participants are likely to have very little knowledge about. The properties are projected from either one or two exemplars to another exemplar of the same category. Participants are asked to judge how likely the conclusion is given the premises, for instance, how likely is it that magpies have a syrinx given that parakeets have a syrinx? The models are tested on four kinds of arguments:

Single Positive:
Parakeets have a syrinx.
Magpies have a syrinx.

Single Negative:
Parakeets **do not** have a syrinx.
Magpies have a syrinx.

Double Positive:
Parakeets have a syrinx.
Penguins have a syrinx.
Magpies have a syrinx.

Mixed Positive & Negative:
Parakeets have a syrinx.
Penguins **do not** have a syrinx.
Magpies have a syrinx.

Note that in the mixed arguments, the negative premise was always the premise presented second.

Method

Participants 76 students from the University of Leuven, Belgium, participated in the study. Participants received course credits in return for participation.

Design Two groups of participants rated the inductive strength of 40 target and 14 filler arguments. Filler items were arguments that were clearly true or false. One group evaluated 20 single positive arguments and 20 mixed positive and negative premise arguments. Fillers for this group consisted of single and double positive arguments. The other group evaluated 20 single negative premise and 20 double positive premises arguments with fillers being single positive and mixed positive and negative premises. The exemplars and properties used were identical for the two groups matching the characteristics across positive and negative arguments.

Materials To create arguments, we selected exemplars from four animal categories (i.e., *birds, fish, insects & mammals*) from the Leuven Concept Norms (DeDeyne, et al., 2008). For each category, the norms contain exemplars generated by participants as well as pair-wise similarity ratings between them. The norms also contain typicality ratings for each exemplar. Exemplars of the two premises and the conclusion were matched for typicality across the single and double premise arguments. The to-be-projected properties were biologically plausible blank properties. For each animal category we selected five kinds of characteristics (i.e., anatomical, behavioral, developmental, metabolic, necessity) that people were likely to have little knowledge about (e.g., Robins require amylase for their digestion). The task was administered in form of a questionnaire. The first page contained a description of the task with the instruction and an example argument. This was followed by 54 arguments starting with 3 warm-up fillers. The remaining 11 fillers were evenly distributed across the items. One random order of items and its reverse was used.

Procedure The induction task was presented as part of a battery of test. Students participated in a large group and took no longer than 10 minutes to complete the task.

Results

Preliminary Analysis Five participants were excluded from the analysis due to a lack of variance in their responses. In a subsequent reliability analysis, the two groups showed high consistency in their responding (Cronbach's alpha of .88 and .95). The data were averaged across participants and subsequent analyses were carried out on the items.

Manipulation Check Each of the 40 target items appeared once with positive and once with negative evidence. Of these, 20 items were single premise and 20 were double premise arguments. Figure 1 shows the average argument strength across those four conditions.

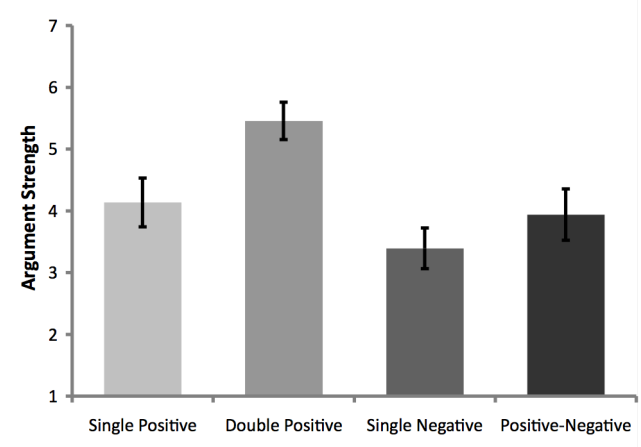


Figure 1: Argument strength for all four types of argument. Error bars are 95% CI.

Arguments containing negative evidence (darker bars) were rated lower in argument strength than those with positive evidence. For positive and negative evidence, arguments having two premises increased argument strength. Note though that in the mixed positive-negative premise arguments the increase in argument strength is due to the addition of a positive rather than negative premise.

The data were submitted to a 2 × 2 mixed factorial analysis of variance with type of evidence (contains negative evidence vs. does not contain negative evidence) as repeated measure and type of argument (single vs. double premise arguments) as between subjects factor. Although the data suggested that adding a positive premise has a greater effect if the first premise is positive as opposed to negative, the interaction between argument type and evidence type was not significant ($F(1, 38) = 3.2$, $p = .08$). Both main effects of type of evidence ($F(1, 38) = 27.8$, $p < .001$) and type of argument were significant ($F(1, 38) = 38.3$, $p = .001$). Single negative premise arguments were rated weaker than single positive premise arguments ($t(19) = 2.2$, $p < .05$). Similarly mixed positive-negative premise arguments were judged less strong than those with two positive premises ($t(19) = 5.9$, $p < .05$). Adding a positive premise to either a positive ($t(38) = 5.2$, $p < .05$) or a negative premise ($t(38) = 2.1$, $p = .05$) increased argument strength.

The data confirmed the intuition that negative evidence should have an adverse effect on argument strength. Arguments involving negative evidence were rated lower than those with positive evidence. For positive evidence, we also found a monotonicty effect (Nisbett, et al., 1983); more premises led to stronger arguments.

Modeling preliminaries In order to evaluate the model fits, we use the correlation between the averaged observed and predicted argument strength within each condition. To derive predicted values from the models, we extracted pair-wise similarity ratings between items from the Leuven Concept Norms (De Deyne, et al., 2008). Although the SimProb model provides predicted values in terms of conditional probabilities the other two models do not and we therefore do not make any claims about the scales of the predicted values and will not discuss differences between the models in those terms.

In terms of model parameters, the Sim model does not contain any free parameters. The SimCov model uses the alpha parameter to determine the relative influence of its two components (i.e., the similarity component and the coverage component). Figure 2 presents model fits (i.e., correlations between predicted and observed) across the whole range of the alpha parameter. In all four conditions a reduction in the alpha parameter led to a reduction in fit indicating that the coverage term did not play a role. Consequently we fixed the alpha parameter at 1.

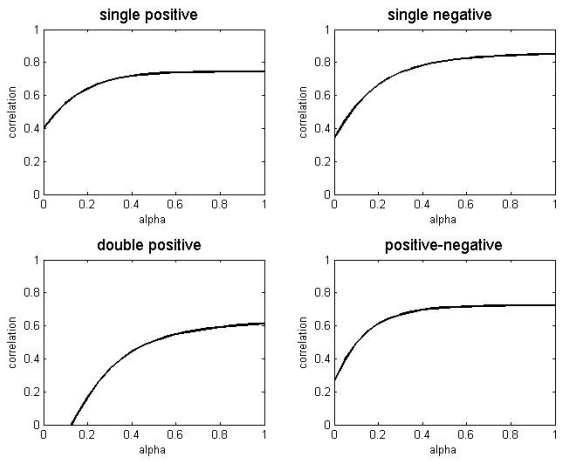


Figure 2: Model fits plotted against the complete range of the alpha parameter of the SimCov model in each condition.

The SimProb model requires prior probability judgments for the properties as input parameter to the model. Nevertheless, Blok et al. (2007) suggest that the SimProb model can handle arguments containing blank properties. They recommend using uniform and low prior probabilities, as this will ensure that the similarity component of their model will do most of the work. We therefore opted for uniform priors across premises and conclusion of .2.

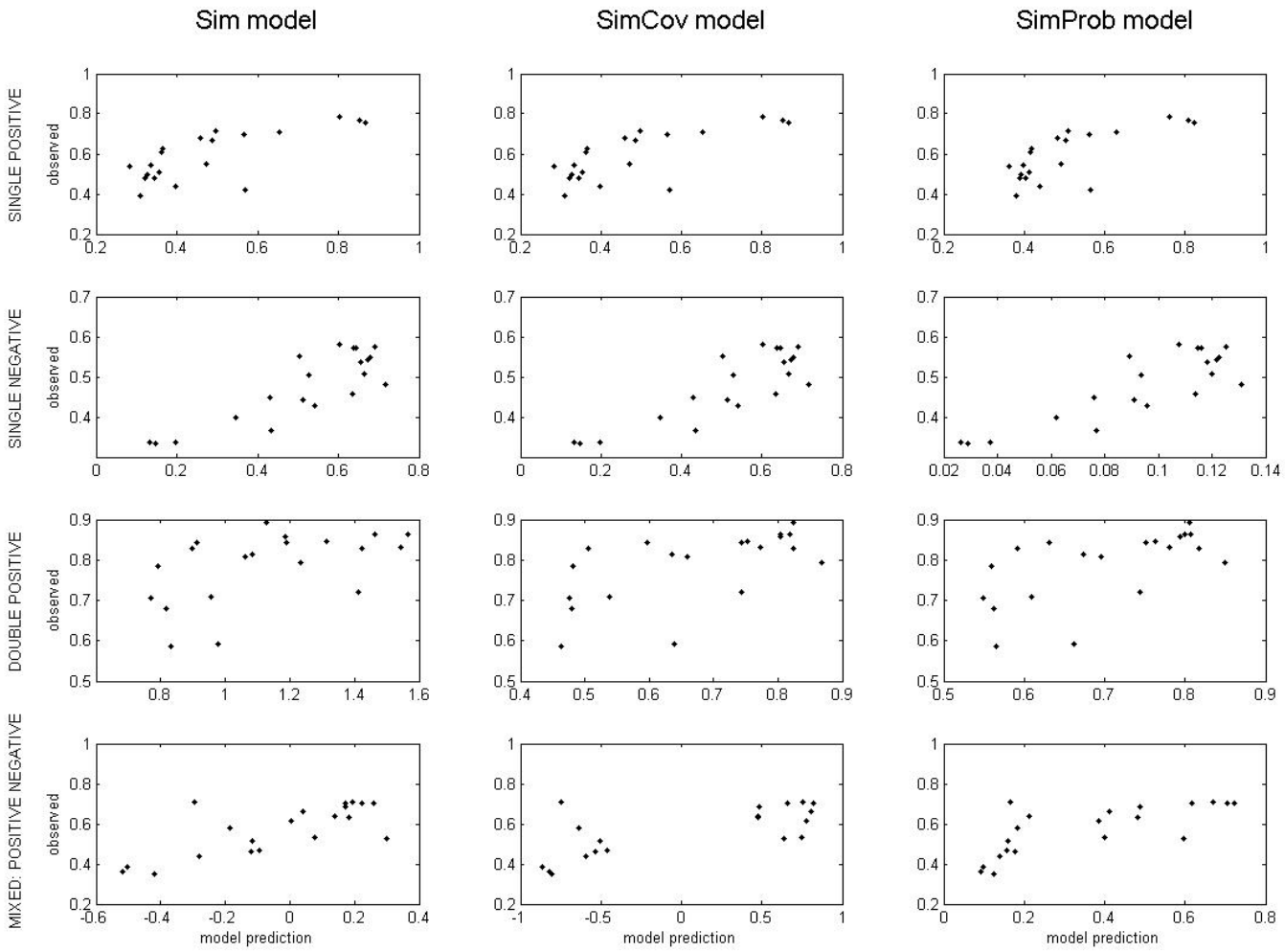


Figure 3: Scatter plots of observed against predicted values for each model across single positive, single negative, double positive and mixed positive-negative arguments.

Modeling results Figure 3 shows the scatter plots of the predicted versus observed values for each of the three models (columns) across the four types of argument (rows). All correlation coefficients were significant at $p < .05$ with n = 20. For single positive premises arguments (top row), the three models showed virtually identical results with a good fit of $r = .74$ for all three models. Looking at single premise arguments with negative evidence (2nd row), the models were equally capable at predicting participants' responses and even showed a better fit ($r = .85$). There was no difference in model predictions or fit across the three models. Thus for single premise arguments the three models can equally well account for argument strength involving positive and negative evidence.

The third row shows that for double positive premise arguments the three models differed in their predictions. The Sim model showed a somewhat weaker fit ($r = .53$) than the SimCov ($r = .61$) or the SimProb ($r = .62$) models. Applying a t-test to the Fisher's Z transformed correlation coefficients however showed that the difference was not significant ($t(17) = .56$, *n.s.*). Overall the fit of the models for double positive premise arguments was not as good as for single premise arguments.

Testing the fit for mixed positive and negative premise arguments (4th row) we find no difference between the models in terms of the correlation coefficient (Sim: $r = .75$; SimCov: $r = .73$; SimProb: $r = .73$). However the scatterplot shows that the SimCov model, unlike the other two, predicts two separate clouds of data points across the range of observed values. The human data clearly showed a continuous distribution across the whole range of possible values without two separate clouds. The difference in overall mean of each cloud in the predicted data seems to drive the correlation. This is due to the max function in the

similarity component choosing the premise (positive or negative) that has the greater similarity and dropping the influence of the other premise. In contrast the Sim model and the SimProb model take both premises into account.

General Discussion

In making an inference, we have to determine whether a piece of information is relevant or not. For evidence in favor of our inference, theories of induction (Blok, Medin, & Osherson, 2007; Osherson, et al., 1990; Rips, 1975; Sloman, 1993) have suggested that the relevance is determined by the similarity between the evidence and the conclusion. In everyday reasoning, however, we often face at least some evidence that is not in line with our favored conclusion. Here we have tested whether models that use similarity to determine relevance are able to handle arguments involving negative evidence.

The model fits showed that for single premise arguments all three models were able to account for the data from both positive and negative premise arguments equally well. This indicates that the relevance of negative evidence can also be modeled using similarity. For double premise arguments all three models did a decent job with positive evidence. However, for mixed positive–negative premise arguments only the Sim and the SimProb model were able to account for the data. Although showing a good fit in terms of the correlation coefficient, the SimCov model showed a pattern of predicted values not reflected in human data. Taken together, two factors can account for the behavior of the SimCov model. First, with our data the coverage component of the SimCov model did not contribute to the prediction of argument strength. One reason for this might be that the generalizations in our arguments were to other exemplars rather than the category itself. Second, the similarity component only takes into consideration the most similar premise disregarding the other. If this happens to be the negative one, predicted values are low. Conversely if the max function selects the positive premise predicted values are high. Without an influence of the coverage terms two clusters of predicted values emerge.

The results from the double premise arguments again support the fact that similarity can be used to determine the relevance of negative just as well as positive evidence. However the results highlight that with several pieces of evidence it becomes important to consider how to model the combination of both positive and negative evidence. Differences in how the models combine the evidence make them better or worse candidates in modeling negative evidence with multiple premises. Disregarding one piece of evidence over another clearly does not resemble participants responses. However similarly a simple additive model like the Sim model becomes less realistic in the case of multiple premises of the same kind, evident in our double positive condition.

The aim of the present study was not to provide a new model of induction but to test whether similarity-based models of induction can handle arguments involving negative evidence. We have shown that similarity can indeed be used to model relevance of negative evidence. In addition, our data highlight the importance of taking all evidence into account. Models of induction that try to account for the influence of negative evidence will need have a specific mechanism to combine positive and negative evidence.

Acknowledgments

This research was supported by a postdoctoral research fellowship within the framework of international mobility awarded to the first author by KU Leuven under the supervision of Gert Storms.

References

Blok, S. V., Medin, D. L., & Osherson, D. (2007). Induction as conditional probability judgment. *Memory and Cognition, 35,* 1353–1364.

Carey, S. (1985). *Conceptual change in childhood*. MIT Press.

De Deyne, S., Verheyen, S., Ameel, E., Vanpaemel, W., Dry, M., Voorspoels, W., & Storms, G. (2008). Exemplar by feature applicability matrices and other Dutch normative data for semantic concepts. *Behavioral Research Methods, 40,* 1030-1048.

Goodman, N. (1955). *Fact, fiction, and forecast.* Harvard University Press.

Heit, E. (2000). Properties of inductive reasoning. *Psychonomic Bulletin & Review, 7,* 569-592.

Hempel, C. G. (1966). *Philosophy of Natural Science*. New Jersey, Prentice Hall.

Kalish, C. W. & Lawson, C. A. (2007). Negative evidence and inductive generalisation. *Thinking & Reasoning, 13*, 394-425.

Kemp, C. & Tenenbaum, J. B. (2009). Structured statistical models of inductive reasoning. *Psychological Review, 116*, 20-58.

Lipton, P. (2004). *Inference to the best explanation*. London, Routledge.

Nisbett, R. E., Krantz, D. H., Jepson, D., & Kunda, Z. (1983). The use of statistical heuristics in everyday reasoning. *Psychological Review, 90*, 339-363.

Osherson, D. N., Smith, E. E., Wilkie, O., Lopez, A., & Shafir, E. (1990). Category-based induction. *Psychological Review, 97*(2), 185–200.

Rehder, B. (2009). Causal-based property generalization. *Cognitive Science, 33*, 301-343.

Rips, L. (1975). Inductive judgments about natural categories. *Journal of Verbal Learning and Verbal Behavior, 14*, 665-681.

Sloman, S. A. (1993). Feature-based induction. *Cognitive Psychology, 25*, 231–280.

The Role of Dynamic Visualizations and Spatial Layout of Static Visualizations for Learning How to Classify Locomotion Patterns

Birgit Imhof (b.imhof@iwm-kmrc.de)[a]

Katharina Scheiter (k.scheiter@iwm-kmrc.de)[a]

Peter Gerjets (p.gerjets@iwm-kmrc.de)[a]

Jörg Edelmann (edelmann@gris.uni-tuebingen.de)[b]

[a] Knowledge Media Research Center, Konrad-Adenauer-Strasse 40, 72072 Tuebingen, Germany
[b] Wilhelm Schickard Institute, University of Tuebingen, Sand 14, 72076 Tuebingen, Germany

Abstract

In two studies the effectiveness of dynamic and multiple static visualizations was investigated for a highly perceptual learning task, namely locomotion pattern classification. In Study 1a, seventy-five students viewed either dynamic, static-sequential, or static-simultaneous visualizations. For tasks with intermediate difficulty dynamic visualizations led to better recognition of the locomotion patterns than static-sequential visualizations, but not than static-simultaneous visualizations. To test whether the presentation of the static-simultaneous visualizations in rows or their permanent visibility was accountable for this effect, three additional static-simultaneous conditions were investigated in Study 1b. Seventy-five students viewed the static-simultaneous visualizations either presented in columns, in matrices, or in circles. The dynamic condition outperformed all three additionally investigated static-simultaneous conditions in the intermediate tasks. Accordingly, for learning how to classify locomotion patterns dynamic visualizations are better suited than most static presentation formats. Nevertheless, presenting static-simultaneous visualizations appropriately can achieve equal results at least for tasks with intermediate difficulty.

Keywords: learning; dynamic visualizations; multiple static visualizations; spatial ability

Learning with Visualizations

Dynamic visualizations have not always been found to lead to better learning than static visualizations (Tversky, Bauer-Morrison, & Bétrancourt, 2002). Bétrancourt and Tversky (2000) have suggested that dynamic visualizations should be superior only for specific tasks. In particular, they will aid learning if understanding the content explicitly requires understanding of its dynamic aspects like trajectory or continuity of changes. These dynamic aspects can be conveyed directly through a dynamic visualization. Thus, in many studies in which dynamic visualizations failed to be beneficial, a direct depiction of the contents' dynamic aspects may not have been necessary (e.g., Byrne, Catrambone, & Stasko, 1999). On the other hand, tasks that require a profound understanding of continuous changes often benefit from dynamic visualizations (e.g., hand manipulation tasks, Ayres et al., 2009; Wong et al., 2009).

Similarly, the current study focuses on a task that explicitly requires identifying the continuity of the depicted dynamics and involves a strong perceptual component, namely recognizing biological locomotion patterns of fish as a basis of species classification. To accomplish this task, it is important that learners correctly perceive the underlying kinematics, for instance, to decide whether a fin moves in a wave-like or a paddle-like manner. The continuity of these dynamics can be shown explicitly only in dynamic visualizations. However, one can argue that multiple static visualizations may also foster the understanding of continuity, but that this is likely to depend on how they are presented. In particular, to foster the understanding of continuity static pictures have to be presented in a way that they facilitate mental animation (e.g., Paas, Van Gerven, & Wouters, 2007). Mental animation is the process of inferring movements from static pictures based on knowledge about relevant components and their causal relations to other components (Hegarty, 1992). We assume that both, temporal as well as spatial aspects of presenting static pictures affect how well they support mental animation.

Temporal Aspects of Presenting Static Pictures

The main difference concerning temporal aspects of presenting multiple static pictures is their sequentiality. They can be depicted either sequentially or simultaneously. In a sequential presentation one picture is shown after another at the same position, whereby later pictures replace former ones. In a simultaneous presentation all pictures are shown next to each other on a single screen. The temporal alignment of visual elements is easier in a sequential presentation because elements that are identical across the pictures are depicted at identical spatial positions (unless they change their position over time). However, to make comparisons between relevant objects the information of earlier pictures has to be memorized until later pictures are shown (Paas et al., 2007). Hence, integrating information across the pictures may be challenging for learners. In contrast, in a simultaneous presentation the depicted information remains visible on the screen and therefore comparisons among discrete steps are enabled. Moreover, in

static-simultaneous visualizations learners can regulate the pacing of their cognitive processing by deciding when to look at a picture and for how long. This all suggests that a simultaneous presentation of static pictures may be better suited to foster mental animation than a sequential one.

This assumption was confirmed by Boucheix and Schneider (2009), who found that static-simultaneous visualizations were as good for understanding a mechanical system as dynamic ones and that they outperformed static-sequential ones. This was especially true for learners with low spatial ability (but see Kim et al., 2007). For the locomotion pattern classification task used in the current study, we found a very similar result pattern, namely that dynamic visualizations outperformed static-sequential ones, whereas static-simultaneous visualizations reached the same performance as dynamic ones (Imhof, Scheiter, Gerjets, 2009). These findings suggest that dynamic visualizations may not be the only solution to convey knowledge about dynamic changes. The first part of the current study (Study 1a) focused on replicating the findings of Imhof et al. (2009) with more standardized visualizations and a broader range of classification tasks at different levels of difficulty.

Spatial Aspects of Presenting Static Pictures

When using static-simultaneous visualizations the question arises of how to arrange the static pictures on the screen to facilitate mental animation. In the study by Imhof et al. (2009) as well as in Study 1a the static pictures were represented in two rows of five pictures each. A row representation requires comparisons between different pictures to be made from left to right or vice versa. This should be advantageous for several reasons: Firstly, it corresponds to the reading order for texts (in Western cultures) and is also common for other static-simultaneous visualizations (e.g., comics). Secondly, eye tracking research has shown that irrespective of the depicted stimulus horizontal eye movements are more likely to occur than vertical ones (Tatler & Vincent, 2008). Finally, arranging multiple visualizations of an object that is moving from left to right in a row corresponds to the moving direction of this object. Taken together, a row presentation should facilitate mental animation, because it better corresponds to the nature of the depicted movement as well as to our typical viewing behavior. This may be why it is also the common presentation format for static-simultaneous visualizations used in former studies (Boucheix & Schneider, 2009; Imhof et al., 2009; Kim et al., 2007). However, it is unclear whether the static-simultaneous presentation formats used so far yield similar performance as dynamic visualizations, because the pictures remain visible all the time or because their spatial arrangement facilitates mental animation. Hence, in Study 1b we compared dynamic visualizations to three additional variants of static-simultaneous ones, namely to column, matrix, and circle presentations (Figure 1).

When depicting pictures in columns comparisons have to be made from upper to lower positioned pictures or vice versa. This spatial layout may yield the advantage that at least for pictures presented in a landscape format the distance between to-be-compared elements in two pictures is smaller. Hence, shorter saccades are required. Moreover, for the current task the elements that need to be compared to each other to determine their relative position (i.e., the fins) and thus to infer the locomotion pattern from it are vertically aligned. Hence, only few visual search processes are needed. On the other hand, this arrangement corresponds neither to the reading order nor to the objects' moving direction. In Study 1b we additionally implemented a matrices presentation of the pictures, where horizontal as well as vertical processing was needed. Finally, the circle presentation took into account that the depicted locomotion patterns are cyclic (i.e., reiterating) so that the last picture of one movement cycle automatically leads to the beginning of a new cycle without forcing the learner to skip back to the beginning of the row or column.

The question of how different spatial layouts of static-simultaneous visualizations influence their effectiveness compared to dynamic visualizations was investigated in Study 1b. If dynamic visualizations were superior to these static-simultaneous variants, this would indicate that the row presentation format used earlier is advantageous because of its specific spatial layout and not just because the pictures are permanently visible, which is also true for the other static-simultaneous variants.

The Role of Spatial Ability

In line with prior research we considered learners' spatial ability as a possible moderator of the effectiveness of dynamic and static visualizations during learning (e.g., Boucheix & Schneider, 2009; Hays, 1996). Hegarty (1992) proposed that learners' spatial ability plays a role for the process of mental animation. Moreover, Hegarty and Sims (1994) showed that high spatial ability learners outperformed low spatial ability learners in mechanical mental animation tasks. Furthermore, Hays (1996) showed that low spatial ability learners particularly benefited from learning with dynamic visualizations compared to static ones or no visualizations suggesting that these learners have fewer abilities to mentally animate the dynamics based on static pictures (Hegarty & Waller, 2005). Whereas low spatial ability learners suffer from "poor" instructions, high spatial ability may compensate for such instructions (cf. ability-as-compensator hypothesis, Mayer & Sims, 1994; see also Boucheix & Schneider, 2009). Accordingly, for the current study benefits in favour of dynamic visualizations (and potentially, static-simultaneous-rows visualizations) should be more pronounced for low rather than for high spatial ability learners.

Hypotheses

For *Study 1a*, in which we addressed the temporal aspects of static visualization formats, we assumed that dynamic visualizations would be superior to static-sequential visualizations, but not to static-simultaneous visualizations presented in rows, thereby replicating findings form earlier

studies with a broader range of recognition tasks and more standardized visualizations (see below). In *Study 1b* we tried to further disentangle temporal and spatial aspects of presenting multiple static pictures by testing whether dynamic visualizations would be superior to other static-simultaneous presentation formats. We assumed that dynamic visualizations would show stronger advantages in this case, thereby suggesting that the benefits of static-simultaneous visualizations presented in rows are not just due to temporal aspects but also due to their spatial layout.

For both studies, we assumed that higher spatial ability would be associated with better learning outcomes than lower spatial ability. Moreover, we proposed that learners with lower spatial ability would benefit stronger from learning with dynamic visualizations compared to static visualizations than those with higher spatial ability.

Study 1a

Method

Participants and Design. We randomly assigned 75 university students (average age: 24.48 years, *SD* = 4.34; 53 female) to one of three visualization conditions: dynamic vs. static-sequential vs. static-simultaneous-rows.



Figure 1: The four to-be-learned locomotion patterns (relevant movements indicated by arrows).

Materials. Participants were asked to learn how to classify fish according to their locomotion patterns based on visualizations that illustrated four different locomotion patterns. These locomotion patterns differed in terms of the used body parts that generate propulsion (i.e., the body itself or several fins) and also in the manner of how these body parts are moving (i.e. wave-like or paddle-like; cf. Figure 1). One of the major challenges in identifying these locomotion patterns is that fish may deploy a variety of other movements in addition, for instance, for navigation. These navigational movements used by a fish displaying a specific propulsion locomotion pattern can easily be confused with movements used for propulsion in another locomotion pattern.

We varied the *presentation format* of the visualizations as independent variable. Dynamic representations were compared to nine either sequentially or simultaneously (in rows) presented static visualizations.

We developed highly realistic 3D-models of fish performing the four to-be-learned locomotion patterns based on which 2D-animations were rendered that were standardized in terms of the perspective, the background and the position of the fish. These animations were used as dynamic learning materials. The static pictures were extracted from these animations by an expert and represented the key states in the movement cycles.

In the *dynamic condition* the movement cycles of the locomotion patterns were presented in loops in the animations (72 s per locomotion pattern). In the *static-sequential condition* the nine static pictures were presented twice successively for 4 s each. In the *static-simultaneous-rows condition* the same pictures were presented in parallel for 72 s. They were arranged in two rows corresponding to the two phases of the locomotion patterns (cf. Figure 2, upper left part). To facilitate the transition from the first to the second row, the fifth picture was depicted twice, once as the last picture of the upper row and once as the first picture of the lower row. The pictures' size was half of the size of the dynamic and the static-sequential conditions. There was no need for the subjects to scroll the page.

During learning the participants saw visualizations for each of the four to-be-learned locomotion patterns in a predefined order. The presentation was system-controlled and accompanied by narration. The narration explained the locomotion pattern in terms of typical fish using this locomotion pattern, body parts involved, kind of movements executed (undulation versus oscillation), parameters of the movements (e.g., amplitude), and maximum velocity.

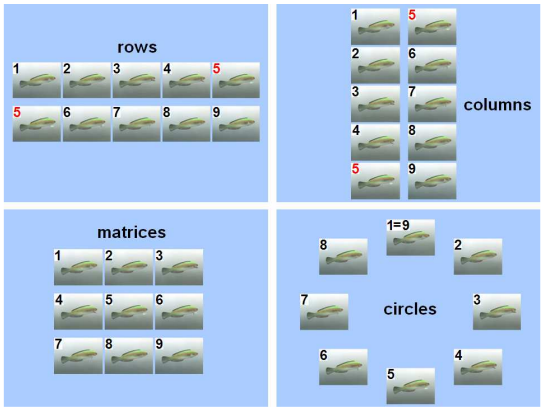


Figure 2: Static-simultaneous presentation formats.

Measures. Learners' spatial abilities were assessed with two different tests, namely the mental rotation test (MRT, Vandenberg & Kuse, 1978), and a short version of the paper folding test (PFT, Ekstrom et al., 1976). Both spatial ability measures were used in the analyses as continuous factors.

To assess learning outcomes a locomotion pattern recognition test consisting of pictorial multiple-choice items was administered. Underwater videos of real fish performing one of the four locomotion patterns were used as test stimuli. The number of test items was constrained by a number of aspects (e.g., resolution, visibility of the fish from a certain perspective, clear depiction of the respective locomotion patterns). For each of the four locomotion patterns seven videos were identified. To choose for each item the kind of locomotion pattern that was depicted, learners had to identify the body parts relevant for

propulsion and their way of moving. Possible answers were the correct terms of the four locomotion patterns and the additional answer "I don't know" (see Figure 3). Each item was awarded one point for the correct answer (max. 28 points). The recognition test items were categorized by two independent domain experts into items with low, intermediate, and high task difficulty. Their decisions were based on the visibility of the relevant parts used for propulsion as well as on the absence or presence of miscellaneous movements of the fish's body parts that could have been mistaken as being relevant for propulsion (e.g., movements only necessary for navigational purposes). Videos that showed the pattern relevant for propulsion continuously and contained no other movements were assigned a low task difficulty (8 items). Videos that showed the pattern relevant for propulsion continuously, but contained movements similar to another locomotion pattern were assigned an intermediate task difficulty (11 items). Videos that either showed the pattern relevant for propulsion continuously, but contained additional movements similar to at least two other locomotion patterns or videos that did not show the relevant propulsion pattern continuously or that did show it in a non-salient manner (whereby all of these videos contained movements similar to at least one other locomotion pattern) were assigned a high task difficulty (9 items). Five cases of disagreement between the two experts were resolved by negotiation.

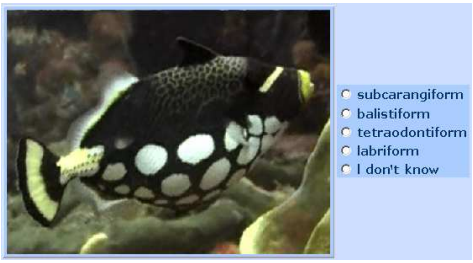


Figure 3: Screenshot of a recognition test example item.

Procedure. After completing paper-based the MRT, PFT, and a demographic questionnaire, participants read an introduction, which was followed by the computer-based learning phase. Finally, learners worked on the computer-based pictorial recognition test.

Results

Performance in the three recognition subtests was analyzed by a MANCOVA with presentation format (dynamic vs. static-sequential vs. static-simultaneous-rows), the MRT, and the PFT as independent variables (Table 1).

There was an overall effect for presentation format (F = 2.28, p = .04) and for the PFT (F = 3.62, p = .02), but no other main effect or interactions. There was an effect for presentation format only for recognition tasks with an intermediate difficulty (F = 4.00, p = .02). Dynamic visualizations were superior to static-sequential visualizations, but not to static-simultaneous-rows visualizations. Higher performance in the PFT was associated with better recognition for tasks with low (F = 7.52, p < .01) and intermediate difficulty (F = 9.18, p < .01).

Table 1: Adjusted means (and standard errors) for recognition performance (in % correct) as a function of presentation format and task difficulty (Study 1a).

	Presentation Format		
Task Difficulty	dynamic (n = 25)	static-sequential (n = 25)	static-simultaneous-rows (n = 25)
low	92.65 (3.90)	84.58 (3.88)	86.43 (3.93)
intermediate	87.83 (4.33)	71.85 (4.30)	74.30 (4.36)
high	71.80 (4.57)	72.67 (4.55)	74.36 (4.61)

Discussion of Study 1a

The results confirmed that dynamic visualizations are better suited to convey knowledge about the continuity of locomotion patterns compared to static-sequential visualizations, but not to static-simultaneous visualizations presented in rows – at least for recognition tasks with an intermediate difficulty level. These findings hence replicate those of a former study, where digital underwater videos as well as black-and-white animated line drawings were used as dynamic visualizations (Imhof et al., 2009). Hence, the results obtained by Imhof et al. were not an artefact of either low visibility of important kinematical aspects in the underwater videos or their potentially oversimplified representation in the animated line drawings, because the visualizations in the current study were of high quality in terms of the visibility and fidelity of important features.

In sum, the results suggest that dynamic visualizations as well as static-simultaneous-rows presentations allow for the construction of an adequate mental representation of kinematics; however, it is yet not clear whether the relative good performance of the latter condition is due to its temporal (permanent visibility) or its spatial aspects (rows), which is why Study 1b was conducted.

Study 1b

Method

Participants and Design. We randomly assigned 75 university students (average age: 23.35 years, *SD* = 3.71, 57 female) to three static-simultaneous conditions, namely a static-simultaneous-columns, a static-simultaneous-matrices, and a static-simultaneous-circles condition, to compare their performance to that of students in the dynamic visualization condition of Study 1a.

Materials. The learning domain, the measures as well as the procedure were identical to Study 1a. In the *static-simultaneous-columns condition* the single pictures were

arranged in two columns corresponding to the two phases of the locomotion patterns (cf. Figure 2, upper right part). To facilitate the transition between the left and the right column the fifth picture was depicted twice, once as the last picture of the left column and once as the first picture of the right column. In the *static-simultaneous-matrices condition* the nine pictures were presented in 3x3 matrices, ordered primarily from left to right and secondarily from top to bottom (cf. Figure 2, lower left part). Contrary to the static-simultaneous-rows and the static-simultaneous-columns condition no pictures were depicted twice. In the *static-simultaneous-circles condition* the single pictures were presented in a clockwise arrangement with the first picture at the 12 o'clock position (cf. Figure 2, lower right part). In this condition the ninth picture was not presented, because it depicted the same state in the locomotion pattern as the first picture. The pictures in all conditions had the same size as those in the static-simultaneous-rows condition in Study 1a.

Results

Performance in the three recognition subtests was analyzed by a MANCOVA with presentation format (static-simultaneous-columns vs. static-simultaneous-matrices vs. static-simultaneous-circles vs. dynamic), the MRT, and the PFT as independent variables (Table 2).

There was an overall effect for presentation format (F = 2.64, p = .01), for the MRT (F = 4.93, p < .01) and for the PFT (F = 2.82, p = .04), but no interactions. There was an effect for presentation format for recognition tasks with low (F = 4.01, p = .01) and intermediate difficulty (F = 6.41, p = .001). Dynamic visualizations led to better recognition for tasks with low difficulty compared to the static-simultaneous-matrices visualizations as well as for tasks with intermediate difficulty compared to all three static-simultaneous conditions. Moreover, higher performance in the MRT was associated with better recognition performance for tasks with low (F = 4.55, p = .04) and intermediate difficulty (F = 14.59, p < .001). Furthermore, higher performance in the PFT was associated with better recognition for tasks with low difficulty (F = 4.63, p = .03).

Discussion of Study 1b

None of the additionally tested spatial layouts of the static-simultaneous visualizations achieved the same recognition performance as the dynamic visualizations for tasks with an intermediate level of difficulty. For recognition tasks with a low level of difficulty we found dynamic visualizations to be superior to static-simultaneous visualizations presented as matrices, showing that this presentation format bears the fewest of all advantages for the task at hand.

The possible advantage of a circular presentation that it adequately represents the cyclic nature of the locomotion patterns might have been cancelled out by the fact that with this presentation format the orientation of the pictures interfered with the swimming direction of the fish. That is, for pictures presented in-between the 3 o'clock and the 9 o'clock position, the next picture is depicted to the left of its previous picture, whereas the swimming direction of the fish still indicates a movement from left to right. Moreover, contrary to the assumption that the spatial contiguity in a column supports the visual alignment of to-be-compared elements and hence might facilitate mental animation, this condition was not any better than the dynamic condition.

In sum, the results suggest that dynamic visualizations are superior to different static-simultaneous presentation formats as long as the spatial layout of the static pictures does not support mental animation processes in a way that corresponds to our reading/viewing behavior and that is in line with the moving direction of the depicted object.

Table 2: Adjusted means (and standard errors) for recognition performance (in % correct) as a function of presentation format and task difficulty (Study 1b).

	Presentation Format			
	static-simultaneous			dynamic
Task Difficulty	columns (n = 25)	matrices (n = 25)	circles (n = 25)	(n = 25)
low	83.85 (4.13)	72.40 (4.47)	79.21 (4.07)	92.78 (4.36)
intermediate	70.26 (4.20)	63.65 (4.55)	66.90 (4.14)	88.36 (4.43)
high	66.69 (4.76)	62.58 (5.16)	61.52 (4.70)	71.77 (5.02)

General Discussion

The superiority of dynamic visualizations over most static presentation formats for learning tasks that explicitly require the identification of the continuity of movements and involve a strong perceptual component was supported in Studies 1a and 1b. However, consistent with prior findings (Boucheix & Schneider, 2009; Imhof et al., 2009) a static-simultaneous presentation of multiple pictures in rows led to the same performance as the dynamic visualizations. Accordingly, for this specific case where the moving direction of the depicted object and the spatial layout of the pictures correspond to each other, learners seem to be well able to mentally animate the sequence of pictures and hence to infer the kinematics from it (Hegarty, 1992). However, this result pattern holds true only for tasks of intermediate difficulty. The fact that we did not find the same results for tasks of low difficulty can be explained in terms of a ceiling effect. The items are maybe so clearly identifiable that learners from all experimental conditions (except for the matrices condition in Study 1b) achieved very good results. According to the expert opinions there were always at least two concurring patterns visible in items with high task difficulty. Which one of these is used for propulsion cannot be answered only on the basis of perceptual input. Rather conceptual knowledge acquired from the spoken explanations, which were identical in all experimental conditions, had to be used to answer these items. Additional design techniques like cueing (De Koning et al., 2009) or enriching static displays (Münzer, Seufert, & Brünken,

2009) could further enhance the effectiveness of static-simultaneous presentation formats.

Astonishingly, there was no moderating effect of spatial ability concerning the effectiveness of different presentation formats of visualizations. Therefore, the assumed ability-as-compensator hypothesis could not be confirmed. In further studies this issue should be addressed in more detail, because there is an ongoing discussion about the separate components that make up the construct spatial ability (for an overview see Hegarty & Waller, 2005). Especially, the dynamic spatial ability component might be a relevant dimension for mental animation in dynamic tasks (D'Oliveira, 2004; Hunt et al., 1988). Hence, it might be that the tests used here may not have addressed those spatial ability components that might be most relevant to mental animation, even tough they are commonly used in visualization research. Despite of these doubts concerning the validity of the measures used, we were nevertheless able to show that irrespective of visualization format higher spatial ability was associated with better learning outcomes than lower spatial ability for tasks with low and intermediate difficulty, thereby replicating the findings of Hegarty and Sims (1994). Hence, we can at least conclude that spatial abilities are relevant to the task at hand. Nevertheless, further studies need to address the question of how mental animation from static-simultaneous visualizations supports learning.

Acknowledgments

The study is part of a research project on the "resource-adaptive design of visualizations for supporting the comprehension of complex dynamics in the Natural Sciences" funded by the Leibniz-Gemeinschaft.

References

Ayres, P., Marcus, N., Chan, C., & Qian, N. (2009). Learning hand manipulative tasks: When instructional animations are superior to equivalent static representations. *Computers in Human Behavior, 25*, 348-353.

Bétrancourt, M., & Tversky, B. (2000). Effect of computer animation on users' performance: A review. *Le Travail Humain, 63*, 311-329.

Boucheix, J.-M., & Schneider, E. (2009). Static and animated presentations in learning dynamic mechanical systems. *Learning and Instruction, 19*, 112-127.

Byrne, M. D., Catrambone, R., & Stasko, J. T. (1999). Evaluating animations as student aids in learning computer algorithms. *Computers and Education, 33*, 253-278.

De Koning, B. B., Tabbers, H. K., Rikers, R. M. J. P., & Paas, F (2009). Towards a framework for attention cueing in instructional animations: Guidelines for research and design. *Educational Psychology Review, 21*, 113-140.

D'Oliveira, T. C. (2004). Dynamic spatial ability: An exploratory analysis and a confirmatory study. *International Journal of Aviation Psychology, 14*, 19-38.

Ekstrom, R., French, J., Harmon, H., & Derman, D. (1976). *Manual for Kit of Factor-Referenced Cognitive Tests*. Princeton: Educational Testing Service.

Hays, T. A. (1996). Spatial ability and the effects of computer animation on short-term and long-term comprehension. *Journal of Educational Computing Research, 14*, 139-155.

Hegarty, M. (1992). Mental animation: Inferring motion from static diagrams of mechanical systems. *Journal of Experimental Psychology: Learning, Memory and Cognition, 18*, 1084-1102.

Hegarty, M., & Sims, V. K. (1994). Individual differences in mental animation during mechanical reasoning. *Memory and Cognition, 22*, 411-430.

Hegarty, M., & Waller, D. (2005). Individual differences in spatial ability. In P. Shah, & A. Miyake (Eds.), *Handbook of Visuospatial Thinking*. Cambridge University Press.

Hunt, E., Pellegrino, J. W., Frick, R. W., Farr, S. A., & Alderton, D. (1988). The ability to reason about movement in the visual field. *Intelligence, 12*, 77-100.

Imhof, B., Scheiter, K., & Gerjets, P. (2009). Realism in dynamic, static-sequential, and static-simultaneous visualizations during knowledge acquisition on locomotion patterns. In N. A. Taatgen, & H. van Rijn (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 2962-2967). Austin, TX: Cognitive Science Society.

Kim, S., Yoon, M., Whang, S., Tversky, B., & Morrison, J. (2007). The effect of animation on comprehension and interest. *Journal of Computer Assisted Learning, 23*, 260-270.

Mayer, R. E., & Sims, V. K. (1994). For whom is a picture worth a thousand words? Extensions of a dual-coding theory of multimedia learning. *Journal of Educational Psychology, 86*, 389-401.

Münzer, S., Seufert, T., & Brünken, R. (2009). Learning from multimedia presentations: Facilitation function of animations and spatial abilities. *Learning and Individual Differences, 19*, 481-485.

Paas, F., Van Gerven, P. W. M., & Wouters, P. (2007). Instructional efficiency of animation: Effects of interactivity through mental reconstruction of static key frames. *Applied Cognitive Psychology, 21*, 783-793.

Tatler, B. W., & Vincent, B. T. (2008). Systematic tendencies in scene viewing. *Journal of Eye Movement Research, 2*, 1-18.

Tversky, B., Bauer-Morrison, J., & Bétrancourt, M. (2002). Animation: Can it facilitate? *International Journal of Human-Computer Studies, 57*, 247-262.

Vandenberg, S. G., & Kuse, A. R. (1978). Mental rotations, a group test of three-dimensional spatial visualization. *Perceptual and Motor Skills, 47*, 599-604.

Wong, A., Marcus, N., Smith, L., Cooper, G. A., Ayres, P., Paas, F., & Sweller, J. (2009). Instructional animations can be superior to statics when learning human motor skills. *Computers in Human Behavior, 25*, 339-347.

Working Memory Constraints on Multiple Center-Embedding

Fred Karlsson (fgk@ling.helsinki.fi)
Department of Modern Languages, PO Box 24
FI-00014 University of Helsinki, Finland

Abstract

Gibson's (1998) theory on the locality of syntactic dependencies claims that multiply center-embedded clauses are unacceptable if they contain a parse-state with at least two long unresolved predicted categories in addition to the top-level verb. 'Long unresolved' means a syntactic prediction spanning at least three intervening new discourse referents. This claim was based on experimental analysis of invented examples. Karlsson (2007b) provided corpus data demonstrating that, contrary to widely accepted views in linguistics and cognitive science, there are well-defined constraints on how many (maximally three) and what types of multiple center-embeddings occur in spoken and written discourse in natural languages. Gibson's theory of the processing of multiple center-embeddings will be evaluated in the light of Karlsson's empirical data. The corpus data do not support the idea of a discrete limit on working memory capacity, because almost one third of the extant examples of multiple center-embedding are more complex than Gibson's acceptability limit stipulates. Spoken language processing complexity is clearly below Gibson's limit, written language is capable of transgressing it.

Keywords: center-embedding; clausal embedding; cognitive explanation; complexity; embedding; multiple center-embedding; recursion; syntactic complexity.

Definition of Center-Embedding

The notion EMBEDDING refers to all types of clauses occurring as subordinate parts of their superordinate clauses (which themselves may be either main or subordinate). The starting point will be the classical view of subordination as expounded in Quirk et al. (1989, Chapter 14). Typical finite sub-clauses are of three types: complement, relative, and adverbial. They are indicated by subordinators or relative pronouns, henceforth called sub/wh-elements.

CENTER-EMBEDDED clauses have words of the superordinate clause both to their left (excluding subordinators and coordinators) and to their right, as the relative clauses in (1, 2) and the *when*-clause in (3). SELF-EMBEDDING is multiple center-embedding invoking two or more clauses of the same type, e.g. two relative clauses as in (4). In the examples, the gross clausal structure is indicated by angular brackets prefixed by the character 'C' for center-embedding and an integer indicating dept of embedding.

(1) Others [$_{C\text{-}1}$ who are attracted to this Mecca of the beat generation] are heroin addicts and small hoodlums. (Brown Corpus)
(2) Another frequent pioneer difficulty, [$_{C\text{-}1}$ caused by wearing rough and heavy shoes and booths,] was corns (Brown).
(3) On March 13, [$_{C\text{-}1}$ when he preached a sermon on the text,] he told his congregation how disappointed he was (Brown).
(4) For an analysis of the possible modifications [$_{C\text{-}1}$ of which the pathological termination of an act [$_{C\text{-}2}$ which is not according to law] are susceptible] we have therefore ... (Jeremy Bentham)

When a sentence contains multiple embeddings of the same type, e.g. two center-embeddings as in (4), the DEGREE OF EMBEDDING is equal to the number of embeddings and occasionally indicated by the character 'C' superfixed with the degree. Thus, (4) is an instance of C^2, double center-embedding.

A clause embedded after the initial subordinating (or coordinating) conjunction of the superordinate clause is not center-embedded but initially-embedded, e.g. the I-2-clauses in (5, 6):

(5) [$_{I\text{-}1}$ If [$_{I\text{-}2}$ what is tantamount to dictatorship ...] continues in a union] it can ... (Lancaster-Oslo/Bergen Corpus = LOB))
(6) c. [$_{I\text{-}1}$ If [$_{I\text{-}2}$ when I'm 38] Metallica ends] I don't think ...] (British National Corpus = BNC)

Here, the subordinating conjunctions of the respective I-1-clauses are not fully integrated syntactic constituents in their clauses and therefore a further clause embedded after them is not center-embedded but initially-embedded. The superordinate clause material preceding a center-embedding must consist of full syntactic constituents, as in (1-4).

Empirical data on multiple center-embedding

By systematically searching the Brown and LOB corpora, by checking the extant scarce empirically-minded literature on multiple center-embedding, by consulting more than 100 older grammars, style manuals and philological studies especially of older forms of German and Latin, and furthermore by manually analyzing 6000 sentences by three 19th century scholars known for their intricate and syntactically complex language use (Jeremy Bentham, John Stuart Mill, C. S. Peirce), Karlsson (2007) established a data pool of 13 triple center-embeddings, C^3, and 104 double center-embeddings, C^2. As every C^3 contains two C^2s, the total number of C^2s is 130. The languages concerned were English, German, Latin, Swedish, Finnish, French and Danish, from Antiquity to the 21st century.

Here are three of the C^3s observed:

(7) In an excellent article ... Salvini draws a parallel between the way [$_{C\text{-}1}$ in which the spoken Latin of the men [$_{C\text{-}2}$ with whom Gregory of Tours, [$_{C\text{-}3}$ whom he has no reason to mention,] must have mixed] eventually became Old French ...,] and the comparable direct development of pre-Romanesque painting ... (L. Thorpe, *Gregory of Tours: The History of the Franks*, 1974; due to Geoffrey Sampson)

(8) The Prime Minister [$_{C\text{-}1}$ who at the height of the crisis had snapped to a junior minister [$_{C\text{-}2}$ who, [$_{C\text{-}3}$ not having seen him for some time,] had approached him in a Westminster corridor with a view to wishing him luck …,] 'If you want to resign, put it in writing',] was unlikely to ... (Patrick Cosgrave 1979; De Roeck et al., 1982)

(9) A person [$_{C\text{-}1}$ who, [$_{C\text{-}2}$ when riding a cycle, [$_{C\text{-}3}$ not being a motor vehicle,] on a road or other public place,] is unfit to ride through drinks or drugs,] shall be guilty of an offence. (*British Road Traffic Act*, 1972; Hiltunen, 1984)

Here are four C^2s ('F' = finally-embedded clause; '&' coordinated clause):

(10) And yet a widow, [$_{C\text{-}1}$ whose pension, [$_{C\text{-}2}$ for which her husband paid,] is wiped out [$_{F\text{-}2}$ because she works for a living wage,]] will now ... (LOB)

(11) At one point in the game [$_{C\text{-}1}$ when the skinny old man in suspenders [$_{C\text{-}2}$ who was acting as umpire] got in the way of a thrown ball] [$_{\&C\text{-}1}$ and took it painfully in the kidneys,] he lay there ... (Brown)

(12) ... the girl ... [$_{C\text{-}1}$ who was clothed in the tightest-fitting pair of slacks [$_{C\text{-}2}$ I had ever seen on a woman] and a sweater [$_{F\text{-}2}$ that showed everything [$_{F\text{-}3}$ there was]]] wanted to be sociable.] (Brown)

(13) But the idea [$_{C\text{-}1}$ that the fact [$_{C\text{-}2}$ that some pain is heading my way] gives me no special reason to avoid it] seems so at odds with ...] (Internet)

On the basis of the material collected, Karlsson (2007) induced the following generalizations:

(14) The maximal degree of multiple center-embedding is three in written language, but C^3 is so rare as to be practically non-existing (only 13 instances found).

(15) The maximal degree of multiple center-embedding is two in spoken language, but it is so rare as to practically non-existing (only three instances found).

(16) Only clauses that postmodify nouns (i.e. relative clauses as in (7, 8, 10 12), complement *that*-clauses as in (13), and indirect questions allow central self-embedding.

(17) A C^2 must contain at least one postmodifying clause.

(18) The typical C^3/C^2 contains a pair of relative clauses, and is located at the end of the grammatical subject immediately before the main verb. Its main function is to aid in the specification of the topic of the sentence.

(19) A/the lower clause in a multiple center-embedding must contain at least one overt pronoun, preferably as grammatical subject.

(20) Direct objects must not be multiply relativized in C^2s or C^3s.

In practice, constraint (15) rules out multiple center-embedding in spoken language. This means that genuine rested syntactic recursion under no circumstances can be considered an important design feature of natural language syntax.

Constraint (18) is explicable by the fact that the S-V junction is the major natural syntactic break in SVO-languages, between the topic (the grammatical subject) and the comment.

Constraint (20) rules out the classical sentence (21), even if (21) is in conformance with constraint (14). Sentence (21) has often been used in the literature as supposed proof of the absence of constraints on the degree of multiple center-embedding.

(21) The rat$_j$ [$_{C\text{-}1}$ the cat$_k$ [$_{C\text{-}2}$ the dog$_m$ chased _$_k$] killed _$_j$] ate the malt.

Not a single genuine example of double object relativization was found in Karlsson's (2007) corpus, nor are there any in the literature known to me. (In (21), the traces of the preposed objects are indicated.)

Gibson's processing theory

Edward Gibson's (1998) Syntactic Prediction Locality Theory (SPLT) has been influential in accounting for the relationship between the sentence processing mechanism and the available (mental) computational resources.

The theory has two components: an INTEGRATION COST component and a component for the MEMORY COST associated with keeping track of obligatory syntactic requirements. The type of memory concerned is, of course, working memory (WM). WM cost is quantified in terms of the number of syntactic categories that are necessary to complete the current input as a grammatical sentence. Both memory cost and integration cost depend on LOCALITY. The longer a predicted category must be kept in WM before the prediction is satisfied, the greater is the memory cost for maintaining that prediction. The greater the distance is between an incoming word and the most local head or dependent to which it attaches, the greater the integration cost.

When a syntactic prediction has been made, a WM cost of one memory unit (MU) is taxed every time a new discourse referent is encountered until the prediction is satisfied. The operational definition of 'new discourse referent' is either introduction of a referent not so far mentioned, or a tensed verb. Because several syntactic predictions may be active

simultaneously, several WM taxation counts may be running simultaneously, consuming WM resources. The main verb is assumed to be cost-free because its existence is taken for granted in every sentence.

SPLT explains the processing difficulties associated with an impressive number of difficult structures, including the unacceptability (I would say: ungrammaticality) of multiple center-embeddings such as the double object relativization (21).

To illustrate, first consider Gibson's examples (23, 24)

(22) The reporter [$_{C\text{-}1}$ who attacked the senator] admitted the error.
(23) The reporter [$_{C\text{-}1}$ who the senator attacked] admitted the error.

In the subject-relativized sentence (22), *who* predicts the occurrence of a predicate and a pronoun gap in the relative clause. When the next word *attacked* is encountered, both predictions are satisfied and no costs incur. *Attacked* next predicts an object but this occurs as the next constituent and therefore no costs incur for this prediction either. The total WM taxation for (22) is therefore 0 MUs.

Next, consider the object relativization in (23). *Who* makes the same predictions as in (22), but the next constituent is the new referent expressed by *the senator*, whereupon both predictions incur a cost of 1 MU, totaling 2M(1) (two predictions having passed one new referent). Next *attacked* resolves the pending predictions and the analysis proceeds as in (22). Thus the correct analysis is made that object relativization is more complex by consuming more WM resources than subject relativization, a fact well established in psycholinguistics.

Now consider Gibson's (made-up) equivalent of (21), sentence (24) with double object relativization:

(24) The administrator [$_{C\text{-}1}$ who the **intern** [$_{C\text{-}2}$ who the **nurse supervised]** had bothered] lost the medical reports.

The syntactic predictions (i.e. the predicate and the relative pronoun gap) of the first *who* will not be satisfied until *had bothered* is encountered, yielding a WM expenditure of 2M(3) MUs, the relevant three crossed new referents pending in WM storage being *intern*, *nurse* and *supervised* (bolded in (24)).

Gibson infers the following generalization "…structures which include a parse state with at least two long unresolved predicted categories in addition to the top-level verb are unacceptable, and those without such a state are usually acceptable. Under the memory cost function assumed here, a 'long' unresolved prediction is one spanning at least three intervening new discourse referents. Thus, sentences whose parses include parse states whose memory cost is 2M(3) MUs or greater are generally not acceptable, while sentences whose parses do not include such a costly parse state are generally acceptable. A reasonable conclusion from this analysis is that linguistic working memory capacity is somewhere around 2M(3) MUs or just below".

Gibson based his analysis on a handful of invented examples. The rest of this paper evaluates Gibson's ACCEPTABILITY LIMIT in the light of my empirical data on multiple center-embeddings. A characterization is also offered of the overall processing complexity of these constructions.

Triple center-embeddings

Of the thirteen observed triple center-embeddings, only one, (25), is clearly below the acceptability limit, by Gibson defined as 2M(3) MUs. (25) consumes only 1M(3) for satisfaction of the prediction at *weil* that C-1 needs a predicate. This prediction is satisfied at the word *verzichtet*, having crossed three new discourse referents, one in each of the embedded clauses (*Mitbewerber*, *angenommen*, *überlegen*). At *angenommen wird* there is a parse state with a cost of 1M(2)+2M(1) MUs, also clearly below the acceptability limit. (Note the use of the plus notation to indicate the sum of differing simultaneous prediction costs.) The consumption of WM resources is low in (25) because all three embedded clauses are short, C-2 contains two pronouns, and C-3 an impersonal passive construction which disposes of its grammatical subject.

(25) Er hat den Preis nur, [$_{C\text{-}1}$ weil ein Mitbewerber, [$_{C\text{-}2}$ welcher ihm, [$_{C\text{-}3}$ wie allgemein angenommen wird,] überlegen ist,] verzichtet hat,] bekommen. (Literal gloss: 'He has the price only, (C-1) because a competitor, (C-2) who over him, (C-3) as is generally presumed, (C-2) is superior, (C-3) gave up, (Main) got.') (Blatz 1896: 1274)

There are two C^3s reaching a maximum of 1M(4), with all prior parse states < 2M(3). Gibson does not discuss instances where only one prediction crosses more than three referents. Assuming for the moment that the effort invested in one syntactic prediction would be equal in WM cost to that of crossing one new referent, we obtain the value 6 for the TOTAL EFFORT invested at Gibson's acceptability limit (2 syntactic predictions * 3 referent crossings = 6). We shall assume that all multiple center-embeddings with a maximal total effort smaller than 6 are below the acceptability limit, in particular 1M(4), 1M(5) and 3M(1), all of which exist, provided they have no prior parse state exceeding 2M(3). Thus three C^3s out of 13 are clearly below the limit, when redefined in terms of total effort = 6.

Sentence (9) is exactly at (or, according to Gibson, perhaps slightly above) the acceptability limit 2M(3), which is reached at the verb *is* in C-1, after crossing of the three referents *riding*, *cycle*, and *road*. In C-3 neither the bleak copula nor the classificatory NP *motor vehicle* were included in the count because C-3 expresses a property, not an independent referent. Note the non-finitenesss of the verbs in C-2 and C-3 and the consequent suppression of two

grammatical subjects by sharing them with the upper clause. There is one more C^3 in the corpus at 2M(3) with an additional parse state at 1M(4):

(26) Der Landvogt ... fand, [$_{C\text{-}1}$ als er, [$_{C\text{-}2}$ von dem, [$_{C\text{-}3}$ was vorgefallen,] benachrichtigt,] in bestürzten Märschen zurückkehrte,] die Stadt in allgemeinen Aufruhr. ('The governor found, (C-1) as he, (C-2) about that, (C-3) which [had] happened, (C-2) notified, (C-1) returned in fast march, (Main) the city in general uproar'. (H. von Kleist, Michael Kohlhaas; Hoffmann-Krayer 1925: 131)

Sentence (8) above requires maximally 1M(6) MUs and is on the same level of processing complexity as those needing 2M(3) MUs when analyzed in terms of total effort.

The remaining seven C^3s are further beyond the acceptability limit. Sentence (7) above and one more claim exactly 2M(4) MUs, one claims 2M(4) with a later parse stage of 1M(5). These sentences were produced by well-known writers and do not intuitively feel (much) more complex than (9, 26), suggesting that 2M(3) MUs is just one point on a more continuous slope of decreasing acceptability. Still more convoluted is the following sentence from von Kleist:

(27) Der Ritter von Malzahn, [$_{C\text{-}1}$ dem der Junker sich als einen Fremden, [$_{C\text{-}2}$ der bei seiner Durchreise den seltsamen Mann, [$_{C\text{-}3}$ den er mit sich führe,] in Augenschein zu nehmen wünschte,] vorstellte,] nötigte ihn ... ('(M) The rider from Malzahn, (C-1) to whom the Junker himself as a stranger, (C-2) who upon his journey (through) the strange man, (C-3) whom he brought with himself, (C-2) to judge by appearance wanted, (C-1) introduced, (Main) forced … (H. von Kleist, *Michael Kohlhaas*; Schneider 1959: 469)

The more verbose the embedded clauses are, and the more full (non-pronominal) constituents they contain, the greater will be the WM expenditure as new referents are crossed. (27) has a parse state peak at *in Augenschein* requiring 1M(5)+2M(4) MUs, where the prediction of the predicate in C-1 (*vorstellte*) has crossed (at least) five referents (it is not always clear what should be counted as a referent, what not), and the predictions of a predicate and subject relative in C-2 have consumed 2M(4) MUs. When the prediction of *vorstellte* in C-1 finally is satisfied, its parse state has risen to 1M(6): the referents crossed are *Junker – Durchreise – Mann – führe – Augenschein – nehmen.*

The three most complex C^3s in my corpus are a Swedish one from 1863 reaching 2M(6), a German one from 1893 peaking at 2M(6)+2M(4) with a later local maximum at 2M(7), and a Danish sentence from a court decision in 1892 containing a maximum of 2M(6)+2M(4) with a later local peak 1M(11), cf. examples (3, 12, 13) in Karlsson (2007b). Such monster sentences are of course incomprehensible.

The conclusion of the analysis of C^3s must be that few of them, only three, are below Gibson's acceptability limit. If the limit reflects a foundational WM restriction, this corroborates the marginality of C^3s as a structural option, already expressed in (14). But the extant C^3s rather seem to populate a gradual slope, where the value 2M(3) MUs does not stand out as being of particular significance.

Double center-embeddings

The 104 C^2s in my corpus distribute themselves over WM cost as shown in Table 1. Columns 1a-b give the WM costs and numbers of those C^2s that clearly are below Gibson's limit 2M(3) MUs (total effort less than 6). Columns 3a-b lists the instances which are above the acceptability limit with a total effort equal to or greater than 6.

Table 1: Working memory cost in 104 C^2s.

1a	1b		2a	2b	Total
Cost	**N**		**Cost**	**N**	
M(0)	1		1M(6)	4	
1M(1)	9		1M(7)	1	
1M(2)	12		1M(8)	2	
1M(3)	10		2M(3)	11	
1M(4)	5		2M(4)	6	
1M(5)	3		2M(5)	1	
2M(1)	13		2M(6)	1	
2M(2)	17		2M(9)	1	
3M(1)	4		3M(2)	1	
			3M(3)	2	
Sum	74			30	104
%	71			29	100

More than two thirds of the C^2s are below the limit and many of them are far from causing overflow in working memory. Here is an assortment, listed according to growing complexity, of those C^2s that are easiest to process and understand, with the WM cost indicated at the end in angular brackets.

(28) The girl ... [$_{C\text{-}1}$ who was clothed in the tightest-fitting pair of slacks [$_{C\text{-}2}$ I had ever seen on a woman] and a sweater [$_{F\text{-}2}$ that showed everything [$_{F\text{-}3}$ there was]]] wanted to be sociable. [M(0)]

(29) We yet looked forward to a time ... [$_{F\text{-}1}$ when the rule [$_{C\text{-}2}$ that they [$_{C\text{-}3}$ who do not work] shall not eat,] will be applied not to paupers only. [1M(1)]

(30) It's ironic [$_{F\text{-}1}$ that I'm here, [$_{F\text{-}2}$ where the man [$_{C\text{-}3}$ the trophy [$_{C\text{-}4}$ I won] is named after] coached. [1M(2)]

(31) The reason [$_{C\text{-}1}$ why this question of [$_{C\text{-}2}$ when the copy was made] is of some interest] is that ... [1M(3)]

(32) He knows ... [$_{F\text{-}1}$ that, for example, [$_{C\text{-}2}$ whereas in 1908 the proportion of his students at Leeds [$_{C\text{-}3}$ who were drawn from within 30 miles] was 78 %,] it was, by 1955, reduced to 40 %. [1M(4)]

(33) Laughland's assertion [$_{C-1}$ that the presence of Delors – 14 years old [$_{C-2}$ when the war began –] in the Compagnons de France, the Vichy youth movement, meant [$_{F-2}$ that he supported fascism]] is ridiculous. [1M(5)]

(34) The two most difficult skills [$_{C-1}$ that everyone [$_{C-2}$ I know] has to learn when they join a team] are… [2M(1)]

(35) All the concern [$_{C-1}$ which he [$_{C-2}$ to whom it belongs by adoption] has in the matter] is the being ... [2M(2)]

(36) But the general principle [$_{C-1}$ that every thing [$_{C-2}$ to which such and such sensation belongs,] has such and such a complicated series of predicates,] is not one determined by reason but… [3M(1)]

In contradistinction to C^3, C^2 is obviously a well-established even if rare construction type especially in written language: there is no question of the grammaticality and acceptability of (28-36) even if it is clear that overall acceptability has a tendency to decrease as the number of constituents pending in WM and the number of new referents crossed increases.

Note that there even are C^2s like (28) that invoke no WM cost at all. This situation is possible in (28) because the subject and predicate of C-1 (*who was clothed*) are immediately available and therefore the do not need to be entered as pending predictions in WM. The predicate of C-1 predicts the occurrence of an adverbial prepositional phrase, but it too (*in the tightest-fitting pair of slacks*) is completed immediately, as the first part of a coordinated construction. C-2 is beneficially inserted before the second, optional part of the coordinated construction in C-1 and therefore does not tax WM at all. – (The C-3 of (7) is also inserted at a coordination junction, corroborating conclusion (18) that multiple center-embedding is preferred at natural syntactic breaks.)

The following sentences exemplify those 30 sentences (29 %) of Table 1 that are at or beyond the acceptability limit, consuming 1M(6) or 2M(3) MUs or more. The examples are listed according to growing complexity.

(37) For the remainder of his industrious life (apart from during the second world war [$_{C-1}$ when he worked in the Ministry of Information [$_{C-2}$ – where he was banished to Belfast [$_{F-3}$ for being lazy and unenthusiastic –]] and the Auxiliary Fire Service)] Quennell ... [1M(6)]

(38) And in particular [$_{C-1}$ when the motives [$_{C-2}$ which are applied] are of the nature of those [$_{F-2}$ which result from a change [$_{f-4}$ made in the condition of the body,]]]] the power may be said to … [1M(8)]

(39) Neither, however, [$_{C-1}$ as their critics and all of those [$_{C-2}$ who subsequently complained about their assault on Heath] always stress,] felt moved to resign. [2M(3)]

(40) The occasion [$_{C-1}$ on which in the nation [$_{C-2}$ of whose language I am writing] the word *repugnancy* has been most frequently made use of] is that where ... [2M(4)]

(41) A number of speeches [$_{C-1}$ into which a great deal of thought and preparation on a level a great deal higher [$_{C-2}$ than is common in modern politics] have gone] are not reported at all ... [2M(5)]

(42) Es wird allgemein angenommen, [$_{F-1}$ dass die Militärs, [$_{C-2}$ die das Land dreizehn Jahre lang mit Unterschiedlichem Erfolg und — mit Ausnahme Murtala Muhammeds, [$_{C-3}$ der erst sieben Monate an der Macht war, [$_{F-4}$ als er im Februar 1976 ermordet wurde]] — ohne Popularität zu erlangen geführt haben,] von sich aus eine Rückkehr an die Macht nicht anstreben. 'It is normally surmised (F-1) that the soldiers (C-2) who ruled the country thirteen years with variable results and – with the exception of Murtala Muhammed, (C-3) who first was seven months in power (F-4) until he was murdered in February 1976 – without achieving popularity, do not themselves strive for a return to power.' [2M(9)]

(43) For an analysis of the possible modifications [$_{C-1}$ of which the pathological termination of an act [$_{C-2}$ which is not according to law] are susceptible] we have therefore only to ... [3M(2)]

(44) (Swedish:) Helt säkra på [$_{C-1}$ vad blandningen, [$_{C-2}$ som de insjuknade har druckit] består av,] var läkarna inte.] '(M) Quite sure of (C-1) what the mixture, (C-2) that the patients had drunk (C-1) consisted of, (Main) the doctors were not'. [3M(3)]

Table 2 displays the data of Table 1 recounted in terms of total effort.

Table 2. Data of Table 1 recalculated in terms of total effort.

Total effort	**N**
0	1
1	9
2	25
3	14
4	22
5	3
6	16
7	1
8	8
9	2
10	1
10+	2
Sum	104

Recall Gibson's definition of the acceptability limit: "… linguistic working memory capacity is somewhere around 2M(3) MUs or just below". Table 2 shows that there are no less than 16 instances of 2M(3) and its equivalents of a total

effort of 6. These instances cannot all be declared unacceptable by intuition alone. This suggests at least that the acceptability limit rather is above than below 2M(3) MUs and its equivalents.

Discussion

The analysis of C^3s and C^2s has shown that the conjecture of a demarcation line at 2M(3) MUs, or slightly below, between acceptable and unacceptable multiple center-embeddings does not find clear support in real language data drawn from genuine written texts or, rarely, from natural spoken discourse. If there is such a demarcation line, it is rather above than below 6 total effort units. But most likely, the overall data speak in favor of a cline of asymptotically decreasing complexity.

There might be systematic flaws in the design of the procedure counting MUs. For example, the Swedish sentence (44) has a WM cost of 3M(3) which is huge. All ten native informants (including myself) I have consulted on the acceptability of (44) confirm that there is nothing weird about this sentence, which appeared in 2001 in the main Swedish newspaper of Finland, *Hufvudstadsbladet*. It is perfectly grammatical, acceptable and understandable. Its WM expenditure is high just because the prediction of its main clause subject is satisfied only after the doubly center-embedded relative clauses have been passed. That is, the prediction of a postposed main clause grammatical subject (here, *läkarna* 'doctors') turns out to be overly costly. The model needs revision.

A similar problem occurs in sentences with an initial modal or frame adverbial and a postposed grammatical subject, like (38, 39). The late grammatical subject causes the WM cost to become unrealistically high.

Overall, the data of this paper are in good conformance with current theories of the nature of working memory, e.g. Cowan's (2000, 2005) theory of the storage limit on WM being around four chunks, or the well-known capability of humans to be able to simultaneously register some four elements in the focus of visual attention. Recall that even C^2 is next to non-existing in spoken language (15). Sentence (30) above is one of the few documented ones from spoken language. Its WM cost is only 1M(2), far below Gibson's acceptability limit. Of course one should not let the most extreme instances of written language, such as (44), define what the bottom line of (spoken) language WM consumption is.

References

Baddeley, A., Gathercole, S., & Papagno, C. (1998). The phonological loop as a language learning device. *Psychological Review, 105,* 158-173.

Blatt, F. (1957). Latin influence on European syntax. *Travaux du Cercle Linguistique de Copenhague, 11,* 33-69.

Cowan, N. (2000). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24,* 87-185.

Cowan, N. (2005). *Working memory capacity*. New York: Psychology Press.

De Roeck, A., Johnson, R., King, M., Rosner, M., Sampson, G., & Varile, N. (1982). A myth about centre-embedding. *Lingua, 58,* 327-340.

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition, 68,* 1-76.

Givón, T., & Shibatani, M. (Eds.) (2009). *Syntactic complexity. Diachrony, acquisition, neuro-cognition, evolution.* Amsterdam and Philadelphia: John Benjamins Publishing Company.

Hiltunen, R. (1984). The type and structure of clausal embedding in legal English. *Text, 4,* 107-121.

Hoffmann-Krayer, E. (1925). *Geschichte des deutschen Stils in Einzelbildern*. Leipzig: Verlag von Quelle & Meyer.

Hurford, J. R. (2007). *The origins of meaning. Language in the light of evolution.* New York: Oxford University Press.

Karlsson, F. (2007a). Constraints on multiple center-embedding of clauses. *Journal of Linguistics, 43,* 365-392.

Karlsson, F. (2007b). Genuine data on multiple center-embedding of clauses. [Available at www.ling.helsinki.fi/~fkarlsso/ceb_data.pdf]

Lewis, R. L. (1996). Interference in short-term memory: The magical number two (or three) in sentence processing. *Journal of Psycholinguistic Research, 25,* 93-115.

MacWhinney, B. (2009). The emergence of linguistic complexity. In Givón & Shibatani (Eds.).

Pawley, A. (2009). On the origins of serial verb constructions in Kalam. In Givón & Shibatani (Eds.).

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1989). *A comprehensive grammar of the English language*. London: Longman.

Schneider, W. 1959. *Stilistische deutsche Grammatik.* Freiburg im Breisgau: Verlag Herder KG.

Tucker, D. M., Luu, P., & Poulsen, C. (2009). Neural mechanisms of recursive processing in cognitive and linguistic complexity. In Givón & Shibatani (Eds.).

Vasishth, S. (1972*). Working memory in sentence comprehension. Processing Hindi center-embeddings*. New York and London: Routledge.

Coordination of Understanding in Face-to-Face Narrative Dialogue

Kathleen M. Eberhard (eberhard.1@nd.edu) **Hannele B. M. Nicholson** (hnichol1@nd.edu)
Department of Psychology, University of Notre Dame
Notre Dame, IN 46556 USA

Abstract

We report the results of a study investigating speakers' and addressees' coordination of understanding in face-to-face narrative dialogue. Analyses of the occurrence of addressees' acknowledgments and exemplifications of understanding showed that nonverbal forms consistently coincided with the speakers' gaze on their face. In contrast, there was less consistent correspondence between the addressees' verbal evidence of understanding and the speakers' gaze on their face. Evidence that speakers gaze off addressees' faces because of the demands of utterance planning or encoding comes from a correspondence between their gaze off the addressee's face and their production of pause fillers (*uh* and *um*), especially at the beginning of clauses.

Keywords: spoken dialogue; disfluency; gaze patterns

Introduction

Conversation is the quintessential form of language use. It is a purposeful activity requiring the coordination of two or more people. The aim of the research presented here was to examine the coordination process in a face-to-face storytelling situation by examining speakers' gaze patterns relative to delays in speaking signaled by pause fillers and relative to addressees' signals of understanding.

According to Clark (1996), language use is a joint project consisting of 4 hierarchical levels of speaker-addressee coordinated actions, which he refers to as an action ladder. Consider the case of a speaker asking an addressee, "What time is it?". At the first level, the speaker is executing a behavior that consists of producing the sounds of the utterance. The addressee, in turn, attends to the behavior (speech). At the second level, the speaker is presenting words and phrases, which the addressee identifies as such. At the third level, the speaker is signaling an intended meaning (requesting the current time), and the addressee is understanding the meaning. At the fourth level, the speaker is proposing a joint project, namely that the addressee inform him of the current time, and the addressee considers accepting the proposal. There are two essential properties of this hierarchy of actions. The first is upward causality: The actions at a lower level cause the actions at the next level up. The second property is downward evidence: Evidence of successful completion of the actions at a higher level constitutes evidence of success at all levels below it.

As Clark (1996, p. 222) states, "A fundamental principle of any intentional action is that people look for evidence that they have done what they intended to do." Furthermore, people strive to provide evidence that is sufficient for current purposes, in a timely manner, and with the least effort. In the example above, the valid, timely, and sufficient evidence of success comes from the addressee responding with the current time soon after the end of the speaker's utterance. In doing so, the addressee provides positive evidence of her acceptance of the speaker's proposed joint project at level 4 as well as positive evidence of her understanding the meaning of the speaker's utterance (level 3), her identification of the speaker's words (level 2), and her attending to the speaker's speech (level 1).

In contrast to interactive conversation, in narrative dialogues, the speaker produces sequences of utterances across an extended time, resulting in minimal turn-taking. Thus, the joint project at level 4 is an extended proposal consisting of multiple iterations through the lower 3 levels. In this situation, the highest level of evidence of successful completion is level 3 (signaling and understanding of meaning). There are two main forms of evidence that are provided by addressees. One form is *acknowledgments*, which are assertions of understanding, also referred to as backchannels (Yngve, 1970) and generic listener responses (Bavelas, Coates, & Johnson, 2000). They may be verbal, e.g., *mhm*, *okay*, *uh huh*, or non-verbal head nods. The second form of evidence is *exemplifications* of understanding, also referred to as specific listener responses (Bavelas et al, 2000). Exemplifications are reactions to the meaning of the speaker's utterance, and, as such, they constitute more valid evidence. They can be verbal, e.g., *wow*, *oh*, *that's awful*, or non-verbal, e.g., facial gestures, such as winces, grimaces, looks of surprise or sadness. Both acknowledgments and exemplifications are brief, requiring little planning and they often overlap with or occur at the end of the speaker's utterance (e.g., Goodwin, 1981).

Evidence that the addressee is attending to the speaker's execution of a communicative behavior (level 1) is provided by his or her maintaining gaze on the speaker's face (e.g., Argyle & Cook, 1976; Ehrlichman, 1981; Goodwin, 1981; Kendon, 1967). In contrast, the speaker often exhibits a pattern of gazing on and off the addressee's face (e.g., Ehrlichman, 1981, Goodwin, 1981; Kendon, 1967). In interactive conversation, this asymmetry in mutual gaze is considered one cue to turn-taking (Duncan, 1974; Kendon, 1967; Maclay & Osgood, 1959). That is, speakers typically gaze at the addressee at the end of their turn, thereby relinquishing the floor.

However, the asymmetry in speaker's and addressee's gaze on the other's face is also observed in narrative dialogues. In this situation, the speaker's gaze on and off the addressee's face is likely to reflect other aspects of

coordination than turn-taking. In particular, following Kendon (1967), Bavelas, Coates, and Johnson (2002) proposed that in narrative dialogue, speakers gaze at addressees for evidence of their understanding. Support for this proposal comes from their finding that addressees' acknowledgments and exemplifications occurred more often when speakers gazed at them than when they gazed away, and, the speakers' gaze away occurred shortly after the occurrence of this evidence of understanding.

The current study sought to replicate and extend Bavelas et al.'s (2002) findings. Specifically, like Bavelas et al., the current study tested the hypothesis that speakers gaze at their addressee's face for evidence of successful understanding (level 3), which is provided by the addressee producing verbal and/or nonverbal acknowledgments and exemplifications. The current study extends Bavelas et al's work by investigating the additional hypothesis that speakers gaze away from their addressee when the resource demands for utterance planning or encoding are high. This second hypothesis was tested by examining the co-occurrence of speakers' gaze off their addressee's face and their production of pause fillers such as *uh* or *um*, which signal a delay in speaking due to planning or encoding difficulties (Bortfeld, Leon, Bloom, Schober, & Brennan, 2001; Brennan & Schober, 2001; Clark & Fox Tree, 2002; Fox Tree, 2001). Studies investigating interactive dialogues with frequent turn-taking provide some evidence for this latter hypothesis by showing that speakers often gaze away from their addressee at the beginning of their turn at talk (Kendon, 1967; Beattie, 1978), which is the point at which speakers produce pause fillers when the high demands of utterance planning and encoding cause a delay in speaking (Smith & Clark, 1993).

As mentioned above, the speaker's gaze away from the addressee at the beginning of a turn in interactive dialogues may also be a procedure by which the speaker establishes his or her turn to talk). The limited turn-taking in narrative dialogue eliminates this possible role of gaze patterns. Furthermore, we examined whether there is a co-occurrence of the speaker's gaze off an addressee's face with pause fillers that occur before a clause/discourse segment, when the demands of utterance planning should be high, as well as within a clause, where the delay may reflect lexical retrieval difficulties.

The speakers in the current study read an obscure Brothers Grimm story, which they then told to an addressee. The story involves three main characters and several subordinate characters. It also has four main scenes, corresponding to different time periods and settings. To ensure that the speakers understood the story and that they would tell it in a relatively uniform way, they completed a quiz after reading it and before telling it to the addressees. The addressees completed the same quiz after listening to the speakers tell the story to them.

Experiment

Method

Participants Seven same sex dyads participated in the experiment in exchange for payment of $10.00 each or extra credit in a course. All were native American English speaking adults with an average age of 21 years. Five dyads were female, two of which were familiar with each other prior to the experiment. One male dyad were familiar with each other prior to the experiment.

Procedure The members of the dyads signed up for an hour-long session with one person, designated as the Speaker, arriving 30 minutes before the other person, who was designated as the Addressee. Upon arriving at the lab, the Speaker was given a consent form to read and sign. Then, he or she read a printed copy of the Brothers Grimm story *Faithful John* in a quiet room. All of the participants were unfamiliar with the story prior to the experiment. After reading the story, the Speaker completed a quiz consisting of 14 multiple-choice questions about the main events and characters in the story. The questions were presented one at a time on a computer screen, and the Speaker was given as much time as needed to select a response, which was made by pressing a key on the keyboard. The Speaker was allowed to consult the printed copy of the story when answering the questions, and the correct response for each question was displayed after the Speaker made his or her response. The Speaker was then seated at a table and fitted with a free-head eye-tracker.

After the Addressee read and signed the consent form, he or she was seated at the table opposite to the Speaker. For the four dyads who were unfamiliar with each other, the Speaker and Addressee were introduced, and each was asked to tell the other about themselves (e.g., where they were living, what year they were in college, and what their major was). The dyads were then given instructions for the task. Specifically, they were told that the experiment investigates conversational interactions between two individuals, and that the Speaker was to tell a Brothers Grimm story, which she or he had just read, to the addressee. The Addressee was told that the Speaker had answered a set of comprehension questions about the story and that he or she would receive the same set of questions after listening to the Speaker tell the story to him or her. Thus, the goal was for the Addressee to understand the story sufficiently well to be able to answer the questions correctly. The Speaker and Addressee were told that they could talk to each other and that the important thing was for them to interact as naturally as possible.

The Speaker and Addressee were informed that the experimenter would remain in the room to monitor the recording equipment. However, she would have her back to them and would listen to music over headphones to prevent her from participating as an "overhearer". The Addressee was instructed to tap the experimenter on the shoulder when the speaker had finished telling the story. The entire story-

telling session was video-taped, and after it was over, the Addressee completed the quiz.

Apparatus The Speaker was fitted with an eye-tracker (Applied Science Laboratories, Model 501) consisting of a lightweight eye camera attached to an adjustable headband. The eye camera was positioned above the Speaker's left eye, and it captured an infrared image of the eye at a 60 Hz sampling rate. The distance between the centers of the corneal and pupil infrared reflections were used to calculate the relative eye-in-head position. The head band also contained a scene camera that captured an image of the Addressee's head and torso across the table. The scene camera's image was displayed on a TV monitor along with a record of the Speaker's eye movements in the form of cross hairs that were superimposed over the scene image. A brief calibration routine was conducted to map nine eye-position coordinates onto nine corresponding scene-image coordinates. The accuracy of the resulting eye fixation record was approximately 0.5° over a range of $\pm 20^{\circ}$. Lapel microphones were attached to the Speaker's and Addressee's shirts and connected to a Hi8 VCR, which also recorded the scene image and eye-movement record displayed on the TV. A Hi8 video camera, which was positioned to the side of the Addressee, recorded an image of the Speaker's head and torso. Responses on a survey administered at the end of the experimental session indicated that the eye-tracking apparatus was not distracting or only minimally distracting to the Speakers, and it was minimally to moderately distracting to the Addressees.

Video Coding The two video-taped recordings of each dyad's experimental session were digitized at a 60Hz NTSC sampling rate and aligned with each other using Final Cut Express (Apple, Inc.). The project files were annotated using frame-by-frame playback of the synchronized audio and video tracks (each frame = 33 msec). Labeled markers were inserted on the first frame of events of interest and extended to the last frame. All coding was done independently by two individuals, with a third individual (KE) reconciling any disagreements. Categories of events of interest that were marked included the following:

(1) *Speaker's gaze*: The Speakers' gaze on and off the Addressee's face was coded in a binary fashion such that the frame that marked the last consecutive fixation on the Addressee's face was followed by the frame that marked the first fixation off the face. The Speaker's gaze on the Addressee's face consisted of two or more consecutive fixations anywhere on the face. The Speaker's gaze off the Addressee's face consisted of one or more fixations in the region surrounding the face, including the Addressee's neck and torso as well as the wall behind the Addressee. In addition, the gaze off the Addressee's face included instances in which there was a loss of the eye-tracking record due to the Speaker looking down or closing his or her eyes for a period longer than a blink.

(2) *Addressee's gaze*: The first and last frames of the Addressee's gaze away from the Speaker's head were marked based on the direction of the Addressee's eye gaze available from the eye-tracker's scene image. The Addressee's gaze away typically involved looking down at the table or to the left or right of the Speaker.

(3) *Addressee's nonverbal responses*: The beginning and end frames of the Addressee's head nods (acknowledgments) and facial gestures (exemplifications) were marked. Facial gestures displayed reactions to the story's content such as surprise or disbelief in the form of eye flashes or raised eyebrows, grimaces, winces, and frowns. Smiles were not included as a nonverbal response.

Utterance Coding The Speakers' and Addressees' utterances were orthographically transcribed using Praat (Boersma & Weenink, 1996). Transcriptions of the Speaker's and Addressee's utterances were created on separate tiers in the textgrid files, with the tiers time-aligned with the digitized audio track (48 kHz sampling rate). Transcriptions were completed independently by two individuals and checked by a third (HN). The Addressee's transcriptions contained boundaries that marked the utterances' onset and offset. The utterances consisted of acknowledgments (e.g., *okay*, *mhm*, *hmmm*, *oh*, *uh huh*) and exemplifications (e.g., *wow*, *that's weird*, *crazy*), as well as requests for clarification.

Two duplicate tiers contained the transcriptions of the Speakers' utterances. One tier contained boundaries that marked intonational phrases, which typically consisted of one or two clauses. The other tier contained boundaries for individual words, which included pause fillers (e.g., *uh*, *um*) as well as silent pauses. A third tier was used to label the pause fillers with respect to whether they occurred at the beginning of a clause, within a clause, or embedded in a larger disfluency involving a repair. As shown in examples (a) and (b) below, clause-initial fillers preceded or followed one or more discourse markers (e.g., *so*, *and*, *then*, etc.). Examples (c) and (d) show fillers that occurred within a clause, and example (e) shows a filler that occurred in the middle of a larger disfluency. The numbers in square brackets show the location of a silent pause and its duration in seconds.

a.) *Clause-initial*: [0.494] um so he knows what his inheritance is except for this one [0.635]
b.) *Clause-initial*: [0.334] and [0.596] um so they know that this princess really likes gold
c.) *Within-clause*: and they take a ship across the um [0.109] sea or something
d.) *Within-clause*: and she's like wow can I [0.426] um get some of that
e.) *Mid-disfluency*: if someone sticks [0.383] um [0.227] if someone makes her lip bleed

Analyses: The markers coding the video recordings were exported from Final Cut Express and imported into Praat as labeled tiers in the textgrid files that were time-aligned with

the transcription tiers and the digitized audio track. Scripts were used to extract frequency and duration information from the tiers. The analyses of the pause fillers and gaze patterns excluded fillers that were part of a larger disfluency (i.e., the mid-disfluency fillers).

Results

As shown in Table 1, the Speakers took an average of 659 seconds, or about ten minutes, to tell the story, and they did so with an average speaking rate of 192 words per minute. The Speakers' accuracy on the quiz was slightly higher than the Addressees (average of 95% vs. 91%, respectively). For three dyads in which both the Speaker and Addressee scored less than 100%, the questions that were responded to inaccurately by the Speaker differed from the questions that were responded to inaccurately by the Addressee.

Table 1: Total time, speech rate, and quiz scores

Dyad	Total time (sec)	Words per min	S's quiz score	A's quiz score
F1	591	202	100%	100%
F2	773	175	100%	86%
F3	531	205	93%	93%
F4*	1047	178	100%	100%
F5*	627	249	93%	86%
M1	551	155	93%	100%
M2*	491	181	86%	75%
Mean	659	192	95%	91%

Note: S = Speaker, A = Addressee, F = female, M = male, * = friends prior to experiment

Table 2 shows the number and mean duration of the Speakers' gaze on and off the Addressee's face for each dyad, as well as the percentage of the total time that the Speakers' gazed off the Addressee's face. Five of the seven Speakers' exhibited the commonly reported pattern of spending more time gazing off their Addressee's face than gazing on their Addressee's face. The other two Speakers, one male (M2) and one female (F5), spent less time gazing off their Addressee's face than on it. As for the Addressees, all five female Addressees gazed at their Speaker's face 97% or more of the storytelling time. The two male Addressees gazed at their Speaker's face 79% (M1) and 46% (M2) of the storytelling time, respectively.

Table 2: Number and average duration (sec) of Speakers' gaze on and off the Addressee's face

Dyad	# Gaze on	Duration gaze on	Duration gaze off	% Total time gaze off
F1	127	1.271	3.380	73%
F2	285	1.209	1.499	55%
F3	211	1.066	1.449	58%
F4	379	0.991	1.805	65%
F5	107	5.405	0.451	8%
M1	189	0.634	2.580	78%
M2	132	2.519	1.203	32%
Mean	204	1.871	1.724	53%

Gaze and Addressees' Responses: Table 3 shows the number of the Addressees' nonverbal and verbal acknowledgments (e.g., head nods, saying *mhm*, *okay*, etc.) and exemplifications of understanding (e.g., looks of surprise, grimaces, saying *wow*, *oh my*, etc.). There was variability across the dyads in the frequency of providing evidence of understanding, with the total number of all forms ranging from 17 (M2) to 201 (F2). However, all 7 Addressees produced more acknowledgments than exemplifications as well as more nonverbal responses than verbal responses.

Table 3: Number of Addressees' acknowledgments and exemplifications and the percentage that overlapped with the Speaker's gazed on their face

	Acknowledgments		Exemplifications	
	Nonverbal	Verbal	Nonverbal	Verbal
F1	40 (75%)	9 (44%)	14 (79%)	6 (17%)
F2	111 (86%)	64 (55%)	21 (95%)	5 (60%)
F3	11 (73%)	5 (80%)	6 (83%)	0
F4	157 (60%)	28 (54%)	4 (100%)	0
F5	76 (99%)	38 (97%)	10 (100%)	11 (100%)
M1	15 (80%)	2 (0%)	0	0
M2	51 (90%)	8 (88%)	1 (100%)	5 (100%)
Mean	66 (80%)	22 (60%)	8 (93%)	4 (69%)

On average 80% of the Addressees' non-verbal acknowledgements (head nods) overlapped with the Speaker's gaze on the Addressee's face (range 60% to 99%), and 93% of their non-verbal exemplifications overlapped with the Speaker's gaze on the their face. In contrast, the average percentages of the Addressees' verbal acknowledgments and verbal exemplifications that overlapped with the Speaker's gaze on their face were less, i.e., 60% and 69%, respectively. For each dyad, the number of the Addressee's non-verbal responses and verbal responses that overlapped with the Speaker's gaze on his or her face was compared to the numbers expected to overlap by chance using the procedure described by Bavelas et al. (2002). Specifically, when the total number of nonverbal

responses or verbal responses was greater than 20, a z-value was calculated and evaluated with the normal distribution using the formula:

$$z = \frac{O - E - .5}{\sqrt{npq}}$$

where, n is the total number of responses, O is the observed number of responses overlapping with the Speaker's gaze on the face, p is the percentage of total time the Speaker spent gazing on the Addressee's face, q is 1-p, and E is the expected number of responses overlapping with a gaze on face by chance (p*n). The subtraction of .5 is a correction for continuity. When the total number of verbal or nonverbal responses was less than or equal to 20, then the combination of n, p, and O were tested for significance using the binomial distribution. The results of the tests for each dyad are given in Table 4.

Table 4: Tests of the significance of the observed number of Addressees' responses occurring with gaze on their face

Dyad	*n* total responses	*O* # with gaze on face	*p* % total time gaze on face	*z*	p-value[a]
		Addressees' Nonverbal Responses			
F1	54	41	0.27	7.95	< .0001
F2	132	116	0.45	9.81	< .0001
F3	17	13	0.42		< .002
F4	161	98	0.35	6.80	< .0001
F5	86	85	0.92	2.14	< .02
M1	15	12	0.22		< .002
M2	52	47	0.68	3.31	< .0001
		Addressees' Verbal Responses			
F1	15	5	0.27		n.s
F2	69	38	0.45	0.56	= .06
F3	5	4	0.42		< .05
F4	28	15	0.35	1.86	< .05
F5	49	48	0.92	1.27	n.s
M1	2	0	0.22		n.s
M2	13	12	0.68		< .05

[a]One-tailed test was used for binomial tests when $n \leq 20$.

Table 4 shows that the number of the Addressees' nonverbal responses that overlapped with the Speaker's gaze on their face was significantly greater than the number expected by chance for all seven dyads. In contrast, the number of the Addressees' verbal responses that coincided with the Speaker's gaze on their face was significantly greater than the number expected by chance for only three of the seven dyads, and it was marginally significant for one other dyad. The results for the nonverbal responses replicates Bavelas et al.'s (2002) findings. The current finding that the Addressees' verbal evidence of understanding less consistently overlaps with the Speaker's gaze on their face is likely due to Speaker's gaze being unnecessary for conveying this form of evidence.

Gaze and pause fillers: The Speakers produced an average of 41 pause fillers (range 15 - 76), at an average rate of 1.87 per 100 words (range 1.0 - 3.4). The correlation between the Speakers' pause filler rate and the average duration of their gaze off the Addressee's face is 0.45. As shown in Table 5, the Speakers produced more clause-initial pause fillers than within-clause ($t(6) = 4.05$, $p < .02$, two-tailed); however, clause-initial fillers were not significantly longer in duration ($t(6) = 2.08$, $p = .08$, two-tailed).

Table 5: Filled pause rate per 100 words, % of all pause fillers (number) and average duration (sec) that were clause-initial or within-clause

Dyad	Rate	Clause-initial Total	Clause-initial Dur. (sec)	Within-clause Total	Within-clause Dur. (sec)
F1	2.57	53% (27)	0.439	33% (17)	0.378
F2	3.37	67% (51)	0.399	25% (19)	0.362
F3	0.99	61% (11)	0.438	11% (2)	0.519
F4	1.45	71% (32)	0.406	20% (9)	0.340
F5	1.67	52% (22)	0.378	21% (9)	0.307
M1	2.81	55% (22)	0.422	43% (17)	0.349
M2	1.01	53% (8)	0.449	33% (5)	0.355
Mean	1.98	59% (25)	0.419	27% (11)	0.373
SD	0.94	8% (14)	0.026	11% (7)	0.068

For 6 of the 7 Speakers, all or all but one of their clause-initial pause fillers coincided with their gazing off their Addressee's face. The Speaker (F5) who spent most of the storytelling time (92%) gazing on her Addressee's face had fewer clause-initial pause fillers (23%) coinciding with a gaze off her Addressee's face than with a gaze on. Nevertheless, a binomial test of the number of the clause-initial pause fillers that coincided with her gaze off the Addressee's face was significantly greater than expected by chance ($p = .02$, two-tailed). Thus, there was a clear correspondence between the occurrence of the Speakers' gaze off their Addressee's face and their production of pause fillers at the beginning of clauses, when the demands of utterance planning and encoding re likely to be highest.

An examination of the within-clause pause fillers also provided evidence that these signals of production difficulty coincided with the Speakers' gaze off their Addressee's face in a narrative situation. Specifically, except for the Speaker (F5) who spent most of the storytelling time gazing on her Addressee's face, the number of within-clause pause fillers produced by the other six Speakers that coincided with their gaze off their Addressee's face was greater than the number expected by chance, which was calculated by multiplying the percentage of the Speaker's total time gazing off the Addressee's face by the Speaker's total number of within-clause pause fillers. Binomial tests were significant for four of the six Speakers (p-values $\leq .05$, one-tailed), and marginally significant for one Speaker (M1) ($p = .08$, one-tailed). The test was nonsignificant for the remaining Speaker (F3) due to a small number of observations i.e.,

only 2 within-clause pause fillers, both of which overlapped with the Speaker's gaze off the Addressee's face. For the Speaker (F5) who spent most of the time gazing on her Addressee's face, only 1 of her 9 within-clause pause fillers coincided with her gaze off the Addressee's face, which was equal to the number expected by chance, albeit not significant by a binomial test ($p > .05$).

Discussion

The results of the current study demonstrated coordination of understanding between Speakers and Addressees during a face-to-face narrative dialogue. Specifically, consistent with Bavelas et al.'s findings, Addressees produced nonverbal acknowledgments and exemplifications of their understanding more often when the Speaker gazed on their face than when the Speaker gazed off their face. However, across all seven dyads, there was less consistent co-occurrence of the Addressees' verbal acknowledgments and exemplifications (e.g., *mhm*, *wow*) with the Speaker's gazed on their face. This finding is likely due to the Speaker's gaze on the Addressee's face being unnecessary for conveying this evidence verbally. The results extended previous findings by providing evidence that Speakers gaze off their Addressee's face in narrative dialogues when they experience a delay in speaking due to utterance planning or encoding. Specifically, for six Speakers, nearly 100% of their pause fillers (*um*, *uh*) that occurred before a clause, when the demands of utterance planning are high, coincided with their gazing off their Addressee's face. For all seven Speakers, the number of their clause-initial pause fillers that coincided with a gaze off was significantly greater than expected by chance. There was some evidence that pause fillers that occurred within a clause also coincide with the Speaker's gaze off the Addressees' face, however this relationship was significant for only four of the six Speakers. Future research will examine the relationship between the Speaker's gaze off the Addressee's face and longer disfluent intervals, such as a syllable prolongation followed by a pause, then pause filler, etc. In addition, coordination may also be reflected in Speakers seeking and Addressees providing evidence that a disfluency involving a repair did not impede the Addressee's understanding.

Conclusion

Although there are a number of studies investigating coordination via gaze patterns, signals of understanding, and disfluencies in interactive conversation (e.g., Bard, Anderson, Chen, Nicholson, Havard, Dalzel-Job, 2007), few studies have investigated coordination in narrative dialogue. The research presented here extends previous findings by demonstrating that Speakers' gaze on and off their Addressee's face when telling a story reflect the demands of encoding meaningful messages in speech, and evidence of its success.

Acknowledgments

We thank Carlene Koken, WonJae Shin, Susan Gundersen for their assistance with data collection and coding.

References

Argyle, M., & Cook, M. (1976). *Gaze and mutual gaze*. Cambridge: Cambridge University Press.

Bard, E. G., Anderson, A. H., Chen, Y., Nicholson, H. B. M., Havard, C., & Dalzel-Job, S. (2007). Let's you do that: Sharing the cognitive burdens of dialogue. *Journal of Memory & Language, 57*(4), 616-641.

Bavelas, J. B., Coates, L., & Johnson, T. (2002). Listener as a collaborative process: The role of gaze. *Journal of Communication, September*, 566-580.

Beattie, G. W. (1978). Sequential temporal patterns of speech and gaze in dialogue. *Semiotica, 23*, 29-52.

Boersma, P. & Weenink, D.. (1996). Praat: A system for doing phonetics by computer. Inst. Phonetic Sci., Univ. Amsterdam, Amsterdam, The Netherlands, http://www.praat.org.

Bortfeld, H., Leon, S. D., Bloom, J. E., Schober, M. F., & Brennan, S. E. (2001). Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender. *Language & Speech, 44*, 123-123.

Brennan, S. E., & Schober, M. F. (2001). How listeners compensate for disfluencies in spontaneous speech. *Journal of Memory & Language, 44*, 274-274.

Clark, H. H. (1996). *Using Language*. Cambridge: Cambridge University Press.

Clark, H. H., & Fox Tree, J. E. (2002). Using uh and um in spontaneous speaking. *Cognition, 84*, 73-111.

Duncan, S. (1974). On the structure of speaker-auditor interaction during speaking turns. *Language in Society, 3*, 161-180.

Ehrlichman, H. (1981). From gaze aversion to eye-movement suppression: An investigation of the cognitive interference explanation of gaze patterns during conversation. *British Journal of Social Psychology, 20*, 233-241.

Fox Tree, J. E. (2001). Listeners' uses of um & uh in speech comprehension. *Memory & Cognition, 29*, 320-326.

Goodwin, C. (1981). *Conversational Organization: Interactions between Speakers and Hearers*. New York: Academic Press.

Kendon, A. (1967). Some functions of gaze-direction in social interaction. *Acta Psychologica, 26*, 22-63.

Maclay, H., & Osgood, C. E. (1959). Hesitation phenomena in spontaneous English speech. *Word, 15*, 19-44.

Smith, V. L. & Clark, H. H. (1993). On the course of answering questions. *Journal of Memory & Language, 32*, 25-38.

Yngve, V. H. (1970). On getting a word in edgewise. *Papers from the sixth regional meeting of the Chicago Linguistics Society* (pp. 567-578). Chicago: Chicago Linguistic Society.

The Interpretation of Null and Overt Pronouns in Japanese: Grammatical and Pragmatic Factors

Mieko Ueno (miueno@ucsd.edu)
Andrew Kehler (akehler@ucsd.edu)
Department of Linguistics; University of California, San Diego
9500 Gilman Drive, La Jolla, CA 92093-0108, USA

Abstract

Pronoun interpretation in English has been demonstrated to be sensitive to an interaction between grammatical and pragmatically-driven factors. This study investigated the interpretation of pronouns in Japanese, which has both null and overt forms. Thirty-two native speakers of Japanese participated in a passage completion experiment with transfer-of-possession contexts, varying prompt type, aspect, and topic/nominative-marking of the previous subject. Two judges annotated the referents of the matrix subjects and coherence relations in the completed passages. Japanese overt pronouns were revealed to pattern closely with English overt pronouns in their sensitivity to pragmatic factors, whereas Japanese null pronouns were predominantly governed by grammatical position. Somewhat surprisingly, topic-marking did not influence reference or coherence relations. The data suggest distinctive patterns of interactions between grammatical and pragmatic factors in the interpretation of null and overt pronouns in Japanese, and cast doubt on the existence of a division of labor between the two forms.

Keywords: Japanese pronoun interpretation, discourse processing, cross-linguistic language processing

Introduction

Previous work (Stevenson, Crawley, & Kleinman, 1994; Arnold, 2001; Rohde, Kehler, & Elman, 2006, 2008) has shown that pronoun interpretation in English is driven by the interaction of grammatical and pragmatic biases. For instance, Rohde et al. (2006) showed that pronoun interpretation differs in transfer-of-possession passages that vary by verbal aspect between perfective (1) and imperfective (2).

(1) John$_{\text{SOURCE}}$ handed a book to Bob$_{\text{GOAL}}$
He _____

(2) John$_{\text{SOURCE}}$ was handing a book to Bob$_{\text{GOAL}}$
He _____

The context sentences in (1) and (2) contain two possible referents for the pronoun, one that appears in subject position and fills the Source thematic role (*John*), and one that appears as the object of a prepositional phrase and fills the Goal thematic role (*Bob*). The results of a passage completion experiment revealed significantly more interpretations of pronouns to the Source referent (the grammatical subject) in the imperfective condition as compared to the perfective condition. Rohde et al. also found that the influence of aspect in pronoun interpretation was correlated with certain relationships inferred to hold between the two clauses (henceforth 'coherence relations'), suggesting that a shift in the distribution of coherence relations induced the shift in the distribution of pronoun interpretations.

Following Stevenson et al. (1994), Rohde et al. (2008) ran passages with pronoun prompts like (1) against those with 'free' prompts (3).

(3) John handed/was handing a book to Bob.
_____

Results showed more references to the Source and more Source-biased coherence relations in the pronoun condition than in the free condition, indicating that pronouns overlay a grammatical subject bias on top of the pragmatic biases that were revealed by the aspect manipulation.

Present Study

Null pronouns in Japanese occur most commonly in subject position, but occasionally in object positions as well (Ueno & Polinsky, 2009). Overt pronouns also exist, but occur less commonly than the null forms (Martin, 1976).[1]

The interpretation of Japanese null pronouns has been claimed to be analogous to the interpretation of overt pronouns in other languages without a null form (e.g., Kuroda, 1965; Kameyama, 1985; inter alia). The GIVENNESS HIERARCHY (GH) of Gundel, Hedberg, and Zacharski (1993) makes this claim as well, and further predicts that the Japanese null and overt forms should display a 'division of labor' effect whereby the preferred referents of the two forms fall into complementary distribution. These predictions result from the fact that the six cognitive statuses that comprise the GH participate in an implicational hierarchy, and are thus expected to give rise to scalar implicatures. According to the GH, English overt pronouns and Japanese null pronouns require referents of the highest status (IN FOCUS), whereas Japanese overt pronouns occupy the second highest status (ACTIVATED).

[1] The Japanese third person overt pronouns (e.g., *kare* 'he', *kanojo* 'she') are generally considered to be direct translations of their English counterparts, and appear to be becoming incorporated into daily Japanese at an increasing rate. A corpus count of *Asahi Shimbun* (popular Japanese newspaper) articles shows that out of 11,073,167 sentences, *kare* was used 28,795 times and *kanojo* 14,209 times (Amano & Kondo, 2000).

Whereas overt pronouns are compatible with both ACTIVATED and IN FOCUS referents, Grice's (1975) Maxim of Quantity ('say as much as you need to say') predicts that the informationally-stronger null form should be used for IN FOCUS referents, in turn predicting that overt pronouns will be used only for referents that are ACTIVATED but not IN FOCUS, creating the division-of-labor effect.

A relatively small number of experimental studies have been performed on the interpretation of null pronouns. Working within Centering Theory, Walker, Iida, and Cote (1994) reported an influence of grammatical/information-structural factors found in a referent-choice experiment, including higher salience for topic-marked (*-wa*; cf. Kuno, 1973) than nominative/subject-marked (*-ga*) referents. A recent study by Christianson and Cho (2009) suggests that topical arguments in Odawa are more likely to be realized as null pronouns than non-topical arguments. Experiments performed by Alonso-Ovalle et al. (2002) offered mixed support for a division-of-labor effect between null and overt pronouns in Spanish. In a written questionnaire study, for instance, null pronouns referred to the previous subject 73.2% of the time whereas overt pronouns did 50.2% of the time; while null pronouns clearly incorporated a stronger subject bias, the referents of the two forms were not strictly in complimentary distribution. Further, whereas an acceptability judgment task found that participants rated sentences with unambiguous references to the previous subject as being more acceptable when a null pronoun was used (4.19 on a 5-point scale) as compared to when an overt pronoun was used (3.57), the overt pronoun cases were still deemed to be relatively acceptable.

Taken together, the foregoing work gives rise to a series of questions that the present study seeks to answer. First, we ask whether the behavior of Japanese null and/or overt pronouns patterns with that of English overt pronouns in displaying sensitivity to pragmatic subsequent-mention biases, or whether their interpretation is determined primarily by other (e.g., grammatical) factors. This question can be addressed by employing a passage completion task that uses the same aspect manipulation employed by Rohde et al. (2006). The second question is whether null and overt Japanese pronouns exhibit a division-of-labor effect such that, for instance, a demonstrated subject bias for null pronouns corresponds to a commensurate non-subject bias for overt pronouns. This question will be answered using a manipulation that varies prompt type between null pronoun, overt pronoun, and free. Third, we ask whether topic-marked antecedents attract more pronominal references than subject-marked antecedents. We will answer this question by varying the morphological marking on the first mentioned referent of the preceding clause, specifically between subject/nominative marking (*-ga*) and topic marking (*-wa*). Lastly, we ask whether any grammatical biases that are revealed to be associated with these referential forms affects the distribution of ensuing coherence relations, as Rohde et al. (2008) found for English. This question will be answered by having judges annotate the completions with respect to coherence relations and comparing the resulting distributions across prompt types.

Methods

We followed the passage completion task design used by Rohde et al. (2006, 2008) using Japanese stimuli.

Participants

Thirty-two native speakers of Japanese recruited from the San Diego area participated in the study. Participants were reimbursed for their time.

Materials

The experiment employed a 3x2x2 design that varied prompt type (Null-pronoun[2]/Overt-pronoun[3]/Free), aspect (Perfective/Imperfective[4]), and topic/nominative-marking of the context sentence subject (*-wa*/*-ga*), as shown in (4).

(4) Stimuli

太郎は/が 次郎に 本を渡した/渡している ところだった。
Taro-wa/ga Jiro-ni hon-o watashita/watashi-te-iru tokoro-datta.
Taro-TOP/NOM Jiro-to book-ACC handed/hand-INF-ASP scene-was
'Taro handed/was handing a book to Jiro.'

主語省略/彼は/自由_____
shugo-shoryaku/kare-wa/jiyu
subject-omission(Null)/he-TOP(Overt)/free(Free)

The 60 experimental stimuli each had context sentences with different transfer-of-possession verbs. The Source referent ('Taro' in (4)) always appeared in subject position, and the Goal referent ('Jiro') was the dative/'to'-marked indirect object of the sentence. All verbs described physical transfer events (e.g., 'hand', 'throw').

Fillers consisted of 40 context sentences, containing transitive or intransitive non-transfer verbs in the perfective

[2] The 'subject-omission' prompt was used to indicate the presence of a null pronoun. A pilot study revealed that most participants were capable of continuing such prompts appropriately, which was confirmed in the actual study.

[3] All overt pronoun prompts were topic-marked. This was done because the pilot study revealed that nominative-marked overt pronouns tend to be used to express an embedded subject of a complex sentence rather than a matrix subject. Topic-marking the pronoun resolved the issue.

[4] Imperfectivity is not as straightforward to encode in Japanese as in English, since *-teita* 'was ~ing' is ambiguous between a perfective and imperfective reading depending on the verb (or VP) with which it co-occurs. Because transfer-of-possession verbs typically express achievement events as a default, the more natural interpretation of *-teita* with these verbs is perfective. We therefore use *tokoro* ('was in the scene of') to 'stretch out' instantaneous events and make an imperfective reading of achievement events possible, in a manner similar to what the English progressive does to achievement events.

or imperfective aspect. The transitive verbs varied between active and passive voice, and adverbs, names, and gender-unambiguous overt pronouns served as prompts. The 100 sets of sentences instantiating the 12 (3x2x2) experimental conditions were placed in a Latin square design to create 12 parallel lists of 100 sentences, such that no one participant saw more than one sentence from each set.

Task

Using a web-based interface, participants were asked to write continuations for the 100 passages. They were instructed to imagine a natural continuation to the story, writing the first continuation that came to mind and avoiding humor.

Data Analysis

Following previous studies on English, we focused our analysis on the interpretation of matrix subjects. Identifying the matrix subject can be less straightforward in Japanese than in English, however, since Japanese clauses may contain multiple null elements, and are characterized by flexible and head-final word order. It therefore proved useful to translate the continuations into English, thereby recovering the referents of null elements. For instance, if the original sentence in Japanese said 'felt happy because passed exam', detectable null pronouns were postulated, as in '(s/he) felt happy because (s/he) passed (her/his) exam' in the relevant English translation. The first author (who is a native speaker of Japanese) then underlined the likely matrix subject of the given sentence for the subsequent annotation processes.

Two trained judges, who were native speakers of Japanese but were blind to the experimental hypotheses, annotated the referent of the matrix subject of each continuation sentence with respect to five categories: Source ('Taro' in (4)), Goal ('Jiro'), Theme ('book'), Other, and Unsure. The judges were instructed to do the annotation separately, without talking to each other. The first author compared their annotations and discarded the cases the judges did not agree on, as well as the cases in which participants did not omit a subject even though they were given a null pronoun prompt. The tokens discarded in this way constituted about 15% of the data.

The remaining tokens were then given back to the judges for annotating the coherence relations that held between each context sentence and continuation, as shown below (Hobbs, 1990; Kehler, 2002; Rohde, 2008).[5]

Elaboration: continuations that provide additional details about the eventuality described in the context sentence
e.g., *Taro handed a book to Jiro. He did so slowly and carefully.*

Explanation: continuations that describe the cause of the eventuality described in the context sentence
e.g., *Taro handed a book to Jiro. He no longer had a use for it.*

Occasion: continuations that describe an eventuality that initiates from the end state of affairs of the eventuality described in the context sentence
e.g., *Taro handed a book to Jiro. He began reading it.*

Result: continuations that describe the effect or result of the eventuality described in the context sentence
e.g., *Taro handed a book to Jiro. He thanked him for the gift.*

The judges resolved disagreements through discussion.

Results

Reference

ANOVAs were run on the percentage of Source referents as a function of the total number of Source and Goal referents. Prompt type, aspect, and topic/nominative marking were used as factors. There was a significant main effect of prompt type [$F_1(2, 31) = 74.11$, $p<.0001$; $F_2(2, 59) = 64.10$, $p<.0001$]. Subsequent Tukey HSD posthoc comparisons found significant differences in order of Null > Overt > Free by both participants and items, i.e., Null pronouns were most Source-biased, followed by Overt pronouns, followed by Free prompt continuations. There was also a significant main effect of aspect [$F_1(1, 31) = 15.81$, $p<.0001$; $F_2(1, 59) = 21.02$, $p<.0001$], indicating that Imperfectives yielded more Source referents than Perfectives. Figure 1 shows the proportion of Source and Goal referents for each prompt type and aspect (collapsed over topic/nominative-marking) averaged across participants. The null pronoun conditions had about 80% Source referents irrespective of aspect, while Overt and Free conditions varied by aspect. Pairwise comparisons between Imperfectives and Perfectives within each prompt type revealed significant differences for Overt pronoun [$t_1(31) = 3.70$, $p=.0008$; $t_2(59) = 3.95$, $p=.0002$] and Free [$t_1(31) = 1.59$, ns; $t_2(52)$[6] $= 3.47$, $p=.0011$][7] conditions, but not for Null conditions.

[5] Although there are several other coherence relations which sometimes occurred – e.g., 'Violated Expectation' and 'Parallel' – we analyzed only these four.

[6] Some degrees of freedom vary due to missing cells.

[7] The lack of significance by subjects in the Free prompt condition was due in part to the fact that the analysis included all continuations, as opposed to only those in which participants started their continuation with something other than a pronoun (i.e., a proper name). When name continuations only were compared, the aspect distinction yielded a marginal effect by participants [$t_1(23) = 2.02$, $p=.0551$] and remained significant by items [$t_2(43) = 2.56$, $p=.0141$].

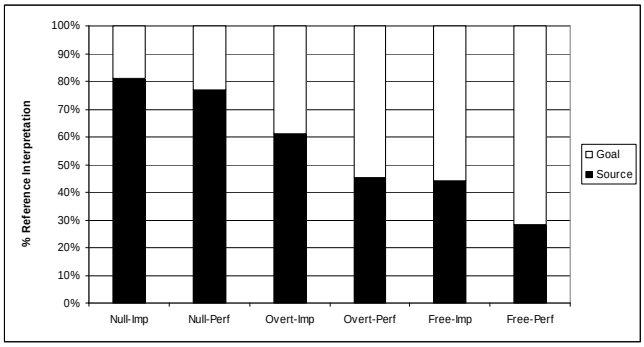


Figure 1: Proportion of Source and Goal referents for all conditions (collapsed over topic-marking).

Mirroring Rohde et al. (2008), overt pronouns led to significantly more subject mentions of the Source than free prompts. We further divided the free prompt continuations according to their matrix subject type, namely, Null pronouns, Overt pronouns, and Names, and performed the same ANOVA as above. The results yielded significant main effects of aspect [$F_1(1, 31) = 4.01$, $p=.0462$; $F_2(1, 59) = 6.02$, $p=.0149$] and subject type [$F_1(2, 31) = 67.59$, $p<.0001$; $F_2(2, 59) = 43.40$, $p<.0001$]. Subsequent Tukey HSD posthoc comparisons found significant differences in order of Null > Overt > Free by participants and Null > Overt, Free by items, which shows the highest proportion of Goal referents for Name continuations, again consistent with Rohde et al.

Unlike aspect, however, there were no significant main effects or interactions involving topic-marking. Figure 2 shows the proportion of Source and Goal referents for each prompt type and topic/nominative-marking (collapsed over aspect) averaged across participants. Pairwise comparisons between Topics and Nominatives within each prompt type revealed no significant differences for any prompt type, but there was a marginal difference in Null continuations [$t_1(30) = 1.70$, $p=.0989$; $t_2(57) = 1.87$, $p=.0665$] that favored subject referents in the topic condition.

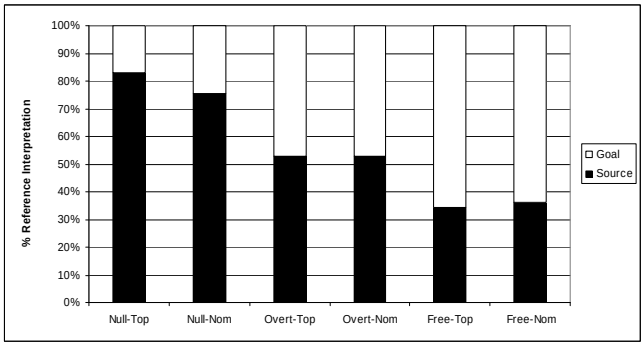


Figure 2: Proportion of Source and Goal referents for all conditions (collapsed over aspect).

Coherence Relations

Figure 3 shows the Source/Goal referent count for each coherence relation (collapsed over 12 experimental conditions) averaged across participants.

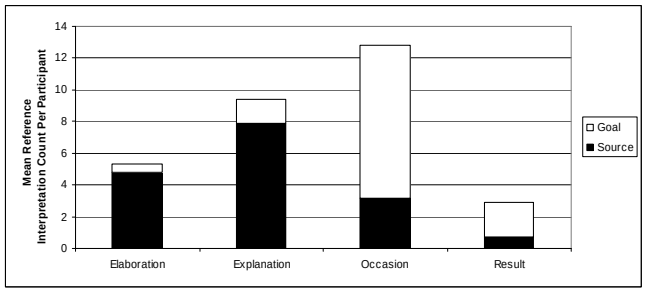


Figure 3: Mean Source/Goal referent count for each coherence relation (collapsed across conditions).

Figure 4 shows the referent biases as proportions between Source- and Goal-referential completions. As has been previously reported for English (Rohde et al., 2006), Elaboration and Explanation are highly Source-biased whereas Occasion and Result are highly Goal-biased.

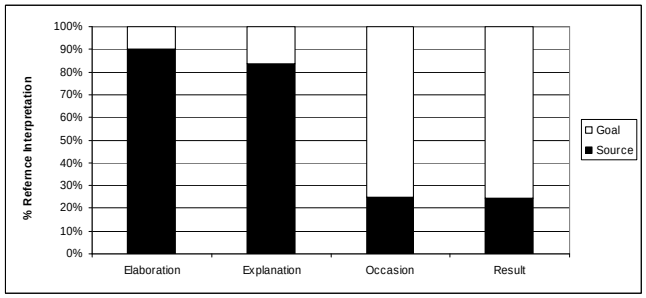


Figure 4: Proportion of Source/Goal referents for each coherence relation (collapsed over conditions).

For our statistical analysis, we collapsed the proportion of Elaboration and Explanation (Source-biased relations) on the one hand and Occasion and Result (Goal-biased relations) on the other hand for each participant's continuations, and conducted repeated measures ANOVAs on the proportion of Source-biased relations over Source- and Goal-biased relations. There was a significant main effect of prompt type [$F_1(2, 31) = 25.94$, $p<.0001$; $F_2(2, 59) = 22.34$, $p<.0001$; Tukey HSD: Null > Overt > Free by participants, and Null > Overt, Free by items], suggesting that Null prompt conditions were most Source-biased. There was also a significant main effect of aspect [$F_1(1, 31) = 9.75$, $p=.0018$; $F_2(1, 59) = 9.54$, $p=.0021$], suggesting that Imperfectives yielded more Source-biased relations than Perfectives (Figure 5). Pairwise comparisons between Imperfectives and Perfectives within each prompt type revealed significant differences for Overt pronoun prompt conditions [$t_1(31) = 3.68$, $p=.0009$; $t_2(59) = 3.82$, $p=.0003$] and marginal by-item significance for Free prompt

conditions [$t_1(31) = 1.64$, ns; $t_2(44) = 1.90$, $p=.0637$][8], but non significance for Null conditions.[9]

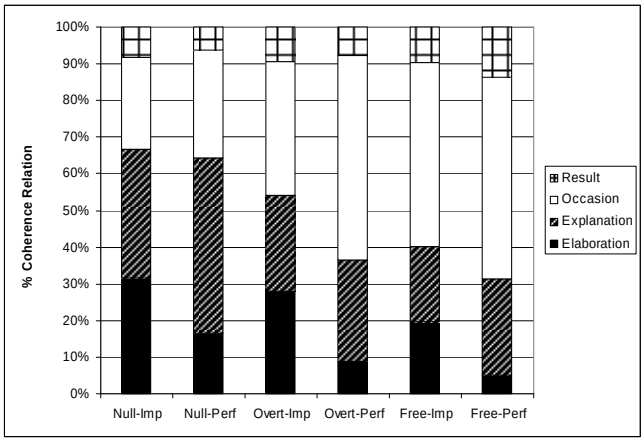


Figure 5: Proportion of coherence relations for all conditions (collapsed over topic-marking).

As was the case for reference, ANOVAs revealed no main effects or interactions involving topic marking. While Figure 6 indicates a small numerical trend of more Source-biased relations for topic-marked than nominative-marked Overt and Free continuations, pairwise comparisons between Topics and Nominatives within each prompt type revealed no significant or marginal differences.

In summary, the distribution of coherence relations generally followed the pattern found for reference, being consistent with previous studies in English (Rohde et al., 2006, 2008).

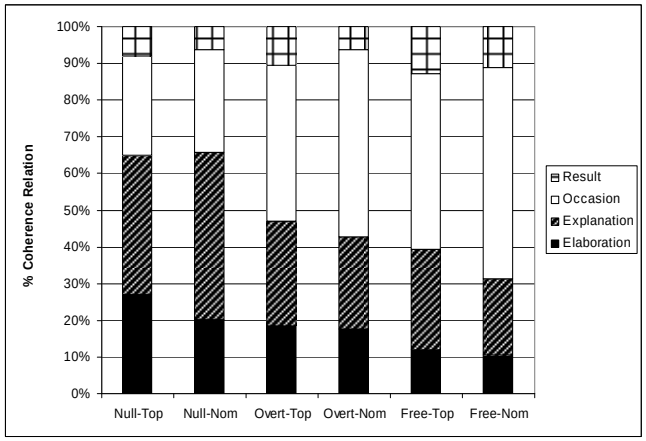


Figure 6: Proportion of coherence relations for all conditions (collapsed over aspect).

Discussion

We are now in a position to answer the four questions posed in the introduction to the paper. First, we asked whether the behavior of Japanese null and/or overt pronouns patterns with that of English overt pronouns in displaying sensitivity to pragmatic subsequent-mention biases, or whether their interpretation is determined primarily by other (e.g., grammatical) factors. The results indicate that Japanese null pronoun interpretation is not analogous to English overt pronoun interpretation as previous researchers have suggested. Instead, null pronouns were most strongly and uniformly Source-biased for both interpretation and coherence relations, apparently being driven predominantly by grammatical position and without showing sensitivity to the aspect manipulation. Instead, overt Japanese pronouns patterned with English in demonstrating such sensitivity, with Imperfective conditions yielding more Source referents and Source-biased coherence relations. Further, overt pronouns led to significantly more mentions of the Source than free prompts, demonstrating that, like English overt pronouns, Japanese overt pronouns overlay a subject bias on top of pragmatically-driven ones. Indeed, the results for Japanese overt pronouns mirrored those of Rohde et al. (2006, 2008) for English pronouns quite closely.

The second question we asked is whether null and overt Japanese pronouns exhibit a division-of-labor effect such that a demonstrated subject bias for null pronouns would correspond to a commensurate non-subject bias for overt pronouns. The answer is no; both null and overt pronouns displayed a subject bias, and hence their referents were not in complimentary distribution. Although the nature of the biases were different – overt pronouns overlay a subject bias on top of pragmatically-driven subsequent-mention biases as measured in the free prompt condition, whereas null pronouns appear to have a more grammaticalized subject bias that is impervious to pragmatic expectations – both pronominal forms were used to refer to Sources more often

[8] As was the case for reference, the mixed results for the Free prompt condition were due in part to the fact that the analysis included all continuations, as opposed to only those in which participants started their continuation with something other than a pronoun (i.e., a proper name). When only Name continuations were compared, there was a marginal difference between Imperfectives and Perfectives by participants [$t_1(20) = 1.98$, $p=.0619$] and a significant one by items [$t_2(32) = 2.05$, $p=.0483$].

[9] Posthoc observation revealed that the proportion of Elaborations within the Source-biased relations was consistently higher for Imperfectives than Perfectives for all prompt types. ANOVAs run on the proportion of Elaboration over Elaboration and Explanation relations revealed a significant main effect of aspect [$F_1(1, 31) = 28.29$, $p<.0001$; $F_2(1, 59) = 30.63$, $p<.0001$] with no other statistically-supported main effects or interactions. Imperfective conditions had a uniformly higher proportion of Elaboration than Explanation relations across prompt types; pairwise comparisons between Imperfectives and Perfectives within each prompt type revealed significant differences for all types except for Free prompts by subjects [Null: $t_1(23) = 2.42$, $p=.0240$; $t_2(39) = 2.54$, $p=.0151$; Overt: $t_1(24) = 3.36$, $p=.0026$; $t_2(39) = 3.06$, $p=.0040$; Free: $t_1(17) = 1.70$, $p = .1069$; $t_2(28) = 2.36$, $p=.0256$]. Participants were therefore more likely to elaborate an event described as ongoing (imperfective) than one described as completed (perfective), indicating an effect of aspect on coherence that is independent of the choice of subsequently mentioned entity.

than Goals. It therefore appears that the use of an overt pronoun does *not* implicate that the referent is an entity other than what the preferred referent would have been if a null pronoun had been used (i.e., the subject). At first blush, these patterns nonetheless appear consistent with those found for Spanish by Alonso-Ovalle et al. (2002), although further comparison is difficult since the experimental tasks and manipulations carried out were very different.

The third question we asked was whether topic-marked antecedents attract more pronominal references than subject-marked antecedents. The answer again was no. Perhaps surprisingly, there was no significant influence of topic marking across prompt types.

Lastly, we asked whether any grammatical biases that are revealed to be associated with these referential forms affects the distribution of ensuing coherence relations, as Rohde et al. (2008) found for English. This is clearly the case. Although the null and overt pronouns were always fully ambiguous between the available Source and Goal referents, their appearance in a prompt biased the continuation toward mentioning the previous subject referent first, which in turn biased the participants toward continuing the story using a Source-biased coherence relation. Further, while the aspect manipulation in the null pronoun condition created differences in the distribution of Source-biased relations – imperfectives resulted in a greater number of Elaborations, at the expense of Explanations (see footnote 9) – it did not change the allocation between Source- and Goal-biased relations, in accord with the fact that the aspect manipulation resulted in no difference in the distribution between Source and Goal referents.

Several experiments suggest themselves as ways of confirming the conclusions arrived at in this paper. One is to see whether the lack of effect of pragmatic bias for null pronouns holds across different verb types. Whereas we manipulated aspect on a single type (transfer of possession), we could also vary the verbs themselves, choosing types that are known to yield substantially different subsequent-mention biases. Contexts employing object-biased implicit causality verbs, for instance, would offer strong test for the subject bias associated with null pronouns. Likewise, the lack of effect of topic-marking could be further examined by comparing reference in contexts in which nominative-marked subject referents compete with topic-marked object referents. Such studies remain for future work.

Acknowledgments

This research was supported by a grant from the UCSD Academic Senate. We thank Shin Fukuda for helpful discussion and his insights in Japanese syntax, Ria Abe and Sho Nakamura for data annotation, and Ryo Goto, Emiko Nakamura, and Susanne Mari Sakai for help in conducting the experiment.

References

Alonso-Ovalle, L., Fernández-Solera, S., Frazier, L., & Clifton, C. (2002). Null vs. overt pronouns and the topic-focus articulation in Spanish. *Rivista di Linguistica, 14.2*, 1-19.

Amano, N., & Kondo, M. (2000). *NTT database series nihongo-no goikokusei: Lexical properties of Japanese* (Vol. 7). Tokyo: Sanseido.

Arnold, J.E. (2001). The effects of thematic roles on pronoun use and frequency of reference. *Discourse Processes, 31(2)*, 137-162.

Christianson, K., & Cho, H.Y. (2008). Interpreting null pronouns (*pro*) in isolated sentences. *Lingua, 119*, 989-1008.

Grice, H.P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Speech acts*. New York: Academic Press.

Gundel, J., Hedberg, N., & Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. *Language, 69*, 274-307.

Hobbs, J. R. (1990). *Literature and cognition*. CSLI Lecture Notes 21. Stanford, CA.

Kameyama, M. (1985). *Zero anaphora: the case of Japanese*. Doctoral dissertation, Stanford University.

Kehler, A. (2002) *Coherence, reference, and the theory of grammar*. CSLI Publications, Stanford, CA.

Kuno, S. (1973). *The Structure of the Japanese Language*. MIT Press, Cambridge, MA.

Kuroda, S-Y. (1965). *Generative grammatical studies in the Japanese language*. Doctoral dissertation, MIT.

Martin, S. (1976). *A reference grammar of Japanese*. Yale University Press.

Rohde, H. (2008). *Coherence driven effects in sentence and discourse processing*. Doctoral Dissertation. UCSD.

Rohde, H., Kehler, A., & Elman, J. (2006). Event Structure and Discourse Coherence Biases in Pronoun Interpretation. In *The Proceedings of the 28th Annual Conference of the Cognitive Science Society*, Vancouver, July 26-29, 2006.

Rohde, H. & Kehler, A. (2008). The bidirectional influence between coherence establishment and pronoun interpretation. Poster presented at the 21st Annual CUNY conference on Human Sentence Processing, March.

Stevenson, R., Crawley R., & Kleinman D. (1994). Thematic roles, focusing and the representation of events. *Language and Cognitive Processes, 9*, 519–548.

Ueno, M. & Polinsky, M. (2009). Does headedness affect processing? A new look at the VO-OV contrast. *Journal of Linguistics, 45*, 675-710.

Walker, M., Iida, M., & Cote, S. (1994). Japanese discourse and the process of centering. *Computational Linguistics, 20*, 193-232.

Novel Words in Novel Contexts: The Role of Distributional Information in Form-class Category Learning

Patricia A. Reeder (preeder@bcs.rochester.edu)
Elissa L. Newport (newport@bcs.rochester.edu)
Richard N. Aslin (aslin@cvs.rochester.edu)
Department of Brain & Cognitive Sciences, University of Rochester
Meliora Hall, Box 270268 Rochester, NY 14627 USA

Abstract

One major aspect of successful language acquisition is the ability to organize words into form class categories and generalize from properties of experienced items to novel items. Furthermore, learners must often determine how to use a new word, when there is very sparse information regarding its acceptable contexts. In this work we employ an artificial language learning paradigm to explore how adult learners, under circumstances of varying distributional cues to category boundaries, apply their knowledge of category properties to a new word. We find that in cases of strong category cues and strong category learning, adults readily generalize all of the distributional properties of the learned category to a word that shares just one context with the other category members. However, as the distributional cues regarding the target category become sparser and contain more systematic gaps, learners show more conservatism in generalizing the allowable distributional properties to the novel word. Taken together, these results show striking flexibility in learners' tendency to generalize, depending on the distributional properties of the input corpus, in a probabilistically rational way.

Introduction

The problem that learners face when they attempt to categorize items in the environment is deciding when they should treat instances as a category (thus generalizing from properties of experienced items to novel ones) and when they should treat instances separately (with no generalization from properties of experienced items to predicted properties of novel items). This problem cannot always be solved on the basis of perceptual similarity, as membership in some categories is independent of the surface features of the members.

The acquisition of grammatical categories is an example of this type of problem, but has some additional complicating factors. We hear individual words in a limited number of specific contexts. However, the rules that languages are built on involve patterns defined over categories of words, not the individual words themselves. Language input is serially presented, so we need to predict the proper contexts for words we have not yet heard. Furthermore, learners never see the entire input corpus, so they must figure out the proper contexts for new words, keeping in mind that sometimes there are lexically specific restrictions on words (such as *give* versus *donate*: despite similar meaning, Joe can *give David a book,* but Joe cannot **donate David a book*). In acquiring grammatical categories, the learner must ask whether contexts are absent by accident, or because they are ungrammatical. This question is particularly difficult to resolve when a new item is encountered in a single context and therefore overlaps only minimally with previously encountered words. For example, consider hearing the sentence: *I remembered to nerk yesterday*. Should one generalize from this context to another context where words of the category 'verb' are grammatical, such as *She will make him nerk tomorrow*, or *I saw the cat nerk earlier*?

One hypothesis about how learners handle this situation is that they have innately defined linguistic categories with featural and contextual information predefined, so that minimal exposure to language is needed to sort out which words belong to each category (e.g., McNeill, 1966). Another hypothesis is that learners use semantic categories to bootstrap the syntactic categories (e.g., Grimshaw, 1981). A third possibility is that learners exploit distributional information in the input to discover the category structure of natural languages (e.g., Braine, 1987). This third hypothesis is what we investigate in the present experiments.

A number of researchers have asked whether there is adequate distributional information in the input to form linguistic categories. This work uses hierarchical clustering and a computational learning mechanism to attempt to deduce grammatical categories from corpora of child-directed speech based solely on distributional analyses of the input (e.g., Mintz, Newport, & Bever, 2002; Redington, Chater, & Finch, 1998). These models have been able to use co-occurrence statistics among words to achieve relatively good categorization performance for frequent target words. To explore whether human learners can actually use this information during language learning, Mintz (2002) tested categorization in an artificial language learning environment, showing evidence that learners did engage in distributional analyses of the input in order to generalize their knowledge of previously encountered strings to grammatical novel strings. Hunt and Aslin (2010) showed that adults could learn categories embedded in sequences of visual symbols during a serial reaction time task when the only cue to category structure was distributional information among the symbol strings.

Building off of these findings on the importance of distributional information for category formation, we have proposed a systematic set of computational variables that can explain the types of distributional information that are

important for categorization. Deciding whether to generalize across words or preserve lexical specificity appears to be determined by (at least) 3 distributional variables: the *number* of linguistic contexts in which each word in the input set occurs, the *density* or proportion of these contexts that are present in the input, and the degree of *overlap* of contexts across words. In previous work (Reeder et al., 2009) we showed that learners are remarkably sensitive to these cues, which interact with each other to determine how basic category and subcategory structure are acquired. To do this, we manipulated the distribution of contexts for a target category in the exposure set to examine how adults determine when to generalize (deciding whether gaps in their input are accidental or systematic). When participants were exposed to a dense sampling of the language where there was rich coverage of contexts for a target category and high overlap in contexts across words, adult learners showed complete generalization to all possible grammatical contexts, even those that were never heard before for particular words. But as the input to the learner became more sparse with less overlap, participants became more conservative in their generalizations. Furthermore, as we increased the frequency of recurring gaps in the input, participants became more certain that the gaps were not accidental but rather part of the structure of the language, and they decreased their generalizations to unseen grammatical contexts. In the present work we ask how, under these same varying circumstances of category strength and category learning, learners will extend their knowledge of the target category to a *novel* word, one for which they have only minimal context information. In particular, is there a point in category learning where hearing one context for a novel word is enough to obtain full category privileges for that word? Or does every novel word need to be heard in a number of overlapping contexts in order to be treated as a member of the category?

Experiment 1

In Experiment 1 of Reeder et al. (2009), the learner was exposed to a very dense sampling of the language space, with all the words in the target category appearing in many highly overlapping contexts. Under these conditions, learners represented the words as a true category, generalizing fully across the gaps in the exposure corpus. In Experiment 1 we ask whether, under the same circumstances, the target category's distributional properties will also generalize to a novel word that they have only heard in a single context. The logic of this paradigm is that, if learners acquire a strong category (called X), then novel sentences which observe even a bit of the category structure of the language might be perceived to be just as grammatical (or familiar) as sentences that have actually been heard during training.

An artificial grammar with the structure (Q)AXB(R) was used, similar to that used in Reeder et al. (2009), where each letter represents a set of 2, 3, or 4 words. In Experiment 1, the Q and R categories had 2 words each, the A and B categories had 3 words each, and the X category had 4 words. The words of the grammar were *spad, klidum, flairb, daffin, glim, tomber, zub, lapal, fluggit, mawg, bleggin, gentif, frag,* and *sep.* The words were not mapped on to any referential world, so there were no semantic cues to categorization. All studies were run with two languages that differed only in which words were assigned to each of the categories in the language, to ensure that obtained results were not due to coincidental preferences for specific sound combinations. As in Reeder et al. (2009), X was the target category of interest, A and B were "context" categories that formed the distributional cues to the category X, and Q and R were optional flanker categories that allowed strings to range from 3 to 5 words in length.

Table 1: Possible AXB strings in Exp. 1-4. Items presented in Exp 1 are denoted *; items presented in Exp 2 are denoted ♦; items presented in Exp 3 & 4 are denoted ○.

$A_1X_1B_1$ *	$A_1X_2B_1$	$A_1X_3B_1$ * ♦ ○	$A_1X_4B_1$ * ♦ ○
$A_1X_1B_2$	$A_1X_2B_2$ * ♦	$A_1X_3B_2$ * ○	$A_1X_4B_2$
$A_1X_1B_3$ * ♦ ○	$A_1X_2B_3$ *	$A_1X_3B_3$	$A_1X_4B_3$
$A_2X_1B_1$	$A_2X_2B_1$ * ♦ ○	$A_2X_3B_1$ *	$A_2X_4B_1$
$A_2X_1B_2$ * ♦ ○	$A_2X_2B_2$ *	$A_2X_3B_2$	$A_2X_4B_2$
$A_2X_1B_3$ * ○	$A_2X_2B_3$	$A_2X_3B_3$ * ♦	$A_2X_4B_3$
$A_3X_1B_1$ * ♦	$A_3X_2B_1$ * ○	$A_3X_3B_1$	$A_3X_4B_1$
$A_3X_1B_2$ *	$A_3X_2B_2$	$A_3X_3B_2$ * ♦ ○	$A_3X_4B_2$
$A_3X_1B_3$	$A_3X_2B_3$ * ♦ ○	$A_3X_3B_3$ *	$A_3X_4B_3$

Method

Participants 16 monolingual native English-speaking students at the University of Rochester participated in Experiment 1, eight in each of the two languages created by different assignments of words to categories. Subjects had not participated in any other categorization experiment and were paid for their participation.

Stimulus Materials Of the possible 36 AXB sentence types in the language, 19 were presented to participants, and the remainder were withheld for testing generalization (see Table 1). The presence of the 2 Q and 2 R words was varied evenly such that the exposure set was expanded to 76 possible (Q)AXB(R) sentences. The exposure set contained only four X_4 strings: $A_1X_4B_1$, $Q_1A_1X_4B_1$, $A_1X_4B_1R_1$, and $Q_2A_1X_4B_1R_2$, which presented the X_4 word in only one context ($A_1X_4B_1$); the remaining 72 sentences included equal numbers of sentences containing X_1, X_2, and X_3. Training consisted of 4 times through this exposure set, forming 22 minutes of exposure. Importantly, every X_1, X_2, and X_3 was seen with every A and every B word, but X_4 was only seen in one context. Thus, the training set for Experiment 1 was dense for X_1-X_3 such that participants were exposed to a high proportion of the possible strings for those three X words, but very sparse for X_4. Additionally, there was complete overlap of contexts among X_1, X_2, and X_3, but X_4 shared only one context with X_1-X_3.

A female native English speaker recorded the words in isolation with both non-terminal and terminal intonation. Words were then adjusted in Praat such that pitch, volume, and duration were roughly consistent. Sentences were constructed by splicing words sequences in Sound Studio such that all words except the last had non-terminal intonation, with 50ms silence between each word. The final word in each sentence had terminal intonation contour. The order of sentences in the exposure set was randomized for each subject and presented via a custom software package on a Dell PC. Each sentence was separated by 1.5s of silence. Participants wore headphones and passively listened to the exposure sentences during training.

Immediately after exposure, participants heard a series of test strings and were asked to rate each on a scale from 1 to 5, where 1 meant it definitely did *not* come from the language they were exposed to, and 5 meant it definitely *did* come from the exposure language. All test strings were 3-word sentences of one of the following forms: a grammatical familiar string (10 AXB strings presented during training), a grammatical novel string (13 AXB strings withheld during training), or an ungrammatical string (of the form AXA or BXB). Of the grammatical novel test strings, 4 of the 13 were strings testing generalization of X_4: $A_2X_4B_2$, $A_2X_4B_3$, $A_3X_4B_2$, and $A_3X_4B_3$. With these strings we can ask whether learners have generalized X_4 to the full range of grammatical contexts for X-words, judging the familiar and novel grammatical sentences for X_4 to be equivalent, even though they have only seen X_4 in one of these contexts. These strings can then be compared to the 6 ungrammatical strings that contain X_4 (3 AX_4A, 3 BX_4B).

Results

A repeated measures ANOVA with condition (familiar, novel, ungrammatical) as the within subjects factor and language as the between subjects factor showed no significant effects of language (F<1). For test items without X_4, the mean rating of grammatical novel strings was 3.87 (SE=0.14), the mean rating of grammatical familiar strings was 3.85 (SE=0.13), and the mean rating of ungrammatical strings was 2.89 (SE=0.15). We found no significant difference between ratings of grammatical novel items and grammatical familiar items (F(1,14)=0.24, p=0.63). These items were rated significantly higher than ungrammatical test strings (F(1,14)=26.40, p<0.005). For the test items that contained X_4, the mean rating of grammatical novel strings was 3.28 (SE=0.18), the mean rating of grammatical familiar strings was 3.59 (SE=0.24), and the mean rating of ungrammatical strings was 2.61 (SE=0.21). These items showed the same pattern as the without-X_4 items: there was no significant difference between ratings of grammatical novel X_4 items and familiar X_4 items (F(1,14)=1.71, p=0.21), however there was a significant difference between these items and ungrammatical X_4 strings (F(1,14)=13.10, p<0.01). [1]

Discussion

As in Reeder et al. (2009), learners strongly preferred familiar and novel grammatical sentences to ungrammatical sentences. Learners also showed generalization to the novel grammatical X_4 strings, but not to the ungrammatical X_4 strings. Thus they generalized X_4 to the full range of grammatical contexts for X words, even though they heard X_4 in only one of these contexts. These results show that, when learners are exposed to a dense sampling of the language space for words in the target category (X_1-X_3) and presented with many overlapping contexts, they generalize their knowledge within the category X_1-X_3 and also extend it to X_4. Importantly, the generalized contexts are novel contexts for X_4, but are strongly represented by the learner's exposure to the permissible contexts for X_1-X_3. Learners did not require semantic or perceptual cues to indicate that the X words form a category.

Experiment 1 provided the learner with a dense sampling of the language space for most of the words in the target category. In the remaining experiments we systematically manipulated the density, overlap, and number of contexts for X_1-X_3 in the exposure set while restricting exposure to contexts for X_4, in order to explore the impact of these distributional variables on the generalization of category knowledge.

Experiment 2: Sparseness

In Experiment 2, we decrease the density of the contexts for X_1-X_3 words, but we keep the number and overlap among X_1-X_3 contexts the same. We still present only one context for X_4 and explore what the increase in sparseness for X_1-X_3 does to learners' generalizations to the novel X_4 item.

Method

Participants 16 monolingual native English-speaking students at the University of Rochester participated in Experiment 2 for payment, eight in each of the two possible languages. Subjects had not participated in any other categorization experiment.

Stimulus Materials The strings of the language were constructed in the same manner as Experiment 1, with two languages that had different assignments of words to categories. Here, however, the exposure set contained only 10 (versus 19 in Exp. 1) of the 36 possible AXB combinations (see Table 1). As in Experiment 1, every X_1-X_3 word was heard in combination with every A and every B. With the addition of AXB strings with optional Q and R

[1] We did not compare ratings of the X_1-X_3 test items with the X_4 items because of the lower statistical power of the X_4 means. For all experiments, we take the pattern of learning for familiar and novel grammatical items to be more informative than the size of the differences between X_1-X_3 and X_4.

flanker words, there were 40 sentences in the exposure set. The exposure set was repeated 4 times through so that each sentence type was presented with the same frequency as in Experiment 1, for an exposure of about 12 minutes. The test phase was the same as described for Experiment 1.

Procedure The procedure was the same as in Experiment 1.

Results

A repeated measures ANOVA with condition as the within subjects factor and language as the between subjects factor revealed no difference between the two languages (F<1). For test items without X_4, the mean rating of grammatical novel strings was 3.55 (SE=0.09), the mean rating of grammatical familiar strings was 3.54 (SE=0.10), and the mean rating of ungrammatical strings was 2.63 (SE=0.14). Just as in Experiment 1, as well as Experiments 1 and 2 from Reeder et al. (2009), we found no significant difference between ratings of grammatical novel items and grammatical familiar items without X_4 (F(1,14)=0.008, p=0.93), but grammatical sentences were rated significantly higher than ungrammatical test strings (F(1,14)=25.37, p<0.001). For the test items that contained X_4, the mean rating of grammatical novel strings was 3.27 (SE=0.15), the mean rating of grammatical familiar strings was 3.53 (SE=0.22), and the mean rating of ungrammatical strings was 2.55 (SE=0.16). This is the same trend as demonstrated by the without-X_4 items and the analyses in Experiment 1. While there was a significant difference between grammatical X_4 strings and ungrammatical X_4 strings (F(1,14)=9.87, p<0.01), there was no significant difference between ratings of grammatical novel X_4 items and familiar X_4 items (F(1,14)=1.59, p=0.23).

Discussion

These results mirror those in Experiment 1, demonstrating that reduced density does not greatly affect learners' performance when there is full overlap of contexts among X_1-X_3 words. The generalization to X_4 is maintained despite greatly reduced exposure due to a sparser sampling of the language space. We next explore how learners behave when there is reduced overlap of X_1-X_3 word contexts.

Experiment 3: Overlap

Similar to Experiment 2, we present the learner with only 10 of the 36 possible AXB combinations. However, in order to test how overlap in contexts influences generalization of category knowledge to new X-words, we now reduce the overlap of contexts among members of X_1-X_3. Individual X-words do not fully share all of their contexts with other X-words, though the set of X-words as a whole occurs in all A and B contexts. By reducing the overlap in contexts across X words, we can assess the degree to which learners restrict generalization within X_1-X_3, and also how they extend the category knowledge to X_4.

Method

Participants 16 monolingual native English-speaking students at the University of Rochester participated in Experiment 3, eight in each of the two possible languages. Participants had not been in any other categorization experiment and were paid for their participation.

Stimulus Materials Strings were assembled in the same way as Experiment 1, with two languages that had different assignments of words to categories. Exposure consisted of only 10 of the 36 possible AXB combinations, as in Experiment 2; however now X_1, X_2, and X_3 were heard with 2 of the 3 A-words and 2 of the 3 B-words each. X_1 occurred with A_1, A_2, B_1, and B_2, but not A_3 or B_3; X_2 was heard with A_2, A_3, B_2, and B_3, but not A_1 or B_1; X_3 was heard with A_1, A_3, B_1, and B_3, but not A_2 or B_2. Thus, the overlap among contexts is maintained over the X_1-X_3 category as a whole, but individual X-words do not have the degree and type of overlap in distributional contexts that they do in Experiments 1 and 2, where each X word occurs with every A and every B. X_4 was still only seen with one context (see Table 1).

Procedure The procedure was the same as in Experiment 1.

Results

A repeated measures ANOVA with condition as the within subjects factor and language as the between subjects factor showed no significant difference between the two languages (F<1). For test items without X_4, the mean rating of grammatical novel strings was 3.71 (SE=0.12), the mean rating of grammatical familiar strings was 3.91 (SE=0.09), and the mean rating of ungrammatical strings was 2.55 (SE=0.15). Unlike Experiments 1 and 2, but in line with results from Reeder et al. (2009), we found significant differences between ratings of grammatical novel items and grammatical familiar items (F(1,14)=9.12, p<0.01). Additionally, both of these items were rated significantly different from ungrammatical test strings (F(1,14)=26.82, p<0.001). For the test items that contained X_4, the mean rating of grammatical novel strings was 3.25 (SE=0.16), the mean rating of grammatical familiar strings was 3.66 (SE=0.24), and the mean rating of ungrammatical strings was 2.21 (SE=0.16). Unlike the without- X_4 items, we do not see any significant difference between novel grammatical X_4 strings and familiar X_4 strings (F(1,14)=2.98, p=0.11), perhaps due to the lower statistical power for these test items; there is still a significant difference between ratings of grammatical and ungrammatical X_4 items (F(1,14)=26.21, p<0.001).

Discussion

In Experiment 3, we reduced the overlap among contexts in the exposure set by a third, but we kept the number of contexts in the input the same as in Experiment 2. The results indicate that despite full coverage of contexts across lexical items, the incomplete overlap between X_1-X_3-words

led to decreased generalization. However, learners still showed a much higher rating for grammatical novel items than ungrammatical items, indicating that they were still willing to generalize, though more conservatively than in Experiments 1 and 2. Additionally, learners were much less likely to generalize their knowledge of grammatical X_1- X_3 contexts to X_4 given the systematic gaps in the Experiment 3 exposure set. Thus, as we move along the dimensions of sparseness and overlap explored in Experiments 2 and 3, we can see how learners weigh the likelihood that X_4 shares the same contexts as X_1- X_3 and use this as a diagnostic for how strongly the X category has been formed.

Experiment 4: Overlap with extended exposure

The decision to generalize over a gap in the input or maintain lexical distinctness may also be influenced by the frequency of contexts (and gaps) in the input. If a context is consistently absent as in Experiment 3, learners start to show conservatism in their generalizations. If this gap is made even more prominent by creating an exposure set that has repeated instances of sparse contextual information, learners might develop even more certainty that gaps in the input are systematic and not accidental (e.g., Wonnacott, Newport & Tanenhaus, 2008; Xu & Tenenbaum, 2007). This will be particularly important with regard to X_4, where we can explore how an increase in the exposure to the one context for X_4 (and potentially a perceived increase also in the gaps at the non-occurring contexts for X_4) affects how learners generalize their knowledge of the category X_1- X_3. If the category X_1- X_3 is strongly defined (as in Experiment 1), we would expect that a very large increase in frequency of the one context of X_4 (and perceived increase in exposure to gaps for X_4) might be required before there is a decrease in generalization and a lessening of X_4 membership in the X-word category. However, if the X-category is weakly defined as in Experiment 3, the small increase in the number of repetitions in Experiment 4 might be enough to make learners conservative in their generalizations.

Method

Participants 16 monolingual native English-speaking students at the University of Rochester participated in Experiment 2, eight in each of the two possible languages. Participants had not been in any other categorization experiment and were paid for their participation.

Stimulus Materials The language was the same as in Experiment 3, except that exposure to the language was tripled by presenting the corpus 12 times rather than 4. Training lasted for approximately 22 minutes (as in Experiment 1), but contained only 10 contexts (as in Experiments 2 & 3). Test strings were the same as in Experiment 3.

Procedure The procedure was the same as in Experiment 1.

Results

A repeated measures ANOVA with condition as the within subjects factor and language as the between subjects factor showed no significant difference between the two languages (F<1). For test items without X_4, the mean rating of grammatical novel strings was 3.86 (SE=0.12), the mean rating of grammatical familiar strings was 4.05 (SE=0.10), and the mean rating of ungrammatical strings was 2.61 (SE=0.21). These results show a significant difference between ratings of grammatical novel items and grammatical familiar items (F(1,14)=8.60, p=0.01). Additionally, these items were rated significantly higher than ungrammatical test strings (F(1,14)=35.83, p<0.001). For the test items that contained X_4, the mean rating of grammatical novel strings was 3.44 (SE=0.19), the mean rating of grammatical familiar strings was 4.06 (SE=0.21), and the mean rating of ungrammatical strings was 2.37 (SE=0.21). Similar to the without-X_4 items, we now find a significant difference between novel grammatical X_4 strings and familiar X_4 strings (F(1,14)=8.33, p=0.011), along with a significant difference between these and ungrammatical X_4 items (F(1,14)=31.04, p<0.001).

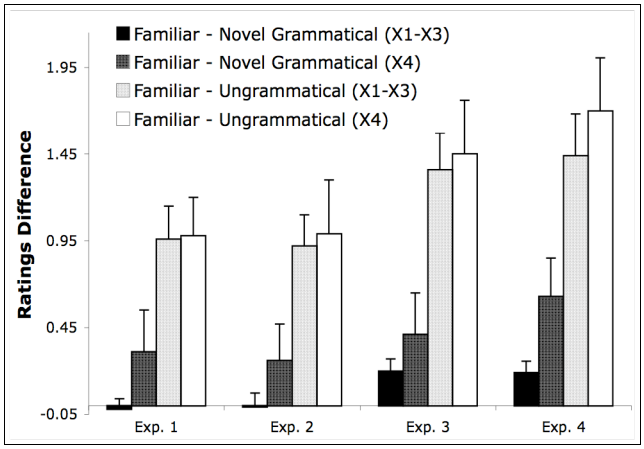


Figure 1: Experiment 1-4 difference scores of ratings of grammatical familiar items and grammatical novel items (for X_1-X_3 words and X_4), and grammatical familiar items and ungrammatical items (for X_1-X_3 words and X_4).

Discussion

These results indicate that, when we increase exposure to the same sparse data (with recurring gaps that may also become more prominent), learners act rationally and are even less likely to generalize over such gaps. Furthermore, learners apparently view the category formed by X_1- X_3 as weakly defined due to the sparse sampling of the language and incomplete overlap among words, which also seems to increase learners' uncertainty about the status of the withheld grammatical X_4 items. While we still see that novel grammatical test strings are judged more grammatical than the ungrammatical strings, we hypothesize that increasing exposure to the sparse input set even longer might push learners to judge all novel items as ungrammatical. In contrast, if we increased the number of

unsystematic gaps in the input, we expect that learners would show more generalization, especially for the X_4 word.

General Discussion

The present experiments add grammatical category learning to a large literature showing that learners are highly sensitive to many types of distributional information in their input. We have replicated Experiments 1-4 of Reeder et al. (2009), demonstrating that learners are able to extract the category structure of an artificial language based on distributional information alone, and we show that learners are quite rational, statistically speaking, in how much and when they generalize across gaps in the input. Importantly, the current experiments also show that learners can skillfully transfer their knowledge of category structure and category cues to a novel item that is only weakly represented in the input. When given a dense sampling of the language space with almost complete overlap of contexts for many words in a target category X, learners generalize a novel word (X_4) to the full range of grammatical contexts of the other X-words, even when they have only seen X_4 in one of those contexts. This willingness to add X_4 to the strongly established X_1- X_3 category is strongest when the X_1- X_3 contexts are dense and overlapping; when contexts are more sparse and less overlapping across different X words, we also see more conservative generalization to a new X_4 word. The most extreme case is when we increase the number of times the learner hears the sparse exposure set, thus increasing also the frequency of recurring gaps in the input for X_1- X_3: learners in this situation rate the withheld X_4 contexts as more unfamiliar, while rating as highly familiar only the one context in which X_4 was actually heard. These findings are in line with results from Wonnacott, Newport and Tanenhaus (2008) in the area of verb-argument learning, where if the language is generally lexically specific, participants do not show generalization of the minimal exposure item (i.e., X_4) to other contexts. In contrast, if the language has the same contexts permitted for all verbs, then participants show strong generalization for the minimal exposure item.

We are in the process of modeling these results to determine the type of information learners might encode in order to accomplish these outcomes; storing any simple statistics – such as word, bigram, or trigram frequencies – would not be adequate to account for generalization to the novel X_4 strings. Instead, learners must be forming a more abstract representation of the data in order to generalize their knowledge to novel strings.

In contrast to our experiments, as learners face the problem of inferring category membership from sparse and incomplete data in natural languages, there are a number of correlated cues that they could use to help them extract category information, such as phonological, prosodic, or semantic cues as well as distributional cues. Indeed, many studies have shown that category learning is enhanced when category membership is correlated with such surface cues (e.g., Monaghan, Chater, & Christiansen, 2005). But an important question in this literature has been whether category learning can utilize distributional information, either alone or when very poorly correlated with other cues. While natural languages do sometimes contain multiple cues to grammatical categories, our work indicates that learners are able to skillfully employ a statistical learning mechanism as a primary tool with which to extract category information from the input, even in cases where other correlated cues are incomplete or absent.

Acknowledgments

We would like to thank Josh Tenenbaum for valuable discussions of this work. This research was supported by NIH Grants HD037082 to RNA and DC00167 to ELN, and by an ONR Grant to the University of Rochester.

References

Braine, M.D.S. (1987). What is learned in acquiring word classes – A step toward an acquisition theory. In B. MacWhinney (Ed.), *Mechanisms of language acquisition.* (pp. 65-87). Hillsdale, NJ: Lawrence Erlbaum Associates.

Grimshaw, J. (1981). Form, function, and the language acquisition device. In C.L. Baker and J.J. McCarthy (Eds.), *The logical problem of language acquisition*. (pp. 183-210). Cambridge, MA: MIT Press.

Hunt, R.H., & Aslin, R.N. (2010). Category induction via distributional analysis: Evidence from a serial reaction time task. *Journal of Memory and Language*, *62,* 98-112.

McNeill, D. (1966). Developmental Psycholinguistics. In F. Smith & G. Miller (Eds.), *The Genesis of Language: A Psycholinguistics Approach* (pp. 69-73). Cambridge, MA: MIT Press.

Mintz, T.H. (2002). Category induction from distributional cues in an artificial language. *Memory and Cognition, 30,* 678-686.

Mintz, T.H., Newport, E.L., & Bever, T.G. (2002). The distributional structure of grammatical categories in speech to young children. *Cognitive Science, 26*, 393-425.

Monaghan, P., Chater, N., & Christiansen, M. (2005). The differential role of phonological and distributional cues in grammatical categorization. *Cognition, 96,* 143-182.

Redington, M., Chater, N., & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science, 22,* 435-469.

Reeder, P.A., Newport, E.L., & Aslin, R.N. (2009). The role of distributional information in linguistic category formation. In N.A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 2564-2569). Austin, TX: Cognitive Science Society.

Wonnacott, E., Newport, E.L., & Tanenhaus, M.K. (2008). Acquiring and processing verb argument structure: Distributional learning in a miniature language. *Cognitive Psychology, 51,* 165-209.

Xu, F., & Tenenbaum, J.B. (2007). Word learning as Bayesian inference. *Psychological Review, 114,* 245-272.

Optimal Language Learning: The Importance of Starting Representative

Anna N. Rafferty (rafferty@cs.berkeley.edu)
Computer Science Division, University of California, Berkeley, CA 94720 USA

Thomas L. Griffiths (tom_griffiths@berkeley.edu)
Department of Psychology, University of California, Berkeley, CA 94720 USA

Abstract

Child-directed speech has a distinctive structure and may have facilitatory effects on children's language learning. We consider these facilitatory effects from the perspective of Marr's levels of analysis: could they arise at the computational level or must they be located at at the algorithmic or implementation levels? To determine if the effects could be due to computational level benefits, we examine the question of what samples from a language should best facilitate learning by identifying the optimal linguistic input for an ideal Bayesian learner. Our analysis leads to a mathematical definition of the "representativeness" of linguistic data, which can be computed for any probabilistic model of language learning. We use this measure to re-examine the debate over whether language learning can be improved by "starting small" (i.e. learning from data that have limited complexity). We compare the representativeness of corpora with differing levels of complexity, showing that while optimal corpora for a complex language are also complex, it is possible to construct relatively good corpora with limited complexity. We discuss the implications of these results for the level of analysis at which a benefit of starting small must be located.
Keywords: language learning; child-directed speech; Bayesian models; representativeness; starting small

Introduction

Child-directed speech is an important source of information for children's language acquisition. Hoff and Naigles (2002) found that the amount of child-directed speech produced by mothers was predictive of the vocabulary of their children, and Cameron-Faulkner, Lieven, and Tomasello (2003) found correlations between the grammatical frames mothers used in speech to their children and the grammatical frames used by the children. Child-directed speech also differs from adult-directed speech in a number of ways. For example, Snow (1972) found that speech to two year olds by caregivers has simplified structure and involves more repetitions than speech to older children or adults, and Sherrod, Friedman, Crawley, Drake, and Devieux (1977) found that the mean length of utterances spoken to a child changed in response to changes in the child's understanding. Overall, child-directed speech tends to be simplified, more grammatically correct, and more repetitive than adult-directed speech (Pine, 1994). This raises an important question: Does the structure of child-directed speech facilitate language acquisition?

There is some evidence for a facilitatory effect of child-directed speech. Furrow, Nelson, and Benedict (1979) found that children's language development was positively correlated with mothers' use of simple constructions, and Newport, Gleitman, and Gleitman (1977) found that acquisition of certain syntactic features was facilitated by characteristics of mothers' speech, such as placement of particular syntactic structures early in sentences. However, Newport et al. (1977) also found that many measures of acquisition were unaffected by characteristics of caregivers' speech, and Huttenlocher, Vasilyeva, Cymerman, and Levine (2002) found that exposing children to more complex speech resulted in the children using more complex syntax.

Previous work has used specific computational models such as associative learning and artificial neural networks to explore the effects of simplified input on language learning (Goldowsky & Newport, 1993; Elman, 1993; Rohde & Plaut, 1999). Elman (1993) found that training a simple recurrent neural network to predict the next word in a sequence using a corpus of limited complexity resulted in better generalization than beginning with the full corpus. However, the effects of "starting small" are far from clear: Rohde and Plaut (1999) subsequently found a disadvantage for language learning that begins with data of limited complexity when using similar models and corpora.

Demonstrating an effect of starting small under specific assumptions about learning leaves open the question of the level of analysis at which there might be an advantage for child-directed speech. Marr (1982) defined three levels at which information processing systems can be analyzed: the *computational* level, where the analysis aims to identify the abstract problem being solved and its ideal solution; the *algorithmic* level, where the focus is on the representation and algorithm being used to implement this solution; and the *implementation* level, which emphasizes the physical hardware on which the algorithm is executed. Facilitatory effects of the structure of child-directed speech could be caused by considerations at any of these levels. At the computational level, data of this kind could provide more statistical evidence for the structure of the language. Alternatively, constraints at the algorithmic or implementation levels might limit the information-processing capacities of children, making simplified input necessary despite the lack of a computational level benefit.

We try to identify the level of analysis at which a benefit from simplified input could be located by asking what characteristics a sample of language should have in order to be most useful for an ideal learner. If simpler corpora are better for this ideal learner, then we can provide a computational-level account of the benefit of starting small. If not, such an effect must be located at a lower level. It is necessary to consider the performance of ideal learners in order to rule out the possibility that starting small provides a computational-level advantage. If this were the case, it would not be necessary to assume algorithmic level constraints are the cause of an

advantage for starting small, as has been done in previous research.

We identify the optimal input for an ideal Bayesian language learner by asking what data maximize the posterior probability such a learner ascribes to the target language. This is a special case of the problem of defining a "representative" sample analyzed by Tenenbaum and Griffiths (2001). Consequently, we define a Bayesian measure of representativeness, and use this measure to give a mathematical characterization of an optimal corpus. We present a general mathematical result characterizing representativeness for discrete probability distributions, which are the basic component of any probabilistic model of language. This result provides the basis for a more detailed exploration of whether language of limited complexity might be as good or better for learning than language of full complexity. We explore the implications of this result by identifying the optimal input for four different learning scenarios, involving estimating probabilistic grammars with varying degrees of knowledge about the structure of a language and estimating n-gram models.

Identifying Optimal Corpora

To understand the characteristics of an optimal sample of language, we formalize the problem of language learning in terms of Bayesian inference. Learning a probabilistic model of language requires estimating the value of a set of parameters θ from observed linguistic data d. Assuming the learner has some initial beliefs about the value of θ, expressed through a *prior* probability distribution $p(\theta)$, the beliefs of a rational learner after observing d are given by the *posterior* distribution $p(\theta|d)$ obtained by applying Bayes' rule,

$$p(\theta|d) = \frac{p(d|\theta)p(\theta)}{\int p(d|\theta)p(\theta)d\theta} \quad (1)$$

where the *likelihood* $p(d|\theta)$ indicates the probability of d under the probabilistic model with parameters θ.

A Measure of Representativeness

Formalizing language learning in this way now allows us to consider what corpora will most strongly support learning. Assume that the true structure of the language is characterized by parameters θ^*; we consider a learner that is simply learning this structure, although more complicated models that also learn other parts of the language, such as semantics, are possible. To maximize the probability of a learner inferring θ^* over other values of θ, a teacher should provide data d that maximize $p(\theta^*|d)$. Examination of the right hand side of Equation 1 shows that this can be done by maximizing

$$R(d,\theta^*) = \frac{p(d|\theta^*)}{\int p(d|\theta)p(\theta)d\theta} \quad (2)$$

with respect to d, as the prior probability $p(\theta^*)$ is constant and thus unaffected by the choice of d. Tenenbaum and Griffiths (2001) suggested that $R(d,\theta^*)$ be considered a measure of the "representativeness" of d relative to θ^*, being an indicator of the strength of evidence that d provides in favor of θ^* relative to other values of θ. Intuitively, a sample is more representative if it is both very probable under the true model (the numerator of Equation 2) and not as probable under a model selected at random (the denominator of Equation 2).

Representativeness for Discrete Distributions

In general, we may not be able to solve the integral in the denominator of Equation 2 exactly. However, we can solve this integral in the case where the model $p(d|\theta)$ is a discrete probability distribution, as is often true with probabilistic models of language. For a multinomial with ordered outcomes, the likelihood is $p(d|\theta) = \prod_{i=1}^{t}(\theta_i^*)^{k_i}$, where t is the number of possible outcomes, θ_i^* is the probability of outcome i, and k_i is the number of times the outcome i occurred. We place a uniform Dirichlet prior on the distribution θ, reflecting no strong expectations about the probabilities of different rules. Thus, the integral in Equation 2 is in this case:

$$\int_\Delta \prod_{i=1}^{t} \theta_i^{k_i} d\theta = \frac{(\prod_{i=1}^{t} k_i!)}{(t-1+\sum_{i=1}^{t} k_i)!} = \frac{(\prod_{i=1}^{t} k_i!)}{(t-1+n)!} \quad (3)$$

where Δ is the simplex of values such that $\sum_{i=1}^{t}\theta_i = 1$, and n is the total number of observations. The representativeness of a corpus with respect to this model with a particular value of θ is then:

$$R(d,\theta) = \frac{(t-1+n)!\prod_{i=1}^{t}\theta_i^{k_i}}{(\prod_{i=1}^{t} k_i!)}. \quad (4)$$

The optimal corpus is that which maximizes this quantity.

We can find an exact expression for the frequencies an optimal corpus would have by maximizing the quantity in Equation 4 with respect to k_i. Since the logarithm is monotonic, the corpus that maximizes $R(d,\theta)$ is also the corpus that maximizes $\log R(d,\theta)$, so we perform our maximization with this transform. Additionally, this is a constrained optimized problem since n must equal $\sum_{i=1}^{t} k_i$. We enforce this constraint with a Lagrange multiplier, and replace the factorials using Stirling's approximation to obtain the objective function:

$$L = (t-1+n)\log(t-1+n) + 1 - t + \sum_{i=1}^{t} k_i\log(\theta_i) - k_i\log(k_i) + \lambda(n - \sum_{i=1}^{t} k_i) \quad (5)$$

where λ is the Lagrange multiplier. To determine the optimum of this objective function, we differentiate with respect to k_i, set the derivative to zero, and solve for k_i. This shows that the optimal value is $k_i = n\theta_i$. Rounding to the nearest integer, this corresponds to what one might intuitively expect: The most representative corpus is that in which the relative frequencies of the outcomes match their probabilities under the target multinomial.

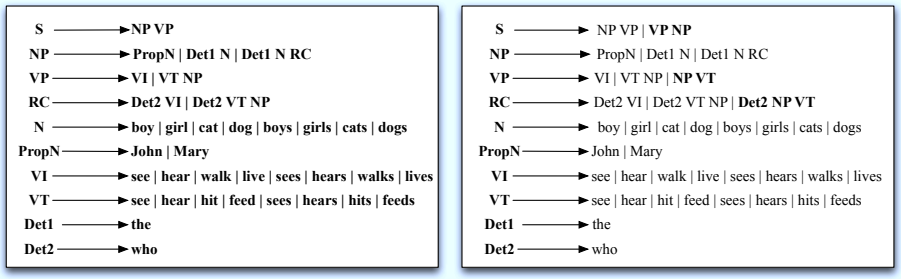


Figure 1: The context-free grammars used in our simulations. On the left, the true grammar; on the right, the overly general grammar with added rules, used in the third simulation. Bolded expansions are those present in the expanded grammar but not in the true grammar. In addition to these rules, subject-verb agreement is enforced, resulting in a much larger PCFG.

Representativeness for Probabilistic Grammars

The results for the representativeness of samples from multinomial distributions can be used to characterize optimal corpora for any probabilistic language model with discrete elements, such as an n-gram model. These results also generalize naturally to a representativeness measure for more sophisticated probabilistic models of language, such as probabilistic grammars. A *probabilistic context-free grammar* (PCFG; Baker, 1979) defines a probability distribution over sentences via a set of expansion rules for non-terminals (e.g. a noun phrase consists of a determiner followed by a noun) and distributions over those rules indicating the probability of a given non-terminal being expanded to a particular sequence (see Figure 1). The distributions over rules are independent multinomials, allowing us to build on the representativeness analysis above. In this case, the parameters θ describe the multinomial distributions associated with each expansion rule.

When the structure of sentences (i.e. the sequence of expansion rules used in generating each sentence) is known, the representativeness of a corpus follows directly from our result for multinomials. Since each rule is associated with an independent multinomial, the representativeness is the product of the representativeness for each multinomial. Thus, a representative corpus is one in which the relative frequencies with which expansion rules are used match the probabilities associated with those expansion rules in the grammar.

When the structure of sentences is unknown, $p(d|\theta)$ is obtained by marginalizing over possible structures. For PCFGs, this can be done efficiently using a dynamic program; in our simulations, we used Mark Johnson's implementation of this algorithm.[1] However, since there is not a closed form for this marginalization, we cannot calculate the denominator of Equation 2 exactly. In this case, we can use a Monte Carlo method to approximate the integral and obtain an estimate of the representativeness of a corpus.

Starting Small

As described in Elman (1993), "starting small" involves showing a learner only a limited number of "complex" sentences from a language first, and gradually exposing the learner to the full language. A sentence is complex if it contains a recursive rule; for example, in both Elman (1993) and Rohde and Plaut (1999), complex sentences are those that contain relative clauses. Both Elman (1993) and Rohde and Plaut (1999) used neural networks that learned to predict the next word in the sentence. Elman (1993) found that starting small was essential for his model; when a corpus of full complexity was used, the learner was never able to predict the next word with satisfactory accuracy. Rohde and Plaut (1999) found, in contrast, that in most cases starting small resulted negative impacts on performance, and none of their simulations showed any advantage to starting small.

We use the analysis of representativeness for an ideal language learner given in the previous section to explore the locus of a potential effect of starting small. Since our analysis focuses solely on the statistical evidence a corpus provides in favor of a particular language, we can examine whether a potential benefit of starting small could arise at the computational level, or must be a consequence of specific information-processing constraints associated with human learning. Thus, we consider two questions: First, does starting small result in particularly good corpora for language learning? And second, can a corpus of limited complexity be as good as a corpus without limited complexity? Clearly, if a starting small corpus is optimal, then such a corpus is as or more representative than a more complex corpus. However, even if a limited complexity corpus is non-optimal, it might be as representative as corpora generated by other means. In particular, we compare corpora of different complexity generated by maximizing representativeness and generated randomly.

As in the analysis in the previous section, we consider two types of corpora: those in which sentence structure is known and those in which structure is unknown. In two simulations, we assume that the learner knows the rules of the grammar, but does not know the frequencies with which they occur. Our third simulation introduces ambiguity about the rules of the grammar, and the fourth considers the possibility that children are not learning a grammar but simply distributions over which words follow one another. We used a PCFG similar to that in Elman (1991). The only instance of recursion was in the relative clause, which occurred in 75% of sentences generated from the grammar, and the grammar enforced the

[1] Version last updated 2 September 20007, and available at http://www.cog.brown.edu/∼mj/Software.htm

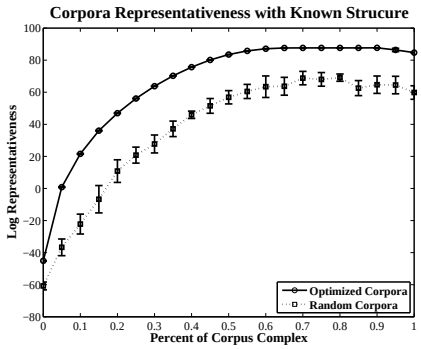


Figure 2: Representativeness of corpora with known structure. As the number of complex sentences increases, the representativeness of the corpora increases non-linearly.

agreement of subjects and verbs[2]. Figure 1 shows the grammar prior to integrating the constraint of verbal agreement; the final grammar consisted of 63 rules and 23 nonterminals.

Representativeness with Known Structure

We first considered the problem of learning from a corpus in which the structures of the sentences are known, allowing us to use the closed form given in Equation (4) to exactly compute the representativeness of the corpus. We sought to quantify how representative a corpus could be given the constraint on complexity and discover how this compared to randomly generated corpora as well as more complex corpora.

To investigate this question, we generated several types of corpora. All corpora were created by choosing a subset of sentences from a large corpus generated by the grammar. *Random* corpora were generated by selecting this subset randomly, subject to a constraint on the number of complex sentences. *Optimized* corpora were collections of sentences chosen to maximize representativeness. An ε-greedy perturbation process was used to maximize representativeness. First, an initial corpus of the target complexity was randomly selected. This corpus was perturbed by adding additional sentences, and then pruning sentences from the augmented corpus. With small probability, a sentence was chosen randomly to add or prune. Otherwise, a sentence was chosen by checking the effect of adding or pruning each possible sentence and greedily adding or pruning the sentence that resulted in the corpus with the largest representativeness. Twenty perturbations of ten sentences each were performed; results were not sensitive to small variations in these parameters.

For both the random and optimized conditions, we created corpora with constrained complexity. Corpora were generated with complexity ranging from 0% to 100% complex sentences, at 5% intervals. A complex sentence was any sentence containing a relative clause. Additionally, random and optimized corpora were generated with no complexity constraint. Each corpus contained 100 sentences.

As shown in Figure 2, this procedure succeeds in finding subsets of sentences that are significantly more representative than randomly generated corpora of the same complexity. However, the limitation on complexity greatly affects representativeness. While limiting complexity significantly impacts the representativeness of only one rule, that which allows the introduction of the relative clause, this impact is severe enough to outweigh the representativeness of the other rules. Thus, an optimized sample of severely limited complexity is much less representative than a random sample in which complexity is not constrained. When the limit is not as severe, though, optimized corpora with somewhat limited complexity and random corpora with greater complexity have equal representativeness, due to the fact that the severity of the complexity constraint has a non-linear effect on representativeness (Figure 2). For corpora of unconstrained complexity, the results mirror the results for corpora with constrained complexity equal to the true base rate of complex sentences for the grammar (75%). The average representativeness of randomly selected corpora was 65.7 ± 4.9, with $76.3\% \pm 4.9$ complex trees, while the average representativeness of corpora selected for representativeness was $87.7 \pm 8x10^{-5}$, with $80.1\% \pm 10.3$ complex trees.

Representativeness with Unknown Structure

The previous simulation assumed that our corpus consisted of the structure of the sentences, from which we could directly compute the representativeness of a given corpus. However, one might alternatively assume that a language learner has only the sentences as data and must consider all possible structures. We examine this possibility by using the same corpora of sentences as in the previous simulation, but assuming the structure of each sentence is unknown.

As mentioned in the previous section, when the structure of sentences is unknown we need to resort to Monte Carlo approximation to compute representativeness. We used importance sampling (Neal, 1993); our proposal distribution was a Dirichlet distribution with parameters equal to the true distribution in the grammar multiplied by ten. Results are averages of 30 iterations of 10,000 samples each; in the case of random corpora, sampling was done for each of ten corpora with the same constraints on complexity. Given that variance for the optimized corpora was much smaller, sampling was done for only one optimized corpus of each level of complexity. We consider the same four levels of complexity as Elman (1993) and Rohde and Plaut (1999): 0%, 25%, 50%, or 75% of the total corpus size. Additionally, we consider corpora of unconstrained complexity.

Figure 3 shows that the general trends from the previous simulation hold, with a few variations. The separation between the optimized corpora and the random corpora is smaller than when the structure is known. This is partially due to the way the corpora were created. Presumably, if it was feasible to optimize over corpora with unknown structure, further separation might be attained. However, these results do suggest that optimizing the input sentences would not greatly help a learner who must consider all possible structures of sentences. Consistent with the previous simulations, repre-

[2]Grammar creation was facilitated by the Simple Language Generator (Rohde, 2003)

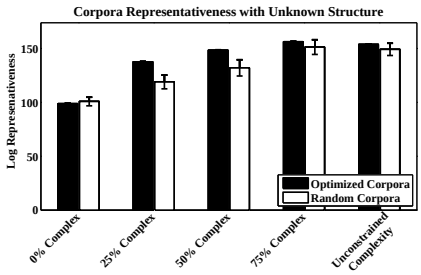


Figure 3: Representativeness of corpora with unknown structure. Limiting the complexity of a corpus limits its representativeness, with the most extreme limitation having the greatest effect.

sentativeness increases non-linearly with complexity. Again, the least complex corpora are not as representative as those that match the base rate of complexity in the grammar.

Using an Overly General Grammar

One might consider the above assumptions too strong: What if the exact structure of the grammar is not known? In this variation, we instead assume the learner has an overly general grammar that includes rules not present in the true grammar (see Figure 1). For example, rather than having only the option of expanding a transitive verb phrase to a verb followed by a noun phrase, the learner's grammar also has the possibility of expanding such a phrase to a noun phrase followed by a verb phrase. This simulates learning a grammar with unknown structure while maintaining a tractable hypothesis space (in this case, not knowing the word order in the language, but knowing the relevant syntactic classes). The extra rules give the learner a larger hypothesis space to consider, and our previous hypothesis space is the subset of the new space in which the probability of each of our newly added expansions was zero. By using an overly general grammar, we introduce much more ambiguity as to the structure of any given sentence. Thus, one might expect different results than in the previous simulation, where the number of possible derivations for any given sentence was relatively small.

The procedure for calculating representativeness in the case of an overly general grammar was very similar to the previous case. We again are considering representativeness for sentences with unknown structure, and thus used importance sampling to calculate the integral. The proposal distribution for sampling was modified so that expansions with the added rules (not present in the true grammar) would be considered. We again used a Dirichlet distribution, but the parameters were equal to ten times the true parameters plus one. Thus, rules that had zero probability in the true grammar had a parameter of one in the Dirichlet prior.

As shown in Figure 4, even with a grammar with extra rules, the results are very similar to the previous simulation. Optimizing the corpora has the strongest effect when the complexity is somewhat limited, but for the greatest representativeness, it is still best to use a corpus with greater complexity. This result suggests that even if a learner does not know the true grammar, it is still better to provide a corpus of full complexity rather than a "starting small" corpus. However, several concerns remain. The way in which we generalized the grammar was limited to switching the orders of verb phrases and noun phrases. This adds significant ambiguity to the grammar, but is not equivalent to considering any arbitrary grammar. For example, one could imagine a grammar that had over-general rules for producing relative clauses. In that case, it is still unclear whether representativeness in a corpus of severely limited complexity could equal the representativeness of a more complex corpus. To fully explore the problem, we would need a tractable way to consider all (infinite) possible grammars that could produce the data.

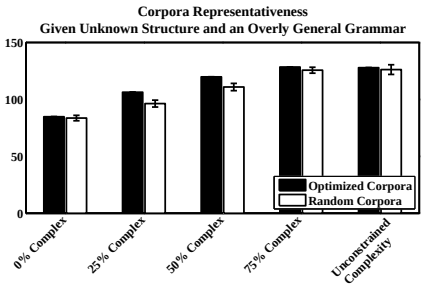


Figure 4: Representativeness of corpora with unknown structure using a grammar with extra rules.

Representativeness with N-Grams

The above simulations assume the learner learns a PCFG, but existing neural network models formulate language learning as learning to predict the next word based on previous words. This corresponds to a model where the learner learns distributions over n-grams rather than rules, and thus we can apply the same mathematical tools to analyze the representativeness of corpora according to an n-gram model.

To calculate representativeness, we can use exact counts as in the first simulation. An n-gram is a sequence of two (bigram) or three (trigram) words, and we assume a language model that estimates the probability of the next word given the previous one or two words. Our target distribution is now the correct proportions for each n-gram, which we estimate by computing the n-grams on the large corpus from which the other corpora were drawn.

Despite the fact that the model of the language has changed significantly, similar results hold in this case as in the other cases. Figure 5 shows the representativeness of the same corpora used in the other simulations with respect to n-grams: the random corpora of full complexity are still more representative than optimized corpora with limited complexity.

Summary

In our analysis, we have shown that starting small limits the degree of representativeness of the corpora and that the effect on representativeness is greatest when the limitations are particularly severe. These results hold regardless of the variations we considered. While our task is not exactly the same as in Elman (1993) or Rohde and Plaut (1999), it has bearing on this debate. From a computational level perspective, the only

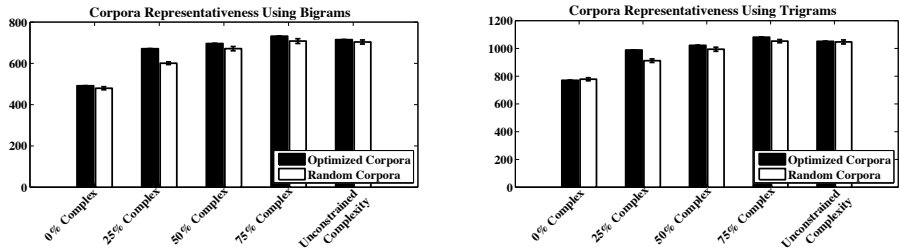


Figure 5: Representativeness of corpora using an n-gram model.

concern for such learning is whether the corpora are representative, and we have shown that starting small (at least in the extreme form) is not compatible with maximizing representativeness. However, if for mechanistic reasons one needs to start small, the above results suggest that starting "smaller" can result in similar representativeness in an optimized corpus to that of a random corpus of full complexity.

Discussion

We have shown how the concept of Bayesian representativeness can be applied to language in order to characterize an optimal sample and presented a case study of how representativeness changes with constraints on the sample. Mathematically, the Bayesian representativeness of language structures matches our intuitive sense of representativeness: a sample of language is most representative if the actual number of occurrences of each structure matches the expected number. While we cannot give a closed form expression for the representativeness of a corpus where the sentences structures are not given, simulations show that the trends concerning representativeness given constraints on complexity hold for these corpora as well. Finally, it is suggestive that given a grammar with overly general rules, we still find a disadvantage for corpora of limited complexity.

Our results suggest that if there is a beneficial effect of starting small, it is not located at the computational level: the statistical evidence a corpus provides in favor of the target language falls off as its complexity deviates from the complexity of the language. However, our results do show how it might be possible to start small in response to mechanistic information-processing constraints and still not impede learning, as it is possible to construct limited-complexity corpora that provide as much evidence as a random sample from the language. While suggestive, we note that these conclusions are tempered by the models we considered, and in particular the space of alternative hypotheses we allow the learner.

Overall, our analysis provides insight into what optimal linguistic input should look like in several interesting cases. A variety of next steps are possible. First, a more detailed exploration of the nature of an optimal sample given unknown rules would illuminate whether the preliminary results we have found hold given a larger space of possible grammars. Additionally, comparing our theoretical results to actual corpora of language acquisition would indicate whether child-directed speech is more representative than randomly selected adult-directed speech. This would suggest a pedagogical role for child-directed speech. This work provides a foundation for addressing these more advanced questions.

Acknowledgements. This work was supported by a Graduate Research Fellowship and grant number SES-0631518 from the National Science Foundation.

References

Baker, J. (1979). Trainable grammars for speech recognition. In J. J. Wolf & D. H. Klatt (Eds.), *Speech Communication Papers presented at the 97th Meeting of the Acoustical Society of America* (pp. 547–550). MIT, Cambridge, Massachusetts.

Cameron-Faulkner, T., Lieven, E., & Tomasello, M. (2003). A construction based analysis of child directed speech. *Cognitive Science*, *27*(6), 843–873.

Elman, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning*, *7*, 195–224.

Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition*, *48*(1), 71–99.

Furrow, D., Nelson, K., & Benedict, H. (1979). Mothers' speech to children and syntactic development: Some simple relationships. *Journal of Child Language*, *6*(3), 423–443.

Goldowsky, B. N., & Newport, E. L. (1993). Modeling the effects of processing limitations on the acquisition of morphology: The less is more hypothesis. In J. M. Mead (Ed.), *The Proc. of the 11th West Coast Conference on Formal Linguistics.* Stanford, CA: CSLI.

Hoff, E., & Naigles, L. (2002). How children use input to acquire a lexicon. *Child Development*, *73*(2), 418–433.

Huttenlocher, J., Vasilyeva, M., Cymerman, E., & Levine, S. (2002). Language input and child syntax. *Cognitive Psychology*, *45*(3), 337–374.

Marr, D. (1982). *Vision*. San Francisco, CA: W. H. Freeman.

Neal, R. M. (1993). *Probabilistic inference using Markov chain Monte Carlo methods* (Tech. Rep. No. CRG-TR-93-1). Dept. of Computer Science, University of Toronto.

Newport, E. L., Gleitman, L. R., & Gleitman, H. (1977). Mother I'd rather do it myself: Some effects and non-effects of maternal speech style. In C. Snow & C. Ferguson (Eds.), *Talking to children: Language input and acquisition* (pp. 31–49). Cambridge, England: Cambridge University Press.

Pine, J. M. (1994). The language of primary caregivers. In C. Gallaway & B. J. Richards (Eds.), *Input and interaction in language acquisition* (pp. 15–37). Cambridge, England: Cambridge University Press.

Rohde, D. L. (2003). *The simple language generator: Encoding complex languages with simple grammars* (Tech. Rep.). Department of Brain and Cognitive Science, MIT.

Rohde, D. L., & Plaut, D. C. (1999). Language acquisition in the absence of explicit negative evidence: How important is starting small? *Cognition*, *72*, 67–109.

Sherrod, K. B., Friedman, S., Crawley, S., Drake, D., & Devieux, J. (1977). Maternal language to prelinguistic infants: Syntactic aspects. *Child Development*, *48*(4), 1662–1665.

Snow, C. E. (1972). Mothers' speech to children learning language. *Child Development*, *43*(2), 549–565.

Tenenbaum, J. B., & Griffiths, T. L. (2001). The rational basis of representativeness. In *Proceedings of the 23rd Annual Conference of the Cognitive Science Society* (pp. 84–98).

A Cognitive Model of Positive and Negative Congruency Effects in Unmasked Priming: The Role of Attentional Limit and Conflict

Ahmad Sohrabi (a.sohrabi@uok.ac.ir)
Department of Psychology, University of Kurdistan, Pasdaran Blvd.
Sanandaj, Kurdistan 66177 15175 Iran

Robert L. West (robert_west@carleton.ca)
Department of Psychology and Institute of Cognitive Science, Carleton University, 1125 Colonel By Drive
Ottawa, Ontario K1S 5B6 Canada

Abstract

Positive priming effect has been found with a short interval between the prime and the target, while negative priming effect (i.e., a congruent prime causes longer RTs) has been found with a long time between the prime and the target. Negative priming effect has been shown mainly using masked priming but some recent studies have shown it without masks (i.e., in unmasked or conscious conditions). We employed our previous model of masked priming for the unmasked condition here, only by removing mask presentation. The model successfully simulated the negative priming effect in unmasked condition found in previous experimental studies.

Keywords: Negative congruency effect; Negative compatibility effect; modeling; attention; consciousness.

Introduction

Studies on priming have long shown reliable positive effects of the congruent prime on target processing. An early study, in the age of using tachistoscopes, was one conducted by Marcel (1983) on word and color naming. The effect of masked priming showed that masked stimuli are indeed processed to the level of response. Later studies on unmasked and masked conditions showed similar results both for masked priming (e.g., Neumann & Klotz, 1994; Dehaene et al., 1998; Eimer & Schlaghecken, 2002) and masked and unmasked priming differences (e.g., Cheesman & Merikle, 1986; Dehaene, Artiges, et al., 2003; Schlaghecken & Eimer, 2002).

In masked priming tasks, a brief masked stimulus (the prime) can affect the processing of the stimulus that follows (the target). A prime, a mask, and a target are presented sequentially and the task is to make a decision on the target. The result is usually a Positive Congruency Effect (PCE), also known as the positive compatibility effect. In PCE, the prime speeds up the performance on the target if they are congruent and slows down the performance if they are incongruent (e.g., Neumann & Klotz, 1994; Dehaene et al., 1998; Eimer & Schlaghecken, 2002; Jaśkowski & Ślósarek, 2007). Conversely, a negative priming effect has been found, called the Negative Congruency Effect (NCE). This effect is also known as the negative compatibility effect, where paradoxically the prime increases the performance on the target if they are incongruent and decreases the performance if they are congruent (e.g., Schleghecken & Eimer, 2000, 2002, 2006; Eimer, 1999; Eimer & Schlaghecken, 1998; 2001, 2002; Lleras & Enns, 2004, 2006; Verleger et al., 2004; Jaśkowski & Ślósarek, 2006). The PCE has been shown with a short mask-target Stimulus Onset Asynchrony (SOA), while the NCE has been shown with a longer mask-target SOA (see below).

The PCE has been found usually with verbal and shape stimuli and a short mask (e.g., 71 ms, as in Dehaene et al., 1998) and no or a small interval between stimuli. In contrast, the NCE has been shown mainly with arrow stimuli and a longer mask (e.g., 100 ms). Recently, it has been replicated with other stimuli, for example shapes (Jaśkowski & Ślósarek, 2006) and faces (Bennett, Lleras, Orient, & Enns, 2007). This effect has been found by using a long mask (about 100 ms) and a long mask-target SOA (>80 ms) or a long (> 30 ms) prime-mask Inter Stimulus Interval (ISI) or mask-target ISI (e.g., Eimer & Schleghecken, 1998, 2002; Jaśkowski & Ślósarek, 2007). These manipulations all increase the prime-target SOA.

In Eimer and Schlaghecken's (2002) aforementioned experiments on the role of prime duration and mask density, participants who were better at detecting the prime showed a later change from positive to negative, and conversely those who were not good in reporting the prime showed an earlier change from positive to negative, showing that there is a close relationship between prime reportability and the direction of priming. Schlaghecken and Eimer (2000) and Eimer and Schlaghecken (2002, see also 2003) found that when there is no mask or the mask is peripheral (i.e., it does not make the prime unreportable), the result is PCE, unlike the situation with masked priming. Using their motor self-inhibition hypothesis, they argued that the inhibition is initiated (as an automatic or evolved process) when visual input disappears, otherwise is blocked by visual input. Therefore, they claimed that an NCE, being a result of this self-inhibition, occurs only in the masked condition because prime input is stopped by the mask. They added that with the reportable prime, motor self-inhibition is prevented by the prime, so a PCE occurs. However, recently Lleras and Enns (2006), by comparing different studies, showed that prime visibility has no linear relationship with NCE, meaning that NCE is not necessarily caused by prime invisibility (see below).

To investigate whether there is any differences between masked and unmasked priming, Cheesman and Merikle (1986) employed Marcel's colour priming task with modifications. They changed the ratio of congruent to incongruent trials, so that in one condition this ratio was 25:75 and in the other one it was 75:25. In the unmasked condition, they found that when the number of congruent trials was high (i.e., the 75:25 condition), the congruency effect was higher than when this number was low (i.e., the 25:75 condition). In other words, when an incongruent trial was frequently preceded by a congruent trial, the congruency effect increased, and conversely, when an incongruent trial was frequently preceded by an incongruent trial, the congruency effect decreased. This difference was not found in the masked condition. They argued that participants can use a strategy based on context only in the unmasked condition.

Jaśkowski (2007) combined Eimer and Schleghecken's paradigm and Merikle and colleagues' (Cheesman & Merikle, 1984, 1986; Merikle & Joordens, 1997) to study the difference between the masked and unmasked conditions. In a congruent to incongruent ratio of 20:80, a PCE was found in the unmasked condition with both medium (100 ms) and long (800 ms) prime-target ISI. While in the congruent to incongruent ratio of 80:20, a PCE was found in medium (100 ms) ISI but an NCE was found, interestingly enough, in long ISI condition. In another experiment, while Jaśkowski found an NCE in the masked condition with a prime-target ISI of 100 ms, he found only a non-significant NCE with a long ISI. Therefore, surprisingly, with the long ISI the NCE for the unmasked condition was larger than it was for the masked condition, ruling out the necessity of the mask and invisibility of the prime in NCE. A similar result had already been found with a Stroop task (Merikle & Joordens, 1997).

In our previous work we have modeled masked priming using a neurocomputational cognitive model (Sohrabi and West, 2009a, b; see also Sohrabi, 2008). We employed that model of masked priming for the unmasked condition here. We only removed the mask presentation to simulate the unmasked condition in human experimental studies (here, Jaśkowski, 2007).

The Model

The model is based on previous neurocomputational modeling and nurophysiological studies (e.g., Usher & Davelaar, 2002; Gilzenrat et al., 2002, see also Aston-Jones & Cohen, 2005). It has been demonstrated that these types of reduced models can resemble the neural computation of a large group of neurons (e.g., Wong & Wang, 2006).

The model has been described previously (Sohrabi and West, 2009a, b; see also Sohrabi, 2008 and Sohrabi and West, 2010). It is a multi-layer dynamic neural model (shown in Figure 1) that consists of a feed-forward component for perceptuo-motor processing from the Input Layer (IL) to the Representation Layer (RL) and Motor Layer (ML, not shown). An assumption is that the cognitive processing, including the response, is modulated by attention. The Alert Attention layer (AA) simulates attentional modulation that is supposed to be a model of Locus Coeruleus (LC) that potentiates cortical areas through norepinephrine (Aston-Jones & Cohen, 2005). The executive attention is only modelled through its effects on AA, using a Task Layer (i.e., TL) for conflict monitoring. The effect of TL on AA simulates direct cortical projections to LC (Aston-Jones & Cohen, 2005). The TL and ML are affected by both prime and target. The ML's architecture is identical to TL's, with the exception that it sends no outputs to AA, is slower, and noisier (see Table 1).

Each condition in a simulation consists of 20,000 trials (200 independent blocks of 100 trials each, with congruent and incongruent trials counterbalanced randomly within each block). A single trial takes 1100 cycles. Each block starts with 500 cycles without changes in IL to let the units in other layers reach a steady state of activation. Similarly the Inter-Trial Interval (ITI) for each trial is 500 cycles, which allows the activation of units to return to baseline following the responses. The prime is presented by clamping one of the two units in the IL to 1, intended to be left or right in the case of arrows. The mask units in IL are set to 1 at the time of mask presentation and are otherwise set to 0. Therefore, the recognition of the stimuli is implemented with a localized representation, for example, the left unit is turned on when the stimulus is an arrow pointing left; otherwise the right unit is turned on. Accordingly, as will be described below, in a congruent trial the two corresponding units (e.g., the left unit of the prime and target in IL) is set to 1 or 0 at the time of stimulus presentation, while in an incongruent trial, one of the two relevant units of the prime or target is set to 1 and the other to 0.

The units in each layer make connections, via excitatory weights, to their corresponding units in other layers. The activations of these units (except IL) are calculated by a sigmoid (logistic) function of the incoming information, and a small amount of random noise. The RL sends excitatory activities to ML and TL continuously but activates AA only if a unit of the prime or target reaches a designated threshold of .62. Similarly, when one of the two units in the ML reaches the same designated threshold it triggers a manual response (i.e., initiating a hand movement). When AA is activated and its activation reaches a threshold, it starts modulating information processing in RL, TL, and ML by making the activation function of their units steeper (see Figure 2, as described below).

As shown in Figure 1, the IL encodes the prime, the mask, and the target, and projects to RL through excitatory connections. For the sake of simplicity, prime and target units, as well as an identical mask unit for each (not activated in this simulation) were implemented in two separate paths. All units in TL have a self-excitation connection, intended to simulate mutual excitation among a group of neurons. Connections between mutual units (for prime and target and to the mask) from IL to TL have small

cross-talks (see Table 1), indicating feature overlaps or similarities among stimuli. The units also have lateral inhibition with neighboring units within the same layer. The mask units are activated after the prime and before the target for a specific time. They have lateral inhibition with prime and target. To simulate unmasked condition here the mask units are not activated (i.e., are not clamped) but units' baseline activities were preserved for the sake of model stability without changing the parameters.

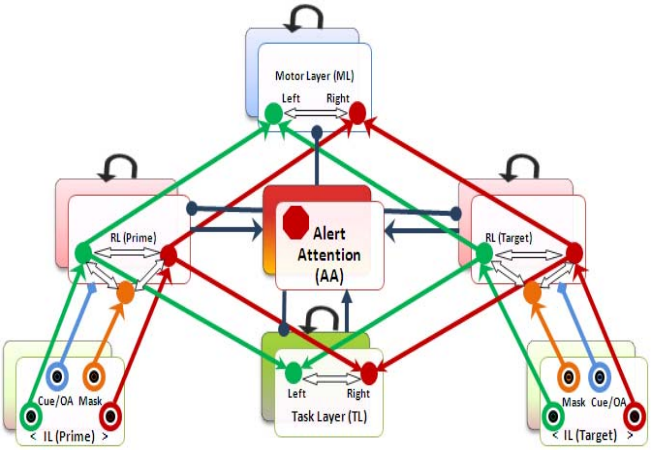


Figure 1. Architecture of the model showing hypothetical networks and connections. *Unit types:* ◉ IL ● TL and ML (not shown here) ● AA. *Attention types:* -◆ Cue/Orient Attention (OA) (not employed here) -▶ Executive (conflict driven) -● Alert. *Activation types:* ↶ Self-excitation and recurrent excitation ↔ Lateral inhibition ⟶ Feed-forward activation.

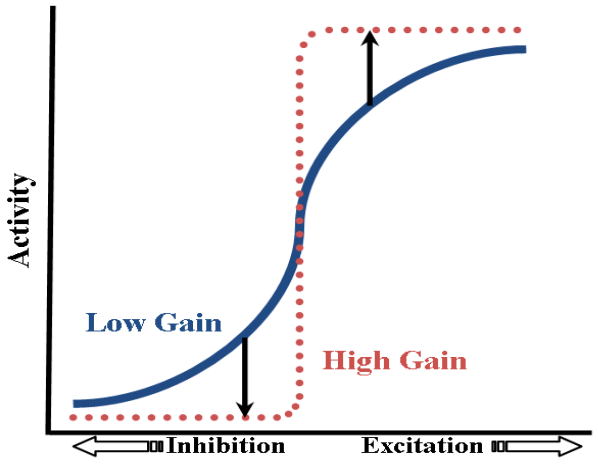


Figure 2. Effect of gain modulation on nonlinear activation function (adapted from Servan-Shreiber et al., 1990, see also Astone-Jones & Cohen, 2005).

The units in all layers (except IL and AA) receive additive Gaussian noise (zero mean and variance σ), intended as general, irrelevant incoming activities. The activations in the model are represented using units with real valued activity levels. The units excite and inhibit each other through weighted connections. Activation propagates through the network when the IL is clamped with input patterns, leading to a final response. As will be described below, the states of units in RL, ML, and TL are adopted in a method similar to a noisy, leaky, integrator algorithm (Usher & Davelaar, 2002; Gilzenrat et al., 2002). These types of models are noisy versions of previous connectionist models.

In each trial or epoch, one of the prime units in the IL is turned on and the network is left active for 43 cycles, then it is turned off for 168 cycles (short prime-target SOA), 234 cycles (long prime-target SOA), or 294 cycles (very long prime-target SOA), followed by turning on the target input in IL for 200 cycles. This is similar to a trial in human data (Dehaene et al., 1998; Eimer & Schleghecken, 2002; Jaśkowski & Ślósarek, 2006; Jaśkowski, 2007).

The prime and target units in the IL are used to represent the stimulus features (here, direction). However, as mentioned before, the recognition of the stimuli is not implemented in detail, but is encoded as a binary code. For example, in the case of arrows here, 1 is used for the left unit if it points left, and 0 is used for the opposite (reciprocal) unit. In the congruent condition, the RL units of the prime and target at the same side (left or right randomly) are turned on (1) or off (0) in each trial at the time of stimulus presentation. By contrast, in the incongruent condition, the two units at the opposite sides are turned on and the other two are left off, with random selection of the two possible cases.

The RL is governed by a modified version of previous models (Usher & Davelaar, 2002; Gilzenrat et al., 2002), which is calculated with discrete integrational time steps using the dynamic equation:

$$X_i(t+1) = \lambda_x X_i(t) + (1-\lambda_x)\, f\,[WX_iX_i\, X_i(t) + WX_iI_i\, I_i(t) - WX_iX_j\, X_j(t) - \theta X_i + \xi X_i] \quad (1)$$

Likewise, ML and TL are modelled in a similar way with their inputs coming from RL:

$$Y_i(t+1) = \lambda_y Y_i(t) + (1-\lambda_y)\, f\,[WY_iY_i\, Y_i(t) + WY_iX_i\, X_i(t)) WY_iY_j\, Y_j(t) - \theta Y_i + \xi Y_i] \quad (2)$$

In equations (1) and (2), *X* and *Y* denote the activity of units through time *t*. *W* is the weight of the connections between units, *I* is the input, and the subscripts *i* and *j* are indexes of the units. The three weight parameters in the brackets correspond to recurrent self-excitation, feed-forward excitation, and lateral inhibition, respectively. However, for the sake of simplicity in equation 1, the lateral

excitation from mask units to the prime and target, WX_iX_j $X_j(t)$, and the cross-talk in prime and target to reciprocal units and mask units, WX_iI_j $I_j(t)$, are not present. The term θ is the bias, the term ξ is noise, and f is a sigmoid function (see equation 3). The term λ represents neural decay which is related to the discrete integrational time steps in the underlying equation (Usher & Davelaar, 2002).

The AA modulates other layers by changing their activation from sigmoid toward binary responses. The activation function, f, transfers the net input, X, of a unit, and modulatory gain, g, to its activity state, implementing the firing rate of a neuron or the mean firing rate of a group of neurons:

$$f(X_i)=1/(1+\exp(-X_i g)) \quad (3)$$

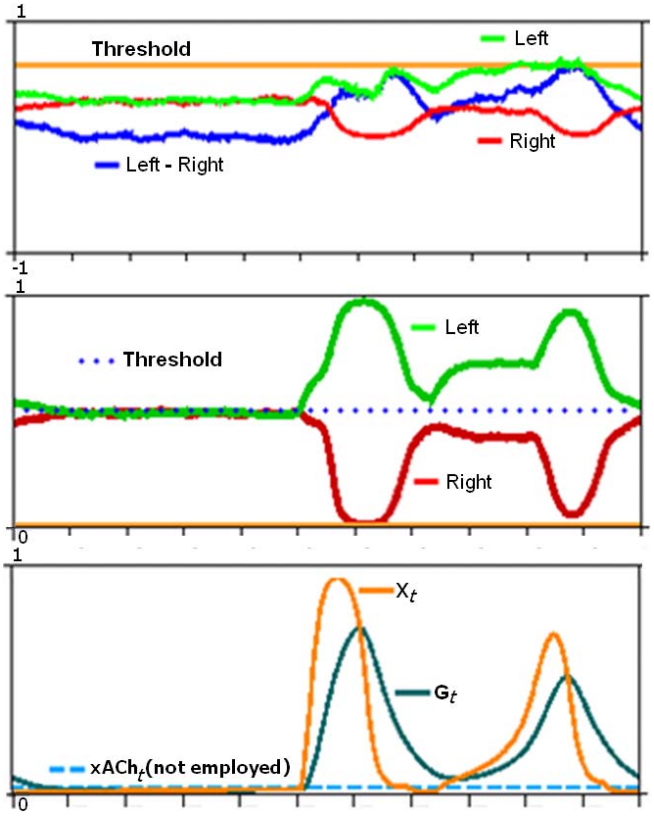


Figure 3. An unmasked congruent trial (where no conflict occurs) of 1100 cycles, including 500 cycles inter-trial interval) with 234 cycles prime-target SOA that crosses the threshold after 876 cycles (including 500 cycles inter-trial interval). From the top, ML, TL, and AA (but RL-prime, RL-target, and IL are not shown).

A conflict-monitoring measurement was employed to take the activations of the units in the TL layer to adjust phasic and tonic response modes of AA. The activation of the TL units was used to measure the Hopfield energy function between units (Hopfield, 1982), as used previously (Botvinick et al., 2001). Conflict can be defined as the concurrent activation of the competing units and as the joint effect of both prime and target in TL. Hopfield energy can be calculated as

$$E = -.5\, X^t W X = -.5 \begin{bmatrix} X_1 & X_2 \end{bmatrix} \begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \quad (4)$$

where E denotes energy, X denotes the activity of a unit, W is the weight of the connection between units, and the subscripts $_1$ and $_2$ are indexes of the two units.

As noted above, TL combines prime and target activations and measures conflict between its two units. When one TL unit is active and the other is inactive, conflict is low. However, when both units are active concurrently, the conflict is high. Activations in TL units are converted to 1 if they are equal to or greater than .5, and to 0 otherwise (i.e., using a threshold function). Also, $E > .5$ is considered as a conflict, otherwise as no conflict. When the activation of a prime or target unit in TL reaches the designated threshold, .62, the AA is activated with a phasic or tonic mode, depending on the absence or presence of conflict in TL. The change in AA response mode usually occurs by the presentation of a target that is incongruent with the prime.

Here the AA is modeled using a reduced or abstracted version of LC neurons in a Willson-Cowan type of system (e.g., Wilson & Cowan, 1972) adopted recently (Usher & Davelaar, 2002) (there are similar models and detailed implementations of this type of attention (Gilzenrat et al., 2002):

$$\begin{aligned} X(t+1) &= \lambda_x X(t) \\ &\quad + (1-\lambda_x)\, f\,[c\,(a_x X(t) - bY(t) + I_x(t) - \theta_x)], \\ Y(t+1) &= \lambda_y Y(t) \\ &\quad + (1-\lambda_y)\, f\,[c\,(a_y X(t) - \theta_y)], \\ G(t+1) &= \lambda_g G(t) \\ &\quad + (1-\lambda_g)\, X(t) \end{aligned} \quad (5)$$

where f is again a sigmoid function (as in equation 3), X is the fast variable representing AA activity and Y is a slow auxiliary variable, together simulating excitatory/inhibitory neuron groups in the LC (Usher & Davelaar, 2002). The X and Y variables have decay parameters λ_x and λ_y, excitatory/inhibitory coefficients, a_x and a_y, as well as thresholds θ_x and θ_y, respectively. The G variable is the output of the AA, which is based on X. The g (used in equation 3) is computed from G: g = G * K. The AA modulates other layers when g crosses a threshold, 1. Its activity modes can be phasic or tonic depending on the conflict state, *low* or *high*, respectively.

In all conditions the TL can change the AA mode according to the conflict between prime and target (i.e., using within-trial conflict). The phasic and tonic modes of AA responses are implemented using high or low c value (*3* or *1)* (see equation 5). The c value is 3 at the beginning of each trial (for the prime), but it is set to 1 (for the target) if conflict occurs. The number of computer simulation cycles from the target onset until one of the ML units reached a designated threshold, .62, was considered as RT. A constant, as other sensory and motor processes, could be added to this RT, to increase the match between simulation and human data.

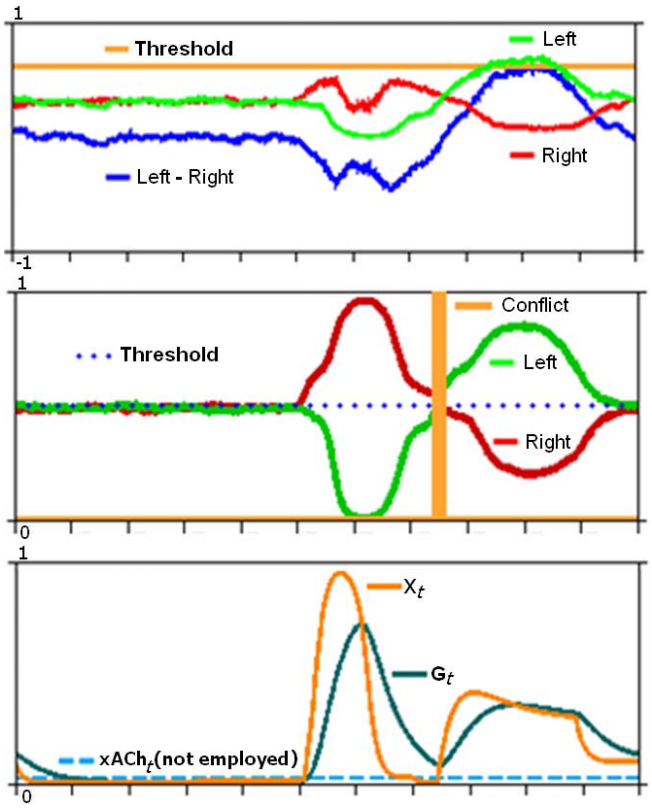


Figure 4. An unmasked incongruent trial (where no conflict occurs) of 1000 cycles, including 500 cycles inter-trial interval) with 234 cycles prime-target SOA that crosses the threshold after 876 cycles (including 500 cycles inter-trial interval). From the top, ML, TL, and AA (but RL-prime, RL-target, and IL are not shown).

Simulation Results

To create the short and long prime-target SOA conditions, a relatively short SOA (168 cycles) and two relatively long SOAs (234 and 294 cycles) were used. As shown in Figure 5, a strong PCE was found at prime-target SOA 168 cycles and a strong NCE was found at SOA 234 and 294 cycles. In the unmasked condition, in the current simulation, NCE remains high with further increases in SOA but it decreases slowly.

The simulation results in Figure 5 show a change from PCE to NCE and a drop in RTs, similar to the human data. However, the SOA in the long condition in Jaśkowski (2007) is much longer than the long conditions in the current simulation, due to a limited time course in the model, as the parameters were set for a short trial.

The activities in three layers (ML, TL, and AA) are shown for a given congruent and incongruent trial in Figures 3 and 4, respectively. There is smaller activation left in AA for the target, but it can be recovered as an effect of conflict in the incongruent condition as the phasic mode becomes tonic.

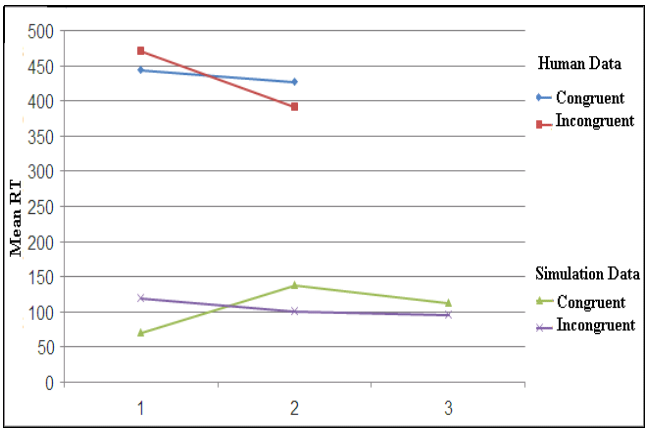


Figure 5. Unmasked priming using 168, 234, and 294 cycles prime-target SOAs, indicated by 1, 2, and 3, respectively, compared to 116.7 and 816.7 SOAs in Jaśkowski (2007), indicated by 1 and 2, respectively.

Discussion

A model that we have used for simulating masked positive and negative priming previously could simulate unmasked priming effect as well. Because there was no interruption by the mask, in the relevant unmasked prime condition, a PCE was found for short prime-target SOA. In this case, the PCE was large, consistent with the unmasked condition in Jaśkowski (2007). We assumed that a relevant or predicting prime as in Jaśkowski (2007) evokes a phasic activation in the so called alert attention to the prime but can lead to a refractory period of attentional response to the target.

An unmasked prime caused large PCE and NCE at short and long prime-target SOA, respectively. A few studies have previously shown an NCE in the unmasked condition. Here it is assumed that this effect was found in those studies because they used a medium (Koechlin et al., 1999) and long (Jaśkowski, 2007) prime-target SOA, and especially the tasks required action on (which requires attention too), or attention to, the prime, respectively. In the former, especially because of controlling physical repetition priming (and an action on the prime was required as on the target), and in the second, especially because of prime relevance (participants were told that prime highly predicts the target), the NCE was large. It could be caused by the strong refractory period created by attention to the unmasked prime. To simulate this phenomenon, in this simulation the prime was unmasked and AA mode for the prime was put in the high phasic mode ($c=3$), as with simulations of masked conditions.

At longer prime-target SOA, the relevant unmasked prime caused an NCE even larger than an equivalent masked condition (see Sohrabi and West, 2009a, b; see also Sohrabi, 2008; Sohrabi and West, 2010), consistent with Jaśkowski & Ślósarek, (2006) and Jaśkowski (2007). Interestingly, the conflict period caused by an unmasked incongruent prime (in all unmasked

conditions) was longer than that of masked incongruent prime consistent with Dehaene et al. (2003) that have shown more brain activations in unmasked incongruent compared to congruent condition.

Table 1. *Parameters in the model, fixed for all conditions, unless otherwise mentioned.*

WX_iI_i (IL to RL) [P & T] & WY_iX_i (RL to ML) [P & T]	3 & 1.5
WX_iI_i (IL to RL) [M] & WY_iX_i (RL to TL) [P & T]	1.5 & 1
WX_iX_i (RL) [P & T], WX_iX_i (RL) [M], WY_iY_i (TL), & WY_iY_i (ML)	1.5, 1.25, 1, & .9
WX_iX_j (RL) & WY_iY_j (ML & TL)	1 & 1
WX_iX_j (RL) [M to P & T] & WX_iI_j (IL to RL)	.75±.1 & .33
K (AA)	4.52
α & β (RL, TL, & ML) [M, P, T]	1 & 1
θ_x, θ_y (AA), θ_x (RL), θ_y (TL), & θ_y (ML)	1.25, 1.5, .5, .85, & 2
b,c, a_x & a_y (AA)	4, 1-3, 2, & 3
λ_x, λ_g, & λ_y (AA)	.92, .98, & .996
λ (TL), λ (ML), & λ (RL)	.75, .925, & .95
σ (CL), σ (RL) [P & T], σ (ML), & σ (RL) [M]	.025, .2, .25, & 1.25

IL=Input Layer; RL=Representation Layer; TL= Task Layer; ML=Motor Layer; AA=Alert Attention; P=Prime; T=Target; M=Mask.

Acknowledgement

This research was supported by the Carleton Cognitive Modelling Lab, Ottawa, Ontario, Canada and the University of Kurdistan, Sanandaj, Kurdistan, Iran.

References

Aston-Jones, G. & Cohen, J. D. (2005). An Integrative Theory of Locus Coeruleus-Norepinephrine Function: Adaptive Gain and Optimal Performance, *Nature Review Neuroscience*, 28, 403-450.

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. C. (2001). Conflict monitoring and cognitive control, *Psychological Review,* 108, 624-652.

Bowman, H., Schlaghecken, F., Eimer, M. (2006). A neural network model of inhibitory processes and cognitive control. *Visual Cognition,* 13, 401–480.

Cheesman, J. & Merikle, P. M. (1984). Priming with and without awareness. *Perception & Psychophysics*, 36, 387–395.

Dehaene, S., Naccache, L., Le Clec'H, G., Koechlin, E., Mueller, M., Dehaene-Lambertz, G., van de Moortele, P. F., and Le Bihan, D. (1998). Imaging unconscious semantic priming. *Nature*, 395, 597-600.

Dehaene, S. et al. (2003). Conscious and subliminal conflicts in normal subjects and patients with schizophrenia: the role of the anterior cingulated, *Proceedings of the National Academy of Sciences of the United States of America,* 100, 13722-13727.

Eimer, M. & Schlaghecken, F. (1998). Effects of masked stimuli on motor activation: Behavioural and electrophysiological evidence. *Journal of Experimental Psychology: Human Perception and Performance*, 24, 1737-1747.

Eimer, M. & Schlaghecken, F. (2002). Links between conscious awareness and response inhibition: evidence from masked priming. *Psychonomic Bulletin & Review,* 9, 514–520.

Gilzenrat, M. S., Holmes, B. D., Rajkowski, J., Aston-Jones, G., & Cohen, J. D. (2002). Simplified dynamics in a model of noradrenergic modulation of cognitive performance. *Neural Networks, 15,* 647–663.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences, USA, 79,* 2554-2558.

Jaśkowski, P. & Ślósarek, M. (2007). How important is a prime's gestalt for subliminal priming? *Consciousness and Cognition*, 16, 2, 485-497.

Lleras, A., Enns, J.T. (2006). How much like a target can a mask be? Geometric, spatial, and temporal similarity in priming. A reply to Schleghecken & Eimer (2006). *Journal of Experimental Psychology: General* 135, 495–500.

Marcel, A.J. (1983). Conscious and unconscious perception: experiments on visual masking and word recognition. *Cognitive Psychology* 15, 197-237.

Merikle, P. M., & Joordens, S. (1997). Parallels between perception without attention and perception without awareness. *Consciousness and Cognition*, 6, 219–236.

Neumann, O., Klotz, W. (1994). Motor responses to unreportable, masked stimuli: Where is the limit of direct motor specification. In C. Umiltà and M. Moscovitch (Eds.), *Attention and Performance XV: Conscious and non-conscious information processing* (pp. 123-150). Cambridge: MIT Press.

Nieuwenhuis, S., Gilzenrat, M. S., Holmes, B. D. & Cohen, J. D. (2005). The Role of the Locus Coeruleus in Mediating the Attentional Blink: A Neurocomputational Theory. *Journal of Experimental Psychology: General*, 134, 3, 291–307.

Schlaghecken, F., & Eimer, M. (2002). Motor activation with and without inhibition: Evidence for a threshold mechanism in motor control. *Perception & Psychophysics, 64,* 148–162.

Schlaghecken, F., Eimer, M., (2000). A central/peripheral asymmetry in subliminal priming. *Perception and Psychophysics,* 62, 1367–1382.

Servan-Schreiber, D., Printz, H., Cohen, J.D. (1990). A network model of catecholamine effects gain signal to noise ratio and behaviour. *Science* 249,892–895.

Sohrabi, A. (2008). Positive and Negative Congruency Effects in Masked and Unmasked Priming: Match of representation strength, Attention, and Consciousness. Ph.D. dissertation, Carleton University.

Sohrabi, A. and West, R. L. (2010). Cognitive Science of Primed Decision Making. *VDM-Verlag Publishing.*

Sohrabi, A. & West, R. L. (2009a). Positive and Negative Congruency Effects in Masked Priming: A Neuro-computational Model Based on Representation, Attention, and Conflict, *Brain Research, 1289, 124-132.*

Sohrabi, A. & West, R.L. (2009b). A Biologically-Plausible Cognitive Model (BPCM) of Positive and Negative Congruency Effects in Masked Priming. Proceedings of the 31st annual meeting of the Cognitive Science Society, Amsterdam, Netherlands, pp. 911-916.

Usher, M., & Davelaar, E. J. (2002). Neuromodulation of decision and response selection. *Neural Networks, 15,* 635–645.

Wilson, H., & Cowan, J. (1972). Excitatory and inhibitory interactions in localized populations of model neurons. *Biological Cybernetics*, 12, 1–24.

The role of action in perceiving and comparing functional relations

Ivan Vankov (i.i.vankov@cogs.nbu.bg)
Central and East European Center for Cognitive Science
New Bulgarian University, 21 Montevideo Street, Sofia 1618, Bulgaria

Boicho Kokinov (kokinov@nbu.bg)
Central and East European Center for Cognitive Science
New Bulgarian University, 21 Montevideo Street, Sofia 1618, Bulgaria

Abstract

There is growing evidence that even the most abstract capacities of human cognition are not entirely amodal and disembodied. The present study presents two empirical studies which aim to demonstrate that relational reasoning is grounded in our sensory-motor experience. Experiment 1 shows that the affordances of tool-like objects have an effect on comparing functional relations. Experiment 2 makes sure that this finding can not be explained by an automatic activation of motor systems. The results are interpreted as evidence that at least certain functional relations are perceived by simulating interactions with the environment. It is also asserted that the process of comparing such relations is constrained by the properties of the human body such as hand-dominance.

Keywords: relations, situated cognition, embodiment, action, simulation, analogy

Introduction

Imagine you are asked to compare the relation between an axe and a wooden log with the relation between a meat chopper and a piece of meat. One way to solve this problem is to find out what relation holds in the first pair of objects, turn it into propositional form (e.g. 'is used to cut(axe, wood)'), do the same for the second pair of objects and than compare the two symbolic structures. This is what many models of relational reasoning do (Gentner, 1983; Falkenhainer, Forbus, & Gentner, 1989). Some models can also establish a correspondence between distinct relational symbols by measuring their semantic similarity (Holyoak & Thagard, 1989; Kokinov, 1994; Hummel & Holyoak, 1997, 2003). However all these models do not address the problem of where the relational meaning comes from (how the propositions are encoded) and they all assume that the process of comparing relations is amodal and disembodied in nature.

On the other hand, there is plenty of evidence that human cognition is inherently modal and constrained by the characteristics of the human body (Glenberg, 1997; Barsalou, 1999; 2008; Lakoff, 1999; Fisher & Zwaan, 2008). For example, it has been shown that the perception of a graspable object immediately activates a potential motor interaction with this object, even when it is task-irrelevant (Tucker & Ellis, 1998, 2004; Beauchamp, Lee, Haxby & Martin, 2002; Beauchamp & Martin, 2007; Buccino, Sato, Cattaneo, Rodà & Riggio, 2009). Proponents of the embodiment theory claim that this phenomenon is not a mere side effect of spreading activation, but that motor programs are part of the representation of objects. These motor programs are used to simulate potential interactions with an object and determine its function.

Similarly, the perception of a functional relation between two objects should require a mental simulation of the relevant interactions with the objects. For example, in order to comprehend the functional relation between an axe and a piece of wood, you would simulate grasping the axe and chopping the wood with it. And in order to compare two instances of functional relations you have to be able to compare the motor dynamics resulting from simulating the actions involved in each of the relations. Such an approach is justified by the study of Klatzky, Pellegrino, McClosky & Lederman (1993), which found that there is remarkable consistency in people's knowledge about the movements underlying functional interactions with objects. There is also evidence that sometimes people consciously try to detect the perceptual motor similarities of different situations in order to evaluate how analogous they are (Clement, 2009).



Figure 1: An example of the stimuli used by V&K. Participants had to compare the relation between the objects in the left part with the relation in the right part of the screen. The affordances of the objects were manipulated by making them easier to be grasped with the left or with the right hand. In this example, both affordances are right.

Recently, Vankov & Kokinov (2009) (henceforth V&K) proposed a model of grounding relational meaning in simulated interactions with the environment. According to the model, the motor dynamics resulting from these interactions is used not only to comprehend relations, but also to solve the role-filler binding problem (Hummel, 1999). The model makes two major predictions. First, it

predicts relation-specific motor effects when relations are perceived, even if the task does not involve any motor activity. The second prediction is that relations are compared most efficiently when it is possible to simulate the underlying interactions simultaneously or in close temporal proximity.

V&K reported an experiment which managed to provide support for both hypotheses. Participants were asked to compare the functional relations in two pairs of objects (Figure 1) by giving a verbal response – pronouncing 'yes' or 'no'. An effect of the affordances of the objects was found. Right-handed participants' response times were faster when the objects in the relation on the left were displayed in such a way, that it was easier to manipulate them with the left hand. The effect of the affordance of the object on the right was reversed in direction and much smaller in size. The very fact that an affordance effect was found is in support of the hypothesis that perceiving relations involves simulating actions. The bigger size of the effect of the affordance which was closer to the subjects' non-dominant hand implied that subjects tried to simultaneously simulate the actions involved in the two relations. A control study ruled out the possibility that this result was due to the mere perception of objects with varying affordances. However it is still possible the effect was due to presenting the two relations at once. Another valid point is that the overall reaction time could have been affected mostly by the affordance of the relation displayed near the subjects' non-dominant hand because it was harder to be simulated. Also the design of the experiment did not allow to control the sequence in which the subjects look at the two relations. Therefore it is possible that the effect of the relation which had been attended last was different (bigger or lesser) from the other one. A new experiment was designed and conducted in order to overcome these problems.

Experiment 1

The experiment used the same stimuli as V&K, but the relations were displayed one by one in the center of the screen in order to control the order in which they were perceived and isolate the effect of the presentation location.

The elimination of the factor of the presentation location served to set apart the effect of the affordances of the stimuli from any spatial compatibility effects. It is well known that people respond faster to stimuli which location is compatible to the response action (Simon & Rudell, 1967). Although the response action in V&K was verbal, it is possible that subjects' reaction times had been affected by the congruence of the presentation location and the affordances of the stimuli. For example, an interaction between objects with a left affordance could be easier to be simulated with the right hand if they are displayed in the right part of the screen.

The new design also allowed testing the effect of a stimulus – the relation which was presented first – which had to be retrieved from memory at the time of the comparison. If any affordance effect was found for the first relation, it would seriously question any disembodied view on relational comparison which assumes that relations are first encoded as symbols and then compared.

However, according to embodied view on relational reasoning, there must be an effect of both affordances because the sensory-motor dynamics of both relations is needed at the time of comparison. Moreover, the embodied view predicts that relations will be compared more efficiently when the underlying interactions with environment could be simulated in close temporal proximity. Therefore it is predicted that subjects will be faster when the affordances differ and they can employ both their hands in the simulations.

Method

Participants 36 right-handed participants (20 females) took part in the experiment for course credit or as volunteers. Their average age was 24.06 years (age range from 18 to 53, *SD* = 5.91).

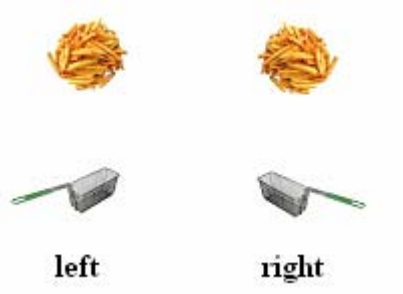


Figure 2a: A left and a right affordance of a pair of objects which make a functional relation.

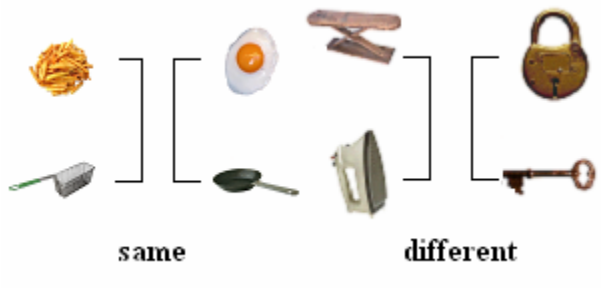


Figure 2b: Examples of the stimuli used in 'same' and 'different' trials.

Stimuli The stimulus set was the same one that was used in V&K. It consisted of 144 photos of various household objects. Each stimulus consisted of two pairs of objects. The objects in each pair participated in a certain functional relation, such as 'hammer' – 'nail', 'key' – 'lock', 'fork' – 'spaghetti', etc. In all pairs, it was possible to manipulate the affordance so that the interaction between the objects could be performed easier either with the left or the right hand

(Figure 2a). The relations in the two pairs were the same in half of the stimuli ('same' trials) and different in the others ('different' trials). A pre-test study was used to organize the objects pairs in such a way that there was maximal agreement among people whether the relations were same or different (Figure 2b). All images were resized to 400x400 pixels. In all pairs there was one tool-like, graspable object (axe, hammer, ironer, fork, etc) and it was always located at the bottom position.

Design The experiment had a 2x2 within subject design. The two independent variables were:

First affordance – left or right, depending on the affordance of the first pair of objects.

Second affordance – left or right, depending on the affordance of the second pair of objects.

The dependent variable was the reaction time of participants' verbal responses ('yes'/'no').

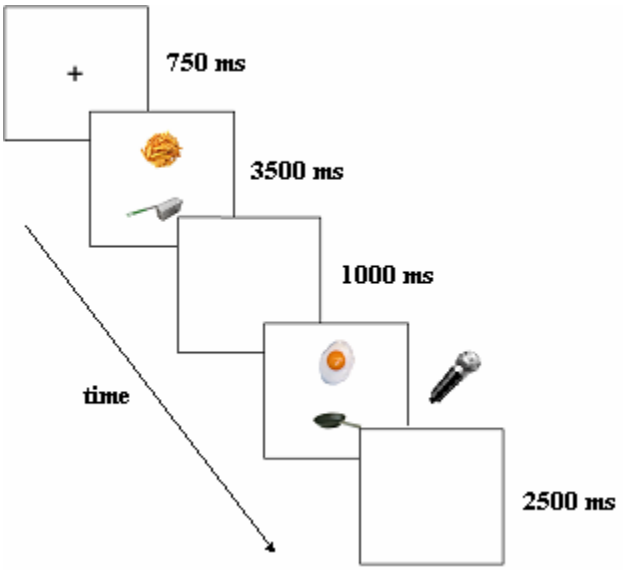


Figure 3: Experimental procedure. Reaction time was measured from the onset of the second relation until a verbal response was given.

Procedure Each stimulus was presented once to each subject. Affordance conditions and the order of presentation of the relations (first or second) were counterbalanced across subjects and it was made sure that the same combination of the affordance factors would not repeat more than 3 times in a row. Same/different trials were pseudo-randomized, so that a given correct response would not repeat more than 3 times. The trial sequence was fixed for all subjects, i.e. they saw the stimuli in the same order.

Participants were tested in a sound-proof booth. The stimuli were presented on 19" computer monitor with a resolution of 1280x1024 pixels. Before the actual experiment all participants went through a microphone training session in order to make sure that they would articulate their responses clearly enough. The experimental session started with 8 practice trials, none of which appeared in the experimental part. Each trial began with a centrally location fixation cross (750ms), followed by the onset of the first pair of objects. The objects were displayed one below the other in the centre of the screen. Subjects were instructed to perceive the relation between the objects without making any response. The first pair of objects was presented for 3500 ms and when it disappeared the screen stayed blank for 1000 ms. After that a second pair of objects was presented at the same position as the first one. The stimuli stayed on the screen for 5000ms or until a response was generated. Participants were instructed to respond by saying 'yes' if the relation between the objects in the second pair was the same as in the first pair and say 'no' otherwise. The subject's response time (RT) was measured since the onset of the second pair of objects till the moment a verbal response was detected. Stimulus presentation and response recordings were controlled by E-prime software (Schneider, Eschman, & Zuccolotto, 2002). The inter-trial interval was 2500 ms. The experiment took about 10 minutes. The total number of test trials for each subject was 36, including 18 'same' and 18 'different' trials.

Results

Trials in which subjects failed to respond or the response was incorrect were excluded from the analysis. An incorrect response was counted when a subject said 'yes' in a 'different' trial or 'no' in a 'same' trial. RT lying more than ±2.0 standard deviations from the mean 'same' and 'different' RT times were also removed. Thus a total of 82.10% of the originally collected RT data were included in the analysis.

Same and different trials were analysed separately. A 2x2 repeated measures ANOVA was performed on subject RT means in 'same' trials (Figure 4) and revealed a significant main effect of the affordance of the first relation ($F(1, 35) = 7.12$, $p < .05$, $\eta^2 = .17$). There was no effect of the affordance of the second relation ($F(1, 35) = 2.02$, $p = .16$, $\eta^2 = .06$). The interaction between the two affordance factors was not significant ($F(1, 35) = .20$, $p = .66$, $\eta^2 = .01$).

An analysis of mean item response times also found a main effect of the first affordance of 'same' items ($F(1, 17) = 9.05$, $p < .01$, $\eta^2 = .35$). The effect of the second affordance ($F(1, 17) = 2.69$, $p = .12$, $\eta^2 = .14$) and the interaction ($F(1, 17) = 2.15$, $p = .16$, $\eta^2 = .11$) were not significant .

Analyses of 'different' trials and items revealed similar patterns of results, but none of the effects reached significance.

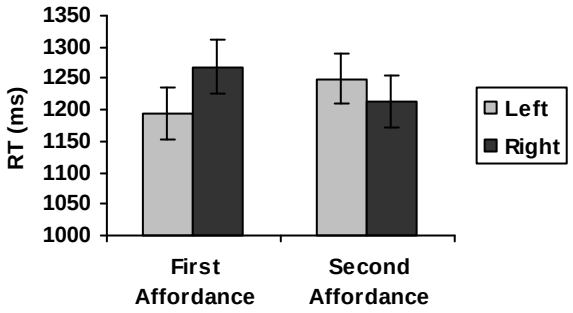


Figure 4: Experiment 1 results for same trials. Subjects' responses were significantly faster when the affordance of the first relation was left although all subjects were right-handed. The tendency for the affordance of the second relation was reversed. Error bars represent standard errors.

Discussion

The results replicated the findings of V&K (2009) as long as an effect of left/right affordances on comparing functional relations was found. Also, the shortest reaction times were in the condition when one of the affordances was left and the other one was right. Another similarity was that the effect size of the first affordance effect was bigger than the effect size of the second affordance.

The major result of Experiment 1 was that the subjects, all of which were right-handed, were faster to respond when the first affordance was left. This result can not be explained by presentation location as all stimuli were presented in the center of the screen. At first glance, there is no reasonable explanation why participants would be faster when one of the stimuli is easier to process by their non-dominant hand. However the results start to seem logical if we assume that subjects tried to run the two simulations of functional interactions simultaneously in order to compare the resulting motor dynamics. It is reasonable to assume that subjects always engaged their dominant right hand in simulating the functional interactions of the visually available second relation, event when the affordance of the objects was congruent to their left hand. Thus, when they had to compare the two relations by running two simulations at once they could use only their non-dominant hand for recalling and simulating the first relation.

The pattern of results of Experiment 1 is inconsistent with any classical encode-and-compare account. If relations are first turned into propositions and then compared, then there would not be any effect of the first affordance. The first relation would have already been encoded by the time the second relation is presented and the response is given. If the effect is due just to the activation of the visual image of the first relation then the direction of affordance effect should be the same for both relations. Yet, we conducted a control study to make sure that main results of Experiment 1 are specific to the relation comparison task.

Experiment 2

Several researchers have shown that mere looking at manipulable objects activates regions of the brain related to action (Beauchamp et al., 2002; Beauchamp & Martin, 2007; Buccino et al., 2009). The goal of this experiment was to make sure that the main findings of Experiment 1 are not due to such kind of automatic motor activation. In particular we wanted to check whether if two objects with varying affordances are presented sequentially and the task is to compare them for some reason, the reaction times will be shorter when the affordance of first object is congruent with the non-dominant hand of the subjects.

Method

Participants 24 right-handed participants (17 females) took part in the experiment for course credit or as volunteers. Their average age was 22.79 years (age range from 17 to 32, *SD* = 3.13).

Stimuli The target stimuli set consisted of the manipulable tool-like objects which used in the 'same' trials of Experiment 1. Each target stimulus consisted of two such objects (Figure 5). Objects were paired in the same way as they were in the previous experiment. There were 18 target trials. An equal number of fillers were compiled using 18 photos of man-made objects, none of which was used in the target trials, and 18 photos of natural objects (fruits, plants, rocks, etc).



Figure 5a: An example of a target stimulus used in Experiment 2. Both objects are artifacts, so the subjects should respond by saying 'Yes'.



Figure 5b: An example of a filler. The correct response is 'No' as one of the objects is of natural origin. Either of the objects could be a natural one.

Design The design was identical to Experiment 1. The affordances of the objects were described by two independent variables – 'first affordance' and 'second affordance'.

Procedure The setting of the experiment was similar to Experiment 1 except for the task. Each trial began by a fixation cross (750 ms), followed by the presentation of the

first object (2000ms). After that, the screen stayed blank for 1000 ms and the second object was presented. Subjects were instructed to say 'yes' if none of the objects was of natural origin and say 'no' otherwise. Response time was recorded since the onset of the second object. All objects were displayed in the center of the screen. The order of presentation of the objects and the affordance conditions were counter-balanced across subjects.

Results

Fillers and trials with invalid or incorrect responses were excluded from the analysis. Response times lying more than ±2.0 standard deviations from the mean RT time were removed. Thus a total of 92.40% of the originally collected non-filler RT data were included in further analysis.

A 2x2 repeated measures ANOVA was performed on subject means. It revealed main effects of the first ($F(1, 23) = 5.18$, $p < .05$, $\eta^2 = .18$) and the second object affordance ($F(1, 23) = 5.36$, $p < .05$, $\eta^2 = .19$). The interaction was not significant ($F(1, 23) = 0.22$, $p = .64$, $\eta^2 = .01$). Response times were faster when the affordances of both objects were right (Figure 6).

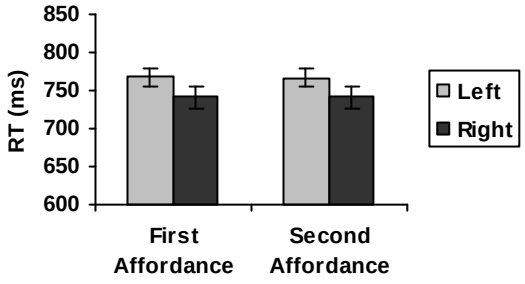


Figure 6: Results for trials with objects which were part of 'same' items in Experiment 1. Subjects were significantly faster when both objects were presented in such a way, that they were easier to be grasped with the right hand. Error bars represent standard errors.

Discussion

Experiment 2 showed that the main findings of Experiment 1 can not be explained by the automatic activation of motor programs by object affordances. Response times were shorter when the affordances of *both* objects were congruent to the subjects' dominant hand. Also, there was no difference between the sizes of the effects of the first and the second affordance. These results are different from what was found in the previous experiment and they show that the results of Experiment 1 are specific to the relation comparison process.

General Discussion

The presented experiments provide further evidence in support of the hypothesis that the meaning of functional relations is grounded in the sensory-motor dynamics resulting from simulated interactions with the environment. The pattern of results is also consistent with the idea that comparing functional relations involves running two or more such simulations simultaneously or in close temporal proximity. The outcome of Experiment 2 rules out the possibility that the results were due to the object affordances per se.

The experiments were designed not to rely on the stimulus-response compatibility paradigm, unlike most other behavior studies of affordances (for instance Tucker & Ellis 1997, 2004; Spivey, Richardson & Cheung, 2001). In this way it was made sure that the results could not be attributed to accidental spreading of activation from conceptual to motor areas of the brain (Mahon & Caramazza, 2009). If the activation of motor areas was just a side effect it would not have had any effect on verbal responses as the motor areas dedicated to hand manipulations and language production are unlikely to be systematically connected. Informal debriefing after the experiments showed that subjects were completely unaware that the task had anything to do with their hands and simulations of actions.

The outcomes of the experiments are clearly in support of an embodied view on cognition. However one may attempt to interpret the results of Experiment 1 without adopting the specific idea of embodying relational representations and relational reasoning by referring to the theory of event coding (Hommel, Müsseler, Aschersleben & Prinz, 2001). According to this theory, elements of perception and action are encoded in a common medium. When the stimulus features related to perception and action are active for a long time period they become bound into an event file. Once bound, these features are less available for planning of other actions. Hence it is possible that a right affordance of the first pair of objects would bind the features representing the right hand of the subject to the stimulus features of the first relation. When the second relation is presented, the right hand of the subject would be less available for simulating the use of the presented objects and the response would be delayed. A result of this kind has been reported by Spivey, Richardson & Cheung (2001). Such an explanation reduces the role of simulated action to the process of object recognition.

However, the theory of event coding contradicts the results of the control study, unless it is assumed that the presentations times were too short for the event filing to happen. Such an assumption is highly unlikely to be true, as in the control study the first object was presented for a fixed period of 2000ms, followed by a 1000 ms inter-stimulus interval before the second object was displayed. This period is much longer than the time which was required for suppression of future actions in the studies of Spivey, Richardson & Cheing (2001) and Stoet & Hommel (1999). Also, there is no evidence so far that such a phenomenon could occur outside the stimulus-response compatibility paradigm and have an effect on verbal reaction time. Hence,

the event filing explanation can not adequately account for the results presented in this paper.

The results of the present study are broadly consistent with the 'body-specificity hypothesis' (Casasanto, 2009), according to which 'people who interact with their physical environments in systematically different ways should form correspondingly different mental representations'. We demonstrated that an asymmetry of our bodies, such as hand dominance, constrains performance in a task which is traditionally thought to be highly symbolic and abstract in nature. It remains however an open question to what extent abstract concepts and reasoning abilities are dependent on our bodies and whether such constraints are the only source of meaning.

Acknowledgements

We would like to thank Armina Janyan and the ANALOGY team for useful discussions and Simona Dobrinova for collecting most of the data.

This work was supported by the Project ANALOGY: Humans–the Analogy-Making Species, financed by the FP6 NEST Programme of the European Commission. (Contr. No 029088).

References

Barsalou L. (1999). Perceptual symbol systems. *Behavioral and brain sciences*, *22*, 4.

Barsalou, L.W. (2008). Grounded cognition. *Annual Review of Psychology, 59,* 617-645.

Beauchamp, M., Lee, K., Haxby, J.,& Martin, A. (2002). Parallel visual motion processing streams for manipulable objects and human movements. *Neuron*, *34-1*, 149-159.

Beauchamp, M., Martin, A. (2007) Grounding object concepts in perception and action: evidence from fMRI studies of tools, *Cortex*, *43*, 461-468.

Buccino, G., Sato, M., Cattaneo, L., Rodà, F.,& Riggio, L. (2009). Broken affordances, broken objects: A TMS study. *Neuropsychologia*, *47-14*, 3074 - 3078.

Casasanto, D. (2009). Embodiment of abstract concepts: Good and bad in right and left-handers. *Journal of Experimental Psychology: General*, *138-3*, 351-367.

Clement, J. (2009). Analogical reasoning via imagery: the role of transformations and simulations. In B. Kokinov, K. Holyoak & D. Gentner (Eds.), *New Frontiers in Analogy Research* (pp. 463 - 472). Sofia, Bulgaria: NBU Series in Cognitive Science.

Falkenhainer, B., Forbus, K., Gentner, D. (1989) The Structure-Mapping Engine: Algorithm and Examples. *Artificial Intelligence*, *41(1)*, 1-63.

Fisher, M., Zwaan, R. (2008). Embodied language: a review of the role of the motor system in language comprehension. *Quarterly journal of experimental psychology*, *61*, *6,* 825-850.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, *7*, 155-170.

Glenberg, A. (1997). What memory is for. *Behavioral and Brain Sciences*, *20*, 1-19.

Holyoak, K., & Thagard, P. (1989). Analogical mapping by constraint satisfaction. *Cognitive Science*, *13*, 295-355.

Hommel, B., Müsseler J, Aschersleben G, & Prinz W. (2001) The theory of event coding (TEC): A framework for perception and action planning. *Behavioral and Brain Sciences*, *24*, 849-878.

Hummel, J. (1999). Binding problem. In R.A. Wilson and F.C. Keil (Eds.), *The MIT Encyclopedia of the Cognitive Sciences*, MIT Press.

Hummel, J., Holyoak, K. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review*, *104*, 427-466.

Hummel, J. Holyoak, K. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review*, *110*, 220-264.

Klatzky, R., Pellegrino, J., McClosky, B., & Lederman, S. (1993). Cognitive representations of functional interactions with objects. *Memory and Cognition*, *21*, 294-303.

Kokinov, B. (1994). A hybrid model of reasoning by analogy. In K. Holyoak & J. Barnden (Eds.), *Advances in connectionist and neural computation theory*: *Analogical connections*. Norwood, NJ: Ablex.

Mahon, B. Z. & Caramazza, A. (2009). A critical look at the embodied cognition hypothesis and a new proposal for grounding conceptual content. *Journal of Physiology – Paris*, *102*, 59-70.

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). *E-Prime user's guide*. Pittsburgh: Psychology Software Tools Inc.

Simon, J., & Rudell, A. (1967). Auditory S–R compatibility: The effect of an irrelevant cue on information processing. *Journal of Applied Psychology*, *51*, 300-304.

Spivey, M., Richardson, D., & Cheung, J. (2001). Motor representations in memory and mental models: The embodied zork. *Proceedings of the 23rd Annual Conference of the Cognitive Science Society* (pp. 867-872), Mahwah, NJ: Lawrence Erlbaum Associates.

Stoet, G. & Hommel, B. (1999). Action planning and the temporal binding of response codes.. *Journal of Experimental Psychology: Human Perception and Performance*, *25*, 1625- 1640.

Tucker, M., & Ellis, R. (1998). On the relations between seen objects and components of potential actions. *Journal of Experimental Psychology: Human Perception and Performance*, *24(3)*, 830-846.

Tucker M., & Ellis R. (2004). Action priming by briefly presented objects, *Acta Psychologica*, *116*, 185-203.

Vankov, I., & Kokinov, B. (2009). Grounding relations in action. In B. Kokinov, K. Holyoak & D. Gentner (Eds.), *New Frontiers in Analogy Research* (pp. 463 - 472). Sofia, Bulgaria: NBU Series in Cognitive Science.

How Causal Reasoning Can Bias Empirical Evidence

Momme von Sydow[1] (momme.von-sydow@bio.uni-goettingen.de)
York Hagmayer[1] (york.hagmayer@bio.uni-goettingen.de)
Björn Meder[1,2] (meder@mpib-berlin.mpg.de)
Michael R. Waldmann[1] (michael.waldmann@bio.uni-goettingen.de)
[1]Department of Psychology, University of Göttingen, Gosslerstr. 14, 37073 Göttingen, Germany
[2]Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany

Abstract

Theories of causal reasoning and learning often implicitly assume that the structural implications of causal models and empirical evidence are consistent. However, for probabilistic causal relations this may not be the case. We propose a causal consistency hypothesis claiming that people tend to create consistency between the two types of knowledge. Mismatches between structural implications and empirical evidence may lead to distortions of empirical evidence. In the present research we used trial-by-trial learning tasks to study how people attempt to create consistency between structural assumptions and learning data. In Experiment 1 we show biasing of empirical evidence with causal chains even after repeated testing of direct and indirect relations. Experiment 2 investigates whether different causal models lead to different judgments, despite identical data patterns. Overall, the findings support the idea that people try to reconcile assumptions about causal structure with probabilistic data, but also suggest that this may depend on the type of causal structure under consideration.

Keywords: causal reasoning; induction; Markov condition; top-down effects; heuristics and biases

Causal Reasoning and Empirical Evidence in Covariation Assessment

Probability judgments about indirect causal relationships may be based on direct observations of covariations between events (empirical evidence) or they may be derived from top-down assumptions about the underlying causal structure (structural knowledge). The crucial advantage of causal model knowledge is that we can make inferences about relations which we have not directly observed. For example, we may first learn about a causal relation $A \rightarrow B$, and later about a causal relation $B \rightarrow C$. By combining the single links into a causal chain $A \rightarrow B \rightarrow C$ we can make inferences regarding the initial event A and the final event C. For example, deterministic causal relations warrant transitive inferences, that is, the occurrence of A allows us to infer that C is present, too (like in logical 'Modus Barbara'). However, most causal relationships tend to be probabilistic: a virus does not always cause a disease; a gene does not always cause a phenotypic trait. Crucially, in the case of probabilistic relations, transitivity relations do not necessarily hold (see Ahn & Dennis, 2000; von Sydow, Meder, & Hagmayer, 2009). However, causal models may nevertheless be used for assessing indirect relations from knowledge of direct relations, in a way that is inconsistent with direct empirical evidence.

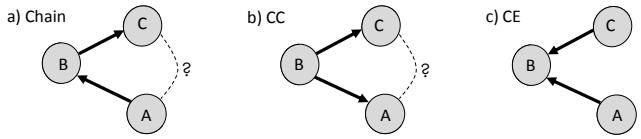


Figure 1: A causal chain, a common cause (CC) and a common effect (CE) model.

The representation of causal relationships in qualitative causal models (Gopnik et al., 2004; Rehder, 2003; Sloman, 2005; Waldmann, Hagmayer, & Blaisdell, 2006; Waldmann & Holyoak, 1992) and in causal Bayes nets (Pearl, 2000; Spirtes, Glymour, & Scheines, 1993) suggests that people only represent direct causal relations and infer other relations from these causal models based on abstract assumptions about the structures. At the center of the Bayes net formalism is the *causal Markov condition*, which states that a variable in a causal network is conditionally independent of all other variables apart from its effects, given its direct causes. If the Markov condition holds, a causal chain (Fig. 1a) with positive direct relations, $A \rightarrow B$ and $B \rightarrow C$, entails a positive contingency between variables A and C. More specifically, the conditional probability of A given C, $P(C|A)$, is given by:

$$P(C|A) = P(B|A) \cdot P(C|B) + P(\neg B|A) \cdot P(C|\neg B) \quad (1)$$

Similarly, other indirect conditional probabilities can be derived from applying the Markov condition to the causal model. If we have a common cause model (CC, cf. Fig. 1b) $A \leftarrow B \rightarrow C$, the Bayes net formalism implies a positive relation between A and C. On the other hand, if the variables are linked in a common effect structure $A \rightarrow B \leftarrow C$ (CE, cf. Fig. 1c), no positive relation between A and C is entailed.

From a computational point of view, the Markov assumption is used as prerequisite for inducing causal structures from conditional dependency and independency relations, and as a basis for probabilistic inferences across complex causal networks (Spirtes et al., 1993; Pearl, 2000). On the other hand, the status of the Markov condition as a necessary and universal feature of causal representations has been criticized (Cartwright, 2001). However,, the status of the Markov condition in human causal reasoning is still under dispute (e.g., Rehder & Burnett, 2005; Mayrhofer, Goodman, Tenenbaum, & Waldmann, 2008).

A Causal Consistency Hypothesis

A number of studies in causal learning have shown that people tend use initial assumptions about causal models and do not tend to necessarily verify whether the assumptions underlying the model hold in the data. For example, Waldmann and Hagmayer (2001) showed that people make use of instructions regarding causal structures when assessing causal strengths, even when the data contradicted the initially suggested causal model. Waldmann, Meder, von Sydow and Hagmayer (2010) connected this research with categorization and demonstrated similar effects of category transfer with variable categorization schemes.

Similar phenomena may arise when participants are requested to make inferences about indirect relations within causal models. Previous research on inferences about indirect relations in causal chains has shown that people have a tendency to assume the Markov condition when making inferences from an initial event *A* to the final event *C*. Ahn and Dennis (2000) and Baetu and Baker (2009) have presented participants with data about direct relations between binary events. Learners' inferences about the indirect relations were consistent with the use of the Markov condition. However, they only investigated inferences in the absence of any evidence regarding the indirect relation.

Von Sydow, Meder and Hagmayer (2009) provided direct evidence about the indirect relation when learning causal chains. They showed that participants reasoned transitively (apparently assuming the Markov condition) even if the learning data provided evidence against transitivity. The present research continues in the wake of this work. Participants are again provided with data about the indirect relation. In addition, the influence of the amount of learning input, task features, and different causal structures are examined. We here particularly focus on the interplay between the implications of causal structures when the Markov condition is assumed and the observed data sample. Consider the data shown in Table 1. In these data, it holds that $P(B|A) = 0.75$ and $P(C|B) = 0.75$. Nevertheless, according to the data there is no contingency between *A* and *C*, since $P(C|A) = P(C|\neg A) = 0.5$. However, if we used these data to parameterize a causal chain $A \rightarrow B \rightarrow C$, and assumed the Markov condition, this causal model would imply that there is a positive contingency between the initial event *A* and the final effect *C* (i.e., $P(C|A) > P(C|\neg A)$, cf. Equation 1). Thus, depending on whether we assess the indirect relation between *A* and *C* directly from the data, or induce a causal model from the data and use the model to make inferences regarding indirect relations, we may arrive at very different conclusions. However, whether there is a potential tension between structural knowledge and data depends on the exact structure of the causal model. For example, a common effect model $A \rightarrow B \leftarrow C$ (Fig. 1c) does not entail a statistical dependency between *A* and *C*, as in this model the two events constitute independent causes of their common effect *B*.

Our causal consistency hypothesis suggests that when there is a mismatch between the causal model's structural implications and the observed data, people will create consistency by aligning the observed evidence with the causal model's implications. As a consequence, for an actually intransitive causal chain one should observe an overestimation of the statistical relations between the indirectly linked events *A* and *C*. For example, if learners assume the Markov condition when inducing a causal chain they should infer $P(C|A) > P(C|\neg A)$. This should also hold for common cause structures (but see von Sydow et al., 2009). By contrast, a common effect model implies no statistical dependency between *A* and *C*, as they represent independent cause of their common effect *B*. Thus, for this model there should be no conflict between the structural knowledge and the observed empirical probabilities.

Table 1: Sample of intransitive data from Experiment 1.

	A	B	C
1	present	present	present
2	present	present	present
3	present	present	absent
4	present	absent	absent
5	absent	present	present
6	absent	absent	present
7	absent	absent	absent
8	absent	absent	absent

Another potentially important factor which may affect how people deal with conflicts between structural implications and empirical evidence are the number and the focus of the test questions. In a previous study (von Sydow et al., 2009), participants were confronted with a causal chain and intransitive data, which did not show a positive statistical relation among the initial cause *A* and the final effect *C*, although the direct causal relations were positive. In these studies participants were first queried about the direct causal links before being asked about the indirect causal relation among *A* and *C*. Although participants had all relevant data available, they misjudged the relation between *A* and *C* to be positive. However, when participants are queried more often about the indirect relation, they may assess the relation directly, thereby arriving at estimates that correspond more closely to empirical probabilities.

Goals of Experiments and Hypotheses

The goal of the first experiment was to investigate how task features affect the integration of structural knowledge and empirical evidence. Participants were either asked frequently or only once about the indirect causal relation. We suspected that frequent queries would direct participants' attention to the empirical evidence regarding the indirect causal relation. Unlike in our previous studies (von Sydow et al., 2009) we presented subjects with trial-by-trial data instead of grouped data. Moreover, we used simpler dichotomous learning items, as opposed to variable category exemplars. Our goal was to find out whether these changes would make the empirical conditional probabilities more salient, thereby leading to judgments corresponding closer to the learning data. The main goal of Experiment 2 was to study other causal structures as well. While keeping the trial-by-trial

contingencies identical, we aimed to investigate whether the different possible causal structures modify the distortion of the empirical evidence. Participants were instructed about a causal chain, a common cause or a common effect model (Fig. 1). Their task was to investigate conditional probabilities between the direct and indirect causal relations. As outlined above, applying the Markov condition to these causal models leads only to a mismatch between data and model-based inferences in the chain model and in the common-cause model, but not in the common-effect model.

Experiment 1

Experiment 1 studied conditional probability judgments after successive trial-by-trial learning of two generative causal relations, $A \rightarrow B$ and $B \rightarrow C$, which were instructed to be part of a causal chain. Assuming the causal Markov condition, these relations imply a positive contingency between *A* and *C*. However, the learning data showed no statistical dependency between *A* and *C*, that is, $P(C|A) = P(C|\neg A) = 0.5$. We explored how the structure of the learning course, such as repeated queries about $P(C|A)$ might affect learners' estimates.

Methods

Design Experiment 1 had three conditions, each of which comprised eight learning phases and up to 12 test phases. Figure 2 depicts the succession of the phases. In all conditions, in the final test phase (P20) we requested estimates of the conditional probability $P(C|A)$ (Fig. 2). The state of all three events *A*, *B*, and *C* was presented simultaneously during learning, although the instructions focused participants on the direct relations of the causal chain, $A \rightarrow B$ and $B \rightarrow C$. Moreover, the directly linked pairs were circled to highlight their causal relation. After each learning phase, participants were requested to give probability estimates of the respective direct causal relation ($A \rightarrow B$ or $B \rightarrow C$). Because of the focus on direct relations we expected a substantial influence of structural knowledge.

In Condition 2 (C2), participants were also focused on the direct causal relations, but the conditional probability estimates of the indirect relation ($P(C|A)$) were additionally requested several times during learning (Fig. 2). This procedural change was intended to draw participants' attention to the indirect relation as well. We expected that repeated testing of the *A*-*C* relation would strengthen the influence of the empirical data.

Condition 3 (C3) served as control condition to ensure that participants used the scales correctly and were able to detect the zero contingency between *A* and *C*. In this condition participants only received the subset of information about the relation between *A* and *C* (cf. Fig. 2, Table 2).

Participants Sixty students from the University of Göttingen took part in the experiment for course credit or were paid 5€. They were randomly assigned to the conditions.

Procedure and Material Participants were instructed to take the role of a developmental biologist investigating newts that undergo a metamorphosis. The metamorphosis proceeded in three stages. In each stage a particular type of carotene (Alpha, Beta, and Gamma; henceforth denoted as *A*, *B*, and *C*) could occur or not occur. These carotenes may or may not affect the presence of other carotenes in a later stage.

Learning and Test Phases

		P1	P2	P3	P4	P5	P6	…	P20
Condition	C1	Learn $A \rightarrow B$, C	Test $P(B\|A)$	Learn A, $B \rightarrow C$	Test $P(C\|B)$	–	Learn $A \rightarrow$B, C	…	Test $P(C\|A)$
	C2	Learn $A \rightarrow B$, C	Test $P(B\|A)$	Learn A, $B \rightarrow C$	Test $P(C\|B)$	Test $P(C\|A)$	Learn $A \rightarrow B$, C	…	Test $P(C\|A)$
	C3	Learn $A \rightarrow C$	–	Learn $A \rightarrow C$	–	Test $P(C\|A)$	Learn $A \rightarrow C$	…	Test $P(C\|A)$

Figure 2: Design of Experiment 1

In the first condition (C1) participants were asked to assess one of the two causal relations after each learning phase (cf. Fig. 2), alternating between the first relation ($A \rightarrow B$) and the second ($B \rightarrow C$). Although in all learning phases all three events were shown, participants were only asked about the indirect relation between *A* and *C* after all eight learning phases. Condition 2 was similar, but here participants were asked to assess the indirect relation between *A* and *C* after every other learning phase (see Fig. 2). In the control condition (C3) participants only observed the relation between *A* and *C*. After every second learning phase learners had to assess $P(C|A)$.

In the two experimental conditions (C1 and C2) information about the state (present vs. absent) of all three types of carotene was presented in a trial-by-trial learning procedure. Table 2 shows the learning input. In each of the eight learning phases, 24 newts were presented in randomized order. In total, 192 newts were shown. The empirical conditional probabilities were: $P(B|A) = P(C|B) = 0.75$, $P(B|\neg A) = P(C|\neg B) = 0.25$, $P(C|A) = P(C|\neg A) = 0.5$. Thus, in the data there was a zero contingency between *C* and *A*. The probabilities entailed by the chain model assuming the Markov condition were $P(C|A) = 0.625$ and $P(C|\neg A) = 0.375$, that is, a positive contingency.

Table 2: Learning data in Experiments 1 and 2.

Pattern			Phase	Total
A	B	C	6	48
A	B	¬C	3	24
A	¬B	¬C	3	24
A	¬B	C	0	0
¬A	B	¬C	0	0
¬A	B	C	3	24
¬A	¬B	C	3	24
¬A	¬B	¬C	6	48
All			24	192

Based on the outlined design (cf. Fig. 2) three types of test phases were used, in which we assessed participants estimates of the relations between *A* and *B*, *B* and *C*, and *A* and *C*. For each judgment we used a rating scale ranging from -100 to +100. For instance, when accessing $P(C|A)$, participants were asked whether newts that had developed Alpha carotene (*A*) in the first stage rather tended to develop

Gamma carotene (*C*) or to develop no Gamma carotene (¬*C*) in the subsequent stage. The scale ranged from -100 ('newts with Alpha carotene never develop Gamma carotene') to +100 ('newts with Alpha carotene always develop Gamma carotene') in steps of 10. The middle point of the scale, 0, was labeled 'Alpha and Gamma carotene occurred together only by chance' (i.e., with $P(C|A) = 0.5$).

Results

Figure 3 shows the means of participants' ratings concerning the probability of *B* given *A*, of *C* given *B*, and of *C* given *A* over the course of learning (the different measurement points are denoted as t1 to t4).

Panels 1 and 2 reveal that participants detected the positive causal relation between the directly linked events quickly and rated the probabilities $P(B|A)$ and $P(C|B)$ roughly correctly ($P(B|A) = P(C|B) = .75$ or +50 on the used scale). Panel 3 shows learners' estimates of the indirect relation $P(C|A)$. The results suggest that in both experimental conditions (C1 and C2) the estimates were affected by structural knowledge. While in the control condition (C3) learners' estimates were around zero (corresponding to a probability of $P(C|A) = 0.5$), a very different pattern of judgments was obtained in the two experimental conditions. In both condition C1 and C2 participants gave judgments above zero; and the obtained estimates also differed from the control condition (C3). The results of C1 complement previous findings by showing that abstract causal knowledge guides learning and reasoning even when people are provided with almost 200 trials on the state of all three variables with a shown objective zero contingency. Nonetheless, the average estimate of $P(C|A)$ was actually about as high as if it were exclusively based on inference assuming the Markov condition (cf. Equation 1, $P(C|A) = .625$, or +25 on the used scale). The second condition (C2) illustrates the interplay between abstract causal knowledge and empirical evidence over the course of learning. From the first (t1) to the last measurement (t4) participants' estimates of the indirect relation *A*-*C* declined, showing the influence of the learning data. Nevertheless, even in the last test phase (t4) the judgments were above zero and higher than in the control condition. An analysis of variance of the final judgments with the three conditions as between subjects factor yielded significant results, $F(2, 56) = 6.95$, $p < .05$, $MSE = 594.0$. Additionally, we computed pair-wise comparisons, with a significant contrast between C1 and C2 ($M_{C1} = 24.5$, $M_{C2} = 8.5$) ($F(1, 56) = 17.21$, $p < .0001$), as well as between C2 and C3 ($M_{C3} = -7.5$; $F(1, 56) = 4.40$, $p < .05$) and between C1 and C2 ($F(1, 56) = 4.31$, $p < .05$).

In sum, Experiment 1 supports the idea that subjects' judgments for indirect causal relations were derived from causal model representations obeying the Markov condition, even when the available evidence indicated that this condition did not hold. C1 shows that even after a long period of learning of zero contingencies, a positive contingency between the initial and final event was inferred.The difference between C1 and C2 shows that the impact of evidence also depends on the attentional focus during learning: when the attention is directed more clearly to the indirect relation, the distortion of the learning by top-down inferences is reduced. But even after almost 200 trials the bias did not disappear completely.

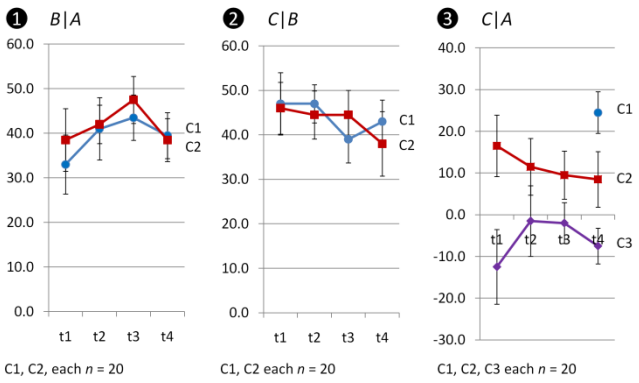


Figure 3: Mean judgments (±SE) in Experiment 1. t1 to t4 denote measurements at different points in time over the course of learning.

Experiment 2 – Causal Models

In Experiment 2 we investigated further causal structures. In addition to a causal chain we also used a common cause and a common effect model (Fig. 1). The learning data presented to participants were identical in all conditions and corresponded to Experiment 1 (Table 2). Although in the experiment participants were confronted with identical data about the three events the mapping of the events to their causal roles differed. In the chain condition *A* caused *B* and *B* caused *C*, in the common cause condition *B* was the common cause of *A* and *C*, and in the common effect condition *A* and *C* were independent causes of their common effect *B* (Fig. 1). If participants' mental causal models obeyed the Markov condition, increased values of $P(C|A)$ should be obtained in the chain and the common cause condition, but not in the common effect condition. Due to the lack of a mismatch between model and data in the CE model, participants should provide ratings corresponding to the empirical conditional probability of $P(C|A) = 0.5$.

Methods

Participants 150 students from the University of Göttingen participated for course credit or 5€. They were randomly assigned to one of the three causal model conditions.

Procedure and Material The procedure was almost identical to Condition 2 of Experiment 1 (Fig. 2), apart from the manipulations of the initial causal model assumptions. A different cover story was used, concerning the development of the metabolism of ravens. As causes and as effects we used three substances, which could be present or absent in different developmental stages of the ravens: Xantan, Yojan, and Zetosan (henceforth denoted as *A*, *B*, *C*). Participants in all condition were informed that they would have to answer questions about the potential direct causal relations (be-

tween *A* and *B*, and between *B* and *C*) as well as about the indirect relation between *A* and *C* after the learning phases. The causal links be present or absent.

The task investigated whether people assumed the Markov condition to hold when integrating single links into a larger causal structure. Although the instruction may well be interpreted to put a higher prior probability on the respective causal structures, the instructions were completely silent on whether one should assume the Markov condition. Hence, this provides a test for whether participants implicitly asserted the Markov condition and distorted the empirical probabilities accordingly.

Like in Condition 2 of Experiment 1 there were eight successive learning phases showing all three events *A*, *B*, and *C*. Again participants were focused on the respective direct causal relationship by the instructions and a circling of the directly related events. The data patterns were randomized within each learning phase; the learning data was identical in all conditions (Table 2).

In the test phases participants were again asked to assess conditional probabilities on a scale between +100 and -100 (cf. Experiment 1). When investigating the direct relations between *A* and *B* and *B* and *C* we assessed conditional probabilities in the causal direction (chain: $P(A|B)$, CC: $P(B|A)$). But note that in our learning data both conditional probabilities were identical ($P(B|A) = P(A|B)$). The wording of the question for $P(C|A)$ was identical, irrespective of condition.

Learning and Test Phases

P1	P2	P3	P4	P5	P6	…	P20
Learn $A \rightarrow B$, C	Test $P(B\|A)$	Learn A, $B \rightarrow C$	Test $P(C\|B)$	Test $P(C\|A)$	Learn $A \rightarrow B$, C	…	Test $P(C\|A)$

Figure 3: Design of Experiment 2.

Results

Figure 4 shows participants' mean estimates in the three conditions (Panel 1 – 3) across the four test phases (t1 – t4). Estimates of $P(B|A)$ and $P(C|B)$ were all positive, although they underestimated the correct value. With regard to the crucial estimate, $P(C|A)$, an inspection of the data reveals that the results of the chain condition replicate the results of Exp. 1, but that the expected effect for the CC model was not obtained. Consistent with our predictions, participants' estimates in the CE condition were close to zero. We conducted an ANOVA with the test phases (t1 to t4) as within-subject factor and causal structure (Chain, CC, CE) as between-subject factor. This resulted in a significant main effect of causal structure, $F(2, 147) = 4.34$, $p < .05$, $MSE = 2312$. No other effects proved significant. The pair-wise contrasts between the chain and the CE condition and between the chain and the CC condition yielded significant differences, $F(1, 147) = 5.31$, $p < .05$ and $F(1, 147) = 7.52$, $p < .01$. However, the contrast between the CC and CE condition was not significant: $F(1, 147) = 0.19$, $p = .66$. A test of the mean estimates of $P(C|A)$ against zero showed that only the chain condition consistently and significantly differed from zero, with no reduction over time. (Chain: t1, $t(50) = 2.49$, $p < .05$; t2, $t(50) = 2.04$, $p < .05$; t3, $t(50) = 2.98$, $p < .01$; t4, $t(50) = 3.27$, $p < .01$; CC: t1, $t(50) = -0.90$, $p = .37$; t2, $t(50) = -0.66$, $p = .51$; t3, $t(50) = .43$, $p = .66$; t4, $t(50) = -0.41$, $p = .68$; CE: t1, $t(50) = -0.19$, $p = .84$; t2, $t(50) = .99$, $p = .32$; t3, $t(50) = -0.49$, $p = .62$; t4, t1, $t(50) = .15$, $p = .88$).

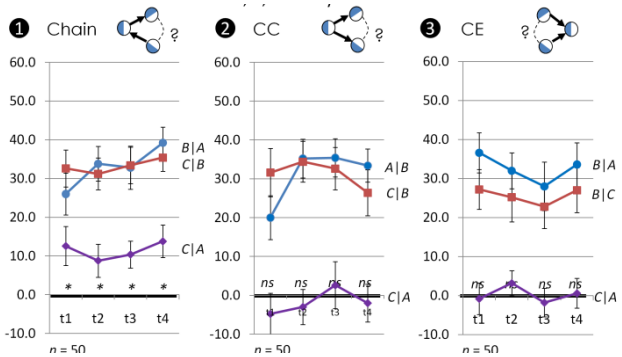


Figure 4: Means (±SE) of conditional probability estimates on a scale from -100 to +100 for the three causal structures (chain, common cause (CC), and common effect (CE)) across the four test phases (t1 to t4).

In sum, Experiment 2 replicated the biasing effect of structural knowledge with causal chains. As predicted by Bayes nets, no such effect was found for the common effect model for which top-down assumptions and empirical evidence were consistent with each other. Interestingly, no effect was obtained for the common cause model.

We can only speculate why we did not find an effect in the CC condition. Maybe the Markov condition is more intuitive in causal chains, in which the intermediate event can be easily represented as separating the initial from the final event. In contrast, screening-off relations may be harder to envision in common cause structures in which the intermediate event simultaneously causes several effects (see Cartwright, 2001). Actually, von Sydow et al. (2009) suggested that CC structures may often be interpreted to violate the Markov assumption, at least if one is concerned with the predication of attributes of a category (without representing alternative causes of the attributes). Attributes of categories are often represented as CC structures (Rehder, 2003). It has been argued that people may represent different kinds of noisy logical interaction patterns of such attributes (including XOR) (von Sydow, 2009). If such judgments correspond to a *causal* logic of CC structures, they would violate the assumption of conditional independence and unconditional positive correlation between effects (the Markov condition). However, further research is needed to connect models of noisy logical predication with theories of causal induction.

Another possibility may be that attentional factors caused the low ratings in the CC and CE condition, since we switched the direction of the question formats for the local causal links (e.g., $P(A|B)$ in the chain and $P(B|A)$ in the CC condition). Although, a predictive question format seemed to be most natural to elicit the causal representations that we

aimed to manipulate, this remains a factor that should be controlled for in future research.

General Discussion

The results of Experiment 1 corroborate our prediction that in a causal chain $A \rightarrow B \rightarrow C$ conditional probability judgments about the indirectly linked events A and C will be distorted by structural assumptions of the underlying causal model. We investigated the influence of causal inferences based on the Markov condition when learning such relations. Going beyond previous studies (Baetu & Baker, 2009; Ahn & Dennis, 2000), we provided data on the indirect relation, which showed a zero contingency. Hence, transitivity did not hold in the data (cf. von Sydow et al., 2009). In Experiment 1 we investigated this issue in a trial-by-trial learning scenario, assessing the role of repeated questions. The conditional probability estimates of $P(C|A)$ matched the values that would have been predicted if people estimated this probability based on their knowledge about the direct relations and structural assumptions about causal models (i.e., the Markov condition). This biasing effect was remarkably stable even if people obtained contradicting empirical evidence in several learning phases and were repeatedly queried about the indirect relation, which was intended to draw participants' attention to the indirect relation. With repeated queries the influence of causal reasoning became smaller, but did not disappear even after almost 200 trials.

Experiment 2 confirmed that chains and common effect structures ($A \rightarrow B \leftarrow C$) led to different judgments of $P(C|A)$ despite identical learning input. As predicted by causal Bayes nets, a biasing effect only occurred in the chain condition in which the data violated the structural constraints underlying chains. Consistent with this idea, no influence of structural knowledge was obtained for the common effect model. Interestingly, in the common cause structure ($A \leftarrow B \rightarrow C$) we did not found an influence of the causal model on participants' judgments. The reasons for this failure are unclear at present. One hypothesis may be that people find the Markov condition less plausible for these models (see also von Sydow et al., 2009). Alternatively, attentional effects during learning may have had an effect.

Taken together, the results provide further evidence for our claim that people try to create consistency between structural top-down knowledge and empirical evidence when making probabilistic causal inferences (von Sydow et al., 2009; cf. also Waldmann et al., 2010).

Acknowledgments

This research was supported by a grant 'Bayeslogik' by the *Deutsche Forschungsgemeinschaft* (DFG, Sy 111/1-2 [MvS]). We thank Johanna Frisch and Deborah Wolff for their help and assistance with the data collection.

References

Ahn, W., & Dennis, M. (2000). Induction of causal chain. *Proceedings of the 22nd Annual Conference of the Cognitive Science Society* (pp. 19-24). Lawrence Erlbaum Associates, NJ: Mahwah.

Baetu, I., & Baker, A. G. (2009). Human judgments of positive and negative causal chains. *Journal of Experimental Psychology: Animal Behavior Processes. 35(2)*, 153-168.

Cartwright, N. (2001). What is wrong with Bayes nets? *The Monist, 84*, 242-264.

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review, 111*, 3-32.

Mayrhofer, R., Goodman, N. D., Waldmann, M. R., & Tenenbaum, J. B. (2008). Structured correlation from the causal background. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 303-308).

Pearl, J. (2000). *Causality: Models, reasoning, and inference.* Cambridge, MA: Cambridge University Press.

Rehder, B. (2003). A causal-model theory of conceptual representation and categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*, 1141–1159.

Rehder, B., & Burnett, R. (2005). Feature inference and the causal structure of categories. *Cognitive Psychology*, *50*, 264-314.

Sloman, S. (2005). *Causal Models. How People Think about the World and Its Alternatives*. Cambridge, MA: Oxford University Press.

Spirtes, P., Glymour, C., & Scheines, R. (1993). *Causation, prediction, and search*. New York: Springer-Verlag.

von Sydow, M. (2009). On a general Bayesian pattern logic of frequency-based logical inclusion fallacies. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 248-253). Austin, TX: Cognitive Science Society.

von Sydow, M., Meder, B., & Hagmayer, Y. (2009). A transitivity heuristic of probabilistic causal reasoning. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 803-808). Austin, TX: Cognitive Science Society.

Waldmann, M. R., & Hagmayer, Y. (2001). Estimating causal strength: The role of structural knowledge and processing effort. *Cognition, 82*, 27-58.

Waldmann, M. R., Hagmayer, Y., & Blaisdell, A. P. (2006). Beyond the information given: Causal models in learning and reasoning. *Current Directions in Psychological Science, 15*, 307-311.

Waldmann, M. R., & Holyoak, K. J. (1992). Predictive and diagnostic learning within causal models: Asymmetries in cue competition. *Journal of Experimental Psychology: General, 121*, 222-236.

Waldmann, M. R., Meder, B., von Sydow, M. & Hagmayer, Y. (2010). The Tight Coupling between Category and Causal Learning. *Cognitive Processing, 11*, 143-158.

Group Stratification and Coordination Failure in a Continuous N-Player Stag Hunt

Seth Frey (sethfrey@indiana.edu)
Cognitive Science Program, 819 Eigenmann, 1910 E. 10th St.
Indiana University, Bloomington, IN 47406

Robert L. Goldstone (rgoldsto@indiana.edu)
Psychology Building, Room 338, 1101 E. 10th St.
Indiana University, Bloomington, IN 47405

Abstract

We reveal spontaneous group formation and differentiation in an online dynamic coordination experiment. We observe increased group stratification and attribute it to increases in pairwise cooperative behavior, rather than uncooperative behavior. Our network analyses document the fine scale structure of coordination failure in the face of many established determinants of coordination success. We explore previous work in coordination failure to frame our own findings. Factors that have been previously shown to improve coordination in discrete-time, forced-decision experimental games do not prevent decisive coordination failure in our real-time, asynchronous group decision-making environment.

Keywords: coordination; coordination failure; n-player games; continuous-time games; stag hunt; functional networks; group structure

Introduction

Sixty years of literature have established cooperation as only a special-case outcome in experiments of economic behavior. Where there is a conflict between individual and group interest, the net benefit of all members may suffer in favor of individual gains. Even in coordination games, where individual payoffs are directly related to the group's outcome as a whole, the maximum levels of coordination and personal benefit have been notoriously difficult to attain in the laboratory (Devetag & Ortmann, 2007), either because of uncertainty inherent in the task or strategic uncertainty with regard to the actions of other participants. However, the majority of these coordination situations involve slow, round-based decisions. By contrast, humans and animals often make decisions within the time scales that characterize processing and decision time. In these environments, decision opportunities may be presented with little notice, and part of the decision process is deciding when to decide. Despite the increased cognitive demands, the ability of animal groups to coordinate successfully in these environments is well established in both experimental and observational experiments (Conradt & Roper, 2005; Petit, Gautrais, & Leca, 2009; Couzin, 2009). Recent research on group behavior has worked to understand how these faster-paced decision environments affect coordination.

We build on this work, applying dynamic network methods to estimate the changes in network structure between participants in an n-person stag hunt coordination game. Though we find support for many structural factors that have been proposed (all else being equal) to promote efficient outcomes, participants fail to coordinate on payoff-dominant equilibria, and coordination continues to decay over the course of the experiment. We take advantage of the rich data available from these fine timescale experiments and introduce functional network measures to estimate the emergent structure of participant interactions.

Literature

Experiments in real-time group decision-making document both coordination failure and success. Dyer et al. report a collective navigation experiment in which human groups walked in the induced direction, despite constrained communication, and conflicting individual information (Dyer et al., 2008). Kearns, Suri, and Montfort examined coordination in a network experiment based on a map-coloring problem to investigate the effects of group topology and information differences on cooperation (Kearns et al., 2006). Roberts and Goldstone also looked at the effect of different information treatments in a collective number-guessing game and found that participants spontaneously adopt specific roles without communication(Roberts & Goldstone, 2009). Furthermore, the extent to which group members differentiated themselves predicted how well they solved the coordination problem. Related results have been found in minority and market entry games (Bottazzi & Devetag, 2007; Duffy & Hopkins, 2005). In another network game, continuous-time decision-making is proposed to improve coordination (Berninghaus, Ehrhart, & Ott, 2006). In all of these domains, participants were generally successful at resolving conflicts and finding solutions that brought a net collective benefit.

With continued ties to the study of coarse time-scale political and macroeconomic decision making (as in Schelling, 1980), experimental economists are traditionally interested in synchronous, normal form games wherein all players choose a strategy without knowledge of any other players' choices. However, there is still rich research in games that elicit decisions at finer time scales, as discussed by Berninghaus, Ehrhart, and Keser (1999). This literature has more often found coordination failure than success. Participants in the three games of E. Friedman, Shor, Shenker, & Sopher (2004) frequently failed to converge on any equilibrium, much less those that dominate with respect to payoff or certainty. However, Bottazzi and Devetag find emergent structure that leads to higher coordinative outcomes (2007) and Cheung and

Friedman observe successful coordination in a more complex continuous-time experiment modeling financial speculative attacks (2009).

From a game-theoretic standpoint, the game in this experiment shares several features with so-called stag hunt games, which have been extensively studied in the lab. In the allegorical stag hunt, two neighbors decide whether to hunt hare separately, or to work together to hunt (the more rewarding) stag. Even though both individuals benefit by coordinating for the larger quarry, they may decide that the costs of coordination, and their uncertainty as to the actions of their neighbor, are too great. In its synchronous form, game theory does not predict which outcome is more likely without refinements like risk and payoff dominance (Harsanyi & Selten, 1988). However, in the asynchronous, extended form, in which one neighbor is forced to make a choice that the other can observe beforehand, theory predicts that the rational player will select the high payoff equilibrium (Kuhn, 2009). Similarly, continuous-time games in the laboratory have led to more effective coordinative behavior than comparable synchronous games (D. Friedman & Oprea, 2009).

While real life uncertainty and decision cost dilemmas modeled by the stag hunt are at least as common for groups as for pairs of people, generalized n-player stag hunts have only been investigated in simulation (Pacheco, Santos, Souza, & Skyrms, 2009). However, as pointed out by Resnick (2007), the important coordination experiments of Van Huyck et al. (1990) are relevant. The precedents in both of these investigations seem to predict that adding players to the stag hunt makes coordination failure more likely in our experiment.

We introduce a coordination game building on previous evidence of emergent social conventions in repeated coordination games (Bottazzi & Devetag, 2007; Rankin, Huyck, & Battalio, 2000; Roberts & Goldstone, 2009; Uzzi & Spiro, 2005). Taken together, the above theoretical and experimental results make contrasting predictions about how participants should perform in the asynchronous environment provided by the real-time nature of our group experiment.

Experiment

Paradigm

Groups of two to six participants played an n-player stag hunt game on networked computers. Each player was shown information about the payoff and location of the other uniquely identified participants (Figure 1). Participants were instructed to "harvest" points from any of twelve game tiles. Participants were awarded a tile's points after waiting for a specified amount of time. To introduce the incentives of the stag hunt, tile payoffs increased with each tile's *coordination number*. The coordination number specified the quorum of participants necessary to harvest the points on a tile. When a tile reached its quorum, a visible timer on the tile would start counting down to zero. During this time, any participant on the tile could leave it and other participants could join. If a timer reached zero, each participant on the tile received the tile's payoff, even if participant presence on a tile was above quorum. If the tile fell below quorum during countdown, the timer was reset and no points were distributed.

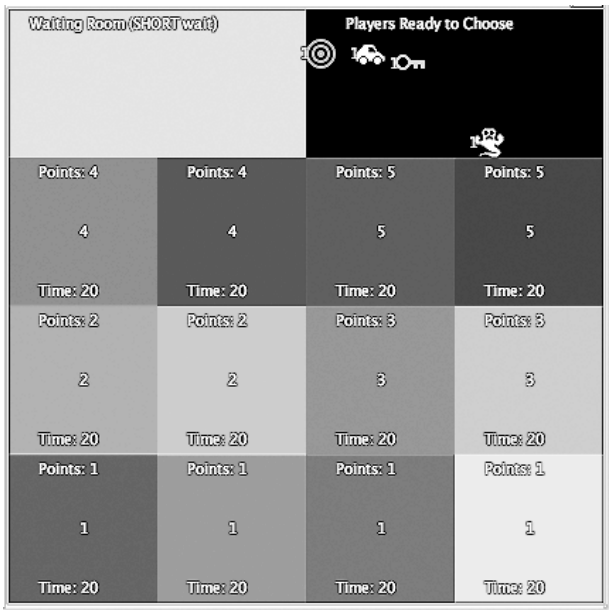


Figure 1: Screenshot of game board. Icons are controlled by participants, each of whom may select any tile. The coordination number in the center of each tile reflects how many participants must wait for the listed Time to each receive the specified number of Points. The number to the left of each player reflects his or her accumulated number of points.

The coordination number ranged from one to five, even in cases where there were more than or fewer than five participants in a session. To allow for the possibility of independent action, tiles at each possible coordination level (of five) were distributed redundantly among all possible tiles (twelve). This allowed participants to harvest points alone on tiles with a coordination number equal to one, or to harvest tiles collectively with other participants on tiles with higher coordination numbers. We may have introduced theoretical difficulties by providing multiple tile choices at a given coordination level, but participants showed an ability to coordinate at higher levels, perhaps due to focal point effects or the cheap decision costs of the environment.

In this game, the most efficient strategy was to repeatedly choose one tile with the coordination number equal to the group size. The strategy least vulnerable to strategic uncertainty was to always choose a tile with coordination number equal to one.

Participants started in a "staging area" on the game board from which they selected a tile to harvest from. After distributing its points, a harvested tile was reset and its participants were held for two seconds before being returned to the staging area. Participants were never forced to make any choice, and the ability of any participant to make a choice did not depend on "turns" or any system of rounds that structured when participants made decisions relative to each other.

We thus distinguish between the real-time, asynchronous and non-forced aspects of decision-making in this environment. This less constrained structure allowed us to explore the effects of changes in pay level, risk, and experience on spontaneous group coordination and structure in an environment that evokes real world, short time-scale decision-making.

Manipulations

After a practice round, participants played eight five-minute rounds of the game (permitting up to 75 consecutive harvesting events). We manipulated *pay level* and *harvest time* over rounds in a $2 \times 2 \times 2$ block factorial design ([low versus high payoff]$\times$[long versus short harvest time]$\times$[first versus second block]). The first four rounds (block one) consisted of the 2×2 set of low and high payoffs and short and long harvest times, randomly permuted, and were followed by a second random permutation of the same four conditions. The blocks controlled for order effects and enabled investigation of the effects of experience with the group.

The two levels of pay level were *low* (tile payoff is the same as the tile coordination number) and *high* (tile payoff is the square of tile coordination number). Scaling payoffs with coordination number created an incentive for participants to risk coordinating with other participants for a higher payoff. The two levels of harvest time, the number of seconds necessary to harvest any tile, were *fast* and *slow* (two and ten seconds, respectively). This factor manipulated the amount of time that a tile had to remain at or above quorum to deliver payoffs, and thus how vulnerable participants were to strategic uncertainty—uncertainty as to the beliefs and future actions of their peers.[1]

Participants were introductory psychology undergraduates receiving course credit. Participants did not receive monetary compensation, but observational reports indicate that participants found earning points intrinsically motivating. Participants often spontaneously cheered when they received a large payoff. There were forty-three participants over twelve experiments. Our experiments ultimately had five levels of *group size*, with four groups of 2 participants, two groups of 3, two groups of 4, three groups of 5, and one group of 6. Though it was not controlled, our analysis will treat group size as an independent, between-participant variable. Each experiment lasted one hour.

Measures

The structure of the experiment allowed us to develop and compare a number of measures of coordination. The game was implemented such that the state of each player was recorded approximately twice every second. This enabled investigation of each individual's decisions as a time series, and of participants' decisions together as a multivariate time series.

[1]Strategic uncertainty is strictly uncertainty as to the beliefs and actions of other participants (Huyck et al., 1990), but this experiment provides full information as to the actions of peers in the experiment.

For each participant, in each condition, we recorded *payoff*, *wait tile*, and *success tile*. Payoff represents how many points a participant earned in a condition. Wait tile represents the mean tile that a participant spent time on. Success tile represents the mean tile that a participant successfully harvested a payoff from. In terms of these measures, highly successful coordination would be reflected by maximum payoffs and wait and success tile values equal to group size, all identical across participants. Conversely, zero-coordination behavior would be reflected in minimal payoffs and wait and success tiles equal to one.

We also used the multivariate nature of the time data to calculate measures of functional proximity between participants in a group. These measures reflect the extent to which any pair of participants coordinated on a tile. We then assembled these dyad weights into fully connected networks representing the pattern of behavioral couplings in the group. Since these networks are fully connected, edge weights replace edge presence in representing heterogeneity within a group. With these graphs of internal group structure, we used a variety of network statistics as additional measures of coordination. These "functional networks" have been used in computational neuroscience to infer how regions of the brain are related, and to determine the extent to which dynamic functional relations correspond to physical anatomical connections (Bullmore & Sporns, 2009; Hagmann et al., 2008). Given the rich data available in the experiment, and the wealth of theoretical issues common to neuroscience and group behavior (Couzin, 2009), the extension of these tools to the study of social behavior was natural. Because functional networks have only recently been applied to the study of group behavior (Nagy, Akos, Biro, & Vicsek, 2010), we implemented three separate measures of proximity: choice distance, mutual information (Cover & Thomas, 2006) and transfer entropy (Schreiber, 2000). Because the three measures gave analogous, consistent results, this investigation will treat only the choice distance, the simplest of the measures.

The *choice distance* between a pair of participants is the number of seconds that they were on different tiles. In the resulting graph, two participants that always made the same choice are *close*, with a choice distance of zero. If their choices were always different, they are *far*, with a distance equal to the number of seconds that they could have been on the same tile. The matrix representing this graph is symmetric, and its diagonal, representing the number of times a participant is on his or her own tile, is always zero.

For every condition, we calculated the mean and variance of each individual's proximity to every other participant in their group. The mean proximity is related to the closeness centrality metric in graph theory and social network theory (Wasserman & Faust, 1994), but it preserves measure units. A change in a participant's mean proximity across some experimental condition entails that the participant coordinated more or less closely, on average, with all other participants in

their experiment. The image of a loaf of raisin bread provides a simple visualization of a net change in average choice distance. If yeast caused the loaf to rise, each raisin will be farther from every other. By analogy, an increase in mean choice distance is a net expansion of the graph. A net increase in the variance of this distance, which we interpret as stratification, corresponds to a case where the raisins in the bread all start equidistant from each other, and differently attract and repel each other over time. Together with wait tile and success tile, these graph statistics provide a powerful toolkit for investigating the internal structure of groups in our coordination game.

Table 1: Payoffs

Payoff	Observed		Min:Max[2]		Efficiency	
Harv. time	slow	fast	slow	fast	slow	fast
Low pay	21.5	62.7	25:99.4	75:298	-5%	-3%
High pay	68.1	181	25:430	75:289	9%	9%

Results

Group structure

Groups stratified with time, as seen in an increase by block in the variance of participants' choice distances from each other ($F(1,38) = 10.51, p < 0.05$). However, despite the reliable decrease in other measures of coordination (below), it does not seem that mean choice distance changed by block. The mean choice distance of participants from each other (the number of seconds that they spent on different tiles) was about 33 seconds in blocks one and two. The significant increase in standard deviation about this mean was from 7.5 seconds in the first block to 8.8 seconds in block two.

A change in the variance in a given participant's choice distances suggests an increase in graph heterogeneity: either a selective decrease in the minimum distances, a selective increase in the largest distances, or a combination of the two. An increase in the maximum distance would be consistent with choice refusal, a phenomenon observed in iterated prisoner's dilemma experiments by which participants choose not to reenter a game with a peer who has defected against them in previous iterations (Stanley, Ashlock, & Tesfatsion, 2004). Additionally, an increase in trust between only some peers in the group, and a corresponding net decrease in the minimum distance, could also account for the increase in variance with time. We found only insignificant evidence that the maximum choice distance increases with block ($F(1,38) = 2.34, p = 0.134$ with sample mean maximum distances of 79.8 and 83.1 for blocks one and two) and slightly stronger evidence that the minimum choice distance decreased with time ($F(1,38) = 3.88, p = 0.056$, with sample mean minimum distances of 54.6 and 50.2 by block). This experiment thus provides only minimal support for choice refusal and somewhat stronger support for trust building.

Coordination

In addition to these changes in group structure with time, we also documented decisive coordination failure. We define minimum efficiency as strictly individual, zero-coordination behavior (all players select tiles with coordination number one), and maximum efficiency as group-wide fully cooperative behavior.[2] Normalizing over these extremes, participants performed at negative, or very low efficiencies. For low pay level, observed efficiency was below that of purely individual behavior, presumably due to an excessive time cost of coordination, since waiting on a tile below quorum can yield no payoff. Similarly at the high pay level, the square scaling of payoffs by coordination number brought efficiency up only 9% above minimum (Table 1). Wait tile was on average 47% of maximum for each group size, and success tile was significantly lower at 35% ($t(656) = 6.49, p < .001$) (Table 2). This difference shows that participants consistently took risks on high coordination tiles that were not rewarded with success, even in the face of increasing coordination failure.

This failure of groups to coordinate at higher equilibria increased over the course of the experiment. Wait tile, success tile, and payoff all decreased slightly with block (all $p < 0.001, F(1,38) = 26.2, F(1,38) = 40.3$, and $F(1,42) = 14.4$). results on Table 3).

Only one exception appeared: one of the four 2-person groups had high coordination in the first block and reached near perfect levels of coordination in the second block. Although some groups did not produce significant decreases in coordination with time, this group was the only one to increase coordination, ultimately to the payoff-dominant equilibrium by the end of the experiment. On average, at group size two, wait and success tile stayed the same over block, and payoff did not increase significantly. This serve as limited evidence for the relative stability of two-player stag hunts in continuous time.

Supporting results

Another observed block effect was a significant reduction in participants' information entropy ($F(1,38) = 4.57, p < 0.01$), from 2.67 to 2.46 bits, suggesting that participants had more predictable behavior in the second block. This supports previous observations in a group minority game in which participants tended towards pure strategies over time (Bottazzi & Devetag, 2007), the key difference being that in the minority game experiment participants successfully coordinated towards benefits exceeding those predicted by the mixed-strategy equilibrium for that environment.

Both larger pay levels and shorter harvest times were strongly associated with higher wait tiles and higher success tiles (all $p < 0.001$, Table 3). Thus participants coordinated more when they were receiving higher payoffs, sup-

[2]Coordination number of all harvest tiles equals group size. Maxima in table calculated over four groups of 2, two groups of 3, two groups of 4, three groups of 5, and one group of 6. For the group of six, full cooperation implies that all six participants are on a tile of coordination five, the highest that we implemented.

porting the results of Brandts and Cooper (2006) in a minimum effort game. Participants' average choice distance correlated with wait tile ($\beta = -13.5, p < 0.05$) and success tile ($\beta = -20.0, p < 0.05$).

Table 2: Observed Wait and Success Tile by Group Size

Group size	2	3	4	5	6	ave	max[3]
Wait tile	1.5	1.7	2.6	2.9	3.1	2.5	3.98
% of max	.53	.37	.54	.49	.42	.47	1.00
Succ. tile	1.4	1.5	2.3	2.3	2.6	2.1	3.98
% of max	.40	.27	.45	.33	.31	.35	1.00

Discussion

Group structure

Investigation of participant time series revealed a detailed perspective into the evolution of group structure during the coordination task. Previous investigations into group behavior over time have focused on the differentiation of individuals into less complex roles (Bottazzi & Devetag, 2007; Roberts & Goldstone, 2009). This investigation extends these results by composing specific relationships into a group-level representation of interaction patterns. Groups tended to stratify in time as participants came to preferentially coordinate with and, to a lesser extent avoid, specific peers. In their model of Broadway musical production teams, Guimera et al. (2005) modeled the preference of past team members to work together in the future, but they included no complimentary aversion mechanism corresponding to choice refusal. Our results support this modeling decision.

Coordination failure

To our surprise, participants failed to coordinate on the most efficient outcomes and the extent of their coordination decreased over time. In their review of laboratory coordination failure, Devetag and Ortman propose a number of efficiency-enhancing design principles for coordination success (Devetag & Ortmann, 2007). Within the controlled factors of this experiment, we directly support two of the proposals in their review. We observed that an increase in pay level corresponded to increases in wait tile and payoff, supporting the efficiency-enhancing effect of "lowering the attractiveness of the secure action relative to the risky action required for the efficient equilibrium." Our manipulation of harvest time corresponded to both "lowering the costs of experimentation" and creating "less stringent coordination requirements," all of which are *ceteris paribus* efficiency-enhancing. In addition to this direct support, our design had a majority of the other features observed to be efficiency-enhancing: participants received zero deviation costs, they had repeated encounters in a fixed-group match, and participants had full information as to their peers' states and action choices. Also, although there was no explicit communication, participants' ability to make and change their decisions at low cost may have permitted signaling—some behavioral equivalent of "cheap talk" as Bottazzi and Devetag propose to explain their observations in a group minority game (2007)—thus providing another efficiency-enhancing factor.

In addition to implementing a majority of Devetag and Ortmann's efficiency-enhancing conditions, we also implemented a number of efficiency-enhancing conditions that they did not report. Our groups operated in a continuous time environment in which they made decisions asynchronously, and in which part of their decision was when, or whether, to make a decision. Continuous time environments have been observed to improve coordination in an experimental prisoner's dilemma (D. Friedman & Oprea, 2009) and a number of other studies (Berninghaus et al., 1999; Cheung & Friedman, 2009; E. Friedman et al., 2004). Asynchronous, extended form decision making is hypothesized to improve coordination in stag hunts (Kuhn, 2009). In the face of all of this previous qualitative evidence for high coordination outcomes, our participants still produced net coordination failure and increasing coordination failure with time, in all groups except one of the smallest.

Devetag and Ortmann make no claim for a simple additive relationship between their thirteen design principles, and they repeatedly stress that each factor is only efficiency-enhancing if all else is equal. Therefore, we think it is most reasonable to look to our experiment's major differences to explain the increasing failure of participants to coordinate on payoff-dominant equilibria. Despite the rich evidence supporting continuous time and asynchronous environments as efficiency-enhancing, we suspect that either of these factors or the third, non-forced decision making, may have overwhelmed the other efficiency-enhancing factors built into the experiment.

Conclusion

In an internet-enabled, computer-run coordination experiment, we show spontaneous group formation and differentiation. We also use functional network methods from computational neuroscience to document the fine scale structure of coordination failure in the face of many established determinants of coordination success. Within the experiment's controlled factors, we support previous claims with regard to the efficiency effects of payoff changes and uncertainty, but in the greater context of a coordination failure that increases over time.

Acknowledgments.

The authors wish to acknowledge Charlene Tay and the NSF, Research and Evaluation on Education in Science and Engineering, DRL-0910218

[3]This maximum is calculated over all groups of all group sizes.

Table 3: Main effects on Wait and Success Tile

	Low Pay Level	High Pay Level	Slow Harvest	Fast Harvest	Groups of Three	Groups of Six	Block One	Block Two
Mean wait tile	2.4	2.7	2.3	2.6	1.7	3.1	2.6	2.4
Mean success tile	2.1	2.3	2.0	2.1	1.4	2.6	2.2	2.0

References

Berninghaus, S., Ehrhart, K., & Keser, C. (1999). Continuous-time strategy selection in linear population games. *Experimental Economics*, *2*(1), 41–57.

Berninghaus, S., Ehrhart, K., & Ott, M. (2006). A network experiment in continuous time: The influence of link costs. *Experimental Economics*, *9*(3), 237—251.

Bottazzi, G., & Devetag, G. (2007). Competition and coordination in experimental minority games. *Journal of Evolutionary Economics*, *17*(3), 241–275.

Brandts, J., & Cooper, D. (2006). A change would do you good.... an experimental study on how to overcome coordination failure in organizations. *The American Economic Review*, *96*(3), 669-693.

Bullmore, E., & Sporns, O. (2009, Jan). Complex brain networks: graph theoretical analysis of structural and functional systems. *Nat Rev Neurosci*.

Cheung, Y., & Friedman, D. (2009, Jan). Speculative attacks: a laboratory study in continuous time. *Journal of International Money and Finance*.

Conradt, L., & Roper, T. (2005). Consensus decision making in animals. *Trends in Ecology & Evolution*, *20*(8), 449–456.

Couzin, I. (2009). Collective cognition in animal groups. *Trends in Cognitive Sciences*, *13*(1), 36–43.

Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory*. Wiley-Interscience.

Devetag, G., & Ortmann, A. (2007). When and why? a critical survey on coordination failure in the laboratory. *Experimental Economics*, *10*(3), 331–344.

Duffy, J., & Hopkins, E. (2005). Learning, information, and sorting in market entry games: theory and evidence. *Games and Economic Behavior*, *51*(1), 31–62.

Dyer, J., Ioannou, C., Morrell, L., Croft, D., Couzin, I., Waters, D., et al. (2008). Consensus decision making in human crowds. *Animal Behaviour*, *75*(2), 461–470.

Friedman, D., & Oprea, R. (2009). A continuous dilemma. *Department of Economics, UCSC*, 657.

Friedman, E., Shor, M., Shenker, S., & Sopher, B. (2004). An experiment on learning with limited information: nonconvergence, experimentation cascades, and the advantage of being slow. *Games and Economic Behavior*, *47*(2), 325–352.

Guimera, R., Uzzi, B., Spiro, J., & Amaral, L. (2005). Team assembly mechanisms determine collaboration network structure and team performance. *Science*, *308*(5722), 697.

Hagmann, P., Cammoun, L., Gigandet, X., Meuli, R., Honey, C., Wedeen, V., et al. (2008). Mapping the structural core of human cerebral cortex. *PLoS Biology*, *6*(7), e159.

Harsanyi, J., & Selten, R. (1988). A general theory of equilibrium selection in games. *MIT Press Books*.

Huyck, J. V., Battalio, R., & Beil, R. (1990). Tacit coordination games, strategic uncertainty, and coordination failure. *The American Economic Review*.

Kearns, M., Suri, S., & Montfort, N. (2006). An experimental study of the coloring problem on human subject networks. *Science*, *313*(5788), 824.

Kuhn, S. (2009). Prisoner's dilemma. In *Stanford encyclopedia of philosophy.* Metaphysics Research Lab, Stanford University.

Nagy, M., Akos, Z., Biro, D., & Vicsek, T. (2010). Hierarchical group dynamics in pigeon flocks. *Nature*, *464*(72907290), 890-893.

Pacheco, J., Santos, F., Souza, M., & Skyrms, B. (2009). Evolutionary dynamics of collective action in n-person stag hunt dilemmas. *Proceedings of the Royal Society of London, Series B: Biological Sciences*, *276*(1655), 315–321.

Petit, O., Gautrais, J., & Leca, J. (2009). Collective decision-making in white-faced capuchin monkeys. *Proceedings of the Royal Society B: Biological Sciences*, *276*(1672), 3495.

Rankin, F., Huyck, J. V., & Battalio, R. (2000). Strategic similarity and emergent conventions: Evidence from similar stag hunt games. *Games and Economic Behavior*, *32*(2), 315–337.

Resnick, E. (2007). Cooperation in multi-player stag hunt games.

Roberts, M. E., & Goldstone, R. L. (2009). Adaptive group coordination. In *Proceedings of the thirty-first annual conference of the cognitive science society.* Amsterdam, Netherlands.

Schelling, T. C. (1980). *The strategy of conflict*. Harvard University Press.

Schreiber, T. (2000). Measuring information transfer. *Physical review letters*, *85*(2), 461.

Stanley, E., Ashlock, D., & Tesfatsion, L. (2004). Iterated prisoner's dilemma with choice and refusal of partners. *Staff General Research Papers*.

Uzzi, B., & Spiro, J. (2005). Collaboration and creativity: The small world problem. *American Journal of Sociology*, *111*(2), 447–504.

Wasserman, S., & Faust, K. (1994). *Social network analysis: methods and applications*. Cambridge University Press.

Learning from Failures for Cognitive Flexibility

Dongkyu Choi (dongkyuc@uic.edu)
Stellan Ohlsson (stellan@uic.edu)
Department of Psychology
University of Illinois at Chicago
1007 W Harrison Street (M/C 285), Chicago, IL 60607 USA

Abstract

Cognitive flexibility is an important goal in the computational modeling of higher cognition. An agent operating in the world that changes over time should adapt to the changes and update its knowledge according to them. In this paper, we report our progress on implementing a constraint-based mechanism for learning from failures in a cognitive architecture, ICARUS. We review relevant features of the architecture, and describe the learning mechanism in detail. We also discuss the challenges encountered during the implementation and describe how we solved them. We then provide some experimental observations and conclude after a discussion on related and future work.

Keywords: cognitive architecture, constraints, constraint violations, learning from failures, skill acquisition

Introduction

In computational models of higher cognition, it is important to simulate the broad human functionality that we call *adaptability* or *flexibility*. Cognitive flexibility is, of course, a multi-dimensional construct, but in this paper, we focus specifically on the ability of humans to act effectively when a familiar task environment is changing, thus rendering previously learned skills ineffective or obsolete.

Traditionally, researchers discussed two types of error correction mechanisms for this problem. *Weakening* (Anderson, 1983, pp. 249–254) assumes that certain knowledge structures like rules, skills, schemas, or chunks have strengths associated with them, and it decreases the strength of the particular structure that generates a negative outcome. However, actions themselves are not typically correct or incorrect, or appropriate or inappropriate. Instead, they are appropriate, correct or, useful in some situations but not in others. The goal of learning from failure is thus to distinguish between the class of situations in which a particular type of action will cause errors and the class of situations in which it does not. Weakening does not accomplish this, because lower strength makes an action less likely to be selected in all situations.

Another mechanism proposed for error correction is *discrimination* (Langley, 1987). The key idea behind this contribution is to compare a situation with a positive outcome and another with a negative outcome to identify discriminating features. If an action generates both positive and negative outcomes across multiple situations, the system identifies any features that were true in one situation but not in the other, and uses them to constrain the applicability of the action. But the computational discrimination mechanism also has several problems including: the lack of criterion for how many instances of either type are needed before a valid inference as to the discriminating features can be drawn; the possible existence of a very large number of potential discriminating features, leading to complex applicability conditions or large numbers of new rules or both; and the inability to identify potential discriminating features with a causal impact from those of accidental correlation.

In response, Ohlsson (1996) developed a *constraint-based specialization* mechanism for learning from negative outcomes. The production system implementation of the mechanism overcomes most of the weaknesses of previous methods. It assumes that the agent has access to some declarative knowledge in the form of *constraints*, which consist of an ordered pair with a relevance criterion and a satisfaction criterion. The system matches the relevance criteria of all constraints against the current state of the world on each cycle of its operation. For constraints with matching relevance conditions, the system also matches the satisfaction conditions. Satisfied constraints require no response, but violated constraints signal a failed expectation due to various reasons including a change in the world or erroneous knowledge. This constitutes a learning opportunity, and the system revises the current skill in such a way as to avoid violating the same constraint in the future. The computational problem involved here is to specify exactly how to revise the relevant skill when an error occurs, and the constraint-based specialization provides a solution to this problem.

Unlike weakening, the constraint-based approach identifies the specific class of situations in which an action is likely (or unlikely) to cause errors. It also differs from the discrimination method, and the mechanism does not carry out an uncertain, inductive inference. Instead, it computes a rationally motivated revision to the current skill. However, these advantages were limited by a simplistic credit/blame attribution algorithm and the lack of serious architectural supports like other learning mechanisms that can operate in parallel. In this paper, we adapt the constraint-based specialization mechanism to a cognitive architecture, ICARUS, and address these limitations. The architecture features hierarchical knowledge structures, and it has a variety of well-developed capabilities including learning from positive outcomes (Langley & Choi, 2006). We first review the relevant features of the ICARUS architecture, and describe the constraint-based specialization mechanism in some detail. Then we identify the challenges we encountered during the implementation in ICARUS, with a particular attention to the credit assignment problem. Finally, we report some experimental observations with the system, and discuss related and future work.

The ICARUS Architecture

Cognitive architectures aim for a general framework for cognition. They include a set of hypotheses covering representation, inference, execution, learning and other aspects of cognition. Soar (Laird et al., 1986) and ACT-R (Anderson, 1993) are some of the well-known cognitive architectures, and the ICARUS architecture exhibits some similarities to them but has some important differences as well (Langley & Choi, 2006). In this section, we review the fundamental aspects of the architecture before continuing our discussion to the specifics of learning from failures in this framework.

Representation and Memories

ICARUS distinguishes conceptual and procedural knowledge. Concepts describe the environment, and enable the system to infer beliefs about the current state of the world. Skills, on the other hand, consist of procedures that are known to achieve certain goals. The architecture also distinguishes long-term, abstract knowledge and short-term, instantiated structures. Long-term concepts and skills are general descriptions of situations and procedures, and the system instantiates them before applying them to a particular situation. Instantiated concepts and skills are short-term structures, in that they are applicable only at a specific moment. ICARUS has four separate memories to support these distinctions.

The architecture encodes concepts with definitions that are similar to Horn clauses. They consist of a head and a body that includes perceptual matching conditions or references to other concepts. Table 1 shows some sample concepts. The first concept has a head, `(same-color ?block1 ?block2)`, and specifies perceptual matching conditions among the variables involved in its `:percepts` and `:tests` fields. It is a *primitive* concept, which does not have any reference to other concepts. The second concept also has a head and some perceptual matching conditions, but it has references to other concepts in the `:relations` field, and therefore, it is a *non-primitive* concept.

Table 1: Some sample ICARUS concepts for the Blocks World. Question marks denote variables.

```
((same-color ?block1 ?block2)
 :percepts  ((block ?block1 color ?color)
             (block ?block2 color ?color))
 :tests     ((not (equal ?block1 ?block2))))

((not-color-sorted ?color)
 :percepts  ((block ?block1 color ?color)
             (block ?block2))
 :relations ((on ?block1 ?block2)
             (not
              (same-color ?block1 ?block2))))
```

On the other hand, ICARUS' skills resemble STRIPS operators. The head of each skill is the predicate it is known to achieve, and therefore, all skills are indexed by their respective goals. Each skill has a body that consists of perceptual matching conditions, some preconditions, and either direct actions to the world or references to its subgoals. Like in concepts, skills with no references to any subgoals are *primitive*, while the ones with them are *non-primitive*. The hierarchical organization provides multiple layers of abstraction in the specification of complex procedures.

In Table 2, the first skill indexed by its goal `(stacked ?block ?to)` has some perceptual matching conditions and a precondition, `(stackable ?block ?to)`. It includes several direct actions in the world (marked with asterisks), and therefore, it is a primitive skill. The second skill, however, is a non-primitive one, with references to subgoals, `(stackable ?block1 ?block2)` and `(stacked ?block1 ?block2)`. The subgoals are ordered, and they invoke other skills that achieve them. For instance, the second subgoal will invoke skills like the first example in the table. In this manner, ICARUS's skills are hierarchically organized.

Table 2: Some sample ICARUS skills for the Blocks World.

```
((stacked ?block ?to)
 :percepts ((block ?block)
            (block ?to xpos ?xpos ypos ?ypos
                   height ?height))
 :start    ((stackable ?block ?to))
 :actions  ((*horizontal-move ?block ?xpos)
            (*vertical-move ?block
                            (+ ?ypos ?height))
            (*ungrasp ?block)))

((on ?block1 ?block2)
 :percepts ((block ?block1)
            (block ?block2))
 :subgoals ((stackable ?block1 ?block2)
            (stacked ?block1 ?block2)))
```

Inference and Execution

The ICARUS architecture operates in cycles. On each cycle, the system instantiates its long-term concepts based on the current situation. The bottom-up inference of concepts creates beliefs in the form of instantiated conceptual predicates. The inference process starts with the perceptual information about objects in the world. The system attempts to match its concept definitions to the perceptual information and, when there is a match, it instantiates the head of the definitions to compute its current beliefs.

Once the architecture computes all its beliefs, it starts the skill retrieval and execution process. ICARUS' goals guide this process, and the system retrieves relevant long-term skills based on the current beliefs. When it finds an executable path through its skill hierarchy, from its goal at the top to actions at the bottom, ICARUS executes the actions specified at the leaf node of the path. This execution, in turn, changes the environment, and the system starts another cycle by inferring the updated beliefs from new data received from the environment.

Problem Solving and Learning

During the execution for its goals, the architecture sometimes encounters a situation where it can not find any executable skill path. When this happens, ICARUS invokes its means-ends problem solver, chaining backward from its goal. It at-

tempts to use two types of *chains*, a skill chain that uses a goal-achieving skill with unsatisfied preconditions and a concept chain that decomposes the goal into subgoals through concept definitions. Once the system finds a subgoal with an executable skill during this process, it immediately executes the skill and continue to the next cycle until it achieves all the top-level goals.

When the architecture finds a solution and achieves a goal (which includes both the top-level goals and any of their subgoals), it learns new skills from the successful problem solving trace. The learned skills differs in their forms based on the type of the problem solving chain. Further discussions on the problem solving and learning capabilities would require more space than we can afford here, but Langley and Choi (2006) covers all the details. In the subsequent sections, we explain the details of the constraint-based specialization mechanism and its implementation in ICARUS.

Learning from Failures

As described in the previous section, ICARUS has hierarchically organized skill knowledge and it can learn from positive outcomes through problem solving. However, the architecture can not adapt to environmental changes when some of its existing skills become incorrect or obsolete. Extending ICARUS with the constraint-based specialization mechanism provides this capability.

Representation of Constraints

The extended architecture stores each constraint as a pair of relevance and satisfaction conditions, following Ohlsson and Rees (1991). Both relevance and satisfaction conditions are conjunctions of predicates, and the ICARUS architecture keeps a list of such pairs in a separate constraint memory.

Table 3 shows some sample constraints we use in the Blocks World domain. For convenience, we store each pair with a name like *color*, *top-block*, or *width*. The first constraint, *color*, says that two blocks should have the same color when they are stacked, which, in effect, enforces all towers to have a single color. Similarly, the other two constraints mean that a block that is designated as a *top-block* should always be clear, and that a block on top of another block should be smaller than the one below, respectively.

Table 3: Some sample constraints for the Blocks World.

```
(color     :relevance    ((on ?a ?b))
           :satisfaction ((same-color ?a ?b)))
(top-block :relevance    ((top-block ?b))
           :satisfaction ((clear ?b)))
(width     :relevance    ((on ?a ?b))
           :satisfaction ((smaller-than ?a ?b)))
```

Detection of Constraint Violations

On each cycle, the system checks if the current belief state satisfies all the constraints. It first attempts to match the relevance conditions of its constraints against the current state, and, if a match is found, verifies that the satisfaction conditions also hold. When it finds an unsatisfied constraint, it attempts to revise the skill that caused this violation.

We distinguish two different types of constraint violations. In the first type, a constraint just becomes relevant after an action but not satisfied at the same time. For instance, when an agent stacks a red block, *A*, on top of a blue block, *B*, it achieves `(on A B)`, so the corresponding instance of the color constraint in Table 3 matches and the constraint becomes relevant by the stacking action. But the satisfaction condition, `(same-color A B)`, is not met in this case, because one of the blocks is red and the other is blue. We refer to violations like this as *type A* violations.

Another type of violations, which we call *type B* violations, involves a constraint that has been relevant and satisfied, but becomes unsatisfied as a result of an action while it still stays relevant. An example of this type occurs when an agent stacks a block *C* on top of a block *TB* that is designated as a top block. In this case, the top-block constraint stays relevant before and after the stacking action, since the predicate, `(top-block TB)` continues to hold. But the satisfaction condition, `(clear TB)` becomes false as a consequence of the action, and the constraint is violated.

Skill Revisions

Once the system detects constraint violations of either type, it randomly chooses one of them and attempts to make revisions to the skill it just used. The revision process shares its basic steps with those used in previous research (Ohlsson, 1996; Ohlsson & Rees, 1991). The goal of this process is to constrain the application of the skill to situations in which it will not violate the constraint.

For a type A violation, where a constraint becomes relevant but violated, one of the revisions forces the constraint to stay irrelevant, and the other ensures that it is both relevant and satisfied. On the other hand, a type B violation, in which a constraint stays relevant but becomes violated, invokes one revision that makes sure the constraint is irrelevant, and another that restricts the use of the skill to cases where the satisfaction is not affected.

The system revises skills by adding preconditions, and Table 4 shows how the system computes the new preconditions for the two types of violations. C_r and C_s represent the relevance and satisfaction conditions. O_a and O_d are the add and delete lists of the executed primitive skill. The rationale for these computations has been developed in detail in prior publications (Ohlsson, 1996; Ohlsson & Rees, 1991).

As an example, let us revisit the Blocks World cases. In the first case, we have a red block, *A*, and a blue block, *B*. The system executes an instance of the first skill shown in Table 2, `(stacked A B)`, which adds the predicate to the state. This implies that the relevance condition of the color constraint, `(on A B)` also becomes true, but the satisfaction condition `(same-color A B)` does not. When detecting this type A violation, the system computes additional preconditions and attempts to make two revisions. The first calculation, $\neg(C_r -$

Table 4: New preconditions created in response to constraint violations.

Type \ Revision	1	2
A	$\neg(C_r - O_a)$	$(C_r - O_a) \cup (C_s - O_a)$
B	$\neg C_r$	$C_r \cup \neg(C_s \cap O_d)$

O_a), leads to `(on A B) − (stacked A B)`, which results in a null precondition. Therefore, the system ignores the first revision and tries the second one. This time, the additional precondition comes from $(C_r - O_a) \cup (C_s - O_a)$, which leads to `(same-color A B)`. The system adds this precondition to the skill that caused the violation, and restricts the execution of the stacking action to the case where two blocks have the same color.

In the second case, we have two blocks, *C*, and *TB*. When the system stacks the block C on top of the block TB using the skill `(stacked C TB)`, the top-block constraint becomes unsatisfied (`(clear TB)` not true in the state) while it stays relevant continuously (`(top-block TB)` true in the state). Upon detecting this type B violation, the system computes two sets of additional preconditions using the formulas, $\neg C_r$ and $C_r \cup \neg(C_s \cap O_d)$. These lead to `(not (top-block TB))` and `((top-block TB) (not (clear TB)))`, respectively, which are added to two separate revisions of the skill. The first revision prevents the use of the stacking action onto a block designated as a top-block. The second revision is a case of over-specialization, which makes it impossible to fire. Nevertheless, the two revisions achieve the proper restriction of the skill for the top-block constraint.

Challenges in Implementation

Although this implementation in the context of ICARUS shares the basic steps with previous systems using constraint-based specialization, various important differences between the ICARUS architecture and production system architectures require some significant changes in the revision process. In this section, we discuss the challenges and our solutions to them.

Hierarchical Organization

First of all, ICARUS' hierarchical organization of skill knowledge poses the most significant change, in relation to the assignment of blame. Production systems have flat structures, and it is mostly the case that the last executed rule caused a violation. But in ICARUS, execution involves a skill path, which may include more than one skill instance. Skill instances near the top of the path are more abstract, and those close to the bottom are more specific. Depending on the level of abstraction at which the violated constraint exists, the skill that needs to be revised can be anywhere on this path, and no simple attribution rule will be sufficient. So, the question is how ICARUS can identify the right skill to revise generally.

An analysis of multiple examples indicates that the architecture should find the highest level in the skill path in which all the variables involved in the additional preconditions for the revision are bound. All the additional preconditions are fully instantiated at this level and, therefore, it is the highest level in which the preconditions become meaningful. This makes it the right level at which to make the corresponding revisions. The results of running ICARUS indicate that this solution is correct. This solution is easily computable and general across domains. The possibility that it applies to other types of hierarchical systems might deserve attention.

Add and Delete Lists

Another problem occurs during the computation of the additional preconditions for skill revisions. Unlike production systems that have explicit and complete add and delete lists associated with actions, the ICARUS architecture has skills associated with goals. Goals typically do not include any side effects we do not care about, and they do not specify any predicates that should disappear after a successful execution. For this reason, the add and delete lists are not explicit in the architecture, and we must compute them from other sources.

Meanwhile, the use of add lists during the revision process is limited to the calculation of logical differences, and we can use goals as if they represent complete add lists. This will make the revised skill more restrictive rather than less so, thus making it safe. However, we should compute the delete list explicitly because of the way it is used during the revision process. We chose to calculate the list by comparing two successive belief states, although this may include some predicates removed by sources external to the agent. Again, however, this makes the revisions more restrictive, rather than more general, keeping the agent safe, because the delete list is negated during the computation of preconditions.

Disjunctive Definitions

ICARUS' support for multiple, disjunctive definitions of concepts adds another layer of complexity. When computing additional preconditions for skill revisions, the system should decompose any non-primitive concepts. Disjunctive concepts create multiple expansions, possibly resulting in more than one set of additional preconditions. The architecture accepts all such expansions and create multiple revisions.

The consequences of this approach are significant. When the system experiences a constraint violation, the situation might involve a particular disjunction of a concept. Nevertheless, the architecture learns multiple revisions from this case, covering all possible disjunctions of the concept. This approach is based on the understanding that there is a good reason why the disjunctive concepts have the same head, and that the system benefits from learning about all such cases. In future tasks, the system might confront a situation in which another one of the disjunctions applies, and, due to its prior learning, the system will already know how to avoid making an error in this situation even though it has never encountered it before.

Experimental Observations

To verify that the system works as intended, we performed experiments in two domains. We give only the basic concept and skill sets to the system at the beginning, along with the information on constraints. This means that the system knows how to operate in the world, but not at the level of expertise that enables it to satisfy the constraints at all times. It is as if humans sometimes know what should happen and what should not, but do not necessarily know how to impose these rules and often make mistakes. As the system learns from its failures, it revises the basic skills to avoid constraint violations in the future.

Blocks World

We modified the familiar Blocks World from the typical setup to include blocks of different colors and sizes. This modified domain supports various constraints like the color and size of the blocks in a tower or the maximum height of each tower. Table 3 shown earlier includes three of the constraints we have in this domain. The *color* constraint says whenever a block is stacked on top of another the two blocks should have the same color. This, in effect, forces any tower to have only the blocks of one color. The *top-block* constraint means any block designated as a top block (according to the system's conceptual knowledge) should always be clear, having no other blocks on top. The last constraint, *width* enforces that a block is smaller than the block underneath it.

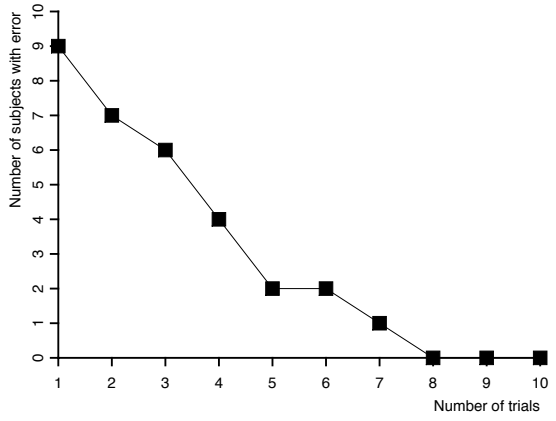


Figure 1: Number of simulated subjects that violated a constraint at each trial.

We ran simulation experiments with several different goals, and Figure 1 shows the result from one of them. In this experiment, we had ten simulated subjects, and each subject performed ten trials of the given task. We recorded the number of subjects that violated any constraints during each trial. The graph clearly shows that the revision process gradually reduces the number of the simulated subjects with constraint violations.

Route Generation

Another domain we used to test the system is a simplified version of route generation between places. Here, in addition to testing the specialization mechanism in ICARUS, we also want to verify that the mechanism can operate in parallel to other learning schemes like learning from success. The agent starts at a certain location, and has the goal of getting to a target location elsewhere. Using the information on connections between neighboring locations, the system performs problem solving to find a route to its target. As a result, it finds one of the several possible routes that involve different waypoints, and ICARUS learns specific route knowledge from this positive experience.

But some of the routes might become unavailable for travel due to various reasons like a broken bridge. At subsequent runs, the agent encounters situations where it can not use routes it learned before. While attempting to get to the target using a learned route, ICARUS recognizes that it gets stuck at a location with no outlet, violating a constraint not to be at a dead end. This failure triggers the system to learn a revised skill, which prevents it from moving to a location without any outlet. On the next trial, armed with this new skill, the system attempts to find another route to get to its target, and learns a skill for an alternate route for later use.

Let us see this behavior in a sample run. We give the system a goal to get to a target location, *B*, starting from the initial location, *A*. The two locations are connected by two alternate routes using waypoints *W1* and *W2*, respectively. The system starts out with two concepts and a skill as shown in Table 5. It also has the connection information between the locations, *A*, *B*, *W1*, and *W2* as some static beliefs. The only constraint it knows of is,

```
(at ?location) → (not-dead-end ?location)
```

which simply says that it should not be at a dead end at any time. During the first trial, the system finds a path, *A - W1 - B* through problem solving, and learns a specific skill for this route. Before we continue to the next trial, we intentionally remove the connection between *W1* and *B*, making the path obsolete. On the next trial, the system attempts to reuse the path, but it finds that it violates the constraint while it is at location *W1*. This violation triggers a revision process, resulting in another new skill. Once the system learns this new skill, it attempts to find an alternate route through yet another problem solving process, resulting in the path, *A - W2 - B*. After ICARUS stores this route as a specific skill, it executes the skill when it encounters the same task at a later time.

Related and Future Work

The current work on the constraint-based specialization has important similarities to some work in the explanation-based learning (EBL) literature (see Ellman, 1989; Wusteman, 1992 for reviews). EBL methods assume a significant amount of domain theories presumed to be perfect. However, in most of the domains, this is not true, and they require some ways

Table 5: Two concepts and a skill given to ICARUS for the route generation domain, and the two skills the system learned. The first skill is learned from problem solving (marked as P-S), and the other is learned from constraint-based specialization (marked as C-S). The additional precondition in this skill is shown in bold face.

```
   Given:      ((at ?location)
                :percepts  ((self ?self location ?location)))
               ((not-dead-end ?location)
                :percepts  ((location ?location))
                :relations ((connected ?location ?to1)
                            (connected ?location ?to2))
                :tests     ((not (equal ?to1 ?to2))))
               ((at ?location)
                :percepts  ((location ?from))
                :start     ((at ?from)
                            (connected ?from ?location))
                :actions   ((*move-to ?location)))
-------------------------------------------------------------
  Learned      ((at B)
from P-S 1:     :subgoals  ((at W1)
                            (at B)))
  Learned      ((at ?location)
 from C-S:       :percepts  ((location ?from))
                 :start     ((at ?from)
                             (connected ?from ?location)
                             (not-dead-end ?location))
                 :actions   ((*move-to ?location)))
  Learned      ((at B)
from P-S 2:     :subgoals  ((at W2)
                            (at B)))
```

to augment or correct the domain theories. There, researchers worked on the similar problems of blame assignment and theory revision, although the exact formulations were different. Unlike most of these work, our approach uses explicit descriptions of constraints, which the system uses to detect failures and revise existing theories accordingly.

With the successful implementation of the constraint-based specialization mechanism in ICARUS, we are able to study the important problem of interactions between two learning mechanisms. People learn in a variety of ways (Ohlsson, 2008) and human-level flexibility is the outcome of the interactions among multiple learning mechanisms. Currently, we have only a limited understanding of how learning mechanisms interact to produce flexible behavior. We intend to add additional mechanisms to ICARUS, including learning from examples or from analogies, and explore the conditions under which multiple mechanisms produce more flexible behavior than individual mechanisms.

Another key problem is how to interleave thinking (search in a mental, symbolic problem space) and action (search in an external, physical environment). The two types of processes differ in a variety of ways, most importantly in that a return to a previous state can be achieved by fiat in the internal search space, but has to be accomplished through physical action in the external environment. We intend to experiment with multiple schemes for controlling the interleaving in multiple task domains.

Conclusions

An intelligent agent cannot be limited to learning from positive experience. When task environments change, the extrapolation of prior experience to cover future situations inevitably leads to errors, mistakes and unacceptable outcomes. To exhibit human-level flexibility, an artificial agent needs learning mechanisms that specify how to change in the face of such negative outcomes. The constraint-based specialization mechanism provided this capability in a production system framework before, and we implemented it with the hierarchical skill representation in the ICARUS architecture successfully, after resolving multiple conceptual problems. We performed some test runs in the Blocks World and a navigation domain, and found the mechanism successfully removes failures after revisions. We also verified that the mechanism works well in parallel to another learning mechanism, allowing further study of human level flexibility in this direction.

Acknowledgments

This research was funded by Award # N0001-4-09-1025 from the Office of Naval Research (ONR) to the second author. No endorsement should be inferred.

References

Anderson, J. R. (1983). *The architecture of cognition*. Cambridge, MA: Harvard University Press.

Anderson, J. R. (1993). *Rules of the mind*. Hillsdale, NJ: Lawrence Erlbaum.

Ellman, T. (1989). Explanation-based learning: A survey of programs and perspectives. *ACM Computing Surveys*, *21*(2), 163–222.

Laird, J. E., Rosenbloom, P. S., & Newell, A. (1986). Chunking in soar: The anatomy of a general learning mechanism. *Machine Learning*, *1*, 11–46.

Langley, P. (1987). A general theory of discrimination learning. In D. Klahr, P. Langley, & R. Neches (Eds.), *Production system models of learning and development* (pp. 99–161). Cambridge, MA: MIT Press.

Langley, P., & Choi, D. (2006). Learning recursive control programs from problem solving. *Journal of Machine Learning Research*, *7*, 493–518.

Ohlsson, S. (1996). Learning from performance errors. *Psychological Review*, *103*, 241–262.

Ohlsson, S. (2008). Computational models of skill acquisition. In R. Sun (Ed.), *The cambridge handbook of computational psychology* (pp. 359–395). Cambridge, UK: Cambridge University Press.

Ohlsson, S., & Rees, E. (1991). Adaptive search through constraint violations. *Journal of Experimental & Theoretical Artificial Intelligence*, *3*, 33–42.

Wusteman, J. (1992). Explanation-based learning - a survey. *Artificial Intelligence Review*, *6*(3), 243–262.

Motor Affordances in Object Perception

Stephen J. Flusberg (sflus@stanford.edu)
Alexia Toskos Dils (atoskos@stanford.edu)
Lera Boroditsky (lera@stanford.edu)
Stanford University, Department of Psychology
Jordan Hall, 450 Serra Mall, Building 420, Stanford, CA 94305 USA

Abstract

Recently, researchers have suggested that when we see an object we automatically represent how that object affords action (Tucker & Ellis, 2001). However, the precise nature of this representation remains unclear: is it a specific motor plan or a more abstract response code? Furthermore, do action representations actually influence *what* we perceive? In Experiment 1, participants responded to an image of an object and then made a laterality judgment about an image of a hand. Hand identification was fastest when the hand corresponded to both the orientation and grasp type of the object, suggesting that affordances are represented as specific action plans. In Experiment 2, participants saw an image of a hand before interpreting an ambiguous object drawing. Responses were biased towards the interpretation that was congruent with the grasp type of the hand prime. Together, these results suggest that action representations play a critical role in object perception.

Keywords: object perception; motor affordances

Background

Traditional approaches to visual perception have assumed that the primary goal of the visual system is to construct a detailed internal picture of the external world based on a noisy retinal image (e.g. Marr, 1980). Recently, however, there has been a growing appreciation for the possibility that visual perception may be equally concerned with how we move around and act in our environment (Milner & Goodale, 1995). The idea that vision and action are intimately linked can be traced to the ecological psychology of James Gibson (1979), who argued that organisms see the world in terms of how it affords *action*. While Gibson eschewed the notion of mental representation, contemporary scholars have suggested that visual perception may be at least partially characterized by a mental representation of the *affordances* in the environment (Tucker & Ellis, 1998). For example, seeing the coffee mug on the desk before me might involve mentally representing how I could reach out and grasp it in order to drink from it. However, the precise nature of these affordance representations and how they relate to object perception remains unclear.

Several researchers have suggested that affordances are represented as action plans in the motor systems of the brain (e.g. Tucker & Ellis, 1998; Chao & Martin, 2000). Tucker and Ellis conducted a series of studies to test whether people automatically generate a motor representation in response to the visual presentation of an object, even when there is no intention to act on the object (e.g. Tucker & Ellis, 1998, 2001). In one experiment, participants had to make a left or right-handed button press to indicate whether an image of an object on the screen was upright or inverted. The objects were chosen to have clear right or left-handed affordance (e.g. a frying pan with a handle oriented to the left affords a left-handed grasp). Participants responded faster and made fewer errors when their responding hand was congruent with the (task-irrelevant) affordance of the object on the screen. Similar stimulus/response compatibility (SRC) effects have been obtained for other types of *micro-affordances*, such as *grasp type* (Tucker & Ellis, 2001) and *wrist orientation* (Tucker & Ellis, 1998).

Others have argued that the motor representations activated when we observe an object form an integral part of our perception of the object (e.g. Helbig, Graf, & Kiefer, 2006). For example, Helbig et al. (2006) presented participants with images of two objects in quick succession. Participants were more accurate in naming the second object when similar actions were required to make use of two objects (e.g. pliers and a nutcracker). These findings raise important questions regarding, (1) the level at and specificity with which affordance information is represented, and (2) the potential causal role this information plays in the process of object perception.

For instance, several researchers have argued that the SRC effects obtained by Tucker and Ellis actually reflect abstract response coding rather than specific motor plans (e.g. Anderson, Yamagishi, & Karavia, 2002). Indeed, even Tucker and Ellis (2001) suggested that these effects could not be explained by appealing exclusively to the neural systems responsible for the on-line control of actions because they were obtained using both images of objects as well as real-world objects that were out of reach of participants. Rather, affordance representations might be more abstract, specifying, for example, the general class of hand shape required to interact with an object rather than precise motor parameters. Further, while the results of Helbig et al. (2006) are consistent with motor affordance representations contributing to object perception, the data could also be explained by the fact that objects that are used in similar ways are typically similar to one another in other ways as well (e.g. semantically).

In this paper we describe two experiments designed to address these issues. First, we wanted to know whether motor affordances might be represented as specific action plans for interacting with an object rather than as abstract response codes. To this end, in Experiment 1 we made use of a dependent measure that is known to draw on very

specific manual action representations, namely the hand identification task (Parsons, 1987). Previous work has shown that the time it takes to identify whether an image is of a left or right hand is directly proportional to the time it would take, and how difficult it would be, to rotate your own hand into that position (Parsons, 1987). In our study, participants first made a response towards an image of an object that afforded a particular grasp type and wrist orientation. They then saw an image of a hand and had to indicate whether it was a right or left hand. If seeing an object leads to the activation of a specific manual action representation, participants should be faster to respond to an image of a hand that matches that object on grasp type and wrist orientation.

Second, we investigated the possibility that action representations actually contribute to the perceptual representation of objects. In Experiment 2, participants were first primed with an image of a hand depicting a specific grasp type. They then saw a drawing of an ambiguous object and had to indicate what they thought it was. The object could be interpreted as affording a power grasp (e.g. a football) or a precision grasp (e.g. a coffee bean). Responses were biased towards the interpretation that was congruent with the primed hand. We also included a control condition designed to rule out task demand and memory-based explanations of our findings. Together, these results suggest that action representations play a critical role in object perception.

Experiment 1

What motor information becomes activated when we look at an object? Previous research suggests that abstract response codes representing individual micro-affordances such as *grasp type* or *wrist orientation* become activated during object perception (Tucker & Ellis, 2001). Experiment 1 makes use of a novel application of the hand identification task in order to test the specificity of motor representations activated during object perception. Participants first made judgments on an image of an object that afforded a particular grasp type (power, precision, or no grasp affordance) at a particular wrist orientation (upright or inverted). Then they made laterality judgments on an image of a hand that was configured in a particular grasp type (power or precision) at a particular orientation (upright or inverted). To the extent that viewing objects activates specific manual motor plans selective for both grasp type and wrist orientation, laterality judgments should be fastest when *both* of the micro-affordances manipulated align between the images of the object and hand.

The degree to which various micro-affordances are activated during object perception might also depend on current goals (Bekkering & Neggers, 2002). Experiment 1 was designed to test whether viewing objects activates motor plans more strongly when participants make grasp-related compared with grasp-unrelated judgments about the objects. Similar reaction time profiles between these conditions would suggest that motor affordances are activated automatically during object perception, regardless of current task goals. Conversely, differences in reaction time profiles between the grasp-related and grasp-unrelated conditions would suggest that the current task goals do affect the kind of motor representation activated.

Methods

Participants Sixty-eight right-handed individuals from the Stanford community were recruited to participate in this study in exchange for payment or class credit.

Stimuli *Object Images*: Objects used in Experiment 1 varied on two dimensions: required *grasp type* (power, precision, or none) and required *wrist orientation* (upright or inverted). The dimensions were fully crossed within-subjects to produce 6 different object types. Two power grasp objects (flashlight and glass), two precision grasp objects (pushpin and tweezers), and four objects with no grasp affordance (desk, bookcase, grandfather clock, and sofa) populated the object categories. Hence, participants saw 16 unique images, each of which was repeated 32 times for a total of 512 object presentations.

The object stimuli were designed to afford right-handed responses because we recruited exclusively right-handed participants. The upright version of each object faced upward and to the left so as to afford an upright right-handed grasp on the part of the observer. The upright version of each object was rotated 90 degrees counter-clockwise in order to create the inverted version, which faced down and to the left. Pilot testing confirmed that right-handed individuals most often reached for real-world objects in both the upright and inverted orientations with their right hands.

Hand Images: The hand images varied on three dimensions: *grasp type* (power or precision), *wrist orientation* (upright or inverted), and *laterality* (left or right). The dimensions were fully crossed within-subjects to produce 8 different hand types. Four images of hands producing a power grasp and four images of hands producing a precision grasp were used to populate each of the hand categories. Hence, participants saw 32 unique hand images, each of which was repeated 16 times for a total of 512 hand presentations. The upright and inverted right and left hand images were generated using the same process described above for the upright and inverted object images.

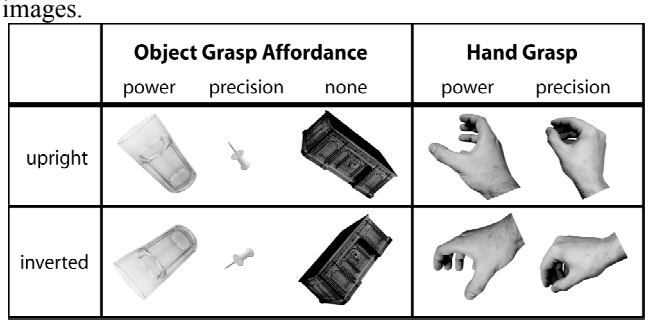


Figure 1: Sample stimuli from Experiment 1

Procedure Each trial in Experiment 1 had two parts. Participants first responded to a picture of an object and then to a picture of a hand. Participants were randomly assigned to one of two conditions: an *Orthogonal Judgment Condition* and a *Non-orthogonal judgment condition*. In the Non-orthogonal Condition, participants made a grasp-related judgment in response to each object ("Can you pick it up with one hand?"). In the Orthogonal Condition, participants made a grasp-unrelated judgment in response to each object ("Is it smaller than a shoebox?"). In both conditions, participants pressed the "j" key with their right index finger to enter a "yes" response, and the "f" key with their left index finger to enter a "no" response. Participants were told to respond as quickly and accurately as possible. Each object in Experiment 1 was preceded by a 500 ms fixation period. The image remained on the screen until the participant responded or the 10-second deadline expired, at which point the trial advanced to the hand portion.

Each hand image was preceded by a 500ms fixation period. Participants pressed the "j" key with their right index finger for pictures of right hands, and pressed "f" with their left index finger for pictures of left hands. The experiment advanced when the participant entered a response or at the end of the 10-second deadline. A black screen appeared for 750 ms to mark the end of each trial. The 16 unique object images were fully crossed with the 32 unique hand images to generate 512 unique experiment trials, each of which participants saw only once.

Results

The data from eight participants were removed because they did not contribute to all cells in the analysis or they had extremely high error rates or reaction times.

Trials analyzed: Only trials in which participants made correct responses to both the object and hand images were analyzed, resulting in the exclusion of 13.4% of trials. Any response times faster than 200 ms. or slower than 5000 ms. were omitted from all analyses, resulting in the removal of 1.4% of remaining trials across conditions. Finally, the stimuli used in Experiment 1 were designed to elicit right-hand affordances from right-handed individuals. As a result, only images of right hands were analyzed.

Coding: In 'Orientation Match' trials the orientation of the object was identical to that of the subsequent hand (collapsing across upright and inverted images). In 'Orientation Mismatch' trials the orientation of the object differed by 90 degrees in angular rotation from that of the subsequent hand.

RT Analyses: Figure 2 illustrates the mean pairwise RT differences (Orientation Match – Orientation Mismatch) across all levels of Object Affordance (power, precision, none) and Hand Stimulus (power, precision). Negative difference scores suggest a match advantage with respect to orientation. Positive difference score suggest a mismatch advantage with respect to orientation. The difference scores were submitted to a 3 (Object Affordance: power, precision, none) x 2 (Hand Stimulus: power, precision) repeated measures ANOVA. The analysis produced no main effects of object affordance ($F(2,57)=1.53$, *ns*) or hand stimulus ($F(1,58)=0.436$, *ns*), but a reliable quadratic interaction between the two variables ($F(1,58)=13.14$, $p< 0.001$).

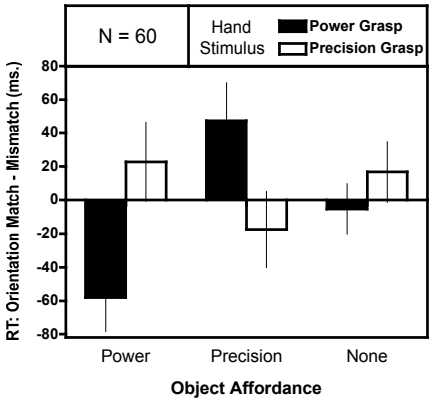


Figure 2: Differences in reaction time (Orientation Match – Mismatch) to the hand stimulus in Experiment 1. Error bars reflect the SE of the mean for each cell.

Participants showed a reliable match advantage to images of hands in a power grasp after having seen an object affording a power grasp (M=-58 ms, SD=158) compared to having seen an object with no manual action affordance (M=-5 ms, SD=116) (t(59)=-2.31, $p<0.05$). Conversely, participants showed a reliable mismatch advantage to images of hands in a power grasp after having seen an object affording a precision grasp (M=47 ms, SD=177) compared to having seen an object with no manual action affordance (M=-5 ms, SD=116) (t(59)=2.05, $p<0.05$). Images of hands in a precision grasp showed analogous trends, but none of the relevant comparisons reached significance (all $p>0.2$). This may be due to the fact that these hands were harder to correctly identify and thus they produced more errors and more variance in RT compared to grasp hands. The effect of orientation for hands in a precision grasp did, however, differ from the effect of orientation for hands in a power grasp both when the preceding object required a power grasp (M=22.80, SD= 23.51) (t(59)=-2.58, $p< .05$) and when the preceding object required a precision grasp (M=-17.54, SD=22.63) (t(59)= 2.05, $p< .05$).

The data appear to follow a Mexican hat distribution (Muller et al., 2005), such that reaction times *increase* when the object affordance only somewhat overlaps with the hand stimulus and *decrease* when the two overlap entirely (relative to trials where the preceding object had no grasp affordance). To further test for such a distribution, trials were binned into five similarity-based categories (Figure 3). For each of the bins, the object affordance on a given trial relative to the subsequent hand stimulus was either: (1) same orientation and grasp type, (2) same grasp type only, (3) same orientation only, (4) different orientation and grasp type, or (5) had no grasp affordance. A repeated measures

ANOVA yielded reliable quadratic ($F(1,58)=8.04$, $p<0.01$) and cubic ($F(1,58)=6.05$, $p<0.02$) effects of similarity.

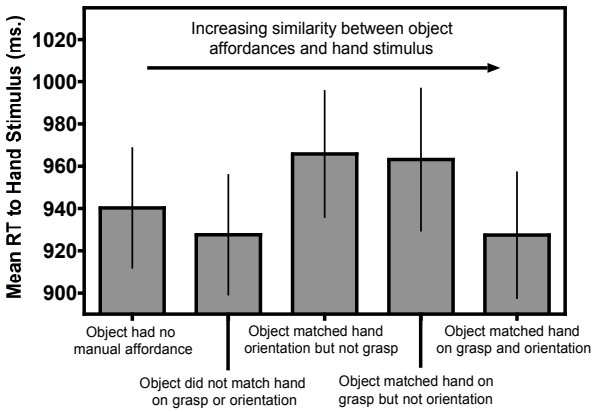


Figure 3: RT to hand stimuli in Experiment 1, binned by how similar the hand was to the set of manual affordances in the previous object

Finally, there was no main effect of the between-subjects condition (Task Goals: grasp-related, grasp-unrelated) ($F(1,58)=1.50$, $p>0.20$), nor did condition interact with the quadratic Object Affordance by Hand Stimulus interaction ($F(1,58)=2.81$, $p>0.10$) or the effect of similarity ($F(1,58)=1.76$, $p>0.15$). As a result, all analyses described above were run on the combined data from the two conditions.

Discussion

In this experiment we asked whether people generate a specific motor plan for interacting with an object when they see that object. RTs for the hand laterality judgment were fastest when the hand corresponded in both grasp type and wrist orientation to the previous object. This suggests that when we look at an object we represent the specific motor parameters necessary for interacting with that object. In other words, when we see a drinking glass we actually simulate reaching out and grasping it.

Interestingly, RTs for the laterality judgment were slowest when the object and hand corresponded in just one micro-affordance dimension (i.e. either grasp type *or* wrist orientation). Researchers have argued that similar reaction time profiles in classic vision and attention tasks suggest an underlying surround inhibition mechanism (Muller et al., 2005; Roeber, Wong, & Freeman, 2008), where activating a particular representation suppresses highly similar but not distantly similar representations. Importantly, motor cortex is believed to have the kind of connectivity that would support surround inhibition (Lukashin & Georgopoulos, 1993; Sohn & Hallett, 2004). Thus it is plausible that viewing an object in the present study activates a highly specific action representation, which in turn spreads inhibition to highly similar but not distantly similar action representations. These patterns of activation and inhibition would result in slower RTs to trials in which the images of the object and hand differ on one but not all dimensions, which is precisely what was found in Experiment 1.

Such connectivity further predicts a full Mexican hat-like distribution of response times such that the representations in similarity space just beyond the inhibited surround should see facilitation that tapers off as distance increases (Muller et al., 2005). That is, responses to hands preceded by objects that afford the wrong grasp in all dimensions should be faster than those preceded by objects that afford no grasp at all. The cubic effect of similarity found in Experiment 1 is driven by that very difference, suggesting a Mexican hat response time profile (Muller et al., 2005). Studies better designed to test for such a pattern are currently underway. As it stands, the pattern observed in Experiment 1 is consistent with the kind of connectivity believed to exist in motor cortex. Furthermore, the hand identification task used in this study was selected precisely because it is believed to be supported by specific motor regions. As a result, the present findings are consistent with the hypothesis that object perception activates highly specific action representations in the motor system and does so in a manner similar to the act of grasping itself.

Finally, varying participants' task goals when viewing the object had no significant effect on these results. Whether participants made a grasp-related ("Can you pick it up?") or grasp-unrelated ("Is it smaller than a shoebox?") judgment towards the object, the same affordance information appears to have been represented. This supports the original findings of Tucker and Ellis (1998), who argued that affordance information is represented irrespective of the intentions of the observer. However, other researchers have found effects of intentions on affordance representation (e.g. Bekkering & Neggers, 2002), and the effects in the present study tended to be more robust in the task-related than the task-unrelated condition, which suggests that more research is called for on this issue.

Experiment 1 supports the idea that motor affordances are represented as specific action plans in the motor system regardless of task goals. However, it is unclear how this action representation relates to our ability to actually perceive the object. We turn to this issue in Experiment 2.

Experiment 2

Does action representation contribute to object perception? One possibility is that the visual and motor aspects of object perception are fairly independent: extracting the visual features of an object occurs in one processing stream while extracting the affordance information relevant for action occurs in a different processing stream (Milner & Goodale, 1995). Alternatively, visual and motor processes may be more interdependent, and currently activated action representations might play a causal role in visual object processing. Experiment 2 was designed to test the latter possibility by priming participants with a specific manual action to see if it would affect their interpretation of an ambiguous object drawing. We also ran a control condition where we presented the ambiguous object image first in order to control for task demand or memory-based explanations for the data.

Methods

Participants 245 individuals from Amazon's Mechanical Turk website participated in exchange for payment.

Stimuli & Procedure The stimuli for this experiment included four photographs of hands taken from Experiment 1 and an ambiguous object line drawing created by the authors. The four hand photographs showed either left or right hands in either a power or precision grasp. Pilot testing suggested that the ambiguous object drawing could be interpreted as an object that afforded a power grasp (e.g. *football*) or as an object that afforded a precision grasp (e.g. *coffee bean*).

In the *experimental condition*, one of the four hand images was randomly selected for each participant and displayed on the screen for three seconds. After this, the ambiguous object drawing was displayed for three seconds. Then, participants were asked to name the object in the line drawing that they had just seen and to identify whether the hand they had seen was a left or right hand. The only difference in the procedure for the *control condition* was that participants were shown the ambiguous object image first and hand image second. Thus participants were not primed with an action representation prior to viewing the ambiguous object, but they saw the same two images prior to making their object interpretation response.

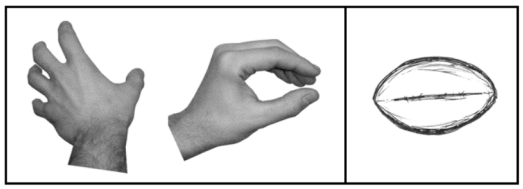


Figure 4: On the left, images of the left hand stimuli used in the experiment displaying precision and power grasps. On the right, the ambiguous object drawing.

Results

The data from 19 participants were removed because they failed to respond to the test questions appropriately (N=8) or because they took the survey more than once (N=11). We then coded the object interpretation responses for the remaining participants in terms of what sort of grasp would be afforded by the perceived object. We used the following coding scheme: Power (e.g. *football* or *coconut*), Precision (e.g. *coffee bean* or *nut*), and None (e.g. *lips* or any response that listed more than one interpretation).

Experimental Condition: For our analyses we collapsed across left and right hand prime stimuli and excluded object interpretation responses coded as *none*. A 2 (hand prime stimulus: power grasp vs. precision grasp) X 2 (perceived object affordance: power vs. precision) chi-square test of independence showed a significant relationship between hand prime stimulus and perceived object affordance, $\chi^2 = 7.04$, $p<0.01$. Participants primed with an image of a power grasp hand were more likely to interpret the ambiguous image as an object affording a power grasp while participants primed with a precision grasp hand were more likely to interpret the ambiguous image as an object affording a precision grasp.

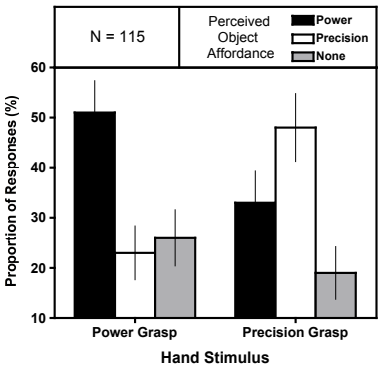


Figure 5: Experiment 2, *Experimental Condition*: Proportion of ambiguous object interpretations coded for perceived object affordance. Error bars are the SE of the proportion.

Control Condition: A 2 (hand stimulus: power grasp vs. precision grasp) X 2 (perceived object affordance: power vs. precision) chi-square test of independence showed no relationship between hand stimulus and perceived object affordance, $\chi^2 = 0.74$, $p>0.38$.

Interaction Analysis: A 2 (interpretation: congruent vs. incongruent) X 2 (condition: experimental vs. control) interaction analysis showed that hand stimuli only affected ambiguous object interpretations in the experimental condition, $\chi^2 = 4.05$, $p<0.05$.

Discussion

In this experiment we asked whether currently active motor representations would influence what participants saw when they looked at an ambiguous object. We found that when participants were primed with a hand displaying a power grasp they were more likely to interpret an ambiguous image as an object that afforded a power grasp (e.g. football). Conversely, when they were primed with a hand displaying a precision grasp, they were more likely to interpret the image as an object that afforded a precision grasp (e.g. coffee bean). This finding suggests that action representations can play a causal role in the process of object perception.

That said, there are a number of possible alternative explanations for these data. First, because of the simple design of this experiment, participants may have simply figured out what we wanted from them and tried to give it to us. The results of the control condition suggest that this is unlikely, however. In that condition, participants also saw both the ambiguous object and the hand stimulus prior to giving their object interpretation response, the only difference being they saw the ambiguous object first. If the results from the experimental condition were due to demand characteristics, we would expect to find the same pattern of results here. However, in the control condition there were

no such effects. This also helps rule out the possibility that the results of the experimental condition were due to associations in memory rather than the online effects of action representation on perception.

Finally, because we used purely visual stimuli in this experiment, it is possible that our results reflect visual priming rather than motor priming. While prior research has demonstrated that visually processing images of hands typically involves activating motor representations of one's own hand (Parsons, 1987), it is difficult to rule out visual priming as an explanatory mechanism at the present time. Research currently underway in our lab is moving away from visual prime images and towards actual motor movements as priming stimuli. Moreover, we are developing additional controls that include reversible images that do not afford grasping in order to rule out alternative mechanisms such as altered scanpaths or attentional patterns.

General Discussion

In this paper we explored the role that action representation plays in visual object processing. In Experiment 1 we took a bottom-up approach, asking whether we generate a specific motor plan or a more abstract response code when we observe an object with a particular set of manual affordances. Participants made a judgment about an image of an object that afforded a particular grasp type and wrist orientation. They then made a laterality judgment about an image of a hand displaying a particular grasp type and wrist orientation. RTs for the laterality judgment were fastest when the hand corresponded in both grasp type and wrist orientation with the previous object. This suggests that when we look at an object we represent the specific motor parameters necessary for interacting with that object within the motor systems of the brain.

Intriguingly, RTs for the laterality judgment were slowest when the object and hand corresponded in just one micro-affordance dimension (i.e. either grasp type *or* wrist orientation). This "Mexican hat" response time function has been found by other researchers studying motor representation in the brain (Loach, Frischen, Bruce, & Tsotsos, 2008; Lukashin & Georgopoulos, 1993; Sohn & Hallett, 2004), providing further support for the idea that affordances are represented as specific action plans in the motor system.

In Experiment 2 we took a top-down approach, asking whether activating a particular manual action representation would influence the perception of an ambiguous object image. The results suggest that action representations can play a causal role in the process of object perception.

All together, the results of these experiments suggest that action representation plays a crucial role in visual object processing. As we look around the world we are not merely constructing an internal picture of what's out there, we are also preparing to act and behave on what's before us. Furthermore, our current action state affects how we process the *what* that is out there.

Acknowledgments

The authors would like to thank the members of Cognation for their helpful feedback. This research was supported by an NSF Career Award Grant given to Lera Boroditsky.

References

Anderson, S. J., Yamagishi, N., & Karavia, V. (2002). Attentional processes link perception and action. *Proceedings of the Royal Society of London B, 269,* 1225-1232.

Bekkering, H. & Neggers, F. W. (2002). Visual search is modulated by action intentions. *Psychological Science, 13,* 370-374.

Chao, L. L. & Martin, A. (2000) Representation of manipulable man-made objects in the dorsal stream. *Neuroimage, 12,* 478-484.

Gibson, J. J. (1979). *The ecological approach to visual perception*. Lawrence Earlbaum: Hillsdale, NJ.

Helbig, H. B., Graf, M., & Kiefer, M. (2006). The role of action representations in visual object recognition, *Experimental Brain Research, 107*(2), 221-228.

Loach, D., Frischen, A., Bruce, N., & Tsotsos, J. (2008). An attentional mechanism for selecting appropriate actions afforded by graspable objects. *Psychological Science, 19*, 1253-1257.

Lukashin, A. V., & Georgopoulos, A. P. (1993). A dynamical neural network model for motor cortical activity during movement: population coding of movement trajectories. *Biological Cybernetics, 69*, 517 – 524.

Marr, D. (1982). *Vision*. San Francisco, CA: Freeman.

Milner, D. & Goodale, M. (1995). *The Visual Brain in Action*. Oxford: Oxford University Press.

Muller, N. G., Mollenhauer, M., Rosler, A., & Kleinschmidt, A. (2005). The attentional field has a Mexican hat distribution. *Vision Research, 45*, 1129-1137.

Parsons, L. M. (1987). Imagined spatial transformation of one's body. *Journal of Experimental Psychology: General, 19,* 178-241.

Roeber, U., Wong, E. M., & Freeman, A. (2008). Cross-orientation interactions in human vision. *Journal of Vision, 8*, 1-11.

Tucker, M. & Ellis, R. (1998). On the relations between seen objects and components of potential actions. *Journal of Experimental Psychology: Human Perception and Performance. 24*(3), 830-846.

Tucker, M. & Ellis, R. (2001). The potentiation of grasp types during visual object categorization. *Visual Cognition, 8*(6), 769-800.

Order Effects in Moral Judgment

Alex Wiegmann (Alex.Wiegmann@psych.uni-goettingen.de)
Jonas Nagel (jnagel1@uni-goettingen.com)
Stefan Mangold (smangol@gwdg.de)
Department of Psychology,
University of Göttingen, Germany

Yasmina Okan(yokan@ugr.es)
Department of Experimental Psychology, University of Granada
Campus Universitario de la Cartuja s/n ,18071, Granada, Spain

Abstract

Explaining moral intuitions is one of the hot topics of recent cognitive sciences. In the present article we focus on a factor that attracted surprisingly little attention so far, namely the temporal order in which moral scenarios are presented. We argue that previous research points to a systematic pattern of order effects that has been overlooked until now: Only judgments of actions that are normally regarded as morally acceptable are affected by the order of presentation. Additionally, this is only the case for dilemmas immediately preceded by a scenario where the proposed action was judged as morally unacceptable. We conducted an experiment that confirmed this pattern and allowed us to analyze the individual level responses it was generated by. We argue that investigating order effects is necessary for approaching a complete descriptive moral theory.

Keywords: moral intuitions; trolley dilemmas; order effects

Introduction

In the past decades, trolley dilemmas have been used extensively for testing philosophical and psychological theories of moral judgments. In the standard description of the trolley dilemma introduced by Philippa Foot (1967), an out-of-control train threatens to kill five people standing on its track. The only way to prevent this is to pull a switch that redirects the train onto a different track where it will kill only one person. In a modification of this scenario (Thomson, 1976), the only possibility to prevent the five people from being killed is to push a heavy person from a footbridge onto the track. This would stop the train but kill the heavy person. Numerous studies (e.g., Hauser, Cushman, Young, Jin, & Mikhail, 2007) have shown that given the same number of people being killed vs. saved, participants approve of acting in the first but not in the second scenario. Several competing descriptive theories explicate psychological principles supposed to underlie this pattern of moral intuitions (e.g. Greene et al., 2001; Hauser, 2006). However, surprisingly little is known about potential effects of the order in which several consecutive scenarios are presented. It is plausible to assume that consecutive scenarios will not be judged independently of each other: A principle or mechanism that is activated when a particular scenario is represented or evaluated might later be applied to a series of subsequent scenarios. However, only few studies have dealt with this issue so far (Petrinovich & O'Neill, 1996; Lanteri, Chelini, & Rizzello, 2008). Their results suggest that under certain circumstances moral judgments can indeed be transferred from one situation to another.

If such order effects could be replicated systematically, this would have important implications for psychological theories aiming to explain patterns of moral reasoning at a descriptive level. Furthermore, relevant practical implications would arise both for methodological considerations inside the research laboratory (in terms of controlling for order effects when designing experiments) and for everyday judgments outside the lab.

The present work has three main goals. First, we will provide the first comprehensive review of previous empirical research on order effects in moral judgments, and we will demonstrate that a systematic pattern of results has been overlooked so far. Second, we will empirically test the existence, extent and direction of order effects in reasoning about moral dilemmas. Finally, we will discuss the theoretical and practical implications of our findings, focusing on psychological theories of moral reasoning.

Order Effects in Previous Research

Speaking of order effects in moral judgment, there are at least two possible interpretations that could be labelled "within-scenario order effect" and "between-scenarios order effect", respectively. The first type of effect results if the order in which information concerning one particular situation is presented affects judgment. If, for example, the task is to judge the permissibility of an action, and the results solely differ as a function of the particular sequence in which positive and negative consequences are presented, a "within-scenario order effect" occurs. Second, a judgment regarding an action in a particular scenario might be influenced by a judgment that had previously been made about a different scenario. To illustrate, consider two conditions in which a given scenario C is preceded by one of two different scenarios (A vs. B). Differences in judgments of the action scenario C between the two conditions would – all other things being equal – instantiate a "between-scenarios order effect". The present research will focus on this second category of order effects.

One of the few studies addressing "between-scenarios order effects" in moral reasoning was conducted by Petrinovich and O'Neill (1996). Their aim was to analyze whether the presentation order of a set of moral dilemmas would affect participants' level of agreement or disagreement with the action proposed in each case. In one condition (standard order), the dilemmas were arranged according to decreasing predicted agreement with the potential action, whereas in the second condition (reversed order) the presentation order was reversed. While Petrinovich and O'Neill (1996) did not report any order effects in an experiment comparing three dilemmas that differed with regards to content (Study 2, Forms 1 and 1R), a reanalysis of their data revealed an order effect for the dilemma with the highest predicted agreeability. In particular, the average agreement rating in this scenario was significantly higher if the scenario had been presented first than if it had been preceded by the other two dilemmas (t_{57}=2.11; p<.05, two-tailed). In contrast, the other two dilemmas received almost equally low ratings in both order conditions. A reanalysis of a similar experiment using a different set of dilemmas (Study 2, Forms 3 and 3R) also revealed that the average rating for one of the positively rated dilemmas varied between the two order conditions. The average rating was lower if the scenario had been directly preceded by a dilemma that received lower (as opposed to higher) ratings (t_{68}=2.88; p<.01, two-tailed).

Another experiment reported by Petrinovich and O'Neill (1996; Study 2, Forms 2 and 2R) compared three different versions of the trolley dilemma. As in the previously reported experiments, a reanalysis revealed order effects for the two scenarios with the highest predicted agreeability (t_{57}=2.93; p<.01, two-tailed, and t_{57}=2.58; p<.05, two-tailed, respectively). However, the third scenario that involved pushing a person from a footbridge in order to stop the train (cf. Table 1) was not affected by the order of presentation.

Similarly, Lanteri, Chelini, and Rizzello (2008) reported order effects for the standard trolley dilemma, but not for the footbridge scenario. In addition, similar order effects were found incidentally in some studies. For example, Nichols and Mallon (2006) found that acting in a case equivalent to standard trolley was marginally more likely to be judged as breaking a rule if the scenario was preceded by a footbridge-equivalent case than if presented in the first position. No analogous effects of a preceding standard trolley-equivalent case on judgments in the footbridge-like case were reported. Recently, Lombrozo (2009) incidentally found results analogous to those obtained by Lanteri et al. (2008). Finally, Alistair Norcross (2008) described an interesting order effect outside of an experimental setting that is nevertheless relevant for the present research. He points out that when he asked his students to evaluate the standard switch-trolley dilemma in the first position, the majority judged that diverting the trolley is permissible. However when this dilemma was preceded by a scenario in which saving the lives of five patients requires to kill a healthy person in order to transplant his organs, the proportion of students judging that diverting the trolley is permissible was considerably lowered.

A Systematic Pattern

We claim that a closer look at the findings reported reveals a systematic pattern: First, all dilemmas that are affected by an order effect were rated positively (in the sense that the proposed action is on average rated as morally right/acceptable). Dilemmas that received a negative rating seem to be unaffected. Second, the dilemmas that were rated positively are only affected if they are directly preceded by a dilemma that was rated negatively. In this case, the ratings were lower or, in those cases in which the response format is dichotomous, the proportion of people that judge the action to be acceptable decreased.

Previous attempts to account for between-scenario order effects failed to fully capture the pattern we are suggesting here. For instance, Petrinovich and O'Neill (1996) argue that the initial strength of the response (agreement vs. disagreement) influences subsequent responses. If this were true, dilemmas that are normally rated negatively should also be affected by the order of presentation. However, this does not seem to be the case, since these dilemmas seem to be rated equally negative in all cases. Lanteri et al. (2008) take this asymmetry into account when explaining their results. However, they focus on properties of specific scenario contents instead of formulating a general pattern.

It is important to note that so far there is no evidence for a major change of people's judgments at a qualitative level. In Petrinovich and O' Neill's study (1996), the ratings for the proposed action do not seem to change enough to be regarded as acceptable in one order condition but as inacceptable in the other. In Lanteri et al. (2008), the percentage of people judging the proposed action as acceptable is indeed lowered, but it still remains above 50%.

Taking into account all the previous points, the main goal of our work will be to empirically test the existence of the pattern described above. If an order effect is present we will aim to determine its strength and, in particular, whether it can be strong enough to lead people to disagree with a proposed action that they would normally (i.e., when evaluated independently) agree with. We will use several variations of the trolley dilemma due to the existence of a large body of previous research establishing how the modification of different factors in these dilemmas affects how they are judged.

Experiment

Subjects

Fifty participants (35 women) were recruited using the lab in the psychology department at the University of Göttingen. They were randomly distributed to the different experimental conditions. The average age was 23 (*SD*=2.83).

Materials

We presented participants a series of five moral dilemma scenarios (the standard switch trolley and four modifications; see Table 1). Each scenario included a brief description of a situation and an action that could potentially be carried out in each case, accompanied by a diagram depicting the situation schematically. The initial description of the situational set-up was identical for all scenarios: An out-of-control trolley rapidly approaches three railroad workers who will die if Karl, the only bystander in the scenario, does not intervene.

Table 1: Summaries of the actions proposed in the five dilemmas

Scenario	Proposed action
Push	Push the large person from the bridge in order to stop the train
Trap	Push a button that will open a trap door in order to let the person on top of the bridge fall onto the track and stop the train
Redirect	Redirect a train containing one person that is on a safe parallel track onto the main track in order to stop the train
Run Over	Redirect an empty train that is on a safe parallel track onto the main track in order to stop the train thereby running over a person that is on the connecting track
Standard	Press a switch that will redirect the out-of-control train onto a parallel track where it will run over one person

This introduction was followed by a description of a specific action that Karl could conduct in order to save the three workers. This action was different for each of the five scenarios, but in all cases it resulted in the death of one innocent person (see Table 1). Instructions were included to ensure that participants assumed that the proposed action was the only available option in each case that, if carried out, would always lead to the described outcome. The number of potential victims (3 vs. 1) was kept constant across scenarios.

In order to establish a baseline of agreement with the action proposed in each of the five different scenarios we conducted a pilot study using a different sample consisting of 100 University of Göttingen students. Participants were individually approached on campus and asked to indicate for one of the scenarios (*n*=20) whether Karl should act in the proposed way or not on a scale from 1 to 6, where 1 was "not at all" and 6 was "absolutely". Table 2 shows the average ratings for the different scenarios.

Table 2: Mean ratings (standard deviations) of agreement and percentage of subjects disagreeing with the proposed action in the five scenarios when evaluated independently.

	Scenario (each *n*=20)				
Measure	Push	Trap	Redirect	Run Over	Standard
Mean Rating	1.95	3.4	4.15	4.4	4.45
(SD)	(1.76)	(1.76)	(1.42)	(1.14)	(1.15)
% Disagreement	80	40	30	10	15

Note. % Disagreement is the percentage of subjects who gave a rating <3,5 on a scale ranging from 1 to 6.

Based on these results we ordered the five scenarios according to level of agreement with the proposed action (i.e., Push<Trap<Redirect<Run Over<Standard). From here onwards we will refer to this ordering as the *level of agreeability* of the scenario, as defined by the extent to which participants agree with the action when the dilemmas are judged independently.

Procedure

The experiments were run individually on computers. Initially, the instructions were presented on the screen, followed by the five different scenarios. After each scenario, participants were requested to rate, on a scale from 1 to 6, whether Karl should act in the proposed way or not, where 1 was "not at all" and 6 was "absolutely". Half of the participants saw the sequence of dilemmas in increasing order of agreeability (Least Agreeable First [LAF] condition, beginning with Push), whereas the other half saw the sequence of dilemmas in the reverse order (Most Agreeable First [MAF] condition, beginning with Standard). The computerized format of the task guaranteed that each dilemma was judged before the following one was presented. Furthermore, there was no possibility for participants to withhold their judgment until the end of the sequence or to switch back in order to change a previously given rating.

Results

To test whether the pattern of ratings of the dilemmas differed in the two orders of presentation, a 2*5 mixed analysis of variance (ANOVA) was conducted, where the first factor was the order of presentation (LAF vs. MAF, between-subjects) and the second factor was the scenario judged (within-subjects). The results are shown in Table 3 and Figure 1. They revealed a main effect for order of presentation. Specifically, average ratings were significantly lower in LAF compared to MAF ($F[_{1,48}]$=8.03; p<0.01). Furthermore, we found a main effect for scenario

($F_{4,192}$=23.44; p<0.001), confirming our expectation of different average agreeability ratings for the scenarios. Crucially, the interaction between order of presentation and scenario was significant ($F_{4,192}$=8,2; p<0.001), suggesting the presence of a strong asymmetric order effect, in line with our predictions.

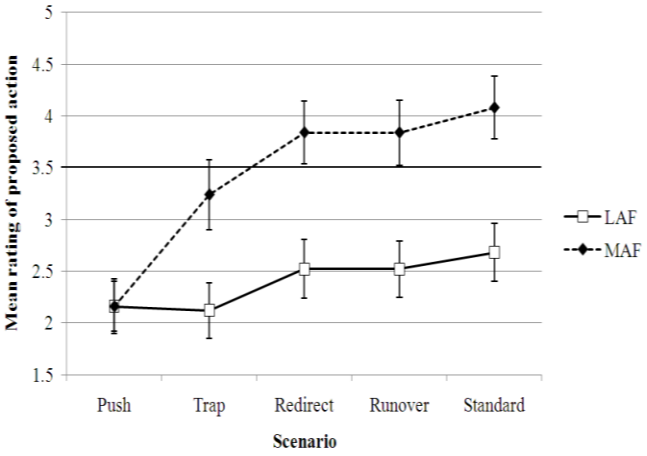


Figure 1: Mean ratings of agreement with the proposed action in the five scenarios when evaluated sequentially, as a function of the order of presentation. Error bars indicate SEM. The bold line at y=3.5 indicates the division between average agreement and disagreement. MAF = Most Agreeable First; LAF = Least Agreeable First.

Table 3: Mean ratings (standard deviations) of agreement and percentage of subjects disagreeing with the proposed action in the five scenarios evaluated sequentially, as a function of the order of presentation.

Order Condition	Scenario				
	Push	Trap	Redirect	Run Over	Standard
	Mean ratings (SD)				
MAF (n=25)	2.16 (1.21)	3.24 (1.69)	3.84 (1.52)	3.84 (1.57)	4.08 (1.53)
LAF (n=25)	2.16 (1.31)	2.12 (1.33)	2.52 (1.42)	2.52 (1.36)	2.68 (1.41)
	% Disagreement				
MAF (n=25)	76	52	40	32	32
LAF (n=25)	80	80	72	72	68

Note. % Disagreement is the percentage of subjects who gave a rating <3,5 on a scale ranging from 1 to 6. MAF = Most Agreeable First. LAF = Least Agreeable First.

In order to test our prediction more specifically, we conducted planned comparisons involving Standard and Push as examples of scenarios typically eliciting high and low agreeability ratings, respectively. The average rating for Standard varied considerably depending on the position in which it appeared. When it had been evaluated first, the average rating was 4.08, while the average was only 2.68 when it appeared at the end of the sequence. This difference was significant ($F_{1,48}$=11,39, p<0.01). In contrast, the average rating for the Push scenario was the same in both orders (2.16). Moreover, after computing the within-subjects differences between the ratings for the Standard and the Push scenarios, it can be shown that the average difference is significantly larger in MAF than in LAF ($F_{1,48}$=14,69; p<0.001), a result that further supports our prediction of an asymmetrical order effect.

It is worth noting that the difference between the ratings for the Standard scenario in the two order conditions is relevant not only in quantitative but also in qualitative terms: Treating ratings below 3.5 as disagreement and above 3.5 as agreement with the action proposed in a particular scenario, the majority of participants' ratings in LAF would fall into the first category (18 out of 25; 72%) whereas the majority of participants' ratings in MAF would fall into the second (18 out of 25; 72%). This difference is significant (χ^2_1=9.68; p<0.01). The same is true for Run Over (χ^2_1=8.01; p<0.01), Redirect (χ^2_1=5.20; p<0.05), and Trap (χ^2_1=4.37; p<0.05), but not for Push (χ^2_1=0.12; p=0.73).

Discussion

In sum, the data were largely in line with the pattern we discovered in previous studies: The judgments of actions that received a positive rating when inquired independently (Standard, Run Over, Redirect) differed significantly in the two order conditions. In contrast, ratings for the action in Push, which was rated negatively when judged independently, did not differ in the two conditions. Furthermore, in the MAF condition, the pattern of the average ratings was very similar to the one obtained when the scenarios were judged independently. In contrast, the average ratings in the LAF condition differed widely from those independent ratings.

It should be noted, however, that the results obtained for one of the scenarios cannot be directly derived from the aforementioned pattern. In particular, Trap was also affected by the order of presentation (both in quantitative and in qualitative terms), even though the proposed action in this scenario was rated slightly negative when judged independently. This finding motivated us to have a closer look at the results at the level of individual participants. In particular, we explored the data treating the ratings as a set of binary choices made by each participant (i.e., treating ratings < 3,5 as indication of disagreement and ratings > 3,5 4 as indication of agreement with the proposed action) and observed the tendency that a disagreement with an action was "transferred" to the judgment of the action in the next scenario. That is, an action receiving a positive rating when judged independently received lower ratings when presented as part of a sequence if the preceding scenario was rated negatively by the same participant. In contrast, positive ratings did not affect the ratings of the next action (i.e. changing them into positive ones) if this action was rated negatively in independent ratings. For instance, in the LAF

condition, only three out of 20 participants who disagreed with the proposed action in the initial Push-scenario changed their rating towards agreement during the whole sequence, resulting in 17 votes against the proposed action in the final Standard scenario. In contrast, when participants started with Standard, eleven out of the 17 participants who voted in favor of the proposed action changed their ratings towards disagreement on the way to Push, resulting in only six positive ratings for the proposed action in this final scenario of the sequence. Reformulating the pattern this way allows order effects to occur not only for actions rated positively when judged independently but also for actions rated negatively on average provided that the number of participants who would disagree with the proposed action in a particular scenario is sufficiently higher than the number of participants who would disagree with the action in the subsequent scenario. Within a sequence of scenarios this excess of "disagreements" can be transferred to the next scenario and cause an order effect. On the flipside, an order effect might also occur when the action to be judged in a particular dilemma is preceded by a dilemma where the proposed action is judged positively. Again, it just has to be the case that the number of disagreements in the preceding scenario would be sufficiently higher than in the following scenario if both scenarios were rated independently.

A similar distribution of nominal data could well underlie the results obtained by Petrinovich & O'Neill (1996). Unfortunately, we cannot conclusively confirm this claim because only aggregated results are reported.

It is not possible to determine from our data why the reported asymmetry occurs. However, a possible explanation is the existence of a difference in the urge to justify prohibitions and permissions. When we, e.g., prohibit a child to play with knives we automatically think of – or already have in mind – a justification for this prohibition. Prohibitions seem to call for a justification. In contrast, we do not think about a justification regarding most things we permit. We do not feel an urge to explain or justify to someone why, e.g., he or she is allowed to walk around. Normally, we only justify or explain permissions when a prohibition is the default case. For instance, we might explain to a child that in the case of an emergency an ambulance is permitted to drive over red lights although it is usually prohibited. Applying this line of reasoning to the asymmetric pattern found in our data it might be the case that because participants prohibited the proposed action in Push they were - consciously or unconsciously - thinking about a justification for their prohibition. If they reach a rough justification like "You must not kill an innocent person", they might keep this principle in mind and apply it to the remaining scenarios. Since an innocent person has to be killed in all scenarios in order to rescue three persons, participants might judge all proposed actions as prohibited. In contrast, when they start with a scenario where they judge the proposed action as permissible it might be the case that no effort is invested in justifying this judgment and, therefore, no such justification is applied to the remaining scenarios.

Implications for Descriptive Moral Theories

An important goal of descriptive moral theories is to provide an explanation of an average person's moral judgments that is as comprehensive as possible. A potential source of variance in moral judgments which has received comparably much attention is the structural set-up of the situations in question (e.g., whether the victim serves as means or side-effect in saving the three workers; see Cushman, Young, & Hauser, 2006). However, the effects generated by the manipulation of these factors are usually fairly small, i.e. they account only for a very limited amount of the total variance in moral judgments and thus leave a good portion of between-subject differences unexplained. Thus, considering only factors concerning the objective situational set-up is by no means sufficient to generate a comprehensive descriptive moral theory. Rather, it seems to be necessary to take into account additional psychological mechanisms that influence how a given situational set-up is apprehended, represented, and evaluated. In our experiment, for example, previously judged scenarios seemed to serve as a reference which influenced the judgment of subsequent scenarios. This reference is exogenous to the subsequent scenarios, but indispensable to predicting and explaining the reactions regarding them. Note that, under a certain order condition (LAF), the effects of objective situational parameters that can usually be found have almost entirely vanished. The strength of this effect demonstrates the large predictive potential of such exogenous factors and underpins the importance of spending more efforts on investigating them in the future.

According to our results, differential experiences prior to a moral judgment can have a profound influence on this judgment. Such effects can be expected to be especially large under conditions that strongly suggest the adequacy of transferring a certain judgment from one scenario to the next. This is the case if one person is required to give several subsequent judgments on various cases similar in structure or content in a within-subjects design. As our results suggest, extreme caution is required if responses generated under such conditions are to be attributed to properties of the scenarios themselves.

Finally, we believe that between-scenario order effects might also play a role under conditions outside the laboratory. The viewpoints taken by people discussing moral issues in everyday life might be highly affected by the issues that have been discussed immediately before. This influence might not only be quantitative, but even qualitative. Possible areas of application might be the design of public opinion polls or surveys that consecutively gauge responses to several (moral) issues. Previous research in other contexts showing that such instruments can be highly sensitive to effects of question positioning (e.g., Benton & Daly, 1991;) in combination with our results from the moral domain support this claim. On the other hand, we

acknowledge that the similarity between the dilemmas used in our study might particularly encourage the transfer of judgments between scenarios. It might be that in cases where the issues in question present a larger variability in structure or content, order effects would diminish, and ratings would be more similar to those made independently.

Summary and conclusions

In this article, we have argued that order effects can have a profound influence on judgments of actions in moral dilemma situations. Amongst order effects, we subsume cases in which a given action is judged differently when rated independently, as compared to when it has been preceded by one or several other scenarios. We began by reviewing the (scarce) literature on order effects in moral psychological research. We then reported the results of an experiment conducted in order to find out whether the pattern of results extracted from the literature reviewed could be replicated. For four out of our five scenarios this was the case: Three scenarios that received positive ratings when evaluated independently received negative ratings when directly preceded by a scenario that had been judged negatively. The ratings for Push were also in line with this pattern, since the proposed action in this case was rated negatively when judged independently and was not affected by the order of presentation.

However, one scenario where the proposed action received slightly negative ratings when judged independently (Trap) was also affected by the order of presentation. This finding motivated us to have a closer look at the results by performing an analysis treating the individual ratings as binary choices. Following this analysis, we reformulated the pattern as follows: In those cases in which a participant disagrees with the action proposed, this judgment is likely to be "transferred" to the judgment of the action in the next scenario, even if this action is rated positively when judged independently. However, positive ratings are not able to change the ratings of the next action into positive ones if people normally disagree with the action proposed in this case.

We went on by speculating what could explain the asymmetry between negative and positive ratings in terms of the potential to be transferred to the next case. One candidate feature discussed was the greater urge to justify prohibitions (negative ratings) compared to permissions (positive ratings). Of course, more research is needed in order to evaluate explanatory mechanisms underlying the observed asymmetry.

In the last section of the paper, we discussed the implications of our findings for descriptive theories of morality. We argued that descriptive moral researchers should be extremely cautious when interpreting results of experiments using within-subjects designs. Furthermore, we contended that they should devote more attention to general psychological mechanisms contributing to moral judgment in addition to focusing on features of particular scenarios.

Overall, the present study should draw the attention of descriptive theories of moral judgment to previously overlooked important sources of variance such as order effects. Due to the crucial implications of these findings, much more empirical and theoretical research needs to be done in the future in order to address determinants, mechanisms, and boundary conditions of the issues discussed here.

Acknowledgements

We thank Johanna Kirchhoff for data collection.

References

Benton, J. E. & Daly, J. L. (1991). A question order effect in a local government survey. *Public Opinion Quartely, 55,* 640-642.

Cushman, F., Young, L., & Hauser, M. (2006). The role of conscious reasoning and intuition in moral Judgment: Testing three principles of harm. *Psychological Science, 17*(12), 1082-1089.

Foot, P. (1967). The problem of abortion and the doctrine of the double effect. *Oxford Review, 5,* 5-15.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science, 293*(5537), 2105-2108.

Hauser, M. (2006). *Moral minds: How nature designed our universal sense of right and wrong*. Ecco.

Hauser, M., Cushman, F., Young, L., Kang-Xing Jin, R., & Mikhail, J. (2007). A dissociation between moral judgments and justifications. *Mind & Language, 22*(1), 1-21.

Lanteri, Chelini, & Rizzello (2008). Lanteri, A., Chelini, C., & Rizzello, S. (2008). An experimental investigation of emotions and reasoning in the trolley problem. *Journal of Business Ethics, 83*(4), 789-804.

Nichols, S., & Mallon, R. (2006). Moral dilemmas and moral rules. *Cognition, 100*(3), 530-542.

Norcross, A. (2008). Off her trolley? Frances Kamm and the metaphysics of morality. *Utilitas, 20*(01), 65-80.

Petrinovich, L., & O'Neill, P. (1996). Influence of wording and framing effects on moral intuitions. *Ethology and Sociobiology, 17*(3), 145-171.

Thomson, J. J. (1976). Killing, letting die, and the trolley problem. *The Monist, 59,* 204-217.

Preferences in Cardinal Direction

Marco Ragni (ragni@cognition.uni-freiburg.de)
Department of Cognitive Science, University of Freiburg
Germany

Benedikt Becker (beckerb@informatik.uni-freiburg.de)
Department of Cognitive Science, University of Freiburg
Germany

Abstract

How do we reason with imprecise spatial descriptions? Do reasoners typically prefer one conclusion (over another) consistent with the imprecise descriptions? Based on empirical findings we are able to give a positive answer for the second question for spatial reasoning with cardinal direction relations. Analyzing further the pattern of the preferred conclusion reflects the idea of informativeness of the description. In consequence, we briefly explain heuristics and present a Bayesian model representing subjective belief of the reasoner.

Keywords: Probabilistic Reasoning; Preferential Reasoning; Qualitative Reasoning

Introduction

Reasoning with spatial information requires sometimes to reason with incomplete information. Take for example,

Berlin is north-east of Paris.
Paris is north-west of Rome.

You can (based on this information alone, e.g. no background knowledge, no map) easily infer that Berlin must be north of Rome. But you cannot infer (based on this information alone) if Berlin is eastern or western of Rome. But if you have to reason without having assumptions about geographic positions – do we prefer certain relations? The question on how humans solve such deduction problems is at the core of qualitative reasoning. In other words, how do we infer new knowledge (a *conclusion*) from given knowledge, and moreover, what are the differences to formal approaches in artificial intelligence?

Formally, there are two main approaches in AI on how such reasoning problems can be solved: By the application of (transitivity) rules or by the construction and inspection of models. Principally, both approaches are equivalent (Russell & Norvig, 2003), i.e. it is not possible to derive more information with each of these methods. This equivalence, however, makes it harder to distinguish which method(s) is applied by humans while solving such problems. Nonetheless, a number of empirical studies investigates this research question by psychological means. The most prominent and best supported theory with respect to the number of effects that can be accounted for is the theory of mental models (MMT) (Johnson-Laird & Byrne, 1991) (to name only a few: the indeterminacy effect (Johnson-Laird & Byrne, 1991), the form of premises and the figural effect (Knauff, Rauh, Schlieder, & Strube, 1998), the wording of conclusions (Van der Henst & Schaeken, 2007), etc.). According to the MMT, linguistic processes are relevant to transfer information from the premises into a spatial array and back again, but the reasoning process itself relies on model manipulation only. A *mental model* is an internal representation of objects and relations in spatial working memory, which matches the state of affairs given in the premises. The semantic theory of mental models is based on the mathematical definition of deduction, i.e. a propositional statement φ is a consequence of a set of premises $\mathcal{P}$, written $\mathcal{P} \models \varphi$, if in each model $\mathcal{A}$ of $\mathcal{P}$, the conclusion φ is true.

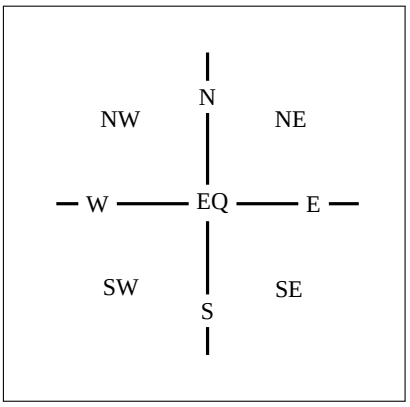


Figure 1: The nine base relations of the cardinal direction calculus in the projection based representation. Other representations are cone-based representations (Ligozat, 1998)

An interesting finding is the so-called preference effect, i.e. in multiple model cases (nearly always) one preferred model is constructed from participants and used as a reference for the deduction process (Rauh, Hagen, Schlieder, Strube, & Knauff, 2000). Further findings showed that during the validation phase alternative models are constructed by small modifications to the initially constructed model. This was the reason why the mental model theory for spatial reasoning was extended within the framework of preferred mental models (Rauh et al., 2000).

A new research line (Oaksford & Chater, 2007) focuses on Bayesian explanations for preferred solutions, e.g. for syllogistic reason. The authors use here the notion of informativeness to explain why a certain quantifier is used. The question is still open, if the Bayesian approach is sufficient to model spatial reasoning.

This paper is structured as follows: First, we will present an empirical investigation analyzing the question about pref-

erences in cardinal direction. Our empirical findings are then analyzed w.r.t. the main theories in the field (Theory of Mental Models, Theory of Mental Logic) with heuristics and we present a Bayesian model representing subjective belief of the reasoner. Finally, we discuss the different findings.

Preferences in Cardinal Direction

The language of cardinal direction consists of points in the euclidean plane $\mathbb{R}^2$. Based on the point algebra it is possible to distinguish 9 base relations $a, b \in \mathbb{R}^2$:

CD	EQ	N	NE	E	SE	S	SW	W	NW
PA	(=,=)	(=,>)	(>,>)	(>,=)	(>,<)	(=,<)	(>,<)	(<,=)	(>,>)

In other words $a\ N\ b := a_x = b_x \wedge a_y > b_y$, b is a northly. An *assignment* of a set of CD constraints C over the vocabulary $\mathcal{B} = \{\mathrm{N}, \mathrm{NE}, \mathrm{E}, \mathrm{SE}, \mathrm{S}, \mathrm{SW}, \mathrm{W}, \mathrm{NW}, \mathrm{EQ}\}$ is a function $\alpha : V(C) \to \mathbb{R}^2$, mapping each variable x, occuring in C to coordinates in the real plane.

Over the euclidean plane these *jointly exhaustive* and *pairwise disjoint*-base relations (cp. Fig. 1) with the composition table (cp. Figure 2) form a relation algebra. In the first experiment discussed here we used relations from the set $\mathcal{B}' := \mathcal{B} \backslash \{\mathrm{EQ}\}$ to construct a type of relational reasoning task that is referred to as three-term-series-problems (3ts-problems) in cognitive research (e.g. (Hunter, 1957)). In these tasks always two statements are used as premises and the task of the participants is to generate a statement that is consistent with the premises – the conclusion. E.g.,

A is northeast of B.
B is west of C.
Which relation holds between A and C?

The 3ts-problems can be formally described by the composition of two base relations and the question for a satisfiable relation. The set of all possible relations with premises $a\ R_1\ b$, $b\ R_2\ c$ are denoted by the composition $R_1 \circ R_2$. Normally, it is presented as a composition table (cf. Figure 2).

For the above example NE$\circ$W contains the following three relations: NE, N, NW. Since CD consists of nine base relations, there are without EQ 64 possible compositions of two base relations. In other words, exactly 64 different three-term-series problems exist. If we omit all one-relation cases (cells with one entry in Figure 2), it results in 40 multiple relation cases out of the 64 possible compositions. The participants of our studies were confronted with all 64 problems and had to infer a conclusion.

Empirical Data

The first central question we are interested in is: How do people reason about cardinal directions? Do they construct preferred mental models, and if so, what are the principles? An answer to this question might give hints of how preferences differ between large-scale spaces and small-scale spaces. For the latter scale of space, preferences have already been identified (Ragni, Fangmeier, Webber, & Knauff, 2007).

Participants. 24 students of the University of Freiburg took part in this web experiment (14m/10f, $M = 23.5/22.1, SD = 2.3/2.1$). They were paid for their participation.

	NW	N	NE	W	E	SW	S	SE
NW	1 [NW] NW 1.0 0.21	1 [NW] NW 1.0 0.21	3 [NW,N,NE] N 0.91 0.21	1 [NW] NW 1.0 0.29	3 [NW,N,NE] N 0.89 0.36	3 [SW,W,NW] W 0.83 0.14	3 [SW,W,NW] W 0.78 0.36	8(9) [B] W 0.29 0.0
N	1 [NW] NW 1.0 0.36	1 [N] N 1.0 0.14	1 [NE] NE 1.0 0.43	1 [NW] NW 1.0 0.29	1 [NE] NE 1.0 0.36	3 [SW,W,NW] W 0.67 0.36	2(3) [N,S] N 0.5 0.14	3 [SE,E,NE] E 0.64 0.21
NE	3 [NW,N,NE] N 1.0 0.29	1 [NE] NE 1.0 0.5	1 [NE] NE 1.0 0.36	3 [NW,N,NE] N 0.89 0.36	1 [NE] NE 1.0 0.29	8(9) [B] N 0.21 0.0	3 [SE,E,NE] E 0.73 0.21	3 [SE,E,NE] E 1.0 0.29
W	1 [NW] NW 1.0 0.29	1 [NW] NW 1.0 0.39	3 [NW,N,NE] N 0.6 0.29	1 [W] W 1.0 0.57	2(3) [W,E] E 0.58 0.14	1 [SW] SW 1.0 0.29	1 [SW] SW 1.0 0.21	3 [SW,S,SE] S 0.5 0.29
E	3 [NW,N,NE] N 0.88 0.43	1 [NE] NE 1.0 0.5	1 [NE] NE 1.0 0.29	2(3) [W,E] W 0.77 0.07	1 [E] E 1.0 0.29	3 [SW,S,SE] S 0.78 0.36	1 [SE] SE 1.0 0.43	1 [SE] SE 1.0 0.21
SW	3 [SW,W,NW] W 0.9 0.29	3 [SW,W,NW] W 0.78 0.36	8(9) [B] W 0.29 0.0	1 [SW] SW 1.0 0.57	3 [SW,S,SE] S 0.44 0.36	1 [SW] SW 1.0 0.43	1 [SW] SW 1.0 0.29	3 [SW,S,SE] S 0.91 0.21
S	3 [SW,W,NW] W 0.7 0.29	2(3) [N,S] S 0.82 0.21	3 [SE,E,NE] E 0.78 0.36	1 [SW] SW 1.0 0.43	1 [SE] SE 1.0 0.64	1 [SW] SW 1.0 0.5	1 [S] S 1.0 0.29	1 [SE] SE 1.0 0.29
SE	8(9) [B] SE 0.43 0.0	3 [SE,E,NE] E 0.73 0.21	3 [SE,E,NE] E 0.9 0.29	3 [SW,S,SE] S 1.0 0.36	1 [SE] SE 1.0 0.36	3 [SW,S,SE] S 1.0 0.21	1 [SE] SE 1.0 0.29	1 [SE] SE 1.0 0.21

Figure 2: The preferred relations in reasoning with cardinal direction. In each cell, the first number gives the number of correct relations and the relations. In the second row we have the preferred relation, then in the indeterminate case the relative frequency of this relation, i.e. how often it was chosen by the participants and then the error rates.

Materials. The experiment used the whole set of Cardinal Direction relations presented in Fig. 1. In the main part of the experiment all participants had to solve the same set of 64 3ts-problems. Here is an example-problem:

A is northwest of B.
B is southeast of C.
Which relation holds between A and C?

In half of the trials we asked for the relation between A and C and in half of the trials between C and A.

Procedure and Design. The experiment was conducted as a web experiment (partially conducted at our site for control) using webexp2. Tasks were presented in a randomized order. The premises were presented sequentially, i.e. the first premise disappeared when the second premise appeared. In other words, the participants were forced to hold the premise information in the working memory. All premises were presented in a self-paced procedure. Finally, the participants had to give a relation as an answer.

Overall, 87% of the problems were correctly solved. The results regarding the preference effects can be found in Figure 2.

As shown in Figure 2 out of the given 64 problems exactly 24 are determinate problems and 40 are indeterminate problems. Most of the indeterminate problems exactly 90% (only 4 relations were not significantly preferred: $N \circ S$, $W \circ SE$, $W \circ E$, $SW \circ E$) were solved with a clear preference for one relation. However, it is remarkable that several relations could have been chosen as a possible conclusion, but, in fact, the participants chose just one of them and their preferences also often corresponded.

Discussion. There are differences between preferred relations in small-scale spaces and in large-scale spaces. Contrary to the small-scale spaces (Ragni, Fangmeier, et al., 2007) where the first-free fit strategy has been identified in relational reasoning in large-scale spaces participants used a first-fit strategy. In other words they inserted the third object C in-between A and B (cp. the relations $S \circ N$ and $E \circ W$ where in the first case S and in the second W has been reported). The inverse composition $N \circ S$ and $W \circ E$ are not statistically significant.

By a formal analysis it was possible to explain the preferred mental model in indeterminate cases by the following distinction

- Principle 1 (*In-between Insertion Principle*): If the two relations of the composition are inverse (e.g. S and N, W and E) then the third object C is inserted in-between A and B,(e.g. A is S of C and B is north of C, and so on).
- Principle 2 (*Cut Principle*): Choose always the relation in the geometrical cut of the two relations, i.e. if $NE \circ NW$ is composed and the relations NW, N, NE are possible than the relation N is chosen.

The participants preferred the cut between relations, e.g. in the composition of $NE \circ NW$ and $NW \circ NE$ they preferred the relation N. The same pattern holds as well for $SW \circ NW$ and so on. This gives an indication that without additional information they use (independently of projection based or cone based representation of Cardinal Direction) similar distances.

Theories of Deduction

In this section we ground the intuitively used theories formally (and mathematically) and analyze them with respect to their reasoning power.

A *relational structure* is a tuple $(D, (R_i)_{(i \in I)})$ consisting of a domain D (sometimes called discourse universe) and a set of (usually binary) relations R_i (Russell & Norvig, 2003). For example, geographic knowledge like *New York is north-east of Washington* can be expressed by cardinal direction relations $N, NE, E, SE, \ldots$ over the domain of cities. More complex expressions can be formed by using connectives like conjunctions (New York is north-east of Washington *and* New York is in the U.S.) and disjunctions ($\ldots$ *or* $\ldots$). By allowing negations, we have a propositional relational language $\mathcal{L}$ over cardinal direction relations. Such relational structures can be used to describe *knowledge representation*. But how can new information be derived?

Theory of mental logic

The theory of mental logic (Rips, 1994) assumes that we use (transitivity) rules to draw conclusions, whereas the classical model theory argues that we use models for this inference process. The classical mental model theory (Byrne & Johnson-Laird, 1989) claims that in multiple model cases (i.e. more than one model is consistent with the premises) other models are inspected.

1. $West(x,y)$ & $North(z,x) \rightarrow West(z,y)$
2. $West(x,y)$ & $North(z,y) \rightarrow West(x,z)$
3. $West(x,y)$ & $West(y,z) \rightarrow West(x,z)$
4. $West(x,y) \leftrightarrow East(y,x)$
5. $(West(y,x)$ & $West(z,x)) \rightarrow (West(y,z)$ *or* $West(z,y))$
6. $(West(y,z)$ *or* $West(z,y))$ & $North(w,z) \rightarrow$ $(West(y,w)$ *or* $West(w,y))$

Figure 3: Set of (incomplete) inference rules specified for spatial reasoning adapted from Van der Henst (2002).

The central idea of this approach can be characterized as follows: "Reasoning consists in the application of mental inference rules to the premises and conclusion of an argument. The sequence of applied rules forms a mental proof or derivation of the conclusion from the premises, where these implicit proofs are analogous to the explicit proofs of elementary logic" (Rips, 1994, p. 40). Hagert (1984) defined a first set of spatial inference rules (cf. Fig. 2). This set of rules has been extended by two additional rules (cf. the rules 5 and 6 in Fig. 2) to deal with indeterminacy by Van der Henst (2002). The rules in Fig. 2 are successively applied to the premises of a problem description.

There is, however, no recent theory in explaining the construction of the preferred relations (Figure 2).

Theory of mental models

The mental model theory assumes that the human reasoning process consists of three distinct phases: The *model generation phase*, in which a first model is constructed out of the premises, an *inspection phase*, in which the model is inspected to check if a putative conclusion is consistent with the current model. In the *validation phase*, finally, alternative models are generated from the premises that refute this putative conclusion. The indeterminacy effect is mainly responsible for human difficulty in reasoning (Johnson-Laird, 2001).

Recent findings indicate a phenomenon encountered in multiple-model cases, namely that humans generally tend to construct a *preferred mental model* (PMM). This model is easier to construct, less complex, and easier to maintain in working memory compared to all other possible models (Knauff et al., 1998). The principle of economicity is the determining factor in explaining human preferences (Manktelow, 1999). This principle also explains that a model is constructed incrementally from its premises. Such a model construction process saves working memory capacities because each bit of information is immediately processed and integrated into the model (Johnson-Laird & Byrne, 1991). In the model variation phase, this PMM is varied to find alternative interpretations of the premises (Rauh et al., 2000). From a formal point of view, however, this theory has not been formalized yet and is therefore not fully specified in terms of necessary operations to process such simple problems as were described above.

A model $\mathcal{A}$ is called *consistent* with a set of premises Φ over a relational language $\mathcal{L}$ (mathematically $\mathcal{A} \models \Phi$) if all

expressions of Φ are true in $\mathcal{A}$. Then a conclusion Ψ can be derived from the premise set Φ (mathematically $\Phi \models \Psi$, whereby $\models$ is called the *consequence relation*) if

$$\begin{aligned} \Phi \models \Psi \quad &\Leftrightarrow \quad \text{All models of } \Phi \text{ are models of } \Psi. \\ &\Leftrightarrow \quad \text{There is no model } \mathcal{A} \text{ with} \\ &\qquad \mathcal{A} \models \Phi \text{ and } \mathcal{A} \models \neg\Psi. \end{aligned}$$

A model $\mathcal{A}$ with the property $\mathcal{A} \models \Phi$ and $\mathcal{A} \models \neg\Psi$ is called *counter-example*. It follows if there is a counter-example to Φ and Ψ then $\Phi \models \Psi$ cannot hold.

This classical (mathematical) consequence relation, however, does not explain how initial mental models are constructed and varied (Rauh et al., 2000). Since there is a huge empirical evidence supporting the preferred mental model theory for different calculi (Rauh et al., 2000; Ragni, Fangmeier, et al., 2007; Ragni, Tseden, & Knauff, 2007) it seems worth to ground this theory mathematically.

A Probabilistic Approach

As already stated, a new approach are probabilistic models (Oaksford & Chater, 2007) to explain preferred relations. Those are based on the consideration to use probabilities instead of truth values as the representation of semantics. This is a valid consideration as a probability might be interpreted in a *subjective* manner describing a subjective degree of belief rather than a relative frequency of an event. Following this subjective interpretation probability theory can be utilized for belief updating and inference. The probabilistic approach to inference is based on:

$$P(\text{“If } p \text{ then } q\text{”}) = P(q|p). \tag{1}$$

Thus, the probability of a conditional proposition is identified with the conditional probability of the proposition. The a-posteriori belief in the fact q in face of certainty about the fact p is given by the a-priori conditional probability: $P_1(q) = P_0(q|p)$, if $P_1(p) = 1$. This is called “conditionalization”. It constitutes the basis of probabilistic inference.

The probabilistic representation of conditionals as given in equation 1 enables the application of Bayes’ theorem:

$$P(q|p) = \frac{P(p|q)\,P(q)}{P(p)} \tag{2}$$

This has two advantages: First, whereas $P(q|p)$ is a rather abstract value, the probabilities of its right hand side can often be derived from the agent’s experience. Second, it implies basic patterns of performance while reasoning with conditional propositions which appear as “errors and biases” from a logicistic standpoint.

Bayesian Rationality arises from a rational analysis of the problem, the environment, and the constraints of an agent while conducting deductive tasks. As such, it is not a theory of the actually psychological processes in use, but a description of general regularities. It is further independent of cognition *about* probabilities. It shows that cognition often obeys the laws of probabilistic theory.

The following models[1] are to reproduce the frequency distribution of the 3ts-task on cardinal directions this way.

Spatial Bayesian Models The spatial reasoning task of the previous section uses the set of cardinal relations $\mathcal{B}' = \{\mathrm{N}, \mathrm{E}, \mathrm{S}, \mathrm{W}, \mathrm{NE}, \mathrm{SE}, \mathrm{SW}, \mathrm{NW}\}$. The statement of an item in the 3ts-task is given by a pair of relations $R_1, R_2 \in \mathcal{B}'$ with aR_1b and bR_2c for three locations a, b, c. The subject’s guess for the relation between a and c is another relation $R_3 \in \mathcal{B}'$. The relative frequency of R_3 for an item R_1, R_2 will be referred to as $f_{R_1,R_2}(R_3)$.

The objective of a Bayesian model for the 3ts-task is to implement a probability distribution of R_3 parametrized by the task item R_1, R_2, i.e. $P_{R_1,R_2}(R_3)$. This probability distribution is assumed to be a prediction of the experiment’s relative frequencies f_{R_1,R_2}. Thus, model M’s preferred relation given the task’s relations R_1 and R_2 is

$$M(R_1, R_2) := \operatorname*{arg\,max}_{R_3 \in \mathcal{B}'} P_{R_1,R_2}(R_3).$$

The per-item probability distribution of R_3 can be identified with the probability of R_3 conditioned by the item’s relations. Therefore, it further allows the application of Bayes’ theorem (equation 2):

$$P_{R_1,R_2}(R_3) := P(R_3|R_1,R_2) = \frac{P(R_1,R_2|R_3)\,P(R_3)}{P(R_1,R_2)} \tag{3}$$

Consequently, it is sufficient for a Bayesian model of the 3ts-task to specify merely the reversed conditional probability $P(R_1,R_2|R_3)$ as well as the marginal probabilities $P(R_3)$ and $P(R_1,R_2)$.

The following sections will describe such implementations. The quality of each model M will be compared to the empirical data by three factors: a) the mean correlation C^M between P_{R_1,R_2} and the empirical data f_{R_1,R_2}, b) the sum E^M of the squared differences between P_{R_1,R_2} and f_{R_1,R_2} and c) the number N^M of correctly predicted preferred relations.

The Unit Layout (Model M_1) The computation of $P^{M_1}(R_1,R_2|R_3)$ is based on a heuristic for detours when moving by R_3 in the so called *unit layout*. R_1 and R_2 describe the detour. The farther the detour the smaller is the conditional probability of R_1, R_2.

The unit layout is a rectangular subset of $\mathbb{Z}^2$ and separately defined for each direction R_3. The brackets $[\cdot]^{R_3}$ map the locations a and c each to a field in $\mathbb{Z}^2$ such that $[a]^{R_3}\,R_3\,[c]^{R_3}$. Each pair of relations R_1, R_2 with $R_3 \subset R_1 \circ R_2$ is likewise mapped to a field in $\mathbb{Z}^2$ by $[\cdot]^{R_3}$ such that

$$[a]^{R_3}\,R_1\,[R_1R_2]^{R_3} \text{ and } [R_1R_2]^{R_3}\,R_2\,[c]^{R_3}.$$

The fields $[a]^{R_3}$ and $[c]^{R_3}$ must be chosen in such a way that each $[R_1R_2]^{R_3}$ is definite. That way, the unit layout is definite in $\mathbb{Z}^2$ up to translations. Figure 4 shows the unit layout for $R_3 = \mathrm{NW}$.

[1] The source code is available at `http://tiny.cc/hmi3f`.

SE-NW	S-NW	SW-NW	SW-N	SW-NE
E-NW	*a*	W-NW	W-N	W-NE
NE-NW	N-NW	NW-NW	NW-N	NW-NE
NE-W	N-W	NW-W	*c*	NW-E
NE-SE	N-SE	NW-SE	NW-S	NW-SE

Figure 4: The *unit layout* for $R_3 = \text{NW}$. Field *a* is to the northwest of *c*. All other field are uniquely labeled with relations R_1-R_2. It holds for each of them that field *a* is R_1-wards of it and it is R_2-wards of *c*.

For $R_3 \subset R_1 \circ R_2$, the unit layout entails the costs of a "detour" moving from field $[a]^{R_3}$ via $[R_1R_2]^{R_3}$ to $[c]^{R_3}$ utilizing a metric d on $\mathbb{Z}^2$.

$$c^{R_3}_{R_1,R_2} := \frac{d([a]^{R_3},[R_1R_2]^{R_3}) + d([R_1R_2]^{R_3},[c]^{R_3})}{d([a]^{R_3},[c]^{R_3})}$$

The costs $c^{R_3}_{R_1,R_2}$ for $R_3 \not\subset R_1 \circ R_2$ are defined by the model parameter distimposs. This cost measure entails the wanted conditional probability:

$$P^{M_1}(R_1,R_2|R_3) := \frac{{c^{R_3}_{R_1,R_2}}^{-1}}{\sum_{R_1',R_2' \in \mathcal{B}'} {c^{R_3}_{R_1',R_2'}}^{-1}}.$$

This points out the influence of the model parameter distimposs: For infinity, the model performs accurate and it simulates errors for positive numbers.

This is how model M_1 computes the conditional probability of the right-hand side of equation 3. The marginal probability of $P^{M_1}(R_3)$ is a unit distribution which can be furnished with a probability gain for the main cardinal directions by cardinalgain and an additional gain towards the west by westgain. The probability of $P^{M_1}(R_1,R_2)$ is assumed to be a unit distribution.

Parameter Variations Varying the metric d between Euclidian, Manhattan, and maximum had no noteworthy effect on the quality estimation factors (C^M, E^M, N^M). So we chose the euclidian metric, as it matches the intuitive concept of distance best. The model parameter distimposs was varied systematically between 20 and 200, the parameters cardinalgain and westgain were varied between 0.1 and 0.9.

We found a maximal convergence against the empirical data with model parameters distimposs $= 150$, cardinalgain $= 0.2$, and westgain $= 0.2$.

It has a mean correlation $C^{M_1} = 0.91$, a summed error of $E^{M_1} = 2.82$ and predicts the preferred relation correctly in $N^{M_1} = 59$ cases. This instance of model M_1 has a mean correlation of 0.96 in 60 items of the task. Nevertheless, the mean correlation for the task items with opposing intermediate directions is as little as 0.17. This suggest the appearance of another strategy in these cases.

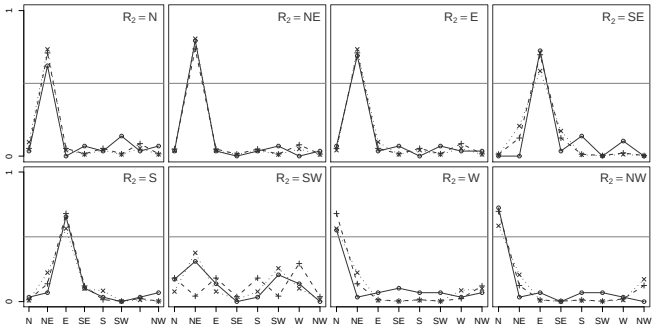


Figure 5: Relative frequencies of R_3 for task items with $R_1 =$ NE from the experiment (circles: $\circ$) as well as M_1's (pluses: $+$), and M_2's (crosses: $\times$) probabilities .

A Secondary Strategy (Model M_2) Model M_2 is an extension of the model presented in the preceding section. It adds a probability gain g_{R_1,R_2} to the value of $P^{M_1}_{R_1,R_2}$. This gain implements priming effects on the relations R_1 and R_2. The amounts of priming towards R_1 and R_2 are controlled by the model parameters firstprim and secondprim, respectively. Values of 0 each void the priming effect.

The extent of this probability gain is in turn controlled per task item by the certainty z_{R_1,R_2} of M_1

$$z_{R_1,R_2} := \max_{R_3 \in \mathcal{B}'} P^{M_1}_{R_1,R_2}(R_3).$$

The (yet to be normalized) probability distribution of M_2 is defined as

$$P^{M_2}_{R_1,R_2}(R_3) := z_{R_1,R_2} \cdot P^{M_1}_{R_1,R_2}(R_3) + (1 - z_{R_1,R_2}) \cdot g_{R_1,R_2}(R_3).$$

It weakens M_1's probability distribution $P^{M_1}_{R_1,R_2}$ and strengthens the priming effect g_{R_1,R_2} as a function of decreasing certainty.

Parameter Variations In a systematic search through the parameters of model M_1 as well as firstprim and secondprim we found an instance of M_2 with mean correlation of $C^{M_2} = 0.94$, summed error $E^{M_2} = 2.67$ and $N^{M_2} = 62$ correctly predicted items. Along with it, this instance has a mean correlation of 0.73 for the task items with opposed intermediate directions. The parameters were distimposs = 180, cardinalgain = 0.1, westgain = 0.2, firstprim = 0.3 and secondprim = 0.2.

Figure 5 shows results both from model M_1 and M_2 for $R_1 =$ NE. M_2's improvement is apparent for $R_2 =$ SW.

Interpretation

The following lines give a clue of how the found model parameters can be read as a hints on the underlying cognitive processes.

Utilizing Experience The first model, M_1, shows that the spatial reasoning task can be modelled by a Bayesian approach. The computation of $P(R_3|R_1,R_2)$ is based on an "intuition of the benefit" to move towards R_1 first and then towards R_2 to attain towards R_3 overall. This intuition might

reflect complying knowledge of the subject arising from basic experience navigating through the world.

It was possible to further increase the convergence of the model towards the empirical data by means of higher marginal probabilities of the cardinal directions, and additionally the west. This might reflect frequency effects for the cardinal directions as well as an effect of the reading direction for the western direction.

Shifting Strategies Whereas model M_1 behaved poorly for tasks with opposed intermediate directions, model M_2's correlation on those could be improved by simulating priming effects on the relations given by the current task item. Those tasks excel in a high uncertainty about the answer. This suggests the subjects shift their strategy to be driven by priming effects under uncertainty.

General Discussion

If incomplete information is available only (i.e. several relations are possible), humans tend to take a relation more into account than others. This finding complements a series of findings for preferred spatial reasoning with intervals (Rauh et al., 2000), with the spatial relations right and left (Jahn, Knauff, & Johnson-Laird, 2007), and with topological relations (Ragni, Tseden, & Knauff, 2007).

Our starting point was the question if it is possible to model preference effects for cardinal directions in both theories (the Mental Model Theory and the Bayesian rationality) based on heuristics. Only by a formalization it is possible to compare human reasoning to approaches in AI. A formal handling of the preferred mental model theory by a consequence relation allows to make precise predictions about which kind of conclusion(s) are drawn (from a given set of premises) and which are neglected. These heuristics can be described by two principles: the in-between insertion principle and the cut principle. Both together can explain the preferences in the composition table (Figure 2) and support the theory of cognitive economicity (Manktelow, 1999).

The primer raised question, if the Bayesian approach is expressible enough to model preference effects in spatial reasoning (with cardinal directions) can be positively answered. Moreover, it reproduces the full frequency distribution quiet well: The first model is based on a heuristic for detours which explains the preferences (Figure 2). It has a mean correlation of 0.91 and predicts the preferred relation correctly in 59 from 64 cases. The second model which adds a priming effect leads to an increase from 0.17 to 0.73 in the correlation in the four cases of opposed intermediate directions.

A possible limitation of the Bayesian model is connected to the certainty of the conclusion. While each statement is given with absolute certainty (Berlin is north-east of Paris) a conclusion has only a degree of certainty. Taken together, the results clearly indicate that the preference effect can be explained by heuristics in both mental models and bayesian approach. Further research necessarily requires an investigation for a general heuristic explaining preference relations for the diverse spatial calculi.

One point, however, is certain: the role of heuristics has been vastly underestimated in explaining the preferences in spatial reasoning.

References

Byrne, R. M., & Johnson-Laird, P. N. (1989). Spatial reasoning. *Journal of Memory & Language*, *28*(5).

Hagert, G. (1984). Modeling mental models: Experiments in cognitive modeling spatial reasoning. In T. O'Shea (Ed.), *Advances in artificial intelligence.* Elsevier.

Hunter, I. M. (1957). The solving of three-term series problems. *Br J Psychol*, *48*(4), 286–98.

Jahn, G., Knauff, M., & Johnson-Laird, P. N. (2007). Preferred mental models in reasoning about spatial relations. *Memory and Cognition*.

Johnson-Laird, P. N. (2001). Mental models and deduction. *Trends in Cognitive Sciences*, *5*(10).

Johnson-Laird, P. N., & Byrne, R. M. J. (1991). *Deduction.* Hove (UK): Erlbaum.

Knauff, M., Rauh, R., Schlieder, R., & Strube, G. (1998). Continuity effect and figural bias in spatial relational inference. In *Proceedings of the 20th cognitive science conference* (pp. 573–578). Mahwah, NJ: Lawrence Erlbaum.

Ligozat, G. (1998). Reasoning about cardinal directions. *Journal of Visual Language Computing*, *9*(1), 23–44.

Manktelow, K. (1999). *Reasoning and Thinking.* Hove: Psychology Press.

Oaksford, M., & Chater, N. (2007). *Bayesian rationality the probabilistic approach to human reasoning*. Oxford University Press.

Ragni, M., Fangmeier, T., Webber, L., & Knauff, M. (2007). Preferred mental models: How and why they are so important in human spatial reasoning. In C. Freksa, M. Knauff, B. Krieg-Brückner, B. Nebel, & T. Barkowsky (Eds.), *Spatial cognition v.* Berlin: Springer.

Ragni, M., Tseden, B., & Knauff, M. (2007). Cross cultural similarities in topological reasoning. In S. Winter, M. Duckham, L. Kulik, & B. Kuipers (Eds.), *Proceedings of 8th International Conference on Spatial Theory, COSIT 2007* (Vol. 4736). Springer.

Rauh, R., Hagen, C., Schlieder, C., Strube, G., & Knauff, M. (2000). Searching for Alternatives in Spatial Reasoning: Local Transformations and Beyond. In *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum Associates.

Rips, L. J. (1994). *The psychology of proof: Deductive reasoning in human thinking*. Cambridge, MA: MIT Press.

Russell, S., & Norvig, P. (2003). *Artificial intelligence: A modern approach* (2nd ed.). Prentice-Hall.

Van der Henst, J. (2002). Mental model theory versus the inference rule approach in relational reasoning. *Thinking and Reasoning*, *8*, 193-205.

Van der Henst, J., & Schaeken, W. (2007). The wording of conclusions in relational reasoning. *Cognition*, *97*, 1–22.

Expanding Retrieval Promotes Long Term Retention by Preventing Rapid Rates of Forgetting

Aimee A. Callender (aac0005@auburn.edu)
Department of Psychology, 226 Thach Hall
Auburn, AL 36849 USA

Abstract

Expanding retrieval, increasing the delay between retrieval attempts of recently studied material, should lead to better memory than equally spaced retrieval, however, recent results have been mixed. Ninety-six participants studied word pairs with the following expansion schedules: 1-2-3; 1-5-9; 3-5-8; 5-8-13. An evenly spaced (5-5-5) condition was also used. A final test was given after a 10 minute or 48 hour retention interval. Performance after the 48 hour retention interval was best in the 5-8-13 condition. The higher level of performance in this condition was due to no forgetting between the final learning trial and the immediate final test.

Keywords: Memory; Retrieval; Expanding Retrieval;

Introduction

Spacing out study trials and retrieval attempts is a simple and effective way to improve memory for verbal material(see Cepeda, Pashler, Vul, Wixted & Rohrer, 2006 for a review). Repeated testing has also produced robust memory effects in the laboratory (Roediger & Karpicke, 2006) and classroom studies of repeated testing. One aspect of repeated testing that has produced mixed results is the effects of timing of the repeated tests or retrieval practice on long term retention. Two basic ways that repeated testing can be implemented are to evenly space out the retrieval attempts or to gradually increase the interval between each retrieval attempt, called expanding retrieval. Intuitively, expanding retrieval is thought to improve long term retention, yet the literature is mixed. This experiment investigated two different aspects of the retrieval schedule, the number of intervening items between the study period and the first retrieval attempt and the number of intervening items in the remainder of the schedule, to further delineate the conditions under which expanding retrieval may or may not improve long term retention.

Expanding Retrieval

Expanding retrieval is a method in which an initial study period is followed by retrieval attempts that are spaced out with increasingly longer intervals between each attempt. For example, a typical expansion sequence used in the literature is a 1-5-9 sequence (see Cull, Shaughnessy, & Zechmeister, 1996; Karpicke & Roediger, 2007), meaning that after the initial study period, there is one intervening item between study and retrieval, then five intervening items between the first and second retrieval attempts, and finally, nine intervening items between the second and third retrieval attempts.

Intuitively, and anecdotally, expanding retrieval is thought to produce better long term retention than evenly spaced retrieval. Theoretically, one would expect expanding retrieval to improve long term memory based on the principle of desirable difficulty (Bjork, 1994). According to Bjork (1994), conditions that introduce difficulty into the learning process increase the likelihood of retrieving the information on a long term retention test. Expanding retrieval is built on the idea of introducing difficulty into the learning process. The initial conceptualization of expanding retrieval posited that the first retrieval attempt should occur soon after the learning trial to ensure successful retrieval. Difficulty is introduced on subsequent retrieval attempts by increasing the number of intervening items between each attempt. Gradually increasing the delay between each retrieval attempt makes each subsequent attempt more difficult than the previous attempt. Surprisingly, recent research investigating the benefits of expanding retrieval has produced mixed results. Some studies have found that expanding retrieval does improve memory when compared to an evenly spaced control condition (Cull et al., 1996) whereas others found that expanding retrieval is no better than evenly spaced retrieval (Logan & Balota, 2008).

Karpicke and Roediger (2007) explored possible reasons for the mixed results and found that the benefits of expanding retrieval depend on two factors: the time of the final test (immediate or delayed) and the number of items between the study trial and the first retrieval attempt. With respect to the time of the final test, expanding retrieval produced better performance than evenly spaced retrieval on a test that occurred 10 minutes after learning. When the final test occurred after a 48 hour retention interval, the effects reversed, and equally spaced retrieval produced superior performance compared to the expanding retrieval conditions. The results of the second factor that was investigated, the delay between the study period and the first retrieval attempt, indicated that expanding retrieval improved retention the most when the first retrieval attempt was delayed regardless of the rest of the sequence (whether it was evenly spaced or expanding).

Although Karpicke and Roediger (2007) addressed and reduced the confusion surrounding expanding retrieval, the extant research on expanding retrieval is restricted by two limitations. One limitation (see Karpicke and Roediger, 2007, Exp. 3) was that the expanding sequences that were used to investigate the number of intervening items between

study and the first retrieval attempt were not true expanding sequences. In order to investigate the interval between the study trial and the first retrieval attempt and to control for the number of intervening items, the sequences were constructed by simply adding an initial retrieval attempt onto the standard 5-5-5- and 1-5-9 sequences. The evenly spaced sequence became a 5-5-5-5 sequence and the expanding sequence became a 5-1-5-9 sequence. The expanding sequence was no longer a true expanding sequence as it contracted before expanding. This experiment addressed this issue by comparing the 5-5-5 sequence against an expanding sequence that started with a first retrieval attempt after 5 intervening items (5-8-13).

The other limitation is that few other studies investigating expanding retrieval have distinguished between the effectiveness of different expansion sequences (but see Logan & Balota, 2008), and most studies have used a 1-5-9 expanding sequence (Cull et al, 1996; Karpicke & Roediger, 2007; Morris, Fritz, Jackson, Nichol, & Roberts, 2005). Other schedules have been used, for example Landauer and Bjork (1978) used 1-4-10 and 0-1-3-8 schedules, and in an experiment with preschool age children, Fritz, Morris, Nolan and Singleton (2007) used a 1 minute-1day-2 day expansion sequence. However, the different schedules have not been compared against one another. Other than Karpicke and Roediger's (2007) investigation of the initial retrieval attempt and Logan and Balota's (2008) study of various schedules of spacing, the remainder of the expansion schedule has not been a variable of interest. Thus, the present experiment investigated different expansion sequences to determine if the benefits of expanding retrieval depend on the particular sequence that is used.

According to Bjork's theory of Desirable Difficulty it is critical to create a task that is challenging and difficult for the individual, but is not so difficult such that the individual cannot complete the task. The difficulty of the expansion sequence can be manipulated by increasing the number of intervening items between the study trial and the first retrieval attempt or by increasing the average number of intervening items across the entire expansion schedule. This experiment manipulated both types of difficulty, but focused on the number of intervening items between the study trial and the first retrieval attempt.

The Current Experiment

The current experiment investigated 5 different sequences: an evenly spaced control (5-5-5) and four expansion sequences (1-2-3; 1-5-9; 3-5-8; 5-8-13). The 1-5-9 sequence was chosen because it has been used in the majority of previous research. Although the expansion sequences may appear to be chosen at random, the three novel sequences used in this study (1-2-3; 3-5-8; 5-8-13) were chosen because they come from the same number sequence that occurs in nature, the Fibonacci sequence. The Fibonacci sequence is a naturally expanding sequence in which each number is determined by adding together the two previous numbers in the sequence. By using portions of the Fibonacci sequence for each of the novel sequences used in this experiment, each expanding sequence expands in the same way.

The expansion sequences used in this study were chosen from the Fibonacci sequence by considering two factors. First, the critical issue in this experiment was to investigate the number of intervening items between the study trial and the first retrieval attempt. Thus, sequences were selected by choosing different starting points in the Fibonacci sequence to correspond to the initial retrieval attempts that have been used in previous literature (1 or 5). This resulted in the 1-2-3 and 5-8-13 sequences. The other factor that was considered was the average number of intervening items across the entire sequence. The average number of intervening items in the 5-5-5 and 1-5-9 sequences that are typically used in the literature is 5. Thus, the 3-5-8 condition was selected.

The comparison of most interest was to investigate the difference in performance between the 5-5-5 condition and the expanding condition that also had five intervening items between the study trial and the first retrieval attempt, 5-8-13. This comparison was vital because Karpicke and Roediger's (2007) investigation of this variable found that delaying the first retrieval attempt improved performance regardless of the rest of the expansion sequence. However, the expansion sequences were different from the expansion sequences used in this experiment as their sequences started over from 1 after the first retrieval attempt (5-1-5-9) whereas the expansion sequence in the present experiment continued to expand from that first retrieval attempt (5-8-13).

Method

Participants

Ninety-six undergraduate students at Auburn University participated in the study in exchange for extra credit for a psychology course. Participant were between the ages of 17 and 25, 74% were female, 26% were male, 88% were white, 9% were African American, and 3% identified themselves as another race/ethnicity.

Design and Materials

The experiment used a 2 (retention interval) x 7 (testing schedule) mixed design. The final retention interval (10 minutes or 48 hours) was manipulated between subjects and the various testing schedules were manipulated within subjects. The testing schedules included an evenly spaced control (5-5-5), 3 expanding schedules based on the Fibonacci sequence (1-2-3; 3-5-9; 5-8-13), a standard expanding schedule (1-5-9), and two single test conditions in which participants took a single test immediately (Single0) or after 2 intervening items (Single2).

The experiment was based on Karpicke and Roediger (2007), using word pairs in which the first word was an unfamiliar (low frequency) word such as *Tumbrel,* and the

second word was a one word synonym or definition, *Cart*. During the study phase, the word pairs were presented together, *Tumbrel-Cart*, and during the testing phases the first word was presented and participants were required to type in the appropriate word pair. Fifty-six word pairs were constructed. Forty of the word pairs were critical word pairs. For each of the 7 testing conditions, there were 5 word pairs to study. The remaining 5 items served as unstudied control items. Eight counterbalancing conditions were constructed to allow for each set of 5 word pairs to be rotated through each of the testing and unstudied conditions. This resulted in 40 critical word pairs (5 for each of the 7 conditions, counterbalanced so remaining 5 were unstudied or control items). Additionally, 16 filler items were included for a total of 56 word pairs. Two filler items were used as buffers at the beginning of the task, and 3 filler items were used as buffers at the end of the task. The remaining fillers were interspersed throughout the list to allow for the appropriate spacing of all of the study and test trials. This resulted in a total of 142 trials in the experiment.

Procedure

Data collection occurred in groups of up to 15 participants in a computer lab. Participants were instructed that they would study word pairs that included one familiar word and one unfamiliar word. They were instructed to study the word pairs and type in the appropriate response during the test trials. Each of the trials (study or test) was 8 s. with a 500 ms. intertrial interval. Participants had to spend the entire 8 s. viewing the study screen, but during the test trials they were allowed to press Enter to move on to the next trial once they entered their response. If no response had been entered within 8 s., the computer program automatically advanced to the next trial. This task (142 trials) generally took between 15 and 20 minutes for participants to complete.

The final retention test tested participants on the 40 critical word pairs. Thirty-two of the participants took the final test after 10 minutes, and 32 took the test after a 48 hour delay. Participants were instructed that they were going to be tested on the words that they had learned previously, and to type in the appropriate word pair. Each trial was 14 s. (participants could press Enter to advance to the next trial once they entered a response) and the interstimulus interval was 500 ms. If no response had been entered in, the computer program automatically advanced after 14 seconds. This task generally took about 10 minutes to complete.

Results

Three separate analyses were conducted. The first analysis investigated performance on the learning trials, that is, each retrieval attempt during the expanding or evenly spaced sequence. A second analysis was conducted on final cued-recall performance. The third analysis investigated forgetting between the last learning trial and the final cued-recall test.

Learning Trials

A 5(spacing condition) x 3 (learning trial) within subjects ANOVA was conducted (the single trial conditions were not included in this analysis). The effect of learning trial was significant, F (2, 464) = 5.82, p = .003, *partial* $?^2$ = .02, which was qualified by a significant interaction of learning trial and spacing condition F (8, 930) = 4.52, p < .001, *partial* $?^2$ = .04. Table 1 shows that performance in the two conditions in which there was only one intervening item between the study trial and the initial retrieval attempt (1-2-3 and 1-5-9) was much higher on the learning trials than the other three spacing conditions (3-5-8; 5-8-13; 5-5-5), indicating that the trials that had more intervening items before the first retrieval attempt were more difficult to learn (See Table 1).

Table 1: Performance on learning trials by spacing condition.

Spacing	Trial 1	Trial 2	Trial 3
1-2-3	.68 (.02)	.56 (.02)	.58 (.02)
1-5-9	.66 (.03)	.52 (.03)	.56 (.06)
3-5-8	.43 (.03)	.48 (.07)	.42 (.03)
5-8-13	.46 (.03)	.44 (.03)	.43 (.04)
5-5-5	.43 (.03)	.44 (.03)	.45 (.03)

Note. Standard errors are in parentheses.

Final Cued Recall

A 2 (delay) x 7 (spacing condition: 1-2-3; 1-5-9; 3-5-8; 5-8-13; 5-5-5; Single0; Single2) repeated measures ANOVA was conducted on the final recall performance. Both the within-subjects effect of spacing and the effect of delay were significant; F (6, 89) = 25.45, p < .001, *partial* $?^2$ = .64, and F (1, 94) = 26.14, p < .001, *partial* $?^2$ =.22, respectively. The spacing by delay interaction was marginally significant, F (6, 89) = 2.15, p = .06, *partial* $?^2$ = .07. As Table 2 shows, performance in the 5-8-13 condition was the highest in both the immediate condition (although not significantly) and was significantly higher than the other spacing conditions (except for the 1-5-9 condition) after the 48 hour delay. The 5-5-5 spacing condition was not significantly better than the other spacing conditions on either the immediate or delayed test.

Table 2: Final recall as a function of spacing condition and delay.

Spacing	10 min.	48 hours
1-2-3	.37 (.03)	.17 (.03)
1-5-9	.37 (.04)	.18 (.04)
3-5-8	.34 (.04)	.15 (.04)
5-8-13	.42 (.04)	.25 (.04)
5-5-5	.38 (.04)	.15 (.04)
Single 0	.14 (.02)	.02 (.02)
Single 2	.23 (.03)	.07 (.03)

Note. Standard errors are in parentheses.

Forgetting

A 2 (retention interval) X 5 (spacing condition: 1-2-3; 1-5-9; 3-5-8; 5-8-13; 5-5-5) repeated measures ANOVA was conducted on the amount of forgetting that occurred between the final learning trial and the final test. There was significantly more forgetting in the delayed condition than the immediate condition, F (1, 94) =55.69, $p < .001$, *partial* $?^2$=.37. Forgetting also depended on the spacing condition, F (4, 91) = 19.28, $p < .001$, *partial* $?^2$=.46, but there was no interaction. Table 3 shows the amount of forgetting in each condition. Comparing the 5-8-13 condition to the 5-5-5 condition, there was less forgetting on the immediate test (although not significantly less), and significantly less forgetting on the delayed final test, $p = .001$.

Table 3: Forgetting as a function of spacing and delay.

Spacing	10 min.	48 hours
1-2-3	.18 (.03)	.45 (.03)
1-5-9	.13 (.09)	.44 (.09)
3-5-8	.08 (.03)	.27 (.03)
5-8-13	.02 (.03)	.18 (.03)
5-5-5	.07 (.03)	.30 (.03)

Note. Standard errors are in parentheses.

Discussion

Expanding retrieval can lead to better long term retention than evenly spaced retrieval when a slight modification is made to the original conceptualization of the method. Long term retention is best when the learning trials are constructed by combining a delayed initial retrieval attempt with an expanding sequence for the remainder of the learning trials. In accordance with Karpicke and Roediger (2007), increasing the number of intervening items between the study trial and the first retrieval attempt is critical to improve learning and memory of the information. Additionally, as the current experiment shows, it is also critical for the sequence to expand for optimal long-term retention.

Several findings from this experiment provide insight into the conditions under which expanding retrieval will produce better long term retention than evenly spaced retrieval. The first relevant finding is, as stated above, the first retrieval attempt after the study period must be sufficiently difficult. Both of the spacing conditions that only had a single item between study and retrieval produced high performance on the initial retrieval attempt, but those conditions also saw a large decline in performance across the learning trials. The 1-2-3 and 1-5-9 conditions, for example, both had a 10 percentage point drop in performance between the first and third retrieval attempts. The conditions that were more difficult and had at least three items between studying and the first retrieval attempt had no forgetting across the learning trials. Thus, increasing the difficultly of the first retrieval attempt produced steady performance throughout the learning trials. Maintaining a steady rate of performance throughout learning may be important in preventing rapid forgetting from occurring.

The next important finding was concerned with forgetting in each of the conditions. Forgetting, as Tables 1 and 3 show, was quite rapid in the 1-2-3 and 1-5-9 conditions both during learning (as performance decreased across learning trials) and during the final test retention intervals. Although there was no forgetting across learning trials for the sequences that began with a larger number of intervening items, performance on the third and final learning trial was still higher in the 1-2-3 and 1-5-9 conditions than the 3-5-8, 5-8-13 or 5-5-5 conditions. However, on the final test given 10 minutes later, the 5-8-13 condition produced numerically better performance than the other conditions, including the evenly spaced (5-5-5) condition. This result was due to almost no forgetting in the 5-8-13 condition. This trend continued, and the 5-8-13 condition produced significantly better performance than evenly spaced retrieval on the final test administered after a 48 hour retention interval. In this retention interval condition, the 5-8-13 sequence produced half the amount of forgetting as the 5-5-5 condition.

The 5-5-5- and 5-8-13 conditions were considered the critical comparison in this study and addressed the question of whether expanding retrieval can produce superior performance on a long term retention test compared to evenly spaced retrieval. On the learning trials, these two conditions produced similar levels of performance. After a 10 minute retention interval the 5-8-13 condition produced slightly better results, and after 48 hours, the 5-8-13 condition produced significantly better memory for the word pairs than the 5-5-5 condition.

Thus far, the question of whether expanding retrieval can produce superior performance compared to evenly spaced retrieval on immediate and long term retention tests has been mixed at best. In fact, equally spaced retrieval has generally produced better performance than expanding retrieval on tests taken at least 24 hours later. Karpicke and Roediger (2007) even noted that, "we know of no existing study using a continuous paired associate learning task…that has shown that expanding retrieval produces greater long-term retention (after delays greater than 24 hr) than equally spaced practice." The present experiment is one that does show that expanding retrieval produces greater long-term retention after 48 hours. Based on the current results, it is not entirely surprising that the previous research has been so mixed. Considering that a key factor in this experiment was the difficulty of the initial retrieval attempt, the 1-5-9 condition that has been used most widely is not a sequence that would be expected to improve long term retention. This also explains why the 5-5-5 condition, which begins with a difficult initial retrieval attempt, has produced such good performance in previous experiments.

The unique 5-8-13 expansion sequence, which combined a difficult first retrieval attempt with expanding retrieval, not only resulted in virtually no forgetting during both the

learning trials and on a test administered after a 10 minute retention interval. This particular sequence prevented the rapid forgetting that normally occurs immediately after learning takes place. Further, and of most importance to the question at hand, is that the combination of a difficult initial retrieval attempt and an expanding sequence resulted in half the amount of forgetting on the long-term retention test when compared to the evenly spaced control.

In summary, increasing the difficulty of the initial retrieval attempt protects against rapid forgetting that can occur within minutes of the study trial. This is evidenced by the small amount of forgetting during learning and on the 10 minute retention interval test in both the 5-5-5 and the 5-8-13 conditions. However, expanding retrieval from a difficult initial retrieval attempt produces better performance on a final test 48 hours. Combining a difficult initial retrieval attempt with even more difficult subsequent attempts reduces the forgetting that normally occurs immediately after learning (described by Ebbinghaus, 1885/1913; Rubin, Hinton & Wenzel, 1999) and up to 48 hours later.

Acknowledgments

This project was funded by Auburn University's College of Liberal Arts.

References

Bjork, R.A. (1994). Memory and metamemory considerations in the training of human beings. I J.Metcalfe & a. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185-205). Cambridge, MA: MIT Press.

Cepeda N., Pashler, H., Vul, E., Wixted, J., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin, 132,* 354-380

Cull, W., Shaughnessy, J., & Zechmeister, E. (1996). Expanding understanding of the expanding-pattern-of-retrieval mnemonic: Toward confidence in applicability. *Journal of Experimental Psychology: Applied, 2,* 365-378

Fritz, C., Morris, P., Nolan, D., & Singleton, J. (2007). Expanding retrieval practice: An effective aid to preschool children's learning. *The Quarterly Journal of Experimental Psychology, 60*(7), 991-1004.

Karpicke, J.D. & Roediger, H. L. (2007). Expanding retrieval practice promotes short-term retention, but equally spaced retrieval enhances long-term retention. *Journal of Experimental Psychology: Learning, Memory & cognition,* 33, 704-719.

Landauer, T.K. & Bjork, R.A. (1978). Optimum rehearsal patterns and name learning. In M.M. Gruneberg, P.E. Morris, & R.N. Sykes (Eds.), *Practical aspects of memory* (pp. 625-632). London: Academic Press.

Logan, J.M. & Balota, D.A. (2008). Expanded vs. equal interval spaced retrieval practice: Explorign different schedules of spacing and retention interval in younger and older adults. *Aging, Neuropsychology and Cognition, 15,* 257-280.

Morris, P., Fritz, C., Jackson, L., Nichol, E., & Roberts, E. (2005). Strategies for Learning Proper Names: Expanding Retrieval Practice, Meaning and Imagery. *Applied Cognitive Psychology, 19*(6), 779-798

Roediger, H., & Karpicke, J. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1,* 181-210.

Rubin, D., Hinton, S., & Wenzel, A. (1999). The precise time course of retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25,* 1161-1176.

Learning Functional and Causal Abstractions of Classroom Aquaria

Ashok K. Goel[1], Swaroop S. Vattam[1], Spencer Rugaber[1], David Joyner[1], Cindy E. Hmelo-Silver[2], Rebecca Jordan[3], Sameer Honwad[2], Steven Gray[3], Suparna Sinha[2]

[1]Design & Intelligence Laboratory, School of Interactive Computing, Georgia Institute of Technology, Atlanta, GA 30332.

[2]Department of Educational Psychology, Graduate School of Education, Rutgers University, New Brunswick, NJ 08901.

[3]School of Environmental and Biological Sciences, Department of Ecology, Evolution, and Natural Resources, and the Program in Science Learning, Rutgers University, New Brunswick, NJ 08901.

Abstract

Structure-Behavior-Function (SBF) models of complex systems use functions as abstractions to organize knowledge of structural components and causal processes in a system. We describe an interactive learning environment called ACT (Aquarium Construction Toolkit) for constructing simple SBF models of classroom aquaria, and report on a case study on the use of SBF thinking and the ACT tool in middle school science classes. We present initial data indicating that SBF thinking supported in part by the ACT tool leads to enhanced understanding of functions and behaviors of aquaria.

Keywords: Science education, Middle school science, Complex systems, Ecological systems, Functional models, Interactive learning.

Motivation and Goals

Understanding of complex systems enables important tasks such as monitoring, measurement, sensemaking, troubleshooting, explanation, prediction, diagnosis, redesign and design. Thus, understanding complex systems has been recognized as a key idea in science education in national science standards (National Research Council, 1996) as well as local standards (e.g., New Jersey Department of Education, 2006).

However, understanding complex systems is cognitively hard not only because of the large number of components and variables in a given system, but also because complex systems are dynamical and contain feedback loops (Forrester 1968) and exhibit hierarchical structure but are only nearly decomposable (Simon 1996); causal processes at one abstraction level in a complex system emerge out of interactions among components and processes at lower levels; and while some components of a complex system may be visible, many components, relations and processes typically are invisible. Thus, understanding complex systems challenges cognitive resources such as attention, memory and perception. The juxtaposition of understanding complex systems as an educational standard and the cognitive difficulty of understanding complex systems in turn poses a practical challenge for cognitive and learning sciences.

Theories of understanding complex systems in terms of functional models use functions as abstractions for organizing knowledge of structural components and causal processes (e.g., Chandrasekaran 1994a, 1994b; Kitamura et al. 2004; Rasmussen 1986). In Structure-Behavior-Function (SBF) models, for example, Structure refers to components of a complex system as well as connections among the components; Behaviors pertain to causal processes in the complex system; and Functions are abstractions of structural components and causal behaviors (Goel et al, 1996; Prabhakar & Goel, 1998; Goel, Rugaber & Vattam 2009). Representations of structural components and causal processes specify the functions they accomplish; representations of functions in turn act as indices into the components and processes that combine to accomplish them.

The SBF theory of understanding complex systems has led to lesson plans and interactive tools for learning about complex systems in science education. Our ongoing ACT project, for example, is an interactive learning environment that enables middle school children to construct and simulate SBF models of classroom aquaria (Vattam et al. 2010). An initial study indicates that teacher-led SBF thinking about aquaria, supported in part by use of ACT by small teams of students, led to significant improvement in understanding the basic structure, behaviors and functions of aquaria. However, we also found that in practice, middle school teachers and students did not use ACT the way we had planned. Instead of using ACT to construct and simulate full SBF models of aquaria, middle school students in our studies used the tool mainly to construct SBF graphical models of aquaria (Jordan et al. 2009).

In this paper, we report on a new study that utilizes a new version of the ACT interactive tool. The new version of ACT (ACT3) directly builds on our observations of SBF thinking practices in middle school science classrooms in the initial studies as well as feedback from the middle school teachers and students on the use of the previous version of ACT (ACT2). Preliminary results from new studies of SBF thinking about aquaria, stimulated, scaffolded and supported in part by the new ACT tool, appear to replicate the findings from the earlier studies with the new and more engaging tool.

The SBF Theory of Understanding of Complex Systems

Narayanan (2007) characterizes complex systems as follows: complex systems exhibit hierarchical structures composed of subsystems and components; subsystems and components exhibit natural behaviors or engineered functions; the subsystem/component behaviors causally influence other subsystems/components; the propagation of these causal influences creates chains of events in the operation of the overall system and gives rise to its overall behavior and function; and these chains of events extend in temporal and spatial dimensions. The origin of both Narayanan's characterization and our SBF models lies in Chandrasekaran's (1994a) Functional Representation (FR) scheme. Chandrasekaran (1994b) traces the development of FR; Goel, Rubager, Vattam (2009) describe the evolution of SBF from FR. Briefly, (1) the structure portion of an SBF model of a complex system specifies the "what" of the system, namely, the components of the system as well as the connections among them. (2) Behaviors specify the "how" of the complex system, namely, the causal processes occurring in the system. A behavior typically comprises of multiple states and transitions among them. The transitions are annotated by causal explanations for them. (3) Functions specify understanding of the "why" of the system. A function is a teleological interpretation of the components and processes in the system. (4) A component of a complex system can itself comprise a system and thus have its own SBF model. (5) The behavior of a system specifies the composition of the functional abstractions of its subsystems into the system functions.

Other researchers have described similar functional models of complex systems, e.g., Rasmussen (1986) and Kitamura et al. (2004). Although the various functional models differ in many features, they typically share some key characteristics, viz., explicit representation of function, use of functional representations to organize knowledge of causal behaviors and structural components, a hierarchical system-subsystem organization of knowledge, a view of causal behavior as an intermediate abstraction between structure and function, and domain-independent vocabularies for representing structure, behaviors and functions of complex systems. Erden et al. (2008) provide a recent survey of functional models of complex systems and their use in design.

Note that in the SBF theory of understanding complex systems, functions are mental abstractions, and thus are not intrinsic to the complex system. In case of designed systems, a functional abstraction corresponds to an intended output or observable behavior of a system, subsystem, or component. However, since functions are abstractions, we have also used the SBF theory to model natural systems including biological systems such as the human heart and ecological systems such as aquaria. Like designed systems, natural systems exhibit the types of causal processes and multiple levels of abstraction that characterize complex systems. We use function as a lens through which to view complex biological systems as well. For example, we may model a pond as being able to regulate the chemicals inside its water to maintain a livable environment for fish and plants. We may also specify the invisible causal process that achieves this self-regulation of the pond. In addition, we may state how this causal process combines functional abstractions of other processes and subsystems into the self-regulation function of the pond. In this functional representation of the pond, functional abstractions provide explanations for the relevance of specific subsystems in the context of a causal process.

Since SBF models explicitly represent functions, they differ fundamentally from causal models of complex systems (e.g., Chi 2005). The interactive tool called Betty's Brain (Biswas et al. 2005) is a good representative of the use of causal models in interactive learning because it too works in the same general domain (ecology) and targets the same general audience (middle school students). The innovation in the system lies in transforming the role of students into teachers of problem-solving software agents (Betty). This role transformation is motivational and engaging to middle school students. The models that students help Betty build, however, are causal graphs, with no mention of function and only implicit specification of structure. Although SBF models also represent behaviors in the form of causal graphs, the behavioral representations are grounded in the structure and indexed by their functional abstractions.

ACT: Interactive Construction of SBF Models

Empirical studies in the SBF framework show that while aquaria experts and hobbyists typically understand aquaria in terms of their structure, behavior and function, novices such as middle school students and pre-service teachers familiar with aquaria focus on the visible structure, show minimal understanding of function, and show little evidence of understanding the invisible causal behaviors (e.g., Hmelo-Silver, Marathe & Liu 2007). Thus, we developed a suite of interactive tools called RepTools that included SBF-inspired function-centered hypermedia (Liu & Hmelo-Silver 2009) as well as NetLogo simulations of aquaria generated by experts (Hmelo-Silver et al. 2007). Using the SBF coding scheme to analyze students' work on pre- and post- tests and the metrics for measuring SBF understanding of complex systems developed earlier (Hmelo, Holton & Kolodner 2000), we showed that the use of RepTools leads to deeper SBF understanding of complex systems in middle school science classrooms.

Although RepTools provided a useful learning environment, it did not provide a knowledge construction facility that allowed students to explicitly articulate their SBF understanding of complex systems. However, we know that scientists *construct* models of complex systems they seek to understand (Clement 2008; Nersessian 2008). From a constructivist perspective, much of learning entails active, social construction of knowledge (Palincsar 1998), and research on interactive learning increasingly emphasizes collaborative construction of external representations (Kozma 2000; Lajoie et al. 2001; Suthers 2006).

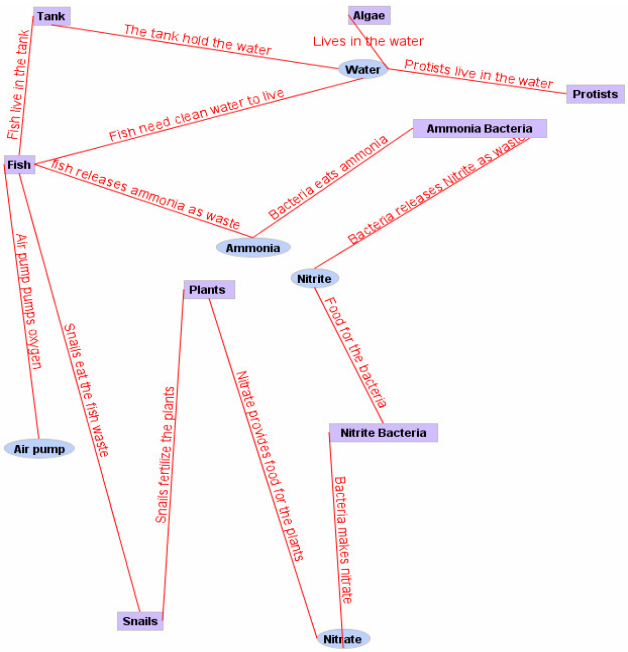


Figure 1: Model Graph of the nitrification process designed by a 7th grade student using ACT3.

Thus, we developed an interactive learning environment called ACT that provided a tool (called SBFAuthor) for constructing SBF models of classroom aquaria in middle school science (Vattam et al. 2010). In order to adapt the SBF modeling language to serve as an effective modeling tool for learners, we augmented it with a visual syntax to obtain vSBF: a visual SBF modeling language. Creating an SBF model of a particular complex system in vSBF now becomes an exercise in drawing an annotated, flowchart-like diagram of the system using the modeling primitives provided by the language. ACT also integrated SBFAuthor with the Netlogo simulation platform (Wilensky 1999; Wilensky & Resnick 1999). In addition, ACT provided access to extant RepTools. The goal was to encourage middle school students to understand complex systems in terms of functional abstractions and casual behaviors. The intended method was teacher-led SBF thinking supported by the use of ACT for construction, simulation and revision of SBF models of classroom aquaria.

In an initial study conducted in 2008, we introduced the original ACT tool (ACT2) into three middle school classrooms consisting of one hundred and fifty seven students (Jordan et al. 2009). One example of SBF thinking used by the three middle school teachers in the initial study pertained to the nitrification process. The nitrification process is the process by which an aquarium cleans itself of waste that is poisonous to fish. Fish release ammonia in their waste, a highly poisonous chemical; nitrosomonas consume this ammonia and output nitrite, while nitrobacters eat this nitrite and release nitrate. Nitrate, though still poisonous to fish in large quantities, is much less dangerous than ammonia. In this example, the structural components in the system are the fish and bacteria. These components serve certain functions; for example, one function of the nitrosomonas is to clean the water of harmful ammonia and provide food for nitrobacters. Of course, this function is merely our teleological interpretation of this action of nitrosomonas, since (insofar as we know) the bacteria do not intentionally set out to serve a purpose to the fish. The behavior by which these nitrosomonas accomplish cleaning is through a natural ingestion/output behavior. In this example, it is also possible to see how SBF models may examine systems at multiple levels of abstraction. One could state that the aquarium as a whole serves the function of cleaning itself, and the behavior by which it accomplishes this is the nitrification process. One can also imagine how a similar analysis could be applied to *how* bacteria eats one chemical and outputs another.

Our initial study indicated that teacher-led SBF thinking, supported in part by use of the ACT tool, led to statistically significant improvement in understanding of classroom aquaria as a complex system (Vattam et al. 2010). The finding appeared robust in that it was independent of the teaching styles of the three middle school teachers in the initial study. We also found the middle school students in our initial study did not use the ACT tool as we had intended. Instead of using ACT to construct and simulate SBF models of the nitrification process described above, middle school students in our studies used the tool mainly to construct simple SBF graphical models of the process (Jordan et al 2009). This may have been in part because the 1-week and 2-week science units in which the ACT tool was used were too short for students to become familiar enough with SBF thinking as well as the ACT tool to construct and simulate SBF models of the nitrification process. It may also partially be due to difficulty in understanding the notions of states and transitions between states. Detailed feedback from some middle school teachers suggested the need for SBF tables that list the structural components, causal behaviors, and their functional abstractions.

The New ACT: Simplification of SBF Models

Given our observations of the practice of SBF thinking and learning in the initial study, as well as the feedback from middle school teachers and students in the study, we redesigned the ACT interactive environment. The new ACT environment (ACT3) supports two tools: SBFAuthor and RepTools. Further, SBFAuthor enables the construction of simple, partial, single-level SBF models through a Model Graph tool and Model Table tool that work in conjunction with each other. These can be seen in Figures 1 and 2.

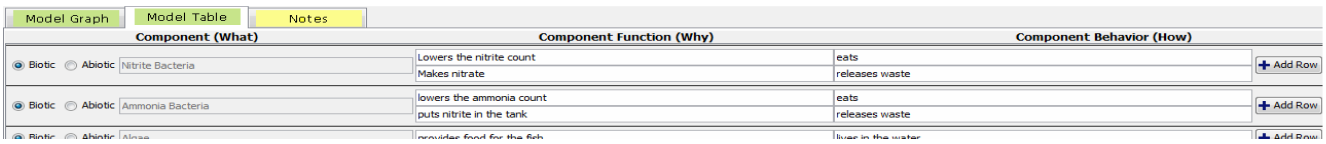


Figure 2: The Model Table derived from the previously shown Model Graph.

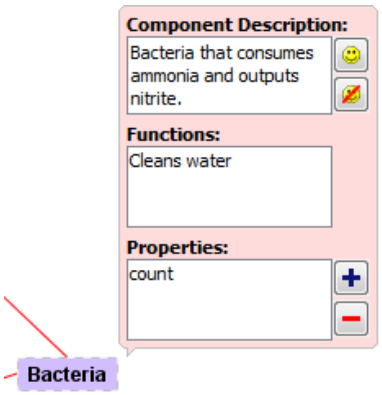


Figure 3: Dialog for adding details to the structure in the model graph. Note the specification of the function of the component.

Model Graph: The Model Graph enables users to create the structural portion of an SBF model in terms of its structures (components and substances) and their associated connections; Figure 1 depicts a Model Graph actually constructed by a student in the classroom. The structure model is presented as a graph. For each component or substance in the structure, a corresponding node is created. Nodes are linked together by behaviors, which are represented by lines drawn between nodes. Functions of the structures and behaviors are added using a dialog window (see Figure 3), as well as the Model Table (see Figure 2). In this way, students can define and connect structures, behaviors and functions in an externalized view, which helps guide them toward a more expert-like understanding. Most importantly, this allows students to explicitly define the functions of the system in order to better understand how larger processes emerge from underlying functions.

Model Table: The Model Table is an organizational tool intended to allow students to engage in their natural thought process when first encountering a complex system. An example can be seen in Figure 2. The Model Table features three columns: one for Structure, one for Behavior, and one for Function. Structures are linked to Behaviors in a one-to-many association, while Behaviors are linked to Functions in a one-to-one association. The Model Table is more than a preliminary brainstorming tool, however. Adding structures to the Model Table will automatically result in their creation on the Model Graph. Behaviors and Functions appear in the Model Graph after their addition to the Model Table, through the Structure's pop-up dialog menu. The control works both ways: new Structures, Behaviors and Functions added on the Model Graph automatically appear on the Model Table.

RepTools: ACT also links to the extant RepTools. RepTools was designed to accompany a physical aquarium installed in each classroom. It provides digital tools that feature function-centered hypermedia from which students can read about the structures, behaviors, and functions occurring within an aquarium system (Liu & Hmelo-Silver 2009). It also includes a micro and macro-level NetLogo-based simulations (Wilensky 1999) developed by experts. The macro-level simulation enables students to test ideas about fish spawning and water quality, and the micro-level simulates the nitrification process that occurs within an aquarium as part of its biological filtration (Hmelo-Silver et al. 2007). In combination, these digital tools allow students to not only test ideas about the aquarium system but also gain insight into the explanations behind the processes and outcomes that occur at multiple levels within the aquarium.

Methods

Setting

Overall, two hundred and seventy three (273) students participated in this 2009 study from four middle schools classrooms in central New Jersey - three from seventh grade and one from the eighth grade. Their science teachers integrated this unit as a part of their regular science instruction. Prior to beginning the study, none of the students were taught to use SBF as a representational tool for complex systems. All four teachers attended an evening workshop where they were introduced to these digital tools prior to implementation in the classroom. The curriculum unit lasted from one to two weeks.

Besides the eighth grade classroom, none of the other classes had a physical model of the aquatic ecosystem (aquarium) as a part of their classroom environment. In order to prepare for the unit, the researchers set up aquariums in the remaining three seventh grade classrooms. Students used the digital tools (ACT, SBFAuthor, RepTools) on laptops while working in small groups, which varied from 2 to 6 students per computer, to generate models for analysis in this study.

Classroom Instruction

The four science teachers appropriated the curriculum and implemented it based on their individual scientific knowledge and learning styles of their students. While all the teachers used the SBF as a representational tool to organize their thinking about complex systems, there were variations within actual implementations of the curriculum.

SBF Introduction: Two teachers decided to begin the instruction with a discussion on the aquarium and focus on SBF as an initial activity using the ACT Model Table. The other teachers adopted the reverse strategy. Their introduction to the unit began with description of the SBF while illustrating it from students' immediate environment (for e.g. the classroom as a complex system). This top down effect was intended for the students to think about the SBF from a micro to macro level.

Modeling Aquatic Ecosystem: While some teachers emphasized the importance of the models as a means to represent ideas in summative fashion, other teachers chose to use the modeling task throughout implementation as a means to continually formulate and refine ideas. Additionally, some

teachers chose to have students model the entire system, while other teachers had students generate a model based on a portion of the system that corresponded quite closely to one of the NetLogo simulations.

Figure 1 illustrates a model graph created in ACT by a 7th grade student as part of an SBF model construction activity in one of the middle school classrooms. This figure shows the one of the systems frequently modeled in the classrooms: the nitrification process, described previously. Structures are shown as nodes (purple for biotic structures, blue for abiotic structures), while behaviors link together structures that directly and relevantly influence one another. Although not depicted in the figure, inside the structure boxes are statements about a component's function as indicated in the dialog box of Figure 3; these functions can also be seen in the Model Table in Figure 2. In this way again, students are encouraged to recognize and explicitly state the functions of the system, reinforcing a functional understanding.

Results

To assess the effectiveness of the SBF-driven curriculum and technology, identical tests were administered before and after engagement in the aquarium unit. These tests asked about the structures, behaviors and functions of the aquaria, and were also given problems to solve regarding aquarium processes. To examine learning with respect to SBF, we coded the pre- and post- tests using an SBF coding scheme (Hmelo, Holton & Kolodner 2000). Structural components, such as fish, plants, filter, was coded as structure. A reference to the mechanisms of how the components worked was coded as behavior. For example, a behavior of the plants could be absorb some of the carbon dioxide in the fish tank and produce oxygen through photosynthesis. Reference to the outcome of a behavior was coded as function. For example, a function of the filter could be to clean and circulate water. All tests were coded blind to condition by one rater.

Table 1: Pre- Posttest Results

	Structure	Behavior	Function
Pretest Mean (SD)	8.08 (2.624)	3.80 (2.107)	4.78 (2.924)
Posttest Mean (SD)	9.33 (2.347)	6.20 (2.766)	8.12 (3.241)
t(273)	5.60*	11.65*	12.55*
Effect size	0.24	0.44	0.47

**All p<.05*

In this preliminary study, the objective was to ensure that the SBF curriculum described here is successfully increasing understanding of functions and behaviors. Since students already are generally familiar with the structure of aquaria, increases in understanding of structure are considered a baseline for comparison of how the curriculum enhances understanding of functions and behaviors. Table 1 shows initial results from the pre- and post- tests collapsed across the four middle school classrooms consisting of 273 students. The first number in the first two rows refers to the Mean and the second number in parentheses to the Standard Deviation. As indicated by the effect sizes, gains in structural understanding were modest, while we saw the greatest effect size for increase in behavioral (or causal) and functional understanding for all groups. These tests suggest that the SBF-driven curriculum and the ACT technology effectively increase understanding in terms of the deeper concepts of functions and behaviors. Thus, these results replicate the findings from our initial study. A sibling paper (Honwad et al. 2010) that too appears in these proceedings focuses on the use of RepTools in the ACT learning environment and reports on more recent data collected in 2010.

Conclusions & Open Issues

Functional models use functions as abstractions to organize knowledge of complex systems. We are pursuing a research program that investigates the use of Structure-Behavior-Function modeling for helping middle school children understand complex systems such as classroom aquaria (Hmelo-Silver et al. 2008; Jordan et al. 2009, Vattam et al. 2010). In this paper we described a new version of an interactive tool called ACT that enables middle school children to author simple SBF models of complex processes such as the nitrification process that results in self-cleansing in aquaria. We also described teacher-led SBF thinking in multiple classrooms supported in part by use of the ACT tool by small teams of middle school children. Preliminary results from the SBF-driven science curriculum in this study indicate significant improvement in understanding of the basic structure, behaviors and functions of aquaria. These results appear to confirm initial results from earlier studies.

Of course, there remain many open issues, including the following three. Firstly, now that we have experimentally affirmed that the SBF curriculum and ACT technology is effective in learning about functions and behaviors of aquaria, there is a need to conduct controlled experiments. In particular, there is a need for finer analysis of the effectiveness of SBF thinking and the ACT tool based experiments featuring many conditions, such as curriculum without software and software without curriculum. Secondly, there is a need to determine whether the improved understanding of the functions and behaviors of aquaria is enabling improved reasoning about tasks such as establishment and maintenance of aquaria. Thirdly, there is growing evidence that middle school teachers on their own are appropriating SBF meta-models and transferring them to other complex systems such as the human digestive system (Hmelo-Silver et al. 2010). There is a need to determine if middle school children too are appropriating and transferring SBF meta-models to other complex systems.

Acknowledgments

This research has been supported by NSF ALT Grant (# 0632519) on "Learning About Complex Systems in Middle School by Constructing Structure-Behavior-Function Models." Vivek Menon contributed to the development of ACT3.

References

Biswas, G., Leelawong, K., Schwartz, D., & Vye, N. (2005) Learning By Teaching: A New Agent Paradigm For Educational Software. Applied Artificial Intelligence 19(3-4): 363-392.

Chandrasekaean, B. (1994a) Functional Representations and Causal Processes. In M. Yovits (editor): Advances in Computers, pp. 73-143.

Chandrasekaran, B. (1994b). Functional Representation: A Brief Historical Perspective. *Applied Artificial Intelligence*, 8(2): 173-197.

Chi, M. (2005) Commonsense Conceptions of Emergent Processes: Why Some Misconceptions are Robust. *Journal of the Learning Sciences*, 14: 161-199.

Clement, J. (2008). *Creative Model Construction in Scientists and Students: The Role of Imagery, Analogy, and Mental Simulation.* Dordrecht: Springer.

Erden, M., Komoto, H., van Beek, T., D'Amelio, V., Echavarria, E., & Tomiyama, T. (2008) A Review of Function Modeling: Approaches and Applications, *AI for Engineering Design, Analysis and Manufacturing,* 22 (2): 147-169.

Forrester, J. (1968) *Principles of Systems*, Productivity Press, 2nd Edition.

Goel, A., Gomez, A., Grue, N., Murdock, W., Recker, M., & Govindaraj, T. (1996) Towards Design Learning Environments - Explaining How Devices Work. *In Proc. International Conference on Intelligent Tutoring Systems*, Montreal, Canada, June 1996.

Goel, A., Rugaber, S., & Vattam, S. (2009) Structure, Behavior & Function of Complex Systems: The SBF Modeling Language. *AI for Engineering Design, Analysis and Manufacturing,* 23: 23-35.

Hmelo, C. E., Holton, D., Kolodner, J. L. (2000). Designing to learn about complex systems. *Journal of the Learning Sciences*, 9, 247–298.

Hmelo-Silver, C. E. Liu, L., Gray, S., Finkelstein, H., & Schwartz, R. (2007). Enacting things differently: Using NetLogo models to learn about complex systems. Presented to biennial meeting of *European Association for Research on Learning and Instruction*. Budapest, Hungary.

Hmelo-Silver, C., Jordan, R., Demeter, M., Gray, S., Liu, L., Vattam, S., Rugaber, S., & Goel, A. (2008). Focusing on Function: Thinking Below the Surface of Complex Natural Systems. *Science Scope* 31(9): 27-35, Summer 2008, NSTA.

Hmelo-Silver, C., Marathe, S., Liu, L. (2007) Fish Swim, Rocks Sit and Lungs Breathe: Expert-Novice Understanding of Complex Systems. *Journal of the Learning Sciences.* Routledge.

Hmelo-Silver, C., Sinha, S., Gray, S., Jordan, R., Honwad, S., Rugaber, S., Vattam, S., Goel, A., Ford, W., & Schmidt, C. (2010) Appropriating Conceptual Representations: A Case of Transfer of in Middle School Science Teacher. Presented to the *Annual Conference of the National Association for Research in Science Teaching,* Philadelphia, March 2010.

Honwad, S., Hmelo-Silver, C., Jordan, R., Eberbach, C., Gray, S., Sinha, S., Goel, A., Vattam, S., Rugaber, S., & Joyner, D. (2010) Connecting the Visible to the Invisible: Helping Middle School Children Understand Complex Ecosystem Processes. In *Proc. 32nd Annual Meeting of the Cognitive Science Society*, Portland, Oregon, August 2010.

Jordan, R., Hmelo-Silver, C., Gray, S., Goel, A., & Rugaber, S. (2009) Modeling Practices as a Function of Task Structure. Presented to *Annual Meeting of the American Educational Research Association*, San Diego, California, April 2009.

Kitamura, Y., Kashiwase, M., Fuse, M., & Mizoguchi, R. (2004). Deployment of an Ontological Framework for Functional Design Knowledge. *Advanced Engineering Informatics*, 18(2).

Kozma, R. B. (2000). The use of multiple representations and the social construction of understanding in chemistry. In M. J. Jacobsen & R. B. Kozma (Eds.), *Innovations in Science and Mathematics Education (pp. 11-46).* Mahwah NJ: Erlbaum.

Lajoie, S. P., Lavigne, N. C., Guerrera, C., & Munsie, S. D. (2001). Constructing knowledge in the context of Bio World. *Instructional Science, 29*, 155-186.

Liu, L., & Hmelo-Silver, C. E. (2009). Promoting complex systems learning through the use of conceptual representations in hypermedia. *Journal of Research in Science Teaching*, *46*, 1023-1040.

Narayanan, N. H. (2007). The impact of cognitively based design of expository multimedia. In D. Alamargot, P. Terrier & J.M. Cellier (Eds.), *Written Documents in the Workplace*, Elsevier Science Publishers, pp. 243-260.

National Research Council (NRC). 1996. *National science education standards*. Washington, DC: National Academy Press.

Nersessian, N.J. (2008) *Creating Scientific Concepts*. Cambridge, MA: MIT Press.

New Jersey Department of Education (2006). *New Jersey Core Curriculum Standards for Science*. Retreived June 19, 2008.

Palincsar, A. (1998) Social constructivist perspectives on teaching and learning. *Annual Review of Psychology.*

Prabhakar, S., & Goel, A. (1998) Functional Modeling for Enabling Adaptive Design of Devices for New Environments. *Artificial Intelligence in Engineering,* 12:417-444.

Rasmussen, J. (1986) Information Processing and Human-Machine Interaction. North-Holland.

Suthers, D. (2006). Technology affordances for intersubjective meaning making. *International Journal of Computer Supported Collaborative Learning, 1*, 315-337.

Vattam, S., Goel, A., Rugaber, S., Hmelo-Solver, C., Jordan, R., Gray, S., Sinha, S. (2010). Understanding Complex Natural Systems by Articulating Structure-Behavior-Function Models. To appear in *Educational Technology & Society*,

Wilensky, U. (1999). NetLogo. http://ccl.northwestern.edu/netlogo/.

Wilensky, U., & Resnick, M. (1999). Thinking in levels: A dynamic systems approach to making sense of the world. *Journal of Science Education and Technology,* 8, 3-19.

The Effects of Work Shift and Strategy on an Orientation Task

Tim Halverson (thalverson@gmail.com)
Oak Ridge Institute for Science and Education
Air Force Research Laboratory
Mesa, AZ 85212 USA

Glenn Gunzelmann (glenn.gunzelmann@mesa.afmc.af.mil)
Air Force Research Laboratory
Mesa, AZ 85212 USA

L. Richard Moore Jr. (larry.moore@mesa.afmc.af.mil)
Lockheed Martin
Air Force Research Laboratory
Mesa, AZ 85212 USA

Hans P.A. Van Dongen (hvd@wsu.edu)
Sleep and Performance Research Center, Washington State University
Spokane, WA 99210 USA

Abstract

Cognitive alertness decreases at night due to circadian rhythms with adverse effects on performance across domains and tasks, including real-world tasks like driving and flying. Additionally, the strategy used on a task may have a substantial effect on performance. However, little is known about whether and how circadian rhythms and strategy interact to affect performance. The current study investigates participants' performance on an orientation task performed over a period of two weeks. Participants were assigned to simulated day or night shift conditions, and were trained to use one of two strategies for the orientation task. The results indicated that shift condition had little impact on a more declarative strategy for the task, but had a significant impact on a more spatial strategy. The results illustrate how different aspects of cognitive functioning may be affected differently by circadian rhythms, and point to some important implications for training and task performance in real-world contexts.

Keywords: spatial; sleep; circadian rhythm; fatigue; learning; shift work

Introduction

Critical, safety-sensitive activities, such as driving and air traffic control, are performed at all times of the day and night. Yet, it is not well understood how nighttime operations affect task performance in contexts such as these. Most research on night and shift work has focused on how shift differences affect sleep and frequency of accidents (e.g., Åkerstedt, 1988). Little work has focused on how shift work and task differences affect different cognitive processes alone or in interaction.

Variations in alertness due to circadian rhythms and sleep loss have been shown to affect various components of cognitive functioning (Jackson & Van Dongen, in press). For example, vigilant attention (Lim & Dinges, 2008), perceptual learning (Mednick, Nakayama & Stickgold, 2003), and motor learning (Walker, Brakefield, Morgan, Hobson & Stickgold, 2003) are all affected by fluctuations in alertness associated with time awake and circadian rhythms.

For shift work, circadian rhythms are particularly important. Circadian rhythms are driven by a biological clock in the suprachiasmatic nuclei of the hypothalamus, which imposes cyclical changes in alertness throughout the day, leading to increased pressure for sleep at night. This leads to nocturnal degradations in cognitive performance (Van Dongen & Dinges, 2005), as demonstrated in a variety of tasks and domains (e.g. Caldwell, 2003; Dinges, 1995).

The present research investigates how strategies recruiting different cognitive-perceptual processes may be differentially affected by fluctuations in alertness resulting from circadian rhythms in laboratory-simulated shift work. This is accomplished within the context of a spatial direction task, where distinct alternative cognitive strategies have been identified (Gunzelmann, Anderson & Douglass, 2004). In this task, participants are presented with two views of a set of objects (Figure 1). One of the views (the left side in Figure 1) is an overhead, ego-oriented perspective, based on a viewpoint at the bottom of the screen. Within the ego-oriented view, one of the objects (small circles) in each trial is filled in to identify it as a target. The other view (the right side in Figure 1) shows a map-like perspective with the viewpoint indicated by the arrow, which may be misaligned relative to the ego-oriented view on the left. The task requires participants to identify the location of the target in the map-like perspective.

In the study described here, participants were taught to use one of two strategies for the spatial direction task: one based on counting and the other on mental rotation, as in Gunzelmann et al. (2004). The strategies are described in more detail below. The key feature is that the strategies emphasize different cognitive functions, declarative and

spatial, and lead to reliably different performance in participants trained to use them.

The alternative strategies for the spatial direction task offer an opportunity to explore how different cognitive capabilities may vary in their susceptibility to fluctuations in alertness. Such variations can be important in naturalistic contexts, where a variety of strategies may be available. To address this issue in the context of a common situation, we compare performance on the spatial direction task between individuals placed on a simulated night shift schedule for two weeks versus individuals sleeping according to a simulated day shift schedule.

Method

This experiment was conducted as part of a larger study to understand how circadian rhythms and sleep disruption affect performance in a variety of domains.

Participants

Twenty-six individuals, 14 female and 12 male, ranging in age from 22 to 39 years old (mean = 27), from the general community of Spokane, Washington participated in the experiment. The participants were screened to be healthy and without sleep disorders, with no evidence of brain damage or learning disabilities, and free of drugs of abuse. Participants gave written informed consent, and were paid for their participation.

Stimuli

Participants completed the task shown in Figure 1. There are 8 possible target locations and 8 possible misalignments (45 degree intervals). However, performance is roughly equivalent for right-left mirrored stimuli (see Gunzelmann et al., 2004). For instance, response times for targets located in the lower-left and lower-right positions are similar for a given misalignment. Likewise, response times are similar for misalignments that differ only in the rotation direction, such as assumed perspectives at positions 4 versus 6 on the map. Because of this correspondence, participants were presented with only one of these trials in each session. There were therefore 25 trials per session — 5 target locations (bottom, near, middle, far, and top) crossed with 5 misalignments (0, 45, 90, 135, and 180 degrees) — which were presented in random order.

Participants responded using the numeric keypad portion of a computer keyboard, which was spatially mapped to the possible response locations on the map view. So, if the correct response was the bottom position on the map (as it is in the sample trial shown in Figure 1), participants responded by pressing the "2" on the numeric keypad.

Procedure

Participants were in the laboratory for fourteen consecutive days. The first day was a baseline day with 10 hours in bed for sleep (22:00–08:00). Subsequently, some of the participants (n = 12) changed to a simulated night shift. Night shift participants were given five hours in bed (15:00–20:00) on the second baseline day, before starting five consecutive work days with 10 hours in bed during the daytime (10:00–20:00) on each day. On the seventh and eighth day, night shift participants had a simulated weekend during which they had 5 hours in bed (10:00–15:00), 7 hours awake, 10 hours in bed during the night (22:00–08:00), 7 hours awake, and then 5 hours in bed (15:00–20:00) before resuming their night shift schedule for the next 5 days. This schedule represented a stereotypical schedule for individuals working a night shift, who frequently shift back to a nighttime sleep schedule during weekends. After the last night shift day, night shift participants received 5 hours in bed (10:00–15:00), 7 hours awake, and then, on the final day of the study,

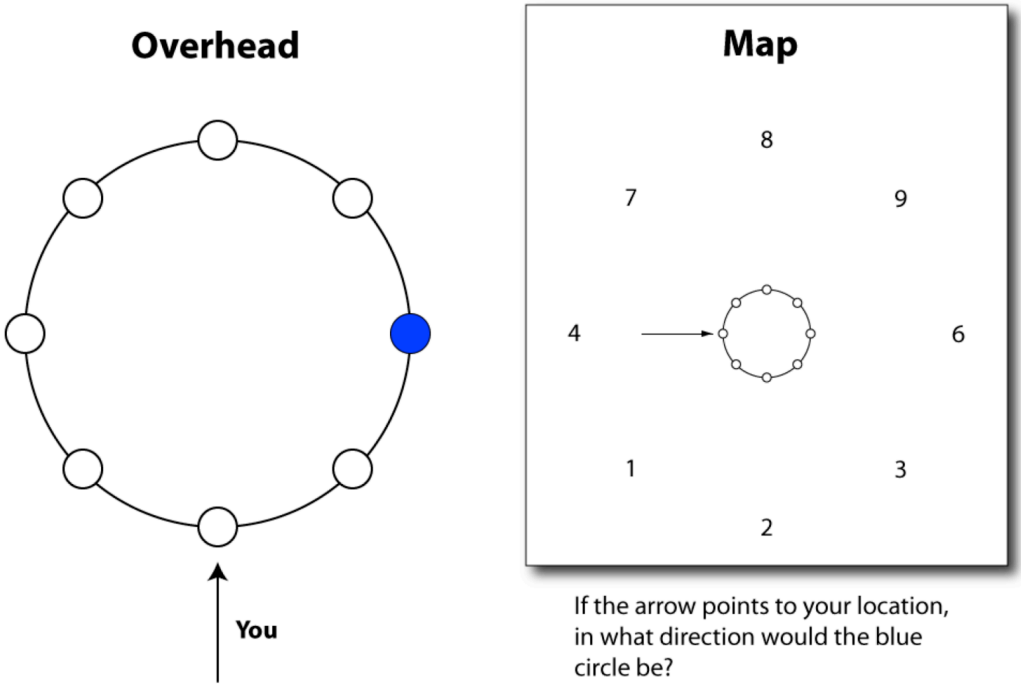


Figure 1: An example trial. The target on the overhead ego-oriented view (left side), indicated by the filled circle, is at middle distance to the right of center. The perspective on the map view (right side), indicated by the arrow, is misaligned by 90° clockwise. The correct response in this example trial is "2."

were given 10 hours in bed (22:00–08:00) for recovery.

Participants on the day shift (n = 14) maintained the same sleep schedule throughout the study, with 10 hours in bed (22:00–08:00) each night. Note that participants on the day shift and night shift schedules were given the same amount of time in bed over the course of the experiment, although it was distributed differently.

Participants completed fifty-one test sessions of the spatial direction task over the fourteen consecutive days, with 2 to 4 sessions per day. On the first baseline day, participants completed three sessions; on the second baseline day, they completed two sessions. On each of the remaining days of the study, participants completed four sessions up until the last day when they completed two sessions.

Before the first session, participants were presented with instructions for the task, including training for either the rotation (n = 13) or counting (n = 13) strategy for which they completed four practice sessions. Training on the rotation strategy encouraged the participants to mentally rotate the relative positions of the viewpoint and the target on the overhead view (left side) to align them with the viewpoint indicated on the map view (right side). Specifically, they were taught to imagine an angle that connects the viewpoint (indicated by the "You" arrow) to the target on the overhead view, with the vertex at the center of the field (a 90 degree angle in Figure 1). They were then told to mentally shift to the map view, and to rotate the angle so that the arrow in the overhead view was aligned with the arrow in the map view (a rotation of 90 degrees clockwise in the trial shown in Figure 1). At this point, the answer could be determined by finding the target end of the angle.

Training on the counting strategy taught the participants to count the number of objects from the arrow at the bottom of the ego-oriented view to the target position (the count is 2 in Figure 1) and note the direction in which the target was located (counterclockwise in Figure 1). They were then told to count the same number of steps around the map view in the appropriate direction from the location indicated by the smaller arrow.

Results

The analyses focused on how the study condition (night shift versus day shift) interacted with the trained task strategy to affect performance. Previous research using this task has shown that some people use special-case strategies when the target is at the top ("across from where I am") or bottom ("where I am") of the ego-oriented view (Gunzelmann et al., 2004). In order to ensure that the analysis truly reflected differences in the use of the counting and rotation strategies, these special cases were removed from the analysis. Additionally, we only included data in the analysis for sessions when the sleep schedules were different for the two groups (i.e., when the night shift group was up at night), that is, days 3 to 7 and days 9 to 13.

Linear mixed-effect models were used for the analysis, using the R environment (R Development Core Team, 2009) with the nlme package (Pinheiro, Bates, DebRoy, Sarkar & the R Core Team, 2009). The skewed distribution of the response time data was corrected using an inverse square root. An alpha level of .05 was used for all statistical tests.

The analysis concentrated on the effects of the *strategy* that the participant was taught (rotation or count), the work *shift* of the participant (day or night), the *day* of participation, the location of the *target* (near, mid, and far), and *misalignment* between camera and target view (0°, 45°, 90°, 135°, and 180°). These were all included in the nlme analysis as multi-level factors, except for day, which was continuous. Participant was used as a repeated-measure grouping factor, and intercept, target and misalignment were included as random factors.

Table 1 shows the mean response times by strategy and shift. Neither the strategy, $F(1, 22) = 0.05$, $p = .83$, nor the shift, $F(1, 22) = 0.47$, $p = .50$, displayed a simple main effect on response time. As seen in Figure 2, participants performed better in later days, $F(1, 15458) = 2{,}300$, $p < .001$, reflecting a learning curve. As seen in Figure 3, targets located further away required more time, $F(2, 15458) = 81$, $p < .001$, and larger misalignments also required more time, $F(4, 15458) = 150$, $p < .001$. Additionally, misalignment had a larger effect when targets were further away, $F(8, 15458) = 26$, $p < .001$.

Performance improved more as time progressed for participants using the rotation strategy than for participants using the count strategy, $F(1, 15458) = 11$, $p < .001$. Performance of participants on the day shift improved faster than that of participants on the night shift, $F(1, 15458) = 21$, $p < .001$. Figure 2 shows the interaction of strategy, shift, and day, which was significant, $F(1, 15458) = 15$, $p = .008$. Up until day six, participants using the rotation strategy were performing worse, no matter which shift they worked, than those using the counting strategy. Later, participants using the rotation strategy on the night shift eventually reached the performance level of those using the count strategy, and participants using the rotation strategy on the day shift outperformed the other groups.

Observed error rates were low ($M = 4\%$, $SD = 3\%$). The error rates tended to correlate with the response time ($r^2 = 0.58$), suggesting that the between-group differences did not stem from a speed-accuracy trade-off.

An analysis of the baseline data alone was conducted to explore the possible influence of differences among the groups at the start on the observed effects. Importantly, neither the strategy, $F(1, 22) < 0.01$, $p = .99$, the shift, $F(1,$

Table 1: Mean (SD) response times (ms) by strategy and shift.

		Shift	
		Day	Night
Strategy	Counting	2016 (802)	2113 (945)
	Rotation	2015 (1033)	2210 (1041)

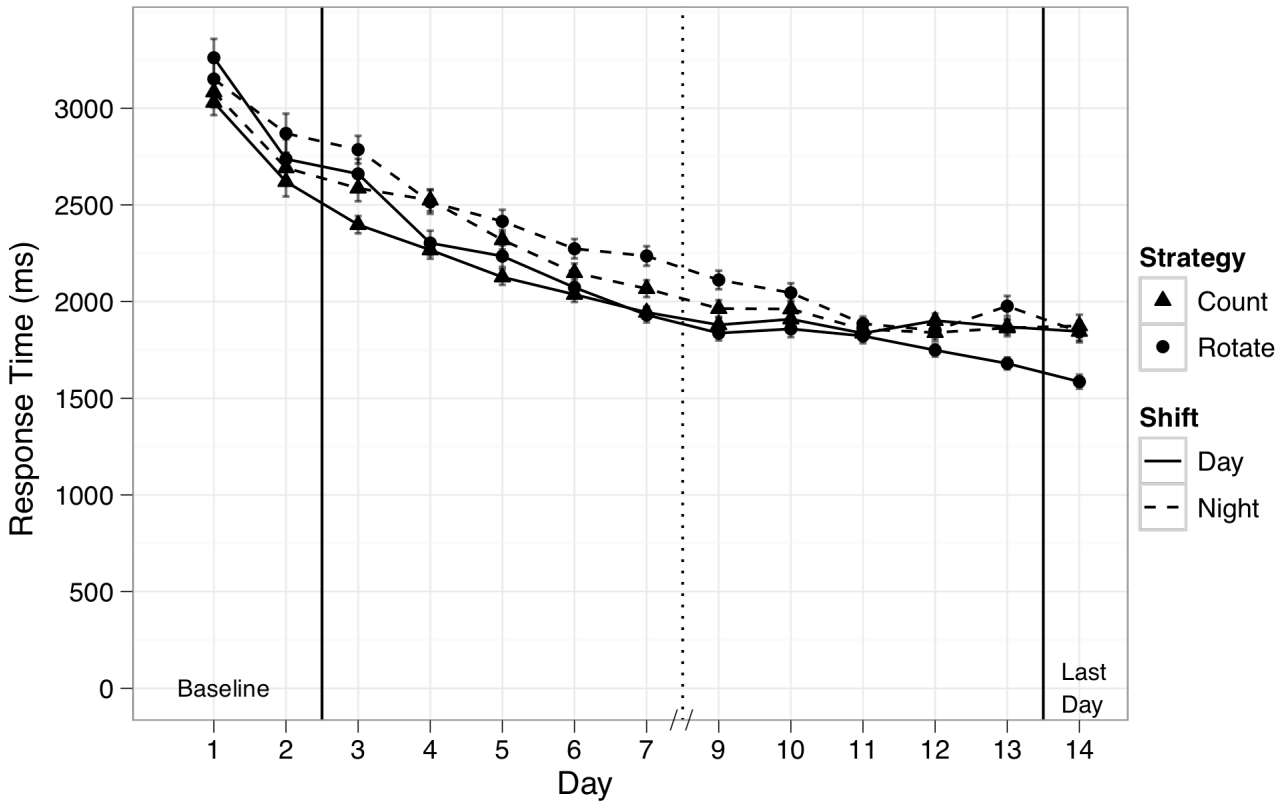


Figure 2. Reaction time as a function of strategy trained, work shift, and day in study. Data from days 1 and 2 (baseline) prior to work shift and day 14 (last day) after work shift are shown for reference, but were not included in the primary analysis. The sleep schedule was interrupted by a simulated weekend on day eight (dotted line), which was not included in the analysis or shown here. Error bars indicate ±1 standard error.

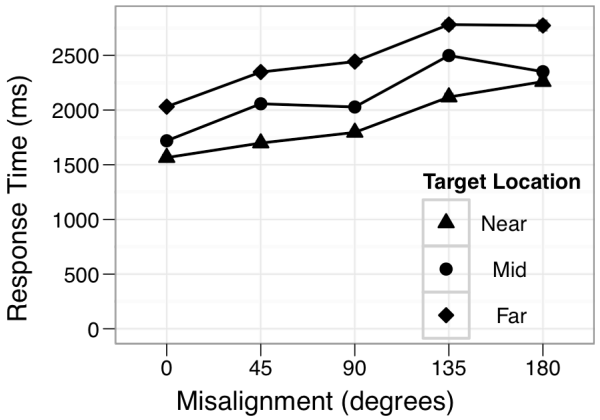


Figure 3. Response time as a function of misalignment and target location. Error bars are too small to be visible.

22) = 0.12, p = .73, nor their interaction, F(1, 22) = 0.04, p = .85, were significant, indicating that the groups were roughly equivalent in their performance at the start.

Discussion

All of the participants gained extensive expertise in the task by performing the task multiple times per day over a two-week period. Still, the strategy the participants trained on and the work shift to which they were assigned had a significant impact on performance.

During the first two days of the experiment (i.e., baseline days), performance was not significantly different across work shift conditions, which supports the conclusion that differences seen in the subsequent weeks were real and not a result of selection bias. Differences seen in the baseline condition with respect to strategy are consistent with previous research using this task (Gunzelmann et al., 2004). As seen in Figure 2, participants trained to use the counting strategy initially performed slightly better (but not significantly better) than participants trained to use the rotation strategy, and, as shown in Figure 3, misalignment angle and target location interacted, both of which are in line with those previous results.

Participants on the night shift tended to perform worse than those on the day shift. Previous research has shown that performance on a variety of tasks tends to be worse at night (e.g., Van Dongen & Dinges, 2005) and a number of commercial and industrial disasters have been attributed to degraded cognitive functioning associated with such shift work (Caldwell, 2003; Dinges, 1995). Further, within each shift condition, participants using the rotation strategy tended to perform worse than those using the counting strategy. As with the baseline data, this was expected, as it is consistent with previous research (Gunzelmann et al.,

2004). However, the results also suggest that, although initially more difficult, the rotation strategy may be a more efficient approach to the task by the end of the experiment (at least in the day shift condition).

Asymptotic performance appears to have been reached earlier when the counting strategy was used. Further, asymptotic performance appears to have been the same for day and night shift when the counting strategy was used. When the rotation strategy was used, the rate of performance improvement was reduced. However, on the night shift, performance using the rotation strategy was eventually equivalent to performance using the counting strategy. Moreover, on the day shift, performance with the rotation strategy continued to improve through the end of the protocol, and was eventually better than the performance in all other conditions. These results suggest that (a) learning occurs faster for the counting strategy than for the rotation strategy, (b) the task is learned equally well when the counting strategy is used whether performed during the day or night, (c) the task is not learned as well at night when the rotation strategy is used, and (d) the rotation strategy may ultimately display the greatest amount of learning, when performed during the day.

So what could cause this interaction of strategy and shift? One possibility is the familiarity of the knowledge and transformations needed for the two strategies. The counting strategy relies heavily on well-known facts: the order of integers. That familiarity may have limited the impact of lower alertness and allowed participants on the night shift to arrive at a level of performance comparable to those on the day shift by the second half of the experiment.

In contrast, the rotation strategy may rely on knowledge that is less well practiced, thus requiring more cognitive or perceptual learning. Mental rotation is often associated with the visual perceptual system (e.g., Kosslyn, Thompson, & Ganis, 2006). While mental rotation is a well-practiced process, it may be stimulus or task specific. For instance, research has shown that the rate of rotation varies with stimulus complexity (Bethell-Fox & Shepard, 1988). While the stimuli in this task are relatively simple, the angle to be rotated by the participants is defined only by the end points, which may have added to the difficulty in maintaining an accurate visualization. Results of this imaginal visualization may be more difficult to learn or recall with a lower level of alertness, thus resulting in slower performance for participants on the night shift.

With practice, specific angles and rotations may be consolidated and stored in declarative memory. Within a session, the same combination of target and misalignment angle was never repeated. However, trials were repeated across sessions. This may have allowed participants to learn the results of mental rotations over days.

In addition, the rotation strategy may allow for more optimization of the procedural knowledge than does the counting strategy. Perhaps because mental rotations require more effort than counting, there was more pressure for additional optimization in the rotation strategy. Initially, the task takes longer to execute using the rotation strategy. This extra time may work as additional pressure to optimize (either explicitly or implicitly) the procedural knowledge brought to bear on the task. Further, variations in alertness may affect the pressure to optimize or the results of the optimization.

If the rotation strategy involves more learning throughout the task, either through declarative or procedural knowledge processes, then this may explain why participants using that strategy on the night shift performed more poorly. It is possible that one effect of decreased alertness is to decrease the effectiveness of learning. Specifically, fluctuations in alertness may affect the encoding, consolidation, or retrieval of declarative knowledge gained through effortful processes, like mental rotations, or interfere with the optimization of procedural knowledge (Jackson & Van Dongen, in press).

Importantly, performance on the last day of the experiment, when all participants performed the task during the day, does not support the argument that memory retrieval was the cause of slowed performance at night. Performance continued to improve only for participants using the rotation strategy on the day shift, but remained fairly consistent with the previous three days for all other participants. If retrieval processes, rather than learning or encoding, were causally involved, we would expect performance for night shift participants using the rotation strategy to improve noticeably on the last day. Additional research is required to determine if declarative knowledge, procedural knowledge, or both are affected by decreased alertness when performing orientation tasks at night.

Conclusion

Performance differences based on strategy and sleep patterns have both real-world and theoretical importance. The results have implications for task training and performance in real-world contexts, and also illustrate how different cognitive processes may be affected differently by circadian rhythms.

This study shows that training must be evaluated in context. The time of day in which the task will be performed and the time allowed for training need to be considered, among other things. If the choice of strategy were based upon the best day shift performance alone, the preferred strategy in this task may be rotation. However, shift alone is only part of the story. The rotation strategy resulted in performance improvements over the counting strategy only near the end of the two-week experiment. If the training period were short or if consistent performance across shifts were an important criterion, a strategy that uses familiar knowledge, as the count strategy does, may be more beneficial.

Choosing the correct strategy for the task environment can help reduce the effects of night shift decrements in alertness. Even small differences in performance can have drastic effects on some tasks. Orientation tasks are commonly performed in parallel with many time-critical tasks, such as driving or flying. Distractions from the

primary tasks of even a couple hundreds of milliseconds can have unwanted consequences, especially when magnified in more complex tasks and environments. This is true in many situations, in addition to orientation tasks, where delays and errors can have severe consequences.

This research also reveals ways in which different components of cognitive functioning, utilized by different strategies, are differentially affected by circadian rhythms. The performance of individuals using the counting strategy did not vary significantly between those on a day shift schedule and those on a night shift schedule. This robustness was likely the result of using familiar knowledge in the strategy, leading to similar learning trends regardless of shift assignment.

In contrast, there was a significant impact of shift on performance for those using the rotation strategy, suggesting that the cognitive processes involved may be less robust to degradations in alertness at night. This vulnerability may be due to a greater reliance on the learning of visual perceptual information (i.e., angles and rotations), which appeared to be hindered by lower alertness.

In conclusion, the findings presented here speak to both the need for considering the strategy set used in a task and the potential for decrements in learning caused by decreased alertness. In other words, when evaluating the effects of cognitive moderators, such as alertness, it is critical to consider the strategy people use to complete tasks.

Acknowledgments

The views expressed in this paper are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government. The research was supported in part by the Air Force Research Laboratory's Warfighter Readiness Research Division and grants 07HE01COR, 09RH06COR, 10RH04COR and FA9550-09-1-0136 from the Air Force Office of Scientific Research (AFOSR). The first author was supported by an appointment to the Postgraduate Research Participation Program at the U.S. Air Force Research Laboratory administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and the Air Force Research Laboratory. The experimental research was supported by FMCSA grant DMC75-07-D-0006.

References

Åkerstedt, T. (1988). Sleepiness as a consequence of shift work. *Sleep*, *11*(1), 17-34.

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review*, *111*(4), 1036-1060.

Bethell-Fox, C. E. & Shepard, R. N. (1988). Mental rotation: Effects of stimulus complexity and familiarity. *Journal of Experimental Psychology: Human Perception & Performance*, 14(1), 12-23.

Caldwell, J. A. (2003, Fall). Wake up to the importance of sleep for air safety. *Flightline*, 30-33.

Dinges, D. F. (1995). An overview of sleepiness and accidents. *Journal of Sleep Research*, *4*(2), 4-14.

Gross, J. B., Gunzelmann, G., Gluck, K. A., Van Dongen, H. P. A., & Dinges, D. F. (2006). Computational modeling of the combined effects of circadian rhythm and sleep deprivation. *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, Vancouver, B.C., Canada, 297-302.

Gunzelmann, G., Anderson, J. R., & Douglass, S. (2004). Orientation tasks with multiple views of space: Strategies and performance. *Spatial Cognition and Computation*, *4*(3), 207-253.

Jackson, M. L. & Van Dongen, H. P. A. (in press). Cognitive effects of sleepiness. In Thorpy, M. & Billiard, M. (editors), *Sleepiness*, Cambridge University Press.

Kosslyn, S. M., Thompson, W. L., & Ganis, G. (2006). *The Case for Mental Imagery*. New York: Oxford.

Lim, J. & Dinges, D. F. (2008). Sleep deprivation and vigilant attention. *Annals of the New York Academy of Science*, *1129*, 305-322.

Mednick, S., Nakayama, K., & Stickgold, R. (2003). Sleep-dependent learning: A nap is as good as a night. *Nature Neuroscience*, *6*(7), 697-698.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & the R Core Team (2009). *nlme: Linear and nonlinear mixed effects models* (R package version 3.1-96) [Computer Software].

R Development Core Team (2009). *R: A language and environment for statistical computing* [Computer Software]. Vienna, Austria.

Van Dongen, H. P. A. & Dinges, D. F. (2005). Sleep, circadian rhythms, and psychomotor vigilance. *Clinics in Sports Medicine*, *24*(2), 237-249.

Walker, M. P., Brakefield, T., Morgan, A., Hobson, J. A., & Stickgold, R. (2003). Practice with sleep makes perfect: Sleep-dependent motor skill learning. *Neuron*, *35*(1), 205-211.

A Distributional Account of Covariance Effects and Talker Adaptation in Infant and Adult Phonetic Category Recognition

Bevan K. Jones (Bevan_Jones@Brown.edu)
Department of Cognitive and Linguistic Sciences
Brown University
Providence, RI 02912, USA

Abstract

Both infants and adults are sensitive to the non-linguistic features of speech, and this sensitivity impacts speech sound categorization, but with somewhat different effects. While both infants and adults sometimes confuse the non-linguistic for the linguistic and are susceptible to categorization problems when the two covary, adults, on the other hand, are often able to exploit non-linguistic features to improve perceptual categorization. We present a Bayesian account of both adult and infant behavior, arguing that differing levels in linguistic maturity correspond to different models of linguistic structure. The infant's task is one of structure learning, adults, on the other hand, are estimating parameters for an already established structure.

Keywords: Speech perception; distributional learning; language acquisition; Bayesian models.

Introduction

Talker variability is a fundamental challenge in speech perception. The same phonetic category as uttered by two different talkers may seem quit different. At the same time, distinct categories produced by two different talkers may be acoustically quite similar (Dorman, Studdert-Kennedy, & Raphael, 1977; Magnuson & Nusbaum, 2007). Unsurprisingly, this variability poses a problem for infants as they acquire their language. In particular, studies have shown that infants are prone to confounding talker-specific characteristics with phonetic categories when the talker covaries with the category during learning (Houston & Jusczyk, 2000; Creel, Aslin, & Tanenhaus, 2008). For instance, when taught to recognize two different categories, one produced exclusively by a female speaker and the other by a male speaker, infants were unable later to identify those phones when spoken by the opposite sex. This suggests that learning not only involves acquiring information about the features of the exemplars of the category, but, more fundamentally, about which features relate to the categorization task at all.

Adults are not immune to talker variability either and can also be misled by talker differences (Kraljic, Brennan, & Samuel, 2008; McQueen, Norris, & Cutler, 2006), but the same studies also demonstrate that adults are able to adapt to the differences. In fact, speaker identity may even be exploited to improve recognition performance at times, as suggested by experiments with episodic memory. Goldinger (1996) showed that words spoken by one speaker can be more easily recognized when uttered by the same speaker even after significant time has elapsed. This suggests that not only do listeners note linguistically weighted cues but also indexical cues that might be used for talker identification.

While both infants and adults are faced with similar input and utilize statistical learning mechanisms, the nature of the problem they each face is quite different. Both face a categorization problem. Infants are still struggling to decide which dimensions in the high dimensional perceptual space are most relevant to the categorization task. Voice onset time, for instance, serves largely to distinguish the words "dime" and "time" since "d" is followed by a much shorter voicing delay than "t". Other features such as fundamental frequency may serve an indexical function (aiding in distinguishing whether the talker is male or female, for instance) but are much less clearly related to the linguistic content in a language like English. Infants are engaged in a kind of feature selection, narrowing down the infinite set of possible features to just those that are most useful. Adults, on the other hand, have already determined which features are linguistic and which are not. However, far from simply discarding the non-linguistic information, adults may employ indexical features to track the talker, allowing them to adapt to the peculiarities of the individual's speech patterns.

We present a Bayesian account for both the infant and adult behavioral results. In the infant's case, the problem can be framed in terms of a model selection problem, a search through some space of models that relate the latent phonetic category to the observed features, both linguistic and non-linguistic. In the adult's case, talker adaptation is more of a problem of parameter estimation given an already learned model relating phonetic category, talker, and the observed linguistic and indexical features.

The models we present fall within the distributional learning paradigm. It is well known that speech sounds of all types tend to fall according to a Gaussian distribution (Peterson & Barney, 1951; Lisker & Abramson, 1964; Espy-Wilson, 1992). Furthermore, Maye, Werker, and Gerken (2002) show that bimodal distributions tend to prompt infants to identify two sounds where unimodal distributions lead to identification of a single category, suggesting that learners may rely to some extent on an assumption of something like a Gaussian distribution. Thus, learning can be characterized as a kind of parametric statistical search over unimodal or, in our case, Gaussian distributions.

We present an array of models to account for the different behaviors, arguing that not one, but several different models of the dependencies between features are required. Linguistic development is characterized under our assumption of multiple models as the selection of one model over another based

on accumulated evidence. In the early days, when infants have little evidence of which model is likely to generalize, infants make decisions based on recent experience. Hence, covarying talker with phonetic category during training results in the infant's selecting a model that does not generalize to a more natural situation where talker and phonetic category do not covary. Similarly, we argue that adult talkers also shift between models depending on the available information. In the adult's case experience is not so acute an issue, but some features are not always present in the input, or are obscured by noise, and thus they must use an alternative model that does not depend on those features.

We argue for a fluid shifting between models over a single monolithic model. Shifts between qualitatively different models, as opposed to a gradual adjustment of a single model, accounts for how distinct situations result in different processes. Yet each model operates on the same basic principles of distributional learning, where even the shift between models may be accounted for within a Bayesian framework.

Model Definitions

Figure 1 presents the four different structural relationships we consider, slight variations but with important implications. At heart, they are all instances of a Gaussian mixture model which attempts to explain the linguistic feature x_i of the ith sound by a distribution indexed by the sound's phonetic category c_i. The more complex models ($\mathcal{M}_3$ and $\mathcal{M}_4$) elaborate on the theme by introducing talker specific distributions over x_i, and introduce an additional latent variable t_i for each sound to represent talker identity. All the models assume exactly two phonetic categories, and the talker specific models in turn assume exactly two talkers, a restriction that is easily relaxed but does not interfere with our purpose: explaining the human behavior in certain psycholinguistic experiments.

In the case of models $\mathcal{M}_1$ and $\mathcal{M}_3$ each speech sound also bears an indexical feature y. The two models treat y quite differently, however. $\mathcal{M}_1$ assumes all features are linguistic, and therefore represents a direct dependency between c_i and y_i, paralleling the dependency between c_i and x_i. $\mathcal{M}_3$, however, distinguishes between linguistic and indexical features, and introduces a direct dependency between the indexical feature and the talker instead of the phonetic category. This change captures the notion that indexical features primarily serve to identify the talker, and only secondarily aid in recognition. This feature could be anything: fundamental frequency, or even an odd way of smacking ones lips at the end of each utterance. Since we are primarily interested in modeling phonetic category learning and not so much talker recognition, we treat this feature as a simple Bernoulli variable with a predefined parameter. That is, while the model learns the parameters for the distributions over x, y is determined by a prespecified Bernoulli parameter.

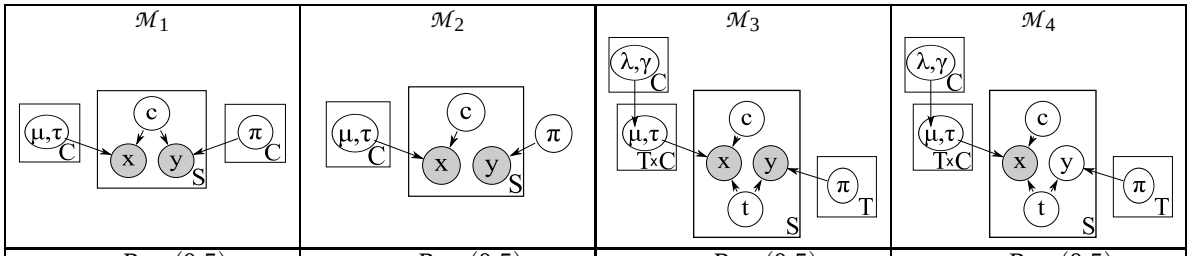
These models attempt to explain the phenomena observed in certain psycholinguistic experiments. Houston and Jusczyk (2000) demonstrated that 7.5 month olds were able to recognize words in a segmentation task when they were produced by a speaker of the same sex during test time as during training, but were unable to generalize across sexes. Singh (2008) demonstrates a similar sensitivity to other covariant non-linguistic features. Model $\mathcal{M}_1$ captures the behavior of infants in these situations, where all features are treated as linguistic. Since the model assumes all features are directly relevant to the categorization task, it will have a tendency to over fit when presented with data where talker and phonetic category accidentally covary (or are contrived to do so by an experimenter). Model $\mathcal{M}_2$, on the other hand, treats the indexical feature as independent, only modeling the dependency between x and c, and is more likely to generalize across speakers.

Models $\mathcal{M}_3$ and $\mathcal{M}_4$ introduce the ability to adapt to individual talkers by providing separate talker-specific distributions for the linguistic feature x. However, the individual talker-specific distributions for a particular phonetic category are related to each other by a distribution for the category common to all talkers. Thus, we introduce a hierarchical Gaussian distribution over linguistic features, capturing the notion that, although each talker may have his own peculiar way of producing a sound, sounds of the same category all tend to be similar across speakers. The hierarchical distribution allows for speech recognition even when faced with a completely unfamiliar talker, since the λ and γ parameters define a prior over talker specific categories, providing a mechanism of generalization from familiar talkers to novel talkers.

Goldinger (1996) showed that adults are better able to understand speech when presented by the same talker. Similarly, Kraljic et al. (2008) noted that adults adapt to speaker-specific idiosyncrasies. In particular, they showed that when presented with speech where the alveolar fricative "s" as in the word "see" was shifted to a more palatal place of articulation resembling "sh" as in "she", subjects were able to adapt and correctly identify the shifted "s" sounds — so long as they were provided with cues as to which variant of "s" was likely to occur. These situations are modeled by $\mathcal{M}_3$ and $\mathcal{M}_4$. $\mathcal{M}_3$ uses the additional cue y to help identify the talker, and hence, the correct distribution for the category over linguistic cue x. This way the indexical feature has an indirect impact on recognition even if there is no direct dependency between c and y. $\mathcal{M}_4$ attempts to adapt to the talker without the aid of the indexical cue. The model assumes such features exist, but are not observed and therefore cannot assist in identifying the talker. The prediction for $\mathcal{M}_4$ is that, like the subjects in the study by Kraljic et al. (2008), the model will perform more poorly and will incorrectly allow talker-specific variation to influence recognition of other talkers.

Inference

The models were implemented using WinBUGS (Spiegelhalter, Thomas, Best, & Lunn, 2003), which uses an automatic Gibbs sampling MCMC approach to estimate parameters and allows rapid prototyping and testing



$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$	$\mathcal{M}_4$
$c \sim Bern(0.5)$	$c \sim Bern(0.5)$	$c \sim Bern(0.5)$	$c \sim Bern(0.5)$
		$t \sim Bern(0.5)$	$t \sim Bern(0.5)$
		$\lambda \sim \mathcal{N}(30, 5 \cdot 10^{-4})$	$\lambda \sim \mathcal{N}(30, 5 \cdot 10^{-4})$
		$\gamma \sim Gamma(0.2, 0.2)$	$\gamma \sim Gamma(0.2, 0.2)$
$\mu_c \sim \mathcal{N}(30, 5 \cdot 10^{-4})$	$\mu_c \sim \mathcal{N}(30, 5 \cdot 10^{-4})$	$\mu_{c,t} \lvert \lambda_c, \gamma_c \sim \mathcal{N}(\lambda_c, \gamma_c)$	$\mu_{c,t} \lvert \lambda_c, \gamma_c \sim \mathcal{N}(\lambda_c, \gamma_c)$
$\tau_c \sim Gamma(0.2, 0.2)$	$\tau_c \sim Gamma(0.2, 0.2)$	$\tau_{c,t} \sim Gamma(0.2, 0.2)$	$\tau_{c,t} \sim Gamma(0.2, 0.2)$
$x \lvert c, \mu, \tau \sim \mathcal{N}(\mu_c, \tau_c)$	$x \lvert c, \mu, \tau \sim \mathcal{N}(\mu_c, \tau_c)$	$x \lvert c, t, \mu, \tau \sim \mathcal{N}(\mu_{c,t}, \tau_{c,t})$	$x \lvert c, t, \mu, \tau \sim \mathcal{N}(\mu_{c,t}, \tau_{c,t})$
$y \lvert c, \pi \sim Bern(\pi_c)$	$y \lvert \pi \sim Bern(\pi)$	$y \lvert t, \pi \sim Bern(\pi_t)$	$y \lvert t, \pi \sim Bern(\pi_t)$

Figure 1: Four Possible Speech Perception Models: $\mathcal{M}_1$ treats all features as linguistic, $\mathcal{M}_2$ distinguishes the true and false linguistic features, $\mathcal{M}_3$ models individual talkers and treats some features as indexical, and $\mathcal{M}_4$ models talkers where the indexical features are absent or obscured. The variables are defined as follows: c is the speech sound category, t is the talker, x is a linguistic feature, y is an indexical feature, and the other variables are distributional parameters, defining talker and category specific distributions. C is the set of categories, T is the set of talkers, and S is the set of all speech sound tokens.

of Bayesian models.

We use an explicit initialization strategy, running the models in a generative mode with no observed variables and drawing category parameters for x at random from a $\mathcal{N}(50, 0.0025)$ for the mean and a $Gamma(2,2)$ distribution for the precision. Using an initialization strategy such as this could speed convergence, since it tends to start the model out in a higher probability space. It also has the effect of reducing problems with numerical underflow error in WinBUGS. We were careful to pick the parameters randomly in such a way as to avoid biasing search in favor of any particular model or clustering, since we are primarily interested in the model properties, not the effects of initialization on convergence.

We find that even the more complex models converge in well under the 30,000 iterations we use. We average over the next 1000 iterations after convergence to measure the various parameters and statistics we report in subsequent sections. We take care in observing performance over these last 1000 iterations for any trends or abrupt changes. These mixture models have multiple symmetric optimal solutions, where "t" may be associated with cluster 1 and "d" with 2, or vice versa. If left to run long enough, the MCMC search strategy tends to switch between these different symmetric configurations every few thousand iterations. Averaging over instances of multiple such symmetric cases results in increased error in measurement. For instance, attempting to estimate the mean x value for phones in a cluster that toggles between "t" and "d" gets an average that is dissimilar to both configurations, and not only results in a measurement that is far from the gold standard but does not even accurately reflect the stationary distribution of the sampler.

Simulations

Data

We run the model on three synthetic data sets, illustrating the contrast between English word initial "t" and "d". The primary difference between the two is in the voice onset time (VOT). We generate 100 sounds. Table 1 shows the model parameters used to generate each of the three data sets. For data set one we generate sounds as though there is only one speaker. For data set two we use two talkers, covarying the category with the talker so that instances of the first phone are produced by talker one and all instances of the second phone are produced by talker two. Finally, for data set three we split the 100 sounds evenly between the two talkers and the two categories, where talker and category are independent.

Simulation 1: The Developmental Situation

To simulate a situation similar to the psycholinguistic experiments of Houston and Jusczyk (2000), we present the models with two different data sets: data set one, where there is only one talker, and data set two, where there are two talkers, each producing just one of the two phones. In the behavioral experiment, it was observed that infants trained with word stimuli in a female voice were only able to reliably recognize words at test time when they were again presented in a female voice, and could not generalize to a male voice. Thus, the infants seem to confuse some non-linguistic feature of the sound, perhaps fundamental frequency, with the linguistic identity of the sounds. In this simulation, we shall

Table 1: Three Synthetic Data Sets

Talker	Parameter	Data Set One	Data Set Two	Data Set Three
One	π_1	0.5	0.8	0.8
	π_2	0.5	0.8	0.8
	μ_1	15	15	0
	μ_2	35	-	35
	τ_1	15^{-2}	15^{-2}	15^{-2}
	τ_2	5^{-2}	-	5^{-2}
Two	π_1	-	0.2	0.2
	π_2	-	0.2	0.2
	μ_1	-	-	15
	μ_2	-	35	65
	τ_1	-	-	15^{-2}
	τ_2	-	5^{-2}	5^{-2}
Talkers Covary		-	Yes	No

say that our indexical feature y corresponds to a thresholded fundamental frequency: sounds with a high fundamental frequency are more likely to be produced by the female talker, and lower fundamental frequency sounds by the male talker.

To simulate the developmental character of an infant's nascent linguistic capabilities, we perform a kind of structure discovery using Bayesian model selection between $\mathcal{M}_1$ and $\mathcal{M}_2$, where the infant is attempting to determine if the indexical feature y is relevant to the linguistic category ($\mathcal{M}_1$) or not ($\mathcal{M}_2$). We do this by introducing an additional latent variable corresponding to the model and define a uniform prior over the model. Then, we compute the probability of the model given the data, integrating out all other variables. To compare the two models, we simply compare the probabilities assigned to each model given the data. Typically, in such cases if the ratio $P(\mathcal{M}_1|D)/P(\mathcal{M}_2|D)$, called the Bayes factor, is greater than one, we say that model one is preferred, and otherwise model two is preferred.

In this case, whether we use data set one or two, virtually all the probability mass (approximately 100%) is placed on exactly one of the two models. $\mathcal{M}_1$ is overwhelmingly preferred when using data set two, the case where talker and phonetic category covary. On the other hand, data set one, the data set where both phonetic categories are produced by the same talker, results in an overwhelming preference for $\mathcal{M}_2$.

Table 2 presents accuracy results for the two models on the two data sets. Note that in general for these sorts of clustering algorithms there is an identifiability problem. That is, we cannot immediately say whether a particular category value $c = 1$ corresponds to the "t" or "d" sound. However, this poses less of a problem for this simple case with only two categories. For our purposes, it seems sufficient to assign the category that achieves highest accuracy.

We observe that while the model that mistakes the indexical for a linguistic feature ($\mathcal{M}_1$) performs very well for the artificially contrived covarying data, it performs worse on the

Table 2: Categorization Accuracy

Model	Data Set One	Data Set Two	Data Set Three
$\mathcal{M}_1$	0.77	**0.89**	0.52
$\mathcal{M}_2$	**0.81**	0.81	**0.76**

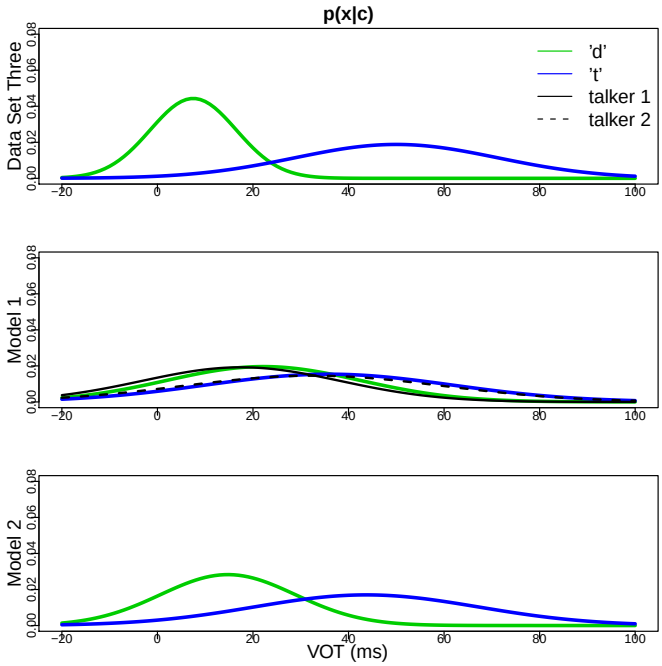


Figure 2: The conditional distribution over x given c for the true data set as compared to the two models $\mathcal{M}_1$ and $\mathcal{M}_2$. The clusters for the two talkers have been merged for ease of comparison. We also compare model 1's clusters against the distribution over x given the talker t.

data set that has only a single talker, and very nearly at chance for the data set with two different talkers that don't covary with the category. Figure 2 depicts the clusterings found by the two models on data set three (the data set with two talkers that don't covary with the phone). While $\mathcal{M}_2$ seems to do as well as can be hoped considering its inability to adapt to individual talkers, $\mathcal{M}_1$ very nearly fails to differentiate at all between "t" and "d". $\mathcal{M}_1$ attempts to cluster according to the indexical, collapsing the two categories together for each talker and clustering by talker instead of by category.

Thus, the model selection approach predicts the psycholinguistic results very well. Training on sounds in one talker's voice, as in the covarying data set, results in the incorrect model being learned, which then fails to generalize to the same sound produced in the other talker's voice.

Simulation 2: Talker Adaptation

Adult talkers actually have the ability to adapt to individual talkers, learning to exploit talker specific variations

(Goldinger, 1996). To simulate this ability, we compare the performance of models $\mathcal{M}_3$ and $\mathcal{M}_4$. Model $\mathcal{M}_3$ corresponds to a case where the subject has learned that the indexical feature y can be used to identify the talker. On the other hand, $\mathcal{M}_4$ corresponds to the case where, although the subject is aware that the sounds may be produced by a different talker, the voice is disguised so that no cue is available for the identification of the talker. The contrast between these two models is similar to that demonstrated by Kraljic et al. (2008), where subjects were presented with ambiguous sounds that, in one condition, were accompanied by an additional cue indicating the ambiguity was result of talker dialect, and, in a second condition, were presented without this cue. This dialectical indicator, based on a phonological context, corresponds to our indexical feature y. Thus, condition one corresponds to $\mathcal{M}_3$ and condition two to $\mathcal{M}_4$. In the behavioral study, it was observed that subjects were much more prone to confusing the two different phonetic categories when the sounds were presented without the additional cue. Thus, we expect $\mathcal{M}_3$ to do much better.

Table 3 contains the categorization accuracy results for $\mathcal{M}_3$ and $\mathcal{M}_4$. Note that these models can theoretically identify the talker as well as the phonetic category, and we report accuracy for both. $\mathcal{M}_3$ does slightly better at clustering the phonetic categories, which is likely due to its much better ability to identify the talker. Note that without the indexical feature, $\mathcal{M}_4$ is at chance with regard to talker identification.

Table 3: Categorization Accuracy for Data Set Three

Model	Category	Talker
$\mathcal{M}_3$	**0.86**	**0.78**
$\mathcal{M}_4$	0.81	0.50

Figure 3 shows the clusters inferred by the two talker adapting models. The inferred Gaussian distributions for the two talkers are much more distinct for $\mathcal{M}_3$ than they are for $\mathcal{M}_4$ and more closely resemble the true distribution.

The inferred clusters, presented in Figure 3, are particularly interesting when compared against the findings of Kraljic et al. (2008), who observed that when the dialectical cue was absent, subjects adjusted their perceptual judgments for all talkers, not just the talker that produced the ambiguous variant. Model 3 makes use of the additional feature y for keeping the two talkers distinct, and therefore is less likely to let experience with the ambiguous talker influence its judgment for the other talker. Similarly, model 4 captures the situation where no additional cues are available. In this case, even if separate clusters are maintained for each talker, the two are functionally identical, falling somewhere in between the two true clusters. The mean is the mean of the two talker specific variants of the category, and, in the case of the "d" sound, the variance is much larger. Thus, the ambiguous talker influences recognition of the other talker when no additional cues are available, but not nearly as much when additional cues are discernible.

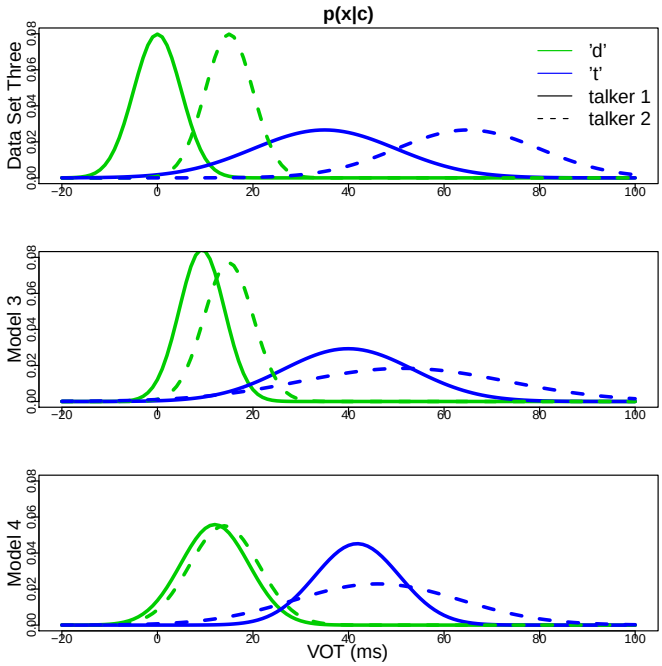


Figure 3: The conditional distribution over x given c for the true data set as compared to the two models $\mathcal{M}_3$ and $\mathcal{M}_4$.

As in the case of the developmental simulation, we see that the alternate performance of two models predicts the empirical results much better than would any one of the two.

Discussion and Conclusion

We have presented a computational model demonstrating a distributional account of certain covariance effects in infant and adult speech perception observed in the psycholinguistic literature. In particular, we found that by modeling the development of infant speech perception as a type of Bayesian model selection, we can account nicely for documented effects of covarying talker and phonetic category on infant confusions between categories (Houston & Jusczyk, 2000). We also found that by modeling talker identity, the same talker-specific features that confused the infant models could be exploited to improve performance, similar to demonstrations of talker adaptation in adult subjects (Kraljic et al., 2008). Also consistent with Kraljic et al. (2008), we found that when the talker adapting models were deprived of observed indexical information, talker specific speech habits influenced the category representations for all talkers not just the talker that produced the offending speech sounds.

While it would be difficult to account for all the phenomena with a single model of the statistical dependencies in the data, multiple models predict the empirical results fairly closely. This raises the question of how human subjects move between models, begging a model of the model selection process itself. Developmental shifts are readily handled in the Bayesian framework as a model selection problem, just the

approach we took for explaining the infant behavior. Though it is beyond the scope of this paper, a similar selection process may account for a shift between the infant and adult stages, perhaps with several additional intermediate structures. We argue that modeling the developmental process as a shift between models rather than a gradual adjustment of a single model better matches the fact that there are distinct developmental stages. One set of models may correspond to a particular stage, where the underlying behavioral causes are made explicit by the dependency structure of the model.

Although the simulations we presented dealt primarily with covariance between talker and phonetic category, we expect that models based on similar principles could explain equally well other kinds of covariance phenomena, such as with speech affect and category (Singh, 2008). Note that the models we presented to explain infant phenomena had no explicit model of talker identity. Thus, the choice between the two models in the developmental case only constituted a feature selection task, where features that clearly covaried with the phonetic category were greatly preferred by the selection criterion. Thus, these simple models, in fact, generalize directly.

Similarly, while the talker adapting models do contain an explicit representation of talker identity, there is nothing that requires that the t variable refer to a talker. Similar variables could represent modes of talking, such as infant directed speech, or happy speech, or to dialectical variations or any number of other categorizable speech types. That is, the talker adapting models present a general adaptation strategy that could be employed with little or no modification.

We argue for the generality of the principles underlying our computational account while stressing that the full speech recognition problem, or even just that of phonetic category recognition, is a difficult one, and we have not attempted to model it in its entirety. In fact, we made several explicit simplifications. First, we assumed there are only two categories and two talkers. Second, we assumed that there are roughly equal numbers of tokens of each category, and that each talker produces about half of the sounds. Also, since we were primarily interested in how phonetic categories are learned, we assumed a simple Bernoulli distribution for the indexical feature, when, in fact, in many cases this feature too may very well be continuous. Furthermore, it was sufficient for our purposes to model a recognition problem along only one or two dimensions of the perceptual space.

These simplifications eased the implementation work but did not interfere with our ability to simulate the behavioral situations in which we were interested. They should not limit the generalizability of the results, and could be relaxed in a fairly straightforward manner if we wished to increase the realism. For instance, the first restriction could be relaxed by allowing the model to infer how many categories there are from the data using an infinite mixture model. We could also use a beta prior to infer relative talker and phonetic category frequency. A similar prior could be used to infer the distribution over the indexical features. Finally, multivariate Gaussians could be used for multiple correlated linguistic features (Vallabha, McClelland, Pons, Werker, & Amano, 2007). These are obvious extensions to consider for future work.

References

Creel, S. C., Aslin, R. N., & Tanenhaus, M. K. (2008). Heeding the voice of experience: The role of talker variation in lexical access. *Cognition*, *106*, 633–664.

Dorman, M. F., Studdert-Kennedy, M., & Raphael, L. J. (1977). Stop consonant recognition: Release bursts and formant transitions as functionally equivalent, context-dependent cues. *Perception & Psychophysics*, *22*, 109–122.

Espy-Wilson, C. Y. (1992). Acoustic measures for linguistic features distinguishing the semivowels /w j r l/ in american english. *Journal of the Acoustical Society of America*, *92*(1), 736–757.

Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 1166–1183.

Houston, D. M., & Jusczyk, P. W. (2000). The role of talker-specific information in word segmentation by infants. *Journal of Experimental Psychology: Human Perception and Performance*, *26*(5), 1570–1582.

Kraljic, T., Brennan, S. E., & Samuel, A. G. (2008). Accommodating variation: Dialects, idiolects, and speech processing. *Cognition*, *107*, 54–81.

Lisker, L., & Abramson, A. A. (1964). A cross-language study of voicing in initial stops: acoustical measurements. *Word*, *20*, 384–422.

Magnuson, J. S., & Nusbaum, H. C. (2007). Acoustic differences, listener expectations, and the perceptual accommodation of talker variability. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(2), 391–409.

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*(3), B101–B111.

McQueen, J. M., Norris, D., & Cutler, A. (2006). The dynamic nature of speech perception. *Language and Speech*, *49*, 101–112.

Peterson, G. E., & Barney, H. (1951). Control methods used in the study of vowels. *Journal of the Acoustical Society of America*, *24*(2), 175–184.

Singh, L. (2008). Influences of high and low variability on infant word recognition. *Cognition*, *106*, 833–870.

Spiegelhalter, D., Thomas, A., Best, N., & Lunn, D. (2003). *Winbugs: User manual, version 1.43*. Cambridge: Medical Research Council Biostatistics Unit.

Vallabha, G. K., McClelland, J. L., Pons, F., Werker, J. F., & Amano, S. (2007). Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Science*, *104*, 13273–13278.

Getting at the Cognitive Complexity of Linguistic Metadata Annotation – A Pilot Study Using Eye-Tracking

Steffen Lohmann
Dept. of Computer Science & Applied Cognitive Science
Universität Duisburg-Essen
Duisburg, Germany

Katrin Tomanek
Language & Information Engineering (JULIE) Lab
Universität Jena
Jena, Germany

Jürgen Ziegler
Dept. of Computer Science & Applied Cognitive Science
Universität Duisburg-Essen
Duisburg, Germany

Udo Hahn
Language & Information Engineering (JULIE) Lab
Universität Jena
Jena, Germany

Abstract

We report on an experiment where the decision behavior of annotators issuing linguistic metadata is observed with an eye-tracking device. As experimental conditions we consider the role of textual context and linguistic complexity classes. Still preliminary in nature, our data suggests that semantic complexity is much harder to deal with than syntactic one, and that full-scale textual context is negligible for annotation, with the exception of semantic high-complexity cases. We claim that such observational data might lay the foundation for empirically grounded annotation cost models and the design of cognitively adequate annotation user interfaces.

Keywords: Natural Language Metadata Annotation; Annotation Behavior; Eye-Tracking; Syntactic Complexity; Semantic Complexity; Cognitive Cost Modeling

Introduction

Supervised approaches to machine learning (ML) are currently very popular in the natural language processing (NLP) community. While linguistic regularities are no longer handcrafted by human experts in this paradigm, human intervention is still required to produce sufficient amounts of reliably annotated training material from which ML classifiers may learn or, considered as empirically valid ground truth, against which NLP systems can be evaluated.

The assignment of linguistic metadata (e.g., related to parts of speech, syntactic parses, or semantic interpretations) to plain natural language corpus data, a process called *annotation*, is a complex cognitive task. It requires a sound competence of the natural language in the corpus, as well as a decent level of domain and even text genre expertise.

Meanwhile lots of annotated corpora have been built which contain these precious human judgments (e.g., PennTreeBank (Marcus, Santorini, & Marcinkiewicz, 1993), PennPropBank (Palmer, Gildea, & Kingsbury, 2005) or OntoNotes (Pradhan et al., 2007)). Almost all of these annotated corpora were assembled by collecting the documents to be annotated on a random sampling basis (once the original document set had been restricted thematically or chronologically).

Only recently, more sophisticated approaches to select the annotation material are being investigated in the NLP community. One of the most promising approaches is known as *Active Learning* (AL) (Cohn, Ghahramani, & Jordan, 1996 ; Tomanek, Wermter, & Hahn, 2007) where an intentional selection bias is enforced and only those linguistic samples are selected from the entire document collection which are considered to be most informative to learn an effective classification model. When different approaches to AL are compared with each other, or with standard random sampling, in terms of annotation efficiency the AL community, up until now, assumed *uniform* annotation costs for each linguistic unit, e.g., words (Ringger et al., 2008 ; Settles, Craven, & Friedland, 2008 ; Arora, Nyberg, & Rosé, 2009). This claim, however, has been shown to be invalid in several studies (Hachey, Alex, & Becker, 2005 ; Settles et al., 2008 ; Tomanek & Hahn, 2010). If uniformity does not hold and, hence, the number of annotated units does not indicate the true annotation efforts required for a specific sample, empirically more adequate cost models have to be developed. Accordingly, we here consider different classes of syntactic and semantic complexity that might affect the cognitive load during the annotation process, with the overall goal to find empirically more adequate variables for cost modeling.

The complexity of linguistic utterances can be judged either by structural or by behavioral criteria. Structural complexity emerges, e.g., from the static topology of phrase structure trees and procedural graph traversals exploiting the topology of parse trees (see Szmrecsányi (2004) or Cheung et Kemper (1992) for a survey of metrics of this type). However, structural complexity criteria do not translate directly into empirically justified cost measures.

The behavioral approach accounts for this problem as it renders observational data of the annotators' eye movements. The technical vehicle to gather such data are eye-trackers which have already been used in psycholinguistics (Rayner, 1998). Eye-trackers were able to reveal, e.g., how subjects deal with ambiguities (Frazier & Rayner, 1987 ; Rayner, Cook, Juhas, & Frazier, 2006 ; Traxler & Frazier, 2008) or with sentences requiring re-analysis, so-called garden path sentences (Altmann, Garnham, & Dennis, 2007 ; Sturt, 2007).

The rationale behind the use of eye-tracking devices for the observation of the annotation behavior is that the length of gaze durations and the behavioral patterns underlying gaze movements are considered to be indicative for the hardness of the linguistic analysis and the expenditures for the search of clarifying linguistic evidence (e.g., anchor words) to solve hard decision tasks such as phrasal attachments or word sense disambiguation. Gaze duration and search time are then taken as empirical correlates of processing complexity and, hence, unveil the *real* costs. We therefore consider eye-tracking as a promising means to get a better understanding of the nature of linguistic annotation processes with the ultimate goal of identifying predictive factors for annotation cost models.

[Federal Aviation Adminstration]$_{ORG}$ investigators were to examine the aircraft, said spokeswoman *[Arlene]*$_{PER}$. She said *[Martinair Holland]*$_{ORG}$ is certified to fly large jet aircraft into the *[US]*$_{LOC}$ as a scheduled passenger service.

When the *[Cessna]*$_{ORG}$ took off in rain and snow from the 6,900-foot runway at *[Cheyenne Municipal Airport]*$_{LOC}$ in *[Wyoming]*$_{LOC}$, *[Reid]*$_{PER}$ was seated at one control panel, *[Jessica]*$_{PER}$ was seated at another and her father was in a passenger seat in a four-seat *[Cessna]*$_{ORG}$ 177B, a 21-year-old single-engine plane owned by *[Reid]*$_{PER}$.

Figure 1: Text snippets taken from MUC7 documents annotated by *LOCation*, *PERson*, and *ORGanization* entity types.

Experimental Design

The focus of our study is on semantic annotation, the annotation of named entity mentions in particular. In this task, a human annotator has to decide for each word in a sentence whether it belongs to one of the entity types of interest or not. For the first time ever to the best of our knowledge, we applied eye-tracking to study the cognitive processes underlying the annotation of linguistic metadata.

We used the English part of the MUC7 corpus (Linguistic Data Consortium, 2001) for our study, which contains *New York Times* articles from the year 1996 reporting on plane crashes. These articles come already annotated with three types of named entities considered important in the newspaper domain, *viz.* "persons", "locations", and "organizations". Figure 1 depicts typical text snippets from these articles along with the available annotations.

Annotation of these entity types in newspaper articles is admittedly fairly easy. We chose this rather simple setting because the participants in the experiment had no previous experience with document annotation and no serious linguistic education background. Moreover, the limited number of entity types reduced the amount of participants' training prior to the actual experiment, and positively affected the design and handling of the experimental apparatus (see below).

We triggered the annotation processes by giving our participants specific *annotation examples*. An example consists of a text document having one single *annotation phrase* highlighted which then had to be semantically annotated for named entity mentions. The annotation task was defined such that the correct entity type had to be assigned to each word in the annotation phrase. If a word belongs to none of the three entity types a fourth class, "no entity", had to be assigned.

The phrases highlighted for annotation were *complex noun phrases* (CNPs), each a sequence of words where a noun (or an equivalent nominal expression) constitutes the syntactic head and thus dominates dependent words such as determiners, adjectives, or other nouns or nominal expressions (including noun phrases and prepositional phrases). CNPs with even more elaborate internal syntactic structures, such as coordinations, appositions, or relative clauses, were isolated from their syntactic host structure and the intervening linguistic material containing these structures was deleted to simplify overly long sentences. We also discarded all CNPs that did not contain at least one *entity-critical* word, i.e., one which might be a named entity given its orthographic appearance (e.g., starting with an upper-case letter). It should be noted that such orthographic signals are by no means a sufficient condition for the presence of a named entity mention within a CNP.

The choice of CNPs as stimulus phrases is motivated by the fact that named entities are usually fully encoded by this kind of linguistic structure. The chosen stimulus – an annotation example with one phrase highlighted for annotation – allows for an exact localization of the cognitive processes and annotation actions performed relative to that specific phrase.

Independent Variables

We defined two measures for the complexity of the annotation examples: The *syntactic* complexity was given by the number of nodes in the parse tree dominated by the annotation phrase (Szmrecsányi, 2004).[1] According to a threshold on the number of nodes in such a parse tree, we classified CNPs as having either high or low syntactic complexity.

The *semantic* complexity of an annotation example is based on the inverse document frequency $df(w_i)$ of each word w_i of the respective CNP according to a reference corpus.[2] We calculated the semantic complexity score as $max_i \frac{1}{df(w_i)}$, where w_i is the i-th word of the annotation phrase. Again, we determined a threshold classifying CNPs as having either high or low semantic complexity. This automatically generated classification was then manually checked and, if necessary, revised by two annotation experts. For instance, if an annotation phrase contained a strong trigger (e.g., a social role or job title as with *"spokeswoman"* in *"spokeswoman Arlene"*; cf. Figure 1), it was classified as a low-semantic-complexity item even though it was assigned a high inverse document frequency due to the infrequent word *"Arlene"*.

Two experimental groups were formed to study different kinds of textual context. In the *document context* condition the whole newspaper article was shown as annotation example, while in the *sentence context* condition only the sentence containing the annotation phrase was presented. The participants[3] were randomly assigned to one of these groups. We

[1] Constituency parse trees were generated using the OpenNLP TreeBank parser (`http://opennlp.sourceforge.net/`).

[2] We chose the English part of the Reuters RCV2 corpus, a collection of over 400,000 news stories from 1996 and 1997, as the reference corpus for our experiments.

[3] 20 subjects (12 female) with an average age of 24 years (mean = 24, standard deviation (SD) = 2.8) and normal or corrected-to-normal vision capabilities took part in the study. All participants were students with a computing-related study background, with good to very good English language skills (mean = 7.9, SD = 1.2, on a ten-point scale with 1 = "poor" and 10 = "excellent", self-assessed), but without any prior experience in annotation practice and without previous exposure to academic linguistic education.

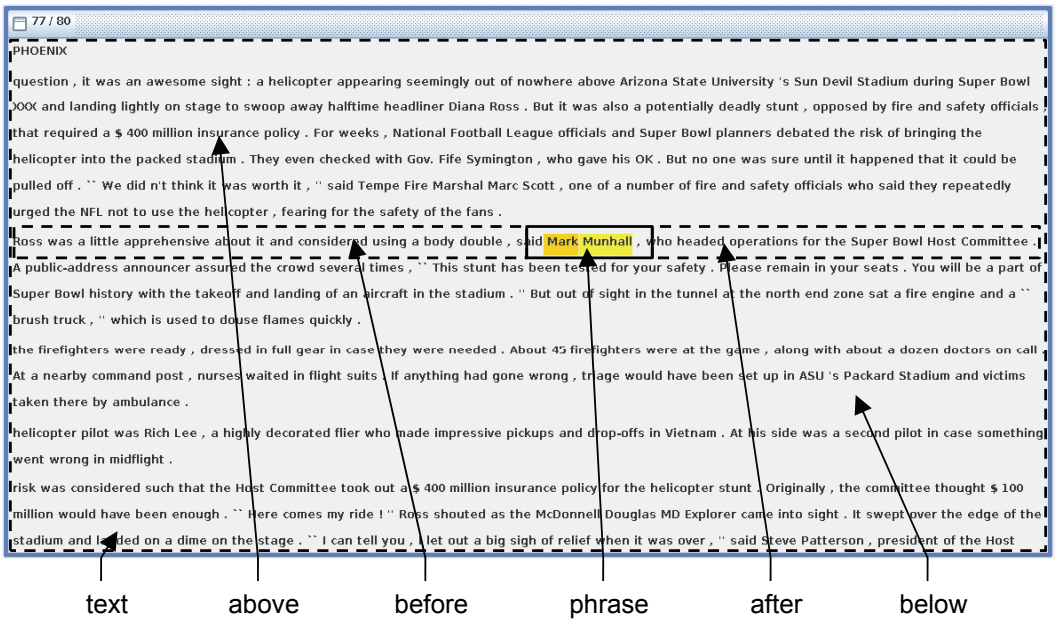


Figure 2: Subareas for the eyetracking analysis. Annotation example is of low semantic and low syntactic complexity.

decided for this between-subjects design to avoid any irritation of the participants caused by constantly changing contexts. Accordingly, the participants were assigned to one of the experimental groups and corresponding context condition already in the second training phase that took place shortly before the experiment started (see below).

Hypotheses and Dependent Variables

We tested the following two hypotheses:

Hypothesis H1: *Annotators perform differently in the two context conditions.*

H1 is based on the linguistically plausible assumption that annotators are expected to make heavy use of the surrounding context because such context could be helpful for the correct disambiguation or de-anaphorization of entity classes. Accordingly, lacking context, an annotator is expected to annotate worse than under the condition of full context. As an adverse effect, the availability of (too much) context might overload and so distract annotators, with a potentially negative effect on annotation performance.

Hypothesis H2: *Annotators' performance is different for varying levels of syntactic and semantic complexity.*

The assumption is that high syntactic or semantic complexity, in contrast to low complexity, for both complexity types significantly lowers the annotation performance.

In order to test these hypotheses we collected data for the following dependent variables: *(a)* the annotation accuracy – we identified erroneous entities by comparison with the original gold annotations in the MUC7 corpus, *(b)* the time needed per annotation example, and *(c)* the distribution and duration of the participants' eye gazes.

Stimulus Material

According to the above definition of complexity, we automatically preselected annotation examples characterized by either a low or a high degree of semantic and syntactic complexity. After manual fine-tuning of the example set assuring an even distribution of entity types and syntactic correctness of the automatically derived annotation phrases, we finally selected 80 annotation examples for the experiment. These were divided into four subsets of 20 examples each falling into one of the following complexity classes:

sem-syn	low semantic – low syntactic complexity
SEM-syn	high semantic – low syntactic complexity
sem-SYN	low semantic – high syntactic complexity
SEM-SYN	high semantic – high syntactic complexity

Experimental Apparatus and Procedure

The annotation examples were presented in a custom-built tool and its user interface was kept as simple as possible not to distract the eye movements of the participants. It merely contained one frame showing the text of the annotation example, with the annotation phrase highlighted (as with *"Mark Munhall"* in Figure 2). A blank screen was shown after each annotation example to reset the eyes and to allow for a break, if needed. The time the blank screen was shown was not counted as annotation time. The 80 annotation examples were presented to all participants in the same randomized order, with a balanced distribution of the complexity classes. A variation of the order was hardly possible for technical and analytical reasons but is not considered as a drawback due to extensive, pre-experimental training (see below). The limitation to 80 annotation examples reduced the chances of errors due to fatigue or lack of attention that can be observed in long-lasting annotation sessions.

subareas	above	left	*phrase*	right	below
percentage of participants looking at a subarea	35	32	100	34	16
average number of fixations in a subarea per participant	2.2		14.1		1.3

Table 1: Distribution of annotators' attention among sub-areas per annotation example.

Five introductory examples (not considered in the final evaluation) were given to get the subjects used to the experimental environment. All annotation examples were chosen in a way that they completely fitted on the screen (text length was limited) to avoid the need for scrolling and thus eye distraction. The contextual position of the CNP was randomly distributed, excluding the first and last sentence.

The participants used a standard keyboard to assign the entity types for each word of the annotation example. All but 5 keys were removed from the keyboard to avoid extra eye movements for finger coordination (three keys for the positive entity classes, one for the "no entity" class, and one to confirm the annotation). Pre-tests had shown that the participants could easily issue the annotations without looking down at the keyboard.

We recorded each participant's eye movements on a Tobii T60 eyetracking device which is invisibly embedded in a 17" TFT monitor and comparatively tolerant to head movements. The participants were seated in a comfortable position with their head in a distance of 60-70 cm from the monitor. Screen resolution was set to 1280 x 1024 px and the annotation examples were presented in the middle of the screen in a font size of 16 px and a line spacing of 5 px. The presentation area had no fixed height and varied depending on the context condition and length of the newspaper article. The text was always vertically centered on the screen.

All participants were familiarized with the annotation task and the guidelines in a pre-experimental workshop where about 60 minutes were spent on annotation exercises. During the next two days, the actual experiments were conducted, each one lasting between 15 and 30 minutes, including calibration of the eye-tracking device. Another 20-30 minutes of training time directly preceded each individual experiment. After the experiment, the participants were interviewed and asked to fill out a questionnaire. Overall, the experiment took about two hours for each participant for which they were financially compensated. The participants were also instructed to focus more on annotation accuracy than on annotation time as we wanted to avoid random guessing. Accordingly, as an extra incentive, we rewarded the three participants with the highest annotation accuracy with cinema vouchers. None of the participants reported serious difficulties with either the newspaper articles or the annotation tool and all subjects agreed that they understood the annotation task very well.

Results

We used a mixed-design analysis of variance (ANOVA) model to test the hypotheses, with the context condition as between-subjects factor and the two complexity classes as within-subject factors.

Testing Context Conditions

To test hypothesis H1 we compared the number of annotation errors on entity-critical words made by the annotators in the two contextual conditions (complete document *vs.* sentence). Surprisingly, on the total of 174 entity-critical words within the 80 annotation examples, we found exactly the same mean value of 30.8 errors per participant in both conditions. There were also no significant differences on the average time needed to annotate an example in both conditions (means of 9.2 and 8.6 seconds, respectively, with $F(1,18) = 0.116$, $p = 0.74$).[4] These results seem to suggest that it makes no difference (neither for annotation accuracy nor for time) whether or not annotators are shown textual context that contains the annotation phrase beyond the sentence.

To further investigate this finding we analyzed the eye-tracking data of the participants gathered for the document context condition. We divided the whole text area into several subareas as shown in Figure 2. We then determined the average proportion of participants that directed their gaze at least once at these subareas. We considered all fixations with a minimum duration of 100 ms, using a fixation radius (i.e., the smallest distance that separates fixations) of 30 px and excluded the first second as it was mainly used for orientation and identification of the annotation phrase.

Table 1 reveals that on average only 35% of the participants looked in the textual context above the annotation phrase embedding sentence, and even less perceived the context below (16%). The sentence parts before and after the annotation phrase were, on the average, visited by one third (32% and 34%, respectively) of the participants. The uneven distribution of the annotators' attention becomes even more apparent in a comparison of the total number of fixations on the different text parts (see Table 1): 14 out of an average of 18 fixations per example were directed at the annotation phrase and the surrounding sentence, the text context above the annotation chunk received only 2.2 fixations on the average and the text context below only 1.3.

Thus, eye-tracking data indicates that the textual context is not as important as might have been expected for quick and accurate annotation. This result can be explained by the fact that participants of the document-context condition used the context whenever they thought it might help, whereas participants of the sentence-context condition spent more time thinking about a correct answer, overall with the same result.

[4] In general, we observed a high variance in the number of errors and time values between the subjects. While, e.g., the fastest participant handled an example in 3.6 seconds on the average, the slowest one needed 18.9 seconds; concerning the annotation errors on the 174 entity-critical words, these ranged between 21 and 46 errors.

experimental condition	complexity class	e.-c. words	time mean	time SD	errors mean	errors SD	errors rate
document condition	sem-syn	36	4.0s	2.0	2.7	2.1	.075
	SEM-syn	25	9.2s	6.7	5.1	1.4	.204
	sem-SYN	51	9.6s	4.0	9.1	2.9	.178
	SEM-SYN	62	14.2s	9.5	13.9	4.5	.224
sentence condition	sem-syn	36	3.9s	1.3	1.1	1.4	.031
	SEM-syn	25	7.5s	2.8	6.2	1.9	.248
	sem-SYN	51	9.6s	2.8	9.0	3.9	.176
	SEM-SYN	62	13.5s	5.0	14.5	3.4	.234

Table 2: Average performance values for the 10 subjects of each experimental condition and 20 annotation examples of each complexity class: number of entity-critical (e.-c.) words, mean annotation time and standard deviations (SD), mean annotation errors, standard deviations, and error rates (number of errors divided by number of entity-critical words).

Testing Complexity Classes

To test hypothesis H2 we also compared the average annotation time and the number of errors on entity-critical words for the complexity subsets (see Table 2). The ANOVA results show highly significant differences for both annotation time and errors.[5] A pairwise comparison of all subsets in both conditions with repeated *t*-test measurements showed non-significant results only between the SEM-syn and syn-SEM subsets.[6] Thus, the empirical data generally supports hypothesis H2 in that the annotation performance seems to correlate with the complexity of the annotation phrase, on the average.

Context and Complexity

We also examined whether the need for inspecting the context increases with the complexity of the annotation phrase. So we analyzed the eye-tracking data in terms of the average number of fixations on the annotation phrase and on its embedding contexts for each complexity class (see Table 3). The values illustrate that while the number of fixations on the annotation phrase rises generally with both the semantic and the syntactic complexity, the number of fixations on the context rises only with semantic complexity. The number of fixations on the context is nearly the same for the two subsets with low semantic complexity (sem-syn and sem-SYN, with 1.0 and 1.5), while it is significantly higher for the two subsets with high semantic complexity (5.6 and 5.0), independent of the syntactic complexity.[7] These results suggest that the need for context mainly depends on the semantic complexity of the annotation phrase, while it is less influenced by its syntactic complexity.

This finding is qualitatively supported by gaze plots we generated from the eye-tracking data. Figure 3 shows such a plot for one participant which illustrates a scanning-for-coreference behavior we observed for many annotation phrases with high semantic complexity. Words were searched in the upper context, which according to their orthographic appearance might refer to a named entity, but which could not fully be resolved only relying on the information given by the annotation phrase itself and its embedding sentence. This is the case for *"Roselawn"* in the annotation phrase *"Roselawn accident"*. The context reveals that Roselawn, which also occurs in the first sentence, is a location. A similar procedure is also performed for acronyms and abbreviations which cannot be resolved from the immediate local context. As indicated by the gaze movements, it also became apparent that texts were rather scanned for hints instead of being deeply read.

complexity class	fixation on phrase mean	fixation on phrase SD	fixation on context mean	fixation on context SD
sem-syn	4.9	4.0	1.0	2.9
SEM-syn	8.1	5.4	5.6	5.6
sem-SYN	18.1	7.7	1.5	2.0
SEM-SYN	25.4	9.3	5.0	4.1

Table 3: Average number of fixations on the annotation phrase and context for the document condition and 20 annotation examples of each complexity class.

Summary and Conclusions

We explored the use of eye-tracking technology to investigate the behavior of human annotators during the assignment of three types of named entities – persons, organizations and locations – based on the eye-mind assumption. We tested two main hypotheses: one relating to the amount of contextual information being used for annotation decisions, the other relating to different degrees of syntactic and semantic complexity of expressions that had to be annotated. We found experimental evidence that the textual context is searched for decision making on assigning semantic meta-data at a surprisingly low rate (with the exception of tackling high-complexity semantic cases and resolving co-references) and that annotation performance highly correlates with semantic complexity and to a lesser degree with syntactic complexity.

[5] Annotation time results: $F(1,18) = 25$, $p < 0.01$ for semantic complexity and $F(1,18) = 76.5$, $p < 0.01$ for syntactic complexity; Annotation complexity results: $F(1,18) = 48.7$, $p < 0.01$ for semantic complexity and $F(1,18) = 184$, $p < 0.01$ for syntactic complexity.

[6] $t(9) = 0.27$, $p = 0.79$ for the annotation errors in the document context condition, and $t(9) = 1.97$, $p = 0.08$ for the annotation time in the sentence context condition.

[7] ANOVA result of $F(1,19) = 19.7$, $p < 0,01$ and significant differences also in all pairwise comparisons.

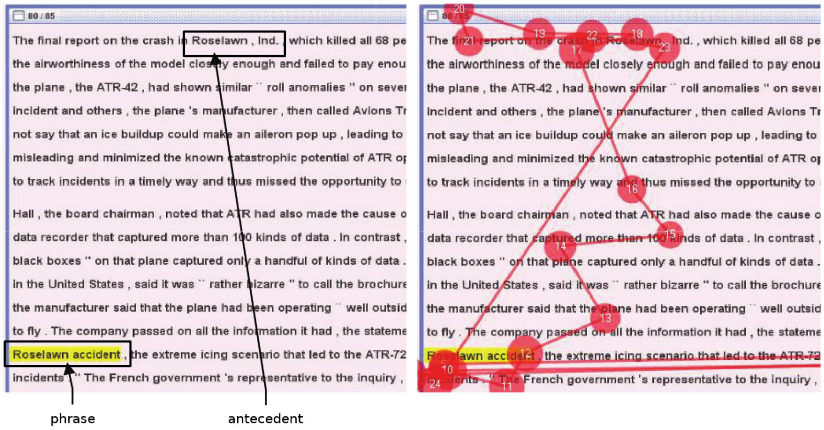


Figure 3: Annotation example with annotation phrase and the antecedent for *"Roselawn"* in the text (left), and gaze plot of one participant showing a scanning-for-coreference behavior (right).

The results of these experiments can be taken as a heuristic clue to focus on cognitively plausible features of learning empirically rooted cost models for annotation (see Tomanek, Lohmann, Ziegler, et Hahn (2010) for more details).

Références

Altmann, G., Garnham, A., & Dennis, Y. (2007). Avoiding the garden path: Eye movements in context. *Journal of Memory and Language*, *31*(2), 685–712.

Arora, S., Nyberg, E., & Rosé, C. (2009). Estimating annotation cost for active learning in a multi-annotator environment. In *NAACL HLT Workshop on Active Learning for Natural Language Processing* (pp. 18–26).

Cheung, H., & Kemper, S. (1992). Competing complexity metrics and adults' production of complex sentences. *Applied Psycholinguistics*, *13*, 53–76.

Cohn, D., Ghahramani, Z., & Jordan, M. (1996). Active learning with statistical models. *Journal of Artificial Intelligence Research*, *4*, 129–145.

Frazier, L., & Rayner, K. (1987). Resolution of syntactic category ambiguities: Eye movements in parsing lexically ambiguous sentences. *Journal of Memory and Language*, *26*, 505–526.

Hachey, B., Alex, B., & Becker, M. (2005). Investigating the effects of selective sampling on the annotation task. In *CoNLL 2005 – 9th Conference on Computational Natural Language Learning* (pp. 144–151).

Linguistic Data Consortium. (2001). *Message Understanding Conference (MUC) 7*.

Marcus, M., Santorini, B., & Marcinkiewicz, M. (1993). Building a large annotated corpus of English: PENN TREEBANK. *Computational Linguistics*, *19*, 313–330.

Palmer, M., Gildea, D., & Kingsbury, P. (2005). The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, *31*(1), 71–106.

Pradhan, S., Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., & Weischedel, R. (2007). Ontonotes: A unified relational semantic representation. In *ICSC 2007 – International Conf. on Semantic Computing* (pp. 517–526).

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *126*, 372–422.

Rayner, K., Cook, A., Juhas, z. B., & Frazier, L. (2006). Immediate disambiguation of lexically ambiguous words during reading: Evidence from eye movements. *British Journal of Psychology*, *97*, 467–482.

Ringger, E., Carmen, M., Haertel, R., Seppi, K., Lonsdale, D., McClanahan, P., et al. (2008). Assessing the costs of machine-assisted corpus annotation through a user study. In *LREC 2008 – 6th International Conference on Language Resources and Evaluation* (pp. 3318–3324).

Settles, B., Craven, M., & Friedland, L. (2008). Active learning with real annotation costs. In *NIPS 2008 Workshop on Cost-Sensitive Machine Learning* (pp. 1–10).

Sturt, P. (2007). Semantic re-interpretation and garden path recovery. *Cognition*, *105*, 477–488.

Szmrecsányi, B. M. (2004). On operationalizing syntactic complexity. In *JADT 2004 – 7th International Conf. on Textual Data Statistical Analysis* (pp. 1032–1039).

Tomanek, K., & Hahn, U. (2010). Annotation time stamps: Temporal metadata from the linguistic annotation process. In *LREC 2010 – 7th International Conference on Language Resources and Evaluation.*

Tomanek, K., Lohmann, S., Ziegler, J., & Hahn, U. (2010). A cognitive cost model of annotations based on eye-tracking data. In *ACL 2010 – 48th Annual Meeting of the Association for Computational Linguistics.*

Tomanek, K., Wermter, J., & Hahn, U. (2007). An Approach to Text Corpus Construction which Cuts Annotation Costs and Maintains Reusability of Annotated Data. In *EMNLP/CoNLL 2007 – Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (pp. 486–495).

Traxler, M., & Frazier, L. (2008). The role of pragmatic principles in resolving attachment ambiguities: Evidence from eye movements. *Memory & Cognition*, *36*, 314–328.

Subject-Object Asymmetries in Korean Sentence Comprehension

Jiwon Yun (jy249@cornell.edu)
John Whitman (jbw2@cornell.edu)
John Hale (jthale@cornell.edu)
Department of Linguistics, 203 Morrill Hall, Cornell University
Ithaca, NY 14850 USA

Abstract

The Entropy Reduction Hypothesis (Hale, 2006) derives the subject-object asymmetry in Korean relative clauses. This asymmetry has been observed by Kwon, Polinsky, and Kluender (2006), among others. Agreement between the Entropy Reduction predictions and the available empirical data suggests that the heightened comprehension difficulty attested in object-extracted relatives is due to distinctive incremental parser states associated with comparatively greater temporary ambiguity.

Keywords: sentence comprehension, relative clauses, Korean, probabilistic grammar, Entropy Reduction, syntax

Introduction

Relative clauses (RCs) have long been objects of fascination for cognitive scientists interested in language comprehension (Kaplan, 1974). In the well-known "subject-extracted" (SRC) and "object-extracted" (ORC) cases, a large literature exists. In languages such as English and French, a processing advantage for SRCs has been confirmed in a wide variety of measures including phoneme-monitoring (Frauenfelder, Segui, & Mehler, 1980), eye-fixations (Holmes & O'Regan, 1981), reading times (King & Just, 1991), PET (Stromswold, Caplan, Alpert, & Rauch, 1996) and fMRI (Just, Carpenter, Keller, Eddy, & Thulborn, 1996). It has been suggested that the SRC advantage may be a processing universal (Lin, 2008). If ORCs are harder than SRCs in all languages, then what is it about human sentence comprehension that makes this so? The Korean language is a key test for any universal processing theory because it is syntactically different from English and French. These differences include verb-final clauses and prenominal RCs.

In this paper, we offer an account of the SRC/ORC asymmetry in terms of the information-processing difficulty of incremental parsing in general. This proposal relates the hardness of parsing to syntactic facts about Korean. A language independent complexity metric known as Entropy Reduction (Wilson & Carroll, 1954; Hale, 2003, 2006) correctly derives the SRC advantage when applied with a Korean grammar. This demonstration supports the claim that human comprehension difficulty reflects the kind of information-processing work that Entropy Reduction quantifies. [1]

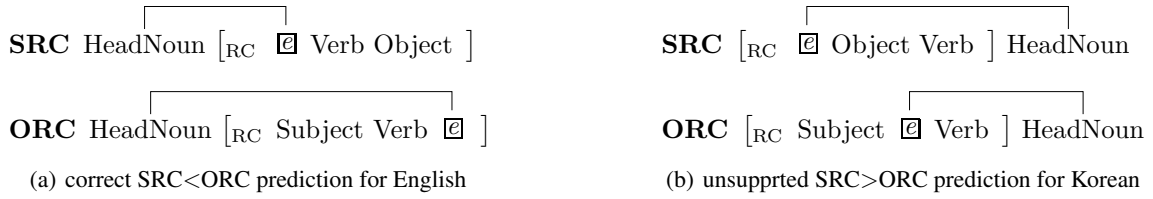
[1] A longer companion paper, Hale (under review), develops an automaton model of the sentence comprehension process. It presents a generalized left-corner parser that operates in accordance with the Entropy Reduction Hypothesis when its decisions about how to resolve nondeterminism are guided by experience.

Theories of the Subject-Object Asymmetry

As an empirical phenomenon, the SRC/ORC processing asymmetry is well-established. However, its implications for the architecture or mechanisms of human language comprehension remain controversial. Three broad classes of theory have been advanced. LINEAR DISTANCE theories, illustrated in Figure 1, point to a greater number of intervening elements between the relativized position and the headnoun to which it is meaningfully related. The boxed *e* notation stands for an "empty" element. Particular theories of LINEAR DISTANCE offer alternative ways of measuring the separation between this omitted position and the headnoun (Wanner & Maratsos, 1978; Gibson, 2000; Lewis & Vasishth, 2005). These theories all provide an adequate account of the English pattern, and in some cases relate this prediction to plausible mechanisms of human sentence comprehension. They are thwarted, however by data that confirm an SRC-over-ORC processing advantage in Korean (O'Grady, Lee, & Choo, 2003; Kwon et al., 2006; Lee, 2007). Figure 1(b) shows how theories of this type derive the wrong prediction for Korean.

The second broad class includes STRUCTURAL DISTANCE theories. The simplest theory of this kind maintains that ORCs are harder because the relativized element is more deeply embedded when it is an object. If ORCs are formed by a movement rule, then this movement would "cross" both a VP node and an S node to arrive at its surface position (O'Grady, 1997, 179). Hawkins (2004, 175) singles-out "a connected path that must be accessed for gap identification and processing." Hawkins' path is shown using dotted branches in Figure 2. This path is shorter for SRCs in both Korean and English. This general account is thus adequate but not very precise. It leaves open, for instance, the question of where exactly greater difficulty should start to accrue during incremental processing.

The third broad class contains the INFORMATION-THEORETICAL approaches. The Entropy Reduction Hypothesis (ERH) fits into this class. It holds that a person's difficulty at a word reflects the amount by which that word helped him or her to ascertain which construction the speaker intends. The ERH uses the concept of entropy to quantify the average uncertainty about derivations consistent with an observed initial string. This entropy is high when there are many equiprobable continuations and low when there are just a few continuations or the probability distribution on them is sharply concentrated. This quantity standsin for the degree of confusion in the comprehender's mind. When it is reduced in the transition from one word to the next, the comprehender



(a) correct SRC<ORC prediction for English (b) unsupprted SRC>ORC prediction for Korean

Figure 1: Predictions of LINEAR DISTANCE

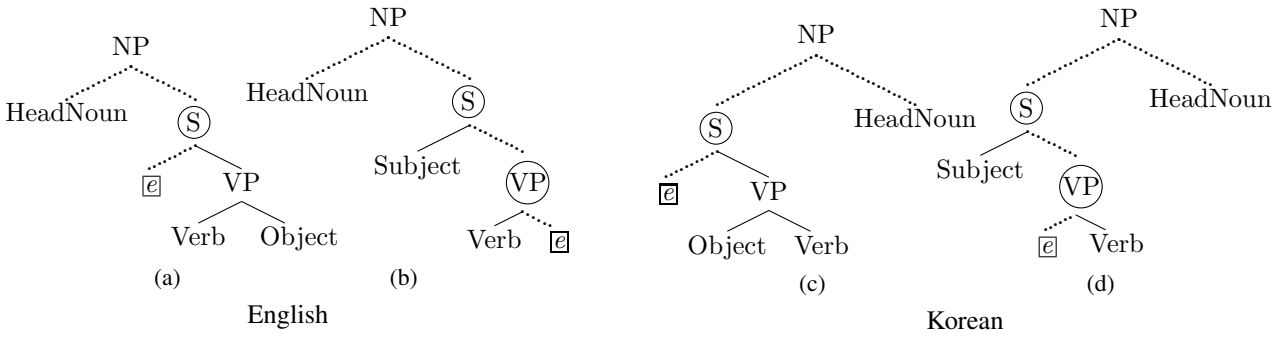


Figure 2: Predictions of STRUCTURAL DISTANCE. In ORCs, ((b),(d)) the pathway between e and HeadNoun crosses two circled nodes whereas in SRCs it crosses just one ((a),(c)). This asymmetry makes the right prediction in both languages.

has accomplished disambiguation work. The ERH interprets this theoretical work as a word-by-word metric of incremental comprehension difficulty.

Hale (2006) derives Entropy Reduction predictions for English relative clauses. Asymmetries between them suggest that relativized non-subjects are harder to comprehend because of greater temporary ambiguity at the embedded verb. While it is well-known that Korean exhibits considerable temporary ambiguity in the middle of sentences, precise levels have not been compared across constructions. Figure 3 illustrates this ambiguity with a prefix string that could signal at least four different clause-types. The ERH offers the possibility of accounting for the SRC/ORC asymmetry in terms of contrasting levels of such ambiguity.

Procedure

We calculate Entropy Reductions at every inter-word point in Korean SRC and ORC sentences using a procedure that mirrors Hale (2006). One of us (JY) prepared a Korean grammar that covers the sentences listed in the Appendix. This grammar is written in Stabler's Minimalist Grammars (MG) formalism (Stabler, 1997). This transformational formalism adopts certain themes of Chomsky's Minimalist Program (1995) and has been shown to be mildly context-sensitive in the sense of Joshi (1985) by Michaelis (2001). We consider subject-extraction and object-extraction in each of the four clause-types shown in Figure 3. Our analysis supposes that the headnoun moves in relativization. We use the MG *move* rule to implement this analysis. Figure 4 shows a structural description generated by this grammar. This grammar analyzes postnominal case markers as separate words and verb suffixes as part of verbs. Here, a coindexed trace, *t(3)* indicates movement of the headnoun *kica* 'reporter' from its base position in a specifier of little *v* to a position outside the RC. Weighting each construction type listed in the Appendix by its attestation count in a Korean Treebank (Han et al., 2006), we estimate a probabilistic context-free grammar (PCFG) of MG derivations. By chart parsing, we recover a new PCFG for each prefix of the sentences of interest. This chart-PCFG is an alternative presentation of the AND-OR graph encoded by the chart (Lang, 1991). It represents all possible analyses that are consistent with the given prefix. We calculate the entropy of the start symbol of this chart-PCFG to arrive at the conditional entropy of the prefix string. This value is a cognitive model of an incremental comprehender's degree of confusion about which construction he or she is in. When it goes down, disambiguation work has occurred.

Results

Table 1 summarizes the ERH predictions: SRCs are easier to comprehend than ORCs. This prediction also follows in noun complement clauses. However, empty elements in subject position are not always easier. In simple matrix clauses and adjunct clauses, no difference is predicted.

Clause type	SBJ Extraction	OBJ Extraction
Matrix Clause	19.6	19.6
Adjunct Clause	34.66	34.66
Complement Clause	32.1	42.98
Relative Clause	27.13	35.65

Table 1: Average Entropy Reduction in bits-per-word

matrix clause ⓔ ***uywon ul*** *kongkyekhayssta* pro senator ACC attack-DECL '(someone) attacked the senator.'	complement clause ⓔ ***uywon ul*** *kongkyekhan sasil* pro senator ACC attack-ADN fact 'the fact that (someone) attacked the senator'
adjunct clause ⓔ ***uywon ul*** *kongkyekhayese* pro senator ACC attack-ADV 'because (someone) attacked the senator,'	relative clause ⓔ ***uywon ul*** *kongkyekhan kica* gap senator ACC attack-ADN reporter 'the reporter who attacked the senator'

Figure 3: The same initial morphemes signal at least four different clause types[2]

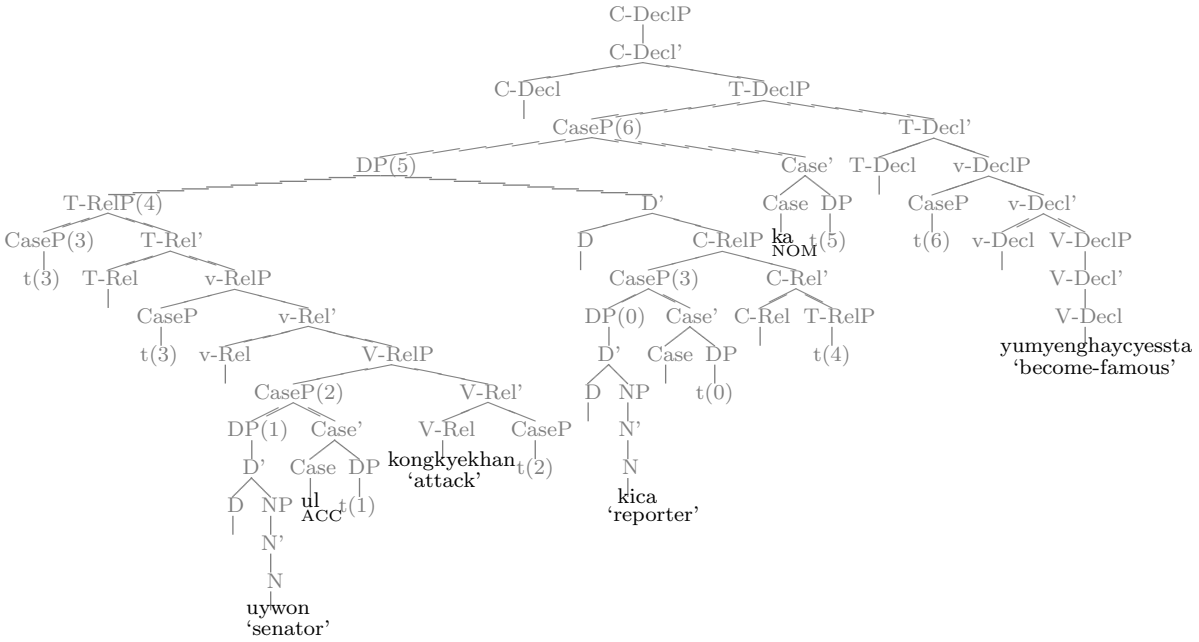


Figure 4: Structural description of SRC example (d) from the Appendix

Word-by-word Entropy Reduction graphs, shown in Figure 6, illustrate how predicted difficulty peaks coincide with the positions that disambiguate clause-type and the role of omitted elements. This is indicated with double-circles in Figure 5. The subject-object asymmetry in RCs is predicted to show up on the headnoun at the position marked N in Figure 6(d). This prediction matches the findings of Kwon et al. (2006), who observe a reading time asymmetry at this point.

Discussion

The Entropy Reduction account of the subject advantage in relative clauses and complement clauses is rooted in contrasting levels of uncertainty about syntactic structure. The crucial position, immediately after the adnominal form of the verb, is marked ❸ in Figure 7. In the ORC case, the conditional entropy at this point is 32.28 bits, while in the SRC case, the corresponding value is only 23.76 bits. The conditional entropy values at ④ are exactly the same — 17.43 bits in both cases. Thus, the ERH models the greater difficulty in the object cases with greater conditional entropy at point ❸.

The disparity between these conditional entropies reflects contrasting numbers of alternative continuations. These continuations correspond to different roles the prefix string might play at the matrix level. Figure 8 shows that the ORC prefix **N NOM V-ADN** could be in fact the beginning of a reading on which the nominative-marked noun is a complete matrix-level subject on its own, where both the subject and the object of the embedded clause are omitted. These properties allows the ORC prefix to have the multiple parses shown in (1-3) below. The disparity derives, ultimately, from syntactic properties of Korean. As we have seen, it is an SOV language with prenominal RCs; crucially, arguments may be freely omitted when they are recoverable in-context. Such additional structures are not acceptable as a continuation of the SRC prefix **N ACC V-ADN**, which cannot be split by additional empty categories.

(1) *kica* *ka* [SRC ⓔ ⓔ *kongkyekhan*] *uywon* *ul*
reporter NOM gap pro attack-ADN senator ACC
manassta.
meet-DECL

[2]Our notational conventions include NOM for nominative case, ACC for accusative, ADV for adverbial, ADN for adnominal and DECL for declarative.

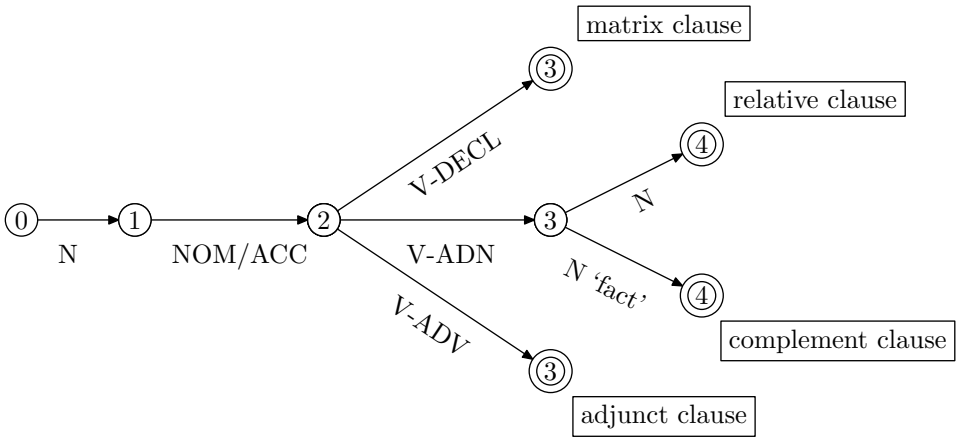
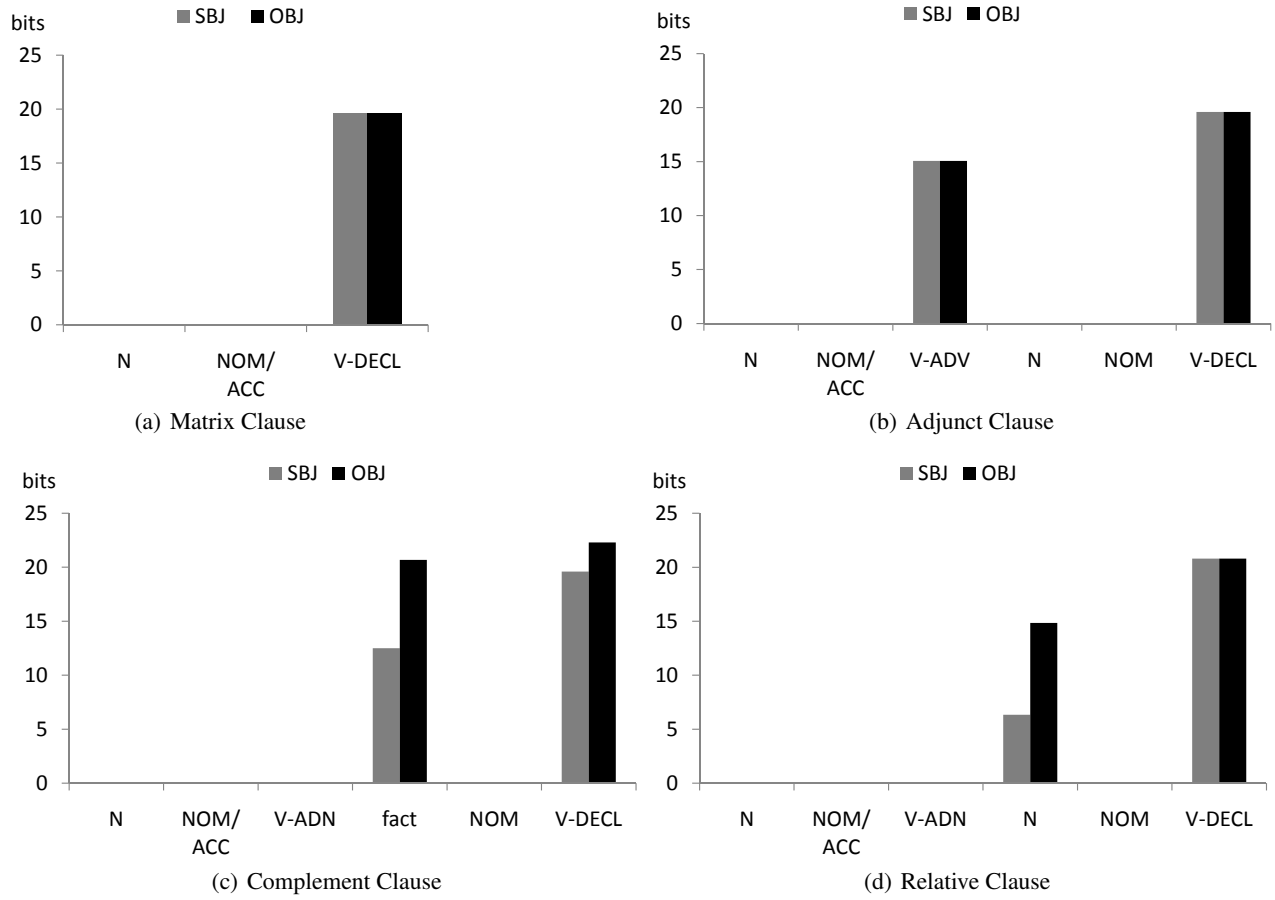


Figure 5: Continuations signal clause-types



(a) Matrix Clause

(b) Adjunct Clause

(c) Complement Clause

(d) Relative Clause

Figure 6: Word-by-word comprehension difficulty predictions derived by the INFORMATION-THEORETICAL Entropy Reduction Hypothesis. Horizontal axes labels name word classes. SBJ abbreviates "subject-extracted", OBJ "object-extracted". Clause-types (a)–(d) are as in Figure 3.

SRC *kica* ① *lul* ② *kongkyekhan* ❸ *uywon* ④
reporter ACC attack-ADN senator
'the senator who attacked the reporter'

ORC *kica* ① *ka* ② *kongkyekhan* ❸ *uywon* ④
reporter NOM attack-ADN senator
'the senator who the reporter attacked'

Figure 7: SRC and ORC. The black circle indicates where the difference of structural uncertainty is observed.

SRC a. [e *kica lul kongkyekhan*]
[SBJ OBJ V-ADN]

ORC a. [*kica ka* e *kongkyekhan*]
[SBJ OBJ V-ADN]
b. *kica ka* [e e *kongkyekhan*]
[SBJ OBJ V-ADN]

Figure 8: Alternative syntactic roles for elements of the two prefix strings. Brackets indicate embedded clauses.

'The reporter met the senator who attacked (someone).'

(2) *kica ka* [ORC e e *kongkyekhan*] *uywon ul*
reporter NOM pro gap attack-ADN senator ACC
manassta.
meet-DECL
'The reporter met the senator whom (someone) attacked.'

(3) *kica ka* [CC e e *kongkyekhan*] *sasil ul*
reporter NOM pro pro attack-ADN fact ACC
alkoissta.
know-DECL
'The reporter knows the fact that (someone) attacked (someone).'

Related work

These results offer a new perspective on the work of Ishizuka, Nakatani, and Gibson (2006). Using Japanese RCs, which are structurally similar to Korean, these authors show that the penalty for ORC processing can be mitigated or even eliminated if certain readings are pragmatically suppressed by prior discourse. The ERH suggests that disambiguating those readings is exactly the source of the ORC penalty. It quantifies the difficulty of coping with all the available alternatives.

Our results also suggest a lack of subject-object asymmetry in adjunct clauses. We would like to emphasize that this does not entail a contradiction with the experimental results of Kwon et al. (2006). The design of this experiment leverages that fact that a matrix clause noun is a felicitous controller of *pro* when it appears in an embedded clause. Indeed, these authors suggest that "the identification of the gap in an adjunct clause does not involve any syntactic operations." It is thus appropriate that our syntax-only approach predicts no distinction between missing subjects and objects in this clause type. The ERH might naturally be combined with a pragmatic component to yield a broader theory. We leave this extension to future work.

Conclusion

The ERH, in conjunction with an appropriate formal grammar, can account for the subject advantage in Korean RCs. Its predictions cannot be summarized by simply saying that missing objects are always harder; for instance both types of main clauses are predicted to be equally easy. However they do include the prediction of a subject-object asymmetry in complement clauses with omitted arguments. The effect should appear on the word *sasil* 'fact'. This prediction would not follow on a STRUCTURAL DISTANCE account, since no movement relation exists between the empty element *pro* and *sasil* in that construction. If a subject-object asymmetry were to be experimentally observed at that point, this would leave the ERH as the only theory able to explain the English as well as the Korean results. We hope that our work encourages empirical investigation of this case.

Acknowledgment

The authors would like to express their appreciation to Nayoung Kwon for her comments on the manuscript. Thanks are due as well to Na-Rae Han for information about the Korean Treebank. This research was supported by a Small Grant from the Cornell University Institute for the Social Sciences.

Appendix: Examples

The Minimalist Grammar used to derive the comprehension-difficulty predictions graphed in Figure 6 covers all of the examples listed below. The combinatorics of the promotion analysis imply the existence of other grammatical strings such as the examples (1)–(3) in discussion.

a. matrix clause with a *pro*-subject

uywon ul kongkyekhayssta.
senator ACC attack-DECL

'Someone attacked the senator.'

b. adjunct clause with a *pro*-subject

uywon ul kongkyekhayse kica ka
senator ACC attack-ADV reporter NOM
yumyenghaycyessta.
become-famous-DECL

'Because someone/he attacked the senator, the reporter became famous.'

c. complement clause with a *pro*-subject

uywon ul kongkyekhan sasil i palkhyecyessta.
senator ACC attack-ADN fact NOM is-revealed-DECL

'The fact that someone attacked the senator was revealed.'

d. subject relative clauses

uywon ul kongkyekhan kica ka
senator ACC attack-ADN reporter NOM
yumyenghaycyessta.
become-famous-DECL

'The reporter who attacked the senator became famous.'

e. matrix clause with a *pro*-object

kica ka kongkyekhayssta.
reporter NOM attack-DECL

'The reporter attacked someone.'

f. adjunct clause with a *pro*-object

kica ka kongkyekhayse uywon i
reporter NOM attack-ADV senator NOM
yumyenghaycyessta.
become-famous-DECL

'Because the reporter attacked someone/him, the senator became famous.'

g. complement clause with a *pro*-object

kica ka kongkyekhan sasil i palkhyecyessta.
reporter NOM attack-ADN fact NOM is-revealed-DECL

'The fact that the reporter attacked someone was revealed.'

h. object relative clauses

kica ka kongkyekhan uywon i
reporter NOM attack-ADN senator NOM
yumyenghaycyessta.
become-famous-DECL

'The senator whom the reporter attacked became famous.'

References

Chomsky, N. (1995). *The Minimalist Program*. Cambridge, Massachusetts: MIT Press.

Frauenfelder, U., Segui, J., & Mehler, J. (1980). Monitoring around the relative clause. *Journal of Verbal Learning and Verbal Behavior*, *19*(3), 328 - 337.

Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. In Y. Miyashita, A. Marantz, & W. O'Neil (Eds.), *Image, language, brain* (pp. 95–126). Cambridge, Massachusetts: MIT Press.

Hale, J. (2003). The information conveyed by words in sentences. *Journal of Psycholinguistic Research*, *32*(2), 101–123.

Hale, J. (2006). Uncertainty about the rest of the sentence. *Cognitive Science*, *30*, 643-672.

Hale, J. (under review). *What a rational parser would do.*

Han, N.-R., Ryu, S., Chae, S.-H., Yang, S., Lee, S., & Palmer, M. (2006). Korean treebank annotations version 2.0 [Computer software manual].

Hawkins, J. A. (2004). *Efficiency and complexity in grammars*. Oxford University Press.

Holmes, V. M., & O'Regan, J. K. (1981). Eye fixation patterns during the reading of relative-clause sentences. *Journal of Verbal Learning and Verbal Behavior*, *20*(4), 417–430.

Ishizuka, T., Nakatani, K., & Gibson, E. (2006). *Processing Japanese relative clause in context.* Paper presented at the 19th Annual CUNY Conference on Human Sentence Processing.

Joshi, A. K. (1985). Tree adjoining grammars: How much context-sensitivity is required to provide reasonable structural descriptions? In D. Dowty, L. Karttunen, & A. M. Zwicky (Eds.), *Natural language parsing: Psychological, computational and theoretical perspectives* (pp. 206–250). New York: Cambridge University Press.

Just, M. A., Carpenter, P. A., Keller, T. A., Eddy, W. F., & Thulborn, K. R. (1996). Brain activation modulated by sentence comprehension. *Science*, *274*(5284), 114–116.

Kaplan, R. M. (1974). *Transient processing load in relative clauses*. Unpublished doctoral dissertation, Harvard University.

King, J., & Just, M. A. (1991). Individual differences in syntactic processing: the role of working memory. *Journal of Memory and Language*, *30*(5), 580–602.

Kwon, N., Polinsky, M., & Kluender, R. (2006). Subject preference in Korean. In D. Baumer, D. Montero, & M. Scanlon (Eds.), *Proceedings of the 25th west coast conference on formal linguistics (WCCFL 25)* (p. 1-14). Somerville, MA: Cascadilla Press.

Lang, B. (1991). Towards a uniform formal framework for parsing. In *Current issues in parsing technology* (p. 153-171). Kluwer Academic Publishers.

Lee, C.-K. (2007). *Relative-clause processing in Korean adults: effects of constituent order and prosody*. Unpublished master's thesis, Rutgers University.

Lewis, R. L., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, *29*(3), 375–419.

Lin, C.-J. C. (2008). The processing foundation of head-final relative clauses. *Language and Linguistics*, *9*, 813–838.

Michaelis, J. (2001). Derivational minimalism is mildly context-sensitive. In M. Moortgat (Ed.), *Logical aspects of computational linguistics* (pp. 179–198). Springer. (Selected papers from LACL98)

O'Grady, W. (1997). *Syntactic development*. University of Chicago Press.

O'Grady, W., Lee, M., & Choo, M. (2003). A subject-object asymmetry in the acquisition of relative clauses in Korean as a second language. *Studies in Second Language Acquisition*, *25*(3), 433-448.

Stabler, E. (1997). Derivational minimalism. In C. Retoré (Ed.), *Logical aspects of computational linguistics.* Springer-Verlag.

Stromswold, K., Caplan, D., Alpert, N., & Rauch, S. (1996). Localization of syntactic comprehension by positron emission tomography. *Brain and Language*, *52*(3), 452–473.

Wanner, E., & Maratsos, M. (1978). An ATN approach to comprehension. In M. Halle, J. Bresnan, & G. A. Miller (Eds.), *Linguistic theory and psychological reality* (pp. 119–161). Cambridge, Massachusetts: MIT Press.

Wilson, K., & Carroll, J. B. (1954). Applications of entropy measures to problems of sequential structure. In C. E. Osgood & T. A. Sebeok (Eds.), *Psycholinguistics: a survey of theory and research.* Indiana University Press.

The Benefit of Imitating Particular Individuals

Yasuaki Sakamoto (ysakamot@stevens.edu)
Hongyuan Shi (hshi@stevens.edu)
Center for Decision Technologies
Howe School of Technology Management, Stevens Institute of Technology
Hoboken, NJ 07030 USA

Abstract

We examined the benefits of different search strategies by testing four computational models. In one model, agents in a group always innovated. The other three models incorporated some mechanisms of imitation. In the second model, each agent imitated the best solution of a random other. In the third model, each agent followed preferential attachment and imitated the best solution of the agent that was asked by many agents. In the fourth model, each agent developed a familiarity with an agent based on how often it asked a certain agent, and imitated this agent. In two simulation studies, following the most popular or the most familiar agent resulted in a good compromise between efficiency and diversity in finding good solutions. People's desire to follow particular individuals may be a key to their adaptive behavior, allowing them to disseminate ideas efficiently while encouraging the exploration of new ideas.

Keywords: Innovation and imitation; computational modeling; social learning; search.

Introduction

How do we search for information? Some individuals like to innovate. Others like to imitate. We all engage in both. Because we are social beings, we often rely on other's behavior to shape our own behavior. By observing and imitating others, people can entertain solutions that they would not have even considered otherwise (Bandura, 1965). The creation of innovative solutions (Kraatz, 1998), the evolution of language (Smith et al., 2003), and the development of culture (Dennett, 1995) all result from the process of iterated learning, in which people learn from the previous outputs of others.

In the current work, we examine the benefits of different types of search strategies through computer simulation. We know that whereas too much innovation results in poor dissemination of good solutions, too much imitation results in under exploration of good solutions (Gureckis & Goldstone, 2006). A group of people needs to both innovate and imitate to prosper. But when should we innovate and when should we imitate? Who should we observe if we decide to imitate?

When people are unsure about the best solution, they use other's information as an indicator of what is best (Cialdini & Goldstein, 2004; Deutsch, & Gerard, 1995; Festinger, 1954; Sherif, 1935). People also adopt other's information due to their desire to be liked and to not appear deviant (Asch, 1956; Deutsch, & Gerard, 1995). This imitation behavior is consistent with the principle of preferential attachment (Barabási & Albert, 1999), in which people are attracted to already popular solutions. For example, people instantly get in line when they see a long line outside of a cupcake store, assuming that the store must be offering some really good cupcakes. If everyone imitates, however, it will be difficult for the group to find another cupcake store that also serves really good cupcakes. Thus, imitation leads to efficient problem solving when there is a single best solution. When there are multiple good solutions, however, imitation can lead the group to quickly converge to a single solution, under-exploring the others: some people need to explore other possibilities.

For studying innovation and imitation, we used a simple search game, inspired by a recent social learning tournament (http://www.intercult.su.se/cultaptation/tournament.php). In our game, five agents guessed an action value between 0 and 100, and received as feedback the number of points obtained from the guess. A function converted the guess to a payoff. The agents did not know the function and did not try to learn it. They simply stored the guessed action value that was associated with the highest payoff. The payoff distributions are displayed in Figures 1 and 2. In one case, the search space had a single peak at action 80 as shown in Figure 1. In another case, the search space had three peaks at action values 10, 50, and 80 as shown in Figure 2. Although the game may seem overly simple and artificial, it is analogous to many tasks we encounter every day (see Page, 2007).

The five agents, A_1, …, A_5, selected to either innovate (randomly select a value between 0 and 100) or imitate (receive another agent's value with the highest payoff) in turn. Four groups are simulated:

1. **Innovate**: Agents only innovated.
2. **Ask Random**: Agents imitated a randomly selected agent. The preference weight of A_i asking A_j, p_{ij}, was equal for all j.
3. **Ask Majority Preference**: Agents imitated another agent who was imitated by many others. That is p_{ij} was determined by the number of times A_j was asked by other agents, m_j. This group followed the principle of preferential attachment, and conformed to the majority's behavior.
4. **Ask Individual Preference**: Agents imitated another agent based on how often they asked a certain agent:

$$p_{ij} = P_1 + \frac{P_2 - P_1}{1 + \exp[-C(f_{ij} - F)]}$$

where $P_1 = 0$, $P_2 = 10$, $C = 0.2$, and $F = 15$. The ask history, f_{ij}, tracked the number of times A_i imitated A_j. Agents maintained a counter for every other agent it had interaction with. They followed the footstep of a particular agent they became familiar with.

The imitating agent always received another agent's current best solution. That is, the asked agent always returned the action value associated with the highest payoff that it previously guessed. In the current simulation, when asked agents returned worse solution than the existing one (i.e., the imitating agent had a better solution than the one asked), the agent innovated on the next round. Likewise, when asking someone does not result in good solution, humans often explore the environment by themselves. After innovating once, the agent tried to imitate again. Without this innovation round, always imitating can quickly converge to an action value regardless of its payoff.

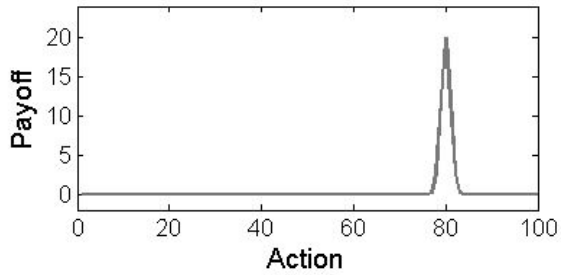


Figure 1: The distribution of payoff in Simulation 1 is shown. There is a single peak at action 80.

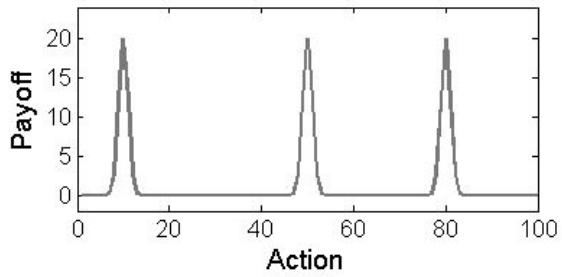


Figure 2: The distribution of payoff in Simulation 2 is shown. There are three peaks at action 10, 50, and 80.

Which group will result in all agents finding the value associated with the highest payoff most efficiently? Imitating others will help disseminate ideas. But which type of asking is best? Humans often conform to the group, similar to the Majority Preference model (e.g., Cialdini & Goldstein, 2004; Sakamoto et al., 2009). They also build familiarity for a particular other, and follow this individual (e.g., Sadlon, et al., 2009; Sakamoto et al., 2008). This following behavior allows us to make near-optimal decision in a limited amount of time in many different social circumstances (Gigerenzer et al., 2000). At the same time, vocal group members often sway the opinion of individuals, and thus the opinions produced by a group may only reflect those of a small subset of the group. Then, imitating others may not be advantageous when there are multiple good solutions to find. In this case, the Innovate group may be successful because the group has no social influence that converges their solutions. Previous work has focused on how given social network structures influence the dissemination of ideas (e.g., Mason & Goldstone, 2008). In the current work, the agent's behavior determines the kinds of social networks built and thus how information is spread within the group.

Simulation Study 1

In Simulation Study 1, the four groups of agents searched for the action with the highest payoff in a space with a single best solution as shown in Figure 1. The agents did not know what the maximum payoff was. Each group had five agents that all followed the same behavioral rule as described previously: innovate, ask random, ask majority preference, or ask individual preference. Each group had 500 cycles to search, each cycle consisting of an agent taking its action. We used 500 cycles so that each group would perform well at the end and we could see the entire course of evolution. Each group was simulated 30 times.

Figures 3 to 6 show the results from the four groups. The left most graph of each figure shows the evolution of total payoff (sum of all agents' payoffs) over the course of 500 cycles, averaged over 30 simulations. During the first 200 cycles, the Innovate model and the Majority Preference model lag behind the Random model and the Individual Preference model. The Innovate model is especially far behind the other models early on. After 300 cycles, the Innovate model is performing the worst, the Random model performing the best, and the two preference models in between. After 400 cycles, the two preference models catch up with the Random model, while the Innovate model is still behind the other models. At 500 cycles, every model has nearly all agents discovering the action with the highest payoff.

The middle three histograms in each figure show the frequency of total payoff for the 30 simulations. At 50 and 100 cycles, the disadvantage of the Innovate model is apparent: no simulation resulted in total payoff of 80 or 100. At 500 cycles almost all 30 simulations for each group result in every agent knowing the best action.

The color map on the right side of each figure shows the evolution of each agent's payoff averaged over 30 simulations. The Innovate model is darker in general, indicating that it took longer to find good solutions than the other models. In addition, the Innovate model has darker horizontal band, indicating that some agents had hard time innovating a good solution. In contrast, the other three models incorporating imitation disseminated the best action efficiently. The Random model was especially quick at disseminating good solutions.

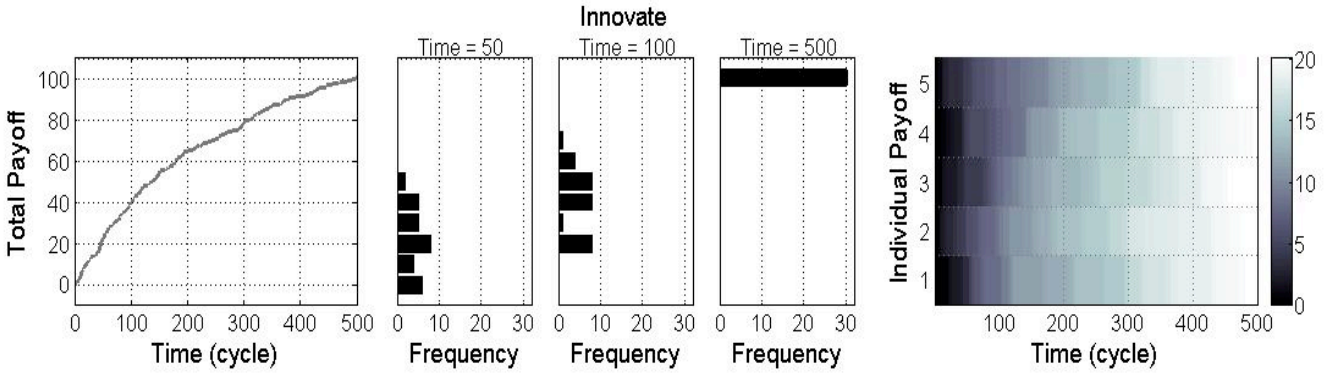


Figure 3: The results from the Innovate model in Simulation Study 1 are shown. The left most graph shows the evolution of total payoff (sum of all agents' payoffs) over the course of 500 trials, averaged across 30 simulations. The middle three histograms show the frequency of total payment for the 30 simulations at 50, 100, and 500 cycles. The color map on the right side shows the evolution of each agent's payoff averaged over 30 simulations.

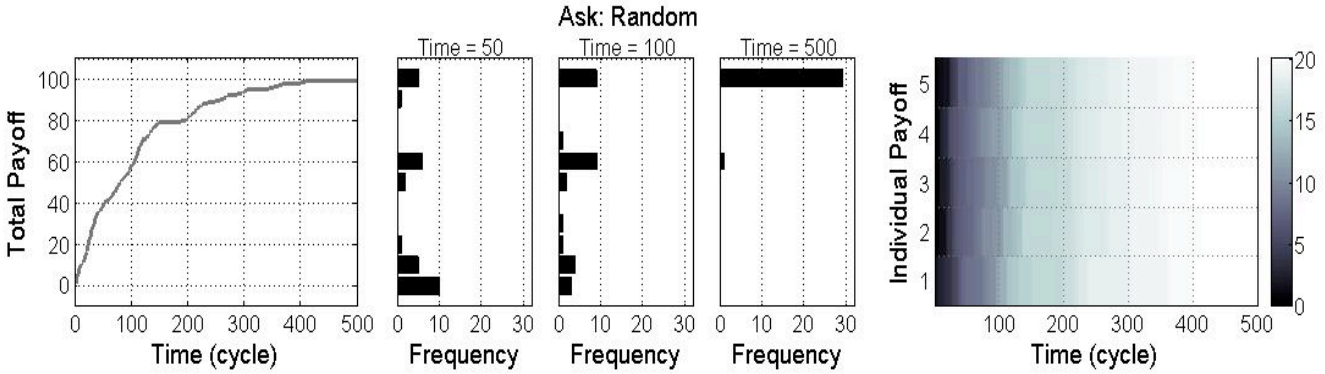


Figure 4: The results from the Random model in Simulation Study 1 are shown. The left most graph shows the evolution of total payoff (sum of all agents' payoffs) over the course of 500 trials, averaged across 30 simulations. The middle three histograms show the frequency of total payment for the 30 simulations at 50, 100, and 500 cycles. The color map on the right side shows the evolution of each agent's payoff averaged over 30 simulations.

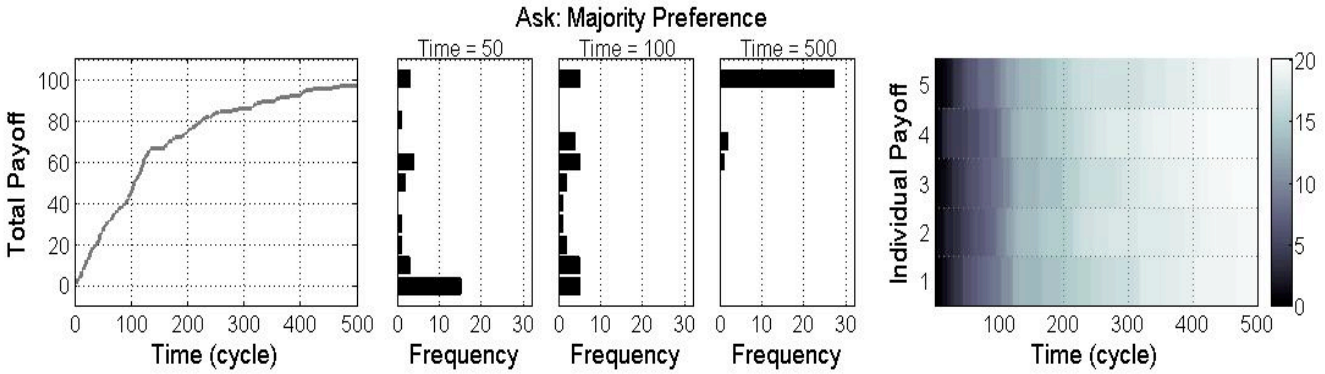


Figure 5: The results from the Majority Preference model in Simulation Study 1 are shown. The left most graph shows the evolution of total payoff (sum of all agents' payoffs) over the course of 500 trials, averaged across 30 simulations. The middle three histograms show the frequency of total payment for the 30 simulations at 50, 100, and 500 cycles. The color map on the right side shows the evolution of each agent's payoff averaged over 30 simulations.

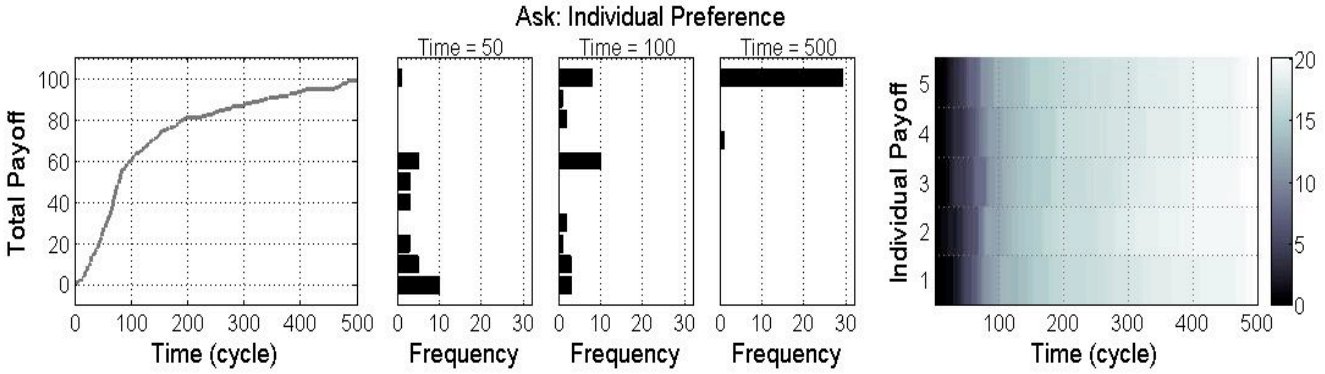


Figure 6: The results from the Individual Preference model in Simulation Study 1 are shown. The left most graph shows the evolution of total payoff (sum of all agents' payoffs) over the course of 500 trials, averaged across 30 simulations. The middle three histograms show the frequency of total payment for the 30 simulations at 50, 100, and 500 cycles. The color map on the right side shows the evolution of each agent's payoff averaged over 30 simulations.

The results from Simulation Study 1 show that for a single peak search space, asking random others can be especially beneficial when the time to search is limited. Every group member innovating can slow the team performance down. If there is a reasonable amount of time, the Majority Preference model and the Individual Preference model work fine. The success of the Random model suggests that we should sometimes observe different, random others, instead of always following the same individuals.

Simulation Study 2

In Simulation Study 2, the search space had three best solutions as shown in Figure 2. In this case, imitating can limit the number of good solutions the group discovers by causing all agents to conform to a single good solution. In contrast, the group can collectively find different solutions if group members innovate.

The procedure for Simulation Study 2 was the same as that for Simulation Study 1. The same four models were evaluated using a diversity metric and a normalized search speed for finding good solutions. The diversity metric was defined as the percentage of the group finding two or more best actions in 30 simulations. The normalized search speed, $\hat{v}_e$, is a relative metric defined by the time required to achieve 70% of the optimal result for a group, T_e. If a constant S quantifies the solution space, behavior model k has an observed average exploration speed, v_e:

$$v_e(k) = \frac{S}{T_e(k)}$$

Then the normalized search speed for model k, $\hat{v}_e(k)$, is:

$$\hat{v}_e(k) = \frac{v_e(k)}{\min_j v_e(j)} = \frac{S/T_e(k)}{\min_j S/T_e(j)} = \frac{\max_j T_e(j)}{T_e(k)}$$

	Number of Solutions 1	2	3	Diversity Metric
Innovate	0%	30%	70%	**100%**
Ask: Random	96.7%	3.33%	0%	**3.33%**
Ask: Majority Preference	70%	30%	0%	**30%**
Ask: Individual Preference	73.3%	26.7%	0%	**26.7%**

Table 1: The results from Simulation Study 2 are shown. The diversity metric shows the percentage of finding two or more best solutions in 30 simulations. The Innovate model was able to find two best solutions 9 times (30%) and three best solutions 21 times (70%), resulting in 100% diversity score. In contrast, the Random model resulted in finding only one good solution in 29 of 30 simulations. The performances of the Majority and Individual Preference models were in between those of the Innovate and Random models.

Table 1 displays the simulation results for the payoff distribution with three peaks. As predicted, the Innovate model was able to find multiple best solutions, resulting in a high diversity score. In contrast, the Random model resulted in the discovery of only one good solution in almost all 30 simulations (96.7%). The Majority Preference model and the Individual Preference Model were in between the Innovate model and the Random model. Although the two

preference models could not find all three best solutions, they were able to find two best solutions in some cases, much more frequently than the Random model.

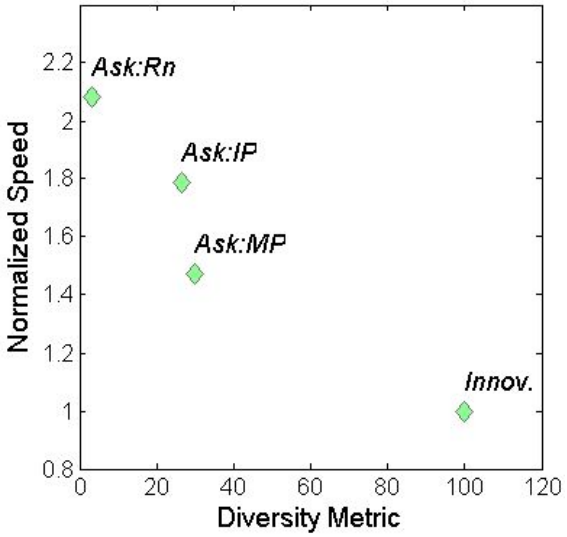


Figure 7: Each group's normalized search speed is shown as a function of its diversity metric. The normalized speed axis shows how quickly the group achieves a high total payoff (higher speed means faster). The diversity metric shows the percentage of finding two or more best actions in 30 simulations. The Innovate model (Innov) results in a high diversity measure but is slow to have all agents finding a good solution, indicated by low normalized speed. The Random model (Ask: Rn) leads to a high normalized speed, but this group converges to a single solution too quickly and thus results in low diversity of good solutions. The Majority Preference model (Ask: MP) and the Individual Preference Model (Ask: IP) were in between the Innovate model and the Random model.

Figure 7 shows each group's normalized search speed as a function of its diversity metric. The normalized speed axis shows how quickly the group achieves a high total payoff. This axis shows that we have essentially replicated Simulation Study 1 in terms of speed of finding a good solution as a group: Random, Individual Preference, Majority Preference, and Innovate, from fastest to slowest.

The diversity metric is the new measure relevant to the multiple best solutions in Simulation Study 2. All agents innovating results in a high diversity measure but, as Simulation Study 1 found, is slow to have all agents finding a good solution. The opposite of the Innovate model is the Random model. Asking random others leads to efficient dissemination of a solution and thus a high normalized speed, but makes the group converges to a single solution too quickly. Thus the Random model under explore the search space. The Majority Preference model and the Individual Preference model are quite efficient in disseminating a solution relative to the Innovate model. At the same time the two preference models have the time to explore the space. This is because always asking a particular individual has a higher chance of resulting in incidental innovation in the next round than asking a random other. When the asked agent does not have a good solution, the asking agent will be dissatisfied and innovate on the next round. When imitating a particular other, the imitating agent will likely keep asking the same agent. If this asked agent does not have a good solution, the asking agent will have many opportunities to innovate. When imitating a random other, the imitating agent will ask different agents at different cycles. There is less chance that the imitating agent always asks another agent with a poor solution in the Random ask model than in the two preference models. Thus the random imitation does not result in innovation as often as the other types of imitation.

Discussion

In the current study, we examined the benefits of different search strategies through computer simulation. We tested four models. In the Innovate model, each of the five agents in the group innovated on each cycle. The other three models incorporated some mechanisms of imitation. In the Random model, each agent imitated the best solution of a random other on each cycle. In the Majority Preference model, each agent imitated the best solution of the agent that was asked by many agents. This group followed the principle of preferential attachment, and conformed to the majority's behavior. In the Individual Preference model, each agent tracked how often it imitated the other agent, and imitated another agent based on how often it asked a certain agent. In this group, agents developed familiarity with a particular agent and followed this agent. We tested these four models in two kinds of search space: single best solution and three best solutions. In the current simulation, when imitating did not result in a better solution than the existing one, the agent innovated on the next time cycle and then resumed the imitation on the following time cycle.

The results from Simulation Study 1 showed that for a single peak search space, asking random others could be especially beneficial if the time to search was limited. In contrast, every group member innovating could take a long time for all the group members to find a good solution. The Majority Preference model and the Individual Preference model found good solutions in a reasonable amount of time.

In Simulation Study 2, the four models were tested under the three-peak environment. All agents innovating resulted in the group finding multiple good solutions, but, as Simulation Study 1 found, was slow to have all agents finding a good solution. In contrast, asking random others led to efficient dissemination of a solution, but the group converged to a single solution too quickly, and thus the Random model under explored the search space. Majority Preference model and the Individual Preference model had

the time to explore the space and were still quite efficient in disseminating a solution relative to the Innovate model.

Taken together, these results suggest that following a particular other, whether the most popular one or the most familiar one, results in a good compromise between speed and diversity in finding good solutions. It is interesting that these models that incorporate characteristics found in humans are most robust in the sense that they work well in different environments, although they may not be optimal in a single environment. Perhaps people's desire to follow particular others is a key to adaptive behavior, allowing people to disseminate ideas efficiently while still encouraging the innovation of new ideas.

Future work should explore more complex models, in which the group can have a mix of innovators and imitators. Individual differences can be useful when the group tends to converge too quickly. When group members converge quickly to an optimal solution, responding to a new situation becomes a problem (Resnick, 1994). For example, if all team members responded to an immediate threat in area X (which happens in the real world), it may take a while for everyone to respond to a new alert in area Y. Analogously, a group may fail to respond to a new and better solution when the group converges to a good solution too quickly. A simple way to avoid such failure to adapt to better solution is to include individuals with different abilities in a team (Sakamoto & Nickerson, 2007). By making some individuals innovate more often than others, we can encourage some learners to focus on disseminating solutions, and others to explore the space for new situations. These models incorporating individual difference can be robust to changing environments, such as when the payoff distribution shifts from time to time. Future work should include these variables, such as changing environment and individual difference, to make the simulation world closer to the world we live in. Future work should also compare these models against people.

In conclusion, the current simulation studies showed that people's natural tendency to follow particular others may have survived for a good reason: It leads to reasonable performances in a reasonable amount of time in different environments. If the dimension to optimize is well defined, one may tailor the search strategy. For example, if the time is not an issue, a group of agents that all innovate can find a diverse set of good solutions as a group. If there is a need to disseminate information widely and quickly, then asking random others will be the way to search the space. If one does not know what to optimize, following the particular others will result in a reasonable performance.

References

Asch, S. E. (1956). Studies of independence and conformity: A minority of one against a unanimous majority. *Psychological Monographs*, *70* (Whole no. 416).

Bandura, A. (1965). Behavioral modification through modeling procedures. In L. Krasner & L. P. Ulmann (Eds.), *Research in behavior modification: New development and implications* (pp. 310–340). New York: Rinehart and Winston.

Barabási, A. L., & Albert, R. (1999, October 15). Emergence of scaling in random networks. *Science*, *286*, 509–512.

Cialdini, R. B., & Goldstein, N. J. (2004). Social influence: Compliance and conformity. *Annual Review of Psychology*, *55*, 591–621.

Dennett, D. C. (1995). *Darwin's dangerous idea*. New York: Touchstone.

Deutsch, M., & Gerard, H. B. (1995). A study of normative and informational social influences upon individual judgment. Journal of Abnormal Social Psychology, 51, 629–636.

Festinger, L. (1954). A theory of social comparison processes. *Human Relations*, *7*, 117–140.

Gigerenzer, G, Todd, P. M., and the ABC Research Group (2000). Simple Heuristics that Make Us Smart. New York: Oxford.

Gureckis, T. M., & Goldstone, R. L. (2006). Thinking in groups. In S. Harnad & I. Dror (Eds.), *Distributed cognition: Special issue of pragmatics & cognition*, *14* (pp. 293–311). Amsterdam, The Netherlands: John Benjamins.

Kraatz, M. S. (1998). Learning by association? Interorganizational networks and adaptation to environmental change. *Academy of Management Journal*, *41*, 621–643.

Mason, W. A., Jones, A., & Goldstone, R. L. (2008). Propagation of innovations in networked groups. Journal of Experimental Psychology: General, 137, 422-433.

Page, S. E. (2007). The Difference. Princeton University Press.

Resnick, M. (1994). Turtles, termites and traffic jams: Exploration in massively parallel microworlds. Cambridge, MA: MIT Press.

Sadlon, E., Sakamoto, Y., Dever, H. J., Nickerson, J. V. (2008). The Karma of Digg: Reciprocity in Online Social Networks. In *Proceedings of the 18th Annual Workshop on Information Technologies and Systems*.

Sakamoto, Y., & Nickerson, J. V. (2007). Social behavior in a team of autonomous sensors. In *Proceedings of the Intelligence and Security Informatics Conference*.

Sakamoto, Y., Sadlon, E., & Nickerson, J. V. (2008). Bellwethers and the emergence of trends in online communities. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society*.

Sakamoto, Y., Ma, J., & Nickerson, J. V. (2009). 2377 people like this article: The influence of others' decisions on yours. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society*.

Sherif, M. (1935). A study of some social factors in perception. *Archives of Psychology*, *27*, 1–60.

Smith, K., Kirby, S., and Brighton, H. (2003). Iterated learning: A framework for the emergence of language. Artificial Life, 9, 371–386.

Arithmetic Notation…*now in 3D!*

David Landy (dlandy@richmond.edu)
Department of Psychology, University of Richmond
Richmond, VA 23173

Sally Linkenauger (sal3g@virginia.edu)
Department of Psychology, University of Virginia
Charlottesville, VA 22904

Abstract

When people reason formally, they often make use of special notations—algebra and arithmetic are familiar examples. These notations are often treated as mere shorthand—a concise way of referring to meaningful mathematical concepts. Other authors have argued that people treat notations as pictures—literal diagrams of an imagined set of objects (Dörfler, 2003; Landy & Goldstone, 2009). If notations depict objects that exist in space, then it makes sense to wonder how they are arranged not just in the two visible dimensions, but in depth. In four experiments, we find a consistent pattern: properties that increase mathematical precedence also tend to make objects appear closer in space. This alignment of formal pressures and informal pressures suggests that perceived depth may play a role in supporting computational reasoning processes. Although our primary focus is documenting the existence of depth illusions in notations, we also evaluate several sources of information that might guide depth judgments: availability of an object for computational actions, formal syntactic structure, relative symbol salience and voluntary attention shifts. We consider relationships between these nonexclusive possible sources of information in guiding how people judge depth in mathematics.

Keywords: Mathematical cognition, embodied cognition, depth perception

Introduction

Special notations are ubiquitous markers of mathematical thinking. Often, these notations are treated as mere conventional patterns, which serve as the target of rule-learning systems such as generic production systems (e.g., (Koedinger, Alibali, & Nathan, 2008; Anderson, 2005; Koedinger & MacLaren, 2002). From this perspective, the exact format and layout of the expression doesn't much affect how reasoning happens—what makes learning difficult is the rules, not the layout. However, growing evidence suggests a different account. It seems formal operations—from reasoning to logic to simple mathematics—are not always computed through the action of an abstract reasoning system, but instead make use of perceptual-motor systems typically involved in real-world perception and action.

In the case of mathematics in particular, reasoning that on the surface appears to require formal operations can be simplified if reasoners treat notations as pictures of a physical scene (Dörfler, 2003; Landy & Goldstone, 2009). For instance, algebraic syntax has a hierarchical structure partially described by the order of operations. In any equation, operations bind in the following order: parentheses, exponents, multiplications and divisions, additions and subtractions. This apparently arbitrary system appears to require explicit memorization, but in fact can be computed using basic perceptual-motor mechanisms such as grouping (Landy & Goldstone, 2007) and automatic attentional biases (Landy, Jones, & Goldstone, 2008). On this account, computing the answer to a math problem involves taking physical actions to transform abstract forms.

If notations really are fundamentally abstract, then their implied physical structure is entirely given by their surface form: there is no sense in which any of the symbols are 'close' or 'far away.' However, if people indeed often reason by treating symbols as pictures of objects in space, then these objects must be laid out in some three-dimensional arrangement. Thus, it is at least possible that different symbols would be seen as closer or further away than others. In the next section, we outline some reasons to expect such differences.

Reasons to Expect Illusions of Differential Depth

Actual equations and expressions are of course purely two-dimensional; thus actual depth experience should not directly inform perceived depth judgments with mathematical forms. Furthermore, accounts that treat notation as basically abstract predict no particular differences in apparent depth of different symbols. However, several factors might affect the perceived depth of symbols seen as objects that exist in space, which can be acted on in particular kinds of ways.

One clear prediction is that symbols that afford action appear proximal relative to those that do not. Several studies have found that depth perception can be affected by the action capabilities of the observer (Linkenauger, Witt, Stefanucci, Bakdash, & Proffitt, 2009; Witt, Linkenauger, Bakdash, & Proffitt, 2008; Witt & Proffitt, 2005; Witt, Proffitt, & Epstein, 2005). Therefore, if solving a mathematical equation requires actions on the part of the solver, high precedence terms—those most available for actions—should generally seem most proximal in arithmetic expressions. Put simply, years of experience acting first on multiplications in expressions like $3 + 4 \times 5 = 23$ will lead the multiplication to appear closer than the addition.

We hypothesized, more generally, that terms and operation signs that were most immediately available for

action would be seen as more proximal. In some cases, general syntactic factors do not align with action. That is, unlike in the stimuli in Figure 1, in some cases low-precedence operations afford more immediate action than high-precedence operations. This issue is taken up again in Experiment 3.

Another reason to expect systematic biases in perceived depth of arithmetic signs comes from the salience structure of those signs. As mentioned above, typical multiplication signs (the dot and the cross) are more salient and readily attended than addition signs. Salience and attention present conflicting pressures in the paradigm employed here. Voluntary attention has been shown to influence the perceived depth of ambiguous figures (Kawabata, 1986). In the case of arithmetic expressions, attention shifts systematically from high-precedence operations to low-precedence over the course of problem solution. Since participants made depth judgments after solving the problem, attention would most recently have been primarily allocated to additions, potentially causing addition signs to seem closer immediately after computation.

Salience also affects perception of figure and ground such that highly salient parts tend to be interpreted as parts of figures (Hoffman & Singh, 1997). Generally speaking, this may imply that salient objects will tend to be seen as proximal (though see Huang & Pashler 2009). The higher salience of multiplications should then also cause them to be perceived as proximal.

In three of the experiments reported below, participants were asked to solve simple mathematical problems, superimposed over two views of a baseball (see Figure 1). After solving the problem, they judged the relative distance of the two baseballs. If the relative availability of computational action affects perceived depth, the baseball associated with the high-precedence sign (the multiplication sign) should appear closer than the baseball under the low-precedence sign. The baseballs were used to ground the participants' judgments and make clear the nature of the task; our assumption is that judged distance of the baseball reflects primarily the relative perceived depth of the symbol superimposed on it.

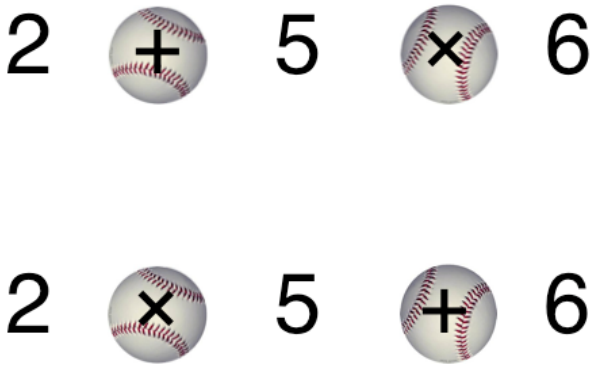


Figure 1: Sample stimuli used in Experiment 1.

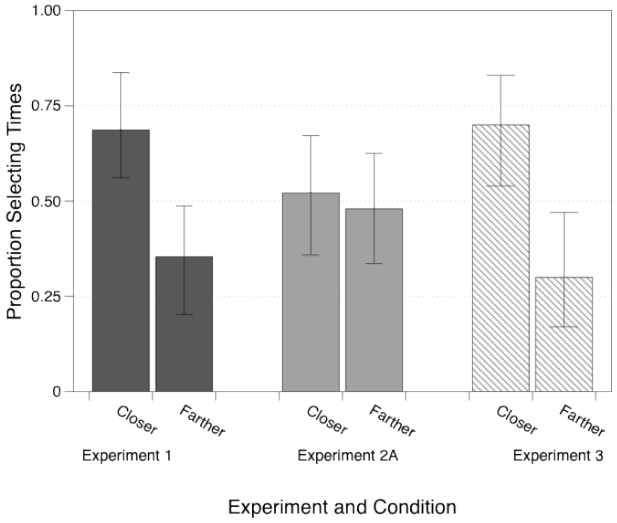


Figure 2: Results of Experiments. Errors represent 95% confidence intervals on proportions.

Experiment 1

Method

Participants Forty-eight students from the University of Richmond received partial course credit for participation. All participants had normal or corrected to normal vision.

Materials Participants viewed simple two-operation arithmetic expressions superimposed over images of a baseball (see Figure 1). The same image that appeared under the plus sign appeared under the multiplication sign, rotated 90°. Instructions informed participants that one baseball was close, and approaching, while the other baseball was farther away, and receding.

Procedure Participants were asked to solve the arithmetic problem. After doing so, they were asked to decide the relative distance of the two baseballs. Half of the participants were asked to circle the baseball that was closer; the other half to circle the baseball that was farther. Each participant responded to two expressions; one was in the format $a+b\times c$, the other $a\times b+c$. The order of presentation of the problems was counterbalanced across participants, as was which image orientation appeared on the left (thus image orientation and operation sign were independent). The task was performed as part of a distracter, between two phases of an unrelated experiment. Several other short problems appeared in between the two baseball judgments.

Results On each trial, the participant circled either the multiplication sign or the addition sign (see Figure 2). These choices were analyzed using two repeated measures logistic regressions: one included just an intercept, while the other also included judgment type (closer vs. farther) as a between-participants factor. Including judgment type

significantly improved the quality of the fit over the baseline model, ($\chi^2(1) = 10.2$, $p<.01$).

Discussion The primary purpose of Experiment 1 was to evaluate whether there was consistency in how depth would be judged in a simple arithmetic expression. The results demonstrated that there is. In a simple two operation problem, multiplications are seen as closer than additions. In simple arithmetic processes, the higher precedence operation is perceived as more proximal to the reasoner than is the lower precedence operation.

Though predicted by the idea that multiplication is more available than addition in this expression, these results could be produced by any combination of an effect of the higher salience of cross signs over plus signs, the syntactic order of operations, and the relative availability of multiplication in this expression. The prediction that voluntary attention toward the plus sign at the end of solving the problem would dominate proximity judgment was not borne out.

Experiment 2a

In this experiment, we tested whether the bias in perceived depth revealed in Experiment 1 was due to particular salience differences between the addition and multiplication sign rather than order of operations. In Experiment 2a, we used the same stimuli and design as in Experiment 1, except that we used parentheses to make the plus sign the first operation rather than the multiplication. If the result in Experiment 1 was due simply to perceptual differences between the multiplication and addition signs, then we should expect to replicate Experiment 1. However, if the result is due to order of operations, adding the parentheses should result in either a null effect or the opposite effect.

Method

Participants Forty-eight students from the University of Virginia received partial course credit for participation. All participants had normal or corrected to normal vision.

Materials Packets were created using the same stimuli as in Experiment 1, except the stimuli were modified so that the two numbers adjacent to the plus sign were enclosed by parentheses.

Procedure The procedure was the same as in Experiment 1 except that rather than serving as a distracter in an unrelated task, these judgments served as the primary experiment. In between the two trials, participants completed a distracter task which involved solving a maze.

Results and Discussion Participant choices were analyzed using a repeated measures logistic regression (see Figure 2). Including judgment type did not improve the fit by a likelihood ratio test over the null model ($\chi^2(1) = .12$, $p\sim.68$).

Unlike Experiment 1, there was no sign of a consistent relationship between perceived relative depth and operation sign in simple expressions with parentheses. One plausible interpretation of the disparity is that the higher precedence of multiplication sign interacts with a pressure to see parenthesized terms as closer. There are (at least) two plausible reasons why this would occur: first, as hypothesized, terms that can be computed early may appear closer than they otherwise would. Second, visual factors intrinsic to parentheses may make them things inside parentheses appear differentially closer.

Consider Figure 3. In the top part, the circle on the left appears to be in front of (and consequently partially occluding) the illusory oval induced by the curved lines. This, in turn, causes it to appear closer than the circle on the right. In a similar manner, it may well be that the parentheses create a (relatively weak) illusory oval. If the symbols are interpreted by the visual system as being in front of that oval, then they may be perceived as being closer than the symbols not inside the parentheses.

Experiment 2b

One difference between Experiment 1 and 2a is that the populations differed, one being drawn from University of Richmond students and the other from the University of Virginia. Therefore, to eliminate the possibility that the null results in Experiment 2a were due to differences between the student populations, a second experiment was run at the University of Virginia to ensure that similar perceptual effects as in Experiment 1 could be

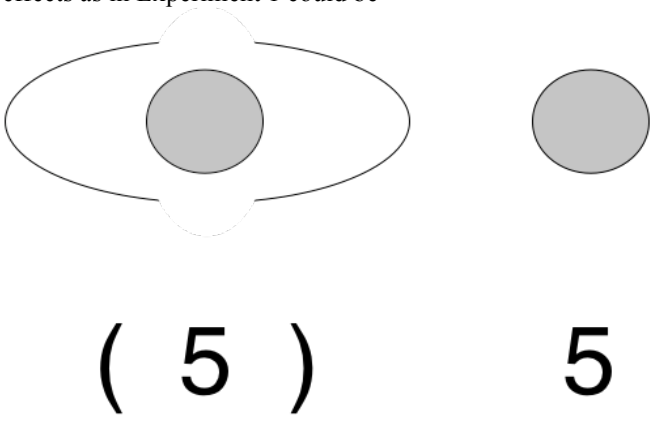


Figure 3: Visual factors could influence the relative perceived proximity of terms inside parentheses. Just as the circle on the left appears to be closer than the circle on the right, so apparent occlusion may cause the 5 on the left to be in closer than the 5 on the right.

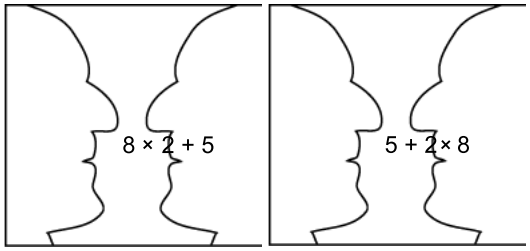


Figure 4. Example Stimuli from Experiment 2b.

found in this population. In addition, this experiment offers a more indirect judgment of relative proximity.

Method

Participants Twenty-two students from the University of Virginia received partial course credit for participation. All participants had normal or corrected to normal vision.

Stimuli and Apparatus Mathematical equations were superimposed onto an illustration of the vase-face illusion so that either the multiplication sign was on the vase and the plus on the face or vice versa (see Figure 4). Instructions printed above the picture directed participants to solve the equation and then indicate whether or not they saw a vase or faces. The equation was either on the right side of the illustration, as in Figure 4, or on the left.

Packets were created which consisted of two face-vase trials. A maze was inserted in between the vase-face trials, which acted as a distracter task. Half of the packets contained vase-face illusions in which equations were located on the right; the other half had the equations located on the left. Within each packet, one illusion had the multiplication sign on the vase, and the other had the multiplication sign on the face. Order was counter-balanced across packets.

Procedure Participants were given a packet and told to complete it in full. They were asked to follow the instructions written on each sheet, and not to look at the subsequent sheets until they had completed their task on the current sheet.

Results and Discussion Arbitrarily, each trial was coded as positive if the participant chose the faces as the foreground. A significantly better fit was found for a logistical regression including position of the multiplication sign as a factor, than for the null model ($\chi^2(1) = 5.1$, $p<0.05$). Overall, when the multiplication sign was over the vases, participants chose the face interpretation less often (M=.5, CI=0.28-0.71) than when the multiplication sign was over the faces (M=.82, CI=0.60-0.95).

These results show that the effect in Experiment 1 can generalize across populations and across tasks. This finding is also a less direct manipulation of perceived depth and is less likely to be affected by demand characteristics. Whether the face or vase is seen in the foreground is indicative of the depth relationship between the vase and the face. Therefore, participants determined which figure they saw instead of directly specifying the depth relationship between the mathematical operators. Interestingly, order of operations influenced the figure-ground and therefore, the depth relationship in an ambiguous figure illusion.

Prior results have shown that fixations (Gibson & Peterson, 1994) and exogenous cues (Vecera, Flevaris, & Filapek, 2004) guide figure-ground segmentation, as long as the cues appear inside figures (as was done here). Thus, these results could result from the greater salience of multiplication signs. However, once again, voluntary attention shifts are unlikely to account for these effects, as voluntary attention is most likely directed toward the addition at the end of computation (when participants were instructed to make their judgments).

The first two experiments indicated that syntactic precedence, available formal actions (computations), and perceived proximity tend to go together. Experiment 3 distinguishes between syntactic precedence and action structure by repeating the structure of Experiments 1 and 2A in the context of an algebraic rather than an arithmetic task. Because linear equations are solved starting from the lowest, rather than the highest precedence operations, in Experiment 3 the two theories make opposite predictions. If availability of formal actions guides perceived depth, then lower precedence items should be seen as proximal in Experiment 3; if syntactic precedence predicts or guides depth, then high precedence operations should appear proximal, as in Experiments 1 and 2B.

Experiment 3: Linear Equations

Method

Participants Twenty students from the University of Virginia received partial course credit for participation. All participants had normal or corrected to normal vision.

Materials Packets were created using a format identical to Experiment 2A. Similar stimuli were used, except the stimuli were modified from an arithmetic computation to the solution of a linear equation (see Figure 5). Participants were instructed to solve the equation before deciding which ball was closer (further).

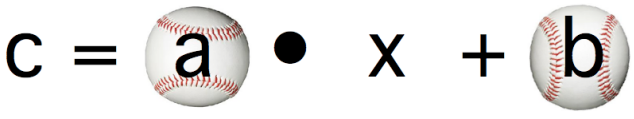


Figure 5. Example Stimulus from Experiment 3.

Two other changes were made to the stimuli: first, the dot notation was used for multiplication in place of the cross. This sign is much more typical in algebraic contexts, and is less likely to be confused for a variable than the cross. Furthermore, crosses and dots have similar salience advantages in mathematical contexts (Landy et al., 2008). Second, the baseballs were moved from behind the operations to behind the operands. This was done because, in an equation solution context, the operands are more relevant units of action. That is, one solves the problem illustrated by cancelling the *b*, and then the *a*. Thus, the effect of perceived action was predicted to be strongest in this configuration.

Procedure The procedure was identical to Experiment 2A.

Results and Discussion Participant choices were analyzed using a repeated measures logistic regression (see Figure 2). Including judgment type significantly improved the fit over the baseline model ($\chi^2(1) = 6.0$, $p<0.05$). Participants were more likely to judge the baseball under the multiplicative term as closer than the baseball under the additive term.

The results of Experiment 3 contradicted our original hypothesis that perceived depth would align with available actions, instead supporting the idea that syntactically central symbols appear to be closer than syntactically peripheral symbols in mathematical expressions.

The results of Experiment 3 are to some degree compatible with the interpretation that low-level visual features guide perceived proximity. In this case, perceived depth of the baseball images would be affected by the terms adjacent to the judged baseballs, rather than those directly behind it. In a pre-cuing task, Baylis and Driver (1995) reported that exogenous cues to attention did not affect figure-ground segregation (which tends to align with depth perception in most cases, though see Huang & Pashler, 2009), when the cue appeared outside the area in which the figure appeared (see also Vecera et al., 2004). Nonetheless, in this case, it is possible that the highly salient dot adjacent to the baseball increases its apparent proximity. It is also possible that the multiplicative terms and the multiplication sign are visually grouped and therefore that the salience of one part of the group (the multiplication sign) causes the entire group to appear closer.

General Discussion

Although mathematical notation is a formal language, and is inherently two dimensional, readers of these notations come to quite consistent judgments about the relative proximity of terms in formal expressions. Three experiments demonstrated that factors that determine formal precedence (operation sign and parentheses) also systematically influence perceived depth.

Variations in perceived depth aligned in our stimuli with formal precedence. The current results do not distinguish between the possibilities that syntactic precedence directly affects perceived depth, and that the low-level visual features of typical mathematical notation determine apparent depth. Although future work should distinguish these two possibilities, we think it notable that mathematical notation is structured in such a way that there is a systematic relationship between low-level visual features affecting and mathematical syntax. This alignment raises the possibility that perceived three-dimensional structure may be used as a cue to mathematical ordering.

As long as episodes of formal reasoning are indeed typically organized by attention-based interactions with external environments (Landy et al., 2008; Patsenko & Altmann), the alignment of perceptual factors such as visual grouping, salience, and depth may be significant factors in making symbolic mathematical notation such a powerful and successful system for supporting reasoning.

Three limitations of the current work are worth noting: one is that it does not indicate the strength of the judgment. Although judgments were significant and consistent in Experiments 1, 2A, and 3, participants made forced choice binary judgments. Thus, while we can conclude that people generally perceive multiplications as closer than additions, the current experiments give no indication of the magnitude of the perceived difference.

Another limitation is that there is no indication in the current studies of whether this perceived difference in depth has any effect on mathematical judgments. Future work should explore whether explicit manipulations of apparent depth disrupt mathematical reasoning processes. Finally, there is a possibly important confound in Experiment 1. In these stimuli, the laces on the baseballs overlap, and are obscured by, the addition sign slightly more than the multiplication sign. This might provide a stronger depth cue, causing subjects to see the baseball under the addition sign as farther away. This confound could not explain the difference between Experiment 1 and 2a, nor could it explain the effect in Experiment 3. In Experiment 3, the baseballs appeared under the letters. These were counterbalanced across condition, and so could not have led to differences in judgment.

Recognizing these limitations, nevertheless the existence of consistent depth cues in mathematical notations bolsters interpretations that treat mathematical reasoning as (sometimes) a form of spatial reasoning over symbolic objects. Accounts that treat symbolic reasoning as abstract rule learning cannot make systematic predictions about depth, such as those seen here. Understanding how and when such factors matter for reasoning promises to further illuminate our understanding of general formal reasoning processes.

Acknowledgments

Thanks to Dennis Proffitt, Lydia Nichols, and Ryan Smout for valuable comments on the development of this project. Thanks to Ryan Smout for data collection.

References

Anderson, J. R. (2005). Human symbol manipulation within an integrated cognitive architecture. *Cognitive Science*, *29*(3), 313–341.

Baylis, G. C., & Driver, J. (1995). One-sided edge assignment in vision: 1. Figure-ground segmentation and attention to objects. *Current Directions in Psychological Science*, *4*, 140-146.

Dörfler, W. (2003). Diagrams as means and objects of mathematical reasoning. In *Developments in mathematics education in German-speaking countries. Selected papers from the annual conference on didactics of mathematics*.

Hoffman, D. D., & Singh, M. (1997). Salience of visual parts. *Cognition*, *63*(1), 29–78.

Huang, L., & Pashler, H. (2009). Reversing the attention effect in figure-ground perception. *Psychological Science*, *20*(10), 1199-1201.

Kawabata, N. (1986). Attention and depth perception. *Perception*, *15*(5), 563-572.

Koedinger, K. R., Alibali, M. W., & Nathan, M. J. (2008). Trade-offs between Grounded and Abstract Representations: Evidence from Algebra Problem Solving. *Cognitive Science*, *32*(2), 366-397.

Koedinger, K. R., & MacLaren, B. A. (2002). *Developing a pedagogical domain theory of early algebra problem solving*. Pittsburgh, PA: Carnegie Mellon University.

Landy, D., & Goldstone, R. L. (2007). How abstract is symbolic thought? *Journals of Experimental Psychology: Learning, Memory, and Cognition*, *33*(4), 720-733.

Landy, D., & Goldstone, R. L. (2009). Pushing Symbols. In N.A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society*. Amsterdam: Cognitive Science Society.

Landy, D., Jones, M. N., & Goldstone, R. L. (2008). How the appearance of an operator affects its formal precedence. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 2109-2114). Austin, TX: Cognitive Science Society.

Linkenauger, S. A., Witt, J. K., Stefanucci, J. K., Bakdash, J. Z., & Proffitt, D. R. (2009). The Effects of Handedness and Reachability on Perceived Distance. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(6), 1649-1660.

Patsenko, E. G., & Altmann, A. M. (2010). How planful is routine behavior? A selective-attention model of performance in the Tower of Hanoi. *Journal of Experimental Psychology: General*, *139*, 95-116.

Vecera, S. P., Flevaris, A. V., & Filapek, J. C. (2004). Exogenous spatial attention influences figure-ground assignment. *Psychological Science*, *15*(1), 20–26.

Witt, J. K., Linkenauger, S. A., Bakdash, J. Z., & Proffitt, D. R. (2008). Putting to a bigger hole: Golf performance relates to perceived size. *Psychonomic Bulletin and Review*, *15*(3), 581.

Witt, J. K., & Proffitt, D. R. (2005). See the ball, hit the ball: apparent ball size is correlated with batting average. *Psychological Science*, *16*(12), 937-938.

Witt, J. K., Proffitt, D. R., & Epstein, W. (2005). Tool use affects perceived distance, but only when you intend to use it. *Journal of Experimental Psychology: Human Perception and Performance*, *31*, 880–888.

Computational Semantic Detection of Information Overlap in Text

Julia M. Taylor (jtaylor@riverglassinc.com)
RiverGlass Inc, 2001 South First St
Champaign, IL 61820 USA

Abstract

This paper is an attempt to investigate whether a computer is capable of finding similar information in structurally different texts, as people do it, without relying on lexical matching and without guessing the meaning of sentences based on word co-occurrence. Considered texts describe the same event, but each text may focus on different parts of the event. The considered texts are not paraphrases, but rather human-produced descriptions of a simple picture. The goal is not to find similar words in texts, which can be easily done, but to meaningfully connect the overlapping concepts and relationships used in the text descriptions. The meaning-based approach does not use any statistical/machine-learning techniques. The performance of a machine in finding similarity is compared to human performance not just in numbers but in the found information. The results show that the machine matches four out of the five human findings.

Keywords: text duplication and similarity, information overlap detection, meaning processing, ontological semantics.

Overview

This paper examines the use of the Ontological Semantic Technology (OST)—a modified version (Raskin et al 2010) of Ontological Semantics (Nirenburg & Raskin 2004)—for processing similar texts and compares it to human processing. Instead of selecting existing texts and assessing their similarity, users were given the same picture to describe. Clearly, the users will emphasize different objects or events on the picture, but at the same time, because they are all looking at the same picture, some of the provided information will overlap. The experiment is done to demonstrate the ability of the technology to understand the meaning of text, regardless of individual words that are used and of the length of the sentences.

The OST claim to fame is that it "understands" the meaning of text. The meaning of text includes paraphrases of sentences or paragraphs. A large number of paraphrases can be produced from a single sentence, an even larger number can be produced from a paragraph. Because of this large number of potential paraphrases, and because it is unclear which ones are good enough, instead of asking people to paraphrase a text, we ask them to describe a picture.

The untested assumption is that looking at the picture should activate the same schema(ta) as reading a paragraph. Thus, the main information received should be approximately the same whether looking at the picture or reading text. Instead of *reconstructing* original sentences after reading or listening to a text, the subjects were asked to *describe* what they see on the picture in their own words. The tasks of paraphrase and describing a picture are by no means identical, even for short sentences when compared to very simple pictures. Several things should be noticed: 1. Length of the sentences in paraphrases has probably some correlation to the length of the original sentences. 2. The choice of words for the description task is not limited by the original sentence, whereas it is possible that, in the paraphrase, the subjects would try to come up with unnatural synonyms in their desire to paraphrase. 3. The order of sentences is free in the picture description, while it is possible that the sentences would be ordered according to the original text in the paraphrase.

While paraphrase detections have received some attention from the machine-learning community (Fernando & Stevenson 2008, Clough et al 2002, Qiu et al 2006, Zhang & Patrick 2005), to the best of our knowledge same picture descriptions have not been addressed. This is surprising because most real life event descriptions are more similar to picture descriptions than to paraphrasing tasks.

The task of paraphrase limits information that is available to the subjects to that in the task, while describing the picture provides more freedom of focus. For example, the sentence *a black ball is on top of a green cube*, can only be paraphrased in term of the provided information. Possible paraphrases are: *a green cube is under a black ball; a black sphere-shaped object is above a green cube; a ball is positioned on top of a cube, the ball is black and the cube is green*. Notice that there may be a considerable variation among paraphrases in terms on the words used, the order in which they are described, and the number of clauses used in the description. What they all have in common, however, is the properties and attributes that connect the described objects: all describe shape either explicitly as in *sphere-shaped object*, or by accessing the knowledge of a shape of a lexical item as in *ball* or *cube*; and all describe color. However, if picture is shown (Figure 1), other things may come into focus for different people, such as relative size of the objects.



Figure 1: A black ball on top of green cube

It would be interesting to see if such unmentioned-in-the text characteristics would ever be brought up by the subjects in the paraphrase generation as unknown. It is, however, not the purpose of this experiment. The only significant assumption for this paper is that the greater variation of text should be encountered in the picture experiment, which in turn tests the machine's capability of catching the overlap to

a much greater extent. On the other hand, it would be interesting to see if a coherent description of a situation could be constructed from a union of all descriptions, as it is likely that these descriptions, to some extent, complement each other.

It is this overlap information in descriptions reported by subjects, as well as the difference or the union, that is captured and analyzed by the machine, as compared to the overlap and difference in information in responses as perceived by human is the subject of the paper. The theoretical knowledge obtained in this kind of research is applicable to an increasingly urgent task of easing the information overload by removing duplicate and overlapping information[1].

Ontological Semantics Technology

OST is an upgraded, much improved and implemented (and, on occasion, perverted) version of (Nirenburg & Raskin 2004) that detects the meaning of text. Ontological Semantics is a theory, methodology and technology for representing natural language meaning, for automatic transposition of text into the formatted text-meaning representation (TMR), and for further manipulation of TMRs for inferencing and more advanced reasoning, both theoretically and in a growing variety of applications. The main knowledge resources in OST are the language-independent ontology and language-specific lexicons.

The OST is not a toy system that works on a handful of examples; instead, it works with unrestricted texts in real-life applications, as well as avoiding the scalability problems (see Raskin et al 2010).

Ontology

The ontology contains information about the world; it is a constructed, engineered model of reality, a theory of the world (Gruber 1993, 1995; Nirenburg & Raskin 2004:138-139). It is a structured system of concepts covering the processes, objects, and properties in all of their pertinent complex relations, to the grain size determined by an application or by considerations of computational complexity. The ontology contains PROPERTIES, EVENTS, and OBJECTS. The concepts are named purely for the convenience of a human: the label itself does not contribute to the information content. Every OBJECT and EVENT is defined with a number of properties, thus allowing the concept to differ not only in label, but also in machine-understandable information. The child concepts inherit properties from the parent concepts.

Formally, the OST ontology is a lattice of conceptual nodes (for a construction of ontology and verification see Hempelmann et al. 2010 and Taylor et al 2010 respectably), each of which is represented as:

concept-name
  (property (facet(property-filler^{+}))$^{+}$)$^{+}$
property-filler
  concept-name | literal value
property
  attribute | relation
facet
  SEM | VALUE | DEFAULT | RELAXABLE-TO[2]

The current implementation of OST uses the following three axioms:

- subClassOf for concepts: IS-A (example: PHYSICAL-OBJECT IS-A OBJECT)
- subPropertyOf for properties: IS-A (example: COLOR IS-A PHYSICAL-OBJECT-ATTRIBUTE)
- inverse for properties: INVERSE (example: THEME INVERSE THEME-OF)

Concept interpretation (without facets, for the ease of reading) can be looked at using the following: given a set of objects $\mathcal{D}$, where $\mathcal{D}$ is the disjoint union of $\mathcal{D}$c (concepts) and $\mathcal{D}$d (literals), and given its interpretation function $\mathcal{I}$, for every atomic concept B, $\mathcal{I}$[B]$\subseteq\mathcal{D}$c; for every literal V, $\mathcal{I}$[V]$\subseteq\mathcal{D}$d ; for every relation R, $\mathcal{I}$[R]$\subseteq\mathcal{D}$c x $\mathcal{D}$c; for every attribute A; $\mathcal{I}$[A]$\subseteq\mathcal{D}$c x $\mathcal{D}$d. Moreover, the following is true for concepts C and D:

$\mathcal{I}$ [ALL] = $\mathcal{D}$
$\mathcal{I}$ [ε] = Ø
$\mathcal{I}$ [C D] = $\mathcal{I}$ [C] $\cup$ $\mathcal{I}$ [D]
$\mathcal{I}$ [and C D] = $\mathcal{I}$ [C] $\cap$ $\mathcal{I}$ [D]
$\mathcal{I}$ [(Rel(D)))]= {x$\in\mathcal{D}$c| y $\in$ $\mathcal{I}$ [D], <x, y>$\in\mathcal{I}$ [Rel]}
$\mathcal{I}$ [(Rel(and C D))] = $\mathcal{I}$ [Rel(C)] $\cap$ $\mathcal{I}$ [Rel(D)]
$\mathcal{I}$ [Rel(C D)] = $\mathcal{I}$ [Rel(C)] $\cup$ $\mathcal{I}$ [Rel(D)]
$\mathcal{I}$ [C(Rel(D))] = $\mathcal{I}$ [C] $\cap$ $\mathcal{I}$ [Rel(D)]
$\mathcal{I}$ [(Att(V)))]= {x$\in\mathcal{D}$c| y $\in$ $\mathcal{I}$ [V], <x, y>$\in\mathcal{I}$ [Att]}

Clearly, concept C is a descendant of D if $\mathcal{I}$[C]$\subseteq\mathcal{I}$[D]; and $\mathcal{I}$[(C(R(D))]$\subseteq\mathcal{I}$[C]. Whenever relation Rel is defined with a domain D and range R, if $\mathcal{I}$[C]$\subseteq\mathcal{I}$[D] and $\mathcal{I}$[E]$\subseteq\mathcal{I}$[R], then C(Rel(E)) is equivalent to $\mathcal{I}$ [C] $\cap\mathcal{I}$ [D(Rel(and E R))].

For the examples in this paper, it is sufficient to mention that when facets are involved, the highest priority facet takes precedence over the lower priority one.

Lexicon

The lexicon is the starting point for machine interpretation of language in OST. Since Ontological Semantics is centered on meaning, we will largely concentrate on the semantic structure (sem-struc) part of the lexicon entries.

In general, the lexicon can be looked at as a collection of words (and phrasals), organized such that each word is

[1] The author believes that whether an overlap indicates an importance of information in text is a separate (to her, dubious) hypothesis, which will not be addressed in this paper.

[2] The list shown has been enriched in the current implementation of OST, but since facets do not contribute much to this paper, the list is left as it was first introduced.

listed with all of its senses. Each sense of the word in a lexicon follows the following structure:

```
(WS-PosNo
(cat(Pos))
(synonyms "WS-PosNo"))
        (anno(def "Str")(ex "Str")(comments "Str"))
        (syn-struc((M)(root($var0))(cat(Con))(M))
        (sem-struc(Sem))
)
```

where the following grammar defines what is allowed:

```
M  → (Srole((root(Var))(cat(Cpos)))
   → (Srole((opt(+))(root(Var))(cat(Cpos)))
   → (M(M))
Pos → N |                (noun)
    → V |                (verb)
    → Adj |              (adjective)
    → Adv |              (adverb)
    →…
Con → NP |               (as defined by rules omitted
    → VP |               here to save space)
    → Con Con |
    → Pos
SRole → subject |                (syntactic roles,
      → directobject |           only some are shown
      → pp-adjunct               to save space)
      → …
No    → [1-9]                    (any digit)
Str   → [A-Z|a-z| |,|.]          (any string)
Var   → $varNo
      → Str
Sem   → C |                      (any ontology concept)
      → ^Var(R(F(C))) |          (R, F, C from ontology)
      → C(R(F(^Var)))            (C, R, F from ontology)
```

When the machine processes text with the help of the resources, the ontological concepts are accessed through the (English) lexicon. For example, a lexical entry for the verb *run* will contain all the possible senses, of which #6 is shown below:

```
(run-v6
    (cat(v))
    (anno
        (comments "...")
        (def "meet unexpectedly")
        (ex "I ran into my teacher at the movies last
night."))
    (syn-struc
        ((subject((root($var1))(cat(np))))
        (root($var0))(cat(v))
        (prep((root(into))(cat(prep))))
        (directobject((root($var2))(cat(np)))))
    )
    (sem-struc
        (meet-with
            (agent(value(^$var1(should-be-
a(sem(human))))))
            (beneficiary(value(^$var2)))
            (intentionality(value(<0.3))(relaxable-to(<0.5)))
        )
    )
)
```

The entry shows that this sense of *run* means 'unexpected meeting event' (from sem-struc), and it needs a preposition *into* (from syn-struc) to be activated. It also shows that in its normalized form the subject is usually the agent of the event, and the direct object is the beneficiary. Optional properties such as time, place, etc are usually not shown in the lexical items.

OST On Black balls and Green Cubes

OST uses the Semantic Text Analyzer (STAn) to interpret the meaning of sentences. The (machine generated) output of STAn is a text meaning representation (TMR) that shows the conceptual representation of the text, regardless of the language of the input. Let us go back to the sentence *a black ball is on top of a green cube*. The resulting TMR is:

```
Event: pred1
  (theme(value (physical-object1
   (shape(value(sphere)))
   (color(value(black)))
   (above(value(physical-object2
        (shape(value(cube)))
         (color(value (green)))
   )))
)))
```

Possible paraphrases from the previous section is: *a green cube is under a black ball:*

```
pred1
  (theme(value (physical-object1
   (shape(value(cube)))
   (color(value(green)))
   (below(value(physical-object2
        (shape(value(sphere)))
         (color(value (black)))
   )))
)))
```

Another interesting paraphrase is: *a ball is positioned on top of a cube, the ball is black and the cube is green*, which will result in the following:

```
put1
  (theme(value (physical-object1
   (shape(value(sphere)))
   (above(value(physical-object2
        (shape(value(cube)))
   )))
)))
pred1
   (theme(value (physical-object1
        (shape(value(sphere)))
        (color(value(black)))
```

```
   )))
pred2
   (theme(value (physical-object2
        (shape(value(cube)))
        (color(value(green)))
   )))
```

Notice that besides the PUT event, corresponding to *is positioned*, and the inverse of the BELOW-ABOVE properties, the rest of the information is identical for any purposes, including reasoning. The third example is especially interesting, as the colors are assigned to the indexed objects, referenced by the previous sentence.

The intersection of the paraphrases, as indicated by the TMRs once the inverse properties are used, are:

```
 pred1
   (theme(value (physical-object1
    (shape(value(sphere)))
    (color(value(black)))
    (above(value(physical-object2
        (shape(value(cube)))
         (color(value (green)))
    )))
)))
```

The union of the TMRs adds information only present in the third example, namely that of PUT, thus, producing

```
put1
   (theme(value (physical-object1
    (shape(value(sphere)))
    (color(value(black)))
    (above(value(physical-object2
        (shape(value(cube)))
         (color(value (green)))
    )))
)))
```

If Figure 1 is described instead of paraphrases, and sentences like a *ball is smaller than a cube* happen to be added to the description, it is easy to see that the intersection of TMRs will remain the same, while the union will add the additional size information.

More Complex Pictures

As demonstrated in the previous sections, OST is capable of understanding the meaning of close paraphrases and represent it in such a way that the differences and similarities are shown. The next experiment aimed at stretching the similarities as far as possible, but asking the user to describe a picture instead of paraphrasing a text.

The picture shown to the user was selected to depict an unambiguous object in the foreground, while the background contains objects that can be described either very briefly, if at all, or be paid as much attention as possible. The hypotheses are:

- The description of the central element of the picture is affected by individual/personal schemata, and therefore will partially differ from person to person. However, there should be an overlap in descriptions, focused on that central object, just as the paraphrases showed.
- The description of the background will differ from person to person to a much greater degree. A very small overlap is expected from pairs of participants since the background is not in focus (metaphorically).
- The activated schemata are not expected to be known to a computer, thus the computer will process only information explicitly stated by the subjects.

This is not at all an attempt to deal with the well-researched figure-ground phenomenon (see Talmy 2000, vol, 1: 311-344). Instead, we are only interested in the foreground display, but the background may provide individual distinctions.

Methodology

Once a picture was chosen, 3 subjects, unfamiliar with an experiment's goals and from unrelated occupations, were asked to describe the picture. The picture was visible to the subject all the time, thus the description is not effected by the accuracy of their recollection of the picture. The instructions requested to describe only what is seen on the picture, without alluding to any inferences or encyclopedic knowledge that the picture may activate. The subjects were not given any specific time frame to complete the task.

The described text was then entered into a machine for processing, and the union and intersection of information in individual texts were computed. Whenever the descriptions contradicted each other, the contradictions were also added to the union as alternative interpretation.

To check the validity of the found union and intersection, a person not participating in the description task and not involved in the OST part of the experimentation was asked to highlight the similarities in text. These similarities were then compared to the intersection of interpretations provided by a computer.

The foreground of a picture showed a moving elephant. The background of the picture contained trees, shrubs and other greenery, as well as a place where several cars were parked, as seen in Figure 2.



Figure 2: An elephant crossing the road

Results of Human Description

The descriptions of the submitted texts varied length (the first text used 54 words, the second text used 124 words,

and the third text used 151 words) and structure of sentences.

The following similarities were noticed by a human in all of the descriptions:

- Elephant's existence.
- Road on which the elephant is located.
- Trees in front of the cars, in some spatial relation to the elephant
- Cars parked in the background

The following information was included in at least one of the texts (author's summary below):

- A large African male elephant is shown on the picture and is moving either on the road or bare ground or crossing the road. The elephant has large tusks, 4 legs, one visible ear, one visible eye, a tail and a trunk. The front right leg of the elephant is bent at the knee.
- There is dust on road and some dirt or hard soil on the edges of the road. The road is wide and paved.
- A row of trees are between the elephant and the cars, past the cars and on the berm. The trees are large with extensive but not overwhelming foliage. The grass is mostly yellowish and dusty.
- Cars, red and light blue or white, are parked on the parking lot. The red car is a hatchback. The cars, either 4 or 2, are all compact models. All cars are parked behind the trees on what may be a parking lot.
- A building that has yellow corner is behind the cars.
- It is a bright sunny day; the sky is blue with light clouds.

From this description, it can be noticed that the hypothesis of the central element of the picture being similarly described between all participants could not be accepted. Interestingly, the descriptions varied in movement information—it could be argued that it is not salient to the central object itself—but not in the elephant's location on the road. The description of the elephant and its body parts did not vary as much between any 2 subjects as between all of them. It should also be noticed that there was no contradictory description of the elephant itself. Thus, perhaps a better metric would be to find overlap used by the majority of the participants, as opposed to all, for real-world applications.

The second hypothesis, namely the difference in the background descriptions due to focus on different elements could not be rejected based on this small set. Between the objects that were noticed by all participants, the description varied more than that of the central object, and often the information was contradictory. For example, there was no agreement on the number of cars in the picture or their colors and very different description of greenery.

Computational Description

Computational overlap, as expected, was clustered around objects. Thus, the following concepts were identified: ELEPHANT, ROAD, CAR, TREE. Additionally, the following descriptions of the concepts were found:

```
undetermined_event
    (agent(value(elephant1)))
    (location(value(road1)))
car1
    (behind(value(tree1(number(greater-than(1))))))
put2
    (instrument(value(car1)))
    (location(sem(parking-lot)))
```

In plain English, it says that there is an elephant that is doing something on the road, there is a car behind trees, and somebody left a car in the parking-lot. Clearly, what is missing here from the overlap found by a human is that there are trees in some special relation to the elephant.

The union of information was not as successful due to coreference resolution mistakes (with STAn's coreference module not yet fully activated), however, the trivial unions of information were found. The number of unconnected clusters of information was small enough, that based on the concepts connected through the overlap above, it is possible to conclude that the three stories described similar information.

Perhaps it is worthwhile to demonstrate the computational process in the discovery of the overlap. Consider the following sentences:

- (1) A large grey elephant is moving on a road or bare ground.
- (2) This is a photograph of an elephant crossing a road. It is a large male African elephant.
- (3) Elephant is on asphalted road.

The sentences result in the following TMRs:

```
(1) land-animal-motion1
      (phase(value(continue)))
      (agent(value (elephant1)))
        (color(value(grey)))
      )))
      (location(value(road1 ground1)))
(2) pred1
      (theme(value(photograph
            (representation-of(value(change-location1
                (agent(value(elephant1)))
                (path(value(road1)))
            )))
      )))
pred2
      (theme(value(elephant1
            (size(value(large)))
            (gender(value(male)))
            (location(pnd(Africa)))
      )))
(3) exist1
      (agent(value(elephant1)))
      (location(value(road1
            (made-of(value(asphalt)))
```

)))

From the above descriptions, we know the following about the elephant:

From (1): <land-animal-motion1, elephant1> $\in \mathcal{I}$ [agent]

From (2): <change-location1, elephant1> $\in \mathcal{I}$ [agent]

From (3): <exist1, elephant1> $\in \mathcal{I}$ [agent]

Taking the intersection of the events for which the elephant is an agent results in x $\in \mathcal{I}$ [event]. Thus, producing undetermined_event(agent(value(elephant1))).

Continuing with each TMR, we find the following:

From (1): <land-animal-motion1, road1> $\in \mathcal{I}$ [location]

From (1): <land-animal-motion1, ground1> $\in \mathcal{I}$ [location]

From (2): <change-location1, road1> $\in \mathcal{I}$ [path]

From (3): <exist1, road1> $\in \mathcal{I}$ [location]

It can be easily noticed that ground1 occurs only in (1), thus the intersection with (2) and (3) results in an empty set. For road1, the calculation is similar to that of an elephant with the only addition of parent-child relationship of location and path.

It should also be noted that if we were to find an overlap of (1) and (2) and discarded (3), the event in question would have a considerably finer grain. According to the ontology, the most specific ancestor of both LAND-ANIMAL-MOTION and CHANGE-LOCATION is CHANGE-LOCATION. This means that while the sentences used different verbs to describe the movement of the elephant (crossing and moving), the OST understands what both mean and finds the general concept for both, as opposed to ignoring the similarity in meaning.

Similar processing is done for all sentences, resulting in the above relationship for car1 and put2 in addition to elephant.

The calculation of overlap is done in a similar manner, with the exception of the selection rules: each pair of concepts does not have to overlap in the found properties, instead uniquely found relationships are added to the existing set.

Conclusion

This paper was an attempt to investigate whether a computer is capable of finding similar information in structurally different texts that describe the same event, each focusing on potentially different parts of the event. The goal was not to find similar words in texts, which can be easily done, but to meaningfully connect the overlapping concepts and relationships used in the text descriptions. The approach is radically different from the machine-learning one. The performance of a machine in finding similarity was compared to human performance. The machine matched four out of five human findings.

It is too early to reach a conclusion that it is possible for computers to find overlap and difference between texts similarly to those that humans find, and, of course, more extensive experiments should be conducted. However, it is promising that the first result is not negative.

Acknowledgements

The author is grateful to Victor Raskin and the anonymous reviewers for their comments and to RiverGlass Inc for permission to use examples from their proprietary resources.

References

Carroll, D. (2004), *Psychology of Language***,** Thompson Wadsworth, Belmont, California, 2004

Clough, P., Gaizauskas, R., Piao, S. & Wilks, Y. (2002) METER: MEasuring TExt Reuse. In Proceedings of the 40th Anniversary Meeting for the Association for Computational Linguistics (ACL-02), pages 152–159, Pennsylvania, PA.

Fernando, S. & Stevenson, M. (2008) A semantic approach to paraphrase identification. In *Proceedings of the 11th Annual Research Colloquium of the UK Special-interest group for Computational Lingusitics*, Oxford, England.

Gruber, T. R. (1993) A translation approach to portable ontology specification. *Knowledge Acquisition*, 5, 199-200

Gruber, T. R. (1995) Toward principles for the design of ontologies used for knowledge sharing. In N. Guarino and R. Poli, (Eds.), Special Issue: The Role of Formal Ontology in the Information Technology. *International Journal of Human and Computer Studies* 43(5-6), 907-928

Hempelmann, C.F, Taylor, J. M., & Raskin, V. (2010) Application-guided Ontological Engineering, In *Proceedings of International Conference on Artificial Intelligence,* Las Vegas, Nevada

Nirenburg S., & Raskin, V. (2004) *Ontological Semantics.* Cambridge, MA: MIT Press

Raskin, V., Hempelmann, C. F., & Taylor, J. M. (2010) Guessing vs. Knowing: The Two Approaches to Semantics in Natural Language Processing, In *Proceeding of Annual International Conference Dialogue 2010*, Moscow, Russia

Qiu, L., Kan, M.Y, & Chua, T. S. (2006) Paraphrase recognition via dissimilarity significance classification. In Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, pages 18–26, Sydney, Australia, July. Association for Computational Linguistics.

Talmy, L. 2000. Toward a cognitive semantics, vols. 1-2. Cambridge, MA: MIT Press

Taylor, J. M., Hempelmann, C. F., & Raskin, V. (2010) On an Automatic Acquisition Toolbox for Ontologies and Lexicons in Ontological Semantics, In *Proceedings of International Conference on Artificial Intelligence,* Las Vegas, Nevada

Zhang, Y. & Patrick, J. (2005) Paraphrase identification by text canonicalization. In Proceedings of the Australasian Language Technology Workshop 2005, pages 160–166, Sydney, Australia, December.

Do Baseball Fans Experience the Fan Effect?

Travis R. Ricks (tricks2@uic.edu)
Jennifer Wiley (jwiley@uic.edu)

[1]Department of Psychology
1007 W. Harrison St. M/C 285
University of Illinois, Chicago
Chicago, IL 60607 USA

Abstract

A set of studies examines whether domain knowledge for baseball will enable participants to overcome the fan effect from baseball-related sentence sets. In a first study, neither high nor low knowledge participants overcame the fan effect when baseball positions and locations were randomly paired together. In a second study, when positions and locations were consistent with baseball expectations, both high and low knowledge participants overcame the fan effect on target sentences. However only high knowledge participants showed no effect on foils. The results suggest that prior knowledge can affect both representation and decision phases underlying recognition memory.

Domain-related Knowledge and Memory

Research on expertise has generally found that possession of domain-related knowledge or experience leads to superior problem solving, learning, and memory performance (see Feltovich, Prietula, & Ericsson, 2006 for review). The superior performance is thought to be due to extensive, easily accessible and well-connected knowledge structures in long-term memory (Bedard & Chi, 1992; Ericsson & Kintsch, 1995; Ericsson & Staszewski, 1989) which allows for more connections or associations to be made with incoming stimuli. Interestingly, another body of research suggests that increasing the number of associations among incoming stimuli can lead to a detriment in memory performance (See Reder et al., 2007 for review). This phenomenon, called the *fan effect*, refers to the slowdown in verification time that occurs as a function of the number of associations with a presented concept (e.g., Anderson & Bower, 1973; Lewis & Anderson 1976; Reder, Donavos, & Erickson, 2002).

What is the Fan Effect?

Typically the *fan effect* is demonstrated by having participants study sets of sentences that vary in the number of associations stated between concepts such as objects and locations (Anderson, 1974; Reder et al., 2007). The number of associations that each object or location is paired with is the "fan" size, and it usually varies between one and three.

For example, participants could be presented with the following sentences:

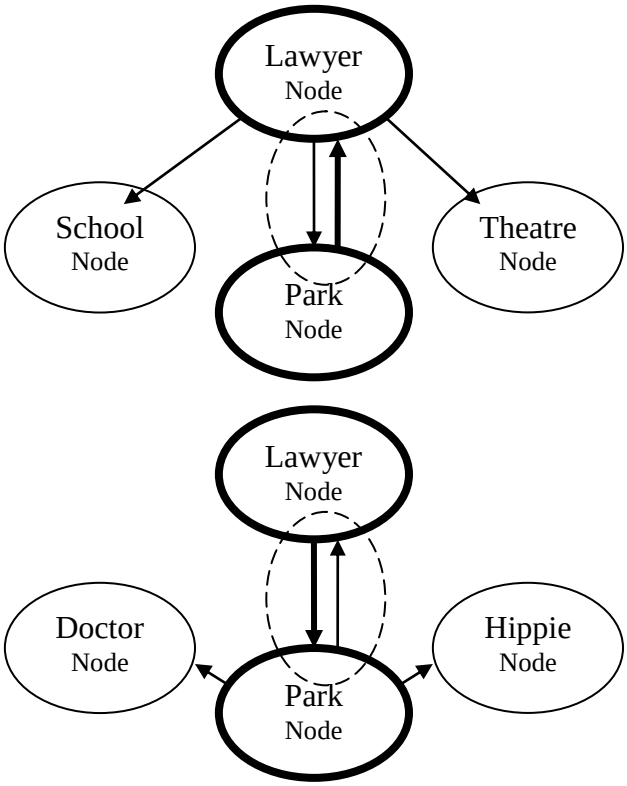
The lawyer is at the school.
The lawyer is at the park.
The lawyer is at the theater.
The doctor is at the museum.

In this example, the lawyer has a fan size of three and the doctor has a fan size of one. Participants are required to memorize these sentences to some criterion during the study phase. Then, after reaching criterion, participants move to a recognition test phase where they are asked to decide as quickly as possible whether or not sentences appeared in the study list. Typically, participants take longer to verify that the statement "The lawyer is at the school" appeared in the study list than "The doctor is at the museum," due to the larger fan of lawyer in this set.

There are two main accounts that have been offered for the *fan effect*: the propositional network theory (Anderson, 1974; Reder et al., 2007) and the situation model theory (Radvansky, Spieler & Zacks, 1993; Radvansky, 1999).

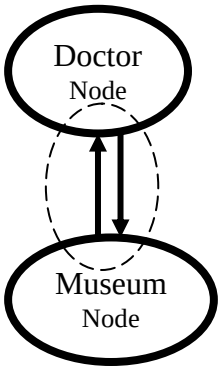
The propositional theory suggests that the fan effect is a function of the number of pathways that branch from a target concept in a memory network and a corresponding reduction in the spread of activation. To illustrate this, imagine that nodes exist in memory that correspond to the presented concepts (e.g. lawyer, park, doctor, school, park, theater, museum). When participants are required to verify if they have seen a sentence such as "The lawyer is in the park," the nodes "lawyer" and "park" are activated. As shown in the top of Figure 1, activation spreads along all connecting pathways that exist, represented by the arrows. For the "lawyer," with the fan of three, activation is diffused among the pathways connected to "school," "park," and "theater." Similarly, if three statements are presented involving "park", as shown in the bottom of Figure 1, then activation would also be diffused in that case.

Figure 1:
Two models with fan sizes of three



In contrast, the sentence, "The doctor is at the museum," has a fan size of one. As shown in Figure 2, there will be no diffusion of activation in this representation because there is only one pathway branching from each of the nodes "doctor" and "museum."

Figure 2:
A model with fan size of one



In the fan of three examples, the partitioning of activation among associations decreases the amount of activation that spreads to each connecting pathway, which increases the amount of time it takes participants to become aware of the target pathway (Jones & Anderson, 1987). In this model, the distribution of activation to irrelevant pathways is called interference (Anderson, 1974), and indeed the empirical results have generally supported that the number of associations predicts verification time (Reder et al., 2007).

However, one exception in previous empirical results, has been the observation that not all fans of three are equal. The situation model account suggests that the type of fan represented by the bottom network in Figure 1 will experience less interference than the type of fan represented in the top panel. The situation model account posits that when people can integrate incoming information into a single representation then they will not be susceptible to the interference due to multiple associations (Radvansky, 1999). In this explanation, slower verification times are not the result of the number of associations that are present, but rather the number of models that need to be searched. From this perspective, one can imagine all three sentences in the second example could refer to a single situation in the park, perhaps with the lawyer meeting the hippie and the doctor. If these three sentences are integrated into one representation in memory, then even though there are three items associated with park, there is only one model to search. Consistent with this approach, several studies have demonstrated that the ability to integrate sentences into a single representation or model can eliminate the fan effect (Gomez-Ariza & Bajo, 2003;Moeser 1979; Myers, O'Brien, Balota & Toyofuku, 1984; Radvansky, Spieler & Zacks, 1993; Smith, Adams & Schorr, 1978).

Although it has not yet been tested, one implication of this model is that participants may be able to overcome the fan effect for domain-related information, as prior knowledge may allow readers to represent and integrate sets of sentences into a single model. Thus, the goal of this research was to investigate if the possession of prior knowledge related to the topic of the sentences would eliminate the fan effect in recognition memory.

Experiment 1

Method

Participants. Participants were 110 students in introductory psychology classes at University of Illinois at Chicago who received course credit for their

participation.

Procedure. Participants were administered a baseball-related fan task in groups ranging in size between 1 to 12. The sessions last approximately 1 hr. The stimuli were created by randomly pairing a type of baseball player (e.g., catcher) with a location on a baseball field (e.g., second base) to create sentences (e.g., The catcher is at second base). The task was analogous to the fan task used by Radvansky and Zacks (1991) with participants being presented 18 sentences and being asked to memorize them. The 18 sentences contained 4 at fan size 1, 4 at fan size 2 and 10 at fan size 3.

Each participant was seated at their own computer. During the study phase, the sentences were presented on a computer screen one at a time for 7-seconds each. After the study phase, participants we retested for their memory of the sentences. If participants were unable to remember 90 percent of the sentences correctly they repeated the study and test phase. Feedback was provided for incorrect answers during the test phase. This cycle was continued until participants reached the 90 percent criterion.

After the participants reached the 90 percent criterion, the completed a speeded recognition task. Twelve target sentences were presented from the studied materials (four sentences at each of the fan sizes of one, two, and three; Similar to Radvansky and Zacks (1993), studied sentences that had both a player and a location with more than one association were not used.).

Twelve foils were created by repairing the studied players and locations. The re-pairing was done within the fan size so fan 1 player/locations were re-paired with fan 1 player/locations. Participants pressed the "Z" key if the sentence was not studied and "M" if it was studied.

At the end of the study, participants completed a 45-item baseball knowledge questionnaire (Spilich, Vesonder, Chiesi & Voss, 1979). Average performance on the baseball questionnaire was 16.47 (SD 12.66) Range was 0 to 41. Two levels of domain knowledge were defined by a median split at 15. All participants' accuracy was above 90% on the speeded recognition task and there was no significant difference for accuracy between high and low knowledge participants.

Results

A 2 X 3 mixed ANOVA was used to assess the effects of Fan Size (one, two or three), and Expertise (high, low baseball knowledge) on correct verification RT. Similar to Radvansky, Spieler, and Zacks (1993) responses that were faster than 500 ms and slower than 10,000 ms were considered errors. The pattern of results is shown in Figure 1. The ANOVA revealed a main effect for Fan Size, $F(2, 216) = 7.36$, $p < .001$, $\eta^2 = .06$, and Expertise, $F(1, 108) = 8.24$, $p < .01$, $\eta^2 = .06$, but not a Fan Size X Expertise interaction, $F(2,216) < 1$, $\eta^2 = .01$. As expected, participants experienced a slowdown in the recognition test as the number of associations increased from 1 to 3. There was also a main effect such that high knowledge participants made faster decisions than low knowledge participants. However, neither high knowledge nor low knowledge participants overcame the fan effect. Both high and low knowledge participants experienced increasing verification times as fan size increased. The same pattern of results was observed for the studied and foil sentences in this study.

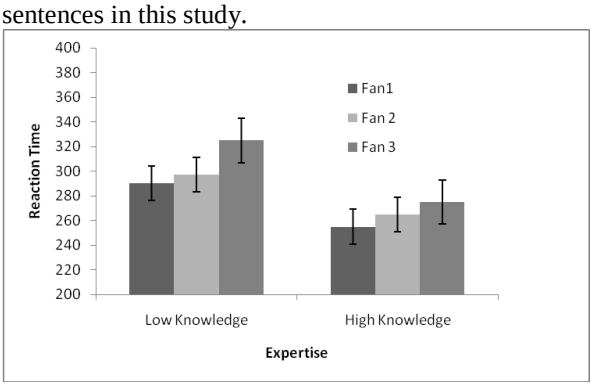


Figure 1: Verification time per syllable (ms) with random pairings in Experiment 1. Error bars represent standard errors.

These results provide a replication of the basic fan effect finding. As fan size increased, so did verification times. Simply having prior knowledge for the topics of the sentences did not change this pattern. However, because the pairings were random, it is possible that this task did not test the situation model account. In order to support the construction of a single model, one may need sentences that "make sense" within the domain. According to the situation model theory, the fan effect should not be eliminated unless participants are able to integrate the multiple players and locations into a single model. Randomly pairing the players and locations together made pairs that were not consistent with baseball experience and which may not have made it any easier for high knowledge participants to integrate sentences into single situations. Thus, in Experiment 2, we presented sentences that were more consistent with baseball situations.

Experiment 2

The goal of this second study was to use sentences that were more consistent with real baseball situations, and to test whether prior knowledge might affect performance under those circumstances. For this study, players and positions were paired to reflect plausible situations, such as:

The reliever is at the mound.
The manager is at the mound.
The catcher is at the mound.

These sentences could represent a pitcher conference, an event that happens in the majority of baseball games.

Method

Participants. Participants were 110 students in introductory psychology classes at University of Illinois at Chicago who received course credit for their participation. These were new participants that had not participated in Experiment 1.

Procedure. Participants were administered a baseball-related fan task almost identical to the one administered in Experiment 1. The only difference was that that the players and positions were not randomly paired together, but were paired to create plausible sentences by the researcher.
The 12 foils were also consistent with baseball expectations. Some example foil sentences were:

The pinch runner is at second base.
The reliever is at first base.
The pitcher is at the bullpen.

Participants again completed a 45-item baseball knowledge questionnaire at the end of the study. Average performance was 14.53 (SD = 13.35). Range was 0 to 41. A median split of 15 was used similar to Experiment 1. All participants' accuracy was above 90% on the recognition task and there was no significant difference between high and low knowledge participants.

Results:

A 2 X 3 mixed ANOVA was used to assess the effects for Fan Size and Expertise on correct verification RT. Again responses that were faster than 500 ms and slower than 10,000 ms were considered errors. The pattern of means can be seen in Figure 2. The ANOVA revealed a main effect for Fan Size, $F(2, 216) = 7.82$, $p<.01$, $\eta^2=.07$, but no main effect for Expertise $F(1, 108) < 1$, $\eta^2=.01$. However, there was a Fan Size X Expertise interaction, $F(2, 216) = 7.16$, $p < 01$, $\eta^2=.07$.

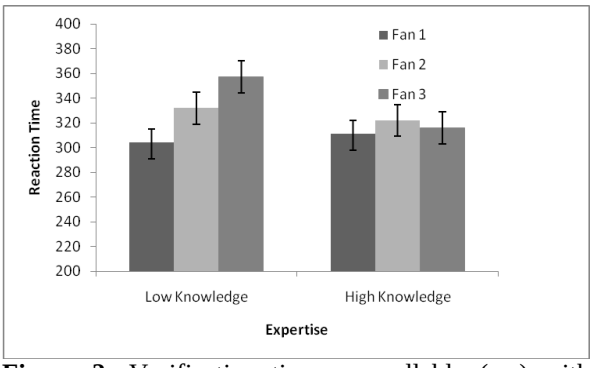


Figure 2: Verification time per syllable (ms) with plausible pairings in Experiment 2. Error bars represent standard errors.

As in Experiment 1, participants experienced an overall slowdown in verification time as the number of associations increased from 1 to 3.

However, this effect was qualified by a significant Fan Size by Expertise interaction, with low knowledge participants showing the typical fan effect, and high knowledge participants showing a diminished fan effect. On the face of it, these results can be seen as consistent with the situation model account. They suggest that, now that the sentences are plausible, participants with prior knowledge may be able to create a single model for each set of sentences, which allows for efficient search of memory, regardless of the number of overlapping associations.

What is responsible for the elimination of the fan effect among high knowledge participants? To further examine this question, we performed some additional analyses and in particular we examined whether performance improved on both target and foil trials. If the better performance among high knowledge participants is due to the efficiency of needing to search only a single model, then this account would predict facilitation for both correct acceptance of targets and rejections of foils. However, as can be seen in Figure 3, different patterns were found across target (top panel) and foil (bottom panel) decisions.

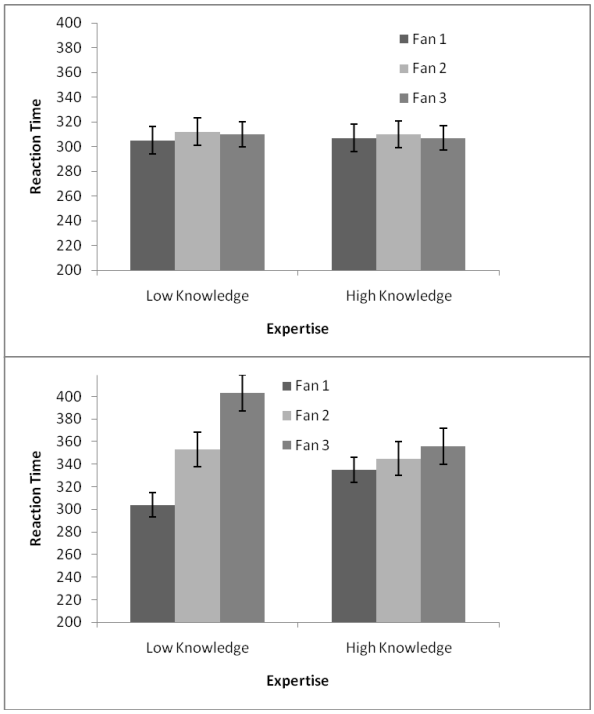


Figure 3: Reaction times per syllable (MSec) for studied sentences (top panel) and foils (bottom panel) . Error bars represent standard errors.

First, looking at performance on the studied sentences, here we can see that in fact, no fan size effect was found for either knowledge group. A 2 X 3 mixed ANOVA showed that there were no significant effects for Fan Size, F (2, 216) = 1.91, p <.15, η^2=.05, or Expertise, F (1, 108) = 1.31, p < .27, η^2=.01. Nor was there a Fan Size X Expertise interaction, F (2,216) = 1.60, p < .21, η^2=.03. Thus, neither high nor low knowledge participants experience a fan effect on correct verifications for plausible sentences.

Quite a different picture is seen when one examines the response times for the foils. Here another 2 X 3 mixed ANOVA revealed a main effect for Fan Size, F (2, 216) = 20.16, p<.01, η^2=.23. Although there was no main effect for Expertise, F (1, 108) < 1, η^2=.02, there was a Fan Size X Expertise interaction, F (2,216) = 8.17, p < 01, η^2=.07. Low knowledge participants were especially vulnerable to the slowdown on foils as fan size increased.

The results show that the fan effect was diminished for experts on both correct verifications and rejections, while novices experienced a fan effect, and this was driven by decision times on the foils.

DISCUSSION

According to the situation model account, the fan effect should be eliminated when participants are able to integrate multiple associations from a set of sentences within a single situation model. Consistent with this approach, it appears that presenting participants with sentence sets representing plausible combinations of baseball players and positions enabled both high and low knowledge participants to respond quickly to studied sentences, regardless of the number of associations among them. This finding is consistent with the situation model account. It is also reminiscent of findings that have demonstrated that the fan effect is diminished when participants are able to integrate the sentence sets into stories (Ariza & Bajo, 2003; Myers, O'Brien, Balota & Toyofuku, 1984; Smith, Adams & Schorr, 1978).

In addition, the further analysis of the studied and foil sentences separately revealed the interesting result that the non-studied foils showed a different pattern of verification times than the studied sentences. When participants were presented with foil sentences that were also consistent with realistic baseball situations, the performance of high and low knowledge participants diverged. High knowledge participants experienced a diminished fan effect. However, low knowledge participants foil response times were more affected by fan size. Thus, it does not appear that the low knowledge participants were able to efficiently reject the foils. If both low and high knowledge participants were able to form single models from the sentence sets, decisions on foils should have been as easy as on targets. Only the high knowledge participants showed this advantage.

Thus, this result highlights another recent perspective from the fan effect literature which emphasizes that recognition memory results need to be thought of as both being a function of differential representation in memory, as well as being a function of decision making processes (Anderson, 1999). In essence, making a recognition judgment requires not just memory retrieval or search, but also an evaluative assessment or decision. This current dissociation between performance on targets and foils suggests that high and low knowledge participants may be achieving fast verification times to studied sentences via different means. While high-knowledge participants may have the advantage of a single model which allows for fast, direct retrieval for each set of sentences, the low-knowledge participants may have used some sort of plausibility heuristic during the verification task. This improved their performance for the studied items, but made it difficult for them to reject the foils.

An alternative explanation is that the quality of the

memory representations for the sentence sets differed among the low and high knowledge participants. It is possible that the high knowledge participants were able to create more detailed or distinctive traces for the sentence sets, which improved their ability to decide both what was studied and what was not (Hunt & Einstein, 1981). Low knowledge participants on the other hand, may have had "good enough" representations to aid performance on the studied items, but perhaps these traces were not detailed enough to aid them on the foils. Such a result would be consistent with a few recent findings that expertise can confer advantages in episodic memory (i.e. memory for words and order in domain-related word lists) (Rawson & Van Overshelde, 2006; Ricks & Wiley, 2009).

While most previous studies have suggested that reductions in the fan effect are due to unitized representations, thepresent results suggest that effects on decision processes are critical to consider. However, decision processes can only be explored when one uses plausible foils that require detailed memory for the studied sentences. In present study, thematic materials allowed all participants to avoid fan effects for the studied sentences, but only high knowledge participants were better able to detect foils. Thus the present design allowed for a clearer understanding of how prior knowledge may support both better integration and discrimination in recognition memory.

Acknowledgements

We would like to thank Allison Jaeger and Kari Andrews for their diligent data collection. Travis Ricks also thanks his wife for her support while running subjects and preparing the manuscript.

References

Anderson, J. R. (1974). Retrieval of propositional information from long-term memory. *Cognitive Psychology*, 6, 451-474.

Anderson, J. R., & Bower, G. H. (1973) *Human associative memory. Washington DC: Winston.& Sons.*

Anderson, J. R., & Reder, L. M. (1999). The fan effect: New results and new theories. *Journal of Experimental Psychology: General*, 128(2), 186-197.

Bedard, J. & Chi, M. T. H. (1992) Expertise. *Current Directions in Psychological Science*, 1(4), 135-139.

Ericsson, K. A., & Kintsch, W. (1995). Long-term working memory. *Psychological Review*. 102(2), 211-245.

Ericsson, K. A., Staszewski, J. J. (1989). Skilled memory and expertise: Mechanisms of exceptional performance. In Complex Information Processing: The Impact of Herbert Simon (183-208). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Feltovich, P. J., Prietula, M. J., & Ericsson, K. (2006). Studies of expertise from psychological perspective. In K.A. Ericsson, N. Charness, P. Feltovich, & R. Hoffman (Eds.), Cambridge Handbook of Expertise and Expert Performance (41-67). New York: New York: Cambridge University Press.

Gómez-Ariza, C. J., & Bajo, M. T. (2003). Interference and integration: The fan effect in children and adults. *Memory*, 11, 505-523.

Hunt, H. R., & Einstein, G. O. (1981). Relational and item-specific information in memory. *Journal of Verbal Learning & Verbal Behavior*, 20, 497-514.

Jones, W. P., & Anderson, J. R. (1987). Short-and long-term memory retrieval: A comparison of the effects of formation load and relatedness. *Journal of Experimental Psychology: General*, 116, 137-153.

Lewis, C. H., & Anderson, J. R. (1976). Interference with real world knowledge. *Cognitive Psychology*, 8, 311-335.

Moeser, S. D. (1979). The role of experimental design in investigations of the fan effect. *The Journal of Experimental Psychology: Human Learning and Memory*, 5, 125-134.

Myers, J. L., O'Brien, E.J., Balota, D. A., & Toyofuku, M. L. (1984). Memory search without interference: The role of integration. *Cognitive Psychology*, 16, 217-242.

Radvansky, G. A. (1999). The fan effect: A tale of two theories. *Journal of Experimental Psychology: General*, 128, 198-206.

Radvansky, G. A., & Zacks, R. T. (1991). Mental models and the fan effect. *Journal of Experimental Psychology Learning, Memory, and Cognition*, 17 (5), 940-953.

Radvansky, G. A., Spieler, D. H. & Zacks, R. T. (1993). Mental model organization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 19, 95-114.

Rawson, K. A., & Van Overshelde, J. P. (2006). How does knowledge promote memory? Contribution of organizational and item-specific processing. *Journal of Memory and Language*, 58(3), 646-668

Reder, L. M., Paynter, C., Diana, R. A., Ngiam, J., & Dickison, D. (2008). Experience is a double-edged sword: A computational model of the on encoding/retrieval tradeoff with familiarity. In Ross, B. & Benjamin, A. S. (Eds.), The Psychological of Learning and Motivation, Academic Press.

Reder, L.M., Donavos, D.K., & Erickson, M.A. (2002). Perceptual match effects in direct tests of memory: The role of contextual fan. *Memory & Cognition*, 30(2), 312-323.

Ricks, T. R., & Wiley, J. (2009). The influence of domain knowledge on the functional capacity of working memory. *Journal of Memory and Language*, 59, 519-537.

Smith, E. E., Adams, N., & Schorr, D. (1978). Fact retrieval and the paradox of interference. Cognitive Psychology, 10, 438-464.

Spilich, G. J., Vesonder, G. T., Chiesi, H. L., & Voss, J. F. (1979). Text processing of domain-related information for individuals with high and low domain knowledge. *Journal of Verbal Learning and Verbal Behavior*, 14, 506-522.

Prior expectations in pedagogical situations

Patrick Shafto[1], Noah D. Goodman[2], Ben Gerstle[1], & Francy Ladusaw[1]
[1] Department of Psychological and Brain Sciences, University of Louisville
[2] Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology

Abstract

Much of human learning occurs in social situations, and among these, pedagogical situations may afford the most powerful learning. In pedagogical situations, a teacher chooses the concept that they are going to teach and the examples that they use to teach the concept. If learners know that a teacher is helpful and understands the implications, this could support strong inferences. In previous work, Shafto and Goodman (2008) proposed and tested a model of pedagogical data selection. We integrate special-purpose pedagogical expectations in this framework, and derive a task that allows independent assessment of pedagogical expectations. Two experiments contrast people's expectations about pedagogical and communicative situations. The results show that people's expectations differ in these situations, and that in pedagogical situations people expect teachers to present generalizable and semantically coherent knowledge. We discuss the implications for modeling learning in pedagogical settings, as well as for understanding human learning more broadly.

Keywords: Pedagogy; Learning; Bayesian Model

Much of human learning occurs in social contexts. We learn from siblings, parents, friends, and teachers by observing, imitating, and teaching. Among these social learning settings, pedagogical situations stand out as potentially the most important. Pedagogical situations are situations in which one person, a teacher, chooses information for the purpose of helping another person, a learner, arrive at some belief. Pedagogical situations might provide uniquely powerful learning situations, especially if learners are privy to, and understand the implications of, teachers' intentions to help.

Indeed, recent theories argue that an intuitive understanding of pedagogical situations may be what sets us apart from other animals (Csibra, 2007). Under this proposal, learners' intuitive understanding of pedagogical situations consists of two components: inferences how teachers choose examples to teach a concept, and expectations about what kinds of concepts teachers are more likely to teach.

The issue of how teachers choose information and learners' understanding of these situations have been investigated in detail (for a review, see Csibra and Gergely, 2009). Recently, Shafto and Goodman (2008) have proposed a computational model of reasoning in pedagogical situations. This account provides a formal explanation of why and how teachers decide which examples to choose, and how learners can capitalize on the teacher's intent to make stronger inferences.

Researchers have also argued that young children come prepared with expectations about what kinds of knowledge to expect in pedagogical situations. Specifically, Csibra and Gergely (2009) argue that very young children expect that knowledge provided in pedagogical contexts is semantically generalizable. For instance, Topal et al. (2008) show that children make A-not-B errors in pedagogical contexts, but not in neutral contexts. They argue that the perseverative errors are a consequence of children misinterpreting initial pedagogical demonstrations as indicating that the A box is where the ball belongs. While these results are quite compelling, they contain influences of both the learner's inference about the teachers' choice of data, and the learners' expectations about what kinds of properties are likely to be taught.

In the current paper, we investigate the hypothesis that people expect semantically generalizable knowledge in teaching situations. We begin by discussing the role of prior knowledge in pedagogical reasoning, and how this can be integrated with Shafto and Goodman's (2008) model of pedagogical reasoning. We then use this framework to develop a method for separating the role of pedagogical priors from pedagogical data selection. Two experiments use this method to investigate whether adults expect generalizable knowledge (Experiment 1) and whether adults expect semantically coherent knowledge (Experiment 2). In each case, we contrast pedagogical situations with communicative situations to address whether these prior expectations are specific to pedagogical contexts. We conclude by discussing implications for modeling human learning and understanding reasoning in social situations.

The role of priors in pedagogical reasoning

The proposal that learners expect generalizable information can be integrated naturally into a Bayesian reasoning framework. From this perspective the problem of learning is one of inferring the probability of different hypotheses, h, given observed data, d. Bayes' theorem provides a way of updating our posterior beliefs about hypotheses, $P(h|d)$, given prior beliefs, $P(h)$, and as-

[1]Please address correspondence to Patrick Shafto, p.shafto@louisville.edu

sumptions about how data are sampled, $P(d|h)$,

$$P(h|d) \propto P(d|h)P(h). \quad (1)$$

In standard Bayesian learning, it is typically assumed that the prior, $P(h)$, is determined by some stochastic generative process, and data are sampled from a hypothesis, $P(d|h)$, randomly; however, these standard assumptions are not appropriate for pedagogical situations. In pedagogical situations, a teacher chooses both the hypothesis that they are going to teach, and the data that they use to teach the hypothesis. It is, therefore, reasonable for the leaner to expect that teachers' choices are not random, but are instead purposeful.

Recent work has formalized pedagogical sampling – how teachers may choose data for the purpose of teaching a hypothesis (Shafto and Goodman, 2008). This model suggests that learners may use the knowledge that the teacher is choosing data for the purpose of teaching them a hypothesis, to replace the random sampling with an assumption that the teacher will choose data that tend to make the learner believe the correct hypothesis,

$$P(d|h) \propto P(h|d). \quad (2)$$

Their results show that this pedagogical sampling assumption allows for stronger inferences than random sampling. They also provided evidence that people's inferences differ in pedagogical and ostensibly random contexts, and the pedagogical model accurately predicted people's intuitive pedagogical inferences.

Here, the question is whether people have prior expectations about which kinds of hypotheses should be taught. That is, are there a specific set of prior probabilities, $P(h)$, that apply to teaching situations. Intuitively, the question is whether learners expect teachers to choose particular hypotheses that are important or worthy of teaching. For example, more general hypotheses might be expected because the knowledge is more likely to be applicable in future situations. Formally, we capture these expectations as a utility function which defines how hypotheses are weighted in pedagogical situations, $U(h; pedagogy)$. Integrating this prior into Equation 1 would allow us to capture how prior expectations, specific to pedagogical situations, affect inferences.

Investigating pedagogical priors

In this paper, we attempt to answer two questions: (1) what are the characteristics of people's pedagogical prior expectations, and (2) are these expectations specific to pedagogical situations? To address the second question, we need to choose appropriate (non-pedagogical) control conditions. To address the first question, we can ask people to make judgments about which of two teachers they would rather have teaching them, while varying the particular hypothesis that each is teaching. However, given our above analysis, this requires that we separate the contribution of pedagogical data sampling from the contributions the pedagogical prior expectations.

Judgments about the teachers can be formalized as an inference about whether the teacher chooses hypotheses

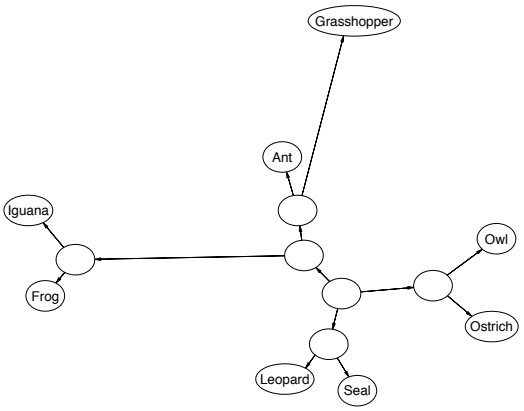


Figure 1: A tree representing the intuitive taxonomic relationships among 8 animals.

that the learner expects to be taught about. More formally, we assume a parameter, θ, which specifies how systematically a teacher chooses her examples. The probability of choosing hypotheses depends on how systematic the teacher is,

$$P(h|\theta) \propto e^{-\theta U(h)}, \quad (3)$$

where systematic teachers tend to choose hypotheses that have higher utility (Luce, 1959).

The learner can then infer how systematic a teacher is, given some data,

$$P(\theta|d) \propto \sum_h P(d|h)P(h|\theta)P(\theta) \quad (4)$$

where $P(\theta)$ is a prior distribution specifying whether people tend to be systematic or not. This equation states that teachers are considered systematic to the degree that they choose hypotheses that agree with the learner's prior expectations, as specified by Equation 3. However, in this inference, the influence of $P(h|\theta)$ is not isolated.

To see how we could isolate the effects of $P(h|\theta)$, consider hypotheses about properties of the the animals in Figure 1. The set of possible hypotheses can be defined extensionally, by enumerating all possible combinations of animals that have or do not have the property. For example, one hypothesis about a property is {ostrich=true, owl=false, grasshopper=false, ant=false, iguana=false, frog=false, leopard=false, seal=false}. Our goal is to eliminate the contribution of the sampling assumption, $P(d|h)$. Assuming that we want to teach the learner the hypothesis that all of the animals have a particular property, how would we choose which animal or animals to provide as examples? By presenting all of the data – each animal labeled as having or not having the property – we essentially choose one hypothesis. Thus, the contribution of the sampling of data is to simply indicate

a particular hypothesis. Formally, the $P(d|h) = 1$ for the true hypothesis, and zero for all others. Equation 4 reduces to,

$$P(\theta|d) \propto P(h|\theta)P(\theta). \quad (5)$$

Given the fully labeled data, the learner's judgments about the teacher's systematicity depend on whether the learner expects that hypothesis to be chosen, and their prior expectations about systematicity.

To isolate the influence of learners' prior expectations about hypotheses $P(h|\theta)$, we can ask learners to choose between two teachers. Because each teacher is equally likely to be systematic *a priori*, judgments about which of two teachers is preferred isolate the effects of a learner's prior expectations. Formally, the judgment becomes a ratio of two inferences, each individually specified by Equation 5,

$$\frac{P(\theta_1|d_1)}{P(\theta_2|d_2)} \propto \frac{P(h_1|\theta)P(\theta)}{P(h_2|\theta)P(\theta)} = \frac{P(h_1|\theta)}{P(h_2|\theta)}. \quad (6)$$

In the following, we present two experiments in which people make judgments about which of two teachers they want to have teaching them in the future (presumably the one that chooses a hypothesis that is more consistent with their expectations). In our investigations, we have two goals: (1) identifying the prior expectations that people bring to pedagogical situations, and (2) establishing whether these expectations are unique to pedagogical situations. The experiments test two claims related to prior expectations about pedagogical situations: that learners expect more generalizable information, and that learners expect semantically coherent information.

Experiment 1: Testing the bias toward generalizability

Experiment 1 investigated whether people have an expectation that teachers would teach generalizable information. To investigate this question, we choose a domain for which we have a good understanding of the possible hypotheses, the domain of animals. Figure 1 shows the animals, and the intuitive taxonomic relations among these animals.[2] We operationalize generalizable concepts here as a concept that is true of a broader class of animals.

To investigate whether people expected generalizable knowledge, we presented participants with scenarios in which pairs of teachers taught concepts of different levels of generality. The generalizable teacher taught a property that was consistent with the tree structure and was true of a greater number of exemplars. For instance, the generalizable teacher might teach a property that was true of all 8 animals, while the less generalizable teacher might teach a property that was true of only ostriches and none of the other animals. If people expect teachers to teach generalizable information, we expect to find that people choose the teacher who teaches properties that were true of broader sets of examples.

[2]The tree was derived using the tree learning algorithm and a subset of the animals used in Kemp and Tenenbaum (2008).

Methods:

Participants: Twenty-four university undergraduates participated in this experiment in exchange for course credit. Participants were randomly assigned to the pedagogical or the communication scenarios.

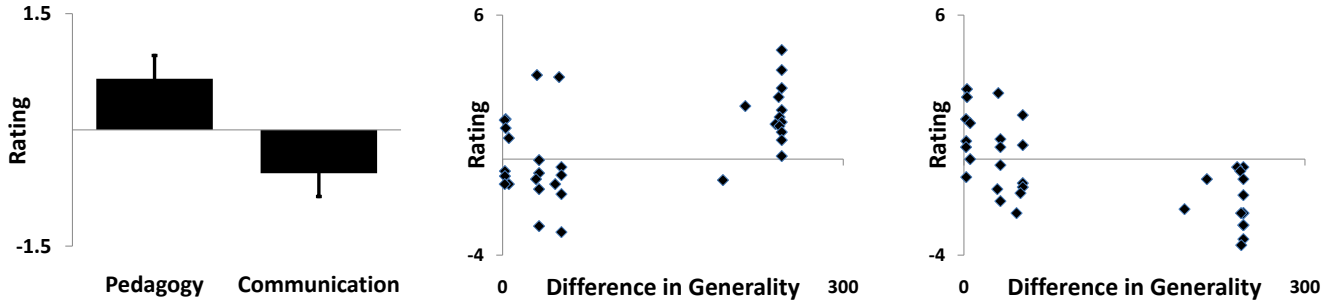
Procedure: In the pedagogical situation, people were presented with a series of questions asking them to decide which of two teachers they would like to learn from in the future. Each teacher was presented as teaching about a novel enzyme, e.g. "Teacher 1 is teaching about enzyme P23T." The names of the enyzmes were random combinations of letters and numbers. This was followed by lists indicating which of the eight animals had the enzyme and which did not. Each question contrasted two teachers, where teachers differed in the generality of the properties taught. For instance, one teacher might teach a property that was true of owls, ostriches, leopards, and seals, but not of grasshoppers, ants, iguanas, and frogs, while the other was teaching a property that was true of all eight animals. Paired teachers always taught concepts where one was a subset of the other, so the more generalizable concept included all of the positive examples of the property in the less generalizable concept, with additional positive examples (e.g. ostrich versus ostrich and owl). Participants indicated which teacher they would rather have teaching them about new enzymes using a Likert scale ranging from -10 to 10, where the extremes indicated the teacher on the left or the right and zero indicated indifference. Participants rated all possible pairings of teachers, resulting in a total of 34 questions. Order of the questions, as well as the side (right or left) of the more general concept, were randomized.

The communication condition was identical to the pedagogical condition, with the exception of some of the wording. Instead of teaching about enzymes, the situations described people who were talking about enzymes. For example, "Person 1 is talking about enzyme P23T." Additionally, participants were asked to provide ratings about which one they would rather talk to in the future. Otherwise, the questions and response sheets were identical.

Results & Discussion

We coded people's judgments as positive if they were in the direction of the more generalizable teacher and negative if they were in the direction of the teacher with the less general property. To test whether people expected more general properties, we compared the average ratings to chance (zero). In the pedagogical condition, people chose the teacher with the more general information, $mean = 0.66, t(407) = 2.06, p < 0.05$. In contrast, in the communication condition, people choose the less general information, $mean = -0.56, t(407) = -2.09, p < 0.05$. The difference between the two conditions was significant, $t(814) = 2.84, p < 0.01$. These results suggest that people expect that more general properties will be taught in pedagogical situations, in contrast with communicative settings, where people expect less general properties.

To follow up on these results, we investigated the pat-



(a) (b) (c)

Figure 2: Experiment 1 results: (a) Average human ratings in the pedagogy and communication conditions. Positive ratings indicate the more generalizable teacher. (b) Scatterplot showing the relationship between the difference in generalizability for pairs of teachers (x axis) and people's ratings (y axis) for the pedagogy condition. The strength of people's ratings increases with an increasing difference in generalizability, $r = 0.51$, indicating that they expect more generalizable concepts in pedagogical settings. (c) Scatterplot showing the relationship between difference in generalizability and people's ratings for the communication condition. The strength of ratings decreases with an increasing generalizability, $r = -0.66$, indicating that people expect less generalizable concepts in communicative settings.

tern of ratings for individual items. If people choose more generalizable concepts, then pairs for which there was the greatest gap between the more and less generalizable teacher should have the strongest ratings. To investigate this question, we needed to quantify how general each hypothesis was. We consider two possible measures of generality: the number of positive examples, and the sum of the distances among items in the tree. To test whether ratings indicated an expectation that properties would be generalizable, we collapsed individual judgments into a single average rating for each question, resulting in 34 ratings. To investigate which measure of generality best predicted people's judgments, we conducted a stepwise regression with item averages as the dependent variable. The independent variables included the number of positive examples in more general concept, the number of positive examples in the less general concept, the difference in number of positive examples, as well as the summed distances for the more and less general concepts, and the difference in the summed distance. The two difference scores allowed us to test whether people's judgments take into account both teachers, or just a single teacher when making their judgments. Stepwise regression greedily selects the variable that accounts for the greatest variance, and iterates until no variables account for significant variance. Analysis of the pedagogy condition showed that the difference in summed distances accounted for the greatest variance, $r = 0.51$, $F(1, 32) = 11.49, p < 0.01$, and that no other variables accounted for significant residual variance. The correlation indicates that the bigger the difference in generalizability was, the stronger people's ratings were toward the more generalizable teacher. In contrast, regression analyses on the communication condition showed that while the difference in summed distance was a significant predictor of ratings, the relationship was negative, $r = -0.66$, $F(1, 32) = 24.52, p < 0.001$. This suggests that in communicative settings, people's expectations about generalizability are the opposite of their expectations in pedagogical settings.

The number of positive examples, while a straightforward measure of generality for this task, is undesirable for two reasons. First, if this leads to an accurate characterization of people's inferences, then one might wonder to what degree the results are a consequence of task demands (given that people were answering questions about lists of animals). Second, the number of positive examples is not a very good measure of generality because it bears no necessary relationship with actual semantic generalizability. As can be seen in Figure 1, many possible sets with the same number of positive examples differ markedly in their coverage of the tree. Instead, we prefer to measure the generalizability of a concept in terms of the sum of distances between all pairs of positive examples. This provides a measure that is not subject to task demands, and is related to the semantic generality of the concept. Our analyses show that distance in the tree provides a better description of people's behavior, providing evidence that people's judgments do not simply reduce to task demands, and that their judgments are based on semantic generalizability.

It appears that people have strong prior expectations that they bring specifically to pedagogical situations. In pedagogical situations, learners expect that teachers will choose to teach generalizable information. In contrast, when in communicative situations, people expect that speakers are likely to talk about specific information. Our analyses showed that people's judgments are better predicted by distance in a semantic tree, consistent with a bias toward semantically generalizable information.

Experiment 2: Testing the bias toward semantic coherence

Experiment 2 investigated whether people have an expectation that teachers will choose semantically coherent concepts. To investigate this, we presented participants with scenarios in which two teachers each taught concepts with two positive exemplars. The semantically coherent teacher taught a property that was true of two tree-consistent exemplars, such as owl and ostrich. They were contrasted with a semantically incoherent teacher who taught a property that was true of two tree-inconsistent exemplars, such as ostrich and leopard. If people expect teachers to teach semantically coherent concepts, we expect to find that people choose the teacher who teaches tree-consistent properties.

Methods:

Participants: Twenty university undergraduates participated in this experiment in exchange for course credit.

Procedure: The procedure was identical to that used in Experiment 1 with the exception of the questions used. Each scenario provided information taught by two teachers. All properties were true of two animals, but were absent in the other six. In each scenario, one teacher taught a property that was semantically coherent – it was consistent with the structure of the tree – and the other taught a property that was semantically incoherent – it was inconsistent with the structure of the tree. For instance, a semantically coherent property might be true of owls and ostriches, but no other animals. Contrarily, a semantically incoherent property might be true of owls and leopards but no other animals. Questions were designed such that semantically coherent pairs were contrasted with all minimally different semantically incoherent pairs that overlapped one animal. For example, owls and ostriches were contrasted with owls and leopards, owls and seals, ostrich and leopards, and ostrich and seals. This resulted in a total of 16 questions. Order of the questions, as well as the side of the semantically coherent pair (left or right), were randomized.

Results & Discussion

Do people expect teachers to teach semantically coherent concepts? To address this question, we coded people's ratings as positive numbers if they were in the direction of the semantically coherent teacher, and negative numbers if they were not. We then ran separate t-tests comparing the means in the pedagogical and communicative conditions to zero. In the pedagogical condition, people tended to choose teachers of semantically coherent concepts, $mean = 0.97, t(159) = 2.04, p < 0.05$, one-tailed. In the communication condition, people also chose teachers of semantically coherent concepts, $mean = 2.21, t(159) = 6.76, p < 0.001$. The difference between the two conditions was also significant, $t(308) = 2.16, p < 0.05$.

To further investigate the role of semantic coherence, we computed the distance between all of the positive examples in each scenario (see Figure 1). If people expect semantically coherent concepts, then more semantically coherent pairs – those with shorter distances – should have the strongest ratings. We ran a stepwise regression with people's ratings as the dependent variable, and independent variables including distance between the positive examples in the more and less coherent sets, and the difference in the distances. For the pedagogical condition, the distance between positive examples in the coherent concept was the only predictor selected, $r = -0.70, F(1, 14) = 13.24, p < 0.01$. Of the coherent hypotheses, the teachers teaching the more coherent concepts were rated more strongly. For the communication condition, regression analyses showed that distance between positive examples of coherent pairs did not strongly predict people's ratings, $r = 0.23, F(1, 14) = 0.80, p > 0.3$. [3]

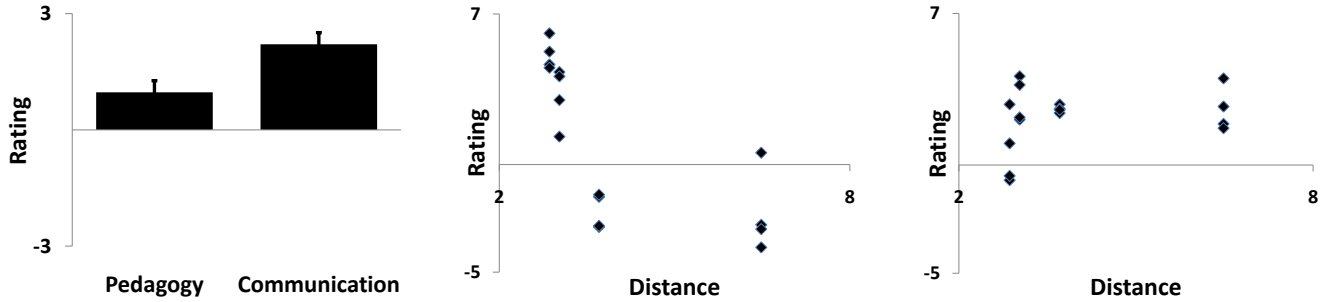
Interestingly, unlike in Experiment 1, people's judgments in Experiment 2 were best predicted by the coherence of the more coherent hypothesis alone (as opposed to the difference in coherence). This suggests that the semantically incoherent hypotheses did not play a large role in people's judgments. This may reflect an explicit judgment that these cases are so unexpected that they, in effect, have zero weight.

The evidence suggests that people expect teachers to teach semantically coherent concepts: overall, people chose teachers of more semantically coherent concepts, and the strength of people's ratings decreased as the strength of coherence decreased. The evidence also suggests that people's expectation of coherence may apply across more than just pedagogical situations. Results from the communication condition showed that people tended to choose the more coherent speaker, but the strength of their ratings was not related to the degree of coherence. These results suggest that people's expectation of semantic coherence may not be limited to pedagogical situations.

Discussion

Pedagogical situations play a central role in human learning. In pedagogical situations, teachers choose which concepts to teach and which examples to use to teach the concept. We have presented an extension of Shafto and Goodman's (2008) model of pedagogical data selection that incorporates specific expectations about pedagogical situations. Using this framework, we have derived a method for isolating the effects of prior expectations about pedagogical situations. The results of Experiment 1 showed that people expect teachers to provide generalizable knowledge, and that this expectation does not apply in more general communicative settings. The results of Experiment 2 showed that people expect teachers to provide semantically coherent information, although this appears not to be specific to pedagogical situations. Taken together, these results provide evidence that people have specific expectations—intuitive

[3] A separate stepwise regression showed that none of the independent variables accounted for significant variance in people's judgments.



(a) (b) (c)

Figure 3: Experiment 2 results: (a) Average human ratings in the Pedagogy and Communication conditions. Positive ratings indicate the more semantically coherent teacher. (b) Scatterplot showing the relationship between distance among positive examples of coherent hypotheses (x axis) and people's ratings (y axis) in the pedagogy condition. People's ratings increase with decreasing distance, $r = -0.70$, suggesting that people expect coherent hypotheses in pedagogical settings. (c) Scatterplot showing the relationship between distance among positive examples of coherent hypotheses (x axis) and people's ratings (y axis) in the communication condition. People's ratings are only weakly related to distance, $r = 0.23$.

theories of pedagogical situations.

Our results provide additional evidence in support of Csibra and Gergely's (2009) claim that people expect generalizable information in pedagogical contexts. Where previous results focused on young children, our results suggest that this expectation continues into adulthood. Our results also provide evidence that semantic coherence, while expected in pedagogical situations, is not specific to these contexts. Rather, the expectation of semantically coherent concepts extends to communicative, as well as pedagogical situations.

Here we have focused on learners' expectations, but for these pedagogical expectations to be reasonable, it is important that teachers meet their expectations. Specifically, do people choose to teach concepts that are more generalizable and more coherent? If so, what are the implications of these matching (or mismatching expectations) in terms of the kinds of concepts that can be learned, the speed at which they are acquired, and the robustness of knowledge transmission? Future research will aim to answer these questions.

Our experiments have provided information about people's prior expectations in pedagogical situations, but it is also important to explain why people have these biases. There is work to be done in formalizing computational models that explain why certain hypotheses would be more or less likely to be taught. This may not turn out to be entirely straightforward because while there is a reasonable motivation for teaching generalizable concepts, there are also motivations for teaching other kinds of concepts. For instance, one might also want to teach sparse concepts because they may be difficult to discover on one's own. Further empirical research may help narrow down the possibilities and provide guidance for more explanatory models.

More generally, previous approaches to modeling human learning have focused on a single unitary set of prior expectations that apply generically across situations (but see Shafto et al., 2006). However, this approach seems obviously too simple. We all intuitively understand that we have different expectations that apply when, for example, we talk to children as opposed to adults. Pedagogical situations are but one case of a more general problem. Understanding how social situations affect learning will require understanding how different contexts affect both learners' prior expectations and learners' assumptions about how information is selected.

Acknowledgments

Thanks to Russell Warner and Carissa Shafto for helpful comments and suggestions during the writing process.

References

Csibra, G. (2007). Teachers in the wild. *Trends in Cognitive Sciences*, 11:95–96.

Csibra, G. and Gergely, G. (2009). Natural pedagogy. *Trends in Cognitive Sciences*, 14:148–153.

Kemp, C. and Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences*, 105:10687–10692.

Luce, R. D. (1959). *Individual choice behavior*. John Wiley, New York.

Shafto, P. and Goodman, N. D. (2008). Teaching games: Statistical sampling assumptions for pedagogical situations. In *Proceedings of the 30th annual conference of the Cognitive Science Society*.

Shafto, P., Kemp, C., Mansinghka, V., Gordon, M., and Tenenbaum, J. B. (2006). Learning cross-cutting systems of categories. In *Proceedings of the 28th annual conference of the Cognitive Science Society*.

Topal, J., Gergely, G., Miklosi, A., Erdohegyi, A., and Csibra, G. (2008). Infants perseverative search errors are induced by pragmatic misinterpretation. *Science*, 321:1831–1834.

A Spiking Neuron Model of Serial-Order Recall

Feng-Xuan Choo (fchoo@uwaterloo.ca)
Chris Eliasmith (celiasmith@uwaterloo.ca)
Center for Theoretical Neuroscience, University of Waterloo
Waterloo, ON, Canada N2L 3G1

Abstract

Vector symbolic architectures (VSAs) have been used to model the human serial-order memory system for decades. Despite their success, however, none of these models have yet been shown to work in a spiking neuron network. In an effort to take the first step, we present a proof-of-concept VSA-based model of serial-order memory implemented in a network of spiking neurons and demonstrate its ability to successfully encode and decode item sequences. This model also provides some insight into the differences between the cognitive processes of memory encoding and subsequent recall, and establish a firm foundation on which more complex VSA-based models of memory can be developed.

Keywords: Serial-order memory; serial-order recall; vector symbolic architectures; holographic reduced representation; population coding; LIF neurons; neural engineering framework

Introduction

The human memory system is able to perform a multitude of tasks, one of which is the ability to remember and recall sequences of serially ordered items. In human serial recall experiments, subjects are presented items at a fixed interval, typically in the range of two items per second up to one item every 4 seconds. After the entire sequence has been presented the subjects are then asked to recall the items presented to them, either in order (serial recall), or in any order the subject desires (free recall). Plotting the recall accuracy of the subjects, experimenters often obtain a graph with a distinctive U-shape. This unique shape arises from what is known as the primacy and recency effects. The primacy effect refers to the increase in recall accuracy the closer the item is to the start of the sequence, and the recency effect refers to the same increase in recall accuracy as the item gets closer to the end of the sequence.

Many models have been proposed to explain this peculiar behaviour in the recall accuracy data. Here we will concentrate on one class of models which employ vector symbolic architectures (VSAs) to perform the serial memory and recall. Using VSAs to perform serial memory tasks would be insufficient however, if the VSA-based model cannot be implemented in spiking neurons, and thus, cannot be used to explain what the brain is actually doing. In this paper, we thus present a proof-of-concept VSA-based model of serial recall implemented using spiking neurons.

Vector Symbolic Architecture

There are four core features of vector symbolic architectures. First, information is represented by randomly chosen vectors that are combined in a symbol-like manner. Second, a superposition operation (here denoted with a $+$) is used to combine vectors such that the result is another vector that is similar to the original input vectors. Third, a binding operation ($\otimes$) is used to combine vectors such that the result is a vector that is dissimilar to original vectors. Last, an approximate inverse operation (denoted with *, such that A^* is the approximate inverse of A) is needed so that previously bound vectors can be unbound.

$$A \otimes B \otimes B^* \approx A \tag{1}$$

Just like addition and multiplication, the VSA operations are associative, commutative, and distributive.

The class of VSA used in this model is the Holographic Reduced Representation (HRR) (Plate, 2003). In this representation, each element of an HRR vector is chosen from a normal distribution with a mean of 0, and a variance of $1/n$ where n is the number of elements there are in the vector. The standard addition operator is used to perform the superposition operation, and the circular convolution operation is used to perform the binding operation. The circular convolution of two vectors can be efficiently computed by utilizing the Fast Fourier Transform (FFT) algorithm:

$$\mathbf{x} \otimes \mathbf{y} = \mathcal{F}^{-1}(\mathcal{F}(\mathbf{x}) \odot \mathcal{F}(\mathbf{y})), \tag{2}$$

where $\mathcal{F}$ and $\mathcal{F}^{-1}$ are the FFT and inverse FFT operations respectively, and $\odot$ is the element-wise multiplication of the two vectors. The circular convolution operation, unlike the standard convolution operation, does not change the dimensionality of the result vector. This makes the HRR extremely suitable for a neural implementation because it means that the dimensionality of the network remains constant regardless of the number of operations performed.

The VSA-based Approach to Serial Memory

There are multiple ways in which VSAs can be used to encode serially ordered items into a memory trace. The CADAM model (Liepa, 1977) provides a simple example of how a sequence of items can be encoded as a single memory trace. In the CADAM model, the sequence containing the items **A**, **B**, and **C** would be encoded as in single memory trace, M_{ABC} as follows:

$$\begin{aligned} M_A &= \mathbf{A} \\ M_{AB} &= \mathbf{A} + \mathbf{A} \otimes \mathbf{B} \\ M_{ABC} &= \mathbf{A} + \mathbf{A} \otimes \mathbf{B} + \mathbf{A} \otimes \mathbf{B} \otimes \mathbf{C} \end{aligned}$$

The model presented in this paper, however, takes inspiration from behavioural data obtained from macaque monkeys. This data suggests that each sequence item is encoded using

ordinal information (Orlov, Yakovlev, Hochstein, & Zohary, 2000), rather than being "chained" together as in the CADAM model. To achieve this, additional vectors are used to represent the ordinal information of each item. In the subsequent equations, this ordinal vector is represented as P_i, where i indicates the item's ordinal number in each sequence. The memory trace M_{ABC} would thus be computed like so:

$$M_A = P_1 \otimes \mathbf{A} \tag{3}$$

$$M_{AB} = P_1 \otimes \mathbf{A} + P_2 \otimes \mathbf{B} \tag{4}$$

$$M_{ABC} = P_1 \otimes \mathbf{A} + P_2 \otimes \mathbf{B} + P_3 \otimes \mathbf{B} \tag{5}$$

The encoding strategy presented above does not seem to have any mechanism by which to explain the primacy or recency effects seen in human behavioural data. In order to achieve these effects, additional components are added to the model. These components are discussed further below.

Neural Representation

To implement any of these models, we need to determine how a vector can be represented by a population of spiking neurons. In 1986, Georgopoulos et al. demonstrated that in the brain, 2D movement directions are encoded by a large population of neurons, with each neuron being most active for one specific direction – their preferred direction. The activity of each neuron would then indicate the similarity of the input vector to each neuron's preferred direction vector. Since the movement direction is essentially a two-dimensional vector, this method of vector representation can be extended to multiple dimensions as well. For a population of neurons, the current J flowing into neuron i can then be calculated by the following equation.

$$J_i(\boldsymbol{x}) = \alpha_i(\tilde{\phi}_i \cdot \boldsymbol{x}_i) + J_i^{bias} \tag{6}$$

In the above equation, the dot product computes the similarity between the input vector $\boldsymbol{x}$ and the neuron's preferred direction vector $\tilde{\phi}$. The neuron gain is denoted by α, while J^{bias} denotes a fixed background input current. The current J_i can then be used as the input to any neuron model $G[\cdot]$ to obtain the activity for neuron i. In this model, we use the leaky integrate-and-fire (LIF) neuron model, characterized as such:

$$a_i(\boldsymbol{x}) = G_i[J_i(\boldsymbol{x})] = \frac{1}{\tau^{ref} - \tau^{RC} \ln\left(1 - \frac{J^{th}}{J_i(\boldsymbol{x})}\right)}, \tag{7}$$

where $a_i(\boldsymbol{x})$ is the average firing rate of the neuron i, τ^{ref} is the neuron refractory time constant, τ^{RC} is the neuron RC time constant, and J^{th} is the neuron threshold firing current. For a time-varying input $\boldsymbol{x}(t)$, the equations remain the same, with the exception that the activity of the neuron is no longer an average firing rate, but rather a spike train:

$$a(\boldsymbol{x}(t)) = \sum_n \delta(t - t_n) \tag{8}$$

Since the spike train represents the neuron's response to the input vector $\boldsymbol{x}$, given the spike trains from all the neurons in the population, it should be possible to derive decoding vectors ϕ that can be used to estimate the original input. Eliasmith and Anderson (2003) demonstrate that these decoding vectors can be found using the following equation.

$$\begin{aligned} &\phi = \Gamma^{-1}\Upsilon, \text{ where} \\ &\Gamma_{ij} = \int a_i(x) a_j(x)\, \mathrm{d}x \qquad \Upsilon_i = \int a_i(x) x\, \mathrm{d}x \end{aligned} \tag{9}$$

By weighting the decoding vectors with the post-synaptic current $h(t)$ generated by each spike, it is then possible to construct $\hat{x}(t)$, an estimate of the input vector. Equation (10) demonstrates how this is achieved. The parameters used to generate the shape of $h(t)$ is determined by the neurophysiology of the neuron population.

$$\begin{aligned} \hat{x}(t) &= \sum_{i,n} \delta(t - t_{in}) * h(t) \phi_i \\ &= \sum_{i,n} h(t - t_{in}) \phi_i \end{aligned} \tag{10}$$

The encoding and decoding vectors also provides a method by which the optimal connection weights between two neural populations can be. If for example, the transformation between two populations of neurons is a simple scaling operation, where the output of the second group of neurons should be $C\boldsymbol{x}$, then the connection weights w between the populations should be

$$w_{ij} = C \alpha_j \tilde{\phi}_j \phi_i \tag{11}$$

Extending Equation (7) for linear operations is also straightforward. Consider three neural populations: one to represent the input x, another to represent the input y, and a third that we wish to have compute the linear combination $Cx + Dy$. The activity of the neurons in final population can be determined by

$$c_k(Cx + Dy) = G_k \left[\sum_i w_{ki} a_i(x) + \sum_j w_{kj} b_j(y) + J_k^{bias} \right], \tag{12}$$

where a_i, b_j, and c_k are the activities of the neurons in the first, second and third neural populations respectively. Employing Equation (11), the synaptic connection weights can also be determined. Letting w_{ki} be the connection weights between the first and third population, and w_{kj} be the connection weights between the second and third population, they work out to be:

$$w_{ki} = \alpha_k \tilde{\phi}_k \phi_i^x \qquad \text{and} \qquad w_{kj} = \alpha_k \tilde{\phi}_k \phi_j^y \tag{13}$$

Note that in the equation above, the superscripts serves to disambiguate the decoders, where ϕ^x signifies the decoders that represent x, and likewise for ϕ^y. Eliasmith and Anderson (2003) go into greater detail on how to use this general framework, known as the Neural Engineering Framework, to derive the appropriate decoders and connection weights to perform arbitrary non-linear operations as well.

The Neural Model

The neural model implemented in this paper is divided into two neural processes. One encodes an item sequence into a single memory trace, and the other decodes an encoded memory trace to retrieve its constituent items.

Sequence Encoder

Analysis of Equations (3) to (5) show that the memory trace for an arbitrary sequence of items can be constructed by computing the convolution of the last item vector with its ordinal vector, and then adding the result of the convolution to the memory trace of the sequence less the final item. From this, a generic sequence encoding equation can be derived (from here on referred to as the *basic encoding equation*).

$$M_i = M_{i-1} + P_i \otimes I_i \tag{14}$$

In the equation above, M_i represents the memory trace after encoding the ith item. P_i and I_i represents the i_{th} item's ordinal vector and item vector respectively.

As mentioned previously, the encoding equation in its basic form does not account for the primacy and recency effects seen in human behavioural data. To achieve the primacy effect, rehearsal is simulated by adding an additional weighted copy of the old memory trace to the memory trace being calculated for the current item. In essence, as each item is rehearsed, a weighted copy of the item is added to the memory trace to "boost" the item's representation within the memory trace. In the equation below, the memory trace of the rehearsal-based encoding is denoted by R_i and the weight applied to the rehearsed contribution of the old memory trace is denoted by α. In the model implemented for this paper, α was set to 0.3.

$$R_i = R_{i-1} + P_i \otimes I_i + \alpha R_{i-1} \tag{15}$$

$$= (1+\alpha)R_{i-1} + P_i \otimes I_i \tag{16}$$

To achieve the recency effect, an separate memory component is added to play the role of a sensory input buffer. The input buffer encodes items in a similar fashion to Equation (14) with a decay added to the old memory trace. This decay causes the input buffer to store only the most recently presented items, thus mimicking the basic recall characteristics of the human working memory system. In the neural implementation of this model, the decay is achieved by tuning the integrators used in the memory modules to slowly drift to zero if no additional input is applied to them. Equation (17) illustrates how this decay can be represented mathematically, with the memory trace of the input buffer represented by B_i, and the rate of decay represented by β.

$$B_i = \beta B_{i-1} + P_i \otimes I_i \tag{17}$$

The final memory trace of the encoded item sequence is then computed by combining the memory trace from the rehearsal component and the memory trace from the input buffer component.

$$M_i = R_i + B_i \tag{18}$$

From the above encoding equations, several issues become evident. First, two operations need to be implemented – a circular convolution and an addition operation. Second, a memory module is needed to hold the value of M_{i-1} while the new memory trace M_i is computed. With these components, and the rehearsal and decay mechanisms described above, a high level block diagram of the complete encoding network can be constructed, as shown in Figure 1.

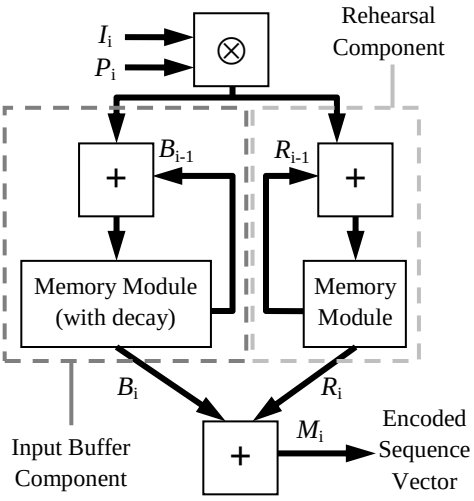


Figure 1: Encoding network functional block diagram.

Sequence Decoder

The decoding process is much simpler than the encoding process. The first step of the decoding process is to convolve the encoded memory trace with the inverse of the desired ordinal vector. For example, if the system is trying to decode the second item in the sequence, the encoded memory trace would be convolved with the inverse of P_2. Next, the result of this convolution is fed to a cleanup memory module. The cleanup memory module contains a copy of all the item vectors in the original sequence, and when given an input, will determine which of the original item vectors best matches the input vector. An example of this decoding process follows. To simplify the example, only the basic encoding equation is used.

$$\begin{aligned} M_{ABC} &= P_1 \otimes \mathbf{A} + P_2 \otimes \mathbf{B} + P_3 \otimes \mathbf{B} \\ C_B &= M_{ABC} \otimes P_2^* \\ &= P_1 \otimes \mathbf{A} \otimes P_2^* + P_2 \otimes \mathbf{B} \otimes P_2^* + P_3 \otimes \mathbf{B} \otimes P_2^* \\ &\approx P_1 \otimes \mathbf{A} \otimes P_2^* + B + P_3 \otimes \mathbf{B} \otimes P_2^* \\ I_B &= cleanup(C_B) \approx B \end{aligned}$$

From the example above, we see that convolving the memory trace M_{ABC} with the inverse of P_2 results in a vector with the desired item vector B combined with the unwanted vectors $(P_1 \otimes \mathbf{A} \otimes P_2^*)$ and $(P_3 \otimes \mathbf{B} \otimes P_2^*)$. However, since the cleanup memory module only contains the item vectors from

the original sequence and not the superfluous vectors, feeding the result of the convolution through the cleanup memory isolates the item vector B, producing the desired result. Figure 2 illustrates the high level block diagram used to implement the decoding network.

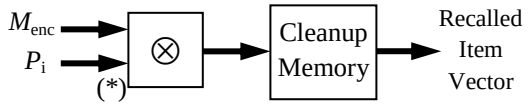


Figure 2: Decoding network functional block diagram.

Performing the Binding Operation Referring back to Equation (2) we see that the binding operation can be calculated using the FFT and IFFT algorithms, so the first step to implementing the binding operation in neurons is to implement these two operations. The equations that compute the FFT and IFFT algorithms are as follows:

$$\begin{aligned} \text{FFT}: &\qquad X_k = \sum_{n=0}^{N-1} x_n e^{-\frac{2\pi i}{N} kn} \qquad k = 0,...,N-1 \\ \text{IFFT}: &\qquad x_n = \frac{1}{N} \sum_{k=0}^{N-1} X_k e^{\frac{2\pi i}{N} kn} \qquad n = 0,...,N-1 \end{aligned} \tag{19}$$

Taking a closer look at the equations above, we see that they can be implemented efficiently as a multiplication between the input vector and a matrix containing the FFT (or IFFT) coefficients. From Equation (11), we can then set the synaptic connection weight matrix as the Fourier transform coefficients to calculate the required FFT and IFFT operations. The one caveat to this approach is that the real and imaginary components of the Fourier transform have to calculated separately and then recombined (with the appropriate sign changes) when the final result is calculated.

With the neural implementation of the Fourier transforms solved, the implementation of the circular convolution binding operation becomes trivial since the only other operation needed is an element-wise multiplication. This can be achieved by utilizing multiple neural populations, each handling one element in the element-wise multiplication.

The Memory Module Since the circular convolution and addition operations are essentially feed-forward neural networks, the memory module in this model needs to be able to drive the network with a constant value and store the new value at the same time. This is achieved by the use of gated integrators. When the integrator is not being gated, it attempts to match the value of the input signal. When the integrator is gated, it no longer responds to the input value, and outputs the previously stored value. By placing two gated integrators in parallel controlled by complementary gating signals, the memory module is able to simultaneously store the new input value while outputting the previously stored value.

Cleanup Memory The cleanup memory network used in this model is an extension of the cleanup memory presented in (Stewart, Tang, & Eliasmith, 2009). In essence, the implementation of cleanup memory involves creating multiple neural populations, each assigned to one item vector from the original item sequence. The preferred direction vectors $\tilde{\phi}$ for each neuron in one population is predefined to match the item vector it is meant to clean up. From Equation (6), we see that the the similarity (dot product) is calculated to determine the activity of the neuron. By predefining $\tilde{\phi}$, we can then determine the similarity of the decoded item vector to each of the original item vectors, thus determining which of the original item vectors best matches the decoded item vector.

Combining the Encoder and Decoder

Getting the spiking neuron model to encode a sequence, and subsequently decode the memory trace is achieved by chaining the encoder and the decoder together. Control signals are used to ensure that the decoding network only commences after the encoder has finished encoding the last item vector. Figure 3 shows the results of the complete network encoding and decoding a example twenty-dimensional 4-itemed sequence.

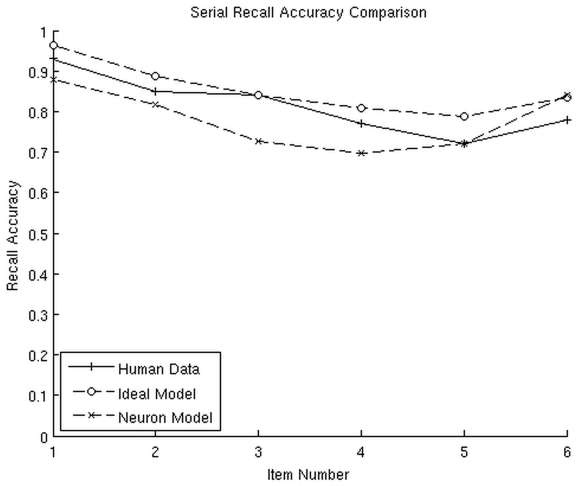


Figure 4: Plot of the recall accuracy data comparing results from human behavioural studies (from Henson et al. (1996), Figure 1), an ideal model implemented in Matlab®, and the spiking neuron model.

Results

The results of the simulation of the spiking neuron implementation of the ordinal serial encoding process is displayed in Figure 4. From the graph it can be seen that both the ideal Matlab®-implemented model and the spiking neuron model are a good match to the human data. The slightly reduced primacy in the neuronal implementation suggests that the simplistic implementation of the rehearsal mechanism can be improved. Figure 5 compares the transposition gradients – which is the count of the recall occurrences of each item for

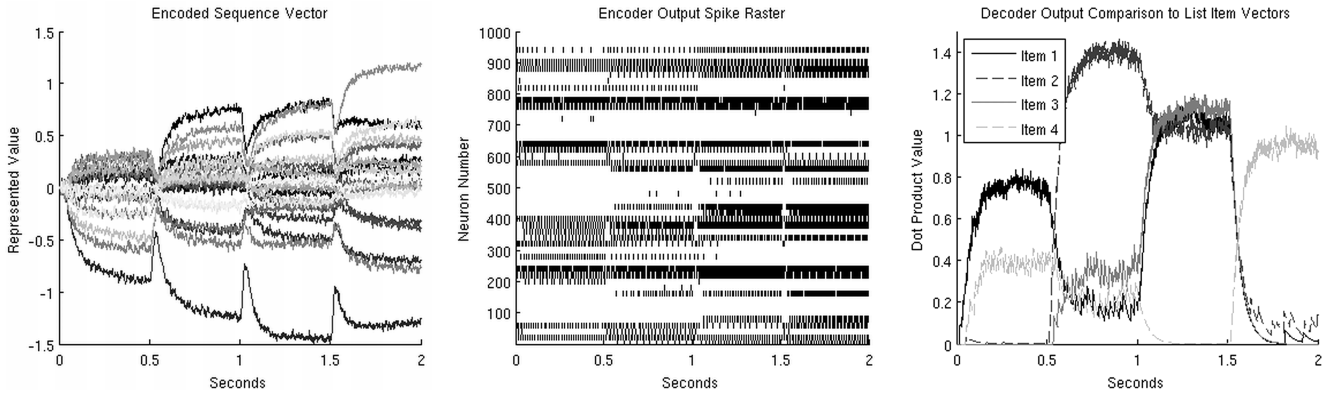


Figure 3: Simulation results from the spiking neuron implementation of the sequence encoder network. A 4-itemed sequence of 20-dimensional item vectors was presented to the network at a half-second interval (two items per second).
(Left) The output of the encoder, M_i, showing the encoded memory trace for each item vector presented. Referring to Equation (18) , the graph at $t = 0.5$ seconds shows $M_1 = R_1 + B_1$, the graph at $t = 1$ second shows $M_2 = R_2 + B_2$, and so forth for. The final encoded memory trace for the entire sequence is the output of the encoder network at $t = 2$ seconds.
(Center) The spike raster plot of the neurons in the output neuron population of the encoder network as it is encoding the sequence in the top figure. The spike raster is displayed for every 20th neuron.
(Right) The similarity plot of each extracted item vector to each one of the four original item vectors. The similarity value between the vectors is obtained using the dot product operation. The graph shows the network correctly identifying the first, second, and last item. The third item is incorrectly identified because the similarity measures of the first three items are too close together for the system to accurately distinguish the correct answer.

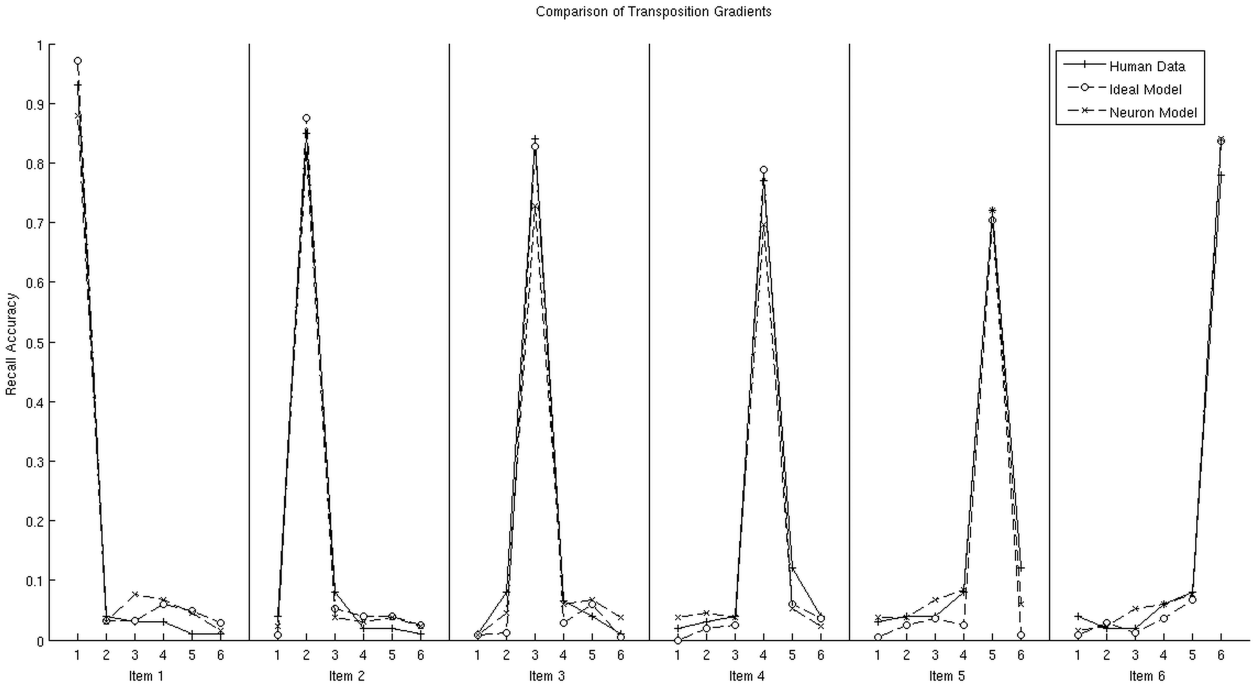


Figure 5: Plot of the transposition gradients comparing results from human behavioural studies (from Henson et al. (1996), Figure 2, for non-confusable items), an ideal model implemented in Matlab®, and the spiking neuron model. Comparing the plots, both the ideal model and the spiking neuron model are able to replicate the transposition curves in the human data.

each serial position – also reveals that both the ideal implementation and the spiking neuron implementation are able to reproduce the transposition effects seen in humans. Both of these simulations were run using six-itemed sequences consisting of fifty-dimensional HRR vectors, and were run for an average of 200 trials each.

Discussion

From the results it can be seen that both the ideal implementation and the spiking neuron model demonstrate the ability to reproduce the primacy, recency, and transposition effects seen in human data. Furthermore, unlike other models which entail a host of tunable parameters to fit the human data, this model only utilizes two tunable parameters: the amount of contribution to the memory trace in the rehearsal component, and the decay rate of the input buffer component.

The model presented here also provides some insight into the neurophysiological requirements of serial memory. It demonstrates the need for a working memory system capable of simultaneous storage and retrieval. This model also maps on very well to Baddeley's model of working memory (Baddeley, 2007), with the input buffer component acting as the phonological loop, and the rehearsal component functioning as the episodic buffer.

Despite their complexity, there are advantages of creating a spiking neuron model in comparison to theoretical models, or models implemented using rate neurons. It provides the ability to compare the spike data of the model to data collected from neural recordings. For example, data collect in Warden and Miller's 2007 paper shows that the neurons change their preferred items as more items are introduced into the system. Although the analysis has yet to be completed at the time this paper was written, it can be inferred that because the encoded sequence vector changes as more items are added, a neuron that is responsive to one configuration of the sequence vector would either be less responsive or not responsive at all when a new item is added – changing the configuration of the encoded sequence vector – as it does in this model.

Several studies (e.g. Chein & Fiez, 2001) have also identified brain areas that are active during serial memory tasks. Moreover, the studies have demonstrated that there are similarities and more importantly, differences, between the areas of activity during the encoding phase and recall phase. By assigning different components of the model to different brain areas (for example, the input buffer component to the temporal lobe, near the auditory cortex, and the rehearsal component to the lateral prefrontal cortex), it would be possible to determine if the pattern of activities recorded in these studies matches the pattern of activities produced by this model.

Future Work

As mentioned in the results section, the performance of the rehearsal component needs to be improved slightly. Possible ways of doing this is by having an active rehearsal mechanism which decodes and then re-encodes the stored memory trace within the inter-item interval (time between each item presentation). Such a rehearsal mechanism will also enable the model to be compared with serial recall studies involving list sizes that exceed the typical human memory span of 4 to 7 items.

Additionally, the current implementation of cleanup memory has a fixed vocabulary of item vectors that is predefined when the network is created. This means that the items in cleanup memory are static and do not change over time. It seems inconceivable that this is what occurs in the brain. Rather, future cleanup memory implementations should be dynamic, with the ability to "load" and "unload" arbitrary item vectors into its vocabulary.

References

Baddeley, A. (2007). *Working memory, thought, and action*. Oxford: Oxford University Press.

Eliasmith, C., & Anderson, C. H. (2003). *Neural engineering: computation, representation, and dynamics in neurobiological systems*. Cambridge, MA: The MIT Press.

Georgopoulos, A. P., Schwartz, A., & Kettner, R. E. (1986). Neuronal population coding of movement direction. *Science*, *233*, 1416–1419.

Henson, R. N., Norris, D. G., Page, M. P., & Baddeley, A. D. (1996). Unchained memory: error patterns rule out chaining models of immediate serial recall. *The Quarterly Journal of Experimental Psychology*, *49A*(1), 80–115.

Liepa, P. (1977). *Models of content addressable distributed associative memory.* (Unpublished manuscript)

Orlov, T., Yakovlev, V., Hochstein, S., & Zohary, E. (2000, March). Macaque monkeys categorize images by their ordinal number. *Nature*, *404*.

Plate, T. A. (2003). *Holographic reduced representations: distributed representations for cognitive structures*. Stanford, CA: CSLI Publications.

Stewart, T. C., Tang, Y., & Eliasmith, C. (2009). A biologically realistic cleanup memory: autoassociation in spiking neurons. *9th International Conference on Cognitive Modelling*.

Warden, M. R., & Miller, E. K. (2007). The representation of multiple objects in prefrontal neuronal delay activity. *Cerebral Cortex*, *17*, 141–150.

Nonverbal Semantic Processing Disrupts Visual Word Recognition in Healthy Adults

Lang Chen (lchen32@wisc.edu)
Department of Psychology, 1202 W. Johnson Street
Madison, WI 53706 USA

Timothy T. Rogers (ttrogers@wisc.edu)
Department of Psychology, 1202 W. Johnson Street
Madison, WI 53706 USA

Abstract

Two experiments examined the effect of semantic interference on visual lexical decision (vLD) in normal skilled readers. Experiment 1 employed a dual-task paradigm to test whether nonverbal semantic processing disrupts visual word recognition when the orthographic structure of words and non-words is controlled. Experiment 2 employed the same paradigm to test whether participants strategically shifted their reliance onto orthographic information when orthographic structure provided a cue to lexicality. The results showed (1) significant semantic interference in the vLD task in normal skilled readers when words and non-words were matched for orthographic well-formedness and (2) no semantic interference when words and non-words differed reliably in their orthographic well-formedness. The results are consistent with the view that accurate lexical decisions depend upon semantic activation, especially when judgments cannot be made on the basis of orthographic structure alone.

Keywords: semantics; lexical decision; dual-task; dual-route models.

Introduction

What is the relationship between semantic and lexical knowledge in the mind and brain? Neuropsychological investigations of this question have led to two contradictory conclusions. One long-standing tradition has emphasized neuropsychological dissociations to support the argument that knowledge of word forms and meanings are supported by functionally independent cognitive systems. For instance, patient EM performed poorly on semantic tasks such as picture naming but perfectly when reading or recognizing even irregular, low-frequency, and orthographically strange words (Blazely, Coltheart, & Casey, 2005; for similar cases, also see Cipolotti & Warrington, 1995; Schwarz, Saffran, & Marin, 1980). For some theorists, such evidence suggests that successful performance in lexical tasks like reading aloud or recognizing words does not depend on intact input from the word-meaning system (Coltheart, 2004).

A different tradition has emphasized that such classical dissociations are observed in only a tiny fraction of patients with semantic impairment, and that, in the vast majority of cases, lexical and semantic impairments go hand-in-hand (Woollams, Ralph, Plaut, & Patterson, 2007). For instance, Patterson et al. (2006) examined performance on four lexical tasks—including reading aloud, lexical decision, spelling, and past-tense inflection—in fourteen patients with semantic dementia (SD), a progressive degenerative syndrome that produces a remarkably pure semantic impairment. Results revealed that, in all four tasks, all fourteen patients were seriously impaired at processing low-frequency items with atypical phonological, orthographic, or syntactic structure. Similarly Woollams and colleagues (2007) reported reading performance in a cohort of 51 patients with semantic impairment and found that only a vanishingly small proportion—3 out of 51—showed spared performance comparable to EM's (and see Graham, Patterson, & Hodges, 2000; Patterson & Hodges, 1992; Patterson, Lambon Ralph, Hodges, & McClelland, 2001; for similar accounts of association between semantic and lexical impairment). For these theorists, the strong association between semantic and lexical impairment suggests that, in most individuals, performance on lexical tasks depends importantly on intact input from the semantic system (Plaut, McClelland, Seidenberg, & Patterson, 1996; Seidenberg & McClelland, 1989).

Differentiating these views on the basis of neuropsychological evidence has proven challenging because both views can account for the major findings, that is, the strong association of lexical and semantic impairment in the majority of reported cases and the occasional dissociation in a small minority. For those who believe semantic and lexical processes are functionally independent, the strong association arises because the disease process in these individuals has affected both systems. Patterson et al. (2006) refer to this as the "Associated but unrelated deficits" (ABUD) view. Under ABUD, only dissociations provide useful information about the functional architecture of the language system, because they straight-forwardly disprove causal necessity: reading, word recognition, spelling, etc, cannot of necessity depend upon intact semantic input, because it is possible for these abilities to be completely spared in the face of degraded semantic knowledge.

The alternative view—that lexical processes depend importantly upon semantic input—was dubbed "It's All Semantics" (IAS) by Patterson et al. (2006). For proponents of IAS, the few cases that show strong lexical-semantic

dissociations are the exceptions that prove the rule. Such cases may deviate somewhat from the more typical pattern of associated deficits because they are exceptional in other ways. For instance, they may have had unusually good lexical skills in their premorbid state, so that, with mild semantic impairment, they remain capable of performing within the established norms for their age group, even if they have declined significantly from their premorbid peak. From this point of view, the fact that EM was a secretary for much of her life is potentially important—she presumably took dictation and as a result may have developed unusually robust orthographic and phonological representations.

Further complicating the picture is the fact noted by Plaut (1997) and others that some patterns of apparent dissociation in the literature may be attributable to poorly controlled stimulus materials. It is now well established that, when semantic knowledge degrades, patients can retain good knowledge of the "surface" structure of different domains. For instance, even when unable to retrieve the meanings of words, patients with semantic impairments can retain knowledge about orthographic structure, that is, which letter sequences are common and which unusual in the language. In tests of word-recognition, such patients can appear completely normal if the target words are all orthographically well-formed and the distractor words are all orthographically strange (Rogers, Ralph, Hodges, & Patterson, 2004). The same patients show serious impairments, however, if the orthographic structure of words and non-words is matched—indeed, some patients judge well-formed non-words to be real words at rates exceeding chance, showing a strong over-reliance on orthographic structure in making their decisions.

Taken together, the evidence from neuropsychological studies is arguably compatible with both ABUD and IAS and it is not clear that further neuropsychological evidence can adjudicate the different positions. Because the status of semantic knowledge cannot be manipulated experimentally in such studies the causal links between semantic and lexical processing are difficult to establish.

Experiment 1 of the present study tests the hypothesis that semantic processing contributes to one kind of lexical process—word recognition—using a dual-task paradigm. Healthy participants performed a visual lexical decision task while simultaneously performing a secondary nonverbal task (sound judgment) that either did or did not tap semantic memory. The key question is whether word-recognition is significantly more disrupted by the semantic than the non-semantic secondary task. According to ABUD, word recognition does not depend upon input from semantics, so there should be no effect of secondary task type as long as the two tasks are equally demanding. According to IAS, word recognition does depend upon semantics, so word recognition should be worse when participants simultaneously perform the nonverbal semantic task. Experiment 2 uses the same methods to test the hypothesis that people show less or even no reliance on input from semantics when lexicality is confounded with orthographic structure—that is, when words and non-words differ reliably in their orthographic well-formedness.

Experiment 1

Method

Participants Fifty-one undergraduate students from UW-Madison participated in Experiment 1 for course credit or monetary compensation. All were native English speakers with normal or corrected-to-normal vision.

Materials and Design Participants were asked to perform two tasks simultaneously: a visual lexical-decision (vLD) task and a sound judgment task. The experimental manipulation concerned whether the sound judgment task did or did not draw upon semantic knowledge. In the non-semantic "Tones" condition, participants listened to a complex tone and judged whether it was ascending in pitch or not. The task is non-semantic because it does not require the participant to consult or draw upon stored knowledge about the sound. In the semantic "Birds" condition, participants listened to an animal sound and judged whether it was produced by a bird or not—hence this task required participants to draw on stored knowledge about the sounds produced by birds and animals.

The stimuli for the vLD task were adapted from a previous study (Hauk, et al., 2006) and consisted of 50 orthographically typical words (TW; e.g., "rot"), 50 orthographically strange words (SW; "yacht"), 50 orthographically typical non-words (TNW; "yot") and 50 orthographically strange non-words (SNW; "racht"). Words and non-words were matched for the goodness of their orthographic structure as measured by summed bigram and trigram frequencies (for details, see Rogers, et al., 2004). This manipulation ensured that participants could not rely on the well-formedness of the letter string to decide whether the item was a word (Blazely, et al., 2005; Plaut, 1997). In all word items, only 11% of them referred to animal names. Since little is known about the semantic interference with non-word stimuli, we will examine the effect of sound-judgment tasks on word and non-word stimuli separately.

The sound judgment task included 50 items in each condition. The tones were complex sounds similar to a dial tone, half ascending in pitch and half descending, and varying in initial pitch and rate of change. The animal sounds included the vocalizations of 25 different birds and 25 non-bird animals. Items from the two conditions were matched on total duration. A pilot study with 28 participants who did not engage in Experiment 1 showed that the two tasks did not differ significantly by items or subjects in

mean accuracy and response time (all *ps* > 0.10). Thus the two sound-judgment tasks were closely matched for overall difficulty.

Procedure The 51 participants were randomly assigned to either condition, resulting in 25 in Tones and 26 in Birds. Every participant was tested individually and began with three short practice sessions. First, participants practiced the vLD task: on each trial they viewed a letter string on the computer monitor and pressed a button with their dominant hand to indicate whether it was a word or not. Next, they practiced the sound-judgment task alone: participants listened to a series of sounds presented over headphones and orally reported their response by saying "Yes" (for ascending tones in the Tones condition or for birds in the Birds condition) or "No" (for descending tones / non-birds). The oral responses were recorded by the experimenter. If any lexical processing was involved in the oral response, it should be equivalent across two conditions. In the third practice phase, participants performed both tasks simultaneously with a small number of stimuli. In this practice phase and in the experiment proper, the onsets of stimuli in vLD and sound tasks were asynchronous so participants could not get into a "rhythm" of doing one task then the other. After participants were familiarized with the dual-task procedure, they continued to the experiment proper, performing both tasks simultaneously until they had responded to all 200 items in the vLD task (presented in random order). In the sound task, sounds were selected randomly with replacement until participants had finished the vLD task. The study took about 40 minutes.

Results

The mean accuracy in the sound judgment tasks was generally high and did not differ significantly between groups: 0.90 (SD = 0.07) for Tones and 0.93 (SD = 0.03) for Birds, $F(1,49) = 2.329$, MSE = 0.003, $p = 0.133$.

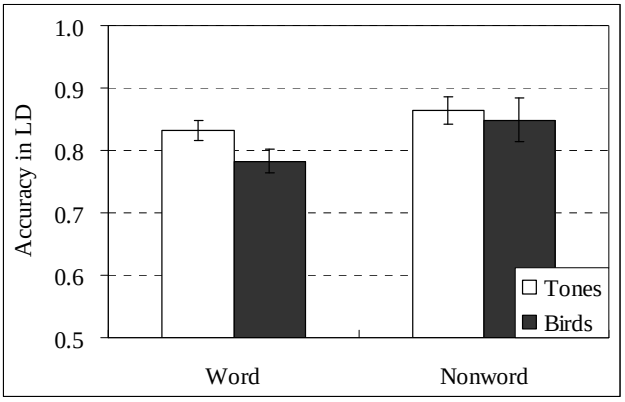


Fig. 1: Mean accuracy of the vLD task in Experiment 1.

Figure 1 shows mean accuracy and standard errors for words and non-words in each condition. A one-way ANOVA revealed that, for word items, accuracy was significantly lower in the Birds than in the Tones condition both by subject and by item (Tones, mean = 0.83, SD = 0.08; Birds, mean = 0.77, SD = 0.10), $F_1(1,49) = 5.410$, MSE = 0.008, $p = 0.024$, $F_2(1,99) = 50.996$, MSE = 0.003, $p < .001$) with no difference in response time (Tones, mean = 1079.36, SD = 358.67; Birds, mean = 1066.43, SD = 439.40, all *ps* > 0.10). For non-words neither accuracy (Tones, mean = 0.86, SD = 0.11; Birds, mean = 0.85, SD = 0.18) nor RT (Tones, mean = 1112.12, SD = 317.14; Birds, mean = 1108.26, SD = 451.54) differed reliably between conditions, all *ps* > .05. Thus, the participants made more errors recognizing words, but not rejecting non-words, when their semantic system was occupied with a secondary nonverbal categorization task compared to an equally-demanding but non-semantic task.

To further test the hypothesis that semantic processing interferes with vLD, we investigated the correlation in overall accuracy between the vLD and the sound judgment task across subjects in each group. If the two tasks do not share a critical resource, we expect a strong positive correlation in accuracy: participants who generally cope well with dual-task situations will perform well on both, whereas those who generally cope poorly with dual tasks will perform poorly on both. If, however, the two tasks share an important resource, this relationship should be altered: allocation of the resource to one task should boost performance in one task but should hinder performance of the other task, attenuating or eliminating the expected positive correlation between the two tasks.

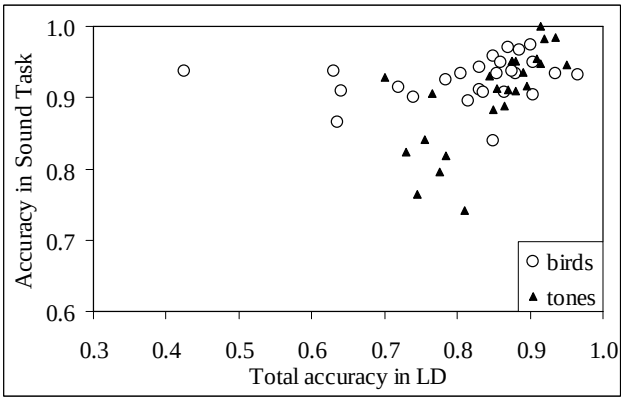


Fig. 2: Correlation between mean accuracy in the vLD task and sound judgment tasks in Experiment 1.

Figure 2 plots the mean accuracy in vLD and the sound-judgment task for the two groups. Performance on vLD and the Tones task was positively correlated (r = 0.700, $p < .001$), while this relationship in the Birds condition was not reliable (r = 0.201, $p = .325$) and was significantly lower than that in Tones condition, $Z = 2.225$, $p = 0.026$. Thus some participants traded off accuracy on vLD for an acceptable level of accuracy on the semantic but not the

non-semantic sound judgment task.

Experiment 2

Experiment 1 found that healthy participants showed worse performance on the vLD task when their semantic knowledge was engaged in a concurrent task. Experiment 2 assessed whether this semantic interference is attenuated when orthographic structure provides a valid cue to lexicality. We hypothesized that, if words and non-words differed reliably in their orthographic well-formedness, participants could rely on this surface cue to guide their decisions, so that reliance on the semantic system would be reduced or eliminated.

Method

Participants Sixty undergraduate students who did not participate in Experiment 1 participated in return for course credit.

Materials and Designs We used identical materials but with two important differences in design. First, stimuli were grouped into two sets in such a way that, within each set, words and non-words differed systematically in their orthographic structure. Thus Set 1 (TW-SNW) included typical words (e.g., rot) and strange non-words (e.g., racht); while Set 2 (SW-TNW) included strange words (e.g., yacht) and typical non-words (yot). Participants completed either Set 1 or Set 2. Second, to maximize our power to detect an influence of semantic interference on word recognition, the secondary task condition (Tones vs. Birds) was manipulated within every subject. Each set was divided into two subsets closely matched for accuracy and response time (all the *p*s > 0.05) in a pilot study with 23 participants who did not participate in Experiment 2. Participants in each group then completed one subset paired with the Tones task and the other subset paired with the Birds task. The order of subsets and their combinations with Tones or Birds condition were counterbalanced across participants.

Procedure Participants were randomly assigned to one of the set conditions resulting in 30 participants in each. The dual-task procedure was identical to that in Experiment 1 except that the participants were exposed to both Tones and Birds conditions in a block design.

Results

Set 1 (TW-SNW) Unexpectedly, the mean accuracy in the sound judgment tasks differed reliably for this group (0.86, SD = 0.10 for Tones and 0.91, SD = 0.04 for Birds), $F(1,29) = 6.447$, MSE = 0.006, $p = 0.017$. Some participants performed especially badly in the Tones task, as implied by the larger SD in this condition. We will return to this issue later in this section.

Neither mean accuracy nor RT in the vLD task differed significantly in the Tones versus Birds conditions—F values ranged between 0.005 and 2.04, all *p*s > 0.16 for all comparisons except response time to reject non-words for tones versus birds. In this contrast there was a trend toward an effect, but with somewhat *faster* response times in the Birds than the Tones condition (Tones, mean = 1026.45, SD = 340.14; Birds, mean = 956.34, SD = 364.01), $F_1(1,29) = 3.424$, MSE = 21558.907, $p = 0.073$, $F_2(1,49) = 2.400$, MSE = 61468.119, $p = 0.128$). Thus there is no evidence that performance of the nonverbal semantic task disrupted word recognition in this condition.

Could this difference from Experiment 1 somehow be attributable to the participants who performed poorly at Tone judgment? To address this question we identified 8 participants with accuracy lower than 0.80 in the Tones task and excluded them from all analyses to see whether the results would differ. With these participants excluded, mean accuracy in Tones condition was 0.91 (SD = 0.07) which was not significantly difference from the Birds condition (mean = 0.91, SD = 0.04), $F(1,21) = 0.242$, MSE = 0.004, $p = 0.628$). In the remaining 22 participants we still observed no reliable effect of sound-judgment task on either accuracy or response time in the vLD task (all the *p*s > 0.05). Thus when words are well-formed and non-words are ill-formed, there is no evidence that participants rely on semantic processing to make lexical decisions.

Set 2 (SW-TNW) For participants who completed Set 2, where words were orthographically ill-formed and non-words were orthographically typical, there was no significant difference in the sound judgment accuracy for Tones versus Birds (Mean accuracy = 0.91, SD = 0.08 for Tones and 0.93, SD = 0.04 for Birds, $F(1,29) = 0.781$, MSE = 0.003, $p = 0.384$).

Just as in Set 1, the mean accuracy and response time for the vLD task did not differ significantly in the Tones versus Birds conditions—all F ratios were between 0.001 and 1.17, all *p*s > 0.28. Thus even when words were orthographically strange and non-words were regular, participants showed no evidence of worse performance when simultaneously performing a semantic relative to a non-semantic task. Experiment 2 thus suggests that, when orthographic structure can serve as a reliable cue to lexicality, participants do not substantially rely upon semantic processing to recognize words.

Discussion

In a dual-task interference paradigm we found that nonverbal semantic processing disrupted word recognition in healthy adults (Experiment 1), especially when orthographic structure did not provide a useful cue to

lexicality (Experiment 2). These results are consistent with the view that word recognition depends upon semantic processing (Patterson, et al., 2006; Rogers, et al., 2004; Woollams, et al., 2007), and they also suggest, in accordance with other work (Plaut, 1997), that such effects can be attenuated in tasks that confound lexicality with orthographic structure.

Our results complement patient studies documenting a strong association between impaired semantic knowledge and disturbed performance on lexical tasks including word recognition (Patterson, et al., 2006; Rogers, et al., 2004; Woollams, et al., 2007). A natural interpretation of this patient work has been that semantic, orthographic and phonological representations of words are all represented within the same interactive system (Dilkina, McClelland, & Plaut, 2008; Plaut, et al., 1996) so that, when semantic representations degrade, so too does the stability of unusual phonological and orthographic forms. This hypothesis has proven difficult to test through patient studies alone, however, because it has been difficult to rule out the alternative hypothesis that lexical and semantic impairments occur as a consequence of a disease process that jointly affects two independent systems. The current study provides a stronger test of the hypothesis because there is no disease process—instead, the contribution of semantic processing to word recognition was functionally disrupted by engaging the semantic system in a secondary task. Moreover, the secondary task was a nonverbal sound-recognition judgment that arguably makes no demands upon lexical processes. Nevertheless, it led to poorer word-recognition when performed simultaneously with vLD.

Our results challenge the view that there exists "an orthographic lexicon that is distinct from the semantic system" (pp1163, Coltheart, 2004). On this view, normal participants with intact orthographic lexicons should show equivalent performance in dual-task conditions, regardless of nature of the secondary task, because accurate word-recognition can be accomplished solely by consultation of the orthographic lexicon.

Others have previously argued that the orthographic structure of targets and distracters might influence the extent to which accurate lexical decisions depend upon semantic processing (Plaut, 1997; Seidenberg & McClelland, 1989), and this hypothesis was corroborated in Experiment 2: using the same materials and procedure as Experiment 1, the semantic interference effect was eliminated simply by blocking stimuli so that orthographic well-formedness provided a reliable cue to lexicality. If participants could perform accurately simply by accepting (for Set 1) or rejecting (for Set 2) all well-formed letter strings, then they relied less or not at all on semantic input.

It is worth noting that this latter result also poses a puzzle for the view that there exists an orthographic lexicon that is independent of semantics. If lexical decisions are "…done at the level of the orthographic lexicon" (pp701, Blazely, et al., 2005), it is not clear why one should observe different patterns of behavior for the exact same set of target words, depending upon how they are blocked with non-word distractors. Besides, the results from Experiment 2 eliminated the possibility that the semantic interference observed in Experiment 1 was due to difference in the extent of covert word reading across conditions. If so, some might expect to observe poorer performance on vLD in the Birds condition as well, since the same paradigm and sound stimuli were used in Experiment 2. However, this prediction is not supported by the result, suggesting that the covert articulation, if any, cannot be the alternative explanation for the observed semantic inference in Experiment 1.

The present study leaves at least one important question unanswered: How does one account for individual cases who, despite serious semantic impairment, can perform within the normal range on tests of word recognition or other lexical tasks? Recent computational modeling work has emphasized that individual differences in linguistic experience can influence the performance of lexical tasks and might account for the occasional lexical/semantic dissociations observed in neuropsychological studies (Dilkina, et al., 2008). For instance, Zevin and Seidenberg (2006) showed that variability in the model training regime can produce individual differences in non-word reading patterns similar to those observed in skilled readers. Dilkina et al. (2008) also demonstrated how differences in the frequency with which a model encounters orthographic versus visual inputs can produce dissociations between word reading and object naming in an interactive model of the lexico-semantic system.

In addition to such differences in experience, our results suggest that individuals may differ in other important respects. In Experiment 1, we found that, whereas some individuals coped well with the dual task scenario—performing near ceiling on both tasks—others struggled considerably and, in the "semantic interference" condition, appeared to trade off the accuracy of one task for another. Previous work (Herdman & LeFevre, 1992) has shown that a dual-task paradigm increases resource demands and affects different aspects of word recognition process, such as speed and efficiency. Presumably, participants with superior cognitive control are better able to manage the resource demands for both tasks and so may show little semantic interference. Understanding how individual differences in linguistic experience and in cognitive control may contribute to differential reliance on the semantic system in the performance of lexical tasks remains a goal for future research.

In conclusion, the present study demonstrates that normal participants' performance on a visual lexical decision task is disrupted by a simultaneous sound judgment task that taxes semantic memory, suggesting that lexical processes draw upon semantic processes. Moreover, the semantic interference was affected by the orthographic structure of

the words and non-words, suggesting that reliance on semantic versus orthographic information in lexical decision is dynamic.

Acknowledgements

This research was funded by a Vilas Fellowship awarded by the University of Wisconsin-Madison to the second author.

References

Blazely, A. M., Coltheart, M., & Casey, B. J. (2005). Semantic impairment with and without surface dyslexia: Implications for models of reading. *Cognitive Neuropsychology, 22*(6), 695 - 717.

Cipolotti, L., & Warrington, E. K. (1995). Semantic memory and reading abilities: A case report. *Journal of the International Neuropsychological Society, 1*, 104-110.

Coltheart, M. (2004). Are there lexicons? *Quarterly Journal of Experimental Psychology: Section A, 57*, 1153-1171.

Dilkina, K., McClelland, J. L., & Plaut, D. C. (2008). A single-system account of semantic and lexical deficits in five semantic dementia patients. *Cognitive Neuropsychology, 25*(2), 136-164.

Graham, N. L., Patterson, K., & Hodges, J. R. (2000). The impact of semantic memory impairment on spelling: Evidence from semantic dementia. *Neuropsychologia, 38*(2), 143-163.

Hauk, O., Patterson, K., Woollams, A., Watling, L., Pulvermüller, F., & Rogers, T. T. (2006). [Q:] When would you prefer a SOSSAGE to a SAUSAGE? [A:] At about 100 msec. ERP correlates of orthographic typicality and lexicality in written word recognition. *Journal of Cognitive Neuroscience, 18*, 818-832.

Herdman, C. M., & LeFevre, J.-A. (1992). Individual differences in the efficiency of word recognition. *Journal of Educational Psychology, 84*(1), 95-102.

Patterson, K., & Hodges, J. R. (1992). Deterioration of word meaning: Implications for reading. *Neuropsychologia, 30*(12), 1025-1040.

Patterson, K., Lambon Ralph, M. A., Hodges, J. R., & McClelland, J. L. (2001). Deficits in irregular past-tense verb morphology associated with degraded semantic knowledge. *Neuropsychologia, 39*(7), 709-724.

Patterson, K., Ralph, M. A. L., Jefferies, E., Woollams, A., Jones, R., Hodges, J., et al. (2006). "Presemantic" cognition in semantic dementia: Six deficits in search of an explanation. *Journal of Cognitive Neuroscience, 18*(2), 169-183.

Plaut, D. C. (1997). Structure and function in the lexical system: Insights from distributed models of word reading and lexical decision. *Language & Cognitive Processes, 12*, 765-806.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review, 103*(1), 56-115.

Rogers, T. T., Ralph, M. A. L., Hodges, J. R., & Patterson, K. (2004). Natural selection: The impact of semantic impairment on lexical and object decision. *Cognitive Neuropsychology, 21*(2-4), 331-352.

Schwarz, M. F., Saffran, E. M., & Marin, O. S. M. (1980). Fractionating the reading process in dementia: Evidence for word-specific print-to-sound associations. In M. Coltheart, K. Patterson & J. C. Marshall (Eds.), *Deep dyslexia*. London: Routledge and Kegan Paul.

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review, 96*(4), 523-568.

Woollams, A. M., Ralph, M. A. L., Plaut, D. C., & Patterson, K. (2007). SD-squared: On the association between semantic dementia and surface dyslexia. *Psychological Review, 114*(2), 316-339.

Zevin, J. D., & Seidenberg, M. S. (2006). Simulating consistency effects and individual differences in nonword naming: A comparison of current models. *Journal of Memory and Language, 54*(2), 145-160.

Effects of Anticipatory Coarticulation on Lexical Access

Stephen J. Tobin (stephen.tobin@uconn.edu)
Department of Psychology, University of Connecticut, Storrs, CT 06269-1020 USA
Haskins Laboratories, 300 George St., New Haven, CT 06510 USA

Pyeong Whan Cho (pyeong.cho@uconn.edu)
Department of Psychology, University of Connecticut, Storrs, CT 06269-1020 USA
Haskins Laboratories, 300 George St., New Haven, CT 06510 USA

Patrick M. Jennett (pjennett@engr.uconn.edu)
Cognitive Science Program, University of Connecticut, Storrs, CT 06269-2054 USA

James S. Magnuson (james.magnuson@uconn.edu)
Department of Psychology, University of Connecticut, Storrs, CT 06269-1020 USA
Haskins Laboratories, 300 George St., New Haven, CT 06510 USA

Abstract

One of the most challenging unsolved problems in cognitive science is lack of invariance in spoken language. We take the view that variability due to coarticulation is systematic and beneficial. Several recent eye tracking experiments have demonstrated listeners' sensitivity to local coarticulatory cues between adjacent phonemes. We examined sensitivity to longer-range, anticipatory vowel-to-vowel coarticulation, which can spread across multiple syllables. Using a variant of the Visual World eye tracking paradigm (Tanenhaus et al., 1995), we conducted the first on-line test of whether lexical access is sensitive to such subtle, long-range cues, and whether the impact of such cues is modulated by the coarticulation resistance of intervening segments. Lexical access was delayed when misleading anticipatory coarticulation was available in cross-spliced materials. This significantly extends the nature and temporal range of subcategorical cues known to influence on-line sentence comprehension, and demonstrates that lexical access is simultaneously constrained by information at multiple temporal grains.

Keywords: Coarticulation; anticipation; garden path; eye tracking.

Introduction

One of the hardest unsolved problems in cognitive science is *lack of invariance* in speech. There is a many-to-many mapping between acoustics and percepts, such that the same acoustic information can map to different speech sounds, while different acoustic information can map to the same speech sounds (depending on phonetic context, speaking rate, physical or indexical characteristics of talkers, etc.). This is true of production and perception even for clearly articulated segments and syllables (Ladefoged & Broadbent, 1957; Liberman, Delattre & Cooper, 1952; Peterson & Barney, 1952). The problem is compounded in mapping to words and beyond in conversational speech, where even more variation occurs. For example, Hawkins (2003) describes radical changes in the acoustics of the message "I do not know" in a progression from careful speech to casual speech ("I dunno", and even more reduced forms). The puzzle, then, is how we reliably map acoustics to words despite (or perhaps with the aid of) all this variation.

In order to make progress in studying spoken word recognition, psycholinguists have made the temporary simplifying assumption that the input to word recognition can be approximated by a phonemic transcription (as though this were the product of a speech perception mechanism). This allows one to sidestep the lack of invariance problem and related complications due to *coarticulation.* Coarticulation refers to the fact that the articulatory gestures of adjacent and even nonadjacent segments overlap, and therefore, so do their acoustic realizations. That is, as you produce one speech sound, you are simultaneously preparing your articulators for upcoming segments, and still experiencing effects of preceding articulations. Coarticulation is often viewed as destructive, as in Hockett's (1955) metaphor of a wringer squishing together a line of easter eggs on a conveyor belt, and a major contributor to lack of invariance.

Even when a scientist is cognizant of the fact that the phonemic input assumption is almost certainly incorrect, and she explicitly considers it provisional (until we solve the lack of invariance problem at the phonological level), it has the potential to hide constraints on word recognition (Magnuson, 2008). For example, Salverda, Dahan and McQueen (2003) reported that listeners use subtle prosodic cues (e.g., vowel duration) to anticipate word length and constrain lexical competition. They tracked eye movements as subjects followed spoken instructions to click on pictures on a computer display. When initial vowel duration was consistent with a bisyllabic word, subjects immediately began looking preferentially at items with bisyllabic names.

A phonemic transcription also abstracts away from coarticulatory information, which can specify qualities of upcoming segments. Dahan, Magnuson, Tanenhaus, and Hogan (2001) demonstrated that listeners are extremely sensitive to such information. They cross-spliced words (e.g., *neck* and *net*) to provide misleading coarticualtory cues to final consonants. Using the visual world paradigm, they found fast, robust effects of such mismatches on lexical activation and competition.

These examples are inconsistent with suggestions that coarticulation has a destructive impact on phonetic

information. Instead, coarticulation systematically provides anticipatory and redundant cues that afford rapid information transmission in speech. This optimistic view, that coarticulation is lawful and informative (Elman & McClelland, 1986; Fowler, 1986), is consistent with evidence that listeners compensate for coarticulation, taking into account the predictable structure of an ongoing speech event at the gestural (Fowler, 1980; Browman & Goldstein, 1992), phonological (Gow, 2001), lexical (Ganong, 1980; Magnuson et al., 2003), and sentential (Gaskell & Marslen-Wilson, 2001) levels.

All of these examples involve cases of local coarticulation, that is, coarticulation between adjacent segments. Subsequent work on speech production has revealed the existence of long-range coarticulation, in some cases spanning multiple segments or even syllables (Heid & Hawkins, 2000; Recasens, 1984; West, 2000). This raises the possibility that listeners have even richer information at their disposal -- cues specifying qualities of upcoming sounds even several segments in advance. However, these effects are subtle and subject to strong constraints. Among these constraints is *coarticulation resistance*, a finding of Bladon and Al-Bamerni (1976), who observed that intervening consonants could modulate the effects of vowel-to-vowel coarticulation. Specifically, light, palatal [l]s reduced coarticulation between surrounding vowels in comparison to dark, velarized [l]s. The articulatory and perceptual effects of coarticulation resistance has been investigated in some detail (Fowler, 2005; Fowler & Brancazio, 2000; Recasens & Espinosa, 2009). Consonants with high coarticulation resistance prevent coarticulation between surrounding vowels. Typically, high coarticulation resistance consonants involve tongue body or tongue tip articulations (e.g., [t], [d]) and/or fine motor control (e.g., [s], [z]). That is, strong constraints on tongue tip or tongue body damp long-range, vocalic coarticulation. Low coarticulation resistance consonants allow coarticulatory information to spread further, as they do not involve the tongue body or tip and do not require particularly fine motor control (e.g., [p], [b], [f], [v]).

The current study is the first on-line study, to the best of our knowledge, to examine whether lexical access is sensitive to long-range coarticulatory information, and whether the impact of such information is modulated by the coarticulation resistance of intervening segments. If so, this will represent a substantial increase in the amount of detail listeners are known to use in order to constrain speech perception and word recognition.

Experiment

We examined whether coarticulatory effects would influence lexical access by manipulating two factors. The first was (*Coarticulatory*) *Match*. At the Match level, two instances of one utterance (e.g., "pick up a pole") were cross-spliced after the word "a". In the Mismatch condition, two utterances with different final vowels were cross-spliced (e.g., "pick up a pole" and "pick up a pail"). This provides a potentially more powerful window on sensitivity to long-range coarticulation; the Mismatch stimuli should slow lexical access of the final target word, since the coarticulation is consistent with another word, which should compete more strongly for lexical access.

The second factor was (*Coarticulation*) *Resistance*. After low coarticulation resistance consonants such as [p], the full vowel in the final word in the utterance, "pick up a pole", is likely to influence the realization of the reduced vowels in “up” and “a.” In contrast, [t] is high coarticulation resistant, so anticipatory coarticulation from the vowel would be less likely if the final word were “toll.” Therefore, we selected words beginning with High Resistance ([t, s, S]) or Low Resistance ([p, f]) segments.

Predictions

We used a variant of the visual world paradigm with two printed words as response choices (e.g., *pole, pail*), as subjects heard sentences like, "pick up a pole". Our first question is exploratory: whether and when subjects might begin to favor one word based on anticipatory coarticulatory cues. Given a low resistance consonant, it is possible that subjects could begin to pick up information about the final vowel as early as the vowel in "up." When Match and Resistance are crossed, an interaction is predicted: the effect of Match should be most apparent at the Low level of Resistance.

Methods

Participants Thirty-one undergraduate students at the University of Connecticut participated in this experiment for course credit. All were native English speakers with normal or corrected-to-normal vision and reported normal hearing.

Table 1: Low and High Coarticulation resistance items. Numbers indicate quadruple set membership.

Low Coarticulation Resistance Pairs		High Coarticulation Resistance Pairs	
1.pail,pole	6.fake,folk	1.tail,toll	6.sake,soak
2.pea,porch	7.fail,foal	2.tea,torch	7.sail,sole
3.paste,post	8.feel,fault	3.taste,toast	8.seal,salt
4.pan,pool	9.feet,fog	4.tan,tool	9.seat,sauce
5.peak,pork	10.feed,ford	5.teak,torque	10.seed,sword
	11.field,fall		11.shield,shawl
	12.feet,fort		12.sheet,short

Materials Twelve quadruples of words were chosen for the study. Each was composed of a pair of words starting with a low coarticulation resistance consonant (e.g., *pail/pole*), and a pair starting with a high coarticulation resistance consonant (e.g., *tail/toll*; see Table 1 for the full set). A number of constraints were observed in the selection of these quadruples. (1) The phonemes /p/ or /f/ were used as low coarticulation resistance consonants, and /t/, /s/, or /ʃ/ were used as the high coarticulation resistance consonants. (2) Highly discriminable front or back vowels were used to maximize acoustic differences

between words and also allow for maximal acoustic difference in anticipatory vocalic coarticulation. The front vowels used were: /i/, /eɪ/ and /æ/. The back vowels were: /əʊ/, /ɑ/, /o/ and /u/. (3) When possible, we used the same final consonant in all words in a quadruple, while also varying length and frequency as little as possible. These constraints had to be relaxed in a few cases in order to find enough items. However, the most critical portion of any noun in our design is the initial consonant and vowel, which constrain the potential for long-range coarticulation. ANOVAs confirmed that items did not differ reliably in any of these characteristics.

Each word in a quadruple was recorded with the same sentence frame (e.g., "Pick up a"). This particular sentence frame was selected to be as naturalistic as possible, while also containing neutral vowels (/ə/ in "up" and "a"), thereby maximizing the chance of observing long-range coarticulatory effects. A male, native English speaker recorded all of the sentences at a moderately fast speaking rate, which produced observable long-range coarticulation. The auditory stimuli were recorded and presented in 16-bit resolution at a 44.1 kHz sampling rate.

The spoken sentences were all cross-spliced at the onset of the noun-initial consonant. In the Match condition, two tokens of the same recording were spliced together, to ensure that any effects in the Mismatch condition were not due to artifacts of cross-splicing. In the Mismatch condition, the carrier phrase from one recording (e.g., the "pick up a" portion of "pick up a pail") was cross-spliced with the noun from another recording (e.g., the "pole" of "pick up a pole"). Average durations were 121 ms for "pick", 171 ms for "up a", and 317 ms for nouns.

Table 2: Mean F2-F1 (Hz) for *up* and *a* by condition.

	Coarticulation Resistance			
	Low		High	
	up	*a*	*up*	*a*
Front Vowels	789	793	811	1064
Back Vowels	772	612	808	938
ΔV	17	181	3	126

Acoustic analysis The formants of vowels of 'up' and 'a' of each target sentence were measured. Vowel formant center frequencies were measured using LPC and FFT spectra with reference to a wideband spectrogram. Measurements were made at the most stable portion of the middle of the vowel. Following Ladefoged (1993) we used F2-F1 (second formant - first formant) as a measure of vowel backness. The results are summarized in Table 2, collapsing over Front (/i, eɪ, æ/) and Back (/əʊ, ɑ, o, u/) vowels. F2 and F1 are more widely spaced for front than back vowels, so F2-F1 should be greater for front vowels. High Resistant consonants should yield smaller vowel backness differences than Low Resistant consonants. All of these differences were observed in the mean F2-F1 values at both 'up' and 'a'. Thus, we were successful in providing long-range anticipatory coarticulation cues in the materials.

Procedure Participants were seated at a comfortable distance from a computer screen (approximately 60 cm). Eye movements were monitored with an SR Research EyeLink 1000 desktop-mounted (remote) system, sampling at 500 Hz. Spoken sentences were presented to participants through headphones.

Each trial started with a drift correction procedure (participants briefly fixated a central dot). Then, a central fixation cross appeared. Participants clicked on it to begin the trial. When the cross was clicked, the members of a word pair appeared on the screen, one on the left and one on the right, with target and distractor position counter-balanced and pseudo-randomized. (We did not use pictures because we were unable to find enough highly imageable words meeting our phonological constraints; see McMurray, Tanenhaus & Aslin, 2009, and Huettig & McQueen, 2007, for precedents of using printed words in the visual world paradigm to obtain fine-grained time course measures of speech perception and spoken word recognition.) After a delay of 500 ms, the spoken sentence was presented over headphones. Participants were instructed to click on the final word of each sentence. The trial ended when the participant clicked on a printed word.

Each participant was presented with all 48 experimental trials and 44 filler trials. Twenty-two fillers consisted of rhyming pairs (to direct attention away from the onset similarity of critical items), and the remaining twenty-two were non-rhyming pairs. Half of each of these sets were presented with an auditory sentence cross-spliced to produce mismatching coarticulation, and the other half were spliced with another token of the same sentence, using exactly the same procedure as for the experimental stimuli (due to space constraints, we do not report analyses of filler items here). Experimental and filler trials were presented in random order following four (filler) practice trials. Half the experimental trials were presented in the Match condition and half were presented in the Mismatch condition. Half the trials in the Match and Mismatch conditions were Low Coarticulation Resistance words and the others were High. Thus, participants were given 12 trials in each of the four possible conditions. The word pairs (*pail/pole*) were counterbalanced between participants, such that a participant heard one of the pair (*pail*) in the Match condition, and the other (*pole*) in the Mismatch condition. Across participants, each word pair was presented the same number of times in Match and Mismatch conditions, and each word appeared equally often on the left or right.

Results

Data from two participants was excluded from analyses because of poor eye tracker calibration. Mean accuracy was at least 0.99 in all conditions. Figure 1 shows the average time course of target and competitor fixation proportions (at 20 ms intervals). Qualitatively, one

Table 3: Growth curve analysis of competitor fixation proportions. See text for details.

	Model fit			Match			Coarticulation Resistance			Match x CR		
	-2LL	ΔD	p	Estimate	*t*	*p*	Estimate	*t*	*p*	Estimate	*t*	*p*
Base	5996.8	-	-	-	-	-	-	-	-	-	-	-
Intercept	6013.8	17.1	0.001	0.074	4.39	<0.001	0.011	0.63	0.528	-0.058	-2.43	0.017
Linear	6018.9	5.1	0.166	-0.138	-2.16	0.031	-0.043	-0.68	0.500	0.130	1.44	0.149
Quadratic	6097.1	78.2	<0.001	-0.122	-7.57	<0.001	-0.001	-0.04	0.971	0.067	2.94	0.003
Cubic	6151	53.9	<0.001	0.111	6.92	<0.001	0.033	2.04	0.041	-0.082	-3.60	<0.001

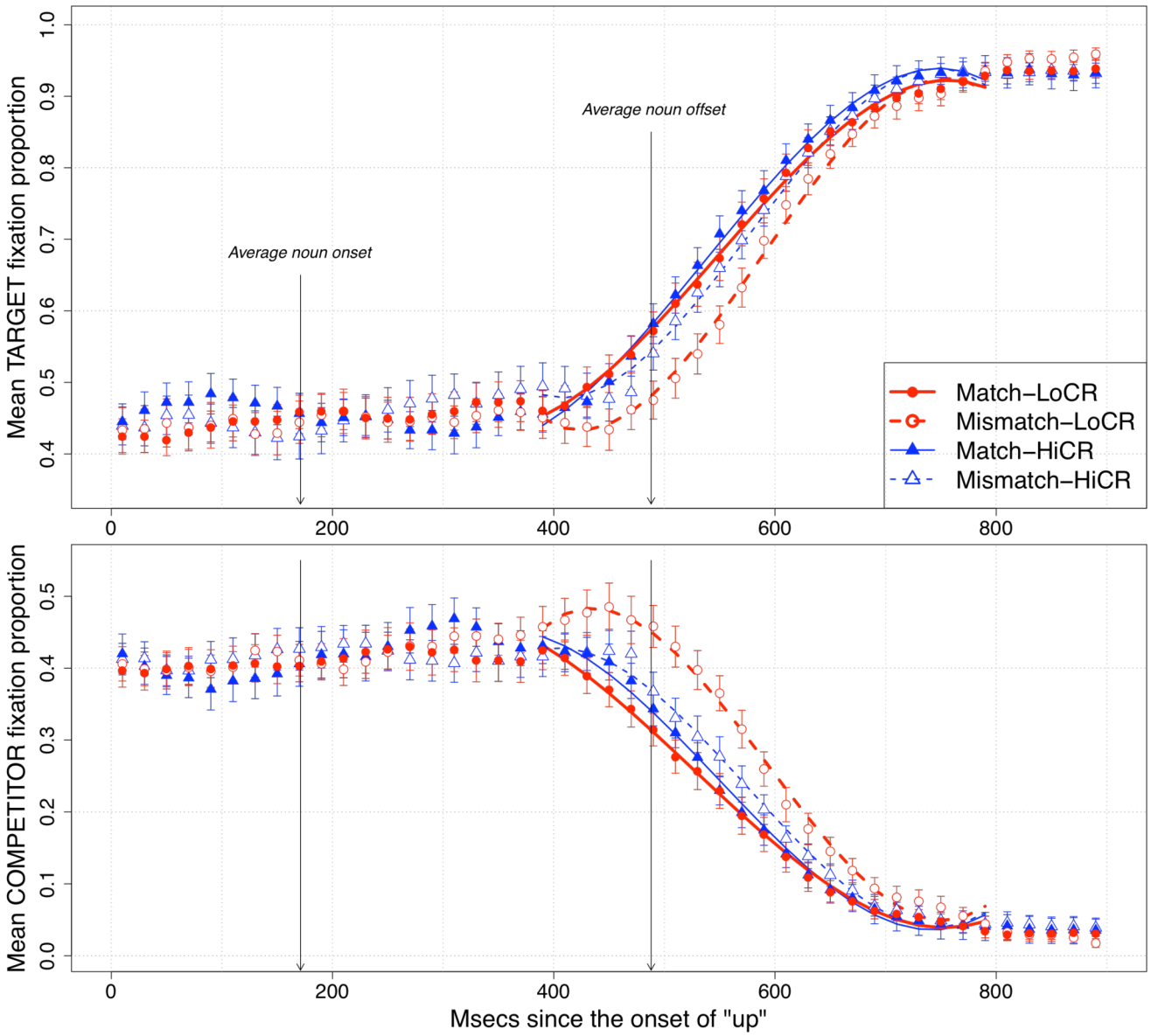


Figure 1: Mean fixation proportions to targets (top) and competitors (bottom). Symbols show observed data. Lines show growth curve fits for the 371-800 ms analysis window. Note different y-axis ranges.

condition stands out. The target is fixated most slowly in the Mismatch, Low Resistance condition, and a complementary increase in competitor fixations is also observed. The effect is not apparent until midway through the noun; however, the difference can only be due to coarticulatory detail available prior to word onset, since the signal in that condition was identical to that in the Match, Low Resistance condition from noun onset onward. There was also a slight trend towards an effect of Match at the High level of Coarticulation Resistance.

For the statistical analyses, we applied *growth curve analysis* (GCA; Mirman, Dixon, & Magnuson, 2008).

GCA is a variant of multi-level modeling that fits orthogonal power polynomial terms to over-time data. Conceptually, curves are fit to subject mean proportions at the lowest level of condition combinations (e.g., the Match level at the Low level of Resistance), and GCA assesses whether the curve parameters differ. Notably, this approach is dynamically consistent: the average of subject-level fits is equivalent to the fit to averaged data.

It is clear from Figure 1 that there were no differences between conditions until midway through the noun. Therefore, we constrained the GCA to a window beginning 200 ms after noun onset, the approximately earliest point where noun-driven changes in fixation proportions could be observed (371 ms after "up" onset), and ending at target asymptote (800 ms).

While any order polynomial can be used, the first three terms are conceptually easiest to link to visual world data. The *intercept* is recentered, such that it is analogous to mean proportion in the analysis window, and so directly analogous to an ANOVA on mean fixation proportion (indeed, although standard ANOVAs are less powerful than the multi-level modeling afforded by GCA, ANOVAs on mean proportion in this analysis window converge with the GCA intercept analyses). The *linear* term is the mean slope over the analysis window, the *quadratic* term reflects bowing of the primary curve inflection, and the *cubic* term capture inflections at the tails (necessary for fitting the s-like curves here).

Table 4: Mean competitor fixation proportions in the 371-800 ms analysis window by condition.

	Coarticulation Resistance	
	Low	High
Match	0.198	0.209
Mismatch	0.271	0.225

Because effects on targets and competitors are logically complementary (since these categories represent the two primary objects of fixation), we will present just the competitor analysis. The analyses are summarized in Table 3, with main effects of Match on intercept (greater mean proportion in Mismatch than Match conditions), linear slope and quadratic curvature (due to more negative slope and greater curvature in Mismatch conditions that follows from the greater lag preceding the drop-off in competitor fixations), and the cubic term (due to the early initial rise in the Mismatch, Low condition). There were no main effects of Coarticulation Resistance, but there were interactions of Match and Coarticulation Resistance on intercept, quadratic, and cubic terms. The intercept interaction is crucial (and space does not permit discussion of the other interactions); simple effects analyses reveal a significant effect of Match at Low Resistance (t=19.5, p<0.001) but not at High (t=1.5). Relevant means are presented in Table 4. Thus, the predicted interaction was observed: the effect of Mismatch was only reliable in the context of Low Coarticulation Resistance consonants.

Discussion

This study addresses the question of whether long-range coarticulatory information influences the time course of lexical activation and competition. We systematically varied (a) the Coarticulation Resistance (high vs. low) of the onset consonant of a monosyllabic target noun and (b) Coarticulatory "Match", i.e., whether the long-range anticipatory coarticulation matched or mismatched the vowel of the target noun that was ultimately heard.

While we observed trends associated with Coarticulatory Match and Resistance, the effects were driven largely by a difference in one condition: responses were slowed in the Mismatch, Low Resistance condition (though there was also a slight trend toward an effect of Mismatch at High Coarticulation Resistance).

The fact that the influence of anticipatory coarticulation was most apparent at low Coarticulation Resistance is consistent with phonetic analyses of long-range coarticulation, as those effects are simply more likely to propagate over segments (like [p]) that do not impose strong constraints on the position of the tongue tip or tongue body. The fact that influences of anticipatory coarticulation were most apparent in the Mismatch conditions is consistent with our expectation that these would be subtle effects and that misleading cues might be required to elicit detectable changes in lexical access (cf. Dahan, Magnuson, Tanenhaus & Hogan, 2001).

Notably, the effects do not emerge until midway through the final noun. This is surprising, as these effects must be due to anticipatory coarticulatory effects, since the nouns in the Match and Mismatch conditions were identical; those conditions differed only the "pick up a" portion of the instruction. It may be that anticipatory effects are detected as they occur, but require combination with confirmatory bottom-up evidence before they have a detectable impact. If front or back vowel qualities are detected on the vowel in 'up', this indicates that such a vowel is forthcoming, but it may still be several syllables away. This could prime appropriate phonological representations without driving strong lexical activation.

We do not mean to imply that lexical access initially proceeds based on local, bottom-up cues and other constraints are integrated after a delay (cf. Swinney, 1979). Instead, we note that when constraints are *gated* by bottom-up information but are integrated continuously, weak effects can appear to be late effects (Dahan, Magnuson & Tanenhaus, 2001; Shillcock & Bard, 1993).

Even though the effects we observed were relatively late and relatively modest, we did find a reliable influence of long-range coarticulatory anticipation. This reveals how extraordinarily sensitive listeners are to the rich sea of subphonemic details (some strong, some subtle) in which the islands of stability we describe as "phonemes" are embedded. This reinforces the view that coarticulation and other sources of variability in speech are not noise or

problems listeners must overcome. Rather, variability is largely lawful, enabling rapid rates of information transmission (Elman & McClelland, 1986; Fowler, 1986) due to local and anticipatory constraints it provides at multiple temporal scales.

Acknowledgments

Supported in part by NIH HD-40353 to Haskins Labs, and NSF grants CAREER 0748684 and 0642300 to JSM.

References

Bladon, R. A., & Al-Bamerni, A. (1976). Coarticulation resistance in English /l/. *J. of Phonetics, 4*, 137-150.

Browman, C., & Goldstein, L. (1992). Articulatory phonology: An overview. *Phonetica, 49*, 155-180.

Dahan, D., Magnuson, J. S., and Tanenhaus, M. K. (2001). Time course of frequency effects in spoken-word recognition: Evidence from eye movements. *Cognitive Psychology, 42,* 317-367.

Dahan, D., Magnuson, J. S., Tanenhaus, M. K., & Hogan, E. M. (2001). Subcategorical mismatches and the time course of lexical access: evidence for lexical competition. *Language & Cog. Processes, 16*, 507-534.

Elman, J. L. & McClelland, J. L. (1986). Exploiting the lawful variability in the speech wave. In J. S. Perkell & D. H. Klatt (Eds.), *Invariance and variability of speech processes*. Hillsdale, NJ: Erlbaum.

Fowler, C. A. (1980). Coarticulation and theories of extrinsic timing. *Journal of Phonetics, 8*, 113-133.

Fowler, C. A., (1986). An *event* approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics, 14,* 3-28.

Fowler, C. A. (2005). Parsing coarticulated speech in perception: Effects of coarticulation resistance. *Journal of Phonetics, 33(2),* 199-213.

Fowler, C. A., & Brancazio, L. (2000). Coarticulation Resistance of American English Consonants and its Effects on Transconsonantal Vowel-to-Vowel Coarticulation. *Language and Speech, 43* (1), 1-41.

Ganong, W.F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception & Performance, 6,* 110-125.

Gaskell, M. G. & Marslen-Wilson, W. D. (2001). Lexical ambiguity and spoken word recognition: bridging the gap. *Journal of Memory and Language,* 44, 325-349.

Gow, J. D. (2001). Assimilation and anticipation in continuous spoken word recognition. *Journal of Memory and Language, 45*, 133-159.

Hawkins, S. (2003). Roles and representations of systematic fine phonetic detail in speech understanding. *Journal of Phonetics, 31*, 373-405.

Heid, S., & Hawkins, S. (2000) An acoustical study of long-domain /r/ and /l/ coarticulation. *Proc. 5th Seminar on Speech Production: Models and Data.* (ISCA). Kloster Seeon, Bavaria, Germany. 77-80.

Hockett, C. F. (1955). A Manual of Phonology. *International Journal of American Linguistics (Memoir II)*. Baltimore: Waverley Press.

Huettig, F., & McQueen, J. M. (2007). The tug of war between phonological, semantic and shape information in language-mediated visual search. *Journal of Memory and Language, 57(4),* 460-482.

Ladefoged, P. (1993). *A Course in Phonetics* (3rd ed.). NY: Harcourt Brace Jovanvich Inc.

Ladefoged, P., & Broadbent, D. E. (1957). Information conveyed by vowels. *J. Acoust. Soc. Am., 29,* 98–104.

Liberman, A. M., Delattre, P. C., & Cooper, F. S. (1952). The role of selected stimulus variables in the perception of unvoiced stop consonants. *American Journal of Psychology, 65*, 497.

Magnuson, J. S. (2008). Nondeterminism, pleiotropy, and single word reading: Theoretical and practical concerns. In E. Grigorenko & A. Naples (Eds.), *Single Word Reading*, pp. 377-404. Mahweh, NJ: Erlbaum.

Magnuson, J. S., McMurray, B., Tanenhaus, M. K., and Aslin, R. N. (2003). Lexical effects on compensation for coarticulation: The ghost of Christmash past. *Cognitive Science, 27,* 285-298.

McMurray, B., Tanenhaus, M., & Aslin, R. (2009). Within-category VOT affects recovery from "lexical" garden paths: Evidence against phoneme-level inhibition. *J. of Memory and Language, 60* (1), 65-91.

Mirman, D., Dixon, J.A., & Magnuson, J.S. (2008). Statistical and computational models of the visual world paradigm: Growth curves and individual differences. *J. of Memory and Language, 59,* 475-494.

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of vowels. *Journal of the Acoustical Society of America, 24,* 175–184.

Recasens, D. (1984b). V-to-C coarticulation in Catalan VCV sequences: An articulatory and acoustical study. *Journal of Phonetics, 12*, 61-73.

Recasens, D. & Espinosa, A. (2009). An articulatory investigation of lingual coarticulatory resistance and aggressiveness for consonants and vowels in Catalan. *J. Acoustical Society of America, 122(4),* 2228-2298.

Salverda, A. P., Dahan, D., & McQueen, J. M. (2003). The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension. *Cognition, 90*, 51-89.

Shillcock, R. C., & Bard, E. G. (1993). Modularity and the processing of closed class words. In G.T.M. Altmann & R. C. Shillcock (Eds.), *Cognitive models of speech processing* (pp. 163–185). Erlbaum.

Swinney, D. A. (1979). Lexical access during sentence comprehension: (Re)consideration of context effects. *J. of Verbal Learning and Verbal Behavior, 18,* 645–659.

Tanenhaus, M. K., Spivey-Knowlton, M., Eberhard, K., & Sedivy, J. C. (1995). Integration of visual and linguistic information is spoken-language comprehension. *Science, 268*, 1632–1634.

West, P. (2000). Perception of distributed coarticulatory properties of English /l/ and /r/. *Journal of Phonetics, 27*, 405-425.

Cross Cultural Differences in Implicit Learning

Sachiko Kiyokawa (kiyo@isc.chubu.ac.jp)
Department of Psychology, Chubu University
1200 Matsumoto-cho
Kasugai, Aichi 487-8501 Japan

Zoltán Dienes (dienes@sussex.ac.uk)
School of Psychology, University of Sussex
Falmer, Brighton BN1 9QH, UK

Daisuke Tanaka (tanaka@rstu.jp)
Faculty of Regional Sciences, Tottori University
Koyama-cho-minami 4-101, Tottori City, Tottori 680-8551, Japan

Ayumi Yamada (ayumi.yamada@gmail.com)
Human Innovation Research Center, Aoyama Gakuin University
Shibuya 4-4-25, Shibuya-ku, Tokyo 150-8366, Japan

Abstract

Previous studies have indicated cross cultural differences in conscious processes such that Easterners have a preference for a more global perspective and Westerners for a more analytical perspective. We investigated whether these biases also apply to implicit learning. In Experiment 1, Japanese and British participants were asked to attend to one of the two aspects of a set of GLOCAL strings, global or local. The results showed that they could learn the AG implicitly only from the attended level in both cultural groups. They also showed that the global superiority in implicit learning was found only for the Japanese. In Experiment 2, these cultural differences were examined without manipulating the participants' attention. The results indicated implicit learning only at the global and not the local level for the Japanese, but equal learning of both levels by the British. We concluded that cultural biases strongly affect the type of unconscious knowledge that people acquire.

Keywords: cultural differences; selective attention; implicit learning; artificial grammar learning; global/local.

Role of Selective Attention in Implicit Learning

When repeatedly exposed to large amounts of information, we can acquire some abstract knowledge, such as rules or covariations between variables, without being aware of it. This phenomenon has been known as *implicit learning*. Since Reber's pioneering work on it (Reber, 1967), implicit learning has been studied using several paradigms, for example, serial reaction time (SRT) task or artificial grammar (AG) learning (for reviews, see Dienes, 2008; Reber, 1989; Shanks, 2005).

Reber (1989) suggested that we can implicitly learn some knowledge with a minimal amount of attention. Several researchers have agreed with the claim (e.g. Perruchet & Vinter, 2002; Whittlesea & Dorken, 1993). Based on this claim, it can be supposed that some attentional selection should occur in implicit learning.

Previous studies on the role of selective attention in implicit learning (e.g. Cock, Berry, & Buchner, 2002; Jiménez & Méndez, 1999; Rowland & Shanks, 2006) have provided supportive evidence to Reber's claim. However, these studies have mainly used the SRT task and few studies have investigated the role of selective attention in AG learning.

Seger (1998) argued that different mechanisms may underlie learning in the SRT task and in AG learning. Specifically, SRT task involves the acquisition of perceptual motor implicit knowledge, whereas AG learning involves acquiring implicit knowledge for the purpose of making judgments. Similarly, Boucher and Dienes (2003) speculated that sequential tasks such as SRT involve error correction mechanisms based on prediction, whereas AG learning may involve an automatic chunking mechanism. Although some researchers suggest that there is a common mechanism in these two tasks (e.g. Perlman and Tzelgov, 2006), the roles of selective attention in implicit learning may differ in SRT and AG learning. This claim needs to be further tested.

The first attempt to investigate the role of selective attention in AG learning was conducted by Tanaka, Kiyokawa, Yamada, Dienes, and Shigemasu (2008). They developed a new method using GLOCAL strings (an example is shown in Figure 1) to manipulate selective attention. GLOCAL strings are chains of compound letters (Navon, 1977). A compound letter represents one large letter (i.e. a global letter) composed of a set of small letters (i.e. local letters). A critical feature of this stimulus is that while a GLOCAL string can be read as one string by using global letters (NVJTVJ in Figure 1), it can also be read as a string using local letters (BYYFLB in Figure 1). Since GLOCAL strings can simultaneously represent two different strings following different AGs, we can examine whether the participants can learn the two AGs—one is attended while the other is unattended—by manipulating their

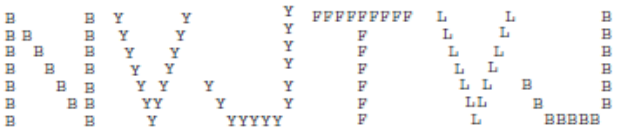


Figure 1. An Example of GLOCAL Strings.

attention. Using the GLOCAL strings, Tanaka et al. revealed that participants could learn an AG only from the attended level of the GLOCAL strings. They concluded that selective attention plays a critical role in AG learning.

Tanaka et al. (2008) also found the global superiority in AG learning. In Experiment 1, the classification accuracy for the attended grammatical strings was higher in the global attention condition than in the local attention condition. In Experiment 2, they examined whether or not the information at the unattended level was encoded by using a Stroop paradigm. They found the global superiority again. These results suggest that there is a global/local asymmetry in implicit learning. This tendency is consistent with the claim for a general preference for processing at the global level (see Navon, 2003, for a review).

Cultural Differences in Attention

Cultural psychology literature has suggested that there are cultural differences in attention between Easterners and Westerners (for reviews, see Nisbett, 2003; Nisbett & Miyamoto, 2005; Nisbett, Peng, Choi, & Norenzayan, 2001). Specifically, Easterners tend to pay attention to a scene globally, whereas Westerners do so locally.

Masuda and Nisbett (2001) examined whether Easterners attend to context more than Westerners do. They presented Japanese and American participants with animated vignettes of underwater scenes (in Study 1) or with photos of an animal in the wild (in Study 2) and asked the participants to report the contents. In a subsequent recognition test, the participants were shown previously seen objects as well as new objects, either in their original setting or in novel settings, and were then asked to judge whether or not they had seen the objects. The results showed that Easterners made more statements about contextual information and relationships than Westerners did. They also found that Easterners recognized previously seen objects more accurately when they saw them in their original settings rather than in the novel settings, whereas this manipulation had relatively little effect on Westerners.

Kitayama, Duffy, Kawamura, and Larsen (2003) developed the framed line test (FLT). In this test, participants were presented with a square frame in which a vertical line was printed. They were then presented with another square frame of a different size and required to draw a line that was the same either in absolute length (absolute task) or in proportional length (relative task). Kitayama et al. (2003) found that the performance of Westerners in the absolute task was better than that in the relative task, whereas for Easterners the pattern was reversed. The results indicated that Westerners are better able to filter out or to suppress contextual frame information, whereas Easterners are better at incorporating contextual information. Ishii and Kitayama (2007) extended the results to non-student participants and to auditory tasks.

Based on these studies, there is a possibility that the global superiority found by Tanaka et al. (2008) is limited to Easterners. In Tanaka et al. (2008), the participants were all Japanese. Because they tended to pay more attention to the information at the global level than that at the local level, they might have had difficulty filtering out the information at the global level when asked to focus on the strings at the local level. As a result, global superiority in AG learning emerged.

Present Study

In the present study, we determined whether or not the global superiority in AG learning found by Tanaka et al. (2008) would be obtained for Western participants. Based on the cross cultural literature, there ought to be cultural differences in attention. Since selective attention plays an important role in AG learning, we hypothesized that the cultural differences in attention would have an effect on AG learning: Easterners could learn AG from the global level more than from the local level, whereas Westerners could not.

We modified the procedures used by Tanaka et al. (2008) in the following ways. The first is the instructions in the learning session. In Tanaka et al. (2008), the participants were asked to write down the strings represented either by global or by local levels during their presentation. This procedure in the learning session might help the participants to learn the attended grammar more than otherwise because they can read the strings that they wrote down on the paper. In the present study, the participants were asked only to look at the strings carefully and sometimes write them down after the GLOCAL string had disappeared.

The second is in the procedure followed in the test session. In the previous study, the participants were not instructed regarding on which AG they should base their judgments. This procedure might cause the degree of each type of AG learning to be underestimated. In the present study, we divided the test into two sessions and the participants were explicitly told to judge the grammaticality based on one of the two AGs in each session. The order of these two test sessions was counterbalanced among participants.

In the third modification, the participants were asked to show the basis of their judgment in each grammaticality judgment trial. Although this point will not be discussed in this paper owing to space constraints, this procedure allows us to examine in more detail whether participants'

grammaticality judgment was based on an implicit or explicit basis.

Experiment 1

This experiment was designed to examine whether or not the global superiority found in Tanaka et al. (2008) could be replicated by Japanese and British participants.

Method

Participants Forty undergraduates from Chubu University and forty-two from the University of Sussex participated in the experiment and received a course credit following the completion of the experimental session. Assignments on types of GLOCAL strings and the order of the tests were counterbalanced. None of the students had previously participated in the same kind of experiment.

Stimuli The same AGs as those in Tanaka et al. (2008) were used. Grammar 1 comprised five letters (J, N, T, V, and X), as did Grammar 2, which used the letters B, F, L, Y, and Z.

Eighteen grammatical strings with a length of three to six letters were constructed from each AG. Two types of GLOCAL strings were constructed from these strings, following the two AGs. One type of GLOCAL string followed Grammar 1 at the global level and Grammar 2 at the local level; this was reversed for the other type of GLOCAL string, so grammar was counterbalanced across levels.

GLOCAL strings were presented as white uppercase letters against a black background. Small letters were used, printed in 12-point MS Gothic font. One large letter was the height of seven small letters. Eight small letters were arranged horizontally to obtain F, J, L, and X; nine to obtain B, N, T, and Y; thirteen to obtain V; and seven to obtain Z. The height of a large letter on the screen was approximately 3.2 cm and the width was approximately 1.8–3.0 cm. The distance between the display and the participants was approximately 60 cm.

Twenty strings following each grammar used in the test phase were composed of five or six letters. These were not GLOCAL but regular letter strings. Half of these were used in the learning phase and will be referred to as 'presented grammatical strings'. The remaining strings were not identical to any of the strings presented in the learning phase and will be referred to as 'novel grammatical strings'. All of these grammatical strings were used to construct nongrammatical strings that violated both of the grammars by placing one or two characters in nonpermissible locations.

Four types of string pairs were constructed for the test phase. The first type—Global_Old—paired a presented grammatical string at the global level of GLOCAL strings in the learning phase with a nongrammatical one based on the AG extracted from the global level of the GLOCAL strings. The second type—Global_New—paired a novel grammatical string at the global level of GLOCAL strings in the learning phase with a nongrammatical one based on the AG that was extracted from the global level of the GLOCAL strings. Similarly, the third type was termed Local_Old, and the fourth Local_New. Each type comprised 20 pairs. Thus, there were 80 pairs in the test phase. Matching pairs of grammatical and nongrammatical strings in each type were randomized for each participant, subject to the constraint that the two strings should have the same length.

Procedure During the learning phase, 18 GLOCAL strings were presented on the display for 6 seconds. Half of the participants were asked to look at and memorize the GLOCAL strings represented by the large letters. The other half were asked to do so with respect to the strings represented by the small letters. The former was a global attention condition and the latter was a local attention condition. The participants were also required to write down the string represented by the attended level when the message was shown on the display. The message was presented about once in ten trials. Each GLOCAL string was presented six times. A mask stimulus comprising many '+' signs in the area where the GLOCAL strings were intended to be displayed was presented for the 1-second interval between the presentation of GLOCAL strings.

At the beginning of the test phase, the participants were informed that two strings would be presented in the upper and lower regions of the display, each of the two levels of the training strings followed a set of rules, and each string of a pair followed one set of rules. The test phase consisted of two sessions: a test on the global level and one on the local level. Half of the participants were required to press the key associated with the string that they judged to be grammatical, extracted from the global aspects of the GLOCAL strings in the first test session and the local in the second one. The remaining participants were asked to do the same thing, first for the local and then for the global level.

Forty pairs were presented to each participant in a random order in each test session. A pair of strings remained on display until the participants pressed one of the two keys. The presentation of strings from a pair in the upper region was also randomized for each participant, subject to the constraint that one type of pair (i.e. the grammatical string) would be presented equally in each region.

After making judgments, the participants were asked what they based their judgments on and were required to choose one of the following five answers:

1. Random responding or guessing: Your judgment had no basis whatsoever; you could have just flipped a coin to make your judgment.
2. Intuition: You have some confidence in your judgment, but you have no idea why.
3. Familiarity: The sequence seemed familiar or unfamiliar for reasons you could not state.
4. Recollection: You recollected or failed to recollect seeing all or part of the sequence in the training phase.
5. Rules: You based the judgment on a rule or rules you could state if asked.

All of the instructions were presented in Japanese for the participants from Chubu University and in English for those from the University of Sussex. The English instructions were back translated and checked to make sure they had the same meaning as those in Japanese.

Design A 2 × 2 × 2 mixed design was employed. The first factor was global/local. The participants were instructed to attend to the global or local level of the learning phase. This was a between-participants factor. The second factor was attended/unattended. In the test phase, half of the pairs could be judged correctly on the basis of the grammar extracted from the attended level of the GLOCAL strings, whereas the other half could be judged correctly on the basis of the grammar extracted from the unattended level. This was a within-participants factor. In addition, the third factor, presentation, indicated whether or not the grammatical string had been presented before in the learning phase. This was also a within-participants factor.

Results and Discussion

Figure 2 shows the mean classification accuracy for each condition in the test phase. First, the proportion of accurate classifications was subjected to a 2 × 2 × 2 mixed ANOVA with global/local, attended/unattended, and presentation (old or new grammatical string) as factors for each cultural group.

For the Japanese participants, the main effect of the attended/unattended level was significant (F (1, 38) = 231.43, $p < .001$). Accuracy concerning the grammar of the attended level was higher than that of the unattended level. The interaction between the global/local and attended/unattended levels was also significant (F (1, 38) = 11.04, $p < .01$). The results of the simple main effect revealed that accuracy in the global attention condition was higher than that in the local attention condition at the attended level (F (1, 76) = 10.67, $p < .01$), whereas this effect disappeared at the unattended level ($F < 1$).

For the British participants, the main effect of the attended/unattended level was significant (F (1, 40) = 69.03, $p < .001$), indicating that accuracy concerning the grammar of the attended level was higher than that of the unattended level. The interaction between the global/local and attended/unattended levels was not significant (F (1, 40) = 1.43).

In order to examine the possibility that the participants could learn the AG from the unattended level to some degree, we compared the proportions accurately classified with chance (.5) in each condition. With respect to the Japanese participants, accuracy for Unattended_Old and Unattended_New in both the global and local conditions was not higher than chance (ts < 1). With respect to the British participants, on the other hand, accuracy for Unattended_Old in the global condition was significantly higher than chance (t (20) = 2.91 $p < .01$).

We replicated the results of Tanaka et al. (2008) for the Japanese participants. They were able to learn the AG from the global level more than from the local level only when they paid attention to the level itself. Global superiority, however, was not found for the British participants. In addition, the result of a t-test showed that they were able to learn the AG not only from the attended level but also from the unattended level when asked to pay attention to the global level. This might indicate that they have a tendency to process more information from the local level than from the global level.

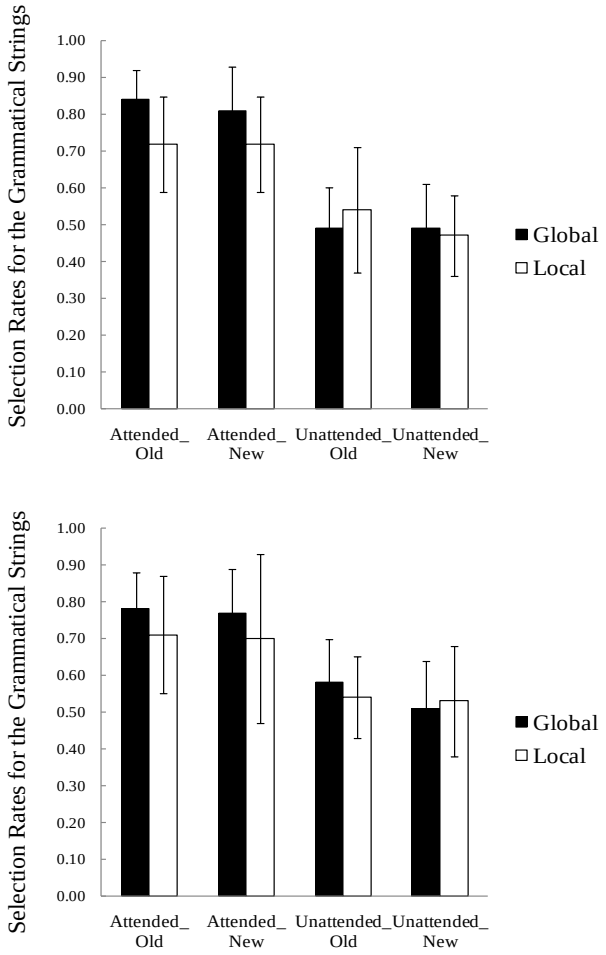


Figure 2. Mean Selection Rates for the Grammatical Strings in the Pairs of Attended_Old, Attended_New, Unattended_Old, and Unattended_New Grammatical Strings with Nongrammatical Strings in Each Attention Condition with Standard Deviations. Top panel: Japan; Bottom panel: UK.

In sum, the results suggest that there are cultural differences in implicit learning such as AG learning. This may be explained by attentional bias between Easterners and Westerners. In Experiment 2, therefore, we examined whether or not there would be cultural differences in implicit learning without manipulating the participants' attention.

Experiment 2

This experiment was designed to examine whether or not there would be cultural differences in attention and AG learning when the participants were free to manage their attention in the learning session.

Method

Participants Twenty undergraduates from Chubu University and eighteen from the University of Sussex participated in the experiment and received a course credit following the completion of the experimental session. Assignments on types of GLOCAL strings and the order of test were counterbalanced. None of the students had previously participated in the same kind of experiment.

Stimuli The same AGs as those in Experiment 1 were used.

Procedure The same procedures were used as in Experiment 1 except for the following points. First, the participants' attention was not manipulated in the experiment. They were asked to look at the GLOCAL strings not at one level but at both levels. Second, two questions were asked at the end of the experiment. The first question was, 'Which aspect—the bigger letters or the smaller letters—did you pay more attention to in the first session?' The second was, 'By how much more do you think you attended to your favorite aspect, e.g. twice as much, three times as much, etc.?'

Design A 2 × 2 × 2 mixed design was employed. The first factor was cultural group. This was a between-participants factor. The second factor was global/local. This was a within-participants factor. In addition, the third factor was presentation. This was also a within-participants factor.

Results and Discussion

Figure 3 shows the mean classification accuracy for each condition in the test phase. First, the proportion of accurate classifications was subjected to a 2 × 2 × 2 mixed ANOVA with cultural group, global/local, and presentation as factors.

The main effect of the global/local factor was significant (F (1, 36) = 12.13, p < .01). The interactions between cultural group and global/local and between cultural group and presentation were also significant (F (1, 36) = 9.52, p < .01; F (1, 36) = 5.50, p < .05, respectively). The results of the simple main effect revealed that accuracy in the global grammar was higher than that in the local one for the Japanese participants (F (1, 36) = 21.57, p < .0001), whereas this effect disappeared for the British participants (F < 1). The results of the simple main effect showed that accuracy in the new grammatical stimuli was higher than that in the old ones for the British participants (F (1, 36) = 3.75, p = .06), whereas this effect was not found for the Japanese participants (F (1, 36) = 1.91, p > .10).

In order to examine the possibility that the participants could learn the AG from each level, we compared the proportions accurately classified with chance (.5) in each condition. With respect to the Japanese participants, accuracy only for the Global_Old and Global_New strings was significantly higher than chance. With respect to the British participants, on the other hand, accuracy only for all types of strings was significantly higher than chance.

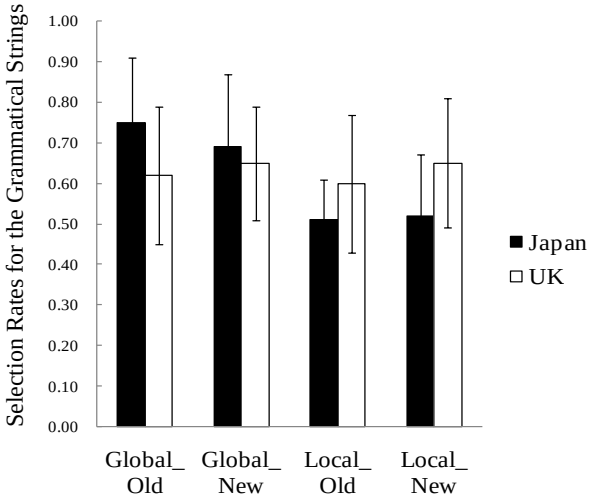


Figure 3. Mean Selection Rates for the Grammatical Strings in the Pairs of Global_Old, Global_New, Local_Old, and Local_New Grammatical Strings with Nongrammatical Strings with Standard Deviations in Each Cultural Group.

Table 1. The ratio of participants who preferred each level in the learning session.

	Japan	UK
Global	90.0	50.0
Local	5.0	44.4
Equal	5.0	5.6

(%)

To examine the attentional bias in the learning session, we compared the ratio of the participants who paid more attention to each level between cultural groups. Table 1 shows the ratio of the participants who preferred each level. A *chi-square* test revealed that more participants preferred the global level to the local one in Japan, whereas this pattern was not found (*chi-square* (N = 38) = 8.36, p < .05). The result indicated that there were cultural differences in attention during learning sessions. It also indicated that this attentional bias might cause the cultural difference in AG learning.

General Discussion

In the present study, we examined whether or not there are cultural differences in implicit learning using an AG learning paradigm with GLOCAL strings. In Experiment 1, the global superiority in AG learning was obtained only for the Japanese participants. This indicated that there was a cultural difference in implicit learning between Easterners and Westerners. However, it was common that selective attention played a critical role in AG learning. Although the British participants could memorize the grammatical strings at the unattended level, in both cultural groups, the participants could learn only the AG extracted from the attended level. The results strongly support the necessity for attention in AG learning suggested by Tanaka et al. (2008).

The results of Experiment 2 revealed that the Japanese participants could learn the AG only from the global level, whereas the British participants could learn from both levels. It was also found that there was attentional bias in the learning session: most of the Japanese participants tended to pay more attention to the global level, whereas half of the British participants tended to pay more attention to the local level. Based on the cultural difference in attention, the results of AG learning should be interpreted as showing not that the British participants could simultaneously learn both AGs, but that some learned the AG only from the global level and others only from the local level, corresponding to their attentional preference.

It is necessary to examine whether or not there are also any cultural differences in learning or judging strategy between Easterners and Westerners based on the participants' judgment bases. Previous studies (e.g. Nisbett, 2003; Nisbett et al., 2001) have suggested that Eastern people prefer holistic processing, whereas Western people prefer analytic. It should be examined whether this tendency can be applied to implicit learning situations such as our task setting.

Conclusion

Selective attention plays a critical role in implicit learning in both Eastern and Western cultural groups. However, there are cultural differences in global/local asymmetry. Specifically, Japanese participants learned the AG extracted from the attended global level better than that from the local one, whereas British participants did not. The cultural difference in AG learning seems to be caused by cultural biases in attention between Easterners and Westerners.

References

Boucher, L., & Dienes, Z. (2003). Two ways of learning associations. *Cognitive Science*, *27*, 807-842.

Cock, J. J., Berry, D. C., & Buchner, A. (2002). Negative priming and sequence learning. *European Journal of Cognitive Psychology*, *14*, 27-48.

Dienes, Z. (2008). Subjective measures of unconscious knowledge. In R. Banerjee & B. Chakrabarti (Eds.), *Models of brain and mind: Physical, computational and psychological approaches*. Amsterdam: Elsevier.

Ishii, K., & Kitayama, S. (2007). Holistic attention to context in Japan: A test with non-student adults. *Japanese Journal of Social Psychology*, *23*, 181-186.

Jiménez, L., & Méndez, C. (1999). Which attention is needed for implicit sequence learning? *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *25*, 236-259.

Masuda, T. & Nisbett, R. E. (2001). Attending holistically versus analytically: Comparing the context sensitivity of Japanese and Americans. *Journal of Personality and Social Psychology*, *81*, 922-934.

Navon, D. (1977). Forest before trees: The precedence of global features in visual perception. *Cognitive Psychology*, *9*, 353-383.

Navon, D. (2003). What does a compound letter tell the psychologist's mind? *Acta Psychologica*, *114*, 273-309.

Nisbett, R. E. (2003). *The geography of thought: How Asians and Westerners think differently*. New York: Free Press.

Nisbett, R. E. & Miyamoto, Y. (2005). The influence of culture: Holistic versus analytic perception. *Trends in Cognitive Science*, *9*, 467-473.

Nisbett, R. E., Peng, K., Choi, I., & Norenzayan, A. (2001). Culture and systems of thought: Holistic vs. analytic cognition. *Psychological Review*, *108*, 291-310.

Perlman, A., & Tzelgov, J. (2006). Interactions between encoding and retrieval in the domain of sequence-learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *32*, 118-130.

Perruchet, P., & Vinter, A. (2002). The self-organising consciousness: A framework for implicit learning. In R. M. French & A. Cleeremans (Eds.), *Implicit learning and consciousness: An empirical, philosophical and computational consensus in the making*. Hove, U.K.: Psychology Press.

Reber, A. S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning & Verbal Behavior*, *6*, 855-863.

Reber, A. S. (1989). Implicit learning and tacit knowledge. *Journal of Experimental Psychology: General*, *118*, 219-235.

Rowland, L. A., & Shanks, D. R. (2006). Attention modulates the learning of multiple contingencies. *Psychonomic Bulletin & Review*, *13*, 643-648.

Seger, C. A. (1998). Independent judgment-linked and motor-linked forms of artificial grammar learning. *Consciousness & Cognition*, *7*, 259-284.

Shanks, D. R. (2005). Implicit learning. In K. Lamberts & R. L. Goldstone (Eds.), *Handbook of cognition*. London: Sage.

Tanaka, D., Kiyokawa, S., Yamada, A., Dienes, Z., & Shigemasu, K. (2008). Role of selective attention in artificial grammar learning. *Psychonomic Bulletin & Review*, *15*, 1154-1159.

Whittlesea, B. W. A., & Dorken, M. D. (1993). Incidentally, things in general are particularly determined: An episodic-processing account of implicit learning. *Journal of Experimental Psychology: General*, *122*, 227-248.

Interaction between lexical and syntactic structures in transcoding from verbal to Arabic numerals

Rafael Hurtado (rghurtadoh@unal.edu.co)
Department of Physics, Universidad Nacional de Colombia, Crr. 30 No 45-03
Bogotá, Colombia

Mariela Orozco-Hormaza
Diego F. Guerrero
Department of Psychology, Universidad del Valle, Calle 13 No 100-00
Cali, Colombia

Keywords: Numerical transcoding; number processing; number representation; network analysis.

Abstract

We explore the relationship between the lexical structure and the syntactical structure of numerical expressions in number transcoding from the oral verbal format to the Arabic digital format. The experimental setup included asking six to eight-year old Spanish-speaking children attending elementary school to write in Arabic format a set of dictated numerals. The method of analysis includes the construction of a relational representation of children's production and the use of clustering techniques to identify patterns. The model relates children and dictated numerals by children's accomplishment and generates a subsidiary similitude relation between dictated numerals with patterns that show differentiated structures. We find that the presence or absence of a verbal expression for the Hundred position digit in the dictated numeral marks one of the structures. The second structure comes from the role of two digit numbers (e.g. 20 or 34): homogeneous in the Decade position and heterogeneous in the Thousand position. We interpret these results as consistent with the semantic-lexical internal number representation model by R.J.D. Power and M.F. Dal Martello, The dictation of Italian numerals, *Language and Cognitive Processes*, 5, 237-254; 1990.

Introduction

McCloskey, Caramazza and Basili made in 1985 (McCloskey, Caramazza, & Basili, 1985) a pioneering proposal in adult neuropsychology of cognitive mechanisms in number processing and calculation. They posted the existence of a number comprehension mechanism, a number production mechanism and a calculation system with a numerical internal representation. The number comprehension mechanism translates a numerical input into the internal representation and the number production mechanism translates from the internal representation into the output format required. Comprehension and production mechanisms process both verbal and Arabic numerals and distinguish lexical-processing and syntactic-processing components. Lexical processing involves comprehension or production of individual elements in a number. Syntactic processing involves the relations among numerical elements in order to comprehend or to produce a number as a whole. The number internal representation in this model is semantic-abstract, with numbers expressed in basic quantities associated to abstract forms of the base-10 representation (e.g. $2x10^1$ and $5x10^0$ for *twenty five*). The model specifies that the internal representation activates the syntactical production mechanism and that the highest power of ten in the central representation generates a syntactic frame with the appropriate number of slots or positions for the production of the Arabic numeral. This representation is independent of the structure of verbal representation. The semantic base-10 representation implies that the peculiarities of verbal stimulus do not exert any influence on the Arabic production mechanisms.

Number transcoding, defined as translating a numeral from one code to another one, has been used to explore number processing in impaired subjects and adolescents with mild retardation (Barrouillet, Camos, Perruchet & Seron, 2004; Granà, Lochy, Girelli, Seron, & Semenza, 2003; McCloskey, 1991; McCloskey & Caramazza, 1987; McCloskey *et al*., 1985; McCloskey & Macaruso, 1995; Seron & Noël, 1995) and in groups of children (Barrouillet *et al*., 2004; Power & Dal Martello, 1990; 1997; Seron & Fayol, 1994). From these works two main explanatory models for number processing in children arise: One proposes the existence of a semantic representation of numerical quantities, which can be either semantic-abstract (Macaruso, McCloskey & Aliminosa, 1993; McCloskey, 1991; McCloskey *et al*., 1985) or semantic-lexical (Power & Dal Martello, 1990); the other perspective proposes an asemantic model where number processing results from assembling rules or algorithms (Deloche & Seron, 1982, 1987; Barrouillet *et al.*, 2004). Some models propose both routes.

In the semantic-lexical model posted by Power and Dal Martello (1990) working with children "the form of the representation reflects the structure of the subject's verbal numeral system.". To these authors the internal representation is due to the interpretation of the verbal numeral and its internal structure tied to the verbal code. They propose sum and product structures as the main relationships between numerical concepts and, from the subject perspective, an overwriting rule for the sum operator and a concatenation rule for the product in Arabic

production. In this model a product relation produces numerals as *200* and *2000* from primitive numerical concepts as Ones, Teens, Tens and Hundred, *a*nd a sum relation to obtain composed numerals as *220*, *2020*, *20220*. For Power and Dal Martello (1990) "Every non-primitive number is represented as the sum or product of two unequal numbers". They show that syntactic errors, related to the numeral structure (e.g. *20045* for *two hundred forty five*), are more frequent than lexical errors related to the exchange of digits (e.g.. *255* for *two hundred forty five*).

An asemantic model that predicts children production was developed by Barrouillet *et al*., (2004), "ADAPT: A Developmental, Asemantic and Procedural Model for Transcoding from Verbal to Arabic Numerals", with two versions: the basic ($ADAPT^{BASIC}$) and the advanced ($ADAPT^{ADV}$). The first version describes the transcoding process for numerals up to 99 and the second up to six digit numbers. The main proposal of the $ADAPT^{BASIC}$ model is that when a verbal string for transcoding corresponds to a representational unit stored in Long Term Memory (LTM), this string is processed as such, whereas its transcription is the result of the direct memory retrieval of its digital form. In the $ADAPT^{ADV}$ version, an algorithm is used to explain transcoding from the verbal numeral format to the Arabic digital format. For writing numerals the model assumes that the verbal expression is coded in a phonological code and analyzed by a parser. The results of the analysis are processing units that can be either elements whose digital transcription can be directly retrieved from LTM or separators, as Hundred and Thousand. These processing units trigger the transcoding rules that in a serial process generate either slots, digits or digit chains. (Barrouillet *et al*., 2004)

The present work heads number processing by identifying patterns of relationship between the lexical (phonetic) structure of verbal expressions and one syntactical structure of the Arabic expressions in children's production transcoding numbers from the oral verbal to the Arabic digital format. Our model includes a positional analysis using similarity, structural equivalence (Borgatti, Everet, Freeman, 2002; Johnson, 1967; Wasserman & Faust, 1994) and betweenness (Freeman, 1979; Girvan & Newman, 2002). The novelty of this analysis is the use of a relational representation children's production. The most relevant findings are two syntactic structures: One pined to the presence or absence of a verbal expression in the Hundred position of the dictated numeral, and other one which differentiates Decade forms (e.g. *20* or *70*) and Decade-Unit forms (e.g. *34* or *76*) when they are at Thousand position (e.g. *30.432* or *37.432*).

Method

In this section we present the experimental setup and the model of analysis used in this study.

Participants

The experimental setup included the dictation of numerals to 207 children attending first, second and third grades of elementary school in Colombia. First grade children (65) were 6 years and 8.3 months old with standard deviation SD=2.9 months, second grade children (74) were 7 years and 7.0 months old (SD=3.0 months) and third grade children (68) were 8 years and 7.8 months old (SD=3.0 months).

Experimental procedure

The experiment consisted of asking children to transcode a set of dictated numerals from the oral verbal format to the Arabic one. Numerals were dictated in Spanish in a 20-minute testing session to each child. Dictated numbers were of a higher order than those traditionally taught at the corresponding school grade. Four dictation lists with the same set of numbers were generated randomly in each grade and assigned randomly to each child. The set of numbers included all lexical and syntactic forms for three, for and five digit numerals in Arabic code. The notation used to classify numerals is: *a, b, c, d* and *e* letters which represent the digital forms different from zero and correspond to the basic quantities of the numeral in the Arabic format (ex. *3789* is *abcd*). In all of the cases *a* represents the highest order quantity of the Arabic format. The *0* digit represents a null quantity. We use *x* to represent either a basic or a null quantity.

Method of analysis

Network analysis perspective uses graphs to represent and study complex systems in terms of relations between elements or parts. It gives a detailed account of structural and dynamical properties of systems by identifying patterns at the microscopic level and macroscopic phenomena. (Freeman, 2004; Wasserman & Faust, 1994)

Our model of analysis defines ties between numerals in terms of correlations in children's production: Two numerals are tied with a strength equal to the number of children with production syntactically and lexically correct or incorrect for both numerals. The relation defined in this way does not have internal structure and therefore it is suitable for exploring elementary structures and patterns. As the relation expresses similarity, we use graph visual representations (Freeman, 2005), Johnson's hierarchical clustering (Johnson, 1967), structural equivalence (Breiger, Boorman, & Arabie, 1975; Burt, 1976) and tie betweenness for cohesive subgroups (Girvan & Newman, 2002). These techniques have been implemented in computer codes as Netdraw (Borgatti, 2002) and UCINET (Borgatti, Everett & Freeman, 2002).

Visual exploration*:* Low dimensional graph representations of networks are often used to organize nodes and ties in landscapes where the location of a node is related to its actual location in the network. We use a spring embedded model with node repulsion, equal edge length and similitude

by geodesic distances. (Borgatti, 2002; Freeman, 2005; Wasserman & Faust, 1994)

Proximity clustering: Johnson's hierarchical clustering identifies clusters from correlations. (Borgatti, Everett, & Freeman, 2002; Johnson, 1967; Wasserman & Faust, 1994) It uses similarities or dissimilarities to find a series of nested partitions by departing from N partitions, each one with a node, and joining partitions in successive steps according to their relative distance from adding individual distances to other nodes.

Structural equivalence: Nodes with structurally equivalent positions in the network have identical relational profiles and are tied to the same nodes. We use two techniques to identify structural equivalence: Convergence of iterated correlations between profile vectors (CONCOR), based on Pearson's correlation (Borgatti, Everett, & Freeman, 2002; Breiger, Boorman, & Arabie, 1975; Lerner, 2005; Wasserman & Faust, 1994), and the Euclidean distance from each node to all other nodes (Borgatti, Everett, & Freeman, 2002; Burt, 1976; Wasserman & Faust, 1994).

Clustering by tie betweenness: Freeman (1979) defined betweenness centrality for a vertex or node as the number of shortest paths between any other two nodes passing by the node. Girvan and Newman (2002) extended this concept in order to identify cohesive subgroups by removing ties or links with the highest values of betweenness. This method relies on the fact that a tie linking two non-overlapping clusters must have the highest betweenness in the graph. In our analysis we combine successive weak strength and high betweenness ties removal to identify clusters.

Results

The results from applying the selected analysis techniques to children's production are presented for each grade including frequency analysis, visual exploration, proximity clustering, structural equivalence and tie betweenness.

First grade children's production included 1.047 (44.7%) correct answers from 2.340 dictated numerals (36 numerals to 65 children). The highest frequency of correct answers was for naught numerals (40%), then for *a0c* (25%), *ab0* (18%) and *abc* (17%) numerals. Table 1. The mean of correct answers for child is 29 (SD 11); for naught numerals 47 (SD 6.5), *a0c* 29 (SD 2.5), *ab0* 21 (SD 2.6), and *abc* numerals 20 (SD 2.7). All clustering techniques confirm the existence of three subgroups: 1) *a00*; 2) *a0c*; 3) *ab0* and *abc*. Figure 1 shows the graph obtained with Netdraw (Borgatti, 2002). The ties between numerals in the graph have at least 53 children each. Shapes of nodes indicate numeral classification: *abc*, *ab0*, *a0c*, *a00*.

Second grade children's production included 1.613 correct answers from 4.366 dictated numerals (37.0%). The highest percentage of correct answers was for naught numerals (*a000*) (78.8%), then for *a00d* (49.9%), *ab00* (45.8%), *abc0* (39.7%), *abcd* (34.0%), *a0c0* (27.4%) and *a0cd* numerals (24.2%). See Table 1. The mean and standard deviation (SD) of correct answers per child are: for naught numerals 51.2 (SD 4.2), for *a00d* 32.4 (SD 3.8), for *ab00* 29.5 (SD 1.4), for *abc0* 25.8 (SD 4.1), for *abcd* 22.1 (SD 3.2), for *ab0d* 19.8 (SD 4.1), for *a0c0* 17.8 (SD 1.8) and for *a0cd* numerals 15.7 (SD 3.1). Numbers with verbal expressions in the Hundred position (*abxx*) have 38.2% of correct answers and numbers without verbal expression in the Hundred position (*a0xx)* have 31.4%. For *abxx* numerals the mean of correct answers per child is 24.8 (SD 4.5) and for *a0cx* numerals 16.4 (SD 2.5).

Table 1: Frequency of correct answers

FIRST GRADE				SECOND GRADE					
100	60	*246*	25	*1000*	61	*8190*	29	*3452*	20
200	51	*240*	24	*2000*	53	*1524*	28	*8367*	19
500	49	*450*	24	*4000*	52	*6900*	28	*9070*	19
300	48	*198*	23	*6000*	51	*7009*	28	*3402*	18
400	46	*810*	23	*3000*	50	*7800*	28	*4030*	18
600	43	*980*	22	*5000*	50	*4730*	27	*4031*	18
700	43	*367*	21	*8000*	50	*9600*	27	*6080*	18
800	40	*190*	20	*7000*	48	*6980*	26	*7019*	18
900	39	*452*	20	*9000*	46	*9670*	26	*8090*	18
603	32	*524*	19	*2008*	40	*2198*	24	*1024*	17
307	31	*730*	19	*1004*	35	*1504*	23	*4731*	17
504	31	*731*	19	*8100*	32	*6985*	23	*7809*	17
809	31	*985*	19	*8300*	31	*7819*	23	*3052*	16
402	30	*360*	18	*1500*	30	*8197*	23	*5040*	16
905	29	*520*	18	*2100*	30	*9673*	23	*9073*	15
108	26	*670*	18	*3002*	30	*5240*	21	*2098*	13
206	26	*819*	18	*4700*	30	*5246*	21	*5046*	13
701	26	*673*	16	*5200*	30	*6085*	21	*8067*	13
				3400	29	*8307*	21	*8097*	13
				8007	29	*2108*	20		

THIRD GRADE					
40000	50	*19603*	39	*90307*	35
50000	50	*40985*	39	*73450*	34
20000	48	*60819*	39	*10673*	33
10000	47	*67819*	39	*20031*	33
70000	47	*73400*	39	*60809*	32
30000	46	*98307*	39	*73452*	32
52198	45	*52190*	38	*70452*	31
80000	45	*70450*	38	*46085*	30
20731	44	*80524*	38	*35040*	29
50190	44	*81524*	38	*40085*	29
52100	44	*90367*	38	*30046*	28
24731	42	*10603*	36	*24030*	27
50198	42	*35246*	36	*46080*	27
90000	42	*46985*	36	*19003*	25
60000	41	*67809*	36	*24031*	25
81500	41	*19673*	35	*35046*	25
30246	40	*80520*	35	*67009*	23
98367	40	*81520*	35	*98007*	23

Frequency analysis, Johnson's hierarchical clustering, structural equivalence from Euclidean distances and tie bewteenness indicates the existence of four main subgroups: Naught numerals, *a00d*, *a0cx*; and *abxx* numerals. However, structural equivalence from CONCOR gives four subgroups: two for naught numerals (one containing *3000*, *6000*, *8000* and *9000*), one for *a0xx* and one for *abxx* numerals. Results from CONCOR have to be considered because the main ability of this technique is to recognize

patterns. Figure 2 shows the graph obtained with Netdraw (Borgatti, 2002). The ties between numerals in the graph have at least 55 children each. Shapes of nodes indicate numeral classification: *abcd*, *abc0*, *ab0d*, *ab00*, *a0cd*, *a0c0*, *a00d*, *a000*.

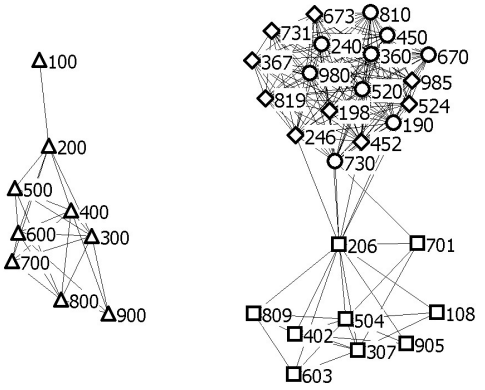


Figure 1: Graph for the relation between numerals in first grade children's productions.

Third grade children's production included 1.992 correct answers from 3.672 dictated numerals (54.3%). The highest achievement was for naught numerals (*a0000*) (73.8%). Numerals with verbal expression in the Hundred position (*axcxx*) are more frequently correct (55.8%) than numerals without verbal expression on this position (*ax0xx*) (39.7%), excluding naught numerals. The mean of correct answers per child for *axcxx* numerals is 38.2 with SD 3.8 and for *ax0xx* is 27.9 with SD 3.3. (See Table 1)

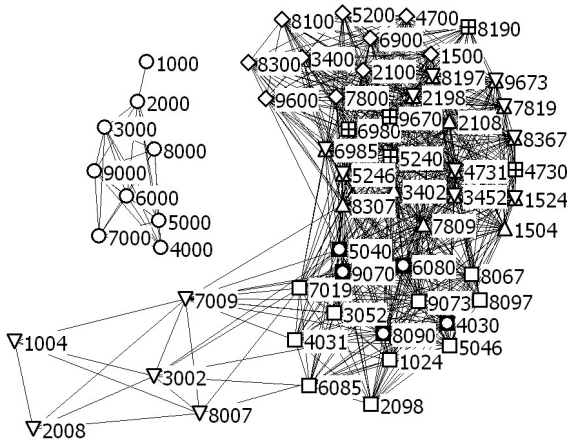


Figure 2: Graph for the relation between numerals in second grade children's productions.

Frequency analysis, visual exploration (Figure 3), Johnson's hierarchical clustering, structural equivalence from Euclidean distances and tie betweeness are consistent with the existence of three main subgroups: Naught numerals, *ax0xx* and *axcxx* numerals. Structural equivalence from CONCOR gives six subgroups: Naught numerals excluding 60000; *ax0xx* excluding naught numerals; *axcxx* excluding *90307*, *24731* and *35246*; *90307*; *60000* and *24731*; and *35246* (See Table 2). This classification is consistent with a differentiation between *axcxx* and *ax0xx* numerals.

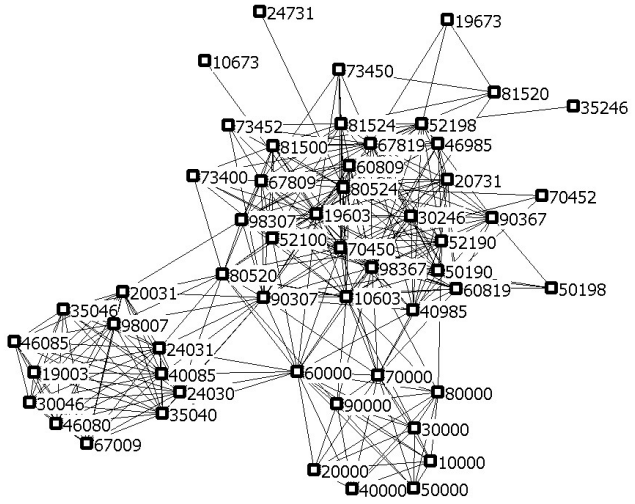


Figure 3: Graph for the relation between numerals in third grade children's productions.

Table 2: Numeral clustering for third grade.

Numeral	Johnson	CONCOR	Euclidean	Betweenness
10000	1	1	1	1
20000	1	1	1	1
30000	1	1	1	1
40000	1	1	1	1
50000	1	1	1	1
60000	1	5	1	1
70000	1	1	1	1
80000	1	1	1	1
90000	1	1	1	1
20031	2	2	2	2
30046	2	2	2	2
40085	2	2	2	2
10603	3	3	3	3
60809	3	3	3	3
90307	3	5	3	3
50190	3	3	4	3
70450	3	3	3	3
80520	3	3	3	3
10673	3	3	3	4
20731	3	3	3	3
30246	3	3	3	3
40985	3	3	3	3
50198	3	3	4	3
60819	3	3	3	3
70452	3	3	3	3
80524	3	3	3	3
90367	3	3	3	3
19003	2	2	2	2
67009	2	2	2	2
98007	2	2	2	2
24030	2	2	2	2
35040	2	2	2	2
46080	2	2	2	2
24031	2	2	2	2
35046	2	2	2	2
46085	2	2	2	2
52100	3	4	3	3
73400	3	4	3	3
81500	3	4	3	3
19603	3	4	3	3
67809	3	4	3	3
98307	3	4	3	3
52190	3	4	3	3
73450	3	4	3	3
81520	3	4	3	4
19673	3	4	3	4
24731	3	6	5	4
35246	3	7	3	4
46985	3	4	3	3
52198	3	4	3	3
67819	3	4	3	3
73452	3	4	3	3
81524	3	4	3	3
98367	3	4	3	3

Discussion

First grade results show the existence of three subgroups of numerals in children's production: naught numerals, *a0c* and *abx* numerals. None of the analysis techniques shows any internal structure in the subgroups indicating that each one is homogeneous. Power & Dal Martello (1990) also

found differences in children's production between numerals ending in *00* (*a00*), numerals with internal zero (*a0c*), and numerals without internal zero (*ab0*, *abc*). Seron and Fayol (1994) predicted the following order of acquisition *a00* > *a0c* > *abc* = *ab0*. They did not find other significant difference between the proposed forms. Their results are consistent with a higher frequency of errors for *ab0* and *abc* numerals than for *a0c* numerals. For Barrouillet *et al.,* (2004) the Decade-Unit form (*bc*) in *abc* numerals and the Decade form (*b0*) in *ab0* numerals are retrieved from LTM as representational units and the error rate did not present any significant difference between the two forms in three and four digit numbers. In our results *ab0* and *abc* numerals are both similar and equivalent.

For second grade, results indicate the existence of four main subgroups: 1) naught numerals, 2) *a00d*, 3) *a0cx*, and 4) *abxx*. One clustering technique, CONCOR, indicates that *a00c* and *a0cx* numerals should belong to the same subgroup, in agreement with Figure 2. Additionally, in our study the *a00d* numeral with the highest rate of correct answers was *2008*, which has the same structure as the number of the year when the experimental testing was done. We consider that this fact might have influenced the production of *a00d* numerals.

We interpret these results as that the presence or absence of a verbal expression in the Hundred position of the dictated numeral determines the digital production. In the case of Spanish, three types of expressions are assigned to the Hundred verbal expression: *cien* for one hundred, *quinientos* for five hundred and *cientos* accompanied by a prefix in all other cases as in *trecientos* for three hundred. Note that *a1xx* numerals were more frequently correct than *abxx* ones (Table 1.).

An important difference between our results and Barrouillet *et al.*, (2004) is that they found higher rates of errors for the Unit form *0d* in *ax0d* than for the Decade form *c0* in *axc0* and the Decade-Unit for *cd* in *axcd*. Our results indicate the opposite situation in agreement with Power and Dal Martello (1990) and Seron and Fayol (1994).

In third grade production frequency analysis, visual exploration, Johnson's hierarchical clustering, structural equivalence and tie betweenness indicate the existence of three main subgroups: Naught numerals, *ax0xx* excluding naught numerals, and *axcxx* numerals. Note that *ax1xx* numerals were more frequently correct than *axbxx* ones (Table 1.). Additionally, *ab00e* numerals obtained the lowest frequency of correct answers, in contrast with *a00d* numerals that obtained a high frequency in second grade children's production.

However, CONCOR neatly differentiates between *abcxx* and *a0cxx* numerals. This result is not consistent with Barrouillet *et al.*, (2004) proposal for Decade-Unit and Decade forms as being retrieved directly from LTM. Instead, it might be explained by considering that the difference between these forms is the number of operations for their construction (one product for Decade forms and one product and one sum for Decade-Unit forms), as proposed by Barrouillet *et al.*, (2004) for complex Decade forms in French (e.g. *soixante-dix* for seventy and *quatre-vingt-dix* for ninety) or by Power and Dal Martello (1990).

Conclusions

First, the presence or absence of a verbal expression in the Hundred position of dictated four and five digit numerals generate differentiated structures in children's written production. This result indicates that the lexical (phonetic) structure in the oral verbal format interacts with the syntactic structure of Arabic digits in children's production in number transcoding.

Secondly, the difference found between *abcxx* and *a0cxx* numerals in children's production calls for new explicative efforts.

Finally, the method of analysis used gives valuable information that can not be obtained from frequency analysis or simple statistics.

Acknowledgments

This research was funded by COLCIENCIAS under grant 110645221365. We thank John Mora for exploratory work on the data sets.

References

Barrouillet, P., Camos, V., Perruchet, P., & Seron, X. (2004) ADAPT: A Developmental, Asemantic, and Procedural Model for Transcoding from Verbal to Arabic Numerals. *Psychological Review*, 111(2), 368-394.

Borgatti, S. P. (2002). NetDraw: *Graph Visualization Software*. Harvard, MA: Analytic Technologies.

Borgatti, S.P., Everett, M.G., & Freeman, L.C. (2002). Ucinet for windows: *Software for Social Network Analysis*. Harvard, MA: Analytic Technologies.

Breiger R., Boorman S., & Arabie P. (1975). An algorithm for clustering relational data, with applications to social network analysis and comparison with multi-dimensional scaling. *Journal of Mathematical Psychology*, 12, 328-383.

Burt, R. (1976). Positions in Networks. *Social Forces*, 55, 93-122.

Deloche, G., & Seron, X. (1982). From Three to 3: A differential analysis of skills in transcoding quantities between patients with Broca's and Wernicke's aphasia. *Brain*, 105, 719-733.

Deloche, G., & Seron, X. (1987). Numerical transcoding: A general production model. In G. Deloche & X. Seron (Eds.), *Mathematical disabilities: A cognitive neuropsychological perspective*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Freeman, L, C. (1979). Centrality in Social Networks: Conceptual Clarification. *Social Networks*, 1, 215-239.

Freeman, L.C. (2004). *The development of Social Network Analysis. A study in the sociology of science*. Vancouver: Empirical press.

Freeman, L.C. (2005). Graphic techniques for exploring social network data. In S. Wasserman, J. Scott and P. J. Carrington (Eds.) *Models and methods in social network analysis*. Cambridge, UK: Cambridge University Press.

Girvan, M., & Newman, M. E. (2002). Community structure in social and biological networks. *Proc. Natl. Accad. Sci. USA* 99, 7821-7826.

Granà, A., Lochy, A., Girelli, L., Seron, X., & Semenza, C. (2003) Transcoding zeros within complex numerals. *Neuropsicología*, 41, 1611-1618.

Johnson, S. C. (1967) Hierarchical clustering schemes. *Psychometrika*, 32, 241-253.

Lerner, J. (2005) Role assignments. In Brandes, U., & Erlebach, T. (2005). *Network Analysis: Methodological Foundations*. Lecture Notes in Computer Science. Tutorial. Berlin: Springer-Verlag.

Lochy, A., Pillon, A., Zesiger, P., & Seron, X. (2002) Verbal structure of numerals and digits handwriting: New evidence from kinematics. *The quarterly Journal of Experimental Psychology*, 55A (1), 263-288.

Macaruso, P., McCloskey, M., & Aliminosa, D. (1993). The Functional Architecture of the Cognitive Numerical-processing System: Evidence from a Patient with Multiple Impairments. *Cognitive Neuropsychology*, 10(4), 341-376.

McCloskey, M. (1991). Cognitive mechanisms in numerical processing: Evidence from acquired dyscalculia. *Cognition* 44, 107-157.

McCloskey, M., & Caramazza, A. (1987). Cognitive mechanisms in normal and impaired number processing. In G. Deloche & X. Seron (Eds.) *Mathematical disabilities: A cognitive neuropsychological perspective*. Hillsdale, NJ: Erlbaum.

McCloskey, M., Caramazza, A., & Basili, A. (1985). Cognitive mechanism in number processing and calculation: Evidence from dyscalculia. *Brian and Cognition*, 4, 171-196.

McCloskey, M. & Macaruso, P. (1995). Representing and using numerical information. *American Psychologist*, 50, 331-363.

Noël, M. P., & Seron, X. (1995). Lexicalization errors in writing Arabic numerals: A single case study. *Brain and Cognition*, 29, 151-179.

Power, R. J. D., & Dal Martello, M. F. (1990). The dictation of Italian numerals. *Language and Cognitive Processes*, 5, 237-254.

Power, R. J. D., & Dal Martello, M. F. (1997). From 834 to Eighty Thirty four: The reading of the Arabic numerals for Seven-years-old children. *Mathematical Cognition*, 3(1), 63-85.

Seron, X., & Fayol, M. (1994). Number transcoding in children: A functional analysis. *British Journal of Developmental Psychology*, 12, 281-300.

Wasserman, S., & Faust, K. (1994). *Social Network Analysis: Methods and Applications*. Cambridge: Cambridge University Press.

Language specific preferences in anaphor resolution: Exposure or Gricean maxims?

Barbara Hemforth (barbara.hemforth@parisdescartes.fr)
Laboratoire de Psychologie et de Neuropsychologie Cognitives, CNRS, Université Paris Descartes,
71 ave Edouard Vaillant, 92100 Boulogne-Billancourt, France

Lars Konieczny (lars@cognition.uni-freiburg.de)
Center for Cognitive Science, University of Freiburg, Friedrichstr. 50, 79098 Freiburg, Germany

Christoph Scheepers (c.scheepers@psy.gla.ac.uk)
Department of Psychology, University of Glasgow, 58 Hillhead Street, Glasgow, Scotland

Savéria Colonna (Saveria.Colonna@univ-paris8.fr)
Laboratoire Structure Formelles du Langage, CNRS, Université Paris 8, 59-61 rue Pouchet, 75849 Paris Cedex 17

Sarah Schimke (sarah.schimke@sfl.cnrs.fr)
Laboratoire Structure Formelles du Langage, CNRS, Université Paris 8, 59-61 rue Pouchet, 75849 Paris Cedex 17

Peter Baumann (p.bau@web.de)
Center for Cognitive Science, University of Freiburg, Friedrichstr. 50, 79098 Freiburg, Germany

Joël Pynte (joël.pynte@parisdescartes.fr)
Laboratoire de Psychologie et de Neuropsychologie Cognitives, CNRS, Université Paris Descartes,
71 ave Edouard Vaillant, 92100 Boulogne-Billancourt, France

Abstract

In this paper we will present evidence for language specific preferences in anaphor resolution from two series of experiments in English, German, and French. For within sentence anaphor resolution with "before" subclauses, we will show that English and German follow the generally assumed preference for the first mentioned NP or subject of the sentence, whereas French shows a clear preference for the object of the matrix clause. We will argue that our data can most easily be explained by a usage-based account, linking comprehension preferences to production preferences.

Keywords: Sentence processing; anaphor resolution; crosslinguistic differences; usage-based preferences

Introduction

It has been shown for many languages that the resolution of non-reflexive pronouns is strongly influenced by pragmatic factors such as topicality (in the sentence or in the discourse; Givon, 1983), the chain of causality, and other kinds of discourse relations (e.g. Kehler, 2002; Sanders & Noordman, 2000). On the sentence level, two of the factors that seem to play a role are a preference for the first mentioned antecedent (Gernsbacher, 1990), and a preference for the subject (Jaervikivi, van Gompel, Hyöna, & Bertram, 2005). These preferences are assumed to be valid across languages so that for subject-verb-object sentences like (1) a preference for the first noun phrase would generally be predicted, given that it is mentioned first and the subject at the same time.

(1) **English:** *The postman met the streetsweeper before he went home.*
French: *Le facteur a rencontré le balayeur avant qu'il rentre à la maison.*
German: *Der Briefträger hat den Strassenfeger getroffen bevor er nach Hause ging.*

More language specific predictions can be derived from accounts based on the availability of alternative constructions in the grammar of a particular language. According to the Gricean Maxim of Manner (Clarity), speakers should avoid ambiguous constructions in choosing unambiguous alternatives if they exist. If for an ambiguous construction an unambigous alternative exists for one of the readings, listeners may thus assume that the speaker would have chosen this alternative for the respective reading. From this reasoning, a preference for the reading without an unambiguous alternative will result for the ambiguous construction.

In this paper, we will compare closely matched constructions in English, French, and German (see examples 2-5) to investigate cross-linguistic differences in pronoun resolution. What makes the comparison of these languages particularly interesting, is the distribution of alternative constructions for the different interpretations of an ambiguous sentence like (1): In French, a highly frequent construction exists for binding an anaphoric pronoun to the subject of the matrix clause (2) which does not exist for German.

(2) **English:** *The street-sweeper met the postman before going home.*
French: *Le balayeur a rencontré le facteur avant de rentrer à la maison.*

Following the Gricean Maxim of Manner, the existence of this alternative predicts a preference for an object antecedent in sentences with full pronouns for French in contrast to the presumably saliency based preference for the subject for German. Listeners hearing a French sentence with "avant que" followed by a full pronoun will assume that the speaker would have used the unambiguous infinitival form in (2) had she intended the temporal clause to relate to the subject of the matrix clause. The pronoun is thus preferentially interpreted as relating to the object of the matrix clause for which no such alternative exists.[1]

English is an interesting case for comparison, given that an alternative construction with a zero anaphor exists for subject antecedents (2). This construction is, however, used less frequently than the infinitival construction in French. Gricean accounts would thus predict that English patterns with French with respect to pronoun resolution.

An unambiguous alternative for one of the readings may also influence frequencies of usage. In a small scale corpus analyses (100 sentences per language) we established the following distribution: 77% subject antecedents for German (Frankfurter Rundschau), 64 % subject antecedents for English (Wall Street Journal) and 100 % (Le Monde) or 85% (Google News groups) object antecedents for French. Frequency based accounts would thus position English between German and French.

Experiments

Series 1: Visual World Experiments

In our first series of experiments, participants (32 native French speakers, 32 native English speakers, and 24 native German speakers) were presented with pictures such as in Figure 1 showing two characters while they listened to sentences such as (3-6). Their task was to judge whether a sentence presented aurally matched the picture or not. All 16 experimental trials were "match" cases. Half of the 4 practice items as well as of the 24 filler items were "mismatch" cases. Mismatches were realized by including characters in the sentence that were not in the picture (such as: "The florist prepared a bouquet for the street-sweeper"). Mismatches were realized at different positions during the sentence.

[1]The same pattern would be predicted by Ariel's (1990) accessibility hierarchy: less informative anaphora are predicted to prefer more salient antecedents. The zero anaphor in the infinitival construction in French, prefers the subject as the most salient antecedent. Using a full pronoun can thus be interpreted as a cue to search for a less salient antecedent which would be the object in sentences such as (1).

Materials: In our experimental materials, the subclause introduced by *before*, *avant que,* or *bevor*, was semantically biased for the High Antecendent (HA, the subject of the sentence which is situated higher in the phrase structural representation of the sentence, 3,5), or the Low Antecedent (LA) the object (4,6) of the main clause as antecedent of the pronoun. To control for visual scanning preferences, the first mentioned character was either on the left (3,4) or on the right (5,6) side of the screen. As a between participants factor, we also switched the position of the characters for half of the participants, so that, for example, the postman was on the right of the screen and the street sweeper on the left.

(3) **French:** *Le facteur a rencontré le balayeur avant qu'il ramasse* ***les lettres.***
English:*The postman met the street-sweeper before he picked up* ***the letters.***
German: *Der Briefträger traf den Straßenfeger, bevor er* ***die Briefe e****insammelte.*

(4) **French:** *Le facteur a rencontré le balayeur avant qu'il ramasse* ***la poubelle.***
English:*The postman met the street-sweeper before he picked up* ***the trash.***
German: *Der Briefträger traf den Straßenfeger, bevor er* ***den Abfall*** *einsammelte.*

(5) **French:** *Le balayeur a rencontré le facteur avant qu'il ramasse* ***les lettres.***
English: *The street-sweeper met the postman before he picked up* ***the letters.***
German: *Der Straßenfeger traf den Briefträger, bevor er* ***die Briefe*** *einsammelte.*

(6) **French:** *Le balayeur a rencontré le facteur avant qu'il ramasse* ***la poubelle.***
English: *The street-sweeper met the postman before he picked up* ***the trash.***
German: *Der Straßenfeger traf den Briefträger bevor er* ***den Abfall*** *einsammelte.*

Eight lists were created such that each item appeared in a different condition across lists, but only once in each list. Participants were first presented with four practice items followed by one of the eight lists of experimental items mixed with 24 filler items. The lists were randomized individually. Participants received course credits for their participations. Each experiment lasted less than 30 minutes including calibration. Eye movements were recorded using the Eyelink II© system by SR research.



Figure 1: Example of the visual stimulus material

Results: We calculated the likelihood of a gaze on either of the two critical picture elements by time steps of 20 ms starting from 500 ms before the onset of the pronoun (*he*/*she*) and ending at 2000 ms after the onset of the pronoun. From these data, we calculated the *logodds* for a gaze on the first-mentioned referent at each time step. Values below zero represent more fixations on the object, values above zero more fixations on the subject. Figure 2 shows the results for English, Figure 3 for German, and Figure 4 for French. HA means High Antecedent and corresponds to the subject, LA means Low Antecedent and corresponds to the object. "Left" and "right" correspond to the position of the subject on the picture.

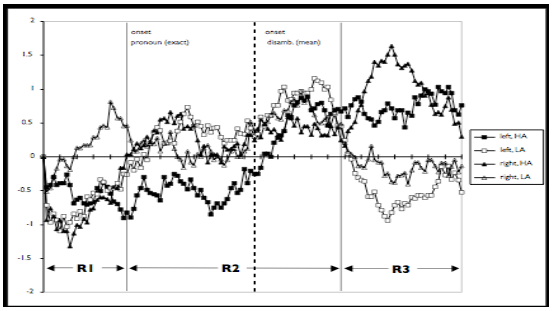


Figure 2: Time course analysis for English; log2(p(sub/p(obj)), HA=High Antecedent, subject; LA= Low Antecedent, object

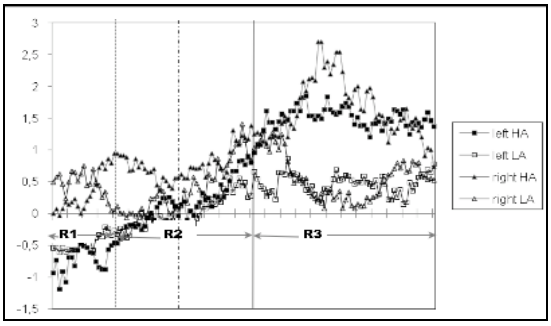


Figure 3: Time course analysis for German; log2(p(sub/p(obj)), HA=High Antecedent, subject; LA= Low Antecedent, object

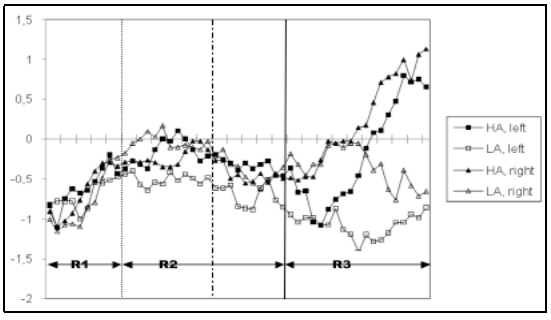


Figure 4: Time course analysis for French; log2(p(sub/p(obj)), HA=High Antecedent, subject; LA= Low Antecedent, object

Before the onset of the pronoun, marked by the first vertical line in Figures 2 to 4, participants had a tendency to fixate the object more often than the subject (for German speakers, this tendency is somewhat modulated by the position of the object). This is not surprising, given that the object was the last mentioned entity in the matrix clause. After the onset of the pronoun, participants did not show any preference for a short period of time. This probably reflects the time needed to process the pronoun plus the time for planning a saccade (at least 230ms + 250 ms = 480 ms). After this period, German and English speakers fixated the subject more often than the object, whereas French speakers fixated the object more often. Disambiguation can only start playing a role after the onset of the disambiguating word plus at least 480 ms (given the time needed for processing and saccade planning). The dotted vertical line reflects the mean onset of the disambiguation, the third vertical line shows the earliest point possible for disambiguation to kick in. Participants start fixating the corresponding character more often after this point. Note that the onset of disambiguating is earlier in German due to German word order.

We defined three critical time periods for each individual trial: the 500 ms period before the onset of the pronoun (R1), the time period from the onset of the pronoun until 480 ms after the onset of the disambiguating region (R2), and the remaining time steps until 2000 ms (R3). For each participant and condition, respectively item and condition, we calculated a single logodds value per time period. The summarized data across conditions for English, German, and French are shown in Table 1.

Table 1: Average log odds for gazes on the first-mentioned referent, broken down by region. Ninety-five percent confidence limits are listed in parentheses (by subjects / by items).

Language	**R1**	**R2**	**R3**
English	**-0.50** (±0.52/±0.43)	**+0.34** (±0.28/±0.19)	**+0.40** (±0.45 / ±0.32)
German	**-0.09** (±0.51 / ±0.32)	**+0.35** (±0.33 / ±0.34)	**+1.14** (±0.22 / ±0.32)
French	**-0.90** (±0.36 / ±0.28)	**-0.38** (±0.30 / ±0.22)	**-0.50** (±0.29 / ±0.29)

The eye movements show clear differences between the languages investigated. In the ambiguous region R2, we find a reliable preference to look at the subject of the matrix clause for English and German. In German, this extends even to the disambiguating region R3. In French, however, participants preferentially fixated the character corresponding to the object of the matrix clause.

One question remains to be answered at this point: Do the French fixation preferences reflect interpretational

preferences or possibly just differences in visual scanning patterns? Since the object of the matrix clause is at the same time the last entity mentioned before hearing the ambiguous pronoun, our French participants may have preferred to continue fixating the entity they just heard of until disambiguating information would be made available by the linguistic input. French participants did actually look at the character representing the object of the matrix clause more often than German and English participants even in Region 1. In order to test this possibility, we ran a further eye-tracking experiment with 32 native French speakers, using constructions with no structural alternative for either of the possible interpretations (7a-d). A subject preference would be predicted for these cases.

(7) a . *Le facteur a rencontré le balayeur. Puis il a ramassé* ***les lettres.***
The postman met the street-sweeper. Then he picked up **the letters.**
b . *Le facteur a rencontré le balayeur. Puis il a ramassé* ***la poubelle.***
The postman met the street-sweeper. Then he picked up ***the trash.***
c . *Le balayeur a rencontré le facteur. Puis il a ramassé* ***les lettres.***
The street-sweeper met the postman. Then he picked up ***the letters.***
d . *Le balayeur a rencontré le facteur. Puis il a ramassé* ***la poubelle.***
The street-sweeper met the postman. Then he picked up ***the trash.***

The set up of the experiment was identical to the earlier experiments. Since the preference for the more local referent could only be established for French, we will only present the French data here (see Figure 5).

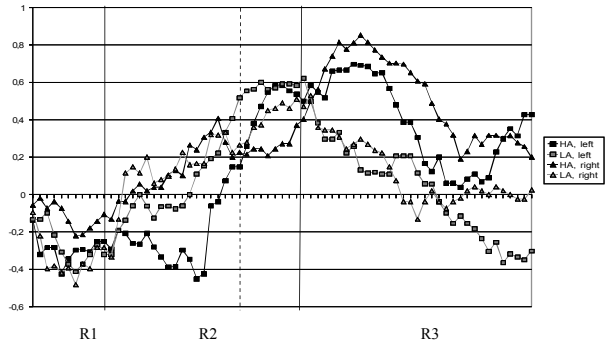


Figure 5: Time course analysis for French between-sentence anaphor resolution; log2(p(sub/p(obj)), HA=High Antecedent, subject; LA= Low Antecedent, object

As in the earlier experiments, the French participants started with an increased number of fixations to the object of the first sentence. However, after a short period without any preferences right after the onset of the pronoun, they look reliably more often at the character representing the subject of the matrix clause. Note, that the pictures we used in this experiment were identical to the ones used before. Clearly, French speakers do not have different visual scanning patterns. In cases where a subject preference is predicted, they clearly look at the character representing the subject more often, although the subject is the less local entity.

Figure 6 summarizes the results of all four experiments: Remember that values above zero reflect more looks to the subject, whereas values below zero reflect more looks to the object of the matrix/first clause. The most striking differences can be found in Regions 2 and 3: For within sentence pronoun resolution the subject is preferred as the antecedent for German and English, and likewise for between sentence anaphor resolution in French. The only deviating cases are French within sentence anaphors, showing a preference for the object.

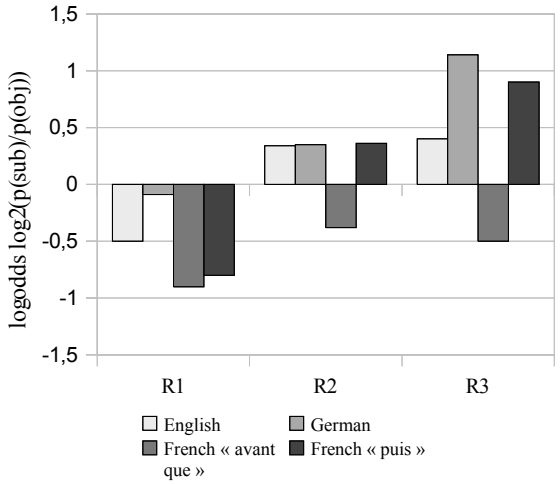


Figure 6: Average log odds for gazes on the first-mentioned referent, broken down by region.

The fixation patterns are thus far compatible with the corpus frequencies mentioned above. We can, however, not be fully sure that they really reflect interpretational preferences and not just fixation preferences. We therefore ran a series of questionnaire studies in all three languages to clarify this issue.

Series 2: Questionnaires

Materials and procedure. In this series of experiments, we presented participants with ambiguous sentences derived from the materials used in the eye tracking experiments and asked them to fill a gap in a paraphrase following each sentence to indicate their interpretation of the pronoun.

(7) **French:** Le facteur a rencontré le balayeur avant qu'il rentre chez lui.
Le _____ rentre chez lui.
English: The postman met the street-sweeper before he went home.
The _____ went home.

German: Der Briefträger hat den Straßenfeger getroffen, bevor er nach Hause ging.

Der _____ ging nach Hause.

We also included a cross-sentence condition (8), where the second sentence always started with "puis", "then", or "dann".

(8) **French:** Le facteur a rencontré le balayeur. Puis il est rentré chez lui.
German: Der Briefträger hat den Straßenfeger getroffen. Dann ist er nach Hause gegangen.
English: The postman has met the street-sweeper. Then he went home.

To control for semantic/pragmatic biases, we switched the grammatical role of the characters as a between participants factor, so that, for example, the postman became the object of the matrix clause and the street sweeper became its subject. We created eight lists so that each item appeared in a different condition but only once in each list. The 16 experimental items where interspersed with 64 filler items mostly from unrelated experiments. Each list was randomized once. 32 native speakers of each language participated in the experiment.

Results. Figure 7 shows the results of the questionnaire experiments. All three languages showed a clear preference for the subject for between sentence pronoun resolution for sentences with "puis", "then", and "dann" (all ps < .01). However, whereas English and German participants chose the subject of the matrix clause more often as the antecedent of the pronoun for within sentence pronoun resolution as well, French participants chose reliably more often the object of the matrix clause.

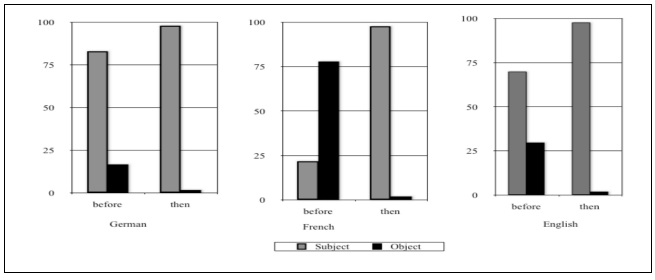


Figure 7: Decisions in % for the subject or the object of the main/first clause as the antecedent of the pronoun.

Discussion

In our experiments, we established the following pattern:

- All three languages show a clear subject preference for between sentence pronoun resolution in sentences like (7) and (9).
- German and English, both show a subject preference for within sentence pronoun resolution (3-6, 7).
- French shows are clear object preference for within sentence pronoun resolution (3-6, 7).

An explanation of the differences between German and French before-sentences could be based on the Gricean Principle of Manner (avoid ambiguity). In French, the temporal clause can be unambiguously related to the subject of the matrix clause using an infinitival construction such as (2). In German, no such alternative construction exists. French listeners or readers might thus apply a Gricean logic taking the object of the matrix clause as the antecedent of the the full pronoun in (1).

A Gricean account is, however, hard to reconcile with the English data: For English, an alternative construction relating the temporal clause to the subject is available as well (2). Still, the full pronoun in (1) consistently shows a clear preference for the subject across experiments. An experience-based account would be fully compatible with the results of the sentences with « before » as can be seen in the small scale corpus study mentioned above (see Figure 8 for a direct comparison of off-line decisions and corpus data).

The Gricean Principle of Manner neither predicts production preferences nor comprehension preferences in English. This finding is very much in line with earlier evidence showing that speakers do not follow the Principle of Quantity (they very often produce more information then necessary in referring expressions, e.g., Pechmann, 1989), neither are they generally cooperative in using unambiguous alternatives for one of the possible interpretations of an unambiguous construction (Ferreira & Dell, 2000). Arnold, Wasow, T., Asudeh, and Alrenga (2004) likewise argue against sentence production as designed to be easily comprehensible for the audience, based on a consistent lack of ambiguity avoidance. The choice of linguistic expressions seems to be more affected by cognitive pressure than by cooperativeness (Wardlow & Ferreira, in press).

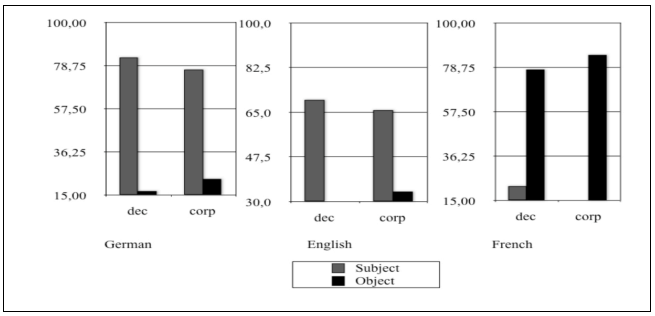


Figure 8: Decisions in % for the subject or the object of the main clause as the antecedent of the pronoun compared to corpus counts

However, we still have to explain why French and English should be different with respect to production

preferences: A reason why French speakers prefer producing an infinitival construction for subject antecedents may be the increased complexity of temporal clauses with « avant que »: The French conjunction "avant que" demands the subjunctive form as do many other conjunctions such as "puisque", "pour que", "bien que", whereas others demand the indicative form, such as "après que", "lorsque", "parce que" and many others. The correct marking of the verb will thus have to be adapted to the respective conjunctions. Using the infinitival form avoids the necessity of checking which verb form to use in the actual utterance.[2] No such checking would be needed for English conjunction-plus-pronoun sentences which consistently demand the indicative form.

The results so far would thus be fully compatible with an approach linking comprehension preferences to production preferences (Cuetos & Mitchell, 1988; Gennari & MacDonalds, 2009, Konieczny, 2000). French speakers prefer using the infinitival form whenever possible, which is the case when the infinitival clause is related to the subject of the matrix sentence. "Avant que" plus pronoun will thus mostly be used in cases where the pronoun is related to a non-subject antecedent. These production preferences will result in the distributions observed in the corpora. Exposure to these distributions will consequently shape preferences in comprehension.

We do, of course, by no means imply that pronoun resolution preferences are based on exposure exclusively. Factors such as information structure, coherence relations and others are most certainly playing a role as well. An interesting question for further research will be, in how far the crosslinguistic differences established in our experiments extend to other conjunctions, and in how far they interact with factors influencing the prominence of antecedents such as first mention, topicality, prominence, and many more (Colonna, Schimke, & Hemforth, 2009; Schimke, Colonna, & Hemforth, 2009). We will also have to extend our research to other languages. Interestingly, European Portuguese provides a combination of alternatives highly comparable to what can be found in French. Recent self-paced reading experiments and questionnaire studies (Baumann, Konieczny, & Hemforth, 2010) show a clear object preference for pronouns in Portuguese constructions parallel to those under investigation in this paper.

References

Ariel, M. (1990) *Accessing Noun-Phrase Antecedents*. Routledge, London.

Arnold, J. E., Wasow, T., Asudeh, A, and Alrenga, P. (2004). Avoiding Attachment Ambiguities: The Role of Constituent ordering. *Journal of Memory and Language.*

Baumann, P., Konieczny, L., & Hemforth, B. (2010). Expecting coreference: the role of alternative constructions. In *Proceedings of the 23d Annual Meeting of the CUNY Conference on Human Sentence Processing*. New York, March 2010.

Colonna, S., Schimke, S., & Hemforth, B. (2009). The role of information structure in pronoun resolution. Talk at the conference on *Linguistic and Psycholinguistic approaches to Text Structuring*, Paris, September 21-23.

Cuetos, F., & D. Mitchell. 1988. Cross-linguistic differences in parsing: Restrictions on the use of the late closure strategy in Spanish. *Cognition,* 3: 73-105.

Ferreira, V. S., & Dell, G. S. (2000). The effect of ambiguity and lexical availability on syntactic and lexical production. *Cognitive Psychology, 40*, 296-340.

Gennari, S. P., and MacDonald, M. C. (2009) Linking production and comprehension processes: The case of relative clauses, *Cognition*, 111, 1-23.

Gernsbacher, M.A. (1990). *Language Comprehension as Structure Building.* Hillsdale NJ: Lawrence Erlbaum.

Givon, T. (1983). Topic continuity in discourse: A quantitative cross-language study. Amsterdam: Benjamins.

Järvikivi, J., van Gompel, R., Hyöna, J., & Bertram, R. (2005). Ambiguous pronoun resolution: Contrasting the first-mention and subject preference accounts. *Psychological Science*, 16, 260-264.

Kehler. A. (2002). , *Coherence, Reference, and the Theory of Grammar,* Stanford: CSLI Publications.

Konieczny, L. (2000). Locality and parsing complexity. *Journal of Psycholinguistic Research, 29-6.* 627-645

Pechmann, T (1989). Incremental speech production and referential overspecification. *Linguistics*,27: 89–110.

Sanders, T. & Noordman, L. (2000). The role of coherence relations and their linguistic markers in text processing. *Discourse Processes*, 29: 37-60.

Schimke, S., Colonna, S., & Hemforth, B. (2009). Discourse prominence and pronoun resolution : Evidence from French. Talk at the conference on *Text and Discourse*, Rotterdam, July 26-28.

Wardlow Lane, L. & Ferreira, V. S. (in press). Speaker-external versus speaker-internal forces on utterance form: Do cognitive demands override threats to referential success? *Journal of Experimental Psychology: Learning, Memory, & Cognition*.

[2]French native speakers are actually not allways fully sure of which form to use. A short questionnaire sent by mail to 20 doctoral students (mostly from the linguistics department) asking for the correct verb to use in sentences like "Le balayeur a appelé le facteur après qu'il _____ rentré à la maison." (The street sweeper called the postman after he _____ gone home.), resulted in 56 % responses using he subjunctive and 44 % using the indicative. Following normative grammar, "après que" does not demand the subjunctive.

Acknowledgements

This work was partially funded by the bilateral research grants "Alliance" and "Procope" attributed to the first three authors. We would like to thank the "Linglunch" participants at Paris Diderot as well as our colleagues from the FRIAS in Freiburg for many helpful discussions.

Designing State-Trace Experiments to Assess the Number of Latent Psychological Variables Underlying Binary Choices

Guy Hawkins (Guy.Hawkins@newcastle.edu.au)
Melissa Prince (Melissa.Prince@newcastle.edu.au)
Scott Brown (Scott.Brown@newcastle.edu.au)
Andrew Heathcote (Andrew.Heathcote@newcastle.edu.au)
School of Psychology, University of Newcastle
University Drive, Callaghan, 2308, NSW Australia

Abstract

State-trace analysis is a non-parametric method that can identify the number of latent variables (dimensionality) required to explain the effect of two or more experimental factors on performance. Heathcote, Brown, and Prince (submitted) recently proposed a Bayes Factor method for estimating the evidence favoring one or more than one latent variable in a state-trace experiment, known as Bayesian Ordinal Analysis of State-Traces (BOAST). We report results from a series of simulations indicating that for larger sample sizes BOAST performs well in identifying dimensionality for single and multiple latent variable models. A method of group analysis convenient for smaller sample sizes is presented with mixed results across experimental designs. We use the simulation results to provide guidance on designing state-trace experiments to maximize the probability of correct classification of dimensionality.

Keywords: State-trace analysis; Bayesian analysis; Bayes Factor; Encompassing prior method; Simulation.

State-Trace Analysis

State-trace analysis (Bamber, 1979), also known as dimensional analysis (Loftus, Oberg, & Dillon, 2004), is a method for determining whether a single latent variable is capable of explaining the joint effect of two experimental factors. Dimensionality is traditionally assessed by testing for an interaction between the two factors. However, interactions can be scale dependent (e.g., distorted by floor or ceiling effects) when response variables are bounded (e.g., accuracy data, see Dunn & Kirsner, 1988; Loftus, 1978). State-trace analysis overcomes this problem by assessing the ordinal relationships between the effects of experimental factors. One factor is comprised of a set of indicator variables, and is referred to as the *state factor*. A second factor is a variable thought to differentially influence performance over levels of the state factor, and is referred to as the *dimension factor*.

State-trace analysis is most easily described by an example. For this purpose we use the *disproportionate face inversion effect* (DFIE), the finding in perceptual and recognition memory studies that stimulus inversion has a more deleterious effect on faces than other mono-oriented stimuli (Valentine, 1988; Yin, 1969). This result has been taken to suggest that faces are encoded along a 'configural' dimension that is not available to mono-oriented non-face stimuli (e.g., houses; Maurer, Le Grand, & Mondloch, 2002). In this example, stimulus type (faces or houses) is the state factor and stimulus orientation (upright or inverted) is the dimension factor, as inversion is thought to differentially affect memory for faces and houses.

State-trace analysis results are shown in a state-trace plot: a scatterplot of the co-variation of performance for the levels of the state factor (e.g., faces or houses). Memory accuracy results for the two levels of the state factor form the two axes of the state-trace plot. Each point on the plot represents a pair of measurements, with a pair of (x, y) coordinates for each level of the dimension factor (e.g., upright and inverted). For our example there would be two coordinate pairs on the state-trace plot, one for upright stimuli and one for inverted stimuli. To infer dimensionality a third variable, referred to as a *trace factor*, is added to the state-trace design. The trace factor sweeps out a set of coordinate pairs for each level of the dimension factor. Levels of the trace factor within each level of the dimension factor are usually connected with a line in the state-trace plot, with each line referred to as a *data trace*.

Latent dimensionality is identified by assessing whether all of the data points in the state-trace plot fall on a single monotonic (i.e., always increasing or always decreasing) function, indicating evidence for a single latent variable. Monotonicity holds if all of the x axis values in a state-trace plot have the same order as the y axis values. Although a monotonic plot is necessary to infer a single latent variable, it is not sufficient: monotonicity cannot be diagnostic unless the data traces overlap on at least one axis. Hence, an assessment of whether data traces overlap is essential to a proper assessment of dimensionality. Similarly, it is important to establish that the trace factor does not itself affect dimensionality, so that results in favor of more than one dimension can be unambiguously attributed to the effect of the dimension factor. This can be checked by determining if the trace factor has a monotonic effect within each level of the dimension factor.

A Bayesian Approach to State-Trace Analysis

Given an observed state-trace plot, where the effects of the underlying latent variable(s) are perturbed by measurement error, how can we determine whether a monotonic curve best describes the data? A number of statistical methods for assessing departures from monotonicity have been suggested (see Loftus et al., 2004; Newell & Dunn, 2008). Recently Heathcote et al. (submitted) proposed a Bayes Factor approach to state-trace model selection, known as Bayesian Ordinal Analysis of State-Traces (BOAST), based on Klugkist, Laudy, and Hoijtink's (2005) encompassing prior method. The encompassing prior method uses Bayes Factors to select

among models defined by inequalities. The advantage of this approach is that it automatically accounts for differences in flexibility amongst models, which is a key issue in state-trace analysis as a one-dimensional model is far less flexible than a multi-dimensional model.

BOAST assumes binomially distributed data (e.g., a binary two-alternative forced choice response used to measure recognition accuracy in the DFIE example), with state-trace models being defined by sets of inequality constraints on binomial probability parameters. For example, we define a 'trace' model, which instantiates the assumption that the trace factor does not change dimensionality, by specifying that the trace factor has a monotonic effect on performance within each level of the dimension factor. This specification implies that, for a trace factor with three levels and an overall increasing effect on accuracy, that accuracy is smaller for the first level of the trace factor than the second level, and smaller for the second level than the third. The trace model is, therefore, an order constrained special case of an 'encompassing' model that places no restrictions on the order of parameters.

When model M_i is an order constrained version of an encompassing model M_k, Bayes Factors can be estimated from prior and posterior samples from the encompassing model (Klugkist, Kato, & Hoijtink, 2005). The proportion of prior ($\hat{\pi}$) and posterior ($\hat{\Pi}$) samples that adhere to the order constraints of the more restricted model M_i are used to estimate a Bayes Factor from the ratio of the two sample counts,

$$BF_{ik} \approx \frac{\hat{\Pi}}{\hat{\pi}}. \quad (1)$$

This Bayes Factor indicates the strength of evidence in favor of M_i over M_k. Intuitively this is the case because it is the ratio of the probability that the model fits the data before the data are observed, which is proportional to the complexity of the model (e.g., the maximally complex encompassing model will always fit any data pattern), to the actual fit of the model to the data. If this ratio is greater than one it indicates that the model fits better than chance.

A set of such Bayes Factors, assuming the same encompassing model, can be used to compare a set of order-restricted models by calculating each models posterior model probability, $p(M_i|D)$, given observed data D. The quantity $p(M_i|D)$ is the probability that model M_i is the 'true' (data generating) model, on the assumption that one model in the set is the true model. Model selection based on $p(M_i|D)$ can also be justified on other grounds, even when the set does not contain the true model (e.g., it selects the model that is most likely to minimize a measure of error in predicting new data), so we refer to it simply as a method of selecting the 'best' model. For a set of models M_i, $1 \ldots m$ that are assumed to have a probability p_i of being the best model prior to observing the data, the posterior model probability for M_i is:

$$p(M_i|D) = \frac{p_i \times BF_{ik}}{\sum_{j=1}^{m} p_j \times BF_{jk}} \quad (2)$$

for any $k = 1 \ldots m$ which includes i. Throughout we assume each model is equally likely to be the best model before observing the data.

Our aim here is to assess, via simulation, how often BOAST analysis selects the correct number of latent variables, either one or more than one. We begin by simulating an individual participant analysis. We then examine a method of aggregating participant results to select the best characterization of dimensionality for a group of participants.

Simulations

Figure 1 shows state-trace data consistent with a single latent variable model (1D) and a two latent variable model (2D). In both cases the trace factor has a clear monotonic effect on performance; that is, as the level of the trace factor increases so too does the dependent variable. The two models also both exhibit moderate and equal data trace overlap. These two patterns were used to generate simulated data (by using their coordinates to specify binomial probability parameters) and we will refer to them as the 1D and 2D models.

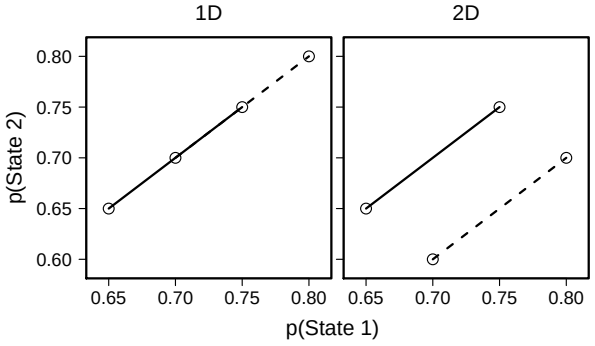


Figure 1: The two models on which simulations were based. p(State 1) and p(State 2) refer to the proportion of correct responses for the first and second level of the state factor, respectively. The two lines on each plot represent data traces, one for each level of the dimension factor. The solid lines are identical for both models, and the dashed line for the 2D model is the same as the dashed line for the 1D model but transposed downward by 0.1.

We next elaborated the 1D and 2D models with 2 trace levels shown in Figure 1, which we call the $T2$ designs, by creating variants with three and four trace levels, $T3$ and $T4$ designs respectively. In $T2$ designs the two levels of the trace factor provided data for the end points of the data traces. For $T3$ and $T4$ designs the additional levels were evenly spaced between the end points of each data trace. One purpose of these simulations was to provide guidance on experimental design in terms of the trade-off between number of trials contributing to the estimates of each point in the state-trace plot and the number of levels in the trace factor. For a fixed sample size (number of trials) there is a trade-off between these two factors, with more trace levels resulting in fewer trials

per point. For each model and each T we explored 6 total trial numbers (n) with total n conserved across each T at 192, 384, 768, 1536, 3072 and 6144. For example, a model with $n = 192$ had 24 observations per coordinate of each point for $T2$, 16 observations for $T3$, and 12 observations for $T4$. In total we performed 36 simulations (2 models $\times$ 3 trace levels $\times$ 6 sample sizes). For each simulation 1000 Monte Carlo replicates were sampled from binomial distributions with parameters determined by the design and model. Sufficient posterior samples were obtained so that posterior proportions of monotonic samples had 90% credible intervals less than 0.025; prior proportions were determined analytically assuming a uniform prior (see Heathcote et al., submitted, for details).

BOAST Results

For each simulation we estimated Bayes Factors to test four mutually exclusive models, which we refer to as the non-trace (NT), no-overlap (NO), unidimensional (UD) and multidimensional (MD) models. Together these models account for all possible orders (i.e., together they constitute the encompassing model). Posterior model probabilities were calculated for each Monte Carlo replicate for each model by dividing each Bayes Factor by the sum of all four Bayes Factors (i.e., Equation 2), which we refer to as p(NT), p(NO), p(UD) and p(MD), respectively. Figure 2 illustrates results in terms of the proportion of comparisons selecting one of the four models (i.e., where the models posterior probability was greatest amongst the set of four models). Figure 2 can be interpreted by comparing the height of corresponding points across the panels in each row. In particular, the 'highest' point indicates which of the four models is most often supported.

The Trace Model

An important first check in any state-trace analysis is to determine whether the trace model is supported. For example, we described study duration as a possible trace factor. In this case the trace model indicates that accuracy increased as study durations became longer for both levels of the state and dimension factors. In contrast, support for the non-trace model indicates that the order dictated by the trace factor was violated. Even when the trace model is the data generating model, measurement noise can cause violations of the trace model (i.e., support for the non-trace model) to arise more frequently when differences between levels of the study duration factor produce only small changes in accuracy. Support for the non-trace model clouds any conclusions about underlying dimensionality of the state factor since the effects of the dimension and trace factors are confounded, and can suggest that the experimental design needs to be improved by using more widely spaced trace factor levels.

The non-trace model results are shown in the left column of Figure 2. The figure demonstrates a number of key points. As expected, evidence for the trace model is similar across both 1D and 2D simulations, since the trace factor should have a consistent effect irrespective of underlying dimensionality. Secondly, as total sample size increases the lines always approach zero, indicating consistent selection of the trace model. That is, BOAST recovers the trace model with increasing reliability as measurement error decreases due to an increase in sample size. Finally, the probability of selecting the non-trace model approached zero with lower total trials for $T2$ compared to $T3$ and $T4$. As seen in Figure 2, selection is approximately zero for $T2$ at $n = 768$, whereas this increased to $n = 1536$ for $T3$ and $T4$ in the 1D model, and to $n = 3072$ for $T4$ in the 2D model. Thus, for smaller n, the trace model had a greater chance of being supported in $T2$ designs compared to $T3$ and $T4$ designs. This occurs because the combination of a smaller sample size (and hence greater measurement noise) and closer spacing between results for adjacent trace levels as T increases makes a violation of monotonicity within a data trace more likely.

The No-Overlap Model

When the trace model holds it implies that one of the three remaining models best describes the data, as they are each trace models. A monotonic state-trace plot is a special case of the trace model where all data points have the same ordering for both levels of the state factor. A non-overlapping monotonic plot is a case where data traces for both levels of the dimension factor do not cross over at any point along either axis of the state-trace plot. In this case, monotonicity is not diagnostic of dimensionality, as both one-dimensional and multi-dimensional data generating models produce monotonic state-trace plots when there is a failure of data trace overlap. Hence, an important second check in a state-trace analysis is to determine whether the no-overlap model holds.

Results for the no-overlap model differed between the 1D and 2D data generating models. The 1D simulations results generally give some support for the no-overlap model, which is perhaps not surprising given the the 1D model produces monotonic data. Of more concern is the fact that this support was inconsistent as a function of sample size, n, for $T4$ and to a lesser degree for $T3$. That is, support for the no-overlap model initially increased with n, but then decreased, from $n = 1536$ for the $T4$ design and from $n = 768$ for the $T3$ design. In contrast, the no-overlap model consistently received little support across all T and n in the 2D simulations. Overall, these results suggest that when there is in fact trace overlap in a one-dimensional data generating model, the no-overlap model is more often rejected in designs with fewer trace levels.

The Unidimensional and Multidimensional Models

For both data generating models the unidimensional and multidimensional posterior model probabilities provided support for the true model dimensionality. For the 1D case support for the unidimensional model (middle right column of Figure 2) increased with sample size, but the level of support was smaller for larger T. For the 2D case support for the multidimensional model (right column of Figure 2) also in-

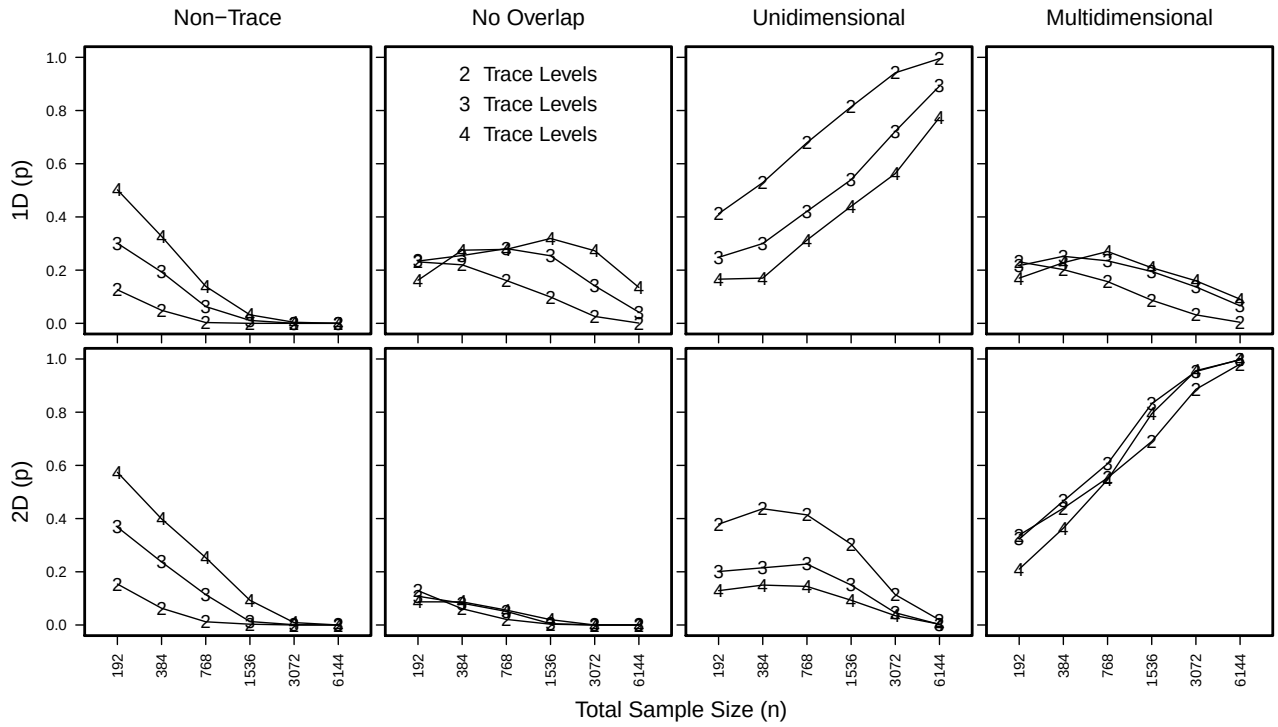


Figure 2: Model selection results for both data generating models, type of comparison, number of trace levels, T, and number of 'trials', n. Columns correspond to each of the mutually exclusive models being tested and rows to the type of the data generating model. On each plot the x axis represents the six levels of n and the y axis represents the proportion of simulations in which posterior model probability favored the model specified for each column. The lines group designs with the same T.

creased with sample size. In contrast to the 1D case, the level of support was similar for all T, although it was slightly less for the $T4$ design for smaller n (likely reflecting the larger level of support for the non-trace model) and slightly less for the $T2$ design for the second and third largest value of n, with all T designs perfectly selecting the true model for the largest sample size. Across the 1D and 2D data generating models support for the wrong dimensionality was generally low and decreased with sample size, although there was some inconsistency for the three smallest sample sizes.

Overall, the results of the simulation study indicate that accurate results for all comparisons can only be guaranteed for quite large sample sizes. This indicates that analysis of individual participant data may not produce clear results in applications where it is not possible to measure performance on a large number of trials for each individual. In such situations it would be desirable to have a method of combining results over participants in a way that improves correct identification at the group level. In the next section we extend the analysis of our simulation results to assess the performance of one such method suggested by Heathcote et al. (submitted), the group Bayes Factor.

Group Bayes Factors

A Bayes Factor for a group of participants, assuming each participant contributes independent evidence, can be obtained by taking the product of each participants Bayes Factor. Hence, a group Bayes Factor for a model M_i is given by $GBF_i = \prod_{j=1}^{N} BF_{ij}$, where N is the number of subjects. Group Bayes Factors can then be combined to obtain a posterior model probability for model M_i at the group level. Again we assume each model is equally likely to be the best model before observing the data, and so:

$$p(M_i|D) = \frac{GBF_i}{\sum_{k=1}^{m} GBF_k} \tag{3}$$

for a set of $k = 1 \ldots m$ models that includes model i.

We examined the utility of group Bayes Factors using the simulations from the previous section. For each simulation we sampled with replacement (i.e., resampled) sets of individual Bayes Factors from the 1000 available. The sets were of sizes (N) 8, 16 and 32, representing experiments with different numbers of participants. These N's cross with total trials n in a balanced manner. For example, a set of $N = 32$ with $n = 192$ trials provides results from a total of 6144 trials, equivalent to the set $N = 16$ with $n = 384$ trials, and $N = 8$ with $n = 768$ trials. The resampling procedure was repeated 500 times for each possible grouping: two data generating models (1D, 2D), with three trace levels ($T2$, $T3$, $T4$), three total trial sizes (n = 192, 384, 768), and three participant sample sizes (N = 8, 16, 32), for each of the four comparisons (non-trace, no-overlap, unidimensional, multidimensional), a

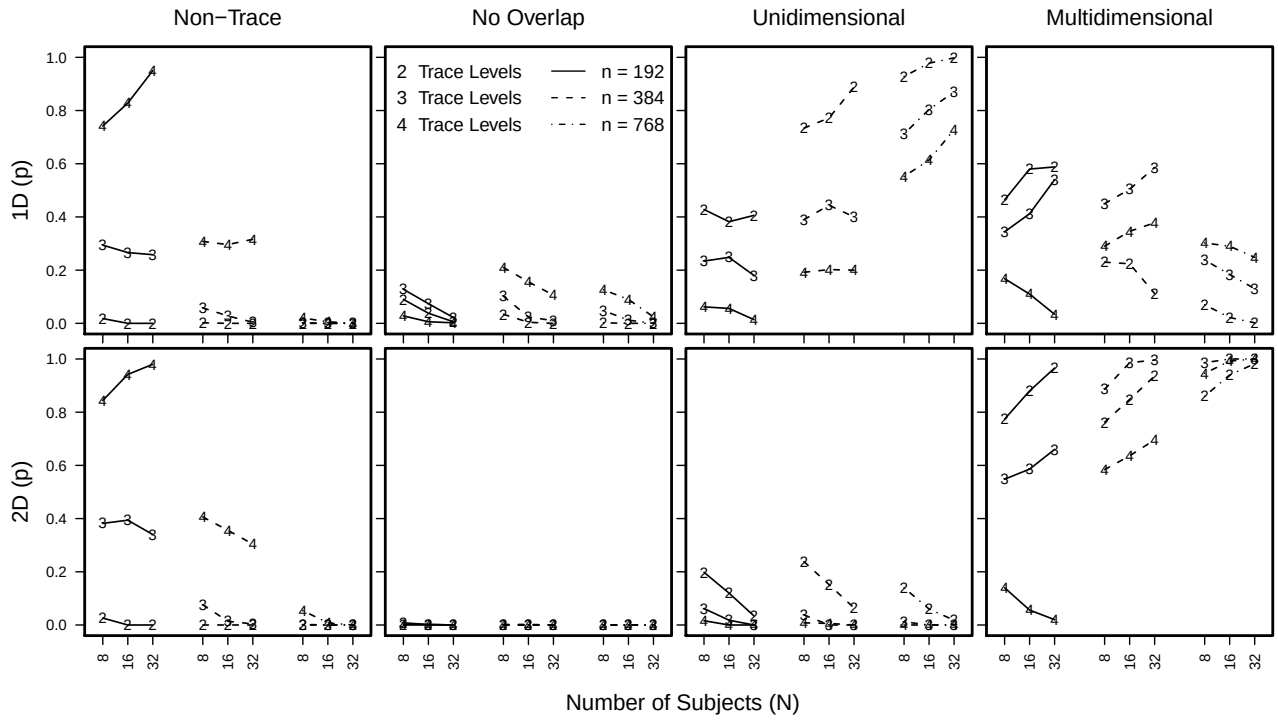


Figure 3: Group level results for the 216 comparisons. The rows and columns represent the same data generating models and comparisons as Figure 2. On each plot, the x axis represents the three levels of N that were resampled for each of n = 192, 384, 768, and the y axis represents the proportion of cases in which the posterior model probability at the group level favored the model specified for each column.

total of 216 combinations ($3^3 \times$ 2 models $\times$ 4 comparisons). For each of the 500 repetitions of the 216 combinations we estimated group Bayes Factors, and then calculated the proportion of comparisons selecting one of the four models (i.e., where the models posterior probability was greatest amongst the set of four models), with results shown in Figure 3.

For the no-overlap model the group Bayes Factors results were much the same as for the individual analysis, except that the inconsistent effect of sample size for the individual analysis of the 1D data generating model disappeared in the group analysis. For the trace model performance was excellent when $n = 768$ but the wrong (non-trace) model received increasing support when there were fewer observations per participant for all but the $T2$ design. These problems with the trace model caused corresponding failures to identify the correct dimensionality for lower values of n, whereas for $n = 768$ performance in identifying dimensionality was similar to that of the largest samples sizes in the individual analysis. In particular, the 2D data generating model was almost perfectly identified, but with higher T designs being slightly better, whereas performance in classifying the 1D data generating model was very good for $T2$ designs but decreased markedly for the $T3$ and $T4$ designs.

Conclusions

We aimed to investigate the ability of BOAST analysis to identify latent dimensionality. The results of individual participant data indicated that large sample sizes produced strong support for the correct outcome for both 1D and 2D data generating models across designs with two, three and four levels in the trace factor. Classification for the 1D data generating model was most reliable in designs with two trace levels, whereas the opposite tendency was evident for the 2D data generating model; dimensionality assessment was more accurate with larger numbers of trace levels. Overall these results indicate that a design with three trace levels provides the best compromise for accurate diagnosis of both single and multiple latent variable data generating models.

We also explored a group analysis procedure that is advantageous where it is practically difficult to obtain a large number of responses from each individual participant, such as in cases where the number of available stimuli is limited, but where larger numbers of participants are available. Generally, this method was found to be very effective in identifying the 2D data generating model. However, our results indicate that it should be used with caution as it could be biased against detecting cases in which only one latent variable is present in certain experimental designs. When each participant contributed a smaller number of responses (192 or 384) results could be inaccurate even for the largest number of participants (32). For 768 observations per participant performance was more accurate and improved with group size for the 1D

data generating model. In contrast to the individual participant results, the group level analyses indicate that designs with two levels in the trace factor produce the best compromise of most accurate classification across number of trials per participant and different numbers of participants for both 1D and 2D data generating models. However, these results should be used with some caution given the three and four trace level designs demonstrated a large proportion of cases supporting the non-trace model (possibly due to the small experimental effects of the trace factor in these larger trace level designs), which had strong consequences for the correct classification of dimensionality.

Our individual and group analyses indicate that the ideal number of trace levels in a state-trace experiment is dependent on the intended approach to data collection. If only a small number of trials per participant are obtainable it seems wiser to use a trace factor with few levels so as to maximise data per point, and then combine across participants with group Bayes Factors. In contrast, if many trials per participant can be obtained, correct classification of dimensionality is possible with a three level trace factor through individual participant analysis, which confers additional benefits such as the exploration of individual differences in performance.

In summary, these results indicate that the success of BOAST analysis, and likely any state-trace analysis method, depends strongly on the particular model producing the state-trace plot. This highlights a caveat on our group analysis, which assumes all participants have an identical underlying model (rather than just having the same dimensionality but possibly different magnitudes of the effects of experimental factors). As well as being unrealistic, this assumption likely magnifies the effects of a particular data pattern. In ongoing research we will simulate groups of participants that vary in the effects of experimental factors (while maintaining a consistent dimensionality) in order to check the generality of the group analysis results reported here.

Acknowledgments

We acknowledge support from the Keats Endowment Research Fund.

References

Bamber, D. (1979). State-trace analysis: A method of testing simple theories of causation. *Journal of Mathematical Psychology*, *19*, 137–181.

Dunn, J. C., & Kirsner, K. (1988). Discovering functionally independent mental processes: The principle of reversed association. *Psychological Review*, *95*, 91–101.

Heathcote, A., Brown, S., & Prince, M. (submitted). The design and analysis of state-trace experiments. *Psychological Methods*.

Klugkist, I., Kato, B., & Hoijtink, H. (2005). Bayesian model selection using encompassing priors. *Statistica Neerlandica*, *59*, 57–69.

Klugkist, I., Laudy, O., & Hoijtink, H. (2005). Inequality constrained analysis of variance: A Bayesian approach. *Psychological Methods*, *10*, 477–493.

Loftus, G. R. (1978). On interpretation of interactions. *Memory & Cognition*, *6*, 312–319.

Loftus, G. R., Oberg, M. A., & Dillon, A. M. (2004). Linear theory, dimensional theory, and the face–inversion effect. *Psychological Review*, *111*, 835–863.

Maurer, D., Le Grand, R., & Mondloch, C. J. (2002). The many faces of configural processing. *Trends in Cognitive Sciences*, *6*, 255–260.

Newell, B. R., & Dunn, J. C. (2008). Dimensions in data: Testing psychological models using state-trace analysis. *Trends in Cognitive Sciences*, *12*, 285–290.

Valentine, T. (1988). Upside-down faces: A review of the effect of inversion upon face recognition. *British Journal of Psychology*, *79*, 471–491.

Yin, R. K. (1969). Looking at upside-down faces. *Journal of Experimental Psychology*, *81*, 141–145.

Simulating individual differences in language ability and genetic differences in *FOXP2* using a neural network model of the SRT task

Joseph C. Toscano (joseph-toscano@uiowa.edu)
Dept. of Psychology, University of Iowa, Iowa City, IA 52242 USA

Kathryn L. Mueller (kathryn-mueller@uiowa.edu)
Dept. of Communication Sciences and Disorders, University of Iowa, Iowa City, IA 52242 USA

Bob McMurray (bob-mcmurray@uiowa.edu)
Dept. of Psychology, University of Iowa, Iowa City, IA 52242 USA

J. Bruce Tomblin (j-tomblin@uiowa.edu)
Dept. of Communication Sciences and Disorders, University of Iowa, Iowa City, IA 52242 USA

Abstract

Recent work has shown that individual differences in language development are related to differences in procedural learning, as measured by the serial reaction time (SRT) task. Performance on this task has also been shown to be associated with common genetic variants in *FOXP2*. To investigate what these differences can tell us about the functional properties of language processing, we present a computational model of the SRT task. We varied parameters in the model to observe their effects on performance in the task. We found that the combined effect of several model parameters produced changes in the learning trajectory that were similar to those observed behaviorally.

Keywords: language processing; specific language impairment; *FOXP2*; procedural learning; serial reaction time task; computational modeling; simple recurrent networks

Introduction

The mechanisms that underlie language use emerge over the course of development through the integration of multiple biological and environmental factors (Elman et al., 1997). Much previous research has focused on whether these mechanisms are language-specific or domain-general (Christiansen & Chater, 2008). Regardless of which is the case, we must specify how different factors interact to give rise to language.

One way to study the mechanisms involved in language is to look at individual differences in language ability. Recently, the use of molecular genetics has emerged as a tool for investigating these differences. However, the use of genetics to study complex cognitive processes, like language, presents a challenge: how do we address questions regarding the role of genes when they are so far removed from language processing? Similarly, how do we assess the role of individual genes when it is unlikely that there is a one-to-one correspondence between genes and specific characteristics of language?

As a first step, we need a way to observe the effects of functional properties of language processing on behavior. Computational models offer a tool for doing this. The units in a neural network model, for instance, correspond to functional (rather than structural) units in the system. Thus, computational models may be useful for examining how genetic factors relate to the functional organization of cognitive systems.

The aim of the current paper is to investigate the relationship between individual differences (both differences in language ability and genetic differences) and functional properties of language processing using a computational model of the serial reaction time (SRT) task. The SRT task measures participants' ability to learn pattern sequences. Variation in performance on the SRT task has been associated with both language ability (Tomblin, Mainela-Arnold, & Zhang, 2007b) and genetic differences (Tomblin, Christiansen, Bjork, Iyengar, & Murray, 2007a). Given this, and the fact that sequence processing is a critical component of language use, this task provides a useful paradigm for studying these relationships.

Individual differences in language abilities

One area in which differences in language ability have been extensively studied is specific language impairment (SLI). SLI is a relatively common developmental disorder characterized by difficulty acquiring language in the absence of gross cognitive or sensory impairments, and despite adequate experience and educational opportunities (Tomblin, Records, & Zhang, 1996). Typically, research criteria for SLI classification require that the individual falls 1.15 SD below the mean on a range of standardized assessments of language while falling in the normal range for non-verbal intelligence (Tomblin et al., 1996).

Children with SLI have deficits in various language abilities, such as mopho-syntactic processing, phonological processing, word learning, and spoken word recognition (Leonard, 1998; McGregor, Newman, Reilly, & Capone, 2002; McMurray, Samelson, Lee, & Tomblin, 2010). In many ways, these children demonstrate language abilities associated with typically developing younger peers. They have smaller vocabularies, use shorter, simpler syntactical constructions, and make more morphological errors than would be expected for children their age (McGregor, Friedman, Reilly, & Newman, 2002).

A range of possible hypotheses for SLI have been proposed, and include deficits in temporal-perceptual processing, generalized slowing, problems with phonological processing, and deficits in working memory (Bishop, North, & Donlan,

1996). Thus, the underlying causes remain unclear, though it is likely that SLI is multiply determined.

Genetic factors and language

Genetics is now commonly employed as a tool for investigating differences in language development. Initial molecular studies centered on the KE family, a multigenerational pedigree that appears to show an autosomal dominant pattern of language impairment (Hurst, Baraitser, Auger, Graham, & Norell, 1990). Affected individuals have been characterized as having apraxia of speech, as well as expressive and receptive language problems (Vargha-Khadem, Watkin, Alcock, Fletcher, & Passingham, 1998). They also have a rare genetic mutation in the *FOXP2* (forkhead box P2) gene (Lai, Fisher, Hurst, Vargha-Khadem, & Monaco, 2001). More recently, Mueller, Bjork, Tomblin, and Murray (in preperation) investigated the role of more common genetic variants in *FOXP2*. These variants were single nucleotide polymorphisms (SNPs), which represent differences in a single base pair in the genome. They examined multiple SNPs in a population with a range of language abilities and found an association between SNPs in the promoter region and language ability as a discrete phenotype. This suggests that these common variants of *FOXP2* also play a role in language development.

FOXP2 is expressed in multiple species as well as several different organs, including the lungs and brain (Shu et al., 2007; Fujita et al., 2008). This has led some to argue that the link between *FOXP2* and language is weak. However, the fact that *FOXP2* is neither species- nor domain-specific means it is likely to play a role in multiple cognitive processes. In addition, since *FOXP2* is a transcription factor (i.e., encodes a regulatory protein that affects gene expression), it is possible to identify other elements of the gene pathway (and therefore the systems) in which it exists (Vernes et al., 2008).

A more general role for *FOXP2* fits with the hypothesis that language itself is shaped by domain-general processes (Christiansen & Chater, 2008). Statistical learning plays an important role in language acquisition (Saffran, Aslin, & Newport, 1996), and it is closely related to procedural learning (Perruchet & Pacton, 2006). *FOXP2* remains a candidate gene involved in language because of its association with procedural learning and the basal ganglia (Enard et al., 2009).

Procedural learning and the SRT task

Given the links between language ability, *FOXP2*, and procedural learning, researchers have examined sequence learning to better understand these relationships and mechanisms associated with language. The SRT task is a sequence learning task designed to measure participants' ability to implicitly learn sequences. Participants are presented with blocks of trials that are either random or repeat in a particular sequence. As sequence processing is fundamental to language and statistical learning provides a useful mechanism for learning language (Saffran et al., 1996), this task allows us to measure some of the key functional properties of language.

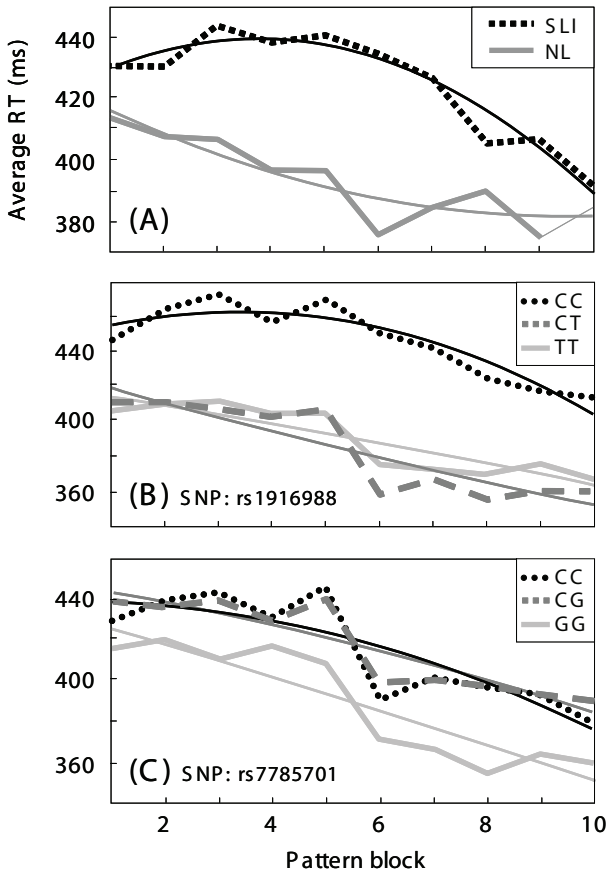


Figure 1: Behavioral data for pattern trials in the SRT task. (A) Data from Tomblin et al. (2007b) comparing SLI and NL groups. (B & C) Data from Tomblin et al. (2007a) for individuals with different genotypes of SNPs rs1916988 and rs7785701.

Tomblin et al. (2007b) used an SRT task to examine differences between children with normal language (NL) and children with SLI. In their task, participants were shown four boxes on a computer screen. On each trial, a picture of a cartoon creature appeared in one of the boxes, and the participant's task was to choose the box containing the picture as quickly as possible.

For the first 100 trials, stimuli were presented randomly. Then, 200 trials were presented in which the sequence [1, 3, 2, 4, 4, 2, 3, 4, 4, 2, 4] was repeated (pattern trials). Finally, 100 additional random trials were presented. Participants were not informed which trials were random and which were pattern trials during the course of the experiment. The experiment was divided into blocks of 20 trials each for data analysis (blocks 1-5 were the first set of random trials, blocks 6-15 were pattern trials, and 17-20 were random trials).

Tomblin et al. found that, overall, the SLI group had longer RTs than the NL group (Fig. 1A). During the pattern trials, performance of both groups improved, indicating that they learned something about the sequence. However, the learn-

ing trajectory differed for the two groups. For the NL group, RT decreased rapidly after the first few blocks of pattern trials and then leveled off. In contrast, for the SLI children, RT remained flat (or increased slightly) during the first few pattern blocks before decreasing. The difference between these two learning trajectories can be approximated by a quadratic function (small differences between the two groups at the first and last blocks; large differences in the middle blocks).

In another study, Tomblin et al. (2007a) examined the relationship between multiple SNPs and performance in this SRT task. They found that the CC genotype of SNP rs1916988 (Fig. 1B) and the CC genotype of SNP rs7785701 (Fig. 1C) were associated with slower RTs over the course of the pattern trials. The CC genotype of SNP rs1916988) was also associated with a learning curve that was similar to the SLI children.

These results suggest that both language impairment and genetic variation in *FOXP2* have similar effects on performance in the SRT task. Given previous work showing a link between *FOXP2* and language, these effects may be related to common functional differences evident in language impairment and some variants of *FOXP2*.

Computational model

We used a neural network to examine whether some of the functional properties of procedural learning are related to the differences observed with human participants. In particular, we would like to capture the difference in the shape of the learning trajectory observed between some of the fast RT groups (children with NL [Fig. 1A] and the CC and CT genotypes of SNP rs1916988 [Fig. 1B]) and slow RT groups (SLI children and the CC genotype of that SNP). By exploring the parameter space of the model, we can determine which functional properties are associated with these differences in the learning trajectories.

Model architecture

The model is a simple recurrent network (SRN; Elman, 1990; c.f. Misyak, Christiansen, & Tomblin, 2009, for an adaptation to the SRT task). The network has three layers: an input layer, an output layer, and a hidden layer with recurrent connections. The input and output layers each have four units (corresponding to the four possible stimulus locations). The hidden layer's recurrent connections provide it with information about the state of the hidden units on the previous trial (context units). This allows the network to learn sequences, like those in the pattern trials of the SRT task. Connection weights are updated using backpropagation (Rumelhart, Hinton, & Williams, 1986). Logistic activation functions are used for the hidden and output units.

Simulation procedure

The network was trained on a task based on the one used by Tomblin et al. (2007b). On each trial, a stimulus was presented to the network by activating a particular input unit and setting the rest to zero, and activation flowed to the output units. Luce choice ratios were computed by dividing each output unit's activation by the total activation. These values were then used to compute an RT for the network according to the equation

$$RT = \frac{1}{C - \frac{\sum I}{n-1}} \quad (1)$$

where C is the activation of the correct output unit, I is the activation of each of the three other output units, and n is the number of output units (four for these simulations). This gives an estimate that is analogous to RT; a lower value corresponds to a lower RT in the SRT task. Thus, when one unit is significantly more active than all the others (i.e., the network is confident in a single response) the RT will be low. When all the units are similarly active (the network is unsure what the response is) the RT will be high.

The correct unit on each trial is the output unit that corresponds to the one that was activated at the input layer. This corresponds to the SRT task in which participants respond by selecting the location containing the stimulus.

For the first 100 trials, a random location was chosen and presented as input. Then, for 200 trials, the sequence [1, 3, 2, 4, 4, 2, 3, 4, 4, 2, 4] was repeated. Finally, an additional 100 random trials were presented. Only trials on which the correct output unit had the highest activation were included in the analysis. The entire simulation run was divided into 20 blocks of 20 trials.

Simulation 1

In the first simulation, we varied several parameters individually to gauge their effect on performance in the SRT task: *context strength*, *input strength*, *learning rate*, *number of hidden units*, and *temperature*.

Context strength determines the strength of the connections from the hidden to context units (i.e., hidden unit activations are multiplied by this amount when setting context unit activations). A lower context strength may have an effect on the network's ability to learn sequences, which could influence learning in the SRT task.

Input strength controls the fidelity of the stimulus presented to the network. The input unit corresponding to the chosen location is set to the value of the input strength and the others are set to zero. A lower input strength makes the stimulus location less distinct from the others.

Learning rate is the value that the weight change term is multiplied by each time the weights are updated. Models with lower learning rates require more trials to learn the task, but may have more stable learning. This could affect the network's ability to learn over the course of the pattern trials.

Number of hidden units affects the amount of information the network can hold about the sequence. If the network has too few, its ability to encode the sequence will be impaired.

Temperature corresponds to the temperature parameter of the logistic activation function. This activation function constrains the hidden and output units to have activations between zero and one. A higher temperature makes the logistic

more nonlinear. Thus, if the correct output unit has the highest activation, a high temperature parameter will make this value more distinct from the values of the incorrect units, resulting in a lower model RT. The temperature parameters for the hidden and output units were varied separately.

Five hundred repetitions of each condition were run.

Results

The network was able to learn the SRT task and showed an overall learning trajectory similar to the ones observed in the behavioral data. The network's performance improved over the course of the simulation and was faster during the pattern trials than the random trials.

Fig. 2 shows the performance of the model on the SRT task for different values of each parameter. A range of values for the parameters were tested to find a set that produced responses similar to those observed for the fast RT groups in the behavioral data. Each parameter was then varied individually, holding the others constant at those values. For example, in Fig. 2A, *context strength* was varied. The other parameters were held constant for both *context strength* conditions at the baseline values (i.e., *learning rate* = 0.10, *hidden units* = 12, *input strength* = 1.0, *hidden unit temperature* = 1.0, *output unit temperature* = 1.0).

Context strength (Fig. 2A) had very little effect on on the network's RT. This suggests that the network can still perform the task with limited information from the previous trial.

Input strength (Fig. 2B) had an effect on overall RT and an effect on the shape of the learning trajectory. Models with a lower input strength showed a small increase in RT at the beginning of the pattern trials, but this did not persist to the middle blocks.

Learning rate (Fig. 2C) also had an effect on the shape of the learning trajectory. This was due to the fact that the network initially shows an increase in RT at the beginning of training. By decreasing the learning rate, this increase was pushed forward in time into the pattern trials. Thus, one reason that some groups show an increase during the pattern trials in the SRT task might be that they are still in this initial learning phase.

Number of hidden units (Fig 2D) had an effect similar to *input strength*. Fewer hidden units resulted in longer overall RTs and a small increase at the beginning of the pattern trials.

Temperature (Figs. 2E & 2F) had an effect on the overall RT at the beginning of the pattern trials, but did not capture the change in the shape of the learning curve.

Discussion

Several parameters produced changes in the network's median RT and learning trajectory. Changes in *input strength*, *learning rate*, and *number of hidden units* can account for some of the changes in the shape of the learning trajectory observed behaviorally. As discussed above, however, specific SNPs and individual differences in language ability are likely to have multiple functional effects. Thus, we may find a better

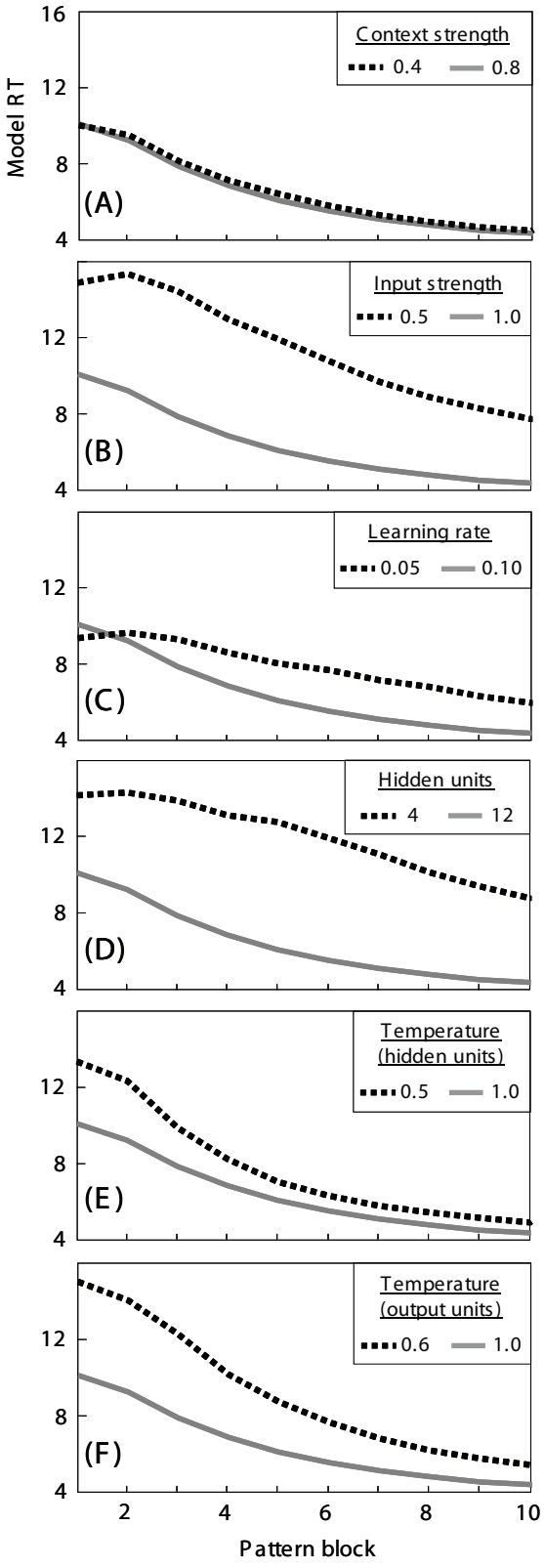


Figure 2: Results of Simulation 1. For each simulation, the set of parameters producing effects similar to those seen in the fast RT groups was used as a baseline (solid lines in figures), and individual parameters were varied (dashed lines).

fit to the behavioral data by examining the combined effects of multiple parameters. This was done in Simulation 2.

Simulation 2

In the second set of simulations, we varied multiple parameters in the model simultaneously, allowing us to explore the parameter space of the network further. Five values were tested for the number of hidden units, and four were tested for each other parameter, yielding a total of 5,120 combinations. The simulation procedure was the same as Simulation 1, except that 50 repetitions of each combination were run.

Results

In order to determine which parameter sets reflected the fast and slow RT groups in the behavioral data, pairwise comparisons were made and the difference scores were fit to quadratic functions (the pattern of the differences in the learning trajectories). Thus, for each comparison there was a set of parameters corresponding to the slow RT groups and a set corresponding to the fast RT groups.

Several pre-processing criteria were used to exclude sets that did not show correct performance on the SRT task (i.e., better performance over the course of the pattern trials) and comparisons that would not yield a pattern consistent with the difference between groups in the behavioral data (i.e., quadratic). The remaining pairs were then fit to quadratic functions using the least squares method, and R^2 was used to determine the goodness of fit.

R^2 values greater than 0.9 were found for 0.47% of the pairs. To determine which parameters drove the effect, we computed the mean parameter values for the slow and fast RT groups for these pairs. The mean values for each parameter for the two groups are shown in Table 1. Some parameters did not differ between the groups, whereas others differed greatly. We found that the parameters in Simulation 1 that produced changes in the learning trajectory (*learning rate*, *number of hidden units*, and *input strength*) had similar effects when varied in conjunction with *temperature*. Fig. 3A shows the responses of the model when these parameters are varied simultaneously.

Adjusting the parameters by hand allowed us to distill the set of parameters down to two, *learning rate* and *temperature*, that accounted for the difference in learning trajectories for the first half of the pattern trials, but not the second half (the slow RT model did not reach the same RT by the end of the pattern trials). Fig. 3B shows the responses of the model when these parameters are varied together.

Table 1: Simulation 2 results.

Parameter	Slow RT	Fast RT
Context strength	0.48	0.53
Input strength	0.83	0.90
Hidden units	5.9	9.5
Learning rate	0.13	0.18
Temperature (hidden)	0.39	0.52
Temperature (output)	0.38	0.44

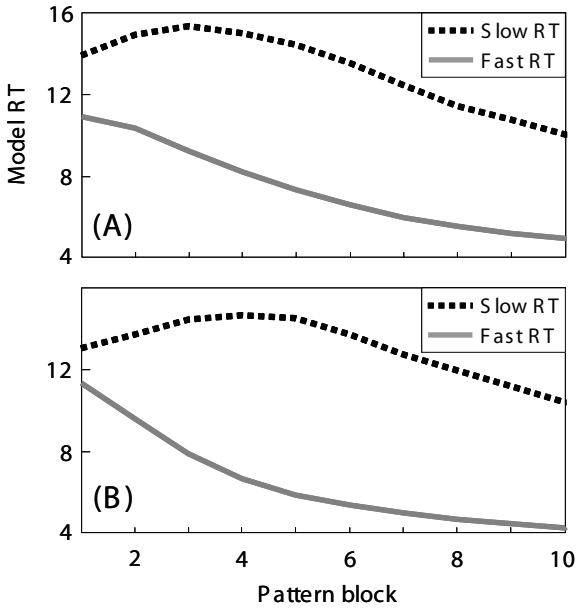


Figure 3: Results of Simulation 2. (**A**) Responses of the model when *learning rate*, *input strength*, *number of hidden units*, and *temperature* are varied simultaneously. (B) Responses when *learning rate* and *temperature* are varied simultaneously. Five hundred repetitions of each condition were run to produce the figures.

Discussion

The results of this simulation show that the combined effects of several parameters together can better approximate the difference in learning trajectories. This suggests that this approach can be used to determine which combinations of parameters mirror the behavioral data. Additional exploration of the parameter space (i.e., testing a larger range of values) may allow us to find a better fit.

General discussion

The results of these simulations suggest that several functional aspects of sequence processing contribute to the differences in SRT performance observed behaviorally and that by examining multiple factors at the same time, we can get a better estimate of the effects of language impairment and genetic variation. This fits with the notion that genetic differences are likely to have multiple functional consequences.

Recently, McMurray et al. (2010) used a similar approach to determine which parameters in TRACE (McClelland & Elman, 1986) corresponded to differences between NL and SLI children in a spoken word recognition task. They found that variation in the network's decay parameter produced differences similar to those between the SLI and NL groups. This parameter is related to competition. In the SRN used here,

the *temperature* parameter corresponds to competition (e.g., a lower temperature parameter for the output unit activation function leads to greater activation for the competitor units). Thus, these two sets of of simulations, modeling different tasks with different networks, provide converging evidence that competition between internal representations may be a critical mechanism in language processing that produces differences between NL and SLI children.

The simulations presented here provide a first step towards assessing the role of genetic variation and language ability in procedural learning, and they suggest several functional properties that may be influenced by these differences. More broadly, they show that exploring the parameter space of a computational model may offer an approach to studying the effects of genetic factors on cognitive systems.

Acknowledgments

We would like to thank Morten Christiansen and Jennifer Misyak for theoretical discussion of these topics and Cheyenne Munson for help preparing the figures.

References

Bishop, D. V. M., North, T., & Donlan, C. (1996). Nonword repetition as a behavioural marker for inherited language impairment: evidence from a twin study. *J Child Psychol Psychiatry, 37*, 391-403.

Christiansen, M. H., & Chater, N. (2008). Langauge as shaped by the brain. *Behav Brain Sci, 31*, 489-509.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science, 14*, 179-211.

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1997). *Rethinking Innateness: A Connectionist Perspective on Development*. Cambridge, MA: MIT Press.

Enard, W., Gehre, S., Hammerschmidt, K., Holter, S. M., Blass, T., Somel, M., et al. (2009). A humanized version of Foxp2 affects cortico-basal ganglia circuits in mice. *Cell, 137*, 961-971.

Fujita, E., Tanabe, Y., Shiota, A., Ueda, M., Suwa, K., Momoi, M. Y., et al. (2008). Ultrasonic vocalization impairment of Foxp2 (R552H) knockin mice related to speech-language disorder and abnormality of Purkinje cells. *Proc Nat Acad Sci, 105*, 3117-3122.

Hurst, J. A., Baraitser, M., Auger, E., Graham, F., & Norell, S. (1990). An extended family with a dominantly inherited speech disorder. *Dev Med Child Neurol, 32*, 352-355.

Lai, C. S. L., Fisher, S. E., Hurst, J. A., Vargha-Khadem, F., & Monaco, A. P. (2001). A forkhead-domain gene is mutated in a severe speech and language disorder. *Nature, 413*, 519-523.

Leonard, L. B. (1998). *Children with specific language impairment*. Cambridge, MA: MIT Press.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychol, 18*, 1-86.

McGregor, K. K., Friedman, R. M., Reilly, R. M., & Newman, R. M. (2002). Semantic representation and naming in young children. *J Speech Lang Hear Res, 45*, 332-346.

McGregor, K. K., Newman, R. M., Reilly, R. M., & Capone, N. C. (2002). Semantic representation and naming in children with specific language impairment. *J Speech Lang Hear Res, 45*, 998-1014.

McMurray, B., Samelson, V. M., Lee, S. H., & Tomblin, J. B. (2010). Eye-movements reveal the time-course of online spoken word recognition language impaired and normal adolescents. *Cognitive Psychol, 60*, 1-39.

Misyak, J. B., Christiansen, M. H., & Tomblin, J. B. (2009). Statistical learning of nonadjacencies predicts on-line processing of long-distance dependencies in natural language. In N. Taatgen, H. van Rijn, J. Nerbonne, & L. Schomaker (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (p. 177-182). Austin, TX: Cognitive Science Society.

Mueller, K. L., Bjork, J. B., Tomblin, J. B., & Murray, J. C. (in preperation). *Common genetic variants in FOXP2 are associated with individual differences in language development.*

Perruchet, P., & Pacton, S. (2006). Implicit learning and statistical learning: One phenomenon, two approaches. *Trends Cognitive Sci, 10*, 233-238.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propogating errors. *Nature, 323*, 533-536.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science, 274*, 1926-1928.

Shu, W., Lu, M. M., Zhang, Y., Tucker, P. W., Zhou, D., & Morrisey, E. E. (2007). Foxp2 and Foxp1 cooperatively regulate lung and esophagus development. *Development, 134*, 1991-2000.

Tomblin, J. B., Christiansen, M. H., Bjork, J. B., Iyengar, S. K., & Murray, J. M. (2007a). Association of FOXP2 genetic markers with procedural learning and language. Poster presented at the Annual Meeting of the American Society of Human Genetics.

Tomblin, J. B., Mainela-Arnold, E., & Zhang, X. (2007b). Procedural learning in adolescents with and without specific language impairment. *Lang Learn Dev, 3*, 269-293.

Tomblin, J. B., Records, N. L., & Zhang, X. (1996). A system for the diagnosis of specific language impairment in kindergarten children. *J Speech Hear Res, 39*, 1284-1294.

Vargha-Khadem, F., Watkin, K. E., Alcock, K., Fletcher, P., & Passingham, R. (1998). Praxic and nonverbal cognition deficits in a large family with a genetically transmitted speech and language disorder. *Proc Nat Acad Sci, 92*, 930-933.

Vernes, S. C., Newbury, D. F., Abrahams, B. S., Winchester, L., Nicod, J., Groszer, M., et al. (2008). A functional link between distinct developmental language disorders. *N Eng J Med, 359*, 2337-2345.

Time Course of Visual Attention in Statistical Learning of Words and Categories

Chi-hsin Chen[1], Chen Yu[2] ({chen75, chenyu}@indiana.edu)
Damian Fricker[2], Thomas G. Smith[2], Lisa Gershkoff-Stowe[1]
Department of Speech and Hearing Sciences[1], Department of Psychological and Brain Sciences[2]
Indiana University, IN 47405 USA

Abstract

Previous research indicates that adult learners are able to use co-occurrence information to learn word-to-object mappings and form object categories simultaneously. The current eye-tracking study investigated the dynamics of attention allocation during concurrent statistical learning of words and categories. The results showed that the participants' learning performance was associated with the numbers of short and mid-length fixations generated during training. Moreover, the learners' patterns of attention allocation indicated online interaction and bi-directional bootstrapping between word and category learning processes.

Keywords: Eye-tracking; statistical learning; word learning; category learning.

Introduction

Over the past few decades, researchers have found that humans are sensitive to statistical regularities in the environment. People are able to use statistical information in non-linguistic tasks, such as making inferences (e.g., Xu & Denison, 2009) or finding predictive features of complex visual scenes (e.g., Fiser & Aslin, 2001). They can use statistical information in linguistic tasks as well, such as learning phonetic distributions (e.g., Maye *et al*., 2002), word boundaries (e.g., Saffran *et al*. 1996b), word and meaning mappings (e.g., Smith & Yu, 2008), and rudimentary syntax (e.g., Gomez & Gerken, 1999). These studies suggest that statistical learning is a domain-general ability in human cognition.

An earlier cross-linguistic study conducted in our laboratory (Chen *et al*., 2009) also showed that adult English and Mandarin speakers were able to use co-occurrence information to learn word-to-object mappings and to form object categories at the same time. However, even though these two groups of learners had comparable performance in learning word-to-object mappings, they showed different levels of sensitivity to the cues associated with category learning. Participants were better at learning the types of regularities that were present in their native language than the ones that were incongruent with their linguistic input. In Experiment 1 of the study, objects from the same category had similar attached object parts and their labels ended with the same final syllable. This syllable-to-category association simulated a prevalent linguistic feature in Mandarin in that the final syllables of object names often indicated category membership. The results showed that Mandarin speakers were able to learn individual word-to-object mappings and to form syllable-to-category associations under cross-situational learning contexts. On the other hand, English speakers tended not to use the final syllables of labels as cues in category learning. In Experiment 2 of that study, the category markers were moved to the beginning of labels to simulate a more frequent feature in English (e.g., the adjectives in noun phrases). As the structures of the training stimuli were more congruent with the input in the naturalistic environment, the English speakers' category learning performance became significantly better. More importantly, they also had better performance in the word learning task. One possible explanation of the improvement of word learning performance is that category learning bootstraps word learning. That is, learning which objects belong to the same category helps the learners to focus on relevant features of the stimuli and to rule out certain distractors as possible referents of a word. However, from the design of that study, we were not able to draw a conclusive link between the English speakers' success in forming categories and their improvement in word learning.

The present study was designed to address this issue by using eye-tracking techniques. Category learning studies using eye-tracking techniques have shown that learners generally attend to all possible dimensions early in learning. But during the process of learning, they gradually shift their attention to relevant dimensions (e.g., Rehder & Hoffman, 2005; Blair et al., 2009). Based on previous studies, similar patterns might be observed in statistical word learning and category learning. Our prediction is that at the beginning of training, learners will pay attention to all objects on the screen when hearing a word. Across learning, they will gradually tune their attention to the most probable referent of a word. Moreover, after successfully forming a few word-to-object mappings, the learners should notice that the objects (and their labels) can be grouped into different categories, each having its own distinctive feature. After establishing primitive category structures, the learners should then use this information to rule out certain distractors as possible referents of a word. The goals of the current study are to examine the dynamics of attention allocation in statistical learning of words and categories and to investigate the real-time interaction between word learning and category formation.

Method

Participants

Participants were 23 undergraduates (14 females, mean age: 19.1 years) who received course credit for volunteering.

None had previously participated in any cross-situational learning experiments.

Design and Stimuli

The experimental design in this study was the same as the one used in Experiment 2 of Chen *et al.* (2009) with slight modification in the length of training trials. Participants were trained under a cross-situational learning paradigm, which was first proposed by Yu and Smith (2007). In each training trial, the participants viewed four novel objects on a computer screen and heard four novel words. However, the temporal order of the word presentations was not related to the spatial locations of the words' target referents. In order to find the correct word-to-object mappings, the participants had to track the co-occurrence regularities between objects and words across different trials. There was a total of 18 object-word pairs to learn. Over the training, there were 12 repetitions per object-word pairing, yielding a total of 54 trials (18 pairs *12 repetitions / 4 pairs per trial). The length of each trial was 14 seconds and the whole training lasted for 12.6 minutes.

The to-be-learned objects were divided into three different categories, with six items in each category. Members in a category had an attached part that looked similar to each other. As an example, Figure 1 shows two items from a category in which all members had an attached spiral part that spread at the end. Moreover, these objects all had labels that began with the same syllable (e.g., *la-* in this case).

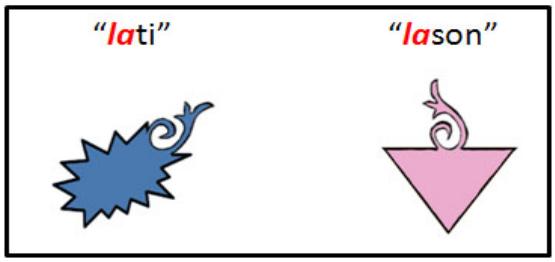


Figure 1 Sample objects and labels used in the study

Apparatus

The course of the experiment was controlled by a computer using E-prime. The visual stimuli were presented on a 17 inch monitor with a resolution of 1280*1024 pixels. The learners' eye gaze was measured by a Tobii 1750 near infrared eye-tracker (www.tobii.se). The eye-tracking system recorded gaze data at 50Hz (accuracy = 0.5°, and spatial resolution = 0.25°).

Procedure

Before the experiment, the eye-tracker system was calibrated. We used a procedure including nine calibration points. The experiment consisted of a Training session, followed by a Testing session. In the Training session, the participants were presented with 4 novel objects and 4 novel words in each trial without any information about which word referred to which object. The learners had to keep track of the co-occurrences between objects and words across trials to find the correct word-to-object mappings. Once they formed several correct word-to-object mappings, we expected they would be able to detect the associations between the first syllables of words and the attached object parts and to form object categories accordingly. The syllable-to-category associations should in turn facilitate word-to-object mappings, because the learners would be able to use the first syllable of a label to determine its possible referents. Eye movements were recorded during the Training session.

There were two tasks in the Testing session, a word-to-object Mapping task and a Generalization task. The Mapping task tested how well the participants learned the names of the training objects. The participants were instructed to select the referent of a training word from 4 alternatives. There were 18 trials in the Mapping task.

In the Generalization task, the participants were asked to select the referent of one novel word from three alternatives, each containing the object-part that corresponded to the particular feature of one category. The first syllable of the novel word was the same as the labels from one of the three categories. If the learners had formed the syllable-to-category associations, they should be able to use the first syllable of the novel word to find its referent. There were 9 trials in the Generalization task (3 for each category).

Eye-tracking dependent variables

To derive eye movement measures, we defined four rectangular region-of-interests (ROIs) that covered the objects displayed on the screen for each trial. We took the onset of a series of gaze data that fell within an ROI as the onset of a fixation and the end of the fixation was determined when the gaze fell outside of the same ROI. The minimum length of a gaze was 20ms (i.e., the length of 1 data point recorded by the eye-tracker). All gaze data outside the ROIs were viewed as saccadic eye movements and not included in the analyses.

Based on the remaining gaze data, we computed two dependent measures. The first variable was the *number of fixations* per trial. We set the thresholds at 100ms, 500ms, and 1000ms and counted the numbers of fixations exceeding these thresholds. Moreover, fixations with a length between 100ms and 500ms were defined as Short fixations; fixations between 500ms and 1000ms were viewed as Mid-length fixations; and those longer than 1000ms were taken as Long fixations. The reason for setting different thresholds was that previous category learning studies using eye-tracking techniques have found that looking more at the correct or relevant features during training was positively correlated with behavioral performance (e.g., Rehder & Hoffman, 2005; Blair et al., 2009). This indicates that more looking at the relevant features during training might lead to better learning. However, more looking could result from either having a few long fixations or having many short fixations combined together. Setting different thresholds would allow

us to examine whether longer looking also leads to better learning.

The second measure was *proportion looking time* (ranging from 0 to 1), which took the time spent fixating on one object divided by total time spent fixating on all objects. Moreover, based on the word being presented, we divided the objects into 3 categories: Correct Object, Within-Category Distractor, and Between-Category Distractor. Because there were 4 objects in each training trial while there were only 3 categories to learn, there could be more than 1 object from a specific category in a trial. Therefore, for each word, the Correct Object was the target referent while a Within-Category Distractor was an object from the same category. On the other hand, the Between-Category Distractors were the ones from a different category. Figure 2 illustrates a situation in which there are two objects from the *la-* category, one from the *jo-* and one from the *mu-* category. The label of each object can be found above it (please note that in real training, the labels were presented auditorily). For the word "*lati*", there is one Within-Category Distractor and two Between-Category Distractors in this trial. In contrast, for the word "*joler*", there are three Between-Category Distractors. However, in this case none of the objects is a Within-Category Distractor for this word. The mean numbers of Correct Object, Within-Category Distractor, and Between-Category Distractor for the training words in each trial are: 1, 0.74, and 2.26, respectively.

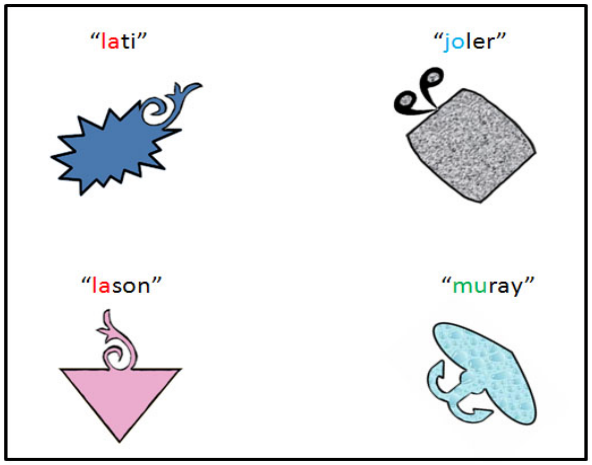


Figure 2 Sample stimuli in Training

Behavioral Results

On average, more than 50% of the participants' responses were correct in the Mapping task and in the Generalization task as well (see Figure 3). Consistent with earlier findings, participants learned more word-to-object mappings than expected by chance ($t(22) = 4.211$, $p < .001$). They also performed significantly above chance in the Generalization task ($t(22) = 3.227$, $p = .004$). That is, they could use the first syllable of a novel label to find its referent. In addition, we found a strong positive correlation between the learners' Mapping and Generalization performance ($r = .773$, $p < .001$). This suggests that the more words participants learned, the more likely they were to use the first syllable as a cue in categorizing novel objects.

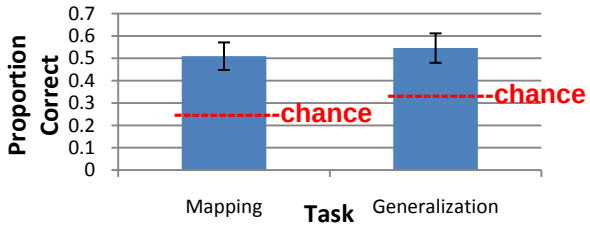


Figure 3: Proportion of accurate responses in Mapping and Generalization tasks

Eye Movement Data Analyses

According to the participants' performance in the Mapping task, we divided them into three groups. The participants that had more than 70% correct responses were viewed as High Learners. The people that made less than 35% correct responses were viewed as Low Learners. People having 35% to 70% correct responses were viewed as Mid Learners. There were 8, 6, and 9 people in the High, Mid, Low group, respectively. We compared the *number of fixations* and *proportion looking time* to different types of objects of the High, Mid, and Low Learners to see if there were differences in their eye movement patterns during the training.

Number of Fixations

As mentioned previously, we counted the numbers of fixations exceeding 100ms, 500ms, and 1000ms for each participant. The results can be found in Figure 4. The solid lines indicate the numbers of fixations exceeding 100ms. The High, Mid, and Low Learners had comparable numbers of fixations at the beginning of training. Across the Training session, the numbers of fixations of the Mid and Low Learners gradually decreased and the decreasing rate was slightly higher for the Low Learners. The dashed lines show that when the threshold was set at 500ms, the High Learners tended to have more fixations than the other two groups, especially in the second half of training. When the threshold was set at 1000ms, there did not seem to be group differences.

The patterns observed above were confirmed by statistical analyses. We compared the numbers of Short (100ms-500ms), Mid-length (500ms-1000ms), and Long fixations (<1000ms) of different groups of learners. With regard to Short fixations, trial-by-trial ANOVAs showed that group differences were significant between Trial 38 and Trial 42 (*p*s <.05). Pair-wise comparisons showed that the High Learners generated more Short fixations than the Low Learners (*p*s <.05). For Mid-length fixations, Trial-by-Trial ANOVAs revealed that significant group differences occurred between Trial 31 and Trial 39 at *p* level of .05. Pair-wise comparisons showed that the High Learners generated more Mid-length fixations than the Mid and Low Learners (*p*s < .05). In addition, the Mid Learners also generated more Mid-length fixations than the Low Learners

in Trial 13, 16, 39 and 40. When the threshold was raised to 1000ms, all three groups had about equal numbers of fixations across trials. Significant group differences were only found at Trial 26, in which the High Learners generated more fixations than the Mid and Low Learners (*p*s<.05).

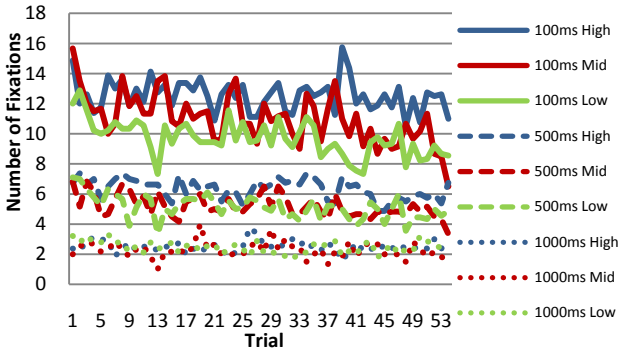


Figure 4 Number of Fixations of High, Mid, and Low Learners. The number of fixations was counted separately with 100ms, 500ms, and 1000ms as thresholds of minimal eye fixation length.

To summarize, the major differences between the High, Mid, and Low Learners were caused by the decreasing Short and Mid-length fixations of the Mid and Low Learners. The High Learners had more Short and Mid-length fixations than the other two groups, especially in the second half of training. The Mid learners also generated more Mid-length fixations than the Low learners.

Proportion Looking Time

Proportion Looking Time By Trial We first looked at the dynamics of attention allocation during the course of statistical learning. For ease of comparison, Figure 5 to Figure 7 present the normalized Proportion Looking Time of the High, Mid, and Low Learners across training trials. The Proportion Looking Time to a certain type of object is normalized so that the chance level is 25%. As can be seen from Figure 5, there was a drastic increase in the High Learners' Proportion Looking Time to the Correct Object. There was also a decreasing trend in their looking at the Between-Category Distractors. Starting from Trial 34, the High Learners looked at the Correct Object significantly more than expected by chance (*p*s < .05). They also looked at the Between-Category Distracters significantly less than chance from Trial 35 on (*p*s < .05). As to the Mid Learners in Figure 6, even though there was an increasing trend in their Proportion Looking Time to the Correct Object, it did not reach statistical significance. As can be seen in Figure 7, the Low Learners had chance level performance across the training. Though they had above- or below-chance performance in a few trials, the patterns were not reliable.

We also conducted trial-by-trial ANOVAs to compare group performance. Starting from Trial 38, the High Learners looked at the Correct Object more than the Mid and Low Learners (at *p*s <.05). The pattern can be seen in Figure 8. There was also a trend that the Mid Learners looked at the Correct Object more than the Low Learners at the last third of training. But the pattern was not reliable. As to Within-Category Distractors, there were significant group differences in a few trials in which the High and Mid Learners looked at the Within-Category Distractors more than the Low Learners. But the patterns were not reliable either. With regard to Between-Category Distractors, there were significant group differences starting from Trial 24. Compared to the High Learners, the Low Learners looked more at the Between-Category Distractors in the second half of training. Additionally, they looked more at the Between-Category Distractors than the Mid Learners in the last third of training.

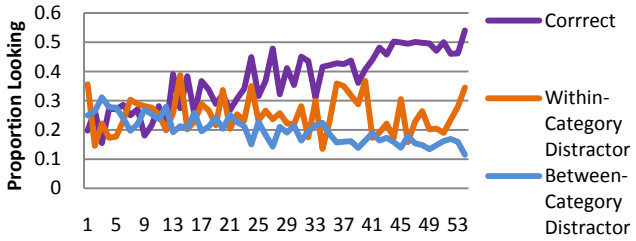


Figure 5 Proportion Looking Time of High Learners

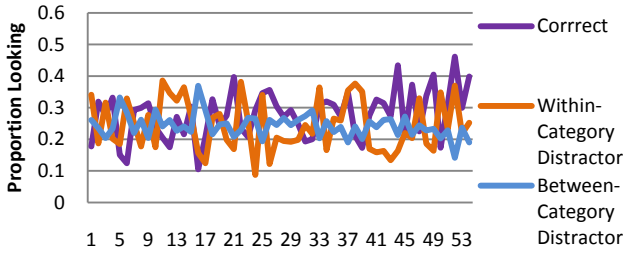


Figure 6 Proportion Looking Time of Mid Learners

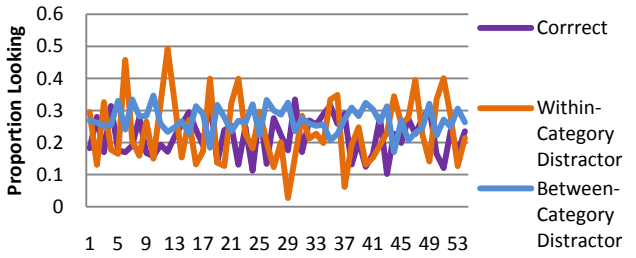


Figure 7 Proportion Looking Time of Low Learners

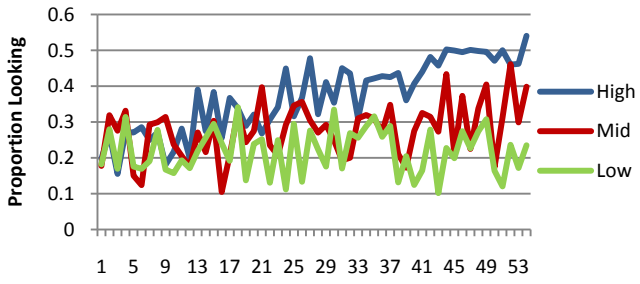


Figure 8 Proportion Looking Time to the Correct Object

Proportion Looking Time By Occurrences Across the Training session, each word-object pair occurred 12 times. For each participant, we calculated the Proportion Looking Time by word-object occurrences. For example, we took their Proportion Looking Time at the first occurrence of individual objects and averaged it across objects to get the Proportion Looking Time at Occurrence 1. This gave us 12 values for each participant. We then compared the High, Mid, and Low Learners' Proportion Looking Time to the Correct Object by occurrence.

Figure 9 illustrates that at about the third time the High Learners heard a word, they looked more at the Correct Objects than the Mid and Low Learners. Trial-by-trial analyses showed that group differences became significant at the third occurrence of a word (*p*s < .05). Except for the 6th occurrence, the High Learners were more likely to look at the Correct Object than the other two groups. The Mid Learners looked more at the Correct Objects than the Low Learners from Occurrence 10 to Occurrence 12.

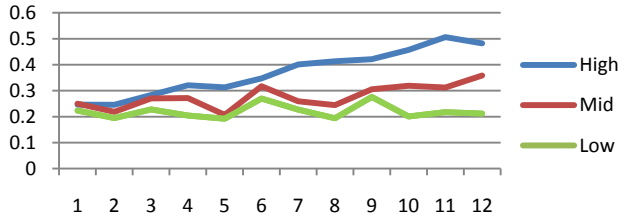


Figure 9 Proportion Looking Time to Correct Object by Occurrences

Compared to chance, the High Learners looked at the Correct Objects significantly above chance from the 7th to the last time they encountered a word (*p*s < .05). The Mid Learners looked at the Correct Objects significantly above chance from the 10th to the last time they heard a word (*p*s < .05). As for the Low Learners, they did not look at the Correct Objects more than chance. This indicates that it took only a few repetitions for the High Learners to detect the word-to-object co-occurrence regularities and that they could quickly tune their attention to the most probable referent of a word. However, it took longer for the Mid Learners to find the correct referent of a word.

Predictive Looking

Because the first syllable of a label indicated an object's membership, another question we were interested in was whether the participants made predictive looking and attended to objects from a relevant category even before the whole word was finished. For example, if the learners formed the association between the syllable *la-* and the spiral part, they might be able to use the syllable *la-* as a cue to rule out Between-Category Distractors even before the word "*lati*" was completed.

We calculated Proportion Looking Time to objects from a relevant category (i.e., the Correct Object and Within-category Distractor) and objects from irrelevant categories between 600ms and 900ms after the onset of a word. We chose the time between 600ms and 900ms based on the approximation that it took at least 200ms to generate stimulus-driven fixations and 600ms is about 200ms after the end of the first syllable while 900ms is about 200ms after the end of the word[1]. The Proportion Looking Time to object from a Relevant Category of the High, Mid, and Low Learners can be seen in Figure 10. For ease of comparison, the results were normalized, so that the chance value was .5. In the first half of training, all three groups had similar performance. In the second half of training, the Mid and the High Learners started to fixate on objects from a Relevant category even BEFORE the whole word was completed. However, for the Mid Learners, the trend was not as reliable as the High Learners.

It is noteworthy that the High Learners' predictive looking could only be reliably observed in the last third of training, which occurred after their reliable above-chance looking at the Correct Objects. This indicates that prior to forming syllable-to-category associations, the learners needed to establish at least a few correct word-to-object mappings in order to extract the regularities across objects.

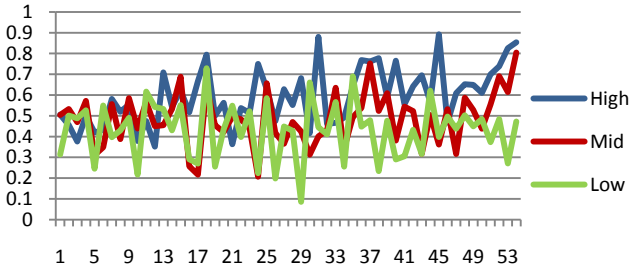


Figure 10 Proportion Looking between 600ms and 900ms after the onset of a word

Predictors of Behavioral Performance

As mentioned, the participants were grouped based on their performance in the Mapping task, which is a behavioral task administered after training. The above analyses showed that group differences could be observed from eye movement data during training. This suggests that eye gaze patterns during training might be used as predictors of behavioral performance.

Table 1: Correlations between Eye Gaze and Behavioral Measures.

		Mapping	Generalization
Number of Fixation	Short	.167	.107
	Mid-length	.339	.400*
	Long	.150	.118
Proportion Looking	Correct	.803**	.586*
	Within-category	.046	.278
	Between-category	-.749**	-.609**

* p < .05
** p < .001

[1] We also tried 500ms-800ms and 500ms-900ms. The trends are similar to the patterns observed here.

To find the best predictor of behavioral performance, multiple linear regression analyses were conducted. As can be seen from Table 1, there is a positive correlation between the number of Mid-length fixations and Generalization performance. The learners' Proportion Looking Time to the Correct Object is positively correlated with their Mapping and Generalization performance. In contrast, Proportion Looking Time to the Between-Category Distractors is negatively correlated with Mapping and Generalization performance. Stepwise regression showed that the best predictor of the Mapping performance is Proportion Looking Time to the Correct Objects during training. Consistent with the findings of previous studies, the more the learners looked at the correct features during training, namely the correct object, the better they performed in the following behavioral task. On the other hand, the best predictor of the Generalization performance is Proportion Looking Time to the Between-Category Distractors. The less the learners looked at the Between-Category Distractors, the better they did in the following Generalization task. This suggests that less looking at the Between-Category Distractors can be viewed as an indicator of category learning.

General Discussion

This study replicates previous findings that adult learners are able to use co-occurrence information to simultaneously learn word-to-object mappings and to form object categories. In addition, the current study shows that the learners' behavioral performance in the Mapping and Generalization tasks can be predicted from their looking patterns during the course of learning. Learners who generated more short- and mid-length fixations tended to perform better in the following behavioral tasks. However, there was no difference in the numbers of long fixations generated by different groups of learners. This indicates that more looking was not due to longer looking. Instead, the good learners tended to shift their attention back and forth among objects to check the possible referents of a word. Thus, rapid gaze shifts between several concurrent visual objects suggest a real time competition process which leads to better learning.

Patterns of attention allocation of the High, Mid, and Low Learners could be detected during the course of learning in addition. After accumulating certain statistical information, learners tended to shift their attention to objects containing relevant features. Moreover, at the third encounter with a word, the High Learners appear to have (partially) formed the association between a word and its referent. On the other hand, it took about 10 times for the Mid Learners to form correct mappings. This suggests that from eye movement data, we might be able to observe the accumulation of partial knowledge and how it leads to successful learning.

After forming a few individual word-to-object mappings, the High and Mid Learners shifted their attention to relevant categories BEFORE a word was completed. This suggests that after establishing syllable-to-category associations, they use the first syllable of a word to eliminate Between-Category Distractors as possible referents of the word. Together, the results of the present study reflect online interaction of word learning and category learning. It also provides evidence that word learning and category learning bootstrap each other.

References

Blair, M. R., Watson, M. R., Walshe, R. C., & Maj, F. (2009). Extremely selective attention: Eye-tracking studies of the dynamic allocation of attention to stimulus features in categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*, 1196-1209.

Chen, C., Yu, C, Wu, C.-Y., & Cheung, H. (2009). Statistical Word Learning and Object Categorization: A Cross-Linguistic Study in English and Mandarin. *Proceeding of the 31st Annual Conference of the Cognitive Science Society*.

Colunga, E. and Smith, L. B. (2008). Knowledge embedded in process: the self-organization of skilled noun learning. *Developmental Science, 11(2)*, 195-203.

Fiser, J., & Aslin, R. N. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological Science, 12(6)*, 499-504.

Gomez, R. L., & Gerken, L. (1999). Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition, 70(2)*, 109-135.

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition, 82(3)*, B101-B111.

Rehder, B. & Hoffman, A. B. (2005). Eyetracking and selective attention in category learning. *Cognitive Psychology*, 51, 1-41.

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996b). Word Segmentation: The role of distributional cues. *Journal of Memory and Language, 35*, 606-621.

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition, 106*, 1558-1568.

Xu, F. & Denison, S. (2009). Statistical inference and sensitivity to sampling in 11-month-old infants. *Cognition, 112*, 97-141.

Yu, C., & Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science, 18*(5), 414-420.

Development in the Estimation of Degree Measure: Integrating Analog and Discrete Representations

Jonathan Michael Vitale (JMV2125@Columbia.Edu)
Department of Human Development, Teachers College, 525 W. 120th Street
New York, NY 10027 USA

John B. Black (Black@Exchange.TC.Columbia.Edu)
Department of Human Development, Teachers College, 525 W. 120th Street
New York, NY 10027 USA

Eric O. Carson (EOC2102@Columbia.Edu)
Department of Mathematics, Science and Technology, Teachers College, 525 W. 120th Street
New York, NY 10027 USA

Chun-Hao Chang (Seedmic@Gmail.com)
Department of Human Development, Teachers College, 525 W. 120th Street
New York, NY 10027 USA

Abstract

We examined adult and child performance on two numerical, geometric estimation tasks. In both tasks adults demonstrated greater accuracy than children as well as more mature representations, in general. Furthermore, evidence from mouse tracking data demonstrates that adult strategy includes the application of discrete landmark values while child strategy, generally, does not. This evidence suggests that adults construct mental representations of landmark values and successfully integrate them into analog tasks. Implications for future intervention studies are discussed.

Keywords: numerical estimation, embodied cognition, mathematical development, cognitive assessment

Introduction

Numerical estimation tasks provide researchers with a powerful means of assessing individuals' mental representation for number. Evidence from brain scans demonstrates that approximate numerical tasks, such as less-than/greater-than judgments, activate cortical regions associated with spatial processes, while activities that rely exclusively upon recall, such as single-digit multiplication, do not (Dehaene, Piazza, Pinel, & Cohen, 2003). According to Dehaene (1997) our ability to map numerical values to spatial magnitudes is what is commonly referred to as "number sense," and grounds all mathematical reasoning.

Yet, the study of number sense extends beyond theoretical interest as recent evidence suggests a link between estimation and mathematical achievement. Along these lines Halberda, Mazzocco, and Feigenson (2008) discovered that 14-year-old's ability to discriminate between dot displays of varying cardinalities was highly correlated with achievement scores extending back to kindergarten. Likewise, Siegler and Booth (2004) found that individual differences on a number line estimation task are correlated with standardized test scores.

In the case of number line estimation individual differences may embody large, qualitative shifts in representation (Siegler & Opfer, 2003). Dehaene (1997) asserts that numerical symbols implicitly recruit a logarithmic representation that is more precise at smaller values. Siegler and Opfer (2003) found that young children, especially with larger numerical ranges, tend to apply this kind of logarithmic representation while estimating the position of a given value on a number line. Specifically, data of estimated magnitude over actual magnitude are best fit by a logarithmic function for these younger children. On the other hand, older children's data, in many cases, is best fit by a straight line.

The emergence of a linear representation has several possible causes and implications. In particular, Siegler and Opfer (2003) differentiate between two models of linear representation. In the *accumulator* model, adopted from Gibbon and Church (1981), noise in the mental representation for a numerical value increases in proportion to the mean. This representation implies increasing variability in the estimates as the magnitude increases. In the *linear-ruler* model – which was found to be a better fit for the data – variability in estimates has a constant relation to magnitude.

The authors suggest that the mature, linear representation is developed through cultural, particularly school-based, experience. Furthermore, evidence of less variability near landmark values along the number line (e.g. quartiles) demonstrates a specific means for implementing the linear-ruler model. One may even speculate that at the lowest-level number representation may be logarithmic or accumulator in nature, but at the level of conscious-level processing number concepts are modulated for specific tasks.

If the appeal of number line estimation tasks is due, in part, to its high ecological validity, one might then find it

surprising that although number lines are a ubiquitous feature of elementary school classrooms, many students maintain immature, logarithmic representations. Yet, recent evidence suggests that the development of mature representations may be promoted through simple, economical interventions, such as playing linear board games (Siegler & Ramani, 2008) or providing corrective feedback (Opfer & Siegler, 2007) on values that maximize the logarithmic-linear difference. In the latter case many children demonstrated a logarithmic to linear shift within a few feedback trials.

Yet, the ease with which some children transition from a logarithmic to linear representation begs the question of whether these children already maintain a linear representation of whole numbers and simply learn to recruit it for the given task. From this perspective, "development" of a linear representation in these interventions may capture only the tail end of this learning process, only made possible through years of informal experience with numerical concepts (Ginsberg, 1983).

At the cost of ecological validity, an alternative approach to cognitive developmental research might imply the adoption of a task utilizing a novel, unique spatial representation of number. Within such a paradigm researchers would observe as children (or adults) struggle to construct meaning out of the task, although the task may be meaningless beyond the research setting. This research model would afford psychological researchers with a level of control that is unobtainable with common concepts.

As a compromise between ecological validity and control this study applies the numerical estimation paradigm to degree measure, which is an important element of mathematics, but is under-utilized in elementary school curriculum and therefore relatively unfamiliar to children (Clements & Battista, 1992). Considering that degree measure does not become a major component of curriculum, generally, until high school, research with degree measure provides an opportunity to study numerical development of older children.

Yet, degree measure is not a unitary concept, but is rather composed of two psychologically distinct spatial representations: degree as angle of intersection between lines and degree as rotation. While some tasks, such as LOGO programming, may confound the two concepts (Clements, Battista, & Sarama, 2001), other activities clearly demonstrate that children perceive physical models of each concept distinctly (Mitchelmore, 1998).

Given the unique spatial qualities of each representation, we should expect courses of development for rotation and angle concepts that may differ from whole number concepts. For example, Clements and Burns (2000) found that fourth grade students physically modeled angle values and curtailed the degree of embodiment with increasing expertise. Furthermore, both students and instructors focused on the representation of "benchmark" (or landmark) magnitudes, such as 90°. Although one might perform a degree estimation task by applying the same linear representation developed for the number line, albeit in a circular form, the emphasis on standard landmarks for degree measure suggests that performance with number lines and degree measure is likely to be quite different. Specifically, the mental representation for degree measure might rely upon the integration of continuous models of numbers and discrete abstractions of landmark values.

While the nature of a mental abstraction is a constant source of debate, the grounded or embodied cognition framework (Barsalou, 2008) asserts that all mental representations are composed of sensory-motor elements of experience. Specifically, perceptual symbols develop from frequent encounters with a meaningful type of object. In turn perceptual simulators develop to provide individuals a means of representing a concept in its perceptual absence (Barsalou, 1999).

In the case of angle and rotation, perceptual symbols are likely to embody landmark values. Given the perceptual salience of perpendicular lines – which can be discriminated from non-perpendicular lines by Amazonian tribesman (Dehaene & Izard, 2006) – we should expect that 90° angles are represented in this form. However, the perceptual symbol encoding 90° angles may only account for a limited range of valid right angles, such as right angles with sides oriented horizontally and vertically from the ground. Thus, perceptual symbols may develop in both robustness for particular symbols, and in number, overall.

In the case of rotation, the spatial mapping of language may play an important role in the embodiment of this numerical concept (Lakoff & Nunez, 2000). For example, the directives "turn around" or "turn to your right" may ground landmark values of 180° and 90°, respectively. Older students may develop other landmark values for common spatial transformations, such as a rotation of 45°.

Yet, all numerical tasks involving degree measure do not, explicitly, require landmark values. An angle measure of 117°, for example, is unlikely to have its own unique representation. However, individuals may shuttle between analog, continuous models and discrete, abstract models (Schwartz & Black, 1996). In this case one is likely to apply this process to numerical estimation by searching for relevant landmarks (e.g. 90°) and then applying an analog procedure, in a form of divide-and-conquer. As stated above, Siegler and Opfer (2003) suggest that this is the specific mechanism used by adults to implement a linear-ruler representation on the number line. However, since number lines may utilize an arbitrary range of values, linear representation may reflect the online development of landmarks, rather than the application of perceptual symbols from memory.

Although evidence for the application of landmark strategies is suggested by the pattern of variability in accuracy across numerical range, these accuracy measures reflect only the final judgment of the participant and may hide strategy-relevant features of the estimation process. On the other hand continuous, online measures of performance afford researchers a view of the specific process undertaken

by a participant (Spivey, 2007). In this study we adapt a mouse-tracking paradigm in which mouse position and time is recorded at continuous, fine intervals. Specifically, as participants perform the estimation task the mouse's rotational orientation about the center of the screen is recorded to facilitate the analysis of inflections near landmark values.

In the following study we analyze estimation performance of relative experts (graduate students in education) to novices (middle-to-upper elementary school students). Given the adults likely experience with relevant geometric concepts, we expect that these participants are likely to use apply a process of shuttling between landmark and analog representations, which should result in a high proportion of mouse stops near landmark values and an overall linear representation.

The children, on the other hand, are less likely to be familiar with landmark values and may struggle to integrate them into the estimation task. Therefore we suspect that linear representations will be rare. Although these students, mostly in fourth grade, should clearly maintain a linear representation on the number line, we expect that, given the novelty of this task, they are likely to adopt a logarithmic representation. Furthermore, we expect there to be clear differences between adults and children in overall accuracy.

Method

Participants

Sixteen adults (mean age = 28.9, SD = 8.7) were recruited from an introductory cognitive psychology course as part of a research requirement. Sixteen children (mean age = 9.5, SD = .73) were recruited from an after-school program located in a low-SES neighborhood of a large city. The children consisted of two third-graders, one fifth-grader, and thirteen fourth graders.

Tasks

Both adults and children performed two distinct numerical estimation tasks on Apple MacBooks. Both tasks were programmed with Adobe Air 1.1 in the Adobe Flash CS3 development suite. The application covered the entire height of the screen (23 cms) and approximately 87% of the width (32.5 cms).

Both tasks required a single click (and release) on a circle within the display to initiate each trial. Upon completing each trial participants were required to click-and-hold the mouse button for a half second to "lock-in" their answer to reduce the frequency of accidental clicking. Although there was no time limit, if the participant made no motion with the mouse for more than ten seconds the trial was terminated.

In *angle construction* (Figure 1) the participant maneuvers the mouse to rotate one leg (9.2 cms long) of an isosceles triangle clockwise about a fixed vertex, while the other leg remained motionless – opening and closing the triangle. Participants were asked to manipulate a target angle, marked with a red arc, to reflect a target number of degrees. At 0° and 180° the figure becomes a straight line. Motion beyond 180° maintained the appearance at 180° and was recorded as 180°. Participants could move directly from 0° to 180° by moving the mouse counter-clockwise from the initial position.

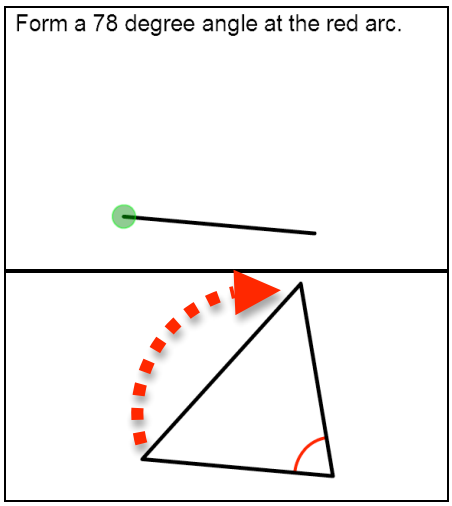


Figure 1: The top pane shows the initial display for an angle construction trial. Participants clicked within the circle to begin the trial. The lower pane shows a triangle that has been formed to match the target value. The arced arrow is superimposed here to demonstrate the vector of motion of the non-stationary vertex from its original position.

In *triangle rotation* (figure 2) the participants maneuvered the mouse to rotate an isosceles triangle about the center of the triangle. Participants were asked to rotate the triangle, clockwise, a target number of degrees from the triangle's initial orientation. A light gray triangle in the initial orientation of the triangle remained throughout the trial to provide a reference. The shape of the triangle was varied between trials by randomizing the angle measure at intersection of the triangles legs from 10° to 170°, although the length of the legs was constant (9.2 cms). Varying the shape was necessary to avoid the use of strategies involving static relationships between the moving triangle and the gray reference triangle. Participants could maneuver the triangle between 0° and 180°. Motion beyond 180° did not affect the appearance of the figure. Like angle construction, in some cases participants moved directly from 0° to 180° by moving the mouse counter-clockwise.

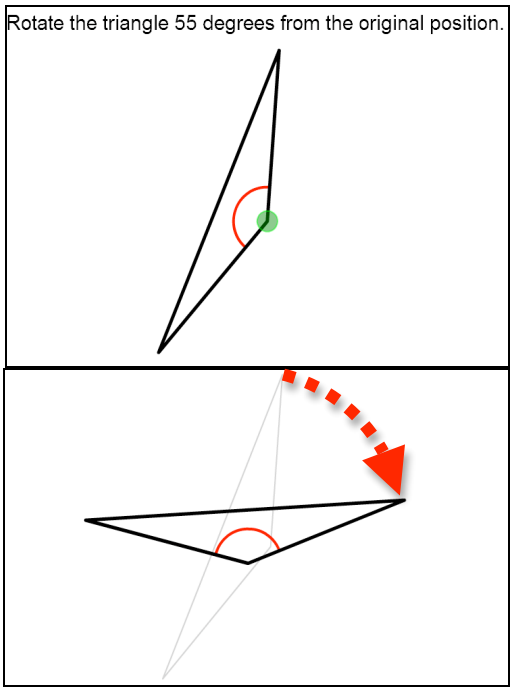


Figure 2: The top pane shows the initial display for a triangle rotation trial. Participants clicked within the circle at the vertex to begin. The lower pane shows how the triangle has been rotated to match the target value.

Procedure

Both children and adults were split into two groups of eight, varying task order. The children received a block of 20 non-feedback trials in both *angle construction* and *triangle rotation*. The adults received 120 trials, organized into six blocks, for each task. However, for the purpose of directly comparing adult and child performance, only the adults' first block for each task were analyzed here. The task was individually administered to adults in a private room. Children performed the task in a dedicated section of a classroom as their classmates completed homework.

Prior to the first block of each task participants received 5 practice trials. Each practice trial required the participant be within 15° of the target and displayed written, verbal feedback suggesting an increase or decrease. The practice values of 90°, 45°, 135°, 15°, and 180° were selected to represent a wide distribution of the range. However, in angle estimation the 180° trial was replaced with a 165° trial to maintain a triangular appearance of the display.

Each block was populated with target values in the range of 10° to 170°. Target values were selected randomly from 20 intervals of 8° over this range. In the interest of directly comparing the two tasks, angles greater than 180° were not used as they cannot form internal angles of a triangle.

Data Analysis

Prior to all analyses outlier trials were removed to eliminate cases of accidental clicks, which prematurely terminated trials. From observation of performance accidental clicks generally occurred near either extreme value (0° or 180°) or within a short time period (e.g. double-clicking). Therefore trials in which the participant moved the mouse less than 5° from the initial position, ended the trial within a degree of the 180° endpoint, or completed the trial in less than one second were removed from analysis. To reduce the likelihood that subjects intended degree measures in the outlier range we only used targets between 10° and 170°.

During each trial the current value of the manipulated degree measure was sampled at approximately every 40 msecs. Degree over time data was fit to a function and smoothed using the 'fda' package within R (Ramsay, Wickham, Graves, & Hooker, 2009). The first derivative of smoothed data, degree change over time, was then searched for values at or near zero for an extended period of time (500 msecs), indicating a stop point.

Stop points within 10° of specific landmark values were tallied and are referred to as *landmark stops*. Likewise stop points in 10° ranges just above and below the landmark ranges were tallied and are referred to as *near-landmark stops*. For example, 90° landmark stops included all stops between 80° and 100°, while 90° near-landmark stops occur in the ranges 70° to 80° and 100° to 110°. For 180° landmark stops were tallied between 170° and 180°, while near-landmark stops range between 160° and 170°.

Although stops in a landmark or near-landmark range could represent random behavior, subjects consistently applying a landmark strategy are more likely to stop within landmark than near-landmark range, while subjects stopping at random should be equally likely to stop within either range. Therefore we suggest that a high proportion of landmark to near-landmark stops indicates the explicit use of a landmark strategy. This landmark-to-near-landmark proportion was calculated as a statistic ranging from -1 (all near-landmark) to 1 (all landmark) by subtracting the count of near-landmark stops from the count of landmark stops and dividing by their total. For example, three landmark stops to one near landmark stop is a value of .50 [i.e. (3-1)/(3+1)]. On the other hand, one landmark stop to two near-landmark stops is a value of -.33 [i.e. (1-2)/(1+2)]. A value of zero indicates either an equal number of landmark and near-landmark stops or no stops.

Results

To determine the nature of a participant's representation of estimated magnitude vs. actual magnitude data was fit to a linear and logarithmic model. Participants were classified to "linear" or "log" representation to indicate the model that accounted for a larger proportion of their variance. In the case where neither model was a significant predictor of estimates participants were classified as "other." Also, the absolute deviations (residuals) from estimated to actual magnitudes of "linear" participants were fit to a linear model to determine the presence of scalar variability. Those

participants with a positive slope, significantly different from zero ($p < .05$) were classified as "accumulator", while those with no trend of increasing deviations were classified as "linear-ruler." Table 1 indicates the distribution of models for adults and children in each task.

Table 1: Model frequencies across task and age.

Model:	Linear-ruler	Linear accum.	Log	Other
Adult–Angle	11	5	0	0
Child – Angle	4	2	5	5
Adult–Rotation	10	2	4	0
Child–Rotation	4	1	1	10

A chi-square test of this distribution indicates that adults and children vary significantly in both angle construction and triangle rotation [$\chi^2(3, N=32) = 14.55$ and 14.7, respectively, $ps < .01$]. Model frequencies between tasks, within age groups, did not differ significantly, $ps > .05$.

Furthermore, the means of all absolute deviation from estimated magnitude to target magnitude (error), indicating overall accuracy, showed a similar pattern (Table 2).

Table 2: Mean and standard deviation of absolute deviations from estimated and actual magnitudes.

	Mean Error	SD
Adult–Angle	9.78°	2.69
Child – Angle	35.7°	16.7
Adult–Rotation	16.5°	4.93
Child–Rotation	41.0°	13.9

A two-way ANOVA with task-type as a within subjects factor and age as a between subjects factor reveals a strong main effect for age [$F(1,31)=29.8$, $p<.000$] and a weaker, yet significant, effect for task type [$F(1,31)=4.6$, $p<.05$], indicating better performance for angle construction. Interaction between task and age was non-significant.

To analyze the use of landmarks each participant's trials were divided into 4 quadrants with target values in the ranges 0°-45°, 46°-90°, 91°-135°, and 136°-180°. For each participant the 180° landmark to near-landmark statistic was calculated for trials in quadrant four (136°-180°) and 90° in quadrant two and three (46°-135°). We then applied t-tests to determine whether these values differed significantly from zero, suggesting explicit use of the landmark. The means and p-values are shown in Table 3 below.

Table 3: Mean ratios of landmark to near-landmark stops and associated p-values.

Angle Construction	**Mean Ratio**		**Triangle Rotation**	**Mean Ratio**
Adult – 90° Q2-Q3	.37*		Adult – 90° Q2-Q3	.19
Child – 90° Q2-Q3	-.06		Child – 90° Q2-Q3	.25
Adult – 180° Q4	.25+		Adult – 180° Q4	.48**
Child – 180° Q4	-.06		Child – 180° Q4	.25*

+ $p < .1$ * $p < .05$ ** $p < .01$

For the rotation task both children and adults applied the 180° landmark strategy, while neither did so in the angle construction task. In angle construction only adults utilized a 90° landmark strategy. We applied two ANOVA models to compare age and task for each landmark within its associated region. With 90° landmarks there was a significant main effect of age [$F(1,31)=5.0$, $p < .05$], but not task-type ($p > .05$). With 180° landmarks there was a significant main effect for task-type [$F(1,31)=4.9$, $p < .05$], but not for age.

Discussion

Clearly, there is a large difference between adult and child performance in both tasks. This is certainly not surprising, given the difference in experience between the adults and children in the domain of geometry. Yet, considering that 10 of the 16 children's data for the rotation task could not be fit by either a linear or logarithmic function, the extent of this difference was surprising.

The graphs for individual subjects who were classified as "other" show either general randomness, crowding towards some arbitrary magnitude, or (in one case) a *negative* linear relationship. For these students, the mapping between numerical value and the chosen spatial representation was either meaningless or completely misconceived.

Another explanation is that these children simply refused to "play the game" correctly, and were simply applying their own, idiosyncratic rules. Yet, considering that only half as many children were classified as "other" for angle construction, it is unlikely that this behavior emerged from general disinterest. Rather, many students expressed their frustrations, especially during the rotation task, by telling test administrators that, "I don't know how to do this." Furthermore, in both tasks, several students were unsure as to which direction represented an increase in value.

Adults and children also differed greatly in the application of landmarks. While adults clearly used 90° landmarks for targets ranging from 45° to 135° in angle construction, children showed little evidence of this strategy. In fact, the children were slightly more likely to stop at near-landmark values (albeit at non-significant level). Both adults and

children used the 180° landmark in the rotation task, but this may be an effect of the particular environment as 180° represents a clear physical boundary.

One may also reasonably claim that near-landmark stops were, in some cases, attempts at using specific landmarks. Yet, due to a general lack of precision, and difficulty with mouse control, children often stopped outside of the accepted landmark range suggesting that, with practice, landmark stops may replace near-landmark stops.

While, this data demonstrates relative extremes in numerical representation it cannot inform us about the path of development from novice to expert. Rather, we are left to ask whether adult performance result from a wealth of exposure to degree measures in particular or from a flexible understanding of the linear nature of numbers? From the latter perspective one might imagine that adults are able to imagine a curved number line with endpoints at 0° and 180° – which would enable linear performance with relatively little experience with degrees, specifically. Furthermore, given the important role that landmark values holds in adult performance, should child instruction focus on strategies incorporating landmark values or will mental representations for these landmarks emerge from exposure to the entire range of magnitudes?

Such questions suggest the potential of intervention studies to elucidate paths of development. Possible interventions to promote understanding of degree measure may include measuring angles, playing games aimed specifically at these numerical constructs, or situated activities such as LOGO programming. In particular our research team is currently investigating the latter two means of developing numerical understanding.

In a preliminary intervention study applying a LOGO-like environment and geometry curriculum with thirteen children (from this study), we have found a trend towards improvement in overall accuracy measure for angle construction [$t(12)= 2.1$, $p = .059$]. Furthermore, of these 13 students the number of students demonstrating a linear representation increased from four to nine.

Although the study of relatively novel numerical concepts is of theoretical interest, one might argue that if these concepts are so under-represented in curriculum then interventions at this level may be unnecessary or inappropriate. However, the National Council of Teachers of Mathematics (2000) stresses the important of geometric and spatial reasoning for children of all ages. We suspect that mastery of basic concepts, such as angle measure, serves as a grounding for higher-level conceptual skills, such as geometric constructions and proofs

References

Barsalou, L. W. (2008). Grounded Cognition. *Annual Review of Psychology , 59*, 617-645.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences , 22*, 577-609.

Clements, D. H., & Battista, M. T. (1992). Geometry and spatial reasoning. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 420-464). New York: Macmillan.

Clements, D. H., & Burns, B. A. (2000). Students' development of strategies for turn and angle measure. *Educational Studies in Mathematics , 41*, 31-45.

Clements, D. H., Battista, M. T., & Sarama, J. (2001). Logo and geometry. *Journal for Research in Mathematics Education Monograph Series , 10*, i-177.

Dehaene, S. (1997). *The number sense: How the mind creates mathematics.* New York: Oxford University Press.

Dehaene, S., & Izard, V. (2006). Core knowledge of geometry in an Amazonian indigene group. *Science , 311* (5759), 381-384.

Dehaene, S., Piazza, M., Pinel, P., & Cohen, L. (2003). Three parietal circuits for number processing. *Cognitive Neuropsychology , 20* (3/4/5/6), 487-506.

Gibbon, J., & Church, R. M. (1981). Time left: Linear versus logarithmic subjective time. *Journal of the Experimental Analysis of Behavior , 7*, 87-107.

Ginsberg, H. P. (1983). *The development of mathematical thinking.* New York: Academic Press.

Halberda, J., Mazzocco, M. M., & Feigenson, L. (2008). Individual differences in non-verbal number acuity correlate with maths achievement. *Nature , 455* (2), 665-668.

Lakoff, G., & Nunez, R. E. (2000). *Where mathematics comes from: How the embodied mind brings mathematics into being.* New York: Basic Books.

Mitchelmore, M. (1998). Young students' concepts of turning and angle. *Cognition and Instruction , 16* (3), 265-284.

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics.* Reston, VA: National Council of Teachers of Mathematics.

Ramsay, J. O., Wickham, H., Graves, S., & Hooker, G. (2009). *fda: Functional Data Analysis.* Retrieved from R package version 2.1.3. http://CRAN.R-project.org/package=fda.

Schwartz, D. L., & Black, J. B. (1996). Shuttling between depictive models and abstract rules: Induction and fallback. *Cognitive Science , 20*, 457-497.

Siegler, R. S., & Booth, J. L. (2004). Development of numerical estimation in young children. *Child Development , 75*, 428-444.

Siegler, R. S., & Opfer, J. E. (2003). The development of numerical estimation: Evidence for multiple representations of numerical quantity. *Psychological Science , 14*, 237-243.

Siegler, R. S., & Ramani, G. B. (2009). Playing linear number board games - but not circular ones - improves low-income preschoolers' numerical understanding. *Journal of Educational Psychology , 101* (3), 545-560.

Spivey, M. (2007). *The continuity of mind.* New York: Oxford University Press.

The Disproportionate Face Inversion Effect in Recognition Memory

Melissa Prince (Melissa.Prince@newcastle.edu.au)
Andrew Heathcote (Andrew.Heathcote@newcastle.edu.au)
School of Psychology, The University of Newcastle
University Drive, Callaghan, 2308, NSW Australia

Abstract

The Disproportionate Face Inversion Effect (DFIE), the finding that inversion disproportionately affects face recognition, provides a primary piece of evidence to suggest that faces are processed in a qualitatively different way to other visual stimuli (i.e., along configural as well as featural dimensions). However, when Loftus, Oberg and Dillon (2004; also Prince and Heathcote, 2009) examined the DFIE using state-trace analysis (Bamber, 1979) they found evidence for a one-dimensional encoding of unfamiliar faces when inversion only occurred during the study phase of a recognition memory task. We further examine this one dimensional result with more precise individual measurement and more specifically, Prince and Heathcote's suggestion that the use of configural encoding may not be automatic in recognition memory.

Keywords: Disproportionate Face Inversion Effect; Recognition Memory; State-trace analysis.

Over the course of a human lifetime, thousands of faces can become so familiar that they can be recognized after only a glance, when seen in an unfamiliar context and even after undergoing significant physical changes (Maurer, LeGrand & Mondloch, 2002). Indeed, the common experience of recognizing a familiar face in a crowd or involuntarily imagining a face in scenic features seems to indicate that humans possess an innate aptitude for face processing. However, this expertise is less evident when the faces are unfamiliar (Hancock, Bruce & Burton, 2000), and even more so when they are viewed upside-down.

It is widely found that perception and memory performance for all mono-oriented stimuli (i.e., objects usually viewed in a specific "upright" orientation) are strongly disadvantaged by inversion; called the *Inversion Effect.* However, in his seminal paper, Yin (1969) observed that this inversion effect was disproportionately stronger for faces compared to mono-oriented control stimuli (e.g., houses) that were matched as closely as possible to faces in terms of familiarity, complexity and difficulty in applying a verbal label; known as the *Disproportionate Face Inversion Effect* (DFIE). Although the inversion effect is taken to indicate there is a general factor affecting the processing of all mono-oriented stimuli, the DFIE suggests there is an additional face specific factor. Hence the DFIE has become one of the primary pieces of evidence to suggest that face processing is "special".

In this paper, we aim to explore the evidence for the DFIE in recognition memory accuracy for unfamiliar faces. In particular, we will focus on an alternate statistical method for testing the effect of inversion called *State-Trace Analysis* (Bamber, 1979). Using this technique, Loftus, Oberg and Dillon (2004) found that, in contrast to results from traditional analyses that revealed a weak DFIE, state-trace results indicated that unfamiliar faces were not special relative to other mono-oriented stimuli when inversion was only manipulated during the encoding stage of a recognition memory task. Loftus et al. therefore suggested that the DFIE only occurs during memory retrieval. Although Prince and Heathcote (2009) replicated this state-trace result, as well as ruling out several potential caveats on Loftus et al.'s methodology and state-trace analyses, they questioned the memory retrieval interpretation. Here we examine an alternate explanation for these results, namely Prince and Heathcote's *Strategic Hypothesis.*

The Disproportionate Face Inversion Effect

Since Yin's (1969) initial demonstration, the DFIE in recognition memory has been replicated numerous times and with various procedural variations. Although many studies have followed Yin's original design where items were studied upright or inverted and tested in the same "matched" orientation, a DFIE has also been found when images were tested using a different viewpoint from study (Valentine & Bruce, 1986) as well as when all images were studied upright but tested upright or inverted (Yarmey, 1971). Consequently, the DFIE has been taken to indicate that face processing is qualitatively different from the processing of other visual stimuli.

It has been suggested that the two factors (or dimensions) underlying the DFIE might be two types of information that can be extracted from the images. The first, *featural information*, is common to all mono-oriented stimuli and refers to the isolated features of an object that can be specified without reference to its other parts. In contrast, the second type, *configural information*, is mostly or only available to faces and enables good discrimination despite the highly similar structure and features that all faces share (McKone & Yovel, 2009). At least three types of configural information have been proposed: (a) *holistic information*, which captures the overall look of a face; (b) *first order relational information*, which refers to the arrangements of features that define a face; and (c) *second order relational information*, which refers to distances between internal features. However, the differences between these sub-types are not of critical importance here. Rather, what is important is the general finding that inversion differentially affects two broad classes of largely independent information.

Although both featural and configural information are affected by inversion, it is typically found that the extraction of configural information is particularly disrupted. Hence, it is often believed that upright faces are processed using both

featural and configural information, whereas only featural information is available for inverted faces. Recent evidence, however, suggests a more graded relationship, such that inversion decreases the rate at which both featural and configural information can be extracted from a face, but to a greater degree for configural information (Valentine & Bruce, 1986).

Identifying Dimensions of the DFIE

Evidence for the DFIE, and hence for the existence of two underlying dimensions for face processing, is traditionally provided by a dissociation quantified by an interaction test comparing the size of the inversion effect for faces (the *face inversion effect*; FIE) to that for a mono-oriented control stimulus, such as houses (the *house inversion effect*; HIE). However, it has been argued that such dissociation logic at best makes the rejection of a one-dimensional account more plausible or parsimonious. Moreover, where response measures are bounded (e.g., accuracy data), interactions may be scale dependent (e.g., influenced by floor and ceiling effects; Loftus, 1978). An alternate method proposed to overcome the caveats on dissociation logic is *State-trace analysis* (e.g., Newell & Dunn, 2008). State-trace analysis provides a rigorous method for determining whether a single dimension (i.e., a single latent variable) is able to explain the joint effect of two or more experimental factors, and assumes only that the mapping between the latent variable and response is monotonic (i.e., that the response and latent variable consistently change in the same direction).

The results from state-trace analysis are assessed using a *state-trace plot*, which is essentially a scatterplot showing the covariation of two factors, namely the *state* and *dimension* factors. As shown in Figure 1, the state factor defines the axes of this plot, while each level of the dimension factor typically defines a set of points within the plot. In particular, the dimension factor is manipulated with the aim of differentially influencing the latent variables. In applications examining the DFIE, the state factor is defined by recognition accuracy for face and house images and the dimension factor manipulated to differentially influence the latent configural dimension is the image orientation.

The crucial diagnostic feature of this plot concerns whether or not the data fall on a single monotonic function; that is, whether the ordering of the x-axis values is the same as the ordering of the y-axis values. At least three data points are required to potentially violate monotonicity, and thus a third factor, called the *trace factor*, is introduced to sweep out a set of points (i.e., a "data trace") within each level of the dimension factor. Importantly, the trace factor must itself have a monotonic effect if we are to unambiguously attribute dimensionality evidence to the interaction between the state and dimension factors (i.e., that A<B and a<b in Figure 1). Loftus et al. (2004), for example, manipulated the study presentation time, which can reasonably be assumed to have a monotonic effect on accuracy; shorter study durations always lead to poor recognition in all conditions (within measurement limits).

If the two levels of the state factor depend on the same underlying dimension, the points on a state-trace plot will fall on a single monotonic function (e.g., in Figure 1a the x- and y-axis order is a,A,b,B). If, however, performance for each state is determined by more than one dimension (e.g., along featural and configural dimensions), the resulting state-trace plot will be non-monotonic (see Figure 1b). It is important to note that although a non-monotonic plot cannot have been produced by a one-dimensional model, the converse does not necessarily hold. Monotonic evidence is only diagnostic of dimensionality when there is overlap of the data traces on at least one axis. Where data-trace overlap fails (such as in Figure 1c), a state-trace plot can be monotonic even if two separate dimensions exist.

Despite the simplicity of state-trace analysis graphically, the best statistical method for testing departures from monotonicity remains an open question (e.g., Newell & Dunn, 2008). Recently Heathcote, Brown and Prince (submitted) proposed a method for assessing dimensionality in state-trace designs based on a Bayes Factor method of selecting amongst models defined by ordinal constraints: namely, (a) a *non-trace* (NT) model, which assumes the trace factor does not have a monotonic effect on performance: that is that the trace model is violated; (b) a *no overlap* (NO) model, which given the trace model holds, assumes the data traces do not overlap and hence cannot be

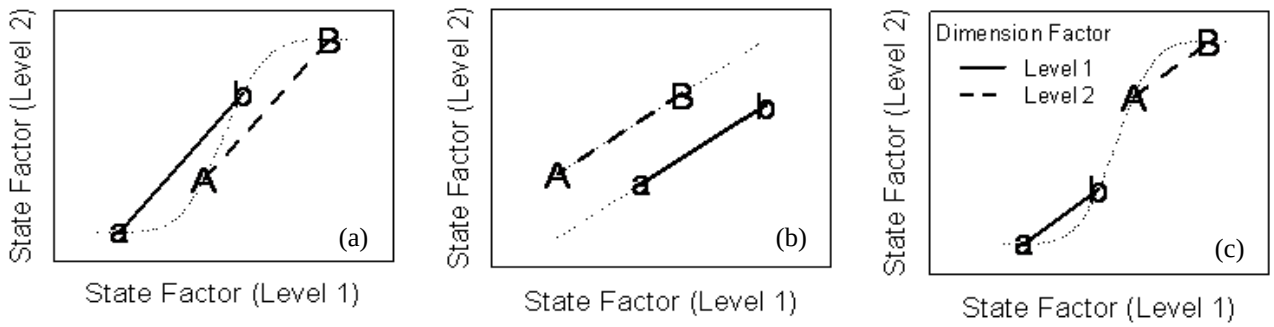


Figure 1: Example state-trace plots for a design where the state, dimension and trace factors each have two levels. The thin dotted lines show the underlying dimension or processes revealed in the plot, with examples of (a) a one-dimensional plot, (b) a two-dimensional plot and (c) a non-diagnostic state-trace plot (i.e., due to the data traces having no overlap)

considered diagnostic of dimensionality; (c) a *uni-dimensional* (UD) model, which assumes the state-trace plot is monotonic, given that both trace monotonicity and data trace overlap hold; and (d) a *multi-dimensional* (MD) model, which assumes the state-trace plot is non-monotonic, again given that both trace monotonicity and data trace overlap hold. Together these four models account for all possible orders. The aim is then to find the model with the highest posterior probability; that is, the model with the highest probability of being the data generating model.

A Memory Retrieval Phenomenon?

Using state-trace analysis, Loftus et al. (2004) reported an apparent exception to the otherwise robust DFIE result; evidence for a single dimension in accuracy averaged over participants in recognition memory for unfamiliar faces (Experiment 1). In contrast, when the faces were famous (i.e., familiar; Experiment 2), they found evidence for more than one dimension. In both experiments, images were studied upright or inverted and all tested upright. This design was utilized to examine Valentine's (1988) assertion that "the orientation of the inspection [study] series does not appear to be critical" (p.474) to produce a DFIE. Loftus et al. concluded that a DFIE would only emerge when inversion was present at the time of memory retrieval, because familiar faces cause memory retrieval at study (and so produce a DFIE when inversion occurs during study), but unfamiliar faces do not (so inversion occurring at study cannot cause a DFIE).

Although, Prince and Heathcote (2009) replicated the one-dimensional state-trace result with unfamiliar faces, the conclusion that a DFIE only occurs at memory retrieval goes against the general opinion in the literature which would suggest the DFIE "is really a perceptual phenomenon rather than a memory phenomenon" (Freire, Lee & Symons, 2000; p.160). Consequently, Prince and Heathcote proposed an alternate explanation more compatible with this perceptual view, whereby participants may be able to strategically use configural information, but only when they know it will improve performance for all items. That is, that the use of configural information may not be automatic in recognition memory.

Here we aim to further examine the one-dimensional state-trace result for unfamiliar faces, as well as Prince and Heathcote's (2009) strategic hypothesis. To do so we ran new experiments that greatly increased the number of observations obtained from each participant (78 observations per design cell), by increasing the number of trials and reducing the number of study durations. Our first condition partly replicated Prince and Heathcote's *Test Upright* design, with both upright and inverted study trials mixed in each study list and all items tested upright. However, it used a two-alternative forced choice (2AFC) recognition memory test, rather than the single-item testing used in the original study (i.e. on each test trial participants chose between a studied and unstudied face, or between a studied and unstudied house). This condition was run to check if Prince and Heathcote's result was replicable with a different testing procedure and with a slightly different, and more powerful, design. We denote this condition *TUM2* (*T*est *U*pright, *M*ixed study lists, *2*AFC).

In *TUM2* (thus also Prince and Heathcote's, 2009, *Test Upright* design), an old item can either be studied and tested upright or studied inverted and tested upright. The former case has a matched (configural) encoding available at study and retrieval. However, when an image is studied inverted it only (or at least mostly) can be encoded using featural information, yet configural information is available from the upright test presentation. As suggested by the encoding specificity effect (i.e., the improvement in memory when study and test conditions match; Tulving & Thomson, 1973), if only featural information was available at study, performance would benefit from a matched (featural) test encoding and be hurt by a mismatched (configural) encoding. Hence it may be detrimental for participants to use configural information when an item had been studied inverted.

In these upright test conditions, upright and inverted items were mixed together at study. Therefore, when all items are presented upright at test, participants have no way of knowing for which test items the use of configural information may be detrimental (i.e., those studied inverted). As these experiments used multiple study-test cycles participants would quickly become aware that all test items were upright. Hence it is possible that they decided to rely purely on featural information, either by not encoding upright study items along a configural dimension, or choosing not to use the configural information available at test. In either case, both faces and houses would only be encoded along a single (featural) dimension, producing the one-dimensional state-trace plots observed by Loftus et al. (2004) and Prince and Heathcote (2009).

To test this possibility, in our second condition, participants viewed two types of study-test lists where: (a) all items were studied and tested upright or (b) all items were studied inverted and tested upright. By blocking study orientation in this manner we hoped that participants would become aware of when configural encoding was advantageous (in type 'a' lists) and hence make use of it. If this occurred, we should observe a multi-dimensional state-trace plot, and hence evidence against Loftus et al.'s (2004) memory retrieval hypothesis. We denote this condition *TUB2* (*T*est *U*pright, *B*locked study lists, *2*AFC).

Method

Participants

The 38 participants were recruited from members of the wider community, who had normal or corrected-to-normal vision. They received cash reimbursement for their time (total AUD$30.00). Two subjects in TUM2 were excluded due to their raw percentage correct falling below 55%, leaving 18 subjects in *TUM2* and *TUB2*.

Stimuli

Stimuli were black and white bitmap images (120 x 105 pixels) displayed at twice their original size. A total of 936 face stimuli were sourced from the FERET database (Phillips, Wechsler, Huang & Rauss, 1998), excluding images with averted gaze, distinctive facial expressions or blemishes (either natural or the result of photographic process). These face stimuli were divided into homogenous blocks based on race, gender and any other distinctive feature (i.e., glasses or facial hair). An additional 36 Caucasian males without facial hair or glasses were included for the practice phase.

A total of 936 house stimuli (with an additional 36 for practice) were sourced using real estate websites and internet search engines. Houses were excluded if located in New South Wales in order to reduce potential familiarity effects given that participants were largely drawn from this region. Following Prince and Heathcote (2009), house stimuli were also divided into homogenous blocks based on their most distinctive feature (e.g., fence, two-storey).

Apparatus

Testing was completed either at individual computer terminals equipped with 17inch LCD monitors or at an external location using laptop computers. All stimuli and text were presented on a black background with white font. Prospective and retrospective confidence judgments were made using the computer keyboard with the keys "z", "x", ".", "/" labeled "1", "2" "3" and "4" respectively.

Procedure

It was emphasized during the instructions for the task, that the orientation of a stimulus was irrelevant to a recognition decision; that is, participants should identify an image as being "old" even if the test item had been studied in a different orientation. In *TUB2*, participants were further informed that study lists would be comprised of either all upright or all inverted images and a warning was displayed prior to each study list indicating the study orientation to be used. Before commencing the main experiment, participants completed two full length practice blocks; one for faces and one for houses, with order counterbalanced over participants.

A study list (comprised of 18 trials) was initiated by pressing the space bar, following which the warning "*Prepare for study … of … Place your fingers on the keys*" was displayed for 2000ms. For each study trial a centrally placed fixation cross was displayed for 1000ms, followed by a 300ms blank screen. The target stimulus was then presented for its designated duration (upright: 33, 100, 267ms; inverted: 267, 800, 2048ms), with durations selected to maximize data trace overlap and each duration level used equally often in every study list. After each study presentation, participants had a maximum of 2500ms to rate their prospective confidence by responding to the question "*How confident are you that you will remember this image later on?*" using a four-point scale from "definitely no" to "definitely yes". The purpose of this prospective confidence judgment was to encourage participants to attend to the stimulus and this data will not be considered further.

The test list (again comprised of 18 trials) was marked by a 300ms blank screen, followed by the warning "*Prepare for test … of … Place your fingers on the keys*" displayed for 2000ms. Each test trial was preceded by a blank screen following which the test item and retrospective confidence response scale were presented for a maximum of 5000ms. For our 2AFC design, a pair of test images (one old and one new, with the old item appearing equally often on the left and right) were presented above the question "*Which image was previously studied and how confident are you that you have seen this image earlier?*" Again participants responded using a four-point scale from "definitely left" to "definitely right". For the entire length of the study and test lists, the words "STUDY" and "TEST" were respectively displayed in the top left corner of the screen.

Following the practice study-test lists, participants received feedback on the number of times they used each of the confidence levels. The purpose of this feedback was to encourage participants to use the full range of the confidence scale.

Participants were required to attend three one hour sessions, preferably on consecutive days. Participants completed 12 study-test lists in their first session and 20 study-test lists in the later two. At the end of each list participants were able to take a self paced break (minimum of 30s), while three longer breaks (minimum of 5min) occurred within each one hour session.

Results

The retrospective confidence rating was used to determine a participant's probability correct (i.e., the number of trials correct divided by total number of trials). Accuracy was then quantified by the inverse cumulative normal probability (*z*) transformation of the probability correct.

We first report a preliminary analysis to ensure the present study was able to replicate previous findings that accuracy is linear as a function of the logarithm of study duration. One-way repeated measures ANOVAs were performed on the effect of the logarithm of study duration for upright and inverted houses and faces in each condition with polynomial trend analyses. Linear trends were all statistically reliable ($p<.05$) and accounted for almost all (minimum 88%) of the variance in accuracy as a function of study duration. The only quadratic trends to approach significance were for *TUM2*'s upright faces ($p=.045$) and upright houses ($p=.075$).

Evidence for the DFIE was first assessed by the traditional test of an interaction between orientation and stimulus type. As the 267ms duration level was the only study duration common to both upright and inverted stimuli, the DFIE was tested by a two-way (orientation by stimulus type) ANOVA using only the 267ms data. Table 1 also shows estimates of the inversion effect (i.e., the difference between upright and inverted) for faces (FIE) and houses

(HIE), the corresponding DFIE estimates (DFIE=FIE-HIE) and the results of associated *t*-tests.

Table 1: Estimates of the FIE, HIE, and DFIE and results associated t-tests, for the 267ms data.

	FIE	HIE	DFIE
TUM2	0.281**	0.270***	0.011
TUB2	0.274***	0.215***	0.059

Note: ***p < .001, **p < .01, *p <.05

For *TUM2* there was no reliable difference in accuracy between houses (*M*=0.479) and faces (*M*=0.411), *p*=.152. However, accuracy was reliably higher for upright items (*M*=0.582) than inverted items (*M*=0.307), *F*(1,17)=30.50, *p*<.001. Although a slightly greater inversion effect was observed for faces than houses (DFIE=0.01), this effect was not statistically reliable, *p*=.91. Similarly for *TUB2*, accuracy was higher for houses (*M*=0.421) than faces (*M*=0.409), but not reliably so, *p*=.83. Upright items were again reliably more accurate (*M*=0.537) than inverted items (*M*=0.293), *F*(1,17)=41.93, *p*<.001. However, there was no reliable DFIE (DFIE=0.059; *p*=.42).

State-trace plots for each condition are shown in Figure 2. Results for upright study are joined, as are the points for inverted study. These lines are clearly monotonically increasing, and consistent with the requirement that the trace factor has a monotonic effect, both conditions' posterior model probabilities favored the trace model being true, p(NT)<.001. The plots also show excellent data trace; for both *TUM2* and *TUB2* p(NO)<.001. In assessing the overall dimensionality, *TUM2* showed positive evidence for a multi-dimensional model, p(MD)=0.910, however, TUB2 showed equivocal evidence suggestive of a one-dimensional account, p(UD)=0.733.

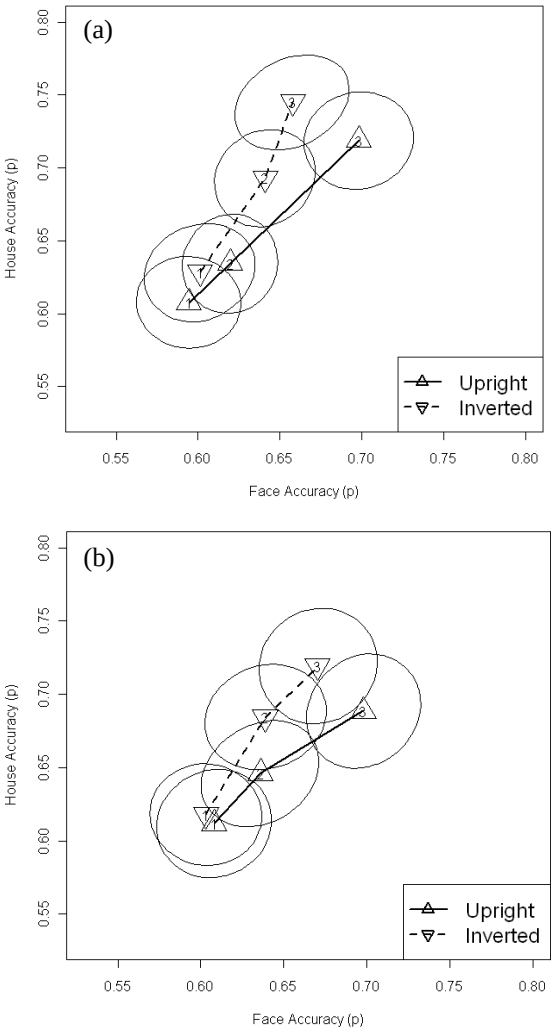


Figure 2: State-trace plots showing the 50% credible regions for the (a) *TUM2* and (b) *TUB2* conditions. The numbers 1…3 indicate shorter to longer study durations.

Discussion

We replicated Loftus et al.'s (2004) and Prince and Heathcote's (2009) finding of a linear increase in accuracy consistent with the suggestion that there was no abrupt change in strategy (i.e., no switch from featural to configural processing) associated with longer study durations. Additionally, we replicated the lack of evidence for a DFIE using the traditional interaction measure (although the DFIE estimates were of the same magnitude as Loftus et al., and Prince & Heathcote). Our state-trace findings, however, were mixed.

We found clear multi-dimensional evidence consistent with the use of both featural and configural information for *TUM2*, where inversion was only manipulated during initial encoding and upright and inverted items were mixed together at study. However, for *TUB2*, where study lists were blocked by orientation, we observed evidence suggestive of a single underlying dimension (although at an equivocal level). In this blocked condition, participants were informed of an item's study orientation if it was old and therefore, according to Prince and Heathcote's (2009) strategic hypothesis, may have been able to reinstate the use of configural information. The observed one-dimensional result, however, does not offer support for this proposal. This is not to say that our results are consistent with Loftus et al.'s (2004) memory retrieval hypothesis. Indeed the strong multi-dimensional result for *TUM2* cannot be explained by a memory retrieval interpretation, as orientation was only manipulated during initial encoding.

It is important to note that the posterior model probabilities on which we are basing our interpretations, are not simply the average result over participants. Rather they examine, for example, the probability that *all* individual state-trace results are one-dimensional versus *all* being multi-dimensional. Hence these probabilities can sometimes be influenced by outlying subjects. To ensure our results were not influenced in this way, we re-examined the dimensionality results, excluding participants with poor evidence (*p*>.5) for trace monotonicity and data trace overlap (four participants from *TUM2* and seven from *TUB2*

were excluded using this criteria). However, both conditions revealed the same pattern of results; that is, multi-dimensional evidence for *TUM2* and equivocal evidence for *TUB2*. Although *TUB2* showed a decrease in the probability supporting a one-dimensional model, p(UD)=0.691.

One possible explanation for observing multi-dimensional evidence, even though inversion was only manipulated during the initial stimulus encoding, is that our more precise individual measurement also produced higher accuracy performance overall and consequently an improved effect size. Although state-trace analysis is not affected by floor and ceiling effects to the same degree as traditional dissociation analyses, if accuracy is not high enough to reveal the decrement caused by inversion then it will also not be able to reveal the underlying dimensionality. Consistent with this suggestion, we can observe from the *TUM2* state-trace plot that the data traces do not depart from monotonicity (indicating multi-dimensional evidence) until the longer study duration levels (where accuracy is also higher). This same pattern can also be seen to a lesser degree in *TUB2*.

Although not reported here, we also ran these same mixed and blocked conditions using a yes/no testing procedure (i.e., participants were shown a single test item and asked to indicate if that item had or had not been studied), and in contrast to our 2AFC results, observed equivocal one-dimensional evidence. Interestingly, it has been found that memory performance is advantaged by a 2AFC procedure over a yes/no procedure (Deffenbacher, Leu & Brown, 1981), which could explain why our 2AFC conditions tended toward multi-dimensional evidence. It should also be noted that recognition memory studies in general tend to show a smaller inversion effect than perceptually based studies (e.g., *TUM2* showed a 9.97% drop in accuracy, but perceptual tasks can show a drop double this magnitude; see McKone & Yovel, 2009). Hence evidence for more than one dimension underlying face processing may only emerge when performance is high enough to reveal the decrement caused by stimulus inversion.

We will pursue two avenues in future research. First, as our results did not offer strong insight into Loftus et al.'s (2004) memory retrieval hypothesis we will examine state-trace evidence for the DFIE in recognition memory using a paradigm in which unfamiliar faces are all studied upright and tested either upright or inverted. In this paradigm, Loftus et al.'s (2004) memory hypothesis predicts that evidence for multiple dimensions should emerge, because inversion occurs at test where memory retrieval is required. Second, we will extend the use of state-trace analysis to a perceptual paradigm, such as a sequential same-different task, in order to investigate whether evidence for more than one dimension emerges with larger inversion effects.

Acknowledgments

Thanks to students enrolled in Psyc4815/Psyc6781 at the University of Newcastle, 2009 for assistance with stimulus preparation and running of participants.

References

Bamber, D. (1979). State-trace analysis: A method of testing simple theories of causation. *Journal of Mathematical Psychology, 19,* 137-181.

Deffenbacher, K.A., Leu, J.R., & Brown, E.L. (1981). Memory for faces: Testing method, encoding strategy, & confidence. *American Journal of Psychology, 94,* 13-26.

Freire, A., Lee, K., & Symons, L.A. (2000). The face inversion effect as a deficit in the encoding of configural information: Direct evidence. *Perception, 29,* 159-170.

Hancock, P.J.B., Bruce, V., & Burton, A.M. (2000). Recognition of unfamiliar faces. *Trends in Cognitive Science, 4,* 330-337.

Heathcote, A., Brown, S. & Prince, M. (submitted). The design and analysis of state-trace experiments.

Loftus, G.R. (1978). On interpretation of interactions. *Memory & Cognition, 6,* 312-356.

Loftus, G.R., Oberg, M.A., & Dillon, A.M. (2004). Linear theory, dimensional theory, and the face-inversion effect. *Psychological Review, 111,* 835-863.

Maurer, D., Le Grand, R., & Mondloch, C.J. (2002). The many faces of configural processing. *Trends in Cognitive Science, 6,* 255-260.

McKone, E., & Yovel, G. (2009). Why does picture-plane inversion sometimes dissociate perception of features and spacing in faces, and sometimes not? Toward a new theory of holistic processing. *Psychonomic Bulletin & Review, 16,* 778-797.

Newell, B.R., & Dunn, J.C. (2008). Dimensions in data: Testing psychological models using state-trace analysis. *Trends in Cognitive Sciences, 12,* 285-290.

Phillips, P.J., Wechsler, H., Huang, J., & Rauss, P.J. (1998). The FERET database and evaluation procedure for face recognition algorithms. *Image and Vision Computing, 16,* 295-306.

Prince, M., & Heathcote, A. (2009). State-trace analysis of the face-inversion effect. In N.A. Taagen & H. van Rijn (Eds.). *Proceedings of the 31st Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Tulving, E., & Thomson, D.M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review, 80,* 352-373.

Valentine, T. (1988). Upside-down faces: A review of the effect of inversion upon face recognition. *Acta Psychologica, 79,* 471-491.

Valentine, T., & Bruce, V. (1986). The effect of race, inversion and encoding activity upon face recognition. *Acta Psychologica, 61,* 259-273.

Yarmey, A.D. (1971). Recognition memory for familiar "public" faces: Effects of orientation and delay. *Psychonomic Science, 24,* 286-288.

Yin, R.K. (1969). Looking at upside-down faces. *Journal of Experimental Psychology, 81,* 141-145.

Decomposing Externally Cued Task Switching Costs

Christina Wasylyshyn (christina.wasylyshyn@nrl.navy.mil)
Naval Research Laboratory, 4555 Overlook Ave., S.W.
Washington, DC 20375 USA

Abstract

The double-cued task switching procedure has recently been introduced as a new way to measure externally cued switch costs. In the present individual differences study, two hundred fifty young adults completed measures of task switching, inhibition, and long-term memory. A latent variable approach was taken to examine the relationships among these cognitive measures. Decomposing the externally cued task switching costs into a cue switch component and a task switch component indicated that individual differences in these costs could be explained by benefits of repeated cues rather than by changes in tasks. Individual differences in the cue switch component were predicted by long-term memory scores.

Keywords: cued task switching; switch costs; individual differences; long-term memory; inhibition

Introduction

Recent interest in understanding how people shift their mental sets in response to external cues has led to the development of a new method of measuring task switching costs: the double-cued procedure. In traditional, single-cued procedures, cues and tasks are mapped one-to-one, leading to the possibility that the components of cue switching and task switching are confounded. When there is a change in cue, there must be a change in task; when there is no change in cue, it follows that the task will repeat from the previous trial. Both cue switching and task switching then contribute to the overall switch cost in an undifferentiated manner.

One way to distinguish between cue switching and task switching is to use a double-cued procedure, that is, to have two cues to indicate each task. This leads to three types of trials: cue repeat, cue switch, and task switch. In cue repeat trials, both the cue and the task repeat; this is a traditional nonswitch trial. In cue switch trials, the cue changes but the task remains the same; this is also classified as a nonswitch trial. In task switch trials, both the cue and the task change. The latency differences between cue switch and cue repeat trials are thought to indicate encoding benefits from repeated cues. The latency differences between task switch and cue switch trials are thought to reflect the act of task switching.

In task switching paradigms, responses to the current stimulus trial are slower (and typically less accurate) if the task differs from that completed on the previous trial. Switch costs are thought to indicate how flexibly one can change his/her cognitive configurations, or task sets, to accommodate newly relevant task demands. In order to establish a task set, one must activate relevant task rules (e.g., Mayr & Kliegl, 2003; Rubinstein, Meyer, & Evans, 2001) and minimize interference from competing task sets, possibly through inhibition processes (e.g., see Mayr & Keele, 2000). It has therefore been suggested that one's task set reflects the interaction of task set inertia from previous trial(s), exogenous task set activation from the stimulus itself, and endogenous control input needed to overcome the other two biases and to reconfigure the cognitive system for a change in task (Aron, Monsell, Sahakian, & Robbins, 2004).

Mayr and Kliegl (2003) suggest that performance in double-cued procedures can be explained by two processes: cue-based retrieval of task rules from memory and the application of task rules to the target. The retrieval of task rules produces the cue encoding benefit, while the time involved in applying the mapping rules produces the actual switch cost. Retrieval and loading of task rules from long-term memory is necessary for both nonswitch and switch trials. A repetition of the immediately preceding cue leads to a reactivation of the most recent retrieval process (i.e., positive priming); a cue change, however, requires activation of a new (or less recently activated) retrieval process.

Logan and Bundesen (2003) offer a similar explanation, but one that assumes there is no endogenous control component. The explicit task cue provides enough information to uniquely indicate a response on each trial. There is no task set reconfiguration process between the cue and the target stimulus. Instead, any switch costs that remain beyond the act of cue switching are the result of encoding benefits on nonswitch trials, or priming from related cues, not task switching (see Logan & Bundesen, 2004, for explanation of the process of mediator repetition, and see Schneider & Logan, 2005, for formal model).

Using double-cued procedures, several studies have shown evidence for cue encoding benefits (i.e., responses were faster for cue repeat trials than for cue switch trials). However, after cue encoding effects have been accounted for, the remaining task switch costs have been negligible in some studies (e.g., Logan & Bundesen, 2003, 2004; Monsell & Mizon, 2006 Experiment 1) and substantial in others (e.g., Mayr & Kliegl, 2003; Arrington & Logan, 2005 Experiment 3; Monsell & Mizon, 2006 Experiments 2-6). There are several procedural differences between these studies that may explain some of the differences in results, including type of task cue and frequencies for task switches. Evidence that switch costs result from the processing of the task cue rather than from the switching of tasks has been shown in studies that use salient verbal or pictorial task cues and/or 50% task switch frequency. Evidence for substantial task switch costs over and above any effect of a cue change

has been shown in studies that employ arbitrary task cues and/or 33% or less task switch frequency.

Most of the research in task switching and executive control functioning has been experimental in nature. The present study, however, takes a novel individual differences approach to decompose switch costs to determine whether a cue switch component can be differentiated from a task switch component. The overall goal is to establish a representation of the structure of individual differences in the double-cued procedure to determine whether switch costs are more likely to reflect processes involved in the interpretation of instructional cues (i.e., trial to trial change in retrieval path) or the switching of task sets (i.e., trial to trial change in the task itself). In this way, it will be possible to test, at the latent level, if individual differences in the costs incurred reflect a benefit for cue repetition instead of, or as well as, a cost for task switching.

Method

Participants

Two hundred fifty Syracuse University students (169 females, 81 males, mean age = 18.92, SD = 1.21) participated. All students were native English speakers and non-colorblind.

Tasks

There were three task categories: task switching, long-term memory, and inhibition.

Task Switching Switch costs were measured in three task domains: digits, shapes, and verbal. For the digits task, magnitude and parity judgments were made on a series of digits (1, 2, 3, 4, 6, 7, 8, 9). In the shapes task, participants determined either the form or color of an image. Stimuli consisted of combinations of two shapes and two colors. For the verbal task, participants determined if a word shown could be classified as an animal or a non-animal or if it could be classified as something that was smaller or larger than a basketball. Stimuli consisted of 64 high frequency and high imagery nouns obtained from the Toronto Word Pool.

Each task consisted of 120 trials. Cues and targets were combined randomly, with the constraint that cue-repeat, cue-switch, and task-switch trials each occurred on one-third of the trials. In each task, there were four cues. Two meaningful and salient cues distinguished each sub-task, so that one cue noted category membership and the other noted response mapping. In the digits task, participants were presented with one of four cues on each trial: Magnitude, High-Low, Parity, or Odd-Even. Depending on the cue shown, participants pressed the 'z' key if the target digit was higher than five or odd, and the '/' key if the target digit was lower than five or even. In the shapes task, one of four cues was presented on each trial: Shape, Triangle-Circle, Color, or Red-Green. Participants pressed the 'z' key if the target image was a triangle or was colored red, and the '/' key if the target image was a circle or was colored green. In the verbal task, either the cue Creature, Animal-Nonanimal, Size, or Smaller-Larger was presented on each trial. Participants pressed the 'z' key if the target word was an animal or was smaller than a basketball, and the '/' key if the target word was a non-animal or larger than a basketball. On cue repeat trials, both the cue and the task repeated from trial *n* to trial *n+1* (e.g., in the digits task, magnitude followed by magnitude). On cue switch trials, the cue changed but the task repeated (e.g., magnitude followed by high-low). On task switch trials, both the cue and the task changed (e.g., magnitude followed by parity). Participants were given 150 ms for preparation during the cue-stimulus interval and 300 ms for passive dissipation during the response-cue interval.

Inhibition In the flanker task (Eriksen & Eriksen, 1974), participants were asked to identify a centrally presented target letter (either an 'S,' 'C,' 'H,' or 'K'). This target letter was either presented alone or with three noise letters flanking it on each side. Participants pressed the 'z' key when the target letter was S or C, and the '/' key when the target letter was H or K, as quickly as possible. There were four stimulus conditions: (1) no noise (e.g., S), (2) noise same as target (e.g., SSSSSSS), (3) noise response compatible (e.g., SSSCSSS), and (4) noise response incompatible (e.g., SSSKSSS). After completing 32 practice trials, participants completed 160 trials (4 blocks of 40 trials). Trials began with a 500 millisecond fixation cross presented in the center of the screen, followed by a blank screen for 50 milliseconds. The stimulus was then presented until a response was made. The latency difference between the noise response incompatible condition and the no noise condition served as the dependent measure.

Long-Term Memory Participants were asked to learn a list of 30 words. Words were presented one at a time in the center of the screen at a rate of 3 seconds each. After approximately 15 minutes, participants were given a recognition test of all 30 words randomly mixed with 30 foil words. Participants pressed the 'z' key if the word was part of the original study list and the '/' key if the word was not presented in the original list. Stimuli were obtained from the Toronto Word Pool and consisted of highly familiar 1 syllable words that were 3, 4, or 5 letters in length. Scores were obtained using a nonparametric form of the discriminability index (i.e., a′), and participants with an a′ score below .5 were excluded from analysis.

Results

Descriptive Statistics

Descriptive statistics are presented in Table 1. All of the measures meet the criteria for univariate normality (Kline, 1998); skews are all less than 3 and kurtosis values are all less than 4. All measures therefore displayed adequate distributional properties for being subjected to latent variable analysis.

Table 1: Descriptive statistics

Variable	M	n	SD	Skew	Kurtosis
Digits					
CR	808.72	241	118.72	0.16	-0.41
CS	1204.26	241	251.99	0.48	0.05
TS	1299.25	241	290.04	0.51	0.49
Shapes					
CR	797.77	245	138.43	0.39	-0.42
CS	1055.56	245	190.16	0.41	0.23
TS	1135.71	245	211.97	0.44	-0.13
Verbal					
CR	934.63	245	149.29	0.44	-0.08
CS	1286.99	245	242.03	0.38	-0.38
TS	1406.19	245	268.89	0.28	-0.16
Delayed Word Recognition	0.83	248	0.08	-0.91	1.29
Flanker	85.39	250	59.92	0.23	0.79

Note: CR = cue repeat trial, CS = cue switch trial, TS = task switch trial

Cue Switching vs. Task Switching: Accounting for Individual Differences in Switch Costs

Figure 1 presents the mean encoding costs and task switching costs (errors bars indicate SE) by task domain (digits, shapes, verbal). Encoding costs were calculated as latency differences between cue switch and cue repeat trials, and task switch costs were computed as latency differences between task switch and cue switch trials. Across task domains (please refer to Table 1), cue repeat trials were the fastest (M = 847 ms), followed by cue switch trials (M = 1182 ms), and finally task switch trials (M = 1280 ms). Cue switch trials were more like task switch trials than cue repeat trials, suggesting that cue repetition effects account for most of the switch cost. Indeed, the majority (77%) of the switch cost is accounted for by the cost of encoding the cue (335 ms cost). However, there is an overall mean residual task switching cost of 98 ms. Planned contrasts indicated that task switch trials were significantly slower than cue switch trials, $t = 8.04$, $p < 0.0001$. Therefore, from the means analysis, we can conclude that there is an effect of task switching. Partial correlations between cue repeat and cue switch trials, controlling for performance on task switch trials, were also computed. Controlling for task switch trials did not significantly attenuate the relationship between cue repeat and cue switch trials in any of the task domains; this suggested that the residual effect of task switching might not be useful as an individual differences variable. The next goal was to use Structural Equation Modeling to test for reliable individual differences.

Two models were contrasted to test whether, after accounting for cue encoding, the process of cue-switching is sufficient to explain individual differences in switch costs, or whether an additional task switching process is needed to fully account for these costs. The former will be called the 2-factor model, and the later will be referred to as the 3-factor model. In both models, the manifest measures are the RTs from the cue repeat, cue switch, and task switch trials from the digits, shapes, and verbal tasks. Residual variances of trial type measures employing the same task domain were correlated.

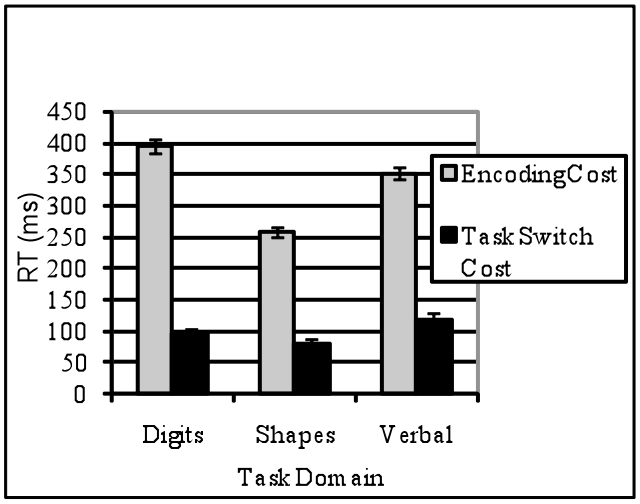


Figure 1: Mean encoding costs and task switching costs (± SE) by task domain (digits, shapes, verbal)

Model fit was assessed using the chi-square test for goodness of fit, the comparative fit index (CFI; Bentler, 1990), the Tucker-Lewis index (TLI; Tucker & Lewis, 1973), and the root mean square error of approximation (RMSEA; Browne & Cudek, 1992). Acceptable model fit is reflected by a nonsignificant chi-square test for goodness of fit, relative fit indices (i.e., CFI and TLI) above .90, and a RMSEA value below .08 (Bentler, 1990; Bentler & Bonnet, 1980). A RMSEA value below .05 indicates excellent fit. Analyses were conducted with AMOS 5 software (Arbuckle, 2003) using maximum likelihood estimation. For all of the SEM models, the factor loadings and interfactor correlations were allowed to vary (Anderson & Gerbing, 1988).

In order to capture what was common, all nine measures were free to load on the first factor, that is, an encoding baseline, because all types of trials (cue repeat, cue switch, and task switch) involve encoding the external cue and thus should have systematic differences in the encoding process. A 1-factor encoding model, however, did not sufficiently account for individual differences in switch costs, $\chi^2(18) = 108.88$, $p = .000$, TLI = .881, CFI = .952, RMSEA = .142. The second factor, cue switch, was then introduced to reflect the systematic individual differences associated only with switching a cue; it represents the common variance of the cue switch and task switch trials, once baseline encoding has been accounted for. The 2-factor model (please see

Figure 2; note that the correlations among the residual variances are not included in the figure for ease of interpretation) provided an acceptable fit to the data, $\chi^2(12) = 22.76$, $p = .030$, TLI = .979, CFI = .994, RMSEA = .060. In Figure 2, the larger circles represent the latent variables, and the rectangles represent the scores on the individual indicator tasks that were used to measure each of the latent variables. The smaller circles represent the residual variances of the indictor tasks.

A third factor, task switch, was then introduced to determine if a task switching process, in addition to encoding baseline and cue switching, could better explain individual differences in switch costs. This third factor was equal to the residual common variance of task switch trials only. The 3-factor model (please see Figure 3) provided an excellent fit to the data, $\chi^2(9) = 12.62$, $p = .181$, TLI = .991, CFI = .998, RMSEA = .040. At first glance, this suggests that there is an effect of task switching, over and beyond the processes of cue encoding and cue switching.

Because the 2-factor and 3-factor models are not nested, the Akaike Information Criterion (AIC) and the Expected Cross-Validation Index (ECVI) were used to compare overall model fit (Note: unlike the chi-square difference test, these indices do not provide a statistical comparison of competing models). In general, a model that has the lowest AIC and ECVI values is judged to fit the data better than the alternative model(s) tested (Brown, 2006). For the 2-factor model, AIC = 106.76 and ECVI = 0.429; for the 3-factor model, AIC = 102.62 and ECVI = 0.412. In terms of the AIC and ECVI indices, the 3-factor model best fits the data, however, the 3-factor model was deemed unacceptable due to non-interpretable and ill-fitting parameter estimates of the task switch factor. The variance of the task switch factor was not significant, $p = 0.272$. Only the path loading for the task switch trials in the Shapes task to the task switch factor was significant ($\beta = 0.18$, $p = 0.015$); the path loadings for the Digits and Verbal tasks were not significant.

Therefore, the 2-factor model was accepted. The lower (i.e., better fitting) AIC and ECVI values for the 3-factor model seem to simply reflect adjustment for model complexity compared to the 2-factor model, so that the 3-factor model does not account for any additional systematic differences in switch costs. The act of task switching does not provide reliable information that was not already available from cue switching.

Using Cognitive Measures to Explain Individual Differences in Task Switching

Further support for this claim comes from additional modeling of possible cognitive predictors (inhibition and long-term memory) that might serve to explain some of the systematic differences in each of the latent factors. Please see Table 2 for standardized effects and model fits. It should be noted that because only single indicator predictors are being employed, the following effects are small, but significant. Individual differences in inhibition significantly predict individual differences in encoding, but not cue switching or task switching. That is, individuals who are quicker at encoding the present, most relevant information are also faster at inhibiting previous and/or distracting information. Individual differences in long-term memory

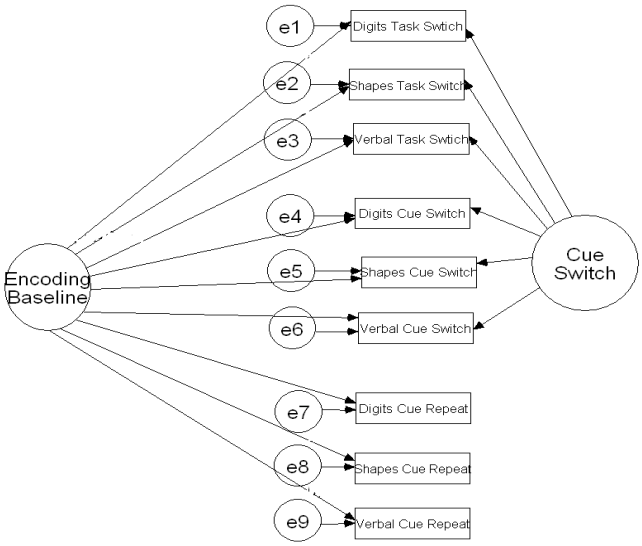


Figure 2: 2-Factor latent variable model to account for switch costs

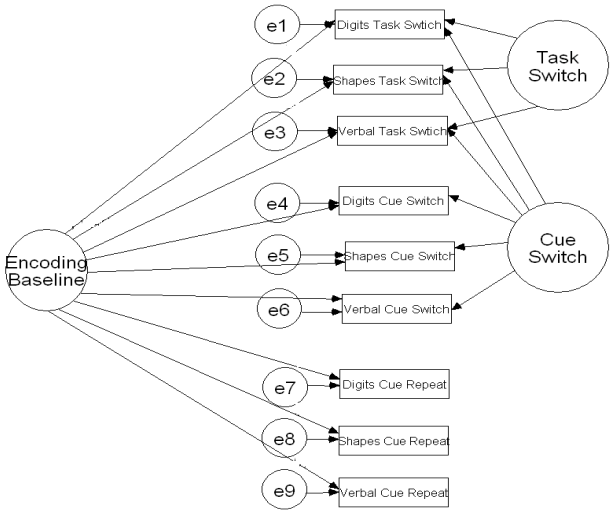


Figure 3: 3-Factor latent variable model to account for switch costs

significantly predict individual differences in cue switching, but not encoding or task switching. That is, individuals who are able to efficiently switch between trials that require a change in information, or change in instructional cue, have higher long-term memory scores.

Table 2: Standardized effects of inhibition and long-term memory on the task switching factor components and model fit statistics

	Inhibition		Long-Term Memory	
	β	p	β	p
Encoding Baseline	**.203***	**.005***	-.039	.586
Cue Switch	-.099	.274	**-.187***	**.036***
Task Switch	.109	.375	-.132	.412
χ^2	22.339		17.37	
df	15		15	
p	.099		.297	
CFI	.996		.999	
TLI	.986		.995	
RMSEA	.044		.025	

Discussion

The overall goal of the present study was to establish a representation of the structure of individual differences in the double-cued procedure. There were two specific aims. The first aim was to examine if individual differences in performance on cue switch and task switch trials could be distinguished at the level of latent variables to determine whether switch costs reflect processes involved in interpreting instructional cues rather than, or in addition to, switching task sets. The second aim was to examine the relationships of the decomposed costs with measures of long-term memory and inhibition to determine the underlying mechanisms or processes that might explain some of the variance in each of the components.

In the aggregate trial-type RT analysis, cue switch trials (M = 1182 ms) appeared more like task switch trials (M = 1280 ms), than like cue repeat trials (M = 847 ms). Although 77% of the switch cost was accounted for by encoding the cue (335 ms), the mean overall residual task switching cost of 98 ms was significant. That is, average performance on task switch trials was significantly slower than average performance on cue switch trials. Using partial correlation, it was found that the first-order correlation between cue repeat and cue switch trials in the digits task was reduced from .76 to .40 after controlling for performance on task switch trials. The correlation between cue repeat and cue switch trials in the shapes task reduced from .80 to .52, and the correlation for the verbal task cue repeat and cue switch trials reduced from .76 to .35. This implied that 63%, 72%, and 59% of the variance shared between the cue repeat and cue switch trials in the digits, shapes, and verbal tasks, respectively, was associated with performance on task switch trials. However, controlling for task switch trials did not significantly attenuate the relationship between cue repeat and cue switch trials in any of the task domains, suggesting that the residual effect of task switching might not be useful as an individual differences variable.

The results of the present study showed that task switching did not serve as a reliable individual differences variable; task switch trials did not provide any additional information that was not already accounted for by the cue switch and cue repeat trials. The residual common variance for the task switch factor was not significant, lending support to the claim made by Logan and Bundesen (2003) that efficient performance does not require an actual act of task switching. It should be noted that this claim can only be made for externally cued paradigms that employ short preparation intervals, as this study only used one interval. Yehene and Meiran (2007) suggest that this may not be the case at longer preparation intervals. However, it should also be noted that in an individual differences study, Friedman and Miyake (2004) could not distinguish switch costs incurred at short preparation intervals from those incurred at longer preparation intervals at the level of latent variables.

That there were no reliable individual differences to account for the act of task switching cannot be attributed to participants' preparatory strategies in response to a high probability of a task switch trial. Recent studies (e.g., Schneider & Logan, 2006; Monsell & Mizon, 2006) have indicated that the frequency of switch trials is related to the magnitude of switch costs, so that the higher the probability of the occurrence of a task switch trial, the smaller the overall switch cost. In the present study, the overall probability of a task switch, *p*(task switch), was 0.33, and the probability of a task switch given a cue switch, *p*(task switch|cue switch) was 0.5. In the Logan and Bundesen (2003) studies, *p*(task switch) = 0.5, and *p*(task switch|cue switch) = 0.67. Unlike the Logan and Bundesen experiments, the present study can rule out the possibility that participants might have strategically controlled their task-set readiness as a function of expectation for a task switch trial, thereby reducing their overall switch costs. Moreover, other procedural precautions were taken in the design of the current study, as suggested by Monsell and Mizon (2006), to capture an endogenous control process, or actual act of task switching, if there was one. For example, the response-stimulus interval was kept constant to avoid confounding active preparation with passive decay, and highly salient cues were used. Finally, the present study can

make the claim that increasing the number of target stimuli from 4 in the shapes task, to 8 in the digits task, to 64 in the verbal task did not lead to a task switching effect; participants did not resort to switching task sets in response to the cue as the mapping combinations between cues, targets, and responses got larger.

Conclusions

In summary, individual differences in switch costs were attributed to changes in cue initiated retrieval; switch costs were a consequence of cue priming effects, not a consequence of task changes. The further modeling of cognitive measures to predict individual differences in the component factors indicated that the single inhibition measure was associated with individual differences in the encoding baseline factor, and the single long-term memory measure was related to individual differences in the cue switching factor. It should be noted that although these effects were small, they were theoretically grounded. These results lend support to the claim that the loading of task rules from long-term memory was necessary even on nonswitch trials (Mayr & Kliegl, 2003). Because only one preparation interval was included, the reduction in the switch cost effect across increasing preparation intervals cannot be measured. Therefore, it is not possible to completely rule out an endogenous task set reconfiguration process. The present study can, however, conclude that at short preparation intervals, reliable variance in switch costs could be explained by a cue repetition benefit; an additional task switching process was not needed to fully account for individual differences.

References

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin, 103,* 411-423.

Arbuckle, J. L. (2003). Amos (Version 5.0) [Computer software]. Spring House, PA: Amos Development Corporation.

Aron, A. R., Monsell, S., Sahakian, B. J., & Robbins, T. W. (2004). A componential analysis of task-switching deficits associated with lesions of left and right frontal cortex. *Brain: A Journal of Neurology, 127(7),* 1561-1573.

Arrington, C. M., & Logan, G. D. (2005). Voluntary task switching: Chasing the elusive homunculus. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 31(4),* 683-702.

Bentler, P. M. (1990). Comparative fit indices in structural models. *Psychological Bulletin, 23,* 611-626.

Bentler, P. M., & Bonnet, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88,* 588-606.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York: Guilford Press.

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods and Research, 21,* 230-258.

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics, 16,* 143-149.

Friedman, N. P., & Miyake, A. (2004). The relations among inhibition and interference control functions: A latent-variable analysis. *Journal of Experimental Psychology: General, 133(1),* 101-135.

Klein, R. B. (1998). *Principles and practice of structural equation modeling*. New York: Guilford Press.

Logan, G. D., & Bundesen, C. (2003). Clever homunculus: Is there an endogenous act of control in the explicit task-cuing procedure? *Journal of Experimental Psychology: Human Perception & Performance, 29(3),* 575-599.

Logan, G. D., & Bundesen, C. (2004). Very clever homunculus: Compound stimulus strategies for the explicit task-cuing procedure. *Psychonomic Bulletin & Review, 11(5),* 832-840.

Mayr, U., & Keele, S. W. (2000). Changing internal constraints on action: The role of backward inhibition. *Journal of Experimental Psychology: General, 129(1),* 4-26.

Mayr, U., & Kliegl, R. (2003). Differential effects of cue changes and task changes on task-set selection costs. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 29(3),* 362-372.

Monsell, S., & Mizon, G. A. (2006). Can the task-cuing paradigm measure an endogenous task-set reconfiguration process? *Journal of Experimental Psychology: Human Perception and Performance, 32(3),* 493-516.

Rubinstein, J. S., Meyer, D. E., & Evans, J. E. (2001). Executive control of cognitive processes in task switching. *Journal of Experimental Psychology: Human Perception and Performance, 27,* 763-797.

Schneider, D. W., & Logan, G. D. (2005). Modeling task switching without switching tasks: A short-term priming account of explicitly cued performance. *Journal of Experimental Psychology: General, 134(3),* 343-367.

Schneider, D. W., & Logan, G. D. (2006). Priming cue encoding by manipulating transition frequency in explicitly cued task switching. *Psychonomic Bulletin & Review, 13(1),* 145-151.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38(1),* 1-10.

Yehene, E., & Meiran, N. (2007). Is there a general task switching ability? *Acta Psychologica, 126(3),* 169-195.

Deconfounding Hypothesis Generation and Evaluation in Bayesian Models

Elizabeth Baraff Bonawitz (liz_b@berkeley.edu)
Department of Psychology, 5427 Tolman Hall
Berkeley, CA 94720 USA

Thomas L. Griffiths (tom_griffiths@berkeley.edu)
Department of Psychology, 3210 Tolman Hall
Berkeley, CA 94720 USA

Abstract

Bayesian models of cognition are typically used to describe human learning and inference at the computational level, identifying which hypotheses people should select to explain observed data given a particular set of inductive biases. However, such an analysis can be consistent with human behavior even if people are not actually carrying out exact Bayesian inference. We analyze a simple algorithm by which people might be approximating Bayesian inference, in which a limited set of hypotheses are generated and then evaluated using Bayes' rule. Our mathematical results indicate that a purely computational-level analysis of learners using this algorithm would confound the distinct processes of hypothesis generation and hypothesis evaluation. We use a causal learning experiment to establish empirically that the processes of generation and evaluation can be distinguished in human learners, demonstrating the importance of recognizing this distinction when interpreting Bayesian models.

Keywords: Approximate Bayesian Inference; Hypothesis Generation; Hypothesis Evaluation; Causal Learning

Introduction

Learning causal relationships, categories, and languages all require solving challenging inductive problems, using limited data to assess underdetermined hypotheses. In the last decade an increasing number of papers have argued that people solving inductive problems act in ways that are consistent with optimal Bayesian inference (e.g., Griffiths & Tenenbaum, 2005; Goodman, Tenenbaum, Feldman, & Griffiths, 2008; Xu & Tenenbaum, 2007). However, most of these analyses operate at what Marr (1982) termed the *computational* level, using Bayesian inference to identify the hypotheses that an ideal learner with particular inductive biases would choose to explain the observed data. An important question for this approach is what learners are doing at the *algorithmic* level: identifying the psychological processes by which learners solve inductive problems, and understanding how these algorithms connect back to the computational level.

Connecting the algorithmic and computational levels involves two challenges: identifying algorithms that can produce behavior consistent with Bayesian inference, and determining how the assumptions of a computational-level analysis relate to the components of these algorithms. In this paper, we take up these two challenges for one class of algorithms for inductive inference. The most naïve translation of Bayesian inference into an algorithm for inductive inference would be to assume that learners implement Bayes' rule directly, having a fixed set of hypotheses and updating a probability distribution over all of those hypotheses simultaneously as data are observed. However, the assumption that learners possess all relevant hypotheses before seeing data is at odds with numerous findings suggesting that generating appropriate hypotheses can be one of the hardest parts of inductive inference (e.g., Kuhn, 1989; Klahr, Fay, & Dunbar, 1993). We thus consider the consequences of separating the processes of generating hypotheses and evaluating those hypotheses, assuming that learners perform Bayesian inference with only the set of hypotheses they generate.

To investigate this, we present a mathematical analysis of a simple algorithm in which hypothesis generation and evaluation are separated. This produces a surprising result: This algorithm results in behavior that can still be analyzed in terms of Bayesian inference, but with a prior that conflates the plausibility of a hypothesis with the ease of generating that hypothesis. This result suggests that we should be cautious when interpreting the priors of Bayesian models estimated from behavioral data. Such priors will always reflect the inductive biases of human learners – those factors that lead people to select one hypothesis over another when both are equally consistent with the data. However, human inductive biases can include components that result from processes at the algorithmic level, such as generating hypotheses.

To demonstrate the importance of taking into account algorithmic-level factors in interpreting Bayesian models, we present an experiment exploring the separability of hypothesis generation and evaluation. In the task, we conduct a causal learning experiment in which we manipulate the hypotheses that people generate: by "priming" an appropriate hypothesis, we increase the probability of people producing responses consistent with that hypothesis; however, when we employ a more standard Bayesian reasoning task, providing a set of hypotheses and asking participants to evaluate them, the effect of priming goes away. A computational-level analysis would require postulating different prior distributions in order to explain behavior on these two components of the task. However, an algorithmic-level analysis shows that this difference can be explained as the result of the separate effects of hypothesis generation and evaluation. Finally, we discuss the implications of this work for future models of human cognition and for studies of developmental changes.

Analyzing inductive inferences

Bayesian inference indicates how a rational learner should change his or her beliefs about a set of hypotheses in light

of observed data. Let h be a hypothesis belonging to a set of hypotheses $\mathcal{H}$. Assume that the learner has different degrees of belief in the truth of these hypotheses, and that these degrees of belief are reflected in a probability distribution $p(h)$, known as the *prior*. Then, the degrees of belief the learner should assign to each hypothesis after observing data d are given by the *posterior* probability distribution $p(h|d)$ obtained by applying Bayes' rule

$$p(h|d) = \frac{p(d|h)p(h)}{\sum_{h' \in \mathcal{H}} p(d|h')p(h')} \quad (1)$$

where $p(d|h)$ indicates the probability of observing d if h were true, and is known as the *likelihood*.

Bayes' rule provides a computational-level theory of inductive inference, being a component of the optimal solutions to a variety of problems of reasoning under uncertainty (Anderson, 1990; Anderson & Schooler, 1991; Freeman, 1994; Geisler, Perry, Super, & Gallogly, 2001; Griffiths & Tenenbaum, 2007; Huber, Shiffrin, Lyle, & Ruys, 2001; Knill & Richards, 1996; Körding & Wolpert, 2004; Shiffrin & Steyvers, 1997; Weiss, Simonvelli, & Adelson, 2002). As an account of inductive inference, the prior $p(h)$ captures the inductive biases of the learner, indicating which hypothesis a learner will favor when multiple hypotheses are equally consistent with the observed data (ie. which hypothesis will have higher probability when multiple hypotheses have equal likelihood). This account is attractive in that it can potentially allow us to identify the inductive biases of human learners, comparing different Bayesian models to find an appropriate prior. However, as we show in the remainder of this section, one should be cautious in interpreting such a prior: Considering algorithms by which people might be making inductive inferences shows that multiple processes can be reflected in a prior estimated from behavioral data.

Inferences with a reduced hypothesis space

As a computational-level theory of inductive inference, Bayesian models make no commitments about the psychological mechanisms by which people actually learn and reason. The most naïve interpretation of experiments demonstrating that people produce behavior consistent with Bayesian inference is that people are actually computing Bayes' rule in their heads. There are many reasons why such an algorithm is implausible, not least the requirement that people have all relevant hypotheses available whenever they make an inductive inference. However, this naïve algorithm provides a good starting point for exploring the consequences of different psychological processes that could play a role in inductive inference. Here we explore the consequences of modifying one aspect of this algorithm: rather than considering all possible hypotheses in the hypothesis space, considering only a subset of these hypotheses.

Research in inductive inference and scientific reasoning has shown that hypothesis generation is a challenging component of solving inductive problems (e.g., Kuhn, 1989; Klahr et al., 1993). Hypotheses can be generated in many different ways, including detecting cues from context, recognizing similarities to previous experiences, and making analogies to other domains (e.g., Gick & Holyoak, 1980; Gentner, 2002; Nersessian, 1992; Koslowski, 1996). We will not attempt to model these processes here, but for our purposes, it is sufficient to assume that the result of all of these processes can be summarized in a single probability distribution over hypothesis spaces. Using this probability distribution, $q(\mathcal{H}^*)$, we define the *Generate-Evaluate* (GE) algorithm for Bayesian inference with a reduced hypothesis space:

Step 1: Generate Sample a reduced hypothesis space $\mathcal{H}^* \subseteq \mathcal{H}$ from the probability distribution $q(\mathcal{H}^*)$.

Step 2: Evaluate Evaluate the hypotheses in the reduced hypothesis space $\mathcal{H}^*$ by applying Bayesian inference, using a prior distribution on $\mathcal{H}^*$ proportional to the prior on the full hypothesis space $\mathcal{H}$. Using $p(h)$ to denote the prior on the full hypothesis space, as in Equation 1 we obtain the reduced posterior distribution

$$p^*(h|d) = \frac{p(d|h)p(h)}{\sum_{h' \in \mathcal{H}^*} p(d|h')p(h')} \quad (2)$$

for $h \in \mathcal{H}^*$, with all other hypotheses receiving a posterior probability of zero. Because we are only sampling a subset of hypotheses, those that are not sampled will never be considered.

Mathematical analysis

Having defined an algorithm that takes into account the process of hypothesis generation, we can now analyze the consequences of using this algorithm. We have two questions of interest. First, will a learner using the GE algorithm produce behavior that appears to be consistent with Bayesian inference? Second, how does the process of hypothesis generation influence the interpretation of the resulting Bayesian model? We can answer both of these questions for a special case of this algorithm by exploiting its relationship to a Monte Carlo method known as importance sampling.

Monte Carlo methods are a class of algorithms that are used to approximate probabilistic computations by substituting samples from a probability distribution for the distribution itself. For example, if we wanted to perform computations involving a distribution $p(x)$, we could instead substitute a set of m values $x_1, \ldots, x_m$ drawn from $p(x)$, each with weight $1/m$. Importance sampling is a Monte Carlo method that takes this one step further, substituting samples from another distribution (the *surrogate* distribution) for samples from the target distribution (for details, see Neal, 1993). Thus, if we wanted to perform computations involving $p(x)$, we would generate a set of samples $x_1, \ldots, x_m$ from the surrogate distribution $q(x)$. We can get away with doing this if we no longer assign those samples equal weights. Instead, we give each sample x_i a weight proportional to $p(x_i)/q(x_i)$. The approxi-

mation to $p(x)$ is thus

$$p^*(x) = \frac{p(x_i)/q(x_i)}{\sum_{j=1}^{m} p(x_j)/q(x_j)} \quad (3)$$

for $x_i \in \{x_1, \ldots, x_m\}$, and zero otherwise. Intuitively, the weights proportion to $p(x_i)/q(x_i)$ reflect the "importance" of each sample. If x_i is more probable under $q(x)$ than $p(x)$, it will be over-represented in the sample, and thus should receive lower weight. If x_i is more probable under $p(x)$ than $q(x)$, there will be fewer such values than there should be, and it receives higher weight to compensate. This yields an asymptotically unbiased approximation to probabilistic computations involving the target distribution, provided certain constraints are observed (for example $q(x)$ has to be greater than zero wherever $p(x)$ is greater than zero).

Importance sampling gives us the tools we need to analyze the GE algorithm. If we assume that the samples are drawn independently, with $q(\mathcal{H}^*) = \prod_{h \in \mathcal{H}^*} q(h)$, then the GE algorithm is an importance sampler for the target distribution

$$\frac{p(d|h)p(h)q(h)}{\sum_{h \in \mathcal{H}} p(d|h)p(h)q(h)} \quad (4)$$

which is the posterior distribution obtained when using a prior proportional to the product of $p(h)$, the prior on the original hypothesis space, and $q(h)$, the probability of generating that hypothesis. It is straightforward to check that this is the case: if we approximate the distribution given in Equation 4 using $q(h)$ as the surrogate distribution, then we should generate a reduced hypothesis space $\mathcal{H}^*$ by sampling from $q(h)$ and then assign each sampled hypothesis a weight proportional to $p(d|h)p(h)q(h)/q(h) = p(d|h)p(h)$. This is exactly the procedure followed in the GE algorithm, with Equation 2 being equivalent to Equation 3.[1]

This analysis answers our two questions about the GE algorithm. First, it shows that a learner using this algorithm will still produce behavior consistent with Bayesian inference, albeit with a modified prior. Second, it indicates how the process of hypothesis generation affects behavior: If we estimate a prior by assuming people are performing Bayesian inference, that prior will reflect both the a priori plausibility of hypotheses, $p(h)$, and the probability of generating those hypotheses, $q(h)$. One needs not consider $q(h)$ when hypotheses are provided to the learner to evaluate, and thus no generation is required. However, the analysis indicates that we should be careful in interpreting priors estimated using Bayesian models: if we do not take algorithmic processes into account, hypothesis generation and evaluation are confounded. This can be problematic, as processes that change the way people generate hypotheses, such as priming a particular hypothesis, will influence the distribution $q(h)$ and hence the estimated prior, without influencing the plausibility of a hypothesis $p(h)$. Critically, ignoring the algorithmic level could therefore lead to counter-intuitive results where we need to use different priors to explain behavior across contexts where all that differs is the ease in which hypotheses are generated.

Generation and evaluation in human inferences

Our analysis assumes that the spontaneous generation of hypotheses can be separated from the evaluation of a given hypothesis. Thus, if the analysis is correct, generation and evaluation should be separable components of human inductive inference. If a learner does not sample the correct hypothesis, she will never consider it and thus cannot evaluate it; however, if a hypothesis is given to her (e.g., supplied by an experimenter), she should be able to evaluate the hypothesis just as if she generated it herself. We can empirically explore whether confounding generation and evaluation is a problem for Bayesian models in practice.

Testing the assumption that generation and evaluation are separable requires finding a task that allows us to manipulate the ease of generating different hypotheses. Previous work suggests that priming of a hypothesis can help people solve complex reasoning tasks. For example, Schunn and Dunbar (1996) found that even though participants do not spontaneously make an explicit analogy between domains, knowledge from one domain can influence reasoning in the other. Encouraged by this finding, we predicted that participants should generate different samples of hypotheses if primed differently. Priming hypotheses would thus modify the probability of generating those hypotheses, $q(h)$. However, such priming should not affect the evaluation of hypotheses provided for a learner.

In order to test whether the processes of generating and evaluating hypotheses are separable, we designed a priming task and a two part causal learning experiment. Prior to the causal learning experiment, half the participants read a vignette that primed them to think about the correct causal rule; the other half of the participants were given a "neutral" vignette. In the causal learning experiment, participants were given experience with sets of blue blocks (individuated by a letter on the block) that sometimes lit up when they interacted with each other. In the first part of the causal learning experiment, as participants encountered data, they were asked to make predictions about the result of new block interactions[2], and following all evidence, participants were asked to describe the rule they had discovered that best captured the pattern of lighting/nonlighting blocks. The actual rule by which evidence was generated was a "rank order" rule, which meant that a latent feature of "block strength" dictated which blocks could light others. The evidence was ambiguous such that the rule was not immediately obvious, but still potentially discoverable. In the second part of the causal learning experiment, participants completed a more standard task, traditionally taken as reflecting the posterior in Bayesian learning paradigms; participants were given several rules and asked

[1] Technically, we also require that $\mathcal{H}^*$ be a multiset, allowing multiple instances of the same hypothesis.

[2] The block task was inspired by a similar method used by Tenenbaum and Niyogi (2003).

to evaluate the degree to which each rule seemed plausible, given the blocks' interactions previously demonstrated in the learning phase.

Note that because the participants are required to discover the correct causal rule in the first part of the causal learning experiment, their ability to produce the correct predictions and the correct description require both steps of the GE algorithm: the subjects must generate a set of possible hypotheses and evaluate those hypotheses to discover the causal rule that best captures the observed data. In contrast, the second part of the causal learning experiment requires only evaluation, because the set of possible hypotheses is already provided for the participant. If generation is an important factor in determining people's inferences, we should observe a difference between the two parts of the experiment, and in particular, a difference in participants' sensitivity to the prime manipulation. Specifically, if the prime affects only generation, it should only affect participant responses in the first part of the experiment: participants given a strong prime should be more likely to generate the correct hypothesis than participants who are given a weak prime, but strong prime and weak prime participants should be equally likely to correctly rate the explanations provided to them in the second part of the experiment because this task only requires evaluation and does not require generation. However, if the prime affects other things, like the prior, it will affect both parts of the experiment: the strong prime participants should not only be more likely to generate the correct causal explanations in the first part of the experiment, but they should also be more likely than the weak prime participants to provide a higher rating of the provided, correct explanation in the second part of the experiment.

Methods

Participants and Design Participants were 40 undergraduates from the University of California, Berkeley who participated either for pay or for course credit. Participants were randomly assigned to either a *Strong Prime* or *Neutral Prime* condition. About half the participants completed an unrelated experiment prior to completing this experiment.

Stimuli The *Strong* and *Neutral Prime* vignettes were given to participants on a single sheet of paper with instructions. The target experiment included six small (6cm $\times$ 6cm) cardboard cutouts that the participants could manipulate as they completed the task and a 12 page booklet that included instructions, descriptions of the blocks, and sections to write in answers (see Figure 1).

Procedure The procedure involved a priming stage and a two part causal learning task, we outline each in turn.

Priming: Participants were first given an "unrelated" survey, which included a vignette about teachers watching children interacting on a playground and learning about rules that governed which children would win a game. In the *Strong Prime* condition the story suggesting that the rule governing which children would win was related to the childrens

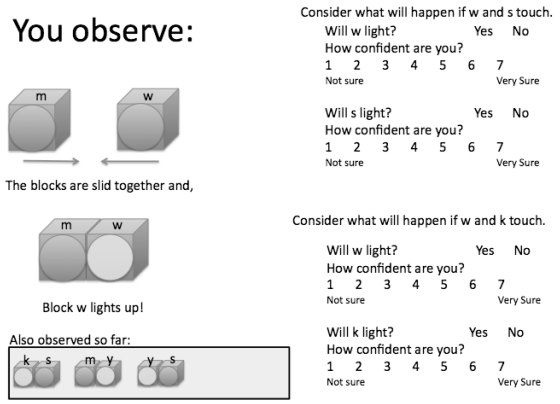


Figure 1: Example page from experiment booklet.

height. The text read, in its entirety: "Teachers at an elementary school taught their students a game for two children to play. They observed the results of pairs of students playing the game and tried to come up with a way to predict (for any given pair of students) who was going to win the game. At first it was difficult for the teachers to notice anything that would help them correctly predict the outcomes of the games. Then the teachers started organizing the children by the height of the children and the pattern of results quickly became apparent. The teachers were able to use the height of the children and make very accurate predictions as to who (for any given pair of students) was going to win the game." The *Neutral Prime* vignette was identical, except that instead of organizing the children by height, children were organized by the shirt color. Shirt color was chosen because pilot work suggested that numerous possible orderings may be plausible (e.g. sorting by the color wheel; bold colors vs. neutral colors; arranging from lightest to darkest colors, etc.), and thus the primed causal rule was somewhat arbitrary. Following the vignettes, participants were asked to respond to two simple questions about the story on the back of the sheet.

Causal Learning: In the first part of the causal learning task, participants saw sets of blue blocks (individuated by a letter on the block) that sometimes lit up when they interacted with each other. The actual rule, unbeknownst to the participants, was that the blocks could be ordered by "strength" with the "stronger" blocks always causing the "weaker" blocks to light (i.e. a variable like "height" given in the *Strong Prime* vignette, that would result in causal relations following a rank order[3]). As participants encountered data, they were asked to make predictions about the result of new block interactions ("Will this block light? Yes or no?") and provided confidence ratings on a scale of 1 to 7 (see Figure 1). Following all evidence, participants were asked to describe the rule they had discovered that best captured the pattern of light-

[3]Pilot work suggested that causal relations that follow a rank-order (e.g. dominance hierarchy) are not immediately obvious, but still potentially discoverable to participants, following suggestive evidence.

ing/nonlighting blocks and whether they could organize the blocks to best capture the rule.

In the second part of the causal learning task, participants were asked to evaluate four different explanations describing how the blocks should interact. Two explanations captured some, but not all of the data (e.g. "The blocks can be organized into two groups: blocks *s*, *k*, & *m* have the power to light up the other blocks (*y*, *w*, & *g*). Blocks in the same group do not light each other.") One explanation was nondescriptive: "The blocks can not be organized. They will light or not light randomly, but only one block can be lit at a time." And the final explanation was the target explanation, which correctly described the data: "The blocks can be organized by 'strength'. The stronger blocks will light the weaker ones. Strongest *s k m y w/g* Weakest". Participants rated the explanations on a scale from 1 ("not good") to 7 ("very good").

Results

Data were coded by the first author and reliability coded by a research assistant blind to condition and hypothesis outcomes. Explanation generation responses were labeled as "correct" or "incorrect". Agreement was 98%; the single disagreement was resolved conservatively with respect to predictions. Two participants were excluded and replaced for failing to provide a sensible response to the comprehension questions. Otherwise, all participants completed the comprehension questions for the priming vignettes.

Results confirmed that the ability to generate a hypothesis is separate from the evaluation of hypotheses. As predicted by Bayesian inference, there were no differences in evaluating the hypotheses between conditions: Both the Strong Prime and Neutral Prime participants readily rated the correct explanation equally likely: (*Strong*: 5.3; *Neutral*: 5.6; $t(38) = .48, p = ns$), and both groups ranked it well above the other (incorrect) provided rules (Strong: 2.8; Neutral: 3.0) (Wilcoxon Signed-Rank: *Strong*, $z = 3.07$, $p = .001$; *Neutral*, $z = 3.60$, $p < .001$) (Figure 2). However, there was a significant effect of condition: Participants in the *Strong Prime* condition were significantly more likely to answer the prediction questions correctly (Wilcoxon Signed-Rank: $w = 45$, $p < .01$; Figure 2a) and were more likely to generate the correct rule, Pearson $\chi^2(N = 40, 1) = 3.6, p = .058$. 65% of the participants in the *Strong Prime* condition provided the correct hypothesis, whereas only 35% of participants in the *Neutral Prime* condition generated the correct hypothesis (Figure 2b). That is, even though participants in the *Neutral Prime* condition were able to correctly evaluated the rules when they were provided, they were not necessarily able to generate the correct rule from the evidence alone.

We also looked at participant explanation ratings with the dependent factor being whether or not the participant generated the correct prediction on their own. Participants who did not generate the correct rule on their own still provided a significantly higher rating to the correct explanation (mean = 4.9) than to the incorrect explanations (mean = 3.3)(Wilcoxon Signed-Rank: $z = 2.63$, $p < .01$). That is, even though these participants were not able to generate the correct rule on their own, they were perfectly able to evaluate the good and bad explanations, being more likely to rate the correct explanation higher than the incorrect explanations.

Discussion

Connecting the computational and algorithmic levels is a significant challenge for Bayesian models of cognition. We have shown that considering the algorithms by which people might perform inductive inference can provide insight into how different psychological processes influence the conclusions that we can draw when using Bayesian models. Mathematical analysis of a simple algorithm in which learners first generate and then evaluate hypotheses indicates that while the resulting behavior is still consistent with Bayesian inference, estimation of a prior distribution from this behavior will confound the probability of generating a hypothesis and its a priori plausibility.

The responses of participants in our experiment provide some empirical support for the assumptions behind our analysis: While priming influenced whether participants could *generate* the correct explanation, it did not affect participants ability to correctly *evaluate* explanations that were provided. That is, one interpretation of our results is that the prime affected the distribution $q(h)$ from which hypotheses are generated, but it did not affect the prior probability of any particular hypothesis $p(h)$, since there were no differences between conditions when participants were asked to evaluate hypotheses that were provided to them. In the remainder of the paper, we consider some of the implications of these results and directions for future work.

Errors and approximations

Approaching inductive inference from the algorithmic level results in additional implications and predictions that may be valuable to explore in future work. For example, the algorithmic approach taken in this paper offers some reconciliation between computational level theories that suggest people are carrying out rational inference, with approaches that show people performing in seemingly "irrational" ways, such as not coming to the correct conclusion despite unambiguous or compelling evidence. By suggesting that people may be *approximating* rational inference by sampling a subset of hypotheses, these failures of inductive inference can be explained as the result of not generating appropriate hypotheses.

This makes predictions about the factors that should influence the errors that people make in inductive inference. For example, as the hypothesis space becomes large, the probability of sampling the correct hypothesis decreases. Thus, we should observe a trade-off between the size of the space and the probability of generating the correct explanation. Similarly, if cognitive limitations are imposed (for example increasing participant computational load, with additional tasks) then the set size of samples generated should decrease, and thus decrease the probability of generating the correct

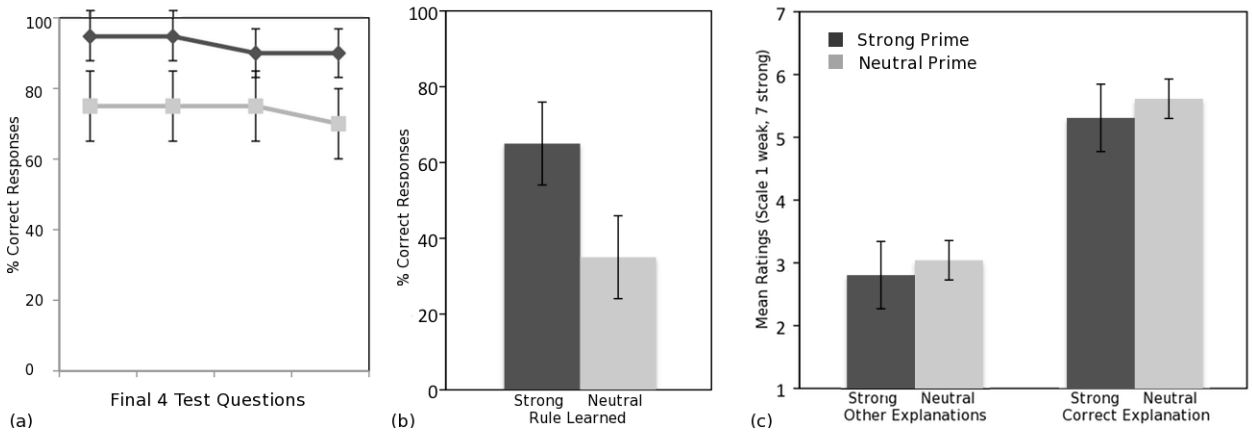


Figure 2: (a) Participants' responses to the final four prediction questions in the Neutral and Strong Prime conditions. (b) Percentage of participants who generated the correct explanation. (c) Average rating (1 weakest - 7 strongest) of the provided explanations by participants in both conditions.

sample. It may also be valuable to explore these questions in a developmental setting, examining how changes in information processing capacity influence the conclusions that children reach.

Conclusion

Bayesian models of cognition provide a computational-level account of inductive inference. Here, we have presented an analysis that shows how taking an algorithmic-level approach can allow us to tease apart two processes that are confounded in computational-level models: hypothesis generation and evaluation. We also present experimental results that suggest that these two processes are separable in human inductive inference. Together, our analysis and empirical findings indicate that we should take both the probability of generating a hypothesis and its a priori plausibility into account when interpreting prior distributions estimated using Bayesian models. More generally, these results illustrate that understanding human inductive inference will require working at both computational and algorithmic levels of analysis, and establishing the connections between them.

Acknowledgments. We thank Nannick Bonnel and Jason Martin for assistance in data collection. This research was supported by the James S. McDonnell Foundation Causal Learning Collaborative and grant number IIS-0845410 from the National Science Foundation.

References

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science*, *2*, 396-408.

Freeman, W. T. (1994). The generic viewpoint assumption in a framework for visual perception. *Nature*, *368*, 542-545.

Geisler, W. S., Perry, J. S., Super, B. J., & Gallogly, D. P. (2001). Edge co-occurrence in natural images predicts contour grouping performance. *Vision Research*, *41*, 711-724.

Gentner, D. (2002). Analogy in scientific discovery: The case of johannes kepler. In L. Magnani & N. Nersessian (Eds.), *Model-based reasoning: Science, technology, values.* New York, NY: Kluwer Academic, Plenum Publisher.

Gick, M., & Holyoak, K. (1980). Analogical problem solving. *Cognitive Psychology*, *12*, 306-355.

Goodman, N., Tenenbaum, J., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, *32:1*, 108-154.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*, 354-384.

Griffiths, T. L., & Tenenbaum, J. B. (2007). Two proposals for causal grammars. In A. Gopnik & L. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation.* Oxford: Oxford University Press.

Huber, D. E., Shiffrin, R. M., Lyle, K. B., & Ruys, K. I. (2001). Perception and preference in short-term word priming. *Psychological Review*, *108*, 149-182.

Klahr, D., Fay, A., & Dunbar, K. (1993). Heuristics for scientific experimentation: A developmental study. *Cognitive Psychology*, *25*, 111-146.

Knill, D. C., & Richards, W. A. (1996). *Perception as Bayesian inference*. Cambridge: Cambrdige University Press.

Körding, K., & Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature*, *427*, 244-247.

Koslowski, B. (1996). *Theory and evidence: The development of scientific reasoning*. Cambridge, MA: MIT Press.

Kuhn, D. (1989). Children and adults as intuitive scientists. *Psychological Review*, *96*, 674-689.

Marr, D. (1982). *Vision*. San Francisco, CA: W. H. Freeman.

Neal, R. M. (1993). *Probabilistic inference using Markov chain Monte Carlo methods* (Tech. Rep. No. CRG-TR-93-1). University of Toronto.

Nersessian, N. (1992). How do scientists think? capturing the dynamics of conceptual change in science. In R. Giere & H. Feigle (Eds.), *Minnesota studies in the philosophy of science.* Minneapolis: Minnesota: University of Minnesota Press.

Schunn, C., & Dunbar, K. (1996). Priming, analogy, and awareness in complex reasoning. *Memory & Cognition*, *24:3*, 271-284.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM: Retrieving Effectively from Memory. *Psychonomic Bulletin & Review*, *4*, 145-166.

Tenenbaum, J. B., & Niyogi, S. (2003). Learning causal laws. In R. Alterman & D. Kirsh (Eds.), *Proceedings of the 25th annual meeting of the cognitive science society.* Hillsdale, NJ: Erlbaum.

Weiss, Y., Simonvelli, E. P., & Adelson, E. H. (2002). Motion illusions as optimal percepts. *Nature Neuroscience*, *5*, 598-604.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, *114*, 245-272.

Finding a Bigger Fish Bowl: Higher Difficulty Helps Transitive Inferences

Sarah Schwind (schwinsr@mail.uc.edu)
University of Cincinnati, Department of Psychology,
Dyer Hall, Cincinnati, OH 45221-0376, USA

Heidi Kloos (heidi.kloos@uc.edu)
University of Cincinnati, Department of Psychology,
230 Dyer Hall, Cincinnati, OH 45221-0376, USA

Abstract

The current study looks at preschoolers' ability to discover higher-order patterns spontaneously, without being explicitly taught to do so. The higher-order pattern of interest was the degree of transitivity among the relations of three arbitrary dimensions. Preschoolers and adults were taught two relations (i.e., A = B; B = C), and they were asked to guess the third relation (i.e., between A and C). In each case, a relation was a perfect correlation between two arbitrary relations (e.g., heavy = large). The crucial manipulation pertained to how difficult it was to learn the two relations. The two relations either matched in direction (which was conceived as low learning difficulty), or they had opposite directions (which was conceived as high learning difficulty). Our prediction was that the higher-order pattern of transitivity becomes apparent when learning difficulty is high. The argument is that a local mismatch makes it difficult for children to focus merely on the isolated relations, and thus sets the stage for higher-order insights. Results confirm our hypothesis, both for preschoolers and adults. Participants were more likely to engage in higher-order transitive reasoning in the case of a local mismatch between the to-be-learned relations than in the case of a local match.

Keywords: learning; preschoolers; reasoning

Introduction

It is commonly believed that young children learn best when the content is broken down into 'digestible' pieces of information. The implicit expectation is that the pieces of information are combined into a whole later on, when the child is thought to be cognitively ready. For example, to teach children about an overarching principle, say in physics, one might introduce children to the constitutive parts of the principle first. When ready, the child might then put the pieces together and infer the overarching principle.

Basic-level research casts doubt on this logic, however. In particular there is evidence that children have difficulty combining pieces of information into larger units (e.g., Morris & Sloutsky, 2002; Ruffman, 1999). Therefore, teaching them piece-meal information might not lead to the desired success. Take for example a context in which participants are presented with the three physical dimensions size, loudness, and grayness (Smith & Sera, 1992). The task is to relate each dimension to the next, such as to determine whether something small goes with something loud or quiet, whether something loud goes with something dark or light, and whether something dark goes with something big or small. Children at preschool age had no difficulty relating the dimensions in a consistent way (e.g., if they decided that small goes with light, they also related dark with big). However, preschoolers were not constrained by the higher-order transitivity among these relations. Children believed, for example, that a big object was related to a loud sound (A = B), that the loud sound was related to the light gray (B = C), and that the light gray was related to the small object (C = not A). Figure 1 shows these three relations in schematic form. While they are normatively possible, they do not respect transitivity.

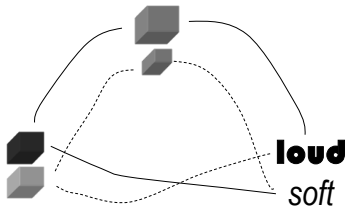


Figure 1. Representation of three features relations (combining two dimensions each) that lack transitivity among each other.

There is another reason why a piece-meal teaching approach might not work. Children not only fail to combine pieces of information into a desired higher-order structure, they impose an incorrect structure, ignoring pieces of information that conflict with it. In other words, children might fail altogether to learn a piece of information if it does not match with other beliefs they might hold. Consider, for example, children's beliefs that heavier objects sink faster in water than lighter objects. When this belief is somehow elicited in a teaching protocol, children will have difficulty learning that small objects sink faster than large objects (Kloos, 2007). There is nothing particularly difficult about the latter volume-speed relation, and children can easily learn it if it is the only thing children think about. But as soon as mass is varied in a salient way, children impose an overarching belief that mass and volume correspond in their

effect. They believe that, if heavy objects sink fastest, large objects should sink fastest too.

If learning the individual parts of the whole does not necessarily set young children up to spontaneously discover an overarching principle, and if young children might even fail to learn individual parts, what could help them learn higher-order patterns?

To address this question, we used a transitive-inference task similar to the one used in Smith & Sera (1992) described. While transitive inference is not a concept commonly taught to children, it is seen as a basic reasoning process that might underlie all learning of higher-order structure (Inhelder & Piaget, 1958). Furthermore, young children are not incapable of making a transitive inference (e.g., Adams, 1978). When preschoolers were taught the length relation between five sticks in a series (e.g., stick A is taller than stick B, stick B is taller than stick C, and so on), they were able to incorporate a new stick into the series and guess its relative size, integrating the sticks into one continuous dimension.

Of course, Adams' transitive-inference task on how sticks relate to each other in their lengths differs from Smith & Sera's transitive-inference task on how the alignment of poles respects higher-order Gestalt. Most notably, transitivity pertains to a logical necessity in the former case, but not in the latter case. If Stick A is larger than Stick B, and Stick B is larger than stick C, then stick A has to be larger than stick C – by logical necessity. Conversely, if big goes with loud, and loud goes with dark, it is not logically required that dark goes with big. Nevertheless, despite these differences in context, findings from Adams (1987) shed light on what it is that might help children discover the higher-order pattern of transitivity.

In particular, preschoolers in Adam's study were more likely to make a transitive inference when the length differences between the sticks were small (approximately 1 cm.). The small difference in length might have allowed children to think of the series as a whole, rather than to focus on each pair individually if the length differences were larger. Based on the findings, we predict that children are more likely to attend to a higher-order pattern when a narrow focus on an isolated pattern hinders learning of isolated parts. In Adams' (1978) transitive-inference task, a 1cm length difference between adjacent sticks made it difficult to narrowly focus on isolated sticks (none of them stuck out as particularly long or particularly small).

Similar arguments have been made in mathematical reasoning, when 11-year-olds spontaneously discovered a mathematical rule after being presented with individual instances (Kaminski, Sloutsky, & Heckler, 2009). Learning was markedly improved when individual instances were maximally abstract, possibly because it made it difficult for children to sustain a local focus on the separate instances. In the current paper, we apply this idea to transitive inferences.

In particular, we adapted a version of the Smith and Sera (1992) task that involved relations between three physical dimensions (size, shading, and depth). Given that dimensions are polar (they have a 'more' pole and a 'less' pole), a relation can be considered positive or negative. For example, 'big' aligning with 'dark' is a positive relation, while 'big' aligning with 'light' is a negative relation[1]. Transitivity exists when the three relations are congruent among each other. For example, in a congruent set, 'big' is aligned with 'dark' (A = B), 'dark' is aligned with 'deep' (B = C), and 'deep' is aligned with 'big' (C = A). Figure 2 shows the congruence among these relations graphically.

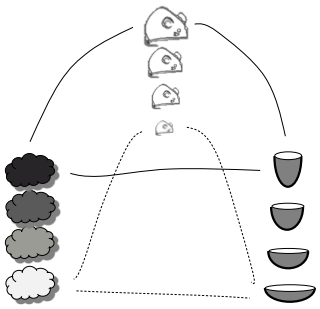


Figure 2. Representation of the three feature relations that are congruent among each other.

Preschool children were taught two of the relations (e.g., how size relates to darkness and how darkness relates to depth), and they had to guess the third relation (e.g., how size relates to depth). Adults were included for comparison purposes.

We tested the idea that children might attend to a higher-order pattern of transitivity in situations in which focusing of isolated parts was hampered in the task. The crucial manipulation was whether the two to-be-learned relations matched in direction or not. Relations that matched in direction were either both positive (e.g., 'big' goes with 'dark', and 'dark' goes with 'deep'), or they were both negative (e.g., 'big' goes with 'light', and 'light' goes with 'deep'). And for relations that did not match in direction, one was positive and one was negative (e.g., 'big' goes with 'dark', and 'dark' goes with 'shallow').

Our reasoning was that children could easily learn relations that match in direction. Children might therefore merely focus on learning the isolated relations, without regard for the higher-order pattern of transitivity. Conversely, children should have more difficulty learning the two relations of opposite direction. As a result, they might be more likely to spontaneously integrate the two into the higher-order patterns of transitivity.

[1] Note that the 'more' pole is ambiguous for shadings (more grey vs. more white), and for depth (deeper vs. wider). On an absolute level, it is therefore arbitrary whether a relation is considered positive (big = more gray) or negative (big = less white). However, the chosen 'more' pole was always labeled as such (i.e., darkest; deepest) resulting in the prescribed direction of the relation.

Method

Participants

A total of 63 children, aged 4 to 5 years (M = 5.0 years, SD = 3.6 months) were recruited from daycares and elementary schools located around the Cincinnati, OH and Northern Kentucky areas. Three children were tested and excluded from the experiment because they did not meet the learning criterion (see Procedure), and five children did not finish due to loss of interest. In addition, we tested 60 undergraduate students (M = 21.4 years, SD = 5.7 years), recruited from the University of Cincinnati, in return for class credit.

Materials

Materials were pictures of four cartoon mice, four clouds, and four bowls, presented on a computer screen. Mice differed in size (from 1 to 4 cm), clouds differed in achromatic color (from the lightest shade of gray to the darkest gray), and bowls differed in depths (from shallow to deep). Figure 2 shows the four pictures of each element. The resulting relations are between mouse size and cloud darkness (MC), mouse size and bowl depth (MB), and between cloud darkness and bowl depth (CB).

Relations between features were labeled either positive or negative, depending on how the poles of features size, darkness, and depth were introduced. For example, for a positive mouse-cloud relation (MC+), the bigger mouse was paired with the darker cloud; for a negative mouse-cloud relation (MC-), the bigger mouse was paired with the lighter cloud.

Design

There were three conditions that differed in the direction of the relations presented to participants. Participants were taught two relations: two positive relations (e.g., MC+CB+) in the Plus-Plus condition, two negative relations (e.g., MC-CB-) in the Minus-Minus condition, or a negative and a positive relation (MC+CB-) in the Plus-Minus condition.

Table 1 shows, in schematic form, how the relations were combined to create the three different conditions. The first column contains the two to-be-learned relations participants were presented with. The second column shows the relation that participants were asked to guess. Finally, the last column shows the expected direction of the third relation if participants pay attention to transitive congruence. For example, if participants learned a positive cloud-bowl relation (CB+) and a positive mouse-cloud relation (MC+), then the direction of the third relation is expected to be positive as well.

Procedure

The cover story involved an explorer, Toto, who found a machine on a far-away planet. The machine was said to transform things. In particular, participants were told that this machine transformed objects: "If something is put it on one end, something completely different comes out on the other end."

Table 1: How the combinations of the relations create each condition.

To-be-Learned Relations	To-be-Inferred Relation	If congruent…
Plus-Plus (two positive relations)		
Minus-Minus (two negative relations)		
Plus-Minus (one positive and one negative relation)		

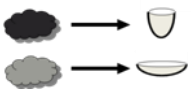
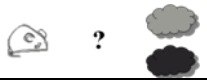
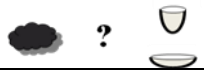
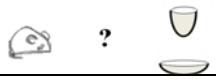
The next step was to introduce the objects that could go into the transformer. Six pictures of differently-sized mice were placed in front of the participant in random order. The difference in dimension was pointed out (e.g., 'See how some mice are big and some are little"), and participants were asked to point to the biggest mouse. Help was provided as needed. The chosen picture was moved to the side, participants had to point to the next biggest mouse, which again was moved to the side, and so on. Next, participants were presented with six pictures of differently colored clouds, and they were asked to order them from darkest to lightest. Finally, participants were presented with six pictures of bowls and the required ordering was from deepest to most shallow. Children and adults had no difficulty completing this task, suggesting that they could focus on the dimensions in question.

To prepare participants for the learning task, the experimenter provided the following information, accompanied by pictures on the computer:

> "A mouse will either turn into a cloud or a bowl, a cloud will turn into either a mouse or a bowl, and a bowl will turn into either a mouse or a cloud. Sometimes the biggest mouse will turn into the darkest cloud, and sometimes the biggest mouse will turn into the lightest cloud. Sometimes the darkest cloud will turn into the deepest bowl, and sometimes the darkest cloud will turn into the shallowest bowl. Sometimes the deepest bowl will turn into the biggest mouse, and sometimes the deepest bowl will turn into the smallest mouse. Toto is very confused and doesn't know what's going on. But he made some movies for us showing us what the transformer is doing. Can you help him figure it out?"

The experiment proper started immediately and had two phases: a demonstration phase and a testing phase (see Table 2). During demonstrations, participants watched a set of movies that conveyed the feature relations. For each movie, two transformers were displayed above each other, in the middle of the screen. Two objects entered simultaneously on one side of each transformer, and another two objects came out simultaneously on the other side of the transformer. For example, a big mouse and a small mouse each entered a transformer, and a dark and a light cloud each come out on the other end.

Table 2: The experiment phases in step-by-step form.

Phase 1: Demonstration	
Relation 1	
3 movies Pretest: 4 trials 3 movies Pretest: 4 trials	
Relation 2	
3 movies Pretest: 4 trials 3 movies Pretest: 4 trials	
Phase 2: Testing	
Learning Trials	
Relation 1: 4 trials	
Relation 2: 4 trials	
Inference Trials	
Relation 3: 4 trials	

To convey a relation, there were two sets of three movies, each followed by pre-testing to gauge initial learning. Movies pertaining to the same relation differed in the way items were combined with each other. The order in which movies were presented was randomized across children. Pre-testing started with a reminder of the relation presented during the set of movies. For example, if the movies showed a bigger mouse turning into a darker cloud, the experimenter explained: "The biggest mouse will always turn into the darkest cloud." Four pre-testing trials followed, each asking what an object turned into. For example, if the movies showed a bigger mouse turning into the darker cloud, the question was: "What cloud did the bigger mouse turn into?" Participants had to perform consistently, either correct or incorrect, on at least three of the four trials. They were excluded otherwise.

After participants watched two sets of movies for one relation and then two sets of movies for the second relation, the testing phase started, with the following instructions:

> "I think you know everything there is to know about the transformer. But just to make sure, Toto wants to ask you a few more questions."

Then four trials per relation were presented. The learning trials, for the first and second relations, were identical to the pre-testing trials presented earlier. The inference trials came last, following the same format as the other trials. Participants were asked to make a guess about the third relation that was not presented. For example, if the mouse-bowl relation was never shown, the experimenter would ask "what will the big mouse turn into?"

The demonstration phase lasted about 10 minutes, with the introduction and testing phases each lasting another 2-3 minutes. Overall, the experiment lasted around 15 minutes.

Results

In a preliminary analysis, we looked at children's learning of the two relations presented to them. For each participant, we calculated an average proportion-correct score across the eight learning trials. A 3 by 2 between-subjects ANOVA was conducted, with condition (Plus-Plus, Minus-Minus, Plus-Minus) and age (preschoolers, adults) as the factors. It revealed a significant effect of age, $F(1,117) = 31.27$, $p < .01$, in that adults performed better on learning trials ($M = .96$) than preschoolers ($M = .77$). There was also a difference in condition, $F(2,117) = 5.49$, $p < .01$, suggesting that participants had some difficulty learning the relations presented to them. However, there was no interaction with age and condition, $F < 1.85$, $p > 0.16$. Figure 3 shows the degree of learning (represented as mean % correct), as a function of age group and condition.

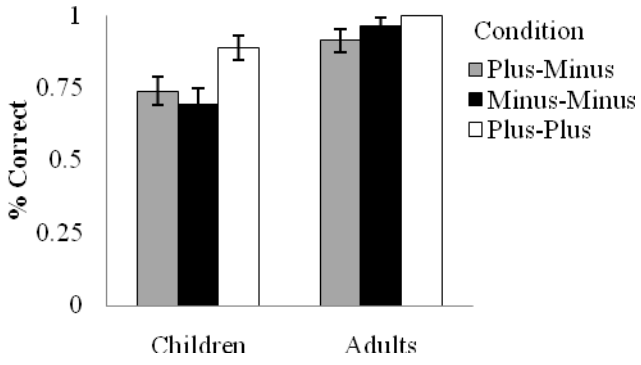


Figure 3. Mean performance correct on learning trials (to test the degree of learning), as a function of age and condition. Standard errors are shown as error bars.

To determine if participants were congruent in their inferences about the third relation, we considered only those participants who performed consistently on each set of four learning trials per relation. 'Consistent' here means either correct performance on at least three learning trials of a

relation, or incorrect performance on at least three learning trials of a relation. Eighteen children (29%) and 4 adults (7%) did not meet this criterion and were not included in the transitivity analysis. Of the included participants, 13 of the children and 3 of the adults performed consistently incorrect on one set of learning trials, and nobody performed consistently incorrect on both sets of learning trials.

If children make congruent inferences, then the inferred relation should be negative if one of the learned relations is positive and the other is negative. The inferred relation should be positive if the learned relations are either both negative or both positive. To what degree did participants' inferences follow this pattern across the four inference trials? Figure 4 shows participants' transitivity performance as a function of age and condition. A score of 1 means that performance was congruent on all four inference trials, while a score of 0 means that performance was incongruent on all four inference trials. As can be seen in the figure, both children and adults were more likely to give congruent answers in the Plus-Plus and Plus-Minus condition than the Minus-Minus condition.

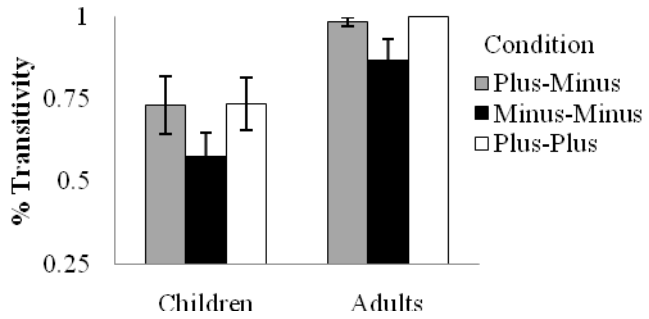


Figure 4. Mean proportion transitive inferences, as a function of age and condition. Standard errors are shown as error bars.

The transitivity scores were submitted to a 3 by 2 between-subjects ANOVA, with conditions (Plus-Plus, Minus-Minus, Plus-Minus) and age (preschoolers, adults) as factors. There was a significant effect of age, $F(1,95) = 30.11$, $p < .01$, with adults having higher transitivity scores ($M = .95$) than children ($M = .69$). More importantly, there was a significant effect of conditions, $F(2,95) = 3.59$, $p = .03$. There was no significant interaction, $F < .05$, $p > .95$, meaning that this pattern stayed the same for both children and adults. For both children and adults, guesses were transitive in the Plus-Minus and Plus-Plus conditions, but less so in the Minus-Minus condition. Learning score was uncorrelated with transitivity score.

Summary & Discussion

Our prediction was that children would be more likely to attend to the higher-order pattern of transitivity when the learning of the local elements (the single relations that make up the whole) did not afford a narrow focus. Learning two positive relations did not interfere with a local focus: children could pay attention to only one of the two relations and still be able to learn the second one (because the direction matched). The same was true for learning two negative relations: focusing locally on one negative relation did not hinder (and might have even helped) the learning of the second negative relations. But when children were asked to learn a positive and a negative relation, a local focus on a single relation hindered learning.

Results support our prediction – with a twist. Inferences of children in the Minus-Minus condition were less transitive than of children in the Plus-Minus condition. And the lower transitivity performance was not related to the participants' learning scores (i.e., the degree of transitivity of the guessed relation cannot be explained by the degree of learning of the two presented relations). This finding is consistent with our hypothesis: when children had to learn non-matching relations that hindered an overly local focus, the overarching pattern of transitivity was likely to emerge. Importantly, the patterns of transitivity appeared spontaneously for an age group that is commonly known for having difficulties with transitive inferences.

Adults were more likely to make a transitive inference than children. However, they were also affected by the learning manipulation. Transitivity was lower in the Minus-Minus condition than the Plus-Minus condition. As was found with preschool children, when single relations were difficult to learn with a narrow focus on each separate relation, adults spontaneously applied the higher-order transitivity to the relations.

A surprising finding pertained to performance in the Plus-Plus condition. We predicted transitivity to be low in this condition, because the two to-be-learned relations matched in direction, and thus afforded a local focus. Nevertheless, children and adults made higher-order transitive inferences when asked to guess the direction of the third relation. What could explain this performance?

A closer look at the specifics of the Plus-Plus condition might shed light on participants' inferences. Recall that two positive relations are congruent with another positive relation. If 'big' goes with 'dark' (positive), and 'dark' goes with 'deep' (positive), then 'deep' should go with 'big' (positive). But guessing a positive relation might be a default (cf., Inhelder & Piaget, 1958). Therefore, participants might have guessed a positive relation in this case with little regard to transitivity among all three relations.

If this is the case, participants' bias toward a congruent set of relations in the Plus-Minus condition is even more impressive evidence of transitive inference. In the case of a positive and a negative relation, the congruent third relation is negative (e.g., if 'big' goes with 'dark', and 'dark' goes with 'shallow', then 'deep' should go with 'little'). Thus, to make a congruent guess, participants (including preschool children) had to go against a default of guessing a positive relation and guessed a negative relation. Note that this

interpretation of the results needs to be qualified until we gain a better understanding of how children match the poles a priori.

Taken together, the results suggest that higher-order transitivity is an emergent property, employed as a means of reducing learning complexity. With higher complexity of individual elements, a local focus was compromised, helping children to note the larger whole. In future studies, it may be useful to follow up with different conditions, such as other cover stories or other objects. It remains to be seen if these claims hold across different domains.

References

Adams, M. J. (1978). Logical competence and transitive inference in young children. *Journal of Experimental Child Psychology, 25*, 477-489.

Inhelder, B., & Piaget, J. (1958). The growth of logical thinking from childhood to adolescence. New York: Basic Books.

Kaminski, J.A., Sloutsky, V.M., & Heckler, A. (2009). Transfer of mathematical knowledge: The portability of generic instantiations. *Child Development Perspectives, 3*, 151-155.

Kloos, H. (2007). Interlinking physical beliefs: Children's bias towards logical congruence. *Cognition, 103*, 227-252.

Morris, B. J., & Sloutsky, V. M. (2002) Children's solutions of logical versus empirical problems: What's missing and what develops? *Cognitive Development, 16*, 907-928.

Ruffman, T. (1999). Children's understanding of logical inconsistency. *Child Development, 70*, 872–886.

Smith, L. B. and Sera, M. D. (1992). A developmental analysis of the polar structure of dimensions. *Cognitive Psychology, 24*, 99-142.

Acknowledgements

This study was supported by grants to HK by the National Science Foundation (DRL 0723638) and the National Institute of Health (1R03HD055324). We thank the children, families, and adults who kindly agreed to participate in this research.

Preschoolers sample from probability distributions

Stephanie Denison (smdeniso@berkeley.edu) **Elizabeth Baraff Bonawitz (liz_b@berkeley.edu)**
Alison Gopnik (gopnik@berkeley.edu) **Thomas L. Griffiths (tom_griffiths@berkeley.edu)**
Department of Psychology, University of California, Berkeley, Berkeley, CA 94720 USA

Abstract

Researchers in both educational and developmental psychology have suggested that children are not particularly adept hypothesis testers, and that their behavior can often appear irrational. However, a growing body of research also suggests that people do engage in rational inference on a variety of tasks. Recently researchers have begun testing the idea that reasoners may be sampling hypotheses from an internal probability distribution when making inferences. If children are reasoning in this way, this might help to explain some seemingly irrational behavior seen in previous experiments. Forty 4-year-olds were tested on a probabilistic inference task that required them to make repeated guesses about which of two types of blocks had been randomly sampled from a population. Results suggest that children can sample from a probability distribution as evidenced by the fact that, as a group, they engaged in probability matching and that the dependency between successive guesses decreased over time.

Keywords: Cognitive Development; Causal Learning; Approximate Bayesian Inference; Sampling Hypotheses

Introduction

Young children are faced with a variety of novel situations on a daily basis. They encounter countless episodes in which they must reason about why particular events unfold the way they do, what this means in terms of how related events might unfold in the future, and how this newly acquired information fits into the knowledge they already possess. Humans revise their beliefs throughout development, often beginning with relatively flawed beliefs and progressing towards an increasingly accurate portrayal of the world. However, no current theory provides a satisfactory explanation of how children decide which hypotheses to test. Somehow they must search through the potentially infinite number of hypotheses that exist at the beginning of the learning process. Here, we investigate this question by asking whether young children can make probabilistic inferences via a process of sampling hypotheses from probability distributions.

The question of whether children and adults are capable of using rational inference to search through a hypothesis space and revise their beliefs has drawn mixed empirical findings. To begin, Piaget noted that children tend not to reason systematically about hypotheses, at least until they reach the formal operational stage in late childhood (Piaget, 1983). Since Piaget, some researchers have found evidence to corroborate this claim, stating that children often appear to navigate randomly through a selection of predictions and explanations (Siegler & Chen, 1998). For example, researchers in educational psychology have revealed evidence suggesting that young children and even non-expert adults are not particularly skilled hypothesis testers (e.g., Kuhn, 1989; Klahr, Fay, & Dunbar, 1993). Furthermore, developmental psychologists have found that children often revise their beliefs surprisingly slowly, suggesting a struggle to efficiently update theories (e.g., Carey, 1991; Wellman, 1990).

On the other hand, at least two kinds of evidence exist to suggest that children might be capable of using rational inference to generate, search through and evaluate hypotheses. First, recent research in cognitive psychology suggests that people reason in ways that are consistent with optimal Bayesian models in a variety of tasks (e.g., Griffiths & Tenenbaum, 2005; Goodman, Tenenbaum, Feldman, & Griffiths, 2008). Although most of this work examines adult reasoning, a growing body of evidence suggests that children can also reason in a way that is consistent with Bayesian inference (e.g., Gopnik et al., 2004; Kushnir & Gopnik, 2005; Schulz & Gopnik, 2004; Schulz, Bonawitz, & Griffiths, 2007; Goodman et al., 2006). For example, Xu and Tenenbaum (2007) found that preschoolers can systematically integrate prior knowledge regarding hierarchical information with evidence in order to apply the correct labels to a variety of objects in a word learning task and Schulz et al. (2007) and Kushnir and Gopnik (2007) found that children's causal inferences rationally depend on both their prior beliefs and the observed evidence. Second, many researchers advocate the theory-theory of conceptual development, which states that children's knowledge is organized into abstract, coherent conceptual systems, similar to those found in science (Carey, 1985; Gopnik & Meltzoff, 1997; Murphy & Medin, 1985; Wellman & Gelman, 1992). This framework predicts that children will engage in hypothesis testing in ways similar to scientists during learning, and much evidence has accumulated in support of this view (e.g. see Karmiloff-Smith & Inhelder, 1974; Bonawitz, Lim, & Schulz, 2007; Legare, Gelman, & Wellman, in press). However, the theory-theory does not specify where the hypotheses are derived from in the first place or how children could be expected (albeit unconsciously) to compute full Bayesian inference over (often) infinite hypothesis spaces.

The sampling hypothesis

Although Bayesian inference corresponds well to the theory-theory of conceptual development, researchers who advocate a rational approach to human inference do not suggest that adults and children actually work through the steps of Bayes' rule in daily life. Evaluating all possible hypotheses each time new data are observed would not be feasible both from a formal and a practical standpoint, given the large number of hypotheses that would require consideration. One way to think about how the mind may be approximating Bayesian inference is to start with good engineering answers to this problem. Techniques for approximating Bayesian inference have

already been developed in the fields of machine learning and statistics and we can see whether humans are also using some version of these strategies.

One strategy for implementing Bayesian inference is sample-based approximation (Shi, Feldman, & Griffiths, 2008; Sanborn, Griffiths, & Navarro, 2006). This approach states that people might be approximating Bayesian inference by evaluating a small sample of the many possible hypotheses. This "sampling hypothesis" has been supported by additional empirical data that suggest people often base their decision on just a few samples (Goodman et al., 2008; Mozer, Pashler, & Homaei, 2008). Indeed, in many cases an optimal solution is to take only one sample (Vul, Goodman, Griffiths, & Tenenbaum, 2009). Sampling partly involves picking a hypothesis at random from the distribution. However, the process is not entirely random in that distributional information may be used to generate hypotheses that are highly likely more often than those that are less likely. This strategy allows the learner to entertain a variety of hypotheses, ensuring that they will spend more resources testing likely hypotheses but will not overlook a lower probability hypothesis that could turn out to be correct.

The sampling strategy predicts "probability matching": aggregating over numerous samples, generated by different individuals in a group, should return the original distribution; as the number of samples approaches infinity, the closer the result will be to the distribution. This benefit of averaging is called the "wisdom of crowds". If instead people generate a "best guess", then aggregating over numerous samples should result in an inaccurate reflection of the distribution, characterized by an overweighting of the most likely hypothesis. Sampling also depends on independence between guesses; the more independent the draws from the distribution, the more accurate the sample will be. However, we might expect that if a single individual is generating multiple guesses, then there may be dependence between guesses, but this dependence may decrease as time between guesses increases.

Recently, Vul and Pashler (2008) tested the sampling hypothesis in adults. They asked individuals to make guesses about a variety of real-world statistics such as: What percentage of the world's airports are in the United States? In an *Immediate* condition, participants were asked to make guesses about a variety of real-world statistics and then asked the questions a second time directly after. In a *Delayed* condition, the question was asked for the second time two weeks later. It was found that an individual's error was reduced when their guesses were averaged compared to each of their individual guesses in both the Immediate and Delayed conditions. There was also a greater benefit of averaging guesses in the Delayed group than in the Immediate group; the independence of guesses and, therefore, accuracy was greater after a time delay. This suggests that adults were most likely sampling guesses from an internal distribution rather than always providing an optimal guess.

The results from Vul and Pashler (2008) suggest that adults may be approximating rational solutions when making guesses about frequencies, in accordance with the sampling hypothesis. We turn to the question of whether or not children are drawing samples from probability distributions in a similar way. We explore two predictions of the sampling hypothesis. First, if children use a strategy of sampling hypotheses from a distribution, we should see that the probability with which they select hypotheses should match the distribution. This contrasts with a strategy of maximizing (always choosing the most likely answer) or guessing (randomly providing responses, independent of their probability), which make different predictions. We will refer to this as the *probability matching* prediction. Second, because sampling depends on independence, we can predict that increasing dependencies between guesses will decrease the degree to which responses accurately reflect the distribution. We will call this the *dependency* prediction.

While results of several studies seem to suggest that children do in fact probability match in numerous situations (e.g. see Kam & Newport, 2009; Kushnir, Wellman, & Gelman, 2008; Bonawitz, Chang, Clark, & Lombrozo, 2008; Sobel, Tenenbaum, & Gopnik, 2004), to our knowledge, no research has demonstrated the dependency prediction, or analyzed results in terms of the sampling hypothesis. While much research has demonstrated the sophisticated graded response of children on average, any particular child's response is often, paradoxically "non-optimal." That is, often developmental studies involve forced-choice responses, and so the predictions of any single child seem in conflict to rational models: Why wouldn't children simply always choose the most likely response, rather than some fraction of children choosing the likely response and some smaller fraction choosing the unlikely response? If children are in fact approximating rational inference by sampling hypotheses at least in some situations, this may provide an account of these data. More importantly, the sampling hypothesis may also provide an account of how children navigate through potentially infinite hypothesis spaces during learning: rather than computing full Bayesian inference over the whole hypothesis space, children sample a subset of hypotheses. We now turn to our experiment to explore this question.

Do children sample hypotheses?

We investigate the sampling hypothesis in preschool-aged children by testing their ability to use probability information to make guesses about which of two colored blocks was most likely to be sampled from a population (consisting of a 4:1 ratio) on a single random draw. This design allows us to investigate whether children demonstrate the first of two sampling signatures: probability matching. First, we predict that if individual children are sampling from a distribution of hypotheses, their responses will be closer to the correct distribution (i.e., 80% red blocks) than would be predicted by random guessing (50% red, blue guesses) or maximizing (100% red guesses). Second, the dependency prediction suggests that

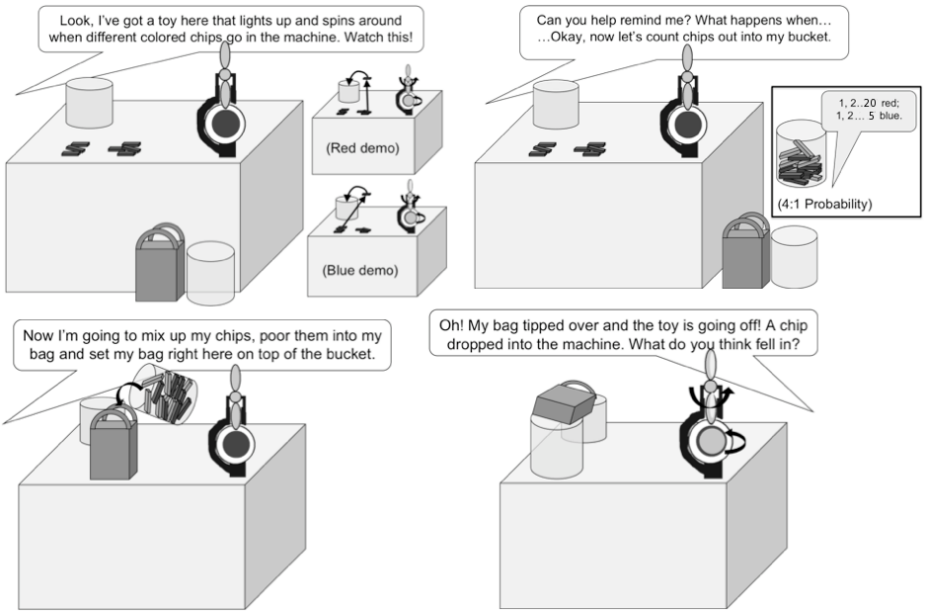


Figure 1: Stimuli and method used to test the sampling hypothesis in children.

children who are given a long delay between guesses will demonstrate greater independence of guesses across trials than children who provide guesses following only a very short delay and we can model these dependencies via a Markov process. As a result of differences in independence between guesses, the distribution over guesses in the *Long Wait* condition should be closer to the predicted distribution than the distribution over guesses in the *Short Wait* condition.

Methods

Participants Forty 4-year-olds were recruited from preschools located on the U.C. Berkeley campus. The children were randomly assigned to one of two conditions, each consisting of 20 children: the *Long Wait* condition, which included 8 males and 12 females (M = 54 months; R = 48 mos – 62 mos) and the *Short Wait* condition, which included 11 males and 9 females (M = 53 months; R = 48 mos – 59 mos). On additional child was tested and excluded due to failure to pass an initial comprehension check (see procedure below). The children's ethnicities reflected the composition of the area.

Stimuli A large box (12in $\times$ 12in $\times$ 18in) constructed out of cardboard and covered in yellow felt previously used in Bonawitz et al. (2008) was used. All five surfaces excepting the back side of the box were intact and covered with felt. A hole was cut out of the top of the box in the front right corner where a toy with a transparent sphere with lights and a spinner inside connected to a cylindrical shaft was inserted such that only the sphere was visible to the children. The toy activated by pressing a button on the shaft, causing it to light up and play music. An opaque activator bin was placed on the back left corner of the box. Additional stimuli included red, blue, and green domino sized wooden blocks; one red, one yellow and one green paper cup; a rigid green bag; and a transparent container. (See Figure 1).

Procedure *Short Wait Condition.* Each testing session was videotaped for data retrieval and a second experimenter recorded all responses online. The experiment began with the child and experimenter sitting across from one another with the large yellow box in between them—the front side facing the child and the back side facing the experimenter. The experimenter introduced children to the large yellow box saying, "This is my big toy and I'm going to show you how it works." The experimenter then took two blocks of each color (red, blue, and green) and placed them on the table. She showed the children that when a red block or a blue block is placed in the activator bin, the toy lights up and plays music and when a green block is placed in the bin, the toy does not activate. In reality, the experimenter was surreptitiously activating the toy by pressing a button. Previous work suggests that children (and even adults) find this manipulation compelling (Bonawitz et al., 2008). A comprehension check was then performed to ensure children remembered that the blue and red blocks make the toy activate and that green blocks did not. Next, the experimenter began Trial 1 by having the child count 20 red blocks and 5 blue blocks one at a time and placing them into a transparent container. The order of block color was counterbalanced. After counting the blocks, the experimenter shook the blocks in the container to mix them and poured them into the rigid bag. She then placed the container upside down in front of the activator bin on the yellow box and placed it on top of the container. She then accidentally knocked it over toward the activator bin. Just after the bag fell

over, the experimenter activated the toy and said, "Oh, I think one of the blocks must have fallen into the toy! Can you tell me which color it was?" Once the child answered the question, the experimenter pretended to remove the block while turning off the toy. Finally she asked, "and why do you think it was a [red/blue] chip?" Once children provided an answer, the experimenter began Trial 2 by saying, "That was kind of funny how I accidentally tipped the bag over and it made the toy go off. Should I try to make that happen again? First we have to count our blocks again." The second and third trials progressed exactly the same as Trial 1. At the end of Trial 3, the majority of children were asked three follow-up questions: the experimenter asked which color they guessed fell into the activator on the first, second, and third trials.

Long Wait Condition. The *Long Wait* condition was identical to the *Short Wait* except that children completed Trial 1 in the first testing session, Trial 2 in a second testing session one week later, and Trial 3 in a third testing sessions one week after Trial 2.

Results

There were no age differences between groups ($t(38) = 0.11, p = ns$). Responses were coded by first author and reliability coded by a research assistant blind to experimental hypotheses. All responses uniquely and unambiguously were either "red" or "blue" and agreement was 100%.

Probability Matching As expected, looking only at the first responses, there were no differences between conditions, $\chi^2(1, n = 40) = 1.9, p = ns$. To assess whether or not children probability matched, we averaged the first response of children in both the *Long Wait* and *Short Wait* condition. Overall, children's responses reflected probability matching (70% providing the more probable chip response). That is, results suggest that children were not simply randomly guessing, as responses were significantly different from chance ($p < .05$; binomial test), but not significantly different from the predicted distribution of .8 (p =*ns*, binomial test). Similarly, children were not "maximizing" by always providing the most probable response (i.e. always choosing the red chip), or responses would have approached ceiling.

Dependency Measures To assess whether children's responses were independent from one another across trials, we first assessed what the independent sampling assumption would predict. That is, given probability θ of sampling a particular chip, what should the distribution of three responses look like? Because there are two possible responses (red (*r*) or blue (*b*)) and there are three trials, there are simply 2^3 or eight possible hypotheses (*rrr*, *rrb*, *rbr*, *rbb*,..., *bbb*). Thus, assuming independence between trials, the probability of any particular hypothesis (e.g., *rrb*) is simply the probability of sampling each chip (i.e. $(.8) * (.8) * (.2)$). In this way, we can compute probabilities for all eight hypotheses. We compared the expectation to the observed distribution of children in the *Short Wait* and *Long Wait* conditions (see Table 1).

Table 1: Pattern of responses expected under independent sampling compared with frequencies in the *Long Wait* and *Short Wait* conditions.

Responses	Expectation	*Long Wait*	*Short Wait*
red,red,red	.512	10	1
red,red,blue	.128	1	1
red,blue,red	.128	2	10
red,blue,blue	.032	3	0
blue,red,red	.128	0	1
blue,red,blue	.032	1	6
blue,blue,red	.032	1	1
blue,blue,blue	.008	2	0

Chi-squared tests revealed a significant difference between children's responses in the *Short Wait* condition to both the *Long Wait* condition, $\chi^2(7, N = 40) = 22.3, p < .05$, and to the expected distribution, $\chi^2(7, N = 20) = 18.6, p < .05$.[1] However, the difference between the *Long Wait* condition and the expected distribution was not statistically significant, $\chi^2(7, N = 20) = 6.57, p = ns$. This suggests that while children in the *Long Wait* condition were providing responses that followed the predictions of independent samples, children in the *Short Wait* condition were doing something else. Indeed, a quick examination of Table 1 suggests that children in the *Short Wait* condition were simply alternating responses. To directly compare the two conditions, we coded children's responses in terms of whether they repeated a response (e.g. "red" then "red" again) or alternated (e.g. "red" then "blue"). Comparing condition by repetition/alternation revealed significant differences both when we coded for repetition/alternation over all three responses, Fisher Exact $(N = 33), p < .0001$, and when we coded for repetition/alternation over two responses, $\chi^2(1, N = 80) = 29.5, p < .0001$.

Another way to think about dependency is to model children's responses as a Markov process and consider the transition matrix. We computed the empirical frequencies with which children moved from a "red chip" response to a "blue chip" response, and so forth (see Table 2). If children are producing independent samples, the probability of producing a particular response should be the same regardless of the previous response. However, this analysis revealed a strong dependency between responses in the *Short Wait* condition, Fisher Exact $(N = 20), p < .0001$, and a much weaker dependency in the *Long Wait* condition, Fisher Exact $(N = 20), p = .03$. These results suggest that although children's pattern of responses in the *Long Wait* condition were close to the predicted distribution, there were still dependencies between a single child's guesses.

[1] Because the approximation to the χ^2 distribution is unreliable with small cell entries, we computed the null distribution numerically. We generated 10,000 contingency tables with these frequencies, computed χ^2 for each, and then computed p values by examining the quantile of the observed χ^2 value.

Table 2: Transition matrices in the two conditions.

	Long Wait		*Short Wait*	
	Next *r*	Next *b*	Next *r*	Next *b*
Current *r*	21	7	4	17
Current *b*	4	8	18	1

We conducted one final analysis to rule out the hypothesis that children in the *Short Wait* condition showed more dependency in responses than children in the *Long Wait* condition purely because children in short wait were more likely to remember their guesses. If children in the *Long Wait* condition had simply forgotten their previous responses more often than children in the *Short Wait* condition, they would be much less likely to show dependencies between guesses simply due to memory differences between trials. Recall that at the conclusion of the experiment children were asked which color block they had said fell in on each of the three previous trials. Looking at whether children answered all questions correctly, we found no difference in memory between conditions, $\chi^2(1, N = 32) = 3.14, p = .08$. However, because there was arguably a marginal difference between conditions, we also gave the children a memory score from 0-3 depending on how many memory questions children answered correctly; comparing memory scores also revealed no significant differences, $t(30) = -1.52, p = ns$.

Discussion

Our experiment examined whether children's responses in a simple causal reasoning task could be accounted for in terms of sampling from a probability distribution. The results of the experiment provide evidence in support of the sampling hypothesis in children. First, children's behavior reflected probability matching. That is, as a group, children provided a percentage of red and blue guesses that corresponded with the actual distribution of red and blue blocks in the population, rather than maximizing and choosing the red block on every guess or randomly guessing 50% of each color. Second, children's responses reflected the predicted patterns of independence and dependence across conditions. After delays of one week, children showed a greater amount of independence between guesses than did children who did not experience a delay. Furthermore, in contrast to results from the *Long Wait* condition, analyses of the *Short Wait* condition revealed that individual children showed strong dependence between their three guesses; thus these children were not randomly sampling from the distribution.

One might ask whether the findings suggesting that children are probability matching in our experiment were an artifact of our particular design. If children were aware that they would be asked the same question multiple times, they might not have been motivated to provide an optimal response, knowing that they would have two more chances to provide guesses. However, children were not aware that they would be playing the game multiple times in either the *Long Wait* or the *Short Wait* conditions. Furthermore, it is unlikely that such young children would be capable of engaging in such sophisticated planning. Moreover, at the conclusion of the three trials we asked children whether they remembered the guesses they had provided on each trial. Across both conditions children's memories were fuzzy, with the majority of them only being able to accurately report one or two of their initial responses in both the *Long Wait* and *Short Wait* conditions and there was no difference between conditions on the memory check.

Although our findings were consistent with the sampling hypothesis in that children both probability matched and displayed greater independence of guesses given a time delay, we did not find evidence for a "wisdom of crowds" effect. The wisdom of crowds predicts that when guesses are aggregated across individuals, this should provide a score that is closer to the actual distribution than the individual guesses alone. Instead, we found no differences between children's first guess and the majority of three responses, $\chi^2(1, N = 40) = .23, p = ns$. Given that Vul and Pashler (2008) found this benefit with adults, we might have expected to find a similar increased advantage of aggregation with children. However, we elected to use a forced choice paradigm due to the young age of our participants, and this may have reduced the sensitivity of our measure such that we were unable to detect the effect. In future work we may explore this further by designing a task that would allow children to make more fine-grained responses.

Future Work

Future work will continue to evaluate the sampling hypothesis in children to investigate the role of evidence in children's hypothesis generation and sampling. For example, we are looking at whether young children are capable of rapidly updating hypotheses based on evidence during a causal learning task. The prediction following the Sampling Hypothesis is that children will update their hypothesis space following either confirming or disconfirming evidence and will adjust their predictions accordingly, and should sample their next hypothesis from the remaining possible hypotheses.

Another future direction will involve investigating the sampling hypothesis in even younger children and current research suggests some possible appropriate methods. In an experiment examining single-event probability, Denison and Xu (in press) used a crawling procedure to show that 13-month-old infants can make predictions about single-event probability. They used two trials, one to establish which of two object-types individual infants preferred and another to test probabilistic inference. They showed infants two large populations of objects, one with a 4:1 ratio of desirable: not-desirable objects and the other with the opposite ratio. The experimenter removed a single item from each of the two populations one at a time and placed them into separate opaque containers. The infant was then encouraged to crawl to the container of their choice. Findings suggested that infants could predict which of the two populations would most likely yield a single-item

sample of their preferred object.

Finally, although other work suggests that children do demonstrate graded sensitivity to probabilities with similar designs (Bonawitz et al., 2008) and we chose a sample probability that maximized the difference between chance response and a strategy of maximizing, further conditions could strengthen our findings here by demonstrating that children's responses match probabilities across an array of values. For example, ongoing studies in our lab suggest that preschool-aged children's first responses do also match to samples where the probabilities are 19:1, 15:5, 12:6, and 10:10. Furthermore, we can demonstrate that children can sample from probability distributions in a more complex hierarchical sampling task. We have adapted the current procedure to show children an overall population of blocks that is physically separated into two sub-populations with different distributions. This design allows assessment of children's ability to make valid probabilistic inferences when they must take into account the condition that the block is being sampled from only one of the two sub-populations.

Conclusions

The current experiment provides a first step in examining the sampling hypothesis in children. Children in our experiment engaged in probability matching and demonstrated increased independence of guesses when given a time delay, suggesting that they may have engaged in a process of sampling from probability distributions. This sampling behavior may begin to explain how children navigate through the potentially infinite number of hypotheses they face at the outset of a learning process. More generally, the sampling hypothesis may also begin to explain how it is that children's behavior can appear irrational when examined individually but may actually reflect a rational strategy overall.

Acknowledgments. Thanks to participating daycares and families, as well as Tiffany Tsai, Madeline Hanson, Beth McCarthy and Jennifer Ng for help with data collection. This research was supported in part by the James S. McDonnell Causal Learning collaborative and grant IIS-0845410 from the National Science Foundation.

References

Bonawitz, E., Chang, I., Clark, C., & Lombrozo, T. (2008). Ockham's razor as inductive bias in preschoolers causal explanations. In *Proceedings of the 7th international conference of development and learning.*

Bonawitz, E., Lim, S., & Schulz, L. (2007). Weighing the evidence: Children's theories of balance affect play. In *Proceedings of the 29th annual conference of the cognitive science society.*

Carey, S. (1985). *Conceptual change in childhood*. Cambridge, MA: MIT Press.

Carey, S. (1991). Knowledge acquisition: Enrichment or conceptual change? In S. Carey & S. Gelman (Eds.), *Epigenesis of mind: Essays on biology and cognition.* Hillsdale, NJ: Erlbaum.

Denison, S., & Xu, F. (in press). Twelve- to 14-month-old infants can predict single-event probability with large set sizes. *Developmental Science*.

Goodman, N., Baker, C., Bonawitz, E., Mansinghka, V., Gopnik, A., Wellman, H., et al. (2006). Intuitive theories of mind: A rational approach to false belief. In *Proceedings of the 28th annual conference of the cognitive science society.*

Goodman, N., Tenenbaum, J., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, *32:1*, 108-154.

Gopnik, A., Glymour, C., Sobel, D., Schulz, L., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and bayes nets. *Psychological Review*, *111*, 1-31.

Gopnik, A., & Meltzoff, A. N. (1997). *Words, thoughts, and theories*. Cambridge, MA: MIT Press.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*, 354-384.

Kam, C., & Newport, E. (2009). Getting it right by getting it wrong: When learners change languages. *Cognitive Psychology*, *59:1*, 30-66.

Karmiloff-Smith, A., & Inhelder, B. (1974). If you want to get ahead, get a theory. *Cognition*, *43:3*, 195-212.

Klahr, D., Fay, A., & Dunbar, K. (1993). Heuristics for scientific experimentation: A developmental study. *Cognitive Psychology*, *25*, 111-146.

Kuhn, D. (1989). Children and adults as intuitive scientists. *Psychological Review*, *96*, 674-689.

Kushnir, T., & Gopnik, A. (2005). Young children infer causal strength from probabilities and interventions. *Psychological Science*, *16:9*, 678-683.

Kushnir, T., & Gopnik, A. (2007). Conditional probability versus spatial contiguity in causal learning: Preschoolers use new contingency evidence to overcome prior spatial assumptions. *Developmental Psychology*, *44*, 186-196.

Kushnir, T., Wellman, H., & Gelman, S. (2008). The role of preschoolers' social understanding in evaluating the informativeness of causal interventions. *Cognition*, *107:3*, 1084-1092.

Legare, C., Gelman, S., & Wellman, H. (in press). Inconsistency with prior knowledge triggers children's causal explanatory reasoning. *Child Development*.

Mozer, M., Pashler, H., & Homaei, H. (2008). Optimal predictions in everyday cognition: The wisdom of individuals or crowds? *Cognitive Science*, *32*, 1133-1147.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, *92*, 289-316.

Piaget, J. (1983). Piagets theory. In P. Mussen (Ed.), *Handbook of child psychology* (4th ed., Vol. 1). New York: Wiley.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2006). A more rational model of categorization. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society.* Mahwah, NJ: Erlbaum.

Schulz, L., Bonawitz, E., & Griffiths, T. (2007). Can being scared give you a tummy ache? naive theories, ambiguous evidence and preschoolers causal inferences. *Developmental Psychology*, *43*, 1124-1139.

Schulz, L., & Gopnik, A. (2004). Causal learning across domains. *Developmental Psychology*, *40:2*, 162-176.

Shi, L., Feldman, N., & Griffiths, T. L. (2008). Performing Bayesian inference with exemplar models. In *Proceedings of the Thirtieth Annual Conference of the Cognitive Science Society.*

Siegler, R., & Chen, Z. (1998). Developmental differences in rule learning: A microgenetic analysis. *Cognitive Psychology*, *36*, 273-310.

Sobel, D., Tenenbaum, J., & Gopnik, A. (2004). Children's causal inferences from indirect evidence: Backwards blocking and bayesian reasoning in preschoolers. *Cognitive Science: A Multidiscipinary Journal*, *28:3*, 303-333.

Vul, E., Goodman, N. D., Griffiths, T. L., & Tenenbaum, J. B. (2009). One and done? optimal decisions from very few samples. In *Proceedings of the 31st annual conference of the cognitive science society.*

Vul, E., & Pashler, H. (2008). Measuring the crowd within: Probabilistic representations within individuals. *Psychological Science*, *19:7*, 645-647.

Wellman, H. M. (1990). *The child's theory of mind*. Cambridge, MA: MIT Press.

Wellman, H. M., & Gelman, S. A. (1992). Cognitive development: Foundational theories of core domains. *Annual Review of Psychology*, *43*, 337-375.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, *114*, 245-272.

Topical Relevance and Information Quality in Cognitive Models of Web Search Behavior: Introducing Epistemic Scent into Information Foraging Theory

Peter Gerjets (p.gerjets@iwm-kmrc.de)
Knowledge Media Research Center, Konrad-Adenauer-Strasse 40, 72072 Tuebingen, Germany

Yvonne Kammerer (y.kammerer@iwm-kmrc.de)
Knowledge Media Research Center, Konrad-Adenauer-Strasse 40, 72072 Tuebingen, Germany

Abstract

Current cognitive models of Web navigation (e.g., Information Foraging Theory, IFT, Pirolli, 2007) are based on the assumption that users' behavior is guided by evaluating the topical relevance of information encountered on the Web. This "information scent" has been successfully used to model Web search behavior. In this paper, however, we claim that topicality-oriented theories like IFT need to additionally consider the evaluation of information quality in order to address a broader class of realistic search tasks. For instance, when search tasks are complex and the quality of available Web information is highly variable, Web navigation will also depend on evaluating information quality, in addition to evaluating topical relevance. In this paper we first provide a theoretical framework of quality evaluation during Web search. Second, we review two experimental studies to substantiate this theoretical framework. Finally, we propose an extension of IFT using the concept of *epistemic scent* to incorporate evaluations of quality into the theory.

Keywords: information scent; evaluation processes; complex search tasks; interface design; epistemological beliefs

Web search and information quality

With the exponential growth of information available on the World Wide Web (WWW), the Web has evolved into one of the most important information sources. Besides searching for simple and uncontroversial facts or researching product purchases, the Web increasingly serves as a rich information source for conducting research on more complex academic or science-related topics (cf. Horrigan, 2006). For instance, in the context of personal concerns of individuals, such as medicine and health care, using the Web as a supplement to the interaction with experts has achieved great popularity (Moharan-Martin, 2004).

However, as anyone can publish virtually any information on the Web, the WWW is characterized by a large variability of information quality with information sources differing dramatically with regard to Web authors' expertise and motives. As a result, the trustworthiness of online information on topics like medicine or healthcare varies considerably, with many Web sites containing misleading or even false information (Eysenbach, Powell, Kuss, & Sa, 2002). Despite this variability, different Web sources (e.g., scientific and other institutions, journalists, lay people, or companies) are usually interspersed in the results lists returned by search engines. Moreover, in many cases popular commercial or social Websites (e.g., shops or forums) that may be doubtful with regard to their motives or expertise fit exactly to search terms entered by users, so that they are listed among the highest-ranked search results on a search engine result page (SERP). Thus, even the information contained in the top search results of a SERP might turn out to be biased and one-sided, leading to premature or even wrong decisions. Accordingly, Web users may not only be required to critically evaluate the topical relevance of search results but also their quality (cf. Taraborelli, 2008) - especially when dealing with controversial issues such as the effectiveness of specific medical treatments. Contrary to this claim, however, most current cognitive models of information search on the Web focus on evaluating the topical relevance of search results, thereby neglecting issues of information quality.

In this paper, we propose an extension of one of the most influential theories of search and navigation on the Web – the Information Foraging Theory (IFT, Pirolli, 2007) – based on the results of two experimental studies. These results will be reviewed following a theoretical introduction of Web-search models and quality evaluations.

Topicality-oriented models of Web navigation

In the last decade, various computational cognitive models of Web navigation have evolved. These models are based on concepts like semantic similarity and topical relevance, such as SNIF-ACT by Fu and Pirolli (2007), CoLiDeS by Kitajima, Blackmon, and Polson (2000), MESA by Miller and Remington (2004), and CoLiDeS+ by Juvina & Van Oostendorp (2008). Although several models exist, they have all ignored the evaluation of information quality.

In this paper we will focus on the SNIF-ACT model, which is based on IFT. IFT postulates that the selection of hyperlinks (e.g., from a SERP or Web page) is determined by the strength of a so called "information scent". Information scent reflects the perceived semantic similarity between *proximal cues* (i.e., keywords or trigger words available in link labels or search results) and the current *search goal* of the user, which is defined by a desired distal information source (e.g., a Web page). A strong information scent of a hyperlink indicates a high likelihood that the source accessible via the hyperlink contains the desired information and thus increases the likelihood that the hyperlink will be selected. As IFT explains Web searching behavior based on this notion of information scent, the theory presupposes that Web searching is exclusively guided by the topical rele-

vance of Web information. The computational modeling of information search in SNIF-ACT uses spreading activation in semantic memory as a mechanism for determining semantic similarity. A strong information scent occurs when the encoding of proximal cues in semantic memory results in a substantial spread of activation to the representation of the current search goal. Activation spread according to the associative strength between concepts in memory is a standard measure to represent semantic similarity in the underlying ACT-R architecture. Figure 1 illustrates the concept of information scent (IS) for a user pursuing the goal of finding information about "medical treatments for cancer" (this is the desired distal information defining the search goal). It is assumed that the user encounters a search result like the one depicted in Figure 2, which includes the terms "cell, patient, dose, beam" (these are the available proximal cues). The arrows represent the spread of activation from the search result to the goal representation, which is used to calculate the information scent of the search result.

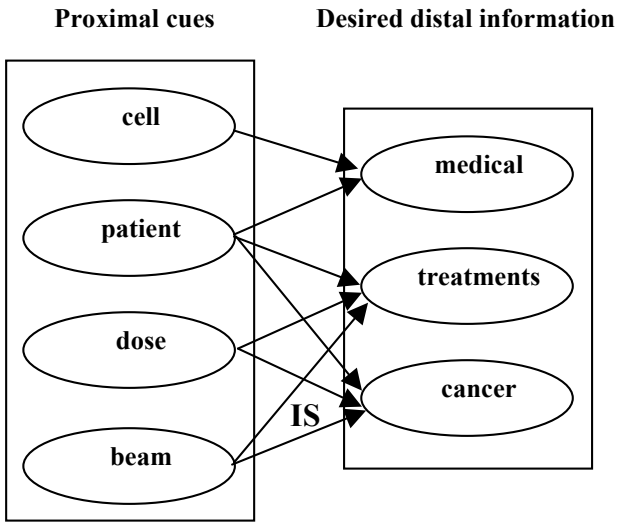


Figure 1: Illustration of information scent (IS), example adapted from Pirolli (2007).

Proton **Beam** Therapy with High-**Dose** Irradiation for Superficial and ...
Table 1 **Patient** characteristics, **dose** of irradiation, efficacy, and toxicity in superficial Supported in part by a grant-in-aid for **Scientific** Research (No. ... Earlam R., Cuhana· Oesophageal squamous **cell** carcinoma. ...
clincancerres.aacrjournals.org/cgi/content/full/9/10/3571 ·

Figure 2: Example of a search result link.

Topically-oriented computational models like SNIF-ACT have been able to predict Web search and link selection in a wide range of different tasks. Thus, at first sight they seem to allow for a successful and precise modeling of Web navigation of any kind. However, we claim that all tasks that were used for modeling forced users to focus their attention on the topical fit of available information: Users either had to engage in simple fact-finding tasks or they had at their disposal a selection of Web information that was restricted to uncontroversial and consistent information of established quality. For these types of task, quality evaluations are not an important issue. Moreover, previous studies used search environments that provided users with more or less salient topicality cues but not with salient cues pointing to the quality of search results.

Preconditions for quality evaluations on the Web

Given the search tasks and search environments used in previous studies on information scent, it seems plausible that users' Web navigation in these tasks was mainly a function of the perceived topical relevance of available information (i.e., its information scent), because quality evaluation are neither required nor supported. However, we hypothesize that the role of quality evaluations on search behavior might change considerably when certain preconditions are given with regard to task characteristics, user prerequisites, and search interface. The hypothesized interplay of these preconditions, which is illustrated in Figure 3, will be used as a theoretical framework throughout this paper.

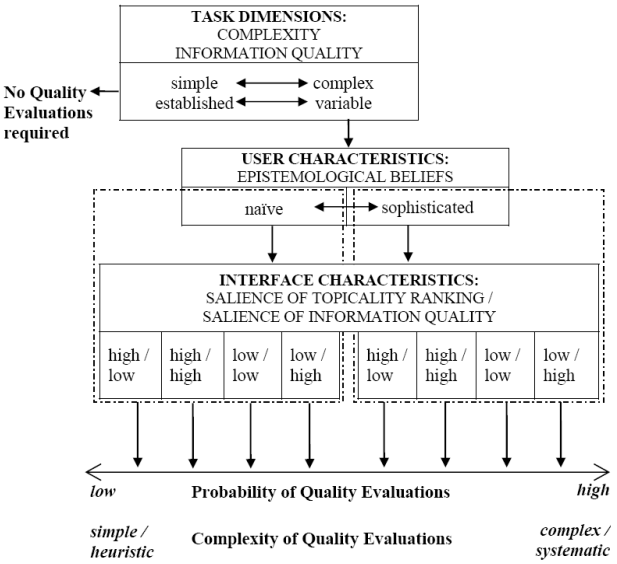


Figure 3: Preconditions of quality evaluations

Task complexity and variability of information quality

We assume that the evaluation of information quality (e.g., in terms of credibility, accuracy, and completeness) becomes of major importance (1) when the search task is sufficiently complex and, even more important, (2) when the available information is highly variable with regard to its quality. Search tasks loading high on these two task dimensions have become an increasingly important activity on the Web, for instance, when users search for controversial science-related topics or personal concerns like medical or health issues. In such search tasks, inconsistent and potentially contradictory Web information of variable quality is often encountered, so that searchers should not take the accuracy of the available information for granted. Despite the growing popularity of research on Web information quality in the last decade (for a review see Rieh & Danielson, 2007), to the best of our knowledge, the fit of topicality-

oriented models of Web navigation with users' search behavior in tasks that require the evaluation of information quality has not yet been investigated.

User prerequisites: Personal epistemology

Once the user is confronted with a search task that requires quality evaluations, both the probability and the complexity of these quality evaluations will strongly depend on the searcher's cognitive prerequisites, for instance on his or her personal epistemology. In line with dual-process theories (e.g., Chen & Chaiken, 1999), quality evaluations can range from simple, non-elaborated, intuitive, and spontaneous "heuristic" evaluations, on the one hand, to complex, cognitively elaborated, conscious, and reflected "systematic" evaluation processes on the other hand. In order to systematically evaluate the quality of Web information, searchers need to consider how credible a source of information is, how certain and consistent with other sources the information itself is, and how strongly the information might be influenced by the motives of the information provider. According to Hofer (2004) this kind of reasoning is closely connected to a person's epistemological beliefs (EBs), that is, to one's personal beliefs about the nature of knowledge and knowing. EBs have been shown to guide users' cognitive and metacognitive activities during Web search (Hofer, 2004). For instance, it has been demonstrated that users with naïve EBs are less critical Web searchers and that EBs influence search techniques and the ability to recognize authority (Hofer, 2004; Whitmire, 2003). Certainly, there exist other important cognitive prerequisites beyond EBs that support systematic quality evaluations of Web information, such as domain expertise or Web expertise. These prerequisites were, however, not investigated as factors in the studies reviewed in this paper and will therefore not be discussed in greater detail.

Search interface: Salience of topicality and quality

A third precondition for quality evaluations – beyond task requirements and user prerequisites – is related to the affordances and information provided by the search interface. We assume that even if a user is able to engage in quality evaluations required by a search task, the concrete enactment of these processes during Web search might depend on two aspects of the search interface: first, whether the search interface affords quality evaluations and second, whether it comprises quality-related information. In our opinion, the interface of popular search engines usually does not support quality evaluations with regard to these two aspects.

First, search engines usually present search results in a list, with the most topically relevant and most popular Web pages being the highest-ranked ones (cf. Cho & Roy, 2004). This list format provides a strong affordance for users to start reading at the top of the list and to follow the strict and non-ambiguous order when reading and selecting the search results presented. Thus, no affordances are provided for users to take over the responsibility for evaluating and selecting search results on their own. Rather, searchers' awareness of the ongoing selection process is hindered by the SERP layout.

Second, search engines usually display only very little information for each search result (e.g., a title, an excerpt from the respective Web page, a URL) on which evaluation processes aimed at deciding which search results to select for further inspection must be based. Moreover, the search result descriptions are typically confined to topical information, whereas quality-related source information is sparse and non-salient. Accordingly, the interface design of standard search engines does not support users to engage in quality-related evaluation processes on their own.

It can be expected that (1) the salience of topicality rankings of search results and (2) the salience of proximal cues in search results pointing to the quality of information are two important factors that determine whether quality evaluations take place or not. We assume that a search interface that provides salient proximal cues for information quality and refrains from making the topicality ranking of search results the most salient feature will stimulate more quality evaluations than a search interface without these characteristics. Thus, within the limits of users' individual cognitive prerequisites, a proper search interface might lead to navigational decisions that are based to a substantial degree on evaluating information quality in addition to evaluating topical relevance.

Hypotheses and review of experimental studies

Based on the framework describing the preconditions of quality evaluations during Web search (Figure 3), a couple of hypotheses were derived and tested in two studies reviewed in this section. In both studies, fine-grained process data (combination of eye-tracking methodology and log file data) were used to test the relationship between the probability and complexity of quality evaluations in a science-related search task and the design of the search interface and users' EBs. The task of both studies addressed a controversial medical topic. The collection of Web pages made available in the studies represented the variability of information quality on the Web and included Web pages provided by official institutions, scientific authorities, journalists, companies, and lay people (e.g., discussion pages). All Web pages were topically relevant to the respective search topic. We hypothesized that a search task with these characteristics would cause users to engage in quality evaluations, at least when their cognitive prerequisites and the search interface used would allow for these processes. Users' EBs were measured to test whether users with naïve and sophisticated EBs differ in the quality evaluations they engage in.

Two different interface design approaches were implemented to test whether they stimulate quality-related evaluation processes. In study 1 (Kammerer, Wollny, Gerjets, & Scheiter, 2009) participants either used a standard Google search result list or an augmented search result list additionally containing source categories for each search results (cf. http://www.clewwa.de/). This approach aimed at providing salient quality-related cues. In Study 2 (Kammerer & Ger-

jets, 2010) a standard list format was compared to a grid format with search results arranged in multiple rows and columns (cf. www.viewzi.com). This approach aimed at decreasing the salience of the topicality ranking and at increasing users' awareness of the selection process.

We hypothesized that both experimental interfaces would lead to more and better quality evaluations than a standard search interface with a high salience of the topicality ranking and a low salience of information quality.

Study 1: Display of search results with source categories

In this study (for details see Kammerer et al., 2009) participants were confronted with a fictitious request from an overweight friend, who wants to loose weight by changing her diet. Participants were asked to conduct a 20-minute Web search to make an informed decision between low fat and low carb diets in order to recommend one of the two diet methods. Participants were provided with three prearranged Google-like SERPs with ten search results each.

Method. Thirty university students participated in the experiment by either using a standard Google search result list or an augmented search result list (15 participants per group). The augmented list additionally contained source category labels printed in bold next to the URL. The labels indicated to which of five different source categories a search result belonged. The five source categories were Science/Institutions, Portals/Advisors, Journalism/TV, Readers' Comments, and Shops/Companies. We assumed that these source categories provided users with cues regarding the quality of the respective Web pages without changing the topical information available for each search result. Furthermore, searchers' EBs were obtained with the Epistemic Beliefs Inventory (EBI; Schraw, Dunkle, & Bendixen, 1995). In order to study participants' evaluation processes, their eye movements and mouse clicks during Web search were captured. We assumed that the amount of attention (i.e. total fixation duration) spent on a search result reflected evaluative processes with regard to this search result. As the topical information did not differ between the experimental conditions we assume that group differences in the amount of attention indicate differences in quality evaluations. Similarly, selection differences between groups cannot be traced back to differences in topicality but indicate that searchers evaluated the quality of sources differently.

Results and discussion. The results showed various differences between the two search interfaces and between naïve and sophisticated users with regard to the attention distribution on SERPs and the selection of search results. First, augmenting SERPs with source categories resulted in less linear viewing sequences than standard SERPs. Second, the availability of source categories influenced students' evaluation and selection behavior, such that they gave less attention to commercial search results ("Shops/Companies") and were more likely to select search results from the category Portals/Advisors. Third, beyond these effects of the interface design, the results revealed that source categories stimulated users with sophisticated EBs to pay more attention than naïve users to search results that were rather ambiguous with regard to their information quality (Portals/Advisors, Journalism/TV, and Readers' Comments) compared to the remaining categories Science/Institutions (high quality) and Shops/Companies (low quality). Fourth, with regard to EB effects on standard SERPs, the results indicated that sophisticated users paid less attention than naïve users to search results linked to social or commercial Websites. A possible explanation is that searchers with sophisticated EBs might be able to identify such search results as being of rather low quality by having only a quick look on the search result descriptions (e.g., the URLs). To conclude, Study 1 revealed several effects of (1) enriching search interfaces with salient quality-related cues and (2) of the personal epistemology searchers bring to the task. These two factors would be difficult to model with topicality-oriented theories of Web navigation like the IFT because the differences in attention distribution and selection behavior were not associated with differences in topical relevance.

Study 2: List interface versus grid interface

In this study (for details see Kammerer & Gerjets, 2010) users had to decide between two competing therapies for Bechterew's disease. They were given eight minutes to conduct a Web search regarding the pros and cons of both therapies and to make an informed decision between them. Participants were provided with two prearranged SERPs, one for each therapy, with nine search results each.

Method. Eighty university students participated in the experiment by either using a standard Google search result list or a grid interface with search results arranged in three rows and three columns. Furthermore, the trustworthiness order of search results on a SERP was experimentally manipulated in order to test participants' sensitivity to information quality (cf. Pan et al., 2007). The trustworthiness order of the search results presented in this study was obtained empirically in a pilot-study. Based on these data, the nine search results per SERP, which were all of high topical relevance, were either presented in an *optimal* order, with the most trustworthy search results presented first and the least trustworthy ones presented last, or in a *reversed* order, so that the least trustworthy search results were presented first. For the grid interface, trustworthiness of search results was arranged line-by-line, that is, from left to right in each of three rows. Twenty participants were assigned to each of the four conditions with trustworthiness order (optimal vs. reversed) and search interface (list vs. grid) varied as between-subjects factors. Searchers' epistemological beliefs were obtained with the Internet-Specific Epistemology Questionnaire (ISEQ, Strømsø & Bråten, 2010). Searchers' eye movements and mouse clicks were captured during Web search. Additionally, retrospective verbal protocols were obtained by asking participants post-hoc to think aloud while watching a replay of their own eye movements during search.

Results and discussion. The results showed numerous differences between the two search interfaces, between the two

trustworthiness orders, and between naïve and sophisticated users with regard to the attention distribution on SERPs, the selection of search results, and the occurrence of quality-related verbal utterances. First, the grid interface caused less homogenous and less linear viewing sequences on SERPs than did the list interface (for both trustworthiness order conditions). Second, when using the list interface most attention was given to the search results on top of the list – independent of their trustworthiness. In contrast, with a grid interface, nearly all search results on a SERP were attended for equivalent durations. Consequently, when search results were presented in a *reversed* order, participants using the list interface attended significantly longer to the least trustworthy search results and selected the most trustworthy search results significantly less often than participants using the grid interface. Third, with regard to verbal utterances, the grid interface stimulated quality-related utterances compared to the list interface, although these utterances mostly reveal simple and heuristic quality evaluations rather complex and systematic ones. Fourth, EB results showed that, with regard to searchers' selection behavior, sophisticated users better identified trustworthy sources than naïve users. With regard to verbal data, naïve users reflected less on the type of sources they had encountered. With regard to attention distribution, naïve users paid less attention to the URLs of the search results. To conclude, Study 2 revealed several effects of (1) the presentation format and presentation order implemented in the search interface and (2) of searchers' personal epistemology. These factors influenced verbal behavior, attention distributions, and selection behavior, providing evidence that at least sophisticated searchers using an interface with a low salience of the topicality ranking (i.e., the grid interface) substantially engaged in quality evaluations to guide their web navigation. Again, because the search results displayed in all experimental conditions were equivalent with regard to topical relevance, the findings obtained would be difficult to model with topicality-oriented theories of Web navigation like the IFT.

Extending Information Foraging Theory

Based on the theoretical framework illustrated in Figure 3 we predicted that – given certain preconditions – Web navigation would be substantially guided by quality evaluations in addition to topicality evaluations. The two experimental studies reviewed confirmed these expectations. Searchers in different experimental conditions were presented with search results that were equivalent with regard to topical relevance. Experimental manipulations involved the presentation format (list versus grid), the trustworthiness order (optimal versus reversed) and the availability of quality-related proximal cues (source categories). Additionally, we distinguished searchers with naïve and sophisticated EBs. The results yielded various effects of quality-related manipulations and of searchers' EBs on attention distribution, selection behavior, and verbal utterances. IFT and other topicality-oriented models of Web search would not have predicted these effects, because the topical relevance of search results remained unaffected by the manipulations. We propose to extend IFT in three ways to account for the data we obtained. Our suggestions are illustrated in Figure 4, which refers to the example introduced in Figures 1 and 2 (medical treatments for cancer).

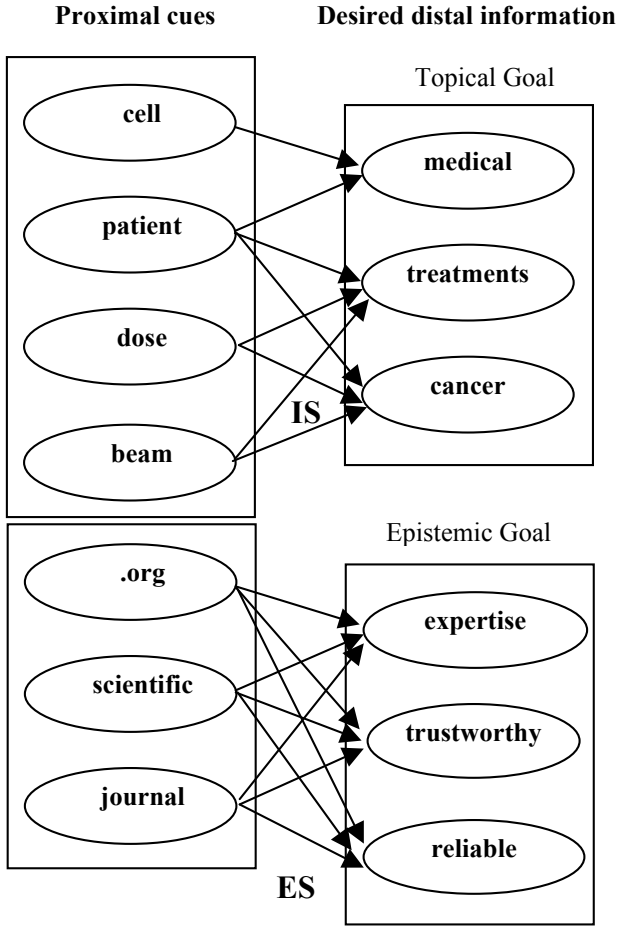


Figure 4: Extension of Information Foraging Theory

Tasks that require quality evaluations. IFT claims that Web search is guided by a topical goal, namely the goal of finding topically relevant information irrespective of its quality. In order to account for our data, however, it is necessary to introduce more complex goal structures that comprise an additional epistemic goal component (e.g., find *trustworthy* information of topical relevance). In order to decide which tasks require an epistemic goal component leading to quality evaluations, additional procedural knowledge is necessary to trigger the epistemic goal component (e.g., in cases in which contradictory information or information of variable quality is encountered during Web search).

Epistemic scent as a guiding parameter. When an epistemic goal component is active due to the characteristics of the search task and the nature of the search results encountered, a second scent parameter becomes available, namely the spread of activation from proximal cues for information

quality (e.g., the words "scientific", "journal" or ".org" in a search result description) to the representation of the current epistemic goal component (e.g., reliable and trustworthy information provided by experts). This epistemic scent (ES) based on information quality, can be taken into account in addition to the topicality-based information scent for guiding Web navigation. An open issue might relate to the integration of information scent with epistemic scent (e.g., by summing up activations, applying "metacognitive" rules).

Epistemic knowledge: Concepts and rules. To account for effects of EBs and quality-related cues on SERPs we suggest not only to model searchers' domain knowledge but also their epistemic knowledge. Epistemic knowledge comprises conceptual knowledge (e.g., knowing that information in a scientific journal provided by experts is trustworthy, see the lower part of Figure 4). Conceptual epistemic knowledge is necessary to interpret quality cues on SERPs and to judge the epistemic scent of search results. Epistemic knowledge also comprises procedural rules that guide systematic quality evaluations (e.g., recognizing good and unbiased Web information) and allow to handle information of variable quality (e.g., selection and attention behavior). These procedural rules will, however, strongly depend on whether search interfaces provide the information necessary for their application. Conceptual and procedural components of epistemic knowledge together can be used to model the influence of searchers' EBs on Web navigation. The proposed extensions of IFT would broaden its scope to include search tasks that require quality evaluations. Based on these extensions, IFT could be used to model aspects of users' Web navigation behavior that are not determined by topicality alone. Furthermore, these extensions are in line with the general assumptions of IFT and with our framework on the preconditions of quality evaluations (Figure 3). Moreover, they are consistent with the pattern of results obtained in the two experimental studies reviewed in this paper. Finally, they would allow for novel predictions on how domain and epistemic knowledge in combination can affect quality evaluations due to their associations in semantic memory.

References

Chen, S., & Chaiken, S. (1999). The heuristic-systematic model in its broader context. In S. Chaiken & Y. Trope (Eds.), *Dual-Process Theories in Social Psychology* (pp. 73-96). New York: Guilford Press

Cho, J., & Roy, S. (2004). Impact of search engines on page popularity. In *Proceedings of the 13th international conference on World Wide Web.* WWW 2004. ACM Press, New York, NY, 20-29.

Eysenbach, G., Powell, J., Kuss, O., & Sa, E.-R. (2002). Empirical studies assessing the quality of health information for consumers on the World Wide Web. A systematic review. *Journal of the American Medical Association, 287*, 2691-2700.

Fu, W-T.F., & Pirolli, P. (2007). SNIF-ACT: a cognitive model of user navigation on the World Wide Web. *Human Computer Interaction, 22,* 355-412.

Hofer, B. K. (2004). Epistemological understanding as a metacognitive process: Thinking aloud during online searching. *Educational Psychologist, 39*, 43-55.

Horrigan, J. (2006). The internet as a resource for news and information about science. *Pew Internet & American Life Project.* Retrieved from http://www.pewinternet.org/ Reports/2006/The-Internet-as-a-Resource-for-News-and-Information-about-Science.aspx

Juvina, I. & Van Oostendorp, H. (2008). Modeling semantic and structural knowledge in Web navigation. *Discourse Processes, 45*, 346-364.

Kammerer, Y., & Gerjets, P. (2010). How the interface design influences users' spontaneous trustworthiness evaluations of Web search results: Comparing a list and a grid interface. In C. Morimoto, & H. Instance (Eds.), *Proceedings of the 2010 Symposium on Eye Tracking Research & Applications ETRA '10 (pp. 299-306)*. New York, NY: ACM.

Kammerer, Y., Wollny, E., Gerjets, P., & Scheiter, K. (2009). How authority-related epistemological beliefs and salience of source information influence the evaluation of Web search results – An eye tracking study. In N. A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 2158-2163). Austin, TX: Cognitive Science Society.

Kitajima, M., Blackmon, M.H., & Polson, P.G. (2000). A Comprehension-based Model of Web Navigation and Its Application to Web Usability Analysis. *Proceedings of CHI 2000,* ACM Press*,* 357-373.

Miller, C. S., & Remington, R. W. (2004). Modeling information navigation: Implications for information architecture. *Human-Computer Interaction*, *19*, 225-271.

Moharan-Martin, J. M. (2004). How internet users find, evaluate, and use online health information: A cross-cultural review. *CyberPsychology & Behavior, 7*, 497-510.

Pan, B., Hembrooke, H., Joachims, T., Lorigo, L., Gay, G., & Granka, L. (2007). In Google we trust: Users' decisions on rank, position, and relevance. *Journal of Computer-Mediated Communication, 12*, 801–823.

Pirolli, P. (2007). *Information Foraging Theory. Adaptive interaction with information*. New York: Oxford University Press.

Rieh, S. Y. & Danielson, D. R. (2007). Credibility: A multidisciplinary framework. In B. Cronin (Ed.), *Annual Review of Information Science and Technology* (Vol. 41, pp. 307-364). Medford, NJ: Information Today.

Strømsø, H.I., & Bråten, I. (2010). The role of personal epistemology in the self-regulation of Internet-based learning. *Metacognition and Learning, 5,* .91-111.

Taraborelli, D. 2008. How the Web is changing the way we trust. In A. Briggle, K. Waelbers, P.A.E. Brey (Eds.), *Current Issues in Computing and Philosophy*. IOS Press, Amsterdam.

Whitmire, E. (2003). Epistemological beliefs and the information-seeking behavior of undergraduates. *Library & Information Science Research, 25*, 127-142.

Simulating the Elimination and Enhancement of the Plosivity Effect in Reading Aloud

Qiang Liu (qxliu@ucsc.edu) & Alan H. Kawamoto (ahk@ucsc.edu)
Department of Psychology, University of California at Santa Cruz
Santa Cruz, CA 95064 USA

Abstract

In this paper, we consider articulatory processes in a connectionist model of reading aloud to account for effects of manner of articulation of the initial segment in a variety of tasks. We first describe experimental results showing how flexibility in articulation can completely eliminate the *a priori* acoustic latency difference between plosives and non-plosives in some tasks, and exaggerate this difference in other tasks. We then simulate an expanded version of the Connectionist Incremental Articulation Model that incorporates Stevens' (1998) 3 phases involved in articulating a speech segment.

Keywords: Phonological Priming; Word Naming; Speech Production; Segment Duration; Jordan Network; Minimal Unit of Articulation; Response Criterion

The Precise Articulatory Sequence in the Production of Individual Sounds

Researchers have known that the onset of the articulatory response occurs long before the onset of the acoustic response (Bell-Berti & Harris, 1981). This asynchrony between the onset of articulatory and acoustic responses is due to the fact that individual sounds are produced in three distinct articulatory phases: (1) the movements of the articulators toward the formation of the oral constriction; (2) the flow of air behind or around the oral constriction; and (3) the release of the current constriction and movement toward the next constriction (Stevens, 1998).

The exact point at which the acoustic event is produced in the sequence above for initial segments that differed in manner can be distinguished according to their specific articulatory requirements. For non-plosive segments such as /m/, /l/, /r/, /s/, etc., acoustic energy can be generated shortly after their respective oral constriction are formed. However, for plosive segments such as /p/, /t/, /k/, etc., acoustic energy can be generated only after (a) the buildup of sufficient intra-oral pressure, and (b) the release of the current oral constriction. That is, acoustic onset for non-plosive segments occurs during the second phase, whereas for plosive segments, it occurs during the third.

The Plosivity Effect

This differential requirement for the production of the acoustic events of plosive and non-plosive segments is the basis for the Plosivity Effect. Because the acoustic event for plosive segments requires the buildup of the intra-oral pressure, the onset of acoustic energy (acoustic latency) for responses beginning with plosive segments is typically 50 – 100 ms slower than responses beginning with non-plosive segments, although this difference can be as short as 20 ms in speeded naming tasks (see Kawamoto, Kello, Jones & Bame, 1998). Notably, the plosivity effect can be completely eliminated in some tasks (Kawamoto, Liu, Mura, & Sanchez, 2008), or enhanced in others (Liu, Kawamoto, & Grebe, 2009).

The Elimination of the Plosivity Effect

In a study examining the temporal relationship between the onsets of the articulatory and acoustic responses in the delayed naming task, participants were presented with the complete stimulus (a monosyllabic word) at the beginning of a trial and were told to respond as quickly and accurately as possible only after a signal to respond was given, which was given after a delay. Using stimuli that began with the segments /p/, /t/, /m/, and /n/, Kawamoto and colleagues found that what constituted the initiation of the response to participants was the onset of the acoustic response, despite the fact that for long delays, the onset of the articulatory response occurred before the signal to respond. In essence, participants were moving their articulators into the optimal position for the phonation of the acoustic response while they were waiting for the signal to respond. For sufficiently long delays, that meant holding the acoustic response in abeyance until the signal to respond was detected. For responses beginning with non-plosive segments (/m/ and /n/), that meant holding the response at the cusp of the second articulatory phase, whereas for responses beginning with plosive segments (/p/ and /t/), it was held at the cusp of the third articulatory phase.

Basically, when participants were afforded the opportunity (i.e., long delays) to not only form the appropriate oral constriction (phase 1), but also build the required intra-oral pressure (phase 2) for plosive initial segments, the plosivity effect disappeared completely. But when there was insufficient time for the first two phases to be completed for plosive initial segments (i.e., short delays), a plosivity effect could still be observed, although the magnitude of the plosivity effect diminishes as a function of delay.

The Enhancement of the Plosivity Effect

In a different study, which examined the minimal unit of phonological information needed to initiate articulation (i.e., minimal unit of articulation), participants again produced monosyllabic utterances that began with the

segments /p/, /t/, /m/, and /n/ under a variant of the delayed naming task, called the pair-wise priming task (Liu, Kawamoto, & Grebe, 2009).

Procedurally, the pair-wise priming task is identical to the delayed naming task except that instead of presenting the complete stimulus at the beginning of the trial, only the initial letter was presented (e.g., *m__*). The complete stimulus (e.g., *mood*) was presented only after a variable delay (i.e., stimulus onset asynchrony or SOA) of 300 ms or 600 ms.

The key issue was whether or not the initial segment was sufficient for a participant to initiate the articulatory response or even the acoustic response. Liu and colleagues (2009) found that participants could in fact initiate the articulatory response on the basis of the initial segment alone. In fact, in the 600 ms SOA condition, articulatory onset for both plosive and non-plosive segments on average occurred before the presentation of the complete stimulus.

The next issue is whether knowledge beyond the initial segment is required to initiate the acoustic response. If the acoustic response can be initiated on the basis on the initial segment alone in certain conditions, one can lengthen the acoustic event for these segments until the subsequent segment becomes available. Indeed, for nasals, a 10.08 ms increase in the mean acoustic segment duration was observed in the 600 ms SOA condition compared with the 300 ms SOA condition (see Figure 1). In fact, in certain trials, acoustic onset occurred prior to the presentation of the complete stimulus. Although the total number of trials where this was observed was limited to 6 out of 474 trials in total, they represent a clear existence proof that for non-plosive initial segments, the acoustic response can be initiated based on knowledge of the initial segment alone.

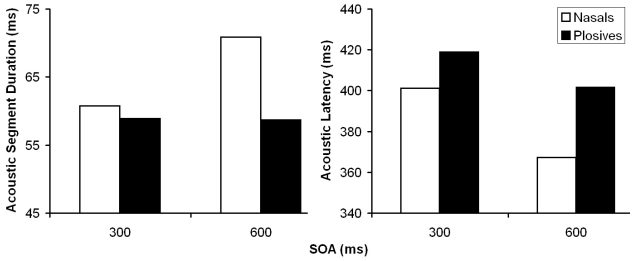


Figure 1: The acoustic segment durations (left) and acoustic latencies (right) of plosive and nasal initial segments across priming conditions reported by Liu et al., (2009).

However, the results for plosive initial segments differed from those for nasals. For plosives, acoustic onset corresponds to the explosive release of pent up pressure that require knowledge of the following segment (i.e., phase 3), and thus acoustic onset always occurs some time after the complete target is presented. Moreover, because this release of pressure occurs more or less in an all or none fashion, plosive segments are relatively resistant to acoustic lengthening. Indeed, acoustic duration for plosive segments were almost identical across the different priming conditions (58.91 ms and 58.72 ms for the 300 ms and 600ms SOA, resp.). So, despite the fact that initiation of the articulatory response for plosive initial segments does not require knowledge of the subsequent segment, the initiation of the acoustic response is contingent on the following segment.

Given that the initiation of the acoustic response for plosive initial segments must wait until the next segment is known but it does not for non-plosive segments, the onset of the acoustic response for non-plosive initial segments can be initiated much earlier than plosive initial segments, particularly in the 600 ms SOA condition. This differential constraint is what was driving the 16.64 ms enhancement of the plosivity effect on acoustic latency observed across the different priming conditions (i.e., the difference between the 17.77 ms plosivity effect in the 300 ms SOA condition and the 34.42 ms plosivity effect in the 600 ms SOA condition as illustrated in Figure 1).

Individual Differences: A Case for multiple Response Criteria. Although the results of Liu and colleagues' (2009) study represent the strongest evidence to date that participants can initiate both the articulatory and acoustic responses before the full phonological code of a word is generated, not all participants behaved in this manner. In fact, there is a wide range of individual differences.

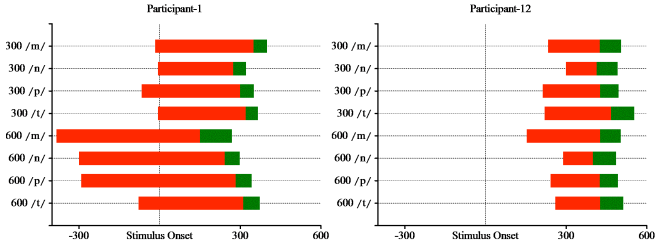


Figure 2: Data for two participants (1 and 12) reported by Liu et al., (2009). The articulatory onset to acoustic onset interval (AAI) is shown in red, and the acoustic segment duration is shown in green.

At one extreme (e.g., participant 1 in Figure 2), there are participants who initiated the articulatory response shortly after the prime is processed and long before the complete stimulus is presented. In a sense, data from these participants typified a response strategy where articulation is based on a segmental criterion (cf. Kawamoto et al., 1998). For these participants, articulation of the initial segment was necessarily lengthened because articulation of the initial segment was initiated before the subsequent segment was available. Since information for the next segment comes much later in the 600 ms SOA condition, these participants showed the largest AAI and acoustic segment duration effects (for nasals). When the difference in the SOA is taken into account, there is virtually no difference in when the articulatory response is initiated (articulatory latency)

Table 1: Examples of the input and output representations used in the current model. Input for the Articulatory Control structure (in bold), with the "-" denoting the Unspecified Segment unit (used only for priming), and the "$" denoting the Metrical Slot unit (specified or not). The first and last sweep represents the neutral state.

The Complete Input Plan for *mood*																																				
	Onset1										Onset2						Vowel								Coda											
Sweep	s	p	m	n	t	f	l	r	-	$	p	t	r	l	-	$	a	e	i	u	I	U	-	$	d	p	t	f	l	r	-	$				
1	0	0	0	0	0	0	0	0	**0**	**0**	0	0	0	0	**0**	**0**	0	0	0	0	0	0	**0**	**0**	0	0	0	0	0	0	**0**	**0**				
2-8	0	0	1	0	0	0	0	0	**0**	**1**	0	0	0	0	**0**	**1**	0	0	0	1	0	0	**0**	**1**	1	0	0	0	0	0	**0**	**1**				
9	0	0	0	0	0	0	0	0	**0**	**0**	0	0	0	0	**0**	**0**	0	0	0	0	0	0	**0**	**0**	0	0	0	0	0	0	**0**	**0**				

Corresponding Target Output Values																			
Sweep	Articulatory Phase of Current Segment												Velum	Tongue Tip	Tongue Body	Lip Vertical	Lip Horizontal	Pressure	Glottis	Acoustic Energy
1	0	0	0	0	0	0	0	0	0	0	0	0	0.1	0.5	0.2	0.8	0.8	0	0.5	0
2	**1**	0	0	0	0	0	0	0	0	0	0	0	0.1	0.45	0.2	0.1	0.7	0	0.4	0
3	0	**1**	0	0	0	0	0	0	0	0	0	0	0.9	0.4	0.15	0.1	0.5	0.4	0.2	0.5
4	0	0	**1**	0	0	0	**1**	0	0	0	0	0	0.5	0.3	0.12	0.15	0.45	0.4	0.2	0.6
5	0	0	0	0	0	0	0	**1**	0	0	0	0	0.1	0.1	0.1	0.4	0.3	0.45	0.1	0.7
6	0	0	0	0	0	0	0	0	**1**	**1**	0	0	0.1	0.5	0.12	0.5	0.4	0.55	0.4	0.5
7	0	0	0	0	0	0	0	0	0	0	**1**	0	0.1	0.9	0.15	0.7	0.7	0.8	0.5	0.1
8	0	0	0	0	0	0	0	0	0	0	0	**1**	0.1	0.5	0.2	0.8	0.8	0.2	0.5	0.8
9	0	0	0	0	0	0	0	0	0	0	0	0	0.1	0.5	0.2	0.8	0.8	0	0.5	0

relative to the presentation of the prime.

At the same time, there are also participants who, despite the fact that the prime provided sufficient information to initiate articulation, chose to wait until more information becomes available. These participants (e.g., Participant 12 in Figure 3) typified a response strategy where articulation is based on the whole word criterion (cf. Kawamoto et al., 1998). For these participants there was little or no difference in when and how the response was executed across priming conditions.

The Incremental Articulation Account: An Expanded Implementation

In this section, we describe an expanded implementation of the Incremental Articulation Account (Kawamoto et al., 1998). The core assumptions of the incremental account are that (1) the minimal unit of articulation is the segment, (2) multiple response criteria can be used, and (3) when articulation is initiated on a segment by segment basis, the articulation of the current segment/s may be lengthened to accommodate the time needed to process the subsequent segments.

The goal of the initial implementation of this model was to account for anticipatory coarticulation and segment duration effects (Kawamoto and Liu, 2007). The goal of the current implementation is to expand the generality of this model by demonstrating how the elimination and enhancement of the plosivity effect can be accounted for when the precise sequence of articulatory events involved in articulation is considered.

The Representations Used

The input representation for the current implementation, as in Kawamoto and Liu (2007), is a slot based local representation scheme that specifies the segmental content, syllabic frame, and encoding status of a metrical slot (see Table 1). The output and state representations in the current implementation correspond to the current articulatory phase, the syllabic position of the segment being produced, and a small set of articulatory dimensions: the velar opening, the positions of the tongue tip and tongue body, the vertical and horizontal lip separations, the degree of intra-oral pressure, the constriction of the glottis, and acoustic intensity. Phase 1 of a word medial segment overlaps in the current output representation with phase 3 of the previous segment to convey the fact that these two phases are one and the same (see Table 1).

Model Architecture

The current model consists of two linked networks — a phonological network that provides the input to an articulatory network (see Figure 3).

The Phonological Network

The phonological network consists of three distinct components: a Phonological Buffer, a Response Criterion layer, and a Buffer Control structure. The Phonological Buffer functions simply as a temporary storage mechanism for the phonological code generated by the preceding processes such as phonological encoding (not modeled here). The Response Criterion layer denotes the particular set of response criteria used. In the current implementation there are three sets of units: (1) the Instruction units, (2) the Lengthening Criterion units, and (3) the Articulation Criterion units.

The Instruction units were designed to mimic the effect of instructions on the precise articulatory juncture that participants decide to hold the response in abeyance (i.e., {1, 0, 0} denotes articulatory onset, {0, 1, 0} denotes the acoustic onset, and {0, 0, 1} denotes vocalic onset). Their specific function in this implementation is to simulate the results of the delayed naming task. The Lengthening Criterion units dictate the manner in which a verbal response will be lengthened (i.e., {1, 0} for articulatory lengthening, and {0, 1} for acoustic lengthening).

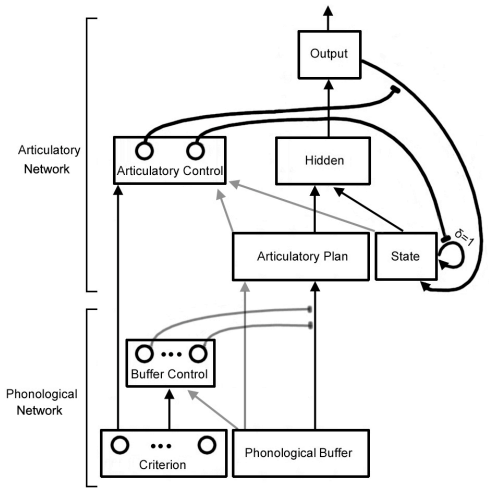


Figure 3: The basic architecture of the current model. The black arrows denote full connectivity with the exception that the connections between Output-to-State and State-to-State are one-to-one. The gray arrows denote that only some inputs are fed into the next layer (see discussion below).

Instruction and Lengthening Criterion units are fed directly into the Articulatory Control structure (see below), while the Articulatory Criterion units are fed into the Buffer Control structure. The Buffer Control structure, which also takes input from the unspecified segment and Metrical Slot units in the Phonological Buffer (i.e., "-" and "$" units), monitors the content of the Phonological Buffer and gates the flow of information to the articulatory network according the specific articulation criterion used.

Phonological Gating. The gating of information from the Phonological Buffer to the articulatory network is carried out by the Buffer Control units through Sigma-Pi connections from the Phonological Buffer to the Articulatory Plan (Buffer-to-Plan connections). All of the Buffer-to-Plan connections will be turned "off" initially for the whole word criterion (i.e., {0, 1}). They are turned "on" only when the segments in each of the specified metrical slots are known (i.e., when all of the Unspecified Segment units, "-", are 0 or off). If the segment criterion is used (i.e., {1, 0}), all of the Buffer-to-Plan connections will be turned "on" from the very beginning. Thus, the flow of information from buffer to plan will always be unrestricted. In this respect, under the segment criterion, an incomplete syllabic code can be fed into the articulator network and incrementally updated to reflect ongoing processing, whereas under the whole word criterion, only the complete syllabic code is fed to the articulatory network.

The Articulatory Network

The articulatory network consists of two primary components, an Articulatory Control structure that is coupled to a Jordan net (Jordan, 1986).

Jordan Net. The basic architecture of the Jordan net in the current implementation remains relatively unchanged from that in Kawamoto and Liu (2007). As in Kawamoto and Liu (2007), the connections between Output-to-State and State-to-State are one-to-one. The decay parameter is set to equal 1 (i.e., δ=1; no decay). Due to the actions of the Sigma-Pi connections from the control structure (see below), only the output of a single sweep will be buffered in the State units at any given time. Accordingly, the values of the State and Output Units are always offset by 1 sweep.

Articulatory Control Structure. The control structure is a simple feed forward network that acts as a monitoring and gating mechanism that takes input from the Instruction units, Articulation Criterion units, Lengthening Criterion units, and partial input from the Plan and State units.

The control structure checks the articulatory phase of the segment currently being produced with the availability of the following segment and gates the progress of the Jordan net according to the instruction or lengthening criterion specified by the criterion units. If information about the following segment is available, the output of the Jordan net will naturally move from producing the articulatory sequence of the current segment to those of the next. If not, the model will simply prolong the production of the articulatory phase/s according to the lengthening criterion specified. This control is accomplished by taking as inputs the Unspecified Segment units and the Metrical Slot units (i.e., the "-" and "$" units) from the Plan units and the Current Articulatory Phase units from the State units. The Unspecified Segment units denote whether or not the identity of a particular segment is unknown (1="on" or unknown), the Metrical Slot units denote whether or not a particular metrical slot is specified in the word frame (1="on" or specified), and the Current Articulatory Phase units denote the syllabic position and articulatory phase of the segment produced in the previous sweep. Together, these units feed into 2 control units, one for the Sigma-Pi connections to the Output-to-State connections and the other for the Sigma-Pi connections to the State-to-State connections, and turn them "on" or "off" accordingly. The specific operation of the Sigma-Pi connections under different situations is discussed in the simulation below.

Training Network

The Jordan net was trained on the Tlearn simulator (Plunkett & Elman, 1997). The training corpus for the Jordan net consisted of the input and output sequences corresponding to a small lexicon consisting of words such as *mood*, *neat*, *pit*, *pale*, etc. Training was carried out for 5000 epochs, with a learning rate = 0.05 and momentum = 0. Because the behavior of the Articulatory Control structure is just that of a simple table lookup, its operation was hardwired in the current simulation. The specific conditions that trigger its operation are discussed below.

Testing the Network

Simulation of certain key results of the delayed naming experiment reported by Kawamoto et al., (2008) and the pair-wise priming task reported by Liu et al., (2009) were carried out. For the delayed naming task, our interest focused on simulating the elimination of the plosivity effect at long delays. For the pair-wise priming task, the result of interest to us was the enhancement of the plosivity effect driven by participants using the segment response criterion. To demonstrate the malleability of the plosivity effect as well as the differential lengthening of the articulatory and acoustic responses, the outputs of the Articulatory Network were computed offline from the weight matrices of the Jordan net and the Articulatory Control Structure.

Delayed Naming. The test sequences for delayed naming were simply the training sequences lengthened to 17 sweeps for test sequences beginning with plosive segments and 18 for non-plosive sequences. For both test sets, the neutral state was presented for the first 4 sweeps to represent the time that it takes for the phonological code to be generated. The complete syllabic code is presented in the 5th sweep and remained so until the 14th sweep for plosive test sequences and the 15th sweep for non-plosive sequences, after which both sets of test sequences returned to neutral state on the final sweep. Input from the Instruction units to the Articulatory Control units were set at {0, 1, 0} (indicating that the acoustic response will be held in abeyance) from sweeps 2-10 to indicate the time for the signal to respond to be detected, after which the input was switched to {0, 0, 0} for the remainder of each test sequence set.

The outputs of these test sequences clearly show that for sequences beginning with plosive and non-plosive segments, the articulatory phase prior to the generation of acoustic energy is lengthened until the signal to respond is detected (i.e., sweep 10). Specifically, the Sigma-Pi connections to the Output-to-State connections were turned "on" and the connections to the State-to-State connections were turned "off" on the 5th sweep for non-plosive sequences and on the 6th sweep for plosive sequences. These actions turn the Jordan net into a feed forward network that simply updates the output of the earlier sweep. In essence, the articulation of the first articulatory phase for non-plosive sequences was lengthened until sweep 10, whereas for plosive sequences it was the second articulatoy phase. When the Sigma-Pi units switch back to their default state on sweep 10, the articulatory network turned back into the Jordan net and produced the next articulation phase (phase 2 for non-plosives and phase 3 for plosives) in sweep 10 and for the remainder of the response in subsequent sweeps.

Since acoustic energy is generated in Phase 2 for non-plosive initial segments, and phase 3 for plosive initial segments, the output of the test sequences demonstrate that, due to the input from the Instruction units, acoustic energy for these test sequences are generated in synchronicity in sweep 8 — an elimination of the plosivity effect.

Pair-wise Priming. Two different sets of test sequences were used for the pair-wise priming task to simulate the behavior of participants using different response criteria: (1) the whole word, and (2) the segment criteria. The sequence lengths for these input sets were 17 and 15, respectively. For both test sets, the first 4 sweeps represent the neutral state. On sweeps 5-9, a fragmentary syllabic code consisting of information for the initial segment only (e.g., *m*___ or *p*__) with the unspecified segment represented by the "-" unit in the appropriate metrical slots. From sweeps 10 to the penultimate sweep,

Table 2: Example of the test sequences *m*___ → *mood* for the pair-wise priming task. The values of the Current Articulatory Phase Units from the Output Layer, the input from the Lengthening Criterion units, and the actions of the Articulatory Control units are to show the sequence of events under the **Segment Criterion**.

	The Pair-wise priming input example for *m*___ → *mood*																															
	Onset1											Onset2						Vowel								Coda						
Sweep	s	p	m	n	t	f	l	r	-	$		p	t	r	l	-	$	a	e	i	u	I	U	-	$	d	p	t	f	l	r	-	$
1-4	0	0	0	0	0	0	0	0	**0**	**0**		0	0	0	0	**0**	**0**	0	0	0	0	0	0	**0**	**0**	0	0	0	0	0	0	**0**	**0**
5-9	0	0	**1**	0	0	0	0	0	**0**	**1**		0	0	0	0	**0**	**0**	0	0	0	0	0	0	**1**	**0**	0	0	0	0	0	0	**1**	**0**
10-14	0	0	**1**	0	0	0	0	0	**0**	**1**		0	0	0	0	**0**	**0**	0	0	0	**1**	0	0	**0**	**1**	**1**	0	0	0	0	0	**0**	**1**
15	0	0	0	0	0	0	0	0	**0**	**0**		0	0	0	0	**0**	**0**	0	0	0	0	0	0	**0**	**0**	0	0	0	0	0	0	**0**	**0**

			Articulatory Control Units		Corresponding Values of the Current Articulatory Phase Units (Within the Output Layer)											
	Lengthening Criterion				Onset1			Onset2			Vowel			Coda		
Sweep			State-to-State	Output-to-State												
1	0	0	0	1	.02	.00	.01	.01	.00	.00	.02	.02	.02	.01	.01	.01
2-4	1	0	0	1	.01	.01	.02	.01	.02	.02	.02	.01	.02	.01	.00	.02
5-7	1	0	**1**	**0**	**.98**	.01	.01	.00	.00	.02	.00	.01	.00	.02	.00	.01
8	0	1	0	1	.01	**1.00**	.01	.01	.01	.02	.01	.01	.01	.02	.02	.00
9	0	1	**1**	**0**	.01	**1.00**	.01	.01	.01	.02	.01	.01	.01	.02	.02	.00
10	0	1	0	1	.00	.01	**.98**	.01	.02	.01	**1.00**	.02	.00	.01	.01	.00
11	0	1	0	1	.01	.02	.01	.01	.00	.01	.00	**1.00**	.01	.02	.02	.00
12	0	1	0	1	.02	.00	.01	.00	.01	.00	.01	.02	**.99**	**.99**	.00	.01
13	0	1	0	1	.01	.01	.01	.01	.02	.01	.01	.01	.02	.01	**1.00**	.02
14	0	1	0	1	.00	.00	.01	.02	.00	.00	.00	.00	.01	.01	.01	**.99**
15	0	0	0	1	.00	.00	.00	.01	.01	.01	.02	.01	.02	.00	.00	.01

the test sequences contained the complete syllabic code (e.g., *mood* or *pit*); and neutral state on the final sweep (see Table 2).

Inputs to the criterion units, specifically, the Articulation Criterion and Lengthening Criterion units, for the whole word criterion test set were set to {0, 1} and {0, 0}, respectively. For the segment criterion test set, the input was set to {1, 0} and {1, 0} on sweeps 1-7, and switched to {1, 0} and {0, 1} for the remaining sweeps.

The output of the whole word test sets show that the first articulatory phase of the initial segment for both plosive and non-plosive test sequences was not produced until sweep 10, and subsequent articulatory phases were produced in each successive sweep until the entire response was completed on sweep 16. This is because the input from the Articulation Criterion units (i.e., {0, 1}) to the Buffer Control units coupled with input from the Unspecified Segment units (i.e., "-" units) in the Phonological Buffer turned "on" the Sigma-Pi connection from the Buffer Control units to the Buffer-to-Plan connections, which effectively shut off input from the Phonological Buffer units to the Plan units until the full phonological code became available on sweep 10. Once the Sigma-Pi connections to the Buffer-to-Plan connections were turned "off", the entire syllabic code was then fed into the articulatory network, and the syllabic response is produced uninterrupted. Because the acoustic event for the initial segment was produced on the 11th sweep for non-plosive sequences, and on the 12th sweep for plosive sequences, a plosivity effect of 1 sweep was observed.

For the segment criterion test sets, the first articulatory phase was produced on sweeps 5-7 for both the plosive and non-plosive sequences because the initial segment only became available on sweep 5. On sweep 8, the input from the Lengthening Criterion units was switched to {0, 1}, at which point the action of the Articulatory Control units reverts the Sigma-Pi connections to their default state, thus, allowing the next articulatory (phase 2) to be produced. However, on sweep 9, based on the input from the Current Articulatory Phase units within the State units (values offset from those within the Output Layer, shown in Table 2, by 1 sweep), the Articulatory Control units again turned "on" the Sigma-Pi connections to the Output-to-State connections and turned "off" the Sigma-Pi connections to the State-to-State connections. Thus, phase 2 was repeated on sweep 9 (see Table 2). When the entire syllabic code became available on sweep 10, the Sigma-Pi connections again revert back to its default state allowing the remainder of the syllabic response to be produced in subsequent sweeps. Accordingly, the acoustic event for non-plosive sequences (phase 2) was produced on sweep 8, whereas, for plosive sequences (phase 3), it was produced on sweep 10, resulting in a plosivity effect of 2 sweeps.

Although the output of these two test sets taken together produced a mean plosivity effect of 1.50 sweeps — a net plosivity enhancement of 0.50 sweeps, the exact magnitude of the enhancement ultimately depends on the proportion of participants using the segment criterion and the whole word criterion.

Conclusion

The results of the current simulations demonstrate that a single network using a sub-syllabic minimal unit and different response criteria can account for both the elimination and enhancement of the plosivity effect, as well as the differential lengthening of the articulatory and acoustic responses when the precise sequence of articulatory events involved in the generation of the individual sounds is taken into consideration. This approach can easily be extended to other latency and duration effects, both articulatory and acoustic, that can arise from a variety of processing difficulties in speech production. Moreover, with slight modifications, the current network can be easily coupled to existing models, such as the one described by Dell, Juliano, and Govindjee (1993) to account for a wider range of empirical data (e.g., latency data). Such an extension provides a way to explore the intricate coordination between different processing stages, and how potential asynchronies in the flow of information may reveal themselves in the interplay between different dependent measures such as articulatory latency, acoustic latency, and the duration of various components of a verbal response.

Acknowledgments

This work was partially supported by a grant from the UCSC faculty senate.

References

Bell-Berti, F., & Harris, K. S. (1981). A temporal model of speech production. *Phonetica, 38*, 9-20.

Dell, G. S., Juliano, C., & Govindjee, A. (1993). Structure and content in language production: A theory of frame constraints in phonological speech errors. *Cognitive Sciences, 17,* 149-195.

Jordan, M. (1986). Serial order: A parallel distributed processing approach. Technical Report 8604. San Diego: Institute for Cognitive Science. University of California.

Kawamoto, A. H., Kello, C. T., Jones, R. M., & Bame, K. (1998). Initial phoneme versus whole word criterion to initiate pronunciation: Evidence based on response latency and initial phoneme duration. *Journal of Experimental Psychology: Learning, Memory, & Cognition. 24*, 862-885.

Kawamoto, A. H., & Liu, Q. (2007). A unified account of segment duration and coarticulation effects in speech production. In D. S. McNamara & J. G. Trafton (Eds.) Proceedings of the 29th Annual Meeting of the Cognitive Science Society (pp. 1151-1156). Austin, TX: Cognitive Science Society.

Kawamoto, A. H., Liu, Q., Mura, K, & Sanchez, A. (2008). Articulatory preparation in the delayed naming task. Journal of Memory and Language, 58, 347-365.

Liu, Q., Kawamoto, A. H., & Grebe, P. R. (2008, May). Incremental articulation: Evidence from onset priming. Poster presented at the 21st Annual Conference of the Association for the Psychological Science. San Francisco, CA.

Plunkett, K. & Elman, J. L. (1997) Exercises in Rethinking Innateness: A Handbook for Connectionist Simulations. Cambridge, MA: MIT Press

Stevens, K. N. (1998). Acoustic Phonetics. Cambridge: MIT.

Ideal representations in a similarity space

Wouter Voorspoels (wouter.voorspoels@psy.kuleuven.be),

Wolf Vanpaemel (wolf.vanpaemel@psy.kuleuven.be),

Gert Storms (gert.storms@psy.kuleuven.be)

Department of Psychology, Tiensestraat 102
3000 Leuven, Belgium

Abstract

The present study provides an empirical evaluation of the ideal representation view of concept representation. We compared the ideal representation view with the more established exemplar and prototype views both in common taxonomic categories and in ad hoc categories. All three views are modeled based on underlying spatial similarity representations. Results suggest that the ideal representation is the better representation in ad hoc categories, and that the exemplar model is the better representation in the common taxonomic categories.

Keywords: concepts; category representation; computational models of concept representation; typicality; ad hoc concepts

An important and robust observation in concept representation research is that not all members of a category are equally representative of the category. For example, while a platypus is a mammal, it is not a good example of a mammal. It has many features that do not fit our image of what a mammal should be like: it has webbed feet, a beak and it lays eggs. A cow on the other hand, is a good example of a mammal to most people. In the same way, a spoon is a bad example of the category *weapons*, and a gun is a good example.

Previous research suggests that people are in agreement as to what are representative, good examples of a certain category and which examples are not (Rosch & Mervis, 1975). This graded membership structure is often referred to as the typicality gradient and has been reliably observed in a broad range of natural language categories, including common taxonomic categories (e.g. De Deyne et al., 2008) and ad hoc categories, such as goal derived categories (Barsalou, 1983, 1985)

Typicality is assumed to be closely linked to the representation of a concept (e.g., Murphy, 2002; Rosch, 1978). Theories of concept representation should therefore be able to explain the observation of a typicality gradient. The observation of a typicality gradient in different kinds of categories however, does not necessarily imply that the same processes and the same kind of concept representation underlies typicality judgments. The present study aims at evaluating different views on concept representation in different kinds of categories.

Kinds of concept representations

Two contrasting views on category representation have dominated the computational research on categories and concepts, each giving a different account of the graded internal structure of categories. In both approaches typicality is related to similarity of a category member to the category representation. The two views differ in what the category representation is assumed to consist of.

On the one hand, the prototype view states that a category is represented by an abstract summary representation, referred to as the prototype (e.g., Hampton, 1979; Posner & Keele, 1968). In this view, the concept *vehicle* is a represented by a summary of what vehicles are like on average, abstracted from specific instances of vehicles, containing information such as 'moves people or cargo from point A to point B'. The typicality of *car* for the category *vehicle* then is the similarity of *car* to this abstract prototype.

On the other hand, the exemplar view proposes that a category is represented by previously encountered instances of the category, instead of an abstract summary (e.g., Brooks, 1978; Medin & Shaffer, 1978). According to this view, typicality is conceptualized as the summed similarity of a category member to all stored members of the category. For example, the concept *vehicle* consists of memory traces of previously encountered instances of vehicles, such as *train*, *plane* and *metro* (i.e. member-categories at a lower level of abstraction). The typicality of *car* is then its summed similarity to all stored instances of *vehicle*.

Barsalou (1985) has proposed a third approach to account for the typicality gradient. Focusing on ad hoc categories – categories constructed ad hoc to serve a specific purpose, for example *things you rescue from a burning house* or *things you eat when on a die* – he proposed the idea of an ideal representation. Like a prototype representation, an ideal representation is a summary representation. Unlike a prototype which is based on average, central tendency values on the stimulus dimensions, an ideal contains extreme values on relevant dimensions. For example, a typical member of the category *things to eat when on a diet* has an extreme value on the ideal dimension 'fat percentage' – typical examples being at the extreme low end of that dimension, with a zero percentage of fat as an extreme ideal representation.

Barsalou (1985) compared a number of determinants of the typicality gradient in both common, taxonomic categories and ad hoc categories – including a prototype measure and an ideal representation measure. He found that whereas in common taxonomic categories the prototype measure was the dominant determinant of typicality, the ideal measure determined the typicality gradient of the ad hoc categories significantly.

This notion of ideal representation provides and excitingly new perspective on concept representation, but, unlike the exemplar and prototype views, it has not yet made its way into a computational model of concept representation. Recently we developed a model that attempts to translate the idea of an ideal representation to a computational model (Voorspoels, Vanpaemel & Storms, submitted) that is based on an underlying spatial similarity representation. To test whether this model is a proper translation of the notion of ideal representations, we aim at replicating the findings of Barsalou (1985) using computational models. We will compare the performance of the model that implements ideal representations to an exemplar model and a prototype model (also based on underlying similarity spaces) in common taxonomic categories and ad hoc categories. If our model is a proper implementation of ideal representations, we expect an interaction between the type of model and the kind of category. The ideal representation model should be the lesser model in the common taxonomic categories and the better model in the ad hoc categories.

Models

The models considered in the present paper are all based on underlying spatial similarity representations. In a spatial representation of a category, the members are represented by points in a M-dimensional space, and the distance between two members (i.e., between two points) is inversely related to the similarity between the two members. Such a representation is typically derived using multidimensional scaling (MDS) techniques, based on pairwise similarity data. The axes that span the similarity space of a category can be considered dimensions that are important to determine the similarity relations between members in the category. In the present study, we do not attempt to interpret the axes.

Ideal Dimension Model

The ideal dimension model (IDM) posits that an ideal dimension exists in the underlying similarity space. Each exemplar of a category has a certain value along the ideal dimension, obtained by an orthogonal projection on this dimension. The further this value is located along the dimension in the ideal direction, the more typical an exemplar is.

It is useful to think of the ideal dimension as a specific combination of (unarticulated) features. The more a member has of this combination of features, the more typical it is for the category. In the case of *things to eat when on a diet*, the ideal dimension possibly is made up by a combination of features such as fat percentage, sweetness and calories. For taxonomic categories, it is more difficult to articulate the specific combination of features that might make up the ideal. To put it somewhat trivially: a car is typical for the category of *vehicles* if it has a lot of the combination of features that make up "vehicle-ness".

Formally, the IDM assumes that judging the typicality of an item i for a category A comes down to evaluating the value of i on a certain dimension V_A. In an M-dimensional space, the typicality of item i for category A, is then given by:

$$T_{iA} = \frac{\sum_{k=1}^{M} x_{ik} x_{Ak}}{\left(\sum_{k=1}^{M} (x_{Ak})^2 \right)^{1/2}}, \qquad (1)$$

where x_{Ak} are the coordinates spanning the ideal dimension V_A, x_{ik} are the coordinates of item i, and M is the number of dimensions. We restrict x_A to be at a fixed distance from the origin. This does not pose a restriction for the ideal dimension.

The model orthogonally projects item i on the ideal dimension V_A , and returns a dimensional value relative to the origin that rises when the projection is farther in the ideal direction (i.e., the direction determined by the vector V_A). This value is considered the typicality of item i for category A.

Generalized Context Model

The generalized context model (GCM; Nosofsky, 1984, 1986) assumes that categorization decisions are based on similarity comparisons with individually stored category exemplars. Originally, the model was developed to account for categorization decisions, but it has successfully been adapted for typicality judgments (Nosofsky, 1991; Voorspoels, et al. 2008a).

Typicality of an exemplar is calculated by summing the similarity of that exemplar to all other exemplars in the category. Formally, the typicality of an exemplar i for category A is then given by:

$$T_{iA} = \sum_{j=1}^{n} \eta_{ij} , \qquad (2)$$

where η_{ij} is the similarity of exemplar i to exemplar j, with j belonging to category A.

The similarity between two exemplars is a function of the distance of the exemplars in the M-dimensional psychological space, adjusted by attentional weights – that specify which underlying dimensions are important in the similarity calculation – and a sensitivity parameter – which magnifies or shrinks the psychological space. Formally, the scaled psychological distance between two exemplars i and j is given by:

$$d_{ij} = c\left(\sum_{k=1}^{M} w_k \left| x_{ik} - x_{jk} \right|^r \right)^{1/r}, \quad (3)$$

where x_{ik} and x_{jk} are the coordinates of exemplars *i* and *j* on dimension *k*, w_k a parameter reflecting the attention weight for dimension *k*, *M* is the number of dimensions, and *c* is the sensitivity parameter. Since Euclidean distances are generally accepted to be more appropriate for integral dimensions (Shepard, 1964), we fixed *r* at 2 for the present studies.

Similarity of a stimulus *i* to another stimulus *j*, is related to psychological distance as follows:

$$\eta_{ij} = \exp\left(-d_{ij}\right), \quad (4)$$

where d_{ij} is the scaled psychological distance between exemplar *i* and *j*. The free parameters in the GCM consist of M-1 dimension weights and a scaling parameter *c*.

MDS-based Prototype Model

Within the framework of the GCM, one can easily define a prototype model (MPM; Nosofsky, 1992). Typicality of a category member then is the similarity towards the prototype of the category:

$$T_{iA} = \eta_{iP_A}, \quad (5)$$

where P_A is the prototype of category *A*. The position of the prototype in the similarity space is determined by averaging the coordinates of all category members on each axis.
The free parameters in the model are identical to the free parameters in the GCM (i.e., M-1 dimension weights and a scaling parameter).

Data

Construction of the psychological space relies on similarity data. Evaluation of the models relies on typicality data. For the common categories we used data from a recent norm study De Deyne et al. (2008). For the ad hoc categories, we collected the data. We will discuss the data for both category types in turn.

Common taxonomic categories

Eleven common taxonomic categories, from two semantic domains (animals and artifacts) were used in the present study (from de Deyne et al., 2008): *birds, fish, insects, mammals, reptiles, clothes, kitchen utensils, musical instruments, tools, vehicles* and *weapons*. The categories contain between 22 and 30 members.

Typicality measure The exemplars of each category, presented as verbal stimuli, were rated by 28 participants for goodness-of-example for the superordinate category they belonged to on a Likert-rating scale ranging from 1 for very bad examples to 20 for very good examples. The reliability of the judgments was evaluated by means of split-half correlations corrected with the Spearman-Brown formula, and ranged from .91 to .98 across the 11 categories (De Deyne et al., 2008, Table 1, p. 1033). The ratings were averaged over participants.

Similarity measure Pairwise similarity ratings were also available in de Deyne et al. (2008). Similarity of each member pair within a category was rated by 15 to 25 participants (varying across categories, not within categories). Estimated reliability of the ratings ranged from .88 and .96 across categories.

Ad hoc categories

Ten ad hoc categories were constructed, including those of Barsalou (1985): *things you put in your car, things you rescue from a burning house, things not to eat/drink when on a diet, wedding gifts, things you use to bake an apple pie, things you take to the beach, means of transport between Brussels and London, properties and actions that make you win the election, weapons used for hunting* and *tools used when gardening*.

For each of the categories, 80 participants generated at least eight members. From the resulting potential members pool, we sampled 20 to 25 members, covering the production frequency dimension.

Typicality measure The members of each category were rated for goodness-of-example by 30 participants on a Likert-rating scale ranging from 1 for very bad examples to 20 for very good examples. The reliability of the judgments was evaluated by means of split-half correlations corrected with the Spearman-Brown formula, and ranged from .94 to .98.

Similarity measure Since the members of an ad hoc category can be very divers and seemingly irrelevant to each other (e.g., tissues and candy), we did not ask participants to directly rate the similarity of each member pair within a category. Participants performed a sorting task, an often applied technique to arrive at a similarity measure for large stimuli sets (e.g., Ameel & Storms, 2006; Van der Kloot & Van Herk, 1991). We will briefly describe the procedure.

For each category, 60 participants sorted the members into piles according to whatever principle they thought was fitting, the only restriction being that there had to be more than one pile and less than the number of members in a category. Following their initial sort, they were asked to either further divide the piles they made in subgroups (when the number of piles in the initial sort was smaller than five), or to join piles together (when the number of piles was larger than five). This procedure resulted in 120 exemplar-by-exemplar matrices (on for each separate sort) for each category, each cell reflecting whether the pair was in the same pile or not. We summed the 120 matrices, arriving at one matrix per category, the summed scores in the cells reflecting the similarity between two members.

Results

The similarity measures for all 21 categories were used as input for a SAS non-metric MDS analyses, resulting in spatial representations in Dimensionalities 2 to 8. Stress values, measuring the badness-of-fit for the resulting geometric representation, showed a monotonically decreasing pattern in each category, indicating that the algorithm did not get trapped in a local minimum. Overall, the stress values dropped below .1 from Dimensionality 4 onwards for the common taxonomic categories and from Dimensionality 3 onwards for the ad hoc categories. Taking into account stress and the number of members of the categories, we will present results for the common taxonomic categories in Dimensionalities 4 to 8 and for the ad hoc categories from Dimensionality 3 to 6 (following generally used rules of thumb regarding number of dimensions and stress).

Recently, increasing attention has been drawn to the importance of a model's flexibility and complexity in model evaluation, and the necessity to penalize models that are more complex (any data pattern can be accounted for perfectly by a sufficiently complex model). Comparing the best fit a model can provide ignores this complexity, while assessing the average fit of the model across all possible parameter values balances model complexity and data fit (e.g., Pitt, Kim & Myung, 2003). This average fit is measured by the marginal likelihood. Given the differences in functional form of the GCM and IDM, the model evaluation in terms of marginal likelihoods is preferable.

The results of the model analyses are reported through model weights. The model weight of a model reflects the relative evidence that the data provide in favor of that model, within the set of all models that are evaluated. The evidence for a model is the marginal likelihood of the model – calculated by sampling the parameter space. For each sampled parameter value, one can calculate the likelihood given the prior distributions of the parameters. After a number of samples, the average of all samples will converge into an estimate of the marginal likelihood of the model.

We relied on standard uninformative priors. For the IDM, this translates to a uniform prior over all points at a certain distance of the origin. For the GCM and the prototype model, a uniform prior over the range 0 to 1 was used for the dimensional weights, adding the restriction that the dimensional weights have to sum to 1. The prior for the sensitivity parameter followed a Gamma(.001,.001) distribution.

We will first present the results of the analyses of the common categories. Then we will present the results for the ad hoc categories.

Common taxonomic categories

Figure 1 presents the model weights for all three models for the common taxonomic categories. For 9 out of 11 categories, the results are highly consistent across dimensionalities. Results are not consistent for *musical instruments* and *vehicles,* consequently making inferences regarding these categories rather difficult. We will consider the results of categories *fish* and *tools* to be consistent, since only in Dimensionality 4 they deviate from the other Dimensionalities. For *tools*, closer inspection of the underlying representation revealed that stress-values dropped below .1 from Dimensionality 5 onwards, possibly explaining the anomaly in the Dimensionality 4.

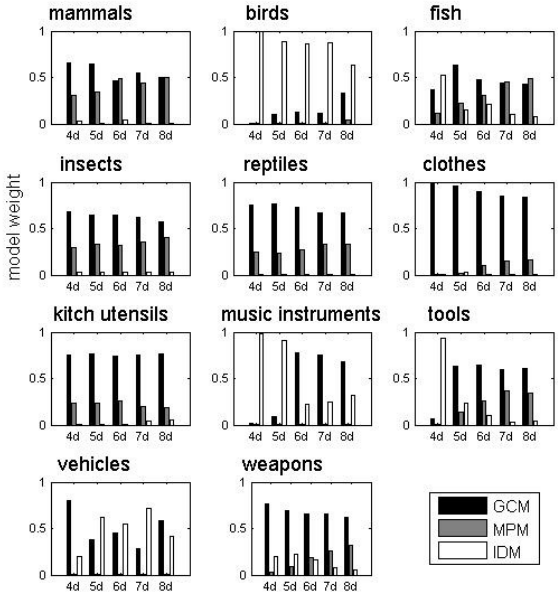


Figure 1. Model weights for the GCM, MPM and IDM for the common taxonomic categories.

It can be seen that for the 9 consistent categories, the GCM gives the better account of the typicality gradient for 8 out of 9 categories. For only 1 out of 9 categories, *birds*, the IDM clearly provides a better account. The MPM is not competitive in the present evaluation. Only for the category *fish*, it seems to provide a viable alternative in higher Dimensionalities (but even there, the MPM is not convincingly better).

In sum, the GCM seems to be the better model for the typicality gradient of the common taxonomic categories. The prototype model is never competitive, performing worse than the GCM in all categories and nearly all dimensionalities. This result confirms results of earlier comparisons between the exemplar view and the prototype view in common taxonomic concepts (e.g., Voorspoels et al. 2008) and artificial category learning (Nosofsky, 1992, Vanpaemel & Storms, 2010). The IDM possibly drives the typicality gradient of a small minority of common taxonomic categories (only *birds* in our set).

Ad hoc categories

Figure 2 presents the model weights of the three models for the ad hoc categories. For 9 out of 10 categories, the results are consistent across dimensionalities. Results are not consistent across dimensionalities for *things you take to a*

beach. Looking at the 9 consistent categories, the evidence is overwhelmingly in favor of the IDM in 7 categories. Only for the categories *hunting weapons* and *things you use when baking an apple pie* the GCM (in close competition with the MPM for the latter) is the best model. In sum, the ideal representation view indeed seems to provide a better account of the typicality gradient of ad hoc categories than the prototype and exemplar view, yet the evidence is not univocal.

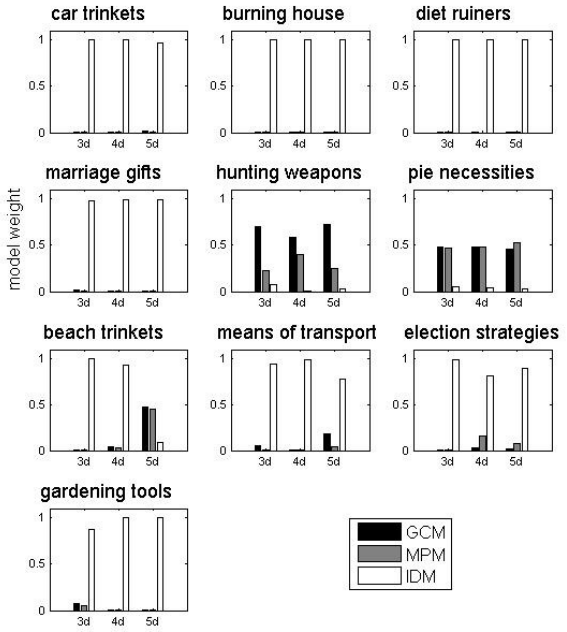


Figure 2. Model weights for the three models for the set of ad hoc categories.

The model weights reported are a relative measure of model performance, i.e., the model weight only reflects the performance of a model relative to a set of competitive models. To our knowledge however, the representational mode and the computational models used in the present study have not been applied to ad hoc categories. It is therefore informative to evaluate whether the models can give a sufficient account of the typicality gradient in absolute terms.

To this end we calculated correlations between observed and predicted typicality scores, using the optimal parameter values for each model. Results of these analyses are presented in Figure 3. It can be seen in Figure 3 that correlations rise above .6 for all categories in which the IDM is to be preferred based on the model weights, except for *properties and actions that make you win the election* and *means of transport between Brussels and London*. For the categories in which evidence based on the model weights was not in favor of the IDM, or the model weights were not consistent across dimensionalities, the optimal correlations are generally somewhat lower.

Discussion

The present study focused on the IDM, a model that provides a computational account of the notion of an ideal representation in the context of spatial similarity representations. The IDM was evaluated in its account of the typicality gradient both common taxonomic categories and ad hoc categories and compared to the GCM, arguably the most successful exemplar model, and the MPM. Following earlier findings by Barsalou (1985), we hypothesized that the IDM would have difficulty accounting for the typicality gradient of the common taxonomic categories, but that it would give a better account of the typicality gradient of ad hoc categories.

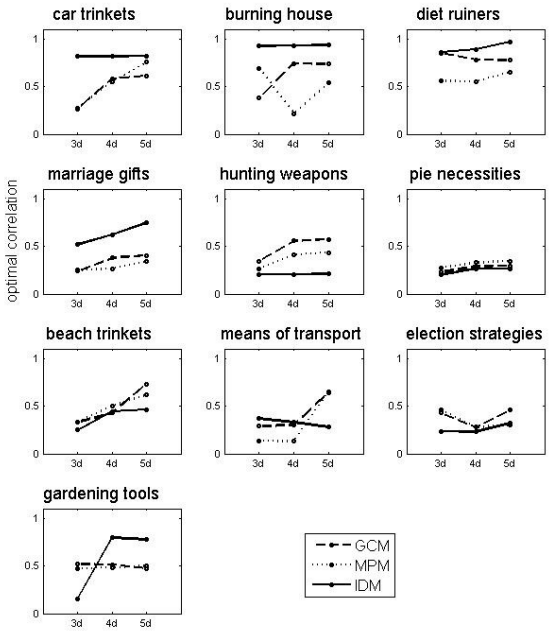


Figure 3. Optimal correlations between observed and predicted typicality ratings as a function of Dimensionality

The results supported the hypothesis. While evidence was not consistent across dimensionalities for 3 out of 21 categories, the overall pattern clearly showed the expected interaction: in the common taxonomic categories, the GCM was the better model – as can be expected based on earlier findings – and in the ad hoc categories the IDM was the better model. The evidence in any case strongly suggests that the typicality gradient of common taxonomic categories and of ad hoc categories is determined by a different representation. Moreover, the results support the reasonableness of the IDM as a formal implementation of Barsalou's (1985) notion of ideal representation.

It is unclear why this pattern broke down in 3 out of the 16 "consistent" categories. For *fish*, the IDM was the better model. In *hunting weapons* and *things you use to make an apple pie*, the GCM (MPM respectively) was the better model. Note however that for *things you use to make an*

apple pie, none of the models could give a good account of the typicality gradient in terms of optimal correlations (see Figure 3). This might suggest that the typicality gradient in this category is driven by yet another process, different from than the ones under consideration. For *hunting weapons*, the category might be considered a well-established category, rather than an ad hoc category.

To a certain extent, this study is a replication of Barsalou's work on ad hoc categories and ideal representations (Barsalou, 1985). There are, however, three crucial differences. First, we compared the ideal dimension approach to (advanced implementations of) both a prototype approach and an exemplar approach. This is important, since in this study, and in previous studies (e.g., Voorspoels et al., 2008) it is found that the exemplar approach is to be preferred over the prototype approach in concept representation.

Second, Barsalou (1985) used a priori ideals, which were generated intuitively by the researchers, for which all members of the relevant category were rated. No such instruction takes place with the IDM.

Third, Barsalou (1985) evaluated the relative contribution of different determinants of typicality, such as ideals and central tendencies, using regression analyses and a number of measures of these determinants. We tested and compared computational models of typicality that are derived from assumptions concerning concept representation. Importantly, we developed a computational model that introduces the notion of ideal representation to the context of underlying spatial representations in an intuitive way. An important finding of the present study is that the IDM indeed can be considered a computational model of ideal representations, which can be usefully applied in the further investigation of differences between concepts in terms of concept representation.

Acknowledgments

The Research in this article is part of research project G.0281.06 sponsored by the Belgian National Science Foundation – Flanders, given to the third author. We want to thank Sander Vanhaute for his help with collecting data.

References

Ameel, E., & Storms, G. (2006). From prototypes to caricatures: Geometrical models for concept typicality. *Journal of Memory and Language, 55*, 402-421.

Barsalou, L.W. (1983). Ad hoc categories. *Memory & Cognition, 11*, 211-227.

Barsalou, L. W. (1985). Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11, 4*, 629 - 654.

De Deyne, S., Verheyen, S., Ameel, E., Vanpaemel, W., Dry, M., Voorspoels, W., & Storms, G. (2008). Exemplar by feature applicability matrices and other Dutch normative data for semantic concepts. *Behavioral Research Methods, 40*, 1030-1048.

Hampton, J. A. (1979). Polymorphous concepts in semantic memory. *Journal of Verbal Learning and Verbal Behavior*, **18**, 441-461.

Medin, D. M., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, **85**, 207-238.

Murphy, G. L. (2002). *The Big Book of Concepts*. Cambridge: MIT Press.

Nosofsky, R. N. (1984). Choice, Similarity, and the context model of classification. *Journal of Experimental Psychology: Learning, Memory and Cognition, 10*, 104-114.

Nosofsky, R. N. (1986). Attention, Similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General, 115*, 39-57.

Nosofsky, R. N. (1991). Typicality in logically defined categories: Exemplar-similarity versus rule instantiation. *Memory & Cognition, 19*, 131-150

Nosofsky, R., N. (1992). Exemplars, prototypes, and similarity rules. In A. F. Healy, & S. M. Kosslyn (Eds.), *Essays in honor of William K. Estes*. Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc.

Pitt, M., A., Kim, W., & Myung, J. (2003). Flexibility versus generalizability in model selection. *Psychonomic Bulletin & Review, 10,* 29-44.

Posner, M.I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology, 3*, 392-407.

Rosch, E. (1978). Principles of categorization. In E. Rosch and B. B. Lloyd (Eds.), *Cognition and categorization (pp. 27-48)*. Hillsdale, NJ: Erlbaum.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology, 7*, 573-605.

Shepard, R. N. (1964). Attention and the metric structure of stimulus space. *Journal of Mathematical Psychology, 1*, 54-87

Van der Kloot, W. A., van Herk, H. (1991). Multidimensional scaling of sorting data: A comparison of three procedures. *Multivariate Behavioral Research, 26 (4)*, 563-581.

Vanpaemel, W., & Storms, G. (in press). Abstraction and model evaluation in category learning. Behavior Research Methods.

Voorspoels,W., Vanpaemel, W., Storms, G. (2008). Exemplars and prototypes in natural language concepts: a typicality based evaluation. *Psychonomic Bulletin & Review, 15, 3*, 630-637.

Wagenmakers, E. J., & Farrel, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review, 11*, 192-196.

Learning Structured Generative Concepts

Andreas Stuhlmüller, Joshua B. Tenenbaum, Noah D. Goodman
Brain and Cognitive Sciences, MIT
{*ast, jbt, ndg*}*@mit.edu*

Abstract

Many real world concepts, such as "car", "house", and "tree", are more than simply a collection of features. These objects are richly structured, defined in terms of systems of relations, subparts, and recursive embeddings. We describe an approach to concept representation and learning that attempts to capture such structured objects. This approach builds on recent probabilistic approaches, viewing concepts as generative processes, and on recent rule-based approaches, constructing concepts inductively from a language of thought. Concepts are modeled as probabilistic programs that describe generative processes; these programs are described in a compositional language. In an exploratory concept learning experiment, we investigate human learning from sets of tree-like objects generated by processes that vary in their abstract structure, from simple prototypes to complex recursions. We compare human categorization judgements to predictions of the true generative process as well as a variety of exemplar-based heuristics.

Introduction

Concept learning has traditionally been studied in the context of relatively unstructured objects that can be described as collections of features. Learning and categorization can be understood formally as problems of statistical inference, and a number of successful accounts of concept learning can be viewed in terms of probabilistic models defined over different ways to represent structure in feature sets, such as prototypes, exemplars, or logical rules (Anderson, 1990; Shi, Feldman, & Griffiths, 2008; Goodman, Tenenbaum, Feldman, & Griffiths, 2008). Yet for many real world object concepts, such as "car", "house", "tree, or "human body", instances are more than simply a collection of features. These objects are richly structured, defined in terms of features connected in systems of relations, parts and subparts at multiple scales of abstraction, and even recursive embedding (Markman, 1999). A tree has branches coming out of a trunk, with roots in the ground; branches give rise to smaller branches, and there are leaves at the end of the branches. A human body has a head on top of a torso; arms and legs come out of the torso, with arms ending in hands, made of fingers. A house is composed of walls, roofs, doors, and other parts arranged in characteristic functional and spatial relations that are harder to verbalize but still easy to recognize and reason about. Besides objects, examples of structured concepts can be found in language (e.g. the mutually recursive system of phrase types in a grammar), in the representation of events (e.g. a soccer match with its fixed subparts), and processes (e.g. the recipe for making a pancake with steps at different levels of abstraction).

Such concepts have not been the focus of research in the probabilistic modeling tradition. Here we describe an approach to representing structured concepts—more typical of the complexity of real world categories—using probabilistic generative processes. We test whether statistical inference with these generative processes can account for how people categorize novel instances of structured concepts and compare with more heuristic, exemplar-based approaches.

Because a structured concept like "house" has no single, simple perceptual prototype that is similar to all examples, learning such a concept might seem very difficult. However, each example of a structured concept itself has internal structure which makes it potentially very informative. Consider figure 1, where from only a few observations of a concept it is easy to see the underlying structural regularity that can be extended to new items. The regularities underlying structured concepts can often be expressed with instructions for *generating* the examples: "Draw a sequence of brown dots, choose a branch color, and for each brown dot draw two dots of this color branching from it."



Figure 1: Three examples of a structured concept described by a simple generative process.

We build on the work of Goodman, Tenenbaum, et al. (2008), who introduced an approach to concept learning as Bayesian inference over a grammatically structured hypothesis space—a "language of thought." Single concepts expressed in this language were simple propositional rules for classifying objects, but this approach naturally extends to richer representations, providing a concept learning theory for any representation language. Here we consider a language for generative processes based on *probabilistic programs*: instructions for constructing objects, which may include probabilistic choices, thus describing distributions on objects—in our case distributions on colored trees. Because this language describes generative processes as programs, it captures regularities as abstract as subparts and recursion.

The theory of concept representation that we describe here shares many aspects with previous approaches to concepts. Like prototype and mixture models (Anderson, 1990; Griffiths, Canini, & Sanborn, 2007), probabilistic programs describe distributions on observations. However, prototypes and mixtures generate observations as noisy copies of ideal prototypes for the concept and thus cannot capture more abstract structures such as recursion. Like rule-based models of concept learning, our approach supports compositionality: complex concepts are composed out of simple ones—but rather

than deterministic rules, our concepts denote distributions. Finally, the probabilistic program approach can be seen as a generalization of previous approaches to generative representations of concepts (Kemp, Bernstein, & Tenenbaum, 2005; Rehder & Kim, 2006; Feldman, 1997).

We investigate human learning for classes of generating processes that vary in their abstract structure, from simple prototypes to complex multiply recursive programs. We compare predictions for categorization judgments based on the true generative model to the predictions of exemplar models, which exploit the relational structure of the examples to varying degrees but cannot detect more abstract structure. We find two regimes: for concepts with simple prototype-like structures, human judgements are well described by a relational exemplar model, but humans can also easily learn more abstract regularities—such as sub-concepts and recursion—which are better captured by a model using more expressive generative descriptions based on probabilistic programs.

Formal Framework

In the following, we first explain the formal language we use to describe generative processes, then the different methods of categorization (or generalization) we compare to subjects' judgements.

Concept Representation

We analyze concepts as generative models, i.e. as formal descriptions of processes that generate observations. We do so within a simple domain where we can fully know and manipulate the actual generating processes behind complex objects. We use tree-structured graphs with colored nodes as observations in our experiments—these are a simple proxy for many real-world concepts, where the dependencies among parts are hierarchical or tree-like. Human bodies, buildings, and events all consist of parts that themselves contain parts, with each part standing in interesting relation to the others.

We represent these trees as nested lists: each list denotes a tree, with the first element in the list specifying the color of the root node and the remaining elements describing the children of this node, each child itself being a list (tree). For example, the second tree shown in figure 1 can be represented as `'(● (●) (● (●) (●)) (●))`.

We formalize the processes that generate these observations using a subset of Church, a Lisp-like stochastic programming language[1] (Goodman, Mansinghka, Roy, Bonawitz, & Tenenbaum, 2008). Programs in Church describe processes that produce values; running a program corresponds to generating a value from such a process. Because Church contains primitive functions that randomly choose from a distribution on values (e.g. the function `flip` that randomly chooses `true` or `false`), Church programs describe *stochastic* processes. The meaning of a Church program is a distribution on return values—which may be complex values such as nested lists—and any given execution results in a sample from this distribution. In what follows we describe Church programs which sample colored trees.

We group generative models into classes by the abstract constructions they use. Table 1 illustrates each of these types using a single concept program and observations drawn from this program. The simplest tree-generating processes in our language use only the stochastic function `node`, which takes as its first argument a color symbol and as its remaining arguments subtrees. With high probability, `node` returns a tree that has the given color symbol at its root and the given subtrees as its children, but with some probability ε, it switches to a noise process that can return any tree, that is, `node` introduces a random noise process into the tree construction. Under the noise process, the number of children for a node is sampled from a geometric distribution with parameter ε and the node color is sampled uniformly.

Programs like `(node ● (node ●) (node ●))` denote stochastic *prototypes*. They are most likely to generate the tree that corresponds to the given colors, in this case `'(● (●) (●))`, but they can return any tree with a certain probability. The more a tree deviates from the prototype, the less likely this process is to generate it. For example, the simple program described above could switch at the third `node` to the noise process and produce `'(● (●) (● (●)))` instead of the prototype. By introducing the noise process, `node` turns a deterministic prototype into a stochastic process.

All of the more abstract ways of formalizing generative models in our tree domain compose these basic processes. *Nested prototypes* formalize the intuition that a concept or a part of a concept can be "either this or that". Running the program `(if (flip .5) (node ●) (node ●))` will flip a fair coin and return a sample from `(node ●)` with probability .5, otherwise a sample from `(node ●)`.

One of the central reasons for analyzing concepts as represented in a language of thought is that they compose analogously to the components of natural and artificial languages—*parts* similarly allow composition through reuse in our domain. A part concept is defined first and can then be used in arbitrarily many places within other concepts. For example, the program `(define (part) (node ● (node ●)))` names a simple part consisting of only two nodes. This part can now be reused in other concepts. For example, the most likely return value for `(node ● (part) (part))` is `'(● (● (●)) (● (●)))`. When parts are defined, they are available to the noise process. This leads to some invariance to the position of parts and captures the idea that a generating process may give rise to observations that contain a part in a different place, although with lower probability compared to an observation with the part in the correct place.

Parameterized parts can capture both deterministic structure and random choices and reuse them in multiple places. When a part like `(define (part x) (node ● x x))` is used, for example in the program `(part (node ●))`, it evaluates

[1] Church uses prefix notation, i.e. function application is written with the operator first, the operands following. For example, `(node x y)` means that the function `node` is called with the arguments `x` and `y`.

the body of the part—here `(node • x x)`—with `x` assigned to its argument, here `(node •)`. Evaluating the program `(part (node •))` is therefore most likely to result in the observation `'(• (•) (•))`.

Allowing parts to call themselves introduces *recursion*, a means to capture a large amount of repetitive observed structure in a single short definition. For example, the part `(define (p) (if (flip) (node •) (node • (p))))` can generate arbitrarily deep lists of single blue nodes, with shorter ones being more likely.

The power of these program constructs is that they may be used compositionally to build more complex concepts, such as those shown in table 1.

Categorization

In order to model generalization and categorization behavior of human subjects, we need not only a way to represent concepts, but also a way to compute the probability of any given observation belonging to a known concept. We analyze our experimental results using four models that differ in how much they make use of representational structure.

On the unstructured end of the scale, we use a model that computes generalization judgements solely by comparing the fraction of nodes that have a given color. On the other end of the scale, a generative Bayesian model uses the likelihood under the true generative process to judge category membership. In between, an exemplar model makes use of tree structure in the observations, but not of the more abstract generative process that led to the observations.

Generative Model In modeling concept learning as Bayesian program induction, we follow the approach taken by Goodman, Tenenbaum, et al. (2008). Since we formalize concepts as probabilistic programs, the likelihood $P(O|C)$ of an observation O under a given concept C corresponds to the probability of the program making its random choices such that it returns the observation as its value (see Goodman, Mansinghka, et al. (2008)). The posterior probability of a concept C given observations O is proportional to this likelihood multiplied by the prior:

$$P(C|O) \propto P(O|C)P(C) \tag{1}$$

In the last section, we described a language for programs which generate trees; a prior $P(C)$ could be derived from this language, as in Goodman, Tenenbaum, et al. (2008). An ideal learner would then infer the posterior distribution $P(\mathbf{C}|O)$ over concepts $\mathbf{C}$ given the observation O and make predictions about whether a new observation t belongs to the category of the observed objects using each concept $C \in \mathbf{C}$ in proportion to its posterior probability:

$$P(t|O) \propto \sum_{C} P(t|C)P(C|O) \tag{2}$$

In order to make computational modeling tractable, we make the simplifying assumptions that (1) subjects' reasoning is dominated by the maximum a posteriori (MAP) estimate of this distribution, i.e. by the single concept that has the highest posterior probability and that (2) the true generating concept C_{true} is a good approximation to the MAP estimate. Thus, for each of the concept types we investigate, we model subjects' behavior using the program from which the training data was sampled. The likelihood of a new observation t belonging to this concept is simply $P(t|C_{true})$ which we compute using an adaptive importance sampling algorithm.

We do not claim that subjects necessarily identify the true generating concept from a few examples; this approximation is made for computational tractability. The full Bayesian model, which maintains uncertainty over generating concepts, can make different predictions in certain cases, but it is not clear whether this represents a bias for or against the approximation—to the extent that people remain uncertain of the concept after a few examples, the Bayesian model would capture human inferences better than our approximation.

Tree Exemplar Model This and the next two models are versions of the exemplar-based generalized context model (GCM) (Nosofsky, 1986). For observations $O_1, \ldots, O_n$ from category C and a new observation t for which we would like to estimate the likelihood under category C, we use $P(t \in C|O_1, \ldots, O_n \in C) \propto \frac{1}{n}\sum_{i=1}^{n} e^{-d(O_i,t)}$ where d is a distance measure that is sensitive to the tree structure of the observations. Starting from the root node, this measure matches the trees as much as possible, incrementing by 1 for each node that differs in color between the two trees and for each node that must be generated because it exists in one tree but not in the other tree. This approach is similar to the structure mapping approach used by Tomlinson and Love (2006).

Frequency-based Exemplar Models As in the tree exemplar model, we use a distance measure d to estimate the likelihood of an observation belonging to a category for which we have only positive examples. In this version of the model, $d(t_1, t_2)$ is the RMSE between the transition count vectors of t_1 and t_2. For each pair of node colors, the transition count vector contains the number of times this pair occurs adjacent (as parent-child) in the given tree. We call this model *Transition GCM*. We also investigate a simplified version that uses the distance between the color count vectors. The length of this vector corresponds to the number of possible node colors, with each entry in the vector denoting how often this node color appears in the tree of interest. We call this *Set GCM*.

Experiment

This experiment is an exploratory investigation into generalization from observations of structured objects. Since our main goal in this study is to investigate the representation of concepts and their use for categorization and generalization rather than the memory aspects of learning, we use a paradigm that minimizes memory demands. By doing so, we hope to focus on how people represent the commonalities between observed instances of a concept and how they use this knowledge to generalize to new instances. We chose a

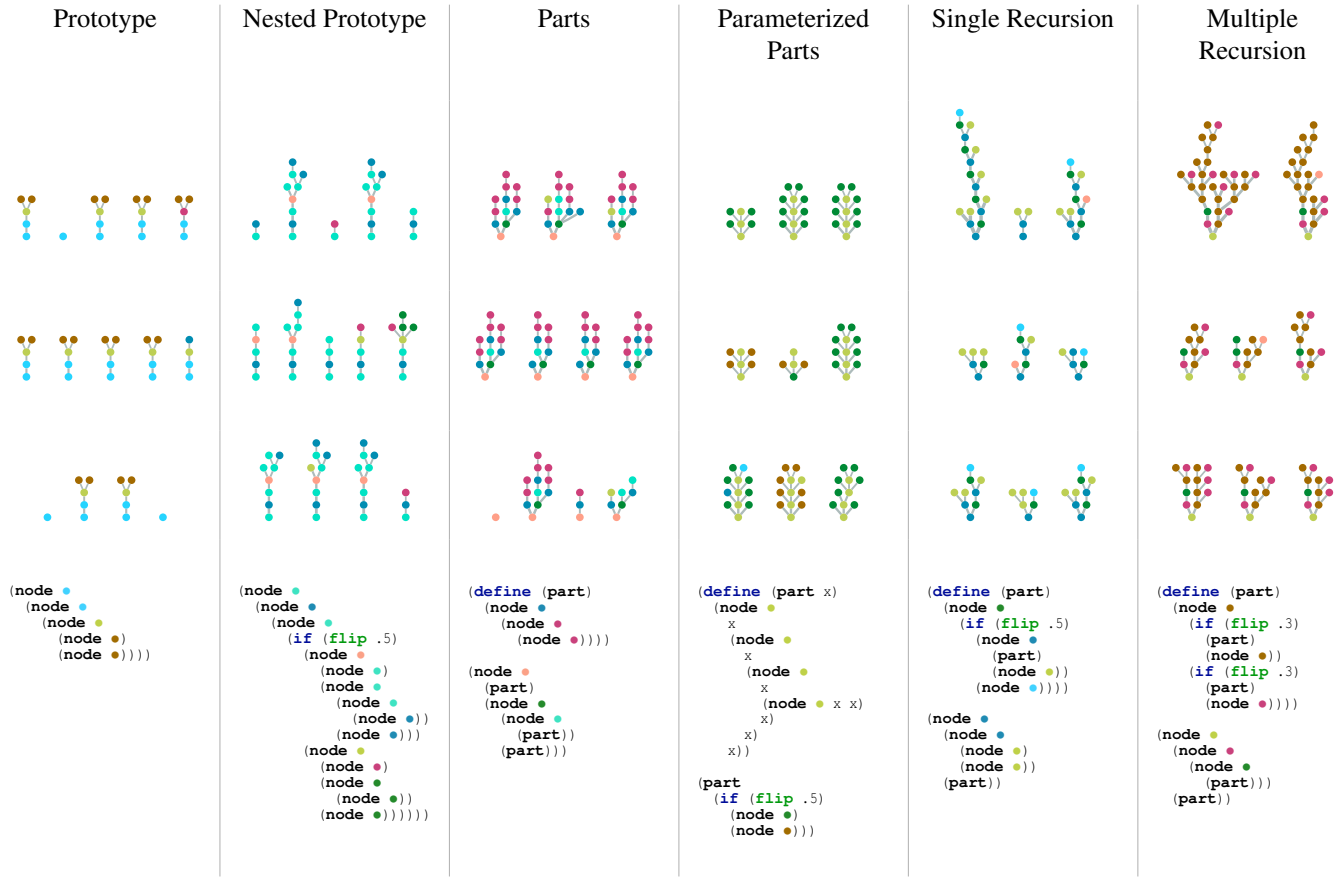


Table 1: This table illustrates the concept types that can be represented within our language for generative models. For each type, an example of a concept (a stochastic program) is shown together with observations drawn from this program. The stochastic function `node` generates a mixture of the subtrees that are passed to it as its arguments and a noise process that, with low probability, can generate any tree. The abstraction methods stochastic branching, (parameterized) parts and recursion compose these stochastic prototypes into more structured generative processes.

domain that both contains observations with simple structure and allows for interesting generative processes—the domain of colored trees generated by probabilistic programs.

Methods

Participants 250 members of Amazon's crowdsourcing service *Mechanical Turk* took part in the online experiment. Subjects were compensated for participation.

Stimuli Subjects were told that they are looking at newly discovered kinds of plants that grow in extreme environments. Each subject saw 18 pages, with each page consisting of 15 training examples, a control question, and a test example together with a classification question. Both training and test examples were images of simple trees with colored nodes drawn from tree-generating programs (see e.g. table 3). For each of the concept types shown in table 1, there were three tree-generating programs, and for each program there were 7 test examples. These test examples were chosen to cover a wide range of both intuitive and model judgements of category membership. Both training example order and stimuli colors were randomized.

Procedure In order to ensure that subjects process the training stimuli, a control question on each page asked how many of the training trees consist of more than 7 dots. 55 subjects answered less than 13 out of the 18 control questions correctly within an error margin of 2. We did not include these subjects in the analysis.

The categorization question asked: "How likely is it that the following plant is the same kind of plant as the plants above?" Subjects chose on a seven-step scale ranging from "certainly the same kind" to "certainly not the same kind". For each subject, the responses were normalized to $[0, 1]$.

Results

Table 2 summarizes the correlation results for all models. Figure 2 shows for each concept type human results and model results for both the exemplar and generative model. For each concept type, three different concepts were part of the experiment, and for each concept, seven different test observations were shown. A single point in the scatterplot contains information on the mean subject response for a single test tree and on the model prediction for this tree.

Neither of the two exemplar models based on simple statis-

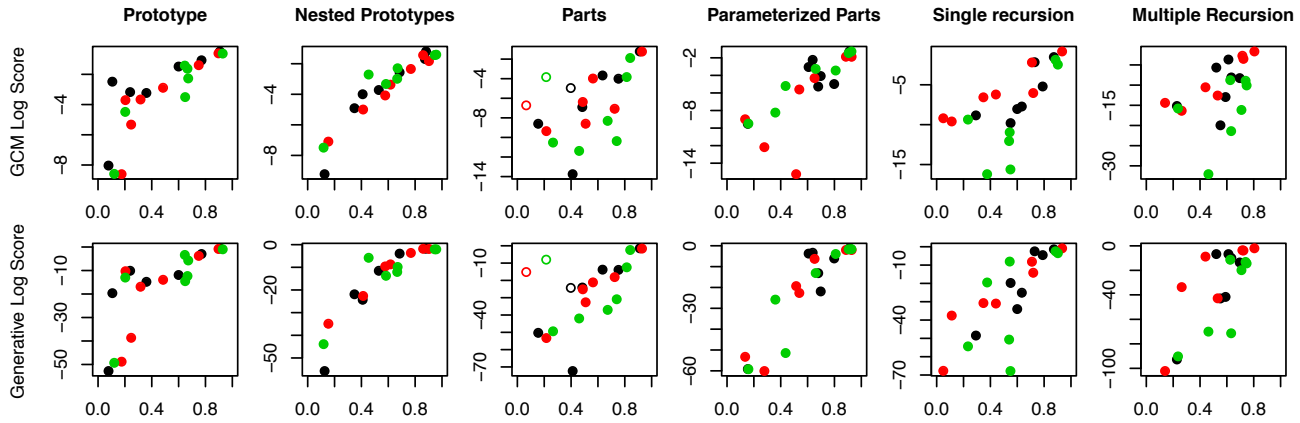


Figure 2: Comparison between human and model responses across concept types for tree exemplar and generative model. For each of the six concept types, three examples were shown; the color of the dots indicates to which example any given datapoint belongs. Empty circles denote isolated part cases that were excluded from the correlation analysis.

	Set GCM	Transition GCM	Tree GCM	Generative Model
Prototype	0.589	0.751	**0.803**	0.748
Nested Prototype	0.544	0.851	**0.937**	0.904
Parts*	0.320	0.617	0.705	**0.835**
Parameterized Parts	0.298	0.591	0.778	**0.911**
Single Recursion	0.284	0.499	0.637	**0.773**
Multiple Recursion	0.505	0.561	0.451	**0.770**

Table 2: Human-model correlations for the experiment. Each row shows how well the different models predicted subjects' performance for a particular concept type. *Correlations excluding isolated part cases (see text).

tics was the best predictor for any of the concept types, with the transition-based exemplar model performing strictly better than the set-based model. An effect that is not accounted for by the less structural exemplar models is illustrated by the nested prototype example in table 3: Subjects generalize significantly more to examples with branches they have seen before than to examples that have a mixture of two known branches. Likewise, subjects seem to generalize significantly more to trees with known branches than to trees that have new branches with similar surface statistics. Both results are expected under the two models that make use of tree structure.

If we group prototype and nested prototype as "less structured" and subconcepts with and without arguments, single recursions, and multiple recursions as "more structured", then the tree exemplar model best predicts human responses for the less structured stimuli whereas the true generative model best predicts performance for the more structured stimuli.

Our generative model makes the simplifying assumption that the learner infers a single generating concept from the examples whereas one interpretation of the tree exemplar model is that it uses each of the training examples as a hypothesis on what the true concept looks like. A fully Bayesian learner, which maintains a distribution over generative processes, may predict human behavior in ways similar to the tree exemplar model for less structured examples and similar to the true generating process model for the more structured examples.

Having seen how different models predict human judgements for different concept types, we will now look at individual response patterns in order to determine ways in which both of the two structural models can be improved.

The part example in table 3 shows how changes to the location of a part can have significantly different effects depending on whether the overall concept is preserved (resulting in high generalization) or the part is moved into a completely different environment (resulting in low generalization). By analogy, a Picasso face, with eyes in odd places, is still more of a face than an eye alone. Parts seen out of context constitute a problem for all models (except for the simplest set-based one): subjects judged these isolated parts as unlikely to come from the concept that included them as subparts whereas the models did give a high score to these examples. Since including these outliers dramatically changed the scores and made the interpretation of the model comparison difficult, we excluded these data points from the analysis in table 2. Without correction, the model-human correlations for the part concepts are: 0.403 for the set-based exemplar model, 0.505 for the transition exemplar model, 0.512 for the tree-based exemplar model, and 0.543 for the generative model (note that rank-order among the models does not change as a result of excluding these data points).

For the parameterized part example in table 3, changing the argument uniformly, i.e. in all places where it occurs, leads to consistently higher scores than changing the argument differently in different places; however, this difference is not significant. This difference is expected if subjects inferred the true generative model, since changes to the argument require only one use of the noise process, whereas nonuniform changes require many different nodes to be generated by the noise process. Future research needs to determine whether this effect is real, perhaps by manipulating the diversity of

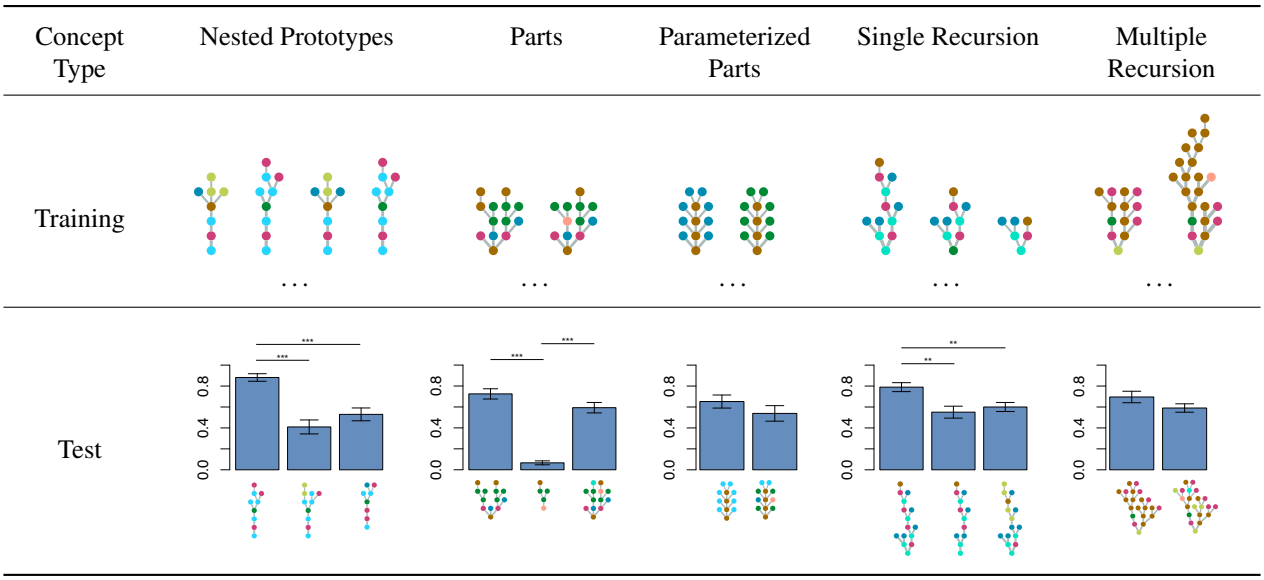


Table 3: This table illustrates a small selection of our experimental results. For five different concept types, training observations from a single concept of this type are shown together with subjects' generalizations for particularly interesting test examples. The error bars are standard errors of the mean.

parameter arguments in the observations.

For the single recursion example in table 3, changing the color of a few nodes within the recursion results in a significantly lower generalization. At the same time, a very similar manipulation does not result in a significant change in the generalization rating for the multiple recursion example. Intuitively, we sometimes see a change as destroying a very obvious pattern structure whereas at other times, the change in structure is not assumed to be relevant. Future research needs to characterize when subjects infer that such a pattern exists, and when they instead assume coincidence.

The comparison between the frequency based exemplar models and the two models that rely on tree structure in the observations makes clear that subjects do make use of the fact that the observations are structured in their generalization judgements. Furthermore, comparing the tree exemplar model to the true generative model that makes use of more abstract structure hints at the possibility that subjects are relying on recursive structure in the observations. The individual response patterns in the results of our exploratory experiment highlight ways in which both the exemplar-based model and the generative model can be improved to more closely reflect human generalization patterns.

Conclusion

Most studies of concept learning have focused on relatively unstructured objects based on simple features. We have suggested viewing concepts as probabilistic programs that describe stochastic generative processes for more structured objects. In this view concepts denote distributions over objects, and these distributions are built compositionally. We explored this idea within a domain of tree-like objects, and carried out a study of human generalization using a broad variety of concepts in this domain. Our results suggest that humans are able to extract abstract regularities, such as recursive structure, from examples, but also that there are many subtle effects to be discovered and accounted for in such domains.

Acknowledgements We thank Frank Jäkel and Brenden Lake for useful comments. This work was funded in part by grants from the ONR (N00014-09-0124) and the AFOSR (FA9550-07-1-0075).

References

Anderson, J. (1990). The adaptive character of thought.

Feldman, J. (1997). The structure of perceptual categories. *Journal of Mathematical Psychology*.

Goodman, N. D., Mansinghka, V., Roy, D. M., Bonawitz, K., & Tenenbaum, J. B. (2008). Church: a language for generative models. *Proceedings of the 24th Conference in Uncertainty in Artificial Intelligence*, 220–229.

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, *32*(1), 108–154.

Griffiths, T., Canini, K., & Sanborn, A. (2007). Unifying rational models of categorization via the hierarchical dirichlet process. *Proceedings of the 29th Annual Conference of the Cognitive Science Society*.

Kemp, C., Bernstein, A., & Tenenbaum, J. (2005). A generative theory of similarity. *Proceedings of the Twenty-Seventh Annual Conference of the Cognitive Science Society*.

Markman, A. (1999). Knowledge representation.

Nosofsky, R. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–57.

Rehder, B., & Kim, S. (2006). How causal knowledge affects classification: A generative theory of categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Shi, L., Feldman, N., & Griffiths, T. L. (2008). Performing bayesian inference with exemplar models. *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, 745–750.

Tomlinson, M., & Love, B. (2006). From pigeons to humans: Grounding relational learning in concrete examples. *Proceedings of the National Conference on Artificial Intelligence*, *21*(1), 199.

Short-Term Word Priming Across Eye Movements

Stephen E. Denton (sedenton@umail.iu.edu) and Richard M. Shiffrin (shiffrin@indiana.edu)
Department of Psychological and Brain Sciences, and Cognitive Science Program, Indiana University
1101 E. 10th Street, Bloomington, IN 47405 USA

Abstract

The authors conducted a short-term repetition priming experiment (using a visual, forced-choice word identification task) that compared a standard priming condition, where prime and target words appeared in the same spatial location, with an experimental condition in which prime and target words were spatially separated enough to necessitate an eye movement. Prime presentation duration was manipulated and, within both eye movement conditions, it was found that short primes produced a preference to choose a primed alternative, whereas for longer duration primes this preference was absent. Based on the similarity between eye movement conditions, it is argued that prime and target features from separate fixations are still confusable and that evidence regarding prime feature must still be discounted. A computation model that includes these offsetting components of source confusion and discounting provides an excellent account of our results.

Keywords: short-term priming; immediate priming; repetition priming; perceptual identification

The term *priming* refers to a well-known information processing effect wherein one stimulus (a *prime*) influences a similar or related stimulus (a *target*) presented at a later time. The influence the prime has on the target is usually one of facilitation. In the typical priming task trial, the prime is presented first followed by a target that is briefly flashed and masked. This paradigm is referred to as *short-term* or *immediate* priming because primes are presented immediately prior to the targets, with inter-stimulus intervals generally less than a second. The first stimuli is called a prime because it is thought to "prime the pump" for a related target, yielding faster and more accurate responding. However, priming does not simply result in facilitation. Previous research has indicated that there is an intricate set priming effects that occur across various situations and experimental conditions (e.g., Huber, Shiffrin, Lyle, & Ruys, 2001; Weidemann, Huber, & Shiffrin, 2005). Some experiments find facilitation, while other fail to find facilitation or even find reliable deficits due to priming.

Huber et al. (2001) began to systematically examine the effects of priming using a two-alternative forced-choice (2-AFC) testing procedure with the goal of separating the perceptual and decisional components involved in priming. Observers were asked to correctly identify a previously flashed target word when given a choice between it and an incorrect foil word. This study indicated that priming largely arises from preferences to choose whatever has been primed. For example, in repetition priming (priming in which the prime can be the same word as either the target or foil), it was found was that with short prime presentation durations, priming with the target increases accuracy and priming with the foil decreases accuracy, when both are compared to a control condition in which the prime is unrelated to either the target or the foil. The corresponding prime conditions are termed *target-primed*, *foil-primed*, and *neither-primed*, respectively.

The preference effects revealed by Huber et al. (2001) proved to be readily manipulatable, changing in magnitude and direction as a function of prime saliency (e.g., Huber, Shiffrin, Quach, & Lyle, 2002; Weidemann et al., 2005; Weidemann, Huber, & Shiffrin, 2008). Thus, when the prime is made more salient, either through a longer presentation duration or though an active task that required responding to the prime, the conventional priming effect diminishes or in some cases reverses. Prime saliency manipulations result in reduced facilitation (or slight deficits) in target-primed conditions, while leading to increased accuracy in foil-primed conditions.

The ROUSE Model

To account for a range of findings within the 2-AFC identification paradigm, Huber et al. (2001) developed a feature-based Bayesian model of short-term priming. The *responding optimally to unknown sources of evidence* (ROUSE) model accounted for experimental priming data by incorporating the two offsetting components of feature source confusion and discounting. The source confusion portion of the model posits that features can be carried over from the prime to the target percept, without source information. Thus, when a choice word is presented, feature activations could be due to the prime presentation, the target presentation, and/or noise without the source of the activation being known to the system. This factor alone can produce the standard priming effect, as it causes a preference toward prime-related choice words. The discounting mechanism in the model can counteract this preference, because this component posits that perceived features are assigned evidence and feature evidence is discounted when known to have come from the prime. A Bayesian decision process then arrives at an optimal response given it is operating on this noisy and discounted evidence. The implication here is that making the prime more salient (e.g., long presentation duration) leads to increased discounting of prime feature evidence. Thus, discounting mechanisms can explain a lack of, or a reversal in, the typical priming preference.

In the ROUSE model, choice words are represented as a feature vector, typically consisting of 20 binary features (Huber et al., 2001). Features can be independently activated by the prime, with a probability α, by the target, with a probability β, or be activated due to noise, with a probability γ. The system is assumed to only have access to what features are active and not the source of their activation (i.e., source confusion), therefore the probabilistic effect of α, if

not counteracted, will result in a preference toward primed choice words. This preference is counteracted by a decision process that assigns lower levels of evidence to features that might have been activated by the prime word (i.e., discounting). To discount features optimally, the system needs to know the probabilities that features are active due to particular sources (i.e., α, β, and γ). However, it is assumed that the decisional system does not have access to the exact probabilities and therefore uses estimates, α', β', and γ'. These estimates, which reflect the amount of discounting applied to particular sources, are used to evaluate evidence. These estimates are theorized to be close to their true values, but the model's behavior depends critically on the magnitude and direction of the difference between α and its estimate. Under-discounting ($\alpha' < \alpha$) results in a preference for primed choice words and over-discounting ($\alpha' > \alpha$) produces a preference against primed choice words.

Given the estimated source probabilities, the optimal response among choice words can be computed by combining feature evidence. Feature evidence takes the form of a likelihood ratio that specifies the probability that a feature is from the target word over the probability the feature is part of the foil, given the feature's current activation state and whether or not the feature appeared in the prime. Assuming feature independence, these likelihood ratios can be multiplied together across all word features to produce an overall likelihood that the choice word is the target. The likelihoods of the two words can then be compared with the larger being identified as the target. In the case of ties, a random selection is made between the choice words.

The above description of the ROUSE model is only intended as a brief summary, which means a number of details, large and small, have been left out. For a more comprehensive presentation of the ROUSE model, the reader is referred to Huber et al. (2001) for the original stochastic version and Huber (2006) for a later analytic version.

Feature carry-over across eye movements

The ROUSE model has been useful in accounting for and predicting priming data, but there is still much uncertainty about how its key components, source confusion and discounting, actually perform their theorized functions. Many questions remain unanswered regarding what causes or allows both feature confusions and discounting. Source confusion could be the result of a number of candidate causes, e.g., spatial or temporal proximity between prime and target, or similarity between prime and target on any number of dimensions. Similarity relations (semantic, orthographic, etc.) between target and prime have received considerable attention elsewhere and will be set aside for present purposes. Temporal proximity does appear to be important as increasing the inter-stimulus intervals (ISIs) between prime and target to be greater than 250 ms has been shown to disrupt and diminish priming (Hochhaus & Marohn, 1991). In the following experiment, we chose ISIs to be less than 250 ms so that we can effectively ignore the temporal dynamics of priming for the purposes of the present study.

The focus of this research is how spatial proximity effects priming. We are interested in knowing if feature confusions, discounting, or both are location dependent. In the present study, the location at which the prime and target were presented were sometimes spatially separated at a 10° visual angle. Visual acuity drops to roughly 25% (of central vision) at a visual eccentricity of 10° (Low, 1951). Given diminished visual acuity with this spatial separation, we are effectively enforcing an eye movement between viewing the prime word and the target word, if both are to be seen. Not only will the target appear in a different location than the prime, but it will also appear in a different eye fixation.

The goal of the present experiment is to see if features are carried over from a prime fixation to a target fixation when the fixation locations are relatively distant spatially. We would like to see if features from the prime join with the target percept at the new location (i.e., will there still be source confusion when an eye movement is made). Discounting could also be affected by our eye movement manipulation. If there is some source confusion across eye movements, it is conceivable that prime features could be discounted completely (i.e., little evidence from the prime would be used in making the target identification decision). Then again, given the presence of some source confusion, it is also possible that better estimates of feature activations due the prime can be made when the prime and target appear at separate and distinct locations. This would reduce or eliminate under-discounting and over-discounting of prime evidence. The ROUSE model will be used to provide an indication of the relative contribution of source confusion and discounting across eye movement conditions.

Priming Across Eye Movements Experiment

We designed the current experiment to investigate how spatial separation between prime and target affects priming. The experiment compared an eye movement condition to an appropriately matched control condition in which all words were presented in the center of the screen (i.e., the stereotypical short-term priming paradigm). For all trials, the prime word was presented in the center of the screen. As with our previous studies, prime salience was manipulated by adjusting the prime presentation duration, which was either short (50 ms) or long (1000 ms). On any particular trial, the prime word itself could be either the same as the target, the foil, or neither, corresponding to the prime conditions discussed earlier. The target word was flashed in the center of the display on half to the trials and on the other half it appeared equally often in one of four locations, directly above, below, left or right of the center. The participant's task was to identify the flashed target given a 2-AFC test at the end of each trial.

The experiment was a within-subject design that crossed the 2 levels of prime presentation duration (short and long) with 3 levels of prime condition (target, foil, or neither primed) and crossed both with 5 different target locations split

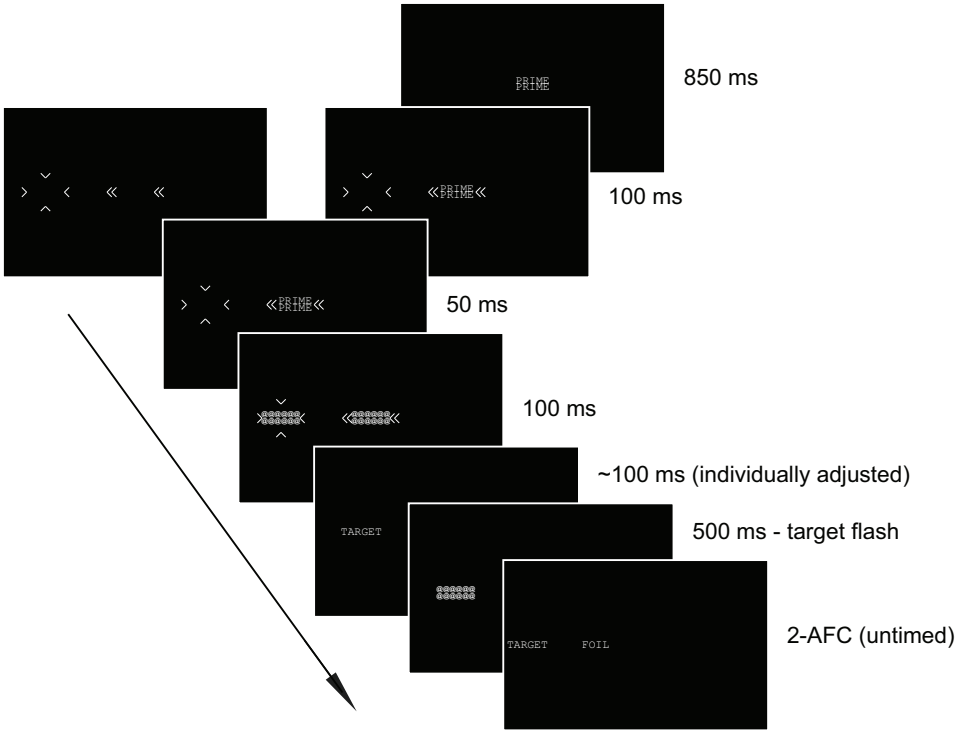


Figure 1: An example sequence of events in the present experiment. This figure shows both short and long prime duration conditions for one particular peripheral target location (left of center). The duration for each frame is presented on the right. The left sequence corresponds to a short prime duration condition, whereas the right sequence shows a long prime duration sequence where the total time the prime is presented is 1000 ms (850 ms+100 ms+50 ms). Both sequences were preceded by a 500 ms presentation of a fixation cross in the center of the screen (not shown). The positions of the target and foil word on the last frame were randomized.

into 2 general categories (center and peripheral [top, bottom, left, and right]). A schematic of two example trials is presented in Figure 1. A more detailed description of our exact experimental procedure is below.

Methods

Participants Fifty one undergraduate students volunteered to participate for partial credit in an introductory psychology course at Indiana University.

Materials and Apparatus We used two pools of 1,100 five-letter and 1,300 six-letter words with a written-language frequency of at least 4 per millions as defined by Kučera and Francis (1967). All words were presented in uppercase using the fixed-width "Courier Bold" 17-point font. Throughout the experiment, stimulus words were sampled randomly without replacement with the only constraint being that 5 and 6 letter word never appeared together in the same block of trials. All masking was done with two rows of six "@" signs presented in "Arial Narrow Bold" 13-point font. This ensured dense coverage of the prime and target. The stimulus words were presented in white against a black background.

All stimuli were displayed on 16-in. (40.6 cm) PC CRT monitors with vertical refresh rates of 120 Hz and a screen resolution of 800 $\times$ 600 pixels. The experiment was programmed using the Vision Egg library for the python programming language (Straw, 2008). The display was synchronized to the vertical refresh of the monitor providing display increments of 8.33 ms.

Participants sat individually in a dimly lit, ventilated, sound-dampened booth. Participants were asked to sit up straight to keep the distance from their eyes to the monitor at approximately 50 cm., but no head restraint was used to enforce this viewing distance. This viewing distance ensured that peripheral targets would appear at least a 10° visual angle away from the center of the screen. Participant responses were collected using a standard computer keyboard. In a 2-AFC test, participants were asked to press the "F" key to choose the left alternative or the "J" key to choose the right.

Procedure The procedure used in the present experiment was carefully designed as to compare an experimental condition involving eye movements to an appropriate control that maintained important aspects of the experimental condition (namely timing and perceptual masking), while not requiring a eye movement. The experiment was designed based on some knowledge of the timing of eye movements or saccades. A 10° saccade would take less than 50 ms to complete

once initiated, but takes more than 150 ms to plan and initiate (Irwin, Brown, & Sun, 1988). As such, having the target flash appear in the periphery immediately after the prime would make it impossible to see. To remedy this, we preceded the target with a cue that indicated the correct location where the target would appear. This target indicator cue appeared 250 ms before the target flash. We left a 100 ms interval between the prime offset and the target onset. During this 100 ms interval the prime was post-masked while the target was pre-masked. This timing is reflected in the example sequences presented in Figure 1.

Each trial began with a fixation cross appearing in the center of the screen for 500 ms. In long prime duration conditions, the prime is presented (in duplicate) for 850 ms then the prime and target indicator appear together for 150 ms. In trials with short prime durations, the target indicator appears first by itself for 100 ms, then the prime and the target indicator appear together for 50 ms. Participants should not be able to plan and initiate a saccade in under 100 ms, therefore the prime will be viewed; the prime is 'snuck-in' before the eyes have a chance to move. After the prime and the 100 ms intervening mask, the target is flashed. The target is post-masked and then the 2-AFC options are presented to the immediate left and right of the location where the target appeared. A peripheral target word can appear at the top, bottom, left, or right of the screen. Center target location trials have the same timing as peripheral target trials. Besides the actual target location, the only differences between center and peripheral trials are that the target indicator appears in the center of the screen (surrounding the prime) and the prime post-mask and the target pre-mask are one in the same. On each trial, once a 2-AFC selection is made, feedback is provided.

Each participant went though 672 priming trials broken into 7 blocks of 96. The first 32 trials were neither-primed practice trials with long target durations (150 ms) to get participants used to the task. These practice trials were followed by 64 neither-primed calibration trials. Target word durations were individually adjusted for each subject such that accuracy was roughly 75% on neither primed conditions. This calibration was done separately for the center and peripheral target locations using a staircase method. As with previous studies (e.g., Huber et al., 2001; Weidemann et al., 2005), there were large individual differences. For center target trials, target flash times ranged from 25 ms to 91.7 ms with a median of 50 ms. For peripheral targets, flash times ranged from 33.3 ms to 200 ms (the maximum allowed) with a median of 91.7 ms. The increased variance in target flash times for peripheral locations is likely due to individual differences in saccade latency on top of individual differences in target processing time, i.e., some participants may have fixated on the target later and took longer to process the target.

Results

Data from all peripheral target locations were combined for the purposes of analysis, creating two target location conditions, central and peripheral, that initially had equal sample sizes. Reaction times were collected and used to eliminate deviant trials in the data. Trials in which the participant responded in less than 100 ms or took more than 3-s to respond were eliminated. Approximately 1% of the data was thrown out by this criterion. The first block of trials, which included practice and calibration trials, were not included in the analysis. Remaining experimental data were analyzed with a $3 \times 2 \times 2$ (Priming Condition $\times$ Prime Duration $\times$ Target Location) repeated measures analysis of variance (ANOVA).

There were large main effects of prime condition, $F(2,100) = 167.6, p < .001$, and prime duration, $F(1,50) = 135.5, p < .001$. Also, these two variables interacted, $F(2,100) = 121.37, p < .001$. The main effect of target location was not significant, however this variable had a significant interaction with both prime condition, $F(2,100) = 20.43, p < .001$, and prime duration, $F(1,50) = 8.637, p < .005$. Finally, there was a significant 3-way (Priming Condition $\times$ Prime Duration $\times$ Target Location) interaction, $F(2,100) = 48.57, p < .001$.

Average accuracy across all conditions is shown graphically in Figure 2. The dots in the figure give ROUSE model predictions which will be discussed in the next section. As can be seen from Figure 2, the typical priming effect was found when the target and a short prime were both presented in the center of the screen. When the prime was presented for a longer duration this prime preference disappeared. When the target was presented in the periphery after a centrally presented prime, the trend remained largely the same, although the magnitude of the priming effect decreased.

Applying the ROUSE model

Fitting the ROUSE model to the current experiment involved estimating the following eight parameters:

1. the probability that a choice word feature is activated by the prime (α),
2. the estimated probability that a feature is activated by a short prime when the prime and target are located in the center ($\alpha'_{S.C}$),
3. the estimated probability that a feature is activated by a long prime located in the center ($\alpha'_{L.C}$),
4. the estimated probability that a feature is activated by a short prime from a previous fixation ($\alpha'_{S.P}$),
5. the estimated probability that a feature is activated by a long prime from a previous fixation ($\alpha'_{L.P}$),
6. the probability that a feature is activated due to the target flash (β),
7. the probability that a feature is activated due to noise given a short prime presentation (γ_S), and
8. the probability that a feature is activated due to noise given a long prime duration (γ_L)[1]

[1]Two separate noise parameters were included to account for trend in the data that performance is better on neither-primed trials when prime duration is long compared to when the prime duration is short. The γ values from the to-be-discussed model fit indicate there is higher noise in the short prime conditions when compared to long prime conditions. This is somewhat sensible given the erratic na-

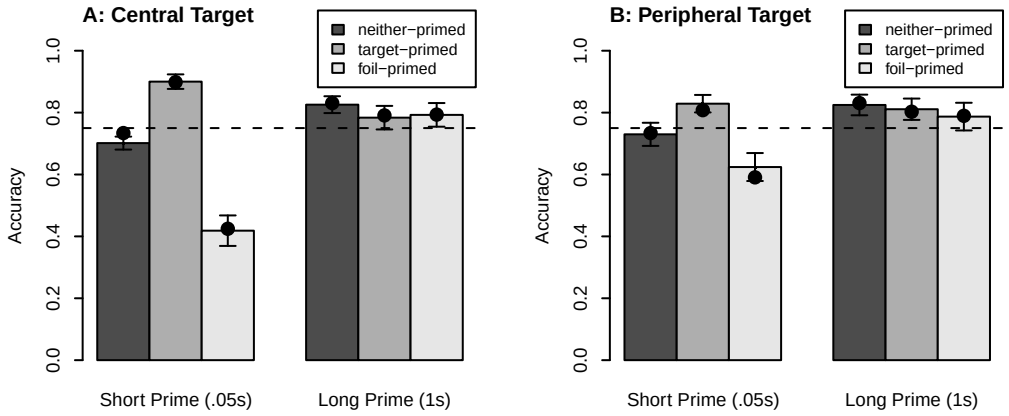


Figure 2: Forced choice performance and corresponding ROUSE predictions (represented by the dots). The bar heights show the mean proportions of correct target identification choices (error bars show $\pm$95% confidence intervals) within each condition. Panel A shows accuracy for centrally located targets and Panel B shows accuracy when the target was presented in a peripheral location. Each panel is further subdivided by prime duration and priming condition. The dashed horizontal line indicates 75% performance; the accuracy participants should roughly achieve on neither-primed trials due to the target duration calibration procedure.

As in previous work (e.g., Huber et al., 2001; Weidemann et al., 2005), it is assumed that estimates of feature activations due to targets and noise are equal to the actual values (i.e., $\beta' \equiv \beta$ and $\gamma' \equiv \gamma$).

The parameters listed above were estimated to generate the ROUSE model fit that appears in Figure 2. The parameter estimates used were: $\alpha = .12$, $\alpha'_{S.C} = .032$, $\alpha'_{L.C} = .65$, $\alpha'_{S.P} = .10$, $\alpha'_{L.P} = .20$, $\beta = .064$, $\gamma_S = .067$, and $\gamma_L = 0.011$. It is important to stress that exact values of these parameters are relatively immaterial, especially the exact values of the α and α' parameters, as it is their relative magnitudes that dictate the behavior of the model. Multiple model fits were done with both more and less parameters free to vary, but this fit provided the most satisfying account to the data. The large number of parameters (8 parameters for 12 data points) and possible over-fitting should be somewhat of a concern here, but since the model is largely descriptive, these complexity concerns are left unaddressed in the present context.

As can be seen in Figure 2, the ROUSE model provides a very good qualitative and quantitative description of the priming data we collected. Interpretation of the fit parameter values can be a tricky, because attending to the exact magnitude can be deceptive. For example, the same α value was used to fit all conditions. When the 2 target location conditions were allowed to have separate freely estimated α parameters it produced virtually no fit improvement. This could be taken as evidence that source confusion was the same at peripheral locations as at the center. However, α' was fit separately for each condition and changes in these values can compensate for possible underlying differences in source confusion (α).

Given the relationship between the fit α' values, it does appear that the magnitude of discounting decreased when the target was moved to a novel location. The estimates of primed feature intrusion contracted closer to optimal in the peripheral conditions (i.e., $\alpha'_{S.C} < \alpha'_{S.\underline{P}} < \alpha$ (optimal) $< \alpha'_{L.\underline{P}} < \alpha'_{L.C}$), which means there was less under-discounting for short primes and less over-discounting for long primes when the prime and the target did not appear in the same spatial location.

Discussion

In the present research, we investigated whether spatial proximity of prime and target was necessary to find priming effects, particularly when there was a large spatial distance between the two requiring an eye movement. Our findings indicated that even after a relatively large eye movement (a 10° visual angle), participants showed similar priming preferences as when they viewed all stimuli in one location. More specifically, in both eye movement and control conditions, participants showed a preference to choose the prime word when the prime duration was short. Further, this prime preference was undetectable when the prime was made more salient by increasing its presentation duration.

The ROUSE model provided an excellent account of our experimental data. The model's success, the resulting best fitting parameters, and data themselves all provide evidence that across eye movements (a) source confusion is still present,

ture of short duration priming trial (i.e., the target indicator, prime, mask, and target all appear in rapid succession) which could induce noise in the system. This seems especially true in neither-primed trials where all prime features presented are effectively noise. Increasing noise (γ) in the model decreases accuracy and uniformly translates all performance predictions down. Separate β parameters could be used similarly to selectively increase and decrease performance across short and long prime duration conditions, but saying that there is an increased probability of target features being activated for long prime durations provides a much less sensible explanation of the results.

and (b) discounting of evidence regarding prime features persists. The evidence that source confusion endures after an eye movement to a novel location implies that prime features appearing in one eye fixation are carried over to the next and join with the target precept at the new location. Further, source confusion from two spatially distinct locations does not appear to be substantially different from what it is when all feature sources are in the same spatial location.

The presence of discounting across eye movements implies that the decisional system is, in some sense, aware that features are not tied to particular locations and therefore must continue to estimate the likelihood that feature activations in choice words are the result of the prime in an effort to produce optimal responses. Even when the prime was in a previous eye fixation, the estimation process evidently succumbs to the same biases that are present without an eye movement, i.e., short, difficult to detect primes are under-discounted and long, salient primes are over-discounted. Since evidence for the prime is only discounted to the extent that features are known to exist in the prime, it is conceivable that the distinct spatial location of the prime provides some additional information to the decisional system that helps it better estimate the actual probability that a feature activation is due to the prime. Consequently, spatially separating the prime and target would make discounting slightly more optimal. Our modeling results hint that this is the case as the model estimates of feature intrusions from the prime better match the actual intrusion probabilities in the peripheral condition.

Our study suggests that the offsetting mechanisms of source confusion and discounting operate on features that are in general not tied to specific locations. Eye saccades are thought to effectively erase iconic memory, which is a visual store with high capacity but with a very limited duration (Irwin, 1992). There are a few visual items (3 to 4) that are retained from one eye fixation to the next in what is termed trans-saccadic memory, which has a limited capacity compared to iconic memory but has a longer duration (at least 750 ms). Nevertheless, an enforced eye movement will eliminate most low-level visual features from the previous fixation. After a saccade, the features that remain are presumably more high-level and location-independent. If source confusion and discounting operate with such features then an eye movement would not eliminate priming effects. This conforms to the findings of the present study. It is possible that source confusion and discounting operate on low-level features as well, and this may provide some explanation as to why eye movements slightly degrade the effectiveness of priming. The extent to which priming involves low-level features, and if it is indeed their suppression during saccades that produce the minor differences across eye movement conditions, are topics for future research.

Acknowledgments

This research was supported by NSF Grant 6804643 awarded to the second author.

References

Hochhaus, L., & Marohn, K. M. (1991). Repetition blindness depends on perceptual capture and token individuation failure. *Journal of Experimental Psychology: Human Perception and Performance, 17*(2), 422–432.

Huber, D. E. (2006). Computer simulations of the ROUSE model: An analytic simulation technique and comparison between the error variance–covariance and bootstrap methods for estimating parameter confidence. *Behavior Research Methods, 38*(4), 557–568.

Huber, D. E. (2008). Immediate priming and cognitive aftereffects. *Journal of Experimental Psychology: General, 137*(2), 324–347.

Huber, D. E., Shiffrin, R. M., Lyle, K. B., & Quach, R. (2002). Mechanisms of source confusion and discounting in short-term priming 2: Effects of prime similarity and target duration. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*(6), 1120–1136.

Huber, D. E., Shiffrin, R. M., Lyle, K. B., & Ruys, K. I. (2001). Perception and preference in short-term word priming. *Psychological Review, 108*(1), 149–182.

Huber, D. E., Shiffrin, R. M., Quach, R., & Lyle, K. B. (2002). Mechanisms of source confusion and discounting in short-term priming: 1. effects of prime duration and prime recognition. *Memory & Cognition, 30*(5), 745–757.

Irwin, D. E. (1992). Memory for position and identity across eye movements. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*(2), 307–317.

Irwin, D. E., Brown, J. S., & Sun, J.-S. (1988). Visual masking and visual integration across saccadic eye movements. *Journal of Experimental Psychology: General, 117*(3), 276–287.

Kučera, H., & Francis, W. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

Low, F. N. (1951). Peripheral visual acuity. *Archives of Ophthalmology, 45*, 577–593.

Straw, A. D. (2008). Vision egg: An open-source library for realtime visual stimulus generation. *Frontiers in Neuroinformatics, 2*(4), 1–10. (doi: 10.3389/neuro.11.004.2008)

Weidemann, C. T., Huber, D. E., & Shiffrin, R. M. (2005). Confusion and compensation in visual perception: Effects of spatiotemporal proximity and selective attention. *Journal of Experimental Psychology: Human, Perception, and Performance, 31*(1), 40–61.

Weidemann, C. T., Huber, D. E., & Shiffrin, R. M. (2008). Prime diagnosticity in short-term repetition priming: Is primed evidence discounted, even when it reliably indicates the correct answer? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*(2), 257–281.

Toward a Large-Scale Characterization of the Learning Chain Reaction

Alexei V. Samsonovich (asamsono@gmu.edu)
Krasnow Institute for Advanced Study, George Mason University, 4400 University Drive MS 2A1
Fairfax, VA 22030-4444, USA

Abstract

Designing an agent that can grow cognitively from a child to an adult human level of intelligence is the key challenge on the roadmap to human-level artificial intelligence. To solve this challenge, it is important to understand general characteristics of the expected learning process at a level of mathematical models. The present work makes a step toward this goal with a simple abstract model of a long-term learning process. Results indicate that this process of learning is characterized by two distinct regimes: (1) limited learning and (2) global learning chain reaction. The transition is determined by the set of initially available learning skills and techniques. Therefore, the notion of a 'critical mass' for a human-level learner makes sense and can be determined experimentally.

Keywords: human-level artificial intelligence; self-regulated learning; teachable systems.

Introduction

Since the 1956 Dartmouth conference, researchers are taking seriously the challenge of creating machines capable of general human-level intelligence, human-like learning and self-improvement (McCarthy et al., 1955), yet the distance toward this goal has hardly decreased since 1955 (in fact, it is practically difficult to evaluate to make any judgment). Main progress made since the beginning is visible in reassessment of the difficulties, in emergence of new approaches, e.g., cognitive architectures (Newell, 1990; Gray, 2007), and in a new understanding of the goal. Yet, despite growing interest to the field, it is not clear what the final goal is and what would be the impact of achieving it.

Here four different views and associated with them schools of thought can be named: (1) computational neuroscience, that tries to understand how the brain works in terms of neurophysiological mechanisms and neuroarchitectures; (2) cognitive modeling, pursuing higher-level computational description of human cognition and behavior based on more abstract cognitive architectures; (3) human-level artificial intelligence, aiming at generally intelligent artifacts that can replace humans at work; and (4) a new emergent paradigm (that can be called "machine consciousness", or "human-like learners", etc.) aiming at artificial minds that can be understood by humans intuitively, that can learn like humans, from humans and for human needs. All these are fundamental scientific problems. While (1) and (2) are focused on understanding the roots of human cognition, and therefore allow for validation of their progress using brain-and-behavior data, (3) and (4) are oriented toward creation of an expected phenomenon that does not exist yet: a computational replica of the human mind capable of human-like learning and cognitive growth.

From this point of view, defining a success criterion for (3) and (4) is the key challenge on the roadmap to a human-level learner. Examples of practically inefficient guiding criteria include ambitious global challenges like the Turing test (Turing, 1950; Korukonda, 2003) and limited challenges like beating humans in chess or poker. As a result of the missing clear understanding of the overarching goal, the original high spirit appears to be lost in specific, incremental steps that together did not produce a quantum leap and seem to lead nowhere. On the other hand, both, computational neuroscience and microelectronics made tremendous progress in recent decades. Today there is no big mystery in functioning of the brain elements, in the sense that the generally accepted view of the brain is that of a natural information processing device. Capabilities of the modern computing hardware are approaching or have already exceed computational resources of the human brain, therefore, hardware is not a bottleneck on the road to human-level intelligence. Then, what are we missing?

Several recent conferences, e.g., the BICA symposia (Samsonovich, 2008, 2009, members.cox.net/bica2010) addressed this question. To summarize the situation: it becomes clear that biologically inspired cognitive architectures (BICA) provide the most promising approach to creating a functional replica of the human mind, and the key to solving this challenge is in replication of the human ability to learn. In other words, having a machine that can be taught virtually anything that a human child can be taught would imply the achievement of the grand overarching goal. Therefore, in order to successfully steer research toward this goal, it is necessary (a) to better understand principles and characteristics of human learning at a big scale: e.g., in educational practice, rather than in limited behavioral laboratory experiments, and (b) to match the same characteristics of learning in artifacts.

From this point of view, the overarching goal should be formulated as a challenge in terms of scalability laws characterizing the learning process over a long period of time, when previously learned knowledge and skills enable the acquisition of new knowledge and new learning techniques, and so on. This large-scale process of bootstrapped learning can be visualized as a chain reaction (e.g., as depicted in Figure 1). Understanding details of this bootstrapped learning process, such as principles of self-regulated learning (SRL: Zimmerman, 2002), is vital for achieving the goal. At the same time, description of the general laws and their mathematical understanding is possible at an abstract level. Indeed, this is a task of high priority on the roadmap toward general human-level artificial intelligence, or artificial human-like minds.

Similarly to the history of the study of the nuclear chain reaction, models and theories are needed to describe and understand the learning chain reaction in detail in order to make it achievable. The key task is to clarify the difference between solutions leading to small incremental steps and solutions leading to a big quantum leap, and to identify conditions (e.g., the 'critical mass' of the initially available capabilities, if it makes sense) that make the leap possible.

The present work presents an attempt to start developing the necessary theory by constructing and simulating a general model of large-scale learning of the above sort, using learning of an abstract curriculum as an example.

Methods

First, in this section a challenge scenario is outlined that can serve as a goal for designing a general-purpose human-like learner and at the same time as an example justifying abstract modeling. Then, based on this scenario, an abstract model is constructed that will be studied numerically in the next section.

An Example Challenge Scenario

The following challenge scenario for an artificial learner can be viewed as an operational definition of certain key aspects of the human learning ability that needs to be replicated in artifacts: in particular, scalability, robustness, cross-domain transferability, and most importantly, its metacognitive nature. This paradigm will also allow experimenters to measure the "critical mass" of initial knowledge and skills that enable human-level bootstrapped learning.

Settings: The agent is embedded in a virtual learning environment with the study material (the textbook and supplementary materials normally available to students) encoded electronically and made available to the agent. All interactions between human instructors and the agent are mediated by an interface implemented at a symbolic level. The agent is expected to make progress in study of the curriculum, being taught by human instructor(s) via the high-level symbolic interface. The domain of study can be limited, e.g., to high school algebra. The agent together with instructors will go through chapters of the textbook (e.g., Algebra 1 and Algebra 2: Larson et al., 2007a,b) in their electronic representation, will learn new concepts, will ask questions, will take quizzes, will do exercises, and will be able to explain the learned material.

Approach: The approach to implement an artificial student capable of learning how to solve high school algebra problems can be based on a biologically-inspired cognitive-metacognitive architecture implementing principles of SRL (e.g., Samsonovich, De Jong and Kitsantas, 2009; Samsonovich, 2010).

Metrics and Criteria: The challenge for the agent is (i) to demonstrate a long-term ability to learn the curriculum step-by-step, starting from basic concepts available initially (examples: an integer number, a variable, a function) and gradually moving to their practical usage and to complex constructs based on them, and (ii) to demonstrate improvement of meta-learning skills over time, in particular, SRL skills (Zimmerman, 2002). The agent performance improvement over time at the task level will be measured using standard tests and metrics used for evaluation of student academic progress. The agent will be evaluated not only based on its task-level learning and problem solving performance, but also based on changes in the approach to problem solving, using general metrics for SRL (Winne & Perry, 2000). A particularly interesting question relates to the measure of initial intelligence (the "critical mass") that enables learning of the curriculum.

Generalizations: This example of a challenge scenario entails a generalization. Taking a step to the abstract level, it makes sense to introduce of elements of an abstract curriculum as a system of interrelated abstract units (skills, facts, etc.) that the agent may have the ability to learn, depending on its current knowledge and experience. This is done below.

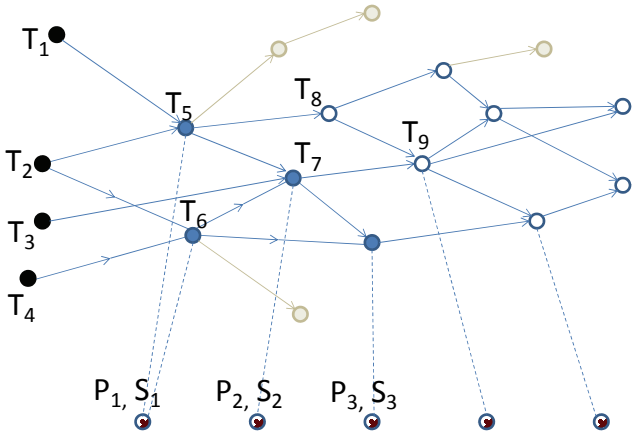


Figure 1: A possible curriculum structure. Solid circles (black and blue) represent available techniques.

Designing an abstract Model

A simple abstract model of a learning process consistent with the above scenario can be defined as follows. Elements of the model are abstract problems P, abstract facts and techniques T applicable to the problems, and abstract solutions S understood as subsets of techniques associated with problems. It is assumed that the set of techniques $\{T\}$ is ordered based on their mutual dependence, excluding a possibility of circular dependence. This ordering can be understood as resulting from a process of adding new techniques one by one to the set, as follows. Given a set of available i-1 techniques, the new technique T_i is defined as an abstract function f_i of a subset of the available i-1 techniques:

$$T_i = f_i\left(\{T_j : \quad 0 < j < i, W_{ij} = 1\}\right), \quad \sum_j W_{ij} = m, \quad (1)$$

where W is a randomly generated, sparse Boolean matrix with zero elements for all $i \leq j$. To simplify the modeling study, it is assumed that each row of W (except the first m rows) has exactly m nonzero elements that are uniformly sampled in the part of the row before the diagonal. Each of the first m rows has the maximal number of nonzero elements satisfying (1). Therefore, a typical matrix W looks like the example plotted in Figure 2 B.

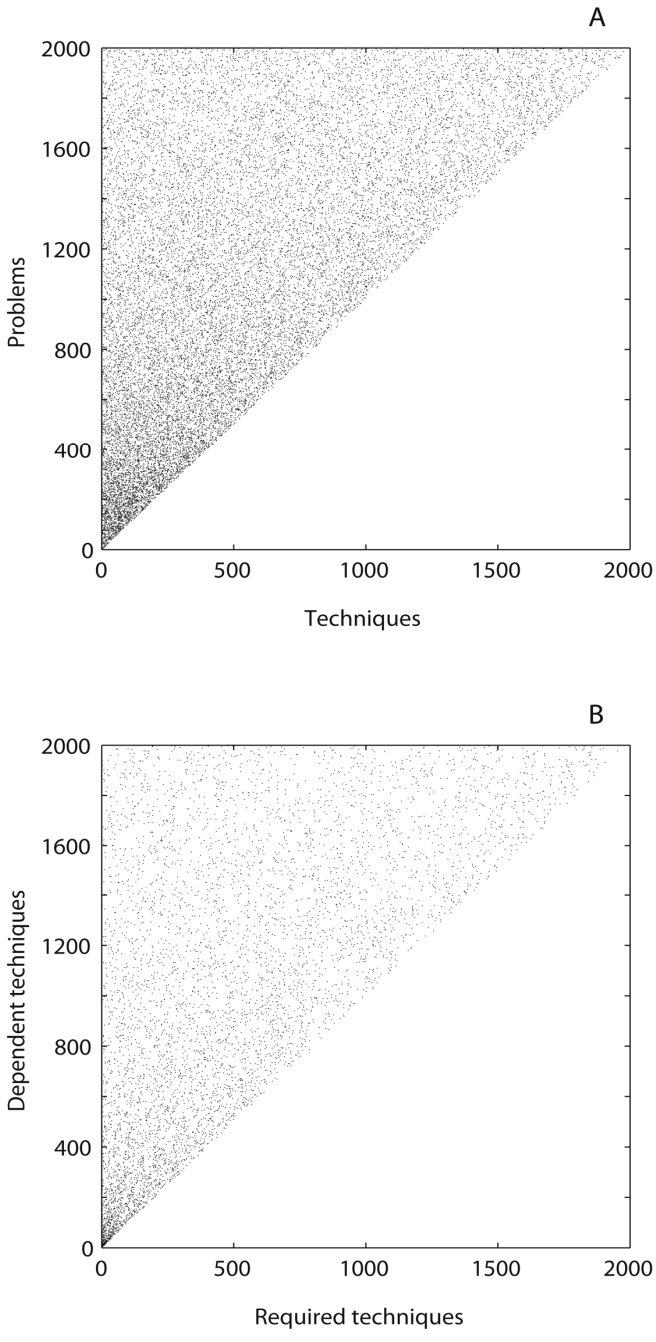


Figure 2: Examples of the abstract curriculum materials used in simulations. A: Example set of solutions for each of the 2,000 abstract problems. B: Example set of dependencies among the 2,000 abstract techniques.

Next, it is assumed that the set of problems $\{P\}$ is ordered similarly, and solutions S to problems are described by a matrix W' defined by a duplicate of (1), only with a different parameter m':

$$S_i = \left\{ T_j : \quad 0 < j < i, W'_{ij} = 1 \right\}, \quad \sum_j W'_{ij} = m', \quad (2)$$

To simplify the study, each problem is assumed to have exactly one solution, which is defined as a certain set of m' techniques. Of course, in real life, knowing a set of techniques that together are sufficient for solving a given problem does not imply the ability to solve the problem: one needs to know in what order and how to apply those techniques. For now, however, we assume that this part of solution is available automatically whenever each of the set of techniques that are necessary to solve the problem is mastered. Therefore, a typical matrix defining solutions of problems looks like the plot in Figure 2 A.

Naturally, there should be also an abstract notion of applicability of techniques to problems: each technique is either applicable to a given problem or not. In this sense, a solution must be a subset of techniques applicable to the given problem. The matrix of applicability $A = \{A_{ij}\}$ tells us whether T_i is applicable to P_j. The matrix A is again generated as a Boolean random matrix with given sparsity.

Model Dynamics

From the learner's perspective, each problem in this simplistic model is either solved or not. Similarly, each technique has three possible states: unavailable, available (yet not mastered), and mastered. The model dynamic rules of learning are defined as follows. A technique becomes (and forever remains) available when all techniques on which it depends are mastered. A technique becomes (and forever remains) mastered when it is successfully used to solve a problem. A technique can be used to solve a problem when it is available. A problem that is already solved is not considered again (indeed, with exactly one solution for each problem, its consideration would change nothing).

Experimental Paradigm

The abstract learning paradigm is the following. Given a set of n techniques and n problems (having the same n just simplifies the consideration), of which initially k techniques are available and zero problems are solved, the learner tries to solve as many problems as possible by randomly picking problems that are not solved yet and trying to apply available techniques to them. This is done in discrete steps. At each step, one problem is selected, and if the solution is contained within the set of available applicable techniques, then the problem is considered solved, and those techniques that constitute the solution become mastered. Simultaneously, states of all techniques are updated based on the above stated dynamic rules: this may result in the emergence of new available techniques. The process ends after a fixed number of steps N_{max}. The progress made in

learning is measured by the final numbers of solved problems, available and mastered techniques.

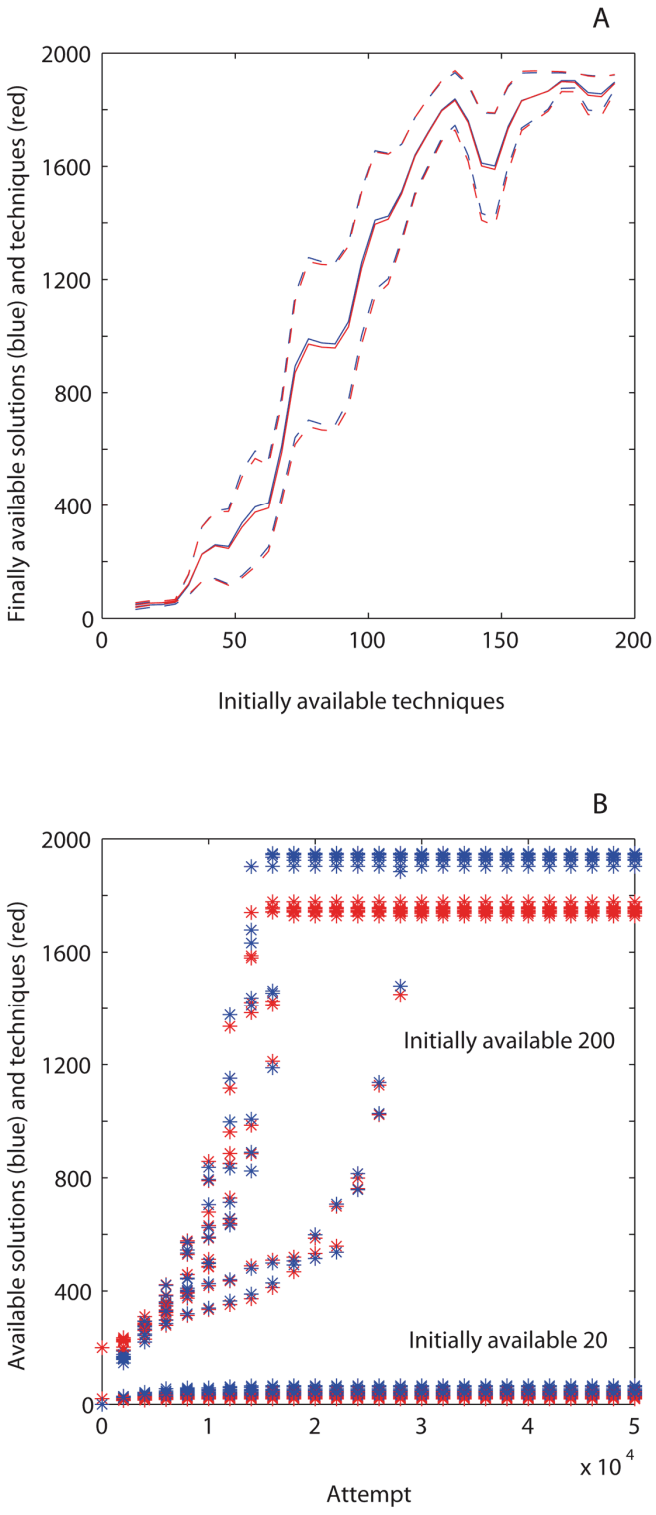


Figure 3: Simulation results. A: Results of learning as a function of initially available techniques. The dotted lines indicate the standard error. B: Learning curves for different numbers of initially available techniques.

Simulation Results and Analysis

The model described above was simulated on a computer with the following parameters. The number of techniques and problems was $n = 2{,}000$, the maximal number of dependencies for techniques $m = 3$, the maximal length of a solution = 10, the sparsity of the applicability matrix = 0.05, the number of steps N_{max} = 50,000, and the number of initially available techniques was $k = 200$ and $k = 20$ in two sets of trials. The results were averaged over 10+10 trials. Results of simulations are summarized in Figure 3.

Figure 3A shows the dependence of the number of finally available solutions (blue) and techniques (red) as functions of the number k of initially available techniques that was varied gradually from 20 to 200. Figure 3B shows multiple learning trajectories for two different values of k: 20 and 200. In this case, the numbers of available solutions (blue) and techniques (red) are plotted as functions of the number of learning attempts.

Results indicate two distinct regimes: "limited learning" ($k < 70$) and "global chain reaction" ($k > 100$). As Figure 3A shows, limited learning is characterized by the number of finally available solutions and techniques that is commensurate with the number of initially available techniques. By contrast, the chain reaction regime is characterized by the number of finally available solutions and techniques that is close to the maximal given number of solutions and techniques n. Figure 3B demonstrates clear clustering of the learning trajectories corresponding to the two regimes.

The qualitative observation of the two regimes is robust with respect to variation of parameters of the model. The transition is clearly visible in Figure 3 and shows a tendency to become sharper as the number of learning steps and the number of problems and techniques n are increased (not shown in Figure 3). The null hypothesis, that dynamics of learning produces results linearly increasing as a function of the amount of initial intelligence, can be ruled out based on the sigmoid appearance of the curve in Figure 3A: residuals of the linear fit exhibit the standard error several times along a substantial fraction of the curve (not shown in Figure 3A). A more rigorous validation of the result will be presented elsewhere.

Speaking generally, the observed chain reaction behavior resembles the phenomenon of percolation, which is characterized by threshold dynamics (a phase transition: see Ziman, 1979). The present limited study, however, does not allow for a detailed investigation of the phase transition characteristics.

Discussion

The presented study and its results are limited in many aspects. For example, from the cognitive psychological perspective, this approach may seem to be fundamentally lacking in an appreciation for the difficulty in understanding human cognition and learning and how they vary as a function of many variables (experience, context, personality differences, etc.). It is nevertheless always reasonable to

start with a simple general model, and then correct the findings by including missing details. This was the motivation behind the choice of the presented model and its simulation. In the present version, the model and its simulation results do not tell us precise details about how humans learn throughout their lifetimes, and this was not the intent. On the other hand, the present work opens a new topic in cross-disciplinary discussions: only through meaningful interactions between hardcore computer scientists and mathematicians and psychologists and cognitive neuroscientists we will be able to achieve the overarcing goals described in the Introduction. Certainly, more formal analysis informed by all related disciplines will benefit artificial intelligence, psychology and neuroscience, as well as other disciplines.

The study and its analysis presented here constitute a first step of its kind, paving the way to finding general scalability criteria for intelligent learning systems intended to achieve the human level of performance. In this respect, it would be very interesting to compare results of this study with available data from educational studies and from machine learning: this will be done elsewhere. The questions to be addressed in future studies include: Do we see the same kinds of limited learning and global learning chain reaction regimes in human children? How this model might be made less abstract and more related to existing specific models of learning? How to include cross-discipline learning into the model? How to specifically describe the role of metacognition within the abstract formalism? And so on.

Returning to the topic of the Introduction, the present study outlined a new kind of scalability criteria that will be useful guiding the design of human-level learners. The knowledge that learning dynamics have a threshold nature allows in principle to identify the threshold (the "critical mass") for a given learning system based on its abstract modeling and then to set the achievement of this "critical mass" as the goal of development.

Conclusions

Designing an agent that can master one specific cognitive skill in specific settings may be feasible today based on traditional approaches in artificial intelligence, regardless of the level of the selected cognitive skill. At the same time, designing an agent that can grow cognitively from a child to an adult human level of general intelligence is challenging. To solve the challenge, it is important to understand the different nature of the two tasks in terms of mathematical characteristics of the expected learning process. The present work addressed this goal with a simple abstract model of a long-term learning process. Results indicate that this process of learning is characterized by two distinct regimes: (1) limited learning and (2) a learning chain reaction that extends through the entire learning material.

The transition between the two regimes is determined by the 'mass' of initially available learning skills and techniques. Therefore, the notion of a 'critical mass' for a human-level learner makes sense and can be determined experimentally. Therefore, this criterion can be used to guide research in human-level teachable systems.

Stated simply, while an intelligent learning agent may be successful in learning starting from any amount of initial knowledge, it needs to begin with a "critical mass" of knowledge and skills in order to successfully self-teach to learn much more than was conceived during its design – up to the human level of general knowledge.

Acknowledgments

I am grateful to members of the GMU BICA team, Drs. Kenneth De Jong and Anastasia Kitsantas, who fueled my thinking about the problem addressed by this study. This work is supported by the Center for Consciousness and Transformation, George Mason University.

References

Gray, W. D. (Ed.) (2007). *Integrated Models of Cognitive Systems. Series on Cognitive Models and Architectures.* Oxford, UK: Oxford University Press.

Korukonda, A. R. (2003). Taking stock of Turing test: a review, analysis, and appraisal of issues surrounding thinking machines. *International Journal of Human-Computer Studies*, 58: 240-257.

Larson, R., Boswell, L., Kanold, T. D., & Stiff, L. (2007a). Algebra 1. Evanston, IL: McDougal Littell.

Larson, R., Boswell, L., Kanold, T. D., & Stiff, L. (2007b). Algebra 2. Evanston, IL: McDougal Littell. ISBN 9780618595419.

McCarthy, J., Minsky, M.L., Rochester, N., & Shannon, C.E. (1955/2000). A proposal for the Dartmouth summer research project on artificial intelligence. In Chrisley, R., & Begeer, S. (Eds.). *Artificial Intelligence: Critical Concepts.* Vol. 2, pp. 44-53. London: Routledge.

Newell, A. (1990). *Unified theories of cognition.* Cambridge, MA: Harward University Press.

Samsonovich, A. V. (Ed.). (2008). Biologically Inspired Cognitive Architectures: Papers from the AAAI Fall Symposium. AAAI Technical Report FS-08-04. Menlo Park, CA: AAAI Press. ISBN 978-1-57735-396-6.

Samsonovich, A. V. (Ed.). (2009). Biologically Inspired Cognitive Architectures II: Papers from the AAAI Fall Symposium. AAAI Technical Report FS-09-01. Menlo Park, CA: AAAI Press. ISBN 978-1-57735-435-2.

Samsonovich, A. V. (2010). A human-inspired cognitive architecture supporting self-regulated learning in problem solving. In: Raja, A. & Josyula, D. (Eds.). *Metacognition for Robust Social Systems: Papers from the 2010 AAAI Workshop, AAAI Technical Report* WS-10-07. Menlo Park, CA: AAAI Press (forthcoming).

Samsonovich, A. V., Ascoli, G. A., & De Jong, K. A. (2006). Human-Level Psychometrics for Cognitive Architectures. In Proceedings of the Fifth International Conference on Development and Learning (ICDL5), CD-ROM. Bloomington, IN: Indiana University, 2006.

Samsonovich, A. V., De Jong, K. A., and Kitsantas, A. (2009). The mental state formalism of GMU-BICA.

International Journal of Machine Consciousness 1 (1): 111-130.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, LIX: 433-460.

Winne, P.H., & Perry, N.E. (2000). Measuring self-regulated learning. In P. Pintrich, M. Boekaerts, & M. Seidner (Eds.), *Handbook of self-regulation*. Orlando, FL: Academic Press.

Ziman, J. M. (1979). *Models of Disorder: The Theoretical Physics of Homogeneously Disordered Systems.* Cambridge, UK: Cambridge UP.

Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory into Practice* 41 (2): 64-70.

The Effects of Communication Medium Upon Collaborative Orientation Task Performance

Laura M. D'Andrea (dandrea1@illinois.edu) & Wai-Tat Fu (wfu@illinois.edu)
Applied Cognitive Science Lab, Human Factors Division & Beckman Institute
405 N. Mathews Ave., Urbana, IL 61801

Abstract

Pairs of dispersed individuals are often forced to solve orientation tasks collaboratively. The present study examines how collaborative orientation tasks are solved when pairs of individuals complete the tasks using one of three computer-based communications (text, audio, and video). Both simple and complex tasks were presented. Pairs in the audio condition outperformed those in the text and video condition overall, and specifically on complex tasks, despite the fact that the video condition allows for the greatest amount of information to be communicated. Analysis of conversations between pairs indicates that those in the video condition had different conversational behavior. Results suggest that social effects of video communication may impair collaborative orientation task performance.

Keywords: Spatial Cognition; Interactive Behavior

Introduction

An understanding of spatial relationships is critical for successful interaction with the world around us, impacting our ability to complete tasks as simple as reaching for a pencil and complex as maneuvering an environment (Taylor & Tversky, 1996). The ability to orient oneself in an environment with the use of a navigational aid, such as a map, is a particularly interesting spatial task. This situation demands that the map-reader's personal perception/view of the environment be aligned with the map's view, a frame of the same environment from a perspective entirely independent of, and depicting locations external to, the reader (Gunzelmann, Anderson, & Douglass, 2004; Klatzky, 1998). These differing perspectives of the same environment are referred to, respectively, as egocentric and allocentric frames of reference (Klatzky, 1998). To orient oneself in an environment, one must recognize how the two frames of reference correspond and depict the same environment; in other words, the reference frames must be aligned. The act of aligning reference frames likely requires somewhat more complex processing than a mental rotation of the two perspectives, because the frames are two distinct formats of information (egocentric vs. allocentric) (Gunzelmann, Anderson & Douglass, 2004). Though various processing strategies are used to mentally coordinate the different perspectives of a scene, including array rotation and viewer rotation, all accomplish orientation within the environment through the same overall strategy - alignment of the differing reference frame types (Gunzelmann, Anderson & Douglass, 2004).

The current ubiquity of communication-oriented technologies has, in some ways, added complexity to the process of orienting oneself in an environment. The aforementioned work on orientation was done with regard to a single individual. However, a lost driver can now easily call a friend for directions rather than look at a map. An astronaut repairing a broken device in space can receive instruction from ground control on how to repair it if the instruction manual is outdated. In these and countless similar situations the orientation task is distributed across multiple geographically distributed individuals, each with information that is crucial to solving the task but insufficient on its own, and each with a different frame of reference.

Disparate, communicating individuals presumably must orient themselves in the same general manner as is done by an individual – by aligning the available egocentric and allocentric reference frames (Gunzelmann, Anderson, & Douglass, 2004). In distributed orientation tasks, however, reference frames cannot be aligned by examining the egocentric and allocentric frames and physically aligning them (as would be possible if an individual were lost and had a map in hand). Therefore, it seems that communicative partners can only overcome the disparity in their reference frames by actively discussing pertinent spatial relationships within the environment, until they are able to align each other's perspectives. The role of communication in distributed collaborative orientation tasks is critical, and the fact that the individuals are not co-located introduces challenges to their ability to effectively communicate. For instance, compared to face-to-face collaborators, dispersed collaborators have been found to often have different understandings of the information/task at hand and of the meaning of their partner's actions (e.g., silence during a conversation), and show a reduced ability to establish and maintain a common understanding of each others' knowledge of the situation or task at hand (Cramton, 2001; Diamant, Fussell, & Lo, 2008). A major contributing factor to these challenges is the inability of computer-mediated communication tools to allow for the same access to social and contextual cues that are visible during face-to-face interaction (Cramton, 2001; Diament, Fusell, & Lo, 2008).

The type of communicative technology being used also introduces potential issues. By nature, different types of computer-mediated technologies (e.g., audio conference, video conference, text communication) convey different levels of cues about one's partner, and can differentially impact how an individual feels about his partner and their task performance (Diamant, Fussell, & Lo, 2008). Diament, Fussell, & Lo (2008) evaluated the impact of three communication mediums – Text, Audio, and Video – to examine in relation to one another, and found that

technology type interacted with the culture of individuals to predict their attributions of performance (Diament, Fussell, & Lo, 2008). Diament, Fussell & Lo's (2008) findings indicate that the affordances of a technology determine the way in which it influences attributions of performance. Perhaps the affordances of those technologies also have differential impacts upon how individuals work together to complete collaborative orientation tasks, tasks in which communication is critical to solving the task. We expect the video condition to generate the best performance, the audio to allow the second best, and the text to result in the worst. This prediction is based upon the amount of information that each communication type provide (i.e., the video condition allowing individuals to not only speak but also use gesture to help describe spatial relationships within their view of an environment. We tested this possibility with an experiment studying the impact of different communication mediums upon collaborative orientation task performance.

Method

Overview

A collaborative orientation task is a type of spatial task that can only be solved when multiple individuals work together, combining their knowledge to deduce the solution. The impact of communication medium type upon collaborative orientation task performance was studied by requiring pairs of individuals to work together to solve spatial tasks while communicating through one of three communication mediums: text, audio, or video chat. Each of our collaborative orientation tasks included two unique displays of task-relevant information, one for either participant in the pair. Both displays contained solution-critical information, and the task demanded that pair members communicate their information in order to ultimately deduce the cardinal direction of the target.

Participants

Participants were 48 adults over the age of 18, recruited from Champaign, IL and paid for their participation. Participants were randomly assigned to a pair. One pair of participants (Audio condition) failed to perform the tasks, and their data was not included in the analysis. Of the remaining 46 adults (mean age=24.6; mean years of education=15.7), 29 were female and 17 were male. Participants were screened for normal or corrected-to-normal vision.

Measures

Spatial abilities were measured with two paper-pencil tasks. Participants' ability to mentally rotate objects was measured with the Mental Rotation Test (MRT) (Vandenberg and Kuse, 1978). Perspective taking ability (i.e., the ability to imagine how a scene looks from a different location in space) was assessed with the Perspective Taking/Spatial Orientation Task (PTSOT) (Hegarty & Waller, 2004; Kozhevnikov & Hegarty 2001).

Collaborative Orientation Task Stimuli

Our stimuli were adapted from those used by Gunzelmann, Anderson, & Douglass (2004). In their study, a single task contained two separate displays of information that needed to be reconciled to solve the task; individuals completed the tasks alone. In our study, a single task contained the two separate displays of information. However, we gave only one display to either member of the pair (one for the Responder and one for the Instructor). We also slightly modified the appearance of the tasks. The Responder was presented with a 2D array of seven images and one target icon, all located in one of the eight cardinal directions (*North, South, East, West, Northeast, Northwest, Southeast, Southwest*). The Instructor was given a display showing two of the seven images seen by the Responder, as well as an arrow indicating North (relative to the center of their screen). In either pair members' display, the icons maintained identical spatial relationships with other icons. However, the entire array was rotated to some degree (rotations of 90° increments), so the Instructor and Responder's displays were not identical. See Figure 1 for examples of each display, as they would appear in an actual task.

For each task, pairs' goal was to deduce the cardinal direction in which the target was located, relative to the X in the center of the Responder's screen. The Responder was ultimately responsible for reporting the direction of the target. Because the Responder was given information regarding the target's location relative to other images, and the Instructor was given the cardinal directions of certain images, pairs needed to discuss their displays (e.g., images, directions of images, relationships between images, etc.) in order to align their perspectives of the displays and deduce the target's direction. Stimuli of two levels of complexity (simple, complex) were displayed. In simple tasks, each icon on the Responder's display was unique. In complex tasks, the Responder's display contained multiples of the two icons that were present in the Instructor's display.

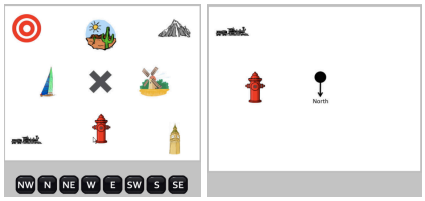


Figure 1. Sample trial displays for a simple task. The left is a display seen by a Responder; the right is a display seen by an Instructor (correct response=Southwest).

Procedure

Within each pair, individuals were randomly assigned to their roles (Responder vs Instructor). Pairs were randomly assigned to one of the three communication conditions. 8

pairs participated in each condition; however, one pair in the Audio condition was not included in analysis due to a failure to perform the tasks. The orientation task portion of the experiment was conducted with both pair members present in the same room, but seated at computers separated by enough distance/barrier so that participants were out of each other's sight and hearing range during the task. Before beginning the collaborative orientation tasks, individuals completed demographic and spatial tasks. They were then shown to their computers and instructed as to what their communication medium would be. On each computer, an instruction screen was presented which explained the tasks and offered a sample display representative of what each individual would view, depending upon their role (Responder vs. Instructor). The pair then performed one practice task before beginning a set of 20 tasks, 10 complex and 10 simple. In each condition, the task workspace took up half of the computer screen; the other half contained the communication tool (*Text condition:* an IM chat box; *Video:* Skype video chat interface (see Figure 2); *Audio*: Skype audio chat interface). Accuracy of task performance and conversations between pair members were recorded.

Across all pairs, the practice trial was identical. However, each of the 20 actual trials was randomly generated for each pair. This randomization included: the 7 icons that appeared on the Responder's display (randomly selected from a master set of 18 icons); the target's location on the Responder's display; the direction of North on the Instructor's display; the relative locations of the two icons appearing on both the Instructor's and Responder's displays; the degree of disparity between the Instructor's and Responder's displays (90 increments); the distribution of complex/simple tasks throughout the 20 overall tasks.

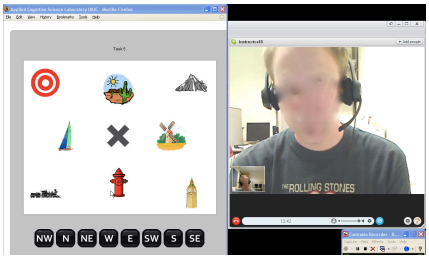


Figure 2. Screenshot from Video condition (Responder's computer screen)

Equipment

Stimuli were presented on each participant's computer screen; responses were made using the mouse. Camtasia screen capture software was used to record audio and video feeds in the Audio and Video conditions.

In the Text condition, participants communicated by typing to each other using AOL Instant Messenger. In the Audio condition, participants wore Logitech ClearChat headphones (with microphone) when performing the tasks. Auditory communication was enabled through the use of Skype's auditory calling feature. In the Video condition, participants wore Logitech ClearChat headphones (with microphone) when performing the task, as there is an auditory component to video chatting. Video chat communication was enabled through the use of Skype's video chat feature. Each computer was supplemented with Logitech Webcam Pro 9000 cameras in order to permit video chatting.

Results

Performance

We performed a two-way ANOVA examining the effect of communication media (video, audio, text) and task difficulty (simple, complex) on overall accuracies. There main effect of communication media was not significant ($p=0.11$), but the main effect of task complexity was significant ($F[1,21]=4.22$, $p<0.05$). There was also a significant communication by complexity interaction ($F[2,21]=3.55$, $p<0.05$). Simple effect analysis between media conditions in simple tasks showed no significant difference, but the difference was significant in complex tasks ($F(2,21)=5.86$, $p<0.05$). Posthoc tests (Fisher's LSD) showed that the audio condition was significantly better than the text condition ($t(6)=4.8$, $p<0.01$) and the video condition ($t(6)=3.1$, $p<0.01$), but the difference between text and video was not significant (see Figure 3). This demonstrated that pairs in the Audio condition performed the tasks better than pairs in either the Video or Text conditions only in the complex tasks.

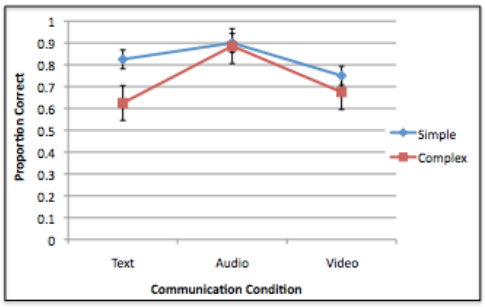


Figure 3. Average proportion of trials correct by communication medium, as a function of trial complexity.

To assess whether these differences in task performance was linked to pairs' spatial aptitude, rather than the communication medium being used, spatial ability task scores were analyzed. Results indicate no significant difference in the abilities of individuals in the three communication conditions. A one-way ANOVA examining MRT scores of individuals in the different conditions (video, audio, text) revealed no significant difference in the MRT scores of individuals in the different communication conditions ($F[2,43]=0.70$, $p=0.50$). Similarly, a one-way ANOVA examining PTSOT scores across conditions (video, audio, text) revealed no significant difference in the PTSOT scores of

individuals in the different communication conditions ($F[2,41]=0.11$, $p=0.90$).

We also examined the relationship between a pair's average spatial ability score and their collaborative orientation task performance. For each pair, a single score was generated for both the MRT and PSOT by averaging the scores of the two individuals within the pair. This score (denoted with "-P") was then correlated with task performance. The MRT-P score was significantly correlated with task performance in the Text condition ($r=0.86$, $p<0.01$, $df=6$) and the Video condition ($r=0.77$, $p<0.05$, $df=6$). However, MRT-P was not correlated with performance in the Audio condition ($r= -0.38$, $p=0.40$, $df=5$). The same trend followed with regard to the PTSOT-P scores. PTSOT-P was significantly correlated with performance in the Text ($r= -0.87$, $p<0.01$, $df=6$) and was marginally correlated with performance in the Video condition ($r= -0.68$, $p=0.09$, $df=5$), but was not correlated with performance in the Audio condition ($r= -0.41$, $p=0.36$, $df=5$). The lack of correlation between spatial ability and performance within the Audio condition may be due to the restricted range of performance observed within this group. Additionally, averaging spatial ability scores is not an optimal approach to examining abilities across conditions, as an average score can obscure potentially interesting information (e.g., relative abilities of the Responder and Instructor in each pair).

The effect of practice on task performance was also assessed. Average scores for each trial were correlated with trial number. Overall performance on the collaborative orientation tasks, regardless of communication medium, was significantly correlated with trial number ($r=0.59$, $p<0.01$). Trial number was not significantly correlated with performance in the Text condition ($r=1.32$, $p=0.20$). It was marginally correlated with performance in the Audio condition ($r=1.98$, $p=0.06$). Performance of pairs in the Video condition, was significantly correlated with trial number ($r=2.46$, $p<0.05$). This finding suggests differential effects of practice depending upon the communication medium being used; thus, the communication mediums, rather than the task itself, are impacting whether practice improves performance.

Conversational Analysis

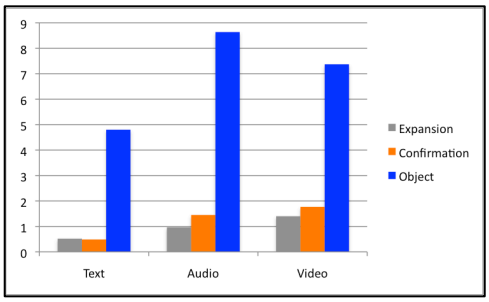
To investigate underlying factors as to why the Audio condition allowed for superior performance, the conversations between each pair were transcribed and coded. A coding scheme was developed post-hoc and addressed four main categories of Utterance Type: Object Description, Revision/Repair, Request for Confirmation, and Request for Expansion. These types are rooted in Clark and Wilkes-Gibbs's (1986) work regarding the ways in which pairs of individuals, during conversation, collaborate to reach agreement on the noun phrase being referred to (the noun phrase being crucial to understanding what each other is trying to communicate).

Within each of these main Utterance Type categories were sub-categories regarding the contents of the utterance. In turn, each of these Utterance Content categories contained more-specific Statement Type categories. Each utterance was coded with regard to the main Utterance Type (e.g., revision/repair, request for expansion), and with respect to the different specific Statement Types within utterance content categories A, B, and C. Each utterance could receive multiple categorizations.

We were interested in whether there were differences across communication conditions in their use of the main Utterance Types, as well as whether the use of different main Utterance Types related directly to task performance. Therefore, a 4 (utterances types) X 3 (communication media) X 2 (correctness of response) ANOVAs with average frequencies of occurrences of utterances as dependent variable was conducted. Occurrences of each utterance type were normalized, taking the frequency of utterance in proportion to the number of trials that had occurred. Results indicated a significant three-way interaction ($F(6,60)=4.33$, $p<0.001$), a significant two-way interaction between correctness and media ($F(2,20)=4.39$, $p<0.05$), and significant main effects of utterance types ($F(3,20)=43.24$, $p<0.001$) and correctness ($F(1,20)=20.42$, $p<0.001$). Given that we observed interactions between media and correctness, the significant 2-way interaction and main effects were likely caused by the different number of correct and incorrect trials in each medium. We therefore focus on the further analyzing the 3-way interaction.

We performed separate 3 (media) x 2 (correctness) ANOVAs on each type of utterances. Results showed significant main effect of media for requests for expansion ($F(2,20)=4.17$, $p<0.05$). Post-hoc comparisons showed that Video had significant more requests for expansion than audio ($t(7)=2.87$, $p<0.05$) and text ($t(7)=3.00$, $p<0.05$) conditions. We found significant main effects of correctness and media for requests for confirmation ($F(1,20)=16.15$, $p<0.05$) and $F(2,20)=3.67$, $p<0.05$ respectively). Posthoc comparisons showed that Video had significant more requests for confirmation than text ($t(7)=3.25$, $p<0.05$). We found significant main effects of correctness and significant interaction between correctness and media for object description ($F(1,20)=23.1$, $p<0.01$) and $F(2,20)=4.7$, $p<0.05$ respectively). Posthoc comparisons showed that *only in incorrect trials,* Video had significant more requests for confirmation than audio ($t(7)=2.47$, $p<0.05$). There was no significant difference in any of the variables for Revision/repair. [See Figure 4].

4a)



4b)

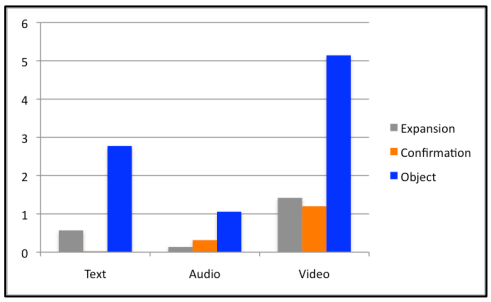


Figure 4. In each communication media, average utterance type per trial: 4a) Correct trials; 4b) Incorrect trials.

The Expansion and Confirmation request types are indicative of unsatisfactory or insufficient information communication between partners, indicating individuals' need/desire to gain more information from their communicative partner (Clark and Wilkes-Gibbs, 1986). Intuitively, one might think that the video condition would result in the fewest requests because it allows for greater amount of information to be communicated. However, participants in the video condition made more requests for expansion than individuals in text or audio, and more requests for confirmation than participants in text or audio (when incorrect performance resulted), indicating that the information communicated was more frequently deemed insufficient by pair members in the video condition than in either text or audio.

Discussion

This study's results were interesting in that, contrary to our predictions, the video condition did not induce the highest task performance level, but was in fact worse than audio and on the same level as text. The spatial abilities of individuals in each communication condition did not differ, indicating that our performance findings did not result from an overload in high or low ability individuals within specific conditions. Further research is needed to uncover the reasons behind observed performance trends, as our study was not geared to investigate several of the underlying factors that may have contributed to the performance differences. For instance, in the future we may match participants on spatial ability across conditions so as to better account for the abilities of pairs (rather than individuals) within conditions, and perhaps examine the impact of the relative ability of each pair member upon performance. However, current findings suggest an interesting social influence on cognition. The remainder of this discussion focuses upon tying together our performance measure and conversational analysis results.

We found evidence that pairs' performance on collaborative orientation tasks is impacted by the type of communication medium used during task solving. Contrary to our expectation that the Video condition would result in the highest level of task performance, pairs in the Audio condition outperformed pairs in the Video and Text conditions both overall and on complex tasks, while performance of simple tasks was no different across communication mediums. So, it appears that all three communication mediums allowed for good performance on easier tasks, but some aspect of the auditory communication medium allows its users to sustain their performance level when tasks increase in complexity. In addition, the finding that the joint measure of pairs' spatial ability (MRT-P, PTSOT-P) was not correlated with task performance in the audio condition, but was in text and video, suggests that some factor inherent to the audio communication medium was at the root of its optimality for solving collaborative orientation tasks.

Our conversational analysis, aimed towards investigating why pairs in the other communication mediums did not show the same performance, revealed differences in how pairs in different communication conditions actually communicated information to each other. Two of the main utterance types – Requests for Expansion and for Confirmation – were used more by pairs in the Video condition than they were by pairs in both Audio or Text. Before moving forth with the discussion of our findings, it is important to reiterate the purpose of these types of utterances. Clark and Wilkes-Gibbs (1986) explain that when two people are speaking, over the course of the conversation one of them will utter a statement (specifically, a noun phrase) that their listening partner deems unacceptable or inadequate. The unacceptability could occur because the listener needs more of a description to understand what their partner is referring to (request for expansion), or because they want to clarify that what they heard is correct (request for confirmation) (Clark and Wilkes-Gibbs, 1986).

Although task performance overall and on complex tasks was significantly better in the audio condition, overall communicative behavior was essentially the same in the audio and text conditions. This suggests that the there were factors inherent to the textual condition that impacted task performance without effecting how pairs actually worked together. Previous research indicates that textual communication is simply more difficult for pairs to work with, which could be the case here. Cramton (2001) discusses how various traits of text-based communication, including the slower rate of information exchange and the demand to communicate typically non-verbal cues with words (i.e., saying 'yes' instead of nodding), impede performance in text communication mediums. Text-based systems do not provide significant feedback, like verbalizing 'yeah' or 'mmhm' to indicate

understanding, and thereby imposes on pairs' ability to develop a shared knowledge of the situation (Cramton, 2001). Our text condition certainly presented these issues, which could have been the root of the resulting poorer performance. Another possibility is that the processing/resource demand of the text condition was higher than in audio, and while pairs could overcome the text condition's inherent difficulties in simpler tasks and they were unable to do so in more complex tasks. If this were the case it would follow predictions of theory on the impact of resources on multiple-task performance (Wickens, C. D., 1991). However, our study did not specifically examine any of these factors; therefore, a conclusion regarding the poorer performance in the text condition cannot be reached.

Performance in the video condition was also significantly poorer than that in the audio condition. The video condition *did* allow for some amount of Cramton's (2001) described non-verbal feedback so a lack of social and verbal cues cannot be entirely blamed for performance (though video communication is still deficient when compared to face-to-face; Cramton, 2001). It is possible that performance in the video and text conditions were both rooted in some cognitive/attentional load issue; however, the load induced by the video condition appears much higher than that of the text (as a large video feed of another individual was in close proximity to the task, compared to a text message box). If it were cognitive demand that decreased performance, one would expect the video condition to have experienced a more significant impact. And if the processing/resource demand had affected conversational behavior, one would expect text and video conditions to have communicated in the same manner. But, only the video condition incited pairs to use more requests for expansion and confirmation during their conversations. The difference in conversational behavior, and more specifically in the types that were differing (requests for confirmation /expansion, both statement types that indicate inadequacy of initial communication/a need to confirm what was said (Clark and Wilkes-Gibbs, 1986), suggesting that the video condition spurs a sub-optimal communicative behavior between partners. The video condition's poor performance levels indicate that this behavior was detrimental in some manner to task performance. The root of this behavior could lie in social effects imposed by the fact that video condition participants could see each other while communicating. For instance, perhaps individuals unfamiliar with each other restrict their display descriptions due to some discomfort felt by knowing that a stranger is watching them while they think. Additional research is needed to further examine the mechanism driving performance in different communication media, as we did not aim to examine such social influences. Follow-up work should include larger sample size and more trials, to better examine the effects of practice in varying conditions, and the aforementioned control of pairs' spatial abilities.

Conclusion

We found that audio communication allowed users to maintain a high level of performance on collaborative orientation tasks of varying complexities. Text and video communication mediums made the tasks more difficult to perform. In the case of video communication, this decrease in performance is likely tied to the style of conversational behavior incited by the communication medium. More research is needed regarding both the cause of decrease in performance observed in textual communication, and the reason behind the shift in conversational behavior observed in the video condition.

Acknowledgements

We thank Miriam Holtzman and Namrata Donthamsetti for help running the experiment and preparing data for analysis.

References

Clark, H. H., Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition, 22,* (1-39).

Cramton, C. D. (2001). The Mutual Knowledge Problem and Its Consequences for Dispersed Collaboration. *Organizational Science, 12*(3) (346-371).

Diamant, E. I., Fussell, S. R., & Lo, F. (2008). Where did we turn wrong? Unpacking the effects of culture and technology on attributions of team performance. *Proceedings of CSCW 2008.*

Gunzelmann, G., Anderson, J. R., Douglass, S. (2004). Orientation Tasks with Multiple Views of Space: Strategies and Performance. *Spatial Cognition and Communication, 4*(3), 207-253.

Hegarty, M. & Waller, D. (2004). A dissociation between mental rotation and perspective-taking spatial abilities. *Intelligence, 32* 175-191.

Klatzky, R. L. (1998). Allocentric and egocentric spatial representations: Definitions, distinctions, and interconnections. In C. Freksa, C. Habel, and K. F. Wender (Eds.). *Spatial cognition: An interdisciplinary approach to representing and processing spatial knowledge* (pp. 1–17). New York: Springer-Verlag.

Kozhevnikov, M., & Hegarty, M. (2001). A dissociation between object-manipulation spatial ability and spatial orientation ability. *Memory and Cognition, 29,* 745–756.

Shepard R. N., Hurwitz, S. (1984). Upward direction, mental rotation, and discrimination of left and right turns in maps. *Cognition, 18*, 161-193.

Taylor, H.A. & Tversky, B. (1996). Perspective in Spatial Descriptions. *Journal of Memory and Language, 35*, 371-391.

Vandenberg S.G., Kuse, A.R. (1978). Mental rotations, a group test of three-dimensional mental rotation. *Perceptual and Motor Skills (47)*2, 599-604.

Wickens, C. D. (2001). Processing resources and attention. In Damos, D. L. *Multiple-task performance.* (pp. 3-34). Taylor & Francis, Inc., PA: Bristol.

Semi-supervised learning is observed in a speeded but not an unspeeded 2D categorization task

Timothy T. Rogers, Charles Kalish, Bryan R. Gibson, Joseph Harrison and Xiaojin Zhu
Departments of Psychology and Computer Science
University of Wisconsin-Madison
Madison, WI 53706 USA

Abstract

Recent empirical studies of semi-supervised category learning—where learners only occasionally receive information about a given item's category membership—have yielded contradictory results, with some studies showing strong effects of unlabeled experience and others little or no effect. We report two experiments designed to help understand this heterogeneity. In both, participants performed a two-category classification task with novel stimuli varying along two psychologically separable dimensions. In semi-supervised conditions, participants categorized and received feedback on 32 "labeled" items intermixed with a large number of "unlabeled" items. In the supervised-only condition, participants viewed the same labeled trials intermixed with a large number of filler trials. Without time pressure participants learned the task equally well in both conditions. When required to respond very rapidly, however, participants performed substantially better in the semi-supervised condition. The discrepant results may indicate a role for selective attention in human semi-supervised learning.

Keywords: semantic interference, visual lexical decision, dual-task, single-system view

Introduction

Most theoretical and computational approaches to human category learning consider fully supervised learning: for every training experience, the learner has access to a representation of the stimulus and to the true category label (e.g. Nosofsky, 1986; Kruschke, 1992; Gluck and Bower, 1988; Anderson, 1991 and many others). Fully unsupervised approaches—where the learner never has access to the true category label but must learn to group items into categories on the basis of their similarity—are less common but have also appeared in the literature (e.g. Fried and Holyoak, 1984). Neither approach seems fully adequate, however, for explaining human categorization. Although a great deal of natural experience is unsupervised—we continually encounter objects in the world without a "teacher" telling us what kind of things they are—we also certainly get a nontrivial amount of "labeled" experience, where a recognized authority provides the true class label either directly in an explicit teaching scenario or indirectly through use of the label in communication. Human category learning may, therefore, involve combining both labeled and unlabeled sources of information—that is, human category learning may be semi-supervised.

The question of how best to combine labeled and unlabeled data has been a topic of considerable investigation in machine learning, where it has been formally shown that, for some kinds of learning problems, a learner can converge much more quickly on an accurate representation of the category structure by combining labeled and unlabeled observations (Chapelle, Zien, and Scholkopf, 2006; Zhu and Goldberg, 2009). In cognitive psychology, the empirical question of how experience with both labeled and unlabeled items might influence category learning has rarely been studied. Some well-known computational approaches to category learning suggest ways in which labeled and unlabeled observations might combine to influence knowledge of category structure (e.g. Nosofsky, 1986; Schyns, 1991; Love et al. 2004), but these ideas have not been linked to the formal analyses offered by machine learning and have not been a focus of much empirical work.

We are aware of only two studies designed to assess whether category learning is influenced by unlabeled experiences, and these come to opposing conclusions. On the positive side, Zhu and colleagues (2007) studied performance in a 1-dimensional 2-category learning task. After learning a category boundary with a small amount of supervised training (ie training with corrective feedback), participants subsequently classified a large number of items with no feedback. These "unlabeled" items were sampled from a bimodal distribution with a trough that was displaced to one side or the other of the original learned category boundary. The authors found that, following the unlabeled experience, participants shifted their mental category boundary toward the trough of the unlabeled distribution. This finding suggests that people expect category boundaries to align with low-density regions in the unlabeled feature space, and use unlabeled observations to adjust their representations of category structure accordingly.

In contrast, Vandist and colleagues (2009) studied a binary classification task with stimuli that varied in two psychologically separable dimensions (the orientation and spatial frequency of Gabor patches). Participants viewed a number of labeled examples intermixed either with additional unlabeled examples or with unrelated filler items. Unlabeled items were sampled from a bimodal distribution in which the trough aligned with the true category boundary. The authors found no difference in the rate of learning or overall performance between these conditions—suggesting that the unlabeled items provided no overall benefit in learning the category structure, even though the distribution

of these items was consistent with the to-be-learned boundary.

In this paper we investigate some of the factors that might explain the different results obtained by these studies. Though both groups focused on semi-supervised learning, there were several key differences in their experiments: (i) Where Vandist et al. used stimuli varying in two psychologically separable dimensions, Zhu et al. employed visually complex shapes varying along a line in a multidimensional feature space. (ii) Where Vandist et al. provided participants with many labeled items, Zhu et al. trained participants with 10 repetitions each of just 2 individual tokens (ie one exemplar of each category). (iii) Vandist et al. employed a task requiring participants to integrate two separable dimensions (ie the category boundary was oblique in the 2D feature space) whereas Zhu et al. employed a simple 1D category learning task. (iv) Vandist et al. provided participants with ongoing labeled training experiences, whereas Zhu et al. performed a short block of supervised learning followed by a long block of unsupervised trials. (v) Vandist et al. compared performance in a semi-supervised condition to performance in a fully-supervised condition, whereas Zhu et al. compared two different semi-supervised conditions.

Thus there are several potential hypotheses as to why different results were obtained in the two studies. We report two experiments designed to narrow the range of possible hypotheses by capitalizing on the positive characteristics of both Zhu et al.'s (2007) and Vandist et al.'s (2009) original designs. Like Vandist and colleagues, our experiments (i) employ stimuli that vary along two obvious and psychologically separable dimensions, (ii) compare a semi-supervised condition to a matched supervised condition, and (iii) provide participants with ongoing exposure to labeled data. Like the experiment described by Zhu et al., (i) our stimuli were more object-like, (ii) participants in the semi-supervised condition received relatively few labeled trials (8%), and (iii) the boundary to be learned did not require integration of the two dimensions. In Experiment 1 we show that, under these conditions, people seem relatively insensitive to unlabeled learning experiences. Experiment 2 then tests a more explicit hypothesis about the conditions under which unlabeled experiences influence performance.

Experiment 1

Method

Participants. 50 undergraduate students from UW-Madison participated in Experiment 1 for course credit or monetary compensation. All had normal or corrected-to-normal vision.

Materials and Design. The stimuli were derived from classic work by Nosofsky (1986). They consisted of circles bisected by an oblique line, and varied in radius (ie circle size) and in the precise angle of the bisecting line. Like the dimensions employed by Vandist et al. (2009), size and line orientation are two psychologically separable dimensions—that is, it is possible to attend selectively to one dimension without processing the other. In our stimuli, circle radius varied from 50 to 120 pixels while line orientation varied from 0 to 90 degrees (measured from the horizontal).

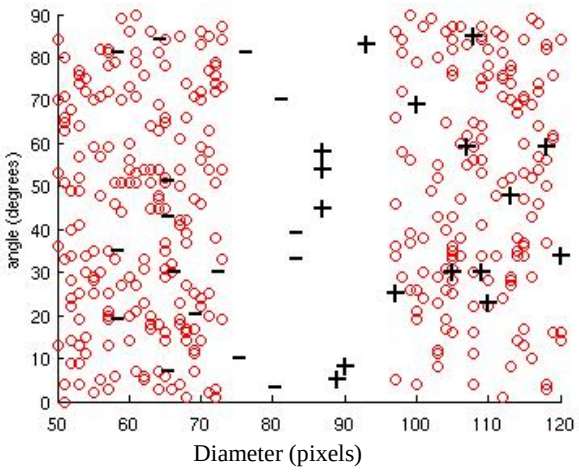


Figure 1. Example of the distribution of labeled (black) and unlabeled (red) items for one participant. Plus signs show labeled items from Category A, minus signs show labeled items from Category B.

Pilot testing with a fully-unsupervised procedure showed a general bias for classifying these stimuli according to the angle dimension—only 35% of participants made unsupervised categorization decisions based on size. Consequently our experiments involved learning to classify these items according to their size. Items larger than or equal to 85 pixels in radius were designated class A while those smaller than 85 pixels were designated class B.

The experiment included two between-subjects conditions. In the *semi-supervised* (SS) condition, participants viewed a total of 32 labeled items—items for which feedback was provided—sampled from a uniform distribution over the space. These were intermixed with 400 unlabeled examples sampled from a bimodal distribution that was uniform along the angle dimension but had a substantial gap along the size dimension (see Figure 1). Thus the gap in the unlabeled distribution provided a potential cue to orientation and location of the true category boundary. In the *supervised-only* (SO) condition, participants viewed the same 32 labeled items as in the semi-supervised condition. In this case, however, these items were intermixed with filler trials in which participants viewed the word "left" or "right" on the screen and pressed the corresponding mouse button. Labeled trials were ordered so that 8 appeared in each block of 100 unlabeled/filler trials. Subjects in the SS and SO

conditions were yoked so that each SO participant viewed exactly the same labeled items in exactly the same sequence and at exactly the same time as a participant in the SS condition. Thus the only difference between conditions was whether the trials interspersed among labeled examples consisted of unlabeled examples or of filler. After experience with the labeled and unlabeled/filler trials, both groups categorized, without feedback, 36 items forming an evenly-spaced "grid" in the stimulus space. Performance was assessed as the mean proportion correct in each successive block of 8 labeled items and on the unlabeled grid items.

If participants use the gap in the unlabeled distribution to form their mental category boundary, their accuracy on the labeled items should increase more rapidly, and their performance on the final grid should be better overall, than participants in the control condition.

Procedure The experiment was carried out on PCs running the DMDX software package under Windows XP. The 50 participants were randomly assigned to either the SS or SO condition with 25 participants in each. Participants in both groups were told that they would view a series of objects and that each belonged to one of two categories. Their job was to learn to classify the objects correctly by pressing one of two buttons on the mouse. Participants in both conditions were told that they would only occasionally get feedback indicating whether their choice was correct, but that they should do their best to categorize all of the items regardless. Participants in the SO condition were additionally told that categorization trials would be interspersed with button-pressing trials in which they would view the word "left" or "right" and must press the corresponding mouse button. The principal dependent measure was the mean proportion correct for each successive block of 8 labeled items and for the 36 unlabeled grid items.

Results

Figure 2 (top) shows means and standard errors of the accuracy for each block of 8 labeled items and for the final unlabeled grid in the two conditions. A repeated measures ANOVA treating time (each block of 8 labeled items plus final grid) as a within-subjects factor and learning condition (SS / SO) as a between-subjects factor revealed a significant main effect of time with performance improving overall ($F(5,192) = 5.36$, $p < 0.001$), but no effect of learning condition ($F(1,48) = 0.29$, $p = 0.59$) and no interaction between these ($F(4,192) = 0.51$, $p = 0.73$).

Performance overall was highly variable, with some participants learning fairly well and others not at all. In fact performance on the final grid was bimodal in both groups, with one subgroup choosing correctly on 67% or more of the grid trials and the other group at chance. We therefore classified each participant as a "learner" or a "nonlearner" based on grid performance, with learners showing accuracy greater than 66%. The number of learners in each condition was comparable (13/25 in the semi-supervised group, 12 /25 in the control group), suggesting that the unlabeled items did not produce a greater likelihood of learning the correct boundary.

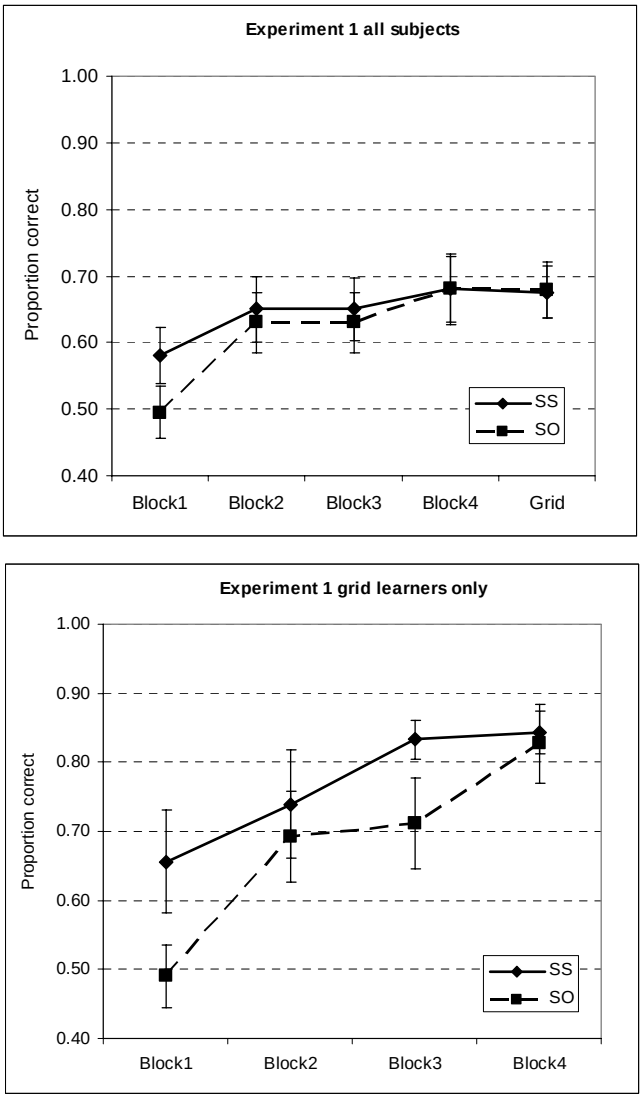


Figure 2. Top: Mean proportion correct for labeled items and grid for all participants in Experiment 1. Bottom: Mean proportion correct for labeled items in each block across participants who performed above criterion on the final grid. Error bars indicate the standard error of the mean.

Finally we investigated the effect of time and learning condition on accuracy for the 4 blocks of labeled items considering just those participants who performed to criterion on the grid items. These data are shown in Figure 2 (bottom). Though learners in the SS condition appeared to perform marginally better, this effect was not statistically reliable. A repeated-measures ANOVA showed a reliable main effect of time ($F(3,69) = 10.9$, $p < 0.001$) but no effect

of condition ($F(1,23) = 2.2$, $p = 0.15$) and no interaction ($F(3, 69) = 1.0$, $p = 0.40$).

In sum, we obtained no evidence for semi-supervised learning in this experiment: though unlabeled items were selected from a distribution with a prominent gap that aligned well with the true category boundary, experience with this distribution did not significantly impact the overall rate of learning, the mean accuracy, or the number of participants who learned successfully.

Experiment 2

Consistent with the observations of Vandist and colleagues (2009), Experiment 1 showed little effect of unlabeled experience on category learning. What then accounts for the strong effects of unlabeled experience previously observed by Zhu et al. (2007)? Experiment 2 tested one hypothesis: perhaps the difference is observed because, in both the current work and in Vandist et al.'s (2009) experiment, the stimuli were composed of two psychologically separable dimensions. A classic tradition of research in *concept attainment* has shown that, for such stimuli, people often adopt a "win-stay-lose-shift" strategy (Bruner, Goodnow and Austin, 1956). That is, they formulate a hypothesis about the relevant dimension for categorization, then make their decision based solely on that dimension until they receive evidence that their hypothesis is wrong, at which point they shift to a new hypothesis. If feedback is very sparse, participants may focus on the dimension they believe to be relevant to the exclusion of other dimensions. That is, participants may not attend to the competing dimension at all on many trials, and so may be exposed to very little information about the distribution on this dimension. Especially for our stimuli, where pilot studies suggest that participants are biased to attend to the irrelevant dimension (angle), such strategic/attentional effects might seriously attenuate any influence of unlabeled experience.

To test this hypothesis, we conducted a second study identical to Experiment 1 in all but one respect: in Experiment 2, participants were required to respond within a deadline of 600ms. With this requirement of a very rapid response, participants have little time to focus their attention on one dimension or the other. Consequently, we predicted that the distribution of unlabeled examples would have a more significant impact on category learning in this paradigm.

Method

Participants 50 undergraduate students who did not participate in Experiment 1 were recruited for this study in return for course credit. All participants had normal or corrected-to-normal vision.

Materials and Designs The materials and design were identical to Experiment 1, except that participants in both groups were told that they would need to respond to each item as rapidly as possible.

Procedure Participants were randomly assigned to one of the 2 conditions, with 25 participants in each group. The procedure was identical to Experiment 1 with the following exceptions. First, each stimulus appeared onscreen for 125ms and was then replaced by a visual mask composed of hash marks. Participants were given 600ms from the onset of the mask to make their response. If the participant did not respond within this window, the computer indicated that the response was too slow. On labeled trials that did not meet deadline, the computer indicated that the response was too slow and also presented the correct category label. In both conditions, the deadline was imposed on both labeled trials and on unlabeled/filler trials.

Results

Trials that did not meet deadline were discarded from the analysis; these included just 5% of trials on average. Thus most participants were able to respond within the time-window on the majority of trials. For the remaining trials, we computed the mean accuracy on each successive block of 8 labeled trials and on the final unlabeled grid. Results are shown in Figure 3.

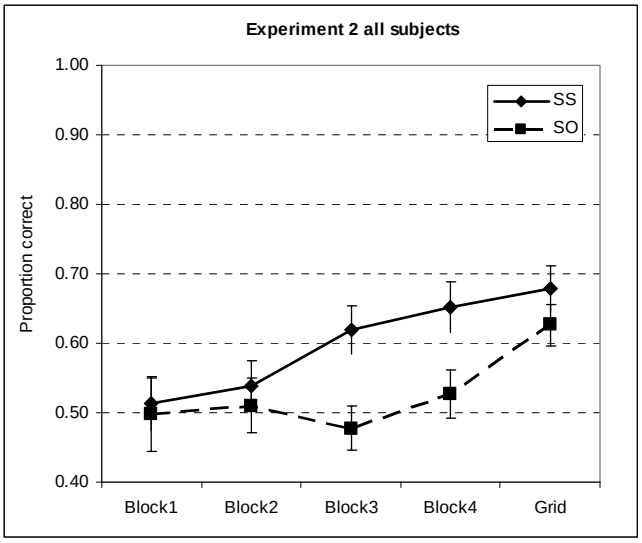


Figure 3. Mean proportion correct across all participants in Experiment 2 for labeled items in each block and grid. Error bars indicate the standard error of the mean.

In contrast to Experiment 1, participants in the semi-supervised condition showed greater accuracy across all blocks and on the final grid. A general linear model treating time (4 successive blocks of 8 labeled items + grid) as a within-subjects factor and learning condition (SS versus SO) as a between-subjects factor revealed reliable main effects of both factors (for time, $F(4,192) = 6.8$, $p < 0.001$; for learning condition, $F(1,48) = 4.32$, $p < 0.05$) and no

interaction between them ($F(4, 192) = 1.2$, $p = 0.32$).

As previously we also computed the number of participants who performed to a criterion of 67% or better on the final grid in each condition. In the SS condition, more than half the participants exceeded this criterion (13/25) whereas less than a third did in the SO condition (8/25). These odds are different with likelihood $p<0.08$ according to a one-tailed test of the log odds ratio.

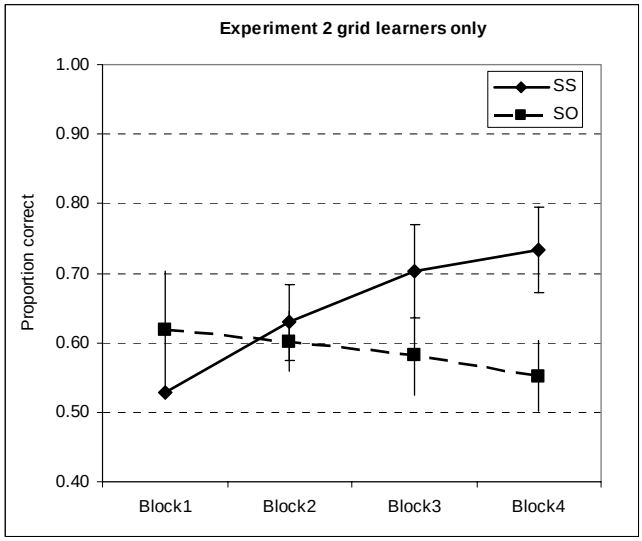


Figure 4. Mean proportion correct in Experiment 2 for participants who performed above criterion in the final grid. Error bars indicate the standard error of the mean.

Finally, we again considered mean accuracy over successive blocks of labeled items in just the participants who performed to criterion according to their grid accuracy. In these participants performance was much better in the SS than the SO group, with accuracy on labeled items improving from 50% to 73% for learners in the SS group but not exceeding chance on any block in the SO group. A general linear model of these data showed no reliable main effect of time or learning condition but these factors did interact significantly ($F(3,60) = 2.8$, $p < 0.05$) . Inspection of Figure 4, which plots these data, explains the absence of any main effect and the interaction: performance did not improve significantly at all for the 8 participants in the SO group who performed above criterion on the final grid, but did improve substantially for those participants in the SS group. Consistent with these observations, oneway repeated-measures ANOVAS conducted separately for the two groups showed significantly different accuracy across blocks for learners in the SS condition ($F(3,36) = 4.6$, $p < 0.009$) but not in the SO condition ($F(3,24) = 0.3$, $p = 0.83$).

In sum, when responses were speeded, providing little time for strategic control of attention, participants in the SS condition performed more accurately overall, were marginally more likely to learn to criterion, and learned labeled items more rapidly than participants in the SO condition.

Discussion

In two experiments we assessed whether the ability to learn a simple 2D binary classification task is influenced by unlabeled experiences. In the first experiment, where participants responded with no time pressure, we observed little evidence that unlabeled data matter: participants performed equally well, were equally likely to learn, and learned equally rapidly regardless of whether they received unlabeled learning items. In the second experiment, which was identical in all respects except that participants were pressured to respond rapidly, we observed a very different pattern: in this case, experience with unlabeled items led to better overall performance, a greater likelihood of learning to criterion, and more rapid learning compared with supervised learning only. Like Vandist et al (2009), we found little evidence that unlabeled data influence category learning when response times were unconstrained. When responses were speeded, we replicated Zhu et al.'s (2007) finding that unlabeled data can produce substantial effects. What accounts for these different patterns?

One possibility concerns the extent to which participants can selectively attend to only some of the stimulus feature dimensions. Prior work has shown that, in categorization tasks where it is possible for participants to form an explicit categorization rule, learning depends importantly upon mechanisms of attention and cognitive control (Ashby and Maddox, 2005). In Zhu et al.'s (2007) work, stimuli varied along a line in a complex multidimensional feature space—therefore it was impossible for participants to selectively attend to information that was irrelevant to the category learning task. In contrast, in Vandist's et al.'s (2009) work and the current study, stimuli varied in two psychologically separable dimensions. If participants selectively attended to only one of these, so that distributional information about the unattended dimension was not available to the learning system, effects of unlabelled data might be attenuated or eliminated—producing the null result in Vandist's (2009) work and in Experiment 1.

On this hypothesis, the robust influence of unlabeled data in Experiment 2 was observed because participants lacked sufficient time to selectively attend to just one feature dimension. If, under speeded conditions, both stimulus dimensions are fully represented, then the unlabeled distribution should have a more robust impact on learning. On this view, it is not the speed of response that matters *per se*, but whether or not the learning system has access to all of the relevant distributional information. If this account is correct, it predicts that unlabeled data should have a stronger effect for multidimensional stimuli where the stimulus dimensions are not psychologically separable, even if response times are unconstrained. We leave this prediction to future work.

We further note that, because there are many factors that differentiate Zhu et al's (2007) study from that of Vandist and colleagues (2009), there remain several additional

hypotheses about the difference in their findings. The current study isolates speed of response as an important mitigating factor, but other potentially important factors—including the orientation of the category boundary in the stimulus space, the ratio of labeled to unlabeled examples, and the temporal distribution of labeled examples over the learning session—should be parametrically explored in future work.

More generally, the question of whether or not people make use of unlabeled observations when learning categories has strong implications for theories of human conceptual knowledge. Many researchers have noted that even young children are able, with just a handful of learning experiences, to infer the extension of many category labels (Hall and Waxman, 2004; Keil, 1979; Markman, 1989). Once they reach the right age, most children need hear the word "horse" only once or twice before being able to make a reasonable guess about which objects in the world are horses and which not. This rapid learning from sparse data is sometimes held to indicate that children bring strong inductive biases to bear on word-learning (Xu and Tenenbaum, 2007).

Semi-supervised learning suggests a different explanation: Maybe children can learn from just a few labeled examples because they are marrying these sparse episodes to knowledge gleaned from a vast amount of unsupervised experience. If children assume that category labels tend to span relatively dense clusters in a conceptual feature space, and that category boundaries follow the low-density valleys in this space, then—to the extent that this assumption holds—they only need a small number of labeled experiences to work out which labels "go with" which clusters. This explanation frees theories of word-learning from having to rely too heavily on strong inductive biases to explain rapid word-learning abilities in children.

Acknowledgements

This work was supported in part by a grant from the Air Force Office of Scientific Research (AFOSR project number FA9550-09-1-0313) and in part by NSF project IIS-0916038.

References

Anderson, J. R. (1991) The adaptive nature of human categorization. *Psychological Review*, 98, 409-429.

Ashby, F. G. and Maddox, W. T. (2005). Human category learning. *Annual Review of Psychology*, 56, 149-178.

Bruner, J. S., Goodnow, J. J. and Austin, G. A. (1956). *A Study of Thinking*. Hoboken NJ: John Wiley and Sons.

Chapelle, O., Zien, A. and Scholkopf, B. (2006). *Semi-Supervised Learning*. Cambridge, MA: MIT Press.

Fried, L. S. and Holyoak, K. J. (1984). Induction of category distributions: A framework for classification learning. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 10 (2), 234-257.

Gluck, M. A. and Bower, G. H. From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General*, 117(3), 227-247.

Hall, D. G. and Waxman, S. R., Eds. (2004). *Weaving a Lexicon*. Cambridge, MA: MIT Press.

Keil, F. C. (1979). *Semantic and Conceptual Development: An Ontological Perspective*. Cambridge, MA: Harvard University Press.

Kruschke, J. K. (1992). An exemplar-based connectionist model of category learning. *Psychological Review*, 99(1), 22-44.

Love, B., Medin, D. L. and Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, 111(2), 309-332.

Markman, E. M. (1989) *Categorization and Naming in Children*: Cambridge, MA: MIT Press.

Nosofsky, R. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, 115(1), 39-57.

Schyns, P. G. (1991). A modular neural network model of concept acquisition. *Cognitive Science*, 15, 461-508.

Vandist, K., de Schryver, M. and Rosseel, Y. (2009). Semi-supervised category learning: The impact of feedback in learning the information-integration task. *Attention, Perception and Psychophysics*, 71(2), 328-341.

Xu, F. and Tenenbaum, J. (2007). Word learning as Bayesian inference. *Psychological Review*, 114, 245-272.

Zhu, X. and Goldberg, A. B. (2009). *Introduction to Semi-Supervised Learning*. San Rafael: Morgan and Claypool.

Zhu, X., Rogers, T. T., Qian, R., and Kalish, C. (2007). Humans perform semi-supervised classification too. Proceedings of *AAAI 2007*.

Embodied Cognition and Virtual Reality in Learning to Visualize Anatomy

Susan Jang (sj306@columbia.edu)
John B. Black (black@exchange.tc.columbia.edu)
Department of Human Development, Teachers College, Columbia University, 525 West 120th St., New York, NY 10027

Robert W. Jyung, MD (jyungrw@umdnj.edu)
University of Medicine and Dentistry of New Jersey, DOC, 90 Bergen St., 8100, Newark, NJ 07103

Abstract

This study examines the facilitative effects of embodiment of a complex internal anatomical structure through three-dimensional ("3-D") interactivity in a virtual reality ("VR") program. Since Shepard and Metzler's influential 1971 study, it has been known that 3-D objects (e.g., multiple-armed cube or external body parts) are visually and motorically embodied in our minds. Such findings confirm the theory that our mental images, and rotations of these images, are in fact confined by the laws of physics and biomechanics, because we perceive, think and reason in an embodied fashion. With the advancement of new technologies, virtual reality programs for medical education now enable users to interact directly in a 3-D environment with internal anatomical structures. Given that such structures are not readily viewable to users and thus not previously susceptible to embodiment, coupled with the VR environment affording all possible degrees of rotation, how people learn from these programs raises new questions. If we embody external anatomical parts we can see, such as our hands and feet, can we embody internal anatomical parts we cannot see? Does manipulating the anatomical part in virtual space facilitate the user's embodiment of that structure and therefore the ability to visualize the structure mentally?

Medical students grouped in yoked-pairs were tasked with mastering the spatial configuration of an internal anatomical structure; only one group was allowed to manipulate the images of this anatomical structure in a 3-D VR environment, whereas the other group could only view the manipulation. The manipulation group outperformed the visual group, suggesting that the interactivity that took place among the manipulation group promoted visual and motoric embodiment, which in turn enhanced learning. Moreover, when accounting for spatial ability, it was found that manipulation benefits students with low spatial ability more than students with high spatial ability.

Keywords: Embodied cognition; Virtual reality; Visualization.

Introduction

Virtual reality programs have the potential to be the most dramatic change in the way anatomy is taught since Vesalius's richly illustrated volumes of the human body based on careful and intricate cadaver dissections. Although computer technology has undoubtedly transformed the manner in which doctors evaluate and treat their patients (e.g., CT scans, robotic surgery), the methods used to teach medical students have been in place for centuries (e.g., lectures, anatomy textbooks, cadaver dissection). Some believe this is all about to change. Virtual reality ("VR") programs for medical education now enable users to interact directly with, as well as view, anatomical parts in three-dimensions, with the potential to change the way medical students learn anatomy, perform dissections and even practice surgical procedures.

The advent of these programs raises questions for cognitive psychologists, some of which this study aims to address. At the broadest level: how are complex, internal anatomical structures learned through 3-D viewing and interactivity? What factors, from both a cognitive and human-computer interaction perspective, contribute to the learning of anatomy through these VR programs?

This study considers the above-mentioned factors under the framework of embodied cognition: that cognition is inextricably linked to our physical interactions with our environment (Wilson, 2002). Using the embodied cognition framework, this study explores the following research questions: 1) Does the physical manipulation of, versus solely viewing, a complex internal anatomical structure in a virtual reality program facilitate a better visualization of the structure? 2) Does spatial ability affect participants' visualizations in this particular study?

Theoretical Background

The theoretical framework underlying and informing the questions in this study bridges two distinct areas of cognitive psychology through the lens of embodied cognition: mental rotation and imagery and multimedia learning.

Studies in mental rotation and imagery provide some of the most compelling evidence of how cognition is rooted in our bodily interactions with the environment. Shepard & Metzler's seminal research showed that people mentally manipulate objects similarly to the way they would with actual objects in physical space, and that the time it takes to rotate the image increases linearly with the degree of rotation (Shepard & Metzler, 1971; Shepard & Cooper, 1982). Subsequent research using the Shepard & Metzler paradigm has confirmed the proposition that motor processes are involved in mental rotation (Wexler et al., 1998) and that motor cortices (primary/M1 or premotor cortex) are activated when performing the task (Kosslyn et al., 1998).

Additional research in mental rotation and imagery has helped to clarify and refine the nature and extent to which motor processes are connected to mental rotation and

imagery. For example, it is known that there are differences in the way we conduct mental rotations of an object as compared with a body part. This difference arises because the trajectory imagined, for example, for the observer's hand or foot is strongly influenced by the biomechanical constraints specific to the actual movement of the hand and foot. For example, people are faster and more accurate at performing mental rotations of drawings of hands or identifying which hand is pictured when they are asked to imagine rotating the hand that does not require difficult bodily movements (Parsons, 1987a, b; Schwoebel et al., 2001). Given the details of the way the body actually works, the motor imagery system actively facilitates or constrains how quickly mental imagery is executed. Neuropsychological studies have proven that motor processes are recruited when we imagine and manipulate complex 3-D structures in our mind but also that the body's biomechanical constraints actually affect our ability to conduct mental rotations (Amorim et al., 2006).

Research in the area of multimedia learning endeavors to complement the research discussed above in embodied cognition and mental rotation and imagery by analyzing how multimedia programs may be designed to maximize learning and understanding. Recent theories and studies have focused on how the motor or haptic channel, through direct tactile manipulation and feedback can aid in deeper learning and understanding and the degree to which interactivity of any kind is productive (Meyer & Kieras, 1997; Chan & Black, 2006; Black, in press). For example, Chan & Black (2006) investigated how seventh graders are better at visualizing complex concepts such as Newtonian mechanics if they are able to interact with a technology-rich environment allowing for direct-manipulation animation ("DMA"). DMA allowed learners to interact directly with navigation controls, determine their viewing direction and to control the pace of the navigation of the content. Chan & Black found that DMA, which incorporated the haptic channel in the learning process, provided learners with a superior learning experience as compared with those who were in the non-haptic groups (narrative-only, narrative-and static visuals, narrative and animation) about causal interactions and functional relations in systems.

Despite the ubiquity of computer programs that exist for 3-D visualizations of anatomy, there are very few empirical investigations on what makes such programs effective. These studies have started to investigate, from a human-computer interaction and cognitive perspective, what factors contribute to developing successful visualizations of complex anatomy (or anatomy-like) structures from various 3-D visualization programs. From the corpus of these studies, the following variables have emerged as being significant: 1) manipulation (or interactivity) of the 3-D object versus just viewing, 2) the importance of having access to certain views and/or orientations of the structure, and 3) spatial ability of the learner. The most significant studies were conducted by Garg and his colleagues (Garg et al., 1999, 2001, 2002) and Keehner and her colleagues (2008a, b), who concluded that developing accurate visualizations of an anatomical (or anatomical-like) structure has more to do with participants' access the critical views and orientations of the structure than being able to interact with it, and furthermore, that such programs should be used carefully with those with lower spatial ability were found to have had a harder time learning from such programs.

Yet, it is curious that the exact opposite findings have been found in some studies where active exploration appeared to benefit those participants with low spatial ability scores over those with higher spatial ability scores. In a study conducted by Luursema et al. (2006), participants were divided into groups viewing the same computer-generated 3-D images of anatomical parts of the abdomen, but with half of the group viewing the images stereoptically (using shutter-glasses), providing actual depth perception, and the other half of the group viewing the images binocularly (without shutter glasses). Luursema et al. found that a "combination of computer-implemented *stereopsis* (visual depth through seeing with both eyes) and *dynamic exploration* (being able to continuously change one's viewpoint with respect to objects studied in real-time) is beneficial to anatomical learning" (p. 455), and that participants with low visuo-spatial ability benefited more from this combination than participants with high visuo-spatial ability. In a more recent study, Meijer & van den Broek (2010) also found that active exploration actually improved low spatial participants' 3-D mental representations of complex 3-D objects (and had no effect on middle or high spatial participants' representations).

Research Design and Questions

This study builds from, as well as aims to overcome some of the potential confounds of the previous studies, in investigating the effects of interactivity and embodiment in a VR system when learning a complex, internal anatomical structure. First, the computer 3-D visualization program used in this study is more intuitive from a visual and motor processing standpoint. This VR system provides the user with stereoscopic vision with 3-D goggles, allowing for full depth perception of the object of study. In addition, this system has a joystick that allows the user to interact physically/motorically with the virtual object in a similar manner as one would outside of a virtual environment. Both these elements would, in theory, foster a stronger sense of embodiment because of the more realistic and natural aspects of visual and motor information in the VR system. Therefore, it is possible that if the interface of the VR program allows for a more intuitive mechanism for viewing and rotating the virtual object, and is compatible with the human body's natural movements, a participant might be able to develop a better internal 3-D visualization of a complex anatomical structure.

Second, the stimulus used in this study is an internal anatomical structure (as opposed to a fictitious structure or external body part) – the inner ear. In Garg et al.'s studies,

it is possible that the findings were confounded by the stimulus material – the carpal bones – because it is a part of the body that people are very familiar with both visually and motorically. That is, the wrist falls on two natural planes, and people are used to seeing as well as feeling their wrists in those two common positions. Therefore, it is not surprising that there are canonical views of the wrist that would naturally transfer to canonical views of the carpal bones within the wrist. Furthermore, using an internal anatomical structure free of any joint articulation or specific visual cues to orient the structure allows us to begin to investigate how (if at all) and which canonical views users develop of this structure during their study time. In essence, it is addressing the issue of whether the user literally embodies (or maps onto him/herself) the internal anatomical structure.

With these changes, this study addresses the following research questions:

1) Does the physical manipulation of, versus solely viewing, a complex internal anatomical structure in a virtual reality program facilitate a better visualization of the structure? If so, is there a difference in visualizing: a) different sub-structures within the larger structure that have different shapes, i.e., line (e.g., the path of a nerve) versus circles (e.g., semi-circular canals protruding off a surface); and b) the structure from different vantage points (i.e., anatomical planes)?

2) Does spatial ability affect participants' visualizations in this particular study? If so, does it have a different effect for: a) participants who manipulate versus view the structure; and/or b) participants with differing spatial abilities (i.e., low versus high)?

Method

Participants

Seventy-six medical students between the ages of 20-38 years at the University of Medicine and Dentistry of New Jersey, Newark, participated in this study. None of the participants had formal instruction of the inner ear or prior exposure to the VR machine.

Materials

The VR system and target anatomical structure

The VR machine is housed at the University of Medicine and Dentistry, Newark. It generates a stereoscopic 3-D environment that is viewed through stereoscopic 3-D goggles. It has a free-moving, non-mounted joystick, enabling the user to hold, control and manipulate the movement (by rotating on an x-, y-, and z- axis) of the 3-D representation of the anatomical structures in a similar manner as one would be able to with a tangible object outside of a virtual environment.

The target anatomical structure is the inner ear. The inner ear is a structurally complex system concentrated in a very small area in the human skull. The virtual ear model was developed by an otolaryngologist at UMDNJ in conjunction with the engineers of the VR program to ensure accuracy of the model.

Pre-test measures

Participants took the following pre-tests prior to working on the VR machine: 1) a background questionnaire which includes questions on comfort level of using a joystick and playing video games, as well as any prior use of working with 3-D modeling programs; 2) an ear anatomy questionnaire; 3) Vandenberg & Kuse (1978) Mental Rotation Test ("MRT"). This is a standardized test of spatial ability that assesses one's ability to rotate and visualize a 3-D structure; and 4) Ekstrom et al.'s Building Memory Test. This standardized test was used assess participants' ability to remember the location of an object within a map.

Post-test measure

A series of snapshots of the virtual ear model were taken in the following anatomical planes: lateral, superior, inferior, anterior and posterior. The purpose of using all these anatomical planes is to create a 3-D "voxel" of the area of study. For each plane, two snapshots were produced, one without the facial nerve and another without the semi-circular canals. Therefore, the post-test consisted of a total of 10 snapshots.

Procedure

Each participant was tested individually. First, each participant completed all four pre-test measures. Next, the participant was randomly assigned to one of the two conditions (*manipulation, visual)*. The *manipulation* participant was given a brief training period with the joystick in the VR machine. Once the participant indicated that he/she felt comfortable using the joystick, the target anatomical structure (inner ear model) was presented. After providing a brief explanation of the inner ear and how it was positioned in a surgical position, the participant was asked to study the spatial configuration of two sub-structures within the inner ear: the facial nerve and the semi-circular canals. The *manipulation* participant was informed that he/she could use the joystick to rotate the ear model, and was given 5 minutes to study. Each *manipulation* participant's study of the inner ear, based on his/her own joystick movements, was recorded in the VR machine, which was then shown to the yoked, *visual* participant. After the study period, each participant was given the 10 post-test snapshots (randomized order by sub-structure) and asked to draw in, to the best of his/her ability, the missing sub-structure.

Coding

The drawings were assessed for accuracy of visual representation on the following criteria: parts, angle and placement and size. The various individual criterions were summed to derive the following TOTAL scores: 1) overall TOTAL; 2) TOTAL for each anatomical plane; 3) TOTAL for facial nerve; 4) TOTAL for semi-circular canal. The researcher and an independent coder coded the post-test.

Both coders were blind to the identity of the participants and condition assignment each coded all 760 drawings.

Analysis and Results

Participants in the *manipulation* condition scored higher than those in the *visual* condition on all the TOTAL scores (Table 1).

Table 1. Mean score analysis on all dependent measures

	Manipulation Mean (s.d.)	Visual Mean (s.d.)	t(37), p
TOTAL	72.47 (7.303)	60.76 (11.391)	6.437, <.001
TOTAL by sub-structure			
facial nerve	32.39 (5.900)	26.71 (6.375)	4.870, <.001
semi-circular canals	39.74 (3.020)	33.82 (6.673)	5.029, <.001
TOTAL by anatomical plane			
lateral	15.11 (2.051)	13.58 (2.937)	3.153, =.003
superior	13.92 (1.440)	11.63 (2.562)	5.663, <.001
inferior	13.79 (2.183)	10.74 (2.565)	6.481, <.001
anterior	14.18 (2.078)	12.13 (2.622)	4.830, <.001
posterior	15.03 (2.175)	12.45 (2.738)	4.700, <.001

A correlational analysis of all the pre-test measures with TOTAL score showed that only MRT was correlated (r = .0325). A one-way analysis of covariance (ANCOVA) was conducted using MRT as the covariate. A significant interaction effect was found between MRT and condition on TOTAL score, F (1, 35) = 5.168, p < .029. Simple group main effects tests were conducted to assess differences between those who scored lower on the MRT (1 *SD* below the mean = 11.274) and those who scored higher on the MRT (1 *SD* above the mean = 27.166) (Figure 1).

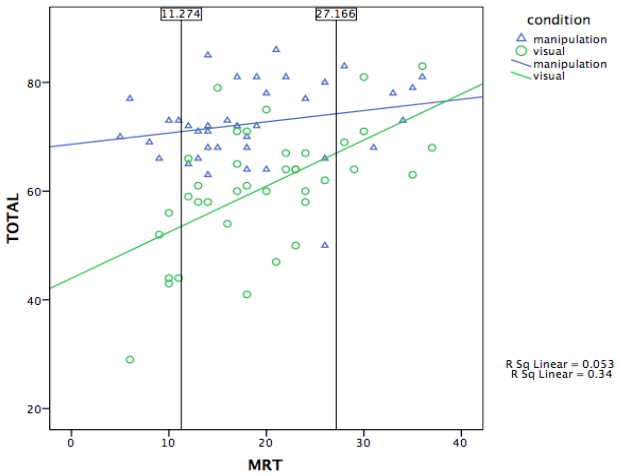


Figure 1. Differences on TOTAL score performance between the two conditions on two levels (high versus low) of the covariate (spatial ability).

Given that the statistical analysis revealed that those in the *manipulation* condition had more accurate 3-D visualizations of the inner ear over those in the *visual* condition, some ancillary questions arose with respect to what the participants in the *manipulation* group were doing. For example, were there certain strategies used by the *manipulation* participants that enabled better embodiments of the inner ear? That is, were there common characteristics of the manner in which *manipulation* participants rotated the structure that led to highly successful performance on the post-test? Or, conversely, what were the common characteristics among *manipulation* participants who performed relatively poorly on the post-test?

A qualitative video profile of the top and bottom performing manipulation participants showed that common characteristics might exist on either end. First, among the top performing manipulation participants, they quickly oriented the structure into the posterior plane, which when put in context of the human body means positioned in an upright manner, standing up and looking forward. In addition to standing the model upright, the top manipulators often went back to this posterior view after exploring other views (as though it grounded them in some way), suggesting this view was the one they were most comfortable with. In contrast, the lowest scoring manipulators did not position the structure in an upright position as quickly as those in the top scoring group. There was no particular familiar or comfortable perspective that developed among the low scorers. Second, all the high scoring participants spent more time studying still positions as opposed to moving the object continuously. In contrast, the low scoring manipulators generally spent their time moving and rotating the object in various, haphazard directions and not holding it still.

Third, and perhaps most interestingly, when the high scoring manipulators moved between these still positions, they moved in a "wiggling" manner between these two planes. There were two kinds of wiggling among the top scoring *manipulation* participants. One type was a wiggling that constituted alternating between two still positions, in what appeared to be a comparison and analysis of the two positions. Another strategy demonstrated a different type of wiggling: choosing one "still" position and varying the view of that position by only a few degrees in either direction.

Discussion and Conclusions

The main finding of this study is that manipulating, rather than viewing, an internal anatomical structure in virtual space strengthens the embodiment of that structure and therefore the ability to visualize the structure. As demonstrated in the analysis (Table 1), the *manipulation* group outperformed the *visual* group regardless of whether the participants were visualizing different anatomical sub-structures or from different orientations (i.e., anatomical planes). Participants who are afforded the opportunity to manipulate in virtual space 3-D images of anatomical structures with which they are not familiar outperform participants who are only given the opportunity to watch the 3-D images being rotated. Such results support the general framework of embodied cognition, that there is an intimate connection between our motor and visual processes, and the more explicit the connection, the better the learning.

Beyond this main finding that the motor and visual processes are connected and provide stronger learning, it is posited that the participants may literally have tried to embody (to varying degrees) the inner ear model by mapping it onto their own bodies. Results of TOTAL scores by anatomical plane, combined with the video analysis, support the theory that the virtual inner ear was embodied by the participants in this study in a more literal sense of the term embodiment – that is, that they mapped the structure onto (or within) their own bodies. The results from the mean score analysis (Table 1) show that regardless of condition, participants performed better on the planes we are more familiar with seeing ourselves and others in (lateral, posterior and anterior) over less familiar planes (superior, inferior). As a general matter, we are much more comfortable and familiar with looking at others face-to-face rather than looking down a person's head (superior) or up a person's chin (inferior). This conclusion is similar to ones reached in studies by Parsons and others who have shown that the real world biomechanical constraints on our physical bodies do in fact constrain our mental abilities – specifically the ability to rotate and visualize a body part in our mind. Therefore, it is possible that participants in *both* the manipulation and visual conditions found that visualizing the ear from the superior and inferior planes was a somewhat physically awkward perspective to embody, as it is rare to look into the top of one's head or look up into one's chin. Even though the virtual ear was displayed in the absence of surrounding physical landmarks that would immediately cause the viewer to orient the image in an upright position, there was a way to orient the image (via embodiment) that made it the anatomical plane more familiar and more comfortable to the participants.

Further support for the embodiment theory is that the video analysis revealed that the top manipulators developed a *canonical viewpoint* (Palmer, Rosch & Chase, 1981) for this model. Palmer et al. coined the term canonical viewpoint to describe perspectives in which identification performance of 3-D objects is best. The canonical viewpoint for the ear model appears to be the posterior plane. The qualitative video analysis revealed that the top scoring manipulators started their study with the structure oriented in the posterior plane and often returned to this position as though it was the most stable position. Given that this is their canonical view, it strongly suggests that the manipulators literally embodied – that is, they mapped onto themselves the inner ear from the perspective of their own body schema, or rather that they projected their bodies onto the object in an embodied fashion, maintaining the body axes (head-feet, front-back, and left-right) when doing so ("bodily projection", Lakoff & Johnson, 1999).

The results from this study also illustrate the facilitative effects of interactivity on embodiment and that the development of an internal visual representation of a 3-D structure depends on the spatial ability of the participant. Specifically, the benefits of embodiment in virtual reality appear to be greater for those participants with lower spatial ability. As shown in Figure 1, those who score lower on tests of rotational spatial ability have more to gain from interacting with a 3-D virtual reality environment than those with high rotational spatial ability. There is a greater difference between the two regression lines at a low MRT score versus at a high MRT score, indicating that manipulation, which strengthens embodiment, may help those with lower spatial ability to perform as well as those who have higher spatial abilities (and may not need the manipulation experience).

It is important to note that the main effect of interactivity runs counter to some of the more relevant studies discussed in the literature review (Garg et al., Keehner et al.) who argue that interactivity, which allows for complete freedom of movement and exposure to views from varying perspectives, may overload the learner and prevent effective visualizations. Why is it then that in this study the participants in the *manipulation* condition outperformed the *visual* participants? Perhaps the answer is that this virtual reality program provided a stronger sense of embodiment or "presence" (Usoh et al., 1999) for the *manipulation* participants with the intuitive interface and stereoscopic depth perception of the target structure. Luursema et al. (2006) used a similar program and found that the combination of stereopsis and dynamic exploration to be beneficial for anatomy learning.

There are some limitations in this study. Regarding the dependent measure, the drawing test, it could be argued that assessing the accuracy of visualization by evaluating a participant's drawings may favor participants who have good drawing skills, while disadvantaging participants who may have successfully understood the visual features of the anatomical structures but were less skilled at transmitting their understanding onto several sheets of paper. One way that a future study might be able to address this limitation is to complement the drawing test with an interview of the participant in which the participant would describe his or her understanding of the anatomical structures and/or explain what he or she was trying to draw.

Another potential confound that exists in this study relates to the yoked-pairs design. Although this design is effective in terms of providing the participants in both conditions with the same visual information, it may be argued that certain strategies employed by the *manipulation* participant may make no difference, or may in fact actually hinder the *visual* participant when studying the model. For example, the wiggling that was used by many top performing *manipulators* may have introduced noise or confusion to the yoked, *visual* participant, which in turn made learning less effective. It is possible that a future study where the *visual* participant watches a recording of a high scoring *manipulation* participant without the wiggling could lead to results where learning is equalized.

The findings from this study present significant implications for the potential role of virtual reality in educational settings generally, as well as in field of medical education. Perhaps most significantly, this study suggests

that it is possible to embody internal anatomical structures that are not generally visible or familiar to people. While it has been known for some time that we embody wrists and hands, it has not previously been shown that we may be able to embody an internal structure we are not even aware of, such as components of the inner ear. It follows logically that if it is possible to embody the inner ear with its substructures (e.g., semi-circular canals and the facial nerve), perhaps it is also possible to embody the spleen or the liver or the heart. While further research in this area is warranted, if is in fact the case that it is possible to embody other parts of our anatomy, then there may be benefits to approaching the teaching of anatomy with an understanding of embodied cognition in mind.

References

Amorim, M., Isableu, B., & Jarraya, M. (2006). Embodied spatial transformations: "Body Analogy" for the mental rotation of objects. *Journal of Experimental Psychology: General*, 135: 327-347.

Barsalou, L.W. (2008). Grounded cognition. *Annual Review of Psychology*, 59, 617-645.

Chan, M.S. and Black, J.B. (2006) Direct-manipulation animation: Incorporating the haptic channel in the learning process to support middle school students in science learning and mental model acquisition. Proceedings of the International Conference of the Learning Sciences. Mahwah, NJ: LEA.

Garg, A., Norman, G. R., Spero, L., & Maheshwari, P. (1999). Do virtual computer models hinder anatomy learning? *Academic Medicine*, 74, S87-S89.

Garg, A., Norman, G. R., Spero, L. (2001). How medical students learn spatial anatomy. *The Lancet*, 357, 363-364.

Garg, A., Norman, G. R., Eva, K., Spero, L., & Sharan, S. (2002). Is there any virtue in virtual reality? The minor role of multiple orientations in learning anatomy from computers. *Academic Medicine*, 77, S97-S99.

Keehner, M., Hegarty, M., Cohen, C., Khooshabeh, P., & Montello, D. R. (2008a). Spatial reasoning with external visualizations: what matters is what you see, not whether you interact. *Cognitive Science, 32*, 1099-1132.

Keehner, M., Khooshabeh, P., & Hegarty, M. (2008b). Individual differences among users: implications for the design of 3D medical visualizations. In F. Dong, G. Ghinea & S. Chen (Eds.), *User Centered Design for Medical Visualization* (pp. 1-24). Hershey, PA: IGI Global.

Kosslyn, S. M., Digirolamo, G. J., Thompson, W. L., & Alpert, N. M. (1998). Mental rotation of objects versus hands: neural mechanisms revealed by positron emission tomography. *Psychophysiology, 35*, 151-161.

Lakoff, G., & Johnson, M. (1999). *Philosophy in the flesh: The embodied mind and its challenge to Western thought*. Chicago: University of Chicago Press.

Luursema, J., Verwey, W. B., Kommers, P., Geelkerken, R. H., & Vos, H. J. (2006). Optimising conditions for computer-assisted anatomical learning. *Interacting with Computers, 18*, 1123-1138.

Meijer, F., & van den Broek, E. L. (2010). Representing 3D virtual objects: interaction between visuo-spatial ability and type of exploration. *Vision Research, 50*, 630-635.

Meyer, D. E. & Kieras, D. E. (1997). A computational theory of executive control processes and human multiple-task performance: Part 1. Basic Mechanisms. Psychological Review, 104, 3-65.

Palmer, S., Rosch, E., & Chase, P. (1981). Canonical perspective and the perception of objects. In J. Long & A. Baddeley (Eds.), *Attention and Performance IX* (pp. 135-151). Hillsdale, NJ: Erlbaum.

Parsons, L. M. (1987a). Imagined spatial transformations of one's body. *Journal of Experimental Psychology: General, 116*, 172-191.

Parsons, L. M. (1987b). Imagined spatial transformations of one's hands and feet. *Cognitive Psychology, 19*, 178-241.

Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science*, 171:701-703.

Shepard, R. N. & Cooper, L. A. (1982). *Mental images and their transformations*. Cambridge, Mass.: MIT Press.

Schwoebel, J., Friedman, R., Duda, N. & Coslett, H. B. (2001). Pain and the body schema evidence for peripheral effects on mental representations of movement. *Brain*, 124: 2098-2104.

Usoh, M., Arthur, K., Whitton, M. C., Bastos, R., Steed, A., Slater, M., et al. (1999). *Walking, walking-in-place, flying, in virtual environments.* Paper presented at the Proceedings of the 26th annual conference on computer graphics and interactive techniques.

Wexler, M., Kosslyn, S. M., & Berthoz A. (1998). Motor processes in mental rotation. *Cognition, 68*, 77-94.

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin & Review, 9*(4), 625-636.

Predicting Students' Retention of Facts from Feedback during Study

Robert Lindsey (robert.lindsey@colorado.edu)
Department of Computer Science, 430 UCB
University of Colorado, Boulder, CO 80309 USA

Owen Lewis (owen.lewis@colorado.edu)
Department of Applied Mathematics, 526 UCB
University of Colorado, Boulder, CO 80309 USA

Harold Pashler (hpashler@ucsd.edu)
Department of Psychology, 0109
University of California, San Diego, La Jolla, CA 92093 USA

Michael Mozer (mozer@colorado.edu)
Department of Computer Science, 430 UCB
University of Colorado, Boulder, CO 80309 USA

Abstract

Testing students as they study a set of facts is known to enhance their learning (Roediger & Karpicke, 2006). Testing also provides tutoring software with potentially valuable information regarding the extent to which a student has mastered study material. This information, consisting of recall accuracies and response latencies, can in principle be used by tutoring software to provide students with individualized instruction by allocating a student's time to the facts whose further study it predicts would provide greatest benefit. In this paper, we propose and evaluate several algorithms that tackle the benefit-prediction aspect of this goal. Each algorithm is tasked with calculating the likelihood a student will recall facts in the future given recall accuracy and response latencies observed in the past. The disparate algorithms we tried, which range from logistic regression to a Bayesian extension of the ACT-R declarative memory module, proved to all be roughly equivalent in their predictive power. Our modeling work demonstrates that, although response latency is predictive of future test performance, it yields no predictive power beyond that which is held in response accuracy.

Keywords: intelligent tutoring, ACT-R, Bayesian inference, fact learning

Introduction

An effective way to teach facts is to test students while they are studying (Roediger & Karpicke, 2006). For example, if a student is learning the meanings of foreign words, an appropriately designed tutoring system would display a foreign word, ask the student to guess the English translation, and then provide the correct answer. In this work, we consider the case where students undergo several rounds of this type of study. By convention, we refer to the group of rounds as a *study session*. At the end of a study session, students have had several encounters with each item being studied. In addition to promoting robust learning, testing students during study provides valuable information that can in principle be used to infer a student's current and future state of memory for the material. Through the use of a student's performance during study to predict recall at a subsequent test, informed decisions can be made about the degree to which individual facts would benefit from further study. In this paper, we explore algorithms to predict a student's future recall performance on specific facts using both the accuracy of the student's responses during study, and their response latencies—the time it took to produce the responses. In principle, other information is available as well, such as the nature of errors made and the student's willingness to guess a response. However, we restrict ourselves to accuracy and latency data because such data are independent of the domain and the study question format. Thus, we expect that algorithms that base their predictions on accuracy and latency data will be applicable to many domains.

Predicting future recall accuracy from observations during study can be posed as a machine learning problem. Given a group of students for whom we have made observations, we divide the students into "training" and "test" groups. The training group is used to build predictive models whose performance is later evaluated using the test group. We developed several predictive models and describe them later in this paper. Of particular interest is a method we call Bayesian ACT-R (BACT-R). It is based on the declarative memory module of the ACT-R cognitive architecture (Anderson, Byrne, Douglass, Lebiere, & Qin, 2004). The module has equations that interrelate response latency during study, accuracy during study, the time periods separating study sessions from one another and from the test, and the probability of a correct answer at test. However, these equations have a large number of free parameters which makes it challenging to use the model in a truly predictive manner. BACT-R is a method for using Bayesian techniques to infer a distribution over the free parameters, which makes it possible to use the ACT-R equations to predict future recall.

This paper is organized as follows: first, we describe the experiment from which we obtained accuracy and latency data for a group of students studying paired associates. Next, we describe BACT-R and three other models we built to predict student recall in the experiment. Finally, we evaluate and discuss the performance of the algorithms.

Data

Our data are from an unpublished experiment by Pashler, Mozer, and Wixted (unpublished) in which 56 undergraduates tried to learn the disciplines of 60 relatively obscure Nobel prize winners. In an initial pass through the material, subjects were shown the names of the prize-winners paired with their disciplines. Each winner-discipline pair was displayed for five seconds. For each prize winner's name, subjects were given either three or six study opportunities during which they could guess the discipline. For each guess, they received auditory feedback that signaled whether or not the guess was correct. If it was incorrect, the correct answer was displayed on the screen. For these study trials, subjects responded by pressing one of four keys on a keyboard (the experiment involved only three disciplines, and a fourth key indicated "no guess"). During study, both the accuracies and latencies of the subjects' responses were recorded. Two weeks following study, subjects were evaluated in a cumulative test over all the material. The cumulative test was given in the same format as the study trials.

Approaches to Predicting Recall Performance

In our machine learning approach to predicting student recall at test, we split subjects into training and test groups. For both the training and test groups, we gave our algorithms access to response accuracies and latencies obtained during the study session. Additionally, we gave the algorithm access to the response accuracies at the cumulative test session for only the training group. In this section, we describe four increasingly complex algorithms designed to learn from the training group in order to make predictions about the test group.

We use the information from the training subjects to build a model that we apply to the test subjects to predict the probability that they will answer correctly when tested. The model is then evaluated on the test subjects: for each subject s in the test group and item i being learned, we use the model to predict the probability that s correctly recalled i, and compare this prediction to the observed accuracy. In the future, we will refer to s and i as a "subject-item pair."

Because all subjects learned the same set of items, it is possible to use the performance of the training group on a particular item to inform the predicted performance of the test group on this item. We chose to avoid methods that do this because they are restricted to situations where data are available for a large number of subjects learning the same set of items. In principle, the methods we explore here might work even if individuals learned different items chosen from the same domain.

Percentage Classifier

This was the simplest method we examined: given a subject-item pair, the predicted probability of a correct answer at test is the simply the fraction of correct answers given during study. Unlike the other methods we describe in this section, the percentage classifier does not use data from the training subjects — the only information came from the subject's own responses during study.

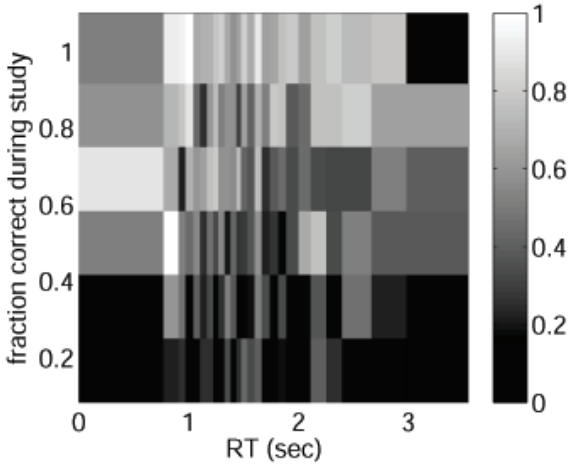


Figure 1: The grid used by the histogram classifier for subject-item pairs that had six study trials. Shading indicates the fraction of those subject-item pairs in the cell that had a correct answer at test. In this figure, the number of bins has been fixed. In practice, it is chosen by cross-validation and is unique to each test subject.

Histogram Classifier

For this method, we specified each subject-item pair by two numbers: the fraction of correct answers during study and the mean latency of the correct answers. We then formed two grids, one for the subject-item pairs that had three trials and another for the pairs that had six. The grids were formed in the following way: one axis had n numbers, such that each interval between two successive numbers contained an equal number of the mean latencies for the training set. n is a parameter of the model and was chosen by cross-validation. The other axis contained either four (for the three-session grid) or seven (for the six-session grid) numbers, such that each interval between two successive numbers contained exactly one of the possible fractions of correct answers. Each training example could then be placed in exactly one of the grid cells. For each cell, we found the number of training examples that fell within the cell and how many of these corresponded to a correct answer at evaluation. This enabled us to find, for each cell, a fraction correct. Given a test subject-item pair, we then found which cell it would fall into based on study performance and predicted that its probability of being correct at evaluation would be that cell's fraction correct. Figure 1 shows the grid for the six-trial case. Note that to display the figure, we had to fix the number of bins. In reality, since this number was chosen by cross-validation, it would be different for each test subject. In the grid shown in the figure, if a subject had a mean RT of 0.5 seconds for their correct answers and answered all study questions correctly, they would fall in the upper left hand cell, and have a predicted probability of future accuracy of about 0.6.

Logistic Regression

Logistic regression is a powerful prediction technique used in statistics and machine learning. In its simplest form, logistic regression takes the values of some number of predictor variables x_i (which may be either binary or continuous) corresponding to an input and then outputs a prediction of the probability that the input belongs to one of two classes. This probability of membership in one of the classes is given by:

$$f(x_1,\ldots,x_n) = \left[1+\exp(-\beta_0 - \sum_{i=1}^{n}\beta_i x_i)\right]^{-1}$$

The weights β_i are to be learned. β_0 is an offset term.

In this application, the predictor variables x_i are the latencies and accuracies obtained during study. More specifically, to predict the probability of a correct response at test for a subject-item pair with three study trials, we use six predictor variables. Three of these are binary and indicate whether each of the three answers given during study were correct or incorrect. The other three variables are the response latencies for the study answers and are therefore continuous. The predictor variables are constructed analogously for the six trial cases. The two classes are "correct answer at test" and "incorrect answer at test."

BACT-R

ACT-R is an influential cognitive architecture whose declarative memory module is often used to model recall follow a series of study sessions (e.g., Pavlik and Anderson (2008)). ACT-R assumes a separate trace is laid down each time an item is studied. Each trace decays according to a power law, t^{-D}, where t is the age of the memory and D is the decay rate. Following N study episodes, the activation for an item combines the trace strength of individual study episodes. It is governed by the equation:

$$A(\mathbf{t},D,B,c) = \log\left(\sum_{j=1}^{N} t_j^{-D}\right) + B + \varepsilon, \qquad \varepsilon \sim f(x;c)$$

where A is activation, B is a base activation level, ε is a noise term drawn from a logistic distribution with mean zero. That is, ε has the density function $f(x;c) = \frac{1}{4c}\operatorname{sech}^2\frac{x}{2c}$, where c is a free parameter. Recall probability is related to activation by:

$$\mathrm{P}(\text{correct recall}\,|\,A;\tau,c) = \left[1+\exp\left(\frac{\tau-A}{c}\right)\right]^{-1}$$

where τ is a free parameter. According to the model, latency (RT) is related to activation by:

$$\mathrm{RT}(A,F,f) = Fe^{-fA}$$

where F and f are free parameters. In total, there are six free parameters whose values we must estimate from the data: D,B,c,τ,F,f. Of these, we assume that c,τ,F,f are to be chosen for each subject-item pair, while the trace decay D and base-level activation term B are fixed for each subject.

For each subject-item pair we have a set of study-trial accuracies and latencies, and we can compute the likelihood of these data for any parameter vector. To do this, we plug the parameters into the equations to generate predictions for study trials and then compare these predictions to actual results of the study trials. More explicitly, we do likelihood-weighted sampling. For a given test subject, we take n_S samples from prior distributions of the six parameters. For each item, we compute the likelihood L of each set of parameters that have been generated. The final prediction of the probability of a correct answer at test is then:

$$\hat{P} = \sum_{i=1}^{n_S} P([D,B,c,\tau,F,f]_i)\ \frac{L([D,B,c,\tau,F,f]_i)}{\sum_{j=1}^{n_S} L([D,B,c,\tau,F,f]_j)}$$

where $\hat{P}$ is the prediction. The likelihood of a set of parameters with respect to a given subject-item pair is given by the product of its likelihood on each study trial:

$$\mathrm{L}(D,B,c,\tau,F,f) = \prod_{i=1}^{n_{trials}} l^i_{acc} l^i_{RT},$$

where i runs over study trials, and l_{acc} and l_{RT} denote the contribution to the likelihood of the accuracy and response latency. The l_{acc}

$$l^i_{acc} = \begin{cases} P(\text{correct recall}|\hat{A};\tau,c) & \text{if response } i \text{ is accurate} \\ 1-P(\text{correct recall}|\hat{A};\tau,c) & \text{otherwise} \end{cases}$$

Here, $\hat{A} = A(\mathbf{t},D,B,c)$.

$$l^i_{RT} = \begin{cases} \frac{1}{4c}\operatorname{sech}^2\frac{\hat{\varepsilon}}{2c} & \text{if response } i \text{ is accurate} \\ 1 & \text{otherwise} \end{cases}$$

where $\hat{\varepsilon} = \log\left(\frac{\mathrm{RT}^i}{\mathrm{RT}(\hat{A},F,f)}\right)$ and RT^i is the observed latency on the ith study trial. The intuition is that for a given set of parameters, we calculate how much noise would be necessary for these parameters to produce the observed latency and then take the likelihood to be the probability of observing this noise level. We used 250 samples for likelihood-weighted sampling. We found that increasing this number did not noticeably improve performance. One implementation detail should be noted: since the interval between study and test is so much larger than the interval between study sessions, we followed Pavlik and Anderson (2008) and compressed the interval between study and test into what they call "psychological time" via a small multiplicative factor.

To define priors for the six parameters, we use the fact that the framework above allows us to find, for each subject, maximum likelihood estimates for the parameter values. We do this for a group of training subjects and compile the results in a histogram. We then fit the results for each parameter to a probability distribution which is then that parameter's prior. In practice, the optimization routine we used to do the likelihood maximization did not converge for all subjects. The

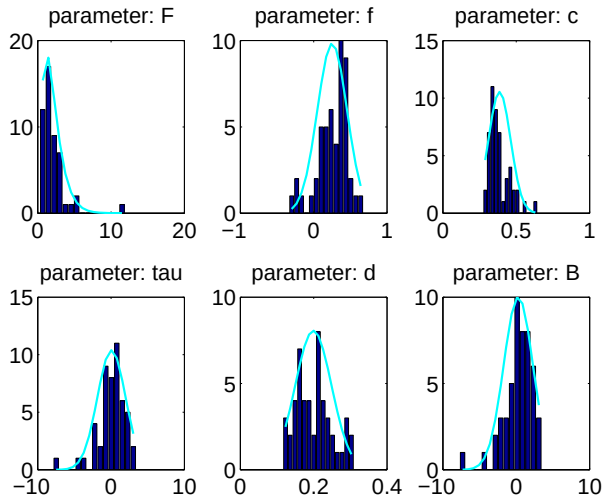


Figure 2: A set priors used in BACT-R. To set these priors, we find the maximum likelihood parameter values for each of the subjects in the training group, compile these estimates into histograms, and then fit the data for each parameter to a continuous probability distribution.

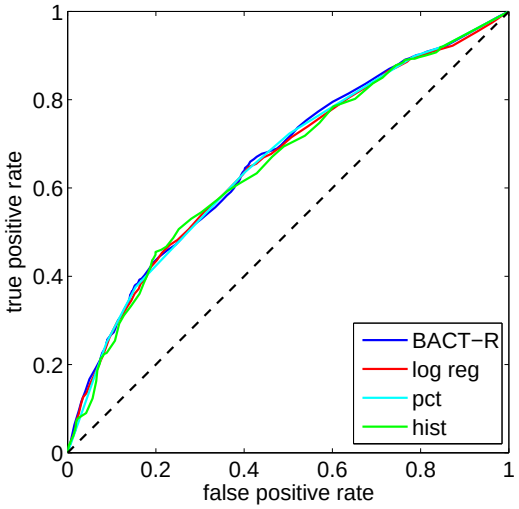


Figure 3: ROC curves for the methods we tried. A comparison shows that all methods perform similarly.

subjects for which it failed to converge were left out of the calculation of the prior. A set of priors, together with the histograms used to define them, are shown in Figure 2. This figure shows that the histograms were generally sharply peaked.

Results

To evaluate the different methods we tried, we used leave-one-out cross-validation. Each subject in turn was held out as a test subject and a prediction for that subject was made by models trained on all the other subjects. This prediction takes the form of a probability between zero and one. Because the data with which we have to compare these predictions are binary — a subject's response is either correct or incorrect — we thresholded the probability so that the predictions also become binary. After thresholding, the models' predictions are either true positive, false positive, true negative, or false negative. Adjusting the threshold changes the number of predictions that fall into each of these categories. In Figures 3-8 (to be described shortly), we summarize the threshold manipulation with an ROC curve, which plots the false positive rate versus the true positive rate for various thresholds. If the ROC curve falls exactly on the dashed diagonal line in the figures, then the method achieves results equivalent to chance prediction. In general, the more bowed the ROC curve, the better the performance of the model.

Comparison of methods

The results obtained by the various methods we tried are shown in Figure 3. As this figure shows, all the methods performed almost equally well. In particular, BACT-R did not outperform other methods we tried. It is interesting to note that this implies that the *order* of correct and incorrect responses, which is information to which BACT-R had access and the percentage classifier did not, seems not to have enabled BACT-R to outperform the percentage classifier. Of course, this does not necessarily mean that there is no useful information contained in the order data.

Relative Importance of Latency and Accuracy Information

We next examined how much information, if any, is contained in the latency data. Our findings are mixed. On the one hand, logistic regression and BACT-R performed just as well with the latency information removed as with it included (see Figures 4 and 5, respectively). On the other hand, when provided only with latency information, logistic regression yielded results significantly better than chance (Figure 4).

We also examined the weights given by logistic regression to latency and accuracy features. The inputs to logistic regression are normalized so that it is meaningful to compare the magnitudes of these weights. The mean magnitudes of the weights for accuracy and latency data are 0.3884 and 0.0751, respectively. The mean weight for the latencies is considerably smaller than the mean weight for the accuracies; it is not negligible. Thus, there is information in the latencies, but it is to a large extent redundant with the information from the accuracies.

The fact that latency information does not improve the performance of our methods may shed some light on our methods performing equivalently: no method took advantage of the latency information; all the information present in the accuracy information reduced to the percentage correct during study. Therefore, all methods did almost exactly as well as the percentage classifier.

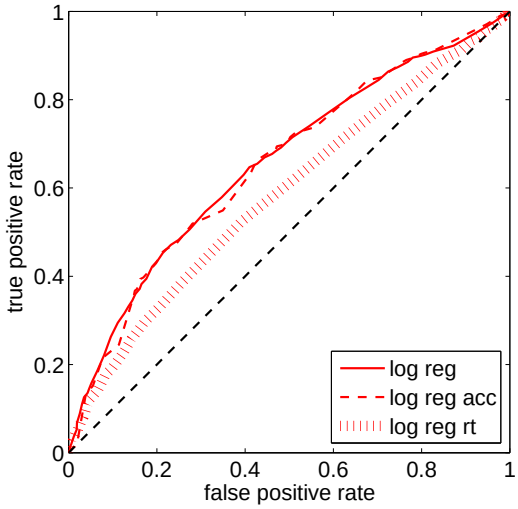


Figure 4: ROC curves for logistic regression, when the model was trained with all available data ("log reg"), only accuracy data ("log reg acc"), and only latency data ("log reg RT"). Removing the latency information does not degrade logistic regression's performance. However, using only the latency information gives results that are significantly better than random. We conclude that the latencies contain information, but that this information is redundant with the accuracy information, and does not help with classification.

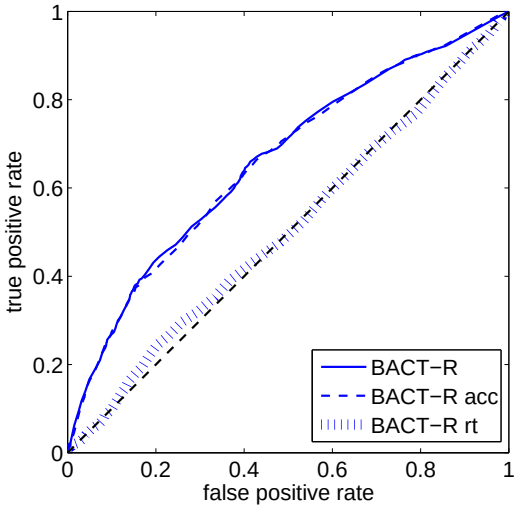


Figure 5: ROC curves for BACT-R when the method uses all available data, only the accuracy data, and only latency data. As with logistic regression (Figure 4), removing latencies does not noticeably hurt the performance of BACT-R. Using only latencies with BACT-R gives worse performance than it does with logistic regression.

Number of Study Trials

Figure 6 shows the performance of BACT-R when restricted to only the three- or the six-trial study conditions.

As expected, BACT-R performed better with six trials than with three, but the difference is not drastic. This is significant because it rules out the possibility that the BACT-R's performance was being dragged down by the three-session cases.

Another experiment we did involved applying logistic regression to only the first study session. In general, we have data from either three or six study trials for each subject-item pair. For this experiment, we used only the first of these. Apart from this, logistic regression was applied in the same way as before. The motivation for this experiment was the hypothesis that even if the accuracy information dominated the latency information when we used all the trials available, perhaps it would contribute more if we used only one trial. In fact, this was what we observed, as is shown in Figure 7, which indicates that, in the one-trial case, adding the latency information to the accuracy information gives a substantial improvement in performance. In addition, we see that it is possible to get reasonably good predictive performance even when we use information from only one trial.

Effect of Priors on BACT-R

In order to examine how much information was contained in the priors we used for BACT-R, we tried replacing the priors chosen by maximum likelihood with uniform priors having mean zero and length four, values that were chosen heuristically based on Figure 2. As Figure 8 shows, the results were noticeably worse than the results obtained with the maximum likelihood priors. This is a validation of the Bayesian approach, since it shows that the performance of the model was due, at least in part, to the knowledge contained in the prior distributions used for the parameters.

Variants

In addition to the methods described above, we tried several variants. For example, we tried replacing raw latencies with z-scores and including latencies from incorrect trials. We also tried assigning greater weight to information from later trials, since these were closer to the test time. No variant we tried significantly altered the performance of the models.

Discussion

Testing students as they study facts is known to be better than just having them reread the facts (Roediger & Karpicke, 2006). Testing has a side benefit: it produces feedback from the student which potentially could inform an intelligent tutoring system about how well the student has learned the facts. In this work, we described an experiment in which feedback was collected from students learning to identify the disciplines of 60 Nobel Prize winners. This feedback took the form of response accuracy and latency during a study session in which each fact was reviewed multiple times. Using data from the study session, we are able to predict memory for individual facts after a two-week retention interval.

We found that latency data alone was predictive. To the best of our knowledge, this finding has not been reported before in modeling literature. However, we also found that

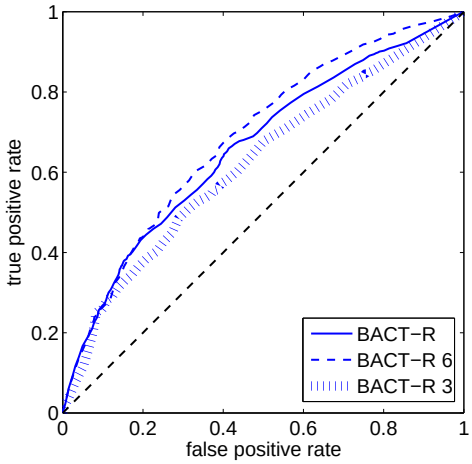


Figure 6: A comparison of the performance of BACT-R on the three-trial and six-trial subject-item pairs. Since six study trials give more feedback than three study trials, we expected BACT-R to perform better for these cases. As the figure shows, this is what we observed. Also as expected, we see that the three-study trial cases gave worse performance. However, BACT-R's performance on the three-study trial cases was not sufficiently degraded to conclude that these trials are responsible for BACT-R's inability to outperform the other methods we studied.

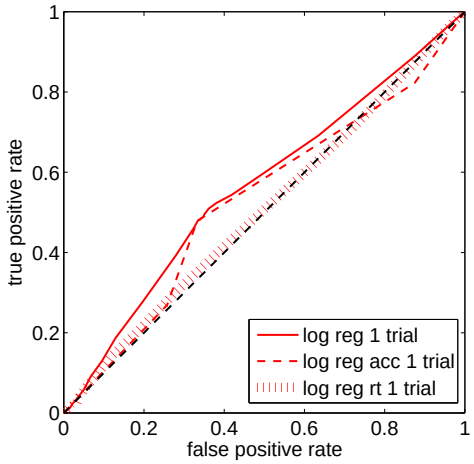


Figure 7: ROC curves for logistic regression when this method was applied to data from only the first study trial for each subject-item pair. If we look at only one study trial, we see that using latency information gives a substantial improvement in performance over the model trained with accuracy data alone. We also observe that, when using both pieces of data, we obtain reasonably good prediction performance, even on the basis of only one study trial.

adding latency data to accuracy data did not improve the performance of our models, suggesting that the latency information was redundant with the accuracy information.

We found that all the predictive models had similar performance, including a model based on ACT-R's declarative memory module, which is one of the best developed and evaluated high-level theories of human memory. Although BACT-R did not outperform other models, we believe that the addition of Bayesian uncertainty integration to the ACT-R framework is a promising idea that should be explored in other contexts. We also believe that the use of latency information for prediction of future recall warrants further study, especially when the feedback data are sparse (e.g., Figure 7, which shows the benefit of latencies when we have feedback from only one trial). Further, it would be interesting to see if the latency information from an experiment specially designed to elicit fast latencies would be more informative than the latencies from this experiment.

In one sense, our conclusions are not astonishing: accuracy of recall during study predicts accuracy of recall at a subsequent test. However, it is important that we have made this intuitively obvious relationship quantitative and that we have explored multiple computational approaches that can exploit the relationship to make concrete predictions of future recall performance.

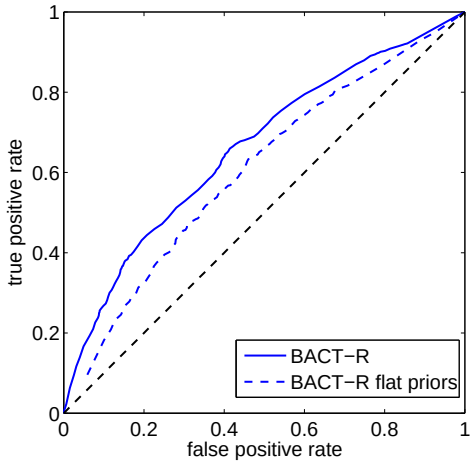


Figure 8: Using the maximum likelihood priors for BACT-R gives substantially better performance than using uniform priors.

References

Anderson, J. R., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review*, *111*(4), 1036-1050.

Pashler, H., Mozer, M., & Wixted, D. (unpublished). *Metrics of forgetting: weakening of associations versus skills.*

Pavlik, P. I., & Anderson, J. R. (2008). Using a model to compute the optimal schedule of practice. *J. Exp. Psych.: Applied*, *14*, 101–117.

Roediger, H., & Karpicke, J. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*, 181–210.

Differences in the Development of Analogy Across Cultures: A Computational Account

Leonidas A.A. Doumas (leonidas@hawaii.edu)
University of Hawaii at Manoa, Department of Psychology
2430 Campus Rd. Honolulu, HI 96822

Robert G. Morrison (rmorrison@luc.edu)
Loyola University Chicago, Psychology Department,
6525 North Sheridan Road Chicago, IL 60626 USA

Lindsey E. Richland (lerich@uci.edu)
University of California, Irvine, Department of Education
2001 Berkeley Place, Irvine, CA 92697-5500

Abstract

Theories of the development of analogical reasoning emphasize either the centrality of relational knowledge accretion or changes in information processing. Recent cross-cultural data collected from children in the United States and China (Richland, Chan, Morrison, & Au, 2010) provides a unique way to test these theories. Here we use simulations in LISA/DORA (Doumas, Hummel, & Sandhofer, 2008; Hummel & Holyoak, 1997, 2003), a neurally-plausible computer model of relational learning and analogical reasoning, to argue that the development of analogical reasoning in children may best be conceptualized as an equilibrium between knowledge accretion and progressive improvement in information processing capability. Thus, improvements in inhibitory control in working memory as children mature enable them to process more relationally complex analogies. At the same time, however, children produce more complex and more accurate analogies in domains in which they have learned richer and more refined representations of relational concepts.

Relational thinking—i.e., thinking based on the roles that objects play rather than the literal features of those things—is a cornerstone of human cognition. It underlies, among many other things, our ability to make analogies, or to appreciate correspondences between domains (e.g., Holyoak & Thagard, 1995).

As with many cognitive processes, our ability to make analogies changes with development. While there is considerable agreement that analogy is a very important process in cognitive development (e.g., Gentner, 2003), there is considerable disagreement as to how the ability to reason analogically develops.

Theories of the Development of Analogical Reasoning

Three primary hypotheses have been put forward to explain age-related differences in analogical reasoning: changes in domain knowledge, a relational shift from object similarity to relational similarity, and increased processing or working memory (WM) capacity.

Goswami and colleagues (Goswami, 1992, 2001; Goswami & Brown, 1989) proposed that the ability to make analogies is present even in early infancy. However, children can only evidence this ability with age and increased knowledge. In other words, the change in children's ability to make analogies is not a function of a developing mechanism, but rather knowledge accretion.

Alternately, Gentner and Rattermann (1991; Rattermann & Gentner, 1998) argued that a domain-specific "relational shift" is responsible for changes in children's analogical abilities. Gentner and Rattermann suggest that as children build knowledge in a particular domain they progress from reasoning about that domain in terms of the perceptual features of objects, to the relations between those objects. For example, 3 year-old children will categorize objects based on overall featural similarity (e.g., they will match apples to red balls rather than bananas), however by age 4 or 5, children will categorize objects based on relational similarity (e.g., matching apples to bananas even in the presence of featural distracters like red balls; Gentner & Namy, 1999). The ability to make analogies based on relational commonalities between domains, therefore, progresses on a domain-by-domain basis with more complex analogies produced in domains in which knowledge is richer.

In contrast to accounts of analogy development based on increases in knowledge, the relational complexity hypothesis of Halford (1993; Andrews & Halford, 2002; Halford et al., 2002) holds that limits in children's WM capacity affects their ability to process relations simultaneously, and therefore their ability to make analogies. According to Halford and colleagues, children can process only specific levels of relational complexity, defined as the number of sources of variation that are related and must be processed together. The simplest level of relational complexity is a binary relation, where only two arguments are sources of variation. The relation, *chase* (dog, cat), for instance, specifies a single relation (*chase*) between two objects (dog, cat). To reason about this relation, a one must keep only the two objects and their relation in mind. A ternary relation (e.g.,

gives-to (boy, girl, book) is more complex, requiring a reasoner to consider three objects and their respective roles. The more complex the relation, the more WM resources are required to process it. As children mature, neural developments leading to increased WM capacities (see, e.g., Diamond, 2002) allow processing situations with greater relational complexity, and, by extension, children are capable of drawing richer and more complex analogies.

Likewise, Richland, Morrison, and Holyoak (2006) proposed that inhibitory control might help to explain the relationship between maturation and the impact of relational complexity on analogical reasoning in young children. While inhibitory control has been a major topic in models of cognitive development (Bjorklund & Harnishbeger, 1990; Diamond, 2002) it has not previously been applied to understanding the development of analogy; however, the hypothesis that inhibitory control is important for the development of analogy is consistent with results from other cognitive tasks (e.g., Diamond, Kirkham & Amso, 2002; Lorsbach & Reimer, 1997; Zelazo et al., 2003).

Multiple Sources in Analogical Development

Richland, Morrison and Holyoak (2006) developed a set of scene analogy problems to investigate relational complexity and featural distraction within a single analogical reasoning task. They found that children from age 3 to 14 steadily improved in their ability to solve more relationally complex problems and resist distraction.

In a follow-up study Richland, Chan, Morrison, and Au (2010) used these same problems with Cantonese speaking 3-4 year old children from Hong Kong. While US children of this age showed main effects of both relational complexity and featural distraction, Chinese children only showed an effect of featural distraction (see Figure 5).

There are several reasons to believe that the Chinese children would score differently on analogical reasoning problems than U.S. children based on their knowledge base and experience with reasoning about relations. Adult studies have shown cultural differences in normative patterns for drawing relational inferences (see Nisbett 2003) such that Chinese and Japanese reasoners may be more attuned to relational correspondences than U.S. participants. These differences also appear cross-culturally in children's socialization and linguistic routines. For example, Asian caregivers use more action oriented language and referential verbs than relatively object-focused U.S. caregivers (e.g., Chinese: (Mandarin) Tardif, 1996; Tardif, Gelman & Xi, 1999; Tardif, Shatz, Naigles, 1997; (Cantonese) Leung, 1998). Chinese children themselves may additionally show a higher relative rate of verb usage in Mandarin (Tardif, 1996; 2006; Tardif, Shatz, & Naigles, 1997; Tardif, Gelman, & Xu, 1999) than U.S. children of comparable ages who show a more pronounced noun bias. In contrast, there is no theoretical reason to expect differences in information processing capacity between the US and Hong Kong (Hedden, et al., 2000).

Accordingly, Richland et al. (2010) reasoned that the US and HK 3-4 year old children each had decreased inhibitory control relative to older children resulting in their distractibility, but that HK children had more sophisticated relational representations than US children resulting in their superior ability to solve more relationally complex problems.

A Computational Account of the Multiple-Source Theory of Analogical Development

Previous Work

Traditionally, researchers have attempted to model the effects of knowledge accretion and increased working memory capacity on analogical development in isolation. For example, Gentner and colleages (e.g., Gentner et al., 1995) used SME (Falkenhainer, Forbus, & Gentner, 1989) to model the relational shift data of Gentner and Rattermann (1991). Gentner et al. captured the differences in analogical reasoning in 4 and 5 year-old children by providing the model with more relational representations at age 5 than at age 4. That is, with limited knowledge of relations, the model behaved like the younger children in Gentner and Rattermann's experiments, making analogies based on over-all perceptual similarity. However, with increased relational knowledge, the model behaved more like the older children, making analogies based on shared relations. Importantly the representations provided to the model had to be hand-coded by the modeler.

More recently, Morrison, Doumas, and Richland (2006), used the LISA model (Hummel & Holyoak, 1997, 2003) in an attempt explain changes in children's analogy making in terms of changes in capacity limits. LISA is a model of analogy-making that relies on time as a signal to bind distributed (i.e., connectionist) representations of objects and relational roles into structured (i.e., symbolic) representations. LISA is powerful, in part, because it benefits from both the flexibility of connectionist approaches and the structure-sensitivity of symbolic approaches (an important property for demonstrating human like relational reasoning; see, e.g., Doumas & Hummel, 2005; Holyoak & Hummel, 2000; Penn, Holyoak, & Povinelli, 2008). In addition, as a consequence of using time to carry binding information, LISA suffers from capacity limitations that mirror those of human WM (Hummel & Holyoak, 2003; Morrison, Doumas, & Richland, 2006; Morrison et al., 2005). LISA relies on lateral inhibition between units to establish the temporal patterns that carry binding information. By decreasing lateral inhibition, LISA's WM is effectively reduced. Morrison et al. (2006), used this property of to capture the pattern of results from Richland et al. (2006).

Approaches using SME and LISA both suffer from limitations, though. First, each approach is based on a

single explanatory variable. As a result, the knowledge accretion approach seems insufficient to explain the results of the Scene Analogy task (see Richland et al., 2006), while the simply changing capacity limits cannot explain the cross-cultural findings of Richland et al. (2010). In addition, both approaches rely on hand-coded relational representations that must be added by the modeler. Neither model makes any claims where these representations, which both models require in order to reason relationally—and that provide the explanatory mechanism in the knowledge accretion case—come from in the first place.

Doumas, Hummel, and Sandhofer (2008) have developed an extension of the LISA model, called DORA (*Discovery of Relations by Analogy*) that learns structured representations of relations from unstructured (i.e., flat feature vector) representations of object properties. DORA provides a means by which the representations used by LISA are learned from examples, and, consequently, provides an opportunity to understand the interplay between the dual sources of knowledge accretion and increasing capacity limits as effectors of the changes in children's analogy making.

The LISA/DORA model

LISA (Hummel & Holyoak, 1997, 2003) is a symbolic-connectionist model of analogy and relational reasoning. DORA (Doumas et al., 2008) is an extension of LISA that learns structured (i.e., symbolic) representations of relations from unstructured inputs. That is, DORA provides an account of how the structured relational representations LISA uses to perform relational reasoning can be learned from examples.

DORA accounts for over 20 phenomena from the literature on relational learning, as well as its development (e.g., Doumas & Hummel, 2010; Doumas et al., 2008). In addition, as DORA learns relational representations, it comes to take LISA as a special case, and can simulate the additional 30+ phenomena in relational thinking simulated by LISA. The description of LISA/DORA that follows is a brief overview due to space constraints. For full details of the models and their operations see Doumas et al. (2008) and Hummel and Holyoak (1997, 2003).

LISAese Representations In LISA (and DORA *after* it has gone through learning) relational structures are represented by a hierarchy of distributed and localist codes (see Figure 1). At the bottom, "semantic" units represent the features of objects and roles in a distributed fashion. At the next level, these distributed representations are connected to localist units (POs) representing individual predicates (or roles) and objects. Localist role-binding units (RBs) link object and predicate units into role-filler binding pairs. At the top of the hierarchy, localist P units link RBs into whole relational propositions.

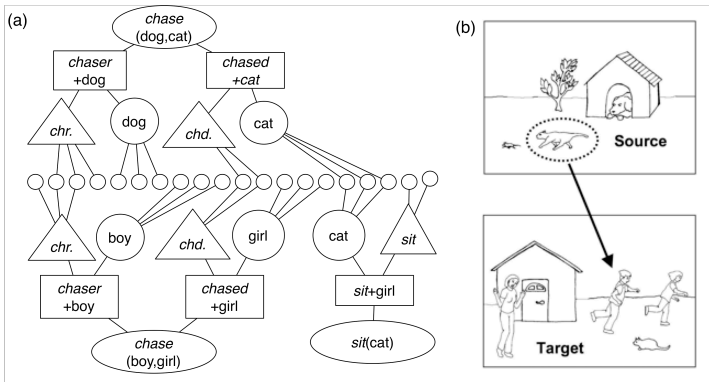


Figure 1. (a) Representation of a LISA/DORA representation like that used to simulate a Scene analogy problem like that in (b). The P (oval), RB (rectangle), and predicate (triangle) units were learned during Simulation Part One. Objects (circles) desribed the objects involved in the Scene Analogy problem. (b) Example of a scene analogy problem from Richland et al., 2006.

Propositions are divided into two mutually exclusive sets: a driver and one or more recipients. In LISA/DORA, the sequence of firing events is controlled by the driver. Specifically, one (or at most three) proposition(s) in the driver becomes active (i.e., enter working memory). When a proposition in the driver becomes active, role-filler bindings must be represented dynamically on the units that maintain role-filler independence (i.e., POs and semantic units; see Hummel & Holyoak, 1997). In LISA binding information is carried by synchrony of firing (with roles firing simultaneously with their fillers). In DORA, binding information is carried by systematic asynchrony of firing, with bound role-filler pairs firing in direct sequence (see Doumas et al., 2008 for details).[1] Activation flows from the driver units to their semantic units. Units in the driver and recipient share the same pool of semantic units. Thus, units in the recipient become active in response to the pattern of activation imposed on the semantic units by the active driver proposition.

Relational Learning Very simply, DORA uses comparison to isolate shared properties of objects and to represent them as explicit structures. DORA starts with simple feature-vector representations of objects (i.e., a node connected to set of features describing that object). When DORA compares one object to another, corresponding features of the two representations fire simultaneously. Any semantic features common to both objects receive twice as much input and thus become roughly twice as active as features connected to one but not the other. By recruiting a new PO unit and learning connections between that unit and active semantics via Hebbian learning (wherein the strength of connections is a function of the units' activation), DORA learns stronger connections between the new PO unit and more active

[1] Asynchrony-based binding allows role and filler to be coded by the same pool of semantic units, which allows DORA to learn representations of relations from representations of objects (Doumas et al., 2008).

semantic units. The new PO thus becomes an explicit representation of the featural overlap of the compared objects. Applied iteratively this process results in explicit and structured representations of object properties and relational roles (see Doumas et al., 2008). Comparison also allows DORA to learn representations of multi-place relations by linking sets of constituent role-filler pairs into relational structures (i.e., to learn the *chases* relation by linking together representations of the roles *chaser* and *chased*; see Doumas et al., 2008 for details).

Mapping For the purposes of analogical mapping, LISA/DORA learns *mapping connections* between units of the same type (e.g., PO, RB, etc.) in the driver and recipient (e.g., between PO units in the driver and PO units in the recipient). These connections grow whenever corresponding units in the driver and recipient are active simultaneously. They permit LISA to learn the correspondences (i.e., mappings) between corresponding structures in separate analogs. They also permit correspondences learned early in mapping to influence the correspondences learned later.

Simulations

Methods

We tested the hypothesis that differences in performance between U.S. and Chinese children were due to differences in relational knowledge. Specifically, we hypothesized that the relational representations of children from Hong Kong were more developed than those of children from the U.S. We used LISA/DORA to test this hypothesis by simulating the results of Richland et al. (2010). The simulation consisted of two complementary parts. In the first part we used DORA to develop representations of relational concepts from examples. We simulated the difference in U.S. and Chinese children by allowing DORA increased learning trials in order to simulate the Chinese children, reflecting the assumption that the experience of children in Hong Kong differs from children in the U.S. We then used the representations that DORA had learned during the first part of the simulation to simulate the Richland et al. (2010) task.

Simulation Part One We used DORA's relational learning algorithm (see Doumas et al., 2008 for details) to develop relational representations from unstructured examples. We started DORA with representations of 100 objects attached to random sets of features (chosen from a pool of 100). We then defined 4 relations (*chase*, *reach-for*, *angry-with*, and *hang-from*). Each relation consisted of two roles, each with three semantic features (e.g., for the *chase* relation, both the roles *chaser* and *chased* were each defined by three specific semantic units). Each of the 100 objects was attached to the features of between 1 and 3 relational roles chosen at random. For example, object1 might be attached to the features for *chaser* (one role of *chases*) and *reaching* (one role of *reach-for*). On each iteration we presented DORA with sets of objects from similar relations, and allowed it to compare the objects and learn from the results (as per DORA's relation learning algorithm). As DORA learned new representations it would also use these representations to make subsequent comparisons. For instance, if DORA learned an explicit representation of the property *chases* (*x*, *y*) by comparing sets objects attached to the roles of *chase* (i.e., *chaser* and *chased*), it could use this new representation for future comparisons. On each trial we selected between 2 and 4 representations and let DORA compare them and learn from the results (i.e., perform predication, and relation learning routines).

We ran 25 sessions each consisting of 800 learning trials During each session, the inhibition parameter was set to a value sampled from a random distribution with a mean of 0.7, and a standard distribution of 0.1. The value of the parameter reflected the reduced WM capacity evidenced in young children (see Morrison et al., 2006).. We measured the quality of the representations DORA had learned during the last 100 trials after each 100 trials. Quality was calculated as the mean of connection weights to relevant features (i.e., those defining a specific transformation or role of a transformation) divided by the mean of all other connection weights + 1 (1 was added to the mean of all other connection weights to normalize the quality measure to between 0 and 1). A higher quality denoted stronger connections to the semantics defining a specific relation relative to all other connections (i.e., a more pristine representation of the relation). Figure 2 shows the quality of the representations DORA learned for each set of 100 comparisons from 100 to 800. As expected, the quality of the representations DORA learns increase as a function of experience (see Doumas et al., 2008 for more details)

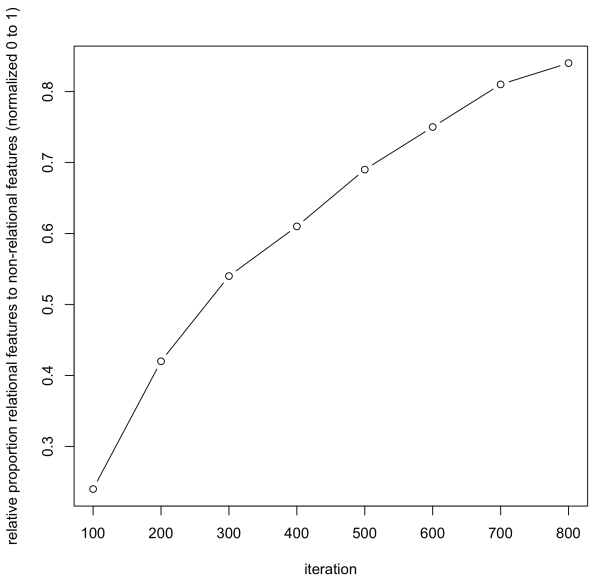


Figure 2. Quality of the representations DORA learned during Simulation Part One.

Simulation Part Two To model the Scene Analogy Problems we used representations of the four problem types (1-relation, no distracter; 1-relation, distracter; 2-relation, no distracter; 2-relation, distracter) composed from the representations DORA had learned during Simulation Part One. For example, to represent the problem from Figure 1, we used a representation of the *chase* relation DORA had learned during Simulation Part One (relational role, RB, and P units) along with object units (e.g, boy and girl) composed of 5 semantic features describing that object (see Figure 1). For 2-Relation problems both relations were represented in LISA's WM together (Hummel & Holyoak, 1997). Vitally, we simulated children from the U.S. by using the representations DORA had learned after only 400 comparisons, and those of the children from Hong Kong using the representations DORA had learned after 600 comparisons.

The laternal inhibition parameter was set exactly as in Simulation Part One. Each simulation run consisted of firing three phase sets in LISA/DORA's working memory, "randomly" assigned by LISA/DORA and allowing LISA/DORA to try to map the representation in the driver to the representation in the recipient. When LISA/DORA failed to determine a stable mapping after firing three phase sets, an answer was selected at random.

Results

The simulation results along with the experimental results from Richland et al. (2010) are presented in Figure 3. LISA/DORA's performance mirrored experimental results for the age groups from both the U.S. and China across conditions.

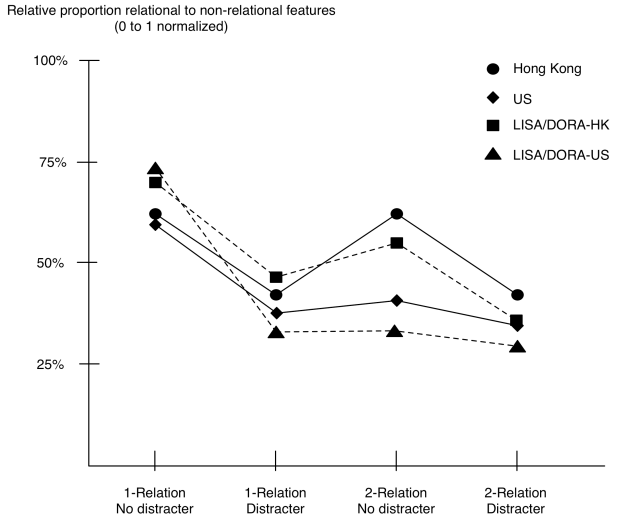


Figure 3: Experimental (Richland et al. 2010) and Simulation,.

General Discussion

In this paper we presented simulations in LISA/DORA that support the hypothesis that both maturation of inhibitory control in working memory and development of knowledge representations is critical for the development of adult-like analogical reasoning. Specifically, we demonstrated that simple changes in inhibition levels in LISA/DORA (i.e., inhibition between elements of competing relational representations in working memory) coupled with DORA's predicate learning routines could account for both relational complexity and featural distraction effects in young children's analogical reasoning performance across cultures. In contrast, approaches based on knowledge accretion and capcity changes in isolation seem unable to capture all of these results.

We conclude that both relational knowledge acquisition and inhibitory control in working memory shape an individual's analogical reasoning performance. We suggest that the development of analogical reasoning in children can be conceptualized as an interaction between these two factors. As children age their knowledge about relations advances while their working-memory capacity as modulated by inhibitory control also improves. At a given time during development, the child is able to perform an analogical task based on both their level of relational knowledge and their working-memory resources. Specifically, the equilibrium operates such that greater relational knowledge can impose fewer processing demands while less knowledge imposes higher demands. Thus, Hong Kong children given the same working-memory resources can better solve relational complex problems. Thus, as relational knowledge increases in a domain, the demands on working memory decline, allowing for more complex reasoning at any given age. This pattern in cognitive development builds on an understanding of working-memory effects in expertise (e.g., Chase & Simon, 1973) where advanced relational knowledge can decrease processing demands and thereby allow experts to accomplish cognitive tasks which novices cannot.

We believe that to truly understand the development of relational reasoning in children, future experimental and computational studies must take into account both advances in relational knowledge and changes in inhibitory control in working memory, and importantly, studying how these two aspects of development interact.

References

Andrews, G. & Halford, G.S. (2002). A cognitive complexity metric applied to cognitive development. *Cognitive Psychology*, *45, 153-219.*

Bjorklund, D. F., & Harnishfeger, K. K. (1990). The resources construct in cognitive development: Diverse sources of evidence and a theory of inefficient inhibition. *Developmental Review, 10,* 48-71.

Chase, W. G., & Simon, H. A. (1973). The mind's eye in chess. In W. G. Chase, (Ed.), *Visual Information* Processing (pp 215–281). New York: Academic Press.

Diamond A. (2002). Normal development of prefrontal cortex from birth to young adulthood: Cognitive functions, anatomy, and biochemistry (pp. 466-503). In D.T. Stuss & R.T. Knight (Eds). *Principles of frontal lobe function*. London: Oxford University Press.

Doumas, L. A. A. & Hummel, J. E. (2010). A computational account of the development of the generalization of shape information. *Cognitive Science, 34*, 698-712.

Doumas. L. A. A., Hummel, J. E., & Sandhofer, C. M. (2008). A theory of the discovery and predication of relational concepts. *Psychological Review, 115,* 1 - 43.

Falkenhainer, B., Forbus, K. D., & Gentner, D. (1989). The structure-mapping engine: Algorithm and examples. Artificial Intelligence, 41, 1-63.

Gentner, D. (2003). Why we're so smart. In D. Gentner and S. Goldin-Meadow (Eds.), *Language in mind: Advances in the study of language and thought*, 95-235. Cambridge, MA: MIT Press.

Gentner, D., & Namy, L. (1999). Comparison in the development of categories. *Cognitive Development, 14*, 487-513.

Gentner, D. , & Rattermann, M. J. (1991). Language and the career of similarity. In S. A. Gelman & J. P.Byrnes (Eds). *Perspectives on thought and language: Interrelations in development (pp. 225-277)*. London, Cambridge University Press.

Gentner, D., Rattermann, M. J., Markman, A. B., & Kotovsky, L. (1995). Two forces in the development of relational similarity. In T. J. Simon & G. S. Halford (Eds.), Developing cognitive competence: New approaches to process modeling (pp. 263-313). Hillsdale, NJ: LEA.

Goswami, U. (1992). *Analogical reasoning in children*. Hillsdale, NJ: Erlbaum.

Goswami, U. (2001). Analogical reasoning in children. In D. Gentner, K. J. Holyoak & B. N. Kokinov (Eds.), *The analogical mind: Perspectives from cognitive science* (pp. 437-470). Cambridge, MA: MIT Press.

Goswami, U., & Brown, A. L. (1989). Melting chocolate and melting snowmen: Analogical reasoning and causal relations. *Cognition, 35*, 69-95.

Halford, G. S. (1993). *Children's understanding: The development of mental models*. Hillsdale, NJ: Erlbaum.

Halford, G. S., Andrews, G., Dalton, C., Boag, C., & Zielinski, T. (2002). Young children's performance on the balance scale: The influence of relational complexity. *Journal of Experimental Child Psychology*, *81*, 383 – 416.

Hedden, T., Park, D. C., Nisbett, R., Ji, L., Jing, Q., & Jiao, S. (2002). Cultural variation in verbal versus spatial neuropsychological function across the life span. Neuropsychology, 16(1), 65-73.

Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. Psychological Review, 104, 427-466.

Hummel, J. E., & Holyoak, K. J. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review*, *110*, 220-264.

Leung, V. W.-Y. (1998). The use of nouns versus verbs in Cantonese-speaking children's early vocabularies and their mothers'speech. PhD dissertation, University of Hong Kong.

Morrison, R.G. (2005). Thinking in working memory. In K. J. Holyoak & R. G. Morrison (Eds.), *Cambridge handbook of thinking and reasoning* (pp. 457-473). Cambridge, UK: Cambridge University Press.

Nisbett, R. E. (2003). *The geography of thought: How Asians and Westerners think differently ... and why*. New York: The Free Press.

Piaget, J., Montangero, J., & Billeter, J. (1977). La formation des correlats. In J. Piaget (Ed.) *Recherches sur l'abstraction reflechissante I* (pp. 115-129). Paris: Presses Universitaires de France.

Rattermann, M.J., Gentner, D (1998) More evidence for a relational shift in the development of analogy: Children's performance on a causal-mapping task *Cognitive Development, 13*, pp. 453-478

Richland L.E., Chan, T-K., Morrison, R.G., & Au, T.K-F. (2010). Young children's analogical reasoning across cultures: Similarities and differences. *Journal of Experimental Child Psychology, 105,* 146-153 .

Richland, L.E., Morrison, R.G., & Holyoak, K.J. (2006). Children's development of analogical reasoning: Insights from scene analogy problems. *Journal of Experimental Child Psychology, 94,* 249–273.

Tardif, T. (1996). Nouns are not always learned before verbs: Evidence from Mandarin speakers' early vocabularies. Developmental Psychology, 32, pp. 492-504.

Tardif, T. (2006). But Are They Really Verbs? Chinese Words for Action. In. K. Hirsh-Pasek, K. Golinkoff, R. Michnick; *Action meets word: How children learn verbs.* New York, NY, US: Oxford University Press, pp. 477-498.

Tardiff, T. Gelman, S. A., Xu, F. (1999). Putting the 'noun bias' in context: A comparison of English and Mandarin. *Child Development, 70*, 620-635.

Tardif, T. Shatz, M. Naigles, L. (1997). Caregiver speech and children's use of nouns versus verbs: A comparison of English, Italian, and Mandarin. *Journal of Child Language,* 24, 535-565.

Discerning Affect in Student Discussions

Jihie Kim, Erin Shaw, Saul Wyner, Taehwan Kim, and Jia Li
(jihie@isi.edu)
University of Southern California/ Information Sciences Institute
Marina del Rey, CA 90292 USA

Abstract

Students' emotions and attitudes are discernible in messages posted to online question and answer boards. Understanding student sentiment may help instructors identify students with potential course issues, optimize help-seeking, and potentially improve student achievement, as well as identify both positive and negative actions by instructors and provide them with valuable feedback. Towards this end, we present a set of context-independent emotion acts that were used by students in a university-level computer science course to express certainty and uncertainty, frustration, and politeness in an online Q&A board and develop viable classification approaches. To explore the potential of sentiment-based profiling, we present a heuristic-driven analysis of thread resolution and detail future research.

Keywords: student online discussions, discourse analysis.

Introduction

Online discussion boards are widely used in higher education, extending the availability of instructors, assistants, and materials to students beyond the traditional classroom. Students use discussion forums to collaborate, exchange information, and seek answers to problems from their instructors and classmates. Discussion board use is generally associated with improved academic performance and greater student satisfaction (Kumrow, 2005; Newman and Schwager, 1995).

Previous work on analyzing student discussions has been based on rhetorical speech acts, course topics, and problem tasks (Kim et al., 2007; McLaren et al., 2007). Classification systems for these features enable researchers to automatically identify student problems. Similarly, understanding student affect may help instructors identify students with potential course issues, optimize help-seeking, and potentially improve student achievement. In addition, by examining the results of different instructor-student interactions in terms of affect, instructors could potentially receive valuable feedback about their online interactions.

In this paper we present a set of dialogue features, or *emotion acts* (EAs), analogous to Speech Acts (Searle, 1969), that characterize student sentiment with respect to 1) frustration and tension, 2) high and low certainty (confidence) and 3) politeness. These sentiments were exhibited in student discourse within a question and answer (Q&A) board in an undergraduate Computer Science course. A discussion corpus consisting of 1,030 student posts was manually tagged with the emotion acts. We describe the first stages of the development of practical classification for emotion acts and explore the potential of sentiment-based student profiling. Specifically, in this paper, we do the following:

1. Identify categories of affect exhibited in an online student discussion in an undergraduate CS course.
2. Examine the frequency of affect in the corpus by gender, role and types of participants.
3. Examine the influence of affect in instructor messages on student responses (discerned by affect).
4. Examine the correlation between affect and type of thread (resolved or unresolved).
5. Illustrate of how emotion acts can be used in assessing and predicting student discussion outcome.
6. Describe our approach to and initial results of automatically classifying three categories of affect.

Identifying Categories of Affect

It is extremely difficult to devise a category of affect labels given the gradations and subtlety of the way feelings and emotions are expressed in language. It is not surprising then that there is no general agreement on how to label affective content and that instead there exist a number of different labeling schemes for different domains (Ordelman and Heylen, 2005). However, previous work suggests that at least some affective content can be identified and selected for, independent of context. For example, acknowledgements are recognizable by the presence of common *politeness* phrases (Brown & Levinson, 1987), and may be used to indicate resolution in Q&A discussion; and *certainty* categorization was shown to assist in distinguishing between editorial and news writing (Rubin et al., 2006), and may be used to distinguish questions and answers by the presence and absence of student confidence.

Identifying a set of categories was an iterative process, and there were three criteria for selection: a) category examples had to be well represented in the corpus, b) researchers had to agree on the categories, and c) categories had to be relevant to student learning. Selection was originally motivated by the desire to identify students' self-efficacy and attitudes, although these categories were too abstract to be practical. We examined discourse that indicated confidence, interest and mastery, and also, urgency, understanding and technicality. Our final categories were *high and low certainty (confidence)*, *tension/frustration*, and *politeness*.

Tension (kappa: 0.74)	**Examples**
Instructor Judgments: Possible student issues with class attendance, judgment or choices	If you really want to do this; I stated in class on at least 2 occasions
Student Judgments: Possible student issues with questioner or target	Result of this sucks; Wow… That was..
Frustration (kappa: 0.92)	**Examples**
Repetitious Actions, Continual Actions: Descriptions of continuous actions without real progress	A lot (15+ times); Never seems to end; High rate of redundancy; Another can of worms
Large Quantities: Descriptions of overwhelming amounts of work and other material	Zillions of references; Super-huge; Simply gargantuan; Monstrous, super-verbose
Difficulty/Impassability, Material Denigration: Statements of explicit difficulty in either solution or understanding of issues, as well as frustration about the material itself	Serious disk quota problems; Severe annoyances; A pain to fix; Makes it really hard
Self-Denigration/Lack of Confidence: Declarations of a personal belief in a lack of ability on the part of the poster	I have spent FAR too long; …I'm stumped; Longer than they should have
High Certainty (kappa: 0.80)	**Examples**
Specificity of Question/Answer: Specific phrasing that concisely explains through examples and pre-conditions	The only way; I found the answer; It only appears
Ease of Understanding/Completeness: Emphasis of the simplicity or completeness of a solution or question	The trick is; Just wait till; Will be simple; All you need to do is
Necessity: Specifically stating that the presented solution is required, or in the case of a question, its importance	Must be able to; Vitally important task; Must have something; You will
Logical Presentation: A method of presenting a proposition, solution, or question that makes it a logical proposal	I assume that; Granted,; Likewise,; On the other hand,
Low Certainty (kappa: 0.95)	**Examples**
Vagueness in Question/Answer: Statements that imply only general or surface understanding of the material at hand by stating personal understandings over factual presentation	What is wrong?; If I understand; Seems to me; Read it somewhere
Lack of Understanding: Statements that clearly state a lack of understanding; differs from other Speech Acts as it implies a continuing lack, rather than an individual issue	I am still confused; Not sure if I understand; I follow most; I'm not sure
Optional Nature: Statements indicating a not strongly recommended or vital issue, solution, or question	Should be compiled from the network directory; You might try; …maybe I'll try making; What is wrong?
Weakened Presentation: Phrases that weaken or justify logical proposal statements	Correct me if I am wrong; Apparently; I am guessing that is the way; As far I know
Politeness categories	**Examples**
Positive (kappa: 0.99): Language strategies used according to formal cultural rules to avoid losing face. Commonly identified as typical polite speech	Thanks; Okay thanks; Good luck with your project
Negative (kappa: 0.99): Dealing with a face-threatening act, by lightening the request or response into a less pressing, informal status.	I was wondering if; Thought I'd throw this out there; Get this cleared up early; Just a head's up,
Bald on record (kappa: 0.84): Dealing with a face-threatening situation by ignoring or emphasizing the consequences of the threat	I question the; don't bzero anything; Change it to this; Do we?
Off record (kappa: 0.82): Attempting to change the request or response into a non-face-threatening statement, i.e., by generalizing a query to a rather than asking for direct help	Has anyone else had this problem; What would do; Asking for answers directly is way easier

Table 1. Categories of affect – description and examples.

The final categories had high agreement among the research team, and thus had potential for use in an automatic classification system.

Annotation Methodology

Annotating affect involved identifying those speech fragments that reliably indicated an identified emotion act in a repeatable fashion throughout the corpus of student discussion board posts. This was complicated by the highly irregular nature of the message content, which was characterized by frequent misspellings and grammar and syntactical errors, stemming from common parlance, simple carelessness, and Computer Science student subculture language use. This necessitated a high level of selectivity and repeatability in all annotations, as well as reliance on specific patterns of distinct phrases and grammar from within the corpus rather than whole statements.

Table 1 lists and describes the final EA categories. A dataset of 1,030 messages in 210 threads from an Operating Systems course was analyzed. Several iterations were performed until we minimized ambiguity among the categories and finalized clear EA definitions. For inter-annotator agreement, we compared two annotators' data on 322 messages in 30 discussion threads. For the current categories, the kappa values are

greater than 0.7. Some Politeness EAs show very high agreement ratios since the annotators consider them very clear and there are only small numbers of cases.

The labeling process consisted of EA classification as well as the marking up of contextual information within the message content. The markup included information about the type of response (question/query or answer/statement) and the role of the author (student, instructor, or TA).

Affect Frequency by Type of Participants

The final frequency distribution of emotion acts for messages posted by different participants within the dataset is shown in Table 2. Of interest are the high occurrences of low certainty and the relatively high frequency of frustration. Female students seemed to present less frustration than male students. Also, females present more positive politeness in their messages. As expected, the instructor's messages show high confidence. Among politeness categories, the instructor presents more bold-on-record politeness (BOR) than students.

We also looked at the presence of emotion acts among high and low frequency contributors. Figure 1 shows the distribution of different emotion acts for seven groups of contributors. As can be predicted from the distribution in Table 2, confidence and polite acts dominate. For the students who post many messages, the number of other emotion acts increases, especially confidence, but also frustration and negative politeness.

Influence of Instructor Affect on Students

The course instructor participated in discussions in many ways; he provided answers directly, gave alternative perspectives, supported student ideas, and elaborated on student answers. It is useful to analyze the influence of the instructor on student dialogue.

In Table 3, we consider what happens when an instructor exhibits emotion. It appears that students tend to express more emotion themselves (certainty, frustration, negative politeness) after an instructor shows emotion. Students appear to express high certainty when they respond to an instructor's high certainty. Similarly, student frustration and low certainty can follow the instructor's expression of low certainty.

While these results show many interesting possible relationships between expressed emotion acts and topic success, the clear and immediate indication shows that emotion acts can show distinctions between different types of posts and threads, which prove their potential usefulness as a profiling mechanism.

Table 2. Distribution of Emotion Acts among participants.

Emotion Act	Percent Emotion Acts found in messages:			
	Total (N=1179)	Male Student (N=782)	Female Student (N=62)	Instruct-or (N=300)
Tension	2%	1%	0%	6%
Frustration	14%	19%	9%	2%
Certainty_High	32%	31%	36%	35%
Certainty_Low	20%	26%	27%	4%
Politeness_Pos	13%	15%	55%	0%
Politeness_Neg	13%	18%	3%	3%
Politeness_OFF	5%	6%	11%	0%
Politeness_BOR	10%	8%	11%	16%

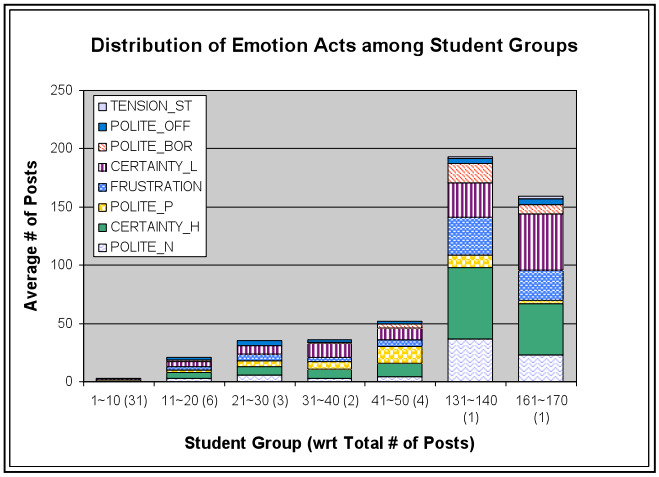


Figure 1: Distribution of Emotion Acts in infrequent and frequent discussion board contributors.

Table 3: Students' EAs following an instructor EAs.

Instructor EAs \ Following Student EAs	#	Tension	Frustration	High_Cert	Low_Cert	Polite_Pos	Polite_Neg	Polite_OFF	Polite_BOR
		0	31	60	44	22	35	6	12
Tension	19	0	2	4	6	1	1	1	2
Frustration	7	0	2	2	1	1	2	0	0
High_Certainty	107	0	16 (15%)	33 (31%)	23 (21%)	5	20(19%)	3	9
Low_Certainty	12	0	5 (42%)	3 (25%)	4 (33%)	2	4 (33%)	0	0
Politeness_Pos	1	0	1	0	1	0	1	0	0
Politeness_Neg	11	0	1	2	1	12	2	1	0
Politeness_OFF	0	0	0	0	0	0	0	0	0
Politeness_BOR	49	0	4	16 (33%)	8	1	5	1	1

Subset for Analysis	(N)	Emotion Act Percentage (%) From Each Group									
		Certainty Level				Frus-tration	Tension	Politeness			
		High	Low	Med	N/A			Bald-On-Record	Positive	Negative	Off-Record
All Posts	(1030)	49.03	13.50	3.79	33.69	24.27	2.72	19.71	17.38	24.66	6.89
All Resolved	(916)	50.11	12.99	3.17	33.73	23.47	2.84	19.43	17.36	23.91	6.33
All Unresolved	(114)	40.35	17.54	8.77	33.33	30.70	1.75	21.93	17.54	30.70	11.40
Resolved Answers	(645)	56.12	10.23	2.95	30.70	18.60	3.72	22.79	11.16	20.31	2.02
Unresolved Answers	(79)	43.04	10.13	8.86	37.97	25.32	1.27	29.11	13.92	32.91	6.33
Resolved Questions	(271)	35.79	19.56	3.69	40.96	35.06	0.74	11.44	32.10	32.47	16.61
Unresolved Questions	(35)	34.29	34.29	8.57	22.86	42.86	2.86	5.71	25.71	25.71	22.86
Resolved Inst. Ans.	(233)	62.66	2.58	0.43	34.33	5.58	8.58	35.19	3.00	7.30	0.43
Unresolved Inst. Ans.	(25)	48.00	0.00	8.00	44.00	0.00	4.00	44.00	8.00	16.00	0.00
Resolved First Posts	(180)	33.89	17.78	2.22	46.11	33.33	0.56	13.89	37.22	32.78	20.00
Unresolved First Posts	(30)	43.33	16.67	3.33	36.67	50.00	0.00	23.33	36.67	40.00	23.33
Resolved Final Posts	(180)	45.56	16.11	4.44	33.89	17.22	2.22	17.22	26.11	24.44	5.00
Unresolved Final Posts	(30)	26.67	13.33	3.33	56.67	20.00	0.00	30.00	6.67	13.33	6.67

Table 5. The distribution of relevant emotion acts in resolved vs. unresolved threads.

Affect Patterns in Threads

While sentiment-based discussion analysis is a significant development, emotion acts represent only the lowest level of potential analysis of student message content. With consistent and functional emotion acts, posters, posts, and entire threads can be analyzed in terms of repeatable EA profiles. As a proof of concept, we wished to develop an independent heuristic to classify threads with a hypothetically robust emotional distinction, and examine the resulting EA profile for such a distinction.

We chose the concept of resolved and unresolved discussion threads, where resolved threads contain a final solution or demonstrable ratification of issues, as well as a beneficial discussion, and open threads are those for which initial questions are not satisfied or which have unresolved issues. The ultimate goal is to identify patterns of affective states that help to discern students who may require further assistance, and topics that may require further clarification. Towards this goal, we experimented with several classification measures based upon observed trends in annotated threads. To fulfill the need for a conclusion, we focused on threads that concluded with an answer, or an acknowledgement of thanks for a provided solution. To ensure a basic level of back-and-forth pedagogic discourse we included only the subset of threads that also contained equal numbers of or more answer/statement posts. The generated results by these criteria were examined by the annotators and found to closely conform to their intrinsic impressions of "resolved" threads. Those threads that were not considered resolved were classified as an "unresolved". This revealed 180 resolved, and 30 unresolved threads.

After this classification, both resolved and unresolved threads were further broken-down into relevant subsets for EA analysis. The analysis was based upon a simple presence test for specific EAs, and the percentage of posts within the subset that contained that emotion act. Certainty, however, as the most common emotion act, was instead calculated as a level, defined by containing over 75% of a specific type of either high or low certainty emotion acts. If the ratio was less than 75%, it was designated as medium certainty. While rudimentary, this examines the potential for more rigorous profiling, by revealing any obvious difference among threads.

The results show a clear distinction between resolved and unresolved threads. Distinctions were noted when there existed at 10% or above difference from resolved vs. unresolved versions of the chosen subset.

Within the certainty measures, high certainty is shown to strongly influence the resolution of a thread with respect to answers, while having little effect on questions. However, in initial posts, high certainty seems to counter-indicate resolution. In contrast, low certainty seems to have minimal effect, except in the case of questions, in which it is strongly represented in unresolved questions. A lack of certainty (both high and low) also strongly differentiates resolved and unresolved questions and initial posts, while it shows the inverse in final posts.

In terms of frustration and politeness, frustration is unsurprisingly well-represented in unresolved posts, though most notably in initial posts. Bald-on-record politeness shows strongly in unresolved instructor answers, original posts, and final posts. Positive politeness is seen greatly in resolved questions and final posts, while negative politeness is greater in resolved final posts. Off-record politeness shows little effect overall.

Automatic Affect Classification

For automatic classification of emotion acts, we followed a similar approach that was previously applied to identify speech acts in student discussions (Kim et al., 2009). We focused on certainty and frustration because they are most relevant to student performance. The annotated discussion threads were first pre-processed: Because student discussions are informal and noisy with respect to grammar, syntax and punctuation, our model fixes common typos, transforms informal words to formal words, and converts apostrophes to their original forms. It replaces some typical words and phrases with fixed keywords; for instance, programming code fragments are replaced with by CODE, and contractions such as "I'm" and "You're" were replaced with "I am" and "You are". The features used include:

F1: Cue phrase and their position in the post
We used n-gram features including unigrams (1 word), bigrams (two word sequence), trigram (three word sequence) and two separate unigrams. We also use position in the post as in the first part, last part or elsewhere. Beginning sentences can have different meanings than those in subsequent sentences. For example, "Thank you" in the beginning sentence position may be an expression of gratitude for previous information, while "thank you" in the last sentence may indicate only politeness.
F2: Message position in the thread: Indicates if the post is the first post, last post or one of the other posts.
F3: The emotion acts of the previous message: EAs in the previous message that the current message is replying to.
F4: Poster class: Defined as either a student or instructor.
F5: Poster change: Checks if the current poster is the same as the previous.
F6: Post length: Categorizes the post as Short (1-5 words), Medium(6-30 words), or Long (>30 words).

Given all combination of features F1-F6, we used Information Gain (Yang and Pedersen, 1997) to prune the feature space and select features. For each Emotion Act, we sorted all features (lexical and non-lexical) by Information Gain and used the top N (=200) features.

We used the Support Vector Machine of Chang and Lin (2001). We did a 5-fold cross validation in the training. RBF (Radial Basis Function) was used as the kernel function. We performed a grid search to get the best parameter (C and gamma) in training and applied them to the test corpus. With the training data of 159 threads and the test data of 52 threads, the initial classification result is shown in Table 4.

The initial results indicate that the EA classification is feasible. Due to the relatively small set size of available manually-annotated training data, the result is not yet at a level where it can be immediately applied in a functional setting. However, we strongly expect these results to improve as more training data becomes available.

	Test Data Results		
Emotion Act	Precision	Recall	F-Score
High Certainty	0.68	0.64	0.65
Low Certainty	0.80	0.83	0.81
Frustration	0.73	0.75	0.73

Table 4. Automatic classification test results for certainty and frustration.

Related Work

Our work builds on prior research on spoken dialogue analysis including dialogue acts (Searle 1969; Hirschberg and Litman 1993; Samuel 2000; Graesser et. al., 2001; Kim et al., 2009), rhetoric analysis (Mann and Thomson, 1988), and surface cue words analysis (Hirschberg and Litman 1993; Samuel 2000). There have also been Dialogue Acts modeling approaches for automatic tagging and recognition of conversational speech (Stolcke et al., 2000) and related work in corpus linguistics where machine learning techniques have been used to find conversational patterns in spoken transcripts of dialogue corpus (Shawar and Atwell, 2005). Although spoken dialogue is different from message-based conversation in online discussion boards, they are closely related to our thread analysis work, and we plan to investigate potential use of conversation patterns in spoken dialogue in threaded discussions.

Carvalho and Cohen (2005) present a dependency-network based collective classification method to classify email speech acts. However, estimated speech act labeling between messages is not sufficient for assessing contributor roles or identifying help needed by the participants. We included other features like participant profiles. Also our corpus consists of less informal student discussions rather than messages among project participants, which tend to be more technically coherent.

Requests and commitments of email exchange are analyzed in (Lampert et al., 2008). As in their analysis, we have a higher kappa value for questions than answers, and some sources of ambiguity in human annotations such as different forms of answers also appear in our data. However, student discussions tend to focus on problem solving rather than task request and commitment as in project management applications, and their data show different types of ambiguity due to the different nature of participant interests.

There has also been work on non-traditional, qualitative assessment of instructional discourse (Boyer et al., 2008; Graesser et al., 2005; McLaren et al., 2007), and results have been used to find features for critical thinking and level of understanding. Similar approaches for classifying speech acts were investigated in Ravi and Kim (2007). This work captures features that are relevant to analyzing

noisy student discussion threads and supports a full automatic analysis of student discussions instead of manual generation of thread analysis rules. Earlier work on annotating emotion in dialogue focused on polarity (positive or negative) and intensity (Craggs and Wood, 2004) but is less useful for analyzing student discussions.

Finally, there have been studies of student affective states in tutorial dialogue, including boredom, confusion, surprise and frustration. These were analyzed and captured using dialogue states with linguistic features such as cohesion measures (D'Mello et al., 2009). Our work focuses on 'threaded' discussions, and is potentially useful for analyzing student discussion outcome.

Summary and Future Work

As the distinctions between resolved and unresolved threads show that profiling and automatic identification by affect is fully possible, it is important to look forward toward methods and directions of higher-level interpretation. The procedure used in investigating closure is only for broad proof-of-concept, rather than developing specific profile criteria for automatic categorization. As such, future development in profiling will require specific categories, defined by interactions within posts between differing affect in a repeatable manner. This can reveal information about important qualities of posts, threads, and students.

We have described an important first step towards the identification and use of emotion acts for instructional analysis of student discussions: We have identified common acts used by students within a course discussion board, developed a promising classification approach, and have shown that these acts are significant within the corpus through an investigation of resolved/unresolved threads. There are many research avenues to explore. In combination with existing metrics based on rhetorical speech acts, contribution quantity and technical depth, the new measures will assist instructors and researchers in understanding how students learn. This study complements prior work on speech acts and discussion topics (Carvalho & Cohen 2005; Feng et al., 2006; Kim et al. 2007).

References

Boyer, K., Phillips, R., Wallis M., Vouk M., Lester, J., Learner Characteristics and Feedback in Tutorial Dialogue (2008), *ACL workshop on Innovative Use of NLP for Building Educational Applications*.

Brown, P. and Levinson, S.C. 1987. *Politeness: Some universals in language usage*. Cambridge: Cambridge University Press.

Carvalho, V.R. and Cohen, W.W., (2005). On the collective classification of email speech acts, *Proceedings of the SIGIR Conference*.

Chang, C. and Lin, C. (2001). *LIBSVM: a library for support vector machines*.

Craggs R., and Wood M. (2004). A Categorical Annotation Scheme for Emotion in the Linguistic Content of Dialogue. *Affective Dialogue Systems*, Elsevier, 89-100.

D'Mello, S., Dowell, N., and Graesser (2009). A.: Cohesion Relationships in Tutorial Dialogue as Predictors of Affective States. *Proceedings of the AI in Education Conference*.

Feng, D., Kim, J., Shaw, E., and Hovy, E. (2006), Towards Modeling Threaded Discussions through Ontology-based Analysis, *Proceedings of National Conference on Artificial Intelligence*.

Graesser, A. C., Olney, A., Ventura, M., Jackson, G. T., (2005). "AutoTutor's Coverage of Expectations during Tutorial Dialogue." *Proceedings of the FLAIRS Conference*.

Graesser, A., VanLehn, K., Rosé, C., Jordan, P., Harter, D. (2001). Intelligent Tutoring Systems with Conversational Dialogue. *AI Magazine*, 22(4).

Hirschberg, J. and Litman, D. (1993). Empirical Studies on the Disambiguation of Cue Phrases, *Computational Linguistics*, 19(3).

Kim, J., Kim, T. and Li, J. (2009). Identifying Student Online Discussions with Unanswered Questions, *Proceedings of the International Conference on Knowledge Capture*.

Kim, J., Shaw, E. Chern, G. and Herbert, R. (2007) Novel tools for assessing student discussions: Modeling threads and participant roles using speech act and course topic analysis, *Proceedings of the AI in Education Conference*.

Kumrow, D. (2005). Student Self-Regulatory Resource Management Strategies and Academic Achievement in a Web-based Hybrid Graduate Nursing Course. *Education and Technology*, Editor(s): T.C. Montgomerie, J.R. Parker, ICET.

Lampert, A., Dale, R., and Paris, C. (2008). The Nature of Requests and Commitments in Email Messages", *AAAI workshop on Enhanced Messaging*.

Mann, W.C. and Thompson, S.A., (1988). Rhetorical structure theory: towards a functional theory of text organization. *Text: An Interdisciplinary Journal for the Study of Text*, 8 (3)

McLaren, B. et al., Using Machine Learning Techniques to Analyze and Support Mediation of Student E−Discussions, *Proceedings of the AI in Education Conference*.

Newman, R. and Schwager, M. (1995). Students' Help Seeking During Problem Solving: Effects of Grade, Goal, and Prior Achievement. *American Educational Research Journal*, v32 n2.

Ordelman, R. and Heylen, D. (2005). *Annotation of Emotions in meetings in the AMI project*.

Ravi, S., Kim, J. (2007) Profiling Student Interactions in Threaded Discussions with Speech Act Classifiers, *Proceedings of the AI in Education Conference*.

Rubin, V., Liddy E., and Kando, N. (2006). Certainty Identification in Texts: Categorization Model and Manual Tagging Results. In Computing Attitude and Affect in Text: Theory and Applications.

Samuel, K., (2000) *An Investigation of Dialogue Act Tagging using Transformation-Based Learning*, PhD Thesis.

Searle, J., (1969). Speech Acts. Cambridge: Cambridge Univ. Press.

Shawar, B. A. and Atwell, E. (2005). Using corpora in machine-learning chatbot systems. *International Journal of Corpus Linguistics*, V10.

Stolcke, A., Coccaro, N. Bates, R., Taylor, P., Van Ess-Dykema, C., Ries, K., Shriberg, E., Jurafsky,D., Martin,R., Meteer,M., (2000). Dialogue act modeling for automatic tagging and recognition of conversational speech, *Computational Linguistics*, v.26 n.3.

Recognizability of Individual Creative Style Within and Across Domains: Preliminary Studies

Liane Gabora (liane.gabora@ubc.ca)
Department of Psychology, University of British Columbia
Okanagan campus, 3333 University Way
Kelowna BC, V1V 1V7, CANADA

Abstract

It is hypothesized that creativity arises from the self-mending capacity of an internal model of the world, or worldview. The uniquely honed worldview of a creative individual results in a distinctive style that is recognizable within and across domains. It is further hypothesized that creativity is domain-general in the sense that there exist multiple avenues by which the distinctiveness of one's worldview can be expressed. These hypotheses were tested using art students and creative writing students. Art students guessed significantly above chance both which painting was done by which of five famous artists, and which artwork was done by which of their peers. Similarly, creative writing students guessed significantly above chance both which passage was written by which of five famous writers, and which passage was written by which of their peers. These findings support the hypothesis that creative style is recognizable. Moreover, creative writing students guessed significantly above chance which of their peers produced particular works of art, supporting the hypothesis that creative style is recognizable not just within but across domains.

Keywords: art; creative writing; creativity; Darwinian theory; expertise; heuristic search; honing; style; voice.

Introduction

The therapeutic nature of the creative process is well known. Eminent creators and laypeople alike often claim that through engagement in creative activities they gain a clearer sense of themselves as unique individuals. By making artistic choices, and observing how these choices guide subsequent thoughts about the work, eventually culminating in original, creative form, they acquire self-knowledge, and often, are left with a sense of completeness. The transformation that occurs on canvas or on the written page is said to be mirrored by a sense of personal transformation and self-discovery that occurs within.

Artists often find a style that feels as if it is 'theirs' only after periods of exploration with different media and established styles and art forms. Similarly, writers speak of transitioning from a stage in which they were merely imitating the styles of authors they admired to a stage in which they discovered their own authentic 'voice'. This sense of self-discovery may seem to the creator as real as anything he or she has ever experienced, and the transition from merely imitating others to finding one's own identifiable style is often evident to anyone exposed to an individual's creative works. But although the phenomenon of recognizable creative style seems obvious to artists themselves, and to those who appreciate what they do, it is not predicted by well-known theories of creativity.

This paper presents the results of preliminary experiments designed to test the hypothesis that creative individuals possess a distinctly recognizable creative style, and that this creative style is recognizable not just within a domain but across domains. We begin by discussing well-known theories of creativity, and how the phenomena of individual style and 'voice' are not predicted by them. Three studies are then presented. The first two studies test the hypothesis that the phenomenon of creative style is real; that is, that creative individuals such as artists and writers genuinely exhibit a creative style that others come to associate with them. The third study tests the hypothesis that an individual's creative style is recognizable not just in one domain, but across different domains. Finally, we discuss how the findings are compatible with a new theory of creativity.

Theories of Creativity

This section very briefly summarizes some leading theories of how the creative process works, and then presents a new theory of creativity referred to as honing theory.

Creativity as Heuristic Search

Inspired by the metaphor of the mind as a computer (or computer program), it was proposed that creativity involves a process of heuristic search, in which rules of thumb guide the inspection of different states within a particular state space (set of possible solutions) until a satisfactory solution is found (Eysenck, 1993; Newell, Shaw & Simon 1957; Newell & Simon 1972). In heuristic search, the relevant variables are defined up front; thus the state space is generally fixed. Examples of heuristics include breaking the problem into sub-problems, hill-climbing (reiteratively modifying the current state to look more like the goal state), and working backward from the goal state to the initial state. A variation on this is the view that creativity involves heuristics that guide the search for, not a possibility within a given state space, but a new state space itself (*e.g.,* Boden, 1990; Kaplan & Simon, 1990, Ohlsson, 1992). That is, it involves switching from one representation of the problem to another, sometimes referred to as *restructuring* (Weisberg, 1995).

The Expertise View of Creativity

Some posit that creativity involves everyday thought processes such as remembering, planning, reasoning, and restructuring; no special or unconscious thought processes need be postulated (Perkins, 1981; Weisberg, 2006). This is sometimes referred to as the *expertise view* of creativity because it stresses the extent to which creative acts draw upon familiarity with a particular domain of knowledge. Thus this view in particular is associated with the notion that creativity is highly domain-specific; expertise in one domain is not viewed as enhancing creativity in another domain. The expertise view is also associated with the notion that the creative process result in products that are largely derivative, or *reproductive* (as Weisberg puts it), as opposed to genuinely new, or *productive*.

The Darwinian Theory of Creativity

Another approach to modeling the creative process involves framing it in Darwinian terms. While some philosophers describe the growth of knowledge as Darwinian merely in the sense that conjectures must be refutable, *i.e.*, able to be selected against (Popper, 1963; Lorenz, 1971), Campbell (1960) goes further, arguing that a stream of creative thought is a Darwinian process. The basic idea is that we generate new ideas through 'blind' variation and selective retention (abbreviated BVSR): 'mutate' the current thought a multitude of different ways, select the fittest variant(s), and repeat this process until a satisfactory idea results. The variants are 'blind' in the sense that the creator has no subjective certainty about whether they are a step in the direction of the final creative product.

Currently the Darwinian view of creativity is most closely associated with Simonton (1998, 1999a,b, 2007a,b), who views creativity as essentially a trial-and-error process in which the most promising 'blindly' generated ideational variants are selected for development into a finished product. It should be noted that the endeavor to apply natural selection to creative thought is not without critics (Dasgupta, 2004; Eysenck, 1995; Gabora, 2005, 2007; Sternberg, 1998, Thagard, 1980; Weisberg, 2000). Nevertheless, the development of a creative idea can be said to be evolutionary in the very general sense that it exhibits descent with adaptive modification.

The Honing Theory of Creativity

Central to the *honing theory* of creativity is the notion of a *worldview,* by which we mean one's internal model of the world, as well as one's values, attitudes, predispositions, and habitual patterns of response (Gabora, 2000, 2004, 2008, Gabora & Aerts, 2009). Honing theory posits that creativity arises due to the *self-organizing, self-mending* nature of a worldview, and that it is by way of the creative process the individual hones (and re-hones) an integrated worldview. Honing theory places equal emphasis on the externally visible creative outcome and the internal cognitive restructuring brought about by the creative process. Indeed one factor that distinguishes it from other theories of creativity is that it focuses on not just restructuring as it pertains to the conception of the task, but as it pertains to the worldview as a whole. When faced with a creatively demanding task, there is an interaction between the conception of the task and the worldview. The conception of the task changes through interaction with the worldview, and the worldview changes through interaction with the task. This interaction is reiterated until the task is complete, at which point not only is the task conceived of differently, but the worldview is subtly or drastically transformed. Thus one distinguishing feature of honing theory is that the creative process reflects the natural tendency of a worldview to seek integration or consistency amongst both its pre-existing and newly-added components, whether they be ideas, attitudes, or bits of knowledge; it mends itself as does a body when injured.

The Recognizability of Creative Style

Theories of creativity based on heuristic search, the acquisition of expertise, or chance, random processes, such as BVSR, give no reason to expect that the act of creation leads to a clearer or more integrated sense of self, or that the works of a particular creator should exhibit a unique and recognizable style. This is particularly so if, as is often claimed, creativity is strongly domain-specific (Baer, 1998; Sawyer, 2006; Weisberg, 2006). If creativity is limited to a particular domain then why should it result in a global sense of wellbeing or integration?

Claims about the domain-specificity of creativity are based largely on findings that correlations amongst alternative measures of creativity are small, and expertise or eminence with respect to one creative endeavor is rarely associated with expertise or eminence with respect to another (e.g. Getzels & Jackson, 1962). Thus, for example, creative scientists rarely become famous artists or dancers. The focus of these studies is squarely on expertise or eminence as evidence of creative achievement. But what if creative achievement is measured not by expertise or eminence but by having found a way to express *what is genuine and unique about us* through whatever media we have at a given time at our disposal? One might expect that an artist's or scientist's personal style comes through in how he or she prepares a meal or decorates a room, what creativity researchers refer to as little-c (Richards, Kinney, Benet, & Merzel, 1988) or mini-c (Beghetto & Kaufman, 2007) creative activities. Findings of domain-specificity in creativity may have more to do with the fact that we focus on creative achievement at a level that takes a decade or more to obtain (Simonton, 2007), as opposed to looking for evidence that creative potential and personal style transcends particular domains. In other words, looking for evidence of exceptional creativity in multiple domains is not the only or necessarily even the best way to address the question of whether creativity is domain-specific. Another way is to look for evidence that an individual exhibits a creative style in one domain that also 'comes through' when engaged in creative activities in other domains.

Although the phenomenon of recognizable style or voice is not predicted by the view that creativity is a matter of heuristic search, expertise, or Darwinian selection, it is predicted by the honing theory of creativity. We have seen that, according to honing theory, creativity is the process by which one hones a worldview, and each idea the creator comes up with is a different expression of the same underlying core network of understandings, beliefs, and attitudes. A worldview has a characteristic structure, and the creator's various outputs are reflections of that structure, and they are related to one another, and potentially pave the way for one another. Thus honing theory predicts that creative individuals have a recognizable style.

There is evidence that human creativity is more consistent with honing theory than with competing theories of creativity with respect to developmental antecedents of creativity, personality traits of creative individuals, and studies of lifespan creativity (Gabora, under revision). This paper reports on the results of creative style experiments that provide further support for the theory. The goal of the first two studies was to find empirical evidence for the common belief that there really *is* such a thing as recognizable style or voice. Although artists have no doubt this is true, it has not been studied by psychologists, and as we have seen, most theories of creativity do not predict it. The goal of the third study was to test a more controversial prediction of honing theory, the prediction that the structure of a worldview manifests in a unique and recognizable way, to varying degrees, through *different* creative outlets. Thus for example, you might recognize someone's art by knowing how they dress or decorate.

Study 1: Within-domain Recognizability of Artistic Style

The first study tested the hypothesis that individuals who are highly familiar with the art of a given artist will recognize other works by that artist that they have not encountered before.

Method

Participants The research was conducted with 10 University of British Columbia undergraduates majoring in art who were highly familiar with five well-known artists, and with each other's art.

Materials and Procedures Prior to the study, participants were instructed to bring from home a recently completed painting that they had never discussed with or shown to any of their classmates. They were asked to hide their signatures or any other identifying feature of the painting. Before the study, the paintings were examined to ensure that signatures and any other identifying features had been covered.

At the beginning of the study, the art students were shown three well-known paintings by each of five well-known artists as a refresher. The well-known artists were Picasso, Monet, Van Gogh, Dahli, and Andy Warhol. These artists were decided upon because previous discussion with the class indicated that all students were highly familiar with them. The students were then shown ten unfamiliar (rare or newly completed) works that they had not studied in class. Signatures on all artworks were covered by black tape. The art students were given a questionnaire and asked to guess which famous artist did each painting. For each answer they were also asked to state how certain they were on a 3-point scale that they had not encountered the work before.

They were also shown the paintings by their fellow classmates that they had never seen before. The rationale for showing classmates' paintings was to control for the possibility that with the well-known artists, a participant who, though not recognizing the creative voice, might guess above chance levels to which era or country the artist belonged. The only sufficiently large number of artists from the same era and locale that the students were familiar with were their own classmates. As with the famous artists, they were asked to guess which classmate did each painting, and to state how certain they were on a 3-point scale that they had not encountered the work before.

The participants were debriefed, and the results were analyzed. If a participant had encountered a work before, or was uncertain about having encountered it before, the score for this question was not included in the analysis. Less than 5% of scores were not included in the analysis.

Analysis The data were analyzed to determine if the participants correctly identified the artists at above-chance levels. First, a proportion correct score for each participant was computed. For example, if a participant correctly identified seven out of 10 possible artists, the proportion correct score for that person was .70. Then, the proportion correct score that would have been obtained on the basis on random guesses for each question was computed. For example, for the well-known artists, since there were 5 of them, the proportion correct based on random guesses was .20. One-sample t-tests were then computed comparing the average proportion correct scores to the proportion correct values that would have been obtained had participants been randomly guessing. A one-sample randomization test (Manly, 2007) was used to compute the p-levels for these t-test values, given the small sample sizes, and .05 was used as the criterion for statistical significance.

Results

The results are divided into two sections: recognition of famous artists, and recognition of classmates' art.

Recognition of Famous Artists For the task in which art students were asked which famous artist painted each painting, the mean proportion correct was .78 (SD = .12). The proportion correct that would have been obtained on the basis of random guesses was .20. This difference is statistically significant, $t(9) = 15.3$, $p < .0001$, r (effect size) = .98. Thus art students were able to distinguish above chance which famous artists created pieces of art they had not seen before.

Recognition of Classmates' Art A similar result was obtained with works of art by the students themselves. The mean proportion correct was .74 (SD = .29). The proportion correct that would have been obtained on the basis of random guesses is .11. This difference is significant, $t(9) = 6.8$, $p < 0.0001$, r = .92. Thus art students also correctly identified their classmates' art above chance.

Study 2: Within-domain Recognizability of the Notion of a Writer's 'Voice'

This study tested the hypothesis that individuals who are highly familiar with the work of a given writer will recognize other works by that writer that they have not encountered before.

Method

Participants The research was conducted with seven University of British Columbia advanced undergraduate creative writing students who were highly familiar with five well-known writers, and with each other's writing.

Materials and Procedures The analogous procedure to that described above for art students was used for creative writing students. Prior to the study, they had been asked to write a passage about a kitchen and a poem about a month of the year. They were explicitly asked to include no immediately identifying content in their writing (*e.g.,* no mention of surfing if it is known that they like surfing). These constituted their two pieces of writing. At the beginning of the study they were given three well-known written passages by each of ten well-known writers as a refresher. The well-known writers were Ernest Hemingway, Douglas Coupland, Emily Dickinson, Walt Whitman, Allen Ginsburg, Jack Kerouac, TS Eliot, Jane Austin, George Orwell, and Franz Kafka. These writers were chosen because previous discussion with the class indicated that all students were highly familiar with them. A sample of one of the written passages by well-known writers (in this case, Ernest Hemingway) that were provided to creative writing students as a refresher is provided in Table 1.

Table 1: Sample of written passage by well-known writer provided to creative writing students as a refresher.

> "If the book is good, it is about something that you know, and is truly written, and reading it over you see that this is so, you can let the boys yip and the noise will have that pleasant sound coyotes make on a very cold night when they are out in the snow and you are in your own cabin that you have built or paid for with your work."

The creative writing students were then shown twenty rare passages that they had not studied in class. A sample of one of the passages by well-known writers (in this case, Jane Austin) is provided in Table 2.

Table 2: Sample of written passage by well-known writer provided to creative writing students as a test of their ability to recognize writer's style.

> "However, here they are, safe and well, just like their own nice selves, Fanny looking as neat and white this morning as possible, and dear Charles all affectionate, placid, quiet, cheerful, good humour. They are both looking very well, but poor little Cassy is grown extremely thin, and looks poorly. I hope a week's country air and exercise may do her good. I am sorry to say it can be but a week. The baby does not appear so large in proportion as she was, nor quite so pretty, but I have seen very little of her. Cassy was too tired and bewildered just at first to seem to know anybody. We met them in the hall -- the women and girl part of us -- but before we reached the library she kissed me very affectionately, and has since seemed to recollect me in the same way."

The creative writing students were also given the two pieces of writing by each of their fellow classmates (the passage about a kitchen and the poem about a month of the year) that they had never seen before. They were given a questionnaire, and asked to guess which famous writer wrote each passage in the first set of passages, and which classmate wrote each passage in the second set. For each answer, they were also asked to state on a 3-point scale how certain they were that they had not encountered the work before.

Participants were debriefed, and the results were analyzed. As in the first study, if the participant had encountered the work before, or was uncertain about having encountered it before, the score for this question was not included in the analysis. Once again, less than 5% of scores were not included in the analysis.

Results

The results are divided into two sections: recognition of famous writers, and recognition of classmates' writing.

Recognition of Famous Writers For creative writing students exposed to passages by famous writers, the mean proportion correct was .34, (SD = .28). The proportion correct that would have been obtained on the basis of random guesses is .10. This difference is significant, $t(7) = 7.0$, $p < 0.0001$, r = .94. Thus creative writing students correctly identified above chance passages by famous writers that they had not encountered before.

Recognition of Classmates' Writing A similar but less pronounced result was obtained with passages written by the students themselves. The mean proportion correct was .27 (SD = .16). The proportion correct that would have been obtained on the basis of random guesses is .14. This difference is significant, $t(7) = 2.3$, $p < 0.05$, r = .66. Thus, creative writing students also correctly identified above chance passages written by classmates.

Study 3: Cross-domain Recognizability of Style

This experiment tested the hypothesis that familiarity with an individual's creative work in one domain facilitates recognition of that individual's creative work in another.

Method

Participants The same seven University of British Columbia advanced undergraduate creative writing students who participated in Study 2 also participated in Study 3. They were highly familiar with each other's writing, but unfamiliar with each other's art.

Materials Each creative writing student brought one piece of covered art to the professor's office. They were asked to hide their signature and any other identifying feature. Before the study, the paintings were examined to ensure that signatures and other identifying features had been hidden.

Procedure The students were shown unsigned art done by classmates. They were given a questionnaire and asked to guess which classmate did which piece of art. As above, for each answer they were also asked to state on a scale of 1-3 how certain they were that they had not encountered the work before. If they had seen the piece before, or thought they might have seen it before, their answer was not included in the analysis. Less than 5% of scores were discarded from the analysis.

Results

The mean proportion correct was .39 (SD = .27). The proportion correct that would have been obtained on the basis of random guesses is .17. This difference is significant, $t(6) = 2.2$, $p < 0.03$, $r = .67$. Thus creative writing students were able to identify above chance which of their classmates created a given work in a domain *other* than writing, specifically art.

Discussion and Conclusions

The experiments with artists and writers reported here provide support for the hypothesis that different works by the same creator exhibit a recognizable style or 'voice', and that this recognizable quality even comes through in different creative outlets. Art students were able to distinguish significantly above chance which famous artists created pieces of art they had not seen before. They also correctly identified their classmates' art significantly above chance. Similarly, creative writing students correctly identified significantly above chance passages by famous writers that they had not encountered before, and correctly identified their classmates' writing significantly above chance. Creative writing students additionally correctly identified significantly above chance works of art produced by classmates. (The opposite study, determining whether art students correctly identify written passages generated by their classmates, has not yet been carried out.)

The higher recognizability of artistic style (study 1) than writer's style (study 2) comes as a surprise. It cannot be entirely due to the famous artists coming from a wider range of eras and locales than the famous writers, for if that were the correct explanation, the recognizability of classmates' art in Study 1 and classmates' writing in Study 2 should have been comparable. Perhaps there are fewer constraints on artists, *i.e*. fewer demands that the work 'make sense', and it need not exhibit plot structure or character development. Thus there may be more acceptable ways of 'doing one's own thing'. The analysis takes into account that there were twice as many writers to choose from as artists, but in future studies the number of famous artists and writers will be the same, in order to make the studies as comparable as possible.

The results support the hypothesis that creators have a recognizable style. These findings are not predicted by theories of creativity that emphasize chance processes or the accumulation of expertise. If creative output is a matter of chance or the acquisition of expertise, then what is the source of this identifiable personal style? These findings are, however, predicted by honing theory, according to which personal style reflects the uniquely honed structure of an individual's worldview. The finding that creative writing students were able to identify above chance which of their classmates created a given work in a domain *other* than writing, specifically art, supports the prediction that creators hone a uniquely structured worldview that exhibits a style that is recognizable not just within a domain but across domains. Further experiments are underway to replicate these findings with larger groups of participants, and adapt the general procedure to study the recognizability of style within and across domains using trained jazz musicians.

It is worth pointing out how this approach, in particular the investigation of recognizable cross-domain style, differs from typical attempts to determine to what extent higher cognition is domain-general versus domain-specific. As mentioned in the introduction, it is commonly assumed that this issue can be resolved by determining to what extent ratings of expertise in one domain are correlated with ratings of expertise in another. An unspoken assumption here is that ratings of expertise are all that is needed to detect any quality that might characterize or unify an individual's creative or intellectual ventures, and indeed that outputs of higher cognitive processes are objectively comparable. In reality, while manifestations of higher cognition are *sometimes* comparable, even quantitatively, often there is little objective basis for comparison. The present results suggest that higher cognition is domain general not in the sense that expertise in one enterprise guarantees expertise in another, but in the sense that there are multiple interacting venues for creative exploration and self-expression open to an individual, and through which that individual's worldview may be gleaned. It may be that our potential for cross-domain learning is only just beginning to be exploited, through ventures such as the Learning through the Arts program in Canada, in which

students, for example, learn mathematics through dance, or learn about food chains through the creation of visual art. It seems reasonable that if knowledge is *presented* in compartmentalized chunks, students end up with a compartmentalized understanding of the world, while if knowledge were presented more holistically, a more integrated kind understanding may be possible.

Acknowledgments

This work is funded by grants from the Social Sciences and Humanities Research Council of Canada (SSHRC) and the GOA program of the Free University of Brussels. The author wishes to thank Brian O'Connor for insightful discussion and statistical help.

References

Baer, J. (1998). The case for domain specificity in creativity. *Creativity Research Journal, 11,* 173–177.

Beghetto, R. A., & Kaufman, J. C. (2007). Toward a broader conception of creativity: A case for mini-c creativity. *Psychology of Aesthetics, Cre- ativity, and the Arts, 1,* 73–79.

Campbell, D. T. (1960). Blind variation and selective retention in creative thought as in other knowledge processes. *Psychological Review, 67,* 380–400.

Dasgupta, S. (2004). Is creativity a Darwinian process? *Creativity Research Journal*, *16*, 403–413.

Eysenck, H. J. (1993). Creativity and personality: Suggestions for a theory. *Psychological Inquiry, 4,* 147-178.

Eysenck, H. J. (1995). *Genius: The natural history of creativity*. Cambridge, England: Cambridge University Press.

Gabora, L. (2000). Conceptual closure: Weaving memories into an interconnected worldview. In (G. Van de Vijver & J. Chandler, Eds.) Closure: Emergent Organizations and their Dynamics. *Annals of the New York Academy of Sciences, 901,* 42–53.

Gabora, L. (2004). Ideas are not replicators but minds are. *Biology & Philosophy*, *19*(1), 127-143.

Gabora, L. (2007). Why the creative process is not Darwinian. Commentary on 'The creative process in Picasso's Guernica sketches: Monotonic improvements versus nonmonotonic variants' by D. K. Simonton. *Creativity Research Journal, 19*(4), 361-365.

Gabora, L. (2008). The cultural evolution of socially situated cognition. *Cognitive Systems Research, 9*(1-2), 104-113.

Gabora, L. & Aerts, D. (2009). A model of the emergence and evolution of integrated worldviews. *Journal of Mathematical Psychology*. 53, 434-451.

Getzels, J. W. & Jackson, P. W. (1962). *Creativity and Intelligence*. New York: John Wiley.

Kaplan, C. A. & Simon, H. A. (1990) In search of insight. *Cognitive Psychology, 22,* 374–419.

Newell, A., Shaw, C. & Simon, H. (1957). The process of creative thinking. In H. E. Gruber, G. Terrell & M. Wertheimer (Eds.) Contemporary approaches to creative thinking (pp. 153-189). New York: Pergamon.

Newell, A. & Simon, H. (1972). *Human problem solving*. Edgewood Cliffs NJ: Prentice-Hall.

Ohlsson, S. (1992). Information-processing explanations of insight and related phenomena. In M. T. Keane & K. J. Gilhooly (Eds.) *Advances in the psychology of thinking* (Vol. 1, pp. 1–44). New York: Harvester Wheatsheaf.

Perkins, D. N. (1981). *The mind's best work*. Cambridge MA: Harvard University Press.

Richards, R., Kinney, D. K., Benet, M., & Merzel, A. P. C. (1988). Assessing everyday creativity: Characteristics of the Lifetime Creativity Scales and validation with three large samples. *Journal of Personality and Social Psychology, 54,* 476–485.

Sawyer, R. K. (2006). *Explaining creativity: The science of human innovation.* New York: Oxford University Press.

Simonton, D. K. (1998). Donald Campbell's model of the creative process: Creativity as blind variation and selective retention. *Journal of Creative Behavior, 32*(3), 153–158.

Simonton, D. K. (1999a). Creativity as blind variation and selective retention: Is the creative process Darwinian? *Psychological Inquiry*, *10*, 309–328.

Simonton, D. K. (1999b). *Origins of genius: Darwinian perspectives on creativity*. New York: Oxford University Press.

Simonton, D. K. (2007). The creative process in Picasso's Guernica sketches: Monotonic improvements versus nonmonotonic variants. *Creativity Research Journal, 19*(4), 329–344.

Simonton, D. K. (2007b). Rejoinder: Picasso's Guernica creativity as a Darwinian process: Definitions, clarifications, misconceptions, and applications. *Creativity Research Journal*, *19*(4), 381–394.

Sternberg, R. J. (1998). Cognitive mechanisms in human creativity: Is variation blind or sighted? *Journal of Creative Behavior*, *32*, 159–176.

Thagard, P. (1980). Against evolutionary epistomology. In P. D. Asquith & R. N. Giere (Eds.) *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, pp. 187-96.

Weisberg, R. W. (1999). Creativity and knowledge: a challenge to theories. In R. J. Sternberg (Ed.) *Handbook of creativity* (pp. 226-259). Cambridge, England: Cambridge University Press.

Weisberg, R. W. (2000). An edifice built on sand? Review of Origins of Genius: Darwinian Perspectives on Creativity, D. K. Simonton. *Contemporary Psychology: APA Review of Books, 45,* 589-593.

Weisberg, R. W. (2003). Creativity: Understanding innovation in problem solving, science, invention, and the arts. Hoboken, NJ: John Wiley.

Weisberg, R. W. (2006). *Creativity: Understanding innovation in problem solving, science, invention, and the arts.* Hoboken, NJ: Wiley and Son.

Computational Modeling of Emotional Content in Music

Kristine Monteith (kristinemonteith@gmail.com)

Tony Martinez (martinez@cs.byu.edu)

Dan Ventura (ventura@cs.byu.edu)

Department of Computer Science
Brigham Young University
Provo, UT 84602

Abstract

We present a system designed to model characteristics which contribute to the emotional content of music. It creates *n*-gram models, Hidden Markov Models, and entropy-based models from corpora of musical selections representing various emotions. These models can be used both to identify emotional content and generate pieces representative of a target emotion. According to survey results, generated selections were able to communicate a desired emotion as effectively as human-generated compositions.

Keywords: Music cognition; computational modeling; learning; music composition.

Introduction

Music and emotion are intrinsically linked; music is able to express emotions that cannot adequately be expressed by words alone. Often, there is strong consensus among listeners as to what type of emotion is being expressed in a particular piece (Gabrielsson & Lindstrom, 2001; Juslin, 2001). There is even some evidence to suggest that some perceptions of emotion in music may be innate. For example, selections sharing some acoustical properties of fear vocalizations, such as sudden onset, high pitch, and strong energy in the high frequency range, often provoke physiological defense responses (Ohman, 1988). Researchers have demonstrated similar low-level detection mechanisms for both pleasantness and novelty. (Scherer, 1984, 1988). There also appears to be some inborn preference for consonance over dissonance. In studies with infants, researchers found that their subjects looked significantly longer at the source of sound and were less likely to squirm and fret when presented with consonant as opposed to dissonant versions of a melody (Zentner & Kagan, 1996).

There are a variety of theories as to what aspects of music are most responsible for eliciting emotional responses. Meyer theorizes that meaning in music comes from following or deviating from an expected structure (Meyer, 1956). Sloboda emphasizes the importance of associations in the perception of emotion in music and gives particular emphasis to association with lyrics as a source for emotional meaning (Sloboda, 1985). Kivy argues for the importance of cultural factors in understanding emotion and music, proposing that the "emotive life" of a culture plays a major role in the emotions that members of that culture will detect in their music (Kivy, 1980). Tolbert proposes that children learn to associate emotion with music in much the same way that they learn to associate emotions with various facial expressions (Tolbert, 2001). Scherer presents a framework for formally describing the emotional effects of music and then outlines factors that contribute to these emotions, including structural, performance, listener, and contextual features (Scherer, 2001).

In this paper, we focus on some of the structural aspects of music and the manner in which they contribute to emotions in music. We present a cognitive model of characteristics of music responsible for human perception of emotional content. Our model is both discriminative and generative; it is capable of detecting a variety of emotions in musical selections, and also of producing music targeted to a specific emotion.

Related Work

A number of researchers have addressed the task of modeling musical structure for the purposes of building a generative musical system. Conklin summarizes a number of statistical models which can be used for music generation, including random walk, Hidden Markov Models, stochastic sampling, and pattern-based sampling (Conklin, 2003). These approaches can be seen in a number of different studies. For example, Hidden Markov Models have been used to harmonize melodies, considering melodic notes as observed events and a chord progression as a series of hidden states (Allan & Williams, 2005). Similarly, Markov chains have been used to harmonize given melody lines, focusing on harmonization in a given style in addition to finding highly probable chords (Chuan & Chew, 2007).

Wiggins, Pearce, and Mullensiefen present a system designed to model factors such as pitch expectancy and melodic segmentation. They also demonstrate that their system can successfully generate music in a given style (Wiggins, Pearce, & Mullensiefen, 2009). Systems have also been developed to produce compositions with targeted emotional content. Delgado, Fajardo, and Molina-Solana use a rule-based system to generate compositions according to a specified mood (Delgado, Fajardo, & Molina-Solana, 2009). Rutherford and Wiggins analyze the features that contribute to the emotion of fear in a musical selection and present a system that allows for an input parameter that determines the level of "scariness" in the piece (Rutherford & Wiggins, 2003). Oliveira and Cardoso describe a wide array of features that contribute to emotional content in music and present a system that uses this

information to select and transform chunks of music in accordance with a target emotion (Oliveira & Cardoso, 2007). The authors have also developed a system that addresses the task of composing music with a specified emotional content (Monteith, Martinez, & Ventura, 2010). In this paper, we illustrate how our system can be interpreted as a cognitive model of human perception of emotional content in music.

Methodology

The proposed system constructs statistical and entropic models for various emotions based on corpora of human-labeled musical data. Analysis of these models provides insights as to why certain music evokes certain emotions. The models supply localized information about intervals and chords that are more common to music conveying a specific emotion. They also supply information about what overall melodic characteristics contribute to emotional content. To validate our findings, we generate a number of musical selections and ask research subjects to label the emotional content of the generated music. Similar experiments are conducted with human-generated music commissioned for the project. We then observe the correlations between subject responses and our predictions of emotional content.

Initial experiments focus on the six basic emotions outlined by Parrott (Parrott, 2001)—love, joy, surprise, anger, sadness, and fear—creating a data set representative of each. A separate set of musical selections is compiled for each of the emotions studied. Selections for the training corpora are taken from movie soundtracks due to the wide emotional range present in this genre of music. MIDI files used in the experiments can be found at the Free MIDI File Database.[1] These MIDI files were rated by a group of research subjects. Each selection was rated by at least six subjects, and selections rated by over 80% of subjects as representative of a given emotion were then selected for use in the training corpora. Selections used for these experiments are shown in Figure 1.

Next, the system analyzes the selections to create statistical models of the data in the six corpora. Selections are first transposed into the same key. Melodies are then analyzed and *n*-gram models are generated representing what notes are most likely to follow a given series of notes in a given corpus. Statistics describing the probability of a melody note given a chord, and the probability of a chord given the previous chord, are collected for each of the six corpora. Information is also gathered about the rhythms, the accompaniment patterns, and the instrumentation present in the songs.

The system also makes use of decision trees constructed to model the characteristics that contribute to emotional content. These trees are constructed using the C4.5 algorithm (Quinlan, 1993), an extension of the ID3 algorithm (Quinlan, 1986) that allows for real-valued attributes. The decision tree classifiers classifiers allow for a more global analysis of generated melodies. Inputs to these classifiers are the default features extracted by the "Phrase Analysis" component of the freely available jMusic software.[2] This component returns a vector of twenty-one statistics describing a given melody, including factors such as number of consecutive identical pitches, number of distinct rhythmic values, tonal deviation, and key-centeredness. These statistics are calculated for both the major and minor scales.

A separate set of classifiers is developed to evaluate both generated rhythms and generated pitches. The first classifier in each set is trained using analyzed selections in the target corpus as positive training instances and analyzed selections from the other corpora as negative instances. This is intended to help the system distinguish selections containing the desired emotion. The second classifier in each set is trained with melodies from all corpora versus melodies previously generated by the algorithm, allowing the system to learn melodic characteristics of selections which have already been

Love:
Advance to the Rear
Bridges of Madison County
Casablanca
Dr. Zhivago
Legends of the Fall
Out of Africa

Surprise:
Addams Family
Austin Powers
Batman
Dueling Banjos
George of the Jungle
Nightmare Before Christmas
Pink Panther
The Entertainer
Toy Story
Willie Wonka

Fear:
Axel's Theme
Beetlejuice
Edward Scissorhands
Jaws
Mission Impossible
Phantom of the Opera
Psycho
Star Wars: Duel of fhe Fates
X-Files: The Movie

Joy:
1941
633 Squadron
Baby Elephant Walk
Chariots of Fire
Flashdance
Footloose
Jurassic Park
Mrs. Robinson
That Thing You Do
You're the One that I Want

Anger:
Gonna Fly Now
James Bond
Mission Impossible
Phantom of the Opera
Shaft

Sadness:
Forrest Gump
Good Bad Ugly
Rainman
Romeo and Juliet
Schindler's List

Figure 1: Selections used in training corpora for the six different emotions considered.

[1] http://themes.mididb.com/movies/

[2] http://jmusic.ci.qut.edu.au/

accepted by human audiences.

For the generative portion of the model, the system employs four different components: a Rhythm Generator, a Pitch Generator, a Chord Generator, and an Accompaniment and Instrumentation Planner. The functions of these components are explained in more detail in the following sections.

Rhythm Generator

The rhythm for the selection with a desired emotional content is generated by selecting a phrase from a randomly chosen selection in the corresponding data set. The rhythmic phrase is then altered by selecting and modifying a random number of measures. The musical forms of all the selections in the corpus are analyzed, and a form for the new selection is drawn from a distribution representing these forms. For example, a very simple AAAA form, where each of four successive phrases contains notes with the same rhythm values, tends to be very common. Each new rhythmic phrase is analyzed by jMusic and then provided as input to the rhythm evaluators. Generated phrases are only accepted if they are classified positively by both classifiers.

Pitch Generator

Once the rhythm is determined, pitches are selected for the melodic line. These pitches are drawn according to the *n*-gram model constructed from melody lines of the corpus with the desired emotion. A melody is initialized with a series of random notes, selected from a distribution that models notes most likely to begin musical selections in the given corpus. Additional notes in the melodic sequence are randomly selected based on a probability distribution of note mosts likely to follow the given series of *n* notes.

For example, with the "joy" corpus, the note sequence (C4, D4, E4) has a 0.667 probability of being followed by an F4, a 0.167 probability of being followed by a D4, and a 0.167 probability of being followed by a C4. If these three notes were to appear in succession in a generated selection, the system would have a 0.167 probability of selecting a C4 as the next note.

The system generates several hundred possible series of pitches for each rhythmic phrase. As with the rhythmic component, features are then extracted from these melodies using jMusic and provided as inputs to the pitch evaluators. Generated melodies are only selected if they are classified positively by both classifiers.

Chord Generator

The underlying harmony is determined using a Hidden Markov Model, with pitches considered as observed events and the chord progression as the underlying state sequence (Rabiner, 1989). The Hidden Markov Model requires two conditional probability distributions: the probability of a melody note given a chord and the probability of a chord given the previous chord. The statistics for these probability distributions are gathered from the corpus of music representing the desired emotion.

For example, C4 is most likely to be accompanied by a C major chord, and F4 is most likely to be accompanied by a G7 chord in selections from the "love" corpus (probabilies of 0.099 and 0.061, respectively). In the "sadness" corpus, C4 is most likely to be accompanied by a C minor chord (probability of 0.060). As examples from the second set of distributions, the G7 chord is most likely to be followed by the G7 or the C major chord in selections from the "love" corpus (both have a probability of 0.105). In selections from the "sadness" corpus, the G7 chord is most likely to be followed by the G7 or the C minor chord (probabilities of 0.274 and 0.094 respectively).

The system then calculates which set of chords is most likely given the melody notes and the two conditional probability distributions. Since many of the songs in the training corpora had only one chord present per measure, initial attempts at harmonization also make this assumption, considering only downbeats as observed events in the model.

Accompaniment and Instrumentation Planner

The accompaniment patterns for each of the selections in the various corpora are categorized, and the accompaniment pattern for a generated selection is probabilistically selected from the patterns of the target corpus. Common accompaniment patterns included arpeggios, block chords sounding on repeated rhythmic patterns, and a low base note followed by chords on non-downbeats.

For example, arpeggios are a common accompaniment pattern in the corpus of selections expressing the emotion of "love." Two of the selections in the corpus feature simple, arpeggiated chords as the predominant theme in their accompaniments, and two more selections have an accompaniment pattern that feature arpeggiated chords played by one instrument and block chords played by a different instrument. The remaining two selections in the corpus feature an accompaniment pattern of a low base note followed by chords on non-downbeats. When a new selection is generated by the system, one of these three patterns is selected with equal likelihood to be the accompaniment for the new selection.

Instruments for the melody and harmonic accompaniment are also probabilistically selected based on the frequency of various melody and harmony instruments in the corpus. For example, melody instruments for selections in the "surprise" corpus include acoustic grand piano, electric piano, and piccolo. Harmony instruments include trumpet, trombone, acoustic grand piano, and acoustic bass.

Evaluation

In order to verify that our system was accurately modeling characteristics contributing to emotional content, we presented our generated selections to research subjects and asked them to identify the emotions present. Forty-eight subjects, ages 18 to 55, participated in this study. Six selections were generated in each category, and each selection was played for four subjects. Subjects were given the list of emotions and asked to circle all emotions that were represented in each

song. Each selection was also played for four subjects who had not seen the list of emotions. These subjects were asked to write down any emotions they thought were present in the music without any suggestions of emotional categories on the part of the researchers. Reported results represent percentages of the twenty-four responses in each category. To provide a baseline, two members of the campus songwriting club were also asked to perform the same task: compose a musical selection representative of one of six given emotions. Each composer provided selections for three of the emotional categories. These selections were evaluated in the same manner as the computer-generated selections, with four subjects listening to each selection for each type of requested response. Reported results represent percentages of the four responses in each category.

Results

Figure 2 outlines the characteristics identified by the decision trees as being responsible for emotional content. For example, if a piece had a Dissonance measure over 0.107 and a Repeated Pitch Density measure over 0.188, it was classified in the "anger" category. Informally, angry selections tend to be dissonant and have many repeated notes. Similar information was collected for each of the different emotions. Selections expressing "love" tend to have lower repeated pitch density and fewer repeated patterns of three, indicating these selections tend to be more "flowing." Joyful selections have some stepwise movement in a major scale and tend to have a strong climax at the end. The category of "surprise" appears to be the least cohesive; it requires the most complex set of rules for determining membership in the category. However, repeated pitch patterns of four are present in all the surprising selections, as is a lack of stepwise movement in the major scale. Not surprisingly, selections expressing "sadness" adhere to a minor scale and tend to have a downward trend in pitch. Fearful selections deviate from the major scale, do not always compensate for leaps, and have an upward pitch direction. Downward melodic trends do not deviate as much from the major scale. Our model appears to be learning to detect the melodic minor scale; melodies moving downward in this scale will have a raised sixth and seventh tone, so they differ in only one tone from a major scale.

Tables 1 and 2 report results for the constrained response surveys. Row labels indicate the corpus used to generate a given selection, and column labels indicate the emotion identified by survey respondents. Based on the results in Table 1, our system is successful at modeling and generating music with targeted emotional content. For all of the emotional categories but "surprise," a majority of people identified the emotion when presented with a list of six emotions. In all cases, the target emotion ranked highest or second highest in terms of the percentage of survey respondents identifying that emotion as present in the computer-generated songs. As a general rule, people were more likely to select the categories of "joy" or "sadness" than some of the other emotions, perhaps

---

Love:
RepeatedPitchDensity <= 0.146
- RepeatedPitchPatternsOfThree <= 0.433: Yes
- RepeatedPitchPatternsOfThree > 0.433: No
RepeatedPitchDensity > 0.146: No

Joy:
PitchMovementByTonalStep <= 0.287: No
PitchMovementByTonalStep > 0.287
- ClimaxPosition <= 0.968
- - ClimaxTonality <= 0: No
- - ClimaxTonality > 0
- - - PitchMovementByTonalStep(Minor) <= 0.535: No
- - - PitchMovementByTonalStep(Minor) > 0.535: Yes
- ClimaxPosition > 0.968: Yes

Surprise:
RepeatedPitchPatternsOfFour <= 0.376: No
RepeatedPitchPatternsOfFour > 0.376
- PitchMovementByTonalStep (Minor) <= 0.550
- - ClimaxPosition <= 0.836: Yes
- - ClimaxPosition > 0.836
- - - LeapCompensation <= 0.704: No
- - - LeapCompensation > 0.704
- - - - KeyCenteredness <= 0.366: No
- - - - KeyCenteredness > 0.366: Yes
- PitchMovementByTonalStep(Minor) > 0.550: No

Anger:
Dissonance <= 0.107: No
Dissonance > 0.107
- RepeatedPitchDensity <= 0.188: No
- RepeatedPitchDensity > 0.188: Yes

Sadness:
TonalDeviation(Minor) <= 0.100
- OverallPitchDirection <= 0.500: Yes
- OverallPitchDirection > 0.500: No
TonalDeviation (Minor) > 0.100: No

Fear:
TonalDeviation <= 0.232: No
TonalDeviation > 0.232
- LeapCompensation <= 0.835
- - OverallPitchDirection <= 0.506
- - - TonalDeviation <= 0.290: Yes
- - - TonalDeviation > 0.290: No
- - OverallPitchDirection > Yes
- LeapCompensation > 0.835: No

Figure 2: Decision tree models of characteristics contributing to emotional content in music.

because music in western culture is traditionally divided up into categories of major and minor. A higher percentage of people identified "joy" in songs designed to express "love" or "surprise" than identified the target emotion. "Fear" was also a commonly selected category. More people identified angry songs as fearful, perhaps due to the sheer amount of scary-movie soundtracks in existence. Themes from "Jaws," "Twilight Zone," or "Beethoven's Fifth Symphony" readily come to mind as appropriate music to accompany frightening situations; thinking of an iconic song in the "anger" category is more of a challenging task. Averaging over all categories, 57.67% of respondents correctly identified the target emotion in computer-generated songs, while only 33.33% of respondents did so for the human-generated songs.

For the open-ended questions, responses were evaluated by similarity to Parrott's expanded hierarchy of emotions. Each of the six emotions can be broken down into a number of secondary emotions, which can in turn be subdivided into tertiary emotions. If a word in the subject's response matched any form of one of these primary, secondary, or tertiary emotions, it was categorized as the primary emotion of the set. Results are reported in Tables 3 and 4. Again, row labels indicate the corpus used to generate a given selection, and column labels indicate the emotion identified by survey respondents.

The target emotion also ranked highest or second highest in terms of the percentage of survey respondents identifying that emotion as present in the computer-generated songs for the open-ended response surveys. Without being prompted or limited to specific categories, and with a rather conservative method of classifying subject response, listeners were still often able to detect the original intended emotion. Once again, the computer-generated songs appear to be slightly more emotionally communicative. 21.67% of respondents correctly identified the target emotion in computer-generated songs in these open-ended surveys, while only 16.67% of respondents did so for human-generated songs.

Listeners cited "fondness," "amorousness," and in one rather specific case, "unrequited love," as emotions present in selections from the "love" category. One listener said it sounded like "I just beat the game." Another mentioned "talking to Grandpa" as a situation the selection called to mind. Reported descriptions of selections in the "joy" category most closely matched Parrott's terms. These included words such as "happiness," "triumph," "excitement", and "joviality." Selections were also described as "adventurous" and "playful."

None of the songs in the category of "surprise" were described using Parrott's terms. However, this is not entirely unexpected considering the fact that Parrott lists a single secondary emotion and three tertiary emotions for this category. By comparison, the category of joy has six secondary emotions and 34 tertiary emotions. The general sentiment of "surprise" still appears to be present in the responses. One listener reported that the selection sounded like an ice cream truck. Another said it sounded like being literally drunken with happiness. "Playfulness," "childishness," and "curiosity" were also used to describe the selections.

Angry songs were often described using Parrott's terms of "annoyance" and "agitation." Other words used to describe angry songs included "uneasy," "insistent," and "grim." Descriptions for songs in the "sad" category ranged from "pensive" and "antsy" to "deep abiding sorrow." A few listeners described a possible situation instead of an emotion: "being somewhere I should not be" or "watching a dog get hit by a car." Fearful songs were described with words such as "tension," "angst," and "foreboding." "Hopelessness" and even "homesickness" were also mentioned.

Table 1: Emotional Content of Computer-Generated Music. Percentage of survey respondents who identified a given emotion for selections generated in each of the six categories. Row labels indicate the corpus used to generate a given selection, and column labels indicate the emotion identified by survey respondents.

	love	joy	surprise	anger	sadness	fear
love	58%	75%	12%	4%	21%	0%
joy	58%	88%	25%	0%	4%	0%
surprise	4%	54%	38%	0%	12%	8%
anger	4%	04%	46%	50%	17%	88%
sadness	0%	8%	25%	42%	62%	58%
fear	17%	21%	29%	12%	67%	50%

Table 2: Emotional Content of Human-Generated Music.

	love	joy	surprise	anger	sadness	fear
love	50%	0%	25%	25%	100%	0%
joy	100%	25%	0%	0%	75%	0%
surprise	0%	0%	50%	75%	50%	50%
anger	25%	25%	0%	25%	50%	50%
sadness	75%	25%	25%	25%	0%	25%
fear	50%	0%	0%	0%	100%	50%

Conclusion

Pearce, Meredith, and Wiggins (Pearce, Meredith, & Wiggins, 2002) suggest that music generation systems concerned with the computational modeling of music cognition be evaluated both by their behavior during the composition process and by the music they produce. Our system is able to successfully develop cognitive models and use these models to effectively generate music. Just as humans listen to and study the works of previous composers before creating their own compositions, our system learns from its exposure to emotion-labeled musical data. Without being given a set of preprogrammed rules, the system is able to develop internal mod-

Table 3: Emotional Content of Computer-Generated Music: Unconstrained Responses.

	love	joy	surprise	anger	sadness	fear
love	21%	25%	0%	0%	0%	0%
joy	0%	58%	0%	4%	0%	0%
surprise	0%	12%	0%	8%	0%	0%
anger	0%	8%	0%	17%	0%	25%
sadness	4%	0%	0%	4%	17%	17%
fear	0%	8%	0%	12%	17%	17%

Table 4: Emotional Content of Human-Generated Music: Unconstrained Responses.

	love	joy	surprise	anger	sadness	fear
love	0%	25%	0%	0%	0%	0%
joy	0%	25%	0%	0%	0%	0%
surprise	0%	0%	0%	0%	25%	0%
anger	0%	0%	0%	0%	25%	0%
sadness	0%	0%	0%	0%	25%	0%
fear	0%	0%	0%	25%	25%	50%

els of musical structure and characteristics that contribute to emotional content. These models are used both to generate musical selections and to evaluate them before they are output to the listener. The quality of these models is evidenced by the system's ability to produce songs with recognizable emotional content. Results from both constrained and unconstrained surveys demonstrate that the system can accomplish this task as effectively as human composers.

Acknowledgments

This work is partially supported by the National Science Foundation under Grant No. IIS-0856089.

References

Allan, M., & Williams, C. K. I. (2005). Harmonising chorales by probabilistic inference. *Advances in Neural Information Processing Systems*, *17*, 25-32.

Chuan, C., & Chew, E. (2007). A hybrid system for automatic generation of style-specific accompaniment. *Proceedings International Joint Workshop on Computational Creativity*, 57-64.

Conklin, D. (2003). Music generation from statistical models. *Proceedings AISB Symposium on Artificial Intelligence and Creativity in the Arts and Sciences*, 30-35.

Delgado, M., Fajardo, W., & Molina-Solana, M. (2009). Inmamusys: Intelligent multi-agent music system. *Expert Systems with Applications*, *36*(3-1), 4574-4580.

Gabrielsson, A., & Lindstrom, E. (2001). The influence of musical structure on emotional expression. *Music and Emotion: Theory and Research*, 223-248.

Juslin, P. N. (2001). Communicating emotion in music performance: A review and a theoretical framework. *Music and Emotion: Theory and Research*, 223248.

Kivy, P. (1980). *The corded shell: Reflections on musical expression*. Princeton, NJ: Princeton University Press.

Meyer, L. (1956). *Emotion and meaning in music*. Chicago: Chicago University Press.

Monteith, K., Martinez, T., & Ventura, D. (2010). Automatic generation of music for inducing emotive response. *Proceedings of the International Conference on Computational Creativity*, 140-149.

Ohman, A. (1988). Preattentive processes in the generation of emotions. *Cognitive perspectives on emotion and motivation*, 127-144.

Oliveira, A., & Cardoso, A. (2007). Towards affective-psychophysiological foundations for music production. *Affective Computing and Intelligent Interaction*, 511522.

Parrott, W. G. (2001). *Emotions in social psychology*. Philadelphia: Psychology Press.

Pearce, M. T., Meredith, D., & Wiggins, G. A. (2002). Motivations and methodologies for automation of the compositional process. *Musicae Scientiae*, *6*(2).

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*(1), 81-106.

Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. San Mateo, CA: Morgan Kaufman.

Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proc. IEEE*, *77*(2), 257-285.

Rutherford, J., & Wiggins, G. (2003). An experiment in the automatic creation of music which has specific emotional content. *Proceedings of MOSART, Workshop on current research directions in Computer Music*, 35-40.

Scherer, K. R. (1984). On the nature and function of emotion: A component process approach. *Approaches to emotion*, 293-317.

Scherer, K. R. (1988). On the symbolic functions of vocal affect expression. *Journal of Language and Social Psychology*, *7*, 79-100.

Scherer, K. R. (2001). Emotional effects of music: Production rules. *Music and Emotion: Theory and Research*, 223248.

Sloboda, J. (1985). *The musical mind: The cognitive psychology of music*. Oxford: Oxford University Press.

Tolbert, E. (2001). Music and meaning: An evolutionary story. *Psychology of Music*, *24*, 103-130.

Wiggins, G. A., Pearce, M. T., & Mullensiefen, D. (2009). Computational modelling of music cognition and musical creativity. *Oxford Handbook of Computer Music and Digital Sound Culture*.

Zentner, M., & Kagan, J. (1996). Perception of music by infants. *Nature*, *383*(29).

Cross-Situational Statistical Learning: Implicit or Intentional?

George Kachergis, Chen Yu, and Richard M. Shiffrin
{gkacherg, chenyu, shiffrin}@indiana.edu
Department of Psychological & Brain Science / Cognitive Science Program
Bloomington, IN 47405 USA

Abstract

For decades, implicit learning researchers have examined a variety of cognitive tasks in which humans seem to automatically extract structure from the environment. Similarly, statistical learning studies have shown that humans can use repeated co-occurrence of words and referents to build lexicons from individually ambiguous experiences (Yu & Smith, 2007). In light of this, the goal of the present paper is to investigate whether adult cross-situational learners require an explicit effort to learn word-object mappings, or if it may take place incidentally, requiring merely attention to the audiovisual stimuli. In two implicit learning experiments with incidental tasks that direct participants' attention to different aspects of the stimuli, we found evidence of learning, suggesting that cross-situational learning mechanisms can be incidental without explicit intention. However, learning was superior under explicit study instructions, indicating that strategic inference may also play a role.

Keywords: implicit learning; language acquisition; cross-situational statistical learning; automaticity

Introduction

Humans have a remarkable capacity to adapt to the regularities in our environment, and our everyday actions—from navigating a room to navigating a conversation—are evidence of our learned skills. Often, we adapt without overt effort or even awareness of the regularity or of our changing behavior. Dubbed *implicit learning* (Reber, 1967), this automatic adjustment to the world is typically studied in cognitive experiments using grammaticality judgments or reaction times to stimuli generated by finite state grammars (see Shanks, 2005 for a review).

The burgeoning statistical learning literature has motivations and predictions that significantly overlap with those of the implicit learning literature, as discussed by Perruchet and Pacton (2006). The seminal work on statistical learning (Saffran, Aslin, & Newport, 1996) demonstrated that infants are sensitive to statistical regularities in a continuous stream of an audible artificial language, enabling them to distinguish probable syllable sequences (i.e., words) from improbable syllable sequences. Newport and Aslin (2004) found that infants are also sensitive to temporally distal regularities, which weighs in favor of a more general statistical learning mechanism, rather than a simple mechanism for associating adjacent sounds. Other studies have found that infants can acquire nouns via the repeated co-occurrence of words and their referents across situations containing multiple words and objects, which are thus separately ambiguous (e.g., Smith & Yu, 2008).

As in adult studies of implicit learning, infant statistical learning studies present participants with structured training data but no explicit learning instructions, and find behavioral differences due to the statistical regularities in the training data. Inspired by this, our aim here is to empirically investigate the automaticity of cross-situational statistical word learning in adults, who are typically given explicit instructions to learn the meaning of the words (e.g., Yu & Smith, 2007). In Experiment 1, we presented participants with a set of spoken words and visual objects with one-to-one mappings between them, but framed the task as one of recognition memory for individual stimuli, and not as one of learning word-object mappings. We then gave participants a surprise test: for each of 54 word-object pairings, they were asked to indicate how often the word and object co-occurred. With their attention focused on memorizing individual words or visual objects, would participants unintentionally learn which words and objects co-occurred more frequently? In Experiment 2, we used a signal detection task as another incidental task to direct participants' attention to both auditory and visual streams, but again with no explicit instructions to learn word-object mappings. After that, we gave them a surprise test to assess their knowledge of word-object mappings. In both experiments, after the initial implicit learning blocks, as a measure of their statistical learning capability (to compare with implicit learning), participants also completed blocks in which they were explicitly instructed to either count word-object co-occurrences, or simply to learn the meaning of the words.

The organization of the paper is as follows: we first introduce the cross-situational learning paradigm, and then discuss the possible learning mechanisms and potential contributions of the present implicit learning studies to advance our understanding of statistical learning. We then present two implicit learning experiments and their results. Finally, we conclude by summarizing the results from the two studies and discussing the connection between statistical and implicit learning.

Cross-Situational Statistical Learning

In a typical version of cross-situational learning, adults are asked to learn which word goes with each object, and are then shown a series of training trials, each of which contains four objects (e.g., a sculpture) and four spoken pseudowords (e.g., "manu"). Because correct word-referent pairings are not indicated, learners can utilize only the repeated co-occurrence of words with their intended referents to learn across many trials. In a typical learning scenario (e.g., in Yu & Smith, 2007), participants attempted

to learn 18 pseudoword-object pairings from 27 12-second trials. This design allowed each stimulus (and hence each correct word-referent pairing) to be presented six times. In one form or another, the learning of a pairing involves the accumulation of word-object co-occurrence statistics across the training trials. Participants acquired, on average, nine of the 18 pairs, as measured by a 4-alternative forced choice (4AFC) referent test for each word.

When each trial contains 16 possible word-referent associations, how might learning proceed? There are at least two distinct approaches that learners may apply. First, an ideal associative learner may maintain a word x object co-occurrence matrix M, incrementing the count in cell $M_{w,o}$ whenever word w and object o appear together in a trial. Table 1 shows such a matrix, which represents the training statistics used in the present study. At test, such a learner may choose the most frequently co-occurring referent for each word. Associative models typically approximate this co-occurrence matrix by strengthening a randomly sampled (perhaps according to current association strengths) subset of pairings on each trial. The association of spatiotemporally proximal stimuli could be carried out by automatic processes that require neither strategy nor intent to learn. Modern memory models such as REM (Shiffrin & Steyvers, 1997) even predict such associations by allowing feature values of nearby items to accidentally be recorded in an item's trace.

Another plausible learning approach is implemented in rule- and inference-based models (e.g., Siskind, 1996), which propose and store a number of hypothesized word-object pairings on each trial. Proposals may be made with respect to constraints such as mutual exclusivity, and hypothesized pairings may be confirmed if consistent evidence is presented later or removed from the lexicon if contradictory evidence is observed. This type of learning is more in accord with a deliberative, strategic learning process. If cross-situational learning is largely automatic, one may expect participants to have some knowledge of which words and objects frequently co-occurred during training, even when they were not explicitly trying to learn these relations. On the other hand, if cross-situational learning relies on more strategic, intentional inferences, then participants may perform much worse in such an incidental learning condition. Thus, the results from incidental learning tasks may shed light on the underlying learning mechanisms that learners use.

In particular, the present study will test participants' knowledge not only of the correct pairings (i.e., the diagonal cells of Table 1) as is typically done, but also of the spurious word-object co-occurrences (non-diagonal cells) that appear during training— the sort of detailed and partial information that is stored by associative models (or an ideal learner), but typically not by rule-based models. We do this by asking participants to rate the strengths of co-occurring word-object pairings for both correct and incorrect pairings.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	6	2	0	1	0	0	1	1	0	0	1	1	3	2	1	4	0	1
2	2	6	1	1	0	1	0	0	2	0	0	0	2	1	2	1	2	3
3	0	1	6	1	2	2	0	0	2	2	1	0	2	1	1	0	2	1
4	1	1	1	6	0	0	0	2	1	1	1	4	0	0	4	0	0	2
5	0	0	2	0	6	1	3	0	1	0	1	1	1	4	0	1	1	2
6	0	1	2	0	1	6	0	1	3	2	0	1	2	0	1	1	3	0
7	1	0	0	0	3	0	6	2	0	2	4	1	0	2	0	1	0	2
8	1	0	0	2	0	1	2	6	0	3	2	2	1	0	2	1	1	0
9	0	2	2	1	1	3	0	0	6	1	0	0	1	1	1	1	2	2
10	0	0	2	1	0	2	2	3	1	6	4	1	1	0	0	0	1	0
11	1	0	1	1	1	0	4	2	0	4	6	2	1	0	0	1	0	0
12	1	0	0	4	1	1	1	2	0	1	2	6	0	0	3	1	1	0
13	3	2	2	0	1	2	0	1	1	1	1	0	6	1	0	2	0	1
14	2	1	1	0	4	0	2	0	1	0	0	0	1	6	0	3	1	2
15	1	2	1	4	0	1	0	2	1	0	0	3	0	0	6	0	2	1
16	4	1	0	0	1	1	1	1	1	0	1	1	2	3	0	6	1	0
17	0	2	2	0	1	3	0	1	2	1	0	1	0	1	2	1	6	1
18	1	3	1	2	2	0	2	0	2	0	0	0	1	2	1	0	1	6

Table 1: Word x Referent matrix with the co-occurrences of each word and object accumulated across the 27 training trials used in each condition in both present experiments.

The testing paradigm allows us to both access participants' knowledge of spurious pairs and to compare that with what they know about correct pairs. Previous work has found evidence that people are sensitive to how often words and objects have co-occurred—even when a single object appears with a few words with differing frequency (Vouloumanos, 2008). However, Vouloumanos presented only a single word-object pair per trial, giving participants no choice as to which pairings to attend. In contrast, our paradigm offers 16 possible pairings per trial. Thus, the presence of four concurrent objects and four successive words per trial demands that participants modulate their attention, possibly forming stronger associations between particular words and objects, or perhaps attending only a subset of possible pairings. Thus, it is unclear how well participants' co-occurrence ratings will be correlated with actual stimuli co-occurrences in the explicit conditions, since inference-based learners may only track a lexicon of the most likely pairs (i.e., high co-occurrence stimuli), rather than a full matrix of associations.

Experiment 1

Every participant went through four blocks of training and testing in a fixed order. Training and testing in block 0 was structured differently than the remaining three. Participants were told that they would see multiple objects and hear multiple words on each trial, and that they should remember each object and word because their memory will be tested at the end. After the brief training period in block 0, they were given a recognition memory test: a single stimulus (word *or* object) was presented, and they were asked to label it *old* or *new*. In block 1, participants were told again that they should remember each object and word for a subsequent memory test. However, after this training period, participants were given a surprise test of their knowledge of stimuli co-occurrences. In block 2, participants were explicitly asked to remember how many times each word and object appeared together during training. They were not

told what type of test to expect, but the co-occurrence rating test given was exactly the same as in block 1. In block 3, participants were simply asked to learn the meanings of the words—explicit learning instructions like those given in previous cross-situational word learning studies.

Subjects

Participants were 35 undergraduates at Indiana University who received course credit for participating. None had participated in other cross-situational experiments.

Stimuli

Verbal stimuli were 72 computer-generated pseudowords that are phonotactically-probable in English (e.g., "bosa"), and were spoken by a monotone, synthetic female voice. Objects were 72 photos of uncommon, difficult-to-name objects (e.g., strange sculptures). Of these sets of objects and words, 54 were randomly assigned to three sets of 18 word-object pairings; one set for each study condition. The remaining 18 words and 18 objects were used for an initial recognition memory test.

In block 0, each trial presented three unusual objects concurrently and three pseudowords heard in succession. Block 0's training consisted of only three 11-second trials, displaying in total nine unique words and objects. After these trials, participants were given a recognition test for each trained word and object, as well as nine new words and objects. On each test trial, a single stimulus (word *or* object) was presented, and participants were asked to indicate if it was *old* or *new*.

In blocks 1-3, each training trial consisted of a display of four objects and four pseudowords were played in succession, and 27 such trials were in each block. Each training trial began with the appearance of four objects, which remained visible for the entire trial. After 2 seconds of initial silence, each word was heard (randomly ordered, duration of one second) followed by two additional seconds of silence, for a total duration of 14 s per trial.

After each training period, participants were tested for knowledge of stimuli co-occurrences. One word and one object were presented on each trial, and participants were asked to indicate how many times [0-6] the given word-object pairing had appeared during training. Each of the 18 words and objects appeared in three test trials, for a total of 54 randomly-ordered trials. The correct (6-co-occurrence) pairings comprised 18 of the test trials (Table 1's diagonal). The remaining 32 trials tested cells in the matrix with 0 (14 trials), 1 (14), 2 (12), 3 (8), and 4 (6) co-occurrences.

Procedure

Condition order was fixed, and each participant took part in all four blocks. Block 0 was a three trial training period with three words and objects per trial, followed by a recognition test of every individual stimulus presented, and nine new words and objects. In block 1, participants were instructed to study individual stimuli for a memory test. However, following the 27 training trials, participants were instead asked to indicate how many times [0-6] each of 54 specific word-object pairings appeared during training. In block 2, participants were asked to track how often each word co-occurred with each object. After the 27 training trials—which had the same co-occurrence statistics as in block 1, albeit different stimuli—participants were again given test trials asking them to rate the same 54 pairings. Finally, in block 3 participants were simply instructed to learn the meanings of the words, given cross-situational training (statistically identical to blocks 1 and 2), and again tested on the same 54 pairings.

Results & Discussion

In block 0, participants recognized a mean of 96% of the objects and 90% of the words, with a low false alarm rate (8%). In both word and object recognition, every participant was at least 77% accurate. It is notable that memory is imperfect for the stimuli, since many models of cross-situational learning assume that learners can absolutely identify each stimulus, which is evidently not the case.

To determine how related participants' co-occurrence ratings were to the actual number of times the tested word-object pairings actually appeared together during training, Kendall's rank correlation coefficient (tau) was calculated for each participant's 54 test trials in each condition. The mean tau values for each condition are shown in Figure 1. In block 1, when participants were studying individual words and objects (but not attending to co-occurrences), their responses in the surprise rating task showed a small but significant positive correlation with the actual number of times the presented pairings co-occurred during training (M = .04, one-sided $t(34) = 1.90$, $p<.05$). In comparison, in the explicit learning conditions in blocks 2 and 3, when participants were respectively told to track all word-object co-occurrences and to learn the meaning of the words, their ratings were significantly more positively correlated than in block 1 (block 2 M = .15, paired $t(34) = 3.82$, $p<.001$; block 3 M = .17, paired $t(34) = 3.86$, $p<.001$). Moreover, the strength of correlations in the two explicit conditions is not significantly different (paired $t(34) = 0.66$, $p>.05$).

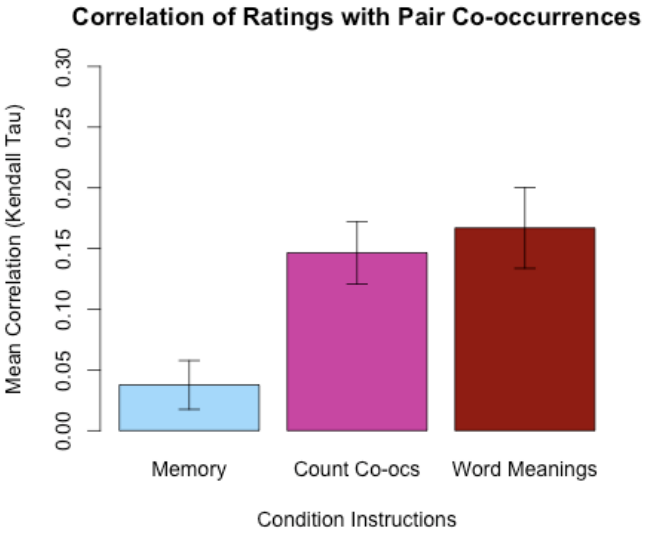


Figure 1: Mean correlation of participants' responses with the actual pair co-occurrences in Exp. 1. Error bars: +/-SE.

Positive correlations between ratings and a broad sample of the actual co-occurrence statistics from training indicate that participants are sensitive to arbitrary stimuli co-occurrences when explicitly told to attend to such correspondences. However, one could imagine that the positive correlations could be due largely to knowledge of some particular subsets of the co-occurrences: e.g., perhaps learners are sensitive to words and objects that never co-occurred, and thus rated these pairings very low, and all others high. To examine performance in more detail, we calculated each participant's d-prime (d′)[1] for the most extreme pairings tested in each condition: stimuli that co-occurred 0 or 6 times. Positive d′ shows sensitivity resulting from a high hit rate and low false alarm rate. As shown in Figure 2, participants only had significant sensitivity for 6-co-occurrence ('correct') pairings in the explicit learning conditions (count co-occurrences $M = 0.64$, one-sided $t(34) = 4.92$, $p<.001$; word meanings $M = 0.81$, one-sided $t(34) = 5.08$, $p<.001$).

Two patterns from this study are noteworthy. First, based on both d′ analysis and correlation measures, the learning that results from the counting co-occurrences condition and the word learning condition were similar. Although not conclusive, this may suggest that participants in the word learning condition may have used an associative learning strategy based on counting word-object co-occurrences.

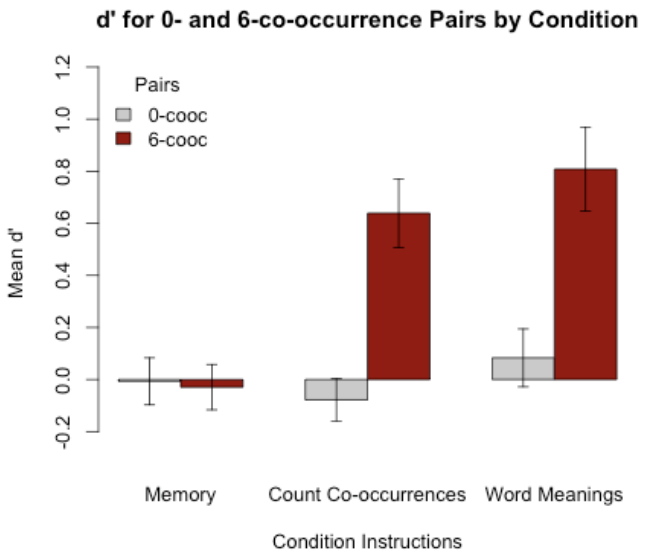


Figure 2: Mean d′ for 0- and 6-co-occurrence word-object pairings in Experiment 1, by condition. Error bars are +/-SE.

Second, despite having good recognition memory for individual words and objects that were presented during training, word learning was very poor in the implicit learning condition, as measured both by correlation of their responses with actual pair co-occurrences, and by d′ for correct pairings and stimuli pairings that never co-occurred. Nonetheless, the positive correlations found in every condition—although smaller in the implicit condition—show that participants do, on average, absorb some of the stimulus co-occurrences in all conditions. However, this sensitivity is not enough to support implicit word learning in our study, as much stronger correlations are shown when learners are instructed to count co-occurrences or learn word meanings. Under these instructions, participants become sensitive to words and objects that frequently co-occur.

Experiment 2 investigates whether a different incidental task, which may direct attention to word-object co-occurrences, rather than the stimuli themselves, may yield automatic word learning.

Experiment 2

Experiment 1 showed that an incidental memory task results in some implicit knowledge of word-referent co-occurrences, but that explicit instructions to learn word-object co-occurrence or to learn word meanings resulted in much greater knowledge. In Experiment 2, we use a different task in the implicit learning condition: instead of asking participants to remember individual stimuli for a later memory task, we give participants a signal detection task to carry out during training. This task—detecting visual noise added to visual objects, and louder auditory stimuli (words)—directed participants to pay attention to both visual and auditory stimuli simultaneously, but gave no directions to engage in learning of word-object pairings.

Subjects

37 undergraduates at Indiana University received course credit for participating. None had participated in previous cross-situational experiments.

Stimuli

The sets of pseudowords and referents for Experiment 2 were identical to those used in Experiment 1. Training trials were the same as those used in Experiment 1, and had the same co-occurrence statistics (shown in Table 1). However, on each training trial in blocks 1 and 2, a random number [0-4] of the words were louder than others, and Gaussian pixel noise was momentarily added to a single object *during a word presentation* a random number of times [0-4] each trial. Thus, for 6.3% of audio stimulus presentations during training, that word would be loud and *one* of the objects would simultaneously have noise added, highlighting a pairing—but only the correct pairing in 25% of these cases.

Procedure

In block 1, participants were told that they would be presented with artificial words and objects on a series of slides, on which some words would be louder than the others and some objects would have multicolored speckles (noise). Their task was to quickly press the mouse button each time a loud word or noisy object was presented. However, after the 27 training trials, participants were given a surprise test, and asked to indicate how many times [0-6]

[1] For example, hits for 0-co-occurrence pairings are responses of 0, and false alarms are responses of 0 for pairings that co-occurred more than never. d′ = Z(p(hit)) – Z(p(false alarm)), where Z is the inverse of the cumulative Gaussian distribution.

each of 54 specific word-object pairings appeared during training. In block 2, participants were asked to track how often each word co-occurred with each object, and were also told to do the same signal detection task during training. Instructions for block 3 asked participants to track word-object co-occurrences without doing the signal detection task, and in block 4 participants were simply told to learn the meanings of the words. The same 54 rating test trials of specific pairings followed the training periods of blocks 2, 3, and 4, though with different stimuli for each block.

Results & Discussion

Experiment 2 used a signal detection (SD) task that required participants to attend to both auditory and visual stimuli, but did not mention that they would need to remember the stimuli later. However, as in Experiment 1, after this first training block participants were given a surprise test for incidental learning. In successive learning conditions, participants were instructed to do both the SD task and to count word-object co-occurrences (SD+CC), to count co-occurrences (with no other task; CC), and finally, to simply learn the meanings of the words (Word Meanings). As in Experiment 1, Kendall's tau was calculated for each participant's 54 test trials in each condition to measure how related their ratings were to the actual number of word-object co-occurrences. As shown in Figure 3, although the SD task resulted in significantly positive correlations (M = .10, one-sided $t(36) = 3.75$, $p<.001$), the explicit learning conditions showed significantly more correlated responses (CC M = .21, paired $t(36) = 3.57$, $p<.01$; SD+CC M = .25, paired $t(36) = 5.12$, $p<.001$; Word Meanings M = .29, paired $t(36) = 4.97$, $p<.001$). Thus, as found in Experiment 1, participants show sensitivity to stimuli co-occurrences in every condition, but greater sensitivity in the explicit learning conditions than in the implicit learning condition.

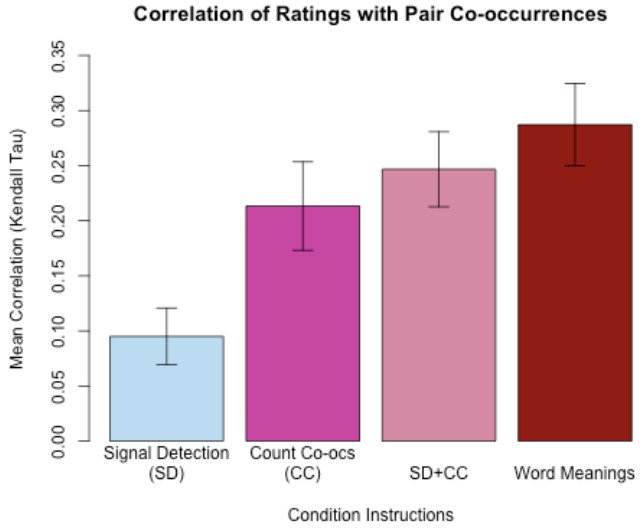


Figure 3: Mean rank correlation of each participant's responses with the actual number of pairing co-occurrences in Experiment 2. Error bars show +/-SE.

As in Experiment 1, we calculated d′ for maximal and minimal co-occurrence pairings by condition to gain insight into the kind of pairings to which participants in Experiment 2 were sensitive. As shown in Figure 4, participants had significant sensitivity for 6-co-occurrence pairings in the implicit learning condition (SD M = .19, one-sided $t(36) = 1.81$, $p<.05$) as well as the explicit conditions, but showed significantly greater sensitivity in the explicit conditions (CC M = .61, paired $t(36) = 2.97$, $p<.01$; SD+CC M = .61, paired $t(36) = 3.44$, $p=.001$; word meanings M = .81, paired $t(36) = 3.58$, $p<.001$). In the explicit conditions, d′ for 0-co-occurrence pairings was significantly positive (CC M = .49, one-sided $t(36) = 1.50$, $p=.07$ (marginal); SD+CC M = .32, one-sided $t(36) = 2.66$, $p<.01$; word meanings M = .30, one-sided $t(36) = 2.20$, $p<.05$), but not in the implicit condition (SD M = .07, one-sided $t(36) = .87$, $p=.19$). Thus, although participants given SD instructions did show some implicit learning of 6-co-occurrence pairings, they were more sensitive to these pairings under explicit instruction.

There are a few intriguing results from this experiment. First, performance in the SD+CC condition was at least as good as CC alone. Thus, participants could handle the two tasks concurrently without hindering performance. We suspected that the signal detection task might encourage participants to attend to both auditory and visual streams simultaneously, perhaps increasing storage of cross-modal associations. Possibly as a result of this focus, in contrast to Exp. 1, participants in Exp. 2 showed significant sensitivity to 0-co-occurrence pairings in the explicit conditions.

Second, word-learning instructions yielded performance as high as found in other explicit instructions (SD+CC and CC). This confirmed our finding from Experiment 1: both counting co-occurrences—as an ideal associative learner might do—and attempting to learn words result in similar performance in humans, both for correct pairs and for spurious co-occurrences.

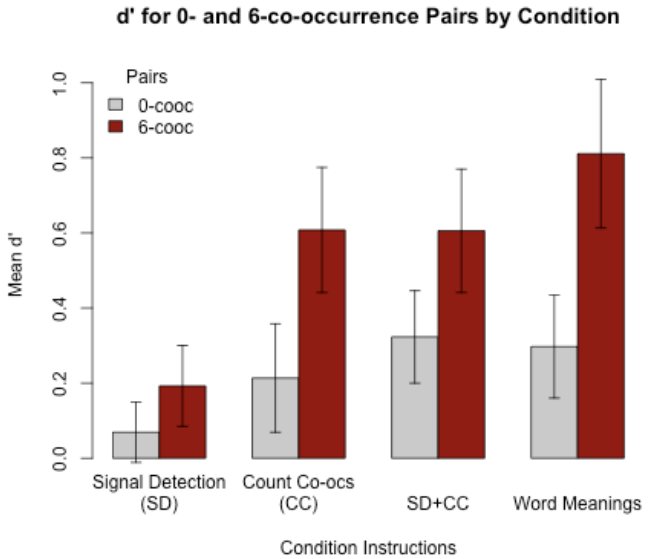


Figure 4: Mean d′ for 0- and 6-co-occurrence word-object pairings in Experiment 2, by condition. Error bars are +/-SE.

General Discussion

Implicit learning and statistical learning both describe an agent's adaptation to regularities in its environment. We set out to determine whether cross-situational word learning can be accomplished by mere exposure to the same type of training used in intentional settings. In Experiment 1's implicit learning condition, participants attempted to remember individual stimuli. In a surprise test of knowledge for word-object co-occurrences, participants' ratings were correlated with the actual number of co-occurrences, meaning that learners had acquired a rough approximation of the real-world statistics, much like associative models predict. However, a signal detection analysis showed no sensitivity to correct word pairings. Moreover, in subsequent explicit conditions, participants showed stronger correlations, as well as sensitivity to correct pairings. Using a signal detection task rather than a memory task in the first block, thus encouraging concurrent attention to both words and objects, Exp. 2 asked again whether participants acquire cross-situational co-occurrence statistics automatically. Participants demonstrated some implicit knowledge as in Exp. 1, but also showed some sensitivity for correct word-referent pairs. However, in explicit conditions participants showed greater sensitivity to such frequently co-occurring stimuli, as well as significant knowledge of spurious co-occurrences. Furthermore, we found that participants' learning when instructed to count co-occurrences looks similar to learning under instructions to merely learn words, which we speculate may mean that participants utilize a similar strategy in both conditions. By asking participants to perform slightly different tasks with the same input and then comparing their resulting learning, it will be possible to determine which regularities are automatically acquired and which must be explicitly attended or inferred.

What do the present results tell us about cross-situational statistical learning? They seem to contradict simple hypothesis-testing mechanisms, which would typically not maintain information about spurious co-occurrences, and which may not operate automatically. However, the results also contradict a strong associative account: learning was greater in explicit conditions than in implicit conditions, suggesting that learning may be in part strategic, or at least modulated by attention. Thus, we may say that cross-situational statistical word learning is neither wholly implicit, nor wholly explicit: some statistics are acquired automatically, and the learning system indubitably uses this information during explicit study, as well. Moreover, the fact that the explicit conditions always produced greater sensitivity for the correct pairings than for pairings that never co-occurred suggests that some mechanism for highlighting stimuli that frequently co-occur is at work.

In summary, although the implicit learning we observed was inferior to the explicit learning, its presence indicates that knowledge of co-occurrence statistics can be acquired incidentally. Since implicit learning requires few resources, it can be carried out minute-by-minute, hour-by-hour and day-by-day. Hence, in the long run, cumulative implicit learning may still play an important role in human language acquisition. Overall, our work suggests that neither simple associative models that approximate ideal observers, nor hypothesis-testing models relying on explicit inferences capture both the implicit and intentional aspects of cross-situational word learning. We hope that this work will motivate researchers to consider hybrid models that include both strategic, inference-based mechanisms as well as automatic, associative ones. Finally, we believe this work represents an early step in linking the implicit learning and statistical learning literatures.

Acknowledgments

This research was supported by National Institute of Health Grant R01HD056029. Special thanks to Jeanette Booher for data collection.

References

Kachergis, G., Yu, C., & Shiffrin, R. M. (2009). Frequency and contextual diversity effects in cross-situational word learning. In N. Taatgen, H. van Rijn, J. Nerbonne, & L. Schomaker (Eds.) *Proceedings of CogSci 31*. Austin, TX: Cognitive Science Society.

Newport, E. L., & Aslin, R. N. (2004). Learning at a distance: I. Statistical learning of non-adjacent dependencies. *Cognitive Psychology*, *48*(2), 127–162.

Perruchet, P. & Pacton, S. (2006). Implicit learning and statistical learning: one phenomenon, two approaches. *Trends in Cognitive Sciences*, *10*, 233-238.

Reber, A. S. (1967). Implicit learning of artificial grammars. *J. Verbal Learn. Verbal Behav*., *6*, 855-863.

Saffran, J.R., Aslin, R.N., & Newport, E.L. (1996). Statistical learning by 8-month old infants. *Science*, *274*, 1926-1928.

Shanks, D.R. (2005). Implicit learning. In *Handbook of Cognition* (Lamberts, K. & Goldstone, R., eds), pp. 202–220, Sage Publications.

Shiffrin, R. M. & Steyvers, M. (1997). A model for recognition memory: REM: Retrieving effectively from memory. *Psychonom. Bulletin and Review*, 4(2), 145-166.

Siskind, J. M. (1996) A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, *61*(1-2):1-38.

Smith, L. & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition, 106,* 1558-1568.

Vouloumanos, A. (2008). Fine-grained sensitivity to statistical information in adult word learning. *Cognition, 107,* 729-742.

Yu, C. (2008). A statistical associative account of vocabulary growth in early word learning. *Language Learning and Acquisition*, *4*(1), 32-62.

Yu, C. & Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science, 18,* 414-420.

Yurovsky, D. & Yu, C. (2008). Mutual exclusivity in cross-situational statistical learning. *Proceedings of CogSci 30*. Austin, TX: Cognitive Science Society.

An Interactive Environment for Explanatory Biological Modeling

Pat Langley (langley@asu.edu)
Computing Science and Engineering
Arizona State University, Tempe, AZ 85287 USA

Abstract

In this paper, we describe an interactive environment for the representation, interpretation, and revision of explanatory biological models. We illustrate our approach on the systems biology of aging, a complex topic that involves many interacting components. We also report initial experiences with using this environment to codify an informal model of aging. We close by discussing related efforts and directions for future research.

Keywords: scientific models, qualitative reasoning, applied cognitive science

Introduction and Overview

There is general agreement that the explosive growth in biological data offers great opportunities but also poses major challenges. Although less widely recognized, the growing complexity of biological models that aim to account for these observations raises a host of other issues. Computational techniques hold promise for mitigating this complexity, but most responses have been driven by algorithmic concerns rather than the cognitive needs of scientists who must develop, interpret, and understand complex models. Biologists would benefit from new computational tools designed with scientific users in mind.

Many efforts in modern science aim to understand complex phenomena from a systems perspective. One important example comes from research on aging, with recent studies suggesting that senescence results from the interaction of many distinct but interconnected processes (Vijg & Campisi, 2008). Individual laboratories report experiments and propose hypotheses to explain them, but there has been little work on how they fit together. The systems biology movement has championed integrative science, but it has emphasized topics like gene regulation and left phenomena like aging understudied.

In this paper, we report an interactive computational framework designed to support modeling of this variety. Our approach relies on three distinct but mutually supportive ideas:

- formal representations of scientific knowledge that make contact with specific fields' terms and concepts;
- methods for reasoning over models cast in these formalisms that provide the same flexibility and draw the same conclusions as scientists;
- techniques that let researchers analyze and update these models in an incremental, cumulative manner.

In the next section, we discuss three computational challenges that these capabilities raise, after which we describe an interactive software environment that embodies our responses. We illustrate the system's abilities with examples from the domain of aging, then report initial experiences with the environment. We conclude with a discussion of related work on scientific modeling, along with directions for additional research.

Some readers may question the relevance of our work to cognitive science. Of course, scientific reasoning has long been a topic of study within this community, but we will not claim our system reasons in precisely the same way as biologists. However, our approach is informed by results from cognitive science that constrain it in important ways. In particular, it borrows from research on qualitative mental models, which has proposed representations and reasoning methods that are consistent with knowledge about human cognition. A good analog comes from work on intelligent tutoring systems (e.g., VanLehn, 2006), which does not model the details of human tutors but takes lessons from them. We view our work on computational aids for biological modeling as another important instance of applied cognitive science.

Challenges in Scientific Modeling

As we have noted, the construction of complex scientific models raises three separate but interrelated challenges. Here we expand upon each of them in turn, placing constraints on the form our responses should take in developing an environment for biological modeling.

The overall aim of science is to produce knowledge, but the social nature of science requires the use of *communicable* formalisms that researchers can exchange and understand (Džeroski, Langley, & Todorovski, 2007). Thus, our first computational challenge involves selecting a communicable formalism for biological models. Over the past decade, computational researchers have proposed many notations for such models, but most utilize notations borrowed from other fields that have questionable relevance to traditional biological thinking. Research in biology generally, and on aging in particular, imposes two constraints on modeling formalism. One is that most accounts of phenomena are qualitative, not because researchers prefer them intrinsically, but because they enable useful claims even when lacking more precise information. A second feature is that biologists often move beyond simple predictive models to posit causal hypotheses or processes that underlie known phenomena.

Science also differs from some areas of inquiry by its concern with observations. However, biologists typically

desire more from their models than simple predictions; they prefer *explanations* that account for observations in terms of concepts and mechanisms they find familiar and plausible. Such explanatory reasoning is common in biology (Darden, 2006), but the growing complexity of models suggests that, without assistance, researchers will otherwise overlook important implications. Thus, a second computational challenge involves supporting reasoning over the communicable scientific formalisms just described. Methods for calculating results from numeric equations are well established, but automated reasoning over the qualitative models that dominate biology requires a different approach. One complication that arises in qualitative models is that two or more causal pathways can predict different relationships between variables. Another is that it can be difficult to reason qualitatively about how a system changes over time.

A third important feature of science is its cumulative character. Historians often focus on conceptual breakthroughs by individuals like Darwin, Pasteur, and Morgan, but the great majority of research involves filling in technical details rather than changing paradigms. This is especially true for biology and medicine, in which scientists devote considerable effort to piecing together complicated models with many interacting parts. Thus, our final computational challenge involves supporting the cumulative improvement of system-level models by biological researchers. A common response is to develop curated knowledge bases (e.g., Karp et al., 2000; Vastrik et al., 2007) that rely on centralized control by a few experts, but the field has also explored community-based approaches. Both require ways to update models incrementally as new knowledge becomes available.

An Interactive Modeling Environment

We have incorporated our responses to the above issues into an interactive software environment for biological modeling. We have implemented the initial system in Lisp and we have used it to formalize four compartments of Furber's (2009) network diagram of aging, which depicts in a graphical but informal way some well-supported hypotheses and phenomena from biogerontology. In this section, we report the environment's response to each of the challenges just described, using examples from aging to clarify its operation.

Representing Biological Models

Recall that our first computational challenge involves encoding explanatory models and presenting them in ways that biologists will understand. Let us review some key features of aging that hold implications for modeling these phenomena:

- Different effects of aging and age-related disease are localized in different portions of body. For instance, some age-linked changes occur in specific parts of the cell, such as the lysosome or the mitochondria.
- Some hypotheses about aging involve transient substances, such as enzymes and reactive oxygen species (ROS), whereas others involve far more stable entities like lipofuscin and mitochondrial mutations that accumulate over time.
- Empirical results generally take the form of qualitative relations between continuous variables. For instance, one robust finding involves a negative influence of caloric intake on lifespan in model organisms.
- Aging takes place over time, but its effects are primarily monotonic in character, with the values of variables increasing or decreasing consistently. For example, lipofuscin in the lysosome is generally observed to increase with chronological age.
- Empirical findings about aging come in two distinct varieties: uncontrolled observations about changes over time and results of controlled experiments that measure the effect of one variable on another.

Taken together, these observations provide both constraints on our approach to modeling aging processes and avenues for making the task more tractable.

Table 1 presents our reformulation of the lysomone compartment of Furber's network diagram. The initial 12 statements in (a) and (b) reflect the first two points above. They declare specific locations – the lysosome, the cytoplasm, and the cell that contains them – along with quantities that are measurable (at least in principle) in those locations. Some quantities refer to stable substances, such as junk protein, oxidized protein, and lipofuscin, which accumulate over time unless actively broken down, whereas others denote transient substances, like Fe, ROS, and lytic enzyme, which are reactive enough to be very short lived.

The table also includes a set of hypotheses (c) about how these quantities influence each other. One claim is that transient ROS increases with transient Fe within the lysosome, whereas another is that stable oxidized protein increases with transient ROS in the same location. Hypotheses may also relate quantities in distinct locations (e.g., that lipofuscin in the cytoplasm increases with damaged membrane in the lysosome). These hypotheses have a clear causal interpretation, in that they state how one variable will change when one alters another. However, although they link continuous quantities, the relations themselves are qualitative in character.

Of course, we should remember the purpose of hypotheses like those in Table 1 (c), which is to explain known empirical results and predict new ones. This in turn requires not only that we represent these empirical findings formally, but also that we distinguish them clearly from the hypotheses themselves. Table 1 (d) shows four facts about aging in the lysosome that illustrate our earlier point about two forms of empirical findings. The first two items clarify both the observational, nonexperimental character of many facts about

Table 1: Formalization of Furber's (2009) lysosome model, including (a) locations, (b) stable and transient quantities in these locations, (c) hypothetical claims about causal influences between these quantities, and (d) empirical facts about relations between quantities.

(a) location cell
location lysosome in the cell
location cytoplasm in the cell
(b) stable junk protein in the lysosome and cytoplasm
transient degradation rate in the lysosome
transient Fe in the lysosome
transient ROS in the lysosome
stable oxidized protein in the lysosome
stable lipofuscin in the lysosome and cytoplasm
transient lytic enzyme in the lysosome
stable damaged membrane in the lysosome
transient H2O2 in the lysosome and cytoplasm
(c) hypothesis junk protein decreases with degradation rate in the lysosome
hypothesis junk protein in the lysosome increases with junk protein in the cytoplasm
hypothesis Fe increases with junk protein in the lysosome
hypothesis ROS increases with Fe in the lysosome
hypothesis oxidized protein increases with ROS in the lysosome
hypothesis lipofuscin increases with oxidized protein in the lysosome
hypothesis degradation rate decreases with lipofuscin in the lysosome
hypothesis lytic enzyme decreases with lipofuscin in the lysosome
hypothesis ROS increases with lipofuscin in the lysosome
hypothesis damaged membrane increases with ROS in the lysosome
hypothesis lipofuscin in the cytoplasm increases with damaged membrane in the lysosome
hypothesis H2O2 in the lysosome increases with H2O2 in the cytoplasm
(d) fact lipofuscin in the lysosome increases with time
fact membrane damage in the lysosome increases with time
fact lytic enzyme decreases with ROS in the lysosome
fact H2O2 does not change with ROS in the lysosome

aging and also their monotonic nature. These explicitly mention time as a variable, which the model hypotheses do not. The other two facts reflect (plausible) results of experimental studies that measure the effect of one quantity's variation on another. The first states that lytic enzyme decreases with ROS in the lysosome. The second states that H2O2 does not vary with of ROS. Such negative results place constraints on models, although hypotheses may contain only positive causal relations.

This notation meets two of the criteria given earlier. It supports qualitative models that nevertheless relate quantitative variables of the type that biologists typically measure, and the hypotheses that make up models

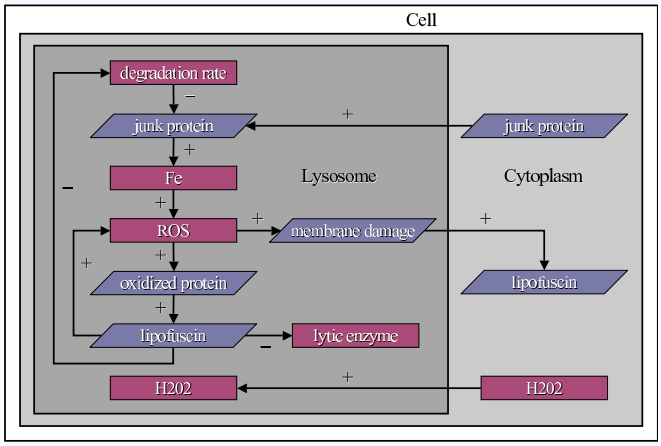


Figure 1: A graphical visualization of the qualitative lysosome model from Table 1, with plus (+) on an arrow denoting that one quantity increases with another and with minus (−) denoting a decreasing relationship.

have a clear causal interpretation. The formalism also lends itself to graphical display, with quantities shown in locations where they occur and with arrows depicting direct causal influences between these variables. Figure 1 shows a graphical version of the lysosome model from Table 1, with the empirical facts omitted. Our implemented system does not yet generate such graphs automatically, but adding this ability should not be difficult.

In addition, our notation lets users specify places, quantities, hypotheses, and empirical facts in constrained English, which we believe will make it more accessible to biologists who are uncomfortable with traditional computer languages. Yet models stated in this notation are well defined and unambiguous about their claims, making them just as formal as ones stated in the more arcane languages typically proposed in computational biology. This also distinguishes our approach from work on qualitative reasoning in cognitive science and AI (e.g., Bredeweg et al., 2007; Forbus, 1984), which has influenced our approach to biological modeling.

Reasoning over Biological Models

Our second computational challenge involves interpreting a given model to account for known phenomena. Scientists regularly engage in such reasoning, but with complex models they can easily overlook some conclusions and incorrectly infer others (e.g., Feldman et al., 1989). Thus, automatically determining a model's implications should be a key part of our scientific modeling environment. Good models should explain known phenomena and predict new ones accurately, while phenomena place constraints on model content. Our framework's formal statement of hypotheses and empirical results has another advantage: it lets one answer questions about how one quantity should affect another and predict the outcomes of thought experiments.

We can clarify this ability by introducing the notion of a *query* about how two quantities are related. This takes the same form as an empirical finding except that it does not state the direction in which one variable influences another or indeed whether an influence occurs at all. Thus, given the hypotheses in Table 1, we might ask "Does lipofuscin in the cytoplasm vary with Fe in the lysosome?" or "Does ROS in the lysosome vary with time?" The first asks a question about how changes to one quantity in a controlled experiment affect another; the second asks how a given quantity changes over time. The reasoning task is simplified by our assumption that effects are monotonic in character, giving behavior that one can describe in terms of a single qualitative state. This differs from much work on qualitative reasoning, which deals with trajectories of such states over time (e.g., Bredeweg et al., 2007; Forbus, 1984).

Because hypotheses take a form similar to facts, we can utilize a relatively straightforward chaining procedure to answer queries. To handle a question about how dependent variable Y varies with independent variable X, other things being equal, one simply finds a causal pathway, typically through other quantities, that starts with Y and ends with X. If no such path exists, then one can conclude that changes to X do not produce changes in Y. If there is such path, then one must still predict the direction of the effect. Briefly, if the path contains an even number of 'decreases' links, then one predicts that Y increases with X; otherwise one predicts that it decreases. For example, the model in Figure 1 lets one conclude that lytic enzyme will decrease with ROS. The justification for this strategy is simple: each 'decreases' link reverses the direction or sign of the path's overall influence, so that an even number of them cancel out.

One complication arises when multiple paths from Y to X make different predictions. Without knowing the functional forms and parameters that produce each causal link, one cannot determine the exact effects of alternative pathways. Given the modeling framework as we have described it, in such cases one can only state that the hypotheses make contradictory predictions. However, we can extend the formalism in a simple way that lets it express another type of hypothesis that biologists regularly make: that the effect of one causal pathway dominates that of another. This requires a way to specify paths between two quantities and note which has the greater or dominating effect. Once included, such dominance relations let a qualitative causal model make unambiguous predictions about how one quantity varies with another, despite its abstract character.

Reasoning about how quantities change over time requires a slightly different approach. We assume that any exogenous variables not influenced by other quantities take on constant positive values. One can then infer the effect of such an exogenous quantity on another variable downstream by finding pathways that connect them and combining the influences on their causal links. One can conclude that 'stable' quantities occurring downstream will increase or decrease over time, depending on their relation to the exogenous term. We can treat causal loops between two variables as special cases of conflicting paths in which a variable influences itself, again provided we specify which path is dominant.

Taken together, these computational mechanisms respond to a number of the issues raised above. They let our biological models move beyond inert structures to become interpretable 'programs' one can use to answer directed queries and make predictions about empirical relations. They also support reasoning about the effects of both controlled manipulation and the passage of time. As we will see shortly, the system can also explain the reasons for its conclusions. Computational aids of this sort should let biologists derive the implications of system-level models of aging that are more complex than ones they can handle without assistance.

Interactive Aids for Model Improvement

Our third computational challenge involves the incremental revision of models to bring them into closer alignment with known phenomena. This depends on the ability to represent such models and reason over them, but it must go beyond to identify portions of models that are problematic and modify them in response. Although there has been some research on automated model revision (e.g., Mahidadia & Compton, 2001), we have chosen to rely on interactive revision under user control. To this end, the system includes a number of commands through which users can update the knowledge base. These are currently available only through a textual interface, but we also plan to embed them in a graphical environment.

Naturally, the most basic commands includes ones for adding new model elements. The user can introduce new locations, quantities, hypotheses, and empirical facts by entering this content in the same format as shown in Table 1. The modularity of the modeling formalism, and its constrained English syntax, make these steps simple to carry out. The environment also includes a display command that presents the user with all elements in the current model or only those of a specified type. These commands provide the basic functionality needed for the cumulative improvement of causal biological models.

However, the system also provides users with additional details about the model's behavior that can inform their revisions. In addition to answering specific queries like "Does ROS in the lysosome vary with time?", users can also ask the environment to compare the current model's predictions to known phenomena. When these predictions disagree with the empirical facts, the user can also ask the system to explain its reasoning. For each explanation, it presents the causal chain between two quantities that, taken together, predicted a partic-

ular outcome. Exceptions occur when the model incorrectly predicts no effect because no causal chain exists or makes an ambiguous prediction when two paths conflict and the user has not specified one as dominant.

The ability to inspect not only predictions but the reasoning behind them provides important insights about a model's strengths and weaknesses. If the model fails to match one or more empirical facts, explanations may reveal the source of the problem and ways to fix it. The user can remedy such situations in two basic ways – by adding new hypotheses, as described above, and by removing existing hypotheses. However, because the impact of deleting an element may be unclear in advance, the environment also lets users disable a model element without removing it entirely, as well as enable it later if that seems desirable. Taken together, these commands provide basic support for the incremental improvement of models, which will continue to be needed as new phenomena become available and demand explanation.

Initial Experiences with the Environment

We selected the systems biology of aging as our initial application domain because it was gaining increased attention within biology and because John Furber (2009) had already developed a network diagram that summarized many hypotheses and phenomena in this complex field. Repeated discussions with Furber let us convert his informal statements into our modeling notation.

We have focused our efforts on four compartments of Furber's diagram. These involve the dysfunction of lysosomes due to the accumulation of indigestible aggregates known as lipofuscin, the degeneration of mitochondrial energy production in the cell as the result of mutations, the shortening of telomeres and decline in Lon protease mRNA over time in the cell nucleus, and the crosslinking of proteins in the extracellular matrix. The lysosomal model, already seen in Table 1 and Figure 1, incorporated three places, nine quantities, and 12 hypotheses. The mitochondrial model included three places, nine quantities, and ten hypotheses, while the nuclear and extracellular models have similar complexities.

Naturally, translation of content from the informal diagram into our logical notation required some care and effort, with certain representational issues becoming apparent only along the way. Interactions with Furber clarified his intentions and usually determined how to proceed. Once we had the initial translation complete, we used the environment to detect and correct problems with these models, much as we intend its use by scientists. Running the reasoning mechanism over these models revealed a number of errors, some in our encoding of Furber's chart but also a few ambiguities in the original aging diagram itself. Formalization of the aging model, combined with the environment's reasoning methods, led to repair of these problems.

Related Work on Scientific Modeling

Our approach to interactive biological modeling borrows ideas from three distinct traditions, but combines them in new ways to produce novel capabilities. The computational biology community has pursued a number of projects that support Web-based access to biological knowledge. For instance, KEGG (Kanehisa, 1997), Reactome (Vastrik et al., 2007), and Metacyc (Karp et al., 2000) let their users explore biological content that curators have extracted from the literature, but they have only limited abilities to reason over their knowledge.

Some other biological modeling efforts come closer to our framework. For example, Genepath (Zupan et al., 2003) offers a Web-based environment that lets users enter qualitative results from genetics experiments and knowledge about gene regulation, but the model construction process is entirely automated. JustAid (Mahidadia & Compton, 2001) supports iterative revision of qualitative causal models, with the system proposing changes but the user selecting which to implement. Racunas et al.'s (2004) HyBrow supports interactive creation of qualitative models and checks their consistency with logical reasoning, but our system provides a more general treatment of explanatory biological models.

Of course, we have also been strongly influenced by research on mental models in cognitive science, especially work on qualitative reasoning and simulation (e.g., Forbus, 1984). Our approach shares some key ideas, especially that models involve qualitative causal relations among continuous variables. One difference is our assumption that behavior is monotonic over time, which simplifies reasoning considerably. Another distinction is our willingness to resolve ambiguity by specifying that one path dominates another. A third lies in our emphasis on predicting relations between pairs of quantities, rather than on model simulation. Our incorporation of qualitative models into an interactive modeling environment is not new. Bredeweg et al.'s (2007) GARP lets users construct qualitative models manually and simulate their behavior, although it focuses on ecology rather than biology, it uses a more complex process ontology, and it does not emphasize incremental revision.

Directions for Future Research

Although our modeling environment shows considerable promise, we need to extend the framework along a number of fronts. Clearly, our first step should be to embed the existing abilities in a graphical interface. This would let users visualize models in a manner similar to Figure 1, but it would also use this display to support query answering, prediction, and explanation, each of which have natural visual analogs. The environment would include templates for creating new locations, quantities, hypotheses, and empirical facts, for disabling and enabling model elements, and for copying and editing entire

models. These features would not change the environment's basic functionality, but they would make it more accessible to many biologists.

We should also expand the representational abilities of the modeling framework. One extension would enable grouping a set of causal links into a process, much as in Forbus' (1984) qualitative process theory. This would let a graphical interface hide model details until a user asks to see individual connections. Another augmentation would allow contextual conditions on causal links that specify the tissues and organisms in which they occur. If queries included similar conditions, then the reasoning system would collect relevant connections to create query-specific models for use in drawing conclusions. Finally, we should explore ways to move beyond the framework's strict assumption of monotonic behavior. One response would involve adding quantitative conditions to causal links and dominance relations that specify when they hold, with the reasoner collecting relevant model elements to make predictions for a specific situation.

Concluding Remarks

In this paper, we reported an interactive approach to the representation, interpretation, and revision of scientific models. Our environment encodes models as sets of qualitative causal influences that relates quantities in particular location, and its reasoning methods answer queries, make predictions, and explain its conclusions. Users can interactively invoke these abilities, which should help them understand a model's behavior and improve it over time. We have carried out initial tests on cellular models of aging, using the environment's interactive character to identify problems in these models and repair them.

Although our approach draws on ideas developed in earlier work, it combines them in novel ways to support three key facets of the scientific enterprise: the formal representation of knowledge and hypotheses, relating that knowledge to observations through explicit reasoning, and the incremental development of knowledge over time. Many projects that formalize biological knowledge have focused on inert structures, rather than offering aids for reasoning over complex models, and most techniques for codifying knowledge rely on curators, rather than giving scientists tools to make their own changes. We believe our interactive environment offers a promising approach that addresses these issues in ways that biologists will find accessible and useful.

Acknowledgements

This research was supported in part by Grant CAA 0113-07 from Science Foundation Arizona and in part by Arizona State University. We thank John Furber for providing feedback about his aging network diagram, along with Durga Bidaye, Rick Chimera, Juraj Dzifcak, Seungchan Kim, Stephen Racunas, David Stracuzzi, and Michael Verdicchio for early contributions to the project.

References

Bredeweg, B., Bouwer, A., Jellema, J., Bertels, D., Linnebank, F., & Liem, J. (2007). Garp3 - A new workbench for qualitative reasoning and modelling. *Proceedings of the Fourth International Conference on Knowledge Capture* (pp. 183–184), Whistler, BC.

Bridewell, W. & Langley, P. (2010). Two kinds of knowledge in scientific discovery. *Topics in Cognitive Science*, *2*, 36–52.

Darden, L. (2006). *Reasoning in biological discoveries*. New York: Cambridge University Press.

Džeroski, S., Langley, P., & Todorovski, L. (2007). Computational discovery of scientific knowledge. In S. Džeroski & L. Todorovski (Eds.), *Computational discovery of communicable scientific knowledge*. Berlin.

Feldman, B. Z., Compton, P. J., & Smythe, G. A. (1989). Hypothesis testing: An appropriate task for knowledge-based systems. *Proceedings of the Fourth Knowledge Acquisition for Knowledge-based Systems Workshop*. Banff, Canada, October 1989.

Forbus, K. D. (1984). Qualitative process theory. *Artificial Intelligence*, *24*, 85–168.

Furber, J. (2009). Systems biology of human aging: Network model of biochemical and physiological interactions in human senescence. http://www.legendarypharma.com/chartbg.html.

Kanehisa, M. (1997). A database for post-genome analysis. *Trends in Genetics*, *13*, 375–376.

Karp, P., Riley, M., Saier, M., Paulsen, I., Paley, S., & Pellegrini-Toole, A. (2000). The EcoCyc and MetaCyc databases. *Nucleic Acids Research*, *28*, 56–59.

Mahidadia, A., & Compton, P. (2001). Assisting model-discovery in neuroendocrinology. *Proceedings of the Fourth International Conference on Discovery Science* (pp. 214–227). Washington, DC: Springer.

Racunas, S., Shah, N., Albert, I., & Fedoroff, N. (2004). HyBrow: A prototype system for computer-aided hypothesis evaluation. *Bioinformatics*, *20*, i257–264.

VanLehn, K. (2006). The behavior of tutoring systems. *International Journal of Artificial Intelligence in Education*, *16*, 227–265.

Vastrik, I., D'Eustachio, P., Schmidt, E., Joshi-Tope, G., Gopinath, G., Croft, D., de Bono, B., Gillespie, M., Jassal, B., Lewis, S., Matthews, L., Wu, G., Birney, E., & Stein, L. (2007). Reactome: A knowledge base of biologic pathways and processes. *Genome Biology*, *8*, R39.

Vijg, J., & Campisi, J. (2008). Puzzles, promises and a cure for ageing. *Nature*, *454*, 1065–1071.

Zupan, B., Bratko, I., Demsar, J., Juvan, P., Halter, J. A., Kuspa, A., & Shaulsky, G. (2003). GenePath: A system for automated construction of genetic networks from mutant data. *Bioinformatics*, *19*, 383–389.

On the limits of dynamic imagination: A mental extrapolation task

Florent Levillain (levillain@lutin-userlab.fr)
Laboratoire "Cognitions Humaine et Artificielle"
Université Paris 8 – 2 rue de la Liberté 93526 Saint-Denis France

Luca L. Bonatti (lucabonatti@mac.com)
ICREA and Universitat Pompeu Fabra
Carrier Roc Boronat, 138, 08018 Barcelona, Spain

Abstract

To mentally extrapolate the trajectory of a moving object which disappears from sight, it is possible to exploit two different sources of information. One source is the memory of the last visible movement of the object, and the other is its inferred movement through time. It is often assumed that these cues are integrated into dynamical analog mental representations. To investigate the nature of the mental representation of imagined movements, we used a new experimental paradigm for which a causality attribution task was combined with motion prediction task. Participants were instructed to imagine the trajectory of a moving object disappearing behind a screen while estimating the degree to which the movement was caused by another moving object.
We show that the predicted movement departs from a correct extrapolation based on accurate memory for velocity. Furthermore the mental representation of the physical and causal structure of the dynamical events did not appear to be as detailed as a theory of mental simulation would predict.

Keywords: mental imagery, prediction of motion, perception of causality.

Introduction

Correctly performing actions on moving objects typically requires a high level of accuracy. Tasks, such as hitting or catching a ball show that humans can accurately and consistently represent the timing of a visible moving object and anticipate its future positions (Regan, 1982).

However, when the stimulus is not visible, such as when it is temporarily occluded, it is not clear how precisely we can time a non visible movement and whether we possess an extrapolation mechanism that can time non visible displacements. Interception tasks are mostly driven by kinematic properties, whereas mental extrapolation may be more influenced by cognitive factors, particularly by how we represent the causal interactions of the objects within a scene.

In studies of mental imagery, it is putatively assumed that the mind builds analog representations that can be used ti estimate possible outcomes of dynamical events (Johnson-Laird, 1983) or to reveal spatial properties of objects (Kosslyn, 1994). Similarly, dynamical analog representations may subserve the ability to represent the timing structure of an invisible dynamical event (Shepard & Cooper, 1982; Schwartz, 1999).

Additionally, dynamical analog representations could integrate variables related to the physical structure of the environment. Results suggesting that humans are capable of recognizing physically correct object movements (Kaiser et al., 1992), along with findings showing that we can perceive high-level properties of these stimuli, such as their causal relations (Leslie, 1994), or agency status (Premack, 1990), support this possibility. Indeed, it has been claimed that internalizing invariant properties of the environment is evolutionarily adaptive (Hubbard, 1995; Shepard, 2001).

Thus, it is plausible to conjecture that information regarding the dynamic properties of a scene that we are capable of representing (for example, their causal relations, or the amount of physical forces acting upon an object) is integrated in a unique mental simulation. This being the case, such a dynamical representation may allow for accurate prediction of future states of invisible events. Alternatively, the prediction of motion and the representation of other forms of physical information may be independent, and hence not merged into a single optimal simulation of dynamical events. In the present article, we aim to determine the ability to accurately estimate motions of invisible objects and to clarify how participants integrate an intuitive causal understanding of the represented events into a mental representation of motion .

Experiment 1

Experiment one determined the accuracy for predicting the position of a moving object that is no longer visible. Participants were required to predict the time-to-arrival of an animated ball at different positions after its disappearance.

We also tested how the representation of causal relations influenced participants' accuracy for predicting invisible dynamical events. If the information used to compute the velocity of an object is integrated with the information used to compute the causal structure of the scene, we would expect that events considered as causally correct are predicted more precisely than events considered as causally anomalous. However, if the two kinds of information are processed separately, we should observe a dissociation between the accuracy of online predictions of imagined position and the perception of causal correctness.

In every experimental condition, there were two moving objects, a launcher and a target, the movement of the target behind the occluder was to be predicted, while the causal relation between the launcher and the target, which could vary both in spatial and temporal contiguity, was to be estimated.

Method

Participants. Nineteen randomly chosen participants completed the experiment (mean age = 24,4; range from 20 to 31 years).

Stimuli. We created video stimuli with the animation software Cinema4D. The animation clips used were created with some intent of realism. For example, objects' shadows cast on the ground, had slight grooves, offering some depth cues. In each clip, a white and a green ball moved onto an earth-ground, below a blue, cloudy sky. A red screen partially covered the movement of the green ball (see Figure 1).

After 1 s, a white ball with a 3° diameter appeared from one side of the scene, and travelled horizontally at a constant speed of either 25,8°/s, 19,3°/s or 12,9°/s toward a green ball, which was stationary at the centre of the scene.

The white ball (the launcher) either did or did not contact the green ball (the target), but the target always started its movement as fast as the launcher and in the same direction.

A red rectangular screen was positioned such that its border contacted the edge of the target and its length covered the entire trajectory of the target. After initiating its movement, the target continued its trajectory behind the screen, until the end of the animation segment. Three vertical black lines were drawn on the red screen, placed at six different positions, yielding two configurations (see Tables 1 and 2). The direction of the balls (movement to the right, or left) was balanced across trials.

Three different spatio-temporal conditions were implemented by either varying the spatial interval between the launcher and the target at the end of the launcher's movement, or by varying the delay between the end of the launchers' movement and the beginning of the target's movement.

Table 1: Angular speed and hypothetical arrival time (s) in bar configuration 1 (in parentheses, distance from the origin of each bar).

Bar Number	12.9°/s	19.3°/s	25.8°/s
1 (44.2°)	0.44	0.28	0.24
4 (60.8°)	1.72	1.16	0.88
6 (71.8°)	2.6	1.72	1.32

Table 2: Angular speed and hypothetical arrival time (s) in bar configuration 2 (in parentheses, distance from the origin of each bar).

Bar Number	12.9°/s	19.3°/s	25.8°/s
2 (49.7°)	0.88	0.56	0.44
3 (55.3°)	1.28	0.84	0.68
5 (66.3°)	2.16	1.4	1.12

In the **Contact** condition, the motion of the launcher immediately ceased after having contacted the target, and the target began to move immediately after contact with the launcher. Neither the launcher nor the target exhibited deformation as a result of contact.

In the **Delay** condition, an interval was introduced at the moment the two balls made contact. The interval was 480 ms for the first condition and 640 ms for the second condition.

In the **Space** condition, although the end of the movement of the launcher and the beginning of the movement of the target were simultaneous, the launcher stopped its trajectory before contacting the target. The space between the endpoint of the launcher's path and the target's starting position was determined according to the delays previously specified: the distance between the two balls was equal to the distance the launcher would have covered during the interval specified in the Delay condition had it continued its movement (a distance of 100 pixels for the first condition and 130 pixels for the second condition).

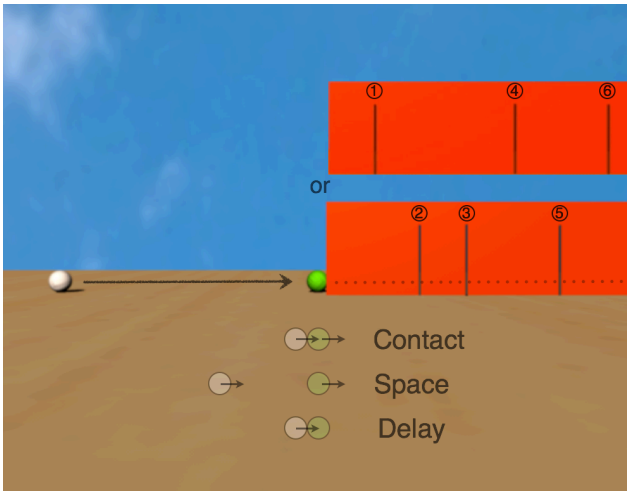


Figure 1. Overall sequence of events in a trial and causal conditions (Contact, Space, Delay).

To reveal the impact of occlusion on the time-to-arrival estimation, we designed another set of video sequences for which the target remained visible. However, in these clips the bars remained in the same positions as in the above described segments, and had the same spatio-temporal properties described previously. Also, the velocity of the target was constant and equal to the velocity of the launcher. Finally, in order to break the monotony resulting from the horizontal movements of the launcher, we intermixed the experimental animations with distractor segments in which the launcher fell from above, landing in the same position that the launcher stopped in the experimental sequences. Finally, for all the animations, the launcher appeared either from the left of the screen and moved right, or vice versa. Thus, in total, 120 experimental animations were created (5 conditions of interest crossed with the other experimental factors: Contact/Delay (x2)/Space (x2), target visibility (x2), bar configuration 1/2, speed (x3), direction of movement (x2) and 40 distractor animations).

A graded scale (from 0 to 9) was employed to collect participants' causal judgments for each clip.

Apparatus. Stimulus animations were displayed and the data were collected using a PowerMac G4 running the GNU software package PsyScope X (http://psy.cns.sissa.it). The animations were projected on a 200x135 cm screen with an Epson EMP 8100 projector. Reaction times were recorded using a Newmicros Button Box. This response box, together with a mouse and a numerical keypad were placed on a table positioned in front of participants.

Procedure. Participants sat in a darkened room, 2.5 meters from the screen. From their position, they could easily press the button box.

Each session began with a practice trial (Contact condition), with the velocity of the balls always set to 19.3°/s. Participants were instructed to visually track the launcher and, after the target disappeared behind the occluder, to press the key on the button box each time they felt the target would reach a bar on the red screen. They were encouraged not to press the key only three times. They were also informed that the balls would move at constant velocity and that they had identical speed. This information could be used to predict the position of the second ball on the basis of the speed of the first ball. Participants were able to move their head freely as they tracked the balls.

Participants were also informed that at the end of each segment a 1-to-9 scale would be projected on the screen. They were instructed to evaluate the perceived strength of the causal relation between the two balls by moving the mouse on the scale and clicking on the appropriate magnitude (1= not at all causal; 9 = completely causal). No explicit relation was drawn between the first online prediction and the second causal judgment tasks.

Each trial was initiated by pressing a button on the response box. The movement of the launcher started one second after the beginning of the scene. According to the velocity of the balls, the trial could last either 10, 11 or 12 seconds. At the end of the trial, a black screen, in which the causality scale appeared, filled the scene. After participants punched a number on a numerical keyboard, the next trial started. The beginning of the novel segment was controlled by participants. No feedback about response accuracy was given. Animation segments were presented in blocks of 80 (60 experimental, 20 distractors), arranged in a semi-random order, with the constraint that the same spatio-temporal condition could not be presented more than three times in a row. The first block contained only animations with occluded targets, and the second block contained the sequences with visible targets. The overall duration of the experiment was one hour, with a pause between the two blocks after thirty minutes.

Results

The mean timing error was computed as the difference between the total response time to a tested position from the beginning of the sequence and the total arrival time of the target, from the beginning of the sequence to the moment the target crossed a bar. The frame in which the invisible target ball reached each bar was determined offline, as the first frame where the target made contact with the bar. Thus, a positive error value indicates that participants entered their response after the target crossed the bar, while a negative error value indicates the response was given before the arrival time.

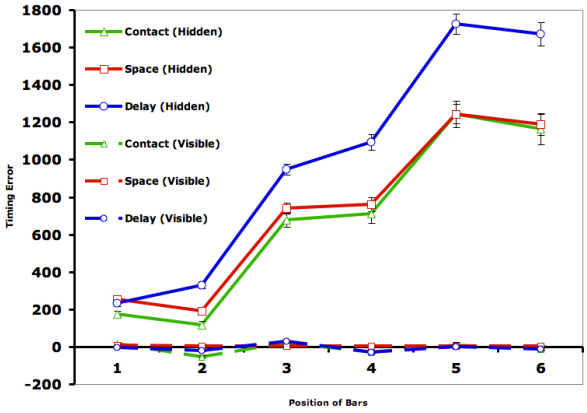


Figure 2: Mean timing error (in ms) for the main experimental conditions (Contact/Space/Delay), separated by target visibility. The horizontal axis indicates the position of the bars on the trajectory of the target.

Trials were excluded from analysis if participants did not press the button exactly three times, if they pressed the button before the disappearance of the target, or if reaction times exceeded 2.5 SD from the mean response time in the relevant conditions.

We initially analyzed how the visible and invisible conditions differ for each tested position. Figure 2 shows the time differences between participants' responses and arrival time of the target ball, plotting together the three tested positions of each experimental animation sequence (1,4,6 and 2,3,5). When the target was visible, participants were accurate at determining the exact moment of arrival. Not only does the result show participants' accuracy at predicting contacts with direct visual feedback, but it also reveals that the task of tracking three successive positions with the spatio-temporal parameters we tested is perfectly feasible. However, when the target was not visible, clearly the average prediction systematically overestimated the time of arrival of the target. A two-way repeated measures ANOVA with individual means of timing error as a dependent variable and object occlusion as independent variable (visible/non visible) reveals an effect of occlusion on timing error ($F_{1,18} = 78.28$, $p < 0.0001$). The error was positive at every tested position, indicating an overestimation of the time needed for the target to cover the distance between the different bars. This response delay increased with occlusion time, but neither linearly nor continuously, as a simulation hypothesis would predict. In fact, the timing error did not differ between any two close bar pairs tested in the two bar position conditions. In other words, participants could not distinguish between any two close positions (as confirmed by post-hoc t-tests with Bonferroni alpha adjustment) when tested separately, but only in the invisible target condition, suggesting that the ability to predict the position of an invisible target is, if at all present, rather coarse.

In successive analyses, we thus collapsed the two bar position conditions, which did not differ. A two-way repeated measures ANOVA, restricted to the invisible target condition, with the sequence of response as independent variable (R1/R2/R3) revealed an effect of response order (F2,18 = 109.04, $p < 0.0001$). For each response, the timing delay increased as confirmed by a post-hoc t-tests (Bonferroni alpha adjustment) on the differences between timing errors in the third, second, and first response (**R2 – R1** = 631.02, $p < 0.0001$; **R3 – R1** = 1348.62, $p < 0.0001$).

Effect of causal conditions on timing accuracy. Figure 2 shows the timing error at each tested bar for the main conditions of the experiment. Timing was not accurate in any of the three tested conditions compared with the error values obtained in the visible movement conditions. There were also differences observed within the three conditions in the target-occluded animations. A two-way repeated measures ANOVA performed on individual error means for the invisible target conditions, with the type of interaction between the balls as the independent variable, indicated a significant effect of the three causal conditions ($F_{2, 18}$ = 84.43, p <0.0001). This main effect depended on the difference between the Contact and the Delay conditions (post-hoc, Bonferroni adjusted t-tests : **Delay - Contact** = 337.84, p <0.0001) and between the Space and Delay conditions (post-hoc, Bonferroni adjusted t-tests : **Delay – Space** = 293.14, p <0.0001), while there was no difference between the Contact and Space condition (post-hoc, Bonferroni adjusted t-tests : **Space – Contact** = 44.7, p = 0.53).
Thus, the Delay condition reveals a timing error that is not only greater than the (almost null) error in the corresponding visible condition, but also greater than both non-visible conditions. Instead, timing errors in the Contact and Space conditions remained relatively similar across the trajectory of the target.

The effect of the spatio-temporal conditions did not depend on the size of the temporal or spatial intervals, as indicated by the lack of interaction between the two factors in a two-way ANOVA restricted to the responses in the Space and Delay conditions ($F_{1, 18}$ = 0.02, p =0.89).

Effect of causal conditions on causal attribution. We analyzed participants' estimates of the causal strength of the scenes. A two-way repeated measures ANOVA with causal attributions as the dependent variable and interval size and causal conditions as independent variables revealed no effect of interval size ($F_{1, 18}$ = 3.39, p = 0.08) and no interaction between size and conditions (F1, 18 = 0.66, p = 0.43). Thus, we collapsed the data across the interval dimension in further analyses.

A two-way repeated measures ANOVA yielded a significant effect of the type of interval introduced (F2, 18 = 66.99, $p < 0.0001$). The effect was mainly carried by the difference between the Contact condition and Space conditions, but all conditions were different (post-hoc, Bonferroni adjusted t-tests : **Contact - Space** = 6.23, p < 0.0001; **Contact - Delay** = 4.52, p=0.03; **Delay - Space** = 2.18 , p <0.01).

As expected, in the Contact condition the relation between the launcher and target was considered to be the causally strongest. Instead in the Space condition causality was considered non existent. Noticeably, causal interaction in the Delay condition was judged higher than in the Space condition. Combining such results with the prediction task, and comparing the two conditions in which causal violations were introduced, one can see that participants were better at predicting the position of an invisible target in the condition (Space) that was judged causally weaker than the other (Time). That is, prediction abilities and perception of causality do not align.

Effect of expertise on timing accuracy. Because many of our participants were highly skilled in physics and had a thorough understanding of real kinematics, we also checked whether expertise had any effect on accuracy. We divided the total number of participants in three groups based on the number of years they received physics education (naive: up to middle school; intermediate: up to high school; high: Masters and Ph.D in Physics).

Overall, expertise had no effect on prediction accuracy, as revealed by a two-way repeated measures ANOVA with individual means of timing error as a dependent variable and levels of expertise as independent variable (naive/ intermediate/high) ($F_{2, 18}$ = 0.35, p = 0.71). Nor did any effect appear when causal attributions were the dependent variable ($F_{2, 18}$ = 1.35, p = 0.29). Expertise did not interact with spatio-temporal conditions in either predictive accuracy or causal attributions.

Discussion

Experiment one revealed that participants were highly accurate when predicting the time of contact of a moving target when the target was continuously visible, regardless of the type of interaction with the launcher. Yet, they were highly inaccurate when the target moved behind an occluder, making errors as high as 70% of the duration of the full scene. Furthermore, the amount of overestimation did not appear to increase continuously as the distance of the arrival point increased, revealing a sort of quantization of the error that is difficult to reconcile with a simulation theory of imagined movement.

This overestimation is difficult to explain by the violations of causal interactions in the events presented, as a large overestimation error was also present when the events were causally correct (Contact condition). Although we cannot be certain that, at a perceptive level, the computation of causal interactions does not interfere with the prediction, we found that the attributions of causality were dissociated from prediction accuracy: participants were better at predicting the position of an unseen object in conditions that they judged causally worse.

It is thus more likely that the variations in the amplitude of timing error have a source in the time necessary to integrate the two successive movements at a purely kinematic level. As such, this experimental situation might reveal particularly interesting in the exploration of movements integration.

Overall, these results suggest that our ability to predict future states in a partially occluded dynamical event is severely limited and probably does not integrate our knowledge about causal interactions.

Experiment 2

An alternative explanation for the large delays observed in experiment one, which maintains the tenet that humans simulate physical events, could be that participants do simulate physical events, but they do it even better than required. We showed animation sequences in which balls rolled over flat terrain. If participants integrate real physical constraints they may not avoid considering friction in their simulation, thus 'mentally slowing down' the speed of an unseen object. This integration of a physical variable might explain why participants delayed their reactions in imagination. Although the size of the delays we found is not easily reconciled with a simple integration of real friction parameters given the terrain in our videos, the point remains valid. Indeed Hubbard (1995) suggested mental analogs of gravity and friction are directly integrated in our simulations of object motion, systematically biasing certain position estimations. So the time-to-arrival overestimation in our experiment could reflect the fact that participants are simulating a deceleration instead of using their memory of a constant velocity.

We tested this possibility by modifying the context of the previous sequences, so as to prime certain physical representations. Specifically, we tilted the slope of the track such that the balls would either roll downwards or upwards. A previous study has shown that such a transformation can bias memory for position in a representational momentum paradigm (Bertamini, 1993).

If the prediction is indeed driven by inferred dynamical properties, we expect the timing error to be modified according to the orientation of the slope. If, instead, the prediction is not affected by the integration of physical variables and the error we found in Experiment one was due to limits in how we can simulate physical events (if we have such an ability), then we expect the timing error to persist unaffected by the conditions of Experiment two.

Method

Participants. Thirteen randomly chosen participants were recruited for the experiment**.** Their ages ranged from 19 to 30 years (mean age = 22,9).

Stimuli. We used the animations with occluded target movement from Experiment one, but modified such that the slope of the track was altered by rotating the images 20° either clockwise or counterclockwise. Thus, in experiment two there were three groups of animation stimuli: two containing balls rolling on an inclined plane, and a third group containing the same sequences used in Experiment one, with the balls rolling on a horizontal plane. This configuration allowed us to determine how gravity modifies the results of Experiment one. No vertical movement distractor was present in this experiment.

Apparatus and Procedure. The same set-up and procedure used in Experiment one were used for Experiment two, with the exception that we did not run the visible target condition, as performance in this condition was previously shown to be accurate.

Results

Data exclusion criteria and error calculations were as in Experiment one.

Effect of the orientation of the slope on timing accuracy. Figure 3 represents the variation of timing error as a function of slope. As in Experiment one, an overestimation of the time-to-arrival, increasing with response order, was observed. There was no obvious difference in timing accuracy between the different slope conditions, although a slight decrease in timing error appeared in the slope downward condition.

A two-ways repeated measures ANOVA performed on the timing error, with speed and slope as independent variables, revealed a main effect of slope ($F2, 12 = 4.61$, $p = 0.02$). Post-hoc t-tests with Bonferroni adjustment revealed that the effect was carried by the difference between the Slope up and Slope down conditions (**Slope up - Slope down** = 175.34, p = 0.02), whereas no difference was found between the two tilted conditions and the horizontal condition. Furthermore, the difference between these conditions only occurred at one speed. Indeed, speed and slope interacted ($F4, 12 = 2,87$, $p = 0.03$); post hoc analyses showed that the difference between Slope up and Slope down was significant only when the balls moved at 19°/s (post-hoc Bonferroni-adjusted t-tests: 19°/s **Slope up – Slope down** = 300,86, $p < 0.01$; 13°/s **Slope up – Slope down**= 102,38, $p = 0.99$; 26°/s **Slope up – Slope down** = 122,78, $p = 0.94$).

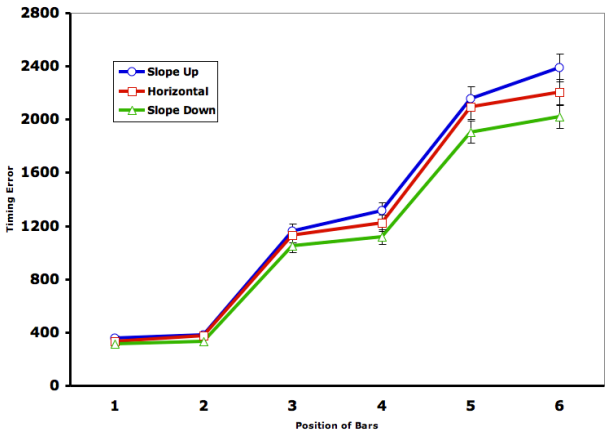


Figure 3. Mean timing error (ms) for the three plane rotations (Slope up, Slope Down, Horizontal). The horizontal axis indicates the position of the bars on the trajectory of the target.

Discussion

In this experiment, we tested whether the prediction of the position of an unseen object was influenced by the

integration of physical variables in a mental simulation of the dynamic of the action.

We observed a slight effect of slope on participants' predictions, but only at one velocity. While the effect is compatible with the mental simulation of physical parameters, it remains mysterious as to why it should occur only in the 19°/s velocity condition. Thus, overall, it is difficult to entirely reconcile the results with the assumption that our mental models faithfully simulate the dynamics of object movements.

General Discussion

How does the cognitive system deal with incomplete information about the trajectory of a moving object? Research on mental imagery and on the prediction of motion frequently appeals to mental analog representations as a potential substrate for spatial computations and dynamics understanding. Here we provided evidence that this conception may not offer an adequate account of how we represent dynamic stimuli.

We devised a task for motion prediction that directly probed participants' ability to estimate the position of a moving object online, as opposed to other known paradigms of motion prediction which test memory for past positions rather than fast prediction of future positions (e.g., Hubbard, 1995). With this task, we demonstrated that estimations of time-to-arrival are inaccurate, with a large overestimation of the time necessary for the target to reach a position (confirming and expanding upon previous results obtained with different paradigms; e.g., Gilden Blake & Hurst, 1995),. This result supports the claim that there is no predictive mechanism to estimate an object's position when it is occluded, when a direct visual evidence is lacking (Keane & Pylyshyn, 2006).

Furthermore, by coupling this task with causal strength judgment task, we showed that intuitive perceptions of causality do not integrate with online prediction of imagined object movements, casting further doubts on the existence of a representation that integrates physical variables into an analog simulation of objects and physical forces in the world. Finally, we showed that the system responsible for the overestimation error we revealed, takes into account very obvious physical properties, such as gravity, only haphazardly. This aspect of our results is difficult to reconcile with evolutionary accounts of cognition, according to which integration of gravity should be a prime candidate for a variable that evolutionary history may have embodied into a mental simulator.

How then can we account for the overestimation error we observed? Some studies suggest that when we track a moving object, our time perception for rapidly moving stimuli is lengthened as compared to static stationary stimuli (Brown, 1995; Kanai et al., 2006). Such a phenomenon could account in part for the present results, and as a consequence it could indicate that rather than extrapolating object position by means of an analog mental simulation of real physical forces, we use an internal clock to make an only coarse estimate of when an invisible object should be at a given location.

As a general conclusion, our results point toward the existence of several independent systems, one of which may compute object velocity, and another that may compute causal relations in the world. Although it may be tempting to unite the two kinds of systems, our results cast doubt on the existence of a common substrate for the extrapolation of trajectories in dynamical sequences of movements. These results also cast doubts on the existence of richly detailed analog representations that could assist us in knowing and understanding the physical world.

Acknowledgments

Our thanks to J. Mehler and the members of the Language, Cognition & Development lab in Trieste for helping us in designing and testing our experimental paradigm. This research was supported by a MICINN PSI2009-08232 grant to L.B.

References

Bertamini, M. (1993). Memory for position and dynamic representations. *Memory & Cognition, 21*, 449-457.

Brown, S. W. (1995). Time, change, and motion: The effects of stimulus movement on temporal perception. *Perception & Psychophysics, 57*, 105-116.

Gilden, D., Blake, R, & Hurst, G. (1995), Neural adaptation of imaginary visual motion. *Cognitive Psychology, 28,* 1-16.

Hubbard, T. L. (1995). Environmental invariants in the representation of motion: Implied dynamics and representation momentum, gravity, friction and centripetal force. *Psychonomic Bulletin & Review, 2*, 322-338.

Johnson-Laird, P. N. (1963). *Mental Models: Towards a cognitive science of language, inference, and consciousness.* Cambridge University Press.

Kaiser, M. K., Proffitt, D. R., & Anderson, K. A. (1985). Judgments of natural and anomalous trajectories in the presence and absence of motion. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11*, 795-803.

Kanai, R., Paffen, C. L. E., Hogendoorn, H., & Verstraten, F. A. J. (2006). Time dilation in dynamic visual display. *Journal of Vision, 6*, 1421-1430.

Keane, B., P., & Pylyshyn, Z., W. (2006). *Is motion extrapolation employed in multiple object tracking? Tracking as a low-level, non-predictive function*. Cognitive Psychology, 52(4), 346-368.

Kosslyn, S.M. (1996). *Image and Brain: The resolution of the imagery debate*. MIT Press, London, England.

Leslie, A. M. (1994). ToMM, ToBy, and agency: core architecture and domain specificity. In L. Hirschfeld & S. Gelman (Eds.), *Mapping the Mind; Domain specificity in cognition and culture*. Cambridge University Press.

Premack, D. (1990). The infant's theory of self-propelled objects. *Cognition, 36*, 1-16.

Regan, D. (1992). Visual judgments and misjudgments in cricket and the art of flight. *Perception, 21*, 91-115.

Schwartz, D. L. (1999). Physical imagery: Kinematic versus dynamic models. *Cognitive Psychology, 38*, 433-464.

Shepard, R. N. (2001). Perceptual-cognitive universals as reflections of the world. *Behavioral and Brain Sciences, 24*, 581-601.
Shepard, R. N., & Cooper, L. (1982). *Mental Images and Their Transformations*. MIT Press.

(Category) Learning by Doing: How Goal Directed Tasks Constrain Conceptual Acquisition

Seth Chin-Parker (chinparkers@denison.edu)
Department of Psychology, Denison University
Granville, OH 43023 USA

Abstract

The current study explores conceptual acquisition that occurs as the result of completing a task in a novel domain. The items encountered in the domain were complex in that there were multiple sources of information that might be used to organize conceptual knowledge related to the domain. I test the hypothesis that goal-directed interactions will constrain the acquisition of knowledge such that functional categories of the items, organized around goal-relevant features, are learned. Converging evidence from two measures provided strong support for the idea that participants organized their knowledge of the domain in terms of goal-relevant features, and the conceptual organization was able to support both the completion of the task and subsequent categorization tasks.

Keywords: category learning, goals, similarity.

Prior experience underlies intelligent behavior – people learn through interactions with the environment what behaviors lead to successful outcomes and what ones do not. An important component of this is recognizing categories of events and items among those experiences, a process that leads to the acquisition of conceptual knowledge, knowledge of those categories. That knowledge can be used to categorize, communicate, reason, and problem solve at later points. A fundamental question then is how coherent categories of items are identified so that the conceptual knowledge can be appropriately applied. Theories of categorization address what ties together items within a category and subsequently coheres the conceptual organization that reflects those categories, and most theories rely on a notion of similarity for at least a component of that cohesion (Hahn & Ramscar, 2001). Thus, the question shifts to how this similarity is determined.

There have been two basic approaches to answering this question (Malt, 1995). The first assumes that the environment constrains the similarity. Rosch et al. (1976) nicely capture this idea by positing that features of items in the world occur in reliable clusters and the conceptual system learns to recognize that structure. They go so far as to illustrate in one "experiment" (1976, Exp. 3) how overlaying tracings of the members of various basic level categories (e.g. cat, shoe, truck) results in greater perceptual overlap within-category than across categories. Similarly, others (e.g. Anderson, 1991) have stressed the role of the environment in determining conceptual structure. That view can be contrasted with one that places much of the emphasis on the individual to constrain the similarity. Murphy and Medin (1985) argued information from the environment has to be situated within the knowledge structures (e.g. inter-category relations and theories) that the individual brings to any interaction with the environment. In this way, the structure of conceptual knowledge is constructed as the individual interprets what is to be considered feature information and how those features relate to one another.

Although most researchers interested in concepts and categories stake out some middle ground in this debate, much of the work in human category learning assumes that the environment provides structure. This assumption has seemingly created a disconnect between work exploring more naturalistic concepts and the basic experimental work examining conceptual acquisition (Murphy, 2005). I identify two critical differences between basic experimental studies and more naturalistic ones and explore them in the current study. First, in most experimental work, the categories are well defined in terms of their features and structure. Second, participants interact with members of those categories with the goal of differentiating the items they encounter based on that structure. In more naturalistic studies, the presence of the categorical structure is less clear and people interact with the items not with the goal of classifying items, but with the goal of accomplishing some other task. I present a study that incorporates a more complex, arguably more naturalistic, structure and vary the interactions that participants have with the items. In this manner, I examine how goal-directed behaviors within a domain affect the structure of the conceptual knowledge acquired about items within that domain.

I begin this research with the assumption that the environment is not a source of simple, unambiguous information about the categories that exist. For instance, people are able to recognize and use information about the taxonomic categories of food items, e.g. breads and vegetables, but they also recognize and use goal-related categories, e.g. snack foods and breakfast foods, to guide inferences and determine appropriate groupings of foods (Ross & Murphy, 1999). Similarly, people can identify and use ad-hoc categories (Barsalou, 1991) to guide behaviors. A study by Medin, Lynch, Coley, and Atran (1997) illustrates how this complexity can be reflected in conceptual knowledge. The experimenters asked various tree experts to sort cards labeled with tree names into groups. Those experts concerned with research and teaching tended to create groups that were highly correlated with the biological taxonomy, but landscapers tended to create groups that reflected the way the trees would be incorporated into landscaping decisions (e.g. a shade tree versus a weed tree). These cross-classifications and the development of ad-hoc categories of items are problematic

for an account that posits that the environment alone provides structure to our conceptual knowledge (although see Anderson, 1991, for a rebuttal). Although we have evidence of the complexity of naturalistic categories, the structure of the categories used in basic experimental work does not reflect this complexity (Murphy, 2005). Most items that comprise the categories are defined by specific feature lists or simple visual features, and the relation of the features to the categories is also carefully controlled. This typically results in a structure with only one "correct" organization for the items. Instead of operating in a situation with multiple possible configurations, participants in experimental studies are placed into a situation that is much more constrained by the information available.

Within these experimental studies, the interactions that people have with the categories are also different from what occurs in more naturalistic situations. In a typical category learning study, an item is presented, the participant predicts the category membership of the item, and feedback is given on the classification judgment. This approach has produced a great deal of information about how people learn to classify items, but may not capture important aspects of how people learn about categories in more naturalistic situations (Ross, Chin-Parker, & Diaz, 2005). Numerous studies have shown that classification learning promotes a near exclusive focus on diagnostic information, the features that distinguish the categories (Chin-parker & Ross, 2004; Rehder & Hoffman, 2005), but it is not apparent whether other means of category learning share this restricted focus (e.g. Minda & Ross, 2004). Arguably, the interactions we have within more naturalistic contexts are more varied and richer than the classification decisions made in a typical experimental setting. Importantly, I note that these interactions occur not with the primary intention to learn about the categories but rather to accomplish some other goal. The importance of goal-directed interactions has been explored by a range of cognitive scientists (e.g. Ram & Leake, 1995), and goals are implicated to some extent in how we come to recognize structure in the environment (Love, 2005). For instance, the naturalistic studies mentioned prior (e.g. Medin, Lynch, Coley, & Atran, 1997) suggest goal-directed interactions give rise to conceptual organizations that are able to support those interactions.

So, it seems that we can begin to bridge the chasm between experimental and naturalistic study of concept acquisition by adopting more complex categorical structures and varying the goals of the participants as they interact with the items that comprise those categories. Recently, Jee and Wiley (2007) did just that. They had participants learn about creatures that could be distinguished in terms of their perceptual features, shown through simple line drawings, or the nutritional value and ability to avoid predators (information about these features was conveyed through a list of features located beneath the picture). In their study, participants initially organized items in terms of similarity of the simple perceptual features, but as they learned about the domain and interacted with items, they either learned to identify the nutritional value or how the creature avoided predators, the goal-relevant information came to be important within the conceptual organization. Subsequent transfer tasks showed that a participant adopted a conceptual organization that reflected the information that was critical to their interactions within the domain, and the participants' similarity judgments were shaped by the presence of that information. Their study provides more clear evidence that the goal-directed interactions caused the shift in the conceptual structure.

In the current study, I examine category learning that occurs as the result of goal-directed interactions with items. Like Jee and Wiley (2007), I have a complex structure and participants interact with the items in accordance with different goals. In our study, the items are Flux Capacitor Boards, actual physical boards with various electrical components (non-functioning) affixed to them. As is described below, I created the boards so that there were two types of the boards that the participant would encounter during their initial task. However, only the classification participants were informed that these categories existed; the other participants were simply asked to complete their assigned task with the boards. Our primary hypothesis is that the conceptual organization adopted by the participants will be organized around the features of the boards that are relevant to the attainment of their goal.

As is described below, the *goal-relevant features* of the boards varied across the conditions. In one condition, the goal-relevant features are the configuration of specific components of the boards. In another condition, the goal-relevant features are relationships that exist between the components of the boards. For the classification condition, there were several possible sources of information that would be considered goal-relevant, or diagnostic. As noted in Jee and Wiley (2007), working towards a specific goal can often lead to information that is not goal-relevant to be left out of the conceptual organization. In this study, I expect that the two conditions with specific goal-relevant information will focus exclusively on that information, like classification learners in previous studies (e.g. Chin-Parker & Ross, 2004). Interestingly, since the classification condition will have multiple sources of information relevant to differentiating the categories, I predict that they will show a more general knowledge of the boards. I have no strong prediction as to whether the difference in the kind of goal-relevant information available to the two non-classification task conditions will affect the participants' acquisition of useful conceptual knowledge.

Experiment

Methods

Participants and Design Fifty-seven participants were randomly assigned to three experimental conditions: 18 participants were assigned to the *flexible condition*, 19 to the *solid condition*, and 20 to the *classification condition*. Two participants in the classification condition failed to show evidence of learning during their initial task, so their data

were removed from all analyses. All participants interacted with the same set of items during the initial task and completed the same two transfer tasks[1]. The presentation order of items during the initial and transfer tasks was randomized for each participant.

Materials and Procedure The primary materials for the study consisted of the Flux Capacitor boards and the connectors used to complete the boards. Each board had a series of nine terminal posts and various electrical components affixed to the board (see Figure 1). The posts were organized into three sets: One set in the upper, left-hand region of the board, one in the middle region, and one in the lower, right-hand region. The other components were placed around these posts according to the parameters described below. The boards were designed so that there were two types of boards that the participants encountered during the initial task, and variations of these two types of boards were created for the transfer tasks. During the initial task, participants in the flexible and solid conditions were given connectors that they placed onto the terminal posts to complete each board. The participants in the classification condition did not use connectors during their initial task

In the flexible condition, the connectors were made of wire and varied in terms of how they fit onto the terminal posts: The connector either fit over an open post or was inserted into a hole drilled into the "capped and drilled" post. As can be seen in Figure 1, each set of terminal posts in the Type A boards featured one post that has been capped and drilled and two posts that were open. In contrast, the Type B boards featured sets consisting of two capped and drilled posts and one open post. The configuration of the posts is considered to be the goal-relevant feature for the flexible condition because they constrain how the flexible connectors can be placed onto the board.

In the solid condition, the connectors were made of inflexible aluminum pieces, and the placement of these connectors was constrained by the presence of components situated near the terminal posts. For the Type A boards, the connectors had to go between components. For the Type B boards, the connectors had to go around the components. Thus, the relationship of the components to the posts is considered the goal-relevant feature for the solid condition because it constrained how the solid connectors could be placed onto the board.

The electrical components were unique to each board. However, each of the boards featured a perceptually salient *correlated component*. The correlated component for the Type A boards was a two-inch section of a computer memory module placed in the near left corner, and the correlated component for the Type B boards was a stack of silver clips with copper wire loops placed in the far right corner. These components were not implicated in how the connectors in either condition could be placed onto the board but were perfectly diagnostic of the two board types.

During the initial task, participants in the solid and flexible conditions were asked to complete the boards by placing three of the six connectors onto the terminal posts. The participants in the classification condition were told that the boards were incomplete and before they could be completed they needed to be identified as "positive flux" or "negative flux" boards. The classification participants were instructed to learn how to identify the two types of boards. There were eight boards used in the initial task phase, half were Type A and half were Type B. Each participant encountered each board twice during this phase.

In the classification condition, a board was placed into the holder, and after the participant responded with either "positive flux" (correct for the Type A boards) or "negative flux" (correct for the Type B boards), the experimenter provided feedback about the classification and allowed the participant to study the board. In the solid and flexible conditions, the experimenter placed a board into the holder, and the participant determined which connectors to place

Figure 1: Example problem boards from Experiment 1



Notes: The boards on the left are Type A boards, and the boards on the right are Type B boards. The top images show boards with no operators present. The center images show boards completed with the flexible connectors. The bottom images show boards completed with the solid connectors.

[1] Participants also completed a sorting task following the same-different task. However, the classification condition was inadvertently given different instructions for the task, so we are unable to compare performance across the groups. The results of the task very closely tracked those of the same-different task.

onto the board. After each trial, the board was removed from sight and a new board was placed into the holder.

After the initial task, all participants completed the same two transfer tasks. The materials for the transfer tasks consisted of photographs of flux capacitor boards the participants had not encountered during the initial task. The boards in the images were designed so that they varied in terms of how they related to the Type A/Type B board distinction that had been present during the initial task. No feedback was given to participants as they completed the transfer tasks.

First, the participants completed the *same-different task*. During each trial, the participant was presented with images of two boards affixed to a piece of paper. She was asked to indicate whether she would consider the two boards pictured to be the same type or different types. Across the sixteen items in the same-different task, I balanced whether the boards matched or mismatched in terms of the goal-relevant features. Eight of the pairs of boards maintained the same structure as the initial tasks boards; four of those pairs matched and four mismatched. All participants regardless of condition should identify the matches as the same and the mismatches as different if they picked up on any of the sources of information that differentiated the Type A and Type B boards during the initial task. The other eight boards were designed so that the goal-relevant features from the solid and flexible conditions were placed into opposition. For instance, if the goal-relevant features for the flexible condition matched what had been seen on the Type A board, the goal-relevant features for the solid condition would match what had been seen on the Type B board. Four of these board pairs were designed so that flexible condition goal-relevant features matched while the solid condition goal-relevant features mismatched. The other four board pairs were designed so the flexible condition goal-relevant features mismatched while the solid condition goal-relevant features matched.

The *category goodness-rating task* was the final task. I balanced whether the Type A or Type B boards were rated first. The participant was first shown a target board, one of the boards solved during the initial task phase, and was told that the board was either an "X-12" (Type A) or "G-59" (Type B) board. She was asked to rate each subsequent board shown in terms of the category indicated by the target board on a scale from one ("excellent example of this board type") to nine ("not this type of board"); also anchored at three ("good example of this board type"), five ("ok example of this board type"), and seven ("poor example of this board type"). After the participant studied the target board for a minute, it was removed, and the items for the goodness-rating task were shown to the participant one at a time. There were five types of boards pictured in the stimuli for this task, and the participant rated two of each type for each of the categories. The *category consistent* boards were structurally identical to the target board. The *category inconsistent* boards were structured like the other type of board; so if the target board was a Type A board, the category inconsistent board was a Type B board. The *correlation violation* boards were the same type of board as the target board, but the correlated feature was replaced by a small, perceptually dissimilar component. The *flexible violation* boards were of the same type as the target board, but were altered so the flexible connectors would not fit onto the posts. The *solid violation* boards were also of the same type as the target board, but they were altered so the solid connectors would not fit. Once the participant completed rating the ten boards for the first type, the target board for the second type was shown to the participant, and the task repeated for the second type.

Results

In the same-different task (Table 1), I found strong evidence that the participants in the flexible and solid conditions organized their knowledge of the domain in terms of the goal-relevant features. Across all items in the task, both the flexible condition, $M = 0.95$, $SD = 0.13$, $t(17) = 14.72$, $p < 0.001$, and the solid condition, $M = 0.87$, $SD = 0.18$, $t(18) = 8.64$, $p < 0.001$, were above chance performance in terms of assigning the pairs as the same or different in terms of the goal-relevant features for their conditions. The difference between the flexible and solid conditions was not significant, $t(35) = 1.57$, $p = 0.12$. A similar summarization of the results for the classification condition is not possible because there was no a priori prediction of how the classification participants would handle the items when the two goal-relevant features were placed in opposition. However, as can be seen in Table 1, when both of the goal-relevant features matched, they considered the boards as the same, and when both did not match, they considered the boards as different. When the goal-relevant features for the flexible and solid conditions were put into opposition (as in the "Flex + / Solid -" and "Flex - / Solid +" items), the participants in the classification condition did not show a preference for one source of information over the other as a group. Within the

Table 1: Proportion of Items (standard deviation) Identified as "the Same" in the Same-Different Task

	Relation of Boards in the Pair			
Condition	Flex + Solid +	Flex - Solid -	Flex + Solid -	Flex - Solid +
Flexible	0.94 (*0.24*)	0.01 (*0.06*)	0.90 (*0.26*)	0.04 (*0.18*)
Solid	0.86 (*0.21*)	0.11 (*0.23*)	0.15 (*0.29*)	0.86 (*0.21*)
Classification	0.83 (*0.33*)	0.19 (*0.24*)	0.46 (*0.39*)	0.36 (*0.36*)

Notes: For each item, the boards pictured either matched in terms of the goal-relevant features of the flexible (Flex +) or solid (Solid +) conditions or mismatched in terms of those features of the flexible (Flex -) or solid (Solid -) conditions.

classification condition, five participants had a pattern of response that indicated that they were using information about the goal-relevant features for the flexible condition, four participants appeared to be using information about the goal-relevant features for the solid condition, and nine participants had a pattern of responding that did not clearly indicate a preference for either source of information.

The category-goodness rating task provided a more specific indication of what information from the domain was being used by the participants in each condition. The data from the task (Figure 2) were analyzed using a series of ANOVAs. I report the results of five one-way ANOVAS that compared the ratings for each items type across the experimental conditions. I also include relevant within-condition comparisons where appropriate (full analyses are not included due to space restrictions).

There were no differences as to how participants in the three conditions rated the category consistent items, $F(2, 54) = 0.01$, $MSE = 0.02$, $p = 0.99$, but there were significant differences within all the other item types. Participants in all conditions rated the category inconsistent items as less good category members compared to all other items. Also, within the ratings for the category inconsistent items, there were some differences between the conditions, $F(2, 54) = 3.41$, $MSE = 1.27$, $p = 0.04$, primarily between the classification and flexible conditions. The ratings of the correlation violation items also varied by condition, $F(2, 54) = 3.72$, $MSE = 3.45$, $p = 0.03$. The classification condition rated these items as significantly worse category members than the category consistent items ($p < 0.01$) but the other two conditions did not. There were significant differences in the ratings of both the flexible violation items, $F(2, 54) = 21.62$, $MSE = 3.69$, $p < 0.01$, and the solid violation items, $F(2, 54) = 40.51$, $MSE = 3.04$, $p < 0.01$. As predicted, the participants in the solid condition rated the solid violation items as significantly worse than the category consistent items ($p < 0.01$), but did not rate the flexible violation items differently than the category consistent items ($p = 0.54$). The participants in the flexible condition rated the flexible violation items as significantly worse category members than the category consistent items ($p < 0.01$), but not the solid violation items ($p = 0.83$). The classification condition did not rate the solid violation items as significantly worse than the category consistent items ($p = 0.13$), but did rate the flexible violation items as worse category members ($p = 0.02$).

Figure 2: Mean Category-Goodness Ratings by Condition and Item Type

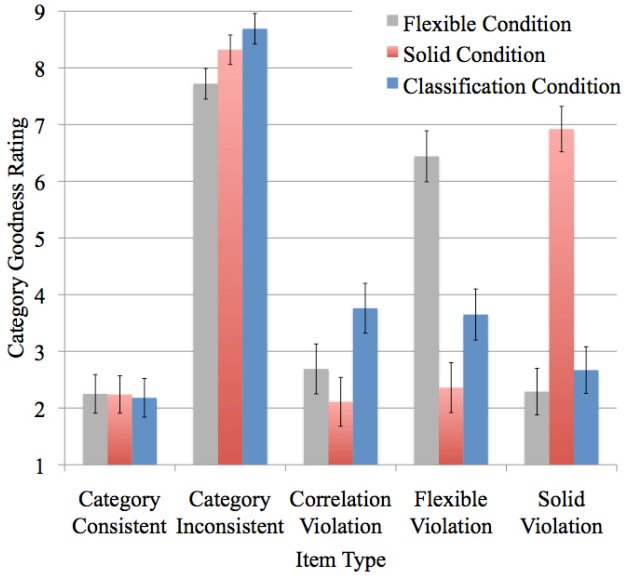


Figure Note: The category-goodness rating scale ranged from one ("excellent example of this board type") to nine ("not this type of board").

Discussion

I return to the question regarding what constrains the similarity underlying conceptual organization. The results of the participants in the flexible and solid conditions clearly show that the goal-relevant features are central to their notion of similarity for the items and thus critical for the organization of categories of boards within the domain. They identify novel boards as the "same" when they match in terms of the goal-relevant features and "different" when those features do not match. They also show a pattern of category goodness ratings that indicates that violating those goal-relevant features makes the boards less good members of the category while violating other sources of information have little or no effect on those judgments. The participants in the classification condition seem to maintain a more diffuse attentional focus during the initial task. This is interesting given earlier studies that show a very narrow focus for classification learners. However, these results fit well together when we consider that the goal of classification learning is to predict the category membership of items. Typically only a subset of the information available within the experimental materials allows those judgments to be accurately made, so the classification learner attends most to that subset of information. In this study, multiple sources of diagnostic information existed, so the classification learners maintained a correspondingly wide attentional focus. It was the participants in the solid and flexible conditions that had a narrow focus in this experiment, and this was due to the fact that only a subset of the information available within the domain was goal-relevant for each condition.

In one sense the results of this study are not surprising – there are numerous models of learning that incorporate an attentional mechanism (as discussed in Kruschke, 2003) to account for shifts across the information available during learning. Although attention obviously plays a critical role in the learning, it is not a sufficient determinant of learning; we need to understand what drives the attention. I propose that attention is guided by comparisons between the boards as the participants interact with them in terms of the goal they have (how to place the connectors or classify the board), and those features that are relevant to the person's

goal are picked out during the comparison process (e.g. Medin, Goldstone, & Gentner, 1993). It is important to note that this approach helps to explain how "simple" perceptual features (e.g. the capped and drilled posts) and relational information (e.g. how components were positioned with regard to the posts) can both be picked up and used in grounding similarity.

One critical question is whether the participants in this study were *really* engaged in category learning, or to put it another way, did they *really* recognize categories of boards during the initial task? In the typical classification learning paradigm, this question is deflected because the participants explicitly know of the presence of the categories and their responses are made in response to those categories. However, as has been noted prior, there are questions as to whether even that *really* constitutes learning a category (Ross, Chin-Parker, & Diaz, 2005). For this study, I would argue that the participants acquire knowledge that is sufficient to support their task (identifying the "type" of board facilitates the placement of the connectors, and there was ample evidence of this facilitation occurring during the learning) and to guide later, more explicitly category-based tasks. It is at least the foundation of category learning.

The current study was not designed to address all facets of this process. For instance, additional study within this paradigm will be able to determine whether the participants came to adopt different representations of the features (e.g. Schyns, Goldstone, & Thibault, 1998) or whether they learned to ignore certain information (e.g. Denton & Kruschke, 2006) as they better discriminated the features during the learning. This paradigm provides a unique way to approach these types of questions and to situate the study of them within a larger framework intended to guide our understanding of the acquisition of conceptual knowledge.

Conducting this type of study within a more complex, and arguably more naturalistic, domain, we can begin to see the interaction between the individual and the environment that helps to shape the acquisition of conceptual knowledge. We can extend our understanding of the ways in which goals are implicated in category learning and how we might bridge the chasm that has separated naturalistic studies of concepts from more experimental studies.

References

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review, 98*, 409-429.

Barsalou, L.W. (1991). Deriving categories to achieve goals. In G.H. Bower (Ed.), The psychology of learning and motivation: Advances in research and theory (Vol. 27, pp. 1-64). San Diego, CA: Academic Press.

Chin-Parker, S., & Ross, B. H. (2004). Diagnosticity and prototypicality in category learning: A comparison of inference learning and classification learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*, 216-226.

Denton, S. E., & Kruschke, J. K. (2006). Attention and salience in associative blocking. *Learning & Behavior, 34,* 285-304.

Hahn, U., & Ramscar, M. (Eds.). (2001). Similarity and categorization. New York : Oxford University Press.

Jee, B. & Wiley, J. (2007) How goals affect the organization and use of domain knowledge. *Memory & Cognition, 35*, 837-51.

Kruschke, J. K. (2003). Attention in learning. *Current Directions in Psychological Science, 12*, 171-175.

Love, B. C. (2005). Environment and goals jointly direct category acquisition. *Current Directions in Psychological Science, 14,* 829-835.

Malt, B. C. (1995) Category Coherence in Cross-Cultural Perspective. *Cognitive Psychology, 29,* 85-148.

Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological Review, 100,* 254-278.

Medin, D. L., Lynch, E. B., Coley, J. D., & Atran, S. (1997). Categorization and reasoning among tree experts: Do roads lead to Rome? *Cognitive Psychology, 32*, 49-96.

Minda, J. P., & Ross, B. H. (2004). Learning categories by making predictions: An investigation of indirect category learning. *Memory & Cognition, 32*, 1355-1368.

Murphy, G. L. (2005). The study of concepts inside and outside the laboratory: Medin versus Medin. In W. Ahn, R. L. Goldstone, B. C. Love, A. B. Markman, & P. Wolff (Eds.), Categorization inside and outside the lab: Festschrift in honor of Douglas L. Medin. Washington, DC: APA.

Murphy, G. L., & Medin, D. L. (1985) The role of theories in conceptual coherence. *Psychological Review, 95*, 289-316.

Ram, A., & Leake, D. B. (Eds.). (1995). Goal-driven learning. Cambridge; MIT Press.

Rehder, B., & Hoffman, A. B. (2005). Eyetracking and selective attention in category learning. *Cognitive Psychology, 51,* 1-41.

Rosch, E., Mervis, C. B., Gray, W., Johnson, D., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology, 8,* 382-439.

Ross, B. H., Chin-Parker, S., & Diaz, M. (2005). Beyond classification learning: A broader view of category learning and category use. In W. Ahn, R. L. Goldstone, B. C. Love, A. B. Markman, & P. Wolff (Eds.), Categorization inside and outside the lab: Festschrift in honor of Douglas L. Medin. Washington, DC: APA.

Ross, B. H., & Murphy, G. L. (1999). Food for thought: Cross-classification and category organization in a complex real-world domain. *Cognitive Psychology, 38*, 495-553.

Schyns, P. G., Goldstone, R. L., & Thibaut, J-P (1998). Development of features in object concepts. *Behavioral and Brain Sciences, 21*, 1-54.

Different Kinds of Pragmatic Factors Explain Failures of Default-to-Stereotype Inferences

Matthias Unterhuber
University of Düsseldorf

Gerhard Schurz
University of Düsseldorf

Abstract: Connolly, Fodor, Gleitman, and Gleitman (2007) present theoretical and empirical evidence against the Default-to-Stereotype (in short: DS) inference and argue that prototype theories of concepts predict (DS)-inferences. Hence, they conclude that prototype theories are inadequate. Jönsson and Hampton (2008) and Hampton, Jönsson, and Passanisi (2009) argue that (1) prototype theories do in general not predict (DS)-inferences, and that (2) Gricean pragmatic effects can largely account for Connolly et al.'s empirical results. They, however, interpret a minor but still substantial part of the findings as a genuine deviation from the (DS)-inference rule. In this paper we first argue that the results of Connolly et al. (2007) pose a greater threat to prototype theories of concepts than Jönsson and Hampton (2008) suggest. Second, we present an experiment which implies that Connolly et al.'s findings can be solely explained by following Gricean pragmatic factors: (a) non-redundancy preferences and (b) informativeness suppositions.

The Effect of Graph Design Type on Word Preferences In the Description of Trend and Cyclic Events

Ozge Alacam
Middle East Technical University

Annette Hohenberger
Middle East Technical University

Kursat Cagiltay
Middle East Technical University

Abstract: This study was conducted as a part of larger study on the effect of graph type on trend and cyclic event comprehension. It aims to present the analysis of subjects' word preferences in the verbal description of trend and cyclic events, given different graph types (linear, round). For this purpose, a novel round graph type was designed. For instance, while in the linear graph timeline for a year, January is located on the left and December on the right side, in the round graph, January and December are located adjacently. As a data collection tool, verbal description task and evaluation forms were used. 40 university students participated in this study.

The results revealed that, although the graph type has no significant effect, the event type conveyed by them modulates word preferences (ex. usage of trend, discrete, and conceptual words) in the description of relations among elements depicted in the graph.

When to Switch? Understanding How Performance Tradeoffs Shape Dual-Task Strategy

Duncan Brumby
University College London

Nina del Rosario
University College London

Christian Janssen
University College London

Abstract: We use a novel dual-task paradigm to investigate how people adapt their strategy for interleaving attention between tasks to meet varying performance objectives. The study required participants to encode and enter a series of route instructions from a secondary display while driving a simulated vehicle. Experimental instructions were given to encourage participants to either prioritize safe driving or the secondary navigation task. Results show that participants met the required task objective by varying the frequency that they interleaved tasks and by varying the amount of time they spent in between visits to the secondary display. We explain these data using a framework for modeling driver distraction effects. The model explained the observed change in performance measures between the two priority conditions and also the observed change in strategy. Taken together these results support the idea that people can strategically allocate attention in multitask settings to meet specific performance criteria.

When More Load Leads to Less Distraction in Multimedia Learning: An Event-Related Potential Approach

Krista DeLeeuw
Knowledge Media Research Center

Richard Mayer
University of California, Santa Barbara

Barry Giesbrecht
University of California, Santa Barbara

Abstract: In multimedia learning, the modality effect occurs when students learn better from a lesson containing graphics accompanied by narration than one accompanied by on-screen text. The redundancy effect occurs when students learn better from a lesson containing graphics accompanied by narration than one accompanied by narration and on-screen text. In order to determine the information-processing mechanisms responsible for these effects, 36 students viewed three multimedia lessons in which the words were presented as narration, on-screen text, or both. During the lessons, brief visual distractors were presented and event-related potentials (ERPs) in response were measured. Learners showed a more positive early (P1) ERP response during the graphics+text lesson than during the graphics+narration+text lesson, indicating that more perceptual processing was required for the latter condition. In general, results suggested that perceptual load plays an important role in the modality and redundancy effects, a useful clarification of the cognitive theory of multimedia learning.

Information Search in Decisions from Experience: Do Our Patterns of Sampling Influence Our Decisions?

Thomas Hills
University of Basel

Ralph Hertwig
University of Basel

Abstract: Does the way we sample information from the environment influence the decisions we make, even when the information we obtain would otherwise be equivalent? In past research, this question has been difficult to answer because the information we obtain is often confounded with its consequences. We investigate this question by analyzing data in a paradigm where exploration comes prior to consequential decision-making, in the binary choice paradigm of decisions from experience. By investigating the relationship between patterns of information sampling and subsequent decisions, we find that individuals who switch more between options are less sensitive to the sample means and more likely to make decisions based on the outcome of pairwise comparisons, especially recent outcomes—choosing options that win most of the time. We further show that such pairwise strategies are associated with the underweighting of rare events.

The sensory nature of knowledge

Lionel Brunel
University lyon 2

Guillaume Vallet
University lyon 2

Benoit Riou
University lyon 2

Mathieu Lesourd
University lyon 2

Elodie Labeye
University lyon 2

Rémy Versace
University lyon 2

Abstract: The aim of the present studies are to assess the sensory nature hypothesis of knowledge through a series of experimental results. Especially, we investigated the links between memory and perception using a short-term priming paradigm based on a previous learning phase consisting of the association between a geometrical shape and a white noise. The priming phase examined the effect of a geometrical shape, seen in the learning phase, on the processing of a target (tones or picture). Our main results demonstrate that memory and perception share some mechanisms and components. These ones are relevant for the processing of each form of knowledge (episodic and semantic). At last, reflections about the implication of this work to study perceptual learning and memory are presented.

Wrong prediction by experts provide more support than that by novices

Kuninori Nakamura
Tokyo Institute of Technology

Abstract: The current research explored whether lay people have a tendency to provide higher support for "wrong" predictions made by experts than those made by novices. Three empirical studies consistently revealed that there indeed exists a preference for wrong predictions even when predictions are made by experts. In addition, the current research also formulates preferences for wrong predictions made by experts in terms of a Bayesian inference and expresses the processes by which one may believe the wrong prediction in the form of two factors—prior odds and likelihood ratio. Finally, I argue that this preference is logical when treated as a result of the comparison between the two competing hypotheses.

Change in Encoding Facilitates Principle Acquisition

Richard Prather
Indiana University

Abstract: This study addresses the interaction between stimulus encoding and learning. Specifically I address the acquisition of arithmetic principle knowledge and its relation to learners arithmetic equation encoding. Arithmetic principle knowledge has been shown to be a key aspect of early mathematical development. Behavioral results suggest that children with experience encoding relevant characteristics show a change in their principle knowledge. Computational results mirror this finding. Model instantiations in which the same equation encoding is used show similar behavior. I take this as preliminary evidence of a direct connection between the encoding of arithmetic equations and knowledge in the arithmetic domain. This has clear applications to both developmental theory and educational practice.

Complementary processing systems: A PDP model of the simultaneous perception of multiple objects

Cynthia Henderson
Stanford University

James McClelland
Stanford University

Abstract: Illusory conjunctions in normal and simultanagnosic subjects are instances where the binding of visual information fails to function correctly. When presented with multiple objects simultaneously, simultanagnosic patients and normal subjects under conditions of attentional loads or brief presentation times often erroneously report miscombinations of features of the objects. A connectionist model of multi-object perception examines how the concurrent perception of more than one object could occur in normal subjects and become deficient with shortened processing times. In this model, the correct identification of two objects is accomplished through lateral connections between the ventral and dorsal pathways. Lesioning of the dorsal pathway produces failures in multi-object recognition characteristic of the effect of parietal damage in simultanagnosia. It is hoped that the functioning of this model might help elucidate possible processes underlying the correct solution of the binding problem in normal subjects.

Ascribing Causality and Intention to 2D Animations

David Pautler
A*STAR Singapore

Bryan Koenig
A*STAR Singapore

Boon-Kiat Quek
A*STAR Singapore

Andrew Ortony
A*STAR Singapore, Northwestern University

Abstract: People routinely ascribe intentions and other mental states to others (partly) on the basis of observed behavior, and research shows that they tend to do this spontaneously, even with simple geometric objects moving in a 2D plane. We believe that 2D animations isolate a critical kind of information — object movement — that avatars and social robots could use when making attributions in social interaction. Our approach uses spatiotemporal, contoured constraints about objects and their movements to identify candidate causes and intentions, and then, based on evidence from background knowledge, infers which is most likely. This approach could eventually be integrated with perceptual information, such as appearance or gaze, as well as richer models of the world and other agents' minds, in order to augment the social intelligence of artificial agents.

Limits in Monitoring and Recall with Constant and Changing Memory Sets

Daniel Cassenti
U.S. Army Research Laboratory

Richard Carlson
The Pennsylvania State University

Troy Kelley
U.S. Army Research Laboratory

Abstract: Monitoring of the environment for consistency with working memory is a common aspect of human performance, as is updating working memory to reflect changes in the environment. In two experiments, we examined limits in monitoring performance and working memory retention as the number of items to be held and monitored varied from one to eight. In Experiment 1, participants monitored a display on the basis of a static memory load. In Experiment 2, participants sometimes updated the memory load by substituting new information in the display. Both monitoring and retention were quite good in Experiment 1. In Experiment 2, monitoring performance was compromised even with a single-item load, and retention was poor for loads greater than 4 or 5 items. We discuss both theoretical and applied implications of these results.

Hick's Law in the Random-Dot Motion Task

Leendert van Maanen
University of Amsterdam

Eric-Jan Wagenmakers
University of Amsterdam

Abstract: In a series of experiments we studied whether Hick's law is present in the random-dot motion task (RDM). Hick's law is the very strong experimental finding in multiple-choice research that mean response time increases with the logarithm of the number of response options. In the RDM task participants have to indicate from a group of moving dots what the dominant direction of movement is. We studied how response times in this task increased as a function of the number of alternative directions of movement presented to the participants. Using a computational model, we show that Hick's law is present, but only if the relative distance between the alternative directions of movement is taken into account.

Are Place Names Merely Referencing Expressions?

Paula Engelbrecht
Ordnance Survey

Michael Tull
University of Portsmouth

Abstract: Previous research found cross-modality priming for the names of well-known people and landmarks, but not for nouns or country names (Hollis & Valentine, 2001). Hollis and Valentine argue that country names were processed like nouns because they have sense (i.e. they can act as adjectives) and are therefore not pure referencing expressions. However, it could be argued that the participants processed country names like nouns because they possessed richer knowledge about countries (e.g. culture, climate, political structure) than about landmarks. The current study used locality names that cannot act as adjectives to test this interpretation. No cross-modality priming was found for locality names, indicating that they are processed like nouns. One possible interpretation of these findings is that the richness of the mental representation associated with a given place name determines whether it will be processed as a noun. © Crown Copyright 2010 Reproduced by permission of Ordnance Survey.

Sociocultural history of the machine metaphor

Vladimir Glebkin
Gymnasium 1514, Moscow

Abstract: Most theories of metaphor in modern cognitive linguistics either highlight the metaphorical system of human beings in general or describe differences in metaphorical systems of various cultures as a fact and do not explain the reason for these differences. When does some metaphor appear in the language? Why does it appear in this period of time but not earlier or later? Such questions sound weird in the context of this discourse, but prove to be important for understanding of sociocultural nature of some basic metaphors. The most obvious answer to these questions is: metaphor appears in language when its source domain begins to play an important role in social life. However, the history of metaphor of machine shows that this answer is not always correct. The metaphor of machine was not known in antiquity, but it had appeared in the Middle Ages long before the machines became widely used.

Agile Software Development Process: A Case Of Collaborative Cognition In Flux

Nik Nailah Binti Abdullah
GRACE Center, National Institute of Informatics

Robert G.M Hausmann
Carnegie Learning Incorporation

Shinichi Honiden
GRACE Center, National Institute of Informatics

Helen Sharp
Centre for Research in Computing, The Open University

Abstract: What role do physical artifacts play in decreasing the flux of an individual's and a group's representation during collaboration? This question was addressed in the case of Agile software development, which emphasizes collaboration. Evidence suggests that the keys to success in Agile are two physical artifacts: the "story card" and the "wall." These artifacts are particularly useful when supported by appropriate social interactions. Thus, we conducted an ethnography study of an Agile team and used situated cognition to study their communication process. We found that members perform a categorization process both at the perceptual and conceptual level, which is akin to structural coupling. The physical artifacts play a mediation role to helped members form and sustain the structural coupling process, both together well as individually. This, in turn, helped them to sustain common ground and decrease the flux of the individual's and groups representation of system design.

The Influence of Prior Knowledge on Recall for Height

Pernille Hemmer
University of California, Irvine

Jenny Shi
University of California, Irvine

Mark Steyvers
University of California, Irvine

Abstract: Many aspects of our experiences do not have to be explicitly remembered, but can be inferred based on our knowledge of the regularities in our environment. Such knowledge can operate at multiple levels of abstractions. For example, this could lead to recall for the height of a particular person to be influenced not only by general knowledge about heights of people, but also by specific knowledge about the height of men and women. We assess the relative contribution of this type of prior knowledge on reconstructive memory. In a series of behavioral studies we first assessed people's a priori expectations of the heights of men and women. We show that people's a priori expectations are in line with the true distribution of heights in the population. We then tested memory performance in a continuous recall task in which subjects had to reconstruct from memory the height of people shown earlier in a sequence. The stimuli were either naturalistic images of males and females or gender-ambiguous silhouettes. Our results suggest not only that prior knowledge can improve average recall, but also that knowledge can come from multiple levels of abstraction such as gender and the overall height of people.

Harry Potter and the Sorcerer's Scope: Scope Biases in Explanatory Reasoning

Sangeet Khemlani
Princeton University

Abigail Sussman
Princeton University

Daniel Oppenheimer
Princeton University

Abstract: What makes a good explanation? We show that individuals prefer explanations with a more narrow scope – those that account for fewer unobserved effects – to broader explanations. In Experiments 1 and 2, participants evaluated a narrow scope and a broad scope explanation of an observed symptom and preferred the former to the latter. In Experiment 3, participants evaluated more natural explanations of unexpected observations, and again displayed a bias for narrow scope explanations. We conclude by considering what this novel bias tells us about how humans evaluate explanations and engage in causal reasoning.

Cognitive Load Measurement through Multimodal Behaviour Patterns

Natalie Ruiz
NICTA, UNSW

Ronnie Taib
NICTA, UNSW

Fang Chen
NICTA, UNSW

Abstract: Our research focuses on extending the accepted benefits of multimodal computer interaction by using the paradigm to detect fluctuations in cognitive load. The advantage of this approach is that cognitive load can be determined implicitly by monitoring variations in specific multimodal input features executed during day to day tasks using a computer interface. Such unobtrusive measures may help determine the user's cognitive load in real-time, and achieve the ultimate goal of adapting information selection and presentation in a dynamic computer interface with reference to load. We identified some correlations between the communicative structure of combined speech and manual gesture input and high levels of cognitive load. The results suggest that semantic multimodal communicative structures are sensitive to cognitive load variations, with multimodal communication becoming half as redundant in high load than in low load. The feasibility of using rates of redundant constructions or complementary constructions in multimodal input as an index of cognitive load is supported by the results of our study.

Conceptual understanding of the relationship between consciousness, memory, and attention

Eunsook Kim
Interdisciplinary Research Program of Cognitive Science, Pusan National University, South Korea

Hyunjung Shin
Department of Psychology, Pusan National University, South Korea

Abstract: Consciousness is regarded as too ambiguous a concept to be understood and accepted as a mental construct without the inclusion of memory and attention in any conceptualization. We need one criterion to count satisfactorily as an explanation of consciousness in information processing. An operational working definition of consciousness could be made in comparison of memory and attention: Consciousness would be a subjective awareness of momentary experience and also have the characteristics of an operating system performing control and consolidation information processing, even though those are not equivalent concepts. This could be called a cognitive consciousness. If cognitive consciousness is postulated as a mental construct characterizing awareness, control and consolidation, the phenomena like word superiority effect, auditory continuity and object categorization, could be understood clearly, which was not the case in the past.

To Be Subtle or To Be Clear?: Comparing Strategies for Changing People's Attitudes Towards Social Groups

M. Afzal Upal
Influence & Effects Research Group Adversarial Intent Section Defence Research & Development Canada Toronto

Abstract: The problem of deciding which strategy to use to influence a target audience's social identity beliefs is of interest to social influence practitioners as well as social cognition researchers. This paper compares the effectiveness of three social influence strategies in terms of their ability to lessen their reader's affiliation for a targeted social group. We designed three messages that vary in terms of (a) how well they hide their persuasive intent and (b) clarity of the message. Our results indicate that message clarity had a stronger impact on people's group affiliation than the persuasive intent of the message. The most rhetorical and blunt Message 1 was more effective in reducing people's affiliation for the targeted group than the more subtle narrative Messages 2 & 3. The most subtle Message 3 was least effective in terms of being able to reduce subject's affiliation for the targeted group.

Category directedness

Steven Verheyen
University of Leuven

Gert Storms
University of Leuven

Abstract: The answer to the question of what constitutes a category generally comes in two guises. The first refers to the set of characteristic features associated with the category (category intension). The second refers to the set of items in the world that is delineated by the category (category extension). Although intension and extension are two complementary depictions of what a category is, little is known about their interrelation. We will present a theory of semantic categories that assumes both exemplars and features to vary along a common latent scale. Evidence for this theory will be provided through an analysis of feature by exemplar applicability matrices with the two parameter logistic model. This item response model for unidimensional data not only fits the applicability matrices, its parameters naturally account for the varying representativeness of the constituting features and exemplars.

'Meryem (reportedly) missed her flight': Cognitive Implications of the Turkish Evidential

Sumeyra Tosun
Texas A&M University

Jyotsna Vaid
Texas A&M University

Lisa Geraci
Texas A&M University

Abstract: Two experiments with adult native users of Turkish and English speaking controls examined cognitive repercussions of obligatory grammatical marking in Turkish of directly vs. indirectly experienced events. Exp. 1 examined recall accuracy of Turkish sentences containing direct vs. indirect past tense suffix markers; equivalent sentences in English contained lexical marking of indirectness (e.g., "reportedly"). Exp. 2 examined incidental recognition memory for sentences containing direct vs. indirect experience markers. Performance in Exp. 1 was uniformly low, indicating a floor effect in sentence information recall. Exp. 2 showed significantly better recognition memory for sentences containing the direct marker vs. the indirect marker in Turkish; no such advantage was observed in English. The findings suggest that obligatory marking of directly experienced events has a privileged status in mental representation.

Abstract Perceptual Learning of Hidden Patterns

Everett Mettler
University of California at Los Angeles

Hongjing Lu
University of California at Los Angeles

Philip Kellman
University of California at Los Angeles

Abstract: 'Perceptual learning' (PL) refers to experience induced improvements in the extraction of information from the environment. Although early work emphasized that PL often involved discovery of abstract invariants from stimulation (Gibson, 1969), most recent work has focused on concrete, low level stimulus properties. We describe research to understand abstract perceptual learning (APL), which requires discovery of structural relations between features, and compare it to concrete PL. Learners discovered hidden targets – 10 squares of the same luminance embedded in 12x12 'grids' of varying luminance noise. Concrete targets maintained pixel position and luminance across trials. Abstract targets changed either position or luminance across trials. In a discovery task, humans were able to discover both concrete and abstract patterns. We show that existing computational models can describe concrete but not abstract learning and we suggest models that can account for abstraction in PL.

Metathesis in English and Hebrew: A Computational Account of Usaged-Based Phonology

Paul De Palma
Gonzaga University University of New Mexico

George Luger
University of New Mexico

Abstract: It is now well understood that language use shapes the acoustic delivery of phonological patterns. One common example of this type of language change-under-use is metathesis, which is the reversal of the expected linear ordering of sounds. The gradual transformation of the Spanish word chipotle to chipolte in the United States is an example of metathetic change. The Genetic Algorithm (GA) is an optimization technique loosely based on the idea of natural selection. This paper shows that the GA can provide a computational model of a usage-based account of examples of metathesis. In the process, it argues that computer models can bring precision to linguistic theory. As an example we create a GA that is able to characterize metathesis in English and then is able to achieve even better results for related expressions in modern Hebrew.

Explaining the Minimal Counterintuitiveness Effect Without Assuming A Strongly Modular Mind

M. Afzal Upal
Influence & Effects Research Group Adversarial Intent Section Defence Research & Development Canada Toronto

Abstract: This paper outlines two approaches to account for the finding that concepts that are minimally counterintuitive are better remembered than intuitive or maximally counterintuitive concepts. The first approach considers such memory advantages to be a property of the concepts themselves while the second approach emphasizes the role played by the context in which such concepts appear in allowing a reader to make sense of them. The context-based view also suggests that counterintuitive concepts lose their advantages as they become widely accepted and embedded in a cultural milieu. In the new context, ideas with enhanced counterintuitivess obtain transmission advantages. This helps explain cultural innovation and dynamism. It also allows us to account for the development and spread of complex cultural ideas such as the overly counterintuitive religious concepts including the Judeo-Christian-Islamic conceptions of God.

Interpretate Novel Conceptual Combinations: Age-Related Impact of Memory Availability

Sandra Jhean-Larose
Université Paris Sorbonne /I.U.F.M. de Paris Equipe CHArt (Cognitions Humaine et Artificielle)- E.P.H.E 41, Rue Gay-Lussac.75005 Paris. FRANCE

Fabiola Martinez
Equipe CHArt (Cognitions Humaine et Artificielle)- E.P.H.E 41, Rue Gay-Lussac.75005 Paris. FRANCE

Abstract: This study looks at how combinations of two French nouns ("Incendie Brûlure"/"Fire Burn" ; "Voiture Tortue"/"Turtle Car") are interpreted. The order of occurrence of the constituents of two types of conceptual combinations, Relation and Property, was manipulated in view of determining how property-based and relation-based interpretations evolve with age. Three groups of French-speaking children (ages 6, 8, and 10) and a group of adults performed an interpretation-selection task. The results for the children indicated that while property-based interpretations increased with age, relation-based interpretations were in the majority for both combination types, whereas for the adults, relation-based interpretations were in the minority for property combinations. For the children and adults alike, the most frequent interpretations were ones in which the head noun came first and was followed by the modifier (the opposite of the order observed for English).

Interactive Representation in the Motor Control

Daniel Hsi-wen Liu
Providence University

Abstract: Within the approach to embodied representation, Bickhard's (1993, 2000) account of interactive representation, like Rosenberg & Anderson's (2004) guidance theory of representation, employs a notion of representation that is not grounded on the standing-in-for relation. However, Bickhard's accounts of interactive representation remains in need of explanation as to why the interactive representation is genuine representation. The present paper aims at this explanation, with the focus of anticipatory motor activities, by employing Merleau-Ponty's notion of 'motor intentionality' (Merleau-Ponty, 2006). For this, the present paper explains how the interactive representation relates to motor performance in its immediate environment, in other words, how the interactive representation gains its intentional content. In addition, the present paper argues that the interactive representation of motor activities provides the guidance of motor actions, and vice versa, hence Bickhard's notion of interactive representation and Rosenberg & Anderson's guidance theory of representation are two versions of the same theory.

The network properties of episodic graphs

Yuwen Zhuang
Department of Computer Science and Engineering, Ohio State University (OSU)

Vishnu Sreekumar
Department of Psychology, Ohio State University (OSU)

Mikhail Belkin
Department of Computer Science and Engineering, Ohio State University (OSU)

Simon Dennis
Department of Psychology, Ohio State University (OSU)

Abstract: We present statistical analyses of the small world properties for two particular types of episodic graphs. One is from the paragraph space of the Internet Movie Database (IMDb) and the other is from images collected as subjects engaged in their activities of daily living. We show that they have a small-world structure which is characterized by sparse connectivity, short average path lengths between nodes, and high global clustering coefficient. However, the degree distribution analyses show that they are not scale-free graphs. For the analyses, we selected edges from different proportions to construct the networks, hence, a series of analyses reveal the growth style of these two episodic graphs.

Keywords: Small World; Episodic Graphs; Scale-Free Graphs; Internet Movie Database (IMDb);

Weakly Supervised Learning: What Could It Do and What Could Not?

Jinhui Yuan
Tsinghua University

Bo Zhang
Tsinghua University

Abstract: Weakly supervised learning is not only a typical way of human concept learning, but also has wide real-world applications. Of particular interest to this paper is the theoretical aspect of weakly supervised learning: (a) Could weakly supervised learning learn the target concept the same as that of fully supervised learning? (b) If yes, under what conditions it will and how to achieve it? In other words, this paper will investigate what weakly supervised learning could do and what could not. The basic idea is, weakly supervised learning could be transformed into an equivalent supervised learning problem, in which way, it could be understood with the tools of supervised learning. The major results of the paper include: (a) the hardness of weakly supervised learning depends on the properties of training data and the adopted feature representation; (b) though there is no theoretical guarantee for a unique identification of the relevant variables, incorporating minimum description length principle may help infer target concept; (c) weakly supervised learning could be solved by EM-style algorithm, which is not a novel idea, however, the theoretical analysis suggests that the E-step and M-step should adopt feature representations with distinct properties rather than using the same feature.

Models of Information Integration in Perceptual Decision Making

Jared Hotaling
Indiana University

Andrew Cohen
University of Massachusetts

Jerome Busemeyer
Indiana University

Richard Shiffrin
Indiana University

Abstract: In cognitive science there is a seeming paradox: On the one hand researchers studying judgment and decision making (JDM) have repeatedly shown that people employ simple and often sub-optimal strategies when integrating information from multiple sources. On the other hand another set of researchers has had great success using Bayesian optimal models to explain information integration in fields such as categorization, perception, and memory. One impediment to reconciling this paradox lies in the different experimental methods each group has used. Recently, Hotaling, Cohen, Busemeyer, & Shiffrin (submitted) conducted a perceptual decision making study designed to bridge this methodological divide and test whether the sub-optimal integration found in verbal problems stated in terms of probabilities may also appear in perceptual tasks. Their results indicate that a classic JDM finding, the dilution effect, does arise in perceptual decision making. Observers were given strong evidence X favoring A over B, and weak evidence Y also favoring A over B. According to Bayesian analysis, the odds in favor of A should be multiplied, resulting in an increased likelihood of A. Instead, Hotaling et al. found that the weak evidence diluted the strong evidence, producing decreased judgments and choice probabilities favoring A, given X & Y, than given X alone. I review these empirical findings and test both rational and cognitive models of the integration process.

The Influence of Integration and Counterintuitiveness on Memory for Text

M. Afzal Upal
Influence & Effects Research Group Adversarial Intent Section Defence Research & Development Canada Toronto

Mary Harmon-Vukic
Psychology Department Providence College

Abstract: Recent studies suggest that counterintuitive ideas embedded in stories facilitate their subsequent recall, thus increasing the likelihood that such stories survive time and space. However, it could be that structure of counterintuitive stories affects memory rather than the distinctiveness of their contents. Indeed, Harmon-Vukic and Slone (2009) demonstrated that integration of story information eliminated the counterintuitiveness effect. The purpose of the present experiment was to further explore the influence of integration on memory for counterintuitive concepts. Participants were presented with a story containing elements that were either intuitive, minimally counterintuitive, or maximally counterintuitive. In addition, the stories were either integrated or not integrated. Participants were asked to recall the material either immediately, or one week later. Consistent with the results of Harmon-Vukic and Slone recall performance was best for integrated stories, regardless of level of intuitiveness. The same effect occurred on week later, although overall memory performance was lower.

An Attention Based Theory to Explore Affordances of Textual and Diagrammatic Proofs

Peter Coppin
University of Toronto

James Burton
University of Brighton

Abstract: Shimojima and Katagiri have demonstrated that diagrams reduce "inferential load" during reasoning by scaffolding visual-spatial aspects of memory. In response, we wondered why, if this is true, that proofs are usually text based? The purpose of this paper is to explore ergonomic affordances of text that may encourage its use in the communication of proofs by building on prior work in attention. We claim that textual notations may focus a reasoner's "spotlight" of attention through serialized sequential chunks, whereas many diagrams may "diffuse" attention and that a diagrammatic notation system that serialized information in chunks amenable to focused attention could leverage the power of textual notations. We present such an example through a case study focused on generalized constraint diagrams, a visual logic with attributes that may support focused attention and extract ergonomic principles that may transcend each notation system.

Socially Facilitated Alignment and Novelty in Separate Channels of Communication

Monica Riordan
University of Memphis

Rick Dale
University of Memphis

Roger Kreuz
University of Memphis

Abstract: We discuss two viewpoints of potential interactive alignment, socially facilitated priming and socially facilitated novelty, and test them by using simulated online conversation. In a computer-based pseudo-interactive environment, participants were led to believe they were interacting with another person or that they were seeing examples from a database and must supply 12 responses. The exchange consisted of a modified game of "I never." In both conditions, nine prompt sets were presented in which the verb, tense, topic, and presence of emoticons varied. Recall was also tested. Results show that those who believed they were conversing with another person aligned less than those who believed they were seeing examples, but recalled more of the prompts. In addition, those who believed they were talking to another person used more emoticons than those who believed they were seeing examples. We suggest that a more complex theory of alignment is necessary in which different levels of alignment, including but not limited to topical and emotional, are modulated differentially to account for the flow and drive of conversation.

The Influence of Causal Information on Memory for Misinformation

Jessecae Marsh
Texas Tech University

Sarah Kulkofsky
Texas Tech University

Abstract: Causal information has been characterized as a "mental glue" that binds ideas together in the mind. This experiment tests the influence of such causal binding in a traditional memory misinformation paradigm. Participants studied information that either appeared as individual traits or as traits connected by causal links. Participants then rated a series of traits that contained both true and "misinformation items." Misinformation items were created to be causally plausible or implausible alternatives to previously learned information. Our question was whether the presentation of causal links in the study phase would affect the reporting of misinformation in later phases. Participants in the causal version of the study phase were more influenced by misinformation items that were causally plausible than items that were causally implausible. Participants in the noncausal version of the study phase were not differentially influenced by the plausibility of misinformation lures. Explanations for these results are discussed.

The Interaction of Age and Skill for Recognition of Chess Positions

Jerad H. Moxley
Florida State University

K. Anders Ericsson
Florida State University

Tyler J. Towne
Florida State University

Ryan Best
Florida State University

Abstract: Even young chess players display superior recall for chess positions (Chi, 1978; Horgan & Morgan, 1990; Schneider, Gruber, Gold, & Opwis, 1993), which has been attributed to their greater chess knowledge and available patterns and chunks (Chase & Simon, 1973). Controlling for skill no effect of age has generally been found in youth chess players on chess tasks. This study demonstrates, a strong effect of age, where older children are better able to recognize chess positions. Additionally this study uncovered an interaction between age and skill whereby older children who are better at chess do much better then other groups. We propose that older children use deeper processing techniques to scaffold the complex memory structures that support chess playing, which may not be automatically applied to an unfamiliar task.

Thinking Ahead: How Children Reason About the Future

Janani Prabhakar
Rutgers University, New Brunswick

Judith Hudson
Rutgers University, New Brunswick

Abstract: Episodic future thinking relies heavily on self-projection, i.e., projecting oneself into the future. It involves the use of episodic and semantic memory in order to plan and anticipate future need. In our study, we focused on understanding the role of self in episodic future thinking in 3- and 4-year old children. Children were asked to make choices either for their own future need (self group) or for another individual's future need (other group). Including these groups allowed us to directly manipulate the role of self. Participants were shown a 3-D model of a neighborhood with several locations (houses and stores) and were asked to navigate around the neighborhood to achieve a future goal requiring a two-step action (go to toy store to buy present and then go to friend's house to give present). In one version of the study, children were given only a few choices of locations in the neighborhood in order to achieve their goal. Preliminary results suggest that both 3- and 4-year-old children demonstrate episodic future thinking by accurately following the actions to achieve the future goal. In the second version of the study, 2 additional choices of location were added to the neighborhood, while keeping the goal the same. These additional items served as distracters in the environment. Results show that while 4-year-old children still demonstrate episodic future thinking skills, 3-year-old children are no long able to accurately follow the actions necessary to achieve the future goal. Further research on optimal performance by 3-year-old in such visual working memory and episodic future thinking tasks is necessary.

When Distance Relationship Contradicts Similarity in SUSTAIN

Chung-Yu Wang
National Cheng-Kung University

Lee-Xieng Yang
National Chengchi University

Abstract: SUSTAIN is an influential multiple clusters (i.e., prototypes) model for categorization, in which the cluster nearer to the stimulus is activated for categorization. Due to that the cluster activation is the average of the dimensional similarity weighted by attention, the choice between clusters is little influenced by the dimension on which the similarity difference to the stimulus between clusters is negligible, even with equivalent dimensional attention weights and the clusters having the same distance difference on dimensions (e.g., 1 vs.2 on dimension 1 and 4 vs.5 on dimension 2). This is evident in modeling Experiment 1 in Erickson and Kruschke (1998) that SUSTAIN activates the rule-category cluster, which is actually farther to the critical item than the exception-category cluster. The computer simulation results suggest that the larger the learning rate, the more likely this similarity-distance contradiction occurs. With the SUSTAIN using the ALCOVE's similarity computation equation, this contradiction disappears.

Adaptive Information Indexing in Re-finding Information

J. Michelle Moon
Carnegie Mellon University

Wai-Tat Fu
University of Illinois at Urbana-Champaign

Abstract: We studied how the human cognitive system adaptively performs information indexing (i.e. knowing where to re-find information without necessarily knowing the information content). Participants searched for information using computer icons with or without locations or luminance cues. The cues represented the history of use of icons and were calculated using the ACT-R memory equation. Results suggest that participants adaptively used the location/luminance cues to offload information indexing from internal memory to external cues. Availability of location cues led to more frequent icon accesses and worse recall of icon titles. Availability of luminance cues led to worse recall of icon titles and locations. Participants adapted to the cost-benefit structure of the environment by strategically shifting the use of different kinds of external cues based on their relative access costs. Results highlight the dynamic interplay between external representations and human information processes in shaping adaptive interactive behavior in information search and indexing.

Variability Helps Children Balance a Beam

David Pfeiffer
University of Cincinnati

Daniel Bullard
University of Cincinnati

Heidi Kloos
University of Cincinnati

Abstract: Basic-level research suggests that learning about an event is a function of what is being attended to and what is being ignored rather than amount of time spent exploring the event. However, what about learning to overcome a misconception, such as the task of balancing an asymmetrically weighted beam away from center? This research investigated the effects of training variability for children trying to balance visually symmetrical yet proprioceptively asymmetrical beams on a fulcrum. Results indicate that (i) older children's judgments of a beam's weight distribution improved with experience, (ii) younger children had particular difficulty distinguishing between the heavier and lighter side when the weight difference was smallest, and (iii) children in the mixed-experience condition scored fewer errors than children who received more extensive experience but in just one type. The findings of a significant quadratic trend for effect of learning underscore the importance of variability in children's experience.

Mutual Alignment Analogy Facilitates Abstraction and Transfer of a Complex Scientific Principle

Judy Orton
Georgia State University

Florencia Anggoro
College of the Holy Cross

Benjamin Jee
College of the Holy Cross

Abstract: Learning about a scientific principle often occurs in the context of unfamiliar examples. Mutual alignment analogy—a type of analogical comparison in which the analogs are only partially understood—has been shown to facilitate learning from unfamiliar examples (Kurtz, Miao, & Gentner, 2001; Loewenstein, Thompson, & Gentner, 1999, 2003). The present study examined the role of mutual alignment analogy in the abstraction and transfer of a complex scientific principle from examples presented in expository texts. The results provide evidence that promoting comparison between two examples and orienting the learner toward the relational commonality between the examples result in greater abstraction and transfer of the principle "convergent evolution". These findings suggest that mutual alignment analogy can promote learning complex scientific principles from texts. Mutual alignment analogy is therefore likely to be a helpful learning aid and pedagogical technique in science education.

A Lexical Gap in the Humor Domain of Japanese and Its Possible Implications for Theories of Conceptual Language

Whitney Vandiver
Purdue University

Abstract: Interviews with native speakers of Japanese reveal that, within the humor domain, Japanese experiences a lexical gap, lacking terms for ethnic, political, and religious humor and their corresponding preference terms. However, Japanese terms do exist for the same concepts in non-humorous domains but yet, they do not cross into the humor domain, thus pointing to a specific phenomenon within the language. These limited data may indicate a variation in the mind's conceptual structure or, perhaps more likely, of the concepts' realizations as lexical items depending on one's native language. Not to fall into Whorf's overgeneralization, this may still suggest that the acceptable application of a concept's expression in a language is more than simply a linguistic variation and but may still be the consequence of a language's influence on conceptual structure, resulting in its deliberate and possibly societaly and ethically conventionalized deviation from the organization of the language-independent ontology.

Reactive Task-Set Switching Ability, Not Working Memory Capacity, Predicts Change Blindness Sensitivity

Robert Youmans
California State University, Northridge

Abstract: Individual differences in working memory capacity (WMC) have long been shown to predict how well people perform tasks that require directed attention, but other individual differences responsible for task-set switching and noticing behaviors are less well understood. In this study, students from California State University, Northridge completed a measure of WMC, a measure of cognitive flexibility, and attempted to identify disappearing objects in change blindness slides. WMC had no relationship with the other measures, but measures of cognitive flexibility were directly correlated with the ability to notice change, and no relationships were established. The author argues that these findings support: 1) new ways of thinking about task-set switching behaviors, and 2) the existence of individual differences in the ability to notice changes in an environment that are independent of WMC.

Perception of Visual Similarity: Modeling Feature-Based Effects

Michael Romano
University of California, Merced

Michael Spivey
University of California, Merced

Abstract: Similarity is central to human cognition. Its relevance is apparent in nearly all theories of cognitive science. Concept acquisition, metaphor, pattern recognition, priming, predictions, inferences; all these processes rely on similarity. Despite its relevance, relatively little is understood about how similarity is processed. In particular, there is a need to better understand the scope in which our perceptual systems constrain our judgments of similarity. The current study investigates this question in the area of visual cognition. By attempting to control for the influence of categorical knowledge, the goal was to understand how different types of feature-dimensions and category boundaries influence the perception of similarity. A connectionist model was developed to explain these findings.

Preschooler's Performance in Three Visual Perspective Taking Tasks

Yue Yu
Department of Psychology and Yuanpei College, Peking University

Yanjie Su
Department of Psychology and Yuanpei College, Peking University

Raymond Chan
Neuropsychology and Applied Cognitive Neuroscience Laboratory, Key Laboratory of Mental Health, Institute of Psychology, Chinese Academy of Sciences

Abstract: Previous studies suggested that visual perspective taking (VPT) requires spatial computation (transformation) and solving conflict between reality and imagination (interference). The current study examined how these two processes may influence preschoolers' performance in different VPT tasks. Eighty-four 3 5-year-old children and 8 adult controls completed three VPT tasks, where different conditions were set to manipulate the requirement to perform transformation and to solve interference. The result showed that 4-year-olds could solve the interference between reality and imagination, but the immature spatial computation ability remained a confining factor throughout preschool years and prevented children from passing more complex VPT tasks.

Implicit Learning of Spatial Context by School-Age Children

Hanako Yoshida
University of Houston

Kevin Darby
University of Houston

Joseph Burling
University of Houston

Abstract: The contextual cuing effect refers to a robust phenomenon in which a repeated context guides attention to relevant information by constraining search. The effect is measured by an object search task in which a target is located within repeated or nonrepeated visual contexts. Shorter response times with repeated configurations indicate that contextual information has facilitated search. Though the effect is robust among adult participants, recent tests of the effect with children yielded mixed results. Because contextual cuing could play a critical role in cognitive development, resolving this issue is important. The present study used child-friendly paradigms to investigate whether children show the effect. The study suggests that adult participants show the effect regardless of stimulus type; 9- to 12-year-old children's contextual cuing effect was limited to certain stimuli types. The results are discussed in terms of the relation between visual complexity of stimuli and the recognition of search items.

Studies of the Effects on Creativity of Having Very Different Parents

Brian O'Connor
University of British Columbia

Liane Gabora
University of British Columbia

Abstract: We tested the hypothesis that having parents who are very different from one another is associated with heightened creativity. In the first study, scores for 591 participants on the five factor model of personality were used to compute the personality vector known to be associated with creativity. Higher scores were significantly correlated with parental difference scores. In the second study, 114 participants were given questionnaires that included a measure of creativity (the Remote Associates Test), and a measure of creative personality (the Creative Personality Scale), as well as measures of parental behavior and personality, and parental conflict. Creativity scores and creative personality scores were significantly correlated with parental differences, but not with measures of parental conflict. We posit that the greater the parental differences, the greater the extent to which the child's worldview contains inconsistencies that invite contemplation, and thereby accustom the child to thinking for him/herself.

A Dynamic Memory Model

Eva Cadez
School of Social Sciences, Humanities, and Arts, University of California, Merced, 5200 North Lake Road Merced, CA 95343 USA

Evan Heit
School of Social Sciences, Humanities, and Arts, University of California, Merced, 5200 North Lake Road Merced, CA 95343 USA

Vladimir Cadez
Astronomical Observatory of Belgrade, Volgina 7 Belgrade, 11060 Serbia

Abstract: We introduce a mathematical model of evolution of a memory trace. The model generalizes similar dynamical systems models used in other research areas of cognitive science as well as physics and other sciences (e.g. transport processes of radiation). We then simulate an experimental data set and argue that the model may well simulate complexity of serial recall reported in Anderson, J. R., Bothell, D., Lebiere, C. Matessa, M. (1998). An integrated theory of list memory. Journal of Memory and Language, 38, 341-380. Finally, we argue that in the future our model may be used to simulate seemingly broad spectrum of memory research data: familiarity effects and rate of presentation effects in list memory, attention, false memory, instructions in memory tasks, and noise and phase transitions in neural networks and other dynamic systems. Our hope is that this kind of work may help integrate data while showing constraints useful for directing future research in cognitive science.

Effects of the Target Distribution on Numerical Prediction

Jason Jones
University of California, San Diego

Abstract: Human subjects repeatedly attempted to predict a four-digit number. The distribution producing the target numbers was varied within subjects. Target distribution had a large effect on prediction performance. The target distribution standard deviation was a good predictor of the slope of the learning curve across trials.

Cognitive leaps and multiple epistemological resources: an agent-based modeling approach

Paulo Blikstein
Stanford University

Abstract: Agent-based modeling (ABM) has been increasingly used by scientists to study a wide range of phenomena in physics, chemistry, and biology. In these models, each element ('agent') follows local, simple rules, and the overall macroscopic pattern emerges from these multiple local behaviors. Despite its roots in the natural sciences, ABM is highly relevant to research in the social sciences. The recent decades have seen a surge in social-science studies employing ABM, and more recently it has also been used to illustrate aspects of cognitive development, collaboration, and group work. In this paper, we describe a model to explain sudden leaps in cognition observed in a science classroom when students switch between the use of two types of epistemological resources. The model confirms that reliance on brute-force search is very sensitive to increased complexity of the content, whereas selective search is initially less efficient but more accurate for complex content.

Representing Conceptual Knowledge: A Network Analysis

Takashi Yamauchi
Texas A&M University

Ricardo Gutierrez-Osuna
Texas A&M University

James Caverlee
Texas A&M University

Abstract: We adopted social network analysis and investigated how concepts related to living things (e.g., organic objects such as dogs, cats, and trees) and artifacts (e.g., desks, tables, and cars) are organized. Our analysis shows that there is a basic division between the two types of concepts (living things and artifacts), and that the division emerges partly from the fact that living things are highly interconnected as compared to artifact concepts in their attributes. Three network measures, density, clustering coefficients, and complete triplets, indicate that organic concepts are heavily clustered by their attributes as compared to non-organic artifacts.

Development of the Semantic Network: From a random to a complex network

Shohei Hidaka
Japan Advanced Institute of Science and Technology

Abstract: In the present study, we investigate semantic knowledge of both adults and children for early learned words. Previous studies have suggested that semantic knowledge forms a network with particular properties, called a complex network. Since, in theory, a particular kind of network structure may be generated by a particular process, the structure of the semantic knowledge network found in semantic tasks has been supposed to reflect the developmental process of semantic knowledge acquisition. However, at this point, no empirical description is available for the development of children's knowledge –adult substitutes are used. We investigate children's semantic knowledge using an alternative-forced-choice association task. The result suggests that the children's semantic network is closely approximated by a random network which is different from adults' network. However, it is not truly random but contains a reliable structure. We discuss a possible developmental trajectory from a children's random-like network to an adult's complex network.

Predicting Coreference: the role of alternative constructions

Peter Baumann
University of Freiburg

Lars Konieczny
University of Freiburg

Barbara Hemforth
CNRS, Universite Paris Descartes

Abstract: Expectations about alternative constructions play a crucial role in anaphora resolution.

In a self-paced reading experiment, we presented Portuguese sentences, consisting of a main clause with two referents, followed by a subclause with a pronoun referring unambiguously to one of the referents. The sentences varied in the kind of subordinating conjunction: 'antes que' (before) vs. 'quando' (when).

On the pronoun and the spill-over, there is a clear decrease in reading times for the conjunction 'antes que' in the object coreference condition, whereas no difference was not found for 'quando'.

These results can be explained by comprehenders using an expectation-based strategy: in Portuguese, for sentences with 'antes que' there is a frequent alternative infinitive construction (antes de abrir: before opening), which allows coreference only with the subject of the preceding clause. Upon seeing the subordinate construction, comprehenders may assume that the speaker intended coreference with an antecedent other than the subject.

The Effects of Transcranial Magnetic Stimulation over Premotor Cortex on the Perception of Biological Motion

Bianca van Kemenade
University College London

Neil Muggleton
University College London

Vincent Walsh
University College London

Ayse Pinar Saygin
University of California, San Diego

Abstract: We investigated the roles of posterior superior temporal sulcus (STS) and premotor cortex in biological motion perception using transcranial magnetic stimulation (TMS). Subjects viewed noise masked point light displays (PLDs) of humans and scrambled figures and determined whether a person was present in each trial. Non-biologically moving PLDs (polygons) served as control stimuli. Theta burst TMS was delivered over left premotor cortex, left STS, or vertex (Saygin, 2007, Brain). Sensitivity and response bias were both affected after premotor TMS (but not STS) which was due to an increase in false alarms. This effect was not found in the control task. These data suggest that the STS and premotor areas play dissociable roles in biological motion perception. The increased false alarms after premotor TMS suggests that this region may normally help refine the computations of posterior areas during biological motion perception.

Questioning the Free Will Comprehension Question

Edward T. Cokely (cokely@mpib-berlin.mpg.de)
Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany
Michigan Technological University, Dept. of Cognitive & Learning Sciences, 49931, Houghton, Mi

Adam Feltz (adfeltz@schreiner.edu)
Philosophy and Interdisciplinary Studies, 2100 Memorial Boulevard
Kerrville, Texas 78028 USA

Abstract

Understanding the folk notion of free will and moral responsibility is important for a host of applied and theoretical issues in psychology, philosophy, and ethics. The bulk of experimental research has focused on folk intuitions concerning determinism's relation to free will and moral responsibility. However, determinism is a difficult term for many folk to understand. Accordingly researchers often use comprehension questions to identify and exclude large proportions of participants who seem to struggle with relevant concepts. Here, we document some of the cognitive mechanisms involved in folk judgments related to comprehension of determinism, and its relations to free will and moral responsibility. Results provide prescriptions for experimental designs that can increase comprehension, potentially decreasing sampling biases. Theoretical and methodological implications are discussed.

Keywords: Free will, moral responsibility, experimental philosophy, judgment, decision making, cognitive control, methodology.

Free Will and Moral Responsibility

One of the most persistent philosophical problems deals with difficult issues surrounding free will and moral responsibility. When grappling with these issues, some philosophers take what ordinary people think about free will and moral responsibility to be important theoretical considerations for a variety of reasons. For example, beliefs in free will and moral responsibility form essential cornerstones of many people's relationships to themselves and others. Some philosophers even go so far as to say that if we find that humans in fact do not have free will and moral responsibility, we should continue to let people indulge in the myth that they can be free and moral responsibility (Smilansky, 2002). But typically, philosophers have not conduced systematic or controlled scientific studies to uncover what ordinary people think about free will and moral responsibility. Over the past decade, empirically minded theorists have begun filling in this lacuna.

Most of the debate about free will and moral responsibility concerns determinism's relation to free will and moral responsibility. The empirical investigation of folk intuitions about free will and moral responsibility has reflected this tradition. Most empirical studies so far have focused on ordinary people's intuitions about determinism's relation to free will and moral responsibility.[1] But "determinism" is a "term of art". Because determinism is a term of art, some have worried that some folk may not be fully internalizing or understand determinism when making judgments about free will or moral responsibility. We agree that this is a problem for any measure of folk intuitions about determinism's relation to free will and moral responsibility. In this paper, we present evidence from two studies suggesting that one can make the deterministic nature of some scenario more transparent to participants with a simple manipulation. We argue that these results provide some important insights into judgments of free will and moral responsibility, the processes responsible for those judgments, and the philosophical importance of those judgments.

Free Will Comprehension Question

In recent years, researchers have attempted to help shed light on folk notions of freedom and moral responsibility. For example, Nahmias, Morris, Nadelhoffer, and Turner (2005; 2006) found that most people judge that people in some deterministic universes are free and morally responsible—a result that has been widely replicated (Feltz & Cokely, 2009; Nahmias, Coates, & Kvaran, 2007; Nichols & Knobe, 2007). These results indicated that many people have *compatibilist* intuitions that free will and moral responsibility are compatible with the truth of determinism. Relatively few people have *incompatibilist* intuitions that free will and moral responsibility are not compatible with the truth of determinism.

However, many theorists have worried that participants do not fully understand the deterministic nature of the commonly used scenario. If participants do not appropriately appreciate the deterministic nature of the scenario, then their responses do little to help illuminate folk intuitions about determinism's relation to free will and moral responsibility. The worry is

[1] There are, or course, some exceptions. For examples, Vohs and Schooler (2008) explore the effects of increasing anti-free will beliefs, Woolfolk, Doris, and Darley (2006) explore intuitions about moral responsibility as a function of constraint, and Miller and Feltz (2009) discuss Frankfurt-style cases. These studies do not obviously gauge people's intuitions about determinism's relation to free will and moral responsibility.

especially problematic because the philosophical sense of determinism is highly technical and nuanced. To illustrate, Alfred Mele (2006) describes determinism as the thesis that "at any instant exactly one future is compatible with the state of the universe at that instant and the laws of nature" (p. 3). This notion is conceptually distinct from other, yet related, notions such as fatalism. Fatalism is "the thesis that whatever happens must happen; every event or state of affairs that occurs, must occur, while the nonoccurrence of every event and state of affairs is likewise necessitated" (Bernstein, 2002, p. 65). One reason why fatalism and determinism are distinct concepts is that fatalism is consistent with determinism being true or false (Bernstein, 2002). That is, all things may happen necessarily even if some things are indeterministically caused (e.g., God may have foreknowledge of all events). Many compatibilists believe that fatalism rules out free will and moral responsibility. Hence, it is difficult but essentially important to convey to non-philosophers an accurate notion of determinism that does not imply something too strong (e.g., fatalism) or something that is too weak (e.g., indeterminism).

To help ensure that participants understand the deterministic nature of the scenarios, many authors include comprehension questions. It is common practice to exclude those who fail the comprehension checks (Nahmias, Morris, Nadelhoffer, & Turner, 2005, 2006). However, one shortcoming of previous research is that often the number of participants who fail the comprehension question is not reported and no empirically tested explanation is offered why participants fail the checks. Feltz, Cokely, and Nadelhoffer (2009) suggest that "many participants often fail the manipulation checks in these kinds of studies" (p. 16).[2] But why do so many people fail the manipulation checks and how can we get more people to pass them? We explore this question in a series of two experiments.

Experiment 1

One possible explanation of why participants do not give the "correct" answer to the comprehension question is that they may not spend the time necessary to internalize the description of determinism. Typically, participants are volunteers or are given for partial course credit for participating. In these situations, there may be some reason for the participants to complete the survey, but there is little reason for them to spend a great deal of time or effort completing the surveys. In these conditions, participants may fail the comprehension question because of a relative lack of concentration, effort, or time needed to understand the deterministic nature of the scenarios. Hence, one hypothesis is that increasing incentives for participants to understand the deterministic nature of the scenarios would increase correct answers to the comprehension question.

An alternative hypothesis is that the order of presentation of questions biases responses to the comprehension question. On a standard "two systems" conception of cognition, System 2 (controlled, deliberative processing) works to override and correct System 1 processing (effortless, quick, effortless processing) (Stanovich & West, 2000; but see Cokely, 2009 or Gigenerenzer & Regier, 1996; Osman, 2004 for critical discussion of "dual systems"). Hence, once impressions are formed (e.g., after making other judgments about crimes that have high emotional valence), one can only attempt to correct previous, perhaps erroneous intuitions. Unfortunately, theory suggests that many people do not have the requisite cognitive control capacities that would allow such an intervention. Fortunately, evidence suggests biases can also be overcome or entirely avoided by shaping the initial interpretation and representation of tasks before alternative intuitions are issued. This shaping requires early intervention (e.g., early selection cognitive control or changes in task orders) and can circumvent the need to correct biased processing (Cokely & Kelly, 2009; Cokely, Parpart, & Schooler, 2009; for related mechanistic accounts see also *Query Theory* by E. Weber, E.J. Johnson, & colleagues).

To be clear, cognitive processes associated with philosophical intuitions (e.g., long term memory activation and retrieval dynamics) may be shaped by participants' representation of the content of the previously viewed scenarios. Such influence is especially likely when the task involves a strong affective component. Evidence already suggests that a scenario's affective strength can alter people's judgments concerning freedom and moral responsibility (Nichols & Knobe, 2007). Those given especially strong affective scenarios tend to give more compatibilist-friendly responses than those given a relatively less affectively charged scenario. Nichols and Knobe (2007) even go so far as to say that these results indicate the existence of an "affective bias" for judgments of freedom and moral responsibility.

We hypothesized that a similar phenomenon might occur with respect to the comprehension question. When affective components are presented before the comprehension question, they may alter the interpre-

[2] In one of the few studies that report the percentage of participants who were excluded, Nahmias, Coates, and Kvaran (2007) reported that they excluded 22 percent of their participants because of comprehension failures. However, in their study, participants first answered the comprehension question. Anecdotal evidence suggests that in other experiments, rates of failure were much higher. These data provide some evidence suggesting that placement of the comprehension question is a key factor in correct responses (see Experiments and Discussion)

tation of the task making it difficult to give the normatively correct answer to the comprehension question (e.g., as this answer could contradict the previous answer that participants feel very strongly about). However, when the comprehension question is presented first and in the absence of affective information, many people may find it relatively easier to answer the comprehension question according to the supplied definition. Hence, this possibility generates the hypothesis that if the comprehension question is presented first and before affective material, then there should be more correct answers to the comprehension question than if the comprehension question is presented after affective material.

Methods and Materials

There were three conditions in Experiment 1. In the Control and Paid condition, participants read the following scenario from Nahmias, Coates, and Kvaran (2007):

> "Most respected psychologists are convinced that eventually we will figure out exactly how all of our decisions and actions are entirely caused. For instance, they think that whenever we are trying to decide what to do, the decision we end up making is completely caused by the specific thoughts, desires, and plans occurring in our minds.
>
> The psychologists are also convinced that these thoughts, desires, and plans are completely caused by our current situation and the earlier events in our lives, and that these earlier events were also completely caused by even earlier events, eventually going all the way back to events that occurred before we were born.
>
> So, once specific earlier events have occurred in a person's life, these events will definitely cause specific later events to occur. For example, one day a person named John decides to kill his wife so that he can marry his lover, and he does it. Once the specific thoughts, desires, and plans occur in John's mind, they will definitely cause his decision to kill his wife.
>
> Assume the psychologists are right that events that occurred before John was born definitely caused his decision to kill his wife. Please rate to what degree you agree with the following statements."

After reading this scenario, participants were asked to rate their level of agreement with the following two sentences (1 = strongly disagree, 4 = neutral, 7 = strongly agree): (1) John decided to kill his wife of his own free will; (2) John is morally responsible for killing his wife. Immediately following those two questions, participants were asked two 'yes' or 'no' questions: (3) Do you personally think the psychologists are right that all of our decisions are ultimately caused by events occurring before our birth? and (4) Regardless of how you answered question 3, if the psychologists are right, is it accurate to say that if the universe were re-created, people would make all the same decisions?

In the Early Selection condition, participants only read the first paragraph and then answers questions (3) and (4) (with appropriate substitution). After answering questions (3) and (4), they went to a separate page where the entire scenario was present and they were asked to answer questions (1) and (2). They were instructed not to go back and change their answers to (3) and (4) after they submitted their answers.

Participants

In the Control and Early Selection conditions, participants completed the surveys at the Philosphical-Personality website. In the Paid condition, participants completed surveys at Schreiner's Behavioral Philosophy Lab. Participants were tested in 6 groups of no larger than 12 and no smaller than 4 participants. They received exactly the same materials as those in the Control condition. Participants completed the survey at computer terminals and were told that they would be paid $2.00 for getting at least 80% of the questions correct.

Following standard practice, we eliminated all those with extensive philosophical training (all those reporting having a B.A. or greater in philosophy), those whose first language was not English, and those who reported themselves to be age 18 or under.

Results and Discussion

Question (3) is one typical comprehension question. Table 1 represents the numbers of those who passed and failed the question for each condition:

Table 1: Comprehension Failures.

	Fail	Pass
Early Selection (N = 71)	23, 32%	48, 68%
Control (N = 115)	54, 47%	61, 53%
Paid (N = 40)	24, 60%	16, 40%

The difference between Early Selection and Control was significant χ^2 (1, N = 186) = 3.837, p = .05. The difference between Early Selection and Paid was significant, χ^2 (1, N = 111) = 7.99, p = .005. However, the difference between Control and Paid was not significant χ^2 (1, N = 155) = 2.02, p = .16, but showed a very small, potential non-significant trend.

These results suggest that a substantial number of participants can pass the comprehension question. In addition, giving participants incentive to answer the comprehension question correct did not have a reliable effect. Rather, the position of the question seemed to be the most relevant factor determining participants judgment accuracy. However, one possible alternative explanation might be that the rather lengthy scenarios complicates interpretation and is the primarily reason participants did not make the correct inference concerning the deterministic nature of the scenarios. Participants may be more likely to lose track of the deterministic nature of the scenarios because they are so long and the comprehension questions late in the series of questions (Nichols and Knobe, 2007). To rule out this possible explanation, we performed a second experiment to replicate and extend the results of Experiment 1.

Experiment 2

Experiment 2 was designed to address two different issues. First, the length of the scenarios was drastically reduced. If we still find an effect with respect to the placement of the comprehension question, then it is not likely that the length of the scenario is what was responsible for differences observed in Experiment 1. Second, we wanted to make perfectly clear what the correct answer to the comprehension question was. So, in one version of the scenario we provided a sentence that clearly states what the correct answer to the comprehension question was. This provides the strongest test of whether the placement of the comprehension question is responsible for correct to that question in comparison to the 'too lengthy' hypothesis. Those who received the extra sentence should get the comprehension question correct more frequently than those who did not receive the extra sentence. In addition, those who receive the extra sentence and the comprehension question first should give the correct response to the comprehension question more often than those who receive the extra sentence with the comprehension question second.

Experiment 2a

Participants Participants completed the short survey on the Philosophical Personality website. We excluded those who reported that their first language was not English, had at least a bachelors degree in philosophy, and who were under 18. After excluding these people, there were 109 participants remaining.

Methods and Materials Participants were randomly divided into two conditions. One group was in the No Sentence condition where they read the following passage: "Most respected psychologists are convinced that our thoughts, desires, and plans are completely caused by our current situation, the earlier events in our lives, and events that occurred before we were born." The other group of participants were in the Extra Sentence condition. Participants in this condition read the following passage in addition to the passage read by those in No Sentence: "That means that if the psychologists are right and the world was exactly re-created, people would make all the same decisions." They were then asked two questions: (a) According to the psychologists, is it accurate to say that if the universe was exactly re-created people would make all the same decisions? and (b) Regardless of how you answered question 1, do you personally think the psychologists are right that all of our decisions are completely caused by our current situation, earlier events in our lives, and events occurring before our birth? They could only respond 'yes' or 'no'. On a separate page, both groups were given an addition sentence: Please imagine that one day a person named John decides to kill his wife so that he can marry his lover, and he does it. Following this sentence, they were asked to rate their level of agreement with two sentences (1 = strongly disagree, 4 = neutral, 7 = strongly agree): (c) John decided to kill his wife of his own free will, (d) John is morally responsible for killing his wife. After submitting their answers to (a) and (b), participants could not go back and change their answers.

Results and Discussion Results for each condition are reported in Table 2.

Table 2: Comprehension Failures.

	Fail	Pass
"No Sentence" (N = 60)	36, 60%	24, 40%
"Extra sentence" (N = 49)	16, 33%	33, 67%

The difference between the two conditions was statistically significant, $\chi^2 = (1, N = 109) = 8.09$, $p = .004$. To test whether order made a difference in these judgments, an additional experiment was conducted.

Experiment 2b

Participants Participants completed the survey on the Philosophical Personality website. The same exclusion criteria that applied in Experiment 2a applied in Experiment 2b. After excluding these participants, 187 participants remained.

Methods and Materials Participants received the exact same materials as those in Experiment 2a except that the order of the questions was reversed. Participants received questions in the following order: (c), (d), (a), (b). Importantly, participants received the sentence concerning John killing his wife before they answered all questions. Questions (c) and (d) appeared on one page. Once participants answered (c) and (d), they then answered on a separate page (a) and (b) and were not able to go back to change their answers to (c) and (d).

Results and Discussion Results for Experiment 2b are reported in Table 3.

Table 3: Comprehension Failures

	Fail	Pass
"No Sentence" (N = 89)	54, 61%	35, 39%
"Extra sentence" (N = 98)	56, 57%	42, 43%

The difference between the two conditions was not statistically significant, $\chi^2 = (1, N = 187) = .24$, p = .62. Importantly, across experiment 2a and 2b, the order had an effect on whether those who received the extra sentence version gave the correct answer, $\chi^2 = (1, N = 147) = 7.84$, p = .005.[3]

General Discussion

Our experiments suggest that the placement of the comprehension question was an important factor in whether participants give the correct answer. Experiment 1 suggested that those who received the comprehension question first and in the absence of affective material were more likely to give the correct answer than those who received it later and after the presence of affective material. Moreover, Experiment 1 suggested that motivation to understand the scenarios was not a reliable, primary factor in whether people were able to make the correct inferences about the deterministic nature of the scenarios (although it is possible, but by no means certain, it could become more influential with higher levels of incentive). Experiment 2 suggested that the length of the scenario was not responsible for comprehension failures. In addition, Experiment 2 suggested that the order of presentation was a factor even when it was very clear what the correct answer to the comprehension question was.

These results lead to some clear prescriptions regarding how to go about measuring comprehension in these scenarios. If one desires to maximize the usable sample (i.e., those who pass the comprehension question) and avoid potential sample bias one should present the comprehension question first and before any potentially biasing, affective content. These results also indicate that people may interpret scenarios and questions differently than intended by the experimenters (Feltz & Cokely, in press; see also Cokely & Feltz, 2009a; 2009b) or at least that the ability to infer correctly the deterministic nature of these scenarios is plastic. A common explanation of failures to the comprehension questions is that the participant did not "care" enough or could not correctly apply or understand the deterministic nature of the scenarios. However, our data suggest something different. Perhaps those who fail the comprehension question simply interpret and represent the question differently. Perhaps they understand the question in a fundamentally different way. Perhaps they are biased by affect. Given their understanding of the task, they may not be giving an incorrect answers at all (Feltz, Cokely, & Nadelhoffer, 2009; but for more general examples from the judgment and decision making see Gigerenzer, 1991, and Gigerenzer & colleagues). In such cases, these responses reveal something important about how plastic our intuitions about freedom and moral responsibility can be. The understanding of determinism appears to bi-directional—correctly understanding determinism could influence judgments about free will and moral responsibility. But importantly, it could be that beliefs about free will and moral responsibility can influence judgments about determinism. If all of this is true, it provides a more plausible description of the rich and nuanced ways people go about making a host of important judgments about themselves, their place in the world, and their relationships with others. Critically, results such as these (and others) necessitate that experimental philosophers and cognitive scientists adopt psychologically sensitive multimethod approaches to the investigation of folk intuitions.

[3] In some of the experimental conditions, we found that answering the control question interacted with judgments about moral responsibility, as expected. Those who failed the comprehension question agreed more strongly that the person is morally responsible than those who passed. In both experiments, people's judgments about free will did not change as a function of answering the comprehension question correctly.

Acknowledgments

Authorship is equal. We would like to thank Joshua Knobe, Shaun Nichols, Mark Isaac, and participants at the NEH Summer Institute for Experimental Philosophy for helpful comments.

References

Bernstein, M. (2002). Fatalism. In R. Kane (Ed.), *The Oxford Handbook of Free Will* (pp. 65-81). Oxford: Oxford University Press.

Cokely, E.T. (2009). Beyond generic dual processes: How should we evaluate scientific progress? *PsycCritiques,*Vol. 54, Release 51, Article 10.

Cokely, E.T., & Feltz, A. (2009). Adaptive variation in judgment and philosophical intuition. *Consciousness and Cognition, 18,* 355-357.

Cokely, E.T., & Feltz, A. (2009). Individual differences, judgment biases, and theory-of-mind: Deconstructing the intentional action side effect asymmetry. *Journal of Research in Personality, 43,* 18-24.

Cokely, E. T., & Kelley, C. M. (2009). Cognitive abilities and superior decision making under risk:

A protocol analysis and process model evaluation. *Judgment and Decision Making, 4*, 20-33.

Cokely, E.T., Kelley, C.M., & Gilchrist, A.H. (2006). Sources of individual differences in working memory: Contributions of strategy to capacity. *Psychonomic Bulletin & Review, 13,* 991-997.

Cokely, E. T., Parpart, P., & Schooler, L. J. (2009). On the link between cognitive control and heuristic processes. In N. A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31 Annual Conference of the Cognitive Science Society* (pp. 2926-2931). Austin, TX: Cognitive Science Society.

Feltz, A., & Cokely, E. (in press). Individual differences and theory-of-mind judgments: Side effects and order effects. *Philosophical Psychology*.

Feltz, A., & Cokely, E. T. (2009). Do judgments about freedom and responsibility depend on who you are?: Personality differences intuitions about compatibilism and incompatibilism. *Consciousness and Cognition, 24*, 342-350.

Feltz, A., Cokely, E. T., & Nadelhoffer, T. (2009). Natural compatibilism v. natural incompatibilism: Back to the drawing board. *Mind & Language, 24*, 1-23.

Gigerenzer, G. (1991). How to make cognitive illusions disappear: Beyond "heuristics and biases". In W. Stroebe & M. Hewstone (Eds.), *European Review of Social Psychology* (Vol. 2, pp. 83-115). Chichester: Wiley.

Gigerenzer, G. (1996). On narrow norms and vague heuristics: A rebuttal to Kahneman and Tversky (1996). Psychological Review, 103, 592–596.

Gigerenzer, G. (1998). Surrogates for theories. Theory & Psychology, 8, 195–204.

Gigerenzer, G., & Regier, T. (1996). How do we tell an association from a rule? Comment on Sloman (1996). Psychological Bulletin, 119, 23–26.

Gigerenzer, G., Todd, P. M., & the ABC Research Group. (1999). Simple heuristics that make us smart. New York, NY: Oxford University

Johnson, E. J., Haubl, G., & Keinan, A. (2007). Aspects of endowment: A query theory of value construction. Journal of Experimental Psychology: Learning Memory and Cognition, 33, 461–474.

Mele, A. (2006). *Free Will and Luck*. New York: Oxford University Press.

Miller, J., & Feltz, A., (2009). *Frankfurt and the folk: An empirical investigation of Frankfurt-style cases.* Manuscript submitted for publication.

Nahmias, E., Coates, J., & Kvaran, T. (2007). Free will, moral responsibility, and mechanism: Experiments on folk intuitions. *Midwest Studies in Philosophy, 31*, 214-242.

Nahmias, E., Morris, S., Nadelhoffer, T., & Turner, J. (2005). Surveying freedom: Folk intuitions about free will and moral responsibility. *Philosophical Psychology, 18*(5), 561-584.

Nahmias, E., Morris, S., Nadelhoffer, T., & Turner, J. (2006). Is incompatibilism intuitive? *Philosophy and Phenomenological Research, 73*(1), 28-53.

Nichols, S., & Knobe, J. (2007). Moral Responsibility and Determinism: The Cognitive Science of Folk Intuition. *Nous, 41*, 663-685.

Osman, M. (2004). An evaluation of dual-process theories of reasoning. Psychonomic Bulletin & Review, 11, 988–1010.

Smilansky, S. (2002). Free will, fundamental dualism, and the centrality of illusion. In R. Kane (Ed.), *The Oxford Handbook of Free Will* (pp. 489-505). New York: Oxford University Press.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences, 23*, 645-726.

Vohs, K., & Schooler, J. (2008). The value of believing in free will: Encouraging a belief in determinism increases cheating. *Psychological Science, 19*, 49-54.

Weber, E. U., Johnson, E. J., Milch, K. F., Chang, H., Brodscholl, J. C., & Goldstein, D. G. (2007). Asymmetric discounting in intertemporal choice: A querytheory account. Psychological Science, 18, 516–523.

Woolfolk, R., Doris, J., & Darley, J. (2006). Identification, situational constraint, and social cognition: Studies in the attribution of moral responsibility. *Cognition, 100*, 283-301.

Verb-action versus role relations congruence effects: Evidence from ERPs in picture-sentence verification

Pia Knoeferle (knoeferl@cit-ec.uni-bielefeld.de)
Cognitive Interaction Technology Excellence Cluster
Bielefeld University, 33615 Bielefeld, Germany

Thomas P. Urbach (turbach@cogsci.ucsd.edu)
Department of Cognitive Science
University of California San Diego La Jolla, USA

Marta Kutas (kutas@cogsci.ucsd.edu)
Department of Cognitive Science
University of California San Diego La Jolla, USA

Abstract

Comprehenders can rapidly use both their linguistic knowledge and different kinds of information in visual context during language comprehension. Little is known, however, about the relative time courses and mechanisms by which different kinds of visual information influence language comprehension. We recorded event-related brain potentials (ERPs) as participants read a subject-verb-object sentence and verified whether or not it matched different (verb-action versus role relations) aspects of a recently viewed picture. When the verb-action did not match the depicted action, we replicated larger N400s (300-500ms) over centro-parietal scalp to the verb (300-500 ms) relative to the responses for matches. In contrast, ERP effects to role-relation mismatches (a person depicted as undergoing an action but described as performing it) qualitatively differed from and occurred prior to the verb-action congruence N400. Our findings implicate at least two temporally distinct mechanisms governing picture-sentence verification processes.

Keywords: sentence-picture verification; visual context effects; event-related brain potentials;

Introduction

Information in visual context can rapidly influence online language comprehension and ambiguity resolution (e.g., Altmann, 2004; Knoeferle, Habets, Crocker, & Münte, 2008; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). Recently, researchers also have begun to examine picture-sentence congruence processes even when sentences are structurally unambiguous and do not necessarily globally match visual context (e.g., Knoeferle, Urbach, & Kutas, 2009; Vissers, Kolk, van de Meerendonk, & Chwilla, 2008; Wassenaar & Hagoort, 2007). The motivation for these studies is that determining the correspondence between what is said and how things are, appears to play a central part in natural language processing: Positive verification may be readily inferred from, e.g., expressions of agreement (*So I heard*) while failures to verify may be inferred from corrections and expressions of disbelief, requests for clarification and the like (e.g., *Well no, actually what happened was ...*, *Are you sure?*).

Psycholinguistic research on verification processes is by no means a recent endeavour: Just and Carpenter (1971) found that participants' verification latencies were shorter when the color of the dots on an image (red vs. black) matched than mismatched the color adjective in a corresponding sentence (henceforth "congruence effects", see also, e.g., Clark & Chase, 1972). To account for these findings and a range of others, Carpenter and Just (1975) developed the Constituent Comparison Model (CCM) of sentence-picture comparison processes. The model operates via a serial mechanism that incrementally compares representations of sentence ([AFF, (RED, DOTS)]) and picture ("black dots") constituents. The comparison proceeds from inner to outer representations (in this case right to left). When a mismatch is found (here for the inner representations), it is indexed "-"; the truth value is changed to "false", and the comparison process is reinitialized, resulting in one extra comparison step (and hence longer response times) relative to a match (e.g., "red dots"). The output of that comparator process is the truth value of the comparison and response time values.

This verification model does not specify the time course of constructing the representations of verbal information as the sentences are read, and it is unclear to which extent the constituent-wise comparator mechanism implies incremental comprehension processes. Recent event-related brain potential (ERP) research, however, suggests that congruence processing is incremental, i.e., ongoing during (not merely after) word-by-word sentence processing, and furthermore can be systematically related to end-of-sentence verification times (Knoeferle et al., 2009). This was evidenced by finding (a) reliable congruence effects in ERPs as soon as a word (e.g., the verb) that mismatched aspects of a preceding visual context (e.g., a depicted action) was encountered; (b) reliable congruence effects in verification time response latencies; and (c) reliable correlations between a participant's ERP and verification time congruence effects.

The incremental ERP congruence effects at the verb are, in principle, compatible with the CCM comparator mechanism. One may question, however, to what extent the CCM can account for verification processes during (rather than after) the sentence. Specifically, it is unclear whether all aspects of picture-sentence mismatch processing are adequately accounted for by a single comparator mechanism as suggested by the CCM.

Findings from two recent studies can be viewed as supporting a single mechanism account: highly similar ERP congruence effects were observed in response to different picture-sentence mismatches (Vissers et al., 2008; Wassenaar & Hagoort, 2007). In Wassenaar and Hagoort (2007), healthy older adults inspected a line drawing of an agent-action-patient event (e.g., man pushing woman vs. a woman pushing a man), and then listened to a spoken utterance in Dutch (e.g,. 'The tall man on [sic] this picture pushes the young woman'). Participants indicated whether the thematic relations of the utterance were congruent with the depicted role relations or not. There was no reliable response time congruence effect[1]. In the ERPs, however, healthy adults exhibited congruence effects. For active sentences as in the above example, there was a larger posterior negativity (with a non-reliable late positivity) to mismatching than matching conditions in the verb region (centro-posterior from 50-450 ms; for anterior sites from ca. 50-300 ms). Irreversible active and reversible passive sentences showed an early negativity for incongruous relative to congruous trials and a subsequent (reliable) late positivity. These effects were interpreted as reflecting thematic role assignment.

In a different study, participants inspected a line drawing containing two objects (e.g., a square followed by a circle, Vissers et al., 2008). They then read a sentence via rapid serial visual presentation (e.g., *De cirkel staat achter het vierkant*, 'The circle stands in front of the square'), and verified whether or not the object arrangement described in the sentence matched the depiction. For a first condition the location mismatch occurred within the same (horizontal) dimension (e.g., the sentence would state that the circle was in front of the square while it was in fact behind it). For a second, mismatch, the incongruence occurred between the horizontal and vertical dimensions: The sentence stated that the circle is below the square while it was in fact behind it. The authors observed an N400-P600 ERP pattern as in Wassenaar and Hagoort (2007) despite differences in the mismatches (object location rather than role relations) and language modality (written versus spoken). They interpreted the mismatch effects as reflecting monitoring of potential processing errors. Cruci ally, mean amplitudes of the ERPs in Vissers et al. did not differ in response to the two kinds of picture-sentence mismatches (200-400 ms; 500-700 ms time-locked to the critical preposition).

Based on these findings, it appears that some picture-sentence mismatches (e.g., role relations versus object location mismatches) elicit similar ERP patterns, providing - at least tentative - support for a single functional brain mechanism dealing with these incongruences (though note the different interpretations of the ERP pattern in these two studies).

In contrast with the Vissers et al. and Wassenaar and Hagoort findings, tentative support for the alternative - multiple mechanism - view comes from Knoeferle et al. (2009) in which participants read a subject-verb-object sentence and verified whether or not the verb matched a previously viewed (depicted) action. When verbs mismatched a depicted action, speeded verification response latencies were reliably longer, N400s over centro-parietal scalp to the verb were larger, and post-verbal potentials up to the time of the response (including an anterior negativity to the object) were more negative relative to the responses for matches. These different negativities across the sentence differ from the ERP congruence effects in response to role relations and object location mismatches per the absence of an ensuing P600-like congruence effect. In either case, however, our knowledge of the *relative* time course and nature of different visual context effects during sentence comprehension is relatively limited and only few studies have directly compared different visual context effects.

The present research further investigates the nature and time course of picture-sentence verification processes by directly comparing visual context effects that require interpreting a written verb in relation to an action with effects that involve interpretation of sentential role relations in relation to depicted role relations. In two Experiments, we analyzed ERPs as participants read a subject-verb-object sentence and verified whether or not the sentence matched a recently viewed visual scene. The verb either matched the previously depicted action or not; and who-does-what-to-whom in the sentence was either congruous with the depicted role relations or not, resulting in 4 (fully counterbalanced) conditions (see Table 1).

If there is a single mechanism for congruence processing we would expect to see similar ERP patterns to role relations and verb-action mismatches. Alternatively, the action and role-relation mismatches involve different mechanisms. Processing a role-relations mismatch involves comparing depicted agents and patients with a compositional interpretation, perhaps requiring more time and processing effort than relating an action to a verb interpretation. Recall, that prior research observed a verb-action congruence N400 effect (Knoeferle et al., 2009). Assuming a larger negativity reflects greater processing difficulty (see Monetta, Tremblay, & Joanette, 2003), and given that Wassenaar and Hagoort (2007) observe their first congruence effects at the verb, such an account predicts greater negative mean ERP amplitudes during the N400 region at the verb for the role relations than verb-action mismatches (and most negative for the combined mismatches), and also later ERP and verification time congruence effects.

Alternatively, role-relation effects would precede verb-action congruence effects: People likely expect the first noun phrase of a sentence to be the agent. When they read the first noun and realize that it does not refer to the character depicted as the agent, they may begin to anticipate incongruence between picture and sentence even though there is no overt mismatch at the first noun phrase; the moment the verb confirms

[1]The failure to replicate the verification time congruence effect could be due to the fact that the verification response occurred well after sentence end (but essentially this requires further investigation.)

this expectancy (specifying agent-action relationships), role (in)congruence could be confirmed and thus might elicit earlier ERP congruence effects than verb-action mismatches. To better delineate any role relations and verb-action congruence effects, we varied stimulus onset asynchrony between Experiment 1 (500 ms) and Experiment 2 (300 ms). The timing of those congruence effects that depend soley on processing associated with the first noun phrase is not expected to change substantially as a function of SOA. Alternatively, the timing of congruence effects related to information provided by the verb, is expected to vary with the interval between the noun and verb.

Experiments 1 and 2

Methods

Participants Thirty-two students of UCSD took part in Experiment 1, and a further thirty-two participated in Experiment 2. All participants were native English speakers, right-handed (Edinburgh Handedness Inventory), and had normal or corrected-to-normal vision. All gave informed consent; the experiment protocol was approved by the UCSD IRB.

Materials, design, and procedure We derived materials for both experiments from a previous study (Knoeferle et al., 2009). The present two experiments had a 2-factor *role-relation congruence* (congruent, Picture 1a/b vs. incongruent, Picture 1c/d) x *action congruence* (congruent, Picture 1a/c vs. incongruent, Picture 1b/d) within-subjects design (Table 1).

Table 1: Example of the four experimental conditions

Condition	Picture		Sentence
full match	1a		*The gymnast punches the journalist*
action mismatch	1b		*The gymnast punches the journalist*
role mismatch	1c		*The gymnast punches the journalist*
combined mismatch	1d		*The gymnast punches the journalist*

The sentence, *The gymnast punches the journalist*, in Table 1 is congruent on both action and role dimensions with Picture 1a, (full match); it is incongruent on the action but congruent on the role-relation dimension with Picture 1b (action mismatch); it is congruent on the action but incongruent on the role relations dimension with Picture 1c (role mismatch); and it is incongruent on both of these dimension following Picture 1d (combined mismatch). The materials were counterbalanced to ensure that any congruency-based ERP differences were not spuriously due to stimuli or to their presentation. There were 80 item sets which, combined with the conditions and further counterbalancing, yielded 16 experimental lists. Each list contained one occurrence of an item sentence/picture, and an equal number of left-to-right and right-to-left action depictions. Each list also contained 160 filler items, of which half were mismatches. These filler sentences had different syntactic structures including negation, clause-level and noun phrase coordination, as well as locally ambiguous reduced relative clause constructions.

Participants inspected the picture on a CRT monitor for a minimum of 3000 ms terminated via a right thumb button press. Next, a fixation dot was presented for a random duration between 500 and 1000 ms, followed by the sentence, one word at a time. Word onset asynchrony was 500 ms in Experiment 1 and 300 ms in Experiment 2; word duration was 200 ms in both. Participants were instructed to examine the picture and then to read the sentence in the context of the preceding picture. Participants indicated via a button press as quickly and accurately as possible after each sentence whether it matched or did not match the preceding picture. After that button press, there was a randomly varying pause between 500 and 1000 ms prior to the next trial.

Analysis We report analyses of variance (ANOVA) on response latencies and mean amplitude ERPs. Time regions for the ERP analyses were: the first noun; the verb, and the post-verbal object noun. We performed omnibus repeated measures ANOVAs on mean ERP amplitudes (averaged by participants for each condition at each electrode site) with role congruence (mismatch vs. match), action congruence (mismatch vs. match), hemisphere (left vs. right electrodes), laterality (lateral vs. medial), and anteriority (5 levels) as factors. The pre-stimulus baseline for all analyses was 200 ms. Time windows (0-100, 100-300, 300-500) were chosen based on prior studies and visual inspection of waveforms.

Results Experiment 1 (500 ms SOA)

Behavioural results Repeated measures ANOVAS for the verification latencies showed that response times were marginally faster for the action match than mismatch conditions (1115 ms vs. 1163 ms, $p = 0.06$), while there was no reliable effect for the role relations factor ($p > 0.2$); the interaction between these two factors was reliable ($p < 0.01$).

The response latency data replicate findings of a verb-action congruence effect (Knoeferle et al., 2009) as well as the absence of verification time congruence effects for role relations mismatches (Wassenaar & Hagoort, 2007).

ERP results We present grand average ERPs at prefrontal, parietal, temporal, and occipital sites for all four conditions (Fig. 2) and for mean amplitude role mismatches versus matches (Fig. 3).

For the role relations factor, differences emerged early, during the first noun phrase. ERPs for role mismatches were more negative beginning about 200 ms after noun onset (Figure 3), with the effect more pronounced at lateral electrodes over right anterior scalp (100-300 ms, $p < 0.05$). In line

with early (200-400ms) mismatch effects observed by Vissers et al., we also measured ERPs from 200-400 ms at the first noun. Analyses revealed more negative going ERPs to role mismatches than matches ($p < 0.01$). Following the anterior negativity, a relative positivity for mismatches was observed, largest over posterior scalp, beginning around 400ms after noun onset and continuing beyond the onset of the subsequent verb. This effect was reliable from 0-100 ms and 100-300 ms following the verb ($p < 0.01$). These role congruence effects were also reliable when analyzed relative to a pre-noun baseline. They did, however, not last into the later portion of the verb (300-500 ms).

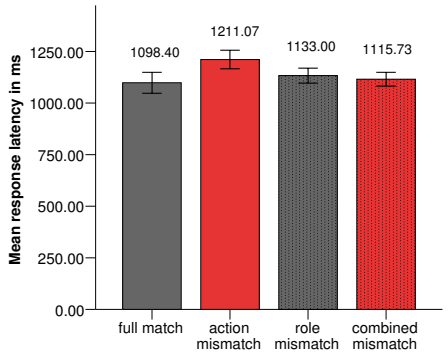


Figure 1: Experiment 1 verification response times (error bars indicate 95% confidence interval)

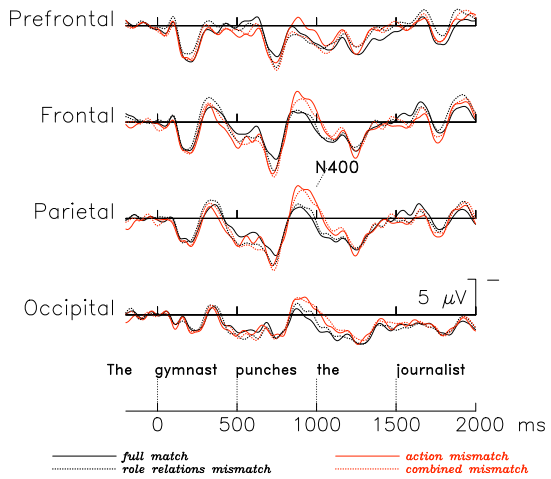


Figure 2: Grand average ERPs (mean amplitude) for all four conditions across the sentence at prefrontal, parietal, temporal, and occipital sites (Experiment 1)

For action mismatches, the first reliable effects occurred at the verb, where we replicated larger mean amplitude ERPs to action mismatches than matches with a a centro-parietal maximum (300-500 ms, $p < 0.001$, see Fig. 2 Knoeferle et al., 2009). The reliable verb-action congruence effect in this window (300-500 ms at the verb and the absence of a

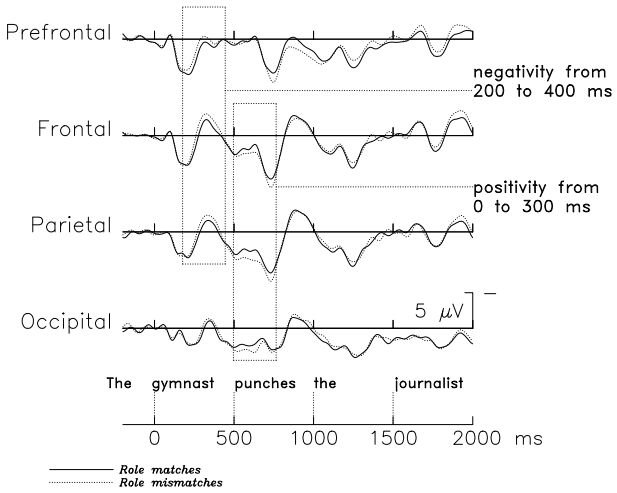


Figure 3: Grand average mean amplitude ERPs for role mismatching conditions versus role matching conditions across the sentence at prefrontal, parietal, temporal, and occipital sites (Experiment 1)

role-relation effect a lead to an interaction between these two factors ($p < 0.05$). During the second noun (300-500 ms), the role mismatches were more negative-going than the role matches ($p < 0.05$).

Results Experiment 2 (300 ms SOA)

Analyses of verification time latencies revealed no reliable effects of the manipulated factors (see Fig. 4).

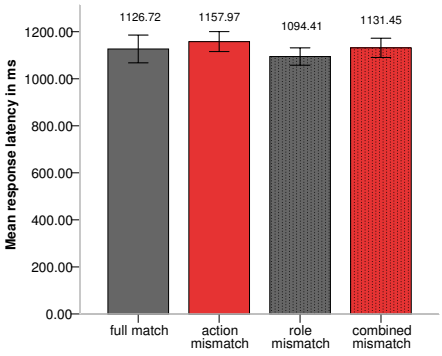


Figure 4: Response latencies in ms for Experiment 2

For the ERPs, the earliest effect of a role mismatch appears to be a broadly distributed relative negativity that reached a maximum between about 300 and 400ms, i.e., shortly after the verb onset (Figure 5). These role congruence effects and occurred from 0-100 and 100-300 ms after verb onset (i.e., 300-600 ms after noun onset).

In these early verb time windows (0-100, 100-300 ms) role mismatches were more negative than role matches ($p < 0.001$, see Fig. 5). That negativity is confirmed when analysing the data re-baselined relative to the first noun (300-500 ms and 200-400 ms ($p < 0.01$). Analysis of the time

window 300-500 ms post-verb found no role mismatch effects ($F < 0.2$). There were no further reliable role relations congruence effects except for more negative ERPs for mismatching than matching trials during the post-verbal object noun (second noun: 400-600 ms, $p < 0.05$).

For the action mismatches, the effects in Experiment 2 appeared after the verb (300-500 ms, $p < 0.001$, see Fig. 6) just as in Experiment1, leading to a reliable interaction of role and action congruence ($ps < 0.01$). Post-verbally, the verb-action congruence negativity continued into the determiner and object noun (see Fig. 6).

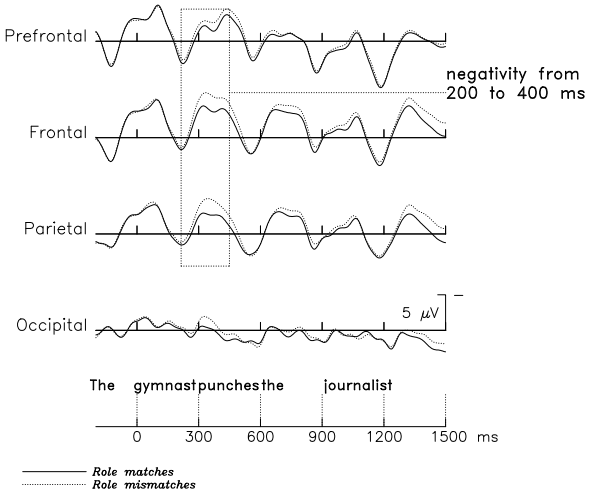


Figure 5: Grand average mean amplitude ERPs scores for role relations mismatches versus matches across the sentence at prefrontal, parietal, temporal, and occipital sites (Experiment 2)

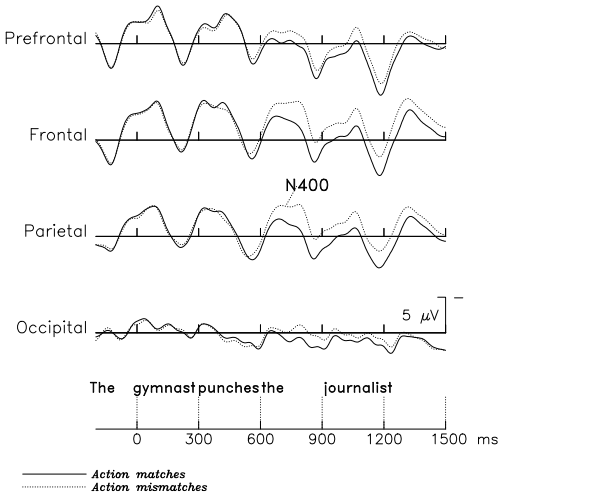


Figure 6: Experiment 2: Grand average mean amplitude ERPs scores for action mismatches versus matches across the sentence at prefrontal, parietal, temporal, and occipital sites (Experiment 2)

In Experiment 2, role mismatch effects again clearly preceded verb-action mismatch effects. Although the role mismatch effect was more broadly distributed and laterally symmetric than the early role congruence negativity in Experiment 1, both effects had the same polarity and a similar time course and neither exhibited the posterior, right lateralized maximum frequently observed for N400 effects.

General Discussion

The present ERP experiments compared role-relation and verb-action congruence processing in a picture-sentence verification task, and examined whether they differed in their natures and/or time courses. Verification time congruence effects for verb-action mismatches (longer response times for action mismatches relative to matches) were replicated (marginal effect) at the longer SOA (Exp 1) and were not reliable at the shorter SOA (Exp 2). By contrast, role match and mismatch response times did not differ at either SOA. ERPs, however, revealed reliable revealed reliable but different effects for both role and action mismatches (*vs* matches)

The earliest role mismatch effects were seen within a few hundred milliseconds of the first noun onset at both SOAs. By contrast, reliable effects of action mismatches were observed only later, a few hundred milliseconds after verb onset. Although the action mismatch effect also was a broadly distributed relative negativity to the mismatches, it tended to be larger over posterior scalp (as is characteristic of visual N400) whereas the role relation mismatch effect was not. At the longer SOA (only) the role relation congruence negativity was followed by a reliable positivity over posterior scalp that continued past the onset of the next word (verb).

As in Knoeferle et al. (2009) we find ongoing ERP congruence effects across the sentence suggest that verification-related processes are part of ongoing incremental sentence interpretation. We observe effects of the action-verb mismatch at the verb, continuing into the second determiner and object noun (see also Ferretti, Singer, & Patterson, 2008; Singer, 2006, for related evidence on text verification). The overall morphology, latency, and centro-parietal distribution of the N400 is similar to that for lexico-semantic anomalies or low cloze probability words in sentences read for comprehension (e.g., Kutas, 1993; Kutas, Van Petten, & Kluender, 2006; Berkum, Hagoort, & Brown, 1999).

Conclusions Our findings are consistent with verification models on which depicted information modulates processing of verbal information as sentences unfold word by word. In the context of a just-viewed depicted action in which a journalist is punching a gymnast, there is nothing incongruous or anomalous about a sentence that begins with *The gymnast* .... People could have waited until they read the verb before assigning a thematic role to the first noun phrase. It seems, however, that when they read the first noun and realized that it referred to a character that had not been depicted as the agent of an action, their expectations of thematic role assignment (i.e., that the first noun in a sentence often refers to the thematic agent) conflicted with their visual context representation (of that character as a patient). Such incongruence may

have led to the larger negativity for role relations mismatches at the first noun. The subsequent centro-parietal positivity elicited by role relations mismatches may be a P600, related to the revision of thematic role assignments though, if so, it is unclear why it did not replicate in Experiment 2.

Furthermore, although action and role relations mismatches were both evident at the verb (mismatching the action; identifying the first noun phrase as a role filler that mismatches its role in the picture, respectively), critically, the time course of their effects differed, as did - at least for the positivity during the early verb in Experiment 1 - polarity. Role mismatch effects were further absent in the later time window at the verb for which we found the verb-action congruence N400 effect. The reliable interaction of role and action congruence suggests these two effects are dissociable. To the extent that a single mechanism account does not straightforwardly predict this dissociation, our findings appear to accord better with the view that multiple functional brain mechanisms govern visual context effects during online language comprehension.

Neither the ERP nor verification time data confirmed the complexity account which predicted substantially longer verification latencies for role than action mismatches. In both studies, verification times to the role relations conditions were no longer than those to action mismatches. A complexity account also predicts larger (and possibly delayed) negative mean ERP amplitudes for role mismatches (combined mismatch and role mismatch) relative to action mismatches (action mismatch and combined mismatch, 300-500 ms at the verb, e.g., Fig. 2). This also was not what we observed.

Why then did we find a difference in ERP effects for a role relations mismatch relative to verb-action mismatch effects, while prior research has failed to find differences between ERP congruence effects in response to such - at first blush - different mismatches as object locations versus role relations (Vissers et al., 2008; Wassenaar & Hagoort, 2007)? First, prior studies did not compare object location with role relations mismatches directly. A theoretically more interesting possibility is that for both the role relations and object location mismatches, re-processing involves restructuring of mental representations (spatially and/or in terms of thematic role relations) whereas for our verb-action mismatches, re-processing concerned lexico-semantic content (rather than the structure) of mental representations.

In sum, we find that the time course of visual context influences on language comprehension can vary as a function of which aspects of a picture (role relations versus actions) mismatch corresponding aspects of a sentence. The findings best align with an incremental account of comprehension in picture-sentence verification.

Acknowledgments

This research was funded by a postdoctoral fellowship to PK (German research foundation, DFG) and by NIH grants HD-22614 and AG-08313 to Marta Kutas. The studies were conducted while PK was at UC San Diego.

References

Altmann, G. T. M. (2004). Language-mediated eye-movements in the absence of a visual world: the 'blank screen paradigm'. *Cognition*, *93*, B79–B87.

Berkum, J. van, Hagoort, P., & Brown, C. M. (1999). Semantic integration in sentences and discourse: evidence from the n400. *Journal of Cognitive Neuroscience*, *11*, 657–671.

Carpenter, P. A., & Just, M. A. (1975). Sentence comprehension: a psycholinguistic processing model of verification. *Psychological Review*, *82*, 45–73.

Clark, H. H., & Chase, W. G. (1972). On the process of comparing sentences against pictures. *Cognitive Psychology*, *3*, 472–517.

Ferretti, T. R., Singer, M., & Patterson, C. (2008). Electrophysiological evidence for the time course of verifying text ideas. *Cognition*, *108*, 881-888.

Just, M. A., & Carpenter, P. A. (1971). Comprehension of negation with qualification. *Journal of Verbal Learning and Verbal Behavior*, *10*, 244–253.

Knoeferle, P., Habets, B., Crocker, M. W., & Münte, T. F. (2008). Visual scenes trigger immediate syntactic reanalysis: evidence from ERPs during situated spoken comprehension. *Cerebral Cortex*, *18*, 789–795.

Knoeferle, P., Urbach, T. P., & Kutas, M. (2009). Is incremental semantic interpretation related to end-of-sentence verication?: Evidence from correlation analyses. In N. Taatgen, H. van Rijn, L. Schomaker, & J. Nerbonne (Eds.), *Proceedings of the Annual Conference of the Cognitive Science Society* (p. 1127-1132). Cognitive Science Society, Inc.

Kutas, M. (1993). In the company of other words: Electrophysiological evidence for single-word and sentence context effects. *Language and Cognitive Processes*, *8*, 533–572.

Kutas, M., Van Petten, C., & Kluender, R. (2006). Handbook of psycholinguistics. In M. Traxler & M. Gernsbacher (Eds.), (2nd Edition ed., p. 659-724). New York: Elsevier.

Monetta, L., Tremblay, T., & Joanette, Y. (2003). Semantic processing of words, cognitive resources and n400: An event-related potentials study. *Brain and Cognition*, *53*, 327–330.

Singer, M. (2006). Verification of text ideas during reading. *Journal of Memory and Language*, *54*, 574-591.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*, 632–634.

Vissers, C., Kolk, H., van de Meerendonk, N., & Chwilla, D. (2008). Monitoring in language perception: evidence from erps in a picture-sentence matching task. *Neuropsychologia*, 967-982.

Wassenaar, M., & Hagoort, P. (2007). Thematic role assignment in patients with broca's aphasia: sentence-picture matching electrified. *Neuropsychologia*, *45*, 716-740.

Integrating Syntactic Knowledge into a Model of Cross-situational Word Learning

Afra Alishahi
Computational Linguistics and Phonetics
Saarland University, Germany
afra@coli.uni-saarland.de

Afsaneh Fazly
Computer Sciences and Engineering
Shiraz University, Iran
fazly@cse.shirazu.ac.ir

Abstract

It has been suggested that children learn the meanings of words by observing the regularities across different situations in which a word is used. However, experimental studies show that children are also sensitive to the syntactic properties of words and their context at a young age, and can use this information to find the correct referent for novel words. We present a unified computational model of word learning which integrates cross-situational evidence with the accumulated semantic properties of the lexical categories of words. Our experimental results show that using lexical categories can improve performance in learning, particularly for novel or low-frequency words in ambiguous situations.

Learning the Meaning of Words

In the course of learning a language, children need to learn mappings between words and their meanings, mostly from noisy and ambiguous contexts. It has been suggested that children learn the meanings of words by observing the regularities across different situations in which a word is used, or the *cross-situational evidence* (Quine, 1960; Pinker, 1989). Experimental studies on children and adult learners have shown that both groups are sensitive to cross-situational evidence, and can efficiently use it to deduce the correct meanings of novel words in ambiguous situations (Smith & Yu, 2007; Monaghan & Mattock, 2009). Moreover, many computational models have demonstrated that cross-situational learning is a powerful and efficient mechanism for learning the correct mappings between words and meanings, and can explain several behavioural patterns observed in children (Siskind, 1996; Yu, 2005; Fazly et al., 2008).

Another valuable source of information for mapping words to meanings is the syntactic structure of the sentence that a word appears in. There is substantial evidence that children are sensitive to the structural regularities of language from a very young age, and that they use these structural cues to find the referent of a novel word (e.g. Naigles & Hoff-Ginsberg, 1995; Gertner et al., 2006), a hypothesis known as syntactic bootstrapping (Gleitman, 1990). The syntactic bootstrapping account is in accordance with children's early sensitivity to distributional properties of language: one-year-old infants can recognize sentences from an artificial grammar after a short period of exposure (Gomez & Gerken, 1999), and 2 to 3-year-olds demonstrate robust knowledge of some of the abstract lexical categories such as nouns and verbs (e.g., Gelman & Taylor, 1984; Kemp et al., 2005). Therefore, it is likely that they draw on their knowledge of the structural regularities of language (and of lexical categories in particular) to facilitate word learning, especially in cases where cross-situational evidence is not reliable. However, a coherent account of word learning that explains the interaction between these two information sources is lacking. Also, despite the extensive body of experimental research on the role of syntactic knowledge in semantics acquisition, few computational models have been developed to explore the usefulness of lexical categories in learning word meanings (but see Yu, 2006).

We present a probabilistic model of word learning which integrates cross-situational evidence and the knowledge of lexical categories into a single learning mechanism. We use an existing computational model of cross-situational learning proposed by Fazly et al. (2008), and augment it with the syntactic categories of words. Our computational simulations show that such information can improve the model's performance in learning words. Especially, the results suggest that the syntactic category of a word and the context the word appears in provide complementary information for the acquisition of word–meaning mappings.

Related Computational Models

A number of computational word learning models have used cross-situational learning as their core mechanism for mapping words to meanings. The rule-based model of Siskind (1996) and the probabilistic models of Yu (2005) and Fazly et al. (2008) all rely on the regularities of the co-occurrences of words and meaning elements, successfully learning word meanings from noisy and ambiguous data. Moreover, these models simulate several behavioural patterns observed in children, such as vocabulary spurt, fast mapping, and learning synonymy and homonymy. However, all these models ignore the syntactic properties of the utterances and treat them as unstructured bags of words.

There are only a few existing computational models that explore the role of syntax in word learning. Maurits et al. (2009) has investigated the joint acquisition of word meaning and word order using a batch model. This model is tested on an artificial language with a simple relational structure of word meaning, and limited built-in possibilities for word order. The Bayesian model of Niyogi (2002) simulates the bootstrapping effects of syntactic and semantic knowledge in verb learning, i.e., the

use of syntax to aid in inducing the semantics of a verb, and the use of semantics to narrow down possible syntactic forms in which a verb can be expressed. However, this model relies on extensive prior knowledge about the associations between syntactic and semantic features, and is tested on a toy language with very limited vocabulary and a constrained syntax. Yu (2006) integrates information about syntactic categories of words into his model of cross-situational word learning, showing that this type of information can improve the model's performance. Yu's model also processes input utterances in a batch mode, and its evaluation is limited to situations in which only a coarse distinction between referring words (words that could potentially refer to objects in a scene, e.g., concrete nouns) and non-referring words (words that cannot possibly refer to objects, e.g., function words) is sufficient. It is thus not clear whether information about finer-grained categories (e.g., verbs and nouns) can indeed help word learning in a more naturalistic incremental setting.

An Overview of Our Integrated Model

Consider a young language learner hearing the sentence *the kittie is playing with the yarn*, and trying to find out the meaning of *yarn*. Usually there are many possible interpretations for *yarn* based on the surrounding scene, and the child has to narrow them down using some learning strategy. One such method is to register the potential meanings in the current scene, and compare them to those inferred from the previous usages of the same word (i.e., cross-situational learning). Another way to make an informed guess about the meaning of *yarn* is to pay attention to its syntactic properties. For example, if the child has already heard some familiar words in a similar syntactic context (e.g., *daddy is playing with the ball*, *the kittie is sniffing the slipper*), she can conclude that a group of words which can appear in the context "*is Xing the –*" usually refer to physical objects. Therefore *yarn* must refer to one of the objects present in the scene, and not for example to an action or a property.

We present a computational model that combines these two complementary approaches into a single mechanism of word learning. Our goal is to examine whether using the knowledge of word categories in addition to cross-situational observations can improve the performance in word learning. We use the computational model of Fazly et al. (2008) as the base model of cross-situational learning: the model learns word meanings as probabilistic associations between words and semantic elements, using an incremental and probabilistic learning mechanism, and drawing only on the word–meaning co-occurrence statistics gradually collected from naturally-occurring child-directed input. This model has been shown to accurately learn the meaning of a large set of words from noisy and ambiguous input data, and to exhibit patterns similar to those observed in children in a variety of tasks (see Fazly et al., n.d., for a full set of experiments on this model).

In order to augment the base model with category knowledge, we assume that an independent categorization module can process each sentence and determine the lexical category for each word, e.g., based on its surrounding context. That is, we make the simplifying assumption that prior to the onset of word learning, the categorization module has already formed a relatively robust set of lexical categories from an earlier set of child-directed data. This assumption is on the basis of previous empirical findings that young children gradually form a knowledge of abstract categories, such as verbs and nouns (e.g., Gelman & Taylor, 1984). In addition, several computational models have been proposed for inducing reliable categories of words by drawing on distributional properties of their context (see, e.g. Parisien et al., 2008). However, children's acquisition of categories is most probably interleaved with the acquisition of word meaning, and these two processes must be studied simultaneously. As a first step, we investigate whether the word learning process can benefit from the knowledge of lexical categories, assuming that such knowledge exists.

In the next sections we sketch the base model of cross-situational learning, and explain how we extend it to integrate lexical categories as an alternative source of guidance. During the course of learning in both models, we use the feedback from the categorization model to detect different senses of the same word. That is, the same word types which belong to different categories are represented as separate lexical items. For example, the verb sense and the noun sense of the word *cry* are mapped to two independent meaning representations.

Cross-situational Learning

This section explains the details of the cross-situational word learning model of Fazly et al. (2008), which we use as our base model.

Representation of Input

The input to our word learning model consists of a set of utterance–scene pairs that link an observed scene (what the child perceives) to the utterance that describes it (what the child hears). We represent each utterance as a sequence of words, and the corresponding scene as a set of semantic features, for example:

He hit the rabbit { ANIMATE, MALE PERSON, ACT, MOTION, CONTACT, FORCE, ANIMAL, MAMMAL, RABBIT }

In the Evaluation section, we explain how the utterances and the corresponding semantic features are selected.

Given a corpus of such utterance–scene pairs, our model learns the meaning of each word w as a probability distribution $p(.|w)$ over the semantic features appearing in the corpus. In this representation, $p(f|w)$ is the probability of feature f being part of the meaning of word w.

In the absence of any prior knowledge, all features can potentially be part of the meaning of all words. Hence, prior to receiving any usages of a given word, the model assumes a uniform distribution over semantic features as its meaning.

The Learning Algorithm

The model proposes a probabilistic interpretation of cross-situational learning (Quine, 1960) through an interaction between two types of probabilistic knowledge acquired and refined over time. Given an utterance–scene pair ($\mathrm{U}^{(t)}$, $\mathrm{S}^{(t)}$) received at time t, the model first calculates an alignment probability a for each $w \in \mathrm{U}^{(t)}$ and each $f \in \mathrm{S}^{(t)}$, using the meaning $p(.|w)$ of all the words in the utterance prior to this time. It then revises the meaning of the words in $\mathrm{U}^{(t)}$ by incorporating the alignment probabilities for the current input pair. This process is repeated for all the input pairs, one at a time.

Step 1: Calculating the alignment probabilities. For a feature $f \in \mathrm{S}^{(t)}$ and a word $w \in \mathrm{U}^{(t)}$, the higher the probability of f being part of the meaning of w (according to $p(f|w)$), the more likely it is that f is aligned with w in the current input. In other words, $a(w|f, \mathrm{U}^{(t)}, \mathrm{S}^{(t)})$ is proportional to $p^{(t-1)}(f|w)$. In addition, if there is strong evidence that f is part of the meaning of another word in $\mathrm{U}^{(t)}$—i.e., if $p^{(t-1)}(f|w_k)$ is high for some $w_k \in \mathrm{U}^{(t)}$ other than w—the likelihood of aligning f to w should decrease. Combining these two requirements:

$$a(w|f, \mathrm{U}^{(t)}, \mathrm{S}^{(t)}) = \frac{p^{(t-1)}(f|w)}{\sum_{w_k \in \mathrm{U}^{(t)}} p^{(t-1)}(f|w_k)} \quad (1)$$

Note that a feature can have a non-zero alignment with more than one word in an utterance. For example, if two concrete nouns occur in a sentence, they both need to be aligned with the single feature ARTIFACT.

Step 2: Updating the word meanings. We need to update the probabilities $p(.|w)$ for all words $w \in \mathrm{U}^{(t)}$, based on the evidence from the current input pair reflected in the alignment probabilities. We thus add the current alignment probabilities for w and the features $f \in \mathrm{S}^{(t)}$ to the accumulated evidence from prior co-occurrences of w and f. We summarize this cross-situational evidence in the form of an association score, which is updated incrementally:

$$\mathrm{assoc}^{(t)}(w, f) = \mathrm{assoc}^{(t-1)}(w, f) + a(w|f, \mathrm{U}^{(t)}, \mathrm{S}^{(t)})$$

where $\mathrm{assoc}^{(t-1)}(w, f)$ is zero if w and f have not co-occurred before. The model then uses these association scores to update the meaning of the words in the current input, as in:

$$p^{(t)}(f|w) = \frac{\mathrm{assoc}^{(t)}(f, w)}{\sum_{f_j \in \mathcal{F}} \mathrm{assoc}^{(t)}(f_j, w)} \quad (2)$$

where $\mathcal{F}$ is the set of all features seen so far. We use a smoothed version of this formula to accommodate noisy or rare input, as explained in Fazly et al. (n.d.).

Word Acquisition Score

To evaluate our model, we need to verify how accurately the model learns the meaning of words. We thus define the *acquisition score* of a word w at time t as an estimation of how closely the meaning probability $p^{(t)}(.|w)$ resembles the *correct* meaning of w, or $\mathcal{T}_w$. The correct meaning of a word is a set of semantic features according to an input-generation lexicon.[1] Ideally, a word is accurately learned when its relevant semantic features (those in $\mathcal{T}_w$) are ranked at the very top of the distribution $p^{(t)}(.|w)$. We use average precision[2] to measure how well $p^{(t)}(.|w)$ separates the relevant features of w from irrelevant ones.

Adding Lexical Categories to the Model

As mentioned before, we assume that prior to the onset of word learning, the child has formed a number of lexical categories, each containing a set of word forms. More formally, we assume that the word learning model has access to a categorization function $\mathrm{cat}(w, \mathrm{U}^{(t)})$ which at any time t during the course of learning can determine the category of a word w in utterance $\mathrm{U}^{(t)}$. We do not make any assumptions about the details of the categorization process, except that it does not rely on the meaning of words in order to find their appropriate category.

As the model learns meanings for words, the categories that these words belong to are implicitly assigned a meaning as well. Once the word learning process begins, we assign a meaning distribution to each category on the basis of the meanings learned for its members. Formally, we define the meaning of a category C as the average of the meaning distributions of its members, as in:

$$p^{(t)}(f|C) = \frac{1}{|C|} \sum_{w \in C} p^{(t)}(f|w) \quad (3)$$

where $|C|$ is the number of word forms in category C, and $p^{(t)}(f|w)$ is the meaning probability of word w for feature f at time t. Prior to observing any instances of the members of a category in the cross-situational input, we assume a uniform distribution over all the possible meaning elements for each category.

[1] The model does not have access to this lexicon for learning; it is used only for input generation and evaluation.

[2] Precision is calculated as the proportion of the number of features from $\mathcal{T}_w$ to the total number of features at each cut-off point in the ranked list $p^{(t)}(.|w)$. The acquisition score is the average over the precisions for all the cut-off points up to the point where all the features in $\mathcal{T}_w$ are included in the ranked list. Note that this score is 1 when the probabilities assigned to all of the relevant features of w are higher than those assigned to the irrelevant features.

Using Categories in Alignment

Knowledge of word categories is integrated into the base model in the alignment phase (i.e. **Step 1** of the learning algorithm), where we decide which semantic feature in an observed scene must be aligned with which word(s) in the accompanying utterance. Given a new utterance–scene pair, we can align words in the utterance with the semantic features in the observed scene based on the cross-situational evidence that we have accumulated so far. Alternatively, we can find the category for each word and use the meaning associated with the word category as a guide to align it with the best matching semantic features from the scene. We can merge these two pieces of information into an extended version of Eqn. (1):

$$\begin{aligned} a(w|f,\, \mathrm{U}^{(t)},\, \mathrm{S}^{(t)}) \quad &= \mathit{weight}(w) \cdot a_w(w|f,\, \mathrm{U}^{(t)},\, \mathrm{S}^{(t)}) \\ + \quad &(1 - \mathit{weight}(w)) \cdot a_c(w|f,\, \mathrm{U}^{(t)},\, \mathrm{S}^{(t)}) \end{aligned}$$

The word-based alignment score $a_w(w|f,\, \mathrm{U}^{(t)},\, \mathrm{S}^{(t)})$ is calculated as in Eqn. (1). The category-based alignment score $a_c(w|f,\, \mathrm{U}^{(t)},\, \mathrm{S}^{(t)})$ is calculated in a similar fashion, except it relies on the meaning of the word category:

$$a_c(w|f,\, \mathrm{U}^{(t)},\, \mathrm{S}^{(t)}) = \frac{p^{(t-1)}(f|\mathrm{cat}(w, \mathrm{U}^{(t)}))}{\sum_{w_k \in \mathrm{U}^{(t)}} p^{(t-1)}(f|\mathrm{cat}(w_k, \mathrm{U}^{(t)}))}$$

where the meaning probability $p^{(t-1)}(f|\mathrm{cat}(w_k, \mathrm{U}^{(t)}))$ is calculated as in Eqn. (3).

The relative contribution of the word-based versus the category-based alignment is determined by the function $\mathit{weight}(w)$. It has been shown that cross-situational evidence is a reliable cue for frequent words: Fazly et al. (n.d.) show that once their model receives a few instances of a word form, it can reliably align it with proper semantic features. On the other hand, the category-based score is most informative when the model encounters a low-frequency word. Therefore, we define $\mathit{weight}(w)$ as a function of the frequency of w:

$$\mathit{weight}(w) = \frac{\mathit{freq}(w)}{\mathit{freq}(w) + 1}$$

Once the overall alignment score is calculated for the new input pair, the meaning probabilities of words are updated through **Step 2** of the original learning algorithm, and the meaning of their corresponding categories are updated accordingly.[3]

Evaluation

The training data for our model consists of a sequence of utterances, each paired with a set of semantic features. We extract utterances from the Manchester corpus (Theakston et al., 2001) in the CHILDES database

[3]For each word w in $U^{(t)}$, the meaning distribution of the corresponding category C is incrementally updated as $p^{(t)}(f|C) = p^{(t-1)}(f|C) + \frac{1}{|C|}(p^{(t)}(f|w) - p^{(t-1)}(f|w))$.

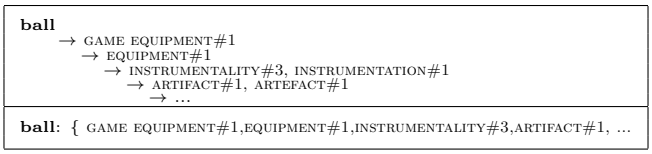


Figure 1: Semantic features for *ball* from WordNet.

(MacWhinney, 1995), which contains transcripts of conversations with children between the ages of 1;8 and 3;0. We use the mother's speech from transcripts of 6 children, remove punctuation and lemmatize the words, and concatenate the corresponding sessions as our test data. We automatically construct a scene representation for each utterance based on the semantic features of the words in that utterance. For nouns, we extract the semantic features from WordNet[4] as follows: We take all the hypernyms (ancestors) for the first sense of the word, and add the first word in the synset of each hypernym to the set of the semantic features of the target word (see Figure 1 for an example). For verbs, we extract features from WordNet as well as from a verb-specific resource, VerbNet.[5] For adverbs, adjectives and closed class words we use the features of Harm (2002). Words not found in these three resources are removed from the utterance.

To form the initial lexical categories, we use a non-overlapping portion of the part-of-speech tagged version of the Manchester corpus. The original corpus has 60 fine-grained tags, which we map to 11 coarser-grained categories, such as Noun, Verb, and Preposition.[6]

Learning Curves

To understand whether category information improves learning of word–meaning mappings, we compare the pattern of word learning over time for two models: the base model which only uses cross-situational evidence, and the extended model which incorporates lexical categories into learning. For each model we measure the average acquisition score (defined on page 3) of all words that the model has encountered up to each point in time.

Figure 2 shows the learning curve for each model over 5000 time units (or processed input pairs). The curves show that the extended model consistently outperforms the base model. The improvement is more pronounced as the model receives more input, since by learning more about the meanings of words the model also forms a more reliable knowledge about the meanings of categories and can use them more efficiently in aligning the novel words with their referents.

[4]`http://wordnet.princeton.edu/`

[5]`http://verbs.colorado.edu/~mpalmer/projects/verbnet.html`

[6]We thank Chris Parisien for providing us with the coarse-grained tagging of the corpus.

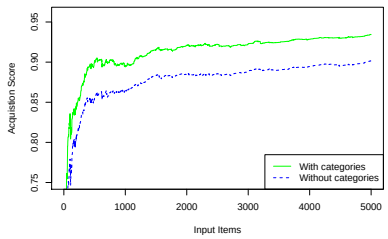


Figure 2: Avg. acquisition score for all words over time, with and without using lexical categories.

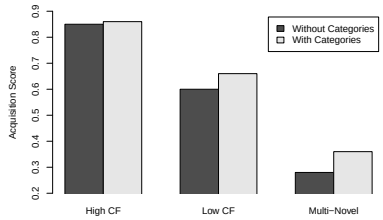


Figure 3: Avg. acquisition score for words in contexts with different degrees of familiarity.

Categories and Context Familiarity

The learning curves presented above show an overall improvement when lexical categories are incorporated in word learning. However, we expect the gain from including categories to vary across different situations. For example, experimental and computational studies have shown that cross-situational learning can account for accurate mapping of a novel word to a novel object in a familiar context (see Fazly et al. (n.d.) for a discussion on this phenomenon). The same pattern is expected in our base model, where the alignment between a word and a semantic feature is in part determined by what the model has learned about the possible meanings of the co-occurring context words (see Eqn.1). Therefore, it can learn a lot about a novel word from a single exposure if that word appears in a familiar context.

We hypothesize that categories can be particularly helpful in cases where a novel word first appears in an unfamiliar context (where not all words in the utterance are accurately learned), or when an utterance contains more than one novel word. To investigate this hypothesis, we introduce a context familiarity measure CF as the mean *familiarity* of all words that co-occur with a target word, where the familiarity of a word is determined by its frequency range. The mappings between familiarity values and frequency ranges are as follows: 0 (0), 1 (1), 2 (2–4), 3 (5–9), 4 (10–29), and 5 ($\geq$ 30), where the numbers in parentheses specify the frequency range.

Figure 3 shows the average acquisition score of words with high and low context familiarity (CF $\leq$ 3 vs. CF $>$ 3), and for novel words which appear in the company of other novel words (this last condition is marked as Multi-Novel in Figure 3). The average scores are calculated by both models after the first occurrence of each word. As can be seen, the inclusion of categories leads to a statistically significant improvement for words in all three conditions ($p < 0.05$).[7] However, the improvement is much more pronounced for words with low context familiarity, and particularly when an utterance includes more than one novel word (i.e. a highly unfamiliar context). These results support our hypothesis, and suggest that in learning the meaning of words, the context of a word and its lexical category can be seen as complementary sources of knowledge.

[7]The p-values are measured according to a two-sided sign test for a confidence interval of 95%.

Comparing Different Categories

To better understand the impact of lexical categories on word learning, we examine the pattern of improvement for words with different parts of speech. Lexical categories differ in their frequency of occurrence and in their semantic properties. For example, open-class categories such as Verb and Noun tend to have lower token frequency, higher type frequency, and more within-class meaning variability compared to closed-class categories such as Determiner and Preposition.

Recall that words in our test corpus are tagged with one of 11 coarse-grained parts of speech. Three of these categories (Auxiliary, Infinitive and Negation) each contain only a single word type, and one (Other) is not a coherent and meaningful category. The average acquisition score in both models for the remaining categories are shown in Figure 4. Out of these seven categories, four are open-class: Noun (599 word types), Verb (261), Adjective (60), and Adverb (25), and three are closed-class: Determiner (23), Preposition (17), and Conjunction (8).

Interestingly, we observe that category information helps more with the acquisition of open-class words, in particular Noun ($p < 10^{-16}$) and Verb ($p < 0.0001$). We believe this difference is due to the high token frequency of closed-class words which makes them very easy to learn, even for the base model that does not take into account the information about their categories. Moreover, using categories does not significantly improve the acquisition of adjectives and adverbs. We suspect that this is a result of the small number and the inconsistent meaning representations of these categories in the resource of Harm (2002). In general, we predict that using better resources for extracting semantic features will boost the contribution of lexical categories in word learning.

Conclusions and Future Directions

Our computational model of word learning demonstrates the advantage of integrating lexical categories into a cross-situational model of word learning. Drawing on

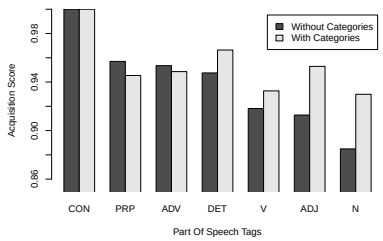


Figure 4: Avg. acquisition score for different categories.

the meaning probabilities of individual words, the model gradually associates each lexical category with a meaning representation, which in turn can boost learning of novel words. Our simulations of the model over the course of acquisition show that using lexical categories consistently improves learning over a base model which only relies on cross-situational evidence. Moreover, our analyses of the results suggest that lexical categories can have a significant impact on the acquisition of open-class words which appear in less familiar context.

The model in its current form makes simplifying assumptions that must be addressed in future work. It is assumed that lexical categories are formed prior to the onset of the word learning process, and that the category of each word can be precisely determined upon its first appearance in the input data. In the future, we intend to use an incremental model of category induction to simultanously learn lexical categories and word meanings. In fact, using a finer-grained set of categories induced by such a model might be more suitable for our purpose, since they can represent more specialized meanings (e.g., fruits and animals instead of nouns). Moreover, the categorization process can benefit from the integration of word meanings in addition to the distributional context. This extension will allow us to study how the early stages of word learning and category formation interact.

Acknowledgment

We would like to thank Grzegorz Chrupała for his invaluable help, and the anonymous reviewers for their insightful comments on our paper. Afra Alishahi was funded by IRTG 715 Language Technology and Cognitive Systems provided by the German Research Foundation (DFG).

References

Fazly, A., Alishahi, A., & Stevenson, S. (n.d.). A probabilistic computational model of cross-situational word learning. *Cognitive Science*. (To appear)

Fazly, A., Alishahi, A., & Stevenson, S. (2008). A probabilistic incremental model of word learning in the presence of referential uncertainty. In *Proc. of CogSci'08.*

Gelman, S., & Taylor, M. (1984). How two-year-old children interpret proper and common names for unfamiliar objects. *Child Development*, 1535–1540.

Gertner, Y., Fisher, C., & Eisengart, J. (2006). Learning words and rules: Abstract knowledge of word order in early sentence comprehension. *Psychological Science*, *17*(8), 684–691.

Gleitman, L. (1990). The structural sources of verb meanings. *Language Acquisition*, *1*(1), 3–55.

Gomez, R., & Gerken, L. (1999). Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition*, *70*(2), 109–135.

Harm, M. W. (2002). *Building large scale distributed semantic feature sets with WordNet* (Tech. Rep. No. PDP.CNS.02.1). Carnegie Mellon University.

Kemp, N., Lieven, E., & Tomasello, M. (2005). Young Children's knowledge of the "determiner" and " adjective" categories. *Journal of Speech, Language and Hearing Research*, *48*(3), 592–609.

MacWhinney, B. (1995). *The CHILDES project: Tools for analyzing talk* (second ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Maurits, L., Perfors, A. F., & Navarro, D. J. (2009). Joint acquisition of word order and word reference. In *Proc. of CogSci'09.*

Monaghan, P., & Mattock, K. (2009). Cross-situational language learning: The effects of grammatical categories as constraints on referential labeling. In *Proc. of CogSci'09.*

Naigles, L., & Hoff-Ginsberg, E. (1995). Input to verb learning: Evidence for the plausibility of syntactic bootstrapping. *Dev. Psychology*, *31*(5), 827–37.

Niyogi, S. (2002). Bayesian learning at the syntax-semantics interface. In *Proc. of CogSci'02.*

Parisien, C., Fazly, A., & Stevenson, S. (2008). An incremental Bayesian model for learning syntactic categories. In *Proc. of CoNLL'08.*

Pinker, S. (1989). *Learnability and cognition: The acquisition of argument structure.* Cambridge, MA: MIT Press.

Quine, W. (1960). *Word and object.* Cambridge, MA: MIT Press.

Siskind, J. M. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, *61*, 39–91.

Smith, L., & Yu, C. (2007). Infants rapidly learn words from noisy data via cross-situational statistics. In *Proc. of CogSci'07.*

Theakston, A. L., Lieven, E. V., Pine, J. M., & Rowland, C. F. (2001). The role of performance limitations in the acquisition of verb-argument structure: An alternative account. *J. of Child Language*, *28*, 127–152.

Yu, C. (2005). The emergence of links between lexical acquisition and object categorization: A computational study. *Connection Science*, *17*(3–4), 381–397.

Yu, C. (2006). Learning syntax–semantics mappings to bootstrap word learning. In *Proc. of CogSci'06.*

Sentence Processing Mechanisms Influence Cross-Situational Word Learning

Judith Köhne & Matthew W. Crocker
Department of Computational Linguistics
Saarland University
Saarbrücken, Germany
{judith, crocker}@coli.uni-saarland.de

Abstract

Word learning has traditionally examined separately the role of constraints provided by the visual context (e.g. *cross-situational learning*) and the linguistic context (e.g. *syntactic bootstrapping*). We suggest that the combined investigation of these learning scenarios is important: Firstly, to determine whether cross-situational word learning applies when words are presented in sentences and, secondly, to illuminate possible interactions of linguistic and situational learning mechanisms. We conducted three experiments to examine the role of visual and linguistic contextual constraints during foreign language word learning. In particular, our studies show that, given a visual context, syntactic verb-argument constraints together with knowledge about plausible real-world action-object relations help to further enhance cross-situational word learning.

Keywords: Cross-situational word learning in context; sentence processing; verb-derived expectations;

Introduction

Adults' foreign language learning often happens in a planned and incremental way to systematically gain increasing command of the language's structures. Parts of the vocabulary are learned very explicitly via vocabulary lists. When it comes to using and improving a foreign language in a natural situation, within an actual speech community, however, the language novice faces a less controllable, more diverse situation. When trying to understand and learn new words, there are two challenges: Firstly, words are often embedded in complex linguistic contexts and, secondly, there is a rich but ambiguous visual context containing possible world referents (*referential uncertainty*, Gleitman, 1990). One important mechanism for dealing with referential uncertainty is to keep track of words and referents co-occurring over different contexts. As previous research shows, adults as well as children are able to exploit such cross-situational learning analyses (*cross-situational word learning*, CSWL, e.g., Quine, 1960; Yu and Smith, 2007; Vouloumanos and Werker, 2009). In a study by Yu and Smith (2007), participants were asked to learn novel names for novel objects. Within a single trial, participants were exposed to 2-4 auditorily presented nouns, unconnected to each other, and the equal number of visually presented objects. Despite the referential uncertainty in each trial, participants were able to learn noun-object mappings by exploiting cross-trial co-occurrences.

In most CWSL studies, words are not presented as parts of sentences. This idealization has drawbacks, however: Firstly, language is not presented in its natural complexity (i.e., with sententially embedded words) and, secondly, possibly useful constraints provided by the linguistic context are intentionally withheld. This means that learning tasks are potentially oversimplified in some respects and overcomplicated in others. There is some evidence that adults are able to make use of the linguistic context that words come together with to understand the language input, instead of being distracted by it. Lee and Naigles (2008), for instance, show that the verb frame of a sentence helps verb learning (*syntactic bootstrapping*, Landau and Gleitman, 1985; Fisher, 2002; Lidz, Gleitman, and Gleitman, 2003). These kinds of studies, however, are usually not visually situated.

Some sentence processing mechanisms that are generally automatically applied by adult native speakers may also interact when dealing with foreign language input. As Altman and Kamide (1999) have shown, for instance, native speakers rapidly make inferences about linguistically upcoming referents in a sentence, given a restrictive verb (such as *eat*) and a visual scene. We investigate the hypothesis that learners may use similar on-line mechanisms when learning novel nouns. Specifically, we investigate whether such on-line predictions influence CSWL by reducing the size of sets of potential world referents a novel noun refers to.

In this paper, three adult language-learning eye-tracking experiments using a pseudo-natural language (modified Indonesian) are presented, addressing three central hypotheses: 1) CSWL mechanisms operate successfully when novel nouns are embedded in sentences. 2) Verb-driven, anticipatory expectations based on semantic verbal restrictions guide learners' (visual) attention. 3) Verb-driven, anticipatory expectations based on semantic verbal restrictions identify subsets of world-referents that novel nouns are likely to denote, thereby constraining CSWL.

Experiment 1

We investigated these issues with a stepwise learning procedure. Participants first learned restrictive verbs and were then exposed to novel nouns, embedded in spoken subject-verb-object (SVO) sentences as syntactic subjects (referring to characters) and syntactic objects (referring to objects), and depicted on scenes.

Methods

Participants 32 German native speakers took part in the experiment, 8 of which had to be excluded due to technical problems. Data of 24 participants was analyzed (17 female).

Design, Materials & Procedure The language consisted of six restrictive verbs (three food verbs like *bermamema*, 'eat', and three clothing verbs like *melimema*, 'iron'), twelve nouns

(six referring to human characters such as *badut*, 'clown', three to food objects such as *sonis*, 'sausage', and three to clothing items, e.g. *oblung*, 't-shirt') and one article that preceded all nouns (*si*, 'the'). The language was based on Indonesian (word order, article, parts of the words).

The experiment comprised three phases: isolated verb learning, situated noun learning, and vocabulary testing. In Phase 1, participants learned verbs by being exposed to static depictions of actions presented together with the corresponding spoken verb. Each action was named ten times. Then knowledge of verbs was tested: Participants were presented a picture not seen before and asked to pronounce the matching verb. Feedback was provided. The eye-tracker was adjusted and verbs together with depictions were presented again.

Phase 2 consisted of the sentence-comprehension and noun-learning phase: Semi-natural scenes and spoken sentences were presented (sentence start 1s after picture). Scenes depicted the target character and the target object (the named referents), as well as one distractor character and one distractor object, together with background. There was always one food item and one clothing item. Sentences were constructed using the already learned verbs and novel nouns. Word order was SVO. The syntactic subject denoted the target character and the syntactic object referred to the target object, either the food or the clothing one, corresponding to verb type (see example in Figure 1). People were not explicitly told the word order. There were 36 trials (randomized in order), each object and each character was named six times (and each one was shown twelve times). Participants were asked to understand the sentences and learn the unknown words. Eye-movements were measured.

In Phase 3, a forced-choice vocabulary test with 12 trials, one for each new noun, was performed. Pictures for the forced-choice vocabulary test showed 4 potential referents (= characters and objects) and were presented together with a spoken noun. Combinations of the four options differed but there was always at least one competitor of the same kind (character, food, or clothing item, respectively). Participants had to mouse-click onto the appropriate picture. Learning performance was the main measurement of interest for Phase 3. The experiment lasted about 30 minutes.

Predictions Hypothesizing that participants understand the SVO-sentence structure and have similar gaze behavior as native comprehenders, we expected more looks to characters than objects during NP1 and more looks to objects than characters during NP2. Secondly, we hypothesized that to identify character referents and learn their names, participants would exploit cross-situational analysis (Hypothesis 1). This predicts differences between looks to target and distractor characters to emerge over time during NP1: While in the very beginning participants have no hint which character NP1 referred to, tracking co-occurences of character names and depictions over trials makes it possible to identify the target. This increase in looks to the target should also become visible in the averaged data. We further hypothesized that verb

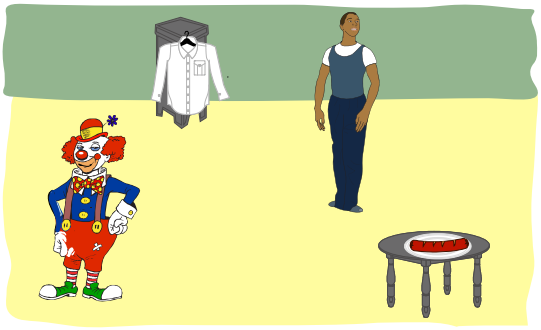


Figure 1: Example Item Experiment 1
Si badut bermamema si worel.
'The clown will eat the sausage.'

restrictions would be exploited quickly to identify object referents and learn their names, possibly additionally to CSWL (Hypotheses 2 & 3). That means that during the verb and NP2, targets should be inspected much more than distractors even early in Phase 2. Our hypothesis that verb restrictions provide additional cues regarding target objects (Hypothesis 3) moreover predicts that object names are learned better than character names overall.

Data Analysis, Results, & Discussion

The vocabulary test revealed a noun-learning rate that is well above chance (about 55% with a baseline of 25%, $t = 9.28, p < .001$). When analyzing only the data of the participants who actually learned all verbs in Phase 1, $N = 15$, it was 64% ($t = 8.59, p < .001$). Numerically, object names were learned better than character names but this difference did not reach significance (all: $t = .90, p = .38$; only good verb learners: $t = 1.38, p = .191$).

For eye-movement analysis we examined trials with at least one inspection on our regions of interest (ROI; target character, distractor character, target object, distractor object) for three time periods linked to the unfolding sentence (from onset of NP1 to onset of verb (V), from onset of V to onset of NP2, and from onset of NP2 to offset of NP2). All time periods were shifted such that they started 200ms later than the actual starting points in the speech stream because planning of saccades takes people about that much time. We conducted logistic regressions by entering the binomial data (fixation or no fixation at certain time to a specific ROI) into linear mixed effect models with logit link function (from the lme4 package in R, Bates, 2005). Participant and item were considered as random factors. To see whether the fixed factor (ROIs) had a main effect (i.e. whether including the factor significantly improved the predictive power of the model, regarding where people looked) we compared between the models that include and exclude this factor with a Chi-Square test (Baayen, Davidson, & Bates, 2008). Contrasts between levels of a factor (i.e. single ROIs) were investigated by studying the ratio of regression coefficients and standard errors since the p-values produced by lmers (Wald z test) are anti-conservative

(Baayen et al., 2008): If the coefficient is greater than the standard error times two, the comparison is considered to be reliable. Tables of these statistical comparisons are given below. The formulas describing the lmer models are of the following form: dependent variable (inspections during time periods) is a function of ($\sim$) the independent variable (ROI) plus random effects (subjects and items).

Table 1: Lmer models for inspections on characters vs. objects (m1) and targets vs. distractors (m2) during time periods (Experiment 1)
m1/m2: $InspectionsduringNP1/V/NP2 \sim 1 + ROI + (1|sub) + (1|item), family = binomial(link = "logit")$

		Predictor	Coef.	SE	Wald z	p
m1						
1	*NP1*	(Int) (char)	1.953	0.150	13.019	$< .001$
2		objects	-0.859	0.129	-6.647	$< .001$
3	*V*	(Int) (char)	0.224	0.116	1.928	$< .100$
4		objects	1.317	0.114	11.548	$< .001$
5	*NP2*	(Int) (char)	-0.253	0.146	-1.732	$< .100$
6		objects	0.492	0.101	4.872	$< .001$
m2						
7	*NP1*	(Int) (targ)	0.925	0.145	6.378	$< .001$
8		distractor	-0.112	0.111	-1.006	$= .315$
9	*V*	(Int) (targ)	0.593	0.122	4.859	$< .001$
10		distractor	-0.382	0.102	-3.734	$< .001$
11	*NP2*	(Int) (targ)	-0.557	0.107	-5.195	$< .100$
12		distractor	-0.257	0.105	-2.452	$< .050$

Eye-movements suggest that participants quickly understood the sentence structure: There were reliably more inspections on the characters than inspections on the objects during NP1 (Table 1, rows 1-2) and reliably more inspections on the objects than on the characters in the V interval (rows 3-4) and NP2 (rows 5-6). Moreover, the target character was inspected more often than the distractor character in NP1 (rows 7-8) and the target object was looked at more than the distractor object during NP2 (rows 11-12). This supports the hypothesis that participants succeeded in identifying the targets over the experiment. Furthermore, during V, the target object was also looked at reliably more than the distractor (rows 9-10). This likely reflects an anticipatory effect based on semantic verb restrictions. Also, the difference between looks to target and distractor objects during NP2 was greater than the difference between target and distractor characters during NP2, suggesting that verb restrictions contributed to an improved identification of objects during on-line processing (see timegraph in Figure 2).

Summarising, we found evidence that adults can learn nouns cross-situationally when words are embedded in sentences and referents are embedded in scenes, and further that they rapidly exploit semantic verb restrictions to identify referents on-line. We replicated these results with even better learning rates ($72\%, t = 8.249, p < .001$) and for another word order (OVS, with a learning rate of 51%, $t = 3.840, p < .001$) in a follow-up experiment. It is unclear, however, whether the on-line predicting of the referent has an effect on noun learning. We take this up in Experiment 2.

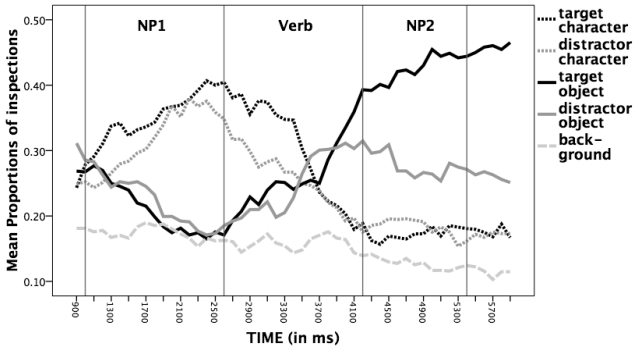


Figure 2: Timegraph Experiment 1

Experiment 2

In Experiment 2 we manipulated the degree of verb restriction to study the interaction of CSWL and verb-derived inference learning (Hypothesis 3). The focus therefore was on object learning rather than character learning.

Methods

Participants 50 German native speakers took part in the experiment (18 excluded due to bad verb learning and technical problems). Data of 32 participants was analyzed (23 female).

Materials & Procedure The language consisted of six verbs, 14 nouns, and the same article as in the other experiments. There were 2 non-restrictive verbs (e.g., *take*) and 4 restrictive verbs: either 2 food verbs (e.g., *eat*) and 2 clothing verbs (e.g., *iron*) or the 2 food verbs and 2 container verbs (e.g., *fill*), depending on list. The nouns denoted 2 characters (man and woman) and 12 objects: 4 food items (e.g., *broccoli*), 4 clothing items (e.g., *trousers*), and 4 container items (e.g., *vase*). Word order was SVO.

The experiment consisted of five parts with very similar procedures as in Experiment 1. The main difference was that instead of one sentence comprehension phase and one vocabulary testing part, there were two each (Blocks 1 and 2). The whole experimental sequence comprised: verb learning and testing, eye-tracker preparation, and verb repetition (Phase 1); sentence comprehension (noun learning) Block 1 (Phase 2); vocabulary test Block 1 (and verb repetition) (Phase 3); sentence comprehension (noun learning) Block 2 (Phase 4); vocabulary test Block 2 (Phase 5).

Phase 1 resembled Phase 1 in Experiment 1, except that we used animated verb-learning and verb-testing pictures to improve recognizability of the actions. In Phase 2 and 4, items were manipulated according to one three-level within-participant factor (Degree of referential uncertainty). There were three conditions: the no-referential-uncertainty condition (Condition 1), the low-referential-uncertainty condition

(Condition 2), and the high-referential-uncertainty condition (Condition 3). Firstly, the conditions differed with regard to verb type: In Condition 1 and 2, a restrictive verb was used, in Condition 3, it was a non-restrictive verb. Secondly, there were differences in the visual scenes. Images always depicted one character and four objects embedded in a simple indoor scene. One of the objects was the target object. The others were competitors (= potential world referent except the target) and distractors (= objects which are not potential world referents). The combination of competitors and distractors depended on the condition the item was in: In Condition 1, there was no competitor since the verb was restrictive (e.g. *eat*) and only the depicted target fulfilled verb constraints (e.g. *food*). In Condition 2, there was one competitor: The verb was restrictive but there was one depicted object in addition to the target which was a member of the required semantic class. In Condition 3, there were three competitors because the verb was non-restrictive and did not semantically constrain the category of potential referents denoted by the post-verbal argument (see Table 2). In Block 1 (Phase 2), no target was a competitor in another trial to make sure participants could not exclude competitors based on other learned words. In Block 2 (Phase 4), however, learning was potentially simplified via the possibility to exclude already learned mappings. The 48 trials (24 per Block) were presented randomized in order with each noun repeated four times. Participants were told that sentences were of the form 'someone VERBs something'. We monitored eye-movements in Phases 2 and 4.

Table 2: Conditions Experiment 2
(Number of potential referents on scene (Column 4) as a result of verb type (Column 2) and number of competitors (Column 3))

Condition	Verb	Competitors	= Pot. referents
1: No-ref. unc.	restr.	0	1
2: Low-ref. unc.	restr.	1	2
3: High-ref. unc.	non-r.	3	4

For the forced-choice vocabulary tests (Phases 3 and 5) there were six depictions presented on the screen: the target (*e.g. tomato*) and another instance of the targets category (e.g. *broccoli*), two objects of one of the two other categories (e.g., *shirt* and *skirt*), and two characters. Additionally to the mouse clicks, we introduced a confidence rating to have another, more sensitive measurement because there were only two nouns to be learned per condition: Participants were encouraged to press a number on the keyboard (between 1-9) to indicate how sure they were about their choice of a referent.

Predictions We hypothesized that selectional verb restrictions help identifying target referents and interact with CSWL (Hypothesis 3). In particular, we hypothesized, firstly, that during Phases 2 and 4, verb restrictions narrow down the search space in Condition 1 (from four to one since there was no competitor) and Condition 2 (from four to two since there was one competitor) (see Table 2); Secondly, that participants additionally conduct CSWL in Condition 2; And thirdly, that participants conduct only CSWL in Condition 3. Our predictions were, therefore, that noun learning rates and confidence ratings, as reflected in the vocabulary tests, would be highest for objects in Condition 1 and lowest for objects in Condition 3. With regard to eye-movements in Phases 2 and 4, our hypotheses predict differences for conditions during NP2: a clear preference for inspecting the target in Condition 1, a preference to inspect the target but a secondary preference to inspect the competitor in Condition 2, as well as a less strong preference for target inspection and an equally strong consideration of all competitors in Condition 3.

Finally, we hypothesized that in Block 2 participants can exclude those objects as potential referents, which have been already linked to a world-word-mapping in Block 1 (assuming the use of the principle of mutual exclusivity, Markman and Wachtel, 1988). This predicts an enhanced noun learning in Block 2 compared to Block 1.

Data Analysis, Results, & Discussion

Noun learning was reliably better than chance (25%) for all groups of interest and correlated positively with confidence ratings ($r = .452, p < .001$, see Table 3).

Learning was clearly better in Block 2 than Block 1 (all conditions: $\chi(1) = 30.77, p < .001$; Condition 1: $\chi(1) = 6.31, p < .05$; Condition 2: $\chi(1) = 10.17, p < .01$; Condition 3: $\chi(1) = 16.57, p < .001$). The same was true for confidence ratings (all conditions: $\chi(1) = 12.85, p < .001$; Condition 1: $\chi(1) = 5.42, p < .05$; Condition 2: $\chi(1) = 5.48, p < .05$; Condition 3: $\chi(1) = 10.69, p < .01$).

The direction of the differences between noun learning success and confidence ratings in the three conditions was as expected: Nouns were learned best and the decisions were rated highest in Condition 1 and worst in Condition 3. This was true for both blocks together as well as for Blocks 1 and 2 separately. We analyzed both values with linear mixed effect models, using logistic regression for the categorical learning rates (logit link function) and linear regression for the continuous confidence ratings, with participant and item as random factors. For confidence ratings we calculated Monte Carlo Marcov Chain values (MCMCs) whose p-values are a good estimate of the factor's significance (but are only applicable for continuous variables), (Baayen et al., 2008). Analyses did not reveal significant main effects for noun learning rates but did for confidence ratings. There were reliable differences in confidence ratings between single conditions: Condition 1 and Conditions 3 in all parts (Block 1, 2, and 1+2), Condition 1 and Condition in 2 in 1+2 and marginally in Block 2, and between Conditions 2 and 3 in all parts (Table 4: numbers in both blocks taken together).

Eye-movements were analyzed as in Experiment 1. Considering all conditions, the eye-gaze pattern for all parts of the

Table 3: Noun learning percentages (t-tests against chance 25%) / confidence ratings, Experiment 2

	Blocks 1+2	Block 1	Block 2
all	$72\%(t(62)=12.18, p<.001)/5.73$	$62\%(t(62)=6.90, p<.001)/5.06$	$83\%(t(62)=14.24, p<.001)/6.39$
Cond1	$77\%(t(62)=10.04, p<.001)/6.98$	$69\%(t(62)=6.24, p<.001)/6.34$	$85\%(t(62)=12.56, p<.001)/7.5$
Cond2	$74\%(t(62)=9.25, p<.001)/6.42$	$64\%(t(62)=5.19, p<.001)/5.88$	$85\%(t(62)=11.34, p<.001)/6.8$
Cond3	$66\%(t(62)=7.43, p<.001)/5.4$	$52\%(t(62)=3.49, p<.001)/4.45$	$80\%(t(62)=8.69, p<.001)/6.02$

Table 4: Lmer models & p-Values from MCMC sampling for confidence ratings, conditions 1-3 (Exp 2, both blocks)
m1: $condition \sim 1 + confidencerating + (1|sub) + (1|item)$

	Predictor	Coefficient	SE	t	MCMCmean	pMCMC	$Pr(>\|t\|)$
confidence ratings	(Intercept) (Condition1)	7.0575	0.3804	18.554	7.0303	$<.001$	0.0000
	Condition2	-0.6371	0.2869	-2.221	-0.58376	$<.100$	0.0272
	Condition3	-1.7530	0.30235	-5.799	-1.7027	$<.001$	0.0000

Table 5: Lmer models for inspections on target vs. distractors (m1) and distractor1/competitor vs. rest (m2) during NP2, conditions 1-3 (Exp 2, both blocks together)
m1/m2: $InspectionsduringNP2 \sim 1 + ROI + (1|sub) + (1|item), family = binomial(link = "logit")$

		Predictor	Coef.	SE	Wald z	p
m1						
1	*Cond1*	(Int) (tar)	-0.314	0.130	-2.406	$<.050$
2		char	-1.310	0.174	-7.540	$<.001$
3		di1	-0.685	0.171	-4.003	$<.001$
4		di2	-0.741	0.172	-4.306	$<.001$
5		di3	-0.506	0.174	-2.910	$<.010$
6	*Cond2*	(Int) (tar)	-0.124	0.121	-1.032	$=.300$
7		char	-1.426	0.174	-8.210	$<.001$
8		di1	-0.509	0.162	-3.134	$<.010$
9		di2	-1.325	0.186	-7.134	$<.001$
10		di3	-1.260	0.190	-6.642	$<.001$
11	*Cond3*	(Int) (tar)	-0.239	0.129	-1.859	$<.100$
12		char	-1.562	0.179	-8.709	$<.001$
13		di1	-0.172	0.156	-1.102	$=.270$
14		di2	-0.505	0.165	-3.058	$<.010$
15		di3	-0.983	0.178	-5.533	$<.001$
m2						
16	*Cond1*	(Int) (di1)	-0.998	0.141	-7.090	$<.001$
17		tar	-0.691	0.171	4.038	$<.001$
18		char	-0.629	0.181	-3.469	$<.001$
19		di2	-0.501	0.180	-0.317	$=.751$
20		di3	0.178	0.182	0.983	$=.326$
21	*Cond2*	(Int) (com)	-0.633	0.131	-4.840	$<.001$
22		tar	0.509	0.162	-8.208	$<.010$
23		char	-0.917	0.181	-5.070	$<.001$
24		di2	-0.816	0.192	-4.249	$<.001$
25		di3	-0.751	0.196	-3.826	$<.001$
26	*Cond3*	(Int) (di1)	-0.411	0.130	-3.171	$<.010$
27		tar	0.172	0.156	1.102	$=.270$
28		char	-1.389	0.180	-7.715	$<.001$
29		di2	-0.333	0.166	-2.001	$<.050$
30		di3	-0.811	0.179	-4.524	$<.001$

experiment resembles that of Experiment 1: We found referential inspections of the character during NP1, verb-driven anticipation of the target object(s) in verb region, and referential inspections of the target object in NP2.

More interestingly, differences between conditions for (referential) inspections in NP2 support the offline results: In Condition 1, the target was inspected reliably more than the character and the distractors (in all parts of the experiment) (Table 5, rows 1-5, for Blocks 1 and 2 together). In Condition 2, the target was inspected most, too (rows 6-10); However, the competitor was also inspected reliably more than the character and the distractors. The difference between looks to target and competitor was not significant in Block 1 but was in Block 2 and both blocks taken together (rows 21-25). For Condition 3, the target was inspected reliably more than the character or the distractors as well, except that the difference between looks to the target and to one distractor was significant only in Block 2, but neither in Block 1 nor in both blocks taken together (rows 11-15). This distractor shared category with the target. There were also significantly more looks to this distractor than to the other distractors and the character in Block 2 (but not for both blocks, see rows 26-30). The gaze pattern for Condition 3 is somewhat unexpected but interesting as it suggests that participants learned a new co-occurrence restriction for verbs in Block 2 (e.g., container objects and *take*) - although the verbs were non-restrictive, the distractor of the target category was preferred over the other distractors (which were of categories associated with other, restrictive, verbs).

The second experiment revealed clear effects of condition in on-line and off-line data showing that, firstly, referents are identified better when verbs provide information about the referent's category (better learning rates and confidence ratings in Condition 1 and 2 than in Condition 3) and, secondly, that cross-situational word learning interacts with the exploration of verb restrictions in that verb restrictions narrow

down the search space, lowering referential uncertainty (better learning rates and confidence ratings for Condition 2 than 3). As in Experiment 1, trials in Condition 1 made clear that verb restrictions can narrow down the number of potential referents to one, which means that there is a situation close to fast-mapping. Eye-movements during NP2 support the results except that there was an unexpected preference to look at both members of target category in Condition 3. We attribute this to spontaneous verb-argument category learning.

Summary & General Discussion

Two foreign-language learning experiments with an incremental learning scenario were conducted in order to study the influence of semantic verb restrictions on identifying world referents and learning world-word mappings. In Experiment 1, we found that nouns which are sententially embedded are successfully learned cross-situationally (with SVO and OVS sentences) and that participants additionally exploited verb restrictions rapidly to identify post-verbal referents. In Experiment 2, we additionally found evidence for the claim that verb restrictions interact with and improve Cross-Situational Word Learning.

With this investigation we have presented evidence for the claim that adult sentence processing mechanisms interact with statistical word learning and that foreign language word learners can benefit from exploiting linguistic and visual contextual cues. In particular, we revealed that semantic verb restrictions together with knowledge about plausible arguments reduces the set of potential (visual) referents, thus simplifying CSWL complexity. This highlights the cooperation of multiple learning mechanisms in situated word learning. Our findings are consistent with recent word learning models which combine co-occurrences frequency analysis with other, in particular situational and knowledge-based cues (Frank, Goodman, & Tenenbaum, 2009; Yu & Ballard, in press).

Acknowledgments

The research reported of in this paper was supported by IRTG 715 "Language Technology and Cognitive Systems" funded by the German Research Foundation (DFG).

References

Altman, G., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73*, 247–264.

Baayen, R., Davidson, D., & Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390–412.

Bates, D. (2005). Fitting linear mixed models in r. *R News*, *5*, 27–30.

Fisher, C. (2002). Structure limits on verb mapping: The role of abstract structure in 2.5-year- olds interpretations of novel verbs. *Developmental Science*, *5*, 55–64.

Frank, M., Goodman, N., & Tenenbaum, J. (2009). Using speaker's referential intentions to model early cross-situational word learning. *Psychological Science*, *20*, 578–585.

Gleitman, L. (1990). The structural sources of verb meanings. *Language Acquisition*, *1*, 3–55.

Landau, B., & Gleitman, L. (1985). *Language and experience: Evidence from the blind child*. Cambridge, MA: Harvard University Press.

Lee, J., & Naigles, L. (2008). Mandarin learners use syntactic bootstrapping in verb acquisition. *Cognition*, *106*, 1028–1037.

Lidz, J., Gleitman, H., & Gleitman, L. (2003). Understanding how input matters: Verb-learning and the footprint of universal grammar. *Cognition*, *87*, 151–178.

Markman, E., & Wachtel, G. (1988). Childrens use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology*, *20*, 121–157.

Quine, W. (1960). *Word and object*. Cambridge, MA.

Vouloumanos, A., & Werker, J. (2009). Infant's learning of novel words in a stochastic environment. *Developmental Psychology*, *45*, 1611–1617.

Yu, C., & Ballard, D. (in press). A unified model of early word learning: Integrating statistical and social cues. *Neurocomputing*.

Yu, C., & Smith, L. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, *18*, 414–420.

Adaptive Constraints and Inference in Cross-Situational Word Learning

George Kachergis, Chen Yu, and Richard M. Shiffrin
{gkacherg, chenyu, shiffrin}@indiana.edu
Department of Psychological & Brain Science / Cognitive Science Program
Bloomington, IN 47405 USA

Abstract

Previous research shows that human learners can acquire word-referent pairs over a short series of individually ambiguous situations each containing multiple words and referents (Yu & Smith, 2007). In this kind of cross-situational statistical learning based on the repeated co-occurrence of words with their intended referents, the application of principles such as mutual exclusivity and contrast can leverage prior experience to reduce the complexity in situations with multiple words and multiple referents. However, these principles can also block the learning of one-to-many mappings. In a study analogous those done in traditional associative learning, we manipulate the early and late evidence for particular pairings in the cross-situational learning paradigm, and examine the effects on learning of both one-to-one and many-to-many mappings. Two major findings are: 1) participants use mutual exclusivity and contrast to facilitate learning; and 2) given sufficient evidence, learners can adaptively disregard these principles and learn many-to-many mappings.

Keywords: statistical learning; language acquisition; cross-situational associative learning; blocking; highlighting

Introduction

Human infants and adults can acquire word-object pairings after experiencing a small number of individually ambiguous situations, each of which consists of several words and referents. The abilities required in cross-situational learning are to remember at least some of the co-occurrence statistics of nouns and their objects and to integrate statistical information across multiple learning situations, if one assumes that these words often occur when their referents are present, and that words and their referents will appear together in different situations. This idea, cross-situational learning, has been proposed as an essential means by which infants acquire language (Pinker, 1989; Gleitman, 1990). Recently, Smith & Yu (2008) empirically demonstrated that young infants can learn nouns through cross-situational learning.

In the more complex adult cross-situational word learning paradigm (Yu & Smith, 2007), participants were instructed to learn which word goes with each object and were then shown a sequence of training trials. Each trial consisted of a display of a few novel objects and a few successively spoken pseudowords. Each word referred to a particular on-screen object, but the correct referent for each word was not indicated, thus making meanings ambiguous on individual trials. In one learning condition with four words and four objects on each trial, participants attempted to learn 18 word-object pairings from 27 12-second trials. Thus, each stimulus—and each correct pairing—appeared six times. Learning was assessed by a 4AFC test of each pseudoword after training, and showed that participants on average acquired slightly more than 9 of the 18 pairs.

How might participants learn so many pairings from such a short series of trials, each of which contains 16 possible word-referent pairings? Some reasonable principles that learners may apply during training can significantly restrict the space of possible pairings. Among others, the mutual exclusivity (ME) constraint, which holds that learners will try to map words to referents in a 1-to-1 way (Markman, 1990), has been demonstrated in various word learning tasks. In the context of cross-situational learning, for example, suppose a learner hears words *A B C D* and sees referents *b a c d* on some trial. However, she realizes that she has heard *A* and *B* on some prior trial, and seen objects *a* and *b*, but the other stimuli are novel. Even if she does not know that *A*-*a* and *B*-*b* are the correct pairings (they may not even be unambiguous at this point), employing ME she may assume that {*A*, *B*} map to {*a*, *b*}, and that {*C*, *D*} map to {*c*, *d*}. Thus, by applying the ME constraint with some minimal previous experience the learner can whittle the 16 possible pairings down to four. This sort of mechanism was used to model cross-situational learning of pairs that appeared in consecutive trials, as well as the pairs that were not temporally contiguous, both of which were learned better than in conditions with no temporal contiguity (Kachergis, Yu, & Shiffrin, 2009b).

The role of prior knowledge has also been investigated (Klein, Yu, & Shiffrin, 2008), and it was found that participants can use pre-studied pairs to facilitate subsequent learning. Their experimental conditions were the same with that of (Yu & Smith, 2007) described above (27 trials, four pairs per trial) except that three pairs were unambiguously pre-trained, learners acquired a mean of 13.7/18 pairs (10.8 of the un-pretrained pairs). Presumably, pre-training helped participants learn more un-pretrained pairs (compared to 9.5 pairs in Yu & Smith) by reducing the number of possible pairings in trials containing any pre-trained pairs. Further evidence of bootstrapping made possible by the assumption of ME was found in a study that varied pair frequency and contextual diversity (which other pairs a given appears with during training). Kachergis, Yu, and Shiffrin (2008a) found that pairs appearing only thrice during training were learned significantly better when they were allowed to co-occur with pairs appearing nine times than when all pairs appeared solely with pairs of the same frequency.

All of these results indirectly imply that learners assume mutual exclusivity during training, and demonstrate the added power it can yield when pairings are 1-to-1. Yurovsky and Yu (2008) gave participants a cross-

situational task with some non-mutually exclusive pairings: halfway through training, half of the referents ceased appearing (e.g., A-a_1), and each was replaced by a second referent (e.g., A-a_2) which henceforth always co-occurred with the original referent's word (e.g., *A*). Thus, by the end of training half the words had both a primacy referent (a_1) and a recency referent (a_2). In separate 4AFC tests of primacy and recency referents, participants learned more than 50% of both the primacy and the recency referents. By the law of the excluded middle, some participants must have learned both pairings (e.g., *A*-a_1 and *A*-a_2), and thus built lexicons that violated ME. In another condition, the trials with primacy and recency referents were randomly interleaved, and learners still acquired nearly as many non-ME pairings as ME pairings (and above 50%, on average).

Ichinco, Frank, and Saxe (2009) studied ME using a different modification of the cross-situational word learning paradigm: Halfway through training, instead of replacing primacy referents with recency referents, they added an extra referent that always co-occurred with a particular old word-referent pair. That is, halfway through training, trials began to contain one more referent (i.e., 4) than words (3). Examining conjoint probability of learning ME-violating associations for each word, Ichinco *et al*. found that most learned pairings respected ME. Similar behavior was found in a condition with an added word instead of an added referent. Thus, although some participants learned some ME-violating pairings, the majority of learning behavior seems to follow a mutual exclusivity bias. This is unsurprising if one views cross-situational learning as a more complex form of traditional associative learning.

In a typical associative learning task, participants are given some subset of cues on each trial, asked to predict an outcome, and then shown the actual outcome. The subject's learning of associations between particular cues and outcomes is tracked over time. In cross-situational learning, the words can be construed as cues, and the objects as outcomes (or vice-versa). No trial-to-trial feedback is given, but the learner may generate it on the basis of the preceding training trials. *Blocking* is an associative learning effect often observed in experiments with two training stages: in the first stage, cue A is repeatedly paired with outcome X, and in the next stage A and B are jointly paired with X. The association between B and X is found to be weaker than when only the second stage training occurs: thus, B has been blocked by A's pretraining. Ichinco *et al*.'s design closely matches a blocking design (see Table 1), and their results—weak learning of the old word (or referent) to new referent (or word) association—are indeed a blocking effect.

Training Stage	Yurovsky & Yu, 2008	Ichinco, *et al.*, 2009	present study
Early	w_1-o_1	w_1-o_1	w_1-o_1
Late	w_1-o_2	w_1-{o_1, o_2}	{w_1, w_2}-{o_1, o_2}

Table 1: Comparison of three cross-situational ME designs. *N.b.*: these examples suppress other concurrent trained pairs.

The goal of the present paper is to systematically investigate how statistical learners accumulate and use current statistical evidence in subsequent learning. In the present study, for the first time we set up a strong test of inference, akin to associative learning's highlighting: will participants use knowledge of pairs acquired early in training, in addition to the principle of ME, to quickly learn pairs introduced late in training? If so, will this mechanism block the learning of many-to-many mappings? With two word-referent pairs sharing the same referent, one mapping appears in the early part of training and the other appears in the late training, will subjects prefer one over the other if we vary the amount of evidence (i.e., co-occurrence statistics) given during the early and late stages of training? Will they learn both pairs eventually? The set of experiments in the present paper allows us to answer those questions and examine how learners may adaptively incorporate evidence to potentially overwhelm biases.

Experiment 1

Participants are tasked with learning many word-referent pairs from a series of individually ambiguous training trials according to the cross-situational word learning paradigm (Yu & Smith, 2007). In the present study, each training trial is composed of two objects and two spoken pseudowords. On any given trial, participants can only guess which word refers to which object, since the order of presentation of the words is randomized, and there is no indication of which word refers to which object. However, since words only occur on trials with their intended referents, the correct pairings are disambiguated over the series of trials.

In the present cross-situational study, we divide each set of learning experiences into an early stage and a late stage, and systematically vary the number of times pairs appear in each stage. Half of the pairs appear in both the early training stage and the late stage, and the remaining half the pairs only appear in the late stage. As shown in Table 2, when a pair w1-o1 from the early stage appears in the late stage, another specific pair only appearing in the late stage (w7-o7) always co-occurs with w1-o1. Thus, in the late stage, {w1,o1,w7,o7} always co-occur, therefore, all of the four possible associations are reasonable (w1-o1, w1-o7, w7-o1, w7-o7). In fact, there is no additional information in the late stage that can be used to identify which ones are better than others. However, the key manipulation in this experiment is to vary the strength of w1-o1 in the early stage. More specifically, the early pairs co-occur 0 (no early training), 3, 6 and 9 times before they appear together with the late pairs. Given that we already know that people can effectively extract co-occurrence statistics in cross-situational learning, it is reasonable to assume that participants in this study would form some form of knowledge about w1-o1 when they enter the late stage trials. The research questions are: 1) whether they would prefer w7-o7 by applying the ME constraint; 2) whether they still learn w1-o7 and w7-o1 – a violation of ME; 3) how the amount of evidence about w1-

o1 may influence how they process the otherwise-ambiguous information in late trials with {w1,o1,w7,o7}.

Training Stage	Repetitions	Example Trials
Early [pairs 1-6]	0, 3, 6, or 9	$\{w_1, w_2, o_1, o_2\}$, …, $\{w_1, w_5, o_5, o_1\}$,
Late [pairs 1-12]	3 (Exp. 1) 6 (Exp. 2) 9 (Exp. 3)	$\{w_1, w_7, o_7, o_1\}$, …, $\{w_1, w_7, o_7, o_1\}$

Table 2: Experiment design, with example trials. Early pairs are indexed 1-6, and late-only pairs are 7-12. Pairs 1-6 also appear in the late stage, and thus occur more than pairs 7-12.

Subjects

33 undergraduates at Indiana University participated to receive course credit. None had participated in other cross-situational experiments.

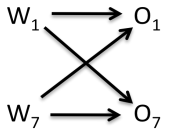
Stimuli

On each training trial, two unusual objects (e.g., sculptures) are simultaneously shown while two pseudowords were sequentially heard. The 48 computer-generated words are phonotactically-probable in English (e.g., "bosa"), and were spoken by a monotone, synthetic female voice. These 48 objects and 48 words were randomly assigned to four sets of 12 word-object pairings, one set for each learning condition. Within each set, 6 pairings only appear in the late training and the other 6 appear in both the early and late trainings. Each 8-second training trial began with the appearance of two objects, which remained visible for the entire trial. After 2 s of initial silence, each word was heard (randomly ordered, duration of 1 s) followed by 2 s of silence.

There were four learning conditions in this study. The late training was the same in those four conditions, and was composed of 18 trials. Each pair appeared 3 times late in training. Therefore, the same trial {w1,o1,w7,o7} also appeared 3 times. Four conditions varied in the early training. There was no early training in condition 1. In condition 2, each of 6 early pairs appeared 3 times, forming 9 early trials before the late training. In conditions 3 and 4, each early pair appeared 6 or 9 times. Accordingly, there were 18 and 27 early training trials in those two conditions.

Procedure

Learners were instructed that they would see a series of trials with two objects and two alien words, and that they should try to figure out what each word means for a test at the end. Participants were not told there were training stages, and there was no perceptible break. After training, their knowledge was assessed using 11-alternative forced choice (11AFC) testing: on each test trial a single word was played, and the participant was instructed to choose the appropriate object from a display of 11 of the 12 trained referents. Participants were instructed to click on the best referent for the word. Each word was tested twice in 11AFC trials: once without its corresponding early referent as one of 11 choices to test its association with the late referent ('early-late' and 'late-late'), and once without its corresponding late referent as one of 11 choices to test its association with the early referent ('early-early' and 'late-early'). In this way, we tested their knowledge of all of the four possible associations showing in Figure 1 (two associations for each word), and we access their knowledge of each of four possible associations individually in this test.



The four arrows show the associations tested (11AFC) for early pair 1 and late pair 7, which always co-occurred in the late training stage.

Figure 1: Example of associations tested by 11AFC.

Training condition order was counterbalanced, and each learner participated in all four conditions.

Results & Discussion

Fig. 2 displays the learning performance[1] in the training conditions with 3, 6, and 9 repetitions of early pairs. A 3x2x2 ANOVA with factors of early repetitions (3, 6, or 9), pair stage (early or late), and pairing type (within-stage or across-stage) showed only a significant main effect of pairing type ($F(1,30) = 8.62$, $p<.01$)[2]. As shown in Fig. 2, learning of within-stage (i.e., early-early and late-late word-object) pairs was much greater than learning of across-stage pairs (within-stage $M = .71$, across-stage $M = .15$). Even this small proportion of mean between-stage pair learning was significantly above chance, by subject (11AFC chance = .091; paired $t(30) = 2.29$, $p<.05$). Thus, although within-stage pairings—those consistent with ME—were clearly favored, participants also learned some ME-violating across-stage pairings. However, the indistinguishable, high level of performance on both early-early and late-late pairings is evidence of strong, ME-based inference: a given late pair could only be unambiguously learned by filtering out the consistently co-occurring early pair.

It is surprising that the number of early pair repetitions did not have a consistent effect on performance ($F(2,30) = .90$, $p>.05$). That is, even three repetitions of each early pair was enough prior experience to allow participants to infer the correct late pairs, thus achieving performance equal to the proportion of early pairs learned—no benefit was conferred by additional repetitions of early pairs (i.e., 6 or 9). In the condition with no early stage (0 early repetitions), participants learned 32% of the 2-to-2 pairings, on average—well above chance (paired $t(30) = 8.83$, $p<.001$).

Experiment 1 demonstrated that participants can efficiently leverage the ME constraint to learn late-appearing pairs that always co-occur with early pairs, and would thus be ambiguous, if not for prior experience. Indeed, performance on pairs learned using this filtering inference technique was no less than the performance on the

[1] Data from two subjects were excluded because their average performance in all four conditions was at chance (11AFC chance = .091). The outcomes of statistical tests were unaffected.

[2] We will report the results of the 0-early-training condition later as those results can be used to compare the data across experiments.

early pairs, which were learned by ambiguous cross-situational training, and which appeared more times, overall. Moreover, in tests of across-stage pairings, participants showed some learning of ME-violating associations.

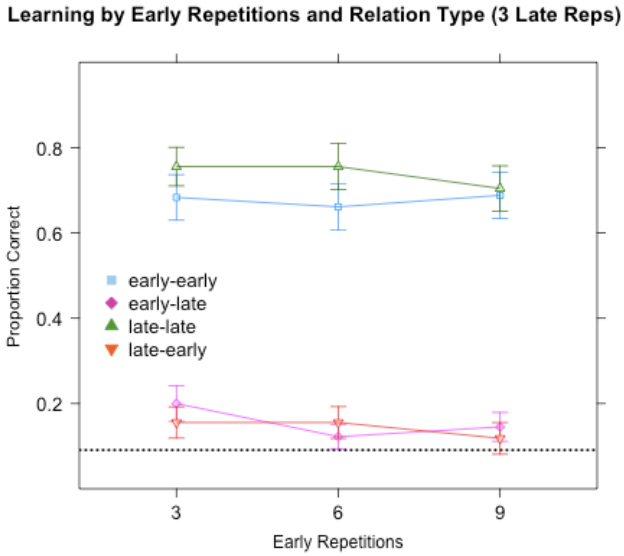


Figure 2: Mean accuracy by condition for the four types of associations (within and between early and late pairs). Error bars show +/-SE and the dotted line shows chance (.091).

In the next experiment, we increase the amount of late evidence, which we expect will cause participants to adaptively relax the ME constraint and learn more across-stage pairings (e.g. w7-o1). If they do relax ME, will they also learn fewer late-late pairings, or can they learn both?

Experiment 2

Experiment 1 showed that participants tend to utilize the principle of mutual exclusivity in combination with prior cross-situational training to quickly infer the referents of late-appearing words. In fact, performance on late and early pairs was undifferentiated, even when early pairings appeared more frequently (in both the early and late stages) than the late pairs (only the late stage). It appears that three more repetitions of early pairs grants enough knowledge to highlight the late-late pairs—which appear only three times—and bring performance up to the same level as the early-early pairs.

Thus, to some degree, Exp. 1 is analogous to blocking studies in associative learning literature, but with multiple co-occurring cues and outcomes on each trial. We show that just three repetitions of early pairs dramatically changed how they processed statistical information later and they apparently applied ME-based learning as evident by very few across-stage pairs learned later (i.e., early- late or late-early). In Experiment 2, we provide learners with more evidence for across-pair associations via additional repetitions of late stage pairs. Will participants adapt to this change, evaluate the statistical evidence in a different way and begin to count more on statistical information in the late part? Will they learn more ME-violating pairings?

Subjects

29 undergraduates at Indiana University received course credit for participating. None had participated in previous cross-situational experiments.

Stimuli & Procedure

The sets of pseudowords and referents used in Experiment 2 are identical to those used in Experiment 1. In each condition, the late stage of training was simply doubled from 18 trials to 36 trials wherein each {w1,o1,w7,o7} appears six times (instead of three times in Experiment 1), yielding three more repetitions of late and early pairs.

Results & Discussion

Figure 3 displays the average[3] levels of learning achieved in Exp. 2. Once again, a 3x2x2 ANOVA with factors of early repetitions (3, 6, or 9), pair stage (early or late), and pairing type (within-stage or across-stage) showed only a significant main effect of pairing type ($F(1,24) = 5.45$, $p<.05$). As in Exp. 1, learning of within-stage (i.e., early-early and late-late) pairs was much greater than learning of across-stage pairs (within-stage $M = .74$, across-stage $M = .28$). Thus, the increase of statistical evidence in the late stage with three more repetitions of both early and late pairs, didn't improve the learning of within-stage pairs (Welch $t(51.8) = .55$, $p>.05$). Nonetheless, learning of across-stage pairings increased—as predicted—due to increased late stage pairings (Welch $t(37.5) = 2.35$, $p<.05$). That is, having six rather than three repetitions of each late pair with its matched early pair in the late stage increased the learning of early word to late object (and vice-versa) pairings; the pairings that violate ME. Thus, although people are initially inclined to assume mutual exclusivity, and are able to use it to quickly infer the meaning of novel words, people will also adaptively relax ME in the face of greater evidence that words are being mapped to additional objects. As in Exp. 1, there was no significant effect of the number of early repetitions on learning ($F(2,24) = .06$, $p>.05$).

In summary, this experiment demonstrates that learners react to increased evidence that late and early pairs go together by learning more pairings. By doing so, they exhibit the ability to violate ME in order to learn 1-to-many mappings without reduced learning of ME-compliant pairings. In Experiment 3, we increase the late stage evidence once more to determine whether learners will continue to adaptively relax the ME constraint and learn more across-stage pairings.

[3] Data from four subjects were excluded because their average performance in all four conditions was at chance (11AFC chance = .091). The outcomes of statistical tests were unaffected.

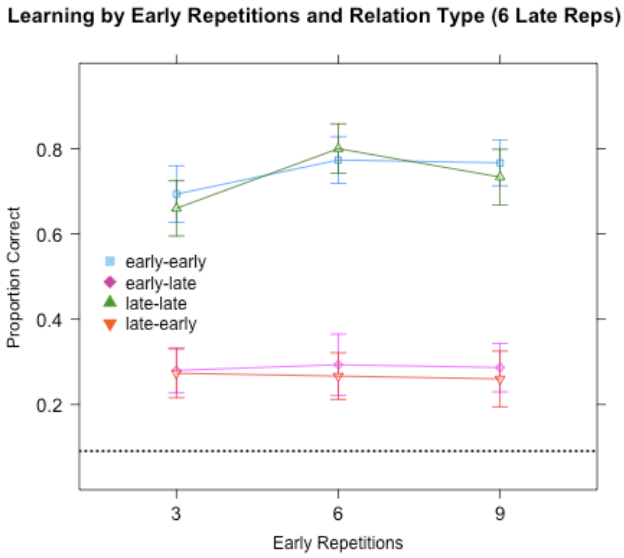


Figure 3: Accuracy by number of early pair repetitions and association type. Error bars show +/-SE.

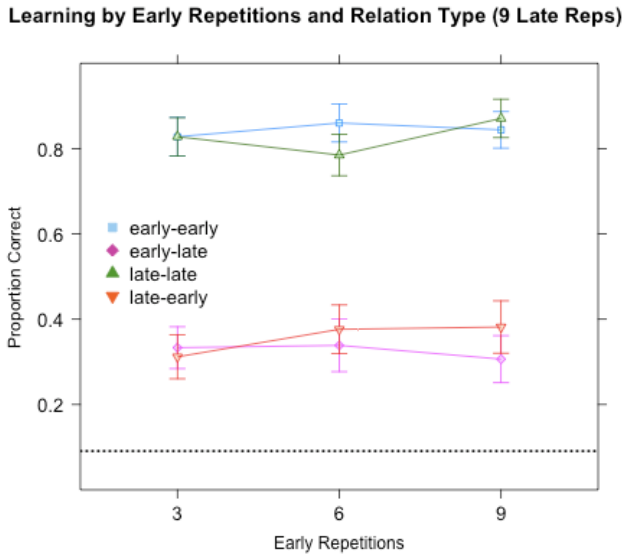


Figure 4: Accuracy by number of early pair repetitions and association type. Error bars show +/-SE.

Experiment 3

Experiment 2 provided six repetitions of late pairs with matched early pairs in comparison to Experiment 1's three repetitions, and learners indeed began to learn more late-early/early-late (i.e., across-stage, ME-violating) pairings. In Experiment 3, we once again increase the late stage by three repetitions (to nine), providing further evidence for cross-association of the matched early and late pairs.

Subjects

34 undergraduates at Indiana University received course credit for participating. None had participated in earlier cross-situational experiments.

Stimuli & Procedure

The same sets of pseudowords and referents were used in Experiment 3 as were used in Experiments 1 and 2. In each condition, there were 54 late-stage training trials, yielding three more repetitions of each late and early pair.

Results & Discussion

Figure 4 shows the mean[4] learning level by condition and pair type in Experiment 3. A mixed ANOVA (3, 6, or 9 late repetitions [between-subjects] x 3, 6, or 9 early repetitions x early or late pair stage x across- or within-stage pairing type) showed a main effect of pairing type ($F(1,101) = 161.76$, $p<.001$) and an interaction between pairing type and the number of late repetitions ($F(2,101) = 10.89$, $p<.01$). In brief, the patterns in Exp. 3 were consistent with what we observed in Exp. 2: participants learned within-stage pairs (e.g. w1-o1, w_7-o_7) quite well and also learned across-state pairs (e.g. w7-o1, w1-o7) with the additional evidence provided in late training.

[4] Data from three subjects were excluded because their average performance in all four conditions was at chance. The outcomes of statistical tests were unaffected.

General Discussion

When attempting to learn word meanings from a series of individually ambiguous trials, applying the mutual exclusivity constraint on each trial can significantly reduce the number of pairings a learner must consider, and even allow learner to quickly infer the meanings of novel words and referents. In the cross-situational experiments presented here, participants used such filtering—akin to highlighting in the associative learning literature—to quickly learn the same proportion of late-late pairings as early-early pairings. Thus, learners can use ME to infer the late-late pairings, even though each late-late pairing always co-occurred with one early pairing.

However, if pairings are not in fact mutually exclusive, or if word-referent mappings may change over time, assuming ME would be maladaptive. Across experiments, we increased the number of repetitions of late pairs, each of which always appeared with a particular early pair (e.g., w_1 and o_1 always appeared with w_7 and o_7, which had never appeared before the late stage). As an early stage pair appears more often with a particular late pair, a flexible learner would relax the assumption of ME.

Exp. 1, with three late pair repetitions, demonstrated that participants learn early-early pairs very well with only three early repetitions, and use this knowledge, in combination with ME, to learn the late-late pairs at a similar rate. However, even in this experiment, participants showed evidence of learning some ME-violating pairings (i.e., late-early and early-late pairings). Results from Experiment 2, with six late pair repetitions, looked very similar, but with slightly higher learning of ME-violating pairings. With nine late pair repetitions, Experiment 3 provided further evidence that participants should disregard ME, and indeed they learned more cross-stage pairings.

Viewing cross-situational statistical word learning as a complex form of associative learning in which there are multiple cues and outcomes on each training trial, it is unsurprising to find evidence for ME in cross-situational experiments, for ME may be responsible for well-known associative learning phenomena such as highlighting and blocking. However, our results demonstrate that participants do not merely use the ME constraint for logical inference. Instead, they utilize an adaptable learning strategy: as statistical evidence for violation of ME increases, they learn more ME-violating pairs. A comparison of the results from the 0-early-stage condition in three experiments in which each early pair (e.g., w_1-o_1) always co-occurred with a particular late pair (e.g., w_7-o_7). Figure 5 shows the mean proportion of learned ME-respecting pairings (i.e., only pairings involving each stimulus once) and ME-violating pairings (i.e., pairings that involve a stimulus twice) that participants learned by the number of late repetitions (i.e., experiment), which increases as more late repetitions give evidence that the ME constraint should be relaxed.

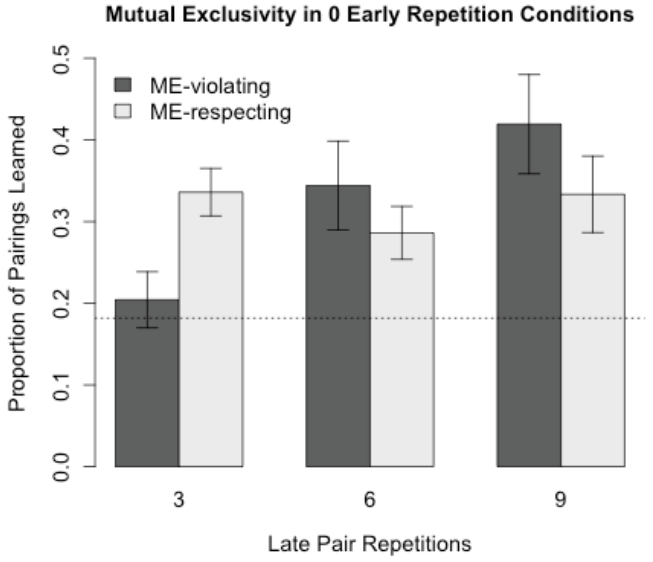


Figure 5: The mean proportion of pairings that each participant learned that violate ME and that comply with ME (across all experiments). Chance for ME-respecting pairings is the dotted line (2/11). Chance for ME-violating pairings is $(1/11)^2 = \sim.01$.

In summary, the results from the present study provide a complete picture of how participants use the ME constraint and how they accumulate statistical evidence over the course of learning. By varying the strengths of word-object associations in both the early and late training stages, we were able to demonstrate various learning behaviors that have been shown in previous studies. In fact, some of those studies have produced incompatible results. Therefore, the contribution of our work is to unify different views inferred from previous results. We argue that human statistical learning is adaptable: learners are able to adjust their learning strategy over the course of learning in response to changing amounts of evidence for particular types of pairings. One possible reason that previous results are not compatible is because each study took a snapshot of a continuous learning process. For example, many ME studies are conducted with one or two simple trials. We observe that learners' cross-situational learning strategies are dynamic and adaptable, and thus cannot be adequately portrayed by one snapshot. Rather, we need to examine learning trajectories by varying the amounts (repetitions) and types (ME-violating or –respecting) of evidence. We believe that the data presented here will quite useful in constraining models of cross-situational language acquisition. Future studies use real-time behavioral data (e.g., eye movements) to provide access online learning strategies.

Acknowledgments

This research was supported by National Institute of Health Grant R01HD056029 and National Science Foundation Grant BCS 0544995. Special thanks to Tarun Gangwani for data collection.

References

Frank, M. C., Goodman, N. D., Tenenbaum, J. B. (2008). A Bayesian framework for cross-situational word-learning. *NIPS 20* (pp. 457-464). Cambridge, MA: MIT Press.

Gleitman, L. (1990). The structural sources of verb meanings. *Language Acquisition*, *1*(1), 1-55.

Ichinco, D., Frank, M. C., & Saxe, R. (2009). Cross-situational word learning respects mutual exclusivity. In N. Taatgen, H. van Rijn, J. Nerbonne, & L. Schomaker (Eds.) *Proceedings of CogSci 31*.

Kachergis, G., Yu, C., & Shiffrin, R. M. (2009a). Frequency and contextual diversity effects in cross-situational word learning. In N. Taatgen, H. van Rijn, J. Nerbonne, & L. Schomaker (Eds.) *Proceedings of CogSci 31*.

Kachergis, G., Yu, C., & Shiffrin, R. M. (2009b). Temporal contiguity in cross-situational statistical learning. In N. Taatgen, H. van Rijn, J. Nerbonne, & L. Schomaker (Eds.) *Proceedings of CogSci 31*. Austin, TX: Cognitive Science Society.

Klein, K. A., Yu, C., & Shiffrin, R. M. (2008). Prior knowledge bootstraps cross-situational learning. *Proceedings CogSci 30*, Austin, TX: Cognitive Science Society.

Markman, E. M. (1990). Constraints Children Place on Word Learning. *Cognitive Science*, 14, 57-77.

Pinker, S. (1989). *Learnability and cognition*. Cambridge, MA: The MIT Press.

Smith, L. & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition, 106,* 1558-1568.

Yu, C. & Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psych. Sci., 18,* 414-420.

Yurovsky, D. & Yu, C. (2008). Mutual exclusivity in cross-situational statistical learning. *Proceedings of CogSci 30*. Austin, TX: Cognitive Science Society.

Desirable Difficulties in Cross-Situational Word Learning

Haley A. Vlach (haleyvlach@ucla.edu)
Catherine M. Sandhofer (sandhof@psych.ucla.edu)
Department of Psychology, University of California, Los Angeles
1285 Franz Hall, Los Angeles, CA 90095 USA

Abstract

The world offers learners a seemingly infinite number of word-to-world mappings (Quine, 1960). In order to account for how learners manage to accomplish such a difficult task, theories of word learning have proposed different tools that make the task of learning words easier. However, we propose that reducing difficulty may be detrimental—difficulty may promote long-term word learning. We tested this hypothesis in a cross-situational paradigm in which object-label mappings were ambiguous during each learning event. The three conditions of learning (2 x 2, 3 x 3, and 4 x 4) varied in the degree of difficulty. Results revealed that, although difficulty deterred immediate performance, difficulty promoted long-term performance. We suggest that theory and research should shift from focusing on in-the-moment learning to examining both immediate *and* long-term learning. A complete theory of word learning not only accounts for word learning in the moment and on each time scale, but also integrates them in order to understand how they influence each other over time.

Keywords: word and category learning; statistical learning; cross-situational learning; long-term memory

Introduction

The world offers learners a seemingly infinite number of word-to-world mappings (Quine, 1960). Thus, an essential question for research on word learning is: How do learners manage to accomplish the difficult task of mapping words to objects, actions, and events in the world?

Theories of word learning typically focus on tools that learners use to make word learning easier. In this study, we examine word learning from the radical perspective that reducing difficulty may be detrimental. This study explores the idea that some difficulty may promote word learning, even in difficult tasks in which learners must track mappings across events, such as cross-situational word learning.

Theories of Word Learning. Three main classes of theories have sought to explain word learning: the Constraints/Principles theories, the Social-Pragmatic theories, and the Domain-General theories. All three of these theories propose tools that make word learning easier but differ in the nature of the task simplification tools.

The Constraints/Principles theories suggest that word learning is made easier and more feasible by constraints that narrow the search space for possible word-to-world mappings, such as mutual-exclusivity (e.g., Markman, 1989) and the novel-name nameless-category assumption (e.g., Golinkoff, Mervis, & Hirsh-Pasek, 1994). These constraints guide learners' interpretations of new words and thus reduce the degree of indeterminacy. For example; the mutual-exclusivity principle (Markman, 1989) proposes words have mutually-exclusive meaning—one object can have only one referent. Consequently, when learners hear an unfamiliar label, they will assign an unfamiliar label to an unfamiliar object rather than an object that has already been named.

A second class of theories, the Social-Pragmatic theories, propose that word learning is simplified because learners are embedded in a social world in which they are guided by expert word learners (e.g., Bloom, 1993; Tomasello & Barton, 1994). Adults, as expert word learners, resolve the ambiguity of the word-learning scenario by guiding children's attention and thus make the task of word learning easier. For example, adults commonly talk about objects, events, and actions that learners are already focused on and consequently make it easier for learners to make word-to-world mappings (Bloom, 1993).

A final class of theories, the Domain-General theories, assert that general cognitive mechanisms such as perceptual saliency, association, and frequency make word learning straightforward (e.g., Smith, 1995)—learners notice objects and actions that are most salient in their environment and pair them with the most frequently associated label. For example, in one study (Samuelson & Smith, 1998), children were able to learn a novel word-novel object link by using saliency cues in the absence of other cues, suggesting that saliency cues alone guided children's word learning.

Word Learning and Memory. Although the three classes of theories make different predictions about many aspects of word learning, in this study we investigated a cognitive mechanism that is inevitably a critical part of each theory: memory. For example; the Constraint/Principles theories argue that constraints promote memory for words—if everything had a multiple unique labels it would be impossible to store and recall all of these words from memory. Social-Pragmatic theories rely on processes of memory and attention for establishing joint attention among two people—learners must attend and remember what others are focusing on in order to adequately label words and actions (e.g., Bloom, 1993). Domain-General theories assert that word learning is guided by global principles of attention, association, and frequency, which are basic cognitive mechanisms associated with memory (e.g., Smith, 1995). In sum, memory is a critical component to word learning theories because it supports every part of the word learning process—learning words requires attending to

words, encoding the properties of the word, binding words to objects in the world, and recalling words when needed in order to communicate with others.

Although it is clear that memory matters for learning words, relatively little work has investigated the role of memory and retention in word learning. In fact, only a handful of studies have imposed a delay between learning and testing (see Horst & Samuelson, 2008, for a discussion of this issue). Consequently, the vast majority of word learning theories are based upon immediate performance rather than performance over time.

Examining word learning both immediately and over time is essential for two reasons. First, a complete theory of word learning accounts for developmental changes in word learning and retention abilities. Moreover, such a theory not only accounts for word learning and retention on each time scale, but also integrates them in order to understand how they influence each other over time.

Second, immediate performance may not be a reflection of long-term performance. The few studies that have examined word learning and retention have yielded mixed results as to whether performance remains constant over time (e.g., Horst & Samuelson, 2008; Markson & Bloom, 1997). Alternatively, the memory literature has provided countless examples of how the factors that promote immediate learning do not necessarily promote long-term learning (e.g., Bjork, 1994). Immediate performance may not predict long-term performance because the degree of difficulty in learning influences long-term performance.

Desirable Difficulties in Learning. There has been a long history of research that has investigated the conditions under which long-term memory is enhanced. The principle goals of this research have been to discover factors that promote adults' ability to (1) produce and store a representation of knowledge and (2) create a representation that remains accessible and recallable over extended periods of time. Research has revealed that several manipulations of learning events can enhance long-term memory, such as varying the conditions of practice (e.g., Smith & Rothkopf, 1984), providing contextual interference (e.g., Mannes & Kintsch, 1987), distributing practice and the spacing effect (e.g., Cepeda et al., 2006), and reducing feedback to the learner over time (e.g., Schmidt, 1991).

These manipulations promote long-term memory because they introduce difficulty for learners while knowledge is being acquired (e.g., Bjork, 1994). Although learning tasks that are designed to make learning easy initially show greater learning, retention tests reveal that more difficult learning tasks promote more long-term memory and learning (and hence the term 'desirable' difficulty is commonly used). Thus, the memory literature suggests that, instead of making tasks easy for learners, the best way to promote long-term memory is to create difficulty for learners during learning.

An example of a desirable difficulty in learning is the spacing effect (e.g., Cepeda et al., 2006). The spacing effect describes the robust phenomenon whereby memory is enhanced when learning events are distributed across time (i.e., spaced), instead of being presented in immediate succession (i.e., massed). Spaced learning is more difficult than massed learning because the time between learning events creates greater opportunities for forgetting (e.g., Bjork & Allen, 1970). Massed presentations prevent forgetting because presentations are in immediate succession. In fact, upon immediate testing, massed presentations lead to a greater amount of learning than spaced presentations. However, if a test is administered following a delay, a spaced presentation schedule will yield more learning than the massed presentation schedule.

Several researchers have long suggested that, although introducing difficulty during memory tasks is beneficial, these difficulties may be detrimental in more difficult cognitive tasks (e.g., Gagne, 1950). For example, spaced learning was thought to be particularly detrimental in generalization tasks. In fact, spaced learning was coined the "enemy of induction" (e.g., Gagne, 1950; see Kornell & Bjork, 2008, for a discussion).

Desirable Difficulties in Word Learning. Despite speculations that desirable difficulties may be the "enemy of induction", recent research suggests that imposing difficulty during learning promotes long-term word learning and generalization (e.g., Kornell & Bjork, 2008; Vlach et al., 2008). For example, one study (Vlach et al., 2008) presented children with novel objects and labels in an object category learning paradigm. Category exemplars were presented on two schedules, massed and spaced. Children were tested after a three minute delay and were required to generalize a word to a novel instance of a category. The results revealed that spaced presentations promoted more learning than massed presentations. Thus, a spaced learning schedule, a more difficult learning schedule, promoted word learning and generalization.

One limitation of research on desirable difficulties in word learning is that all of the studies have been artificially simplistic—a linguistic label could only be mapped onto one object. In real word learning contexts, mapping words to objects is generally not this straightforward. Word learners must figure out what words map onto in the world (Quine, 1960). Thus, because learners must track possible mappings across learning events, real world word learning is much more difficult than tested in recent research on desirable difficulties in word learning.

Research on cross-situational learning has indicated that the more objects and labels in each learning event, the more difficult it is for learners to determine mappings (e.g., Smith & Yu, 2008; Yu & Smith, 2007; Yurovsky & Yu, 2008). For example, when adult learners are presented with two words and two objects in learning events, they demonstrate relatively high performance, ~90% correct mappings on an immediate test. However, when learners are presented with four objects and four labels in each word learning event, learners perform significantly lower, ~55% correct

mappings on an immediate test (Yu & Smith, 2007). Thus, it appears that the more objects and labels in each learning event, the more difficult it is to track mappings across learning situations.

The memory literature would suggest that increasing the number of objects and labels in each word learning event presents several forms of desirable difficulty. First, increasing the number of object and labels in each learning event creates more spaced learning because each object-label pairing is interleaved between more possible pairings (e.g., Cepeda et al, 2006; Vlach et al., 2008). Second, an increase in the number of objects and labels in each learning event creates more contextual variation and interference between word learning events (e.g., Mannes & Kintsch, 1987). Both of these factors have been shown to promote long-term retention (e.g., Bjork, 1994).

Although recent research suggests that difficulty may promote word learning, this hypothesis has only been tested in artificially simple tasks where object-label mappings are straightforward. It may be the case that adding more difficulty to an already difficult task of mapping words to objects is not beneficial. Consequently, too much difficulty may deter both in-the-moment and long-term word learning. The current study investigates this possibility.

Current Study. The current study investigated the role of difficulty during word learning in a cross-situational word learning paradigm. Participants were presented with word learning events in which determining the object-label mappings were increasingly difficult. In the 2 x 2 condition, each trial presented two words and two objects. In the 3 x 3 condition, each trial presented three words and three objects. Finally, in the 4 x 4 condition, each trial presented four words and four objects. There were also three testing delay conditions: immediate, 30 minute delay, and one week delay. These conditions allowed for a direct comparison of the effects of varying degrees of difficulty in both in-the-moment and long-term word learning.

Method

Participants

Participants were 95 undergraduates at University of California, Los Angeles. Participants received course credit for their participation.

Design

This study used a 3 x 3 design. Learning Condition (2 x 2, 3 x 3, and 4 x 4) and Testing Delay (immediate, 30 minutes, and one week) were both between-subjects factors. Participants were randomly assigned to one of the nine conditions of the study.

Stimuli

Pictures of objects were presented on a 15-inch computer screen and the sound for the labels was presented by the computer's speakers. As Figure 1 shows, the objects were pictures of novel objects. There were a total of 18 objects. The labels were novel words following the phonotactic probabilities of English (e.g., 'blicket', 'dax'). There were a total of 18 labels. Objects and labels were randomly paired together, for a total of 18 object-label pairs. In all conditions, there were a total of 6 presentations of each of the 18 object-label pairs. There were also an additional four objects and four labels presented during the training trial. These objects and labels were not used during the learning phase of the experiment.

In the 2 x 2 condition, two objects and two words were presented in each learning trial (see Figure 1). In the 3 x 3 condition, three objects and three labels were presented. In the 4 x 4 condition, four objects and four labels were presented.

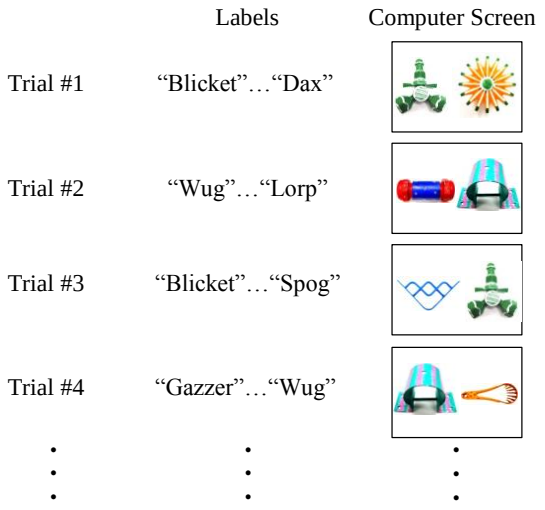


Figure 1. Example stimuli from the 2 x 2 condition.

Because the same number of object-label pairs (18 pairs) were presented in each condition, the same number of times (6 presentations each), other presentation factors varied across conditions in order to ensure equivalent exposure to the object-label pairs. Table 1 outlines these variations, which were adapted from Yu and Smith (2007). Although the number of trials and time per trial varied, the total exposure time remained constant across the conditions (see Table 1).

Table 1. Three Learning Conditions

Condition	Number of Trials	Time per Trial (in secs)	Total Time (in secs)
2 x 2	54	6	324
3 x 3	36	9	324
4 x 4	27	12	324

Procedure

Participants were told that they would be shown children's toys and it was their job to figure out which word went with which toy. They were also instructed that it would be

ambiguous as to which words went with which objects on each trial. Participants were then given a brief training exercise to demonstrate what the experiment would be like. The training consisted of three learning trials, each with two objects and two labels, immediately followed by a forced-choice test. Objects and labels used during training were not included during the rest of the experiment.

After the training trial, participants were informed that they would now be beginning the learning phase of the experiment. Participants were presented with learning trials according to the condition in which they were assigned (2 x 2, 3 x 3, or 4 x 4). The number and length of trials was also set according to the condition (see Table 1).

After viewing all of the trials, participants were given a forced-choice test, depending upon the testing condition in which they were assigned. In the immediate condition, participants were given a test immediately following learning. In the 30 minute delay condition, participants were asked to play tetris for 30 minutes, and then were given a test. In the one week delay condition, participants were asked to come back exactly 7 days after the learning session and complete a test.

The test consisted of four force-choice questions. Each question presented one label and asked participants to identify the corresponding object among four objects. The three foil objects were other objects used in the experiment. No one object was repeated in the tests. Thus, 16 of the 18 objects were used in the test. The labels and objects used during the test were randomly assigned.

Results

We asked whether difficulty would promote learners' long-term word learning in a cross-situational learning paradigm. If difficulty promoted word learning, we would expect to see lower performance immediately, but stronger performance long-term. However, if difficulty did not promote word learning, we would expect to see lower performance regardless of testing delay.

We first conducted a 3 (Learning Condition) x 3 (Testing Delay) ANOVA, with the number of correct responses as the dependent measure. Results of this test revealed a significant main effect of learning condition, $F(2, 86) = 20.582$, $p < .001$, a significant main effect of testing delay, $F(2, 86) = 17.294$, $p < .001$, and a significant interaction of learning and testing delay, $F(4, 86) = 2.542$, $p = .045$.

First, three univariate ANOVAs were conducted within each testing condition. We then computed three planned comparisons using t-tests with Bonferroni corrections ($p < .05$) to determine the nature of the differences between learning conditions within each testing delay condition. If difficulty promoted word learning, we expected there to be differences in performance between learning conditions across the testing conditions.

When the immediate testing condition, there was a main effect of learning condition, $F(2, 32) = 10.997$, $p < .001$. Participants in the 2 x 2 condition ($M = 3.85$ correct mappings out of 4, $SD = .376$) had significantly higher performance than in the 4 x 4 condition ($M = 2.00$ correct mappings out of 4, $SD = 1.195$), $p < .001$. Performance was also marginally higher in the 2 x 2 condition than the 3 x 3 condition ($M = 3.07$ correct mappings out of 4, $SD = .997$),

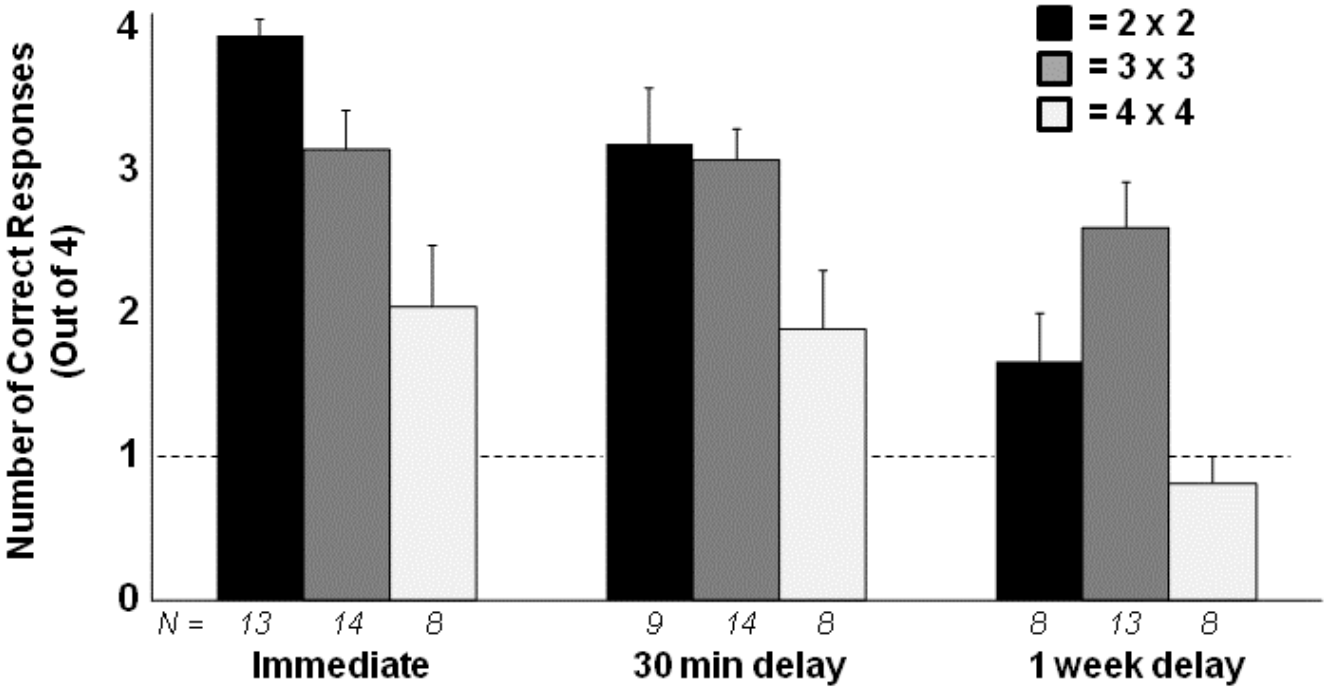


Figure 2. Average number of correct responses (out of 4) by learning condition (2 x 2, 3 x 3, 4 x 4) and testing condition (immediate, 30 minute delay, one week delay). The dashed line represents chance performance.

p = .086. Finally, performance in the 3 x 3 condition was significantly higher than the 4 x 4 condition, p =.029. Thus, greater the number of object-label pairings in each learning trial, the lower the performance.

However, there was a different pattern of results in the 30 minute delay condition. There was a main effect of learning condition, $F(2, 28) = 5.304$, $p = .011$. Participants in the 2 x 2 condition (M = 3.11 correct mappings out of 4, SD = 1.167) had similar performance to participants in the 3 x 3 condition (M = 3.00 correct mappings out of 4, SD =.784), $p > .05$. Participants in the 2 x 2 and 3 x 3 conditions both had significantly higher performance than participants in the 4 x 4 condition (M = 1.75 correct mappings out of 4, SD = 1.035), $p = .022$ and $p = .021$.

In the one week testing delay condition there was a particularly interesting pattern of results. There was a main effect of learning condition, $F(2, 26) = 11.286$, $p < .001$. Participants in the 3 x 3 condition (M = 2.54 correct mappings out of 4, SD = 1.127) had higher performance than both the 2 x 2 condition (M = 1.62 correct mappings out of 4, SD = .518), $p = .071$, and 4 x 4 condition (M = .75 correct mappings out of 4, SD = .463), $p < .001$. Participants in the 4 x 4 condition performed similarly to participants in the 2 x 2 condition, $p > .05$. Thus, although initially participants in the 3 x 3 condition had lower performance than the 2 x 2 condition, one week later participants in the 3 x 3 condition had higher performance than participants in the 2 x 2 condition.

In addition to examining the differences within each testing condition, we also examined differences in each learning condition across the testing conditions using ANOVAs and three planned comparisons with Bonferroni corrections ($p < .05$). In the 2 x 2 condition, there was a main effect of testing delay, $F(2, 27) = 12.255$, $p < .001$. Immediate performance was marginally higher than performance in the 30 minute delay condition, $p = .085$, and the performance in the 30 minute delay condition was significantly higher than performance in the 1 week condition, $p = .001$. Thus, there was significant decrease in retention across each of the testing delay conditions.

There was also a main effect of testing delay in the 4 x 4 condition, $F(2, 21) = 3.868$, $p = .037$. There was not a significant difference in performance between immediate and 30 minute delay conditions, $p > .05$, or the 30 minute delay and one week delay conditions, $p > .05$. However, there was a significant difference in performance between the immediate and one week delay condition, $p = .047$. Thus, there was a significant decrease in retention between the immediate test and one week delayed test. Finally, in the 3 x 3 condition, there was not a main effect of testing delay, $F(2, 38) = 1.172$, $p > .05$. Thus, there was not a significant decrease in retention between the immediate and delayed tests.

Discussion

The results of this study support the idea that difficulty imposed during learning can promote long-term word learning (e.g., Vlach et al., 2008). Moreover, difficulty promoted word learning in the already difficult task of cross-situational word learning, in which learners must track mappings across events. We found that, when tested immediately, learners had the highest performance in the 2 x 2 condition and the lowest performance in the 4 x 4 condition. Performance in the 3 x 3 condition was somewhere in between. These findings replicate that of Yu & Smith (2007). However, when tested 30 minutes later, there were no differences in the performance between the 2 x 2 and 3 x 3 conditions. Finally, when tested a week later, performance in the 3 x 3 condition was higher than performance than in both the 2 x 2 and 4 x 4 conditions. Thus, although difficulty yielded lower immediate performance (i.e., the 2 x 2 condition had higher performance than the 3 x 3 condition), there was a benefit of difficulty for long-term performance (i.e., one week later the 3 x 3 condition had higher performance than the 2 x 2 condition). This study demonstrates that, even in the seemingly difficult task of mapping words to objects (Quine, 1960), adding difficulty promoted long-term word learning.

The findings from this study also have implications for research on cross-situational word learning and, more generally, statistical word learning. Recent research on statistical word learning has focused on the factors that promote immediate performance in order to discover the mechanisms by which words are acquired over time (e.g., Kachergis, Yu, & Shiffrin, 2009; Lany & Saffran, 2010). However, this study suggests that this may not be the best approach for describing long-term trajectories of word learning. This study clearly demonstrates that immediate performance does not always reflect long-term performance. Thus, in order to assert that a mechanism promotes word learning over time, evidence should be provided from not just an immediate test, but an immediate *and* delayed test.

The broader impact of this study is that it highlights the intimate relationship between word learning and memory. Learning new words and categories requires perceiving an object, attending to relevant features, mapping a label to the object, binding this mapping to other instances of the label and object, abstracting across instances, and, finally, generalizing to novel objects. Memory is a critical factor in this process, both during category formation (e.g., remembering relevant features and binding instances and labels) and recall (e.g., retrieving stored instances and categories).

Despite the clear relationship between word learning, memory, and retention, we have failed as word learning researchers and developmentalists to explore the mechanisms underlying this relationship. Fundamental questions have remained unexamined. For example, the few studies that have asked whether children retain words over time have provided conflicting evidence. While one study finds children retain words for a month (e.g., Markson & Bloom, 1997), other studies have found that children forget words in a matter of minutes (e.g., Horst & Samuelson,

2008; Vlach et al., 2008). What are the implications of our research if participants do not remember words after a few minutes? Why are we speculating about long-term word learning from immediate performance, rather than empirically investigating word learning over time? Are we really uncovering the mechanisms of word learning?

In sum, future research should investigate both in-the-moment and long-term word learning. Exploring in-the-moment word learning is essential for understanding how words and categories are initially encoded. However, in theories of word learning, the common assumption is that performance will remain constant over time. This study clearly demonstrates that this is not always the case.

In order to account for real-world word learning, research should incorporate testing over longer time-scales—over the course of weeks, months, and years. A complete theory of word learning not only accounts for word learning in the moment and on each time scale, but also integrates them in order to understand how they influence each other over time.

Acknowledgments

We thank the undergraduate research assistants of the Language and Cognitive Development Lab at UCLA for their assistance in collecting the data for this study.

References

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalf and A. P. Shimura (Eds.), *Metacognition: Knowing about knowing* (pp. 185-205). Cambridge, MA: MIT Press.

Bjork, R. A., & Allen, T. W. (1970). The spacing effect: Consolidation or differential encoding? *Journal of Verbal Learning and Verbal Behavior, 9*, 567-572.

Bloom, L. (1993). *The transition from infancy to language: Acquiring the power of expression*. Cambridge, England: Cambridge University press.

Bloom, P. (2000). *How children learn the meaning of words*. Cambridge, MA. MIT Press.

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin, 132*, 354-380.

Gagne, R. M. (1950). The effect of sequence of presentation of similar items on the learning of paired-associates. *Journal of Experimental Psychology, 40*, 61-73.

Golinkoff, R. M., Mervis, C. B., & Hirsh-Pasek, K. (1994). Early object labels: The case for a developmental lexical principles framework. *Journal of Child Language, 21*, 125-155.

Horst, J. S., & Samuelson, L. K. (2008). Fast mapping but poor retention by 24-month-old infants. *Infancy, 13*, 128-157.

Kachergis, G., Yu, C., & Shiffrin, R. M. (2009). Frequency and contextual diversity effects in cross-situational word learning. In N. Taatgen, H. van Rijn, J. Nerbonne, & L. Schomaker (Eds.) *Proceedings of the 31st Annual Conference of the Cognitive Science Society* . Austin, TX: Cognitive Science Society.

Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories: Is spacing the "enemy of induction"? *Psychological Science, 19*, 585-592.

Lany, J., & Saffran, J. R. (2010). From statistics to meaning. *Psychological Science*, published online January 2010.

Mannes, S. M., & Kintsch, W. (1987). Knowledge organization and text organization. *Cognition and Instruction, 4*, 91-115.

Markman, E. M. (1989). *Categorization and naming in children: Problems of induction*. Cambridge, MA: MIT Press.

Markson, L., & Bloom, P. (1997). Evidence against a dedicated system for word learning in children. *Nature, 285*, 813-815.

Quine, W.V.O. (1960). *Word and object*. Cambridge, MA: MIT Press.

Samuelson, L. R., & Smith, L. B. (1998). Memory and attention make smart word learning: An alternative account of Akhtar, Carpenter, and Tomasello. *Child Development, 69*, 94-104.

Schmidt, R. A. (1991). Frequent augmented feedback can degrade learning: Evidence and interpretations. In G. E. Stelmach & J. Requin (Eds.), Tutorials in motor neuroscience (pp. 59-75). Dordrecht: Kluwer.

Smith, L. B. (1995). Self-organizing processes in learning to learn words: Development is not induction. In C. A. Nelson (Ed.), *Basic and applied perspectives on learning, cognition, and development: The Minnesota Symposia on Child Psychology* (Vol. 28, pp. 1-32). Mahwah, NJ: Erlbaum.

Smith, L. B., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, 106, 333-338.

Smith, S. M., & Rothkopf, E. Z. (1984). Contextual enrichment and distribution of practice in the classroom. *Cognition and Instruction, 1*, 341-358.

Tomasello, M., & Barton, M. (1994). Learning words in non-ostensive contexts. *Developmental Psychology, 30*, 639-650.

Vlach, H. A., Sandhofer, C. M., & Kornell, N. (2008). The spacing effect in children's memory and category induction. *Cognition, 109*, 163-167.

Yu, C., & Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science, 18*, 414-420.

Yurovsky, D. & Yu, C. (2008). Mutual Exclusivity in Cross-Situational Statistical Learning. In B. C. Love, K. McRae, & V.M. Sloutsky (Eds.). *Proceedings of the 30th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

The Goldilocks Effect: Infants' preference for stimuli that are neither too predictable nor too surprising

Celeste Kidd (ckidd@bcs.rochester.edu)
Brain & Cognitive Sciences, Meliora Hall
Rochester, NY 14627 USA

Steven T. Piantadosi (piantado@mit.edu)
Brain & Cognitive Sciences, 43 Vassar Street
Cambridge, MA 02139 USA

Richard N. Aslin (aslin@cvs.rochester.edu)
Brain & Cognitive Sciences, Meliora Hall
Rochester, NY 14627 USA

Abstract

Even before birth, infants attend to the statistical properties of their sensory environments to learn about events in world. Tracking these statistics is crucial to mastery of visual, social, linguistic, and cognitive tasks. However, the degree to which their sampling follows prescriptions of rational statistical inference is unclear. Do infants' attentional preferences reflect efficient information gathering? We investigated using an ideal observer model (a Markov Dirichlet-multinomial). We predicted infants' attention to sequential events would be moderated by information content. We tested infants (7-8 months) with 32 unique event sequences (objects popping out of boxes) on a Tobii eye-tracker. Each sequence continued until look-away. Controlling for other variables, we found infants were significantly more likely to look away at either highly informative or uninformative events according to the model. This suggests infants allocate visual attention to maintain intermediate rates of information processing, avoiding committing cognitive resources to either overly predictable or surprising events. This "Goldilocks effect" may reflect a general strategy for efficient learning from environmental statistics.

Keywords: Statistical learning; statistical inference; idealized learner; infant gaze behavior; infant methods; infant eye-tracking; Bayesian modeling; information theory; infant visual attention.

Introduction

Infants have a lot to learn in the first few years of life, and a limited set of resources with which to do it. The world is brimming with potential sources of information, but where among this spatiotemporal array of events should infants begin their learning? From birth, infants survey their sensory environments, sampling the visual data that surrounds them at the incredibly rapid pace of two or more fixations a second during 90% of their waking hours (Haith, 1980). This process, of surveying and sampling, provides infants with rich information from which they can start to learn about the world.

Previous empirical work has demonstrated that infants are able use the statistical properties of their environment in a diverse array of learning tasks pertaining to sounds, words, people, shapes, and objects (Fiser & Aslin, 2002; Maye, Weiss, & Aslin, 2008; Saffran, Aslin, & Newport, 1996; Saffran, Johnson, Aslin, & Newport, 1999; Yu & Ballard, 2007)[1]. In many complex cognitive systems (e.g., object recognition, language), infants obtain the representations of higher-level structures by tracking the low-level statistical coocurrences. For example, newborn infants must track the distribution of acoustical properties of speech sounds in their target language in order to infer its phonological categories (White, Peperkamp, Kirk, & Morgan, 2008). Researchers of visual, conceptual, and social learning find similar patterns. Recent technologies (e.g. eye-tracking, brain imaging techniques) have elucidated many of the mechanisms infants employ in building high-level structures from low-level environmental statistics. Though the topic has been of great interest to researchers, the mechanisms and representations infants employ during the process of collecting environmental statistics are still not well understood.

Amid the unbinned and unsorted masses of sensory data available in the world, an undirected search would be inefficient. Infants have too many things to do—motor actions to program, words to learn, categories to form—to waste time. How then should the infant allocate her visual attention?

Several researchers attempted to unify this work by identifying an overarching stimulus feature that could generally account for all of infants' preferences for various stimulus properties. Sokolov (1960) postulated that the primary driver of infants' attention is stimulus novelty. Consistent with this theory, infants commonly prefer the novel stimuli in preferential looking/listening tasks such as those use in the Fantz paradigm (Fantz, 1964), high-amplitude sucking procedure (Siqueland & DeLucia, 1969), and head-turn preference procedure (Kemler Nelson et al., 1995) . The novelty account is also consistent with habituation behavior, during which infants' attention to recurring stimuli decreases over time. However, the novelty hypothesis does not account for infants' *familiarity* preferences in many preferential looking and listening studies. Notable examples include infants' affinity for their native languages and for faces, especially those of their mothers.

Roder and others attempted to reconcile infants' preference patterns by relating preference and processing load (Hunter & Ames, 1988; Roder, Bushnell, & Sasseville, 2000; Wagner & Sakovits, 1986). Roder suggested that the process of memory formation was responsible for preference. Under this theory, infants would be expected to exhibit a familiarity preference early in processing as they form a memory of

[1]The literature detailing these statistical learning abilities is so large, in fact, that if it were printed and stacked in a pile, it would be more than 140 infants tall (based on 161,000 unique articles cataloged by Google Scholar at time of publication and 26-inch mean height of 8-month-olds).

the stimulus, and a novelty preference later after memory formation was complete. While this account correctly predict age and experience-related shifts in visual preference, it does not on its own account for all types of visual preferences. It does not, for example, make clear predictions of why infants would prefer one novel object to another entirely novel object, since infants would not possess a memory for either item. Kinney and Kagan similarly suggested a processing-based account of preference. Their *moderate discrepency hypothesis* states that infants will preferentially attend to stimuli that are "optimally discrepant", meaning those that are most distinct from the representations they already possess. Like Roder's memory-based account, Kinney and Kagan's theory relates stimulus preferences to stimulus representations established by past experiences (Kinney & Kagan, 1976). The moderate discrepancy hypothesis has the added advantage of accounting for preferences among completely novel stimuli, since it defines the representation formation process as pertaining to the infants' existing representations. Unfortunately, attempts to test this theory behaviorally were hindered by methodological difficulties. First, researchers had no direct access for determining the type, quantity, and nature of infants existing representations, which are crucial to the theory for generating testable predictions. Second, manipulating the identity of stimulus items to test for visual preferences forced researchers to rely on *qualitative* judgments of discrepancy rather than a quantitative metric.

Yet another account, Dember and Earl's *theory of choice/preference*, suggests that *stimulus complexity* drives looking behavior. Dember and Earl posited that every stimulus contains a certain "complexity value, and that each individual[2] has a certain preferred complexity level it seeks to maintain (Dember & Earl, 1957). In this context, complexity can be thought of as information content. The theory predicts that individuals will seek out stimuli containing the ideal level of complexity with respect to their own preferred complexity rates. The amount of information an individual will derive from a stimulus decreases as experience accumulates. Thus, like other processing-based accounts, this complexity-driven one can theoretically predict age and experience-related shifts in visual preference. Berlyne noted that a complexity-driven preference would be an optimal strategy for learning (Berlyne, 1960). It provides a rational solution to the infant learner's problem of deciding where best to allocate attention in the world. As with the attempts to test memory-based theories of attention, the collection of empirical evidence for this theory was hindered by the use of stimuli varied along *qualitative* dimensions rather than quantitative metric.

All prior models of infant visual attention, using the standard 2-second look-away criterion, have been based on hypothetical underlying processes such as information complexity, processing speed, and stimulus salience. Unfortunately, none of these underlying processes were validated by an independent assessment. As a result, the precise way in which these processes were combined could not be estimated, except by observing the outcome of their integrated effect on gaze durations. Here we seek to provide a quantitative model of visual attention to sequential events by systematically manipulating information complexity while holding processing speed and stimulus salience constant.

We used an idealized statistical model—a Dirichlet-Multinomial Markov model—to predict infant looking behavior to a display of sequential events. Our results suggest that infants' behavioral responses to a stimulus are influenced by its information content. Further, we find evidence that infants allocate their attention to maintain a certain information rate under a statistical model of the world. We present this as evidence that infants use rational statistical inference in understanding the world and deciding where to allocate attention and other cognitive resources.

Infant Behavioral Data

Participants

Twenty-five infants (mean = 7.9 months, range = 7.0 - 8.8) were tested. All infants were born full-term and had no known health conditions, hearing loss, or visual deficits, according to parental report. All participating infants completed the study.

Stimuli

We presented each infant with 32 unique event sequences, with the order of the sequences randomized across infants. The events each sequence consisted of were three unique objects that were animated to pop out from behind three occluding surfaces, which simulated an array of boxes. The sequences of object "pop ups" were chosen to vary in their information-theoretic properties (e.g., entropy, surprisal). Thus, some sequences were highly predictable (e.g., *AAAAAAAA*), and others were less predictable (e.g., *CAAABBCABAC*).

For each infant, the Matlab script generated an animated scene based on each of the 32 event sequences. Each event sequence was implemented by creating a scene consisting of three uniquely patterned and colored boxes, each concealing a unique familiar object (e.g., a cookie). The locations of the three boxes for a given sequence were chosen randomly but remained static throughout a scene. The box locations were randomly shuffled between event sequences, but no more than two boxes appeared on either half of the screen. Neither the patterns on the boxes nor the objects were repeated across event sequences so that each object-box pair was independent and unique.

The objects, boxes, and the order in which the 32 event sequences were presented were randomized across infants. The same 32 event sequences were presented to every infant. This design ensured that differences in looking time across event sequences were not driven by differences in scene items or

[2]Individuals referred to not only baby humans, but also adults and animals

presentation order. Each event in a sequence consisted of an object that popped out of a box (1 s), and then back into the box (1 s). The total duration of each event was 2 s, and events were presented sequentially with no overlap or delay.

Procedure

Each infant was seated on his or her parent's lap in front of a table-mounted Tobii 1750 eye-tracker. The infant was positioned such that his or her eyes were approximately 23 inches from the monitor, the recommended distance for accurate eye-tracking. At this viewing distance, the 17-inch LCD screen subtended 24 X 32 degrees of visual angle. Each of the 3 boxes was 2 X 2 inches. To prevent parental influence on the infant's behavior, the parent holding the infant was asked to wear headphones playing music, lower their eyes, and abstain from interacting with their infant throughout the experiment.

The experiment consisted of 32 trials, one for each event sequence. Each trial was preceded by an animation designed to attract the infant's attention to the center of the screen—a laughing and cooing baby. Once the infant looked at the attention-getter, an experimenter who was observing remotely pushed a button to start the trial.

For each trial, an animated scene depicting one of the event sequences was played. The animated sequence of events—objects popping out of boxes one at a time—continued until the infant looked away continuously for 1 sec, or until the sequence timed out at 60 sec. The 1-sec look-away criterion for trial termination was automatically determined by the Tobii eye-tracking software. If the infant looked continuously for the entire 60-sec sequence, the trial was automatically labeled as a "time out" and discarded before the analysis (3.5% of trials). If the trial was terminated before the infant actually looked away, the trial was labeled by an experimenter as a "false stop" and also discarded. False stops occurred as a result of the Tobii software being unable to detect the child's eyes continuously for 1 sec, usually due to the infant inadvertently blocking the his or her own eyes with head or arm movements (18.5% of trials).

Every infant saw all 32 event-sequence trials. The dependent measure for the subsequent computational modeling was the event at which the infant looked away in each trial; that is, at what point in the sequence did infants look away from the display for more than 1 consecutive second? We predicted that infants were more likely to look away during events that contained either too little or too much information for a particular infants' preferred information-intake rate. We predicted infants would be least likely to look away during events that were "just right"—those that were neither too predictable, nor too surprising. Our Ideal Observer Model was used to determine the amount of information for each event in the event sequences (i.e., which event contained more or less information). If infants' attention to a stimulus is governed by the amount of information it contains, we would expect infants' look-aways to be predictable given the model. We tested our hypothesis by comparing the model's predicted probabilities of an infant looking away for each event in the sequence to the infants' actual look-aways in test.

Ideal Observer Model

We used a Markov Dirichlet-multinomial model (MDM) to evaluate the relationship between the statistical properties of the event sequences and infants' attention to events in that sequence. The model allows us to test the best-fitting set of parameters for predicting from the event sequence whether the infant will continue looking or terminate a trial by looking away from the display. The MDM is a general-purpose statistical model that infers an underlying (multinomial) probability distribution on events, using the history of how many times each event has been observed. The MDM makes parametric assumptions about the form of the prior probability of an event and the likelihood of the event, and is often used in Bayesian statistics because it is computationally simple. Intuitively, infants observe how many times each event in the world occurs, and then use these event counts to infer an underlying probability distribution on events, just as readers extract an underlying word frequency distribution using a set of observations of individual words. An observer who sees only a single event happen would not likely infer that that single event is the only possible event (e.g, has probability 1.0). Instead, observers likely bring expectations to the task. In the version of a MDM used here, this prior expectation is parameterized by a single free parameter, α which controls the prior degree of belief that the distribution of events is uniform (e.g., that all unobserved events are equally likely). As α gets larger, the model has stronger prior beliefs that the distribution of events in the world is uniform; as $\alpha \rightarrow 0$, the model believes more strongly that the distribution is closer to empirically observed counts on events.

Formally, if there are three events, A, B, and C, which have been observed to occur c_A, c_B, and c_C times respectively, then the model assigns probability to a distribution on these three events proportional to

$$P(A)^{c_A+alpha-1}P(B)^{c_B+alpha-1}P(C)^{c_C+alpha-1}, \qquad (1)$$

where P is a hypothesized distribution on the events A, B, and C. That is, after observing each event occur some number of times, the infant may form a representation P, which gives the true underlying distribution of events. Every distribution can be "scored" according to Equation 1, allowing one to compute a distribution of beliefs about the state of the world according to the model. This simple model allows us to quantify an ideal observer's degree of belief that any given distribution on events is the true one. Importantly, because of the parametric form of the MDM, statistical measures such as the most likely true distribution of events, can be computed analytically.

We used two different forms of the MDM. In the first, the events A, B, and C correspond to events in the behavioral experiment (objects appearing from behind the occluding boxes). This model does not represent the transitions between events in the world; that is, the sequence *AAABBBCCC*

would have the same expectation as *ABCABCABC*. In the second model, we treated the events *A*, *B*, and *C* as transitions (or bigrams): for each object, we created a separate MDM for events that happen next. This model represents three separate MDMs that capture the transitions between events.

Both of these forms of the MDM provide an estimate of what an ideal observer would infer about the structure of the world. However, a model of infant's beliefs alone is not sufficient to predict their behavior: what is needed additionally is a set of linking assumptions that relate beliefs to actions. Here, we assume that the infant's looking behavior is at least partially determined by the information-theoretic properties of the model. Specifically, we test whether the predictability of a stimulus according to an idealized learning model influences infants' looking behavior. Formally, we use the negative log probability of the current event according to the model, conditioned on observing all the previous events. As this negative log probability increases, the current event is more surprising: for instance, after seeing a long sequence of *A*s, a *B* would have a high negative log probability. Negative log probability is a convenient measure because it corresponds to the number of *bits* of information conveyed by the stimulus. Thus, negative log probability provides a measurement at each point in time of the unpredictability of an event, using a measure that is typically used as a measure of information content. Because of the form of the MDM, the model roughly predicts that events in the future will tend to occur with their already-observed probability. However, the model essentially adds a small amount of smoothing—parameterized by alpha—that prevents unseen events from having probability zero.

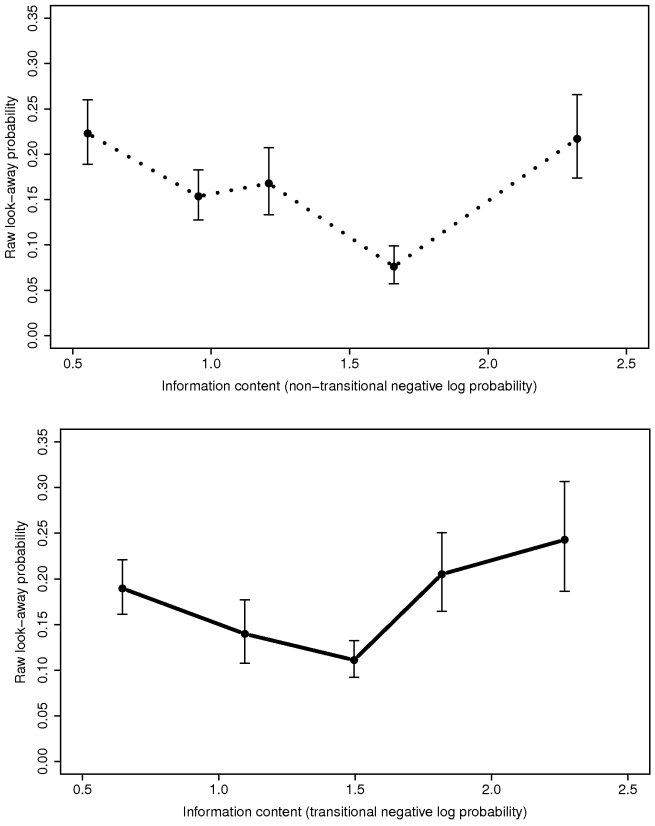


Figure 1: Infant look-away probabilities as a function of non-transitional surprisal (top) and transitional surprisal (bottom).

Results & Analysis

At each event in a sequence, infants make an implicit decision to either look away or keep looking at the scene. Figure 1 shows their raw probability of looking away at each item, as a function of that item's negative log probability according to the model, and collapsing across infants and sequences. The blue line shows the results for the non-transitional model, and the red line shows results for the transitional model. Both show a U-shaped relationship between raw look-away probability and model-based estimate of surprisal, with infants looking away to events that are especially surprising or especially predictable. There is a "Goldilocks" value of surprisal around 1.5, corresponding to infants' preferred rate of information in this task[3] which corresponds roughly to the point in the graph where infants have the lowest raw probability of looking away.

Survival analysis

Although the MDMs in Figure 1 provide a revealing picture of the relationship between indexes of surprisal and looking durations, there are likely other factors that influence infant look-aways. For instance, it might be the case that events in sequences generally become higher probability as infants form a picture of the statistical properties of the stimulus. If infants generally looked for a fixed amount of time, rather than paying attention to the statistical properties of the stimulus, then generally increasing predictability could make it look as though they preferred a certain information rate. To address this, we performed a regression analysis to control for the influence of other factors on look-away probability.

When infants look away, their trial ends and they provide no more additional data for that sequence [4]. This means that there is only a data point for an infant at time t if they have not looked away before t. We used a type of regression that respects this statistical relationship between look-aways and future data called a *survival analysis*. The type of survival analysis we used, Cox regression, measures the log linear influence that predictors have on the probability of a look-away at each point in time, but controls for a baseline look-away distribution. In the variety of survivial analysis we used, the baseline looking distribution is fit nonparametrically to the data, meaning that the analysis conservatively removes the

[3]This information rate must be interpreted relative to the frequency with which events in the sequence are presented, one every 2 seconds

[4]In the statistical literature, this type of data is called *censored*.

influence of an "average" distribution of looking times, before testing the significance of predictors.

We used a stepwise procedure for the Cox regression that tested whether each of several variables improved the model fit (AIC). Thus, at each iteration, the regression only added variables if they contributed positively, and at the same time removed variables if they contributed negatively. We included the following predictors in the survival analysis as control covariates.

- TRIAL-NUMBER: The number of sequences the child has already observed
- FIRST-APPEARANCE: A boolean factor corresponding to whether this event is the first time an object has been observed.
- UNSEEN-ITEMS: The number of objects that have not yet been observed.
- SAME-EVENT : A boolean factor for whether or not the current event is the same as the one that just happened.

The primary predictors we included in the survival analysis is the negative log probability of the event according to the MDM. Table 1 revealed that this variable is likely related to look-away probability quadratically, so we also included the squared negative log probability of the event according to the model[5]. A significant effect of squared predictability tests the significance of the U-shaped effect observed in Figure 1. As discussed above, we formed both transitional and non-transitional versions of the model, corresponding to models that treat each event independently, or each transition independently. Because the predictions of these two models are highly-correlated, we performed separate analyses on each.

Figure 2 shows the results of the survival analysis, including all predictors that were added via the stepwise procedure. These results can be interpreted by multiplying each coefficient by the value of the covariate and then exponentiating. This number represents an amount by which the probability of looking away is scaled, according to the best-fitting model. For instance, the coeffient of TRIAL-NUMBER is 0.033, meaning that by the 10'th sequence the child sees, they have a $exp(10 * 0.033) = 1.39$ greater factor of looking away. This effect of TRIAL-NUMBER is a plausible effect of fatigue. The results also show a significant effect of SAME-EVENT: children are a factor of $exp(0.316) = 1.37$ more likely to look away when the event is a repeat of the most recent event. This effect is also plausible: infants search for other things to keep their interest when the experiment shows a repeating–and therefore boring–event.

The regression results also reveal significant effects of NEG-LOG-PROB-SQUARED. Because these variables were standardized, the outcome can be interpreted as the response to changing the negative log probability by one standard deviation from those seen throughout the entire experiment. If the negative log probability of the event changes by one standard deviation, the probability of looking away changes by a factor of $exp(0.099) = 1.10$ for the non-transitional model and $exp(0.194) = 1.21$ for the transitional model. That is, infants are a factor of 1.1 to 1.21 more likely to look away on events that are either highly surprising or highly non-surpising according to an idealized statistical model for learning the structure of the sequences they observe.

The predictions of the the transitional and non-transitions models are difficult to distinguish because they are closely related: the information content of both models are correlated at $R = 0.62$ ($p < 0.001$). However, if both are entered into a stepwise Cox regression, the transitional NEG-LOG-PROB-SQUARED is significant at $p < 0.001$ (coef=0.25, $z = 5.74$)[6], while the non-transitional information content is not significant $p > 0.1$. This provides strong evidence that infants track transitional probabilities, but the null result for the non-transitional model is difficult to interpret due to its correlation with the transitional model and the noise inherent in infant data.

Conclusions & Discussion

These results have explicitly tested two interrelated hypotheses related to infants' looking behavior. First, we constructed a rational, statistical model that used observed events or transitions between events to form probabilistic expectations about what events are most likely in the future. This model embodies a simple, but non-trivial learning theory under which infants follow at least approximately rational statistical inference in inferring properties of the world. Second, we used this model to test whether infants have a preferred information rate in deciding where to allocate attention. The model was necessary in determining what information content a stimulus should convey, to an idealized observer. A failure of either theses assumptions—the probabilistic model or the linking assumption of the relevance of information content—would have yielded a null result.

In our analysis, we we used a Cox regression survival analysis, which allowed us to test the predictions of the model controlling for potential confounds such as the number of items that have not appeared yet, item repeats, and an arbitrary baseline distribution of look-away probabilities. To our knowledge, the hypothesis that infants prefer a fixed information rate has not been tested controlling for these other variables; nor has previous work used this type of formal model in measuring information rate. As such, this work provides several methodological advances. Rather than predicting infants' average looking time to a stimulus, our analysis attempted to predict the precise item in a sequence that an infant would look away on. We found that the information-theoretic properties of a formal model were a significant predictor of infant

[5]Covariates were standardized before including them in the analysis and before squaring them.

[6]The Variance Inflation Factors are small for these variables (< 3.1), suggesting that collinearity is not a substantial problem in computing statistical significance.

Non-transitional model				
Variable	**Coef.**	**Std. Error**	z	**p-value**
TRIAL-NUMBER	0.033	0.006	5.867	0.000
SAME-EVENT	0.213	0.100	2.140	0.032
NEG-LOG-PROB-SQUARED	0.099	0.049	2.024	0.043
Transitional model				
Variable	**Coef.**	**Std. Error**	z	**p-value**
TRIAL-NUMBER	0.033	0.006	5.791	0.000
SAME-EVENT	0.316	0.114	2.772	0.006
NEG-LOG-PROB-SQUARED	0.194	0.047	4.134	0.000
UNSEEN-ITEMS	-0.175	0.089	-1.959	0.005

Figure 2: Included variables using a stepwise Cox regression analysis to predict infant look-aways. In predictions of both transitional and non-transitional models, the squared (standardized) negative log probability is a significant predictor of look-aways.

look-aways, over and above the effects of other variables, but that their effect was U-shaped. Thus, the Cox regression validates the trend observed in Figure 1, showing that it does not result from other confounds.

We take these results as strong evidence for the theory that infants are the Goldilocks of the "blooming, buzzing confusion," preferring stimuli with a certain moderate level of information, and are at least approximately rational in their decisions about where to allocate attention.

Acknowledgments

The first and second authors were supported by Graduate Research Fellowships from the National Science Foundation. The research was supported by a grant from the National Institute of Health (HD-37082). We thank Johnny Wen for his help with Matlab; Holly Palmeri, Laura Zimmermann, and Kathryn Schuler for their help preparing stimuli and collecting infant data; Suzanne Horwitz, Katheryn Lukens, Alyssa Thatcher, Lindsay Woods, and Rosemary Ziemnik for their help recruiting and scheduling subjects; and Collin Bannard, Michael S. DeFreitas, Noah Goodman, T. Florian Jaeger, Elissa Newport, Josh Tenenbaum, Ed Vul, Katherine S. White and members of CoCoSci, and the Aslin and Newport labs for comments and suggestions.

References

Berlyne, D. (1960). *Conflict, arousal, and curiosity*. New York: McGraw-Hill.

Dember, W. N., & Earl, R. W. (1957). Analysis of exploratory, manipulatory, and curiosity behaviors. *Psychological Review*, *64*, 91-96.

Fantz, R. L. (1964). Visual experience in infants: Decreased attention to familiar patterns relative to novel ones. *Science*, *1964*, 668-670.

Fiser, J., & Aslin, R. N. (2002). Statistical learning of new visual feature combinations by infants. *Proceedings of the National Academcy of Sciences*, *99*(24), 15822-15826.

Haith, M. M. (1980). *Rules that babies look by: The organization of newborn visual activity*. Lawrence Erlbaum Associates.

Hunter, M. A., & Ames, E. W. (1988). Advances in infancy research. In L. P. Lipsitt (Ed.), (p. 69-95). New York: Academic Press.

Kemler Nelson, D. G., Jusczyk, P., Mandel, D. R., Myers, J., Turk, A., & Gerken, L. (1995). The headturn preference procedure for testing auditory perception. , *18*, 111-116.

Kinney, D. K., & Kagan, J. (1976). Infant attention to auditory discrepancy. *Child Development*, *47*(1), 155-164.

Maye, J., Weiss, D. J., & Aslin, R. N. (2008). Statistical phonetic learning in infants: facilitation and feature generalization. *Dev Sci*, *11*(1), 122–134.

Roder, B. J., Bushnell, E. W., & Sasseville, A. M. (2000). Infants' preference for familiarity and novelty during the course of visual processing. *Infancy*, *1*(4), 491-507.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928.

Saffran, J. R., Johnson, E. K., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition*, *70*(1), 27-52.

Siqueland, E. R., & DeLucia, C. A. (1969). Visual reinforcement of nonnutritive sucking in human infants. *Science*, *165*(898), 1144–1146.

Wagner, S. H., & Sakovits, L. J. (1986). Advances in infancy research. In L. Lipsitt & C. Rovee-Collier (Eds.), (Vol. 4, p. 195-217). Nordwood, NJ: Ablex.

White, K. S., Peperkamp, S., Kirk, C., & Morgan, J. L. (2008). Rapid acquisition of phonological alternations by infants. *Cognition*, *107*(1), 238-265.

Yu, C., & Ballard, D. H. (2007). A unified model of word learning: integrating statistical and social cues. *Neurocomputing*, *70*(13-15), 2149-2165.

Infants' Visual Processing of Faces and Objects: Age-Related Changes in Interest, and Stability of Individual Differences

Marybel Robledo (marobledo@ucsd.edu)[1]
Gedeon O. Deák (deak@cogsci.ucsd.edu)[1]
Thorsten Kolling (T.Kolling@psych.uni-frankfurt.de)[2]

[1]Department of Cognitive Science, University of California
San Diego, San Diego, CA 92093-0515 USA

[2]Johann Wolfgang Goethe-Universität, Institut für Psychologie
D-60054 Frankfurt/Main Germany

Abstract

Longitudinal measures of infant visual processing of faces and objects were collected from a sample of healthy infants (*N*=40) every month from 6 to 9 months of age. Infants performed two habituation tasks each month, one with novel female faces as stimuli, and another with novel complex objects. Different individual faces and objects served as habituation (i.e., visual learning) and dishabituation (i.e., novelty response) stimuli. Measures included overall looking time to the habituation stimuli, slope of habituation, and recovery to the dishabituation stimuli. Infants were more interested in faces than objects, but this was contextualized by task order. The order effect suggests a "habituation of habituation" effect. Infants showed an age-related decrease in interest in objects, but no decrease in interest in faces. This contradicts claims that infants shift around 6-7 months from interest in faces to interest in objects. The results showed modest between-month stability of interest in faces, but little stability in any other behavioral measures. This implies that habituation is driven more by unexplained subject x session x stimulus variance than by "infant IQ."

Keywords: Infant habituation; face processing; longitudinal studies; object perception; infant cognition; stimulus effects; visual preferences.

Introduction

Visual stimulus processing in infants is typically studied in a habituation paradigm. An infant is presented with a stimulus repeatedly until she or he habituates (i.e., meets some criterion of diminished looking time). A novel stimulus is then presented. If the new stimulus is perceived as different, the infant increases the duration of looking at the stimulus. This paradigm is a robust, reliable way to assess visual discrimination in the first year (Fagan, 1970; Fantz, 1964).

Habituation is used to assess more than stimulus discrimination. Psychologists commonly use habituation to estimate infants' cognitive capacity. Total looking time, or longest look to a stimulus, are considered inversely related to cognitive efficiency (Borstein, Pêcheux, & Lécuyer, 1988). Whereas processing speed in children and adults is used as a proxy for overall cognitive efficiency (Salthouse, 1996), there is no analogous measure of cognitive speed in infants. Thus, speed of habituation is taken to indicate how quickly infants process a stimulus. Also, dishabituation might relate to infants' interest in novelty, which might reflect curiosity. These ideas are bolstered by findings that infant habituation predicts later cognitive skills. For example, Thomson, Faulkner, and Fagan (1991) found a correlation between infants' novelty preference and Bayley Scales of Infant Development scores (BSID, a standardized test of cognitive, language, and social skills) at 12 and 24 months of age. Also, a meta-analysis by McCall and Carriger (1993) showed a consistent relation between habituation in the first year and IQ from 1 to 8 years. Thus, there is correlational evidence of a relation between infant habituation speed and later cognitive performance.

If this correlation is the result of some broad factor such as cognitive efficiency, we might expect individual infants to show consistent habituation speed (relative to their cohort) across time and task. However, few studies have tested longitudinal stability of habituation. In one study of infants at 3, 4, 7, and 9 months, the strongest long-term stability was found in longest-look (i.e., peak) duration (Colombo, Mitchell, O'Brien, & Hotowitz, 1987). However, cross-age stability was modest. Also, Bornstein and Suess (2000) found low stability of total looking time over several months. Thus, it is unclear how stable individual infant's rate of habituation is.

A complication in addressing this question is that infants might habituate differently to different stimuli (Arteberry & Bornstein, 2002). For example, infants like to look at high-contrast, colorful, moderately complex objects (e.g., baby toys) (Fantz, 1964). They also like to look at faces (Johnson, Dziurawiec, Ellis, & Morton, 1991). However, it is not clear whether infants like to examine pictures of objects and faces to the same degree, or prefer one to the other. If the latter is true, we do not know how uniform these preferences might be across infants. For instance, children with autism spectrum disorder spend less time looking at faces than age-matched controls (Hutt & Ounsted, 1966). Perhaps some children are relatively faster to habituate to only one kind of stimulus (e.g., faces) but not the other. A related question is, how stable are individual differences in preferences? Does an infant who strongly prefers faces show a long-lived face-preference in looking time?

The answers to these questions will affect how we interpret infant habituation. If habituation is an index of individual differences in cognitive speed, we should find consistent performance from month to month. We might also find consistency in dishabituation (akin to curiosity or attraction to novelty). However, there is no guarantee of stability across different classes of stimuli. For example, perhaps infants show stable processing time (or interest) for faces, but no month-to-month consistency in looking times to objects. Thus, one of our goals was to address how infants' stable or changeable interests or preferences impact their information-processing speed.

Method

Participants

Forty healthy infants (18 girls, 22 boys) were recruited between three to four months of age to participate longitudinally at six months (mean age = 188 days), seven months (mean age = 219 days), eight months (mean age = 249 days), and nine months (mean age = 278 days). Each monthly visit was scheduled within a 10-day window based on the child's birthday. Infants were recruited through announcements and flyers at local hospitals, and visits and flyers at mother-infant recreational groups and infant play groups in San Diego, CA. Infants and parents were of middle-class socioeconomic level; 88% were white and 12% were of African American, Asian American or Hispanic descent. Parents' mean age was 32.4 years and mean education was 16.6 years. Recruitment and testing procedures were approved by the UCSD Human Research Participants Protection committee.

Stimulus and Apparatus

Pictures of 8 faces and 8 objects were used as habituation and dishabituation stimuli. Faces were taken from the Computer Vision Center's AR Face Database (Martinez & Benavente, 1998). We selected faces of young women of apparent Euro-American ethnicity, with mildly pleasant expressions but not full smiles. Lighting, angle, image size, and image resolution are all controlled in the database. All faces are photographed in front of a light background and are stripped of salient non-facial objects like large jewelry or eyeglasses. (See Figure 1.)

Object stimuli were pictures of unfamiliar geometric objects. All objects were colorful and had similar levels of detail. (See Figure 2.) A group of parents had rated a large set of candidate object pictures for familiarity and attractiveness to infants. Low-familiarity but attractive objects were chosen based on these results. All objects were photographed on a white background.

Different face and object stimuli were presented at each testing session. The stimuli were projected onto a white screen, and measured 30.5cm^2. The room lights were kept low while the infant and parent were seated and prepared for testing; the light was gradually dimmed to near-darkness for the test session. Infants requiring postural support sat in a Bumbo® placed on the caregivers lap. Caregivers wore shaded glasses and earphones playing music so that they could not see the pictures or hear infants' vocal reaction. A Cannon GL camera placed directly in front of infants was used to capture a zoomed-in frontal view of infants' faces. (See Figure 3.)



Figure 1: Example of habituation and novel face stimuli (from Martinez & Benavente, 1998)



Figure 2: Example of habituation and novel object stimuli



Figure 3: Infant participant orienting attentively to stimulus image. Even though infant has shifted away from the center of the camera field and twisted her body, the stimulus image is clearly reflected (as a white box) in the center of her cornea.

Design and Procedure

During each visit infants performed two habituation tasks. At 6 and 8 months object-habituation was the first task and face-habituation was the second task (see Figure 1 and 2). At 7 and 9 months face-habituation was the first task and object-habituation was the second task.

Parents were seated 91 *cm* from the projector screen and instructed to secure the infant on their lap. The experimenter then placed the earphones and glasses on the parent and exited. Another experimenter (E2) in an adjacent room then darkened the room and began the task.

Training and Coding. E2 watched the infant's face from a monitor in the control room, and recorded the infants' fixations and look aways. E2 was trained extensively to record infant looking by watching previously taped habituation sessions, until a high accuracy criterion was

attained. E2 watched for white squares reflected on infants' corneas; these are reflections of the stimulus that are centered on the cornea when the infant looks at the stimuli. E2 depressed a button whenever an infant was looking, and released the button when the infant looked away or most of the stimulus square was outside of the pupil.

To calculate reliability, 15% randomly chosen sessions were coded frame-by-frame off-line. A different coder found the first and last frames for the fixation in each trial. Correlations between online and offline coding were r = .998 (p < .01) for peak looking times, and r = .995 (p < .01) for total looking time to the habituation stimulus.

Task and Trials For each habituation task, infants were presented with one stimulus for a maximum of 12 trials. Custom software (InfAttend) tracked the looking time on each trial while E2 depressed the button, and ended when the button had been released (i.e., look-away periods) for 1 s. The program then imposed a 1 sec ISI and advanced to the next presentation. The presentation automatically ended if the infant looked for 20 sec. Habituation was monitored automatically: when looking duration of the last two trials averaged less than 50% of the mean of the two peak (i.e., longest) trials, the next trial presented the novel (dishabituation) stimulus.

After the first task was completed, the infant and parent took a 1-3 min break to have a snack or drink, or diaper change, so they would be comfortable for the second task.

When infants participated at 6 months they had already been to the lab twice to participate in other cognitive tests, including habituation. Thus, in every session infants were already familiar with laboratory settings and personnel, procedures in the testing room, and even habituation tests. Thus, task or setting familiarity cannot explain age differences. Also, infants saw different stimuli in each session, so stimulus novelty was constant across sessions.

Measures

Several habituation measures were calculated for both tasks (i.e., object; face): looking durations on each trial, and total looking time until habituation; number of trials to habituate (i.e., slope); and peak looking duration. Novelty response was calculated as looking time to the dishabituation stimulus, compared to the mean of the two shortest looking times to the habituation stimulus.

Results

Stimulus Type Effects

Infants were more interested in faces than in objects. Although this difference was mediated by an interaction with order (see below), the face preference was reflected in total looking time (Figure 4). To show this we examined total looking time in a 4 (Age) X 2 (Stimulus) MANOVA. The multivariate age effect was not significant, $F(3, 12) = 1.24$, $p = .337$. However, the effect of Stimulus was, $F(1,14) = 6.60$, $p = .022$ ($\eta^2 = .32$), as was the Age X Stimulus interaction, $F(3,12) = 4.45$, $p = .025$ ($\eta^2 = .53$).

Within-subjects contrasts reveal a cubic age X stimulus effect, $F(1,14) = 12.00$, $p = .004$ ($\eta^2 = .46$), related to the order-related interaction described below.

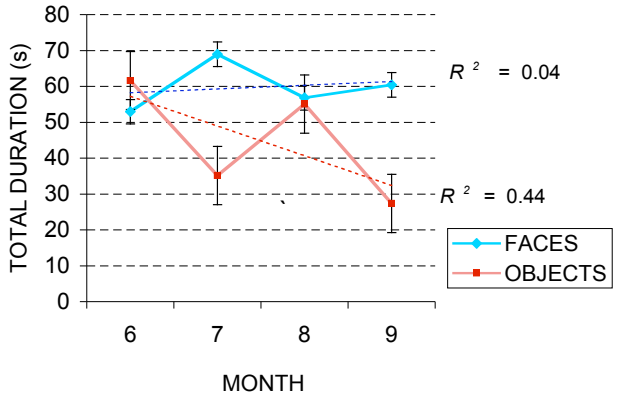


Figure 4: Infants' looking time to faces and objects at each month. Note that presentation order reversed monthly; this explains the oscillation across months of looking times to objects. Bars = *SE*. Best-fitting linear regression lines are shown for each stimulus, with R^2 indicating the age effect for each stimulus type. Note the significant age-related trend of declining attention to objects.

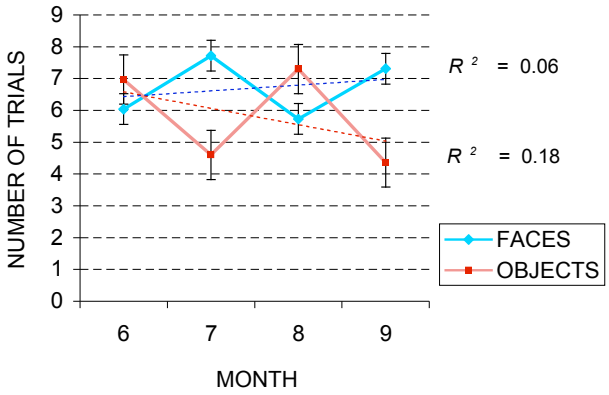


Figure 5: Trials to habituate to faces and objects, by age. Bars = *SE*. Best-fitting linear regression lines are shown with R^2 indicating the age effect for each stimulus type.

The same pattern of effects was apparent in the number of trials to habituation. A 4 X 2 MANOVA found a non-significant age effect, $F(3, 12) = < 1$, but a significant Stimulus effect, $F(1,14) = 12.6$, $p = .003$ ($\eta^2 = .47$), and Age X Stimulus interaction, $F(3,12) = 10.67$, $p = .001$ ($\eta^2 = .73$). Again, within-subjects contrasts reveal a cubic Age X Stimulus effect, $F(1,14) = 29.4$, $p < .001$ ($\eta^2 = .68$).

For both the face habituation and object habituation task, there are no significant total looking time differences between the months that had the same order of presentation (i.e. 6 and 8; 7 and 9 months). Similarly, for both face and

object habituation, there are no significant number-of-trials differences between months with the same stimulus order.

Task Order Effects

Order effects (i.e., which test was first or second) were found in total looking time to objects. At 6 and 8 months, when objects were first, infants looked longer at the object than they did at 7 and 9 months (Tables 1 and 2) when objects were second. Follow-up *t*-tests showed that total looking times for objects significantly differed between 6 and 7, 6 and 9, 7 and 8, 8 and 9 months (all with a value of $p < .005$). Total looking times to faces also was lower when the face task was second; however, the difference was not significant.

Table 1: Mean total looking time to habituation stimulus at 6 and 8 months (*SD* in parentheses).

	6 months	8 months
Task 1: Objects	61.62 *s* (50.58)	55.06 *s* (28.35)
Task 2: Faces	52.93 *s* (36.44)	56.81 *s* (41.48)

Table 2: Mean total looking time to habituation stimulus at 7 and 9 months (*SD* in parentheses).

	7 months	9 months
Task 1: Faces	68.99 *s* (38.17)	60.44 *s* (35.61)
Task 2: Objects	35.13 *s* (18.86)	27.36 *s* (17.83)

This task order effect was also found for both objects and faces in number of trials to habituate. For each stimulus type, it took more trials to habituate if that stimulus was used in the first task than in the second (Tables 3 and 4). Follow-up *t*-tests showed that number of trials to habituate to objects differed between 6 and 7, 6 and 9, 7 and 8, and 8 and 9 months (all *p*s < .05).

Table 3: Mean total number of trials to habituate at 6 and 8 months (*SD* in parentheses).

	6 months	8 months
Task 1: Objects	6.97 (2.97)	7.30 (2.81)
Task 2: Faces	6.04 (2.22)	5.93 (2.62)

Table 4: Mean total number of trials to habituate at 7 and 9 months (*SD* in parentheses).

	7 months	9 months
Task 1: Faces	7.72 (2.77)	7.31 (2.47)
Task 2: Objects	4.60 (1.61)	4. 36 (1.56)

The second- versus first-task differences shows that overall interest in visual examination of stimuli declined across time in the experimental context. This can be interpreted as a "habituation to habituation" effect (see Sirois & Mareschal, 2002). However, the effect is not uniform: it is modulated by infants' face-preference. Infants' interest is maintained or renewed if after habituating to an object, they are shown a face. In contrast, infants' interest significantly decreases when an object is presented after the face habituation task. We do not know what stimulus properties or biases produced this difference in habituation-of-habituation, but it highlights the importance of examining stimulus-by-task interactions in infant habituation.

Individual Stability: Looking Time

Significant individual stability for total looking time to habituation faces was found between 6 and 8 months (r = .56, $p = .002$), 7 and 8 months ($r = .43$, $p = .020$), 7 and 9 months ($r = .49$, $p = .004$), and 8 and 9 months ($r = .43$, $p = .020$). Stability for looking time to objects was not significant between any pair of months.

There were no significant correlations between sessions in the number of trials to habituate to faces or objects.

Individual consistency across the object and face task within a month was found at 6 months for total looking time to habituate ($r = .53$, $p <= .002$). No significant across-task correlation was found in later months.

Individual Stability: Dishabituation

Stability of dishabituation, or recovery of looking-time to a novel stimulus, was tested. At 6 months, total looking at the habituation face moderately predicted a greater novelty response to the new face ($r = .39$, $p = .024$). The same effect was found at 8 months ($r = .52$, $p = .002$) and at 9 months ($r = .41$, $p = .015$). For objects, the same effect was present at 6 months ($r = .44$, $p = .009$), at 7 months ($r = .43$, $p = .017$), and at 8 months ($r = .36$, $p = .033$).

Gender Differences: Stimulus Preference

Connellan, Baron-Cohen, Wheelwright, Batki, and Ahluwalia (2000) argue that male newborns prefer objects whereas female newborns showed preference for faces (mean age = 36. 7 hrs). Baron-Cohen (2002) claims that there are deep-seated gender differences in social inference and intelligence. However, we did not find gender differences in interest to faces and objects at any age.

Connellan et al. (2000) suggest that the gender effect for stimuli preference present in 1-day-old infants could be due to a biological nature since none have yet had exposure to stimuli. However, infants learn prenatally as well as in their first minutes and hours (Butko, Fasel & Movellan, 2006). Infants between 6 and 9 months of age have had much more exposure to social and non-social stimuli, which might be expected to amplify any nascent gender-based preferences suggested by Connellan et al. (2000). However, no such difference was found.

Discussion

These findings suggest that infant habituation to faces is moderately stable between 6 and 9 months of age, at least with respect to total looking time. However, there was no stability across months in looking-time to objects. Also, there was no stability in the slope of habituation (i.e., number of trials to habituate) for either faces or objects.

These results complement Colombo et al.'s (1987) findings of moderate stability in visual habituation from 3 to 9 months. They found stability in duration of looking to faces, but not in trials-to-habituate or in dishabituation. Consistent with Colombo et al.'s findings, we found stability in looking time to faces, but not in trials-to-habituate or in dishabituation. Our results establish that those effects are somewhat specific to faces. However, we found moderate stability between habituation time and dishabituation at some months, for faces and objects. The reason for this is unclear; perhaps general attention or arousal states contribute to consistent patterns of visual examination from habituation to dishabituation trials.

Some studies of habituation use peak (i.e., longest single trial) looking time as a measure of processing efficiency or of interest. We focused on total looking time on the assumption that it would carry less trial-by-trial error variance. However, we did examine peak fixation times (not reported here). This revealed very limited stability across months.

One implication is that longitudinal prediction of infant cognitive efficiency is stimulus-dependent. Researchers have not known how different stimuli in habituation tests predict individual differences in infants' cognitive speed. Our data show that stability is dependent on the type of stimulus tested, as well as the type of response measured and the age of the infant. It is unclear why stability is greater for faces than for objects. It might be that interest in faces is related to dimensions of temperament that relate to sociability; these dimensions show some stability in infants (Garcia-Coll et al., 1992). Conversely, interest in objects—particularly pictorial representations of objects, which do not allow typical multimodal exploration—might be highly subject to episodic and stimulus-specific preferences. It was not true, contrary to claims by Baron-Cohen (2002) and colleagues, that gender predicted stimulus-interest.

It should be noted that "stimulus" here, and in most studies, refers to pictorial representations, which are unnatural in many ways. The use of live models and real objects might considerably alter these patterns, and this would be an intriguing direction for future research. Notably, in ongoing research we are testing whether the face-habituation trends generalize from static faces to dynamic faces (i.e., videos of rotating faces).

Our findings do not support the claim that infants' interest in faces declines, and interest in objects increases, after about 6 or 7 months (Adamson & Bakeman, 1991). We found no decrease in face interest, but a mild decline in object interest from 6 to 9 months. Although the claim that interest in faces declines during this period is based on very different types of data, our results suggest that there is not a general reversal of interest, but perhaps only a task-specific one. Currently there are no data or theories to explain this. One possibility is that as infants' response capabilities in dynamic environments expand, their relative interests in faces and objects start to differentiate. Their interest might become primed by the response "channels" that become viable in a given situation. Notably, during this age range infants gain response capabilities for object manipulation and for social interaction. These capabilities can be enacted only when responding to real, near-at-hand objects, on the one hand, and live, interactive people, on the other. Thus, we would expect infants' interests to be governed by situations that permit these expanding response channels. By contrast, in a narrow response channel like looking time, with stimuli that are static and non-interactive, we might detect only muted effects of changing interests in people and objects. Thus, infants' expanding action repertoire might influence their interest in people and objects.

Infants attend to objects and faces for approximately equal durations when presented first. However, this is not true when they are presented second. The comparable first-task interest to faces and objects could be due to the novelty of the task. It is known that habituation itself declines with repeated testing (Thompson & Spencer, 1966); this is known as "habituation of habituation." However, these data suggest that infants from 6 to 9 months show more habituation of habituation when a more interesting stimulus, a face, is followed by a less (or less-consistently) interesting stimulus, an object.

For this reason, we cannot make broad generalizations about infants' relative interest in faces versus objects. The differences depend on the sequential context of exposure, as well as the infant's age. Also, infants' familiarity with faces and objects cannot be controlled in any obvious way. We used novel exemplars of faces and objects, but it is unlikely that unfamiliar objects are novel to infants in the same way as unfamiliar faces. Infants have much experience with face processing, and are likely able to make fairly fine discriminations. They also have fast-growing experience with objects, but the nature of their experience is quite different. We can, nonetheless, compare the same infant on the same stimulus types across months, and see if they show parallel stability across stimulus types. The current data show that they do not

The results show that it is not possible to use a single—or even several—visual habituation tasks to draw valid inferences about individual infants' visual information-processing traits. Stable traits, such as they are, appear to be conditional and subtle.

Acknowledgments

This project was funded by a grant from the National Science Foundation (HSD SES-0527756) to G. Deák. We thank the families who participated, and Ana Ramundo, Jenny Nowinski, and Colleen Sheh for assistance in collecting and coding data.

References

Adamson, L., & Bakeman, R. (1991). The development of shared attention during infancy. *Annals of Child Development*, 8, 1-41.

Arterberry, M.E., & Bornstein, M.H. (2002). Variability and its sources in infant categorization. *Infant Behavior and Development*, 25(4), 515-528.

Baron-Cohen, S. (2002). The extreme male brain theory of autism. *Trends in Cognitive Science, 6,* 248-254.

Bornstein, M.H., & Benasich, A.A. (1986). Infant habituation: Assessments of individual differences and short-term reliability at five months. *Child Development*, 57(1), 87-99.

Bornstein, M.H., Pêcheux, M., & Lécuyer, R. (1988). Visual habituation in human infants: development and rearing circumstances. *Psychological Research*, *50*(2), 130-133.

Bornstein, M.H., & Suess, P.E. (2000). Physiological self-regulation and information processing in infancy: Cardiac vagal tone and habituation *Child Development, 71*, 273-287.

Butko, N., Fasel, I., & Movellan, J. (2006). Learning about humans during the first 6 minutes of life. *Proceedings of the 5th International Conference on Development and Learning*, Bloomington, Indiana.

McCall, R.B., & Carriger, M.S. (1993). A meta-analysis of infant habituation and recognition memory performance as predictors of later IQ. *Child Development, 64*, 57-79.

Colombo, J., and others. (1987). The stability of visual habituation during the first year of life. *Child Development*, *58*(2), 474-87.

Colombo, J., Shaddy, D.J., Richman, W.A., Maikranz, J. M., & Blaga, O.M. (2004). The developmental course of habituation in infancy and preschool outcomes. *Infancy*, *5*(1).

Connellan, J., Baron-Cohen, S., Wheelwright, S., Batki, A., & Ahluwalia, J. (2000). Sex differences in human neonatal social perception. *Infant Behavior and Development*, 23(1), 113-118.

Fagan, J.F. (1970). Memory in the infant. *Journal of Experimental Psychology*, *9*, 217-226.

Fagan, J.F. (1972). Infants' recognition memory for faces. *Journal of Experimental Child Psychology*, 14, 453–476.

Fagan, J.F. (1973). Infants' delayed recognition memory and forgetting. *Journal of Experimental Child Psychology*, 16, 424–450.

Fantz, R.L. (1964). Visual experience in infants: Decreased attention to familiar patterns relative to novel ones. *Science,* 146, 668–670.

Feldman, J.F. (1997). Memory and speed: their role in the relation of infant information processing to later IQ. *Child Development*, *68*(4), 630-41.

Feldman, R., & Mayes, L.C. (1999). The cyclic organization of attention during habituation is related to infants' information processing. *Infant Behavior and Development*, *22*(1), 37-49.

Garcia Coll, C.T., Halpern, L.F., Vohr, B.R., Seifer, R., & Oh, W. (1992). Stability and correlates of change of early temperament in preterm and full-term infants. *Infant Behavior And Development 15*, 137-153.

Hutt, C., & Ounsted, C. (1966). The biological significance of gaze aversion with particular reference to the syndrome of infantile autism. Behavioral Science, 11 (5), 346–356.

Johnson, M.H., Dziurawiec, S., Ellis, H.D., Morton, J. (1991). Newborns preferential tracking of facelike stimuli and its subsequent decline. *Cognition, 40*, 1 - 21.

Martinez, A.M. & Benavente. R. (1998, June). *The AR Face Database*. CVC Technical Report #24.

Robertson, S.S., Bacher, L.F., & Huntington, N.L. (2001). The integration of body movement and attention in young infants. *Psychological Science*, *12*(6), 523–526.

Roder, B., Bushnell, E., & Sasseville, A. (2000). Infants' preferences for familiarity and novelty during the course of visual processing. *Infancy*, (4), 491–507.

Rose, S. A., & Feldman, J. F. (1997). Memory and speed: their role in the relation of infant information processing to later IQ. *Child Development*, *68*(4), 630-41.

Salthouse, T.A. (1996). The processing speed theory of adult age differences in cognition. Psychological Review, 103, 403-428.

Sirois, S., & Mareschal, D. (2002). Models of habituation in infancy. *Trends in Cognitive Sciences, 6* (7), 293–298.

Tamis-LeMonda, C.S., & Bornstein, M.H. (1989). Habituation and maternal encouragement of attention in infancy as predictors of toddler language, play, and representational competence. *Child Development*, 738-751.

Thompson, L.A., Fagan, J.F., & Fulker, D.W. (1991). Longitudinal Prediction of Specific Cognitive Abilities from Infant Novelty Preference. *Child Development*, *62*(3), 530-538.

Thompson, R.F. & Spencer, W.A. (1966). Habituation: a model phenomenon for the study of neuronal substrate of behavior. *Psychological Review, 73*, 16–43.

Mechanisms of Sustained Selective Attention in 3- to 5-year-old Children: Evidence from a New Object Tracking Task

Anna V. Fisher (fisher49@andrew.cmu.edu)
Department of Psychology, 5000 Forbes Avenue
Pittsburgh, PA 15213 USA

Abstract

Sustained selective attention is a crucial component of many higher-order cognitive processes; yet there is little research into the mechanisms of this ability early in development. One of the challenges in investigating mechanisms of sustained selective attention in young children is lack of appropriate experimental paradigms. This paper reports findings from a novel paradigm designed to investigate mechanisms of sustained selective attention in young children - the Object Tracking task. Results of two experiments with 3- to 5-year-old children provided support to the notion that development of the endogenous component of selective sustained attention lags behind the development of the exogenous component of this process. Importantly, the Object Tracking paradigm allowed investigating both of these components within the same task, thereby making it possible to attribute changes in performance to different mechanisms of attentional control rather than to differences in the level of motivation and engagement in different tasks.

Keywords: Selective attention; Sustained attention; Focused Attention; Cognitive Development.

Introduction

The ability to selectively sustain attention is crucially important because it is an essential component of most higher-order cognitive processes, such as categorization, language comprehension, reasoning, and problem solving. For example, it takes preverbal infants as little as 500 ms to locate a target object among eight distracters (Adler & Orprecio, 2006), whereas it takes them approximately twenty times longer to categorize a single object (Quinn & Eimas, 1996). Similar latency differences between simple visual search tasks and higher-order categorization tasks are also present in older children and adults (Fisher, in press; Gerhardstein & Rovee-Collier, 2002; Trick & Enns, 1998). However, development of the mechanisms of sustained selective attention, also referred to as focused attention, has been sparsely investigated. The goal of the research presented below was to investigate mechanisms of sustained selective attention in 3- to 5-year-old children.

When several objects are present in a scene and one needs to focus attention on a single object, how is the competition resolved? One of the paradigms that has been widely used to explore this question in the domain if visual attention is the visual search paradigm pioneered by Treisman and Gelade (1980). The classic finding from this paradigm is that when adults are asked to search a visual array for a target object defined by a conjunction of features (e.g., color and shape), their reaction time increases with the increase in the number of distracter objects in the display. However, when displays contain target objects defined by a single feature (e.g., color) visual search reaction times remain constant regardless of the number of distracters, as the target object seems to instantly "pop-out" from the display.

While there is no consensus on the mechanisms of visual search, many theories distinguish between two broad ways in which competition between multiple objects in a scene can be resolved. One way has been characterized as stimulus-driven, effortless, bottom-up, and passive, whereas the other way has been characterized as participant-driven, effortful, top-down, and active (Lavie, 2005; Lavie & Tsal, 1994; Kastner & Undergleider, 2000; Norman & Shallice, 1986; Schneider and Shiffrin, 1977; Schneider & Chein, 2003).

Research on the development of selective attention indicates that even newborns are not indifferent to what they attend to, and prefer to attend to some stimuli over others (Colombo, 2001; Fantz, 1963). However, this selectivity has been characterized as stimulus-driven or automatic (i.e., driven by exogenous factors), rather than participant-driven or voluntary (i.e., driven by endogenous factors). In particular, selective attention in newborns and young infants is driven to a large degree by the properties of the stimulus, such as its frequency and duration (for auditory stimuli) and intensity and brightness (for visual stimuli), rather than by infants' intentions (Bornstein, 1990; Ruff & Rothbart, 1996).

By the time infants reach seven months of age, their allocation of attention is driven by a complex interaction of exogenous and endogenous factors (Oakes, Kannass, & Shaddy, 2002). For instance, exogenous factors, such as stimulus brightness and complexity still exert a strong pull on attention allocation; however, reorientation to salient distracters is less likely when infants are in a state of focused attention (i.e., concentrating on a particular toy or activity) than when infants are in a state of casual attention – suggesting that internal state of an infant (an endogenous factor) plays a role in how attention is allocated (Tellinghuisen, Oakes, & Tjebkes, 1999).

Considerable evidence suggests that when several objects compete for attention and one of these objects is defined by a unique feature, similar to adults, infants as young as 3 ½-months of age exhibit the "pop-out" effect (Adler & Orprecio, 2006; Gerhardstein & Rovee-Collier, 2002; Treisman & Gelade 1980). Search for objects defined by a conjunction of features has not been studied with preverbal infants, however findings with 12- to 36-month old toddlers

indicate that their response latency increases with increased number of distracters in the display – a pattern that is similar to that in adults (Gerhardstein & Rovee-Collier, 2002; Scerif, Cornish, Wilding, Driver, & Karmiloff-Smith, 2004; Treisman & Gelade 1980). Despite considerable quantitative differences in performance of children and adults persisting until at least until ten years of age (Trick & Enns, 1998), the qualitative pattern of results from the visual search tasks with young children is similar to that of adults.

However, higher-order cognitive processes (such as categorization, language comprehension, and reasoning among many others) impose greater demands on attention than simply selecting an object for processing. One of these demands is sustaining attention to the selected object for at least brief periods of time. Development of this ability has been often examined in natural settings (such as free play) in prior research as well as computerized vigilance-type tasks (Oakes, Kannass, & Shaddy, 2002; Ruff & Lawson, 1990; Sarid & Breznitz, 1997; Tellinghuisen, Oakes, & Tjebkes, 1999). These studies indicate dramatic improvements in this ability between 12 months and six years of age. For example, studies utilizing the context of free play suggest that duration of focused free play increases from approximately four minutes in 2- and 3-year-old children to over nine minutes in 5- and 6-year-olds (Ruff & Lawson, 1990; Sarid & Breznitz, 1997). Furthermore, these studies indicate that older children are markedly less distractible than younger children, and also more likely to return to an interrupted activity.

Another kind of paradigm that has been successfully used to investigate sustained selective attention in young children is a Continuous Performance Test – a vigilance-type task modeled after tests used with adults (Warm & Jerison, 1984). In this task participants are asked to attend to a stream of visual stimuli and to respond to a target stimulus while withholding response to non-target stimuli. For example, participants might be presented with a series of images depicting ducks and turtles, and instructed to press a button every time they see a duck and avoid pressing the button when they see turtles (Akshoomoff, 2002). The goal of this task is to investigate whether participants can remain alert for prolonged periods of time (e.g., 5- to 9-minute intervals) and accurately detect infrequently appearing target objects. A typical finding of such studies is that approximately 50% of 3 ½-year-old children fail to complete this task, indicating difficulty in sustaining their attention (Akshoomoff, 2002; Corkum, Byrne, & Ellsworth, 1995). Those 3-year-olds who can complete the task (thus demonstrating their ability to maintain attention for prolonged periods of time) exhibit high rates of both misses and false alarms, suggesting difficulty with the voluntary control of selectivity. Marked improvement on this task (in terms of proportion of children completing the task, response time, and accuracy) is observed between four and five years of age.

Studies of sustained selective attention in the context of free play and vigilance-type tasks provide valuable insights regarding the milestones in the development of this important ability. However, these studies are limited in their ability to assess the mechanisms of sustained selective attention in young children and changes in these mechanisms in the course of development. One of the challenges in investigating this question stems from the lack of appropriate experimental paradigms. For example, it has been argued that differences in the level of performance on existing tasks of focused attention between younger and older children may arise as a result of differential levels of motivation and engagement in the task rater than developmental changes in mechanisms of attentional control (Ruff & Rothbart, 1996). Furthermore, there is currently no task that makes it possible to assess contribution of exogenous and endogenous factors to selective sustained attention within the same task, thus making it difficult to uniquely attribute changes in performance to different attentional mechanisms rather than to task-specific factors. The goal of the present research was to develop a task suitable for investigation of the mechanisms of sustained selective attention in young children, and to use this task to investigate the contribution of exogenous and endogenous factors to sustained selective attention in 3- to 5-year-old children.

The Object Tracking Task

The Object Tracking task is reminiscent of the Multiple Object Tracking (MOT) task used with adults to study properties of visual attention (Pylyshyn & Storm, 1988; Yantis, 1992). In the MOT task participants are asked to visually track several identical target objects moving along random trajectories among a larger set of identical objects, also moving along random trajectories. In this paradigm target objects are distinct only at the beginning of each trial (all target objects pulsate for a brief period of time at the onset of each trial), however adult participants (often to their own surprise) are capable of tracking four targets in the field of eights distracters with accuracy approaching 90% (Pylyshyn & Storm, 1988). While this paradigm has been successfully used to investigate properties of object-based attention in adults for over twenty years, our pilot testing suggested that the task is prohibitively complex for young children. Furthermore, the MOT paradigm does not allow assessment of automatic and voluntary components of sustained selective attention within the same task. The new Object Tracking task was created specifically to investigate mechanisms of sustained selective attention with young children.

In the Object Tracking task participants are presented with a three by three grid, with each of the nine grid locations identified by a popular cartoon character, and a target object moving on the grid along a random trajectory. Participants are asked to visually track the target and identify the grid location last visited by the target before it disappears. The moving target in this task can be accompanied by zero to eight distracters, also moving along a random trajectory. Target and distracter objects are randomly selected on each trial from a pool of nine different

geometric shapes. At the beginning of each trial participants are presented with still objects, and the object designated as the target is clearly marked at the beginning of each trial by being encircled in red (see Figure 1 for a schematic depiction of the task).

There are no restrictions on the motion paths of distracter objects, but there are two restrictions on the motion paths of the target objects. First, the target object has to disappear in the middle of one of the nine cells to reduce possible confusion if the target diapers on the border of two or more cells. Second, the target object must visit all nine screen locations at least once before disappearing. In all the experiments presented below, the speed of motion for all target and distracter objects was set at 800 pixels per frame at 30 frames per second (this speed was chosen during pilot testing with a separate group of 3- to 5-year-old children). Average trial duration was approximately 11 sec (a more detailed description of trial duration is provides in the Methods section).

When presented with the task, participants are explained that (1) objects will start moving when the experimenter pushes a button, (2) the goal of the task is to watch the object encircled in red, (3) the red circle will disappear as soon as objects start moving, and (4) once all objects disappear from the screen participants will need to point to the grid location last visited by the target object. Notice, that participants are not asked to perform visual search since the target is clearly marked at the beginning of each trial.

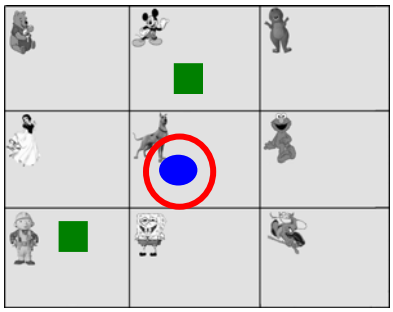


Figure 1. Schematic depiction of the Object Tracking task.

It has been demonstrated that salient objects engage attention automatically, whereas less salient objects may require voluntary processing (Koch & Ullman, 1985; Smith, et. al., 1996; Trick & Enns, 1998; Underwood, et. al., 2006). Therefore, distracter manipulations in the Object Tracking task can allow assessment of the contribution of exogenous and endogenous factors to selective sustained attention in the visual domain. In particular, it is expected that target objects will be more salient when distracters are identical to each other and different from the target (All Same Distracters condition) than when distracter objects are different from the target and from each other (All Different Distracters condition) (Treisman & Gelade, 1980). Thus, tracking accuracy in the All Different Distracters condition will reflect the contribution of predominantly endogenous factors, whereas tracking accuracy in the All Same Distracters condition will reflect the contribution of both exogenous and endogenous factors. The difference in performance between these conditions will be reflective of the unique contribution of exogenous attention. Experiment 1 investigated mechanisms of sustained selective attention in 3- to 5-year-old children using the Object Tracking task in which target objects were accompanied by two distracters, and Experiment 2 investigated performance with six distracters.

Experiment 1

Method

Participants

Participants were 15 3-year-old children (M = 3.66 years, SD = .28 years; 5 females and 10 males), 17 4-year-old children (M = 4.49 years, SD = .25 years; 5 females and 12 males), and 18 5-year-old children (M = 5.23 years, SD = .23 years; 7 females and 11 males).

Design and Procedure

There were two within-subject conditions in Experiment 1: All Same Distracters and All Different Distracters condition. The order of these two conditions was counterbalanced across participants; both conditions were completed on two separate testing sessions that were spaced one to two weeks apart.

As described in the introduction, the Object Tracking task is designed such that the target objects have to appear at least once in each of the nine on-screen locations and disappear in the middle of one of these locations. Due to these restrictions, trial duration is not fixed but varies slightly from trial to trial. In Experiment 1, minimum trial duration was set to 10 s and mean trial duration was 11.00 s (SD = 0.95 s) in the All Same condition and 10.98 s (SD = 1.03 s) in the All Different condition.

To control for the possibility that any observed differences in tracking accuracy may stem from children being more likely to remember what object they were supposed to track in the All Same Distracters condition than in the All Different Distracters condition, at the end of each trial participants were asked to identify which object served as target on the trial they had just completed. Children were presented with a card depicting all nine shapes that could serve as target objects in this task, and asked to point to the shape they had been tracking.

All participants were tested by hypothesis-blind experimenters in quiet rooms in their day care centers. Participants completed 11 trials of the Object Tracking task in each condition. The first trial was completed with assistance from the experimenter who traced the moving target with their index finger. Participants were then explained that they needed to complete the rest of the task by themselves, tracking the target objects only with their

eyes. Data from the first experimenter-assisted trial were discarded from the analyses.

Results

Memory Accuracy

Accuracy with which children recognized the target object (among 9 possible objects) at the conclusion of each trial is presented in Table 1. Memory scores were submitted to a mixed ANOVA with age as a between-subject factor and condition (All Same vs. All Different) as a within-subject factor. This analysis indicated a main effect of age, F (2, 47) = 5.77, p < .01, η2 = .20. Post-hoc Tukey HSD tests indicated that overall memory accuracy was lower in 3-year-old children (M = .67) than in both older age groups (p < .05), and statistically equivalent in 4- and 5-year-old children (M = .83 and M = .86, respectively). Most importantly however, there was no effect of condition and no age-by-condition interaction, both Fs < 1, ns. Therefore, results of the memory check indicate that if any significant differences in object tracking accuracy are observed between the All Same and the All Different Distracters conditions, these differences are unlikely to stem from differential demands on working memory.

Table 1: Memory accuracy in Experiments 1-2 (standard deviations in parentheses).

	All Same Distracters	All Different Distracters	t-test *p*-values
Experiment 1 (2 Distracters)			
3-y.o.	.69 (.22)	.65 (.22)	p > .53
4-y.o.	.83 (.17)	.83 (.17)	p > .83
5-y.o.	.86 (.13)	.86 (.22)	p > .95
Experiment 2 (6 Distracters)			
3-y.o.	.47 (.31)	.44 (.39)	p > .75
4-y.o.	.75 (.29)	.73 (.29)	p > .83
5-y.o.	.79 (.19)	.85 (.2)	p > .44

Object Tracking Accuracy

Tracking accuracy scores were averaged across 10 trials for each participant and submitted to a mixed ANOVA with experimental condition (All Same and All Different Distracters) as a within-subject factor and age (3-, 4-, and 5-years of age) as a between-subject factor. Results of this analysis revealed a main effect of experimental condition, F(1, 47) = 11.46, p < .002, η2 = .19 and age F (2, 47) = 8.04, p < .002, η2 = .25. These main effects were qualified by an age by condition interaction, F (2, 47) = 3.43, p < .05, η2 = .13.

Planned comparisons indicated that participants in all conditions in all age groups identified the final location of the target object at above chance level (chance = 11% given nine response options), all one-sample ts > 5.85, ps < .0001. Five-year-old children were equally accurate in both conditions (83% and 84% of correct in the All Same and All Different condition, respectively) paired-sample t (17) < 1, *ns*. However, younger children exhibited higher tracking accuracy in the All Same than in the All Different condition (see Figure 2): 4-year-olds averaged 76% and 65% of correct responses, respectively, paired-sample t (16) = 2.26, p < .05, Cohen's d = .57; and 3-year-olds averaged 67% and 48% of correct responses, respectively, paired-sample t (14) = 2.63, p < .05, Cohen's d = .77.

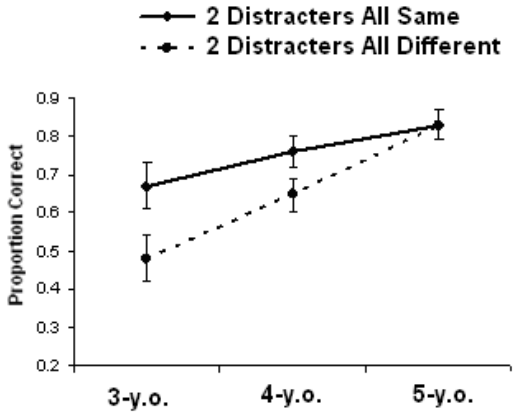


Figure 2. Tracking accuracy scores in Experiment 1.

Overall, results of Experiment 1 suggest that the ability to accurately track an object amidst heterogeneous distracters shows more protracted development than the ability to accurately track an object amidst homogenous distracters. Notice however, that 5 year-old children in Experiment 1 exhibited no effect of condition on tracking accuracy. At the same time, it has been shown that voluntary control of attention continues to mature well beyond the preschool years (Casey, Tottenham, & Fossella, 202; Trick & Enns, 1998). It is possible therefore, that condition differences in tracking accuracy will emerge in 5-year-old children if the task difficulty is increased (e.g., by increasing the number of distracters in the task). This possibility was investigated in Experiment 2.

Experiment 2

Method

Participants

Participants were 15 3-year-old children (M = 3.33 years, SD = .27 years; 6 females and 9 males), 16 4-year-old children (M = 4.41 years, SD = .32 years; 8 females and 8 males), and 20 5-year-old children (M = 5.33 years, SD = .37 years; 11 females and 9 males).

Design and Procedure

Design and procedure of Experiment 2 were identical to that of Experiment 1 with one important exception: the number of distracter objects was increased to six (compared to two distracters in Experiment 1). Mean trial duration was 11.00s

(*SD* = .94s) in the All Same condition and 10.92s (*SD* = .86s) in the All Different condition.

Results

Memory Accuracy

Memory accuracy data are presented in Table 1. Memory scores were submitted to a mixed ANOVA with age as a between-subject factor and condition (All Same vs. All Different) as a within-subject factor. The analysis indicated a main effect of age, F (2, 48) = 11.08, $p < .0001$, $\eta 2 = .33$. Post-hoc Tukey HSD tests indicated that overall memory accuracy in 3-year-old children ($M = .45$) was lower than in older children ($p < .05$), and statistically equivalent in 4- and 5-year-old children ($M = .74$ and $M = .82$, respectively). Similar to Experiment 1, there was no effect of condition and no age-by-condition interaction, both Fs < 1, *ns*.

Object Tracking Accuracy

Tracking accuracy scores were submitted to a mixed ANOVA with experimental condition (All Same and All Different) as a within-subject factor and age (3-, 4-, and 5-years of age) as a between-subject factor. Results of this analysis revealed a main effect of experimental condition F (1, 48) = 23.40, $p < .0001$, $\eta 2 = .32$, and a main effect of age F (2, 48) = 8.93, $p < .005$, $\eta 2 = .27$. Unlike Experiment 1, the age by condition interaction did not reach significance, F (2, 48) < 1, *ns*.

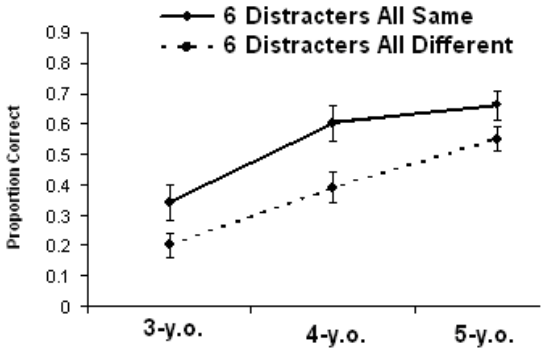


Figure 2. Tracking accuracy in Experiment 2.

Similar to Experiment 1, participants in all conditions in all three age groups identified the final location of the target object at above chance level (chance = 11%), all one-sample *t*s > 2.44, *p*s < .03. However, unlike Experiment 1, 5-year-old children exhibited the effect of condition with higher accuracy in the All Same condition (66%) than in the All Different condition (55%), paired-sample t (19)= 2.52, $p < .05$, Cohen's d=.35. Similarly, 4-year old children exhibited higher accuracy in the All Same condition compared to the All Different condition (60% and 39%, respectively), paired-sample t (15)= 3.71, $p < .005$, Cohen's d=.97, as did 3-year-old children (34% and 20%, respectively), paired-sample t (14)= 2.07, $p = .05$, Cohen's d =.69.

Across the two experiments, it is appears that overall level of performance in all three age groups was lower in Experiment 2, when six distracters were present, than in Experiment 1, when two distracters were present. Indeed, when the data from both experiments were submitted to a mixed ANOVA with age and number of distracters (two vs. six) as between-subject factors and type of distracters (All Same vs. All Different) as a within-subject factor, the analysis revealed a main effect of the number of distracters, F (1, 95) = 34.09, $p < .0001$. This main effect was qualified by the distracter number by distracter type interaction, F (1, 95) = 33.88, $p < .0001$, suggesting that decrease in accuracy with the increase in the number of distracters was greater in the All Different condition than in the All Same condition (mean decrease in accuracy across all age group was 27% and 22%, respectively).

General Discussion

This paper presents findings from a novel task in which children were asked to track a moving target object accompanied by distracters that varied in type (all same versus all different) and number (two versus six). The results pointed to several novel findings. First, tracking accuracy improved with age. Second, tracking accuracy was higher in the All Same Distracters condition than in the All Different Distracters condition for all age groups when target objects were accompanied by six distracters; a similar difference between conditions was observed in 3- and 4-year-old children when targets were accompanied by two distracters. Third, unlike the visual search tasks, increase in the number of homogenous distracters resulted in lower accuracy for all three age groups tested in this study. Finally, there was no effect distracter type on children's ability to remember which object they were supposed to track; therefore, effects reported in this paper can not be attributed to differences in memory demands in different conditions.

The central finding reported in this paper is that preschool-age children are more successful at tracking targets moving among homogeneous than among heterogeneous distracters. This pattern of performance may arise for two different reasons. Consistent with the notion that the speed of engaging attention (or attention-getting) and speed of releasing attention (or attention-holding) are separate factors (Cohen, 1972), one possibility is that homogeneous distracters provide less competition for attentional resources and therefore children are less likely to glance away from the target moving amidst identical distracters. In other words, low competition for attentional resources may enhance attention-holding properties of the target. Alternatively, it is possible that children are equally likely to glance away from the target regardless of the type of distracters; however children are more successful in locating the target after glancing away in the homogeneous than in the heterogeneous distracter condition. In other words, low competition for attentional resources may

enhance attention-getting properties of the target. These possibilities remain to be addressed in future research.

Overall, findings presented above support the notion that development of endogenous attention (probed by the All Different Distracters condition) lags behind the development of exogenous attention (probed by the All Same Distracters condition). Importantly, the Object Tracking task makes it possible to assess both mechanisms within the same paradigm and quantify this lag in terms of the differences in tracking accuracy. Therefore, this new paradigm allows attributing changes in performance to different mechanisms of attentional control rather than to differences in the level of motivation and engagement in different tasks.

Acknowledgments

I thank children, parents, teachers, and administrators for their participation. This research was supported by NICHD though Grant 1RO3HD060086-01A1.

References

Adler, S.A., & Orprecio, J. (2006). The eyes have it: Visual pop-out in infants and adults. *Developmental Science*, 9, 189-206.

Akshoomoff, N. A. (2002). Selective attention and active engagement in young children. *Developmental Neuropsychology*, 22, 625-642.

Bornstein, M. H. (1990). Attention in infancy and the prediction of cognitive capacities in childhood. In J. Enns (Ed.) *Development of Attention: Research and Theory*. Amsterdam: Elsevier.

Casey, B.J., Tottenham, N., & Fossella, J. (2002). Clinical, imaging, lesions, and genetic approaches toward a model of cognitive control. *Developmental Psychobiology*, 40, 237-.254.

Cohen, L. B. (1972). Attention-getting and attention-holding processes of infant visual preferences. *Child Development*, 43, 869–879.

Colombo, J. (2001). The development of visual attention in infancy. *Annual Review of Psychology*, 52, 337-367.

Corkum, V., Byrne, J. & Ellsworth, C. (1995) Clinical Assessment of Sustained Attention in Preschoolers. *Child Neuropsychology*, 1(1), 3-18.

Culham, J. (2003). Attention-grabbing motion in the human brain. *Neuron*, 40, 451-452.

Fantz, R.L. & Miranda, S. B. (1975). Newborn infant attention to form and contour. *Child Development*, 46, 224-228.

Fisher, A.V. (in press). Mechanisms of Induction Early in Development. In M. Banich & D. Caccamise (Eds.) *Generalization of Knowledge: Multidisciplinary Perspectives*. New York: Psychology Press.

Gerhardstein, P., & Rovee-Collier, C. (2002). Visual search in infants and very young children. *Journal of Experimental Child Psychology*, 81, 194-215.

Kannass, K.N., Oakes, L.M., & Shaddy, D.J. (2006). A longitudinal investigation of the development of attention and distractibility. *Journal of Cognition and Development*, 7, 381-409.

Kastner S. & Ungerleider L.G. (2000). Mechanisms of visual attention in the human cortex. *Annual Review of Neuroscience*, 23, 315-341.

Koch, C. & Ullman, S. (1985). Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiology*, 4, 219-227.

Lavie, N. (2005). Distracted and confused? Selective attention under load. *Trends in Cognitive Sciences*, 9, 75-82

Lavie, N. & Tsal, Y. (1994). Perceptual load as a major determinant of the locus of selection in visual attention. *Perception and Psychophysics*, 56, 183-197.

Norman, D. A. & Shallice, T. (1986). Attention to action: Willed and automatic control of behaviour. In R.J. Davidson, G.E. Schwartz, & D. Shapiro (Eds.), *Consciousness and self-regulation*, Plenum Press.

Pylyshyn, Z. W., & Storm, R. W. (1988). Tracking multiple independent targets: Evidence for a parallel tracking mechanism. *Spatial Vision*, 3, 179-197.

Quinn, P. C. & Eimas, P. D. (1996). Perceptual cues that permit categorical differentiation of animal species by infants, *Journal of Experimental Child Psychology*, 63, 189–211.

Ruff, A.H., & Lawson, K.R. (1990). Development of sustained, focused attention in young children during free play. *Developmental Psychology*, 26, 85-93.

Ruff, H., & Rothbart, M. K. (1996). *Attention in early development*. New York: Oxford University Press.

Sarid, M., & Breznitz, Z. (1997). Developmental aspects of sustained attention among 2- to 6-year-old children. *International Journal of Behavioral Development*, 21, 303-312.

Scerif, G., Cornish, K., Wilding, J., Driver, J., & Karmiloff-Smith, A. (2004). Visual search in typically developing toddlers and toddlers with Fragile X or Williams Syndrome. *Developmental Science*, 7, 116-130.

Schneider, W. & Chein, J.M. (2003). Controlled and automatic processing: Behavior, theory, & biology. *Cognitive Science*, 27, 525-559.

Schneider, W. and Shiffrin, R.M. (1977). Controlled and automatic human information processing: Detection, search, and attention. *Psychological Review*, 84, 1-66.

Stechler, G., & Latz, E. (1966). Some observations on attention and arousal in the human infant. *Journal of the American Academy of Child Psychology*, 5, 517-525.

Treisman, A. M., & Gelade, G. (1980). A Feature-integration theory of attention. *Cognitive Psychology*, 12, 97-136.

Trick, L.M. & Enns, J.T. (1998) Lifispan changes in attention: The visual search task. *Cognitive Development*, 13, 369-386.

Warm, J.S., & Jerison, H. J. (1984). The psychophysics of vigilance. In J.S. Warm (Ed.) *Sustained attention in human performance* (pp. 15-59). Chichester, UK: Wiley.

The Influence of Route Planning and its Execution on Spatial Learning

Kayoko Ohtsu (id-ant@moegi.waseda.jp)
Graduate School of Education, Waseda University, 1-6-1, Nishi Waseda, Shinjuku-ku Tokyo, Japan

Yoshihiro Ouchi (oouchi@teikyo-gjc.ac.jp)
Teikyo-Gakuen Junior College, 615-1, Kobutizawa-cho, Hokuto City, Yamanashi, Japan

Abstract

We propose that spatial inferences made during planning and executing a route influence the learning of relative locations through wayfinding. In Experiment 1, separate and combined route plans were compared. The results suggest that inferring multiple directions during the initial stage of planning leads to more accurate representations of relative locations than planning a single route. In Experiment 2, regular and irregular updating modes during the execution phase were compared. The results suggest that irregular updating, which involves multidirectional self-to-object updating, also leads to more accurate representations than regular updating. We conclude that the requirement to make spatial inferences about multiple multidirectional metric interconnections in egocentric reference frames during wayfinding facilitates spatial learning.

Keywords: spatial learning; route planning; wayfinding; egocentric reference frames

Introduction

The means by which humans and animals develop knowledge about their surrounding environments has been a controversial topic for a long time. One theory of the development of spatial knowledge assumes a qualitative change from route knowledge to survey-type knowledge over time (Siegel & White, 1975), and thus the knowledge should become more elaborate as experiences of traveling increase. It is also thought that the qualitative change could occur by automatic and unconscious reorganization of the route knowledge (Thorndyke & Hayes-Roth, 1982). However, there are studies that showed that experiences of an environment do not facilitate spatial learning automatically (Moeser, 1988; Rossano & Reardon, 1999), and repetitive learning does not always efficiently promote the accurate development of knowledge (Ishikawa & Montello, 2006).

The present study examines the relationship between human spatial learning and route planning during wayfinding. Though wayfinding includes a wide range of cognitive activities and behaviors (Gärling, Böök, & Lindberg, 1984), after a destination has been set the basic process of wayfinding is planning and executing of a route in which one decides on and follows between a point of origin and a destination (Golledge, 1999). Specifically, we focused on route planning when moving through environmental spaces such as cities or the interior of buildings. In previous studies, route planning, which incorporates factors such as "short cuts", is often used as a dependent variable that changes with the development of spatial cognition. However, to our knowledge, no study has yet examined the effects of route planning on spatial learning.

Here we assume that spatial inferences during planning and executing a route facilitate the learning of relative locations. This might sound paradoxical because knowledge of relative locations is often thought to be a precondition for planning. Spiers & Maguire (2008) pointed out that when planning a route, the relative direction from the origin to the destination is determined before a specific path can be chosen. In the case when very little is known about a particular environment, how is it possible to find the way to a destination that is out of sight? Given that, to facilitate wayfinding, spatial knowledge of a particular environment is manipulated using rules of inference (Kuipers, 1978). A relative direction must be inferred by representing and manipulating the incomplete knowledge that has already been acquired. For example, when one is not sure which path to take at a four-way intersection in an unfamiliar environment, he or she can express a vague direction to a destination by pointing a finger, which is a spatial inference that people make routinely in their daily lives. The core idea in this study is that the inference of this type will be effective to develop spatial knowledge.

The relationship between relative locations can be described in either an environmental reference frame (object-to-object relations) or an egocentric reference frame (self-to-object relations). However, when deciding on a direction of movement within an environment during wayfinding, it is necessary for a traveler to represent one's body and the destination in an egocentric reference frame (Sholl, 1996) in order to translate one's spatial knowledge into action. On theoretical grounds, self-to-object relations can be represented in a number of ways, for example, as location-dependent reference direction (Poucet, 1993) or in a network of reference frames (Meilinger, 2008). However, the representations commonly contain metric information, defined as the direction and distance from one place to another.

Our expectation was that spatial inferences about self-to-object metric relations would have a facilitating effect when planning a route and updating self-position and orientation at the decision point (e.g. intersections). Gärling et al. (1984) suggested that metrical relations only between important reference points are represented for travel. Naturally, an origin and a destination are such reference points for determining a route at the initial stage. In addition, the decision point should also be the key reference point for following the route. Unlike on-line-type spatial inferences such as narrowly defined path integration, which are based

on continuous updating, people pay attention to metric relations during wayfinding mostly when the need arises such as when one chooses a path at the decision point.

Two experiments were conducted to compare incidental learning outcomes when planning and executing different types of the routes using a direction estimation post-test that reflects the structure of self-to-object representations.

In Experiment 1, separate and combined route plans were compared. We assumed that number of the goal directions that participants were required to infer at the start would affect their learning of relative locations within an environment. When a traveler is visiting multiple places, if he or she makes separate route plans (i.e. plan a route to the first place, move to that place, and then plan the route to the next one), one will just compute one direction each time. In contrast, to make a combined route plan for the complete round of visits, the traveler would have to consider multiple interconnections between the origin and the destinations at the same time and effectively learn the interconnections.

In Experiment 2, two types of order of visiting, which led to regular or irregular updating, were compared. In regular updating one constantly updated one's position to destinations situated in the same self-to-object relation. In irregular updating the destinations were situated in multidirectional self-to-object relations. We assumed that a requirement for different types of directional inference when updating would also affect learning relative locations. If a traveler has to infer multidirectional self-to-object relations through the updating process, rather than constantly updating, they would be able to utilize egocentric reference frames over a wide range of the environment.

The Environments and Settings

A real environment was used to observe spontaneous spatial inferences. Additionally, to achieve a natural response from the participants, we set up the wayfinding task as a role-playing game that involved stories (Appendix A). The experiments took place on the campus of Waseda University with participants aged 18 and older attending a school festival and agreeing to participate in the experiments.

Experiment 1

Method

Participants Out of a total of fifty-six participants, who were randomly assigned to each group (Single-Goal or Multiple-Goals), fifty people (mean age 22.0) were included in the analyses. Three women in the Multiple-Goals group made errors in the wayfinding task and were excluded from the analyses. Thus the last three female participants in the Single-Goal group were also excluded, so that both groups contained 25 people with the same male-to-female ratio (9:16).

Materials Labyrinth 1 (7 by 7 meters) was built in a classroom using identical fiberboard sheets (Figure 1, left panel; each panel was 2 meters long and 1 meter wide). Figure 2 shows the layout of the labyrinth and the locations of the four targets, which corresponded to computer displays that showed illustrations of the four residents in the story (Appendix A) and instructions for the wayfinding task. No two displays could be seen at the same time. We developed two programs that were written in Visual Basic for Applications: one controlled the task and recorded the responses, and the other was used for the post-test. The left panel in figure 3 is an example of the operation screen used in the post-test to record the judgments of the participants.

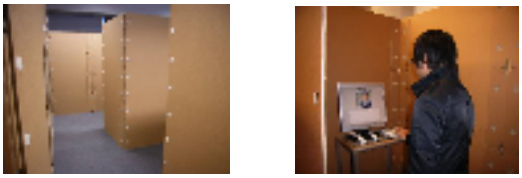


Figure 1: Images of the Labyrinth

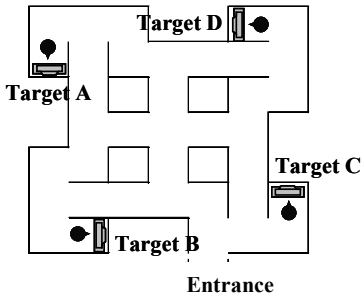


Figure 2: Layout of Labyrinth 1

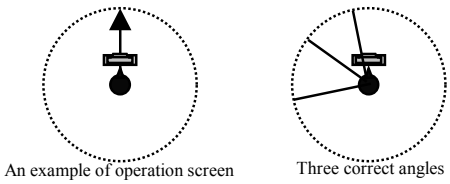


Figure 3: An operation screen and correct angles

Order of Visiting The orders were devised so that the participants did not encounter the same positional relations (Table 1).

Table 1: Orders of Visiting

First round	Second round
1. Target A→B→C→D→A and	A →D→C→B→A
2. Target A→D→C→B→A and	A →B→C→D→A
3. Target A→B→D→C→A and	A →C→D→B→A
4. Target A→C→D→B→A and	A →B→D→C→A
5. Target A→C→B→D→A and	A →D→B→C→A
6. Target A→D→B→C→A and	A →C→B→D→A

Procedure The wayfinding task consisted of two phases: (i) exploring and (ii) visiting (two rounds). The participants were escorted individually from an anteroom to the labyrinth by an experimenter, who monitored the progress of the task from outside the labyrinth.

During the exploring phase, the participants walked around freely and found the four computer displays. When they found a display, they pressed a keypad that was placed in front of each display (Figure 1, right panel). The visiting phase started when the participant found the last display.

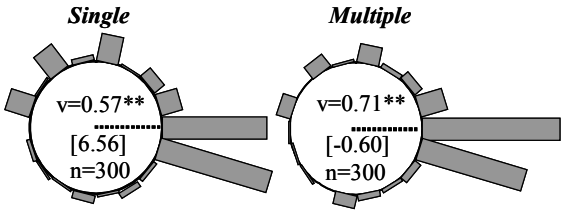
This formed the point of origin of the visiting phase; the point of origin varied depending on how the individual had explored, but because the labyrinth was fully symmetric, each order of visiting involved similar components regardless of the location of the origin. The participants were asked by the resident in the last display to revisit the other residents. In the Single-Goal procedure, only the first target goal was given at the point of origin, and when the participants reached that goal they received the next one. In contrast, in the Multiple-Goals procedure, all three target goals and the order in which they should be visited were given at the point of origin. When the participants reached a target, they pressed the keypad. The task ended after two rounds of revisiting.

After the task, the participants were escorted to another room and took the post-test. They were informed that the experiment included "easy quizzes about your memory and sense of direction". After five filler questions that asked about the story, they were asked to indicate 12 relative directions in the following manner: "if you were standing and facing the target X, indicate the direction of target Y". A computer display used in the test was placed horizontally on a table. The participants viewed the operation screen (Figure 3, left panel) from above and indicated the directions by turning the arrow clockwise or counterclockwise using keypads. The graphic shows a birds-eye view of a participant standing in front of a computer display. Instructions of 12 combinations of X and Y were presented one by one randomly at the top of the screen. Three solid lines in the right panel of Figure 3 shows 3 correct angles for 12 relative directions (there were four groups of 3 directions that had the same correct angle).

Results

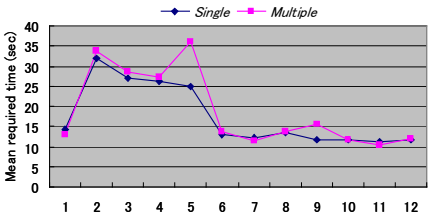
To analyze the 12 relative directions for each group as one data set for each condition, all judgments were adjusted such that the correct angle was 0 degrees. The twelve judgments by each person were analyzed individually to avoid cases where the mean angle corresponded to the correct angle fortuitously (for instance, if two judgments were +120 degrees and –120 degrees, the mean angle would be 0 degrees, the correct angle). Figure 4 shows the mean angles, values for v (a measure of the clustering around a correct direction that decreases as the dispersion increases and varies from –1 to 1), and the results of the V-tests that revealed each data set clustered around the correct angle. The Watson–Williams test revealed that there was no significant difference between the mean angles for the two groups. The accuracy of the judgments was represented by the amount of dispersion, because a greater degree of dispersion meant that more data departed from the correct angle than for a lower value. We compared the dispersions of the two groups by the Mann–Whitney Test, as suggested by Batschelet (1981), and found that the dispersion of Multiple-Goals was smaller than that of Single-Goal ($Z=-2.29$, $p<0.05$).

A T test revealed there was no significant difference between the mean total required times for the two groups. Next, we conducted a two-way analysis of variance (ANOVA) on the required time using the following factors: (F1) the number of goals and (F2) sections (see Figure 5). An effect of F2 ($F(11,528)=76.40$, $p<0.01$) was observed, together with an interaction between the two factors ($F(11,528)=3.05$, $p<0.01$). Student–Newman–Keuls test revealed that among the comparisons between all possible pairs of factor levels, there was only a significant difference between the groups for Section 5 (the first section of the visiting phase). There was no correlation between individual values for v (using 12 judgments per person) and those for the total required time ($r=0.13$ in Single-Goal and -0.33 in Multiple-Goals).



····· : Correct angles *p<.05 **p<.01
[ ]: mean angles (degree)
plus—counterclockwise direction minus—clockwise direction

Figure 4: Frequency distribution graphs



The exploring phase
Section 1: from the start of the task, until a keystroke at 1st target
2: until a keystroke at 2nd target
3: until a keystroke at 3rd target
4: until a keystroke at 4th target (instructions)
The visiting phase (First round)
Section 5: until a keystroke at the first target visited
6: until a keystroke at the second one
7: until a keystroke at the third one
8: until a keystroke at the 4th target (instructions)
(Second round)
Section 9, 10, 11 and 12: the same as 5, 6, 7 and 8

Figure 5: Mean required time for 12 sections

Discussion

The result that the participants in the Multiple-Goals group performed better in the post-test than those in the Single-Goal group supported the hypothesis that combined route planning facilitates the learning of relative locations. We conclude that inferences about multiple interconnections that were made when planning the route improved the accuracy of the judgments made in the post-test. Detailed analysis of the required time showed that participants assigned to the Multiple-Goals group spent more time on Section 5, during which the participants received their first instructions for the visiting phase and reached the first target.

Presumably, this was because the participants in Multiple-Goals had to recall the relative locations of three targets to make a combined-route plan, as well as having to absorb instructions that contained the next three target goals and the order of visiting. Simultaneously, they had to infer three self-to-object relations from their current position to the targets. In contrast, those in the Single-Goal group had to infer just one direction to the next goal. Thus, the difference in time taken shows the complexity in processing the additional directional inferences in Multiple-Goals as compared with Single-Goal.

In both groups a large proportion of clockwise errors appeared (Figure 4) because angles A and C tended to be considered just in front of and on the right hand side respectively from the imagined standing points of the participants. Though there was no difference between mean angles for the two groups, the angle in Single-Goal containing clockwise error shows that more participants in the group had this tendency than in Multiple-Goals.

After the participants became aware of the position of the target at the initial stage of the planning, they could revisit the targets relatively easily and without taking the wrong path because the shape of the labyrinth gave them a reasonably good view of the access aisles and there were multiple accessible paths. Although the specific pathways that were taken during the task were not recorded in Experiment 1, both conditions involved a similar amount of walking because there was no difference in the total required times between the two groups.

Experiment 2

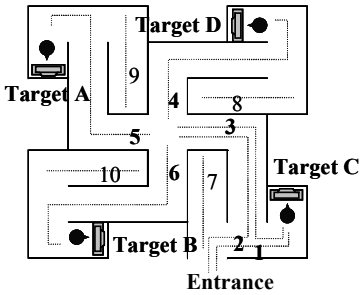
Method

Participants Out of the forty-four participants who were randomly assigned to each group (Circle-Order and Non-Circle-Order), thirty-eight people (mean age, 20.7; male-to-female ratio in each condition, 11:8) were included in the analyses. Three women in each group who made errors in the wayfinding task were excluded from the analyses.

Materials We partially rearranged Labyrinth 1 into the format shown in Figure 6 without changing the locations of the targets and set it up in the same classroom used in Experiment 1. Camcorders were used to record the paths taken by the participants. The other basic materials were the same as those in Experiment 1, except the post-test program the filler questions were redrafted to correspond to the new story (Appendix A).

Order of Visiting Visiting orders 1 and 2 (Table 1) corresponded to the Circle-Order procedure in which the participants visited three targets in a clockwise or counter-clockwise order, for example visiting A→B→C→D→A in the first round, and then A→D→C→B→A in the second round, so that they turned constantly to the right or left at a decision point during each round. The other visiting orders, for example visiting A→B→D→C→A in first round, and then A→C→D→B→A in second round, represented the Non-Circle-Order procedure in which the participant turned right and left turns and going straight ahead at decision points.

Procedure The basic procedure was the same as that of Experiment 1, except that all participants were informed of the three target goals with their visiting order at the point of origin and they carried camcorders during the wayfinding task.



Value for leg 1= 0.5, leg 2=1.5, leg 3, 4, 5, 6=1

Figure 6: Layout of Labyrinth 2

Results

All the judgments were analyzed in the same way as Experiment 1. Figure 7 shows the mean angles, values for v (refer to results of Experiment 1), and the results of the V-tests that revealed each data set clustered around the correct angle. The Watson–Williams test revealed there was no significant difference between the mean angles for the two conditions. The result of the Mann–Whitney Test showed that the Non-Circle-Order group had a smaller degree of dispersion than the Circle-Order group ($Z=-3.13$, $p<0.01$).

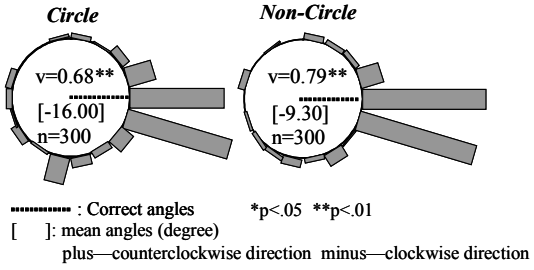


Figure 7: Frequency distribution graphs

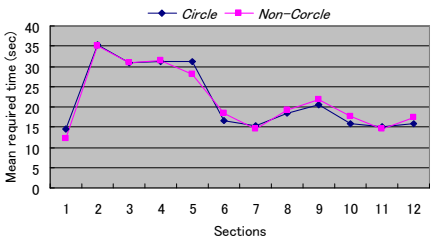


Figure 8: Mean required time for 12 sections

A T test revealed that there was no significant difference between the total required times for the two conditions. We conducted a two-way ANOVA on the required time using

the following factors: (F1) the orders of visiting, and (F2) sections (12 levels). We only detected an effect of F2 ($F(11,396) = 43.67$, $p<0.01$) (Figure 8).

The paths that each participant took inside Labyrinth 2 were detailed by video. The aisles of Labyrinth 2 were divided into 10 legs, and to count and compare the amount of walking, we assigned a value to each leg (Figure 6) and summed the values based on the paths taken by each participant, with the exception of legs 7, 8, 9, and 10, which no one walked. The averages of the total were 23.97 in Circle-Order and 23.87 in Non-Circle-Order. T tests revealed that there was no significant difference between them. During the exploring phase, one person in each group seemed to locate the targets first without pressing the keypads; during the visiting phase 12 people (5 in Circle-Order and 7 in Non-Circle-Order) realized that they had gone the wrong way and retraced their steps. We did not exclude these people from the analyses because, during the exploring phase, only one person in each group did not press the keypads immediately, and during the visiting phrase, all the participants remembered the required order and corrected their course. The other participants took the shortest paths of which the total was 22.5. There was no correlation between individual values for v and the following two values: the total required time ($r=0.08$ in Circle-Order and 0.31 in Non-Circle-Order) and the average of the values based on the paths ($r=-0.23$ in Circle-Order and -0.26 in Non-Circle-Order).

Discussion

The result that the participants in the Non-Circle-Order group performed better in the post-test than those in the Circle-Order group supported the hypothesis that irregular updating facilitates the learning of relative locations. We conclude that multidirectional self-to-object updating at decision points improved the accuracy of the judgments made in the post-test.

It is noteworthy that the value for v of the Circle-Order group was equivalent to that of the Multiple-Goals group in Experiment 1. This can be interpreted as a replication of the effect of multiple goals because the participants in both groups planned combined routes. Unlike Experiment 1, it was not possible to execute the plan only with an awareness of the targets' locations because of the fylfot-shaped labyrinth. To reach the next target goal, the participants had to choose one path by updating their positions to the next target at the decision point. Whereas those in the Circle-Order group inferred the same direction to the targets that were always to the right or left of their body in a given round, those in the Non-Circle-Order group had to infer multiple directions to the targets that were backward right, to the right and to the left in a given round.

The length of time spent in the labyrinth and the amount of walking did not show a direct correlation to performance. There were no differences between the groups with respect to, total required time, time in each section and the number of legs that the participants walked during the task.

General Discussion

The results of the experiments revealed that, regardless of physical experience (e.g. the amount of walking and minor differences in both the migration pathways taken and the number of legs walked), the need to infer metric interconnections between multiple points during the initial stage of planning and while executing a route plan improves the accuracy of representations of relative directions within an environment.

The effect of the initial planning in Experiment 1 is consistent with spatial theories and models that propose that the acquisition of representations about spatial structures through wayfinding involves the integration of local perspectives and views that a traveler has learned independently (e.g. Meilinger, 2008; Poucet, 1993; Sholl & Nolin, 1997). The improved accuracy can be interpreted as the consequence of profound and extensive integration because combined route planning involved the representation of greater amounts of local information and the computation of more metric relations in egocentric reference frames at one time than separate route planning. There are navigational strategies that do not involve inferring metric relations. However, in the experimental situations described here the participants were instructed unexpectedly to revisit three unfamiliar targets in a specific order in a completely new environment. They had to recall which display corresponded to which resident and where it was located. In addition, they had to consider object-to-object positional relations between three targets in order to plan a combined route. These combined-route plans, which contained more directional components than those of the simple plans, appeared to facilitate the integration of the local perspectives and views.

We have addressed the question of why regular updating facilitated learning while irregular updating did not in the discussion of Experiment 2. The effect of irregular updating was different from the effect of direct directional inferences to the targets, which was observed in Experiment 1. In Experiment 2, the participants updated their positions relative to the targets at the decision point in the center of Labyrinth 2, and not in front of the displays; however, in the post-test, they were required to estimate directions from the displays. Thus it can be interpreted that the inferences through the updating had a spillover effect on the estimation of self-to-object directions between the targets. We assume that this effect was due to strong interconnections between the decision point and the targets' locations. During irregular updating the decision point, which was one of the key reference points in a spatial structure of Labyrinth 2, was far more important than that of regular updating (as discussed in the next paragraph), and thus it would be strongly interrelated to the other reference points. If we compare the reference point and path to a node and edge, respectively, in a graph, the participants in Non-Circle-Order might have recognized the interrelation of the points as a graph with five nodes and eight edges (e.g. like a square with diagonal lines), while those in Circle-Order have

recognized that as a graph with four nodes and four edges (i.e. just a square). Though we do not make a decisive conclusion here, it seems reasonable that the former structure of interrelation would have been more advantageous in representing relative locations in an egocentric reference frame than the latter.

The ineffectiveness of regular updating might be caused by a difference in the hierarchical levels of navigational strategies between the two conditions. Trullier, Wiener, Berthoz, & Meyer (1997) proposed a classification of strategies that is based on levels of complexity of required processing and the information that is perceived, represented, and processed. According to the classification, route following that involves regular updating can be substituted with a lower level strategy that requires the participant to regularly turn left or right at the decision point rather than having to compute a metric relation to choose a path at the decision point each time. Thus, this type of regular decision-making during wayfinding might be ineffective at improving representations of relative directions.

Our findings reflect the natural behavior of humans because our participants in the game-like experiments did not know that they were going to be asked the directions in the post-test. The utilization of inferences for planning and executing a route might be one of the key mechanisms by which individuals refine and modify their representations of relative locations in an environment. Differences in the inferences made might be one of the reasons why "individuals with equal levels of exposure to a place will differ in the extent and accuracy of their spatial knowledge (Montello, 1998)".

References

Batschelet, E. (1981). *Circular statistics in biology*. New York: Academic Press.

Gärling, T., Böök, A., & Lindberg, E. (1984). Cognitive mapping of large-scale environments: The interrelationship of action plans, acquisition, and orientation. *Environment and Behavior*, 16, 3–34.

Golledge, R. G. (1999). Human wayfinding and cognitive maps. In R. G. Golledge (Ed.), *Wayfinding behavior: Cognitive mapping and other spatial processes* (pp. 5–45). Baltimore, Maryland: Johns Hopkins University Press.

Ishikawa, T., & Montello, D. R. (2006). Spatial knowledge acquisition from direct experience in the environment: Individual differences in the development of metric knowledge and the integration of separately learned places. *Cognitive Psychology*, 52, 93–129.

Kuipers, B. (1978). Modeling spatial knowledge. *Cognitive Science*, 2, 129–153.

Meilinger, T. (2008). The network of reference frames theory: A synthesis of graphs and cognitive maps. In C. Freksa, N. S. Newcombe, P. Gärdenfors, & S. Wölfl (Ed.), *Spatial Cognition VI* (pp. 344–360). Berlin: Springer.

Moeser, S. D. (1988). Cognitive mapping in a complex building. *Environment and Behavior*, 20, 21–49.

Montello, D. R. (1998). A new framework for understanding the acquisition of spatial knowledge in large-scale environments. In M. J. Egenhofer, & R. G. Golledge (Ed.), *Spatial and temporal reasoning in geographic information systems* (pp. 143–154). New York: Oxford University Press.

Poucet, B. (1993). Spatial cognitive maps in animals: New hypotheses on their structure and neural mechanisms. *Psychological Review*, 100, 163–182.

Rossano, M. J., & Reardon, W. P. (1999). Goal specificity and the acquisition of survey knowledge. *Environment and Behavior*, 31, 395–412.

Sholl, M. J. (1996). From visual information to cognitive maps. In J. Potugali (Ed.), *The construction of cognitive maps* (pp. 157–186). Netherlands: Kluwer Academic Publishers.

Sholl, M. J., & Nolin, T. L. (1997). Orientation specificity in representations of place. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 23, 1494–1507.

Siegel, A. W., & White, S. H. (1975). The development of spatial representations of large-scale environments. In H. W. Reese (Ed.), *Advances in child development and behavior vol. 10* (pp. 9–55). New York: Academic Press.

Spiers, H. J., & Maguire, E. A. (2008). The dynamic nature of cognition during wayfinding. *Journal of Environmental Psychology*, 28, 232–249.

Thorndyke, P. W., & Hayes-Roth, B. (1982). Differences in spatial knowledge acquired from maps and navigation. *Cognitive Psychology*, 14, 560–589.

Trullier, O., Wiener, S. I., Berthoz, A., & Meyer, J. (1997). Biologically based artificial navigation systems: Review and prospects. *Progress in Neurobiology*, 51, 483–544.

Appendix A

Experiment 1 Labyrinth 1 was set in an imaginary town where a cat and four residents lived: an old lady, an elementary school girl, a vegetable shop owner, and a middle-aged lady. The last resident found by the participants in the exploring phase was determined to be the cat owner. Participants were told by the owner that their cat was missing and were asked to revisit the other residents and get information about the cat.

Experiment 2 Labyrinth 2 was set in an imaginary rural town in Asia where four residents lived: a village headman, a Buddhist monk, an elephant driver, and an old lady. Participants were told by the last resident found that a hidden gem had been stolen by a monkey. Then, the resident asked them to revisit the other residents and get information about the monkey.

The Direction Bias and the Incremental Construction of Survey Knowledge

Tobias Meilinger (tobias.meilinger@tuebingen.mpg.de)
Max Planck Institute for Biological Cybernetics
Spemannstr. 38, 72076 Tübingen, Germany

Heinrich H. Bülthoff (heinrich.buelthoff@tuebingen.mpg.de)
Max Planck Institute for Biological Cybernetics
Spemannstr. 38, 72076 Tübingen, Germany
Department of Brain and Cognitive Engineering, Korea University,
Anam-dong, Seongbuk-gu, Seoul, 136-713 Korea

Abstract

This study examines how spatial memory acquired from navigation is used to perform a survey task involving pointing. Participants learned a route through a virtual city by walking it multiple times in one direction on an omnidirectional treadmill. After learning, they were teleported to several locations along the route, self-localized and pointed to multiple other locations along the route. Pointing was done away from or towards the current location. Preliminary data show that participants were faster in pointing away. This suggests that pointing was based on an incremental process rather than an all-at-once process which is consistent with mentally walking through a cognitive map or constructing a mental model of currently non-visible areas of the city. On average participants pointed faster to targets located further down the route towards the end than to targets located route upwards towards the start. Analysis of individual performance showed that more participants than expected by chance showed such an effect of target direction also in their pointing accuracy. The direction of this effect differed between participants. These direction biases suggest that at least some participants encoded the environmental space by multiple interconnected locations and used this representation also for pointing.

Keywords: Reference frame; environmental space; spatial memory; survey knowledge; cognitive map; mental walk; mental model; pointing; virtual environment

Introduction

When navigating through an environmental space such as a city or a building we experience multiple views of parts of this environment from various perspectives (Montello, 1993). The knowledge acquired from these experiences can be used to retrace familiar routes, plan novel routes, point to distant locations, or look for shortcuts. The last two tasks are examples of survey tasks (Ishikawa & Montello, 2006; Siegel & White, 1975). To solve a survey task, one has to consider metric relations (distance, relative direction) between two locations not mutually visible. Often, these two locations are our current location and a target location we want to point to, estimate the distance to, or find a shortcut towards. In order to do so at least our current location and the target location have to be represented within a single reference frame. This could be our egocentric reference frame within which the direction and distance of the target is represented in relation to our body. It could also be an allocentric world-centered reference frame within which our current location and the target are represented. Unless we obtain our environmental knowledge from a map which already provides this information within a single reference frame we have to integrate the multiple pieces of information acquired during navigation to represent them within one single reference frame. This work aims to cast some light on how this integration process might work. We will introduce theories of survey knowledge, derive predictions from these theories, and test them in an experiment.

Theories of Survey Knowledge

Most spatial memory theories which explain survey knowledge assume that navigators form a global world-centered reference frame within which all relevant locations are represented. Such a global reference frame might be formed very quickly, with all novel locations represented within it (Mou, McNamara, Valiquette & Rump, 2004; O'Keefe, 1991; Stachniss, 2009). Alternatively, this global reference frame is eventually formed from multiple local representations (Kuipers, 2000; Mallot & Basten, 2009; McNamara, Sluzenski & Rump, 2008; Poucet, 1993; Trullier, Wiener, Berthoz & Meyer, 1997). It then either works as an additional layer embedding local representations within a metric reference frame or as the top-level in a hierarchical memory structure subsuming lower level reference frames. In the following a global world-centered reference frame will be called a cognitive map.

Survey relations can be obtained from a cognitive map in several ways. The easiest way is to simply *read out* the coordinates of the relevant locations (e.g., the current location and the target location) and compute the relative direction, the distance between the locations, etc., by subtracting these coordinates from each other. If required by the task these parameters are then transformed into an egocentric reference frame, for example, when pointing to a target.

Alternatively, navigators could *mentally walk* through a cognitive map. While mentally moving from one point to another, they integrate the metric survey relation between the start and the mental position in the map until reaching the target (Byrne, Becker & Burgess, 2007). Thus the relative direction, the distance, etc. are derived. The activation pattern of hippocampal place cells is a plausible mediator for this process – although the conscious imagery of the mental walk might take place in posterior parietal cortex.

Place cells represent locations within an environment (O'Keefe & Nadel, 1978). Even in the absence of sensory stimulation (e.g., during sleep) they can fire in an ordered fashion as they would do when walking a route (Skaggs & McNaughton, 1996). Similar neural processes might happen during mental walks when performing a survey task.

A different position assumes that an environmental space is not represented within a single global reference frame (i.e., a cognitive map), but by multiple local interconnected reference frames (Meilinger, 2008). The integration within a single reference frame which is required for survey tasks happens during retrieval by constructing a *mental model* of the non-visible environment (a related model was presented for updating by Fujita, Klatzky, Loomis & Golledge, 1993). For example, navigators imagine what the environment would look like if the surrounding walls were transparent. First, they imagine the adjacent street from their current position, then they add the street branching off from it, etc. In this way all locations from the current location along a route leading towards the target location are imagined step-by-step within the current egocentric reference frame (this could also be done from a different imagined viewpoint). The mental model of the non-visible environment is constructed piecewise from a certain perspective. No one mentally walks through this constructed environment and the underlying memory structure is no cognitive map, but a network of reference frames interconnected by directed links (i.e., the links point in certain direction). The construction of the mental model is assumed to be easier when done along the direction of the links (i.e., imagine a distant location the link point towards). Otherwise these links have to be inverted which is computationally costly.

The Prediction of Performance Differences

The three positions, read out from a cognitive map, mentally walking through a cognitive map, and constructing a mental model from a network of reference frames predict specific performance differences due to incremental vs. all-at-once process of deriving survey relations and due to direction biases in the underlying memory.

All-at-once vs. Incremental Estimation of Survey Relations Reading out coordinates of two locations from a cognitive map and subtracting them is an all-at-once process in the sense that the survey relation (e.g., the relative direction of the target from a current location) is determined as whole. Contrary, mentally walking to a target or extending a mental model of the environment until it includes the target are incremental processes. The further we walk and the further the model is constructed the better we can estimate the direction and distance towards our target. Due to the incremental character locations in-between have to be represented during this process. This is not the case for reading out. One way to test this is to have navigators point to multiple locations in an ordered way. For example, they point to all locations along a route from the current location to a location B. When they do so in an order away from the current location (i.e., first point to the adjacent location, then the second closest, etc., until finally pointing to B) they can mentally walk or construct a model up to the first location, point there, extend this model or mentally walk to the second location, point there, etc., until mentally reaching location B. In the opposite case when they point in an order towards the current location (i.e., first to location B, then the second last location until finally pointing to the location closest to the current location) they first have to construct the whole model up to B, respectively mentally walk the entire distance up to location B. Then they either shift their attention to the second last target in the model, mentally walk back to the second last target or do it all over again from the current location to the second last location. No matter how navigators precisely do this, this process should last longer and/or be more error prone than pointing to targets in an order away from the current location. When reading out locations from a cognitive map navigators cannot profit from their last pointing. They have to compute the survey relation for each target individually no matter in which order they point to the locations. Order thus should not lead to different performance as in the case of a mental walk or a mental model.

The Direction Bias A cognitive map does not show direction specificity between locations (although the whole map might be oriented in a certain way such as north-up in a paper map). That means that no matter whether one points from A to B or from B to A the result should not differ in performance. This is just the same for reading out as well as for mental walk. On the contrary, a direction bias is expected in certain cases for the mental model explanation, because of the underlying memory. The mental model is based on directed interconnections between local reference frames. Constructing a model in the direction of the interconnection is easier as no inversion is required. It should yield better performance.

In order to predict performance differences one has to know where the directedness in memory originates from. According to Meilinger (2008) navigators encode local reference frames during navigation (e.g., corresponding to a street or a room). The interconnections between these local reference frames represent the metric relations (i.e., relative direction, distance, and orientation) between them. They can be derived in at least two ways. First, navigators might obtain interconnections from their visual input. They see that a street branches off to the right in 20 meters. The reference frame of this street is located 20 meters to the front and is oriented 90° to the right. This results in a forward connection, for example, expressed by vector pointing forwards. Alternatively, they could walk up to the next street while updating the origin of their current street (i.e., the origin of the memory reference frame representing the street). The interconnection to the last reference frame is the updated vector pointing back to the last street (i.e., a backwards interconnection). Here an individual navigator is expected to apply only one kind of strategy (i.e., either forward or



Figure 1: The virtual city as seen from navigation perspective (left side) and from bird's eye view with the route marked in red (right side). During learning the start, the end and each of the six intersections in-between were marked with white crosses on the floor. They worked as pointing locations and targets during the test phase.

backward encoding), at least over some time interval such as an experiment. Thus walking a route in one direction will result in directed interconnections (either forwards or backwards). Using these interconnections for constructing a mental model is easier along the direction of interconnections and should lead to better pointing performance. Depending on the encoding strategy this direction bias should be in forward or backwards direction.

Methods

For the experiment we used an immersive virtual city environment presented via a head-mounted display (HMD). In the learning phase, participants experienced the virtual environment by walking through it on an omnidirectional treadmill. They only walked the route in one direction. In the testing phase, participants were teleported to different locations in the environment, without walking physically. They were then asked to identify their location and heading and were instructed to point towards multiple targets on the route. Pointing order could be either towards their current location or away from it. Direction biases were examined by comparing pointing performance for pointing to targets located route upwards (to the start) with pointing performance for targets route downwards (to the end).

Participants

So far eleven participants (5 females and 6 males) aged between 21 and 34 (M = 26.6 years, SD = 4.5 years) participated in the experiment. They were recruited via a subject database and were paid for their participation. All participants signed an informed consent approved by an ethical committee before participating in the experiment.

Material

The Virtual City In the learning phase, participants had to learn a route through a virtual city. Figure 1 shows a snapshot of the city as seen during walking, as well as a bird's eye view of the route. The route consisted of a start, six intersections and an end. During learning, all eight locations were marked with a white X on the floor, as were all intersections visible from this route. The type of houses changed along the route, as did street width and the heights of houses. In addition, individual houses ensured sufficient landmark information to identify each location.

The Setup Participants walked on a 4x4 meters omnidirectional treadmill (Figure 2 left side). It allowed them to walk for infinite distances in any direction by moving them back to the centre of the treadmill. This unique interface allows for realistic proprioceptive and vestibular feedback as well as efference copies while walking in virtual environments. Participants wore a climbing harness for the unlikely event of falling and hurting themselves on the moving platform. To obtain participants' location on the treadmill their head position was tracked by 16 high-speed motion capture cameras at 120 Hz (Vicon® MX 13). This data was used both to control the treadmill and to update the visualization of the virtual environment. The visual surrounding at a location was rendered in real time (60Hz) using a NVIDIA Quadro FX 4600 graphics card with 768 MB RAM in a standard PC. Cables connected the PC to the display via the ceiling. Participants viewed the scene in stereo using a nVisor SX head-mounted display that provided a field of view of 44×35 degrees at a resolution of 1280×1024 pixels for each eye with 100% overlap. The setup thus also provided important visual depth cues such as stereo images and motion parallax.



Figure 2: The virtual reality setup. The left image depicts a participant walking on the omnidirectional treadmill during the learning phase. The right image shows a participant pointing to a target during the testing phase by facing the target and pressing a button on a gamepad.

Procedure

In the *learning phase*, participants walked the route at least six times from start to end. They were instructed to first learn the route, and secondly be able to self-localize when teleported to an X along the route after the learning phase. Participants were free to look around as long as they wanted. In their first run, they walked up to an intersection, looked around, and the experimenter pointed out the street to take when the participant looked down the correct street

by stating "the route is this direction" (the experimenter was in the same room and could task with the participant). No verbal turning information (e.g., "left", "straight on", etc.) was given. When reaching the end and having looked around participants were teleported back to the start. From the second run onwards participants were asked to approach an intersection, look into the direction the route was going on and say "this way". The experimenter gave feedback whether this was right or wrong, before participants proceeded. They were not allowed to leave the route. For each new run, the virtual environment was rotated 90° clockwise relative to the lab. Sound sources within the lab could thus not be used to derive global orientation. The learning phase ended when participants walked the route at least six times and at least two runs were error-free. This criterion ensured comparable levels of route knowledge for all participants. Participants briefly trained walking on the treadmill before starting the experiment.

In the following *test phase*, participants were teleported to locations on the route formerly marked by an X (i.e., the start, the end or one of the six intersections in between). They were now asked to self-localize and then successively point to multiple target locations which had all been formerly marked by an X. For self-localization, participants could look and rotate around, but not walk around – a circular handrail around them with 0.48 meters diameter prevented them leaving their location during the test phase (Figure 2 right side). As soon as they subjectively knew their location and orientation, they were asked to press a button on a gamepad they were holding. Then they pointed to multiple targets. Pointing was done by turning on the spot until a vertical black line in the middle of the display matched the direction in which the participant thought the target was located. They thus would look directly at the target location if the surrounding houses were transparent. When participants thought they faced the target, they pressed a button and then pointed to the next target. No feedback was provided. After they had pointed to all targets, participants pressed a second button on the gamepad and were teleported to a new position.

Four conditions determined the targets and the order in which participants were asked to point towards them (Table 1). They should point either (1) first to the start and then to all locations between start and the current location in the order of walking (i.e, start, 1st intersection, 2nd intersection, etc.). (2) They should point to the same locations, but in reverse order (i.e., first the intersection before the current location, then the second last, etc. until finally pointing to the start). (3) They should point to the next intersection along the route after the current location, then the second next, etc. until pointing to the end. Or they should (4) point first to the end, then the last intersection, the second last intersection, etc. until pointing to the intersection after the current location. Consequently, we varied the two factors 'target direction' (route upwards to start vs. route downwards to end) and 'pointing order' (away vs. towards the current location; see Table 1). Please note that the adjacent intersections to point towards were always visible during pointing (although the Xs were removed). From the eight locations on the route (including start and end) participants pointed to every other location twice (in the order away and towards the current location). All 28 pointing sets were presented in random order for each participant (pointing downwards from seven locations, upwards from seven locations, both in two orders). This whole procedure was repeated resulting in 56 pointing sets altogether. After finishing a pointing set participants received feedback about the number of pointing targets they pointed towards: whether they pointed towards the right number of targets, how many targets they missed; or how many superfluous targets they pointed towards. No feedback about pointing accuracy was provided. Pointing sets with too few or too many pointings were not analyzed as the target locations could not be assigned to pointings. We recorded self-localization time (not further reported), pointing time and pointing direction for each pointing in a complete pointing set. After pointing participants drew a sketch map and we asked for subjective strategies with a questionnaire. The whole experiment lasted approximately two hours.

For the analysis we used pointing time and computed the absolute pointing error (i.e., the deviation between correct and estimated pointing direction irrespective of the direction of the error). Values deviating more than two standard deviations from a participant's mean were not analyzed. Only if a participant's mean absolute pointing error significantly differed from 90°, indicating that some survey knowledge was acquired, data were analyzed (90° error is the average error you get when randomly pointing in all directions). For analyses within participants we used t-tests. For analyses across participants we computed mean values per participant and condition and used within-participants ANOVA and t-tests. Cohens *d* and partial eta-square (η_p^2) are presented for the estimation of effect sizes.

Table 1: The Four Pointing Conditions

Target direction on the route	*Pointing order (relative to the current location)*	*Instruction: Point from…*
Upwards	Away	current location to start
Upwards	Towards	start to current location
Downwards	Away	current location to end
Downwards	Towards	end to current location

Results

For all but one participant pointing accuracy differed significantly from chance (t's > 10.9, p's < .001). They did acquire survey knowledge and were thus further analyzed. Their average absolute pointing error was 19.6°; mean pointing time was 2.8 seconds per pointing.

Mental walk and mental model theories of survey knowledge predicted performance differences for pointing order.

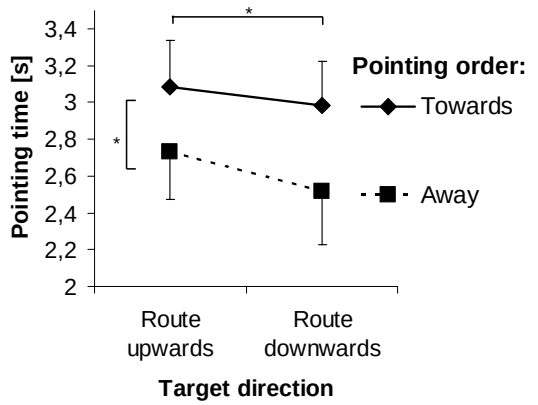


Figure 3: Average pointing time as a function of target direction and pointing order. Both main effects were significant as indicated by the asterisks. Means and (between participants) standard errors are displayed.

Indeed, participants pointed faster away (M = 2.6s) than towards the current location (M 3.0s; see Figure 3; $F(1, 9) = 9.50$, $p = .013$, $\eta_p^2 = .51$; accuracy: towards $M = 22°$, away $M = 18°$, $F(1, 9) = 2.21$, $p = .171$, $\eta_p^2 = .20$). This difference was not predicted by a process of reading out from a cognitive map.

According to the mental model of survey knowledge, participants' performance should differ as a function of target direction – although the direction of the effect might differ between participants. When averaging across all participants they pointed faster to targets located further down the route to the end (M = 2.7s) than to upward targets (*M = 2.9s;* $F(1, 9) = 8.22$, $p = .019$, $\eta_p^2 = .48$; accuracy: upwards $M = 17°$, downwards $M = 23°$, $F(1, 9) = 2.21$, $p = .151$, $\eta_p^2 = .22$). Looking at the effect of target direction on pointing accuracy for each participant individually a more differentiated picture emerges: Six out of the ten participants showed an effect of target direction in their pointing accuracy (i.e., their pointing accuracy differed between pointings upwards to the start vs. downwards to the end t's > 1.99, p's < .049, d's > 0.19). Three of them pointed more accurately downwards the route (*$M = 7.0°$ vs. $M = 10.5°$)*, three participants pointed more accurately upwards (*$M = 23.6°$ vs. $M = 45.1°$)*. Four participants did not show a significant effect of target direction (upwards $M = 16.2°$, downwards $M =$,18.7°; t's < 1.88, p's > .063, d's < 0.21). If there was no target direction effect each participant has a chance of 5% to (erroneously) be classified as being direction biased in any direction by a t-test. The observed proportion of 6 out of 10 participants showing such an effect is highly unlikely to originate from such a 5% chance rate (binomial test $N = 10$, $\pi = 5\%$: $p < .001$). Consequently, the null-hypothesis that there is no effect of target direction on accuracy is rejected. Since individual participants showed opposite direction biases, we observed no average global bias in pointing accuracy in one specific direction. When looking at differences in pointing time on the level of individual participants only one participant significantly pointed faster downwards the route ($t(207) = 2.29$, $p = .023$, $d = 0.22$). This proportion (one out of 10) does not significantly differ from a 5% chance rate (binomial test $N = 10$, $\pi = 5\%$: $p = .401$).

We found no effect of pointing in walking order which is expressed by the interaction between target direction and pointing order (time and accuracy both $F(1, 9) < 1$). Pointing to multiple targets in walking order (i.e., from start to current location or from the current location to the end) did not differ significantly from pointing in opposite walking order (i.e., from end to current location or from current location to start).

Discussion

The present study examined predictions from three different theories about how survey relations are derived from spatial memory. The three positions (read out from a cognitive map, mentally walking through a cognitive map, and constructing a mental model from a network of reference frames) predict specific performance differences for target directions and pointing order.

We found an effect of pointing order. Participants pointed faster to targets in the order away from the current location than towards the current location. This result suggests that pointing is based on an incremental rather than an all-at-once process. Navigators might mentally walk through a cognitive map and integrate the walked distance (Byrne et al., 2007) or they could stepwise construct a mental model of the non-visible environment until this model includes the target (Meilinger, 2008).

There was also an effect of target direction. On average, participants pointed faster to targets further down the route, than to targets route upwards to the start. When looking at target direction effects for each individual, more participants than expected by chance showed a significant effect of target direction in their pointing accuracy. Half of these pointed more accurately towards locations further down the road, the other half pointed more accurately towards targets upwards the route. These results in pointing accuracy suggest different strategies in the encoding of an environment. Some participants might have encoded multiple local environments (e.g., rooms, streets, etc.), updated the last environment while walking to the next environment and stored the updated vector pointing backwards to the last environment. Deriving survey relations from this string of backwards connected locations should be easier in a backwards direction. For locations route downwards the connection would have to be inverted which is an additional process and thus an additional source of errors. Another group of participants seems to have encoded multiple local environments in the opposite direction (i.e., in the direction they walked the route). They could have derived the interconnections from their visual input: they saw how the route was going on (e.g., 30 meters straight on, then turn to the right) and used this information for connecting encoded locations, thus resulting in a forward connection. For them, constructing a mental model in forward direction did not involve in-

version of the interconnection and thus resulted in more accurate pointing. The third group of participants did not show a significant effect of target direction on the level of the individual. They might have formed a cognitive map and used this representation for pointing (likely by mental walk). Alternatively, their orientation bias was not strong enough to reach the significance level. The time advantage for pointing route downwards when averaging across participants might simply be an effect of averaging across the groups and could suggest that forward encoding was more likely than backward encoding.

The results reported here were found in a virtual reality setup. Therefore, we cannot exclude the possibility that participants might behave differently in real environments. However, the present setup provided most of the bodily and visual cues also available when walking through a real environment (proprioceptive feedback, efference copy, vestibular stimulation, motion parallax, stereo vision, texture gradient, familiar size cues, etc.). Also, on average pointings were quite precise. A generalization to real environments does, thus, not seem implausible.

One major limitation is the small sample size. More participants are needed to see whether the effects observed are really stable. With more participants we will also be able to directly compare the different subgroups in target direction and have a closer look at their strategies.

This study examined how navigators derive survey relations used for pointing or short cutting from memory of an environmental space which they have to navigate through in order to experience it. Our results suggest that pointing is based on an incremental process as predicted by mentally walking a cognitive map or by constructing a mental model of the non-visual environment. At least for some participants we found indications for a direction specific encoding of such an environment (i.e., a string of location representations connected via directed links). Their pattern of performance is consistent with a mental model construction based on such a memory. Future experiments will have to clarify the exact circumstances which yield which kind of memory.

Acknowledgments

This research was supported by the DFG grant "The functional, computational and neural basis of human survey knowledge – comparing mental maps and mental graphs", the Max Planck Society and by the WCU (World Class University) program through the National Research Foundation of Korea funded by the Ministry of Education, Science and Technology (R31-2008-000-10008-0). The authors thank Jan Souman for help in planning the experiment, discussing the results and writing the paper, Nadine Simon for help in data collection, Joachim Tesch, Michael Kerger, and Harald Teufel for intensive technical support, as well as Johanna Steffen for proof reading.

References

Byrne, P., Becker, S. & Burgess, N. (2007). Remembering the past and imagining the future: a neural model of spatial memory and imagery. *Psychological Review, 114,* 340-375.

Fujita, N., Klatzky, R. L., Loomis, J. M., & Golledge, R. G. (1993). The encoding-error model of pathway completion without vision. *Geographical Analysis, 25,* 295-314.

Ishikawa, T. & Montello, D.R. (2006). Spatial knowledge acquisition from direct experience in the environment: individual differences in the development of metric knowledge and the integration of separately learned places. *Cognitive Psychology, 52,* 93-129.

Kuipers, B. (2000). The spatial semantic hierarchy. *Artificial Intelligence, 119,* 191-233.

Mallot, H.A., & Basten, K. (2009). Embodied spatial cognition: biological and artificial systems. *Image and Vision Computing, 27,* 1658-1670.

Meilinger, T. (2008). The network of reference frames theory: a synthesis of graphs and cognitive maps. In C. Freksa, N.S. Newcombe, P. Gärdenfors, & S. Wölfl (Eds.), *Spatial Cognition VI* (pp.344-360). Berlin: Springer.

McNamara, T.P., Sluzenski, J. & Rump, B. (2008). *Human Spatial Memory and Navigation.* In H.L. Roediger, III (Ed.), Cognitive Psychology of Memory. Vol. 2 of Learning and Memory: A Comprehensive Reference (pp.157-178). Oxford: Elsevier.

Montello, D. R. (1993). Scale and multiple psychologies of space. In A.U. Frank & I. Campari (Eds.), *Spatial information theory: A theoretical basis for GIS* (pp. 312-321). Berlin: Springer.

Mou, W., McNamara, T.P., Valiquette, C.M. & Rump, B. (2004). Allocentric and egocentric updating of spatial memories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30,* 142-157.

O'Keefe (1991). An allocentric spatial model for the hippocampal cognitive map. *Hippocampus, 1,* 230-235.

O'Keefe, J., & L. Nadel (1978). *The hippocampus as a cognitive map.* Oxford: Clarendon Press.

Poucet, B. (1993). Spatial cognitive maps in animals: Newhypotheses on their structure and neural mechanisms. *Psychological Review, 100,* 163-182.

Siegel, A. W., & White, S. H. (1975). The development of spatial representations of large-scale environments. In H. Reese, (Ed.), *Advances in Child Development and Behavior, Vol. 10,* (pp 10–55). New York: Academic Press.

Skaggs, W.E. & McNaughton, B.L. (1996). Replay of neuronal firing sequences in rat hippocampus during sleep following spatial experience. *Science, 271*, 1870-1871.

Stachniss, C. (2009). *Robot Mapping and Exloration.* Berlin: Springer.

Trullier, O., Wiener, S.I., Berthoz, A. & Meyer, J.-A. (1997). Biologically based artificial navigation systems: Review and prospects. *Progress in Neurobiology, 51,* 483-544.

Alignment of Spatial Perspective

Elena Andonova (andonova@uni-bremen.de)
SFB/TR8, University of Bremen, Cartesium, 7 Enrique-Schmidt Str.,
Bremen, Germany

Abstract

Most of the experimental research on dialogue that has provided evidence for interactive alignment focuses on speakers aligning at the lexical and syntactic levels of representations and dialogic contributions, i.e., having converging choices of lexical and syntactic means of referring to pictured objects and events. Less is known about alignment at the conceptual level, or situation models. This paper addresses alignment in spatial perspective (route vs. survey perspective) between speakers in a confederate experimental task taking turns in describing routes on schematic maps. The findings of two experiments show that speakers' spatial perspective choices are aligned with those of their partners both before and after partners switch perspective. Furthermore, this alignment effect holds both if partners show consistency adhering to the same perspective for a sequence of descriptions and when they display inconsistency by switching spatial perspective for every new description they provide.

Keywords: spatial perspective; interactive alignment.

Introduction

Imagine asking someone on the phone for directions on how to go some place while looking at a simple map. Now imagine being told to go 'left' while your current orientation is facing 'downward' on the map. This is potentially a problem because of the ambiguity inherent in this term. It is unclear if *left* is mapped onto your perspective and orientation at a given time as situated in the environment or to be interpreted from the external viewpoint of looking at the map as if from above. Now imagine further that it is your turn to make a suggestion for a route to the person on the phone. How likely are you to use the same perspective your partner used just now vs. another? The inherent ambiguity of terms such as *left* and *right* when produced and understood within different perspectives and frames of reference is an excellent testing ground for frameworks of interaction and coordination in a dialogic situation.

The interactive alignment model (Pickering & Garrod, 2004) posits that much of speech production choices in dialogic situations can be explained via an automatic mechanism involving priming at multiple levels of linguistic representation and percolation between these levels. Furthermore, alignment of situation models is achieved on the basis of such lower-level alignment of representations. While the model also allows for alignment via explicit reasoning and modeling of the partner's mental states and mental model updating, it places a particular emphasis on these low-level mechanisms. Alternative accounts of dialogue behavior question the explanatory power of automatic priming in dialogic convergence and underline the role of (explicit) modeling of partners and their mental states of representation. Common conversational ground is the outcome of a joint effort on behalf of interlocutors who attend to the degree to which information is mutually shared (Clark, 1996).

Research on dialogue has addressed how speakers deal with variability and ambiguity in order to achieve alignment of situation models. One and the same object or event can trigger multiple perceptual and conceptual representations. For example, in a study of goal-directed dialogue (Garrod & Anderson, 1987), different description schemes were used by speakers referring to a maze and movement in it (path, coordinate, line, figural schemes). Similarly, in a study of how people describe complex multiple-object scenes, speakers' choices varied significantly depending on the nature of the array (Andonova, Coventry, & Tenbrink, in press).

Multiple perspectives, or ways of speaking about the world, are reflected on different levels of language but also in variation at a conceptual level. In spatial reference, different conceptualizations can be found in the choices of spatial perspective and frames of reference. In particular, perspective taking involves abstracting from the visual scene and organizing and packaging information in accordance with one or another type of viewpoint. Spatial perspective varieties can be characterized in different ways. Here we will adopt a binary distinction which is a simplified yet common typology. A route or environment can be described from an embedded (route or egocentric) perspective, that is, from within the environment, based on the way-finder, as embedded in the path, or from an external (survey or allocentric) perspective, that is, a viewpoint external to the environment, commonly associated with maps and cardinal directions, the way people would look at a map or a drawing of a route. For the sake of brevity and simplicity, here we will refer to these as the route perspective and the survey perspective. Previous studies have demonstrated that a number of individual, environmental, and learning factors are sources of variation in spatial perspective in verbal descriptions. Mode of acquisition has been shown to affect perspective choices in spatial memory; for example, participants who studied maps gave more accurate responses later in survey perspective tasks whereas those who were navigating gave more accurate responses to route perspective tasks (Taylor, Naylor, & Chechile, 1999). Taylor & Tversky (1996) tested

the influence of four environmental features on spatial perspective choices and found that although overall most participants' descriptions followed a survey or a mixed perspective, preference for the use of route perspective was enhanced in environments that contained a single path vs. multiple paths and environments that contained landmarks of a single size scale vs. landmarks of varying size. Bugman, Coventry, and Newstead (2007) found that context of retrieval (frequency of visitation vs. importance of activities) can affect spatial perspective choices, too.

Variability in spatial perspective choices is frequently accompanied with perspective switching behavior– participants tend to mix perspectives quite regularly, for example, 27 out of 67 participants in Taylor & Tversky's (1996) first experiment and 74 out of 192 participants in their second experiment mixed perspectives in their descriptions. There are multiple reasons why a speaker may switch from one perspective to another, for example, because of some features of the environment or the task. However, although most studies have researched spatial perspective choices in a monologue setting, one important reason for initial perspective choice and subsequent switches may be the behavior of the interlocutor (conversation partner) in a typical dialogue setting of giving road instructions, for example. Two exceptions to the dominant monologue settings of spatial perspective research are a study by Schober (1993) which showed that speakers set spatial perspectives differently with actual addressees than with imaginary ones and another by Striegnitz, Tepper, Lovett, & Cassel (2008) in which there was an increased use of survey perspective in response to clarification questions and in re-phrasal of previously given route descriptions.

The variability of spatial perspective and perspective switching make this phenomenon a suitable testing ground on coordination of speakers' choices in dialogue. Thus, two strands of research and related questions are in the combined focus of this paper—spatial perspective use and interactive alignment.

When dialogue partners refer to the same scene, they select a frame of reference or a perspective for the description. Thus, in dialogue, perspective use and perspective switching are part of the overall process of coordination. Does choice of perspective depend then on the previous use or preference for a certain perspective shown by one's dialogue partner, i.e., do speakers align in their choices of a spatial descriptive schema? If so, to what extent can this influence be modulated by the degree of consistency of partners' choices? Furthermore, how flexible is this process of coordination and perspective choice? Does the first 'conceptual pact' one strikes implicitly with one's partner remain dominant throughout an interaction, or alternatively, if the partner switches perspective, is one more likely to adhere to the previously used perspective, or to switch along, and re-align?

In the studies presented here, there were two clearly possible perspectives on the scene and route to be described: survey and route perspective. Route perspective is by far the more natural way to describe routes whereas survey perspective is more typical of location descriptions. In order to enhance the probability of use of survey perspective and to bring the two more into balance, the maps to be described were positioned vertically, which also corresponds to viewing maps on a screen.

In the first experiment, we ask first whether speakers align choices of spatial perspective when their partner follows one perspective consistently in a short sequence of descriptions (four maps with routes). We also ask whether spatial perspective alignment persists even when the partner switches perspectives and offers a subsequent series of descriptions in an alternative perspective.

Experiment 1

As stated above, this experiment was designed to examine two related questions. First, whether speakers align on spatial perspective, and second, if they continue to align with their partners even when their partners switch perspective between an early and a later experimental block. If speakers rely only on a general model of partner preferences built on the basis of their experience during the early block, then perspective switch by the confederate would not reverse speakers' choices in accordance with the new spatial perspective bias exhibited in the later block. If speakers are sensitive not only to initial partner preferences but they also update their model of their partner (after the switch), then they should also show a tendency to switch perspective in a similar way. A third possibility also exists—the fact that their partners have used both route and survey perspectives and that they switched between them may reduce speakers' preferences for either perspective and lead them to choose between perspectives more or less randomly.

Method

The design of the experiment included prime perspective (route vs. survey) and experimental block (early vs. later) as independent variables and mean percent choice of route perspective on each experimental block as the dependent variable.

Participants 24 participants (3 male) took part in the experiment. They were university students with a mean age of 21.08 years (range 19 – 31) who received course credit or were paid for their participation. All were native German speakers.

Stimuli Thirty-two simplified map drawings were used in the study. Six different maps were created and a total of 16 different routes. Stimuli were pseudo-randomized with the constraints that no two consecutive maps should be the same, and neither the start nor the end points of the routes on consecutive maps should be the same. Routes were pre-drawn on the maps so as to exclude a route planning component in the task and focus exclusively on choice of spatial perspective (see Fig.1 for an example). There were

16 experimental trials (8 prime-target pairs) and 16 fillers. The maps and routes on the experimental prime-target trials were designed to be compatible with both route and survey perspective descriptions. Confederates' descriptions of routes were either in a route perspective or a survey perspective. Filler maps and routes were drawn in such a way as to minimize the use of spatial perspective, for example, a circular trajectory. Furthermore, confederates' scripted descriptions on these trials did not contain any indication of spatial perspective.

Each experimental prime-target pair was preceded by two filler items. There were two blocks of experimental pairs, an early and a later one. In accordance with the design of the experiment, the perspective of the confederate primes was consistent within each block and was either route or survey. However, confederates' scripted descriptions on the two blocks differed in spatial perspective, i.e., the confederate switched perspectives between the early and the later block of trials.

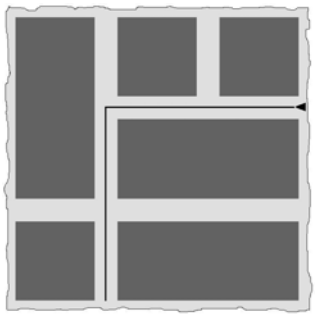


Figure 1: Example of a map and its pre-drawn route. The triangle indicates the position and orientation at the start.

Procedure Each participant was seated across a desk from the confederate and a visual barrier was placed between them and a stack of cards with identical maps and routes drawn were placed in front of them on a vertical stand. Cards were positioned vertically to motivate more the use of the generally weaker and less common survey/gaze perspective. In addition, the confederate used a list of pre-scripted descriptions matching their cards in either route or survey perspective. The scripted responses of the confederate were not visible. We took special care to minimize possible suspicions on behalf of the participants that their partner in the experiment may not be a naïve participant such as they were, including greetings, familiarization procedures, instructions, etc. Confederates were student assistants of the same age and population generally who were trained to act naïve. Participants and confederates took turns in describing the routes on these cards. A red and a green dot marked on the back of each card were used to indicate whose turn it was to speak. Confederates were the first to speak, thus ensuring that their utterances (primes) precede those of the participants on target trials. Participants were instructed to monitor the descriptions of their partner for accuracy and to offer a correction whenever they noted an incorrect description. Three deliberate errors were built into the script on filler items. This instruction ensured that participants were attending to their partners' descriptions and choice of perspective. At the end of the experimental session, participants filled out a questionnaire which included questions asking participants to say what they thought the experiment was about and what they thought about their partner's behavior. As nobody indicated any suspicions that their partner may not have been a naïve participant such as they were, the data of all participants were accepted for analysis.

This procedural setup is a close replica of the procedure in Branigan, Pickering, McLean & Cleland (2007) which studied the effects of participant role on syntactic alignment.

Results

The pre-analysis procedure was identical for the data in Experiment 1 and Experiment 2 and will be described here jointly. Participants' responses were classified according to the spatial perspective used as belonging to one of three categories: route perspective, survey perspective, or mixed perspective. Experimental pairs on which the confederate made a mistake (1.99%) were excluded from the analysis, as well as those when the participant offered a correction to their partner's description of an experimental item (20.39%). Route perspective was the preferred default option and in the majority of these cases participants offered a 'correction' of the confederate's survey perspective description into a route perspective one. The following are examples of participant responses to the map and route in Fig.1 coded as route perspective (a), survey perspective (b), and mixed perspective (c) in their original German and in translation:

(a) *hier gehst du geradeaus und biegst dann links ab*
E. here you go straight and then turn left

(b) *hier gehst du erst nach äh links und dann nach unten*
E. here you first go uhm left and then down

(c) *hier geht's geradeaus und dann nach unten*
E. here one goes straight and then down

The data for each participant for each block (early and late) were converted into mean percent use of route perspective.

The hypothesis that speakers align at the conceptual level of spatial perspective was tested in a 2 (prime: route vs. survey) x 2 (block: early vs. later) analysis of variance on the mean percent use of route perspective. A main effect of prime perspective was found, $F(1, 44)=12.49$, $p=.001$, $\eta_p^2=.22$. No effect of experimental block (early vs. later) was found, and there was no interaction between experimental block and prime perspective. On the early block, participants in the *survey* prime condition described the routes drawn on

their maps in the route perspective only 54.92% of the time while those in the *route* perspective condition did so 88.92% of the time (Table 1). On the later block, following their partner's switch, participants primed by the *survey* perspective also produced significantly fewer descriptions in the route perspective than those who were primed by the route perspective (69.42% vs. 95.17%, respectively).

Table 1: Mean percent use of route perspective before and after prime perspective switch in Experiment 1.

	Early block (pre-switch)	Later block (after switch)
Prime: Route	88.92	95.17
Prime: Survey	54.92	69.42

Experiment 2

The results of the first experiment provided evidence for speakers' alignment with their partner at the conceptual level of spatial perspective both before and after their partner switched from route to survey perspective or the other way round. However, within each of the two experimental blocks, confederates adhered consistently to one perspective only. Thus, when they switched perspective on the later block, prime perspective also remained constant for all four experimental pairs within that block. It is not clear, however, whether speakers' alignment on spatial perspective may have been influenced by this high degree of consistency within an experimental block. The second experiment set out to test whether speakers would also show conceptual alignment of spatial perspective with their partner even if the partner showed high inconsistency and switched perspective all the time, that is, between trials rather than between experimental blocks (as in Experiment 1). Constantly switching perspective may make the confederate's choices appear more random and may thus lead participants to adopt a generally 'random' choice approach themselves. To distinguish between this possible outcome and systematic alignment with one's partner even in the face of the partner's inconsistency, a second experiment was conducted in which speakers' choices were analysed as a function of the immediately preceding prime for each target item.

Method

The design of the second experiment was basically the same with one exception. It included prime perspective (route vs. survey) and experimental block (early vs. later) as independent variables and mean percent choice of route perspective on each experimental block as the dependent variable. However, prime perspective in this case was inconsistent, i.e., constantly alternating between trials.

Participants 19 participants (3 male) took part in the experiment. They were university students with a mean age of 21.32 years (range 19 – 28) who received course credit or were paid for their participation. All were native German speakers.

Stimuli The same visual stimuli were used as in Experiment 1. However, in this second experiment, the confederate switched between route and survey perspective on each trial. The first description they gave was route in one of the experimental lists and survey in the other. Thus, the perspective of the confederate primes was inconsistent.

Procedure The procedure was identical to the one used in Experiment 1.

Results

Participants' responses were classified according to the spatial perspective used as in Experiment 1. The data for each participant for each block (early and late) and for each prime condition (survey vs. route) were converted into mean percent use of route perspective.

A 2 (prime: route vs. survey) x 2 (block: early vs. later) analysis of variance on the mean percent use of route perspective revealed a main effect of prime perspective, $F(1, 61)=5.47$, $p=.023$, $\eta_p^2=.08$. No effect of experimental block (early vs. later) was found, and there was no interaction between experimental block and prime perspective. On average across early and later blocks, participants in the *survey* prime condition described the routes drawn on their maps in the route perspective 68.52% of the time while those in the *route* prime condition did so 88.16% of the time (Table 2).

Table 2: Mean percent use of route perspective on the early and later block in Experiment 2.

	Early block	Later block
Prime: Route	84.21	92.11
Prime: Survey	73.33	62.50

Discussion and conclusions

Experiment 1 showed that speakers do align spatial perspectives with their partners. Those who heard their partner use a survey perspective consistently on the early block of four consecutive experimental trials were less likely to adhere to the otherwise preferred default of route perspective and used instead survey perspective themselves or a mix of the two perspectives in their descriptions. The magnitude of this alignment effect on the early block was 34% and although it was reduced somewhat on the second block to 26%, it nevertheless occurred on this later block of four experimental trials as well. What is more, the 8% reduction was not so considerable as to produce a statistically significant interaction between prime perspective and experimental block, i.e., the alignment

tendency appeared to be equally strong across blocks. This is particularly striking in view of the nature of the second (later) experimental pairs. During those trials, the confederate used the alternative perspective to the one he or she used on the early block, thus displaying a switch from survey to route perspective or vice versa. In this sense, although on each set of four consecutive trials the confederate had made consistent perspective 'choices' in their descriptions, across the two experimental blocks their behavior appeared inconsistent, and yet, participants had the same tendency to align with their partners later as well as earlier during the experimental session. This is notable for two reasons. First, it shows that spatial perspective is used flexibly, and that speakers make use of the possibility to switch perspective with relative ease. Second, it also shows that speakers were not entrained on the first perspective only that they heard their partner use but that they updated. In this sense, this experiment has provided evidence for speakers' sensitivity to their partners' changes in behavior and preference for a representation scheme.

Participants' alignment with their partners in spatial perspective in the first experiment was not significantly reduced on the later post-switch experimental block. However, one good reason for this persistence of alignment even after the switch *between* experimental blocks may have been that the behavior of their partners *within* experimental blocks remained consistent. Experiment 2 put this possibility to the test. Here confederates' pre-scripted descriptions switched between the two perspectives constantly, i.e., if their first, third, fifth, etc. utterances were in a route perspective, then their second, fourth, sixth descriptions were in survey perspective, and vice versa. The analysis of the data revealed that participants aligned even in this case, i.e., they were more likely to use a survey perspective description after they heard their partner use one than if they heard their partner use a route perspective description, an alignment effect of almost 20% difference in choices. Furthermore, this effect did not interact with experimental block (early vs. late), that is, the alignment tendency did not become weaker as time went on. Although the interaction did not reach statistical significance, it is worth noting here that numerically the perspective alignment effect in the later experimental block was much greater (almost 30%) than in the early block. If nothing else, the tendency to align appeared to have been enhanced later. Note that there was no general difference between the early and the later block in this second experimental design, i.e., no sudden change of partner behavior unlike the switch between blocks in the first experiment. In this sense, the growing alignment tendency could be interpreted not as enhanced activation of one of the spatial perspective schemes but more of a general (perspective non-specific) convergence across speakers and accumulation of priming. However, this interpretation can only be offered with a proviso. As described earlier, items where the participant objected to the description used by their partner were not included in data for analysis as priming could not be tested because of an interruption of direct the prime-target sequence and possible interference from self-priming by the alternative 'correction' that participants used in both experiments, although such trials occurred less frequently in the second experiment. Nevertheless, the important finding from Experiment 2 was that speakers aligned in spatial perspective even in cases where their partners exhibited a highly inconsistent descriptive behavior by constantly switching between the two perspective schemas. Such inconsistency by the partner did not lead participants to view either perspective as equally suitable and then adopt one of the two as the easy, less effortful strategy. It did not lead them to make random choices, either. Instead, participants aligned systematically with their partners, i.e., they were prepared to switch perspectives regularly.

Further research into spatial perspective alignment will help solve more mysteries. A memory task experiment (Andonova & Coventry, 2009) has revealed spatial perspective priming. A comparison of the two studies indicates common underlying mechanisms that need to be explored further.

The main conclusions of the two experiments described here are as follows. We found evidence for spatial perspective alignment across speakers in a route description task. Perspective alignment was sensitive to consistency of use by one's partner in the early stages of the interaction (a much weaker alignment tendency of approximately 10% on the early block in Experiment 2 in comparison with the robust 34% effect in Experiment 1). Perspective alignment persisted even after a switch in partner behavior, i.e., alignment persisted even when perspective did not.

Acknowledgments

Thanks go to the German Science Fund (DFG) for funding provided to SFB/TR8 Spatial Cognition at the University of Bremen where this research was conducted, to the research assistants and confederates, and the research team of I5-DiaSpace. The comments and suggestions of the anonymous reviewers of the manuscript are gratefully acknowledged.

References

Andonova, Coventry, & Tenbrink (in press). Function and context affect spatial information packaging at multiple levels. *Psychonomic Bulletin & Review*.

Andonova, E., & Coventry, K. (2009). Alignment and Priming of Spatial Perspective. In: Edlund, J., Gustafson, J., Hjalmarsson, A., & Skantze, G., (Eds.), *Proceedings of Diaholmia*, Workshop on the Semantics and Pragmatics of Dialogue. Stockholm: KTH.

Branigan, H. P., Pickering, M. J., McLean, J. F., & Cleland, A. A. (2007). Syntactic alignment and participant role in dialogue. *Cognition, 104,*163–197.

Bugmann, D., Coventry, K.R., & Newstead, S.E. (2007). Contextual cues and the retrieval of information from cognitive maps. *Memory & Cognition*, 35(3), 381-392.

Clark, H. H. (1996). *Using Language.* New York: Cambridge University Press.

Garrod, S.C., & Anderson, A. (1987). Saying what you mean in dialogue: a study in conceptual and semantic co-ordination. *Cognition*. 27, 181–218.

Pickering, M., and Garrod, S. 2004. The Interactive Alignment Model. *Behavioral and Brain Sciences,* 27(2):169-189.

Schober, M.F. (1993). Spatial perspective-taking in conversation. *Cognition*, 47, 1-24.

Striegnitz, K., Tepper, P., Lovett, A., Cassel, J. (2008). Knowledge representation for generating locating gestures in route directions. In: Spatial Language in Dialogue. Oxford University Press, Oxford.

Taylor, H.A., Naylor, S.J., & Chechile, N.A. (1999). Goal-specific influences on the representation of spatial perspective. *Memory & Cognition*, 27(2), 309-319.

Taylor, H.A., & Tversky, B. (1996). Perspective in Spatial Descriptions. *Journal of Memory and Language*, 35, 371-391.

Spatial Representations with Conflicting Intrinsic Frames of Reference

Franklin P. Tamborello, II (franklin.tamborello@uth.tmc.edu),
Yanlong Sun (yanlong.sun@uth.tmc.edu), &
Hongbin Wang (hongbin.wang@uth.tmc.edu)
School of Health Information Sciences,
University of Texas Health Science Center at Houston
7000 Fannin, Ste. 600
Houston, TX 77030 USA

Abstract

Establishing and updating spatial relationships between objects in the environment is vital to maintaining situation awareness. Wang et al. (2005) found that updating of spatial representations in the intrinsic frame of reference (IFOR) can be prioritized based on salience of task demands. But their study used a task environment with only one IFOR. Often a task environment has several objects in it which may be task-relevant, and they may conflict with each other in one or more ways such as by being oriented in differing directions. Two experiments manipulated relative spatial orientation and task salience of two task-relevant objects such that the objects' orientations conflicted with each other and the task probabilistically demanded response based on the orientation of one or the other object. It was found that spatial updating in the IFOR was constrained by the limits of human attentional processes. Furthermore those constraints can be relaxed with practice.

Keywords: prioritized representation updating; conflicting spatial representations; spatial cognition; intrinsic frame of reference.

Introduction

Spatial cognition is crucial to our everyday interactions with our environment and other people including, for instance, maintaining awareness of one's task environment. A large body of evidence suggests that people organize spatial representations and reason about spatial relationships using frames of reference (FORs, (Levinson, 1996; Wang, Johnson & Zhang, 2001). FORs can be based on our own viewpoints, expressing spatial representations that are centered on ourselves (the egocentric FOR, "EFOR"). EFORs represent spatial affordances within our immediate vicinity, such as a pencil that is within reach. FORs based on navigable environments (the allocentric FOR, "AFOR"), such as rooms, buildings, or cities, represent the shapes of those environments and what affordances they give to wayfinding (Klatzky, 1998; Mou & McNamara, 2002; Wang & Spelke, 2002).

Most research in spatial cognition has focused on the EFOR and AFOR (May & Klatzky, 2000; Shelton & McNamara, 2001) though it is possible to distinguish a third type of FOR, the intrinsic (IFOR), so named because it is intrinsic to the person or object of focus (Mou & McNamara, 2002; Wang, Sun, Johnson & Yuan, 2005). The IFOR is a unique FOR that brings the spatial representation affordances of the EFOR outside the observer's body. The IFOR enables us to imagine spatial relationships from positions other than the one we currently occupy, including the positions of other people. This is important for action planning, interpersonal communication of specific spatial representations, and even theory of mind. For instance, spatial relationships such as, "John is sitting to Mary's right." are represented in the IFOR. Here Mary is the reference anchor around which the framework for the spatial relationship of John's position is based (Levinson, 1996).

Given the importance of IFOR-based spatial representations in everyday tasks, one fundamental question is how easily IFOR representations can be updated within the context of a changing environment. It has been shown that egocentric representations can be updated fairly easily whereas updating allocentric representations other than self-locations often requires effort. Wang, Sun, Johnson, and Yuan (2005) studied IFOR spatial representations and how they may be updated to reflect changes in the task environment, particularly as a function of a target object's task salience. They found that updating of spatial representations in the IFOR can be prioritized based on salience to task demands and that IFOR updating is often, but not always, easy for those salient objects. However their study used a task environment with only one IFOR-supporting object. Often a task environment has several such objects in it which may be task-relevant and they may conflict with each other in one or more ways such as by being oriented in differing directions. In the above example regarding John and Mary, we may also notice that "John is sitting to Sam's left." In this case, John's spatial location is represented in an IFOR centered on Sam and that spatial relationship to John is not the same for Mary and Sam. When John moves, both Mary's and Sam's spatial representations should be updated.

Presumably increasing the number of task relevant IFOR-supporting objects would increase task complexity and consequently demand more attentional resources. At a certain point people will have to prioritize not only their updating of spatial representations of the targets of their actions but also the reference anchors of those spatial representations. In other words, if there are multiple IFOR-supporting objects that must be attended in a task environment then people will need to prioritize their updating not only of the action-target objects but also of the IFOR-supporting objects.

A "Two Cannons" pointing task was designed to test hypotheses regarding updating priority in a two-IFOR spatial task environment. In this task participants needed to determine which way a depicted cannon should turn to point at a designated target. Salience of the two cannons varied so as to make one or the other more important to the completion of the task. If people can attend to only one

IFOR at a time then in a an environment where multiple IFORs may exist updating those representations must be prioritized somehow. If priority of updating between IFORs goes according to salience, as Wang et al. (2005) found for updating target priority within one IFOR, then response time should vary with the targeted IFOR's relative salience. That is, when two IFORs conflict the conflict should be resolved most easily in favor of the more salient IFOR. Furthermore, if people can only form one IFOR at a time they may wait to see which IFOR to use before they invest the time in forming it. Then we would expect an effect of conflict as they wait to see which IFOR to use but no effect of relative IFOR anchor angle since anchor angle would be irrelevant to IFOR selection. On the other hand if people can form and maintain multiple IFORs simultaneously then when the IFOR anchor objects conflict with each other on some dimension (e.g., orientation) there should be an effect of relative IFOR orientation angle such that at one angle the irrelevant IFOR may be easier to inhibit than at another angle.

Experiment 1

We designed a "two cannons" turning response task to investigate how people represent spatial information with multiple conflicting IFORs and how they resolve the conflicting attentional demands of updating spatial relationships involving multiple IFORs. The task required participants to determine the location of a designated target relative to a matching-color IFOR anchor, one of the two cannon stimuli. Orientation of the cannon stimuli varied so that turning direction responses dependent upon those orientations would conflict based on the orientations of the two IFOR anchors, the cannons, with respect to the indicated target. Task salience of the two anchors also varied so as to weight the conflict in favor of one IFOR or the other. If IFOR updating tends to be prioritized according to salience as Wang et al. (2005) found, then response time for each IFOR should be a function of that IFOR's relative salience. That is, when two IFOR spatial relationships conflict, the conflict should be resolved most readily for the more salient IFOR.

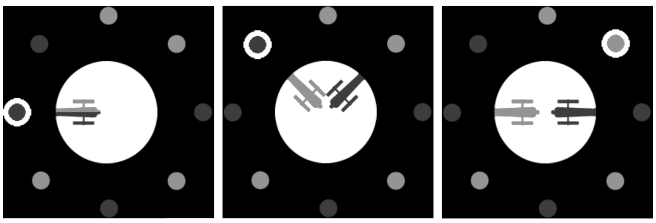


Figure 1. Experiment displays depicting 0° relative cannon angle (A, left), 90° cannon angle (B, center), and 180° cannon angle (C, right). Here blue is depicted as dark gray and red is light gray. In A both cannons (depicted as half red, half blue) are at 270° orientation. The highlighted blue dot indicates the target. In this case the correct response would be to punch the up arrow key to indicate that no turn is required. In B the correct response would be to punch the left arrow key, indicating that the blue cannon would have to turn to its left to face the target. In C the correct response would be to punch the right arrow key, indicating that the red cannon would have to turn to its right to face the target.

Method

Participants Ten graduate students and postdoctoral fellows were paid to participate in Experiment 1. Subjects had a mean age of 32.1 years (SD = 7.75) and five were female.

Design Table 1 enumerates the conditions for Experiment 1. We established conflict in the cannon IFORs by manipulating relative angle between the two cannons so that the two cannons either were on top of each other as in Figure 1A, at 90° to each other (Figure 1B), or 180° to each other (Figure 1C). We used 90° & 180° to compare degree of conflict. The ratio of blue dots to red dots varied sequentially within each trial block, always starting at 8 blue to 0 red and transitioning in increments of 2 dots to 0 blue to 8 red. Color ratio conditions occurred as sub-blocks of eight trials, each of which exhausted the set of eight possible target locations. Thus each block of 40 trials exhausted each of five color ratio conditions once and each of eight target location conditions five times for one combination of relative cannon angle and cannon orientations. Relative cannon angle and cannon orientations varied randomly by block.

Materials The experiments ran on a PC in E-Prime version 1.2. The two cannons subtended a viewing angle of

Table 1. Combinatorial table of conditions of Experiment 1. All factors varied within-subjects, except color ratio.

Relative Cannon Angle	Cannon Orientation (specific to each cannon angle condition)	Target Position (all cannon angle conditions)	Target Color (all cannon angle conditions)	Dot Color Ratio (all cannon angle conditions)
0°	90°(blue & red)	0° – 315°, with 0° being up and incrementing clockwise in steps of 45°. 8 positions total. Varied randomly, without replacement, within each color ratio cycle.	red or blue, varied randomly within each color ratio cycle, constrained by color ratio condition.	8 blue : 0 red – 0 blue : 8 red, in increments of 2 dots. 5 color ratios total. Varied sequentially within each trial block.
	270°(blue & red)			
90°	45° (blue) & 315° (red)			
	135° (red) & 225° (blue)			
180°	90°(blue) & 270°(red)			
	90°(red) & 270°(blue)			

approximately 6° while the entire display of cannons and surrounding dots subtended a viewing angle of approximately 12°. The cannon stimuli were constructed such that they each had an obvious intrinsic orientation (Figure 1). They appeared as though viewed from above, with wheels at their rear and a barrel in the middle, extending far forward.

Procedure Upon onset of the stimulus display, the experiment paused for one second before it flashed a yellow ring around one dot to indicate that it was the target. The matching-color cannon thus became task-salient. Participants were to then respond as quickly and accurately as possible which way the salient cannon should turn to face the dot: left, right, or no turn. Responses had the stipulation that the turn was to be the shortest way round. Participants indicated their responses with the left, right, and up arrow keys, respectively. In the case wherein the target was directly behind the indicated cannon, participants could respond either left or right as turning either way would result in a change in cannon orientation of 180°. The experiment played a "zap" sound as feedback for a correct trial. In the case of incorrect trials the experiment paused for two seconds to discourage random guessing and it played a distinctive "uh oh" sound. Subjects erred on fewer than 5% of trials on average.

Results and Discussion

Data from both experiments were filtered for subject error and outliers, outliers being outside the subject's mean ± 3 standard deviations. This removed approximately 5% of observations. Figure 2 depicts effects on response time of the interaction of target color and dot color ratio. Again, in Experiment 1 color ratio sequence began with all blue dots

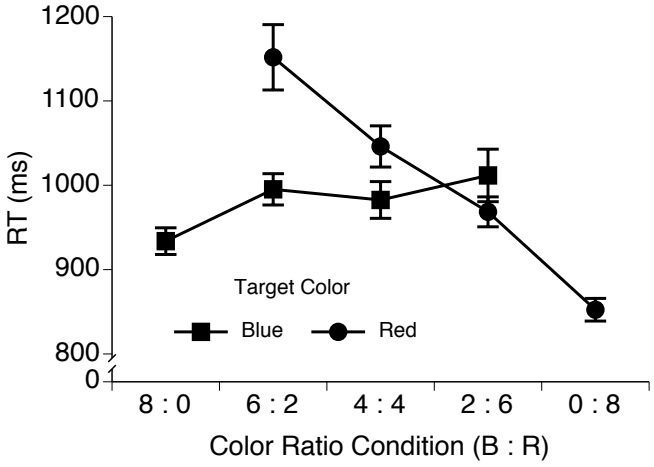


Figure 2. Experiment 1 response times as a function of the interaction of target color and dot color ratio. Within-block color ratio sequence progressed left-to-right along the x-axis. Error bars depict standard error of the mean.

and transitioned gradually to all red dots (i.e., from 8 blue : 0 red to 0 blue : 8 red, hereafter abbreviated #blue:#red for Experiment 1). Repeated measures ANOVA found that target color by color ratio linear by linear interaction contrast was reliable, $F(1, 9) = 39.25$, $p < .001$, meaning that the RT function of blue targets and red targets over color ratio differed. In addition, Color ratio's main effect was reliable, $F(3, 27) = 8.38$, $p < .001$.

The results suggest that for the blue targets as the blue cannon became less salient as the number of dots transitioned from blue to red, the ability of subjects to respond to the blue cannon did not fall off, it stayed the

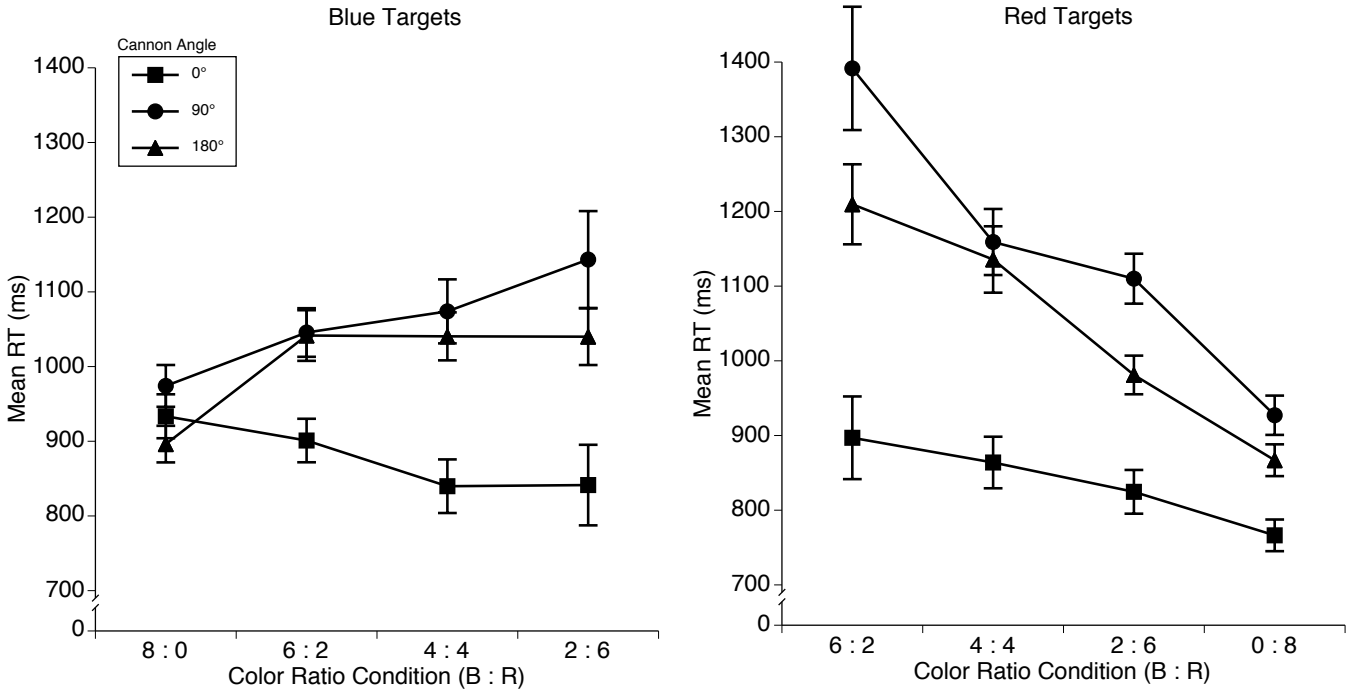


Figure 3. Experiment 1 RTs by target color, cannon angle, and color ratio condition. Error bars depict SEM.

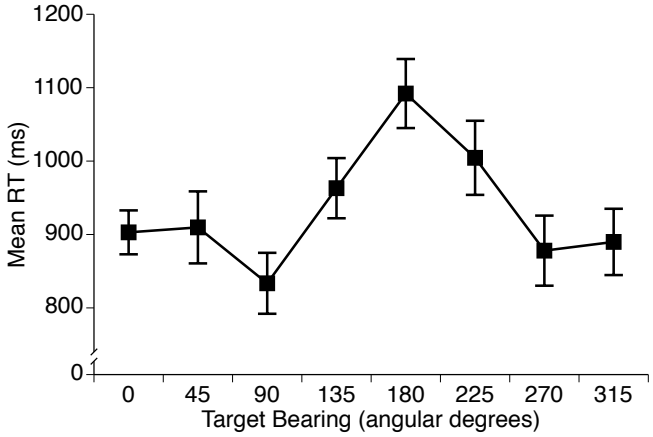


Figure 4. Experiment 1 RTs by target bearing at color ratio 8:0. Error bars depict SEM.

same. This is surprising given that as the blue cannon decreased in salience subjects should have paid less attention to it, thus taking longer to recognize and respond to a rare blue target trial. Additionally, the preservation of the ability to respond to blue did not come at the cost of responding to red, as red targets showed a dramatic decrease in response time across the color ratio condition progression. This indicates little or no strategic trade-off of prioritizing one IFOR over the other.

Breaking the interaction down to more specific experimental conditions, it is clear that IFOR conflict mattered when interacting with color ratio (Figure 3). Collapsing across 90° and 180° cannon angles gives an abstracted conflict versus no conflict (0° cannon angle) contrast. With five post hoc comparisons for this family of tests of the target color by color ratio by conflict interaction, the Bonferroni-corrected α' = .01. Conflict by target color was not reliably different for blue targets ($t(9) = 3.028$, $p = .014$), but was for red ($t(9) = 3.821$, $p = .004$). This means that conflict in IFORs was a product of task demands, which in turn was a combination of IFOR spatial properties and probabilistic IFOR selection properties. The 90° versus 180° cannon angle difference was not reliable (for red targets in the 2:6 color ratio condition, $t(9) = 3.095$, $p = .013$). Furthermore the slope of the conflict versus no conflict by color ratio interaction function was different for blue targets, but not for red targets: $t(9) = 3.903$, $p = .004$; and $t(9) = -1.779$, $p = .109$, for blue and red respectively. This means that for blue targets as the color ratio progressed from blue to red RTs got slightly faster for 0° cannon angle trials but slower for 90° and 180° cannon angle trials. This indicates that participants really saw the two overlapping cannons as one IFOR in the 0° cannon angle condition, whereas the 90° and 180° cannon angle conditions worked well as a manipulation to induce IFOR conflict. When the color ratio was 6:2 the cost to switch attention on the IFORs can be calculated as the difference between the RTs for the blue and red targets within the conflict condition. That switching cost was 257 ms.

The 8:0 condition might be taken as a base case of the two cannons task in that the color ratio of the dots perfectly predicts target color, and therefore which cannon participants should attend. Here, then, we can get a sense for target bearing's RT function (Figure 4). It shows that targets at 135° and 225° bearing took longer to respond to than targets at other bearings (except 180°, which was subject to Hick's Law since participants could respond either direction to this target bearing), $t(9) = -3.848$, $p = .004$.

Kessler & Thomson (2010) used a similar response scheme in their perspective alignment task. They found a flat target bearing function except for longer RTs at 135° rotation in either direction. They speculated that visual comparisons could be made up to about 90° of rotation but that greater degrees of rotation required complex imaginal transformations that took longer. Presumably the same cognitive and perceptual-motor processes take place with the two cannons task since it also requires the alignment of perspectives with the IFOR of the designated cannon.

However, the target bearing function went flat for conditions where a switch of attended cannon was likely, namely when the target belonged to the non-salient IFOR. For instance, when the color ratio was 2:6, a red target was more likely to appear than a blue target. That probability difference made the red cannon more task-salient. Subjects could therefore save some response time by attending the red cannon during the SOA. But if the target turned out to be blue then subjects would have to move attention to the blue cannon and establish a new IFOR around it. When this happened RTs were not longer for 135° or 225° target bearing, contrast for blue targets at 2 blue to 6 red $t(7) = -.997$, $p = .352$; contrast for red targets at 6 blue to 2 red, $t(7) = -2.087$, $p = .075$. Note that df = 7 for these two analyses as two subjects were missing data for these cells, likely due to subject error or outlier RTs. The 135°/225° target bearing effect probably went away for these two conditions because target position would already have been known when it became clear that the non-salient cannon must form the basis of the response. Target bearing could then be integrated during the IFOR attention switch latency rather than, as in the 8:0 color ratio condition, having all other representations formed before target onset and being the last representation left to be formed before responding.

It could be that each piece of the spatial information is acquired and represented as it becomes available, and that pieces are retrofitted into the rest of the representation as needed. This could mean that in 6:2 with a red target, for example, the potential targets have their representations built first (maybe in association with the more likely IFOR), and after the target onset the targeted IFOR is built and retrofitted to the extant spatial environment representation.

Experiment 2

The asymmetry of the target color interaction with color ratio found in Experiment 1 was unexpected, and if real, could imply that people, with practice, may be able to maintain representation more than one IFOR at a time. As trial blocks progressed and blue became less salient then response times for blue targets should have become longer as the blue cannon reduced in updating priority relative to the red cannon. Instead a practice effect on the blue IFOR was apparently sufficient to cancel the expected probability

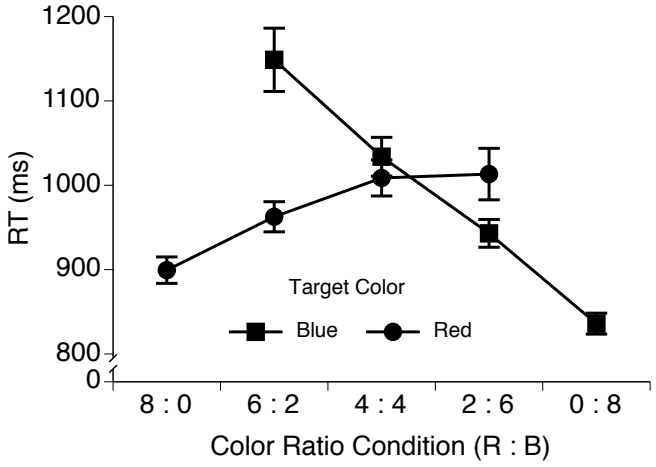


Figure 5. Experiment 2 RTs as a function of the interaction of target color and dot color ratio. Within-block color ratio sequence progressed left-to-right along the x-axis. Error bars depict SEM.

matching effect for the blue IFOR. However, it is also possible that the interaction effect could be due to sequence effects of color ratio presentation order. Experiment 2 was designed to test this possibility by replicating Experiment 1 except that dot color ratios were sequenced in the opposite order, this time going from red to blue within each block.

Method

Ten graduate students and postdoctoral fellows were paid to participate in Experiments 2. Subjects had a mean age of 32.1 years (SD = 6.03) and four of them were female. Experiment 2's design duplicated Experiment 1's except that color ratios incremented from all red to all blue rather than blue to red as in Experiment 1. Experiment 2 replicated Experiment 1's materials and procedures identically.

Results and Discussion

The target color by conflict by color ratio interaction of Experiment 1 replicated in Experiment 2 (Figures 5 and 6), the different color ratio sequence not withstanding (linear by linear interaction contrast of color ratio with target color $F(1, 9) = 96.6$, $p < .001$). With four post hoc comparisons for this family of tests, the Bonferroni-corrected $\alpha' = .0125$. Blue targets with conflicting IFORs took longer for subjects to respond to than blue targets with no conflict, contrast $t(9) = 4.629$, $p = .001$; and likewise for red targets $t(9) = 4.119$, $p = .003$. Also for red targets with conflict RTs got longer as the color ratio transitioned to more blue dots while the RTs became shorter for red targets with no conflict ($t(9) = -5.865$, $p < .001$), but the same was not true for blue targets ($t(9) = 3.035$, $p = .014$), all blue target RTs got shorter regardless of conflict status. This means that when subjects had to choose between IFORs (conflict), practice effects interacted with the time costs associated with switching attention between IFORs and the fact that target color was selected probabilistically from the set of dots. The time costs existed in turn because only one IFOR could be attended at any one time and moving attention from one to the other cost 215 ms.

As for target bearing, 135° and 255° were again slower than other target bearings, 180° excluded, in this experiment's color ratio and target color "base case," $t(9) = -2.706$, $p = .024$ (Figure 7). The two target bearings were not reliably different for blue targets at 2 blue to 6 red ($t(6)$

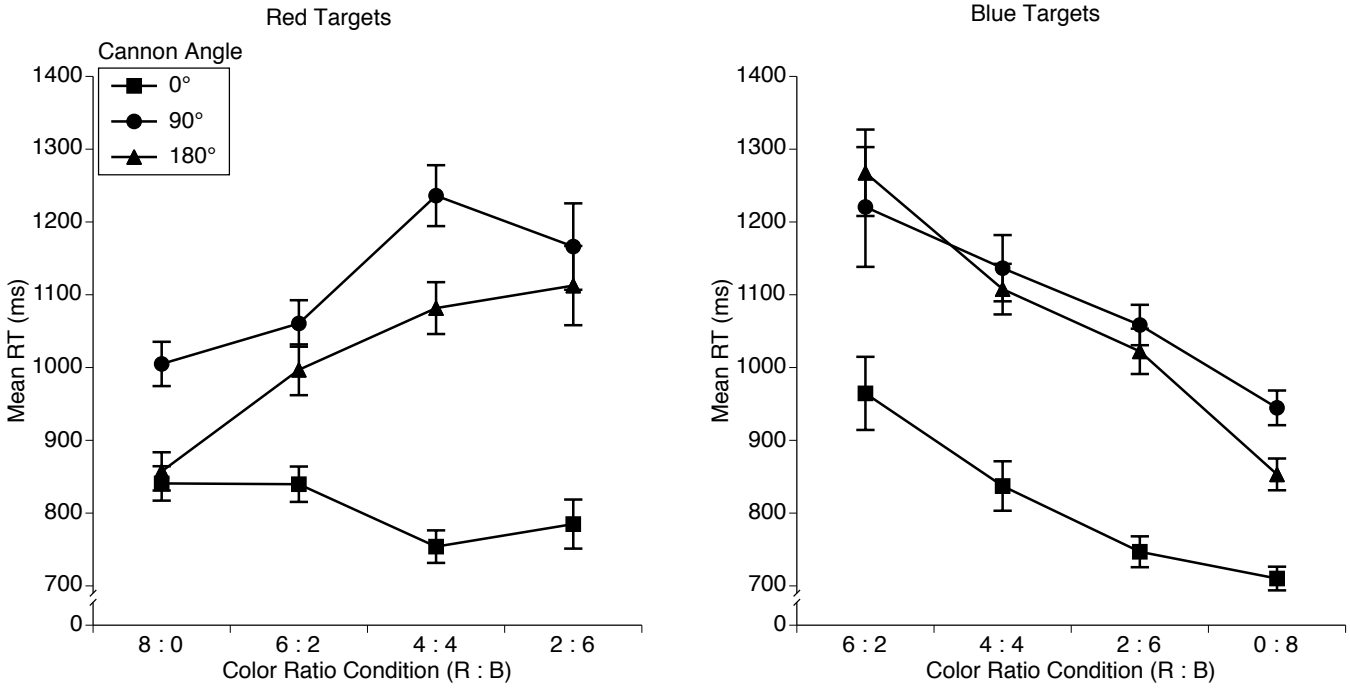


Figure 6. Experiment 2 RTs by target color, cannon angle, and color ratio condition. Error bars depict SEM.

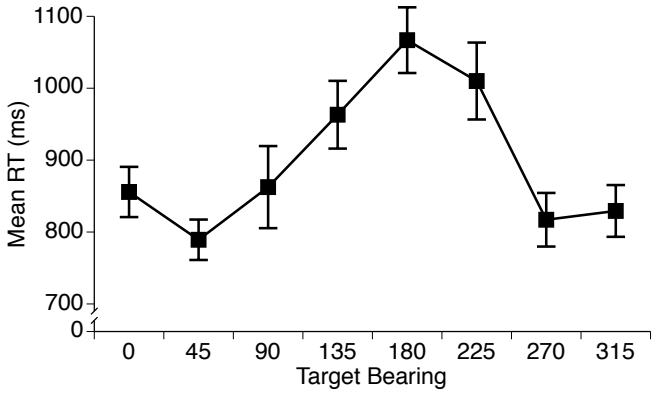


Figure 7. Experiment 2 RTs by target bearing at color ratio 0:8. Error bars depict SEM.

= -1.193, p = .278), nor for red targets at 6 blue to 2 red (t(9) = -1.721, p = .119). So again, having to switch IFORs precludes the kind of processing that yields the Kessler and Thomson (2010) type of target bearing effect.

General Discussion

People use IFORs every day. In social contexts, for example, we often infer various inter-personal relationships based on the spatial relations (such as position and distance) among people. Humans must know how to to prioritize updating of IFORs so that the most relevant IFORs are represented sufficiently richly to support the task at hand. The present study extends the findings of Wang et al. (2005) to study prioritization and updating among multiple IFORs. Spatial updating was found to be significantly constrained by the limits of human attentional processes as evidenced by a switching cost of approximately 236 ms (averaged across the two experiments). But those constraints can be relaxed somewhat with practice, as our interaction effect of target color with color ratio showed. The results indicated that subjects engaged in little strategic trade-off of prioritizing one IFOR over the other, as apparently the cost of attending one IFOR-supporting object or the other, and engaging the target bearing representation based on that IFOR, decreased somewhat with practice. In a larger sense this suggests that since the cognitive mechanisms supporting IFOR-type representation are apparently susceptible to practice effects, that reasoning comes relatively late in the human attentional stream.

Meanwhile the conspicuous absence of a target bearing effect for the conditions in which a switch of attended IFOR was likely hints that the establishment of an IFOR is able to take advantage of spatial representations already in working memory. IFORs, therefore, may be limited to one instantiation at a time as a function of human working memory capacity, but they may be disbanded and instantiated dynamically, taking advantage of whatever spatial representation may be available at the time to be incorporated into the IFOR to support task performance as necessary

It is clear that attention plays a central role in mediating IFOR representational conflicts and in modulating salience. Future work should clarify the computational mechanisms underlying spatial salience and people's capacity to effectively process spatial relationships in multi-IFOR task environments.

Acknowledgments

This work was supported by Office of Naval Research Grant #N000140110132 and a training fellowship from the Keck Center NLM Training Program in Biomedical Informatics of the Gulf Coast Consortia (NLM Grant #T15LM007093).

References

Kessler, K., & Thomson, L. A. (2010). The embodied nature of spatial perspective taking: Embodied transformation versus sensorimotor interference. *Cognition, 114 (1)*, 72-88.

Klatzky, R. L. (1998). Allocentric and Egocentric Spatial Representations: Definitions, Distinctions, and Interconnections. In *Spatial Cognition - An Interdisciplinary Approach to Representation and Processing of Spatial Knowledge (Lecture Notes in Artificial Intelligence 1404)*.

Levinson, S. C. (1996). Frames of reference and Molyneux's question: Crosslinguistic evidence. In P. Bloom, M. A. Peterson, L. Nadel, & M. Garrett (Eds.), *Language and space*. Cambridge, MA: MIT Press.

May, M., & Klatzky, R. L. (2000). Path integration while ignoring irrelevant movement. *Journal of Experimental Psychology: Learning, Memory and Cognition, 26(1)*, 169-86.

Mou, W., & McNamara, T. P. (2002). Intrinsic Frames of Reference in Spatial Memory. *Journal of Experimental Psychology: Learning, Memory and Cognition, 28(1)*, 162-70.

Shelton, A. L., & McNamara, T. P. (2001). Systems of spatial reference in human memory. *Cognitive Psychology, 43(4)*, 274-310.

Wang, H., Johnson, T. R., & Zhang, J. (2001). The mind's views of space. In *Proceedings of the Third International Conference of Cognitive Science*.

Wang, H., Sun, Y., Johnson, T. R., & Yuan, Y. (2005). Prioritized Spatial Updating in the Intrinsic Frame of Reference. *Spatial Cognition & Computation, 5(1)*, 89-113.

Wang, R. F. W., & Spelke, E. S. S. (2002). Human spatial representation: insights from animals. *Trends in Cognitive Science, 6(9)*, 376 – 382.

Learning from Environmental Regularities is Grounded in Specific Objects not Abstract Categories

Lauren L. Emberson (lle7@cornell.edu)
Sackler Institute for Developmental Psychobiology; Weill-Cornell Medical School
Box 140, 1300 York Ave., NY, NY 10065 USA

Dani Rubinstein (rubinsteind@mail.nih.gov)
MEG Core Facility, National Institute of Mental Health
Building 10 - Magnuson Clinical Center, B1D65B, 10 Center Drive, Bethesda, MD 20892 USA

Abstract

This paper examines statistical learning in the presence of predictive regularities at multiple levels of abstraction. Participants were presented with streams of pictures where picture order was predicted by both object identity and the categories these objects belong to. In Experiment 1, we establish that participants do learn based on the specific objects and not solely at the abstract, categorical level. In Experiment 2, we discount the possibility that participants gain abstract knowledge in addition to more concrete, object-based knowledge. Moreover, we consistently find equal learning in those who viewed the atypical exemplars and those who viewed the typical exemplars of the categories. Overall, our results suggest that when learning from environmental regularities, object-specific information takes precedence over more abstract, category level information when both are predictive.

Keywords: Statistical learning; environmental learning; visual development; perceptual learning; object perception; categorization.

Introduction

Throughout our lifetimes, it is clear that experience shapes our mental model of the world. Focusing on adulthood, adults learn to recognize new objects and categories as well as new properties of familiar objects; they learn new words and adapt to changing patterns in the ambient language, all by adapting future behavior based on experience. Despite the clear importance of learning from information in the environment, the nature of the mechanisms that support learning from real-world experience is largely unknown. A central problem in this literature is how learning mechanisms operate given the richness of the information we get from the world. Are learning mechanisms *a priori* constrained to learn particular patterns? Can learning proceed along many types of perceptual information and/or at different levels of abstraction[1] simultaneously?

In this paper, we focus on a type of learning called "statistical learning" where participants passively learn from stimuli embedded with probabilistic information[2]. Previous research has supported the view that these experiential learning mechanisms are unconstrained: statistical learning has been demonstrated in multiple sensory modalities (Conway & Christiansen, 2005), across a wide range of perceptual input. For example, in the visual modality, learning can occur from sequences of gestures (Baldwin, Anderrson, Saffran, & Myers, 2007) as well as abstract shapes (Fiser & Aslin, 2001). While the majority of these studies have focused on learning probabilistic relations of individual items or objects, there is evidence that learning can occur at higher levels of informational abstraction including over new categories of nonsense words (Saffran, 2002) and based on familiar semantic categories (Brady & Oliva, 2008).

Overall, these studies support the view that environmental learning is unconstrained. That is, if there is any reliability probabilistic information in the environment, humans can learn from it regardless of level of abstraction or perceptual properties. If learning is entirely unconstrained, it is unclear how learning mechanisms operate in complex environments where information from multiple sources and at many levels of abstraction abounds.

However, these behavioral demonstrations of an entirely unconstrained learning mechanism arise from paradigms in which information is only predictive at a single perceptual and/or informational dimension. For example, while Brady and Oliva (2008) demonstrate learning of categories of scenes, participants were presented with a new scene from the category during each successive presentation. In this paradigm, individual scenes (e.g. $beach_1$ and $beach_2$) are not predictive of picture order, only the category of pictures are (e.g. a beach predict a kitchen but $beach_1$ does not predict $kitchen_1$), thus it would be impossible for participants to learn based on individual scenes. Thus, these results provide an existence proof of an unconstrained learning mechanism but they arise under specific, restricted conditions.

In actuality, environmental stimuli exhibit statistical regularities at many levels of abstraction, simultaneously.

[1] By "levels of abstraction" we are broadly referring to the multiplicity of ways in which a cognitive system can represent a given object or experience: e.g. your pet could be "Rex", a beagle, a dog, an animate being, a brown object etc.

[2] In the current paradigm, a stream of pictures is embedded with regularities that predict picture order—predictive regularities. If participants learn from this probabilistic environmental information, they should be able to distinguish picture orders that they observed from scrambled or foil orders of pictures.

For example, the predictive relationship between dogs and leashes exists based on abstract categories as well as in the actual objects or exemplars seen in the world (e.g. dogs have their specific leashes). The learning paradigms reviewed above do not reflect this important aspect of information that we receive from the world: information is often redundant across multiple levels of abstraction.

The current paper systematically investigates learning where participants are exposed to environmental regularities at multiple levels of abstraction. Do participants learn from the multiple levels of predictive dependencies simultaneously or are they biased to information at a certain level of abstraction? To address this question, we devised a novel statistical learning task where predictive regularities are learnable and redundant at multiple levels of abstraction. Specifically, participants were presented with sequences of new exemplars from known categories. Both the categories (e.g. dogs-fish, flowers-birds) and the individual exemplars of these categories (e.g. dog_1-$fish_1$, dog_2-$fish_2$) were predictive of picture order (see Figure 1). In two studies, we examined whether participants learn simultaneously based on both types of information or whether participants learn preferentially based on categorical or object-based regularities.

We believe that the current experimental design provides ample opportunity for learning at the abstract, categorical level. First, previous research has established that the categories used in the current experiment are initially processed at the basic-level (dog as opposed to the subordinate level of beagle; Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976) which is the level of categorical regularities of the picture stream. Second, we employed the same stimulus timing (short durations and long inter-stimulus-intervals) as employed by Brady and Oliva (2008) which will likely tap into the fast, gist-based recognition of the pictures. Finally, the stream has fewer pairs of categories than objects (see Figure 1). Thus, category level learning is, in some sense, easier than object-based learning.

Using the same methodology as Brady and Oliva (2008) as described above, pilot testing confirmed that when categorical regularities are predictive of picture order but individual objects or exemplars are not participants can learn based on categorical regularities: mean = 62.6%, $t(13) = 2.80$, $p < 0.05$. These results confirm that if object-based regularities are not present, category-level statistical learning is possible using the current stimuli and categories.

Finally, in order to more closely examine how learning proceeds at the categorical level of information, we manipulated the typicality of the exemplars that participants viewed: roughly half the participants were familiarized with typical exemplars of the categories and the rest were familiarized with atypical exemplars (see Appendix 1 for the atypical exemplars). Research has consistently shown that atypical exemplars are processed differently from typical exemplars (Dale, Kehoe, & Spivey, 2007) and tend to be more quickly processed below the basic-level categories (e.g. penguin as opposed to bird; Jolicoeur, Gluck, & Kosslyn, 1984). Thus, we expect the participants familiarized with atypical exemplars to have weaker learning at the category-level but equivalent learning at the object or exemplar specific level. This typicality manipulation provides another way to examine performance for evidence of learning across different levels of abstraction.

Experiment 1: Testing for Object-Level Learning

The first experiment examines learning based on regularities of individual objects where both objects and object categories are predictive of picture order. Figure 1 illustrates a sample familiarization stream. We employed a testing procedure that is well-established in the statistical learning literature (e.g. Brady & Oliva, 2008; Fiser & Aslin, 2001): participants were asked to distinguish pairs of pictures from familiarization (e.g. $bird_1$-dog_1) from a foil pair created from the same pool of pictures but which violates contingency pattern of the familiarization stream. To isolate knowledge at the object-specific level, the foils were designed to violate object-based regularities while maintaining categorical regularities (e.g. $bird_1$-dog_2, see top panel of Figure 2). Thus, participants *require* object-level knowledge of the familiarization stream in order to distinguish the foils from the pairs. Given this experimental design, if participants are able to consistently distinguish pairs from foils, this is evidence for learning based on the objects and not the categories presented during familiarization.

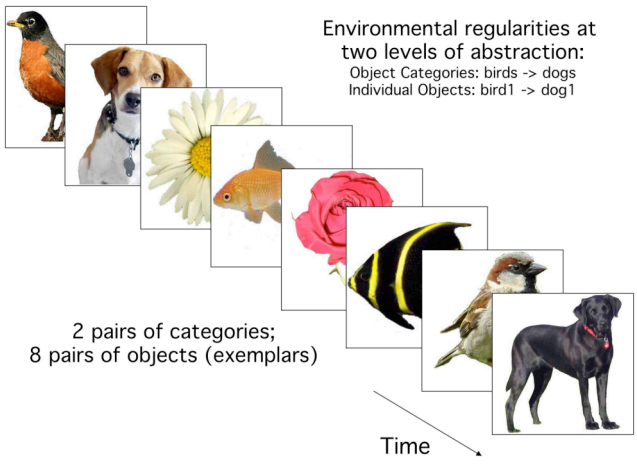


Figure 1: A sample familiarization stream. Pictures were organized into pairs of categories (e.g. birds > dogs) as well as specific objects within these categories (e.g. robin > beagle). Thus, predictive regularities were redundant across multiple levels of abstraction resulting in two pairs of categories and eight pairs of objects or exemplars of these categories.

Methods

18 undergraduate students participated in Experiment 1 (age: mean = 20.7, std = 1.75; 2 left handed; 10F) and randomly assigned to each condition: 10 participants viewed the typical pictures, and 8 viewed the atypical pictures. All participants were from Cornell University, participated in exchange for course credit, and provided informed consent.

Familiarization A statistically-structured familiarization sequence was presented, using PsyScope X B53 on a MacMini computer with a 17in CRT monitor. Each picture was displayed for 300ms with a 700ms inter-stimulus interval (Brady & Oliva, 2008)

There were 4 categories of pictures: birds, dogs, fish, and flowers. For each category 4 different exemplars were used (dog_1, dog_2, etc.). The pictures were grouped into 8 pairs such that both the categories and the specific exemplars were predictive of picture order. For example, $bird_1$-dog_1 would always occur as a pair, as would $bird_2$-dog_2, $bird_3$-dog_3, and $bird_4$-dog_4. Thus, the familiarization stream contains multiple, redundant levels of predictive information: both the exemplar level and the more abstract category level of information are predictive. See Figure 1 for an illustration of the familiarization sequence. To ameliorate any effect of specific pairings on learning, different categories and object pairings were employed across participants. Participants saw each pair 28 times presented in random order without pairs repeating each other and were simply instructed to look at the pictures.

Testing After familiarization, the participants performed a test in which two pairs of pictures were presented sequentially: 700ms between pictures in the same pair, and 1200ms separating the pairs. One pair was from the familiarization (e.g. $bird_1$-dog_1), and one was a foil pair (e.g. $bird_1$-dog_2; see Figure 2). The foils were designed to violate the structure *only* at the exemplar level, and not the category level. Thus, participants require exemplar level knowledge of the familiarization stream in order to distinguish the foils from the true pairs. This test determines whether participants learn the familiarization sequence at the level of exemplars or specific figures or at the level of abstract categories. The participants were instructed to choose which of the pairs seemed more familiar, based on the familiarization task. No time constraint was imposed for their responses. There were 64 test trials.

After the experiment, the participants completed a survey in which they rated the pictures they had seen on a scale of 1-5 for interestingness and typicality. They were also asked to repeat the instructions of each task, to ensure they understood them correctly. Finally, they were asked whether they noticed any patterns during the familiarization sequence, to check for explicit knowledge of the sequence structure.

Results and Discussion

The current experiment was designed such that only exemplar specific knowledge could distinguish pairs seen during familiarization and foils. Performance was evaluated against chance (50%) for evidence of learning. Overall, participants demonstrate evidence of significant learning (mean = 72.7%; std = 23.2; $t(17) = 4.15$, $p < 0.0001$) indicating that participants acquired object-specific knowledge. See the bottom panel of Figure 2 for a graphical presentation of the results of this experiment.

12 participants reported evidence of explicit knowledge via the post-test questionnaires. The majority of these reports involved category level knowledge, some with knowledge of specific pairings within these categories (e.g. "particular flower with certain fish" and "maybe bird w/dog, flower w/fish"). A very small number of reports were exclusively at an object level ("white bird with white flower combo" and "black lab, sunflower, etc").

Data were submitted to an ANOVA examining the effects of exemplar typicality (Atypical vs. Typical) and explicit knowledge on test performance. Consistent with the findings mentioned in the introduction, we hypothesized that any contribution of categorical knowledge would be modulated by the typicality of the exemplars. We report no main effect of exemplar typicality ($F(1,14 = 0.307$; $p > 0.5$) nor interaction of typicality and explicit knowledge. The uniform performance across atypical and typical groups, as indicated in Figure 2, suggests no contribution from category-level knowledge in the current experiment.

We do, however, report a marginal effect of explicit knowledge of sequence structure ($F(1,14) = 3.96$; $p < 0.07$). We will address this issue more deeply in the results section of Experiment 2. Given that the current experiment was designed such that categorical knowledge could not be used to distinguish foils from pairs, and most evidence for explicit knowledge came as a report of predictive dependencies involving category level knowledge, it is unclear how explicit knowledge boosts performance. One possibility is that participants who achieve a high level of knowledge also achieve lexical access to the categories. Possibly knowledge of many of the pairs of exemplars induces category-level explicit knowledge.

In sum, participants were exposed to a sequence of pictures containing predictive dependencies redundant at the level of individual object and at a more abstract level of the categories these objects belonged to. Test performance indicates that participants gained object-specific knowledge. In addition, results suggest that participants do not acquire additional knowledge from more abstract, categorical regularities. We hypothesized that if participants do acquire categorical level knowledge, it would be modulated by object typicality. Results indicate no difference in learning between participants who received exposure to typical or atypical exemplars. Failing to find any difference between these groups suggests that participants learned from object-level regularities exclusively.

However, some participants do report explicit knowledge of the sequences. The majority of these reports included and sometimes were exclusive to abstract, object categories. These results indicate some awareness of the abstract properties of the stream. Moreover, we find that explicit knowledge has a marginally significant effect on test performance. It is unclear how explicit knowledge of this kind could aid in performance given that the experiment was designed to tap into object-specific knowledge only. Thus, while Experiment 1 provides strong evidence for object level learning, it does not entirely exclude the possibility that participants acquire some more abstract knowledge. In the second experiment, we more directly examine the possibility that participant learn from both categorical and object level predictive dependencies.

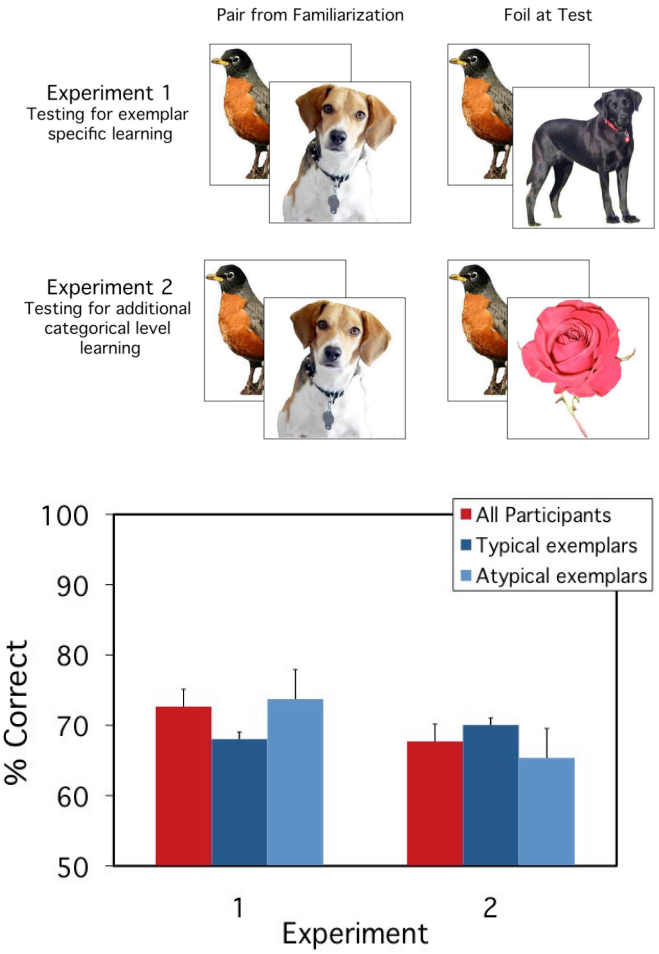


Figure 2: **Top Panel:** The sole difference between experiments was the composition of foils used at test. In Experiment 1, foils were designed to assess learning at the object or exemplar-specific level, while Experiment 2 foils allow for knowledge at both levels of abstraction (object and category) to influence test performance.
Bottom Panel: Results across Experiments 1 and 2 indicate no effect of exemplar typicality or foils on test performance.

Experiment 2: Testing for Additional Category-Level Knowledge

The current experiment addresses whether participants learn from the predictive dependencies at multiple levels of abstraction simultaneously (e.g. objects: $bird_1$-dog_1; abstract, categories: birds-dogs). To this end, we modified the foils used in Experiment 1 while keeping all other aspects of the experiment the same (e.g. $bird_1$-dog_2). The foils in Experiment 1 violated the statistical regularities at the level of individual objects but preserved categorical regularities. Thus, object-specific knowledge but not category level knowledge would be essential in order to distinguish the pair from the foil.

In Experiment 2, we changed the foils to violate both object-level and category-level statistical regularities (e.g. $bird_1$-$flower_3$). Therefore, category knowledge as well as object-specific knowledge could be used at test. If it were the case that participants learn from predictive dependencies at *both* levels of informational abstraction, we hypothesize that it would be easier to distinguish foils in the current experiment, which violate both forms of statistical regularities, compared to the foils used in Experiment 1, where only object-level regularities were violated. However, if participants do not acquire abstract knowledge during familiarization, they will still be able to perform the test in the same manner as Experiment 1. Thus, if participants acquire abstract knowledge, we hypothesize a significant increase in performance in Experiment 2 from Experiment 1, and failure to observe a significant increase in test performance would indicate that learning does not occur at the abstract categorical level.

As in Experiment 1, participants viewed either typical or atypical exemplars. If participants acquire categorical knowledge during familiarization, this knowledge will likely be modulated by the typicality of exemplars. In Experiment 1, we did not observe any asymmetry of performance between these groups; however, categorical knowledge would interfere with test performance in this case. In the current experiment, categorical knowledge would be of benefit. Thus, we hypothesize that, if participants have access to category level knowledge after familiarization, participants who view typical exemplars will have a greater boost in test performance than those who view atypical exemplars.

Methods

Another 24 participants were recruited from the same subject pool and randomly assigned to each condition (16F, 1 left handed, age: mean = 19.6, std = 1.28): 12 viewed the typical pictures, and 12 viewed the atypical pictures. The procedure in this experiment differed from Experiment 1 in only one respect: the foil pairs during the test were designed to violate the statistical structure of the familiarization sequence at the exemplar *and* the category level.

Results and Discussion

We report significant learning overall in Experiment 2 (mean = 67.7%, std = 21.2; $t(23) = 4.10$, $p < 0.0001$). The data were submitted to a two-way ANOVA to evaluate the effects of typicality of exemplars and explicit knowledge. There is a main effect of explicit knowledge ($F(1, 20) = 110.2$; $p < 0.001$) and, as seen in Experiment 1, we report no main effect of typicality ($F(1, 20) = 0.537$, $p > 0.5$) or interaction between these factors. We hypothesized that if categorical knowledge was acquired during exposure that it would be modulated by the typicality of exemplars. The consistent null effect of exemplar typicality indicates that participants do not acquire abstract, category level knowledge during exposure to environmental regularities at multiple levels of abstraction.

We also hypothesized that if participants learned from statistical regularities about objects as well as categories, participants in Experiment 2 would performance better at test than participants in Experiment 1. Results from both experiments were analyzed in a 3-way ANOVA, to test for effects of experiment, typicality, and explicit knowledge on test performance. This analysis confirmed the pattern of results seen in the bottom panel of Figure 2: there is no main effect of Experiment ($F(1,35) = 0.440$, $p > 0.5$). Additionally, we confirm that across both experiments there is no main effect of typicality of exemplars ($F(1,35) = 0.081$, $p > 0.5$) and no interaction between these factors. Thus, test performance is equivalent across experiments indicating that participants likely did not acquire categorical knowledge during exposure to the familiarization stream.

Consistent with results found in both experiments separately, there is a main effect of explicit knowledge: $F(1,35) = 35.9$, $p < 0.001$. Pooling participants from both experiments, we find that participants with explicit knowledge performed better than those without (mean performance: 85.3% vs. 54.4%). However, both groups performed significantly better than chance (with knowledge: $t(20) = 8.07$, $p < 0.001$; without knowledge: $t(20) = 2.21$, $p < 0.02$). Thus, regardless of explicit knowledge there is evidence for learning in both groups.

To determine whether explicit knowledge is related to any of our experimental manipulations (e.g. typicality of exemplars), we examined whether number of participants who demonstrate explicit knowledge is biased towards either a particular experiment (Exp. 1 or 2) or typicality of the objects seen. Of the 42 subjects in both experiments, 21 reported knowledge of the structure, while 21 reported no such knowledge. Chi-square tests show that the proportion of participants who had explicit knowledge of the sequence structure was not significantly different between any of the experimental factors: Experiment 1 vs. Experiment 2: $c2(1, N = 42) = 3.5$, $p > 0.05$; typical vs. atypical: $c2(1, N = 42) = 1.09$, $p > 0.25$. These results indicate that explicit knowledge, while a significant factor affecting performance, is equally distributed across groups and thus should not disproportionally bias overall performance.

Finally, all participants rated both typical and atypical pictures on "interestingness" and typicality. T-tests comparing ratings within categories revealed that participants rate atypical and typical exemplars distinctly and also rate the atypical exemplars as more interesting ($t(134) > 3.5$; $p < 0.001$ within categories for both typicality and "interestingness"). These results validate the assumption that participants view atypical and typical exemplars differently.

Along with Experiment 1, these results support the view that participants learn from statistical regularities at the lowest level of representational abstraction even when more abstract statistical regularities are available to any learning mechanism. Specifically, the results from Experiment 2 cast doubt on the possibility that participants learn from predictive regularities at both levels abstraction.

General Discussion

Humans are able to learn from experience where complex regularities are present. We investigated behavior in a novel learning task designed to investigate a key aspect of the complexity of daily experience: participants viewed streams of pictures with predictive dependencies at multiple levels of abstraction. Specifically, both individual objects or exemplars and the semantic categories that these objects belonged to predicted picture order, thus both object and categorical information could be used determine the structure of the familiarization stream. We consistently find evidence for learning at the lowest level of abstraction: participants respond at test according the predictive dependencies of specific objects or category exemplars and do not show evidence of having learned at the more abstract level of categories even when abstract knowledge could aid test performance. Moreover, we find no modulation of learning by exemplar typicality. These findings suggest that while participants can learn from regularities of categories, they do not learn from more abstract regularities when less abstract, more grounded statistical information is present.

Interestingly, while we systematically find that categorical knowledge has no influence on test performance, some participants acquire explicit knowledge of the categorical knowledge of the sequence. This result is strikingly similar to Brady and Oliva (2008): in their Exp. 3, after viewing streams with regularities present solely at the categorical level, participants were able to perform consistently in a test where pictures were replaced with category labels. In Exp. 4, Brady & Oliva (2008) include regularities at the scene specific or object level in addition to categorical regularities and again find evidence for lexical access. We argue that lexical access results in Exp. 3 and 4 of Brady & Oliva (2008) are similar to the demonstration of abstract level explicit knowledge in the current experiment.

While demonstration of lexical access to categories is interesting and important, we repeatedly show that abstract knowledge does not have a clear effect in test performance, raising questions about the nature and function of this lexical knowledge. To date, there has been no demonstration

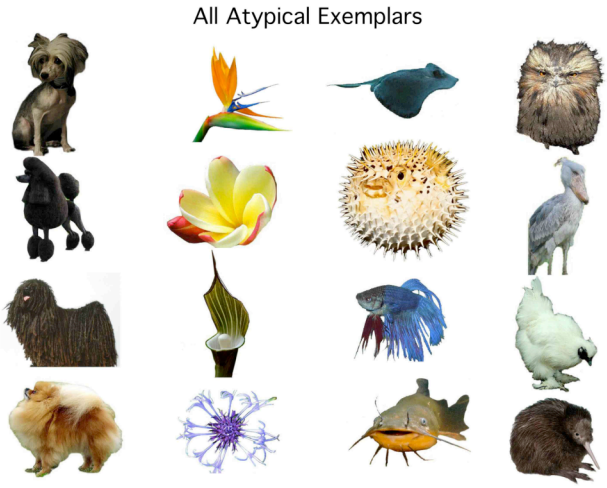
of generalization, an important hallmark of abstract categorical knowledge, when less abstract regularities are present. This is an important avenue for future study to clarify this lexical result. An alternative possibility is that the lexical access is a byproduct of the participant strategy of using mental labels for the familiar objects and scenes as they are being presented. Previous statistical learning studies have been careful to avoid recognizable visual objects for this reason (Conway & Christiansen, 2005; Fiser & Aslin, 2001).

Despite our findings that adults did not learn from abstract, category level regularities when object-based regularities were present, it is nevertheless clear that in a natural environment we do acquire knowledge of higher order regularities. Thus, our results may simply point to a direction of how this learning occurs: learning starts in the relation of specific objects when statistical regularities are comparable at multiple levels of abstraction. One possibility is that when once the least abstract regularities have been mastered, learning can proceed along more abstract dimensions. Nevertheless, this finding may have important implications for more efficient teaching methods and could inform computational modeling of learning and development of human cognitive processes where the abstraction of representation is often an assumption built into the model.

Overall, this study aims to uncover how simple learning mechanisms operate in complex, naturalistic environments. We increased the complexity of the learning task, relative to previous experiments, by having predictive dependencies at multiple levels of abstraction. Results indicate that participants learned based on the more concrete, less abstract predictive dependencies. Results also suggest that participants did not additionally learn the more abstract relationships as this knowledge consistently did not influence test performance. These results inform the on-going debate as to whether domain-general learning mechanisms are largely unconstrained, as previous behavioral studies would have suggested. We believe that these results show some level of constraint on learning where more grounded, less abstract statistical relationships are learned preferentially when categorical and object specific knowledge is redundant.

Acknowledgments

Dr. Dima Amso, Dr. Rick Dale and Jordan DeLong for helpful conversations, Claire Schmidt for data collection and in particular, we thank Dr. Michael Spivey for his support.



Appendix 1: All atypical exemplars used in the current paper, organized by category (from left: dog, flower, fish, bird).

References

Baldwin, D. Anderrson, A, Saffran, J. & Myers, M. (2007). Segmenting dynamic human action via statistical structure. *Cognition, 106,* 1382-1407.

Brady, T. F. & Oliva, A. (2008): Statistical learning using real-world scenes: Extracting categorical regularities without conscious intent, *Psychological Science, 19,* 678-685.

Conway, C. & Christiansen, M. H. (2005). Modality-constrained statistical learning of tactile, visual, and auditory sequences, *Journal of Experimental Psychology: Learning, Memory, & Cognition, 31,* 24-39.

Dale, R., Kehoe, C.E. & Spivey, M. J. (2007). Graded motor responses in the time course of categorizing atypical exemplars. *Memory and Cognition*, 35, 15-28.

Fiser, J. & Aslin, R. A. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological Science, 12,* 499-504.

Jolicoeur, P., Gluck, M. A., & Kosslyn, S. M. (1984). Pictures and names: Making the connection, *Cognitive Psychology, 16,* 243-275.

Marcus, G. F., Fernandes, K.J. & Johnson, S.P. (2007). Infant rule learning facilitated by speech. *Psychological Science, 18,* 387-391.

Saffran, J. R. (2002). Constraints on statistical language learning. *Journal of Memory and Language, 47,* 172-196.

Rosch, E., Mervis, C., Gray, W., Johnson, D., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology, 8,* 382-439.

Adult language learners under cognitive load do not over-regularize like children

Amy Perfors (amy.perfors@adelaide.edu.au)
School of Psychology, University of Adelaide

Nicholas Burns (nicholas.burns@adelaide.edu.au)
School of Psychology, University of Adelaide

Abstract

The "less is more" hypothesis suggests that one reason adults and children differ in their language acquisition abilities is that they also differ in other cognitive capacities: for instance, the relatively poor memory and/or processing abilities of children may make them more likely to over-regularize inconsistent input (Singleton & Newport, 2004; Hudson Kam & Newport, 2005). We investigate this hypothesis by placing adults under a high cognitive load using a standard task. Does their tendency to over-regularize in a simultaneous language-learning task increase? Results indicate that although the cognitive load is high enough to impair overall learning, neither the presence of load nor poor working memory predicts greater over-regularization. This suggests that if the "less is more" hypothesis explains over-regularization in children, the relevant cognitive capacity is not one that was impaired by our load task.
Keywords: language acquisition; over-regularization; statistical learning; memory; processing; development

Introduction

Children and adults differ both qualitatively and quantitatively in their ability to acquire a new language. Adults have difficulty with many aspects of language acquisition, from phonetic perception (Werker & Tees, 1984; Werker & Lalonde, 1988; Kuhl, 2004), to language processing (Clahsen & Felser, 2006), to certain aspects of syntax (e.g., Johnson & Newport, 1989; Birdsong, 2006). Scientists have proposed many theories to account for the difference between children and adults; these theories differ in both the degree and type of contribution made by pre-existing language-specific biases. Although nearly everyone agrees that (due to the inherent logical problem of induction posed by language learning) some bias must be necessary to explain successful language acquisition, explanations about the nature of the bias – and the difference between children and adults – vary considerably.

Some argue that there is a fundamental difference between first- and second-language acquisition. They posit that acquisition in children is guided by an innate Universal Grammar and by language-specific acquisition procedures, whereas adult acquisition is directed by more domain-general learning mechanisms (e.g., Bley-Vroman, 1990). However, there are many other possibilities, since children and adults also differ profoundly in their cognitive capabilities, knowledge, assumptions, and typical linguistic input. For one thing, learning a second language is made more difficult by interference from the first language; the evidence that experience with a first language influences acquisition of a second language is extensive (e.g., Mayberry, 1993; Iverson et al., 2003; Tan, 2003; Weber & Cutler, 2003; Hernandez, Li, & MacWhinney, 2005). This observation overlaps considerably with the related point that adult brains are less malleable than the brains of children (Elman et al., 1996; MacWhinney, 2005). Adults and children also differ in their style of learning (Ullman, 2004) and in the nature of the social support (Snow, 1999) and linguistic input (Fernald & Simon, 1984) they receive.

The observation that children perform more poorly than adults across most domains of cognitive ability, including memory and processing speed, has led to another hypothesis, often called "less is more." It suggests that the relative cognitive deficits in children may actually *help* with language acquisition by enabling them to isolate and analyze the separate components of a linguistic stimulus (Newport, 1988), or by leading them to over-regularize inconsistent input (Hudson Kam & Newport, 2005; Singleton & Newport, 2004). Indeed, it is apparent that children over-regularize while adults often do not. Deaf children exposed to the inconsistent sign language of hearing parents will over-regularize that language and produce regular grammatical forms (Singleton & Newport, 2004), as will children exposed to inconsistent input in an artificial language (Hudson Kam & Newport, 2005; Goldowsky, 1995). By contrast, adult language learners are known to produce highly variable, inconsistent utterances, even after years of experience with the language and after their grammars have stabilized (Wolfram, 1985; Johnson, Shenkman, Newport, & Medin, 1996).

The difference between children and adults has also been found in non-linguistic domains. If adults must predict some phenomenon (e.g., a light flashing or a certain card being drawn from a deck), they will tend to probability match: if the phenomenon occurs 70% of the time, they will expect it 70% of the time they are asked (see, e.g., Myers, 1976; Shanks, Tunney, & McCarthy, 2002, for an overview). Children are more likely to predict that the phenomenon will occur closer to 100% of the time (e.g., Weir, 1964; Derks & Paclisanu, 1967). A similar pattern has been found in causal reasoning: children over-regularize by assuming that causes are deterministic, while adults do not (Schulz & Sommerville, 2006).

Although the tendency toward over-regularization is well-established, the reason for the difference between adults and children is far from clear. As previously mentioned, the "less is more" hypothesis suggests that over-regularization may be due to some aspect of children's cognitive capacities, such as their poorer memory or slower processing speed (Newport, 1988). Adults do tend to over-regularize more when the input is complex, when the probabilities involved are small (Gardner, 1957; Weir, 1964; Gluck & Bower, 1988; Hudson Kam & Newport, 2009), or when lexical retrieval is more difficult (Hudson Kam & Chang, 2009). This may be because

more complex input imposes more of a load on their cognitive resources. The hypothesis is also supported by empirical (Kersten & Earles, 2001) and computational (Elman, 1993) work suggesting that learning is easier when early input is simpler (although that work does not speak directly to the issue of over-regularization). In general, there has been little research that directly measures or manipulates memory or processing speed and evaluates whether these are associated with different degrees of over-regularization in adults.

Here we begin to investigate this question more directly. Our goal is to evaluate whether we can effectively turn adults into children by placing them under cognitive load. If deficiencies in the particular capacities involved in the load tasks are what cause children to over-regularize, then adults under heavy load should behave more like children in their pattern of over-regularization. We find that, although the cognitive load is high enough to impair adult performance in other ways – and although their working memory capacity predicts overall performance on the task – neither increased cognitive load nor poor working memory predicts or leads to increased over-regularization. This suggests that, if the "less is more" hypothesis is the explanation for childrens' tendency to over-regularize, the cognitive capacity that is "less" in children is not one that is impaired by the load tasks we used.

Method

75 adults were recruited from the University of Adelaide and surrounding community and were paid $10 for their participation. In the first part of the experiment, individual differences in working memory capacity were measured using a standard complex span task (Conway, Jarrold, Kane, Miyake, & Towse, 2007; Unsworth, Redick, Heitz, Broadway, & Engle, 2009). In the second part of the experiment, subjects completed a word-learning task (modelled on the paradigm described by Hudson Kam and Newport (2009)) in which they were taught 10 two-word labels from a new language. Interspersed with the word-learning task, participants in the OPERATIONAL LOAD and VERBAL LOAD conditions completed an interference task (involving either solving equations or reading sentences aloud, respectively). In a control condition, the NO LOAD condition, participants performed the word-learning task only. Specific details of the initial complex span task and the subsequent word-learning task follow.

Complex span task

Complex span tasks are widely used to measure the capacity of the working memory system (Conway et al., 2005; Unsworth et al., 2009). In a complex span task, items to be remembered (e.g., random letters, digits, shapes, or spatial locations) are interspersed with an unrelated cognitive activity (e.g., solving equations, reading sentences, or evaluating the symmetry of patterns). After several trials, participants are asked to recall the items to be remembered in the correct serial order. This sort of task is differentiated from a simple span task (e.g., Digit Span from the Wechsler scales), which only includes the memorization component; it has been argued that complex span tasks provide a measure of working memory (as opposed to span memory) because they entail the requirement to process as well as to store information. Complex span tasks have been shown to correlate with cognitive processes that are believed to depend on working memory (Conway et al., 2007; Unsworth & Engle, 2007), and are linked to disorders including Alzheimer's disease (Rosen, Bergeson, Putnam, Harwel, & Sunderland, 2002). They have also been widely used to explore age differences in working memory capacity (Case, Kurland, & Goldberg, 1982; Salthouse & Babcock, 1991).

Two common span tasks incorporate demands on either operational span (Turner & Engle, 1989) or on verbal span (Daneman & Carpenter, 1980), respectively. In an operational span task, participants are presented with equations such as $4/2+2=3$ and told to say, as quickly as possible, whether the equation is correct. In a typical verbal span task, subjects are presented with an 11-15 word sentence and told to say, as quickly as possible, whether the sentence makes sense. In order to enable comparison across participants, in the first part of the experiment all participants were presented with an operational span task regardless of condition. On each trial people first saw an equation and were asked whether it was correct or not. After each response, a random letter was shown. At the end of a set of *n* letters, participants were asked to repeat the list of letters in order, given unlimited time to do so. To make sure that they understood the task, they were first trained on two sets of two trials each. The full task comprised two sets each of sizes ranging from an *n* of three to an *n* of seven, for a total of 50 trials. For each participant a working memory capacity score was calculated, reflecting the number of correct letters recalled in the correct position.

Word-learning task

After the complex span task, all participants took part in an artificial language learning task modelled after a similar task described by Hudson Kam and Newport (2009). Their language contained 51 words, including 36 nouns and 12 verbs, among other lexical items, taught over the course of eight separate sessions extending for 9-12 days. Of critical interest in their study was the evaluation of performance on the determiners, which were associated with nouns in an inconsistent fashion: participants heard the main determiner only 60% of the time. In one condition, they heard nothing the other 40% of the time; in four other conditions, they heard increasingly more *noise* determiners (e.g., two determiners (each 20% of the time), and so forth up to 16 determiners (each 2.5% of the time)). Performance was measured in a sentence completion task in which participants had to provide the noun and determiner associated with a scene and sentence.

We sought to remove extraneous elements of the task so as to focus on the determiner-production aspect while still retaining the important details. We therefore presented participants with a "language" of 10 nouns, all two-syllable non-

sense words[1] mapped to images representing common objects.[2] Each noun was followed by a one-syllable determiner:[3] the *main* determiner occurred 60% of the time, and each of the four *noise* determiners occurred 10% of the time. The specific mapping of the word to the meaning and which determiner was the *main* determiner were randomized for each participant.

Over the course of the task, participants saw 200 trials of image-label pairs. On each trial, an image appeared on the computer screen and, at the same time, the person heard a female voice provide the label: for instance, they might see a picture of a baby and hear churbit mog. In the NO LOAD condition, participants went to the next trial by clicking a next button; in the two load conditions, the image remained visible for 1.5 seconds and then the next phase of the trial began automatically (as explained below). In all conditions, learning was tested with 10 questions every 50 trials, for a total of 40 test questions. At each test, the participant was presented with an image and asked to verbally produce the label for it, which the experimenter wrote down. No feedback was given.

Subjects in the two load conditions completed the same word learning task, except that after each image-label pair, they were asked to perform an unrelated task designed to increase their cognitive load. In the OPERATIONAL LOAD condition, the task was modelled after the operational span test (Turner & Engle, 1989): participants were presented with an equation and told to respond as quickly as possible whether it was correct or not. Half of the equations were correct, and half gave an answer that was one digit away from correct. In order to encourage them to be as fast and correct as possible, a running total of their number correct and elapsed time was displayed on the screen. In the VERBAL LOAD condition, the task was modelled after the verbal span test (Daneman & Carpenter, 1980): participants were presented with an 11-15 word sentence, told to read it aloud, and then asked to respond as quickly as possible whether it was sensible or not. Half of the sentences were sensible, and half were made non-sensible by replacing a content word with a semantically inappropriate one.[4] As before, accuracy and elapsed time was displayed in order to encourage peak performance.

Results

There are three natural questions we must answer in order to properly understand this experiment. First, is the load task difficult enough? Second, did participants in either of the load conditions over-regularize by producing the *main* determiner more than 60% of the time? Third, did individual differences in performance on the initial complex span task predict performance on the word-learning task? The answer to the first question is an essential pre-requisite to interpreting the answers to the other two because if the load task was not challenging enough, comparisons between conditions are meaningless. The answers to the other two bear directly on the questions motivating this work: does putting adults under cognitive load cause them to make the same over-regularization errors that children do? Are adults with poorer performance on the complex span task (and hence lower working memory capacity) more likely to make those errors? We address each of these questions in turn.

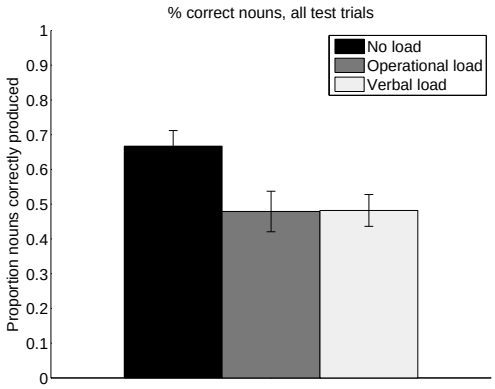


Figure 1: Performance by condition in the noun-learning task. Participants in the two load conditions learned significantly fewer nouns, indicating that the load task provided sufficient cognitive challenge to impair performance.

Was the load task difficult enough?

There are several ways to evaluate whether the load tasks were sufficiently challenging to the cognitive capacities of our participants, whilst still being easy enough so that people could acquire at least some of the image-label mappings in the word-learning task. One indication is that participants in both conditions scored far above chance on the load items, suggesting that they took that task seriously.

To evaluate the degree of difficulty the tasks imposed, we can compare how well participants in each of the three conditions learned the correct noun-image mappings. One would expect that performance would be substantially worse in the two load conditions if the secondary task provided a sufficient challenge to the cognitive capacities of our participants. To explore this, we coded each person's answers as *correct* if the noun they produced was identical to or phonologically similar (e.g., wolin instead of wolid) to the correct noun for that image. Figure 1 demonstrates that participants in both load conditions got fewer nouns correct than in the NO LOAD condition, indicating that the interference tasks were, indeed, imposing significant strain on their cognitive resources. There was no difference in the number of nouns correct between the OPERATIONAL LOAD and VERBAL LOAD conditions.[5]

[1] Noun words used were: dragnip, raygler, churbit, tramdel, shelbin, pugbo, wolid, foutray, nipag, and yeetom.

[2] Objects used were: babies, balls, beds, birds, books, cars, cats, cups, dogs, and shoes.

[3] The five determiners were: mot, ped, sib, kag, and zuf.

[4] For example, a typical sentence is "Cats really love to sit in the sun, since they are desert animals" while the corresponding non-sensible sentence would replace animals with chimneys.

[5] A one-way Anova on nouns correct by condition was significant: $F(2,72) = 4.63, p = 0.0129$. Post-hoc comparisons using the Tukey-Kremar test indicated that the mean score for the NO LOAD condition (M=0.667, SD = 0.05) was significantly different than the

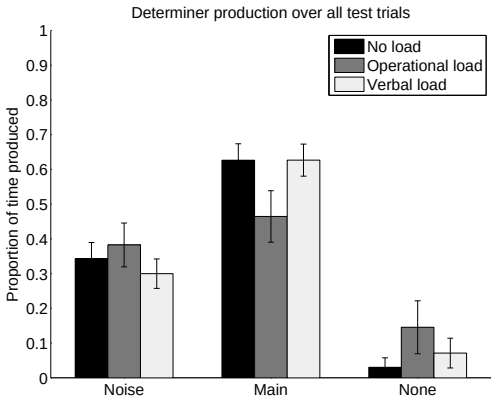


Figure 2: Performance by condition in determiner production. There was no significant difference between conditions in tendency to over-regularize, and in no condition did people produce the main determiner beyond the 60% it appeared in the input.

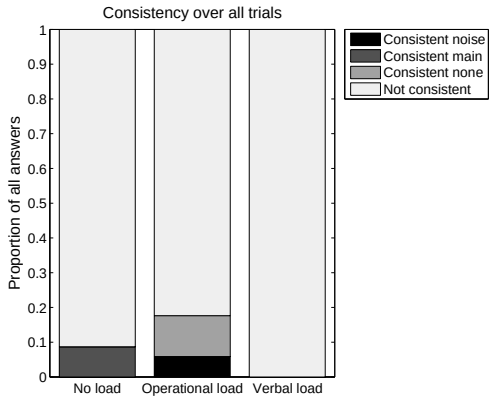


Figure 3: Individual consistency in determiner production by condition. For the most part, few participants showed any consistency in their pattern of determiner usage, and those in the load conditions did not tend to be more consistent.

Did adults over-regularize more when under cognitive load?

The central question motivating this research was whether adults placed under cognitive load could be made to look more like children. To evaluate this, following Hudson Kam and Newport (2009), we excluded all participants who did not get at least 9 out of the final 20 nouns correct on the test trials.[6] Then, on every valid trial (i.e., every trial for which a correct noun was produced), we calculated the percentage of time either the *main* determiner, a *noise* determiner, or *no* determiner was produced. Figure 2 demonstrates that there were no significant differences between conditions in terms of *main* determiner production: that is, participants in the load conditions did not over-regularize.[7] If anything, participants in the OPERATIONAL LOAD condition tended to *under*-regularize, which is the opposite of what one would expect if limited available memory or processing power was the driving force behind over-regularization.

This is suggestive, but because it is an analysis of mean performances this outcome may be hiding individual over-regularization in different directions. To evaluate this possibility, we followed Hudson Kam and Newport (2009) and set a "consistency threshold" of 90%: each participant was coded as *consistent main*, *consistent noise*, or *consistent none* if they produced the determiner type in question on at least 90% of the valid trials, and *not consistent* if they did not.[8] Figure 3 shows that few participants were consistent in any way, and differences between conditions were minor. In order to determine if the tendency to over-regularize changed as they acquired more of the language, we repeated the analyses shown in both Figure 2 and 3 at each of the four stages of testing. There were no differences in behavior at any stage.

Does working memory span have any effect on performance?

The results presented thus far suggest that people with less available working memory capacity (i.e., those in the two load conditions) did not over-regularize the main determiner more than did those in the control condition. Our experiment also provides another way to evaluate how working memory capacity affects over-regularization: by analyzing whether individual differences in performance on the initial complex span task predicts differential performance on the word-learning task. As one would expect, performance on the complex span task is positively and significantly correlated with the ability to learn the noun-image mappings ($\rho = 0.3811, p = 0.0013$): participants with greater working memory capacity learned more noun labels. However, there is no relation between working memory capacity and the tendency to produce the main determiner ($\rho = 0.1066, p = 0.387$), nor do the scatterplots indicate a non-linear relationship.

Discussion

On first glance, our findings might appear to contradict those of Hudson Kam and Chang (2009), who found that over-regularization in adults could be diminished by improving the ease of lexical retrieval. There are three notable differences here. First, they aimed to make adults *less* like children by making the cognitive load easier, rather than to make adults act *more* like children by making it harder. It is possible that there is an inherent asymmetry to adults' performance: that it is relatively easy to make adults over-regularize less, but that getting them to regularize more is difficult. This is certainly the case in the decision-making literature, in which great efforts have been made to stop adults from probability match-

mean for the OPERATIONAL LOAD ($M = 0.479, SD = 0.05$) and VERBAL LOAD ($M = 0.482, S = 0.05$) conditions, but the latter two were not significantly different from each other.

[6] This resulted in 23 subjects in the NO LOAD condition and 17 in each of the others. We ran each of these analyses without this exclusion and results were qualitatively identical in all cases.

[7] One-way Anova on main determiner production by condition: $F(2,54) = 2.64, p = 0.0806$. To further explore this outcome, a post-hoc comparison using Tukey-Kramer indicated no significant difference between any of the conditions compared pairwise.

[8] Results are qualitatively identical even with thresholds of 70% or 80%: there are more consistent participants in those cases, but still no difference between conditions.

ing (e.g., Shanks et al., 2002). Second, and more importantly, the study by Hudson Kam and Chang (2009) examined a different aspect of cognitive load (lexical retrieval rather than working memory capacity). It is possible that differences in lexical retrieval abilities are related to differences in over-regularization between children and adults, but that differences impaired by our load task were not. Third, our language was far simpler than theirs; it is possible that our participants treated the task like paired-associate learning rather than like learning a language with rich internal structure, unlike adult learners in other studies that tasked load (Pitts Cochran, McDonald, & Parault, 1999). We think this would be a rather surprising explanation of our findings, given that the task itself (learning determiner-noun pairings) was the same in both studies, and the main difference was the complexity of the rest of the system they were embedded in; however, it is an open question that we seek to resolve with future work.

The central issue, of course, is what abilities *were* impaired by our load task? In many ways, the two load tasks were quite different: one involved solving equations, while the other involved reading sentences aloud and answering questions about them. Despite this, it has been shown that the complex working memory tests related to these tasks tend to load highly on the same broad working memory factor (e.g., Oberauer, Süß, Wilhelm, & Wittmann, 2003). It may therefore not be a surprise that both load tasks had similar effects. The interesting aspect of this is that these tasks were specifically designed to create a load on multiple different cognitive capacities at once: unlike simple span tasks (such as Digit Span on the Wechsler), which capture only the storage component of memory, these require processing as well. In general, these load tasks should be disrupting many aspects of cognition: among other things, they require people to retrieve information from long-term memory (word meanings in the VERBAL LOAD condition, number and symbol meanings in the OPERATIONAL LOAD condition), to store information in short-term memory (the words in the current sentence or numbers in the current equation), to manipulate representations (to determine the correct answer to the questions), to regulate attention, and to perform the load task while simultaneously learning word-referent mappings. It is interesting that, despite their generality, the load tasks still did not lead to over-regularization in word learning.

How might we interpret these results? One possibility is that the "less is more" hypothesis is incorrect: that children's tendency to over-regularize does not stem from differences in cognitive capacity. Such a possibility is consistent with previous studies finding no effect of load on adult learners (Ludden & Gupta, 2000) as well as other empirical findings in language acquisition showing that children with better memories or faster processing speed actually do *better* at learning language (e.g., Fernald, Perfors, & Marchman, 2006; Rose, Feldman, & Jankowski, 2009).

That said, we cannot be certain that "less is more" is incorrect. It is in theory possible that our load tasks did not sufficiently challenge our subjects enough, and that more difficult ones would result in more over-regularization. This is unlikely, not only because the participants anecdotally seem to have found the task extremely difficult (one person called it the hardest psychology experiment he had ever done), but also because the load tasks had such strong effects on the ability to learn the nouns in the first place. The task would somehow have to be difficult enough to cause over-regularization but not so difficult as to render the task impossible: a balancing act that, if nothing else, seems unlikely to precisely describe the state of child language learners.

This point, however, raises the converse possibility: perhaps our language-learning task was so difficult (such that even in the no-load condition, participants were only about 70% correct overall[9]) that with longer training, the pattern we observed might change. While always a possibility, we think this is more unlikely than other explanations, since we observed no detectable change in tendency to over-regularize over the course of the experiment.

Another possibility is that, because our load task items were interspersed rather than concurrent with the words to be learned, it was less of a burden on concurrent memory and processing speed, and more of a burden on executive control. If so, this would suggest that the differential abilities between children and adults is not due to cognitive control, as has been suggested in a different context (Thompson-Schill, Ramscar, & Chrysikou, 2009). We plan to explore this issue in future work using a concurrent load task like verbal shadowing.

Even if our load task does impair memory and processing speed, there remain some likely possibilities for how the "less is more" hypothesis might be correct and still be consistent with our results. In addition to memory and processing speed, children and adults also differ in the ability to use metacognitive strategies (e.g., Flavell, Green, Flavell, Harris, & Astington, 1995). It may be that adults' ability to introspect and reason about their own cognition makes them more likely to rely on explicit rather than implicit learning (Ullman, 2004) – a difference that has been hypothesized to be the root of child-adult differences in language acquisition. Such metacognitive ability might also make adults more likely to try to capture or imagine patterns in the input that do not exist; this tendency has been suggested as an explanation for why adults probability match in non-language tasks (Estes, 1976). It might result from a generalized preference for simplicity (or tendency to ignore exceptions) on the part of children. It is also possible that having limited memory or processing abilities is especially important for language learning *as a child* but not as an adult, analogously to a similar hypothesis found in other developmental domains (Turkewitz & Kenny, 1982). A great deal of work remains to be done to investigate the many possibilities that remain open.

[9]Keeping in mind that, since there were 10 objects and it was a free-response task, this is actually far above chance performance.

Acknowledgments

We thank Natalie May for her invaluable help recruiting participants and running the experiment.

References

Birdsong, D. (2006). Age and second language acquisition and processing: A selective overview. *Lang. Learning*, *56*(1), 9–49.

Bley-Vroman, R. (1990). The logical problem of foreign language learning. *Linguistic Analysis*, *20*, 3–49.

Case, R., Kurland, D., & Goldberg, J. (1982). Operational efficiency and the growth of short-term memory span. *Jn. of Exp. Child Psych.*, *33*, 386–404.

Clahsen, H., & Felser, C. (2006). How native-like is non-native language processing? *TiCS*, *10*(12), 564–570.

Conway, A., Jarrold, C., Kane, M., Miyake, A., & Towse, J. (2007). *Variation in working memory*. NY: Oxford Univ. Press.

Conway, A., Kane, M., Bunting, M., Hambrick, D., Wilhelm, O., & Engle, R. (2005). Working memory span tasks: A methodological overview and user's guide. *Psych. Bull. & Rev.*, *12*, 769–786.

Daneman, M., & Carpenter, P. (1980). Individual differences in working memory and reading. *Jn. of Verbal Learning & Verbal Behavior*, *19*, 450–466.

Derks, P., & Paclisanu, M. (1967). Simple strategies in binary prediction by children and adults. *Jn. Exp. Psych.*, *73*(2), 278–285.

Elman, J. (1993). Learning and development in neural networks: The importance of starting small. *Cognition*, *48*, 71–99.

Elman, J., Bates, E., Johnson, M., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking innateness: A connectionist perspective on development*. Cambridge, MA: MIT Press.

Estes, W. (1976). The cognitive side of probability learning. *Psych. Review*, *83*, 37–64.

Fernald, A., Perfors, A., & Marchman, V. (2006). Picking up speed in understanding: Speech processing efficiency and vocabulary growth across the 2nd year. *Dev. Psych.*, *42*(1), 98–116.

Fernald, A., & Simon, T. (1984). Expanded information contours in mothers' speech to newborns. *Dev. Psych.*, *20*, 104–113.

Flavell, J., Green, F., Flavell, E., Harris, P., & Astington, J. W. (1995). Children's knowledge about thinking. *Monographs of the SRCD*, *60*(1).

Gardner, R. (1957). Probability-learning with two and three choices. *American Jn. of Psych.*, *70*, 174–185.

Gluck, M., & Bower, G. (1988). From conditioning to category learning: An adaptive network model. *Jn. of Exp. Psych.: Gen.*, *117*, 227–247.

Goldowsky, B. (1995). *Learning structured systems from imperfect information*. PhD dissertation, University of Rochester.

Hernandez, A., Li, P., & MacWhinney, B. (2005). The emergence of competing modules in bilingualism. *TiCS*, *9*(5), 219–224.

Hudson Kam, C., & Chang, A. (2009). Investigating the cause of language regularization in adults: Memory constraints or learning effects? *Jn. of Exp. Psych.: Lng., Mem., & Cog.*, *35*(3), 815–821.

Hudson Kam, C., & Newport, E. (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Lang. Lng. & Dev.*, *1*(2), 151–195.

Hudson Kam, C., & Newport, E. (2009). Getting it right by getting it wrong: When learners change languages. *Cog. Psych.*, *59*, 30–66.

Iverson, P., Kuhl, P., Akahane-Yamada, R., Diesch, E., Tokura, Y., Kettermann, A., et al. (2003). A perceptual interference account of acquisition difficulties with non-native phonemes. *Cognition*, *87*, B47–B57.

Johnson, J., & Newport, E. (1989). Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. *Cog. Psych.*, *21*, 60–99.

Johnson, J., Shenkman, K., Newport, E., & Medin, D. (1996). Indeterminacy in the grammar of adult language learners. *JML*, *35*, 335–352.

Kersten, A., & Earles, J. (2001). Less really is more for adults learning a miniature artificial language. *JML*, *44*, 250–273.

Kuhl, P. (2004). Early language acquisition: Cracking the speech code. *Nature Reviews Neuroscience*, *5*, 831–843.

Ludden, D., & Gupta, P. (2000). Zen in the art of language acquisition: Statistical learning and the less is more hypothesis. *22nd Annual Conference of the Cognitive Science Society*.

MacWhinney, B. (2005). A unified model of language acquisition. In J. Kroll & A. De Groot (Eds.), *Handbook of bilingualism: Psycholinguistic approaches* (pp. 49–67). Oxford Univ. Press.

Mayberry, R. (1993). First-language acquisition after childhood differs from second-language acquisition: The case of american sign language. *Jn. of Speech and Hearing Res.*, *36*, 1258–1270.

Myers, J. (1976). Probability learning and sequence learning. In W. Estes (Ed.), *Handbook of learning and cognitive processes: Approaches to human learning and motivation* (pp. 171–205). Hillsdale, NJ: Erlbaum.

Newport, E. (1988). Constraints on learning and their role in language acquisition: Studies of the acquisition of American Sign Language. *Language Sciences*, *10*, 147–172.

Oberauer, K., Süß, H.-M., Wilhelm, O., & Wittmann, W. (2003). The multiple faces of working memory – storage, processing, supervision, and coordination. *Intelligence*, *31*, 167–193.

Pitts Cochran, B., McDonald, J., & Parault, S. (1999). Too smart for their own good: The disadvantage of a superior processing capacity for adult language learners. *JML*, *41*, 30–58.

Rose, S., Feldman, J., & Jankowski, J. (2009). A cognitive approach to the development of early language. *Ch. Dev.*, *80*(1), 134–150.

Rosen, V., Bergeson, J., Putnam, K., Harwel, A., & Sunderland, T. (2002). Working memory and apolipoprotein E: What's the connection? *Neuropsychologia*, *40*, 425–443.

Salthouse, T., & Babcock, R. (1991). Decomposing adult age differences in working memory. *Dev. Psych.*, *27*, 763–777.

Schulz, L., & Sommerville, J. (2006). God does not play dice: Causal determinism and preschoolers' causal inferences. *Ch. Dev.*, *77*(2), 427–442.

Shanks, D., Tunney, R., & McCarthy, J. (2002). A re-examination of probability matching and rational choice. *Jn. of Behavioral Decision Making*, *15*, 233–250.

Singleton, J., & Newport, E. (2004). When learners surpass their models: The acquisition of American Sign Language from inconsistent input. *Cog. Psych.*, *49*, 370–407.

Snow, C. (1999). Social perspectives on the emergence of language. In B. MacWhinney (Ed.), *The emergence of language* (pp. 257–276). Mahwah, NJ: Lawrence Erlbaum Associates.

Tan, L. (2003). Neural systems of second language reading are shaped by native language. *Human Brain Mapping*, *18*, 158–166.

Thompson-Schill, S., Ramscar, M., & Chrysikou, E. (2009). Cognition without control: When a little frontal lobe goes a long way. *Curr. Dir. in Psych. Sci.*, *18*(5), 259–263.

Turkewitz, G., & Kenny, P. (1982). Limitations on input as a basis for neural organization and perceptual development: A preliminary theoretical statement. *Dev. Psychobiol.*, *15*(4), 357–368.

Turner, M., & Engle, R. (1989). Is working memory capacity task dependent? *JML*, *28*, 127–154.

Ullman, M. (2004). Contributions of memory circuits to language: The declarative/procedural model. *Cognition*, *92*, 231–270.

Unsworth, N., & Engle, R. (2007). The nature of individual differences in working memory capacity: Active maintenance in primary memory and controlled search from secondary memory. *Psych. Review*, *114*, 104–132.

Unsworth, N., Redick, T., Heitz, R., Broadway, J., & Engle, R. (2009). Complex working memory span tasks and higher-order cognition: A latent-variable analysis of the relationship between processing and storage. *Memory*, *17*(6), 635–654.

Weber, A., & Cutler, A. (2003). Lexical competition in non-native spoken word recognition. *Jn. Mem. & Lang.*, *50*, 1–25.

Weir, M. (1964). Developmental changes in problem-solving strategies. *Psych. Review*, *71*, 473–490.

Werker, J., & Lalonde, C. (1988). Cross-language speech perception: Initial capabilities and developmental change. *Dev. Psych.*, *24*(5), 672–683.

Werker, J., & Tees, R. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, *7*, 49–63.

Wolfram, W. (1985). Variability in tense marking: A case for the obvious. *Language Learning*, *35*, 229–253.

Descriptive Assessment of Jeffrey's Rule

Jiaying Zhao (jiayingz@princeton.edu)
Department of Psychology, Green Hall,
Princeton University, NJ 08540 USA

Daniel Osherson (osherson@princeton.edu)
Department of Psychology, Green Hall,
Princeton University, NJ 08540 USA

Abstract

Jeffrey (1983) proposed a generalization of conditioning as a means of updating probability distributions when new evidence drives no event to certainty. His rule requires the stability of certain conditional probabilities through time. We tested this assumption ("invariance") from the psychological point of view. In Experiment 1 participants offered probability estimates for events in Jeffrey's candlelight example. Two further scenarios were investigated in Experiment 2, one in which invariance seems justified, the other in which it does not. Results were in rough conformity to Jeffrey (1983)'s principle.

Keywords: Jeffrey's rule; invariance; probability updating;

Introduction

Consider an idealized agent whose beliefs are represented by a (subjective) probability distribution Pr_1 over an outcome space $\mathcal{S}$. Let $B \subseteq \mathcal{S}$ be such that $Pr_1(B) > 0$ and suppose that experience intervenes to convince the agent that B is certainly true. What probability distribution Pr_2 should represent the agent's new state of belief? The Bayesian answer (Hacking, 2001, Ch. 15) identifies Pr_2 with the result of conditioning Pr_1 on B, that is, $Pr_2(\cdot) = Pr_1(\cdot|B)$. Much can be said in favor of the latter equation from the normative perspective. For example, it follows from compelling axioms on belief change (Gardenfors, 1988, §5.2), and its violation exposes the agent to sure-loss betting contracts (Harman, 1999, §4.12). Updating has also been examined from the psychological perspective with focus on the use of Bayes' Theorem to compute conditional probability (see Stanovich (2010, Ch. 3)).

Conditioning is not always suited, however, to represent the impact of new information. In particular, Jeffrey (1983, §11.1) notes that the passage of experience need not raise the probability of any event to one. He gives the example of examining cloth by faint candlelight. The cloth's potential colors might correspond to different events over $\mathcal{S}$ but none may become certain as a result of the examination. Nor is it feasible to augment $\mathcal{S}$ to include visual sensations, with the idea of setting one of them to unity. Such sensations are too difficult to express and individuate. Instead, says Jeffrey, "the best we can do is to describe, not the quality of the visual experience itself, but rather its effects on the observer," for example, that the probability of blue has shifted to .75 from its original value.

To fill in the rest of Pr_2 after experience has set the value of $Pr_2(B)$, Jeffrey relies on the law of total probability. Let $G \subseteq \mathcal{S}$ be given, and suppose that $0 < Pr_2(B) < 1$. Then:

$$(1) \quad Pr_2(G) = Pr_2(G|B)Pr_2(B) + Pr_2(G|\overline{B})Pr_2(\overline{B}).$$

If experience has not influenced the conditional probability of G given B nor that of G given $\overline{B}$ then *invariance* is said to hold (Jeffrey, 2004, §3.2). That is:

$$(2) \quad Pr_2(G|B) = Pr_1(G|B) \quad \textit{and} \quad Pr_2(G|\overline{B}) = Pr_1(G|\overline{B}).$$

Substituting (2) into (1) yields:

$$(3) \quad Pr_2(G) = Pr_1(G|B)Pr_2(B) + Pr_1(G|\overline{B})Pr_2(\overline{B}).$$

(3) is known as "Jeffrey's rule." It shows how change in the probability of B is propagated to G, without experience directly affecting G. It is straightforward to generalize (3) to finer partitions, in place of the binary partition $B, \overline{B}$.

Assuming invariance, it is easy to show that (3) defines a genuine probability distribution Pr_2, and in the special case of $Pr_2(B) = 1$, that it agrees with conditionalization. Moreover, Williams (1980) proves that Pr_2 as given by (3) is the closest distribution to Pr_1 that yields the new probability of B, where "closeness" is measured by cross-entropy with respect to Pr_1. The normative status of Jeffrey's rule has nonetheless been questioned because successive uses produce distinct distributions depending on the order in which events are considered (Doring, 1999). In our view, such doubts disappear upon closer inspection of the evidential weight of probability judgments (Wagner, 2002; Osherson, 2002).

It is important to observe that Invariance (2) is not normatively justified in every situation. Sometimes the conditional probabilities are shifted by experience. For example, let G represent vigorous growth of a potted plant, and let B represent the decision to place it in the bedroom. Then noticing ample sunshine in the bedroom would increase not just the probability of B but also the probability of G given B. In contrast, if all you notice is the absence of plants in the bedroom then the probability of B increases without a change in the probability of G given B, yielding invariance.

To decide whether invariance is warranted in a given situation, we rely on an observation due to Pearl (1988). Given an experience e that intervenes between times 1 and 2, we expect invariance to hold if at time 1, G is conditionally independent of e given B, that is:

$$(4) \quad \textit{Conditional independence of } G \textit{ from } e$$

given B: $Pr_1(G \mid B, e) = Pr_1(G \mid B)$.
To see that invariance depends on (4), observe that $Pr_2(GB) = Pr_1(GB, e)$ since e is what transpires between times 1 and 2, and the latter term equals $Pr_1(GB)$ just in case (4) holds.

The focus of the present paper is whether invariance is descriptively accurate when mandated normatively. In Experiment 1 undergraduates offer probabilities for events in Jeffrey's candle example. Of particular interest is the extent to which invariance is honored. Two further scenarios are then investigated, one in which invariance seems justified, the other in which it does not. The results show modest deviations from invariance where it seems justified normatively.

Experiment 1

Participants

Ninety-six undergraduates from Princeton University participated in exchange for partial course credit (41 female, mean age 19.4 yrs, SD = 1.0).

Materials

We simulated Jeffrey's candle example by having participants examine colored paper cards with a dim flashlight. There were 12 blue cards and 38 purple cards. Each card was marked with either a hippopotamus or a giraffe on one side. Of the blue cards, eight were marked with a hippo and four with a giraffe. Of the purple cards, 24 were marked with a hippo and 14 with a giraffe. We chose *giraffe* and *blue* as the categories G and B evoked in the Introduction. Table 1 summarizes the objective probabilities figuring in the experiment.

Procedure

Sixty-four participants performed the *experimental* condition, and 32 performed the *control* condition. The purpose of the control condition was to assess the impact of being asked to evaluate the same probabilities a second time. In the experimental condition, the experimenter first shuffled the cards and showed each to the participant. Then the experimenter turned away from the participant, drew one card from the shuffled deck, and put it in her pocket. The draw appeared to be random but in fact was guaranteed across participants to deliver equal numbers of blue and purple cards (for statistical purposes). The participant was informed that the card was randomly chosen, and then answered the following questions about the card, via computer interface (the order was randomized for each participant):

PROBABILITY QUESTIONS:

$Pr(G)$	What's the probability that there is a giraffe on the card?
$Pr(B)$	What's the probability that the card is blue?
$Pr(G\|B)$	What's the probability that there is a giraffe on the card assuming that the card is blue?
$Pr(G\|\overline{B})$	What's the probability that there is a giraffe on the card assuming that the card is purple?
$Pr(B\|G)$	What's the probability that the card is blue assuming that there is a giraffe on the card?
$Pr(B\|\overline{G})$	What's the probability that the card is blue assuming that there is a hippo on the card?

The estimates $Pr(B|G)$ and $Pr(B|\overline{G})$ served as a contrast with $Pr(G|B)$ and $Pr(G|\overline{B})$. Given our procedure (see below), the former estimates were not expected to be invariant across the flashlight experience whereas the latter were.

The participant was then informed that s/he would briefly see the card under dim light. The card would be placed face down on the table so that the participant would not see the animal but only the color of the card. The experimenter then turned off the lights in the room, moved the card from her pocket to the table, and flashed the light for about one second. The card was then returned to the experimenter's pocket, and the participant answered the same set of questions shown above (in a different random order). Since participants had to give their estimates twice to the same questions, we informed them that they were free to provide the same estimate or a different estimate the second time around.

In the control condition, the procedure was the same except that the light was applied to the chosen card immediately after its draw. Participants in the control condition thus answered the questions shown above just once, after briefly seeing the card under the dim light.

Results

Average responses. We separately analyzed results for participants exposed to blue cards (Blue group) and those exposed to purple cards (Purple group). In the experimental condition we use Pr_1 to refer to probability estimates before the light experience and Pr_2 for estimates after the experience. For each condition and each color group, we averaged Pr_1 and Pr_2 estimates. The results are shown in Table 1.

Table 1: Average probably estimates by participants and objective probabilities in Experiment 1 (standard deviations in parentheses).

	$Pr(G)$	$Pr(B)$	$Pr(G\|B)$
Blue Pr_1	0.36(0.12)	0.33(0.11)	0.36(0.16)
Blue Pr_2	0.38(0.15)	0.71(0.26)	0.37(0.19)
Purple Pr_1	0.37(0.12)	0.31(0.12)	0.34(0.15)
Purple Pr_2	0.37(0.12)	0.17(0.16)	0.28(0.20)
Blue control	0.37(0.18)	0.70(0.26)	0.39(0.16)
Purple control	0.35(0.09)	0.22(0.11)	0.35(0.12)
Objective	0.36	0.24	0.33
	$Pr(G\|\overline{B})$	$Pr(B\|G)$	$Pr(B\|\overline{G})$
Blue Pr_1	0.34(0.16)	0.39(0.20)	0.36(0.14)
Blue Pr_2	0.39(0.16)	0.51(0.27)	0.53(0.25)
Purple Pr_1	0.36(0.14)	0.34(0.16)	0.34(0.17)
Purple Pr_2	0.36(0.14)	0.23(0.25)	0.27(0.26)
Blue control	0.30(0.16)	0.58(0.26)	0.48(0.27)
Purple control	0.31(0.11)	0.32(0.20)	0.24(0.15)
Objective	0.37	0.22	0.25

Control vs. experimental conditions. As experimental participants provided estimates twice to the same questions, Pr_2 estimates were compared to those of control group. Independent-sample t-tests for each of the six questions revealed no reliable differences between Pr_2 and the control estimates. Thus, these participants seem not to have been influenced by having to evaluate the same probabilities twice.

Analysis of the Experimental Blue group. We report the Experimental Blue and Experimental Purple participants separately, starting with Blue. As a manipulation check, we first determined whether $Pr(B)$ increased after participants saw the blue card under dim light. As expected, $Pr_2(B)$ was reliably larger than $Pr_1(B)$ [paired $t(31) = 7.0$, $p < .01$, Wilcoxon $p < .01$].

To see whether invariance holds as described in (2), we compared $Pr_1(G|B)$ to $Pr_2(G|B)$ via paired t-test and found no reliable difference [$df = 31$, $p > .05$]. Of the 32 Blue participants, 18 offered a different $Pr_2(G|B)$ estimate from $Pr_1(G|B)$. For these 18 participants, the average signed difference between $Pr_1(G|B)$ and $Pr_2(G|B)$ is -0.02 which is not reliably different from 0 [$t(17) = 0.3$, $p > .05$]. $Pr_1(G|\overline{B})$ was likewise found to be close to $Pr_2(G|\overline{B})$ [paired $t(31) = 1.3$, $p > .05$]. Eighteen out of 32 participants gave a different $Pr_2(G|\overline{B})$ estimate from $Pr_1(G|\overline{B})$. The average signed difference between $Pr_1(G|\overline{B})$ and $Pr_2(G|\overline{B})$ is -0.09 which is reliably different from 0 [$t(17) = 2.2$, $p < .05$].

To more precisely quantify violation of invariance, for each participant we calculated the absolute movement between two estimates as a percentage of the original estimate, via:

$$(5) \quad \textit{Invariance violation} = \frac{\left|Pr_2(G|B) - Pr_1(G|B)\right|}{Pr_1(G|B)}$$

We compared invariance violations to the movement of the converse probability (i.e., blue given giraffe), computed via:

$$(6) \quad \textit{Converse movement} = \frac{\left|Pr_2(B|G) - Pr_1(B|G)\right|}{Pr_1(B|G)}$$

Following an analysis due to Pearl (1988), we expected the converse movements to exceed the invariance violations. This is because giraffe seems to be conditionally independent of the light-experience given blue (once you know that the card is blue, the light provides no further information) whereas blue seems not to be conditionally independent of the light-experience given giraffe (the light here provides additional information about the color). Consistent with this expectation, the means for invariance violations and converse movements were 37.6% and 88.3%, respectively. This difference is reliable by paired t-test ($p < .05$) and Wilcoxon test ($p < .01$).

For each participant we also computed invariance violation for $Pr(G|\overline{B})$ via the following:

$$(7) \quad \textit{Invariance violation for } \overline{B} = \frac{\left|Pr_2(G|\overline{B}) - Pr_1(G|\overline{B})\right|}{Pr_1(G|\overline{B})}$$

The mean invariance violation of $Pr(G|\overline{B})$ was 45.9% and was not reliably different from the 37.6% violation of $Pr(G|B)$ reported above [paired $t(31) = 0.6$, $p > .05$].

From the results above, invariance seems to hold at least approximately. We therefore asked about its use in updating the probability of G. Specifically, for each participant, we computed the value of $Pr(G)$ via the law of total probability (1), relying on the participant's estimates for the quantities at the right of the equality. We will call this value the *total* probability of G, or $Pr_{total}(G)$ for short. Likewise, for each participant we computed the value of $Pr(G)$ via Jeffrey's rule (3). We will call this value the *Jeffrey* probability of G, or $Pr_{Jeff}(G)$ for short. The latter estimates were compared to the participant's direct evaluation of $Pr_2(G)$ via absolute difference:

$$(8) \quad \begin{aligned} total \quad error &= \left|Pr_2(G) - Pr_{total}(G)\right| \\ Jeffrey \quad error &= \left|Pr_2(G) - Pr_{Jeff}(G)\right| \end{aligned}$$

The means for total and Jeffrey error were .07 and .10, respectively, not reliably different via paired t-test [$t(31) = 1.7$, $p > .05$] or Wilcoxon test ($p > .05$).

Analysis of the Experimental Purple group. We first determined whether $Pr(B)$ decreased after participants saw the purple card under dim light. As expected, $Pr_2(B)$ was reliably less than $Pr_1(B)$ [paired $t(31) = 5.6$, $p < .01$, Wilcoxon $p < .01$].

We compared $Pr_1(G|B)$ to $Pr_2(G|B)$ via paired t-test and found no reliable difference [$df = 31$, $p > .05$]. Of the 32 Purple participants, 21 offered a different $Pr_2(G|B)$ estimate from $Pr_1(G|B)$. The average signed difference between $Pr_1(G|B)$ and $Pr_2(G|B)$ was 0.10 which was not reliably different from 0 [$t(20) = 1.8$, $p > .05$]. $Pr_1(G|\overline{B})$ was also found to be close to $Pr_2(G|\overline{B})$ [paired $t(31) = 1.6$, $p > .05$]. The average signed difference between $Pr_1(G|\overline{B})$ and $Pr_2(G|\overline{B})$ was -0.01 which was not reliably different from 0 [$t(15) = 0.3$, $p > .05$].

Just as for the Blue group, we computed invariance violations of $Pr(G|B)$ using (5) and converse movement of $Pr(B|G)$ using (6). The means for invariance violations and converse movements were 48.6% and 60.3%, respectively. This difference is reliable by paired t-test ($p < .05$) and Wilcoxon test ($p < .05$). The mean invariance violation for $Pr(G|\overline{B})$ via (7) was 18.9%, reliably smaller than the 48.6% violation of $Pr(G|B)$ [paired $t(31) = 2.9$, $p < .01$].

Once again, invariance seems to hold at least approximately so for each participant, we computed the value of $Pr(G)$ via the law of total probability (1), again denoting this value by $Pr_{total}(G)$. Likewise, for each participant we computed the value of $Pr(G)$ via Jeffrey's rule (3), denoting this value by $Pr_{Jeff}(G)$. The latter estimates were compared to $Pr_2(G)$ via the absolute differences shown in (8). The means for total and Jeffrey error were .05 and .07, respectively, not reliably different via paired t-test [$t(31) = 1.5$, $p > .05$] or Wilcoxon test ($p > .05$).

Discussion of Experiment 1

In the procedure of Experiment 1, respect for the invariance principle (2) seems normatively mandated inasmuch as experience with the light provides no further information about G once it is granted that the card is blue. In other words, $Pr_2(G|B) = Pr_1(G|B,\ell) = Pr_1(G|B)$, where ℓ is the experience provided by the light (as discussed in the Introduction).

A majority of participants, in contrast, gave different estimates for $Pr_2(G|B)$ compared to $Pr_1(G|B)$ after gaining new information about color via the light.

As a percentage of $Pr_1(G|B)$, the absolute difference between $Pr_2(G|B)$ and $Pr_1(G|B)$ was not trivial but nonetheless reliably smaller than the absolute percentage difference for the converse probabilities $Pr_2(B|G)$ and $Pr_1(B|G)$. Normatively, invariance is not expected with respect to $Pr(B|G)$. When used to estimate $Pr_2(G)$ via the law of total probability, we saw that $Pr_1(G|B)$ could be substituted for $Pr_2(G|B)$ with little loss of accuracy [total versus Jeffrey error, as in (8)]. This provides another indication of the relative modesty of invariance violations in Experiment 1.

Experiment 2

For another assessment of invariance, we asked a new group of participants to estimate probabilities for events in two different scenarios. The *lottery* scenario was designed to justify invariance whereas the *ultimatum game* scenario was not.

Participants

One hundred undergraduates from Princeton University participated in exchange for partial course credit (58 female, mean age 19.5 yrs, SD = 1.1). None had served in Experiment 1.

Materials and procedure

Fifty participants served in the *experimental* condition, and another 50 in the *control* condition. As in Experiment 1, the purpose of the control condition was to assess the impact of being asked to evaluate the same probabilities a second time. Each participant in both conditions was presented with the lottery and the ultimatum game scenarios (the order was counterbalanced).

Lottery scenario. In this scenario we substitute C (*buying a car*) for G and W (*winning the lottery*) for B. The following description was presented on the computer screen for each participant:

> Imagine that a randomly chosen adult (call him Mr. X) in New Jersey has just purchased the *Jersey Cash 5* lottery for this week. In this lottery, there are 5 numbers to be drawn, each from 1 to 40. Each number is drawn from the bowl and then put aside. The lottery jackpot is $240,000 which will be shared by players who have all 5 winning numbers (the order of the numbers doesn't matter). The numbers on Mr. X's lottery ticket are 12 17 24 32 39.

In the experimental condition, the participant answered the following questions before the lottery numbers were drawn (the order was randomized for each participant):

LOTTERY PROBABILITY QUESTIONS:

$Pr(C)$	What's the probability that Mr. X will buy a new car in the next two years?
$Pr(W)$	What's the probability that Mr. X will win the jackpot?
$Pr(C\|W)$	What's the probability that Mr. X will buy a new car in the next two years assuming that he wins the jackpot?
$Pr(C\|\overline{W})$	What's the probability that Mr. X will buy a new car in the next two years assuming that he does NOT win the jackpot?

The participant was then presented with the following additional information.

> It's the night of the lottery, and the numbers are being drawn. Mr. X becomes excited because the first four draws are 32, 12, 24, and 17. In other words, the first four numbers drawn match the numbers on his ticket.

The participant answered the same set of questions shown above (in a different random order) before the last number was drawn. Since participants had to give their estimates twice to the same questions, we informed them that they were free to provide the same estimate or a different estimate the second time around. In the control condition, participants saw the description of the lottery immediately followed by the results of the first four numbers, and then answered the questions shown above just once.

In the lottery scenario knowing the results of the first four draws provides no further information about W once it is granted that Mr. X wins the lottery. In other words, $Pr_2(C|W) = Pr_1(C|W,f) = Pr_1(C|W)$, where f is the experience of knowing the results of the first four draws. For this reason, invariance seems justified.

Ultimatum game scenario. Here we use A (*accepting the offer*) and O (*offering at least $4*) to replace G and B. The following description was presented on the computer screen for each participant:

> Imagine that two undergraduate students are randomly chosen from Princeton University to play a game. The game works as follows. The two students are given the opportunity to split $10. One student is the proposer and the other is the responder. The proposer makes an offer as to how $10 should be split between the two. The responder can either accept or reject this offer. If the responder accepts the offer, the money is split as proposed, but if the responder rejects the offer, then neither of them receives anything. The students have just finished the first trial of the game.

In the experimental condition, the participant was informed that the two students were about to play the second trial. The participant then answered the following questions (randomly ordered) about the second trial *prior to learning the outcome of the first trial*.

ULTIMATUM GAME PROBABILITY QUESTIONS:

$Pr(A)$	What's the probability that the responder will accept the offer from the proposer in the second trial?
$Pr(O)$	What's the probability that the proposer will offer AT LEAST \$4 to the responder in the second trial?
$Pr(A\|O)$	What's the probability that the responder will accept the offer assuming that the proposer offers AT LEAST \$4 in the second trial?
$Pr(A\|\overline{O})$	What's the probability that the responder will accept the offer assuming that the proposer offers LESS THAN \$4 in the second trial?

The participant was then presented with the following additional information about the scenario:

> Now you learn that in the first trial the responder rejected the proposer's offer and neither of them received anything. They are about to play the second trial.

The participant answered the same set of questions shown above (in a different random order) about the second trial of the game. Again we informed participants that they were free to provide the same estimate or a different estimate the second time around. In the control condition, participants saw the description of the ultimatum game immediately followed by the outcome of the first trial, and then answered the questions shown above just once.

In this scenario invariance is not normatively required because the outcome of the first trial suggests that the responder is sensitive to the fairness of offers. Thus, $Pr_2(A|O) = Pr_1(A|O,t) < Pr_1(A|O)$, where t is the experience of knowing the outcome of the first trial.

Results

Average responses. In each scenario we use Pr_1 to refer to probability estimates before the experience and Pr_2 for estimates after the experience. Average probabilities are shown in Table 2.

Table 2: Average estimates in Experiment 2 (standard deviations in parentheses).

lottery	$Pr(C)$	$Pr(W)$	$Pr(C\|W)$	$Pr(C\|\overline{W})$
Pr_1	0.29(0.22)	0.01(0.01)	0.77(0.22)	0.21(0.14)
Pr_2	0.33(0.25)	0.03(0.02)	0.76(0.20)	0.20(0.15)
Control	0.31(0.19)	0.04(0.04)	0.70(0.29)	0.29(0.20)
ultimatum	$Pr(A)$	$Pr(O)$	$Pr(A\|O)$	$Pr(A\|\overline{O})$
Pr_1	0.65(0.22)	0.61(0.29)	0.75(0.19)	0.44(0.30)
Pr_2	0.68(0.22)	0.74(0.24)	0.72(0.20)	0.38(0.30)
Control	0.62(0.25)	0.63(0.28)	0.70(0.25)	0.36(0.29)

Control versus experimental conditions. We found no reliable differences between Pr_2 and the control estimates for each scenario using independent-sample t-tests. Thus, experimental participants seem not to have been influenced by having to evaluate the same probabilities twice.

Analysis of the lottery scenario. We first determined whether $Pr(W)$ increased after participants saw the results of the first four draws. As expected, $Pr_2(W)$ was reliably larger than $Pr_1(W)$ [paired $t(49) = 7.7$, $p < .01$, Wilcoxon $p < .01$].

To see whether invariance holds, we compared $Pr_1(C|W)$ to $Pr_2(C|W)$ via paired t-test and found no reliable difference [$df = 49$, $p > .05$]. Of the 50 participants, only 12 offered different values for $Pr_2(C|W)$ versus $Pr_1(C|W)$. The 12 non-invariant participants made highly variable estimates, with average signed difference of 0.15 between $Pr_1(C|W)$ and $Pr_2(C|W)$ (SD = .57), not reliably different from 0 [$t(11) = 0.9$, $p > .05$]. $Pr_1(C|\overline{W})$ was likewise found to be close to $Pr_2(C|\overline{W})$ [paired $t(49) = 1.0$, $p > .05$]. Fifteen out of 50 participants gave a different $Pr_2(C|\overline{W})$ estimate from $Pr_1(C|\overline{W})$. For these 15, the average signed difference between $Pr_1(C|\overline{W})$ and $Pr_2(C|\overline{W})$ was 0.04 (SD = 0.12), again not reliably different from 0 [$t(14) = 1.0$, $p > .05$]. The average invariance violation was only 1.42% with a median violation of 0. The mean invariance violation for $Pr(C|\overline{W})$ was 1.33% with a median of 0. Since we obtained no estimate for $Pr(W|C)$, converse movement was not computed.

From the results above, invariance seems to hold rather well. For each participant, we therefore computed $Pr_{total}(C)$ via (1) and $Pr_{Jeff}(C)$ via (3), with G and B substituted by C and W. These estimates were compared to the participant's direct evaluation of $Pr_2(C)$ via absolute difference. The means for total and Jeffrey error were .15 and .13, respectively, not reliably different via paired t-test [$t(49) = 1.5$, $p > .05$] or Wilcoxon test ($p > .05$).

Analysis of the ultimatum game scenario. We first determined whether $Pr(O)$ increased after the reported rejection in the preceding trial. As expected, $Pr_2(O)$ was reliably larger than $Pr_1(O)$ [paired $t(49) = 4.2$, $p < .01$, Wilcoxon $p < .01$].

Of the 50 participants, 33 offered a different $Pr_2(A|O)$ estimate from $Pr_1(A|O)$. Thirty-eight out of 50 participants gave a different $Pr_2(A|\overline{O})$ estimate from $Pr_1(A|\overline{O})$. The average invariance violation was 18.71% with a median of 12%. As expected, this violation was reliably greater than that in the lottery case (1.42%) [paired $t(49) = 3.7$, $p < .01$]. For $Pr(A|\overline{O})$ the mean invariance violation was 17.38%, reliably greater than $Pr(C|\overline{W})$ in the lottery case (1.33%) [paired $t(49) = 2.9$, $p < .01$]. Thus, invariance was violated to a greater extent here than in the lottery scenario.

Discussion of Experiment 2

In the lottery scenario, invariance held for a majority of participants, and the absolute difference between $Pr_2(C|W)$ and $Pr_1(C|W)$ as a percentage of $Pr_1(C|W)$ was quite small. When used to estimate $Pr_2(C)$ via the law of total probability, we saw that $Pr_1(C|W)$ could be substituted for $Pr_2(C|W)$ with little loss of accuracy. In the ultimatum scenario, however, a majority of participants gave different estimates for $Pr_2(A|O)$ compared to $Pr_1(A|O)$ after learning the outcome of the first trial. Thus, invariance seems not to hold for the ultimatum scenario, as it ought not on normative grounds.[1]

[1] Without giving details, we note that Experiment 2 was repeated with 330 participants recruited over the internet via *Amazon Turk*.

General Discussion

In Experiment 1, experience with the light changed the probability that the chosen card was blue, but had only mild impact on the probability of the giraffe *given that* the card was blue. That is, $Pr_2(G|B) \approx Pr_1(G|B)$ as well as $Pr_2(G|\overline{B}) \approx Pr_1(G|\overline{B})$. These results conform to Jeffrey (1983)'s invariance requirement for updating a distribution on the basis of events whose probabilities are modified without reaching certainty. As a result, the updated probability $Pr_2(G)$ was equally well predicted from the law of total probability on the basis of $Pr_1(G|B)$ and $Pr_1(G|\overline{B})$ versus $Pr_2(G|B)$ and $Pr_2(G|\overline{B})$.

The invariance documented in Experiment 1 was selective inasmuch as greater movement was seen between the converse probabilities $Pr_1(B|G)$ and $Pr_2(B|G)$ than between $Pr_1(G|B)$ and $Pr_2(G|B)$. The difference in movement makes normative sense because the giraffe is conditionally independent of the light given the color of the card whereas the color of the card is not conditionally independent of the light given the giraffe. Experiment 1 thus provides evidence that the participants were sensitive to the normative appeal of Jeffrey's rule, distinguishing (at least partially) between situations where it legitimately applies and where it does not.

The same conclusion is suggested by the results of Experiment 2. Only one of the two scenarios — involving the state lottery rather than the Ultimatum game — gave grounds for invariance, and participants honored the principle more in the lottery context. In the latter setting, $Pr_2(C)$ (the revised probability of a car purchase) was predicted equally well from the law of total probability based on $Pr_1(C|W)$ and $Pr_1(C|\overline{W})$, as it was from $Pr_2(C|W)$ and $Pr_2(C|\overline{W})$.

Although the experiments support the hypothesis of (tacit) respect for Jeffrey's rule, the fact remains that a majority of our participants (51 of 100) changed their estimate of $Pr(G|B)$ or $Pr(C|W)$ between times 1 and 2. A slightly larger majority (54 of 100) did so for $Pr(G|\overline{B})$ or $Pr(C|\overline{W})$. [The events G(iraffe), B(lue), C(ar), and W(in) come from the flashlight and lottery scenarios, where invariance is warranted.] In percentage terms, these shifts were sizeable in Experiment 1 (averaging around 47%) although much smaller in Experiment 2 (less than 2%). The psychology of updating is incomplete without an explanation of why invariance is not respected scrupulously in settings where it seems to be required normatively.

The mere fact of evaluating the same probabilities twice might explain some of the violation of invariance. The results of our control conditions, however, suggest that this effect was minor. Recall that control participants responded just once, in the phase 2 setting (e.g., after the light), yet produced estimates that were not reliably different from those gathered in phase 2 of the experimental condition. Another source of invariance violation might be illicit conversion of conditional probability statements, e.g., evaluating $Pr(B|G)$ in place of the requested $Pr(G|B)$. This explanation is consistent with studies that highlight such conversion (as in Dawes, Mirels, Gold, and Donahue (1993)), but inconsistent with our own examination of conditional probability judgments (Zhao, Shah, & Osherson, 2009) in which little conversion was observed. Perhaps the variability of previous findings about conversion is somehow connected to the difference between the flashlight and lottery studies in their conformity to invariance.

A third possibility is that the flashlight procedure drew attention to the color dimension of the stimulus, reminding the participant of its predictive value. This realization might have been translated into more extreme conditional probabilities (higher for blue, lower for purple). The less vivid experience in the lottery scenario (merely being told about the first four numbers) would have had less impact, explaining the difference between the two experiments. As an alternative to vivacity, the greater impact of the flashlight might be relateda to the ineffable character of sensory impressions (as stressed by Jeffrey (1983, §11.1)); there is no such difficulty for the event of matching the first four lottery numbers. Of course, more data are needed to test hypotheses such as these.

The results were in close agreement with those reported here.

Acknowledgments

Osherson acknowledges support from the Henry Luce Foundation.

References

Dawes, R., Mirels, H. L., Gold, E., & Donahue, E. (1993). Equating inverse probabilities in implicit personality judgments. *Psychological Science*, *4(6)*, 396–400.

Doring, F. (1999). Why bayesian psychology is incomplete. *Philosophy of Science*, *66*, 379-389.

Gardenfors, P. (1988). *Knowledge in flux: Modeling the dynamics of epistemic states*. Cambridge MA: MIT Press.

Hacking, I. (2001). *An introduction to probability and inductive logic*. Cambridge UK: Cambridge University Press.

Harman, G. (1999). *Reasoning, meaning and mind*. Oxford UK: Oxford University Press.

Jeffrey, R. C. (1983). *The logic of decision (2nd edition)*. Chicago IL: The University of Chicago Press.

Jeffrey, R. C. (2004). *Subjective probability: The real thing*. Cambridge UK: Cambridge University Press.

Osherson, D. (2002). *Order dependence and jeffrey conditionalization.* (Available via http://www.princeton.edu/osherson/papers/jeff3.pdf)

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems*. San Mateo, CA: Morgan Kaufamann.

Stanovich, K. (2010). *Decision making and rationality in the modern world*. Oxford UK: Oxford University Press.

Wagner, C. G. (2002). Probability kinematics and commutativity. *Philosophy of Science*, *69*, 266-278.

Williams, P. M. (1980). Bayesian conditionalisation and the principle of minimum information. *British Journal for the Philosophy of Science*, *31*, 131–144.

Zhao, J., Shah, A., & Osherson, D. (2009). On the provenance of judgments of conditional probability. *Cognition*, *113*, 26–36.

Effects of Varied Priority Training on Complex Perceptual-Motor Learning

[1]Yi Wang, [2]J. Michelle Moon, [1]Wai-Tat Fu, [3]Walter Boot, [4]Kirk Erickson, [1]Arthur Kramer
[1]Beckman Institute, University of Illinois at Urbana-Champaign, Urbana, IL 61801 USA
[2]Department of Psychology, Carnegie Mellon University, Pittsburgh, PA 15213 USA
[3]Department of Psychology, Florida State University, 600 W. College Avenue, Tallahassee, FL 32306 USA
[4]Department of Psychology, University of Pittsburgh, 3107 Sennott Square, Pittsburgh, PA 15260 USA

Abstract

We reported results from a study on the effects of different training methods on complex perceptual-motor skill acquisition using a version of the Space Fortress game, which was originally designed to study the acquisition of complex perceptual-motor and cognitive skills in a multi-tasking environment. Participants were randomly assigned to the Fixed Priority (FP) and Varied Priority (VP) training conditions. Action sequences for controlling the spaceship in a frictionless environment using a joystick were analyzed and compared across conditions. Consistent with the previous findings, VP training was in general more successful than FP training. However, we found that VP training benefited participants more in the low performance group than in the high performance group. Participants in the VP training condition showed faster learning of optimal action sequences and faster reduction of suboptimal action sequences. In addition, results showed that in the high performance group, participants in the VP training condition used significantly more optimal action sequences than in the FP training condition. The findings have important implications on how the effectiveness of different training methods can be optimized for people with different cognitive abilities.

Keywords: Space Fortress game; Fixed Priority training; Varied Priority training, High performance group, Low performance group.

Introduction

Effectiveness of training perceptual-motor skills in complex, multi-tasking environments has been an important cognitive science research topic and has been studied for decades. Examples of tasks in complex, multi-tasking environments involved flying a military jet, driving a vehicle, operating a machine, etc. Operators in these environments are required not only to learn the necessary information regarding operation modes, control procedures, regulations and limitations, but also to apply these details under real-time constraints with competing cognitive demands.

Although in complex multi-tasking environments, practice generally improves performance in different training methods, researchers have found that practice time alone is not sufficient to explain differences in effectiveness of these methods. This has directed more focus towards comparing different training methods through computer-based cognitive simulations such as the 'Space Fortress' game (Mane & Donchin, 1989). This kind of synthetic training environment not only allows careful manipulation of multiple variables to carefully tease apart the multiple cognitive processes that interact dynamically to influence performance, but also allow direct measurement of how performance improves in different training environments.

Among the different training methods, the differences between whole-task training (e.g. learning to steer a bicycle and operate the pedals simultaneously) and part-task training (e.g. separately learning to steer a bicycle and operate the pedals) have been studied most extensively by researchers. In general, research shows that whole-task training is ineffective because the trainee may be overwhelmed by the complexity of the task; while part-task training is ineffective because the trainee may not have sufficient experience in coordinating between different sub-components of the tasks (Ioerger et al., 2003). As a result, a hybrid training method, often called part-whole training, was proposed. Under this approach, the whole task is decomposed into segments. Participants are trained on each of the segments separately before moving to practice the total task as a whole. Although part-whole training has shown to be effective for training in complex, multi-tasking environments (Adams 1987, Wightman & Lintern 1985, Schneider 1985), it still has two problems. First, it is difficult to select the parts to train. Second, by isolating segments, it still suffers from the same problem as in part training, in which training effectiveness may decrease because of the removal of the broader context in which the parts were performed (Gopher et al., 1989).

Varied Priority (VP) training (e.g., Kramer et al., 1995) is a training method that manipulates only the relative emphasis of selected subcomponents in the multi-tasking environment and leaves the whole task intact (Gopher et al., 1989). Gopher et al. showed that systematically varying levels of priorities on attentional control through instruction and feedback could lead to better learning and performance in multi-tasking tasks. They argued that VP training enabled participants to explore different strategies and thus develop a better match between the requirements of the tasks and the efficiency of their efforts. They suggested that participants under VP training condition not only could receive more information on their performance on the emphasized element, but could also learn the costs to performance decrement on the de-emphasized task. As a result, VP training makes people better able to strategically allocate attention to multiple components of the task to comply with the change in emphases during training.

Although benefits of VP training on global performance have been demonstrated through a number of studies, there is still a lack of understanding on the specifics of how it promotes learning of perceptual-motor control. The current

study used a version of the original Space Fortress game to study the impact of VP training on learning a complex perceptual-motor skill. The goal is to understand the impact of VP training on learning of action sequences in a dynamic multi-tasking environment.

The Space Fortress Game

The Space Fortress game was originally developed to study the acquisition of complex perceptual-motor and cognitive skills in fast-paced multi-tasking environments (Mane & Donchin, 1989). The main objective of the game was to maximize the total scores by shooting missiles at and destroying the space fortress, while maintaining a spaceship within a certain velocity limit and pre-specified boundaries on the screen. Missiles were fired from the spaceship, whose movement was controlled by the participant. In addition to destroying the fortress, the participant had to protect his/her spaceship against damage from the fortress and mine. Participants used a joystick to control the spaceship. Forward movement (thrust) of the stick caused the spaceship to accelerate. Left and right movements caused the spaceship to rotate counter-clockwise and clockwise respectively. Because the spaceship flied in a frictionless environment, it would continue to fly in the direction to which it was pointing unless it was rotated and a thrust was applied. In that case, the spaceship would change its direction of movement. This change of movement was essential not only in controlling the spaceship within boundaries, but also in maintaining its velocity within limits because of the frictionless environment (accelerating in a frictionless environment would lead to higher and higher velocity unless there was a change in flying direction).

Participants were instructed to learn to control and maintain the spaceship within a particular range of velocity and a bounded area on the screen. These two subtasks were reflected by the velocity and control scores respectively, which were continuously updated on the screen. Participants also had to protect the spaceship from being hit by bombs emitted from the fortress and mines that periodically emerged on the screen. Participants could also shoot the mines to gain points. The four subscores: points, control, velocity, and speed added up to the total scores, which were also continuously displayed on the screen.

A cognitive task analysis (Schraagen et al., 2000) was conducted to identify major components of the task and to explicate the hierarchical relationship between internal goals and external cues. Figure 1 shows the overall structure of the cognitive task analysis of the Space Fortress game. The overall objective of the game is shown at the function purpose level. The four major subscores are shown at the abstract function level, and each of the subscores is mapped to one or more generalized functions. These generalized functions were assumed to the major subgoals that participants had when they were learning to do the task, and they were explicitly taught how to accomplish these subgoals before they began the training. Each generalized function is then mapped to the various state indicators (e.g., PNTS indicated the points subscore, CNTRL indicated the control subscore, etc.) at the physical function level, which were continuously updated on the display as participants interacted with the task, and they were directly influenced by moment-to-moment actions (joystick or mouse) executed by the participants at the physical form level.

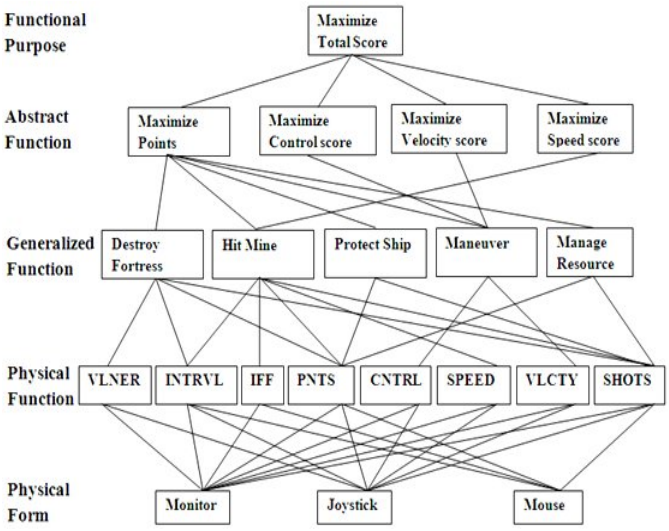


Figure 1: Cognitive task analysis of the Space Fortress game.

In the Fixed Priority (FP) training condition, participants were instructed to give equal weight to the subscores throughout the sessions. In the Varied Priority (VP) training condition, participants were instructed to emphasize one of the four subscores in each game, and the emphasis changed throughout the sessions. Due to space limitation, we will focus on effects of the training conditions on the velocity subscore, which reflected how well the participants could successfully control the velocity of the spaceship. This subscore was also the most predictive of overall performance for all participants.

Method

Participants

Thirty-nine participants recruited from University of Illinois community were randomly assigned to either the Fixed Priority (FP) training or the Varied Priority (VP) training condition. Participants had no more than a moderate amount of video game experience.

Tasks

Figure 2 shows the Space Fortress game display. The starting position of a computer-controlled fortress was centered within two concentric hexagons. And a spaceship controlled by a joystick was located between two hexagons. The fortress rotated to track and fire shots at the spaceship. The small diamond between two hexagons is a shot from the fortress. The arrow is a missile from the spaceship. The

larger diamond is a mine, which appeared every few seconds. The dollar sign indicated an opportunity for bonus.

A control panel was shown below the area in which the spaceship and mines flew. It displayed four subscores, including points (PNTS), control (CNTRL), velocity (VLCTY), and speed (SPEED). It also displayed the vulnerability (VLNER) of the fortress, an indicator to identify friend or foe (IFF), an interval (INTRVL) which indicated the time between IFF responses, and shots (SHOTS) which indicated the number of missiles remaining on the spaceship.

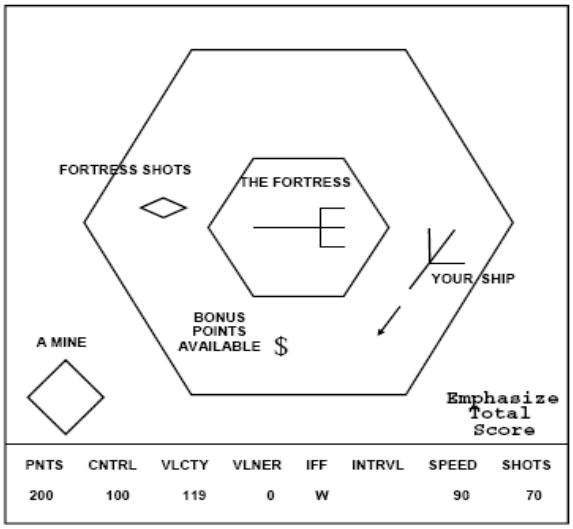


Figure 2: The Space Fortress game display

Participants in both FP and VP conditions initially completed the same three trials of an aiming task to destroy as many mines as possible. This aiming task was designed to demonstrate how to use the joystick to control the spaceship in a frictionless environment. The total aiming score was a function of the number of destroyed mines and the speed with which they were destroyed.

After completing the aiming task, participants in each condition received instructions for the actual Space Fortress game. Participants were instructed that the main objective of the game was to maximize the total score, and this was the same for both conditions. However, participants in the FP condition were told to emphasize each of the four main subscores (points, control, velocity, and speed) equally throughout the whole experiment. On the other hand, participants in the VP condition were told to improve and monitor only one particular subscore while maintaining focus on other subscores in any one of the trials.

Procedure

All participants completed the training in 10 consecutive days. Each day they did a 2-hour session, with each session consisting of 7 blocks. The first and last blocks are test blocks in which participants are required to emphasize total scores. Participants are told these are not practice blocks.

There were 5 emphasis (practice) blocks between the test blocks. For the VP group, in each emphasis block participants were asked to emphasize some aspect of the game in the order of control, velocity, speed, points, and total score, and every other day, the reverse order. All emphasis conditions were communicated to participants by pop-up windows between sessions. Additionally, for the VP group, reminder text appeared at the corner of the display telling participants what they should be focusing on (see Figure 2). For the FP group, participants did the same amount of trials but are told to always emphasize total score.

Results

Due to technical difficulties, two participants did not complete all of the tasks. The total score of one participant was 3 standard deviations away from the mean and was excluded from further analysis. We therefore had data from 36 participants in the following analyses.

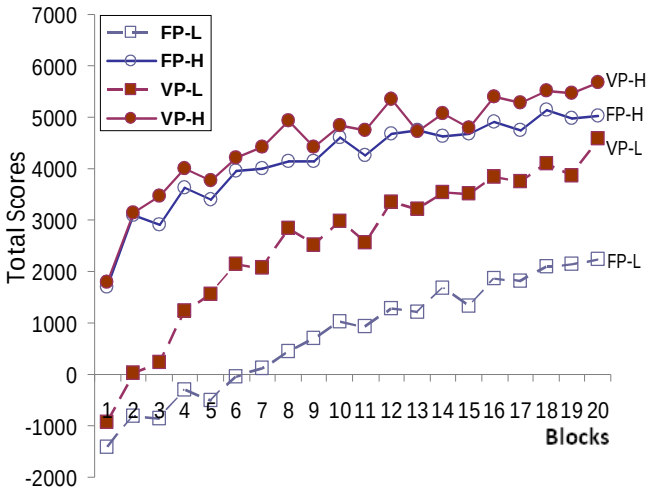


Figure 3: Average total scores across test blocks for the High (H) and Low (L) groups in each condition

Based on previous findings that VP training had different benefits for low and high ability participants (Gopher et al., 1989), we performed a median split on the total scores of the first test block to identify the High (H) and Low (L) performance groups in each condition. Figure 3 shows the total scores for each group across the 20 test blocks. Analysis of Variance (ANOVA) showed a significant main effect of blocks ($F(19, 627) = 106.946$, $p<.001$), H-L ($F(19, 627) = 106.946$, $p<.001$), but not for conditions (FP vs. VP). However, there was a significant interaction between blocks and H-L ($F(19,627) = 3.891$, $p<.001$), and between blocks and conditions ($F(19,627) = 1.745$, $p<.05$). Participants in the High and VP groups learned significantly faster across blocks than the Low and FP groups, respectively. The three-way interaction conditions x HL x blocks was marginally significant ($p=0.18$).

The results showed that, in general, VP training was more successful than FP training. Interestingly, the difference was

larger in the Low performance group, in which participants started with a much lower score and was consistently lower throughout the 20 test blocks. In fact, Figure 3 shows that for the High performance group, participants in the VP condition were only slightly better than those in the FP conditions. However, for the Low performance group, the total scores for participants in the VP condition increased to almost the same level as the High performance group at the last block, but participants in the FP condition had a much lower total score even after 20 hours of training.

Analysis of Action Sequences

To further understand the effects of VP training on perceptual-motor skill acquisition, action sequences were extracted from the game and compared across conditions. The game was designed such that clockwise rotation of the spaceship was better than counter-clockwise direction. Participants were informed of this optimal strategy upfront before the training began. To study how well participants learned to use this optimal strategy, we focus on the analysis of the number of clockwise rotation (CW), counter-clockwise rotation (CCW), and thrust (T) actions across blocks. Given that participants were instructed to control their spaceships by clockwise rotation, it was expected that participants would performed more CW and fewer CCW actions across blocks. In addition, given that the velocity score would decrease when velocity of the spaceship was too high, it was also expected that the number of T actions would decrease across blocks.

First-order Action Sequences Figure 4 shows that the number of T (thrust) actions in each condition. ANOVA showed significant main effect of H-L (F (1, 33) =26.313, p<.001), but not for conditions. The interaction between conditions and H-L was marginally significant (F(1, 33) =3.849, p=.058). As shown in Figure 4, participants used significantly fewer T actions in the High than the Low group. Participants in the FP-L group used much more T actions than those in the VP-L group, but the difference was much smaller between the FP-H and VP-H groups. ANOVA also showed that the main effect of blocks was significant (F(19,627) =3.331, p<.001), confirming the obvious downward trend, suggesting that participants were successful in reducing the use of thrust in controlling the spaceship. The interaction between blocks and H-L was also significant (F(19,627) =3.859, p<.001). The interaction between blocks and conditions and the three-way interaction was not significant.

Figure 5 shows the number of CCW (counter-clockwise) actions across 20 test blocks. ANOVA on the number of CCW actions showed significant main effects of conditions (F(1,33)=6.842, p<.05) and H-L (F(1,33)=28.116, p<.001). The interaction between conditions and H-L was also significant (F(1, 33)=5.08, p<0.05). The main effect of blocks and the interaction between H-L and blocks was significant (F(19,627)=11.306, p<.001 and F(19,627)=3.161, p<.001 respectively). However the interaction between conditions and blocks was not significant, nor was the three-way interaction.

Participants in the FP condition and Low group used significantly more CCW actions than the VP condition and High group, respectively. As shown in Figure 5, the number of CCW actions in the FP-L group was significantly higher than the VP-L group, but there was almost no difference between the FP-H and VP-H groups.

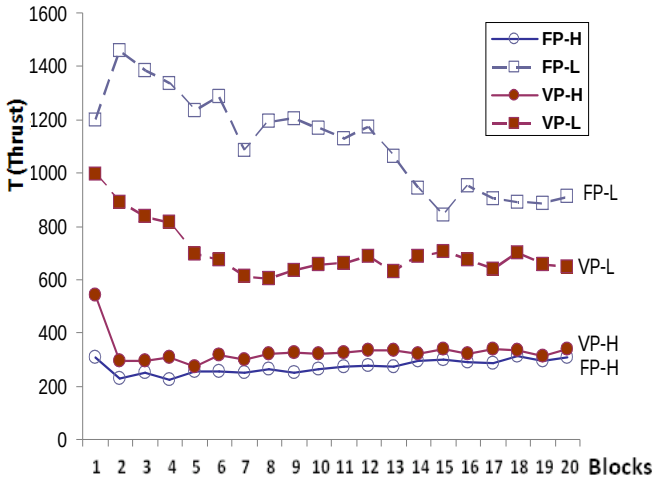


Figure 4: The number of T actions across test blocks for the High (H) and Low (L) groups in the Fixed Priority (FP) and Varied Priority (VP) conditions.

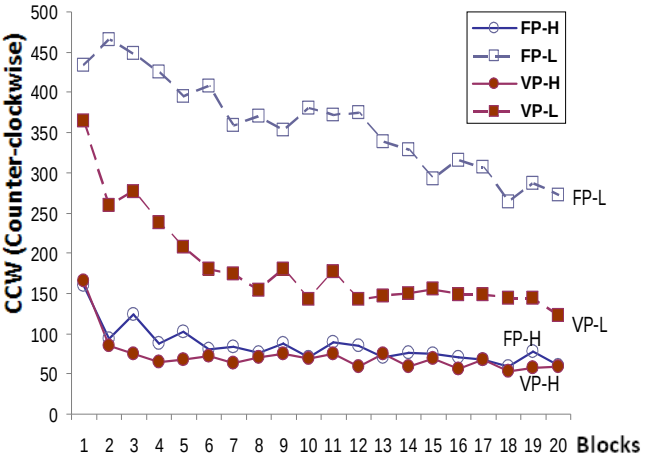


Figure 5: The number of CCW actions across test blocks for the High (H) and Low (L) groups in the Fixed Priority (FP) and Varied Priority (VP) conditions.

The overall behavioral patterns shown in Figure 4 and 5 were very similar, both showing that the number of "suboptimal" actions decreased across test blocks. However, similar to the improvements in total scores (Figure 3), participants in the High performance groups did not differ between conditions. On the other hand, participants in the Low performance groups showed a large difference: the FP-L group used significantly more "suboptimal" actions than the VP-L group. This pattern of results again supported the notion that VP training was more effective than FP training for the Low performance group. Apparently, participants

who were already good at controlling the joystick did not benefit much from either training method.

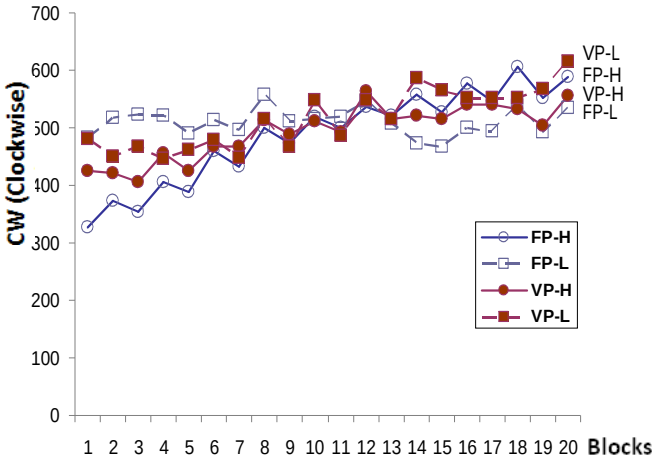


Figure 6: The number of CW actions across test blocks for the High (H) and Low (L) groups in the Fixed Priority (FP) and Varied Priority (VP) conditions.

Figure 6 shows the number of CW (clockwise) actions across test blocks. ANOVA on the number of CW actions showed that main effects of conditions, H-L and their interaction were not significant. However, the main effect of blocks was significant (F(19,627)=10.210, p<.001). The interaction between conditions and blocks was significant (F(19,627)=2.358, p<.001). As shown in Figure 6, although all participants used more CW actions across test blocks, the improvement differed across conditions. The improvement was bigger in the VP than the FP condition in the Low performance group, but not in the High performance group. Results again supported the notion that VP training was more effective to learn the optimal strategy to control the spaceship. Overall, we see that participants not only learned to reduce the number of actions needed to control the spaceship, but also learned to use more effective actions and reduced the use of suboptimal actions.

Higher-order Action Sequences We also extracted the transitions between actions to investigate further how the different training conditions influence learning of these higher-order action sequences. We extracted all second and third order transitions among the three actions CW, CCW, and T (e.g., CW-CW indicates a clockwise rotation followed by another clockwise rotation). There were a total of 9 second order and 27 third order transitions. None of the third order transitions showed significant differences between conditions. Due to space limitation, we will focus on two most frequent second-order transitions that showed significant differences between conditions.

Figure 7 shows the number of CW-T (clockwise-thrust) actions across test blocks. One major function of this action sequence was to change direction of the spaceship, and to control the spaceship to rotate in a clockwise direction and aim (and fire) at the fortress or mines to gain more points. ANOVA on the number of CW-T actions showed that the main effects of conditions (FP vs. VP) was marginally significant (p=.085). Three-way interactions blocks x conditions x H-L was significant (F(19,627) = 2.178, p<.005). No other effect was significant.

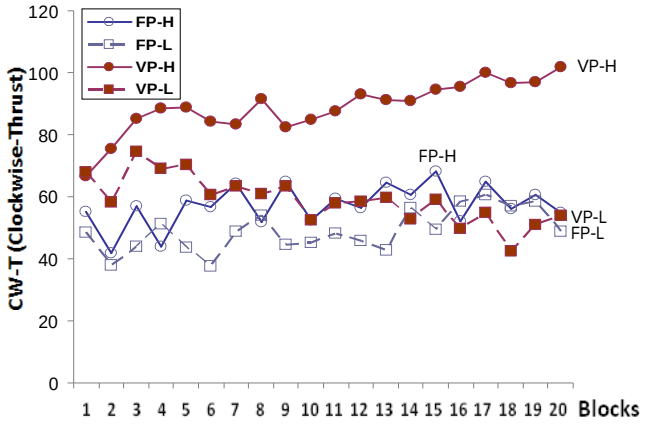


Figure 7: The number of CW-T actions across test blocks for the High (H) and Low (L) groups in the Fixed Priority (FP) and Varied Priority (VP) conditions.

Figure 7 clearly shows that the three-way interaction was caused by the higher number of CW-T in the VP-H group. Although we did not see major differences between the VP-H and FP-H in previous analyses, the higher number of CW-T showed that participants in the VP-H group not only were successful in controlling the spaceship (like the FP-H group), but they were also better at chunking the actions required to control and aim than the other groups.

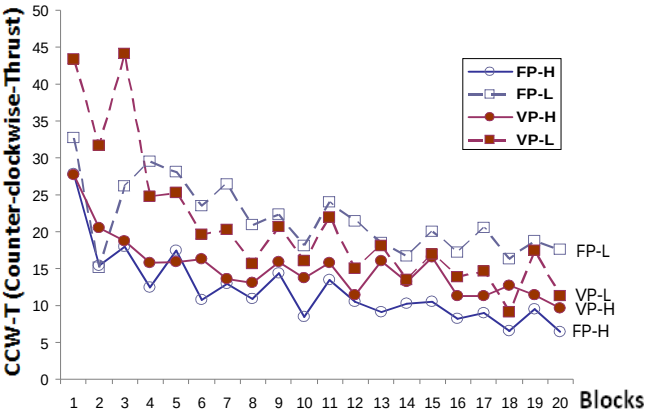


Figure 8: The number of CCW-T actions across test blocks for the High (H) and Low (L) groups in the Fixed Priority (FP) and Varied Priority (VP) conditions.

Figure 8 shows the number of CCW-T (counter-clockwise-thrust) actions across test blocks. CCW-T allowed participants to rotate in a counter-clockwise direction (which was suboptimal) and aim at the fortress or mines. ANOVA on the number of CCW-T actions showed that the main effect of H-L, and the two-way interaction blocks x conditions were marginally significant (p=.054 and p=.108 respectively). The main effect of blocks was significant (F(19,627) = 11.804, p<.001), so was the three-

way interaction blocks x conditions x H-L ($F(19,627) = 1.620$, $p<.05$). Results showed that participants were better at reducing the use of the suboptimal action sequences, and this improvement was faster for the VP-L than the FP-L group.

Conclusion and Discussion

The current results in general provided further support for the VP training method for perceptual-motor skill learning in complex, multi-tasking environments. In the Space Fortress game, perceptual-motor skill learning (controlling the spaceship) was the most difficult and critical skill for performance. The total scores across test blocks showed that for the High performance group, participants in the VP condition were only slightly better than those in the FP conditions. However, for the Low performance group, participants in the VP condition were much better than those in the FP condition.

CW actions were designed to be better than CCW actions in the game, and T actions were expected to decrease across test blocks to obtain a higher velocity score. Therefore CW actions were identified as optimal first-order action sequence, and T actions, CCW actions, and CCW-T actions were separated into suboptimal action sequences. All participants used more CW actions (optimal) across test blocks, and similar to the improvements in total scores, the increases in CW actions were bigger in the VP than the FP condition in the Low performance group, but not in the High performance group. The results of T actions, CCW actions, and CCW-T actions showed that the number of suboptimal action sequences decreased across test blocks, and participants in the VP-L group used much fewer suboptimal action sequences than those in the FP-L group, but the difference was much smaller between the FP-H and VP-H groups. In addition, the analysis of CW-T action sequences showed that participants in VP-H group not only could successfully control the spaceship, but also perform better at chunking the actions required to control and aim than the other groups.

Research has shown that VP training is often better than whole-task, part-task, and part-whole training because VP training not only can reduce task complexity but also can keep the task components as a whole. Our results showed that VP training was more effective for people who started off with a lower performance level. Given that the Space Fortress game is a difficult task that requires efficient attention allocation strategies, it was possible that performance were largely limited by cognitive resources available to the individuals. Participants in the Low performance group were therefore likely reached the resource limits earlier than the High performance group. Given that under FP training, the trainees have to simultaneously split their resources over different subcomponents, but under VP training, the trainees can invest all resources in one subcomponent at one time and then shift to other subcomponents in other trials, participants in the VP group would therefore more likely able to practice each subcomponent with more resources available, and thus would more likely to acquire better action sequences than in the FP group.

In addition to more resources available for each subcomponent, experiences of how different subcomponents were dynamically related to each other were also important in the game. Under FP training, participants received feedback based on the total score that represented the sum of subcomponents (control, velocity, speed, points); while under VP training, participants received feedback on different subcomponents in different trials. Thus, in VP training, participants obtained more diverse feedback and a wider range of experiences of different attention allocation strategies than the FP group. In other words, not only did participants in the VP group able to learn to improve each subcomponent better, they were also more likely to learn when and how to shift attention to different subcomponents and experience the performance consequences. Participants in the VP group would therefore more likely learn to acquire the better action sequences than in the FP group.

In general, participants in the Low performance group tended to benefit most from the VP training, as they showed the biggest overall improvement through faster learning of optimal action sequences and reduction of suboptimal action sequences. However, even in the High performance group, participants were better at acquiring complex action sequences in the VP condition. Our studies complement previous research by showing exactly how the training method has an impact on the acquisition of optimal action sequences in a complex multi-tasking environment, and highlight how the method interacts with the initial learning ability of participants, which is important for realistic training consideration. Future research will further investigate the effectiveness of different training methods for people with different cognitive profiles to understand how these methods can be optimized for them.

References

Adams, J.A. (1987). Historical review and appraisal of research on learning, retention and transfer of human motor skills. *Psychological Bulletin 101*, 41-77.

Gopher, D., Weil, M., & Siegel, D. (1989). Practice under changing priorities: An approach to training of complex skills. *Acta Psychologica, 71*, 147-179

Ioerger, T.R., Sims, J., Volz, R.A, Workman, J., and Shebilske, W.L. (2003). On the Use of Intelligent Agents as Partners in Training Systems for Complex Tasks. *Proceedings of the 25th Annual Meeting of the Cognitive Science Society (CogSci'03).*

Kramer, A.F., Larish, J.F., & Strayer, D.L. (1995). Training for Attentional Control in Dual Task Settings: A Comparison of Young and Old Adults. *Journal of Experimental Psychology: Applied, 1* (10), 50-76.

Mane A., & Donchin, E. (1989). The space fortress game. *Acta Psychological, 71*, 17-22.

Schneider, W. (1985). Training high performance skills: Fallacies and guidelines. *Human Factors 27*, 285-301.

Schraagen, J. M., Chipman, S. F., & Shalin, V.L. (2000). *Cognitive Task Analysis*. Mahwah, NJ: LEA.

Wightman, D.C. and G. Lintem. (1985). Part-task training for tracking and manual control. *Human Factors 27*, 267-284.

Non-Verbal Responses to Being Ignored: Evidence of Cognitive Deconstruction?

Emiel Krahmer (E.J.Krahmer@uvt.nl)
Juliette Schaafsma (J.Schaafsma@uvt.nl)
Marc Swerts (M.G.J.Swerts@uvt.nl)
Martijn Balsters (M.J.H.Balsters@uvt.nl)
Tilburg Centre for Cognition and Communication (TiCC), Tilburg School of Humanities, Tilburg University
PO BOX 90153, NL-5000 LE Tilburg, The Netherlands

Ad Vingerhoets (Vingerhoets@uvt.nl)
Tilburg School of Social and Behavioural Sciences, Tilburg University
PO BOX 90153, NL-5000 LE Tilburg, The Netherlands

Abstract

This study examined people's non-verbal reactions to being ignored or included during a social interaction. It was hypothesized that external judges could determine, on the basis of non-verbal cues, whether a person was ignored or included. Moreover, we expecteded that people who were ignored would become less non-verbally expressive, which could be indicative of cognitive withdrawal. It was found that persons who had been ignored reported lower average mood scores than included persons. External judges were, on average, also able to distinguish individuals who were ignored from those who were included. In terms of people's specific non-verbal behaviors, however, the findings are less clear. Even though persons who were ignored engaged less in affiliative behaviors than included persons, they did not display more non-verbal behaviors that are indicative of withdrawal than included persons (e.g., flight). Limitations of the study and future directions are discussed.

Keywords: Exclusion; non-verbal behaviors; cognitive deconstruction

Introduction

Human beings are deeply motivated to form stable, lasting connections with other people. They strongly desire social attachments and seem inclined to form relationships even in the absence of ulterior motives. Moreover, they are willing to spend considerable time and effort in fostering supportive relationships with others and are generally reluctant to end relationships, even when these relationships have become unnecessary or dysfunctional. This tendency to strive for strong social attachments presumably has an evolutionary basis. There is evidence that, over evolutionary time, human beings who were well-integrated into social groups were most likely to survive, reproduce, and succesfully raise their offspring (Baumeister & Leary, 1995; Leary, 2001).

When people's belonging needs are threatened, they respond in a variety of negative ways. Laboratory studies show that being excluded or rejected, even if it is for only a short period of time, is a painful experience (e.g., Eisenberger, Lieberman, & Williams, 2003) that increases self-defeating behaviors (e.g., Twenge, Catanese, & Baumeister, 2002) and may lead to aggression toward others (e.g., Twenge, Baumeiser, Tice, & Stucke, 2001; Twenge & Campbell, 2003). For example, Twenge et al. (2001) found that participants who had been excluded by other participants or who had been told that they would have a lonely future administered more unpleasant noise blasts to others than those who had been included or who had been told that they would have rewarding relationships throughout their life

Surprisingly, however, researchers have not always found relationships between social exclusion and emotional distress (e.g., Twenge et al., 2001; Twenge et al., 2003). Instead, several laboratory studies suggest that people seem to respond to social exclusion in a detached and emotionally indifferent manner. To account for these findings, it has been hypothesized that social exclusion or rejection may initially lead to feelings of inner numbness or a state of cognitive deconstruction. For example, Twenge et al. (2003) found that rejected participants were more lethargic, displayed slower reaction time, were more likely to agree that life is meaningless and avoided self-focused attention. According to Twenge et al. (*ibid.*) such a deconstructed state may serve as a temporary defense against the negative experience of social rejection.

Most studies to date, however, have only relied on self-reports of people's affective states or moods. So far, little research has focussed on the non-verbal behaviors of excluded persons. The purpose of the current investigation was to how people respond non-verbally when they are being ignored during a conversation with others. It was hypothesized that judges could determine, on the basis of non-verbal cues, whether a person is ignored or included. In line with the 'numbness hypothesis', it was also expected

that, compared to included persons, people who were ignored would become less non-verbally expressive, which could be indicative of cognitive withdrawal. One could argue that this could be an adaptive response to cope with the emotional stress caused by the exclusion. For example, according to Engel (1962, 1975) psychological and physical inactivity during stressful situations (e.g., the withdrawal of attention, self-preoccupation or sleep) may protect individuals from overstimulation or excessive trauma. Persons who are being ignored or excluded may also engage in displacement behaviors. For example, Troisi (2002) argued that displacement activities may be adaptive in that they reduce autonomic activation.

It is also possible, however, that people's non-verbal reactions during social interactions are influenced by how socially anxious they are. For example, previous research has shown that individuals with a higher fear of negative evaluation try to avoid being evaluated unfavorably (Watson & Friend, 1969), generally feel worse about receiving negative feedback (Friend & Gilbert, 1973), and are also more concerned with and try harder to make a good impression on others during interactions (Leary, 1983). In our analyses, we therefore controlled for fear of negative evaluation.

Method

Participants and Design

Participants were 58 undergraduate students (37 women) from the University of Tilburg who participated for partial course credit (*M* age = 20.8, *SD* = 2.4). Participants were randomly assigned to the inclusion or exclusion condition (29 in each condition).

Procedure

The experiment was presented as a study on group decision-making under time pressure/stress, and participants were led to believe that they would be engaging in a decision-making discussion with two other participants. In reality, they would communicate with a pair of actors (one male, one female) operating on an elaborate script.

At the start of the experiment, participants were led into a room and told that the other two "participants" were in separate rooms as well. After the global procedure was explained, participants signed a consent form, and six electrodes were applied to measure heart rate. Following the APA guidelines for ethics, participants were informed that they could stop their participation at any moment, without having to give a reason. None of our participants used this right.

After having received the instructions abou the experiment, participants filled out a first questionnaire to assess their mood. Subsequently, they were exposed to a 7-minute film fragment, consisting of underwater scenes filmed in the Red Sea and accompanied with relaxing music to make sure that participants in both conditions were in a comparable state of mind at the start of the discussion. To check whether this was indeed the case, participants were asked to fill out the same mood questionnaire a second time. Subsequently, participants were accompanied to the discussion room, where they met with the other two "participants" (the confederates). All three were seated at a hexagonal table, so that each person had one conversation partner on the left-hand and one on the right-hand side, and each had a digital DV camera (25 fps) in front. Both the participant and the confederates were recorded, and participants were told that these recordings would be needed to analyze the decision making process afterwards.

At this point, participants read a text about the case to be discussed, containing the description of a communication problem in a local sport school. Participants were instructed to collectively answer two questions (How did the problems arise? And how could they be solved?), and they were given 4 minutes to answer each one. The actual experimental manipulation occurred during the discussion of the second question. In the inclusion condition, the confederates continuously focussed on the contributions of the participant and emphasized how much they appreciated these ("yes, that's an excellent suggestion!"); in the exclusion condition, the confederates discussed the case solely among themselves, ignoring any contributions from the participant.

After 2 x 4 minutes, the experimenters re-entered the discussion room, and each guided one conversation partner (the participant or one of the two confederates) back to one of the individual rooms. Once there, participants were asked to fill in the mood questionnaire once more. After this, they were shown a second, 7-minute Red Sea underwater scene with soothing music, in an attempt to bring the participants' mood back to more neutral levels. Finally, participants filled out the mood questionnaire one last time.

Subsequently, the participants were fully debriefed about the experiment. None of them was suspicious about the experimental set-up; in particular, all believed that they had been interacting with other, "real" participants. Participants also signed a non-disclosure agreement, to make sure that future participants were uninformed about the actual nature of the experiment. Overall, the experiment lasted about one hour.

Measures

Nonverbal measures. The non-verbal behaviors of the participants were analyzed in two ways. To examine whether outside observers could actually see whether a person is included or excluded, 25 undergraduate students (8 women) judged, on the basis of two fragments, whether participants had been included or excluded. For each of the 59 participants in the experiment, two fragments of 8 seconds (200 frames) were selected. One fragment was selected from the beginning of the four minutes experimental manipulation (frames 1000 - 1200, i.e., 0.40 - 0.48 minutes), and one from the second half (frames 4000 - 4200, 2.40 - 2.48 minutes). This resulted in 59 x 2 = 118 stimuli. We opted for fragments of 8 seconds to keep the overall length of the judgement study within reasonable limits. The stimuli were presented to the individual judges in one of two random orders, to control for potential learning effects. Judges had to indicate by forced choice for each fragment whether they believed the person in the film-clip was included or excluded, and on a five point scale how certain they were of their choice. For data processing perceived inclusion was mapped to "1" and perceived exclusion to "-1", and these scores were multiplied with the certainty score. This resulted in a score ranging from -5 ("very certainly excluded") to +5 ("very certainly included"). The evaluation of the fragments was preceded by a short training session of five stimuli (consisting of random 8 second fragments not used in the actual experiment), to make participants acquainted with the experimental setting.

To examine the specific non-verbal behaviors of the participants, two independent raters who were blind to the experimental manipulation coded two 30-second fragments (0.30 - 1.00 and 2.30 - 3.00) for each participant. Fragments of 30 seconds were choosen to obtain a good estimate of the different non-verbal behaviors that participants showed. These selections were coded using the Ethological Coding System for Interviews (ECSI); see, for example, Troisi (2002) or Troisi & Moles (1999). The ECSI is a validated non-verbal behavior scale, consisting of 8 behavioral categories and a total of 37 easy to code nonverbal cues. We selected four behavioral categories for coding, namely "Affiliation" (which is associated with ECSI behaviors 2-6, e.g., smile, head tilt, eyebrow flash), "Flight" (behaviors 10-15, e.g., look away/down, chin to chest), "Displacement" (24-32, e.g., hand-face touching, yawning), and "Relaxation" (33-37, e.g., settle, fold arms, laugh). Coding was done blind to condition, and without sound (as required by the ECSI guidelines). For each individual ECSI behavior, the agreement between the raters was measures using Cohen's kappa. We found that kappa scores for the different behaviors indicate moderate to substantial agreement, where discrete behaviors (e.g., fold arms) generally resulted in higher kappa scores than continuous ones (e.g., head tilt). Disagreements between raters were resolved after discussion.

Fear of Negative Evaluation. Participants completed the brief Fear of Negative Evaluation Scale (Leary, 1983). This scale consists of 10 items (e.g., "I am afraid others will approve of me") that were measures on scales ranging from 1 (not at all characteristic of me) to 5 (extremely characteristic of me). Cronbach's alpha was .83.

Control measures. At several points throughout the experiment, participants were asked to fill out a self-report mood scale derived from Mackie and Worth (1989) and Krahmer et al. (2004). The scale consisted of six 7-point bipolar semantic differential scales ("At this moment, I feel . . . "), using the following adjective pairs (English translations of Dutch originals): happy/ sad, pleasant/ unpleasant, satisfied/ unsatisfied, content/ discontent, cheerful/ sullen and in high spirits/ low-spirited. Alpha's were > .80.

Results

Manipulation Check

To check whether the experimental manipulation worked, we analyzed the self-reported mood scores. Table 1 contains the average scores for the four mood measurements.

Table 1: Average mood scores (standard deviations between brackets)

	Ignored	Included
Mood 1: Initial	5.09 (.81)	5.21 (.76)
Mood 2: After film 1	5.41 (.73)	5.41 (.61)
Mood 3: After manipulation	4.92 (.83)	5.69 (.64)
Mood 4: After film 2	5.73 (.78)	5.56 (.68)

The average mood scores were submitted to a within-subjects Analysis of Variance, with the experimental manipulation (Ignored vs. Included) as a between-subjects factor. Most relevant for our current purposes is that a significant interaction was found between Condition and Time, $F(1, 57) = 7.69$, $p < .001$. In particular, as can be seen in Table 1, average mood scores for the two conditions are exactly the same after the first film fragment (as intended), but after the experimental manipulation a clear difference between the conditions can be observed: participants who had been ignored reported lower average mood scores than participants who had been included. After watching the second film fragment this effect

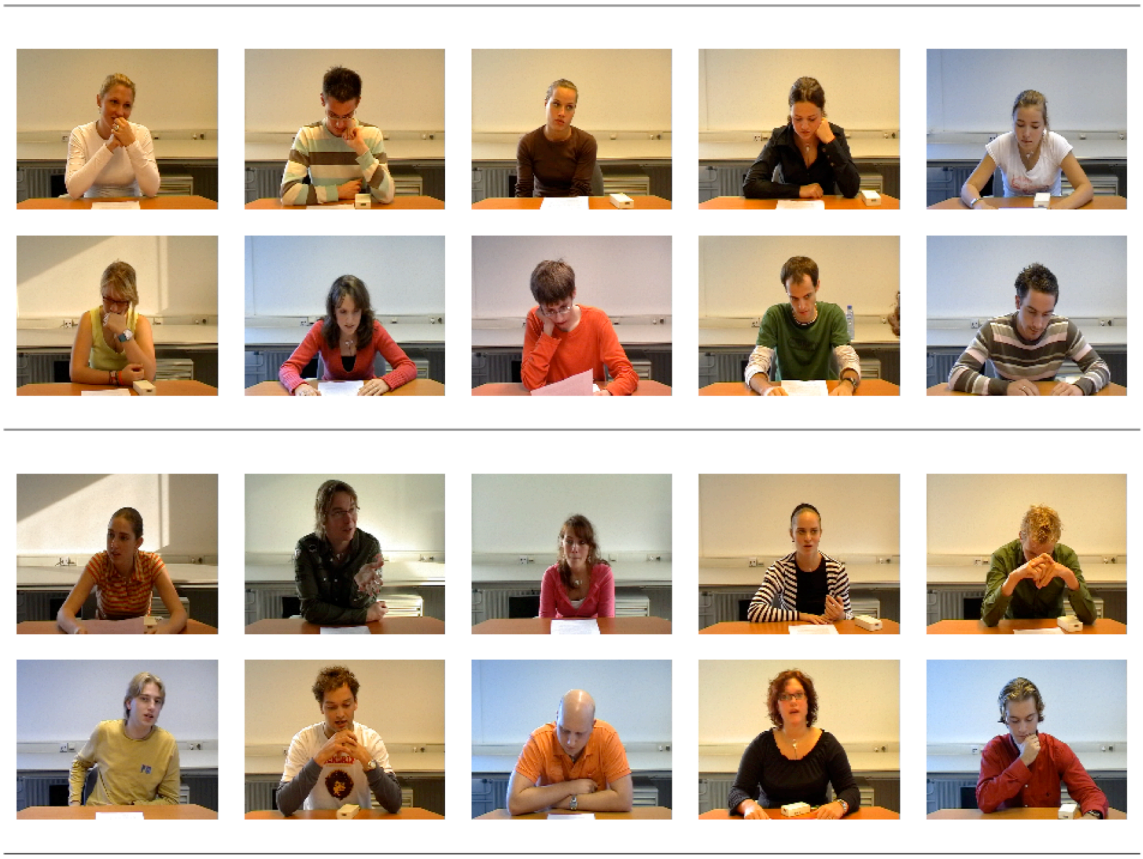


Figure 1: Randomly selected stills of 10 participants that were ignored (top panel) and of 10 participants that were included (bottom panel).

disappeared again. Interestingly, there was also a main effect of Time of the mood measurement, and inspection of Table 1 reveals that participants actually felt better after than before the experiment (irrespective of the condition they had been in), $F(1, 57) = 5.01$, $p < .01$. In fact, various participants indicated during the debriefing phase that they though the experiment was about the effects of watching underwater scenes, which indeed took up a large part of the experimental procedure.

Perceptions of External Raters

To examine whether external raters could, on the basis of non-verbal cues, determine whether participants were included or ignored, we subjected their evaluations of the second film fragments to an Analysis of Variance. Raters' evaluations of the first film fragments were included as a covariate. We also controlled for participant's sex and their fear of negative evaluation. This analysis revealed a main effect for the experimental manipulation, $F(1, 54) = 10.62$, $p < .01$. Participants who were ignored were perceived as more excluded ($M = -1.36$, $SD = 2.48$) than included participants ($M = 1.26$, $SD = 2.54$). Note, however, that the standard deviations are relatively large. There was also a main effect for fear of negative evaluation, $F(1, 54) = 5.21$, $p < .05$. External raters more often considered persons with a lower fear of negative evaluation as included than persons with a higher fear of negative evaluation.

We also examined whether there was an interaction between the experimental manipulation and fear of negative evaluation. We also found an interaction that was significant at the .10 level, $F(1, 54) = 3.04$, $p < .09$). Persons with a lower fear of negative evaluation were perceived as more included in the inclusion condition. There was no difference in how persons with a higher or lower fear of negative evaluation were rated in the exclusion condition.

Non-verbal behaviors

We also examined the specific non-verbal behaviors of the excluded and included participants. For this purpose, we first conducted a MANOVA, with the four ECSI-categories (Affiliation, Displacement, Relaxation, and Flight) as the dependent variables. This analysis only revealed a main effect for Affiliation, $F(1, 55) = 12.76$, $p = .001$). Included

participants displayed more non-verbal affiliative behaviors (M = 1.60) than participants who were ignored (M = .88). The means for all the four categories are presented in Table 2.

Table 2: Average number of non-verbal behaviors across the 4 ECSI-categories

Category	Ignored	Included
Affiliation	0.88 (0.80)	1.60 (0.97)
Flight	1.20 (0.44)	1.17 (0.38)
Displacement	2.12 (1.39)	2.28 (1.46)
Relaxation	0.85 (0.48)	0.83 (0.50)

We then examined the non-verbal behaviors of the participants in more detail. We conducted a series of MANOVAs on the specific non-verbal behaviors within each category of the ECSI coding system. These analyses showed that included participants more often quickly raised and lowered their eyebrows (p < .01) or kept their eyebrows up for some time (p < .05), and smiled more often than participants who were ignored (p < .01). They also seemed somewhat more relaxed (p < .10) and more often displayed a neutral face (p < .05). Compared to included participants, persons who were ignored touched their face more often (p < .05) and twisted their mouth (p < .01), licked their lips (p < .10) or bit their lips more often (p < .10).

Conclusion and Discussion

Being excluded, rejected or ignored is a painful experience that can evoke a host of negative reactions within individuals, ranging from sadness to anger. These reactions, however, may not always occur immediately. Instead, it has been argued that people's immediate reaction to exclusion may be cognitive withdrawal. This withdrawal may be adaptive, in the sense that it may protect individuals from the pain of being excluded.

In this study, we examined whether persons who are being ignored also become less non-verbally expressive. For this purpose, we compared the non-verbal behaviors of persons who were being ignored to the non-verbal behaviors of persons who were being included. We expected that external raters could reliably determine, on the basis of a person's non-verbal behaviors, whether he or she was being ignored or included and we also expected that participants who were being ignored would display non-verbal behaviors that reflect a tendency toward withdrawal. Moreover, we expected that being ignored would result in more displacement behaviors.

The results of this study are somewhat mixed. On the one hand, we found that persons who had been ignored reported lower average mood scores than included persons. External judges also rated participants who were ignored as more excluded than included participants. Generally, persons with a lower fear of negative evaluation were more often perceived to be included than persons with a higher fear of negative evaluation. In terms of people's specific non-verbal behaviors, however, the findings are less clear. For example, even though persons engaged in less affiliative behaviors when they were ignored, they did not display more non-verbal behaviors that are indicative of withdrawal than included persons (e.g., flight). To some extent, however, they did engage in more displacement behaviors.

The data suggest that persons who are being ignored do, in terms of their non-verbal behaviors, seem to become somewhat more lethargic than included persons but do not entirely disengage from the interaction. It is possible, however, that participants may have evoked display rules to mask or neutralize their feelings because they found themselves in the presence of others. The fact that the interaction was recorded may have also contributed to this. Moreover, the ECSI-coding system that we used to analyze people's non-verbal behaviors may not have been sufficiently detailed to assess the more subtle non-verbal behaviors of our participants.

Nevertheless, the results of the study indicate that even though it is distressful to be ignored or excluded, it does not lead to an overt outburst of emotional distress. Moreover, the tendency of individuals to display less affiliative behaviors while they are being ignored may also be useful, because it may help them avoid doing or saying anything that would make things worse. Future studies should, however, examine in more detail how people cope, cognitively and emotionally, with the stress they experience while they are being ignored or excluded.

Ackowledgements

Thanks are due to Lennard van de Laar and Caifeng Shan for technical assistance and to Rian Blankenstein, Lotte Oostrom, Bregje van Rijbroek, and Marjolein de Vries for their help with conducting the experiments. Krahmer thanks the Netherlands Organization of Scientific Research (NWO) for Grant 277-70-007.

References

Baumeister, R. F., & Leary, M. R. (1995). The need to belong: Desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin*, *117*, 497-592.

Engel, G. L. (1962). Anxiety and depression - withdrawal: The primary effects of unpleasure. *International Journal of Psychoanalysis, 43*, 89-97.

Engel, G. L. (1975). Perspectives in depression. *Science, 190,* 453-455.

Friend, R. M. & Gilbert, J. (1973). Threat and fear of negative evaluation as determinants of social comparison. *Journal of Personality*, *41*, 328-340.

Eisenberger, N. I., Lieberman, M. D., & Williams, K. D. (2003). Does rejection hurt? An fMRI study of social exclusion. *Science*, *302*, 290-292.

Krahmer, E., Dorst, J. van, & Ummelen, N. (2004). Mood, persuasion and information presentation, *Information Design Journal*, 12, 40–52.

Leary, M. R. (1983). Social anxiousness: The construct and its measurement. *Journal of Personality Assessment*, *47*, 66-75.

Leary, M. R. (2001). Toward a conceptualization of interpersonal rejection. In M. R. Leary (Ed.), *Interpersonal Rejection* (pp. 3 – 20). New York: Oxford University Press.

Mackie, D. & L. Worth (1989). Processing deficits and the mediation of positive affect in persuasion. *Journal of Personality and Social Psychology*, *57*, 27-40.

Troisi, A. (2002). Displacement activities as a behavioral measure of stress in nonhuman primates and human subjects, *Sress*, *5*, 47-54.

Twenge, J. M., Baumeister, R. F., Tice, D. M., & Stucke, T. S. (2001). If you can't join them, beat them: Effects of social exclusion on aggressive behaviors. *Journal of Personality and Social Psychology*, *81*, 1058-1069.

Twenge, J. M., & Campbell, W. K. (2003). Isn't it fun to get the respect that were going to deserve? Narcissism, social rejection, and aggression. *Personality and Social Psychology Bulletin*, *29*, 261-272.

Twenge, J. M., Catanese, K. R., & Baumeister, R. F. (2002). Social exclusion causes self-defeating behavior. *Journal of Personality and Social Psychology*, *83*, 606-615.

Twenge, J. M., Catanese, K. R., & Baumeister, R. F. (2003). Social exclusion and the deconstructed state: Time perception, meaninglessness, lethargy, lack of emotion, and self-awareness. *Journal of Personality and Social Psychology*, *85*, 409-423.

Watson, D., & Friend, R. (1969). Measurement of social-evaluative anxiety. *Journal of Consulting and Clinical Psychology*, *33*, 448-457.

The Bodily Movements of Liars

Natasha M. Eapen (nme@eyethink.org)
Sam Baron (sambaron@ eyethink.org)
Chris N.H. Street (chris@eyethink.org)
Daniel C. Richardson (dcr@eyethink.org)
Cognitive, Perceptual & Brain sciences, University College London
Gower Street, London WC1E 6BT, UK

Abstract

We measured the continuous bodily motion of participants as they lied to experimenters. These lies were spontaneous rather than elicited, and occurred for different motivations. In one situation, participants were given the opportunity to lie about their performance on a maths test in order to win money. In another, they witnessed one experimenter accidentally break a laptop. When asked what had happened, participants were motivated to lie and deny any knowledge. Across these situations, participants lied 61% of the time, allowing us to contrast the body movements of liars with truth tellers as they answered neutral and critical questions. Those who lied had significantly reduced bodily motion. In one case this motion appeared before the experimenter had even asked the critical question. We conclude that a person's bodily dynamics can be indicative of their cognitive and effective states, even when they would rather conceal them.

Keywords: deception; social cognition; action; body motion

Introduction

While running in between parallel sessions at the next Cognitive Science conference, you catch the eye of a colleague from your university. They smile and ask what you thought of the talk they gave that morning, the one you promised to attend. Since you know that the sessions were both crowded and dimly lit, you take a chance, smile, and say, 'Of course I was there. It was fantastic, as always." Will they believe your deliberate falsehood? What in your choice of words, gestures and behaviour will make you sound either convincing or conniving?

The good news for you is that it is very unlikely your colleague will be able to tell that you are lying. Regardless of their training or self belief, people are able to detect deception in others at only a fraction above chance levels (Bond & DePaulo, 2006, 2008; DePaulo, Stone, & Lassiter, 1985; Köhnken, 1987; Vrij 1993, 2000). The bad news for your colleague, professionals who need to know if people are lying, and those studying deception is that many researchers have found no unique and reliable behavioural signature for deception (Vrij, 2008; Ekman, 1992; Buller & Burgoon, 1996; DePaulo et al., 1985, 2003).

One reason is that deception is a daily activity and a diverse phenomenon (DePaulo et al., 1996). Sometimes a lie means making up a story, sometimes it means a simple denial. Lying can be done out of kindness or out of self interest. It can have different consequences and place different emotional and cognitive demands on the liar (Vrij & Mann, 2004).

It may not be surprising, then, that a clear link between deceit and bodily activity has proven elusive (DePaulo et al., 2003). Whilst some researchers have demonstrated increased movements of the fingers, arms, hands, legs and feet whilst deceiving (McClintock & Hunt, 1975), even if only anecdotally (Porter & ten Brinke, 2009), others report decreased limb movements (Vrij, 2008).

Set against this long tradition of deception research are more recent findings in cognitive science regarding the relationship between motor control and cognitive processing. Thinking about the past versus the future shifts the direction that your body tilts (Miles, Nind, & Macrae, 2010) moving marbles upwards rather than downwards changes the emotional content of the memories you recall (Casasanto & Dijkstra, in press), swaying in time with another person and mimicking their actions they make causes you to like each other (Chartrand & Van Baaren 2009), and how you move your mouse cursor when asked 'Do you like Black people?' reveals the influence of negative stereotypes (Wojnowicz et al., 2009). Many of these recent findings rely on fine grained, continuous measures or manipulations of motor activity, rather than discrete, categorical behaviours such as button presses (see Spivey, 2007 for a motivation). Is it possible to use these continuous methods to detect a behavioural signal to deception?

Recently, Duran, Dale & McNamara (in press) adapted a standard paradigm in deception research, and asked people to respond yes or no to certain questions about themselves. They were instructed to give false answers in some cases and truthful answers in others. In this experiment, they signalled their responses using a Wiimote-controller that translated their hand movements into movements of a cursor. They found that deceitful answers had a characteristic movement trajectory, with increased complexity and competition from other responses. This exciting evidence suffers from only one flaw - it is rare that people ask us to lie to them. Regardless, Duran et al's findings suggest that perhaps there is more in motor behaviour than has typically been measured by deception researchers.

In deception research, bodily behaviour is usually videotaped and coded by experimenters. However, this is a laborious, costly and inaccurate method as reliability between coders must be established, extensive training can be necessary and only bodily movements discernible to the human eye can be analysed. Furthermore, because it is necessary to be selective in such a coding method, it requires an understanding of what bodily movements are potentially interesting to examine before it is possible to begin coding. As we have noted, the consensus is that there is no reliable behavioural profile for deceit. An ideal, for

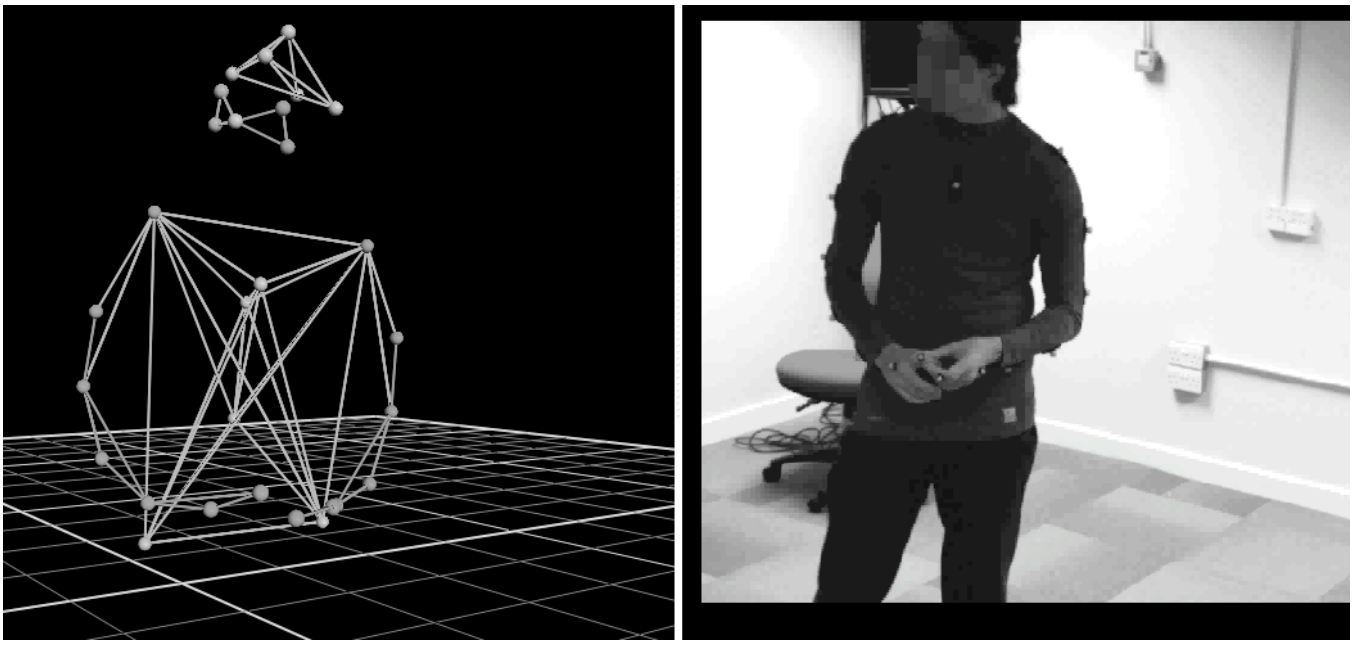


Figure 1. A participant wearing the motion tracking shirt and hat (right) and the 3D position of the markers reconstructed by the motion tracking system (left)

both practical and theoretical goals, is an automated system which will record accurate bodily changes and provide an objective approach to the study of deception (Burgoon et al., 2005; Vrij & Mann, 2004). Such a method, we believe, must be able to maintain the complexity of continuous variables.

Our goal in this experiment was to use continuous, objective measures of the bodily activity of participants who were engaged in spontaneous deceptive behaviour. Participants were told they were in a study about posture and mental arithmetic. This gave us an excuse to measure their body position at 24 locations 200 times a second. It also provided the opportunity for them to lie. We offered them £5 for improving their score in a second round of the maths test. Although the task became much harder, so that score was most likely to decrease, only the participants could know how well they had performed. Therefore, participants were given the opportunity to tell a lie for a monetary reward, without the risk of being caught out.

During the experiment, the participants witnessed the junior experimenter accidentally knock a laptop off of a table. The laptop belonged to a senior experimenter who was absent at the time. Later, when he was unable to turn on his laptop, he asked the participants if they had seen anything happen to it. The junior experimenter was very friendly and the senior experimenter quite unpleasant, and we assumed that this would provided a second motivation to deceive. We hoped that the participants would be motivated to cover up for her, and falsely deny knowledge of the incident. During these moments, when participants thought that the real experiment had finished, and when they could freely chose to tell the truth or lie, we captured their bodily motion.

Methods

Participants

32 participants took part in two experiments, run in a single session. They were UCL students or members of UCL's subject pool, and received course credit or a payment for their participation. There were 20 females and 12 males, with a mean age of 22.5 years old.

Apparatus

The experiment took place in UCL's Multimodal Lab. Six high speed infrared cameras were mounted on a rail around the perimeter of a 5m square area. Participants wore a tight, stretchable shirt and a cap which had 20 plastic markers arrayed over them (see *Figure 1*). The markers were approximately 2cm in diameter and are highly reflective in infrared light. Additional markers were attached to the hands, tips of the index fingers and the face. Image data from the cameras was passed to the Vicon Nexus motion tracking system at a rate of 200 Hz. The 3D position of each marker was reconstructed with an accuracy of 0.1mm. A digital camera recorded a view of the participants' actions, and a *ladybug 2*, 360º panoramic camera recorded all events in the lab. A ceiling mounted omnidirectional mic provided a sound recording. Participants carried out the experiment sat 50cm or stood 200cm away from a Mac laptop.

Procedure

Three experimenters ran the study. One operated the motion tracking systems and did not interact with the participant. The other two experimenters dealt with the participants according to a well rehearsed script. The senior, male experimenter acted in a cold and unpleasant manner

throughout. The junior female experimenter, who was an undergraduate like the participants, was friendly and engaging. The experimental procedure was designed to put participants in two situations in which they might chose to lie spontaneously to the experimenters. For clarity, we will describe those two situations separately, even though the events they describe partially overlapped with each other. In each, motion is captured in two periods: as the participants reply to a neutral question, and as they reply to a question that has a motivation for deception. Participants were unaware that each of these periods were the critical portions of the experiment and that their behaviour was being recorded.

Participants were told the study was investigating the relationship between body sway and mathematical ability. After donning the motion capture clothes, they stood in a T-pose with arms outstretched for a brief calibration recording. The participants then took part in a simple maths test. After the experiment they were debriefed to the true aim of the study, and gave additional retrospective consent with the option to withdraw their data, which none chose to do.

The Maths Test Following calibration, participants were seated at a table and given a maths test on a laptop. There were 30 multiplication questions with three multiple choices. Participants had a limited time to respond and were given feedback on their answer. A pilot test showed that people scored around 75% on the test.

After completing the test, participants were shown their score. The junior experimenter told them that they were now required to repeat the test, but this time standing up while the motion tracking system measured their balance. She explained that our hypothesis was that standing would improve maths performance. She said that was what we had found so far, and hoped to prove conclusively. In violation of good experimental practice, she deliberately increased the demand characteristic of the 'experiment' by telling participants how they should perform. In addition, she explained that participants would receive a £5 reward if their results followed the hypothesised pattern and they scored better while standing.

The participants were told that since they were standing out of reach of the keyboard, they couldn't enter their answer. They were instructed not to voice their answers aloud but keep count of how many they had calculated correctly. The time given for participants to respond in the standing phase gradually reduced. Norming tests confirmed that this made the test considerably harder, but since they were not inputting their answers, only the participants themselves could ever know their score on the standing phase.

Once they had completed the study, the junior experimenter asked two questions, with the order counterbalanced between participants. The neutral question was "Did you feel the second stage took more or less time to complete?" The critical question was "Did your performance improve on the second test?". Participants' body motion was captured from the time she began asking the question to the end of their reply.

Participants who answered 'yes' to the critical question received their £5 reward and were categorised as liars. Even though it is possible that the participants scored higher on the second test, the increased difficulty made this unlikely. Therefore overall, we assumed that people who said yes were more likely to be deceptive than those who said no.

The Accident At the start of the experiment, while the participant was signing the consent form, the senior researcher precariously placed a laptop down on a table saying, "I've got that report of yours on my laptop. Remind me about it at the end". After the first stage of the maths test, the senior experimenter left the lab and the junior experimenter prepared them for the standing phase. While walking backwards, the junior experimenter knocked into the laptop that had been left on the table edge, and sent it crashing to the floor. She exclaimed loudly, made eye contact with the participant and said, "Thank God the cameras were off". Therefore, only the participant witnessed this 'accident'.

After the second maths test, the senior experimenter came back to the lab and told the participant that he needed to take a backup copy of the data. While the junior experimenter was stood in a corner of the lab preparing herself to leave, he asked the participant the neutral question, "Did the maths experiment run okay?" He then opened his laptop and attempted to turn it on without success. He then turned to the participant and asked the critical question, "Did you see anything happen?" During both questions and the participant's replies, their body motion was recorded. Participants were categorised as liars if they denied knowledge about the incident, and as truth tellers if they made any reference to the accident or the junior participant.

Debriefing Following the experiment, participants were fully debriefed about the true nature of the experiment and asked if they suspected that deception was being investigated. We framed all their behaviour in a positive light. For example, if they chose to deceive the senior experimenter about the accident, we referred to this as their choice to 'protect' the junior researcher. Contact details of psychological services were provided in the event that they felt concerned about deceiving or being deceived.

Data Analysis

Marker positions were reconstructed offline using the Vicon Nexus software. Standard procedures were used for identifying markers and excluding noise. For each marker we calculated the distance it moved in 5msec. We summed those values for series of 20 frames to get the total number of millimeters travelled in each 100ms period. Finally, across every marker and across every 100ms period during the data capture, we averaged those values. Our measure of general body motion is operationalized as the average distance in millimeters that a marker traveled every 100ms.

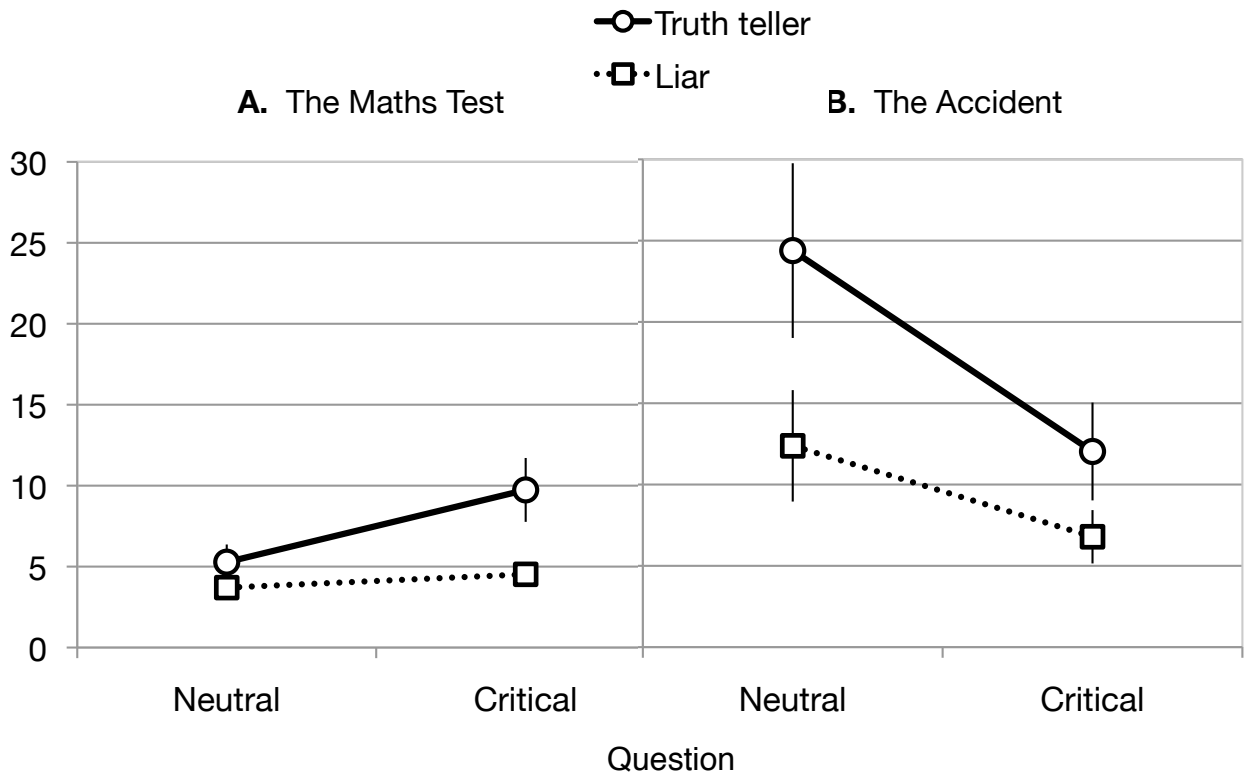


Figure 1. Total upper body motion during neutral and critical questions in the maths test and accident situations

Results

Participants data was discarded if they guessed our real hypothesis, in cases of experimenter error or deviations from the script. The complexity of the equipment and the spontaneous nature of the behaviour meant that data loss was a risk in this experiment. In the event, 34 (out of 128) trials produced unusable data. Since a full analysis of the data would require that we have data for all four periods for all subjects, we carried out planned comparisons on the maths deception and the accident data separately, giving us 18 and 23 subjects respectively. However, the same pattern of means and significance was still found when we analysed the smaller set of 16 subjects who had a full set of data for each cell.

This study is a quasi experiment, as participants placed themselves in the truth or lie conditions rather than being assigned. Overall, they chose to lie 61% of the time, allowing a between subjects comparison of the body motion of truth tellers and liars. Their (dis)honesty was consistent: 73% of participants either lied or told the truth on both occasions. Body motion data were analysed in a 2 (deception: truth teller/liar) x 2 (question type: neutral/ critical) ANOVA.

The Maths Test

Participants moved less if they attempted to deceive the experimenter and lie about their maths score. They moved more when asked the critical question than when asked the neutral question, though the difference between truth tellers and liars was greater in response to the critical question. This pattern of results is shown in panel A of *Figure 2*, and supported by a main effect of deception ($F(1,21)=7.97$, $p=.01$), a main effect of question ($F(1,21)=9.04$, $p=.007$) and a significant interaction ($F(1,21)=6.24$, $p=.021$). Post hocs show that the participants did not significantly differ in their motion in response to the neutral question, but moved significantly more when telling the truth in response to the critical question.

The Accident

Participants moved less if they attempted to deceive the experimenter and deny knowledge of the accident. They also moved less overall in response to the critical question. This pattern of results is shown in panel B of *Figure 2*, and supported by a main effect of deception ($F(1,21)=7.97$, $p=.01$) and a main effect of question ($F(1,21)=9.04$, $p=.007$). The interaction was not significant ($F(1,21)=1.11$).

General Discussion

When people spontaneously lied to us, they reduced their bodily motion. Behind this simple finding, robustly supported by our data, lie two more nuanced stories. One concerns the directionality of the relationship between body motion and deceptive behaviour. The other concerns the difference between the two situations and the two types of lie.

In the maths test, there was no obvious negative consequence to the participants' lie. The junior experimenter gave clear signs that she would prefer a particular answer, and participants were given a monetary reward for providing it. It was clear from the situation that there was no way in which participants could be found out if they lied, since they could be the only people who knew the truth. Therefore, participants had everything to gain and nothing to lose by lying. In this situation, participants' body motion was not different in response to the neutral question, but those who chose to lie for the reward showed less motion when answering the critical question.

This result aligns with other findings (Vrij, 2000) that during cases of deception, bodily motion can decrease. One explanation is that even though there is no logical way that another person can know that we are telling a lie, we suffer from an 'illusion of transparency' (Gilovich, Medvec & Savitsky, 1998). We are prone to thinking that since our own internal mental states are highly salient to us, they must be at least partially visible to others. Therefore, when we think about something that we don't want others to know, we try to suppress our overt actions in an attempt to suppress the (nonexistent) cues to our mental states. As often happens, it is not the lie that causes people to be caught, but the cover-up.

In the accident situation, the motivation and the consequences for lying are quite different. By lying, participants are acting in the interest of another person, the junior experimenter. Our intention was that the participants would feel some affiliation with her, due to her overt friendliness to the participants and their similarity in age and position. In sharp contrast, the senior experimenter asserted his authority over everyone else in the lab, and was curt and unpleasant when speaking to the participant. By lying to him, the participants are aligning their interests with their in-group, which is a strong motivation for social behaviour. However, unlike the maths test situation, there could conceivably be negative consequences to this lie. Something did indeed happen to the laptop, and it's possible that the senior experimenter could find out what happened in the future - perhaps the junior experimenter would confess. In this case, the participants would be discovered to have lied to someone in authority.

Both the motivation (DePaulo et al., 2003) and the possible consequences of the social lie are related to affective outcomes. In contrast to the maths lie, where material reward is at stake, in the accident situation, participants lie to foster an affiliation between themselves and the junior experimenter, but risk the aversive consequences of lying to an authority figure. We hypothesise that these differences produced an unusual feature of our data.

For the accident situation, the difference between liars and truth tellers emerged in responses to both the neutral and the critical questions. The neutral question was always asked before the critical question in the accident situation, as pilot studies showed that it seemed very unnatural for the experimenter to interrogate the participant about his broken laptop, and then switch to innocuous questions.

So why is it that participants who are going to lie to the experimenter in the near future already show a distinctive pattern of body motion when answering his neutral question? In looking back over the situation we constructed for participants, it seems that they may have already been thinking about the laptop and the incident they witnessed during the neutral question. While asking the neutral question, "Did the maths experiment run okay?" the senior experimenter was walking towards the table where the broken laptop was sat. This, coupled with the fact that he mentioned he was taking a backup of the data, makes it plausible to suggest that the participants realised he was about to use his laptop. At that point, perhaps they were considering the affective consequences of him discovering the accident, accusing the junior researcher, and asking them for information. In other words, even during the neutral question, participants' body motion was revealing their relation to the whole scenario of the accident and the two different researchers, and their own potential involvement.

This claim brings us to the second issue raised by these findings. Is it the case that there are some individuals, or some individuals' moods, that correlate with higher levels of bodily motion and higher levels of honesty? Or does the act of forming a lie or preparing to tell the truth produce a particular pattern of bodily motion? In the context of the accident situation, for example, it could be that during the neutral question some individuals are feeling heightened anxiety (because of what they witnessed), and that anxiety produces more bodily motion and higher rates of telling the truth. Alternatively, it could be that during the neutral question, some participants are already acting to suppress their overt behaviour as they prepare to tell a lie to the experimenter, or at least, distance themselves from the awkward situation.

In short, does body motion reveal differences between *people* who tell the truth and people who lie, or does it reveal differences in the *process* of lying and truth telling? There is some evidence for the latter proposal in our maths test situation. If it were true that some people simply move more and are more disposed to honesty, then we would expect to see truth tellers moving more in response to the neutral maths question as well as the critical question. Furthermore, supporting evidence for a direct link between motion and deception comes from experiments which instruct participants to lie or tell the truth, and thereby cause differences in bodily movements (Duran et al., in press). At present though, our data are equivocal on this point, and calls for further investigation.

Conclusion

People who spontaneously lie, or are about to lie, showed reduced body motion in our experiment. Though this pattern was found across two different types of situation, we are not rushing to make any claims to have found a unique bodily signature for deceptive behaviour. Differences in the two types of situation produced distinct patterns in degree of bodily motion and the conditions under which it emerged. We have speculated that these bodily differences are related to differences in the underlying motivations and consequences of deception in the two situations. We take

this work as establishing that motion capture systems are a new telescope that we can point at deception, and reaffirming the complexity of cognitive and affective states that underlie spontaneous deceptive behaviour.

It remains the case that almost no-one is much better than a coin toss at detecting deception in others (Bond & DePaulo, 2006). A notable exception are the FBI interrogators trained and tested by Ekman and O'Sullivan (1991). In their article, 'Who can catch a liar?' they reported a deception detection rate of 64%. Recently Bond (2008) came across a passage in Ekman's (1992) book giving further details of that experiment. As it is described, the study has a striking similarity to our own maths test situation that was designed to evoke spontaneous lies:

> "Immediately after taking the test I would give the correct answers. Then I asked them to raise their hands if they got all ten correct, nine correct, and so forth. I tallied the results on a blackboard so that they could evaluate their own performance against that of their group. . . . In September 1991, our findings on these professional lie catchers were published"
>
> (Ekman, 1992; pp. 282–285)

As Bond (2008) concluded, "Who can catch a liar? It would appear to be Secret Service agents who get to score their own tests." It is an intriguing thought that these FBI agents, might themselves have displayed signature patterns of bodily movement that betrayed the fact that they were actually lying about their ability to detect liars.

Acknowledgments

We would like to thank Geoff Bird, Rick Dale, Natasha Kirkham and Michael Spivey for many insightful discussions. This research was carried out as part of NME and SB's undergraduate dissertations, and authorship is equal.

References

Bond, C.F. (2008). A few can catch a liar, sometimes: Comments on Ekman and O'Sullivan (1991), as well as Ekman, O'Sullivan, and Frank, (1999). *Applied Cognitive Psychology, 22*, 1298-1300.

Bond, C.F., & DePaulo, B.M. (2006). Accuracy of deception judgments. *Personality and Social Psychology Review, 10*(3), 214-234.

Bond, C.F., & DePaulo, B.M. (2008). Individual differences in judging deception: Accuracy and bias. *Psychological Bulletin, 134*(4), 477-492

Buller, D. B., & Burgoon, J. K. (1996). Interpersonal deception the- ory. *Communication Theory, 6,* 203–242.

Burgoon, J., Jensen, M.L., Meservy, T., Kruse, J., & Nunamaker, J.F. (2005). Augmenting human identification of emotional states in video. *Proceedings of the International Conference on Intelligent Data Analysis.*

Casasanto, D., & Dijkstra, K. (in press). Motor action and emotional memory. *Cognition*.

Chartrand, T.L., & Van Baaren, R. (2009). Mimcry. In M.P. Zanna (Ed.), *Advances in Experimental Social Psychology, Volume 41*. Ocford, UK: Academic Press.

DePaulo, B. M., Kashy, D. A., Kirkendol, S. E., Wyer, M. M., & Ep- stein, J. A. (1996). Lying in everyday life. *Journal of Personal- ity and Social Psychology, 70*, 979–995.

DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to deception. *Psycho- logical Bulletin, 129,* 74–118.

DePaulo, B. M., Stone, J. L., & Lassiter, G. D. (1985). Deceiving and detecting deceit. In B. R. Schenkler (Ed.), The self and social life (pp. 323-370). New York, NJ: McGraw-Hill.

Duran, N. D., Dale, R., & McNamara, D. S. (accepted). The action dynamics of overcoming the truth. *Psychonomic Bulletin & Review*.

Ekman, P. (1992). Telling lies: Clues to deceit in the marketplace, politics and marriage. New York: Norton.

Ekman, P. & O'Sullivan, M. (1992). Who Can Catch a Liar? *American Psychologist,* 46(9), 913-920

Gilovich, T., Savitsky, K., & Medvec, Victoria H. (1998). The illusion of transparency: Biased assessments of others' ability to read one's emotional states. *Journal of Personality and Social Psychology, 75*(2), 332-346.

Kohnken, G. (1987). Training police officers to detect deceptive eye-witness statements:Does it work? *Social Behaviour, 2 ,* 1-17.

McClintock, C. C., & Hunt, R. G. (1975). Nonverbal indicators of affect and deception in an interview setting. *Journal of Applied Social Psychology,* 5, 54-67.

Miles, L.K., Nind, L.K. & Macrae, N.C.(2010). Moving Through Time. *Psychological Science,* published online January 8, 2010

Porter, S., & ten Brink, L. (2009). The truth about lies: What works in detecting high-stakes deception. *Legal and Criminological Psychology, 00,* 1-21.

Spivey, M.J. (2007). *The Continuity of Mind*. Oxford: OUP.

Vrij, A. (1993). Credibility judgements of detectives: The impact of nonverbal behavior, social skills, and physical characteristics on impression formation. *Journal of Social Psychology, 133*(5), 601-610.

Vrij, A. (2000). *Detecting Lies and Deceit: The Psychology of Lying and Implications for Professional Practice*. Chichester: John Wiley & Sons.

Vrij, A. (2008). *Detecting Lies and Deceit: Pitfalls and Opportunities*. Chichester: John Wiley & Sons.

Vrij, A., & Mann, S. (2004). Detecting deception: The benefit of looking at a combination of behavioral, auditory and speech content related cues in a systematic manner. *Group Decision and Negotiation, 13*(1), 61-79.

Wojnowicz, M.T., Ferguson, M.J., Dale, R., & Spivey, M.J. (2009). The self-organization of explicit attitudes. *Psychological Science, 20*(11), 1428-1435.

Exploring the mental space of autonomous intentional agents

Peter C. Pantelis and Jacob Feldman
(petercp@eden.rutgers.edu, jacob@ruccs.rutgers.edu)
Department of Psychology, Center for Cognitive Science, Rutgers University-New Brunswick
152 Frelinghuysen Road, Piscataway, NJ 08854 USA

Abstract

How do we use the motion of animate objects to make inferences about their intentions? We investigate this question using displays containing a number of autonomous, independently programmed agents moving about the screen and interacting with each other. Each agent behaves according to an independent autonomous program, controlled by a small number of parameters that define its "personality." We probe subjects' impressions of the similarities among the behaviors of the various agents, and then use multidimensional scaling to recover the subjective parameters defining the mental space of agent types. The most important variable turns out to be one that determines how the agent reacts to a nearby agent at one critical distance. A followup experiment suggests that variation along this parameter contributes to modulating a higher-level percept of how "hostile" or "friendly" the agents appear to be.

Keywords: animate motion perception; theory of mind; intentionality; action understanding; goal inference.

Introduction

Intelligent agents can and must distinguish between animate and inanimate objects that they encounter. Even infants make this distinction, and apparently possess a naïve theory of other beings' mental states and intentions (Gergely, Nádasdy, Csibra, & Bíró, 1995; Keil, 1994; Johnson, 2000). Socially intelligent agents naturally conceive of other humans as animate, mentalistic agents with independent perceptions and motivations. We further benefit from being able to infer the *intentions* of other agents in the environment. This is essential for understanding and predicting others' behavior, a prime skill both for chess players contemplating their moves, and gazelles and lions engaging in mutual scrutiny on the African plain.

This research explores how adult subjects use an observed agent's motion to make inferences about its mental architecture. For this task, motion is only one cue among many (Gelman, Durgin, & Kaufman, 1995), but it is a particularly salient one, with subjects readily ascribing intentionality even to simple moving geometric figures (Heider & Simmel, 1944). A handful of studies have shown that varying the motion of simple geometric figures along certain parameters (e.g. speed, trajectory) can influence the perception of animacy and intentions (Dittrich & Lea, 1994; Tremoulet & Feldman, 2000, 2006). But the factors determining these percepts are still very poorly understood.

Baker, Tenenbaum, and Saxe (2006) and Baker, Saxe, and Tenenbaum (2009) have proposed a Bayesian framework for "inverse planning," that is, inferring or estimating the goals or intentions of an agent assumed to be rational. Sloman, Fernbach, and Ewing (2009) use Bayesian belief networks to describe causal reasoning in the domain of morality. We too presume a Bayesian formulation of the problem, in which the goal is to assign a posterior probability to the mental state (behavioral disposition, goal set, payoff matrix, or some other representation of the other agent's mind) A on the basis of its motion:

$$p(A|\text{motion}) \propto p(\text{motion}|A)p(A). \quad (1)$$

Ultimately, such an inference maps a visual input (the motion observed) onto a distribution of possible agent types. The prior $p(A)$ is defined over the set of possible agent types, that is, the space of behavioral dispositions the observer is in principle willing to entertain as explanations for the observed motion. The nature and structure of this space have been discussed only very speculatively in the literature; Barrett, Todd, Miller, and Blythe (2005) have argued that it probably includes such natural action classes as chasing, courting, following, guarding, fighting, and playing. Some studies have presented subjects with scenes constructed to resemble these different "natural categories" of dyadic interaction, and demonstrate that subjects are reliably able to categorize these scenes, even in degraded forms for which motion is the only salient cue (Barrett et al., 2005; McAleer & Pollick, 2008).

In contrast to most previous experiments, the scenes we present to subjects have *not* been pre-constructed to convey particular categories of interaction. Our aim is to show subjects a broad array of agent interactions—from a richer and more general collection of possibilities—in an attempt to allow subjects' minds to impose *their own* structure on the agent space. The way we produce the desired scenes is also novel: We program the agents inhabiting these scenes to behave *autonomously*, which results in often chaotic multi-agent interactions that we cannot predict in advance.

In Exps. 1 and 2, we use multidimensional scaling (MDS) in an attempt to extract the natural clusters and cleavages present within this stimulus space of intentional behavior. Exp. 3 is explicitly designed to help clarify the results of Exps. 1 and 2 by unraveling the "semantics" of the features uncovered by the MDS. Displays were programmed using the breve Simulation Environment (Klein, 2002), an open-source software package freely available at `http://www.spiderland.org`.

Programming Lifelike Automata

In designing and coding the agent behaviors, we aimed to employ a simple programming scheme that would impose minimal structure on the agents' interactions but, nonetheless, would be capable of producing a rich variety of lifelike agent

behaviors.[1] We programmed the triangular agents to behave autonomously, each running its own independent program. Inspired by the work of Braitenberg (1984), we aimed to create rule-governed agents which, notwithstanding the simplicity of their programs, yield vivid and lifelike behaviors that give subjects a strong impression of intentions.

Agent design Rather than presenting subjects with prefabricated animations, we populate simulations with autonomous agents and then allow these simulations to run for a predetermined length of time (15 seconds). Each agent starts off in the simulation environment with a randomly-assigned velocity and location. The agent always orients one vertex of its triangular body (that which lay on its axis of symmetry) in the direction of its movement, inducing the impression that the front end is the agent's "head" (see Tremoulet & Feldman, 2000). When an agent either collides with another agent or the edge of the scene, it "bounces off" for one iteration of the simulation.[2]

At each iteration of a simulation, an agent finds the nearest *other* agent within the scene and then accelerates toward or away from it to an extent determined by a set of six parameters contained in its program. The parameters control the direction and magnitude of the agent's acceleration—relative to the nearest other agent—at six respective distances from this other agent: 0-5 "units", 5-10, 10-20, 20-40, 40-70, or $>$ 70. A schematic of these 6 radii around an agent, along with a snapshot of Experiment 1, is shown in Figure 1.

One example agent might approach another agent from afar but then veer away as it gets to a closer radius. Others might consistently accelerate away from another agent. Depending on how this other agent is programmed, their interaction might resemble chasing/fleeing, or one pushing the other, or even one agent circling the other.

We constructed a pool of 12 agents, each with 6 randomized parameters within the programming scheme.

Experiment 1

Method

Subjects Eight students between the ages of 18 and 24 participated in an approximately one-hour experimental session in exchange for course credit.

Stimuli Scenes were presented to subjects on a 1440 x 900 LED display, on a 15 inch MacBook Pro laptop with a 2.2 GHz dual core processor. The simulation environment itself measured 33.0 x 16.5 cm, and the viewing distance was approximately 45 cm. The programming library employed units that were equivalent to 8.7 units/cm. The triangular agents had bases of 1 unit length and heights of 4 unit length.

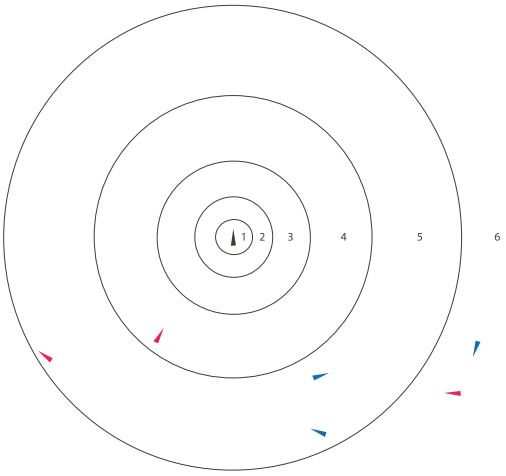


Figure 1: Screenshot from Experiment 1 (colors inverted), with black circles and numbers superimposed onto the scene to help illustrate the programming scheme for the automata. The black automaton in this scene accelerates toward or away from the nearest other agent in the scene. The direction and magnitude of this acceleration depends on the distance to this nearest other agent, with possible distances divided into six zones. Zone #5 seems to be most psychologically relevant.

Procedure In each 15 s scene, the subject observed 7 agents interacting: 3 red, 3 blue, and 1 white. The reds behaved according to the same parameters as the other reds, the blues according to a different set of parameters, and the lone white according to a third set of parameters. The agents were drawn from a larger 12 agent pool; thus, there were 220 possible triads of these 12 agents.[3] For each scene, one of these 220 triads was selected at random, and each of the three programs in the selected triad was randomly assigned to either red, blue, or white. Each subject saw 220 such scenes, exhausting the possible triads.

Subjects were openly encouraged to construe the triangular agents as animate. At the end of each scene, they were asked "Is the white agent behaving more like a red, or more like a blue?" They answered by clicking on a button in a dialog box.

We constructed a 12 x 12 symmetric distance matrix for each subject, to be fed into the individual differences multidimensional scaling (MDS) algorithm (INDSCAL/ALSCAL; Takane, Young, & Leeuw, 1977). Within this matrix, an agent was assigned a distance of 0 from itself. As two different agents appeared in the same trial of an experimental session 10 times, the distance in this matrix between any two agents was initially set at 11.

[1] It is important to note that the programming scheme we employ here is only one possible choice among many. The design of lifelike agents is a complex and multifaceted problem that extends far beyond the scope of our research. For us, these simple automata are merely tools for aiding an empirical study of the perception of intention.

[2] In Experiment 1, this sometimes resulted in jerky and unnatural-looking behaviors at agent collisions, so in Experiments 2 and 3 we changed collision behavior slightly: agents in these experiments bounced off each other for a full .2 s at some random velocity vector.

[3] Strictly speaking, because the status of the white agent in each trial is special, and, as a result, during a given trial the subject cannot respond that he actually believes the blue and red agents to be most alike, 660 possible arrangements actually exist. Rather than show all 660 possibilities, we randomized the procedure so that no agent type would be more or less likely to be "white" during a trial. Nevertheless, this presents a source of noise in the data, and we altered the procedure in Experiment 2 to address this issue.

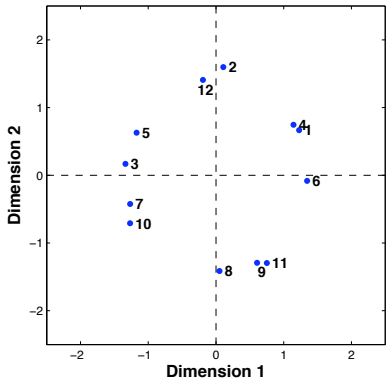


Figure 2: The 2-dimensional MDS solution for the 12 agents, fitting data from Experiment 1.

If the subject chose "red," then the agent whose programming was used for the red agents in this trial was made to be closer together (more similar) in this distance matrix with that of the white agent, and likewise for if the subject chose "blue." That is, the distance between these two agents in the matrix was reduced by 1. Previous studies have used similar methodologies to gauge subject similarity ratings of visual stimuli (e.g., Kahana & Bennett, 1994; Pantelis, van Vugt, Sekuler, Wilson, & Kahana, 2008).

Results and Discussion

We derived a 2-dimensional (2D) MDS solution in order to visualize the space of agents that subjects (on average) perceived (see Figure 2). For this amount of points in the space, the INDSCAL algorithm allows for fits of 2-5 dimensions. Deriving higher-dimension solutions will always result in better fits to the experimental data.[4] However, a higher number of dimensions would be even more difficult to interpret than the 2 condensed dimensions we present, and even a 5-dimensional fit would probably be a condensed version of the true amount of psychologically relevant dimensions in this agent space (which could hypothetically be even higher than the total number of agents in our sample).

A 2D solution allows for the easiest visualization of the inter-agent distances, an important motivation for using the MDS analysis in the first place. If interesting structure emerged only in higher-dimensional fits for these data, this might have justified using these MDS solutions. However, we actually found the clearest and most interesting structure within a 2D fit.

The most striking aspect of the space is its ring-like structure, similar to what one would observe in a 2D MDS plot of the color wheel (see Shepard, 1980). The significance of this ring structure was not immediately clear, in part because MDS dimensions are in general not self-explanatory but rather pull out subjectively primitive parameters. Exp. 3, presented below, was designed to help clarify the nature of the parameter exhibited in this ring.

[4]While we examined a scree plot of the pooled data from Experiments 1 and 2, we do not display it here due to space considerations. This scree plot does not demonstrate a clear "elbow" favoring one particular number of dimensions over another.

Table 1: Correlations (r[10]) between programmed parameters (rows) and MDS dimensions (columns). Bold font represents $p <$ .01

	MDS Dimension 1	MDS Dimension 2
Parameter 1	-.070	.384
Parameter 2	-.275	-.074
Parameter 3	.527	.199
Parameter 4	.411	-.375
Parameter 5	**-.801**	.093
Parameter 6	.459	.197

The goal of the present experiments was not, per se, to see how the somewhat arbitrary parameters with which we programmed the agents mapped to subjects' percepts of the agents' behaviors. Rather, we had aimed to infer the structure of the perceptual space itself. Nonetheless, relating these parameters to the MDS dimensions was a useful step in understanding the 2D MDS space.

Subjects' perception of the agents' behaviors arises from some complex interaction of its underlying programming and the chaotic interaction with other agents that arises during each unique simulation. This contributed to there being many individual differences between subjects' results; few subjects' distance matrices showed obvious correlation. However, one of the 6 parameters with which we programmed each agent was indeed strongly correlated with one of the MDS coordinates (see Table 1). This parameter controlled how an agent behaved when the nearest other agent was between 40 and 70 units (4.6 to 8.1 cm) away from it. This finding is addressed further in the Experiment 2 discussion.

Experiment 2

In Exp. 2, we adjusted the basic methodology of Exp. 1 in hopes of reducing the amount of noise in the data. The most significant change was to allow the subject to control one of the agents in each simulation via the mouse. The chance to interact with the simulated agents would, we expected, allow the subject to glean more information about the other agents' behaviors during the short 15-second display time and thus promote stronger impressions of the agents' "personalities" than was possible in Exp. 1.

Method

Subjects Seven students between the ages of 18 and 23 participated in an approximately one-hour experimental session in exchange for course credit.

Stimuli We presented scenes to subjects on an eMac with a 17 inch (16 inches viewable) monitor and a 1152 x 864 display. The monitor refresh rate was 80 Hz and the computer had a 1.25 GHz processor. The simulation environment it-

self measured 25.4 x 16.5 cm, and the viewing distance was approximately 45 cm.

Exp. 2's scenes were populated with triangular agents of the same size and programmed under the same scheme as in Exp. 1. We used the same pool of 12 agents from Exp. 1, each which had been created with 6 randomized parameters within the programming scheme.

Additionally, the subject controlled one agent with the mouse: a white circular agent 4 units in diameter. The automatic agents reacted to the subject-controlled agent in the same manner as any other triangular agent in the simulation.

Procedure In each 15 s scene, the subject observed 6 agents and controlled 1 agent. 2 agents were red, 2 were green, 2 were blue, and the subject-controlled agent was white. The reds would behave according to the same parameters as the other reds, the greens according to a different set of parameters, and the blues according to a third set of parameters. The agents were drawn from a larger 12 agent pool; thus, there were 220 possible triads of these 12 agent programs. For each scene, one of these 220 triads was selected at random, and then each of the three programs in the selected triad was randomly assigned to either red, green, or blue. Each subject saw 220 such scenes, exhausting the possible triads.

Subjects were openly encouraged to construe the triangular agents as animate, and were instructed that how agents of a certain color behaved during one trial would have nothing to do with how they behaved in subsequent trials. At the end of each scene, they were asked to determine which color of agent behaved *least* like the other two—that is, which was most different: red, green, or blue? They responded by key press, at which point the next trial began.

As in Exp. 1, we constructed a 12 x 12 symmetric distance matrix for each subject, to be fed into the individual differences multi-dimensional scaling (MDS) algorithm. For each trial, the two non-chosen agents in the odd-one-out procedure were made more similar within this distance matrix.

Results and Discussion

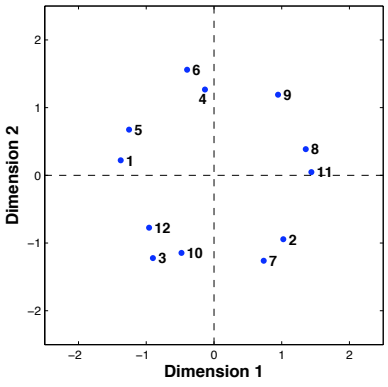


Figure 3: The 2-dimensional MDS solution for the 12 agents, fitting data from Experiment 2.

Table 2: Correlations ($r[10]$) between programmed parameters (rows) and MDS dimensions (columns). Bold font represents $p < .05$

	MDS Dimension 1	MDS Dimension 2
Parameter 1	-.129	-.147
Parameter 2	-.529	-.050
Parameter 3	.096	.195
Parameter 4	.548	.200
Parameter 5	-.310	**-.619**
Parameter 6	.249	.233

Once again, we derived a 2D MDS solution in order to visualize the space of agents that subjects (on average) perceived, and we once again observed a ring-like structure in the space (Fig. 3).

The MDS solutions for the two experiments—processed representations of subjects' raw similarity matrices—were correlated with each other. Dim. 1 of Experiment 1's MDS was strongly correlated with Dim. 2 of Experiment 2's MDS [$r(10) = .713, p < .01$]. Dim. 2 of Experiment 1's MDS was weakly (and negatively) correlated with Dim. 1 of Experiment 2's MDS [$r(10) = -.551, p = .063$]. (The direction of these correlations is arbitrary and unimportant, but helpful in relating the 2D MDS spaces presented in Figures 2 and 3.) These correlations provide some assurance of the robustness and psychological reality of the subjective mental spaces that we have uncovered.

As shown in Table 2, Dim. 2 in Experiment 2 correlated significantly with parameter 5 of the agents' programming. This is consistent with the results of Experiment 1, where Dim. 1 had been correlated with this same parameter. Apparently, how an automaton reacted to (i.e. accelerated toward or away from the direction of) the nearest other agent in the simulation when that agent was 40-70 units away (10 to 17.5 times the length of an agent) was a psychologically important variable.

We wondered if the prominence of this parameter in subjects' judgments was actually an artifact of the frequency with which interactions at this distance actually occurred in the displays. But the data do not bear this out. Because the entire displays were recorded (10 frames/second), we could assess the proportion of the time the inter-agent distance between any automaton and its nearest other agent was within each of the six intervals corresponding to the six underlying programmed parameters. The two most common distances between an automaton and its nearest other agent during a simulation were 0-5 units (0-1.25 agent lengths) and 20-40 units (5-10 agent lengths). 40-70 units (10-17.5 agent lengths) was only the fourth most common inter-agent distance. The pivotal role of this inter-agent distance is not an artifact, but rather reflects a genuine cognitive focus on behavioral interactions at this distance.

Experiment 3

The results of the first two experiments were qualitatively similar, and we therefore choose to pool data from all 15 subjects for the following analysis and discussion. The 2D MDS solution for these pooled subjects reveals an even cleaner ring structure (see Figure 4). But what does it mean as we travel around this ring?

In the combined MDS, Dim. 1 is connected to how an agent behaves when the closest other agent is between 10-17.5 agent lengths away (i.e. programmed parameter #5). Agents low on Dim. 1 all tend to accelerate away from the nearest other agent; agents high on Dim. 1 tend to accelerate toward the nearest other agent. The meaning behind Dim. 2 is less straightforward. While this dimension is clearly not independent from Dim. 1, it is uncorrelated with any of the programmed agent parameters. Hence we turn to further psychophysics to provide evidence about its meaning.

We hypothesize that a potential "friendly" versus "hostile" dimension emerges from the interaction of these two MDS dimensions. This hypothesized dimension would be neither orthogonal nor redundant with whether an agent accelerates toward or away from another agent at a certain distance—say, the distance with which programmed parameter #5 is concerned. When an agent moves in the direction of another, it may, for instance, appear to be aggressive or merely curious.

Method

Subjects Seven students between the ages of 18 and 24 participated in an approximately half-hour session in exchange for course credit.

Stimuli and Procedure We presented scenes to subjects under the same viewing conditions as Exp. 1. We again populated the simulations with the pool of 12 agents employed in Exps. 1 and 2. During each trial, the subject watched 7 agents interacting for 15 seconds. Six of the agents were colored red and behaved under programs randomly selected from the pool of 12. The seventh, critical agent was colored blue, and the subject was instructed to attend to it. At the end of each trial, the subject was asked, "On a scale of 1-5, 1 being most hostile, and 5 being most friendly, how do you rate the blue agent?" The subject indicated his response on the keyboard. Each of the 12 agents in the pool was assigned the blue color for 8 of the session's trials, for a total of 96 trials presented in random order.

Results We first normalized each subject's responses, then calculated each subject's mean normalized response for each of the 12 agents observed over the experimental session. Then, averaging across subjects, we were able to get a sense of how friendly versus hostile subjects perceived each of the 12 autonomous agents. Figure 4 shows, on a gradient from red to green, what these perceptions were. The most hostile agents seem to be those which were high on MDS Dim. 1 and low on Dim. 2, while the friendlier agents tended to be low on Dim. 1 and high on Dim. 2. Agents low on both dimensions were quite neutral. Fig. 4 shows the space in a rotated coordinate frame so that the horizontal dimension optimally reflects the friendliness vs. hostility dimension. (All of the inter-agent distances and relationships have been preserved; only the "ring" has been rotated.) In the rotated space the projection of each agent's position onto the horizontal (i.e. the cosine of its angle relative to the horizontal) reflects its position along the friendly/hostile dimension. We regressed the subjects' mean friendliness rating against this variable and found a close fit ($r(10) = -.768, p < .01$, Figure 5). These data corroborate our hypothesis that the ring variable essentially reflects the degree of perceived friendliness or hostility each agent exhibited.

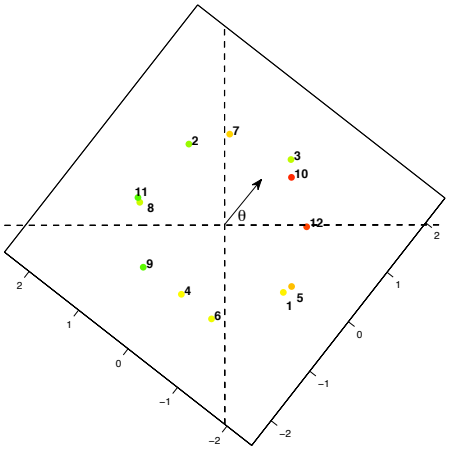


Figure 4: Pooled 2-dimensional MDS from Experiments 1 and 2. The 2D space is rotated about the origin so that the cosine of the agent angle (relative to the horizontal) best predicts how subjects rated the agents along the "hostility" versus "friendliness" dimension. "Hostility" versus "friendliness" is represented for each agent here with a color gradient from red (most hostile) to green (most friendly), with yellow being neutral.

General Discussion and Conclusions

These experiments were designed to probe the underlying structure of the agent space perceived by subjects as they watched autonomously programmed agents interacting in a dynamic scene. In Experiments 1 and 2, the MDS approach succeeded in revealing certain aspects of this perceptual space: a ring-like structure, which—in Experiment 3—we attempted to connect to a dimension of perceived hostility versus friendliness in the agents. One of the low-level parameters controlling the behaviors of the agents contributed to this more abstract percept: that which controlled inter-agent reactive behavior at one critical distance. We conclude that this reflected one perceptually critical inter-agent zone upon which subjects based their interpretations of the agents' intentional behavior.

From the results of Experiment 3, we further conclude that "hostility" versus "friendliness," or something akin to this di-

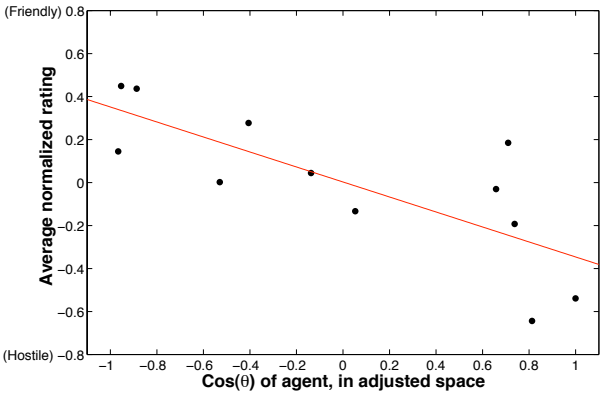


Figure 5: Cosine of the agent's angle in MDS space (see Fig. 4), plotted against how subjects, on average, rated them (from hostile to friendly). Best linear fit is drawn in red.

chotomy, appears to produce an especially salient partition in subjects' perceptual space. In other words, after first surmising that an object in the world has intentions (i.e., is animate), a next step for the cognitive machinery might be an attempt to guess whether these intentions are bad or good.

This work represents one step in what we hope is a fruitful new direction. Programming agents automonously, and asking how subjects' interpretations of these agents' behavior relates to the actual programs they are carrying out, allows one to pursue a true "psychophysics of intention," in which we explore the relationship between the perceived intention and the "actual" intention present in the agent's autonomous program. In future experiments, employing displays of potentially far more complex behavioral interactions, we hope to uncover correspondingly more complex structures in the intentionality percept.

Acknowledgments

This research was supported by the National Institutes of Health (NIH EY15888), the NSF IGERT program in Perceptual Science (NSF DGE 0549115), and a grant from the Hellenic University Club of New York.

References

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*, 329–349.

Baker, C. L., Tenenbaum, J. B., & Saxe, R. R. (2006). Bayesian models of human action understanding. *Advances in Neural Information Processing 18*.

Barrett, H. C., Todd, P. M., Miller, G. F., & Blythe, P. W. (2005). Accurate judgments of intention from motion cues alone: A cross-cultural study. *Evolution and Human Behavior*, *26*, 313–331.

Braitenberg, V. (1984). *Vehicles*. Cambridge, MA: MIT Press.

Dittrich, W. H., & Lea, S. E. G. (1994). Visual perception of intentional motion. *Perception*, *23*, 253–268.

Gelman, R., Durgin, F., & Kaufman, L. (1995). Distinguishing between animates and inanimates: not by motion alone. In D. Sperber, D. Premack, & A. J. Premack (Eds.), *Causal cognition: A multidisciplinary debate* (pp. 151–184). New York, NY: Oxford University Press.

Gergely, G., Nádasdy, Z., Csibra, G., & Bíró, S. (1995). Taking the intentional stance at 12 months of age. *Cognition*, *56*, 165–193.

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *American Journal of Psychology*, *57*, 242–259.

Johnson, S. C. (2000). The recognition of mentalisic agents in infancy. *Trends in Cognitive Development*, *16*, 637–656.

Kahana, M. J., & Bennett, P. J. (1994). Classification and perceived similarity of compound gratings that differ in relative spatial phase. *Perception & Psychophysics*, *55*, 642-656.

Keil, F. C. (1994). The birth and nurturance of concepts by domains: The origins of concepts of living things. In L. Hirschfeld & S. A. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture.* New York: Cambridge University Press.

Klein, J. (2002). breve: a 3d simulation environment for the simulation of decentralized systems and artificial life. *Proceedings of Artificial Life VIII, the 8th International Conference on the Simulation and Synthesis of Living Systems.*.

McAleer, P., & Pollick, F. E. (2008). Understanding intention from minimal displays of human activity. *Behavior Research Methods*, *40*(3), 830–839.

Pantelis, P. C., van Vugt, M. K., Sekuler, R., Wilson, H. R., & Kahana, M. J. (2008). Why are some people's names easier to learn than others? the effects of face similarity on memory for face-name associations. *Memory & Cognition*, *36*(6), 1182–1195.

Shepard, R. N. (1980). Multidimensional scaling, tree-fitting, and clustering. *Science*, *210*(4468), 390–398.

Sloman, S. A., Fernbach, P. M., & Ewing, S. (2009). Causal models: the representational infrastructure for moral judgment. In D. Bartels, C. W. Bauman, L. J. Skitka, & D. Medin (Eds.), *Moral judgment and decision making: The psychology of learning and motivation* (Vol. 50). San Diego, CA: Elsevier.

Takane, Y., Young, F. W., & Leeuw, J. de. (1977). Nonmetric individual differences multidimensional scaling: an alternating least squares method with optimal scaling features. *Psychometrika*, *42*(1), 7-67.

Tremoulet, P. D., & Feldman, J. (2000). Perception of animacy from the motion of a single object. *Perception*, *29*, 943–951.

Tremoulet, P. D., & Feldman, J. (2006). The influence of spatial context and the role of intentionality in the interpretation of animacy from motion. *Perception & Psychophysics*, *68*(6), 1047–1058.

Actor-Observer Differences in Intentional Action Intuitions

Adam Feltz (ADFeltz@schreiner.edu)
Philosophy and Interdisciplinary Studies, 2100 Memorial Blvd.
Kerrville, TX 78028, USA

Maegan A. Harris (MAHarris@schreiner.edu)
Schreiner University, 2100 Memorial Blvd.
Kerrville, TX 78028, USA

Ashley L. Perez (ALPerez@schreiner.edu)
Schreiner University, 2100 Memorial Blvd.
Kerrville, TX 78028, USA

Abstract

Empirically minded researchers have begun exploring the "folk" notion of intentional action, often with surprising results. In this paper, we extend these lines of research and present new evidence from a radically new paradigm in exploration of folk intuitions about intentional action. Our results suggest that in some circumstances people make strikingly different judgments about intentions and intentionality as a function of whether the person brings about or observes an event. Implications for action theory and the experimental study of folk intuitions are discussed.

Keywords: Experimental Philosophy; Intentional Action; Actor-Observer Differences; Side-Effect Effect

Determining whether a person's behavior was intended or intentional is crucial for a host of important judgments such as assigning blame and praise. This part of human experience has been of central concern for philosophers of action (Mele, 1992). Many of these philosophers take themselves to be exploring the everyday or "folk" concept of intentional action (Adams, 1986; McCann, 1986, 2005; Mele, 1992). Some philosophers even write that "a philosophical analysis of intentional action that is wholly unconstrained by that [folk] concept runs the risk of having nothing more than a philosophical fiction as its subject matter" (Mele, 2001, 27). Empirically minded researchers (e.g., experimental philosophers) have helped shed light on this folk notion of intentional action, often with surprising results. In this paper, we extend these lines of research and present evidence using a new paradigm to study folk intuitions about intentional action. Our results suggest that in some circumstances people make strikingly different judgments about intentions and intentionality partially as a function of whether a person brings about or observes an event. Implications for traditional action theory and the study of philosophically relevant folk intuitions are discussed.

Experimental Philosophy and Action Theory

Arguably the best known studies in the experimental investigation of intentional action intuitions are Knobe's (2003a) harmful (underlined) and helpful (bracketed) chairman cases:

Harm/Help: The vice-president of the company went to the chairman of the board and said, "We are thinking of starting a new program. It will help us increase profits, but [and] it will also harm [help] the environment." The chairman of the board answered, "I don't care at all about harming [helping] the environment. I just want to make as much profit as I can. Let's start the new program. They started the program. Sure enough, the environment was harmed [helped]. (191)

The only difference between the two cases is the moral valence of the consequence of the chairman's decision. Remarkably, this shift in the moral valence of the consequence drastically changed people's intentionality judgments about the consequence: 82% of participants judged that the chairman brought about the harm to the environment intentionally whereas only 23% judged the chairman brought about the help to the environment intentionally. This general effect (the side effect-effect or the Knobe effect) has been replicated using similar scenarios (Cushman and Mele, 2008; Knobe, 2003a, 2003b, 2004a, 2004b) across cultures (Knobe and Burra, 2006), as well as across ages (Leslie, Knobe, and Cohen, 2006).[1]

Knobe-style cases feature *side effects*. If a consequence of an intended action is foreseen but not intended, then that consequence is a side effect of the intended action.[2] Side effects have been considered important test cases of some theories of intentional action. Just to take one example, Knobe-style cases have been argued to challenge a prominent view in intentional action—the Simple View (SV). According to the SV, if an agent intentionally performs an action *A* then the agent intends to *A*. Some philosophers have argued that the SV is supported by folk intuitions (Adams, 1986; McCann 1986, 2005). However, as judgments in Harm suggest, sometimes the folk make judgments that are contrary to the SV. If the harm to the environment is a side effect, then it is

[1]See Feltz (2007b) for a more detailed overview.

[2]See Cushman & Mele (2008) for a detailed definition of a side effect.

not intended. But, most people think that the harm is brought about intentionally. Hence, in some circumstances, many have the intuition that one can harm the environment intentionally without intending to do so. This pattern of intuitions seemingly falsifies that the SV is supported by folk intuitions (Nadelhoffer, 2006).

In the next two sections, we suggest that folk intuitions surrounding intentional action may be much more complicated than originally thought and may be influenced by a variety of factors including one's perspective.

Actor-observer differences

Actor-observer differences refer to a common effect where people who engage in behaviors (actors) see things differently than those who watch behaviors (observers). The traditional conception of the actor-observer asymmetry posits that the "actor's view of his behavior emphasizes the role of environmental conditions at the moment of action. The observer's view emphasizes the causal role of stable dispositional properties of the actor" (Jones & Nisbett, 1972, 80). While it is debatable whether this traditional conception is completely accurate (Malle, Knobe, & Nelson, 2007), some actor-observer asymmetries have been revealed in decisions made in risky environments (Fernandez-Duque & Wifall, 2007), moral judgments (Nadelhoffer & Feltz, 2008), and action explanations (Malle & Knobe, 1997; Malle, Knobe, & Nelson, 2007). To illustrate, consider one case from Nadelhoffer and Feltz (2008) where an actor-observer asymmetry was found:

Trolley: A trolley is hurtling down the tracks. There are five workers on the track ahead of the trolley, and they will definitely be killed if the trolley continues going straight ahead since they won't have enough time to get out of harm's way. There is a spur of track leading off to the side where another person is working. The brakes of the trolley have failed and there is a switch which can be thrown to cause the trolley to go to the side track. Imagine that you are an innocent bystander who happens to be standing next to the switch. You realize that if you do nothing, five people will definitely die. On the other hand, you realize that if you throw the switch, you will definitely save the five workers. However, you are also aware that in doing so the worker on the side track will definitely be killed as the result of your actions.[3]

Observers received the same scenario except 'you' was replaced with 'John' (along with appropriate verb conjugations). Participants were asked if flipping the switch is morally permissible and rated how much control over the situation one has. People given the 'John' version were more likely than those given the 'you' version to judge (a) that flipping the switch was "morally permissible" and (b) that John had control over the events.[4]

But why do actors and observers sometimes display this asymmetry? According to Malle, Knobe, & Nelson, one reason is that "we can expect that actors normally have better access to their own reasons than observers do and that they are normally more motivated to portray themselves as active, conscious, and rational agents" (2007, 508). Hence, because actors may be motivated to (a) portray themselves in a positive light and (b) have special access to their own reasons, they are prone to judge their own behaviors differently from others' behaviors. This explanation could account for the asymmetry in the Trolley example. Because actors are motivated to portray themselves in a positive light and flipping the switch results in the awful killing of a person, they are less likely to judge it permissible for them to flip the switch. However, because they are relatively less interested in portraying others in a positive light, they judge that it is permissible for others to flip the switch. But actors who realize that flipping the switch is the optimal decision even if it kills a person may excuse themselves by judging they had no control over the situation.

Given that there are actor-observer differences in a wide variety of contexts, we thought that similar actor-observer differences would be found in judgments about intentions and intentionality. In our first experiment, we used a new method in the study of folk intuitions about intentional action borrowed from experimental economics. We had participants engage in a real decision making process with real rewards and penalties. Because participants *actually* became actors, we hypothesized this methodology would have the greatest chance of revealing actor-observer differences in intentional action intuitions.

Experiment 1

We constructed a decision making environment where participants could (a) engage in helpful and harmful behaviors and (b) observe others' helpful and harmful behaviors. We call *Actors* those who generate a behavior. We call *Observers* those who watch a behavior. In the *Harm* condition, an actor generates a harm to one other person. In the *Help* condition, an actor generates a benefit to one other person. We hypothesized that actors would judge behaviors as (a) less intended and (b) less intentional than when they judge behaviors as observers.

Participants

Participants (N = 40) were recruited via email at a small southern university.[5] Participants were tested in 6 groups consisting of no more than 12 participants and no fewer than 4. Participants received $10 for attending. They also had the opportunity to earn an additional $10 depending on their performance in the experiment (Range = $16-$20). Participants were told that they would be paid as a function how many Experimental Currency Units (ECUs) they earned in the experiment. The payoff function was not disclosed.

[3]These scenarios modified cases used by Petrinovich and O'Neil (1996), but Trolley Problem cases are well known in the literature.

[4] Ninety percent in the 'John' version thought it was permissible versus 65% in the 'you' version. Also, the mean control rating in the 'John' version was 4.28 and 5.12 for the 'you' version (on a 7 point ascending scale).

[5]The expense of the experiment necessitated a small sample size.

Each participant was an actor and an observer (counterbalanced for order). However, each participant was in only one of the Help or Harm conditions. Because we were interested in intuitions about actions, all participants who did not perform the desired action (contributing to Account A, see below) were excluded. Five participants were thereby excluded in Harm. For the purposes of analyzes, there were 20 participants in Help and 15 in Harm.

Methods and Materials

Participants completed the experiment on a computer programed using Z-Tree software (Fischbacher, 2007). Actors in the Harm condition were instructed to indicate how many of their 10 "tokens" they wished to invest in 'Account A'. They were told that for every token they invested in Account A, they would earn 12 ECUs. For every token they did not invest in Account A, they would earn 10 ECUs. However, for every token invested in Account A, they would generate a 3 ECU penalty to one other person in the experiment. Actors in Help were given the same instructions as Actors in Harm but instead of generating a 3 ECU penalty, the actor generated a 3 ECU bonus by contributing to Account A. Observers in Harm read a display indicating that somebody else had contributed 10 tokens to Account A generating a 30 ECU penalty to them. Observers in Help read a display stating that another participant contributed 10 tokens to Account A generating a 30 ECU bonus for them. There was one unpaid practice round followed by one paying round in each condition.

After each instance of acting or observing, participants were asked to rate on a 7 point scale (1=disagree, 7=agree) their level of agreement with the appropriate version of each of the following sentences: 1. You/the other participant intended to generate the penalty/bonus; 2. You/the other participant intentionally generated the penalty/bonus; 3. You/the other participant are/is blameworthy/praiseworthy for generating the penalty/bonus. Participants were also given the opportunity to explain their answers in a few sentences. So, each participant answered 3 actor questions and 3 observer questions in only one of Harm or Help conditions and had the opportunity to explain their answers in each condition.

Results and Discussion

To test our hypothesis, a mixed-model Analysis of Variance (ANOVA) was preformed with Harm/Help and observer order as between participants variables and answers to the Actor Intended and Observer Intended prompts as within participants variables.

Table 1: Actor/Observer

Actor Intended	$M = 3.3$, $SD = 2.08$
Actor Intentionally	$M = 3.5$, $SD = 2.17$
Observer Intended	$M = 4.0$, $SD = 2.8$
Observer Intentionally	$M = 4.11$, $SD = 2.15$

The predicted difference in actor/observer judgments was found for intention judgments, F (1, 31) = 4.51, p =. 04, η_p^2 = .13. Neither order F (1, 31) = 1.12, p =. 29, η_p^2 = .04 nor condition interacted with judgments $F < 1$. A similar mixed-model ANOVA found the predicted differences in intentionality judgments, F (1, 31) = 4.14, p =. 05, η_p^2 = .12. Order did not interact with judgments, $F < 1$.

Theoretically, there should be differences in people's Harm and Help judgments (Knobe, 2003a) and a moderately sized non-significant trend toward an interaction for Harm/Help was observed, F (1, 31) = 2.61, p = .12, η_p^2 =.08. To help illuminate these possible differences, each condition (Harm or Help) was selected and four mixed-model ANOVAs were conducted with order as between participants factors and judgments about (1) Actor Intention/Observer Intention and (2) Actor Intentional/Observer Intentional as within participants factors.

In Harm, predicted differences were found for Intention judgments, F (1, 13) = 5.63, p= .03, η_p^2 = .3. Order did not interact with judgments F (1, 13) = 1.05, p = .33, η_p^2 = .07. Predicted differences were also found for Intentional judgments F (1, 13) = 9.15, p = .01, η_p^2 = .41. Order did not interact with judgments ($F < 1$). In Help, no actor-observer differences were detected (all F's < 1).

Table 2: Harm Actor/Observer

Actor Intended	$M = 2.13$, $SD = 1.13$
Actor Intentionally	$M = 2.5$, $SD = 1.96$
Observer Intended	$M = 3.2$, $SD = 2.01$
Observer Intentionally	$M = 3.73$, $SD = 2.25$

This experiment also allowed us to explore some other possibly interesting actor-observer differences. We thought that actors would display a reversed side effect-effect while observers would display the traditional side effect-effect. As side effects can occur when a behavior is judged intentional but not intended, we selected only those participants who did not judge the behavior in the relevant condition to be intended. After excluding those who did not intend the behavior (responding 4 or less), 14 participants remained in Harm and 10 remained in Help. A univariate ANOVA indicated the predicted shift in judgments in Harm that trended toward significance: Harm $M = 2.35$, $SD = 1.2$, Help M = 3.1, SD = 1.97, F (1, 23) = 2.46, p =. 13, η_p^2 = .11. However, order appeared to interact with judgments, F (1, 23) = 2.46, p = .13, η_p^2 = .11. To eliminate any possible order effect, only first responses were analyzed.[6] After eliminating those who were in the actor condition second, did not contribute to Account A, and responded that they intended the bonus or penalty, a very large marginally significant dif-

[6] Participants could not go back to the previous condition after they had entered their answers. Once participants gave their actor judgments, they could not go back and change them after they entered the observer condition.

ference was observed: Harm (N = 7, M = 2.43, SD = 1.9), Help (N = 3, M = 5.0, SD = 1.0), F (1, 8) = 4.68, p = .06, η_p^2 = .37.[7] However, we did not find the predicted side effect-effect for observers (all $F's$ < 1).

Finally, previous research indicates that some intentional action intuitions are predictable by the global personality trait extraversion (Cokely & Feltz, 2009a). Extraversion is a member of the Big Five personality model and is represented in almost all modern personality models (John, 1999). The current experiment allowed us to test for possible actor-observer differences in relation to extraversion. To this effect, participants also completed the Brief Big Five Inventory at the end of the experiment (Gosling, Rentfrow, & Swan, 2003). Extraversion was negatively correlated with judgments in Harm for Actor Intention, r (15) = -.64, p = .01 and Actor Intentionally, r (15) = -.55, p = .03 but was not correlated with Observer Judgments p < .05. To illustrate the difference, a rough median split of extraverts was created. Those who were relatively more introverted (scoring 9 or less) were more likely than extraverts (scoring higher than 9) to respond that they intended (Introverted M = 2.63, SD = 1.3, Extraverted M = 1.57, SD = .53, F (1, 14) = 3.92, p = .07, η_p^2 = .26) or intentionally (Introverted M = 3.5, SD = 2.2, Extraverted M = 1.29, SD = .49, F (1, 14) = 5.67, p = .04, η_p^2 = .34) brought about the harmful behavior. Order did not interact with judgments ($F's$ < 1). Of note, there was a strong overall correlation of intention and intentionality judgments: Other Intention/Other Intentional r (35) = .83, p < .001, Self Intention/Self Intentional, r (35) = .77, p < .001.

Experiment 2

Experiment 1 suggested that providing the right environment could engender an Actor-Observer difference in judgments about intentions and intentionality. However, a question remains whether Actor-Observer differences can occur in traditional pencil-and-paper surveys where participants are asked to imagine themselves in the role of the chairman. To address this possible worry that the effect found in Experiment 1 is not the result of the testing environment but rather is a more general phenomenon, Experiment 2 was designed to suggest that Actor-Observer differences are not likely to be found when participants are merely asked to imagine that they are the chairman.

Participants

One hundred and one participants were recruited from Amazon's Mechanical Turk to complete the survey for a small reward ($0.15). Participants were excluded if they reported that their first language was not English or if they failed the comprehension question. After excluding these participants, 95 remained.

[7] The small sample size and unequal cells are problematic. The small sample size in Help was anticipated because it is unlikely that good behaviors would be judged unintended by actors. See Feltz (2007a) and Nadelhoffer (2007) for a discussion.

Methods and Materials

Participants were redirected from Amazon's Mechanical Turk to complete the surveys at SurveryMonkey.com. There were four different scenarios: 1. Harm Observer, 2. Harm Actor, 3. Help Observer, and 4. Help Actor. The following were the Help and Harm cases in the Actor condition:

Actor Harm/Help: Imagine that you are the chairman of the board. The vice-president of a company comes to you and says, "We are thinking of starting a new program. It will help us increase profits for this year's balance sheet, but in ten years it will start to [harm/help] the environment." Imagine that you answered, "I don't care at all about [harming/helping] the environment. I just want to make as much profit for this year's balance sheet as I can. Let's start the new program." The program was started. Sure enough, ten years later, the environment started to be [harmed/helped].

Immediately following the scenario, participants were asked to rate their level of agreement with the following sentences (1 = strongly disagree, 4 = neutral, 7 = strongly agree):

1. You intended to [harm/help] the environment;
2. You intentionally [harmed/helped] the environment;
3. You are [blameworthy/praiseworthy] for harming/helping] the environment.

Participants were also asked the following comprehension question:

4. How long did it take before the [harm/help] began?

The following were Help and Harm in the Observer condition:

Observer Harm/Help: The vice-president of a company went to the chairman of the board and said, "We are thinking of starting a new program. It will help us increase profits for this year's balance sheet, but in ten years it will start to [harm\help] the environment." The chairman answered, "I don't care at all about [harming/helping] the environment. I just want to make as much profit for this year's balance sheet as I can. Let's start the new program." They started the new program. Sure enough, ten years later, the environment started to be [harmed/helped].

Immediately following the scenario, participants were asked to rate their level of agreement with the following sentences (1 = strongly disagree, 4 = neutral, 7 = strongly agree): 1. The chairman intended to [harm/help] the environment; 2. The chairman intentionally [harmed/helped] the environment; and 3. The chairman is [blameworthy/praiseworthy] for [harming/helping] the environment. Participants were also asked the following comprehension question: 4. How long did it take before the [harm/help] began?

Each participant was an Actor and an Observer in only one of the Harm (N = 46) or Help (N = 49) conditions. The order of presentation was counterbalanced. Once participants completed their responses to one condition, they could not go back and change their answers.

Results and Discussion

Univariate ANOVAs found a large Knobe-like effect for Actor judgments about intentions, F (1, 93) = 231.92, p < .

001, $\eta_p^2 = .71$. Order did not interact with judgments $p > .22$. A similar effect was found for Observer judgments about intentions, $F(1, 93) = 64.54$, $p < .001$, $\eta_p^2 = .41$. Order did not interact with judgments $F < 1$. A Knobe-like differences was also found for intentionality judgments for Actors, $F(1, 93) = 369.8$, $p < .001$, $\eta_p^2 = .80$ and Observers, $F(1, 93) = 122.29$, $p < .001$, $\eta_p^2 = 57$. Order did not reliably interact with judgments for Actors ($F < 1$) or Observers ($p > .09$).

A mixed-model repeated measures ANOVA with Actor-Observer judgments as within-participants variables and order as between participants factor did not reveal a reliable Actor-Observer in Harm, all *F's* < 1. However, a significant difference was found for judgments in Help Intention $F(1, 47) = 9.77$, $p = .003$, $\eta_p^2 = .17$ and Help Intentionally $F(1, 47) = 8.46$, $p = .006$, $\eta_p^2 = .15$.

Table 3: Means for Paper Survey

	Harm	Help
Actor Intend	*M* = 4.94 *SD* = 1.65	*M* = 1.14 *SD* = 0.54
Actor Intentionally	*M* = 5.83 *SD* = 1.36	*M* = 1.35 *SD* = 0.88
Observer Intend	*M* = 4.94 *SD* = 1.65	*M* = 1.96 *SD* = 1.94
Observer Intentionally	*M* = 5.83 *SD* = 1.24	*M* = 2.04 *SD* = 1.99

The results of Experiment 2 suggest that the Actor-Observer asymmetry produced in Experiment 1 is not likely to exist when participants are only encouraged to imagine they are the chairman.

General Discussion

Consistent with and extending previous research, our results suggest that in some circumstances people tended to judge their own behaviors differently than they judge the identical behavior of others. In addition, our evidence suggests that a well-known result in experimental philosophy—the traditional side effect-effect—can be reversed. Finally, replicating previous work (Cokely & Feltz, 2009a), extraversion was systemically and predictably related to some intention and intentionality judgments.

These results provide further evidence that impression management can play a key role in people's intention and intentionality judgments. An important clue for this interpretation comes from the results of the Harm case. Participants were much less likely to judge that they intended the Harm or intentionally brought it about compared to their judgments as observers. Presumably, participants did not want to be a "bad guy" by bringing about the bad side effect whereas they were relatively less interested in managing their impression of others. Hence, they were more motivated to respond that they did not intend or intentionally bring about the Harm. In addition, extraverts were much more likely to respond this way in Harm. Because extraverts are socially minded individuals, they would be relatively more concerned with possible social aspects of their behavior. However, because the behavior in Help is beneficial, there is less motivation to mitigate possibly negative implications of that behavior. So in Help, the responses between actors and observers would be similar.

These data also provide some important insights into the side effect-effect. We found strong correlations between people's intention and intentionality judgments. Those who favor the SV may take these as supporting data. However, defenders of the SV should be cautious for two reasons. First, correlation indicates that there is *some* relation between intention and intentionality judgments. These correlations *do not* necessarily indicate that an intention to *A* is a *necessary* condition for *A-ing* intentionally. These results are equally consistent with intending to *A* is a *sufficient* condition for *A-ing* intentionally when one *A's*—a condition that most theories of intentional action would endorse under normal conditions (e.g., no causal deviance). Second, we have some evidence that a new but equally problematic side effect-effect exists. For actors who did not think they intended to bring about the penalty or bonus, the moral valence of the consequence influenced their intentionality judgments. Specifically, participants were more likely to judge they brought about the beneficial consequence intentionally than the harmful consequence. These results suggest that at least some folk do not treat an intention to *A* as necessary for *A-ing* intentionally, contrary to the SV.

Third, our results reinforce the importance of individual differences in judgments about intentions and intentionality and provide more evidence that philosophically relevant intuitions are systematically fragmented (Feltz & Cokely, 2009; Cokely & Feltz, 2009b). Those who were extraverted were less likely to judge that they intended or intentionally brought about the penalty. Importantly, we were able to predict a priori who were likely to make those judgments. If there are predictable and systematic differences in intuitions regarding intentions and intentionality, then perhaps there is not a single folk concept of intentional action, but several (Cushman & Mele, 2008).

Finally, we would like to note one limitation of previous work in experimental philosophy that has relied on "pencil and paper" surveys. Rather than simply asking participants to respond to a scenario they read, we asked participants to perform an action and observe an action. We find that participants are less likely to think that a harm they actually bring about is intentional compared to a harm somebody else brings about to them. Hence, using this alternative method uncovered actor-observer differences in intuitions about intentions and intentionality, found an intriguing possible reversal of the side effect-effect, and provided additional evidence that folk intuitions about intentional action are predictably fragmented. We hope that the present experiments open up new methodological avenues for the experi-

mental investigation of folk intuitions about intentional action.

Acknowledgments

We would like to thank Al Mele, Joshua Knobe, Shaun Nichols, Mark Isaac, and participants at the NEH Summer Institute for Experimental Philosophy for very helpful comments.

References

Adams, F. (1986). Intention and intentional action: The Simple View. *Mind and Language, 1,* 281-301.

Cushman, F., & Mele, A. (2008). Intentional action: Two-and-a-half folk concepts? In J. Knobe & S. Nichols (Eds.), *Experimental Philosophy*. Oxford: Oxford University Press.

Cokely, E.T., & Feltz, A. (2009a). Individual differences, judgment biases, and Theory-of- Mind: Deconstructing the intentional action side effect asymmetry. *Journal of Research in Personality, 43,* 18-24.

Cokely, E. T., & Feltz, A. (2009b). Adaptive variation in folk judgment and philosophical intuition. *Consciousness and Cognition, 18*, 355-357.

Feltz, A. (2007a). Knowledge, moral praise, and moral side effects. *Journal of Theoretical and Philosophical Psychology, 27,* 123-126.

Feltz, A. (2007b). The Knobe effect: A brief overview. Journal of Mind and Behavior. *The Journal of Mind and Behavior, 28*, 265-277.

Feltz, A., & Cokely, E.T. (2008). The fragmented folk: More evidence of stable individual differences in moral judgments and folk intuitions. In B. C. Love, K. McRae & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Feltz, A., & Cokely, E.T. (2009). Do judgments about freedom and responsibility depend on who you are?: Personality differences intuitions about compatibilism and incompatibilism. *Consciousness and Cognition, 18,* 342-350.

Fernandez-Duque, D., & Wifall, T. (2007). Actor/observer asymmetry in risky decision making. *Judgment and Decision Making, 2,* 1-8.

Fischbacher, U. (2008). z-Tree: Zurich Toolbox for Ready-made Economic Experiments. *Experimental Economics, 10,* 171-178.

Gosling, S. D., Rentfrow, P. J., & Swann, W. B. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality, 37*, 504-528.

John, O. (1999). The Big-five Trait Taxonomy: History, Measurement, and Theoretical Perspectives. In L. Pervin & O. P. John (Eds.), *Handbook of Personality: Theory and Research*. New York: Guilford.

Jones, E., & Nisbett, R. (1972). The actor and the observer: Divergent perceptions of the causes of behavior. In E. Jones, D. Kanouse, H. Kelly, R. Nisbett, S. Vallins & B. Weiner (Eds.), *Attribution: Perceiving the Causes of Behavior*. Morristown, NJ: General Learning Press.

Knobe, J. (2003a). Intentional action and side-effects in ordinary language. *Analysis, 63*, 190-194.

Knobe, J. (2003b). Intentional action in folk psychology: An experimental investigation. *Philosophical Psychology, 16,* 309-323.

Knobe, J. (2004a). Folk psychology and folk morality: Response to critics. *Journal of Theoretical and Philosophical Psychology, 24,* 270-279.

Knobe, J. (2004b). Intention, intentional action and moral considerations. *Analysis, 64,* 181-187.

Knobe, J., & Burra, A. (2006). Experimental philosophy and folk concepts: Methodological considerations. *Journal of Cognition and Culture, 6,* 331-342.

Leslie, A., Knobe, J., & Cohen, A. (2006). Acting intentionally and the side-effect effect: Theory of mind and moral judgment. *Psychological Science, 17*, 421-427.

Malle, B., & Knobe, J. (1997). Which behaviors do people explain? A basic actor-observer asymmetry. *Journal of Personality and Social Psychology, 72,* 288-304.

Malle, B., Knobe, J., & Nelson, S. (2007). Actor-observer asymmetries in explanations of behavior: New answers to an old question. *Journal of Personality and Social Psychology, 93,* 491-514.

McCann, H. (1986). Rationality and the range of intention. *Midwest Studies in Philosophy, 10,* 191-211.

McCann, H. (2005). Intentional action and intending: Recent empirical studies. *Philosophical Psychology, 18,* 737-748.

Mele, A. (1992). Recent work on intentional action. *American Philosophical Quarterly, 29,* 199-217.

Mele, A. (2001). Acting intentionally: Probing folk intuitions. In B. F. Malle, L. J. Moses & D. A. Baldwin (Eds.), *Intentions and Intentionality*. Cambridge: MIT Press.

Nadelhoffer, T. (2007). Fringe benefits, side effects, and intentional actions: A reply to Feltz. *The Journal of Theoretical and Philosophical Psychology, 27,* 801-809.

Nadelhoffer, T. (2006). On trying to save the simple view. *Mind and Language, 21,* 565-586.

Nadelhoffer, T., & Feltz, A. (2008). The actor-observer bias and moral intuitions: Adding fuel to Sinnott-Armstrong's fire. *Neuroethics, 1,* 133-144.

Petrinovich, L., & O'Neill, P. (1996). Influence of wording and framing effects on moral intuitions. *Ethology and Sociobiology, 17,* 145-171.

Learning from Errors by Counterfactual Reasoning in a Unified Cognitive Architecture

Andreea Danielescu (lavinia.danielescu@asu.edu)
David J. Stracuzzi (david.stracuzzi@gmail.com)
Nan Li (nanan9177@gmail.com)
Pat Langley (langley@asu.edu)
Computing Science and Engineering
Arizona State University, Tempe, AZ 85287 USA

Abstract

A key characteristic of human cognition is the ability to learn from undesirable outcomes. This paper presents a computational account of learning from errors based on counterfactual reasoning, which we embed in ICARUS, a unified theory of the cognitive architecture. Our approach acquires new skills from single experiences that improve upon and mask those that initially produced the undesirable behavior. We demonstrate the operation of this model in a simulated urban driving environment. We also relate our approach to other research on error-driven learning and discuss possible improvements to the framework.

Keywords: cognitive architecture, learning from error, counterfactual reasoning, problem solving

Background and Motivation

The ability to acquire knowledge from experience is a fundamental component of human intelligence. There exist many accounts of learning from positive experiences, most often based on successful problem-solving attempts (Anzai & Simon, 1979; Laird, Rosenbloom, & Newell, 1986). In this paper, we focus instead on learning from undesirable outcomes, an ability that plays an important role in human cognition by providing a mechanism for avoiding past failures in the future. We provide a computational model for one type of error-driven learning that uses counterfactual reasoning to determine both the error's cause and the correct behavior.

Counterfactual reasoning is a strategy that considers what might have occurred if causal events were changed in some way. Psychological studies suggest that people employ counterfactual reasoning in a variety of situations (Roese, Hur, & Pennington, 1999). Byrne and McEleney (2000) also show that they tend to employ counterfactual reasoning mainly in response to negative outcomes, such as failure to achieve or maintain goals. Finally, Epstude and Roese (2008) make the connection to learning, based on their theory that the primary motivation for counterfactual reasoning is to improve future performance.

The work described here offers a computational account of the role of counterfactual reasoning in learning from failures. We embed this account within ICARUS (Langley & Choi, 2006), a unified theory of the human cognitive architecture that makes a commitment to hierarchical, composable knowledge structures. We claim that these structures, along with the mechanisms for using and acquiring them, provide ICARUS with basic support for benefiting from undesirable outcomes. Our approach to learning from errors responds to a single negative experience, which distinguishes it from connectionist, reinforcement-based, and Bayesian techniques, which typically require many experiences.

We begin our discussion with a motivating task domain and a review of the ICARUS architecture. After this, we present our approach to learning from errors via counterfactual reasoning, including methods for determining the source of the error, acquiring new concepts and skills in response, and utilizing these structures in future behavior. We then describe the extended architecture's operation in the task domain, discuss connections to other work on error-driven learning, and consider directions for further research in this important area.

An Illustrative Domain: Urban Driving

In modern society, the task of operating a vehicle in an urban setting is both common and cognitively challenging. People perform a variety of tasks in this context, such as navigation, obstacle avoidance, and signal response, along with higher-level tasks such as package delivery. Successful performance relies on substantial domain expertise, making urban driving a useful domain in which to study embedded cognition and learning.

For this reason, we have developed a three-dimensional urban driving environment based on the Torque Game Engine produced by Garage Games.[1] The driving simulator provides the driver with control over the gas pedal, brake and steering, with objects obeying realistic laws of physics. The simulator also generates detailed perceptual information, in egocentric polar coordinates, about nearby entities, including road segments, intersections, lane lines, buildings, pedestrians, and other vehicles.

The driving task we examine here requires the agent to overtake a stalled vehicle, which it decides to passs on the left. However, in taking this step, the agent crosses a double yellow line, thereby violating the rules of driving and risking collision with oncoming traffic. The problem is not that the agent lacks knowledge about this constraint; the error occurs because it was focusing on a different goal that interacts with the one it violates. We

[1] http://www.garagegames.com

maintain that learning to avoid such errors relies on a form of counterfactual reasoning, which we describe in detail later. Our analysis is not limited to urban driving, nor do we intend it to model how humans learn to operate a vehicle, but it provides a useful setting to illustrate our ideas. However, we must embed our account within some theoretical framework, to which we now turn.

A Review of the ICARUS Architecture

We have explored these issues within ICARUS, a theory of the cognitive architecture (Newell, 1990) that makes commitments to the representations, performance mechanisms, and learning processes that underlie intelligence. Like Soar (Laird et al., 1986) and ACT-R (Anderson, 1993), ICARUS encodes content as symbolic list structures, matches long-term structures against short-term ones in a recognize-act cycle, combines goal-driven with data-driven processing, and interleaves an incremental form of learning with performance. Key differences include separate memories for concepts and skills, the hierarchical organization of knowledge, and cascaded integration in which problem solving builds on skill execution, which in turn relies on conceptual inference. In this section, we review ICARUS' structures and processes, drawing brief examples from urban driving. Langley and Choi (2006) describe the framework in more detail.

The most basic process in ICARUS is inference, which matches the agent's perceptions against long-term conceptual structures to produce beliefs about the environment. On each cycle, the environment deposits descriptions of perceived objects into the *perceptual buffer*, with each percept giving the object type, a unique identifier, and (typically numeric) attribute-value pairs. ICARUS links these perceptions to long-term structures in its *conceptual memory*. Each concept describes a class of situations in logical form that includes the concept's name, its arguments, and the conditions under which the concept applies. Conditions may refer to percepts or to simpler concepts, thus creating a hierarchical organization on memory. For instance, the conceptual structure

```
((aligned-in-lane ?agent ?line1 ?line2)
 :percepts  ((self ?agent) (lane-line ?line1))
            (lane-line ?line2 angle ?angle)
 :relations ((steering-straight ?agent)
            (in-lane ?agent ?line1 ?line2))
 :tests     ((≥ ?angle -3) (≤ ?angle 3)))
```

states when one should infer that `?agent` is driving parallel to the lane lines on either side, with the conditions referring to percepts, numeric tests among perceived attributes, and other conceptual relations.

On each cognitive cycle, ICARUS matches these concepts against elements in its perceptual buffer to produce inferences that it deposits in a *belief memory*, which in turn produce higher-level inferences. These typically describe relations among objects in the environment, with each belief being an instance of some defined concept. For example, the belief `(aligned-in-lane me line23 line24)`, that the agent is aligned in a lane bounded by `line23` and `line24`, is an instance of the `aligned-in-lane` concept above. A recent extension of ICARUS includes time stamps on beliefs to indicate when the architecture inferred them and when they became false, with the symbol `now` indicating that a belief holds on the current cycle. These time-annotated beliefs serve as a simple episodic trace to which we will return later.

ICARUS includes a separate long-term memory for skills. These are similar in form to concepts but specify methods for achieving goals rather than conditions for recognizing their achievement. Each skill includes a generalized goal (which must refer to a defined concept), a set of perceptual and conceptual conditions that must hold for the skill to match, and ordered subgoals that, once satisfied, should achieve the parent goal. For instance, the skill clause

```
((driving-well-in-lane ?agent ?line1 ?line2)
 :percepts  ((self ?agent) (lane-line ?line1)
            (lane-line ?line2))
 :start     ((in-lane ?agent ?line1 ?line2))
 :subgoals  ((at-speed ?agent)
            (centered-in-lane ?agent ?line1 ?line2)
            (aligned-in-lane ?agent ?line1 ?line2)))
```

specifies three subgoals the agent should achieve, once it is in a lane, to be driving well in that lane. Primitive skills have the same structure but replace subgoals with a set of executable actions the agent should carry out.

The ICARUS execution process uses the beliefs produced during conceptual inference to determine which skills to select. This process begins by choosing an unsatisfied goal – a concept instance the agent wants to be true – stored in a separate *goal memory*. The architecture retrieves all skills indexed by this goal, then attempts to find an applicable path downward through the skill hierarchy. A skill path is applicable if, for each skill on the path, the associated goal is not satisfied and the associated conditions are met. Such a path must terminate in a primitive skill with executable actions that affect the environment. On each cycle, ICARUS selects the first such path through the skill hierarchy that it finds, incorporating a preference for continuing an activity it has initiated over starting new ones.

When ICARUS can find no applicable skill in memory to achieve its current goal, it resorts to a form of means-ends problem solving (Newell & Simon, 1961; Carbonell, Knoblock, & Minton, 1990) that attempts to dynamically compose known skills, which it executes as they become applicable. The process begins when the architecture cannot find an applicable path through the skill hierarchy, in which case it attempts to retrieve a skill that would achieve the goal, then creates subgoals for its unsatisfied preconditions. If ICARUS cannot find such

a skill, it instead retrieves the definition for the goal concept and creates subgoals for each of its unsatisfied conditions. This process continues recursively, chaining backward off subgoals, until it retrieves a relevant applicable skill, which it then executes in the environment. Upon achieving the current subgoal, the system turns its attention to others, continuing this activity until achieving the top-level goal that initiated problem solving.

This mechanism lets ICARUS overcome situations in which it lacks skills to achieve its goals, but it often requires substantial search. However, the architecture also includes a learning process that generalizes and stores solutions for future use in similar situations. Briefly, this creates a skill whenever problem solving achieves a goal or subgoal, with the new structure being indexed by that goal, including subgoals that it satisfied along the way, and having conditions that were present when it began working on the problem. The details differ depending on whether the problem solver chained off skills or concept definitions, but the results are similar for both cases. ICARUS can then apply the new skills in the same manner as the other, older skills during subsequent execution.

Learning from Undesirable Outcomes

We have used ICARUS to develop cognitive models in a number of complex domains, including urban driving (Langley & Choi, 2006) and American football (Li, Stracuzzi, Cleveland, et al., 2009). Nevertheless, the architecture lacks some important functionalities, including the ability to learn from undesirable outcomes. The work we report here has started to address this limitation by adding a mechanism for learning from errors. More specifically, we consider the case in which the agent manages to achieve a given goal, but in which this causes it to inadvertently violate another, higher-priority, goal, which it must then attempt to repair.

As we discuss in detail below, the extended ICARUS responds to such situations in three steps. First, it determines which goals conflicted and constructs a new concept that it uses to encode the combined goal. Next, the architecture employs counterfactual reasoning to identify the primitive skill that produced the error and to determine another sequence that achieves this goal. Note that the skill which led to the conflict may have been executed many cycles prior to the actual goal violation. Finally, ICARUS learns new skills that mask the original structures and achieve the joint goal in similar situations. These mechanisms are not completely new, in that they build upon many existing ICARUS processes, making them more elaborations of the architectural framework than separate modules.

Combining Interacting Goals

The first step toward learning from errors is to determine which goals conflicted to cause the failure. A conflict here refers to a case in which the system violates a previously satisfied, higher-priority goal in the course of pursuing its current goal. For example, suppose a driving agent (`me`) has two goals, `(on-right-side me)` and `(at-speed me)`. The agent first maneuvers the vehicle onto the right side of the road and then begins accelerating toward its desired speed. If another, slower-moving vehicle then enters the road in front of the agent, it may execute a skill for passing on the left. This violates the agent's first goal, and causes it to abandon the second goal in an effort to restore the first.

If the agent prefers to pass on the left when approaching a slow vehicle, it will never satisfy both goals in this situation. To address this stalemate, the architecture constructs a new goal by conjoining the two concepts that supported the original goals, then replacing the goals with one based on the new concept. Returning to the driving example above, ICARUS first creates a concept `(on-right-side-and-at-speed ?agent)` with relations `(on-right-side ?agent)` and `(at-speed ?agent)`. Note that the new concept extends the current hierarchy by building on existing concepts. ICARUS then replaces the two original goals with `(on-right-side-and-at-speed me)`, giving the agent a new goal for which it can learn more specific skills.

Assigning Blame and Finding Alternatives

After identifying which goals conflicted and creating a revised top-level goal, ICARUS attempts to understand the reasons for its failure and how it might have been avoided. To this end, it attempts to identify the most recently executed primitive skill that, if replaced by a better choice, would avoid violating the high-priority goal and achieve the newly constructed one. This constitutes a form of counterfactual reasoning. The system begins by considering each primitive skill executed in the episode in reverse chronological order. For a selected skill, it rolls back episodic belief memory to the cycle on which it first selected and executed the skill.[2] ICARUS does not consider non-primitive skills, as they could lead it to backtrack farther than necessary.

After rolling back the episodic trace, ICARUS invokes problem solving with the new, conjoined goal created earlier. Recall that the architecture normally interleaves problem solving with execution, but here the agent can only suppose what might have happened if it had taken another path. For this reason, we introduced a more traditional version of problem solving that uses mental simulation based on skills' effects. Each time the problem solver selects a skill for imaginary execution, it updates belief memory with the expected changes, then triggers inference, which updates belief memory as though it had received new percepts.

[2] Primitive skills may be durative, which means they may require several cycles to achieve their intended goals. Backtracking over such a skill may therefore involve jumping backward in time by several cognitive cycles.

If the problem solver fails to achieve the combined goal, then the system rolls back another step, past the preceding primitive skill in the original episodic trace. The process outlined above then begins again with a new round of problem solving. Note that this approach to error localization and counterfactual reasoning is well integrated with existing facets of the architecture, and it depends critically on results produced by the modules for inference and problem solving. This provides further evidence that ICARUS offers a unified theory of cognition.

Learning and Selecting New Skills

Having determined a sequence of primitive skills that would have achieved both of its original goals, the extended architecture must generalize these results and store them in skill memory for use in guiding future behavior. We want the acquired skills to apply in more constrained situations than those which produced the undesirable outcome, so they should mask the old skills to prevent their selection in these cases. For our driving example, the system should select the new skill only if it wants to achieve *both* `(on-right-side ?agent)` and `(at-speed ?agent)`.

To learn skills from the results of counterfactual reasoning, ICARUS uses the same mechanism as during normal problem solving. As the agent works toward its top-level goal, it acquires a new skill as means-ends analysis achieves each subgoal. The use of counterfactual reasoning and mental simulation, rather than execution in the environment, has no effect on this process. However, to make effective use of this learned knowledge, we must modify ICARUS' execution module. The standard mechanism selects the unsatisfied goal with highest priority, then finds a path through the skill hierarchy that should achieve it. This scheme works well when there are no goal interactions, but, as we have seen, it can lead to problems when they exist.

In response, we modified the execution module to prefer skills that, other things being equal, would let the agent achieve multiple goals. For example, it selects a skill that addresses two goals with first and second priority over one that would achieve only the first goal. However, priority still plays a key role; the system prefers a skill that tackles a first priority goal over one that would achieve goals with second and third priority. This approach lets more specific skills acquired through counterfactual reasoning mask the original, more general skills that caused the undesired behavior, while letting the older skills remain available for situations in which only they apply.

Demonstration on Urban Driving

We tested this extension to ICARUS in the urban driving domain described earlier. Here we consider a single run at length to clarify the architecture's operation. Our aim is not to match human behavior in detail, but to show that the new system exhibits an important capacity of human cognition that its predecessor lacked. In this run, we placed the ICARUS driving agent[3] in the leftmost lane on the right side of the road heading east. Another vehicle in the same lane was stalled in the road ahead. The agent's initial goals were `(on-right-side me)` and `(avoid-obstacle me)`. As the agent drives, it realizes that it is approaching the car ahead too rapidly and avoids colliding with it by swerving left. The agent swerves left rather than right simply because the associated skill happens to prefer that option, but this causes it to violate the high-priority goal, `(on-right-side me)`, by crossing to the left side.

At this point the agent realizes that it had ignored this high-priority goal while focusing on another one. Drawing its counterfactual reasoning abilities, the system creates a concept,

```
((on-right-side-and-avoid-obstacles ?self)
 :percepts ((self ?self))
 :relations ((on-right-side ?self)(avoid-obstacles ?self)))
```

that can serve as a new conjoined goal to direct its analysis. Next, the reasoning system backtracks through the episodic trace to the previously executed primitive skill, `(throttle-special-value me)`. Using time stamps on beliefs to reconstruct its mental state at that point, it invokes problem solving and mental simulation to search for another sequence of skills that achieves the goal `(avoid-obstacle me)` without violating the other one. In this case, the problem solver cannot find a solution that begins with this state, so the reasoner continues to backtrack through the earlier skill, `(crossing-into-left-lane me)`, which also fails to solve the problem. Eventually, after returning mentally to the state before `(wheels-straight me)`, problem solving finds a sequence of primitive skill instances,

```
(crossing-into-right-lane2 me), (wheels-straight me),
(on-right-side-lane2 me), (lane-aligned me),
(wheels-straight me)
```

that, if executed, would have achieved the conjoined goal `(on-right-side-and-avoid-obstacles me)`. Analysis of this solution using the adapted means-end problem solver leads to creation of a single new skill

```
((on-right-side-and-avoid-obstacles ?self)
 :percepts  ((self ?self))
 :start     ((on-right-side-lane1 ?self)
             (drone-ahead ?self ?drone ?dist ?angle))
 :subgoals  ((avoid-obstacles-by-right ?self)))
```

which is indexed by the new conjoined concept that, if satisfied, ensures that its component concepts are met.

On a subsequent run after learning with the same initial situation, the agent makes a different choice when

[3]The ICARUS agent for this task included 23 skill clauses and 48 conceptual clauses, both organized hierarchically.

it approaches the stalled vehicle, swerving into the right lane rather than crossing over to the left side. The reason is that the architecture prefers skills that are indexed by more specific goals, thus masking the original preference for veering left over right. The result is that the agent still avoids hitting the stalled car ahead of it, satisfying the goal `(avoid-obstacle me)`, without violating the even higher-priority goal, `(on-right-side me)`.

Discussion

Counterfactual reasoning has been implicated in humans as a mechanism for establishing the cause of particular events (Roese, 1997), for identifying errors of both omission and commission (Byrne & McEleney, 2000), and for learning from errors (Roese & Olson, 1995; Wells & Gavanski, 1989). Our work with ICARUS has focused on using counterfactuals to establish the cause of negative events (violations of maintenance goals) and to replace incorrect actions with proper ones. Although we have not yet shown that it can recover from errors of omission, we believe that the same mechanisms will support such learning.[4] Our account of counterfactual reasoning makes clear contact with psychological literature on the topic and, although our model makes some implausible assumptions (e.g., about memory), its main features are consistent with key theories and empirical findings.

Few computational models have made use of counterfactuals in the context of learning. One example, Mueller and Dyer's (1985) DAYDREAMER, uses them more broadly than ICARUS but in a less directed manner. The system learns from both positive and negative experiences by postulating alternative actions and considering their consequences, but it proposes scenarios based on control goals, episodic memory contents, and emotional state. This strategy can produce a variety of outcomes, some substantially removed from reality, while ICARUS pursues a single goal until achieving it. Pearson's (1996) IMPROV also makes use of counterfactuals to improve procedural knowledge. Like ICARUS, it considers alternative action sequences starting from the last state before the error occurred, then working backward until it finds a solution. However, IMPROV focuses on revising skills that fail to achieve intended goals, while ICARUS specializes skills that violate other goals it achieved previously. In addition, IMPROV revises its knowledge by modifying skill preconditions, rather than learning new skills that achieve more specific goals.

Other research on learning from errors has also focused on detecting and resolving errors, most on ones that stem from overly general rules. For example, early versions of the SWALE system (Schank, 1986) adapt explanations to unanticipated situations when its expectations are violated. Similarly, Ohlsson (1996) shows how to correct errors that stem from overly general rules by comparing the actual and intended outcomes of selected actions, while Holland et al. (1986) describe a mechanism for specializing rules using counterexamples. Langley (1987) also reports an approach which compares similar situations that produce positive and negative outcomes to improve upon overly general rules. Work on analytical learning typically focuses on learning from success, but a few efforts (Carbonell et al., 1990; Laird et al., 1986) address learning from failure. These share our concern with explaining reasons for errors, but they produce control rules that specify what to avoid, while our approach instead acquires skills that mask the undesired behavior.

In addition to testing the architectural extensions in other domains that involve goal interactions, we should also improve our account of counterfactual reasoning along other dimensions. One involves increasing its psychological plausibility by placing realistic limits on the contents of ICARUS' perceptual buffer and its episodic memory, which currently contain far more than their human analogs. We should also expand the generality of our counterfactual reasoning framework to learn from other types of errors, such as Ohlsson's constraint violations. In addition, we should extend the architecture's representation and its inferential abilities to let it reason about the goals and beliefs of other agents, since many of the most interesting errors that humans exhibit, and from which they are driven to learn, occur during their interpersonal interactions.

The main contribution of our work has been a computational account of skill learning through counterfactual reasoning. This involves three major steps: detecting that pursuit of one goal has violated another, reasoning backwards from the conflict to identify the choice that caused it and finding an alternative path that would have avoided it, and storing a specialized skill that produces the desired actions and masks the original behavior. Although we have embedded our account within ICARUS, one could also incorporate it into other architectures, although some details would certainly differ. And although we have illustrated these mechanisms in the context of urban driving, they seem relevant to any domain in which goal conflicts can arise. We will not claim that our account covers all forms of learning through counterfactual reasoning, which may also support revision of incorrect concepts, skills, and beliefs, but we believe it advances our understanding of this complex ability, and thus our grasp of human cognition.

Concluding Remarks

In this paper, we presented a set of interacting computational mechanisms that support learning from undesirable outcomes via counterfactual reasoning. We embedded this account within ICARUS, a theory of the cognitive architecture that placed strong constraints on our

[4] Ginsberg (1986) discusses the use of counterfactual reasoning in identifying subgoals during problem solving, which we have not addressed here.

approach to the problem. After reviewing the structures and processes that ICARUS assumes, we presented new mechanisms that identify the violation of previously satisfied goals, localize the cause of this event by inspecting an episodic trace, invoke problem solving to find alternative steps that would have avoided the error, and learn specialized skills from this analysis that generate the desired behavior in the future.

We demonstrated these interacting mechanisms in the context of a simulated urban driving environment, showing that they behave as intended in a complex scenario that requires multi-step reasoning. We also considered earlier work on error-driven learning that bears similarities to our own, but that has addressed different issues, and promising directions for extending our approach. It seems clear that counterfactual reasoning plays an important role in human learning and, although our current model explains only certain forms of such cognitive behavior, it nevertheless provides a novel account of the mechanisms that underlie support this complex ability.

Acknowledgements

This material is based on research sponsored by ONR under agreement N00014-09-1-0123. The U. S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are the authors' and should not be interpreted as representing the official policies or endorsements, either expressed on implied, of ONR or the U. S. Government.

References

Anderson, J. R. (1983). *The architecture of cognition*. Cambridge, MA: Harvard University Press.

Anderson, J. R. (1993). *Rules of the mind*. Hillsdale, NJ: Lawrence Erlbaum.

Anzai, Y., & Simon, H. A. (1979). The theory of learning by doing. *Psychological Review*, *86*, 124–140.

Byrne, R., & McEleney, A. (2000). Counterfactual thinking about actions and failures to act. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 1318–1331.

Carbonell, J., Knoblock, C., & Minton, S. (1990). PRODIGY: An integrated architecture for planning and learning. In K. VanLehn (Ed.), *Architectures for intelligence*. Hillsdale, NJ: Lawrence Erlbaum.

Epstude, K., & Roese, N. J. (2008). The functional theory of counterfactual thinking. *Personality and Social Psychology Review*, *12*, 168–192.

Ginsberg, M. (1986). Counterfactuals. *Artificial Intelligence*, *30*, 35–79.

Holland, J., Holyoak, K., Nisbett, R., & Thagard, P. (1986). *Induction: The process of inference, learning, and discovery*. Cambridge, MA: MIT Press.

Laird, J. E., Rosenbloom, P. S., & Newell, A. (1986). Chunking in Soar: The anatomy of a general learning mechanism. *Machine Learning*, *1*, 11–46.

Langley, P. (1987). A general theory of discrimination learning. In D. Klahr, P. Langley, & R. Neches (Eds.), *Production system models of learning and development* (pp. 99–161). Cambridge, MA: MIT Press.

Langley, P., & Choi, D. (2006). A unified cognitive architecture for physical agents. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence* (pp. 1469–1474). Boston: AAAI Press.

Li, N., Stracuzzi, D. J., Cleveland, G., Könik, T., Shapiro, D., Molineaux, M., et al. (2009). Constructing game agents from video of human behavior. In *Proceedings of the Fifth AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment* (pp. 64–69). Stanford, CA: AAAI Press.

Mueller, E. T., & Dyer, M. G. (1985). Towards a computational theory of human daydreaming. In *Proceedings of the Seventh Annual Conference of the Cognitive Science Society* (pp. 120–129). Irvine, CA.

Newell, A. (1990). *Unified theories of cognition*. Cambridge, MA: Harvard University Press.

Newell, A., & Simon, H. A. (1961). GPS: A program that simulates human thought. In H. Billing (Ed.), *Lernende automaten*. Munich: Oldenbourg KG. (Reprinted in E. A. Feigenbaum & J. Feldman (Eds.), *Computers and thought*. New York: McGraw-Hill.)

Ohlsson, S. (1996). Learning from performance errors. *Psychological Review*, *103*, 241–262.

Pearson, D. J. (1996). *Learning procedural planning knowledge in complex environments*. Unpublished doctoral dissertation, Computer Science and Engineering Division, University of Michigan, Ann Arbor, MI.

Roese, N. (1997). Counterfactual thinking. *Psychological Bulletin*, *121*, 133–148.

Roese, N., Hur, T., & Pennington, G. (1999). Counterfactual thinking and regulatory focus: Implications for action versus inaction and sufficiency versus necessity. *Journal of Personality and Social Psychology*, *77*, 1109–1120.

Roese, N., & Olson, J. (1995). *What might have been: The social psychology of counterfactual thinking*. Hillsdale, NJ: Erlbaum.

Schank, R. (1986). *Explanation patterns: Understanding mechanically and creatively*. Hillsdale, NJ: Erlbaum.

Stracuzzi, D. J., Li, N., Cleveland, G., & Langley, P. (2009). Representing and reasoning over time in a symbolic cognitive architecture. In *Proceedings of the 31st Annual Meeting of the Cognitive Science Society* (pp. 2986–2991). Amsterdam.

Wells, G., & Gavanski, I. (1989). Mental simulation of causality. *Journal of Personality and Social Psychology*, *55*, 161–169.

Optimal Inference and Feedback for Representational Change

Yun Tang (tang.162@osu.edu)
Department of Psychology, The Ohio State University,
1835 Neil Ave, Columbus, OH 43210 USA

Christopher J. Young (young.1202@osu.edu)

Jay I. Myung (myung.1@osu.edu)

Mark A. Pitt (pitt.2@osu.edu)

John E. Opfer (opfer.7@osu.edu)

Abstract

Knowledge representations are central to many cognitive processes, and how these representations change is a central issue in learning and cognitive development. Here we developed and implemented a Bayesian inferential procedure to detect and elucidate representational change in numerical estimation. The proposed procedure of an adaptive numerical experiment both infers a learner's representation and predicts the feedback that is likely to induce representational change. We provide an application of this procedure using simulated subjects and demonstrate its effectiveness in inferring representational state and inducing change.

Keywords: representational shift; numerical estimation; adaptive experiment; Bayesian inference.

Introduction

Knowledge representations play a large role in cognitive processes such as learning, memory, and problem-solving (Markman, 1999), and a central problem in cognitive development concerns how representations change with age and experience (Carey, 1985; Dixon & Bangert, 2002; Siegler & Opfer, 2003). A striking example of representational change occurs in developing numerical magnitude representations. These representational changes are apparent across a wide range of tasks where numbers are quantified along a range, whether by categorizing numbers by magnitude (Opfer & Thompson, 2008), estimating numerosity (Booth & Siegler, 2006), measurements (Booth & Siegler, 2006), or positions of numbers on number lines (Dehaene, Izard, Spelke, Pica 2008; Siegler & Opfer, 2003).

Studies on development of numerical representations typically find that young children initially estimate numerical magnitudes to increase logarithmically with actual value before later learning the decimal system (Siegler & Opfer 2003, Booth & Siegler 2004; Opfer & Thompson 2007). This change is interesting theoretically because the logarithmic representation is implicit in speeded magnitude comparisons (Moyer & Landauer, 1967) and generation of random numbers (Banks & Hill, 1974) despite explicit judgments of numerical magnitude. This shift is also widespread across cultures, occurring relatively early in cultures that emphasize children's mathematical education (Siegler & Mu, 2008) and delayed in cultures that lack formal schooling (Dehaene, Izard, Spelke, & Pica, 2008). Recent evidence also suggests that this representational shift can be induced in situ by providing examples (Izard & Dehaene 2007; Opfer & Siegler, 2007). That is, feedback on a few key numbers that are highly discrepant between logarithmic and linear functions causes rapid and broad adoption of linear representations (Opfer & Siegler, 2007).

Ideally feedback should take into account a child's current and target representational states. To do so, one must first infer, from a few noisy examples, the model that best describes the child's perception of numerical magnitude. This inference may be viewed as a model selection problem in which candidate models are evaluated and compared for their ability to capture the regularities underlying the data (Pitt & Myung, 2002). With the underlying representation having been inferred, one is now in a position to determine feedback that is most likely to induce representational changes in learners. This latter perspective proposes hypotheses for the ideal training regimen; feedback given to a child will be the most effective when it maximally discriminates between a logarithmic and linear representation while tracking the learner's current representation. These ideas can be formalized in a statistical framework, which is described in detail in a later section. This formal approach should have benefits to the theoretical questions that motivate research on the shift in numerical estimation, i.e. what is the path and source of change in numerical estimation abilities? We will be able to measure more precisely what about a child's representation changes to and what types of feedback are most likely to elicit it. The fruits of this approach could lead to the introduction of more effective teaching and training regimens.

In the present paper we propose a procedure that both (1) adaptively infers a learner's most likely representation and (2) predicts the feedback that will most likely induce representational shifts through what we call a cognitive tutor. We will demonstrate how this procedure is performed using computer simulations with information drawn from previous experimental data. We will also show the advantages of this procedure over traditional training studies in efficiency and the likelihood of inducing change.

Given our present focus on simulations of the above procedure, the purpose of this simulation study is three-fold. First, before implementation in experimental settings, it is necessary to run simulations to check the performance and accuracy of the method. Second, simulations could demonstrate the advantages of the cognitive tutor over the traditional paradigm. Finally, we are able to generate hypothesis for later experiments from simulation results. We use the topic of numerical estimation as a running example, and then discuss the potential to transfer the technology to other domains.

Adaptive Numerical Experiment

For representational shift problems, specifically in the domain of numerical representation, we propose an adaptive numerical experiment which infers the representation and performs the role of cognitive tutor. The procedure takes a perspective of model selection and distinguishes between the following models:

$$y_i = ax_i + b + e_i \ (i = 1,\ldots,n) \quad (1)$$

$$y_i = a\log x_i + b + e_i \ (i = 1,\ldots,n) \quad (2)$$

where x denotes the presented stimuli, y denotes the perceived numerical magnitude, and e is a normally distributed error with mean 0 and standard deviation σ.

In the experiment, we follow the paradigm used in Opfer and Siegler (2007), which shows the importance of choosing feedback. The same Number-Line Task is used as the numerical estimation task in our experiment. In each experiment trial, the child is shown a number between 0-100 or 0-1000 and is asked to estimate its position on a line.

The experiment is split into three sessions, as illustrated in Figure 1, and mirrors previous number line studies. In the pre-test session (Session 1), the Number-Line Task is performed to infer the child's existing representation model; each trial children are shown a number and asked to estimate its corresponding position on a line. Next in the feedback session (Session 2), children respond as in pre-test, but after each response are shown the (correct) linear position of each number. The post-test session (Session 3) is similar to the pre-test session, which examines whether any shift occurred in the child's representation model, with no feedback provided.

The proposed adaptive numerical experiment applies the Adaptive Design Optimization (ADO) method and reorganizes the three sessions into two processes, the adaptive inference and the adaptive tutoring. In what follows we define the two processes and describe how ADO works and how it is incorporated into the processes.

Adaptive Inference Process

The adaptive inference process (AIP) takes place in the pre-test session and infers a child's most likely representation model (e.g. linear). It conducts a series of experiment trials and presents the numerical stimuli sequentially. Within each trial, the observed response is analyzed and the next stimulus is provided based on the analysis. It is adaptive in that it tailors the test procedure to individual state from trial to trial. Consequently, it obtains sufficient evidence to make inference within the fewest possible trials.

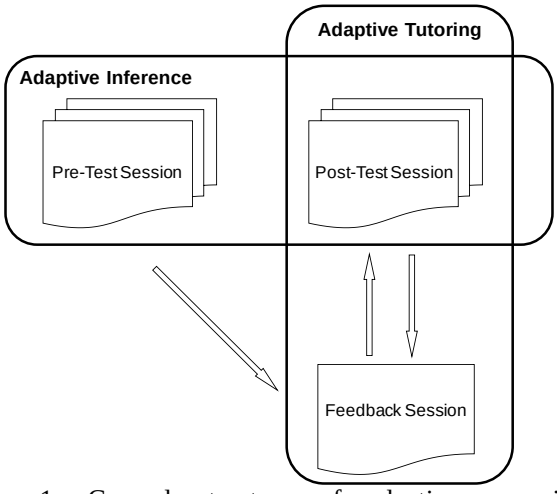


Figure 1: General structure of adaptive numerical experiment consisting of the adaptive inference and the adaptive tutoring processes.

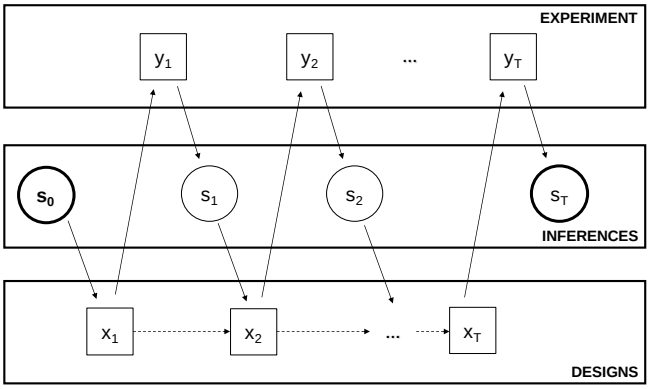


Figure 2: Flowchart of ADO process including repeated sessions of design optimization (designs), data collection (experiment), and model updating (inferences).

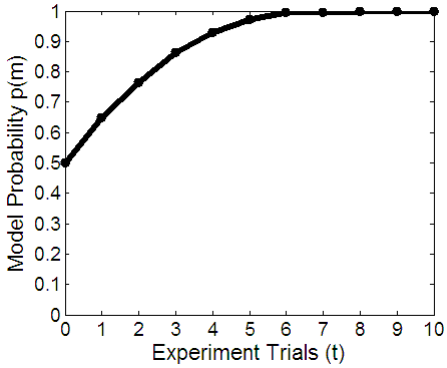


Figure 3: A typical curve of model probability change in ADO experiments.

The adaptive choice of numerical stimuli is formally done via experiment design optimization methods, where the numerical stimuli are the designs of interest. The idea of design optimization in this task is to find a numerical stimulus that is the most informative in distinguishing among alternative representational formats (i.e., logarithmic vs. linear). This method of adaptive design optimization (ADO) is developed and performed in a Bayesian framework (Myung & Pitt, 2009). In ADO, design optimization (designs), data collection (experiment), and model updating (inferences) are repeatedly performed, as illustrated in the flowchart in Figure 2. In the process, x denotes the numeric value presented to the child and is the design variable to be optimized. The symbol y denotes the child's response, and s denotes the current inference about the child's underlying representation state, such as the relative likelihood of candidate models and their parameters, which are formally defined later. The numbers the child sees in the session are updated trial by trial along the experiment.

The ADO process is performed as follows. At the beginning, the experimenter has some prior information s_0 about the child's model, from which the initial number x_1 is drawn and the response y_1 is observed. With s_0 and y_1, the posterior s_1 is obtained by Bayes theorem. For the next trial, s_1 serves as the prior and the above process is repeated. The process continues until the model information s_T after T trials meets certain stopping criterion. Such an adaptive approach bases the later designs upon previous experimental results and makes better use of individual data. Hence, it is more efficient compared to the traditional manner of using the same designs for every individual. Figure 3 shows a typical curve of model probability obtained from an ADO simulation. It indicates that the predicted model probability of the true underlying model reaches as high as .9 within four trials. To summarize, ADO-embedded adaptive inference process could find the optimal designs (i.e. numerical values to estimate) that tailor to individual state, thus could permit efficient inference from the results.

Adaptive Tutoring Process

After inferring the child's representation model through AIP, we may know that the child uses some undesired logarithmic or linear model. The next concern is to find appropriate feedback stimuli that will be most likely to induce representational shift. For this purpose, we combine the feedback session and the post-test session to form what we call the adaptive tutoring process (ATP). Design optimization methods are also applied in ATP. In the feedback session, the choice of the feedback stimuli is optimized in order to teach the child most effectively. For this purpose, we make the assumption that the effectiveness of the design is determined by the maximum discrepancy between the child's model and the target model (e.g., an accurate line $y_i = x_i$). After the optimal feedback stimulus is found and provided to the child, ATP moves to the post-test session. The post-test session infers the child's model again and checks if he has changed the model. If the child retains a logarithmic model or changes to an undesired linear model (e.g. a linear model with slope smaller than .5), the feedback and post-test sessions are repeated until the child has acquired the target model. Generally speaking, adaptive inference is also performed within the adaptive tutoring process.

The adaptive tutoring process starts from the information s_T obtained at the end of the adaptive inference process. In determining the numbers to be used for teaching, our assumption is that the most informative feedback stimuli for the child lie in the region where the target model and the child's current representation model have the largest discrepancy. The target model is assumed as a fixed, correct model. Hence, we are not adapting to the child's representation states, but are optimizing to the difference between the child's current status and the target model. Formally, we are maximizing the informativeness of the feedback stimulus described as the discrepancy between its true value and its value in the child's representation. The child is tested with the optimal feedback and is corrected with the true position. Then the experimenter obtains the updated information about the child's numerical representation model using the same process as in AIP. The updated information can be used to find the next optimal feedback stimulus, if necessary. The process runs back and forth until the child has shown acquisition of the target model by giving accurate linear responses to the numerical stimuli. In all, the adaptive tutoring process tailors to the child's learning progress and provides a way to combine optimal teaching and progress verification.

Bayesian Framework of Design Optimization

In this section, we provide a brief description of the ADO framework implemented in this paper. For fuller technical details and applications, the reader is directed to Myung and Pitt (2009) and Cavagnaro, Myung, Pitt and Kujala (2010). In ADO, each experimental design is assigned a utility describing the value of a hypothetical experiment with that design. It is analogous to choosing among a set of gambles whose payoff is determined by the risks and rewards of each type of gamble. The set of all possible designs that could be used in a given experiment consist of the design space (Amzal, Bois, Parent, & Robert, 2006; Pitt & Myung, submitted). The goal of ADO is to search the entire design space and find the most informative design(s).

The problem of design optimization is formally expressed as finding an optimal design d* over the design space, which maximizes the expected utility function U(d). U(d) typically takes into consideration of all unknown but possible conditions. If multiple models are plausible for describing the underlying process in an experiment, U(d) could be defined as:

$$U(d) = \sum_{i=1}^{K} p(m_i) \iint u(d, m_i, \theta_{m_i}, y)\, p(y \mid \theta_{m_i}, d)\, p(\theta_{m_i})\, dy\, d\theta_{m_i} \quad (3)$$

In the above equation, m_i (i = {1, …, K}) is one of K models under consideration, d is a design, y is the outcome of an experiment with design d under model m, θ_m is the

parameter of model m, and finally, u(d, θ_m, y) is the "local" utility function of deign d, parameter θ_m and experimental outcome y. In general, U(d) represents the expected value of local utility functions in which the expectation is taken over all possible models and their parameters and over all possible experimental observations given the models and parameters.

In *adaptive* design optimization, the optimization of U(d) is repeated over a series of experimental stages. At each stage, the model and parameter priors, p(m) and p(θ_m), are updated upon the specific outcome observed in an actual experiment carried out with the optimal design d*. This updating is performed via Bayes rule and Bayes factor calculation (Gelman, Carlin, Stern & Rubin, 2004).

Simulations

Pre-test Simulations and Results

In this section, we describe the computer simulations that demonstrate the performance and advantages of the adaptive numeric estimation experiment. The purpose of conducting simulations is to guarantee that the processes work as expected, as well as to show the efficiency of the methodology.

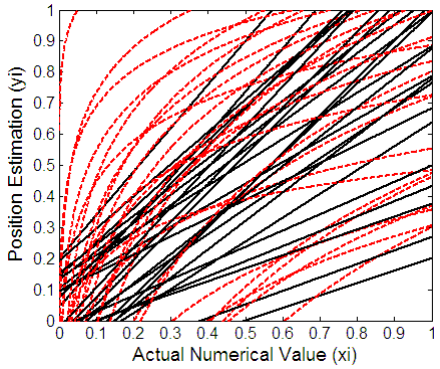


Figure 4: Sample curves of linear (black solid lines) and logarithmic (red dashed curves) models randomly generated from the priors.

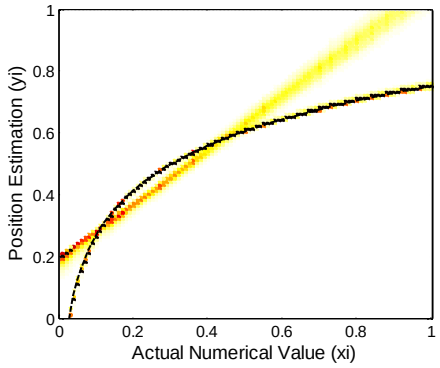


Figure 5: Prediction density scatter plot of linear and logarithmic model predictions at the end of the pre-test. The darkness of each dot indicates the probabilities of a response y given the presented number x. Black dots indicate the highest probabilities and the yellow dots indicate the lowest probabilities.

In order to run the simulations, we first chose the priors on the basis of previous experiment data and experts' beliefs, so that the priors covered a reasonable range of numerical representation models. Several data sets (e.g. Opfer & Siegler, 2007, Siegler & Opfer, 2003) were fitted and the parameter ranges of the models were obtained. Uniform priors over the parameter ranges were then used for intercept, slope, and error variance. Figure 4 shows a sample of possible models under the priors, in which the linear models and logarithmic models are mixed with each other. It also suggests the difficulty of depicting intuitive designs for distinguishing between the two sets of models.

The simulation first implemented the pre-test session with the above priors. The data-generating model, which was assumed to be the child's true model in the simulations, took the following logarithmic form:

$$y_i = 0.21 \cdot \log x_i + 0.75 + e_i,\ e_i \sim N(0, 0.005^2)$$

Within each simulation, we ran 10 trials (number of trials fixed for convenience purposes) of the Number-Line Task in the pre-test session. Results showed that after 6 trials, we had already obtained sufficient evidence to conclude that the logarithmic model was over 90% likely to be the data-generating model. Meanwhile, we also narrowed down the range of model parameters as shown in the prediction density scatter plot in Figure 5. The darkness of each dot indicates the probabilities of a response y given the presented number x. Figure 5 shows that the predictions from possible linear models are more widely spread than the predictions from possible logarithmic models. It suggests that the predictions from the logarithmic model posteriors are highly concentrated and have higher probabilities, which provides strong evidence that the true model takes a logarithmic form.

Feedback and Post-test Simulations

After the pre-test session, we simulated the adaptive tutoring process. The first step was to choose an optimal feedback stimulus that maximized the discrepancy between the target model and what we knew about the child's existing model. Formally, the utility of the feedback design accounted for the prediction probabilities of both models, as well as the parameter range of both models. For the specific simulated learner, the optimal feedback design was found at x = 0.354. That is, the child would be most "surprised" for this stimulus when he sees the difference between his response and the correct answer. Figure 6 shows the location of the optimal feedback and its relationship with the child's model and the target model.

To simulate the post-test session, we needed to assume a learning mechanism that caused the representational shift and generated the post-test experiment results. An intuitive assumption was a conservative learning mechanism in which a child learner made the smallest change to accommodate the feedback. Suppose the child could change to any models within the range of the priors. Among these models, there were a subset of linear models and a subset of logarithmic models that were consistent with the learned

feedback. A conservative learner would estimate the amount of overall discrepancy between these candidate models and the current model and choose the one that has the smallest discrepancy. That is, the conservative learning mechanism assumed the child to be an ideal learner. To demonstrate another plausible mechanism, we also assumed a less ideal learner, the model-conservative learner. The model-conservative learning mechanism assumed that the child only considered a subset of logarithmic models that were consistent with the learned feedback and chose one that required the smallest change from the previous model. In both mechanisms, the winning model was used as the data-generating model for the post-test session. Figure 7 shows representational shifts of the two hypothesized learners. After learning the optimal feedback, the conservative learner changes to a linear model $y_i = 0.758 \cdot x_i + 0.086 + e_i$, and the model-conservative learner changes to another logarithmic model $y_i = 0.218 \cdot \log x_i + 0.580 + e_i$. The two models intersect at the point of optimal feedback because they both accommodate the feedback.

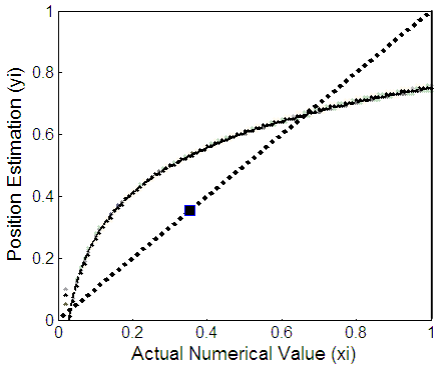


Figure 6: Optimal feedback for the simulated learner indicated by the square at x = 0.354. The prediction density scatter plot shows the inference of the child's representation at the end of the pre-test session. The dotted line shows the target model ($y_i = x_i$).

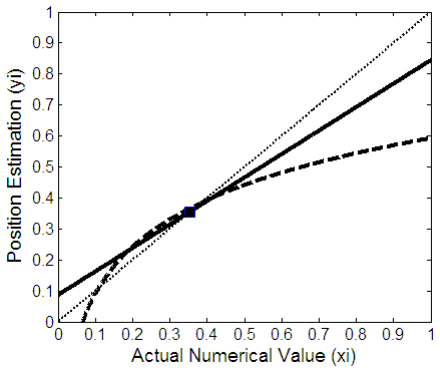


Figure 7: Predicted representational shift to the linear model (solid line) and the logarithmic model (dashed curve) caused by the two learning mechanisms. Both models intersect with the target model at the feedback (the square).

The post-test session simulation started from the same priors used for the pre-test session (shown in Figure 4). It was because the data-generating model had changed and the posterior information from the pre-test session was no longer valid. For convenience purpose, we simulated 5 trials of Number-Line Task in the post-test session. For the conservative learner, there was sufficient evidence to conclude that linear model was over 90% likely to be the data-generating model after 4 trials. For the model-conservative learner, it took 5 trials. The range of parameter estimates for the data-generating model was also narrowed down at the end of the post-test session. Hence, results from the post-test simulations showed that the post-test session made quick and reliable inferences about the new data-generating model.

In general, simulation results of the pre-test, feedback, and post-test sessions demonstrated the validity and the efficiency of the adaptive numerical experiment. We further discuss its practical applications and theoretical implications in the next section.

Discussion

Previous feedback studies have demonstrated that providing children with data that is incommensurate with their current numerical representation can promote a representational shift. In the current paper we improved upon this design using an adaptive design optimization procedure to perform an adaptive-inference, adaptive-tutoring process. This process infers the most likely dominant numerical representation and provides the optimal feedback to elicit a shift to an accurate linear representation. We simulated this process for a logarithmic learner using parameters from previous empirical experiments. Finally we predicted the learner's updated numerical representation based on two possible learning mechanisms.

We established the plausibility of the algorithm for the problem at hand. The adaptive design optimization procedure was able to infer the data generating function in each simulation by optimizing across the design space. The procedure was more efficient than traditional feedback studies in inferring the simulated child's representational state in a few trials. This efficiency in turn suggests that a shorter pre-test phase is less likely to reinforce the learner's initial representation. Shorter testing and feedback phases also provide obvious benefits to both experimentation and real world application for testing children; fewer trials reduce the overall attentional costs to children and thereby reduce the influence of attention-related noise in their responses.

The adaptive tutoring process also proved useful in determining optimal feedback. Feedback points have previously been chosen to maximize the discrepancy between an ideal logarithmic and linear function (Opfer & Siegler, 2007), while our cognitive tutor chooses personalized feedback based on the individual learner's most likely logarithmic or linear representation. This generates very informative results about the ideal feedback points.

The magnitudes chosen by the adaptive tutor are approximately 30% of the range for a simulated learner based on the parameters of children from previous studies. They are near to the previously chosen points (15% of the range), but are clearly not the same. These optimal feedback points may prove to vary widely in actual children, highlighting the need for the adaptive tutoring process to control for individual differences in representations.

The adaptive numeric estimation experiment clearly needs to be run on children to determine its external validity, which we plan to carry out. Nevertheless, we were able to use the adaptive experiment to accurately infer the representational state of a simulated learner. A byproduct of this process was the implementation of two potential learning mechanisms to test the end-state representation of the simulated learner. The conservative and model-conservative learning mechanisms were used to produce quantitative predictions. A conservative model that uses optimal feedback to adjust parameters and the model form with the least amount of change showed a shift to a more accurate linear function with parameters near to the ideal model. The model-conservative mechanism resulted in a preserved logarithmic function with an overall decrease in the model parameters.

If these results can be extended to children, they would support a perspective that learner will behave as a modeler and update his dominant representation with ideal feedback. We might then further test whether the child learner is engaging in Bayesian learning; specifically whether the different learning mechanisms can be seen as a variation in the learner's likelihood ratio. Conservative learning asserts equal likelihood to the representations, while model-conservative learning gives weight only to the dominant representation. These may be plausible mechanisms of cognitive change based on culture and the strength of each representation, with emphasis on mathematical education directly affecting the learner's likelihood ratio of a linear representation.

Adaptive inference of the probability that a learner is linear or logarithmic in representation and an adaptive tutor function that maximizes the effect of feedback are necessary to understand the learner's representation which might apply to many types of representations in diverse areas. The process could easily be extended to similar numerical estimation tasks that use a variety of presented numerical stimuli to determine perceived magnitude. It is possible to extend this design to other areas in which representational shifts are seen, whether to determine children's past tense verb use and predict errors in overgeneralization (Marcus, 1995) or function learning to predict attention to relevant cues (Kruschke 1996). The adaptive design optimization procedure is of obvious use as a means of better modeling the learner and refining training.

Acknowledgment

This research is supported in part by the NIH Grant R01-MH57472 to JIM and MAP.

References

Amzal, B., Bois, F. Y., Parent, E., & Robert, C. P. (2006). Bayesian-optimal design via interacting particle systems. *Journal of the American Statistical Association, 101(474),* 773-785.

Banks, W. & Hill, D. (1974). The apparent magnitude of number scaled by random production. *Journal of Experimental Psychology, 102(2),* 353-376.

Carey, S. (1985). *Conceptual change in childhood.* Cambridge, MA: MIT Press.

Cavagnaro, D. R., Myung, J. I., Pitt, M. A., & Kujala, J. V. (2010). Adaptive design optimization: A mutual information-based approach to model discrimination in cognitive science. *Neural Computation, 22(4),* 887-905.

Dehaene, S., Izard, V., Spelke, E., & Pica, P. (2008). Log or linear? Distinct intuitions of the number scale in Western and Amazonian Indigene Cultures. *Science, 320,* 1217-1220.

Dixon, J., & Bangert, A. (2002). The prehistory of discovery: Precursors of representational change in solving gear system problems. *Developmental Psychology, 38(6),* 918-933.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis. (2nd ed.).* Boca Raton, FL: Champan and Hall/CRC.

Kruschke, J. (1996). Dimensional relevance shifts in category learning. *Connection Science, 8(2),* 225-247.

Marcus, G. (1995). The acquisition of the English past tense in children and multilayered connectionist networks. *Cognition, 56,* 271-279.

Markman, A. B. (1999). *Knowledge representation.* Mahwah, NJ: Lawrence Erlbaum Associates.

Moyer, R. & Landauer, T. (1967). Time required for judgments of numerical inequality. *Nature, 215,* 1519-1521.

Myung, J. I. & Pitt, M. A. (2009). Optimal experimental design for model discrimination. *Psychological Review, 116(3),* 499-518.

Opfer, J. & Siegler, R. (2007). Representational change and children's numerical estimation. *Cognitive Psychology, 55,* 169-195.

Opfer, J. & Thompson, C. (2008). The trouble with transfer: Insights from microgenetic changes in the representation of numerical magnitude. *Child Development, 79(3),* 788-804.

Pitt, M. A. & Myung, I. (2002). When a good fit can be bad. *Trends in Cognitive Sciences, 6,* 421-425.

Pitt, M. A. & Myung, J. I. (submitted). Designing a better experiment. *Manuscript submitted for publication.*

Siegler, R. & Mu, Y. (2008). Chinese children excel on novel mathematics problems even before elementary school. *Psychological Science, 19(8),* 759-763.

Siegler, R. & Opfer, J. (2003). The development of numerical estimation: evidence for multiple representations of numerical quantity. *Psychological Science, 14 (3),* 237-43.

Learning from Errors in Game-Based versus Formal Mathematics Contexts

Lori A. Petersen (lpeters4@nd.edu)
University of Notre Dame
Department of Psychology, 118 Haggar Hall
Notre Dame, IN 46556 USA

Jennifer K. Heil (jheil1@nd.edu)
University of Notre Dame
Department of Psychology, 118 Haggar Hall
Notre Dame, IN 46556 USA

Nicole M. McNeil (nmcneil@nd.edu)
University of Notre Dame
Department of Psychology, 118 Haggar Hall
Notre Dame, IN 46556 USA

Gerald J. Haeffel (ghaeffel@nd.edu)
University of Notre Dame
Department of Psychology, 118 Haggar Hall
Notre Dame, IN 46556 USA

Abstract

Research suggests that educational games may be particularly useful for helping children learn STEM concepts; however, the mechanisms involved in game-based learning are not well understood. The present study tested the hypothesis that games are effective because they provide a supportive learning context that allows children to react adaptively to errors. Children (*M* age – 7 yrs, 6 mo) were given two half-hour learning sessions in which they solved nontraditional arithmetic problems (e.g., __ = 3 + 4) in game and formal contexts. In a third session, children were given a transfer test in which they solved mathematical equivalence problems (e.g., 1 + 5 = __ + 2). Children who committed more of their learning errors in the game context solved a greater number of problems correctly on the transfer test than did children who made more of their errors in the formal context. Moreover, children reacted less negatively to errors made in the game context than in the formal context. These findings suggest that educational games may be an effective learning tool because they provide a supportive context that allows children to learn from errors.

Parents and teachers often use educational games (e.g., computer games, card games, board games, etc.) to help children learn important academic skills. This strategy is intuitively appealing because educational games are widely available, and they seem to make learning fun. The use of educational games is also backed by research in psychology and education. Indeed, many prominent researchers throughout history have suggested that games and other "play" activities facilitate children's learning and cognitive development (Hirsh-Pasek & Golinkoff, 2003; Piaget, 1962; Ramani & Siegler, 2008; Schultz & Bonawitz, 2007; Vygotsky, 1967).

Research suggests that educational games may be particularly useful for helping children learn science, technology, engineering, and mathematics (STEM) concepts (Annetta, Minogue, Holmes, & Cheng, 2009; Barab, Thomas, Dodge, Carteaux, & Tuzun, 2005; Ke, 2008; Ramani & Siegler, 2008; Siegler & Ramani, 2008; Wilson, Revkin, Cohen, Cohen, & Dehaene, 2006). For example, in a series of recent experiments, Siegler and Ramani (2008, 2009; Ramani & Siegler, 2008) demonstrated that the numerical knowledge of children from low-income backgrounds could be improved substantially by playing numerical board games with equal-sized spaces that are linearly arranged and consecutively numbered. Other studies have demonstrated that children who learn STEM concepts via computer games show more motivation, more engagement, and more positive attitudes toward learning than children who learn STEM concepts via formal instruction (Annetta et al., 2009; Coller & Scott, 2009; Ota & DuPaul, 2002). Taken together, the evidence suggests that educational games have the potential to promote learning and engagement in STEM domains.

Although it is widely acknowledged that educational games can be a useful tool for learning STEM concepts, the mechanisms involved in the benefits of game-based learning are not well understood. In the present study, we focused on one potential mechanism involved in the benefits of game-based learning. Specifically, we hypothesized that games promote learning, in part, because they provide a supportive learning context that allows children to react adaptively to and to learn from errors.

All children inevitably make errors when they are learning something new, and the way that they react to these errors has the potential to affect the learning process (Baker, D'Mello, Rodrigo, & Graesser, in press; Elliot & Dweck, 1988; Dweck, 2000). Specifically, negative reactions to errors such as frustration, anxiety, or helplessness are likely to hinder learning (Ashcraft & Kirk, 2001; Baker et al., in press; Dweck, 2000).

Importantly, research suggests that the nature of the learning context can influence how children react to their errors (Mueller & Dweck, 1998; Okolo, 1992). Some

learning contexts are more supportive than others. Supportive learning contexts are non-evaluative and deemphasize the association between errors and intelligence (Burhans & Dweck, 1995; Elliot & Dweck, 1988; Dweck, 2000). Such contexts buffer children from reacting negatively to errors and encourage children to persist longer in the face of errors (Okolo, 1992).

We propose that the benefits of games may derive, at least in part, from the supportive learning context they provide. Games are less evaluative than formal learning contexts. Children's performance is typically not graded during games, and failure during games can often be attributed to luck. Thus, games may help deemphasize the association between errors and intelligence. For these reasons, games should help children learn because they remove the evaluative factors that often cause children to lose motivation for learning. If children do not feel like they are being evaluated, then they may react more adaptively to their errors.

In contrast, formal contexts may make children feel more evaluated. When children err in a formal context, their sense of intelligence may be threatened, and they may respond with helpless behaviors. For example, children might stop trying to solve the problems correctly so that poor performance can be attributed to lack of effort rather than to low intelligence. If children's focus is on being evaluated instead of on learning, then they may react more negatively to their errors.

In the present study, we tested these ideas by studying a group of children who were learning to solve mathematics problems in the context of both games and formal flashcards. If games provide a supportive learning context for making errors, then children should be less likely to react negatively to the errors they make in a game context versus a formal context. Thus, we hypothesized that the proportion of errors that children reacted to negatively during the games would be lower than the proportion of errors that children reacted to negatively during the flashcards. Moreover, if games facilitate learning *because* they provide a supportive context for making errors, then children's learning should benefit from erring more in a game context relative to a formal context. Thus, we hypothesized that children who committed more of their errors during the games would learn more than children who committed more of their errors during the flashcards.

Method

Participants

This study used existing data from a larger study that tested how various ways of solving addition problems affect children's understanding of mathematical equivalence. The participants of interest were 37 children who participated in two sessions in which they learned to solve addition problems that were presented in a nontraditional format (e.g., __ = 3 + 4; 10 = 6 + __). Children were recruited from a diverse range of public and private elementary schools in a mid-sized city in the midwestern United States. One child was excluded because he did not make any errors over the course of the learning sessions. Thus, the sample contained 36 children (*M* age = 7 years, 6 months; 19 boys, 17 girls; 3% Asian, 3% Hispanic or Latino; 11% African-American or black; 83% white).

Procedure

Children participated individually in three half-hour sessions. During the first two sessions, children learned to solve nontraditional addition problems (e.g., __ = 3 + 4; 10 = 6 + __) by playing games one-on-one with a tutor (i.e., game context) and by answering flashcards (i.e., formal context). All children participated in both the game and formal contexts in alternating order during both sessions. Each session started with games, continued onto flashcards, and then ended with more games. During a third session, children were introduced to a new experimenter who assessed their learning by giving them a transfer test. All three sessions were video recorded.

Learning sessions The learning sessions were designed to help children solve single-digit addition facts with two addends (e.g., 17 = 9 + 8, 14 = 8 + 6). All problems were presented in a nontraditional format with the operation on the right side of the equal sign. This format is considered to be "nontraditional" because arithmetic problems are traditionally presented with the operations on the left side of the equal sign. Children learned via two main types of activities: (a) two-player games involving cards, dice, or the computer, and (b) flashcards. Children received feedback about correctness throughout the sessions in both the game and formal contexts, and any errors were corrected.

Game context. Children played several two-player games over the course of the learning sessions with the experimenter. One game was a modified version of "Snakey Math" by Curry K. Software. In this computer game, an addition problem was presented at the bottom of the computer screen (e.g., __ = 3 + 4), and several possible numbers (e.g., 7, 1, 12, 8) were scattered in random locations on the screen. The child and the tutor each controlled an animated snake, and the goal was to be the first snake to "eat" the number that correctly solved the addition problem.

Another game was called "Smack it!" In this card game, the child and tutor each used a swatter with a suction cup at the end. At the beginning of the game, four addition problems were placed face-up on the table, and a pile of number cards were placed face down. To start each round, the tutor turned over one of the number cards to serve as the target number. The goal was to be the first player to "smack" the addition problem that should have the target number in the blank. Children also played other two-player games that were similar in content and scope. Most of the games were rigged so the child would win; however, some games involved luck, so the tutor occasionally won. Overall, children solved an average of 46.03 problems in the game context across the two learning sessions.

Formal context. The formal context consisted of flashcards presented in succession. Before completing the

flashcards, children received a brief demonstration on how to solve the flashcards. Children solved an average of 45.11 flashcards in total across the two learning sessions. Thus, there was not a significant difference in the number of problems that children solved in the game and formal contexts, $F(1, 35) = 0.21$, $p = 0.65$.

Transfer test Children solved four mathematical equivalence problems (1 + 5 = __ + 2, 7 + 2 + 4 = __ + 4, 2 + 7 = 6 + __, 3 + 5 + 6 = 3 + __). Similar to the addition problems solved during the learning sessions, these problems do not correspond to the traditional "operations on left side" format, so they drew on the knowledge that children had gained from the learning sessions. However, they were much more difficult than the problems solved in the learning sessions because they have operations on both sides of the equal sign. Children never saw problems with operations on both sides of the equal sign during the learning sessions. Previous research has shown that most children in this age range in the U.S. have trouble solving mathematical equivalence problems correctly in the absence of special instruction (Alibali, 1999; Falkner, Levi, & Carpenter, 1999; McNeil & Alibali, 2005; Perry, Church, & Goldin-Meadow, 1988). We limited the transfer test to four problems for the sake of efficiency because previous research has shown similar performance on mathematical equivalence problems regardless of whether children solve three, four, or more than four problems (e.g., Alibali, 1999; Perry, 1991; Rittle-Johnson & Alibali, 1999; Siegler, 2002).

When each problem was presented, the tutor told the child to figure out what number to put in the blank to make the right side of the equal sign the same amount as the left side of the equal sign. If the child provided the correct number, the tutor gave positive feedback, such as "good job" and then moved on to the next problem. However, if the child provided an incorrect number, the tutor provided the feedback as follows: "No, that's not the number that goes in the blank. The correct number is *x* because *a* plus *b* is equal to *x* plus *y*" (the actual numbers in the problem were used in the place of *a*, *b*, *x*, and *y*).

Coding

Errors during the learning sessions Children's errors during the learning sessions were tallied, and the total number of errors made in the game context was compared to the total number of errors made in the formal context.

Reactions to errors Children's immediate reactions to hearing that they had made an error were coded as "negative" or "not negative." Reactions were coded as "negative" if children said something negative (e.g., "this is hard," "I'm getting really messed up," "no fair") or exhibited negative behaviors (e.g., whining, growling, huffing, rolling their eyes, or withdrawing). Reliability was established by having a second coder code the reactions of 20% of the children. Agreement between coders was 81.5%.

Transfer performance Children's solutions on the transfer test were coded as correct or incorrect based on a system used in prior work (e.g., Alibali, 1999; Perry et al., 1988; McNeil & Alibali, 2004; Rittle-Johnson, 2006). Children were given a point for every correct solution. Scores ranged from 0-4.

Results

Performance during the learning sessions was highly variable across children. Collapsing across the game and formal contexts, children made an average of 13.70 (SD = 11.01) errors. To test if children made more errors in the game or formal context, we performed a repeated measures analysis of variance (ANOVA) with context (game or formal) as the independent variable and number of errors as the dependent variable. There was no statistical difference in the number of errors that children made in the game context ($M = 6.53$, $SD = 4.73$) versus the formal context ($M = 7.19$, $SD = 7.95$), $F(1, 35) = 0.32$, $p = .58$.

Although there were not general patterns in terms of which context elicited more errors, there were individual differences in which context elicited more errors. Some children made more of their errors in the game context (n = 20), whereas some children made more of their errors in the formal context (n = 16). We predicted that children who made more of their errors in the game context would learn more than and perform better on the transfer test than children who made more of their errors in the formal context.

To test our hypothesis, we performed a between-subjects ANOVA with error group (more errors in game context or more errors in formal context) as the independent variable and number correct on the transfer test (out of 4) as the dependent variable. Consistent with our predictions, there was a significant main effect of error group, $F(1, 34) = 5.99$, $p = .02$, $\eta^2 = .15$. Children who made more of their errors in the game context performed better on the transfer test (M = 2.70, SD = 1.75) than did children who made more of their errors in the formal context (M = 1.13, SD = 1.62). These results held even when controlling for the total number of errors made across contexts (total number of errors was not a statistically significant predictor of transfer performance, $F < 1$).

Results also held when the independent variable was treated as a continuous predictor and a regression analysis was performed. For the regression analysis, we calculated a difference score by subtracting the total number of errors each child made in the formal context from the total number of errors that child made in the game context. Thus, a positive difference score reflects more errors made in the game context relative to the formal context. This difference score was then used to predict number correct on the transfer test (out of 4). As predicted, the difference score was positively associated with performance on the transfer test, $b = 0.12$, $t(34) = 3.00$, $p = 0.005$. The greater the difference between the errors made in the game versus formal context, the greater the number of transfer problems solved correctly. More specifically, for every additional

error made in the game context versus the formal context, the number correct on the transfer test increased by 0.12 (out of 4). The effect was moderate, with the difference score accounting for 21% of the variance in transfer performance.

Finally, we hypothesized that it would be more beneficial for children to make their errors in the game context versus the formal context because games provide children with a more supportive context for making errors. According to this account, children should be less likely to react negatively after making an error in the game context than after making an error in the formal context. To test this prediction, we calculated the proportion of errors that children reacted to negatively in the game context and the proportion of errors that children reacted to negatively in the formal context. Five children were excluded from this analysis because they did not make at least one error in both contexts. We then performed a repeated measures ANOVA with context (game or formal) as the independent variable and proportion of errors that children reacted to negatively as the dependent variable. Consistent with predictions, there was a significant main effect of context, $F(1, 30) = 5.47$, $p = .03$, $\eta^2 = .15$. The proportion of errors that children reacted to negatively was lower in the game context ($M = .14$, $SD = .18$) than it was in the formal context ($M = .28$, $SD = .30$).

Discussion

Games are widely used to teach children STEM concepts because they are intuitively appealing, and they promote learning and motivation. The results of the present study suggest that games may be an effective instructional tool for learning mathematics concepts because they provide a supportive context for making errors. Children who made more of their errors in the game context learned more than did children who made more of their errors in the formal context. This was confirmed by superior performance on the transfer test. Moreover, children had fewer negative reactions to the errors they made in the game context than they did to the errors they made in the formal context. This suggests that games may provide children with a supportive context that allows children to react adaptively to errors, which promotes learning.

Errors are inevitable during the learning process, and how children react to these errors may have important implications for learning. When children react negatively to errors, they may exhibit frustration, anxiety, or helplessness. Such behaviors reduce the probability of learning (Baker et al., in press; Elliot & Dweck, 1988; Dweck, 2000). In contrast, when children do not react negatively to errors, they may be more likely to persist in the face of challenge and regard errors as an opportunity to learn (Dweck, 2000; Okolo, 1992). Such behaviors increase the probability of learning. The present results suggest that games may facilitate learning, in part, because they buffer children from reacting negatively to errors.

Although the results of this study supported our hypotheses, it is important to note that this study was not designed specifically to test the mechanisms by which game contexts outperform formal contexts. The data were collected as part of a larger study that was designed for a different purpose, so future studies will be needed to corroborate the results and rule out alternative explanations. For example, it is possible that the difference between game and formal contexts could be due to an individual difference variable that leads children to perform worse in both the formal context and the transfer test. Specifically, children who have mathematics anxiety may have made more errors in the formal context and on the transfer test because both of these contexts resemble traditional school contexts, and thus, might have been viewed as an evaluative, anxiety-provoking situation.

Alternatively, it is possible that children who committed more errors in the game context (versus the formal context) performed better on the transfer test not because they were buffered from negative reactions in the game context, but because the specific act of playing a game made them more engaged in learning. If children learn to solve the problems correctly in the game context, then they will be more likely to win the game. Thus, it is possible that learning is more instrumental in the game context than in the formal context. Future research should control for this potential confound.

Future research should also examine whether the present results generalize to the classroom setting. In the present study, children learned in game and formal contexts while working one-on-one with a "tutor" who stuck to a meticulous script. More typical learning environments are often less structured and less conducive to one-on-one instruction. In order to determine the practical effectiveness of the game context on learning, future studies should investigate whether the results generalize to the types of game and formal contexts that are used in classroom environments.

Overall, the present results are consistent with prior research suggesting that educational games can be helpful for learning STEM concepts. Results suggest that games are helpful not just because they are fun and engaging, but also because they provide a supportive context for making errors. Future work should continue to investigate the benefits of educational games and other innovative contexts that facilitate children's learning.

Acknowledgments

This paper is based, in part, on a senior honors thesis conducted by Heil under the direction of McNeil and Haeffel. Thanks to April Dunwiddie, Tom Merluzzi, Emily Fyfe, Crysta Sulaiman, Megan Heil, members of the Cognition Learning and Development Lab, and members of the Cognition and Emotion Lab.

References

Alibali, M. W. (1999). How children change their minds: Strategy change can be gradual or abrupt. *Developmental Psychology, 35*, 127-145.

Annetta, L. A., Minogue, J., Holmes, S. Y., & Cheng, M. (2009). Investigating the impact of video games on high

school students' engagement and learning about genetics. *Computers & Education, 53*, 74-85.

Ashcraft, M. H., & Kirk, E. P. (2001). The relationships among working memory, math anxiety, and performance. *Journal of Experimental Psychology, 130*(2), 221-237.

Baker, R., D'Mello, S. K., Rodrigo, M., & Graesser, A. C. (in press). Better to be frustrated than bored: The incidence and impact of learners' affect during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies*.

Barab, S., Thomas, M., Dodge, T., Carteaux, R., & Tuzun, H. (2005). Making learning fun: Quest Atlantis, a game without guns. *Educational Technology Research & Development, 53*(1), 86-107.

Burhans, K. K., & Dweck, C. S. (1995). Helplessness in e arly childhood: The role of contigent worth. *Child Development, 66*, 1719-1738.

Coller, B. D., & Scott, M. D. (2009). Effectiveness of using a video game to teach a course in mechanical engineering. *Computers and Education, 53,* 900-912.

Dweck, C. S. (2000). *Self-theories: Their role in motivation, personality, and development.* Philadephia, PA: Psychology Press.

Elliott, E. S., & Dweck, C. S. (1988). Goals: An approach to motivation and achievement. *Journal of Personality and Social Psychology, 54*, 5-12.

Falkner, K. P., Levi, L., & Carpenter, T. P. (1999). Children's understanding of equality: A foundation for algebra. *Teaching Children Mathematics, December*, 232-236.

Hirsch-Pasek, K., & Golinkoff, R. M. (2003). *Einstein never used flash cards: How our children really learn and why they need to play more and memorize less.* Emmaus, PA: Rodale Press.

Ke, F. (2008). Alternative goal structures for computer game-based learning. *Computer-Supported Collaborative Learning, 3*, 429-445.

McNeil, N. M., & Alibali, M. W. (2004). You'll see what you mean: Students encode equations based on their knowledge of arithmetic. *Cognitive Science, 28,* 451-466.

McNeil, N. M., & Alibali, M. W. (2005). Why won't you change your mind? Knowledge of operational patterns hinders learning and performance on equations. *Child Development, 76*, 883-899.

Mueller, C. M., & Dweck, C. S. (1998). Praise for intelligence can undermine children's motivation and performance. *Journal of Personality and Social Psychology, 75*(1), 33-52.

Okolo, C. M. (1992). The effects of computer-based attribution retraining on the attributions, persistence, and mathematics computation of students with learning disabilities. *Journal of Learning Disabilities, 25*(5), 327-334.

Ota, K. R., & DuPaul, G. J. (2002). Task engagement and mathematics performance in children with attention deficit hyperactivity disorder: Effects of supplemental computer instruction. *School Psychology Quarterly, 17*, 242-257.

Perry, M. (1991). Learning and transfer: Instructional conditions and conceptual change. *Cognitive Development, 6*, 449-468.

Perry, M., Church, R. B., & Goldin-Meadow, S. (1988). Transitional knowledge in the acquisition of concepts. *Cognitive Development, 6,* 449-468.

Piaget, J. (1962). *Play, Dreams and Imitation in Children.* New York, NY: W. W. Norton & Co.

Ramani, G. B., & Siegler, R. S. (2008). Promoting broad and stable improvements in low-income children's numerical knowledge through playing number board games. *Child Development, 79*, 375-394.

Rittle-Johnson, B. (2006). Promoting transfer: Effects of self-explanation and direct instruction. *Child Development, 77*, 1-15.

Rittle-Johnson, B., & Alibali, M. W. (1999). Conceptual and procedural knowledge of mathematics: Does one lead to the other? *Journal of Educational Psychology, 91*, 175-189.

Schulz, L. E., & Bonawitz, E. B. (2007). Serious fun: Preschoolers engage in more exploratory play when evidence is confounded. *Developmental Psychology, 43*(4), 1045-1050.

Siegler, R. S. (2002). Microgenetic studies of self-explanations. In N. Granott & J. Parziale (Eds.), *Microdevelopment: Transition processes in development and learning* (pp. 31-58). New York: Cambridge University.

Siegler, R. S., & Ramani, G. B. (2009). Playing linear number board games – but not circular ones – improves low-income preschoolers' numerical understanding. *Journal of Educational Psychology, 101*(3), 545-560.

Siegler, R. S., & Ramani, G. B. (2008). Playing linear numerical board games promotes low-income children's numerical development. *Developmental Science, 11*, 655-661.

Vygotsky, L. S. (1967). Play and its role in the mental development of the child. *Soviet Psychology, 5*, 6-18.

Wilson, A. J., Revkin, S. K., Cohen, D., Cohen, L., & Dehaene, S. (2006). An open trial assessment of "The Number Race," an adaptive computer game for remediation of dyscalcula. *Behavioral and Brain Functions, 2,* 1-16.

Collaborative Facilitation through Error-Detection: A Classroom Experiment

Soniya Gadgil (smg58@pitt.edu)
Timothy J. Nokes (nokes@pitt.edu)
University of Pittsburgh
Learning Research and Development Center, 3939, O'Hara Street
Pittsburgh, PA, 15260 USA

Abstract

Prior work has shown that individuals working in groups often perform worse than individuals working alone, a finding commonly referred to as collaborative inhibition. In the current work we examine whether engaging in error correction processes can mitigate or eliminate the collaborative inhibition effect and perhaps even facilitate collaborative facilitation. Participants engaged in a writing error-detection and revision task while working either with a partner or individually. On the error-detection task, dyads found more structural flaws in the text, whereas individuals found more surface flaws. Moreover, when comparing dyads nominal groups the dyads did not show the collaborative inhibition effect. A similar pattern of results was found on the revision task. The results are discussed in terms of the underlying cognitive and social processes that support successful collaboration.

Keywords: collaborative learning; error-detection; instruction.

Introduction

When does collaboration lead to robust performance and learning outcomes? A large amount of evidence from past research shows that when individuals collaborate with one or more partners, it leads to better performance outcomes when compared to the average individual (see Hill, 1982; Kerr & Tindale, 2004 for reviews). This result has been found in number different tasks and domains. It is hypothesized that groups are able to "pool" their resources and knowledge to perform better than the average individual whether brainstorming, memorizing lists of words, or solving puzzle problems.

Although groups tend to perform better than the average individual, individuals working in groups often do not perform up to their predicted potential. An extremely robust finding in the collaboration literature is that individuals working in groups actually perform worse than individuals working alone (Andersson & Ronnberg, 1995; Weldon & Bellinger, 1997). This has been referred to as "collaborative inhibition" or "process loss" (Steiner, 1972). It is often measured by comparing the dyad or group performance to nominal group performance. For example, when comparing dyads and individuals, a nominal dyad is formed by randomly pairing two individuals who did not collaborate, and their joint performance is considered, as if they were a collaborating dyad. In a simple list learning task, if a dyad recalled the following letters from a list (a, b, c, d, e, f, g) and two individuals recalled the following: individual 1 (a, b, e, g, h) and individual 2 (c, d, e, f, i) the dyad performs better than the average individual (7 vs. 5) but worse than the nominal dyad (pooled performance: 7 vs. 9).

Much research has focused on trying to understand the causes of collaborative inhibition. Both cognitive and social factors have been advanced to explain it. Social factors include the free-rider effect or social loafing (Karau & Williams, 1993), evaluation anxiety (Collaros & Anderson, 1969; Mullen, 1983), and diffusion of responsibility. Cognitive factors include cognitive overhead of coordination during collaboration (Steiner, 1972) and disruption of retrieval strategy due to interference caused by the collaborators' input (Basden, Basden, Bryner, & Thomas, 1997; Finlay, Hitch, & Meudel, 2000; Weldon, Blair, & Huebsch, 2000). Each of these factors has been shown to contribute to the collaborative inhibition effect.

In addition to identifying factors that increase or decrease collaborative inhibition a few studies have shown an elimination of the collaborative inhibition effect or even an advantage for collaborative groups over nominal groups. For example, Wright and Klumpp (2004) compared individuals and two collaborative group conditions during free recall: a "see" condition in which people in the pairs took turns recalling items from a previously studied list and showed each other the words as they were being recalled, and a "no see" condition, in which the participants again took turns recalling the previously seen list, but neither knew which items the other person had recalled. Thus, the "no see" condition was effectively the same as a nominal group, as the participants did not engage in any form of interaction while recalling the items. Not surprisingly, Wright and Klumpp found that the "no see" group performed significantly better than the "see" group, and equal to nominal groups.

In another demonstration of collaborative facilitation, Takahashi and Saito (2004) compared recall of studied story materials by nominal dyads and collaborators. When tested immediately, they found that nominal dyads performed better than collaborators, however, when tested after a one-week delay, collaborators recalled more than nominal dyads.

These findings suggest that there must be some aspect of the task structure in which collaborators engage that play a part in determining collaborative inhibition or

advantage. We propose that if the task structure facilitates the cognitive mechanisms hypothesized to underlie the collaborative advantage, we should be able to overcome the collaborative inhibition effect.

One of the primary mechanisms suggested to underlie successful collaboration is error-detection (Shaw, 1932; Sniezek & Henry, 1989). Groups are hypothesized to engage in a higher degree of error detection and correction compared to individuals. It has been widely documented that detecting your own errors is an important metacognitive skill, however, not many learners have such skills. Moreover, in order to detect an error, it is necessary to have the requisite domain knowledge, which an individual may not possess, but a collaborator may. This results in more errors being detected and corrected. Further, being in an interactive situation, dyads are more likely to engage in constructive processes such as explanation, and therefore more likely to detect errors when things don't compute. There is evidence that by scaffolding learners' interactions to encourage explanation, they were able to form more coherent representations of science concepts (Coleman, 1998).

If error-detection is indeed a mechanism underlying collaborative facilitation, then engaging in an error-detection task collaboratively should help mitigate the collaborative inhibition effect. In other words, the task of error-detection should lead to dyads performing at least as well as nominal dyads. Past studies have proposed error-detection as a mechanism, but not studied it as an aspect of the task structure. In the current experiment, we employed error-detection as the task in which participants would engage either collaboratively or individually.

We decided to test this hypothesis in a college classroom in the context of writing summaries of empirical articles. One reason for choosing this domain was that it provided an ideal open-ended task for students to work on in dyads or individually. Second, research in writing instruction has consistently shown how generating a coherent summary of read material is a challenging task for most students (e.g., Flower, 1979). In any college course with a substantial writing component, especially research reporting, students have the most difficulty summarizing related research succinctly and relating it to their own ideas. The errors that students make are due to imperfect understanding of what constitutes a good summary. It is not intuitive for students to understand the difference between a good summary and a bad one, without engaging in deliberate cognitive processing.

Most of the past work that has found a collaborative advantage has been with simple tasks such as list learning and tested with recall or recognition judgment tasks. We wanted to extend this further to a task involving higher order processing than simple recall.

Finally, testing this paradigm in a real classroom also gave us increased ecological validity. This paradigm has not yet been explored in a controlled experimental way in a real classroom. Investigations of collaborative learning have been conducted either in a lab setting, where a degree of strict experimental control is possible or in educational settings where factors such as random assignment have been implemented due to various constraints of working in a classroom. Recent endeavors have taken findings from cognitive science and attempted to apply them in authentic learning situations (e.g., Nokes & VanLehn, 2008). We followed in this tradition, and explored this paradigm in a cognitive psychology lab classroom, without sacrificing experimental rigor.

We propose that by collaborating with a peer, students will be more likely to detect flaws in a given summary. Collaborating peers will bring different knowledge to bear on the issue, not all of which will be overlapping. As stated before, we hypothesized that working with a peer will be able to detect a greater number of errors than those working individually. Moreover, we expect that by collaboratively engaging in error-correction, collaborative dyads will outperform nominal dyads, or at least equal them on performance.

We also wanted to see whether the benefits of collaborative error-detection extend to a subsequent on the revision task. In the revision task, students were asked to revise the initial error-ridden summary with the same partner or individually. We hypothesized that because dyads will uncover a greater number of errors to begin with, they will be more likely than individuals to correct those errors, and will perform better than individuals.

Method

Participants

Fifty students from University of Pittsburgh (32 females and 18 males) participated in the study. These students were from three of the lab sections of the course Cognitive Psychology for Majors. Most of the students were upperclassmen (juniors or seniors).

Design

The design was between subjects and students were randomly assigned to either the individual condition or were randomly paired with a partner from the same section without considering gender or ability, and assigned to the collaborative condition. The two main dependent variables of interest were the performance on the error-detection task as measured by number of errors identified and performance on the revision task.

Materials and Procedure

The experiment was conducted over a three-week period, and comprised of homework assignments and in-class activities. The flowchart shown in Figure 1 describes the activities that students performed.

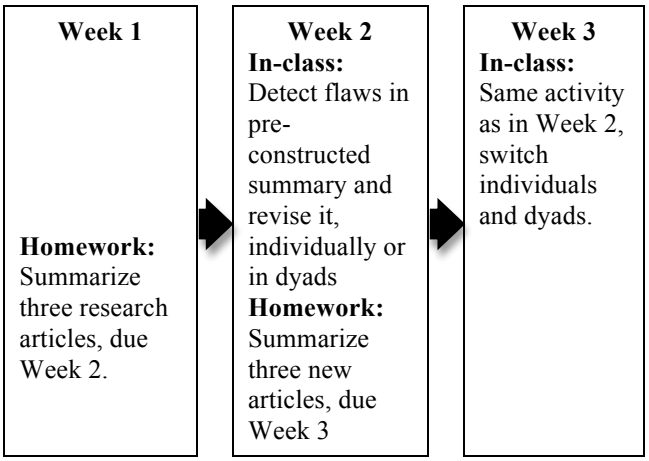


Figure 1: Flowchart of procedure

During week 1, students were asked to summarize three articles on a topic in cognitive psychology. They could choose out of six articles, but one of them was mandatory, because the in-class activity in the following week would be based on that article. The articles were abridged versions of published research articles and consisted of just the Abstract, Method, and Results section. Assigning the summarizing homework prior to the in-class activity ensured that participants were familiar with the task before they worked on it in class.

During the in-class activity in week 2, students were given a pre-constructed summary of the mandatory research article that they had read and summarized in their homework. This summary contained a number of errors. Students were given the same article they had summarized in their homework in order to cut down on time needed to read the article in order to summarize it. During the in-class activity, students were randomly assigned to either the individual or dyadic condition. Each person or dyad got the pre-constructed summary as well as the original abridged article. They were told that this summary was constructed by another student, and their job was to list the flaws in that summary and then rewrite the summary to revise it. At the end of the class, they were assigned a new homework activity in which they summarized three new articles. This homework was due on the class of week 3.

The experiment took place during a regular weekly lab as part of their normal instruction. Students were given 50 minutes to enlist the flaws in the summary and write a revised version. No other scaffolding was provided during the experiment.

A rubric was developed to score students' completed worksheets. First, students' list of flaws was examined to determine how many flaws they could correctly identify. This was compared with a list of all flaws in the document, which could be either structural level flaws or surface level flaws. Structural level flaws included flaws such as "research question not stated" or "participant characteristics absent". Surface level flaws were stylistic flaws, for example, "summary was not indented" or "italicization in reporting of statistics was incorrect." There were a total of 11 structural level flaws and 6 surface level flaws in each summary. See appendix A for a list of flaws.

Next, a rubric was created to score the revised summaries that students had developed. There were 12 criteria that needed to be fulfilled in order to get full credit. See appendix A for a list of criteria.

Results

We will first describe the performance of dyads and individuals on the error detection task. We will then see whether there is a difference in performance when individuals are randomly paired with another individual to form nominal dyads. This will be followed by an analysis of the scores that dyads or individuals received on the revision task, and subsequently whether dyads and nominal dyads differed on the revision task. Finally, we will see whether the effects of the error-detection activity transferred to a new but related situation, by examining students' performance on the homework assignment immediately after the in-class activity.

As we had hypothesized, dyads performed better than individuals on the error detection task. That is, dyads could detect a higher number of structural-level flaws in the summaries compared to individuals. Dyads could detect 2/3 of the total number of flaws, whereas individuals could detect only ½ of them. See Figure 2 for means and standard errors. A 2 (collaboration: dyads versus individuals) x 2 (error: structural versus surface) mixed ANOVA showed no effect of collaboration, F (1, 33) = 2.33, ns. However, there was a main effect of error type with structural errors being better identified than surface level errors, In addition, there was a significant interaction of collaboration by error type, such that dyads were better at detecting structural level flaws than individuals whereas there was no difference between them in detecting surface level flaws, F (1, 34) = 10.83, $p < .05$.

Next, we looked at dyads versus nominal dyads. We used Kelley and Wright's (2010) procedure to form nominal dyads[1], and looked at the unique number of

[1] Most studies that have compared nominal dyads and collaborators in the past have randomly paired individuals to form nominal dyads. This introduces an unnecessary source of errors, and it is advisable to use all possible pairs of nominal dyads to reduce this error. However, with a sample size of 20, one would need to look at 2×10^{24} pairs of nominal dyads, which is computationally almost intractable. Kelley and Wright (2010) have written a program that randomly selects 10,000 pairs of nominal dyads and then generates a list of nominal dyads with a mean and standard deviation closest to the true mean.

errors identified by each nominal dyads. For example, if one member of the nominal dyad identified errors 1, 2, 3, 4, and 5 and the other identified 4, 5, 6, and 7, their total score was 7. The means and standard errors are shown in Figure 2.

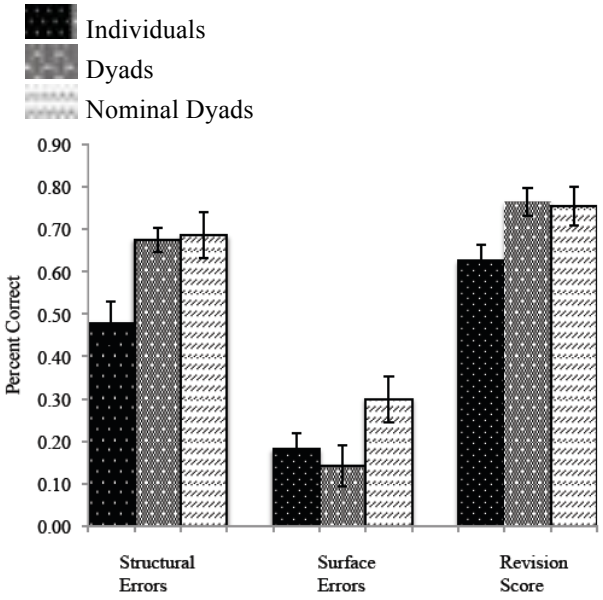


Figure 2: Means and standard errors for individuals, dyads, and nominal groups.

A mixed ANOVA with follow-ups using the LSD procedure (alpha = .05) was performed to examine the effects of collaboration on the number of structural and surface level errors detected. There was a significant main effect of error type such that participants found significantly more structural errors compared to surface errors, $F(1,23) = 83.64$, $p = 0.00$. There was also a significant main effect of condition such that nominal dyads detected a significantly more number of errors overall, compared to collaborators, $F(1,23) = 17.70$, $p = 0.048$. There was an marginally significant interaction between condition (nominal vs collaborative) and error-type (structural vs surface), $F(1,20) = 3.45$, $p = 0.076$. Follow-up tests using Fischer's LSD (LSD = .15) showed that the difference between nominal dyads and collaborators was significant only for surface level errors. The two groups were not different on structural level errors. However, both groups found a significantly larger number of structural level errors compared to surface level errors.

Thus, although dyads did not outperform nominal dyads in detecting structural level flaws, they were equally good in terms of their individual performance within the dyad. We found evidence for collaborative inhibition on the surface level flaws. This might be due to social pressure to identify only those flaws that the students' thought would be considered most important, such as structural-level flaws and that perhaps surface-level flaws were not considered important or so obvious to be easily fixed.

Next, we looked at the performance of dyads and individuals on the revision task. The revised summaries were scored on a rubric where the maximum possible score was 20 points. The means and standard deviations for the revision score are displayed in the last column of Figure 2. A one-way ANOVA revealed that dyads significantly outperformed individuals on revising the summaries, $F(1, 34) = 6.57$, $p < .05$. Thus, the benefit of error detection activity extended to the actual revision of summaries, and reinforcing the collaborative advantage.

We then compared scores on the revision task for nominal dyads and dyads. Similar to the error-detection task, we awarded one point for every criterion that either or both of the two partners got correct in a nominal dyad. A one-way ANOVA revealed that the dyads and nominal dyads were not significantly different from one another, $F(1, 23) = .03$, ns.

To understand how the in-class error-detection activity impacted students' performance on subsequent writing assignments, we looked at their scores on homework assignments immediately following the in-class session. This is a transfer task, because we expected students to apply what they had learned during the in-class activity (error-detection) to generating their own summary of an article.

We expected students who found a greater number of errors to score better on the homework assignment, because they would be less likely to commit the same errors while summarizing an article. We found a marginal correlation on the subsequent homework such that the score on the homework assignment correlated with the number of errors that they detected during the in-class activity $r(44) = .28$, $p = .058$. There was however no difference by condition, that is the scores of the collaborative participants and individual participants did not differ significantly, after controlling for their performance on the earlier homework, $F(1, 43) = .421$, ns.

Discussion

In the present study, we investigated whether by promoting the mechanisms underlying collaboration, we can overcome the collaborative inhibition effect reported widely in the literature. Our results from this experiment are very encouraging, and provide evidence that by structuring collaborative learning activities according to the cognitive processes underlying it, we can get collaborative learners to perform at least as well as nominal dyads.

We found that engaging in an error-detection task with a partner led to better performance on detecting structural level errors than doing so individually. Even more important is the finding that when the dyads were compared with nominal dyads, they did not do worse, than the nominal dyads unlike many past studies (e.g

Andersson & Ronnberg, 1995). However, individuals found a greater number of surface level errors. One of the possible explanations for this is that dyads focused on the structural level features and ran out of time before getting to the surface level features. Individuals on the other hand, because they could find only a certain number of structural errors, moved on to the surface level errors, and were able to detect more of them. However, as noted before, the overall rate of detection of surface level errors was low, indicating that both individuals and dyads focused more on the structural features.

The other important finding from this study was that dyads performed significantly better than individuals when they revised the flawed summaries. When comparing revision scores of collaborators and nominal dyads, we found no difference between the two. Thus, we have evidence that benefit of error-detection extended to the revision task as well.

We also tested the effects of collaborative error-correction on a measure of transfer when we looked at whether the students' performance on the error-detection task affected their performance on a subsequent homework, which involved generating their own summaries. We found that the number of errors detected during the in-class activity was correlated with their score on the homework assignment. Although we did not find a significant difference between scores of individuals and collaborators, the correlation indicates that students who detected more errors were more likely to perform better on the summarizing task, regardless of condition. There are some caveats to our findings. The first is that since this experiment was conducted in a classroom setting, we could not control all variables as strictly as we would have liked to, in a laboratory setting. We therefore aim to replicate this in a more stringently controlled environment, and understand collaborative error-correction at a more fine-grained level.

Next, we need to replicate this finding in a different domain, and find out whether the effects of collaborative error-detection are robust enough to be found across various domains, such as conceptual physics or mathematics problem solving.

Several issues still need to be addressed in understanding why error-detection leads to better collaborative outcomes. It is clear that error-detection encourages some kind of constructive activity in collaborators that causes them to perform better than individuals. Process data such as verbal protocols can help us better understand what these constructive activities are.

For example, the study by Okada and Simon (1997) found that dyads were more likely than individuals to generate explanations. It would be helpful to analyze process data from collaborative error-detection and understand whether collaborators are more likely to generate explanations for the errors they detected, which in turn leads to benefits in learning and transfer, and not remain confined to performance alone.

In recent years, scripting of collaborative interaction had been found to be beneficial especially in computer-mediated settings. Understanding how to encourage constructive processes like explanation through collaboration can help create better scripts for collaborative learning.

It is also important to better understand the social dynamics of collaborative learning. For example, what is the role of grounding in collaboration? In our present study, the participants had the required in the task. However, will we find the same effects if less skilled participants are given the same task? What amount of shared knowledge is necessary for successful collaborative learning? All these are open questions that future work needs to address.

In conclusion, we found a robust effect of collaborative error-correction such that collaborators showed better performance compared on a subsequent revision task, and performed as well as nominal dyads. This can have strong educational implications, ranging from applications to classrooms to computer-mediated learning environments.

Acknowledgements

We would like to thank the lab teaching fellows—Daniel Belenky and Jooyoung Jang for their assistance in collecting data for this experiment. Also, thanks to Melissa Patchan for her insightful suggestions during conceptualizing this study and Linnea Warren for help scoring data.

References

Andersson, J., & Rönnberg, J. (1995). Recall suffers from collaboration: Joint recall effects of friendship and task complexity. *Applied Cognitive Psychology*, *9*, 199-211.

Basden, B. H., Basden, D. R., Bryner, S., & Thomas, R. L., III (1997). A comparison of group and individual remembering: Does collaboration disrupt retrieval strategies? *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *23*(5), 1176-1189.

Coleman, E. B. (1998). Using explanatory knowledge during collaborative problem solving in science. *Journal of the Learning Sciences, 7*(3/4), 387-427.

Collaros, P. A., & Anderson, L. R. (1969). Effect of perceived expertness upon creativity of members of brainstorming groups. *Journal of Applied Psychology, 53,* 159-163.

Finlay, F., Hitch, G. J., & Meudell, P. R. (2000). Mutual inhibition in collaborative recall: Evidence for a retrieval-based account. *Journal of Experimental Psychology: Learning Memory And Cognition, 26*(6), 1556-1567.

Flower, L. (1979). Writer-based prose: A cognitive basis for problems in writing. *College English*, 19-37.

Hill, G. W. (1982). Group versus individual performance: Are n+ 1 heads better than one. *Psychological Bulletin, 91*(3), 517-539.

Karau, S. J., & Williams, K. D. (1993). Social loafing: A meta-analytic review and theoretical integration. *Journal of Personality and Social Psychology, 65,* 681-706.

Kelley, M. R., & Wright, D. B. (2010). Obtaining representative nominal groups. *Behavior Research Methods, 42*(1), 36-41.

Kerr, N. L., & Tindale, R. S. (2004). Group performance and decision-making. *Annual Review of Psychology, 55*(1), 623-655.

Mullen, B. (1983). Operationalizing the effect of the group on the individual: A self-attention perspective. *Journal of Experimental Social Psychology, 19,* 295-322.

Nokes, T. J., & VanLehn, K. (2008). Bridging principles and examples through analogy and explanation. In the *Proceedings of the 8th International Conference of the Learning Sciences*. Mahwah. NJ: Erlbaum.

Rajaram, S., & Pereira-Pasarin, L. P. (2007). Collaboration can improve individual recognition memory: Evidence from immediate and delayed tests. *Psychonomic Bulletin and Review, 14*(1), 95.

Shaw, M. E. (1932). A comparison of individuals and small groups in the rational solution of complex problems. *American Journal of Psychology, 44,* 491–504.

Sniezek, J. A., & Henry, R. A. (1989). Accuracy and confidence in group judgment. *Organizational Behavior and Human Decision Processes, 43,* 1–28.

Steiner, I. D. (1972). *Group processes and productivity.* New York: Academic Press.

Weldon, M. S., & Bellinger, K. D. (1997). Collective memory: Collaborative and individual processes in remembering. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23,* 1160-1175.

Weldon, M. S., Blair, C., & Huebsch, P. D. (2000). Group remembering: Does social loafing underlie collaborative inhibition? *Journal of Experimental Psychology: Learning, Memory, & Cognition, 26*(6), 1568-1577.

Wright, D. B., & Klumpp, A. (2004). Collaborative inhibition is due to the product, not the process, of recalling in groups. *Psychonomic Bulletin & Review, 11*(6), 1080-1083.

Appendix A

List of flaws in summary:
Structural level:
H1. Directly copied from text (plagiarized)
H2. Details of procedure not clear
H3. Gives actual statistics
H4. Does not explain results in plain language/ does not define terms
H5. References table that is absent in summary
H6. Hypothesis not stated
H7: Subject characteristics not present
H8: IV & DV not clear
H9: Experiment design (Between or within not clear)
H10: Limitations/ confounds not mentioned
H11: Does not interpret results/mention implications for further study
Surface level:
L1. Statistics not formatted correctly
L2. Reference absent
L3. Mentions five conditions instead of six
L4. Not indented
L5. Does not separate paragraphs
L6. APA formatting issues

Appendix B

Criteria for scoring revised summaries:

1. What is the research question?
2. What is the hypothesis being tested?
3. Were participant characteristics (number, age, gender, education etc.) correctly stated
4. Was the experimental task clear?
5. Was the experimental design (between or within subjects) correctly stated
6. Are the dependent variables correctly stated?
7. Are the independent variables correctly stated?
8. What were the important points of procedure
9. What were the major finding/s?
10. Are confounds/limitations pointed out?
11. Are findings interpreted in own language and a conclusion stated?
12. Mechanics (spelling, grammar) and Conciseness/ No unnecessary detail

A Double Causal Contrast Theory of Moral Intuitions in Trolley Dilemmas

Michael R. Waldmann (michael.waldmann@bio.uni-goettingen.de)
Alex Wiegmann (Alex.Wiegmann@psych.uni-goettingen.de)
Department of Psychology, University of Göttingen,
Gosslerstr. 14, 37073 Göttingen, Germany

Abstract

In trolley dilemmas a train is about to kill several victims who could be saved if instead a different victim is harmed. A number of theories have been proposed which assume that permissibility judgments in these harm-based moral dilemmas are mediated by an analysis of the underlying causal structure. For example, it has been postulated that it is permissible to harm people as a side effect but not as a means. We have developed a different causal theory which claims that moral judgments are influenced by two contrasts, the global contrast between the number of victims in the presence and absence of the act, and an additional local contrast that compares the fates of the morally relevant target (i.e., threats, victims) of the proposed intervention in the presence versus absence of the act. This double causal contrast theory explains intuitions in various types of trolley dilemmas better than its competitors.

Keywords: moral reasoning; trolley dilemmas; causal reasoning; doctrine of double effect

Introduction

Trolley dilemmas have become the drosophila for testing alternative philosophical and psychological theories of moral judgments in harm-based moral dilemmas (see Kamm, 2007). In the philosopher's Judith Thomson's (1986) version of the trolley dilemma, a situation is described in which a trolley whose brakes fail is about to run over five workmen who work on the tracks. However, the trolley could be redirected by a bystander on a side track where only one worker would be killed (bystander problem). Is it morally permissible for the bystander to throw the switch or is it better not to act and let fate run its course? Most people seem to have the intuition that throwing the switch is morally required or at least permissible. However, the intuitions change in another of Thomson's (1986) examples, in which the train could be stopped by throwing a fat person from a footbridge on the tracks, thus stopping the train with his body (footbridge dilemma). Most people find this act outrageous, even though again one person is sacrificed to save five. For philosophical theories these two intuitions present a puzzle. The intuitions in the bystander dilemma seem to be in line with utilitarian or consequentialist theories that focus on the favorable outcome of the act in contrast to not acting (1 vs. 5 dead people). However, the footbridge dilemma yields the same outcomes. The intuitions in this dilemma seem to be more consistent with non-consequentialist reasoning, which focuses on the impermissibility of the act of killing a person.

Not only in philosophy but also in psychology the trolley dilemmas have attracted interest as test cases for psychological theories of moral intuitions. Some have derided this research as trolleyology because of the artificiality of the task. It is certainly true that most people never will be in a situation that mimics the trolley problem. However, we would like to defend this paradigm as a valuable tool to study the cognitive basis of moral intuitions. People care about how society should deal with violent death, severe illness, terrorism, or emergency, even though they may never be involved in a dilemma involving these events. Nevertheless, these intuitions influence how our society and law functions. Thus, it is important to understand the mechanisms that underlie people's moral intuitions.

Threat vs. Victim Interventions

From a psychological point of view, the philosophical comparisons between bystander and footbridge trolley versions are flawed because of the various confounds. The footbridge dilemma differs in a number of relevant features from the bystander problem, including the act (re-directing a train vs. pushing a person), the physical distance between agent and victim, the directness, and the saliency of the death, or the degree of intentionality (see also Greene et al., 2009; Waldmann & Dieterich, 2007, for evidence). Unfortunately, in the early research on trolley dilemmas psychologists have often adopted close variants of Thomson's (1986) versions, which makes it hard to interpret the results of these studies (Greene et al., 2001; Mikhail, 2007). In our own research we have therefore tried to create variants of trolley dilemmas, which are better controlled so that some of the already well known factors affecting moral intuitions (e.g., distance, violence of act) are kept constant (Waldmann & Dieterich, 2007). We will first present a new, better controlled experiment which highlights the structural differences between different variants of trolley dilemmas. This experiment will serve as the base example for presenting competing theories, which then will be tested in additional experiments.

General Procedure Unless otherwise noted all experiments were run in groups (including seminars and lectures) with students from the University of Göttingen, Germany. Participants came from various fields, but we excluded philosophy and economics to avoid prior exposure to relevant philosophical positions. Subjects were handed booklets in which they were told that they are going to read about a situation which mentions two options of an agent in the story. All dilemmas used a format in which a fictitious agent in a remote control room of a train company is presented with two alternatives with outcomes, which lie in the future. The outcomes were clearly stated and characterized as cer-

tain. In the instructions it was pointed out that participants should carefully read the stories and attempt to empathize with the situation of the agent. The story was presented in a brief story that described the moral dilemma and the future options. Additionally, images were shown that presented the two options (acting vs. non acting)(see figures for examples). Subsequently, a rating scale was presented. Generally participants were asked to rate whether the agent should act or not in the described situation. The scale ranged from 1 ("not at all") to 6 ("definitely") with separated numbered boxes.

Experiment 1

In Experiment 1 we compared two parallel versions in which we manipulated the locus of intervention, threat versus victim. In the threat intervention condition the threatening train is redirected, in the victim intervention condition the train in which the single alternative victim is sitting is targeted. In both variants of the trolley problems all trains are moving and can only be redirected by employees of the train company who are sitting in a remote control room. The workers on the trains did not have any control over the trains. In the threat intervention condition (n=15)(Condition I), which corresponds to the bystander problem, five track workers sit on train A and one on train B. The empty train C, which represents the threat, is, due to a signaling defect, running behind train A and cannot be stopped. Soon it would hit train A with the five workers. However, the control room could throw the switch and redirect the train on the parallel track where it would hit train B. In both cases the victims would be seriously hurt (see Fig 1, I).

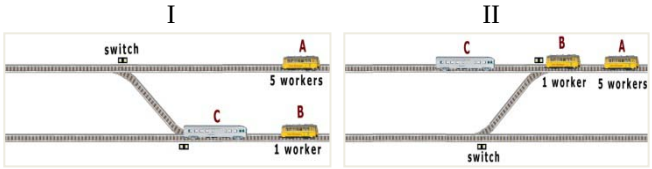


Fig. 1. Illustrations of the consequences of the proposed interventions in the threat (I) versus victim (II) intervention conditions in Experiment 1. In Condition I the threatening train C has been redirected to the side track, in Condition II train B with the victim has been redirected to the main track where it stops train C.

In the victim intervention condition (n=14)(Condition II), the first part of the story is identical. However, here the option is to redirect train B by throwing the switch on the parallel track. This way train B would go up to the track where the other two trains are running and would end up in between train C and train A (see Fig. 1, II). Now train C would hit train B which would stop the threatening train C. This would seriously hurt the one worker in train B, but save the five in train A. Consistent with the findings about bystander and footbridge dilemmas, the threat intervention option was rated more acceptable (M=4.93, SD=0.79) than the victim intervention option (M=2.57, SD=1.02), F(1, 27)=48.8, p=0.00.

Causal Theories of Moral Intuitions

How can the different moral assessments of threat and victim interventions be explained? We kept various familiar factors constant so that some simple accounts are ruled out. In both conditions the distance between the intervention and harm is roughly the same, the initial act (re-directing a train via remote control) is identical; in both cases the act only indirectly affects the fate of the victim, and there is no physical closeness or personal force. Moreover, none of the passengers has control over the train so that there are no differences in responsibility. What are the structural differences that may account for the different intuitions? Given that moral judgments are primarily about evaluating the moral quality of acts or interventions which lead to outcomes, causal theories seem to be a prime candidate for an analysis of such differences.

All theories that can be described as causal include the contrast between the outcomes in the presence versus absence of the intervention, and predict that the size of the contrast influences moral permissibility judgments. However, different theories postulate different representations of the acts and focus on different causal features.

Consequentialism Consequentialism is primarily interested in the contrast between the outcomes. Thus, consequentialist approaches choose a fairly abstract level of describing the acts as acting versus not acting, which blurs the differences between threat and victim interventions. This level of representation in both conditions yields the global outcome contrast between one dead person when the agent acts and five dead people when she refrains from acting (i.e., 1:5). Therefore, this theory predicts generally high acceptability ratings for the act. This may be acceptable as a normative principle (see Unger, 1996), but fails as a psychological account. The theory correctly predicts the intuitions in the threat intervention condition but makes wrong predictions for the victim intervention condition.

Doctrine of Double Effect Traditional non-consequentialist or deontological theories focus on moral rules permitting or prohibiting acts. For example, harmful acts, such as killing, are prohibited. However, simply prohibiting such acts also does not explain the intuitions in trolley dilemmas because apparently people find killing in the threat intervention condition acceptable. A more promising variant of a non-consequentialist theory accounting for trolley intuitions is the doctrine of double effect (DDE), an old deontological rule that is based on a causal analysis and also includes contrasts. A number of psychologists have proposed this rule as a moral heuristic (Royzman & Baron, 2002), or part of an innate moral grammar (Hauser, 2006; Mikhail, 2007). According to the dominant reading of the DDE it is permitted to do a neutral or good act as a means to a greater good, although we foresee lesser harm as a side effect, assuming that there are no better alternatives. However, it is impermissible to bring about lesser harm as an end in itself or as a means to a greater good. Thus, the DDE contains two stages: First a global favorable contrast needs to be ascertained ("greater good")(i.e., 1:5 in the trolley dilemmas). We know

already that this global contrast does not explain the effect, although it is certainly the case that the size of this contrast influences judgments (Nichols & Mallon, 2006). The main focus of the DDE is on the causal processes entailed by the proposed act. Here the doctrine distinguishes between two types of causal processes involving the single victim. If the victim is harmed as a side effect, as in the threat intervention condition, the act is permitted. However, if the victim is used as a means to save the five, as in the victim intervention condition, the act is prohibited. Thus, this rule explains the intuitions in the two conditions of Experiment 1. Importantly the DDE explains the different intuitions by analyzing the causal processes in the *presence* of the proposed intervention, whereas a contrast with events in the *absence* of the intervention does not play a role after the initial evaluation stage.

A Double Contrast Theory We are going to propose and test another variant of a causal contrast theory, our double contrast theory, which is an extension of Waldmann and Dieterich's (2007) proposal. Our main assumption is that subjects choose a level of abstraction of the act that brings out the specific causal characteristics of the proposed intervention. Contrasting the two interventions on the abstract level as presence or absence of acting or as killing and saving is too abstract because it does not reveal the differences between the scenarios. Using a very low-level description, such as button pressing on a remote control, also blurs the differences. We believe the most natural basic level description in the scenarios refers to the kind of intervention and the morally relevant target of the intervention. Morally relevant targets in trolley dilemmas are threats or victims, which can be stopped, redirected, derailed and so forth by the interventions. This is also the level of description that is used in the stories describing trolley dilemmas. For example, a natural description of the interventions in Experiment 1 might state that in Condition I the threatening train is redirected, whereas in Condition II the train with the single victim is set into motion towards the threatening train. Thus, in Condition I the threatening train is the target of intervention, whereas in Condition II the train with the single victim is the target of intervention.

Our main claim is that people will focus on the target of intervention and assess the harm directly caused by intervening in this target in contrast to the harm *the target* would cause in the absence of the intervention. This local contrast which focuses on the target of intervention rather than the global outcomes will, according to our theory, heavily influence the acceptability rating.

How does the double contrast theory explain the two standard dilemmas? In general, the morally relevant targets of intervention in our trolley dilemmas are either the trains which pose a threat, or the trains which house a potential victim. In the threat intervention condition (I) the proposed act can be summarized as re-directing the threat. Thus, the morally relevant target is the threatening trolley C. To assess the local contrast we need to focus on the direct harm caused by the target of intervention, train C, which is one seriously harmed person. This outcome is contrasted with the direct harm caused by the target of intervention (i.e., train C) in the absence of the intervention, which in Condition I are five people who are harmed by train C in the absence of an intervention. Thus, the local and global contrasts are the same in this case (1:5), both favoring the proposed intervention.

In contrast, in the victim intervention condition (II) the proposed act can be described as re-directing train B with its potential victim towards the threatening train C. Thus, train B with its potential victim is the target of intervention, and the local contrast will therefore focus on train B with its single potential victim. Setting this train into motion will directly cause harm to this victim. The fact that five people are saved further in the future is an indirect, more remote consequence of the act and therefore not part of the local contrast. To compute the local contrast the harm caused by the target of intervention in the absence of the act also needs to be considered. Train B with its single passenger, the target of intervention, would safely stay on the side track so that its passenger would not be harmed. Thus, the local contrast focusing on train B would amount to 1:0 (1 harmed vs. 0 harmed). The local contrast implies that the act is harmful, which predicts the lowered acceptability ratings.

As in the other theories we also believe that the global contrast (1:5) additionally plays a role, which explains why the ratings are not at a minimum. However, we assume that these global contrasts are backgrounded. In this regard, the double contrast theory makes similar assumptions as the DDE. But whereas the DDE explains differences of intuitions by focusing on the causal structure entailed by the acts, the double contrast theory focuses on the *contrast* of the fate of the target of intervention. In sum, both the double contrast and the doctrine of double effect explain the patterns in the standard trolley cases (e.g., Experiment 1).

Evidence for the Double Contrast Theory

In order to test our double contrast theory against its competitors we started to look for alternative versions of the trolley problem that better distinguish between the theories. In previous trolley research the target of intervention and the location of the alternative victim were often confounded. Whereas threat interventions typically redirect empty trains, victim interventions more directly intervene in the alternative victim. Other variants of the trolley problem allow us to disentangle these and other confounds, and provide informative tests for the alternative theories.

Experiment 2

In Experiment 2 we ran four conditions with 20 participants in each condition. Condition I is a standard threat intervention condition in which an empty threatening train can be redirected away from five victims towards one. All victims are sitting in trains, as in Experiment 1. As usual, this condition yielded relatively high mean ratings (M=4.6, SD=1.57), which signals high acceptance for the act. Our theory predicts this pattern as a consequence of the 1:5 contrast (see

above). Condition II is new (see Fig. 2): Here a passenger is sitting in the threatening train C. According to the instructions this passenger has no control over the train. The train is about to kill the five on the main track if nothing is done. However, in our instructions we stated that the passenger on the threatening train will be able to jump off the train before it crashes into the train with the five, and save himself. Thus, in the absence of an intervention five people would die, as in Condition I. Alternatively the threatening train could be redirected. Unfortunately, the train needs to be redirected to a side track which traverses a bridge. This bridge prevents the passenger on the threatening train from jumping off so that he will be killed in the collision between train C and the empty train B, which is parked on the side track behind the bridge. This is a novel condition because the intervention targets a threatening train which also transports a potential victim. Thus, this is a case of both a threat and a victim intervention. Interestingly, this condition descriptively received slightly (although not significantly) higher acceptability ratings (M=5.0, SD=1.3) than Condition I, which means that most subjects opted for sacrificing the one. Although in this condition a train with a single victim is the direct target of a harmful intervention, this variant of victim intervention is not aversive.

How does our theory explain this finding? According to the double contrast theory subjects will compute a local contrast on the morally relevant target of intervention. In both Conditions I and II the target is the threatening train C, which in one condition is empty and in the other houses a potential victim. In both conditions, train C directly harms one person in the presence of the intervention but harms five people in the absence of the intervention. Thus, both Conditions I and II yield the same 1:5 local (and simultaneously global) contrast, which favors acting.

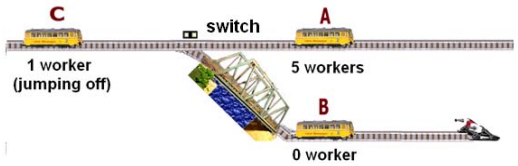


Fig. 2: Illustration of Condition II in Experiment 2 (see text for details).

We ran two more conditions. The most interesting condition of this experiment is Condition III. Here the threatening train C again carries a passenger who has no control over the train, and who is about to jump off (see Fig. 2). In the absence of the intervention, the five in train A at the end of the main track would be killed. On the side track an empty train B is parked, which could be directed upward toward the threatening train C. This empty train would stop the threatening train C on the main track but would kill its single passenger, who, according to the instructions, would not have sufficient time left to jump off. Note that killing the one with the empty train is on the causal path of preventing harm to the five. Thus, the train with its single passenger is used as a means to prevent harm from the five. Harming people and using them as means against their will should, according to the DDE, be aversive (see Experiment 3 for further discussions of the concept of means). In contrast to the predictions of DDE, however, we got again high acceptability ratings (M=4.7, SD=1.69), which in fact are statistically equivalent to the ones in the standard threat intervention condition (I).

Condition IV is a standard victim intervention condition, which serves as a control. An empty threatening train C is heading toward a train (A) with five passengers. At the end of a side track, which leads over a bridge, a train (B) with a single passenger is parked. This train B with its passenger can be set in motion in the direction of the main track where it would arrive in time to stop the threatening train C, however with fatal consequences for the single passenger. This condition yielded the expected low ratings (M=3.15, SD=2.01). In fact, these ratings proved significantly lower than the ratings in the three other conditions, $F(1, 76)=16.2$, p=0.00, which were not significantly different from each other.

How does our double contrast theory explain the difference between Condition III and the superficially similar standard victim intervention, Condition IV? Note that in both conditions the train that is parked on the side track is set in motion, and directed towards the threatening train on the main track. Thus, at first sight one might conclude that this empty train is in both scenarios the target of intervention. However, this is wrong according to our theory. In the victim intervention condition (III) the morally relevant target of intervention is indeed the train on the side track with its potential victim, who would either be killed or would stay alive. Thus, the local contrast favors inaction (1:0). However, although the act seems superficially similar in Condition IV, in this condition the train that is being moved is empty. Thus, it neither represents a threat nor is a victim located inside the train. This train is therefore not a morally relevant target of the intervention; it rather plays the causal role of an instrument to stop the threatening train. In this regard the empty train is similar to other morally irrelevant instruments, such as the remote control or button presses. As a consequence, the threatening train C, not the empty train B is the morally relevant target of the intervention in Condition IV. Computing the local contrast over the harmful outcomes train C is causing in the presence versus absence of the intervention yields a 1:5 local (and global) contrast, which favors the intervention. In sum, the results of the experiment favor our double contrast theory over the DDE and related principles (Kamm, 2007).

Experiment 3

In Experiment 3 we ran different variants of some of the conditions in Experiment 2 along with new conditions. This experiment provides further tests of the DDE and our double contrast theory. Again we used the standard trolley instruction about a threatening train on a test site which, due to a brake failure, is about to hit a train with five track workers

at the end of the main track. These five workers would be killed. As in Experiment 1 there is also a parallel side track, which is connected to the main track via a connecting track (see Fig. 3). We ran four conditions. As in the other experiments the passengers inside the trains had no control over the trains and therefore were not responsible for the outcomes in all conditions. In both Conditions I and II we placed the single victim inside the threatening train B in a safe location in the rear of the train. Thus, unlike in the last experiment the passenger does not need to jump off the train to save himself. Doing nothing leads to the death of the five in train A at the end of the main track, but would spare the passenger in the safe location inside the threatening train B. In Condition I (n=58) , the instructions propose as an alternative that the agents in the remote control station could redirect an empty train C located on the parallel side track up to the main track, thus hitting the threatening train in the rear section and thereby leading to the death of the single passenger (see Fig. 3). However, the threatening train B would be derailed saving the five. This scenario led to fairly high ratings (M=4.4, SD=1.28).

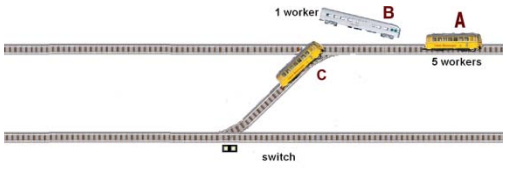


Fig. 3: Illustration of consequence of proposed act in Conditions I and II of Experiment 3 (see text for details).

Condition II was similar. However, to provide a clear cut case against the DDE, we made the role of the victim as a means more salient. Technically one could argue that in all our experiments the means of saving the five were the redirected trains, not the bodies of the passengers, whose deaths could be construed as side effects. However, such an argument would not save the DDE because then the difference between the threat and victim intervention in Experiment 1, for example, would be a puzzle. Moreover, we doubt that people would construe their harm as a side effect if they were sitting in a vehicle that is being used without their consent to save others (see also Kamm, 2007). Anyhow, in Condition II (n=54) we stated again that a single passenger in train C, who is unfamiliar with the steering and brake system, is sitting in the rear of the train in a safe location. Now employees in the control station, guided by a camera inside the train, notice that by hitting the train, the passenger would fortuitously be pushed against the brake system, which would lead to a derailment of the train. The passenger would be killed by this act but the five would be saved. In this instruction the body of the victim is clearly specified as a necessary means for the goal to derail the train and save the five. Interestingly, similarly high ratings as in Condition I were obtained (M=4.81, SD=1.04). In fact, descriptively these were the highest ratings in this experiment. Clearly participants were not sensitive to whether the body of the victim was causally necessary for saving the five or not.

Both Conditions I and II refute the DDE as a theoretical account. Although in both conditions the single victim was used as a means to save the five, subjects found the intervention highly acceptable. This finding is explained by the double contrast theory. As in Condition III in Experiment 2, in Conditions I and II of Experiment 3 the empty train C plays the role of an instrument, the morally relevant target of intervention is train B, which both constitutes a threat and houses a potential victim. In the presence of the intervention train B, the target of intervention, is involved in the death of one victim while in the absence of the intervention the five passengers in train A die. Thus, this conditions leads to a 1:5 local and global contrast.

To ascertain that the high ratings in Conditions I and II are indeed different from predictably aversive conditions, we also ran Condition III as a control, which is the standard victim intervention condition (n=49). In this condition train C on the side track which transports a single passenger is redirected through the connecting track to the main track where the train would hit and derail the empty threatening train B, thus leading to the death of the one in train C, but saving the five in train A. As usual, this intervention was given fairly low ratings (M=3.76, SD=1.64), which is predicted by our theory as a result of the 1:0 local contrast.

Finally, in Condition IV (n=51), a fourth train D in which one worker is sitting was introduced which is parked on the connecting track, thus blocking the way to the main track. The proposed intervention was to send an empty train C located on the parallel side track up the connecting train, thus derailing train D on the connecting track, and thereby killing its passenger. After stating this fact, the instruction mentioned that this event will open up the way to the main track where train C from the side track could derail the empty threatening train B on the main track, thus saving the five in train A. This intervention also yields fairly low ratings (M=3.88, SD=1.37). How does our theory explain the finding in Condition IV? The initial morally relevant target of intervention in this condition is train D, which is parked with its potential victim on the connecting track. This victim dies in the presence but would be alive in the absence of the intervention, thus creating a 1:0 local contrast.

The general pattern is confirmed by an ANOVA: Conditions I and I, which are statistically equivalent, yielded significantly higher acceptability ratings than Conditions III and IV, $F(1, 208)=11.30$, $p<0.001$.

General Discussion

The goal of our studies was to test theories of moral acceptability in harm-based moral dilemmas. Certainly there are other types of moral problems which might require different theories (Haidt, 2007). Trolley dilemmas represent interesting test cases for cognitive theories because they show that our moral intuitions are influenced by structural factors which go beyond simple comparisons between outcomes (e.g., numbers of victims) or acts (e.g., killing, saving).

Despite identical outcomes and the identical conflict between saving and harming, our moral intuitions differ depending on various factors including the kind of act, distance, intention, contact, legal responsibility, personal force, or the framing of the outcomes (e.g., Greene et al., 2009; Rai & Holyoak, 2010). In our studies we tried to control for these already known factors in order to focus on the remaining structural causal differences between types of scenarios, which pose a puzzle for both psychologists and philosophers.

A number of moral theories focus on causal structures and are therefore candidates for explaining effects of such structural differences. These theories differ in the choice of the level of description and in the postulated relevant causal features. *Consequentialism* focuses on outcomes, and therefore uses abstract descriptions of acts. The moral analysis contrasts global outcomes in the presence and absence of the act. This theory fails as a psychological account.

A second causal account, the non-consequentialist *doctrine of double effect* also tests for a favorable global contrast first, but then focuses on the causal paths entailed by the act under consideration. Here the distinction between harming people as a means versus as a side effect carries most of the weight in explaining differences in intuitions in trolley dilemmas.

A third theory, our *double contrast theory*, also starts by considering the global contrast. But then a local contrast is computed using basic level descriptions of the interventions targeting threats or victims. For example, in the victim intervention conditions people represent the intervention as redirecting the victim, and consider what will happen to this victim in the presence versus absence of the proposed act.

Three experiments have shown that the double contrast theory wins over the doctrine of double effect. People clearly find it acceptable to use people as means without their consent when the local contrast favors the act.

Directions for Future Research

More research is needed on how people choose the level of description in moral dilemmas. It would be interesting to present subjects with still movies, and have them describe the scenarios in moral and non-moral settings.

Another interesting goal would be to further explore the factors influencing local contrasts. In our experiments we have chosen interventions in which the acts were morally innocuous (e.g., throwing a switch). In the contrast between re-directing a victim and not re-directing the victim, the morally relevant contrast is surely about what happens to the victim. However, if the intervention was shooting a victim versus not shooting her, the contrast between shooting and not shooting would certainly impact on the moral evaluation of the contrast. A clear example of this case is, for example, the famous *Jim and the Indians* dilemma, in which Jim is given the choice of watching twenty Indians be shot or shoot one of these twenty Indians himself, thus saving the rest (Williams, 1973). Although the local contrast for the Indian, Jim could shoot, would be 1:1 (he is dead regardless of the act), the act is certainly aversive because of the shooting component of the contrast shooting the Indian vs. not shooting the Indian.

Finally it would be interesting to get a more quantitative assessment of the relative weight between global and local contrasts. Global contrasts surely affect moral assessments, as can easily be seen if we consider a 1:1.000.000 contrast in a disaster variant of a trolley problem (Nichols & Mallon, 2006). Note that none of the previous theories includes assumptions about how global contrasts quantitatively affect judgments because moral philosophers typically ask about permissibility, not about degree of permissibility. Our experiments clearly suggest that local contrasts dominate judgments but they do not allow us to answer the question how much weight these contrasts have relative to the global contrast.

References

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science, 293*, 2105-2108.

Greene, J. D., Cushman, F. A., Stewart, L. E., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2009). Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition, 111*, 364-371.

Haidt, J. (2007). The new synthesis in moral psychology. *Science, 316*, 998-1002.

Hauser, M. D. (2006). *Moral minds: How nature designed our universal sense of right and wrong*. New York: Ecco Press.

Kamm, F. M. (2007). *Intricate ethics: Rights, responsibilities, and permissible harm*. Oxford: Oxford University Press.

Mikhail, J. (2007). Universal moral grammar: Theory, evidence, and the future. *Trends in Cognitive Sciences, 11*, 143-152.

Nichols, S., & Mallon, R. (2006). Moral dilemmas and moral rules. *Cognition, 100*, 530-542.

Rai, T. S., & Holyoak, K. J. (2010). Moral principles or consumer preferences? Alternative framings of the trolley problem. *Cognitive Science, 34*, 311-321.

Royzman, E. B., & Baron, J. (2002). The preference for indirect harm. *Social Justice Research, 15*, 165-184.

Thomson, J. J. (1986). The trolley problem. In J. J. Thomson, *Rights, restitution, and risk. Essays in moral theory* (pp. 94-116). Cambridge, MA: Harvard University Press.

Unger, P. (1996). *Living high and letting die.* New York: Oxford University Press.

Waldmann, M. R., & Dieterich, J. (2007). Throwing a bomb on a person versus throwing a person on a bomb: Intervention myopia in moral intuitions. *Psychological Science, 18*, 247-253.

Williams, B. (1973). A critique of utilitarianism. In J. J. C. Smart & B. Williams, *Utilitarianism: For and against* (pp. 82-117). Cambridge: University Press.

Deconfounding Distance Effects in Moral Reasoning

Jonas Nagel (jnagel1@uni-goettingen.de)
Michael R. Waldmann (michael.waldmann@bio.uni-goettingen.de)
Department of Psychology, University of Göttingen, Germany

Abstract

A central question of moral philosophy and moral psychology is whether spatial distance is morally relevant (Kamm, 2007). Does spatial distance reduce our sense of obligation to help strangers in great need? One problem of assessing this question is that distance between agent and victim is typically confounded with other factors, such as saliency of the victim's need, costs for the agent, or location of the agent's means. The goal of our experiments is to find out whether spatial distance *per se* matters in people's intuitions. Whereas the first two experiments seem to indicate that spatial distance between the agent and the victim or between the agent's means and the victim affect subjects' intuitions, Experiment 3 and a closer look at Experiment 2 both reveal that the assumed distance effects disappear if the compared cases are properly deconfounded. Implications of these findings for theories of psychological distance are discussed.

Keywords: moral reasoning; moral intuitions; distance; obligation to help; human experimentation

Introduction

The present research aims at exploring the role of spatial distance in moral judgments: Does spatial distance reduce our sense of obligation to help strangers in great need? The normative relevance of this factor has been heavily disputed in philosophy. Thus, we will set out by first reviewing some of the philosophical debate about whether distance *ought to* matter morally. The aim of this section will not be to contribute to this normative issue, but instead to motivate our empirical investigation and to introduce the thought experiments on which our experimental materials are based. Unlike philosophers we do not want to address the question whether spatial distance ought to matter, but rather aim at finding out whether spatial distance *per se* is psychologically relevant in moral judgments. Alternatively, distance may only appear to be descriptively relevant due to factors with which it is typically confounded. After a brief discussion of relevant empirical work in psychology, we will report three experiments which explore whether our sense of obligation to help strangers is affected by distance per se. In the concluding section, we briefly discuss the implications of our findings for theories of psychological distance.

Distance and the Obligation to Help in Philosophy

In his famous article *Famine, Affluence, and Morality*, the philosopher Singer (1972) argues for an intuitive moral principle: "If it is in our power to prevent something bad from happening, without thereby sacrificing anything of comparable moral importance, we ought, morally, to do it" (p. 231). In a case example supposed to illustrate this principle, a child is drowning in a shallow pond. Intuitively, a person walking past this pond has a strong obligation to rescue the child, even if this means that she will spoil her clothes. Singer then argues that there is no justification to mitigate this principle on the grounds of increased distance between the victim and the potential agent, for such reasoning would clash with "any principle of impartiality, universalizability, [or] equality" (p. 232). Therefore he believes that we are obligated to help distant strangers as much as physically close strangers, for example by donating a good proportion of our assets to the needy.

The philosopher Unger (1996) agrees with this conclusion, and similarly does not view distance as a normatively relevant factor. For him, physical proximity is merely a factor increasing the conspicuousness of a victim's needs to a potential agent. This conspicuousness, while responsible for our increased urge to help near as opposed to far strangers, is not itself given any moral weight by Unger. He supports this view by contrasting several versions of two cases, *The Vintage Sedan* and *The Envelope*. In *Sedan*, the agent refuses to pick up a man with a self-inflicted injury and to drive him to a hospital, because he fears that the victim's blood will spoil the leather-seating of his car, leading to a $5000 damage. As a consequence, the victim loses a leg. In *Envelope*, the agent refuses to respond to a letter from UNICEF which informed him that 30 children could be saved from death if he sent in a check for $100. As a consequence, 30 more children lose their lives than if the agent had donated the money. According to Unger, our intuitions tell us that the agent's behavior is severely wrong in *Sedan*, but not so much in *Envelope*, although there are many features suggesting that the *Envelope*'s behavior is actually much worse (e.g., more victims, each of them suffering a greater loss, smaller costs for the agent, etc.).

To show that the difference in our intuitions between these cases is not primarily grounded in physical proximity, Unger (1996) then discusses both a version of *Sedan* in which physical distance is increased (*The CB Radios*, in which the agent is informed via a radio in his car about the victim's bad condition while he is ten miles away from him), and a version of *Envelope* in which distance is decreased (*The Bungalow Compound*, in which the agent receives the UNICEF mail while he is on holiday, and the children are suffering in his immediate neighborhood). Unger's intuitions (which can in our view be debated) is that we condemn the agent's behavior in *CB Radios* as strongly as in *Sedan*, and that we judge his behavior in the *Bungalow* as leniently as in the *Envelope*. Therefore, our diverging intuitions toward *Sedan* and *Envelope* cannot be accounted

for by the difference of physical distance between agent and victim.

Recently, the philosopher Kamm (2007)[1], who in contrast to Singer and Unger endorses a nonconsequentialist ethical position, has presented a different view on these matters. Part of her argument against Unger's (1996) claims is as follows: If one wants to show that distance per se *never* matters morally, it does *not* suffice to provide a couple of sets of cases in which it does not matter morally, for there might be different equalized contexts in which it does. For example, in both the *Envelope* and the *Bungalow*, the children's bad condition is caused by a lack of basic social justice, and it might be that an individual's obligation to help in such cases is not tracked by distance. However, this does not imply that the same holds true for cases involving accidents, for example. On the flipside, Kamm argues, if one wants to show that distance per se *does* matter morally, it suffices to provide one single set of perfectly equalized cases in which it does. Her approximation of such a set of cases (with the contrast case in parentheses) is as follows:

> Near [*Far*] Alone Case. I am walking past a pond in a foreign country that I am visiting. I alone see many children drowning in it, and I alone can save one of them. [*I alone know that in a distant part of a foreign country that I am visiting, many children are drowning, and I alone can save one of them.*] To save the one, I must put the $500 I have in my pocket into a machine that then triggers (via electric current) rescue machinery that will certainly scoop him out. (p. 348)

Kamm's intuition is that she has a stronger obligation to the child in *Near Alone* than to the child in *Far Alone*. As she notes, in this set of cases most of the factors normally confounded with distance are held constant. Among them are, for example, the numbers of victims, the seriousness of their suffering and how it came about. Further factors are the costs for the agent and the way in which they arise, as well as the agent's means of helping and their probability of success. Moreover, the number of potential alternative helpers typically increases with distance. Because all these confounded factors are held constant, Kamm believes that spatial distance alone is responsible for the difference in our sense of moral obligation between *Near Alone* and *Far Alone*.

In summary, the question of whether we *ought to* help needy strangers who are near us more than those who are far is controversial among philosophers. In the current research we are more interested in the question whether spatial distance per se affects intuitive judgments of laypeople if, like in Kamm's cases, potentially confounded variables are controlled. Surely, the intuition that we have a greater responsibility to take care of what is going on near us rather than far from us is shared by most people. But why is this? Is this intuition entirely explainable in terms of distinct, confounded factors like conspicuousness of need, as Unger (1996) claims? Or does distance possess some moral weight of its own in our intuitive judgments, even if all confounding factors are controlled?

[1] All following references to Kamm refer to this volume.

Distance and Obligation to Help in Psychology

Before we present our experiments, we want to take a look at previous relevant research in psychology. We are primarily interested in the determinants of moral intuitions rather than in what people actually do. Of course, there is an enormous amount of social psychological studies on determinants of actual (im)moral behavior, some of which also involve investigations of distance effects (e.g., Milgram, 1965). However, such behavior is obviously determined by many more factors than moral judgment alone (e.g., Latané & Darley, 1970).

To our knowledge, only a few studies have directly investigated the influence of distance on people's sense of obligation to help. One study is by Gillis and Hagan (1983), in which participants reported that they were more likely to intervene to prevent criminal behavior if the incident occurred close to their own home as opposed to a distant part of their hometown. In this case, distance refers to the proximity of a threat to the center of an agent's territory, whereas the distance between agent, threat, and victim at the time of the incident is constantly small. Hence, the results indicate that some types of spatial distance may influence people's sense of obligation.

Levine and Thompson (2004) presented a British sample of participants with two scenarios describing the aftermath of a natural disaster. One was about an earthquake in Eastern Europe, the other about a flood in South America. Additionally, the instructions highlighted for half of the participants their British identity, whereas for the other half their identity as Europeans was emphasized. Participants responded to be more likely to offer financial help as well as political engagement if the disaster happened in Europe rather than in America. However, this main effect was qualified by an interaction with the highlighted identity: The difference was greater when the European identity was salient, in which case the comparison between Eastern Europe and South America involved an ingroup/outgroup contrast. For this reason, Levine and Thompson (2004) argue that social categorization of the self relative to the victims rather than absolute geographical distance between them crucially affects whether people feel obligated to help. Note, however, that the distance between agent and victims, while differing in relative terms, is very large in both location conditions. Thus, these results do not rule out that distance effects could be found if the contrast involved one case in which the victim is near the agent in absolute terms and one case in which she is far. As Kamm argues, it might be really spatial *proximity* or absolute nearness which makes a moral difference, rather than any difference in relative distance.

Finally, Baron and Miller (2000) explored how people deal with the fact that, in principle, they have an unlimited amount of opportunities to help others in great need at little costs to themselves. They considered several factors that people might use to limit the scope of their positive duties, among them spatial distance. They found in both an American and an Indian sample that people find it more

wrong that an agent does not donate bone marrow to a sick patient if this patient lives in the same town as opposed to on the other side of the world. Moreover, significantly more subjects feel that the agent has a responsibility to donate in the near rather than in the far condition. Whereas the contrast in this study contains a genuine difference of proximity between agent and victim, it is again confounded with a difference in shared group membership. In fact, Baron and Miller (2000) themselves explicitly make the ingroup/outgroup contrast accountable for the distance effect they found.

In sum, there is some evidence compatible with the hypothesis that spatial distance might play a role when people consider whether they ought to help needy others. However, there is no previous study that thoroughly deconfounded distance from other factors naturally varying with distance, such as group membership. Moreover, in all studies reviewed so far the distance factor was varied within subjects only. Since people had to compare cases that were otherwise very similar, the salience of the varied factor was probably artificially increased. Thus, demand characteristics may have distorted the results. While the within-subjects component is actually typical for the setting in which philosophers usually form their intuitions (see above), we believe a stronger empirical case for a true influence of spatial distance on laypersons' moral intuitions could be made if effects were found in a properly deconfounded between-subjects design.

Experiment 1

We take Kamm's *Near Alone* and *Far Alone* cases as a starting point for our investigation. As noted above, these cases are equalized in many important respects. Consequently, confounds contained in previous studies can largely be avoided. Moreover, past research (Miller, Bersoff, & Harwood, 1990) has shown that members of different cultures unanimously regard helping strangers in life-threatening situations as a genuine moral obligation rather than as a matter of social convention or personal choice. That is, helping in such cases is considered both as an "objective" duty (i.e., existing not just because of a law) and as legitimately regulated by society. This indicates that most subjects will evaluate the selected cases in moral terms. If a lack of mere spatial proximity between agent and victim *decreases* people's sense of obligation to help, subjects should judge the agent's obligation in *Far Alone* to be somewhat lower than in *Near Alone*. Experiment 1 tests this hypothesis. Note that, since both cases involve an action that is generally considered to be driven by a strong moral duty, a small effect size near the ceiling is to be expected if distance turns out to be relevant.

Method

Participants 62 Göttingen University students (48 women) with a mean age of 23 years participated voluntarily. The experiment was conducted either in a class room before a lecture, or subjects were individually approached on campus.

Design, materials, and procedure Each participant individually filled out a questionnaire consisting of two pages. The first page contained general instructions explaining the task and asking the participant to try to empathize with the scenario's agent, even though, for methodological reasons, the scenario content would not be realistic. After turning the page, half of the participants (n=31) read a variant of *Near Alone*, the other half (n=31) a variant of *Far Alone*. The wording of both cases was kept as close as possible to Kamm's original formulation (see above), but we decided to include the description of a mechanism by which the agent could possibly have learned about the victims in *Far Alone*. The text of our Near [*Far*] case was as follows (translated from German):

> You are on holiday in a foreign country. There, you take a walk past a pond. You alone see many children drowning in it [*While you take a walk there, you alone learn via an information service on your cell phone that many children are drowning in a pond situated in a distant part of the country*], and you alone can save one of them. To save the one, you must put the €500 you have in your pocket into a machine that accidentally is situated right next to you. This machine then triggers a remote-controlled rescue machine in the pond which will definitely pull one of the children out of the water and save her life. There is no other possibility to save one or more of the children.

This case description was followed by an assessment of the participants' sense of obligation to help. The wording of the question was: "How strongly do you feel obligated to put your €500 into the machine in order to save one of the children?," highlighting both consequences and costs of the action. Finally, participants were asked to indicate their judgment on a 6-point rating scale, labeled "not at all" at the left-hand end (1) and "very strongly" at the right-hand end (6).

Results

The mean rating for sense of obligation was 5.61 (SD=.67) in the Near condition, and 4.97 (SD=1.22) in the Far condition. This difference was statistically significant ($t[df_{corr}=46.37]=2.58$, one-tailed, $p<.01$, $d=.65$).[2]

Discussion

Our participants seem to share Kamm's intuitions regarding the *Near Alone* and *Far Alone* cases. Even though, as expected, ratings were very high in both conditions, participants reported a higher sense of obligation to rescue a child drowning near them rather than far from them. This effect cannot be accounted for by most confounds usually associated with spatial distance, such as social distance, number of potential saviors, urgency, probability of success, or type, and size of costs for the agent.

[2] In none of the experiments there was a significant effect of sex on sense of obligation to help. Therefore, this factor is excluded from all analyses.

This result encouraged us to test further factors proposed by Kamm. In particular, an important claim she makes is that not only proximity between agent and victim might be of moral importance, but also proximity between the victim and the agent's items that are efficacious in helping the victim. In other words, Kamm's intuition is that an agent is more strongly obligated to let his means be used if they are situated near the victim rather than if they are far, even if he is far from the victim himself in both cases. As an example she uses drowning scenarios in which the distance between agent and victim is kept constantly high, but the distance of the means of saving, a boat the agent owns, is either near or far the victim. Kamm's intuition is that this distance is morally relevant.

Experiment 2

In Experiment 2 we seek to test the hypothesis that proximity between an agent's means and the victim increases the agent's sense of obligation to help, even if the agent himself is constantly far away from the victim. We construed a scenario in which an agent has the opportunity to donate money in order to save sick children in Kenya from early death. His means to this end is money on a bank account which is either located close to him but far from the victims (in Göttingen, Germany) or close to the victims (in Kenya). Additionally, we anticipated that subjects might infer that they are in some way involved with Kenya from the fact that their money is there already. To control for this obvious confound, we decided to include previous personal involvement with Kenya as an additional independent variable.

Method

Participants 80 Göttingen University students (48 women) with a mean age of 24 years participated voluntarily after being approached individually on campus.

Design, materials, and procedure The two independent variables yielded a 2 (distance between victim and agent's means: Near vs. Far) × 2 (involvement: High vs. Low) between-subjects design (each n=20). The questionnaires had the same format as in Experiment 1. The case vignette in the Near/High [*Near/Low*] condition read as follows (translated from German):

> A couple of years ago, you have opened a bank account in Kenya while you spent your holidays there [*because you found out about the high interest rates there*]. Since then, you have returned there a couple of times. This is why you are still maintaining this account today. [*Since this proved of value, you are still maintaining this account today. Neither have you ever been to Kenya yourself, nor have you had any other connection to that country.*]
>
> One day, while you are in Göttingen, you hear in the news that several children in Kenya have been infected with a rapidly progressing disease. If these children do not receive medical treatment, they will die within the next few days. However, there is a lack of money for the urgently needed treatment. You could effectively contribute to saving the children by transferring €30 via internet from your Kenyan bank account to a local donation account.

The respective Far conditions were identical, except that the agent's bank account was located in Göttingen. Sense of obligation was assessed using the same scale as in Experiment 1. The wording of the question was: "How strongly do you feel obligated to perform the proposed action?"

Results

The results are summarized in Figure 1. A two-way ANOVA revealed a main effect of involvement on sense of obligation to help (F[1,76]=11.25; p<.01; η_p^2=.13), indicating that stronger previous involvement with Kenya led participants to report that they feel more strongly obligated to donate the money. Moreover, there was a main effect of distance (F[1,76]=4.31; p<.05; η_p^2=.05), showing that participants reported feeling more strongly obligated to help if their bank account was in Kenya. The interaction between both independent variables was not significant (F[1,76]=1.25; p=.27).

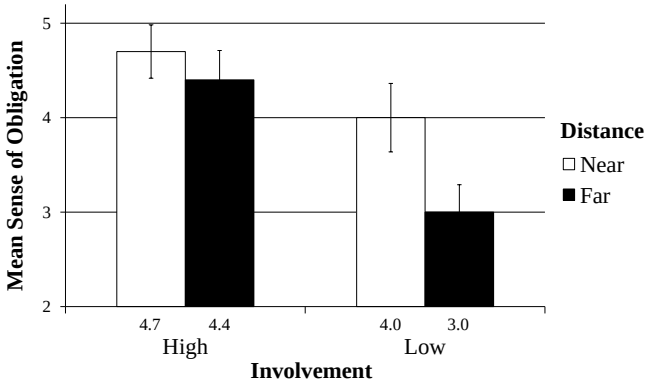


Figure 1: Mean ratings of sense of obligation in Experiment 2. Error bars indicate SEM.

Discussion

As expected, previous personal involvement with the home country of the children in need heavily increased respondents' sense of obligation to help. More interestingly, we found an independent effect of the location of the bank account: Participants felt more obligated to help if the means by which they could do so were already located close to the victims, even in the case in which the agent had never visited the country and only had opened an account because of the favorable interest rate. The effect of the spatial distance of the means is especially interesting since the action required to help (i.e., instructing a transfer of funds via internet) is virtually identical in both conditions. Moreover, the location of the means is actually merely symbolic in this scenario, since the agent's money does not have real physical presence either in Kenya or in Göttingen prior to being withdrawn from a cash machine. Still, it seems that even under these conditions participants share Kamm's intuition that proximity of means increases obligation to help.

So far, Kamm's propositions about the impact of mere spatial distance on moral intuitions are mirrored in our

participants' judgments in both experiments. However, despite all the effort invested in making the cases maximally parallel, both experiments still contain remaining confounds. In particular, it cannot be ruled out that in the Near/Low condition in Experiment 2, the knowledge that the agent is somehow profiting from the Kenyan financial system is the source of increased obligation, even if all other personal involvement is explicitly ruled out. In fact, in informal discussions with participants who had completed this condition, quite a few of them spontaneously mentioned that such considerations had influenced their judgment. Moreover, even in the more tightly controlled and therefore more artificial cases of Experiment 1 there is a remaining potentially relevant confound (see also Kamm): In the Near case, the agent directly sees the drowning children with her own eyes, whereas in the Far case the information is mediated by an electronic device. Therefore, in Experiment 3, apart from trying to replicate the results from Experiment 1, we aim to go one step further and try to experimentally control for informational directness in order to find out whether an independent effect of spatial distance can still be found.

Experiment 3

Experiment 3 aims at replicating the results of Experiment 1 while controlling for the previously confounded factor of informational directness. Additionally, we seek to find out whether a distance effect can be found if there are no considerable costs to the agent. Kamm's intuition is that distance does not matter under such conditions of no cost: If all I need to do to save someone's life is to pull a switch, I ought to do so regardless of whether the victim is near me or not.

Method

Participants 240 Göttingen University students (133 women) with a mean age of 24 years participated voluntarily after being approached individually on campus.

Design, materials, and procedure We orthogonally manipulated three independent variables, yielding a 2 (distance between agent and victim: Near vs. Far) × 2 (informational directness: Direct vs. Mediated) × 2 (costs: Zero vs. High) between-subjects design (each n=30). The case vignettes were closely matched to Near and Far in Experiment 1, but to control for informational directness we made some changes. To be able to construe a case in which the agent has direct information despite large physical distance (by means of binoculars), we decided to move the victims somewhat closer to the agent, so that now the distance was about five kilometers in all Far conditions. In all Mediated cases, the information was again transmitted via cell phone in the form of a video (to keep the visual modality constant). In the Near/Mediated conditions, there was a high wall between agent and victims to avoid direct visual contact. Moreover, the pond was replaced by a thunderous river in all conditions to prevent participants in this condition from assuming that the agent could hear the children screaming. In Near/Direct, we included a fence instead of a wall to make sure that participants in this condition would not believe the agent could simply jump into the river. Finally, in all Zero cost conditions, the action no longer consisted of putting money (in the costly case €300, being closer to Kamm's $500 in terms of actual worth) into the machine, but rather of pulling a switch. Sense of obligation was assessed using the same scale and wording of question as in Experiment 2.

Results

The results are summarized in Figure 2. In order to test whether the results of Experiment 1 could be replicated, we first conducted a planned contrast between the conditions Near/Direct/High and Far/Mediated/High, which correspond to Near and Far in Experiment 1. This contrast was significant (t[232]=2.41, one-tailed, p<.01, d=.62)[3] and the respective means were almost identical with those obtained in Experiment 1, thus neatly replicating its results.

Afterwards, we conducted a three-way ANOVA which revealed a main effect of costs on sense of obligation to help (F[1,232]=15.77; p<.001, η_p^2=.06), indicating that participants reported feeling more strongly obligated to help at zero costs than at high costs. Moreover, there was a main effect of informational directness (F[1,232]=4.53; p<.05; η_p^2=.02), showing that participants reported feeling more strongly obligated to help if they witnessed the incident with their own eyes. Crucially, there was no main effect of distance (F[1,232]<1), nor were any of the interactions between the three independent variables statistically significant.

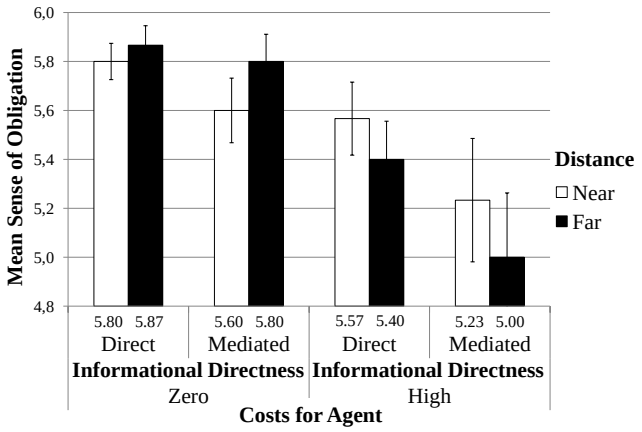


Figure 2: Mean ratings of sense of obligation in Experiment 3. Error bars indicate SEM.

Discussion

The findings from this experiment indicate that the assumed distance effect from Experiment 1 can be attributed to an effect of informational directness. If directness is kept constant, distance does not have an effect anymore. The fact

[3] This difference remains significant if the t-test is based exclusively on the two compared groups and corrected for their unequal variances (t[df_{corr}=45.95]=1.88, one-tailed, p<.05, d=.48).

that we were able to exactly replicate the mean ratings of Experiment 1 in the corresponding conditions of Experiment 3 makes it seem unlikely that any of the small changes we introduced to the material (e.g., slightly lower costs, video-mediated information, more abstract question wording) is responsible for the absence of a distance effect. Thus, we are confident that informational directness also caused the effect in Experiment 1.

The strongest effect we found was that our subjects' sense of obligation to help was lowered when there were substantial costs for the agent. None of our subjects may have actually believed that a child's life is not worth $300. Rather, some of them may have felt that a shady machine taking $300 to rescue a child is itself immoral. More interestingly, we did not find an interaction of costs and distance, as Kamm would have predicted. Distance indeed did not affect the ratings when there weren't any costs for the agent, but neither did it when there were.

General Discussion

Kamm supports her claim that distance per se matters morally with her intuitions regarding her sense of obligation to help needy strangers in well equalized scenarios. Experiment 1 showed that laypersons share her intuitions on one of her central set of cases. Experiment 2 indicated, as Kamm has proposed, that not only distance between agent and victim, but also distance between an agent's means and the victim may affect our moral intuitions. It seems likely, though, that distance effects here were mediated by assumptions about different amounts of social responsibility. The interpretation that distance effects may be generally reducible to other confounded factors is bolstered by Experiment 3, which additionally controlled for informational directness. This experiment revealed that the assumed distance effect from Experiment 1 disappears if the compared cases are properly deconfounded. Thus, while we find that our participants' responses to specific cases are largely in line with Kamm's intuitions, we also find, contrary to what Kamm argues, that these intuitions are informed by factors typically confounded with distance rather than by distance per se. In this sense, our data are more in line with Unger's (1996) behavioral predictions. Moreover, they align nicely with recent findings by Greene et al. (2009) who did not find spatial distance to influence judgments of moral dilemmas when this factor was carefully separated from related factors such as personal force or physical contact.

This pattern of results indicates that a purely spatial notion of distance does not seem to affect moral judgment of laypersons. That, of course, is not to say that what is commonly experienced as spatial distance in everyday life does not influence people's moral judgments. In fact, as Experiment 2 has shown, in naturalistic settings, people are sensitive even to very subtle manipulations of distance. However, psychologically relevant distance seems to be a broad concept naturally enriched with many covariates, such as informational directness or personal involvement. The difficulty of isolating spatial distance from its typically associated dimensions becomes evident in the highly artificial scenarios that result from our attempts to hold the confounded variables constant.

Future research could aim at investigating psychological mediators of effects of (enriched) distance. Previous studies in the framework of Construal Level Theory (CLT) have demonstrated the impact of psychological distance on the intensity of moral judgments (Eyal, Liberman, & Trope, 2008). The present findings constitute an interesting anomaly from the perspective of CLT, which predicts the impact of abstract, high-level moral values (such as "it is good to help others in need") on the intensity of moral judgment to *increase* with increasing psychological distance. While this seems to be true for temporal distance (Eyal et al., 2008), our results indicate that sense of obligation is not affected by spatial distance per se, and that it, if anything, *decreases* with increasing enriched distance. This might indicate that, at least in the realm of morality, the processes through which different distance dimensions operate are not as similar as CLT commonly assumes.

Acknowledgments

We thank Claudia Schwarz and Carina Suhr for their help with data collection.

References

Baron, J. & Miller, J. G. (2000). Limiting the scope of moral obligations to help: A cross-cultural investigation. *Journal of Cross-Cultural Psychology*, *31*, 703-725.

Eyal, T., Liberman, N., & Trope, Y. (2008). Judging near and distant virtue and vice. *Journal of Experimental Social Psychology*, *44*, 1204-1209.

Gillis, A. R. & Hagan, J. (1983). Bystander apathy and the territorial imperative. *Sociological Enquiry*, *53*, 449-460.

Greene, J. D., Cushman, F. A., Stewart, L. E., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2009). Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition*, *111*, 364-371.

Kamm, F. M. (2007). *Intricate Ethics*. Oxford: Oxford University Press.

Latané, B. & Darley, J. M. (1970). *The unresponsive bystander: Why doesn't he help?* Englewood Cliffs: Prentice-Hall.

Levine, M. & Thompson, K. (2004). Identity, place, and bystander intervention: Social categories and helping after natural disasters. *Journal of Social Psych.*, *144*, 229-245.

Milgram, S. (1965). Some conditions of obedience and disobedience to authority. *Human Relations*, *18*, 57-76.

Miller, J. G., Bersoff, D. M., & Harwood, R. L. (1990). Perceptions of social responsibilities in India and in the United States: Moral imperatives or personal decisions? *Journal of Personality and Social Psychology*, *58*, 33-47.

Singer, P. (1972). Famine, affluence, and morality. *Philosophy and Public Affairs*, *1*, 229-243.

Unger, P. (1996). *Living high and letting die.* New York: Oxford University Press.

Developing Notions of Free Will: Preschoolers' Understanding of How Intangible Constraints Bind Their Freedom of Choice

Nadia Chernyak (nc98@cornell.edu), Tamar Kushnir (tk397@cornell.edu)
Department of Human Development, Martha Van Rensselaer Hall
Ithaca, NY 14853 USA

Henry Wellman (hmw@umich.edu)
Department of Psychology, University of Michigan
530 Church Street
Ann Arbor, MI 48109 USA

Abstract

Our folk psychology involves the ability to reason about free will. In a series of experiments, we looked at young children's ability to reason about their own freedom of choice, and contrast this with their ability to reason about situations that constrain it. We asked preschoolers (Range: 4 y; 1 mo. – 5 y; 7 mo.) whether they had the choice to *have done otherwise* when they did not have the necessary knowledge to do so (epistemic constraint), had the moral duty not to do so (moral constraint), preferred not to do so (preference constraint), were told not to do so (permissive constraint), or were told that everyone else did not do so (conformist constraint). Results suggest that while preschool children generally believe their actions are freely chosen, they also understand how psychological, social and moral considerations may constrain their actions. These results have implications for children's developing notions of free will and moral reasoning.

Keywords: preschoolers, freedom of choice, morality, epistemic states

Introduction

Free will has long been studied in the field of philosophy, social psychology, and more recently, cognitive neuroscience (Baer, Kaufman, & Baumeister, 2008; Kane, 2002; Soon, Brass, Heinze, & Haynes, 2008; Wegner, 2003). Recent work has also begun to investigate how this important intuition develops and takes form in young children's reasoning (Kushnir, Wellman, & Chernyak, 2009; Nichols, 2004; Seiver, Kushnir, & Gopnik, 2009).

For example, Nichols (2004) found that six-year-old children ascribe the choice to *have done otherwise* to an agent, but not an inanimate object. Therefore, Nichols (2004) posits an *agent-causal view* of free will in which children believe that agents have indeterminate choice which is unbound by outside forces. This is contrasted with children's beliefs about physical causation, namely that, unlike agents, inanimate objects are *not* free to choose their own course of action and are wholly governed by outside forces.

However, the distinction between agents and inanimate objects is only part of our adult intuitions about freedom of choice. More central to our mature understanding – and to the important role that intuitive notions of free will play in our social and moral reasoning – is the ability to contrast situations in which agents are free to choose and situations in which agents are constrained in their choices. In other words, to adults, "free will can't really mean that at any moment a person's behavior is totally unpredictable (and therefore entirely unconstrained)" (p.4; Baer et al., 2008). Therefore, understanding free will implies understanding the complementary notion of *constraint*.

Kushnir et al. (2009) asked four- and five-year old children if they *could have done otherwise* in two situations. One in which they were free to draw a picture and one in which they were physically prevented from doing so (i.e., the experimenter held the child's hand so that it was stuck in one place). Children overwhelmingly responded that they had freedom of choice when they were physically unbounded, but responded that they did not have that freedom when they were physically constrained. Therefore, preschoolers may already know that their agency, and therefore their freedom of choice, is limited by the physical world.

However, the physical world is just one type of force that may constrain one's free will. One's freedom to choose may also be constrained, or at least limited, by non-physical phenomena, such as beliefs, knowledge states, desires, and social and moral obligations. Research on children's social cognition shows that preschoolers have a rather firm grasp of how constraints which come from the mind differ from those of the physical world (Inagaki & Hatano, 1999, Wellman, 1990). In the current investigation, we explore two related questions about such "intangible" constraints: First, do young children understand that these constraints bind their freedom of choice? Or alternatively, do they believe that their ability to *have done otherwise* is unbounded by psychological and social forces, and is subject only to the laws of the physical world? Second, can children distinguish between intangible constraints which fully determine behavior (and thus fully constrain free will) and those which only influence it (and thus do not fully constrain free will)?

Experiments 1 and 2 explored the first question by asking older and younger preschool children whether they believed they *had the choice to do otherwise* when they didn't have the necessary knowledge to do so. We chose this epistemic constraint – that seeing leads to knowing – because it is one

with which children are quite familiar (Wellman, 1990). Critically, this constraint fully limits one's free will, much like a physical constraint. Thus, we predict that, if children understand intangible (non-physical) constraints, the results should replicate Kushnir et al.'s (2009) findings.

Experiment 3 explored the second question by asking preschoolers about their freedom to act against constraints which, by adult intuitions may influence behavior, but do not fully constrain one's free will. Therefore, we asked children whether they believed they *had the choice to do otherwise* when bound by moral considerations, personal preference, permission, and conformity.

Experiment 1

In Experiment 1, a group of older preschoolers (4.5- 5-year-olds) were asked to reproduce two shapes from a modeled drawing. Across two trials, we varied when each child had the ability to see (thus, to know about) a modeled shape. In the *Constrained Drawing* trial, the modeled shape was hidden from the child's view behind an occluder. In the *Free Drawing* trial, the modeled shape was visible. After drawing, children were asked if they *could have done otherwise* – that is, if they could have drawn the shape they didn't see (and therefore didn't draw) in the Constrained Drawing trial, or if they could have drawn the shape that they did see (but didn't draw) in the Free Drawing trial. We also asked them to explain their responses. If children understand the epistemic constraint binding their free will, then their responses and explanations should differ across the two trials.

Method

Participants 22 four- and five-year-old children (Mean age = 4 y; 11 mo.; SD = 6 mo.) were recruited from preschools in Ithaca, NY.

Procedure Children were interviewed individually in a separate room in the preschool by a female experimenter. Four colored placemats (randomly chosen and ordered from a set of red, orange, green, yellow, blue, and brown), were used to distinguish between the individual trials. The occluder was a black piece of construction paper.

The set-up is shown in Figure 1. The experimenter began by first showing children a drawing of a dot (Shape B) and asking the child to label it. This was followed by the Free and Constrained Drawing trials, order counterbalanced. Each of these trials consisted of an action (drawing a shape), an outcome (the shape) and two critical questions (Alternate Choice Judgment and explanation).

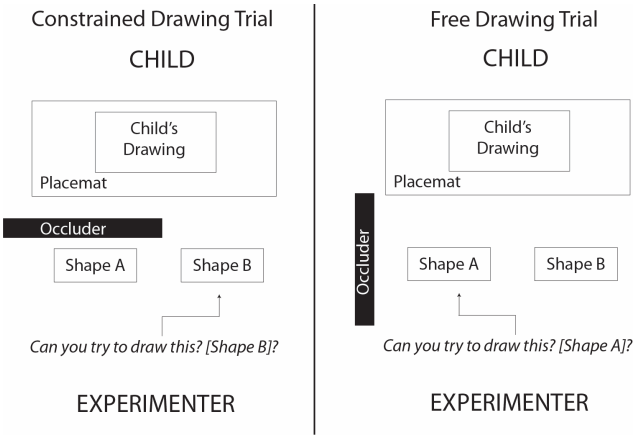


Figure 1: Set-up of Experiment 1.

Action: The experimenter drew Shape A, hidden by the occluder, saying "And now, I'm going to put the paper up like this and draw a different shape." Shape A was either a line or a circle (randomly chosen).

Outcome: The Experimenter then asked the child to draw the hidden shape ("Can you try to draw this?"). If the child refused to draw, the experimenter encouraged them to draw Shape B. Ten children drew Shape B, and 12 drew something on their own.[1] After the child finished drawing, the experimenter revealed the hidden shape ("Now I'm going to show you what I drew!")

Questions: The colored mat was then set aside and children were asked the *Alternate Choice Judgment*: "Last time, on the [blue] mat…could you have drawn the [line]?" The child was then asked to *explain* his/her response.

Coding Explanations were coded and classified into the following four categories: References to Epistemic Constraints ("because the paper was up and I couldn't see it"; "because this time the paper wasn't up"); Enactments ("by going like this"), Non-Explanations ("because there was a dot there"; "I don't know"), and References to Other Constraints ("because you told me to draw this one.").

Results and Discussion

Figure 2 shows that children's responses to the Alternate Choice Judgment were marginally different between conditions. In the Free Drawing trial 12/22 (54.5%) children indicated that they could have drawn the other shape. In contrast, only 8/22 (36%) of the children said they could have drawn the hidden shape in the Constrained Drawing trial (McNemar's $p = .07$, one-tailed).

[1] Analyses revealed no differences between these two groups

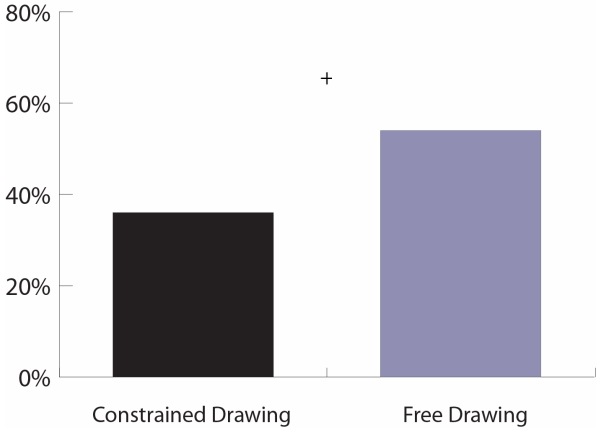


Figure 2: Percentage of children who said that they "could have drawn something else" in Experiment 1.

Figure 3 shows the same pattern in children's explanations. The majority of explanations (54.5%; 12/22) in the Constrained Drawing trial appeal to the epistemic constraint imposed in the task. Epistemic explanations were provided more often than non-explanations, χ^2 (1, N = 15) = 5.40, $p < .05$, enactments, χ^2 (1, N = 16) = 4.00, $p < .05$, and other constraints, χ^2 (1, N = 15) = 5.40, $p < .05$. In contrast, in the Free Drawing trial children mostly provided enactments and non-explanations. In the Free Drawing trial, enactments were provided most often, significantly more often than references to epistemic constraints, χ^2 (1, N = 11) = 4.46, $p < .05$ and the proportion of enactments was not significantly different from the proportion of non-explanations and references to other constraints (all *p*s non-significant).

Like children's judgments, children's explanations differed significantly between trials. Children were significantly more likely to provide epistemic explanations in the Constrained Drawing trial than in the Free Drawing trial (McNemar's $p = .001$, one-tailed). Similarly, a greater proportion of children in the Free Drawing trial explained their response by enactment (demonstrating the alternate action) (McNemar's $p < .05$, one-tailed). The proportion of non-explanations and references to other constraints did not differ significantly between trials.

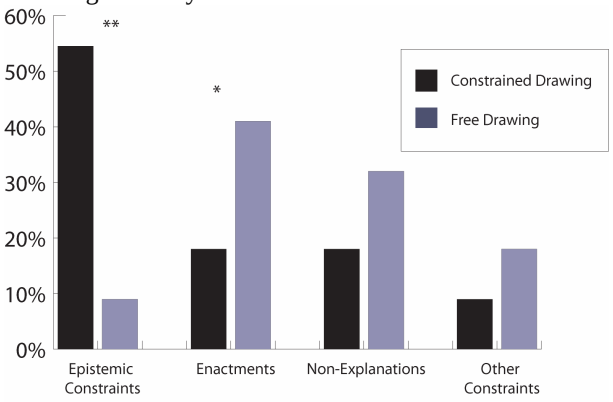


Figure 3: Proportion of Explanation Types within Each Trial in Experiment 1

Experiment 2

In Experiment 1, older preschoolers' judgments and explanations revealed some ability to reason about epistemic constraints on free will. In Experiment 2, we replicated the task with a sample of younger preschoolers, and also made a few critical modifications to the procedure. First, we included a warm-up to prime children to think about knowledge states. Second, we eliminated the ambiguity in the Constrained Drawing trial of what the child was supposed to draw by doing away with Shape A (see Figure 4). Thus, the experimenter had only one drawing in front of her (hidden or visible, depending on the condition). Note also that, in this modified procedure, children were free to draw whatever they wanted to in both trials except, of course, the picture they could not see.

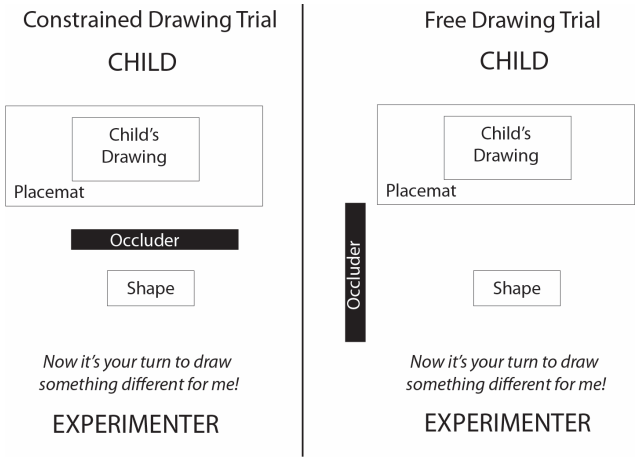


Figure 4: Set-up of Experiment 2

Method

Participants 26 four-year-old children (M=4 y; 6 mo; SD=4.8 mo) were recruited from preschools in Ithaca, NY, Cortland, NY, and New York, NY. The ages of these children was significantly lower than of those in Experiment 1, $t(46) = 3.58$, $p < .01$. All preschools were roughly matched for socioeconomic status and demographic population.

Procedure *Knowledge Access Warm-Up:* In order to prime children to think about knowledge states, we began with a knowledge access task (Wellman & Liu, 2004). In this procedure, children were shown a drawer with hidden contents, asked to guess the contents of the drawer, and then prompted to open the drawer, revealing a toy dog. The drawer was then closed and a doll ignorant to the contents of the drawer was introduced ("Now Polly has never ever seen inside this drawer. Here comes Polly!"). Children were then asked two questions pertaining to the doll's knowledge state: "Does Polly *know* what's in the drawer?" and "Has Polly *seen* inside the drawer?" 85% (22/26) of the children

answered both questions correctly. Corrective feedback was not provided.

The experiment then continued with the same two trials (Free Drawing and Constrained Drawing) as in experiment 1. As shown in Figure 4, the set-up was simplified to include only one drawing in front of the experimenter (either hidden or visible) and a blank sheet of paper in front of the child. The experimenter first drew her shape, then she asked the child "can you see it?" She then prompted the child to draw by saying, "Now it's your turn to draw something different for me!" After both drawings, the experimenter revealed her shape (if hidden) and asked the Alternate Choice Judgment and explanation questions. Coding was the same as in Experiment 1.

Results and Discussion

The results replicate the findings of Experiment 1 with younger preschoolers. Figure 5 shows that children's responses to the Alternate Choice Judgment were significantly different across conditions. In the Free Drawing trial, 17/26 (65%) of children answered that they could have drawn the other shape, whereas only 9/26 (35%) did so in the Constrained Drawing trial (McNemar's $p < .05$, one-tailed).

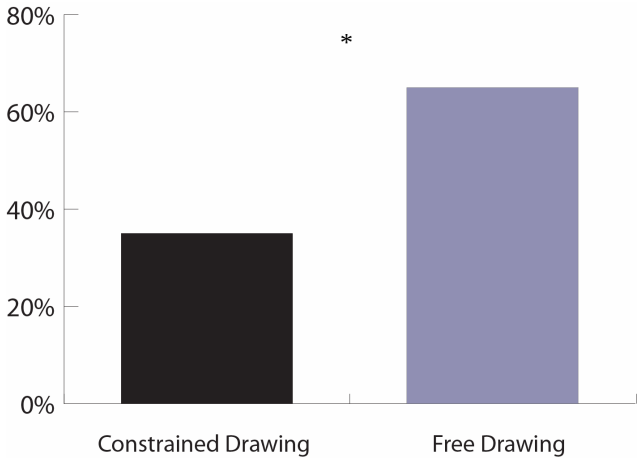


Figure 5: Percentage of children who responded they "could have drawn something else" in Experiment 2

Though younger children provided fewer explanations overall, of the 54% (14/26) of children who provided explanations, 36% (5/14) referred to epistemic constraints in the Constrained Drawing trial, whereas no child referred to epistemic constraints in the Free Drawing trial (McNemar's $p < .05$, one-tailed). This difference is consistent with the pattern of explanations provided by the older preschoolers in Experiment 1. The difference between the proportion of non-explanations, enactments, and references to other constraints in the Free and Constrained drawing trials were not significant.

Experiment 3

Experiments 1 and 2 revealed that 4- and 5-year-olds can reason about their own free will and epistemic constraints on their free will. These results provide initial evidence that young children may already understand that their freedom to act can be restricted by non-physical, intangible constraints.

Experiment 3 focused on other intangible constraints which influence, rather than fully limit, free will – moral considerations, personal preferences, permission, and conformity. Research has shown that even three-year-olds are sensitive to moral rules (e.g., Smetana, 1981) and the subjective nature of preferences (Wellman, 1990, Wellman & Woolley, 1990; Repacholi & Gopnik, 1997). Young children are further able to reason about how rules of permission (Kalish & Shiverick, 1995); and conformist considerations (Kalish, 1998; Racoczy, Warneken, & Tomasello, 2008) may determine one's actions. Do children also believe that these factors constrain their freedom of choice? If so, do they consider these four influences to be equally constraining, or do they distinguish among them?

Participants Participants were 15 four- and five-year olds (Mean age = 4 y; 7 mo.; SD = 4.5 mo) recruited from preschools in Ithaca, NY, Cortland, NY, and New York, NY. Preschools were roughly matched for socioeconomic status and demographic population.

Procedure All children completed four trials (randomly ordered): Moral Trial, Preference Trial, Permissive Trial, and Conformist Trial. In each trial, children began by being shown two shapes (randomly chosen from a set of 8: a dot, a line, a circle, a square, a triangle, a squiggly line, an X, and a U). Each child was then given a white piece of paper on a colored mat and introduced to one of four puppets (a dog, a cat, a pig, or an elephant; randomly chosen).

In the *Moral Trial*, children were asked to act in accordance with a moral obligation: "This is [Doggie]. [Doggie] *hates* [triangles]. [Triangles] remind him of something *really* sad, and sometimes, when he sees them, he even cries! Can you draw the [circle (i.e, other shape)]? In the *Preference Trial*, children were told: "This is [Piggy]. [Piggy] really likes to watch people draw! She wants you to draw whichever one of these shapes you like the best. Can you draw the one you like the best?" Children were then prompted to draw one of the two shapes they had just seen. In the *Permissive Trial*, the experimenter asked the child to act in accordance with a non-moral rule: "This is [Kitty]. [Kitty] says the rule is you *have* to draw a [squiggly]. She says that's the rule and you have to do it. Can you draw a [squiggly]?" Finally, in the *Conformist Trial*, children were asked to do as everyone else has done: "This is [Ellie]. [Ellie] just played with lots of boys and girls and all of them drew a [line]. She says *every* one of them drew a [line]. Can you draw a [line]?"

After each trial, the colored placemat was set aside, and children were asked the Alternate Choice Judgment and explanation questions (as in Experiments 1 and 2).

Coding

Explanations were coded into the following six categories: references to Moral Constraints ("because it would make Doggie sad"), Preferences ("because I wanted to draw the square"), Permissive Constraints ("because Doggie said to draw the line"), Conformist Constraints ("because all of my friends did it"), Enactments ("by going like this"), and Non-Explanations ("because there was a dot there"; "I don't know").

Results and Discussion

The results show that, to a large extent, 4- and 5-year-old children believe their free will is constrained by all four contexts. Overall, 77% (46/60) responses to the Alternate Choice Judgment question were "no's" and 67% (40/60) of the explanations refer to one of the coded constraints. However, there were also important differences between the four constraints in both judgments and explanations.

Figure 6 shows that a significant majority of children (87%; 13/15) indicated that they did not have the choice to act immorally (Binomial $p < .05$) or against conformity (87%; Binomial $p < .05$). A non-significant majority indicated that they could not act against permission (60%; 9/15), or their own preference (73%; 11/15).

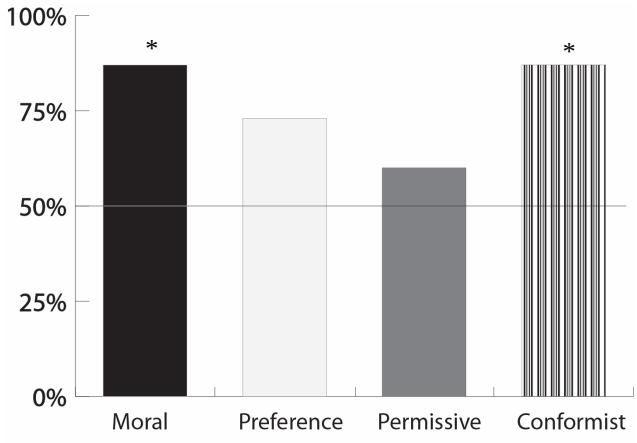


Figure 6: Percentage of children who answered they "could *not* draw something else" in Experiment 3

Figure 7 shows the proportion of each type of explanation that children gave in each trial. In the Moral Constraint trial, the majority (87%; 13/15), of children appealed to moral considerations in their explanations. Also, moral constraints were most often referenced in the moral trial than each of the other three trials (all McNemar's p's $< .01$, one-tailed).[2] In the Preference Constraints trial, approximately half (53%; 8/15) of children referred to preference considerations in their explanations. Preference constraints were references more often in this trial than each of the other three (all McNemar's p's $< .05$, one-tailed). In the permissive trial, only 33% (5/15) children referenced constraints of permission, and in the conformist trial, only 13% (2/15) referenced conformist constraints in their explanations. Moreover, the number of permissive and conformist explanations was low overall and did not significantly vary between trials. Enactments and non-explanations also did not significantly vary between trials.

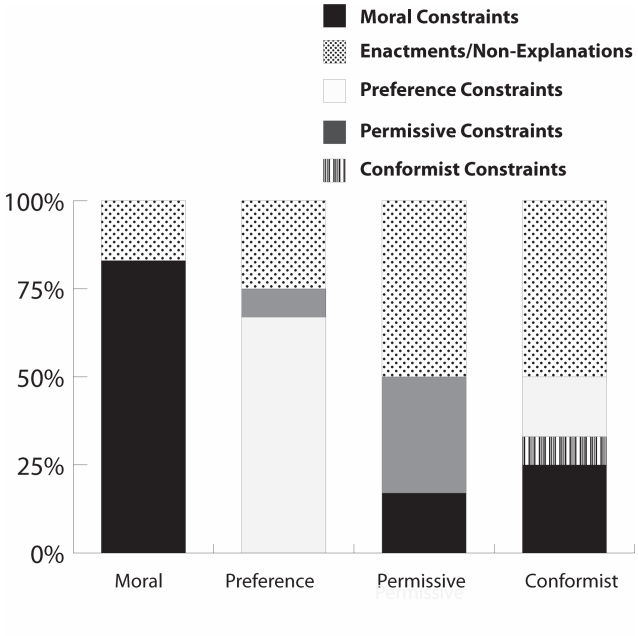


Figure 7: Proportion of Explanation Types within Each Trial in Experiment 3

The results suggest that children responded quite differently to each type of constraint. These differences are further illuminated by analyzing the relationship between children's judgments and their explanations. Only moral constraints were overwhelmingly judged and explained consistently and appropriately – 80% (12/15) of the children responded that they could not draw the other shape ("no" judgment) because it would make the puppet cry (Moral Constraint explanation). By the same analysis, preference constraints were also somewhat consistently evaluated – 47% (7/15) of the children said that they could not draw the other shape because they didn't like it as much. On the other hand, children's overwhelming "no" judgments in the Conformist trial were almost never followed by conformist constraint explanations – only 13% (2/15) of the children said they could not draw the other shape because no one else did. Also, only 20% (3/15) of the children said they could not draw it because those were the rules. Further research is needed to understand the reasons for these differences.

The most critical finding, then, is that children overwhelmingly said that they were not free to act to harm another person. One potential interpretation might be that

[2] Of those that referred to Moral Constraints in the other (non-moral) trials, all children experienced the Moral Trial before the trial in which they referenced the moral constraint, suggesting the presence of an order effect.

children felt "pressured" to state that they could not act immorally because a moral rule presented a "permissive" rule in some sense (Piaget, 1932/1997). However, children's explanations reveal that this is not the case – children referenced moral considerations ("because it would make Doggie cry") rather than permissive ones. Also, they were clearly less likely to say they were constrained by a simple rule (Permissive trial).

General Discussion

The results of these three studies show that by the time children are five years old they have an intuitive notion of free will that is sensitive to certain intangible constraints. Importantly, in contrast to the fact that children overinflate their own abilities (e.g., Stipek, 1984) preschool-aged children do not simply believe that their freedom to choose is limitless. Instead, preschool-aged children already appear to have notions of freedom of choice that are in-line with "compatibilist" (Hume, 1910) views (i.e., that some actions are fully or partially determined while others may be entirely unconstrained).

We also found that preschool children can reason about both wholly constraining (Experiments 1 and 2) and limiting (Experiment 3) influences on their past actions. Moreover, their responses indicate that they distinguish between different types of constraints. This is consistent with past work showing that young children understand the limiting nature of morality (Smetana, 1981; Yamada, 2008) and the nature of social norms (Kalish, 1998; Kalish & Shiverick, 1995). In adults, freedom of choice is linked to moral and normative behavior (Phillips & Knobe, in press; Vohs & Schooler, 2008). The current study suggests that this link is already present in very young children.

In the real world, social and psychological factors often come in conflict. For example, the desire to have your sister's toy may conflict with the moral judgment that grabbing it from her would make her cry. Future work could study how preschoolers reason about freedom of choice when these social and psychological factors conflict.

Acknowledgments

We gratefully thank Heather Baker-Carr, Sarah Fogel, Andy Hsia, Kristyn Herlihy, Laurel Hollenbaugh, Niki Klein, Lauren Latella, Tanya Perry, Carolina Romero, Lauren Schneider, Katie Weidlein, and Frank Wilbourn, for assistance with data collection and transcription.

References

Baer, J., Kaufman, J. C., & Baumeister, R. F. (Eds.) (2008). *Are We Free? Psychology and Free Will*. New York, NY: Oxford University Press.

Hume, D. (1910). *An Enquiry Concerning Human Understanding*. New York: P. F. Collier & Son Corporation.

Inagaki, K., & Hatano, G. (1999). Children's understanding of mind-body relationships. In Siegal, M., & Peterson, C. C. (Eds.) *Children's Understanding of Biology and Health*. Cambridge, UK: Cambridge University Press.

Kalish, C. W. (1998). Reasons and causes: Children's understanding of conformity to social rules and physical laws. *Child Development, 69*, 706-720.

Kalish, C. W., & Shiverick, S. M. (1995). Children's reasoning about norms and traits as motives for behavior. *Cognitive Development*, 19, 401-416.

Kane, R. H. (Ed.) (2002). *The Oxford Handbook of Free Will*. New York, NY: Oxford University Press.

Kushnir, T., Wellman, H. M., & Chernyak, N. (2009). Preschoolers' Understanding of Freedom of Choice. *Proceedings of the Thirty-First Annual Conference of the Cognitive Science Society*, 87-92.

Nichols, S. (2004). The folk psychology of free will: Fits and starts. *Mind and Language, 19*, 473-502.

Piaget, J. (1997). *The Moral Judgment of the Child*. New York, NY: Free Press Paperbacks. (Originally published in 1932).

Phillips, J., & Knobe, J. (in press). Moral judgments and freedom. *Psychological Inquiry*.

Repacholi, B. M., & Gopnik, A. (1997). Early reasoning about desires: Evidence from 14- and 18-month-olds. *Developmental Psychology, 33*, 12-21.

Racoczy, H., Warneken, F., & Tomasello, M. (2008). The sources of normativity: Young children's awareness of the normative structure of games. *Developmental Psychology, 44*, 875-881.

Seiver, E., Gopnik, A. & Kushnir, T (2009, October). Children As Philosophers: Differing Conceptualizations Of Free Will At Ages 4 And 6. Poster presented at the biennall meeting of the Cognitive Development Society. San Antonio, TX.

Stipek, D. J. (1984). Children's perceptions of their own and their classmates' ability. *Journal of Educational Psychology, 73*, 404-410.

Smetana, J. G. (1981). Preschoolers' understanding of moral and social rules. *Child Development, 52*, 1333-1336.

Soon, C. S., Brass, M., Heinze, H., & Haynes, J. (2008). Unconscious determinants of free will decisions in the brain. *Nature, 11*, 543-545.

Vohs, K. D. & Schooler, J. W. (2008). The value of believing in free will: Encouraging a belief in determinism encourages cheating. *Psychological Science, 19*, 49-54.

Warneken, F. & Tomasello, M. (2008). Extrinsic rewards undermine altruistic tendencies in 20-month-olds. *Developmental Psychology, 44*, 1785-1788.

Wegner, D. (2003). *The Illusion of Conscious Will*. Cambridge, MA: MIT Press.

Wellman, H. M. (1990). *The Child's Theory of Mind*. Cambridge: MIT Press

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development, 72*, 655-684.

Yamada, H. (2009). Japanese children's reasoning about conflicts with parents. *Social Development, 18*, 962-977.

Symposium: Dynamic Decision Making

Todd M. Gureckis (*Moderator*) and Douglas Markant
Department of Psychology, New York University

Jared M. Hotaling, Eric Dimperio, and Jerome R. Busemeyer
Department of Psychological and Brain Sciences, Indiana University

Michael D. Lee, Shunan Zhang, and Mark Steyvers
Department of Cognitive Sciences, University of California at Irvine

Bradley C. Love and A. Ross Otto
Department of Psychology, University of Texas at Austin

Dylan A. Simon and Nathaniel Daw
Center for Neural Science, New York University

Keywords: sequential decision making, computational modeling, cognitive neuroscience

The experimental study of decision making has historically focused on simple single-trial judgement or reasoning tasks. However, real world behavior often necessitates online decision making, planning, and sequentially organized behavior. The goal of the proposed symposium is to bring together researchers who are working to understand the cognitive processes underlying this kind of "dynamic decision making" (defined as tasks or contexts that are structured as a sequence of interdependent decisions).

A symposium on this topic is particularly timely since research in the area of dynamic decision making is having a tremendous impact on the field of psychology as a whole. First, researchers are converging on a set of novel computational modeling approaches that explain how decision makers plan sequences of multiple actions, take into account future contingencies, and react in real time to continually changing environmental dynamics. Second, many of the proposed algorithms and models are closely linked to neurobiological correlates (e.g., the recent explosion of research on neurobiology of reinforcement learning). Third, many of the tasks that are being developed for evaluating these models also appear to relate to important individual differences in real-world decision-making. The goal in the symposium is to 1) highlight some of the best work in this area, 2) to facilitate communication between researchers working on these problems from varying perspectives, and 3) to provide an excellent showcase of this area for members of the cognitive science community who may not yet be familiar with this work.

The speakers who agreed to participate are all accomplished researchers in this area but each approach the set of problems involved in sequential decision making and learning from a slightly different perspective. The key topics covered include 1) how people plan sequences of actions to accomplish goals (Hotaling, Dimperio, & Busemeyer, Simon & Daw), 2) the underlying neurobiology of sequential decision making and planning (Simon & Daw), 3) how cognitive representations of the task or environment supporting planning and decision-making (Gureckis & Markant, Love & Otto, and Simon & Daw), and 4) how people balance exploration and exploitation in order to arrive at effective decision strategies in an unknown environment (Lee, Zhang, and Steyvers and Gureckis & Markant). In addition to these overlapping psychological themes, the researchers all share a core approach of applying sophisticated computational models to understand human behavior (including Bayesian approaches, reinforcement learning, and Markov Decision Processes).

Todd Gureckis & Douglas Markant (New York University)

Exploring to Exploit: Modeling the Process of Information Search and Planning

Effective learning often involves actively querying the environment for information that can be exploited at a later point in time. However, the space of observations available in any situation can vary greatly in potential "informativeness" and relative cost. How do people decide which observations to make at any point in time given their future goals? We describe a series of studies looking at how people plan sequences of information collection actions in a cognitive search task based on the children's game Battleship. Participants made sequences of observations to disambiguate between a large number of potential game configurations subject to information-collection costs. Computational models are developed which predict which observations people will make on any given trial and when they should stop collecting information and exploit their current knowledge. In particular, the models measure the degree to which individuals take into account future consequences when planning immediate actions. In our second study, we explore how people generate hypotheses consistent with their prior

beliefs and how these hypotheses in turn influence search behavior.

Jared M. Hotaling, Eric Dimperio, & Jerome R. Busemeyer (Indiana University)

Cognitive Models of Planning Behavior in Multi-Stage Risky Decision Making

Much research into risky decision making has traditionally presented individuals with choice alternatives that provide an immediate reward or punishment based on the outcome of a random event. This allows researchers to understand how the values of choice alternatives and the probabilities associated with risk can influence an individuals choices. We present recent work that extends that research by manipulating the outcome probabilities and rewards involved in a multistage decision task where some rewards are only possible after a sequence of decisions. Our results show individual differences, with some participants being sensitive to possible future rewards and likelihoods, and others appearing not to plan ahead. A comparison of multiple competing models helps identify decision processes when planning ahead did occur.

Michael Lee, Shunan Zhang, & Mark Steyvers (Univ. of California, Irvine)

Human and optimal exploration and exploitation in sequential decision-making

In bandit problems, a decision-maker chooses repeatedly between a set of alternatives. They get feedback after every decision, either recording a reward or a failure. They also know that each alternative has some fixed unknown probability of providing a reward when it is chosen. The goal of the decision-maker is to obtain the maximum number of rewards over all the trials they complete. Bandit problems provide an interesting formal setting for studying the balance between exploration and exploitation in decision-making. In early trials, it makes sense to explore different alternatives, searching for those with the highest reward rates. In later trials, it makes sense to exploit those alternatives known to be good, by choosing them repeatedly. How exactly this balance between exploration and exploitation should be managed, and should be influenced by factors such as the distribution of reward rates, the total number of trials, and so on, raises basic questions about adaptation, planning, and learning in intelligent systems. In this talk, we present a series of models, both Bayesian and heuristic, aimed at understanding how people balance exploration and exploitation, and how their strategies relate to optimal decision-making.

Brad Love & A. Ross Otto (University of Texas at Austin)

You Don't Want To Know What You're Missing: When Information about Forgone Rewards Impedes Dynamic Decision Making

When learning to make decisions from experience, one reasonable intuition is that adding relevant information should improve performance. In contrast, we find that additional information about foregone rewards (i.e., what could have gained at each point by making a different choice) severely hinders participants ability to repeatedly make choices that maximize long-term gains. We conclude that foregone reward information accentuates the local superiority of short-term options (e.g., consumption) and consequently bias choice away from productive long-term options (e.g., exercise). These conclusions are anticipated by a standard reinforcement learning mechanism that processes information about experienced and forgone rewards. In contrast to related contributions using delay-of-gratification paradigms, we do not posit separate top-down and emotion-driven systems to explain performance. We find that individual and group data are well characterized by a single reinforcement learning mechanism that combines information about experienced and foregone rewards. These findings will be situated within a broader research program that aims to characterize how people explore and exploit environments with unknown rewards and poorly understood states. Finally, interventions for improving human performance will be discussed.

Dylan Simon & Nathaniel Daw (New York University)

Neural correlates of decision evaluation by forward planning in sequential tasks

Theoretical models of reinforcement learning are commonly applied within neuroscience to explain the neural processes involved in learning and decision making. However, the approaches used are predominately "model-free," such as temporal difference learning which learns action values or policies directly from reinforcement without explicitly representing or utilizing any information about task structure. While these theories have shed light on observed neural activity in simple "bandit" tasks involving repeated choices rewarded independently and immediately, it is at odds with a long line of behavioral evidence from psychology and cognitive science for more flexible, goal-directed forward planning processes. We show how a different set of "model-based" reinforcement learning algorithms can be used to account for these phenomena, and test this framework in humans using a number of dynamic, continuous, sequential decision tasks. The models can account both for observed choice behavior and fMRI BOLD signals in decision-related brain areas. Consistent with cognitive theories, both behavior and neural activity show evidence of flexible learning and forward planning, indicating that existing neural models provide an incomplete picture of learning and decision making in dynamic tasks.

The Active Role of Partial Knowledge in Cross-Situational Word Learning

Daniel Yurovsky, Damian Fricker, Chen Yu, and Linda B. Smith
{dyurovsk, dfricker, chenyu, smith4} @indiana.edu
Department of Psychological and Brain Science, and Cognitive Science Program
1101 East 10th Street Bloomington, IN 47405 USA

Abstract

A number of modern word learning theories posit statistical processes in which knowledge is accumulated across many exposures to a word and its potential referents. Accordingly, words do not go directly from unknown to known, but rather pass through intermediate stages of partial knowledge. This work presents empirical evidence for the existence of such partial knowledge, and further demonstrates its active driving role in cross-situational word learning. Subsequently, an incremental model which leverages its partial knowledge of word-object mappings from trial to trial is shown to account well for the data. In contrast, models which do not do so cannot explain the data. These results confirm crucial assumptions made by statistical word learning models and shed light on the representations underlying the acquisition of word meanings.

Keywords: word learning; language acquisition; computational modeling; statistical learning

Introduction

We have a tendency to characterize word learning as an all-or-none process: either a child knows a given word, or she has not yet learned it. This is apparent in our methodology (e.g. forced-choice tests, preferential looking), and assessment of vocabulary size via MCDI (Fenson, Dale, Reznick, Bates, Thal, & Pethick, 1994), as well as some theoretical claims. But this implicit all-or-none characterization may stymie our thinking about potential word-meaning representations.

For almost a century we have known that human learning and memory are not binary phenomena (Ebbinghaus, 1913). In learning lists of paired associates, for instance, a failure to recall the correct pair for a prompt does not imply no knowledge of the mapping. Evidence of this knowledge can be recovered using a different test paradigm (e.g. recognition or savings). The knowledge is not absent, but rather partial or sub-threshold. The central idea motivating this work is that such sub-threshold knowledge may play a profound in the course of language acquisition.

Several recent theoretical and computational approaches to word learning have made explicit use of partial knowledge. For instance, McMurray (2007) modeled the learning of a word's meaning as the acquisition of partial meaning tokens. Yu and Smith (2007) argued that early word learning can be thought of as the accumulation of co-occurrence statistics between words and objects across multiple situations. These theories suggest that a word can be learned in bits rather than in a single perfect moment.

Other models make an even stronger claim: not only can one build lexical knowledge by accumulating parts; this partial knowledge is an active driver of the learning system (Blythe, Smith, & Smith, in press, McMurray, Horst, Toscano, & Samuelson, in press, Fazly, Alishahi, & Stevenson, in press, Yu, 2008). These models have been tested predominantly on large corpora, reproducing qualitative patterns found in children's word learning. If they are correct about the presence and role of partial knowledge, however, then we should be able to find empirical evidence for the role of partial knowledge in human word learners.

Yurovsky and Yu (2008) presented indirect evidence of the active role of partial knowledge in cross-situational learning. They exposed participants to a series of individually ambiguous learning trials consisting of multiple words and multiple objects. At the end of each trial, participants were asked to indicate how sure they were (1-10) that they knew the correct label for each object. Yurovsky and Yu showed that a given object's rating could be predicted from the ratings given to the other objects on the same trial, even after the object's rating on its previous exposure was taken into account. Thus, participants seemed to be using partial knowledge of word-object mappings to reduce the set of candidates for other labels.

This analysis, while promising, was performed on participants' subjective knowledge ratings. In the present work, we propose to offer stronger and more direct evidence that partial knowledge plays an active role in word learning. To this end, we expose participants to two consecutive blocks of cross-situational learning. Crucially, half of the words and objects in the second block are those which participants failed to learn in the first block. Comparing the results of block 2 to those of several control conditions, we can determine the role of partial knowledge in cross-situational learning. First, we can ask whether partial knowledge exists in the system, whether learners are really accumulating bits of sub-threshold knowledge.

At a deeper level, we pursue a more interesting question: does partial knowledge of *individual* word-referent pairs – interacting in a *system* with partial knowledge of other word-referent pairs – facilitate the acquisition of new words. To answer this question, a set of computational models are fit to the data to understand the underlying learning mechanisms which give rise to the empirical results. We compare a *simple associative model*, a *biased associative model* which increments associations in proportion to their current strength, and a *competitive associative model* which adds within-trial competition. In the simple associative model, partial knowledge is not used in learning. In the biased associative model, partial knowledge of a word-

referent pair drives learning of that individual pair. Finally, in the competitive associative model, partial-knowledge of multiple word-referent pairs interacts and, by so doing, facilitates the learning of other pairs and thus the whole system of words and referents.

Experiment 1

To demonstrate the role of partial knowledge in word learning, we used the cross-situational word learning paradigm (Yu & Smith, 2007). In this task, participants are exposed to a series of individually ambiguous learning trials, each of which contains multiple co-occurring words and potential referents. While each trial is individually unambiguous, words always co-occur with their correct referent, and thus participants who correctly track co-occurrence between words and objects across trials can learn the correct pairings.

In Experiment 1, participants were exposed to two consecutive blocks of cross-situational word learning. At the end of block 1, participants were asked to select the correct referent for each of the trained words. For participants in the *unlearned* condition, half of the stimuli in block 2 were word-object pairs from block 1 for which they selected incorrect referents. For participants in the *new* condition, all stimuli in the second block were new.

If participants encoded nothing about words for which they selected incorrect referents in block 1, participants in the *unlearned* and *new* conditions should learn equally well in block 2. Alternatively, since no feedback is provided at test, if participants who selected incorrect referents did so as the result of binary hypotheses, and carried these wrong hypotheses to block 2, we might expect participants in the *unlearned* condition to underperform those in the *new* condition. However, if participants who selected incorrectly possess sub-threshold knowledge of the correct referent, we would expect participants in the *unlearned* condition to perform better than *new* participants in block 2. Most interesting would be if sub-threshold knowledge of one pair interacted with sub-threshold knowledge of other word-referent pairs to facilitate learning new pairs in block 2.

Method

Participants. Ninety-two Indiana University undergraduates participated in exchange for course credit; 50 in the *unlearned* condition and 42 in the *new* condition. However, to ensure a fair comparison across conditions, data from only a subset were analyzed (criteria explained in procedure). The final analysis was conducted on 23 participants in the *unlearned* condition, and 10 participants in the *new* condition.

Stimuli. Referents were represented by pictures of unusual objects which were easy to distinguish from each other, but difficult to name. Words were 1-2 syllable synthesized nonsense words constructed to be phonotactically probable in English. All words and objects have been used in previous cross-situational learning experiments (Yu & Smith, 2007, Yurovsky & Yu, 2008). Forty-two unique words and objects were used in total – 24 in block 1 and 18 in block 2.

Training slides for block 1 presented two pictures – one on each side of the screen – and played two labels, following Yu and Smith's (2007) 2x2 condition. Training slides for block 2 presented four objects – one in each corner of the screen – and played four labels, following Yu and Smith's (2007) 4x4 condition. Test slides for each block displayed all of the objects from that block (24 for block 1, 18 for block 2) in random positions and played one label.

Procedure. Each participant was exposed to two blocks of cross-situational learning – first a 2x2 block and then a 4x4 block. Each block consisted of a training phase followed by a test phase. The training phases consisted of a series of trials each displaying a set of objects and playing an equal number of words. Screen position and word order were randomized, such that they provided no information about which word labeled which object.

Following training, participants were given a series of alternative forced choice tests in which they were asked to select the correct referent for each label. Each word was tested once, and all objects from a block were presented on each test trial, so the content of test trials was uninformative as to correct mappings.

Block 1 contained 24 novel words, each of which occurred 5 times with its correct referent and less often with other objects. This resulted in 60 2x2 trials in total. Block 2 contained 18 words, each of which occurred with its correct referent 4 times and less often with other objects. This resulted in 18 4x4 trials. Word-object pairings and trial orders were selected randomly for each participant.

Block 1 was identical for participants in both the *new* and *unlearned* groups. The stimuli for block 2 differed across conditions. In the *new* condition, block 2 consisted entirely of novel stimuli – 18 words and their associated objects. For participants in the *unlearned* condition, however, 9 of the words and objects in the second block were those for which they had selected the incorrect response at test in block 1 (see Figure 1). Thus, for participants in the *unlearned* group, half of the stimuli in block 2 were words and objects for which they had not successfully learned correct associations. We will refer to the words and referents carried over from block 1 as old and those which are seen for the first time in block 2 as new.

Since participants could complete block 2 of the *unlearned* condition only if they had selected incorrect referents for at least 9 words in block 1, we could analyze participants who learned at most 15 of the 24 possible mappings. However, this could produce a skewed measure of average learning performance in block 2 since we would be rejecting data from those who learned "too much" in block 1. To help compensate, we also excluded participants who learned less than 9 correct pairings. Thus, only participants who learned between 9 and 15 correct pairings in block 1 continued on to block 2 of either condition.

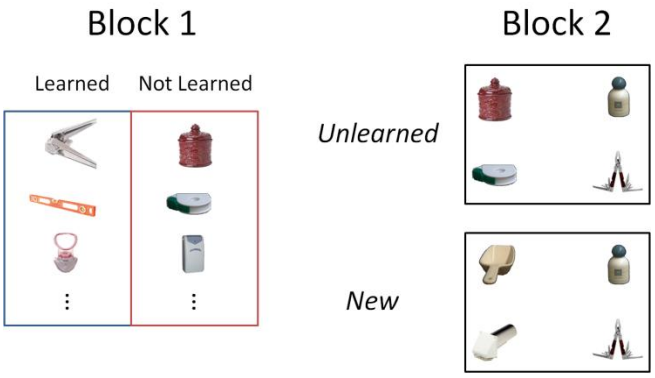


Figure 1: Selection of stimuli for block 2. In the *unlearned* condition, half of the items on each trial of block 2 were those for which the participant had given the incorrect response in block 1. The other half were new. In the *new* condition, all stimuli were new.

Results and Discussion

As described above, only a subset of the participants run in block 1 of either condition proceeded on to block 2.

Importantly, the proportion of words learned in block 1 did not differ between the selected subset and the set of all participants in the *unlearned* condition (M_s = .51, M_a = .50, t = .234, *n.s.*) nor in the *new* condition (M_s = .47, M_a= .43, t = .470, *n.s.*). Neither was there a significant difference between the proportion of words learned in block 1 by the selected participants in the *unlearned* vs. the *new* condition (M_u = .51, M_n= .47, t = 1.62, *n.s.*). This is to be expected given that block 1 was identical across conditions. Thus, all further analysis will be performed on selected participants.

In the second block, participants in both the *unlearned* and *new* conditions learned a significant proportion of word-object pairings (M_u = .56, t_u = 9.74, p < .001, M_n= .27, t_n= 6.62, p < .001, *chance* = .056). However, as shown in Figure 2, participants in the *unlearned* condition successfully mapped more than twice as many words to their correct referents as those in the *new* condition (t = 3.63, p = .01). Further, this benefit was not only for the 9 old pairings carried over from block 1 (M_u = .6, M_n = .26, t = 3.69, p < .001), but for the 9 new pairings as well (M_u = .51, M_n = .27, t = 2.48, p < .05).

Thus, partial knowledge of word-object pairings in block 1 allowed participants in the *unlearned* condition to learn significantly more mappings in block 2 than participants in the *new* condition. Further, the benefit was not just for the pairings for which participants had partial knowledge, but for novel pairings as well. This suggests that partial knowledge plays an active role in organizing cross-situational learning. Even though knowledge of word-object pairings was below threshold in block 1, it was sufficient to drive learning of novel pairings in block 2.

These findings provide initial support for the idea that sub-threshold knowledge of word-object mappings drives cross-situational learning. Partial knowledge of some pairs may influence the learning of other pairs on a trial-to-trial basis by constraining the pairs that are associated within a trial. An alternative explanation, however, is that participants in this experiment are benefitting from knowledge of which of the stimuli in block 2 had been seen previously in block 1. This could allow participants in block 2 of the *unlearned* condition to actively reduce the ambiguity of each training trial by mapping old words to old objects and new words to new objects. Some evidence for this second hypothesis comes from the errors made by participants in block 2 of the *unlearned* condition. When participants made errors in selecting referents for new words, they selected new referents at a probability significantly different from chance (M = .70, t = 3.73, p < .01, *chance* = .44). To provide further insights into the nature of the partial knowledge and its role in learning novel items, we constructed a new condition that was designed to assess the influence of sub-threshold mappings over and above possible knowledge of old/new.

Experiment 2

In Experiment 1 we tested the role of partial knowledge in word-referent mapping by exposing participants to two consecutive trials of cross-situational learning. Crucially, half of the pairings in block 2 were pairings for which participants failed to learn correct mappings in block 1. Learning results in block 2 showed that partial knowledge of these word-object pairings allowed participants to perform more than twice as well as participants exposed to a second block consisting of all new pairings. One possibility is that this benefit is entirely due to participants preferentially mapping old words to old objects and new words to new objects because they categorized them into two groups by mere exposure.

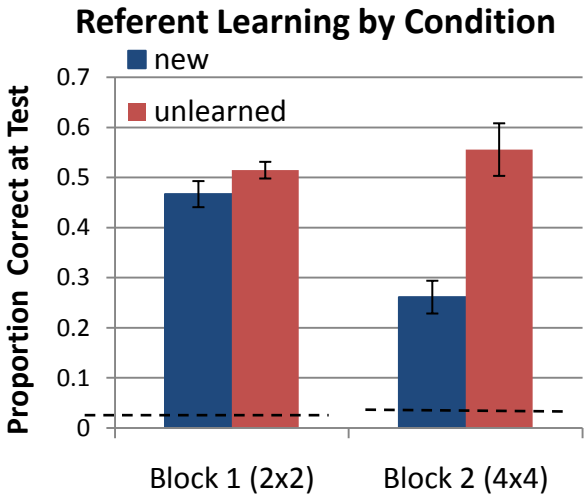


Figure 2: Proportion of word-referent pairings learned by participants in each condition across Blocks 1 and 2. Dotted lines indicate chance levels.

To establish a more stringent baseline for comparison, in Experiment 2 we constructed a *control* condition in which participants were exposed to the same word and object stimuli as participants in the *unlearned* condition, but without any opportunity to learn their associations. These same stimuli then appeared again in block 2. The *control* condition allows us to determine a second baseline – the effect of mere exposure to the stimuli of block 1.

Method

Participants. Ten Indiana University undergraduates participated in exchange for course credit. None had previously participated in Experiment 1.

Stimuli. Stimuli for Experiment 2 were identical to those for Experiment 1.

Procedure. The procedure for the *control* condition was similar to that used in the *unlearned* condition of Experiment 1. The crucial difference, however, was in the co-occurrence statistics of the words and objects of block 1. Whereas all words co-occurred with their correct referents 5 times in Experiment 1, in Experiment 2 half of the words occurred at most one time with each possible referent. Thus, there was essentially no correct referent for these 12 words. These unlearnable words and objects were matched for frequency of occurrence with those in block 1 – only co-occurrence statistics changed.

After the test phase of block 1, participants were exposed to a second cross-situational learning task as before. This time, however, 9 of the words and objects in block 2 were drawn randomly from the set of 12 unlearnable words and objects of block 1. In the second block these words each occurred 4 times with a single correct referent just like the 9 novel words. Thus, participants could distinguish the old words from the new words by their appearance in block 1, but they could not use potential partial knowledge of word-referent mappings to bootstrap their learning in block 2.

Results and Discussion

Because half of the words in block 1 of the *control* condition were unlearnable, it is unsurprising that these participants learned less words in block 1 than those in Experiment 1 (M_1 = .5, M_2 = .28, t_u = 7.21, p < .001). However, when only those words for which there was a correct answer in both Experiments are considered, participants performed equally well in both Experiment 1 and 2 (M_1 = .49, M_2 = .48, t = 0.19, *n.s.*). It is thus reasonable to compare block 2 performance across conditions.

In Experiment 2, we test the hypothesis that the benefit experienced due to participants in the *unlearned* condition of Experiment 1 was due not to partial knowledge, but to the ability to partition stimuli into two sets: old and new. If this is the case, mere exposure to the stimuli of block 1 – without the underlying co-occurrence statistics – should have been sufficient to reproduce this benefit. This is precisely the condition experienced by participants in the *control* condition. However, counter to this hypothesis, participants in the control condition did not outperform those in the *new* condition (M_c = .27, M_n = .26, t = .16, *n.s.*). They did, however, significantly underperform those in the unlearned condition (M_c = .27, M_u = .56, t = -3.22, p < .01). This difference was separately significant for old (M_c = .2, M_u = .6, t = -4.06, p < .001) and trending in the right direction for new (M_c = .34, M_u = .51, t = -1.55, p = .13) words. Figure 3 shows these results. This weighs against the hypothesis of mere exposure and lends credence to the hypothesis that partial knowledge is an active driver of cross-situational learning.

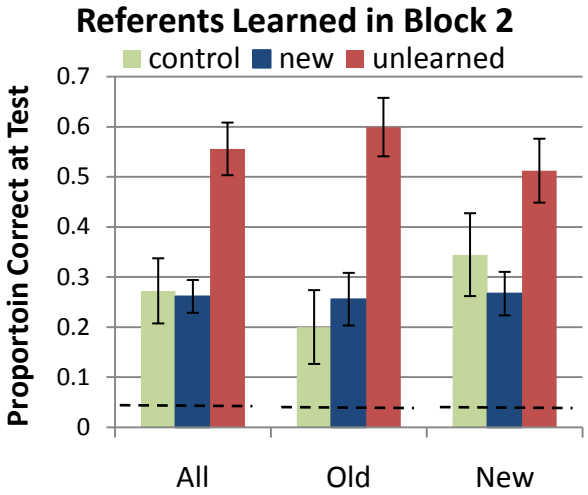


Figure 3: Proportion of word-referent pairings learned by participants in each condition. Old words are those which have been carried over to block 2 from block 1. In the *new* condition there are no old words, so the old words are those which fill the same slots in the training trials as the old words in the *unlearned* and *control* conditions. Dotted lines indicate chance.

Computational Models

To more fully analyze the role of partial knowledge of word-referent mappings in driving cross-situational learning, we implemented three incremental associative models that were exposed to simulated trials identical to those seen by experimental participants. The three models allow us to explicitly test hypotheses about how partial knowledge is used.

The first model – the *simple associative model* – maintains a word x object co-occurrence matrix and simply increments the cell corresponding to a word-object association each time the pair appears on a trial. This model thus learns the pure frequency of each of the possible word-object pairs.

The second model – the *biased associative model* – similarly maintains a word x object co-occurrence matrix. However, instead of incrementing the association strength

between a word and object by one whenever they co-occur, it increments their association by the strength of the current association. Whereas the *simple* model produces linear growth, the *biased* model produces geometric growth. This rich-get-richer scheme capitalizes on partial knowledge of a pairing in order to learn that pairing.

The final model – the *normalized associative model* – adds a competitive process to the *biased* model. On each trial, the increase in association between words and objects are computed as in the *biased* model, but the increment for a given word-object pair is normalized by the sum of all increments made for that object on that trial. This implements competition between all of the words in one trial. Intuitively, as one word accounts better for the presence of an object, the association between other words and that object are depressed. This mechanism is similar to the alignment mechanism used by Fazly et al.'s (in press) iterative version of the IBM Machine Translation Model (Brown, Pietra, Pietra, & Mercer, 1994).

The models are each tested for their knowledge of word-object associations in the same way as experimental participants. At the end of training, they are exposed to a series of alternative-forced choice tests and make their selections using the Shepard-Luce Choice Rule (Luce, 1959, Shepard, 1957). The simulated participant selects each alternative with a probability proportional to the exponential function of the strength of its association with the tested word.

Each model has only one parameter: a sensitivity parameter (λ) which weights each of the exponentiated probabilities in the Shepard-Luce Choice rule. Higher values of λ indicate that participants are more sensitive to differences in associative strengths between alternatives. To simulate Experiments 1 and 2, we exposed simulated participants to exactly the same stimuli as real participants. For instance, simulated participants in the *unlearned* condition were exposed to all of the training trials of the first block one at a time. Then, each simulated participant made selections at test using the Shepard-Luce Choice Rule. Nine of the items for which the model gave the wrong answer were then carried over to block 2, which were once again presented to the participant one trial at a time. Finally, the same decision rule was used to select a referent for each tested word. One thousand simulated participants were run in each of the three conditions using each model.

As can be seen in Figure 4, all of the models make essentially the same predictions for block 1. However, they make differing predictions for block 2 – the block during which partial knowledge may play a role. Figure 5 shows that the *simple associative* model is unable to produce the trend found in the data at even a qualitative level. It predicts that participants in the *unlearned* condition should underperform those in the *control* and *new* conditions. The other two models produce qualitatively similar trends. The *competitive* model, however, performs quantitatively better than the *biased* model (SSE_c = .0071, SSE_b = .0191, Bayes Factor = 2.69). As both have an equal number of

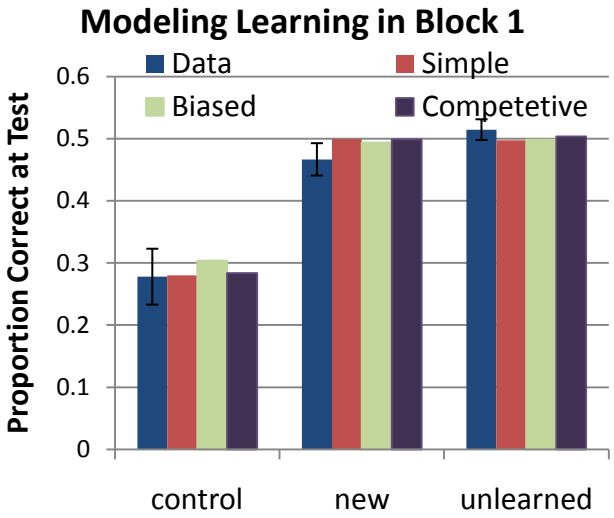


Figure 4: Proportion of word-referent pairs learned in block 1 by experimental participants and each of the three models across all experimental conditions.

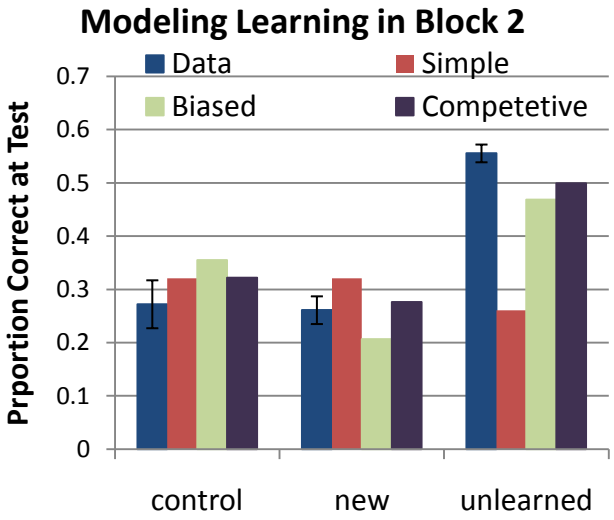


Figure 5: Proportion of word-referent pairs learned in Block 2 by experimental participants and each of the three models across all experimental conditions.

parameters, we can conclude that the *competitive model* is the better model for this empirical data. This supports the hypothesis that partial knowledge plays an active role in cross-situational learning, with partial-knowledge of multiple word-object associations interacting to support the acquisition of new word-object associations.

General Discussion

Whereas many methods for measuring word learning treat it as if it were binary – either the correct referent of a word is known or it is not – recent theoretical and computational models have argued that it is a gradual, accumulative process. Rather than learning a word's referent from a single perfect moment, learners may hone in on the correct referent through exposure to environmental statistics.

Empirical work has demonstrated that co-occurrence statistics alone are sufficient for learning word-object pairings (Yu & Smith, 2007). Furthermore, Vouloumanos (2008) showed evidence that learners are not only sensitive to the most frequently associated object for a given word, but also show deep knowledge of the statistical structure. Still, these results probed statistically acquired word-object knowledge only in its final state – producing a binary learned/unlearned data point for each potential pairing. Empirical evidence of graded states of partial knowledge has been indirect at best (Yurovsky & Yu, 2008).

The present work provides direct empirical evidence of not only the presence of such partial knowledge, but also its active role in driving word learning from exposure to exposure. The compared incremental models of statistical word learning show that partial knowledge may be leveraged on a trial-to-trial basis to bootstrap learning. Crucially, the better quantitative fits of the *competitive* model suggest that partial knowledge of a word-object association does not merely facilitate learning of that one association, but also combines with partial knowledge of other word-referent pairs to bootstrap learning of the whole system of words and referents. When words are learned as an interacting system, partial knowledge of one component gives a learner a leg up on acquiring others (Landuaer & Dumais, 1997).

While there is no denying the importance of word learning models at the computational level (Frank, Goodman, & Tenenbaum, 2009, Xu & Tenenbaum, 2007, Yu, 2008), this work again underscores our need to understand the continuous interaction of knowledge and learning on a moment-to-moment basis. Word learning is a constructive process, with initial successes cascading on themselves to empower even more successful learning (Smith, 1999). By digging deeper into word learning – understanding the latent representations that drive the system – we can hope to come to terms with its incredible complexity.

Acknowledgments

This work was supported by a National Science Foundation Graduate Research Fellowship to the first author and National Institute of Health Grant R01HD056029.

References

Brown, P. F., Pietra, S., Pietra, V., & Mercer, R. L. (1994). The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics, 19*, 263–311.

Blythe, R. A., Smith, K., & Smith, A. D. M. (in press). Learning times for large lexicons through cross-situational learning. To appear in *Cognitive Science*.

Ebbinghaus, H. (1913). Memory. A Contribution to Experimental Psychology. New York: Teachers College, Columbia University.

Fazly, A., Alishahi, A., & Stevenson, S. A probabilistic computational model of cross-situational word learning. To appear in *Cognitive Science*.

Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., & Pethick, S.J.(1994). Variability in early communicative development. *Monographs of the Society for Research in Child Development*, 59. Chicago: University of Chicago Press.

Frank, M. C., Goodman, N. D., & Tenenbaum, J. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science, 20*, 579-585.

Landauer, T. K. & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*, 211-240.

Luce, R. D. (1959). *Individual choice behavior*. New York: Wiley.

McMurray, B. (2007) Defusing the childhood vocabulary explosion. *Science, 317*, 631.

McMurray, B., Horst, J., Toscano, J., & Samuelson, L. (in press). Towards an integration of connectionist learning and dynamical systems processing: case studies in speech and lexical development. In J. Spencer, M. Thomas, & J. McClelland (Eds.), *Toward a new grand theory of development? Connectionism and Dynamic Systems Theory reconsidered*.

Shepard, R. N. (1957). Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika, 22*, 325-345.

Smith, L. B. (2000). Learning how to learn words: An associative crane. In R.M. Golinkoff, K. Hirsh-Pasek, L. Bloom, L. Smith, A. Woodward, N. Akhtar, M. Tomasello, & G. Hollich (Eds.), *Becoming a word learner: A debate on lexical acquisition* (pp. 51-80). New York, NY: Oxford Press.

Vouloumanos, A. (2008). Fine-grained sensitivity to statistical information in adult word learning. *Cognition, 107*, 729-742.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review, 114*, 245–272.

Yu, C. (2008). A Statistical Associative Account of Vocabulary Growth in Early Word Learning. *Language Learning and Acquisition, 4*, 32-62.

Yu, C. & Smith, L. B. (2007). Rapid Word Learning under Uncertainty via Cross-Situational Statistics. *Psychological Science, 18*, 414-420.

Yurovsky, D. & Yu, C. (2008). Mutual Exclusivity in Cross-Situational Statistical Learning. In B. C. Love, K. McRae, & V.M. Sloutsky (Eds.). *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 715-720). Austin, TX: Cognitive Science Society.

Cross-situational Learning of Low Frequency Words: The Role of Context Familiarity and Age of Exposure

Afsaneh Fazly, Fatemeh Ahmadi-Fakhr
Computer Sciences and Engineering
Shiraz University
Shiraz, Iran
{fazly,ahmadifakhr}@cse.shirazu.ac.ir

Afra Alishahi
Computational Linguistics and Phonetics
Saarland University
Saarbrücken, Germany
afra@coli.uni-saarland.de

Suzanne Stevenson
Computer Science
University of Toronto
Toronto, Canada
suzanne@cs.toronto.edu

Abstract

Higher frequency has been shown to have a positive effect on the acquisition of words and other linguistic items in children. An important question that needs to be answered then is how children learn low frequency items. In this study, we investigate the acquisition of meanings for low frequency words through computational modeling. We suggest that for such words, the familiarity of the context they appear in has an important effect on their acquisition. We note that context familiarity is confounded with another factor, namely the age of exposure to a word, and hence examine the independent role of each of the two factors on word learning.

Cross-situational Word Learning

Learning the meaning of words is a challenging task for young children, especially given that most words are learned from noisy and ambiguous contexts. Many specific word learning biases and constraints, as well as general learning mechanisms, have been suggested to be at work in the course of child lexical development. In particular, the learning of word–meaning mappings has been suggested to be based on cross-situational observation (Quine, 1960; Pinker, 1989) — that is, the meaning of a word can be learned by detecting the common set of meaning elements across all situations in which the word occurs. Psychological experiments on adults and children show that they are capable of learning word–referent mappings from their co-occurrences over time, even when each single occurrence of a word–referent pairing is ambiguous (Yu & Smith, 2007; Smith & Yu, 2007). The learning process seems to be sensitive to the statistical properties of the input, such as word frequency, the degree of ambiguity of presentation (i.e., how many words appear together), and the familiarity of the context. However, the relevant properties and their precise impact on learning word meanings are not well understood.

For example, higher frequency has been shown to have a positive effect on learning many linguistic constructions (Huttenlocher et al., 1991; Naigles & Hoff-Ginsberg, 1998). But frequency must be investigated carefully, since in many learning situations — including word learning — it may be confounded with other factors, such as the diversity of the context that a word can appear in (see, e.g., Kachergis, Yu, & Shiffrin, 2009). Moreover, many experimental studies have shown that children can acquire the meaning of a novel word with only one or a few exposures (i.e., when the word has very low frequency), especially if it is presented in a familiar context, a phenomenon known as *fast mapping* (Carey & Bartlett, 1978; Gershkoff-Stowe & Hahn, 2007; Alishahi et al., 2008). These observations about the interactions of various factors with frequency are especially important given that words in the input children receive have a Zipfian distribution (Zipf, 1949) — that is, a large proportion of words have a very low frequency of occurrence, yet generally they are successfully learned. Therefore, the relevant statistical features of the input data and their independent effect on learning need to be carefully investigated.

The experimental study of Kachergis et al. (2009) on adult subjects is one such attempt to identify and study the role of some of the important statistical properties of input on word learning. By varying the frequency of co-occurrence of different word–referent pairs, Kachergis et al. examine the independent and differential role of frequency, contextual diversity (i.e., diversity in the co-occurring words across usages of a target word), and within-trial ambiguity (i.e., the number of co-occurring words in each sentence) in cross-situational word learning. In particular, Kachergis et al. suggest that high contextual diversity and low within-trial ambiguity can boost the acquisition of low frequency words. However, some of the experimental results in their study cannot be explained by the factors they propose. Also, contextual diversity cannot explain children's ability to easily learn the referent of a novel word from a single exposure to that word, since it captures a property of the input across multiple exposures to a word. These observations suggest that other factors may be at play, especially for the acquisition of low frequency words.

Our goal is to investigate what other factors may have an effect on learning the meanings of words in general, and on the acquisition of low frequency words in particular. We use an existing computational model of cross-situational word learning to simulate the experiments of Kachergis et al. (2009), and to examine the effect of two additional factors in word learning: the familiarity of the context that a word appears in (i.e., how well the model/learner knows the other words in the sentence), and the age of exposure to a word. The computational simulations of our model show a matching behavioural pattern with that of adult word learners in the experiments of Kachergis et al. Moreover, our results suggest that for low frequency words, it is not the contextual diversity that helps learn their meaning, but the degree of familiarity of their context. We further test this claim by applying our model to a large corpus of child-directed speech, and examine the role of the proposed factors in the learning performance of the model.

Statistical Properties of the Input

Contextual Diversity and Within-trial Ambiguity

Kachergis et al. (2009) report a series of studies on adult subjects learning word–referent mappings from ambiguous utterance–scene pairs, where the utterance contains a bag of words, and the scene is the set of their referents. They investigate the effect of frequency by having some words and referents appear more often than others. They also investigate the interaction between word–referent frequencies and the diversity of the contexts that a word appears in, to examine the independent effect of each of the two factors on word learning. In these experiments, contextual diversity is varied by manipulating either the overall rate of co-occurrence among words, or the number of co-occurring words within a trial. More precisely, Kachergis et al. study the interactions among the following three factors:

- Word frequency:
  F(w) = total #occurrences of w in the input
- Contextual diversity:
  CD(w) = total #words co-occurring with w across all usages of w in the input
- Within-trial ambiguity:
  WA(w) = mean #words co-occurring with w in each utterance

Their results show that a higher F often leads to better learning of a word, and to boosting the learning of other words. However, F is usually confounded with CD, which can be seen as an alternative explanation for learning facilitation. When F is controlled for, a higher CD improves learning (i.e., more word–meaning pairs are learned), whereas a higher WA harms learning. Similarly when CD is controlled for, a higher F improves learning. Most interestingly, when a higher CD is achieved by interleaving high frequency words in the presentation of low frequency words, the learning of the low frequency words is improved.

Age of Exposure and Context Familiarity

As described in the previous section, the results of Kachergis et al. (2009) suggest that contextual diversity (CD) is particularly important for the acquisition of low frequency words. However, some of their results show a boost in the acquisition of low frequency words where there is no notable difference in CD. In an attempt to explain these results, and inspired by the well-studied fast mapping effect (Carey & Bartlett, 1978), we study two additional statistical factors that might play a role in cross-situational learning:

- Age of exposure:
  AE(w) = time at which w first appears in the input
- Context familiarity:
  CF(w) = mean *familiarity* of words co-occurring with w, averaged across all usages of w in the input

where *familiarity* of a word is determined by its frequency of occurrence prior to its current appearance.

Computational Analysis

We investigate the role of each of the above factors in cross-situational learning through two sets of experiments. First, we replicate the results of Kachergis et al. (2009) using the computational model of Fazly et al. (2008) (briefly explained in the next section), and examine the impact of our proposed factors as well as the ones proposed by Kachergis et al. on learning. Second, we apply our model on a larger corpus of actual child-directed speech to better understand how the model learns the meaning of low frequency words in a more naturalistic situation, and to study the impact of the statistical factors and their interaction during the course of learning.

Overview of the Computational Model

We use an incremental probabilistic word learning algorithm, explained in full detail in Fazly et al. (n.d.). Here we repeat a brief explanation of how the model works.

Utterance and Meaning Representations

The input to our word learning model consists of a set of utterance–scene pairs that link an observed scene (what the child perceives) to the utterance that describes it (what the child hears). We represent each utterance as a set of words, and the corresponding scene as a set of meaning symbols.

Utterance: { *Joe*, *rolled*, *the*, *ball* }
Scene: {`joe`, `roll`, `the`, `ball`}

Given a corpus of such utterance–scene pairs, our model learns the meaning of each word w as a probability distribution, $p(.|w)$, over the semantic symbols appearing in the corpus. In this representation, $p(m|w)$ is the probability of a symbol m being the meaning of a word w. We assume that in the absence of any prior knowledge, all symbols are equally likely to be the meaning of a word. Hence, prior to receiving any usages of a given word, the model assumes a uniform distribution over all semantic symbols as its meaning.

Meaning Probabilities

Our model combines probabilistic interpretations of cross-situational learning (Quine, 1960) and a variation of the principle of contrast (Clark, 1990), through an interaction between two types of probabilistic knowledge acquired and refined over time. Given an utterance–scene pair received at time t, i.e., $(\mathrm{U}^{(t)}, \mathrm{S}^{(t)})$, the model first calculates an alignment probability a for each $w \in \mathrm{U}^{(t)}$ and each $m \in \mathrm{S}^{(t)}$, using the meaning probabilities $p(.|w)$ of all the words in the utterance prior to this time (Step 1 below). The model then revises the meaning of the words in $\mathrm{U}^{(t)}$ by incorporating the alignment probabilities for the current input pair (Step 2). This process is repeated for all the input pairs, one at a time.

Step 1: Calculating the alignment probabilities. We estimate the alignment probabilities of words and meaning symbols based on a localized version of the principle of contrast: that a meaning symbol in a scene is likely to be highly associated with only one of the words *in the corresponding utterance*. For a symbol $m \in \mathrm{S}^{(t)}$ and a word $w \in \mathrm{U}^{(t)}$, the

higher the probability of m being the meaning of w (according to $p(m|w)$), the more likely it is that m is aligned with w in the current input. In other words, $a(w|m, \mathrm{U}^{(t)}, \mathrm{S}^{(t)})$ is proportional to $p^{(t-1)}(m|w)$. In addition, if there is strong evidence that m is the meaning of another word in $\mathrm{U}^{(t)}$ — i.e., if $p^{(t-1)}(m|w')$ is high for some $w' \in \mathrm{U}^{(t)}$ other than w — the likelihood of aligning m to w should decrease. Combining these two requirements:

$$a(w|m, \mathrm{U}^{(t)}, \mathrm{S}^{(t)}) = \frac{p^{(t-1)}(m|w)}{\sum_{w' \in \mathrm{U}^{(t)}} p^{(t-1)}(m|w')} \quad (1)$$

Step 2: Updating the word meanings. We need to update the probabilities $p(.|w)$ for all words $w \in \mathrm{U}^{(t)}$, based on the evidence from the current input pair reflected in the alignment probabilities. We thus add the current alignment probabilities for w and the symbols $m \in \mathrm{S}^{(t)}$ to the accumulated evidence from prior co-occurrences of w and m. We summarize this cross-situational evidence in the form of an association score, which is updated incrementally:

$$\mathrm{assoc}^{(t)}(w, m) = \mathrm{assoc}^{(t-1)}(w, m) + a(w|m, \mathrm{U}^{(t)}, \mathrm{S}^{(t)}) \quad (2)$$

where $\mathrm{assoc}^{(t-1)}(w, m)$ is zero if w and m have not co-occurred before. The association score of a word and a symbol is basically a weighted sum of their co-occurrence counts.

The model then uses these association scores to update the meaning of the words in the current input:

$$p^{(t)}(m|w) = \frac{\mathrm{assoc}^{(t)}(m, w)}{\sum_{m_j \in \mathcal{M}} \mathrm{assoc}^{(t)}(m_j, w)} \quad (3)$$

where $\mathcal{M}$ is the set of all symbols encountered prior to or at time t. We use a smoothed version of the above formula, as described in (Fazly et al., n.d.).

Word Comprehension Score

Our model updates the meaning of a word every time it is heard in an utterance. The strength of learning of a word at time t is reflected in $p^{(t)}(m = m_w|w)$, where m_w is the "correct" meaning of w according to a gold-standard lexicon. We refer to $p^{(t)}(m_w|w)$ as the comprehension score (Comp) of word w at time t. Ideally, a word is accurately learned when the probability distribution $p(.|w)$ is highly skewed towards the correct meaning m_w. In our experiments reported in the following sections, we first train our model on a number of utterance–scene pairs, and then examine the comprehension scores of words as an indirect way of measuring the performance of our model in selecting referents of words.

Analysis of Artificial Word Learning Data

Here we report the results of our simulations on artificially-generated data similar to that of Kachergis et al. (2009). Their (human) experiments examine the effect of three factors on word learning: frequency (F), contextual diversity (CD), and within-trial ambiguity (WA), as defined on page 2. The artificial input data set used in our simulations is explained next, and then the results of the experiments are presented.

Input Data

The artificial data set consists of randomly-generated sequences of utterances in the form of an unordered bag of novel words, each paired with a set of novel meaning symbols. In the artificial data, one of the three factors under study is changed while the other factors are kept constant, in order to better understand the role each plays in learning, as well as the interactions among the different factors. We use nine sets of artificial data (each containing 18 word–meaning pairs), one set for each experimental condition of Kachergis et al. (2009). The first experiment investigates the role of F: one condition divides words into two frequency groups (F=3,9), the other into three frequency groups (F=3,6,9). The second experiment examines the role of context by manipulating either CD or WA, while keeping F constant. One condition manipulates CD by dividing words into two unequal-sized groups (with 6 and 12 words, respectively), and allowing words in each group to co-occur only with other words from the same group. In two other conditions, a word appears with either 2 or 3 other words in each trial (WA=3 and WA=4, respectively). The third experiment studies the interaction between F and CD by controlling the co-occurrence among words from three frequency groups (F=3,6,9), resulting in four conditions: In Low CD condition, words from each frequency group co-occur only with other words from the same group. In Med CD conditions, low frequency words (F=3) are allowed to either co-occur with words in F=6 (Med CD-3&6), or with those in F=9 (Med CD-3&9). In High CD condition, there is no restriction on the co-occurrence of words from different frequency groups. For each experimental condition, we randomly generate 30 different artificial input. Results presented here are averages over 30 different simulations, each using a different input.

Modeling Effects of Frequency and CD

Figures 1 to 3 present the performance of our model on the artificially-generated input in three experiments analogous to those of Kachergis et al. (2009).

Our findings in Experiment 1 (Figure 1) are generally in line with those of Kachergis et al. (2009): that higher frequency does not seem to have a consistently positive effect on word learning. As noted by Kachergis et al., frequency might be conflated with other factors, and thus we cannot make a decisive conclusion only on the basis of this experiment.

Figure 2 (left half) shows that contextual diversity (CD) has a significant positive effect on word learning ($p \ll .001$).[1] Figure 2 (right half) shows that an increased WA has an adverse effect on word learning, even though it also increases CD (difference is significant; $p \ll .001$).

Recall that in Experiment 3 the interaction between CD and F is examined by looking at the learning performance of low (F=3), medium (F=6), and high (F=9) frequency words in

[1] All statistical significance tests reported in this paper are for paired t-tests with a 95% confidence interval, and are performed using the R statistics package (`http://www.r-project.org`).

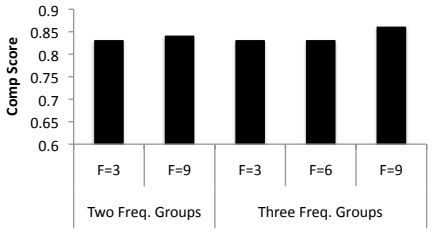


Figure 1: Average Comp scores for words from different frequency ranges (Experiment 1).

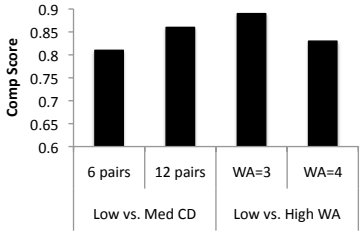


Figure 2: Average Comp scores for words with different contextual diversity (CD), or within-trial ambiguity (WA) (Experiment 2).

conditions with varying degrees of CD. The overall pattern of our results across the four conditions of this experiment, presented in Figure 3, matches those reported by Kachergis et al. for humans. Specifically, we note that, as for human subjects, the overall learning is significantly greater for our model in the High CD condition ($p \ll .001$). Moreover, we observe that, similar to human behaviour, the acquisition of low frequency words in our model is better when they are allowed to co-occur with higher frequency words (Conditions: High CD, Med CD-3&6, and Med CD-3&9); differences between each of these three conditions and the Low CD condition are statistically significant ($p \ll .001$). Whereas Kachergis et al. attribute this behaviour to an increased CD, we suggest that there is another factor (namely context familiarity), which is responsible for this boost of performance in the acquisition of low frequency words.

Context Familiarity as the Explanatory Factor

As discussed above, Kachergis et al. (2009) suggest that contextual diversity is especially important for the acquisition of low frequency words. However, there are cases (in our experiments and in those of Kachergis et al.) where we see a boost in the acquisition of low frequency words, with no notable difference in CD. Instead, as we show now, differences in context familiarity (CF) can explain the pattern of results.

Consider again the results of Experiment 3 shown in Figure 3. We also summarize some properties of the input in the four conditions of that experiment in Figure 4. For each condition we select one simulation such that the overall pattern of results (e.g., with respect to the learning of low frequency words) for these simulations match that of the average performance given in Figure 3. For each input used in the selected simulations, we then calculate the average CD and CF values for words in each of three frequency groups (i.e., F=3,6,9). To

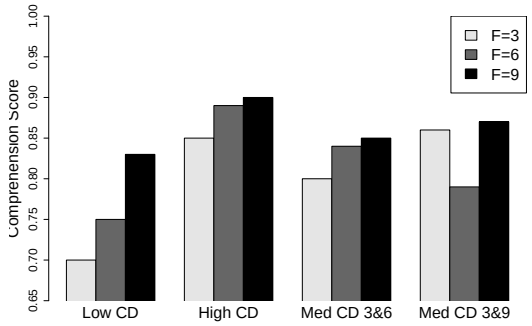


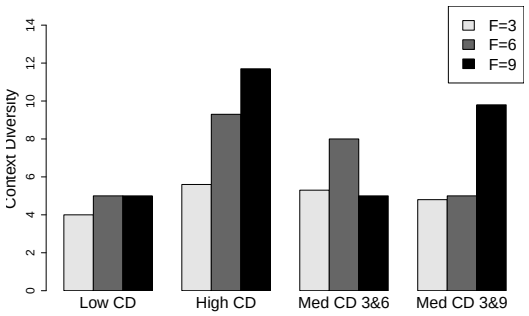
Figure 3: Average Comp scores for low, medium, and high frequency words (left to right bars) within each of the four conditions of contextual diversity (Experiment 3).

calculate CF for a word in an utterance, we set the familiarity of each co-occurring word to its frequency of occurrence prior to the current appearance.

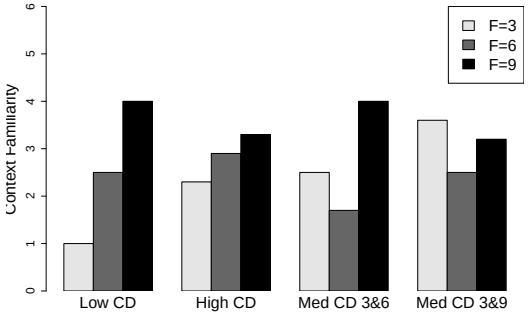
Figure 3 shows that the highest performance in learning of low frequency words is achieved when these words are allowed to co-occur with words with F=9 (Condition: Med CD-3&9). Kachergis et al. attribute this high learning performance to an increased CD. There is indeed *overall* a higher CD value for the High and Med CD conditions. However, when they are separated by the frequency of the items as in Figure 4(a), we observe that the CD values for the low frequency words do not substantially change across the four conditions. (The overall pattern of CD values in Figure 4(a) match very closely those presented in Table 3 on page 5 of Kachergis et al. CD values range from 4 to 5.6 in our experiments, and from 4 to 5.5 in those of Kachergis et al.)

Interestingly, however, if we look at the pattern of the CF values in Figure 4(b), we see that it conforms to the learning performance of our model (and those of humans) on low frequency words (compare the lightest bars in each of the two figures across the four conditions). We can explain this effect of CF in the learning of our model as follows: When low frequency words are allowed to co-occur with words with F=9, we expect the contexts of the low frequency words to be, on average, more familiar than in other conditions. Since our model is expected to have learned something about the possible meanings of a familiar word, this in turn decreases the degree of ambiguity in an utterance–scene pair, making the acquisition of a novel low frequency word easier. This result is a direct consequence of the interactions between the two sets of probabilities accumulated over time in our model, namely the alignment and the meaning probabilities. When aligning a word in an utterance to a referent/meaning in the corresponding scene, our model uses its acquired knowledge about the meaning of the co-occurring words (according to the meaning probabilities). The more familiar the co-occurring words are, the more reliable the meaning probabilities for these words will be, and this in turn makes it easier for the model to align the target word to its correct referent.

For higher frequency words (F=6 and F=9), we can see a



(a) Average CD for different frequency ranges across different conditions



(b) Average CF for different frequency ranges across different conditions

Figure 4: Average CD and CF values for low, medium, and high frequency words (left to right bars) within each of the four conditions of contextual diversity (Experiment 3).

clear effect of CD (compare darkest bars in Figure 3 and Figure 4(a)). These results together suggest that CD positively affects the learning of medium and high frequency words (as also noted by Kachergis et al.), but for low frequency words it is the familiarity of the context that is the key factor in their acquisition (in contrast to the suggestion of Kachergis et al.).

Analysis of Naturalistic Child-directed Speech

As noted previously, children are exposed to a great number of low frequency words in the input they receive (due to the Zipfian distribution of words in a language). We thus further investigate the effects of context familiarity (CF) on the acquisition of low frequency words in a more naturalistic setting, by performing experiments on a large corpus of actual child-directed speech. The child-directed corpus and the details of the experiments are explained below.

Input Data

The child-directed corpus consists of 10,000 utterance–meaning pairs, where the utterances are taken from the Manchester corpus (Theakston et al., 2001) in the CHILDES database (MacWhinney, 2000), and the corresponding meanings are artificially generated by including a distinct meaning symbol for each word in the utterance. Our focus in the following experiments is on F, CD, CF, and another factor usually confounded with CF, namely age of exposure to a word or AE. We control for the effect of within-trial ambiguity (WA) in our experiments by considering only those utterances whose length is between 5 and 7 (inclusive).

Table 1: Average frequency (F), CD, Comp, CF, and AE values for two groups of low frequency words: High Comp vs. Low Comp. Number of words in each group is given in parenthesis.

	High Comp Comp ≥ 0.9 (877)	Low Comp Comp < 0.9 (258)
F	1.50 ± 0.73	1.49 ± 0.73
CD	6.61 ± 2.82	6.70 ± 2.77
CF	4.64 ± 0.40	3.58 ± 0.62
AE	9.39 ± 5.55	6.36 ± 6.19
Comp	0.93 ± 0.02	0.52 ± 0.15

We measure the factors CD, CF and AE for each word according to the definitions on page 2. Here we measure the *familiarity* of a word slightly differently from on the artificial data. Since the frequency of words in the child-directed corpus is on a different scale and varies a lot, we set familiarity of a word to a value between 0 and 5 according to the frequency range it belongs to. The mappings between familiarity values and frequency ranges are: 0 (0), 1 (1), 2 (2–4), 3 (5–9), 4 (10–29), and 5 (≥ 30), where the numbers in parentheses specify frequency ranges. Similarly, we re-scale AE for a word to be the sequence number of the utterance in which the word is encountered for the first time, divided by 500 (e.g., all words in utterances 1 to 499 will have an AE of 0).

Modeling Effects of Context Familiarity

AE has been identified as an important factor in word learning (Carey & Bartlett, 1978; Gershkoff-Stowe & Hahn, 2007). However, it is usually confounded with CF since a later AE entails that there are generally more familiar words in the input. It is thus important to examine the independent role of CF and AE on word learning, as we see below.

After training our model on the 10,000 utterances in our child-directed corpus, we divide low frequency words (those with F < 4) into two groups according to how well they are learned: one group with a high comprehension score (Comp ≥ 0.9), and another group with a lower comprehension score (Comp < 0.9). Table 1 summarizes the averages of the different factors for the two groups. Interestingly, although F and CD are similar for both groups, we observe a substantial difference in the average Comp scores (0.93 vs. 0.52), suggesting that a factor other than F and CD must be responsible for this difference in learning. Looking at CF and AE, we can see an effect for both: words that have a high Comp score also tend to have higher CF and AE. That is, the words that are learned more confidently are those that have occurred in contexts with greater familiarity and that are first seen at a later age (i.e., when more words have been learned).

We now examine the independent effects of CF and AE on the acquisition of low frequency words, by holding one fac-

Table 2: Average AE, CF, and Comp for two groups of low frequency words: High CF vs. Low CF when AE is held constant (top part); and High AE vs. Low AE, when CF is held constant (bottom part). Number of words in each group is given in parenthesis.

	High CF	Low CF
	CF $\geq$ 4.5 (313)	CF $<$ 4.5 (160)
AE	9.90 $\pm$ 2.60	9.68 $\pm$ 2.49
CF	4.84 $\pm$ 0.17	3.98 $\pm$ 0.38
Comp	0.93 $\pm$ 0.02	0.77 $\pm$ 0.22

	High AE	Low AE
	AE $\geq$ 9 (78)	AE $<$ 9 (143)
CF	3.50 $\pm$ 0.38	3.43 $\pm$ 0.41
AE	13.62 $\pm$ 2.95	2.15 $\pm$ 2.22
Comp	0.60 $\pm$ 0.20	0.62 $\pm$ 0.21

tor constant (fixed within a range), and looking at the effect of the other factor. First, we consider low frequency words with AE values within a fixed range (here $5 < \text{AE} < 15$), and divide them into two groups based on their CF (Table 2: top part). Second, we hold CF constant within a fixed range ($2 < \text{CF} < 4$), and divide words into two groups with high and low AE (Table 2: bottom part). (Note that F and CD are the same for the two groups in both conditions.) We find that words that have occurred with differing CF values (top of Table 2) show an effect on their Comp score, with much better learning when the context familiarity is higher. On the other hand, words that have occurred with differing AE values (but with similar CF; bottom of Table 2) show no difference in learning at the different ages of exposure. These results show that CF has an independent and positive effect on the acquisition of low frequency words, whereas AE does not. We suggest that the effect we previously observed for AE (Table 1) is mostly through its effect on CF: since the model/learner learns more and more words over time, words encountered later (with higher AE) are in general more likely to appear with other familiar words, and thus to have a higher CF.

Conclusions

We have used an incremental probabilistic model of cross-situational word learning to study the effects of various statistical properties of the input on the acquisition of low frequency words. This is especially important since a large proportion of words in the input children receive have a very low frequency of occurrence. Replicating the results of a set of psychological experiments on artificial word learning (Kachergis et al., 2009), we argue that different factors affect the acquisition of high and low frequency words. These results and our findings through further experiments on natural child-directed utterances suggest that, for medium and high frequency words, the diversity in the context has a positive effect on learning (as also noted by Kachergis et al.), whereas for low frequency words it is the familiarity of the context that greatly impacts their acquisition.

These effects can be explained as a natural consequence of the interactions between two sets of probabilities that our model acquires over time. Through these interactions, our model draws on its own acquired knowledge of word meanings to boost the learning of other (novel) words. Thus, the acquisition of a set of high frequency words helps learn low frequency words by increasing their context familiarity. Generally, our model learns word meanings by drawing on the statistical regularities found in the input, and without incorporating any specific word learning biases or constraints, thus making the model appropriate for conducting studies on the relation between input properties and word learning.

References

Alishahi, A., Fazly, A., & Stevenson, S. (2008). Fast mapping in word learning: What probabilities tell us. In *Proceedings of CoNLL'08* (pp. 57–64).

Carey, S., & Bartlett, E. (1978). Acquiring a single new word. *Papers and reports on Child Lang. Dev.*, *15*, 17–29.

Clark, E. V. (1990). On the pragmatics of contrast. *Journal of Child Language*, *17*, 417–431.

Fazly, A., Alishahi, A., & Stevenson, S. (n.d.). A probabilistic computational model of cross-situational word learning. *Cognitive Science: An Interdisciplinary Journal*.

Fazly, A., Alishahi, A., & Stevenson, S. (2008). A probabilistic incremental model of word learning in the presence of referential uncertainty. In *Proceedings of CogSci'08*.

Gershkoff-Stowe, L., & Hahn, E. R. (2007). Fast mapping skills in the developing lexicon. *Journal of Speech, Language, and Hearing Research*, *50*, 682–697.

Huttenlocher, J., Haight, W., Bryk, A., Seltzer, M., & Lyons, T. (1991). Early vocabulary growth: Relation to language input and gender. *Developmental Psych.*, *27*(2), 236–248.

Kachergis, G., Yu, C., & Shiffrin, R. M. (2009). Frequency and contextual diversity effects in cross-situational word learning. In *Proceedings of CogSci'09*.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk* (3rd ed., Vol. 2: The Database).

Naigles, L., & Hoff-Ginsberg, E. (1998). Why are some verbs learned before other verbs? effects of input frequency and structure on children's early verb use. *Journal of Child Language*, *25*, 95–120.

Pinker, S. (1989). *Learnability and cognition: The acquisition of argument structure*. Cambridge, MA: MIT Press.

Quine, W. (1960). *Word and object*. The MIT Press.

Smith, L. B., & Yu, C. (2007). Infants rapidly learn words from noisy data via cross-situational statistics. In *Proceedings of CogSci'07*.

Theakston, A. L., Lieven, E. V., Pine, J. M., & Rowland, C. F. (2001). The role of performance limitations in the acquisition of verb–argument structure: An alternative account. *Journal of Child Language*, *28*, 127–152.

Yu, C., & Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, *18*(5), 414–420.

Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Cambridge, MA: Addison-Wesley.

Object and Word Familiarization Differentially Boost Retention in Fast-Mapping

Sarah C. Kucker (Sarah-Kucker@uiowa.edu)
Department of Psychology, E11 Seashore Hall
Iowa City, IA 52242 USA

Larissa K. Samuelson (Larissa-Samuelson@uiowa.edu)
Department of Psychology, E11 Seashore Hall
Iowa City, IA 52242 USA

Abstract

Recent research demonstrated that although twenty-four month-old infants do well on the initial pairing of a novel word and novel object in fast-mapping tasks, they are unable to retain the mapping after a five-minute break. The current study examines the role of familiarity with the objects and words on children's retention in fast-mapping tasks. Twenty-four month-old infants were familiarized with either a series of novel objects or a series of novel names prior to the referent selection portion of a fast-mapping task. Infants retained the novel mapping after a delay when familiarized with the novel objects, but did not demonstrate retention when familiarized with the novel words. The results suggest familiarity with the object or word-form lead to differential encoding of the name-object link and altered subsequent word learning.

Keywords: language acquisition; fast-mapping; word learning

Introduction

Fast-mapping, or the ability to quickly link a novel word to a novel referent is perhaps the canonical example of young children's word learning prowess. In Carey's (1978) original demonstration of this phenomenon, preschoolers correctly determined that the novel word "chromium" referred to a novel olive-green colored tray rather than a familiar blue-colored tray. This result has been replicated and extended many times (see, for instance, Golinkoff, Hirsch-Pasek, Bailey, & Wenger, 1992; Mervis & Bertrand, 1994; Wilkinson, Ross, & Diamond, 2003). Fast-mapping has been demonstrated in infants as young as 17 months (Halberda, 2003), and 30-month-olds have been shown to fast-map as many as six novel items in a single session (Golinkoff et al, 1992). On the basis of such results, there has been a general tendency in the literature to equate fast-mapping and word learning and to see fast-mapping as the basis for children's rapid word learning (see Horst & Samuelson, 2008 for discussion). However, retention has only rarely been examined in fast-mapping paradigms, and recent work suggests that children do not retain the links between a novel name and object formed in these tasks.

Horst and Samuelson (2008) examined retention of name-object links presented in a fast-mapping context with 24-month-old infants. Their fast-mapping paradigm included both referent selection and retention trials. On referent selection trials, infants were presented with two known objects ("get the block"); on other trials, infants were asked for the novel object ("get the roke"). On retention trials, which followed five minutes after referent selection, infants were presented with two objects that had been fast-mapped in the referent selection trials, and a third, previously seen but not named object. During these trials, infants were asked to get one of the previously fast-mapped objects by name. Because all three objects presented on retention trials were equally novel, Horst and Samuelson's task is very stringent. In this carefully controlled environment, infants performed well in the referent selection trials – choosing the known object 73% of the time when requested, and the novel object 69% of the time it was requested. However, retention of the fast-mapped name-object link was no higher than chance after the 5-minute delay (Horst & Samuelson, 2008; Experiment 1A).

The fact that the children in Horst and Samuelson's (2008) study did not retain the newly fast-mapped words contradicts some prior findings of retention following fast mapping. For example, Carey & Bartlett (1978) examined children's memory for "chromium" a week after the original presentation and found that the majority of children retained the link between the word "chromium" and some form of the color green. Likewise, Markson and Bloom (1997) demonstrated retention of novel fast-mapped words in 3-4 year-old children. However, as Horst and Samuelson (2008) point out, many of these prior studies did not use as stringent a measure of retention. For example, Carey and Bartlett (1978) presented the novel name and referent during a very familiar sequence of events (setting up for snack time), thus allowing for many possible contextual supports for retention. Other work demonstrating retention has isolated the target so that it is the only object named during test (Markson & Bloom, 1997) or used ostensive naming in conjunction with fast-mapping (Mervis & Bertrand, 1994), thus failing to provide a stringent test of retention.

Furthermore, Horst and Samuelson's data does fit with Carey's (1978) original proposal of a slow-mapping process that follows the initial fast-mapping of a word to an object. In particular, Carey proposed that after children initially map the novel object and name (fast-mapping), further experience and exposure is required to fully learn the new word and referent (slow-mapping). Subsequent studies have examined this slow-mapping process, demonstrating that depth of semantic representation (Capone & McGregor, 2005), lexical practice (Gershkoff-Stowe, 2002; Gershkoff-Stowe & Hahn, 2007), and word segmentation (Graf Estes,

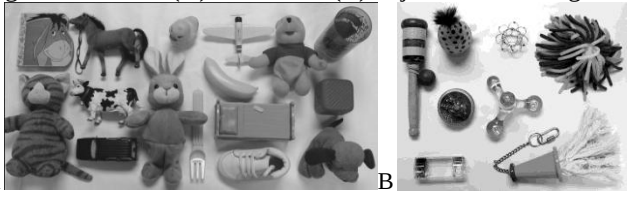
Evans, Alibali, & Saffran, 2007) all play a role in successful word retention and retrieval. The current study follows this line of work, examining the role of prior familiarity with the components of the mapping on retention of newly fast-mapped words.

It seems like familiarity with the components to be mapped may aid children's formation of a lasting association between a novel word and object by aiding in the creation a fairly robust, stable representation of each component. Horst, Samuelson, and McMurray (under review), have recently demonstrated that visual familiarity influences the process of referent selection. Likewise, Capone and McGregor (2005) demonstrated that ostensively highlighting the visual properties of objects (i.e. cueing shape) boosts infants' fast-mapping of object labels and their referents, whereas Graf Estes et al. (2007) demonstrated that statistical segmentation of auditory word forms can play a role in subsequent referent selection.

In the present experiments we used a stringent version of the standard forced-choice referent selection and retention task, modeled after Horst and Samuelson (2008), but added a minimal familiarization period prior to the referent selection task. We used the 3-trial version of Horst and Samuelson's task (2008, Experiment 1C) to reduce the chance of fatigue that might be caused by the time added by the familiarization period. Using this procedure, Horst and Samuelson (2008) found that only 60% of infants in their 3-trial experiment succeeded in the initial mapping of the name and object during referent selection. While this was a statistically significant level of mapping, it means that retention could only be tested in 12 infants. We found similar levels of mapping in pilot testing. Thus, in an effort to boost the number of infants who initially map the novel word to the novel object, we used the same three known items throughout the warm-up and referent selection trials (rather than using different known objects on each of the referent selection trials).

In Experiment 1 we examined the role of minimal familiarity with the objects or word-forms in infants' retention of fast-mapped words. Half the infants were given the novel objects to explore freely for two minutes prior to the referent selection task. The other half of the infants heard the novel word multiple times prior to the referent selection trials. As in Horst and Samuelson (2008), there was a five minute delay between the referent selection and retention trials. Only infants familiarized with the objects demonstrated significant retention. Experiments 2 and 3 serve as controls to ensure our findings were not due to our use of the same known objects on all referent selection trials (Experiment 2), or the use of a highly salient favorite novel item as the target (Experiment 3). Taken together, then, these experiments probe the degree to which prior encoding of either the word or object boosts the retention of fast-mapped words.

Figure 1: Known (A) and Novel (B) objects used during test



Experiment 1: Object and Word Familiarization

Methods

Participants Forty 24-month-old-infants (20 girls, M = 24 months, 26 days; range = 24 months, 10 days – 25 months, 13 days) with a mean vocabulary of 303 words (range = 21-672) participated. All infants were recruited through county birth records and were native English speakers. Participant's parents provided informed consent prior to the start of the study. Participants received a small toy for participation.

Stimuli Each infant saw a random selection of three out of sixteen possible known items and three or six of eight possible novel items over the course of the experiment (see Figure 1). Parents confirmed the status of each object as known or novel prior to the experiment. Substitute items were used if the infant was unfamiliar with any of the known items or familiar with any of the novel items. During the session, stimuli were presented on a 24x45cm white tray divided into three equal sections. Up to six possible novel non-words (Horst & Samuelson, 2008) were randomly selection for use with each child.

Procedure During the study, infants were seated across a white table from the experimenter in a booster seat next to their parents or in their parent's lab. Parents completed the MacArthur-Bates Communicative Development Inventory: Words and Sentences (MCDI; Fenson et al, 1994) during the session and were instructed to avoid interacting with their child, only offering encouragement if necessary.

Pre-familiarization. Half the children began the session with a one minute familiarization period with six novel objects. The experimenter drew the infant's attention to each object by picking it up or by pointing to it and saying "Look." Once the infant had explored each object, the experimenter lined all six items along the middle of the table and asked the infant to pick their favorite item. The favorite item was then given back to the infant to explore briefly. This was repeated twice more and the remaining three non-favorites were then removed from the table. The three favorite items were then used as the novel objects in the experiment with the favorite item selected first being the target during the novel referent selection trial.

The other half of the children began the session with a familiarization period in which they were exposed to six possible novel non-words. A 19-inch, 1280x1024 pixel touch screen computer was presented on the table approximately 24 cm in front of the child. The computer screen showed six 241x241 pixel basic shapes (i.e. circle, triangle, diamond, cross, square, octagon) in six different

basic colors (i.e. red, purple, orange, green, yellow, blue) in a 2x3 matrix with each item roughly 130 pixels away from each other. The trial began when the buttons appeared on the screen. The experimenter directed the infant's attention to the screen saying "Look! You can push the buttons!" and then randomly touched a button, producing one of six possible novel words. This was repeated until all buttons had been pushed and thus all six of the novel words were produced one time each. The experimenter then directed the infant to push the buttons by asking "Can you push the buttons?" If the infant did not respond, the experimenter again demonstrated by randomly pushing each button once. If the infant again did not respond, the experimenter demonstrated the buttons a third time and encouraged the infant to try themselves. At this point, if the infant refused to push the buttons themselves, the experimenter then randomly chose a button and pushed it multiple times to familiarize the infant with one of the six novel names. There were eight possible examples of each novel word varying in intonation, pitch, and frequency, which were randomly selected from at each button press. All words were spoken by the same female experimenter who was running the experimental session. After two minutes of familiarization with the sounds, the computer was removed and the experiment continued with the warm-up trials. The novel name that was produced the most during the familiarization period was used as the target name during the novel referent selection trial.

Warm-up trials. For each infant, three known objects were randomly chosen for use throughout the warm-up trials. The experiment placed each of the items in a slot on the tray, keeping the tray out of sight of the infant. The trial began with the experimenter placing the tray on the table and allowing the infant to examine the objects for three seconds. The experimenter then asked the infant to get an object ("Can you get the block") and slid the tray forward. Infants were prompted up to three times until a response was given. Responses on these warm-up trials were corrected or praised heavily as necessary. Infants were asked for a different object in a different location across the three warm-up trials.

Referent Mapping Trials. The referent selection trials immediately followed the warm-up trials, proceeding in the same manner except that no corrections or praise was given. Each infant was presented with three sets of objects, each of which included two known objects and one novel object. The same three known objects used during warm-up were used. On the first and third trials, infants were asked to get a known object. On the second trial, infants were asked to get a novel object (i.e. "Can you get the roke?"). Location of the target item was counterbalanced across trials and randomized across infants.

Delay Period. A five-minute delay followed the referent selection trials. During the delay, the infant was allowed to play in the waiting room. None of the items used during the experiment were present during the delay.

Retention Trial. The delay period was immediately followed by a single warm-up trial that proceeded in the same manner as the previous warm-up trials and used the same three known objects. Praise was given and infants were corrected as needed. The warm-up trial was immediately followed by the retention trial in which the infant was presented with the three novel objects present during the referent mapping trials, one of which had been named in the second trial and two of which were distracters present when the experimenter had asked for a known objects on trials one and three. The position of items was randomized across infants with the target item never being in the same location it had been during the referent selection trial.

Results

Infants chose the target significantly more than would be expected by chance on novel referent selection trials in both conditions, as seen in Table 1. In particular, 13 out of 19 infants familiarized with the novel object selected it when asked during referent selection as did 18 out of 20 familiarized with the word form; exact binomial, p <.01 and p<.001 respectively, see Table 1. These results are similar to those of Horst and Samuelson (2008; see also Mervis & Bertrand, 1994; Wilkinson et al, 2003). In contrast to Horst and Samuelson (2008), however, infants familiarized with the object prior to referent selection chose the target object at levels significantly greater than chance on the retention trials (10 out of 13, exact binomial, p<.01, note that only data from the infants who correctly mapped in the novel referent selection trials were included in this analysis). Infants familiarized with the word prior to referent selection, in contrast, performed at chance levels on retention trials (6 out of 18, exact binomial, p ns).

Table 1: Referent selection (RS) and retention (Ret) performance in Horst & Samuelson (2008) and current work

		KnownRS	NovelRS	Ret.
Horst	# Correct		12	3
&Sam	N		20	12
(2008)	% map		0.60	0.25
	p		<.01	*ns*
Exp 1	# Correct	27	13	10
Favorite	N		19	13
Object	% map	0.71	0.68	0.77
Familiariz.	*p*	<.001	<.01	<.01
Exp 1	# Correct	24	18	6
Word-Form	N		20	18
Familiariz.	% map	0.63	0.90	0.33
	p	<.001	<.001	*ns*
Exp 2	# Correct	29	18	8
No	N		20	18
Familiariz.	% map	0.73	0.90	0.44
	p	<.001	<.001	*ns*
Exp 3	# Correct	28	14	10
Non-Fav	N		20	14
Object	% map	0.70	0.70	0.71
Familiariz.	*p*	<.001	<.001	<.01

To directly examine the difference between conditions, we performed X^2 tests of homogeneity of proportions. These revealed no differences in referent selection performance across conditions, X^2 (1, N=39), *ns*, however, performance in retention did differ significantly between conditions, X^2 (1, N=31) <.*05*. Thus, it appears that familiarization with the novel object, but not the novel word, prior to the formation of a novel word-object mapping boosts retention of that mapping.

However, before accepting this conclusion we examine the possible role differences between our task and that of prior studies, as well as differences between our conditions, may have had on our findings.

Experiment 2: No Familiarization

One difference between our current procedure and that of Horst and Samuelson (2008) was our use of the same three known objects during warm-up and test. Pilot testing demonstrated that with both a familiarization period and different known objects on every trial, infants could not succeed in referent selection and thus, retention could not be analyzed. Thus, in Experiment 1 we had used the same known objects on each trial in an effort to direct children's attention to the novel object even more, thereby boosting infants' initial mapping during referent selection. However, it is possible that our repeated use of the same known objects on every trial also served to boost retention. We examine this possibility by testing retention in our procedure without the familiarization period, thus demonstrating that using the same three known items serves to boost referent selection but not retention.

Method

Participants Twenty 24-month-old infants (9 girls, *M* = 24 months, 19 days; range = 23 months, 20 days – 25 months, 4 days) with a mean vocabulary of 342 words (range = 134-536) participated. All infants were recruited through county birth records and were native English speakers. Participant's parents provided informed consent prior to the start of the study. Participants received a small toy for participation. Data for one additional infant was not included due to a recording error.

Stimuli The same novel objects and novel names from Experiment 1 were used (see Figure 1).

Procedure The procedure was identical to that of Experiment 1, with the exception that there was no pre-familiarization period.

Results

Infants chose the target significantly more than would be expected by chance on novel referent selection trials (18 out of 20, exact binomial, p<.001, see Table 1). In contrast to infants in Experiment 1 who were familiarized with the object prior to referent selection, infants in this experiment did not retain the novel object-word mapping over the delay; they selected the target object at chance levels during the retention test (8 out of 18, *ns*, note that again, only data from infants who correctly mapped in the novel referent selection trials were included in this analysis). Chi-square tests of homogeneity of proportions revealed that while there was a difference in referent selection performance between infants in Horst and Samuelson (2008) and infants here, X^2 (1, N=40), <.05, there was no significant difference in retention between the two groups, X^2 (1, N=28), *ns.* With respect to Experiment 1, then, these results indicate that easing the task by using the same known stimuli throughout did boost children's mapping ability during initial referent selection, but it was likely not responsible for the boost in retention seen when infants were familiarized with the novel objects.

Experiment 3: Non-Favorite Novel Target

One possible explanation for the difference in retention performance seen for children familiarized with the objects versus the words in Experiment 1 has to do with our use of the child's favorite object as the novel target. Recall that during familiarization we asked children for their three favorite items from the set of six novel objects, using these as the novel items present during referent selection. When then asked to find the target item during the retention trial when all three were present, children would be scored as correct if they chose their overall favorite item, even if they did not recall its link to the novel name. To test this possibility, we re-ran the object familiarization condition of Experiment 1, but instead used the non-favorite items as the novel items during referent selection.

Method

Participants Twenty 24-month-old infants (11 girls; *M* = 24 months, 22 days; range = 23 months, 21 days-25 months, 3 days) with a mean vocabulary of 272 words (18-567) participated. All infants were recruited through county birth records and were native English speakers. Participant's parents provided informed consent prior to the start of the study. Participants received a small toy for participation.

Stimuli The same novel objects and novel names from Experiment 1 were used (see Figure 1).

Procedure The procedure was identical to the visual condition in Experiment 1, with the exception that when asked to pick their favorite novel item during familiarization, that item was then removed from the table. This was repeated until three non-favorite items were remaining. These three remaining items were then used as the novel referents during the experiment.

Results

Infants chose the target significantly more than would be expected by chance on novel referent selection trials (14 out of 20, exact binomial, *p*<.001, see Table 1). Like infants in Experiment 1 who were familiarized with the object prior to referent selection, infants in this experiment also retained the novel word-object mapping over the delay, selecting the target object the majority of the time (10 out of 14, exact

binomial, $p<.01$, note that again, only data from infants who correctly mapped in the novel referent selection trials were included in this analysis). Chi-square tests of homogeneity of proportions revealed that there was no significant difference in referent selection, X^2 (1, N=39), *ns*, or retention X^2 (1, N=27), *ns*, between infants familiarized with the object in Experiment 1 and infants here. These results then indicate that the use of the infant's favorite novel object as the target during referent selection did not alter the infant's ability to retain the word-object link after a delay.

General Discussion

Despite the complexity of word learning, young children are remarkable at quickly mapping a novel word to a novel object. However, recent work has suggested that this mapping may not be as robust as previously thought, and thus, not the basis of children's rapid acquisition of new words. The goal of the present set of experiments was to probe how prior familiarity with the parts of a novel name-object mapping may help children retain novel name-object links. The results indicate that children given prior familiarity with the novel object to be mapped retained the mapping between the object and a novel word following a delay. In contrast, children given prior familiarity with the word-form mapped the novel word to a novel object during referent selection, but did not retain this mapping over a delay. Even when repetition of known objects and novel item preference were controlled for, children still demonstrated retention of a word-object mapping when familiarized with the object prior to test. Thus, our results indicate an important difference in the word-learning boost given by familiarity with the objects verses familiarity with the words in a fast-mapping task.

Importantly, the results of Experiments 1 and 3 support previous suggestions that a slow-mapping process (Carey, 1978; Carey & Bartlett, 1978) is needed for a robust mapping between a word and object. Notably, however, the results also demonstrate that prior familiarity with the object, but not the word, to be mapped helps this process. A similar idea has been presented in a recent model of word learning proposed by Mayor and Plunkett (2010). In this model, fast-mapping is facilitated by a well-developed representation of the object category prior to the actual name-object mapping. Likewise, our findings are also consistent with previous work by Smith and Yu (2008) suggesting that multiple exposures to a novel name and object are necessary for learning (see also McMurray, Horst & Samuelson, in prep; and Horst, McMurray, & Samuelson, 2006). It is clear from the literature that with more experience or information, children's ability to make specific word-object mappings is heightened (see also, Horst, 2007; and McMurray, Horst & Samuelson, in prep; Horst, Samuelson, & McMurray, 2010). One implication of the current study is the suggestion that across multiple exposures, visual and auditory components may not have been encoded equivalently. The literature presents several interesting suggestions as to why this might be the case.

One possible interpretation of the differential effects of word and object familiarization in our results comes from Sloutsky and colleagues' proposal of auditory dominance (Robinson & Sloutsky, 2004; Robinson & Sloutsky, 2007; Robinson & Sloutsky, 2008; Sloutsky & Napolitano, 2003; Sloutsky & Robinson, 2008). This is the suggestion that when both auditory and visual information are given to infants simultaneously, the auditory information receives preferentially processing. Support for this idea comes from studies in which infants were trained that a particular combination of auditory and visual stimuli indicated the location of a prize. When presented with either the trained auditory or visual cue paired with a competing auditory or visual cue, infants relied more on the auditory modality to anticipate the location of the prize (Robinson & Sloutsky, 2004; Robinson & Sloutsky, 2007). This theory would suggest that in Horst & Samuelson's (2008) referent selection task when infants were only given a single exposure to the novel object and name, they preferentially processed the auditory information and thus, only encoded half of the mapping – the novel name, not the physical referent. In the current study, the theory of auditory dominance would suggest that the familiarization period had a differential effect on infant's processing of novel names and objects at the point of referent selection. If infants come to the task with an auditory processing bias and are given additional familiarity with the visual component prior to test, when the word and object were presented during referent selection, both components could be processed at equivalent levels, allowing both the word and object to be encoded robustly. On the other hand, when infants were pre-familiarized with the word-form, the auditory processing was boosted even further, thus overshadowing the encoding of the visual object during referent selection.

It is also possible, however, that the apparent difference in visual and auditory familiarization stems more from task demands than differential processing of each component. That is, perhaps the use of a comprehension task to test retention creates the appearance of processing differences. In the traditional fast-mapping task, the experimenter provides the word during testing. When the infants are pre-familiarized with the objects and the experimenter provides the word during the retention task, children would then have both components necessary to demonstrate robust retention of the word-object link. On the other hand, when infants are pre-familiarized with the words and again given the word at test, the infants only have a rich encoding of the auditory component and do not demonstrate retention of the link. By this view then, the object familiarization condition did provide a boost to word learning, not because infants are biased to process the word form, but rather, because the task privileged the modality in which the children would subsequently use to find the referent.

It may be possible to discriminate between these explanations by examining the strength of the representations of the word and object following the initial referent selection trials without a pre-familiarization period.

Sloutsky's auditory dominance proposal would suggest that in a recognition test following referent selection, infants should show recognition of the words, but not the objects. In contrast, if the differences in results in our experiment are due to task effects, infants should show no encoding of the auditory information following referent selection. These results would also give insight to the extent to which familiarization might boost the representation of the category, as Mayor and Plunkett (2010) predict in their model. We are currently examining this possibility.

While further research is required to elucidate the exact depth to which object and word forms are processed by infants in a fast-mapping task, the current study makes it clear that the novel words and objects presented for mapping play different roles in the establishment of that mapping and in its retention. Thus, our finding that infants retain novel word-object mappings when familiarized with the objects but not the words reinforces Horst & Samuelson's (2008) and Carey's (1978) point that fast mapping is not equivalent to word learning. Our results also point to the importance of further work into the incremental process by which representations of words, objects and their mappings are created on the way to word learning.

Acknowledgments

This work was supported by NICHD grant (R01 RHD045713) to LS. We thank the parents and children who participated in these studies and the members of the Language and Category Development Lab.

References

Capone, N.C., & McGregor, K.K. (2005). The effect of semantic representation on toddlers' word retrieval. *Journal of Speech, Language, and Hearing Research, 48*, 1468-1480.

Carey, S. (1978). The child as word learner. In M. Halle, J. Bresnan, & A. Miller (eds.), Linguistic theory and psychological reality. (pp. 264-293). Cambridge, MA: MIT Press.

Carey, S., & Bartlett, E. (1978). Acquiring a single new word. Proceedings of the Standford Child Language Conference, 15, 17-29.

Fenson, L., Dale, P.S., Reznick, J.S., Bates, E., Thal, D., & Pathick, S. (1994). Variability in early communicative development. *Monographs of the Society for Research in Child Development, 59*(5), 173.

Gershkoff-Stowe, L. (2002). Object naming, vocabulary growth, and the development of word retrieval abilities. *Journal of Memory and Language, 46*, 665-687.

Gershkoff-Stowe, L., & Hahn, E.R. (2007). Fast mapping skills in the developmeitng lexicon. *Journal of Speech, Language, and Hearing Research, 50*, 682-697.

Golinkoff, R.M., Hirsch-Pasek, K., Bailey, L.M., & Wenger, N.R. (1992). Young children and adults use lexical principles to learn new nouns. *Developmental Psychology, 28*(1), 99-108.

Graf Estes, K., Evans, J.L., Alibali, M.W., & Saffran, J.R. (2007). Can infants map meaning to newly segmented words: Statistical segmentation and word learning. *Psychological Science, 18*(3), 254-260.

Halberda, J. (2003). The development of a word-learning strategy. *Cognition, 87*, B23-B34.

Horst, J.S., McMurray, B., & Samuelson, L.K. (2006). *Online processing is essential for learning: Understanding fast mapping and word learning in a dynamic connectionist architecture*. In R. Sun (Ed.), Proceedings of the Twenty-Eighth Annual Conference of the Cognitive Science Society (pp.339-344) LEA.

Horst, J.S., & Samuelson, L.K. (2008). Fast-mapping but poor retention by 24-month-old infants. *Infancy, 13*(2), 128-157.

Horst, J.S., Samuelson, L.K., & McMurray, B. (2010). When word learning is not about words: Fast mapping and visual familiarity. Manuscript submitted for publication, under review.

Mayor, J., & Plunkett, K. (2010). A neural computation account of taxonomic responding and fast-mapping in early word learning. *Psychological Review, 117*(1), 1-31.

Markson, L., & Bloom, P. (1997). Evidence against a dedicated system for word learning in children. *Nature, 385*, 813-815.

McMurray, B., Horst, J.S., & Samuelson, L.K. (in prep). Using your lexicon at two timescales: Investigating the interplay of word learning and recognition.

Mervis, C.B., & Bertrand, J. (1994). Acquisition of the novel name-nameless category (N3C) principle. *Child Development, 65*(6), 1646-1662.

Robinson, C.W., & Sloutsky, V.M. (2004). Auditory dominance and its chance in the course of development. *Child Development, 75*(5), 1387-1401.

Robinson, C.W., & Sloutsky, V.M. (2007). Visual processing speed: Effect of auditory input on visual processing. *Developmental Science, 10*(6), 734-740.

Robinson, C.W., & Sloutsky, V.M. (2008). Effects of auditory input in individual tasks. *Developmental Science, 11*(6), 869-881.

Sloutsky, V.M., & Napolitano, A.C. (2003). Is a picture worth a thousand words? Preferences for auditory modality in young children. *Child Development, 74*(3), 822-833.

Sloutsky, V.M., & Robinson, C.W. (2008). The role of words and sounds in infant's visual processing: from overshadowing to attentional tuning. *Cognitive Science, 32*, 354-377.

Smith, L.B., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition, 106*, 1558-1568.

Wilkinson, K.M., Ross, E., & Diamond, A. (2003). Fast mapping of multiple words: Insights into when the "information provided" does and does not equal "the information perceived." *Applied Developmental Psychology, 24*, 739-762.

Attentional Control and Early Word Learning

Hanako Yoshida (yoshida@uh.edu)
Duc N. Tran (crystalxlite@yahoo.com)
Department of Psychology, University of Houston
Houston, TX 77204-5022 USA

Viridiana Benitez (vlbenite@indiana.edu)
Megumi Kuwabara (mekuwaba@indiana.edu)
Psychological and Brain Sciences, Indiana University
Bloomington, IN 47405 USA

Abstract

There has been increasing interest in the role of early attention in the context of word learning. There has also been growing interest in attentional differences between bilinguals and monolinguals. The present study examined the relationship between mutual exclusivity and attentional control by comparing bilingual children whose attentional control is relatively advanced to age-matched monolingual children. The novel adjective learning paradigm was the word-learning measure and the Attention Network Test was the measure of attentional control. Three-year-old monolingual and bilingual children with similar vocabulary development participated. The results replicate earlier work on advanced attentional control among bilingual children and suggest that better performances in novel adjective learning by bilingual children might be due to attentional control. These findings support the importance of attention in early word mapping. The results add to a growing body of literature on the potential relevance of bilingualism in early word learning.

Keywords: Attentional learning; early word learning; adjective learning paradigm; mutual exclusivity; selective attention

Attentional Shifting and Word Learning

A growing body of experimental literature (mostly concerning adults) indicates that effective attentional control optimizes learning, especially in complex scenes (e.g., Cowan, Fristoe, Elliot, Brunner, & Saults, 2006). In the developmental literature, there has been interest in the role of effective attention shifting in learning, particularly in the domain of word learning. Because different kinds of words refer to different kinds of properties (e.g., nouns to shapes or whole objects, adjectives to properties such as color or texture), being able to shift attention seems an important aspect of word learning (Au, 1990). Indeed, a number of studies have documented that by the time children are 2 years old, they shift attention to different kinds of properties for different kinds of entities and in the context of different kinds of words (Graham, Williams, & Huber, 1999; Landau, Smith, & Jones, 1988, 1998; Soja, 1992; Soja Carey, & Spelke, 1991; Yoshida & Smith, 2003, 2005).

Some accounts of this effective attentional shifting refer to mapping principles for learning words (e.g., Bloom, 2000; Carey, 1978; Markman, 1989). One such principle is mutual exclusivity (Markman & Wachtel, 1988): In the context of a known word and referent, a novel word shifts attention to a novel referent. Mutual exclusivity is pervasive in early word learning, helping the learning of new nouns, but it has also been suggested as a reason why young word learners have difficulty learning adjectives (e.g., Carey, 1978; Markman, 1989; Regier, 1996). For example, young learners who know that a horse is called "horse" might reject the label "brown" being applied to it. This one label–one object constraint has long been considered a positive constraint on early word learning that promotes the learning of nouns. Also, it may help more advanced word learners learn adjectives. Older children, when challenged with two labels for a single object, will effectively shift attention to a nonshape property (if "horse" means HORSE, then "brown" must mean something about the horse; see Waxman, 2001; Waxman & Klibanof, 2000). Although the earlier view saw these constraints as lexically specific and possibly innate (Markman, 1989), more recent work suggests that the effect is related to learning through competitive attentional processes (Halberda, 2009; Hollich, Hirsh-Pasek, Tucker, & Golinkoff, 2000; Plunkett, 1998; Smith, 2000; Yoshida & Hanania, 2007).

If mutual exclusivity emerges because of competitive processes among words and referents that arise in on-line comprehension, then these processes—and their relation to effective attentional control in word learning—should be related to the learner's history of experiences in resolving competitions among words and referents. This, in turn, suggests that the development of mutual exclusivity may benefit from bilingualism and attentional control more generally.

Executive Control

Bilingual children have been characterized as developing cognitive flexibility earlier than monolingual children and demonstrating more robust self-control, including attentional control, throughout their lives (e.g., Bialystok, 1992, 1999; Bialystok & Martin, 2004; Bialystok & Senman 2004; Carlson & Meltzoff, 2008). Such positive effects are seen most profoundly in what are known as executive-function (or self-control) tasks. These are tasks that require the individual to inhibit preferred or prepotent patterns of responding (e.g., not jump up when one should be sitting, not take the candy when told not to, do a task in a new way not an old way; see Beaver & Wright, 2007; Kochanska, Murray, & Harlan, 2000; Luria, 1966; Luria, Pribram, & Homskaya, 1964; Mischel, Shoda, & Rodriguez, 1989; Zelazo & Frye, 1998). A number of recent studies have shown that these effects are also evident in executive function relevant to controlling attention and in the suppression and separation of languages to avoid interference. Indeed, the current consensus is that the bilingual advantage in executive control derives from the history of switching between languages (i.e., Costa, Hernandez, & Sebastian-Galles, 2008; Martin & Bialystok, 2003; Mezzacappa, 2004).

The idea that bilinguals are able to control the choice of their speech via well-developed processes of executive control is supported by their better performance in attentional tasks such as the Attention Network Test (ANT), which was developed by Fan, McCandliss, Sommer, Raz, and Posner (2002). This test was designed to measure the functionality of the attentional network: alerting, orienting, and executive control. Children are asked to find a fish facing a certain direction among other fishes on a computer screen. The direction the fish faces does not change throughout the task, but the facing direction of other fish does change and thus the task requires effective attention control. The response time for searching is often used to measure the attentional control. Costa et al. (2008) reported that bilinguals performed this task faster and more efficiently than monolinguals. Furthermore, when the task was broken down into the attentional network components, bilinguals performed significantly better in the alerting and executive control components. The bilingual advantage has also been reported in studies of bilingual children who have significantly lower English proficiency than the group of comparison English monolinguals (using ANT; Yang, 2004). This is an intriguing finding with a potentially widespread impact: Children who speak more than one language seem to show developmentally advanced attentional control.

What is not known is the extent of the advantage in attentional control or the role it plays in language learning. If this advantage emerges in young bilingual children as a consequence of learning two languages, then it seems its core function might be to support language learning itself. The experiment reported here seeks to link differences in attentional control between monolingual and bilingual children to attention shifting in word learning, and more specifically to mutual exclusivity in the context of learning a novel adjective.

Experiment

Method

Participants

Participants were 20 monolingual English learners with a mean age of 36.66 months (range: 29.47 to 43.16) and 20 bilingual learners (e.g., English–Spanish, English–Bengali, English–Chinese, English–Russian, English–Urdu, English–Vietnamese) with a mean age of 38.86 months (range: 30 to 45.53). The criteria of bilingual status was determined by a demographic questionnaire. A bilingual questionnaire was used to ensure that the language spoken at home was primarily not English.

Stimulus Materials

Vocabulary Assessment (MCDI) Eight sections from the MacArthur–Bates Communicative Development Inventories (MCDI) were selected and used to measure productive vocabulary. For English monolingual children, the English version was used, and for bilingual learners, their dominant language (if reported by their parents) was measured. We also used the Spanish MCDI. Adaptations of the MCDI in Chinese and Vietnamese were used when possible. Monolingual children's total vocabulary was measured as the number of words parents reported in their productive vocabulary in English; bilingual children's total vocabulary was measured as the number of words parents reported in their dominant language (i.e., the language used most often by parent report).

MacArthur Socioeconomic Status Parents were asked to fill out a demographic questionnaire to control for the influence of socioeconomic status (SES) in bilingual and monolingual participants. All participants were matched and came from the same SES background.

Novel Adjective Learning Task Each of the eight trials in this task used three objects (one exemplar, two test objects); the objects in each trial were unique. All were instances of familiar animate objects (e.g., ducks) and inanimate objects (e.g., trucks) with distinctive colors. As shown in Figure 1, each exemplar was presented with a property that was highly novel (*sticky*). The two test objects for each trial had the same shape as the exemplar, but different colors. One test object presented

a target property match of the novel texture (e.g., red *sticky* duck), and one presented a non-property-matching texture (e.g., red *bumpy* duck). Within all trials, all objects—exemplars and test objects (property matching and non-property-matching)—had the same shape.

Familiar Adjective Learning Task The same three-dimensional object form was used for exemplars and test objects (e.g., ducks, trucks). The properties, familiar and likely to be receptively known by the children (e.g., bumpy, spotted, shiny, holey), can be seen in Figure 2. Two types of test objects were presented: one with a property match of a familiar/known texture (e.g., red *bumpy* duck), and one with a nonmatching property where texture did not match the exemplar (e.g., red *shiny* duck).

All objects were approximately 10 cm^3. Textures—the intended target property—were chosen to be highly novel and included a stringy pattern, a wire pipe-cleaner surface, a sponge-like surface, and a Velcro surface. These properties were named by novel labels such as *blickish, dakish, talish,* and *wuggish*, respectively. For familiar textures, stimuli were *holey*, *shiny, bumpy,* and *spotted*.

Attention Network Test (ANT) We used the original "child version" of Dr. Jin Fan's ANT (http://www.sacklerinstitute.org/users/jin.fan/). The children were asked to watch a computer screen where five fish lined up horizontally. The task was to point to the mouth of the "hungry fish," which was defined as the fish always in the middle. The direction the hungry fish faced changed throughout the task, but the facing direction of other fishes changed. Children were required to shift their attention effectively to detect the direction of the hungry fish's mouth. We used a touch-screen laptop to measure accuracy in this task.

Procedure

All children participated in the Novel Adjective Learning Task, Familiar Adjective Learning Task (control), and the ANT (in randomized order) in their dominant language; the task order was counterbalanced. Caregivers were asked to fill out the SES and MCDI forms. Parents of bilingual children were asked to fill out two MCDI forms, one in English and one in their second language. Parents were asked to go through the list and specify all the words they had heard their children use. The Novel Adjective Learning Task, Familiar Adjective Learning Task, and ANT trials were administered in a quiet, controlled room (both at the laboratory and at daycare centers) by trained research assistants fluent in the child's dominant language.

Novel Adjective Learning Task Participants were presented with an exemplar and told the name along with a novel adjective (e.g., "See this? This is a *blickish* duck!"). After the exemplar was removed from view, the participants were then presented with two test objects. They were asked to give the experimenter the one to which the novel adjective could apply (e.g., "Now, can you give me a duck that is *blickish*?). The order of the trials was randomized and the children's selection of the test object—whether a property-matching object or a non-property-matching object—was recorded for all trials for later analysis.

Familiar Adjective Learning Task The same procedure was administered, only now the adjectives presented were familiar/known and not novel (e.g., "See this? This is a *bumpy* duck!" "Now, can you give me a duck that is *bumpy*?").

Attention Network Test (ANT) The ANT trials were administered using E-Prime software on a 15" touch-screen laptop computer. The children sat at a comfortable distance from the screen and used their index finger to touch the fish displayed on the screen. The children were instructed to help feed the hungry (target) fish as fast as they could by touching the mouth of the fish on the screen, according to which direction the hungry fish was oriented. They were told that sometimes the fish would appear alone, and other times it would swim together with other fish. In all cases, they were instructed to concentrate on the one fish in the middle—the hungry fish. They were also asked to keep their eyes on the fixation point during the task. The completion time was approximately 10 min. Their accuracy (percent correct) and reaction times (RTs) were recorded for later analysis.

Results

All bilingual and monolingual participants came from the same SES background (i.e., middle class) and were matched on vocabulary production through parental reports. Table 1 shows vocabulary size, dominant language, and age.

Novel Adjective Learning Task

As can be seen in Figure 3, bilingual children performed better than monolingual children, $t(19)=3.92$, $p<.05$, in the Novel Adjective Learning Task, selecting property-matching objects with high accuracy, whereas the monolingual children performed at chance.

Familiar Adjective Learning Task

In terms of overall accuracy on the Familiar Adjective Learning Task, bilingual and monolingual children performed similarly and above chance, $t(19)=2.75$, $p<.05$, and $t(19)=3.18$, $p<.05$, in selecting property-matching objects (see Figure 3). These results indicate

that the participants were able to comprehend the task at hand and that bilinguals had no special advantage in this task. Thus the advantage observed in the Novel Adjective Learning Task must be due to mapping novel adjectives to the novel properties of known things.

Attention Network Test

Bilingual children performed better than monolingual children on the ANT, $t(19)=3.74$, $p<.05$, (Figure 4). More critically for the present hypothesis, children's success in the Novel Adjective Learning Task was significantly correlated with scores from attentional control ($r=.480$, $p<.05$).

General Discussion

These results replicate the bilingual advantage in attentional control tasks that has been reported by others and tie this effect to attentional strategies in word learning. The findings promise new insights about the cognitive consequences of learning and speaking two languages and the role of attention in using and learning language. Attention is a process that changes itself through its own activity, a fact that has far-reaching importance for understanding the learning of words and referents by both monolingual and bilingual children.

The Consequences of Bilingualism

The default assumption in the study of bilingual children has been that their cognitive systems are no different than those of monolingual children, and thus speaking two languages has often been viewed as a source of developmental delays (e.g., Doyle, Champagne, & Segaloqitz, 1978). However, we now know there are significant positive consequences that extend beyond language itself and appear to involve executive control processes across many domains—from not taking candy, to sitting still when one should, to—in the present study—shifting attention to novel words and properties of well-known objects. In this way the present study connects the bilingual advantage in executive control to language learning—the context in which that advantage emerged in the first place.

Attention in Word Learning

By tying the bilingual advantage in executive control to attention shifting in the learning of novel adjectives, the results also suggest that the competitive and attentional processes that are studied in early word learning (in monolinguals as well as bilinguals) may be fundamentally linked to general processes of executive and attentional control. There is a large body of literature in this domain (Diamond, 1990) showing—in monolingual children—incremental advances from late infancy to the school-age years in the ability to switch attention and inhibit prepotent but irrelevant information. The present results highlight the importance of studying the codevelopment of these processes with word-referent learning in both monolingual and bilingual children. In brief, we may be able to mechanistically ground word-learning strategies in more general attentional processes.

There are certainly intriguing indications that there is still much to be learned from taking this approach. For example, whereas the present task asked children to learn a property label for a known category—and bilingual children showed an advantage—other studies have asked whether bilingual and monolingual children differ in their ability to learn two different names for the same thing. Depending on how one conceptualizes the task, bilingual children either show an advantage at learning two names or exhibit weaker mutual exclusivity constraint in this context (Au & Glusman, 1990; Davidson, Jergovic, Imami, & Theodos, 1997; Davidson & Tell, 2005; Merriman & Kutlesic, 1993). Much of the previous work on this "two names for one thing" task used labels from different languages with different phonological properties. This provides a potentially useful way to understand the microprocesses and context cues that elicit and resolve competitions within and across languages.

In sum, the present work supports the importance of attention in word learning and its link to general processes of attentional switching and executive control. Systematic comparisons of monolingual and bilingual children in both word learning and attentional control tasks offer a new window on these fundamental processes, their development, and their relation to word learning.

Table 1: Productive vocabulary of dominant language based on the MCD.

group	age	noun	verb	adjective	Total
monolingual	36.7	177.9	81.6	48.5	308.0
bilingual	38.9	178.6	83.7	45.9	308.2

Figures

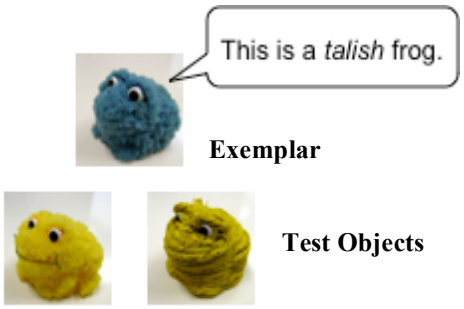


Figure 1: A set of stimulus objects used in the Novel Adjective Learning Task.

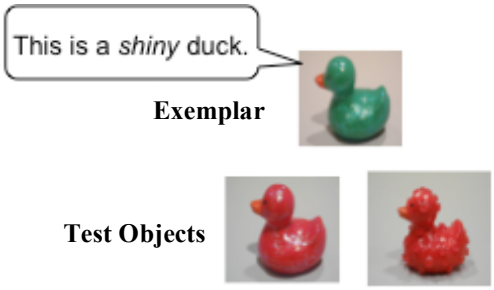


Figure 2: A set of stimulus objects used in the Familiar Adjective Learning Task.

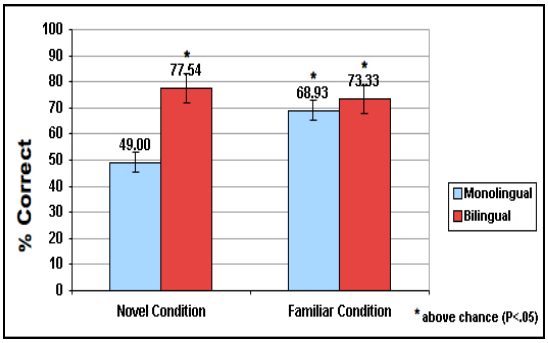


Figure 3: Monolingual and bilingual children's percent correct on mapping novel labels to novel properties (left) and familiar labels to familiar properties (right) in adjective learning tasks.

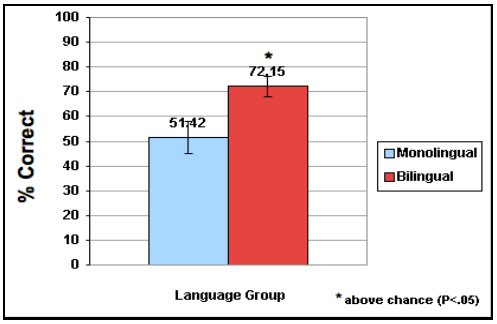


Figure 4: Monolingual and bilingual children's percent correct on the Attention Network Test.

Acknowledgments

This research is supported in part by a National Institutes of Health grant (R01 HD058620), the Foundation for Child Development, and University of Houston's Grants to Enhance and Advance Research (GEAR) program. We thank the children and parents who participated in this study.

References

Au, T. K. (1990). Children's use of information in word learning. *Journal of Child Language, 17*, 393–416.

Au, T. K., & Glusman, M. (1990). The principle of mutual exclusivity in word learning: To honor or not to honor? *Child development, 61*, 1474–1490.

Beaver, K. M., & Wright, J. (2007). The stability of low self-control from kindergarten through first grade, *63*, 86.

Bialystok, E. (1992). Selective attention in cognitive processing: The bilingual edge. In R. J. Harris (Ed.), *Cognitive processing in bilinguals*. Amsterdam: Elsevier Science.

Bialystok, E. (1999). Cognitive complexity and attentional control in the bilingual mind. *Child Development, 70*, 636–644.

Bialystok, E., & Martin, M. M. (2004). Attention and inhibition in bilingual children: Evidence from the dimensional change card sort task. *Developmental Science, 7,* 325–339.

Bialystok, E., & Senman, L. (2004). Executive processes in appearance-reality tasks: The role of inhibition of attention and symbolic representation. *Child Development, 75,* 562–579.

Bloom, P. (2000). *How children learn the meanings of words*. Cambridge, MA: MIT Press.

Carey, S. (1978). The child as word learner. In M. Halle, J. Bresnan, & G. A. Miller (Eds.), *Linguistic theory and psychological reality*. Cambridge, MA: MIT Press.

Carlson, S. M., & Meltzoff, A. N. (2008). Bilingual experience and executive functioning in young children. *Developmental Science, 11*, 282–298.

Costa, A., Hernández, M., & Sebastián-Gallés, N. (2008). Bilingualism aids conflict resolution: Evidence from the ANT task. *Cognition, 106*, 59–86.

Cowan, N., Fristoe, N. M., Elliot, E. M., Brunner, R. P., & Saults, S. (2006). Scope of attention, control of attention, and intelligence in children and adults. *Memory & Cognition, 34,* 1754–1768.

Davidson, D., Jergovic, D., Imami, Z., & Theodos, V. (1997). Monolingual and bilingual children's use of the mutual exclusivity constraint. *Journal of Child Language, 24*, 3–24.

Davidson, D., & Tell, D. (2005). Monolingual and bilingual children's use of mutual exclusivity in the naming of whole objects. *Journal of Experimental Child Psychology, 92,* 25–45.

Diamond, A. (1990). Developmental time course in human infants and infant monkeys, and the neural bases of inhibitory control in reaching. *Development and Neural Bases of Higher Cognitive Functions, 608,* 637–676.

Doyle, A., Champagne, M., & Segalowitz, N. (1978).

Some issues in the assessment of linguistic consequences of early bilingualism. In M. Paradis (Ed.), *Aspects of bilingualism*. Columbia, Hornbeam Press.

Fan, J., McCandliss, B. D., Sommer, T., Raz, A., & Posner, M. I. (2002). Testing the efficiency and independence of attentional networks. *Journal of Cognitive Neuroscience, 14,* 340–347.

Graham, S. A., Williams, L. D., & Huber, J. F. (1999). Preschoolers' and adults' reliance on object shape and object function for lexical extension. *Journal of Experimental Child Psychology, 74,* 128–151.

Halberda, J. (2009, April). *Mutual exclusivity as logical inference: Evidence for domain general disjunctive syllogism in 2-3 year olds.* Talk presented at the Society for Research in Child Development, Denver, CO.

Hollich, G., Hirsh-Pasek, K., Tucker, M. L., & Golinkoff, R. M. (2000). A change is afoot: Emergentist thinking in language acquisition. In P. Anderson, C. Emmeche, N. O. Finnemann, & P. V. Christiansen (Eds.), *Downward causation*. Oxford, England: Aarhus University

Kochanska, G., Murray, K. T., & Harlan, E. (2000). Effortful control in early childhood: Continuity and change, antecedents, and implications for social development. *Developmental Psychology, 36,* 220–232.

Landau, B., Smith, L. B., & Jones, S. (1988). The importance of shape in early lexical learning. *Cognitive Development, 3,* 299–321.

Landau, B., Smith, L. B., & Jones, S. (1998). Object perception and object naming in early development. *Trends in Cognitive Sciences, 2,* 19–24.

Luria, A. R. (1966). *Higher cortical functions in man*. New York: Basic Books.

Luria A. R., Pribram K. H., & Komskaya E. D. (1964). An experimental analysis of the behavioral disturbance produced by a left frontal arachnoidal endotheliomia (meningioma). *Neuropsychologia, 2,* 257–280.

Markman, E. M. (1989). *Categorization and naming in children: problems of induction.* Cambridge, MA: MIT Press.

Markman, E. M., & Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology, 20,* 121–157.

Martin, M. M., & Bialystok, E. (2003, October). *The development of two kinds of inhibition in monolingual and bilingual children: Simon vs. Stroop*. Poster presented at the meeting of the Cognitive Development Society, Park City, Utah.

Merriman, W. E., & Kutlesic, V. (1993). Bilingual and monolingual children's use of two lexical acquisition heuristics. *Applied Psycholinguistics, 14,* 229–249.

Mezzacappa, E. (2004). Alerting, orienting, and executive attention: Developmental properties and sociodemographic correlates in an epidemiological sample of young, urban children. *Child Development, 75,* 1373–1386.

Mischel, W., Shoda, Y., & Rodriguez, M. L. (1989). Delay of gratification in children. *Science, 244,* 933–938.

Plunkett, K. (1998). Language acquisition and connectionism. *Language and Cognitive Processes, 13,* 97–104.

Regier, T. (1996). *The human semantic potential: Spatial language and constrained connectionism*. Cambridge, MA: MIT Press.

Soja, N. (1992). Inferences about the meanings of nouns: The relationship between perception and syntax. *Cognitive Development, 7,* 29–46.

Soja, N., Carey, S., & Spelke, E. (1991). Ontological categories guide young children's inductions of word meanings: Object terms and substance terms. *Cognition, 38,* 179–211.

Smith, L. B. (2000). Learning how to learn words: An associative crane. In Golinkoff, K. Hirsh-Pasek, L. Bloom, L.B. Smith, A.L.. Woodward, N. Akhtar, et al. (Eds) *Becoming a word learner*. New York: Oxford University Press.

Waxman, S. R. (2001). Word extension: A key to early word learning and domain-specificity. Commentary on P. Bloom. *Behavioral and Brain Sciences, 24,* 1121–1122.

Waxman, S. R., & Klibanoff, R. S. (2000). The role of comparison in the extension of novel adjectives. *Developmental Psychology, 36,* 571–581.

Yang, S. (2004). *Testing bilingual children's cognitive advantages in executive attention*. Unpublished master's thesis, Cornell University.

Yoshida, H., & Hanania, R. (2007). Attentional highlighting as a mechanism behind early word learning. In D. S. McNamara & J. G. Trafton (Eds.), *Proceedings of the 29th Annual Cognitive Science Society*. Nashville, TN: Cognitive Science Society.

Yoshida, H., & Smith, L. B. (2003). Shifting ontological boundaries: How Japanese- and English-speaking children generalize names for animals and artifacts. *Developmental Science, 6,* 1–34.

Yoshida, H. & Smith, L. B. (2005). Linguistic cues enhance the learning of perceptual cues. *Psychological Science*, 16 (2), 90–95.

Zelazo, P. D., & Frye, D. (1998). II. Cognitive complexity and control: The development of executive function. *Current Directions in Psychological Science, 7,* 121−126.

A Bayesian Nonparametric Approach to Multisensory Perception

İlker Yıldırım (iyildirim@bcs.rochester.edu)
Robert A. Jacobs (robbie@bcs.rochester.edu)
Department of Brain & Cognitive Sciences, University of Rochester, Rochester, NY 14627, USA

Abstract

We propose a Bayesian nonparametric model of multisensory perception based upon the Indian buffet process. The model includes a set of latent variables that learn multisensory features from unisensory data. The model is highly flexible because it makes few statistical assumptions. In particular, the number of latent multisensory features is not fixed a priori. Instead, this number is estimated from the observed data. We applied the model to a real-world visual-auditory data set obtained when people spoke English digits. Our results are consistent with several hypotheses about multisensory perception from the cognitive neuroscience literature. We found that the model obtained the statistical advantages provided by sensory integration. We also found that the model acquired multisensory representations that were relatively sensory invariant. Lastly, we found that the model was able to associate unisensory representations based on different modalities.

Keywords: multisensory perception; Bayesian modeling; rational analysis; Indian buffet process

Introduction

We learn about our environments from many different senses. Objects can be seen, heard, touched, tasted, and smelled. How are our mental representations based on these different sensory modalities structured, combined, and coordinated?

Cognitive neuroscientists have recently studied three important hypotheses about multisensory perception. First, researchers have conjectured that multisensory representations are advantageous because sensory integration ameliorates the effects of bias and noise contained in representations based on single modalities. Multisensory representations are, therefore, able to convey more accurate and reliable information than the unisensory representations from which they are derived. Consider an observer that sees and touches a surface slanted in depth. Suppose that the observer's slant estimates based on the visual cue and on the haptic cue are each corrupted by sensory noise with some variance. It is easily shown that the maximum likelihood estimate of surface slant obtained by combining information from both cues has a lower variance, and is thus more reliable, than estimates based on either cue alone. Evidence that the brain is able to combine sensory information in such a manner was obtained by Ernst and Banks (2002), for example, who found that people's estimates of object height based on both visual and haptic information was more reliable than their estimates based on either visual or haptic information alone.

Second, researchers have hypothesized that our neural representations of objects are often sensory invariant, meaning they are the same (or at least similar) regardless of the sensory modalities through which we perceive those objects. Evidence consistent with this hypothesis was obtained by Amedi et al. (2001). They showed that a neural region known as the lateral occipital complex (LOC) shows similar patterns of activation regardless of whether an object is seen or touched.

Third, researchers have speculated that representations based on different modalities are associated with each other. Suppose that an observer sees, but does not hear, an object. A visual representation of that object will be active in the observer's brain, and this representation will often predict or activate an auditory representation of the object even though the object is not heard. Evidence consistent with this hypothesis was obtained by Calvert et al. (1997). They found that viewing facial movements associated with speech (lipreading) leads to activation of auditory cortex in the absence of auditory speech sounds.

Here, we propose a model of multisensory perception that learns about its multisensory environment in an unsupervised manner. In unsupervised learning, the data provided to a learner are unlabeled. The goal of the learner is to discover patterns and structure within the data set. There is a dichotomy in the cognitive science and machine learning literatures between parametric and nonparametric unsupervised learning methods. A parametric method uses a fixed representation that does not grow structurally as more data are observed. Examples include factor analysis, where the number of latent variables is fixed a priori, and cluster analysis, where the number of clusters is fixed a priori. In contrast, a nonparametric method uses representations that are allowed to grow structurally as more data are observed. These methods are often used when the goal is to impose as few assumptions as possible and to "let the data speak for themselves" (Blei, Griffiths, & Jordan, 2010). Examples include Dirichlet process mixture models (or Chinese restaurant processes) and Indian buffet processes.

The proposed model of multisensory perception is an instance of a Bayesian nonparametric model. It "explains" the unisensory representations arising from different modalities through the use of a set of latent or hidden variables that learn multisensory representations. The number of latent variables is not fixed. Instead, this number is treated as a random variable whose probability distribution is estimated based on the unisensory data. Because the size of the latent multisensory representations are estimated from the observed unisensory data, nonparametric statistical methods are required for inference. We use a Bayesian nonparametric framework developed by Griffiths and Ghahramani (2005, 2006) known as the Indian buffet process. Due to its Bayesian foundations, the proposed model can be regarded as an ideal observer model inferring optimal features of its multisensory environment (Austerweil & Griffiths, 2009).

We applied the proposed model to a visual-auditory data set obtained when people spoke different digits. Our results are consistent with the three hypotheses from the cognitive neuroscience literature described above. It was found that the model obtained the statistical advantages provided by sensory integration: categorization of objects was more accurate based on its latent multisensory representations than on the latent features of unisensory models. In addition, the model's latent or multisensory representations were relatively sensory invariant. That is, similar representations of an object were formed regardless of whether an object was seen or heard. Lastly, the model was able to associate representations based on different modalities. In other words, it could use one type of unisensory representation to predict or activate another type of unisensory representation.

Visual-Auditory Data Set

The multisensory perception model was applied to a visual-auditory data set known as the Tulips1 data set (Movellan, 1995). Twelve people (9 adult males, 3 adult females) were videotaped while uttering the first four digits of English twice.

In each video frame, the image of a speaker's mouth was processed to extract 6 visual features: the width and height of the outer corners of the mouth, the width and height of the inner corners of the mouth, and the heights of the upper and lower lips. The auditory signal corresponding to a frame was processed to extract 26 features: 12 cepstral coefficients[1], 1 log-power, 12 cepstral coefficient derivatives, and 1 log-power derivative. Because speech utterances had different durations, we sampled 6 frames for each utterance spanning the entire duration of the utterance in a uniform manner. In summary, each data item contained values for 36 visual features (6 frames $\times$ 6 visual features per frame) and 156 auditory features (6 frames $\times$ 26 auditory features per frame).

Training and test sets were created as follows. For the first eight speakers, one utterance of each digit was used for training and the other utterance was used for testing. For the remaining speakers, both utterances were used for training. Thus, the training set contained 16 data items for each digit, and the test set contained 8 data items for each digit.

Multisensory Perception Model

We describe the proposed model in the context of a visual-auditory environment, though we note that the model is equally applicable to other sensory modalities and to any number of modalities. A coarse schematic of the model is illustrated in Figure 1. It contains three sets of nodes or variables corresponding to visual features, auditory features, and multisensory features. The visual and auditory features are statistically dependent. However, they are conditionally independent given values for the multisensory features. The values of the visual features are observed when an object is viewed. When an object is not viewed, the visual features are latent, and their distributions can be inferred. Similarly, the values of the auditory features are observed when an object is heard. Otherwise, the auditory features are latent, and their distributions can be inferred. The multisensory features are always latent variables. Whereas the numbers of visual and auditory features are fixed, the number of multisensory features is not. Consistent with the nonparametric approach, this number is a random variable whose distribution is inferred from the data.

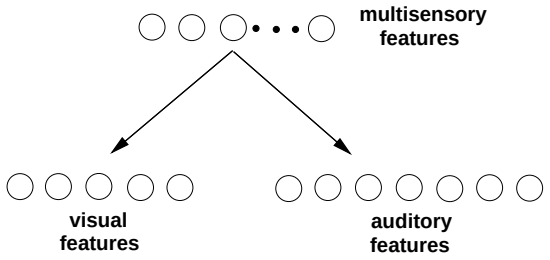


Figure 1: A coarse schematic of the multisensory perception model.

Formally, the model is a straightforward extension of the Indian buffet process (Griffiths & Ghahramani, 2005, 2006). A detailed graphical representation of the model is shown in Figure 2. An important goal of the model is to find a set of latent multisensory features, denoted Z, "explaining" a set of observed visual and auditory features, denoted X_V and X_A, respectively. Assume that a learner both sees and hears a number of objects. Let Z be a binary multisensory feature ownership matrix, where $Z_{ij} = 1$ indicates that object i possesses multisensory feature j. Let X_V and X_A be real-valued visual and auditory feature matrices, respectively (e.g., $X_{V_{ij}}$ is the value of visual feature j for object i). The problem of inferring Z from X_V and X_A can be solved via Bayes' rule:

$$p(Z|X_V, X_A) = \frac{p(X_V|Z)\ p(X_A|Z)\ p(Z)}{\sum_{Z'} p(X_V|Z')\ p(X_A|Z')\ p(Z')}$$

where $p(Z)$ is the prior probability of the multisensory feature ownership matrix, and $p(X_V|Z)$ and $p(X_A|Z)$ are the likelihoods of the observed visual and auditory feature matrices, respectively, given the multisensory features. We now describe the prior and likelihood distributions.

The multisensory feature ownership matrix is assigned a Bayesian nonparametric prior distribution known as the Indian buffet process (Griffiths & Ghahramani, 2005, 2006). It can be interpreted as a probability distribution over feature ownership matrices with an unbounded (infinite) number of features. The distribution is written as:

$$p(Z) = \frac{\alpha^K}{\prod_{h=1}^{2^N-1} k_h!} \exp\{-\alpha H_N\} \prod_{k=1}^{K} \frac{(N-m_k)!(m_k-1)!}{N!}$$

where N is the number of objects, K is the number of multisensory features, K_h is the number of features with history h (the history of a feature is the matrix column for that feature interpreted as a binary number), H_N is the Nth harmonic

[1] Cepstral coefficients are the coefficients of the Fourier transform representation of the log magnitude spectrum.

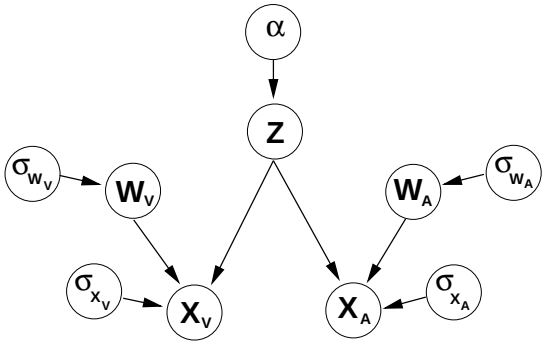


Figure 2: A Bayesian network representation of the multisensory perception model.

number, m_k is the number of objects with feature k, and α is a variable influencing the number of features.

The visual and auditory likelihoods are each based on a linear-Gaussian model. Let z_i be the multisensory feature values for object i, and let $x_{i\beta}$ be the feature values for object i where β is set to either V or A depending on whether we are referring to visual or auditory features. Then $x_{i\beta}$ is drawn from a Gaussian distribution whose mean is a linear function of the multisensory features, $z_i W_\beta$, and whose covariance matrix equals $\sigma^2_{X_\beta} I$, where W_β is a weight matrix (the weight matrices themselves are drawn from zero-mean Gaussian distributions with covariance $\sigma^2_{W_\beta} I$). Given these assumptions, the likelihood for a feature matrix is:

$$p(X_\beta|Z,W_\beta,\sigma^2_{X_\beta}) = \frac{1}{(2\pi\sigma^2_{X_\beta})^{ND_\beta/2}} \times \exp\{-\frac{1}{2\sigma^2_{X_\beta}}\mathrm{tr}((X_\beta - ZW_\beta)^T(X_\beta - ZW_\beta))\}$$

where D_β is the dimensionality of X_β, and $\mathrm{tr}(\cdot)$ denotes the trace operator.

Simulation Results

The multisensory perception model was applied to the visual-auditory data set. To better understand its performances, we also consider the performances of two other models. The vision-only model is identical to the multisensory model except that it contains only two sets of variables corresponding to visual and latent features. When applied to the visual-auditory data set, it received only the visual features. Similarly, the auditory-only model contains only two sets of variables corresponding to auditory and latent features. It received only the auditory features from the data set.

Because exact inference in the models is computationally intractable, approximate inference using Markov chain Monte Carlo (MCMC) sampling methods (e.g., Gelman et al., 1995) was performed based upon the training data following Griffiths and Ghahramani (2005). A single chain of each model was simulated. Each chain was run for 5000 iterations. The first 3000 iterations were discarded as burn-in. To reduce correlations among variables at nearby iterations, the remaining iterations were thinned to every 10th iteration (i.e., only variable values at every 10th iteration were retained). Thus, the results below are based on 200 iterations.

Posterior distributions over latent features

Recall that the number of latent features in each model is not fixed a priori. Instead, it is a random variable whose distribution is inferred from the training data. The three graphs in Figure 3 show the distributions of the numbers of latent features in the visual-only, auditory-only, and multisensory models. The visual-only model used relatively few latent features, the auditory-only model used more latent features, and the multisensory model used the most latent features. This result confirms that the models are highly flexible. Their nonparametric nature allows them to adapt their representational capacities based on the complexities of their data sets.

Categorization performances

We evaluated each model's ability to categorize the speech utterances as instances of one of the first four digits in English based upon its latent feature representations. At each iteration of an MCMC chain, a model sampled a latent feature representation for each data item in the training set. Using these representations, we performed k-means clustering with four cluster centers. We then performed an exhaustive search of assignments of clusters to English digits (e.g., cluster A $\rightarrow$ digit 3, cluster B $\rightarrow$ digit 1, etc.) to find the assignment producing the best categorization performance. Performances were averaged across iterations of a chain.

The results are shown in the leftmost graph of Figure 4. The horizontal axis gives the model, and the vertical axis plots the percent of data items in the training set that were correctly classified (error bars indicate the standard deviations of these percents across iterations of an MCMC chain). As expected, the vision-only model showed the worst performance, the auditory-only model showed better performance, and the multisensory model showed the best performance.

Its possible that the multisensory model showed the best performance solely due to the fact that it received both visual and auditory features and, thus, received a richer set of inputs than either the visual-only or auditory-only models. To evaluate this possibility, we simulated a model, referred to as a 'mixed' model, that resembled the multisensory model in the sense that it received both visual and auditory features. However, for the mixed model, these features were not segregated into separate input streams. Instead, the mixed model contained a set of latent features that received inputs from a set of undifferentiated perceptual features, namely a concatenation of the visual and auditory features. The results for the mixed model on the training set are also shown in the leftmost graph of Figure 4. The mixed model showed significantly poorer performance than the multisensory model, thus suggesting the statistical advantages of segregating perceptual inputs into separate streams.

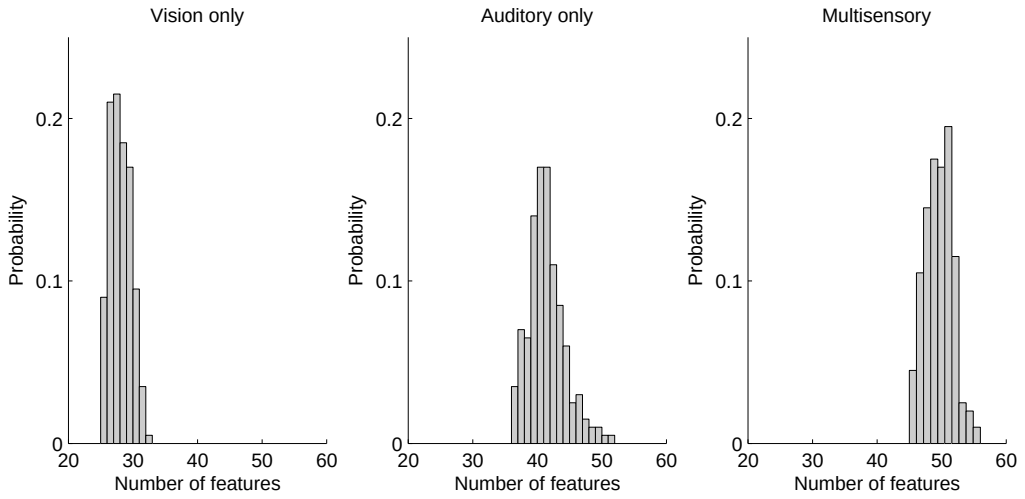


Figure 3: The distributions of the numbers of latent features in the visual-only (left), auditory-only (middle), and multisensory (right) perception models.

This analysis was repeated using the data items in the test set. Performing the analysis on test items presents unique challenges. Although it is reasonable to sample variables' values, and thus estimate variables' distributions, on the basis of training items, models are not meant to learn from test items. Consequently, we could not run our MCMC sampler on a model using the test items to evaluate the model's categorization performance. Doing so would erase the distinction between training and test data items.

Instead, we proceeded as follows. For a given model, consider the latent feature representations obtained on iteration i of the MCMC sampler when the model was trained on the training data. There is one such representation for each training item. These are the latent representations with non-zero probability based solely on iteration i. Let $\mathcal{L}_i$ denote this set of representations. For each data item in the test set, we searched $\mathcal{L}_i$ to find a latent representation that was most probable given the item. This was repeated for every item in the test set. Using these representations, the analysis of the test set is identical to the analysis of the training set described above: latent representations were clustered using k-means clustering, and an exhaustive search of assignments of clusters to digits was performed to find the assignment producing the best categorization performance. Performances were averaged across iterations.

The results are shown in the rightmost graph of Figure 4. Again, the multisensory model showed the best performance.

In summary, the multisensory perception model showed the best categorization performance on both training and test data sets. We conclude that its superior performance is due to both its rich set of inputs (it receives both visual and auditory features) and due to its internal structure (visual and auditory features are segregated perceptual streams). Clearly, this model received the statistical benefits of sensory integration.

Sensory invariance

As discussed above, neural representations of objects are often sensory invariant. That is, the same (or at least similar) neural representations arise regardless of the modality through which an object is sensed. Does the multisensory perception model show this same property?

We investigated this question as follows. As above, let $\mathcal{L}_i$ denote the set of multisensory feature representations obtained on iteration i of the MCMC sampler when the model was trained on the training data. Recall that these are the latent or multisensory representations with non-zero probability based solely on iteration i. For each data item in the training set, we calculated the probability distribution of the multisensory representation given an item's visual features, and the distribution of the multisensory representation given an item's auditory features where $\mathcal{L}_i$ was the set of possible multisensory representations. When all training items are taken into account, these distributions are denoted $p(Z|X_V)$ and $p(Z|X_A)$, respectively. We then calculated the Battacharyya distance between $p(Z|X_V)$ and $p(Z|X_A)$.[2] On every iteration, this distance was zero.

We repeated this analysis using the data items in the test set. Again, we computed $p(Z|X_V)$ and $p(Z|X_A)$ where X_V and X_A refer to the visual and auditory features of test items, and where $\mathcal{L}_i$ is the set of possible multisensory representations. The Battacharyya distances between $p(Z|X_V)$ and $p(Z|X_A)$ are always small values—the distribution of these distances has values of 1.51, 1.55, and 1.68 as its 25th, 50th, and 75th percentiles, respectively. By way of comparison, we also computed the distance between $p(Z|X_A)$ and a uniform distribution over multisensory representations. The distribu-

[2] We also considered the Kullback-Leibler distance but use of this metric led to numerical instabilities.

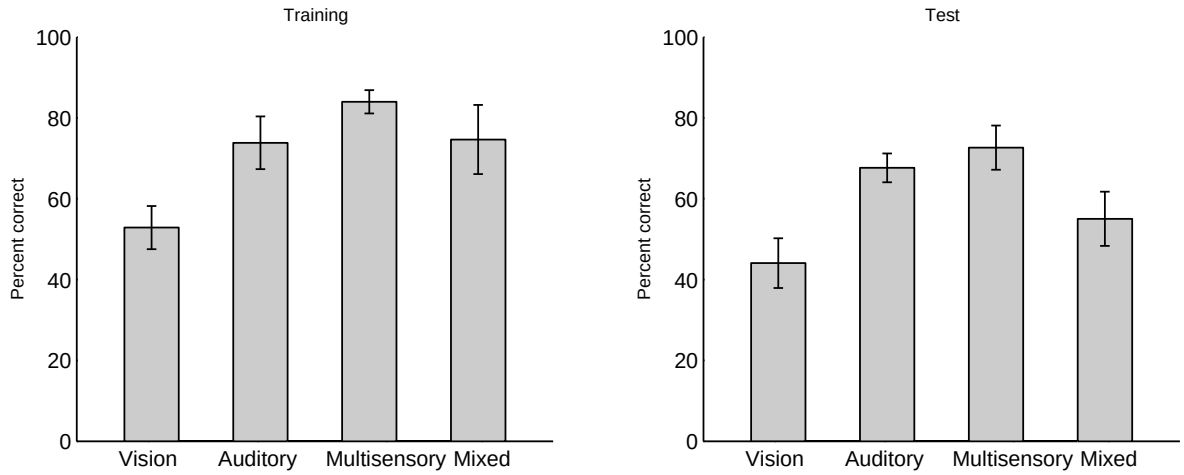


Figure 4: Categorization performances of the vision-only, auditory-only, multisensory, and mixed models on the training set (left) and on the test set (right). The horizontal axis of each graph gives the model, and the vertical axis plots the percent of data items correctly classified (error bars indicate the standard deviations of these percents across iterations of an MCMC chain).

tion of these distances has values of 3.49, 7.83, and 19.04 as its 25th, 50th, and 75th percentiles.

In summary, both training and test sets suggest that the multisensory perception model did indeed acquire sensory invariant representations. Its latent multisensory features had the same or similar distributions regardless of whether a speech utterance was seen or heard.

Predicting sensory representations in missing modalities

Above, we reviewed evidence of activity in people's auditory cortices when they viewed speech utterances but did not hear those utterances (Calvert et al., 1997). This result is consistent with the hypothesis that sensory representations in one modality can predict or activate representations in other modalities. Does the multisensory perception model show this behavior?

This question was studied using the data items in the test set. Let $\mathcal{V}$ and $\mathcal{A}$ denote the sets of visual and auditory feature representations for the data items in the training set. Once again, let $\mathcal{L}_i$ denote the set of multisensory representations obtained on iteration i of the MCMC sampler when the model was trained on the training data. For each test item, we computed the probability distribution of an auditory representation given a test item's visual features. This was accomplished by first calculating a conditional joint distribution over both multisensory and auditory representations, and then by marginalizing over the multisensory representations where the set of possible auditory and multisensory representations were given by $\mathcal{A}$ and $\mathcal{L}_i$. Analogous computations were carried out to compute the distibution of a visual representation given an item's auditory features.

Representative results are shown in Figure 6. Four test items (items 1, 12, 24, and 28) were selected at random with the constraint that one item corresponded to each spoken digit. The four graphs in the top row of the figure show the distributions of the visual representations given the auditory features of the test items. More precisely, the graphs show that when presented with the auditory features corresponding to one of the digits, the model's distribution of visual representations was tightly peaked at a representation corresponding to the same digit. The four graphs in the bottom row show analogous results for distributions of auditory representations given test items' visual features.

In summary, the multisensory perception model learns to associate unisensory representations from different modalities. It successfully predicts representations from missing modalities based on features from observed modalities.

Conclusions

Bayesian nonparametric approaches to modeling are becoming increasingly popular in the cognitive science and machine learning literatures. We regard this approach as an important advance over conventional parametric approaches in which a researcher sets the number of latent variables by hand, often in an ad hoc or unprincipled manner. How can a researcher be sure that the number of latent features should, for example, be exactly 10? Shouldn't the number of latent features be determined by the structure of the task or data set? The Bayesian nonparametric approach is also an advance over modeling approaches that define a set of models, each with a different number of latent features, and perform "model comparison" to select the best model. Typical model comparison techniques are computationally expensive and, thus, only practical for comparing small numbers of models. How should a researcher pick a small number of models to consider? The Bayesian nonparametric approach eliminates (or at least ameliorates) the problems associated with model comparison.

We have proposed a Bayesian nonparametric model of multisensory perception. The model includes a set of latent vari-

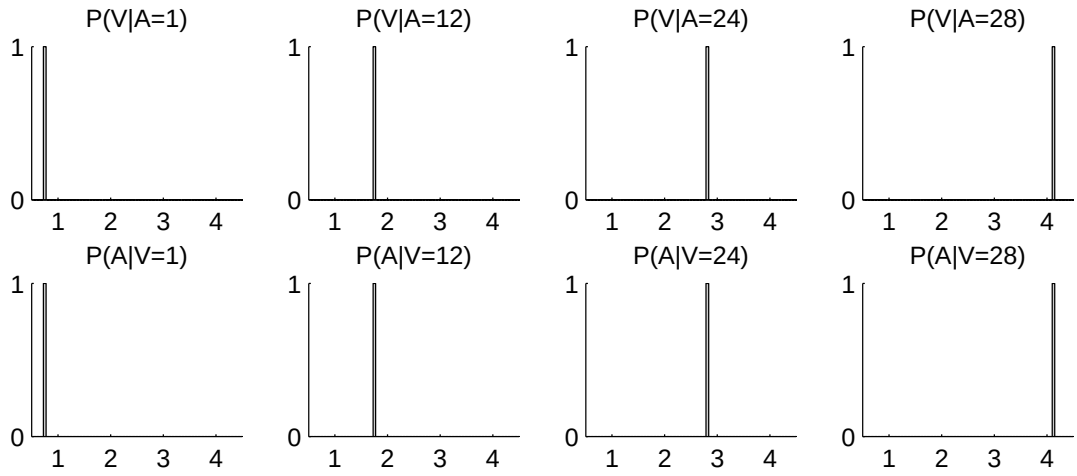


Figure 5: Graphs in the top row demonstrate that when presented with auditory features of a test item corresponding to one of the digits, the multisensory perception model's distribution of visual representations was tightly peaked at a representation corresponding to the same digit. Graphs in the bottom row show analogous results for distributions of auditory representations given test items' visual features.

ables that learn multisensory features from unisensory data. The model is highly flexible because it makes few statistical assumptions. In particular, the number of multisensory features is not fixed a priori. Instead, this number is estimated from the data.

We applied the model to a real-world visual-auditory data set obtained when people spoke English digits. Our results are consistent with several hypotheses about multisensory perception from the cognitive neuroscience literature. We found that the model obtained the statistical advantages provided by sensory integration. We also found that the model acquired multisensory representations that were relatively sensory invariant. Lastly, we found that the model was able to associate unisensory representations based on different modalities.

Because the multisensory perception model is based on Bayesian statistics, it can be regarded as an ideal observer inferring optimal multisensory features from unisensory data (Austerweil & Griffiths, 2009). As such, it provides a basis for a rational analysis of multisensory perception. This analysis suggests that the acquisition of latent multisensory representations that are sensory invariant and more statistically robust than latent features from unisensory models is a rational response of an agent attempting to learn the structure of its multisensory environment. It also suggests the rationality of acquiring associations among unisensory representations.

Acknowledgments

We thank J. Drugowitsch, A. E. Orhan, and C. Sims for many helpful discussions, and J. Movellan for making the Tulips1 data set available on the web. This work was supported by a research grant from the National Science Foundation (DRL-0817250).

References

Amedi, A., Malach, R., Hendler, T., Peled, S., & Zohary, E. (2001). Visuo-haptic object-related activation in the ventral visual pathway. *Nature Neuroscience*, *4*, 324-330.

Austerweil, J. L. & Griffiths, T. L. (2009). Analyzing human feature learning as nonparametric Bayesian inference. In D. Koller, Y. Bengio, D. Schuurmans, & L. Bottou (Eds.), *Advances in Neural Information Processing Systems 21*. Cambridge, MA: MIT Press.

Blei, D. M., Griffiths, T. L., & Jordan, M. I. (2010). The nested Chinese restaurant process and Bayesian inference of topic hierarchies. *Journal of the ACM*, *57*, 1-30.

Calvert, G. A., Bullmore, E. T., Brammer, M. J., Campbell, R., Williams, S. C. R., McGuire, P. K., Woodruff, P. W. R., Iversen, S. D., & David, A. S. (1997). Activation of auditory cortex during silent lipreading. *Science*, *276*, 593-596.

Ernst, M. O. & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*, 429-433.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian Data Analysis.* London, UK: Chapman & Hall.

Griffiths, T. L. & Ghahramani, Z. (2005). Infinite latent feature models and the Indian buffet process. Gatsby Unit Technical Report GCNU-TR-2005-001.

Griffiths, T. L. & Ghahramani, Z. (2006). Infinite latent feature models and the Indian buffet process. In B. Schölkopf, J. Platt, & T. Hofmann (Eds.), *Advances in Neural Information Processing Systems 18*. Cambridge, MA: MIT Press.

Movellan J. R. (1995). Visual speech recognition with stochastic networks. In G. Tesauro, D. Toruetzky, & T. Leen (Eds.), *Advances in Neural Information Processing Systems 7*. Cambridge, MA: MIT Press.

Attention and cross-modal processing: Evidence from heart rate analyses

Christopher W. Robinson (robinson.777@osu.edu)
Center for Cognitive Science
The Ohio State University
208F Ohio Stadium East, 1961 Tuttle Park Place
Columbus, OH 43210, USA

Vladimir M. Sloutsky (sloutsky.1@osu.edu)
Center for Cognitive Science
The Ohio State University
208C Ohio Stadium East, 1961 Tuttle Park Place
Columbus, OH 43210, USA

Abstract

The study of cross-modal processing has generated two seemingly contradictory sets of findings. Studies examining cross-modal processing in infants often find evidence that auditory input interferes with visual processing, whereas studies with adults often find evidence for visual input interfering with auditory processing. However, in the absence of amodal measures of auditory processing, it is possible that visual input also interferes with auditory processing in young infants. The primary goal of the current study was to examine this issue by focusing on Heart Rate (HR) to assess discrimination of unimodal auditory stimuli (Experiment 1), and to examine how visual stimuli affect auditory discrimination (Experiment 2). The results indicate that the presence of visual stimuli facilitated, rather than interfered with, auditory processing.

Keywords: Cognitive Development, Attention, Heart Rate, Psychology, Human Experimentation.

Introduction

There are many tasks that require people to integrate information across sensory modalities (e.g., associating words with objects and categories, learning the sounds that objects make, etc.). While simultaneously presenting information to different sensory modalities can sometimes facilitate learning, there are also many occasions when presenting stimuli to one modality interfere with learning in a different modality (i.e., modality dominance). Interestingly, the study of modality dominance has generated seemingly inconsistent findings.

On the one hand, there is more that 30 years of research on the Colavita effect in adults (Colavita, 1974; Colavita & Weisberg, 1979; Klein, 1977; Posner, Nissen, & Klein, 1976, see Sinnett, Spence, & Soto-Faraco, 2007 for a review). The main finding of these studies is that visual information often interferes with the detection of auditory input, hence the "visual dominance effect". On the other hand, studies with infants and young children often demonstrate the opposite finding: auditory input often interferes with visual processing, hence the "auditory dominance effect" (Lewkowicz, 1988a; 1988b; Robinson & Sloutsky, 2004; 2007; 2010; Sloutsky & Napolitano, 2003; Sloutsky & Robinson, 2008).

Although the asymmetry between infant and adult literatures may reflect genuine developmental differences, it is also possible that the asymmetry stems from a lack of appropriate measure of auditory processing. In particular, most infant studies use visual fixations to examine auditory and cross-modal processing. For example, infants in many of the studies reported above were familiarized or habituated to an auditory stimulus, visual stimulus, or to a cross-modal stimulus. Infants in the cross-modal condition often failed to increase looking to a novel visual stimulus when it was paired with an old sound, suggesting that they did not discriminate the visual stimuli. This finding is noteworthy given that infants ably discriminated the same visual stimuli when they were presented unimodally.

In contrast, there were no costs of cross-modal presentation on auditory processing: infants equally discriminated auditory stimuli when presented unimodally and cross-modally. However, auditory processing was never measured independently of visual processing (i.e., auditory processing was assessed by examining infants' visual fixations). In the absence of a true measure of auditory processing, it is possible that visual dominance was missed, with visual input interfering more with auditory input than the reverse. The goal of the present research was to address this issue.

The achievement of this goal requires an amodal measure of auditory processing. While sucking procedures and ERP tasks can provide modality-independent measures of auditory processing (e.g., Eimas, Siqueland, Jusczyk, & Vigorito, 1971; Nelson & deRegnier, 1992), sucking procedures are not appropriate for older infants and children and ERP tasks often require a large amount of trials. The present study uses infants' Heart Rate (HR) to

measure auditory and cross-modal processing. HR has provided researchers with a powerful tool for examining the dynamics of visual attention. The gist is that HR decelerates while participants are actively processing visual input, and combining HR and visual fixations can delineate various stages of visual attention (Colombo, et. al., 2004; Courage, Reynolds, & Richards, 2006; Richards & Casey, 1992). Panneton and Richards (2002) used HR to assess how 4- to 6-month-old infants attend to auditory, visual, and cross-modal stimuli. This study demonstrated that HR decelerates more to dynamic and cross-modal stimuli than to static and unimodal stimuli. The current study expands on this research by using HR to examine the effects of visual input on auditory processing. In Experiment 1, we presented participants with unimodal auditory stimuli and measured auditory oddball detection. In Experiment 2, we examined how visual input affected the detection of auditory oddballs.

Experiment 1

Method

Participants Twenty-four 10-month-olds (16 boys and 8 girls, *M* = 301 days, *SD* = 49.94 days) participated in this experiment. A majority of infants were Caucasian and none of the infants had auditory or visual deficits, as reported by parents. No infants were excluded from the final sample.

Apparatus Infants sat on parents' laps 100 cm away from a 52" Sony LCD television. Two Boston Acoustics 380 speakers were 76 cm apart from each other and mounted in the wall (concealed by black felt). A pan-tilt-zoom camera was mounted above the television to capture a video stream of the infant, and a Sony DCR-TRV40 camcorder was located behind the infant to capture the AV stimulus presentation. These two video streams were overlaid using a Kramer PIP 200 picture and picture mixer, and videos were saved as mpg video files on a Dell Optiplex 755 computer.

In an adjacent room, a Dell Optiplex 745 computer with *E-prime* software was used to present stimuli to the infants, and a Dell Optiplex 755 computer with Mindware software was used to record electrocardiograms. Two Ag-AgCl electrodes were placed on the infants' right collar bone and left, lower rib, and a reference electrode was placed on the infants' right, lower rib. Electrocardiograms were collected using a BioNex acquisition unit with a BioNex Impedance Cardiograph and GSC amplifier. Electrocardiograms were time-locked with stimulus presentation and saved on the Dell Optiplex 755 computer.

Stimuli Auditory stimuli were seven nonsense words (e.g., vika, leru, kuna, etc.) that were recorded by a female speaker using infant-directed speech. Each nonsense word was edited in CoolEdit 2000 and saved as a 44.1 kHz, 16-bit stereo wav file. Nonsense words were each 1 s in duration and were presented to infants at approximately 68-70 dB. One nonsense word served as the standard (presented 60% of the time) and the remaining nonsense words served as oddballs. While infrequent stimuli were presented for the remaining 40% of the experiment, six different oddballs were presented throughout the experiment. Thus, across the entire experiment the same standard was presented for approximately 60 s, whereas each individual oddball was only presented for 7 s.

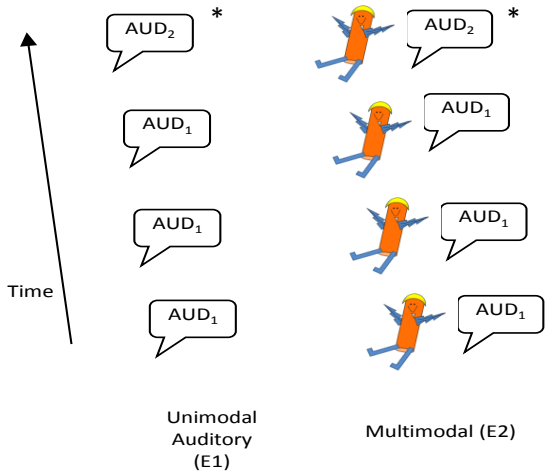


Figure 1: Overview of stimulus presentation in Experiments 1 and 2. Standards were presented five times in a row, followed by four oddballs. Note: "*" denotes an example of oddball in both experiments.

Procedure Infants sat on parents' lap in a quiet, dimly lit room. A picture of a baby playing with toys was presented on the LCD television while the experimenter attached the electrodes to the infant. The experimenter left the room and started the experiment by pressing the spacebar on the Dell Dimension 8200 computer. At this point, the picture of the baby and toys disappeared and a white screen was presented throughout the entire experiment. Infants were presented with alternating standards and oddballs until the infant either became fussy or until all of the stimuli were presented (approximately 1.5 minutes). Stimuli were presented in Trials (i.e., Trial 1 = five presentations of standard → four presentations of oddball 1, Trial 2 = five presentations of standard → four presentations of oddball 2, etc.), such that the same standard was presented throughout the entire experiment and the oddballs changed on every trial. Auditory stimuli were presented for 1 s with a 0.75 s ISI. Thus, within each Trial, the standard was presented for 8.75 s (5 x 1.75 s) and then the oddball was presented for 7 s (4

x 1.75 ms). E-prime sent a pulse to BioNex every time the stimulus changed. For example, E-prime sent a pulse at the onset of the first standard presented in Trial 1. The next pulse was sent at the onset of the first oddball presented in Trial 1, etc. The experiment was not contingent on infants' looking, thus, auditory stimuli were presented as long as the infant was not fussy or interacting with the parent.

Results and Discussion

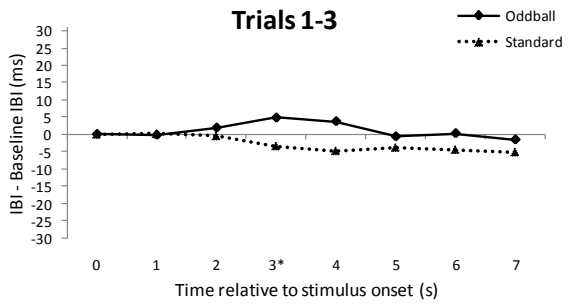
Analyses focused on changes in infants' HR to standards and oddballs across time. Artifacts were corrected using Mindware software, and HR data were transformed to Inter-Beat-Interval (IBI). IBI is inversely related to HR. In particular, as HR slows down, the time between heart beats (distance between R waves) increases. Thus, longer IBIs correspond with slower HR. IBIs were computed by averaging IBIs within a one second bin and baseline corrected. For example, to determine how HR changed 1 s after stimulus onset, we subtracted baseline IBI (IBI 1 s pre-stimulus) from IBI at 1 s post stimulus. To examine how HR changed 2 s after stimulus onset, we subtracted baseline IBI from IBI at 2 s post stimulus. Thus, values greater than zero denote that HR slowed down after stimulus onset and values less than zero denote that HR sped up after stimulus onset.

To examine discrimination of standards and oddballs, we compared IBIs to standards and oddballs averaged across Trials 1-3 (Figure 2a) and averaged across Trials 4-6 (Figure 2b). Paired-sample *t* tests were conducted comparing IBIs to standards and oddballs at each point in time. Reliable differences between standards and oddballs are denoted with a "*" on the x axis. For example, Figure 2a shows that IBIs to standards and oddballs only differed 3 s after stimulus onset, $p < .05$. However, as can be seen in Figure 2b, these differences became more pronounced in Trials 4-6. Furthermore, examination of Figure 2b also shows that the difference between standards and oddballs was not solely driven by greater deceleration to oddballs. Rather, HR also accelerated to standards. Examination of video streams suggests that this acceleration may be related to increased infant fidgeting rather than from auditory stimuli startling infants.

To examine how quickly oddballs engaged attention we identified the point for each infant when two consecutive IBIs exceeded baseline (zero). Eight of the 24 infants did not meet this criterion. Averaged across the remaining infants, it took approximately 2.3 s for HR to decelerate. Finally, we examined dwell time of attention to the oddballs (i.e., how long did the oddball hold infants' attention). For example, one of the infants' first of two consecutive IBIs exceeded zero 1 s after stimulus onset and returned to zero 6 s after stimulus onset. Thus, this infants' dwell time of attention was 5 s (HR was decelerated from 1 s – 5 s). On average infants' HRs were decelerated to oddballs for 5 s. However, it is important to note that many of infants' HRs were still decelerated at the end of the trial. Thus, the value of 5 s underestimates how long the oddballs actually held infants' attention.

In summary, the findings from Experiment 1 demonstrate that HR can serve as a modality-independent measure of attention to assess auditory discrimination in a relatively short period of time, and these discriminations appeared to develop gradually across the experiment. In addition to providing time course information across trials, changes in HR can also provide a measure of speed of engagement and dwell time of attention within trials.

a.



b.

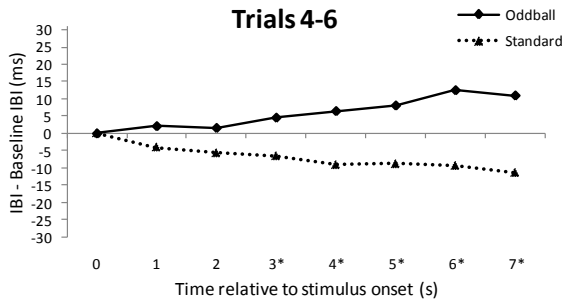


Figure 2: Change in IBI to standards and oddballs in Trials 1-3 (a) and Trials 4-6 (b). Note: "*" on the x axis denote means at that point in time were reliably different, *p*s < .05 (one-tailed).

Experiment 2

The goal of Experiment 2 was to examine how visual input affects discrimination of the auditory stimuli presented in Experiment 1. More specifically, we examined how pairing an old visual stimulus with a novel auditory oddball would affect discrimination, speed of engagement, and dwell time of attention.

Method

Participants Eight 10-month-olds (3 boys and 5 girls, *M* = 309 days, *SE* = 56 days) participated in this experiment. Demographics were identical to Experiment 1. Five infants were tested but were not included in the final sample due to fussiness (n=3) and experimenter error (n = 2).

Stimuli and Procedure The auditory stimuli were identical to Experiment 1, however, in the current experiment, auditory stimuli were paired with a visual stimulus (see Figure 1). The visual stimulus consisted of a novel creature that was created in PowerPoint and saved as a 400 x 400 pixel jpg. The visual stimulus was presented on the 52" Sony LCD and pulsated at the same rate as the auditory stimulus (1 s stimulus duration with a 0.75 s ISI). The procedure also differed from Experiment 1 in one important way. In the current experiment, the procedure paused and the screen darkened when infants looked away. The experiment started back up again when the infant looked to the darkened screen. This manipulation was important because we were interested in how the presence of an old visual stimulus affected auditory processing. Therefore, we only examined discrimination of auditory stimuli on those trials where the infants were looking to the visual stimulus. Trials where the infant looked away were discarded.

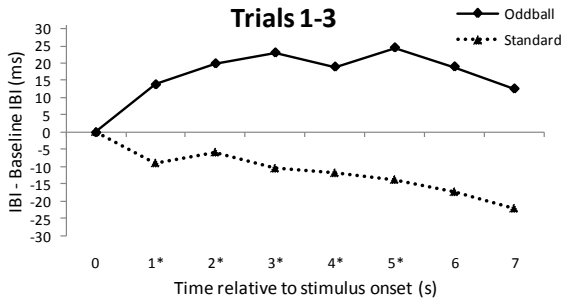
Results and Discussion

As in Experiment 1, we examined discrimination of standards and oddballs in Trials 1-3 (Figure 3a) and in Trials 4-6 (Figure 3b). Paired-sample *t* tests (one-tailed) were conducted to compare discrimination of standards and oddballs at each point in time. In contrast to Experiment 1, infants reliably discriminated auditory standards and oddballs in Trials 1-3 (see asterisks on the x axis to determine which means reliably differed from each other). This suggests that the presence of the visual stimulus actually facilitated auditory discrimination, with infants discriminating oddballs and standards early in the course of processing. Discrimination was also robust in the last three trials of the experiment (see Figure 3b).

As in Experiment 1, we also examined how quickly oddballs engaged attention and how long oddballs held attention. Two of the 8 infants never had two consecutive IBIs exceed zero. Averaged across the remaining infants, it took approximately 1.1 s for the oddballs to engage attention. Recall that infants in Experiment 1 took approximately 2.3 seconds. Therefore, the old visual stimulus did not appear to slow down the detection of the auditory oddballs. Furthermore, infants' HR in the current experiment was decelerated to oddballs for approximately 5.8 s, which was slightly longer than in Experiment 1. However, as in Experiment 1, many infants' HRs were still decelerated at the end of the trial. Therefore, it is unclear if differences between Experiments 1 and 2 would have emerged if infants would have been given more time for HR to return to baseline levels.

In summary, Experiment 2 demonstrates that visual stimuli did not attenuate discrimination of auditory input or slow down the speed in which auditory oddballs engaged attention. Rather, cross-modal presentation in the current experiment actually facilitated auditory processing. Recall that infants in the current experiment (but not in Experiment 1) reliably discriminated standards from oddballs in Trials 1-3. Furthermore, these effects were much stronger in Experiment 2, with reliable discrimination occurring with a sample size of only eight infants. Finally, it is worth noting that all data reported in Experiment 2 came from trials when infants were looking throughout the entire trial. Therefore, analysis of looking data would suggest no discrimination of standards and oddballs, whereas HR data clearly demonstrate that infants discriminated these stimuli.

a.



b.

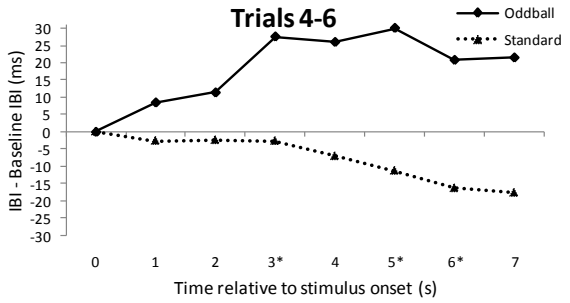


Figure 3: Change in IBI to standards and oddballs in Trials 1-3 (a) and Trials 4-6 (b). Note: "*" on the x axis denotes mean at that point in time were reliably different, *p*s < .05 (one-tailed).

General Discussion

The current study reveals several important findings. First, Experiment 1 demonstrates that HR can provide a powerful tool for examining auditory processing. In particular, changes in HR to frequent and infrequent stimuli can provide a measure of auditory discrimination. Furthermore, speed of engagement and dwell time of attention can also be estimated by examining when HR decelerates compared to pre-stimulus levels and by examining how long HR remains decelerated. More importantly, Experiment 2 demonstrates that visual input facilitated, rather than interfered with, auditory processing.

These findings suggest that the differences in modality dominance between infants and adults do not stem from an underestimation of visual interference with auditory processing in infants. Rather, these findings suggest that the difference may actually reflect a real developmental phenomenon, with allocation of attention to multimodal stimuli changing in the course of development.

Acknowledgments

This research has been supported by grants from the NSF (BCS-0720135), NIH (R01HD056105) and from the US Department of Education (R305H050125 and R305B070407) to Vladimir Sloutsky and from NIH (RO3HD055527) to Chris Robinson.

References

Colavita, F. B. (1974). Human sensory dominance. *Perception & Psychophysics*, *16*, 409-412.

Colavita, F. B., & Weisberg, D. (1979). A further investigation of visual dominance. *Perception & Psychophysics*, *25*, 345–347.

Colombo, J., Shaddy, D. J., Richman, W. A., Maikrantz, J. M., & Blaga, O. (2004). The developmental course of habituation in infancy and preschool outcome, *Infancy*, *5*, 1-38.

Courage, M.L., Reynolds, G.D., & Richards, J.E. (2006). Infants' visual attention to patterned stimuli: Developmental change and individual differences from 3- to 12-months of age. *Child Development*, 77, 680-695.

Eimas, P.D., Siqeuland, E.R., Jusczyk, P., & Vigorito, V. (1971). Speech perception in infants. *Science*, *171*, 303-306.

Klein, R. M. (1977). Attention and visual dominance: A chronometric analysis. *Journal of Experimental Psychology: Human Perception & Performance*, *3*, 365-378.

Lewkowicz, D. J. (1988a). Sensory dominance in infants: 1. Six-month-old infants' response to auditory-visual compounds. *Developmental Psychology*, *24*, 155-171.

Lewkowicz, D. J. (1988b). Sensory dominance in infants: 2. Ten-month-old infants' response to auditory-visual compounds. *Developmental Psychology*, *24*, 172-182.

Nelson, C.A., & deRegnier, R.A. (1992). Neural correlates of attention and memory in the first year of life. *Developmental Neuropsychology*, *8*, 119–134.

Pannenton, R., & Richards, J.E. (2002). *Differential heart-rate activity in infants to uni- and multimodal events. Society for Psychophysiological Research*, Washington, D.C.

Posner, M. I., Nissen, M. J., & Klein, R. M. (1976). Visual dominance: An information-processing account of its origins and significance. *Psychological Review*, *83*, 157-171.

Richards, J. E., & Casey, B. J. (1992). Development of sustained visual attention in the human infant. In B. A. Campbell, H. Hayne, & R. Richarson (Eds.), *Attention and information processing in infants and adults*, pp. 30-60. Hillsdale, NJ: Erlbaum.

Robinson, C. W., & Sloutsky, V. M. (2004). Auditory dominance and its change in the course of development. *Child Development*, *75*, 1387-1401.

Robinson, C. W., & Sloutsky, V. M. (2007). Linguistic labels and categorization in infancy: Do labels facilitate or hinder?. *Infancy*, *11*, 233-253.

Robinson, C. W., & Sloutsky, V. M. (2010). Development of Cross-modal Processing. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*, 135-141.

Sinnett, S., Spence, C., & Soto-Faraco, S. (2007). Visual dominance and attention: Revisiting the Colavita effect. *Perception & Psychophysics*, *69*, 673–686.

Sloutsky, V. M., & Napolitano, A. (2003). Is a picture worth a thousand words? Preference for auditory modality in young children. *Child Development*, *74*, 822-833.

Evidence for auditory dominance in a passive oddball task

Christopher W. Robinson (robinson.777@osu.edu)
Center for Cognitive Science
The Ohio State University
208F Ohio Stadium East, 1961 Tuttle Park Place
Columbus, OH 43210, USA

Nayef Ahmar (ahmar.1@osu.edu)
Center for Cognitive Science
The Ohio State University
208F Ohio Stadium East, 1961 Tuttle Park Place
Columbus, OH 43210, USA

Vladimir M. Sloutsky (sloutsky.1@osu.edu)
Center for Cognitive Science
The Ohio State University
208C Ohio Stadium East, 1961 Tuttle Park Place
Columbus, OH 43210, USA

Abstract

Simultaneous presentation of auditory and visual input can often lead to visual dominance. Most studies supporting visual dominance often require participants to make an explicit response, therefore, it is unclear if visual input disrupt encoding/discrimination of auditory input or results in a response bias. The current study begins to address this issue by examining how multimodal presentation affects discrimination of auditory and visual stimuli, while using a passive oddball task that does not require an explicit response. Participants in the current study ably discriminated auditory and visual stimuli in all unimodal and multimodal conditions. Furthermore, there was no evidence that visual stimuli attenuated auditory processing. Rather, multimodal presentation sped up auditory processing (shorter latency of P300) and slowed down visual processing (longer latency of P300). These findings are consistent with research examining modality dominance in young children and suggest that visual dominance effects may be restricted to tasks that require an explicit response.

Keywords: Attention, Cross-modal Processing, Electroencephalograph (EEG), Neurophysiology, Psychology.

Introduction

Most of our experiences are multimodal in nature. The objects and events that we encounter in the environment can be seen, touched, heard, and smelled. The fact that the brain can integrate this knowledge into a coherent experience is amazing given that each modality simultaneously receives different types of input, and this information is processed, at least in the early stages of processing, by dedicated sensory systems.

While multimodal presentation can sometimes facilitate learning, there are many occasions when presenting information to one sensory modality interferes with learning in a second modality. These modality dominance effects can occur on detection tasks and on more complex discrimination tasks, with auditory input often attenuating visual processing in young children (Sloutsky & Napolitano, 2003; Robinson & Sloutsky, 2004) and visual input often attenuating auditory processing in adults (Colavita, 1974; Colavita & Weisberg, 1979).

Support for visual dominance in adults comes from a long history of research examining how multimodal stimuli affect the detection of auditory and visual input (Colavita, 1974; Colavita & Weisberg, 1979; Klein, 1977; Posner, Nissen, & Klein, 1976; see also Sinnett, Spence, & Soto-Faraco, 2007; Spence, Shore, & Klein, 2001, for reviews). For example, in a classic study Colavita (1974) presented adults with a tone, a light, or the tone and light paired together. Participants had to press one button when they heard the tone and a different button when they saw the light. While participants were accurate when the tone and light were presented unimodally, they often responded to the visual stimulus when the stimuli were paired together, with many adults failing to detect the auditory stimulus. This finding has been replicated using a variety of stimuli and procedures, with little evidence demonstrating that auditory input attenuates visual processing in adults (see Sinnett, Spence, & Soto-Faraco, 2007 for a review).

There appears to be an attentional component underlying visual dominance (Posner, Nissen, & Klein, 1976). In particular, the underlying idea is that the auditory and visual modalities share the same pool of attentional resources. While auditory stimuli

automatically engage attention, visual stimuli often have poor alerting abilities. To compensate for the poor alerting ability of visual input, adults endogenously direct attention to visual stimuli. This increased attention to the visual modality comes with a cost – attenuated auditory processing.

While there is much support for visual dominance, it is important to note that this support primarily comes from studies examining response latencies and response accuracies. Therefore, it is possible that visual input have no effect on encoding or discrimination of auditory stimuli. Rather, these effects may stem solely from visual input dominating the response. The current study begins to address this issue by examining processing of auditory, visual, and multimodal stimuli in a task that does not require an explicit response.

Participants in the current study were presented with a passive oddball task where they were presented with auditory, visual, or multimodal stimuli. Event Related Potentials (ERPs) were recorded as adults passively attended to frequent stimuli (standard) and infrequent stimuli (oddballs). The signature pattern of discrimination is a P300. P300 is a positive component with a peak latency occurring between 300-800 ms after stimulus onset and is strongest over the temporal, parietal, and fronto-central regions (see Polich & Criado, 2006 for a review). The amplitude of P300 is larger for novel or infrequent stimuli (Sutton, Braren, Zubin, & John, 1965), and the latency of P300 can be used as a measure of processing time (Kutas, McCarthy, & Donchin, 1977). In particular, experimental manipulations that affect the processing leading up to classification and responding should affect the latency of P300. The same underlying idea is guiding the current research: multimodal facilitation and interference should manifest themselves by affecting the latency (and possibly amplitude) of P300.

Previous studies have used oddball tasks to examine unimodal and multimodal processing and to examine effects of response on ERP components. However, these procedures differed from the ones reported here in several important ways. First, ERP studies that have directly compared unimodal and multimodal conditions either focused on early ERP components associated with stimulus detection or they required participants to make a response to oddballs (e.g., Brown, Clarke, & Barry, 2007; Fort, Delpuech, Pernier, & Giard, 2002; Giard & Peronnet, 1999; Vidal, Giard, Roux, Barthelemy, & Bruneau, 2008). In contrast, the current study focused exclusively on discrimination of standards and oddballs (P300), and participants did not make an explicit response to these stimuli. Second, the studies that have examined the effects of explicit response on oddball tasks were not interested in modality dominance, thus, they did not examine discrimination of the same auditory and visual stimuli when presented unimodally and multimodally (Mertens & Polich, 1997; Wronka, Kaiser, & Coenen, 2008).

Thus, to the best of our knowledge this is the first study to use a passive oddball task to examine how multimodal presentation affects auditory and visual processing. If visual stimuli interfere with the encoding and/or discrimination of auditory stimuli, then the latency of P300 should occur later in the multimodal condition than in the unimodal condition. However, if visual stimuli only affect the response, then no effects should be found or auditory input may attenuate visual processing (auditory dominance).

Methods

Participants

Thirty-nine undergraduate students from The Ohio State University (23 men and 16 women, *M* = 19.5 years, *SD* = 3.9 years) participated in this experiment for course credit. Prior to the experiment all participants gave informed consent and provided basic personal information (handedness, age, medical history). All participants had normal hearing and normal (or corrected to normal) vision.

Stimuli

The stimuli and cover story were designed for young children. The visual stimuli consisted of six novel creatures that were created in PowerPoint and exported as 400 x 400 pixel jpeg images (see Figure 1 for examples).

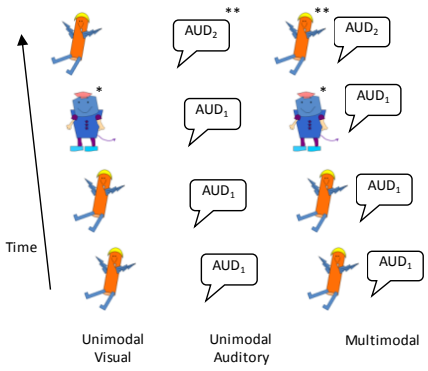


Figure 1: Example of stimuli and overview of the visual, auditory, and multimodal conditions. Note: "*" denotes visual oddball and "**" denotes auditory oddball.

Visual stimuli were presented centrally on a Dell 17" LCD monitor for 480 ms. The interstimulus interval (ISI) randomly varied from 1000 ms - 1520 ms. Auditory stimuli were also 480 ms in duration, with a 1000 ms - 1520 ms ISI. The auditory stimuli

were dynamic sounds that changed in pitch and amplitude across time. The sounds were created in CoolEdit 2000 by using preset functions (e.g., DTMF signal, out of control, etc.). Stimuli in the multimodal condition were constructed by pairing the auditory and visual stimuli together (see Figure 1 and Table 1). Thus, the same stimuli were used in the unimodal and multimodal conditions, therefore, any differences found between these conditions cannot be accounted for by properties of the unimodal stimuli.

Procedure

Three different oddball tasks were used (see Figure 1 and Table 1), and task order was pseudo-randomized for each participant. Approximately half of the participants were presented with the unimodal oddball tasks (order of auditory and visual was randomized for each participant), and then they participated in the multimodal task. The remaining participants were presented with the multimodal task, followed by the two unimodal tasks (order randomized for each participant). The multimodal task took approximately 40 minutes, and each unimodal task took approximately 20 minutes.

As can be seen in Table 1, each task consisted of one standard (presented approximately 80% of the time), four oddballs (each presented approximately 4% of the time), and one novel (presented approximately 4% of the time). Participants were instructed to press a button every time they saw/heard the novel, and to not respond to the standards and oddballs (see cover story). The novel trials were presented to keep participants engaged, and ERPs from these trials were discarded. Four oddballs were used to keep the task interesting for participants and to maintain a low probability of oddballs (each oddball was only presented 4% of the time). In each of the unimodal conditions there were four oddballs, and in the multimodal condition, there were eight oddballs (four auditory and four visual). To examine how multimodal stimuli affect auditory processing, we compared auditory oddballs in the silent condition (e.g., A2, A3, etc.) with the same auditory oddballs in the multimodal condition (e.g., A2V1, A3V1, etc.). To examine how multimodal stimuli affect visual processing, we compared visual oddballs in the silent condition (e.g., V2, V3, etc.) with the same visual oddballs in the multimodal condition (e.g., A1V2, A1V3, etc.).

Prior to each task participants were told a short cover story. For example, in the unimodal visual task, participants were told: *You are going to see creatures from a far away place. Most of the creatures that you will see eat vegetables. However sometimes you will see this creature (novel was presented). This creature eats cookies. In this game you have to press a button every time you see this creature that eats cookies (novel was presented). Do not press any buttons when you see any of the other creatures.* In the auditory condition they were told that they would hear the sounds of creatures eating vegetables and cookies, and in the multimodal condition they were told that they would see creatures and hear the sounds that they make while eating vegetables and cookies.

	Unimodal Auditory	Unimodal Visual	Multimodal
Standard	A1 (280)	V1 (280)	A1V1 (560)
Oddballs (A)	A2 (16)		A2V1 (16)
	A3 (16)		A3V1 (16)
	A4 (16)		A4V1 (16)
	A5 (16)		A5V1 (16)
Oddballs (V)		V2 (16)	A1V2 (16)
		V3 (16)	A1V3 (16)
		V4 (16)	A1V4 (16)
		V5 (16)	A1V5 (16)
Novel	A6 (16)	V6 (16)	A6V6 (16)

Table 1. Overview of stimuli and tasks (frequency of each stimulus).

Participants were presented with a warm up task where they were given 10 standards and 3 novels. ERPs from the warm up task were not included in the final data. Feedback was provided throughout the entire experiment. Feedback was provided if participants: (a) responded to a standard or oddball or (b) failed to respond to a novel. All data were recorded with eyes open and participants in the unimodal auditory condition were asked to fixate on a square taped to the top of the LCD monitor.

Recording Conditions and Data Acquisition

Experiments were conducted in a sound-attenuated, illuminated, and well-ventilated presentation chamber which housed a Dell 17" monitor, two Polk PLKRC65I wall mount speakers, and a response pad. In the experimenter room, a Dell Optiplex 755 computer with E-prime software v.2.0.8.22 was used to present stimuli to participants, and a Harman Kardon AVR-154 receiver was used to amplify the sounds. Timing tests were conducted to ensure that auditory and visual stimuli were presented simultaneously. Offsets between trigger registration and stimuli presentation were measured for unimodal and multimodal conditions and were adjusted during analysis. A PowerPC G5 Mac with Netstation software was used to record and store ERP data.

ElectroEncephalography (EEG) brain activity was recorded using a 128-channel HydroCel Geodesic Sensor Net (Electrical Geodesics, Inc., Eugene, OR). Scalp-electrode impedances were kept below 50 kOhms. All channels were referenced to Cz during

acquisition. EEG was recorded using a 0.1 to 100 Hz band-pass filter (3 dB attenuation), amplified at a gain of 1000, sampled at a rate of 250 Hz, and digitized with a 24-bit A/D converter.

Data analysis

Participants ably discriminated novels in all of the conditions (proportion of hits to novels – proportion of false alarms to standards + oddballs > .99). Because auditory and visual components both changed on novel trials and participants made a response, it is unclear if ERP waveforms reflect auditory discrimination, visual discrimination, or the response. Therefore, data from novel trials were not included in any of the analyses.

ERPs to standards and oddballs were processed using Netstation waveform tool. EEGs were band-passed between 0.1 Hz and 30 Hz and segmented between 100 ms pre-stimulus onset and 1000ms post stimulus onset. ERPs were referenced with respect to the average of all channels after correcting for bad trials using neighboring channels. Trials were then baseline corrected with respect to the 100 ms pre-stimulus and then exported to Matlab.

Initially, we looked at 8 different scalp regions, each comprising of 6 or 7 channels from the 10/20 system representing: F3, F4, P3, P4, T3, T4, Pz, and Oz. However, in the current study we focused exclusively on Pz; the region that provided the best measure of discrimination in all conditions (see Figure 2). In each of the unimodal conditions, participants provided two ERP waveforms (one for the standard and one for the oddball). To equate the number of standards and oddballs, we randomly picked and averaged 64 of the 280 standards and we averaged across the four different oddballs. In the multimodal condition, adults provided a waveform for the standard, a waveform for auditory oddballs, and a waveform for visual oddballs (see Table 1).

Results and Discussion

A reliable P300 was found at Pz, P3, and P4, however, as mentioned above, discrimination was most pronounced at Pz. Thus, analyses reported below focus on Pz between 250-650 ms after stimulus onset. Waveforms for the unimodal conditions are presented on the left side of Figure 2 and waveforms for the multimodal conditions are presented on the right side of Figure 2. As can be seen in Figure 2, participants ably discriminated auditory and visual stimuli when presented unimodally and multimodally. Mean averages were computed for each participant. For example, to assess discrimination of the auditory stimuli in the unimodal auditory condition, we computed a mean average for the standard (between 250-650 ms) and a mean average for the oddball (between 250-650 ms) for each participant. These means were then submitted to a one-way ANOVA with trial type (standards vs. oddball) as a repeated measure. The same analyses were conducted in the four conditions (i.e., Unimodal Auditory, Unimodal Visual, Multimodal Auditory, and Multimodal Visual). All ANOVAs were significant, Fs > 20, ps < .0001.

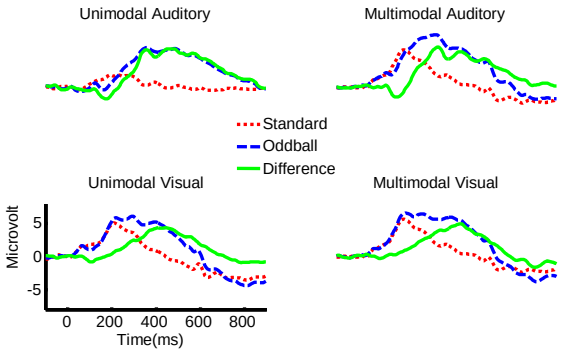


Figure 2. ERP waveforms for Standards and Oddballs across conditions. Solid line represents difference waves (Oddball – Standard).

To examine the effects of visual input on auditory discrimination, we compared the auditory difference waveform in the multimodal condition to the auditory difference waveform in the unimodal condition (see Figure 3a). A one-way AVOVA revealed that mean amplitude between 250-650 ms did not differ between the unimodal and multimodal conditions. We also examined how the presence of auditory input affected visual discrimination by comparing the visual difference waveform in the multimodal condition to the visual difference waveform in the unimodal condition (see Figure 3b). As in the auditory conditions, mean amplitude did not differ between the unimodal and multimodal conditions.

To statistically find and quantify any significant displacement of P300 between unimodal and multimodal conditions, we computed the fractional area latency. In particular, for a predefined window, we measured the area under the curve, and then we found the latency that divided that area into two equal parts (see Hansen & Hillyard, 1980). Using this measure for a window between 250 ms and 650 ms, we found that multimodal presentation sped up auditory discrimination by 26ms and slowed down visual discrimination by 12 ms.

However, as can be seen in Figure 3a, there are multiple peaks in both auditory conditions that could be the result of multiple underlying components. Therefore, we ran sliding windows of 200 ms, 300 ms, and 400 ms for each participant's data covering the whole time range of interest (250ms to 650ms). That is, for each window length, centered at a time sample, we computed the 50% area latency for both

the multimodal difference waveform and for the unimodal difference waveform. We then calculated a difference wave (Difference Multimodal – Difference Unimodal) to denote the displacement. We kept doing this while sliding the window at 4ms increments. Figure 4a – 4c plot the displacement waveforms for the 200 ms, 300 ms, and 400 ms windows, respectively. Values greater than zero denote that multimodal presentation increased the latency of P300 and values less than zero denote that multimodal presentation shortened the latency of P300.

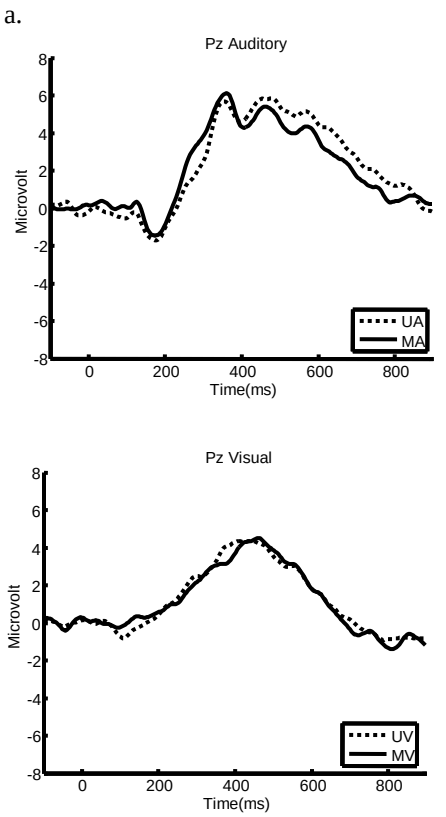


Figure 3. (a) Difference waves for Unimodal Auditory (UA) and Multimodal Auditory (MA), (b) Difference waves for Unimodal Visual (UV) and Multimodal Visual (MV). All data are averaged across participants.

As can be seen in Figures 4a-4c, across all windows, multimodal presentation sped up auditory processing and slowed down visual processing. ANOVAs were conducted for each window at every 4 ms increment. Using a window size of 200 ms, auditory and visual displacement waves differed from 370 ms to 514 ms, *p*s < .05. Using a window size of 300 ms, auditory and visual displacement waves differed from 362 ms to 534 ms, *p*s < .05. Finally, using a window size of 400 ms, auditory and visual displacement waves differed from 370 ms to 554 ms, *p*s < .05. These findings suggest the multimodal presentation had different effects on auditory and visual processing, with multimodal presentation increasing the latency of the visual P300 and shortening the latency of the auditory P300.

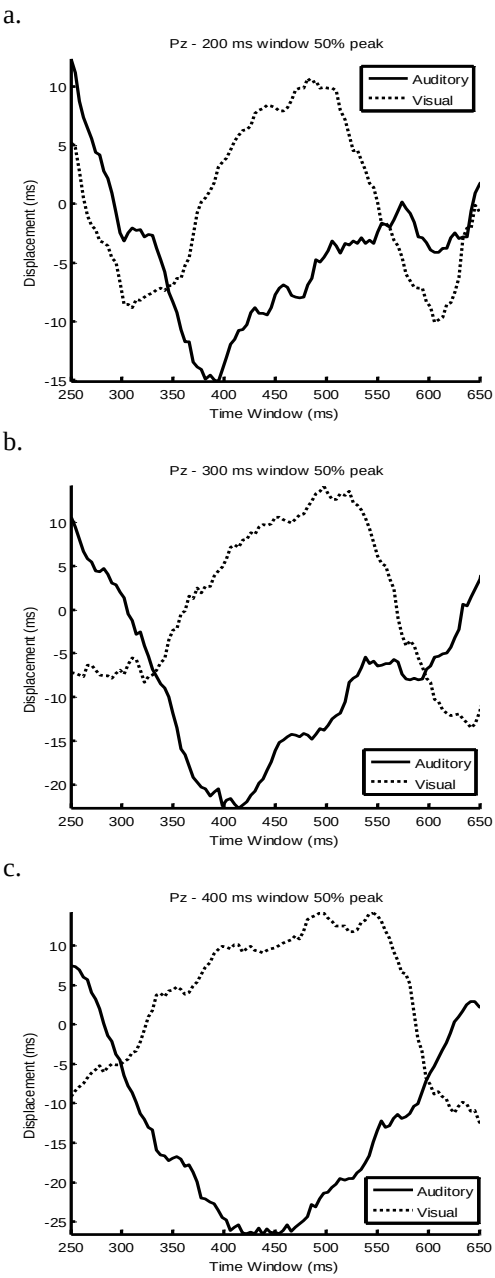


Figure 4. Displacement for (a) 200 ms window, (b) 300 ms window, (c) 400 ms window.

General Discussion

The current study used a passive oddball task to examine the time course of auditory and visual processing when stimuli were presented unimodally and multimodally. As can be seen in Figures 2, 3a, and 4a-4c, there was no evidence that visual input attenuated discrimination of auditory stimuli. Rather,

multimodal presentation appeared to speed up auditory processing and slow down visual processing. These findings have important implications for understanding the underlying mechanisms and time course of modality dominance. In particular, the current findings suggest that some of the effects of visual dominance may stem from visual input dominating the response. However, future research will need to make direct comparisons on tasks that do and do not require explicit responses before any strong conclusions can be drawn.

The novelty of the current research is that we examined the time course of auditory and visual processing on a task that did not require an explicit response. The results replicate auditory dominance effects found in young children, with multimodal presentation attenuating visual processing, and having no effect or facilitating auditory processing (see Robinson & Sloutsky, 2010 for a review). This interaction suggests that effects cannot solely stem from increased tasks demands, otherwise processing in both modalities would have been delayed. Rather, we believe this interaction stems from the dynamics of cross-modal processing. According to this account (Robinson & Sloutsky, 2010), auditory stimuli quickly engage attention and processing of the details of a visual stimulus does not begin until the auditory modality releases attention. While this account has received some support in young children, the finding that auditory input can also slow down visual processing in adults is novel.

Acknowledgments

This research has been supported by grants from the NSF (BCS-0720135), NIH (R01HD056105) and from the US Department of Education (R305H050125 and R305B070407) to Vladimir Sloutsky and from NIH (RO3HD055527) to Chris Robinson.

References

Brown, C.R., Clarke, A.R., & Barry, R.J. (2007). Auditory processing in an inter-modal oddball task: Effects of a combined auditory/visual standard on auditory target ERPs. *International Journal of Psychophysiology*, *65*, 122-131.

Colavita, F. B. (1974). Human sensory dominance. *Perception & Psychophysics*, *16*, 409-412.

Colavita, F. B., & Weisberg, D. (1979). A further investigation of visual dominance. *Perception & Psychophysics*, *25*, 345–347.

Fort, A., Delpuech, C., Pernier, J., Giard, M.H., (2002). Early auditory–visual interactions in human cortex during nonredundant target identification. *Cogn. Brain Res*, *14*, 20–30.

Giard, M.H., Peronnet, F., 1999. Auditory–visual integration during multimodal object recognition in humans: A behavioural and electrophysiological study. *J. Cogn. Neurosci. 11* (5), 473–494.

Hansen, J.C., & Hillyard, S. A. (1980). Endogenous brain potentials associated with selective auditory attention. *Electroencephalography and Clinical Neurophysiology*, *49*, 277-290

Klein, R. M. (1977). Attention and visual dominance: A chronometric analysis. *Journal of Experimental Psychology: Human Perception & Performance*, *3*, 365-378.

Kutas, M., McCarthy, G., Donchin, E. (1977). Augmenting mental chronometry: the P300 as a measure of stimulus evaluation. *Science, 197*, 792–795.

Mertens, R., & Polich, J. (1997). P300 from a single-stimulus paradigm: Passive versus active tasks and stimulus modality. *Electroencephalography & Clinical Neurophysiology*, *104*(6), 488-497.

Polich, J., & Criado, R. (2006). Neuropsychology and neuropharmacology of P3a and P3b. *International Journal of Psychophysiology*, *60*(2), 172-185.

Posner, M. I., Nissen, M. J., & Klein, R. M. (1976). Visual dominance: An information-processing account of its origins and significance. *Psychological Review*, *83*, 157-171.

Robinson, C. W., & Sloutsky, V. M. (2004). Auditory dominance and its change in the course of development. *Child Development*, *75*, 1387-1401.

Robinson, C. W., & Sloutsky, V. M. (2010). Development of Cross-modal Processing. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*, 135-141.

Sloutsky, V. M., & Napolitano, A. (2003). Is a picture worth a thousand words? Preference for auditory modality in young children. *Child Development*, *74*, 822-833.

Sinnett, S., Spence, C., & Soto-Faraco, S. (2007). Visual dominance and attention: Revisiting the Colavita effect. *Perception & Psychophysics*, *69*, 673–686.

Spence, C., Shore, D. I., & Klein, R. M. (2001). Multisensory prior entry. *Journal of Experimental Psychology: General*, *130*, 799-832.

Sutton, S., Braren, M., Zubin, J., John, E. (1965). Evoked potential correlates of stimulus uncertainty. *Science, 150*, 1187–1188.

Vidal, J., Giard, M-H., Roux, S., Barthelemy, C., & Bruneau, N. (2008). Cross-modal processing of auditory-visual stimuli in a no-task paradigm: A topographic event-related potential study. *Clinical Neurophysiology*, *119*(4), 763-771.

Wronka, E., Kaiser, J., & Coenen, A. M. L. (2008). The auditory P3 from passive and active three-stimulus oddball paradigm. *Acta Neurobiologiae Experimentalis*, *68*(3), 362-372.

Effects of Problem Difficulty and Student Expertise on the Utility of Provided Diagrams in Probability Problem Solving

Eliza J. Bobek (eb357@columbia.edu)
Teachers College, Columbia University
525 W. 125th Street, New York, NY 10027 USA

James E. Corter (jec34@columbia.edu)
Teachers College, Columbia University
525 W. 125th Street, New York, NY 10027 USA

Abstract

This study investigated the use of schema-specific diagrams in probability problem solving. Graduate students enrolled in an introductory probability and statistics course solved four probability problems, with and without instructor-provided diagram hints. Participants' solutions were examined and coded for correctness, use of provided diagrams, and use of student-generated external visual representations. Results show that provided diagram hints helped low-ability students on all but the most difficult problem, while high-ability students were aided by diagrams on the most difficult problem. Implications for the use of diagrams in the development of problem solving proficiency are discussed.

Keywords: probability problem solving; diagrams; visual representations; trees; Venn diagrams; contingency tables

Introduction

Learning probability concepts and solving probability problems can be challenging for students (Garfield & Ahlgren, 1988; Konold, 1989; O'Connell, 1999). Successful probability problem solving requires that students understand complex concepts, and also that they master how and when to use specific formal and procedures that are particular to this domain. External visual representations are commonly used in many types of mathematical problem solving, including probability problem solving (PPS). These representations may promote solution success and student comprehension in several ways: by making abstract concepts visible and manipulable, or by organizing the subparts of a problem in a format that can be tied to solution procedures. Using structured visual representations may help problem solvers invent, retrieve, or apply formal solution schemas, increasing the rate of successful solution.

Why are external visual representations useful in the problem solving process?

Tversky (2001) suggests that external visual representations can serve many purposes, including recording information, relieving working memory, communicating to others, and facilitating inference and discovery. Drawing a diagram can reveal implicit information that is not readily available in a written description and make some pieces of information more explicit. It can also give a problem solver unique insight into the problem's structure or schema. Van Essen & Hamaker (1990) report that the "construction of a drawing might increase the chance that the problem situation is recognized and that the correct schemata is identified among other competing schemata" (p. 311). Other researchers (e.g., Larkin & Simon, 1987; Koedinger & Anderson, 1990) propose that problems solvers possess stronger schemas for diagrams than for words that contain the same information. These diagrammatic schemas may thus have an advantage in problem solving over verbal solution methods. Diagrams may also contribute to a fuller understanding through the use of multiple representations when they are used in conjunction with mental images. Several researchers have suggested that multiple representations may lead to increased "depth" of processing (e.g. Logie & Baddeley, 1990; Mayer & Gallini, 1990; Ainsworth, 2007).

However, a diagram may not help every student at every stage of expertise. Lowrie & Kay (2001) suggest that for students who already have schemas in long-term memory, using a diagram may not be particularly helpful; instead these students are able to generate their own representations. Problem difficulty may also play a role in when students choose to create external visual representations. In their study, Lowrie & Kay (2001) found that elementary age students tended to create external visual representations for especially difficult problems. Since problem difficulty is relative to student ability level, this suggests that individual differences may play a role in diagram use.

Furthermore, the types of external visualizations used in scientific reasoning and problem solving may differ, and be used in different ways (e.g., Edens & Potter, 2008; Van Meter & Garner, 2005). Some representations are relatively abstract, and are commonly used to represent schematic or abstract aspects of the problem. These diagrammatic or schematic representations include tree diagrams, and Venn diagrams used to represent part-whole relationships. These general-purpose diagrams should be distinguished from problem-specific representations that include concrete components of the problem itself.

Other external visual representations may be iconic rather than schematic, including pictures that represent the problem context and sketches that display and/or reorganize the information presented in the problem. The type of

external visual representations used may be influenced by the problem goals and other specific problems features.

Use of external visual representations in probability problem solving (PPS)

In the specific area of probability problem solving, Zahner and Corter (2010) provide evidence that particular types of external visual representations are more often used in particular stages in the problem solving process. Some representations, such as reorganizing the problem information or drawing sketches, are more often used early in the solution process when a problem solver is trying to build a mental model of the problem text. Schematic diagrams, such as an outcome tree, can facilitate abstraction of the text, building a mathematical representation of the problem, and planning of the solution process. In general, different external visual representations may come into play in different stages of the problem solving process because their specific structures afford particular functions.

Quite often, the external inscriptions created by students solving probability problems provide evidence of spontaneously created diagrams and other visual representations. Three types of external visual representations often depicted in textbooks and used in PPS are Venn diagrams, outcome trees, and contingency tables. The structure of these diagrams allows for elements of the problem to be represented externally in an organized way. In an outcome tree, for example, each branch can be used to represent individual probabilities of events. The combination of events can be calculated by multiplying the values assigned to each branch. A Venn diagram is used to organize problem information, typically with overlapping circles to show the union and intersection of events. A contingency table is a matrix structure that shows the frequencies or probabilities of events, to show combinations of events.

Previous studies have pointed to the use of specific visual representations appropriate to specific problem types. Russell (2000) found that the use of outcome trees was correlated with improved performance specifically on conditional probability problems. He also found that students in a probability course used outcome trees more often than contingency tables or Venn diagrams. Interestingly, instructing students to draw an outcome tree did not affect performance. However, students who did draw outcome trees outperformed students who did not.

Zahner & Corter (2010) found that particular external visual representations were associated with specific probability topics, and that particular representations were associated with higher rates of solution correctness for some problem types. Their study suggests that using correct external visual representations may generally be facilitative in problem solving, but this facilitation is difficult to detect because students must first choose the correct diagrammatic representation. Other researchers too have noted this challenge (Novick 1990; 2001, Novick & Hmelo, 1994).

The evidence above shows that spontaneous student use of diagrams is associated with higher rates of solution success. But the causal direction is not entirely clear. Evidence that asking students to create diagrams may have a facilitative effect is provided by Schwartz & Martin (2004), who found that student understanding of statistical concepts was influenced by experimenter-prompted "invention" activities (activities in which students created representations.)

The Present Study

While probability problems can be solved using formulas, we hypothesize that using external visuals may help students overcome comprehension difficulties and may lead to greater problem solving success. From an educational standpoint, we would like to better understand the positive correlation between use of diagrams and problem-solving success. The question is whether drawing correct diagrams leads to better understanding, which facilitates problem solution, or if better understanding enables both the creation of correct diagram and problem solving success. Previous work in this area has shown that a major barrier to success in PPS is problem comprehension and representation. Choosing an appropriate representation is a significant factor in problem solving success, and should be viewed as a skill unto itself (Novick & Hmelo, 1994; Edens & Potter, 2007; Uesaka, Manalo, & Ichikawa, 2007). We hypothesize that cuing or providing diagram "hints" appropriate to the problem type may aid students in the problem comprehension phase because the diagram provides a structure upon which problem components can be mapped. They may also help students to recognize the structure of the problem.

In particular, we are interested in the use of schema-specific external visual representations (distinguished from other representations, such as drawing a picture and reorganizing the given information), because we believe they have a special role in PPS. Thus, three common diagrams used in PPS were selected for use in the study: contingency tables, outcome trees, and Venn diagrams. The study attempted to investigate the role of correct external visual representations by providing appropriate but "generic" diagrams ("diagram hints") directly to students. Each problem was chosen as a prototype of a specific problem topic/type and matched to diagram hints that are commonly used in probability curricula. The problems in this study were typical of those presented in the curricula and students had prior exposure to using specific diagrams for specific problem topics. For example, problem 4 is a conditional probability problem, for which outcome trees are an appropriate representation.

We have three main research questions:

1. Do instructor-provided diagram hints (e.g. a correct but unlabeled Venn diagram) increase the probability of problem solving success on specific problems?

2. Do students actually use the diagram "hints," or are they ignored, or are different external visual representations spontaneously created by students?

3. Does student ability mediate the effectiveness of the use of diagrams as a solution strategy?

Method

Participants. Participants were 129 students recruited from introductory probability and statistics classes at Teachers College, Columbia University. Participants were graduate students in education and social sciences, with a broad range of experience in mathematics.

Materials. Each participant was given four probability problems to solve as a problem set (Figure 1). Half of the participants received blank diagrams for problems 1 and 3 (Version A, n=64); the other half of the participants received diagrams for problems 2 and 4 (Version B, n=65). The diagrams were an outcome tree for problem 1, a contingency table for problem 2, a Venn diagram for problem 3, and an outcome tree for problem 4.

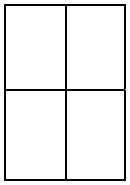
1. A bag of candy contains a mix of jelly beans that includes lime, cherry, and orange flavors. Five jelly beans are cherry, three are orange, and two are lime. Two jelly beans are randomly selected from the bag. What is the probability that the two selected jelly beans include exactly one cherry and one orange?	
2. A survey is conducted on attitudes towards handgun control. 42% of respondents to the survey are urban residents and the rest are rural residents. The results show that 33% of survey respondents are urban residents who support strict handgun controls, while 30% of survey respondents are rural residents who support strict handgun controls. What is the probability that a randomly chosen respondent is a rural resident, given that they support strict handgun control?	
3. A and B are mutually exclusive events. The probability of event A is .3, and the probability of event B is .25. What is $P(A \cap B^C)$?	
4. The weather forecast says that the probability of having good weather tomorrow is .60. If the weather is good, the probability that Eva will go out biking is .80. If it is not good weather, the probability is .20 that she will go out biking. What is the probability that Eva goes out biking tomorrow?	

Figure 1: Probability problems and provided diagrams

Procedure. Participants were allowed to use their class notes to solve the problems, which is standard practice in the course for completing homework assignments and exams. They were given approximately 20 minutes to solve the problems. This time limit was based on a pilot study and was imposed to discourage participants from either quickly scanning the problems or taking an inordinate amount of time.

Coding of participant solutions. Written solutions were coded for several features. First, we coded whether or not the participant gave a correct answer to the problem. Problems were given a score of "0" if incorrect, and a score of "1" if correct. We also totaled student scores for the four problems. Second, we coded whether or not the participant used the instructor-provided diagram hint. Next, we coded for any other type of external visual representation created by the student. The following categories, developed through previous research in our lab, were used to code for the different types of external visual representations: pictures, outcome listings, outcome trees, contingency tables, Venn diagrams, reorganization of given information in the problem, and novel schematic representations (Corter & Zahner, 2007; Zahner & Corter, 2010).

Results

An initial analysis found that over 80% of the participants made use of the instructor-provided diagram for each of the four problems. Analyzing student responses found that the four problems varied in difficulty. Comparing student performance on Version A and Version B allowed us to examine the effect of a diagram hint on the proportion of participants who correctly solved each problem. Table 1 shows the means and standard deviations of participants who correctly solved each problem. Any differences in performance between problems with and without a provided diagram are not statistically significant.

Table 1: Means and standard deviations of correct responses

Problem	Total		Diagram		No Diagram	
	M	SD	M	SD	M	SD
1	0.539	0.500	0.619	0.489	0.462	0.502
2	0.398	0.492	0.339	0.477	0.460	0.502
3	0.305	0.462	0.365	0.502	0.246	0.434
4	0.773	0.420	0.769	0.425	0.778	0.419

We also examined participants' self-generated external visual representations, since we were interested in whether the specific diagram hint we chose to provide was also spontaneously used by students who were not provided with a diagram hint. The problems in this study were typical of those presented in the curricula and students had prior exposure to using specific diagrams for specific problem topics. For example, problem 4 is a conditional probability problem, for which outcome trees are an appropriate representation. Table 2 shows that for all four problems, participants reorganized the given problem information

Table 2: Percentage of participants generating each type of external visual representation for each problem. Dashed lines indicate a cell with fewer than 3 participant uses.

Representation	Problem 1		Problem 2		Problem 3		Problem 4	
	Diagram	No Diagram	Diagram	No Diagram	Diagram	No Diagram	Diagram	No Diagram
Reorganization	43.8	38.4	18.5	75.0	43.8	18.5	35.4	53.1
Pictures	18.8	32.3	--	--	--	--	--	--
Outcome Trees	--	18.5	7.69	6.25	--	--	3.07	45.3
Contingency Tables	--	--	3.07	78.1	--	--	4.61	7.19
Venn diagrams	--	--	--	--	--	--	--	--
Outcome Listings	15.6	15.4	--	--	--	--	--	--
Novel schematic	--	--	--	--	--	--	--	--

more often than any other representation. For problem 1, the provided diagram hint was an outcome tree. Although 18.5% of participants without the diagram also created an outcome tree, 15% of participants given each version also generated outcome listings. For problem 2, the provided diagram was a contingency table, and 78.1% of participants not provided with this diagram chose to create one in solving the problem. A Venn diagram was provided for problem 3; although only 2 participants spontaneously created one, no other representations were used. Finally, problem 4 was accompanied by an outcome tree. 45.3% of students not given an outcome tree created their own in solving the problem.

In order to investigate the role of student ability / problem difficulty on diagram use, we performed a median split on participants' total scores on the four problems, defining two groups of students, low-ability and high-ability. An ANOVA analyzing the effect of provided diagrams showed different effects for these two groups. We hypothesized that the diagram hints might show a facilitative effect only for problems that are hard, but not too hard. Indeed, the pattern of results shows that for both the low-ability and high-ability groups, problems of moderate difficulty were aided by diagrams (Figure 2). "Moderate difficulty" was defined operationally as any diagram showing an overall proportion correct between .3 and .7 for a given ability group. For the below-median group, the problems of moderate difficulty were problems 1 and 4. As seen in Figure 2, problem solving was aided by provided diagrams in these problems, but not for problems 2 and 3. A different pattern emerges for the above-median group. For this group of participants, a facilitative effect is shown for only problem 3, the most difficult problem. An interesting finding is shown for problem 1 in the above-median group. For this problem, providing a diagram (outcome tree) resulted in lower performance than not. It is possible that the outcome tree was not recognized by participants as an appropriate diagram for this problem; indeed outcome listings were spontaneously generated by students, both in the presence and absence of an outcome tree.

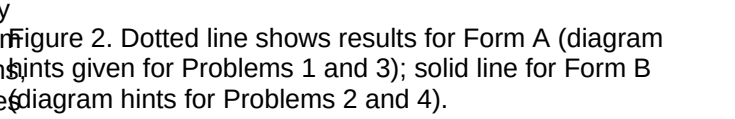


Figure 2. Dotted line shows results for Form A (diagram hints given for Problems 1 and 3); solid line for Form B (diagram hints for Problems 2 and 4).

Discussion

Successful problem solving in mathematics, and especially in PPS, depends on the construction of appropriate representations. External visual representations, including diagrams, are often used to aid in the comprehension and representation of problem information. Diagrams and other external visual devices that are used to comprehend and solve problems are commonplace in the field of

mathematics (e.g. Mayer, 1992). Previous research has revealed that using these schema-specific diagrams can facilitate successful probability problem solving (Russell, 2000; Zahner & Corter, 2010). In this study, we investigated whether the presentation of a diagram was related to problem solving success for each problem.

The participants in this study were novice probability problem solvers; they had only received instruction in PPS for a portion of the semester. We hypothesized that providing them with a schema-specific diagram would influence their success with the comprehension and representation phases of problem solving. The vast majority of participants interacted with the provided diagram in some way. Some participants made marks on the diagram, and also drew a diagram of their own, and some participants filled in the diagram with appropriate numbers and calculations. Many of the students used the diagram to organize and rewrite information. However, the students that used the diagram did not necessarily progress to the stage of comprehending the problem sufficiently to plan a solution. We do not have sufficient evidence to conclude that providing a device particular to solving the problem is necessarily an aid to students at all ability levels. Overall, our results show that providing diagrams does not necessarily help students solve a problem successfully.

Our results show that provided diagrams are able to help low-ability and high-ability students differently (cf. Lowrie & Kay, 2001; Uesaka et al., 2007). High-ability participants may not have been helped by a diagram hint because they already possessed a schematic understanding of the problem. They may have generated their own diagram or used a mathematical formula to solve the problem. Low-ability participants, on the other hand, were helped on the less difficult problems only. We posit that providing a diagram hint helped them form a more complete schematic understanding of the problem and helped them achieve a correct solution. For problems beyond the participants' grasp, however, providing a diagram hint did not help. Students must still know how and when to use the diagram in order for it to be an effective tool. Low-ability students may not have been able to associate the diagrams with a schema appropriate for the problem. We interpret these results in the context of Vygotsky's zone of proximal development (Vygotsky, 1978). Providing a diagram hint on problems of moderate difficulty may be sufficient in helping students relate their current schematic understanding of specific types of probability problems to a solution schema, while more assistance may be needed on problems of greater difficulty.

Many students generated their own diagrams when attempting to solve the problems, even when they were given a correct diagram. In their study of mathematical problem solving, van Essen & Hamaker (1990) found that many students generated external visual representations when solving problems. They posit that "generating a drawing does not guarantee that one finds the correct solution, but merely increases the chance that a problem will be conceptualized correctly" (p. 309). Clearly, the nature of instruction contributes significantly to how students use formulas or diagrams. Students must learn how to use diagrams correctly to solve problems and also receive sufficient practice in order to apply the use of those diagrams to new problems. Previous researchers (e.g., Lewis, 1989; Van Meter & Garner, 2005) argue that learning how to represent a problem is essential, and that it can be taught. This study found that when unsuccessful problem solvers were taught how to represent word problems, their scores on a post-test improved significantly.

Understanding the problem schemas and choosing an appropriate representation is a major barrier to successful problem solving. In other words, diagrams must be understood in order to be helpful. A majority of the participants who were not given a diagram and solved the problem correctly generated their own diagram. Future studies need to examine why people drew fewer diagrams on problems that the data indicate to be more difficult, when research suggests they should do the opposite (e.g. van Essen & Hamaker, 1990).

The participants were all enrolled in a probability course which taught problem solving using the types of diagrams chosen for this study, and these problems were typical of those presented in the course. As students become more proficient in PPS, their associations between problem topic and appropriate diagram use likely become stronger. A notable limitation of this study is that the number of problems studied leaves an alternative explanation of the results, namely that certain diagrams may be easier for students to learn to use and associate with problem structure. For example, problems 1 and 4 could both be solved using an outcome tree; these were also the only problems in which a diagram helped the low-ability group. Thus it may be that outcome trees in particular are helpful to low-ability students. Thus the learnability of the diagrams must be considered as a factor when using them to support students. Future studies examining one type of diagram at a time could help provide information about the properties of the diagrams that make them more or less useful for particular problems. To further support our findings, further research examining the use of diagrams and problem difficulty within a given probability topic is warranted.

Novick (2001) argues that spatial diagrams are "tools for thinking" and that successful construction of these diagrams can lead students to see deep similarities among problem situations. These similarities might otherwise not be prominent. It is important to understand the possible advantages of using external visual representations, as they may help problem solvers to build a mental model and to formulate problem schemas. Further research could explain the choices problem solvers make when solving probability problems. Additionally, this research should explore the role of problem difficulty, problem type, and background knowledge.

Acknowledgments

Support for the investigators' research and dissemination activities has been provided by grants IIS-0725223 and IIS-0855995 from the National Science Foundation.

References

Ainsworth, S. (2006). DeFT: A conceptual framework for considering learning with multiple representations. Learning and Instruction, 16(3), 183-198.

Corter, J. E., & Zahner, D. C. (2007). Use of external visual representations in probability problem solving. Statistics Education Research Journal, 6(2), 22-50, http://www.stat.aukland.ac.nz/serj

Edens, K., & Potter, K. (2008). How students "unpack" the structure of a word problem: Graphic representations and problem solving. School Science and Mathematics, 108(5), 184-196.

Garfield, J., & Ahlgren, A. (1988). Difficulties in learning basic concepts in probability and statistics: Implications for research. Journal of Research in Mathematics Education, 19(1), 44-63.

Koedinger, K., & Anderson, J. (1990). Abstract planning and perceptual chunks: Elements of expertise in geometry. Cognitive Science, 14, 511-550.

Konold, C. (1989). Informal conceptions of probability. Cognition and Instruction, 6, 59-98.

Larkin, J. H., & Simon, H. A. (1987). Why a diagram is (sometimes) worth ten thousand words. Cognitive Science, 11, 65-99.

Lewis, A. B. (1989). Training students to represent word problems. Journal of Educational Psychology, 81(4), 521-531.

Logie, R., & Baddeley, A. (1990). Imagery and working memory. In P. J. Hampson, D. E. Marks, & J. T. E. Richardson (Eds) Imagery: Current Developments (pp. 103-128). New York, NY: Routledge.

Lowrie, T., & Kay, R. (2001). Relationship between visual and nonvisual solution methods and difficulty in elementary mathematics. Journal of Educational Research, 94(4), 248-255.

Mayer, R. (1992). Mathematical problem solving: Thinking as based on domain specific knowledge. In R. E. Mayer (Ed), Thinking, Problem Solving, Cognition (pp. 455-489). New York, NY: W. H. Freeman & Co.

Mayer, R., & Gallini, J. (1990). When is an illustration worth ten thousand words? Journal of Educational Psychology, 82, 715-726.

Novick, L. (1990). Representational transfer in problem solving. Psychological Science, 1, 128-132.

Novick, L. (2001). Spatial diagrams: Key instruments in the toolbox for thought. In D. L. Merlin (Ed.), The psychology of learning and motivation, 40, 279-325.

Novick, L., & Hmelo, C. (1994). Transferring symbolic representations across nonisomorphic problems. Journal of Experimental Psychology: Learning, Memory, and Cognition, 20, 1296-1321.

O'Connell, A. A. (1999). Understanding the nature of errors in probability problem solving. Educational Research and Evaluation, 5, 1-21.

Russell, W. E. (2000). The use of visual devices in probability problem solving. (Doctoral Dissertation, Columbia University, 2000). New York: Academic Press.

Schwartz, D. L., & Martin, T. (2004). Inventing to prepare for future learning: The hidden efficacy of encouraging original student production in statistics instruction. Cognition and Instruction, 22(2), 129-184.

Tomlinson, S. & Quinn, R. (1997). Understanding conditional probability. Teaching Statistics, 19(1), 2-7.

Tversky, B. (2001). Spatial schemas in depictions. In M. Gattis (Ed.), Spatial schemas and abstract thought. Cambridge, MA: MIT Press.

Uesaka, Y., Manalo, E., & Ichikawa, S. (2007) What kinds of perceptions and daily learning behaviors promote students' use of diagrams in mathematics problem solving? Learning and Instruction, 17, 322-335.

Van Essen, G. & Hamaker, C. (1990). Using self-generated drawings to solve arithmetic problems. The Journal of Educational Research, 83(6), 301-312.

Van Meter, P., & Garner, J. (2005). The promise and practice of learner-generated drawing: Literature review and synthesis. Educational Psychology Review, 17(4), 285-325.

Vygotsky, L. S. (1978). Interaction between learning and development. In M. Cole, V. John-Steiner, S. Scribner, & E. Souberman (Eds.), Mind in Society: The development of higher psychological processes. Cambridge, MA: Harvard University Press.

Zahner, D., & Corter, J. E. (2010). The process of probability problem solving: Use of external visual representations. Mathematical Thinking and Learning, 12(2), 177-204.

Interactive Effects of Diagrammatic Format and Teleological Beliefs on Tree Thinking

Brenda C. Phillips (b.c.phillips@vanderbilt.edu)
Dept. of Psychology & Human Development, Vanderbilt University
Nashville, TN 37203 USA

Laura R. Novick (Laura.Novick@vanderbilt.edu)
Dept. of Psychology & Human Development, Vanderbilt University
Nashville, TN 37203 USA

Kefyn M. Catley (kcatley@wcu.edu)
Dept. of Biology, Western Carolina University
Cullowhee, NC 28723 USA

Daniel J. Funk (daniel.j.funk@vanderbilt.edu)
Dept. of Biological Sciences, Vanderbilt University
Nashville, TN 37235 USA

Abstract

A common misconception regarding evolutionary history is that the tree of life depicts the progression of species over time from least complex to most complex, ending with our own species at the pinnacle of evolution. The current study examined the diagrammatic factors that may impact the effect this misunderstanding has on students' ability to correctly interpret evolutionary trees. Students with weaker and stronger backgrounds in biology were presented with two cladograms, each featuring a different focal taxon (human or honeybee). The evolutionary relationships among the taxa were presented in four diagrammatic formats. Students reasoned in qualitatively different ways when asked about the human species as opposed to the honeybee, with specific diagrammatic formats facilitating anthropomorphic views, particularly among weaker background students.

Keywords: spatial cognition, teleological explanations, evolutionary diagrams, evolutionary misconceptions, cladograms, macroevolution

Introduction

There is a wealth of evidence that indicates students have great difficulty acquiring evolutionary concepts, particularly concepts regarding macroevolution and the origin of species. These studies have demonstrated that misconceptions are prevalent even among students with substantial training in the biological sciences (Ferrari & Chi, 1998; Greene, 1990; Samarapungavan & Weirs, 1997). A pertinent question is whether the tools that scientists use to study macroevolution are cognitively accessible and transparent to students of varying abilities, and whether there are perceptual or diagrammatic factors that potentially impede students' understanding of these tools in the absence of explicit instruction.

Tree thinking is a tool that professional biologists use to describe and classify species according to patterns of *most recent* common ancestry and to make inferences in the absence of data (e.g., Angielczyk, 2009). Evolutionary trees, or cladograms, are based on hypotheses regarding the distribution of derived characters among a set of taxa; they provide biologists with a conceptual framework for understanding the historical processes that promote and maintain the biodiversity of our planet. Although intensive instruction on macroevolution and tree thinking is largely absent from high school and college biology classes (Catley, 2006), a recent analysis of textbooks indicates that biology students at both levels are exposed to cladograms (Catley & Novick, 2008). This poses a potential problem if students reason incorrectly about the evolutionary relationships depicted in those diagrams.

Researchers have only recently begun to examine what information students are able to extract from these diagrams, both in the absence of explicit instruction as well as after instruction. This research has focused on assessing tree-thinking skills when cladograms are drawn in the familiar hierarchical tree format or in an alternative ladder format (Catley, Novick, & Funk, accepted; Meir, Perry, Herron, & Kingsolver, 2007; Novick & Catley, 2010; Sandvik, 2008). The results indicate that students, regardless of instruction, find the tree format much easier to understand.

The current study builds upon this prior research by examining how the particular taxa depicted in the cladograms, and especially students' knowledge and/or beliefs about those taxa, affects tree thinking (i.e., cladogram interpretation). We used the simpler-to-understand tree format and manipulated how the cladograms were oriented and how the taxa were arranged (keeping the underlying structure—evolutionary relationships—constant across cladogram versions). In particular, this study explores the misconception that the tree of life depicts the progression of the evolution of taxa over time from least complex to most complex. If students reason incorrectly about the evolutionary relationships among taxa, they may

state that the most cognitively complex taxon (i.e., the human) is the most highly evolved. Feeding into this misconception is the widely held belief that humans are not subject to the same evolutionary pressures as other organisms because we were created intentionally by an outside agent (Evans, 2001; Greene, 1990) or created intentionally to fulfill a purpose (Kelemen, 1999; Kelemen & Rosset, 2008).

These considerations raise several questions: In the absence of explicit instruction, are students likely to perceive the evolution of specific taxa, such as the human, in teleological, or goal-directed, terms? Do students' responses differ depending on their biology background? Additionally, are students more likely to provide teleological responses when the focal taxon is located at the end of the cladogram versus when it is located in the center position? How does the vertical or horizontal orientation of the cladogram influence students' judgments?

Study Overview

The data presented here are part of a larger study that was designed to assess college students' reasoning about evolutionary history among several different subsets of taxa from the tree of life. The questions assessed, in several different ways, students' understanding of the evolutionary relationships among hierarchically nested sets of taxa. We limit our presentation here to one question that examined whether students' misinterpreted the information depicted in the cladograms by stating that the focal taxon was the most highly evolved. This question was asked about two cladograms, which differed in the focal taxon highlighted for subjects (human or honeybee).

The cladograms were drawn in four different ways: The cladogram itself was oriented either horizontally or vertically, and the focal taxon was situated either at one end of the cladogram (top or right) or in the center position among the set of nine taxa. Given students' teleological beliefs and misconception that humans are evolutionarily special, we predicted that students would be more likely to state that the focal taxon was the most highly evolved taxon when it was the human rather than the honeybee (see Figures 1 and 2). We also predicted that students would be more likely to make this claim when the focal taxon occupied the end (top or far right) rather than the middle position because a teleological construal would lead one to expect the most complex taxon to be at the end. Finally, because Franklin and Tversky (1990) have found that the vertical dimension is the most salient of the three spatial dimensions, we predicted that responses indicating that the human is the most highly evolved taxon would be most prevalent for the vertical orientation when human was situated at the top.

The main study included a sample of college students with weaker and stronger backgrounds in biology. In a follow-up study, a subset of the stronger background students received two days of instruction on phylogenetics (i.e., understanding evolutionary trees). They were tested before and after instruction.

Method

Subjects

The subjects in the main study were 112 Vanderbilt University undergraduates. Most students (34 females, 33 males, 2 unknown sex) were recruited from a paid subject pool in the psychology department. The remaining students (23 females, 20 males) were currently enrolled in the evolution class at Vanderbilt (taught by the fourth author).

We divided the subjects into two groups based on their background in biology: Students who had completed at least the two-semester introductory biology sequence for biology majors and pre-med students were assigned to the stronger background group; the remaining subjects were assigned to the weaker background group. The 52 stronger background students (28 females, 24 males) completed an average of 3.09 semesters of biology classes that were chosen from a list of classes presented on a background questionnaire. The 60 weaker background students (29 females, 29 males, 2 unknown sex) had completed an average of only 0.40 semesters of such coursework. This is nearly an 8:1 difference in coursework between the groups.

Materials and Procedure

All students received a 4-page booklet that included one cladogram and two to three questions about the information in that cladogram on each page. The presentation order of the cladograms was counterbalanced. Students completed this booklet, as well as several other booklets, in one session that took approximately 50-75 min.

Each cladogram included nine taxa. One taxon was the focal taxon, so named because the first question for each cladogram asked students what the diagram shows about the evolution of that taxon. (Subjects provided a written response to that question.) We limit our discussion here to the third question that was asked about the two cladograms for which human and honeybee were the focal taxa. This question asked students which taxon/taxa was/were the most highly evolved. (The second question asked students to evaluate the relative evolutionary distance between pairs of taxa. This question did not reference the focal taxon and did not bear on the present results).

Design

We examined three factors in the present study. One factor was weaker versus stronger biology background. We were interested in whether a year-long introductory class (and perhaps subsequent biology coursework) would countervail stronger background students of a teleological perspective on evolution.

The remaining two factors pertained to the visual presentation of the cladograms. The first of these factors was the orientation of the cladogram, which was

manipulated between subjects. The terminal branches were either located at the right side of the cladogram (vertical orientation, Figure 1) or at the top of the cladogram (horizontal orientation, Figure 2). The second factor was the rotation of the branches of the cladogram. The branches were rotated, without altering the depicted relationships (i.e., the underlying topology), so that the focal taxon was located either at the end (far right or top, depending on the orientation; see Figure 1) or at the center position (see Figure 2). In Rotation Set 1, human was located at the end position whereas the honeybee was in the middle. In Rotation Set 2, the human was in the middle and honeybee was at the end. Rotation set was manipulated between subjects. We fully counterbalanced the orientation and rotation of the cladograms (see Figures 1 and 2 for two of the four possible combinations).

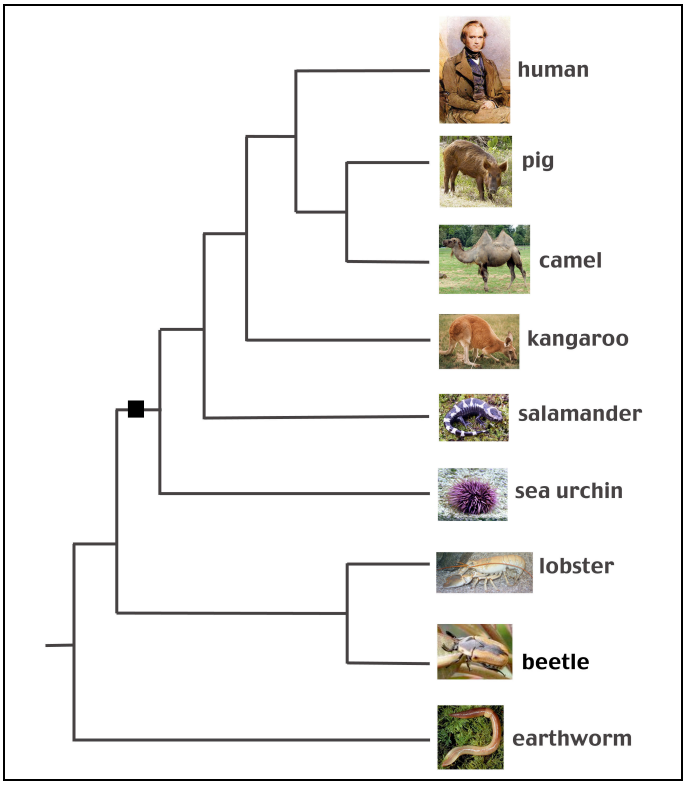


Figure 1: Human cladogram—vertical orientation, focal taxon at the end.

Follow-Up Study

We also examined whether students' evolutionary concepts were amenable to instruction by providing a subset of the students in the stronger biology background group (N = 42) with two days of instruction in phylogenetics (i.e., understanding cladograms). These students were recruited from their evolutionary biology course and completed the cladogram booklets at mid-semester (Time 1) and again 4.5-5 weeks later (Time 2). Students received the same booklet at both times.

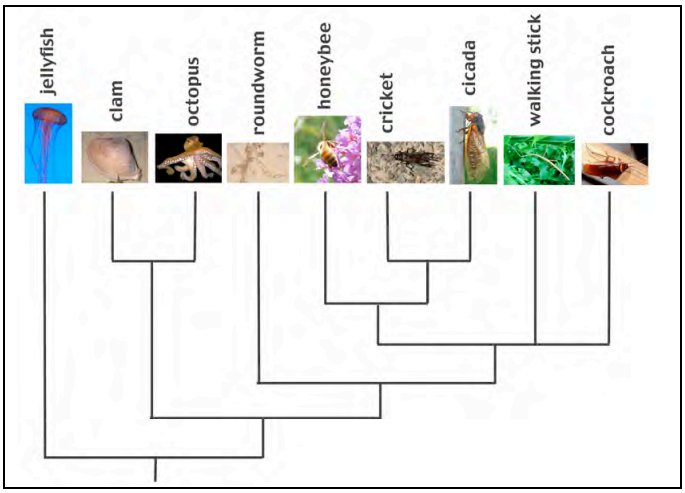


Figure 2: Honeybee cladogram—horizontal orientation, focal taxon in the middle.

Results

Are Humans Most Highly Evolved?

Students received a score of 1 if they indicated that the focal taxon (i.e., human or honeybee) was the most highly evolved species or a score of 0 for any other response. Overall, only 6% of students (all stronger background) responded correctly that the human cladogram did not reveal that any taxon was more highly evolved than any other taxon. In comparison, 5% of students (all stronger background) provided a correct response to this question regarding the honeybee.

As discussed earlier, we expected more responses that the focal taxon was most highly evolved when that taxon was the human as opposed to the honeybee. The results support this prediction, with 35% of students providing this response for the human cladogram, compared with only 2% (2 students, both from weaker backgrounds) for the honeybee cladogram, (χ^2=42.32, p <.001). Because students essentially never said that the honeybee was the most highly evolved taxon, we restricted our analysis of the effects of the diagrammatic factors on these responses to the human cladogram.

We conducted a 2 (biology background; between) x 2 (orientation; between) x 2 (rotation set = human at the end vs. in the middle; between) analysis of variance (ANOVA) on the responses that the human was the most highly evolved taxon. The main effect of biology background, F(1, 104) = 17.20, p < .001, MSE = 0.17, partial η^2 = .14, indicated that weaker background students were more likely to make this incorrect claim than were stronger background students (M = 0.50 vs. M = 0.19, respectively). The main effect of focal taxon location (rotation set), F(1, 104) = 20.17, p < .001, partial η^2 = .16, indicated that a higher proportion of students made this claim when the human was

positioned at the end of the array than at the center (M = 0.52 vs. M = 0.20, respectively).

There was also a biology background x orientation interaction, $F(1, 104) = 6.55$, $p < .05$, partial $\eta^2 = .06$ This interaction was subsumed by a three-way interaction between biology background, focal taxon location, and orientation, $F(1, 104) = 4.02$, $p < .05$, partial $\eta^2 = .04$ (see Figure 3). When the human was located in the middle of the cladogram, weaker background students said humans are most highly evolved 35% of the time, compared with 0% of the time for stronger background students. Cladogram orientation had little effect. When the human was located at the end, however, both groups of students said that the human was most highly evolved, with such responses being especially prevalent for weaker background students who received the vertical orientation. Indeed, 92% of these students said that humans were most highly evolved, compared with only 40% for the other three groups combined.

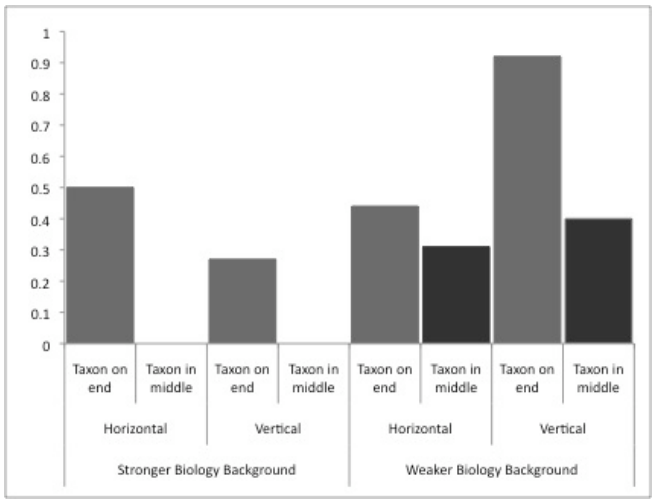


Figure 3: Proportion of students who claimed the human was the most highly evolved taxon as a function of biology background, focal taxon location, and cladogram orientation.

To examine the effects of instruction on students' phylogenetic conceptions, we conducted a 2 (orientation; between) x 2 (rotation set = human at the end vs. in the middle; between) x 2 (test: Time 1 vs. Time 2; within) mixed ANOVA on the responses that the human is the most highly evolved taxon. These students comprised a subset of the stronger background students from the main study. The analysis revealed a main effect of time of test, $F(1, 38) = 6.18$, $p < .05$, $MSE = 0.05$, partial $\eta^2 = .14$. Students were less likely to claim that the human is the most highly evolved taxon after having received two days of instruction on phylogenetics (M = 0.17 vs. M = 0.05, respectively, for before vs. after instruction). There was also a time of test x focal taxon location interaction, $F(1, 38) = 6.18$, $p < .05$, $MSE = 0.05$, partial $\eta^2 = .14$. Students only claimed that the human was the most highly evolved taxon when it was presented at the end (top or right) of the cladogram. Under these conditions, students were less likely to state that the human was the most highly evolved taxon after instruction (M = 0.10) than before (M = 0.33). Students never responded that the human was the most highly evolved taxon when it was presented in the middle position.

Students' Justifications

After indicating which taxon was most highly evolved, students were asked to provide an explanation for their response. We are in the process of devising a coding scheme to examine these qualitative data. In the following paragraphs, we provide a subset of the responses students wrote for the explanation question for illustrative purposes.

Consistent with our hypotheses, students who indicated that the human was the most evolved taxon frequently stated that a) the cladogram presented the progression of evolution across species and time, and b) presented an array of organisms, from least complex to most complex. For example, students provided statements such as, "The general assumption is that with every further deviation from the evolutionary chain, organisms develop more complete biological systems (esp. nervous systems)", "We have complex language & highly developed social systems", or "we are the only sentient beings on earth." Students also made comparative statements such as "I'm arogant [*sic*] enough to believe [that] I'm more evolved than livestock" or "they are the last animal in the chart. I am, as a person, more evolved than a pig."

Students also provided evidence that they were reasoning about phylogenetic concepts, albeit incorrectly: "humans have diverged from the most basic common ancestor the most times out of all the animals shown", "humans are at the top of the diagram and they display the most specified method of evolution in the diagram", or "humans are the organism which most recently evolved." Interestingly, sometimes students provided conflictive statements such as, "from the chart I would say pig & camel, but I'm biased to say human" or "humans have to be the most highly evolved (regardless of the structure of the diagram)." Additional analyses will examine whether stronger and weaker background students provided different types of justifications for incorrect responses.

The aforementioned statements were qualitatively different from those that students provided for correct responses. For example, students who indicated that no taxon was more highly evolved than any other taxa made statements such as, "the diagram only shows the evolutionary relationships not how much each species has changed over time" or "these trees just show genetic similarity and hypothetical common ancestors. All the organisms have radiated into different niches, from the labeled hypothetical common ancestor." As stated previously, only a very small minority of students with stronger backgrounds in biology provided correct responses.

Discussion

The current study provides critical information regarding the use of cladograms for educational purposes. In the absence of instruction, both students with weaker and stronger backgrounds in biology misinterpreted the information depicted in cladograms when asked to evaluate which taxon was the most highly evolved. An important finding is that the cladograms had different effects on students' reasoning depending on the format in which they were presented and the biology background of the students.

As expected, students provided more teleological responses and explanations for the human cladogram than the honeybee cladogram. In fact, students essentially never stated that the honeybee was the most highly evolved taxon despite the fact that the two taxa occupied identical locations in their respective cladograms. Students provided justifications that indicated that they perceived the human as the most complex organism in the array, and therefore the most highly evolved.

Previous research has found that college students endorse scientifically unwarranted explanations for the occurrence of natural phenomena (e.g., "Finches diversified in order to survive"), especially when placed under a high cognitive load (Kelemen & Rosset, 2009). These studies indicate that adults, like children (see Carey, 1985; Keil, 1994; Kelemen, 1999), ascribe to teleological explanations for the existence of biological natural kinds and prefer these explanations to physical-causal explanations. These beliefs are suppressed under certain conditions, such as when students are provided with alternative explanations and are provided with ample time to think about the phenomena in question. However, under cognitively demanding circumstances, these unwarranted scientifically beliefs prevail.

Our results are consistent with these earlier studies and provide new information concerning the perceptual or diagrammatic factors that either promote or lessen students' appeal to teleological interpretations of evolutionary diagrams. We reasoned that students who conceived of evolutionary processes as goal-directed would expect the most complex taxon to occupy an end position. As predicted, students were more likely to state that the human was the most evolved taxon when it occupied the end position rather than the center position. Students with stronger backgrounds in biology only said that the human is the most highly evolved taxon when it was depicted at the end of the set of taxa. Instruction in phylogenetics reduced such responding to only 10% of students.

Teleological responses were most prevalent for weaker background students when interpreting the vertically oriented cladogram with the human located in the top position. One possible interpretation of these results is that students used spatial location to evaluate evolutionary relatedness; that is, they inferred the taxon at the highest vertical point was the most complex. These results are consistent with the embodied cognition perspective that states that individuals orient themselves vertically in reference to elements of the environment, such as the sky and ground (Franklin and Tversky, 1990).

Given that high school and college students in the United States are currently exposed to cladograms in their biology textbooks, and perhaps from their instructors in class as well, our results indicate that it is essential that textbook illustrators and instructors consider the perceptual or diagrammatic factors that impact students' understanding of evolutionary processes. In particular, our results indicate that the horizontal cladogram format is preferable to the vertical format. Moreover, because cladogram branches can be rotated without changing the underlying structure (i.e., the evolutionary relationships depicted; just as the turning branches of a mobile in the wind do not change the structure of the mobile), when cladograms include taxa that may play into students' teleological misconception of evolution, it is critically important to present those taxa in a horizontal order that suppresses activation of this misconception. For example, more complex taxa should be located in the middle rather than the end, and there should be little or no correlation between the linear ordering of the taxa across the terminal branches of the cladogram and students' conceptions of complexity.

References

Angielczyk, K. D. (2009). *Dimetrodon* is not a dinosaur: Using tree thinking to understand the ancient relatives of mammals and their evolution. *Evolution: Education and Outreach, 2,* 257-271.

Carey, S. (1985). *Conceptual change in childhood*. Cambridge, MA: MIT Press.

Catley, K. M. (2006). Darwin's missing link: A new paradigm for evolution education. *Science Education, 90,* 767-783.

Catley, K. M., & Novick, L. (2008). Seeing the wood for the trees: An analyseis of evolutionary diagrams in biology textbooks. *BioScience, 58*, 976-987.

Catley, K. M., Novick, L. R., & Funk, D. J. (accepted pending revision). The Promise and Challenges of Introducing Tree Thinking into Evolution Education. In K. Rosengren, E. M. Evans, S. Brem, & G. Sinatra (Eds.), *Evolution Challenges: Integrating Research and Practice in Teaching and Learning about Evolution.*

Evans, M. (2001). Cognitive and contextual factors in the emergence of diverse belief systems: Creation versus evolution. *Cognitive Psychology, 42*, 217-266.

Ferrari, M., & Chi, M. (1998). The nature of naïve explanations of natural selection. *International Journal of Science Education, 20*, 1231-1256.

Franklin, N., & Tversky, B. (1990). Searching imagined environments. *Journal of Experimental Psychology: General, 119,* 63-76.

Greene, E.D. (1990). The logic of university students' misunderstanding of natural selection. *Journal of Research in Science Teaching, 27*, 865-885.

Keil, F.C. (1994). The birth and nurturance of concepts by domains: The origins of concepts of living things. In L.A.

Hirschfeld & S.A. Gelman (Eds), *Mapping the Mind: Domain Specificity in Cognition and Culture*. Cambridge: Cambridge University Press.

Kelemen, D. (1999). Function, goals, and intention: Children's teleological reasoning about objects. *Trends in Cognitive Science, 3(12)*, 416-468.

Kelemen, D., & Rosset, E. (2009). The human function compunction: Teleological explanation in adults. *Cognition, 111*, 138-143.

Meir, E., Perry, J., Herron, J. C., & Kingsolver, J. (2007, September). College students' misconceptions about evolutionary trees. *The American Biology Teacher Online, 69 (7)*.

Novick, L. R., & Catley, K. M. (2010). *Understanding the Tree of Life: Exploring Tree-Thinking Skills in College Students.* Manuscript under revision for an invited resubmission.

Samarapungavan, A., & Weirs, R.W. (1997). Children's thoughts on the origins of species: A study of explanatory coherence. *Cognitive Science, 21*, 147-177.

Sandvik, H. (2008). Tree thinking cannot taken for granted: challenges for teaching phylogenetics. *Theory in Biosciences, 127,* 45–51. Retrieved from: http://www.springerlink.com/content/eu62420p381402xr/ March 19, 2008.

Thinking with Networks

Jeffrey V. Nickerson (jnickerson@stevens.edu)
Stevens Institute of Technology, Castle Point on Hudson,
Hoboken, NJ 07030

Barbara Tversky
Teachers College, Columbia University,
525 West 120th Street, New York, NY 10027

James E. Corter
Teachers College, Columbia University,
525 West 120th Street, New York, NY 10027

Lixiu Yu
Stevens Institute of Technology, Castle Point on Hudson,
Hoboken, NJ 07030

Yun Jin Rho
Teachers College, Columbia University,
525 West 120th Street, New York, NY 10027

David Mason
Teachers College, Columbia University,
525 West 120th Street, New York, NY 10027

Abstract

Many planning tasks involve complex reasoning about time: what must happen in sequence and what may happen in parallel. One hundred ten online participants were provided with a simple planning scenario (to design a calling tree) and asked to manipulate different diagrammatic representations of the problem. More important than the initial representation was the participants' transformed representations: if time was encoded in the lengths of tree links then inference was more accurate. This finding suggests that diagram transformation may be a useful way to elicit representation strategies, and that such transformations from different starting conditions may be useful as diagnostics and as design aids.

Keywords: diagram understanding, design, topological diagrams, representation of time, distributed computing

Introduction

To solve problems or to simply organize information, people often make diagrams. Diagrams can aid problem solving and information organization by spatializing the essential concepts and relations among them. One of the most abstract kinds of diagrams is a network, where nodes are concepts and links are relations. Because of their generality, network diagrams appear in many diverse domains.

The advantage of networks, their ability to represent so many different relations, is also a disadvantage, because they may not make problem constraints apparent. Often, problems and information have more constraints than the simple binary relations used in networks, constraints that would allow inferences, for example, asymmetric relations. Certain variants of networks can represent such constraints. Trees, for example, are commonly used to represent asymmetric or hierarchical relations, notably for structural relations such as organization charts or phylogenetic relations. They are also used to represent temporal relations, as in decision trees or flow diagrams. For structure, the links indicate an asymmetric structural relation, such as control in corporations or *kind of* in phylogenies. For time, the links indicate asymmetric temporal relations, at an ordinal level: this, then this, then this.

However, there are situations where representing both structural and temporal relations is desired, for example, in coordination situations where a set of agents carry out a temporally constrained set of actions. Representing both structural and temporal organizations simultaneously presents a challenge. The structure – who contacts whom – needs to be represented. Also the timing – the temporal ordering of contacts – must be shown. Links can be used for the structural information, but some other aspect of the diagram needs to be used for the temporal. Representing both structure and time simultaneously can be all the more challenging when *metric* properties of time are important because links in networks are typically used to indicate a relationship, but not the degree of a relationship.

These problems of representation are, more broadly, problems of cognition: as the data will show, reasoning

about coordination is difficult. Given the high cost of coordination failure in a number of different fields, and the everyday importance of coordination in computational fields, the problem deserves attention. Previously, studies have been performed on the way network diagrams convey information. For example, it has been shown that distance along network links is used to evaluate content similarity (Fabrikant, Montello, Ruocco, & Middleton, 2004). That study focused on the way distance and topology map to similarity; here we focus on the way distance and topology map to structure and time.

We turned to users to see how they would represent space and time simultaneously. Often, users turn out to be good designers, inventing clever devices to represent abstract information (e. g., Kessell and Tversky, 2008; Tversky, in press). Furthermore, their visualizations of thought are a window to thought (e. g., Tversky and Lee, 1999). The other side of successful design is comprehension. We have begun exploring how people design and comprehend diagrams and solutions for a class of problems that requires representing both structure and time (see Figure 1).

The paradigm we have been using is based on a distribution tree. Because the problem is a general one of transmission of something from one party to many, it applies to many situations when information or goods are distributed. For example, a telephone tree can be used to distribute information about a school closing due to weather conditions. For speed of transmission, it is better to distribute the callers; for reliability, it is better to minimize the number of callers. Solutions, then, depend both on structure and on time. Although some forms of trees, such as decision trees and flow diagrams, are used to represent time, they only represent temporal *order*. In contrast, optimizing a distribution tree depends on *metric* properties of time as well. Thus, diagramming a distribution tree solution not only requires representing both structure and time, it also requires representing time metrically. An added difficulty for designers and for users in producing or interpreting designs for distribution trees is that several calls can happen at the same time. That is, both sequence and parallelism need to represented and understood.

In extensive pilot work, we have found that people spontaneously create trees to solve these problems, but that their trees usually represent structure, that is, who contacts whom, and rarely represent time. In fact, representing or grasping structure from diagrams is easier and more straight-forward than representing or grasping changes in structure, such as changes in time (e. g., Suwa & Tversky, 1997; Tversky, Heiser, Lozano, Mackenzie, & Morrison, 2008). More generally, space seems to serve as a metaphor for time more readily than time for space (Boroditsky, 2000).

For the telephone tree problem, structure is ordinal, but time is metric. There are several ways to superimpose time onto a network representing structure. Telephone trees are tricky because a single agent can make only one call at a time, but several agents can call simultaneously. One way to represent time, illustrated in Figure 2, is to use length of link, as in additive similarity trees (Sattath & A. Tversky, 1977; Corter, 1996). In this representation, the lengths of the links emanating from any one agent indicate the sequence of that caller's calls. For large trees, this can be visually confusing. Another method to represent time is a combination of using levels of a tree to distinguish when a caller is first notified, and within levels, showing the sequence of calls made by a caller using a left-to-right first-to-last convention; this is visually more organized but requires keeping track of two spatial mappings to assess time. Both methods have been invented by our participants. Here, we investigate solution success when time is or is not represented by length of line.

As noted, users can be effective designers of visualizations of problems. Does the very process of designing visualizations facilitate using them? Architects and other designers sketch designs, study their sketches, get new insights, and revise them, a positive, productive cycle that has been likened to a conversation (Schon, 1983). Creating and revising visualizations of a range of complex concepts, for example, scientific ones, has been shown to increase depth of understanding (e. g., diSessa, Hammer, Sherin, & Kolpakowski, 1991; Schwartz, 1995).

The present experiment examines the dual roles of kind of design and act of designing for solving telephone tree problems. Participants were given a problem analogous to the guard problem and then asked to provide an optimal diagram and to compute the amount of time it should take to notify everyone. They were given an initial diagram, one they could alter, to create a diagram they regarded as optimal. For some participants, the initial diagram represented structure but not time. For others, the initial diagram represented time using proportional length of line. For a third group, the initial diagram varied in line length but not proportionally to time; thus this diagram provides a hint that line length might be helpful, but not how.

This design allows asking a set of questions. We can ask whether representing time explicitly in a visualization makes for a more effective diagram that better helps users to solve a telephone tree problem. We can ask whether time is more likely to be explicitly represented in user diagrams when the starting diagram provided to them uses variable line lengths, either compatible with time or incompatible with time.

Method

One hundred ten participants accepted and completed an assigned task in return for payment on Amazon's crowdsourcing marketplace. The participants in Amazon's pool have been characterized extensively in several previous studies: the pool is 55% female with a mean age of 31 (Kittur, Chi, & Su, 2008; Ross, Irani, Silberman, Zaldivar, & Tomlinson, 2010). Participants were presented with the following textual description:

Please read the following question and then make changes to the diagram.

Hart has the job of notifying 4 other parents in the event that school is called off due to weather conditions. Hart has created a plan for a sequence of phone calls:

Hart calls Dean and then Lane. Dean calls Boyd and then Ward.

Assuming that each phone call lasts one minute, please go to the website below to make changes to the diagram to meet the plan description.

Once they had saved the diagram, they were asked:

Assuming that each call lasts one minute, how many minutes will elapse before all parents know about the school cancellation?

Participants were randomly assigned one of the following three diagrams shown in Figures 1, 2 and 3. Figure 1 is a typical tree structure. The connections indicate who calls whom, using uniform line lengths. In a pilot study, most participants drew such a diagram.

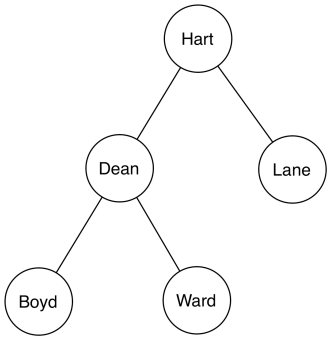


Figure 1: A uniform tree with no time encoding

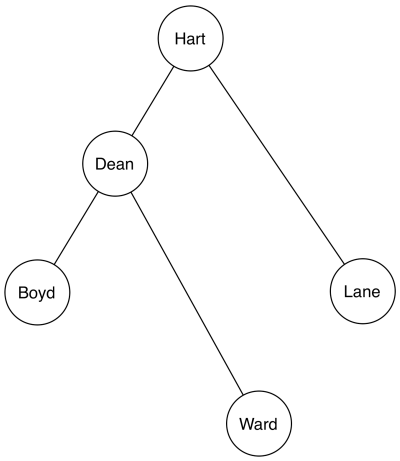


Figure 2: A time-encoded tree with edge lengths consistent with the problem description

In Figure 2, time has been encoded into the lengths of the connections between nodes. That is, after one minute, Hart has managed to talk to Dean. After two minutes, Hart has also managed to talk to Lane, and Dean has talked to Boyd.

The lengths of the connections reflect the constraint of the problem, that a person can only have one phone conversation, and so time has elapsed after each conversation. We found in a paper and pencil pilot study that some participants invented or at least used this representation. It is similar to diagrams used in transportation systems called space-time networks, in which nodes are lined up according to elapsed time (e.g., Pallottino & Scutella, 1998).

In Figure 3, variable edge lengths are used, but the vertical position of a node is not consistent with elapsed time in the problem. For example, in the problem statement, Dean calls Boyd before calling Ward, and the vertical arrangement of Figure 3 implies the opposite order. Thus, the diagram may cue individuals to the possibility of using connection length to represent time, but is not useful in representing the problem (and may even be misleading) unless it is transformed.

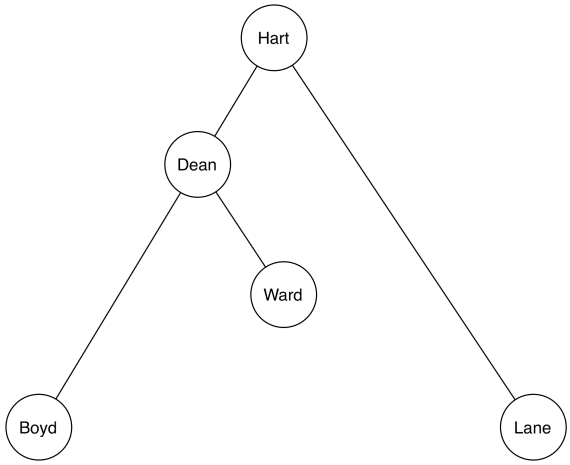


Figure 3: A time-encoded tree with edges inconsistent with the problem description

After being randomly assigned one of the three diagrams above, participants were provided instructions on how to use a customized web-delivered vector-based drawing tool to move nodes, thereby manipulating spacing in the diagram. In the tool, the connections between the nodes are preserved as the nodes are moved. The participants' mouse movements were recorded. Thus, the experiment allows us to study the effect of the initial diagram provided, the cuing representation. In addition, the participants' transformed diagrams can be classified, and the relationship between these produced diagrams and the accuracy of problem solving shown.

Results

A total of 32 participants were cued with the uniform tree of Figure 1, 38 with the consistent time tree of Figure 2, and 40 with the inconsistent time tree of Figure 3. The overall accuracy of their answer to the time question as a function of cuing diagram is shown in Figure 4 as a proportion. The consistent time tree was associated with the highest accuracy (.66), and the inconsistent time tree with lowest accuracy (.48); the uniform tree produced an intermediate level of accuracy (.53). In a logistic regression model comparing accuracy for these three conditions, cuing with the time tree yielded marginally higher accuracy than the uniform tree (Wald = 2.618, .05<p<.10, one-tailed) and cuing with an inconsistent time tree produced marginally lower accuracy than the other two trees (Wald = 1.983, .05<p<.10, one-tailed)

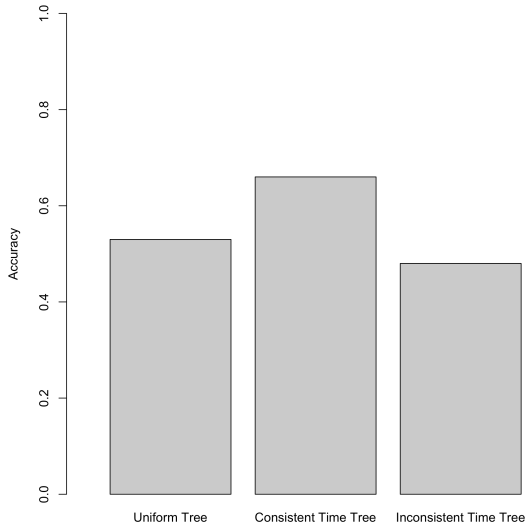


Figure 4: Accuracy of the answer depending upon the randomly provided starting diagram

The final tree diagrams produced by the participants were then classified into three sets: Uniform Trees, Time Trees, and Wrong Trees. Wrong trees could only result if participants used the drawing tool to change the topology of the graphs by adding or subtracting nodes. For example, one participant directly linked Dean to Lane, as in Figure 5.

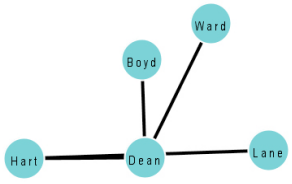


Figure 5: A topologically incorrect tree (Wrong Tree)

The rest of the participants created two kinds of topologically valid trees. Those producing Uniform Trees showed no attempt to encode time through distance, while those producing Time Trees did. Classifying the trees was straightforward, because the uniform trees tended to have uniform distances, and in particular equal distances between parents and direct descendants. We checked inter-rater reliability of the coding of the produced graphs by training two raters on 15 graphs and then testing on 43 graphs: Cronbach's alpha = .99. Figure 6 shows examples of the produced trees.

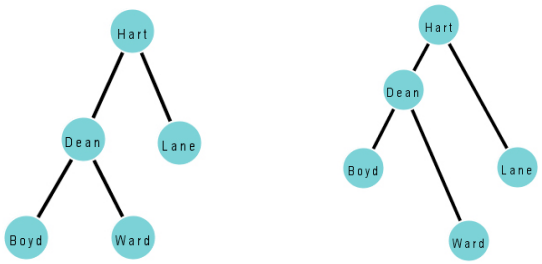


Figure 6: On the left, a produced uniform tree transformed from a given consistent time tree, and on the right, a consistent time tree transformed from a uniform tree.

Figure 7 shows accuracy, as a proportion of participants' time estimates by type of produced diagrams. For example, the left-most blue and tan bars show that 73% of those who produced time trees when provided with uniform trees calculated the correct answer, whereas 50% of those who produced a uniform tree in that same condition calculated the correct answer.

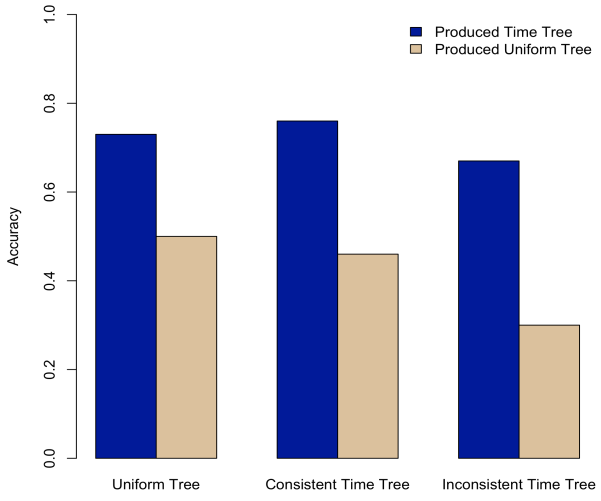


Figure 7: Accuracy of the answer depending upon the diagram produced by the participant, grouped by the starting diagram.

The number of participants in each category can be found in Table 1, which lists the accuracy for each category, the number of participants in each category, and the totals. The

accuracy of those who produced Time Trees was significantly greater than those who produced Uniform Trees, $\chi^2(1) = 7.58$, $p < .01$.

Table 1: Mean accuracy (and frequency) for combinations of given and produced trees

		Starting diagram			
		Uniform Tree	Consistent Time Tree	Inconsistent Time Tree	*Overall (Total)*
Produced diagram	Uniform Tree	.50 (18)	.46 (13)	.29 (17)	.42 (48)
	Consistent Time Tree	.73 (11)	.76 (21)	.85 (13)	.78 (45)
	Inconsistent Time Tree	-- (0)	-- (0)	.00 (4)	.00 (4)
	Wrong Tree	.00 (3)	.75 (4)	.50 (6)	.46 (13)
	Overall (Total)	*.53 (32)*	*.66 (38)*	*.48 (40)*	

Even when presented with a consistent time tree, six participants altered the tree into a uniform tree. That is, some participants went out of their way to reconfigure the most effective diagram type to a simpler type. On the other hand, eleven participants changed the inconsistent time tree to a consistent time tree, and these participants achieved the highest accuracy shown in the table: 85% got the problem right.

Starting from the inconsistent time tree condition, four subjects produced inconsistent time trees. An example is shown in Figure 8: The tree is inconsistent because Boyd is called before Ward, yet Boyd is placed farther away from Dean than is Ward. These inconsistent trees occurred in no other condition. More broadly, from an examination of the drawing logs, we found that participants will sometimes just modify a diagram slightly, as opposed to drastically or not at all.

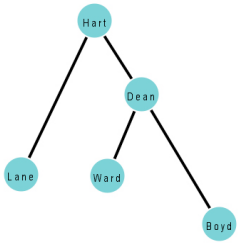


Figure 8: An inconsistent time tree, transformed from an inconsistent time tree.

Discussion

One hundred and ten participants were asked to diagram and solve a telephone tree problem, that is, determine the structure of a call tree that would notify everyone the fastest, and then to use the call tree to compute the total time to call everyone. The implicit challenge was to design a diagram that simultaneously represented both the structure of the telephone tree and the time to accomplish the plan. To assess the effects of cuing, participants were given one of three starting diagrams representing the structure of the plan: one that did not represent time; one that represented time with line lengths proportional to time; one that represented time with line lengths inversely proportional to time. There were two critical questions. Would diagrams that represent time lead to better solutions? Would cuing with a time diagram improve designs and solutions?

Those participants who created a diagram that represented time with line length were far more successful at computing the total time to call all agents than those who produced diagrams that did not represent time. Although diagrams not representing time and even some that represented time in a confusing way could be used to compute the correct solution, explicitly representing time led to large increases in correct solutions. Thus, using a diagram that directly represents all the information needed to compute the answer facilitates computation and performance.

Cuing participants with starting diagrams that did or did not represent time in a compatible way, that is, using line length proportional to time, had effects, if small, on successful solution, mediated by the final diagram participants produced.

The results show that reasoning about parallel and sequential events is difficult. Presented with a simple example involving a small number of nodes, participants in the study failed to infer the total time of a process about half the time. Presenting a diagrammatic representation that encodes time helped some, as did manipulating a diagram into a representation that encoded time.

There are implications for diagram design as well as diagram use. First, people do not always design diagrams that capture all the essential components of a situation or a problem. The present project elucidates one reason for the failure: some features of situations or problems are more readily spatialized than others. Importantly, space and structure, static relations, are more likely to be represented in diagrams and more likely to interpreted correctly than more abstract features such as time. Representing structure and time simultaneously is especially difficult, all the more so because independently, each would select the same diagrammatic feature, lines linking nodes to nodes. Finding a second diagrammatic feature to represent the second variable, in this case time, is a challenge if only because time is unidimensional, best represented as a single line. Producing the right diagram, just like producing the right mental representation, facilitates problem solving.

Acknowledgments

Portions of this research were supported by grants from National Science Foundation IIS-0725223, IIS-0855995, and REC-0440103, the Stanford Regional Visualization and Analysis Center, and Office of Naval Research NOOO14-PP-1-O649, N000140110717, and N000140210534.

References

Boroditsky, L. (2000). Metaphoric structuring: understanding time through spatial metaphors. *Cognition, 75*, 1-28.

Corter, J. E. (1996). *Tree Models of Similarity and Association*. (Sage University Papers, series: Quantitative Applications in the Social Sciences, series no. 07-112). Thousand Oaks CA: Sage.

diSessa, A. A., Hammer, D., Sherin, B., & Kolpakowski, T. (1991). Inventing graphing: Meta-representational expertise in children. *Journal of Mathematical Behavior*, 10, 117-160.

Fabrikant, S. I., Montello, D. R., Ruocco, M., & Middleton, R. S. (2004). The distance-similarity metaphor in network-display spatializations. Cartography and Geographic Information Science, 31, 237-252

Kessell, A. M. & Tversky, B. (2008). Cognitive methods for visualizing space, time, and agents. In G. Stapleton, J. Howse, and J. Lee (Editors), *Theory and application of diagrams*. Dordrecht, NL: Springer.

Kittur, A., Chi, E.H., & Suh, B. (2008). Crowdsourcing user studies with Mechanical Turk. *Proc. CHI 2008*, ACM 453-456.

Pallottino, S. & Scutella, M. F. (1998). Shortest Path Algorithms in Transportation Models: Classical and Innovative Aspects. In Marcotte, P. and Nguyen, S. (Eds), Equilibrium and Advanced Transportation Modelling, Boston: Kluwer.

Ross, J., Irani, L., Silberman, M.S., Zaldivar, A., & Tomlinson, B. (2010). Who are the Crowdworkers? Shifting Demographics in Mechanical Turk. In: *alt.CHI session of CHI 2010 Extended Abstracts on Human Factors in Computing Systems*.

Sattath, S., & Tversky, A. (1977). Additive similarity trees. *Psychometrika, 42*, 319-345.

Schon, D. A. (1983). *The reflective practitioner.* New York, NY: Basic Books.

Schwartz, D. L. (1995). The emergence of abstract representations in dyad problem solving. *The Journal of the Learning Sciences, 4*, 321–354.

Suwa, M. & Tversky, B. (1997). What do architects and students perceive in their design sketches? A protocol analysis. *Design Studies,* 18(4), 385-403.

Tversky, B. (In press). Visualizing thought. *TopiCS in Cognitive Science*.

Tversky, B., Heiser, J., Lozano, S., MacKenzie, R., & Morrison, J. (2007). Enriching animations. In R. Lowe and W. Schnotz (Editors). *Learning with animation*. Cambridge: Cambridge University Press.

Tversky, B., & Lee, P. U. (1999). Pictorial and verbal tools for conveying routes. In Freksa, C., & Mark, D. M. (Eds.). Spatial information theory: cognitive and computational foundations of geographic information science. Berlin: Springer.

Constructing internal diagrammatic proofs from external logic diagrams

Yuri Sato, Koji Mineshima, and Ryo Takemura
Department of Philosophy, Keio University
{sato, minesima, takemura}@abelard.flet.keio.ac.jp

Abstract

Internal syntactic operations on diagrams play a key role in accounting for efficacy of diagram use in reasoning. However, it is often held that in the case of complex deductive reasoning, diagrams can serve merely as an auxiliary source of information in interpreting sentences or constructing models. Based on experiments comparing subjects' performances in syllogism solving where logic diagrams of several different forms are used, we argue that internal manipulations of diagrams, or what we call internal constructions of diagrammatic proofs, actually exist, and that such constructions are naturally triggered even for users without explicit prior knowledge of their inference rules or strategies.
Keywords: External representation; Diagrammatic reasoning; Logic diagram; Deductive reasoning

Introduction

People have tried to enhance reasoning abilities by the use of artificial devices since ancient times. In particular, symbol manipulation is a distinctive tool-use of human beings. Certainly, symbolic logic may be considered to be a tool for deductive reasoning. However, it should be noted that symbolic logic (e.g., first-order logic) is not always a usable system for untrained people. By contrast, visual-spatial representations are considered to be much more intuitive and effective for novices' actual reasoning. Consequently, over the past few decades, many researchers have shown an interest in the efficacy of diagrammatic reasoning (e.g. Allwein & Barwise, 1996; Glasgow, Narayanan, & Chandrasekaran, 1995).

An important assumption in the study of diagrammatic reasoning is that diagrams are syntactic objects to be manipulated in certain ways; we make an inference about a diagram itself, transforming it into another form or combining it with other diagrams. Such syntactic manipulations of diagrams play a crucial role in accounting for their efficacy in deductive problem solving. For example, consider the following process of checking the validity of a syllogism using Euler diagrams.

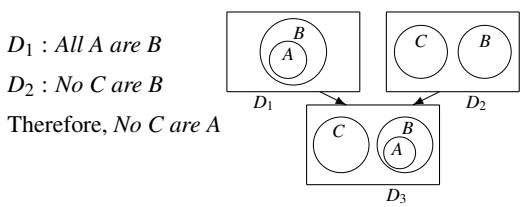


Figure 1: A diagrammatic proof of syllogism *All A are B, No C are B; therefore No C are A* with Euler diagrams.

The premise *All A are B* is represented by D_1, and the premise *No C are B* by D_2. By unifying D_1 with D_2, we can obtain diagram D_3. Here the exclusion relation holds between circles *A* and *C*, from which we can extract the correct conclusion "No *C* are *A*". In what follows, we call such a syntactic manipulation of diagrams to derive a conclusion of deductive reasoning a *construction of a diagrammatic proof*. The point here is that by unifying two diagrams in premises and observing the topological relationship between the circles, one can *automatically read off* the correct conclusion. Shimojima (1996) calls this a "free ride" property, and shows that it can be seen to exist in other kinds of diagram use in reasoning and problem solving.

In general, a deductive reasoning task would be easy if it could be replaced with a task of constructing a concrete diagrammatic proof. Typically, such a construction is supposed to be triggered by external diagrams and carried out internally, without actual drawing or movement of physical objects. However, the existence of such internal manipulations of diagrams has been the subject of controversy (see, e.g. Schwartz, 1995). Indeed, it is widely held that diagrams can serve merely as a memory-aid or an auxiliary source of information in deductive problem solving. Thus, Larkin and Simon (1987) argue that reasoning is largely independent of ways of representing information, hence diagrams are less beneficial in reasoning. Bauer and Johnson-Laird (1993) discuss efficacy of diagrams in deductive reasoning with double disjunction and argue that diagrams are used to keep track of alternative models, as postulated in mental model theory. Also in many logic textbooks, diagrams are used to depict models and aid understanding of logical representations, rather than as objects of syntactic manipulations.

In view of this situation, it is of central importance to investigate whether internal manipulations of diagrams really exist in actual reasoning with diagrams. Trafton and Trickett (2001) argue that there are mental processes of "spatial transformations" to extract information from graphs or visualization, based on an analysis of how expert scientists collect data in their researches. Shimojima and Fukaya (2003) and Shimojima and Katagiri (2008) argue for the existence of "inference by hypothetical drawing", internal transformations of external diagrams, based on eye-tracking data of subjects working with position diagrams in transitive inferential tasks. In this paper, we focus on more complex deductive reasoning tasks, namely, syllogistic reasoning tasks, and on the effects of logic diagrams externally given therein. We present evidence for the existence of internal constructions of diagrammatic proofs, on the basis of experiments comparing subjects' performances in syllogism solving where logic diagrams of several different forms are given. Our claim is consistent with the influential view in the study of external representations in general, namely, that (a) external representations can be used without being interpreted, and that (b) they can change

the nature of tasks, namely, tasks with and without external representations are completely different from users' point of view (see Zhang & Norman, 1994; Scaife & Rogers, 1996).

The efficacy of logic diagrams has been investigated in the context of the studies of logic teaching method (Stenning, 1999; 2002; Dobson, 1999). In these studies, subjects are provided with substantial training in ways of manipulating diagrams. In contrast to this, our interest is in the question whether diagrams can be useful for those who are not trained in rules or strategies of diagrammatic deductive reasoning. This question is important because, in contrast to logical formulas in symbolic logic, logic diagrams in general have been expected to be much more intuitive and effective for novices' reasoning, not for experts' nor for machine reasoning. In view of the complexities of solving processes of deductive reasoning (e.g. Levesque, 1988), it is interesting to ask whether logic diagrams can have this surprising property.

Logic diagrams have also been studied in the field of formal diagrammatic logic since the 1990s (e.g. Shin, 1994; Howse, Stapleton & Taylor, 2005), and inference systems for various diagrams such as Euler and Venn diagrams have been developed. Currently, however, there are few empirical researches to investigate their cognitive foundations. Our study is also intended to provide a bridge between logical and cognitive studies of diagrammatic reasoning.

Cognitive model for reasoning with diagrams

Typical examples of deductive reasoning problems with external logic diagrams are shown in Figure 2.

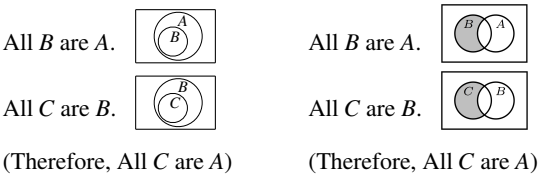


Figure 2: Examples of syllogistic reasoning tasks with diagrams

Here a syllogism is presented with logic diagrams (Euler and Venn diagrams). How can such diagrams contribute to checking the validity of a deductive argument? Let us first hypothesize a cognitive model of deductive problem solving with diagrams. The model is shown in Figure 3. This model highlights two roles of diagrams in deductive reasoning.

Regarding sentential reasoning, we assume a standard two-staged framework in natural language semantics (see, e.g. Blackburn & Bos, 2005), according to which sentences are first associated with semantic information, and then the validity of the argument is checked using some inferential mechanisms (such as model-theoretical or proof-theoretical ones). The details and precise nature of such linguistic comprehension and inference are not our concern here.

Diagrams are also associated with semantic information, but at the same time they are syntactic objects to be manipulated in reasoning processes. We distinguish two ways in which diagrams can be effective in deductive reasoning.

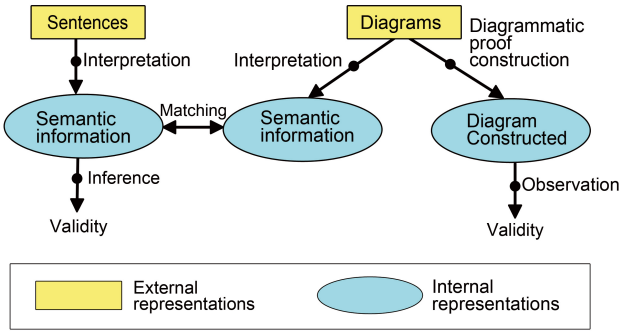


Figure 3: Cognitive model for diagrammatic reasoning

Interpretational efficacy Firstly, diagrams can help fix the correct interpretations of sentences and thereby avoid deductive reasoning errors due to misinterpretation. For example, a sentence "All *A* are *B*" is sometimes misinterpreted as equivalent to "All *B* are *A*". This is known as *illicit conversion error* in the literature (e.g. Newstead & Griggs, 1983). Subjects presented with diagrams such as the ones in Figure 2 could immediately see that the diagrams corresponding to these two sentences are topologically different, and hence deliver different semantic information. In our model, such processes are formulated as processes of matching the semantic information obtained from diagrams with the one obtained from sentences. In this case, the validity of an argument is checked based on the same kind of process as the one in linguistic reasoning. Here diagrams are used in a static way, merely as a record of information (Barwise & Etchemendy, 1991).

Inferential efficacy Secondly, and more importantly, diagrams can play a crucial role in reasoning processes themselves. More specifically, the solving processes of deductive reasoning tasks can be replaced with internal manipulations of diagrams. In other words, one can check the validity of a deductive argument by means of constructions of diagrammatic proofs. The above model assumes that such constructions are conducted through a proof-theoretical component of diagrammatic reasoning. If a task of constructing diagrammatic proofs consists of simple and intuitive steps, it is expected to be more tractable than usual linguistic inferences.

It seems to be generally agreed that logic diagrams have interpretational efficacy. For example, Stenning (2002) argues that (a)symmetricity of diagrams can aid processing of the meaning of quantified sentences in syllogisms. Mineshima, Okada, Sato, and Takemura (2008) presented experimental evidence for such interpretational effects, based on a comparison of the performances of syllogistic solving tasks with and without Euler diagrams. In what follows, we assume that logic diagrams can have interpretational efficacy, and investigate whether they can have inferential efficacy as well.

General Hypothesis

Based on the above model, we propose the following general hypothesis: (1) logic diagrams can have inferential efficacy, that is, internal constructions of diagrammatic proofs occurs

in deductive reasoning with external logic diagrams, and (2) certain diagrams would naturally trigger such constructions so that even users without explicit prior knowledge of inference rules or strategies could correctly manipulate diagrams.

One way to test our hypothesis is to compare performances of deductive problem solving with several distinct diagrams which are equivalent in semantic information but are of different forms, namely, ones that have a form suitable for diagrammatic proof constructions and ones that do not. A basic assumption here is that the existence of internal constructions depends on the forms of diagrams given, and on the simplicity or naturalness of the required diagrammatic proofs. If subjects' performance with diagrams of a form suitable for diagrammatic proof constructions would be significantly better, it could count as evidence for the existence of such constructions in subjects' reasoning.

To test the claim in (2), subjects in our experiments were presented with instructions on the meaning of categorical sentences and diagrams used, but not with any instruction on rules or strategies of constructing diagrammatic proofs. We expect that if certain diagrams have inferential efficacy, it would be exploitable based on their natural properties or constraints, rather than extra conventions. In other words, processes of constructing diagrammatic proofs as postulated in our cognitive model could be conducted without explicit knowledge of the underlying rule or strategies; such internal constructions could be naturally triggered based on the correct understanding of the meaning of diagrams.

Task Analysis

Conventional devices in Euler and Venn diagrams

In our experiment, we use the following three types of diagrams: Euler diagrams, Venn diagrams having two circles, which we call "2-Venn diagrams", and Venn diagrams having three circles, which we call "3-Venn diagrams". Typical examples are shown in Figure 4.

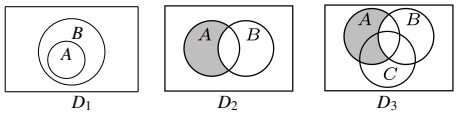


Figure 4: Representations of *All A are B* in Euler diagram (D_1), 2-Venn diagram (D_2), and 3-Venn diagram (D_3).

Euler diagrams used in our experiment are the ones introduced in Mineshima et al. (2008). Our system has the following features: (i) it uses a named *point* 'x' to indicate the existence of objects; (ii) it adopts a convention of *crossing*, according to which two circles which are indeterminate with respect of their relationship are put to partially overlap each other. Consequently, a single categorical statement can be represented by just a single diagram (see D_7 in Figure 5). This contrasts with another version of Euler system, which requires more than one diagrams to represent some categorical sentences, and hence has the well-known problem of combinatorial complexities (see chapter 4 of Johnson-Laird, 1983).

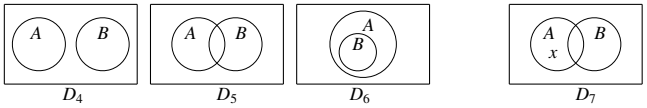


Figure 5: The diagrams corresponding to "Some *A* are not *B*" in traditional Euler system (D_4, D_5 and D_6) and the one in our Euler representation system (D_7)

In Venn diagrams, every circle partially overlaps each other, as in D_2 and D_3 in Figure 4 and D_5 in Figure 5. Such diagrams do not convey any specific information about circles, hence are subject to the convention of crossing. Meaningful relations among circles are then expressed using a novel device, *shading*, by the convention that a shaded region denotes an empty set. For example, the statement "All *A* are *B*" is represented as D_2 or D_3 in Figure 4. Note that the same information can also be conveyed by the Euler diagram D_1. Furthermore, 3-Venn diagrams use a link to connect points, which represent the disjunctive information about a point (see D_4^v and D_5^v in Figure 7 below).

Constructions of diagrammatic proofs

We compare syllogism solving tasks using these three types of diagrams in terms of difficulties in constructing the corresponding diagrammatic proofs. Deductive reasoning generally requires combining information in premises. Such a task could naturally be replaced by a task of combining presented diagrams. We expect that an inference process of combining diagrams is relatively easy to access, and accordingly, that an internal construction of a diagrammatic proof is naturally triggered if it consists only of such combining processes.

Reasoning with Euler diagrams As an illustration, consider a syllogism *All B are A, Some C are B; therefore Some C are A*. A solving process of this syllogism using Euler diagrams is shown in Figure 6.

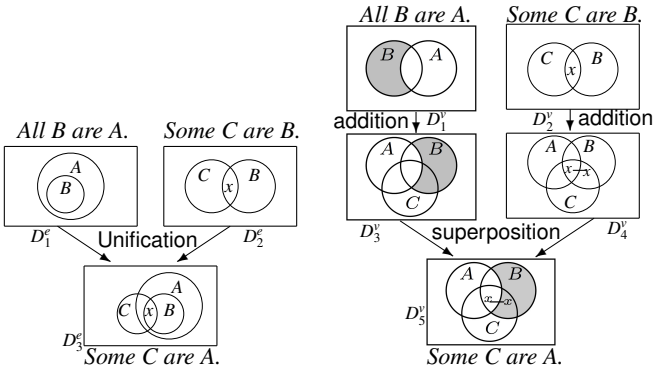


Figure 6: A proof of syllogism using Euler diagrams.

Figure 7: A proof of syllogism using Venn diagrams.

In general, a diagrammatic proof of a syllogism with Euler diagrams consists of a step of combining premise diagrams, which we call a *unification* step. It is expected that unification steps are relatively easy to access, so that such constructions of diagrammatic proofs are naturally triggered.

Reasoning with 2-Venn diagrams A solving process in 2-Venn diagrams is illustrated in Figure 7, where the premise *All B are A* is represented by D_1^v, and the premise *Some C are B* by D_2^v. Since these two diagrams contain a different circle, in order to combining them while preserving the syntax of Venn diagrams, one needs to accommodate their circles. In this case, circle C is added to D_1^v, and A to D_2^v. Then, by superposing the shaded region of D_3^v on D_4^v, one obtains diagram D_5^v, from which the conclusion "Some C are A" can correctly be read off. Here we can see that a solving process in 2-Venn diagrams consists of two steps, which we call *addition* and *superposition*. Note that a process of combining Euler diagrams, namely *unification*, can exploit the movements of circles as in Figure 6, whereas a process of combining Venn diagrams, namely *superposition*, operates on premise diagrams with the same number of circles, and hence does not involve any movement of circles.

Reasoning with 3-Venn diagrams If we start the proof with the 3-Venn diagrams D_3^v and D_4^v in Figure 7, we can skip the steps of adding a circle as required in the case of 2-Venn diagrams. The only step needed is the *superposition* step, which is expected to be relatively easy to access.

Predictions

Intuitively, 2-Venn diagrams seem to be relatively difficult to handle in solving syllogisms. For in order to construct diagrammatic proofs from 2-Venn diagrams, one has to know the relevant inference rules and strategies in advance. More specifically one has to know the successive processes of adding a circle and superposing two diagrams, as indicated in Figure 7. We expect that those who are ignorant of such a solving strategy could not appeal to concrete manipulations of the diagrams. They seem to have to draw a conclusion solely based on usual linguistic inference, with the help of semantic information obtained from 2-Venn diagrams.

To test this point, we introduce set-theoretical expressions corresponding to Venn diagrams, such as $A \cap \overline{B} = \emptyset$ for "All A are B" and $A \cap B \neq \emptyset$ for "Some A are B", as a control condition. We assume that they could only contribute to interpreting premises of syllogisms. Thus, they are used to check whether the effects of Euler, 3-Venn, and 2-Venn diagrams are interpretational or not.

In contrast to 2-Venn diagrams, both Euler diagrams and 3-Venn diagrams seem to be relatively easy to handle even for those users who are not trained to manipulate them in syllogism solving. The essential steps involved are unification and superposition steps. Given the fact that deductive reasoning generally requires combining the information in premises, these processes seem to be natural enough so that they would be immediately accessible to users. We expect that users could exploit natural constraints of diagrams and extract the right rules to draw a conclusion from Euler diagrams and 3-Venn diagrams themselves.

We will say that diagrammatic representations are *self-guiding* if the constructions of diagrammatic proofs are automatically triggered even for subjects without explicit prior knowledge of inferential rules or strategies. Then, our hypothesis amounts to saying that in syllogistic reasoning tasks, Euler diagrams and 3-Venn diagrams are self-guiding, whereas 2-Venn diagrams are not.

Based on these considerations, we predict that the performance in syllogism solving would be better when subjects use Euler diagrams or 3-Venn diagrams than when they use symbolic (set-theoretical) representations. We also predict that there would be little difference between the performance with 2-Venn diagrams and with the symbolic representations.

Method

Subjects are provided with instructions on the meanings of diagrams and then required to solve syllogistic reasoning tasks with diagrams. We conducted a pretest to check whether subjects understood the instructions correctly. The pretest was designed mainly to see whether subjects correctly understood the conventional devices of each diagram, in particular, the convention of crossing in both Euler and Venn diagrams and shading and linking in Venn diagrams.

Participants

365 undergraduates (mean age 19.78 $\pm$ 2.69 SD) in six introductive philosophy classes took part in the experiments. They gave a consent to their cooperation in the experiments, and were given small reward after the experiments. The subjects were native speakers of Japanese. The sentences and instructions were given in Japanese. The subjects were divided into four groups: Symbolic, 2-Venn, 3-Venn, and Euler groups. The four groups in this order consisted of 90, 95, 114, and 66 students, respectively. From each we excluded 26, 27, 35, 3 students (those who gave up before the end), respectively. It is notable that fewer students in the Euler group gave up compared to the other three groups.

Materials

The experiment was conducted in the booklet form.

Pretest The subjects of all groups were presented with ten representations (ten diagrams or ten set-theoretical expressions). They were asked to choose, from a list of five possibilities, all sentences which correspond to a given representation. The highest possible score on the pretest of the Symbolic group was ten and the cutoff point was set to be five. The highest possible score on the pretests of the 2-Venn, 3-Venn, and Euler groups was twelve, because there were two correct answers in two of the ten problems. Their cutoff point was set to be eight. These cutoff points were chosen carefully, based upon the results of our pilot experiments. The total time in Symbolic, 2-Venn and Euler groups was 5 minutes. The total time in the 3-Venn group was 6 minutes, since the instruction was longer than those of the other three groups. Before the pretest, the subjects in each group were presented with three examples.

Syllogistic reasoning tasks The subjects in the Symbolic group were given syllogisms with set-theoretical representations (such as the one in Figure 8). The subjects in the 2-Venn group were given syllogisms with Venn diagrams having two circles in premises (such as the one in Figure 9). The subjects in the 3-Venn group were given syllogisms with Venn diagrams having three circles in premises (such as the one in Figure 10). The subjects in the Euler group were given syllogisms with Euler diagrams (such as the one in Figure 11). We gave 31 syllogisms in total, out of which 14 syllogisms had a valid conclusion and 17 syllogisms had no valid conclusion. The subjects were presented with two premises and were asked to choose, from a list of five possibilities, a sentence corresponding to the valid conclusion. The list consists of *All-*, *No-*, *Some-*, *Some-not*, and *NoValid*. The subject-predicate order of each conclusion was *CA*. The test was a 20-minute power test, and each task was presented in random order (10 patterns were prepared). Before the test, the examples in Figure 8, 9, 10, and 11 were presented to each group.

All B are A. $B \cap \overline{A} = \emptyset$

All C are B. $C \cap \overline{B} = \emptyset$

1. All C are A.
2. No C are A.
3. Some C are A.
4. Some C are not A.
5. None of them.

Correct answer: 1

Figure 8: Example of reasoning task of Symbolic group

All B are A. 

All C are B. 

1. All C are A.
2. No C are A.
3. Some C are A.
4. Some C are not A.
5. None of them.

Correct answer: 1

Figure 9: Example of reasoning task of 2-Venn group

All B are A. 

All C are B. 

1. All C are A.
2. No C are A.
3. Some C are A.
4. Some C are not A.
5. None of them.

Correct answer: 1

Figure 10: Example of reasoning task of 3-Venn group

All B are A. 

All C are B. 

1. All C are A.
2. No C are A.
3. Some C are A.
4. Some C are not A.
5. None of them.

Correct answer: 1

Figure 11: Example of reasoning task of Euler group

Procedure

All four groups were first given 1 minute 30 seconds to read one page instructions on the meaning of categorical sentences. In addition, the Symbolic group was given 2 minutes to read two pages instructions on the meaning of set-theoretical representations. The 2-Venn and Euler groups were given 2 minutes to read two pages instructions on the meaning of diagrams. The 3-Venn group was given 3 minutes to read two pages instructions on the meaning of diagrams. Before the pretest, all groups were given 1 minute 30 seconds to read two pages instructions on the pretest. Finally, before the syllogistic reasoning test, all four groups were given 1 minute 30 seconds to read two pages instructions, in which the subjects were warned to choose only one sentence as answer and not to take a note. These time limits were set based upon the results of our pilot experiments.

Results and Discussion

Pretest

In the Symbolic group, 39 students scored less than 5 on the pretest. In the 2-Venn group, 38 students scored less than 8 on the pretest. In the 3-Venn group, 41 students scored less than 8 on the pretest. In the Euler group, 18 students scored less than 8 on the pretest. These students are excluded from the following analysis.

Syllogistic reasoning tasks

Figure 12 shows the average accuracy rates of the total 31 syllogisms in each group. The rate for the Euler group was 85.2%, the rate for the 3-Venn group was 75.2%, the rate for the Venn group was 66.6%, and the rate for the Symbolic group was 58.7%.

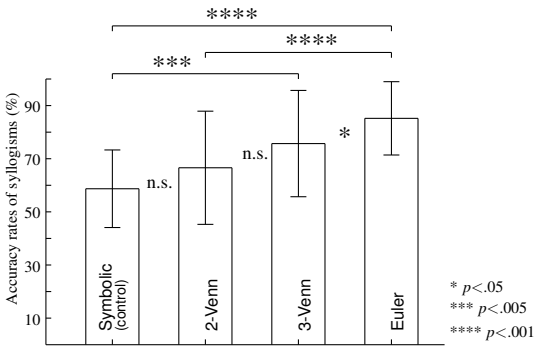


Figure 12: The average accuracy rates of 31 total syllogisms in the Symbolic, 2-Venn, 3-Venn, and Euler groups (error-bar refers to SD).

These data were subjected to a one-way Analysis of Variance (ANOVA). There was a significant main effect, F (3, 134) = 13.680. $p < .001$. Multiple comparison tests by Ryan's procedure yield the following results: (i) There was a significant difference between the Symbolic group and the Euler group, F (1, 68) = 5.935, $p < .001$. (ii) There was a significant difference between the Symbolic group and the 3-Venn group, F (1, 61) = 3.578, $p < .005$. (iii) There was no significant difference between the Symbolic group and the 2-Venn group. (iv) There was a significant difference between the 2-Venn group and the Euler group, F (1, 68) = 4.397, $p < .001$. (v) There was no significant difference between the 2-Venn group and the 3-Venn group. (vi) There was a significant difference between the 3-Venn group and the Euler group, F (1, 81) = 2.537, $p < .05$. It should be noted that if we include those subjects who failed the pretest, we still obtain similar results in each comparison: for (i), (iv) and (v), there were significant differences, $p < .001$; for (ii), there was a significant difference, $p < .01$; for (iii) and (v), there were no significant differences.

The results show that the performances of the Euler group and the 3-Venn group were better than that of the Symbolic group. This provides evidence for our hypothesis that Euler and 3-Venn diagrams have inferential efficacy and are self-guiding in the sense specified above. This means that as far as these diagrams are concerned, the internal constructions of diagrammatic proofs exist, and they can be naturally triggered for subjects without prior knowledge of inference rules or strategies. By contrast, there was little difference between the performance of the 2-Venn group and that of the Symbolic group. This suggests that 2-Venn diagrams have only interpretational efficacy, and are not self-guiding in our sense.

The results shown in Figure 12 indicate that the performance of Euler group was better than that of 3-Venn group. In particular, there was a significant difference with respect to a particular type of syllogism, namely, invalid syllogisms having an existential sentence as one of their premises. The data was subjected to a 4×2 ANOVA. As a main result, (i) there was a significant difference between this type of syllogisms in the 3-Venn group (57.8%) and the other types in the same group (83.4%), F (1, 134) = 29.434. $p < .001$. (ii) there was a significant difference between this type of syllogism in the 3-Venn group (57.8%) and that in the Euler group (84.2%), F (1, 81) = 4.926. $p < .005$. (iii) there was no significant difference between this type of syllogisms in the 3-Venn group (57.8%) and that in the Symbolic group (48.8%).

The relative difficulty in syllogism solving with 3-Venn diagrams could seem to be attributed to the difficulty in the process of drawing a conclusion from an internally constructed diagram. Such a process of extracting information may be formulated as a process of *deletion*. A deletion step in an Euler diagrammatic proof (as illustrated to the left in Figure 13) is simple in that it only requires to remove a circle without adjusting any other part of the diagram.

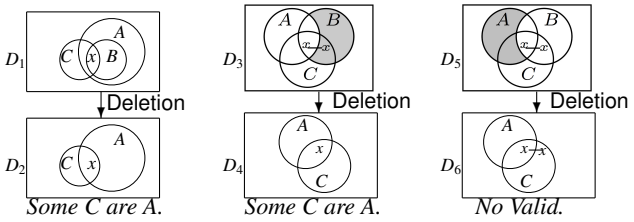


Figure 13: Deletion steps in Euler and Venn diagrams.

By contrast, deletion steps in 3-Venn diagrammatic proofs are somewhat complicated. Especially, in the step from D_5 to D_6 in Figure 13, which is an instance of invalid syllogisms having an existential sentence as one of its premise, one has to remove a circle and shading and to leave a linking point at the same region. Such complexities in deletion steps seem to reflect complexities of the processes of observing conclusions, and hence cause the difficulty in this type of syllogisms.

If our analysis is correct, the complexity of diagrams could make difficult the processes of *extracting* information. On the other hand, our results of the total 31 syllogisms suggest that 3-Venn diagrams have inferential efficacy, while 2-Venn diagrams do not. This in turn suggests that conventional devices such as shading and linking points could facilitate the processes of *combining* information by means of superposition of two diagrams. Thus, we could say that the availability of the process of combining information in diagrams depends on the complexity of the inference processes involved, whereas the availability of the process of extracting information in diagrams depends on the complexity of the conventional devices involved. Stenning and Oberlander (1995) point out that efficacy of diagrams can be ascribed to "specificity" of diagrammatic representations, and argue that diagrams could be effective because of their limited expressive power, in particular of the inability to express indeterminate or disjunctive information. In view of this, our findings are particularly interesting since they show that conventional devices to deal with indeterminacy sometimes facilitate internal manipulations of diagrams, hence contribute to their efficacy.

References

Allwein, G. & Barwise, J. (Eds.). (1996). *Logical Reasoning with Diagrams*. New York: Oxford Univ. Press.

Barwise, J., & Etchemendy, J.(1991). Visual information and valid reasoning. In G. Allwein & J. Barwise (Eds). *Logical Reasoning with Diagrams* (pp.3-26). New York: Oxford Univ. Press.

Bauer, M. & Johnson-Laird, P. N. (1993). How diagrams can improve reasoning. *Psychology Science*, 4(6), 372-378.

Blackburm, P. & Bos, J. (2005). *Representation and Inference for Natural Language*. Stanford: CSLI Publications.

Dobson, M. (1999). Information enforcement and learning with interactive graphical systems. *Learning and Instruction*, 9, 365-390.

Glasgow, J., Narayanan, N.H., & Chandrasekaran, B. (Eds.). (1995). *Diagrammatic Reasoning*. Cambridge, MA: MIT Press.

Howse, J., Stapleton, G., & Taylor, J. (2005). Spider diagrams. *LMS Journal of Computation and Mathematics*, 8, 145-194.

Levesque, H.J. (1988) Logic and the complexity of reasoning. *Journal of Philosophical Logic*, 17, 335-389.

Larkin, J. & Simon, H. (1987). Why a diagram is (sometimes) worth 10,000 words. *Cognitive Science*, 11, 65-99.

Mineshima, K., Okada, M., Sato, Y & Takemura, R. (2008). Diagrammatic reasoning system with Euler circles: theory and experiment design. In G. Stapleton et al. (Eds.), *The Proceedings of Diagrams 2008*, *LNAI 5223* (pp.188-205), Heidelberg: Springer.

Newstead, S. & Griggs, R. (1983). Drawing inferences from quantified statements *J. Verbal learning & verbal behavior*, 22, 535-546.

Scaife, M. & Rogers, Y. (1996). External cognition. *International Journal of Human-Computer Studies*. 45, 185-213.

Schwartz, D.L. (1995). Reasoning about the referent of a picture versus reasoning about the picture as the referent: an effect of visual realism. *Memory and Cognition*, 23(6), 709-722.

Shimojima, A. (1996). *On the Efficacy of Representation*. PhD thesis, Indiana University.

Shimojima, A. & Fukaya, T. (2003). Do we really reason about a picture the referent. In *Proceedings of the 25th Annual Conference of the Cognitive Science Society* (pp. 1076-1081).

Shimojima, A. & Katagiri, Y. (2008). Hypothetical drawing in embodied spatial reasoning. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 2247-2252).

Shin, S.-J.(1994). *The Logical Status of Diagrams*. Cambridge U.P.

Stenning, K. (1999). The cognitive consequences of modality assignment for educational communication: the picture in logic teaching. *Learning and Instruction*, 9, 391-410.

Stenning, K. (2002). *Seeing Reason*. Oxford Univ. Press.

Stenning, K., & Oberlander, J. (1995). A cognitive theory of graphical and linguistic reasoning. *Cognitive Science*, 19, 97-140.

Trafton, J.G. & Trickett, S.B. (2001). A new model of graph and visualization usage. *Proceedings of the 23th Annual Conference of the Cognitive Science Society* (pp. 1048-1053).

Zhang, J., & Norman, D.A. (1994). Representations in distributed cognitive tasks. *Cognitive Science*, 18(1), 87-122.

Learning verb alternations in a usage-based Bayesian model

Christopher Parisien and **Suzanne Stevenson**
Department of Computer Science, University of Toronto
Toronto, ON, Canada
{chris, suzanne}@cs.toronto.edu

Abstract

One of the key debates in language acquisition involves the degree to which children's early linguistic knowledge employs abstract representations. While usage-based accounts that focus on input-driven learning have gained prominence, it remains an open question how such an approach can explain the evidence for children's apparent use of abstract syntactic generalizations. We develop a novel hierarchical Bayesian model that demonstrates how abstract knowledge can be generalized from usage-based input. We demonstrate the model on the learning of verb alternations, showing that such a usage-based model must allow for the inference of verb class structure, not simply the inference of individual constructions, in order to account for the acquisition of alternations.

Keywords: Verb learning; language acquisition; Bayesian modelling; computational modelling.

Introduction

An important debate in language acquisition concerns the nature of children's early syntax. On one side of the debate lies a claim that children develop their syntactic knowledge in an item-based manner. This claim of usage-based learning argues that very young children associate verb argument structure with specific lexical items, only gradually abstracting syntactic knowledge after four years of age (*e.g.*, Tomasello, 2003). An alternative claim suggests that young children do indeed possess abstract syntactic representations—*i.e.*, generalizations about the structure of their language that are not necessarily tied to lexical items (*e.g.*, Fisher, 2002).

Syntactic alternation structure is often considered to be a central phenomenon in this debate. Consider the following example of the English dative alternation:

(1) I gave a toy to my dog.

(2) I gave my dog a toy.

These sentences mean roughly the same thing, but are expressed in different ways. The first, a *prepositional dative*, expresses the theme (*a toy*) as an object and the recipient (*my dog*) in a prepositional phrase. The second, a *double-object dative*, expresses both the theme and recipient as objects and reverses their order.

Verbs that allow similar alternations often have similar semantics (Levin, 1993), which suggests that alternations reflect much of our cognitive representations of verbs. Furthermore, these regularities appear to influence our language use. In word learning experiments, children as young as three years of age appear to use abstract representations of the dative alternation (Conwell & Demuth, 2007). While this is evidence of abstract syntax at a very young age, it does not necessarily invalidate the usage-based hypothesis, since the abstractions may originate from item-specific representations.

One way to bring these opposing positions together is to demonstrate, using naturalistic data, how to connect a usage-based representation of language with abstract syntactic generalizations. We argue that alternation structure can be acquired and generalized from usage patterns in the input, without a priori expectations of which alternations may or may not be acceptable in the language. We support this claim using a hierarchical Bayesian model (HBM) which is capable of making inferences about verb argument structure at multiple levels of abstraction simultaneously. We show that the information relevant to verb alternations can be acquired from observations of how verbs occur with individual arguments in the input. In this sense, we present a *competency* model showing what can be acquired, but we do not make claims regarding the specific processing mechanisms involved.

From a corpus of child-directed speech, our model acquires a wide variety of argument structure constructions over hundreds of verbs. Moreover, by forming classes of verbs with similar usage patterns, the model can generalize knowledge of alternation patterns to novel verbs. This stands in contrast to earlier models which have focused on either the acquisition of the constructions themselves, or the formation of classes over given constructions. The integration in our model of these two important aspects of verb learning has implications for current theories of language acquisition, by showing how abstract syntactic knowledge can be acquired and generalized from usage-level input.

Related work

Previous computational approaches to language acquisition have used HBMs to represent the abstract structure of verb use. Alishahi and Stevenson (2008) used an incremental Bayesian model to cluster individual verb usages (or *tokens*), simulating the acquisition of verb argument structure constructions. Using naturalistic input, the authors showed how a probabilistic representation of constructions can explain children's recovery from overgeneralization errors. In another Bayesian model of verb learning, Perfors et al. (2010) cluster verb *types* by comparing the variability of constructions for each of the verbs. The model can distinguish alternating from non-alternating dative verbs and can make appropriate generalizations when learning novel verbs.

Both of the above models show realistic patterns of generalization, but they operate at complementary levels of abstraction. The model of Alishahi and Stevenson does not capture the alternation patterns of verbs, while Perfors et al. assume that the individual constructions participating in the alternation have already been learned. Furthermore, Perfors et

al. limit their model to only consider two possible constructions (the prepositional and double-object dative), and only the verbs that participate in those constructions.

In this work, we address both levels of abstraction of the above models. We cluster individual verb usages to learn argument structure constuctions and their patterns of use across many verbs, and we also cluster verb types to learn alternation behaviour, generalizing that behaviour to novel verbs. Moreover, we use representative corpora of child-directed speech to model the acquisition of verb alternation behaviour in the context of many constructions, verbs, and alternations.

Vlachos et al. (2009) used a Dirichlet Process mixture model to cluster verb types by their subcategorization preferences, but did not address learning the argument structures themselves. Other work has modelled different aspects of the dative alternation, such as how discourse features affect the expression of dative constructions (de Marneffe et al., submitted), yet did not consider how these preferences are learned.

Model description

We discuss the feature representation of a verb usage and develop two contrasting models to show how alternation classes contribute to generalization in verb learning. Model 1 is an adaptation of an existing probabilistic topic model, the Hierarchical Dirichlet Process (HDP; Teh et al., 2006), to the problem of learning verb argument structure. Model 2, a novel extension to the HDP, addresses the limitations of Model 1 by learning verb alternation classes, allowing regularities in construction use to be transferred to novel verbs.

Verb features

Following from existing approaches (as in Joanis, Stevenson, and James (2008)), we use syntactic "slot" features to encode basic argument information about a verb usage. Table 1 presents the 14 features used in our representation. The first 12 (up through "PP") are binary features denoting the presence or absence of the stated syntactic slot, such as an object (OBJ) or a prepositional phrase (PP); the slots are indicated by labels used by the CHILDES dependency parser (Sagae et al., 2007).[1] When a PP is present, the nominal feature PREP denotes the preposition used. Such syntactic slot features are easier to extract than full subcategorization frames. We make the assumption that children at this developmental stage can distinguish various syntactic arguments in the input, but may not yet recognize recurring patterns such as transitive and double-object constructions. The following examples show this representation used with a double-object dative and a prepositional dative, respectively:

(3) I sent my mother a letter.
⟨ OBJ, OBJ2, PREP = null, NSLOTS = 2 ⟩

(4) I sent a letter to my mother.
⟨ OBJ, PP, PREP = to, NSLOTS = 2 ⟩

[1]We consider only the slots internal to the verb phrase, for now ignoring syntactic subjects. We also do not attempt to distinguish true arguments from adjuncts, a very difficult distinction to make.

Features	Description
OBJ, OBJ2	Objects
COMP, XCOMP	Clausal complements
PRED, CPRED, XPRED	Predicate complements
LOC	Locatives
JCT, CJCT, XJCT	Adjuncts
PP	Prepositional phrases
PREP	Preposition (nominal value)
NSLOTS	Number of slots used

Table 1: Slot features.

Model 1: Argument structure constructions

Like other topic models, the HDP (Teh et al., 2006) is essentially a model of category learning: the model clusters similar items in the input to discover structure. Adopting a usage-based approach to language (*e.g.*, Goldberg, 2006), we view the acquisition of verb argument structure as a category-learning problem. In this view, structured verb knowledge translates well to the hierarchical nature of the model.

Model 1 is a straightforward adaptation of the HDP to verb argument structure, which we will use as a point of comparison for an extended model. Figure 1(a) provides an intuitive description of the hierarchical levels of inference in Model 1. At level 1, the lowest level of abstraction, individual verb usages y_i are represented by sets of features as described above.

At level 2, the model clusters similar usages together to form argument structure constructions, where a construction is represented by a set of multinomial distributions, one for each feature. Since the clustering mechanism is *nonparametric*, we need not specify the total number of constructions to learn. Each of these constructions, denoted by its multinomial parameters θ, probabilistically represents a pattern such as a simple transitive or a prepositional dative. While a construction here encodes only syntactic information, with no semantic elements, the model can be generalized to a combined syntactic/semantic input representation.

At level 3, a multinomial distribution for each verb (π) represents the range of constructions that tend to occur with the verb. For example, in Figure 1(a), *give* (π_2) would have a high probability for the double-object dative and prepositional dative constructions (θ_2 and θ_3, respectively), but a low probability for the transitive construction, θ_1. Let y_{ij} denote feature j of usage i. Levels 1 through 3 are given by the following:

$$\begin{aligned} \pi_v &\sim \text{Dirichlet}(\alpha \cdot \beta) \\ z_i &\sim \text{Multinomial}(\pi_v) \\ \theta_{jz_i} &\sim \text{Dirichlet}(1) \\ y_{ij} &\sim \text{Multinomial}(\theta_{jz_i}) \end{aligned}$$

The indicator variable z_i selects a cluster (*i.e.*, a construction, one of the θ) for usage i. Given a verb v, this is drawn from a multinomial distribution which includes a small probability of creating a new construction.

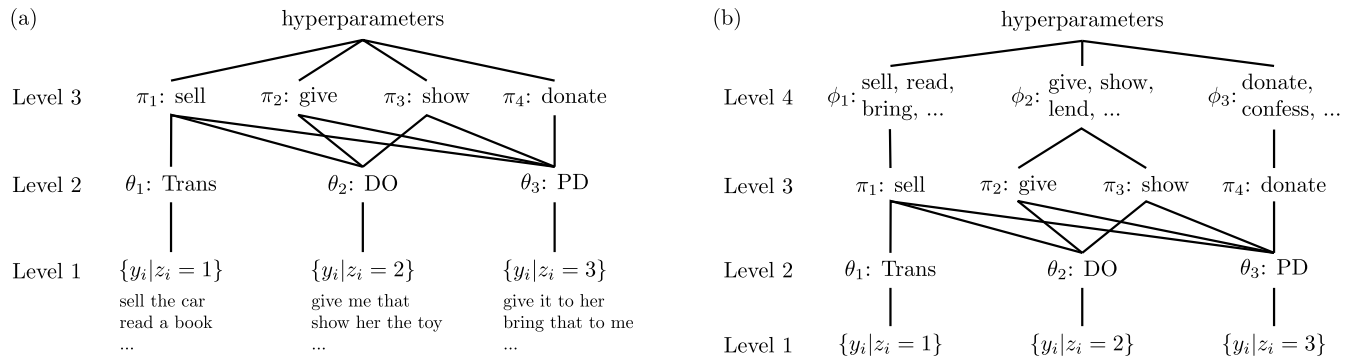


Figure 1: (a) Model 1, a Hierarchical Dirichlet Process applied to learning verb argument structure constructions. (b) Model 2, an extension of Model 1 to learn verb alternation classes.

The verb-specific distributions π_v depend on hyperparameters which encode expectations about constructions in general, across all verbs. They represent acquired knowledge about the likely total number of constructions, which constructions are more likely to occur overall, and so on:

$$\begin{aligned} \gamma &\sim \text{Exponential}(1) \\ \alpha &\sim \text{Exponential}(1) \\ \beta &\sim \text{Stick}(\gamma) \end{aligned}$$

As with lower-level parameters, these are influenced by observed structure in the input. β, drawn from a stick-breaking process (Stick), encodes how many constructions will be used and which constructions are more likely overall. α affects the variability of π_v. Large values of α push π_v closer to β, the global distribution over constructions, while smaller values encourage more variation among verbs. γ affects the total number of constructions; small values of γ correspond to fewer constructions. By drawing α and γ from an exponential distribution, we give a weak preference for verb-specific behaviour and for solutions with fewer constructions. These preferences are effectively designed into the model; they may be informed by general human category-learning behaviour. For further details of this model, see Teh et al. (2006).

Model 2: Alternation classes

Model 1 acquires argument structure constructions from individual verb usages, and learns how those constructions are used by individual verbs, but it is unable to recognize that certain *kinds* of verbs behave differently than others. Competent language speakers regularly use this kind of information. For example, if a verb occurs in a double-object dative construction, then we should infer that it is also likely to occur in a prepositional dative. We develop a novel extension of the above model to capture this phenomenon by learning clusters of similar verbs.

Recall that we represent a verb by a probability distribution over the constructions in which it may occur. In the example shown in Figure 1(a), *give* and *show* both tend to occur with a double-object dative and a prepositional dative, but are less likely to occur as simple transitives. By recognizing the similarity of π_2 and π_3, we can create a cluster containing *give*, *show*, and other similar verbs. Figure 1(b) presents this intuition in Model 2. We extend Model 1 by introducing a fourth level of abstraction, where we represent clusters of similar verbs. For each verb cluster c, we use ϕ_c to represent the range of constructions that tend to occur with any of the verbs in that cluster. By serving as a prior on the verb-level parameters π_v, ϕ_c directly influences each verb in the cluster.

The lower levels of this model are the same as in Model 1. In addition, the verb representations, π_v, depend on the alternation classes in level 4:

$$\begin{aligned} \phi_{c_v} &\sim \text{Dirichlet}(\alpha_0 \cdot \beta_0) \\ \pi_v &\sim \text{Dirichlet}(\alpha_1 \cdot \phi_{c_v}) \\ z_i &\sim \text{Multinomial}(\pi_v) \\ \theta_{jz_i} &\sim \text{Dirichlet}(1) \\ y_{ij} &\sim \text{Multinomial}(\theta_{jz_i}) \end{aligned}$$

Each verb v belongs to a cluster of verbs, denoted c_v. Now, π_v depends on ϕ_{c_v}, which gives a distribution over constructions for all the verbs in the same cluster.

As before, these parameters themselves depend on top-level hyperparameters:

$$\begin{aligned} \gamma_0 &\sim \text{Exponential}(1) \\ \alpha_{0,1} &\sim \text{Exponential}(1) \\ \beta_0 &\sim \text{Stick}(\gamma_0) \end{aligned}$$

These hyperparameters serve similar roles to those in Model 1. β_0 gives a global distribution over all the constructions in use. γ_0 affects the total number of constructions overall. α_1 affects the variability of a verb compared with its class, and α_0 affects the variability of verb classes.

To group verbs into alternation classes, we use a mechanism similar to the way we group individual verb usages into constructions. Recall that c_v acts as an indicator variable, selecting a class for verb v from the available classes in level 4. This is drawn from a multinomial distribution σ which includes a small probability of creating a new verb class:

$$\begin{aligned}\gamma_1 &\sim \text{Exponential}(1)\\ \sigma &\sim \text{Stick}(\gamma_1)\\ c_v &\sim \text{Multinomial}(\sigma)\end{aligned}$$

As with earlier uses of the stick-breaking construction, γ_1 affects the expected total number of verb classes. This method of clustering verb types is similar to Wallach (2008).

Parameter estimation

Models 1 and 2, as written, each specify a prior distribution over the complete set of possible parameters to the models (*i.e.*, all possible values for θ, $\mathbf{z}$, ϕ, and so on). We update these distributions using the observed verb usage data, thus obtaining posterior distributions over parameters.

We estimate the posterior distributions using Gibbs sampling, a Markov Chain Monte Carlo (MCMC) method (Teh et al., 2006). Model parameters are initially set randomly, then iteratively adjusted according to the observed data. We randomly set each z_i to one of 10 initial constructions, and each c_v to one of 10 verb classes (if applicable). We set the remaining parameters to random values drawn from the distributions specified in the model descriptions. We then iteratively update each model parameter *individually* by drawing it from a posterior distribution conditioned on the data and all the *other* parameters in the model. As we iterate through the parameters many times, we collect samples of their values. Over time, this set of samples converges on the posterior distribution—*i.e.*, the model parameters given the observed data. In the experiments, we average over this set of samples to estimate what each model has learned about the input.

Experimental set-up

We use child-directed speech from the Manchester corpus (Theakston et al., 2001), part of the CHILDES database (MacWhinney, 2000). The corpus covers 12 British English-speaking children between the ages of approximately 2 and 3 years. Using CLAN, we extract all child-directed utterances containing at least one verb. We parse the utterances with the MEGRASP dependency parser (Sagae et al., 2007), then reserve every second usage for an evaluation dataset, using the remainder for development. As described above, we extract 14 slot features for each verb usage. The datasets corresponding to each child contain between 4,400 and 10,700 usages and between 239 and 479 verb types. All reported results are obtained using the evaluation data.

Due to flaws in the automatic part-of-speech tagging and parsing, the data contains many errors, particularly in ditransitive constructions. We manually correct the portion of the input related to the dative alternation. For each verb in the development set that occurs with at least one prepositional or double-object dative (as given by the automatic parsing), we draw a sample of up to 50 usages. We repair any cases of incorrectly parsed dative constructions, then duplicate the corrected samples as necessary. Since manual annotation is so labour-intensive, we use this same sample to correct the data for corresponding verbs in the evaluation set. We assume that the proportions of various usages are identical for these verbs across the development and evaluation sets.

We implement both learning models using an adaptation of the NPBayes package (Release 1).[2] For each of the 12 children in the input, we run 10 randomly initialized simulations. The parameters appear to converge within 3,000 iterations, so we run each simulation for 5,800 iterations, discarding the first 3,300 as burn-in. We record a sample of the model parameters on every 25th iteration after the burn-in, giving 100 samples per simulation, 1,000 per child. By averaging over these samples, we can examine the models' behaviour.

Experiments

We compare the ability of our two models to acquire knowledge about the usage patterns of verbs in the input and generalize that knowledge to new verbs. Firstly, we examine construction preferences in two related classes of verbs. Secondly, we test whether the models use an abstract representation of the dative alternation to help learn new verbs.

Verb argument preferences

We examine how our models acquire the usage patterns of verbs in the input by looking at verbs that participate in two different alternation patterns. Earlier, we demonstrated the dative alternation in examples (3) and (4). The benefactive alternation is a related pattern, in which verbs alternate between a double-object form and a *prepositional benefactive* form, as in the following examples:

(5) John made his friend a sandwich.
⟨ OBJ, OBJ2, PREP = null, NSLOTS = 2 ⟩

(6) John made a sandwich for his friend.
⟨ OBJ, PP, PREP = for, NSLOTS = 2 ⟩

We consider all verbs involved in the dative and benefactive alternations, as listed by Levin (1993, Sections 2.1 and 2.2). We test three constructions: the prepositional dative (PD); the double-object construction (DO), whether dative or benefactive; and the prepositional benefactive (PB). Using the samples of the model parameters, we estimate the posterior predictive likelihood of each of these frames for each of the verbs in the given classes. For a given test frame $\mathbf{y}_0$, using verb v, and the observed data $\mathbf{Y}$,

$$\begin{aligned}P(\mathbf{y}_0|\mathbf{Y}) &= \sum_k P(\mathbf{y}_o|k,\mathbf{Y})P(k|v,\mathbf{Y})\\ &= \sum_k \prod_j P(y_{0j}|\theta_{jk})P(k|\pi_v) \qquad (1)\end{aligned}$$

This likelihood is averaged over all 1,000 samples per child.

Figure 2 shows the behaviour of both models. We average the likelihoods over all 12 children, and over all verbs in the following cases: (a) verbs listed as dative but not benefactive, (b) verbs listed as benefactive but not dative, and (c) verbs in both classes. In both models, both dative and benefactive verbs show a high likelihood for the DO frame, and a somewhat higher likelihood for the appropriate prepositional frame

[2] http://www.gatsby.ucl.ac.uk/~ywteh/research/software.html

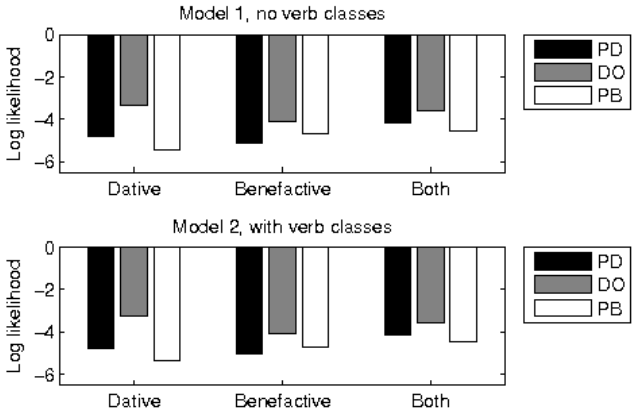


Figure 2: Argument preferences for known dative and benefactive verbs in Models 1 and 2. Shorter bars indicate higher likelihood. The two models show similar behaviour.

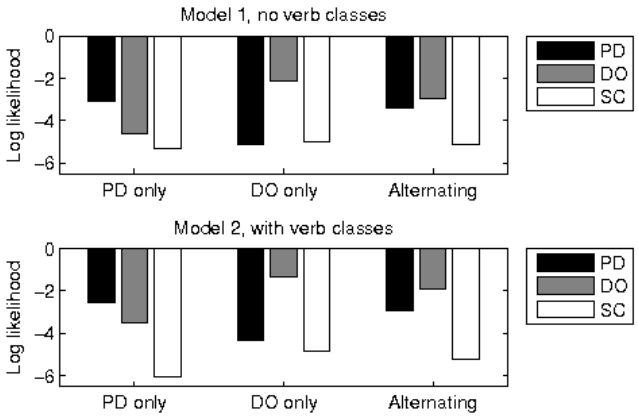


Figure 3: Generalization of novel dative verbs in Models 1 and 2, under various training conditions. Shorter bars indicate higher likelihood.

(PD and PB, respectively) than for the inappropriate one (PB and PD, respectively). Verbs that occur in both classes show closer likelihoods for all three frames.

These results suggest that both models can acquire the argument structure preferences of verbs in the input. In this case, the ability of Model 2 to acquire verb alternation classes is not necessary. Both models are able to cluster verb usages into a range of constructions and acquire appropriate usage patterns over a range of verbs. Both models acquire approximately 20 different constructions. Model 2 acquires 35-40 verb classes, depending on the child.

Novel verb generalization

Children as young as three years of age have been shown to use abstract representations of the dative alternation (Conwell & Demuth, 2007). When young children hear a sentence like *I gorped Charlie the duck*, they appear to know that the same meaning can be expressed by saying *I gorped the duck to Charlie*. We test this generalization in our models by presenting a novel verb in one form of the dative and measuring the likelihood of the alternating form.

We test each model by independently presenting it with a novel verb in three different situations: (a) two instances of the prepositional dative, (b) two instances of the double-object dative, or (c) one instance of each. Only in case (c) is the verb explicitly seen to be alternating. We test the ability to generalize alternation behaviour by comparing the likelihood of the unseen *alternating* form with an unseen form unrelated to the alternation. The non-alternating frame is the sentential complement (SC) frame, which occurs in 1-1.5% of the input, approximately the same overall frequency as either of the two dative frames. For example, if we train the novel verb using only the PD, yet the DO frame shows a higher likelihood than the unrelated SC frame, then we can say that the model has generalized the dative alternation.

Since the novel verbs are *not* in the observed data, we must further iterate the Gibbs sampler, using the new data, to obtain the appropriate samples of the verb-level distribution π_v. For each of the 1,000 parameter samples per child we obtained from the original simulations, we re-initialize the model with the parameters from the sample, add in the novel data for case (a), (b), or (c), then do a further 350 iterations, recording 10 new samples of the model parameters. This gives 10,000 new samples per test case, per child. Using equation (1) and the new samples, we estimate the posterior predictive likelihood of each of the three constructions. This gives an estimate of the relative preferences for a verb's usage and is a direct measure of the acquired lexicon. Translating this estimate to production, as seen by Conwell and Demuth (2007), would require a model of how discourse and other factors influence dative production (*e.g.*, de Marneffe et al., submitted). This is beyond the scope of this paper.

Figure 3 shows how the ability to acquire verb classes aids generalization. In Model 1, without verb classes, only the frames already seen with the novel verb are highly likely. This means that Model 1 is unable to generalize beyond observed data. In contrast, Model 2 shows appropriate generalization for the dative alternation. When the novel verb is trained with the prepositional dative, the double-object dative shows a much higher likelihood than the unrelated SC frame. A similar effect occurs with DO-only training: the PD frame is now more likely than the SC frame, although only slightly. Compared with Model 1, *both* dative frames obtain a higher likelihood across all three training cases, while the SC likelihood remains low. The ability to acquire alternation classes improves the ability to learn *both* alternating constructions.

One aspect of our results differs from the behaviour observed in children. Our verb-clustering model is more likely to generalize to the double-object form when trained only on a prepositional form, than the other way around (i.e., generalizing from a DO to a PD). However, three-year-old children seem to be biased to the prepositional form, the opposite effect (Conwell & Demuth, 2007). We suggest that this is a result of our small corpora. High-frequency dative verbs

tend to be biased toward the double-object form (Campbell & Tomasello, 2001). However, Gries and Stefanowitsch (2004) show that out of 40 alternating verbs in the larger ICE-GB corpus, 19 are prepositional-biased. This strongly suggests that more low-frequency verbs are prepositional-biased than otherwise. A small corpus will likely over-represent a double-object bias because of undersampling of low-frequency verbs. By applying Model 2 to larger corpora of child-directed speech in future work, we hope to correct this issue.

Conclusions

In this paper, we show how verb alternation classes contribute to generalization in verb learning. We develop a hierarchical Bayesian model, Model 2, that is capable of acquiring knowledge of verb argument structure at multiple levels of inference simultaneously. We demonstrate this using the wide range of verbs and constructions contained in a corpus of naturalistic child-directed speech.

By clustering individual verb usages, both of our models acquire a variety of argument structure constructions and learn their patterns of use over hundreds of verbs. Furthermore, Model 2 learns groups of verbs that occur with similar usage patterns. Using the dative alternation as a key example, we demonstrate how this knowledge of alternation classes can be generalized to novel verbs, as observed in the behaviour of children and adults. This verb class model can acquire and apply this knowledge without any prior expectation of which constructions and alternations may or may not be relevant.

In contrast to previous analyses of the dative alternation (Perfors et al., 2010; de Marneffe et al., submitted), we demonstrate its acquisition in the context of many other constructions, verbs, and alternations. Despite the low frequency of the participating constructions, our model successfully acquires the dative alternation. This is a strong endorsement of hierarchical Bayesian models of language acquisition.

This approach offers a potential bridge between differing theoretical positions in language acquisition. By simultaneously learning at multiple levels of abstraction, our model connects a usage-based representation of language, as proposed by Tomasello (2003), with weak abstract representations similar to those championed by Fisher (2002). Other usage-based Bayesian models, such as that of Alishahi and Stevenson (2008), offer a similar opportunity, although our model develops higher-level abstractions regarding the structured knowledge of verbs.

One of the key features of usage-based constructions is that they couple form to meaning (Goldberg, 2006). Moreover, Fisher argues that abstract syntactic representations influence semantics in verb learning, and vice-versa. By augmenting our model's input with semantic properties, we will examine the interaction of syntax and semantics in verb alternations. We will investigate how an argument alternation may convey semantic information, as in Scott & Fisher's (2009) demonstration of 28-month-old children inferring causation in transitivity-alternating verbs.

Acknowledgments

We thank Yee Whye Teh for valuable discussions, and NSERC and the University of Toronto for financial support.

References

Alishahi, A., & Stevenson, S. (2008). A probabilistic model of early argument structure acquisition. *Cognitive Science*, *32*(5), 789-834.

Campbell, A. L., & Tomasello, M. (2001). The acquisition of English dative constructions. *Applied Psycholinguistics*, *22*(02), 253–267.

Conwell, E., & Demuth, K. (2007). Early syntactic productivity: Evidence from dative shift. *Cognition*, *103*(2), 163–179.

de Marneffe, M. C., Grimm, S., Arnon, I., Kirby, S., & Bresnan, J. (submitted). A statistical model of the grammatical choices in child production of dative sentences.

Fisher, C. (2002). The role of abstract syntactic knowledge in language acquisition: a reply to Tomasello (2000). *Cognition*, *82*(3), 259-278.

Goldberg, A. E. (2006). *Constructions at work: the nature of generalization in language*. Oxford University Press.

Gries, S. T., & Stefanowitsch, A. (2004). Extending collostructional analysis: A corpus-based perspective on alternations. *Intl. J. Corpus Linguistics*, *9*, 97–129.

Joanis, E., Stevenson, S., & James, D. (2008). A general feature space for automatic verb classification. *Natural Language Engineering*, *14*(3), 337-367.

Levin, B. (1993). *English verb classes and alternations: A preliminary investigation.* University of Chicago Press.

MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk* (3rd ed., Vol. 2: The Database). Erlbaum.

Perfors, A., Tenenbaum, J. B., & Wonnacott, E. (2010). Variability, negative evidence, and the acquisition of verb argument constructions. *J. Child Language*, *37*(3), 607–642.

Sagae, K., Davis, E., Lavie, A., MacWhinney, B., & Wintner, S. (2007). High-accuracy annotation and parsing of CHILDES transcripts. In *Proc. ACL-2007 Wkshp on Cognitive Aspects of Computational Language Acquisition.*

Scott, R., & Fisher, C. (2009). Two-year-olds use distributional cues to interpret transitivity-alternating verbs. *Language and Cognitive Processes*, *24*, 777–803.

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006). Hierarchical dirichlet processes. *Journal of the American Statistical Association*, *101*(476), 1566-1581.

Theakston, A. L., Lieven, E. V. M., Pine, J. M., & Rowland, C. F. (2001). The role of performance limitations in the acquisition of verb-argument structure: an alternative account. *J. Child Language*, *28*, 127-152.

Tomasello, M. (2003). *Constructing a language: A Usage-Based theory of language acquisition*. Harvard U. Press.

Vlachos, A., Korhonen, A., & Ghahramani, Z. (2009). Unsupervised and constrained Dirichlet process mixture models for verb clustering. In *Proceedings of the EACL Workshop on Geometrical Models of Natural Language Semantics.*

Wallach, H. M. (2008). *Structured topic models for language*. Unpublished doctoral dissertation, Univ. of Cambridge.

Frequent Frames as Cues to Part-of-Speech in Dutch: Why Filler Frequency Matters

Richard Eduard Leibbrandt (Richard.Leibbrandt@flinders.edu.au)

David Martin Ward Powers (David.Powers@flinders.edu.au)
School of Computer Science, Engineering and Mathematics,
Flinders University, Adelaide, Australia

Abstract

The Frequent Frames model (Mintz, 2003) attempts to assign words to word categories based on their distributional patterns of usage. This model is highly successful in categorizing words in child-directed speech in English, but has been shown by Erkelens (2008) to be less effective with Dutch material. We show that extending the amount of contextual information in a frame by making use of the full utterance context does not improve categorization performance, but that constraining the fillers of Frequent Frames to be relatively less frequently occurring words does improve categorization significantly. We connect the latter result to a basic dichotomy in some languages between function words and content words, and conclude that, at least for English and Dutch, paying attention to this dichotomy is of greater importance for distributional bootstrapping proposals than the specific distributional contexts that are used to categorize words.

Keywords: Language learning; Distributional bootstrapping; Parts-of-speech; Function words; Frequent frames

Introduction

The parts-of-speech of a language (word classes such as nouns, verbs and adjectives) are of crucial importance in describing the grammar of the language. A vast amount of research has aimed to delineate the processes by which children learn to categorize words into the parts-of-speech of their native language. Researchers favouring *semantic bootstrapping* approaches (Grimshaw, 1981; Pinker, 1984) have proposed that early word categories are formed by grouping together words that refer to the same dimensions of concrete meaning, such as actions or objects. On the other hand, following early proposals by Maratsos & Chalkley (1980), proponents of *distributional bootstrapping* have argued that word categories can be induced by observing that certain groups of words are used in similar linguistic contexts, whether these contexts are defined at the level of words, morphemes, or even phonological or prosodic phenomena.

In recent years, it has become feasible to implement specific distributional bootstrapping proposals as computer algorithms that attempt to categorize words purely by analysing distributional patterns in large corpora of natural utterances (Cartwright & Brent, 1997; Redington, Chater & Finch, 1998). For instance, Redington et al. (1998) found that words in child-directed English speech could be categorized with a high level of success by considering only very local utterance contexts made up of words that occur in close proximity to the target word.

A particularly successful distributional model has been the *Frequent Frames* model of Mintz (2003, 2006a, 2006b). Frequent frames are defined as a disjunct frame occurring around a target word, made up of the word immediately preceding and the word immediately following the target, so that all frequent frames have the form *a _ b*, with *a* and *b* standing for specific words, and the underscore representing a slot that can accept a variety of filler words. For example, in the three-word sequence "a house and", the frame is "a _ and", and the filler is "house". Once all frames of this form have been collected from a corpus, only the most *frequent* ones are retained for the purpose of categorization. This reflects the intuition that, if two words co-occur frequently on either side of another word across several utterances, this is likely to be due to some meaningful linguistic relationship between them. All words occurring in the same frequent frame are assigned to the same category, and frames that have more than 20% overlap in their set of slot fillers have their categories amalgamated into larger, more general categories. This amalgamation step is crucially important: by grouping together frames that accept similar sets of words, the child may be able to hypothesize that a word used in one verb frame may also be legitimately used in another verb frame; without amalgamation, this kind of generalization is not possible.

Frequent Frames provide a very successful categorization of the words that occur in them, with Mintz (2003) reporting values greater than 0.9 for the evaluation measures *accuracy* and *completeness* when the model was implemented on a set of English corpora. However, recently Erkelens (2008) has shown that, in the case of child-directed speech in Dutch, Frequent Frames provide a less accurate basis for part-of-speech categorization than they do for English: whereas the use of Frequent Frames in English yielded an accuracy figure that exceeded the random baseline by 0.52 for tokens and 0.46 for types, a replication with a Dutch corpus could attain an improvement in accuracy over baseline of only 0.33 for tokens and 0.25 for types.

Full-Utterance Frames As Distributional Contexts

An important issue in distributional bootstrapping is to decide on the most appropriate usage contexts to consider for the purpose of categorization. One possible reason for the purported lower utility of Frequent Frames in Dutch

may be that Dutch simply allows a greater amount of flexibility in the range of structures in which particular words are able to occur. This raises the possibility that the contextual window employed by Frequent Frames may simply have been too small, and that it may be necessary to consider a wider amount of lexical context around a word in order to distinguish between different constructions. The maximum amount of context for a word used in an utterance is arguably the entire utterance, and so from a practical point of view it may be useful to explore the use of frames that comprise a full utterance at a time.

Tomasello (2006) has suggested a prominent role in language development for *utterance-level constructions*, expressions that can be used as complete utterances and are associated in a routinized way with certain communicative functions. Pine & Lieven (1993) provide evidence that some children assemble their earliest multi-word utterances by starting with "frozen", unanalysed phrases and proceeding to analyse these into fixed parts with variable slots into which various elements can ultimately be inserted (although some children instead form multi-word utterances by combining familiar single words together).

Given both these pragmatic and theoretical considerations, Leibbrandt and Powers (2008) evaluated a distributional bootstrapping proposal that makes use of schematic representations of complete utterances, with most words in the utterance lexically specified and one or two additional word positions serving as slots, for example "Are you going to X it?" or "That's the X". Under Leibbrandt & Powers's proposal, words that occur in the same full-utterance frame slot are categorized as belonging to the same word category. This approach was highly effective for categorizing word tokens in a natural corpus of child-directed English speech, attaining levels of correctness in part-of-speech classification that were comparable to those achieved by Frequent Frames (Mintz, 2003).

The Function Word - Content Word Dichotomy

Another important factor in distributional bootstrapping proposals is the basic dichotomy that exists in many languages between content word classes (the classes that carry lexical meaning, such as nouns, verbs and adjectives) and function word classes (the classes that are more closely involved with grammar, such as determiners, conjunctions and prepositions).

Attending to the positional relationships between function and content words has been proposed to be of importance to the language-learning child. For instance, Valian and Coulson (1988) found that learning an artificial language is made easier by increasing the frequency of function words that can serve as anchor points for distributional analysis, and suggest that children may seek out the most frequent elements in language in order to learn about the patterns in which parts-of-speech are allowed to occur in a language.

Gerken, Landau & Remez (1990) point out that function words could be crucial in the two tasks of *word segmentation* and *word labeling* (category assignment). Function words are potentially useful in segmentation because recognizing the relatively small number of function words makes it easier to separate out the far more heterogeneous open-class words that are interspersed between them. Function words could also aid labeling, because they occur in very stereotypical positional relations to open-class words, for instance, "the" is often followed by a noun (or sometimes by an adjective which is followed by a noun), and "-ing" is usually preceded by a verb root.

Because it cannot be assumed that children know *a priori* which words are function words and which are content words, this distinction would have to be learned on the basis of perceptible cues in the language spoken to children. English function words can be identified by a number of phonological cues, including syllable complexity, stress and vowel quality (Morgan, Shi & Allopenna, 1996), and even newborn infants are able to distinguish English function words from content words (Shi, Werker & Morgan, 1999).

Another feature of the distributional approach of Leibbrandt and Powers (2008) that may have contributed to its successful categorization performance is that it attempts to take the function word - content word dichotomy into account by making use of another cue that may plausibly be available to children: most function word types occur more *frequently* in speech than most content word types. Leibbrandt and Powers attempted to approximate the distinction between function words and content words in English by sharply distinguishing between the two sets of frequent and less frequent words, defined as respectively the set of the top *N* most frequently-occurring words in a corpus, and the set of all other words. When creating full-utterance frames for their distributional analysis, they applied the constraint that only frequent words could be used as the lexically-specific words in a frame, and only less-frequent words could be used as the slot fillers that were embedded in the frames.

Adapting The Frequent Frames Model

Erkelens (2008) argues that different cues are useful to differing extents in different languages, and that the occurrence of a word in a frequent frame is not as useful a cue to part-of-speech for the Dutch-learning child as it is for English. While we agree with the former point, we will attempt to show that the utility of Frequent Frames for categorization in Dutch may have been underestimated.

In the remainder of this paper, we report on a series of four experiments intended to investigate whether the Frequent Frames model can be modified to deal successfully with Dutch material. In Experiment 1, we replicate the results of Erkelens (2008) with a larger corpus and Frequent Frame set, and confirm that the unmodified Frequent Frames model is less useful for categorization in Dutch than in English. In Experiment 2, we investigate whether the distributional model of Leibbrandt & Powers (2008) is able to improve over the categorization results of Frequent Frames (we preempt our results here by confirming that it does). As Leibbrandt and Powers's approach differs in two

ways from the Frequent Frames approach, it then becomes important to investigate whether this improved performance is due to both differences, or only one. In Experiment 3, therefore, we modify the Frequent Frames approach to use the complete utterance as the context for categorization, and in Experiment 4, we constrain Frequent Frames to be composed of only frequent (function) framing words, and to take only less-frequent (content) filler words.

Experiment 1

Method

The corpus used in these experiments was the Groningen corpus (Wijnen & Bol, 1993), taken from CHILDES (MacWhinney, 2000), and consisting of data from seven Dutch-learning children in the Groningen area, recorded between the ages of 1;05 and 3;07. The corpus was minimally preprocessed for computer-readability, and all sentences uttered by adults were used. Data from all seven children were merged together in order to increase the data set size for the purpose of data clustering.

Frequent Frames were extracted according to the method used by Mintz (2003). Candidate frames were extracted from each utterance in the corpus, by forming a frame from every three consecutive words in each utterance and replacing the middle word with a slot marker. The frames with the highest frequency of occurrence in the corpus were retained as the set of *frequent* frames. Frequency statistics were collected on how often each word occurred in the slot position of each frequent frame throughout the corpus.

The studies by Mintz (2003) and Erkelens (2008) made use of a set size of 45. Because it was desirable in the present work to apply clustering to the data, a slightly larger frame set of 250 frames was used. These top 250 frames were grouped into clusters of frames by means of average-linkage hierarchical clustering (Sokal & Sneath, 1963), with the initial distances between frames given by Spearman's ranked correlation coefficient. Frames were clustered together if they occurred in the corpus with similar sets of slot-filler words.

Clustering makes it possible to make generalizations about the acceptability of words in frames in which they have not been attested in the corpus. If the clustering algorithm produces K clusters of frames, these clusters correspond to K hypothesized categories. Any word token which occurs in any frame belonging to a particular cluster is then assigned to the category corresponding to that cluster.

Evaluation Measures And Significance

All the experiments reported here involve the task of categorizing words into word categories based on the context in which they are used. In each case, there is an empirical allocation of words to unlabelled categories, which needs to be evaluated by a comparison with the "true" distribution of word tokens into their parts-of-speech. This "true" distribution was created by us after manually inspecting each of the particular word tokens in contextual usage. We made use of the same categories that were used by Erkelens (2008) in her "standard analysis", namely: verbs (including auxiliaries and copula), nominals (nouns, proper names and pronouns), adjectives, prepositions, adverbs, determiners, WH-words, conjunctions and interjections.

Unsupervised categorization models such as Frequent Frames are usually evaluated by means of the mathematical measures *accuracy* and *completeness*, using a *pair counting* approach. A formal definition of these two measures falls outside the scope of this paper, but they can be intuitively understood as expressing the extent to which word tokens assigned to the same category by the model do in fact belong to the same part-of-speech, and the extent to which word tokens which belong to the same part-of-speech were in fact categorized together by the model, respectively.

It is possible to report accuracy and completeness both in terms of the number of word *tokens* correctly categorized and in terms of the number of word *types* correctly categorized; results reported here are based on word type categorization only.

It should be noted that one cannot simply compare accuracy and completeness scores between experiments that make use of different sets of data. Any comparison has at least to take into account the magnitude of the *difference* between accuracy (or completeness) attained by the model, and the baseline accuracy (or completeness) attained by randomly allocating of words to categories.

In order to address this difficulty, we make use of *permutation tests* to assess the significance of differences between evaluation measures, both within and between experiments. Within an experiment, it is possible to assess whether the value of an evaluation measure is significantly higher than the baseline value, by generating a randomized sample of values for that measure and determining how often an equal or higher value occurs in the sample. Between experiments, it is possible to determine whether an obtained value for an evaluation measure in one experiment is significantly better than a value for the same measure in another experiment, by generating a random sample for each experiment separately, taking the differences between pairs of values from the two samples, and comparing this sample of differences to the difference between the originally obtained values.

There is also typically a trade-off between accuracy and completeness, and it is possible to artificially inflate one measure at the expense of the other. For this reason, it is necessary to considered both values together when evaluating the results of an experiment. In addition, we will also report values for the *F measure*, calculated as the harmonic mean of accuracy and completeness, which summarizes both measures and takes on a high value only when they are both high in value.

Results

Firstly, an analysis of the 250 most frequent frames in the pooled corpus confirmed the conclusion drawn by Erkelens

(2008) that frequent frames are not as reliable a basis for the categorization of Dutch words as they are for English words. The categorization accuracy for the top 250 frames from the pooled corpus (displayed in Table 1) was 0.60, against a random baseline of 0.31, i.e. categorizing on the basis of frequent frames rather than by randomly assigning categories improved accuracy by only 0.29. This result is comparable with Erkelens's (2008) 0.25 increase in accuracy, suggesting that the differences between that study and the current experiment (a different corpus and a larger set of frequent frames) did not materially affect the results. As might be expected, completeness was almost equal to the random baseline (and near zero), as frame clusters have not yet been created. These results confirm that the individual frames have some utility in predicting the part-of-speech of their slot-filler words, but that their accuracy is far from perfect.

Table 1: Accuracy and completeness for the top 250 frequent frames before clustering, against results from Mintz (2003) and Erkelens (2008). Random baseline figures in italics.

Study	*Language*	*Measure*	*Value*
Mintz (2003)	English	Accuracy	0.93 *(0.47)*
Erkelens (2008)	Dutch	Accuracy	0.58 (*0.33)*
Experiment 1	Dutch	Accuracy	0.60 *(0.31)*
Experiment 1	Dutch	Completeness	0.01 *(0.01)*

Hierarchical clustering was applied to the frames based on the distributional patterns of their filler words[1]. The results are shown in Table 2. Accuracy decreased sharply, as should be expected, as assigning every frame to its own unique category corresponds to the maximum attainable accuracy value. While completeness increased by a large amount in absolute value, it did not exhibit a large advantage over the random baseline completeness value.

Table 2: Evaluation of word token categorization after hierarchical clustering of the top 250 frequent frames into 12 clusters. Random baseline figures in italics.

Accuracy	0.429 *(0.327)*
Completeness	0.405 *(0.308)*
F	0.417 *(0.317)*

Discussion

These results replicate the findings of Erkelens (2008) that frequent frames have some utility as a basis for the prediction of the part-of-speech of Dutch words, but that they are not nearly as reliable as they are in English.

[1] The number of clusters produced by hierarchical clustering affects the obtained results. Procedures exist for choosing an optimal number of clusters, but for the sake of consistent comparison across the four experiments described here, the number of clusters produced was fixed at 12 in each experiment.

Experiment 2

Method

As in Experiment 1, we made use of the Groningen corpus. Here, however, we attempted to apply the lexically-specific frame approach proposed by Leibbrandt and Powers (2008). A list was compiled of the most frequently occurring word types in the Groningen corpus. This requires a choice of an arbitrary frequency cutoff point, and in this experiment the top 300 most frequent words were selected as the frame-building words. This set included the most common function words in Dutch, including pronouns (*ik*, *hij*, *ze*), determiners (*een*, *de*, *het*, *deze*, *dat*), and forms of the copula (*ben*, *zijn*) as well as a number of common content words.

All utterances were rewritten as lexically-specific frame candidates, by replacing every word that was not on the frequent-word list by a placeholder symbol X. From this set of candidate *dichotomous full-utterance frames*, the 250 frames with the highest frequency of occurrence were retained for analysis.

As in Experiment 1, co-occurrence data was collected about the frequency with which different words occurred in each of the frames, and the set of frames was clustered based on similarity in their sets of filler words. Note that, because of the way in which the frames were constructed, all slot fillers were taken from the set of less-frequent words.

Results

A number of intuitively sensible Dutch full-utterance frames were produced by this process, for example "Daar is de X" ("There's the X"), "Gaat ie X?" ("Is he going to X?") and "Heel X" ("Very X"), frames which could reasonably be expected to take noun, verb and adjective fillers respectively. Note that none of these example utterance structures could have been covered by the Frequent Frames approach, as the slot word occurs at the end of the frame in each case.

Table 3: Evaluation of word token categorization after hierarchical clustering of the top 250 dichotomous full-utterance frames. Random baseline figures in italics.

Accuracy	0.752 *(0.431)*
Completeness	0.407 *(0.233)*
F	0.528 *(0.302)*

The results of categorization evaluation after clustering are shown in Table 3. Accuracy, completeness and F were all significantly higher than baseline, as assessed by a permutation test ($p < 0.01$). Furthermore, categorization performance was significantly better than in the Frequent Frames approach of Experiment 1, as assessed by a permutation test of F value differences ($p < 0.01$).

Discussion

This experiment has shown that the frame selection approach used by Leibbrandt and Powers (2008) produces frames that are far more reliable indicators of the part-of-speech of a word in Dutch than the standard Frequent Frames proposed by Mintz (2003). As stated, this approach differs from Frequent Frames in two ways, making use of full-utterance contexts and accepting only less-frequent word (content word) fillers. In the next two experiments, we attempt to determine whether the improved performance shown here is due to one, or both, of these properties.

Experiment 3

Method

From each utterance in the corpus, candidate frames were extracted that contained all the words in the utterance except for one target word, which was turned into a variable slot (so that each utterance yielded as many candidate frames as there were words in the utterance). For example, the utterance "Dat is het vliegtuig" yielded the frames "X is het vliegtuig", "Dat X het vliegtuig", "Dat is X vliegtuig" and "Dat is het X". The most frequently occurring of these candidate *frequent full-utterance frames* were selected for evaluation. As before, word occurrence frequencies were calculated for each frame, and frames were clustered together based on the patterns of words that occurred in their slots.

Results

Evaluation results are shown in Table 4. While all evaluation measures were significantly greater than baseline ($p < 0.01$), the full-utterance frames did not provide a better basis for categorization than Frequent Frames; on the contrary, the categorization using Frequent Frames in Experiment 1 performed significantly *better* than the categorization with full-utterance frames, as assessed on a permutation test of differences in F measures ($p < 0.01$).

Table 4: Evaluation of word token categorization after hierarchical clustering of the top 250 frequent full-utterance frames. Random baseline figures in italics.

Accuracy	0.428 *(0.267)*
Completeness	0.249 *(0.156)*
F	0.315 *(0.197)*

Discussion

Clearly, the relatively low utility of Frequent Frames in categorizing Dutch words shown in Experiment 1 was not merely due to an insufficient amount of contextual information. In this experiment, increasing context to comprise the whole utterance did not improve categorization, as one might have expected from the superior results with dichotomous full-utterance frames in Experiment 2, and it seems that the success of those frames had nothing to do with their being based on full utterances.

Experiment 4

Method

Candidate frames were extracted from the corpus in the same way as for Experiment 1, i.e. the candidate frames were all Frequent Frames. However, following the approach taken in Experiment 2, we retained frames for the final evaluation set only if both the frame-building words (i.e. the first and third words) occurred in the list of the most frequent words in the corpus. Equally importantly, only words that were not in the frequent-word list were accepted as slot fillers for the frames. The most frequent such *frequent dichotomous frames* were selected and clustered as in the previous experiments.

Results

Evaluation results are shown in Table 5. It is immediately noticeable that the values of all measures are higher than for any of the other 3 experiments. For instance, comparison with Table 1 reveals that accuracy in this experiment is similar to the level of accuracy attained by Mintz (2003) with English frames. All measures are significantly above their baseline ($p < 0.01$), and categorization performance is significantly greater than for Experiments 1 and 3 ($p < 0.01$), but not significantly different from Experiment 2.

Table 5: Evaluation of word token categorization after hierarchical clustering of the top 250 frequent dichotomous frames. Random baseline figures in italics.

Accuracy	0.921 *(0.611)*
Completeness	0.513 *(0.340)*
F	0.659 *(0.437)*

Discussion

In this experiment, we have seen evidence that, contrary to the results of Erkelens (2008), Frequent Frames may yield high accuracy and completeness in categorizing Dutch words, *provided* that the frames are composed of the set of frequent words in Dutch (usually function words) while the categorization targets are taken from the relatively less frequent words, i.e. essentially content words.

Every one of the 250 frames in Experiment 1 was already composed of two frequent words. However, only 105 of those frames were retained in the evaluation set for Experiment 4. Therefore, the large improvement in categorization performance was due to the requirement that fillers should be less-frequent words, thereby effectively dropping function word fillers from the categorization. This seemed to result in a great number of frames being added to the evaluation set that were strongly associated with only one content word class (verbs, nouns, or adjectives).

General Discussion

The experimental results in this paper show that a simple distributional approach is effective in categorizing words in Dutch child-directed speech. Both the dichotomous full-utterance frame approach of Leibbrandt & Powers (2008) used in Experiment 2 and the Frequent Dichotomous Frames approach of Experiment 4 yielded significantly better categorizations than either Frequent Frames or Frequent Full-Utterance Frames, with no significant difference between the two models relative to their random baselines (although Frequent Dichotomous Frames achieved a higher F value in absolute terms, and so may arguably be preferred). By contrast, extending the context used for distributional analysis to a full utterance as in Experiment 3, paradoxically decreased performance. This suggests that the extent of context used in distributional bootstrapping may be less crucial than the kinds of words used for frames and fillers respectively

A simple modification to the Frequent Frames model is therefore able to overcome the shortcomings with Dutch material identified by Erkelens (2008). We suggest that the reason why the Frequent Dichotomous Frames model yields such a successful categorization is that the less-frequent words being categorized are mostly content words. In other words, it may be the case that the only useful targets for distributional analysis are the content word classes such as noun, verb, adjective and adverb. While pronouns, auxiliary verbs, etc. can also be identified by their distribution, these words may simply be learned on a one-by-one basis.

The weaker results of the original Frequent Frames model may have been due to a conflation of legitimate content word contexts with other cases where a function word in the frame slot indicates a different linguistic construction. For instance, the Frequent Frame "die _ wel" accepts a variety of verbs, and the pronoun "die" serves as the subject of the verb. When the slot is filled by the adverb "ook" to form "die ook wel", however, this is a different construction where "die" is the object of a verb outside the frame, or else the subject of an unstated verb. Eliminating function-word fillers avoids the conflation of contexts.

While the results from this corpus analysis speak less directly to how children *actually* learn parts-of-speech than the results of an experimental study would, they demonstrate the feasibility of exploiting a particular form of information in child-directed language. In concurrence with the proposal by Valian & Coulson (1988) that function words serve as anchor points indicating the structure of an utterance and facilitating distributional analysis, we suggest that it would be useful for children to make a distinction between function and content words, based on various cues such as phonology, greater occurrence frequency, etc. When children encounter a function word occurring in the slot position of what would normally be a distributional frame, they would then be able to avoid carrying out the normal process of categorizing the word on the basis of the frame, and to treat the function word as part of the structural information in the utterance only.

References

Cartwright, T. A., & Brent, M. R. (1997). Syntactic categorization in early language acquisition: formalizing the role of distributional analysis. *Cognition, 63*, 121-170.

Erkelens, M. (2008). Restrictions of frequent frames as cues to categories: the case of Dutch. *Supplement to the Proceedings of the 32nd Boston University Conference on Language Development (BUCLD 32)*.

Gerken, L., Landau, B., & Remez, R. E. (1990). Function morphemes in young children's speech perception and production. *Developmental Psychology, 26*(2), 204-216.

Leibbrandt, R. E., & Powers, D. M. W. (2008). Grammatical category induction using lexically-based templates. *Supplement to the Proceedings of the 32nd Boston University Conference on Language Development (BUCLD 32).*

MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk.* (3 ed. Vol. 2: The database). Mahwah, NJ: Lawrence Erlbaum.

Mintz, T. H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition, 90*(1), 91-117.

Mintz, T. H. (2006a). Finding the verbs: Distributional cues to categories available to young learners. In K. Hirsh-Pasek & R. M. Golinkoff (Eds.), *Action meets word: How children learn verbs*. Oxford: Oxford University Press.

Mintz, T. H. (2006b). Frequent frames: Simple co-occurrence constructions and their links to linguistic structure. In E. V. Clark & B. F. Kelly (Eds.), *Constructions in acquisition*. Stanford: CSLI Publications.

Morgan, J. L., Shi, R., & Allopenna, P. (1996). Perceptual bases of rudimentary grammatical categories: Toward a broader conceptualization of bootstrapping. In J. L. Morgan & K. Demuth (Eds.), *Signal to syntax* (pp. 263-283). Hillsdale, NJ: Lawrence Erlbaum Associates.

Pine, J. M., & Lieven, E. (1993). Reanalysing rote-learned phrases: Individual differences in the transition to multi-word speech. *Journal of Child Language, 20*, 551-571.

Redington, M., Chater, N., & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science, 22*(4), 425-469.

Shi, R., Werker, J. F., & Morgan, J. L. (1999). Newborn infants' sensitivity to perceptual cues to lexical and grammatical words. *Cognition, 72*, B11-B21.

Tomasello, M. (2006). Acquiring linguistic constructions. In D. Kuhn & R. Siegler (Eds.), *Handbook of Child Psychology*. New York: Wiley.

Valian, V. & Coulson, S. (1988). Anchor points in language learning: the role of marker frequency. *Journal of Memory and Language, 27*, 71-86.

Wijnen, F. & Bol, G. (1993). The escape from the optional infinitive stage. In A. de Boer, J. de Jong & R. Landeweerd (Eds.) *Language and Cognition 3*, University of Groningen, Dept. of Linguistics.

When 'More' in Statistical Learning Means 'Less' in Language: Individual Differences in Predictive Processing of Adjacent Dependencies

Jennifer B. Misyak (jbm36@cornell.edu)
Morten H. Christiansen (christiansen@cornell.edu)
Department of Psychology, Cornell University, Ithaca, NY 14853 USA

Abstract

Although statistical learning (SL) is widely assumed to play a key role in language, few empirical studies aim to directly and systematically link variation across SL and language. In this study, we build on prior work linking differences in nonadjacent SL to on-line language, by examining individual-differences in *adjacent* SL. Experiment 1 documents the trajectory of adjacency learning and establishes an individual-differences index for statistical bigram learning. Experiment 2 probes for within-subjects associations between adjacent SL and on-line sentence processing in three different contexts (involving embedded subject-object relative-clauses, thematic fit constraints in reduced relative-clause ambiguities, and subject-verb agreement). The findings support the notion that proficient adjacency skills can lead to an over-attunement towards computing local statistics to the detriment of more efficient processing patterns for nonlocal language dependencies. Finally, the results are discussed in terms of questions regarding the proper relationship between adjacent and nonadjacent SL mechanisms.

Keywords: Predictive Dependencies; Sentence Processing; Bigrams; Serial Reaction Time; Artificial Grammar

Introduction

With the expansion of studies on statistical learning (SL) over the past decades, focus has intensified towards probing the potential role for probabilistic sequence learning capabilities in acquiring and using linguistic structure (e.g., Gómez, 2002; Saffran, 2001). A clearer understanding has in turn begun to crystallize about the ways in which SL mechanisms may underpin language across various levels of organization—phonetic, lexical, semantic, syntactic—and across differing timescales—phylogenetic, ontogenetic, and microsecond unfoldings. Largely missing from this picture, however, is empirical evidence that directly links language and SL abilities within the typical population.

There are, though, a few recent studies that address the issue of whether better statistical learners are indeed better processors of language. In a small-scale study of individual differences, Misyak and Christiansen (2007) observed that standard measures of SL performance are positively associated with comprehension accuracy for various sentence-types in natural language. Conway, Bauernschmidt, Huang and Pisoni (2010) reported that better SL performance correlates with better processing of perceptually-degraded speech in highly-predictive lexical contexts. Misyak, Christiansen and Tomblin (2010) found that more-skilled statistical learners of nonadjacent structure were also more adept at the on-line processing of long-distance dependencies in natural language. Thus far, these results would support the general assumption that SL and language processes are systematically interrelated, with positive correspondence in intraindividual variation across them. But is it always the case that greater SL is associated with better language functioning? Or, may excelling at one of these implicate poorer performance at the other?

Such ability-linked reversals in performances within a cognitive domain would not be unprecedented. As an example, bilingual individuals appear to possess more efficient 'inhibitory control' processes than their monolingual peers across a number of studies, which has usually been imputed in some manner to bilinguals' greater experience with 'control' processes for suppressing irrelevant information in the course of successfully using two languages (see Bialystok et al., 2004). However, in a negative priming paradigm where distractor locations that were supposed to be previously ignored became relevant for facilitating responses to a current trial (as they do for monolinguals), bilinguals are at a disadvantage in the cognitive control task, with decreases from a neutral baseline in performance accuracy (Trecanni et al., 2009). Analogously then, might there be natural language contexts in which superior SL skill also becomes disadvantageous?

One possibility is that a statistical learner may focus too much on computing certain statistics, while ignoring others, with repercussions for their linguistic processing. For example, language embodies predictive dependencies that can be broadly characterized as involving either *adjacent* or *nonadjacent* temporal relationships. Thus, a good adjacency learner might perform poorly on nonadjacent dependencies in language. Introducing a new task for documenting micro-level trajectories and individual differences in SL, Misyak et al. (2010) were able to link variation in *nonadjacent* SL positively to signature differences in reading time patterns for the complex *nonlocal* dependency structure of center-embedded object-relative clause sentences. However, this study raises a new set of questions, including ones that directly bear on the above hypothetical, namely: Does the timecourse of *adjacent* SL differ from that of nonadjacent SL? Can substantial differences in adjacent SL also be empirically related to on-line sentence processing? And if so, might this differ from the kinds of positive correlations observed for nonadjacency processing?

We investigated these questions by adapting the AGL-SRT paradigm from Misyak et al. (2010) to isolate the learning of adjacent dependencies. The task implements an artificial grammar (AG) within a modified two-choice serial reaction-time (SRT) layout, using auditory-visual sequence-

strings as input. Experiment 1 thus documents the group trajectory and range of individual differences for adjacency learning obtained from this task. A 'bigram index' reflecting individual differences in adjacency learning is then used to probe relationships to the processing patterns observed in our subsequent natural language experiment (Experiment 2).

Experiment 1: Statistical Learning of Adjacencies in the AGL-SRT Paradigm

The ability of humans to use adjacent statistical information has been demonstrated across various studies. As early as two months of age, humans can identify bigrams, or first-order adjacent pairs, from the co-occurrence frequencies of elements within a constrained temporal sequence (Kirkham, Slemmer & Johnson, 2002). Throughout later development and adulthood, humans can also use adjacent conditional probabilities to locate relevant constituent-boundaries in a continuous stream composed of nonwords, tones, visual elements, or nonlinguistic sounds (see Gebhart, Newport & Aslin, 2009, for a review). And further, both children and adults can learn adjacent predictive dependencies that signal the underlying phrase structure of an artificial language (Saffran, 2001).

Below, we adapt the biconditional grammar of Jamieson and Mewhort (2005) to examine adults' SL of bigrams. This grammar was chosen since it is defined by first-order transitions only, imposes no positional constraints on element placement, and generates strings of equal length. These merits thereby permit us to effectively isolate the learning of predictive adjacencies by our participants.

Method

Participants Thirty native English speakers from the Cornell undergraduate population (15 females; age: *M*=19.4, *SD*=0.8) were recruited for course credit.

Materials Participants observed sequences of auditory-visual strings generated by an eight-element grammar in which every element could be followed by one of only two other elements, with equal probability. Each string consisted of 4 elements, with adjacent probabilities between them as shown in Table 1.The nonwords (*jux, tam, hep, sig, nib, cav, biff,* and *lum)* were randomly assigned to the stimulus tokens (*a, b, c, d , e, f, g, h*) for each participant to avoid

Table 1: Transition probabilities for elements at positions *n* and *n* + 1 of a string, with *n* as an integer from (0, 4).

Element at *n*	Element at position *n* +1 of string							
	a	*b*	*c*	*d*	*e*	*f*	*g*	*h*
a	0	**.5**	**.5**	0	0	0	0	0
b	0	0	**.5**	**.5**	0	0	0	0
c	0	0	0	**.5**	**.5**	0	0	0
d	0	0	0	0	**.5**	**.5**	0	0
e	0	0	0	0	0	**.5**	**.5**	0
f	0	0	0	0	0	0	**.5**	**.5**
g	**.5**	0	0	0	0	0	0	**.5**
h	**.5**	**.5**	0	0	0	0	0	0

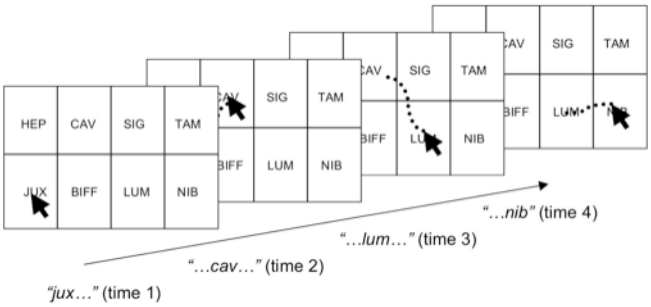


Figure 1: The pattern of mouse clicks for a single trial with the auditory target string *"jux cav lum nib."*

potential learning biases due to specific sound properties of words. Auditory versions of the nonwords were recorded from a female native English speaker and length-edited to 550 ms. Written versions of nonwords were presented with standard spelling in Arial font (all caps) and appeared within the rectangles of a 2 x 4 computer grid (see Figure 1). Each of the 4 columns of the computer grid, from left to right, displayed the nonword options corresponding to the 1st thru 4th respective elements of a string. Ungrammatical strings were created by introducing an incorrect element at the 2nd or 3rd string position, with the next element being one that legally followed the incorrect one (e.g., as in "*a *d e g*").

Procedure Each trial corresponded to a different configuration of the grid, with each of the eight written nonwords centered in one of the rectangles. Every column contained a nonword (target) from a stimulus string, as well as a foil. The first column contained the selection for the first element of a string, the second column contained the selection for the second element, and so on. For example, a trial with the stimulus string *jux cav lum nib*, as shown in Figure 1, might contain the target *jux* and the foil *hep* in the first column; the target *cav* and the foil *biff* in the second column; the target *lum* and the foil *sig* in the third column; and the target *nib* and the foil *tam* in the fourth column. Each nonword appeared equally often as target and as foil within and across the columns. The top/bottom locations of targets and foils were randomized and counterbalanced.

Participants were informed that the purpose of the grid was to display their selections and that a computer program randomly determines a target's location within either the top or bottom rectangle. On every trial, participants heard an auditory stimulus string composed of four nonwords and were instructed to respond to each nonword in the sequence as soon and as accurately as possible by using the computer mouse to select the rectangles displaying the correct targets.

Thus for any given trial, after 250 ms of familiarization to the visually presented nonwords, the first nonword of a string (the target) was played over headphones. Next, the second, third, and fourth words of a given string were each played after a participant had responded in turn to the prior nonword. For example, on a trial with the stimulus string *jux cav lum nib*, the participant should first click the rectangle containing JUX upon hearing *jux* (Fig. 1, left), CAV upon next hearing *cav* (Fig. 1, center-left), LUM upon hearing *lum* (Fig.

1, center-right), and NIB upon hearing *nib* (Fig. 1, right). After a participant had responded to the last nonword, the screen cleared for 750 ms before a new trial began.

An intended consequence of this design is that, for any given trial, the first element of a string cannot be anticipated in advance of hearing the auditory target. However, all subsequent string transitions might be reliably anticipated using statistical knowledge of the bigram structure. Thus, as participants become sensitive to the bigrams, they should be able to anticipate the string transitions, which should be evidenced by faster response times (following standard SRT rationale). Accordingly, our dependent measure on each trial was the reaction time (RT) for a predictive target, subtracted from the RT for the non-predictive initial-column target (which serves as a baseline and controls for practice effects). The predictive target used in this calculation was equally distributed across all non-initial columns across trials. Analogously, for an ungrammatical string trial, if participants are sensitive to the bigrams, then their RTs for incorrect, or violated, elements should be slower; thus, the DV for ungrammatical trials was the RT for the illegal target subtracted from the initial-target RT.

There are 64 unique strings (8 x 2 x 2 x 2) defined by the grammar; these were all randomly presented once each for each grammatical block of trials. Training consisted of six grammatical blocks, followed by an ungrammatical block of 16 trials and then a single grammatical ('recovery') block. Transitions across blocks were seamless and unannounced.

After these eight blocks, participants were informed that the strings had been generated according to rules specifying the ordering of nonwords and were asked to complete two tasks involving prediction and bigram recognition, respectively. The prediction task consisted of 16 trials that were procedurally similar to the trials observed during training, but with the omission of the auditory target for the final column.[1] Instead, participants were told to select that nonword in the final column that they believed best completed the sequence.

In the bigram task, participants were randomly presented with 32 test items of auditory nonword-pairs. They were requested to judge whether each pair followed the rules of the grammar by pressing 'yes'/'no' computer keys. Half of the test items were the 16 bigrams licensed by the grammar (e.g., *a b*); the remaining half were illegal pairings formed by reversing each bigram (e.g., *b a*). Thus, successful discrimination reflects knowledge of the conditional bigrams, rather than only sensitivity to co-occurrences.

Results and Discussion

Analyses were performed on only 'good' trials—that is, accurate string-trials with only one selection for each target.

[1] Instructing participants to complete string *endings* allows for maximal procedural similarity to the speeded training trials without introducing additional cue prompts that would be needed if the aurally-omitted element varied across non-initial columns. It also avoids any indirect feedback effects from presenting the next element after a participant's correct/incorrect medial selection.

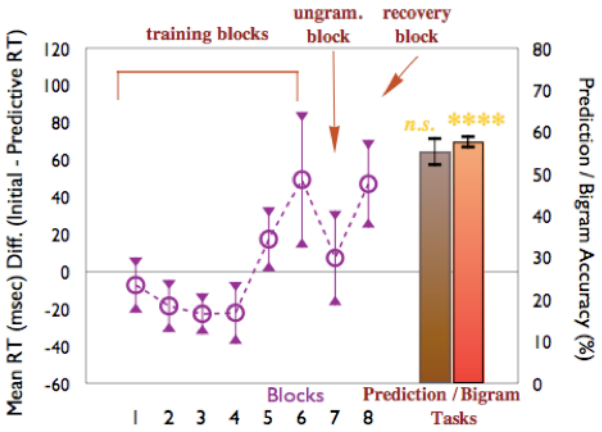


Figure 2: Group learning trajectory (mean RT difference scores per block) and accuracy for prediction (left bar) and bigram (right bar) tasks.

Prior to analysis, the data from five participants were omitted (2 for withdrawing participation; 2 for improperly performing the task, with less than 40% good trials; and 1 for abnormally elevated RTs, averaging in excess of 1470 ms per single response). For remaining participants, good trials averaged 88.2% (*SD*=5.9) of training block trials.

Mean RT difference scores, as described above (i.e., for grammatical trials: initial-target minus predictive-target RT; for ungrammatical trials: initial-target minus illegal-target RT) were computed for each block and submitted to a one-way repeated-measures analysis of variance (ANOVA) with block as the within-subjects factor. Since the assumption of sphericity was violated ($\chi^2(27) = 113.27$, $p < .001$), degrees of freedom were corrected using Greenhouse-Geisser estimates ($\varepsilon = .33$). Results indicated a main effect of block on RT difference scores, $F(2.31, 55.36) = 3.82$, $p = .02$. As seen in Figure 2, mean RT difference scores appear to increase by the final training block, decrease in the ungrammatical block, and increase once again in the recovery block. As RT difference scores measure the amount of facilitation from the predictive targets, an improvement in scores across blocks (as seen here) reflects sensitivity to the adjacent dependencies.

Planned contrasts between the ungrammatical block and preceding/succeeding grammatical blocks confirmed a performance decline for the ungrammatical trials (Block 6 minus Block 7: $M = -42.0$ ms, $SE = 19.6$, $t(24) = 2.14$, $p = .04$; Block 8 minus Block 7: $M = 39.8$ ms, $SE = 17.8$ ms, $t(24) = 2.23$, $p = .04$). This provides evidence for participants' learning of the sequential dependencies, consistent with standard interpretations in the sequence learning literature for comparing RTs to structured versus unstructured material (e.g., Thomas and Nelson, 2001).

Since the amount of exposure to the dependencies during training is equivalent to that which a similar number of participants ($n = 30$) received in the Misyak et al. (2010) study of nonadjacent SL, this invites a comparison of group learning trajectories. The RT timecourse pattern documented here for adjacent SL is very similar to that observed for nonadjacent SL, but with greater variance in

performance for the final training block and with ostensibly more modest (albeit not statistically different) performance in the recovery block. In both cases, sensitivity to the statistical structure does not show signs of emerging until after considerable exposure (the 5th block of training).

Mean accuracy on the prediction task was 55.3% (*SD*=17), which was not above chance ($t(24) = 1.51$, $p = .14$)—despite 20% of participants scoring at or above 75%. However, accuracy on the bigram task reflected adjacency learning ($t(24) = 4.66$, $p < .0001$), with a mean of 57.6%. This performance level is consistent with participants' judgment accuracy in an AGL study with manipulations of this same type of grammar when participants are tested with ungrammatical items containing few rule violations (Jamieson & Mewhort, 2009). Bigram scores further ranged from 37.5 – 71.9%, but with less variance (*SD*=8) than that observed in the prediction task. In post-study questioning, only four participants disclosed that they had noticed any general pattern in the sequence but were unable to verbalize at least one instance of a bigram, suggesting that their performance in the bigram task was not the product of explicit recall or well-formulated meta-knowledge. Next, we use scores on this bigram index to assess whether and how variation in adjacent SL may be associated with differences in processing local and nonlocal language dependencies.

Experiment 2: Individual Differences in Language Processing and Statistical Learning

Sensitivity to both local and long-distance relationships is indispensable to processing natural language, and pervades basic aspects of our everyday sentence comprehension and production—such as those involved in relating the modified subject/object of a described action or state to the main event of a sentence (embedded relative clauses), in identifying whether someone is the recipient or doer of an action (agent-patient thematic roles), and in correctly linking subjects with their verbs (number agreement). The aim of Experiment 2 is to investigate whether predictive processing as exemplified by adjacent SL is empirically related to the on-line processing of such natural language contexts. Consider the following examples of the sentence-types that constitute the focus of the current experiment.

(1a-b) The reporter [that attacked the senator / that the senator attacked] admitted the error.
(2a-b) The [crook/cop] arrested by the detective was guilty of taking bribes.
(3a-b) The key to the [cabinet/cabinets] was rusty from many years of disuse.

In the first sentence example, the subject-relative (SR; 1a) and the object-relative (OR; 1b) versions differ with respect to the manner in which the embedded verb *attacked* relates to its object. This involves a more complex, backwards-tracking long-distance dependency (to the head-noun) for ORs. In prior studies using materials resembling those in (1a-b), greater processing difficulty is elicited at the main verb of ORs compared to that of SRs, with considerable individual differences in the magnitude of this effect (e.g., Wells, Christiansen, Race, Acheson & MacDonald, 2009).

Next, consider the sentence pair (2a-b), which is temporarily ambiguous between a main verb (MV) and a reduced relative (RR) clause interpretation. Its resolution is influenced by the constraint of thematic fit—the fit between the head noun phrase (*the crook* or *the cop*) and the verb-specific roles of the verb (*arrested)*. Given verb-specific conceptual knowledge, the reader knows that *cop* is a typical agent of *arrested*, whereas *crook* is a typical patient. Controlling for animacy, thematic fit functions as an immediately integrated constraint computed over the noun and adjacent verb—with its effect on RTs occurring in the subsequent agent NP region (McRae, Spivey-Knowlton & Tanenhaus, 1998). Thus, the second condition (2b) in which the initial noun is a typical agent for the adjacent verb will elicit greater processing difficulty for the RR interpretation than that for the corresponding patient condition (2a). For our purposes, this provides an example of sensitivity to a local relation relevant for on-line sentence processing.

Lastly, (3a-b) illustrate subject-verb number agreement. In English, it is required that a number-marked subject (*key*) agrees with the number-marking of its verb (*was*). This is the case irrespective of the numerical marking of any intervening material (e.g., *to the cabinet/s*), and individuals are sensitive to this fact during reading. When a sentence's head noun is singular, individuals read longer at the MV in a condition where the 'distracting' local noun (*cabinets*) mismatches in number (i.e., is plural) than in a condition where the local noun matches the head noun's number (i.e., is singular); shorter reading latencies are also found for the word after the verb in the match condition (Pearlmutter, Garnsey & Bock, 1999). Although subject-verb agreement may occur locally between adjacent constituents, materials in the literature (and here) have involved a nonlocal dependency created from interposing a prepositional phrase.

Method

Participants The same participants from Exp. 1 participated directly afterwards in this experiment for additional credit. Because the analyses reported below involve correlations with the bigram index from Exp. 1, data was omitted for those participants already excluded in Exp. 1 analyses and from three others (2 for bilingual status and 1 for declining to participate in the second task).

Materials There were four sentence lists, each consisting of 9 practice items, 60 experimental items, and 50 filler items. The experimental items were sentences drawn from previous studies of sentence processing: 20 subject-object relative clauses (SOR; Wells et al., 2009), 20 reduced relative ambiguities influenced by thematic fit (TF; McRae et al., 1998), and 20 subject-verb agreement transitives (S-V; Pearlmutter et al., 1999). A yes/no comprehension probe followed each item. Item conditions within sentence sets were counterbalanced across lists.

Procedure Each participant was randomly assigned to a list, whose items were presented in random order using a

a standard word-by-word, moving window, self-paced reading paradigm. Millisecond reading times (RTs) per word and accuracy were recorded for analyses.

Results and Discussion

Overall comprehension accuracy across participants was high, *M*= 87.4%, *SD*=7.6. RTs in excess of 2500 ms (0.2% of data) were removed, and remaining RTs were then length-adjusted for the number of characters in a word using a standard procedure (Ferreira & Clifton, 1986). Unless otherwise noted then, all RTs reported below for each of the sentence sets have been length-adjusted, with the same sentence regions examined as those in the original studies. RTs connected with relevant effects for each of the sets were then used to probe for associations with individuals' bigram scores from Experiment 1, as summarized below.

Subject-Object Relatives. Results replicated the main effect for clause-type at the MV from Wells et al. (2009), $F(1, 21) = 5.55$, $p = .03$. OR MVs were read reliably longer (91 ms) than SR MVs. However, there was no signification correlation between bigram scores and MV RTs for either SR ($r = .04$, $p = .85$) or OR ($r = -.16$, $p = .47$) sentences. Thus, differences in adjacent SL did not appear to directly map onto differences in processing long-distance dependencies in these relative clauses.

Thematic Fit. The influence of TF was replicated at the 2-word MV region (e.g., *was guilty*), $F(1, 21) = 6.42$, $p = .02$, albeit not at the directly preceding agent NP region.[2] Agent conditions were read 39 ms longer than patient conditions at the MV region. The correlation between bigram scores and unadjusted RTs at the MV of the 'congruent' patient condition was not significant ($r = .29$, $p = .19$); but for the 'incongruent' agent condition, the correlation reached marginal significance ($r = .40$, $p = .06$), with better adjacent statistical learners taking *longer* to read the disambiguating verb phrase. This suggests a tendency for greater bigram sensitivity (in adjacent SL) to negatively correspond with resolving nonlocal ambiguity when the local TF constraint provides an opposing bias to the RR clause interpretation.

Subject-Verb Agreement. A 34 ms effect of match (i.e., the difference between match and mismatch conditions) was obtained at the verb, $F(1, 21) = 31.28$, $p < .0001$, which replicated Pearlmutter et al.'s (1999) findings. There was a smaller effect of match (23 ms) at the post-verb region, $F(1, 21) = 4.48$, $p = .05$, which was also numerically present but not reliable in Pearlmutter et al. Additionally, the correlation between bigram scores and RTs was significant for the effect at the verb ($r = .51$, $p = .02$), with better bigram learning corresponding to a larger effect of match condition. To further examine differences in processing patterns according to SL status, a median-split was performed on bigram scores, establishing 57.8% as the cut-off for defining membership in either a "high" bigram (*n*=11, *M*= 63.9%, *SD*=4.0) or "low" bigram group (*n*=11, *M*= 51.4%, *SD*=5.8). Significant bigram-group differences emerged for the effect of match condition across regions (as shown in Figure 3). While the low-bigram group did not elicit a significant effect of match condition at either the verb or post-verb region ($p = .13$ and $p = .91$, respectively), the high-bigram group showed a clear effect in both regions (both p's $< .001$). As apparent in Fig. 3, the high-bigram group demonstrated greater sensitivity to the interference created by the locally mismatched marking of the noun in the prepositional phrase (which was irrelevant for computing agreement). Thus, the better adjacent SL of the high-bigram group was related to generally less efficient processing than that by their low-bigram peers of the long-distance dependency entailed by the initial noun and verb. Since bigram groups did not differ in comprehension accuracy for any sentence-types in the experimental sets (all p's $> .15$), nor fillers ($p = .83$), these RT patterns were not the result of a speed-accuracy tradeoff.

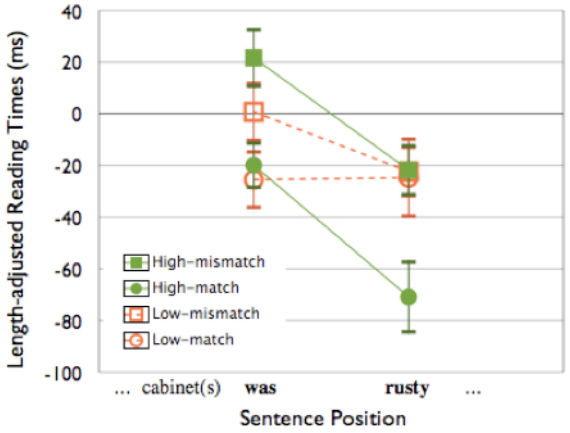


Figure 3: RT patterns on the S-V agreement sentences by bigram group (high/low) and condition (match/mismatch).

Our findings suggest that adjacent SL skill may not directly tap into the processes most relevant for handling long-distance dependencies in natural language—even though nonadjacent SL abilities appear to do so. Thus, while Misyak et al. (2010) reported a positive association between differences in nonadjacent SL and processing for the same SOR clauses as used here, no correlation was detected for adjacent SL. More generally, this is consistent with the lack of within-subjects correlation found between adjacent and nonadjacent SL in Misyak and Christiansen (2007).

However, while 'high' bigram learners may not differ from 'low' learners on processing long-distance relations as such, their increased sensitivity to local relations might interfere with the processing of the longer-distance elements within the sentence. This tendency is seen in the TF set, where above-average bigram tracking abilities seem to have a negative effect for processing the MV—the site where the initial, nonlocal ambiguity must be resolved. Similarly, too much sensitivity to local information is clearly evidenced within the last sentence set, where the irrelevant marking of an adjacent noun negatively affects better bigram learners' resolutions of S-V agreement, with protracted RTs also at the MV site of integrating the long-distance dependency.

[2] The later-occurring but nonetheless reliable effect of thematic fit is likely due to differences in the length of the moving window used in this study (1-word) and that by McRae et al. (2-word).

General Discussion

This study investigated the processing of adjacent predictive dependencies to address questions related to the timecourse of adjacent SL and the nature of any empirical association to natural language variation. While a learning trajectory similar to nonadjacent SL was documented in Exp. 1, findings from Exp. 2 indicated that above-average gains in adjacent SL performance do not necessarily translate to gains in language processing. Notably, those individuals who were strongly attuned to tracking statistical bigrams exhibited a *negative* pattern of correlations to tracking longer-distance aspects of language when either countervailing adjacent constraints or nearby distractive elements were present. This inverse pattern was not evidenced, though, when processing long-distance relations without conflicting local information (in the SOR clauses).

Instances where better bigram learners were worse language processors (or tended towards less efficient RT patterns) occurred when the integration of adjacent information (between a head-noun and part-participle verb) induced greater difficulty for resolving an ambiguity as a RR (the TF constraint in Exp. 2)—or when locally irrelevant information disrupted agreement computations between a nonlocal subject and verb (S-V agreement in Exp. 2). It would appear in these situations that those better in adjacent SL, although excelling at bigram pattern recognition in the SL task, are overly attuned to adjacency patterns and become more susceptible to local 'garden-paths'; in such cases, it may be the 'over-focus,' rather than any preexisting weakness in processing long-distance dependencies (as evidenced by parallel performance of groups in the SOR set) that hinders efficient resolution of nonlocal relationships.

This interpretation of our findings suggests that intraindividual differences in processing biases for the integration of competing constraints among adjacent- and nonadjacent dependencies may contribute to variation across SL-linked language processing skills. As such, it speaks to an open issue regarding whether different systems or different processing biases may be entailed by adjacent and nonadjacent processing capabilities in humans. It has been proposed, for instance, that the two forms of processing may be subserved by separate brain areas (Friederici et al., 2006), or that the two types of SL are only nominally distinct as the outcome of task-specific attention processes that may selectively hone in on adjacent or nonadjacent statistics (cf. Pacton & Perruchet, 2008). The findings here, of negative and specific associations between adjacent SL and aspects of language processing, suggest that future individual differences research incorporating careful attention to a diversity of natural dependency-structures may be needed to help establish the proper relation between these two manifestations of SL and the extent to which they may 'tap' into the same underlying mechanisms.

Acknowledgments

Thanks to Parry Cadwallader, Becky Fortgang and Stephan Spilkowitz for assistance with running participants.

References

Bialystok, E., Craik, F.I.M., Klein, R. & Viswanathan, M. (2004). Bilingualism, aging, and cognitive control: Evidence from the Simon task. *Psychology and Aging, 19*, 290-303.

Conway, C.M., Bauernschmidt, A., Huang, S.S. & Pisoni, D.B. (2010). Implicit statistical learning in language processing: Word predictability is the key. *Cognition, 114*, 356-371.

Ferreira, F. & Clifton, C. (1986). The independence of syntactic processing. *Journal of Memory and Language, 25*, 348-368.

Friederici, A.D., Bahlmann, J., Heim, S., Schibotz, R.I. & Anwander, A. (2006). The brain differentiates human and non-human grammars: Functional localization and structural connectivity. *Proceedings of the National Academy of Sciences, 103*, 2458-2463.

Gebhart, A.L., Newport, E.L. & Aslin, R.N. (2009). Statistical learning of adjacent and nonadjacent dependencies among nonlinguistic sounds. *Psychonomic Bulletin & Review, 16*, 486-490.

Gómez, R. (2002). Variability and detection of invariant structure. *Psychological Science*, *13*, 431-436.

Jamieson, R.K. & Mewhort, D.J.K. (2005). The influence of grammatical, local, and organizational redundancy on implicit learning: An analysis using information theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*, 9-23.

Jamieson, R.K. & Mewhort, D.J.K. (2009). Applying an exemplar model to the artificial-grammar task: Inferring grammaticality from similarity. *Quarterly Journal of Experimental Psychology, 62*, 550-575.

Kirkham, N.Z., Slemmer, J.A. & Johnson, S.P. (2002). Visual statistical learning in infancy: Evidence for a domain general learning mechanism. *Cognition, 83*, B35-B42.

McRae, K., Spivey-Knowlton, M.J. & Tanenhaus, M.K. (1998). Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language, 38*, 283-312.

Misyak, J.B. & Christiansen, M.H. (2007). Extending statistical learning farther and further: Long-distance dependencies, and individual differences in statistical learning and language. In *Proceedings of the 29th Annual Cognitive Science Society* (pp. 1307-1312). Austin, TX: Cognitive Science Society.

Misyak, J.B., Christiansen, M.H. & Tomblin, J.B. (2010). Sequential expectations: The role of prediction-based learning in language. *Topics in Cognitive Science, 2*, 138-153.

Pacton, S. & Perruchet, P. (2008). An attention-based associative account of adjacent and nonadjacent dependency learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*, 80-96.

Pearlmutter, N.J., Garnsey, S.M. & Bock, K. (1999). Agreement processes in sentence comprehension. *Journal of Memory and Language, 41*, 427-456.

Saffran, J.R. (2001). The use of predictive dependencies in language learning. *Jrnl of Memory and Language, 44,* 493-515.

Thomas, K.M. & Nelson, C.A. (2001). Serial reaction time learning in preschool- and school-age children. *Journal of Experimental Child Psychology, 79,* 364-387.

Treccani, B., Argyri, E., Sorace, A. & Della Sala, S. (2009). Spatial negative priming in bilingualism. *Psychonomic Bulletin & Review, 16*, 320-327.

Wells, J.B., Christiansen, M.H., Race, D.S., Acheson, D.J. & MacDonald, M.C. (2009). Experience and sentence processing: Statistical learning and relative clause comprehension. *Cognitive Psychology, 58*, 250-271.

Statistical Learning of Complex Questions

Hartmut Fitz (h.fitz@rug.nl)
Center for Language and Cognition Groningen, Oude Kijk in't Jatstraat 26
9712EK Groningen, the Netherlands

Abstract

The problem of auxiliary fronting in complex polar questions occupies a prominent position within the nature versus nurture controversy in language acquisition. We employ a model of statistical learning which uses sequential and semantic information to produce utterances from a bag of words. This linear learner is capable of generating grammatical questions without exposure to these structures in its training environment. We show that the model performs superior to *n*-gram learners on this task. Implications for nativist theories of language acquisition are discussed.

Keywords: Language acquisition; complex syntax; poverty of the stimulus; statistical learning; distributional information.

Introduction

It is a central question in language acquisition which aspects of our knowledge of language are learned from experience and which are part of our biological endowment for language. Nativist arguments often identify some property of a language and argue that it is not learnable from typical child-directed speech. By abductive reasoning, innate language-specific knowledge is offered as the best explanation of why children come to know this property regardless. The problem of auxiliary fronting in so-called complex polar questions (CPQ hereafter) is a key issue in this nature versus nurture debate.

According to Chomsky (1980), English yes/no-questions are formed from declaratives by displacing an auxiliary. The sentence "The man is happy" transforms into a question by subject-auxiliary inversion: "Is the man happy?". Declaratives with a relative clause can contain two identical auxiliaries as in "The man that is hungry is happy". Chomsky asked how children could learn that the main clause auxiliary should be placed in front, rather than the auxiliary which comes first. Only the former rule yields a grammatical CPQ.

(1) a. Is the man that is hungry happy?
b. *Is the man that hungry is happy?

He claimed that children have no basis in experience to adopt the correct rule since examples such as (1-a) do not occur in child-directed speech. In addition, children should adopt the rule which generates (1-b) because (i) it is supported by experience of simple yes/no-questions and (ii) the correct rule is "far more complex" in that it requires sensitivity to the hierarchical structure of a sentence. But children rarely, if ever, make mistakes as in (1-b) (Crain & Nakayama, 1987; Ambridge, Rowland, & Pine, 2008). They do not seem to generalize in a structure-independent way. To explain this error-free behavior, Chomsky postulated innate structure-dependent constraints on learning.

The above formulation of this poverty-of-the-stimulus argument makes a number of controversial assumptions. There is accreting evidence, for instance, that learning the syntax of questions does not involve learning movement rules (Dabrowska & Lieven, 2005; Estigarribia, 2009). An inadequate description of the learning target in terms of transformational rules might obscure empiricist solutions to the problem. Secondly, auxiliary fronting has been isolated from all the rest of language. Although there is some consensus that structures (1-a) are highly infrequent, the input environment of a child might provide other sources of indirect evidence for the correct rule (Pullum & Scholz, 2002). Another critical assumption is that the structure-independent rule (1-b) is simpler and *should* be preferred in the absence of innate constraints. Yet, if there is no reason to believe that children should overgeneralize there is no explanatory necessity for such constraints; the nativist argument would be preempted.[1]

Despite these reservations, it is clear that any theory of language acquisition which places more emphasis on the role of experience needs to explain how the syntax of complex questions can be acquired. Ideally, such an explanation demonstrates that a concrete, implemented learning mechanism built on justifiable assumptions can acquire auxiliary fronting from plausible input distributions.

Linear versus hierarchical models

Several models of language learning have recently been proposed which explicitly address the issue of auxiliary fronting. These models can roughly be divided into linear and hierarchical approaches. Linear models do not explicitly represent the hierarchical structure of a sentence's organization into phrases and clauses. All models briefly discussed here share the assumption that CPQs do not occur in child-directed speech, they learn solely from indirect evidence.

In the framework of data-oriented parsing, Bod (2009) showed that derivations of parse trees for grammatical CPQs are shorter (or more probable) than those for ungrammatical CPQs given the training data. The model assumes that subtrees are representational primitives in the mind of a human learner. Structure-dependent processing is built into the learning mechanism but it is still a question of linguistic experience whether the correct generalizations are supported in this model. Perfors, Tenenbaum, and Regier (2006) demonstrated that an ideal Bayesian learner favors a hierarchical over a linear grammar to fit a training corpus. This grammar could parse grammatical CPQs while the linear grammar could not. The model, however, did not strictly learn grammars from data, but rather selected one from a given set. How grammar selection bears on the process of child-language acquisition

[1] A more detailed discussion of the assumptions behind this nativist argument can be found in Fitz (2009).

needs to be elucidated. They argued that linear models have little to contribute to the auxiliary fronting debate because structure-dependent processing requires hierarchical representations. This assumption has been challenged by a number of linear approaches. If a linear model learning auxiliary fronting behaved in a manner consistent with structure-dependent processing, this would suggest that explicit representations of hierarchical structure might be superfluous. Clark and Eyraud (2006) proposed a linear alignment learner which substituted relativized NPs for simple NPs if they occurred in identical contexts in the corpus. As a result, the learner could generate grammatical CPQs if and only if their component clauses occurred in training. A simple recurrent network was used by Lewis and Elman (2001) to successfully learn CPQs from artificial language input. The scope of this approach is difficult to assess since the model seems to have been tested on a single item only. The most widely received linear approach used *n*-gram learners on untagged corpora of child-directed speech (Reali & Christiansen, 2005). The authors showed that a Bigram model could reliably classify pairs of grammatical/ungrammatical CPQs by assigning higher sentence probability to the former on 96% of the tested items. They suggested that indirect statistical information extracted from strings of words might be sufficient for children to infer the correct rule of auxiliary fronting. These results were scrutinized by Kam, Stoyneshka, Tornyova, Fodor, and Sakas (2008) who argued that the success of the Bigram model was largely due to a single distinguishing bigram which was supported by accidental phonological facts about English. When they added structural and lexical diversity to the test items, the model failed. Moreover, they argued that the bigram approach might not be valid cross-linguistically.

The Adjacency-Prominence learner

In our own work we aimed at showing that these difficulties could be overcome by a linear statistical model which in addition to *n*-gram based sequence learning uses meaning to constrain sentence production. The statistical information on which the learner draws had two components. The *adjacency* statistic was collected over bigrams in the training corpus. It measured how often two words which co-occurred in sentences, occurred adjacent to each other. The key addition over *n*-gram models was the *prominence* statistic. The learner tracked which words frequently preceded other words in the input environment. Words which on average were found earlier in a sentence than other words were considered more prominent. Using this statistic, a hierarchy was created which ordered words in an utterance in terms of their prominence. More prominent words then tended to be sequenced earlier in production. While the adjacency statistic selected words based on the previous word in an utterance, the prominence statistic selected words based on their prominence relation with remaining words in an utterance. This process is illustrated schematically in Figure 1. Both statistics were combined into the Adjacency-Prominence learner (AP-learner for short). This model of syntax learning was introduced in Chang, Lieven, and Tomasello (2008) where it was tested on a variety of typologically-distinct languages. Formal definitions of the two kinds of statistics are given in Table 1. Note that the adjacency statistic differs from forward transitional probabilities because bigram counts are normalized by the frequency of word pairs instead of the first unigram. The prominence statistic of a word is a sum over its prominence relation with other words. To give a comparable weight to the adjacency statistic, it was multiplied by the number of remaining words in an utterance.

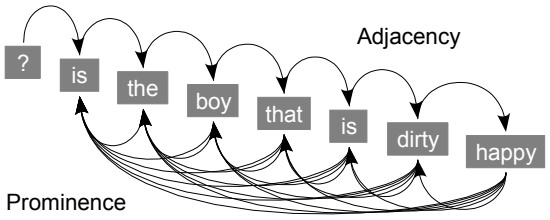


Figure 1: Adacency-prominence statistics for the CPQ *Is the boy that is dirty happy?* (adapted from Chang et al. (2008)).

Table 1: AP-learner statistics.

$C(w_{n-1}, w_n)$	Frequency of bigram $w_{n-1}w_n$
Pair(w_a, w_b)	Frequency of words w_a, w_b occurring together in the same sentence in any order
$P(w_a, w_b)$	Frequency of word w_a occurring before w_b in a sentence at any distance
Length	Number of words in bag-of-words
η	Balance parameter[2]
Adjacency	$\text{Adj}(w_n) = C(w_{n-1}, w_n)/\text{Pair}(w_{n-1}, w_n)$
Prominence	$\text{Pro}(w_n) = \sum_{w_b} P(w_n, w_b)/\text{Pair}(w_n, w_b)$ where w_b are all words in the bag (except w_n)
Adjacency-Prominence	$\text{AP}(w_n) = \text{Length} \times \text{Adj}(w_n) + \text{Pro}(w_n) \times \eta$

Evaluation

The performance of the AP-learner was evaluated in a sentence generation task. We assumed that speakers aim to produce utterances which express the meaning they intend to convey. To approximate constraints that meaning places on sentence production, a target utterance was split into an unordered bag-of-words. The learner then had to use its syntactic knowledge, extracted from the training corpus, to order this bag-of-words. Sentences were produced incrementally one word at a time. At each word position, all words in the bag were competing for the next slot in the utterance. The

[2]This parameter was used to calibrate the contribution of both statistics to word choice. It was held fixed across experiments.

learner could use forward probabilities from the preceding word (adjacency) but also the prominence ordering over the words left in the bag to predict the next word. The prominence value for a given word could dynamically change as the set of word options diminished during production.

Training and test items were identified as questions or declaratives by prepending a marker `quest`/`decl` to each sentence. Utterance generation was initialized by creating a bag-of-words including the marker for the target sentence. For each word in the bag, the adjacency and prominence statistic was collected and the word with the highest combined value was selected (see Table 1). The word was appended to the marker and removed from the bag-of-words. This procedure continued recursively until the bag was empty. The string of words produced by the learner was compared with the target utterance and its grammatical alternatives. For instance, the bag-of-words obtained from "Is the dog that is running happy?" also generated "Is the dog that is happy running?". If the learner produced either form, the sentence prediction accuracy count was incremented. Likewise, if either of the ungrammatical alternatives (with a displaced embedded clause auxiliary) was produced, the output was counted as a structure-independent generalization error.

Reali and Christiansen (2005) tested their *n*-gram learners in a grammaticality judgement task in which CPQs with lower cross-entropy were classified as grammatical. Our learner, in contrast, had to actually produce sentences from a bag-of-words and not merely classify them. Statistical information sufficient for classification might not be suitable for production. Chang et al. (2008) argued that bag-of-word generation is an adequate task to assess and compare statistical learners across languages.

The remainder of this paper is organized as follows. First, we demonstrate that the AP-learner can learn the syntax of complex questions in the absence of positive evidence and that overgeneralization does not occur. Then we compare the AP-learner with *n*-gram models and show that it performs superior. Finally, we identify conditions under which the AP-learner does make structure-independent errors. Such conditions arguably do not obtain in child-language acquisition. We conclude with a discussion of the results.

Method

Language input

The AP-learner was trained on an artificial English-like language with transitives and intransitives as basic construction types. From these constructions, simple declaratives, simple polar questions, complex declaratives, and polar questions with relative clauses could be generated (see Table 2). The language had number and noun-verb agreement, tense (past/present) and aspect (progressive/simple). Nouns could be animate and inanimate, or substituted by pronouns. Over a lexicon of 104 words and inflectional morphemes the language generated approximately 2.8×10^9 distinct sentences. It was suggested by Ambridge et al. (2008) that structure-independent generalizations such as

Table 2: Structures generated by the artificial language.

Sentence type	Example
Simple declarative	The guys buy it.
Simple polar question	Was the dog sleeping?
Complex declarative	A girl that is hitting him plays.
Complex polar question	Is a cat that is grumpy thirsty?

(2) Are the boys **that running** are eating?

may not occur in development because they violate word co-occurrence patterns in English (boldface bigram). In similar vein, Kam et al. (2008) argued that the good performance of the Reali and Christiansen (2005) model was due to these relative clause initial bigrams. To ensure that our learner could, in principle, generalize erroneously, we separated plural markers and inflectional morphemes for tense and aspect from the word stem. Thus sentence (2) was represented in our artificial language as

(3) Are the boy -s **that run** -ing are eat -ing?

The boldface bigram occurred frequently in the training corpus, for example in sentences such as "The boy -s that run are kick -ing the toy". This made it more difficult for the AP-learner to retain the embedded clause auxiliary in CPQs.

Results

Experiment 1

The first experiment tested whether the AP-learner was able to produce correct CPQs when trained only on simple declaratives, simple polar questions and declaratives with relative clauses. The learner was trained on 20.000 sentences randomly generated from the artificial language. 50% of these were simple sentences, the others were complex. 50% of the simple sentences were questions, the others were declaratives. Crucially, the training corpus did not contain any instance of a CPQ or any other question with a relative clause. Thus, it was tested whether the statistical information contained in the trained structures was sufficient for the AP-learner to generalize to the syntax of the novel CPQs. If so, this would support the idea that indirect evidence from frequent structures which are attested in child-directed speech might be sufficient to learn the correct subject-auxiliary inversion rule for complex polar questions.

The test set contained 40 CPQs randomly generated by the artificial grammar. All CPQs had an intransitive main clause. 20 had a center-embedded intransitive relative-clause (II), and 20 had a transitive relative-clause. Half of the transitive embeddings were subject-relativized (ITS), the other half were object-relativized (ITO). All tested CPQs were ambiguous in that the main clause auxiliary was identical with the embedded clause auxiliary. Auxiliaries could be singular or plural, past or present tense. Three actual test questions are listed

in Table 3. In contrast to the study of Reali and Christiansen

Table 3: Sample test questions.

Type	Example
II	Were the boy -s that were dirty play -ing ?
ITS	Was a brother that was push -ing them hungry ?
ITO	Is a cat that a boy is chase -ing jump -ing ?

(2005), the set of tested CPQs was structurally diverse (intransitive and transitive embeddings, subject- and object-relativized) and not limited to the auxiliary "is".

When evaluating the learner's output for ITS and ITO questions, only those grammatical alternatives were considered which preserved clause type and the grammatical role of the relativized constituent. For instance, when tested on ITOs, the learner's utterance had to have an intransitive main clause, and the transitive embedding had to have an object gap in order to count as an accurate production. The results of this experiment are shown in Figure 2.[3] The mean sentence pre-

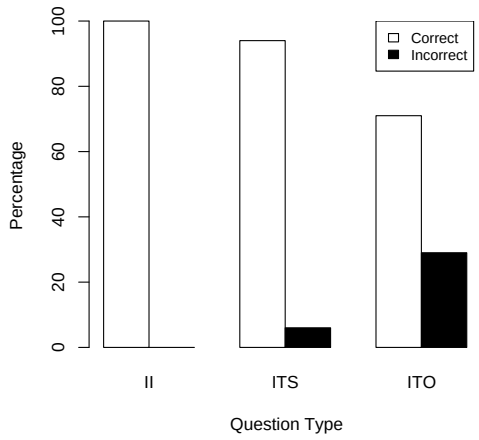


Figure 2: AP-learner tested on three kinds of CPQs.

diction accuracy was 91.25% versus 8.75% incorrect productions. On CPQs of type II, the AP-learner reached 100% accuracy. Slightly lower was the accuracy on ITS (94%) and ITO structures (71%). This difference between subject- and object-relativized transitives is consistent with developmental data on relative-clause acquisition in English-speaking children (Diessel & Tomasello, 2005). The AP-learner made mistakes on this task, it did not produce all test questions correctly. Importantly, however, none of the learner's incorrect productions matched an ungrammatical CPQ which would reflect structure-independent generalization. Although the AP-learner did not experience any instance of a CPQ in training, it correctly generalized the syntax of subject-auxiliary inversion from simple polar questions and declaratives with relative clauses to the formation of complex questions. When we added either ambiguous CPQs or CPQs with mixed number, tense and aspect (or both) to the training set, the learner's performance did not improve on any of the tested question types. These results suggest that the distributional information contained in simple polar questions and complex declaratives support the learning of structure-dependent generalizations even if the learner does not explicitly represent the hierarchical organization of CPQs into clauses and phrasal units. Since both these structures—simple questions and relative clause constructions—typically occur in child-directed speech, children might be exposed to sufficient indirect evidence to induce the syntax of auxiliary fronting in the absence of positive examples.

Experiment 2

In the previous experiment, the AP-learner showed differences in production accuracy between II, ITS and ITO questions. To trace the origin of differential performance, it was helpful to compare the AP-learner with Bi- and Trigram models of statistical learning. Both these models were trained, tested and evaluated in exactly the same way as the AP-learner. Figure 3 shows the prediction accuracy of the different models by CPQ type. All models displayed the same qualita-

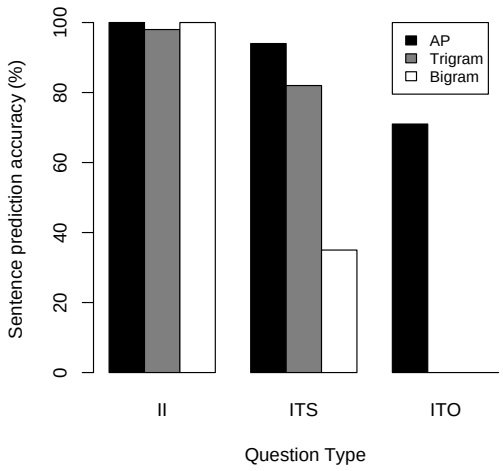


Figure 3: AP-learner in comparison with *n*-gram models.

tive behavior in that II questions were easier to produce than ITS, which were easier than ITO. Both *n*-gram models performed similar to the AP-learner on II questions. These CPQs were shorter than the other question types and thus had fewer choice points for prediction error. Moreover, ungrammatical II questions frequently contained word sequences which were not supported by the training corpus (e.g., "that happy"). The models followed a principle of non-monotonic learning to produce grammatical II questions: in the absence of evidence to the contrary, embedded clause auxiliaries should not be omitted. The Trigram model came close to the AP-learner on ITS questions (82%), whereas the Bigram model dropped below 40% accuracy. Errors made by the Bigram model mostly occurred sentence-initially (e.g., "`quest` Is chase"), whereas

[3] All modelling data reported here are averaged over 10 randomly generated training sets to ensure that results were robust with regard to the artificial language used to create input environments.

Trigram model errors mostly occurred in the relative clause (e.g., wrong verb type). The AP-learner was less vulnerable to these kinds of errors because it did not rely exclusively on co-occurrence frequencies. In addition to adjacency, the model could also use the prominence statistic which informed it that a subject should precede a verb form in the main clause and that a transitive verb should be produced in the relative clause (instead of an intransitive) when there was a direct object (e.g., a pronoun) left to sequence in the bag-of-words. Neither *n*-gram model produced any correct ITO question, whereas the AP-learner produced 71% correct ITOs. The Bigram model made the same errors as in ITS questions and sequenced a verb form after the initial auxiliary. The Trigram model often converted ITOs into grammatical ITS questions. The AP-learner also made such conversion errors, but less frequently. Again, the prominence statistic helped the model to place subject noun phrases before the verb form in transitive embeddings and this information was not available to the other models.

Kam et al. (2008) argued that Bigram models are not sufficient to learn the syntax of complex questions from noisy, realistic corpora. Our results support their findings for idealized input environments. The AP-learner was superior to both *n*-gram models when tested CPQs could not reliably be generated from a bag-of-words based on forward probabilities alone.

Experiment 3

As mentioned in the introduction, Chomsky's argument for the innateness of structure-dependent constraints on language learning has two prongs. Children have no basis in experience to infer the correct rule for auxiliary fronting, and they should overgeneralize by displacing the linearly-first auxiliary, as witnessed in simple polar questions in their language input. In Experiment 1, we found no evidence for either claim. The AP-learner could produce more than 90% grammatical CPQs without having experienced such structures in training. Although the model made some mistakes, it never produced ungrammatical CPQs in which the embedded clause auxiliary was omitted. In a third experiment we attempted to elicit overgeneralizations by creating input conditions which mislead the AP-learner into producing structure-independent errors. To do this, multiple word tokens were distinguished with markers in forward order of their occurrence within one sentence. Question (3), for instance, was now represented as

(4) are1 the1 boy1 -s1 that1 run1 -ing1 are2 eat1 -ing2 ?

After the model had produced a CPQ from a marked bag-of-words, the markers were removed and the output was compared with the equally unmarked target questions (grammatical and ungrammatical versions).

Distinguishing constituents in this way created clause-specific similarities between auxiliaries in different structures. The auxiliary are1 in test item (4) resembled the auxiliary in simple polar questions and the embedded clause auxiliary in complex declaratives from the training set. The auxiliary are2 resembled the main clause auxiliary in complex declaratives. These similarities were picked up by the adjacency-prominence statistics, as shown in Figure 4. Now the AP-

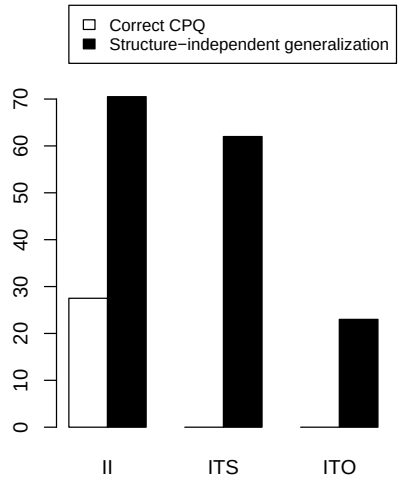


Figure 4: Structure-independent errors occurred when multiple auxiliaries were distinguished in the corpus.

learner produced only 13.75% correct CPQs. Out of the total incorrect CPQs, 65.5% were structure-independent errors in which the question-initial auxiliary was omitted from the relative clause rather than the main clause. Hence, the AP-learner could be forced to generalize erroneously when constituents were forward marked. Children, however, learn the syntax of questions from input which is not marked in this way. It is therefore not self-evident, as Chomsky suggests, that children should adopt the wrong auxiliary fronting hypothesis in the absence of innate constraints. In order to substantiate this claim, one would have to argue that children perceptually distinguish and track multiple auxiliary tokens in a way similar to the AP-learner in the above experiment. Unless this can be done convincingly, there is no reason to believe that children should overgeneralize. As a consequence, it is no longer puzzling that they in fact rarely do (Crain & Nakayama, 1987; Ambridge et al., 2008). Moreover, there is no need to posit innate constraints on learning as the best explanation of why they do not. One crucial premiss of the poverty-of-the-stimulus argument breaks away. Experiments 1 & 3, we believe, jointly shift the burden of proof back to those who claim that a biological endowment for structure-dependent processing is necessary to block overgeneralization.

Discussion and conclusions

Using a statistical model of syntactic development adapted from Chang et al. (2008), we demonstrated that the syntax of complex polar questions was learnable to a high degree of accuracy even when these structures were not present in the language input to the model. The tested questions were more diverse, both lexically (auxiliaries) and structurally (rel-

ative clause types), than the items used in Reali and Christiansen (2005) which may answer to some of the criticism posed by Kam et al. (2008). Our learner, however, was collecting more than *n*-gram statistics to accomplish this task. In addition to adjacency, it used a prominence ordering over words that were left to sequence. Words which were more prominent in sentences of the learner's experience were more accessible for production. Thus, the AP-learner was not relying on the presence (or absence) of particular bigrams to produce grammatical questions and it outperformed several *n*-gram models. Importantly, it was also shown that errors the learner made did not reflect structure-independent generalizations. To elicit these errors, the learning environment had to be manipulated such that it no longer resembled natural language input to children. This casts some doubt on the claim that children should overgeneralize in the absence of innate constraints.

On the downside, the AP-learner was trained on an artificial English-like language which did not exhibit the noisiness, diversity and distributional properties of child-directed speech. Our results should therefore be interpreted as a proof-of-concept that under idealized conditions a statistical learner which draws on sequential and semantic information can learn the syntax of complex polar questions from simpler and similar structures in the input. It remains to be tested whether this approach scales to real corpora and in particular whether it works for different languages which permit complex polar questions other than auxiliary-initial ones (Kam et al., 2008).

We do not suggest here that the adjacency-prominence statistic is all that is required to learn the syntax of complex questions. For one thing, the learner made mistakes where adults do not. The inclusion of meaning constraints (bag-of-words) into a statistical learning model was not sufficient to guarantee error-free learning or rule out the production of grammatical alternatives. Tighter semantic constraints and additional sources of information might be necessary.

Compared with other models which have previously been proposed to show the data-driven learnability of auxiliary fronting, the AP-learner did not make assumptions about the nature of syntactic representations in children, or the operations performed on such representations. The model learned from untagged raw text by means of simple, domain-general mechanisms and did not incorporate language-specific knowledge or biases. The model's task to produce rather than classify sentences is closer to experimental paradigms in developmental psychology than grammaticality judgement and incremental word prediction is consistent with current theories of language processing (Pickering & Garrod, 2007). Furthermore, the evaluation standard did not depend on language-specific assumptions about syntactic categories or on sentence probabilities which are difficult to interpret. Even though the AP-learner did not explicitly represent the hierarchical structure of complex questions or syntactic rules operating on such representations, it performed as if it respected the structure-dependence of auxiliary fronting. Thus, surface distributional information might be sufficient for a statistical learner to resolve the Chomskyan challenge.

Acknowledgments

Thanks to Franklin Chang for helpful discussions.

References

Ambridge, B., Rowland, C. E., & Pine, J. M. (2008). Is structure dependence an innate constraint? New experimental evidence from children's complex-question production. *Cognitive Science*, *32*, 222–255.

Bod, R. (2009). From exemplar to grammar: A probabilistic analogy-based model of language learning. *Cognitive Science*, *33*, 752–793.

Chang, F., Lieven, E., & Tomasello, M. (2008). Automatic evaluation of syntactic learners in typologically-different languages. *Cognitive Systems Research*, *9*, 198–213.

Chomsky, N. (1980). *Language and learning: The debate between Jean Piaget and Noam Chomsky* (M. Piattelli-Palmarini, Ed.). Cambridge, MA: Harvard University Press.

Clark, A., & Eyraud, R. (2006). Learning auxiliary fronting with grammatical inference. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society.* Vancouver, BC.

Crain, S., & Nakayama, M. (1987). Structure dependence in grammar formation. *Language*, *63*(3), 522–543.

Dabrowska, E., & Lieven, E. (2005). Towards a lexically specific grammar of children's question constructions. *Cognitive Linguistics*, *16*, 437–474.

Diessel, H., & Tomasello, M. (2005). A new look at the acquisition of relative clauses. *Language*, *81*(4), 882–906.

Estigarribia, B. (2009). Facilitation by variation: Right-to-left learning of English yes/no questions. *Cognitive Science*, *34*, 68–93.

Fitz, H. (2009). *Neural syntax*. ILLC dissertation series, University of Amsterdam.

Kam, X., Stoyneshka, I., Tornyova, L., Fodor, J., & Sakas, W. (2008). Bigrams and the richness of the stimulus. *Cognitive Science*, *32*, 771–787.

Lewis, J., & Elman, J. (2001). Learnability and the statistical structure of language: Poverty of stimulus arguments revisited. In *Proceedings of the 26th Annual Boston University Conference on Language Development.* Somerville, MA.

Perfors, A., Tenenbaum, J., & Regier, T. (2006). Poverty of the stimulus? A rational approach. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society.* Vancouver, BC.

Pickering, M. J., & Garrod, S. (2007). Do people use language production to make predictions during comprehension? *Trends in Cognitive Sciences*, *11*, 105–110.

Pullum, G. K., & Scholz, B. C. (2002). Empirical assessment of stimulus poverty arguments. *The Linguistic Review*, *19*, 9–50.

Reali, F., & Christiansen, M. H. (2005). Uncovering the richness of the stimulus: Structure dependence and indirect statistical evidence. *Cognitive Science*, *29*, 1007–1028.

Effector-specific Motor Interference in Action Simulation

Peggy Tausche, Anne Springer and Wolfgang Prinz
(Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany)

Abstract

Neuroscientific findings suggest that observing temporally occluded actions evokes a mental simulation of the occluded action part. This action simulation may involve corresponding motor programs in the observer and is suggested to run in real time. The present study aimed to investigate whether real-time action simulation relies on effector-specific motor representations. Our participants watched transiently occluded actions performed either with the arms or the legs and had to predict the action course after occlusion. Participants also responded to the task with a movement involving either their arms or legs. Simulation performance broke down when the observed effector and the moved effector corresponded. In contrast, simulation was intact when the effectors did not correspond. The results are in line with previous research and extend it by showing that interference effects can occur within the real-time course of action simulation. Furthermore, shared representations between action simulation and action execution are effector specific.

Introduction

In everyday life, humans experience hundreds of situations in which other people's actions are temporally or partially occluded. Nevertheless, observers perceive the actions in a fluent manner. It is suggested that humans fill the perceptual gap with a mental simulation of the unseen action parts. This action simulation implies the establishment of a mental representation of the unseen part that is equivalent to the visual representation during visual perception.

In the Common Coding framework, it has been argued that action execution and action perception share a common coding system (Prinz, 1990, 1997). This might enable observers to understand, anticipate and predict others' ongoing behavior (Blakemore & Frith, 2005; Prinz, 2006; Wilson & Knoblich, 2005; Wolpert & Flanagan, 2001). Behavioral studies supported this assumption by showing that concurrent action execution and action observation can interact with each other (Brass, Bekkering, & Prinz, 2001; Brass, Bekkering, Wohlschlager, & Prinz, 2000; Kilner, Paulignan, & Blakemore, 2003; Stürmer, Aschersleben, & Prinz, 2000). An influence of action observation on action execution was shown by Brass et al. (2000). They showed that the observation of a lifting movement of the index finger led to faster execution of a lifting movement with the index finger relative to the middle finger, even when the observed movement was irrelevant to the task. Other studies have investigated the influence of action execution on action perception (Daprati, Wriessnegger, & Lacquaniti, 2007a, 2007b; Jacobs & Shiffrar, 2005). Jacobs and Shiffrar (2005) showed that the ability to discriminate between two observed walking speeds is selectively impaired in walking observers as compared to cycling and standing observers. Taken together these findings propose a bi-directional link between action perception and action production.

While several studies support a bi-directional link between action perception and action production on the level of movements (Brass, et al., 2001; Brass, et al., 2000; Kilner, et al., 2003; Stürmer, et al., 2000), others were able to provide evidence for a link on the level of goals (Bekkering, Wohlschlager, & Gattis, 2000; Hamilton & Grafton, 2006; Woodward, 1998). This suggests that a common representational system of action execution and action observation might be hierarchically organized (1990).

Neurophysiologic findings support this idea by showing a different nature of the so called mirror neurons. These neurons are located in area F5 in the macaque monkey brain (Gallese, Fadiga, Fogassi, & Rizzolatti, 1996) and fire both when the monkey observes an action and when it performs this action on its own. Studies using functional magnet resonance imaging (fMRI) and transcranial magnetic stimulation (TMS) have provided significant evidence that such a mirror neuron system also exists in the human brain (Fadiga, Fogassi, Pavesi, & Rizzolatti, 1995; Grezes, Armony, Rowe, & Passingham, 2003; Rizzolatti & Craighero, 2004). Moreover, there is evidence that there are "strictly congruent" mirror neurons which fire only when the observed and the executed action correspond in means and goals. In contrast, "broadly congruent" neurons generalize across different means and goals (Gallese, et al., 1996).

Other studies have addressed the question of the underlying processes of action simulation itself (Graf, et al., 2007; Prinz & Rapinett, 2008). For instance, Graf and colleagues (2007) recently proposed that the internal simulation of observed actions runs in real-time. The authors used a paradigm in which the participants perceived temporally occluded sequences of point-light actions. In their studies they presented the beginning of an action sequence which was interrupted by an occluder. The occluder was followed by a static test posture. Two independent variables

were manipulated: occluder time (100, 400 or 700 ms) and test posture time (the time which would pass behind the occluder) (100, 400 or 700 ms). The participants' task was to decide whether the test posture was a continuation of the previous seen action in the same visual angle, or whether it was rotated in depth. In accordance with the real-time simulation hypothesis, the participants showed best performance when occluder and test posture time corresponded. Furthermore, performance decreased as the time distance between occluder and test posture time increased. The authors argued that the internal representation of the action is updated in real time and that this leads to high task performance when the upcoming test posture corresponds to a real-time outcome. Furthermore, task performance decreases with increasing dissimilarity in the internal representation and the test posture.

Our study connects to this work by investigating the role of motor representations in this action simulation process. In extension to previous research on motor interference in action observation and action discrimination (Daprati, et al., 2007a, 2007b; Jacobs & Shiffrar, 2005), we focussed on motor interference effects *within the real-time course* of action simulation (as proposed by Graf et al., 2007). Moreover, we wondered whether motor representations, which might be used in action simulation, are organized on an effector-specific level.

It has been suggested that humans have a long-term body representation which contains the basic spatial arrangement of different body parts (Reed & Farah, 1995). A structural overlap between one's own and another person's body enables humans to represent visual, motor and proprioceptive inputs from both bodies within a common code in a shared representational system which in turn would lead to more interactions between both processes. When a common code is used for one process (e.g., action execution), it is not or less available for the other process (e.g., action perception), which should lead to interference. Accordingly, we hypothesized that a structural overlap on the effector-specific level (i.e., the same effector is involved in action simulation and action execution) would lead to increased interference effects as compared to no structural overlap (i.e., different effectors are involved in action simulation and action execution).

In order to investigate this, we adopted the action prediction task used by Graf et al. (2007) and combined it with a secondary motor task. This motor task was performed simultaneously to the action prediction task and involved either the same effector as the relevant effector in the point-light action or a different effector.

Methods

Participants: Thirty right-handed participants (mean 25 years; range 20 – 35 years; 14 female) were tested. One participant's data had to be excluded from the analysis, because of a faulty response device which caused the loss of a part of the data set. Thus, data analysis was based on a total number of 29 participants. All participants reported normal or corrected-to-normal vision and were naive with respect to the purpose of the study. They were paid for their participation. Informed written consent was obtained from each participant prior to the experiment.

Material: We used six film sequences showing a point-light character (Johansson, 1973, 1975) performing familiar actions. These were three arm-related actions (tennis, throwing something with one hand, throwing something with both hands) and three leg-related actions (knee-bends, standing up from a chair, standing up from the floor). We chose actions which were rated on a visual analogue scale as being highly arm- or leg-related by an independent sample (N = 15). All actions of the present study were familiar everyday actions and all participants could easily recognize and name them. This is unique and contrasts with other studies using very simplistic and artificial movements (Brass, et al., 2000; Kilner, et al., 2003; Reed & Farah, 1995; Reed & McGoldrick, 2007), thus allowing us to investigate the involvement of effector-specific motor representations in action simulation of complex and familiar actions.

We used point-light stimuli (instead of real films), because these stimuli are known to emphasize motion information instead of alternative sources of information like social information. The videos were taken from a stimulus set provided by Graf et al. (2007) and showed a male right-handed agent recorded using a motion capture system (Vicon Motion Systems Ltd., Oxford, UK). Each point-light display consisted of 13 black dots that were located at the major joints and were 2 mm in diameter. The point-light character was about 7 cm in height and actions were performed within an area of 340 pixels width and 312 pixels height at the center of the screen. An occluder of the same size was presented as a square.

Design and Data Analysis: As in the original paradigm of Graf et al. (2007), we manipulated the factors occluder time (100, 400 and 700 ms) and test posture time (TPT; 100, 400 and 700 ms). A combination of each level of both factors resulted in a condition in which occluder time and TPT correspond (i.e., time distance of 0 ms) and conditions in which occluder time and TPT did not correspond (i.e., time distance of 300 ms and 600 ms, respectively). Participants had to decide whether the test posture was a continuation in the same visual angle, or whether it was rotated in depth. In accordance with Graf et al. (2007), we used

this task, because no explicit judgments about the timing of the actions were requested. Therefore subjects were explicitly instructed to decide whether the test posture was a correct or rotated continuation at any point in time which avoids that task instruction generates potential real-time effects. In order to investigate effector-specific interference effects in action simulation, we introduced a secondary motor task, which was either performed with the arms or with the legs. Participants were instructed to hold their hands/feet on two home buttons during the action sequence of the action prediction task and to perform a discrete bimanual/bipedal movement in order to provide a response to the action prediction task. The movement was a reaching movement towards two diagonally opposite buttons of an arrangement of four different target buttons (e.g., pressing the right upper key with the right hand and pressing the left lower key with the left hand simultaneously in order to give a "correct continuation" response) (Figure 1). The location of the target keys were randomized across participants. Participants were asked to respond immediately when the static test posture appeared and a time out for their response was set at 4000 ms (time out trials were excluded from data analysis).

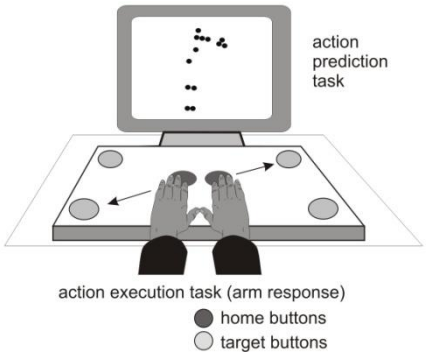


Figure 1: Schema of the experimental setting (exemplary for arm responses). The actions for the action prediction task were presented on the screen. The hands or feet rested on the home buttons (dark gray). The motor task involved a discrete bimanual/bipedal movement towards two diagonally opposite target keys in order to give a response to the action prediction task.

The experiment consisted of 648 trials (3 occluder times x 3 test posture times x 2 response devices [same, rotated] x 2 video effector [arm versus leg] x 2 response effector [arm versus leg] x 9 repetitions) divided into two experimental sessions with a break of one to two hours in between. Each session consisted of 324 trials divided into 12 blocks. The factor response effector (arm versus leg) was constant within one session. The order of the sessions was randomized across participants. The factor occluder time was blocked and the order of blocks was balanced across participants, with the restriction that two identical occluder times did not follow each other. The factors video effector (arm versus leg) and test posture time were completely randomized. Prior to the first session, participants received an initial familiarization phase where all actions were presented twice. This was followed by a practice phase containing different actions as in the experiment (knee-bends, leapfrog, basketball). The practice phase consisted of 30 trials and was performed with the same effector that was required in the first experimental session. Prior to the second session, a practice phase of 15 trials was performed again using the other effector, which was required in the second session. The experimental sessions lasted about 1.5 and 1 hour, respectively. Feedback was given to the participants during the practice and the experimental phases.

Data analysis focused on error rates and reaction times (RTs). RTs were defined as the time between TPT onset and leaving the home buttons. RTs were only analyzed for correct responses. Due to the fact that spatial and temporal aspects were mixed in the rotated trials, the analysis included only unrotated trials.

Our analyses were based on compatibility between the relevant effector in the action prediction task and the effector in the action execution task. Compatible trials were those trials in which the video effector and the response effector corresponded (arm/arm and leg/leg); incompatible trials were the trials in which the video effector and the response effector did not correspond (arm/leg and leg/arm). Compatibility was considered to be an adequate factor for the analysis because participants were required to predict exactly the same actions and to answer with their arms and legs in both compatible and incompatible trials. This allowed us to control for stimulus-dependent effects (due to variability within the point-light actions) and to control for response-dependent effects (due to variability between arm and leg responses), which are not in the center of interest in this study.

Results and Discussion

Error rates: We performed an analysis-of-variance (ANOVA) with the factors occluder time, test posture time (TPT) and compatibility. Error rates showed a significant main effect of TPT ($F(2, 56) = 38.395$; $p < .001$; $Eta^2 = .578$) with significantly higher error rates in the long TPT as compared to short and medium TPTs (ps < .001; Bonferroni corrected). No main effect of occluder time and no main effect of compatibility were found (Fs < .1). A significant two-way interaction between the factors occluder time and test posture time was found ($F(2, 112) = 2.835$; $p < .05$; $Eta^2 = .092$). In line with Graf et al. (2007), lowest error rates were found when occluder time and test posture time corresponded. No other two-way interaction reached significance (Fs < 1). Most importantly, the three-way interaction between the factors occluder time, TPT and

compatibility was significant ($F(4, 112) = 2.664$; $p < .05$; $Eta^2 = .087$). There was no significant occluder time x TPT interaction in the compatible condition ($F < 1.8$), while this interaction was reliable in the incompatible condition ($F(4, 112) = 3.660$; $p < .01$; $Eta^2 = .116$). This indicates that incompatible trials, in contrast to compatible trials, involved real-time action simulation processes.

A way of confirming this effect and to increase power is the analysis of the time distance effect. In a further step, we collapsed the different occluder times and TPTs across time distances (time distance of 0 ms, 300 ms and 600 ms, respectively) and performed an ANOVA with the factors time distance (0, 300 and 600 ms) and compatibility. Data showed a significant main effect of time distance ($F(2, 56) = 11.235$; $p < .001$; $Eta^2 = .286$). As put forward in the real-time hypothesis, error rates were significantly higher in the greatest time distance as compared to the short and medium time distance (ps < .001; Bonferroni corrected). Again, there was no main effect of compatibility ($F < .1$). Most interestingly, data showed a significant interaction between the factors time distance and compatibility $F(2, 56) = 5,050$; $p < .05$; $Eta^2 = .153$), with a significant main effect of time distance in the incompatible trials ($F(2, 56) = 13.461$; $p < .001$; $Eta^2 = .325$), while there was no reliable effect of time distance present in the compatible trials ($F < 1.9$) (Figure 2). Again, this indicates that incompatible trials, in contrast to compatible trials, involved real-time action simulation processes.

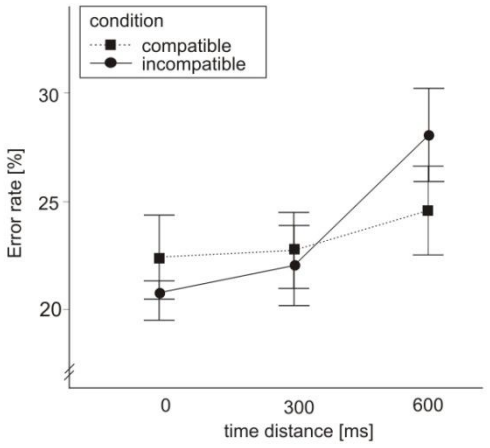


Figure 2: Error rates plotted as a function of time distance dependant on the compatibility between video effector and response effector. Error rates showed a significant time distance effect in incompatible trials, while no reliable time distance effect was present in compatible trials.

Reaction times: We performed an ANOVA with the factors occluder time, test posture time (TPT) and compatibility. Reaction times showed a significant main effect of occluder time ($F(2, 56) = 4.745$; $p < .05$; $Eta^2 = .145$), with significantly shorter RTs in the medium as compared to the short occluder time ($p < .05$; Bonferroni corrected). A significant main effect of TPT ($F(2, 56) = 23.389$; $p < .001$; $Eta^2 = .455$) was found with increasing RTs with increasing TPT (ps < .01; Bonferroni corrected). Furthermore, a significant main effect of compatibility was found ($F(1, 56) = 6.362$; $p < .05$; $Eta^2 = .185$), with shorter RTs in compatible as compared to incompatible trials. A significant two-way interaction between the factors occluder time and test posture time was found ($F(2, 112) = 6.568$; $p < .001$; $Eta^2 = .19$), with longest RTs when occluder time and test posture time did not correspond. Neither other two-way interactions nor the three-way interaction reached significance (Fs) < 1.

Again, we collapsed the different occluder times and TPTs across time distances (time distance of 0 ms, 300 ms and 600 ms, respectively) and performed an ANOVA with the factors time distance (0 ms, 300 ms and 600 ms) and compatibility. Data showed a significant main effect of time distance ($F(2, 56) = 5.337$; $p < .01$; $Eta^2 = .160$). RTs were significantly higher in the long time distance as compared to the short and medium time distance ($p < .05$; Bonferroni corrected). There was a main effect of compatibility ($F(1, 56) = 6.123$; $p < .05$; $Eta^2 = .179$), with significantly faster RTs in compatible as compared to incompatible trials. No significant time distance x compatibility interaction was found ($F < .4$) (Figure 3).

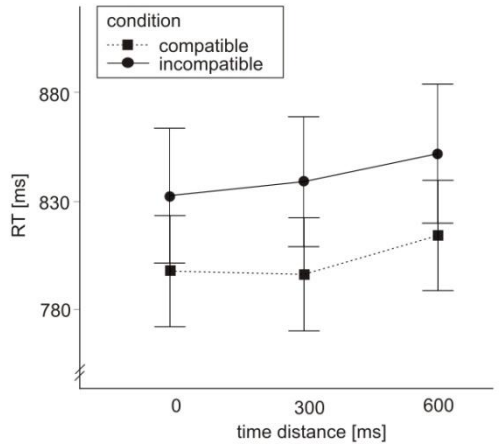


Figure 3: Reaction times plotted as a function of time distance dependant on the compatibility between video effector and response effector.

General Discussion

The present study aimed to investigate the role of motor representations in action simulation by focussing on motor interference within the real-time course of action simulation. Furthermore, it was investigated whether motor representations, which might be used in

real-time action simulation, are organized on an effector-specific level.

Overall, the results showed two major findings: First, our data showed that action simulation and action execution share a common representational system. Action simulation is considered to run in real time (Graf, et al., 2007), but we could show interference in the time course of action simulation in an online action prediction task. To our knowledge, our results are the first to demonstrate that a secondary motor task leads to an interference effect within the real-time course of action simulation. We assume that the preparation of the to-be-executed action takes resources of the same representational system as action simulation, which in turn leads to a lack of resources which might be necessary for an action simulation to run in real time.

Second, we were successfully able to show that this motor interference effect was effector-specific. That is, real-time action simulation broke down when the action prediction task and the action execution task involve the same type of effector as compared to a different type of effector, although the predicted and the executed actions differed in terms of the kind of action, the exact trajectory and action goals. This allows us to specify that shared representations are coded on an effector-specific level and that they can generalize across different kinds of actions, trajectories and goals. This finding is in line with other studies (Reed & Farah, 1995; Reed & McGoldrick, 2007). For example, Reed and McGoldrick (2007) showed that the task performance in a body posture memory task is selectively impaired when a concurrent movement task is applied that involves the same type effector as the body posture memory task as compared to a different type of effector. In this study the only structural overlap regards the effector while the kind of the action, the trajectory and the goals differs between both tasks. The idea that shared representations might be organized hierarchically is also supported by imaging studies showing that parts of the mirror neuron system are organized in a somatotopic pattern which resembles the classical motor homunculus (Buccino, et al., 2001; Buccino, et al., 2004). In line with the common coding framework (cf. Introduction), these results suggests the existence of a hierarchically organized matching system of action observation and action execution.

As mentioned above, real-time simulation was no longer applied in trials, in which the relevant effector in the action prediction task and in the action execution task corresponded. Nevertheless, it is reasonable to assume that real-time action simulation was replaced by another process because task accuracy was comparable between compatible and incompatible trials. Possible candidates of processes are either the memorizing of the arrangement of certain points of the point-light display and matching them onto the test posture which was presented after the occluder or memorizing the test postures and the according feedback and applying a memory process without any simulation process. Although we cannot make any assumptions about the type of process which was applied in compatible trials, we can completely rule out a real-time simulation process. One could speculate that real-time simulation is the default process and that a blocking of a common coding system by a secondary motor task leads to a breakdown of such a process and requires the application of an alternative process.

However, the motor interference effect in real-time action simulation was statistically reliable only in error rates. There are several reasons which might account for that fact. First, we used quite a demanding task. Participants' error rates were relatively high (about 23 percent). This is in contrast to other studies that involve very simplistic and easy tasks showing a compatibility effect in the RTs, which could show floor effects in error rates (Brass, et al., 2001; Brass, et al., 2000). Second, we used a decision task. It is likely that this requires higher cognitive processes in order to reach the decision rather than a reaction towards a certain stimulus. An indication of this is the fact that RTs were much longer (average of 820 ms) than RTs for simple reactions towards a certain stimulus (about 300 ms) (Brass, et al., 2001; Brass, et al., 2000). Third, although we instructed our participants to respond as fast and as accurately as they could, it is possible that they focused more on task accuracy. They had a quite long time in which to give a response (time out of 4000 ms) and participants received feedback on the basis of accuracy while no explicit feedback was given regarding speed. It is likely that this caused participants to focus on task accuracy rather than task speed, which in turn lead to a visible effect of compatibility on action simulation in error rates.

In conclusion, this study demonstrates for the first time motor interference *within the real-time course* of action simulation. This indicates that real-time action simulation of temporally occluded actions and action execution share a common representational system. Preparation for action execution leads to the activation of these shared representations, which in turn leads to a lack of representations for action simulation. This, in turn, causes interference in the real-time cause of action simulation. Finally, we were successful in showing that this representational system is specific on the level of effectors, even when the actions differ in terms of the kind of the movement, trajectories and goals.

Acknowledgments

We are thankful to Wiebke Berger and Christina Schuster for assistance in data acquisition and to Nadine Diersch for fruitful discussion. Special thanks to the Max-Planck-Society for founding this research.

References

Bekkering, H., Wohlschlager, A., & Gattis, M. (2000). Imitation of gestures in children is goal-directed. *Quarterly Journal of Experimental Psychology. A, Human Experimental Psychology, 53*(1), 153-164.

Blakemore, S. J., & Frith, C. (2005). The role of motor contagion in the prediction of action. *Neuropsychologia, 43*(2), 260-267.

Brass, M., Bekkering, H., & Prinz, W. (2001). Movement observation affects movement execution in a simple response task. *Acta Psychologica, 106*(1-2), 3-22.

Brass, M., Bekkering, H., Wohlschlager, A., & Prinz, W. (2000). Compatibility between observed and executed finger movements: comparing symbolic, spatial, and imitative cues. *Brain and Cognition, 44*(2), 124-143.

Buccino, G., Binkofski, F., Fink, G. R., Fadiga, L., Fogassi, L., Gallese, V., et al. (2001). Action observation activates premotor and parietal areas in a somatotopic manner: an fMRI study. *European Journal of Neuroscience, 13*(2), 400-404.

Buccino, G., Lui, F., Canessa, N., Patteri, I., Lagravinese, G., Benuzzi, F., et al. (2004). Neural circuits involved in the recognition of actions performed by nonconspecifics: an FMRI study. *Journal of Cognitive Neuroscience, 16*(1), 114-126.

Daprati, E., Wriessnegger, S., & Lacquaniti, F. (2007a). Kinematic cues and recognition of self-generated actions. *Experimental Brain Research, 177*(1), 31-44.

Daprati, E., Wriessnegger, S., & Lacquaniti, F. (2007b). Knowledge of one's kinematics improves perceptual discrimination. *Consciousness and Cognition, 16*(1), 178-188.

Fadiga, L., Fogassi, L., Pavesi, G., & Rizzolatti, G. (1995). Motor facilitation during action observation: a magnetic stimulation study. *Journal of Neurophysiology, 73*(6), 2608-2611.

Gallese, V., Fadiga, L., Fogassi, L., & Rizzolatti, G. (1996). Action recognition in the premotor cortex. *Brain, 119 (Pt 2)*, 593-609.

Graf, M., Reitzner, B., Corves, C., Casile, A., Giese, M., & Prinz, W. (2007). Predicting point-light actions in real-time. *Neuroimage, 36 Suppl 2*, T22-32.

Grezes, J., Armony, J. L., Rowe, J., & Passingham, R. E. (2003). Activations related to "mirror" and "canonical" neurones in the human brain: an fMRI study. *Neuroimage, 18*(4), 928-937.

Hamilton, A. F. D., & Grafton, S. T. (2006). Goal representation in human anterior intraparietal sulcus. *Journal of Neuroscience, 26*(4), 1133-1137.

Jacobs, A., & Shiffrar, M. (2005). Walking perception by walking observers. *Journal of Experimental Psychology: Human Perception and Performance, 31*(1), 157-169.

Johansson, G. (1973). Visual-Perception of Biological Motion and a Model for Its Analysis. *Perception & Psychophysics, 14*(2), 201-211.

Johansson, G. (1975). Visual Motion Perception. *Scientific American, 232*(6), 76-&.

Keele, S. W., Cohen, A., & Ivry, R. (1990). Motor Programs - Concepts and Issues. *Attention and Performance*(13), 77-110.

Kilner, J. M., Paulignan, Y., & Blakemore, S. J. (2003). An interference effect of observed biological movement on action. *Current Biology, 13*(6), 522-525.

Prinz, W. (1990). A common coding approach to perception and action. In O. Neumann & W. Prinz (Eds.), *Relationships between perception and action: Current approaches* (pp. 167-201). New York: Springer.

Prinz, W. (1997). Perception and action planning. *European Journal of Cognitive Psychology, 9*, 129-154.

Prinz, W. (2006). What re-enactment earns us. *Cortex, 42*(4), 515-517.

Prinz, W., & Rapinett, G. (2008). Filling the Gap: Dynamic Representation of Occluded Action. In F. Morganti, A. Carassa & G. Riva (Eds.), *Enacting Intersubjectivity: A Cognitive and Social Perspective on the Study of Interactions.* (pp. 223-236). Amsterdam: IOS Press.

Reed, C. L., & Farah, M. J. (1995). The psychological reality of the body schema: a test with normal participants. *Journal of Experimental Psychology: Human Perception and Performance, 21*(2), 334-343.

Reed, C. L., & McGoldrick, J. E. (2007). Action during body perception: processing time affects self-other correspondences. *Social neuroscience, 2*(2), 134-149.

Rizzolatti, G., & Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience, 27*, 169-192.

Stürmer, B., Aschersleben, G., & Prinz, W. (2000). Correspondence effects with manual gestures and postures: a study of imitation. *Journal of Experimental Psychology: Human Perception and Performance, 26*(6), 1746-1759.

Wilson, M., & Knoblich, G. (2005). The case for motor involvement in perceiving conspecifics. *Psychological Bulletin, 131*(3), 460-473.

Wolpert, D. M., & Flanagan, J. R. (2001). Motor prediction. *Current Biology, 11*(18), R729-732.

Woodward, A. L. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition, 69*(1), 1-34.

Simulation from Schematics: Dorsal Stream Processing and the Perception of Implied Motion

Kevin J. Holmes (kevin.holmes@emory.edu)
Phillip Wolff (pwolff@emory.edu)
Department of Psychology, Emory University
36 Eagle Row, Atlanta, GA 30322

Abstract

Schematic language (e.g., prepositions) and depictions (e.g., line drawings) reduce the rich detail of the visual world to a coarser level of description. We investigated how these schematic forms may be represented in the brain. Recent neural evidence suggests that such representations may be computed in the dorsal pathway of the visual system, the same pathway involved in processing motion, including simulated motion in static scenes. Drawing on this association, we examined the stimulus conditions and mental sets that give rise to simulation, and by hypothesis, representations in the dorsal stream. Simulated motion was evident for scenes that were highly schematic, as opposed to highly realistic (Experiment 1), and when realistic scenes were processed schematically (Experiment 2). The results suggest that dorsal stream representations capture the schematic aspects of visual experience, rather than more fine-grained information. In affording simulation, these representations may facilitate certain types of reasoning and inference.

Keywords: schematic representations; mental simulation; dorsal stream; implied motion; word meaning.

Introduction

In physics and engineering textbooks, simple line drawings are often used to illustrate complex physical phenomena. These drawings tend to be highly schematic, representing idealized examples of the processes in question. Schematic depictions of this sort may be useful not only because of their visual simplicity, but also because they have a fundamental cognitive basis. In particular, they may map onto mental representations that are themselves schematic in nature and that may afford certain perceptual and cognitive advantages over representations that more veridically capture the rich detail of the visual world. In this research, we investigate the nature of these hypothesized schematic representations and how they might be realized in the brain.

A distinction between representations that are more detailed or featural and those that are more schematic or configural has been proposed to underlie the meanings of words. Landau and Jackendoff (1993) argued that the representations associated with the meanings of object nouns, which encode detailed featural information, differ from those associated with the meanings of prepositions, which encode coarser configural properties. Moreover, they hypothesized that these different types of representations are computed in different processing pathways in the brain. A highly influential model originally proposed by Ungerleider and Mishkin (1982) points to two separate streams for the processing of visual information: a ventral stream, responsible for the identification of objects on the basis of visual properties such as shape, size, color, and texture (the "what" system), and a dorsal stream, responsible for the localization of objects in space (the "where" system). Landau and Jackendoff proposed that the meanings of object nouns are processed in the "what" system and the meanings of prepositions in the "where" system.

While several of Landau and Jackendoff's (1993) conjectures have been supported by subsequent neural research, recent work suggests that the dichotomy between object nouns and prepositions may not adequately capture processing differences in the two streams. Beyond localizing objects in space, the dorsal stream appears to be responsible for certain aspects of object perception. For example, several areas of the dorsal stream are activated during the passive viewing of objects. The caudal part of the intraparietal sulcus (CIP) shows sensitivity to the shapes of objects even when their location is unspecified (Grefkes & Fink, 2005). Similarly, activity in the V5/MT complex has been linked to differences in the shapes of objects in static images (Chandrasekaran et al., 2006). These findings suggest that the ventral stream is not the only pathway in which objects are processed; the dorsal stream is also sensitive to certain object properties, notably shape.

Nonetheless, the two streams appear to differ in the level of abstraction at which they process objects. Whereas the dorsal pathway is primarily concerned with identifying the principal axes, surfaces, and dimensionality of an object, the ventral pathway fills in featural details such as size, color, and texture (Farivar, 2009). Consistent with this characterization of the two streams, Lehky and Sereno (2006) observed that neurons in the dorsal area LIP were sensitive to shape but less able to differentiate shapes than neurons in the ventral area AIT (see also Chandrasekaran et al., 2006). These findings suggest that the ventral stream makes fine-level distinctions, while dorsal stream processing is at a coarser, more schematic level.

Intriguingly, the dorsal stream is also invoked in the perception of implied motion; that is, the kind of motion suggested by frozen-action photographs or speed lines in cartoons. In an imaging study, Kourtzi and Kanwisher (2000; see also Senior et al., 2000) observed activation in V5/MT in response to still photographs of agents or objects in motion (e.g., an athlete about to throw a discus). These findings suggest a way in which dorsal stream processing might be examined behaviorally. When people perceive implied motion from a static scene, it is highly likely that

they are processing the scene in the dorsal stream. Hence, the perception of implied motion can be used as an index of dorsal stream processing, and by hypothesis, of the schematic representations that support such processing.

A necessary condition for taking advantage of this association is to find a way to measure the perception of implied motion. An experimental paradigm developed by Freyd, Pantzer, and Cheng (1988) offers such a measure. In Freyd et al.'s study, participants were presented with a line drawing of a scene depicting a potted plant supported by a pedestal. The scene was then replaced by one in which the pedestal was removed, but the plant was in exactly the same position as it had been previously. This second scene was then replaced with a third scene in which the plant's position was shifted slightly (higher or lower) or remained the same. The participants' task was to indicate whether the plant in the third display was in the same position as in the second. Freyd et al. reasoned that if people viewed the pedestal as exerting a force on the pot, they might (implicitly) expect the plant to move downward due to the influence of gravity. As predicted, participants were more likely to report "same" to a downward shift than an upward one. These results support the hypothesis that motion will sometimes be perceived when a force acting on an object is suddenly removed. This phenomenon of implied motion from disequilibrium is one of several types of *displacement*, in which the mental representation of a target's location is displaced in the direction of (implied) target motion (see Hubbard, 2005, for a review).

Predictions. Based on subsequent neural research, it is highly likely that the implied motion perceived by participants in Freyd et al.'s (1988) study involved processing in the dorsal stream (in particular, area V5/MT). If so, it should be possible to modulate displacement by varying the properties of the visual stimulus. Lobmaier et al. (2008) employed this technique in an fMRI study of face processing, observing greater dorsal (V5/MT) activation to blurred faces (which preserved configural information) than to scrambled faces (which disrupted configural information but preserved detailed featural information) and greater ventral activation to scrambled than to blurred faces. Thus, changing the properties of the visual stimulus changed which pathway was primarily used to process the stimulus.

The findings of Lobmaier et al. (2008) suggest that the perception of implied motion in static scenes will be more pronounced when stimuli are highly schematic, as opposed to highly realistic. Highly schematic stimuli are more likely to be processed in the dorsal stream than in the ventral stream; processing in the dorsal stream should produce larger effects of implied motion, and hence a stronger displacement effect. If this initial prediction is supported, we might find that displacement can be modulated in other ways as well. In particular, it might be possible to influence how a stimulus is processed by varying the observer's mental set. Because relational words like verbs and prepositions encode the world in a relatively schematic fashion, describing a scene by using a high proportion of such words (as opposed to words that encode featural information, such as adjectives) should engage the dorsal stream and result in greater displacement. Drawing a scene might also modulate one's mental set, with more schematic drawings leading to greater displacement. We tested these predictions in the following two experiments.

Experiment 1

In our first experiment, we investigated whether implied motion would be perceived in scenes that varied in realism. We contrasted realistic scenes that resembled photographs with schematic scenes that resembled line drawings, similar to those used by Freyd et al. (1988). Our prediction was that the schematic scenes would engage the dorsal stream more than the realistic scenes, and hence that there would be greater displacement for the schematic scenes than for the realistic ones. Following Freyd et al., we also varied whether the initial picture in the sequence showed a support relation (e.g., a pedestal supporting a plant vs. a plant floating in mid-air), in order to confirm that displacement was due to the perceived removal of a force rather than some perceptual bias to infer that unsupported objects will move downward. Thus, we predicted that displacement would be more likely when the initial picture depicted a support relation than when it did not.

Method

Participants. Fifty-nine Emory University undergraduates received course credit for participating in the experiment.

Materials. We created a set of materials based on the scenes shown in Figure 1. The scenes depicted a room either rich in photorealistic detail (Realistic format) or schematically sketched, as in a line drawing or diagram (Schematic format). The Schematic scene was a contoured rendering of the Realistic scene, with all fine detail removed so that only the basic outline of the objects was visible. All other aspects of the two display formats were identical. Each display was 27.3 cm x 15.7 cm (45.5° x 28.9° visual angle).



Figure 1: The Realistic (top) and Schematic (bottom) support displays used in Experiment 1.

There were four variants of each display format. In the original version shown in Figure 1, a potted plant (height: 2.3 cm / 4.3°) is supported by a marble pedestal at the center of the room (*support* display). In the other three versions, the pedestal was removed and the plant was either in exactly the same position (*no-support* display), slightly raised (*up* display), or slightly lowered (*down* display). In the latter two displays, the plant was 0.15 cm (0.3°) higher or lower, respectively, than its original position. All displays were created using a graphics package called Discreet 3D Studio Max, version 7.

Design and Procedure. Participants were randomly assigned to either the Realistic or Schematic display format and to either the Support or No Support trial type, in a fully crossed between-subjects design with four conditions: Realistic-Support, Realistic-No Support, Schematic-Support, and Schematic-No Support. Figure 2 depicts the trial structure. In the Support conditions, each trial began with the presentation of the support display, which remained on the screen for 250 ms. Following a 250-ms interstimulus interval (ISI), the no-support display appeared for 250 ms. Another 250-ms ISI was followed by one of three test displays: no-support (showing the plant in the same position as it had been previously), up, or down. The test display remained on the screen until participants made a response. The No Support conditions were identical, except that the first stimulus of each trial was the no-support display.

As in Freyd et al. (1988), participants were asked to indicate whether the plant in the test display was in the same position as it had been in the previous (no-support) display. They were instructed to press the 'S' key for *same* and the 'D' key for *different*. The instructions emphasized both speed and accuracy. Participants were also told that they should not expect an equal number of same and different trials, and that they should process the entire display rather than the plant alone. There were a total of 60 randomly ordered trials, 20 with each test display.

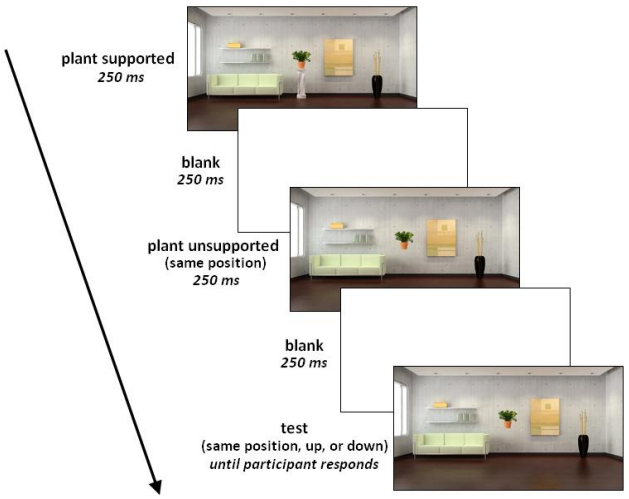


Figure 2: The structure of individual trials, shown with stimuli from the Realistic-Support condition of Experiment 1 and all conditions of Experiment 2.

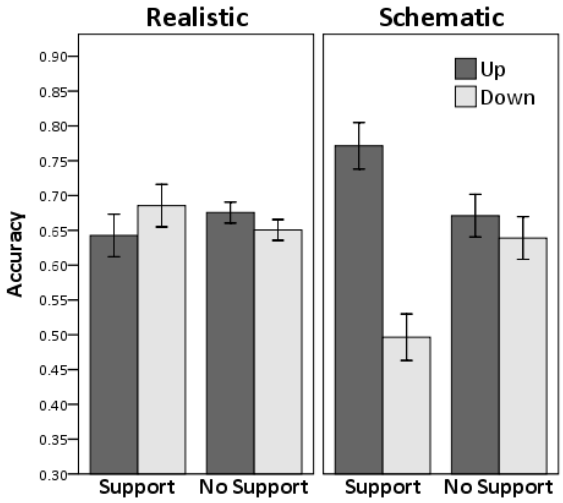


Figure 3: Accuracy on up and down trials across conditions in Experiment 1 (error bars are +/- 1 SEM).

Results

The main finding was that displacement occurred only for schematic scenes that depicted an initial support relation. As shown in Figure 3, participants in the Schematic-Support condition were more likely to indicate "same" when the plant was shifted down than when it was shifted up. No such asymmetry was observed in the other three conditions.

These findings were supported by a mixed ANOVA on participants' accuracy patterns in which format (realistic vs. schematic) and support (initial display showed vs. did not show a support relation) were between-subjects factors and target position (up vs. down) was a within-subjects factor. [The data of 3 participants were excluded from analyses for making *same* responses on greater than 75% of the trials, leaving 14 participants in each condition.] There was a significant main effect of target position [$F(1,52) = 7.01$, $p < .02$], with accuracy lower for down trials ($M = 62\%$) than for up trials ($M = 70\%$). However, this asymmetry between up and down depended on both format and support, as shown by a significant interaction between target position and format [$F(1,52) = 8.78$, $p < .005$] and a significant three-way interaction [$F(1,52) = 7.01$, $p < .02$]. Accuracy was significantly lower for down than for up trials only in the Schematic-Support condition (up: $M = 77\%$, down: $M = 50\%$), $t(13) = 4.11$, $p < .005$. There was no asymmetry in the other three conditions, and no other main effects or interactions were significant (all ps > .2).[1]

Discussion

The results of Experiment 1 replicate the findings of Freyd et al. (1988) in confirming that people simulate

[1] The RT data showed the same general patterns as the accuracy data across both experiments, though some analyses did not reach statistical significance. In this paradigm, as noted by Freyd et al. (1988), there are often too few correct responses to calculate a reliable RT for some trial types (e.g., down trials in the Schematic-Support condition of Experiment 1).

motion in static scenes only when there is perceived removal of a force. However, the results also highlight an important caveat to this conclusion. The simulation processes associated with the perception of implied motion are engaged more when visual stimuli are schematic, as opposed to realistic. We suggest that displacement varied as a function of realism because the properties of the schematic materials reflected the kinds of representations that are hypothesized to exist in the dorsal stream to a greater extent than did the properties of the realistic materials.

Although there was no evidence of mental simulation in the Realistic conditions, this does not imply that realistic materials cannot lead to the simulation of motion. The materials in the Realistic conditions consisted of certain features (e.g., color, texture) that could be processed only in the ventral stream, but they also included features that could be processed in the dorsal stream (e.g., shape). Because schematic language (e.g., prepositions) and depictions (e.g., line drawings) reflect a relatively coarse level of description, activities that promote the use of such forms might induce a more schematic conceptualization of experience. If sufficiently biased through such activities, people might focus on the schematic aspects of realistic materials, in which case even realistic materials might lead to the perception of implied motion.[2] This possibility was examined in the next experiment.

Experiment 2

Experiment 2 examined whether a prior task prompting people to focus on the schematic properties of a realistic stimulus might induce greater simulated motion. Participants completed the same task as in Experiment 1, but this time they were shown only the realistic stimuli. Prior to this task, participants engaged in activities designed to vary the mental set they used when subsequently processing the realistic scene. Half of the participants were asked to describe the scene in writing, while the other half were asked to draw the scene. Within each of these groups, half of the participants were asked to describe or draw the scene in a realistic manner, while the other half were asked to describe or draw the scene in a schematic manner. The key prediction was that schematic processing, whether induced by describing or drawing, would engage the dorsal stream to a greater extent, and hence lead to greater displacement, than would realistic processing.

Method

Participants. Seventy-nine Emory University undergraduates participated in the experiment as part of a course requirement.

Materials, Design, and Procedure. The materials included the same photorealistic stimuli used in Experiment 1. Participants were randomly assigned to either the Describe or Draw condition and to either the Realistic or Schematic format in a fully crossed between-subjects design with four conditions: Describe-Realistic, Describe-Schematic, Draw-Realistic, and Draw-Schematic.

In all conditions, participants were shown the support display, in which the plant is supported by the pedestal. In the Describe-Realistic condition, participants were asked to describe the room "in rich detail, as if describing the details of a photograph." In the Describe-Schematic condition, participants were asked to describe the room "schematically, as if describing the details of a diagram." Similarly, in the Draw-Realistic condition, participants were asked to depict the room "in rich detail, as if your drawing were a photograph," whereas in the Draw-Schematic condition, they were asked to depict the room "schematically, as if your drawing were a diagram." Participants were given 5 minutes to describe or draw the room. Then they completed the implied motion task using the materials from the Realistic-Support condition of Experiment 1 (see Figure 2).

Results

The results showed that varying the mental set of the observer modulated the perception of implied motion. Displacement was observed for realistic scenes when a prior task induced participants to process the scenes schematically, but not when the task induced them to process the scenes realistically.

These findings were supported by a mixed ANOVA on participants' accuracy patterns. [The data of 7 participants were excluded from analyses for making *same* responses on greater than 75% of the trials, leaving 18 participants in each condition.] There was a significant main effect of target position [$F(1,68) = 7.51$, $p < .01$], with lower accuracy for down trials ($M = 62\%$) than for up trials ($M = 72\%$), just as would be expected if participants were simulating downward motion. A significant interaction between target position and format [$F(1,68) = 5.13$, $p < .03$] showed that the asymmetry between down and up trials was larger in the Schematic conditions than in the Realistic conditions. Within the Schematic conditions (collapsing across Describe and Draw), accuracy on down trials ($M = 61\%$) was significantly lower than on up trials ($M = 78\%$), $t(35) = 3.64$, $p < .001$. Within the Realistic conditions, the difference in accuracy between down ($M = 63\%$) and up ($M = 65\%$) trials was not significant ($p > .7$). No other main effects or interactions were significant (all ps $> .09$).

The lack of a three-way interaction between target position, format, and medium [$F(1,68) = 1.15$, $p > .2$] suggests that the down-up asymmetry for the Schematic format (relative to the Realistic format) was comparable in both the Describe and Draw conditions. However, the Schematic format showed a greater asymmetry than the Realistic format only for participants who had produced drawings, $F(1,34) = 6.12$, $p < .02$ (see Fig. 4). The difference between the two formats was not significant for participants who had written descriptions ($p > .4$).

Post-hoc analysis indicated that the magnitude of displacement correlated positively with the proportion of relational terms (prepositions and verbs describing spatial relations) in participants' descriptions ($r = .45$, $p < .01$), but

[2] This prediction is consistent with findings showing that displacement can be influenced by variables such as observers' conceptual knowledge and expectations (see Hubbard, 2005).

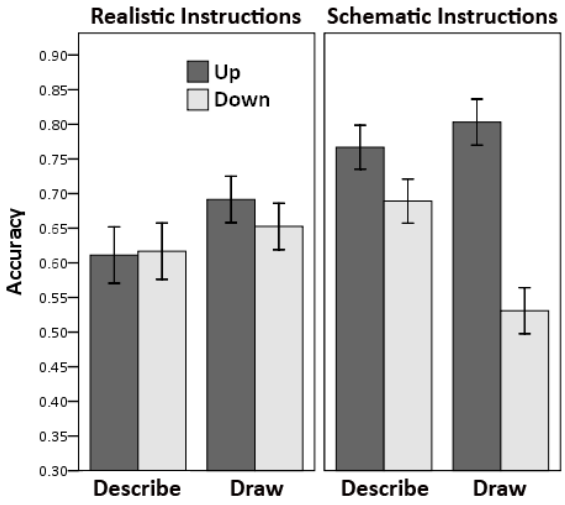


Figure 4: Accuracy on up and down trials across conditions in Experiment 2 (error bars are +/- 1 SEM).

did not correlate with the proportion of adjectives (r = -.26, p > .1). In addition, descriptions from the Describe-Schematic condition had a significantly higher proportion of relational terms [$t(35)$ = 3.02, p < .005] and a marginally lower proportion of adjectives [$t(35)$ = 1.74, p = .09] than descriptions from the Describe-Realistic condition. Ratings of participants' drawings (by a separate group, N = 15) on a 1-to-9 Likert scale of "realism," defined as the extent to which a drawing included cues to 3D properties such as depth and texture, were also collected. On average, raters assigned significantly higher realism ratings to drawings from the Draw-Realistic condition (M = 5.0) than drawings from the Draw-Schematic condition (M = 4.7), $t(14)$ = 2.32, p < .04. Thus, participants who showed greater displacement were those who had used more schematic language or produced more schematic drawings.

Discussion

The results of Experiment 2 provide further support for the idea that implied motion is more likely to be perceived when a scene is conceptualized in a schematic fashion. When conceptualized schematically, the scene may be processed primarily in the dorsal stream, which is largely responsible for the mental simulation of motion. While we found clear effects of drawing on simulation, the effects of verbal description were less compelling. However, an association between simulation and relatively schematic aspects of language in participants' descriptions suggests that verbal description can in fact modulate processing. In particular, the positive correlation between relational terms and the displacement effect is exactly what would be predicted if relational language leads people to process visual stimuli in a schematic fashion, presumably in the dorsal stream. Further, the lack of correlation with adjectives is not surprising, as adjectives encode information presumably processed in the ventral stream.

In sum, the results suggest that everyday activities such as writing and drawing can direct attention to different aspects of visual stimuli and influence how they are processed. Schematic processing may cause the visual world to be represented more like a line drawing than a photograph, and this format of representation may invoke simulation processes in the dorsal stream.

General Discussion

The results from this research suggest that the mental simulation of motion in static scenes depends on the realism of the scenes and the observer's mental set when processing them. Experiment 1 showed that simulation occurred during the processing of highly schematic scenes resembling line drawings, but not highly realistic scenes resembling photographs. Experiment 2 showed that simulation can occur even for highly realistic scenes when they are processed schematically; that is, when prior activities induce the observer to focus on their schematic properties. Because the simulation of motion is strongly associated with processing in the dorsal visual pathway, the conditions under which implied motion is perceived offer a window into the kinds of representations associated with dorsal stream processing. Consistent with previous evidence indicating that the dorsal stream operates at a relatively coarse level in the perception of objects, our findings are suggestive of a format of representation in which the rough contour of objects and the spatial relations among them are preserved, but detailed featural information is lacking. The sparseness of such representations, much like the line drawings in physics textbooks, may be especially suited for the mental operations at work in the simulation of motion.

This link between schematic representations and simulation highlights the potential utility of such representations for reasoning. In particular, reasoning about physical systems sometimes involves forming a mental image of a system and then "running" it (Hegarty, 2004). For example, when solving problems involving interlocking sequences of gears, people often mentally rotate the gears before discovering the abstract rule that governs how they turn, namely that odd and even gears turn in different directions (Schwartz & Black, 1996). Our findings suggest that more schematically rendered or imagined gears may be easier to mentally rotate, which could influence the tendency to re-represent the problem in terms of a rule. Thus, the use of schematic representations may be beneficial for certain types of problem solving and inference.

One key question concerns exactly what visual properties constitute a "schematic" representation, as opposed to a "realistic" one. In future work, we plan to employ the same behavioral paradigm used in the present experiments to specify which aspects of visual stimuli give rise to simulation, and hence reflect properties of schematic representations in the dorsal stream. If, for example, displacement is minimized or eliminated when visual properties such as depth cues or surface gradients are absent, it would imply that schematic representations include such information. Similarly, if displacement persists even when the stimuli are primitive 3D shapes (e.g., spheres, cylinders), it would imply that schematic representations need not have any shape detail beyond simple geometric forms.

Although we found no evidence of simulated motion with realistic materials under neutral conditions, other studies (e.g., Kourtzi & Kanwisher, 2000; Senior et al., 2000) have used realistic materials specifically to identify the neural correlates of simulated motion. However, these studies used single static stimuli in which motion was strongly implied (e.g., frozen-action photographs), whereas our stimuli invoked more subtle forms of motion (slight changes in spatial position) solely through the sequential nature of their presentation. Our findings suggest that the use of schematic stimuli in the former paradigm might lead to even greater simulated motion. Interestingly, displacement effects in a handful of studies using realistic stimuli have been regarded as validating the widespread use of more impoverished stimuli (Hubbard, 2005), but to our knowledge, the current study is the first to manipulate realism directly. Our findings caution against the assumption that simulation for schematic materials will carry over to more ecologically rich contexts.

Together with recent neural work, our findings have implications for models of the neural bases of word meaning. While Landau and Jackendoff (1993) argued that the dorsal and ventral streams map onto different grammatical categories (preposition vs. noun), it is likely that certain aspects of the meanings of object nouns are represented in the dorsal stream as well. Processing differences in the two streams may be better accounted for by a distinction often made in lexical semantics between structural and idiosyncratic aspects of word meaning (Levin & Rappaport Hovav, 2009). Words for spatial relations, for example, can be divided into a structural component, which specifies the abstract geometry of a spatial relation, and a more idiosyncratic component, which distinguishes spatial terms on the basis of more fine-grained geometric information. We suggest that schematic representations computed in the dorsal stream may reflect structural components of word meaning.

Our findings also suggest a novel perspective on the interface between language and thought (Wolff & Malt, 2010). Recent research has focused on how language might augment thought by putting in place representational systems essential for certain kinds of abstract thinking (e.g., reasoning about exact quantities; Gordon, 2004; see Wolff & Holmes, in press, for a review). In our second experiment, more schematic language was associated with greater simulation, suggesting instead that language may serve as a vehicle to abstraction, promoting the use of schematic representations rather than directly instantiating them. Importantly, however, language may be just one of many vehicles to abstraction. Other types of processing (e.g., drawing) may be just as likely to induce a schematic conceptualization of experience. Thus, it may be the schematic representations themselves, rather than the means by which they are recruited, that offer especially powerful tools for thinking.

Acknowledgments

The authors wish to thank Larry Barsalou, Stella Lourenco, Laura Namy, Marjorie Pak, and Grace Song for helpful discussion, and Tonia Davis, Savina Nikolova, Seho Park, Sam Ritter, and Meredith West for assistance with data collection. This research was supported by a William Orr Dingwall Foundation Neurolinguistics Fellowship to KJH.

References

Chandrasekaran, C., Canon, V., Dahmen, J. C., Kourtzi, Z., & Welchman, A. E. (2006). Neural correlates of disparity-defined shape discrimination in the human brain. *Journal of Neurophysiology*, *97*, 1553-1565.

Farivar, R. (2009). Dorsal-ventral integration in object recognition. *Brain Research Reviews*, *61*, 144-153.

Freyd, J. F., Pantzer, T. M., & Cheng, J. L. (1988). Representing statics as forces in equilibrium. *Journal of Experimental Psychology: General*, *117*, 395-407.

Grefkes, C., & Fink, G. R. (2005). The functional organization of the intraparietal sulcus in humans and monkeys. *Journal of Anatomy*, *207*, 3-17.

Gordon, P. (2004). Numerical cognition without words: Evidence from Amazonia. *Science*, *306*, 496-499.

Hegarty, M. (2004). Mechanical reasoning by mental simulation. *TRENDS in Cognitive Sciences*, *8*, 280-285.

Hubbard, T. L. (2005). Representational momentum and related displacements in spatial memory: A review of the findings. *Psychonomic Bulletin & Review*, *12*, 822-851.

Kourtzi, Z., & Kanwisher, N. (2000). Activation in human MT/MST by static images with implied motion. *Journal of Cognitive Neuroscience*, *12*, 48-55.

Landau, B., & Jackendoff, R. (1993). "What" and "where" in spatial language and spatial cognition. *Behavioral and Brain Sciences*, *16*, 217-265.

Lehky, S. R., & Sereno, A. B. (2007). Comparison of shape encoding in primate dorsal and ventral visual pathways. *Journal of Neurophysiology*, *97*, 307-319.

Levin, B., & Rappaport Hovav, M. (2009). Lexical conceptual structure. In K. von Heusinger, C. Maienborn, & P. Portner (Eds.), *Semantics: An international handbook of natural language meaning*. Berlin: Mouton de Gruyter.

Lobmaier, J. S., Klaver, P., Loenneker, T., Martin, E., & Mast, F. W. (2008). Featural and configural face processing strategies: Evidence from a functional magnetic resonance imaging study. *NeuroReport*, *19*, 287-291.

Schwartz, D. L., & Black, J. B. (1996). Shuttling between depictive models and abstract rules: Induction and fallback. *Cognitive Science*, *20*, 457-497.

Senior, C., Barnes, J., Giampietro, V., Simmons, A., Bullmore, E. T., Brammer, M. et al. (2000). The functional neuroanatomy of implicit-motion perception or 'representational momentum'. *Current Biology*, *10*, 16-22.

Ungerleider, L. G., & Mishkin, M. (1982). Two cortical visual systems. In J. Ingle, M. A. Goodale, & R. J. W. Mansfield (Eds.), *Analysis of Visual Behavior*. Cambridge, MA: MIT Press.

Wolff, P., & Holmes, K. J. (in press). Linguistic relativity. *Wiley Interdisciplinary Reviews: Cognitive Science*.

Wolff, P., & Malt, B. C. (2010). The language-thought interface: An introduction. In B. C. Malt & P. Wolff (Eds.), *Words and the mind: How words capture human experience*. New York: Oxford University Press.

Assessing Behavioral and Computational Approaches to Naturalistic Action Segmentation

Meredith Meyer[1] (mermeyer@umich.edu), Philip DeCamp[2] (decamp@media.mit.edu), Bridgette Hard[3] (martin@psych.stanford.edu), Dare Baldwin[4] (baldwin@uoregon.edu), Deb Roy[2] (dkroy@media.mit.edu)

[1]Department of Psychology, University of Michigan, Ann Arbor, MI 48103 USA
[2]Media Lab, Massachusetts Institute of Technology, Cambridge, MA 02139 USA
[3]Department of Psychology, Stanford University, Stanford, CA 94305 USA
[4]Department of Psychology, University of Oregon, Eugene, OR 97403 USA

Abstract

Recognizing where one action ends and another begins is an automatic and seemingly effortless process that supports understanding of goal-directed action. One characteristic of such action segmentation is that it is hierarchical; it reflects the goals and sub-goals of an actor, which correspond to coarse- and fine-grained action units respectively. We report on the success of one method of assessing hierarchical segmentation of naturalistic footage taken from an extensive corpus of unscripted human action (Speechome project, e.g., Roy et al., 2006). Results indicate that hierarchical segmentation occurs in an on-line fashion, with event boundaries marked by surges in attention that are modulated based on whether a boundary marks a fine, intermediate, or coarse unit. We also describe a method by which objective changes in an actor's movement can be measured and analyzed as a predictor of participants' segmentation behaviors.

Keywords: action segmentation; event processing

Drawing inferences and generating predictions about others' actions are processes most people undertake every day. The ways in which people use such inferences and predictions to make sense of others' action is supported in part by the ability to segment continuous action into discrete units. For instance, while observing an individual preparing dinner, we might identify and recognize individual units of action such as chopping a carrot, opening a refrigerator, or rinsing off a dish. Investigations of action segmentation have suggested that people are highly consistent in where they judge event boundaries to exist; people typically report dynamic human action to consist of units corresponding to initiation or completion of goals, with considerable agreement across individuals regarding where event boundaries are located (Baldwin & Baird, 1999; Newtson, Engquist, & Bois, 1977; Zacks, Tversky, & Iyer, 2001). Further, action segmentation is seemingly spontaneous and automatic, engaged in as a routine and ongoing component of perception (Hard, 2006; Zacks & Swallow, 2007).

The apparent ease with which people recognize breakpoints in action is remarkable given the complexity of the action stream itself. Human action is unquestionably a rich and highly variable stimulus; it is evanescent, often proceeds without pauses to mark the completion of individual units, and frequently features occlusion of relevant objects and body parts. Further, the underlying structure of action is also complex, typically characterized by a hierarchy reflecting the goals and sub-goals of an actor (e.g., Schank & Abelson, 1977).

Notably, human observers' skill in segmenting the action stream has been observed on a variety of different levels in line with this hierarchical structure. For example, segmentation of "chop carrot" can be on a *coarse* level, with event boundaries noted at the onset and offset of the entire chopping event, or it can be on a *fine* level, with each vertical movement of the knife noted as marking a discrete unit. In tasks assessing hierarchical segmentation, here again a high degree of consistency has been observed in people's segmentation behaviors (e.g., Hard, 2006; Zacks et al., 2001a), and fMRI studies have revealed differing activation levels in frontal and posterior areas in response to fine and coarse event boundaries, suggesting that the distinction between fine and coarse units is psychologically real on a neural level (e.g., Zacks et al., 2001b).

The ability to determine when one action has ended and another has begun, as well as segmenting action on multiple levels, supports how we make sense of the goal-directed action we observe in others. The fact that hierarchical event segmentation appears to be a relatively effortless process despite the complexity of the action stream itself suggests the workings of an equally complex system enabling this segmentation. Of particular relevance for the current studies, work by Hard and colleague (e.g., Hard, 2006; Hard & Recchia, 2006) suggests that event boundaries are processed differently than within-unit moments, with the detection of boundaries associated with a transient increase in cognitive processing load.

The idea that event boundaries might elicit an upsurge in cognitive processing is consistent with a comprehensive account of action segmentation put forth by Zacks and colleagues. These authors (e.g., Kurby & Zacks, 2007; Zacks et al., 2007) describe the Event Segmentation Theory, an account of how the human observer perceives and conceptualizes action in terms of events. A crucial component of Event Segmentation Theory rests on the observer's ability to make predictions about upcoming action. Such prediction generation is considered a spontaneous, online process that integrates incoming sensory information with prior knowledge and learning in an attempt to create a stable "event model." Event units correspond to periods in which prediction error rate is low; the observed action is consistent with the predictions being

made by the perceptual system, and the event model is stable. For example, within the event of cleaning off plates at the kitchen sink, the predictive system is able to generate accurate predictions of further plate cleaning based on such cues as the person's movements and prior knowledge about kitchen clean-up. Event boundaries, in contrast, are experienced when prediction error rate is high; to extend the example above, such boundary moments are likely to occur at the completion of a task (e.g., cleaning off plates in the kitchen) and before the initiation of another task (e.g., wiping the countertop), because these moments correspond with a reduced ability to predict the onset and content of the second event.

In order to update the event model at moments of reduced predictability, the system is believed to increase attention to the perceptual characteristics of the action stream and to activate new event schemata to replace the prior unsuccessful one. Hard and colleague (Hard, 2006; Hard & Recchia, 2006) provided an empirical test of whether boundaries were indeed associated with differential degrees of cognitive processing. As their methodology formed the basis of the first experiment in the current study, an in-depth explanation of their methods is in order. These authors reasoned that well-known paradigms developed for investigations of hierarchical processing of text would also be suitable for revealing aspects of hierarchical processing of action. In one such text processing study, individuals saw one word at a time from a passage of text and advanced themselves through word-by-word by pressing a button. The length of time between button presses was the primary dependent variable in this "moving window" method, with the idea being that longer reading times would be indicative of increased cognitive load associated with integration of past elements within and across text units into comprehensible larger units. Results indicated that participants tended to spend longer periods of time on words located at the ends of unit boundaries. Further, this "wrap up" effect was modulated by the level of any given unit; reading times were longer for words located at the ends of clauses and longer still for words located at the ends of sentences (Haberlandt & Graesser, 1989).

To study processing of hierarchical action using a similar technique, Hard and colleague adapted the moving window method for use with human action by asking participants to advance through a sequence of still-frame images. These images were taken from regular time intervals of footage of scripted human goal-directed action (e.g., one still-frame image sampled every second). Following this "slideshow" viewing phase, participants watched the live action footage from which the still images had been sampled and marked with a button press the locations of action boundaries (hereafter, 'breakpoints'). Participants completed this segmentation task a total of three times, providing judgments on fine, intermediate, and coarse levels.

Results from the slideshow task indicated that participants tended to spend a longer period of time looking at images close in time to moments judged to be breakpoints in comparison to images taken from within action units, suggesting that breakpoints elicited surges in attention. Further, paralleling results observed in text processing, the effect was modulated by the level of the action breakpoint, with slides close in time to moments judged as coarse-grained breakpoints receiving the longest looking times and those near fine-grained breakpoints receiving the least. This phenomenon, dubbed the dwell time effect, provided evidence that hierarchical segmentation occurs as part of real-time perception, without requiring explicit after-the-fact judgments of breakpoint locations. It further demonstrated the cognitive importance of action breakpoints; heightened attention was associated with moments participants explicitly judged to be breakpoints, and this effect was modulated based on whether that breakpoint was judged to be coarse, intermediate, or fine.

In the current paper, we report on another study that investigated hierarchical processing of action, this time using in vivo recordings collected from the Human Speechome Project. Audio-video data was collected from from the home of a single child using 11 ceiling mounted cameras and 16 boundary layer microphones. Over the first three years of the child's life, 90,000 hours of video was collected, representing roughly 70% of the child's waking experience (Roy et al., 2006).

As described above, past work has made much progress on elucidating the cognitive processes that make up the system enabling segmentation; however, these studies have examined segmentation of either scripted or animated scenes (e.g., Hard, 2006; Hard & Recchia, 2006; Zacks, 2004; Zacks et al., 2001a; Zacks, Kumar, & Abrams, 2009). The use of Speechome footage has the advantage of providing unscripted activity, allowing a test of the validity of methods that have been successful in revealing aspects of hierarchical segmentation of more artificial action scenes. Validation of the dwell time paradigm in Speechome footage additionally provides opportunities for the assessment of automated means of detecting action units, the topic taken up in Study 2.

Study 1 Method

Stimuli

Images for a slideshow viewing task were created by extracting one image every second from a 108-second movie clip take from the Speechome corpus (e.g., see Figure 1). The clip selected depicts an adult male preparing a meal. This video clip also served as the live action footage for which participants provided explicit segmentation judgments. For the explicit segmentation task, a different, 40-second clip of a woman cleaning the kitchen was used for training purposes.

Participants and Procedure

Participants were 28 university students (14 male) receiving class credit for participation. The experiment had two major phases, the *slideshow viewing task* and the



Figure 1: Sample image from slideshow depicting a person preparing food.

explicit segmentation judgment task. All participants began the session with the slideshow viewing task, in which they were instructed to advance at their own pace through the 108 still-frame images. Participants were told to click a mouse to advance the pictures. A Macintosh G4 computer was used to present stimuli on a 19.5" x 12" monitor, and Psychtoolbox (Brainard, 1997) was used to record participants' responses.

Following the slideshow, participants heard a brief description of how action can be seen as consisting of units, and examples of fine, intermediate, and coarse units in actions unrelated to those displayed during test were provided in these instructions. Participants then provided explicit judgments of where they believed breakpoints to be located, first providing judgments for the training video and then for the 108-second test (Speechome) video. Participants indicated their judgments with a key press. Participants were asked to provide segmentation judgments on fine, intermediate, and coarse levels, resulting in a total of three viewings of the movie clip. Half of the participants were asked to segment on a fine level on their first viewing of the clips, followed by segmenting on an intermediate level, and finishing with segmenting on a coarse level (fine-to-coarse order). The other half was asked to segment in the reverse order (coarse-to-fine order). Assignment of participants to these orders was random.

Study 1 Results

Do participants' explicit segmentation judgments reflect understanding of hierarchical structure?

One important preliminary question to answer is whether participants understood our instructions regarding segmentation on fine, intermediate, and coarse levels. Because we planned to compare the dwell times provided by each subject to their explicit breakpoint judgments made afterwards, it was important to ensure that participants differentiated among fine-, intermediate-, and coarse-level breakpoints during the explicit segmentation task.

Evidence for this understanding comes in part from results indicating that participants provided significantly different numbers of judgments for breakpoints at different levels, with fine-level breakpoints receiving the most judgments (M fine = 39.04 [SD = 23.32]), intermediate-level breakpoints receiving the next most (M intermediate = 12.68 [SD = 8 86]), and coarse-level breakpoints receiving the least (M coarse = 5.75 [SD = 2.81]), F (1.13, 30.42) = 61.44, $p < .0001$. (Greenhouse-Geisser statistics are reported due to violations in sphericity.) A significant linear trend characterized these data, F (1, 27) = 64.18, $p < .0001$. Thus, participants were clearly capable of recognizing breakpoints on different levels, providing the predicted differences in number of judgments according to level (fine vs. intermediate vs. coarse). As well, although individual differences in number of judgments were substantial (particularly in fine and intermediate judgments, as evidenced by the large standard deviations), 100% of participants provided the most judgments for fine breakpoints and the least for coarse breakpoints (sig. by a binomial test, $p < .0001$).

Participants were also fairly consistent in where they marked the locations of breakpoints. Figure 2 displays the number of fine, intermediate, and coarse level judgments across the 108 seconds of footage, with judgments "binned" into one-second intervals. As demonstrated by the distinct peaks and valleys reflecting moments commonly judged and rarely judged as breakpoints, respectively, it is apparent that participants frequently marked the same moments for all three levels of judgments, a pattern largely consistent with past studies using the same explicit segmentation method (e.g., Hard, 2006; Zacks et al., 2001a; Zacks et al., 2009).

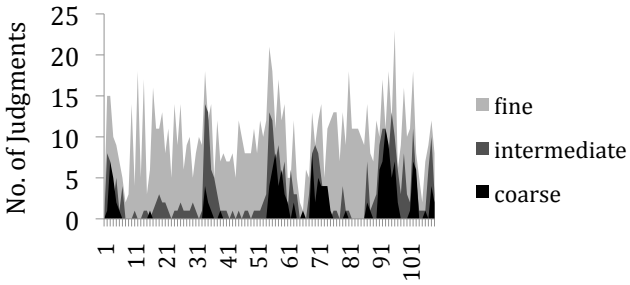


Figure 2: Participants' explicit judgments of fine, intermediate, and coarse level boundaries.

Does dwell time increase at breakpoints?

We next turned to one of the major hypotheses guiding Study 1, namely that participants' dwell time would be longer for images judged to be breakpoints compared to those that weren't. We used the participants' own explicit segmentation judgments, provided during the segmentation task, as the basis for determining which slides were considered breakpoints. Specifically, we applied a binning method, splitting the 108-second test clip into 1 second intervals, each corresponding to a single slide. Breakpoint judgments that fell into a given interval were matched to the corresponding slide, allowing us to classify breakpoint vs. non-breakpoint slides for each participant.

We then treated participants' raw dwell times to individual slides according to the following steps. Outliers (>3 SD above an individual's mean dwell time to all 108 slides) were removed from the data. Data were positively skewed, and thus a log transformation was applied. Due to

participants' tendency to dwell longer on slides at the beginning of the sequence and to speed up as the task continued, most participants' data were consistent with a power function. Significant portions of the variance were accounted for by the model for all participants (highest p value was .02). Thus, data were de-trended, and the residuals calculated based on the power function were used for analysis.

Because there were unequal numbers of slides in the different classifications (e.g., far fewer slides classified as breakpoints vs. non-breakpoints), means for each type were divided by standard deviations of that type, producing an effect size. All reported analyses are on these scores, hereafter referred to as dwell time scores. (Note that dwell time scores can be zero or negative since the residuals represent the difference between actual dwell time and times predicted by the power function; however, it is still the case that higher dwell time scores indicate overall longer dwelling on any given slide.)

A 2 (breakpoint status: breakpoint vs. non breakpoint) x 2 (segmentation order: fine-to-coarse vs. coarse-to-fine) mixed ANOVA (with breakpoint status as a within-subjects variable and segmentation order as a between-subjects variable) revealed only the predicted breakpoint status effect. Dwell time scores for breakpoint slides (M = .124, SEM = .046) were higher than for non-breakpoint (within-unit) slides (M = -.044, SEM = .026), F (1,26) = 6.40, p = .02. The main effect for segmentation order was not significant (M fine-to-coarse = .01, SEM .03; M coarse-to-fine = .07, SEM = .02), F (1, 26) = 3.1, p > .05, nor was the segmentation order x breakpoint status interaction significant, F (1, 26) = .03, p > .05. Dwell time scores were thus higher for breakpoints than non-breakpoints, supporting the first hypothesis.

Do dwell times vary according to fine, intermediate, and coarse levels?

Using the same binning method used to distinguish between breakpoint and non-breakpoint slides for each participant, classification of slides as breakpoints vs. non-breakpoints for each individual participant, slides were additionally categorized as falling at fine, intermediate, and coarse level boundaries. We then examined whether the dwell time effect was modulated based on whether a breakpoint was judged to be on a fine, intermediate, or coarse level. A 3 (segmentation level: fine, coarse, intermediate) x 2 (order: fine-to-coarse vs. coarse-to-fine) mixed between-within ANOVA was run, with segmentation level as the within-subjects variable and order as the between-subjects variable. Because of sphericity violations, we report Greenhouse-Geisser statistics. The predicted main effect for segmentation level was found, F (1.52, 39.43) = 16.17, p < .0001 (see Figure 3 for means). These differences were characterized by a significant linear trend, F (1, 26) = 21.20, p < .0001, with coarse-level breakpoints receiving the longest dwell times, intermediate-level breakpoints receiving the next longest, and fine-level breakpoints receiving the shortest dwell-times. The main effect for order was not significant (M coarse-to-fine = .161, SEM = .057; M fine-to-coarse = .087, SEM = .073), F (1, 26) = 1.23, p > .05; there also was no order x segmentation level significant interaction (F (1.57, 39.43) = .95, p > .05.

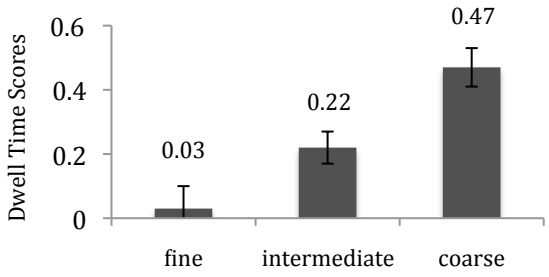


Figure 3: Dwell-time scores to slides designated as fine, intermediate, and coarse breakpoint. Data were characterized by a linear trend, p < .0001.

Study 2

Another line of investigation in action segmentation has focused on determining what perceptible features in the movement stream are relevant to segmentation. For instance, in the same study in which Hard and Recchia (2006) showed attentional differences to event boundaries, they additionally found that greater body movements on the part of the actor (as measured by overall pixel change between slides) significantly predicted observers' segmentation behavior. Similarly, in Zacks and colleagues' (2009) investigation of live action, the authors studied how changes in movement features such as the actor's acceleration and speed were predictive of observers' explicit segmentation judgments. In that study, an actor wore magnetic tracking devices on his hands while filming an action sequence, allowing for later extraction and calculation of the relevant movement features. The authors found that several movement features, including speed, acceleration, and change in distances among the actor's hands and head were predictive of observers' segmentation judgments, particularly for fine-grained event markings (see also Zacks 2004 for similar analyses with animated figures).

The ability to predict event boundaries based on perceptible features that can be extracted from video has great relevance to designers of informational systems that use identified actions as units of analysis. In addition to testing the validity of the dwell-time methodologies in naturalistic action, another goal of the current paper was to assess whether features visible in the action input were predictive of individuals' segmentation judgments. In Study 2, we extracted a set of predictive features, then analyzed how well these predictors correlated to the human judgments collected for Study 1

Study 2 Method

A set of motion features was extracted from the Speechome test clip using an accurate, semi-automatic

tracking system to annotate the positions of the body and hands of the actor appearing in the video (DeCamp & Roy, 2009). Positions were recorded as image coordinates (2D positions on the image, as compared to 3D positions in real space). Body position was defined as the center of the visible portion of the actor's head and torso. The positions of the hands were defined relative to the position of the body in order to reduce the covariance between them. After the position information was collected from the test video, it was used to compute the speed and acceleration of each body part, resulting in six features (see Table 1). The first and last seconds of data were also removed from analysis at this point because it was not possible to robustly define speed and acceleration at these points.

Kernel density estimation was applied to the breakpoints at each granularity level (i.e., fine, intermediate, and coarse). While this process smoothed the data, it also provided a continuous distribution of the breakpoints over time, which was more convenient for analysis than the raw judgment counts. Density estimation was performed with a Gaussian kernel. Bandwidths at were selected for each level using unbiased cross-validation, resulting in 0.92 s for fine breakpoints, 1.13 s for intermediate, and 1.27 s for coarse.

Study 2 Results

We found that each of the six features was significantly correlated to each breakpoint distribution (all p's < .001, see Table 1). The body speed feature achieved the highest correlation (r = 0.71) when correlated with coarse-grained judgments (see Figure 4). Right and left hand speeds had maximum correlations of 0.64 and 0.35, respectively. The acceleration features performed slightly worse, but were nevertheless significant.

Table 1: Correlations Between Visual Features and Breakpoint Distribution

	Correlation		
	Fine	Intermed	Coarse
Body Speed	0.49	0.65	0.71
Right-Hand Speed	0.47	0.64	0.64
Left-Hand Speed	0.45	0.44	0.35
Body Accel	0.40	0.52	0.54
Right-Hand Accel	0.35	0.51	0.47
Left-Hand Accel	0.34	0.40	0.36

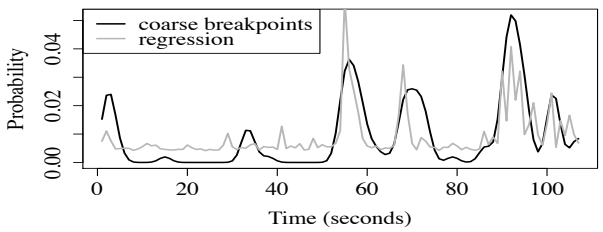


Figure 4: Univariate linear regression on coarse breakpoint distribution using body speed as predictor.

Discussion

In Study 1, we examined human observers' segmentation of naturalistic action, taking our stimuli from a large corpus of unscripted action (Speechome, e.g., Roy, 2006). Participants tended to dwell on images depicting breakpoints longer than non-breakpoints, and this difference was modulated based on whether a breakpoint was judged to be marking the completion of a fine-, intermediate-, or coarse-level unit. Despite the fact that our stimuli depicted naturalistic action, as well as the fact that participants had a decidedly different viewpoint of the action sequence itself than past studies of action (i.e., a ceiling-mounted camera provided the stimuli, and thus participants saw the actor from above), we replicated past findings of the dwell time effect (e.g., Hard, 2006; Hard & Recchia, 2006). Our findings suggest that the dwell time effect is a robust and valid phenomenon, capable of providing another window into the cognitive processes underlying segmentation.

The fact that participants' implicit behavior (dwell time) was associated with their explicit segmentation judgments also offers an exciting direction for future research within the developmental domain. There is clear indication already that infants as young as nine months can segment an action stream, a remarkable finding given infants' relatively impoverished understanding of goals and intentions (e.g., Baldwin et al., 2001; Saylor et al., 2007). Although this work represents an important demonstration of infants' action processing skill, the adaptation of dwell time methodology to this population has the potential to further expand our understanding of the developmental trajectory characterizing the segmentation process. The looking time methods used in these past developmental studies were not suitable for discerning hierarchical processing; further, the work examining hierarchical processing in adults has largely relied on participants' explicit understanding of what constitutes fine, intermediate, and coarse units (e.g., Zacks et al., 2001a, 2001b; Zacks et al., 2009), a task that is clearly beyond the capacity of infants and young children. We are actively pursuing adapting dwell time techniques for use both with preverbal infants as well as young preschool-aged children (e.g., Meyer, Hard, & Baldwin, 2009), a methodological advance that will allow us to study hierarchical processing across the lifespan.

In Study 2, we examined how perceptible movement features predicted human observers' judgments. Our results demonstrated that specific sources of information (i.e., head and hand speed and acceleration) were significantly associated with participants' segmentation judgments. Our results are consistent with similar movement change analyses performed by Zacks et al. (2009), suggesting that analysis of movement features may have broad utility in the design of automated systems of action analysis.

Notably, we additionally observed results that differed from those of Zacks et al., (2009); whereas we observed lower correlations as the judgment granularity was increased (i.e., correlations were highest when examining coarse-grained judgments and lowest when examining fine-grained

judgments), Zacks and colleagues actually observed the opposite. We speculate that this might be attributed to the differences between videos; in our footage the actor had no discernible facial features, and local movements of the hands and fingers were difficult to see; this may have reduced the ability of subjects to identify breakpoints as consistently at finer granularities. As well, the actor in our video moved his entire body through space (e.g., walking from a kitchen island to the sink), whereas the actor in Zacks et al.'s videos was seated. These gross bodily movements were frequently judged as coarse breakpoints and were clearly associated with several of our movement cues. Finally, the use of 2D video annotations in place of 3D motion sensor features may have provided less accurate measures that limited our ability to predict finer-grain events. In any event, the differences we observe offer inviting topics for future investigation relevant to the development of automated action analysis.

To summarize, we both validated the dwell time effect in naturalistic stimuli as well as found objective movement parameters predictive of individuals' segmentation behavior. The latter finding is of great relevance for researchers developing automated action analysis systems. Given that tracking whole people is now feasible for many types of video, current tracking technologies may enable the first steps towards systems that can automatically segment and identify actions from raw video, opening up new possibilities for human behavioral analysis.

Human action is an undeniably rich and complex stimulus. Yet, as we parse the events of our daily lives with little thought or apparent effort, the process may strike us as trivially easy. Nevertheless, the complexity of human action is apparent upon any attempt at formalization, and it poses a considerable challenge towards understanding human cognition. In this paper, we supply part of the solution by demonstrating how the human mind reacts and imparts structure to action sequences as they unfold. We also provide promising results from attempts to predict and model these reactions, suggesting future possibilities for the data driven analysis of events at a massive scale.

Acknowledgements

This research was supported by the U.S. Office of Naval Research, award no. N000140910187.

References

Baldwin, D., & Baird, J. A. (1999). Action analysis: A gateway to intentional inference. In P. Rochat (Ed.), *Early social cognition*, (pp. 215–240). Hillsdale, NJ: Lawrence Erlbaum Associates.

Baldwin, D., Baird, J., Saylor, M. M., & Clark, M. A. (2001). Infants parse dynamic action. *Child Development, 72*, 708–718.

DeCamp, P., & Roy, D. (2009). Human-machine collaborative approach to tracking human movement in multi-camera video. *Proceedings of the 2009 International Conference on Content-based Image and Video Retrieval (CIVR)*.

Haberlandt, K., & Graesser, A. C. (1989). Processing of new arguments at clause boundaries. *Memory & Cognition, 17*, 186-193.

Hard, B. (2006). Reading the language of action: Hierarchical encoding of observed behavior. Doctoral dissertation, Stanford University.

Hard, B., & Recchia, G. (2006). Reading the language of action. In N.A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, (pp. 1433-1439), Vancouver, CA.

Kurby, C. A., & Zacks, J. M. (2008). Segmentation in the perception and memory of events. *Trends in Cognitive Sciences, 12,* 72-79.

Meyer, M., Hard, B., & Baldwin, D. (2009, October). Children's processing of action boundaries. Poster presented at Cognitive Development Society, San Antonio, TX.

Newtson, D., Engquist, G., & Bois, J. (1977). The objective basis of behavior units. *Journal of Personality and Social Psychology, 35*, 847–862.

Roy, D., Patel, R., DeCamp, P., Kubat, R., Fleischman, M., Roy, B., Mavridis, N., Tellex, S., Salata, A., Guinness, J., Levit, M., & Gorniak, P. (2006). The human speechome project. *Symbol Grounding and Beyond: Proceedings of the Third International Workshop on the Emergence and Evolution of Linguistic Communication.* (pp. 192-168).

Saylor, M. M., Baldwin, D., Baird, J. A., & LaBounty, J. (2007). Infants' on-line segmentation of dynamic human action. *Journal of Cognition and Development, 8,* 113-128.

Schank, R. C., & Abelson, R. P. (1977). Scripts, plans, goals, and understanding: An inquiry into human knowledge structures. Hillsdale, NJ: Lawrence Erlbaum Associates.

Zacks, J. M. (2004). Using movement and intentions to understand simple events. *Cognitive Science, 28*, 979–1008.

Zacks, J. M., Braver, T. S., Sheridan, M. A., Donaldson, D. I., Snyder, A. Z., Ollinger, J. M., et al. (2001b). Human brain activity time-locked to perceptual event boundaries. *Nature Neuroscience, 4*, 651–655.

Zacks, J. M., Kumar, S., & Abrams, R. A. (2009). Using movement and intentions to understand human activity. *Cognition, 201,* 201-216.

Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S., & Reynolds, J. R. (2007). Event perception: A mind/brain perspective. *Psychological Bulletin, 133*, 273-293.

Zacks, J. M. & Swallow, K. M. (2007). Event segmentation. *Current Directions in Psychological Science, 16,* 80-84.

Zacks, J. M., Tversky, B., & Iyer, G. (2001a). Perceiving, remembering, and communicating structure in events. *Journal of Experimental Psychology: General, 130*, 29–58.

The Perception of Humans and Robots: Uncanny Hills in Parietal Cortex

Ayse Pinar Saygin (saygin@cogsci.ucsd.edu)
Department of Cognitive Science, University of California, San Diego
La Jolla, CA 92093-0515 USA

Thierry Chaminade (tchamina@gmail.com)
Mediterranean Institute for Cognitive Neuroscience, Aix-Marseille University CNRS
Marseille, 13402 France

Hiroshi Ishiguro (ishiguro@ams.eng.osaka-u.ac.jp)
Department of Adaptive Machine Systems, Osaka University
Suita, Osaka, Japan

Abstract

We report on a functional magnetic resonance imaging (fMRI) study of the perception of human and artificial agents. Participants viewed videos of familiar body movements enacted by the android Repliee Q2, the human after whom it was modeled, and the "skinned" version of Q2 revealing its mechanical parts. We used a neural adaptation (repetition suppression) analysis to reveal brain areas sensitive to body movements, and explored whether the identity of the perceived agents modulated these responses. We found significantly higher activity in a distributed network of brain areas for the android, most notably in anterior intraparietal cortex. The responses for the human and the robot with the mechanical appearance resembled each other. We interpret these results within the framework of predictive coding and suggest that the "uncanny valley" phenomenon may have its roots in processing conflicts within the brain's action perception system.

Keywords: action perception; body perception; biological motion; social robotics; artificial agents; neuroimaging; fMRI; uncanny valley

Introduction

In the near future, artificial agents and humanoid robots are expected to be part of our daily lives, not only in entertainment and retail, but also in important domains such healthcare and education (Billard, Robins, Nadel, & Dautenhahn, 2007; Dautenhahn, 2007; Kanda, Ishiguro, Imai, & Ono, 2004). Thus, exploring human factors in interactive robot design and development is crucial (Ishiguro, 2007; MacDorman & Ishiguro, 2006). Conversely, experiments using artificial agents can address questions about the functional properties of mechanisms involved in the perception of others' actions (Blake & Shiffrar, 2007; Rizzolatti & Craighero, 2004). Here, we summarize a neuroimaging study that we performed as part of an interdisciplinary research program that aims to reveal factors that can guide the design of future artificial agents, as well as to improve our understanding of action and body movement perception more generally.

In primates, the perception of body movements is supported by network of lateral superior temporal, inferior parietal and inferior frontal brain areas (Rizzolatti & Craighero, 2004). Here we will refer to this network as the Action Perception System (APS). The frontal and parietal nodes of the system are known to contain mirror neurons, which respond not only when the monkey executes a particular action, but also when it observes another individual perform the action. The existence of a similar system in humans has been suggested by several neuroimaging and lesion studies (e.g., Fadiga, Fogassi, Pavesi, & Rizzolatti, 1995; Grafton, Arbib, Fadiga, & Rizzolatti, 1996; Hari et al., 1998; Iacoboni et al., 1999; Saygin, 2007; Saygin, Wilson, Dronkers, & Bates, 2004).

The neural activity in premotor and parietal regions during action perception is often interpreted within the framework motor resonance, where "an action is understood when its observation causes the motor system of the observer to 'resonate'" (Rizzolatti, Fogassi, & Gallese, 2001). But what are the boundary conditions for this resonance?

There is a small neuroscience literature on the perception of artificial agents, including robots (Chaminade & Hodgins, 2006; MacDorman & Ishiguro, 2006). Unfortunately, the results are not consistent. Some experiments have reported that robot actions affect the observers' own motor processing or the activity of the APS, whereas others have argued that the APS does not respond, or responds weakly if the perceived actor is an artificial agent (Catmur, Walsh, & Heyes, 2007; Chaminade, Hodgins, & Kawato, 2007; Gazzola, Rizzolatti, Wicker, & Keysers, 2007; Kilner, Paulignan, & Blakemore, 2003; Oberman, McCleery, Ramachandran, & Pineda, 2007; Press, Gillmeister, & Heyes, 2007; Tai, Scherfler, Brooks, Sawamoto, & Castiello, 2004). Furthermore, the specific roles of biological appearance or biological motion have not been sufficiently explored in these experiments, but is an area of interest in social robotics, cognitive neuroscience, and vision science (Chaminade, Hodgins, & Kawato, 2007; Cook, Saygin, Swain, & Blakemore, 2009; Kanda,

Miyashita, Osada, Haikawa, & Ishiguro, 2008; Minato, Shimada, Itakura, Lee, & Ishiguro, 2006; Oyedele, Hong, & Minor, 2007; Saygin, Wilson, Hagler, Bates, & Sereno, 2004).

On the one hand, it seems reasonable that the closer the match between the observed action and the observers' own sensorimotor representations, the more efficient the simulation will be. In support for this, the APS is modulated by whether the observer can in fact perform the seen movement (Calvo-Merino, Grezes, Glaser, Passingham, & Haggard, 2006; Casile & Giese, 2006). The appearance of the observed agent may be additionally important (Buccino et al., 2004; Chaminade, Hodgins, & Kawato, 2007).

On the other hand, human resemblance is not necessarily always a positive feature in robots. The "uncanny valley" phenomenon points out that as a robot is made more human-like in its appearance, the reaction to it becomes more and more positive and empathetic, until a point is reached at which the robot becomes oddly repulsive (Mori, 1970). The effect is well-known in robotics and animation. For example, the movie *Polar Express* (Warner Bros) was criticized for the characters that viewers found creepy and disturbing. The more recent feature *Avatar* (20th Century Fox) received praise for animations that did not fall into the uncanny valley. Despite such well-known examples, and significant anecdotal evidence, there is little scientific data to characterize the uncanny valley (MacDorman, Green, Ho, & Koch, 2009; Steckenfinger & Ghazanfar, 2009).

The Present Study

This paper briefly describes the approach we took to this topic and summarizes the data from an fMRI repetition suppression study. We performed fMRI as participants viewed video clips of human (H) and robotic agents carrying out recognizable actions. We used Repliee Q2, a humanoid robot developed at Osaka University in collaboration with Kokoro Ltd (Ishiguro et al., 2006). This robot has a very human-like appearance (Figure 1b). In order to achieve this, the robot's face was modeled after an adult Japanese female (Figure 1a). Importantly, Repliee Q2 was videotaped both in its original human-like appearance (the Q2H condition, Figure 1b) and in a modified, more mechanical appearance (the Q2R condition, Figure 1c). In this latter condition, we removed as many of the surface elements as possible in order to reveal the electronics and mechanics underneath. The silicone covering the face and hands could not be removed, so we used a custom mask and gloves to change the appearance of these body parts. The end result was that the robot's appearance became obviously mechanical (e.g., metal arms and joints).

There were three conditions: human (H), robot with human appearance (Q2H) and robot with mechanical appearance (Q2R). However, since the Q2H and Q2R are in fact the same robot, the kinematics are identical for these two conditions. In terms of appearance, H and Q2H are very close to each other, whereas Q2R lies on the mechanical end. In terms of kinematics, H represents truly biological motion and Q2H and Q2R are identical, both with mechanical kinematics.



Figure 1. Still frames from the videos used in the experiments depicting the three agents.

The articulators of Repliee Q2 were programmed over several weeks at the Intelligent Robotics Laboratory at Osaka University. The same movements were videotaped in both appearance conditions (Q2R and Q2H). The human (the same female adult to whom Repliee Q2 was designed to resemble) was asked to watch each of Repliee Q2's actions and then perform the same action naturally. All agents were videotaped in the same room and with the same background. A total of 8 actions per actor were used in the experiment, including both transitive (drinking water from a cup, picking up a piece of paper from a table, grasping a tube of hand lotion, wiping a table with a cloth) and intransitive actions (waving hand, nodding affirmatively, shaking head (negative), and introducing self). Video recordings were cut into 2 second long clips, were converted to grayscale, cropped to a uniform size.

20 right handed healthy adults participated. We used a 3T Siemens Allegra scanner at the Wellcome Trust Centre for Neuroimaging in London, UK and a standard T2* weighted gradient echo pulse sequence to obtain functional images (TR=2340 ms, TE=65 ms). 36 slices were acquired at an in-plane resolution of 3 x 3 mm and a through plane resolution of 2 mm and 1 mm gap. Each participant was given exactly the same introduction to the study and the same exposure to the videos prior to scanning since prior knowledge can affect attitudes to artificial agents differentially (Saygin & Cicekli, 2002). Participants were told whether each agent was a human or a robot such that by the time scanning started, they were not uncertain about the identity of the android.

A limitation of previous neuroimaging studies on this topic is that they explored the BOLD fMRI response (Logothetis, 2008). Repetition suppression (henceforth RS, also called fMRI adaptation) is a method applied to fMRI from neurophysiology and refers to the phenomena of reduced neural response to a repeated stimulus compared to the response to a novel stimulus (Grill-Spector & Malach, 2001; Henson & Rugg, 2003; Krekelberg, Boynton, & van Wezel, 2006). RS affects neurons sensitive to the repeated stimulus, so it can be used as a means to explore functional properties of brain areas. In recent years, RS has been applied to the study of action perception (Chong, Cunnington, Williams, Kanwisher, & Mattingley, 2008; Dinstein, Gardner, Jazayeri, & Heeger, 2008; Dinstein, Hasson, Rubin, & Heeger, 2007; Fujii, Hihara, & Iriki,

2008; Hamilton & Grafton, 2006, 2008; Kilner, Neal, Weiskopf, Friston, & Frith, 2009; Lestou, Pollick, & Kourtzi, 2008). This approach was well-suited to our goals as it allows us to test whether neurons in the APS code for biological appearance or biological motion.

Participants watched the action videos in 30 second blocks. There were 12 videos in each block with a 500 ms ISI. Each video was preceded by the same video and the other videos equal number of times and orders were counterbalanced across runs. Each video was preceded by

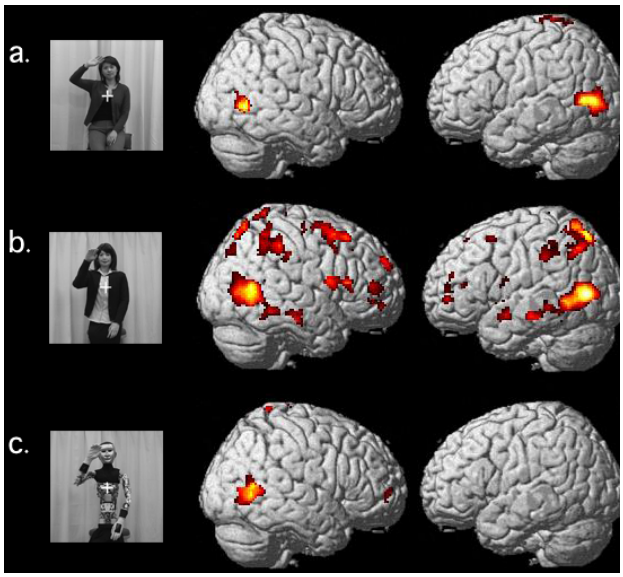


Figure 2. Repetition suppression results for the human (a), Q2H (b), and Q2R(c).

the same video (Repeat) or a different video (Non-repeat). To make sure subjects attended throughout, every 30-seconds, they were presented with a statement about which they made a True/False judgment using a button box (e.g., "I did not see her waving her hand"). The fMRI data were analyzed with SPM5 using standard procedures (http://www.fil.ion.ucl.ac.uk/spm).

Results

For each agent, we identified regions showing a repetition suppression effect at $p<0.05$ and cluster size of > 30 voxels. The effect of repetition suppression differed between the agents (Figure 2). Posterior temporal cortex showed suppression for all agents, but in the left hemisphere there was significantly less response to Q2R. This area corresponds the Extrastriate Body Area or EBA (Peelen, Wiggett, & Downing, 2006), which responds to the visual perception of the human body.

We otherwise did not find evidence for APS coding for the biological appearance or biological movement of the perceived agents. Instead, in comparison to H and Q2R, a larger network showed suppression for Q2H, despite the use of the same procedures and thresholds. This of course, brings to mind the uncanny valley, except we observed "hills" in the form of increased neural responses rather than valleys. Although we cannot include the details here due to space constraints, a region of interest (ROI) analysis further quantified these results, revealing a significant interaction in the inferior parietal lobule between the agents.

Discussion

We interpret these data within the predictive coding framework, which is based on minimizing prediction error though recurrent interactions among levels of a cortical hierarchy (Bar, 2009; Friston, 2005; Kilner, Friston, & Frith, 2007). During the perception of H and Q2R, where there is no mismatch between the appearance and the movement of the agent. For Q2H on the other hand, there is a human-like appearance that leads to a conflict when this information is integrated with the movement kinematics of the agent. This will lead to the generation of a prediction error, which is propagated in the network until the errors of each node are minimized. It is possible to measure prediction errors using neuroimaging (Friston, 2010). It is not possible from the current data to know the exact source and time course of error propagation, but it is clear that the cortical network is engaged strongly during the perception of Q2R compared with the agents that lead to less prediction error. The effect is largest in parietal cortex, which is the node of the network that links the posterior, visual components of the APS and the frontal, motor components (Matelli & Luppino, 2001; Seltzer & Pandya, 1994).

The present study is only a beginning. This framework provides hypotheses that we are testing in new studies. We are now utilizing animation to modulate the appearance and movement parameters more precisely (although this may lead to decrease in presence (Sanchez-Vives & Slater, 2005), whose importance in modulating APS is currently not known). We also need to use other neuroimaging and psychological methods in addition to, or in conjunction with fMRI to study the temporal dynamics of action processing.

With brief exposure times, Repliee Q2 can be mistaken for a human being, but longer exposure usually triggers the feeling of repulsion or discomfort characteristic of the uncanny valley (Ishiguro, 2006). While we did not explicitly assess the uncanny valley in this study, our results suggest an intriguing relationship between the APS and this phenomenon. We are currently exploring this in more sophisticated analyses as well as with new experiments.

In summary, we found that a robot with very humanlike appearance can cause differential responses compared with the same robot with a mechanical appearance, or with a human being that maximally resembles the robot. These differences were found in a network of brain areas, but most prominently in inferior parietal cortex, which connects the posterior areas involved in the visual perception of actions and biological motion to premotor areas in frontal cortex. We propose these "hills" in the brain activity reflect the prediction error that is generated as the brain processes these stimuli. We suggest that the uncanny valley may arise from

processing conflicts in the APS, and can be investigated using fMRI.

Acknowledgments

This research was supported by an innovative research grant to A.P. Saygin from the Kavli Institute for Brain and Mind (UCSD). Additional support was contributed by the European Commission and by the Wellcome Trust. We are grateful to members of the Intelligent Robotics Laboratory for their help in creating the experimental stimuli and to Jon Driver, Chris Frith, James Kilner and members of the Wellcome Trust Centre for Neuroimaging for their support of the fMRI study. We also appreciate discussions with Karl MacDorman, Takashi Minato, Javier Movellan, and Marty Sereno in the early stages of this project.

References

Bar, M. (2009). Predictions: a universal principle in the operation of the human brain. Introduction. *Philos Trans R Soc Lond B Biol Sci, 364*(1521), 1181-1182.

Billard, A., Robins, B., Nadel, J., & Dautenhahn, K. (2007). Building Robota, a mini-humanoid robot for the rehabilitation of children with autism. *Assist Technol, 19*(1), 37-49.

Blake, R., & Shiffrar, M. (2007). Perception of human motion. *Annu Rev Psychol, 58*, 47-73.

Buccino, G., Lui, F., Canessa, N., Patteri, I., Lagravinese, G., Benuzzi, F., et al. (2004). Neural circuits involved in the recognition of actions performed by nonconspecifics: an FMRI study. *J Cogn Neurosci, 16*(1), 114-126.

Calvo-Merino, B., Grezes, J., Glaser, D. E., Passingham, R. E., & Haggard, P. (2006). Seeing or doing? Influence of visual and motor familiarity in action observation. *Curr Biol, 16*(19), 1905-1910.

Casile, A., & Giese, M. A. (2006). Nonvisual motor training influences biological motion perception. *Curr Biol, 16*(1), 69-74.

Catmur, C., Walsh, V., & Heyes, C. (2007). Sensorimotor learning configures the human mirror system. *Curr Biol, 17*(17), 1527-1531.

Chaminade, T., Hodgins, J., & Kawato, M. (2007). Anthropomorphism influences perception of computer-animated characters' actions. *Social Cognitive and Affective Neuroscience, 2*(3), 206-216.

Chaminade, T., & Hodgins, J. K. (2006). Artificial agents in social cognitive sciences. *Interaction Studies, 7*(3), 347-353.

Chong, T. T., Cunnington, R., Williams, M. A., Kanwisher, N., & Mattingley, J. B. (2008). fMRI adaptation reveals mirror neurons in human inferior parietal cortex. *Curr Biol, 18*(20), 1576-1580.

Cook, J., Saygin, A. P., Swain, R., & Blakemore, S. J. (2009). Reduced sensitivity to minimum-jerk biological motion in autism spectrum conditions. *Neuropsychologia*.

Dautenhahn, K. (2007). Socially intelligent robots: dimensions of human-robot interaction. *Philos Trans R Soc Lond B Biol Sci, 362*(1480), 679-704.

Dinstein, I., Gardner, J. L., Jazayeri, M., & Heeger, D. J. (2008). Executed and observed movements have different distributed representations in human aIPS. *J Neurosci, 28*(44), 11231-11239.

Dinstein, I., Hasson, U., Rubin, N., & Heeger, D. J. (2007). Brain areas selective for both observed and executed movements. *J Neurophysiol, 98*(3), 1415-1427.

Fadiga, L., Fogassi, L., Pavesi, G., & Rizzolatti, G. (1995). Motor facilitation during action observation: a magnetic stimulation study. *J Neurophysiol, 73*(6), 2608-2611.

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions B, 360*(1456), 815.

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nat Rev Neurosci, 11*(2), 127-138.

Fujii, N., Hihara, S., & Iriki, A. (2008). Social cognition in premotor and parietal cortex. *Soc Neurosci, 3*(3-4), 250-260.

Gazzola, V., Rizzolatti, G., Wicker, B., & Keysers, C. (2007). The anthropomorphic brain: the mirror neuron system responds to human and robotic actions. *Neuroimage, 35*(4), 1674-1684.

Grafton, S. T., Arbib, M. A., Fadiga, L., & Rizzolatti, G. (1996). Localization of grasp representations in humans by positron emission tomography. 2. Observation compared with imagination. *Exp Brain Res, 112*(1), 103-111.

Grill-Spector, K., & Malach, R. (2001). fMR-adaptation: a tool for studying the functional properties of human cortical neurons. *Acta Psychol (Amst), 107*(1-3), 293-321.

Hamilton, A. F., & Grafton, S. T. (2006). Goal representation in human anterior intraparietal sulcus. *J Neurosci, 26*(4), 1133-1137.

Hamilton, A. F., & Grafton, S. T. (2008). Action outcomes are represented in human inferior frontoparietal cortex. *Cereb Cortex, 18*(5), 1160-1168.

Hari, R., Forss, N., Avikainen, S., Kirveskari, E., Salenius, S., & Rizzolatti, G. (1998). Activation of human primary motor cortex during action observation: a neuromagnetic study. *Proc Natl Acad Sci U S A, 95*(25), 15061-15065.

Henson, R. N., & Rugg, M. D. (2003). Neural response suppression, haemodynamic repetition effects, and behavioural priming. *Neuropsychologia, 41*(3), 263-270.

Iacoboni, M., Woods, R. P., Brass, M., Bekkering, H., Mazziotta, J. C., & Rizzolatti, G. (1999). Cortical mechanisms of human imitation. *Science, 286*(5449), 2526-2528.

Ishiguro, H. (2006). Android science: conscious and subconscious recognition. *Connection Science, 18*(4), 319-332.

Ishiguro, H. (2007). Projects and Vision in Robotics. *Lecture Notes in Computer Science, 4314*, 451.

Ishiguro, H., Asada, M., Shapiro, S. C., Thielscher, M., Breazeal, C., Mataric, M. J., et al. (2006). Human-Inspired Robots. *IEEE Intelligent Systems, 21*(4), 74-85.

Kanda, T., Ishiguro, H., Imai, M., & Ono, T. (2004). Development and evaluation of interactive humanoid robots. *Proceedings of the IEEE, 92*(11), 1839-1850.

Kanda, T., Miyashita, T., Osada, T., Haikawa, Y., & Ishiguro, H. (2008). Analysis of humanoid appearances in human-robot interaction. *IEEE Transactions on Robotics, 24*(3), 725-735.

Kilner, J. M., Friston, K. J., & Frith, C. D. (2007). The mirror-neuron system: a Bayesian perspective. *Neuroreport, 18*(6), 619-623.

Kilner, J. M., Neal, A., Weiskopf, N., Friston, K. J., & Frith, C. D. (2009). Evidence of mirror neurons in human inferior frontal gyrus. *J Neurosci, 29*(32), 10153-10159.

Kilner, J. M., Paulignan, Y., & Blakemore, S. J. (2003). An interference effect of observed biological movement on action. *Curr Biol, 13*(6), 522-525.

Krekelberg, B., Boynton, G. M., & van Wezel, R. J. (2006). Adaptation: from single cells to BOLD signals. *Trends Neurosci, 29*(5), 250-256.

Lestou, V., Pollick, F. E., & Kourtzi, Z. (2008). Neural substrates for action understanding at different description levels in the human brain. *Journal of Cognitive Neuroscience, 20*(2), 324-341.

Logothetis, N. K. (2008). What we can do and what we cannot do with fMRI. *Nature, 453*(7197), 869-878.

MacDorman, K. F., Green, R. D., Ho, C. C., & Koch, C. T. (2009). Too real for comfort? Uncanny responses to computer generated faces. *Computers in Human Behavior, 25*(3), 695-710.

MacDorman, K. F., & Ishiguro, H. (2006). The uncanny advantage of using androids in cognitive and social science research. *Interaction Studies, 7*(3), 297-337.

Matelli, M., & Luppino, G. (2001). Parietofrontal circuits for action and space perception in the macaque monkey. *Neuroimage, 14*(1 Pt 2), S27-32.

Minato, T., Shimada, M., Itakura, S., Lee, K., & Ishiguro, H. (2006). Evaluating the human likeness of an android by comparing gaze behaviors elicited by the android and a person. *Advanced Robotics, 20*(10), 1147.

Mori, M. (1970). The uncanny valley. *Energy, 7*(4), 33-35.

Oberman, L. M., McCleery, J. P., Ramachandran, V. S., & Pineda, J. A. (2007). EEG evidence for mirror neuron activity during the observation of human and robot actions: Toward an analysis of the human qualities of interactive robots. *Neurocomputing, 70*, 2194-2203.

Oyedele, A., Hong, S., & Minor, M. S. (2007). Contextual factors in the appearance of consumer robots: exploratory assessment of perceived anxiety toward humanlike consumer robots. *Cyberpsychol Behav, 10*(5), 624-632.

Peelen, M. V., Wiggett, A. J., & Downing, P. E. (2006). Patterns of fMRI activity dissociate overlapping functional brain areas that respond to biological motion. *Neuron, 49*(6), 815-822.

Press, C., Gillmeister, H., & Heyes, C. (2007). Sensorimotor experience enhances automatic imitation of robotic action. *Proc Biol Sci, 274*(1625), 2509-2514.

Rizzolatti, G., & Craighero, L. (2004). The mirror-neuron system. *Annu Rev Neurosci, 27*, 169-192.

Rizzolatti, G., Fogassi, L., & Gallese, V. (2001). Neurophysiological mechanisms underlying the understanding and imitation of action. *Nat Rev Neurosci, 2*(9), 661-670.

Sanchez-Vives, M. V., & Slater, M. (2005). From presence to consciousness through virtual reality. *Nat Rev Neurosci, 6*(4), 332-339.

Saygin, A. P. (2007). Superior temporal and premotor brain areas necessary for biological motion perception. *Brain, 130*(Pt 9), 2452-2461.

Saygin, A. P., & Cicekli, I. (2002). Pragmatics in human-computer conversation. *Journal of Pragmatics, 34*(3), 227-258.

Saygin, A. P., Wilson, S. M., Dronkers, N. F., & Bates, E. (2004). Action comprehension in aphasia: linguistic and non-linguistic deficits and their lesion correlates. *Neuropsychologia, 42*(13), 1788-1804.

Saygin, A. P., Wilson, S. M., Hagler, D. J., Jr., Bates, E., & Sereno, M. I. (2004). Point-light biological motion perception activates human premotor cortex. *J Neurosci, 24*(27), 6181-6188.

Seltzer, B., & Pandya, D. N. (1994). Parietal, temporal, and occipital projections to cortex of the superior temporal sulcus in the rhesus monkey: A retrograde tracer study. *The Journal of Comparative Neurology, 343*(3).

Steckenfinger, S. A., & Ghazanfar, A. A. (2009). Monkey visual behavior falls into the uncanny valley. *Proceedings of the National Academy of Sciences of the United States of America*., *106* (43), 18362-18366

Tai, Y. F., Scherfler, C., Brooks, D. J., Sawamoto, N., & Castiello, U. (2004). The human premotor cortex is 'mirror' only for biological actions. *Curr Biol, 14*(2), 117-120.

Modeling Cognitive-Affective Dynamics with Hidden Markov Models

Sidney K. D'Mello (sdmello@memphis.edu)
Institute for Intelligent Systems, University of Memphis,
Memphis, TN 39152 USA

Art Graesser (a-graesser@memphis.edu)
Department of Psychology, University of Memphis
Madison, TN 38152 USA

Abstract

We present and test a theory of cognitive disequilibrium to explain the dynamics of the cognitive-affective states that emerge during deep learning activities. The theory postulates an important role for cognitive disequilibrium, a state that occurs when learners face obstacles to goals, contradictions, incongruities, anomalies, uncertainty, and salient contrasts. The major hypotheses of the theory were supported in two studies in which participants completed a tutoring session with a computer tutor after which they provide judgments on their cognitive-affective states via a retrospective judgment protocol. Hidden Markov Models constructed from time series of learners' cognitive-affective states confirmed the major predictions as well as suggested refinements for the theory of cognitive disequilibrium during deep learning.

Keywords: affect dynamics, hidden markov model, learning.

Introduction

Deep learning and problem solving are emotionally rich experiences. Students experience boredom when the material does not appeal to them, confusion when they have difficulty comprehending the material and are unsure about how to proceed, frustration when they make mistakes and get stuck, and perhaps even despair and anxiety when their efforts seem to be futile and the big exam is creeping around the corner. This negative picture of the emotional experiences that accompany learning has a complimentary positive side. Students experience curiosity when they encounter topics that interest them, eureka moments when insights are unveiled and major discoveries made, delight when challenges are conquered, and perhaps even flow-like states (Csikszentmihalyi, 1990) when they are so engaged in learning that time and fatigue disappear.

There have been several theories that link cognition and affect very generally (Bower, 1981; Mandler, 1984; Ortony, Clore, & Collins, 1988; Russell, 2003; Stein & Levine, 1991). While these theories convey general links between cognition and emotions, they do not directly explain and predict the sort of emotions that occur during complex learning, such as attempts to master physics, biology, or computer literacy. Researchers in many different fields are familiar with Ekman's work on the detection of emotions from facial expressions (Ekman, 1984). However, the emotions that Ekman intensely investigated (e.g., sadness, happiness, anger, fear, disgust, surprise) have minimal relevance to learning in typical academic settings (D'Mello, Craig, Sullins, & Graesser, 2006; Kort, Reilly, & Picard, 2001; Lehman, D'Mello, & Person, 2008). Instead, the pervasive cognitive-affective states during complex learning include confusion, frustration, boredom, flow/engagement, and sometimes delight, surprise, anxiety, and curiosity (D'Mello et al., 2006; Lehman, Matthews, D'Mello, & Person, 2008).

The identification of the cognitive-affective states that occur during learning is critical, but it could be argued that merely knowing *what* states occur has limited utility. What is missing is a specification of *how* these states evolve, morph, interact, and influence learning and engagement. What is required is a fine-grained analysis of the rapid dynamics of the cognitive-affective processes that naturally occur during effortful learning activities.

Although affect dynamics has been generally ignored by theories that link affect and cognition during learning, one theory, called the cognitive disequilibrium theory, does address transitions between states. The theory postulates an important role for *cognitive disequilibrium* in comprehension and learning processes, a notion that has a long history in psychology (Berlyne, 1960; Festinger, 1957; Piaget, 1952). Cognitive disequilibrium is a state that occurs when learners face obstacles to goals, contradictions, incongruities, anomalies, uncertainty, and salient contrasts (Graesser, Lu, Olde, Cooper-Pye, & Whitten, 2005; Otero & Graesser, 2001; Piaget, 1952).

The cognitive disequilibrium theory is depicted in Figure 1 as a state transition network. The nodes (circles) in the figure represent the cognitive-affective states (in parentheses) and their presumed causes (in bold). Links represent situations that trigger transitions between the different states.

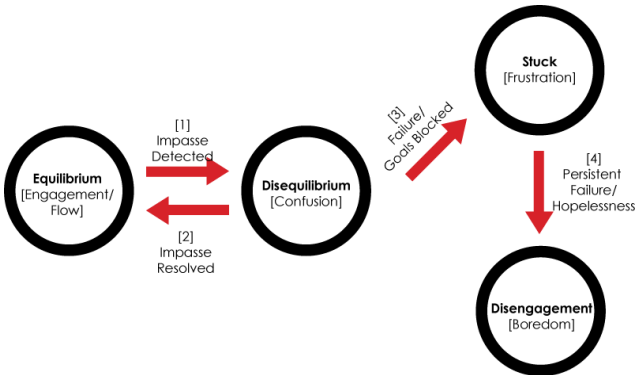


Figure 1. Cognitive Disequilibrium Theory

The theory assumes that learners are in a base state of engagement (perhaps a degree of flow) until they are confronted with a contradiction, anomaly, system breakdown, or error, and when they are uncertain about what to do next (Forbes-Riley & Litman, 2009; Graesser et al., 2005; Siegler & Jenkins, 1989; VanLehn, Siler, Murray, Yamauchi, & Baggett, 2003). Confusion is a key signature of the cognitive disequilibrium that occurs when an impasse is detected (Link 1). Learners must engage in effortful problem solving activities in order to resolve the impasse and restore equilibrium. Equilibrium is restored when the source of the discrepant information is discovered and the impasse is resolved, thereby causing learners to revert back to the engaged state (Link 2).

However, this form of productive confusion associated with impasse resolution can be contrasted with hopeless confusion. This occurs when the impasse cannot be resolved, the student gets stuck, and important goals are blocked. The theory hypothesizes that learners will experience frustration in these situations (Link 3). Furthermore, persistent frustration may transition into boredom, a crucial point at which the learner disengages from the learning process (Link 4).

We have confirmed some of the predictions of the theory in previous publications (D'Mello & Graesser, in review; D'Mello, Taylor, & Graesser, 2007). In particular, we have assessed the presence of oscillations between flow and confusion as well as transitions from confusion to frustration and frustration to boredom. However, verifying the presence of these transitions represents only one important component of the theory. The other crucial component that has not been yet empirically supported pertains to the internal causes that give rise to the observed cognitive-affective patterns. These include an *equilibrium* state that presumably activates the flow/engaged experience, a *disequilibrium* state that causes confusion, a *stuck* state that causes frustration, and a *disengaged* state that emits boredom. Our previous analyses so far have exclusively focused on transitions between the cognitive-affective states but have not explicitly addressed their causes. It is important, however, that both components of the theory be verified before it can be accepted as a useful explanation of the cognitive-affective phenomena that underlies deep learning.

Unfortunately, it is difficult to model the causes that underlie the cognitive-affective expressions. These states can be *observed* via facial expressions, body movements, and contextual cues, but the internal causes are *hidden* (i.e. they cannot be directly observed). This limitation can be alleviated via modeling techniques that permit the simultaneous modeling of both hidden and observed variables. In particular, the present paper describes a study in which Hidden Markov Models (HMMs) were used to model both the observed cognitive-affective states (confusion, frustrations, etc) and their hidden causes (equilibrium, stuck, etc), thereby testing the two components of cognitive disequilibrium theory. The HMMs were parameterized from learners' self reports on their cognitive-affective states via a retrospective judgment protocol after a tutorial session with AutoTutor, an Intelligent Tutoring System with conversational dialogues (Graesser et al., 2004).

Brief Description of HMMs

Hidden Markov Models are valuable tools for modeling system with sequential observable outcomes when the states producing the outcomes cannot be directly observed (i.e. they are hidden). They are widely used to model complex phenomenon with applications in a variety of disparate domains, such as automatic speech recognition, tutorial discourse, computational biology, financial economics, computer vision , and earthquake detection (Jurafsky & Martin, 2008; Rabiner, 1989).

HMMs are characterized by a set of parameters that can be estimated from available data. If there are m hidden states ($H = h_1, h_2, h_3, \ldots . h_m$) and n observable states ($O = o_1, o_2, o_3, \ldots . o_n$), then the parameters include a $m \times n$ emission probability matrix (E) and a $m \times m$ transition probability matrix (T). The emission probability matrix specifies the conditional probability of emitting an observed state o_t at time t given that the system is a hidden state h_t at the same time point [$\Pr(o_t|h_t)$]. On the other hand, the transition probability matrix specifies the conditional probability of transitioning from the current hidden state h_t to the next (or same) hidden state at the next time interval h_{t+1} [$\Pr(h_{t+1}|h_t)$].

As an example consider a simplified model of two hidden states for equilibrium (E) and disequilibrium (D) and two observed states for flow (F) and confusion (C). Here, $m = n = 2$ and both matrices are of size 2×2. The emission probability matrix would consist of the following four conditional probabilities: $\Pr(F|E)$, $\Pr(C|E)$, $\Pr(F|D)$, and $\Pr(C|D)$. Since it is assumed that a given hidden state emits one of the observable states, $\Pr(F|E) + \Pr(C|E) = 1$ and $\Pr(F|D) + \Pr(C|D) = 1$.

The transition probability matrix would also consist of four probabilities: $\Pr(E|E)$, $\Pr(F|E)$, $\Pr(D|D)$, and $\Pr(E|D)$. Once again, $\Pr(E|E) + \Pr(D|E) = 1$ and $\Pr(D|D) + \Pr(E|D) = 1$. Hence, given that a learner is in one of the hidden states, we can probabilistically determine which cognitive-affective state is most likely to be observed as well as what the next hidden state is likely to be.

Methods

Study 1

Participants. 28 undergraduate students (5 male and 23 female) from a large mid-south university participated for extra credit in their psychology courses.

Interaction with AutoTutor. Participants interacted with AutoTutor for 32 minutes on one of three randomly assigned topics in computer literacy: hardware, Internet, or

operating systems. AutoTutor is a validated intelligent tutoring system that helps learners construct explanations by interacting with them in natural language with adaptive dialogue moves similar to human tutors (Graesser et al., 2004). AutoTutor's dialogues are organized around difficult questions, such as why, how, what-if, what if not, how is X similar to Y, that require answers involving inferences, explanations, and deep reasoning. Although each question requires 3-7 sentence-like ideas in a correct answer, learners rarely give the complete answer in a single conversational turn. Therefore, the tutor scaffolds the construction of an answer by an adaptive dialogue with pumps for information, hints, prompts, assertions, summaries, and feedback. AutoTutor delivers its dialogue moves via an animated conversational agent that speaks the content of the tutor's turns.

A video of the participant's face and computer screen was recorded during the tutorial session (see Figure 2). Gross body language was tracked using Tekscan's Body Pressure Measurement System (not described here).



Figure 2. Learner interacting with AutoTutor

Judging Cognitive-Affective States. Participants provided self-judgments of their cognitive-affective states immediately after the tutorial session; learning activities during the session were not interrupted. Similar to a cued-recall procedure (Rosenberg & Ekman, 1994), the judgments for a learner's tutoring session proceeded by playing a video of the face along with the screen capture video of interactions with AutoTutor on a dual-monitor computer system (see center and right monitor in Figure 2). The screen capture included the tutor's synthesized speech, printed text, students' responses, dialogue history, and images, thereby providing the context of the tutorial interaction.

Participants were instructed to make judgments on what affective states were present at any moment during the tutoring session by manually pausing the videos (called *spontaneous* judgments). They were also instructed to make judgments at each 20-second interval; the video automatically stopped every 20 seconds (called *fixed* judgments). If the learner was experiencing more than one affective state, the learner was instructed to mark each state and indicate which was most pronounced. However, only the first choice (more prominent) affective states were included in the subsequent analyses.

Participants were provided with a checklist of seven states (boredom, flow/engagement, confusion, frustration, delight, surprise, and neutral) for them to mark along with definitions of the states. Hence, judgments were made on the basis of the participants' facial expressions, contextual cues via the screen capture, and the definitions of the cognitive-affective states.

Study 2

The participants were 30 undergraduate students (13 male and 17 female) from a mid-south university in the U.S. who participated for extra course credit.

Study 2, was similar to Study 1, but with two important differences. While participants in Study 1 interacted with the traditional typed-input version of AutoTutor, Study 2 participants spoke their responses to a new spoken-input AutoTutor. In addition to changing the input modality, there were a number of technical improvements in the new version of AutoTutor (version 3.1). These include improvements in conversational smoothness via a contextually-sensitive dialogue management module, state-of-the-art semantic and statistical natural language understanding mechanisms (Jurafsky & Martin, 2008), and an updated domain knowledge base for computer literacy.

The second difference between the two studies pertains to the retrospective affect judgment protocol. While participants in Study 1 provided affect judgments every 20 seconds and in-between each 20 second block, participants in Study 2 provided judgments at three pre-selected points plus some random points in the tutorial session. These included: (1) a few seconds after AutoTutor completed a dialogue move, (2) immediately before the learner started expressing his or her spoken response to the tutor, and (3) other randomly selected points in the dialogue. Participants provided approximately 30-35 cognitive-affective ratings at each of these three judgment points. These constituted the fixed judgment points. Similar to Study 1, the participants could stop the video at any time and make spontaneous judgments.

Results and Discussion

The retrospective affect judgment procedure yielded 2967 and 3099 self reported cognitive-affect judgments for Studies 1 and 2, respectively. A time series that preserved the temporal ordering of the cognitive-affective states was constructed for each participant. On average, there were 106 states (SD = 9) per time series for Study 1 and 103 states (SD = 14) for Study 2.

Since the goal of this paper is to investigate transitions between different states, and not persistence in the same state, the data was recoded to eliminate repetitions between states. For example, the sequence $X \rightarrow Y \rightarrow Y \rightarrow Z$ was converted to $X \rightarrow Y \rightarrow Z$. This process reduced the length of the time series to a mean of 64 states for both studies (SD_1 =

19, SD_2 = 15.24). On average, there was a state transition every 32.38 and 32.77 seconds for Studies 1 and 2, respectively (SD_1 = 11.17, SD_2 = 9.58). The recoding process did not alter the distribution of the cognitive-affective states.

Estimating Parameters of HMMs

The current analyses focused on discovering the parameters of HMMs that best explain the relationship between observable cognitive-affective states and the hidden variables that presumably govern their behavior. In particular, we estimated the parameters of an HMM with six observable states and four hidden states. The hidden states were equilibrium, disequilibrium, stuck, and disengaged, whereas the observable states were boredom, flow/engagement, confusion, frustration, delight, and surprise. Although the theory does not explicitly address the presence of delight and surprise, the states were included in the present analyses because they occasionally occur during learning sessions with AutoTutor (Graesser et al., 2006).

The present analyses constructed separate HMMs for each study from the time series of the cognitive-affective states. Parameters of the two matrices of each HMM were estimated with the Baum-Welch algorithm, which is the standard procedure used to train HMMs (Jurafsky & Martin, 2008; Rabiner, 1989). The algorithm begins with a set of initial parameters and then iteratively improves the estimates of these parameters by comparing how well the model constructed at each iteration fits the data. The algorithm converges when the discrepancy between the predictions made by the model and the training data minimally vary (i.e. within a preset threshold).

The choice of initial parameters plays an important role in the estimation process (Jurafsky & Martin, 2008). The initial parameters can be randomly seeded if there is no prior theory guiding their selection. In our case, the cognitive disequilibrium theory provides some important guidelines for initial parameter selection. For example, the theory hypothesizes that flow/engagement is expected to accompany the equilibrium state. Hence, the initial emission matrix was seeded such that the Flow|Equilibrium probability was slightly higher $(.18)$ than the other emissions stemming from the equilibrium state. In particular emissions for Boredom|Equilibrium, Confusion|Equilibrium, etc, were set to .164 $[(1-.180)/5 = .164]$. In this fashion, a small increase in emission probabilities was provided to confusion in the disequilibrium state, frustration in the stuck state, and boredom in the disengaged state.

The initialization process for the transition probability matrix was quite different. Here, transitions into the same hidden states were set to zero (because we are interested in modeling transitions to other states), while transitions to other hidden states were set to .333. Hence, each hidden state had an equal probability of transitioning to any other hidden state. The HMMs were seeded in this fashion to test whether hidden state transitions in the converged HMMs aligned with predictions of the cognitive disequilibrium theory. For example, equilibrium should transition into disequilibrium more frequently then stuck and disengaged.

It should be noted that the initial distribution of hidden states were also set to .25. The initial parameters of the HMM's are listed in Table 1 (see *Init* band). HMMs initialized on the basis of these parameters converged in 30 and 29 iterations for Study 1 and Study 2, respectively.

Exploring the Structure of the Converged HMMs

Before delving into the structure of the HMMs, we first evaluated how well the HMMs captured the dynamics of the state transitions in the two sets of analyses. In the first analysis, we compared each HMM to its random surrogate, which was an HMM that was seeded with the same initial parameters but was trained on randomly shuffled time series. Random surrogate comparisons provide a convenient face-validity test for time series analyses, because random shuffling eliminates all temporal dependencies between events while preserving the priori probabilities of individual events. The results indicated that the log-likelihood (LL) for HMM's constructed on the basis of a randomly shuffled time series was significantly ($p < .05$) lower than the LL for HMMs constructed from the original time series (d = 1.36 and 1.33 for Study 1 and Study 2, respectively).

The second analysis focused on the generalizability of the HMMs. Here we compared HMMs constructed and validated on the entire training set to HMMs constructed on partial data sets using a leave-one-out cross validation procedure (LVOCV). LVOCV involves constructing N HMMs, where each HMM is trained on time series from $N-1$ participants and tested on the time series of the remaining one participant. Correlations between the LL of LVOCV HMMs and HMMs trained on the entire data set were almost perfect (r = .99 for both Studies).

Table 1 lists the parameters of the HMMs for Study 1 and Study 2. As could be expected, the parameters of the emission matrix indicate that the flow state is emitted during equilibrium, confusion during disequilibrium, frustration when stuck, and boredom when disengaged. Hence, the converged emission matrix accurately models the hypotheses of the cognitive disequilibrium theory.

Although the transition matrix was seeded such that transitions between the hidden states were equivalent (.333), a different distribution of transitions emerged after training. In particular, consistent with the theory, the equilibrium state is more likely to transition into disequilibrium than the other states. As predicted, the disequilibrium state is more likely to transition back into equilibrium and the stuck state than the disengaged state.

The patterns were somewhat more murky for the stuck state. Although we hypothesized that stuck should transition into disengagement more frequently than equilibrium or disequilibrium, this pattern was not observed in the HMM for Study 1. The results were more in line with the theory for the HMM for Study 2, where stuck was equally likely to transition into disengagement and disequilibrium, but not

equilibrium. Finally, the theory does not explicitly address transitions from the disengaged state, and the HMMs did not reveal any clear transition pattern for this state.

It is also important to indicate that we constructed two additional HMMs for Studies 1 and 2. These HMMs were identical to the HMMs listed in Table 1 but were seeded with randomly initialized parameters instead of the theoretically derived initial parameters. The structure of these randomly-seeded HMMs were quite similar to the HMMs listed in Table 1, indicating that our theoretically derived initial parameters did not bias the models.

Table 1. Parameters of HMMs

		Emission Matrix						**Transition Matrix**			
HMM	Current Hidden State	Current Observed State *Bor*	*Con*	*Del*	*Flo*	*Fru*	*Sur*	Next Hidden State *Eq.*	*Dq.*	*St.*	*Dg.*
Init	Equilibrium	.16	.16	.16	.18	.16	.16	.00	.33	.33	.33
	Disequilibrium	.16	.18	.16	.16	.16	.16	.33	.00	.33	.33
	Stuck	.16	.16	.16	.16	.18	.16	.33	.33	.00	.33
	Disengaged	.18	.16	.16	.16	.16	.16	.33	.33	.33	.00
S1	Equilibrium	.00	.00	.02	**.96**	.00	.01	.00	**.42**	.28	.30
	Disequilibrium	.00	**.94**	.02	.00	.00	.04	**.43**	.00	**.33**	.24
	Stuck	.00	.00	.15	.00	**.79**	.06	.35	.33	.00	.33
	Disengaged	**.80**	.00	.06	.00	.00	.15	.38	.31	.31	.00
S2	Equilibrium	.00	.00	.06	**.89**	.00	.05	.00	**.46**	.27	.27
	Disequilibrium	.00	**.90**	.07	.00	.00	.03	**.37**	.00	**.35**	.27
	Stuck	.00	.00	.13	.00	**.83**	.03	.29	.36	.00	.36
	Disengaged	**.90**	.00	.04	.00	.00	.06	.32	.33	.35	.00

Notes. Eq. = equilibrium, *Dq* = disequilibrium, *St* = stuck, *Dg.* = disengagement

Discussion

The present paper used HMMs to test a theory of cognitive disequilibrium that is applicable to the dynamics of cognitive-affective states in deep learning environments. The major predictions of the theory were verified via the emission and transition matrices of the HMM which aligned with different aspects of the theory. In particular, the results supported an equilibrium state that emitted flow/engagement, a disequilibrium state that emitted confusion, and transitions between the equilibrium and disequilibrium states. These results support the assertion that students in the state of engagement/flow are continuously being challenged within their zones of optimal learning (Brown, Ellery, & Campione, 1998; Vygotsky, 1978) and are experiencing two-step episodes alternating between confusion and insight.

The HMMs confirmed the presence of a transition from disequilibrium to the stuck state that emitted frustration. However, the prediction of a transition from the stuck state to the disengaged state was only partially supported. The converged HMMs suggest that in addition to the predicted transition from the stuck to disengaged states, transitions from stuck to the disequilibrium and even the equilibrium states are permissible.

These transitions from frustration suggest that it is important to differentiate between different exemplars of frustration. Similar to the discrimination between productive and hopeless episodes of confusion, there might also be different manifestations of frustration. For example, being stuck for a short period of time and then obtaining an insight might trigger delight and cause a transition into the equilibrium state. Some evidence for this assertion can be obtained from the emission matrix which indicates that delight is sometimes emitted from the stuck state. Alternatively, the stuck state can transition into the disequilibrium state when an additional impasse is detected. The third manifestation of frustration is one that is predicted by the theory. Here, persistent failure and hopelessness from being stuck will eventually trigger disengagement, where the learner detaches from the learning session.

In summary, there appear to be three alternatives for transitions from frustration and the stuck state: (a) frustration is alleviated when a resolution is reached, (b) frustration oscillates with confusion when a stuck student detects an additional impasse, and (c) frustration transitions into boredom when a hopelessly stuck learner disengages from the learning session. Testing the fidelity of these transitions will require further empirical research.

Acknowledgments

This research was supported by the National Science Foundation (REC 0106965, ITR 0325428, HCC 0834847). Any opinions, findings and conclusions, or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of NSF.

References

Berlyne, D. (1960). *Conflict, Arousal, and Curiosity*. New York: McGraw-Hill Inc.

Bower, G. (1981). Mood and memory. *American Psychologist, 36*, 129-148.

Brown, A., Ellery, S., & Campione, J. (1998). Creating Zones of Proximal Development Electronically in Thinking Practices in Mathematics and Science Learning. In J. Greeno & S. Goldman (Eds.), (pp. 341-368). Mahawah, NJ: Lawrence Erlbaum.

Csikszentmihalyi, M. (1990). *Flow: The Psychology of Optimal Experience*. New York: Harper and Row.

D'Mello, S., Craig, S., Sullins, J., & Graesser, A. (2006). Predicting affective states expressed through an emote-aloud procedure from AutoTutor's mixed-initiative dialogue. *International Journal of Artificial Intelligence in Education, 16*(1), 3-28.

D'Mello, S., & Graesser, A. (in review). Dynamics of Cognitive-Affective States during Deep Learning.

D'Mello, S., Taylor, R., & Graesser, A. (2007). Monitoring affective trajectories during complex learning. In D. McNamara & G. Trafton (Eds.), *Proceedings of the 29th Annual Cognitive Science Society* (pp. 203-208). Austin, TX: Cognitive Science Society.

Ekman, P. (1984). Expression and the nature of emotion. In K. Scherer & P. Ekman (Eds.), *Approaches to emotion* (pp. 319-344). Hillsdale, NJ: Erlbaum.

Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford, CA: Stanford University Press.

Forbes-Riley, K., & Litman, D. (2009). Adapting to student uncertainty improves tutoring dialogues. In V. Dimitrova, R. Mizoguchi & B. Du Boulay (Eds.), *Proceedings of the 14th International Conference on Artificial Intelligence in Education* (pp. 33-40). Amsterdam: IOS Press.

Graesser, A., Lu, S., Olde, B., Cooper-Pye, E., & Whitten, S. (2005). Question asking and eye tracking during cognitive disequilibrium: Comprehending illustrated texts on devices when the devices break down. *Memory and Cognition, 33*, 1235-1247.

Graesser, A., Lu, S. L., Jackson, G., Mitchell, H., Ventura, M., Olney, A., et al. (2004). AutoTutor: A tutor with dialogue in natural language. *Behavioral Research Methods, Instruments, and Computers, 36*, 180-193.

Graesser, A., McDaniel, B., Chipman, P., Witherspoon, A., D'Mello, S., & Gholson, B. (2006). *Detection of emotions during learning with AutoTutor*. Paper presented at the 28th Annual Conference of the Cognitive Science Society, Vancouver, Canada.

Jurafsky, D., & Martin, J. (2008). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River, NJ: Prentice-Hall.

Kort, B., Reilly, R., & Picard, R. (2001). *An affective model of the interplay between emotions and learning*. Paper presented at the Proceedings of the International Conference of Advanced Learning Technologies, Madison, WI.

Lehman, B., D'Mello, S., & Person, N. (2008). *All Alone with your Emotions: An Analysis of Student Emotions during Effortful Problem Solving Activities*. Paper presented at the Workshop on Emotional and Cognitive issues in ITS at the Ninth International Conference on Intelligent Tutoring Systems.

Lehman, B., Matthews, M., D'Mello, S., & Person, N. (2008). What are you feeling? Investigating student affective states during expert human tutoring sessions. In B. Woolf, A. E., N. R. & L. S. (Eds.), *Proceedings of the 9th International Conference on Intelligent Tutoring Systems* (pp. 50-59).

Mandler, G. (1984). *Mind and Body: Psychology of Emotion and Stress*. New York: W.W. Norton & Company.

Ortony, A., Clore, G., & Collins, A. (1988). *The cognitive structure of emotions*. New York: Cambridge University Press.

Otero, J., & Graesser, A. (2001). PREG: Elements of a model of question asking. *Cognition and Instruction, 19*(2), 143-175.

Piaget, J. (1952). *The origins of intelligence*. New York: International University Press.

Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE, 77*(2), 257-286.

Rosenberg, E., & Ekman, P. (1994). Coherence between expressive and experiential systems in emotion. *Cognition & Emotion, 8*(3), 201-229.

Russell, J. (2003). Core affect and the psychological construction of emotion. *Psychological Review, 110*, 145-172.

Siegler, R., & Jenkins, E. (Eds.). (1989). *Strategy Discovery and Strategy Generalization*. Hillsdale, NJ,: Lawrence Erlbaum Associates.

Stein, N., & Levine, L. (1991). Making sense out of emotion. In A. O. W. Kessen, & F, Kraik (Eds.) (Ed.), *Memories, thoughts, and emotions: Essays in honor of George Mandler* (pp. 295-322). Hillsdale, NJ: Erlbaum.

VanLehn, K., Siler, S., Murray, C., Yamauchi, T., & Baggett, W. (2003). Why do only some events cause learning during human tutoring? *Cognition and Instruction, 21*(3), 209-249.

Vygotsky, L. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.

Productive Failure in Learning the Concept of Variance

Manu Kapur
National Institute of Education, Singapore

Abstract
In a study with 140, ninth-grade mathematics students on learning the concept of variance, students experienced either direct instruction (DI) or productive failure (PF), wherein they were first asked to generate a quantitative index for variance without any guidance before receiving direct instruction on the concept. Whereas DI students relied only on the canonical formulation of variance taught to them, PF students generated a diversity of representations and formulations for variance but were ultimately unsuccessful in developing the canonical formulation. On the posttest however, PF students performed on par with DI students on procedural fluency, and significantly outperformed them on data analysis, conceptual insight, and transfer items. These results challenge the claim that there is little efficacy in having learners solve problems targeting concepts that are novel to them, and that direct instruction alone is the most effective approach for teaching novel concepts to learners.

Introduction

Proponents of direct instruction bring to bear substantive empirical evidence against un-guided or minimally-guided instruction to claim that there is little efficacy in having learners solve problems that target novel concepts, and that learners should receive direct instruction on the concepts before any problem solving (Kirschner, Sweller, & Clark, 2006). Kirschner et al. (2006) argued that "Controlled experiments almost uniformly indicate that when dealing with novel information, learners should be explicitly shown what to do and how to do it" (p. 79). Commonly-cited problems with un-guided or minimally-guided instruction include increased working memory load that interferes with schema formation (Tuovinen & Sweller, 1999; Sweller, 1988), encoding of errors and misconceptions (Brown & Campione, 1994), lack of adequate practice and elaboration (Klahr & Nigam, 2004), as well as affective problems of frustration and de-motivation (Hardiman et al., 1986).

Klahr & Nigam's (2004) often-cited study compared the relative effectiveness of discovery learning and direct instruction approaches on learning the control of variable strategy (CVS) in scientific experimentation. On the acquisition of basic CVS skill as well as ability to transfer the skill to evaluate the design of science experiments, their findings suggested that students in the direct instruction condition who were explicitly taught how to design un-confounded experiments outperformed their counterparts in the discovery learning condition who were simply left alone to design experiments without any instructional structure or feedback from the instructor (I will return to this study in more detail in the discussion section). Further experiments by Klahr and colleagues (Chen & Klahr, 2008; Strand-Cary & Klahr, 2008), and others as well have largely bolstered the ineffectiveness of discovery learning compared with direct instruction (for reviews, see Kirschner et al., 2006).

Be that as it may, the above findings do not necessarily imply that there is little efficacy in having learners solve novel problems, that is, problems that target concepts they have not learnt yet (Schmidt & Bjork, 1992). To determine if there such an efficacy, a stricter comparison for direct instruction would be to compare it with an approach where students first generate representations and methods on their own followed by direct instruction. Expectedly, the generation process will invariably lead to failure, that is, students are rarely able to solve the problems and discover the canonical solutions by themselves. However, this very process can be productive for learning *provided* direct instruction on the targeted concepts is subsequently provided (Kapur, 2008; Koedinger & Aleven, 2007; Schwartz & Bransford, 1998; Schwartz & Martin, 2004).

As a case in point, I present evidence from an on-going research program on *productive failure* (Kapur, 2008; Kapur & Kinzer, 2009; Kapur et al., 2007).

Designing for Productive Failure

There are at least two problems with direct instruction in the initial phase of learning something new or solving a novel problem. First, students often do not have the necessary prior knowledge differentiation to be able to discern and understand the affordances of the domain-specific representations and methods underpinning the targeted concepts given during direct instruction (e.g., Schwartz & Martin, 2004). Second, when concepts are presented in a well-assembled, structured manner during direct instruction, students may not understand why those concepts, together with their representations, and methods, are assembled or structured in the way that they are (Chi et al., 1988; diSessa et al., 1991; Schwartz & Bransford, 1998).

To overcome these two problems, a learning design should focus squarely on first engaging students in processes that serve two critical cognitive functions, which in turn, prepare students for subsequent direct instruction: a) activating and differentiating prior knowledge in relation to the targeted concepts, and b) affording attention to critical features of the targeted concepts.

Productive failure is one such learning design. It comprises two phases—a generation and exploration phase followed by a direct instruction phase. In the generation and exploration phase, the focus is on affording students the opportunity to leverage their formal as well as intuitive prior knowledge and resources to generate a diversity of *structures*—concepts, representations and solution methods—for solving a complex problem; a problem that targets concepts that they have not been formally taught or learnt yet[1]. Research suggests that students do have rich *constructive resources* (diSessa & Sherin, 2000) to generate a variety of structures for solving novel problems (diSessa et al., 1991; Schwartz & Bransford, 1999). At the same time,

[1] The complexity of the problem is in relation to the learner. The problem is complex to the learner because the learner does not know the canonical representations and methods for solving it. To someone who knows these, the problem is no longer complex.

research also suggests that one cannot expect students, who are novices to the target content, to somehow generate or discover the canonical representations and domain-specific methods for solving the problem (Kirschner et al., 2006).

However, the expectation for the generation and exploration phase is not for students to be able to solve the problem successfully. Instead, it is to generate and explore the affordances and constraints of a diversity of structures for solving the problem. To the extent that students can persist in this process, the process not only activates but also differentiates their prior knowledge (as evidenced in the diversity of student-generated concepts, representations and methods). Furthermore, a comparison and contrast between the various structures also affords opportunities to attend to critical features of the targeted concepts (more on this in results section). Consequently, the generation and exploration phase provides the necessary foundation for developing deeper understanding of the canonical concepts, representations, and methods during direct instruction.

Empirical evidence for PF comes from a series of design experiments in grades seven through nine in Singapore mathematics classrooms (Kapur, 2009a, 2009b; Kapur et al., 2008; Kapur & Lee, 2009). Working with approximately 300 students from four public schools, the studies compared PF and DI designs for a two-week, curricular unit on average speed. Findings suggested that PF students produced a diversity of linked problem representations and methods for solving the problems but were ultimately unsuccessful in their efforts. Despite seemingly failing in their problem-solving efforts, PF students significantly outperformed DI students on both procedural fluency and complex analysis problems on the posttests. Furthermore, PF students also demonstrated significantly better transfer performance in adapting and building upon the targeted concepts to learn new concepts on their own.

These findings are consistent with other research programs that suggest that conditions that maximize performance in the shorter term are not necessarily the ones that maximize learning in the longer term (Clifford, 1984; Schmidt & Bjork, 1992). Examples of such research programs include VanLehn's (2003) work on *impasse-driven learning*, Schwartz and Bransford's (1998) work on *preparation for future learning*, Schwartz and Martin's (2004) work on *inventing to prepare for learning*, diSessa's (1991) work on *meta-representational competence*, Koedinger and Aleven's (2007) work on the *assistance dilemma*, among others (Kapur & Rummel, 2009).

Collectively, these research programs support the argument for designing conditions for learners to persist in the process of solving novel, complex problems without instructional support structures initially. Even though such a process invariably leads to failure in the shorter term, the extent to which this process affords learners opportunities to explore and generate a variety of representations and methods, the process can be germane for learning.

The purpose of this paper is to report findings from an on-going, classroom-based research program on productive failure in a public school in Singapore.

Method

Participants

Participants were 140, ninth-grade mathematics students (14-15 year olds) from an all-boys pubic school in Singapore. Students were almost all of Chinese ethnicity. Students were from four mathematics classes; three classes taught by one teacher (teacher A), and the fourth class by another teacher (teacher B). Students had no instructional experience with the targeted concept—variance—prior to the study, although they had learnt the concepts of mean, median, and mode in grades 7 and 8.

Research Design

A quasi-experimental, pre-post design was used with two classes (n = 31, 35) taught by teacher A assigned to the 'Direct Instruction' (DI) condition, and the other two classes (n = 35, 39), under teachers A and B, assigned to the 'Productive Failure' (PF) condition.

First, all students took a five-item paper and pencil pretest (α = .75) on the concept of variance. Not surprisingly, not a single student demonstrated canonical knowledge of the concept, and there was no significant difference between the four classes either, $F(3,136) = 1.665$, p = .177. Next, all classes participated in four, 55-minute periods of instruction on the concept as appropriate to their assigned condition. After the second and fourth periods, students from all classes took a five-item, five-point (1(low) - 5(High)) Likert scale engagement survey (α = .79). Finally, all students took a six-item, paper and pencil posttest (α = .74) comprising items on procedural fluency, data analysis, conceptual insight, and transfer.

In the DI condition, the teacher first explained the concept of variance and its canonical formulation as the square of the standard deviation ($SD^2 = \frac{\sum_{1}^{n}\left(x_i - \bar{x}\right)^2}{n}$) using a data analysis problem. Next, the teacher modeled the application of the concept by working through several data analysis problems, highlighting common errors and misconceptions, and drawing attention to critical features of the concept in the process. The data analysis problems required students to compare the variability in 2-3 given data sets, for example, comparing the variability in rainfall in two different months of a year, or comparing the consistency of performance of three soccer players, and so on. Thereafter, students worked face-to-face in triads on more data analysis problems. The teacher then discussed the solutions with the class. After each period, students were given similar data analysis problems for homework, which the teacher marked and returned to the students, usually by the following period.

The PF condition differed from the DI condition in only one important aspect. Instead of receiving direct instruction upfront, students spent two periods working face-to-face in triads to solve one of the data analysis problems on their own. The data analysis problem presented a distribution of goals scored each year by three soccer players for a twenty-year period. Students were asked to generate a quantitative index to determine the most consistent player. During this

generation phase, no instructional support or scaffolds were provided. Following this, two periods were spent on direct instruction just like in the DI condition. Note that because students in the PF condition spent the first two periods generating an index for variance, they solved fewer data analysis problems overall than their counterparts in the DI condition. To make this contrast even sharper, PF students did not receive any data analysis problems for homework.

Hypothesis The hypothesis tested was that *productive failure will be more effective than direct instruction in learning the concept of variance.* That is, expecting to replicate earlier work on productive failure (Kapur, 2008, 2009; Kapur & Lee, 2009), I hypothesized that students from the PF condition will be able to generate and explore various representations and methods for generating an index for variance (diSessa et al., 1991), but will not be successful in developing or discovering the canonical formulation on their own (Kirschner et al., 2006). However, this seeming failure would be integral for: a) engendering the necessary prior knowledge differentiation (evidenced in the diversity of student-generated structures), and b) drawing attention to critical features of the concept of variance (evidenced in the comparisons between the student-generated structures), which may help students better understand the concept when presented by the teacher during direct instruction subsequently (Schwartz & Bransford, 1998). This better understanding would result in better procedural fluency, data analysis, conceptual insight, and transfer.

Process Results

Process data included group-work artifacts produced on A4 sheets of paper. These provided a rich source of data about the nature of problem representations and methods generated by the students in the PF and DI conditions.

In the PF condition, groups produced four major and *progressively sophisticated* categories of methods and representations. The four categories were: a) central tendencies, b) qualitative methods, c) frequency methods, and d) deviation methods.

Category 1: Central Tendencies. Groups started by using mean, median, and in some cases, mode for data analysis. This was not surprising because students had been taught these concepts in the earlier grades. However, relying on central tendencies alone, it was not possible to generate a quantitative index for variance because the problem was designed in a way to keep the central tendencies invariant.

Category 2: Qualitative methods. Groups generated graphical and tabular representations that organized the data visually and were able to discern which player was more consistent. The visual representations (see Figure 1) afforded a qualitative comparative analysis between the players, but did not provide a quantitative index for measuring consistency even though the ideas of spread and clustering are quite evidently important qualitative conceptual underpinnings for the concept of variance.

Category 3: Frequency methods. Groups built on the qualitative methods to develop frequency-based measures of consistency. For example in Figure 2, groups used the frequency of goals scored within certain intervals to argue that the player with the highest number of goals in the interval containing the mean was the most consistent. Other groups counted the frequency with which a player scored above, below, and at the mean. Frequency methods demonstrated that students could quantify the clustering and bunching up trends in the qualitative representations.

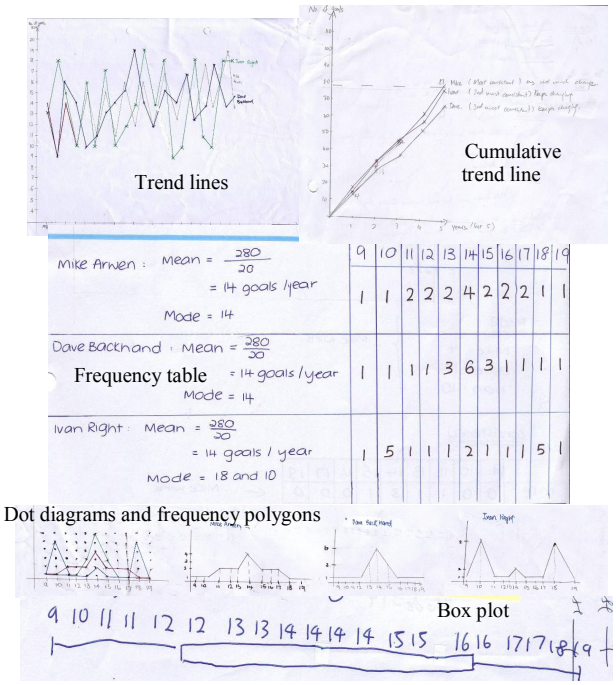


Figure 1 Examples of qualitative representations/methods

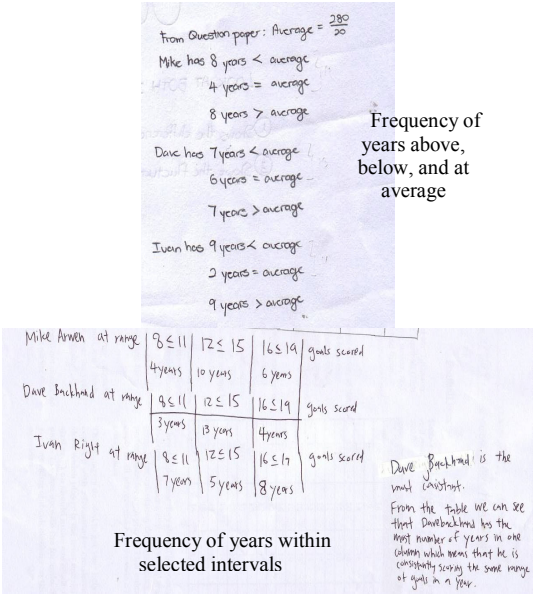


Figure 2 Examples of frequency representations/methods

Category 4: Deviation methods. Figure 3 presents some examples of the deviation methods. The simplest deviation method generated was the range (Deviation method 1, or simply D1). Some groups calculated the sum of year-on-year deviations (D2) to argue that the greater the sum, the

lower the consistency. Among these, there were those who considered absolute deviations (D3) to avoid deviations of opposite signs cancelling each other—an important conceptual leap towards understanding variance. Finally, there were some groups who calculated deviations about the mean (D4) only to find that they sum to zero. For both the D3 and D4 categories, some groups further refined their method to consider not the sum of the deviations, but the average (D5).

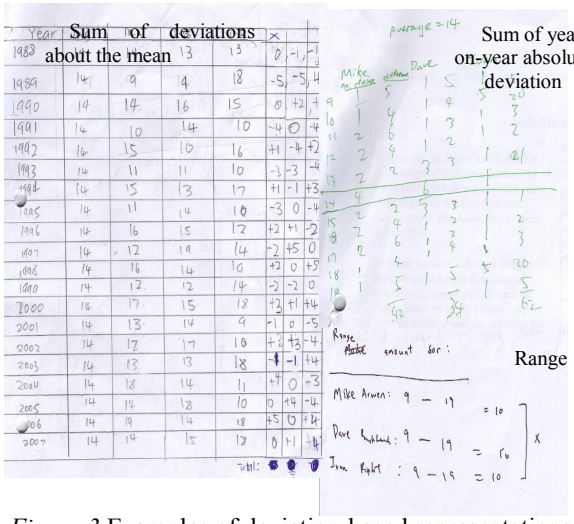


Figure 3 Examples of deviation-based representations and methods

In both the PF classes, all groups demonstrated representational competence at the Category 3 level or greater. Only 2 groups from PF-A and 1 group from PF-B did not reach Category 4. Consistent with the hypothesis, none of the groups were able to develop let alone use the canonical formulation on their own.

More importantly, note that these structures evidence the hypothesis that students will in fact be able to generate a rich diversity of structures to solve the problem without having first learnt the targeted concept of variance, and that comparisons between these structures will afford students the opportunities to attend to deep conceptual features of the concept. The latter needs more elaboration:

i. Comparing central tendencies with qualitative representations afforded an opportunity to attend to the feature that central tendencies alone cannot convey information about variance, and that different distributions with the same mean can have different variance.
ii. A comparison between the frequency methods and the qualitative methods afforded the opportunity to attend to the quantification of qualitative data into a mathematical index that returns a value for consistency.
iii. Because the deviation methods consider the relative position of a data point, a comparison with the frequency methods afforded students the opportunity to attend to the feature that, for consistency, it is not only important to count a point but also consider its position in relation to other points.
iv. Range (D1) afforded students the opportunity to attend to the feature that considering just the extreme points may not be a good measure of consistency, because it tells us nothing about the distribution in the middle. Comparing D1 with any of the qualitative representations easily afforded attention to this feature.
v. A comparison between D2 and D3 afforded students the opportunity to attend to the feature of why deviations must be positive. The comparison clearly shows that when deviations are left with their signs intact, positive and negative deviations cancel out resulting in a case where the variance could be highly underestimated.
vi. A comparison of D3 and D4 methods afforded students the opportunity to attend to the feature of why the reference point must be a fixed point (e.g., the mean), or else the index is sensitive to ordering of data. If the reference point for the deviation is not a fixed point, then a re-ordering of the data will result in a different value of consistency for the same formulation.
vii. A comparison between the sum and the average afforded the opportunity to attend to the feature of how dividing by the number of data points helps compare samples of different sizes.

In the DI condition, analysis of students' classroom work revealed that all students relied *only* on the canonical formulation to solve data analysis problems. This was not surprising given that the canonical formulation is relatively easy to compute and apply, and was corroborated with data from homework assignments. The average performance (i.e., percentage of problems solved correctly) on the homework assignments was high, M = 93.2%, SD = 5.3%. Finally, on the mean of the two self-reported engagement ratings, there was no significant difference between the PF condition, M = 3.84, SD = .51, and the DI condition, M = 3.82, SD = .43, $F(1, 138) = .035$, $p = .852$.

These process findings serve as a manipulation check demonstrating that students in the PF condition experienced "failure" at least in the conventional sense. In contrast, DI students were not only just as engaged as PF students but also demonstrated successful application of the canonical formulation to solve several data analysis problems. The high engagement ratings and performance results also suggest that the DI condition was not simply a case of poor instruction.

Outcome Results

Post-test The six-items on the posttest comprised:

i. one item on procedural fluency (calculating SD for a given data set),
ii. two items on data analysis (comparing means and SDs of two samples; these items were similar to the data analysis problems covered during instruction),
iii. two items on conceptual insight (one item dealing with sensitivity to ordering of data points, and another with outliers), and
iv. one item on transfer (item requiring the development of a normalized score for comparing incommensurable distributions. Note that normalization was not taught

during instruction, and therefore, students needed to flexibly adapt and build upon what they had learnt.).

Maximum score for each item was 10; two raters independently scored the items using a rubric with an inter-rater reliability of .96. Performance on the four types of items formed the four dependent variables. Controlling for the effect of prior knowledge as measured by the pretest, $F(4, 134) = 1.890$, $p = .112$, a MANCOVA revealed a statistically significant multivariate effect of condition (PF vs. DI) on posttest scores, $F(4, 134) = 16.802$, $p < .001$, partial $\eta^2 = .33$. There was no significant difference between the classes *within* the PF or DI conditions, nor was there any significant interaction between prior knowledge and experimental condition.

i. On the procedural fluency item, there was no significant difference between the PF condition, $M = 7.66$, $SD = 3.97$, and the DI condition, $M = 7.98$, $SD = 3.89$, $F(1, 137) = .819$, $p = .367$.
ii. On the data analysis items, students from the PF condition, $M = 14.11$, $SD = 4.20$, significantly outperformed those from the DI condition, $M = 11.38$, $SD = 4.86$, $F(1, 137) = 10.290$, $p = .002$, partial $\eta^2 = .07$.

It is important to note that PF students who were not given any homework and exposed to fewer data analysis problems still managed to perform on par with DI students on procedural fluency, and better than DI on data analysis in spite of DI students receiving homework and more practice and feedback on data analysis problems during instruction.

iii. On the conceptual insight items, students from the PF condition, $M = 16.40$, $SD = 6.41$, significantly outperformed those from the DI condition, $M = 8.20$, $SD = 6.15$, $F(1, 137) = 51.359$, $p < .001$, partial $\eta^2 = .27$.
iv. On the transfer item, students from the PF condition, $M = 4.93$, $SD = 2.99$, significantly outperformed those from the DI condition, $M = 3.07$, $SD = 2.35$, $F(1, 137) = 14.505$, $p < .001$, partial $\eta^2 = .10$.

Discussion

These findings are consistent with previous studies on productive failure with other mathematical topics and profile of students (Kapur, 2009a, 2009b; Kapur et al., 2008; Kapur & Lee, 2009), and also with other studies (e.g., Schwartz & Bransford, 1998; Schwartz & Martin, 2004). Notwithstanding the limitations of what can be achieved in a single study carried out within a particular domain, context and classroom-based setting, implications arising from the findings are simple and significant: There is indeed an efficacy in having learners generate and explore representations and methods for solving problems on their own even if they do not formally know the underlying concepts needed to solve the problems, and even if such un-supported problem solving leads to failure initially. The process analysis showed that this seeming failure was integral for: a) engendering the necessary prior knowledge differentiation (evidenced in the diversity of student-generated structures), and b) drawing attention to critical features of the concept of variance (evidenced in the comparisons between the student-generated structures), which may help students better understand the concept when presented by the teacher during direct instruction subsequently (Schwartz & Bransford, 1998).

This study contributes to the ongoing debate comparing the effectiveness of direct instruction with discovery learning approaches (e.g., Kirschner et al., 2006; Klahr & Nigam, 2004; Dean & Kuhn, 2007); discovery learning being often epitomized as the constructivist ideal. It is perhaps worth clarifying that a commitment to a constructivist epistemology does not necessarily imply a commitment to discovery learning. Simply leaving learners to generate and explore without consolidating is unlikely to lead to learning, or at least learners cannot be expected to "discover" the canonical representations by themselves as indeed our findings suggest. Instead, a commitment to a constructivist epistemology requires that we build upon learners' prior knowledge. However, one cannot build upon prior knowledge if one does not know what this prior knowledge is in the first place. It follows that at the very least the burden on the designer (e.g., teacher, researcher) is to first understand the nature of learners' prior knowledge structures; the very structures upon which the claimed "building" will be done. Designing for productive failure presents one way of doing so, wherein students first generate and explore representations and methods, and in the process externalize their prior knowledge structures, before direct instruction.

Interestingly, one could argue that Klahr & Nigam's (2004) study supports the above contention although it is often cited as a stellar example of the superior effectiveness of direct instruction over discovery learning. A careful reading of the study suggests that before assigning students to either a direct instruction or a discovery learning condition, Klahr and Nigam conducted a baseline assessment where they asked students to design four experiments on their own. As expected, only 8 out of the 112 students were able to design four un-confounded experiments, that is, the success rates before any instruction on the control of variables strategy (CVS) were very low. Students who were subsequently assigned to the discovery learning condition simply continued to design these experiments but without any instruction on CVS or any feedback. However, for students in the direct instruction condition, the instructor modeled and contrasted the design of both confounded and un-confounded experiments with appropriate instructional facilitation and explanation to make them attend to critical features of why CVS, unlike confounded experiments, helps isolate the effects of a factor. It was not surprising therefore that Klahr and Nigam found direct instruction to be more effective than discovery learning as described earlier in this paper.

From the perspective of productive failure however, the baseline assessment in Klahr and Nigam's (2004) study seems to function very much like the generation and exploration[2] phase where students generate their own structures (in this case, experiments) to solve a problem that targets a concept (in this case, CVS) that they had not learnt yet. If so, the very effects that Klahr and Nigam attribute to

[2] Indeed, Klahr & Nigam (2004) themselves termed it the "exploration phase."

direct instruction *alone* seem more appropriately attributed to a generation and exploration phase (their baseline assessment) followed by direct instruction. Therefore, much as Klahr and Nigam set out to show, in part, that there is little efficacy in students exploring and solving problems requiring concepts they have not learnt yet, their findings can be reinterpreted to support precisely the opposing contention that such exploration can in fact be efficacious provided some form of direct instruction follows, for without it, students may not learn much (as indeed the performance of the students in the discovery learning condition revealed). Thus argued, designing for a certain level of failure (as opposed to minimizing it) in the initial learning phase may well be productive for learning in the longer run. Future research would do well not to (over)simplistically compare discovery learning with direct instruction, but instead understand conditions under which these approaches can complement each other productively.

Acknowledgements

The research reported in this paper was funded by grants to the first author from the National Institute of Education of Singapore.

References

Brown, A., & Campione, J. (1994). Guided discovery in a community of learners. In K. McGilly (Ed.), *Classroom lessons: Integrating cognitive theory and classroom practice* (pp. 229–270). Cambridge, MA: MIT Press.

Chen, Z., & Klahr, D. (2008). Remote transfer of scientific reasoning and problem-solving strategies in children. In R. V. Kail (Ed.), *Advances in child development and behavior* (pp. 419–470). Amsterdam: Elsevier.

Chi, M. T. H., Glaser, R., & Farr, M. J. (1988). *The nature of expertise*. Hillsdale, NJ: Erlbaum

Clifford, M. M. (1984). Thoughts on a theory of constructive failure. *Educational Psychologist*, *19*(2), 108-120.

Dean, D., & Kuhn, D. (2007). Direct instruction vs. discovery: The long view. *Science Education*, *91*, 384–397.

diSessa, A. A., Hammer, D., Sherin, B. L., & Kolpakowski, T. (1991). Inventing graphing: meta-representational expertise in children. *Journal of Mathematical Behavior*, *10*(2), 117–160.

diSessa, A. A., & Sherin, B. L. (2000). Meta-representation: An introduction. *Journal of Mathematical Behavior*, *19*(4), 385–398.

Hardiman, P., Pollatsek, A., & Weil, A. (1986). Learning to understand the balance beam. *Cognition and Instruction*, *3*, 1–30.

Kapur, M. (2009a). Productive failure in mathematical problem solving. *Instructional Science*. doi: 10.1007/s11251-009-9093-x.

Kapur, M. (2009b). The role of productive failure in mathematics teaching and learning. In B. Kaur, B. H. Yeap, & M. Kapur (Eds.), *Mathematical Problem Solving* (pp. 43-68), Singapore: World Scientific.

Kapur, M. (2008). Productive failure. *Cognition and Instruction*, *26*(3), 379-424.

Kapur, M., Dickson, L., & Toh, P. Y. (2008). Productive failure in mathematical problem solving. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 1717-1722). Austin, TX: Cognitive Science Society.

Kapur, M., Hung, D., Jacobson, M., Voiklis, J., & Kinzer, C., & Chen, D-T. (2007). Emergence of learning in computer-supported, large-scale collective dynamics: A research agenda. In C. A. Clark, G. Erkens, & S. Puntambekar (Eds.), *Proceedings of the International Conference of Computer-supported Collaborative Learning* (pp. 323-332). Mahwah, NJ: Erlbaum.

Kapur, M., & Kinzer, C. (2009). Productive failure in CSCL groups. *International Journal of Computer-Supported Collaborative Learning* (*ij*CSCL), *4*(1), 21-46.

Kapur, M., & Lee, J. (2009). Designing for productive failure in mathematical problem solving. In N. Taatgen & V. R. Hedderick (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 2632-2637). Austin, TX: Cognitive Science Society.

Kapur, M., & Rummel, N. (2009). The assistance dilemma in CSCL. *Proceedings of the Computer-Supported Collaborative Learning Conference*, Rhodes, Greece.

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work. *Educational Psychologist*, *41*(2), 75-86.

Klahr, D., & Nigam, M. (2004). The equivalence of learning paths in early science instruction: Effects of direct instruction and discovery learning. *Psychological Science*, *15*(10), 661–667.

Koedinger, K. R., & Aleven V. (2007). Exploring the assistance dilemma in experiments with Cognitive Tutors. *Educational Psychology Review, 19*(3), 239-264.

Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science*, *3*(4), 207-217.

Schwartz, D. L., & Bransford, J. D. (1998). A time for telling. *Cognition and Instruction, 16*(4), 475-522.

Schwartz, D. L., & Martin, T. (2004). Inventing to prepare for future learning: The hidden efficiency of encouraging original student production in statistics instruction. *Cognition and Instruction, 22*(2), 129-184.

Strand-Cary, M., & Klahr, D. (2008). Developing elementary science skills: Instructional effectiveness and path independence. *Cognitive Development*, *23*(4), 488-511.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*, 257–285.

Tuovinen, J. E., & Sweller, J. (1999). A comparison of cognitive load associated with discovery learning and worked examples. *Journal of Educational Psychology*, *91*, 334–341.

Van Lehn, K., Siler, S., Murray, C., Yamauchi, T., & Baggett, W. B. (2003). Why do only some events cause learning during human tutoring? *Cognition and Instruction, 21*(3), 209-249.

Finding the Sweet Spot: Is There a Fixed Template for Culturally Successful Counterintuitive Narratives?

M. Afzal Upal (Afzal.Upal@drdc-rddc.gc.ca)
Adversarial Intent Section
Defence Research & Development Canada (DRDC) Toronto
1133 Sheppard Ave W, Toronto, M3M 3B9

Abstract

This article reports an investigation involving a series of studies carried out to critically examine the hypothesis that presence of 2 or 3 counterintuitive concepts in a story makes it more memorable than stories containing fewer or more such concepts. Our results paint a more complicated picture involving a number of interacting factors with contribution of the counterintuitive concepts to the global story cohesion emerging as a key factor.

Keywords: Memory, culture, folktales, concept learning.

Introduction

A number of recent studies have found that minimally counterintuitive concepts are recalled better than intuitive and maximally counterintuitive ideas (Barrett & Nyhof, 2001; Boyer, 1994, 2001; Boyer & Ramble, 2001). Better memorability for minimally counterintuitive concepts, these researchers argue, explains why such concepts form part of widespread religious beliefs and other widely shared cultural beliefs. However, as Atran (2003) has argued, these findings on their own are not sufficient to explain why most of the widespread cultural folktales contain only a small number of counterintuitive concepts[1] and are mostly composed of intuitive concepts. How and why do the apparently less memorable intuitive concepts continue to be successfully transmitted along with a small number of counterintuitive concepts? Does the presence of counterintuitive concepts improve overall recall for a story? If so, would an even larger number of counterintuitive concepts make the story even more memorable or would memorability drop of if counterintuitive concepts are added beyond a certain number?

Norenzayan, Atran, Faulkner, and Schaller (2006) report on an investigation carried out to study these questions. They selected 42 Grimm Brothers folktales such that half of the stories were judged to be "culturally successful" (they attracted more Google hits) and the other half were considered to be "culturally unsuccessful" (because they received fewer Google hits). Counterintuitive concepts present in each story were then counted. They found that a vast majority of the culturally successful folk tales had two or three counterintuitive ideas whereas counterintuitive ideas were more evenly distributed among the unsuccessful folktales. Subjects were then asked to read the stories and answer a number of questions to determine if the subjects thought that the stories were familiar, memorable, easy to understand, easy to transmit, and interesting enough to tell others. Their results show that stories with more Google hits were judged by the subjects to be more memorable and worth telling their friends. On the basis of this evidence, Norenzayan *et al.* argued that stories that contain two or three counterintuitive ideas enjoy memorability advantages over stories that have fewer (0 or 1) or more (4, 5, 6, or larger) counterintuitive ideas. They further argue that this should be true for all stories and not just Grimm Brother's tales or just Northern European folktales from the 19th century, or just for narratives of a certain length. They call stories containing 2-3 counterintuitive concepts as *MCI narratives* and state, "we propose that MCI narratives are culturally successful partly because they enjoy a stronger cognitive advantage in recall than other narrative templates" (Page 549)(Norenzayan, Atran, Faulkner, & Scaller, 2006). Let us call the hypothesis that stories containing 2 or 3 counterintuitive ideas are more memorable than stories containing fewer or more concepts as the MCI-hypothesis.

The objective of this paper is to carefully examine the MCI-hypothesis and its implications. This is accomplished through a series of studies. Initially, we replicate Norenzayan et al.'s methodology but then complement it with other techniques.

Study I

This study replicates Norenzayan *et al.*'s methodology for a different set of folktales. Aesop's fables are folktales credited to a Greek slave named Aesop who is thought to have lived from 620 to 560 BC. Most of the short stories contain between 50 and 500 words and are organized around moral themes. A number of stories contain counterintuitive concepts such as anthropomorphic animals. While Aesop's fables have survived for hundreds (if not thousands) of years and are widely known around the world, not all tales are equally well known. This study used George Fyler Townsend's collection (1867) containing 350 fables. Using Norenzayan *et al.*'s methodology, Google hits were computed for all 350 fables by querying for "Aesop" and the title of a story (e.g., "The Hare and the Tortoise"). Besides Google's initial estimate of the number of matching documents (which was the only measure used by Norenzayan *et al.*), this study also computed the actual

[1] The rest of the article uses the terms MCI concepts or simply counterintuitive concepts when referring to minimally counterintuitive concepts.

number of documents returned once Google was asked to retrieve all of the matching documents. Unfortunately, the rankings on the two counts did not match. The present study used the actual number of documents found as a more reliable indicator of a fable's popularity. The top 21 most popular tales had an average of 488 actual and 6321 estimated hits while the bottom 21 least popular tales had 80 actual and 197 estimated hits.

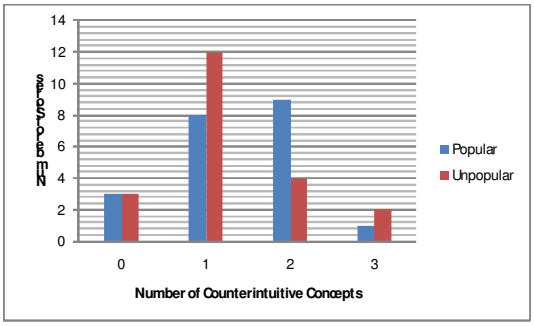


Figure 1: Distribution of counterintuitive concepts among the popular and unpopular Aesop's fables.

Next, a hypothesis-blind coder and the author coded the stories for the number of counterintuitive concepts in them. We agreed on 100% of the initial coding shown in Figure 1. It shows that contrary to predictions of the MCI hypothesis, a majority of popular fables do not have 2 or 3 counterintuitive concepts. Instead, 11 of the 21 popular stories contain 0 or 1 counterintuitive concept while remaining 10 have 2 or 3 counterintuitive concepts. A majority of unpopular stories (16 out of 21) also had 1-2 counterintuitive concepts and only 5 unpopular stories had 2-3 counterintuitive concepts.

A problem with studies reported so far is that they do not directly measure the memorability and are therefore unable to directly test the MCI hypothesis. The next study was designed to directly test the hypothesis that having 2-3 counterintuitive concepts makes a story more memorable than stories containing fewer or a larger number.

Study II

Material & Method

We decided not to use an existing set of stories (such as Grimm Brother's stories or Aesop's fables) because we wanted better control over (a) the number of concepts embedded in each story, and (b) subject's prior exposure to the stories. We designed three short stories containing 300-400 words each. Two of the stories, namely, "The Journey Home" and "The Trader" had been used in previous experiments (Barrett & Nyhof, 2001; Boyer & Ramble, 2001; Upal, 2005; Upal, Gonce, Tweney, & Slone, 2007) while the third story "The Night" was designed specifically for this experiment. Three versions of each story were created. Version I had one counterintuitive idea, while the second version had three and the third version had six counterintuitive ideas in it. Six packet-groups were then designed such that each packet-group contained all three stories and all three story types.

The balanced Latin square experiment required creation of thirty six distinct packets. Thirty six University of Toledo undergraduate and graduate students ranging in age from 18 to 24 were recruited to participate in the experiment. Subjects were asked to carefully read all three stories so that they could answer some questions about them. Next they were asked to solve simple arithmetic problems for one minute. Following that they were asked to write down as much of each story as they could remember. Story recall was measured by dividing each story into individual idea units constituting each story. The ideas roughly corresponded to the sentences in each story, although this wasn't always the case as some sentences were judged to have multiple concepts in them. "The Trader" was determined to have significantly smaller number of ideas (around 30) than "The Journey Home" or "The Night" each of which had roughly the same number of idea units (around 50 each).

Subject responses were coded using a binary coding scheme to measure whether a subject had recalled an element in the story or not. Story recall was measured by dividing the number of ideas a subject recalled by the total number of ideas in the story. Thus a perfectly recalled story would be assigned the recall value of 1 while a story that is not recalled at all would get the recall value of 0. The author and a hypothesis blind coder created two initial codings. We agreed on 89% of the initial coding. Disagreements were resolved through discussion to create one final coding.

Results

The recall rates for 1, 3, and 6 counterintuitive versions of the stories (Figure 2) show that story recall does not significantly vary as a function of the number of embedded counterintuitive concepts. This is true for both the overall story recall rates and also for each of the individual stories we studied.

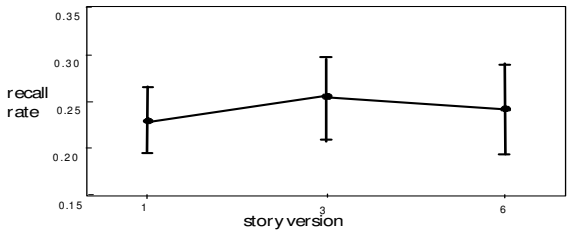


Figure 2: Overall story recall rates for 1, 3 and 6 concept versions of the stories.

Discussion

Our results not only call into question the MCI-hypothesis, they also indicate the need to seek an alternative answer to the question of what distinguishes memorable counterintuitive stories from forgetful ones? Previously (Upal, 2005, 2010; Upal, et al., 2007), I have argued that in order to answer these questions, we need to pay attention to cognitive processes involved in comprehension of text (Graesser, Singer, & Trabasso, 1994; Kintsch, 1998). Discourse analysis researchers and psycholinguists have

identified global cohesion among the elements of a text as a key factor in memorability (Halliday & Hasan, 1976).

Cohesion of a piece of text is defined as connections among various elements of the text and is not just a function of the text itself but also of the background knowledge that the reader possesses. The connections that make a text more or less cohesive include coreferences as well as causal and logical connections among its various elements. A text is better remembered by a reader if its constituents can be made coherent by the reader (Trabasso, Suh, Payton, & Jain, 1995). Furthermore, the more effort a reader spends in making a text coherent, the more memorable the text (Kim, 1999). Building on this and other work in cognitive science (Schank, 1999; Schank & Abelson, 1977) and humour research (Suls, 1983), I proposed a hypothesis that emphasizes the role played by the context in which counterintuitive concepts are embedded in making those concepts more or less memorable. This account suggests that, similar to other expectation-violating and schema-incongruent concepts, counterintuitive ideas are better remembered because they attract a reader's attention by violating the reader's expectations about what is to come next in the text. When a reader's expectations are violated, she attempts to resolve the situation by reasoning to justify the inclusion of expectation-violating information in the text by invoking a variety of knowledge that the reader possesses. If this postdiction effort is successful, the expectation-violating concepts become richly linked to the reader's existing mental representations, which were retrieved to explain the inconsistency to derive a coherent theme. They also become richly connected to the derived story theme itself. This may make counterintuitive elements of a narrative more likely to be recalled when the story title is provided as a cue.

This view suggests that memorability for a story should be mediated through story cohesion. Thus counterintuitive stories should only be remembered well if they can be made coherent by a reader. If a counterintuitive story is too incoherent (or judged too difficult to make coherent given a reader's motivation level) then it should not be well remembered.

The next study was designed to test this hypothesis. I wanted to know whether inclusion of various types of counterintuitive concepts equally affects story memorability. Depending on the context, inclusion of some counterintuitive concepts may, for instance, increase cohesion of a story while addition of other counterintuitive concepts may decrease it. Would inclusion of both types of concepts equally affect story memorability? The above account would suggest that stories including cohesion-enhancing concepts should be remembered better than stories that contain cohesion suppressing concepts.

Study III

Material and Method

I designed three short (95-125 words) Aesop-like fables. Each story involves two human or animal protagonists who happen to meet. At the end, the moral lesson of the story (the same as the story title) is uttered by one of the main characters. Four versions of each story were designed: (1) Coherent-Counterintuitive (CC), (2) Coherent-Intuitive (CI), (3) Incoherent-Counterintuitive (IC), and (4) Incoherent-intuitive (II).

In the coherent-counterintuitive version, both of the main characters are counterintuitive but their counterintuitiveness is *causally relevant* for making sense of the story and for connecting various elements of the story and for deriving the coherent theme that is the story title. For instance, in the CC version of "obscurity brings safety", the protagonists are an invisible-man and an all-seeing-woman. The counterintuitive property of each character is causally relevant because it allows a reader to make sense of the events to follow and to connect them to the moral lesson of the story. For instance, all-seeing-ability of the woman allows a reader to understand why she is able to see an otherwise invisible man. Man's invisibility is needed to understand woman's advice to him to become visible to make his life more enjoyable and why he decides to paint himself skin-tone and then why, on being mugged after becoming visible, he regrets his actions and utters, "obscurity brings safety." These particular counterintuitive properties are causally relevant because, without them, the story and its title make little sense and are not as coherent.

In the coherent-intuitive version, the protagonists are replaced by intuitive beings. However, their intuitive properties are still causally relevant to explaining the events in the story. For instance, in the CI version of "obscurity brings safety", the invisible-man is replaced by a reclusive man and the all-seeing-woman is replaced by a kind-but blunt woman. The man's reclusiveness allows the reader to understand why he is advised by the caring woman to go out and why the man regrets following her advice.

In the incoherent-counterintuitive version, the main characters are counterintuitive but their counterintuitiveness is irrelevant to the events in the story and does not help a reader in her attempt to derive a coherent theme from the story. For instance, the IC version of "obscurity brings safety" includes "a man who has feet for hands" and a "woman who is made of iron". These properties do not help the reader to make sense of why the woman asks the man to stop being invisible and why he decides to paint himself or why he proclaims that "obscurity brings safety" upon unfolding of the story's events.

In the incoherent-intuitive version, the main characters are intuitive beings whose explicitly mentioned intuitive properties are irrelevant to the events in the story and do not allow the reader to derive the moral lesson in the story's title. For instance, the II version of "obscurity brings safety" features "a man with brown hair" and "a woman with dark circles around her eyes." Both properties have little to do with the woman's advice, the man's actions, or the story title/theme.

Each subject packet included three stories. Varying the story order and story type yielded 192 possible packets. Out

of these, 40 packets were randomly selected to be given to 40 Occidental College male and female Cognitive Science and Psychology undergraduates who participated in the experiments for extra credit. After reading all three stories, subjects were instructed to solve simple arithmetic problems for one minute. Following that they were asked to write down as much of each story as they could remember. The subject responses were coded for recall by the author and a hypothesis blind coder following the same methodology as in Study 2. We also measured the number of words recalled and also recall rates for counterintuitive and intuitive descriptions of the protagonists. The two coders agreed on 96% of the initial coding. Disagreements were resolved through discussion to create one final coding.

Results & Discussion

The results are shown in Table 1. There was a significant effect of story cohesion while there was no significant effect of the number of counterintuitive concepts.

Table 1: (a) The first three tables show story recall data for individual stories, (b) the last table shows the overall results. The leftmost column shows the mean recall rates for propositions describing story protagonsists. The middle column shows the recall rate for all of the story elements including the protagonists. The rightmost column shows the recall rate for the rest of the story elements.

(a)

	Obscurity Brings Safety		
	Protagonist recall	Overall Story Recall	Story Minus Protagonist recall
Coherent-Counterintuitive	100	82	77.5
Coherent-Intuitive	64.4	87.5	59.5
Incoherent-Counterintuitive	86.4	62.7	56.8
Incoherent-Intuitive	65	55	52.5

	Never Laugh at Someone		
	Protagonist recall	Overall Story Recall	Story Minus Protagonist recall
Coherent-Counterintuitive	100	100	100
Coherent-Intuitive	100	100	100
Incoherent-Counterintuitive	57.1	57.1	57.1
Incoherent-Intuitive	54.5	54.5	54.5

	No Gratitude From the Wicked		
	Protagonist recall	Overall Story Recall	Story Minus Protagonist recall
Coherent-Counterintuitive	72	72	72
Coherent-Intuitive	45.8	45.8	45.8
Incoherent-Counterintuitive	33.3	33.3	33.3
Incoherent-Intuitive	90	90	90

(b)

	Overall		
	Protagonist recall	Overall Story Recall	Story Minus Protagonist recall
Coherent-Counterintuitive	92.4	92.4	92.4
Coherent-Intuitive	74.1	74.1	74.1
Incoherent-Counterintuitive	61.1	61.1	61.1
Incoherent-Intuitive	65	65	65

The coherent stories were significantly better recalled than incoherent stories ($F(1, 117) = 15.019$ $p = 0.00018$). Contrary to predictions of the MCI hypothesis, stories containing 2 counterintuitive concepts were not better recalled than stories containing 0 counterintuitive concepts ($F(1, 117) = 0.38129$ $p = 0.53811$). In fact, while the differences were not statistically significant, stories containing 2 counterintuitive concepts were less well recalled than stories without any counterintuitive concepts in them. If we control for cohesiveness and vary the number of counterintuitive ideas in a story, we get two distinct trends. As shown in **Error! Reference source not found.**, when counterintuitive ideas enhance cohesion, their addition makes a story more memorable (although not significantly so). However, when counterintuitive concepts cannot be easily integrated to derive the story theme, their addition results in lower recall (again differences are not statisticallysignificant).

Table 3(b) shows that coherent-counterintuitive stories were best recalled, followed by coherent-intuitive stories, which were better recalled than incoherent-intuitive stories. However, only recall for incoherent-counterintuitive stories was significantly lower than recall for coherent-counterintuitive and coherent-intuitive stories ($F(3, 115) = 6.3828$ $p = 0.00049$). The subjects recalled only half of the ideas from the stories in which incoherent protagonists were not causally relevant to the story theme.

Incoherent stories also prompted some subjects to add unsolicited comments to their written responses such as, “the story was unclear”, “this was a weird story” and “I didn't understand the story at all.” Incoherent-counterintuitive stories solicited more (2) comments than incoherent-intuitive stories (1 comment). There was also some evidence to suggest that subjects were attempting to make sense of the incoherent stories. For instance, consider the incoherent-counterintuitive version of Gratitude, where the man decides to go home and mow his lawn and have dinner with his family but the wolf is still mysteriously saved. Two subjects inferred that the man saved the wolf by helping it before going home while another subject said that the man saw the wolf on his way back and saved it! Three subjects made the incoherent version of Laugh coherent by changing it. Instead of the man making fun of the woman’s body and then surprisingly telling the woman never to laugh at people’s body, two of the subjects changed the story so that the man realizes on his own that he should never have

laughed at the woman's body. Another subject changed the story to suggest that the woman made fun of the man!

Results of this study further call into question the notion that inclusion of 2-3 counterintuitive concepts makes a story more memorable and more transmissible. Our results indicate that counterintuitive concepts only make a story more memorable if they can be easily integrated to make the story coherent. Having gathered some support for our hypothesis that story cohesion is key to explaining story recall, I wanted to see whether difference between story cohesion could account for difference in popularity for Aesop's Fables. The final study was designed to investigate this possibility.

Study IV

Material and Method

I designed 32 study packets by randomly ordering the 42 (21 popular and 21 unpopular) stories selected in Study I. Each story was followed by seven randomly ordered questions. Replicating Norenzayan *et al.*'s methodology for their Study 2, I asked subjects to first rate each tale on the following six attributes on 7-point scale (anchored by endpoints labeled s*trongly disagree* to s*trongly agree*). Subject responses were used to measure their perception of each story's:

- *familiarity* ("I have heard this story before"),
- *memorability* ("Right now if someone asked me to close my eyes and tell them the story that I just read, I think I could recall all or most of the critical elements of the story"),
- *likelihood of transmission* ("If I told a 7-year-old this story, he or she would tell it to other children"),
- *interest value* ("This story was interesting"),
- *understandability* ("This story was easy to understand"), and
- *moral lesson* ("This story has a strong moral lesson").

In addition to the above six factors measured by Norenzayan *et al.*, I added the query "I could easily make a few modifications to the story (such as changing the main characters) to make the story's moral lesson even more apparent". Believing that incohesive stories should be judged by adult English readers as more amenable to a change than cohesive stories, I thought that subject responses to this question should be inversely related to story cohesion.

Thirty two adult male and female subjects from DRDC-Toronto participated in this experiment for remuneration. These experiments were individually conducted by a Research Assistant.

Results & Discussion

As shown in Table 2, subjects rated popular and unpopular stories differently on all of the dimensions we measured. Subjects were more familiar with fables that attracted a higher number of Google hits than those that attracted fewer hits. This provided independent support for labeling of the stories mentioned on more Google-indexed websites as popular. This suggests that using Google to measure popularity of an idea is a valuable tool identified by Norenzayan *et al*. This should address lack of availability of data and should prompt more research in this area.

The results also provide some justification for the assumption that memorability had something to do with the popularity of the widespread Aesop fables, as subjects rated popular stories as more memorable than unpopular ones. These results are similar to those of Norenzayan *et al.* who also found that their subjects rated popular and unpopular Grimm Brother's tales to vary significantly along the dimensions of memorability, understandability, and likelihood of transmission.

Table 2: Mean subject ratings on various psychological variables as a function of whether a fable is popular or not.

Subject Ratings	Popular	Unpopular	t	p
Familiarity	–1.08	–2.49	14.51	<.001
Memorability	2.02	1.46	9.57	<.001
Likelihood of transmission	0.06	–0.68	10.16	<.001
Interest value	0.77	-0.09	11.27	< .001
Understandability	2.11	1.47	9.88	<.001
Moral Lesson	1.40	0.54	9.68	<.001
Cohesion	1.14	0.25	10.45	<.001

Unlike Norenzayan *et al*., who did not find significant differences between subject's ratings of the popular and unpopular Grimm Brother's tales along dimension of interest value and moral lesson, our subjects rated popular stories as significantly more interesting and as significantly more likely to have "a strong moral lesson" than unpopular stories. The difference between our results and theirs could be due to the differences in the materials used (Aesop's fables versus Grimm Brother's folks tales) or due the experimental design factors such as differences in sample size (32 subjects × 42 stories = 1342 sample points in our experiment versus 65 subjects × 6 stories = 390 sample points for their experiment).

Our results also support the hypothesis that motivated this experiment, namely, that popular and unpopular stories differ along the dimension of story cohesion. Subjects not only rated popular stories higher on the dimension of "having a strong moral lesson," they also thought that popular stories were harder to modify to make story's "moral lesson more apparent" as compared to unpopular stories. To see whether differences in story cohesion can account for differences in memorability between stories, we computed an aggregated cohesiveness measure by combining the subject ratings in response to the moral lesson and "needing modification" questions, and performed a correlational analysis of aggregated cohesiveness and story memorability. We found that cohesiveness was

strongly correlated with memorability (Pearson Correlation Coefficient $r = 0.71$, $N = 42$, $p < 0.001$). This suggests that cohesiveness of Aesop's fables can explain most of the difference in memorability among Aesop's fables while the number of counterintuitive concepts present in a story cannot. Furthermore, correlation between cohesiveness and memorability becomes even stronger when only counterintuitive stories are considered. For stories containing at least one counterintuitive concept, the correlation is stronger ($r = 0.75$, $N=34$, $p < 0.001$), it is even stronger for stories containing at least 2 counterintuitive concepts ($r = 0.81$, $N = 16$, $p < 0.001$), and it is higher still for stories containing 3 counterintuitive concepts of which there were only three ($r=0.90$, $N = 3$, $p < 0.001$). These results suggest that counterintuitive elements added to a story have to make sense in the context of the story for it to be memorable and that this is especially true as more and more counterintuitive concepts are added to a story. To the extent that the inclusion of counterintuitive concepts can be justified in the context of a story, there may not be a fixed upper limit to the number of counterintuitive concepts that can be included in a memorable story. A writer's creative ability to imagine counterintuitiveness-justifying contexts may be the real limiting factor. If the context in which counterintuitive concepts are embedded does not allow a reader to justify the inclusion of those concepts and make the story cohesive, then that story will not be remembered well. This also answers Norenzayan *et al.*'s question as to why despite all of their memorability advantages counterintuitive concepts never appear alone and are always communicated along with an even larger number of intuitive concepts. The paper suggests that this may be because a context built by intuitive concepts is needed to justify, make sense of, and give meaning to the counterintuitive concepts.

Conclusion

The results of studies reported here call into question the notions that (a) there is a single cognitively optimal template for all narratives, and that (b) inclusion of 2-3 counterintuitive concepts makes a story more memorable and hence more transmissible. This paper suggests that relationship between inclusion of counterintuitive concepts and memory for narratives may be more complicated than previously suggested. The experiments reported here support the hypothesis that inclusion of counterintuitive concepts can make a story more memorable only if they allow a reader to use her/his background knowledge to make the story more coherent. These results have important implications not only for those interested in understanding how elements of culture become widespread but also for those interested in designing memorable messages for influencing target audiences. Thus, cultural scientists cannot ignore the socio-cultural context at the time of diffusion if they want to understand how certain folktales came to be widely distributed in a population. Marketing professionals cannot just throw in a certain number of counterintuitive concepts (or more generally expectation violating or schema incongruent elements) into a message to make it more sticky. For such elements to add value to a message, one must carefully consider all aspects of the context which include both the cultural knowledge that members of the target audience bring to the table and the structure and content of the story to which these concepts are being added.

References

Barrett, J., & Nyhof, M. (2001). Spreading non-natural concepts. *Cognition and Culture, 1*, 69-100.

Boyer, P. (1994). *The Naturalness of Religious Ideas: A Cognitive Theory of Religion.* Berkeley, CA: University of California Press.

Boyer, P. (2001). *Religion Explained: The evolutionary origins of religious thought*. New York, NY: Basic.

Boyer, P., & Ramble, C. (2001). Cognitive templates for religious concepts. *Cognitive Science, 25*, 535-564.

Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructng inferences during narrative text comprehension. *Psychological Review, 101*, 371-395.

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English* London: Longman.

Kim, S. (1999). Causal bridging inference: A cause of story interestingness. *British Journal of Psychology, 90*, 57-71.

Kintsch, W. (1998). *Comprehension.* Cambridge, MA: Cambridge University Press.

Norenzayan, A., Atran, S., Faulkner, J., & Scaller, M. (2006). Memory and Mystery: The Cultural Selection of Minimally Counterintuitive Narratives. *Cognitive Science, 30*, 531-553.

Schank, R. C. (1999). *Dynamic Memory Revisited.* New York: Cambridge University Press.

Schank, R. C., & Abelson, R. (1977). *Scripts, plans, goals, and understanding: An inquiry into human knowledge*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Suls, J. (1983). Cognitive Processes in humor appreciation. In P. E. McGhee & J. H. Goldstein (Eds.), *Handbook of Humor Research*. New York: Springer-Verlag.

Townsend, G. (1867) Aesop's Fables, London, Routledge.

Trabasso, T., Suh, S., Payton, P., & Jain, R. (1995). Explanatory inferences and other strategies during comprehension and their effects on recall. In R. F. Larch & E. J. O'Brien (Eds.), *Sources of Coherence in Reading*. Hillsdale, NJ: Erlbaum.

Upal, M. A. (2005). *Role of Context in Memorability of Intuitive and Counterintuitive Concepts.* Paper presented at the in Proceedings of the 27th Annual Meeting of the Cognitive Science Society, Stressa, Italy.

Upal, M. A. (2010). An Alternative Account of the Minimal Counterintuitiveness Effect. *Cognitive Systems Research*, 11(1), 194-203.

Upal, M. A., Gonce, L., Tweney, R., & Slone, D. J. (2007). Contextualizing counterintuitiveness: How context affects comprehension and memorability of counterintuitive concepts. *Cognitive Science, 31*(3), 415-439.

Fortune Favors the Bold (and the Italicized): Effects of Disfluency on Educational Outcomes

Daniel M. Oppenheimer (doppenhe@princeton.edu)
Princeton University, Department of Psychology
Green Hall, Princeton, NJ 08540 USA

Connor Diemand Yauman (cdiemand@princeton.edu)
Princeton University, Department of Psychology
Green Hall, Princeton, NJ 08540 USA

Erikka B. Vaughan (ebvaugha@umail.iu.edu)
Indiana University, Department of Psychology
1101 E. 10th Street, Bloomington, IN 47405

Abstract

Research has shown that disfluency – the metacognitive experience of difficulty associated with a cognitive task – engenders deeper processing. Since deeper processing typically leads to better retention, this paper examined whether decreasing perceptual fluency of educational materials would improve retention. Study 1 found that harder to read fonts led to increased retention in a controlled laboratory setting. Study 2 extended this finding to real-world classroom environments. It appears as though perceptual disfluency can function as a desirable difficulty in education. Implications and caveats are discussed.

Introduction

It seems logical that to effectively communicate an idea, one should present it in a manner which is clear and easy to follow. Educators follow this principle when designing textbooks—the order, wording, and formatting is designed to help students read the information with minimal effort. Indeed, there is evidence to support the notion that students benefit from decreased cognitive demands when learning new concepts (Sweller and Chandler, 1994).

While it is commonly accepted that reducing extraneous cognitive load is beneficial to student learning, there is some research that seems to suggest there are exceptions to this rule. In fact, research shows that in certain instances, it may be beneficial to *increase* extraneous cognitive load (e.g. Bjork 1994). These aptly named "desirable difficulties" create additional cognitive burdens but nonetheless improve learning.

For example, in one experimental paradigm (Hirshman & Bjork, 1988), participants are asked to remember pairs of words, such as "bread : butter." Hirshman and Bjork found that requiring subjects to mentally generate missing letters in a word pair, such as "bread : b_tt_r," leads to improved recall performance over participants who read the word pair without any missing letters. Bjork extended this strategy to realistic educational settings, finding that students who complete simple fill-in-the-blank sentences are better able to retain information than students who read the same sentences with the key words filled in and underlined for them (Richland, Bjork, Finley, & Linn, 2005).

It seems counterintuitive that imposing unnecessary strain on students' limited cognitive capacity would actually improve performance, yet desirable difficulties seem to exploit nuances in our cognitive systems. Importantly, these instructional techniques *appear* sub-optimal. Without conscious recognition and implementation on behalf of cognitive psychologists and educators, it is likely that these techniques would not even be considered for use.

It is important to explore such techniques and seek out new methods of presentation that better reflect or utilize the way we process information. One such technique may come from explorations on the metacognitive experience of fluency—the subjective feeling of ease or difficulty which is associated with almost any mental task (Alter & Oppenheimer, 2009). For instance, a blurry photograph is disfluent because it is difficult to discern, a whisper is disfluent because it is difficult to hear, and a foreign word may be disfluent because it is difficult to pronounce. Fluency has been shown to influence our judgments in a variety of ways, including our judgments of truth, confidence, intelligence, or familiarity (for a review, see Alter & Oppenheimer, 2009). Importantly, recent studies have begun to explore how fluency influences cognitive processing in ways that might yield positive educational outcomes.

Recent work in fluency has demonstrated that when a problem is disfluent, people adopt a more deliberate processing strategy (Alter, Oppenheimer, Epley, & Eyre, 2007). In one experiment, participants were asked to read logical syllogisms and indicate whether they were true or false. Participants who read the syllogisms in a difficult to read (i.e. *disfluent*) font performed significantly better on the task than those who read the syllogisms in a clear, easy to read font. The authors replicated this result in three distinct cognitive domains. In this way, disfluency may be categorized as a desirable difficulty and can be used to improve student learning by encouraging them to select more accurate problem solving strategies.

Contemporary educational reform measures strive to create learning environments that encourage students to engage deeply with course content because of the numerous forms of evidence suggesting that deep processing increases learning. Most importantly, deeper processing facilitates later recall. For example, participants who are asked whether words appear in capital or lower case letters (low level of processing) do worse at later recall than participants who construct a rhyme for the words (moderate processing) or are asked to define the words (deep processing) (Craik & Tulving, 1975). Therefore, if disfluency facilitates deeper processing, then there is reason to expect that disfluent educational materials will lead to improved retention in the classroom.

The simplest and most standard fluency manipulation is a font manipulation. Information presented in an easy to read font is more fluent than *information presented in a difficult to read font*. The beauty of this manipulation is that it is so easy and cost effective to implement. To the extent that disfluency yields better learning outcomes, the intervention could be implemented on a wide scale with limited logistical or financial challenges.

The purpose of the present research is to empirically examine whether fluency can operate as a desirable difficulty to improve retention in classroom environments.

Study 1

First we aimed to show that disfluency led to better retention in a highly controlled laboratory environment. Twenty eight participants were recruited through the Princeton University paid subject pool and compensated $8 for their time. Participants' ages ranged from 18-33. Participants were given 90 seconds to learn about three species of aliens. Each alien species had seven features, for a total of 21 features that needed to be learned (see Figure 1 or examples of the features to be learned). This task was meant to approximate taxonomic learning that might occur in a biology classroom; fictional alien species were used so that participants had no prior knowledge that might contaminate results.

In the disfluent condition, the stimuli were presented in either 12 point Comic Sans MS 75% greyscale (see Figure 1a) or 12 point *Bodoni MT* 75% grayscale font. In the fluent condition, the stimuli were presented in 16-point Arial 100% black font (See Figure 1b). A between-subjects design was used, such that each participant was only exposed to one font. As is evident from the examples below, while the disfluent text is obviously harder to read than the fluent text (when they are presented side by side) in a between subject design reader's in the disfluent condition were unlikely to even consciously notice the added difficulty the disfluent text engendered.

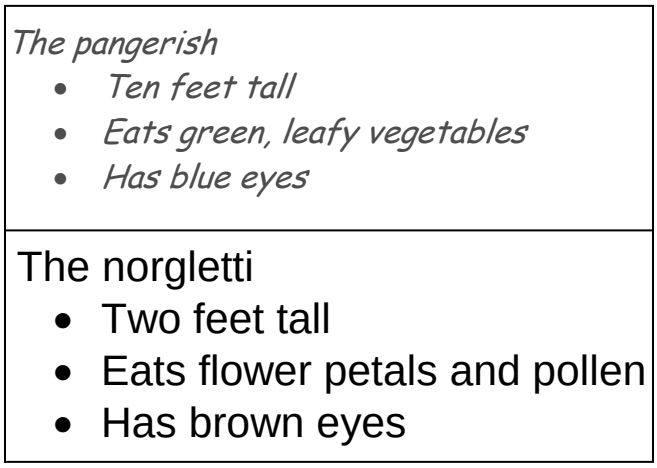


Figure 1: Example stimuli from Study 1. The top panel shows the disfluent font, and the bottom panel shows the fluent font.

After studying the material for 90 seconds, participants were distracted for 15 minutes with unrelated tasks. Participants' memory for the material was then tested. For each participant, seven of the features were randomly asked about. For example "how tall is the pangerish?" or "what color eyes does the norgletti have?"

One outlier was eliminated from consideration for being more than 3 standard deviations from the mean. Participants in the fluent condition were accurate 72.8% of the time. Meanwhile, participants in the disfluent condition successfully remembered the information 86.5% of the time. This difference was statistically significant ($t(26) = 2.3$, $p < .05$). There were no differences in retention between the different disfluent fonts (Comic Sans vs. Bodoni), suggesting that it was not the specific font that led to the difference, but rather the disfluency. In sum, after a 15-minute delay, participants in the recalled nearly 15% more information when the material was presented disfluently than fluently. Moreover, as learning time was constrained, this cannot be due to longer study times for the disfluent materials, which suggests that instead more effective learning strategies were adopted.

While this provides strong preliminary evidence that fluency could be a desirable difficulty in education, there are several reasons why we might be concerned about its generalizability to actual classroom environments. First, the materials we used, while tightly controlled, were not the sorts of materials that would be used in real classroom settings. Different types of materials might elicit different effects. Second, while the effects in Study 1 persisted for 15 minutes, the time between learning and testing is typically much longer in the real world.

Further, while paid laboratory participants may be willing to persist in the face of challenging fonts for 90 seconds, added difficulty may undermine motivation for actual students. Students may just give up, rather than deeply processing the material – particularly as the semester progresses and stress levels rise. Therefore, we ran a large

field study to determine whether these results would persist outside of the lab.

Study 2

222 high school students (ages 15-18) from a public school in Chesterland, Ohio participated in the study. This school accommodates approximately 930 students from grades 9-12 and reported a 98.6% graduation rate in 2008. The school's grades 9-12 are taught by 54 teachers.

Classes were selected for this research using the following criteria: the same teacher must have been teaching at least two classes of the same subject and difficulty level with the same supplementary learning material (PowerPoint presentations or handouts). Six classes met these criteria and agreed to participate. These classes were AP English, Honors English, Honors Physics, Regular Physics, Honors US History, and Honors Chemistry.

The different sections of each class were randomly assigned to either a disfluent or control category. Teachers were instructed to send all relevant supplementary learning materials to the experimenters prior to distributing them to students. At no point did the experimenters ever have face-to-face contact with the students or teachers; editing was done by proxy in Princeton, New Jersey. The fonts of the learning material in the disfluent category were either changed to **Haettenschweiler,** *Monotype Corsiva,* or *Comic Sans Italicized* or copied disfluently (by moving the paper up and down during copying) when electronic documents were unavailable. In the control category, no edits were made to the materials before returning them to teachers. The font size of the supplementary material was not changed unless the original size when converted to disfluent font made the font illegible, in which case the font size was increased until it was readable. One teacher refused to administer Haettenschweiler and so that class was changed to Comic Sans Italicized.

No other changes were made to the students' learning environments, materials, curricula, or to the teachers' classroom routine. To determine the effects of disfluency, the results of the normal assessment tests for the class were collected and analyzed.

The z-scores of the students' test performance were used as a common metric to compare students across different courses. As shown in table 1, average z-scores of the students were higher in the disfluent condition than in the control.

An independent samples t-test of the average z-scores revealed a significant improvement of the students' test scores in the disfluent condition ($t(220) = 3.38$, $p < .001$): students in the disfluent condition scored higher on their tests ($M = .164$, $SD = .1.03$) than those in the control ($M = -.295$, $SD = 1.05$). There were no reliable differences between the different disfluent fonts. That is, it was not the specific of the font that mattered, but rather the fact that it was disfluent.

	Control	Disfluent
AP English	-.058	.135
Honors English	-.175	.131
Physics Honors	-.251	.215
Physics Normal	-1.13	.42
History	-.177	.112
Chemistry	.023	-.017
Total	-.295	.164

Table 1: Average z-score for fluent and disfluent supplementary materials across the 5 usable classrooms. Note that the z-scores do not sum to 0 across conditions because of unequal sample sizes by condition.

The effects of different kinds of disfluent material were examined using a two-level ANOVA to compare the effects of disfluent worksheets and PowerPoint presentations. This test revealed that the PowerPoint presentations were significantly more effective than the documents in improving student performance when presented in a disfluent format ($F(1, 184) = 9.38$, $p < .01$). However it is difficult to read too deeply into this latter finding, as the only classes that used powerpoint materials were the physics classes. As such, we cannot know if the difference was due to the type of material that was being studied, or the manner in which it was presented. Nonetheless, the difference highlights possible future avenues of exploration.

Discussion

In two studies we showed that making the text disfluent by using a hard to read font improved learning. In Study 1, participants recalled 14% more material when the material was initially presented in a disfluent font. In Study 2, students performed better on exams in actual classrooms the fonts of the supplementary materials were harder to read. This occurred for both science and non-science courses, and for different difficulty levels (AP, honors, and regular). This provides strong preliminary evidence that disfluency can indeed function as a desirable difficulty in educational settings.

There are, however, some important caveats that need to be considered in relation to these findings. First, while a small amount of disfluency was able to improve performance, at some level disfluency will necessarily impair functioning. After all, if the font is impossible to read, then the information cannot be encoded, let alone retained. It is unclear from these studies what the optimal level of disfluency is, nor the relative detriment that being overly disfluent might engender.

Secondly, there is the issue of adaptation. One reason that disfluent text might lead to better retention is that it serves as an alarm signal that this material is challenging and merits extra consideration (c.f. Alter et al., 2007). To the extent that students become used to disfluency, they might

no longer adopt deeper processing strategies when exposed to hard to read font. Study 2 was limited to a single semester's worth of materials for logistical reasons. It is unclear whether these effects would persist over longer periods of time.

Third, a large literature has demonstrated that disfluent materials are liked less than fluent materials (for a review see Alter and Oppenheimer, 2009). It may be that while students retained the information better under disfluency, they also liked the material less. This could mean that they are less likely to pursue further studies in the topic (e.g. in college) or are otherwise demotivated in subsequent educational situations. Of course, it is also possible that the increased effort necessary to engage with this material will create cognitive dissonance, which will cause them to like the material more (c.f. Cooper, 2007). This is an empirical question, that will require additional research to resolve.

Fourth, it is quite possible that there are moderators for this effect that these initial studies did not detect. Other forms of desirable difficulties have been shown to be moderated by factors such as the nature of the materials (McDaniel et al., 2000) the nature of the testing (Thomas & McDaniel, 2007), and the abilities of the learner (Macnamara et al., 1996). One could imagine that less motivated students from a less successful school might be more inclined to give up on the material rather than persist and encode it more deeply. Future investigation should look into these issues.

Despite these potential drawbacks, disfluency is a promising form of desirable difficulty because it requires no retraining of teachers, no restructuring of curricula, and can be implemented with minimal cost. Given the results of the present studies, it seems worthwhile to investigate disfluency as an educational intervention further.

Acknowledgments

Thanks to Dr. Beth Yauman, a Princeton Undergraduate research fellowship and the PSURE research fellowship for supporting the second and third authors respectively. Thanks to Michelle Ellefson, Alan Castel , Adam Alter, Anuj Shah, Jeff Zemla, and the Opplab for advice and support.

References

Alter, A.L., & Oppenheimer, D.M. (2009). Uniting the tribes of fluency to form a metacognitive nation. Personality and Social Psychology Review. *Personality and Social Psychology Review. 13(3),* 219-235.

Alter, A. L., Oppenheimer, D. M., Epley, N., & Eyre, R. N. (2007). Overcoming intuition: Metacognitive difficulty activates analytic reasoning. *JEP: General*, *136*, 569-576.

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe and A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp.185-205). Cambridge, MA: MIT Press.

Cooper, J. (2007). *Cognitive dissonance: 50 years of a classic theory*. London: Sage publications.

Craik, F. I. M., & Tulving, E. (1975). Depth of processing and retention of words in episodic memory. J*EP: General, 104,* 268-294

Hirshman, E., & Bjork, R. A. (1988). The generation effect: Support for a two-factor theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*, 484-494.

McDaniel, M.A., DeLosh, E.L., & Merritt, P.S. (2000). Order information and retrieval distinctiveness: Recall of common versus bizarre material. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 1045-1056.

McNamara, D. S., Kintsch, E., Butler-Songer, N., and Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction* , *14* (1), 1-43.

ODE - Home. Web. 21 Jan. 2010. <http://www.ode.state.oh.us>.

Richland, L. E., Bjork, R. A., Finley, J. R., & Linn, M. C. (2005). Linking cognitive science to education: generation and interleaving effects. In B. G. Bara, L. Barsalou and M. Bucciarelli (Eds.) *Proceedings of the Twenty-Seventh Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum.

Sweller, J., & Chandler, P. (1994). Why some material is difficult to learn. *Cognition and Instruction*, *12*, 185-233.

Thomas, A.K., & McDaniel, M.A. (2007). Metacomprehension for educationally relevant materials: Dramatic effects of encoding-retrieval interactions. *Psychonomic Bulletin and Review, 14,*212-218.

Relational Versus Attributional Mode of Problem Solving?

Svetoslav Bliznashki (valsotevs@gmail.com)
New Bulgarian University, 2 Montevideo Street
Sofia 1618, Bulgaria

Boicho Kokinov (bkokinov@nbu.bg)
New Bulgarian University, 2 Montevideo Street
Sofia 1618, Bulgaria

Abstract

We argue that the concept of relational priming (e.g. Schunn 1996, Day 2007) can be extended from priming of specific relations to generating a cognitive state during which subjects are particularly likely to encode and use relations. We conducted an experiment in which three groups of subjects did different tasks before a target matching to sample task was introduced which contrasted a relationally versus an attributionally similar alternative. Subjects in one condition were asked to solve tasks involving relational reasoning while subjects in another condition were asked to tasks involving only attributes. As expected subjects in the first condition were more likely to pick up the relationally similar alternative while in the second condition the results reversed relative to a control group. In conclusion we argue that this study shows that encoding of relations can be a subject to unconscious context influence.

Keywords: relational priming; context dependence; encoding of relations, cognitive state

Introduction

Since Gentner's Structure Mapping theory (Gentner 1983) a great deal of research has been concentrated upon analogical reasoning in terms of mapping of higher-order relations. Although the mechanisms employed by the mapping process have been extensively studied little is currently known about the nature of the processes involved in relational encoding. This study attempts to scratch the surface of this complex matter by asking the question of whether the process of relational encoding is subject to certain external and internal context influences. While the answer to this question certainly would not reveal the nature of the encoding process it would hopefully tell us something about certain specific aspects of its functioning.

Currently there is some agreement that the phenomenon of relational priming exhibits a somehow automatic (i.e. not subjected to voluntary control, external influence and conscious experience) nature (e.g. Kokinov 1996, Schunn 1996, Day 2007, Hristova 2009)[1]. The abovementioned studies employ different methodologies ranging from naturalistic-like settings (Schunn 1996) to Stroop-like interference Reaction Time paradigms (Hristova 2009). There is however a common thread among these research projects – the use of specific relations. In other words all these (and other) studies concentrated on exploring the relational priming by using concretely represented, nameable relations. The same naturally holds true for relational priming in psycholinguistic research (e.g. Gagne 2005, Estes 2006).

Instead of continuing this well established line of research we concentrated on the question of whether a global cognitive state can be induced in which people are more likely to encode relations in general. It can be said that we are still concerned with relational priming but we employ a rather broad, holistic and abstract definition of the phenomenon.

We hypothesized that subjects confronted with tasks explicitly involving relational reasoning will subsequently be more likely to continue this style of reasoning when dealing with completely different tasks. We also hypothesized that subjects forced to encode and use attributes of objects will be considerably less likely to encode (and use) relations in subsequent tasks. In other words we argue that not only specific relations can be primed with similar other relations but also a "relational mode of thinking" can be induced by use of specific task requirements. Thus we claim that relational priming (and consequently "attributional priming") is a much more complicated and abstract phenomenon than currently conceived by traditional research in the area.

Another hypothesis related to the current study concerns the subjects' ability to cope with the particular priming task. Since we argue that task requirements can possibly induce a particular cognitive state it follows that the degree to which this actually happens should depend upon a subject's particular ability to successfully cope with the task at hand (the priming task).

Experiment

In the current experiment we tested three different groups of subjects in order to see whether prior tasks influence significantly relational and attributional reasoning during a target task. In the first condition subjects solved six different mental rotation tasks, in the second condition subjects solved six items from Raven Progressive Matrices test (e.g. Raven 2003) and in a third condition no task preceded the target task. The three groups are called attributional, relational and control conditions respectively.

[1] But see Spellman et al. (2001) whose results indicate that relational priming took place only when participants were explicitly instructed to pay attention to the relations existing between the stimuli (words) involved in the studies.

The target task was a single matching to sample task borrowed from Medin et al. (1990). During this task subjects were required to choose the more similar from two alternative figures to a target figure. One of the alternatives embodied a unique common relation with the target (we called this one the relational alternative) while the other shared a unique attribution with the target (we called it the attributional alternative).

We hypothesized that subjects in the relational condition would be more likely to pick up the relational alternative in the target task (compared to the control group where no priming task was present) because the Raven Progressive Matrices test requires subjects to encode and map complex higher-order relations. On the other hand since the mental rotation task involved dealing with attributes and first order relations between parts of objects we expected that subjects from the attributional condition would be more likely to choose the attributional answer to the matching to sample task (again compared to the non primed control condition). As mentioned above we also hypothesized that there would be a correlation between subjects' levels of performance on the prior tasks and the degree of subsequent relational priming. Moreover since subjects in the attributional task were expected to be les likely to give relational answers to the target task we expected a negative correlation between levels of performance during the priming task and the proportion of relational answers to the target task in the attributional condition. By the same logic we expected a positive relationship to exist in the relational condition.

Design. A simple between group design was employed which involved three independent groups of subjects allocated to the attributional, the relational and the control conditions. The three levels of our independent variable were defined by the task the subjects in the respective condition had to solve before the target matching to sample task. The dependent measure was defined as whether a given subject gave a relational or attributional answer to the target task. The target task was the same for all participants.

Stimuli. The stimuli for the attributional condition consisted of six mental rotation tasks. Each task involved sixteen versions of a particular letter from the Latin alphabet. Thus there were six letters in that condition and each letter appeared sixteen times. Each version of a letter was presented in a rotated position. For eight of the versions it was possible to obtain the original letter via mental rotation (these represented the so called true versions of a particular task since the subjects' task was to indicate whether the particular version could or could not be rotated in order to arrive at the original letter) and for the other eight versions it was impossible to do so for these versions were rotated mirror images of the original letter (these were called the false versions). Each letter (both the true and the false versions) was rotated at eight different angles. The degrees of rotation were 40, 80, 120, 160, 200, 240, 280 and 320 degrees. The six letters used for each of the six tasks were Z, R, F, N, P, S. Each individual task was represented as the sixteen versions of a particular letter arranged in a 4x4 matrix printed on an A4 portrait sheet of paper. The order of the true and false versions as well as the order of the eight different angles of rotations was randomized across the six tasks. The order of the six tasks was randomized across participants. Since subjects were required to make a judgment for each letter version in each task (i.e. subjects were asked to indicate whether a particular version was a rotated original letter or a rotated mirror image of the original letter) there were 6x16=96 judgments made by each participant in the attributional condition. A sample of three letters is presented at figure 1. The top three letters represent instances of the false alternatives and the bottom three letters represent instances of the true alternatives.



Figure 1. Examples of the mental rotation task.

Subjects from the relational condition were presented with six of the Raven Progressive Matrices items. These items were the odd numbered items from series E (the last series) from the test. Thus subjects had to solve items E1, E3, E5, E7, E9 and E11. The items were presented in this ascending order for all participants in this condition. Subjects in this condition were asked to fill the blank in each item with one of the options available at the bottom of the page. The original instruction from the test was given to each participant. The original test panes were used.

The target matching to sample task presented to the participants at the end of the experiment was borrowed from Medin et al. (1990). It is depicted in figure 2. The target is at the top of the figure and is denoted with T. The two alternatives are denoted with B1 and B2 respectively. The subjects' task was to indicate which one of the two options was more similar to the target. As already mentioned the B1 option shared a unique attribute with T (a checked circle) while the B2 option shared a unique relation with T (same shading of the objects).

Procedure. In both the attributional and the relational conditions subjects were given a maximum amount of time of one minute for each individual task (one item from the Raven test or one mental rotation task consisting of sixteen individual versions of a letter). In the attributional condition subjects were instructed to make as many accurate judgments as possible for each task for one minute. In the relational condition subjects were instructed to try to solve each item correctly for one minute. Prior to the experiment subjects from the attributional condition were given a practice trial consisting of sixteen versions of the letter L.

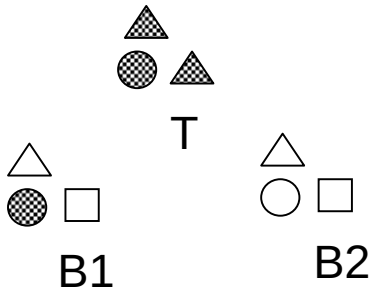


Figure 2. Example of the matching to sample task.

The subjects from the relational condition were given the original instruction from the Raven test as well as items C9, D3 and D8 as complementary practice trials. In each condition after the instruction the experimenter encouraged the participants to try to solve the practice task all by themselves and explained their errors as well as the correct solutions when needed after the one minute maximum time interval. After the experimenter was convinced that subjects understood the procedure the real study began.

The experimenter used a stop watch in order to keep track on time for each task.

In both the relational and the attributional conditions participants indicated their responses verbally and the experimenter wrote down their answers on a scoring sheet. In the attributional condition subjects were instructed to indicate whether a version of a letter could be rotated to its original position by moving from the top row down and moving from left to right within a particular row of a given matrix of sixteen versions of a letter. In case a participant failed to answer to all versions of a mental rotation task within a minute the sheet containing the matrix was removed out of her sight but the participant was asked to try to guess the correct answers for the remaining versions of the letter. Similarly in the relational condition if a person didn't answer to a Raven item within one minute the pane was taken out of her sight but the participant was asked to try to guess the correct answer anyway.

A thirty seconds interval separated the six priming trials from each other in each experimental condition.

Immediately after the end of the initial stage the target stimulus was presented in an ostensibly unrelated task and the subject was asked to indicate whether B1 or B2 option was "more similar" to the T figure (see figure 2). No time limit was present during the final task.

The subjects from the control condition proceeded immediately to this final stage of the experiment, i.e. they were not involved in any prior task.

After the participants indicated their answers they were asked whether they spotted the relational similarity between the target and the B2 option. After their answer was written down by the experimenter the subjects were debriefed and the experiment finished.

Note that the two priming task are both quite different from the target task. Thus it seems rather unlikely that some specific features of the priming tasks may have influenced subjects' judgments during the final matching to sample task.

Subjects. 110 students from New Bulgarian University participated in the study for partial course credit. Thirty five participated in the attributional condition, thirty five participated in the relational condition and forty participated in the control condition. Overall there were 62% females and 38% males in the study which were allocated proportionally to all three conditions.

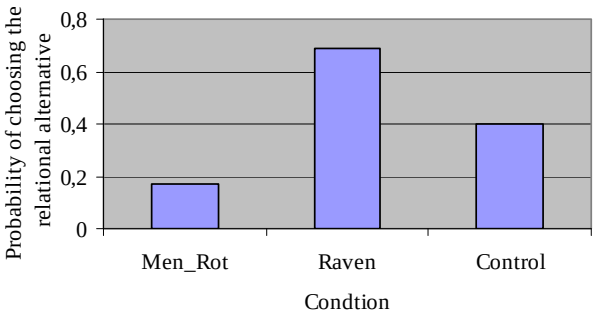
Results and Discussion. Table 1 below shows the raw number of subjects within each condition which gave the attributional and the relational answers to the target task. Numbers in parentheses represent the respective percentages.

table 1.

	#Attrib. Answers	# Rel. Answers	Total
Attrib. Cond.	29 (83%)	6 (17%)	35 (100%)
Rel. Cond.	11 (31%)	24 (69%)	35 (100%)
Contr. Cond.	24 (60%)	16 (40%)	40 (100%)

The results are summarized in figure 3 below. The bars represent the proportion of people giving the relational answer in each condition.

figure 3



As can be seen the data are in line with our hypotheses: subjects from the relational condition were more likely to pick up the relational answer during the matching to sample task compared to subjects in the control condition. Also the subjects from the attributional condition picked up the attributional answer more frequently compared to the baseline control condition. In order to asses the significance of our results we conducted a series of statistical analyses. First we fitted a logistic regression model to the data with our three experimental conditions treated as a single

categorical predictor and subject's answer as a categorical dependent variable (a relational answer was coded as 1 while an attributional answer was coded as 0 for each participant). The model including our independent variable significantly outperformed the null model (including only the intercept): Chi Square of Likelihood Ratio Change (2) = 20.049, $p<0.001$. Thus we see that there is a highly significant effect of our independent variable. Since we defined our control condition as the reference condition for the analysis the b coefficients of the model represent the difference between the other two conditions to the control condition. These were b = -1.17 (1), p = 0.034 for the attributional condition and b = 1.19 (1), p = 0.015 for the relational condition. Thus we see that the relational task significantly increases the probability of relational answer while the attributional task decreases it relative to the control condition.

The pseudo R^2 estimate (Nagelkerke) for the effect of the independent variable was equal to 0.224 – a reasonably high estimate.

Since both conditions were significantly different from the control condition and since their coefficients were with opposite signs it logically follows that the two experimental conditions were significantly different from each other.

In order to further support our results we conducted a series of chi squared analyses. First we assessed the significance of our independent variable as a whole. As with the regression analysis the results were highly significant – chi square (2) = 19.109, $p<0.001$. We proceeded with three post hoc comparisons which compared the proportions of relational answers between all three groups. The results showed that the relational and the attributional conditions were significantly different from each other – chi square (1) = 18.9, $p<0.001$. Both the conditions were also significantly different from the control condition – chi square (1) = 4.705, p = 0.03 for the difference between the control and the attributional conditions and chi square (1) = 6.122, p = 0.013 for the difference between the control and the relational conditions. Thus we see that all our conditions exhibited different proportions of relational answers[2]. Looking back at figure 3 we see that subjects from the relational condition were most likely to give a relational answer to the target task while those from the attributional condition were least likely to do so. The control condition was somewhere in between the other two.

These results strongly support our main hypothesis about the possibility to induce a cognitive state which enhances subjects' ability to encode relations. However there still exists the possibility of people encoding the relation embodied in the target task with approximately equal frequency but for some reason being more prone to choose it in the relational condition. When asked about whether they had spotted the "same shading" relation, however, only two participants from the control condition claimed they had and only one participant from the relational condition did so (these numbers refer only to subjects who gave the attributional answer to the target task, of course; all subjects who responded relationally reported spotting the unique relation). Thus such an alternative explanation seems highly unlikely. Overall the results support our hypothesis of relational priming being an abstract and profound phenomenon with deep impact on cognitive functioning.

Previously we stated our additional hypothesis that subjects' ability to cope with the priming tasks at question should correlate with the degree to relational priming they exhibit. Moreover we hypothesized that there should be a positive correlation between the number of correctly solved trials in the relational condition and the proportion of relational answers to the target task and a negative correlation between the number of correctly solved trials in the attributional condition and the proportion of relational answers to the target task. In both conditions we expressed the number of correctly solved trials as percentages from the overall number of trials. The overall number of trials was six in the relational condition and ninety six for the attributional condition. We calculated the point biserial correlations between these measures and the dependent variables separately for each experimental condition. The results indicated a significant positive relationship in the relational condition – r_{pbis} = 0.36, p=0.018 (one tailed). There was also a significant negative relationship in the attributional condition – r_{pbis} = -0.31, p=0.037. Thus it seems that our hypothesis is supported from the data[3]. These results seem reasonable since we can not expect subjects to be primed by task requirements if they are unable to fulfill the particular task.

Conclusion

In this study we successfully demonstrated that relational priming extends beyond the use of particular relations. It

[2] Technically speaking we should decrease our significance levels when performing these kinds of multiple comparisons in order to keep the type 1 error probability equal to 0.05 for all comparisons simultaneously. This was achieved by adopting a 0.033(3) level of significance for each comparison (we assumed a directed alternative hypotheses because we had strong prior expectations about the results from the study). We see that all our comparisons fall below this level of significance. Here we reported the probabilities from two-tailed tests which should be divided by a factor of two in order to obtain the one-tailed probabilities which fall way below the adopted significance level (although the chi square tests are regarded as inherently two-tailed a test of equality of proportions can be performed which has a one-tailed version; for the case of 2x2 tables the equality of proportions and the chi square tests are mathematically equivalent). Thus we can be confident that all three groups differ significantly from each other.

[3] We tested this hypothesis further by conducting logistic regression analyses with the percentage of correct responses as independent covariate and the response to the target task as a dependent variable. In the case of the relational condition the full model significantly outperformed the null model - Chi Square of Likelihood Ratio Change (1) = 4.63, p=0.031. In the case of the attributional condition the results were marginally significant - Chi Square of Likelihood Ratio Change (1) = 3.833, p=0.05. These results however test a two tailed hypothesis.

seems fertile to talk of cognitive states which enhance subjects' ability to encode relations. Moreover it appears that such cognitive states may be induced through external context factors. We consider our results relevant to the area of analogical mapping research since encoding of relations is obviously a prerequisite for subsequent mapping and transfer.

We also demonstrated that individual differences in terms of subjects' ability to cope with a particular task is a relevant variable for it significantly mediates the task's ability to induce the desired cognitive state for a particular subject.

We would like to stress that items from the Raven's test didn't embody the "same shading" relation of the target task and thus could not have possibly primed the relational answer directly. Also the tasks from the mental rotation condition did not involve any different or specific textures and consequently could not have primed the uniquely shared attribute of the attributional option of the matching to sample task.

Also few of the subjects who chose the attributional answer claimed to have spotted the uniquely shared relation so our results are likely to have arisen from influencing relational encoding rather than from manipulating subjects' relation vs. attribute preference.

Prior to the experiment we felt that using many matching to sample tasks (which would have enabled us to use parametric statistical analyses on one hand and would have granted our results with additional validity on the other) was not as warranted as it may appear at first. The reasons for this are straightforward – we suspected that once a particular subject have spotted the unique shared relation in one item they would search and easily find these relations on subsequent items. Thus we were afraid that no matter how many items we used our dependent measure would basically degenerate to a dichotomy. In such a case using parametric statistical analyses would be faulty and misleading. Another reason for avoiding the use of several different matching to sample tasks was that we speculated that our priming effect may exhibit a limited time duration and thus only the first few items would experience the effect. In case of counterbalancing the order of items across participants this effect might easily be obscured if we decided to run some comparisons at the items level.

Trying to replicate our results with different target item(s) and different priming tasks is a part of our future research agenda. Another part is exploring the duration of the priming effect.

Acknowledgements

This research was financially supported by the ANALOGY project (NEST program, contract 29088) funded by the EC.

We are grateful to Douglas Medin, Robert Goldstone and Dedre Gentner for publishing and making available online the stimulus material part of which we used in our study.

We would also like to thank Veselina Feldman and Kalina Bojadjieva whose usage of the Raven Progressive Matrices test as filler in another experiment dealing with analogical reasoning inspired a discussion which led to the current study.

References

Day, S. & Gentner, D. (2007). Nonintentional Analogical inference in text comprehension. *Memory and Cognition, 35 (1), 39 – 49.*

Estes, Z. (2003). Attributive and relational processes in nominal combination. *Journal of Memory and Language, 48, 304 – 319.*

Estes, Z., and Jones, L. L. (2006). Priming via relational similarity: A copper horse is faster when seen through a glass eye. *Journal of Memory and Language, 55, 89 – 101.*

Gagne', C. L. (2001). Relation and lexical priming during the interpretation noun-noun combinations. *JEP: Learning, Memory and Cognition, 27, 236 - 254.*

Gagne', C. L. (2002). Lexical and relational influences on the processing of novel compounds. *Brain and Language, 81, 723 – 735.*

Gagne ', C. L., Spalding, T. L., & Ji, H. (2005). Re-examining evidence for the use of independent relational representations during conceptual combination. *Journal of Memory and Language, 53, 445 – 455.*

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science, 7(2), 155 – 170.*

Hristova, P. (2009). Unintentional and unconscious analogies between superficially dissimilar but relationally similar simple structures. In: *Proceedings of the Second International Conference on Analogy.*

Hristova, P. (2009). Unconscious analogical mapping? In: *Proceedings of 31st Annual Conference of the Cognitive Science Society.*

Kokinov, B. & Yoveva, M. (1996). Context effects on problem solving. In: *Proceedings of the 18th Annual Conference of the Cognitive Science Society.* Erlbaum, Hillsdale, NJ.

Medin, D. L., Goldstone, R. L., & Gentner, D. (1990). Similarity involving attributes and relations: Judgments of similarity and difference are not inverses. *Psychological Science, 1,* 64-69.

Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological Review, 100(2), 254 – 278.*

Raven, J., Raven, J.C., & Court, J.H. (2003). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 1: General Overview.* San Antonio, TX: Harcourt Assessment.

Schunn, C. D. & Dunbar, K. (1996). Priming, analogy and awareness in complex reasoning. *Memory and Cognition, 24, 271 – 284.*

Spellman, B. A., Holyoak, K. J. and Morrison, R. G. (2001). Analogical priming via semantic relations. *Memory and Cognition, 29, 383 – 393.*

Executive Control in Analogical Mapping: Two Facets

Anna Chuderska (ania.chuderska@gmail.com)
Institute of Psychology, Jagiellonian University
al. Mickiewicza 3, 31-120 Krakow, Poland

Abstract

In recent studies, analogy-making has been shown to depend on the ability to resist interference within working memory (WM). Less evidence refers to the other facets of executive control (EC), especially to the goal-directed selection of relational information. In this study, the load on two above mentioned EC functions and on WM capacity was manipulated in a single picture mapping task. Next to replicating the previous findings on the importance of dealing with distracter interference within WM, the current results demonstrate that the efficiency of relational mapping also depends on the goal-directed search for WM input, and that these two EC functions may be dissociable in mapping. Moreover, it was found that the impact of distraction can be linked to whether relations, in which distracters occur, have not exceeded WM capacity.

Introduction

Analogical reasoning is a flagship example of the human ability to flexibly form and manipulate explicit representations of structure (Hummel & Holyoak, 2003). Making an analogy requires identifying systematic relational correspondences between two analogs (e.g., situations), irrespective of superficial similarities (if they conflict with relational ones) or differences (especially, if they are huge) between them (e.g., Gentner, 1983). This structure-mapping process allows one to infer new goal-relevant information about one analog (*target*) from the second analog (*source*). Thus, analogy is an important tool for dealing with novelty and one of the major vehicles of human intelligence (e.g., Holyoak, 2005).

The inherent computational challenges of processing relational representations (Doumas & Hummel, 2005), and the fact that variance in the efficiency of analogical reasoning is only partially explicable by knowledge accretion (e.g., Doumas, Morrison & Richland, 2009), has made many researchers and theorists postulate that the emergence of analogy is underlain by the efficiency of some constitutional cognitive capacities (or parameters). Of these, working memory (WM) was considered to be the most important (see Morrison, 2005 for review). Yet, explaining analogy-making by WM constraints seems to be quite intricate.

More than WM Capacity

WM is a capacity-limited system responsible for active maintenance, rapid access and easy updating of goal-relevant information (Cowan, 2005). If WM is overloaded by a parallel task (Waltz, Lau, Grewal & Holyoak, 2000), impaired by brain damage (Waltz et al., 1999), or if the number of variables interacting in relational representtation grows (Halford, Baker, McCredden, & Bain, 2005), relational reasoning becomes less efficient. According to relational complexity theory, the load of relational representation on WM increases exponentially with the number of interacting variables that must be concurrently manipulated (i.e., relationally integrated). Human WM is probably typically limited to the parallel processing of up to one quaternary relation, that is a relational representation with four variables (Halford, Wilson & Phillips, 1998).

However, when reflecting on the limited nature of WM capacity, it is important to understand internal cognitive constraints on the proper selection of WM input, by means of which humans single-handedly, but with various degrees of success (e.g., Chuderska & Chuderski, 2009), abstract structural similarity between analogs from other structural and (sometimes very compelling) semantic information. It seems plausible that relational integration might be influenced by the efficiency of earlier goal-driven attentional selection (see Awh, Vogel & Oh, 2006 for a discussion of attention as the "gatekeeper" for WM), or by the efficiency of managing subsequent reasoning steps (e.g., Carpenter, Just & Shell, 1990), which are necessarily isolated out for reduction of complexity (Halford et al., 1998).

On the other hand, the content of variables integrated in WM may perceptually or semantically conflict with the structural information they convey (e.g., Markman and Gentner, 1993). Since processing many distracters leads to no success, the need to deal with distraction within WM, while analogizing, seems indispensible (e.g., Viskontas, Morisson, Holyoak, Hummel, & Knowlton, 2004).

The potential causes for processing irrelevant information by WM, or doing it inefficiently, are delineated in LISA – an artificial neural network model of relational reasoning (Hummel & Holyoak, 2003). LISA dynamically binds roles (i.e., variables) and fillers (i.e., their content) into relations by the synchrony of firing their distributed semantic (featural) and localist (structural) representations. The model contains an intrinsic capacity limit, since only a confined number of such role-filler bindings can oscillate cleanly asynchronously in one processing cycle. The weaker the inhibitory competition between active units, the less role-filler bindings are cleanly discriminated. The strength of inhibitory competition between propositions in problem representation also determines which of them will enter WM and in what order; this is critical for mapping performance (Kubose, Holyoak & Hummel, 2002). The weaker the inhibition, the less reliance LISA has on the importance assigned to propositions, and the less accurate will be its eventual mapping.

Thus, an important source of cognitive constraints in relational reasoning might come from the effectiveness of executive control. Executive control (EC) can be defined as a set of cognitive processes that, instead of representing mental states directly, influence and organize such states in the context of some internal goal. Recent theories assume that EC is an emergent process arising from the dynamic interaction of several independent, elementary control mechanisms (Braver, Gray, & Burgess, 2007; Engle & Kane, 2004). There is some evidence that these functions significantly correlate with abstract reasoning (see Chuderski & Nęcka, 2010, for a review).

Executive Control in Analogical Reasoning

The fact that the maintenance and proper application of a reasoning goal is critical for analogical reasoning might be inferred from findings which show that the frequency of recognizing relational similarity is higher when multiple, instead of single, objects are to be mapped across analogs (Markman and Gentner, 1993; Waltz et al., 2000). Such a manipulation might make the goal of relational processing more salient to participants and aid (or substitute) selection of what should enter WM for structural alignment. It could also be hypothesized that the overriding initial mappings, if they turn out to be incorrect (Keane, 1997), might call not only for inhibition, as proposed in LISA, but also for some goal management mechanisms. More directly, it was shown that mapping performance correlates with most of the proposed executive functions, with three of them (WM updating, switching, and dual-tasking) being accounted for through the monitoring and application of goal and through response inhibition (Chuderska & Chuderski, 2009).

Another function of control within analogical reasoning relates to resolving conflicts and coping with (distracter) interference. For example, Gray, Chabris, and Braver (2003) observed that brain activity in neural structures, recruited by a high-interference condition of a WM updating task, correlated with relational reasoning performance. Some evidence for links between abstract reasoning tests and response inhibition and interference resolution was reviewed by Dempster and Corkill (1999). If superficially similar objects are placed in different relational roles (i.e., are cross-mapped) in structures that are to be mapped, effective interference resolution seems necessary to overcome the observed relational mapping impediment, (e.g., Markman & Gentner, 1993). Cho, Holyoak, and Cannon (2007) manipulated the level of internal complexity and interference of a simple analogical mapping task, demonstrating that young participants' reaction times overadditively increased with relational complexity and interference. Similar decreases in performance by manipulating these two factors were observed in older adults (Viskontas et al., 2004). Richland, Morrison and Holyoak (2006) found that as children get older they are more efficient in dealing with both relational complexity and distraction, which was computationally accounted for by inhibitory competition in LISA (Morrison, Doumas and Richland, 2006).

It seems that goal-driven selection of relevant information for relational processing, as well as the inhibition of irrelevant information, constitute two sides of a "control coin" in analogical reasoning. No study to date has addressed both sides of the coin within a single task. For example, in the studies by Viskontas et al. (2004) and Cho et al. (2007) subjects were provided with all the relevant dimensions and were required to integrate them in WM while ignoring unequivocally irrelevant dimensions. In studies where similar relations were to be induced by the subjects themselves (e.g. Markman and Gentner, 1993; Waltz et al., 2000; Richland et al. 2006), no manipulation of the need for selectiveness occurred.

The goal of the presented study is to extend the empirical data on the role of EC in managing WM content in analogical reasoning with semantically meaningful material, which lacks the predetermination of a relevant relational structure. This will be done by attempting to manipulate experimentally the processing requirements for the above two mentioned aspects of EC. The load of WM capacity will also be varied. Unlike in any previous study known to the author, the needs for attentional selection, interference resolution and relational integration will all be varied in a single analogical mapping task. This procedure should allow one to explore, whether the two postulated EC faculties have dissociable or interacting effects on WM performance in analogical mapping.

The Study

The picture-mapping paradigm was used, as introduced by Markman and Gentner (1993) together with a cross-mapping procedure, advanced by Richland et al. (2006) *inter alia* by relational complexity manipulation, and applied in numerous other studies of analogical mapping (e.g., Tohill and Holyoak, 2000; Waltz et al., 2000). The task consists in analyzing the two scenes, presented to participants at once, and then deciding which object from the target scene best goes with the indicated object from the source scene. The subjects are instructed to search for a common "pattern" in the two pictures. The scenes usually depict simple causal relations, such as "towing" (see example in Fig. 1 from the current study).

The relational complexity (RC) was operationalized as the number of relational arguments (i.e., objects forming a relevant relational structure) to be processed in parallel for successful mapping. Thus, it was slightly different from the study of Richland et al. (2006), where RC was manipulated by necessarily repeating the same relation in a scene. There were either binary (involving two objects) or quaternary (four objects) relations to be mapped.

Unlike in any previous study, the total number of objects in the scene, and therefore the saliency of relevant relations, was factorially varied. It was thought of as an operationalization of the need for a goal-directed selection of structure mapping input. The relevant relations were "hidden" among five or ten objects in total. All other relations than those that were relevant ones, which could be possibly identified among the objects in a scene, were unique to only one scene. Assuming that more overt relational similarity in relatively semantically impoverished analogs constitutes a cue for engaging in relational mapping (i.e., it reminds task's goal), respective enriching the scenes (independently of relational complexity) seems to be a clear-cut way to make this cue less direct and thus more dependent on internal activation. Moreover, having to search for a relevant structure through the number of propositions, clearly exceeding WM capacity, seems to be more dependent on the quality of goal monitoring over the necessarily sequenced reasoning steps.

Since Cho et al. (2007) demonstrated that distracting information is detrimental only if attended to and actively maintained in WM, the manipulation of the need for interference was constrained to the cross-mapping procedure. That is, the presence of semantically (and to some, but never to the full, extent also featurally) similar object in different relational roles was varied always within relevant relations - like in the studies by Markman and Gentner (1993), but unlike in those by Richland et al. (2006). This objective was to ensure that the subjects' attention was not diverted from the relevant relational structure by a distracter external to it, but rather to increase the probability that distraction will affect the attempted structure mapping.

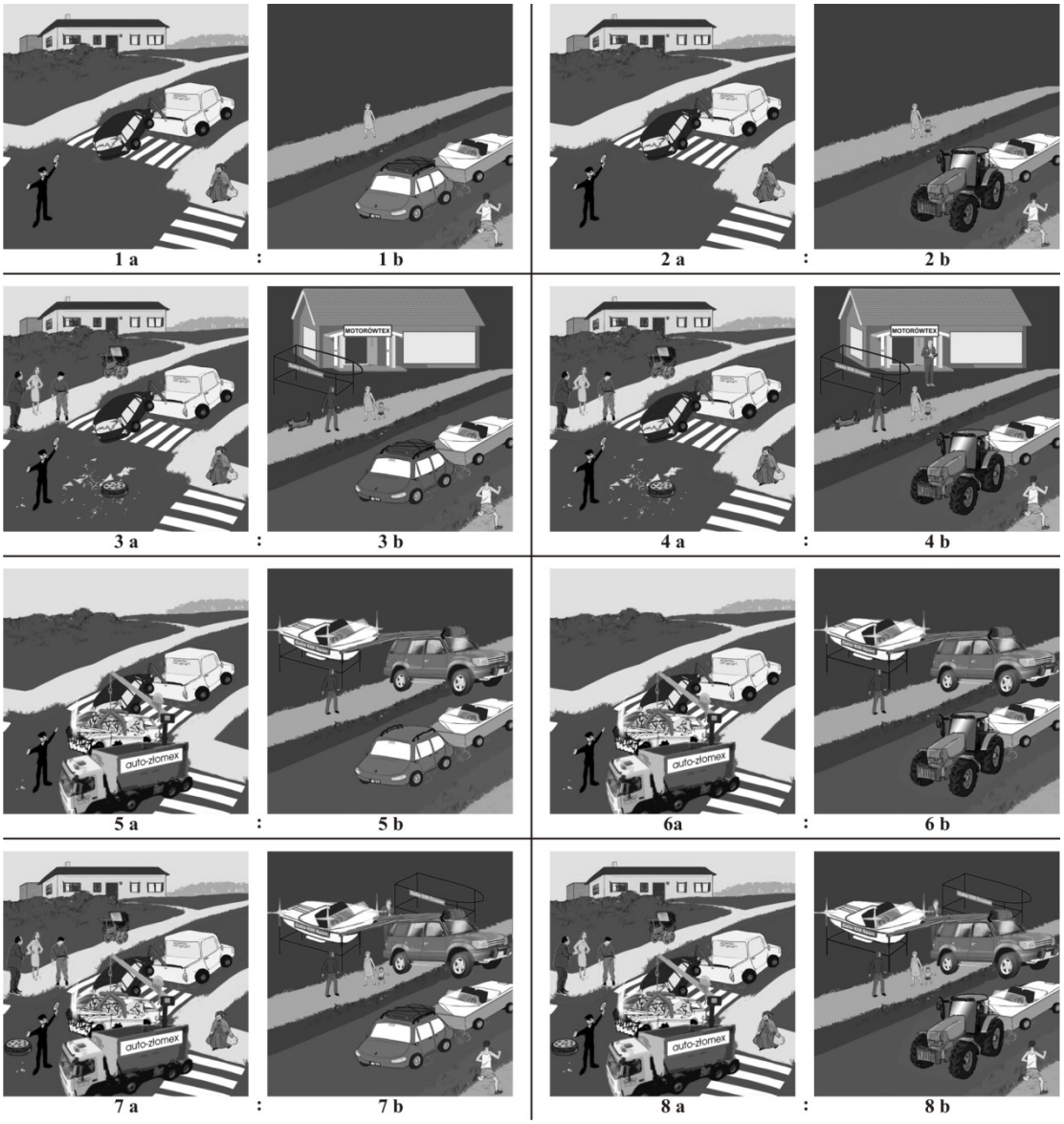


Figure 1. The example of one set of pictures of the analogical mapping task. The towed passenger car (with the *boat-wagon* object as a counterpart), and the towing lorry (pairs N° 1-4) or the loading lorry (N° 5-8) were highlighted for mapping. Odd numbers label pairs with a distracter (the towing passenger car in *b*s). Pairs in third and fourth rows contain quaternary relations (towing and loading are to be integrated). First and third row present high-saliency pairs.

It was expected that all three above manipulations will decrease the subjects' ability for relational mapping, but that their impact would be differential due to tapping into qualitatively distinct, although highly intertwined, cognitive capabilities. Viskontas et al. (2004) and Cho et al. (2007), from their results obtained in a similar, relatively simple mapping task, argued that the overadditive effects of RC and distraction suggest that relational integration and inhibition depend on the common pool of WM resources. However, some researchers suggest there is no reason for EC to operate more or less strongly in different WM load conditions (e.g., Embretson, 1995; Unsworth and Engle, 2005). Also, in the scene-mapping study on children by Richland et al. (2006) RC x distraction interaction occurred only in a group of 3-4 year olds. Thus, it seemed worth re-examining the RC - distraction relationship in a picture mapping task of more realistic complexity and administered to adults. As to the manipulation of relevant relations' (goal's) saliency, it was hypothesized that it will result in relational mapping decrements due to the worse discriminability of relevant relations.

However, no interaction between RC and saliency was expected. Although it appears that the whole scene has to be initially placed in WM to screen out irrelevant information, the impact of the difficulty of this selection process should not be different when more or less complex relations have to be integrated in WM for structure mapping. This is because the selection of input can be done incrementally, while RC taps into the exact WM capacity limits (Halford et al., 1998). Yet, RC and need for more rigid selection should additively affect the overall mapping performance.

Also, no interaction between saliency and distraction was expected due to the assumption that they reflect two different facets of EC, which are functionally distinct although highly related faculties (Braver et al., 2003). Thus, both EC manipulations in this study were expected to have additive influence on mapping performance.

The experiment reported here was a part of a bigger study to be reported elsewhere. Each participant solved the task reported here as their first in the whole session.

Method

Participants The participants were 122 inhabitants of Częstochowa, Poland (age = 16-44 years, M = 22.15, S.D. = 3.77, 62 females) recruited by flyers, newspaper and Internet ads. Each participant was paid 50 PLN (~10 EUR) and received a CD gift for their participation.

Materials and design. The scene mapping test contained a set of fifty-six picture pairs depicting every-day instances of common relations (e.g., destroying, giving) among conventional objects (e.g. a ceiling, money). All test items were similarly colorful and detailed, and were chosen from one hundred and three pilot items according to the items' reliability. No relation or object was repeated across the test. The quaternary (or equivalent) relations were created from the binary ones by extending the critical structure by two more objects necessary to be included in successful mapping. For instance, the relation of towing one object by another was extended by the third object, which remained behind for some reason, being loaded for transport by a fourth object (Fig. 1. 5-8). The example of distraction manipulation might be the changing role of a passenger car in a towing relation (Fig. 1., odd numbers). The spatial location of the corresponding objects was carefully varied within and across the pairs so as not to cue mapping. The pictures contained either five (Fig. 1., 1,2,5 & 6) or ten (Fig. 1., 3, 4, 7 & 8) objects in total. 2 × 2 × 2 repeated-measures design, with three factors: relational complexity (bi- vs. quaternary relations), relational saliency (five vs. ten objects within a scene), and distraction (absent vs. present), resulted in eight fully balanced experimental conditions. In order to control for the difficulty of specific scenes, like in the Richland et al. (2006) study, counterbalanced versions of each scene were created to match each experimental condition. The assignment of an item's versions to a test's version, as well as of test's versions to participants, was randomized. Each participant solved 56 different scene pairs, seven per condition, and the other twelve items in the training set representative for all conditions. The items' presentation order was fully randomized.

Procedure The task was administered on laptop computers (1280 × 800 pix. display resolution) in a group of four to five participants accompanied by the experimenter. The pairs of pictures were presented horizontally, each 5×5 inches large, with the source picture always on the left. The administration software allowed for visual separation of objects to be taken into account by the participants. This was done by covering the rest of the picture, apart from a particular object, with a semi-transparent filter, when a mouse cursor was over this object. Objects sometimes were elements of bigger objects. For example, a hand was separated from "the rest" of a person in a relation where soiling a hand, was critical; or it allowed to impose consideration of a pair of people as one entity, where the relation between pairs of people was a part of the to-be-mapped structure. Two objects were to be mapped for each pair of scenes.

Each participant received the same oral, detailed self-paced written and movie instructions. Each of the subsequent training items were followed by precise feedback. The instruction was to carefully explore pairs of pictures in order to first analyze what links exist between objects within each scene, and then to search for repeated pattern of these links across two scenes. The concept of the same relational role was carefully explained to participants and they learned that they will be required to indicate objects in the same roles in the other picture. They also learned that there might be two or four objects involved in a pattern, so that they should always search for the most complex pattern. Participants were instructed to first detect objects that need to be taken into account and then to verify if they recognized them correctly. A one-word name of an object appeared in the panel right under the picture when the cursor was over this object. The time for exploration was limited to 100 s, which had been validated as sufficient in pilot study. However, participants were encouraged to press a space bar as soon as they knew what the common pattern and object correspondences were. Once the time limit was reached or the space bar pressed, the first object in the source scene was highlighted and this picture became "frozen" for further exploration. The participants were to quickly click on this object with their mouse in the target scene, if they believed it played the same role as the highlighted one. As soon as they clicked in their chosen target object, a second object in the source scene was highlighted and its best counterpart in the target scene was also to be mouse-clicked. The choice of the first object excluded it from options for the second choice. There was a five second limit for a particular object's choice. Which relational role (i.e., agent or patient) was to be first placed in correspond-dence was randomized; in the distraction condition, however, the distracting object was always highlighted first. One object was highlighted across all versions/ conditions, but the other object varied between RC-conditions of a scene, in the way that in quaternary relations the second highlighted object was always from the "extended" part of the structure. The participants took one refreshment break (max. seven minutes) after completing 28 test items. Together with instruction and training, the task took up to two hours, depending on participant speed and the duration of the break.

The dependent variable was correct choice for both objects counted on an all-or-none basis.

Results

All analyses were done with Statistica 8.0 software. Nondirectional null hypothesis significance tests (with α value adopted at .05) and their p values are reported.

Mean correct responses for all conditions are depicted in Table 1. Paired *t*-tests showed that performance in all conditions was above chance level, conservatively defined as .2. In the *high RC/low saliency/distraction* condition, the value of this statistic was: $t\,(121) = 4{,}83$, $p < .001$.

Table 1. Mean correct responses in all exp. conditions

Saliency	Distraction	Relational Complexity	
		Binary	Quaternary
High (5 objects)	No	.62	.50
	Yes	.40	.36
Low (10 objects)	No	.58	.40
	Yes	.32	.28

Each factor yielded a main effect, thus validating the experimental manipulation. A 2 (RC) × 2 (saliency) × 2 (distraction) MANOVA revealed that accuracy of relational mapping decreased: as RC increased (F [1, 121] = 88.14, $p < .001$, $\eta^2 = .42$), as saliency decreased (F [1, 121] = 50.89, $p < .001$, $\eta^2 = .30$), and when distraction occurred (F [1, 121] = 292.98; $p < .001$, $\eta^2 = .71$). The only reliable interaction was two-way RC × distraction interaction, F (1, 121) = 28.872, $p < .001$, $\eta^2 = .19$. Post hoc tests (Tukey's HSD) indicated that differences between all means of this interaction were reliable, $p < .01$. As illustrated in Figure 2., the mapping accuracy dropped when distraction occurred, but the detrimental effect of cross-mapped foil was smaller in the high relational complexity than in the low relational complexity condition.

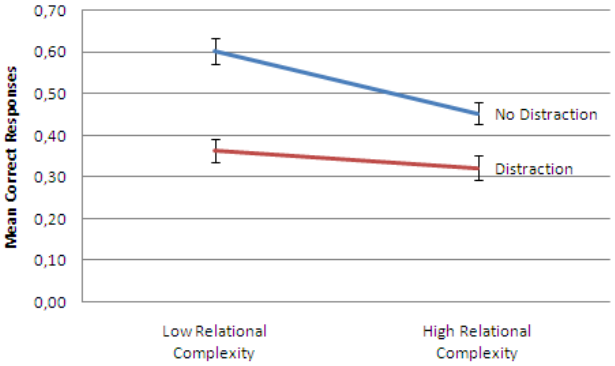


Figure 2. Interaction between Relational Complexity and Distraction. Error bars indicate standard error of the mean.

Discussion

Using a modification of Richland et al. (2006) scene mapping task the role of EC and WM as constraints on structure mapping was examined in adults. Unlike in any previous study the requirements for goal maintenance and application, for dealing with distraction and for relational integration were factorially varied in a single task.

First of all, the results replicate the previous findings in children that with increasing number of variables to be integrated and with similar objects appearing in different relational roles the level of mapping performance drops (Richland et al., 2006). Importantly, however, the presented study extends this evidence by showing that decreasing the saliency of relevant relational structure in a task also reliably impedes structure mapping. Together, these outcomes clearly exemplify the role of two EC faculties in relational mapping. Namely, EC might not only be reflected in interference resolution (Cho et al., 2007) or inhibition (Viskontas et al., 2004) within WM. It seems that EC is also involved during prior goal-directed search through structurally and semantically complex information to select WM input, as relevant for mapping.

The reason that the need for goal management and application was pronounced in this study was probably due to the lack of the predetermination of relevant relational structure and to minimizing the probability of non-relational cues to correct objects' correspondences. This explanation seems to be in line with LISA incremental mapping algorithm (Hummel & Holyoak, 2003), which makes the mapping in the model very dependent on the importance assigned to propositions (Kubose et al., 2002).

The internal activation of a task's goal and proper application of this goal to the necessarily incremental reasoning steps seems to be critically in play during selection and encoding of relational information, thus also before structure mapping is initiated within WM. The partial support for this conjecture comes from Gordon & Moser's (2007) study of eye movements' paths in a picture scene mapping task, in which people first scanned each scene in a given pair for meaningful relations and engaged in structure mapping only thereafter. The strong effect of distraction obtained in this study, together with a lack of reliable interaction between distraction and saliency of relevant relations, suggest that the two manipulated control requirements imposed qualitatively separate constraints on WM during relational mapping. These constraints pertain to the ability to select information for the purpose of identifying relevant relations to reason with, and to the ability to deal with interference when processing these relations.

Further, the lack of reliable interaction between relational complexity and saliency of relevant relations gives a hint that the ability to select relations for analogy might be qualitatively distinct from the ability to integrate these relations within WM. Although both processes are about abstraction, which definitely requires WM resources, only the second seems critically dependent on the capacity of this system. Further research is needed to resolve this issue.

Finally, the reliable underadditive interaction between relational complexity and distraction was a surprise. This finding counters previous results of the opposite direction of this interaction in (Cho et al., 2007; Viskontas et al., 2004 and Richland et al., 2006). It is neither in line with LISA, in which WM capacity and efficiency of dealing with distraction both depend on the same inhibitory competition algorithm (Hummel & Holyoak, 2003), nor does it support the hypotheses that EC operates equally strong in different WM load conditions (e.g. Embretson, 1995). The possible explanation for this result could relate to the limitation of WM capacity (Cowan, 2005). Accordingly, since critical processing takes place only in the highly limited, most active part of WM, it could be speculated, that the strength of the detrimental effect of distraction on mapping is linked to the probability of distracters entering WM. Thus, the interference, which was caused by the "reversed" object-role bindings in the distraction-conditions of this experiment, should have been stronger, if the distracters were a part of the structure successfully accommodated into WM. This was surely more the case for easier (i.e. binary), than for more complex (quaternary) relations.

Summary and Future Directions

The current study sheds some new light on the nature of EC constraints in relational reasoning. The results demonstrate that next to dealing with distraction within WM, the goal-directed selection of information to enter structure mapping is an important, and to some extent, maybe a dissociable constraint. They also hint at a possibility that cross-mapping is only detrimental when affected structure is successfully accommodated within WM. Further research on the intricate contributions of EC to relational reasoning could combine measuring of individual differences in EC functions with experimental manipulation of their load in relational reasoning task. This could provide precise tests of plausibility of computational models of analogy-making.

Acknowledgments

This work was sponsored by the Polish Ministry of Science and Higher Education (grant 1422/B/H03/2009/36, years 2009-2010, awarded to Edward Nęcka and Anna Chuderska). The author wishes to thank Katarzyna Musiał for her assiduous help in data collection, Jarosław Kwiatkowski for implementing the test and Maciej Bernaś for reliable IT support in the lab.

References

Awh, E., Vogel, E.K., Oh, S.-H. (2006). Interactions between attention and working memory. *Neuroscience*, *139*, 201-208.

Braver, T. S., Gray, J. R., & Burgess, G. C. (2007). Explaining the many varieties of working memory variation: Dual mechanisms of cognitive control. In A. R. A. Conway. C. Jarrold, M. J. Kane, A. Miyake, & J. N. Towse, *Variation in working memory* (pp. 76-108). Oxford: Oxford University Press.

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices test. *Psychological Review, 97*, 404-431.

Cho, S., Holyoak, K.J., & Cannon, T.D. (2007). Analogical reasoning in working memory: resources shared among relational integration, interference resolution, and maintenance. *Memory and Cognition, 35*, 1445-1455.

Chuderska, A., Chuderski, A. (2009). Executive control in analogical reasoning: Beyond interference resolution. In N. Taatgen, & H. van Rijn (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 1758-1763). Mahwah, NJ: Erlbaum.

Chuderski, A., & Nęcka, E. (2010). Intelligence and cognitive control. In A. Gruszka, G. Matthews, & B. Szymura (Eds.), *Handbook on individual differences in cognition*. New York: Springer Verlag.

Cowan, N. (2005). *Working memory capacity*. New York: Psychology Press Taylor & Francis Group.

Dempster, F. N., & Corkill, A. J. (1999). Individual differences in susceptibility to interference and general cognitive ability. *Acta Psychologica, 101*, 395-416.

Doumas, L.A.A., Hummel, J.E. (2005). Approaches to modeling human mental representations: what works, what doesn't and why. In K.J. Holyoak, R.G. Morrison (Eds.), *The Cambridge Handbook of Thinking and Reasoning*. New York: Cambridge University Press.

Doumas, L.A.A., Morrison, R.G., & Richland, L.E. (2009). The development of analogy: Task learning and individual differences. *Proceedings of the Thirty-First Annual Conference of the Cognitive Science Society* (pp. 3133-3138). Mahwah, NJ: Erlbaum

Embretson, S. E. (1995). The role of working memory capacity and general control processes in intelligence. *Intelligence*, 20, 169–189.

Engle, R. W., & Kane, M. J. (2004). Executive attention, working memory capacity, and a two-factor theory of cognitive control. In B. Ross (Ed.), *The psychology of learning and motivation, 44, (pp. 145-199)*. New York, NJ: Elsevier.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science, 7*, 155-170.

Gordon, P.C., Moser, S. (2007). Insight into analogies: Evidence from eye movements, *Visual Cognition*, 15, 20-35.

Gray, J. R., Chabris, C. F., & Braver, T. S. (2003). Neural mechanisms of general fluid intelligence. *Nature Neuroscience, 6,* 316-322.

Halford, G. S., Baker, R., McCredden, J. E., & Bain, J. D. (2005). How many variables can humans process? *Psychological Science, 16*, 70-76.

Halford, G. S., Wilson, W. H., & Phillips, S. (1998). Processing capacity defined by relational complexity: Implications for comparative, developmental, and cognitive psychology. *Behavioral & Brain Sciences*, 21, 803-864.

Holyoak, K. J. (2005). Analogy. In K.J. Holyoak, R.G. Morrison (Eds.), *The Cambridge Handbook of Thinking and Reasoning*. New York: Cambridge Univ. Press.

Hummel, J. E., & Holyoak, K. J. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review, 110*, 220-264.

Keane, M. T. (1997). What makes analogy difficult? The effects of order and causal structure on analogical mapping. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*, 946-967.

Kubose, T. T., Holyoak, K. J., & Hummel, J. E. (2002). The role of textual coherence in incremental analogical mapping. *Journal of Memory and Language*, 47, 407–435.

Markman, A., & Gentner, D. (1993). Structural alignment during similarity comparisons. *Cognitive Psychology*, 25, 431-467.

Morrison, R.G. (2005). Thinking in Working Memory. In K.J. Holyoak, R.G. Morrison (Eds.), *The Cambridge Handbook of Thinking and Reasoning*. New York: Cambridge University Press.

Morrison, R. G., Doumas, L. A. A., & Richland, L. E. (2006). The development of analogical reasoning in children: A computational account. In *Proceedings of the twenty-ninth annual conference of the cognitive science society*. Mahwah, NJ: Erlbaum.

Richland, L.E., Morrison, R.G., & Holyoak, K.J. (2006). Children's development of analogical reasoning: Insights from scene analogy problems. *Journal of Exp Child Psychology*, 94, 249–273.

Tohill, J. M., & Holyoak, K. J. (2000). The impact of anxiety on analogical reasoning. *Thinking & Reasoning*, 6, 27-40.

Unsworth, N., & Engle, R. W. (2005). Working memory capacity and fluid abilities: Examining the correlation between Operation Span and Raven. *Intelligence*, 33, 67-81.

Viskontas, I. V., Morrison, R. G., Holyoak, K. J., Hummel, J. E., & Knowlton, B. J. (2004). Relational integration, inhibition and analogical reasoning in older adults. *Psychology and Aging*, *19*, 581-591.

Waltz, J. A., Knowlton, B. J., Holyoak, K. J., Boone, K. B., Mishkin, F. S., de Menzes Santos, M. et al. (1999). A system for relational reasoning in human prefrontal cortex. *Psychological Science*, 10, 119–125.

Waltz, J. A., Lau, A., Grewal, S. K., & Holyoak, K. J. (2000). The role of working memory in analogical mapping. *Memory & Cognition*, 28, 1205–1212.

Selective Attention by Structural Alignment: An Eyetracking Study

Aaron B. Hoffman (aaron.hoffman@mail.utexas.edu)
Department of Psychology, 1 University Station A8000
Austin, Texas 78712-0187

Bradley C. Love (brad_love@mail.utexas.edu)
Department of Psychology, 1 University Station A8000
Austin, Texas 78712-0187

Arthur B. Markman (markman@psy.utexas.edu)
Department of Psychology, 1 University Station A8000
Austin, Texas 78712-0187

Abstract

A potential determinant of people's selective attention is offered by the structural-alignment view of comparison. This view holds that objects are compared via structured representations that align sets of features that share relational roles. A central claim of this account is that the comparison process directs attention towards alignable features. This prediction has been supported by offline measures by Markman and Gentner (1997), who showed that alignable features serve as better cues for recall than nonalignable features. The present study provides the first online test of the structure-alignment theory's claim that alignability drives selective attention. Consistent with this, we show that in addition to serving as better cues for recall, alignable differences are attended more than nonalignable differences. Within-trial attention dynamics revealed that attention to alignable differences increases over the course of the comparison process.

Keywords: comparison, alignment, attention, recall, eye movements, eye tracking

Introduction

The amount of information that inundates people's perceptual systems creates a significant challenge. As people move through their environment, they are faced with thousands of decisions about which information they should selectively attend and which they should filter out. They must decide that certain things are worth remembering and that others are not. How are such decisions made?

There are a variety of factors that influence selection of parts of the stimulus stream. Early work examining how people attend to complex visual scenes showed that people will fixate the most informative elements (Buswell, 1935; Mackworth & Morandi, 1967; see Henderson & Hollingworth, 1999 for review). Subsequent work explored people's tendency to attend to the most perceptually salient features (e.g., Henderson, Weeks, & Hollingworth, 1999; Parkhurst, Law, & Niebur, 2002). Work on schemata and memory suggests that semantic consistency with a schema determines what is later recalled (Bransford & Johnson, 1972; 1973; Brewer & Dupree, 1983; Rummelhart, 1980). Finally, recent eye tracking work in categorization (Rehder, Colner, & Hoffman, 2009) and in natural scene perception (e.g., Hayhoe, Shrivastava, Mruczek, & Pelz, 2003) proposes that the information demands of the task are the biggest influences on what people selectively attend.

In the present study we test the idea that yet another determinant of people's selective attention is the comparisons they make. We will first review comparison processes and then evidence from Markman and Gentner (1997) showing that people have better recall when they are cued by elements from scenes that are part of structural alignment. Then, by replicating Markman and Gentner (1997) with an eyetracker, we provide an online test of the idea that structural alignment can drive selective attention.

Comparison

The ability to compare is an integral part of human cognition. Category membership is determined by the degree of similarity to category representations (Medin & Schaffer, 1978; Nosofsky, 1984). In problem solving, people find solutions by comparing new problems to previously solved problems (Chi, Feltovich, & Glaser, 1981; Ross, 1987). In episodic memory, probes are compared to memory traces (Hintzman, 1986). In analogy people compare base and target domains. (Falkenhainer, Forbus, & Gentner, 1989; Gentner, 1983; Hummel & Holyoak, 1997)

There have been a few approaches to modeling the comparison process, including computing distances in multidimensional space using feature vectors, (Shepard, 1962) or comparing features using set operations, Tversky (1977). And yet to account for human comparison of complex stimuli with relational structure, a third approach has been used. Borrowing from models of analogy (Falkenhainer et al., 1989), the structure-alignment account (Gentner, 1983) represents objects as features inside structures of relations. For example, structure-alignment theory posits that people will encode features (e.g., the people and objects in Figure 1A) as arguments to relational predicates: smokes(man, cigar) or paints(painter, model). On this account, significant processing is applied to building a representation of the relations between features in a scene

or object, and into determining which objects match on the basis of shared roles.

With structure alignment, a great deal more information is both represented and processed than what is proposed by simpler accounts. Rather than comparing sets of features alone, comparisons are made over features and their relations. To accomplish this, objects with the same relational role in both scenes are placed in correspondence, while objects with different roles in their respective scenes are not.

Alignable Differences and Attention

Structure alignment has the ability to represent and calculate similarity over structured representations. However, this ability comes at a processing cost; the alignment process must build structurally consistent matches that satisfy parallel connectivity and one-to-one mapping. Parallel connectivity requires that matching relations have matching arguments. For example, in Figure 1A and 1B, if the photographer is aligned with the painter, then the man with the backpack is aligned with the model. One–to-one mapping states that across representations each object can be aligned to at most one other object—the boy with the backpack cannot also be aligned with the man and the cigar. Thus, the mapping process in structural alignment involves more than simple feature comparisons.

As a result of the more extensive processing involved in structural alignment, three different kinds of output are produced (Markman & Gentner, 1993). Whereas the feature-based approaches distinguish only between commonalities (matching features) and differences (mismatching features), structural alignment produces commonalities on one hand, and two types of differences. Differences that are linked to the commonalities, or *alignable differences*, and those that are not, *nonalignable differences*. For example, the female figure in Figure 1A is an alignable difference with the boy in Figure 1B. However, the man in the chair is a nonalignable difference, since there is no corresponding object in 1B. Thus, instead of just two kinds information used in the similarity calculation, the structural alignment approach has three.

The three types of output allow structure alignment to make the unique prediction that comparisons will focus people's attention on alignable differences. There are two reasons for this. First, it has been shown that people tend to weigh commonalities more heavily than differences in similarity judgments (Tversky, 1977). Since alignable differences are a type of commonality (on the basis of the relational structure) they should receive more attention.

The second reason for additional focus on alignable differences is that the entire alignment process is geared towards building up relational structure. Since alignable differences are what compose that structure, they should receive a significant amount of attention.

Over the last decade there has been a growing amount of evidence that alignable differences in fact receive more weight than nonalignable differences. Markman and Gentner (1996) showed that when given a choice, subjects were more likely to select scenes with nonalignable differences as being more similar to a base scene than scenes with alignable differences. In a second experiment they showed that similarity ratings were more affected by variability in alignable differences than by variability in nonalignable differences. Markman and Gentner (1993) showed that people tend to list more alignable differences than nonalignable differences.

In another demonstration of the importance of alignable differences, Markman and Gentner (1997) had subjects rate the similarity of ten pairs of scenes, like those in Figure 1. Later, subjects were either given probes that were part of an alignable or nonalignable difference, as in Figure 2. They found that on average, subjects recalled 2.35 pieces of information when memory probes were part of an alignable difference versus just 1.3 when the probes were part of a nonalignable difference. Thus, across a range of studies, people seemed to place more weight on alignable differences.

The critical implication of these findings is the idea that structural alignment can be one of the determiners by which people select relevant aspects of their environment. The most direct test of this idea is an online measure of people's selective attention behavior as they make comparisons.

Eyetracking and Selective Attention

It has been well established that eye movements and selective attention are closely linked. For example, Shepard, Findlay, and Hockey (1986) demonstrated that although attending without making corresponding eye movements is possible, it is not possible to make an eye movement without shifting attention. Since high quality visual information is acquired only from a limited spatial region surrounding the fovea, we move our eyes three times each second through high-velocity saccades to position the fovea on what seems important.

It is no surprise then that eye tracking has enjoyed success in numerous research areas that appeal to the construct of selective attention. For example, Rehder and Hoffman (2005a) showed that learning a category corresponded to abrupt shifts in fixations towards relevant information. Later, Rehder and Hoffman (2005b) replicated Medin and Schaffer's (1978) 5-4 category structure with an eye tracker and found that fixation times to stimulus dimensions matched the decisions weight estimated from behavioral responses.

More recently, researchers have begun to leverage the flexibility that eye movement analysis offers in terms of experimental design. It is now possible to examine how attention is allocated across different kinds of tasks (Rehder, Colner, & Hoffman, 2009) and across different stimuli and categories (Blair, Watson, Walshe, & Maj, 2009). The close link between attention and eye movements has been shown across a variety of cognitive tasks (see Liversedge & Findlay, 2000 and Rayner, 1998 for reviews).

Of course, the key advantage to using eye tracking for the present purposes is that it provides an online measure of what people attend to during the comparison process. While recall behavior, verbal protocols, and similarity ratings all point to the conclusion that alignable differences have a greater impact than nonalignable differences on comparison, these are all offline measures. Testing recall performance, for example, occurs well after the comparison process has taken place. Although offline measures can indicate what subjects preferred to encode, they can't tell us about processing dynamics as they unfold over time.

Finally, one of the key claims of structural alignment is that the comparison process can help people determine what information is worth attending to. If in fact alignable differences do not receive more attention than nonalignable differences, then the validity of this claim is called into question. The present study will provide an online test of whether people allocate more attention to alignable features than to nonalignable features.

Experiment

The goal of the present experiment is to use eye tracking as a source of data to measure how comparison processes direct people's attention to important pieces of information, and how that in turn relates to recall of that information. According to the structural-alignment approach, the process of comparison should lead people to attend to alignable over nonalignable differences. As a result of this boost in attention, alignable differences should serve as better cues for recall later on. To test this, we replicated Markman and Gentner (1997), using an eyetracker to monitor subjects' attention allocation. Subjects were fit with a head-mounted eye tracker and we recorded their eye movements to alignable and nonalignable differences as they rated the similarity of ten pairs of scenes.

The main result of interest is whether subjects tend to allocate a greater amount of attention to alignable differences than to nonalignable differences. The structure-alignment approach predicts that subjects' fixation times will be greater on average for alignable differences than for nonalignable differences. Such a finding supports the idea that comparison via structural alignment helps focus people on what's important in the environment.

We will also examine how attention to alignable differences unfolds over the comparison process. Such dynamics will have implications for models of comparison.

Method

Participants Twenty-eight University of Texas students participated for course credit. They were tested individually and assigned to a random order of items. For each item, half of the subjects saw one comparison scene, and half saw the other. At the same time, the assignment of aligned and nonaligned recall cues to each comparison scene was counterbalanced across subjects. This designed allowed us to separate out effects of alignability on attention allocation and memory from any specific object-salience effects, or differences in subjects' ability to recall particular objects from the scenes.

Materials The stimuli in the current study were based on the Markman and Gentner (1997) materials, but were made more suitable for eyetracking by (1) removing unnecessary textures and (2) increasing the distances between objects to more clearly distinguish which were fixated.

Figure 1 shows an example stimulus. As in the original study, there were ten sets of picture triads (one base, and two comparison pictures). The base picture had two relational scenes within it and each comparison picture matched one of the relational scenes. For example, Figure 1A is a base picture. It contains a portrait relation (the artist is painting a portrait of the model on the right), and there is a burning-dropping relation on the left (the man is dropping ash from a lit cigar) on the left. Each comparison matched one of the relational scenes. For example, Figure 1B matches the portrait relation, and Figure 1C matches the base picture on the burning-dropping relation. On a given trial, the base scene and (one of the) comparison scenes are presented together on screen. Later, one object from each relational structure in the base scene was used as a recall cue. For example, as shown in Figure 2, the painter and the man in the chair from Figure lA were used as recall cues.

The eye tracker was an SMI Eyelink II, which was set to track one eye at 250 Hz.

Procedure Subjects were first fitted and calibrated to the eye tracker. Items (i.e., a pairing of a base and one comparison scene) appeared on the screen. At their own pace, subjects rated the similarity of the base picture to the comparison picture (on a 1-to-9 scale). Before each item

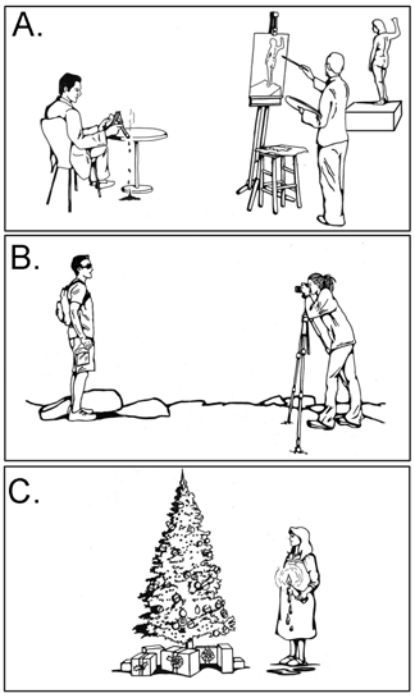


Figure 1. Example stimuli. Panel A is the base picture. Panels B and C are the two comparison scenes.

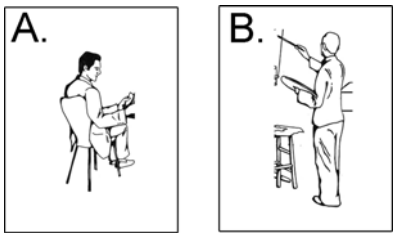


Figure 2. Two example recall cues. Depending on the comparison picture, either cue can be an alignable or nonalignable difference.

presentation subjects were asked to fixate a small circle in the center of the monitor. This was used both as a drift correction and as an indication that they were ready for the next trial. Subjects recorded their rating by typing one of the corresponding numbers keys on the keyboard.

After subjects provided the ten ratings they engaged in a reading task for 30 minutes.

During the recall phase subjects were presented with one of the recall cues. Half of the recall cues were from alignable differences and the other half were from nonalignable differences. Subjects' verbal responses were recorded by a computer microphone.

Results

Recall We first set out to test whether we replicated the basic finding from the original Markman and Gentner (1997) study that alignable cues yield better recall than nonalignable cues with the revised stimuli. Therefore, we examined the effect that alignability had on subjects' recall of the scenes, by counting the number of pieces of information recalled from the base scene as a function of whether they received an alignable or nonalignable cue during recall. The data were first transcribed from the voice recordings and then rated by a single rater. The instructions to the rater were that each proposition (adjective, noun, or verb) about the scene counted as a piece of information.

The average number of correctly recalled pieces of information for the alignable cues (M = 1.8, SD = 1.2) was reliably greater than the number of pieces of information recalled for the nonalignable cues (M = 1.3, SD = 0.92), $t(27) = 2.44$, $p < .05$. The analysis was also carried out by item, and the result was marginally reliable $t(19) = 1.84$, $p = .081$. Thus, the basic findings found by Markman and Gentner were replicated here.

Fixations For our initial analysis, we constructed heat maps of eye fixations to get a sense for where people were looking while judging picture similarity. Figure 3 shows heat maps of fixations superimposed over one of the items, with both comparison scenes. To construct these heat maps, each x-y coordinate of the fixations were weighted by their total fixation time and summed over all subjects for each item. The weighted fixation coordinates were then processed by a Gaussian kernel density estimator, with bandwidth estimation (Jones, Oliphant, & Peterson, 2001). The red spots of the heat map reflect greater average amounts of fixation time, and as a result, where subjects were attending. Overall, and as expected, in both panels of Figure 3 fixations were centered directly over the objects in

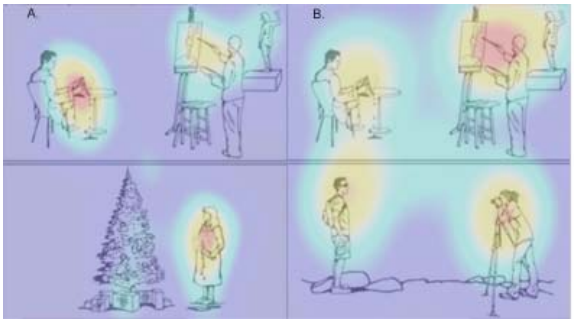


Figure 3. Example heatmap of fixations to an item, averaged over subjects for one of the ten items, with comparison scenes. In Panel A there are more fixations to the man smoking, but in Panel B, there are more fixations to the painter and model.

the scenes. However, the heat maps also show that the allocation of attention is very different depending on which comparison scene the subjects saw.

According to structure-alignment theory, more fixations should land near the objects that align with the comparison picture. For example, the comparison scene in Figure 3A aligns with the man smoking in the left half of the base picture whereas the comparison scene of Figure 3B aligns with the portrait relation on the right hand side of the base image. In fact, the heat maps in Figure 3 show the result predicted by structural alignment. There are more intense and concentrated hot spots over the man in the chair in Figure 3A, and lesser hot spots over the painter and the model. The reverse is true for Figure 3B, there are more intense hot spots over the painter and the model, and weaker hot spots over the man in the chair. The heat map presented in Figure 3 provides a clear illustration of how subjects allocate greater attention allocation to alignable differences in the scene.

Next, we extended the above analysis to all items. For this purpose we coded fixations according whether they were to an alignable difference or to a nonalignable difference in the base picture. We then computed the total fixation time for alignable differences across all items, for each subject. The average total fixation time to alignable differences (M = 1473, SD = 814) was greater than that for nonalignable differences (M = 1272, SD = 690), $t(27) = 2.25$, $p < .05$. (Although item analysis was not statistically reliable $t(19) = 1.2$, $p = .28$., seven out of ten of the items showed the effect in the expected direction). Thus, as structure-alignment predicts, subjects allocated more fixation time to alignable differences as compared to nonalignable differences.

The above results showed that overall, the comparison process engaged by subjects in determining the similarity of two images caused them to fixate alignable differences over nonalignable differences. But how does the comparison process direct attention to important features in a scene, and at what point are people drawn to alignable differences? Figure 4 shows the probability of fixating alignable differences, nonalignable differences, and to the comparison scene as a function of time, for ten seconds of the trial.

To construct Figure 4 we determined, for each 50-ms interval, whether a subject was fixating one of those three

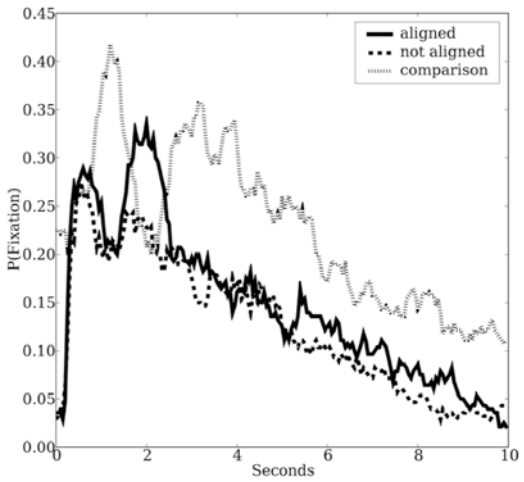


Figure 4. Probability of fixating the aligned, nonaligned, and comparison objects in the scene, as a function of time (seconds).

locations. We then averaged over all trials and subjects to examine attention allocation over the course of the trial.

The figure shows that as expected, subjects showed no immediate preference for the alignable or nonalignable differences in the base scene. (The initial preference for the comparison scene in the first 50 ms reflects that the comparison scene was a much larger area of interest than the individual alignable and nonalignable differences, and there's a greater baseline chance that eye fixations will happen to be there first.)

Figure 4 then shows that during the next second, there was a dramatic increase in fixations to all three locations, but especially to the comparison scene. In fact, after one second, there is a sudden decrease in fixations to the alignable and nonalignable differences. Fixations then shift from the comparison scene to the alignable differences in the base picture, until fixations to alignable differences peak, at around the two-second mark. After this, fixations gradually dropped off for all locations (as more and more subjects have already responded), with the most fixations allocated to the comparison scene. On average, subjects did not allocate more fixations to the nonalignable differences at any point in the trial.

Discussion

Markman and Gentner's (1997) result that people have greater recall performance when cues are part of alignable differences replicated in the present study. These results were consistent with other previous work showing that alignable differences have a greater impact than nonalignable differences on people's comparison behavior.

The main contribution here was that we were able to observe the structural alignment process online. The predictions for the eyetracking results, that more fixations should be allocated to the alignable features obtained. The unfolding of attention allocation over the course of the comparison process also appeared to make sense. As soon as subjects allocated a significant amount of attention to the comparison and base scenes, attention was allocated to the alignable differences, as predicted.

Our results have clear implications for cognitive models. First, mechanisms of comparison need to represent relational structure to explain selective attention behavior towards stimuli with any high level of complexity. Standard models in category learning that contain geometric (Kruschke, 1992) or feature-based (Lee & Navarro, 2002) similarity metrics need to be modified to account for people's ability to represent and attend to relational semantics.

Models that already have the ability to represent relations are consistent with the eye tracking results from the present study. For example, Hummel and Holyoak's, (1997; 2003) LISA and Larkey and Love's (2003) CAB models look for surface-feature similarities between items and only later try to match lower- and higher-order relations. Such mapping patterns reflect the selective attention behavior of our subjects because subjects required two seconds on average to focus primarily on alignable differences.

That subjects in our experiment attended differentially to objects according to their placement in the relational structure provides a proof of concept for using eye movements for more detailed tests of computational models, including those that already have the ability to represent relational structure. Additional eyetracking data can be collected to constrain the various components, for example, by having people make comparisons over objects that with different levels of relations (e.g., higher order versus lower order), or by manipulating subjects working memory, models' changes in selective attention can be related changes in selective attention to humans directly.

One of the most interesting implications for our results is derived from considering the working memory constraints of models like CAB and LISA. Working memory functions in such models to constrain the types of relations considered. With less working memory only lower-order relations or superficial feature matches will be represented by the model. This predicts that the details of the relational structure that people can maintain will also be influenced by working memory constraints. As a result, another potential determiner of what people selectively attend to in a scene is their working memory. If their working memory is compromised, they will not be able to use relational structure to guide their selective attention. Thus, the present data provide clear predictions for future eyetracking studies.

The rich source of data provided by eyetracking was able to confirm predictions of structural alignment and shows promise for constraining and developing more detailed processing accounts of existing computational models of comparison. In addition, there are potential future directions for empirical studies that follow from the present work to explain how it is that people decide what to selectively attend in an information-rich world.

References

Blair, M. R., Watson, M. R., Walshe, R.C., & Maj, F. (2009). Extremely selective attention: Eye-tracking studies on dynamic attentional allocation to stimulus features. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*, 1196-1206.

Bransford, J.D., & Johnson, M.K. (1972). Contextual prerequisites for understanding some investigations of comprehension and recall. *Journal of Verbal Learning and Verbal Behavior, 11*, 717-726.

Bransford, J.D., & Johnson, M.K. (1973). Considerations of some problems of comprehensioin In W.G. Chase (Eds.), *Visual information processing* (pp. 383-438).

Brewer, W. F. , & Dupree, D. A. (1983). Use of plan schemata in the recall and recognition of goal-directed actions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *9*, 117–129.

Buswell, G. T. (1935). How people look at pictures: A study of the psychology of perception in art.Chicago: University of Chicago Press.

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5*, 121-152.

Falkenhainer, B., Forbus, K., & Gentner, D. (1989). The Structure-Mapping Engine: Algorithm and examples. *Artificial Intelligence, 41*, 1-63.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science, **7,*** 155-170

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*, 327–352.

Hayhoe, M., Shrivastava, A., Mruczek, R., & Pelz, J.B. (2003). Visual memory and motor planning in a natural task. *Journal of Vision, 3*, 46-63.

Henderson, J, M., & Hollingworth, A. (1999). High-level scene perception. *Annual Review of Psychology, 50, 243-271.*

Henderson, J, M., Weeks, P.A., & Hollingworth, A. (1999). The effects of semantic consistency on eye movements during scene viewing. *JEP: Human Perception and Performance, 25*, 210-228.

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review*, *93*, 411-428.

Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review, 104*, 427–466.

Jones, E., Oliphant, T., Peterson, P. and others (2001). *SciPy*. http://www.scipy.org/SciPy.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*, 22-44.

Larkey, L.B., & Love, B.C. (2003). CAB: Connectionist Analogy Builder. *Cognitive Science, 27*, 781–794.

Liversedge, S. P., & Findlay, J. M. (2000). Saccadic eye movements and cognition. *Trends in Cognitive Science, 4*, 6-14.

Mackworth, N. H., & Morandi, A. J. (1967). The gaze selects informative details within pictures. *Perception & Psychophysics,2,* 547-552.

Markman, A. B. (1997). Constraints on analogical inference. *Cognitive Science, 21*, 373–418.

Markman, A. B. & Gentner, D. (1993). Splitting the differences: A structural alignment view of similarity. *Journalof Memory and Language, 32*, 517–535.

Markman, A. B., & Gentner, D. (1996). Commonalities and differences in similarity comparisons. *Memory and Cognition, 24*, 235–249.

Markman, A. B., & Gentner, D. (1997). The effects of alignabilty on memory. *Psychological Science, 8*, 363-367.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 85*, 207-238.

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 10, 104-114.

Parkhurst, D., Law, K., & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Research, 42*, 107–123.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124*, 372-422.

Rehder, B., & Hoffman, A. B. (2005a). Eyetracking and selective attention in category learning. *Cognitive Psychology*, 51, 1-41.

Rehder, B., & Hoffman, A. B. (2005b). Thirty-something categorization results explained: Selective attention, eyetracking, and models of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31, 811-829.

Rehder, B., Colner, R.M., & Hoffman, A.B. (2009). Feature inference learning and eyetracking. *Journal of Memory & Language, 60*, 394-419

Rumelhart, D.E. (1980). Schemata: The building blocks of cognition. In R.J. Spiro, B.C. Bruce, & W.F. Brewer (Eds.), *Theoretical issues in reading comprehension* (pp. 33-58). Hillsdale, NJ: Erlbaum.

Shepherd, M., Findlay, J. M., & Hockey, R. J. (1986). The relationship between eye movements and spatial attention. *The Quarterly Journal of Experimental Psychology, 38*, 475-491.

Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function, I. *Psychometrika, 27*, 125-140.

A Structure-Mapping Model of Raven's Progressive Matrices

Andrew Lovett (andrew-lovett@northwestern.edu)
Kenneth Forbus (forbus@northwestern.edu)
Jeffrey Usher (usher@cs.northwestern.edu)
Qualitative Reasoning Group, Northwestern University, 2133 Sheridan Road
Evanston, IL 60208 USA

Abstract

We present a computational model for solving Raven's Progressive Matrices. This model combines qualitative spatial representations with analogical comparison via structure-mapping. All representations are automatically computed by the model. We show that it achieves a level of performance on the Standard Progressive Matrices that is above that of most adults, and that the problems it fails on are also the hardest for people.

Keywords: Analogy, Spatial Cognition, Problem-Solving

Introduction

There is increasing evidence that visual comparison may rely on the same structural alignment processes used to perform conceptual analogies (Markman & Gentner, 1996; Lovett et al., 2009a; Lovett et al., 2009b). An excellent task for exploring this is the Raven's Progressive Matrices (RPM) (Raven, Raven, & Court, 2000) In RPM problems (Figure 1), a test-taker is presented with a matrix of images in which the bottom right image is missing, and asked to pick the answer that best completes the matrix. Though RPM is a visual task, performance on it correlates highly with other assessment tasks, many of them non-visual (e.g., Snow & Lohman, 1989; see Raven, Raven, & Court, 2000, for a review). Thus, RPM appears to tap into important, basic cognitive abilities beyond spatial reasoning, such as the ability to perform analogies.

This paper presents a computational model that uses analogy to perform the RPM task, building on existing cognitive models of visual representation and analogical comparison. Our claims are:

1) Tasks such as RPM rely heavily on qualitative, structural representations of space (e.g., Biederman, 1987; Forbus, Nielsen, & Faltings, 1991). These representations describe relations between objects in a visual scene, such as their relative location. Importantly, these representations are hierarchical (Palmer, 1977); they can also describe larger-scale relations between groups of objects or smaller-scale relations between parts of an object.

2) Spatial representations are compared via structure-mapping (Gentner, 1983), a process of structural alignment first proposed to explain how people perform analogies. Structure-mapping is used here to compute the similarity of two images, to identify corresponding objects in the images, and to generate abstractions based on commonalities and differences.

We previously (Lovett, Forbus, & Usher, 2007) described a model based on these principles that achieved human adult-level performance on two sections of the Standard Progressive Matrices test. That model was unable to handle the more difficult sections of the test because it only considered differences between pairs of images. This paper describes a more advanced model which performs at an above-average level on the hardest four sections of the test. It remains grounded in the same principles but is able to identify patterns of differences across rows of images. Like before, all inputs are automatically computed from vectorized input.

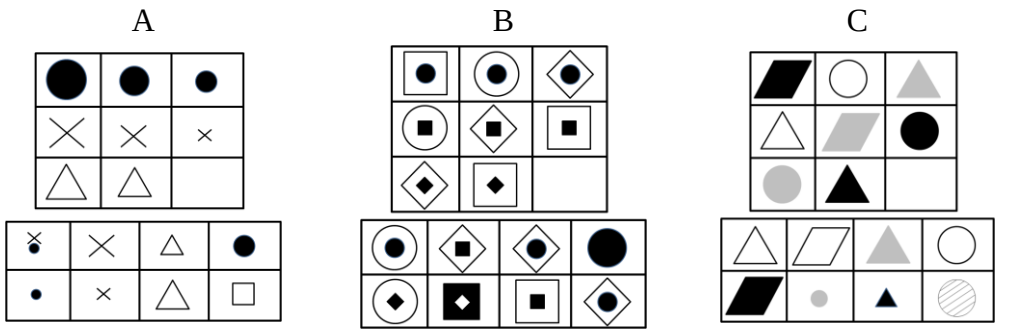
We first discuss Carpenter, Just, and Shell's (1991) computational model of the RPM. We then describe our model and its results on the Standard Progressive Matrices test. We end with conclusions and future work.

Background

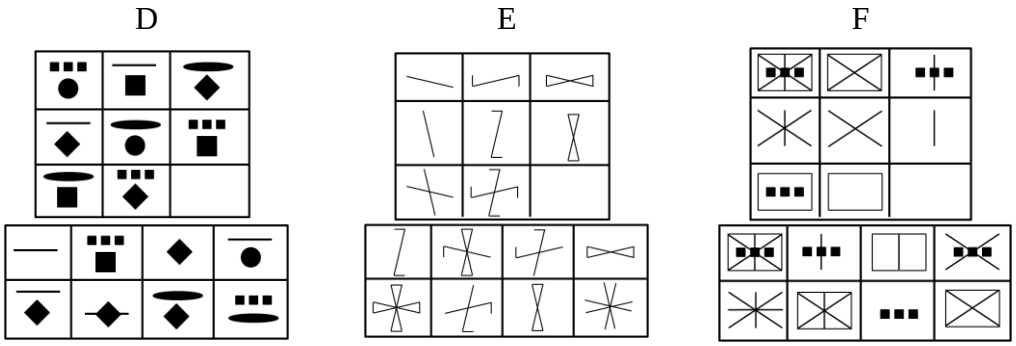
The best-established model of Raven's Progressive Matrices was developed by Carpenter, Just, and Shell (1991). It was based on both analysis of the test and psychological studies of human performance. The analysis led to the observation that all but two of the problems in the Advanced Progressive Matrices, the hardest version of the test, could be solved via the application of a set of six rules (see Figure 1 for examples). Each rule describes how a set of corresponding objects vary across the three images in a row. The simplest, Constant in a Row, says that the objects stay the same. Quantitative Pairwise Progression (Figure 1A) says that one of the object's attributes or relations gradually changes. The other rules are more complex, requiring the individual to align objects with different shapes (Distribution of Three), or to find objects that only exist in two of the three images (Figure Addition or Subtraction, Distribution of Two).

The psychological studies suggested that most people solved the problems by studying the top row, incrementally generating hypotheses about how the objects varied across that row, and then looking at the middle row to test those hypotheses. This process began by comparing consecutive pairs of images in a row.

Armed with their observations, Carpenter et al. built two computational models to solve the Advanced Progressive Matrices: FAIRAVEN and BETTERAVEN. Both models used hand-coded input representations. They solved a problem by: 1) identifying which of the six rules applied to the first two rows, and 2) computing a mapping between those two rows and the bottom row to determine how to apply the same rules in that row.



Carpenter Rules	Quantitative Pairwise Progression	Constant in a Row + Distribution of Three	Distribution of Three (applies twice)
Our Classification	Differences	Literal	Advanced Literal
Answer	3	5	2



Carpenter Rules	Distribution of Three (applies twice)	Figure Addition or Subtraction	Distribution of Two (applies two or three times)
Our Classification	Advanced Literal	Advanced Differences	Advanced Differences
Answer	4	5	7

Figure 1: Several examples of RPM problems. To protect the security of the test, all examples were designed by the authors. Included are the rules required to solve the problems according to Carpenter et al.'s (1991) classifications.

BETTERAVEN differed from FAIRAVEN in that it possessed better goal-management and more advanced strategies for identifying corresponding objects in a row. Whereas FAIRAVEN could perform at the level of the average participant in their subject pool, BETTERAVEN matched the performance of the top participants.

Since BETTERAVEN's development, studies (Vodegel-Matzen, van der Molen, & Dudink, 1994; Embretson, 1998) have suggested that Carpenter et al.'s rule classification is a strong predictor of the difficulty of a matrix problem: problems that involve the more advanced rules, and that involve multiple rules, are more difficult to solve. In this respect, the models have had an important, lasting legacy. Unfortunately, they have two limitations. First, they operate on hand-coded input, hence the problem of generating the spatial representations is not modeled. Carpenter at al. justify this by pointing to the high correlation between RPM and non-spatial tasks, suggesting that perceptual encoding must not play an important role in the task. However, an alternate explanation is that the problem of determining the correct spatial representation for solving a matrix relies on encoding and abstraction abilities shared with other, non-visual modalities. The second drawback is that the six rules identified by Carpenter et al. were hard-coded into their models. Thus, the models tell us little about how people discover those rules in the first place. That is, how do people progress from comparing pairs of images to understanding how objects vary across a row?

Our model addresses these limitations by using existing models of perceptual encoding and comparison. Spatial representations are automatically generated using the CogSketch (Forbus et al., 2008) sketch understanding system. These representations are compared via the Structure-Mapping Engine (SME) (Falkenhainer, Forbus, & Gentner, 1989) to generate representations of the pattern of variance across a row. We describe each of these systems, beginning with SME as it plays a ubiquitous role in our models of perception and problem-solving.

Comparison: Structure-Mapping Engine

The Structure-Mapping Engine (SME) (Falkenhainer, Forbus, & Gentner, 1989) is a computational model of comparison based on Gentner's (1983) structure-mapping theory. It operates over structured representations, i.e., symbolic representations consisting of entities, attributes, and relations. Each representation consists of a set of

expressions describing attributes of entities and relations between entities. For example, a representation of the upper-left image in Figure 1B might include an expression stating that the square **contains** the circle.

Given two such representations, a *base* and a *target*, SME aligns their common relational structure to generate a mapping between them. Each mapping consists of: 1) correspondences between elements in the base and target representations; 2) *candidate inferences* based on expressions in one representation that failed to align with anything in the other; 3) a similarity score between the two representations based on the quantity and depth of their aligned structure. For this model, we normalize similarity scores based on the overall size of the base and target.

SME is useful in spatial problem-solving because a mapping between two spatial representations can provide three types of information. First, the similarity score gives the overall similarity of the images. Second, the candidate inferences identity particular differences between the images. Third, the correspondences can be useful in two ways. (a) Correspondences between expressions identify commonalities in the representations, and (b) correspondences between entities identify corresponding objects in the two images, a key piece of information for determining how an object varies across a row of images.

Finally, SME can take as input constraints on its mappings, such as requiring particular correspondences, excluding particular correspondences, or requiring that certain types of entities only map to similar types. While the psychological support for these constraints is not as strong as the overall psychological support for SME, we have found previously (Lovett et al., 2009b) that constraints can be useful for simulating a preference for aligning similar shapes when comparing images.

Perceptual Encoding: CogSketch

We use CogSketch (Forbus et al., 2008) to generate spatial representations. CogSketch is an open-domain sketch understanding system. Given a sketch consisting of line drawings of a set of objects, CogSketch automatically computes qualitative spatial relations between the objects, generating a spatial representation. This representation can then serve as the input to other reasoning systems.

There are two ways of providing input to CogSketch. A user can either draw out a sketch within CogSketch, or import a set of shapes created in PowerPoint. In either case, it is the user's responsibility to segment an image into objects—CogSketch does not do this automatically.

Essentially, the user is performing part of the job of perceptual organization (Palmer & Rock, 1994), the low-level visual operation that creates a set of entry-level units for processing. We focus on modeling the ways one must reorganize these units—via grouping and segmentation—during the problem-solving processes.

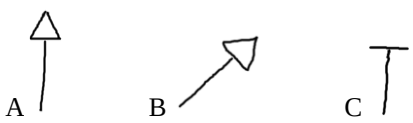


Figure 2. A,B: Two arrow shapes. C: Part of an arrow.

Sketches can be further segmented by using a *sketch lattice*, a grid which indicates which objects should be grouped together into images. For example, to import the Raven problems in Figure 1 into CogSketch, one would create one sketch lattice for each of the two matrices in a problem, then import the shapes from PowerPoint and place them in the appropriate locations in each lattice. In this way, a user can specify an RPM problem for CogSketch to solve.

Generating Representations

Given a sketch, CogSketch automatically generates a set of qualitative spatial relations between the objects in it. These relations describe the relative position of the objects and their topology—i.e., whether two objects intersect, or whether one is located inside another. CogSketch can also generate attributes describing features of an object, such as its relative size or its degree of symmetry.

CogSketch is not limited to generating representations at the level of objects. It is generally believed that human representations of space are hierarchical (Palmer, 1977; Palmer & Rock, 1994). While there may be a natural "object" level of representation, we can also parse an object into a set of parts or group several objects into a larger-scale set. Similarly, CogSketch can, on demand, generate representations at two other scales: edges and groups.

To generate an edge-level representation, CogSketch parses the lines that make up an object into edges. It does this by identifying discontinuities in a line's curvature that indicate the presence of corners (see Lovett et al., 2009b for details). CogSketch then generates qualitative spatial relations between the edges in a shape, describing relative orientation, relative length, convexity of corners, etc.

To generate a representation at the level of groups, CogSketch groups objects together based on proximity and similarity. It can then identify qualitative spatial relations between groups, or between groups and individual objects.

Interactions with SME

We believe structural alignment plays an important role in comparing visual stimuli. CogSketch employs SME to determine how images relate to each other. However, the use of hierarchical representations means that SME can also compare two objects' edge-level representations to determine how the objects relate to each other. Our model uses this capability in two ways, discussed next.

Finding Shape Transformations CogSketch can compare two objects' shapes to identify transformations between them, e.g., the rotation between the arrow shapes in Figure 2. It does this via a simple simulation of mental-rotation (Shepard & Metzler, 1971): (1) Two objects' edge-level representations are compared via SME. SME's mapping identifies the corresponding edges in the two objects. (2) Pairs of corresponding edges are quantitatively compared to determine whether there is a consistent transformation

between them. In Figure 2, CogSketch could identify a rotation or a reflection between the arrows shapes.

CogSketch can identify two types of shape transformations: equivalence transformations (henceforth called simply *transformations*) and deformations. Transformations (rotation, reflection, and changes in overall size) leave an object's basic shape unchanged. Deformations (becoming longer/shorter, becoming longer/shorter in a part, adding/losing a part) change the object's shape.

Based on shape comparisons, a given set of objects can be grouped into *equivalent shape classes*—groups of objects that have a valid transformation between them, such as equilateral triangles of all sizes and orientations—and *strict shape classes*—groups of objects that are identical, such as upright, equilateral triangles of a particular size.

Comparison-Based Segmentation CogSketch can dynamically segment an object into parts based on comparisons with other objects. For example, to determine the relationship between the images in Figures 2A and 2C, it segments each object into its edges, and uses SME to identify corresponding edges. Grouping only edges in 2A that correspond to edges in 2C enables it to segment 2A into two objects, one of which is identical to 2C. The difference between 2A and 2C is then represented as: *A contains the same object as 2C, but with a second, angular object located above it.*

Our Model

Our model is based on Carpenter, Shell, and Just's (1991) finding that people generally begin solving a matrix problem by comparing adjacent pairs of images in each row of the problem. Our model begins by comparing the images in a row via SME. Based on the mappings between images, it generates a *pattern of variance*, a representation of how the objects change across the row of images. The model then computes a second-order comparison (Lovett et al., 2009B), using SME to compare the patterns for the top two rows and rate their similarity. If the rows are sufficiently similar, the model builds a generalization representing what is common to them; it then looks for an answer that will allow the bottom row to best match this generalization. If the top two rows are not sufficiently similar, the model makes a change to its problem-solving strategy.

Instead of identifying RPM-specific rules as Carpenter et al. did, we utilize two general classes of strategies (four strategies in all) for how a person might go about building patterns of variance. We believe these strategies should be applicable to a variety of spatial problems.

The two classes of strategies are Differences and Literal. Differences involves representing the differences between adjacent pairs of images in a row. For example, in Figure 1A the object is gradually getting smaller. Literal involves representing what is literally true in each image of the row. In Figure 1B, every row contains a square, a circle, and a diamond. There are also advanced versions of each strategy, described below. We now describe each strategy in detail.

Differences Strategy

1) Generate Representations CogSketch generates a spatial representation for each object in a row. While CogSketch can generate representations at multiple levels, the model begins with the highest-scale, and thus simplest, representation. Objects consisting of a single edge—or objects consisting of multiple edges that don't form a closed shape—are grouped together based on connectedness to form a single object, e.g., in the first image of Figure 1F, the vertical and diagonal edges are grouped to form a single object. Objects consisting of closed shapes are combined based on proximity and similarity to form *groups*, e.g., the sets of three squares in Figure 1F are grouped together. CogSketch then computes spatial relationships between the objects, and between objects and groups. It also computes object attributes, describing their shape, color, texture, etc.

2) Compute a Basic Pattern of Variance Consecutive pairs of images in the row are compared via SME to identify the corresponding objects. If there are leftover, unmatched objects in both the first and last images of the row, then these images are also compared. Corresponding objects are then compared to identify transformations between their shapes. Based on these comparisons, the model generates one of the following expressions to describe how an object varies between each pair of images: (*a) Identity*: The object remains the same. *(b) Transformation*: A transformation exists between the shapes. (*c) Deformation*: A deformation exists between the shapes. (*d) Shape Change*: The shapes change entirely. Shape changes are represented as a change between two strict shape classes. Essentially, this is equivalent to a person keeping "square changes to circle" in working memory. (*e) Addition/Removal*: An object is added or removed.

If an object is identical in every image in the row, then this is deemed unimportant, and not explicitly represented[1]. The rest of these expressions are supplemented by any changes in the spatial relations and colors of the images, as identified by SME's candidate inferences, to produce a representation of the pattern of variance across the row.

3) Comparison-Based Segmentation For some problems, the appropriate set of objects to consider only becomes clear after images are compared. For example, in Figure 1E, one discovers after comparison that the third object in the row can be segmented into two parts, such that these parts correspond to the previous two objects in the row. Our model attempts comparison-based segmentation for a set of corresponding objects when: (a) The objects can be broken down into edges, i.e., they aren't filled-in shapes. (b) There is at least one total shape change between the objects, suggesting that they currently don't align well. (c) The changed shapes share some similar parts, i.e., edges with

[1] Carpenter, Just, and Shell (1991) found that the Constant in a Row rule, in which an object remains identical across a row, did not contribute to the difficulty of problems, suggesting that people simply ignore objects that don't change.

similar lengths and orientations. (d) There are no identity matches between objects.

Comparison-based segmentation is performed by breaking the objects into their edges, comparing their edges in a new pattern of variance, and then grouping the edges back together based on which sets of edges correspond across the images. This approach is key in solving Figure 1E. It also allows the model to determine that the vertical line and "X" shape are separate objects in Figure 1F. A similar approach is used to segment groups into subgroups or individual objects when they misalign.

4) Compute Final Pattern of Variance Repeat step 2) after segmentation and regrouping.

Advanced Differences Strategy

The advanced differences strategy is identical, except that in steps 3-4, SME mapping constraints are used so that objects only map to other objects in the same strict shape class (i.e., identical objects). Additionally, objects consisting of single edges (as when the shapes in Figure 1E are broken down into their edges) can only map to other single-edged objects at the same relative location in the image. This means the model will never find object transformations, but it will often find object additions/removals, making it ideal for solving problems like 1E and 1F, in which each object is only present in two of the images in a row.

Literal Strategy

The literal strategy represents what is present in each image in a row, rather than what is different between images. It begins by comparing images to identify any features found in all three images (e.g., the inner shapes in Figure 1B). It abstracts these features out, representing only the features in each image that are not constant across the row. If an object has a different shape from other corresponding objects in the row (e.g., the outer shapes in Figure 1B), then the model includes that object's strict shape class in the representation.

Advanced Literal Strategy

The advanced literal strategy begins by applying the basic literal strategy. It then removes any references to the images in which the objects are found. Spatial relations between objects are also abstracted out. Thus, each object is represented independently, and allowed to match independently from the other objects in its image (e.g., Figure 1D). Alternatively, if each image contains only a single object, then an object is split up and each of its attributes are represented as a separate entity (Figure 1C).

Choosing the Best Strategy

Our model evaluates a strategy by computing patterns of variance for the top two rows and using SME to compare them and rate their similarity. If the similarity is above a threshold, the strategy is deemed a success. If not, a different strategy is tried. The strategies are tried in the following order, which approximates simplest to most complex: Differences, Literal, Advanced Literal, Advanced Differences. If no strategy meets criterion, the model picks whichever Differences strategy receives the highest score—Literal strategies that fail to meet criterion are not considered, since by definition they expect a near-identical match between rows.

Selecting an Answer

Once a strategy is chosen, the model compares the pattern of variance for the top two rows to construct an analogical generalization (Kuehne et al., 2000), describing what is common to both rows. The model then scores each of the eight possible answers. An answer is scored by inserting that answer into the bottom row, computing a pattern of variance, and then using SME to compare this to the generalization for the top two rows. The highest-scoring answer is selected. In cases of ties, no answer is selected.

Solving 2x2 Matrices

The easier RPM sections involve 2x2 matrices. The model solves these by simply computing a Differences pattern of variance for the top row, and then selecting the best answer for the bottom row. If no answer scores above a criterion, the model attempts one strategy change: looking down columns, instead of across rows, to solve the problem.

Evaluation

We evaluated our model by running it on sections B-E of the Standard Progressive Matrices test, for a total of 48 problems. Only section A was not attempted, as this section relies more on basic perceptual ability and less on analogical reasoning. While section B uses 2x2 matrices, sections C-E use 3x3 matrices of increasing difficulty.

Each problem from the test was recreated in PowerPoint and then imported into CogSketch. The experimenters segmented images into objects based on the Gestalt grouping principles (Palmer & Rock, 1994).[3] Recall that the model reorganizes the images into new sets of objects as necessary to solve a problem.

Results

Overall, the model correctly solved 44/48 problems. To compare this level of performance to people, we converted this score to a 56/60 on the overall test, as individuals who performed this well on the later sections typically got a 12/12 on section A (Raven et al., 2000, Table SPM2). A score of 56/60 is in the 75th percentile for American adults, according to the 1993 norms (Table SPM13).

If our model captures the way people perform the test, then problems that are hard for the model should also be hard for people. The four missed problems were among the six hardest problems for human participants, according to 1993 norms (Raven, et al., 2000, Table RS3C3).

[3] In one problem, a dotted line was replaced with a gray line for simplicity.

Discussion

Overall, our model matched the performance of above-average American adults on the Standard Progressive Matrices, both in the problems that it got right and the problems that it missed. Thus, it demonstrates that qualitative representations and the Structure-Mapping Engine can be used to model the performance of typical participants on this task. Importantly, structure mapping played a ubiquitous role in the model; it was used to compare objects, images, and patterns of variance. Additionally, these comparisons were used to rate similarities, identify differences, find corresponding elements, and produce generalizations. Thus, the simulation demonstrates that a single mechanism can be used to perform all the necessary comparisons in this complex task.

Direct comparison with BETTERAVEN (Carpenter, Just, & Shell, 1990) is impossible, as it was only built for, and run on, the Advanced Progressive Matrices. However, if we apply the principles of the model and assume perfect performance, it would achieve a 59/60, missing one of the problems missed by our model. Of the other three problems our model missed, two were due to insufficiencies in its representations of object and group attributes. Because it computes its own representations, our model provides a reason that these problems are more difficult for people, i.e. they require encoding more advanced attributes. Thus, while our model might solve fewer problems, its failures predict and explain human performance.

Future Work

We have shown that our approach is sufficient for modeling human performance on Raven's Progressive Matrices. An important further step is to use the model to make new discoveries about how people perform spatial problem-solving. In a previous study (Lovett & Forbus, 2009), we used a similar model to identify possible cultural differences in the ways people represent space. RPM provides a number of unique opportunities to look at both spatial representation and analogical comparison, due to the complexity and diversity of the problems. By classifying problems based on the model strategies and model components required to solve them, we hope to gain a better understanding of both the factors that make one problem harder than another, and the cognitive abilities that make one person better than another at spatial problem-solving.

Acknowledgments

This work was supported by NSF SLC Grant SBE-0541957, the Spatial Intelligence and Learning Center (SILC).

References

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review, 94,* 115-147.

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices test. *Psychological Review, 97*(3), 404-431.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods, 3*(3), 380-396.

Falkenhainer, B., Forbus, K., & Gentner, D. (1989). The structure mapping engine: Algorithm and examples. *Artificial Intelligence, 41,* 1-63.

Forbus, K., Nielsen, P., and Faltings, B. (1991). Qualitative spatial reasoning: The CLOCK project. *Artificial Intelligence, 51*(1-3), 417-471.

Forbus, K., Usher, J., Lovett, A., Lockwood, K., & Wetzel, J. (2008). CogSketch: Open-domain sketch understanding for cognitive science research and for education. *Proceedings of the Fifth Eurographics Workshop on Sketch-Based Interfaces and Modeling.*

Kuehne, S., Forbus, K., Gentner, D., & Quinn, B. (2000). SEQL: Category learning as progressive abstraction using structure mapping. *Proceedings of the 22nd Annual Conference of the Cognitive Science Society.*

Lovett, A., Lockwood, K., & Forbus, K. (2008). Modeling cross-cultural performance on the visual oddity task. *Proceedings of Spatial Cognition 2008.*

Lovett, A. Forbus, K., & Usher, J. (2007). Analogy with qualitative spatial representations can simulate solving Raven's Progressive Matrices. *Proceedings of the 29th Annual Conference of the Cognitive Science Society.*

Lovett, A., Gentner, D., Forbus, K., & Sagi, E. (2009a). Using analogical mapping to simulate time-course phenomena in perceptual similarity. *Cognitive Systems Research 10*(3): Special Issue on Analogies - Integrating Cognitive Abilities, 216-228.

Lovett, A., Tomai, E., Forbus, K., & Usher, J. (2009b). Solving geometric analogy problems through two-stage analogical mapping. *Cognitive Science 33*(7), 1192-1231.

Markman, A. B., & Gentner, D. (1996). Commonalities and differences in similarity comparisons. *Memory & Cognition*, 24(2), 235-249.

Palmer, S. E. (1977). Hierarchical structure in perceptual representation. *Cognitive Psychology, 9*, 441-474.

Palmer, S., and Rock, I. 1994. Rethinking Perceptual Organization: The Role of Uniform Connectedness. *Psychonomic Bulletin & Review* 1(1): 29-55.

Raven, J., Raven, J. C., & Court, J. H. (2000). *Manual for Raven's Progressive Matrices and Vocabulary Scales.* Oxford: Oxford Psychologists Press.

Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science, 171*, 701-703.

Snow, R.E., & Lohman, D.F. (1989). Implications of cognitive psychology for educational measurement. In R. Linn (Ed.), *Educational Measurement, 3rd ed.* (pp. 263–331). New York, NY: Macmillan.

Vodegel-Matzen, L. B. L., van der Molen, M. W., & Dudink, A. C. M. (1994). Error analysis of Raven test performance. *Personality and Individual Differences, 16*(3), 433-445.

Bridging the Gap: From Cognitive Anthropology to Cognitive Science

Organizers

Andrea Bender (bender@psychologie.uni-freiburg.de)
Sieghard Beller (beller@psychologie.uni-freiburg.de)
Department of Psychology
University of Freiburg, Germany

Presenters

Giovanni Bennardo (bennardo@niu.edu)
Department of Anthropology
Northern Illinois University, USA

Asifa Majid (Asifa.Majid@mpi.nl)
Max Planck Institute for Psycholinguistics
Nijmegen, The Netherlands

James S. Boster (jboster@cognition.clas.uconn.edu)
Department of Anthropology
University of Connecticut, USA

Douglas L. Medin (medin@northwestern.edu)
Department of Psychology
Northwestern University, USA

Although cognitive anthropology once was a pioneer in the cognitive revolution and a founding member of the cognitive science, over the years its participation and influence have diminished—to the detriment of both cognitive anthropology and cognitive science more generally. Meanwhile, though, interactions between culture and cognition are increasingly recognized as being of prime interest for cognitive science. Among the most important issues that call for anthropological expertise is the question of cognitive and/or linguistic universals (Evans & Levinson, 2009; Henrich, Heine & Norenzayan, in press; Norenzayan & Heine, 2005). Anthropology, with its expertise in culture and language, thus becomes an invaluable partner for respective research. But only recently, initiatives have been launched to re-calibrate the relationship among the subfields of cognitive science (Bender, Hutchins & Medin, in press).

This symposium is intended as one step in this direction, bringing together scholars from different disciplinary backgrounds (e.g., anthropology, linguistics, and psychology) to present what they regard as the main strengths of their respective disciplines and why and how this could be useful for each other.

The symposium is co-organized by an anthropologist and a psychologist who will give an introduction to the symposium's topic by summarizing some of the evidence for the cultural constitution of cognition (e.g., Beller & Bender, 2008; Beller, Bender & Song, 2009). The presenters are among the leading scientists in their fields. Besides striving for the re-integration of anthropology into cognitive sciences, each of them has contributed considerably to our expanding knowledge on the cultural constitution of cognition (for instance, in comprehensive monographs or articles in high ranking journals):

- Giovanni Bennardo of Northern Illinois University, having a background in anthropology, linguistics, and cognitive science, seeks to model cognitive conceptualizations for various cultural domains (e.g., Bennardo, 2009; Bennardo & Read, 2007).
- Anthropologist and ethnolinguist James Boster of the University of Connecticut is an expert on methodology in cultural research and on intracultural variation (e.g., Boster, 1999, in press) and has published extensively on semantic categories (e.g., Majid, Boster & Bowerman, 2008).
- Asifa Majid from the MPI for Psycholinguistics in Nijmegen combines approaches from cognitive science, psychology, linguistics, and anthropology for her research into the semantic categorization of so far unquestioned domains as body categorization or sensory experiences (e.g., Majid, 2006; Majid et al., 2008).
- And Douglas Medin, being one of the leading scholars on categorization, learning, and decision making, has for many years now scrutinized the cultural constitution of cognition (e.g., Atran & Medin, 2008; Medin & Atran, 2004; Medin, Bennis & Chandler, in press).

Based on own cross-cultural (and often interdisciplinary) research, each presenter in this symposium will argue why anthropology is necessary for cognitive science and how it can contribute to a more comprehensive understanding of cognition (cf., d'Andrade, 1995; Hutchins, 1995). In particular, they will address the question of universals, from the level of syntax through semantic categories and sensory experiences to the relationship between human and nature.

Word order and a cultural model: From universal mind to cultural mind

Giovanni Bennardo

Goldin-Meadow et al. (2008, p. 9167) suggest that SOV (subject – object – verb) is the "natural [mental] order for humans" and that "as a language community grows and its functions become more complex, additional pressures may exert their influence on language form, in some cases pushing the linguistic order away from the semantically clear ArPA (actor, patient, action or SOV) order". Tongan (in Polynesia) is typically regarded as a Verb-Initial language and specifically a VSO language. In this talk, a frequency analysis will be presented of a good number of Tonga texts that partially challenges this assumption. Besides, a founda-

tional cultural model 'radiality' (Bennardo, 2009) in Tongan cognition will be proposed as the engine that might be responsible for the move from 'natural' SOV to Tongan V-initial.

Are translation equivalents referential equivalents?

James S. Boster

Sets of translation equivalent emotion terms were identified in Polish and English. These terms (and others) were used in two tasks, one naming the emotion expressed in facial gestures of emotion, the other naming the emotions elicited by affectively evocative scenarios. In neither case were the translation equivalent terms referentially equivalent. However, treating the question as one requiring a yes/no answer does not do it justice. This paper measures degrees of translation and referential equivalence and compares those measures.

The senses in mind and culture

Asifa Majid

The cognitive sciences aim to understand the human mind but too often fall prey to unwarranted generalizations from a narrow subset of the population: Western, Educated, Industrialized, Rich, Democratic societies. Anthropologists provide one kind of corrective to this bias, providing ethnographies of many alternative ways of thinking. But we still struggle to grasp what is common across cultural groups, and what truly exceptional. I propose that large-scale cross-cultural comparison can bridge this gap between the fields. For example, it has been assumed that sensory experiences are differentially accessible to language. That is, it is easier to describe distal senses (vision, audition) than proximal senses (olfaction, taste). Current theories assume this to be an established fact on the basis of English data alone. In a large-scale collaborative project, involving 25 researchers and 22 languages, we have found the codability of the senses is culturally-relative. This is a challenge to existing theories.

Cognition in context: Why anthropology and the rest of cognitive sciences need each other

Douglas L. Medin,
Megan Bang, Ananda Marin & Sandra Waxman

There is a great deal to be said about the lack of interaction between Anthropology and the other cognitive sciences. Such analyses can be constructive. Our present focus leaves the abstract issues behind to focus on a set of empirical issues linked to psychological distance and how humans are conceptualized in relation to the rest of nature. Native-American and European-American perspectives are contrasted. The research we report begins with ethnographic observations and interviews and then shifts to an analysis of cultural artifacts (children's books). We show how these data can be used in conjunction with the Trope and Liberman (2003) temporal construal theory to predict a number of related cultural differences. The punch line is that Anthropology and the other cognitive sciences need each other if we are to understand cognition in context.

References

Atran, S., & Medin, D.L. (2008). *The native mind and the cultural construction of nature.* Boston: MIT Press.

Beller, S., & Bender, A. (2008). The limits of counting: Numerical cognition between evolution and culture. *Science, 319,* 213-215.

Beller, S., & Bender, A., & Song, J. (2009). Weighing up physical causes: Effects of culture, linguistic cues, and content. *Journal of Cognition and Culture, 9,* 347-365.

Bender, A., Hutchins, E., & Medin, D.L. (in press). Anthropology in cognitive science. *Topics in Cognitive Science.*

Bennardo, G. (2009). *Language, space, and social relationships*. Cambridge: Cambridge University Press.

Bennardo, G., & Read, D.W. (2007). Cognition, algebra, and culture in the Tongan kinship terminology. *Journal of Cognition and Culture, 7,* 49-88.

Boster, J. S. (1999). Cultural variation. In R.A. Wilson & F.C. Keil (Eds), *MIT Encyclopedia of the Cognitive Sciences* (pp. 217-218). Cambridge, MA: MIT Press.

Boster, J.S. (in press). Data, method, and interpretation in cognitive anthropology. In D. Kronenfeld, G. Bennardo, V. C. de Munck, & M. Fischer (Eds.), *The Blackwell companion to cognitive anthropology.* Cambridge: Blackwell.

d'Andrade, R.G. (1995). *The development of cognitive anthropology.* Cambridge: Cambridge University Press

Evans, N., & Levinson, S.L. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences, 32,* 429-492.

Goldin-Meadow, S., So, W.C., Özyürek, A., & Mylander, C. (2008). The natural order of events: How speakers of different languages represent events nonverbally. *PNAS, 105,* 9163-9168.

Henrich, J., Heine, S.J., & Norenzayan, A. (in press). The weirdest people in the world? *Behavioral and Brain Sciences*.

Hutchins, E. (1995). *Cognition in the wild.* Cambridge, MA: MIT Press.

Majid, A. (2006). Body part categorisation in Punjabi. *Language Sciences, 28,* 241-261.

Majid, A., Boster, J.S., & Bowerman, M. (2008). The crosslinguistic categorization of everyday events: A study of cutting and breaking. *Cognition, 109,* 424-441.

Medin, D. L., & Atran, S. (2004). The native mind: Biological categorization, reasoning and decision making in development across cultures. *Psychological Review, 111,* 960-983.

Medin, D.L., Bennis, W., & Chandler, M. (in press). Culture and the home-field disadvantage. *Perspectives on Psychological Science*.

Norenzayan, A., & Heine, S.J. (2005). Psychological universals: What are they and how can we know? *Psychological Bulletin, 131,* 763-784.

Trope, Y., & Liberman, N. (2003). Temporal construal. *Psychological Review, 110,* 403-421.

Semantic Network Connectivity is Related to Vocabulary Growth Rate in Children

Nicole M. Beckage (nbeckage@indiana.edu)
Linda B. Smith (smith4@indiana.edu)
Department of Cognitive Science, Indiana University, 819 Eigenmann, 1910 E. Tenth Street
Bloomington, IN 47406 USA

Thomas Hills (thomas.hills@unibas.ch)
Department of Psychology, University of Basel, Missionsstrasse 64A
4055 Basel, Switzerland

Abstract

Adult semantic networks show small-world structural properties that are believed to support language processing and word retrieval. The focus of this paper is to understand when these properties emerge in lexical development. We believe that they relate to the rate of word acquisition and vocabulary size. To address this, we examine the connectivity patterns of semantic networks of individual children and compare children on faster and slower vocabulary growth trajectories. The results show that small-world properties emerge early. However, children on slower growth trajectories, who are at risk for significant language delay, do not show these properties. The differences between typical and these so-called "late-talkers" persist, even when vocabulary size is equated. Late talkers' vocabularies are not only acquired later, but also less cohesively, a fact that may relate to future language processing difficulties for these children. In brief, the results suggest that properties of network connectivity may play a role in early lexical development.

Keywords: semantic networks, language acquisition, corpus analyses, late talkers

Words connected to other words

Words exist in a sea of other words. The semantic relations among these words play an explanatory role in language comprehension and processing (e.g., Lund & Burgess, 1996; Jones & Mewhort, 2007). These relations are often studied in terms of semantic networks (Collins & Quillian, 1969; Steyvers & Tenenbaum, 2005). Recent advances in graph theory reveal that adult semantic networks have properties that may be important to language processing, and potentially also to word learning.

Graph theory, or network analysis, can be applied to any structure that consists of nodes connected to each other through links or edges. For example, nodes might be cities and links might be roads; or nodes might be proteins and links might be the molecules that bind with and activate them; or, nodes might be words and the links indices of semantic connectedness such as association strength or co-occurrence.

The semantic networks may be built from various sources, including corpora collected from written or spoken language, free association data, and hand-coded collections of words (e.g., Steyvers & Tenenbaum, 2005; Hills et al., 2009b). As such, they describe the typical mature language user. These mature semantic networks exhibit what is known as small world properties (see Steyvers & Tenenbaum, 2005; Hills et al., 2009a). Small world characteristics allow for local structure but global access. In a network with small world characteristics, there are often clusters of densely connected nodes. The connections between the nodes of a cluster tend to connect to nodes in the same cluster. This contributes to the high local structure. However, there are also a few nodes in these dense clusters that have connections to nodes in other potentially distant clusters. This is the global access that allows easy movement and transition from one cluster to another. Quantitatively, these features are apparent in a high clustering coefficient (a measure of local connectivity) and an average geodesic distance (the shortest path between two nodes) on par with a random network of similar size and connection density. To aid in exposition, these properties are illustrated in Figure 1. Small-world properties are believed to support efficient processing, word retrieval, categorization and robustness to damage and deletion (Hills et al., 2009a; Griffiths, Steyvers & Firl, 2007, Steyvers & Tenenbaum, 2005).

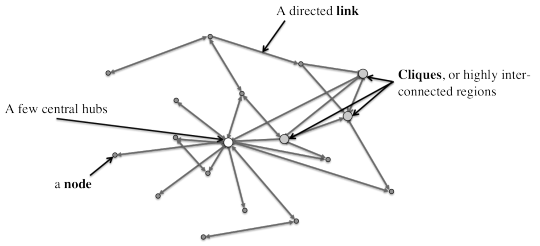


Figure 1: Characteristics of small world structure.

Although it is known that adult semantic networks have small world characteristics, only a few studies have addressed their development and the role of network structure in language acquisition (e.g. Vitevitch 2008, Hills et al., 2009a, Hills et al., 2009b). Here, for the first time, we examine the network structures of the vocabularies of individual children at different points in development. We ask whether small-world properties are dependent on acquiring some number of English words and whether, for any vocabulary size, some children's networks might show more robust connectivity patterns than other children's

networks. Is network connectivity a general fact about the structure of language, or can we show that it is a relevant property at the scale of an individual? Finally, is the connectivity pattern for individual children related to rate of vocabulary growth?

To these ends, we examine the connectivity within the semantic networks of individual children who –by normative standards –are on a path of typical development and children who are on a slower path and one that past research shows is predictive of later language difficulties (e.g., Thal et al., 1997; Bishop & Leonard, 2000; Heilmann, et al, 2005).

Trajectories of Early Vocabulary Growth

Early word learning is first slow and then accelerates (Bloom 2000; Dale & Fenson, 1996), a fact that suggests that already learned words help new word acquisition (see Mitchell & McMurray, 2009). Vocabulary size at any point in development is thus a predictor of future vocabulary growth rates (Dupuy, 1974; Raven, 1948; Bates et al., 1992; Fenson et al., 1993; Thal et al., 1997). Figure 2 illustrates the normative vocabulary size as a function of age for children at the 50th percentile and the 20th percentile (Fenson et al, 1993; see also Dale & Fenson, 1996). Percentile is calculated by considering a child's age, number of words in their productive vocabulary and gender.

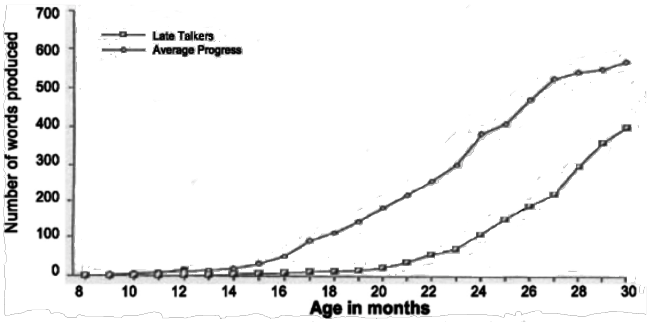


Figure 2:Trajectories of early vocabulary growth, representing children in the 50th percentile and children in the 20th percentile (drawn from Fenson et al, 1993).

The trajectory at the bottom–for children whose vocabulary size falls at or below the 20th percentile of children their same age–has attracted considerable attention in the study of early word learning. Many of these children not only stay on this slower trajectory, but about half go on to have serious deficiencies in language processing and even those who might seem to "catch up" often have measurable difficulties in language tasks (including reading) when they reach school age (e.g Anderson & Freebody, 1981; Bishop & Leonard, 2000; Thal et al., 1997; Moyle et al, 2007). Moreover, early as well as later in development, these children show retrieval errors and word-finding difficulties (Bishop & Leonard, 2000).

Accordingly, we ask how vocabulary size in young children relates to the structure of semantic relations within those vocabularies and whether this structure is related to individual children's rate of vocabulary growth. We examine a broad sample of children and specifically compare vocabularies of children not at risk for language deficits with children whose vocabulary size for their age puts them at risk for language difficulties. In the literature, these at-risk children are often called "late talkers;" we will also use that term although it is somewhat of a misnomer because they are not simply "late" but rather on a slower path of vocabulary growth. If small-world properties are important to the efficiency of language use–and perhaps also to new word acquisitions–then vocabulary structure and not just vocabulary size may be different for these children. Does the connectivity of words in the emerging semantic networks of late talkers differ from the network structure of children whose vocabulary has grown at a more typical pace?

Rationale for the Approach

We analyzed vocabularies from a broad sample of children who differed in age and vocabulary size but whose vocabulary size for age was above the 20th percentile and also from a sample of children, also varying in age and vocabulary size, whose vocabulary size fell below the 20th percentile for age at the time the vocabulary was collected. A semantic network was built for each vocabulary yielding a large set of individual networks that could be ordered by age and separately by vocabulary size.

To build individual networks, we connected the words in an individual's vocabulary, using co-occurrence in a large corpus of child-directed speech as the index of semantic relatedness. The co-occurrences in this corpus of child directed speech is presumed to index the relatedness of the individual words in the language (and that part of the language relevant to children) and in the learning environment in general. This measure of semantic relatedness is *not* the co-occurrences in the specific learning environments of individual children, a key point we will consider in the general discussion. Co-occurrences of words within the corpus formed the edges or links of a semantic network and the nodes were based on the words in each individual child's productive vocabulary.

In sum, the key question is whether and how semantic network connectivity changes as children's vocabularies grow and whether this differs for children whose vocabulary growth rate is sufficiently slow that they are considered at risk for language disorders.

Methods

Vocabularies.

Vocabularies from 73 children ranging in age from 16.2 to 34.6 months were selected for this study. These vocabularies derive from one-time visits of children to the Cognitive Development Laboratory at Indiana University and are measures of productive vocabulary via the Bates-MacArthur Communicative Developmental Inventory (Toddler or Infant form as appropriate to the child's age, Fenson et al, 1993). This is a parent checklist and parents were asked to indicate

which words on the checklist their child produced (Fenson et al, 1993). Total vocabulary as indicated by the parent was used to determine the percentile of the vocabulary size for the child's age. From this repository of child vocabularies we selected a random sample of vocabularies of children whose vocabulary size for their age fell above the 20th percentile and as large a sample as possible of children whose vocabulary size for their age fell below the 20th percentile (see Fenson et al, 1993). Table 1 provides the number of children in each group, means and ranges of their vocabulary size, age, and percentile.

Table 1: Age and percentile of children in study

	# children	Age range in months (mean)	Percentile range (mean)
All children	73	16.2-34.6 (22.1)	5-99 (25.6)
Late talkers	38	16.3-34.6 (24.3)	5-20 (12)
Typical talkers	35	16.2-26.6 (19.8)	25-99 (40.4)

Words.

For the network analysis, only the 291 words that are on both the Toddler and Infant forms were used. This allowed for a more accurate comparison across ages. Of the included words, 204 are nouns, 51 are verbs and the remaining 36 are adjectives, adverbs and function words.

Networks.

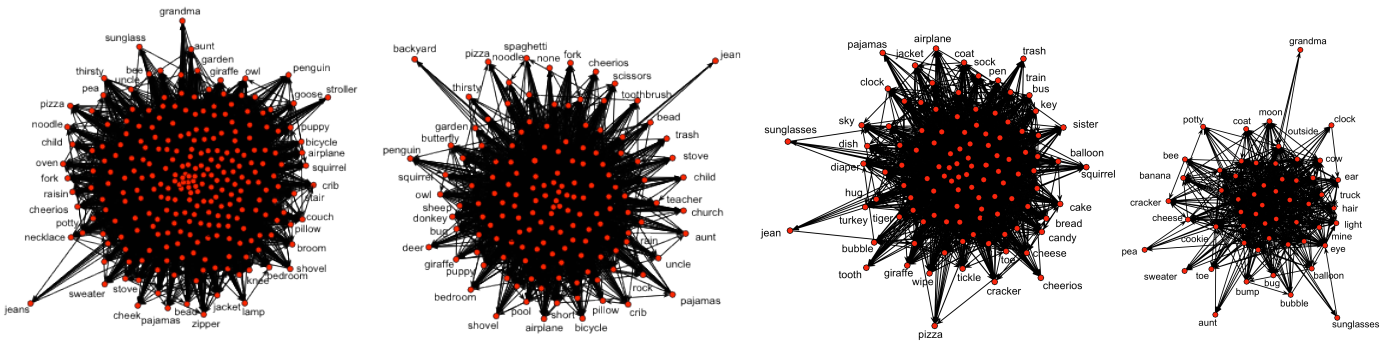
To build the networks, links between words were defined in terms of co-occurrences in the CHILDES database (MacWhinney, 2000). The co-occurrence method was taken from prior analyses by Riordan and Jones (2007) and related lemmas (cat, cats, hit, hitting) were counted as instances of the same lexeme. The matrix of co-occurrences was built using a process similar to the Hyperspace Analogue to Language (HAL) (Lund & Burgess, 1996) and the word co-occurrence detector (Li, Farkas & MacWhinney, 2004). For the 291 unique words, we formed a *291x 291 matrix*, where each cell, *ij*, is filled according to the following rule: a moving window of size 15 moves word-wise through the corpus, with each cell *ij*, changed to a value of 1 if word *j* occurs both downstream and together in the same window with word *i*. This produces a directed network where each word is connected to another word by a directed link if it co-occurs downstream of that word in child directed speech. Frequency counts were taken as the number of occurrences of a given word in the corpus.

Results

The analyses reported here use four network statistics: *median in-degree*, *global clustering coefficient*, *redundancy*, and *geodesic distance*. Each provides a means of assessing connectivity within networks. Figure 3 shows four networks for four typically developing children and the index of connectivity for each of these networks. The four individual networks show considerable small-world structure with as few as 106 (or even 55) words. This suggests that these properties–characteristic of mature semantic networks–are evident even from the earliest stages of lexical development. This could merely reflect the structure of language such that any learner (or random sample of words from early vocabularies) would show these properties. Or, these properties could be more fundamentally related to how individual children build semantic structures for efficient language learning and processing. The comparison of typically-developing and late-talking children provides the relevant evidence.

In-degree.

In-degree is a measure that captures how many connections each node has directed towards it from other nodes. In the present case, the in-degree of the target word or node is the number of distinct words that occurred 15 or fewer words after the target in the CHILDES corpus. The median in-degree provides an overall picture of how sparse or dense a network is. In a sparse network, the words in the



# of words	256	149	106	55
Median in-degree	53.5	29	36	22
Clustering coefficient	.472	.494	.555	.606
Geodesic dist.	1.720	1.737	1.608	1.566
Redundancy	16.04	14.15	9.23	4.84

Figure 3: These semantic networks of typically developing children show that children develop small world structure even with relatively few words. Throughout development the semantic networks of children show high clustering coefficients and low average geodesic distance. The networks also quickly develop a high number of connections and multiple traversable pathways.

vocabulary are not as related to each other and so there can be more words than connections. In a dense network, many words are connected to each other; e.g., the median in-degree is nearly equal to the total number of words or nodes, many words in the network are semantically related and co-occur frequently in speech.

Regression analysis, with median in-degree as the independent variable and the child's MCDI percentile as the dependent, yielded a significant relation between in-degree and percentile with lower median values characterizing late-talkers even when age ($p<.001$) and vocabulary size ($p=.0162$) were controlled. The relation between in-degree and vocabulary size for the two groups is shown in Figure 4.

This indicates that there are more links in a typical talker's network than in a late talker's network even when the networks have the same number of nodes. Typical talkers learn words that are semantically connected to each other but late talkers are less likely to do so, as if perhaps, they learn words as individual islands, as if the *next* word learned is somehow independent of the prior learned words.

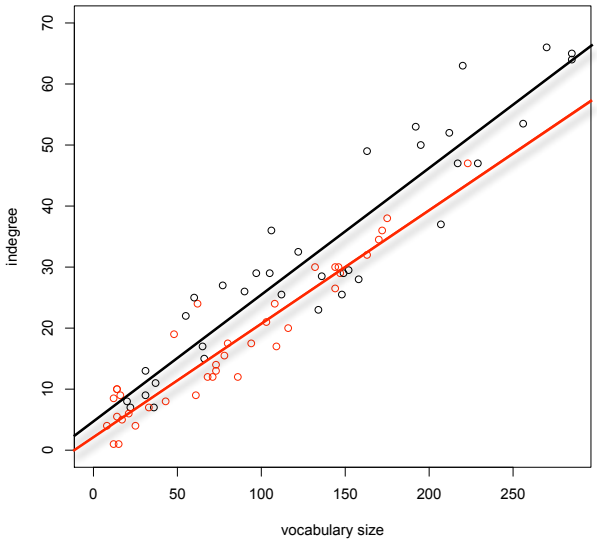


Figure 4: A graph of the median in-degree as a function of vocabulary size. The black line indicates typical talkers and the lighter line, the late talkers. ($p=0.016$).

Global clustering coefficient.

The clustering coefficient provides a measure of how well connected a node's neighbors are to each other. Small-world networks have high clustering coefficients, relative to networks of the same size (number of nodes) and density (ratio of observed links to possible links). A clustering coefficient of 1 indicates that all of a node's neighbors are themselves connected. A clustering coefficient of 0 indicates that none of a node's neighbors are connected to one another. This provides a measure of local clustering, as opposed to more global measure of density assessed with in-degree above. The late-talkers in the present study show a lower average clustering coefficient than late talkers when age is controlled ($ß=-54.6$, $SE=21.991$, $p=0.0154$) and a near significant effect when vocabulary size is controlled ($ß=34.53$, $SE=18.33$, $p=0.0638$). Figure 5 shows the clustering co-efficient as a function of vocabulary size for the two groups. As is apparent from the data points, there is both more variability by this measure among the youngest later-talkers than typically developing children and typically developing children appear to move toward a stable clustering coefficient earlier than do late talkers. The lower average clustering coefficient of late talkers suggests that they are less likely to learn words that fill out categories of closely related words that they already know, a result that again suggests that there may be fewer dependencies between new acquisitions and already learned words. Being unable to fill out categories of closely related words, these late talkers may have trouble reorganizing their current semantic understanding to create new categories and concepts.

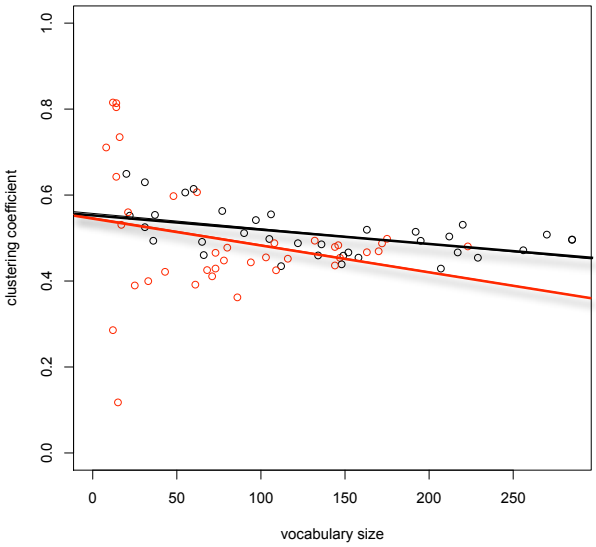


Figure 5: A graph of the clustering coefficient as a function of vocabulary size. The black line indicates typical talkers and the lighter line, late talkers (lm, $p=0.064$).

Redundancy.

Redundancy captures the robustness of the network: in a highly redundant network, if a random connection is deleted, the deleted link will not alter the likelihood of a connected path between two words. For example, with a road network, if there are multiple ways to get between two places, then a road closure is not an insurmountable problem. However, if only one road connects two locations, then a closure of that road makes the two locations inaccessible to one another. Higher redundancy means more possible paths. As opposed to clustering coefficient and in-degree, redundancy provides a measure of the ease of accessibility in the network (from one node or word to another). Compared with the clustering coefficient, this provides a more global measure of cohesion across the network.

Regression analyses yielded significant differences between the two groups, with late talkers having less redundant networks when controlling for age ($p<.001$) and

vocabulary size (p=0.015). To quantify this, for a given network of 200 words, a late talker would have on average 11 possible pathways compared to 13 possible paths in a network for a typical talker (t-test, p= 0.016, comparing all late talkers/typical talkers). Though this difference is small, the actual implication of this difference is that late talkers have many words that have only one or two connections, whereas typical talkers have fewer words with low redundancy suggesting a network more robust to change.

The difference in the number of possible pathways between nodes across the two groups suggests that the robustness of the two groups is also different. The typical talkers, building more redundant networks, are less likely to have trouble transitioning from one area of the network to another. The fluidity of their productive speech also would be less hampered by the forgetting of a few words. By having multiple ways of getting from one word to another, the typical talkers may more easily access one word following another. These differences may relate importantly to the word-finding and word-retrieval difficulties of late talkers, an important question for future work.

Average geodesic distance.

Geodesic distance represents shortest path length between two nodes. We computed the geodesic distance between all nodes excluding isolates or unconnected nodes. We then averaged the geodesic distance for all nodes, further excluding all cases in which there was no traversable path between two nodes. As networks grow larger, more connections are possible and the geodesic distance, or the shortest distance between two nodes, will often trend toward less than 2. This happens when a word that connects to all other words, such as "you", is added to the semantic network. If word A is not directly connected to word B, word A is connected to word B through "you", resulting in an average geodesic distance of approximately 2.

Late talkers have significantly different geodesic distances from typical talkers. When considering networks of similar size (i.e. words known), we see that typical talkers having a mean geodesic distance of 1.82 and late talkers having a mean geodesic distance of 2.55 (t-test, p=0.0276).

Another indication that these at-risk children are building networks with less global structure is the number of components in a network. Components are isolated clusters or words of a network that do not connect to other components in a network. Early on in vocabulary learning, it is possible to learn a word, or words, in complete isolation that is not semantically related to any other word or cluster. For example a child might learn a bunch of animals and a bunch of food words but be missing words like milk that would link the two clusters. Of the children in this study only 17 children showed networks that had more than one component, 14 of which are classified as late talkers.

The difference in geodesic distance and number of components suggests that late talkers are not building networks that allow for the same level of global access.

Discussion

The present study is the first analysis of the network structures of early vocabularies for individual children and the first to reveal potentially meaningful individual differences in the structures of these emerging networks. As such, there are still open questions and limitations that will need to be addressed. These include comparisons to randomly selected vocabularies of different sizes, linking of these differences in vocabulary structure to performance (such as word retrieval), and following individual children's vocabulary growth. Nonetheless, the results provide three new insights: (1) Small-world properties are evident in the network structure of even very small and early vocabularies; (2) these properties are not the consequence of just learning any subset of early English words since–at any vocabulary size–there are individual children with more robustly connected networks than other children; and (3) the structure of these individual differences in network connectivity appears related not just to vocabulary size but to the rate of vocabulary development with children at risk for serious language deficiencies (by normative standards) showing less cohesive and less efficiently structured networks.

The broad sample of typically-developing children, children above the 20th percentile and who are not at risk for language deficiencies, show less variance in network structure, specifically clustering coefficient in our analysis, than do the late-talking children, a remarkable fact in its own right. These typically-developing children seem to be building semantic networks with many of the small-world properties found in adult semantic networks, showing higher in-degree, clustering coefficient, and redundancy, indicating that typical talkers are learning words more cohesively, with more semantic connectivity between learned words—both globally and locally—than do the networks of late talkers. Late talkers are not only learning more slowly but appear to be learning differently. One possibility consistent with the present pattern is that typically developing children build their vocabularies in ways such that learning itself is dependant on the semantic relations among already learned words or the semantic relations in the learning environment (Hills et al, 2009b) whereas late talkers just learn words, adding words as individual and unrelated items, not picking up on the semantic relations in the learning environment.

Because the semantic relations in these networks are themselves normative–reflecting the structure of the general learning environment and not the child's specific learning environment-it is also, in principle, possible that these children's learning environments present less semantic connectivity. Previous research has shown that learning environments, in terms of the kind and number of words that are spoken to children, do influence the kinds and number of words that children learn (e.g., Hurtodo, Marchman & Fernald, 2008; Rowe, 2008; Hoff & Naigles, 2002: Huttenlocher et al, 1991). However, contemporary understanding of language-delayed children suggests that this may not be the sole factor in these delays (see Bishop &

Leonard, 2000). Still, a more detailed examination of individual language learning environments is in order.

Our evidence suggests that typical talkers are more likely to acquire words that share semantic associations with words they already know. This may be a consequence of the fact that they are more sensitive to semantic associations in the environment (what has been called *preferential acquisition*), or that they are more likely to use known words to direct the acquisition of new words (called the *lure of the associates*). Previous work has shown that both of these processes are predictive of word acquisition (Hills et al., 2009b), but these processes may also represent individual strategies for learning. This suggests an interesting direction for future research in individual differences in language acquisition.

Acknowledgments

This research was supported by NICCHD grant HD028675 to Linda Smith.

References

Anderson, R. C., & Freebody, P. (1981). Vocabulary knowledge. In J. T.Guthrie (Ed.), *Reading comprehension and education* (pp. 77–117). Newark, DE: International Reading Association.

Bates, E., Marchman, V., Thal, D., Fenson, L., Dale, P., Reznick, J. S., Reilly, J. & Hartung, J. (1992). Developmental and stylistic variation in the composition of early vocabulary. CRL Technical Report, UCSD.

Bishop, Dorothy V. M. (Ed) (1), & Leonard, L. B. (. (Eds.). (2000). *Speech and language impairments in children: Causes, characteristics, intervention and outcome*. New York, NY, US: Psychology Press.

Bloom, P. (2000). *How children learn the meanings of words*. Cambridge, MA: MIT Press.

Collins, A.M., & Quillian, M.R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior, 8*, 240-248.

Dale, P.S., & Fenson, L. (1996). Lexical development norms for young children. *Behavior Research Methods, Instruments, & Computers*, *28*, 125–127.

Dupuy, H. P. (1974). *The rationale, development and standardization of a basic word vocabulary test (DHEW Publication No. HRA 74-1334)*. Washington, DC: U.S. Government Printing Office.

Fenson, L., Dale, P. S., Reznick, J. S., Thal, D., Bates, E. and Hartung, J. P., et al. 1993, The MacArthur Communicative Development Inventories: User's Guide and Technical Manual (San Diego: Singular)

Griffiths, T.L., Steyvers, M., & Firl, A. (2007). Google and the mind: Predicting fluency with PageRank. *Psychological Science, 18*, 1069–1076.

Heilmann, J., Weismer, S. E., Evans, J., & Hollar, C. (2005). Utility of the MacArthur-bates communicative development inventory in identifying language abilities of late-talking and typically developing toddlers. *American Journal of Speech-Language Pathology, 14*(1), 40-51.

Hills, T., Maouene, J., Sheya, A., Maouene, M., and Smith, L. (2009a). Emergent categories in the feature structure of early-learned nouns. *Cognition*, 112, 381-396.

Hills, T., Maouene, J., Sheya, A., Maouene, M., and Smith, L. (2009b). Longitudinal analysis of early semantic networks: preferential attachment or preferential acquisition? *Psychological Science*, 20, 729-739.

Hoff, E., & Naigles, L. (2002). How children use input to acquire a lexicon. *Child Development*, 73, 418-433

Hurtado, N., Marchman, V. A., & Fernald, A. (2008). Does input influence uptake? links between maternal talk, processing speed and vocabulary size in spanish-learning children. *Developmental Science,* 11, F31-F39.

Huttenlocher, J., Haight, W., Bryk, A., Seltzer, M., & Lyons, T. (1991). Early vocabulary growth: Relation to language input & gender. *Developmental Psychology,* 27, 236-248.

Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, *114*(1), 1–37

Li, P., Farkas, I., MacWhinney, B. (2004). Early lexical development in a self-organizing neural network. *Neural Networks* 17, 1345-1362.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavioral Research Methods, Instruments, and Computers, 28*(2), 203-208.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk* (3rd ed.). Mahwah, NJ: Erlbaum.

Mitchell, C., & McMurray, B. (2009). On leveraged learning in lexical acquisition and its relationship to acceleration. *Cognitive Science*, 33, 1503-1523.

Moyle, M. J., Weismer, S. E., Evans, J. L., & Lindstrom, M. J. (2007). Longitudinal relationships between lexical and grammatical development in typical and late-talking children. *Journal of Speech, Language, and Hearing Research, 50*(2), 508-528.

Raven, J. C. (1948). The comparative assessment of intel-lectual ability. *British Journal of Psychology*, *29*, 12–19.

Riordan, B., and Michael N. J.. (2007). Comparing semantic space models using child-directed speech. In D. S MacNamara & J. G. Trafton (Eds.), *Proceedings of the 29th Annual Conference of the CogSci*. 599-604.

Rowe, M. L. (2008). Child-directed speech: Relation to socioeconomic status, knowledge of child development & child vocabulary skill. *Journal of Child Language,* 35, 185-205.

Steyvers, M., & Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: statistical analyses and a model of semantic growth. *Cognitive Science, 29*, 41-78

Thal, D., Bates, E., Goodman, J., & Jahn-Samilo, J. (1997). Continuity of language abilities: An exploratory study of late and early-talking toddlers. *Developmental Neuro-psychology, 13*, 239–273.

Vitevitch, M.S. (2008). What can graph theory tell us about word learning and lexical retrieval? *Journal of Speech, Language, and Hearing Research, 51*, 408–422.

Effects of Maternal Input on Language in the Absence of Genetic Confounds: Vocabulary Development in Internationally Adopted Children

Carissa L. Shafto[1], Joy Celeste Geren[2], & Jesse Snedeker[2]

[1]**(carissa.shafto@louisville.edu)** Dept. of Psychological and Brain Sciences, University of Louisville, Louisville, KY 40292
[2] **(geren@fas.harvard.edu, snedeker@wjh.harvard.edu)** Dept. of Psychology, Harvard University, Cambridge, MA 02138

Abstract

Parents provide children with both genes (nature) and linguistic input (nurture). A growing body of research demonstrates that individual differences in children's language are correlated with differences in parental speech. Although this suggests a causal link between parental input and the pace of language development, these correlations could reflect effects of shared genes on language, rather than a causal link between input and outcome. We explored effects of maternal input on English vocabulary development in internationally-adopted (IA) children—a population with no genetic confound. IA preschoolers demonstrated some of the same correlations with input as in previous studies; specifically, measures of input quality were significantly correlated with vocabulary. However, IA infants did not demonstrate this pattern. Differences between the age groups may be related to the pace of acquisition; more rapid vocabulary development in the preschoolers suggests that access to, and children's ability to make use of input, may be a limiting factor for the infants.

Introduction

There is a growing body of research demonstrating that individual differences in children's linguistic abilities are correlated with differences in parental speech (e.g., Hart & Risley, 1992, 1995; Hoff, 2003b; Zimmerman et al., 2009). While these studies and others strongly suggest that variation in parental language input contributes to variability in language development, such studies have an unavoidable confound: biological parents provide children with linguistic *and* genetic input. In fact, twin studies consistently find that language skills have moderate to high heritability (Stromswold, 2001) and Plomin and Dale go so far as to say "a case could be made that verbal measures are among the most heritable traits" (2000, p. 39). Rather than a direct causal link between input and outcome, these correlations between parental input and child outcomes could potentially reflect direct effects of shared genes on the verbal abilities of both parties. Here we investigate the role of maternal input in children's vocabulary acquisition when the influence of genetics is absent.

We start by discussing the existing literature on variability in maternal input and evidence for relations between input and child language outcomes. Then we present two experiments with IA children adopted at different ages to explore potential differences in uptake related to different paces of language acquisition. Then we conclude by discussing recent findings on the role of genetics in language development and how our results reconcile the gene-environment confound present in previous studies.

Variability in Maternal Language Input

An early study of differences in caregiver input (Elardo, Bradley, & Caldwell, 1977) investigated the home environment and language abilities of 74 typically developing children living in an urban setting. The majority of the children were African-American and one-third were on welfare at the time of the study. Caregiver input was measured via a home environment assessment (the Home Observation for Measurement of the Environment; Caldwell, Heider, & Kaplan, 1966) and children's language abilities were assessed with the Illinois Test of Psycholinguistic Abilities (Kirk, McCarthy, & Kirk, 1968). The study found that maternal involvement, maternal responsiveness, and providing appropriate play materials had the strongest correlations with children's language.

This study is part of a growing body of research linking individual differences in caregiver demographics to differences in their speech (e.g., Hart & Risley, 1992, 1995; Hoff, 2003b). In their seminal paper, Hart and Risley (1992) described the qualitative aspects of parental speech in 40 diverse families. The qualitative aspects of the parents' speech to their children were strongly related to socio-economic status (SES); parents of higher SES were more verbal and had higher quality verbal interactions with their children. Hoff (2003b) found that mothers' mean length of utterance, number of word types, and number of tokens were each uniquely correlated with SES. Hoff also found that mothers' speech to adults varied with SES (2003a).

More recently Huttenlocher and colleagues examined caregiver speech to young children from 50 ethnically and economically diverse families via home video recordings (Huttenlocher, Vasilyeva, Waterfall, Vevea, & Hedges, 2007). Data were presented from 5 different time points collected when the target children were between 14 and 30 months old. The authors analyzed the composition of speech, the diversity of speech, and the quantity of speech. The results suggest that caregivers' education levels were significantly predictive of the quantity of spoken language and that this relation was more predictive than family income level. They also found that the complexity and diversity of caregiver speech increased linearly over time, while input quantity remained relatively stable.

Effects of Input on Language Development

One might expect such significant SES-related differences in maternal speech to affect children's language development; this is precisely what is found (see Whitehurst, 1997 for review). In an early study with middle-

class mothers Huttenlocher et al. found that the quantity of maternal language spoken significantly correlated with children's vocabulary growth from age 14 to 26 months (Huttenlocher, Haight, Bryk, Seltzer, & Lyons, 1991).

In a study of 22 mother-child dyads varying in SES Hoff-Ginsberg (1986) found several aspects of mothers' speech to correlate with children's language outcomes during the third year of life. Both functional and structural characteristics of maternal speech were predictive of children's language outcomes (e.g., the average number of noun phrases per utterance in maternal speech was predictive of the same feature in children). Another study by Hoff-Ginsberg (1998) found differences between siblings relating to birth order, though genetic influences are similar among siblings.

More recently, in a study of 33 high-SES families and 30 low-SES families Hoff (2003b) found that maternal mean length of utterance, number of word types, and number of word tokens were each uniquely predictive of children's vocabulary size. SES-related differences in maternal speech mediated children's language development such that children with low-SES mothers heard (on average) less rich language input and consequently had less developed language abilities. This finding is supported by a recent review of the literature on how SES relates to brain development (Hackman & Farah, 2009). Hackman and Farah reviewed studies of SES effects on neurocognitive development and found the strongest effects of SES on the brain areas associated with language and executive function.[1]

In a more recent study focused on children learning Spanish as a first language, Hurtado and colleagues found that maternal input correlated with children's vocabulary growth from 18 to 24 months (Hurtado, Marchman, & Fernald, 2008). In addition, speed of word recognition at 24 months was related to quantity of maternal input. The effects of maternal speech on vocabulary size and word recognition speed overlapped considerably, suggesting that these abilities work together in lexical acquisition. Taken together these findings suggest that the observed correlations between parental input and child output may reflect a causal role of the input in language development (see also Weizman & Snow, 2001).

The Current Study

The current study extends this work by exploring the effects of maternal input on early vocabulary development in internationally-adopted (IA) children—a population which eliminates genetic confound. We previously demonstrated that early language acquisition in this population shows the same qualitative patterns that characterize typical language development, suggesting that similar learning processes may be at work (Snedeker, Geren, & Shafto, 2007).

In a more recent study (Snedeker, Geren, & Shafto, in press) we found that the rate of vocabulary acquisition in IA infants was explained primarily by chronological age, while the rate of acquisition in IA preschoolers was explained primarily by time spent learning (i.e., months in the U.S.). Additionally, the preschool-aged IA children acquired English significantly faster than the IA infants, suggesting that for children adopted at older ages the developmental patterns in the early stages of English acquisition occur on an accelerated time table. The quantity and nature of language input may be even more critical to the pace of acquisition in older learners. The current study explored this possibility through experiments with children adopted in two distinct age groups.

In Experiment 1 we assessed English vocabulary in IA children adopted during the preschool years. In Experiment 2 we assessed English vocabulary in IA children adopted as infants, who may not learn English at such an accelerated rate due to less advanced cognitive abilities (e.g., memory). We tested two age groups to explore potential differences in the effects of maternal input due to differences in the pace of language development.

Experiment 1

Method

Participants Twenty-nine children aged 2;9 to 5;2 years who were adopted from Eastern Europe and China between the ages of 2;5 and 4;11 (M: 3;1 years) and had been in the U.S. for 0.5–6 months (M: 3.4 months) at the first assessment. All children were adopted by monolingual English speakers and were typically developing.[2]

All of the children were adopted into upper-middle class homes, with the majority of mothers having earned graduate or professional degrees (N=17). The other mothers earned a college degree (N=9) or attended some college (N=3).

Materials & Procedure Parents participated in monthly sessions until their child had been in the U.S. for 6 months; thus each child had 1–6 sessions (total=63). For each session parents completed the Words and Sentences form of the MacArthur-Bates CDI[3] (CDI-2; Fenson et al., 2006) and recorded a language sample in their home. Families were sent a standard box of toys to use for the language sample, which were an average of 27 minutes long and were transcribed and analyzed using the CLAN program (MacWhinney, 2000).

Measures Once the language samples were transcribed, maternal utterances were coded for quantitative and qualitative features. The maternal input *quantity* variable was the number of words spoken per minute. Maternal input *quality* variables included: mean length of utterance (MLU), the number of word types spoken per minute (a measure of input diversity), percentage of utterances that were yes/no

[1] These findings are preliminary and do not preclude effects of SES on other cognitive domains.

[2] According to a parent report.

[3] We validated the use of the CDI-2 with this population in a previous study (Snedeker et al., in press).

questions (*Is that your crayon?*), percentage of utterances that were wh-questions (*What color is that?*), and the percentage of utterances that were alternative questions (*Do you want to play with the truck or the car?*).

Because children had varying numbers of sessions, the maternal input variables were calculated for each session, and the average values for each variable were used as the predictors for that child. This means that for children with more than one recording session, no particular data point was chosen for use (which could have biased the results), and no data points were represented more than once in the analyses.

Results

CDI-2 'norms' were calculated using data from a larger study of IA preschoolers (N=182). Stepwise regressions were conducted on CDI-2 vocabulary score with Time in the U.S. (R^2=.54, p<.001) and Age of Arrival (R^2=.03, p<.001) as predictors.[4] Results were used to calculate standardized residual scores (SRSs) for vocabulary for the final session of the 29 participants in the current study. Specifically, the SRSs were used as a measure of how different children's reported vocabularies were from their predicted vocabulary. Thus a negative SRS would indicate that a child's reported vocabulary was lower than would be predicted by their Age of Arrival and Time in the U.S.

As a first pass raw correlations were conducted between the maternal input variables and children's SRS (see Table 1). Then step-wise regressions were conducted on children's SRSs using the maternal input variables (averaged across sessions) as predictors. At Step 1 maternal word types per minute was a significant predictor of SRS (adjusted R^2=.56, p<.001; see Figure 1). This suggests mothers with more diverse input had children with higher SRSs; their children exceeded their predicted vocabulary by larger amounts.

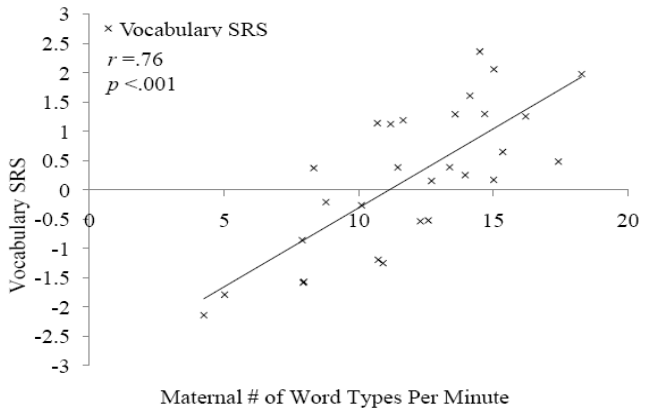


Figure 1: IA Preschoolers' Vocabulary SRSs by Diversity of Maternal Input (Experiment 1).

Percentage of maternal utterances that were yes/no questions accounted for additional variance (adjusted R^2=.07, p<.001) suggesting that mothers who asked more yes/no questions had children with higher SRSs. This suggests that higher levels of prompting or engagement facilitated vocabulary growth.

Contrary to previous findings, words per minute (*quantity*) was not a reliable predictor of SRS (partial R^2=.002, p=.70). However, this may be due to the high correlation between words per minute and word types per minute (see Table 1).

Experiment 2

Method

Participants Seventeen children aged 1;7 to 2;8 who were adopted from China between 8 and 16 months old (M: 12 months) and had been in the U.S. for 7–20 months (M: 15 months) at the first session. All children were adopted by monolingual English speakers and were reported to be typically developing.

All children were adopted into upper-middle class homes, with the majority of mothers having earned graduate or professional degrees (N=11). The other 6 mothers had all earned a college degree.

Materials & Procedure Parents completed monthly sessions until their child was 32 months old; thus each child had 1–12 sessions (total=71). Three of the children had some of their language samples recorded with their father instead of their mother. In order to maximize homogeneity across the language samples for all participants, individual sessions that were recorded with the father were excluded from analyses. This left a total of 64 sessions for analyses.

Measures The same as in Experiment 1.

Results

First CDI-2 'norms' were calculated for the IA infants using data from a larger study of IA infants (N=223). Step-wise regressions were conducted on CDI-2 vocabulary score with Age at Test (R^2=.45, p<.001) and Age of Arrival (R^2=.03, p<.001) as the predictors.[5] Results were used to calculate standardized residual scores (SRSs) for the final session of the 17 IA infants in the current study. As a reminder, a child's SRS represents the difference between their reported and predicted English vocabulary (i.e., a *z* score).

As in Experiment 1, raw correlations were first conducted to determine the relations between the maternal input variables and children's SRSs (see Table 2). Step-wise regressions were then conducted on children's SRSs using maternal input variables (averaged across sessions) as predictors. The percentage of alternative questions

[4] Children's Age at Test was not used as a predictor because it was significantly positively correlated with both Age of Arrival (r=.95, p<.001) and Time in the U.S. (r=.19, p<.05).

[5] Time in the U.S. was not used as a predictor because it was significantly correlated with both Age at Test (r=.89, p<.001) and Age of Arrival (r=-.28, p<.001).

Table 1: Correlation matrix for Experiment 1 (IA preschoolers).

Measure	Words per minute	MLU	Word types per minute	Yes/no questions (% of utterances)	Wh-questions (% of utterances)	Alternative questions (% of utterances)
Words per minute (word tokens)	---					
MLU	.56**	---				
Word types per minute	.82**	.57**	---			
Yes/no questions (% of utterances)	.04	.14	.06	---		
Wh-questions (% of utterances)	-.29	-.03	-.05	.50**	---	
Alternative questions (% of utterances)	-.14	.02	-.03	.51**	.65**	---
Standardized residual vocabulary score (SRS)	.64**	.53**	.76**	.33	.01	.07

**p<.01

Table 2: Correlation matrix for Experiment 2 (IA infants).

Measure	Words per minute	MLU	Word types per minute	Yes/no questions (% of utterances)	Wh-questions (% of utterances)	Alternative questions (% of utterances)
Words per minute (word tokens)	---					
MLU	.61**	---				
Word types per minute	.57*	.62**	---			
Yes/no questions (% of utterances)	-.14	-.19	-.09	---		
Wh-questions (% of utterances)	-.12	.00	-.26	.52*	---	
Alternative questions (% of utterances)	.48^T	.41	.09	.02	-.12	---
Standardized residual vocabulary score (SRS)	.40	.16	.18	.20	-.21	.70**

$^T p$<.06, *p<.05, **p<.01

(e.g., "*Do you want to play with the truck or the car?*") was the only significant predictor of SRS (adjusted R^2=.47, p<.01). Mothers who asked more alternative questions had children with higher SRSs—children who exceeded their predicted vocabulary by greater amounts (see Figures 2 and 3).

Contrary to Experiment 1, maternal word types per minute and percentage of yes/no questions were not reliable predictors of SRS (adjusted R^2=-.03, p=.51; adjusted R^2=-.001 p=.33, respectively). This suggests that the features of maternal speech that seem to influence English vocabulary growth in IA children adopted as preschoolers might be less influential for IA children adopted as infants. Alternatively, the smaller sample size in the infant group may have made any additional effects of maternal input variables undetectable.[6] As in Experiment 1, but contrary to previous findings, words per minute (input *quantity)* was not a reliable predictor of SRS (incremental R^2=.005, p=.72). Unlike in Experiment 1, the raw correlation with words per minute was not significant either (see Table 2). However, the correlation value was moderate (.40), so one possibility is that the effect was suppressed by the variability present in our small sample.

[6] For a moderate correlation (r=.5) with power of 80% a minimum sample size of 28 children is needed.

General Discussion

There were significant relations between some *qualitative* aspects of maternal input—maternal word types and yes/no questions—and English vocabulary ability for the preschool-aged IA children. This is in accord with previous findings of a positive relation between maternal input and children's vocabulary development. Curiously, the relation between input and outcome differed in the two age groups. This difference occurred despite the fact that both age groups were adopted into families with similar SES (high), and thus were like the professional families from Hoff (2003b) who received quite rich language input.

One possible explanation for the difference between age groups is that perhaps older children are more sensitive to variation in input. IA children in both age groups are receiving input that is likely greater in quantity and quality than the general population (due to their high SES environment). However, the IA infants may be immersed in such a rich language environment that their maturational status may be limiting their ability to take advantage of the high quality and quantity of the input they are receiving. Specifically, the IA preschoolers may be more ready to make use of the input because their other cognitive skills (e.g., memory) are more fully developed. They can learn

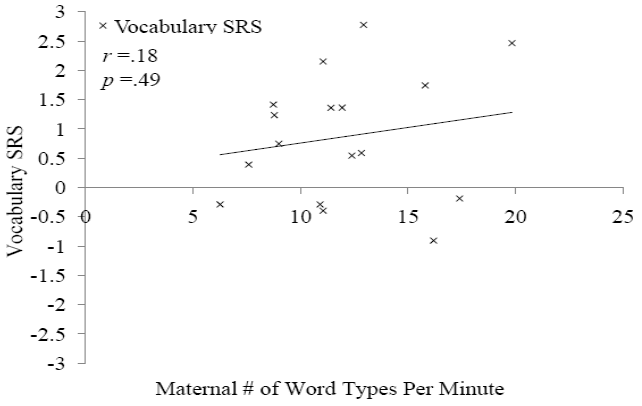


Figure 2: IA Infants' Vocabulary SRSs by Diversity of Maternal Input (Experiment 2).

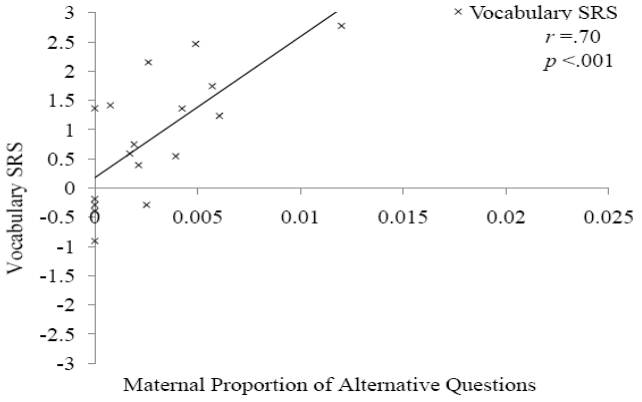


Figure 3: IA Infants' Vocabulary SRSs by Alternative Questions in Maternal Input (Experiment 2).

faster so input is more likely to be a rate-limiting factor. It may also feel more natural to speak more to an older child (IA preschooler) and more input likely results in greater variability in input that the child can exploit. There may be less variation in the input for the IA infants, providing less for the child to exploit.

Contrary to previous studies, we did not find a significant effect of maternal input *quantity* on vocabulary. However, the raw correlations between input quantity and vocabulary were significant for the IA preschooolers. The correlation likely disappears when put into the regression due to the high correlation between number of words and number of word types, with number of word types soaking up all the variance in children's SRS. This suggests that the amount of input may have an effect, but it is suppressed by effect of the input diversity. Another possible explanation for the discrepancy between the IA children in this study and prior findings is that there is a ceiling effect for the effect of environment. Specifically, there is evidence suggesting that environmental contributions may be greater in low-SES samples where environment is likely to be the limiting factor, and smaller in high-SES samples where it is less likely to be the limiting factor (Turkheimer, Haley, Waldron, D'Onofrio, & Gottesman, 2003).

However, there is an important difference between the current study and previous ones. Previous studies have an unavoidable confound of environmental (often indicated by SES) and genetic influences on children's language development. Variation in parental language input may contribute to variability in language development, but biological parents provide their children with both linguistic and genetic input. Thus it is possible that correlations between parental input and child outcomes in previous studies reflect direct effects of shared genes on verbal abilities, and not a direct causal link between input and outcome. So what is the role of genetics in language development?

The Role of Genetics in Language Development

As part of the Twins' Early Development Study (TEDS) thousands of twins were studied to investigate the roles of environmental and genetic factors in children's language development (Oliver & Plomin, 2007; Plomin & Dale, 2000). One motivation for TEDS was a consistent set of findings from adoption and twin studies suggesting a significant effect of genetics on language ability. Although the early findings suggest that nonverbal and verbal abilities have a similar genetic correlation and are moderately correlated with each other (Plomin & Dale, 2000), later studies suggest a stronger environmental influence (Spinath, Ronald, Harlaar, Price, & Plomin, 2003). The myriad of studies published on TEDS data also suggest that the relative potency of genetic and environmental influences changes over time (Oliver & Plomin, 2007).

The genetic confound present in many studies of the effect of input on children's language development was removed in another recent study, which explored effects of teacher input on syntactic development (Huttenlocher, Vasilyeva, Cymerman, & Levine, 2002). The study measured children's syntactic growth over a school year and found it was predicted by qualitative aspects of their preschool teacher's syntactic input, suggesting a direct effect of input on acquisition. According to the authors, this pattern of findings suggests that observed correlations between language input and output in prior research may reflect a causal role of input in language acquisition. While it is true that these results cannot be explained by genetic factors, the focus was only on syntax. Also, when thinking about the effects of input over time it is likely that any effect of input would be compounded and thus we should expect significant predictive links with overall ability as well.

Conclusions

Like populations of children learning their first language from birth, maternal input (nurture) significantly correlated with English vocabulary development in IA children. These relations were strong despite the fact that IA children share no genetic influence (nature) with their adoptive parents. This reconciles the gene-environment confound present in previous studies and provides additional support for the role

of maternal input in children's vocabulary development. In addition, the inclusion of two different age groups provided insight into the contexts in which effects of language input are likely to be largest.

The development of language depends on many things including input, general cognitive skills (e.g., memory), etc. When cognitive skills are well developed and language acquisition is rapid, then the pace of language development is most likely to depend on the variation in input. Thus we see maternal input effects in the preschool-aged IA children even though the amount of (and variation in) input for all children was quite high. In contrast, when the pace of language acquisition is slower because cognitive skills are still developing, then language input may be less likely to be a limiting factor—particularly for children who are in input-rich environments (i.e., IA children).

Acknowledgments

We would like to thank Abbie Claflin, Nicole Gavel, Ellen Godena, Candice Ishikawa, Corinne Jones, Eva Liggett, Angela Lou, John Ste Marie, Ryan Sykora, Cathy Tillman, and K. Yvonne Woodworth for assistance with data collection and transcription; Katie Felkins for her endless assistance over the years; and the families who participated for their altruism. This project was generously supported by a grant from the NSF (No. 0418423) to the third author.

References

Caldwell, B., Heider, J., & Kaplan, B. (1966, September). *Home observation for measurement of the environment.* Paper presented at the meeting of the American Psychological Association.

Elardo, R., Bradley, R., & Caldwell, B. (1977). A longitudinal study of the relation of infants' home environments to language development at age three. *Child Development, 4*, 595-603.

Fenson, L., Marchman, V., Thal, D., Dale, P. S., Bates, E., & Reznick, J. S. (2006). *MacArthur-Bates communicative development inventories (CDIs)* (2nd ed.). Baltimore, MD: Brookes Publishing.

Hackman, D. A., & Farah, M. J. (2009). Socioeconomic status and the developing brain. *Trends in Cognitive Sciences, 13*(2), 65-73.

Hart, B., & Risley, T. R. (1992). American parenting of language-learning children: Persisting differences in family--child interactions observed in natural home environments. *Developmental Psychology, 28*, 1096-1105.

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experiences of American children*. Baltimore, MD: Brookes Publishing.

Hoff-Ginsberg, E. (1986). Function and structure in maternal speech: Their relation to the child's development of syntax. *Developmental Psychology, 22*, 155-163.

Hoff-Ginsberg, E. (1998). The relation of birth order and socioeconomic status to children's language experience and language development. *Applied Psycholinguistics, 19*, 603-629.

Hoff, E. (2003a). Causes and consequences of SES-related differences in parent-to-child speech. In M. H. Bornstein & R. H. Bradley (Eds.), *Socioeconomic status, parenting, and child development* (pp. 147-160). Mahwah, NJ: Erlbaum.

Hoff, E. (2003b). The specificity of environmental influence: Socioeconomic status affects early vocabulary development via maternal speech. *Child Development, 74*, 1368-1378.

Hurtado, N., Marchman, V. A., & Fernald, A. (2008). Does input influence uptake? Links between maternal talk, processing speech and vocabulary size in Spanish-learning children. *Developmental Science, 11*, F31-F39.

Huttenlocher, J., Haight, W., Bryk, A., Seltzer, M., & Lyons, T. (1991). Early vocabulary growth: Relation to language input and gender. *Developmental Psychology, 27*, 236-248.

Huttenlocher, J., Vasilyeva, M., Cymerman, E., & Levine, S. (2002). Language input and child syntax. *Cognitive Psychology, 45*, 337-374.

Huttenlocher, J., Vasilyeva, M., Waterfall, H. R., Vevea, J. L., & Hedges, L. V. (2007). The varieties of speech to young children. *Developmental Psychology, 43*, 1062-1083.

Kirk, S., McCarthy, J., & Kirk, W. (1968). *Examiner's manual: Illinois test of psycholinguistic abilities* (Rev. ed.). Urbana, IL: University of Illinois Press.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk*. Mahwah, NJ: Erlbaum.

Oliver, B. R., & Plomin, R. (2007). Twins' Early Development Study (TEDS): A multivariate, longitudinal genetic investigation of language, cognition and behavior problems from childhood through adolescence. *Twin Research and Human Genetics, 10*, 96-105.

Plomin, R., & Dale, P. S. (2000). Genetics and early language development: A UK study of twins. In D. V. M. Bishop & B. E. Leonard (Eds.), *Speech and language impairments in children: Causes, characteristics, intervention and outcome.* (pp. 35-51). Hove, UK: Psychology Press.

Snedeker, J., Geren, J., & Shafto, C. L. (2007). Starting over: International adoption as a natural experiment in language development. *Psychological Science, 18*, 79-87.

Snedeker, J., Geren, J., & Shafto, C. L. (in press). Disentangling the effects of cognitive development and linguistic expertise: A longitudinal study of the acquisition of English in internationally-adopted children. *Cognitive Psychology*.

Spinath, F. M., Ronald, A., Harlaar, N., Price, T. S., & Plomin, R. (2003). Phenotypic *g* early in life: On the etiology of general cognitive ability in a large population sample of twin children aged 2 - 4 years. *Intelligence, 31*, 195-210.

Stromswold, K. (2001). The heritability of language: A review and meta-analysis of twin, adoption and linkage studies. *Language, 77*, 647-723.

Turkheimer, E., Haley, A., Waldron, M., D'Onofrio, B., & Gottesman, I. I. (2003). Socioeconomic status modifies heritability of IQ in young children. *Psychological Science, 14*, 623-628.

Weizman, Z. O., & Snow, C. E. (2001). Lexical input as related to children's vocabulary acquisition: Effects of sophisticated exposure and support for meaning. *Developmental Psychology, 37*, 265-279.

Whitehurst, G. J. (1997). Language processes in context: Language learning in children reared in poverty. In L. B. Adamson & M. A. Romski (Eds.), *Research on communication and language disorders: Contribution to theories of language development* (pp. 233-266): Brookes Publishing.

Zimmerman, F. J., Gilkerson, J., Richards, J. A., Christakis, D. A., Xu, D., Gray, S., et al. (2009). Teaching by listening: The importance of adult-child conversations to language development. *Pediatrics, 124*, 342-349.

Who's afraid of similarity? Effects of phonological and semantic similarity on lexical acquisition.

Sarah Devi Sahni (sdsahni@gmail.com)
University of Wisconsin – Madison, Department of Psychology

Abstract

Children are sensitive to statistical regularities in speech and likely use these regularities when learning their native language. A central goal of current research is to understand which statistical regularities support different aspects of language acquisition and processing. In the current work we explore phonological and semantic similarity effects on early lexical acquisition. Using a computational model, behavioral findings from word learning studies are simulated and explored. With this model we demonstrate that acquisition can be facilitated by the distinctiveness of individual lexical mappings.

Introduction

Language acquisition research has robustly shown that children are sensitive to statistical regularities in speech, and utilize these regularities when learning their native language (for a review see Saffran & Sahni, in press). A central goal of current research in language acquisition is to understand which statistical regularities support different aspects of language acquisition and processing. Research on adult language processing has revealed that statistical regularities across words can affect lexical access and recognition (Dahan & Magnuson, 2006). Much of this work has examined effects of phonological similarity. Nevertheless, researchers have also examined the effects of semantic similarity along with phonological similarity (e.g. Mirman & Magnuson, 2008).

Phonological and semantic effects in lexical *acquisition* have also been examined. However, little of this work has simultaneously examined phonological and semantic effects in the same set of stimuli or set of studies. In the current work we used a computational model of word learning to investigate the influence of phonological similarity and semantic similarity on early word learning.

Phonological Similarity

Numerous researchers have shown that phonological similarity influences lexical recognition, recall, and access in adults (Dahan & Magnuson, 2006; Luce & Pisoni, 1998; Vitevitch & Luce, 1998; Vitevitch, Luce, Pisoni, & Auer, 1999). Luce's work demonstrates how lexical items that differ by a single phoneme (phonological neighbors) can be simultaneously activated and compete with spoken input (Luce & Pisoni, 1998). While this adult work suggests that phonological similarity impedes lexical processing, developmental work on phonological neighbors suggests that phonological similarity may aid typical lexical acquisition. Storkel (2004) examined whether phonological neighborhood density (together with word frequency and word length) could predict the age of acquisition of early vocabulary items from the Macarthur-Bates Communicative Development Inventory (MCDI) lexical production norms (Dale & Fenson, 1996). She found that words with more phonological neighbors were acquired earlier than words with fewer phonological neighbors, even after accounting for effects of frequency and length. These results suggest that sound similarity (high phonological density) facilitates lexical acquisition.

In contrast with Storkel's work (2004), many nonce word learning studies suggest that infants struggle to learn words that are phonologically similar to one another or to words they already know. Using a habituation task, Stager and Werker (1997) found that 14-month-old infants were able to associate two novel labels with novel objects, but only when the labels were phonologically distinct, like *lif* and *neem*. Infants were unable to map phonologically similar labels *bih* and *dih* to separate objects. This result was quite surprising because using a similar task infants could discriminate the phonemic */b/-/d/* contrast at 8 months (Stager & Werker, 1997). Yet, it was not till 20-months that infants showed clear evidence of learning labels that differed on this contrast (Werker, Fennell, Corcoran, & Stager, 2002).

What can account for these disparate research findings? One important aspect of the child's environment that was not examined in this work is the referent or concept that labels map to. As similarity between labels affects lexical acquisition, it is likely that similarity between referents also affects acquisition. While there has been a significant amount of work investigating how young children will extend category labels based on referent properties', little work jointly examines the role of the label and the role of the referent in lexical acquisition.

Semantic Similarity

Some of the most interesting and revealing work on semantic development investigates label extension and categorization (Quinn & Johnson, 1997; Rakison & Oakes, 2003; Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976). Much of this work has emphasized how the structure of the environment enables infants to group objects and apply category labels to those groups. Their world is well structured; meaningful correlations occur and reoccur, while arbitrary correlations are rarely repeated. This experience allows children to tune into the meaningful and useful correlations in their world.

Research on the shape bias in categorization elegantly demonstrates how the structure of the environment can facilitate language learning. Many of the first words infants learn refer to categories of objects organized by shape.

Experience with these words seems to facilitate categorization abilities. Infants who know 150 words or more, can readily generalize names for newly learned objects to other objects with similar shapes while infants with less than 150 words cannot (Samuelson & Smith, 1999). Samuelson & Smith hypothesized that as children learn more words they extract organizing regularities and form generalizations. These generalizations may initially be restricted to a specific category (e.g. all spherical objects are balls). Then with increased exposure to labeled categories infants form second-order generalizations (e.g. things that are the same shape share a label). These generalizations allow infants to learn the category structure of the objects in their world. Crucially, it is only through sufficient exposure that infants' acquire higher-order generalizations and learn that objects that are the same shape are likely to share a label. If children acquire this bias due to the statistical regularities of the words they know, they must have significant experience with words organized by shape.

Storkel and Adolf (2009) assessed the effect of semantic set size on preschoolers' ability to learn new words. Semantic set size was defined as the number of objects that are meaningfully related to the target word. Subjects showed no difference in initial acquisition of items with large and small set sizes. However, one week after the initial test subjects showed better memory for objects with smaller set sizes. These results suggest that children can learn words more easily when they have a smaller semantic set size and the objects are more unique.

Rogers and McClelland's (2004) categorization model similarly predicts that it will be difficult to learn unique names for items that share many features with other items. Rogers and McClelland hypothesized that infants are sensitive to correlations among different types of directly observable features. These features, which co-occur in the exemplars of a single category, cannot individually define a category. Nor can a specific set of necessary and sufficient features define any category, there are always exceptions. However, the features that consistently co-occur, though not necessarily in every instance of a category, can define a category. For example, birds tend to fly, and have feathers, wings, and beaks. While these features do not always co-occur (penguins have wings but cannot fly) they frequently do and are said to coherently covary with one another. As infants interact in and explore their world they are naturally exposed to these correlations and regularities. Infants are sensitive to the coherent covariation and can use these constellations of features to identify new members of a category. Based on this work, two objects that share many properties will easily map to the same label. While this is beneficial when forming categories, it may be an impediment to children learning the names of similar objects, like "cup" and "glass".

In the current work we use a computational model of word learning to explore effects of phonological and semantic similarity on word learning. Research on phonological similarity is unresolved and suggests similarity facilitates lexical acquisition in some situations but hinders acquisition in others. We propose that by using a computational model to explore effects of phonological and semantic similarity in a single task, we will be able to better understand this phenomenon.

Methods

The main goal of the model was to simulate behavioral experiments that tested infants' abilities to learn similar sounding labels (Werker & Fennell, 2004). In these studies, infants viewed novel objects on a video screen that were audibly labeled with a nonce word. Infants were repeatedly shown these stimuli until their interest had decreased and they were habituated. After habituation, infants received "same" and "switch" test trials. The same trials were the same as habituation trials. In switch trials the objects paired with each label were switched. That is, in switch trials *dih* was paired with the *bih* object, and *bih* was paired with the *dih* object. Longer looking times to switch trials were interpreted as dishabituation and evidence that children learned the mappings.

Architecture

The architecture of the model is presented in Figure 1. The model was composed of three layers: semantic, hidden and phonological. The phonological layer was the input layer and had 192 units (16 units coding phonetic features for each of 12 possible phonemes), the hidden layer had 200 units, and the semantic layer was the output layer and had 135 units. The semantic and phonological layers had recursive units as well as lateral connections between units within the layers. The semantic layer was the output layer over which targets were set and error was calculated.

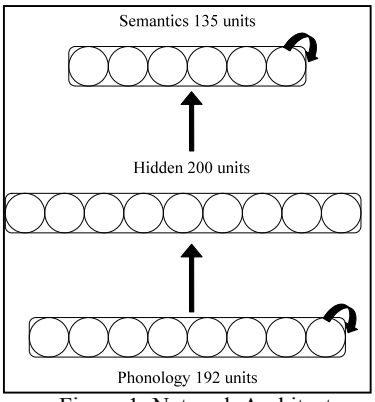


Figure 1: Network Architecture

Training

Three networks initialized with different small random weights, were trained on 332 nouns from the MCDI production checklist (Fenson, et al., 1994). The networks' task was to learn the mapping from phonological labels to semantic referents. Networks were presented with a label on the phonological layer and were to activate the correct set of semantic features describing the referent on the semantic

layer. For example, networks that had learned the word dog would activate the 44 semantic feature units that describe a dog (i.e., eats, has tail, is fun, is lovable, etc.) when presented with the phonological representation of *dog* across the phonological input layer.

Networks were trained using standard backpropagation (Rumelhart, Hinton, & Williams, 1986), with cross-entropy error calculated across output units. The learning rate was set to .005 with no momentum. Networks were trained in batches of 20 words. Output activations and weight matrices were saved every 500 training trials to evaluate the course of learning. Training for each word continued until the activation of each semantic output unit was within 0.2 of its target value or training was manually halted for testing.

Testing

To simulate the behavioral experiments, an analog of habituation and the same-switch procedure was used to test the networks. The networks were trained to differing levels of vocabulary size to simulate the different ages at which infants succeed and fail at the task. At these different stages of training the habituation and same-switch test procedures were simulated in the models.

In the behavioral work by Werker and colleagues (Werker & Fennell, 2004), infants were initially habituated to the stimuli. That is, they were repeatedly exposed to label-object pairs until their looking time decreased by 50%. They were next shown "same" and "switch" test trials, in which the label-object pairing from habituation was either preserved or switched. An increased looking time to the switch trials indicated dishabituation and acquisition of the label-object pairings.

As with infant participants the networks were habituated to the stimuli. Error across the output layer served as the model analog to looking time (Schafer & Mareschal, 2001). To establish the baseline error rate for the habituation phase, models were presented with correct label-object pairings for either *bih-dih* or *lif-neem*. After the first presentation of the novel words, activation on the semantic layer was recorded and compared to the semantic representation of the appropriate referent. This error value provided the baseline error rate for the habituation phase. Models were trained on the pair of novel words until error on the output layer reduced by 50% of baseline. Models were next tested with same and switch trials. On both same and switch test trials error across the semantic output layer was recorded. This error represented the mismatch between a model's expectations and the semantic target of the nonce label. As with infant looking times, larger error indicates surprise and dishabituation from training (Schafer & Mareschal, 2001).

Phonological Representations

Phonological representations of the MCDI nouns and nonce words were based on representations from Joanisse and Seidenberg (1999). See the appendix for a list of features used to represent the phonemes of each word. These representations were slot-based and centered on the first vowel such that when words were compared, phonemes in the same slot position were compared with one another. For example, the words /sta:r/ and /ka:r/ were aligned in vowel-centered slots such that the /a:r/s were aligned even though /sta:r/ has two initial consonants while /ka:r/ only has one.

Slot-based representations have known limitations and can cause delays in training (Plaut, McClelland, Seidenberg, & Patterson, 1996). In these representations phonemes across slots are independent from one another, and cannot facilitate learning across slots. Therefore though knowing the word *pencil* may facilitate acquisition of *penguin* because of the word-inital overlap*;* knowledge of neither *penguin* nor *pencil* can facilitate learning *playpen*, which has a word-final *pen.* Despite these limitations, vowel-centering has been shown to minimize this problem (Harm & Seidenberg, 1999).

Semantic Representations

Semantic representations of the MCDI nouns were taken from Howell, Jankowicz & Becker (2005). Howell et al. used a set of 97 perceptually grounded features to code each word in the MCDI (see the appendix for a list of all features). These features were a subset of the McRae, de Sa, and Seidenberg (1997) empirically derived feature set. Howell et al. chose to use only features that were directly observable by children 8 to 28 months old. They then gathered ratings on these 97 features from human raters for each concept on the MCDI. The final vector for each concept was created by averaging raters' scores.

Howell et al.'s patterns were composed of graded values that varied between 0 and 1, but the majority of features in the set were binary in nature (e.g., "is solid", "is young" etc.). Therefore, all of the conceptually binary features were re-coded as 1's and 0's, with values above .5 becoming 1 and the remaining becoming 0. There were an additional 19 features that coded continuous dimensions (e.g., size, speed, colorfulness, etc.). These features were split into three units representing low, medium and high values of the feature. If a concept had a 0 on one of these continuous dimensions, the high, medium, and low units for that feature were all set to 0. This transformation resulted in semantic patterns using 135 units.

In addition to referents of words from the MCDI, representations for novel referents were created. To create these semantic representations an adult coder, blind to the hypotheses of the studies, looked at pictures and read descriptions of stimuli from published papers. Based on these pictures and/or descriptions each semantic feature was coded as 1 or 0, present or not present, for each novel object.

Results

Word learning experiments conducted by Werker and colleagues (2004) tested children between 14 and 20 months of age. To simulate results over this age range, we used the MCDI norms (Fenson, et al., 1994) to calculate the average number of words children at 14 months can understand. The

norms indicate that the majority of 14-month-olds know at least 64 words. The models reached this level of comprehension at 2500 weight updates. The MCDI comprehension norms do not have data on children older than 16 months, therefore a point later in training that corresponded to a larger vocabulary, 6500 weight updates and 306 known words, was used to simulate the 20-month data point.

We began by simulating the 14-month old studies. Weight matrices produced after 2500 training updates with the full MCDI vocabulary were loaded onto the models. Representations of the two nonce objects were paired with one label from each pair. As with the behavioral studies, the same nonce objects were used for *bih-dih* and *lif-neem*. Networks were habituated and tested with the same-switch procedure as described in the methods sections. All three models showed a larger switch preference when learning *lif* and *neem*, compared to *bih* and *dih* (see Figure 2). This was consistent with 14-month behavioral data (Werker & Fennell, 2004).This indicates that similar to children, the models found the switch trials to be a greater mismatch from what was expected when learning *lif* and *neem,* than when learning *bih* and *dih*.

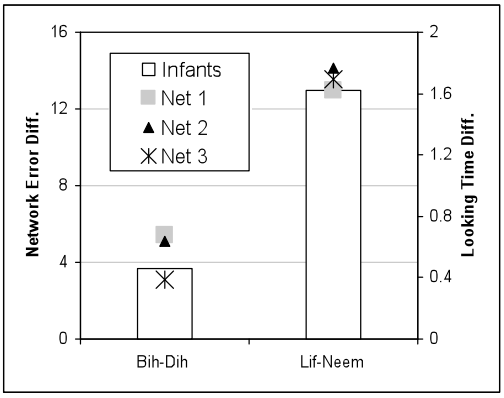


Figure 2: Switch preference for the three networks and infants from Stager and Werker (1997). Difference in error for the networks is labeled on the left y-axis and difference in looking time in seconds for behavioral data is labeled on the right y-axis.

A repeated measures 2 (trial type: same, switch) x 2 (nonce pair: *bih-dih*, *lif-neem*) ANOVA was run on output error from test trials. The main effect of trial type [$F(1,4)=529.571$, $p<.001$] was significant, showing increased error on switch trials for both pairs. There was also a significant interaction between trial type and nonce pair [$F(1,4)=132.42$, $p<.001$]. This result revealed that the switch preference for *lif-neem* was significantly greater than that for *bih-dih.* This replicates the crucial finding that dishabituation is significantly greater for labels that are distinct. The interaction between nonce pairs and test item type is a crucial replication of the Stager & Werker (1997) data.

This computational model of word learning maps phonological representations of labels to semantic feature representations of referents through a 200 unit hidden layer. Weights coming in and out of the hidden layer are adjusted via the backpropagation algorithm. As the model is trained the hidden layer magnifies differences from the input that map to the correct set of semantic features. The activation across the hidden layer can be thought of as an internal representation of the input that maps to the correct features in the output. If two phonological labels produce similar patterns across the hidden layer, the model will more readily map these to similar referents.

To better understand the models' behavior, hidden layer activations of the nonce words were examined prior to habituation. These activations represent the model's ability to discriminate the nonce labels based on current vocabulary size and composition, but prior to training on the nonce items. Weight matrices produced after 2500 and 6500 training trials on the nouns from the MCDI were loaded onto the networks. The networks were then tested on the *bih-dih* and *lif-neem* mappings. Activations produced on the hidden layer were recorded and the distance between patterns for labels in each pair was calculated. That is, for each model we compared activation patterns produced across the hidden layer for the label *bih* with the activation pattern produced by *dih*. Similarly, the hidden layer activation pattern produced by *lif* was compared to the pattern produced by *neem*. Euclidean distance between the two patterns was calculated to assess the model's ability to represent the input as two separate items (see Table 1).

Weight Update	Label	Distance between labels		
		Net 1	Net2	Net 3
2500	*bih-dih*	0.93	0.78	0.78
2500	*lif–neem*	2.07	2.025	2.13
6500	*bih-dih*	1.80	1.58	1.74

Table 1: Euclidean distance between hidden representations of yoked label pairs.

After 2500 weight updates, the distance between hidden layer representations of *lif* and *neem* was greater than the difference between *bih* and *dih*. This greater difference shows that the model is better able to represent *lif* and *neem* as distinct labels. Hidden layer representations were also compared at 6500 weight updates when the model successfully maps *bih* and *dih* to distinct referents. With a larger and more diverse vocabulary, the difference between hidden layer representations of *bih* and *dih* is much greater, indicating that the more experienced model is better able to represent them as separate labels. However, the difference is still not as large as between *lif* and *neem* after 2500 updates, indicating that learning *bih* and *dih* when more experienced is possibly still harder than learning *lif* and *neem* at younger ages. This analysis indicates that with more experience the model is better able to represent the important differences between *bih* and *dih*.

Mapping to Distinctive Referents

A major goal of the current work was to examine the role of semantic similarity on lexical acquisition. In addition to

phonological similarity affecting acquisition it is likely the similarity of referents also affects word learning. To test this hypothesis, we created two new semantic patterns that were completely unique. Both patterns had 36 active semantic units, none of which overlapped. The units were chosen pseudo-randomly, and so patterns do not represent any real-world object. Using the same/switch method, we tested lexical acquisition of *bih* and *dih* and *lif* and *neem* paired with the distinct objects after 2500 updates. If semantic distinctiveness does not affect lexical acquisition, the interaction between test item type and label pair (*bih-dih* vs. *lif-neem*) should persist. Alternatively, if semantic distinctiveness can help to differentiate the label-object pairs, there should be no difference in the acquisition of *bih-dih* and *lif-neem*.

As seen in Figure 3, changing only the distinctiveness of the referents allows the model to learn *bih* and *dih* just as well as *lif* and *neem*. By making the referents of the two labels more distinct, similar-sounding labels are acquired as easily as distinctive sounding labels. A repeated measures 2 (test item type) x 2 (label pair) ANOVA was conducted to examine whether the acquisition of *bih-dih* differed from the acquisition of *lif-neem,* when they were mapped to distinct referents. While the significant main effect of test item type [$F(1,4)=482.437$, $p<.001$] persists, the interaction between test item type and label pair is no longer significant [$F(1,4)=.158$, $p=.711$]. Additionally, as seen in Table 2, hidden layer representations are further differentiated after training with distinct objects. This is true for both *bih-dih* and *lif-neem*.

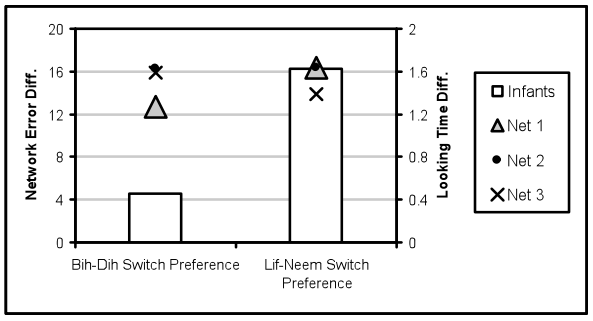


Figure 3: Switch preference for three networks mapping to distinct objects and infants from Stager and Werker (1997). Difference in error for the networks is labeled on the left y-axis and difference in looking time in seconds for behavioral data is labeled on the right y-axis.

Conclusions

The natural world provides infants with strong correlations between linguistic structure and object properties. This structure supports the young child's difficult task of mapping labels to concepts and referents in their world.

In the present work we examined how structure among word forms and words referents can influence word learning. Word learning studies by Werker and colleagues (2004) suggested that high phonological density inhibits acquisition, while Storkel (2004) suggests that in some contexts, phonological density should facilitate acquisition. Using a computational model of word learning, we explored the role that semantic referents of novel words may play in these findings.

Label	Training	Net 1	Net 2	Net 3
bih-dih	Prior to habituation	.832	.783	.78
lif–neem	Prior to habituation	2.01	2.02	2.11
bih-dih	Post habituation	2.43	2.57	2.53
lif–neem	Post habituation	3.8	3.93	3.36

Table 2: Euclidean distance between hidden representations of yoked label pairs when mapping to **distinct** referents.

The computational model examined effects of semantic and phonological similarity on the *process* of word learning. Using model analogs to habituation, we simulated the basic finding that it is difficult to learn similar sounding labels like *bih* and *dih*. By examining the hidden layer representations of these items we found that the surface similarity of the labels affected the model's ability to treat them as separate items. However, models were able to successfully map *bih* and *dih* to separate objects when the objects were completely distinct. Training with these distinct objects allowed the models to pull apart representations of words that had similar labels, as shown in Table 2.

Importantly, this simulation showed that the referents of labels, and their relationship to other items in the input, can affect word learning. This finding brings to light the need to consider the effects of semantic structure when studying word learning.

Acknowledgments

This work has been generously supported, in part, by the 2008 APA dissertation award, and pre-doctoral grant F31DC008737 from the National Institutes of Health. The author wishes to thank four anonymous Cognitive Science reviewers for helpful comments and suggestions.

Appendix: Sound & Semantic Feature Sets

Sound features: voiced, consonantal, vocalic, sonorant, lateral, continuant, noncontinuant, advanced tongue root, nasal, labial, coronal, anterior, high, distributed, dorsal, radical.

Semantic features: size, weight, strength, speed, temperature, cleanliness, tidiness, brightness, noise, intelligence, goodness, beauty, width, hardness, roughness, height, length, scariness, colorfulness, is black, is blue, is brown, is gold, is green, is grey, is orange, is pink, is purple, is red, is silver, is white, is yellow, is conical, is crooked, is curved, is cylindrical, is flat, is liquid, is rectangular, is round, is solid, is square, is straight, is triangular, has feather, has scales, has fur, is prickly, is sharp, is breakable, made of china, made of cloth, made of leather, made of metal, made of plastic, made of stone, made of wood, climbs, crawls, flies, leaps, runs, swims, breathes, drinks, eats, makes animal noise, singles, talks, has four legs, has beak, has door, has shell, has eyes, has face, has fins, has handle, has leaves, has legs, has paws, has tail, has teeth, has wheels, has whiskers, has wings, is

annoying, is comfortable, is fun, is musical, is scary, is strong smelling, is young, is old, is comforting, is lovable, is edible, is delicious.

References

Dahan, D., & Magnuson, J. S. (2006). Spoken-word recognition. In M. J. Traxler & M. A. Gernsbacker (Eds.), *Handbook of Psycholinguistics* (pp. 249-283). Amsterdam: Academic Press.

Dale, P. S., & Fenson, L. (1996). Lexical development norms for young children. *Behavior Research Methods, Instruments & Computers*(28), 125-127.

Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., & Pethick, S. J. (1994). Variability in early communicative development. *Monographs of the Society for Research in Child Development, 59*(5), v-173.

Harm, M. W., & Seidenberg, M. S. (1999). Phonology, reading acquisition, and dyslexia: Insights from connectionist models. *Psychological Review, 106*(3), 491-528.

Howell, S. R., Jankowicz, D., & Becker, S. (2005). A model of grounded language acquisition: Sensorimotor features improve lexical and grammatical learning. *Journal of Memory and Language, 53*(2), 258-276.

Joanisse, M. F., & Seidenberg, M. S. (1999). Impairments in verb morphology following brain injury: A connectionist model. *Proceedings of the National Academy of Sciences of the United States of America*, 7592-7597.

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing, 19*, 1-36.

McRae, K., de Sa, V. R., & Seidenberg, M. S. (1997). On the nature and scope of featural representations of word meaning. *Journal of Experimental Psychology: General, 126*(2), 99-130.

Mirman, D., & Magnuson, J. S. (2008). Attractor dynamics and semantic neighborhood density: Processing is slowed by near neighbors and speeded by distant neighbors. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*(1), 65-79.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review, 103*(1), 56-115.

Quinn, P. C., & Johnson, M. H. (1997). The emergence of perceptual category representations in young infants: A connectionist analysis. *Journal of Experimental Child Psychology, 66*(2), 236-263.

Rakison, D. H., & Oakes, L. M. (Eds.). (2003). *Early category and concept development: Making sense of the blooming, buzzing confusion*: Oxford University Press.

Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach*: MIT Press.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology, 8*(3), 382-439.

Rumelhart, D. E., Hinton, G. E., & Williams, R. (1986). Learning internal representations by error propagation. In D. E. Rumelhart, J. L. McClelland & P. R. Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition.* (Vol. 1). Cambridge, MA: MIT Press.

Saffran, J. R., & Sahni, S. D. (in press). Learning the sounds of language. In M. Joanisse, M. Spivey & K. McCrae (Eds.), *Cambridge Handbook of Psycholinguistics*. Cambridge: Cambridge University Press.

Samuelson, L. K., & Smith, L. B. (1999). Early noun vocabularies: Do ontology, category structure and syntax correspond? *Cognition, 73*(1), 1-33.

Schafer, G., & Mareschal, D. (2001). Modeling infant speech sound discrimination using simple associative networks. *Infancy, 2*(1), 7-28.

Stager, C. L., & Werker, J. F. (1997). Infants listen for more phonetic detail in speech perception than in word-learning tasks. *Nature, 388*(6640), 381-382.

Storkel, H. L. (2004). Do children acquire dense neighborhoods? An investigation of similarity neighborhoods in lexical acquisition. *Applied Psycholinguistics*(25), 201-221.

Storkel, H. L., & Adolf, S. M. (2009). The effect of semantic set size on word learning by preschool children. *Journal of Speech, Language, and Hearing Research, 52*(2), 289-305.

Vitevitch, M. S., & Luce, P. A. (1998). When words compete: Levels of processing perception of spoken words. *Psychological Science, 9*(4), 325-329.

Vitevitch, M. S., Luce, P. A., Pisoni, D. B., & Auer, E. T., Jr. (1999). Phonotactics, neighborhood activation, and lexical access for spoken words. *Brain and Language, 68*(1), 306-311.

Werker, J. F., & Fennell, C. (2004). Listening to sounds versus listening to words: Early steps in word learning. In D. G. Hall & S. R. Waxman (Eds.), *Weaving a lexicon.* (pp. 79-109): MIT Press.

Werker, J. F., Fennell, C. T., Corcoran, K. M., & Stager, C. L. (2002). Infants' ability to learn phonetically similar words: Effects of age and vocabulary size. *Infancy, 3*(1), 1-30.

A SOM Model of First Language Lexical Attrition

Benjamin D. Zinszer (bdz107@psu.edu)
Ping Li (pul8@psu.edu)
Department of Psychology, Pennsylvania State University
University Park, PA 16802 USA

Abstract

First language lexical attrition remains a difficult phenomenon to study empirically, due to its long-term and dynamic effects. Based on observations from existing case studies, we propose a connectionist model to simulate the effects of first language lexical attrition. The model exhibits a plausible time-course for first language lexical comprehension, highlights the independence of productive and receptive attrition trajectories, and predicts an age of onset effect for early cases of L1 lexical attrition.

Keywords: language attrition, lexicon, modeling, self-organizing map, connectionism

Introduction

Many people learn and forget a second or third language during the course of their lifetimes. Less often, a major migration may cause someone to forget all or part of his native language. While a great deal of research has been dedicated to the first and second language acquisition, relatively little is known about language loss (hereafter, attrition) in the individual speaker.

Lexical Attrition

In the last decade, there has been an increasing amount of work devoted to the study of language attrition, specifically in L1 or first language attrition. Apparent age-related effects have been observed in the attrition of L1 phonology (Hytenstam et al, 2009; Pallier et al, 2003). However, long-term lexical attrition has remained largely undocumented, partly due to the lack of rigorous experimental methodologies for the study of lexical attrition.

Nonetheless, one could reasonably expect the long-term course of lexical attrition to differ from that of phonology. Previous research examining the interplay between language learning and cognitive functions has identified differing memory stores for lexical and phonological acquisition (Hernandez & Li, 2007; Ullman, 2001). To the extent that continued performance in the L1 depends on different memory representations, the effects of attrition on phonology and the lexicon may be independent.

The current body of L1 lexical attrition research provides some general observations about the relationship between age of onset (AoO)[1], length of residence (LoR) and the degree of attrition. A case study of letters written by an L1-German immigrant to the United States revealed an ongoing process of lexical attrition even fifty years after AoO (Hutz, 2004). In another case study, an L1-German speaker with a similarly long LoR of 47 years in the United States demonstrated substantial lexical relearning in a natural conversational setting (Stolberg & Münch, 2010). That relearning is possible after such a long time raises the question of whether lexical attrition is truly a case of forgetting, in which L1 knowledge is destroyed in memory, or whether it is the access to L1 knowledge that is primarily affected by attrition.

The most evident problem in current L1 attrition research is the difficulty of reliably measuring change across time. As demonstrated in the case studies, loss of L1 abilities may be a slow and gradual process spanning years or decades. As a result, even longitudinal studies over a few years capture only a snapshot of a highly dynamic language system. The limited span of longitudinal data provided by any single study makes it extremely difficult or statistically impossible to identify the time course of development. While large samples with cross-section age variables (such as age of acquisition in the L2 literature) can mitigate these problems, advanced language users who experience L1 attrition are relatively scarce, making a cross-sectional sample nearly impossible.

One small-scale quantitative study has tested L2 lexical attrition through the relearning paradigm. De Bot, Martens, & Stossel (2004) found a relearning advantage in foreign language study for forgotten words over new words, revealing that the forgotten words, though inaccessible, persisted in memory. While this study found a general adherence to an exponential forgetting curve in which relearning savings are possible below the productive threshold, the findings are difficult to generalize due to the limited size of the vocabulary and the limited scope of the study.

Given the difficulties in systematic control of important learning variables (such as age, language proficiency, and L2 exposure), language attrition research has remained mostly a descriptive enterprise. Computational modeling may serve to turn language attrition research to an experimental science, due to its flexibility in parametric manipulation of the relevant variables and in testing relevant theoretical hypotheses. To date, very little work has been done in the computational modeling of language attrition. The goal of this study is to make a first attempt in providing a detailed computational account of the developmental time course of language attrition in the lexical domain.

[1] "Age of onset" here refers specifically to the beginning of attrition. Due to the difficulty of identifying this event, AoO is typically marked by the change of language environment (e.g., geographic migration), prior L2 exposure notwithstanding.

Computational Models

To our knowledge there has been only one computational model specifically designed to address lexical attrition. Meara's (2004) Boolean model of lexical attrition used a simple connectionist paradigm to simulate the effect of intra-lexical relationships on the time course of attrition. Meara's model exhibited self-organized criticality, that is, the wide-spread and sudden deactivation of lexical nodes at unpredictable intervals. This effect may be interpreted as largely a product of the inter-node dependencies inherent to Boolean models, but more importantly, Meara found that when the mean activation was taken across ten models, the resulting curve showed a gradual decline. This study highlights the troubling possibility that empirical research of lexical attrition in human subjects is hiding potential criticality effects. Increasingly sophisticated computational models may yet fill this gap.

Self-organizing feature maps (SOM) are a promising option in modeling lexical attrition. SOM is a connectionist modeling paradigm which represents data in a network of clustered nodes. Previous research has established the utility of SOM in producing cognitively plausible models of language development (see Li, 2009, for a review; see also Richardson & Thomas, 2008 and Mayor & Plunkett, 2010).

The potential for extending SOM to lexical attrition is suggested by its flexibility in simulating the effects of competing input sets. Age-related dynamic cross-linguistic competition in L2 learning has been demonstrated with other SOM-based models (Li & Farkas, 2002; Zhao & Li, 2007). Furthermore, effects of sensitive period or catastrophic interference have also been shown with the manipulation of learning parameters in SOM (Richardson & Thomas, 2008).

Computational modeling offers the possibility of a unified account of language learning, attrition, and relearning phenomena, integrating empirical research in these fields under more durable hypotheses. The present study aims to produce a SOM model: (1) to replicate the sustained gradual erosion of L1 lexical knowledge in both production and comprehension, (2) to compare the respective rates of attrition for comprehension and production, (3) to produce a plausible time course for long-term L1 lexical attrition, and (4) to reveal age of onset effects in L1 lexical attrition.

Method

In this study, a dual self-organizing feature map (SOM) model is trained in a first language (L1) and at varying ages of onset (AoO) in a second language (L2) while L1 training decreases or stops. Performances of the model in comprehension and production are tracked throughout training.

The Model

The self-organizing feature map (SOM) is a connectionist modeling paradigm wherein each node contains a vector of weights corresponding to each member of the input vector (see Kohonen, 2001 for a detailed explanation of SOM). Node weights falling within a defined neighborhood around the input vector are adjusted towards the input based on their distance from it. Over many epochs of training, this adjustment results in topography-preserving orders, such that similar inputs are represented by nearby clusters of nodes in the map while dissimilar inputs by distinct and distant clusters. The typography-preserving characteristics of SOM are particularly well suited for examining the effects of cross-language lexical competition in a dynamically evolving system as in lexical attrition.

Architecture The model designed for this study employs two such SOMs (see Figure 1). The first SOM was trained on the phonological representations of words. This phonological map self-organizes according to the basic phonemic elements in a word, clustering words of a similar sound together. The phonological map was composed of 1600 nodes on a 40 by 40 rectangular grid. The second SOM was trained using the semantic representation of words. The semantic map clustered words of similar meaning, category, and part of speech. The semantic map was composed of 900 nodes arranged on a 30 by 30 rectangular grid. The semantic map was designed to be smaller than the phonological map because it received half as many unique input representations (see *Stimuli and Training*). The two maps were joined by Hebbian connections (see Hebb, 1949 for model and biological basis). A single Hebbian connection is represented by a weight that multiplies activation between the two nodes it connects. Every node on one map was connected to every node on the opposite map, for a total of 1.44 million (1600 x 900) Hebbian connections.

Functions and Parameters After the presentation of each input stimulus, the maps and Hebbian connections were updated according to a set of learning functions. These functions defined which sets of nodes and weights are adjusted and how much they are adjusted.

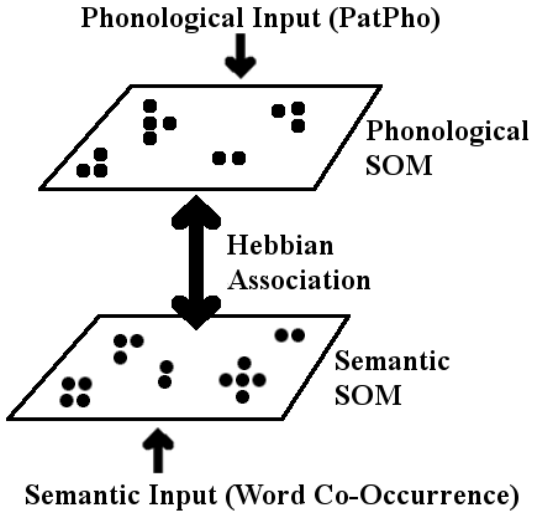


Figure 1: The model is composed of two self-organizing feature maps. Activation in each map is propagated to the other by means of Hebbian Connections.

On the phonological and semantic maps, the node whose weights most closely match those of the input set (measured as minimal Euclidean distance between input and each node) is designated as the Best Matching Unit or BMU. The nodes around the BMU are updated according to a neighborhood function approximating a Gaussian curve with a maximum value of one at the BMU.

The radius of the neighborhood is variable between trials and measured in terms of the Cartesian distances between nodes on the rectangular grid. In this study, the radius was initially set at one half the size of the smaller map (15) to allow maximum adjustment in early trials. With each epoch the radius was allowed to decrease by one if the quantization error was less than in the preceding trial. With this approach, performance of the model was not directly tied to a manipulation of the radius size, but rather the radius size and model performance were allowed to covary through early training stages.

Updates to SOM weights were proportional to a node's value on the Gaussian neighborhood curve, resulting in a smaller change for more distant nodes, and no change for nodes outside the neighborhood. All updates were also multiplied by the SOM's learning rate, a value between zero and one which limits the amount of change that can occur in a single trial. A learning rate of 0.2 was set for both maps.

Hebbian connection adjustment was determined by co-activation in both maps. Activation for each node within the BMU's neighborhood was inversely proportional to Euclidean distance between the node's weights and the input vector. Each Hebbian connection was then adjusted by multiplying the activation of the nodes on each map and the Hebbian learning rate. The Hebbian learning rate was set to 0.1 in this model. Following each trial, Hebbian weights were normalized to values between zero and one.

Stimuli and Training

Two types of stimuli were provided to the model for training. Vectors containing phonological representations of words were presented one at a time to the phonological map. Simultaneously, vectors containing semantic representations of the same words were presented to the semantic map. This paired presentation allowed each map to organize around its respective input and then form connections between the phonological and semantic representations on their respective maps.

Phonological input vectors were generated using the PatPho system for English (Li & MacWhinney, 2002) and Mandarin Chinese (Zhao & Li, 2009). The dimension of each phonological vector was 63 units. Semantic vectors were obtained from the English stimulus set used to train the DevLex-II model. Each semantic input vector was 200 units long, derived from word co-occurrence patterns (see Li, Zhao, & MacWhinney, 2007 for details). In order to help the model discriminate between highly similar words (such as *red* and *blue* or *grandma* and *grandpa*) a nominal amount of noise was randomly added to the semantic data before training began for each model.

Most importantly, the English semantic representations were paired with both Chinese and English phonological representations during training. While emergentist models of bilingualism such as the Unified Competition Model have accounted for semantic and lexical transfer in second language acquisition (MacWhinney, 2005), prior computational models of language acquisition have failed to account for the largely shared conceptual space between two languages. Due to the importance of L2 negative transfer in L1 lexical attrition (Hutz, 2004; Schmitt, 2010) a computational account would be incomplete without a common semantic representation.

Words for the training set were selected from the MacArthur-Bates Communicative Developmental Inventories (English: Dale & Fenson, 1996; Chinese: Hao et al, 2008). Originally, 140 rough translation equivalents were obtained by comparing the English index with the English glosses in the Chinese index. Because intonation was not coded in the phonological representation, several words were eliminated as homophones. A few other words were removed because they could not fit the PatPho template for phonological encoding or did not have readily available co-occurrence data for semantic input. In total 116 English and 116 Chinese words were phonologically and semantically encoded for input to the model.

All instances of the model were trained for 500 epochs. The L1 (Chinese) was trained first, and at varying numbers of epochs (AoO) L1 input ceased and L2 (English) input began. AoO was varied in intervals of 50 epochs from 50 to 400. Ten models were trained for each of the eight AoOs.

Performance Tests

Following each training epoch, production and comprehension of the L1 was tested throughout the entire lifespan of the model.

For modeling purposes, comprehension was defined as the activation of the correct BMU on the semantic map when a phonological stimulus was presented to the phonological map. This activation was achieved by means of the Hebbian connections. After presentation of the stimulus, activation on the phonological map was calculated by the same method described in Functions and Parameters (above). Activation levels in the phonological map were then multiplied through their Hebbian connections. The incoming activation on the semantic map was summed for each node, and the most activated node on the semantic map was found. This most activated node was then compared to a list of semantic BMUs. If the most activated node was also the correct BMU, comprehension had occurred. If no BMU occupied the most activated node, the most activated node was compared to the closest BMU (by Cartesian distance) on the map. In the event that two or more BMUs on either map occupied the same node, all of these BMUs were disqualified from the comparison, preventing their corresponding words from passing the comprehension measure.

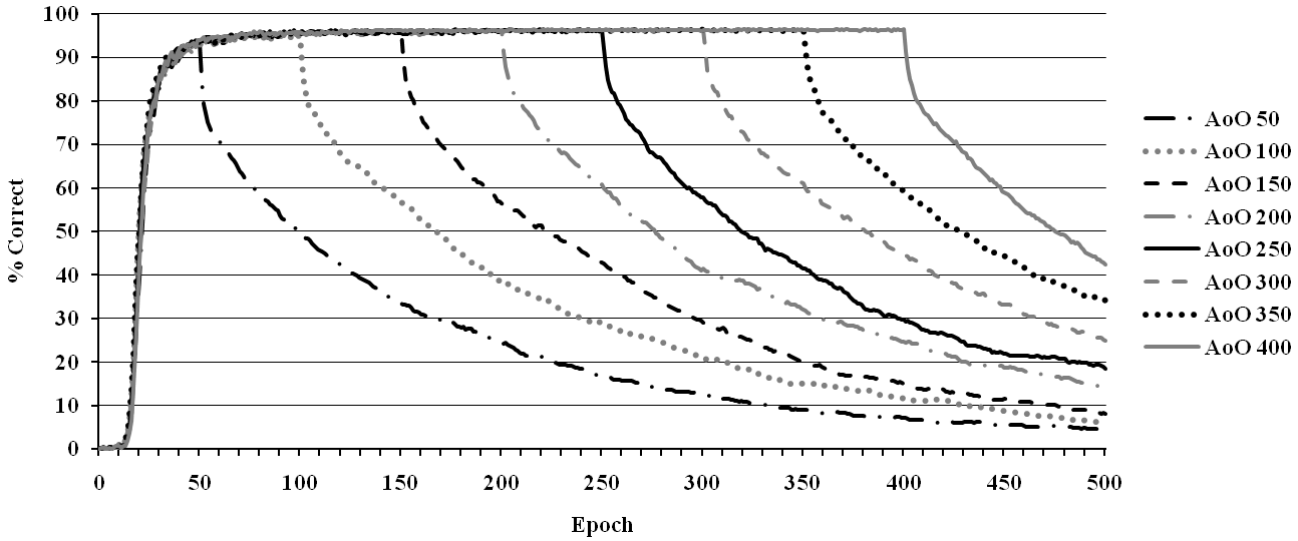


Figure 2: Mean comprehension scores in each AoO condition are graphed across the duration of the models (measured in training epochs). AoO values are also measured in training epochs, as depicted along horizontal axis.

Production was defined by the reverse process of comprehension. A stimulus input was provided to the semantic map, and activation was propagated by the same Hebbian connections to the phonological map, and the most activated unit was compared by the same criteria to the phonological BMUs. One important distinction in the case of production is that the most activated unit was only compared to the L1 BMUs to avoid inter-language confusion.

Results

L1 Comprehension

Mean comprehension curves were calculated across ten models for each AoO condition. Performance for each condition exceeded 92% by 50 epochs (the earliest AoO). AoO conditions later than 50 epochs exceeded 95% comprehension by 100 epochs. Maximum L1 comprehension after 100 epochs was 96.6% (112 out of 116 L1 words) for all models (un-averaged) with an AoO greater than 100. After AoO began, L1 comprehension decreased monotonically. Figure 2 shows the L1 comprehension curves for all eight AoO conditions. At the onset of L2 training, L1 comprehension seems to approximate an exponential decay for each AoO condition. Differences between L1 curves are described below (see section *Age of Onset Effects*).

L1 Production

L1 production declined severely and immediately for all AoO conditions. All models across all conditions performed below 5% correct productions within four epochs of the AoO and remained low throughout L2 training. Due to the low performance, no further analysis was applied to these data. See the *Discussion* section for a further treatment of this topic.

Age of Onset Effects

By visual inspection, AoO was inversely related to rate of attrition for the earlier AoO conditions. To quantify this relationship, the number of epochs required for each AoO condition to drop below 75% comprehension was calculated. Many models in the AoO 50 and 100 conditions did not reach maximum L1 comprehension performance by L2 onset. Therefore the performance calculation compensated by adding to performance measures the difference between each model's maximum L1 comprehension and the overall maximum (112) before calculating the number of epochs necessary to reach the threshold.

Figure 3 approximates the rate of attrition for each AoO by showing the number of epochs elapsed after L2 onset before the 75% L1 comprehension threshold was reached. Error bars indicate the two standard errors of the mean for

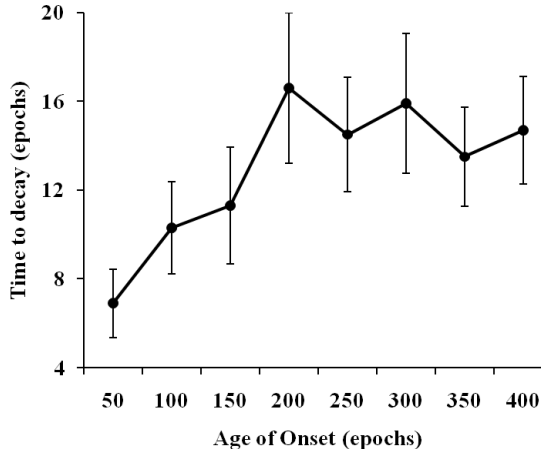


Figure 3: Mean number of epochs to reach 75% L1 comprehension or less, adjusted for incomplete learning. Error bars are two standard errors of the mean.

each AoO condition. ANOVA revealed a highly significant difference ($p<0.001$) in mean decay rates between conditions. Post-hoc tests (Tukey, with a family alpha of 0.05) showed that the AoO 50 condition was significantly different than AoO 200-400 (but not 100 and 150), while AoO 100 was also significantly different from 200.

Discussion

While an examination of learning in connectionist models may be interesting in its own right, the results of this study are most informative with regard to the dynamic trajectories of human first language attrition. Prior studies in L1 attrition have found age effects in phonological attrition, but no such effect has been demonstrated for lexical attrition. Nonetheless, a review of the current L1 lexical attrition literature reveals that lexical attrition is a long-term and dynamic process.

Performance measures for L1 comprehension and production after AoO indicated great instability in the production while comprehension declined more gradually. A potential source of declining performance in both measures was the changing Hebbian weights. Because the weights were normalized with each trial, the magnitude of change to the Hebbian connections due to a stimulus is not strictly dependent on activation levels. This effect is analogous to a decay (or forgetting) rate, as all connection weights were reduced relative to the learning rate.

A major source of instability, and a probable driving factor behind the rapid decline in production, was the reorganization of the phonological SOM. The operational definition of comprehension assumed activation of the correct phonological representations (if present) and tested the consequent activation on the semantic map, rendering comprehension relatively resistant to changes in the phonological map. By contrast, production required that the static semantic representations correctly activate the highly plastic phonological representations. Faced with moving targets, productive performance was at a distinct disadvantage, even when activation was artificially restricted to L1 candidates and criteria were loosened to allow for "close enough" matches.

Although the degree and rate of decline for production may be exaggerated by the model, this finding does reinforce the dissociation of receptive versus productive abilities. Due to this dissociation, studies which primarily measure productive errors in speakers undergoing L1 attrition may overestimate the degree of loss. Stolberg and Münch (2010) found that lexical/semantic production errors decreased by approximately half over the course of 15 conversations in the subject's L1. In light of the dissociation between comprehension and production, the degree to which these errors represent receptive L1 lexical attrition remains in question.

The relearning demonstrated in Stolberg and Münch's study points to the possibility of persistent, though temporarily inaccessible L1 representations. Results from the described model suggest that these representations do persist in memory, reactivated with relatively little practice long after becoming unavailable for production. De Bot et al (2004) confirmed the presence of latent lexical representations in the L2 through a short term relearning task. Foreign language students showed a relearning advantage for words to which they had been previously exposed but forgotten over learning new words. Our model stands to bridge these studies by demonstrating that these latent representations may also explain the observed L1 lexical attrition phenomena, further guiding L1 attrition studies toward seeking L1 representations that may have fallen below the threshold of retrieval for production.

The model also exhibited a highly plausible decay function for first language lexical comprehension. Previously only retrospective analyses, such as that by Hutz (2004), have been available for lexical attrition across a lifetime. Semantic transfer errors identified in Hutz's case study (e.g. *"Das ist feine mit mir*" which is a literal translation of the English idiom "That's fine with me", rather than the equivalent German idiom "*damit bin ich einverstanden*") grew at a diminishing rate over 55 years. The decay of comprehension in this model is highly compatible with Hutz's findings in semantic transfer, indicating that the model's performance curves may represent a component of the generalized time course for L1 lexical attrition.

Moreover, variation of age of onset revealed a possible inverse relationship with the rate at which the comprehension decay occurred. Particularly in the 50 AoO condition, we observed attrition occurring at a higher rate than for later AoOs. This rate, coupled with the slightly lower L1 pre-attrition performance (92% versus 97%), points to the effects of incomplete learning for early onset attrition. Empirical studies have shown that early rather than late exposure to L2 may lead to stronger influence from L2 to L1, causing certain elements of L1 to give way to L2 patterns more easily (e.g., in object naming patterns and categorization; see Pavlenko & Malt, in press). On the other hand, the stronger AoO effects at early stages may be accounted for by the substantial brain plasticity for new languages within the critical period (Pallier et al., 2003).

In the model, it is apparent that the importance of AoO is diminished in cases of later onset. The ostensible leveling-off may be attributable to the limitations placed on Hebbian entrenchment by the normalization. The strength of early AoO effects and high variability in later AoOs reflects Johnson and Newport's (1989) observation of age-related effects in second language acquisition. Like Johnson and Newport's data, our findings are at best ambiguous about the role of age in late second language onset. To what degree the performance of our model was due to incomplete L1 learning versus age-related acceleration of decay requires further investigation.

Conclusion

In empirical literature the study of language attrition has remained a qualitative and descriptive enterprise, due to the

lack of rigorous experimental methodologies for reliably measuring change across time. Coupled with the difficulty of finding a sufficient number of language users who experience L1 language attrition, the extant research makes it difficult to identify any time course of development. In this study, we provided a SOM-based computational model of lexical attrition as a first attempt to systematically investigate mechanisms of language attrition. Specifically, our model is able to produce a gradual decline in L1 lexical performance, suggesting a plausible course of decay in first language comprehension that is compatible with the observations of existing case studies. Furthermore, our model highlights the potential for independent effects on comprehension and production within a single language user. Finally, our model shows age of onset effects in relation to the rate of attrition and points to the possible role of incomplete L1 learning. Such effects are important for understanding the dynamic changes in the competition of two languages during learning.

Acknowledgements

This research was supported by a University Graduate Fellowship for BDZ and a grant from the National Science Foundation (No. 0642586) to PL. We are grateful to Jon-Fan Hu and Xiaowei Zhao for their invaluable discussion and collaboration.

References

Dale, P.S., & Fenson, L. (1996). Lexical development norms for young children. *Behavior Research Methods, Instruments, & Computers, 28*, 125-127.

de Bot, K., Martens, V., & Stoessel, S. (2004). Finding residual lexical knowledge: The Savings approach to testing vocabulary. *International Journal of Bilingualism, 8* (3), 373-382.

Hao, M., Shu, H., Xing, A., & Li, P. (2008). Early vocabulary inventory for Mandarin Chinese. *Behavior Research Methods*, *40* (3), 728-733.

Hebb, D. (1949). *The organization of behavior: A neuropsychological theory*. New York, NY: Wiley.

Hernandez, A. E., & Li, P. (2007). Age of acquisition: Its neural and computational mechanisms. *Psychological Bulletin, 133* (4), 638-650.

Hutz, M. (2004). Is there a natural process of decay? A longitudinal study of language attrition. In M. S. Schmid, B. Köpke, M. Keijzer, & L. Weilemar, *First language attrition: Interdisciplinary perspectives on methodological issue*. Amsterdam: John Benjamins Publishing Company.

Hyltenstam, K., Bylund, E., Abrahamsson, N., & Park, H. (2009). Dominant-language replacement: The case of international adoptees. *Bilingualism: Language and Cognition, 12*(2), 121-140.

Johnson, J., & Newport, E. (1989). Critical Period Effects in Second Language Learning: The Influence of Maturational State on the Acquisition of English as a Second Language. *Cognitive Psychology*, *21*, 60-99.

Kohonen, T. (2001). *The self-organizing maps* (3rd ed.). Berlin: Springer.

Li, P. (2009). Lexical Organization and Competition in First and Second Languages: Computational and Neural Mechanisms. *Cognitive Science*, *33*, 629-664.

Li, P., & Farkas, I. (2002). A self-organizing connectionist model of bilingual processing. In R. Heredia & J. Altarriba, *Bilingual sentence processing, 17,* 59–85. North-Holland.

Li, P., & MacWhinney, B. (2002). PatPho: A phonological pattern generator for neural networks. *Behavior Research Methods Instruments and Computers*, *34* (3), 408–415.

Li, P., Zhao, X., & MacWhinney, B. (2007). Dynamic Self-Organization and Early Lexical Development in Children. *Cognitive Science*, *31* (4), 581-612.

MacWhinney, B. (2005). A unified model of language acquisition. In J. F. Kroll & A. De Groot, *Handbook of bilingualism: Psycholinguistic approaches*. Oxford University Press.

Mayor, J., & Plunkett, K. (2010). A neuro-computational account of taxonomic responding and fast mapping in early word learning. *Psychological Review*. (in press).

Meara, P. (2004). Modelling vocabulary loss. *Applied linguistics*, *25* (2), 137–155.

Pallier, C., Dehaene, S., Poline, J.-B., LeBihan, D., Argenti, A.-M., Dupoux, E., & Mehler, J. (2003). Brain imaging of language plasticity in adopted adults: can a second language replace the first? *Cerebral cortex*, *13* (2), 155-61.

Pavlenko, A. & Malt, B. (2010). Kitchen Russian: Cross-linguistic differences and first language object naming by Russian-English bilinguals. *Bilingualism: Language and Cognition*, (in press).

Richardson, F. M., & Thomas, M. S. (2008). Critical periods and catastrophic interference effects in the development of self-organizing feature maps. *Developmental science*, *11* (3), 371-89.

Schmitt, E. (2010). When boundaries are crossed: Evaluating language attrition data from two perspectives. *Bilingualism: Language and Cognition*, 13 (1), 63-72.

Stolberg, D., & Münch, A. (2010). "Die Muttersprache vergisst man nicht" –or do you? A case study in L1 attrition and its (partial) reversal,. *Bilingualism: Language and Cognition*, *13* (1), 19-31.

Ullman, M. T. (2001). The neural basis of lexicon and grammar in first and second language: the declarative/procedural model. *Bilingualism: Language and Cognition*, *4* (2).

Zhao, X., & Li, P. (2007). Bilingual lexical representation in a self-organizing neural network. *Proceedings of the Twenty-Fifth Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum.

Zhao, X., & Li, P. (2009). An online database of phonological representations for Mandarin Chinese. *Behavior Research Methods*, *41* (2), 575-83.

The Encoding of Spatial Information During Small-Set Enumeration

Harry Haladjian (haladjian@ruccs.rutgers.edu)
Manish Singh (manish@ruccs.rutgers.edu)
Zenon Pylyshyn (zenon@ruccs.rutgers.edu)
Randy Gallistel (galliste@ruccs.rutgers.edu)
Center for Cognitive Science
Rutgers University – New Brunswick
152 Frelinghuysen Road, Piscataway, NJ 08854 USA

Abstract

Using a novel enumeration task, we examined the encoding of spatial information during subitizing. Observers were shown masked presentations of randomly-placed discs on a screen and were required to mark the perceived locations of these discs on a subsequent blank screen. This provided a measure of recall for object locations and an indirect measure of display numerosity. Observers were tested on three stimulus durations (50, 200, 350 ms) and eight numerosities (2-9). Enumeration performance was high for displays containing up to six discs—a higher subitizing range than reported in previous studies. Error in the location data was measured as the distance between corresponding stimulus and response discs. Overall, location errors increased in magnitude with larger numerosities and shorter display durations. When errors were computed as disc distance from display centroid, results suggest a compressed representation by observers. Additionally, enumeration and localization accuracy increased with display regularity.

Keywords: spatial attention; enumeration; subitizing; visual indexing.

I. Introduction

When presented with a set of objects, humans can estimate quickly the set's numerosity with reasonable accuracy. This estimate of number supports various cognitive processes and assists decision-making and action-planning. Given the importance of such abilities, it would be reasonable to expect that a cognitive system employs several methods to obtain numerosity information. The challenge, however, lies in clearly identifying the possible mechanisms involved and determining the conditions under which they are employed.

The primary mechanism responsible for numerosity perception is the nonverbal mental magnitude system that also has been observed in animals and preverbal infants. Magnitudes are inferred mental entities that represent the numerosity or magnitude of things in the world via a mental "accumulator" or "number line" (Dehaene, 1992; Gallistel & Gelman, 1992). An accumulator mechanism is thought to enable the precise representations of duration and numerosity in rats by accumulating neural signals (Meck & Church, 1983). In humans, this accumulator system may represent discrete numerosities through an incrementing process that produces a preverbal count (Gallistel & Gelman, 1992, 2000). Although analog magnitudes are argued to underlie most numerical abilities, an alternate mechanism may be employed for smaller numerosities. The term *subitizing* is used to describe the fast and accurate enumeration of 1-4 objects (Kaufman, Lord, Reese, & Volkmann, 1949). Trick & Pylyshyn (1989, 1994) proposed that a *visual indexing* mechanism may be utilized for subitizing. Visual indexes are "pointers" that automatically pick out and stick to visual items displaying characteristics of "objecthood" (e.g., good continuation, cohesion). Each item that is to be tracked or enumerated is assigned an index in a bottom-up manner, enabling a simultaneous selection of four objects (Pylyshyn, 1989). Subitizing is thought to be the rapid enumeration of these active indexes. When a precise count is required for larger sets, this mechanism can be used to keep track of items that have been counted already, which increases the time required to make a numerosity judgment.

There are theoretical disagreements on the interpretation of the performance differences between small and large sets. Some studies attribute the change in the reaction times to the capacity limitations of information transfer into short-term memory (Cowan, 2001; Klahr, 1973) or a shifting of enumeration strategies (Mandler & Shebo, 1982). The rapid identification of small-set numerosity also can be attributed to the fast mapping of a label to the discrete increments on a mental magnitude (Gallistel & Gelman, 1991) or the fast counting of active indexes (Trick & Pylyshyn, 1994). Whether two systems are responsible for enumeration has yet to be determined conclusively, and this area of research continues to provide evidence supporting both perspectives.

Regardless of the mechanism responsible for subitizing, accurately enumerating a set requires the selection of each visual object. If an indexing mechanism is responsible for subitizing, observers would be able to report on four objects even under time constraints, but with poor memory for locations. Alternatively, if each object must be encoded into working memory for recall, then errors in enumeration and location recall should be similar. Numerosity perception has been studied extensively but little is known about the spatial information that is encoded when enumerating. To address this topic, the current study examines the location encoding that occurs in subitizing.

Studies on the spatial coding of object locations have shown that observers tend to remember locations by using spatial cues to categorize locations according to geometric "prototypes" (Huttenlocher, Hedges, & Duncan, 1991). When presented with a dot inside a geometric shape,

children remembered the location as being further away from the midline and edges of that shape—a bias towards the central tendency of the shape category, or prototype (Huttenlocher, Newcombe, & Sandberg, 1994). In adults, the representation of locations also was biased towards the prototype of spatial categories and these biases increased as memory became less certain over extended response delays (Spencer & Hund, 2002). These studies suggest that a single system for representing space is likely to serve both verbal and motor responses that are spatial in nature (Spencer, Simmering, & Schutte, 2006).

One potentially useful approach to understanding enumeration is to apply statistical and computational methods used in the study of visual perception. For example, one recent study used an information theoretic framework to model the human ability to learn statistical regularities from object features in visual displays, and tested whether observers used this information to enhance their ability to identify the locations of specific colors (Brady, Konkle, & Alvarez, 2009). The authors hypothesized that if there were more redundancies in the information input, then more content can be stored (as predicted by information theory). Their results indicate that more regular displays did in fact facilitate the encoding of information, which increased color recall performance in a way that could be predicted by a Bayesian learning model.

The primary goal of the current study is to characterize the spatial encoding during the enumeration of small sets of dots that were randomly placed on a computer screen and to determine if location and enumeration accuracy can be predicted by the statistical or geometric properties of these displays. To investigate this possibility, we devised an enumeration task that presented a display with randomly-placed small black discs. After a mask, observers marked the perceived location of each disc, which also served as their numerosity response (see Figure 1). Three stimulus durations (50, 200, or 350 ms) and eight numerosities (2-9) were tested. These stimuli were presented very briefly in order to prevent verbal counting and the response method allowed for a nonverbal report of numerosity and location (similar to a reporting methodology described in Dent & Smyth, 2006). *Enumeration accuracy* is measured as the percent of trials with an accurate numerosity report and the average (absolute) number of miscounts. For each trial, each disc on a response display was paired with a disc on the stimulus display to determine *location accuracy*, which is the distance between these corresponding discs.

The location data from this experiment was used to characterize observers' representations of objects selected for enumeration. The properties of the disc configurations in the test displays were compared to those in the observers' responses. This enabled quantitative comparisons between the actual stimulus and its representation. One testable prediction is that a display with more regularity would allow more content to be encoded more accurately into working memory, leading to better enumeration performance and object localization. Display regularity was obtained by applying Delaunay Triangulation methods to identify "simplexes"—triangles with vertices comprised of display discs without other discs inside them (Kendall, 1989). This triangulation was applied to the elements in both the test and response displays, and the average area and side lengths of the resulting triangles were computed for each display. "Maximal circles", which connect the vertices of each triangle simplex, have also been used to study regularity in the spacing between dots (Fidopiastis, Hoffman, Prophet, & Singh, 2000). Similarly, maximal circles were identified and the average radii of these circles was computed and compared to observer responses. Another form of statistical summary examined was the centroid of disc configurations. Humans can estimate the center-of-mass of an array of randomly arranged dots on a display with high accuracy (Juni, Singh, & Maloney, 2008; Zhou, Chu, Li, & Zhan, 2006). The computation of this centroid estimate may prove to be crucial when representing individual locations. For each display, we computed the centroid and the distances of each element on the display from its centroid. We then compared the values between the stimulus and response data in order to estimate variability and compression.

The various regularity measures described above may be used to develop a model that predicts enumeration and localization performance. The current study aims to contribute to this goal by characterizing the spatial encoding during enumeration. This can lead to a better understanding of the nature of numerosity representations obtained under brief viewing conditions and help identify the mechanisms that contribute to this process. Using the characteristics of possible mechanisms—such as the Weberian nature of a magnitude mechanism or the set-based limitation of an indexing mechanism—we can test which model best explains the current data and identify the properties that are better predictors of accurate enumeration.

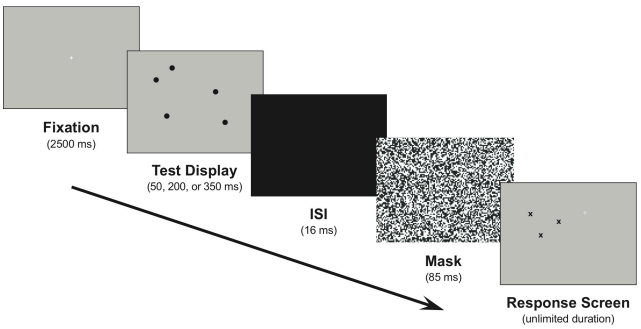


Figure 1. Schematic of this enumeration experiment.

II. Methods

Participants: 24 Rutgers University undergraduates participated in one session for course credit or payment.

Apparatus: The experiment was programmed in MATLAB with Psychophysics Toolbox 3.0.8 (Brainard, 1997) and presented using a desktop computer running Windows XP

(Intel Pentium 4 processor). The stimuli were displayed on a 19" color CRT monitor with a resolution of 1280 x 1024 pixels and a refresh rate of 70 Hz; contrast was set to 100% and brightness was set to 50%. The screen dimensions were approximately 35° by 27° in visual angle.

Stimuli*:* Test displays contained 2-9 identical black discs (35 pixels in diameter, or ~1°) presented on a gray screen for 50, 200, or 350 ms. The discs were randomly placed on the screen with the following constraints: discs could not appear within 115 pixels (~3°) or more than 715 pixels (~20°) of each other, or within ~200 pixels of the screen edges. This produced an effective viewing display of 21° by 16° (768 x 614 pixels). Adequate separation of objects was emphasized to ensure "preattentive" object discriminability, since more attentional resources are required for accurate discrimination when separated by less than 1° (Bahcall & Kowler, 1999). The test display was masked using a random-dot texture created by randomly assigning a white or black value to a grid of 4 x 4 pixel squares.

General procedure: Observers sat approximately 60 cm from a computer screen in a darkened room. They were given instructions by the experimenter and performed six practice trials to ensure understanding of the task. Each trial began with a 2,500 ms presentation of a gray screen with a white central fixation cross. The stimulus screen was then flashed for a designated duration. A black screen appeared for one frame (16 ms) before a mask comprised of a random-dot texture was presented for 85 ms. Finally, a gray input screen with a crosshair pointer appeared and remained until observers made their responses by placing markers ("X") on each of the perceived disc locations. Pressing the space bar initiated the next trial. It was emphasized to the observers that the number of markers placed on the screen should represent the number of discs seen on the test display, even if they were unsure about the exact location. Response coordinates were recorded by the program. See Figure 1 for a diagram of a trial.

Processing the location data: The location data was comprised of two files, one for the stimulus display and another for the response display. In order to analyze the accuracy of location representations, stimulus and response coordinates (x-y values) were paired using the following procedure. When a trial had the same number of stimulus and response elements (i.e., correctly enumerated displays), a Procrustes analysis on the convex hulls of the element locations was used to identify the best fit of the response to the stimulus coordinates for each trial. Procrustes analysis determines the similarity between two shapes by estimating the best fit of one set of points to a comparison set by factoring out variations in scaling, rotation, and translation (Goodall, 1991). After applying the relevant scaling, rotation, or coordinate position transformations, Delaunay Triangulation and nearest-neighbor methods were used to identify stimulus-response pairs. For calculating pattern regularity on a display, the mean and variance values were computed for the areas of triangle simplexes (identified by the triangulation), connecting edges, and the radii of the maximal circles that circumscribe the triangle simplexes. Trials with unpaired discs, which primarily occurred when displays were under- or over-counted, were not included in the location analysis (15% of possible data points).

III. Results

Enumeration Accuracy

The enumeration results replicate previous studies, with the highest accuracy observed in low numerosities. This range was maintained for six items—better than in previous studies where accuracy declines after four items. A follow-up experiment was conducted that included a control where numerosity was reported using Arabic numerals (Haladjian, Pylyshyn, & Gallistel, 2009). Observers performed better in the location-marking block (six items) than the control block (four items), supporting the current results.

Analysis of variance was conducted on the enumeration performance with observer included as a random variable. The largest numerosity condition of nine discs was excluded to control for anchoring effects. Analyzing the proportion of trials with perfect enumeration revealed main effects for display duration (F=34.7(2,276), p<.01) and numerosity (F=68.8(6,276), p<.01), with interactions (F=7.7(12,276), p<.01). Analyzing the absolute value of miscounts for each condition also revealed main effects for display duration (F=36.1(2,276), p<.01) and numerosity (F=51.2(6,276), p<.01), with interactions (F=11.8(12,276), p<.01). Figure 2 depicts the proportion of trials correctly enumerated and Figure 3 depicts the average absolute number of miscounts. Errors increased with larger numerosities but fewer errors were found with longer display durations. When observers made errors, they were generally underestimates (84% of errors were underestimates). Performance in the 50-ms display was significantly worse than the 200- and 350-ms durations for the 6-9 disc displays in both these analyses.

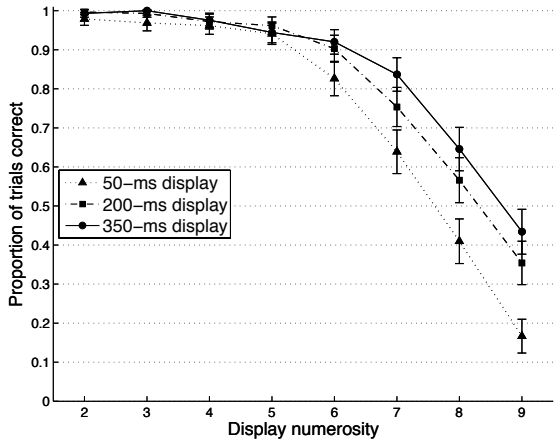


Figure 2. Proportion of trials with correct enumeration.

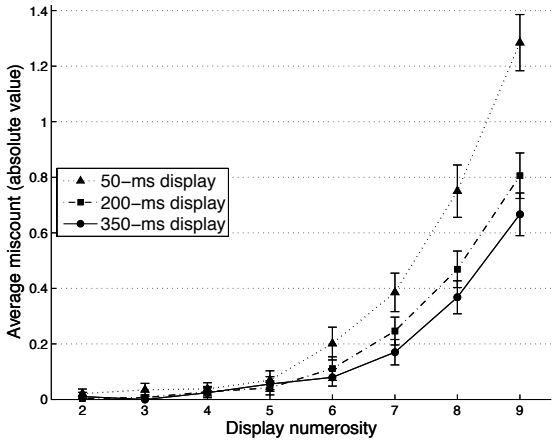


Figure 3. Average counting errors.

Location Accuracy

Location error is reported as the Euclidean distance between the coordinates of stimulus-response pairs for each trial. ANOVA results indicate main effects for display duration (F=27.2(2,276), p<.01) and numerosity (F=81.4(6,276), p<.01), with no interactions (F=1.4(12,276), p=.15). Errors increased with larger numerosities and generally decreased with longer display durations (see Figure 4). The mean and variance of the following variables were computed to estimate display regularity: 1) area of Delaunay "simplex" triangles; 2) length of the triangle segments (shared edges were counted only once); 3) radii of the maximal circles that circumscribed the simplexes; 4) distance between each disc and the display centroid; and 5) radius of the enclosing "circumcircle" around the display elements (to estimate disc dispersion). Since performance was significantly worse in the 50-ms displays, only data from the 200- and 350-ms display durations (combined) are reported here.

The centroid (or center-of-mass) for each display was computed by calculating the mean x- and y-coordinate of all discs on a display. The compression measure is shown in Figure 5 as the average centroid-to-disc distances, that is, the average distance from discs on a display to the centroid. The substantially smaller distances in the observers' responses suggests that their representation is compressed around the centroid of the display. The average dispersion (minimum enclosing circle radius) of the discs on a stimulus display ranged from 203 pixels (SD=73) in 2-numerosity displays to 358 pixels (SD=19) in 9-numerosity displays; for response data, this dispersion ranged from 185 pixels (SD=73) to 314 pixels (SD=44), indicating compression.

Display regularity was measured in terms of the variability in the size of the Delaunay simplexes and the size of the maximal circles that circumscribe these triangles. Here we report the effects of regularity as measured by the variability in the edge lengths of Delaunay simplexes; however, similar patterns of results were obtained with the area of the simplexes and the size of the maximal circles. Figure 6 depicts the average segment lengths and also suggests a compression of these representations. To compare levels of display regularity, the standard deviation of the triangle segments in the test displays were grouped into quartiles, where 25% of the trials with least variation are in the first quartile and 25% of trials with the most variation are in the last quartile. This allowed us to plot location errors as functions of increasing variability (decreasing regularity) in Figure 7 and counting errors in Figure 8. These two charts show that displays with lower variability produce lower errors in both counting and localization (counting performance for displays <6 items are not shown since observers performed almost perfectly).

To compare the regularity of the test and response patterns, the overall compression in the response patterns was first undone using the scaling estimate from the Procrustes analysis. The variance in the simplex segment length for these "uncompressed" response patterns was then compared to, and found to be lower than, the variance in the corresponding stimulus patterns. This suggests that observers imposed regularity on the response patterns than there was not present in the stimulus patterns. Figure 9 plots stimulus and response data from two representative trials, which illustrates the imposed compression and regularity.

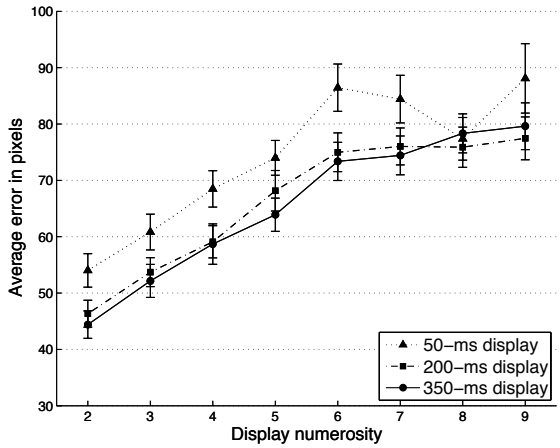


Figure 4. Average location errors in pixels.

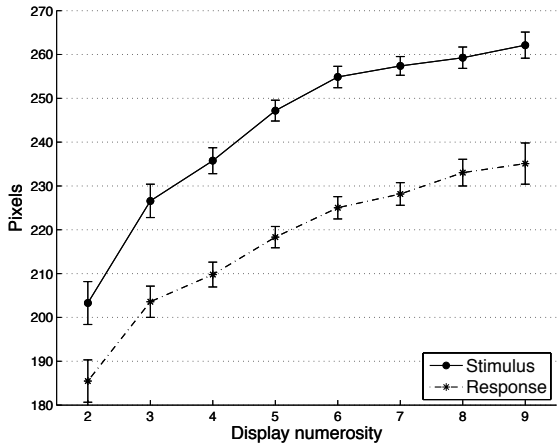


Figure 5. Average centroid-to-disc distance in pixels (200 & 350 ms displays combined).

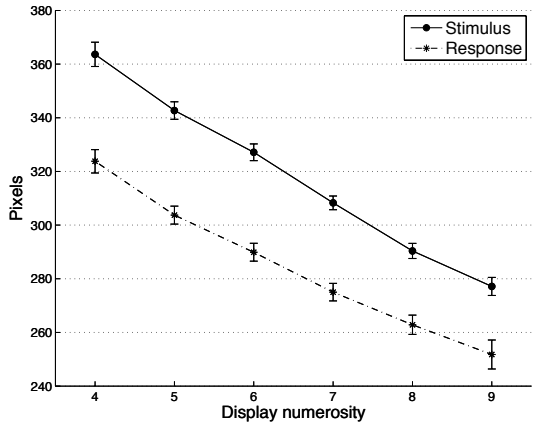


Figure 6. Average segment lengths of Delaunay triangle simplexes (200 & 350-ms displays combined).

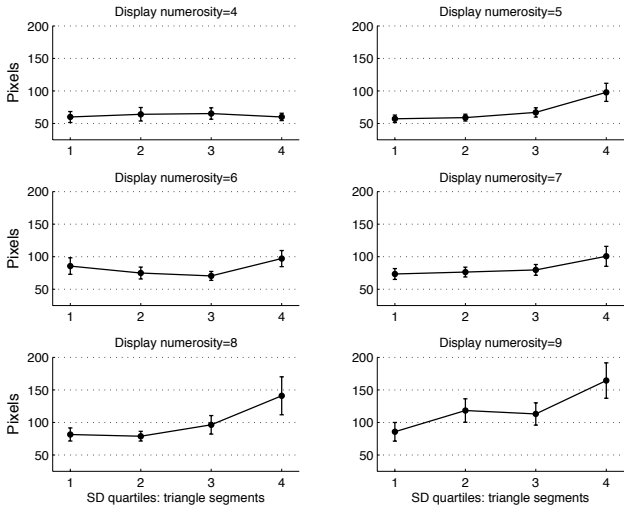


Figure 7. Location errors as a function of increasing triangle segment variability (200 & 350-ms displays combined).

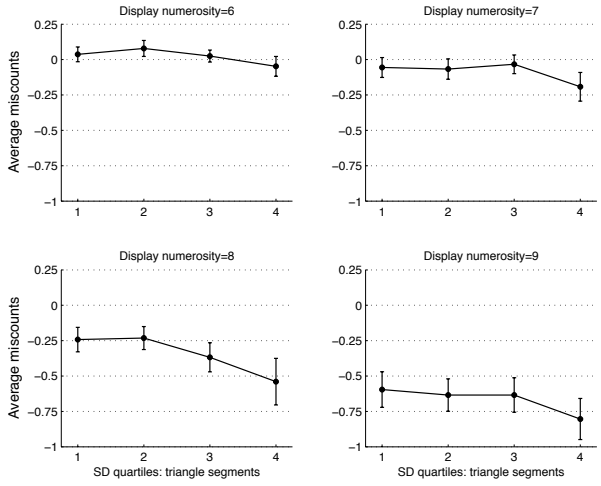


Figure 8. Counting errors as a function of increasing triangle segment variability (200 & 350-ms displays combined).

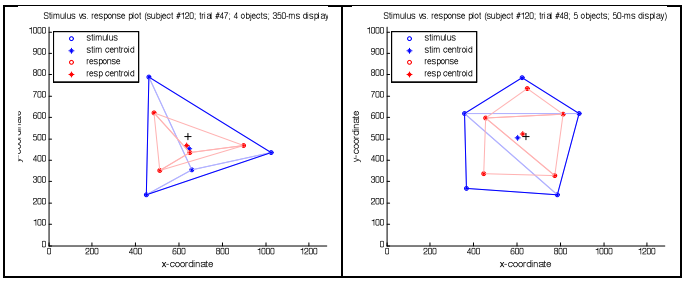


Figure 9. Representative samples of location data with the triangulation simplexes drawn.

IV. Discussion

The visual system is thought to use redundancies from visual stimuli in order to encode information efficiently, as proposed by information theory applications to perception (Attneave, 1954). The current results showing better performance in displays with more regular patterns indicates a more efficient encoding of object locations that may be supported by an information theory of perception. When the triangle simplexes of a display have less variance, observers are more accurate in representing these more regular displays and exhibit better enumerating and localization performance. Additionally, there appears to be a tendency for compressing distances around the centroid. Even after factoring out the overall compression in the response patterns, these distances were found to be less variable in the response configurations than in the test configurations. This could indicate that observers are either assuming there is more regularity when they reconstruct the image, or representation errors are biased towards less variability or towards more "prototypical" representations of shape. This observed tendency to impose regularity on variable displays supports findings from previous studies (e.g., Taylor, 1961).

Increasing stimulus exposure durations from 50 ms to 200 ms produced more accurate enumeration for numerosities greater than six and more accurate location encoding for all numerosities. This suggests a coarse location-estimation process that occurs initially and is updated over time. The disassociation in enumeration and location performance for the smaller numerosity range also suggests that enumeration occurs independent of location-encoding: attention may be required to effectively encode locations but subitizing may be preattentive. This may indicate that visual indexes are responsible for subitizing, since location information does not need to be encoded initially to assign an index, but over time information can be bound to these indexes in order to build more accurate feature representations, including locations (Pylyshyn, 1989). The current results suggest that the indexing mechanism is implemented for smaller numerosities, but further experiments to support this conclusion are required.

The current experiment describes a novel methodology that implements a nonverbal report of numerosity, which appears to enable high enumeration accuracy of six items.

Allowing observers to enumerate by location may be a more accurate demonstration of selection abilities during fast enumeration, and this type of selection is sensitive to the geometric and statistical properties of the visual input. The observed location errors occur systematically and may benefit from inherent geometric regularities. Further analyses of these location data from a statistical perception or information theoretic perspective promise to reveal important information about the spatial nature of numerosity representations.

Acknowledgments

This research was supported by NSF 0549115: IGERT program in perceptual science at Rutgers University (HHH).

References

Attneave, F. (1954). Some informational aspects of visual perception. *Psychological Review, 61*(3), 183-193.

Bahcall, D. O., & Kowler, E. (1999). Attentional interference at small spatial separations. *Vision Research, 39*(1), 71-86.

Brady, T. F., Konkle, T., & Alvarez, G. A. (2009). Compression in visual working memory: Using statistical regularities to form more efficient memory representations. *Journal of Experimental Psychology: General, 138*(4), 487-502.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10*(4), 433-436.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24*(1), 87-114; discussion 114-185.

Dehaene, S. (1992). Varieties of numerical abilities. *Cognition, 44*(1-2), 1-42.

Dent, K., & Smyth, M. M. (2006). Capacity limitations and representational shifts in spatial short-term memory. *Visual Cognition, 13*(5), 529-572.

Fidopiastis, C., Hoffman, D. D., Prophet, W. D., & Singh, M. (2000). Constructing surfaces and contours in displays of color from motion: The role of nearest neighbors and maximal disks. *Perception, 29*(5), 567-580.

Gallistel, C. R., & Gelman, R. (1991). Subitizing: The preverbal counting process. In W. Kessen, A. Ortony & F. I. M. Craik (Eds.), *Memories, thoughts, and emotions: Essays in honor of george mandler* (pp. 65-81). Hillsdale, N.J.: L. Erlbaum Associates.

Gallistel, C. R., & Gelman, R. (1992). Preverbal and verbal counting and computation. *Cognition, 44*(1-2), 43-74.

Gallistel, C. R., & Gelman, R. (2000). Non-verbal numerical cognition: From reals to integers. *Trends in Cognitive Sciences, 4*(2), 59-65.

Goodall, C. (1991). Procrustes methods in the statistical analysis of shape. *Journal of the Royal Statistical Society. Series B (Methodological), 53*(2), 285-339.

Haladjian, H. H., Pylyshyn, Z. W., & Gallistel, C. R. (2009). *Enumerating by location increases the subitizing limit*. Paper presented at the 17th Annual Object Perception, Attention, and Memory (OPAM) Meeting of the Psychonomic Society.

Huttenlocher, J., Hedges, L. V., & Duncan, S. (1991). Categories and particulars: Prototype effects in estimating spatial location. *Psychol Rev, 98*(3), 352-376.

Huttenlocher, J., Newcombe, N., & Sandberg, E. H. (1994). The coding of spatial location in young children. *Cogn Psychol, 27*(2), 115-147.

Juni, M. Z., Singh, M., & Maloney, L. T. (2008). Testing for robustness in visual localization of dot clusters without part structure [abstract]. *Journal of Vision, 8*(6), 1014a.

Kaufman, E. L., Lord, M. W., Reese, T. W., & Volkmann, J. (1949). The discrimination of visual number. *American Journal of Psychology, 62*(4), 498-525.

Kendall, D. G. (1989). A survey of the statistical theory of shape. *Statistical Science, 4*(2), 87-120.

Klahr, D. (1973). Quantification processes. In W. G. Chase (Ed.), *Visual information processing* (pp. 3-34). New York: Academic Press.

Mandler, G., & Shebo, B. J. (1982). Subitizing: An analysis of its component processes. *Journal of Experimental Psychology: General, 111*(1), 1-22.

Meck, W. H., & Church, R. M. (1983). A mode control model of counting and timing processes. *Journal of Experimental Psychology: Animal Behavior Processes, 9*(3), 320-334.

Pylyshyn, Z. W. (1989). The role of location indexes in spatial perception: A sketch of the finst spatial-index model. *Cognition, 32*(1), 65-97.

Spencer, J. P., & Hund, A. M. (2002). Prototypes and particulars: Geometric and experience-dependent spatial categories. *Journal of Experimental Psychology: General, 131*(1), 16-37.

Spencer, J. P., Simmering, V. R., & Schutte, A. R. (2006). Toward a formal theory of flexible spatial behavior: Geometric category biases generalize across pointing and verbal response types. *Journal of Experimental Psychology: Human Perception and Performance, 32*(2), 473-490.

Taylor, M. M. (1961). Effect of anchoring and distance perception on the reproduction of forms. *Perceptual and Motor Skills, 12*, 203-230.

Trick, L. M., & Pylyshyn, Z. W. (1994). Why are small and large numbers enumerated differently? A limited-capacity preattentive stage in vision. *Psychological Review, 101*(1), 80-102.

Zhou, X., Chu, H., Li, X., & Zhan, Y. (2006). Center of mass attracts attention. *Neuroreport, 17*(1), 85-88.

Multiple visual cues enhance quantitative perception in infancy

Joseph M. Baker (Joseph.Baker@aggiemail.usu.edu)
Utah State University, Department of Psychology
2810 Old Main Hill, Logan UT, 84322

Jessica M. Feigleson (Jessica.Feigleson@gmail.com)
Utah State University, Department of Psychology
2810 Old Main Hill, Logan UT, 84322

Kerry E. Jordan (Kerry.Jordan@usu.edu)
Utah State University, Department of Psychology
2810 Old Main Hill, Logan UT 84322

Abstract

Infants possess basic capabilities to assess various quantitative properties such as number, size, and time. Preverbal discriminations are approximate, however, and similarly limited by ratio across these dimensions. Here, we present the first evidence that redundant quantitative unisensory information—namely, simultaneous visual cues to both number and size—accelerates six-month-olds' quantitative competence. Using a habituation-dishabituation paradigm, results demonstrate that, when provided with synchronous visual cues to different quantitative properties, infants make more precise discriminations than when they receive information about a single cue alone. Such redundant conceptual information may be more salient than non-redundant information, which could better recruit attention and result in more precise learning and remembering than when such information is presented through only one cue.

Introduction

Even before they acquire language, infants are capable of perceiving various quantitative dimensions such as time, size, and number (e.g., Brannon, Lutz, and Cordes, 2006; Jordan, Suanda, and Brannon, 2008; Lipton and Spelke, 2003; Wood and Spelke, 2005; Xu and Spelke, 2000; Xu, 2003; Xu et al., 2005). This ability is approximate in that it follows predictions made by Weber's Law: infants' ability to distinguish between two magnitudes is a function of the ratio between the competing magnitudes (e.g., Xu, & Spelke, 2000; Bijeljac-Babic, Bertioncini, & Mehler, 1993; Brannon, Abbot, & Lutz, 2004; Gao, Levine, & Huttenlocher, 2000; Izard, Sann, Spelke, & Streri, 2009). For example, six month-olds successfully distinguish a 1:2 size change of a visual object, but fail to notice a 2:3 size change (Brannon, Lutz, & Cordes, 2006). Similarly, when tested with visual stimuli, six-month-old infants require a 1:2 ratio for successful discrimination of large numerical sets, failing to discriminate a 2:3 ratio change (e.g., Wood and Spelke, 2005; Xu and Spelke, 2000; Xu, 2003; Xu et al., 2005). By 9 months of age, however, infants successfully discriminate a 2:3 change in number (Lipton, & Spelke, 2003). This pattern suggests that discrimination abilities of infants show similar perceptual and cognitive limits across various dimensions of visual quantity.

When provided with redundant information about quantity through multiple sensory modalities, however, infants' discrimination abilities improve to a level previously thought attainable only after additional months of development. For example, although six-month-old infants fail to distinguish a 2:3 ratio change in number when provided with a single visual cue to number, they succeed when given cues simultaneously in both the visual and auditory modalities (Jordan, Suanda, & Brannon, 2008). Five-month-olds discriminate differing rhythmic patterns when presented audiovisually, but fail when presented only with a cue from the auditory or visual modality alone (Bahrick and Lickliter, 2000). Similarly, 3-month-olds are capable of discriminating differing tempos only when they are presented redundantly across multiple modalities (Bahrick, Flom, and Lickliter, 2002).

This tendency for infants' perceptual abilities to improve when provided with multiple synchronous sensory cues has been explained by the Intersensory Redundancy Hypothesis (see Bahrick and Lickliter, 2000 for review). The hypothesis states that redundant stimulation from multiple sensory modalities efficiently recruits infants' attention by providing overlapping sensory information, thereby causing the redundantly specified property to become perceptual "foreground", while other sensory stimulation remains "background". This, in turn, fosters perceptual differentiation, learning, and memory for redundant, amodal properties before other unisensory, modality-specific stimulus properties.

The effect of *intrasensory* redundancy on infants' cognitive abilities, however, remains to be empirically tested. Multisensory stimulation in the form of intersensory redundancy may not be the only way to boost infant quantitative competency; the multiple numerical cues provided by intersensory redundancy may be more important than the multisensory nature of the stimulation itself. Redundant conceptual information, regardless of

sensory modality, may be more salient than non-redundant information, which could better recruit attention and result in more precise learning and remembering than when such information is presented through only one cue. Here, we test the hypothesis that redundant *visual* stimuli will improve infants' quantitative discrimination abilities.

Methods

Participants

Twenty eight full-term six-month-old infants were tested (female= 14, mean age = 6 months 2 days; range: 5 months 15 days to 6 months 14 days). Participants were recruited from birth records obtained from the Utah Department of Health.

Design

Infants were randomly assigned to one of two conditions. Thirteen infants were habituated to a silent movie in which a ball bounced 8 times, while the remaining 15 infants were habituated to a silent movie in which a ball bounced 12 times. The size of the ball which bounced 8 times was exactly two-thirds the surface area of the ball that bounced 12 times. Following habituation, infants were tested with novel silent movies in which the ball bounced 8 or 12 times in alternation for six trials (order counterbalanced). The relative size difference of the balls bouncing 8 versus 12 times remained during test. Importantly, these stimuli closely mirror previous studies (see Jordan, Suanda, and Brannon, 2008 for test of 6-month-olds with these exact stimuli on number discriminations; see Brannon, Lutz, and Cordes, 2006 for test of 6-month-olds on discriminating this ratio of surface area), which demonstrated that 6-month-old infants fail to discriminate this ratio when either number or surface area are tested in isolation.

Stimuli

Infants were habituated to silent movie events of a ball that appeared to drop and then bounce up after making contact with a surface (see Jordan, Suanda, and Brannon, 2008). The size of the ball differed depending on its number of bounces: A ball which bounced 8 times (14.07 cm) covered exactly two-thirds of the total surface area as a ball which bounced 12 times (21.11 cm). The size of the individual ball was fixed and did not vary during the movie. Movies were constructed using Macromedia Flash and displayed on a computer within a 19 x 23 cm area.

Temporal parameters (Table 1) were controlled following Wood and Spelke (2005). During habituation, rate, duration, inter-event interval, and height of individual ball bounces were approximately equal for the 8- and 12-bounce sequences and were constant within trial but varied across trial. Therefore, on average during habituation 12-bounce sequences lasted longer and contained more motion than 8-bounce sequences. In contrast, during test sequences, total sequence duration, cumulative height of ball bouncing, and total inter-event interval were approximately equal for the 8-bounce and 12-bounce sequence.

There were six distinct habituation sequences for 8- and 12- bounce events. During the habituation phase, the six movies in each condition repeated in random order for 16 trials, or until the infant met the habituation criterion. The six test trials consisted of novel 8- and 12-bounce movies shown in alternation, and occurred randomly without replacement for the first six habituation trials.

Table 1. Habituation and test trial parameters

	Bounce Duration (ms)	Interbounce Interval	Total Bounce Duration	Total Interbounce Interval	Total Duration of Sequence	Bounce Height (cm)	Total Bounce Height
Test Trials							
8	471	367	3768	2569	6337	10	80
12	315	233	3780	2563	6343	6.67	80
Habituation Trials							
8	67	333	536	2664	3200	1	8
8	200	300	1300	2400	4000	4	32
8	333	233	2664	1864	4528	8	64
8	400	267	3200	2136	5336	7	56
8	533	200	4264	1600	5864	11	88
8	667	167	5336	1336	6672	14	112
Mean	367	250	2933	2000	4933	7.5	60
12	67	333	804	3996	4800	1	12
12	200	300	2400	3600	6000	4	48
12	333	233	3996	2796	6792	8	96
12	400	267	4800	3204	8004	7	84
12	533	200	6396	2400	8796	11	132
12	667	167	8004	2004	10008	14	168
Mean	367	250	4400	3000	7400	7.5	90

Surface area, circumference, and diameter remained constant within each trial type for 8 and 12 bounce videos respectively. Surface Area (8 = 14.07cm, 12 = 21.11cm); Circumference (8 =13.30cm, 12 = 16.28cm); Diameter (8 = 4.23cm, 12 =5.18cm).

Apparatus and Procedure

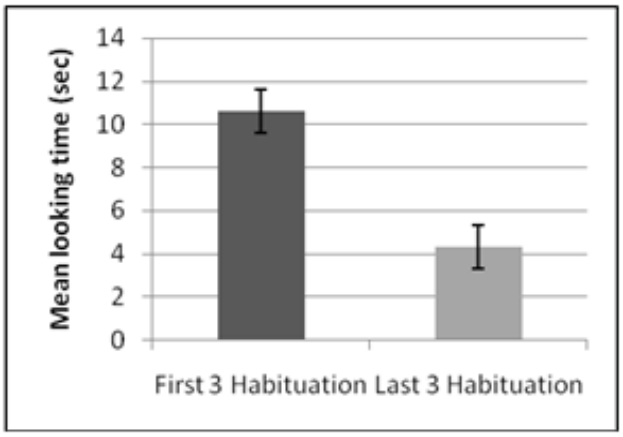
Infants sat on their guardian's lap approximately 60 cm away from a large monitor. Each test trial was initiated by the experimenter when the infant was looking in the direction of the monitor. Following the movie sequence, the last frame of the video was held fixed and remained on the screen for the remainder of the trial. Looking time to the fixed frame was recorded until the infant looked away for a continuous 2 seconds after looking at the fixed image for a minimum of 1 s, or after a maximum of 60 s. The habituation phase continued until the infant met the habituation criterion, defined as a 50% reduction in looking time over 3 consecutive trials relative to the first 3 trials that summed to at least 12 s, or until 16 trials were completed. The test phase consisted of silent movies, three of which

consisted of 8-bounce and three of which consisted of 12-bounce events.

A micro-camera monitoring the infant's face and a feed from the stimulus presentation computer were multiplexed onto a TV monitor and VCR. Each session was recorded for later reliability coding. Inter-rater reliability averaged 93.63% for all infants. Twenty-five percent of infants were coded by a single trained coder.

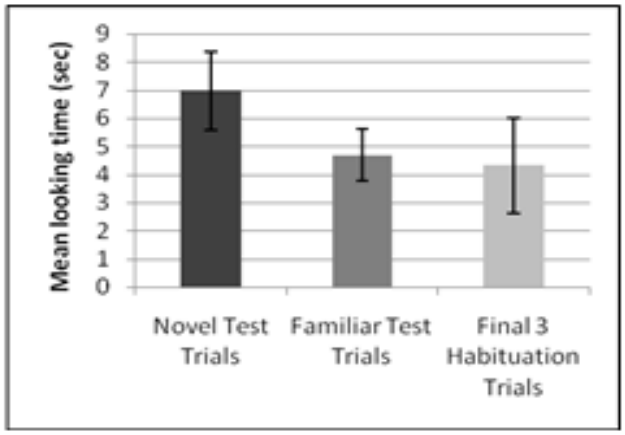
Results

An alpha of .05 was used for all comparisons. Infants required an average of 9.68 trials to reach habituation. Twenty five of the 28 infants reached habituation before all 16 habituation trials were complete. A paired-sample *t* test indicated that infants viewed the first 3 habitation trials significantly longer than the final 3 habituation trials, $t(27) = 9.46$, $p < .001$, Cohen's d = 3.71.

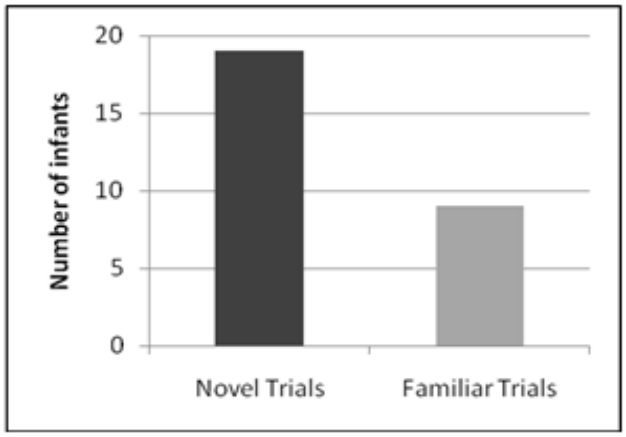


A 2 x 2 x 2 repeated measures ANOVA was used to examine the relationship between the between-subject factors, gender and habituation condition (eight vs. twelve), and the within-subject factor, test trial type (novel vs. familiar). This analysis identified a significant main effect of test trial type, indicating that looking time within novel test trial types (μ = 6.92 seconds) were significantly longer than looking times within familiar test trial types (μ = 4.62 seconds; $F(1, 24) = 6.459$, $MSE = 12.02$, $p = .018$, Cohen's $d = .689$). No other main effects or interactions were significant.

A priori paired samples *t* test comparisons between pairs of novel and familiar test trials identified significant looking time differences between the first pair of novel and familiar test trials ($t(27) = 2.474$, $p = .02$, Cohen's $d = .970$), and the second pair of novel and familiar test trials ($t(27) = 2.196$, $p = .037$, Cohen's $d = .861$). However, looking time between novel and familiar test trials within the third pair of test trials were not significantly different ($t(27) = -.858$, $p = .399$, Cohen's $d = .336$). Infants looked significantly longer at novel test trials than the final three habituation trials ($t(27) = 3.26$, $p = .003$, Cohen's $d = 1.28$). However, looking times for familiar test trials did not differ from the final three habituation trials ($t(27) = .641$, $p = .527$, Cohen's $d = .251$).



Binomial sign tests indicate that the number of infants who looked longer at novel test trial types (n = 19) was significantly larger than the number of infants who looked longer at familiar test trial types (n = 9) (sign-test, $p = .043$).



Discussion

These results demonstrate that infants' quantitative visual discriminations do not adhere to one strict ratio limit. With visual stimuli providing two simultaneous quantitative cues, six-month-old infants make more precise discriminations than previously reported in other studies when they received information about either number or surface area alone (Jordan, Suanda, and Brannon, 2008; Brannon, Lutz, and Cordes, 2006). Our experiment is therefore the first to demonstrate an increase in quantitative precision resulting from redundant information provided by synchronous visual cues.

Past research has demonstrated that multiple sources of information presented across sensory modalities must be synchronous in order to be beneficial to infant learning in various domains (Bahrick, & Lickliter, 2000). In the current experiment, the change in number and surface area occurred in synchrony; however, the change did not occur across sensory modalities. Thus, multisensory stimulation in the

form of intersensory redundancy is not the only way to boost infant quantitative competence (Jordan et al., 2008). Instead, the multiple numerical cues provided by intersensory redundancy may be more important than the multisensory nature of the stimulation itself. Redundant conceptual information, regardless of sensory modality, may be more salient than non-redundant information, which could better recruit attention and result in more precise learning and remembering than when such information is presented through only one cue.

The exact mechanism through which infants accomplish this feat remains unknown. Recent epigenetic theories suggest that infants are endowed with basic capabilities to detect and process intersensory stimuli at birth, but that experiences throughout development are necessary to enhance these abilities (Lewkowicz, 2000). These experiences may begin in utero (Kisilevsky, 1995; Schaal, Orgeur, & Rognon, 1995; Bekoff, 1995), and continue throughout infancy and into childhood (Milner, & Bryant, 1968). However, as suggested by Lewkowicz (2000) only relatively recently have studies begun acknowledging and investigating the pervasive effects of multimodal stimulation on infants' perceptual abilities.

One area that has until recently been largely overlooked are infants' abilities to learn from synchronous stimulation within a single sensory modality. In their everyday environments, it is possible that infants experience synchronous unisensory stimulation no less often than they experience synchronous multisensory stimulation. Therefore, given the proposed importance of experience on the development of perceptual capabilities, the capacity of infants to perceive synchronous unisensory cues may be similar to their capacity to perceive synchronous multisensory cues in many domains. The current results mark an initial step in better understanding how infants utilize multiple temporally and spatially synchronous cues from the same sensory modality.

Our results are the first to indicate that infants' abilities to discriminate between quantities are improved by the presentation of redundant *intrasensory* cues. Redundant visual stimulation may cause more effective encoding of quantity by selectively recruiting infants' attention to visual properties of magnitude, thereby resulting in increased neural responsiveness to synchronous, redundant quantitative information. Therefore, when given multiple synchronous intrasensory cues, infants may have experienced greater signal strength and decreased variance for their ratio-dependent quantitative representations in memory.

It is necessary, however, to investigate the physiological bases of this enhanced quantitative ability in order to clarify potential causes and correlates. Recent findings have shown promise in identifying patterns of brain activation specifically involved in numerical understanding, even in infancy (Libertus, Pruitt, Woldorff, & Brannon, 2009). It has been hypothesized that there may be similar mental algorithms and neural areas devoted to common magnitude processing, as opposed to completely discrete, compartmentalized areas responsible for processing specific quantitative properties individually (e.g., Cantlon et al., 2009).

Much work is still needed now to determine the extent of these findings. For instance, what other synchronous visual stimulation is capable of improving infants' quantitative abilities? Do these findings generalize outside of the domain of quantity, across other amodal properties? Does discrimination of properties in other sensory modalities such as audition benefit from synchronous intrasensory presentation? To better understand the role of synchrony in producing such effects, it would also be informative in the future to provide infants with the same overall amount of information about size and number changes, but to present these changes asynchronously. These questions should be addressed before we can begin to understand the limits of infant intrasensory processing capabilities.

References

Bahrick, L. E., Flom, R., & Lickliter, R. (2002). Intersensory redundancy facilitates discrimination of tempo in 3-month-old infants. *Developmental Psychobiology, 41*, pp. 352-363.

Bahrick, L. E., & Lickliter, R. (2000). Intersensory redundancy guides attentional selectivity and perceptual learning in infancy. *Developmental Psychology, 36(2)*, pp. 190-201.

Bekoff, A. (1995). Development of motor behavior in chick embryos. In J. P. Lecanuet, W. P. Fifer, N. A. Krasnegor, & W. P. Smotherman (Eds.), *Fetal development: A psychobiological perspective* (pp. 191-204). Hillsdale, NJ: Erlbaum.

Bijeljac-Babic, R., Bertioncini, J., & Mehler, J. (1993). How do 4-day-old infants categorize multisyllabic utterances? *Developmental Psychology, 29(4)*, pp. 711-721.

Brannon, E. M., Abbot, S., & Lutz, D. J. (2004). Number bias for the discrimination of large visual sets in infancy. *Cognition, 93*, pp. B59-B68.

Brannon, E. M., Lutz, D., & Cordes, S. (2006). The development of area discrimination and its implications for number representation in infancy. *Developmental Science, 9(6)*, pp. F59-F64.

Cantlon, J., Platt, M.L., & Brannon, E.M. (2009). Beyond the number domain. *Trends in Cognitive Sciences 13*, 83-91.

Gao, F., Levine, S. C., & Huttenlocher, J. (2000). What do infants know about continuous quantity? *Journal of Experimental Child Psychology, 77*, pp. 20-29.

Izard, V., Sann, C., Spelke, E. S., & Streri, A. (2009). Newborn infants perceive abstract numbers. *Proceedings of the National Academy of Sciences, 106(25)*, pp. 10382-10385.

Jordan, K. E., Sudana, S. H., & Brannon, E. M. (2008). Intersensory redundancy accelerates preverbal

numerical competence. *Cognition, 108,* pp. 210-221.
Kisilevsky, B. S. (1995). The influence of stimulus and subject variables on human fetal responses to sound and vibration. In J. P. Lecanuet, W. P. Fifer, N. A. Krasnegor, & W. P. Smotherman (Eds.), *Fetal Development: A psychobiological perspective* (pp. 263-278). Hillsdale, NJ: Erlbaum
Lewkowicz, D. J. (2000). The development of intersensory temporal perception: An epigenetic system/limitations view. *Psychological Bulletin, 126(2),* pp. 281-308.
Libertus ME, Pruitt LB, Woldorff MG, Brannon EM. (2009). Induced alpha-band oscillations reflect ratio-dependent number discrimination in the infant brain. *Journal of Cognitive Neuroscience 21,* 2398-406.
Lipton, J. S., & Spelke, E. S. (2003). Origins of number sense: Large-number discrimination in human infants. *Psychological Science, 14(5),* pp. 396-401.
Milner, A. D., & Bryant, P. E. (1968). Cross-modal matching by young children. *Journal of Comparative and Physiological Psychology, 71(3),* pp. 453-458.
Schall, B., Orgeur, P., & Rognon, C. (1995). Odor sensing in the human fetus: Anatomical, functional, and chemoecological bases. In J. P. Lecanuet, W. P. Fifer, N. A. Krasnegor, & W. P. Smotherman (Eds.), *Fetal development: A psychobiological perspective* (pp. 205-237), Hillsdale, NJ: Erlbaum.
Wood, J. N., & Spelke, E. S. (2005). Infants' enumeration of actions: Numerical discrimination and its signature limits. *Developmental Science, 8(2),* pp. 173-181.
Xu, F. (2003). Numerosity discrimination in infants: Evidence for two systems of representations. *Cognition, 89,* pp. B15-B25.
Xu, F., & Spelke, E. S. (2000). Large number discrimination in 6-month-old infants. *Cognition, 74,* pp. B1-B11.
Xu, F., Spelke, E. S., & Goddard, S. (2005). Number sense in human infants. *Developmental Science, 8(1),* pp. 88-101.

Multidimensional Scaling Methods for Absolute Identification Data

Pennie Dodds (Pennie.Dodds@newcastle.edu.au)
School of Psychology, University Drive
Callaghan, NSW, Australia

Chris Donkin (cdonkin@indiana.edu)
Department of Psychological and Brain Sciences, 1101 E. 10th S.
Bloomington, Indiana

Scott Brown (Scott.Brown@newcastle.edu.au)
School of Psychology, University Drive
Callaghan, NSW, Australia

Andrew Heathcote (Andrew.Heathcote@newcastle.edu.au)
School of Psychology, University Drive
Callaghan, NSW, Australia

Abstract

Absolute identification exposes a fundamental limit in human information processing. Recent studies have shown that this limit might be extended if participants are given sufficient opportunity to practice. An alternative explanation is that the stimuli used – which vary on only one physical dimension – may elicit psychological representations that vary on two (or more) dimensions. Participants may learn to take advantage of this characteristic during practice, thus improving performance. We use multi-dimensional scaling to examine this question, and conclude that despite some evidence towards the existence of two dimensions, a one dimensional account cannot be excluded.

Keywords: absolute identification; unidimensional stimuli; multidimensional scaling; MDS; learning

A typical Absolute Identification (AI) task uses stimuli that vary on only one physical dimension, such as loudness, brightness, or length. These stimuli are first presented to the participant one at a time, each uniquely labeled (e.g. #1 through to *n*). The participant is then presented with random stimuli from the set, without the label, and asked to try and remember the label given to it previously.

This seemingly simple task exhibits many interesting benchmark phenomena. The one of most concern for the current paper is the apparent limitation in performance. The maximum number of stimuli that people were previously thought to be able to perfectly identify was only 7±2 (Miller, 1956). Performance was thought to improve slightly with practice and then reach a low asymptote (Pollack, 1952; Garner 1953).

This finding was particularly surprising given that this limit appeared to be resistant to practice (Garner, 1953; Weber, Green & Luce, 1977), and was generally consistent across a range of modalities (e.g. line length: Lacouture, Li & Marley, 1998; tone frequency: Pollack, 1952; Hartman, 1954; tone loudness: Garner, 1953; Weber, Green & Luce, 1977). In addition, this limitation appears to be unique to unidimensional stimuli. For example, people are able to remember hundreds of faces and names, and dozens of alphabet shapes. It is generally accepted that this is because objects such as faces, names, and letters vary on multiple dimensions. Performance generally increases as the number of dimensions increase (Eriksen & Hake, 1955). This makes intuitive sense when one considers the individual dimensions on a multidimensional object. For example, if people are able to learn to perfectly identify 7 lengths, and 7 widths, they could potentially learn to identify 49 rectangles formed by a combination of lengths and widths.

Despite decades of research confirming this limit in performance for unidimensional stimuli, more recent research has suggested that we may be able to significantly increase this limit through practice (Rouder, Morey, Cowan and Pfaltz, 2004; Dodds, Donkin, Brown & Heathcote, submitted). For example, given approximately 10 hours of practice over 10 days, Dodds et al.'s participants learned to perfectly identify a maximum of 17.5 stimuli (out of a possible 36), a level significantly beyond the 7±2 limit suggested by Miller (1956). From 58 participants that took part in a series of AI tasks, 22 exceeded the upper end of Miller's limit range (nine stimuli).

Other Stimulus Dimensions

The results from Dodds et al. (submitted) were not limited to the identification of lines varying in length. Dodds et al. also used a wide range of other stimuli, and found similar learning effects. For example, dots varying in separation, lines varying in angle and tones varying in pitch all demonstrated similar results. Participants learned to perfectly identify a maximum of 12.6 stimuli using dots varying in separation, 10.4 using lines varying in angle and 17.5 using tones varying in frequency, all exceeding Miller's (1956) upper limit of 9 stimuli.

The learning effects from Rouder et al. (2004) and Dodds et al. (submitted) may be attributed to the type of stimuli employed. The existence of severe limitations in

performance is unique to unidimensional stimuli, and since multiple dimensions are commonly associated with improved performance (Eriksen & Hake, 1955) it may be argued that the stimuli vary on multiple dimensions. Tones varying in frequency for example, are generally viewed as multidimensional. While Dodds et al. employed *pure* tones, leaving the stimuli to vary on only one *physical* dimension (wavelength), our perception of loudness increases as a function of increasing frequency. Therefore as frequency increased, participants would perceive the tones as being of different loudness, creating a greater number of perceived dimensions. This is not an uncommon phenomenon, as a similar effect is found in colour perception. Different colours are generated by a manipulation which is *physically* unidimensional (wavelength change), but the psychological representation of colour is generally considered to consist of three dimensions (e.g., MacLeod, 2003). Therefore it may be possible that the internal psychological representation of different line lengths used in both Rouder et al. (2004) and Dodds et al. (submitted) varied on more than one dimension.

In order to examine this theory using the same stimuli employed by Dodds et al. (submitted), we use Multidimensional Scaling (MDS) methods to examine the structure of similarity ratings generated using these stimuli. MDS refers to a broadly used range of statistical techniques, designed to allow the examination of relationships between objects of interest. Given a matrix of proximity data, MDS uncovers a spatial arrangement of objects in a manner that best reconstructs the original proximity data. For example, given a matrix of data with the distances between *n* cities, MDS analysis would present a spatial 'map' that would arrange the cities in the most likely location, given the distances provided by the data. Because we use subjective "similarity ratings", rather than actual measured distances, we employ non-metric MDS, which does not assume a linear mapping between similarity ratings and distances.

Typically, MDS is employed after one has already assumed the number of dimensions on which the stimuli might vary. In the current experiment however, we use MDS to determine the number of dimensions that best describe Dodds et al.'s (submitted) stimuli.

Method

Participants

The 27 participants, recruited from an introductory psychology course at the University of Newcastle, Australia, took part in exchange for course credit.

Stimuli

Stimuli were 16 lines varying in length (Figure 1). See Table 1 for pixel lengths. Lines were 11 pixels in width and were black, presented on a white background. Stimuli were log spaced, and were separated by a distance substantially greater than the Weber fraction for length (2%; Laming, 1986; Teghtsoonian, 1971).

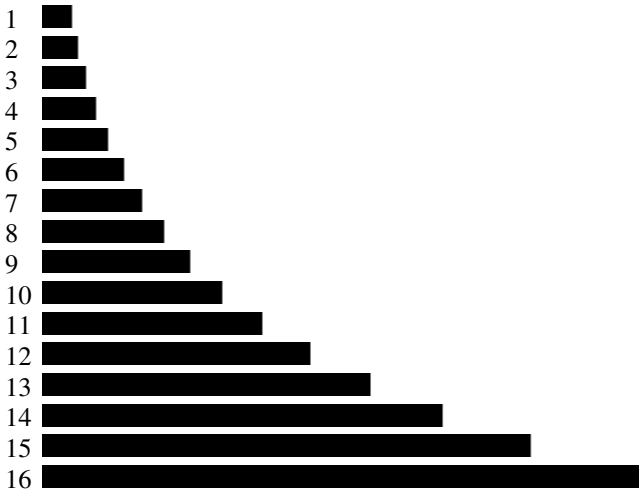


Figure 1. Unidimensional stimuli (line lengths) used in the Experiment. On any single trial, two of these stimuli were presented consecutively. All possible pairs of stimuli, including identical stimuli, were presented twice during the Experiment.

Table 1. Pixel lengths of the 16 lines used as stimuli

Pixel Lengths							
15	18	22	27	33	41	50	61
74	90	110	134	164	200	244	298

Procedure

Participants were instructed to rate the similarity of two stimuli that appeared on a computer monitor, on a scale of 1 to 100. On each trial, a single line would appear on the screen for 1 sec, followed by another line for 1 sec. The position of each line was jittered randomly on every presentation. After the two stimuli had been removed from the screen, a slider panel appeared at the bottom of the screen, allowing the participant to move a scrolling bar along a scale of 1 to 100 (where 1 = *dissimilar* and 100 = *similar*). Every possible pair of stimuli from the set, including identical pairs were presented twice. This resulted in 8 blocks of 64 trials, or a total of 512 trials (i.e., where n=16 stimuli and r=2 replications, number of trials = rn^2). A mandatory 30 sec break was taken between each block.

Each participant was given five practice trials at the beginning of the experiment, where they were asked to complete an identical task to the one above, with the exception that the stimuli were circles varying in diameter. The purpose of the practice trials was only to familiarize the participant with the response method. Different stimuli were used to prevent additional exposure to experimental stimuli.

Results

The main objective of our analysis is to determine whether the stimuli used by Dodds et al. (submitted) are represented internally by one or multiple dimensions. Initial descriptive analysis suggested that the data were consistent with a one-dimensional explanation: Figure 2 shows the average similarity ratings across participants, plotted as function of stimulus magnitude for each stimulus in the rating pair. Note that identical stimuli are rated as very similar (along the central diagonal), and rated similarity decreases monotonically with the rank-distance between the stimuli (at the left and right corners).

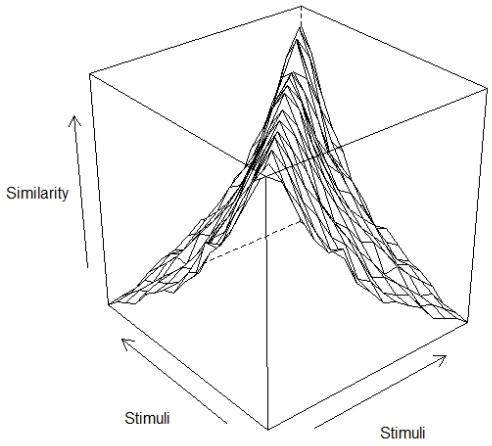


Figure 2. 3D structure of similarity ratings of all 16 stimuli.

Although Figure 2 indicates that the similarity ratings are consistent with a 1D psychological representation, they could nevertheless hide very subtle effects in the data, or large effects for individuals that average out in the group. In order to test this, we calculated non-metric MDS analyses for individual data. Each participant's data were transformed into a single symmetric dissimilarity matrix by subtracting the average similarity rating for each pair of items from 100 and averaging across reversed presentations (e.g., stimulus pair #1-#7 with stimulus pair #7-#1). This matrix was submitted for MDS analyses using both 1D and 2D representations for the data.

Deciding which of the 1D and 2D MDS analyses provides the best account of the data is not trivial. Various ad hoc methods have been used, including examining a goodness of fit measure, or examining the spatial arrangement the points in proximity plots. We applied both methods to our data. In MDS, goodness of fit between the reconstructed and observed dissimilarity matrices is typically measured by sum-squared error, which is called the *stress* value. Smaller stress indicates a better fit; however the MDS models are *nested* meaning that stress must always decrease as more dimensions are included. This means that stress must always be smaller for the 2D than the 1D model. Statistical tests on the magnitude of decrease in stress are not easily constructed, because the key properties of non-metric MDS make it difficult to assume a distributional model for the data. Figure 3 graphs the average stress value, across participants, for MDS fits with dimensions from 1 to 10 (a *scree plot*).

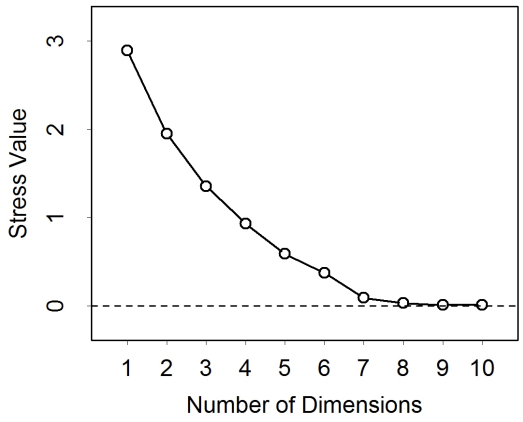


Figure 3. Scree plot showing the decrease in stress value as the number of dimensions increase.

Some authors recommend determining the number of dimensions from a scree plot by finding its "elbow"; a sharp drop in stress value, followed by a relatively flat continuation. Such a pattern could suggest that the latter dimensions fail to provide sufficiently better fit to warrant adding more dimensions to the model. Unfortunately, this method fails to provide any insight into the number of dimensions that best describe the stimuli, as there is no obvious elbow in the scree plot. This is a common problem (e.g., Grau & Nelson, 1988; Lee, 2001). In addition, the use of such methods has been criticized as placing unreasonable emphasis on a numerical measurement. Such methods to determine dimensionality are often used to the exclusion of other, more meaningful aspects of analysis, such as simply the interpretability of results (Shepard, 1974).

A more appropriate method to determine whether a two dimensional model provides a sensible description of the stimuli might be to examine the spatial relationship between objects in the purported 2D psychological space. This can be investigated with a "proximity plot", where each of the points provided in the similarity matrix are physically arranged in a manner that best satisfies the distances (or similarities) provided in the original data. Figure 4 shows two examples of these proximity plots, for two participants, from MDS analyses with two dimensions.

The philosophy of using MDS to recover internal structure relies on the assumption that, if the psychological representation of the stimuli was truly two dimensional, these 2D MDS proximity plots should reconstruct the internal representation. Because of the nature of the models under consideration (e.g. of categorization and absolute identification), this internal representation should have some relatively smooth and systematic shape. On the contrary, if the internal representation of the stimuli is truly one dimensional, these 2D MDS proximity plots should

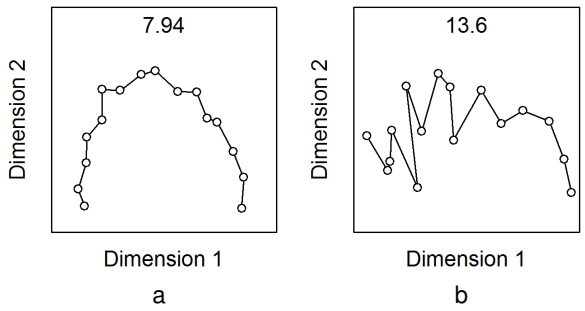


Figure 4. Two proximity plots of individual fits of a two dimensional model. Each of these graphs is the resulting proximity plot from a single participant in the Experiment. Each point represents a single stimulus in 2D space. Lines connect adjacent stimuli in the set. The value at the top of each graph is the stress value, a goodness of fit measure.

illustrate the 1D structure (a straight line) possibly along with some meaningless noise.

However, these interpretations of the proximity graphs are only appropriate when examining the results of metric MDS analyses (using true, quantitative distances). In the current case, where non-metric MDS analyses must be used, patterns that may normally suggest a two dimensional internal representation, might actually arise from data that are truly *one dimensional*. This problem stems from the monotone transformations allowed by non-metric MDS, between the observed similarity data and the internal psychological distances (as noted originally by Shepard, 1974). Since non-metric MDS analyses only preserve the rank order of the similarity ratings, leaving the exact similarity values to vary in systematic ways that best suit the data, there is considerable flexibility in the spatial arrangements that might arise from a single underlying dimension. Therefore both Figure 4a and Figure 4b could be construed as evidence favouring a single underlying dimension. Whilst the two proximity plots demonstrate distinctly different patterns, both provide evidence to suggest that our stimuli vary on only a single dimension.

Even though smooth C- or U-shaped proximity plots are *consistent* with one dimensional internal representations, they are also consistent with two dimensional internal representations – that is, truly C- or U-shaped underlying structures. We attempt to resolve this ambiguity using a simulation study comparing MDS outputs from 1D and 2D fits to truly 1D data, in the presence of noise. These simulations provide a metric for interpreting the stress values from our fits to data.

Simulation Study

We investigated this problem of dimensionality with a simulation study. We generated synthetic data from a similarity matrix that was truly one dimensional (the rated distance between each stimulus was a linear function of their ranked difference in the set). We scaled this generating similarity matrix to be as similar to the observed data as possible; we used 16 stimuli, with maximum and minimum similarity ratings of 95.91 and 6.88, respectively. Similarity between stimuli *i* and *j* could then be set as:

$$sim_{max} - (sim_{max} - sim_{min})*(abs(i-j)/15)$$

From this true similarity matrix, we generated synthetic data sets that matched the characteristics of the real data. Noise was added to the matrix using a normal distribution with standard deviation 12.18, and sampled similarity values outside [0,100] were truncated. These settings resulted in synthetic similarity matrices that were nearly identical to the human data, on average, for the range and variance of similarities, and also for the variance of similarity values across participants, conditioned on each stimulus pair.

We generated 1000 such matrices, and fit each with MDS using both 1D and 2D settings. The lower panel of Figure 5 shows the difference in stress values between these two fits for each simulated data set (negative values indicate a better fit for 2D than 1D).

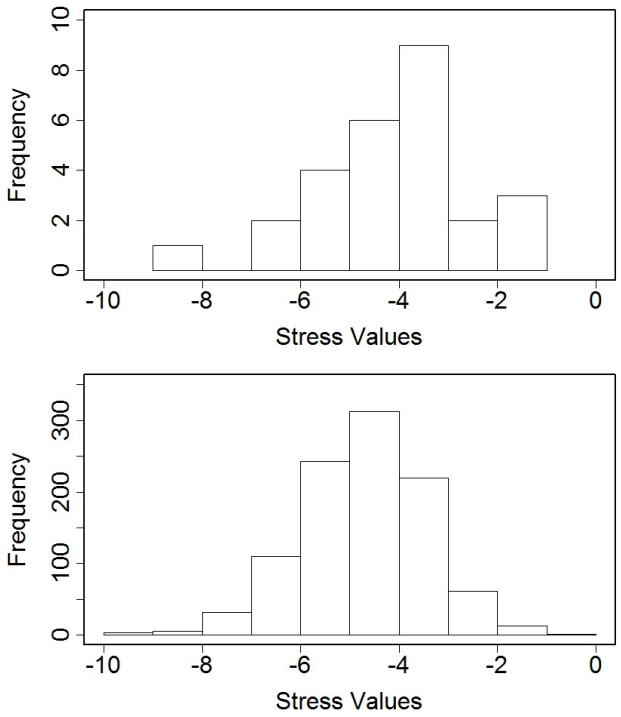


Figure 5. Difference in stress values for between 2D and 1D fits of the original data (top panel) and the true 1D data (bottom panel)

The upper panel of Figure 5 shows the difference between 2D and 1D stress values for the fits to our human data. The important thing to take from these graphs is that the decrease in stress generated by moving from a 1D to a 2D fit is about the same for our human data as it is for our synthetic data. Since the synthetic data were generated by a truly 1D process, this means that the stress values calculated for our human data are entirely consistent with a 1D

account. This provides further support to the evidence provided by the MDS analysis of our own data – that our stimuli may vary on only a single dimension.

Discussion

The purpose of the current experiment was test line-length stimuli commonly used in AI and always assumed to be unidimensional (e.g., Dodds et al., submitted; Rouder et al., 2004; Lacouture & Marley, 1995; Lacouture, 1997). Dodds et al. found that contrary to previous research, their participants were able to substantially improve their performance at the task when given significant practice. Although the stimuli used in their experiment varied on only one physical dimension, the results were more reminiscent of experiments using multiple dimensions, where it is more common to find substantial improvement with practice.

Although the stimuli used in Dodds et al. (submitted) varied on only one physical dimension, it is possible that they may vary on multiple psychological dimensions. In order to examine how many psychological dimensions underpin these stimuli, we used two methods; 1) using MDS techniques we examined similarity data taken using these same stimuli and 2) compared the structure of our data to simulated one dimensional data. MDS proximity graphs suggested that the stimuli may vary on a single dimension, and our simulation study provided further support for this, showing that these fits could be consistent with a one dimensional data generating process, when noise is added.

When examining individual proximity graphs taken from MDS analysis assuming two dimensions, a C (or U) shaped pattern often emerged, which is commonly assumed to provide evidence towards a 2D solution (Shepard, 1974). While this may be appropriate for a metric MDS analysis, the monotonic transformations unique to *non-metric* MDS allow some flexibility in the position of the objects in the final proximity graph. Despite this difference required in interpretation of metric vs. non-metric proximity graphs, it is possible that the two types of proximity graphs generated by our data (Figure 4) were genuinely representative of one vs. multiple dimensions, and that the action of specifying the number of dimensions to examine, forces the model to fit, sporadically producing evidence for and against a two dimensional solution. In support of a one dimensional solution however, our simulated data demonstrate a similar structure to our original similarity data, suggesting that the stimuli used in Dodds et al. (submitted) vary on only a single dimension.

Therefore it appears that the interpretation of MDS output for the number of underlying dimensions in the data is difficult. While we were able to gather evidence using a variety of techniques to suggest that our data were consistent with a single dimension, MDS could not provide a definitive answer. Lee (2001) showed that it is possible to reliably determine dimensionality from MDS analysis, but only when the determination is between larger numbers of dimensions. Like us, he found much poorer reliability when the choice was between lower numbers of dimensions. Hence, the task of choosing between a low number of dimensions remains very subjective, and users should take care not be misled by "overfitting", where a complex model imitates data from a simpler underlying data generating process. Furthermore, in the case of determining dimensionality, one should take care not to focus solely on quantitative results such as the stress value, but also take into consideration the pattern of data in the original similarity matrix (such as in Figure 2) or even simply the interpretability of results (Shepard, 1974).

Both the MDS analysis of the similarity data for Dodds et al.'s (submitted) lines of varying length and our simulation study were consistent with a 1D psychological representation. This finding makes it less likely that the substantial improvement with practice observed by Rouder et al. (2004) and Dodds et al. (submitted) in absolute identification of line lengths was due to participants learning to take advantage of a multi-dimensional psychological representation. This finding may also extend to the other stimuli that Dodds et al. employed. Similar learning effects to that of lines varying in length suggest that modality, or specifically, the number of dimensions that stimuli vary within, cannot be the sole cause of the improvement in performance. Hence, investigation of alternative explanations for the improvement they observed seems warranted.

References

Dodds, P., Donkin, C., Brown, S. D., Heathcote, A. *Practice Effects in absolute identification*. Manuscript submitted for publication

Eriksen, C. W., & Hake, H. W. (1955). Multidimensional stimulus differences and accuracy of discrimination. *Journal of Experimental Psychology, 50*(3), 153-160.

Garner, W. R. (1953). An informational analysis of absolute judgments of loudness. *Journal of Experimental Psychology, 46*(5), 373-380.

Grau, J. W., & Nelson, D. G. K. (1988). The distinction between integral and separable dimensions: evidence for the integrality of pitch and loudness. *Journal of Experimental Psychology: General, 117*(4), 347-370.

Hartman, E. B. (1954). The influence of practice and pitch distance between tones on the absolute identification of pitch. *The American Journal of Psychology, 67*(1), 1-14.

Lacouture, Y. (1997). Bow, range, and sequential effects in absolute identification: A response-time analysis. Psychological Research, 60, 121-133.

Lacouture, Y., & Marley, A. A. J. (1995). A mapping model of bow effects in absolute identification. Journal of Mathematical Psychology, 39, 383-395.

Lacouture, Y., Li, S., & Marley, A. A. J. (1998). The roles of stimulus and response set size in the identification and categorisation of unidimensional stimuli. *Australian Journal of Psychology, 50*(3), 165-174.

Laming, D. (1986). Sensory Analysis. London: Academic Press.

Lee, M. D. (2001). Determining the dimensionality of multidimensional scaling representations for cognitive modeling. *Journal of Mathematical Psychology, 45*, 149-166.

MacLeod, D. I. A. (2003). New dimensions in color perception. *Trends in Cognitive Sciences, 7*(3), 97-99.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits in our capacity for processing information. *Psychological Review, 63*(2), 81-97

Pollack, I. (1952). The information of elementary auditory displays. *The Journal of the Acoustical Society of America, 24*(6), 745-749.

Rouder, J. N., Morey, R. D., Cowan, N., & Pfaltz, M. (2004). Learning in a unidimensional absolute identification task. *Psychonomic Bulletin & Review, 11*(5), 938-944.

Shepard, R. N. (1974). Representation of structure in similarity data: Problems and prospects. *Psychometrika*, *39*(4), 373-421.

Teghtsoonian, R. (1971). On the exponents in Stevens' Law and the constant in Ekman's Law. Psychological Review, 78(1), 71-80.

Weber, D. L., Green, D. M., & Luce, R. D. (1977). Effects of practice and distribution of auditory signals on absolute identification. *Perception and Psychophysics, 22*(3), 223-231

Mind Reading by Machine Learning: A Doubly Bayesian Method for Inferring Mental Representations

Ferenc Huszár (fh277@eng.cam.ac.uk)
Computational and Biological Learning Lab, Dept Engineering, U Cambridge, Cambridge CB2 1PZ, UK

Uta Noppeney (uta.noppeney@tuebingen.mpi.de)
Max Planck Institute for Biological Cybernetics, Spemannstrasse 41, Tübingen 72076, Germany

Máté Lengyel (m.lengyel@eng.cam.ac.uk)
Computational and Biological Learning Lab, Dept Engineering, U Cambridge, Cambridge CB2 1PZ, UK

Abstract

A central challenge in cognitive science is to measure and quantify the mental representations humans develop – in other words, to 'read' subject's minds. In order to eliminate potential biases in reporting mental contents due to verbal elaboration, subjects' responses in experiments are often limited to binary decisions or discrete choices that do not require conscious reflection upon their mental contents. However, it is unclear what such impoverished data can tell us about the potential richness and dynamics of subjects' mental representations. To address this problem, we used ideal observer models that formalise choice behaviour as (quasi-)Bayes-optimal, given subjects' representations in long-term memory, acquired through prior learning, and the stimuli currently available to them. Bayesian inversion of such ideal observer models allowed us to infer subjects' mental representation from their choice behaviour in a variety of psychophysical tasks. The inferred mental representations also allowed us to predict future choices of subjects with reasonable accuracy, even in tasks that were different from those in which the representations were estimated. These results demonstrate a significant potential in standard binary decision tasks to recover detailed information about subjects' mental representations.

Introduction

Cognitive science studies the mental representations humans (and other animals) develop and the way these representations are used to perform particular tasks. A central challenge is to measure and quantify such mental representations experimentally – in other words, to 'read' subjects' minds. A classical approach to this is to ask subjects directly to report their mental contents verbally. Unfortunately, this procedure is prone to introducing biases arising from verbal processing, and from the educational and cultural backgrounds of subjects (Ericsson & Simon, 1980; Russo et al., 1989). In order to eliminate these biases, an alternative approach is to limit subjects' responses to simple binary decisions or discrete choices that do not require conscious reflection upon their mental contents. However, it is unclear what such impoverished data can tell us about the potential richness and dynamics of subjects' mental contents.

A powerful computational framework formalises the goal of learning as estimating the probability distribution or density of stimuli (Hinton & Sejnowski, 1986; Dayan & Abbott, 2001). This motivates many formal theories of human learning and cognition to model the relevant mental content of a subject either implicitly or explicitly as a 'subjective' distribution over possible stimuli (Chater et al., 2006; Sanborn & Griffiths, 2008). In this study we adopted this representation, and our goal was to estimate subjects' subjective distributions solely from their responses in simple binary decision tasks without making any assumptions about the process by which those subjective distributions were acquired, i. e. learning.

Ideal observer models are widely used for explaining human behaviour in various psychophysics tasks (Geisler, 2003). They formalise (quasi-)optimal decision making strategies given the information available to subjects and their background knowledge about the task, which in our case includes their subjective distributions. While previous studies mostly used ideal observer models to determine optimal performance in particular tasks to which human performance could then be compared, we treat them as stochastic models formalising the link between subjective distributions (the unobserved variable), and test stimuli and responses (the observed variables). Our main observation is that such models can be used to provide the likelihood in a Bayesian statistical analysis of subjective distributions, thus enabling one to infer mental contents from task responses in a principled way.

We term our approach *doubly Bayesian*, as we assume that subjects act as quasi-ideal observers, which entails Bayesian inference on their side; and then we use these ideal-observer models in a Bayesian framework to infer a posterior distribution of possible subjective distributions.

Inferring subjective distributions

The graphical model (Koller & Friedman, 2009) in Fig. 1A describes our model of a subject's behaviour in a session of a psychophysics experiment. We assume that the subject entertains a subjective distribution $\mathcal{P}$ over possible stimuli, and that this distribution does not change over the analysed session. In trial i of the experiment, the subject is presented a set of test stimuli $\mathcal{S}_i$ and gives a response r_i. The value of r_i depends on the current stimuli $\mathcal{S}_i$, the subjective distribution $\mathcal{P}$, and 'link' parameters $\Theta_{\mathcal{O}}$ describing further aspects of observation and decision making, such as attention, perceptual noise, etc.

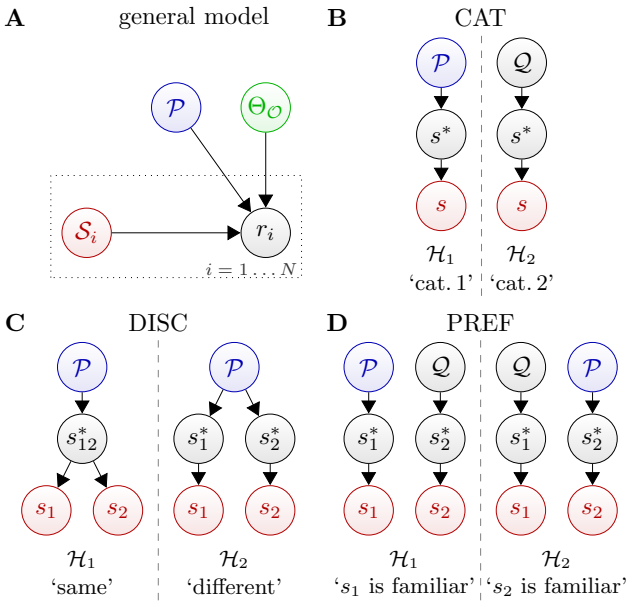


Figure 1: **A,** Graphical model describing the subject's behaviour in an experimental session of N consecutive trials. We assume that the subject represents a subjective distribution, $\mathcal{P}$, over possible stimuli, and in trial i their response r_i depends on the currently observed test stimuli $\mathcal{S}_i$, their subjective distribution, $\mathcal{P}$, and some other parameters influencing their responding, $\Theta_\mathcal{O}$. Our goal is to infer $\mathcal{P}$ and $\Theta_\mathcal{O}$ from the observed sequence of stimulus-response pairs. **B-D,** Generative models for the three task types (CAT, DISC, and PREF, see descriptions under 'Experimental data sets'). Subjects assume that their observations, s, are perceptual noise-corrupted versions of the 'true' stimuli, s^*, sampled by the experimenter from a distribution that is the same as their subjective distribution, $\mathcal{P}$, or an alternative distribution, $\mathcal{Q}$ (which is assumed to be uniform for tractability), depending on the particular hypothesis, $\mathcal{H}$.

In order to quantify the dependence between subjects' choices and their subjective distributions, response probabilities, $p(r_i|\mathcal{S}_i, \mathcal{P}, \Theta_\mathcal{O})$, were specified by quasi-ideal observer models. These models formalise subjects' choices as functions of the posterior probabilities of the two hypotheses corresponding to either response being correct.

Each hypothesis amounts to a different model of how stimuli might have been generated, and so the posterior over hypotheses is inferred by a Bayesian inversion of these generative models. Fig. 1B-D shows such generative models in three tasks considered later in this paper (for more detail, see the supplementary material[1]). Once posterior probabilities are available, the statistically optimal, although psychologically unrealistic, strategy would be to deterministically choose the response with the maximal posterior probability. As a more realistic model of human decision making we used a soft-max function (parametrised by $\Theta_\mathcal{O}$) of log posterior probabilities, that describes quasi-optimal decision making (Sanborn & Griffiths, 2008; Orbán et al., 2008).

Our goal is to estimate latent parameters $\mathcal{P}$ and $\Theta_\mathcal{O}$ from a series of stimulus-response pairs $\{\mathcal{S}_i, r_i\}_{i=1}^N$. As responses given in subsequent trials of the experimental session are assumed to be conditionally independent, the likelihood of latent parameters becomes

$$p(r_{1:N}|\mathcal{S}_{1:N}, \mathcal{P}, \Theta_\mathcal{O}) = \prod_{i=1}^{N} p(r_i|\mathcal{S}_i, \mathcal{P}, \Theta_\mathcal{O})$$

To allow for full Bayesian inference we specified prior distributions over the subjective distribution, $\mathcal{P}$, and link parameters, $\Theta_\mathcal{O}$. We chose to model subjective distributions as mixtures of Gaussians (MoG's). This parametric family of distributions is flexible enough to model complex subjective distributions in low dimensional feature spaces and allows for analytical computation of likelihood ratios in the binary tasks considered here. Importantly, this prior reflected no information about the distribution of stimuli with which subjects were trained (i. e. the distribution to which their subjective distributions could be expected to be close), except for the general domain of possible stimulus values. The MoG representation is not a vital part of our general approach: other representations and priors may be more appropriate in some cases.

Given the prior and the likelihood defined above, we inferred a posterior over $\mathcal{P}$ and $\Theta_\mathcal{O}$ via Bayes' rule:

$$p(\mathcal{P}, \Theta_\mathcal{O}|r_{1:N}, \mathcal{S}_{1:N}) \propto p(\mathcal{P})p(\Theta_\mathcal{O}) \prod_{i=1}^{N} p(r_i|\mathcal{S}_i, \mathcal{P}, \Theta_\mathcal{O})$$

Unfortunately, calculating the posterior exactly is intractable, so we have to resort to approximate inference techniques, for which we implemented a Hamiltonian Monte Carlo algorithm (Neal, 2010).

Experimental data sets

Two experimental data sets were analysed, each collected using simple visual stimuli and requiring binary responses from subjects.

One-dimensional feature space The first set of experimental data was the fish categorisation data set collected by Sanborn & Griffiths (2008). In this experiment, the stimuli used were schematic images of fish of fixed length and variable height, i. e. the relevant feature space was one dimensional (see Fig. 2A). Subjects were trained (with corrective feed-back) in a supervised binary categorisation task (CAT) to distinguish fish drawn from a Gaussian training distribution from fish drawn from a uniform distribution. The mean and variance of the

[1] available online at mlg.eng.cam.ac.uk/ferenc/mindreading

training distribution was varied across four conditions (Fig. 2B, red curves), with 9-11 subjects in each condition. Subjects also performed a stimulus preference task (PREF), in which they had to choose the stimulus which seemed more likely to be drawn from the training distribution. In this task, no feedback was provided. The experiment started with an initial block of 120 CAT trials (to train subjects) followed by four blocks of PREF task alternating with four blocks of CAT task, each block consisting of 60 trials. In a final block of CAT trials, no feedback was provided. For our analysis we neglected the initial training session. We used the next 180 PREF and 180 CAT trials to infer subjects' subjective distributions and reserved the last 60 PREF and 60 CAT trials for cross-validation.

Three-dimensional feature space The stimuli in the second experiment were trapezoids with three features varying systematically: colour (gray-scale), size, and shape (ratio of parallel sides), each parametrised by continuous values between 0 and 1 (Fig. 3A). This experiment involved one-back discrimination (DISC) and stimulus preference tasks (PREF). During DISC trials, which also served to train subjects on a particular distribution of stimuli, subjects were presented with one stimulus per trial, and had to judge (without feedback) whether it was the same or different than the one presented in the previous trial. In actuality, 10% of stimuli were exact repetitions of stimuli presented in the previous trial, the rest was sampled independently from the training distribution. Two different training distributions were used in the two conditions (Fig. 3B, left panels), with six subjects in each condition. During PREF trials subjects had to choose (without feed-back) the stimulus which appeared to be more familiar based on the stimuli they had seen during training. The experiment started with 300 DISC training trials, followed by 100 PREF trials and another 200 DISC trials. In our analysis we neglected the first 100 DISC trials, used 300 DISC and 50 PREF trials to infer subjective distribution and preserved 100 DISC and 50 PREF trials for cross-validation.

Results

Inferring subjective distributions After extensively validating out method on synthetic datasets (supplementary material[1]), we inferred human subjective distributions from the two experiments described earlier. Fig. 2B shows results on the experiment with a one-dimensional feature space. The inferred subjective distributions reflected qualitative aspects of the distributions of stimuli on which subjects were trained in different conditions. This match between inferred and training distributions became especially clear in the categorisation task.

Fig. 3B shows results on the experiment with a three-dimensional feature space. These results suggest that subjects did not learn the training distribution in this experiment very well (see also below), although some resemblance between training and inferred subjective distributions were recovered for a few subjects (e. g. subjects 1, 4, 11 and 12). The subjective distributions inferred for the same subject in the two different tasks also revealed some consistency of these distributions.

Figs. 2-3B illustrate the primary goal of our study: to provide a method for inferring and visualising subjective distributions based on subjects' responding in psychophysics experiments. However, as subjective distributions cannot be observed or measured directly, there is no obvious way to assess the degree to which these inferences are 'correct'. One possibility, pursued above, is to compare the inferred distributions to the distributions subjects were trained on (assuming that subjects are approximately ideal learners and decision makers). While a match between the inferred subjective distribution and the training distribution (Fig. 2B) can be taken as indicative of valid inferences, a lack of match (Fig. 3B) is harder to interpret. In particular, one cannot distinguish between the algorithm giving incorrect results or subjects behaving sub-optimally (because of a failure to learn, or a failure to use learned information to direct choices). Therefore we sought to establish the quality of the inferences of our method in a more reliable way.

Predicting human behaviour A standard way to assess the quality of a statistical model of a data set is to test its predictive performance in cross-validation: infer its parameters (hidden variables) based on a subset of the data, and measure how well it predicts the held-out part of the data set. Our method is readily amenable to this cross-validation approach since it defines an explicit statistical model for predicting subjects' responses based on the stimuli they see (Fig. 1A). Making such predictions is not only important for validation purposes in the context of the present study, but may also be relevant in its own right in applications in which e. g. customer choices need to be predicted based on their previous choices.

For cross-validation, we inferred subjects' subjective distributions and link parameters from the first blocks of trials of a task and based on the inferred model predicted their responses in the final block of trials in the same task (Fig. 4, *double Bayes*). Ideally, subjective distributions are independent of the type of task subjects are performing, and hence one would even expect to be able to infer the subjective distribution from behaviour in one task and, based on that, predict choices in an other task. Thus, we also performed a stronger cross-validation test in which we measured such across-task predictive performance (Fig. 4, *double Bayes-CT*).

Subjects' responding is inherently stochastic, therefore the absolute predictive performance of our model is not particularly informative in itself. In order to establish some relevant baseline performance, we implemented al-

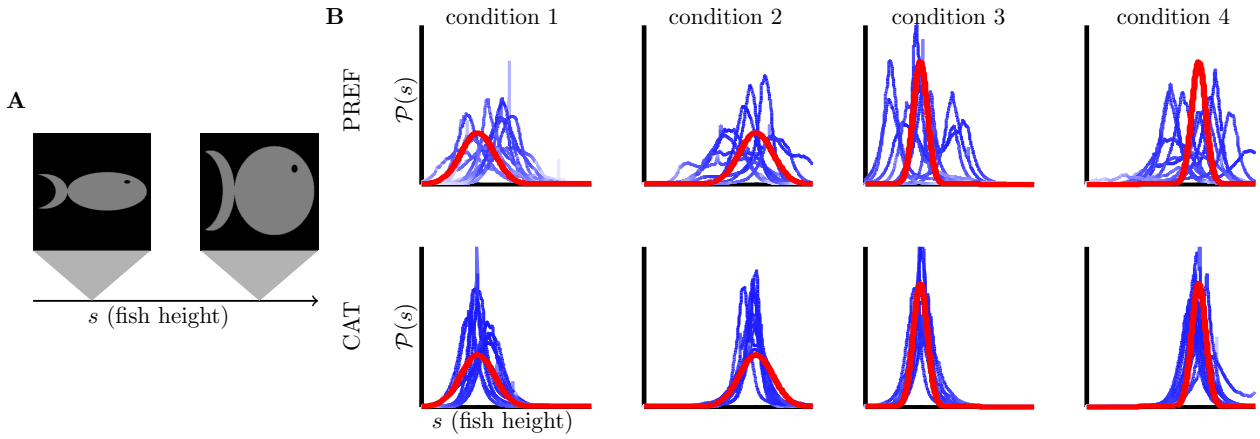


Figure 2: **A,** One dimensional stimuli used in the first set of experiments. **B,** Subjective distributions in a one-dimensional feature space. *Rows* correspond to task types, *columns* correspond to experimental conditions using training distributions with different means (1 & 3 vs. 2 & 4) or variances (1 & 2 vs. 3 & 4). *Red lines* show training distributions, *blue lines* show the posterior mean subjective distribution of each subject. *Shading of blue lines* indicates point-wise marginal posterior uncertainty: lighter means higher uncertainty (s.e.m. divided by the mean).

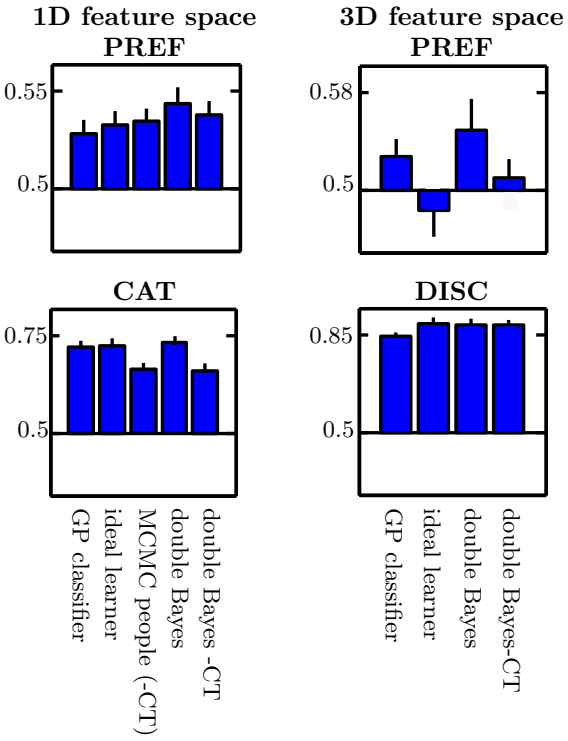


Figure 4: Predicting human responses by alternative methods. *Bars* show across-subject averages (± s.e.m.) of probabilities of correct predictions. In the PREF tasks our method *double Bayes* significantly outperformed both the *GP classifier* and the *ideal learner* in both experiments and also *MCMC people* in the 1D experiment ($p < 0.05$). In the CAT task, the *MCMC people* method was used for across-task predictions (*-CT*).

ternative models for predicting subjects' responses. Since the task of predicting responses based on the stimuli that subjects see is formally equivalent to a binary classification task (see supplementary material[1]), we implemented a Gaussian process classifier (Fig. 4, *GP classifier*) (Rasmussen & Williams, 2006). The GP classifier is a particularly powerful algorithm applicable for such classification tasks, but it is also a black-box model in the sense that it has no explicit notion of subjective distributions. Therefore, it provides an interesting baseline by giving about the best predictive performance that can be achieved without modelling subjects' mental representations.

As an alternative method that did have an explicit notion of subjective distributions, we implemented an 'ideal learner' version of our model, which has the training distribution as its subjective distribution for all subjects, but its link parameters (parametrising stochasticity in decision making) are still fitted to each subject's data individually (Fig. 4, *ideal learner*). This model controls for the importance of individual differences in the inferred subjective distributions in our method, and also tests the validity of the assumption that subjects act as ideal learners in these experiments.

Finally, we also implemented as an alternative method a previously published algorithm ('MCMC with people') to infer subjective distributions (Sanborn & Griffiths, 2008). Although this algorithm can only be applied to specifically designed stimulus preference experiments, one of our data sets includes data from such an experiment, so we tested the performance of the algorithm on that data set by performing both within-task and across-task cross-validation (Fig. 4, *MCMC people (-CT)*).

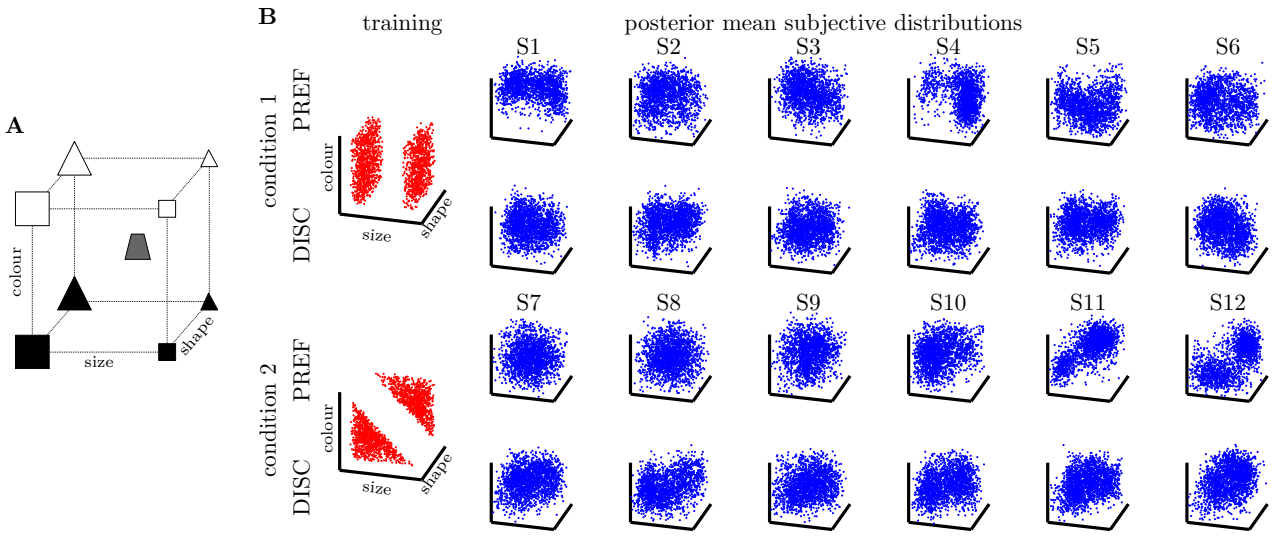


Figure 3: **A,** Stimuli used in the second experiment had three continuous features, size, colour and shape. **B,** Subjective distributions in the three-dimensional feature space. *Left, in red:* training distributions; *right, in blue:* posterior mean subjective distributions *(blue)* for each subject individually *(columns, S1-S12). Rows* correspond to different conditions, using different training distributions, and different task types to test subjects' subjective distributions. In the discrimination task, subjects 7 and 8 responded irrespective of stimuli.

Fig. 4 shows the predictive performances of these methods. The absolute difficulty of predicting responses greatly varied across tasks, the discrimination task and the categorisation tasks being considerably easier than the preference task, but the relative performances of the different methods showed consistent patterns across the different experiments and tasks. When comparing within-task predictive performances, our method was the best, or among the best, in all tasks. Notably, it outperformed the 'MCMC with people' method even in the case when that method was applicable at all.

In most cases, the three subjective distribution-based methods (*double Bayes*, *ideal learner*, *MCMC people*) outperformed the GP classifier, showing that making predictions about subjects' responding benefits substantially from representing and inferring subjective distributions explicitly. This is especially true in across-task cross-validation which is impossible with a GP classifier in lack of any parameter that could be shared between tasks. Yet, in two out of four cases our method had higher accuracy even when comparing its across-task performance against within-task performance of the GP classifier.

Methods using subject-specific subjective distributions (*double Bayes*, *MCMC people*) also performed at least as well as the ideal learner, confirming the validity of the individual differences in subjective distributions these methods inferred, and showing that the poor match found between training and subjective distributions in some cases (Fig. 3B) were real and not a failure of our algorithm to recover 'better looking' subjective distributions.

Discussion

We have presented a new computational method for inferring subjects' mental representation of stimuli from their responses on simple binary decision tasks. Since Bayesian inference was intractable, we implemented a Hamiltonian Markov chain Monte Carlo method for numerical analysis, which we have extensively validated and tested on real-world data sets. We found that the method was able to recover subjective distributions of humans when they were trained on stimuli with known structure and to predict future responses better than other model-based and 'black-box' methods. We have also shown that – using our method – information gained in one type of task could be transferred and applied to predict responses in another task which we take as further evidence for the veridicality and task-invariance of the mental representations we inferred. These results also offer a way to reconcile cognitivism with behaviourism inasmuch as they demonstrate that even when the only goal is to predict responses from stimuli, modelling mental representations explicitly is quantifiably useful.

There is a long tradition in experimental psychology and cognitive science to use simple statistics of task performance, such as percent correct rates, or reaction times, as indices of learning (Gallistel, 1993). These 'naïve' methods, even in their statistically most sophisticated forms (Gallistel et al., 2004; Kakade & Dayan, 2002; Smith et al., 2005; Preminger et al., 2009; Katkov et al., 2007), boil down to estimating a single (time-dependent) scalar measure of memory strength, i. e. the degree of

match between subjects' mental representations and that required by the experimenter (which would presumably allow subjects to perform perfectly). However, by reducing mental contents to simple memory strength measures, these methods fail to provide a detailed picture of structured mental representations which is what we aimed to achieve in the present study.

While structured probabilistic models of cognition have become mainstream more recently (Chater et al., 2006), they have mostly been used in normative theories to account for general, qualitative principles of learning (e. g. patterns of generalisation) rather than to quantitatively estimate individual subjects' mental representations in specific experiments. Our approach is complementary to these as it makes no assumptions about learning itself.

Our work is most closely related to more recent work by Paninski (2006) and Sanborn & Griffiths (2008) who both used ideal observer models to infer subjective distributions. In the paper by Paninski (2006) continuous decision tasks were considered (in which subjects' responses are analogue rather than discrete), and the method developed there does not seem to generalise well to the binary decision tasks considered here (and used extensively in experimental psychology), because the linear programming problem that needs to be solved becomes seriously under-constrained. Our analysis of the preference task is taken from previous work by Sanborn & Griffiths (2008), but they used it to construct a particular kind of stimulus preference task in which subjects' responding itself implements a Markov chain Monte Carlo sampler. This is a most elegant idea, but does not translate in any obvious way to other task types, or indeed to preference tasks which were not constructed according to their particular rules. Our method does not suffer from these limitations because of its doubly Bayesian nature: once ideal observer behaviour based on Bayesian analysis is formalised, the method offers an automatic and principled way of inferring subjective distributions.

A natural way to extend our work in the future will be to consider dynamical priors over subjective distributions in order to track their temporal evolution, inferring changes brought about by learning. The machine learning literature offers powerful tools for carrying out inference in such dynamical models.

Acknowledgements

We thank P. Dayan, C. Rasmussen, and A. Sanborn for useful discussions and K. Jucicaite and S. Ölschläger for help with acquiring psychophysics data. We are grateful for A. Sanborn and T. Griffiths for providing access to their experimental data. This project was supported by the Wellcome Trust (FH, ML), the Gatsby Charitable Foundation (FH), and the Max Planck Society (UN).

References

Chater, N., Tenenbaum, J. B., & Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. *Trends Cogn Sci*, *10*(7), 287–291.

Dayan, P., & Abbott, L. F. (2001). *Theoretical neuroscience.* Cambridge, MA: The MIT Press.

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychol Rev*, *87*, 215–251.

Gallistel, C. R. (1993). *The organization of learning.* Cambridge, MA: The MIT Press.

Gallistel, C. R., Fairhurst, S., & Balsam, P. (2004). The learning curve: Implications of a quantitative analysis. *Proc Natl Acad Sci USA*, *101*(36), 13124–13131.

Geisler, W. S. (2003). Ideal observer analysis. In L. Chalupa & J. Werner (Eds.), *The visual neurosciences.* The MIT Press.

Hinton, G. E., & Sejnowski, T. J. (1986). Learning and relearning in Boltzmann machines. In *Parallel distributed processing.* MIT Press.

Kakade, S., & Dayan, P. (2002). Acquisition and extinction in autoshaping. *Psych Rev*, *109*(33), 533–544.

Katkov, M., Tsodyks, M., & Sagi, D. (2007). Inverse modeling of human contrast response. *Vision Res*, *47*(22), 2855–67.

Koller, D., & Friedman, N. (2009). *Probabilistic graphical models: Principles and techniques.* MIT Press.

Neal, R. M. (2010). MCMC using hamiltonian dynamics. In S. Brooks *et al.* (Ed.), *Handbook of Markov chain Monte Carlo.* Chapman & Hall / CRC Press.

Orbán, G., Fiser, J., Aslin, R. N., & Lengyel, M. (2008). Bayesian learning of visual chunks by human observers. *Proc Natl Acad Sci USA*, *105*(7).

Paninski, L. (2006). Nonparametric inference of prior probabilities from Bayes-optimal behavior. In Y. Weiss *et al.* (Ed.), *NIPS 18.* MIT Press.

Preminger, S., Blumenfeld, B., Sagi, D., & Tsodyks, M. (2009). Mapping dynamic memories of gradually changing objects. *Proc Natl Acad Sci USA*, *106*(13).

Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian processes for machine learning.* The MIT Press.

Russo, J. E., Johnson, E. J., & Stephens, D. L. (1989). The validity of verbal protocols. *Memory and Cognition*, *17*(6), 759–769.

Sanborn, A., & Griffiths, T. (2008). Markov chain Monte Carlo with people. In J. Platt *et al.* (Ed.), *NIPS 20.*

Smith, A. C., Stefani, M. R., Moghaddam, B., & Brown, E. N. (2005). Analysis and design of behavioral experiments to characterize population learning. *J Neurophysiol*, *93*(3), 1776–1792.

The Development and Assessment of Cross-Sectioning Ability in Young Children

Kristin R. Ratliff (krratliff@uchicago.edu)
Department of Psychology, 5848 S. University Ave.
Chicago, IL 60637 USA

Charlotte R. McGinnis (cmcginnis@uchicago.edu)
Department of Psychology, 5848 S. University Ave.
Chicago, IL 60637 USA

Susan C. Levine (s-levine@uchicago.edu)
Department of Psychology, 5848 S. University Ave.
Chicago, IL 60637 USA

Abstract

In two experiments, we investigated the development of cross-sectioning ability using either three-dimensional (3D) or two-dimensional (2D) stimuli. Three to 9 year old children visualized cross-sections of either real 3D geometric shapes (Experiment 1) or 2D photographs of the shapes (Experiment 2). Performance on the 3D task was also analyzed to determine to what extent cross-sectioning ability is related to performance on more widely used spatial tasks including mental rotation and the water-level task. We found that performance on the cross-sectioning and mental rotation tasks were significantly correlated, and the 2D and 3D tasks were both successful in assessing cross-sectioning ability in young children. As expected, we also found a significant increase in cross-sectioning performance across age groups.

Key Words: spatial development; cross-section; education.

Introduction

Spatial ability is important for success across a variety of academic subjects, particularly in the science, technology, engineering, and mathematics (STEM) disciplines. Spatial ability is also related to choosing technological and science-related careers and predicts the choice of math and science as college majors (Shea, Lubinski, & Benbow, 2001), as suggested by the fact that individual differences across verbal, quantitative, and spatial abilities at age 13 were predictive of educational and vocational group membership 20 years later. However, despite the importance of spatial ability, spatial training is not a regular part of school curricula and there are no national or state standards for spatial intelligence. Consequently, many students have difficulty with spatial tasks and lack the opportunity to improve their spatial reasoning skills.

Spatial ability can refer to a wide range of skills, some of which focus on how individuals perceive and act on objects in space while others focus on how individuals orient and navigate within space. One category of spatial ability of particular interest is spatial visualization, or the ability to understand, mentally encode, and manipulate 3D forms (Carroll, 1993; Hegarty & Waller, 2004).

Cross-sectioning, also referred to as "penetrative" thinking (Kali & Orion, 1996) is a particular spatial visualization skill that involves inferring a 2D representation of a 3D structure, and vice versa (Cohen & Hegarty, 2007). This imaginary slicing of a 3D object to a 2D plane is an essential skill for many of the sciences, ranging from anatomical cross-sections in biology and neuroscience to cross-sections of landforms in geology (Cohen & Hegarty, 2008). Conversely, in order to understand what is under a microscope, students must also be able to mentally reconstruct a 3D object from a given 2D image.

Spatial visualization requires performing multistep manipulations of spatial representations, such as a paper-folding task that requires the ability to work quickly, rotate figures, and keep track of multiple operations. This is thought to be distinct from other spatial tasks such as spatial perception and mental rotation (Linn & Petersen, 1985). For example, the water-level task, which requires subjects to draw a horizontal line in a tilted bottle where they believe the water level would be, is categorized as a spatial perception task because it requires determining spatial relationships with respect to a given frame of reference. Linn and Petersen define mental rotation as a Gestalt-like analogue process that involves accurately mentally rotating a 2D or 3D figure. However, the development of cross-sectioning ability has not been compared to these other measures of spatial ability, in part because of a lack of adequate measures and the unknown age at which this ability emerges. Thus, we do not know whether it is more related to spatial visualization, spatial perception, or mental rotation.

Cross-Sectioning Ability of Young Children

There is disagreement about the age at which children are able to reason about cross-sections of 3D objects. In contrast to Piaget and Inhelder's (1956) view that children should have achieved mastery of geometric sectioning by 12 years old, many studies have found that spatial visualization involving cross-sections does not develop until the teenage years. For example, most students do not accurately predict the appearance of a geometric plane intersecting a simple cone or sphere until sometime between the ages of 11 and 15 (Russell-Gebbett 1984, 1985), while even students in grades 8, 10, and 12 have difficulty accurately choosing a cross-section of simple geometric line drawings (Boe, 1968; Davis, 1973).

The difficulty older children and adolescents have with these assessments may be in the presentation of the test

items themselves rather than a lack of underlying cognitive processes supporting cross-sectioning skills. Assessments involving cross-sections are often based on 2D diagrams and complex figures that represent 3D objects. Although these have been shown to successfully measure spatial visualizations of cross-sections among adults (e.g., Santa Barbara Solids Test, Cohen & Hegarty, 2007) and adolescents (e.g., Mental Cutting Test "Schnitte," Quaiser-Pohl, 2003), these assessments are too advanced for use with younger children.

One factor impacting success when measuring other spatial skills in young children has been using more familiar, salient, and concrete stimuli. For example, tasks have used pictures of humans and animals to successfully measure mental rotation ability in young children (Quaiser-Pohl, 2003; Wiedenbauer & Janesen-Osmann, 2008). Similarly, by using basic 2D geometric shapes, Levine and colleagues (1999) were able to successfully assess mental transformation ability in preschool children.

In the present study, we created a new method for assessing cross-sectioning skills in young children by using brightly colored foam shapes as the stimuli. We contrasted this 3D method with a 2D method using photographs of the actual shapes. Thus, we aimed to successfully measure children's cross-sectioning skills to determine a) how cross-sectioning skills develop between the ages of 3 and 9 years, b) the association between cross-sectioning skills and other spatial reasoning tasks, and c) how the method of assessment impacts performance.

We expected that using salient and familiar objects, such as the foam shapes, would make the task accessible for preschool to early elementary children. We also predicted that there would be an increase in spatial ability across the age range. We explored the relation between cross sectioning and two other measures of spatial ability that would engage similar yet categorically distinct spatial operations (see Linn & Petersen, 1985): mental rotation and the water-level task. We predicted that cross-sectioning would correlate with these more established measures of spatial reasoning but that the strength of the correlations would vary depending on the spatial processes required. Specifically, given that cross-sectioning involves manipulating mental images and possibly rotation, we predicted that performance on the cross-sectioning tasks would be significantly correlated with performance on a mental rotation task. However, as the water-level task has been shown to demonstrate distinctly different spatial operations from spatial visualization tasks (see meta-analysis by Linn & Petersen, 1985), we expected this task might not correlate as strongly with cross-sectioning as mental rotation.

Additionally, spatial ability has been shown to develop even through early adolescence (Vasta & Liben, 1996). Therefore, we expected to find an effect of age such that performance on cross-sectioning improves over time. Interestingly, using 3D objects adds complexity, which some researchers have shown negatively impacts performance on mental rotation tasks (e.g. Rosser, 1980). Since cross-sectioning ability involves the interface of 2D and 3D representations we might expect that this task would be more difficult for young children because of the increased complexity of the stimuli. Consequently, in Experiment 2 we contrasted the presentation of the stimuli between actual three-dimensional geometric shapes (3D) and photographs of the real shapes (2D). Our expectation was that young children would be more successful when they were presented with problems involving cross-sections of actual 3D objects than when these same cross-sectioning problems were presented as photographs on a computer screen.

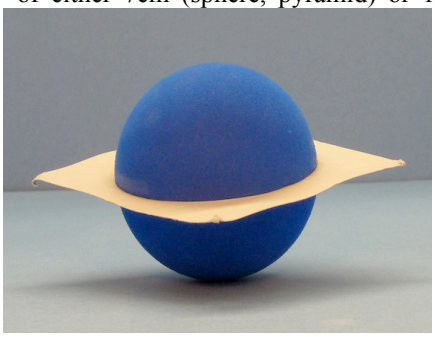
Experiment 1

In this study, we developed an assessment of cross-sectioning ability to determine if this task was suitable for young children. Experiment 1 used real objects (e.g., geometric foam shapes) and compared performance on the 3D cross-sectioning task to performance on two other standard measures of spatial ability (mental rotation and the water-level task) to determine the trajectory of cross-sectioning development during the early elementary years in relation to other spatial skills.

Method

Participants. Fifty-one elementary students (17 boys, 34 girls) ranging in age from 5 years 0 months to 9 years 0 months (M=7.35 years, SD=1.16), were recruited from the Chicago area. Participants were compensated $10 for their time and travel and were also given a t-shirt for participating. We constructed four age groups from the data collected: 5 year olds (n=8, M=5.58 years, SD=0.32), 6 year olds (n=11, M=6.43 years, SD=0.31), 7 year olds (n=16, M=7.32 years, SD=0.32), and 8 to 9 year olds (n=8, M=8.69 years, SD=0.27).

Apparatus and materials. Stimuli for the cross-sectioning task consisted of six solid foam geometric shapes. Each solid had a base edge length or diameter of 7cm and a height of either 7cm (sphere, pyramid) or 14cm (cone, cylinder, rectangular prism, triangular prism).





(a) (b) (c) (d)

Figure 1. The sample cross-sectioning item, showing a sphere bisected by an intersecting plane. Participants were asked to choose among four options to identify the resulting cross-section.

To create the 12 test items, each shape was bisected with an intersecting plane of sturdy, gray stock paper (see Figure 1).

Design and Procedures.

Participants were tested on three spatial reasoning tasks in a set order: cross-sectioning, mental rotation, and the water-level task. All participants were tested individually in a laboratory at the University of Chicago.

Participants first completed a cross-sectioning task adapted from Piaget and Inhelder (1956) and Boe (1968). In this task, children were presented with one sample item and twelve test items composed of familiar, colorful 3D foam shapes that were cross-sectioned by an intersecting plane. A sphere was used as the sample item and five different shapes (a cone, cylinder, pyramid, rectangular prism, and triangular prism) were used to construct the test items. Each shape was used twice, with one test item depicting a cut along the horizontal axis and the other along the vertical axis. An additional horizontal cut cone and vertical cut rectangular prism of different colors were used to complete the 12 test items.

The experimenter first showed each solid foam shape to the child, rotating the object so they could view all sides of the object, and then identified it (e.g. "This is a cylinder") in order to familiarize the child with the stimuli. Next, the experimenter showed the sample item. A sphere was bisected with a piece of sturdy stock paper between the two halves demonstrating where and how the object had been cut (see Figure 1). The participant was told to imagine what the inside of the sphere would look like if we were to pull it apart at the cut point. The experimenter stated that the cut side would be flat and may make a shape that is different than the shape of the whole object that we see. The participants were then asked to point to the resulting cross-section shape from an array of four 2D shape choices. All shapes were cut either symmetrically (i.e., along the center) or asymmetrically along the longitudinal or horizontal axis. The stimuli were shown to the participants from one of two orientations: half of the participants viewed the objects so that they were looking at the intersecting plane from an approximate 90 degree angle (e.g., the plane of the paper was parallel to the ground such that the edges were more visible, see Figure 1) and the other half viewed the intersecting plane face-on (e.g., the paper was perpendicular to the ground such that the surface around was fully visible).

Participants also completed the Primary Mental Abilities (PMA) spatial relations test (Thurstone & Thurstone, 1963), where they were asked to pick from a four-choice array the shape that would make a square if it were put together with the target shape. Participants were instructed that this task was like a puzzle where shapes could be rotated to fit but could not be flipped over.

Then, participants performed the water-level task adapted from Piaget and Inhelder (1956). In this task, participants were presented with a line drawing of a half-full bottle of water. They were then given four pictures of empty bottles tilted upright, to the left or to the right, 30°, 45°, and 60° from horizontal. Participants were asked to draw the resulting water line if the bottle was half full and was tilted.

The cross-sectioning and mental rotation tasks were scored such that each participant received an accuracy score (e.g., mean proportion correct), whereas the water-level task measured the angular disparity from 0 degrees (e.g., mean error). First, we present our analysis on the cross-sectioning task as a new method to assess spatial visualization skill in young children. Then, we compare performance across the three tests.

Results

Table 1 presents the mean performance for each spatial task by age group. Note, for the cross-sectioning task, there was no significant difference based on the viewing orientation of the test items (90 degrees: n=26, M=.69, SD=.21; facing: n=25, M=.76, SD=.18). Thus, we collapsed across this factor in our analysis.

For the cross-sectioning task, a 2x4 ANOVA (gender by age group) revealed a significant effect of age group, $F(3,43)=7.98$, $p<.001$, but no effect of gender, $F(1,43)=0.07$, $p=.79$, and no interaction, $F(3,43)=1.79$, $p=.16$. Specifically, cross-sectioning performance significantly improved at each age compared to the last, except for between ages 6 and 7 years (all p's $< .02$, using Bonferroni adjustment). This reflects a developmental trend, such that participants improve with age, starting at 5 years old, and attain very good performance on this task (88% correct) around 8 years of age.

We conducted an item analysis to determine item difficulty. The most difficult test items across the entire sample were generally those where the cross-section resulted in a different shape from the whole, which we call

Table 1. Mean performance on spatial tasks by age group.

Task	Age	*M*	*SE*	*N*
Cross-Sectioning (proportion correct)	5 yrs	0.53	0.07	8
	6 yrs	0.70	0.05	11
	7 yrs	0.72	0.04	16
	8-9 yrs	0.88	0.03	16
	Total	0.74	0.03	51
Mental Rotation (proportion correct)	5 yrs	0.34	0.06	8
	6 yrs	0.54	0.06	11
	7 yrs	0.50	0.03	16
	8-9 yrs	0.68	0.04	16
	Total	0.54	0.03	51
Water-Level (angular disparity)	5 yrs	172	5.5	8
	6 yrs	178	12	11
	7 yrs	167	15	16
	8-9 yrs	113	40	16
	Total	163	9.5	51

Table 2. Mean accuracy (standard error) of cross-section task items across the entire sample (*N*=51).

Item type	Item Shape	Cross-section	% correct
Congruent	Triangular Prism	Triangle	94
	Cylinder	Circle	92
	Rectangular Prism (2)	Rectangle	84
	Cone (2)	Triangle	78
	Pyramid	Triangle	59
		Total	81.4 (6.3)
Incongruent	Rectangular Prism	Square	84
	Cylinder	Rectangle	80
	Triangular Prism	Rectangle	49
	Cone	Circle	39
	Pyramid	Square	31
		Mean Total	56.6 (10.8)

incongruent cross-sections (Table 2). For example, performance on the pyramid cut horizontally to reveal a square cross-section was the most difficult item (31% accuracy rate overall). Conversely, shapes with congruent cuts were much easier for children to grasp (e.g., the pyramid cut vertically to reveal a triangle, 59% answered correctly). Overall, children scored significantly higher on the congruent items than the non-congruent items, $t(50)=3.72$, $p=.001$.

For the mental rotation and water-level tasks, we conducted an analysis of variance (ANOVA) with the mean proportion correct (or the mean deviation score in the case of the water level test) by gender and age group. We found a significant age group effect for the mental rotation task, $F(3,43)=8.26$, $p<.001$, and the water-level task, $F(3,43)=3.10$, $p = .04$.

The cross-sectioning and mental rotation tasks were significantly correlated across the entire sample, $r(49)=.47$, $p=.001$. However, when controlling for age (in months), a multiple regression model revealed that mental rotation score was not a significant predictor for cross-sectioning performance, $\beta = .23$, $p = .09$, $R^2 = .40$, $\Delta R^2 = .04$. Further, when collapsing across age groups, we found no significant correlation between cross-sectioning and water-level task performance, $r(49)=.20$, $p=.23$.

In summary, children successfully completed the cross-sectioning task, suggesting that children as young as 5 years old are capable of performing basic cross-sections given the appropriate stimuli. Further we found an increase in performance with age. Difficulty of test items generally represented two categories: congruent items were easier in that the cross-section resulted in a similar shape to the overall object, whereas incongruent items were harder due to the cross-section resulting in a different shape than the overall object. Positive correlations between the mental rotation and cross-sectioning tasks were present across the 5 to 8 year age range. However, when controlling for age, mental rotation was not a significant predictor of cross-sectioning performance, which suggests these tasks are not measuring *identical* skills but rather related spatial skills, particularly in children younger than 8 years old. Further, there was no significant correlation between performance on the cross-sectioning and water-level tasks. Thus, cross-sectioning ability is somewhat independent of both spatial perception and mental rotation. We are currently examining cross-sectioning performance in relation to another spatial visualization task using a paper folding task that is appropriate for young children.

Experiment 2

In order to examine the effects of presentation on cross-sectioning ability, we contrasted performance using 3D and 2D stimuli. Hence, half of the participants saw real three-dimensional geometric shapes (3D), while the other half of participants viewed 2D photographs of the shapes on a computer screen. We also investigated whether preschool children as young as 3 years old would succeed at the task. If successful, the cross-sectioning assessment would be useful in a variety of settings outside of a laboratory, as well as with a greater age range.

Method

Participants. Sixty-nine elementary students (37 boys, 32 girls) ranging in age from 3 years 1 month to 9 years 3 months (*M*=5.82 years, *SD*=1.66) were recruited as previously described and randomly assigned to two groups: 3D stimuli (19 boys, 16 girls; age, *M*=5.72 years, *SD*=1.67, range 3yrs1mos to 8yrs11mos) and 2D stimuli (19 boys, 16 girls; age, *M*=5.47 years, *SD*=1.74, range 3yrs1mos to 9yrs3mos).

Apparatus, Design and Procedures. All participants received the same familiarization and testing procedure for the cross-sectioning task only as described in Experiment 1. However, participants were randomly assigned to either the 3D or 2D stimuli group, which determined the type of objects they saw during the cross-sectioning test (either real 3D foam shapes used in Experiment 1 or 2D photographs of the shapes, see Figure 1). Additionally, as viewing orientation did not impact performance in Experiment 1, all stimuli were held by the experimenter (for 3D) or presented on a computer screen (for 2D) such that the intersecting plane was at an approximate 90 degree angle to the child.

Results

Table 3 presents the mean proportion correct across age groups within the 2D and 3D conditions. A 2x2x6 ANOVA (condition by gender by age group) revealed a significant interaction between condition and age group, $F(5,46)=5.16$,

3D	*M*	*SE*	*n*	2D	*M*	*SE*	*n*	*p*
3 years	0.47	0.05	6	3 years	0.50	0.05	5	.70
4 years	0.45	0.08	8	4 years	0.54	0.05	6	.35
5 years	0.65	0.06	6	5 years	0.52	0.02	7	**.043**
6 years	0.72	0.03	6	6 years	0.48	0.08	7	**.016**
7 years	0.80	0.03	5	7 years	0.63	0.05	4	**.023**
8-9 years	0.75	0.10	4	8-9 years	0.92	0.04	5	.12
Total	0.62	0.03	35	Total	0.58	0.03	34	.20

Table 3. Mean proportion correct (standard error) on the cross-sectioning task for each condition (2D vs. 3D) by age group.

p=.001. Specifically, there was a benefit for those in the 3D condition in the 5, 6 and 7 year age groups (see Table 3), but not in 3-4 year olds or 8-9 year olds. There was also a main effect of age, $F(5,46)=8.42$, $p<.001$, such that performance significantly increased with age overall, from early (4 years) to late (8 years), $p<.01$ Bonferroni (3 yrs=48% correct, 4 yrs=49%, 5 yrs=58%, 6 yrs=59%, 7 yrs=72%, and 8-9yrs=84%). This replicates the developmental trend found in Experiment 1 that children improve basic understanding of cross-sections over time, and extends the earliest age tested successfully to 3 years old.

Additionally, we compared performance on individual test items between the 3D and 2D versions of the task (Table 4). Again, the most difficult test items were incongruent cross-sections (e.g, the pyramid cut horizontally to reveal a square cross-section), while shapes with congruent cuts were much easier for children to grasp (e.g., the pyramid cut vertically to reveal a triangle). A 2x2 ANOVA (condition by item type), revealed significantly higher performance for congruent compared to incongruent items, $F(1, 67)=107.21$, $p<.001$, but no effect of condition (p=.34) or interaction between condition and item type, (p=.17).

Table 4. Mean percent correct (standard error) of cross-section task items for 3D (n=34) and 2D (n=35) stimuli.

Item type	Item Shape	Cross-section	3D	2D
Congruent	Triangular Prism	Triangle	66	65
	Cylinder	Circle	77	76
	Rectangular Prism (2)	Rectangle	86 67	87
	Cone (2)	Triangle	74 77	71
	Pyramid	Triangle	89	88
		Total	76.7 (4.3)	77.7 (3.5)
Incongruent	Rectangular Prism	Square	77	53
	Cylinder	Rectangle	60	53
	Triangular Prism	Rectangle	43	18
	Cone	Circle	34	26
	Pyramid	Square	11	9
		Total	41.1 (4.3)	31.2 (3.3)

Discussion

In the present experiments we found that young children do reason about cross-sections and this ability can be assessed successfully using a task that involves either three-dimensional simple geometric shapes or two-dimensional photographs of simple geometric shapes. This ability develops over time, such that basic understanding of cross-sections improves from 3 to 8 years of age. Further, cross-sectioning ability is independent from other spatial skills, but is related to mental rotation more so than the water-level task.

According to Linn & Petersen (1985), spatial visualization tasks require maintaining mental representations and performing multistep manipulations on them. Thus, cross-sectioning skills, which involve such complex mental operations and rotations, would likely be categorized as a spatial visualization skill. As such, we found that cross-sectioning is distinct from mental rotation, as assessed by the Thurstone mental rotation test, and spatial perception, as assessed by the water-level task. Although some studies have not successfully measured cross-sectioning ability in children younger than adolescence, we found that a basic understanding of cross-sections emerges as young as preschool.

Further, it is possible to assess cross-sectioning ability in children using either real objects or photographs of real objects. Although using 3D objects provided a significant advantage for children between 5 and 7 years of age, performance across the 2D group was still above chance levels. The absence of a 3D advantage in the youngest children (3 and 4 year olds) may be due to the use of a simple shape matching strategy for both 3D objects and 2D

photographs. For example, the triangular shaped cone matches the isosceles triangle. However, we included at least one foil item that had a similar shape as the correct answer to prevent this strategy always leading to the answer. In contrast, the absence of a 3D advantage in 8-9 year olds may reflect the development of the ability to think about 2D images as 3D objects. When asking about cross-sections of any stimuli presented in 2D, one must successfully infer the object as 3D prior to performing mental operations. However, if children are unable to accurately process 2D information into 3D structures, they are already starting at a disadvantage. Further study is needed to examine possible strategy differences in children with lower cross-sectioning ability compared to those with more advanced skills. Also, we aim to assess various methods for improving cross-sectioning ability across the preschool to early elementary ages.

References

Boe, B.L. (1968). A study of the ability of secondary school pupils to perceive the plane sections of selected solid figures. *Mathematics Teacher, 61*, 415-421.

Carroll, J.B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge University Press Cambridge; New York.

Cohen, C. A., & Hegarty, M. (2007). Individual differences in use of external visualisations to perform an internal visualisation task. *Applied Cognitive Psychology, 21,* 701-711.

Cohen, C. A., & Hegarty, M. (2008). Spatial visualization training using interactive animation. Conference on Research and Training in Spatial Intelligence. Sponsored by National Science Foundation, Evanston, IL. June 13-15, 2008.

Davis, E.J. (1973). A study of the ability of school pupils to perceive and identify the plane sections of selected solid figures. *Journal for Research in Mathematics Education, 4,* 132-140.

Hegarty, M, & Waller, D. (2004). A dissociation between mental rotation and perspective-taking spatial abilities. *Intelligence, 32*, 175-191.

Kali, Y., & Orion, N. (1996). Spatial abilities of high-school students in the perception of geologic structures. *Journal of Research in Science Teaching. 33,* 369-391.

Levine, S.C., Huttenlocher, J., Taylor, A. & Langrock, A. (1999). Early sex differences in spatial ability. *Developmental Psychology*, *35*, 940-949.

Linn, M.C. & Petersen, A.C. (1985). Emergence and characterization of sex differences in spatial ability: A meta-analysis. *Child Development, 56*, 1479-1498.

Piaget, J., & Inhelder, B. (1956). *The child's conception of space* (F.J. Langdon & J.L. Lunzer, Trans.). London: Routledge & Kegan Paul.

Quaiser-Pohl, C. (2003). The mental cutting test "schnitte" and the picture rotation test-- two new measures to assess spatial ability. *International Journal of Testing, 3*, 219-231.

Rosser, R. (1980). Acquisition of spatial concepts in relation to age and sex (Final Report on Grant No. NIE-6-79-0091 from the National Institute of Education, Department of Education). Tucson: University of Arizona.

Russell-Gebbett, J. (1984). Pupils' perceptions of three-dimensional structures in biology lessons. *Journal of Biological Education, 18,* 220–226.

Russell-Gebbett, J. (1985). Skills and strategies: Pupils' approaches to three-dimensional problems in biology. *Journal of Biological Education, 19,* 293–297.

Shea, Daniel L., Lubinski, David, Benbow, Camilla P. (2001). Importance of assessing spatial ability in intellectually talented young adolescents: A 20-year longitudinal study. *Journal of Educational Psychology, 93*, 604-614.

Thurstone, L.L., & Thurston, T.G. (1963). Manual, *PMA*, Primary Mental Abilities, for Grades. 2-4. Chicago, IL: Science Research Associates.

Vasta, R. & Liben, L.S. (1996). The water-level task: An intriguing puzzle. *Current Directions in Psychological Science, 5*, 171-177

Wiedenbauer, G, & Jansen-Osmann, P. (2008). Manual Training of Mental Rotation in Children. *Learning and Instruction, 18*(1), 30-41.

What can Information Extraction from Scenes and Causal Systems Tell us about Learning from Text and Pictures?

Alexander Eitel (a.eitel@iwm-kmrc.de)
Knowledge Media Research Center, Konrad-Adenauer Str. 40,
72072 Tübingen, Germany

Katharina Scheiter (k.scheiter@iwm-kmrc.de)
Knowledge Media Research Center, Konrad-Adenauer Str. 40,
72072 Tübingen, Germany

Anne Schüler (a.schueler@iwm-kmrc.de)
Knowledge Media Research Center, Konrad-Adenauer Str. 40,
72072 Tübingen, Germany

Abstract

Numerous studies have shown that the gist in photorealistic pictures of scenes is extracted after very short presentation times. So far, the investigation of gist extraction has been limited to pictures of scenes. The present study investigated whether the gist in pictures of causal systems, which are typically used as instructional material, is extracted as fast as the gist in pictures of scenes, and whether more than just the gist is rapidly extracted from a causal system (i.e., information concerning its details and functioning). Schematic and photorealistic pictures of scenes and causal systems were presented to subjects (N = 24) at different presentation times. Results showed that the gist in causal systems is extracted as fast as in scenes, and that an initial understanding of the functioning of schematic causal systems is also rapidly acquired. Results are discussed in the light of their implications for learning from text and pictures.

Keywords: Gist; Scene Perception; Causal Systems; Learning from Text and Pictures

Learning from Text and Pictures

In multimedia research it is a well known finding that learning from text and pictures leads to better retention and recall than learning from text alone (Levie & Lentz, 1982). Moreover, when students have to learn about causal systems, they are better able to apply their knowledge to produce creative solutions to problem-solving questions after learning from text and pictures than after learning from text alone (Mayer, 1989). Accordingly, learning from both text and pictures leads to higher comprehension than learning from text alone. Despite the fact that the beneficial effects of pictures for learning are well established in the research literature, far less is known concerning how the pictorial information is processed during learning.

An exception is an early study by Hegarty and Just (1993), in which comprehension was assessed via questions about the kinematics of different pulley systems. Students were better able to infer motion in the pulley system when previously learning from text and pictures than when learning from text alone or picture alone. Eye tracking data furthermore revealed that during learning from text and pictures of pulley systems, subjects first processed text information, and then switched to the picture in order to integrate the information from both sources.

Unlike Hegarty and Just (1993), various studies showed that when subjects were confronted with information from text and pictures, they often initially looked at the picture for a short time before they started to read the text. This pattern of processing has been shown for advertisements (Rayner et al., 2001), comics (Carroll, Young, & Guertin, 1991), real-world scenes (Underwood, Jebbett, & Roberts, 2004), and biology schoolbooks (Mak, 2008). In a study by Stone and Glock (1981), in which subjects had to learn how to build a cardboard loading cart from text and schematic pictures, subjects first looked at the picture for 1000 to 2000 ms, before they started to read the text. According to the authors, subjects initially looked at the picture in order to get a first impression (i.e., gist) of what the material was about. However, it is yet unclear what role looking briefly at a picture prior to reading a text may play for understanding the presented content. At least this phenomenon has not been directly addressed in research on learning from text and pictures. However, there has been ample research in basic cognitive psychology about the extraction of information from briefly looking at pictures of scenes.

Extraction of Information from Scenes

In an early study of Biederman and colleagues (1974), subjects had to select one out of two labels, which they judged to better describe a picture of a jumbled vs. coherent scene. In coherent scenes, when the two labels were similar (e.g., “shopping plaza” vs. “busy road and stores”), accuracy of selecting the right label was at 100% for the majority of subjects after 300 ms of presentation. When the two labels were dissimilar, a ceiling effect in accuracy of selecting the right label occurred after only 100 ms. The authors concluded that information about the gist of a scene is already extracted after a single fixation, which enabled subjects to perform the task correctly. This is in line with

the findings from Henderson and Hollingworth (1999), who state that the average fixation duration during scene viewing is about 330 ms.

Similarly, Loftus, Nelson, and Kallman (1983) conducted a study in which subjects were asked to decide whether the picture of a scene had already been presented or not. Subjects were told to base their decision either on general properties of the picture or on detail information. When the decision was based on general properties of the picture, performance increased much less between 250, 500, and 1000 ms presentation time than when the decision was based on detail information. The authors concluded that most holistic information in scenes is extracted from the first fixation (about 330 ms; Henderson & Hollingworth, 1999) and subsequent fixations have the primary purpose of identifying relevant details.

Castelhano and Henderson (2008) also provided evidence for a rapid extraction of holistic information from pictures of scenes by presenting photos of scenes to subjects for a short time (25 – 250 ms) and later asking them whether a specific detail had been depicted in the scene. The detail in question was either consistent (e.g., fire hydrant) or inconsistent (e.g., tea set) with the gist of the scene (e.g., street scene) but was never actually present. Between 42 and 250 ms presentation time, subjects more often affirmed that the detail in question was present in the scene when the detail was consistent than when it was inconsistent with the gist of the scene. The authors concluded that a rapidly acquired (42 – 250 ms) scene gist was responsible for more affirmative responses to details consistent with scene gist by activating information about the scene's content and basic-level category.

To conclude, studies in basic cognitive psychology consistently demonstrate that information about the gist (e.g., general topic) in photos of scenes is extracted within the first fixation. Later fixations are presumably made to scan the scene for details.

Aims of the Current Study

The aim of the current study was to apply and compare findings from basic cognitive research on gist extraction from scenes to learning from text and pictures to better understand the role that pictures might play during the latter.

Unlike with scenes, there has yet not been much research about the extraction of information from instructional material. As mentioned before, in the study from Stone and Glock (1981), looking at the picture for 1000 to 2000 ms was interpreted as the time it took subjects to extract the gist. This is much longer than the time it takes subjects to extract the gist from scenes (< 250 ms). However, Stone and Glock interpreted the time subjects initially looked at the picture before reading the text as the time required to extract its gist. Subjects could also have extracted the gist within the first fixation (about 330 ms) as in scenes and looked at the picture up to 2000 ms only in order to scan it for details. Thus, it is still unclear when information about the gist and details is extracted in pictures of instructional material.

Information extraction in pictures of causal systems was investigated, since they are often used as instructional material in studies on learning from text and pictures (e.g., Hegarty & Just, 1993; Mayer, 1989). It was expected that once the gist of a causal system has been extracted, subjects would use the remaining time to understand the functioning of the depicted system. Hence, with longer presentation times knowledge about the functioning of the system should improve. In the aforementioned studies on learning about causal systems (e.g., Mayer, 1989), mostly schematic pictures of causal systems have been used (e.g., line drawings). On the other hand, gist extraction from scenes has been investigated by presenting photorealistic pictures to subjects (e.g., Castelhano & Henderson, 2008). In the present study, it was investigated whether these findings on gist extraction could be extended to schematic pictures of causal systems from studies on learning from text and pictures (e.g., Hegarty & Just, 1993). To overcome the confound that in previous research mostly photorealistic pictures of scenes and schematic pictures of causal systems were used, schematic and photorealistic depictions of both, scenes and causal systems were directly compared to each other in the current study. The degree of realism is considered to be a continuum with schematic line drawings on the one end and photos of natural objects on the other end. The less similar an illustration is to its real-world referent with respect to shape, details, color, and texture the more schematic it is. It can be expected that in general information will be extracted more easily from schematic depictions than from realistic ones, because the prior do contain fewer elements which can be recognized more easily due to better contrasts etc. However, effects of realism were not the focus of the study; rather this variable was solely introduced to bridge the gap between prototypical materials used in scene perception research (photorealistic scenes) and research on learning from text and pictures (schematic depictions of causal systems).

Hence, the current study addressed the question whether the gist in causal systems would be extracted as fast as the gist in scenes. Further, details were assumed to be better extracted at longer presentation times compared to shorter ones. Finally, it was investigated whether the functioning of causal systems would be understood and whether this depended on presentation time.

Method

Participants and Design

Twenty-four students (15 female, 9 male, average age: M = 23.83 years, SD = 3.50) from the University of Tuebingen, Germany, took part in the experiment for either payment or course credit. The experiment followed a 2 × 2 × 4 design, with *Type* (scene vs. causal system), *Realism* (schematic vs. realistic) and *Presentation Time* (150 vs. 600 vs. 2000 vs. 6000 ms) serving as within-subjects factors.

Materials and Procedure

The materials in the experiment comprised 80 pictures of scenes and 80 pictures of causal systems. In a pilot study, subjects had to rate the number of objects in each picture, and to categorize the pictures with respect to their degree of realism and type (scene vs. causal system); the rated number of objects was the same for realistic and schematic pictures, and only unambiguous illustrations were used in the study. A scene depicted an everyday situation. A causal system always had a certain purpose (e.g., pulling weight). It consisted of multiple components, where at least one component was influenced by another – hence, removing one component would have changed the functioning of the system. In the experiment, for both scenes and causal systems, half of the pictures were schematic, the other half realistic. This led to four different categories of pictures in the experiment (see Figure 1). Each picture appeared in the center of the computer screen and covered nearly the whole screen size. An experimental session consisted of 8 training trials and 160 experimental trials.

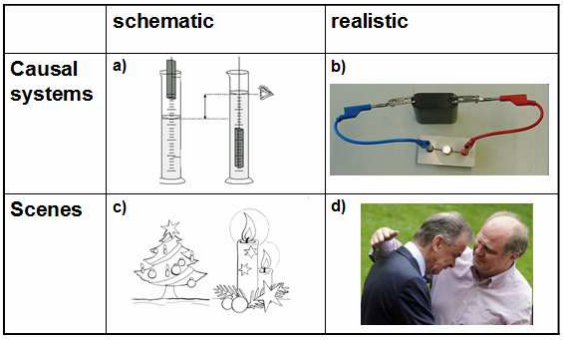


Figure 1: Categorization of pictures used in the experiment. Pictures could either depict a scene or a causal system and could be either schematic or realistic.

Each experimental trial started with the presentation of the word "ready?", which remained on the screen until a key was pressed. After pressing a key, the word "ready?" was replaced by the fixation cross, which was displayed for 800 ms. Then a picture (scene vs. causal system, schematic vs. realistic) appeared for either 150, 600, 2000 or 6000 ms, respectively, and was immediately masked afterwards. Both pictures and presentation times were presented in a randomized order. After each picture, a statement about the gist, then about details, and then about the functioning of the picture was presented and students were asked to respond to these statements (see Measures section for details). The statement concerning the functioning was presented only after pictures of causal systems. After responding to the last statement (detail or functioning), the trial was over and the word "ready?" reappeared, which marked the beginning of a new trial. An experimental trial for a single picture lasted about 15 seconds. The whole experimental session lasted approximately 45 minutes.

Measures

After viewing each picture, participants had to respond to either two or three statements about the picture depending on the experimental condition. All statements were in a two-alternative-forced-choice format, where students had to choose between a "yes" and a "no" response by pressing one of two keys on a keyboard. In half of the trials, "yes" was the correct response, in the other half of the trials "no" was correct. The first statement was about the gist of the picture. For instance, students were asked to decide whether a scene could be identified as "happy people" (see Figure 1d) or whether a causal system could be identified as "electric circuit" (see Figure 1b). Statements about the gist always consisted of only one to three words. In the second statement, participants had to judge whether specific details had been present in the scene (e.g., "presents are lying under the tree"; see Figure 1c) or in the causal system (e.g., "an eye is depicted"; see Figure 1a) just seen. Details were not relevant to either the meaning of the scene or the functioning of the causal system. Moreover, details were depicted in the periphery rather than in the center of the picture so that they were less likely to be seen within the first fixation. The third statement was presented only after pictures of causal systems, and was about the functioning of the depicted system. In order to be able to answer statements about the functioning correctly, inferences were required (e.g., "If the block is pulled out of the test tube, then liquids are at the same level in both test tubes"; see Figure 1a). It is important to note that statements concerning the functioning could be answered correctly only by relating multiple objects from the picture to each other; they could not be answered correctly solely based on prior knowledge that might have been activated once the causal system had been recognized correctly. The detail and functioning statements consisted of one sentence each.

As the main dependent variable, the percent correct was computed. Each correct response (both hits and correct rejections) was coded with 1, each incorrect response with 0. Multiplied by 100, percent correct was 100% at maximum and 50% at chance level and was computed separately for the three types of statements (gist, details, and functioning). Mean reaction times (RT) for responses to the different statements served as a second dependent variable in the experiment. Eye tracking data were assessed as well, but will not be reported here for space reasons.

Results

Overall, results revealed that there was no speed-accuracy trade-off, since there was no significant negative correlation between accuracy and RT (r = .24, p = .26). Thus, only accuracy to statements about the gist, details, and the functioning will be analyzed here.

Gist

T-tests revealed that both in scenes ($t(23) = 31.32$, $p < .001$) and in causal systems ($t(23) = 11.42$, $p < .001$), accuracy to

gist statements was above chance level at the shortest presentation time (150 ms), which speaks in favor of an early extraction of gist from both, scenes and causal systems (see Figure 2).

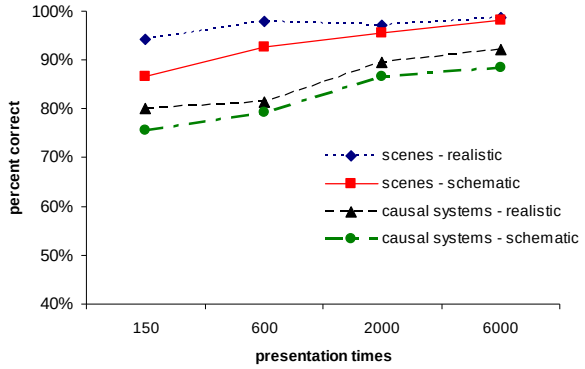


Figure 2: Accuracy to statements about the gist in schematic and realistic pictures of scenes and in schematic and realistic pictures of causal systems.

A 2 (*Type*: scenes vs. causal systems) × 2 (*Realism*: realistic vs. schematic) × 4 (*Presentation Time*: 150 vs. 600 vs. 2000 vs. 6000 ms) repeated-measures ANOVA was conducted to analyze accuracy for statements about gist. There was a significant main effect of *Type*, indicating that statements about the gist were answered more accurately ($F(1, 23) = 96.68$, $p < .001$) in scenes than in causal systems (see Figure 2), which is probably due to a higher difficulty in recognizing the general topic of causal systems. There were also significant main effects of *Realism* ($F(1, 23) = 8.41$, $p = .01$) and *Presentation Time* ($F(3, 69) = 17.51$, $p < .001$) meaning that gist extraction was better in realistic than in schematic pictures, and improved with longer presentation times. There were no interactions (all $p_s > .05$).

Details

T-tests revealed that both in scenes ($t(23) = 5.04$, $p < .001$) and in causal systems ($t(23) = 2.31$, $p = .03$), accuracy to statements about details was above chance level at 150 ms (see Figure 3).

A 2 (*Type*) × 2 (*Realism*) × 4 (*Presentation Time*) repeated-measures ANOVA revealed significant main effects of *Type* ($F(1, 23) = 25.63$, $p < .001$) and *Presentation Time* ($F(3, 69) = 12.46$, $p < .001$) on accuracy to detail statements. As expected, details were recognized more accurately at longer presentation times both in scenes and in causal systems. While there was no main effect of *Realism* ($F(1, 23) = 1.79$, $p = .19$) it interacted significantly with *Type* ($F(1, 23) = 13.08$, $p < .001$). Bonferroni tests showed that detail extraction was better in realistic than in schematic pictures of scenes ($p = .03$), whereas it tended to be worse in realistic than in schematic pictures of causal systems ($p = .065$). There were no further interactions (all $F_s < 1$).

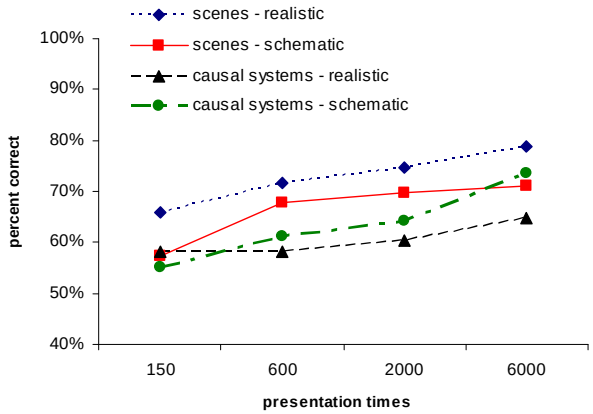


Figure 3: Accuracy to statements about details in schematic and realistic pictures of scenes and in schematic and realistic pictures of causal systems.

Functioning

T-tests revealed that accuracy to statements about the functioning of realistic pictures of causal systems was at chance level for 150, 600 and 2000 ms (all $p_s > .05$). Only at the longest presentation time of 6000 ms, accuracy was above chance level ($t(23) = 5.29$, $p < .001$). On the other hand, accuracy to statements about the functioning of schematic causal systems was already above chance level ($t(23) = 3.86$, $p = .001$) at 600 ms presentation time (see Figure 4). Only at the shortest presentation time of 150 ms, accuracy was at chance level ($p > .05$).

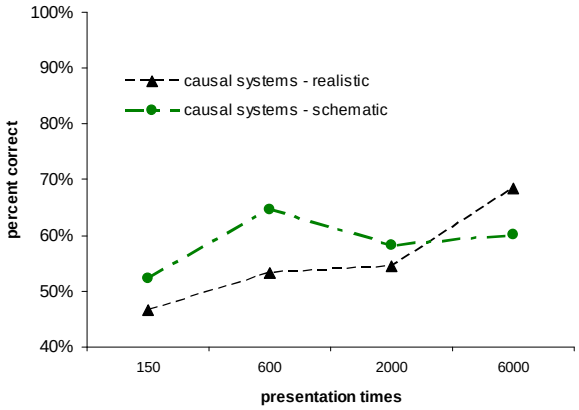


Figure 4: Accuracy to statements about the functioning in schematic and realistic pictures of causal systems.

A 2 (*Realism*) × 4 (*Presentation Time*) repeated-measures ANOVA revealed a significant main effect of *Presentation Time* ($F(3, 69) = 9.20$, $p < .001$) and a significant interaction *Realism*Presentation Time* ($F(3, 69) = 3.13$, $p = .03$), meaning that for realistic pictures of causal systems the functioning was understood better at longer presentation times, which was not the case for schematic ones.

Bonferroni comparisons confirmed that in schematic causal systems, longer presentation times (2000, 6000 ms)

did not lead to further improvements in understanding of the functioning (both $p_s > .05$) compared to 600 ms presentation time. Thus, it can be concluded that in schematic causal systems, an initial understanding was rapidly acquired (at 600 ms), and at longer presentation times schematic causal systems might have solely been scanned for details. In realistic pictures of causal systems there was no understanding of the functioning but at the longest presentation time of 6000 ms. Bonferroni tests showed that understanding of the functioning still improved between 2000 and 6000 ms ($p = .02$). Thus, it took subjects longer to understand the functioning of realistic pictures of causal systems (6000 ms) than to understand the functioning of schematic ones (600 ms). Subjects probably still attended to realistic pictures of causal systems at 6000 ms in order to extract their functioning. This possibly led to less attention to details at 6000 ms, which could have resulted in the marginally lower performance in detail extraction ($p = .054$) for realistic compared to schematic pictures of causal systems after 6000 ms presentation time.

Discussion

The present study aimed at investigating the extraction of different information (gist, details, and the functioning) from briefly attending to schematic and realistic pictures of scenes and causal systems.

The results demonstrate that the gist was rapidly extracted (< 150 ms) in both scenes and causal systems, confirming prior research from gist extraction in scenes (e.g., Castelhano & Henderson, 2008) and expanding it to instructional material. Moreover, details were recognized more accurately at longer presentation times, which is in line with prior research from detail extraction in scenes (Loftus et al., 1983). Comprehension of the functioning quickly reached an asymptote in schematic pictures of causal systems (at 600 ms). In realistic pictures of causal systems, however, subjects needed more time to understand the functioning, which might have impaired detail extraction at longer presentation times because subjects might have split their attention between details and objects that they assumed to be relevant for understanding the functioning of the system. The analysis of the eye tracking data will reveal whether these assumptions hold true.

Influences on Comprehension of the Functioning of Causal Systems

More familiarity can possibly account for the faster comprehension of the functioning in schematic than in realistic pictures of causal systems. Schematic causal systems often appear in textbooks, but students are seldom faced with and almost never learn from photorealistic pictures of causal systems. Hence, a lack of familiarity with realistic pictures of causal systems could explain why understanding of schematic causal systems reached an asymptote very quickly (600 ms), whereas understanding of realistic pictures of causal systems was still at chance level at 600 ms and at 2000 ms presentation time.

Moreover, there might be an influence of domain-specific knowledge on comprehension of the functioning of causal systems. Unfortunately, in the present study no prior knowledge test could be administered, because causal systems were from many different domains (biology, chemistry, physics, engineering, and mechanics) and thus a prior knowledge test for each domain would have been too long. However, a demographic questionnaire was presented to participants that assessed their prior knowledge with regard to their last school grades in the respective school subjects and their general interest in the different domains. No participant had both very good school grades and a high interest in each of the aforementioned domains. Thus, it is highly unlikely that a participant could answer to all statements about the functioning of causal systems solely by relying on high prior knowledge. To test the influence of prior knowledge on the comprehension of causal systems in the respective domain on a more fined-grained level, further studies will be conducted.

Does the Gist of Causal Systems Help in Learning from Subsequent Text?

Studies that experimentally varied the sequence of presenting a text and a corresponding picture (Kulhavy et al., 1993; Ullrich & Schnotz, 2008) have shown that processing of a picture before the corresponding learning text can foster learning. Kulhavy and colleagues (1993) obtained better learning outcomes when a map was presented before a text. According to the authors, the structure of the map helped subjects in learning from subsequent text. However, in these studies, a picture was presented for either three to five minutes, or even without time constraints, which presumably led to a detailed mental model of the picture that was later integrated with the text and thus resulted in higher learning outcomes.

Results of the present study suggest that the gist in causal systems is extracted after very short presentation times (< 150 ms). It is unlikely that the short presentation (150 ms) of a picture already leads to a detailed mental model of the picture. Presumably, it rather acts as a scaffold (Friedman, 1979). Friedman (1979) assumed that subsequent information can then be added to that scaffold, thereby facilitating incremental mental model construction from pictures (and text, cf. Hegarty & Just, 1993). Moreover, Castelhano and Henderson (2007) suggested that gist extraction leads to priming of the spatial structure of a picture, and this spatial structure "lingers in memory and can facilitate later perceptual and cognitive operations and behavior" (p. 760). If the gist already provided a scaffold of a picture that can be held in memory for some time, then later information from the text could be added to that scaffold, which could result in better learning. To test this assumption, we further plan to conduct studies, in which the picture of a causal system is presented for a short time (e.g., 150 ms) before a subsequent learning text.

Does Comprehension of the Functioning of Causal Systems Help in Learning from Subsequent Text?

The current results suggest that 600 ms can be enough to gain a preliminary comprehension of the functioning of schematic causal systems. As mentioned before, Stone and Glock (1981) showed that when subjects learned from text and pictures, they first attended to the picture for 1000 to 2000 ms before they started to read the text. Thus, this initial attention on the picture was probably long enough not only to extract the gist but also to gain a preliminary comprehension of the pictures' functioning, which in turn could have led to better learning from subsequent text.

However, it is not yet clear whether subjects in the study from Stone and Glock (1981) actually gained a preliminary comprehension of the picture after initially attending to. Subjects could also have attended to the picture in the first place because it merely was more visually appealing than the text. In this case, the initial attention to the picture possibly might not have been helpful for learning. Thus, further studies will have to investigate whether attending to a causal system for the time necessary to understand its functioning (i.e., 600 ms) fosters learning from subsequent text when subjects are instructed to attend to the system to understand its functioning versus when they are not.

From Basic Cognitive Research to Educational Settings

The study demonstrated that a well established effect in basic cognitive research (rapid gist extraction in scenes) can be found with instructional material as well, thereby providing further insights into the roles that pictures may play in learning from text and pictures. As such, this study can be considered as a starting point of an interdisciplinary approach that tries to better understand the processes that take place during learning from pictures and text through systematically applying findings from basic cognitive psychology to educational scenarios. Besides leading to a better understanding of the learning process, in the long run this approach may also provide recommendations for efficient instructional designs in educational settings.

References

Biederman, I., Rabinowitz, J. C., Glass, A. L., & Stacy, E. W. (1974). On the information extracted from a glance at a scene. *Journal of Experimental Psychology: General, 103*, 597-600.

Carroll, P. J., Young, J. R., & Guertin, M. S. (1991). Visual analysis of cartoons: A view from the far side. In K. Rayner (Ed.), *Eye movements and Visual Cognition: Scene Perception and Reading*. New York: Springer.

Castelhano, M. S., & Henderson, J. M. (2007). Initial scene representations facilitate eye movement guidance in visual search. *Journal of Experimental Psychology: Human Perception and Performance, 33*, 753-763.

Castelhano, M. S., & Henderson, J. M. (2008). The influence of color on the perception of scene gist. *Journal of Experimental Psychology: Human Perception and Performance, 34*, 660-675.

Friedman, A. (1979). Framing pictures: The role of knowledge in automatized encoding and memory for gist. *Journal of Experimental Psychology: General, 108*, 316-355.

Hegarty, M., & Just, M. A. (1993). Constructing mental models of machines from text and diagrams. *Journal of Memory and Language, 32*, 717-742.

Henderson, J. M., & Hollingworth, A. (1999). High-level scene perception. *Annual Review of Psychology, 50*, 243-271.

Kulhavy, R. W., Stock, W. A., Verdi, M. P., Rittschoff, K. A., & Savenye, W. (1993). Why maps improve memory for text: The influence of structural information on working memory operations. *European Journal of Cognitive Psychology, 5*, 375–392.

Levie, W. H., & Lentz, R. (1982). Effects of text illustrations: A review of research. *Educational Communication and Technology, 30*, 195–232.

Loftus, G. R., Nelson, W. W., & Kallman, H. J. (1983). Differential acquisition rates for different types of information from pictures. *Quarterly Journal of Experimental Psychology-A, 35*, 187-198.

Mak, P. (2008). Effects of references from text to picture on the processing of school texts: Evidence from eye tracking. In A. Maes & S. Ainsworth (Eds.), *Proceedings EARLI Special Interest Group Text and Graphics: Exploiting the opportunities - Learning with textual, graphical, and multimodal representations*. Tilburg: Tilburg University.

Mayer, R. E. (1989). Systematic thinking fostered by illustrations in scientific text. *Journal of Educational Psychology, 81*, 240-246.

Rayner, K., Rotello, C. M., Steward, A. J., Keir, J., & Duffy, S. A. (2001). Integrating text and pictorial information: Eye movements when looking at print advertisements. *Journal of Experimental Psychology: Applied, 7*, 219-226.

Stone, D. E., & Glock, M. E. (1981). How do young adults read directions with and without pictures? *Journal of Educational Psychology, 73*, 419-426.

Ullrich, M., & Schnotz, W. (2008). Integration of picture and text: Effects of sequencing and redundancy on learning outcomes. In A. Maes & S. Ainsworth (Eds.), *Proceedings EARLI Special Interest Group Text and Graphics: Exploiting the opportunities - Learning with textual, graphical, and multimodal representations*. Tilburg: Tilburg University.

Underwood, G., Jebbett, L., & Roberts, K. (2004). Inspecting pictures for information to verify a sentence: Eye movements in general encoding and in focused search. *Quarterly Journal of Experimental Psychology-A, 57*, 165-182.

Does Spatial Verbal Information Interfere with Picture Processing in Working Memory? The Role of the Visuo-spatial Sketchpad in Multimedia Learning

Anne Schüler (a.schueler@iwm-kmrc.de)
Knowledge Media Research Center, Konrad-Adenauer-Strasse 40,
72072 Tuebingen, Germany

Katharina Scheiter (k.scheiter@iwm-kmrc.de)
Knowledge Media Research Center, Konrad-Adenauer-Strasse 40,
72072 Tuebingen, Germany

Peter Gerjets (p.gerjets@iwm-kmrc.de)
Knowledge Media Research Center, Konrad-Adenauer-Strasse 40,
72072 Tuebingen, Germany

Abstract

The reported study examined whether the processing of spatial verbal information interferes in the visuo-spatial sketchpad with the execution of eye movements, associated with viewing pictures and reading. Seventy-four students were randomly assigned to six groups, resulting from a 2×2×2 mixed design, with spatial secondary task (with vs. without), text contents (visual vs. spatial), and text modality (spoken vs. written) as independent variables. Consistent with our assumptions, learners with text containing spatial contents showed worse recall performance than those with text containing visual contents. Furthermore, written presentation of text containing spatial contents loaded the visuo-spatial sketchpad to a higher extent than spoken presentation. Implications of these results for learning with multimedia are discussed.

Keywords: multimedia; working memory; modality effect; spatial verbal information; secondary task

Introduction

In the last two decades, a lot of research has been conducted on how people learn from multimedia, that is, from the presentation of texts together with pictures (Mayer, 2009).

The Cognitive Theory of Multimedia Learning (CTML; Mayer, 2009) is one of the most important theories concerning multimedia learning. One of its theoretical foundations is an older version of Baddeley's working memory model (1992). According to this model, working memory consists of three systems: The phonological loop (PL), where all verbal information is processed, the visuo-spatial sketchpad (VSSP), where visual and spatial information is processed, and the central executive, which governs the functioning of the phonological loop and the VSSP. Accordingly, the CTML assumes that texts are processed in the phonological loop, whereas pictures are processed in the VSSP. The working memory systems are limited in the amount of information that can be processed in parallel. Accordingly, processes accomplished within the same system can interfere with each other and hinder learning. Therefore, text-picture presentations should be designed in a way that a learner can make optimal use of the cognitive resources so that an overload in one or both systems can be avoided.

However, since Baddeley's first comprehensive descriptions of his model there have been numerous new findings concerning the functioning of working memory that have been considered in newer versions of the Baddeley model, but have not yet been incorporated into the CTML. In particular, the structure of the VSSP has been further specified. According to our view, these specifications may play an important role in multimedia learning. Thus, the aim of this paper is to have a closer look at the VSSP and its implications for learning with multimedia.

A Closer Look at the Visuo-spatial Sketchpad

According to Logie (1995), the VSSP can be divided into a visual and a spatial part. Whereas the visual part deals with information like an object's color or form, the spatial part handles information like spatial sequences or spatial configurations (e.g., Darling, Della Sala, & Logie, 2007; Della Sala et al., 1999). Whereas Logie and colleagues focused on pictorial stimuli, other researchers have addressed the question whether the VSSP may also be involved in the processing of text. This research suggests that if text contains information about spatial and/or visual configurations, it will not be processed only in the PL but also in the respective part of the VSSP, whereas if it contains more abstract information, it will be processed in the PL alone (De Beni et al., 2005; Deyzac, Logie, & Denis, 2006). Another line of research on the spatial VSSP has also shown that this structure is not responsible only for the processing of spatial information but also for the control of movements, for example arm or eye movements (e.g., Postle et al., 2006).

Although from a theoretical perspective the VSSP should play a crucial role in multimedia learning, its involvement has not often been considered empirically in multimedia learning. One method to measure the involvement of the spatial VSSP in task performance is the secondary task paradigm. In this paradigm, two tasks are combined, a primary and a secondary task. The primary task is the main

task, for example a multimedia learning task, whereas the secondary task is a task that loads one of the working memory systems. If both tasks rely on the same working memory systems, they will compete for its limited resources. As a consequence, primary task and/or secondary task performance will decrease compared to a control condition in which participants perform the two tasks separately. A secondary task that is assumed to load the spatial VSSP is the spatial tapping task. In this task, participants have to press buttons in a predefined order on a keyboard, which is hidden from view (e.g., Della Sala et al., 1999). Because the spatial VSSP controls the execution of movements, the continuous tapping interferes with the processing of spatial information (Farmer, Berman, & Fletcher, 1986).

Another way of assessing the involvement of the VSSP focuses on determining the learner's capacity of the spatial and visual VSSP and relating them to learning outcomes. Two tasks have been used to measure the capacities of the spatial and the visual VSSP, respectively, the Corsi block task (Milner, 1971) and the Visual Pattern Test (VPT; Della Sala et al., 1997), respectively. In the Corsi block task, the instructor taps fixed spatial sequences of cubes on a wooden board, which the participant has to recall afterwards. In the VPT, the participant has to recall abstract visual patterns. These patterns are presented in two-dimensional matrices in which a random selection of half of the cells is colored black.

Implications for Multimedia Learning

Figure 1 shows which parts of the VSSP are needed to represent different combinations of pictures and text contents, different amounts of eye movements, and a spatial secondary task. Whereas pictures are assumed to be processed in the visual and spatial VSSP because they contain visual as well as spatial information, texts load the visual or spatial VSSP as a function of their contents. Text containing no visuo-spatial information loads neither the visual nor the spatial VSSP, whereas text containing visual contents loads the visual VSSP (Figure 1, upper row), and text containing spatial contents loads the spatial VSSP (Figure 1, bottom row). Furthermore, as the spatial VSSP controls the execution of eye movements, viewing pictures and reading written text will result in an additional load of the spatial part (Figure 1, b, d, f, h). Moreover, the load of the spatial VSSP can be increased by implementing a spatial secondary task (Figure 1, right column).

In the current paper we focus on three implications that result from this analysis and that will be outlined in the following.

First implication: A Spatial Secondary Task Interferes with Picture Processing, Text Containing Spatial Contents, and Eye Movements. The first implication of the preceding analysis refers to the effects of a spatial secondary task on learning. It is presupposed that the spatial secondary task loads the spatial VSSP but not the visual VSSP (Figure 1, compare left vs. right column). Therefore, the spatial secondary task should interfere with the processing of the picture, the processing of text containing spatial contents, and the execution of eye movements associated with reading. On the other hand, it should not interfere with the processing of texts containing visual contents, and it should interfere less with spoken than with written text, because no eye movements are required to listen to text.

Second Implication: Text Containing Spatial Contents interferes with Picture Processing. When presenting pictures together with text containing spatial contents, one would expect interference in the spatial VSSP, because the processing of the spatial picture and spatial text contents as well as the control of eye movements both take place here (see Figure 1, bottom row). When presenting pictures together with text containing visual contents, one would expect less interference because the load is distributed more equally (see Figure 1, upper row). Accordingly, pictures presented together with text containing spatial contents should result in worse learning outcomes than pictures presented together with text containing non-spatial contents. A study conducted by Schmidt-Weigand and Scheiter (2008) confirms this assumption by showing that pictures are helpful for learning only, when they accompany text with a low degree of spatial information compared to text with a high degree of spatial information.

Third Implication: Written Text Containing Spatial Contents Interferes more with Picture Processing than Spoken Text Containing Spatial Contents. A third implication of the preceding analysis refers to the modality of the text: Because eye movements are not needed only for picture inspection, but also for reading, one might expect worse performance with written text than with spoken text when processing text containing spatial contents. Figure 1 (bottom row) shows that the spatial part is less loaded with spoken text containing spatial contents than with written text containing spatial contents, because more eye movements are required to read the text and to switch between text and picture. This load difference might result in worse learning outcomes for written text containing spatial contents than for spoken text containing spatial contents. For text containing non-spatial contents the difference between written text and spoken text is not expected to be equally harmful, because the text contents are not processed in the spatial VSSP and therefore no interference with the control of eye movements is expected. Note that a general superiority of spoken over written text presentations has been acknowledged for a long time already in multimedia research (i.e., modality effect, Moreno & Mayer, 1998); however, its explanation is different from the one presented here and in particular does not depend on the text content. Hence, we will not address this effect here any further.

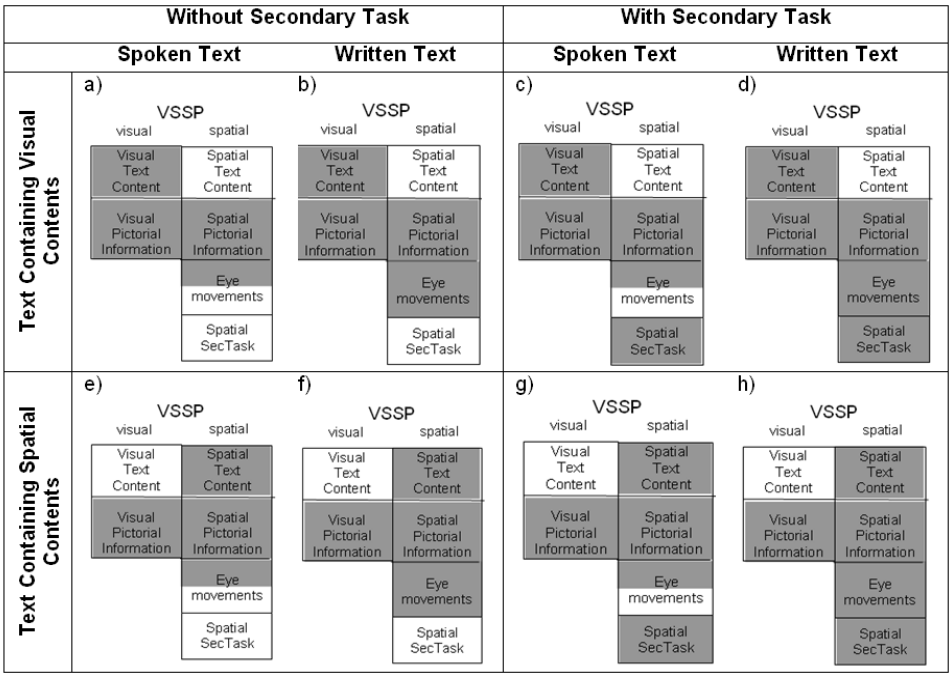


Figure 1. The load (grey color) of the VSSP as a function of the processing of pictures, the processing of different text contents, the processing of a spatial secondary task (Spatial SecTask), and the control of eye movements.

There is some evidence for the prediction that the text contents may moderate the modality effect. Kürschner, Schnotz, and Eid (2007) showed modality effects only with spatial information but not with non-spatial information. One purely text-based study (Glass et al., 1985) explicitly examined the influence of text modality on the processing of text containing visual versus spatial contents. Whereas with regard to sentences about spatial relations a modality effect occurred, this was not the case with regard to sentences about visual characteristics like color.

Experiment

The aim of the current study was to investigate whether processing texts containing spatial contents would interfere with picture processing and whether reading written text containing spatial contents would interfere more with picture processing than listening to the same text. Furthermore, it was investigated whether a spatial secondary task would interfere with the processing of pictures, text containing spatial contents, and eye movements.

Method

Participants and Design. Seventy-four students of the University of Tuebingen (62 female, average age: $M = 21.89$ years, $SD = 3.08$ years, 6 left-handed) participated in the study. They were randomly assigned to one of four conditions, which resulted from a 2×2×2 mixed design, with spatial secondary task (with vs. without) and text contents (visual vs. spatial contents) as between-subject factors and text modality (spoken vs. written text) as within-subject factor. Due to the mixed design, between 18 and 19 students were assigned to one cell (see Table 1).

Materials. The materials were presented in a computerized learning environment. The system-paced learning phase consisted of six static pictures of fictitious fish accompanied by six corresponding texts. Each fish was presented on a single slide. The pictures were identical in all groups, whereas the texts differed with regard to contents and modality as a function of the experimental condition. The lengths and the Flesch reading ease scores (Flesch, 1948) of the two text versions were equivalent indicating that there were no differences in text difficulty across the two versions. The pace of presentation was determined by the duration of the spoken text conditions.

The independent variables were varied between groups in the learning phase as follows: Learners with secondary task had to press different buttons in a predefined order on a keyboard hidden from view during learning. Learners without secondary task learned without performing a secondary task. Learners with text containing visual contents received information about visual features of the depicted fish species, that is, the color or form of specific body parts (e.g., "The pectoral fin has the same light brown color as the dorsal fins"). Learners with text containing spatial contents received information about spatial features of the fish species, that is, the location of a body part or its spatial relation to other parts (e.g., "The pectoral fin lies between the two dorsal fins"). Text modality was varied within the learning environment. Three of the six fish were accompanied by spoken text, the other three fish by written

text (partially balanced design). In the conditions with spoken text, learners listened to the text while the picture was presented on the screen. In the conditions with written text, the text was presented below the picture (see Figure 2).

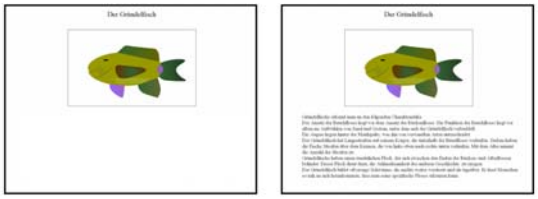


Figure 2: Presentation of the learning materials with spoken (left) and written (right) texts.

Measures. The test phase consisted of four open recall questions, which measured text or picture recall. To measure text recall, learners had to write down everything they remembered from the texts regarding two of the presented fish. To measure picture recall, learners had to draw two of the fish (only information not mentioned in the text was analyzed). Two independent raters blind for experimental condition scored the open recall questions afterwards with an interrater reliability of Cohen's kappa = .79 for text recall and Cohen's kappa = .73 for picture recall. With regard to picture recall, we distinguished between the recall of visual versus spatial picture information. Additionally, the VPT (Della Sala, et al., 1997) to measure the capacity of the visual VSSP and the Corsi Block test (Milner, 1971) to measure the capacity of the spatial VSSP were administered.

Procedure. Participants were tested individually. First, participants were given a short written instruction about the experiment (i.e., about the learning domain as well as the procedure of the experiment). Second, participants in the secondary task conditions were introduced to the task and practiced it for two minutes. Third, all participants entered the system paced learning phase that was subject to experimental manipulation. Fourth, they responded to the open recall questions. Finally, they performed the VPT and Corsi block test. A single experimental session lasted about 60 minutes.

Results

Because it could not be excluded that gender or handedness interacted with the processing of spatial information, we conducted prior analyses in a first step. For gender, no significant interactions were observed, indicating that gender did not influence learning outcomes. Regarding handedness, the corresponding analyses were not possible because of an insufficient number of left-handed participants. However, due to the small number of left-handed participants we did not expect an influence on learning outcomes.

Because of the results of the prior analyses, we collapsed across gender and handedness for the following analyses. For text recall, an ANOVA was conducted. For spatial and visual picture recall, the corresponding variables were analyzed by means of a MANOVA. In all analyses, secondary task and text contents were incorporated as between-subject factors and text modality was incorporated as within-subject factor. To control for individual differences in the capacity of the visual and spatial VSSP, the Corsi block scores and VPT scores were incorporated as covariates. Note that there were no interactions between the two capacity measures and any of the experimental factors. In the following, the statistical details are only reported for significant results, because of space limitations. Adjusted marginal means and standard errors corrected for the influence of the visual and spatial VSSP capacity are reported in Table 1.

With regard to *text recall*, the results showed an effect of the VPT, $F(1, 68) = 4.11$, $p = .047$, $\eta^2_p = .06$: The higher the capacity of the visual VSSP was, the better learners recalled the text information ($r = .21$, $p = .08$). Furthermore, in line with the second implication, learners with text containing visual contents ($M = 32.82\%$, $SE = 3.37$) outperformed learners with text containing spatial contents ($M = 18.95\%$, $SE = 3.36$), $F(1, 68) = 8.29$, $p = .01$, $\eta^2_p = .11$. This indicates that text containing spatial contents and picture processing interfere in the spatial VSSP, resulting in worse learning outcome for the recall of spatial text information compared to visual text information.

With regard to *picture recall*, the MANCOVA showed a significant difference between learners with texts containing visual and spatial contents, $V = .46$, $F(3, 67) = 28.20$, $p < .001$, and an influence of the secondary task on learning outcomes, $V = .09$, $F(3, 67) = 3.21$, $p = .046$. Also the three-way interaction text modality × text content × secondary task was significant, $V = .09$, $F(3, 67) = 3.26$, $p = .045$.

Table 1: Adjusted marginal means and standard errors as a function of the experimental condition.

	text containg visual contents				text containing spatial content			
	without secondary task		with secondary task		without secondary task		with secondary task	
	spoken $n = 18$	written $n = 18$	spoken $n = 19$	written $n = 19$	spoken $n = 18$	written $n = 18$	spoken $n = 19$	written $n = 19$
Recall of text information (%)	35.91 (7.08)	35.30 (6.50)	30.32 (6.69)	29.79 (6.14)	20.48 (6.85)	20.12 (6.30)	15.82 (6.87)	19.37 (6.31)
Recall of visual picture information (%)	57.52 (4.79)	49.97 (4.37)	52.09 (4.52)	46.83 (4.13)	32.90 (4.63)	33.24 (4.23)	36.61 (4.64)	27.90 (4.24)
Recall of spatial picture information (%)	53.32 (5.43)	44.59 (5.87)	38.80 (5.13)	42.63 (5.54)	42.45 (5.26)	58.04 (5.68)	44.81 (5.27)	34.58 (5.69)

Follow-up three-way ANCOVAs confirmed the expected main effect of text content with regard to the recall of visual picture information: Learners with text containing visual contents ($M = 51.60\%$, $SE = 2.06$) recalled the visual aspects of the pictures (like color or form) better than learners with text containing spatial contents ($M = 32.67\%$, $SE = 2.06$), $F(1, 68) = 41.16$, $p < .01$, $\eta^2_p = .38$). This confirms the second implication that text containing spatial contents interferes with picture processing, whereas text containing visual contents does not. With regard to the recall of spatial picture information the effect of the secondary task, $F(1, 68) = 6.24$, $p = .02$, $\eta^2_p = .08$, as well as the three-way interaction, $F(1, 68) = 5.76$, $p = .02$, $\eta^2_p = .08$, were confirmed: In line with the first implication, learners, who performed a spatial secondary task during learning, recalled the spatial picture information ($M = 40.21\%$, $SE = 2.58$) worse than learners who did not perform a secondary task ($M = 49.60\%$, $SE = 2.65$). This indicates that the spatial picture contents are processed in the spatial VSSP. However, the Bonferroni tests of the three-way interaction text modality × text content × secondary task, showed that this main effect of the spatial secondary task on spatial picture recall was due to interference between the spatial secondary task and the processing of written text containing spatial contents ($p = .01$, see Figure 3). This result supports the third assumption, because it indicates a higher load of the spatial VSSP with written than with spoken text presentation, when spatial text contents are presented.

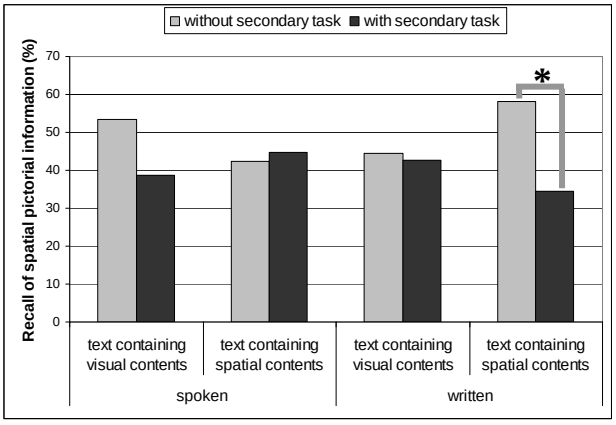


Figure 3: The pattern of results for spatial picture recall (adjusted means). *$p < .05$

Summary and Discussion

One purpose of the reported study was to examine whether pictures, text containing spatial contents as well as eye movements load the spatial VSSP. Furthermore, the hypotheses were tested that text containing spatial contents would interfere with picture processing and that this interference would be affected by the modality of the presented text.

The first assumption concerning interference between the spatial secondary task and the processing of pictures, text containing spatial contents as well as the execution of eye movements was only partially confirmed. Performance decrements while conducting a spatial secondary task were observed with regard to the recall of spatial picture information, especially for learners with written text containing spatial contents (see Figure 3). As mentioned before, the load of the VSSP is assumed to be extremely high in this specific case (see Figure 1, h). This may explain why the secondary task interfered particularly with the recall of spatial picture information accompanied by written text containing spatial contents. Contrary to our assumptions, the secondary task did not hinder the recall of spatial verbal information or the recall of written text in general. These findings imply that the processing of text containing spatial contents and the control of eye movements in general did not load the spatial VSSP to such a high degree that interference with a secondary task was observed.

The second assumption concerning worse learning outcomes with pictures accompanied by text containing spatial contents as compared to text containing visual contents was confirmed: Learners, who received pictures together with text containing spatial contents, showed overall worse performance in recalling text-based and visual picture-based information. Furthermore, learners with text containing visual contents recalled spatial picture-based information to the same extend as did learners with text containing spatial contents. Thus, text containing spatial contents did not support the recall of spatial picture information: These results indicate that learners with text containing visual contents processed the picture more thoroughly than learners with text containing spatial contents. How can these results be explained? In the theoretical part of the paper we assumed that text containing spatial contents leads to an additional load of the spatial VSSP, resulting in worse learning outcomes for text and picture recall. However, in total, the secondary task did not reduce the performance of learners with text containing spatial contents, which may indicate that spatial text contents do not increase the load of the spatial VSSP. Instead, as mentioned above, the secondary task interfered with the processing of spatial picture information only when the load of the spatial VSSP was assumed to be extremely high (see Figure 3). Thus, it cannot be definitely concluded that the observed performance decrement with text containing spatial contents is due to a higher load of the spatial VSSP. Rather, it is also possible that a spatial secondary task does not reduce performance when the spatial VSSP gets simply loaded but only if it gets *over*loaded.

An alternative explanation for the found performance decrement with text containing spatial contents might be the text difficulty. With regard to recall of text contents, one might argue that text containing visual information, that is, information about color and form, is easier to

recall than text containing spatial information, that is, information about spatial relationships or the position of a certain characteristic. Thus, the fact that learners with text containing visual contents performed better, when they had to recall text-based information might potentially be simply explained by differences in text difficulty and not by interference in the spatial VSSP. On the other hand, the Flesch scores indicated the same reading ease for both texts. Furthermore, one may ask why text difficulty should influence the processing of the pictures, which were the same in all groups. One might argue that because text containing spatial contents is more difficult to process, learners might concentrate more on the text and neglect the picture. This in turn might result in worse recall performance for pictures. However, a further study where we used eye tracking methodology to assess the amount of attention devoted to text and pictures showed no differences between learners with different text contents with regard to their viewing behavior. Thus, it seems as if text difficulty is not responsible for the results.

The third assumption concerning a modality effect that would occur only with text containing spatial contents was confirmed for the recall of spatial picture information, when learners additionally performed a secondary task. Thus, under extreme load conditions the eye movements necessary to read the text interfered with picture processing. This implies that written text can decrease performance when the load of the spatial VSSP is already high.

To conclude, these results show that the presentation of text containing spatial contents together with pictures might be detrimental to learning under certain circumstances such as restricted learning time or system-paced presentations. Under these conditions it might be better to convey spatial information only through visualizations, because visualizations are more efficient than texts for accomplishing tasks that require the processing of visuo-spatial properties (Larkin & Simon, 1987). If it is not possible to convey the spatial information only via picture, we recommend presenting spoken texts, because otherwise the eye movements associated with reading may decrease performance when the load of the spatial VSSP is high.

To get deeper insights into the interplay of working memory and multimedia learning, further research is needed that addresses more fine-grained processing aspects (e.g., measuring the amount of eye movements and relate it to spatial text processing). This is in line with our conviction that more cognitive basic research is needed to develop more precise theoretical frameworks for explaining how multimedia learning works.

References

Baddeley, A. D. (1992). Working memory. *Science, 255*, 556-559.

Darling, S., Della Sala, S., & Logie, R. H. (2007). Behavioural evidence for separating components within visuo-spatial working memory. *Cognitive Processing, 8*, 175-181.

De Beni, R., Pazzaglia, F., Gyselinck, V., & Meneghetti, C. (2005). Visuospatial working memory and mental representation of spatial description. *European Journal of Cognitive Psychology, 17*, 77-95.

Della Sala, S., Gray, C., Baddeley, A., Allamano, N., & Wilson, L. (1999). Pattern spans: A tool for unwelding visuo-spatial memory. *Neuropsychologia, 37*, 1189-1199.

Della Sala, S., Gray, C., Baddeley, A., & Wilson, L. (1997). *Visual Pattern Test: a test of short-term visual recall.* London: Harcourt Assessment.

Deyzac, E., Logie, R. H., & Denis, M. (2006). Visuospatial working memory and the processing of spatial descriptions. *British Journal of Psychology, 97*, 217-243.

Farmer, E. F., Berman, J. V. F., & Fletcher, Y. L. (1986). Evidence for a visuo-spatial scratch-pad in working memory. *Quarterly Journal of Experimental Psychology, 38,* 675-688.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology, 32*, 221-233

Glass, A., Millen, D., Beck, L., & Eddy, J. (1985). Representation of images in sentence verification. *Journal of Memory and Language, 24*, 442-465.

Kürschner, C., Schnotz, W., & Eid, M. (2007). Welchen Einfluss haben die Präsentationsmodalität und Repräsentationsmodalität auf die kognitive Verarbeitung von Text mit Bildern. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 39,* 70-83.

Larkin, J. H., & Simon, H. A. (1987). Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science, 11*, 65-99.

Logie, R. H. (1995). *Visuo-spatial working memory*. Hove, England: Erlbaum.

Mayer, R. E. (2009). *Multimedia Learning.* Second edition. Cambridge: Cambridge University Press.

Mayer, R. E., & Moreno R. (1998). A split-attention effect in multimedia learning: Evidence for dual processing systems in working memory. *Journal of Educational Psychology, 90*, 312-320.

Milner, B. (1971). Interhemispheric differences in the localization of psychological processes in man. *British Medical Bulletin, 27*, 272-277.

Postle, B. R., Idzikowski, C., Della Sala, S., Logie, R. H., & Baddeley, A. (2006). The selective disruption of spatial working memory by eye movements. *The Quarterly Journal of Experimental Psychology, 59*, 100-120.

Schmidt-Weigand, F., & Scheiter, K. (2008). The Influence of Spatial Text Information on the Multimedia Effect. In J. Zumbach, N. Schwartz, T. Seufert, & L. Kester (Eds.) *Beyond knowledge: The legacy of competence*. Dordrecht, The Netherlands: Springer.

Impact of placing icons next to hyperlinks on information-retrieval tasks on the web

Saraschandra Karanam (saraschandra@research.iiit.ac.in)
Cog Sci Lab, International Institute of Information Technology, Hyderabad, India.

Janhavi Viswanathan (janu111989@gmail.com)
Mahatma Gandhi Institute of Technology, Hyderabad, India.

Anand Theertha (anand.dharoor@gmail.com)
Mahatma Gandhi Institute of Technology, Hyderabad, India

Bipin Indurkhya (bipin@iiit.ac.in)
Cog Sci Lab, International Institute of Information Technology, Hyderabad, India

Herre van Oostendorp (herre@cs.uu.nl)
Center for Content and Knowledge Engineering, Institute of Information and Computing Sciences,
Utrecht University, Utrecht, Netherlands.

Abstract

Though several studies have demonstrated the usefulness of pictures in multimedia learning, memory, cognitive load and visual search, there have been very few attempts to study their impact in the web-navigation scenario. Also, cognitive models of web-navigation (like CoLiDeS, CoLiDeS+) ignore the information from visual modality and focus solely on the information from text. We conducted an experiment to study the extent to which providing icons next to hyperlinks facilitates information retrieval tasks on the web. Three different versions of navigation styles were created: Hyperlinks with Icons, Hyperlinks alone and Icons alone. Users took significantly less time, were significantly less disoriented and made fewer clicks to finish their tasks when icons were provided along with hyperlink text. These results suggest that it is important for a cognitive model on web-navigation to include information from pictures. An important practical implication is to provide meaningful icons next to hyperlinks for better navigation.

Keywords: Web-navigation, text, pictures, icons, web-usability, cognitive model

Introduction

With the advancement of web-technology, the World Wide Web (WWW) now has evolved into a complete hypermedia environment, i.e. information is spread across all modalities – text, picture, audio and video. This adds to the complexity and the *lost in hyperspace* phenomenon of the users (Conklin, 1987). Pirolli and Card (1999) found that a user always follows the path that gives highest *information scent*; i.e., a user estimates the cost and value of taking a particular action like clicking on a hyperlink, and comparing several such actions always picks the action that has the highest value or information scent (Chi *et al*., 2000, 2001).

Inspired by this information scent model, several cognitive models of web-navigation (CoLiDeS, CoLiDeS+, SNIF-ACT and MESA) have been developed. Comprehension based Linked model of Deliberate Search (CoLiDeS) developed by Kitajima, Blackmon and Polson (2000) divides the process of navigating a website into four steps – *parsing, focusing, comprehension and elaboration* and *selecting.* A user first parses the web page into 5-10 top-level schematic regions, focuses on one of the sub-regions, comprehends and generates an elaborated representation of each object in the sub-region based on his or her background knowledge and finally selects one object in that sub-region. This final step of selection is based on the computation of semantic similarity between the user's goals and the elaborated representations developed by the user.

CoLiDeS+ developed by Juvina and Oostendorp (2005, 2008) improved CoLiDeS by incorporating contextual information, i.e. information from previously visited web pages. It computes *path adequacy* as the semantic similarity between the user goal and the navigation path. CoLiDeS+ selects the incoming information only if it increases path adequacy. When the incoming hyperlink does not increase path adequacy, other hyperlinks with lower information scent are considered. Further, it considers backtracking to other regions in the same page and then to other pages.

Miller and Remington (2004) proposed a *Method for Evaluating Site Architectures* (MESA). This model focuses on the quality of link labels and the effectiveness of various link-selection strategies. By varying the link quality and using links that are not fully descriptive of the target goals, user behaviour is modelled. The common condition when the user is not sure of his goal or is not knowledgeable enough to assess the relevance of the link texts to the goal is also modelled.

SNIF-ACT (*Scent based Navigation and Information Foraging in the ACT Architecture*) developed by Pirolli and Fu (2003), predicts user-navigation behaviours when they perform unfamiliar information retrieval tasks. It also predicts that users would leave a site when the information scent falls below a threshold value. SNIF-ACT is based on an algorithm called Web User Flow by Information Scent

(WUFIS) developed by Chi *et al.* 2001. It combines both information retrieval and spreading activation techniques to arrive at the probabilities associated with each hyperlink that specify the proportion of users who will navigate through it.

All these models compute information scent by calculating the semantic similarity between the user goal and the hyperlink text. Although a web page contains much more information than just hyperlink text, the models ignore all of it. Actually, there has been very little research on the impact of such complex hypermedia environments on navigation in such web scenarios. However, extensive research from the fields of multimedia learning, memory, cognitive load and visual search give us an insight into the positive impact of multimodal representations, and potentially on web-navigation. We will present a study that examines the extent to which providing icons next to hyperlinks facilitate information retrieval.

Overview of research on impact of pictures

Mayer and Moreno (2003a, 2003b, 2004) demonstrated that meaningful learning that involves attending to important aspects of material, organizing it into a coherent structure, comprehending and understanding it and integrating with already known knowledge can happen better with content that has both visual and verbal format. Their theory is based on the dual-channel assumption of Paivio (1986), according to which humans possess separate channels for processing verbal and visual material, and the working memory theory of Baddeley (1998), which states that only a limited amount of processing can take place in any channel at any point of time. Of the seven principles that Mayer came up with, three are relevant to the web-scenario as well: coherence principle (present relevant pictures and avoid unnecessary information), spatial contiguity principle (present on-screen text nearer to the corresponding picture) and personalization principle (use words that are familiar to the user).

Larkin and Simon (1987) differentiate between diagrammatic and textual representations. They demonstrated that by preserving topological and spatial relations, diagrammatic representations make it easier to solve certain geometric problems. Scaife and Rogers (1996) proposed that graphical representations bring advantages for learning by reducing the amount of effort required to solve a problem and reducing ambiguity by limiting the range of inferences that can be drawn. It is known that students develop deeper understanding of material through self-explanations. Ainsworth and Loizou (2003) showed that this phenomenon is stronger when material with diagrams is presented. Sweller and Chandler (1994) showed that by physically integrating text and corresponding pictures, working memory load is reduced as it reduces redundancy and the effort in integrating information from various sources. Levie and Lentz (1982) have shown that information is remembered and retrieved better when accompanied by relevant pictures.

An eye tracking study by Namatame *et al.,* 2008 showed that for a directory-based search in a computer, fixation time and number of saccades are the least for labelled pictogram condition. This is also evident from the research on visual attention (Desimone and Duncan, 1995, Treisman and Galade, 1980), which demonstrated that any object significantly different from its surrounding objects along intensity, colour, orientation, and motion direction will be perceived by the human brain at very early stages of visual processing. Pictures, being the most salient objects on a web page, are thus attended to by the user before text. Although all this literature points to the positive influence pictures have on learning, memory and cognitive load, little effort has been put to study their impact in the domain of web-navigation.

Impact of graphics on web-navigation

Finding information on the web can at times be a daunting task given its vastness and non-linear nature. Further, recent research by Ruddle (2009) showed that even frequent visitors remember very little of the content and structure of a website. This further warrants the need to study factors affecting web-navigation behavior in detail. It has been found by Carnot *et al.,* (2001) that using a concept-map-based browser, which is a hierarchical organization of a set of concepts and relations between the concepts gives much more accurate search performance compared to a normal browser. The position of the navigation bar was found to significantly affect the mean time spent on a page (Petrie *et al.,* 2009).

Hinesley (2005) studied the impact of graphics in locating web page widgets by taking two versions of a page – one with original text intact and the other with all text replaced with character 'X' (greeked pages). She found that it was more difficult to find textual widgets on greeked pages than graphical widgets. Also, Hinesley and Blackmon (2008) investigated the interaction between location expectations and graphics with greeked pages. They found yet again that the performance detriment for graphical widgets was less when location expectations were violated. Hinesley thus claimed that it is graphics that play an important role in identifying web page widgets.

We scrutinized this claim in Karanam *et al.,* (2009) by examining the interaction of text and graphics completely. While Hinesley manipulated text with graphics intact in the first experiment, she manipulated graphics on greeked text in the second. We took four versions of each web page by systematically varying text and graphics. Our major result was that in the absence of graphics, having textual information was better than having no text. When there were no graphics, textual information significantly reduced the user's efforts in finding a widget. Thus, we argued that both text and graphics play an equally important role in a web page and it is important for cognitive models of web-navigation to take this into account.

Both Hinesley and Karanam restricted the tasks in their experiments to locating a widget on a web page. They did not involve any navigation or information retrieval. All that the user had to do was to locate the widget and click on it. What is the impact of graphics on navigation performance? Do graphics and text influence navigation for some information retrieval in the same way as they influence the task of locating a widget? Generally, we assume that users would visit a website with a predefined goal in mind. Thus, what would be the impact on their accuracy of finding the correct target page? This formed the starting point of our next study in which we wanted to investigate if providing meaningful icons next to hyperlinks would aid the user in navigating better. Would it help the user in finding their answers to their goals quicker? Van Oostendorp and Holzel (2005) for instance did find that presenting icons of the participants next to labels of messages on a bulletin board during collective problem solving had a positive effect on communication and performance. We will examine these questions by presenting participants with three different navigation structures – hyperlinks with icons, only hyperlinks and only icons. We will measure the influence of these three different structures on task completion times, number of clicks, task accuracy and user disorientation.

Method

Participants

Forty-five graduate and undergraduate students from Mahatma Gandhi Institute of Technology participated in this experiment. The mean age of the participants was 23. A pre-test was administered before the experiment on the content of the materials used. Three questions each on the country "Georgia" and "Musical Instruments" were given. Each question had four multiple choices, and the participants had to choose one of them as the answer. Correct answers were scored as 1 and wrong answers as 0. For each participant, total score was then calculated by taking the sum of individual scores of each question. All participants scored low on the final scores. (M=1.31, SD=1.05). It can be inferred that their prior knowledge of the subject was quite low.

Material and Apparatus

Website We used two topics for our websites: "Encyclopaedia of Georgia" (25 pages) and "Musical Instruments" (31 pages). Website on Georgia had content describing its geography, society, culture and religion. Website on Musical Instruments had content on different methods of classifying musical instruments – the Western System, the Hornbostel Sachs system and classification by Nationality.

Each website was 4 levels deep. Three versions of each website were created based on three different navigational styles: only hyperlink text (L), hyperlink text with icons (LI) and only icons (I), as shown in figure 1. For each hyperlink, at least five icons were collected, and two authors of this paper scored them for relevancy on a 7-point Likert scale. The icons with highest scores for relevancy were chosen as icons.



Figure 1: Three versions of a webpage with different navigational styles

Information Retrieval Tasks A total of eight user goals were generated: four for each website, one for each level. An example user goal for Georgia website could be – *"A traditional Georgian confection made of caramelized nuts, usually walnuts, fried in honey, is served exclusively on New Year's Eve and Christmas. Name it".*

Table1: Information Retrieval Tasks

"Georgia" Website	
Level	**Task**
1	The Georgian school system is divided into four stages, what are they?
2	Name two traditional Georgian feast songs.
3	A traditional Georgian confection made of caramelized nuts, usually walnuts, fried in honey, is served exclusively on New Year's Eve and Christmas. Name it.
4	Name three trees that cover the northern slope of Greater Caucasus Mountains?
"Musical Instruments Website	
Level	**Task**
1	According to Hornbostel Sachs system of Musical Instruments Classification, which mode of sound production do Idiophones use?
2	In the ancient system of musical instruments classification, xylophone belongs to which category of percussion instruments?
3	The Russian musical instrument – 'Ghusli' has similarities with other instruments in China, Japan and Baltic countries. What are they?

4	In Ancient system of classifying musical instruments, saxophone is categorized as woodwind instrument and not brass instrument. Why?

Measures Our dependent variables were mean task-completion time, number of clicks, task accuracy and disorientation.

Mean Task-Completion Time: The time taken by the user to finish the task was measured. There was a time limit of 5 minutes for each task.

Number of clicks: The number of clicks made by the user before reaching the target page.

Task Accuracy: Task accuracy was measured by scoring the answers given by users. A correct answer from correct page was scored 1. A wrong answer from correct page was scored 0.5. Wrong answers from wrong pages and answers beyond time limit were scored 0.

Disorientation: An objective measure of disorientation was used: It was computed using Smith (1996)'s L measure.

$$L = \sqrt{((N/S - 1)^2 + (R/N - 1)^2)}$$

Where:
R = number of nodes required to finish the task successfully (thus, the number of nodes on the optimal path);
S = total number of nodes visited while searching;
N = number of different nodes visited while searching.

Design

We used a between-subjects design with fifteen participants in each group. Every participant answered all eight questions. The order of questions was randomized.

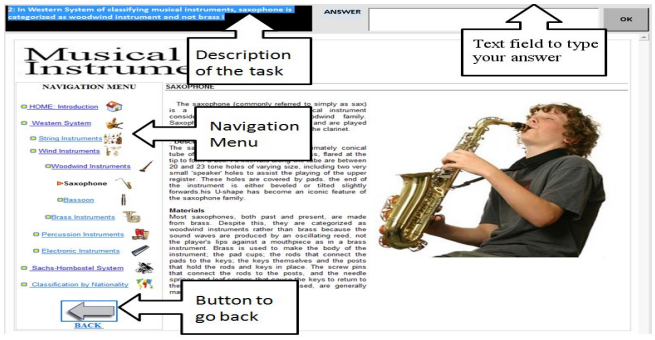


Figure 2: Layout presenting information retrieval tasks

Procedure

After the pre-test to check their prior knowledge about the domain, participants were presented with eight information retrieval tasks on the two different websites in random order. Their task was to locate the target page that contains the answer to these questions; type the answer in the box provided and proceed to the next task. Figure 2 shows the layout of the screen presenting information retrieval tasks. Participants first saw the task description on the screen and then the website was presented in a browser. The task description was always present in the top-left corner, in case the participant wished to read it again.

Results

We did not find a significant impact of condition on task accuracy (p>.05) and therefore we only report the results on other variables – task-completion time, clicks and disorientation.

We first did a mixed ANOVA analysis with our experimental condition as a between-subjects variable and Website as a within-subjects variable. We got a very strong main effect of website for all three of our measures but the interaction effect was not significant. We interpret that the website on musical instruments was relatively unfamiliar and new compared to the website on Georgia, to most of our participants and therefore they took more time, more clicks and were more disoriented in performing their tasks.

Task-Completion Time The number of time-outs in each of the three conditions was computed. The time-out percentages were: Hyperlinks with icons (7.5%), only hyperlinks (10%) and only icons (12.5%). We considered only the tasks for which participants gave correct answers for this analysis. A between-subjects one-way ANOVA with the three experimental conditions as independent variable and mean task completion time as dependent variable was conducted. Results show a main effect of condition (F(2,36) =4.427, p<.05). Figure 3 shows the means plot.

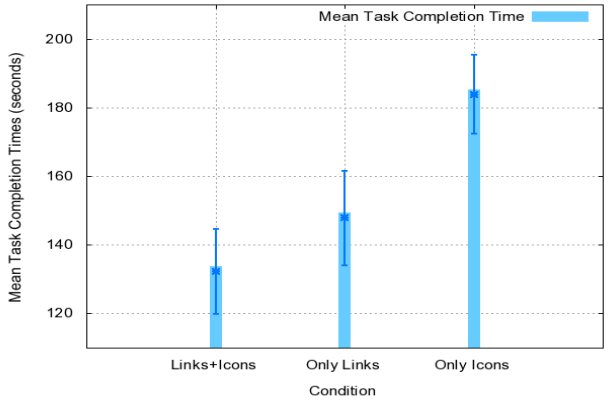


Figure 3: Mean Task-Completion Times

Post-hoc tests reveal significant differences between the groups – (LI)-(I) p<.01. That is, the task-completion times were significantly less when both links and icons were together when compared to only icons. The difference between (L) and (I) groups was not significant p>.05.

Number of Clicks A similar between-subjects one-way ANOVA with the three experimental conditions as independent variable and the number of clicks taken to find the target page with correct answer as dependent variable was conducted. The main effect of Condition is highly significant ($F(2,36) = 43.239$, $p<.001$). Figure 4 depicts this relationship for average number of clicks.

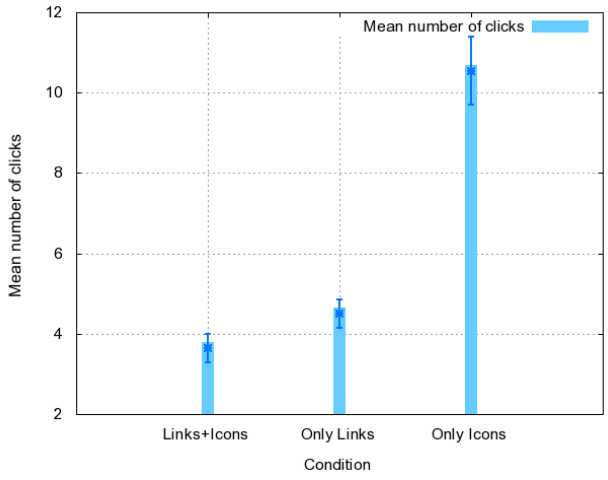


Figure 4: Mean number of clicks

Post-hoc tests reveal significant differences between (LI)-(I) $p<.05$ and (L)-(I) $p<.05$ groups. The number of clicks users took when there were both icons and hyperlink text was significantly less than when there were only icons. Similarly, the number of clicks users took to finish their tasks when there was only hyperlink text was also significantly less than when there were only icons. In other words, having only icons took the maximum number of clicks to finish the tasks, suggesting that only icon-based navigation might not be advisable. The difference in number of clicks between (LI) and (L) groups was not significant ($p>.05$).

Disorientation A between-subjects one-way ANOVA with the experimental condition as independent variable and mean objective disorientation measure as dependent variable was conducted. Results reveal a very significant main-effect of condition $F(2,36)=33.598$, $p<.001$.

Post-hoc tests reveal significant differences between the pairs – (LI)-(I) $p<.05$ and (L)-(I) $p<.05$. Having only icons induced the maximum disorientation in users. Users deviated from the optimum path the most under this condition. They took the maximum number of de-tours and returned to the same page visited earlier frequently. Their disorientation significantly reduced when icons were replaced with corresponding hyperlink text. Users could navigate much better compared to the condition when there were only icons. Further, when both hyperlink text and icons were provided, there was even more significant decrease in disorientation. Figure 5 shows the graph.

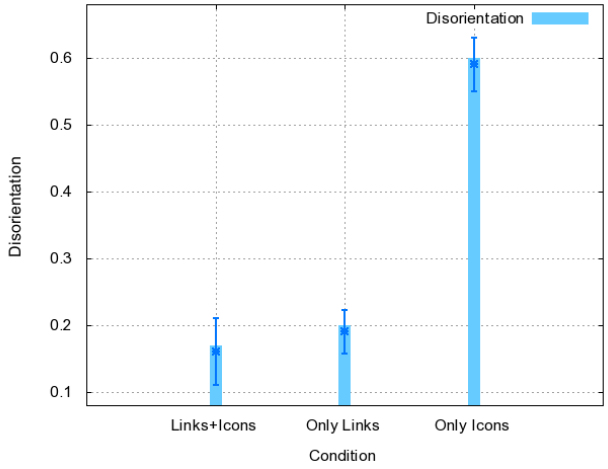


Figure 5: Disorientation

Discussion

In this research, we focused on the impact of providing icons next to hyperlink text in the main navigation menu of a page on user's search and information retrieval performance. Overall, we found that providing icons next to hyperlink text is very helpful for users in reducing the amount of time they take to finish their task, number of clicks they take to reach their target page and in optimizing the path they take to the target page.

It has been found that users take significantly less time when both hyperlink text and corresponding icons are provided compared to the conditions when only hyperlink text or only icons are present. Number of clicks also shows a similar pattern: Users take significantly less number of clicks to find their target pages with icons and hyperlink text present when compared with the pages with only hyperlinks or only icons. Users were disoriented the most when there were only icons present. It was hard to navigate with only icons. This phenomenon of disorientation decreased significantly when there was only hyperlink text and further decrease was effectuated by placing meaningful icons next to the text.

In general, our results support the positive impact of pictures found elsewhere in other domains like multimedia learning, cognitive load and visual search. We have shown that pictures / graphical information together with textual information play an important role in improving overall user-performance in not only locating their target on a web page but also navigating through a website and finding their target pages.

One practical implication of this study is to use meaningful icons next to hyperlink text to improve the overall usability of a website. Also, on basis of this study and our previous study (Karanam *et al.*, 2009), we argue strongly for inclusion of pictorial and graphical information in cognitive models of web-navigation. We have already

shown that semantic information derived from pictures can be included into CoLiDeS (Karanam *et al.,* 2009). We also demonstrated that such a model would predict the correct hyperlink more frequently and more accurately.

References

Ainsworth, S., & Loizou, A.T. (2003). The effects of self-explaining when learning with text or diagrams. *Cognitive Science,* 27(4), 669-681.

Baddeley, A. (1998). *Human Memory.* Boston: Allyn & Bacon

Carnot, M.J., Dunn, B., Canas, A.J., Gram, P., & Muldoon, J. (2001). Concept maps vs. Web pages for information searching and browsing. Institute for Human and Machine Cognition, University of West Florida, USA. http://cmap.coginst.uwf.edu.

Chi, E. H., Pirolli, P., Chen, K., & Pitkow, J. (2001). Using information scent to model user information needs and actions and the Web. *Proceedings of CHI 2001,* ACM Press, 490–497.

Chi, E., Pirolli, P., & Pitkow, J. (2000). The scent of a site: A system for analyzing and predicting information scent, usage, and usability of a website. *Proceedings of CHI 2000*, ACM Press, 161–168.

Conklin, J., 1987. Hypertext an introduction and survey. Computer September, 17-41.

Desimone, R., & Duncan, J., (1995). Neural Mechanisms of Selective Visual Attention, *Annual Review of Neuroscience*, *18,* 193-222.

Hinesley, G.A. (2005). The impact of graphical conventions and layout location on search for webpage widgets. Unpublished Dissertation, University of Colorado, Boulder.

Hinesley, G.A., & Blackmon, M.H. (2008). The Impact of Graphics and Location Expectations on the Search for Webpage Widgets. *Workshop on Cognition and the Web*, Granada, Spain.

Juvina, I., Oostendorp, H. van, Karbor, P., & Pauw, B. (2005). Toward Modeling Contextual Information in Web Navigation. *XXVII Annual Conference of the Cognitive Science Society,* Stresa, Italy.

Juvina, I. & Oostendorp, H. van (2008). Modeling Semantic and Structural Knowledge in Web Navigation. *Discourse Processes, 45*(4-5), 346-364.

Karanam, S., Oostendorp, H. van, Puerta Melguizo, M.C., & Indurkhya, B. (2009). Integrating graphical information into cognitive modeling of web-navigation. *31st Annual Conference of the Cognitive Science Society.* Amsterdam, Netherlands.

Kitajima, M., Blackmon, M.H., & Polson, P.G. (2000). A Comprehension-based Model of Web Navigation and Its Application to Web Usability Analysis. *Proceedings of CHI 2000,* ACM Press, 357-373.

Larkin, J.H., & Simon, H.A. (1987). Why a diagram is (sometimes) worth ten thousand words. C*ognitive Science*, *11*, 65–99.

Levie, W. H. & Lentz, R. (1982). Effects of text illustrations: A review of research. *Educational Communication and Technology*, 30(4), 195-233

Mayer, R. E. & Moreno, R. (2003a). Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist*, *38*, 43-52.

Mayer, R.E. (2003b). The promise of multimedia learning: using same instructional design methods across different media. *Learning and Instruction, 13*, 125-139.

Mayer, R.E., & Moreno, R. (2004). Animation as an aid to Multimedia Learning. *Journal of Educational Psychology Review, 14*(1), 87-99.

Miller, C. S., & Remington, R. W. (2004). Modelling Information Navigation: Implications for Information Architecture. *Human-Computer Interaction, 19*(3), 225-271.

Namatame, M., & Kitajima, M. (2008). Suitable Representations of Hyperlinks for Deaf Persons: An Eye-tracking Study. *Proceedings of the 10th International ACM SIGACCESS Conference on Computers and Accessibility*, 247-248

Paivio, A (1986). Mental representations: a dual coding approach. Oxford. England: Oxford University Press.

Petrie, H., Papadofragkakis, G., Power, G., & Swallow, H. (2009). Navigational Inconsistency in Websites: What does it mean to users? In T. Gross et al. (Eds.): *INTERACT 2009*, Part I, LNCS 5726, pp. 423–427.

Pirolli, P., & Card, S.K. (1999). Information Foraging. *Psychological Review*, *106(4)*, 643-675.

Pirolli, P., & Fu, W.T. (2003). SNIF-ACT: a model of information foraging on the World Wide Web. *9th International Conference on User Modeling (UM 2003);* Johnstown; PA. Berlin: Springer Verlag; LNCS 2702: 45-54.

Ruddle, R. A. (2009). How do people find information on a familiar website? *Proceedings of the 23rd BCS Conference on Human-Computer Interaction (HCI'09)*, 262-268.

Scaife, M & Rogers Y. (1996) External cognition: How do graphical representations work? *International Journal of Human Computer Studies*, *45*, 185-213.

Smith, P. (1996). Towards a practical measure of hypertext usability. *Interacting with Computer, 8* (4), 365-381.

Sweller, J., & Chandler, P. (1994). Why some material is difficult to learn. *Cognition and Instruction, 12*, 185–233.

Treisman, A.M., & Gelade, G. (1980). A Feature Integration Theory of Attention. *Cognitive Psychology, 12*, 97-136.

Van Oostendorp, H. & Holzel, N. (2005). Supportive collective information processing in a web-based environment. In H. van Oostendorp, L. Breure & A. Dillon (Eds.), *Creation, Use and Deployment of Digital Information* (pp. 145-155). Mahwah, NJ: Lawrence Erlbaum Associates.

Theory Acquisition as Stochastic Search

Tomer D. Ullman, Noah D. Goodman, Joshua B. Tenenbaum
{tomeru, ndg, jbt}@mit.edu
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology
Cambridge, MA 02139

Abstract

We present an algorithmic model for the development of children's intuitive theories within a hierarchical Bayesian framework, where theories are described as sets of logical laws generated by a probabilistic context-free grammar. Our algorithm performs stochastic search at two levels of abstraction – an outer loop in the space of theories, and an inner loop in the space of explanations or models generated by each theory given a particular dataset – in order to discover the theory that best explains the observed data. We show that this model is capable of learning correct theories in several everyday domains, and discuss the dynamics of learning in the context of children's cognitive development.

Introduction

As children learn about the world, they learn more than just a large stock of specific facts. They organize their knowledge into abstract coherent frameworks, or *intuitive theories*, that guide inference and learning within particular domains (Carey, 1985; Wellman & Gelman, 1992). Much recent work in computational cognitive modeling has attempted to formalize how intuitive theories are structured, used and acquired from experience (Tenenbaum, Griffiths, & Kemp, 2006), working broadly within a hierarchical Bayesian framework shown in Figure 1 (and explained in more detail below). While this program has made progress in certain respects, it has treated the problem of theory acquisition only in a very ideal sense. The child is assumed to have a hypothesis space of possible theories constrained by some "Universal Theory", and to be able to consider all possible theories in that space, in light of a given body of evidence. Given sufficient evidence, and a suitably constrained hypothesis space of theories, it has been shown that an ideal Bayesian learner can identify the correct theory underlying core domains of knowledge such as causality (Goodman, Ullman, & Tenenbaum, 2009), kinship and other social structures (Kemp, Goodman, & Tenenbaum, 2008). These Bayesian computational analyses have not to date been complemented by working algorithmic models of the search process by which a child can build up an abstract theory, piece by piece, generalizing from experience. Here we describe such an algorithmic model for Bayesian theory acquisition. We show that our algorithm is capable of constructing correct if highly simplified theories for several everyday domains, and we explore the dynamics of its behavior – how theories can change as the learner's search process unfolds as well as in response to the quantity and quality of the learner's observations.

At first glance, the dynamics of theory acquisition in childhood look nothing like the ideal learning analsyes of hierarchical Bayesian models – and may not even look particularly rational or algorithmic. Different children see different random fragments of evidence and make their way to adult-like intuitive theories at different paces and along different paths. It seems unlikely that children can simultaneously evaluate many candidate theories at once; on the contrary, they appear to hold just one theory in mind at any time. Transitions between theories appear to be local, myopic, and semi-random, rather than systematic explorations of the hypothesis space. They are prone to backtracking or "two steps forward, one step back". We suggest that these dynamics are indicative of a stochastic search process, much like the Markov chain Monte Carlo (MCMC) methods that have been proposed for performing approximate probabilistic inference in complex generative models. We show how a search-based learning algorithm can begin with little or no knowledge of a domain, and discover the underlying structure that best organizes it by generating new hypotheses and checking them against its current conceptions of the world using a hiearchical Bayesian framework. New hypotheses are accepted probabilistically if they can better account for the observed data, or if they compress it in some way. Such a search-based learning algorithm is capable of exploring a potentially infinite space of theories, but given enough time and sufficient data it tends to converge on the correct theory – or at least some approximation thereof, corresponding to a small set of abstract predicates and laws.

The plan of the paper is as follows. We first introduce our framework for representing and evaluating theories, based on first-order logic and Bayesian inference in a hierarchical probabilistic model that specifies how the theory's logical structure constrains the data observed by a learner. We then describe our algorithmic approach to theory learning based on MCMC search, using simulated annealing to aid convergence. Finally we study the search algorithm's behavior on two case studies of theory learning in everyday cognitive domains: the taxonomic organization of object categories and properties, and a simplified version of magnetism.

Formal framework

We work with the hierarchical probabilistic model shown in Figure 1, based on those in (Katz, Goodman, Kersting, Kemp, & Tenenbaum, 2008; Kemp et al., 2008). We assume that a domain of cognition is given, comprised of one or more systems, each of which gives rise to some observed data. The learner's task is to build a theory of the domain: a set of abstract concepts and explanatory laws that together generate a hypothesis space and prior probability distribution over candidate models for systems in that domain. The laws and

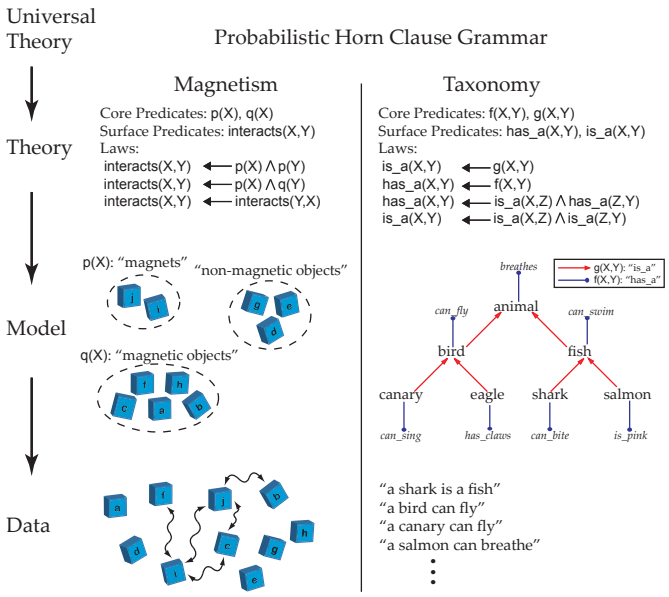


Figure 1: A hierarchical Bayesian framework for theory acquisition

concepts are written in logical form, a "language of thought", typically a subset of first-order logic. The learner's model of a system specifies what is true of that system, and thereby generates a probability distribution over possible observations that can be made for that system.

For example, consider a child learning about the domain of magnetism. She might begin by playing with a few pieces of metal and notice that some of the objects interact, exerting strange pulling or pushing forces on each other. She could describe the data directly, as "Object i interacts with object f", "Object i interacts with object j", and so on. Or she could form a simple theory, in terms of abstract concepts such as magnet, magnetic object and non-magnetic object, and laws such as "Magnets interact with other magnets", "Magnets interact with magnetic objects", and "Interactions are symmetric" (but no other interactions take place). Systems in this domain correspond to specific subsets of objects, such as the set of objects $\{\mathsf{a},\ldots,\mathsf{i}\}$ in Figure 1. A model of a system specifies the minimal facts needed to apply the abstract theory to the system, in this case which objects are magnetic, which are magnets, and which are non-magnetic. From these core facts the laws of the theory determine all other true facts – in our example, this means all the pairwise interactions between the objects: e.g., objects i and j, being magnets, should interact, but i and e should not, because e is non-magnetic. Finally, the true facts generate the actual data observed by the learner via a noisy observation process.

While the abstract concepts in this simplified magnetism theory are attributes of objects, more complex relations are possible. Consider for example a domain of taxonomy, as in Collins and Quillian's classic model of semantic memory as an inheritance hierarchy (Collins & Quillian, 1969). Here the abstract concepts are is_a relations between categories and has_a relations between categories and properties. The theory underlying taxonomy has two basic laws: "The has_a relation inherits down is_a relations" and "The is_a relation is transitive" (laws 3 and 4 on the right side of Figure 1). A system consists of a specific set of categories and properties, such as salmon, eagle, breathes, can fly, and so on. A model specifies the minimal is_a and has_a relations, typically corresponding to a tree of is_a relations between categories with properties attached by has_a relations at the broadest category they hold for: e.g., "A canary is a bird", "A bird is an animal", "An animal can breathe", and so on. The laws then determine that properties inherit down chains of is_a relations to generate many other true facts that can potentially be observed, e.g., "A canary can breathe".

Equipped with this hierarchical generative model, a learner can work backwards from observed data to multiple levels of latent structure. Given the correct theory, the learner can infer the most likely model underlying a set of noisy, sparse observations and predict facts that have not been directly observed (Katz et al., 2008; Kemp et al., 2008). If the true theory is unknown, the learner can consider a hypothesis space of candidate theories, generated by higher-level "Universal Theory (UT)" knowledge. UT defines a distribution over the space of possible theories, $P(T|UT)$, which can then be used by a learner to infer the correct theory describing a domain, according to the standard Bayesian formulation:

$$P(T|D,UT) \propto P(D|T)P(T|UT) \quad (1)$$

Bayes' rule here captures the intuition of Occam's razor. The theory that best explains the data, or has highest posterior probability $P(T|D,UT)$, should be based on two considerations: how well the theory fits the data, as measured by the likelihood $P(D|T)$, and how simple or short is the theory, as measured by the prior $P(T|UT)$. We now define these hypothesis spaces and probabilities more formally, and then describe a learning algorithm that searches the space of theories by proposing small random changes to the current theory and accepting changes stochastically based on whether they are likely to lead to higher overall probability.

A language for theories. Following (Katz et al., 2008) we represent the laws in a theory as Horn clauses: logical expressions of the form $t \leftarrow (p \wedge q \wedge \ldots \wedge r)$. Horn clauses express logical implications – a set of conjunctive conditions under which t holds – but can also capture intuitive causal relations under the assumption that any propositions not generated by the theory are assumed to be false. In our formulation, the clauses contain two kinds of predicates: "core" and "surface". Core predicates are a minimal set of predicates that determine all other predicates when combined with the theory's laws. Surface predicates are derived from other predicates, either surface or core, via the laws. Predicates may or may not be directly observable in the data. The core predicates can be seen as compressing the full model into just the minimal bits necessary to specify all true facts. In the magnetism example above, the core could be expressed in terms of two predicates $\mathsf{p}(\mathsf{X})$ and $\mathsf{q}(\mathsf{X})$. Based on an assignment of truth values to these core predicates, the

Top level theory

(S1)	S	$\Rightarrow$	(Law) $\wedge$ S
(S2)	S	$\Rightarrow$	(Tem) $\wedge$ S
(S3)	S	$\Rightarrow$	Stop

Random law generation

(Law)	Law	$\Rightarrow$	($P_{left} \leftarrow P_{right} \wedge$ Add)
(Add1)	A	$\Rightarrow$	P $\wedge$ Add
(Add2)	A	$\Rightarrow$	Stop

Predicate generation

(P_{left} 1)	P_{left}	$\Rightarrow$	$surface1()$
$\vdots$			
(P_{left} α)	P_{left}	$\Rightarrow$	$surface\alpha()$
(P_{right} 1)	P_{right}	$\Rightarrow$	$surface1()$
$\vdots$			
(P_{right} α)	P_{right}	$\Rightarrow$	$surface\alpha()$
($P_{right}(\alpha+1)$)	P_{right}	$\Rightarrow$	$core1()$
$\vdots$			
($P_{right}(\alpha+\beta)$)	P_{right}	$\Rightarrow$	$core\beta()$

Law templates

(Tem1)	Tem	$\Rightarrow$	template1()
$\vdots$			
(Temγ)	Tem	$\Rightarrow$	templateγ()

Figure 2: Production rules of the Probabilistic Horn Clause Grammar. S is the start symbol and Law, Add, P and Tem are non-terminals. α, β, and γ are the numbers of surface predicates, core predicates, and law templates, respectively.

learner can use the theory's laws such as $\mathsf{interacts}(X,Y) \leftarrow \mathsf{p}(X) \wedge \mathsf{q}(Y)$ to derive values for the observable surface predicate $\mathsf{interacts}(X,Y)$. Notice that $\mathsf{p}(X)$ and $\mathsf{q}(X)$ are abstract predicates, which acquire their meaning as concepts picking out magnets or magnetic objects respectively in virtue of the role they play in the theory's laws. In constructing such a theory the learner essentially creates new concepts (Carey, 1985). Entities may be typed and predicates restricted based on type constraints: e.g., in taxonomy, $\mathsf{has_a}(X,Y)$ requires that X be a category and Y be a property, while $\mathsf{is_a}(X,Y)$ requires that X and Y both be categories. Forcing candidate models and theories to respect these type constraints provides the learner with a valuable and cognitively natural inductive bias.

The theory prior $P(T|UT)$. We posit UT knowledge in the form of a probabilistic context-free Horn clause grammar (PHCG) that generates the hypothesis space of possible Horn-clause theories, and a prior $P(T|UT)$ over this space (Figure 2). This grammar and the Monte Carlo algorithms we use to sample or search over the theory posterior $P(T|D,UT)$ are based heavily on Goodman, Tenenbaum, Feldman, and Griffiths (2008), who introduced the approach for learning single rule-based concepts rather than the larger theory structures we consider here. We refer readers to Goodman et al. (2008) for many technical details. Given a set of possible predicates in the domain, the PHCG draws laws from a random construction process (Law) or from law templates (Tem; explained in detail below) until the Stop symbol is reached, and then grounds out these laws as horn clauses. The prior $p(T|UT)$ is

P(X,Y) $\leftarrow$ P(X,Z)$\wedge$P(Z,Y)	P(X,Y) $\leftarrow$ P(X)$\wedge$P(Y)
P(X,Y) $\leftarrow$ P(Z,X)$\wedge$P(Z,Y)	P(X,Y) $\leftarrow$ P(Y,X)
P(X,Y) $\leftarrow$ P(X,Z)$\wedge$P(Y,Z)	P(X,Y) $\leftarrow$ P(X,Y)
P(X,Y) $\leftarrow$ P(Z,X)$\wedge$P(Y,Z)	P(X) $\leftarrow$ P(X)
P(X,Y) $\leftarrow$ P(X,Y)$\wedge$P(X)	P(X) $\leftarrow$ P(X,Y)$\wedge$P(X)
P(X,Y) $\leftarrow$ P(Y,X)$\wedge$P(X)	P(X) $\leftarrow$ P(Y,X)$\wedge$P(X)
P(X,Y) $\leftarrow$ P(X,Y)$\wedge$P(Y)	P(X) $\leftarrow$ P(X,Y)$\wedge$P(Y)
P(X,Y) $\leftarrow$ P(Y,X)$\wedge$P(Y)	P(X) $\leftarrow$ P(Y,X)$\wedge$P(Y)

Figure 3: The list of templates available to in the PHCG.

the product of the probabilities of choices made at each point in this derivation. All these probabilities are less than one, so overall the prior favors simpler theories with shorter derivations. The precise probabilities of different rules in the grammar are treated as latent variables and integrated out, favoring re-use of the same predicates and law components within a theory (Goodman et al., 2008).

Law templates. We make the grammar more likely to generate useful laws by equipping it with templates, or canonical forms of laws that capture structure likely to be shared across many domains. While it is possible for the PHCG to reach each of these law forms without the use of templates, their inclusion allows the most useful laws to be invented more readily. They can also serve as the basis for transfer learning across domains. For instance, instead of having to reinvent transitivity anew in every domain with some specific transitive predicates, a learner could recognize that the same transitivity template applies in several domains. It may be costly to invent transitivity for the first time, but once found – and appreciated! – its abstract form can be readily reused. The specific law templates used are described in Figure 3. Each "$\mathsf{P}(\cdot)$" symbol stands for a non-terminal representing a predicate of a certain -arity. This non-terminal is later instantiated by a specific predicate. For example, the template $\mathsf{P}(X,Y) \leftarrow \mathsf{P}(X,Z) \wedge \mathsf{P}(Z,Y)$ might be instantiated as $\mathsf{is_a}(X,Y) \leftarrow \mathsf{is_a}(X,Z) \wedge \mathsf{is_a}(Z,Y)$ (a familiar transitive law) or as $\mathsf{has_a}(X,Y) \leftarrow \mathsf{is_a}(X,Z) \wedge \mathsf{has_a}(Z,Y)$ (the other key law of taxonomy, stating that "has_a is transitive over is_a").

The theory likelihood $P(D|T)$. An abstract theory makes predictions about the observed data in a domain only indirectly, via the models it generates. A theory typically generates many possible models: even if a child has the correct theory and abstract concepts of magnetism, she could categorize a specific set of metal bars in many different ways, each of which would predict different interactions that could be observed as data. Expanding the theory likelihood,

$$P(D|T) = \sum_M P(D|M)P(M|T) \tag{2}$$

we see that theory T predicts data D well if it assigns high prior $P(M|T)$ to models M that make the data probable under the observation process $P(D|M)$.

The model prior $P(M|T)$ reflects the intuition that a theory T explains some data well if T requires few additional degrees of freedom beyond its abstract concepts and laws to make its predictions. That is, few specific and contingent facts about the system under observation are required

in addition to the theory's general prescriptions. This intuition is captured by a prior that encourages the core predicates to be as sparse as possible, penalizing theories that can only fit well by "overfitting" with many extra degrees of freedom. Formally, following (Katz et al., 2008), we model all values of the core predicates as independent Bernoulli random variables with conjugate beta priors encouraging most variables to have the same value (on or off). We assume that any proposition potentially in the model M is false unless it is a core predicate turned on by this Bernoulli process or is derived from the core predicates through the theory's laws (the *minimal model* assumption of logic programming).

Finally, the model likelihood $P(D|M,T)$ comes from assuming that we are observing randomly sampled true facts (sampled with replacement, so the same fact could observed on multiple occasions), which also encourages the model extension to be as small as possible.

Stochastic search in theory space: a grammar-based Monte-Carlo algorithm. Following Goodman et al. (2008), we use a grammar-based Metropolis-Hastings (MH) algorithm to sample theories from the posterior distribution over theories conditioned on data, $P(T|D,UT)$. This algorithm is applicable to any grammatically structured theory space, such as the one generated by our PHCG. The MH algorithm proceeds by randomly proposing changes to the current theory, and accepting or rejecting these changes. Each proposed change to the current theory corresponds to choosing a grammatical constituent of the theory then regenerating it from the PHCG. For example, if our theory of magnetism includes the law $\mathsf{interacts}(\mathsf{X},\mathsf{Y}) \leftarrow \mathsf{p}(\mathsf{X}) \wedge \mathsf{q}(\mathsf{Y})$, the MH procedure might propose to add or delete a predicate (e.g., $\mathsf{interacts}(\mathsf{X},\mathsf{Y}) \leftarrow \mathsf{p}(\mathsf{X}) \wedge \mathsf{q}(\mathsf{Y}) \wedge \mathsf{p}(\mathsf{Y})$ or $\mathsf{interacts}(\mathsf{X},\mathsf{Y}) \leftarrow \mathsf{p}(\mathsf{X})$), to change one predicate to an alternative of the same form (e.g., $\mathsf{interacts}(\mathsf{X},\mathsf{Y}) \leftarrow \mathsf{p}(\mathsf{X}) \wedge \mathsf{p}(\mathsf{Y})$) or a different form if available (e.g., $\mathsf{interacts}(\mathsf{X},\mathsf{Y}) \leftarrow \mathsf{p}(\mathsf{X}) \wedge \mathsf{r}(\mathsf{X},\mathsf{Y})$); to resample the law from a template (e.g., $\mathsf{interacts}(\mathsf{X},\mathsf{Y}) \leftarrow \mathsf{r}(\mathsf{X},\mathsf{Z}) \wedge \mathsf{r}(\mathsf{Z},\mathsf{Y})$); or to add or delete a whole law.

These proposals are accepted with probability equal to the minimum of 1 and the MH acceptance ratio,

$$\frac{P(T'|D,UT)}{P(T|D,UT)} \cdot \frac{Q(T|T')}{Q(T'|T)} \tag{3}$$

where T is the current theory, T' is the new proposed theory, and $Q(\cdot|\cdot)$ is the transition probability from one theory to the other, derived from the PHCG (Goodman et al., 2008). To aid convergence we raise the posterior ratio to a power greater than 1, which we increase very slightly after each MH step in a form of simulated annealing. The learner initially explores alternative theories freely, but with time becomes increasingly likely to reject theory changes unless they lead to an improved posterior probability.

While this MH algorithm could be viewed merely as a way to approximate the calculations necessary for a hierarchical Bayesian analysis, we suggest that it could also capture in a schematic form the dynamic processes of theory acquisition and change in young children. Stochastic proposals to add a new law or change a predicate within an existing law are consistent with some previous characterizations of children's theory learning dynamics (Siegler & Chen, 1998). These dynamics were previously proposed on purely descriptive grounds, but here they emerge as a consequence of a rational learning algorithm for effectively searching an infinite space of logical theories.

Approximating the theory score. Computing the theory likelihood $P(D|T)$, necessary to compare alternative theories in Equation 3, requires a summation over all possible models consistent with the current theory (Equation 2). Because this sum is typically very hard to evaluate exactly, we approximate $P(D|T)$ with $P(D|M^*)P(M^*|T)$, where M^* is an estimate of the maximum a-posteriori (MAP) model inferred from the data: the most likely values of the core predicates. The MAP estimate M* is obtained by running a Gibbs sampler over the values of the core predicates, as in (Katz et al., 2008), annealing slightly on each Gibbs sweep to speed convergence and lock in the best solution. The Gibbs sampler over models generated by a given theory is thus an "inner loop" of sampling in our learning algorithm, operating within each step of an "outer loop" sampling at a higher level of abstract knowledge, the MH sampler over theories generated by UT knowledge.

Case Studies

We now explore the performance of this stochastic approach to theory learning in two case studies, using simulated data from the domains of taxonomy and magnetism introduced above. We examine the learning dynamics in each domain and make more explicit the possible parallels with human theory acquisition.

Taxonomy

Katz et al. (2008) defined a similar hierarchical Bayesian framework and showed that a theory of taxonomic reasoning about properties and categories in an inheritance hierarchy could be correctly selected from among several alternatives, on the basis of data. However, they did not address the harder challenge of constructing the theory from the ground up, or selecting it from an effectively infinite hypothesis space of theories (which could be used to describe many other domains). That is our goal here. Following Katz et al. (2008), we take the correct theory to have two unobservable core predicates, $\mathsf{g}(\mathsf{X},\mathsf{Y})$ and $\mathsf{f}(\mathsf{X},\mathsf{Y})$, and two observable surface predicates, $\mathsf{is_a}(\mathsf{X},\mathsf{Y})$ and $\mathsf{has_a}(\mathsf{X},\mathsf{Y})$. There are four laws:

Law 1: $\mathsf{has_a}(\mathsf{X},\mathsf{Y}) \leftarrow \mathsf{f}(\mathsf{X},\mathsf{Y})$

Law 2: $\mathsf{is_a}(\mathsf{X},\mathsf{Y}) \leftarrow \mathsf{g}(\mathsf{X},\mathsf{Y})$

Law 3: $\mathsf{has_a}(\mathsf{X},\mathsf{Y}) \leftarrow \mathsf{is_a}(\mathsf{X},\mathsf{Z}) \wedge \mathsf{has_a}(\mathsf{Z},\mathsf{Y})$

Law 4: $\mathsf{is_a}(\mathsf{X},\mathsf{Y}) \leftarrow \mathsf{is_a}(\mathsf{X},\mathsf{Z}) \wedge \mathsf{is_a}(\mathsf{Z},\mathsf{Y})$

Laws 1 and 2 set up the core predicates to represent the mini-

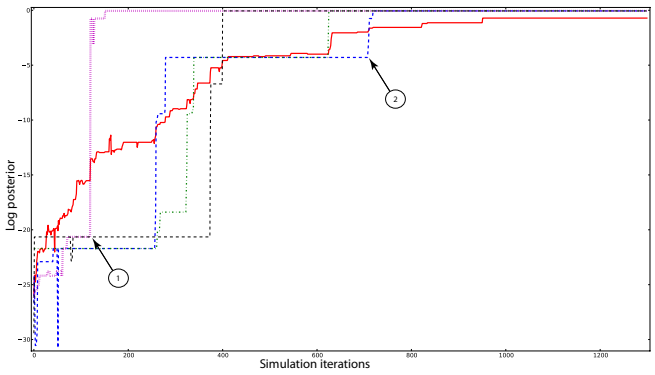


Figure 4: Log posterior score for representative runs of theory learning in Taxonomy. Dashed lines show different runs. Solid line is the average across all runs. Node 1 marks the acquisition of law 3, node 2 marks the acquisition of law 4.

mal is_a and has_a links, on top of which are defined the laws of property inheritance (Law 3) and transitive category membership (Law 4). We take laws 1 and 2 as given, assuming the structure and meaning of the core predicates as Katz et al. did, and ask whether a learner can successfully construct laws 3 and 4. Following Katz et al., we consider a concrete domain with 7 categories and 7 properties in a balanced taxonomy, as shown in Figure 1. Observations include all positive facts asserting that a property is true of a category, as in "An eagle has claws". (The data used for this section and the following case study can be found at http://web.mit.edu/tomeru/www/tlss.) We ran 10 simulations for 1300 iterations of the outer MH loop. Learning curves for representative runs as well as the average over all runs are shown in Figure 4. Out of 10 simulations, 8 found the correct theory within the given number of iterations, and 2 discovered a partial theory which included only law 3 (property inheritance). Several observations are worth noting.

Abstract learning is possible. Using only stochastic local search moves, a learner can navigate the space of potential theories to discover the laws underlying the domain. Even a relatively small dataset (with 7 categories and 7 properties) is sufficient to learn the correct abstract domain theory.

Individual learning curves show sudden changes and high variability in what is learned when, while on average learning is smooth and follows a characteristic timecourse. The learning algorithm's local dynamics are highly stochastic and variable across runs, because of the randomness in what theory changes are proposed when, and the fact that a small theory change can make a big difference in predictive power. Yet there is still a meaningful sense in which we can talk about "typical" learning behavior, even though any one learner may not look much like this average. If stochastic local search is a key component in children's theory construction, it could explain why cognitive development shows this same dual nature: systematic and graded progression at the population level, despite random, discontinuous and highly variable learning rates in any one child.

Although proposals are random, there is a systematic and rational order to learning. While there are many routes through theory space to a given endpoint, a sequence of random MH proposals may still prefer some orders of knowledge acquisition over others. Here, when law 4 is discovered (on 8/10 runs), it is always acquired after law 3. This is because law 4 (transitivity of category membership) provides much more explanatory power – and hence is more stable under our stochastic theory-learning dynamics – given law 3 (property inheritance) and a reasonable domain model specifying which properties hold for which categories. This order is also consistent with the order of acquisition in human cognitive development (Wellman & Gelman, 1992): children learn to generalize properties of biological categories to instances well before they learn that categories can be arranged in a multilevel hierarchy supporting transitive inferences of category membership.

Magnetism

After showing that stochastic search can learn the correct laws in a domain theory, we now consider a second case study in which the acquisition of new laws corresponds to a shift in the meaning of the core predicates, and new (i.e., previously unassigned) core predicates are introduced during learning – akin to some of the conceptual changes described by Carey (1985). Our domain here is the simple version of magnetism described above, with two unobservable core predicates: $\mathsf{p}(\mathsf{X})$ and $\mathsf{q}(\mathsf{X})$, and one observable surface predicate: $\mathsf{interacts}(\mathsf{X},\mathsf{Y})$. There are three laws:

$$\text{Law 1:} \quad \mathsf{interacts}(\mathsf{X},\mathsf{Y}) \leftarrow \mathsf{p}(\mathsf{X}) \wedge \mathsf{p}(\mathsf{Y})$$
$$\text{Law 2:} \quad \mathsf{interacts}(\mathsf{X},\mathsf{Y}) \leftarrow \mathsf{p}(\mathsf{X}) \wedge \mathsf{q}(\mathsf{Y})$$
$$\text{Law 3:} \quad \mathsf{interacts}(\mathsf{X},\mathsf{Y}) \leftarrow \mathsf{interacts}(\mathsf{Y},\mathsf{X})$$

We consider a concrete system with 3 magnets, 5 magnetic objects and 2 non-magnetic objects. These concepts are initially unknown to the learner. The core predicates $\mathsf{p}(\mathsf{X})$ and $\mathsf{q}(\mathsf{X})$ are completely abstract and initially uninterpreted. They will acquire their meaning as concepts picking out magnets and magnetic objects respectively in virtue of the role they play in the theory's laws, specifying that objects in one subset (the p's) interact with each other and with objects in a second set (the q's), but q's do not interact with each other. In constructing a theory, the learner introduces these abstract predicates via new laws, or new roles in existing laws, and thereby essentially creates these concepts where she did not have them before (Carey, 1985).

We ran 10 simulations for 1600 iterations of the outer MH loop. Representative runs are displayed in Figure 5, as well as the average over all the runs. The results were similar to the taxonomy case study in several respects, which we also expect to hold for a variety of other domains. The correct theory was usually learned, with some variation: 9/10 simulations found the correct theory or a variant of it, and one discovered a partial theory containing only law 1. Only some runs learned the exact form of law 3, asserting that interactions are symmetric. Others found variants that were extensionally equivalent to symmetry in this domain, but slightly more

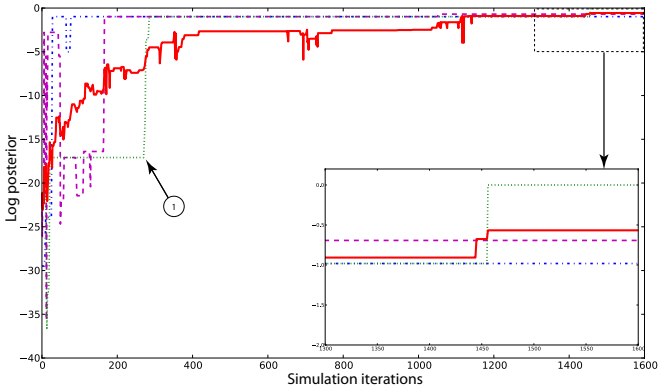


Figure 5: Representative runs of theory learning in Magnetism. Dashed lines show different runs. Solid line is the average across all runs. Node 1 marks the acquisition of law 1 and the confounding of magnets with magnetic objects. Lower right panel zooms into the end of the simulation, showing acquisition of the final correct theory.

complex in their logical form. Individual runs of learning showed discrete jumps with high variability, while average-case behavior was smooth, with systematic order effects. Law 3 is never learned first, because alone it has no explanatory power. Either law 1 or the combination of laws 2 and 3 tend to be learned first, followed by the other, although sometimes laws 1 and 2 are learned first, followed by law 3. Law 1 tends to be learned first overall because it is most likely under the prior (which is also the proposal distribution for local search moves), and also because, as explained below, it represents a reasonable first approximation to the domain's structure.

The algorithm's learning dynamics in this case study are particularly interesting for how they parallel key transitions in childrens' cognitive development: restructuring or construction of new concepts, as when one concept differentiates into two (Carey, 1985). When our simulations of learning about magnetism construct law 1 first, without laws 2 and 3, they find a simpler theory capturing many of the observed facts at the cost of over-generalizing. That is, under law 1 alone, the optimal setting of the core predicates – the most probable model – equates magnets and magnetic objects, making $p(X)$ true for both. This is a good first approximation, even as it collapses two categories of objects with fundamentally different causal properties: the generators of magnetic force (the "magnets") and the objects on which that force acts (the "magnetic objects"). Only once all three laws have been constructed does the learner come to distinguish between magnets and magnetic objects, reflected in the difference between the roles played by the two core predicates $p(X)$ and $q(X)$. Only once law 2 is available does the learner have reason to restrict the extension of $p(X)$ to just magnets, excluding other magnetic objects.

Conclusion and Future Directions

We have presented an algorithmic model of theory acquisition as stochastic search in a hierarchical Bayesian framework and explored its dynamics in two case studies. We were encouraged by the general pattern of successes on these examples and by several qualitative parallels with phenomena of human cognitive development. These results suggest that previous ideal learning analyses of Bayesian theory acquisition can be realized approximately by algorithms that are cognitively plausible for child learners, and indeed potentially descriptive of the dynamics of development.

Previous hierarchical Bayesian analyses of learning abstract knowledge have focused on the role of accumulating data in driving changes to the learner's hypotheses (Kemp & Tenenbaum, 2008). In contrast, here we have focused on how changes to the learner's theories and abstract concepts are driven by a different source, the stochastic dynamics of the learning algorithm. Data-driven and algorithm-driven theory change can have a similar character, first discovering simpler, rougher approximations to reality and then refining those to more complex, accurate representations; sometimes changing by adjusting small details, but other times by making large qualitative transitions or discoveries. In future work we plan to explore further the similarities, differences and interactions between these two drivers of learning dynamics, both in computational analyses and experimental work. We hope to establish tighter quantitative correspondences with human learning curves in development, as well as with controlled laboratory studies of theory learning in adults, where some of the same mechanisms might be at work. We will also consider a broader range of algorithmic approaches, stochastic as well as deterministic, evaluating them both as behavioral models and as effective computational approximations to the theory search problem for larger domains.

Acknowledgments. We thank Yarden Katz for valuable discussions. This work was funded in part by grants from the McDonnell Causal Learning Collaborative, AFOSR (FA9550-07-1-0075), and ONR (N00014-09-0124).

References

Carey, S. (1985). *Conceptual change in childhood*. Cambridge, MA: MIT Press/Bradford Books.

Collins, A., & Quillian, M. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, *8*, 240–247.

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, *32*(1), 108—154.

Goodman, N. D., Ullman, T. D., & Tenenbaum, J. B. (2009). Learning a theory of causality. In *Proceedings of the 31st annual conference of the cognitive science society.*

Katz, Y., Goodman, N. D., Kersting, K., Kemp, C., & Tenenbaum, J. B. (2008). Modeling semantic cognition as logical dimensionality reduction. *Proceedings of the Thirtieth Annual Conference of the Cognitive Science Society*.

Kemp, C., Goodman, N. D., & Tenenbaum, J. B. (2008). Theory acquisition and the language of thought. In *Proceedings of thirtieth annual meeting of the cognitive science society.*

Kemp, C., & Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences*, *105*(31), 10687–10692.

Siegler, R. S., & Chen, Z. (1998). Developmental differences in rule learning: A microgenetic analysis. *Cognitive Psychology*, *36*, 273-310.

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, *10*, 309–318.

Wellman, H. M., & Gelman, S. A. (1992). Cognitive development: Foundational theories of core domains. *Annual Review of Psychology*, *43*, 337-375.

Is it me or the world? 16-month-olds distinguish competing hypotheses about the cause of failed interventions

Hyowon Gweon (hyora@mit.edu) and Laura E. Schulz (lschulz@mit.edu)

Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology
Cambridge, MA 021 USA

Abstract

When an agent fails to make an object function properly, there are two possibilities: the agent did something wrong or something is wrong with the object. As in all problems of confounding, these hypotheses can be disambiguated by varying one factor and holding the other constant: in this case, either by holding the object constant and varying the agent (e.g., by asking for help from others) or by holding the agent constant and varying the object (e.g., by trying another object). Here we show that 16-month-old infants engage in distinct patterns of behavior depending on the relative probability of the competing hypotheses: they ask for help more often when they (rather than the object) are the probable cause of failure; they reach for a new object more often when the object (rather than themselves) is the probable cause of failure.

Keywords: infants, confounding, exploratory behavior, ambiguous evidence, asking for help.

Imagine that you are trying to get into a new office but the key doesn't work. You jiggle it for a minute and then assess your options. What you choose to do next depends on what hypothesis you think is most probable: if you think you are having trouble positioning the key, you might ask a friend for help; if you think you picked up the wrong key, you will probably try a different one.

As intentional agents, we frequently plan and carry out goal-directed actions. Most of the time, these actions are successful. However, when we experience failure, we can experience not only the frustration of our intentions but also a problem of confounded evidence. Did we do something wrong or was something wrong in the world?

This problem of "me or the world" is perhaps the most common example of confounded causal variables that we encounter in everyday life. Of course, the variables "me" and "the world" can sometimes be more precisely defined. In the key example for instance, if you are the problem, you might have put the key in upside down, turned it in the wrong direction, or lack fine motor coordination. Alternatively, if you believe the problem lies in the world, the door might be jammed, or the lock might have been changed. However, these distinctions are subordinate to the primary problem of discovering whether you or the world is the culprit. When things go wrong, how do we identify the locus of failure?

As in any causally confounded situation, changing one variable at a time can disambiguate the evidence. Assuming that changing either variable is possible and equally costly, a rational agent who wants to generate the effect should change the variable that seems most likely to be the source of the failure. If I think I'm the problem, I should hold the object constant and vary the agent (e.g., ask my friend to help); if I think the object poses the problem, I should hold myself constant and vary the object (e.g., try a new key).

There are many reasons to believe that recognizing, let alone solving, problems of confounding between the self and the world might require substantial expertise. Indeed, previous research suggests that both children and adults have difficulty recognizing when information is ambiguous (Penner & Klahr, 1996) and designing experiments that could generate informative evidence (Chen & Klahr, 1999; Koslowski, 1996; Kuhn, 1989). Such studies of formal scientific reasoning however, typically involve many hypotheses, including those that conflict with the learners' prior beliefs. In contexts where there are only two competing hypotheses and both are familiar and plausible, children seem to be sensitive to confounding at a much younger age (Gweon & Schulz, 2008; Kushnir & Gopnik, 2005; Schulz & Bonawitz, 2007; Sodian, Zaitchik, & Carey, 1991). Thus in simple cases, even very young children might be sensitive to competing hypotheses.

While a sensitivity to confounded variables might enable young children to recognize the "me vs. the world" problem, deconfounding the self and the world requires children to understand *how* to intervene on each variable. For example, imagine a simple confounded situation where a child tries a novel toy and fails to make it work; the child needs to understand that other people ('the agent' variable), and other toys of the same kind ('the world' variable), can both serve as useful sources of information. Here we briefly review some previous studies that suggest that even very young children might be capable of such an understanding.

A large body of literature on social referencing in infancy suggests that infants readily treat their caregivers as sources of information about the emotional valence of events (Sorce, Emde, Campos, & Klinnert, 1985; Walden & Ogan, 1988) and the referent of adults' attention (Baldwin, 1993; Carpenter, Nagell, Tomasello, Butterworth, & Moore, 1998). Moreover, infants use the information to regulate their own behavior. In particular, O'Neill (1996) showed that two-year-olds will request help from a knowledgeable (but not ignorant) parent in retrieving a hidden object, suggesting that toddlers not only look to parents for the information they might provide but also actively solicit such information.

Children's imitation of object-directed actions is also often interpreted as an indication that children perceive others as agents like themselves (the 'like me' hypothesis; Meltzoff & Brooks, 2001) and use adult actions for information about how to interact with an object (Gergely, Bekkering, & Kiraly, 2002; Gopnik & Meltzoff, 1994). Notably, children are more likely to imitate an adult's goal-directed action if they themselves have previously failed to generate a target outcome than if they have succeeded (Williamson, Meltzoff, & Markman, 2008) which suggests that young children can use adult actions as evidence about the cause of their own failures, and modify their own actions accordingly.

Such studies speak to children's understanding of other agents as potential sources of information about objects in the world. What about children's understanding that one object can be informative about other members of the object kind? Previous research has shown that preschoolers generalize non-obvious properties (like squeaking or magnetism) from one member of a kind to others (Gopnik & Sobel, 2000; Nazzi & Gopnik, 2000). Moreover, children maintain this expectation even when one exemplar fails to function as expected (Schulz, Standing, & Bonawitz, 2008). Indeed, 9-month-old infants can generalize a property of an object to other identical-looking object after a single exposure (Baldwin, Markman, & Melartin, 1993), and by 15 months infants can even integrate information about how the exemplars are sampled in their inferences about object properties (Gweon, Tenenbaum, & Schulz, 2010). These studies establish that children expect object properties to generalize across similar-looking objects, maintain that expectation even when they themselves fail to elicit the expected property, and, having experienced failure, can both solicit help from caregivers and act on other similar objects. This study however, is the first to investigate the possibility that infants might be sensitive to competing hypotheses for why their actions fail, and might rationally trade-off actions directed towards agents and actions directed towards objects. Here we ask whether 16-month-old infants implicitly recognize the ambiguity in a failed attempt to activate a toy. We predict that children should be more likely to ask for help when they themselves are the probable source of failure and more likely to test another toy when the probable cause of failure lies in the toy itself.

Experiment

In the current study, we introduce infants to three identical-looking toys, differing only in color (Green, Yellow, and Red). The experimenter shows the child that she can push a button on the Green toy and the toy will make music. In the Agent condition, the experimenter then hands the child the Green toy; in the Object condition, the experimenter hands the child the Yellow toy. All the children are allowed to try to activate their toy. However, because the Green toy is actually activated by a hidden switch and the Yellow toy is inert, the toys never activate for the infants.

The condition manipulation is designed to affect the relative probability of the two hypotheses for why the toy fails to activate. In the Agent condition, the hypothesis that the toy is broken is relatively improbable given that the toy had just worked moments before; the hypothesis that the child herself is doing something wrong should seem more probable. By contrast, in the Object condition, where the child's toy has never activated, the hypothesis that the toy doesn't work should seem more probable than the hypothesis that the child herself is doing something wrong (given that the button is conspicuous and easy to press).
Thus, we created a situation in which infants might differentially weigh the two hypotheses about the cause of the failure. In both cases however, infants had identical sources of information that they could use to resolve the ambiguity. All children were seated next to their parents. By turning and asking their parent for help with the toy they had, they could test the 'agent' variable. All children could also reach for the Red toy, which sat on the end of a piece of felt cloth. By pointing to the Red toy or pulling the piece of felt cloth they could try to retrieve a toy of the same kind and test 'the object' variable. (We placed the Red toy at a distance to ensure that all infants would initially attend to the toy they were given.) Because previous research indicates that infants reliably understand the intentional structure of action in a cloth-pulling sequence by the age of 12 months (e.g., Sommerville & Woodward, 2005), we recruited infants slightly older than this age and verified in a warm-up period that they can pull the cloth to retrieve a toy.

We hypothesized that the infants' behavior would be sensitive to the relative probability of the competing hypotheses. Therefore, we predicted that infants in the Agent condition should be more likely to appeal to their parents; infants in the Object condition should be more likely to reach for the other toy.

Methods

Participants Thirty infants (mean: 16 months, 10 days; range: 14 – 20 months; 47% girls) were recruited from a local children's museum; infants were randomly assigned to an Agent condition or an Object condition (n = 15/condition). Six infants were replaced due to parental interference or experimental error. Two additional infants were replaced because they did not pull the cloth to retrieve a toy during the warm-up procedure. (See Procedure.) Finally, two infants (one in each condition) were excluded from analyses because they never showed any of the target behaviors. (See Results.)

Materials One commercially available toy (a plastic fish) was used during the warm-up period. Three similar-looking novel toys were built by attaching a wooden stick (10 cm in length) to a round plastic container (4 inches in diameter). The toys resembled small hand drums with handles. A square-shaped button (2 x 2 x 1 cm) was attached to the top of the container. This button was inert. Each object was covered with green, red, or yellow electrical tape and felt

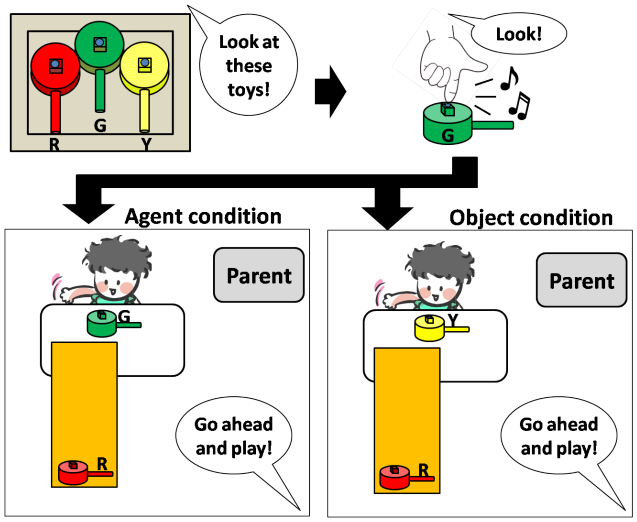


Figure 1: Schematic drawing of the experimental setup and procedure. R (Red), G(Green), Y (Yellow) refer to the color of the toys.

cloth. The Green toy had a small battery-powered circuit that was operated by a hidden switch at the bottom of the container: when the toy was laid flat on a hard surface and the fake button was pressed down, the real switch depressed and the toy played a musical tune (creating the appearance that pushing the fake button activated the toy). Children sat in a highchair. The tray on the high chair was covered with white felt, creating a surface that was too soft to activate the real switch at the bottom of the Green toy. The Green toy never worked on this tray when the fake button was pressed. The Red and Yellow toys did not have a musical mechanism inside, but contained play-dough so that all three toys were matched in approximate weight.

Procedure All children were tested individually in a quiet room inside the museum. The children sat in the highchair and the parents sat next to them on a chair. (See Figure 1 for experimental setup and stimuli.) Parents were instructed not to interact with the toys and only to smile and nod if the child addressed them. They were given a brochure about the study and asked to read it during the experimental procedure. Once the child was positioned in the highchair, the experimenter put a piece of orange felt cloth (approx. 20 x 75 cm) on the table and placed the warm-up toy on one end of the cloth. She pulled the cloth towards herself and retrieved the toy. Then she encouraged the infant to pull the cloth. Infants who did not pull the cloth and retrieve the toy after two demonstrations were excluded from analysis and replaced.

The experimenter removed the warm-up toy and introduced the child to a basket containing the Green, Red, and Yellow toys. She took the Green toy out, put it on the table, and pressed the button on top of the toy to play the music. She demonstrated this three times. Then she showed the child the basket containing the other two toys. She took out the Red toy and placed it on one end of the felt cloth. The toy was approximately 70 centimeters away from the child and was not within direct reach of the child's hands. She placed the other end of the felt cloth on the child's tray within easy reach of the child. Then, the experimenter handed the child either the Green toy (Agent condition) or the Yellow toy (Object condition) and said, "Here you go, you can go ahead and play!" She took the basket with the remaining toy (the Yellow toy in the Agent condition: the Green toy in the Object condition) out of the child's line of sight. The child's behavior was videotaped for 90 seconds (24 children) or until the child fussed-out (6 children); all but one of the infants who stopped playing before 90 seconds played for at least 60 seconds. The remaining infant was in the Object condition and played for 35 seconds. There was no difference between conditions in children's mean length of free play (Agent Condition: mean 89 seconds; Object Condition: mean 84 seconds, $p = ns$).

Results

For our preliminary analyses, we looked at whether all the children imitated the experimenter's action on the toy and whether they were equally persistent in the Agent condition (where they were given the same toy on which the action had been modeled) and the Object condition (where they had to make an inductive generalization from the Green toy to the Yellow toy). Given previous research suggesting that even 9-month-olds readily make such generalizations (Baldwin et al., 1993), we did not expect any difference in their button-pushing behavior. Indeed, all but one infant immediately (within two seconds) pressed the inert button on the toy in front of them. There was no difference in the frequency of children's button-pushing attempts in the two conditions (Agent Condition: mean 3.0 times; Object Condition: mean 3.2 times, $p = ns$).

We also used two different measures to look at whether parents differentially cued the infants to ask for help in the two conditions. Two coders, blind to hypotheses and conditions, coded from videotape; the monitor was partially covered with a cardboard occluder so that only the parent was visible. One coder was asked to make Yes/No judgments about whether the parent ever encouraged the child to ask for help. A second coder rated parents' attempts to initiate communication on a scale from 1 (no attempts to communicate) to 7 (repeated attempts to communicate). There was no difference between conditions on either measure (% of Yes judgments = 7% in both conditions; mean rating: 1.71 (Agent) vs. 1.79 (Object), $p = ns$).

The primary measure of interest was whether children's first response to failure was directed towards their parents or to the other toy. To determine this, we coded three target behaviors: Ask, Point, and Pull. The criteria for coding a behavior as Ask was that the child turned to the parent[1] and

[1] The infants could also have asked the experimenter for help. However, the experimenter stood behind the high chair during the free play period and acted busy (i.e., by writing something on a clipboard). Therefore, although there were a few cases where the infants looked as if they wanted the experimenter's attention, we did not include these attempts as one of our target behaviors.

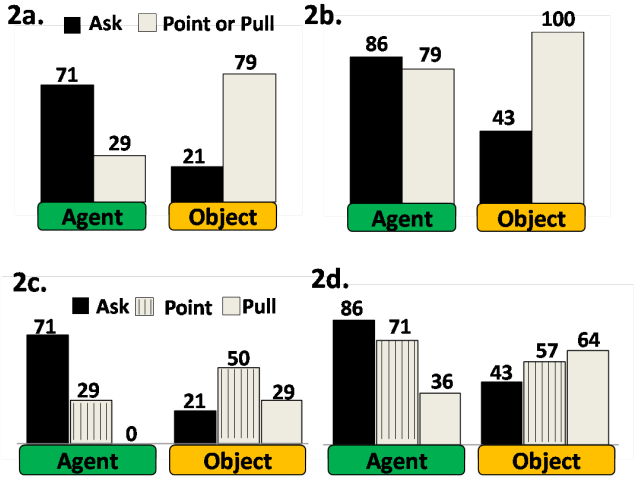


Figure 2: Experimental results. 2a. % infants Asking *first* or Pointing/Pulling *first* in each condition. 2b. % infants *ever* Asking or Pointing/Pulling in each condition (infants can perform more than one behavior, so the combined percentage within a condition can exceed 100).
2c and 2d. Same as 2a and 2b, respectively, but showing Point and Pull separately.

tried to give her the inoperative toy or grabbed the parent's hand and tried to bring it towards the inoperative toy (the Green toy in the Agent condition; the Yellow toy in the Object condition). We coded Point as a finger point to the Red toy or a direct reach for the Red toy. We coded Pull as pulling on the cloth and successfully retrieving the Red Toy. We predicted that children's first response to failure would be to Ask for help in the Agent condition but to Point or Pull in the Object condition. Additionally we coded whether infants ever showed the three target behaviors during free play, the latency to their first target behavior, and what they did with the Red toy if they retrieved it. The data for three target actions were originally coded by the experimenter, but also coded by an observer blind to hypotheses and conditions. The inter-coder reliability was high (Cohen's kappa = 92.6); using the data from the blind-coder did not change the results. All but two infants (one in each condition) performed at least one of the three target behaviors during the course of their free play; these two children were eliminated from subsequent analyses.

As predicted, children were significantly more likely to Ask First than Point or Pull first in the Agent condition than the Object condition (χ^2 (1, N = 28) = 7.04, $p < 0.01$). In the Agent condition, 10 infants (71%) Asked first and 4 infants (29%) Pointed or Pulled first; in the Object condition, 3 infants Asked first (21%); 11 infants (79%) Pointed or Pulled first. Within conditions, children were marginally more likely to Ask first than Point/Pull in the Agent condition ($p < 0.10$ by binomial test) and more likely to Point/Pull first than Ask in the Object condition ($p < 0.05$ by binomial test). See Figure 2a.

We also looked at how many infants in each condition exhibited each of the target behaviors at least once during

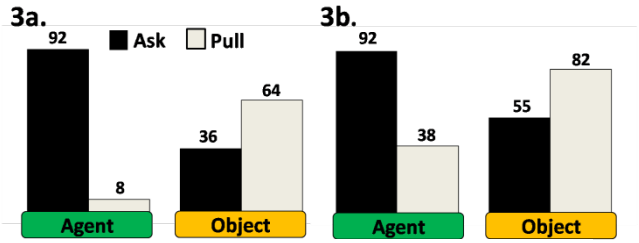


Figure 3: Data excluding Point behavior. 3a. % infants who Asked or Pulled *first* in each condition. 3b: % infants who *ever* Asked or Pulled during free play in each condition.

the course of their free play. Children were more likely to Ask for help over the course of their free play in the Agent condition than the Object condition (χ^2 (1, N = 28) = 5.6, $p < 0.05$, see Figure 2b.). Twelve infants (86%) in the Object condition Asked at some point; only 6 infants (43%) did so in the Object condition. Similarly, there was a trend for children to be more likely to Point to or Pull the red object in the Object condition than the Agent condition: 11 infants (79%) Pointed/Pulled at some point in the Agent condition whereas all 14 infants (100%) did so in the Object condition (χ^2 (1, N = 28) = 3.36, $p = .07$). See Figures 2c and 2d for the first target action and any instance of the target actions broken down by each of the three target behaviors.

There was no difference between conditions in the mean latency to the first target action (Agent condition, mean: 19.3 s; Object condition, mean: 25.4 s, $p = ns$). This suggests that the children in the two conditions were approximately matched in their motivation to act. There was also no difference in latency between the agent-directed and object-directed actions (Ask: 20.5 s; Point/Pull: 26.6 s, $p = ns$). This suggests that the agent-directed and object-directed actions were equivalently easy for the children to perform.

Although there was no overall latency difference, the Pulling action occurred (non-significantly) later than the Ask or Point actions (because most children in both conditions pointed before they pulled). Prima facie, Point is a less complex action than either Ask or Pull. Point required only a finger movement whereas Ask required the child to try to hand the object to the parent or to try to place the parent's hand on the object and Pull required a means/end sequence. We believe the collapsed Point/Pull measure is the correct measure of children's interest in the distal object as there is little doubt from the videotapes that infants coded as Pointing were unambiguously asking for the Red toy. However, to match for the overall complexity of the action sequence, we looked at whether infants were more likely to Ask or Pull first if the Point measure is excluded. Under this analysis, and excluding infants whose only target behavior was pointing (one child in the Agent condition; three children in the Object condition) infants were more likely to Ask than Pull in the Agent condition compared to the Object condition (χ^2 (1, N = 24) = 11.7, $p < 0.001$; see Figure 3a). Within conditions, infants in the Agent condition were more likely to Ask than Pull (12 Ask first, 1 Pull first; $p < 0.01$ by binomial test); infants in the Object condition were equally

likely to Ask and Pull (4 Ask first, 7 Pull first p = ns, by binomial test). Looking at any instance of Asking or Pulling over the course of free play, infants in the two conditions again tended to show different patterns of behavior (χ^2 (1, N = 24) = 3.03, p = .08; Agent condition: 12 Ask overall, 5 Pull overall; Object condition: 6 Ask overall, 9 Pull overall; See Figure 3b).

Finally, of those infants who pulled the cloth and successfully retrieved the Red toy (5 infants in the Agent condition; 9 agents in the Object condition), all but one immediately (within 2 seconds) pressed the button on the Red toy, suggesting that infants did indeed retrieve the toy in order to see whether they could make the toy go.

Discussion

These results suggest that when an object fails to function as expected, 16-month-old infants entertain competing hypotheses about the cause of the failure and act on the most probable hypothesis. Not only did almost every child (28 out of 30) actively try to elicit information from the available sources (another agent or another object), they selectively accessed different sources of information given different evidence about the likely cause for the failure. When the hypothesis that the agent caused the failure was more probable than the hypothesis that the toy was broken (because infants were given a toy that worked for the experimenter), the majority of infants asked their parents for help. In this condition, varying the 'agent' variable is the most effective strategy: if you're doing something wrong, doing the same thing with a different object will not solve the problem. By contrast, when the hypothesis that their toy was broken was more probable than the hypothesis that they were doing something wrong (because infants were given a similar but non-identical toy), the majority of infants reached for a new object. In this condition, trying another exemplar is the most effective strategy: if a toy is broken, asking someone else to act on the broken toy will not solve the problem. These results suggest that infants rationally trade-off help-seeking and object-exploration behaviors depending on the relative probability of the two hypotheses.

Are there alternative ways of accounting for the results? One possibility is that infants' differential behavior across conditions might reflect different affective responses to differentially frustrating situations rather than active requests for information. The manipulation was set up so that infants in the Agent condition would have a stronger expectation that their toy should work than infants in the Object condition. Arguably, infants in the Agent condition might have been more frustrated by their failure, and more likely to turn to their parents than infants in the Object condition. Conversely, infants in the Object condition arguably had a more "boring" toy than infants in the Agent condition (because they had never seen their toy activated). They thus may have been more motivated to discard it and reach for a new toy than infants in the Agent condition.

Further research is needed to definitively rule out these accounts but we believe that the current data renders both explanations unlikely. First, differential frustration or boredom might be indicated by a difference in children's overall playtime between conditions but children played just as long in the Agent condition as the Object condition. Second, infants in the Agent condition who asked their parents for help did not show any signs of upset and did not look for comfort. They handed their parents the toy or tried to place their parents' hands on the toy but they did not cling to their parents or fuss out. Similarly, there was no indication that infants in the Object condition were more bored by the toy than infants in the Agent condition. Infants in the Object condition were just as likely to push the button on the toy as infants in the Agent condition, and they pushed the button just as persistently. Moreover, infants in the Object condition who retrieved the red toy immediately tried the button on the red toy. These behaviors suggest that infants in the Object condition expected that the toys would work and were strongly motivated to try to activate them. Thus the alternative accounts are inconsistent with how infants used the two different means: rather than reflecting frustration or boredom, infants' behavior is consistent with an attempt to generate an effective intervention. As noted however, conclusively distinguishing these possibilities requires further research. If for instance, an irrelevant distracter toy (rather than the Red toy) is placed at the end of the cloth, there should be no differences between the two conditions. We are currently running this control.

We note that the current study falls short of looking at whether children *learn* from the source of information they choose. That is, we cannot distinguish between the possibility that infants are taking the most rational steps to try to generate an outcome and the possibility that infants are (additionally) using the disambiguating evidence to determine the cause of the initial failure. In the current study, we deliberately asked the parents not to touch the toy, and the Red toy on the cloth was always inert. In future studies, we aim to look at whether learning occurs by studying infants' responses to different information that other people or other toys might provide. Imagine for instance, that if children retrieve the Red toy, it works for half the children and is inert for the other half. This evidence should give the children different information about the Yellow toy: if the Red toy works, the yellow toy is probably broken; if the Red toy does not work, it is now more probable that the child herself is the source of the failure. Thus if the Red toy is removed and the Yellow toy is returned to the children, they should be more likely to discard it if the Red toy worked, and more likely to ask for help if the Red toy failed. This would suggest that the infants' interventions not only serve the purpose of helping them make things happen but also help children disambiguate evidence to support causal learning. This research is also currently underway.

The study proposed above might also help clarify whether infants actually entertain the two competing hypotheses simultaneously, or whether only one of the two is considered in a given context. In the current study, some children who pulled the cloth and confirmed that the Red

toy did not work subsequently turned to their caregivers and asked for help. This suggests the possibility that children might indeed consider both hypotheses and act on the one favored by the evidence. Whether infants' choices of actions change dynamically given evidence that favors different hypotheses is an exciting topic for future research.

The current results however, already reveal impressive abilities in 16-month-old infants. There is abundant evidence that young children both ask adults for help (Dunham, Dunham, & O'Keefe, 2000; O'Neill, 1996) and explore objects in the world (Bonawitz, Shafto, Gweon, Spelke, & Schulz, submitted; Gweon & Schulz, 2008; Piaget, 1930; Gweon, Tenenbaum, & Schulz, 2010). This study goes beyond previous work in suggesting that infants' actively trade-off these two alternatives. Infants not only consider competing hypotheses about the failure of goal-directed actions, but also choose different means to resolve the ambiguity depending on which hypothesis is more probable. In the face of failure to achieve a goal, 16-month-old infants do not simply look to their parents nor do they simply move on to a new toy. Instead, they are able to infer the likely cause for their failure, and flexibly and rationally adjust their behavior. In solving the problem of assigning causal responsibility to themselves or the world, infants might lay the earliest foundations for scientific inquiry.

Acknowledgments

Thanks to Stephanie Tong for help with data collection. This research was supported by an NSF Faculty Early Career Development Award, a John Templeton Foundation Award, and James S. McDonnell Foundation Collaborative Interdisciplinary Grant on Causal Reasoning to L.S.

References

Baldwin, D. A. (1993). Infants' ability to consult the speaker for clues to word reference. *Journal of Child Language, 20*(2), 395-418.

Baldwin, D. A., Markman, E. M., & Melartin, R. L. (1993). Infants' ability to draw inferences about nonobvious object properties: evidence from exploratory play. *Child Development, 64*(3), 711-728.

Bonawitz, E. B., Shafto, P., Gweon, H., Spelke, E. S., & Schulz, L. E. (submitted). The double-edged sword of pedagogy: Teaching limits children's spontaneous exploration and discovery.

Carpenter, M., Nagell, K., Tomasello, M., Butterworth, G., & Moore, C. (1998). Social cognition, joint attention, and communicative competence from 9 to 15 months of age. *Monographs of the Society for Research in Child Development, 63*(4).

Chen, Z., & Klahr, D. (1999). All other things being equal: Acquisition and transfer of the control of variables strategy. *Child Development, 70*(5), 1098-1120.

Dunham, P., Dunham, F., & O'Keefe, C. (2000). Two-year-olds' sensitivity to a parent's knowledge state: Mind reading or contextual cues? *British Journal of Developmental Psychology, 18*(4), 519-532.

Gergely, G., Bekkering, H., & Kiraly, I. (2002). Rational imitation in preverbal infants. *Nature, 415*(6873), 755.

Gopnik, A., & Meltzoff, A. N. (Eds.). (1994). *Minds, bodies, and persons: Young children's understanding of the self and others as reflected in imitation and "theory of mind" research*. New York: Cambridge University Press.

Gweon, H., Tenenbaum, J. B., & Schulz, L. E. (in press). Infants consider both the sample and the sampling process in inductive generalization. *Proceedings of the National Academy of Sciences.*

Gweon, H., & Schulz, L. E. (2008). Stretching to learn: Ambiguous evidence and variability in preschoolers' exploratory play. *Proceedings of the 30th annual meeting of the Cognitive Science Society*, 570-574.

Koslowski, B. (1996). *Theory and evidence: The development of scientific reasoning*. Cambridge, MA: The MIT Press.

Kuhn, D. (1989). Children and adults as intuitive scientists. *Psychological Review, 96*(4), 674-689.

Kushnir, T., & Gopnik, A. (2005). Young children infer causal strength from probabilities and interventions. *Psychological Science, 16*(9), 678-683.

Meltzoff, A., & Brooks, R. (2001). "Like me" as a building block for understanding other minds: bodily acts, attention, and intention. In B. F. Malle, L. J. Moses & D. A. Baldwin (Eds.), *Intentions and intentionality: Foundations of social cognition*. Cambridge MA: The MIT Press.

O'Neill, D. (1996). Two-year-old children's sensitivity to a parent's knowledge state when making requests. *Child Development, 67*(2), 659-677.

Penner, D., & Klahr, D. (1996). The interaction of domain-specific knowledge and domain-general discovery strategies: A study with sinking objects. *Child Development, 67*(6), 2709-2727.

Piaget, J. (1930). *The child's conception of physical causality*: New York: Harcourt, Brace.

Schulz, L. E., & Bonawitz, E. B. (2007). Serious fun: preschoolers engage in more exploratory play when evidence is confounded. *Developmental Psychology, 43*(4), 1045-1050.

Sodian, B., Zaitchik, D., & Carey, S. (1991). Young children's differentiation of hypothetical beliefs from evidence. *Child Development, 62*(4), 753-766.

Sommerville, J., & Woodward, A. (2005). Pulling out the intentional structure of action: the relation between action processing and action production in infancy. *Cognition, 95*(1), 1-30.

Sorce, J., Emde, R., Campos, J., & Klinnert, M. (1985). Maternal emotional signaling: Its effect on the visual cliff behavior of 1-year-olds. *Developmental Psychology, 21*(1), 195-200.

Walden, T., & Ogan, T. (1988). The development of social referencing. *Child Development*, 1230-1240.

Williamson, R., Meltzoff, A., & Markman, E. (2008). Prior experiences and perceived efficacy influence 3-year-olds' imitation. *Developmental Psychology, 44*(1), 275-285.

Developmental differences in learning the forms of causal relationships

Chris Lucas (clucas@berkeley.edu)
Alison Gopnik (gopnik@berkeley.edu)
Thomas L. Griffiths (tom_griffiths@berkeley.edu)
Department of Psychology, University of California
Berkeley, CA 94720, USA

Abstract

Children learn causal relationships quickly, and make far-reaching causal inferences on the basis of what they see. In order to be such efficient learners, they must bring abstract knowledge to bear on their problems. This paper addresses children's ability to acquire that knowledge. We present evidence that children can learn about the abstract properties of causal relationships using only a handful of events, and – consistent with a hierarchical Bayesian model of casual inference – children can be more sensitive to evidence than adults.

Introduction

Recent work suggests that children are skilled at inferring specific causal relationships from patterns of data (Gopnik et al., 2004; Sobel, Tenenbaum, & Gopnik, 2004). For example, they can infer which blocks will activate a machine based on the contingencies between the blocks and the machine's activation. But an additional question is whether children can infer more abstract causal principles from patterns in data, and use those principles to shape their subsequent predictions. For example, can a child infer that a particular type of machine activates reliably, or requires only a single cause to activate? Will those abstract discoveries bias the child's interpretations of new data?

Developmental data suggest that children do have broad inductive biases. For example, in language learning the shape bias and the mutual exclusivity principle influence more specific inferences about word meaning (Smith, Jones, Landau, Gershkoff-Stowe, & Samuelson, 2002; Markman & Wachtel, 1988). However there is debate about whether these biases are the result of innate constraints or are themselves the product of learning (Elman et al., 1996; Leslie, 1994). Recent formal work on hierarchical Bayesian models suggests that, at least in principle, the shape bias may itself be learned as a result of normative inferences from patterns of data (Kemp, Perfors, & Tenenbaum, 2007). Similar high-level biases apply to causal learning, and we know that children can learn about causal types (Schulz, Goodman, Tenenbaum, & Jenkins, 2008), and the plausibility of cross-domain relationships (Schulz, Bonawitz, & Griffiths, 2007). In this paper, we explore whether children can learn abstract principles about the forms of causal relationships themselves.

The hierarchical Bayesian approach suggests that the nature of inductive biases may change as evidence accumulates. Absent evidence, a learner without strong built-in biases should assign similar probabilities to a wide range of hypotheses. As data accumulate, the abstract hypotheses consistent with those data become more probable, and the learner discounts any hypotheses that fit the current data but are less compatible with past experience. If this is correct, then we might expect to see different patterns of inductive bias in adults and children. In particular, children might rely less on past experience and more on present evidence than adults. This is a possibility that has not previously been explored in the causal learning literature, and one that we examine through head-to-head (or prior-to-prior) comparison of children and adults in a causal learning task that requires making an abstract generalization about the nature of causal relationships.

We test the high-level generalizations made by children and adults by contrasting two abstract "overhypotheses" (Goodman, 1955; Kemp et al., 2007) about how a causal system works. One is a noisy-OR model, in which each object has a certain independent probability of bringing about an effect. This model is pervasive in the literature on adult causal inference (e.g., Cheng, 1997; Griffiths & Tenenbaum, 2005). The other is an AND model in which individual causes are unable to produce an effect, but multiple causes in conjunction can produce an effect. We provided children and adults with evidence for either an AND or OR relationship and then examined how this evidence biased their judgment of a novel, ambiguous pattern of evidence. Would seeing several instances of a machine activated by a conjunction of causes lead them to assume that this would be the case for a new set of blocks? By comparing how children and adults respond to data that support these different overhypotheses, we can examine first whether children are capable of forming appropriate abstract generalizations, and second whether they are more willing to make these generalizations than adults.

The plan of the paper is as follows. First, we consider how an ideal Bayesian learner can gather evidence for overhypotheses relevant to causal induction. We then discuss the specific overhypotheses about the functional form of causal relationships that we contrast in this paper, together with a method that can be used to diagnose whether learners infer these overhypotheses from data. We go on to use this method to compare the abstract generalizations of children and adults in a causal learning task, finding support for the hypothesis that children are more willing to adopt a novel overhypothesis than adults. We close by discussing the implications of these results.

Causal overhypotheses

Children can identify causes using only a handful of observations (Gopnik et al., 2004), but the extent to which they learn

about the abstract properties of causal relationships remains largely unexplored. From a Bayesian standpoint, learning about causal structure requires having *a priori* beliefs – or priors – about what items are plausible causes, and expectations about how a given causal structure leads to different observable events. These expectations can be expressed formally using a *likelihood* function, which specifies the probability of observing a particular set of events based on the underlying causal structure.

Most work on probabilistic models of causal learning has assumed a specific kind of likelihood function. This likelihood function is based on causes and effects interacting in a "noisy-OR" manner, each having an independent opportunity to produce the effect (Cheng, 1997; Griffiths & Tenenbaum, 2005; Glymour, 1998). More precisely, a noisy-OR relationship implies that the probability that an effect E occurs given the presence of a set of causes $C_1, \ldots, C_N$ is

$$P(E|C_1, ..., C_N) = 1 - \prod_{i=1}^{N} (1 - w_i) \tag{1}$$

where w_i is the probability that C_i generates the effect in the absence of other causes.

Despite the popularity of the noisy-OR in models of causal learning, other kinds of causal relationships are clearly possible. For instance, a noisy-OR model cannot describe an AND relationship, where an effect only occurs when multiple causes are present. This might be the case in an electrical circuit where multiple switches are wired in series, and a light only turns on when all of the switches are flipped. It is important, then, for models of causal inference to accommodate flexible beliefs about the forms relationships can take. Formalizing inferences about the form of a relationship is straightforward, using an expanded likelihood function, $P(E|C_1, ..., C_N, F)$, where F captures information about the form of the causal relationship. For example, F could indicate that the relationship has a noisy-OR form, but another value of F might indicate that a causal relationship has an AND form.

Learning the form of a causal relationship and generalizing that discovery when reasoning about other causal relationships requires inference at multiple levels of abstraction. This kind of inference, in which lessons from one context can be carried forward for future learning, is easily captured by using a hierarchical Bayesian model (Tenenbaum, Griffiths, & Kemp, 2006; Kemp et al., 2007). A learner's abstract beliefs, or overhypotheses, determine the probabilities of more-concrete hypotheses, each encoding specific causal structures and the form a relationship takes. These hypotheses, in turn, determine the likelihood of different patterns of events.

Formally, we can imagine an inference involving variables at three levels: the observed data D, hypotheses about the causal structure underlying those data H, and overhypotheses (or a "theory", as in Griffiths & Tenenbaum, 2009) T representing generalizations relevant to evaluating those hypotheses (see Figure 1). Bayes' rule then specifies how the events

T : Theory/overhypothesis

H : Hypothesis

D : Data

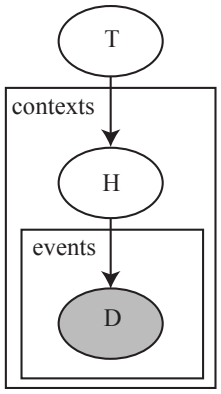


Figure 1: The structure of a hierarchical Bayesian model.

a learner sees (D) should change the learner's beliefs, both about the casual system at hand (H), and about the higher-level properties of that kind of system (T). Formally, we have

$$p(T|D) = \frac{p(D|T)\,p(T)}{p(D)} \tag{2}$$

where $p(T)$ is the prior probability of the overhypothesis T, $p(T|D)$ is the posterior probability, and $p(D)$ is obtained by summing the numerator over all overhypotheses T. The probability of the data given an overhypothesis is obtained by summing over all hypotheses consistent with that overhypothesis,

$$p(D|T) = \int p(D|H)\,p(H|T)\,dH, \tag{3}$$

and can be interpreted as an average of the probability of the observed data under those hypotheses weighted by the extent to which each hypothesis is consistent with the overhypothesis.

Intuitively, this hierarchical Bayesian approach provides a way to explain how learners can form and use abstract generalizations about causal systems. For example, if a child sees events that are likely under an AND relationship, such as a machine activating only when pairs of causal objects are placed on it, then the probability of an overhypothesis predicting future AND relationships increases. This is because the best hypotheses for explaining the observed events are those that are most likely under this overhypothesis, so Equation 3 yields a high value. Incorporating this value into Equation 2, the posterior probability for that overhypothesis will increase.

As the evidence supporting a particular overhypothesis increases, it will be easier to learn about the structure and form of causal systems that are consistent with that overhypothesis. This comes with a cost: if a causal system has strange or rare abstract properties, such as an unlikely functional form, much more evidence will be necessary to learn about it. The implication is that adults, who have seen a great deal of evidence, should find it very easy to learn about the structure and form of causal relationships that have typical properties. Conversely, children, with their limited experience, should be more sensitive to evidence when learning about relationships

that have unusual properties. In the following section, we discuss an experimental design for testing this idea.

The functional form of causal relationships

If children update their abstract beliefs about causal systems in a manner consistent with Bayesian inference, then the events they see should influence their judgments about different sets of events and prospective causes. To test this hypothesis, we used an experiment with two phases, each with a distinct set of objects. In the first phase, children saw a set of events designed to be likely under one of two abstract overhypotheses about the forms of causal relationships. In the second phase, they saw events where different beliefs about the form of the causal relationship should lead them to make different judgments about which objects are causes.

The specific evidence we provided to participants was very similar to that given to adults in Lucas and Griffiths (2009), where the task was to identify the blickets within a set of objects, knowing only that blickets have "blicketosity". Prospective blickets could be placed on a "blicketosity meter", causing it to either activate by lighting up and playing music or do nothing. People might entertain a variety of expectations about the relationship between the blickets and the machine, determining how they interpret different events. For example, if they think that two blickets are necessary to activate the machine, seeing a single object fail to activate it provides no information. At the same time, their expectations about the form of the relationship between blickets and the blicketosity meters can be shaped by the events they observe. For instance, seeing two objects fail to activate the machine separately but succeed together suggests that two blickets are necessary for activation.

We used events from two conditions from Experiment 2 of Lucas and Griffiths (2009). Since this experiment is closely related to the approach we take here, we will recapitulate the method and results. In the *AND*[1] condition of the experiment, participants saw a training block of events where objects labeled A, B, and C were placed sequentially on the machine, which failed to activate in all cases. Next, all pairs of objects were placed on the machine sequentially, with only A and B together causing activation. See Figure 2 for a summary of the events in the training and test blocks. Participants were then asked to rate the probability that A, B, and C were blickets on a 0-10 scale, with 0 indicating the object was definitely not a blicket, a 10 indicating it definitely was, and 5 indicating it was as likely to be a blicket at not.

After making these judgments, participants saw three new objects, D, E, and F, which they had never seen before, and a series of test events intended to be ambiguous, leading to different judgments about which of D, E, and F were blickets, depending on participants' expectations about the form of the relationship. If people expect that a single blicket suffices to activate the machine, they should believe then F is likely to be a blicket, while D and E are not. If, in contrast, people exploit the information provided by the training block so they conclude that two blickets are necessary to activate the machine, then they should think that objects D and F are blickets, and be uncertain about object E.

In the *OR* condition, participants saw a different set of events in the training block, which were chosen to indicate that an OR relationship applied (see Figure 2). Then they saw the same test events that the participants in the *AND* condition saw. Based on the training evidence, participants in this condition were predicted to say that only object F was a blicket.

As predicted, people in the *AND* condition assigned significantly higher probabilities to object D being a blicket, giving a mean score of 3.08 (SD=3.32), versus 0.23 (SD=0.99) in the *OR* condition. The mean rating was less than 5 in the *AND* condition, consistent with the idea that adults believe that disjunctive relationships are more probable, and could interpret the *AND* condition events in several ways, including as evidence for a noisy relationship in which the machine happened to fail to activate when a single, normally sufficient blicket was placed on it.

In summary, Lucas and Griffiths (2009) showed that people's inferences about causal structure are driven by their beliefs about the probable forms of causal relationships, which are in turn influenced by events they have seen in the past. The specific pattern of judgments is consistent with the predictions of a hierarchical Bayesian model given priors reflecting a strong bias in favor of disjunctive (OR) and deterministic relationships. Such priors are also consistent with adults' performance in other experiments (Lu, Yuille, Liljeholm, Cheng, & Holyoak, 2006). This prior could be chiefly due to adults' experiences revealing that OR relationships are more common, or an innate bias. By comparing the judgments of 4-year-old children to those of adults, we aim to answer that question and better understand the origins of the abstract knowledge that drives efficient causal inference.

Causal overhypotheses in children and adults

We used the experimental design from Lucas and Griffiths (2009) to explore two questions about the use of causal overhypotheses by children and adults. The first question was whether children, like adults, can use events to update their knowledge about the likely forms of causal relationships, and apply that knowledge to learn the causal structure behind new and ambiguous sets of events. The second question was whether children are more or less sensitive to evidence supporting such high-level generalizations, as opposed to their prior beliefs.

If children are more likely than adults to call objects D and E blickets in the *AND* condition, we can conclude that much of the bias we see in adults is due to learning during and after childhood, including, for instance, experience with machines to which OR relationships apply. If children's judgments are

[1] Lucas and Griffiths labeled their conditions *conjunctive* and *disjunctive* rather than AND and OR, to highlight a hypothesis space that included a wide range of functional forms, including AND and OR as special cases. We use AND and OR here for the sake of simplicity.

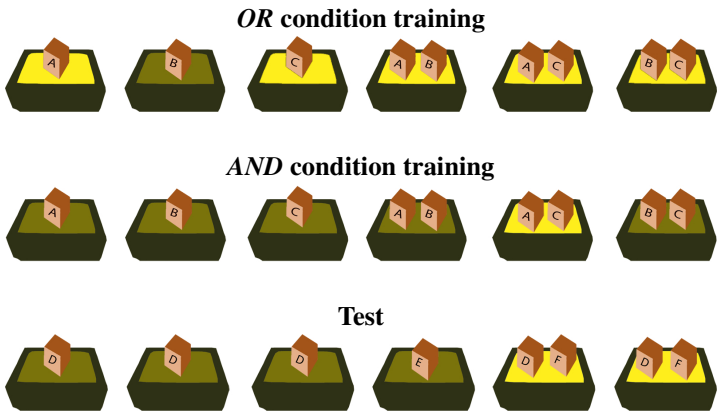


Figure 2: Evidence presented to participants in the two training phases, as well as the subsequent test phase which all participants saw. Events are given as a set of prospective causes and the presence or absence of an effect. The bright-paneled machines represent events in which the effect occurs and the dark-paneled machines represent events in which the effect does not occur.

indistinguishable from adults', we have evidence that learning about the forms of causal relationships occurs early, or plays a minor role in driving our expectations. Finally, if there is no effect of training evidence on test-block judgments, we should question the applicability of the model used by Lucas and Griffiths (2009) to causal inference in children.

We can generate more detailed predictions by speculating about the priors that children bring to the problem of identifying blickets. It seems unlikely that children are constrained to a small set of discrete overhypotheses – it is more natural to suppose that they consider a space of possibilities that includes both OR and AND relationships as special cases. Following Lucas and Griffiths (2009), we use a sigmoid family of likelihood functions, where the probability of the machine's activation given that n blickets are present is

$$P(\text{effect}|N_{\text{blickets}} = n) = \frac{1}{1+\exp\{-g(n-b)\}}. \quad (4)$$

The overhypotheses determine the probability of different values of the gain g and the bias b. The gain specifies how many blickets are necessary to activate the machine, and the bias reflects how noisy the relationship is. Lucas and Griffiths found that exponential priors predicting a high mean gain (3.34) and a low mean bias (0.23) – or reliable OR relationship – lead to model predictions that closely match adults' judgments. If children are happier believing that a relationship could be conjunctive or noisy, the priors that best capture their inferences should lead to *a priori* gains and biases closer to 1. This space of likelihood functions is intended to cover a range of relationships that are appropriate to the cover story and participants' prior knowledge, and we do not claim it includes all relationships that people could conceivably learn, such as those in which blickets prevent the machine from activating.

Participants

Children Thirty-two children were recruited from university-affiliated preschools, divided evenly between the *AND* and *OR* conditions. Children in the *AND* and *OR* conditions had mean ages of 4.46 (SD=0.27) and 4.61 (SD=0.31) years, respectively.

Adults UC Berkeley undergraduates received course credit for participating during lectures of an introductory psychology course. There were 88 participants in the *AND* condition and 55 in the *OR* condition. Five participants in the *AND* condition were excluded for declining to answer one or more questions.

Methods

Children Each child sat at a table facing the experimenter, who brought out three objects, each painted a different color, as well as a green box with a translucent panel on top, describing the box as "my blicketness machine".

At the beginning of the experiment, children were prompted to help the experimenter name the objects using their colors, e.g., "red". They were then told that the goal of the game was to figure out which of the objects were blickets, that blickets have blicketness inside them, and blickets cannot be distinguished from non-blickets by their appearance. No other information was provided about the relationship between blickets and the activation of the machine.

The children then observed a set of training events in which the experimenter placed objects alone or in pairs on the machine, which activated in some cases by lighting up and playing music. These events corresponded to either the *OR* condition or *AND* condition training given in Figure 2. After the children saw these events, they were asked whether each object was a blicket or not. Next, the experimenter brought out three objects that the children had not seen before. After the children named the new objects, the experimenter demonstrated the test events listed in Figure 2 and asked whether

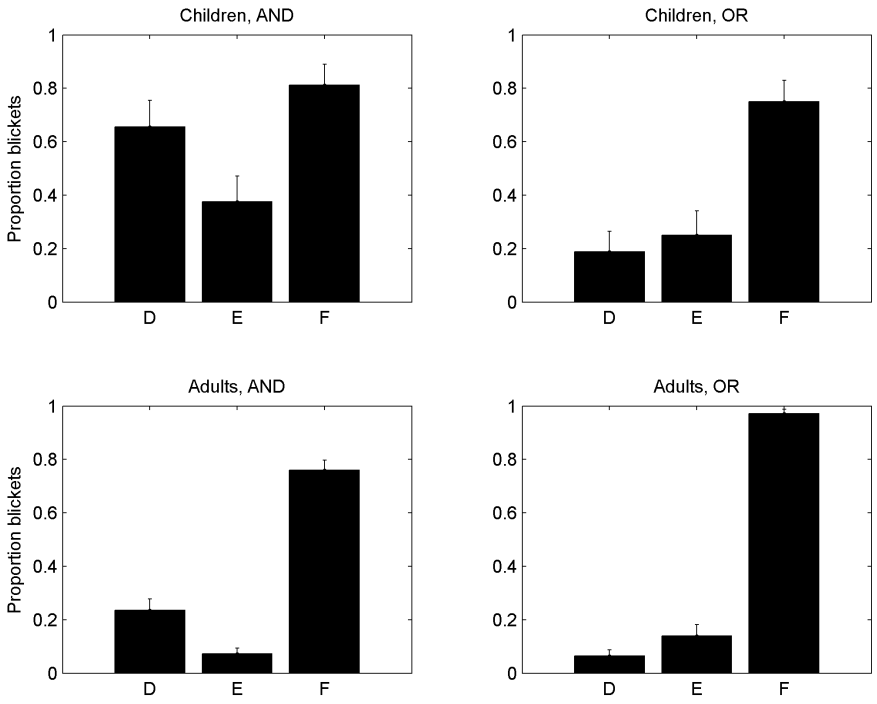


Figure 3: Proportions of objects that were judged to be blickets for children (top row) and adults (bottom row) for the *AND* (left column) and *OR* (right column) conditions. Error bars represent standard error of the mean.

each of these new objects was a blicket or not. In a departure from Lucas and Griffiths's design, the experiment was repeated a second time for each child, using the same patterns of evidence, but with a distinct set of objects that varied by shape and had a uniform gray color. The identities of the individual objects were counterbalanced, as was whether the children saw the different-shaped or different-colored objects first.

Adults The adults were tested in groups, and saw demonstrations that were almost identical to what the children saw in the corresponding conditions. Unlike the children, the adults were not asked to name the objects, and they recorded their judgments on sheets of paper rather than responding verbally.

Results

Children The critical prediction was that children would be more likely to judge object *D* to be a blicket in the *AND* condition than in the *OR* condition, indicating that they were (1) learning about the form of the relationship between blickets and the machine's activation, and (2) transferring that abstract knowledge to make better inferences about novel objects and otherwise ambiguous events.

Children were more likely to judge object D to be a blicket in the *AND* condition than in the *OR* condition ($p < 0.005$, two-tailed permutation test). There was also a change in the predicted direction for object E, albeit non-significant.

Adults Adults were also more likely to judge object D to be a blicket in the *AND* condition than in the *OR* condition ($p < 0.005$, two-tailed permutation test), consistent with the results in Lucas and Griffiths (2009). See Figure 3, bottom row, for a summary of their judgments for the test objects.

Differences In the *AND* condition, the adults judged object D to be a blicket less frequently than children ($p < 0.005$, Fisher's exact test). See Figure 3 for a summary of ratings in the three conditions. Children's ratings were also higher for object E ($p < 0.001$, two-sided permutation test), which is consistent with their being quicker to learn that an AND relationship applies: under an AND relationship, the event where E fails to activate the machine is uninformative, so judgments of E being a blicket should reflect the base rate of blickets occurring. The high frequency of other objects being blickets under an AND relationship (4 of 5), plus a belief that blickets are not rare, should lead a learner to expect that a novel object is somewhat likely to be a blicket.

Model fits We converted children's judgments about blickets to probabilities in order to examine them using the previously-mentioned hierarchical Bayesian model and sigmoid space of hypotheses. We treated is-a-blicket judgments as assertions that objects were definitely blickets, and not-a-blicket judgments as assertions that objects were definitely not blickets. Lucas and Griffiths (2009) found that priors favoring disjunctive, deterministic relationships – predicting a mean gain of 3.34 and a mean bias of 0.23 – fit adults' judgments closely, with a mean squared error of 0.29 per judgment on a zero to ten scale. We found that similar priors best

captured adults' judgments in our experiment, giving a mean squared error of 0.80 with a mean gain of 5.30 and bias of 0.11.

These same priors were wildly inconsistent with children's inferences, giving a mean squared error of 6.12. In contrast, priors giving a mean *a priori* gain and bias of 1 – favoring neither AND nor OR relationships – were much more accurate, with a mean squared error of 0.58. The priors that best fit the children's judgments gave a mean gain and bias of 1.45 and 0.85, respectively, with mean squared error of 0.15.

Discussion

Our experiment was designed to explore two questions: whether children could make high-level generalizations about the form of causal relationships, and whether they were more willing to do so than adults. Our results show that children are capable of making such inferences, and that their judgments were more strongly influenced by the available evidence than adults, whose inferences reflected a bias toward OR relationships. Our results thus support the view that when learning about cause and effect, children are flexible learners whose inexperience may sometimes let them learn better from sparse evidence, especially in novel situations. These results are also consistent with treating the acquisition and application of causal knowledge as a matter of hierarchical Bayesian inference, where a learner has beliefs expressed at multiple levels of abstraction, with abstract theories driving specific hypotheses which, in turn, enable prediction and categorization.

Before closing, we will address two alternative explanations for our results. The first is that children are more likely than adults to judge any object to be a blicket. This is less consistent with the data than our interpretation, given that adults were more likely than children to call object *F* a blicket in the *OR* condition, and nearly as likely in the *AND* condition (75 percent of the objects versus 81 percent). A second alternative is that the children were confused by the training data in the *AND* condition, and responded to the novel objects by guessing randomly. This explanation can be ruled out by noting that children judged objects *D* and *F* to be blickets more often than chance would predict ($t(15) = 3.529$, $p < 0.005$).

The results of our experiment have implications for understanding causal learning, and for understanding cognitive development more generally. In terms of causal learning, these results suggest that the fundamental biases that lie beneath causal inference are more subtle and abstract than *a priori* preferences for specific kinds of causal relationships. We believe that trying to understand these biases is fertile ground for future research. For cognitive development, the idea that children are more flexible in their commitments about the way that causal systems tend to work seems like not just a necessary consequence of a hierarchical Bayesian approach, but an important insight for understanding how it is that children see the world differently from adults. The plasticity of beliefs that this implies helps to explain the bold exploration and breathtaking innovation that characterizes children's interactions with the world.

Acknowledgments. This research was supported by the James S. McDonnell Foundation's Causality Collaborative Initiative and the Air Force Office of Scientific Research, grant FA9550-07-1-0351.

References

Cheng, P. (1997). From covariation to causation: A causal power theory. *Psychological Review*, *104*, 367-405.

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking innateness: A connectionist perspective*. Cambridge, MA: MIT Press.

Glymour, C. (1998). Learning causes: Psychological explanations of causal explanation. *Minds and Machines*, *8*, 39-60.

Goodman, N. (1955). *Fact, fiction, and forecast*. Cambridge: Harvard University Press.

Gopnik, A., Glymour, C., Sobel, D., Schulz, L., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review*, *111*, 1-31.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*, 354-384.

Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-Based Causal Induction. *Psychological Review*, *116*(4), 661–716.

Kemp, C., Perfors, A., & Tenenbaum, J. (2007). Learning overhypotheses with hierarchical bayesian models. *Developmental Science*, *10*(3), 307–321.

Leslie, A. M. (1994). ToMM, ToBY, and agency: Core architecture and domain specificity. In L. A. Hirschfeld & S. A. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture.* Cambridge: Cambridge University Press.

Lu, H., Yuille, A., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2006). Modeling causal leanring using bayesian generic priors on generative and preventive powers. In R. Sun & N. Miyake (Eds.), *Twenty-eighth conference of the cognitive science society* (p. 519-524). Erlbaum.

Lucas, C., & Griffiths, T. (2009). Learning the Form of Causal Relationships Using Hierarchical Bayesian Models. *Cognitive Science*, *34*(1), 113–147.

Markman, E., & Wachtel, G. (1988). Childrens use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology*, *20*(2), 121–157.

Schulz, L. E., Bonawitz, E. B., & Griffiths, T. L. (2007). Can being scared make your tummy ache? naive theories, ambiguous evidence, and preschoolers' causal inferences. *Developmental Psychology*.

Schulz, L. E., Goodman, N., Tenenbaum, J., & Jenkins, A. (2008). Going beyond the evidence: Abstract laws and preschoolers' responses to anomalous data. *Cognition*, *109*(2), 211–223.

Smith, L. B., Jones, S. S., Landau, B., Gershkoff-Stowe, L., & Samuelson, L. (2002). Object name learning provides on-the-job training for attention. *Psychological Science*, *13*(1), 13–19.

Sobel, D. M., Tenenbaum, J. B., & Gopnik, A. (2004). Children's causal inferences from indirect evidence: Backwards blocking and Bayesian reasoning in preschoolers. *Cognitive Science*, *28*, 303-333.

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Science*, *10*, 309-318.

Children's Imitation of Action Sequences is Influenced by Statistical Evidence and Inferred Causal Structure

Daphna Buchsbaum, Alison Gopnik, Thomas L. Griffiths
{daphnab, gopnik, tom_griffiths}@berkeley.edu
Department of Psychology, University of California, Berkeley, Berkeley, CA 94720 USA

Abstract

Children are ubiquitous imitators, but how do they decide which actions to imitate? One possibility is that children might learn which actions are necessary to reproduce by observing the contingencies between action sequences and outcomes across repeated observations. We define a Bayesian model that predicts that children will decide whether to imitate part or all of a sequence based on the pattern of statistical evidence. To test this prediction, we conducted an experiment in which preschool children watched an experimenter repeatedly perform sequences of varying actions followed by an outcome. Children's imitation of sequences that produced the outcome increased, in some cases resulting in production of shorter sequences of actions that the children had never seen performed in isolation. This behavior is consistent with our model's predictions, and suggests that children attend to statistical evidence in deciding which actions to imitate, rather than obligately imitating successful actions.

Keywords: Cognitive development; Imitation; Statistical learning; Causal inference; Bayesian inference

Introduction

Learning the causal relationships between everyday sequences of actions and their outcomes is a daunting task. How do you transform a package of bread, a jar of peanut butter and a jar of jelly into a peanut butter and jelly sandwich? Do you cut the bread in half before or after you put together the sandwich? Can you put the peanut butter on first, or does it always have to be jelly first? In order to achieve desired outcomes – from everyday goals such as eating a tasty sandwich to distinctive human abilities such as making and using tools – children need to solve a challenging causal learning problem: observing that the intentional actions of others lead to outcomes, inferring the causal relations between those actions and outcomes, and then using that knowledge to plan their own actions.

To learn from observation in this way, children cannot simply mimic everything they see. Instead, they must segment actions into meaningful sequences, and determine which actions are relevant to outcomes and why. Recent studies of imitation in children have produced varying answers to the question of whether children are capable of solving this problem. While children sometimes selectively reproduce the most obviously causally effective actions (Williamson, Meltzoff, & Markman, 2008; Schulz, Hooppell, & Jenkins, 2008), at other times they will "overimitate", reproducing apparently unnecessary parts of a causal sequence (Whiten, Custance, Gomez, Teixidor, & Bard, 1996; Lyons, Young, & Keil, 2007), or copying an actor's precise means, when a more efficient action for accomplishing the same goal is available (Meltzoff, 1995). Sometimes children may produce both kinds of behavior in the same study. In the "rational imitation" studies by Gergely, Bekkering, and Kiraly (2002), children saw an experimenter activate a machine with hands free or hands confined. Children both produced exact imitations of the actor (touching their head to a machine to make it go) and produced more obviously causally effective actions (touching the machine with a hand), though the proportion of such actions differed in the different intentional contexts.

We suggest that these different results reflect the multiple sources of information that contribute to a rational statistical inference about causally effective actions. Children need to balance their prior knowledge about causal relations, the new evidence that is presented to them by the adult, and their knowledge of the adult's intentions. Moreover, in the case of imitation there is often no single "right answer" to the question of what to imitate. After all, a longer "overimitation" sequence might actually be necessary to bring about an effect, though that might seem unlikely at first. The imitation problem can be expressed as a problem of Bayesian inference, with Bayes' rule indicating how children might combine these factors to formulate different causal hypotheses and produce different action sequences based on those hypotheses. It is difficult to test this idea however, without knowing the strength of various causal hypotheses for the children. Since previous studies involved general folk physical and psychological knowledge (such as removing a visibly ineffectual bolt to open a puzzle box) it is difficult to know how strong those hypotheses would be. By giving children statistical information supporting different hypotheses we can normatively determine how probable different hypotheses should be, and then see whether children's imitation reflects those probabilities.

It is also independently interesting to explore the role of statistical information in imitation. Recent studies show that children are surprisingly sophisticated in their use of statistical information such as conditional probabilities in a range of domains, from phonology (Saffran, Aslin, & Newport, 1996), to visual perception (Fiser & Aslin, 2002; Kirkham, Slemmer, & Johnson, 2002), to word meaning (Xu & Tenenbaum, 2007). Such information plays a particularly important role in both action processing (Zacks et al., 2001; Baldwin, Andersson, Saffran, & Meyer, 2008; Buchsbaum, Griffiths, Gopnik, & Baldwin, 2009) and causal inference (Gopnik et al., 2004; Gopnik & Schulz, 2007), and allows adults to identify causal subsequences within continuous streams of action (Buchsbaum et al., 2009). Varying the probabilities of events within action sequences may thus provide a way to vary the statistical evidence those sequences provide in favor of different causal hypotheses.

Statistical inference might be particularly important to

Observed Action Sequence	Potential Causal Sequences
ABC+	ABC, BC, C
DBC+	DBC, BC, C
Total Potential Causes	ABC, DBC, BC, C

Table 1: Example demonstrations, and the associated set of potential causal sequences. Letters represent unique observed actions, a + indicates a causal outcome.

imitation because it could allow children to not only determine the causal relationship between action sequences and outcomes, but to identify irrelevant actions within causally effective sequences. Imagine that I am making a peanut butter sandwich, and that between opening the jar, and spreading the peanut butter, I get peanut butter on my hands, so I wipe them on a paper towel. If this is the first time you've seen me make a sandwich, you might mistakenly think that hand-wiping is a necessary step. However, after watching me make a sandwich a couple of times, you might notice that while opening the jar always predicts spreading the peanut butter, it doesn't always predict hand-wiping, and could infer that this step is extraneous. In most previous work on children's imitation of casual sequences, children observed only a single demonstration of how to generate the outcome (e.g. Whiten et al., 1996; Lyons et al., 2007).

In this paper, we look at whether children use statistical evidence from repeated demonstrations to infer the correct causal actions within a longer sequence and imitate them. We present a Bayesian analysis of causal inference from repeated action sequence demonstrations, followed by an experiment investigating children's imitative behavior and causal inferences. We showed preschool children different sequences of three actions followed by an effect, using our Bayesian model to guide our manipulation of the probabilistic evidence, such that the statistical relations between actions and outcomes differed across conditions in ways that supported different causal hypotheses. We then examine which sequences the children produced themselves, and compare children's performance to our model's predictions. We conclude by discussing our results in the context of broader work on imitation, and causal and intentional inference.

Bayesian Ideal Observer Model

In many real world situations, the causal structure of a demonstrated sequence of actions is not fully observable. In particular, which actions are causally necessary and which are superfluous may be unclear. One way children may overcome this difficulty is through repeated observations. By watching someone make a sandwich or turn on a lightbulb on multiple occasions, children can pick up on which actions consistently predict the desired outcome, and which do not.

While it is intuitively plausible that children can use the statistical evidence in repeated demonstrations to infer causal structure, we would like to verify that normative inferences from repeated observations of action sequences and their outcomes vary in a systematic way with different patterns of data. One way to derive what the normative distribution over

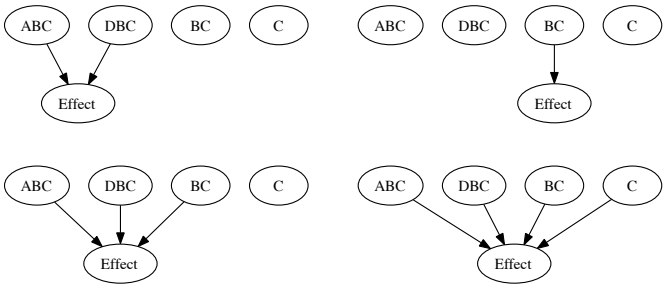


Figure 1: A subset of the hypothesis space. Each box represents a hypothesis about which action sequences are causal.

causes should be is through a Bayesian model (Gopnik et al., 2004; Griffiths & Tenenbaum, 2005). The Bayesian formalism provides a natural way to represent the roles of children's prior assumptions and the observed data in forming their beliefs about which action sequences are likely to be causal.

Model Details

Given observations of several sequences of actions, we assume that children consider all sequences and terminal subsequences as potentially causal. These include both sequences that generate the outcome and those that do not. For instance, if the sequence "squeeze toy, knock on toy, pull toy's handle" is observed, then squeeze, followed by knock, followed by pull handle would be one possible causal sequence, and knock followed by pull handle would be another. Given all of the observed sequences, we can enumerate the potential causes (see Table 1 for an example set of demonstrations and potential causes). As in previous work on children's causal inference, we use a Deterministic-OR model (c.f. Cheng, 1997; Pearl, 1996), in which any of the correct sequences will always bring about the effect. To capture the intuition that there may be more than one sequence of actions that can bring about an effect, we consider all of the potential causes (such as in Table 1), as well as all disjunctions of these causes. The base causes, together with the disjunctions form the space of potential hypotheses, H (see Figure 1).

The learner wants to infer the set of causes, h, given the observed data, d, where the data are composed of an observed sequence of actions, a, and an outcome, e. Bayes' theorem provides a way to formalize this inference. Bayes' theorem relates a learner's beliefs before observing the data, their prior $p(h)$, to their beliefs after having observed the data, their posterior $p(h|d)$,

$$p(h|d) \propto p(d|h)p(h), \tag{1}$$

where $p(d|h)$ is the probability of observing the data given the hypothesis is true. For Deterministic-OR causal models, this value is 1 if the sequence is consistent with the hypothesis, and zero otherwise. For example, given the hypothesis that squeeze is the cause, a consistent observation would be, knock then squeeze followed by music, and an inconsistent observation would be squeeze followed by no music. When multiple sequences of actions and effects are observed, we assume that these sequences are independent.

A key element in this inference is the learner's prior expectations, $p(h)$. Children could have a variety of different beliefs about the kinds of sequences that bring about effects. For instance, they could believe that longer sequences, that include more of the demonstrated actions, are more likely to bring about effects. Or, they could believe that there tends to be only one correct sequence, as opposed to many possible sequences, that cause an effect. We capture these intuitions with a prior that depends on two parameters, β and p, which correspond to the learner's expectations about the length of causal sequences and number of ways to generate an effect.

We formalize the prior as a generative model. Hypotheses are constructed by randomly choosing causal sequences, a. Each sequence has a probability p_a of being included in each hypothesis and a probability $(1-p_a)$ of not being included,

$$p(h) \propto \prod_{a \in h} p_a \prod_{a* \notin h} (1 - p_{a*}) \tag{2}$$

where the probability of including causal sequence a is

$$p_a = \frac{1}{1 + \frac{1-p}{p}\exp(-\beta(|a|-2))}, \tag{3}$$

and $|a|$ is the number of actions in the sequence a. Values of β that are greater than 0 represent a belief that longer sequences are more likely to be causes. Values of p less than 0.5 represent a belief that effects tend to have fewer causes. Together, Equations 1, 2 and 3 provide a model of inferring hypotheses about causes from observed sequences and their effects.

In our experiments, rather than probing children's beliefs directly, we allow children to play with the toy. Therefore, to complete the model, we must specify how children choose action sequences, a, based on their observations, d. Intuitively, we expect that if we know the set of causes of the effect, h, we will randomly choose one of these actions. If we were unsure about which of several possible causes was the right one, then we may choose any of the possible contenders, but biased toward whichever one we thought was most likely. We capture these intuitions formally by choosing an action given the observed data, $p(a|d)$, based on a sum over possible hypotheses,

$$p(a|d) \propto \sum_{h \in H} p(a|h)p(h|d), \tag{4}$$

where $p(a|h)$ is one if a is a cause under h, and zero otherwise, and $p(h|d)$ is specified in Equation 1.

A Simple Modeling Example

We can now verify that the model makes distinct inferences from repeated demonstrations. In the first example, the demonstrated action sequences are ABC+, DBC+ as in Table 1. That is, a sequence of three actions A, B and C is followed by an effect. Subsequently, a different sequence of three actions, D, B, and C is followed by the same effect. In the second example, the observed sequences are ABC+, DBC. Here, the second three-action sequence is not followed by the effect.

Observed Sequences	ABC	DBC	BC	C
ABC+, DBC+	0.23	0.23	0.27	0.27
ABC+, DBC	1.0	0.0	0.0	0.0

Table 2: Example model results, $p = 0.5$ and $\beta = 0$.

Observed Sequences	ABC	DBC	BC	C
ABC+, DBC+	0.26	0.26	0.35	0.13
ABC+, DBC	1.0	0.0	0.0	0.0

Table 3: Example model results, $p = 0.1$ and $\beta = 1.0$.

Using values of $p = 0.5$ and $\beta = 0$ results in a prior that assigns equal probability to all possible causal hypotheses – a uniform prior. With this uniform prior, we can now find the probability of choosing to perform each action sequence to bring about the effect given the observed data, $p(a|d)$, as described in Equation 4. Our model infers that, in the first case, all the sequences are possible causes, with BC and C being somewhat more likely, and equally probable. Notice that the model infers that the subsequences BC and C are the most likely causes, even though neither was observed on its own. The second case is quite different. Here the model sees that DBC and its subsequences BC and C did not lead to the effect in the second demonstration, and infers that ABC is the only possible cause among the candidate sequences (see Table 2).

We now use values of $p = 0.1$ and $\beta = 1.0$ leading the model to favor simpler hypotheses containing fewer causes, and causes that use more of the observed demonstration.[1] This prior does not change results in the second case, where ABC is still the only possible cause. However, in the first case, the model now infers that the subsequence BC is the most likely individual cause, since it is the longest observed sequence to consistently predict the effect (see Table 3).

Model Predictions for Children's Inferences

Our rational model makes differential predictions based on repeated statistical evidence, and is able to infer subsequences as causal without seeing them performed in isolation. We can now use the model to help us construct demonstration sequences that normatively predict selective imitation in some cases, and "overimitation" in others. If children are also making rational inferences from variations in the action sequences they observe, then their choice of which actions to imitate in order to bring about an effect should similarly vary with the evidence. We test our prediction that children rationally incorporate statistical evidence into their decisions to imitate only part of an action sequence versus the complete sequence in the following section.

Experiment

Method

Participants Participants were 81 children (M = 54 months, $Range = 41-70$ months, 46% female) recruited from local preschools and a science museum. An additional

[1] These parameter values qualitatively fit children's imitative behavior, as we discuss later in the paper.

"ABC" Condition	"BC" Condition	"C" Condition
ABC+	A**BC+**	AB**C+**
DEC	ADC	AD**C+**
ABC+	D**BC+**	DB**C+**
EDC	AEC	AE**C+**
ABC+	E**BC+**	EB**C+**

Table 4: The demonstration sequences for "ABC" , "BC" and "C" conditions. Each child observed the experimenter performing all 5 action sequences in their condition.

18 children were excluded from the study because of demonstration error (4), equipment failure (3), lack of English (1), unavailable birth date (1), did not try toy (6), extreme distraction (2), never performed trial termination action (1).

Stimuli There were two novel toys: a blue ball with rubbery protuberances, and a stuffed toy with rings and tabs attached to it. Six possible actions could be demonstrated on each toy. Children were assigned to one of three experimental conditions. In each condition, they saw a different pattern of evidence involving five sequences of action and their outcomes. Each individual action sequence was always three actions long. In the "ABC" pattern, the same sequence of three actions (e.g. A=Knock, B=Stretch, C=Roll) is followed by a musical effect three times, while in the "BC" pattern a sequence composed of a different first action, followed by the same two-action subsequence (e.g. A=Squish, B=Pull, C=Shake and D=Flip, B=Pull, C=Shake) is followed by the effect three times (see Table 4). In both patterns, two additional sequences that end in C and do not contain BC fail to produce the effect. Finally, in the "C" pattern the sequences of actions were identical to those in the "BC" pattern, but the outcome was always positive. The number of times each individual action is demonstrated in each sequence position is identical in all three patterns. As we show later in the paper, our Bayesian ideal observer model confirms that the statistical evidence in each pattern supports different causal inferences.

Procedure The experimenter showed the child one of the toys, and said: "This is my new toy. I know it plays music, but I haven't played with it yet, so I don't know how to make it go. I thought we could try some things to see if we can figure out what makes it play music." The experimenter emphasized her lack of knowledge, so that the children would not assume she knew whether or not any of her actions were necessary. She then demonstrated one of the three patterns of evidence, repeating each three-action sequence (and its outcome) twice. The experimenter named the actions (e.g. "What if I try rolling it, and then shaking it, and then knocking on it?"), acted pleasantly surprised when the toy played music ("Yay! It played music'!'), or disappointed when it did not ("Oh. It didn't go"), and pointed out the outcome ("Did you hear that song?" or "I don't hear anything. Do you hear anything?"). After she demonstrated all five of the 3-action sequences, she gave the child the toy and said "Now it's your turn! why don't you try and make it play music". Throughout the experiment the music was actually triggered by remote activation. To keep the activation criteria uniform across conditions, the toy always played music the first time a child produced the final C action, regardless of the actions preceding it, terminating the trial. Only this first sequence of actions was used in our analysis.

Condition	Triplet	Double	Single	Other
"ABC"	20	1	2	4
"BC"	10	7	0	10
"C"	8	0	8	11

Table 5: Number of children producing each sequence type

Children were videotaped, and their actions from the time they were handed the toy to trial termination were coded by the first author, and 80% of the data was recoded by a blind coder. Coders initially coded each individual action as one of the six demonstrated actions, or as "novel". These sequences were then transferred into an "ABC" type representation, and subsequently coded as one of four sequence types: Triplet, Double, Single or Other (defined below). Inter-coder reliability was very high, with 91% agreement on the "ABC" type representations, and 100% agreement on sequence types.

Results and Discussion

Overall results are shown in Table 5. Children produced significantly different types of sequences across the three conditions, $p < 0.001$ (two-sided Fisher's exact test). We will discuss results for the "ABC" and "BC" conditions first, and then return to the "C" condition.

Effect of Statistical Evidence on Imitation In their imitation, children could either exactly reproduce one of the three-action sequences that had caused the toy to activate (that is, ABC in the "ABC" condition or ABC, DBC or EBC in the "BC" condition), or they could just produce BC in isolation. We refer to these successful three-action sequences as "triplets", and to the BC subsequence as a "double".

Both a triplet and a double reflect potentially correct hypotheses about what caused the toy to activate in both conditions. It could be that BC by itself causes the toy to activate in the "ABC" condition and the A is superfluous, or it could be that three actions are necessary in the "BC" condition, but the first action can vary. In both conditions BC is followed by the effect three times.

If children automatically encode the adult's successful actions as causally necessary, then they should exclusively imitate triplets in both conditions. However, if children are also using more complex statistical information, they should conclude that the BC sequence by itself is more likely to be causal in the "BC" condition than in the "ABC" condition, and that the triplet sequence is more likely to be causal in the "ABC" condition than in the "BC" condition. This is in fact what we found – the number of children producing triplets and doubles varied by condition, $p < 0.01$ (two-sided Fisher's exact test), and differed significantly between the "ABC" and "BC" conditions $p < 0.05$ (two-sided Fisher's exact test).

Effect of Differing Causal Outcomes on Imitation The pattern of evidence in the "BC" condition is more complex than in the "ABC" condition. This may have confused children, leading them to produce a variety of random actions, including BC. The "C" condition controls for this possibility. In this condition the sequences of actions were identical to those in the "BC" condition, but the outcome was always positive. As we show later, our Bayesian ideal observer model confirms that this provided statistical evidence for the hypothesis that C alone was sufficient to produce the effect.

In all three conditions, imitation of just the final C action in isolation was coded as a "single". As in the "ABC" and "BC" conditions, only the subsequence BC was coded as a double in the in the "C" condition. Also consistent with the "ABC" and "BC" conditions, in the "C" condition all five demonstrated successful sequences (ABC, ADC, DBC, AEC and EBC) were coded as triplets.

The "C" condition is as complex as the "BC" condition. However in the "C" condition the final action C produced by itself reflects a likely causal hypothesis. If children selectively imitate subsequences based on the data, then children in the "C" condition should produce C more frequently than children in the "BC" condition, and children in the "BC" condition should produce BC more frequently than children in the "C" condition. Our results support this hypothesis. Children in the "BC" and "C" conditions differed significantly in the overall types of sequences they produced, $p < 0.001$ (two-sided Fisher's exact test), and the number of children producing doubles and singles in the two conditions also varied significantly, $p < 0.001$, (two-sided Fisher's exact test).

Performance of "Other" Actions Across all three conditions, children did not just obligately imitate one of the successful sequences or subsequences they observed – they also produced new combinations of actions. Overall, the types of "other" sequences produced did not qualitatively differ across conditions, and appear to be a mix of exploratory behavior and genuine errors. There was a trend towards children in the "BC" and "C" conditions performing more of these "Other" sequences than children in the "ABC" condition $p = 0.10$, (two-sided Fisher's exact test). This difference becomes statistically significant when two children who imitated unsuccessful triplets are excluded from the analysis, $p < 0.05$, (two-sided Fisher's exact test). This result is compatible with findings that children tend to increase their exploratory behavior when the correct causal structure is more ambiguous (Schulz & Bonawitz, 2007; Schulz et al., 2008). Finally, four children performed completely novel actions they had never seen demonstrated. All of these children were in the "BC" or "C" conditions, consistent with these conditions eliciting more exploratory actions.

Model Results

Supporting our experimental results, our model makes distinct predictions in each of the three experimental conditions, showing that the data lead to differential causal inferences. Parameter values of $p = 0.1$ and $\beta = 1.0$ were chosen because they produced a qualitatively good match to children's performances, as shown in Figure 2. The relatively high value for β suggests that children prefer longer (complete) causal sequences, perhaps representing a pre-existing belief that adults usually don't perform extraneous actions. The relatively low value for p suggests that children employ a causal Occam's razor, assuming that simpler hypotheses, which require fewer causes to explain the data, are more likely. Overall, these results suggest that children's imitative choices conform closely to normative predictions.

Finally, while children performed similarly to our model's predictions, there were some differences in performance as well. Children produced more triplets than our model predicted, especially in the "ABC" condition. One reason for this discrepancy may be that children are able to use information about the knowledge state and intentional stance of the demonstrator that our current model cannot take into account. Models that can incorporate intentional and pedagogical information, in addition to statistical evidence are an important area of future work (Goodman, Baker, & Tenenbaum, 2009; Bonawitz et al., 2009). We are currently developing such a model, and exploring the role of pedagogical cues in children's imitation (Buchsbaum, Gopnik, Griffiths, & Shafto, submitted).

General Discussion

In this paper, we examined whether children are sensitive to statistical evidence in choosing the actions they imitate. We demonstrated that children can use statistical evidence to decide whether to imitate a complete action sequence, or to selectively imitate only a subsequence. In particular, children in the "ABC" condition imitated the complete sequence ABC more often than children in the "BC" condition, while children in the "BC" condition imitated the subsequence BC more often than children in the "ABC" condition. Children's performance in the "C" condition demonstrated that the differential imitation in the "ABC" and "BC" conditions could not be explained as a result of task complexity.

The design of this experiment also eliminated other simple explanations for these results. There were the same absolute number of BC demonstrations followed by effects in all three conditions, but children only produced doubles in the second condition. Similarly, the absolute number of positive triplet demonstrations was the same in the "ABC" condition and the "BC" condition, and was smaller than in the "C" condition, but children produced more triplets in the first condition than in the other two conditions. Finally, the actual sequence of actions was the same in the "BC" and "C" conditions but children behaved differently in the two cases. Children appeared to selectively imitate by considering the conditional probability of the various events and outcomes, and formulating a set of causal hypotheses based on that data. They then produced responses that matched the probability distribution of the hypotheses, at least qualitatively.

It is also worth noting the information-processing com-

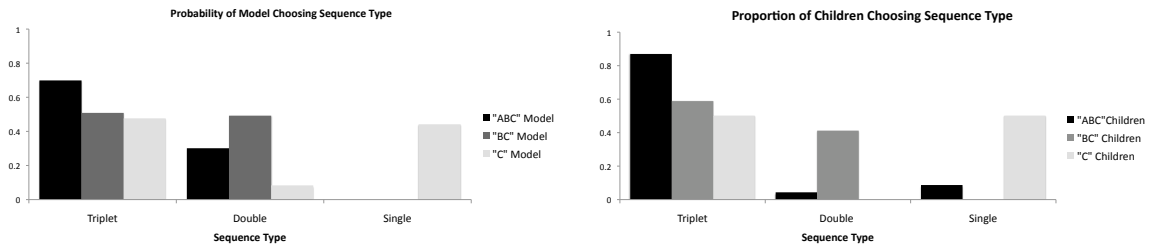


Figure 2: Left: Predictions of our Bayesian model. Right: Children's actual performance in Experiments 1 and 2.

plexity of this task. Children saw thirty similar actions and ten outcomes in each condition, and yet they appeared to track and use this information in deciding which actions to produce. This is consistent with other studies in which children and adults show surprising if implicit capacities to track statistical regularities.

These results extend earlier findings that show children take causal and intentional information into account appropriately in their imitation. They show that children also take into account statistical information about the conditional probability of events and do so in an at least roughly normative way. The studies also suggest a rational mechanism for the phenomenon of "overimitation" In particular, the "triplet" responses could be thought of as a kind of overimitation, reproducing parts of a causal sequence that are not actually demonstrably necessary for the effect. These results suggest that this behavior varies depending on the statistics of the data and the probability of various hypotheses concerning them.

Other factors may also influence the child's judgment of various causal hypotheses. For example, knowing that the adult is knowledgeable about the causal system, and is taking a "pedagogical stance" towards the evidence, may lead the child to different causal conclusions (Bonawitz et al., 2009). We are currently investigating the effect of pedagogical cues on imitation of causal action sequences (Buchsbaum et al., submitted). Similarly, seeing a repeated sequence of actions with no obvious physical causal outcome may lead children to suspect that the actions are intended to have a social or psychological rather than physical effect. Both these processes might lead to greater "overimitation" which would nonetheless be rational.

In general however, this study shows that children are sensitive to statistical information in determining which sequences of actions to imitate. Along with other studies, they support the idea that Bayesian procedures of statistical learning, procedures that allow the construction of causal models from statistical patterns, may play a significant role in many important kinds of early learning.

Acknowledgments. We thank Pat Shafto and Cari Kaufman for discussions on the model design, and Kimmy Yung, Mia Krstic, and Elon Ullman for help with data collection and coding. This material is based upon work supported by a National Science Foundation Graduate Research Fellowship, the McDonnell Foundation Causal Learning Initiative and Grant FA9550-07-1-0351 from the Air Force Office of Scientific Research.

References

Baldwin, D., Andersson, A., Saffran, J., & Meyer, M. (2008). Segmenting dynamic human action via statistical structure. *Cognition*, *106*, 1382-1407.

Bonawitz, L., Shafto, P., Gweon, H., Chang, I., Katz, S., & Schulz, L. (2009). The double-edged sword of pedagogy: Modeling the effect of pedagogy on preschoolers' exploratory play. *Proceedings of the 31st annual meeting of the Cognitive Science Society*.

Buchsbaum, D., Gopnik, A., Griffiths, T. L., & Shafto, P. (submitted). Children's imitation of causal action sequences is influenced by statistical and pedagogical evidence.

Buchsbaum, D., Griffiths, T. L., Gopnik, A., & Baldwin, D. (2009). Learning from actions and their consequences: Inferring causal variables from continuous sequences of human action. *Proceedings of the 31st Annual Conference of the Cognitive Science Society*.

Cheng, P. (1997). From covariation to causation: A causal power theory. *Psychological Review*, *104*, 367-405.

Fiser, J., & Aslin, R. N. (2002). Statistical learning of higher-order temporal structures from visual shape sequences. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *28*, 458-467.

Gergely, G., Bekkering, H., & Kiraly, I. (2002). Rational imitation in preverbal infants. *Nature*, *415*, 755.

Goodman, N. D., Baker, C. L., & Tenenbaum, J. B. (2009). Cause and intent: Social reasoning in causal learning. *Proceedings of the 31st Annual Conference of the Cognitive Science Society*.

Gopnik, A., Glymour, C., Sobel, D., Schulz, L., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review*, *111*, 1-31.

Gopnik, A., & Schulz, L. (2007). *Causal learning: Psychology, philosophy, computation*. New York: Oxford University Press.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*, 354-384.

Kirkham, N. Z., Slemmer, J. A., & Johnson, S. P. (2002). Visual statistical learning in infancy: evidence of a domain general learning mechanism. *Cognition*, *83*, B35-B42.

Lyons, D. E., Young, A. G., & Keil, F. C. (2007). The hidden structure of overimitation. *Proceedings of the National Academy of Sciences*, *104*, 19751-19756.

Meltzoff, A. N. (1995). Understanding the intentions of others: Re-enactment of intended acts by 18-month-old children. *Developmental Psychology*, *31*, 838-850.

Pearl, J. (1996). Structural and probabilistic causality. In D. R. Shanks, K. J. Holyoak, & D. L. Medin (Eds.), *The psychology of learning and motivation: Causal learning* (Vol. 34). San Diego: Academic Press.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month old infants. *Science*, *274*, 1926-1928.

Schulz, L. E., & Bonawitz, E. B. (2007). Serious fun: Preschoolers engage in more exploratory play when evidence is confounded. *Developmental Psychology*, *43*(4), 1045-1050.

Schulz, L. E., Hooppell, C., & Jenkins, A. C. (2008). Judicious imitation: Children differentially imitate deterministically and probabilistically effective actions. *Child Development*, *79*(2), 395-410.

Whiten, A., Custance, D. M., Gomez, J. C., Teixidor, P., & Bard, K. A. (1996). Imitative learning of artificial fruit processing in children (homo sapiens) and chimpanzees (pan troglodytes). *Journal of Comparative Psychology*, *110*(1), 3-14.

Williamson, R., Meltzoff, A. N., & Markman, E. (2008). Prior experiences and perceived efficacy influence 3-year-olds' imitation. *Developmental Psychology*, *44*, 275-285.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as bayesian inference. *Psychological Review*, *114*(2).

Zacks, J. M., Braver, T. S., Sheridan, M. A., Donaldson, D. I., Snyder, A. Z., Ollinger, J. M., et al. (2001, June). Human brain activity time-locked to perceptual event boundaries. *Nature Neuroscience*, *4*(6), 651-655.

Thinking with the Body

David Kirsh (kirsh@ucsd.edu)
Dept of Cognitive Science
University of California, San Diego

Abstract

To explore the question of physical thinking – using the body as an instrument of cognition – we collected extensive video and interview data on the creative process of a noted choreographer and his company as they made a new dance. A striking case of physical thinking is found in the phenomenon of marking. Marking refers to dancing a phrase in a less than complete manner. Dancers mark to save energy. But they also mark to explore the tempo of a phrase, or its movement sequence, or the intention behind it. Because of its representational nature, marking can serve as a vehicle for thought. Importantly, this vehicle is less complex than the version of the same phrase danced 'full-out'. After providing evidence for distinguishing different types of marking, three ways of understanding marking as a form of thought are considered: marking as a gestural language for encoding aspects of a target movement, marking as a method of priming neural systems involved in the target movement, and marking as a method for improving the precision of mentally projecting aspects of the target.

Keywords: Marking; multimodality; thinking, embodied cognition, ethnography.

1. Introduction

This paper explores how dancers and choreographers use their bodies to think about dance phrases. My specific focus is a technique called 'marking'. Marking refers to dancing a phrase in a less than complete manner. See fig. 1 for an example of hand marking, a form that is far smaller than the more typical method of marking that involves modeling a phrase with the whole body. Marking is part of the practice of dance, pervasive in all phases of creation, practice, rehearsal, and reflection. Virtually all English speaking dancers know the term, though few, if any, scholarly articles exist that describe the process or give instructions on how to do it.[1]

When dancers mark a phrase, they use their body's movement and form as a *representational vehicle*. They do not recreate the full dance phrase they normally perform; instead, they create a simplified or abstracted version – a model. Dancers mark to save energy, to avoid strenuous movement such as jumps, and sometimes to review or explore specific aspects of a phrase, such as tempo, movement sequence, or underlying intention, without the mental complexity involved in creating the phrase 'full-out'.

Marking is not the only way dancers 'mentally' explore phrases. Many *imagine* themselves performing a phrase. Some of the professional dancers we studied reported visualizing their phrase in bed before going to sleep, others reporting mentally reviewing their phrases while traveling on the tube on their way home. Our evidence suggests that marking, however, gives more insight than mental rehearsal: by physically executing a synoptic version of the whole phrase – by creating a simplified version externally – dancers are able to understand the shape, dynamics, emotion, and spatial elements of a phrase better than through imagination alone. They use marking as an anchor and vehicle for thought. It is this idea – that a body in motion can serve as an anchor and vehicle of thought – that is explored in this paper.

It is a highly general claim. It has been said that gesture can facilitate thought, [Golden Meadow 05]; that physically simulating a process can help a thinker understand a process [Collins et al 91], and that mental rehearsal is improved by overt physical movement. [Coffman 90] Why? What extra can physical action or physical structure offer to imagination? The answer, I suggest, is that creating an external structure connected to a thought – whether that external structure be a gesture, dance form, or linguistic structure – is part of an interactive strategy of bootstrapping thought by providing an anchor for mental projection. [Hutchins, 05, Kirsh 09, 10]. Marking a phrase provides the scaffold to mentally project more detailed structure than could otherwise be held in mind. It is part of an interactive strategy for augmenting cognition. By marking, dancers harness their bodies to drive thought deeper than through mental simulation and unaided thinking alone.

Hand Marking



Fig 1a



Fig 1b

In Fig 1a an Irish river dancer is caught in mid move. In 1b, the same move is marked using just the hands. River dancing is a type of step dancing where the arms are keep still. Typically, river dancers mark steps and positions using one hand for the movement and the other for the floor. Most marking involves modeling phrases with the whole body, and not just the hands.

[1] Search by professional librarians of dance in the UK and US has yet to turn up scholarly articles on the practice of marking.

2. Methodology

To explore the role of physical activity in dance cognition we were fortunate to study the creation of a new dance piece by the noted choreographer Wayne McGregor, the resident choreographer of the Royal Ballet in London. WM created the dance we studied with his own company, Random Dance, a group of ten extremely talented dancers. An eleventh dancer from a different company in Europe joined the group for the first period of dance creation.

The dance company's process of creation occurred in two phases: a three week episode at the University of California, San Diego (UCSD) in the winter of 2009; and a second period in London, in the late summer of 2009, just preceding the official première at Sadler's Wells Theater.

Method: During each phase, written notes were taken in real-time. During the UCSD phase, fifteen students took notes; during the London phase, a single experienced ethnographer took notes. Both phases, UCSD & London, were exhaustively videotaped using five high definition video cameras placed on the walls, and, whenever possible, two standard video cameras were placed on the ceiling. The whole rehearsing process, 11AM to 5PM, five to six days a week was captured. Video footage exceeds 110 hours (times 5-6 cameras) and captures all scheduled interactions between choreographer and dancers during the dance making process.

Cognitive ethnography requires acquiring a detailed knowledge of a community of practice, and then using that knowledge to illuminate specific episodes of activity. [Williams 06]. To acquire knowledge of the community of practice we interviewed the choreographer as well as the dancers repeatedly. We also reviewed all notebooks, and used our interviews as an opportunity to discuss specific moments of creative activity. The choreographer was interviewed for between forty and sixty minutes on digital video each morning and night. The dancers were interviewed at the end of each rehearsal, Our aim with the dancers was to have them reflect on specific elements of the rehearsal that day, and wherever possible, to show us through movement the dancerly decisions they made. Four dancers were selected and interviewed for thirty minutes each day. About 70 hours of interviews, in total, were videotaped.

To code the video we used ELAN, a free software system developed by the Max Planck Institute for Psycholinguistics, designed originally for studying gesture and small-scale interactions. Systematic audiovisual analysis depends on having a well-defined vocabulary of coding – a classification of activity and phenomena. After a few days of ad hoc coding a formal vocabulary was established by the whole team (20 people) to characterize ongoing activity. After the UCSD phase of capture, we reviewed the video data and selected special phenomena, such as marking for more detailed coding. In the London phase, we interviewed dancers explicitly about marking to probe them on their own views about marking. These interviews were undertaken in addition to the normal 30 minute ones we conducted. In several such sessions, we had the dancers come before the camera and dance *in full* a phrase they knew well; we then asked them to show us several ways they might *mark* that same phrase, and to describe the reasons they would mark one way versus another. We also interviewed them in a less structured manner, often returning to the question: "When do you mark, and how?" which led to multiple follow up questions and nuances of speech, as well as spontaneous performances from the dancers. The videotaped answers, with the corresponding gestures and markings, were transcribed and analyzed in detail with ELAN. On this basis, we created a hierarchical taxonomy of marking, yielding the three parent groups reported below. Intercoder reliability in distinguishing these parent marking types exceeded .9, on a sample of 25 video snippets of marking among our most experienced coders (n=3).

3. The Gross Function and Structure of Marking

At the highest level, three *functions* of marking can be distinguished.

1. *Marking-for-self:* dancers use their body to encode an aspect of a phrase for themselves. This may be for reinforcing memory, reflecting on sequence, or for scrutiny of spatial relations, among other reasons.
2. *Marking-for-others*, dancers use their bodies to encode an aspect of a phrase that others can focus attention on. For example, before a new performance, choreographer, choreographic assistant, and lighting manager review all phrases on stage for space.
3. *Joint-marking*: two or more dancers run through a phrase as a tightly coupled team, verifying timing and grips jointly for each other.

Small vs. Large Marking

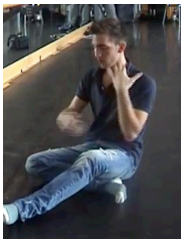


Fig 2a

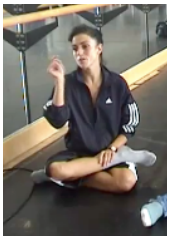


Fig 2b

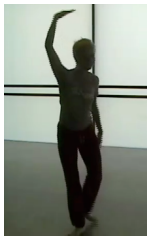


Fig 2c.

Figs 2a, 2b, 2c show the contrast between small and large marking. In 2a, a male dancer is remembering a step, using his hand to small mark it. In 2b, a female dancer is showing how she marks a pirouette. She uses a formal gesture for a pirouette that she learned as a ballet dancer. Her marking is small and conventional. In 2c, a second female dancer marks a phrase using movements that are of comparable size to those in the full phrase. She is clearly modeling the phrase.

There are also a few things to note, at the highest level, about the *structure* of marking.

Variability of size: Marking comes in a continuum of sizes, from very small to full size (but less energetically). In 'small marking', the amount of movement is minimal; the marking movements tend to be in the upper body (hands and head mainly), and the objective is to review the steps, the relationship between simultaneous movements (arm and leg together), and occasionally to attend to timing. See figs 2a and 2b. In extreme cases, such as Irish river dancing (fig 1), marking may be done exclusively with two fingers marking foot rhythm, position, and movement. When marking is very small, it is a form of gesture. In larger marking, especially when the function is to show the floor space required by a movement, or to show off the structure of a phrase to someone else, the movements may be full size but with less intent, emotion, or energy than the real movement (fig 2c). They are imperfect models of the complete phrase, but lacking certain attributes, such as intensity, motion dynamics, or fine detail.

Substitutability: A movement in one body part can represent the movement in another. Hand movements and head tilts regularly stand for the motion of different body parts: a hand movement may represent a leg movement, a head turn may represent a torso turn or a whole body turn; if the legs perform in parallel, one leg may stand in for two. This too is shown in figs 2a and 2b. See figs 3a, b for a standing version and fig 1 for finger version.

Idiosyncratic vs. Conventional Marking



3a. 3b.

Fig 3. In 3a a dancer marks a leg movement with his hands in his own idiosyncratic manner that is a hybrid of conventional ballet marking and personal style. In 3b A dancer from a strong ballet tradition offers a conventional small marking with her hands.

Conventional: In classical ballet and other formalized dance forms, dancers are taught to use specific gestures as ways of marking certain moves. These are a conventionalized form of small markings. For instance, as seen in fig 2b, the female dancer marks for the interviewer with her hand to show that, at a certain point in the phrase, a pirouette is required. In fig 3b she shows us a gesture for a *pas de bourrée.* These small gestures refer to a complex sequence of full moves well known by ballet dancers. We observed that dancers who do not rely on a ballet vocabulary still mark in a way that is reminiscent of ballet marking; but each dancer has personal idiosyncrasies that violate convention. In fig 3a, for instance, a dancer with deep training in both modern and ballet represents a leg movement with his arms, a hybrid marking that is part conventional and part personal gesture.

Aspectival: Marking typically represents an aspect of the full phrase, with some forms of marking focusing solely on tempo, others focusing on sequence, still others focusing on spatial position. For instance, when dancers mark for space they will keep the scale of the full phrase, but other aspects will be ignored or only partially represented, such as the dynamics of the phrase. At other times, just the movement of the upper body or the torso orientation may be marked and the movement of a leg or arm is left completely unmarked. Evidently, when dancers mark they are attending to only certain aspects of the phrase.

4. Analysis

Is it plausible to see marking as a vehicle of thought? There are a few promising ways to approach this question. Perhaps the most obvious line is that marking is a type of gestural semiotic system, possibly like a linguistic code. If gesture can function as a vehicle of thought, as some have argued, then why not marking?

It is useful to classify gestures according to where they lie on 'Kendon's Continuum' (McNeill 92). At one extreme, there are "gestures of the kind that Kendon has called 'quotable' … gestures that must be configured according to pre-established standards of form in order for them to function as signs, such as the OK sign among North Americans" (McNeill & Duncan 2000). These are compositional and behave in many respects like words or phrases in a language. At the other extreme are 'gesticulations'. These are idiosyncratic, created on the fly, and motivated by imagery rather than convention.

In dance, marking in the classical tradition of ballet is convention-driven and quotable. Despite individual differences in marking style, dancers still conform to general norms. Although marking conventions vary from ballet company to company, it does not take long for a professional dancer to pick up the idiosyncrasies of a company. This suggests there are rules determining the structure of ballet marking, and that local differences in marking style should be viewed as akin to differences in accent or handwriting. They need to be learned but are not different in principle than dialects of a common language.

In contemporary dance, the *reference* of marking – the phrases full-out, or aspects of those phrases – are not easily segmented. Movements in contemporary dance are freer, often novel. There are also far fewer conventions governing how dancers should mark. But not none. In the group we studied, for instance, there were quite strict rules about how to mark for the choreographer or his assistant. The spatial

dimensions of the phrase were to be preserved, though energy, and pace could be lessened.

The implication is that marking might well lie nearer the language side on Kendon's continuum than the gesticulation side. This needn't be a surprise. If there are written notation systems for encoding dance, such as Laban notation, then as long as marking is as expressive as these notation systems, anything that can be encoded on paper can be encoded through marking. The one requirement is that there be semantic rules for interpreting the paper notation and semantic rules for interpreting marking.

It is here, however, that the analogy with language fails. Marking is a reliable language only when a) dancers are *marking for others* – the other forms of marking lack adequate semantic rules*;* and, b) only when the point of marking is to display space, position, and structural form, all aspects of the full-out phrase that the choreographer or his assistant can directly see in the marking itself. If the point of marking were to call attention to movement sequence or to motor preparation, external observers would often be unable to infer the movements being sequenced or prepared for.

This is perhaps the key point. If someone states, "there is a circle with radius 30 meters", a competent interpreter need not have seen such a circle beforehand to know what the sentence means. It is enough to know the meaning of the terms 'circle', 'radius', '30 meters' to generate an interpretation. That is what semantic rules are for. By contrast, in marking, because there is so much idiosyncrasy in marking when dancers are marking for themselves, or when marking an aspect of a phrase that is not visibly similar to the full-out phrase (space), observers cannot 'see' the full-out move 'in' a marked version unless they already know what the full-out looks like. This explains why dancers rarely, if ever, mark a phrase they do not already know, and why choreographers never request dancers to show them novel phrases by marking – they insist on a full-out. Evidently, both parties need a clear idea of the target in advance of the marking. They have to have seen the full-out phrase to be able to 'project' it from its marking.

I believe this proves that much if not the majority of marking is not language like. It relies on prior acquaintance with the target, and then matching the mark to its target. That process more closely resembles a pattern completion process than a generative process of constructing the target. Languages are essentially generative, the point of marking is to avoid generating the whole target.

But if marking does not behave as a language this raises a paradox: if a dancer, or an observer, needs a clear idea of the full-out phrase in order to correctly interpret its marked version, why bother with the marking? How can marking ever be more powerful than inner visualization or imagination alone? What more can the physical manifestation of a movement add to the target already 'mentally grasped' through imagination?

One answer is that physical movement is helpful when one wants to measure the distance covered in a phrase. External distance is not guaranteed to be accurate in a mental representation. [Ledermen 87]. And there may be other physical dimensions available in the physical execution of a phrase that are only implicit in its mental representation (for instance, the physical tension in leaping off the floor or lifting another person).

But, beyond making physical attributes measurable, [see Kirsh 10], what extra *cognitive benefits* can physical marking provide that surpass mental rehearsal?

Here are two possibilities. They offer a different take on how marking might serve as a vehicle of thought.

1. Marking is a way of *anchoring projection* to a target. By providing a marked version of a target, a dancer can project a better representation of the target than imagination unaided. Marking, therefore, is a causally important way of augmenting thought. It is a component of a *distributed vehicle of thought,* consisting of an inner part and an outer part, which enables clearer thoughts. (cf. Hutchins 05)
2. Marking is a way of *priming* the neural system of a dancer, thereby enhancing imagination (or projection) by activating cortical elements that would be involved in the full-out movement. Marking is a way of enhancing the vividness and detail of imagination.

Marking as a method of anchoring projection. In the phenomenology of perception, a distinction can be drawn between perception, projection, and imagination. See fig 4.

- When we *perceive* an object, our experience is that we are seeing an object that is really there; we feel it is what causes our perception.
- When we *project* onto an object, we experience ourselves intentionally augmenting the object; we feel we partially cause our experience.
- When we *imagine* an object, we feel as if we are the sole cause of our imagined experience.

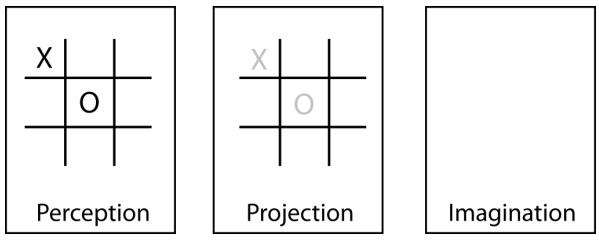


Fig 4. The difference between perception, projection, and imagination is represented here by three conditions of a tic-tac-toe game. Perception: subjects *see* moves. Projection: subjects see only the tic-tac-toe grid, and mentally *augment* it with moves. Imagination: subjects see a blank page and all aspects of the game are imagined – no external stimuli to scaffold or structure imagination.

The application to marking is shown in Fig 5. If the full-out phrase is represented by the complete triangle in 5a, marked versions are represented by 5b – 5e. The marked versions are either fractions or distortions of fractions of the full. But they support projection to full-out, if one has been exposed to the full-out already.

This form of projection is not a standard completion process. In completion, the target is a superset of the fragment. For example, *tang_ _ _* is a *stem* that supports completions like *tangent*. The fragment *ta_ g_ _ s* supports the completions *targets* or *tangles*. In both cases, the target completes the fragment. In projection, the structure that augments the fragment need not complete it because it may produce a new structure that has none of the subset structure. For instance, in 5c, the completion is larger in all dimensions except corner angle. In 5d and 5e, even the angles are not preserved. Projection is not completion.

Kirsh [09] showed that it is easier to conceptualize a target, or recover more memory of a target's structure, if there is something outside that one can 'lean on' for support. It is easier to project than to imagine if there is something helpful outside to support the projection. Recall is better for projected imagery than imagined imagery [ibid].

Marking as Projection



5a **5b** **5c** **5d** **5e**

Fig 5. The idea of marking as a sequence of illustrations of decreasing verisimilitude to the full phrase. 5a: a complete path at full scale. 5b: same path, full scale, shown by vertices and directions. 5c: smaller path, the interpreter must now know the scaling function. 5d: a stylized version of 5a. 5e, a smaller version of 5d, interpreter must project both shape, angles, and know the scaling function.

The relevance to marking is that when dancers mark, they may be creating a physical scaffold that facilitates projection. This would explain what 'extra' a dancer gets by physically marking a phrase rather than mentally rehearsing it. They get an external structure they can extrapolate from. This enables them to generate a conception of the final target that is more vivid, complete, and requiring less mental effort, than when they mentally rehearse without the support of overt movement. Moreover, dancers are able to choose how much extra memory support they want, just by marking more completely. When their mental image of the target is already clear, their marking may be minimal. When they have a weak mental image of the target, they may mark it more extensively, thereby increasing the vividness and control over their conception of the target.

Marking as a method of priming. A second benefit of marking may be that it involves more brain activity than mental rehearsal alone. It may facilitate muscle memory of details or deeper processing of movement goals.

The importance of muscle memory in dance is part of standard teaching. Muscle memory refers to the system of motor procedures – motor schemata – that have been stabilized through practice and are activated during performance. [Krakauer 06] Initial movements prime later movements. Priming also facilitates projection. Priming refers to an increased sensitivity to a stimulus due to prior exposure to a related stimulus. For instance, subjects who recently hear, see, think, and especially perform a particular movement will recognize aspects of that movement, sooner than those who have not. (Koch et al 04) The extent of priming is also a function of the depth of processing involved in the earlier exposure. [Challis, 92, Smith et al 83]. A person who thinks hard about a dance phrase – its energy, sequence, rhythm or spatial extent – will prime more choreographic relatives of the phrase, and prime them more deeply, than someone who merely sees the phrase briefly. Since motor preparation, spatial planning, and proprioceptic monitoring are involved in marking, it is likely that even more areas of cortex are involved in marking than in mental rehearsal alone. This suggests that during marking, there will be more opportunities for deeper processing – more chance to see deeper relations among movement components – than during mental rehearsal. Marking should prime the phrase more deeply, making it easier to remember it in the future.

If marking helps a dancer to envision the target phrase better, it helps to explain why marking is beneficial. Given the importance of internal processes, however, marking is best understood as the external part of an internal-external process. It is best seen as the external part of a distributed vehicle of thought.

5. Conclusion

I have argued that marking is a form of physical thinking. A dancer creates a partial version of a phrase, attends to it while creating it, and because of processes like priming and projection, the dancer is able to understand something deeper about the phrase's structure than through imagination alone. When dancers mark, they are closely coupled with the dance product they are externalizing. *They rely on that product to* think with. Their performance of the marked phrase is part of their ongoing process of grasping the phrase. In some ways, their relation to marked material is reminiscent of what E. M. Forster (27) said about language: "How can I know what I'm thinking until I see what I say". For Forster, the external vehicle of a thought – its linguistic formulation – was a real time achievement of putting the thought into words. It made the thought more precise in virtue of the constraints of language. There was no point asking whether the articulated content was the same as some internal version already encoded in an internal language intrinsically understood, as suggested by Fodor (75) and others. For Forster, as well as for Wittgenstein (51), the articulation is part of the thinking process.

My suggestion, here, is that for a dancer, Forster's rhetorical question can be rephrased as: "How can I know

what my phrase really is until I see what I do?" A dancer's thought of his or her phrase is partly shaped by what is marked. Dancers do think about their phrases without dancing them or marking them. But, by *marking-for-self* dancers *think better* about their full-out phrase. Physical movement replaces mental computation. Instead of imagining transformations, they execute them externally. Marking is part of a distributed vehicle of thought with internal and external parts closely coupled.

Acknowledgements: I gratefully acknowledge help from Dr. Dafne Muntanyola on ethnographic analyses of the dance data, from twenty five students in my class on Creative Cognition in Dance, from Wayne McGregor and the dancers in Random Dance, from Scott delaHunta the director of Random Research, and from the Committee on Academic Research, UCSD for a seed grant for expenses.

References

Challis, B; Brodbeck, D. (1992). Level of processing affects priming in word fragment completion. Journal of Experimental Psychology: Learning, Memory, and Cognition. Vol 18(3), May 1992, 595-607

Clark, A. (2008). Supersizing the Mind. Oxford: Oxford University Press.

Coffman, D., (1990). Effects of Mental Practice, Physical Practice, and Knowledge of Results on Piano Performance. Journal of Research in Music Education, Vol. 38, No. 3, 187-196

Collins, A., Brown, J. S., A Holum. (1991) Cognitive apprenticeship: Making thinking visible. American Educator, 1991

Fodor, Jerry A. (1975). The Language of Thought, Cambridge, Massachusetts: Harvard University Press.

Forster, E.M. (1927) Aspects of the Novel. Orlando: Harcourt.

Golden-Meadow (2005) Hearing Gestures: How Our Hands Help Us to Think. Harvard University Press.

Hutchins, E. Material anchors for conceptual blends. Journal of Pragmatics Volume, 37, Issue 10, 2005, pp. 1555-1577

Kirsh, D. (2009a). Projection, Problem Space and Anchoring. In N. A. Taatgen & H. van Rijn (Eds.), Proceedings of the 31st Annual Conference of the Cognitive Science Society (pp. 2310-2315). Austin, TX: Cognitive Science Society.

Kirsh, D. (2010). Thinking With External Representations. AI and Society: Journal of Knowledge, Culture and Communication. London Springer. Vol 25.4

Kirsh, D. (2009a). Problem Solving and Situated Cognition. in Phillip Robbins & M. Aydede (eds.) The Cambridge Handbook of Situated Cognition. Cambridge: CUP.

Kirsh, D, et al. (2009c). Choreographic Methods for Creating Novel, High Quality Dance. Proceedings, DESFORM 5th International Workshop on Design & Semantics & Form.

Koch I., Keller P., Prinz W. (2004). The Ideomotor Approach To Action Control: Implications For Skilled Performance. Int. Journal of Sport and Exercise Psychology, 2, 362-375

Krakauer,J.W., & Shadmehr, R. (2006). Consolidation of motor memory. Trends in Neurosciences, 29: 58-64.

Lederman, S. J.; Klatzky, R.; Collins, A; Wardell, J. (1987). Exploring environments by hand or foot: Time-based heuristics for encoding distance in movement space. Journal of Experimental Psychology: Learning, Memory, and Cognition. Vol 13(4), Oct 1987, 606-614

McNeill, D. (1992). Hand and mind: What gestures reveal about thought. Chicago: University of Chicago Press

McNeill D. & Duncan S, (2000) Growth Points In Thinking-For-Speaking, D. McNeill (ed.), Language and Gesture, pp. 141-161. Cambridge University Press

Muntanyola, D, Kirsh, D. (2010) Marking as Physical Thinking: A Cognitive Ethnography of Dance. Donosti.

Newell, A, Simon, H., (1976), "Computer Science as Empirical Inquiry: Symbols and Search", Communications of the ACM, 19

Pylyshyn, Zenon, (1986) Computation & cognition: Toward foundation for cognitive science. Cambridge: MIT Press.

Smith, M C., et al., (1983). The relationship between contextual facilitation and depth of processing. Journal of Experimental Psychology: Learning, Memory, & Cognition. Vol 9(4), 697-712.

Williams, R.F. (2006). Using Ethnography to Study Instruction. Proceedings 7th International Conference of the Learning Sciences. Mahwah: Lawrence Erlbaum.

Wittgenstein, L (1951). Philosophical Investigations. Oxford: Basil Blackwell.

The More the Merrier? Examining Three Interaction Hypotheses

Min Chi (minchi@cs.cmu.edu)
Machine Learning Department, Carnegie Mellon University, 5000 Forbes Avenue
Pittsburgh, PA 15213 USA

Kurt VanLehn (Kurt.Vanlehn@asu.edu)
School of Computing, Informatics and Decision Systems Engineering, Arizona State University
Tempe, AZ 85287 USA

Diane Litman (litman@cs.pitt.edu)
Department of Computer Science, University of Pittsburgh, 210 South Bouquet Street
Pittsburgh, PA 15260 USA

Abstract

While high interactivity is one of the key characteristics of one-on-one human tutoring, a great deal of controversy surrounds the issue of whether interactivity is indeed the key feature of tutorial dialogue that impacts students' learning. In this paper we investigate three interaction hypotheses: a widely-believed monotonic interactivity hypothesis, a better supported interaction plateau hypothesis, and our tactical interaction hypothesis. The monotonic interaction hypothesis predicts that increasing interactivity causes an increase in learning; the plateau hypothesis states that increasing interactivity yields increasing learning until it hits a plateau, and further increases in interactivity do not cause noticeable increases in learning. Finally, the tactical interaction hypothesis predicts that interactivity only increases learning when interactions are guided by effective tutorial tactics. In this paper, we examine each hypothesis in the context of an empirical study, the results of which support the tactical interaction hypothesis.

Keywords: machine learning; reinforcement learning; pedagogical strategy; Intelligent Tutoring Systems.

Introduction

One-on-one tutoring is a highly effective educational intervention. Tutored students often perform significantly better than students in classroom settings (Bloom, 1984). Computer learning environments that mimic aspects of human tutors have also been highly successful. Intelligent Tutoring Systems (ITSs) have been shown to be highly effective at improving students' learning in real classroom settings (Koedinger et al., 1997; VanLehn, 2006). A key characteristic of one-on-one tutoring, with both human and computer tutors, is high interactivity.

A common assumption, often referred as the *monotonic interaction hypothesis* (VanLehn, Graesser, et al., 2007) is that greater interactivity leads to greater learning.

However, several studies have failed to confirm this hypothesis. Experiments with human tutors found no significant differences in learning gains when content was carefully controlled and interactivity was directly manipulated (M. T. H. Chi et al., 2001, 2008; Rose et al., 2001). Experiments that compared human tutors and several Natural Language dialogue-based computer tutors also found no significant differences in learning as interactivity varied across students (Evens & Michael, 2006; VanLehn, Graesser, et al., 2007; Reif & Scott, 1999; Katz et al., 2003; Fossati et al., 2008). In a meta-analysis of the tutoring literature, VanLehn found little support for the monotonic interactivity hypothesis and instead proposed the interaction plateau: the hypothesis that increased interactivity increases learning up to a point (roughly, the level of interactivity afforded by conventional step-based ITSs); beyond that threshold, however it does not yield any noticeable increases in learning (VanLehn, submitted).

On the other hand, for any form of tutoring the tutor's behaviors can be viewed as a sequential decision processes wherein, at each discrete step, the tutor is responsible for selecting the next action to take. Each of these tutorial decisions affects successive actions. Some existing theories of learning suggest that when making tutorial decisions, a tutor should adapt its actions to the students' needs based upon their current knowledge level, affective state, and other salient features (Vygotsky, 1971; Collins et al., 1989; Koedinger & Aleven, 2007). Most studies cited above made use of human tutors for their highly-interactive condition, simply assuming that expert tutors will take optimal actions. However, Chi et al. and others have argued that human tutors may not always make optimal tutorial decisions (M. T. H. Chi et al., 2001, 2008). Given that tutoring is a rather complex procedure and tutors have to make many decisions fairly rapidly, even expert human tutors may not take the full advantage of the tutorial alternatives.

Therefore, in this paper we propose a third hypothesis: the *tactical interaction hypothesis*. It states that interactivity only increases learning when interactions are guided by effective tutorial tactics. By "tutorial tactics" we refer to the policies used for selecting the tutorial action taken at each step when there are multiple actions available. In other words, we hypothesize that the tutors' success will not be governed by *how often* they give interactive prompts or ask the students questions but *how well*.

To investigate the three hypotheses, we focused on two tutorial actions: elicit and tell. During the course of one-on-one tutoring, tutors often face a simple question, should they *elicit* the next step information from the student, or should they *tell* the student the next step directly? There are many theories, but no widespread consensus on how or when an elicit or a tell

(a) Elicit Version	(b) Tell Version
1. **T:** So let's start with determining the value of v1. 2. **T:** Which principle will help you calculate the rock's instantaneous magnitude of velocity at T1? **{ELICIT}** 3. **S:** definition of kinetic energy 4. **T:** Please write the equation for how the definition of kinetic energy applies to this problem at T1 **{ELICIT}** 5. **S:** ke1 =0.5*m*v1^2 6. **T:** From KE1 = 0.5*m*v1^2, $\cdots$	1. **T:** So let's start with determining the value of v1. 2. **T:** To calculate the rock's instantaneous magnitude of velocity at T1, we will apply the definition of kinetic energy again. **{TELL}** 3. **T:** Let me just write the equation for you: KE1 = 0.5*m*v1^2. **{TELL}** 4. **T:** From KE1 = 0.5*m*v1^2, $\cdots$

Figure 1: Elicit vs. Tell

should be taken (Vygotsky, 1971; Aleven et al., 2004; Collins et al., 1989). Generally speaking, eliciting more information from the student will result in a more interactive tutorial dialogue. Figure 1 compared a pair of dialogues extracted from logs in this study. Both dialogues begin and end with the same tutor turn (lines 1 and 6 in (a) and 1 and 4 in (b)). In dialogue (a) the tutor chooses to elicit twice (lines 2-3 and 4-5 respectively). Dialogue (b), by contrast, *covers the same domain content* with two tell actions (lines 2 and 3). As a consequence, dialogue (a) is more interactive than (b).

In this paper, we quantify the *interactivity* of a dialogue via the Interactivity ratio (I-ratio) which we define as the number of elicitation decisions divided by the total number of elicit or tell decisions in a given dialogue. The higher this value, the more interactive the tutorial dialogue.

$$\mathbf{I-ratio} = \frac{N_{Elicit}}{N_{Elicit} + N_{Tell}} \quad (1)$$

Unlike the monotonic and plateau hypotheses, validation of the tactical interaction hypothesis requires effective tutorial tactics. In most computer learning environments the pedagogical tutorial tactics are hard-coded rules designed to implement preexisting cognitive and/or pedagogical theories. Typically, these theories are considerably more general than the specific interaction decisions that designers must make. This makes it difficult to tell if a specific policy is consistent with the theory. Moreover, it is often difficult to empirically evaluate these tactics because the tutor's overall effectiveness depends upon many factors, such as the usability of the system, how easily the dialogues are understood, and so on. Ideally, several versions of a system are created, each employing different tutorial tactics. Data is then collected with human subjects interacting with these different versions of the system and the results are compared. Due to the high cost of experiments, however, only a handful of policies are typically explored. Yet, many other reasonable policies are possible.

In recent years, work on the design of dialogue systems has involved several data-driven methodologies. Among these, Reinforcement Learning (RL) has been widely applied(Singh et al., 2002). In this work, rather than implementing pedagogical policies drawn from human experts or theories, we applied and evaluated RL to derive pedagogical tutorial tactics using pre-existing interactivity data.

General Approach

For this study, we induced two sets of tutorial tactics: the Normalized Gain (NormGain) tactics, derived with the goal of making tutorial decisions that contribute to students' learning, and the Inverse Normalized Gain (InvNormGain) tactics, induced with the goal of making less beneficial, or possibly useless, decisions. The two sets were then compared with human students on Cordillera (VanLehn, Jordan, & Litman, 2007), a Natural Language Tutoring System teaching students introductory college physics. Using Cordillera in lieu of human tutors allowed us to rigorously control the content and vary only the interactivity. In order to avoid artifacts due to imperfect natural language understanding, Cordillera incorporated a human wizard whose sole task was to rapidly match students' actual utterance to one of the expected student utterances displayed in a menu. The wizard made no tutorial decisions.

In the learning literature, it is commonly assumed that relevant knowledge in domains such as math and science is structured as a set of independent but co-occurring Knowledge Components (KCs) and that KC's are learned independently. A KC is "a generalization of everyday terms like concept, principle, fact, or skill, and cognitive science terms like schema, production rule, misconception, or facet" (VanLehn, Jordan, & Litman, 2007). For the purposes of tutoring, these are the atomic units of knowledge. It is assumed that a tutorial dialogue focusing on a single KC will not affect the student's understanding of any other KC. This is an idealization, but it has served developers well for many decades, and is a fundamental assumption of many cognitive models (Anderson, 1983; Newell, 1994). When dealing with a specific KC, the expectation is that the tutor's best policy for teaching that KC (e.g., when to Elicit vs. when to Tell) would be based upon the student's mastery of the KC in question, its intrinsic

difficulty, and other relevant, but not necessarily known, factors specific to that KC. In other words, an optimal policy for one KC might not be optimal for another. In this study, we focused on eight KCs. We induced eight policies and conducted eight tests of the three hypotheses, one per KC.

Later results indicated that on average the percentage of elicit prompts students received during the tutoring is more than 70% for both groups in this study, thus based on the standard set in (VanLehn, submitted) the tutorial dialogues reported here are well beyond the threshold of the level of interactivity afforded by conventional step-based ITSs. Therefore, we expect that on each KC:

1. If the monotonic hypothesis is correct, the group that learned more would have a higher I-ratio.
2. If the interaction plateau hypothesis is correct, both NormGain and InvNormGain students would learn equally well regardless of interactivity difference.
3. If the tactical interaction hypothesis is correct and our RL-based tutorial tactics are indeed effective, NormGain students would learn more than InvNormGain peers regardless of interactivity difference.

First we will briefly describe how we apply machine learning to induce tutorial dialogue tactics. Then we will describe our study and its results.

Applying RL to Induce Tutorial Tactics

Much of the previous research on the use of RL to improve dialogue systems has typically used Markov Decision Problems (MDPs) (Sutton & Barto, 1998) to model dialogue data (Singh et al., 1999). An MDP formally corresponds to a 4-tuple (S,A,T,R), in which: $S = \{S_1, \cdots, S_n\}$ is a state space; $A = \{A_1, \cdots, A_m\}$ is an action space represented by a set of action variables; $T : S \times A \times S \rightarrow$ *[0, 1]* is a set of transition probabilities $P(S_j|S_i,A_k)$, which is the probability that the model would transition from state S_i to state S_j after the agent takes action A_k; $R : S \times A \times S \rightarrow R$ assigns rewards to state transitions. Finally, $\pi : S \rightarrow A$ is defined as a policy, which determines which action the agent should take in each state in order to maximize the expected reward.

The central idea behind our approach is to transform the problem of inducing effective pedagogical tactics into computing an optimal policy for choosing actions in an MDP. Inducing pedagogical tactics can be represented using an MDP: the states S are vector representations composed of relevant student-tutor interaction characteristics; $A = \{Elicit, Tell\}$ in this study, and the reward function R is calculated from the system's success measures and we used learning gains. Once the (S, A, R) has been defined, the transition probabilities T are estimated from the training corpus, which is the collection of dialogues, as: $T = \{p(S_j|S_i,A_k)\}_{i,j=1,\cdots,n}^{k=1,\cdots,m}$. More specifically, $p(S_j|S_i,A_k)$ is calculated by taking the number of times that the dialogue is in state S_i, the tutor took action A_k, and the dialogue was next in state S_j divided by the number of times the dialogue was in S_i and the tutor took A_k. Once a complete MDP is constructed, a dynamic programming approach can be used to learn the optimal control policy π^* and here we used the toolkit developed by Tetreault and Litman (Tetreault & Litman, 2008).

In this study, the reward functions for inducing both the NormGain and the InvNormGain sets were based on Normalized Learning Gain (NLG) defined as: $NLG = \frac{posttest - pretest}{1 - pretest}$ because it measures a student's gain *irrespective of his/her incoming competence*. Here *posttest* and *pretest* refer to the students' test scores before and after the training respectively; and 1 is the maximum score. More specifically, the NormGain tutorial tactics induced by using the student's $NLG \times 100$ as the final reward while the InvNormGain ones was induced by using the student's $(1 - NLG) \times 100$ as the final reward. Apart from the reward functions, the two sets were induced using the same general procedure.

In order to learn a policy for each KC, we annotated our tutoring dialogues and action decisions based on which KCs a tutor action or tutor-student pair of turns covered (kappa ≥ 0.77 for each of the eight KCs). Additionally, we have mapped students' pre- and post-test scores to the relevant KCs for each test item. The rest of this section presents a few critical details of the process, but many others must be omitted to save space. Overall, the RL approach in this study differed from that of the previous study (M. Chi et al., 2009) in many aspects. First, we have three training corpora in this study: the Exploratory corpus collected in 2007, the DichGain corpus collected in 2008, and a Combined training corpus. Second, in order to examine a range of possible tactics we included 50 features based upon six categories of features considered by previous research(Moore et al., 2004; Forbes-Riley et al., 2007) to be relevant. Additionally, we also used a different method of searching the power set of the 50 features. Finally we directly used the $NLG \times 100$ for inducing NormGain policies and$(1 - NLG) \times 100$ for inducing InvNormGain ones instead of dichotomizing the NLGs when inducing policies previously.

Figure 2 shows an example of a learned NormGain policy on one KC, "Definition of Kinetic Enegy". The policy involves three features:

[StepDifficulty:] encodes a step's difficulty level. Its value is estimated from the students' log files based on the percentage of correct answers given on the step.

[TutorConceptsToWords:] which represents the ratio of the physics concepts to words in the tutor's dialogue. This feature also reflects how often the tutor has mentioned physics concepts overall.

[TutorAvgWordsSession:] The average number of words in the tutor's turn in this session. This feature reflects how verbose the tutor is in the current session.

MDP generally requires discrete features and thus all the continuous features need to be discretized. The top half of Figure 2 lists how each of the three features was discretized. For example, For StepDifficulty, if its value is above 0.38, it

[Feature:]

StepDifficulty:	$[0, 0.38) \rightarrow 0$;	$[0.38, 1] \rightarrow 1$
TutorConceptsToWords:	$[0, 0.074) \rightarrow 0$;	$[0.074, 1] \rightarrow 1$
TutorAvgWordsSession:	$[0, 22.58) \rightarrow 0$;	$[22.58, \infty) \rightarrow 1$

[Policy:]
Elicit: 0:0:0 0:0:1 1:0:1 1:1:0 1:1:1
Tell: 0:1:0
Else:0:1:1 1:0:0

Figure 2: A NormGain Policy on KC_{20} For ET Decisions

is 1 (difficult) otherwise, it is 0 (easy). The lower half of Figure 2 shows there are 8 rules learned: in 5 situations the tutor should elicit, in one situation it should tell; in the remaining 2 cases either will do. For example, when all three features are zero (which means when the step is easy, the tutor ratio of physics concepts to words so far is low, and the tutor is not very wordy in the current session), then the tutor should elicit as 0:0:0 is listed next to the [elicit]. As you can see, three features already provide relatively complex tutorial tactics and the induced policies were not like most of the tutorial tactics derived from analyzing human tutorial dialogues.

The resulting NormGain and InvNormGain policies were then implemented back into Cordillera yielding two new versions of the system, named NormGain-Cordillera and InvNormGain-Cordillera respectively. The induced tutorial tactics were evaluated on real human subjects to see whether the NormGain students would out-perform the InvNormGain peers.

Methods

Participants

Data were collected over a period of two months during the summer of 2009. Participants were 64 college students who received payment for their participation. They were required to have a basic understanding of high-school algebra. However, they could not have taken any college-level physics courses. Students were randomly assigned to the two conditions. Each took from one to two weeks to complete the study over multiple sessions. In total, 57 students completed the study (29 in the NormGain condition and 28 in the InvNormGain condition).

Domain & Procedure

The tutoring addressed work-energy problem solving from a first-year college physics course. The eight primary KCs were: the weight law (KC_1), definition of work (KC_{14}), Definition of Kinetic Energy (KC_{20}), Gravitational Potential Energy (KC_{21}), Spring Potential Energy (KC_{22}), Total Mechanical Energy (KC_{24}), Conservation of Total Mechanical Energy (KC_{27}), and Change of Total Mechanical Energy (KC_{28}).

All participants in the study followed the same procedures and used the same training and testing materials as were used when collecting the training corpora. More specifically, the participants all: completed 1) a background survey; 2) read a text covering the target domain knowledge; 3) took a pretest; 4) solved the same seven training problems in the same order on Cordillera; and 5) finally took a posttest. The pretest and posttest were identical. Except for following the policies (NormGain vs. InvNormGain), the remaining components of Cordillera, including the GUI interface, the same training problems, and the tutorial scripts, were identical for all students.

Grading

The tests contained 33 test items covering 168 KC occurrences. Each occurrence was graded by a single experienced grader who was not aware of the study condition from which it arose. These were then summed and normalized to the range of [0,1]. Other grading rubrics were also tried. They presented the same pattern of results as the ones presented next.

Results

No significant difference was found between the two conditions in terms of the total training time spent on Cordillera: $t(55) = 0.27, p = .79$. The NormGain group spent ($M =$ 259.98 mins, $SD = 59.22$) and the InvNormGain group spent ($M = 264.57$ mins, $SD = 67.60$). For each student, Cordillera had made on average 260 decisions on whether to Elicit or to tell during the training and on a KC by KC basis, the number of such decisions varies from 4 on KC_1 to 72 on KC_{20}.

Learning Performance

First, we investigated whether students learned by training on Cordillera. A one-way ANOVA was used to test for learning performance differences between the pre- and posttests. Both groups made reliable learning gains from pre-test to post-test: $F(1,56) = 31.34$, $p = .000$ for the NormGain condition and $F(1,54) = 6.62$, $p = .013$ for the InvNormGain condition respectively. On a KC by KC basis, the NormGain conditions learned reliably on all the eight primary KCs while the InvNormGain learned reliably on five primary KCs save for KC14, KC22, and KC28.

Next, we compared the learning performance between the two conditions. Random assignment appears to have balanced the incoming student competence across conditions. There were no statistically significant differences between the two conditions on the mathSAT scores nor in the pretest scores: $t(55) = 0.71, p = .48$. On a KC by KC basis, no significant difference was found between the two conditions across all eight primary KCs except that on KC27, the NormGain group score marginally higher than the InvNormGain group: $t(55) = 1.74, p = 0.088$ (see Table 1). In order to account for varying pretest scores, the adjusted Post-test scores were compared between the two conditions by running an ANCOVA using the corresponding pre-test score as the covariate.

The NormGain condition out-performed the InvNormGain on the overall adjusted posttest scores: $F(1,54) =$

Table 1: Between-Group Comparison on Pre-Test and Adjusted Post-Test Scores Across Primary KCs

KC	TestScore	NormGain	InvNormGain	Stat	d
KC_1	Pretest	0.42 (0.15)	0.39 (0.22)	$t(55) = 0.66, p = 0.51$	0.16
	Adjusted Posttest	0.64 (0.12)	0.54 (0.12)	$F(1,54) = \mathbf{9.80}, p = \mathbf{0.0028}$	**0.85**
KC_{14}	Pretest	0.43 (0.23)	0.44 (0.25)	$t(55) = -0.17, p = 0.86$	-0.04
	Adjusted Posttest	0.65 (0.17)	0.53 (0.17)	$F(1,54) = \mathbf{6.47}, p = \mathbf{0.014}$	**0.72**
KC_{20}	Pretest	0.38 (0.17)	0.37 (0.22)	$t(55) = 0.31, p = 0.76$	0.05
	Adjusted Posttest	0.67 (0.11)	0.58 (0.11)	$F(1,54) = \mathbf{10.30}, p = \mathbf{0.002}$	**0.83**
KC_{21}	Pretest	0.45 (0.20)	0.43 (0.24)	$t(55) = 0.35, p = 0.72$	0.09
	Adjusted Posttest	0.75 (0.13)	0.65 (0.13)	$F(1,54) = \mathbf{7.62}, p = \mathbf{0.008}$	**0.78**
KC_{22}	Pretest	0.42 (0.25)	0.39 (0.26)	$t(55) = 0.41, p = 0.68$	0.12
	Adjusted Posttest	0.63 (0.17)	0.51 (0.17)	$F(1,54) = \mathbf{7.77}, p = \mathbf{0.007}$	**0.72**
KC_{24}	Pretest	0.46 (0.15)	0.41 (0.23)	$t(55) = 0.89, p = 0.38$	0.26
	Adjusted Posttest	0.64 (0.11)	0.58 (0.11)	$F(1,54) = \mathbf{4.22}, p = \mathbf{0.045}$	**0.56**
KC_{27}	Pretest	0.53 (0.21)	0.42 (0.24)	$t(55) = 1.74, p = 0.088$	0.5
	Adjusted Posttest	0.74 (0.18)	0.63 (0.18)	$F(1,54) = \mathbf{5.88}, p = \mathbf{0.019}$	**0.62**
KC_{28}	Pretest	0.37 (0.20)	0.36 (0.26)	$t(55) = 0.13, p = 0.90$	0.04
	Adjusted Posttest	0.53 (0.17)	0.47 (0.17)	$F(1,54) = 1.61, p = 0.21$	0.36

$10.689, p = .002, d$[1] $= 0.86$. On a KC by KC basis, Table 1 summarize the comparisons on the pre-test and adjusted posttest scores between the two conditions. The third and fourth columns in Table 1 list the means and SDs σ of the NormGain and InvNormGain groups' pretest or adjusted posttest scores on the corresponding KC. The fifth column lists the corresponding statistical comparison and the sixth column lists the Cohen's d of the comparison. Table 1 shows that the NormGain condition out-performed the InvNormGain across all primary KCs (in bold) except for KC28, on which no significant difference was found between the two groups.

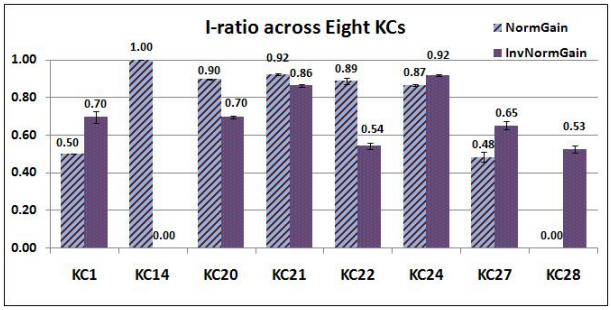


Figure 3: Compare I-ratio Across KCs

I-ratios

We next investigated the interactive characteristics of the derived tutorial tactics by comparing the tutorial dialogues' I-ratios between the two groups. Surprisingly, there were no significant differences between the two groups on the overall I-ratio: $t(55) = -0.395, p = 0.694$. More specifically, we have $M = 0.758, SD = 0.073$ (maximum is 1) for the NormGain group and $M = 0.763, SD = 0.018$ for the InvNormGain group respectively.

However, once the results were examined on a KC by KC basis there were significant differences between the two groups on each of the eight primary KCs . Figure 3 shows that the NormGain condition was more likely to get elicits than the InvNormGain condition on KC_{14}, KC_{20}, KC_{21}, and KC_{22}; and the InvNormGain condition was more likely to get elicits than the NormGain condition on KC_1, KC_{24}, KC_{27}, and KC_{28}.

[1]Cohen's d, which is defined as the mean learning gain of the experimental group minus the mean learning gain of the control group, divided by the groups' pooled standard deviation.

Examining The Three Interaction Hypothesis

The monotonic interactivity hypothesis states that more interactivity should lead to increased learning. Because the NormGain group learned more than the InvNormGroup across all eight KCs except KC_{28}, which was a null result, the NormGain group should also have a larger I-ratio on all seven KCs. From Figure 3, it was shown that this was *not* the case for KC_1, KC_{24} and KC_{27}. Thus, our data are not consistent with the monotonic interactivity hypothesis.

The interaction plateau hypothesis states that increasing interactivity yields increasing learning until it hits a plateau, and further increases in interactivity do not cause noticeable increases in learning. The main difference between this hypothesis and monotonic interactivity hypothesis is once beyond a certain level of interactivity whether increasing interaction would impact students' learning gain or not. In order to test this hypothesis, we mainly focused on the six KCs (all but KC_{14} and KC_{28}). This is because on these six KCs both NormGain and InvNormGain groups' I-ratios were more than 48% (see Figure 3) which is well beyond the threshold of the level of interactivity afforded by conventional step-based ITSs based on the definition set in (VanLehn, submitted). If

the interaction plateau hypothesis is true, then the NormGain group should learn just as much as the InvNormGain group on each of the six KCs. Table 1 however shows that the NormGain group learned more than the InvNormGain group across all six KCs. Thus, the interaction plateau hypothesis is not consistent with our data.

Finally, the tactical interaction hypothesis states that interactivity does not increase learning unless they are governed by effective tutorial tactics. If this is true and all our derived RL-based policies were indeed effective, the NormGain group would learn more than the InvNormGain group across all KCs. This hypothesis was supported by seven of the KCs, and on KC_{28} there was only an unreliable trend in the expected direction. Thus, of all three hypotheses, the tactical interaction hypothesis receives the most support from our data.

Discussion

Overall, our results inform the ongoing discussion of Socratic vs. didactic tutoring by suggesting that a tutor's success is not governed by *how often* they prompt or ask the students questions but *how well*. In particular, the reason human tutors so often failed to be more effective than simple, unoptimized dialogue-based tutors in those previous studies may be that effective policies for tutorial interaction are complex and not easily derived from the tutors' experience. This in turn suggests that an optimized dialogue-based tutoring system, such as NormGain-Cordillera, would be potentially even more effective than expert human tutors. Although controlling for content is difficult when human tutors are involved, testing this speculative hypothesis would certainly be interesting.

Finally, this study suggests that instead of using an overall tutorial tactics for all KCs, inducing KC-based tutorial tactics seems is necessary in that the induced tutorial tactics seems generated different tutorial decisions for different KCs in this study. Additionally, our results demonstrate that RL may be fruitfully applied to derive adaptive pedagogical tutorial tactics from student-computer interactivity data. However, this technique is not yet well understood. It is not completely clear to us, for instance, why our first attempt at inducing policies was suboptimal. In future work, we plan to explore the use of richer POMDP models, and do additional empirical evaluation of the RL approach.

Acknowledgments NSF (#0325054) supported this work and NSF (#SBE-0836012) supported its publication. We thank LRDC for providing all the facilities for this work.

References

Aleven, V., Ogan, A., Popescu, O., Torrey, C., & Koedinger, K. R. (2004). Evaluating the effectiveness of a tutorial dialogue system for self-explanation. In *Intelligent tutoring systems* (Vol. 3220, p. 443-454). Springer.

Anderson, J. R. (1983). *The architecture of cognition*. Cambridge, Mass. : Harvard University Press.

Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher*, *13*, 4-16.

Chi, M., Jordan, P. W., VanLehn, K., & Litman, D. J. (2009). To elicit or to tell: Does it matter? In V. Dimitrova, R. Mizoguchi, B. du Boulay, & A. C. Graesser (Eds.), *Aied* (p. 197-204). IOS Press.

Chi, M. T. H., Roy, M., & Hausmann, R. G. M. (2008). Observing tutorial dialogues collaboratively: Insights about human tutoring effectiveness from vicarious learning. *Cognitive Science*, *32*(2), 301-342.

Chi, M. T. H., Siler, S., Jeong, H., Yamauchi, T., & Hausmann, R. G. (2001). Learning from human tutoring. *Cognitive Science*, *25*, 471-533.

Collins, A., Brown, J. S., & Newman, S. E. (1989). Cognitive apprenticeship: Teaching the craft of reading, writing and mathematics. In L. B. Resnick (Ed.), *Knowing, learning and instruction* (p. 453-494).

Evens, M., & Michael, J. (2006). *One-on-one tutoring by humans and machines.* Mahwah, NJ: Erlbaum.

Forbes-Riley, K., Litman, D. J., Purandare, A., Rotaru, M., & Tetreault, J. R. (2007). Comparing linguistic features for modeling learning in computer tutoring. In R. Luckin, K. R. Koedinger, & J. E. Greer (Eds.), *Aied* (Vol. 158, p. 270-277). IOS Press.

Fossati, D., Eugenio, B. D., Brown, C., & Ohlsson., S. (2008). Learning linked lists: Experiments with the ilist system. In *Intelligent tutoring systems* (Vol. 5091, p. 80-89). Springer.

Katz, S., Allbritton, D., & Connelly, J. (2003). Going beyond the problem given: How human tutors use post-solution discussions to support transfer. *International Journal of AI and Education*, *13*, 79-116.

Koedinger, K. R., & Aleven, V. (2007). Exploring the assistance dilemma in experiments with cognitive tutors. *Educational Psychology Review*, *19*(3), 239-264.

Koedinger, K. R., Anderson, J. R., Hadley, W. H., & Mark, M. A. (1997). Intelligent tutoring goes to school in the big city. *International Journal of AI in Education*, *8*(1), 30-43.

Moore, J. D., Porayska-Pomsta, K., Varges, S., & Zinn, C. (2004). Generating tutorial feedback with affect. In V. Barr & Z. Markov (Eds.), *Flairs conference.* AAAI Press.

Newell, A. (1994). *Unified theories of cognition*. Harvard University Press; Reprint edition.

Reif, F., & Scott, L. A. (1999). Teaching scientific thinking skills: Students and computers coaching each other. *American Journal of Physics*, *67*(9), 819-831.

Rose, C. P., Moore, J. D., VanLehn, K., & Allbritton, D. (2001). A comparative evaluation of socratic versus didactic tutoring. In *Proc. of cognitive sciences society* (p. 869-874).

Singh, S. P., Kearns, M. J., Litman, D. J., & Walker, M. A. (1999). Reinforcement learning for spoken dialogue systems. In S. A. Solla, T. K. Leen, & K.-R. Müller (Eds.), *Nips* (p. 956-962). The MIT Press.

Singh, S. P., Litman, D. J., Kearns, M. J., & Walker, M. A. (2002). Optimizing dialogue management with reinforcement learning: Experiments with the njfun system. *J. Artif. Intell. Res. (JAIR)*, *16*, 105-133.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning*. MIT Press Bradford Books.

Tetreault, J. R., & Litman, D. J. (2008). A reinforcement learning approach to evaluating state representations in spoken dialogue systems. *Speech Communication*, *50*(8-9), 683-696.

VanLehn, K. (2006). The behavior of tutoring systems. *International Journal AI in Education*, *16*(3), 227-265.

VanLehn, K. (submitted). The two-sigma effect revisited: A meta-analysis of human tutoring and several types of computer tutoring.

VanLehn, K., Graesser, A. C., Jackson, G. T., Jordan, P., Olney, A., & Rose, C. P. (2007). When are tutorial dialogues more effective than reading? *Cognitive Science*, *31*(1), 3-62.

VanLehn, K., Jordan, P., & Litman, D. (2007). Developing pedagogically effective tutorial dialogue tactics: Experiments and a testbed. In *Proc. of slate workshop on speech and language technology in education isca tutorial and research workshop.*

Vygotsky, L. (1971). Interaction between learning and development. In T. M. Cole (Ed.), *In mind in society.* (p. 79-91). Harvard University Press: Cambridge Massachusetts.

Comparing Worked Examples and Tutored Problem Solving: Pure vs. Mixed Approaches

Rob Weitz (weitzrob@shu.edu)
Department of Computing and Decision Sciences, Seton Hall University
South Orange, NJ 07079 USA

Ron J. C. M. Salden (rons@cs.cmu.edu)
Human-Computer Interaction Institute, Carnegie Mellon University, 5000 Forbes Avenue
Pittsburgh, PA 15213 USA

Ryung S. Kim (rkim@wpi.edu)
Department of Epidemiology and Population Health, Albert Einstein College of Medicine
1300 Morris Park Ave, Bronx, NY 10461 USA

Neil T. Heffernan (nth@wpi.edu)
Department of Computer Science, Worcester Polytech Institute, 100 Institute Road
Worcester, MA 01609 USA

Abstract

This paper extends our previous work (Kim, Weitz, Heffernan & Krach, 2009) which compared a "classic" worked examples (WE) condition with a tutored problem solving (TPS) condition. By classic we mean the WE condition does not include tutoring, a self-explanation component, or fading. The aim of the current study was to compare the WE and TPS conditions with a mixed condition, which presents students with WE-TPS pairs. More specifically, for conceptual problems a pure WE condition was compared with a WE-TPS condition and for procedural problems a pure TPS condition was compared with a WE-TPS condition. While overall learning occurred in all conditions no significant differences were found between conditions. Further, our findings echo the results of earlier studies, that students who receive worked examples learn more efficiently – that is, they need significantly less time to complete the same learning material. This is an important finding for educators because building classic worked examples is considerably easier than building tutoring.

Keywords: tutored problem solving; worked examples

Introduction

Research on worked examples (e.g., Sweller & Cooper, 1985; Ward & Sweller, 1990) has demonstrated that when students were presented with example-problem pairs rather than problems only, they could attain higher learning outcomes because their working memory capacity was not overloaded. Worked examples reduce problem solving demands by providing worked-out solutions. Therefore, more of the learners' limited processing capacity (i.e., working memory capacity) can be devoted to understanding the domain principles and their application to the problem at hand (Renkl & Atkinson, 2007).

In recent years, a considerable number of studies have explored the conditions under which examples aid in acquiring cognitive skills (for a review, see Atkinson, Derry, Renkl, & Wortham, 2000; Renkl, 2005, 2009). While the impressive body of research on worked examples to date has been quite successful, it also has two important shortcomings. Firstly, the studies are mostly conducted in a laboratory setting without being extended to the more challenging authentic classroom setting and secondly, the studies have almost exclusively compared learning by studying examples to untutored problem solving.

One very successful tutored problem-solving approach is the use of Cognitive Tutors (Koedinger, Anderson, Hadley, & Mark, 1997; Koedinger & Aleven, 2007). These computer-based tutors provide individualized support for learning by doing (i.e., solving problems) by selecting appropriate problems to be solved, by providing feedback and problem-solving hints, and by on-line assessment of the student's learning progress. Because such a tutored environment offers a significant amount of guidance it is a much more challenging control condition than traditional problem solving against which to measure the possible beneficial effects of worked examples. Additionally, research on Cognitive Tutors aims to be examined in the authentic classroom setting (*in vivo* experimentation) which creates a much richer and challenging testing environment compared to a laboratory setting.

Several recent studies have embedded worked examples in a variety of Cognitive Tutors and investigated whether the examples still had beneficial effects over the tougher tutored control condition (e.g., Salden, Aleven, Schwonke, & Renkl, in press; Schwonke et al., 2009). More specifically, these studies proved that replacing some problems with worked examples further enhances student learning by reducing instructional time to the same outcome and/or increasing student outcomes than tutored problem solving.

Of particular interest for the current paper are the studies by McLaren, Lim, and Koedinger (2008) which compared

worked examples with pure TPS (tutored problem-solving) within a Stoichiometry Cognitive Tutor. The results across three studies showed that the students who received worked examples did not learn more than the students who received pure TPS. This reinforces the prior claim that TPS poses a new challenge for the research on worked examples in being a much harder control condition. However, an important consistent finding in the McLaren et al. studies is that the students who received worked examples did learn more efficiently, using 21% less time to complete the same problem set. If these results were to scale across a 20-week course, students could save 4 weeks of time – yet learn just as much.

Another educational system that provides tutored problem solving in classroom settings is the Assistment system (e.g., Razzaq & Heffernan, 2009). Additionally, a further similarity between the Cognitive Tutors and Assistment is their focus on *in vivo* experimentation which allows for an examination of student learning in its most authentic environment. In a previous *in vivo* study Kim, Weitz, Heffernan and Krach (2009) explored the benefits and limitations of worked examples by comparing a "pure worked-example" (pure WE) condition with a pure TPS condition on conceptual and procedural learning. "Pure" means that students in the TPS condition received only TPS remediation while students in the WE condition received solely WE remediation. Note that in contrast to the Cognitive Tutor studies cited above, neither condition included a self-explanation component. The results showed that for conceptual problems students learned more in the pure WE condition and for procedural problems students learned more in the pure TPS condition. In agreement with the findings by McLaren et al. (2008), pure WE was more efficient – that is, it took students less time to do pure WE than TPS.

The current paper addresses a study which extends this research by comparing the best pure condition from the previous study with mixed approaches. That is, for conceptual problems we compare learning resulting in a pure WE condition to one that mixes WE and TPS. For procedural problems we compare a pure TPS approach to a condition that mixes WE and TPS. With these conditions we examine whether the findings of the previous study will still hold. More specifically, if pure WE is better than WE-TPS for conceptual problems and pure TPS is better than WE-TPS for procedural problems it could provide further evidence that examples are always better for conceptual learning and tutored problem solving is always better for procedural learning.

Overall, the outcomes of this study will suggest important guidelines for designing intelligent tutors and provide meaningful insights into the students' learning process. In practical terms, building worked examples is significantly less time consuming than building tutoring; if worked examples are as good as or better than traditional intelligent tutoring – and more efficient – this is valuable information.

The Experiment

Our study involved college students taking an introductory statistics course. Statistics is a good domain for this research as it includes both procedural and conceptual components.

Student Characteristics

Participating students were enrolled in an introductory statistics course at Worcester Polytechnic Institute (WPI), a private university specializing in engineering and the sciences. Eighty-four students, mostly first-year engineering students, participated in the experiment, which was conducted as one of the course's regular lab session.

Design

The tutorial and test problems were typical of problems given in introductory statistics courses. The subject matter concerned one-sample confidence intervals of the mean and was taught on days preceding the experiment. There were no assignments or tests on these topics due before the experiment. At the start of the experiment, students were randomly assigned to one of four groups with equal probability; the resulting student numbers are outlined in Table 1. Note that the mild non-uniformity in numbers is caused by randomness.

Table1: Initial Student Allocation to Groups

Group	Procedural Problem Tutorials	Conceptual Problem Tutorials	No. Students
1	WE-TPS	WE-WE (pure WE)	29
2	WE-TPS	WE-TPS	21
3	TPS-TPS (pure TPS)	WE-WE (pure WE)	17
4	TPS-TPS (pure TPS)	WE-TPS	17

This design allows the comparison of WE-TPS with pure TPS on procedural problems by comparing the performance of students in groups 1 and 2 with that of students in groups 3 and 4. Likewise pure WE may be compared with WE-TPS for conceptual problems by comparing the performance of students in groups 1 and 3 with that of students in groups 2 and 4.

An example of a procedural problem is one that asks the student to calculate a confidence interval. A conceptual problem might ask about the impact on the width of a confidence interval if the sample size is doubled. Procedural problems align with the NSF-Funded ARTIST project guidelines (https://app.gen.umn.edu/artist/glossary.html) for "statistical literacy," and conceptual problems with "statistical reasoning" and "statistical thinking" (delMas, 2002).

Of the eighty-four students that participated we excluded ten students who spent less than 5 minutes in the post-test

from our analysis due to time and motivation issues. Further, we eliminated eleven students from the conceptual part of the analysis as they did not complete the conceptual problems in the tutorial. Note that the conceptual problems were towards the end of the tutorial. The final number of students used in each condition of the analysis is provided in Table 2. In both instances where we eliminated students from the analysis, roughly the same numbers were removed from each group.

Table 2: No. Students in Each Group

Group	Procedural Problem Tutorials	No. Students	Conceptual Problem Tutorials	No. Students
1	WE-TPS	25	WE-WE (pure WE)	21
2	WE-TPS	19	WE-TPS	18
3	TPS-TPS (pure TPS)	13	WE-WE (pure WE)	10
4	TPS-TPS (pure TPS)	15	WE-TPS	12
Total		72		61

The ASSISTment System

Our experiment was conducted via the ASSISTment intelligent tutoring system (http://assistment.org). It is similar to the CTAT system (Koedinger Aleven, Heffernan, McLaren, & Hockenberry, 2004), used in some of the previously mentioned studies (McLaren, et al., 2008), in that the system provides the student with tutoring on the individual steps of a problem, generally breaking a problem down into 3-4 steps. For each step, a student is asked to provide an answer, and receives feedback on their answer until they get it correct. Our system differs from the CTAT structure in several ways including that there is only one solution path and the intermediate solution goals are highlighted. A further difference is that our system does not contain a self-explanation component.

The tutorials were comprised of three pairs of problems.[1] Each pair was comprised of two isomorphic problems. The first two pairs were procedural problems and the last problem pair was conceptual in nature.

TPS-TPS (Pure TPS) Condition

For this study the ASSISTment system was modified to force students to work through the TPS for the first problem of each pair. This "forced TPS" approach ensures that each student experiences tutoring. After completion of the first problem of the pair, the student is presented with an isomorphic problem and is asked by the system to provide the answer. If the student gets this second problem correct, the student is done with the problem. If the student gets the answer incorrect or indicates that s/he needs help solving the problem, the system provides TPS support.

In terms of tutoring, the system gives immediate corrective feedback for each attempt at solving a problem. The student can choose to answer the problem or ask the system to break it into steps. However, if the student answers incorrectly the system automatically breaks the problem into steps. For each step, the student will receive immediate feedback and has the possibility to request hints.

WE-WE (Pure Worked Example) Condition

Firstly, it should be noted that "classic" worked examples are used which do not contain tutoring, a self-explanation component, or fading.

The student is presented with the same first problem as in the TPS condition, and a worked solution including the necessary steps to take in that problem. After studying the worked example, the student is then presented with an isomorphic problem, the exact same second problem as in the TPS condition, which the student is expected to solve. The student has access to the first WE while trying to solve the second. If the student gets this second problem correct, the student is done with the problem. If the student gets the answer incorrect or indicates that s/he needs help solving the problem, the system provides the worked solution for the problem for review by the student.

WE-TPS (Mixed) Condition

The student is presented with the first problem and a worked solution to that problem, similar to the WE-WE condition. After studying the worked example, the student is then presented with the second problem. If the student gets this isomorphic problem correct, the student is done with the problem. If the student gets the answer incorrect or indicates that s/he needs help solving the problem, the system provides TPS support. See Table 3 for an overview of the problem pairs for each experimental condition.

Table 3: A Comparison of Intelligent Tutoring and Worked Examples

	Pure TPS (TPS-TPS)	Pure WE (WE-WE)	Mixed (WE-TPS)
First Problem	Student studies with forced TPS	Student studies WE	Student studies WE
Second Problem	Student is given opportunity to solve the problem. If student answer is incorrect, the problem is marked incorrect and,		
	TPS is provided	WE is provided	TPS is provided

The students were allowed to work though both tutorials at their own pace. One week before the experiment students were given a ten minute tutorial on how to use the ASSISTment software for which they were allowed to work through at their own pace. They created an account for

[1] All of our materials are available at http://teacherwiki.assistment.org/wiki/index.php/CogSci2010 so other researchers can inspect them.

themselves, and enrolled in their professor's class. They got a few minutes of practice with the system during which they did one worked example and one tutored problem solving.

The experiment consisted of three parts: pre-test, tutorial, and post-test. The pre-test and post-test were identical, and were comprised of four procedural problems and three conceptual problems.

The students were given 20 minutes to go through the pre-test without any feedback, 40 minutes for their tutorials, and 20 minutes for the post test (see Table 4). In order to control time, students were not supposed to be allowed to move to the next part of the experiment until a designated time passed. However, in practice we actually had some students not following the directions when asked to move to the next part of the experiment.

Table 4: Outline of Experiment

One Sample Confidence Interval for the Mean

Several Days Prior to Lab Session

- Lecture on the topic

During Lab Session

1. Pre-Test (20 min; students' initial knowledge)
   - 20 minutes
   - Four procedural and three conceptual.
2. Condition (Tutorials)
   - 40 minutes
   - 3 pairs of Problems: 2 procedural, one conceptual (3 parts)
3. Post-Test (20 min; students' knowledge after trial)
   - Same problems as Pre-Test

Results

Learning by Problem

Table 5 provides the percentage of students across all conditions getting each problem correct on the pre- and post-tests. Student learning is clearly evident for all items ($z = 3.78$, $p < .001$, $d = 1.36$).

Following the approach in item response theory (Embretson & Reise, 2002), throughout the remainder of this section, we summarize student performance on a problem or on a category of problems by the adjusted percent correct, that is, the percent correct adjusted by problem difficulty. We then define learning for problems as the difference in adjusted percent correct between post-test problems and the corresponding pre-test problems.

Qualitatively speaking, this means that students who correctly answer harder items will get more credit than students who correctly answer easier items.

We determined these adjusted values using a generalized linear mixed effects model, also referred to as a generalized linear multilevel model (Bates & Sarkar, 2007; Rabe-Hesketh, Skrondal, & Pickles, 2005).

Table 5: Learning by Problem

Problem	**Percent Students Correct**	
Procedural	**Pre-Test**	**Post-Test**
1	5.6%	58.3%
2	16.7%	73.6%
3	15.3%	43.1%
4	30.6%	54.2%
Conceptual		
1	11.5%	27.9%
2	73.8%	91.8%
3	37.7%	52.5%

Learning by Condition

Table 6 below summarizes the learning results by type of tutorial. So, for example, for procedural problems, students in the WE-TPS improved their performance by 40.1% (54.9% - 14.8%).

Table 6: The Adjusted Percent Correct

		Percent Correct
Procedural Problems	Pre-Test	14.8%
	WE-TPS	54.9%
	TPS-TPS	63.3%
Conceptual problems	Pre-Test	37.8%
	WE-TPS	61.2%
	WE-WE	61.7%

For procedural problems, students in the pure TPS condition outperformed students in the WE-TPS condition. However, this difference (63.3% vs. 54.9%) is not significant ($p = 0.23$). Likewise, for conceptual problems, the results indicate a small benefit for the pure WE condition over the WE-TPS condition (61.7% vs. 61.2%); these results are clearly not statistically significant ($p = 0.95$).

Learning Time

As noted earlier, previous research has consistently indicated that doing worked examples requires significantly less time for students than tutored problem solving.

Table 7: Times for Students to do the Tutorial Problems

		n	Mean	SD
Procedural Problems	WE-TPS	44	18.03	7.79
	TPS-TPS	28	26.00	10.63
Conceptual Problems	WE-TPS	30	6.70	2.53
	WE-WE	31	6.60	3.17

Table 7 provides the mean and standard deviation of student times in each group for both types of problems in the tutorial. Focusing on the procedural problems, we see the same pattern here with students in the WE-TPS condition taking less time than those in the pure TPS

condition. These results are statistically significant ($t = 3.42$, $p < .01$, $d = 0.86$).

As the conceptual problems were placed after the procedural problems in the tutorial (condition), the above-reported conceptual times may have been artificially constrained. We observed that procedural times and conceptual times are negatively correlated – an indication that individuals who spent a lot of time on the procedural problems ran out of time on the conceptual problems. Note (again) that we excluded students who did not finish the conceptual part of the tutorial from our post-test results.

Discussion

This paper extends our previous work (Kim et al., 2009) comparing pure WE with pure TPS approaches where the results showed that pure WE was more effective for conceptual problems, while pure TPS was more effective for procedural problems. Furthermore, pure WE was more efficient in that students took less time to work through the WE condition than the TPS condition. The aim of the current study was to compare these pure WE and TPS conditions with a mixed condition, which presents students with WE-TPS pairs. More specifically, for conceptual problems a pure WE condition was compared with a WE-TPS condition and for procedural problems a pure TPS condition was compared with a WE-TPS condition.

While overall learning occurred in all conditions and the pure methods come out ahead in terms of student learning, the results are not statistically significant. More specifically, there were small non-significant differences favoring the pure WE condition for conceptual problems and the pure TPS condition for procedural problems. Furthermore, the efficiency effect of the previous study was replicated meaning that students needed less time to complete the WE tutorial than the TPS tutorial. These results are similar to the findings of McLaren et al. (2008) who also did not find significant differences in student learning but who also found that students who received worked examples did learn more efficiently, using 21% less time to complete the same problem set.

It should be noted that McLaren et al. (2008) and other studies use worked examples in combination with tutoring, a self-explanation component, and/or fading. In contrast to those studies, the worked examples used in our experiments are "classic" worked examples which do not include these extra elements. While these elements can undoubtedly improve learning our studies shows that the use of classic worked examples in tutored problem solving can still result in similar outcomes without any detrimental effect on student learning. As such, the replication of the time efficiency effect makes a strong case for the use of classic worked examples in tutored problem solving.

A possible explanation for the lack of significant main differences could be offered by Rittle-Johnson, Siegler, and Alibali (2001) who stated that effects of worked examples on procedural tasks might be more indirect and need more time to materialize. In fact, other studies (e.g., Anthony, 2008; Salden, et al., 2009) that compared TPS and WE also did not find significant differences on the post-test but they did find positive effects favoring the WE conditions on a delayed post-test.

A further explanation might be found in the time limit that we imposed on the students. We had to exclude eleven students from our data analysis because they did not have enough time to complete the conceptual problems in the tutorial. Had we given them more time then we might have been able to observe possible conceptual learning differences.

For future studies we would like to explore other factors which could deepen the insights on the beneficial effects of worked examples in TPS. One possible factor is students' prior knowledge which can have a mediating influence on their learning progress if students with differing prior knowledge levels work through the same training material. In line with the expertise reversal effect (Kalyuga, 2007), students who have a high knowledge level could even experience detrimental effects of worked examples. In future studies we could use the pre-test scores to check if such differences in prior knowledge exist and use this information to determine what experimental condition a student ought to be in.

Furthermore, in accordance with Schwonke et al. (2009) we could try to add thinking aloud to differentiate learning effects. In their first study Schwonke et al. also did not find student learning differences but they used thinking aloud protocols in their second study which subsequently showed a higher learning gain in terms of conceptual knowledge for the example-enriched TPS condition. It is plausible that students who were thinking aloud about the worked examples engaged in deeper processing of conceptual knowledge than the students in the control TPS condition without examples. Consequently, being able to talk aloud about the worked examples might have led to the observed higher learning gain.

Finally, adding a delayed post-test to our future studies might also enable us to differentiate differences between TPS and WE-TPS conditions. Rittle-Johnson et al.'s statement that the effects on procedural tasks might need time to materialize has been proven to be accurate in other studies compared tutored problem solving and worked examples (e.g., Anthony, 2008; Salden, et al., 2009). More specifically, if worked examples support students in engaging with the conceptual knowledge more deeply but only over longer period of time then this has significant implications for developing computer-based learning programs which use worked examples.

In conclusion, our results extend the previous findings of TPS and WE-TPS comparisons. The tutored problem solving environment poses a more challenging control condition than traditional problem solving conditions. Yet across two studies and in line with the McLaren et al. (2008) studies we consistently found that students needed less time to complete the training phase when being presented with worked examples without any loss of student learning on

the post-test. These results are even more impressive as our experiments used classic worked examples, which do not offer tutoring or a self-explanation component, as those used by McLaren et al. (2008).

This is an important finding for educators because building classic worked examples is considerably easier than building tutoring and in fact is easier than building worked examples with more features. Future studies are needed to further investigate under what circumstances classic worked examples can make computer-based instructional materials more efficient.

Acknowledgments

This works was funded in part by the National Science Foundation CAREER Award the US Department of Education (R305K030140) but the opinions are solely those of the authors. Furthermore, the authors would like to express their gratitude to Professor Joseph D. Petruccelli for his invaluable help in running the study.

References

Anthony (2008). *Developing handwriting-based Intelligent Tutors to enhance mathematics learning*. Unpublished doctoral dissertation, Carnegie Mellon University, USA.

Atkinson, R. K., Derry, S. J., Renkl, A., & Wortham, D. W. (2000). Learning from examples: Instructional principles from the worked examples research. *Review of Educational Research, 70*, 181-214.

Bates, D. & Sarkar, D. (2007). *lme4: Linear mixed-effects models using S4 classes.* R package version 0.9975-12, http://CRAN.R-project.org/.

delMas, R. C. (2002). Statistical literacy, reasoning, and learning: A commentary. *Journal of Statistics Education*, 10 (3). Retrieved from www.amstat.org/publications/jse/v10n3/delmas_discussion.html

Embretson, S. E. & Reise, S. (2002). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Kalyuga, S., (2007). Expertise reversal effect and its implications for learner-entailed instruction. *Educational Psychology Review, 19*, 509-539.

Kim, R, Weitz, R., Heffernan, N. & Krach, N. (2009). Tutored problem solving vs. "pure" worked examples. In N. A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Koedinger, K. R., & Aleven, V. (2007). Exploring the assistance dilemma in experiments with Cognitive Tutors. *Educational Psychology Review, 19*, 239-264.

Koedinger, K. R., Aleven, V., Heffernan, N. T., McLaren, B. M., & Hockenberry, M. (2004). Opening the door to non-programmers: Authoring intelligent tutor behavior by demonstration. In Lester et al (Eds.) *Proceedings of 7th Annual Intelligent Tutoring Systems Conference*, Springer, 162-173.

Koedinger, K. R., Anderson, J. R., Hadley, W. H., & Mark, M. A. (1997). Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education, 8*, 30-43.

McLaren, B. M., Lim, S., & Koedinger, K. R. (2008). When and how often should worked examples be given to students? New results and a summary of the current state of research. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 2176-2181). Austin, TX: Cognitive Science Society.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2005). Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects. *Journal of Econometrics, 128*, 301-323.

Razzaq, L., & Heffernan, N.T. (2009). To tutor or not to tutor: That is the question. In Dimitrova, Mizoguchi, du Boulay and Graesser (Eds.), *Proceedings of the Conference on Artificial Intelligence in Education*. pp. 457-464.

Renkl, A. (2005). The worked-out-example principle in multimedia learning. In R. Mayer (Ed.), *Cambridge Handbook of Multimedia Learning*. Cambridge, UK: Cambridge University Press.

Renkl, A. (2009). *Towards an instructionally oriented theory of example-based learning*. Manuscript submitted for publication.

Renkl, A., & Atkinson, R. K. (2007). An example order for cognitive skill acquisition. In F. E. Ritter, J. Nerb, E. Lehtinen, T. O'Shea (Eds.), *In order to learn: How the sequence of topics influences learning* (pp. 95-105). New York , NY : Oxford University Press.

Rittle-Johnson, B., Siegler, R. S., & Alibali, M. W. (2001). Developing conceptual understanding and procedural skill in mathematics: An iterative process. *Journal of Educational Psychology, 93*, 346-362.

Salden, R. J. C. M., Aleven, V., Renkl, A., & Schwonke, R. (2009). Worked examples and tutored problem solving: Redundant or synergistic forms of support? *Topics in Cognitive Science, 1*, 203-213.

Schwonke, R., Renkl, A., Krieg, C., Wittwer, J., Aleven, V., Salden, R. J. C. M. (2009). The worked-example effect: Not an artefact of lousy control conditions. *Computers in Human Behavior, 25*, 258-266.

Sweller, J., & Cooper, G. A. (1985). The use of worked examples as a substitute for problem solving in learning algebra. *Cognition and Instruction, 2*, 59-89.

Ward, M., & Sweller, J. (1990). Structuring effective worked examples. *Cognition and Instruction, 7*, 1-39.

Interleaving Worked Examples and Cognitive Tutor Support for Algebraic Modeling of Problem Situations.

Albert Corbett (corbett@cmu.edu)
Human-Computer Interaction Institute, Carnegie Mellon University
Pittsburgh, PA 15213 USA

Stephen K. Reed (sreed@sunstroke.sdsu.edu)
Department of Psychology, San Diego State University
San Diego, CA 92182 USA

Bob Hoffmann (bob.hoffman@sdsu.edu)
Department of Educational Technology, San Diego State University
San Diego, CA 92182 USA

Ben MacLaren (maclaren@andrew.cmu.edu)
Human-Computer Interaction Institute, Carnegie Mellon University
Pittsburgh, PA 15213 USA

Angela Wagner (awagner@cmu.edu)
Human-Computer Interaction Institute, Carnegie Mellon University
Pittsburgh, PA 15213 USA

Abstract

Integrating worked examples with problem solving yields more effective and efficient learning, as does intelligent tutoring support for problem solving. This study examines the impact of integrating worked examples and intelligent tutor support for algebra modeling problems. Students in three conditions alternately studied worked examples (either static graphics, interactive graphics or static tables) and solved Algebra Cognitive Tutor problems. A control group solved all the problems with the Cognitive Tutor. Students in the four groups developed equivalent problem-solving skills, but students learned more efficiently in the interleaved worked example conditions, requiring 26% less time to complete the problem set. There were no differences among the four groups in two measures of robust learning – a retention test and a transfer test. But students in the static table condition could more accurately describe what algebraic model components represent in problem situations than could students in the other three conditions.

Keywords: Education; Problem solving; Learning; Classroom Study; Intelligent Tutors; Worked Examples.

Introduction

Extensive research has documented the beneficial impact on learning of interleaving worked examples with problem solving (Kalyuga, et al 2001; Pashler, et al, 2007; Sweller & Cooper, 1985; von Gog, Paas, & Van Merrienboer, 2004). Novices learn more quickly and deeply from a sequence of problems if they are asked to alternate between explaining worked-out examples of problem solutions and solving problems than if they are asked to solve all the problems in the sequence.

Typically in this research problem solving is supported by whole-answer feedback. After students complete a problem solution, whether successfully or not, they are given an example of a correct solution. This comparison condition is relatively weak, since step-by-step assistance in problem solving has been shown to be both more effective (improved learning outcomes) and more efficient (less learning time to achieve the same learning outcome) than whole answer feedback. For instance, Corbett & Anderson (2001) compared step-by-step feedback and whole-answer feedback in the Lisp Programming Cognitive Tutor and found that students in the former condition finished a fixed set of problems in one-third the time required by those in the latter condition, and made 40% fewer errors on posttests.

As a result, the question arises whether interleaving worked examples with problem solving scaffolded by intelligent tutoring systems might also yield improved learning outcomes and/or improved learning efficiency. McLaren, Lim and Koedinger (2008) examined this question in an intelligent tutor for chemistry problem solving and found that interleaving worked examples with problem solving yielded the same learning outcome as the baseline problem-solving condition, but in less time, thereby increasing learning efficiency.

Several studies have examined the impact of incorporating "faded" worked examples into Geometry Cognitive Tutor (GCT) modules in which students solve geometry problems and justify each step with a problem-solving principle (Aleven & Koedinger, 2002). In example fading (Renkl & Atkinson, 2003) the first problem is presented as a complete worked example, and in successive

problems students complete progressively more steps themselves until students are finally solving complete problems. When faded worked examples were incorporated into GCT, learning was more efficient (students spent less time to reach the same level of skill) and some evidence was obtained that the worked-example condition yielded deeper understanding (Salden, et al, 2008; Schwonke, et al, 2009).

The present study examines the impact of interleaved worked examples in a Cognitive Tutor (CT) module for Algebra problem solving. The study has two purposes. First, the study examines the impact of interleaving worked examples on students' learning time, their problem-solving skill and their depth of understanding. Second, the study evaluates three alternative types of worked examples: (1) Static Graphics in which problem components are represented graphically; (2) Interactive Graphics in which students participate in constructing the graphical problem representation; and (3) Static Tables in which problem components are represented symbolically in a table, analogous to the problem-solving interface.

This study compares four learning conditions; three conditions in which each type of worked example is interleaved with Cognitive Tutor problem solving and a fourth, Cognitive Tutor problem-solving baseline condition.

The following sections describe the problem solving domain, the Cognitive Tutor problem-solving environment and the three types of worked examples.

The Domain: Algebraic Modeling

In this study students are asked to solve "mixture problems," for example:

> You have an American Express credit card with a balance of $715 at an 11% interest rate and a Visa credit card with a 15% interest rate. If you pay a total of $165 in annual interest, what is the balance on your Visa card?

The problem-solving goal is to construct a symbolic model of the situation that can be used to solve the problem, e.g.:

$$(.11 \times \$715) + (.15 \times V) = \$165$$

The problem-solving curriculum consists of four problem types: Two types of "arithmetic problems," in which the unknown value is naturally represented as an isolated variable on one side of the equation, and two types of "algebra problems" in which the unknown quantity is more naturally represented as a variable that is embedded in one or in two expressions in the equation. See Figure 1 for an example of each type.

Cognitive Tutor Problem Solving

Figure 2 displays the interface for the Cognitive Tutor at the end of a problem. Each problem describes a mixture scenario and provides a table to scaffold the relationship between the scenario components and the mathematical representations of the components. Students enter a number, variable or operation into each cell. After completing the table, the student enters an equation to model the situation in the text cell at the bottom of the screen. The activities were created with the Cognitive Tutors Authoring Tools (CTAT) environment (Aleven, et al, in 2009). As in all cognitive tutors, students received accuracy feedback on each step, could request advice on any step, and were required to complete a correct solution to each problem.

[Arithmetic Type 1] You have a MasterCard with a balance of $532 at a 21% interest rate. You also have a Visa credit card with a balance of $841 at a 16% interest rate. How much money are you paying in total interest?

$$(.21 \times \$532) + (.16 \times \$841) = T$$

[Arithmetic Type 2] Shelly owed $475 in total interest on her MasterCard and Visa accounts. Her MasterCard charges 19% interest and her Visa Card charges 22% interest. She paid the interest on her Visa Card debt of $1100. How much interest does she still owe on her MasterCard?

$$\$475 - (.22 \times \$1100) = M$$

[Algebra Type 1] You have an American Express credit card with a balance of $715 at an 11% interest rate and a Visa credit card with a 15% interest rate. If you pay a total of $165 in annual interest, what is the balance on your Visa card?

$$(.11 \times \$715) + (.15 \times V) = \$165$$

[Algebra Type 2] You have a total balance of $1405 on two different credit cards— an American Express credit card with a 12% interest rate and a Discover credit card with a 24% interest rate. If you owe a total of $224 in annual interest, what is your balance on the Discover card?

$$(.24 \times D) + (.12 \times [\$1405 - D]) = \$224$$

Figure 1: An example problem situation and symbolic model for each of the four problem types.

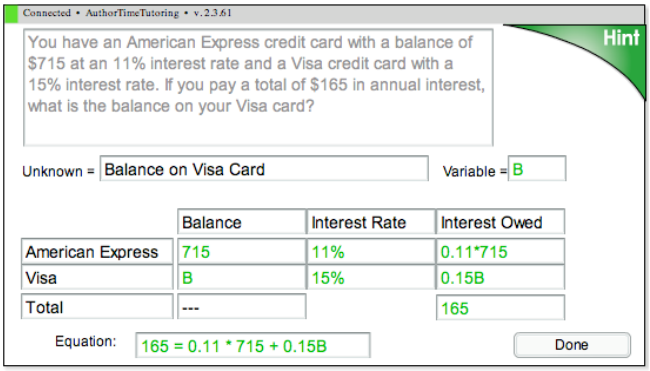


Figure 2: The Cognitive Tutor interface at the completion of a problem.

Worked Examples

Three types of worked examples were developed, in the Animation Tutor environment (Reed, 2005), each consisting

of multiple successive screens. In each case the first screen presented a problem statement alone. Successive screens developed an analysis of the problem's component structure in graphical or tabular form.

(1) Static Graphics (SG). Figure 3 shows the final screen of a static graphics worked example. The first screen displayed just the problem statement at the top. Students successively press the Continue arrow to see (1) the first stack of money which represents an account balance and interest owed, (2) the second stack of money which represents the second account balance and interest owed, and (3) both the third stack, which represents the total interest, and the symbolic model at the bottom of the screen.

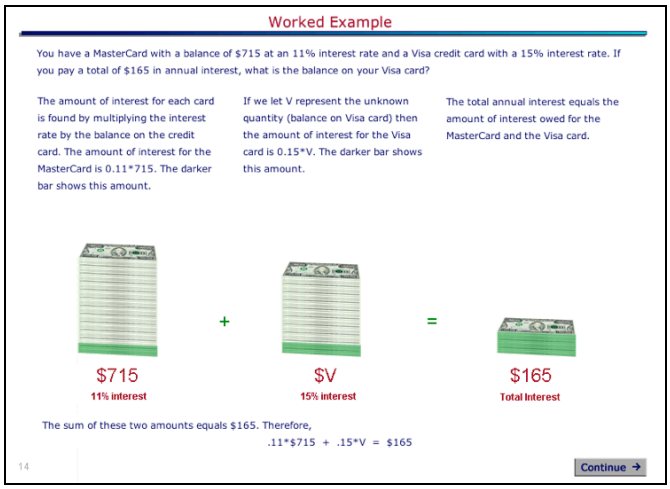


Figure 3: A static graphics worked example at the completion of the example.

(2) Interactive Graphics (IG). Interactive graphics worked examples are the same as the SG worked examples, except that students construct the total interest stack. Students click on the interest component at the bottom of each of the other two stacks and drag that component over to the total interest stack to add up the total interest. Interactive worked examples were developed for all the algebra problems and introduced with a single arithmetic problem. Students in the IG condition viewed static graphic examples for the other arithmetic problems.

(3) Static Table (ST). Figure 4 displays the final screen of a static table worked example. As with the graphics examples, the first screen displays the problem statement alone. Students successively click the Continue arrow to see (1) the column labels and first row of the table, which represents an account balance and interest owed, (2) the second row of the table which represents the second account balance and interest owed, and (3) the symbolic model of the situation beneath the table.

Design Principles. The three types of worked examples all follow two principles of multimedia design (Sweller, 2003; Mayer 2001; Moreno & Mayer, 2007). The first is the proximity principle that different media be closely integrated in space. Verbal explanations are therefore placed immediately above, and the equation immediately below, either the bars or the table in the worked examples. The second principle, minimize cognitive load, is achieved by presenting the solution in successive segments.

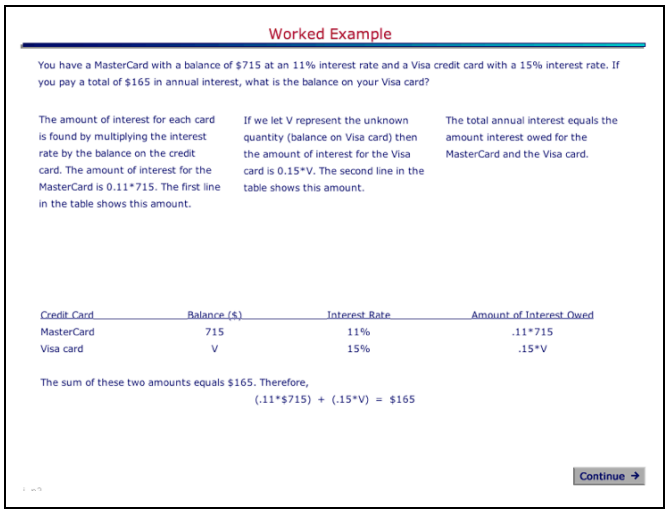


Figure 4: A static table worked example at the completion of the example.

Predictions

Time and Learning efficiency. Time-on-task in learning is expected to be less in the worked example conditions than in the problem-solving condition. Students typically study worked examples in less time than they can generate problem solutions, even with intelligent tutoring support (McLaren, et al, 2008; Salden, et al, 2008; Schwonke, et al, 2009). However, interleaved worked examples are only more efficient if students in those conditions acquire as good, or better, problem-solving skills as students in the problem solving condition.

Robust Learning. There are several reasons to expect that students may acquire a deeper understanding of problem solving in the interleaved worked example conditions. Cognitive Load theory (Sweller, 2003) suggests that worked examples can eliminate the cognitive load associated with generating problem solutions, and free up capacity that students can devote to understanding the solutions. In this study, all the worked-example conditions describe the mapping between the mathematical representations and the problem situations, so students may acquire a better understanding of the underlying semantics, an understanding that should support better retention and transfer to novel problem situations. In addition, the two graphics conditions may promote better retention than the other two conditions, since they encourage visual thinking (Reed, 2010), thereby creating multiple memory codes, both graphical and symbolic (Mayer, 2001; Paivio, 1986). Finally, interactive graphics may foster still better retention than static graphics, since interactively constructing key quantities in the graphics representation, (Moreno & Meyer, 2007), creates a third, motor code (Engelkamp, 1998; Glenberg, et al, 2004; Reed, 2006, 2008).

Robust Learning Measures

A problem-solving pretest and posttest were employed to measure gains in students' algebra problem-solving skills. In addition, three "robust learning" tests were employed to measures students' depth of understanding.

(1) Retention. A retention test examined students' arithmetic and algebra problem solving skills after a one-week interval.

(2) Transfer. A transfer test described "mixture" situations with novel quantitative structures and asked students to generate mathematical models of the situations, which also had novel structures.

(3) Model Description. The Cognitive Tutor Model Analysis Tool (Corbett, et al, 2000, 2007; Corbett, Wagner & Raspat, 2003) was employed to ask students to explain the structure of arithmetic and algebraic models. As displayed in Figure 5, each problem presents a problem description and a mathematical model of the situation. Students select entries from menus to describe what each hierarchical component of the symbolic model represents in the problem situation. As in all Cognitive Tutors, students receive feedback on each problem step, can request advice on each step, and are required to complete a correct solution to the problem.

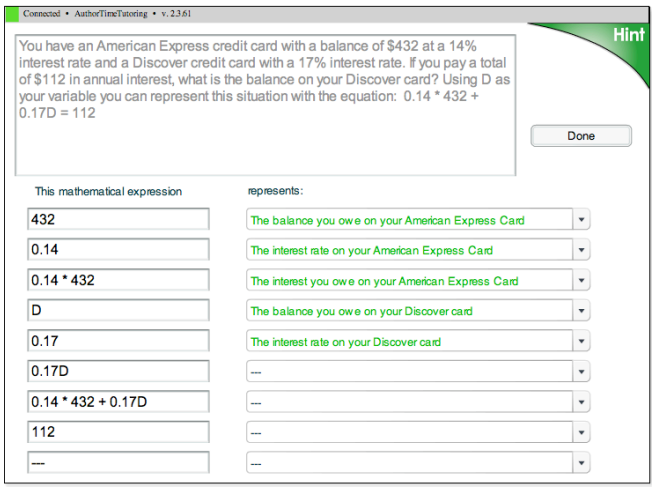


Figure 5: The Model Analysis tool partway through a problem.

Method

Participants

128 students enrolled in Cognitive Tutor Algebra courses in three Pittsburgh-area high schools participated in the study.

Design

The study was completed over the course of three computer sessions in the students' Algebra Cognitive Tutor courses. In the first two sessions, students completed 16 mixture problems, eight problems per day. The students in each of the three courses were randomly assigned to one of four learning conditions. Students in the three worked example conditions studied example solutions for the odd numbered problems and solved the even numbered problems with the Cognitive Tutor each day. Students in the fourth condition solved all the problems each day with the Cognitive Tutor.

Learning Materials

Four types of mixture problems were developed, two "arithmetic" types and two "algebraic" types, as displayed in Figure 1. Four problems of each type were developed, for a total of 16 problems. Two problems of each type involved interest payments on two credit cards, as displayed in the figures. The other two were mining problems, about extracting metals from two ores of different quality. The four problems of each kind were presented in succession, with the two equivalent interest problems first, followed by the two equivalent ore problems.

Test Materials

Four test measures of student learning were developed.

Day-2 Problem-Solving Test. Paper-and-pencil tests were developed consisting of two problems, equivalent to the two types of algebra problems students solved with the online tutor that day. Each problem presented a mixture problem situation and students were asked to generate an equation to model the situation. Two test forms were developed and within each condition, each form served as the pretest for half the students, who then switched to the other form for the posttest, so that the pretests and posttests were matched across the full set of students, but for each student the pretest and posttest were different.

Day-3 Retention Test. This test consisted of four problems, equivalent to the four types of problems students had solved with the online tutor. Again, each problem presented a mixture problem situation and students were asked to generate an equation to model the situation.

Day-3 Transfer Test. The Day-3 transfer test consisted of an arithmetic problem and an algebra problem in which students were asked to generate symbolic models of situations with novel structures.

Day-3 Model Component Descriptions. Four Model Analysis problems were developed. Each problem corresponded to one of the four problem types students had solved on the prior two days of the study. Each problem presented a mixture scenario and presented a symbolic model of the scenario. Students were asked to describe what each hierarchical component of the equation represents in the real-world situation, by selecting entries from menus.

Procedure

In the first session, the online problem solving and worked example activities were introduced, then students worked through the eight arithmetic mixture problems. In the second session, students completed a two-problem paper pretest, worked through eight algebraic problems, then completed a two-problem paper posttest. In the third session, which followed a week later, students completed the four-problem paper retention test, followed by the two-problem paper transfer test and finally the four Model Analysis problems.

Results and Discussion

Four students were excluded from the analyses because they missed the second session and seven others were excluded for talking to others as they worked on the problems.

Day-2 Pretest-Posttest Learning Gains

As displayed in Table 1, there were substantial pretest-posttest learning gains in all four learning conditions, averaging 26 percentage points. In an analysis of variance, this main effect of test type was significant $F(1,105) = 52.14$, $p < .001$. There was no significant difference of learning condition $F(3,105) < 1$, and no significant interaction of test type and learning condition $F(3,105) < 1$.

Table 1: Learning Time per problem for Day 1 and Day 2 (minutes) and Day-2 pretest and posttest accuracy (percent correct).

Learning Conditions	Day 1 Time	Day 2 Time	Pretest %correct	Posttest % correct
CT	2.30	2.15	7	37
IG	1.52	1.68	7	28
SG	1.68	1.52	4	34
ST	1.75	1.72	8	28
Mean	1.81	1.77	6	32

Learning Efficiency

Table 1 displays average learning time per problem for the first two sessions. Elapsed time was not measured for the first worked example in each session (since the environment did not directly record time), so the first pair of equivalent problems in each session is excluded from this analysis for all four groups. In addition, 13 students were excluded from the Day-1 analysis and 16 students from the Day-2 analysis because of missing data. While there were no differences in skill acquisition outcomes among the four conditions, students in the three interleaved worked example conditions spent less time in learning, and so learned more efficiently.

Students in the three worked example conditions averaged 28% less time per problem on Day 1 than students in the problem solving condition (1.65 vs 2.30) and 24% less time per problem on Day 2 (1.64 vs. 2.15). The main effect of condition is significant for Day 1, $F(3,100) = 6.88$, $p < .001$ and for Day 2, $F(3,97) = 6.33$, $p < .001$. Bonferroni comparisons revealed that the CT group differed from each one of the three worked example groups both on Day 1 and on Day 2, $p < .02$ in each case. The three worked example groups did not differ from each other.

These average times mask a highly significant Group x Problem interaction on Day 1, $F(3,100) = 93.12$, $p < .001$, and on Day 2, $F(3,97) = 90.19$, $p < .001$. On Day 1 the three worked example (WE) groups averaged 0.78 min. on the worked examples, while the CT group averaged 2.98 min. solving the corresponding problems. The WE groups averaged 2.53 min. on solving the subsequent equivalent problems, while the CT group averaged 1.63 min. on those problems. On Day 2, the WE groups averaged 0.62 min. on the worked examples and the CT group averaged 2.82 min. solving those problems. The WE group averaged 2.67 min. solving the subsequent problems and the CT group averaged 1.50 min. on those problems.

Robust Learning

Of the 117 students included in the study, 102 completed the day 3 robust learning activities. Table 2 displays results of the three robust learning measures included in the study: (1) retention of problem-solving skill; (2) transfer of problem-solving skill; and (3) explanations of symbolic model components.

Retention Test. Table 2 displays students' test accuracy on the one-week retention test of problem-solving skill. Retention test accuracy did not vary significantly across the four learning conditions, $F(3,90) < 1$.

Transfer Test. As can be seen in Table 2, students in the four learning conditions averaged 17% correct on the transfer test of problem-solving skill. The main effect of learning condition was not significant $F(3,90) < 1$.

Model Component Descriptions. The model analysis task required students to describe what a total of 31 hierarchical equation components represented in the four real-world problem situations. Table 2 displays the average percentage of these 31 descriptions on which students' first menu selection was correct. There was no significant difference among the groups in an ANOVA, $F(3,97) < 1$. But the ST group performed consistently best in describing the model components, achieving the highest accuracy for 18 of the 31 components (vs. 5 for the IG and SG groups and 3 for the CT group). This difference is significant in a Friedman two-way ANOVA of rank ordering, $\chi^2(3) = 20.00$, $p < .001$.

Table 2: Day-3 Robust learning measures: Retention, transfer and model analysis accuracy (percent correct).

Learning Conditions	Retention %correct	Transfer % correct	Model Analysis % correct
CT	32	15	52
IG	29	18	52
SG	29	21	53
ST	26	13	58
Mean	29	17	54

Conclusion

The main results confirm earlier conclusions in chemistry and geometry that incorporating worked examples into intelligent tutor-supported problem solving can improve learning efficiency. While students developed similar problem-solving skills across the four conditions, students spent 26% less time completing the sixteen problems in the three interleaved worked-example conditions than in the problem-solving comparison condition.

However, there is relatively thin evidence that incorporating worked examples yielded a deeper understanding of problems solving, as expected by Cognitive Load theory. Students in the static table worked example condition demonstrated a better understanding of the referential semantics that link the mathematical representations and real-world problem situations than students in the problem solving condition. However, this deeper knowledge did not support greater problem solving accuracy, retention or transfer. Students in the two graphics worked example conditions also did not show more robust learning than students in the problem solving condition.

Acknowledgments

This project was funded by the Pittsburgh Science of Learning Center, NSF awards SBE-0354420 and SBE-0836012.

References

Aleven, V. & Koedinger, K.R. (2002). An effective metacognitive strategy: Learning by doing and explaining with a computer-based Cognitive Tutor. *Cognitive Science, 26*, 147-179.

Aleven, V., McLaren, B.M., Sewall, J., & Koedinger, K.R. (2009). A new paradigm for intelligent tutoring systems: Example-tracing tutors. *International Journal of Artificial Intelligence in Education, 19, 105-154.*

Corbett, A.T. & Anderson, J.R. (2001). Locus of feedback control in computer-based tutoring: Impact on learning rate, achievement and attitudes. In J. Jacko, A. Sears, M. Beaudouin-Lafon & R. Jacob (Eds.), *Proceedings of ACM CHI'2001 Conference on Human Factors in Computing Systems* (pp, 245-252).

Corbett, A., McLaughlin, M., Scarpinatto, K. C., & Hadley, W. (2000). Analyzing and generating mathematical models: An Algebra II Cognitive Tutor design study. In G. Gauthier, C. Frasson & K. VanLehn (Eds.), *Intelligent tutoring systems: 5th international conference* (pp. 314-323). New York: Springer.

Corbett, A.T., Wagner, A., Lesgold, S., Ulrich, H. & Stevens, S. (2007). Modeling students' natural language explanations. *UM2007 User Modeling: Proceedings of the Eleventh International Conference*, 117-126.

Corbett, A., Wagner, A., & Raspat, J. (2003). *The impact of analysing example solutions on problelm solving in a pre-algebra tutor.* Paper presented at the AIED 2003: The 11th International Conference on Artificial Intelligence and Education.

Engelkamp, J. (1998). *Memory for actions*. Hove, England: Psychology Press.

Glenberg, A. M., Gutierrez, T., Levin, J. R., Japuntich, S., & Kaschak, M. P. (2004). Activity and imagined activity can enhance young children's reading comprehension. *Journal of Educational Psychology, 96,* 424-436.

Kalyuga, S., Chandler, P., Tuovinen, J., & Sweller, J. (2001). When problem solving is superior to studying worked examples. *Journal of Educational Psychology, 93*, 579-588.

Mayer, R. E. (2001). *Multimedia learning*. Cambridge: Cambridge University Press.

McLaren, B. M., Lim, S.-J., & Koedinger, K. R. (2008). *When is assistance helpful to learning? Results in combining worked examples and intelligent tutoring.* Paper presented at the Proceedings of the 9th International Conference on Intelligent Tutoring Systems.

Moreno, R., & Mayer, R. E. (2007). Interactive multimodal learning environments. *Educational Psychology Review, 19*, 309-326.

Paivio, A. (1986). *Mental representations: A dual coding approach*. New York: Oxford University Press.

Pashler, H., Bain, P., Bottge, B., Graesser, A., Koedinger, K., McDaniel, M., & Metcalfe, J. (2007). *Organizing Instruction and Study to Improve Student Learning*. Washington, DC: National Center for Education Research, Institute of Education Sciences, U.S. Department of Education.

Reed, S. K. (2005). From research to practice and back: The Animation Tutor project. *Educational Psychology Review, 17*, 55-82.

Reed, S. K. (2006). Cognitive architectures for multimedia learning. *Educational Psychologist, 41*, 87-98.

Reed, S. K. (2008). Manipulating multimedia materials. In R. Zheng (Ed.), *Cognitive effects of multimedia learning* (pp. 51-66). New York: IGI Global.

Reed, S. K. (2010). *Thinking Visually*. New York: Taylor & Francis.

Renkl, A. & Atkinson, R. K. (2003). Structuring the transition from example study to problem solving in cognitive skills acquisition: A cognitive load perspective. *Educational Psychologist, 38*, 15-22.

Salden, R., Aleven, V., Renkl, A., & Schwonke, R. (2008). Worked examples and tutored problem solving: Redundant or synergistic forms of support? In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 589-594).

Schwonke, R., Renkl, A., Krieg, C., Wittwe, J., Aleven, V., & Salden, R. (2009). The worked-example effect: Not an artifact of lousy control conditions. *Computers in Human Behavior, 25*, 258-266.

Sweller, J. (2003). Evolution of human cognitive architecture. In B. Ross (Ed.), *The psychology of learning and motivation* (Vol. 43, pp. 215-266). San Diego: Academic Press.

Sweller, J. & Cooper, G.A. (1985). The use of worked examples as a substitute for problem solving in learning algebra. *Cognition and Instruction*, 1985, *2*, 59-89.

von Gog, T., Paas, F., & Van Merrienboer, J. J. G. (2004). Process-oriented worked examples: Improving transfer performance through enhanced understanding. *Instructional Science, 32*, 83-98.

Learning during Intelligent Tutoring: When Do Integrated Visual-Verbal Representations Improve Student Outcomes?

Kirsten R. Butcher (kirsten.butcher@utah.edu)
Department of Educational Psychology, University of Utah, 1705 E. Campus Center Drive, MBH 327
Salt Lake City, UT, 84108 USA

Vincent Aleven (aleven@cs.cmu.edu)
Human Computer Interaction Institute, Carnegie Mellon University, 5000 Forbes Avenue
Pittsburgh, PA 15213 USA

Abstract

Research has shown that integration of visual and verbal information sources during learning promotes successful student outcomes. However, it is unclear whether it is better to provide students with integrated visual-verbal representations, or to require them to build such integrated representations themselves. In a classroom study, three conditions were used to explore the impact of integrated visual-verbal representations that emphasized rule-diagram mappings in geometry. Students viewed highlighted rule-diagram mappings during learning, generated these mappings themselves, or saw only numerical information embedded in diagrams (control). Students' problem-solving knowledge was measured at posttest and delayed posttest. Overall, students who generated rule-diagram mappings during intelligent tutoring demonstrated better long-term understanding of geometry principles, but effects were only visible at delayed posttest. Results show that integrated visual-verbal representations best support deep learning when they help the learner make connections between features of a visual representation and relevant domain information, and student interactions can be an effective method to scaffold these connections.

Keywords: Intelligent tutoring; Diagrams; Problem solving; Long-term retention; Visual representations

Introduction

Research in multimedia learning has demonstrated that adding visual representations to text materials frequently improves students' learning outcomes (e.g., Carney & Levin, 2002). Studies of cognitive processing with multimedia materials have demonstrated that visual materials support learning by increasing students' generation of effective self-explanations during study (Ainsworth & Loizou, 2003; Butcher, 2006).

However, not all visual representations are equally effective in supporting learning. Diagrams have been shown to be more effective when verbal materials (such as textual labels for diagrams) are integrated directly into the visual representation (e.g., Hegarty & Just, 1993) before they are presented to students. Other research has shown that student-driven integration of visual and verbal materials supports learning with complex materials and may promote goal-oriented behaviors during subsequent, self-directed learning (Bodemer, Ploetzner, Bruchmüller, & Hacker, 2005). Together, these research results suggest clear benefits of integrated visual-verbal representations for learners. However, they also raise the question of whether learners should be provided with integrated visual representations or if it is better to require learners to generate the integrated representations themselves.

The question of whether or not to provide students with integrated visual-verbal representations highlights the *assistance dilemma* (Koedinger & Aleven, 2007). The assistance dilemma refers to the difficulty of deciding when interactive learning environments should provide vs. withhold information in order to support optimal student learning. The assistance dilemma reflects a technology-based application of the long-standing instructional concept of *desirable difficulty* (e.g., Bjork, 1994). Desirable difficulty refers to the finding that increasing the difficulty of a learning activity can improve long-term knowledge outcomes, even though performance during training may suffer. Desirable difficulty argues against a common assumption that optimal learning is facilitated when instructional materials are designed to ease student comprehension and increase successful performance. Thus, a key question for intelligent tutoring systems using visual representations is: when should intelligent tutoring systems provide integrated visual-verbal support vs. withhold this support in order to optimize student learning outcomes?

Connecting Diagrams to Domain Knowledge

The assistance dilemma and the concept of desirable difficulty raise the important question of when to provide vs. withhold integrated visual-verbal representations for optimal learning. However, a central question is what type of integrated representation is most beneficial to learners.

Much of the research on integrated visual representations has made use of visuals that physically embed additional information into a visual representation. Multimedia presentations have been shown to support deeper understanding of instructional materials when they provide students with diagrams into which textual labels and definitions have been embedded (e.g., Hegarty & Just, 1993). In geometry, research has shown that students learn more when they are provided with representations in which numerical measures have been integrated into diagrams (Tarmizi & Sweller, 1988) or when they are provided with color-coded highlighting that links text references (e.g., a

reference to angle ABC) with relevant diagram elements (Kalyuga, Chandler, & Sweller, 1999). Overall, research shows clear benefits for integrated visual-verbal representations during learning. However, integrated visual-verbal representations may not, in and of themselves, prompt learners to make connections to key domain ideas.

Evidence suggests that individuals with deep domain understanding tend to exhibit strong connections between domain concepts and visual representations. For example, experts in geometry use key diagram configurations to cue relevant geometry knowledge (i.e., theorems and principles) during problem solving (Koedinger & Anderson, 1990). During mathematical problem-solving, mathematicians repeatedly analyze connections between generated visual representations, changing goals, and the emerging problem situation (Stylianou, 2002).

Unlike experts, novices do not demonstrate close connections between visual representations and domain knowledge during problem solving. In geometry, novices tend to process diagrams in isolated ways, focusing on visual features without considering their relationship to deeper, conceptual aspects of problems (Lovett & Anderson, 1994). Ainsworth (2006) argues that a central cognitive task in learning with multiple representations is developing an understanding of the relationship between a visual representation and relevant domain information.

One way to support novice learning in geometry, then, may be to scaffold student interactions with visual representations in a way that improves their understanding of the relationship between visual features of geometry problems (i.e., geometry diagrams) and the geometry principles/rules used in problem solving. In geometry, problem solving requires that learners connect meaningful diagram configurations to relevant geometry principles. For example, in Figures 2 and 3, angle ABC is an *interior angle, same side* to angle BCD. Learners should recognize that the diagram contains two parallel lines (AB, DC) intersected by a transversal (BC). Angles ABC and BCD are on the interior of the parallel lines, and on the same side of the transversal. Thus, they are *interior angles, same side* and can be solved using this rule. In this study, we used highlighted diagram features to demonstrate the mapping between diagrams and relevant geometry principles in the domain (see Figures 2 and 3); hereafter, these are referred to as diagram-domain representations.

Integrated Diagrams in Intelligent Tutoring

In previous research (Butcher & Aleven, 2007, 2008), we explored the use of interactive visual diagrams as a method to support the development of integrated visual-verbal knowledge during intelligent tutoring in geometry. The research vehicle for this work was the Geometry Cognitive Tutor, an intelligent tutoring system (ITS) grounded in cognitive theory that provides multiple forms of support for student learning by doing: tracking students' knowledge development using a model of student competency, selecting problems for students to complete that match identified learning needs, structuring problem-solving steps for students, giving feedback on all student actions, and providing hints upon student request or when the student makes repeated errors. Details about Cognitive Tutor features are available elsewhere (e.g., Anderson, Corbett, Koedinger, & Pelletier, 1995).

Butcher and Aleven (2007, 2008) varied the site of student interaction during geometry problem solving in an intelligent tutoring system: students used either an *interactive diagram* or a *solutions table* version of the intelligent tutoring system (see Figure 1). Students using the interactive diagram tutor clicked directly on diagram elements to enter answers and receive feedback, thus creating an integrated representation in which numerical answers were embedded in the visual representation. Students in the control condition used the solutions table to enter their answers and receive feedback. Although the solutions table kept a running record of students' answers, numerical values were not integrated directly into the diagram. Results showed that students who interacted with the diagrams to develop an integrated representation learned geometry principles more deeply (as evidenced by transfer task performance: Butcher & Aleven, 2007) and retained their problem-solving skills for longer periods of time (Butcher & Aleven, 2008).

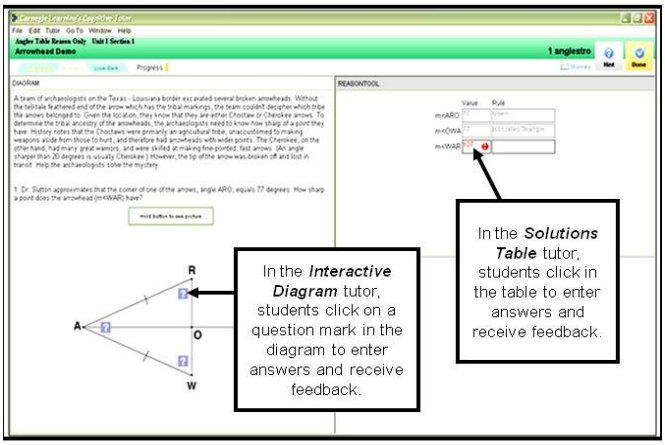


Figure 1: Condition-based differences in interactions with an intelligent tutor (Butcher & Aleven, 2007, 2008).

Despite the success of these diagram interactions in an already-successful intelligent tutoring system, there was still ample room for student improvement at assessment. It is possible that diagram interaction helped students focus on relevant visual elements during problem solving, but the integrated representations that students developed did not make it clear how diagram elements mapped onto domain information.

Student Generation of Integrated Representations
Simply providing students with visual representations that connect diagrams features to domain information may not be optimal for learning. Research has shown that requiring

students to actively integrate visual and verbal information (i.e., using a drag-and-drop interface to produce a labeled diagram) improves learning outcomes and increases the quality of students' self-directed learning behaviors (Bodemer et al., 2005). However, it is unclear whether interactions that emphasize diagram-domain mappings during problem-solving practice can improve learning more than interactions that build integrated visual-verbal representations (cf., Butcher & Aleven, 2007, 2008).

The purpose of this study was to explore the potential benefits of providing students with integrated representations that emphasized the mapping between diagram elements and domain information during intelligent tutoring vs. requiring students to generate these integrated representations. Both conditions were compared to a control condition in which students interacted with diagrams to embed numerical information into the diagrams (i.e., student interactions created an integrated, visual-verbal representation that did not emphasize domain connections).

Method

Participants

Eighty-three students from five 10th grade geometry classrooms at a vocational school in rural Pennsylvania participated in the study as part of their normal classroom curriculum, which included practice with the Geometry Cognitive Tutor once a week (one 75 min session per week).

Grade-matched triplets of students were identified within each class, using students' first semester geometry grades as a measure of prior knowledge. From every grade-matched triplet, one student was randomly assigned to each of the three experimental conditions described below.

Materials

Student-Highlighting Condition The purpose of the student highlighting condition was to require student interactions with the intelligent tutor that generated integrated diagram-domain representations during problem-solving practice in the Cognitive Tutor. In this condition, if a student entered an incorrect answer or reason during practice, s/he was locked out of the numerical answer field until s/he identified the correct geometry principle needed to solve the problem-solving step. Once the correct principle was identified, students highlighted the diagram features relevant to that principle (see Figure 2). These highlights created an integrated diagram-domain representation of the problem situation. As seen in Figure 2, highlighting was scaffolded by a list of diagram features that appeared after students entered a correct geometry principle for a problem-solving step. Students were required to highlight each diagrammatic feature in the list (e.g., for *Interior Angles, Same Side*, students were prompted to highlight the parallel lines, the transversal, and the two relevant angles).

Students highlighted a diagrammatic feature by clicking directly on it; students could deselect a highlighted feature by clicking on it again. Students received immediate feedback on each highlighted feature in the diagram. Incorrect highlights turned red on the diagram and in the accompanying answer area. Correct highlights were kept on the screen until the problem-solving step was completed.

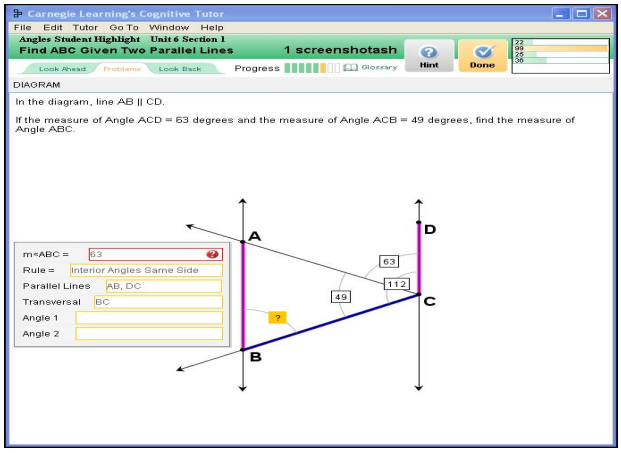


Figure 2: In-progress student highlighting of diagram for *interior angles, same side* rule. Parallel lines and transversal have been highlighted so far.

Tutor-Highlighting Condition This condition utilized the same representations as the student-highlighting condition, but in this case the tutor *provided* students with the highlighted diagram-domain representation. Following a problem-solving error and student identification of a relevant geometry principle, the tutor automatically highlighted the diagram. The screen shot in Figure 3 shows the result of the tutor highlighting; it is important to note that the final representations in the student- and tutor-highlighting conditions were equivalent, differing only in whether the student or tutor generated the representation.

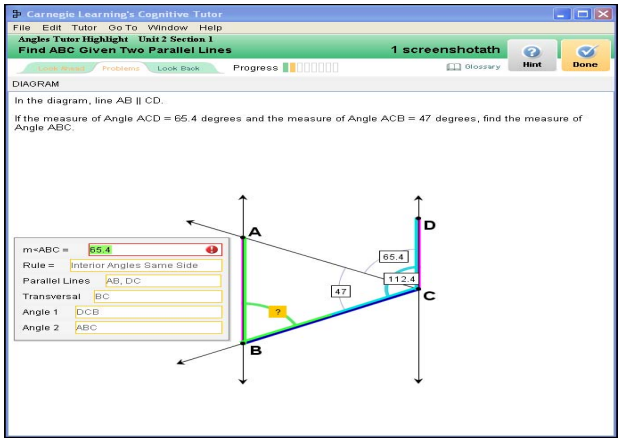


Figure 3: Tutor-highlighted diagram following student error

No Highlighting (Control) The control condition was the successful interactive-diagram version of the Geometry Cognitive Tutor from Butcher and Aleven (2007). This condition did not involve any highlighting of visual diagram features by either students or the intelligent tutoring system.

However, students entered answers directly into the geometry diagram; this created an integrated visual-verbal representation in which numerical values were embedded in the visual diagram.

Assessments *Problem-Solving Pre- and Posttest* The problem-solving pre- and posttest consisted of 16 total items. For each to-be-solved item, students needed to provide a numerical answer (e.g., 65º) and the geometry principle that was used to derive the numerical answer (e.g., Vertical Angles). The problem-solving posttest was the same as the pretest, but problems appeared in a different order. One point was given for each correctly-solved angle and correctly-identified principle. Due to a technical error and student absences, data was collected from 68 participants at pretest and from 70 students at posttest.

Delayed Posttest The delayed posttest was given on the computer, four weeks following the posttest. The delayed posttest followed the same format as the pre- and posttest, but with less complex problems. Students received one point per correctly-solved angle and correctly-identified geometry principle, for a maximum of 8 points on each dependent measure. Due to high numbers of student absences in the week that the delayed posttest was given (near the end of the school year), 41 students completed the delayed posttest.

Procedure

Participants were given up to 30 minutes to complete the pretest during their geometry class. Pretests were delivered via computer; students were instructed to try their best to complete the problems, and to take a guess if they were not sure of an answer. After completing the pretest, students worked with their assigned tutor version for four weeks during a 75-minute, weekly computer lab. This computer lab was a normal part of the students' geometry classes, and all students had used non-experimental versions of the Geometry Cognitive Tutor during previous sessions in the computer lab. The Geometry Cognitive Tutor used in each condition did not differ in problem content, the number of required problems, or the knowledge models used by the Cognitive Tutor.

One week after completing the study, students were given up to 45 minutes to complete the posttest during their geometry computer lab. A delayed posttest was administered one month following the posttest. Participants had up to 30 minutes to complete the delayed posttest.

Results and Discussion

Training Performance

In the Geometry Cognitive Tutor, learners provide a numerical answer and a geometry principle (aka "rule) that justifies the numerical answer for each problem-solving step. Log data from student practice with the Geometry Cognitive Tutor were analyzed to assess performance on the first answer and geometry rule attempted by a learner for each problem step during practice. Data were calculated only for problem steps that were not given in the problem statement. That is, data were analyzed only for problem steps in which students needed to apply a geometry principle in order to calculate a correct answer. Student progress in the Geometry Cognitive Tutor was self-paced and, in general, was slower than anticipated by either the experimenters or the students' classroom teachers. Because the intelligent tutoring system requires mastery learning before students can continue to the next instructional unit, not all students completed the three instructional units in the experimental version of Geometry Cognitive Tutor. In total, 72 students produced tutor log data in unit 1 (control: n = 23, tutor-highlighting: n = 25, student-highlighting: n = 24). Forty-five students reached unit 2 (control: n = 14, tutor-highlighting: n = 16, student-highlighting: n = 15), but only 27 students reached unit 3 (control: n = 10, tutor-highlighting: n = 8, student-highlighting: n = 9).

Due to the drop in student numbers at each instructional unit, three multivariate analyses of covariance (MANCOVAs) were used to assess student performance in each unit of the tutor. Dependent variables were the percent correct of students' initial attempts at numerical answers and geometry rules for each not-given problem-solving step. Students' pretest scores on numerical answers and geometry rules were used as covariates to control for prior knowledge. As seen in Figure 4, unit 1 data demonstrated no significant differences in practice performance on numerical answers or geometry rules (Fs < 1).

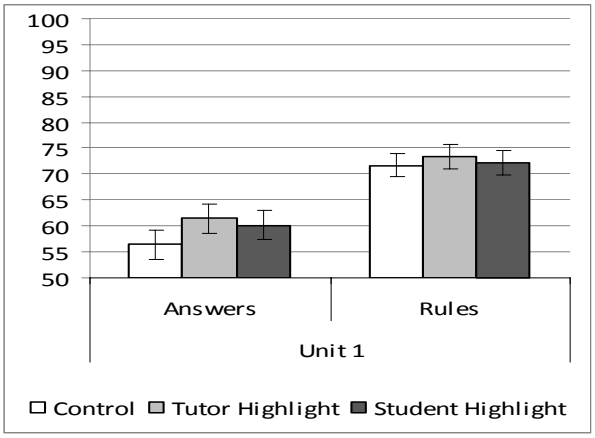


Figure 4: *M* (and *SE*) percent correct for answers and geometry rules in unit 1 during intelligent tutor practice

Unit 2 also failed to show any significant condition differences in problem-solving performance on answers or geometry rules (Fs < 1). For the few students who reached unit 3, students who interacted with the tutor to generate integrated diagram-domain representations had a slight, though non-significant, advantage on numerical answers ($F_{(2, 22)} = 2.7$, $p < .09$). However, as seen in Figure 5, there were no differences in students' accuracy in using geometry rules to justify their problem-solving steps during practice.

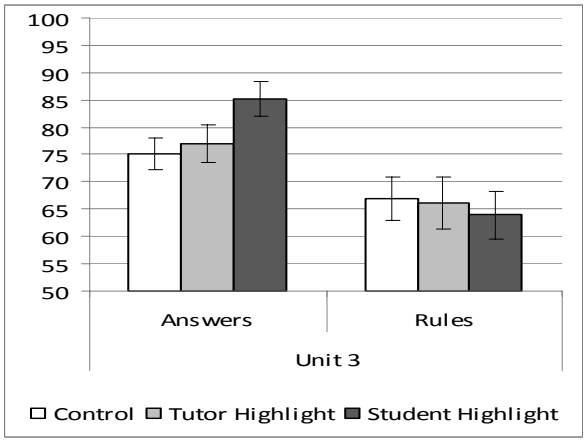


Figure 5: *M* (and *SE*) percent correct for answers and geometry rules in unit 3 during intelligent tutor practice

Problem-Solving Performance

Overall, 33 students completed all three assessments (control: n = 13, tutor-highlighting: n = 11, student-highlighting: n = 10). Data were analyzed using a repeated-measures MANOVA, where test time (pretest, posttest, delayed posttest) was the repeated factor.

For numerical answers, results showed no test time by condition interactions (Linear: $F < 1$; Quadratic: $F_{(2, 31)} = 1.48$, $p > .24$). However, as seen in Table 1, students' performance on geometry principles showed a significant test time by condition interaction (Linear: $F_{(2, 31)} = 4.97$, $p = .01$, $\eta_p^2 = .24$; Quadratic: $F_{(2, 31)} = 3.28$, $p = .05$, $\eta_p^2 = .18$). Students in the student-highlighting condition were best able to justify their problem-solving steps with geometry rules at delayed posttest; however, no differences were seen at the short-term posttest. Figures 6 and 7 show the pattern of means on the posttest and delayed posttest, respectively, adjusted for pretest performance.

Table 1: *M* (and *SD*) percent correct on geometry rules

	Pretest	Posttest	Delayed Posttest
Control	18.2 (16. 7)	25.5 (14.3)	19.4 (18.0)
Tutor-Highlighting	11.1 (10.6)	23.2 (21.2)	17.7 (11.4)
Student-Highlighting	12.4 (7.2)	16.3 (13.6)	31.7 (14.2)

As seen in Figure 6, there were no significant condition differences at posttest. If anything, the pattern of results at posttest was consistent with a disadvantage for students who highlighted diagrams during practice. Although this may seem inconsistent with the overall pattern of performance in unit 3 during intelligent tutoring practice (see Figure 5), one should remember that not all students taking the posttest reached unit 3 in the intelligent tutor.

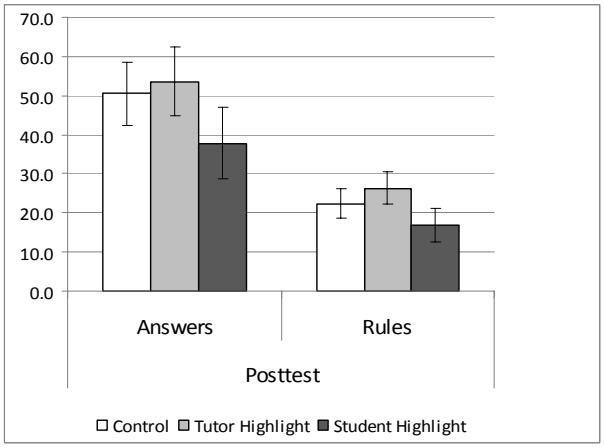


Figure 6: *M* (and *SE*), adjusted for pretest performance, on posttest numerical answers and geometry rules.

One month later, at delayed posttest, the data paint a different picture Although there were no differences in students' accuracy in providing numerical answers at delayed posttest, students who generated integrated diagram-domain representations during practice were better able to justify their problem-solving steps with relevant geometry rules. It is important to note that this advantage was found even though control students made use of integrated diagrams with embedded numerical answers. Moreover, the advantage cannot be attributed to additional information in the diagram-domain representations, as students who were provided with these representations by the tutor did not outperform the control group in correctly using geometry rules (see Figure 7).

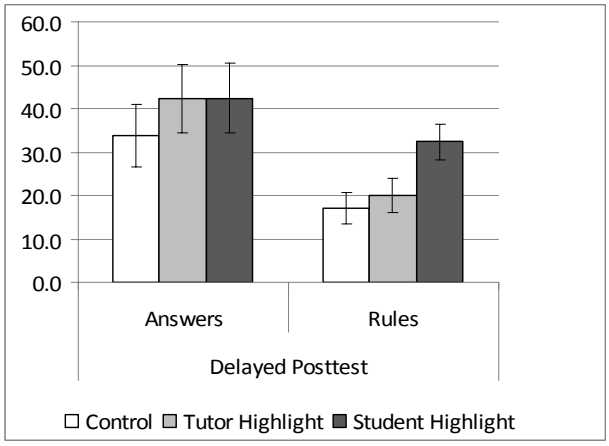


Figure 7: *M* (and *SE*), adjusted for pretest performance, on delayed posttest numerical answers and geometry rules.

To confirm results obtained from the small group of students with full assessment data, two additional analyses were conducted. First, a MANCOVA was used to assess performance changes for all 68 students with pre- and posttest data (control: n = 23, tutor-highlighting: n = 24, student-highlighting: n = 21). Dependent variables were

performance on numerical answers and geometry rules at posttest; covariates were students' performance on answers and rules at pretest. Results were consistent with the small sample, showing no condition differences for either numerical answers or geometry rules (Fs < 1). A second, similar MANCOVA was conducted for all 41 students with pre- and delayed posttest data (control: n = 14, tutor-highlighting: n = 13, student-highlighting: n = 14). Results again were consistent with the small sample, showing a significant advantage of the student-highlighting condition for geometry rules ($F_{(2, 36)} = 4.04$, $p = .03$, $\eta_p^2 = .18$), but not numerical answers ($F_{(2, 36)} = 1.38$, $p > .26$).

General Discussion

Overall, results show that *providing* integrated visual-verbal materials to students during intelligent tutoring does not improve students' learning outcomes. However, findings show that using *interactions to build* integrated diagram-domain representations can support long-term understanding. Students who generated integrated representations that emphasized diagram-domain mappings during problem-solving practice showed no performance advantages in using geometry principles at practice or posttest, but were best able to apply these principles one month following instruction.

Results are consistent with the idea that student interactions can support deep learning with visual information. However, results also argue that integrated visual-verbal representations best support deep learning when they help the learner make connections between features of the visual representation and relevant domain information. The current study shows that student interactions can be an effective method to scaffold these connections. Findings also demonstrate the importance of measuring long-term knowledge gains, as student performance during practice and short-term assessments may not provide an accurate picture of deep understanding.

Acknowledgments

The authors thank Octav Popescu, Carl Angiolillo, Michael Nugent, Grace Leonard, and Thomas Bolster for their contributions. Special thanks to Mark Schoming and Colleen Conko for assistance in the classroom. This work was supported in part by the Pittsburgh Science of Learning Center through funding from the National Science Foundation (SBE# 0354420, 0836012). The opinions and conclusions presented here are those of the authors and do not necessarily reflect the funding agency.

References

Ainsworth, S. (2006). DeFT: A conceptual framework for considering learning with multiple representations. Learning & Instruction, 16(3), 183-198.

Ainsworth, S., & Loizou, A. T. (2003). The effects of self-explaining when learning with text or diagrams. Cognitive Science, 27, 669-681.

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. Journal of the Learning Sciences, 4(2), 167-207.

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. Shimamura (Eds.), Metacognition: Knowing about knowing (pp. 185-205). Cambridge, MA: MIT Press.

Bodemer, D., Ploetzner, R., Bruchmüller, K., & Hacker, S. (2005). Supporting learning with interactive multimedia through active integration of representations. Instructional Science, 33, 73-75.

Butcher, K. R. (2006). Learning From Text With Diagrams: Promoting Mental Model Development and Inference Generation. Journal of Educational Psychology, 98(1), 182-197.

Butcher, K. R., & Aleven, V. (2007). Integrating visual and verbal knowledge during classroom learning with computer tutors. In D. S. McNamara & J. G. Trafton (Eds.), Proceedings of the 29th Annual Cognitive Science Society (pp. 137-142). Austin, TX: Cognitive Science Society.

Butcher, K. R., & Aleven, V. (2008). Diagram interaction during intelligent tutoring in geometry: Support for knowledge retention and deep understanding. In B. C. Love, K. McRae & V. M. Sloutsky (Eds.), Proceedings of the 30th Annual Conference of the Cognitive Science Society (pp. 1736-1741). Austin, TX: Cognitive Science Society.

Carney, R. N., & Levin, J. R. (2002). Pictorial illustrations still improve students' learning from text. Educational Psychology Review, 14(1), 5-26.

Hegarty, M., & Just, M. A. (1993). Constructing mental models of machines from text and diagrams. Journal of Memory and Language, 32, 717-742.

Kalyuga, S., Chandler, P., & Sweller, J. (1999). Managing split-attention and redundancy in multimedia instruction. Applied Cognitive Psychology, 13(4), 351-371

Koedinger, K. R., & Aleven, V. (2007). Exploring the assistance dilemma in experiments with cognitive tutors. Educational Psychology Review, 19(3), 239-264.

Koedinger, K. R., & Anderson, J. R. (1990). Abstract planning and perceptual chunks: Elements of expertise in geometry. Cognitive Science, 14(4), 511-550.

Lovett, M. C., & Anderson, J. R. (1994). Effects of solving related proofs on memory and transfer in geometry problem solving. Journal of Experimental Psychology: Learning, Memory, and Cognition, 20(2), 366-378.

Stylianou, D. A. (2002). On the interaction of visualization and analysis: the negotiation of a visual representation in expert problem solving. Journal of Mathematical Behavior, 21, 303-317.

Tarmizi, R. A., & Sweller, J. (1988). Guidance during mathematical problem solving. Journal of Educational Psychology, 80(4), 424-436.

Explanatory Reasoning for Inductive Confidence

David Landy (dlandy@richmond.edu)
Department of Psychology, M03B Richmond Hall
University of Richmond, VA, 23173 USA

John E. Hummel (jehummel@illinois.edu)
Department of Psychology, 603 E. Daniel St.
Champaign, IL 61820 USA

Abstract

We present Explanatory Reasoning for Inductive Confidence (ERIC), a computational model of explanation generation and evaluation. ERIC combines analogical hypothesis generation and justification with normative probabilistic theory over statement confidences. It successfully captures a broad range of empirical phenomena, and represents a promising approach toward the application of explanatory knowledge in new situations.

Keywords: induction; analogy; probabilistic reasoning

Introduction

We are constantly making guesses. When we come across something new, we know about it in part from its relations to other things and we attribute to the novel the properties of the familiar. For instance, when Apple announced the iPad, technology reporters alternately compared it to tablet PCs, which are similar in size and function, and the iPhone, which is similar in appearance and operating system. In each case, the game was to predict the features of the new object on the basis of the old ones.

Property inductions of this kind—extending known properties of one category to other categories—have been heavily studied in experimental psychology (see Heit, 2000, for a review). Such inductions seem to take advantage of taxonomic knowledge about category structures as well as specific knowledge about particular categories (Shafto, Kemp, Bonawitz, Coley, & Tenenbaum, 2008).

One intuition, pursued here, is that people make inductions by adapting explanations for known properties to novel categories. People are habitual generators of explanations: Scientists explain natural phenomena; engineers explain why structures will or will not support various loads; mathematicians explain why a formal property does or does not hold of a particular situation or object; and everyone routinely explains much more mundane things such as why the doorbell rang, why we smell gas in the kitchen and why a child has a fever. Explanations serve many cognitive functions, but perhaps none is more important than their ability to support inductive inferences: A person who can explain a novel observation can have much greater confidence in their inferences about the circumstances under which that observation is likely to be repeated than a person who cannot explain it—which is why, for example, your auto mechanic is better than you are at knowing whether that strange noise you car is making is likely to be dangerous.

In order to apply explanations of past experiences to novel situations, a cognitive architecture must solve several problems. First, it must be able to generate and retain explanations in the first place. Second, it must have a way to generate novel hypotheses about a current situation from its beliefs about past circumstances. Finally, it must be able to distinguish when a novel explanation is plausible in the current situation, and when it is not.

Bayesian models, and particularly hierarchical Bayesian models, are adept at the last of these goals. For example, the model of Kemp and Tennenbaum (2009) carves known situations into disjoint domains, and applies to novel situations the domain assumptions that appear most appropriate. However, human reasoners also adapt explanation patterns across multiple, dissimilar domains (Medin, Coley, Storms, & Hayes, 2003). Although such cross-domain reasoning is the sine qua non of analogical approaches to reasoning (Falkenhainer, Forbus, & Gentner, 1989; Hummel & Holyoak, 1997), models of analogy generally provide no basis for generating probabilistic estimates of confidence in their inferences.

In this paper, we present the model ERIC, Explanatory Reasoning for Inductive Confidence (see also Landy & Hummel, 2009). ERIC uses a combination of analogical and probabilistic reasoning to (a) generate explanations for newly learned facts, (b) evaluate the plausibility of those explanations in light of its existing knowledge, (c) use those explanations to update its confidence in its existing knowledge and (d) make judgments about the plausibility of new inferences. The resulting model accounts for a large body of empirical findings from the literature on inductive confidence (e.g., Heit, 2000; Shafto et al., 2008).

A central tenet of the model is that the mind uses analogy to adapt old explanations to new situations and then uses those new explanations both to determine its confidence in the new observation and to update its confidence in its existing knowledge—both existing basic facts and existing explanations. The knowledge updated includes both the source analogs (i.e., the old explanations used to generate the new ones) and the analogies themselves (i.e., the mappings from the old [source] explanations to the new [target] explanations). As a result, if an analogy results in a good explanation, then the model becomes more convinced both that the source was true and that the analogy was good.

A second central tenet is that the mind generates these explanations permissively and habitually: Presented with any new "fact" or observation, the mind will generate as many potential explanations of that fact as possible and assign a likelihood or confidence value to each; in turn, these values are used to update its confidence in the very facts that participated in the explanations themselves.

Some models of induction (e.g., Kemp & Tennenbaum, 2009) explicitly carve knowledge into separate domains, and assume that categorically different processes apply to situations attributed to those domains (e.g., reasoning in one way about ontological knowledge and in a different way about geographical knowledge). A third tenet of the model is that knowledge, including knowledge about generating processes, is applied to relevant situations regardless of domain. That is, the processes underlying explanation and confidence estimation are the same across and within all areas of knowledge: Any differences between, say, ontological knowledge and other knowledge domains (e.g., geographical location, diet or behavioral traits) emerge as a natural consequence of the relationship between individual sets of facts, and not through an explicit and absolute categorization into domain.

Finally, in line with other integrative general knowledge models, the goal of ERIC is not to be entirely formally consistent (Wang, 2009). For instance, it will not necessarily be the case that $a \wedge \sim a$ is guaranteed to be false.

Property Induction

We report a collection of simulations using ERIC to perform a property induction task (Osherson, Smith, Wilkie, López, & Shafir, 1990; Rips, 1975). In this task, a subject (or ERIC) is given a *premise*, which is assumed to be true (e.g., "robins get disease *d*"), based upon which they are asked to estimate the likelihood of a *conclusion* (e.g., "birds get *d*"). The dependent measure of interest is the estimated likelihood of the conclusion as a function of the relation between the major term in the premise (here, "robins") and that in the conclusion ("birds"), and of the relation between these categories and the property induced ("disease *d*").

ERIC

Overview

ERIC is based on the following assumptions about the nature of the property induction task:

1. A person enters the laboratory with knowledge (facts, explanations, theories) believed in with varying degrees of confidence.
2. Faced with the premise, the subject tries to explain it by building a fairly large set of potential explanations by analogy to known cases.
3. Each explanation is assigned an inductive confidence that combines confidence in the knowledge involved in the explanation and confidence in the generating analogies.
4. These explanations are added (provisionally) to knowledge, and the confidence of existing statements is updated using Bayesian inference.
5. Faced with a conclusion, the subject repeats process of explanation and confidence updating.
6. Confidence in the conclusion is high to the degree that the explanations are strong.

As input, ERIC takes an explanandum—either a premise or a conclusion. As output, it generates potential explanations, each with an assigned confidence, and an estimate of the confidence in the explanandum itself. Applied to property induction, the mechanism operates in two stages: First, ERIC explains the premise(s) and any knowledge gleaned from those explanations is added to the knowledge base. Next, it explains the conclusion using that augmented knowledge. The result of these processes is an estimate of the likelihood that the conclusion is true.

Knowledge Representation

All of ERIC's knowledge is represented in standard propositional notation, augmented to capture the logical and causal relations that link propositions into explanations. Atoms are of the form *f(a), g(a, b, c)*, and so on. Connectives $\wedge$, $\vee$, and $\sim$ are used in their usual sense to mean *and*, *or*, and *not*.

Two less universal connectives provide a language for representing explanations and analogical mappings. The connective $\Rightarrow$ denotes an explanatory or causal relationship. For example, $q \Rightarrow r$ should be read as "*q* (if true) would tend to explain (cause) *r*." In contrast to some prior models (e.g., Falkenhainer et al., 1989; Hummel & Holyoak, 1997), causal connections are treated as special types, and not as generic two-place predicates (see also Hummel & Landy, 2009). Syntactically, they are equivalent in ERIC to a material conditional.

The second novel connective is the *mapping relation*, $q \leftrightarrows r$, which asserts that *q* and *r* map to each other in some analogy, and provides ERIC's initial estimate that *q* and *r* might map to each other in some future analogy. Mapping connections have learned confidences.

Confidence Each statement, *q*, is assigned a *confidence* value between 0 and 1, which is intended to work much like an intuitive probability that the statement is true. Indeed, we will refer to the confidence as "the probability of *q*," or p(*q*).

Statements in the initial knowledge set have a preset initial confidence. Regular property statements and cause relations (e.g., $q \Rightarrow r$) that do not appear in the initial knowledge have a confidence set to arbitrary low values (0.1 and 0.001).

Explanations An explanation is a recursive binary modal structure, with the pattern *E(explanation; explanandum)*, where the explanandum is a statement, and the explanation is a set of statements. They have the form of a modus ponens: Some set of (possibly recursively justified) causes

and an explanatory connective statement justify the effects. For instance, the explanation:

$$E_1(p,q,E_2(r,r \Rightarrow q;q),p \wedge q \Rightarrow s;s).$$

asserts that "*p*, *q* (where *q* is explained by *r*), and [*p* and *q* cause *s*] jointly cause *s*."

An explanation differs from a causal connective in several ways. First, a causal connective is purely dispositional, while an explanation asserts that *in fact*, the explanation explains the explanandum. An explanation thus encodes a derivation pattern, rather than a potential relationship. Further, an explanation carries its own internal semantics; it denotes a possible state of affairs.

Knowledge base ERIC's knowledge consists of three major classes of statements: simple property statements, such as *eats(Robin, Worm)*; simple explanations, such as generic taxonomic explanations of the form *isa(A, B) ^ x(B)* ⇒ *x(A)*; and taxonomic assertions, of the form *isa(Robin, Bird)*. It is worth noting here that taxonomic assertions are simply property statements, and not a special part of the model mechanism.

Justification

ERIC revises its beliefs (e.g., explanations) using two kinds of justification: analogical and explanatory. For either, the effect of a justification, *j*, on an explanandum, *i*, is to update the probability of *i* according to a probabilistic-OR rule:

$$p(s) \leftarrow p(j) + (1 - p(j))p(s) \qquad (1)$$

Intuitively, (1) can be read as meaning that if the justification, *j*, is correct, then the assertion, *s*, it justifies must be correct, but if it is not, then *s* might still be correct with (base rate) probability *p*(*s*).

The initial confidence of an explanation is simply the probability that all the statements in the explanation are true:

$$p(j) = \prod_{e \in E} e \qquad (2)$$

Analogical Justification Intuitively, an analogy, $r \leftrightarrows q$, justifies *q* to the extent that the source analog (*r*) is true, and the mapping is reliable. Thus,

$$p(j) = p(r)p(r \leftrightarrows q) \qquad (3)$$

The target of an analogical justification is always a causal statement. These are updated by applying the justification to the cause statement via equation (1), just as with explanatory justification.

Explanation Generation

When a new explanandum, *q*, is presented to ERIC, two steps are recursively applied to generate new explanations of *q*. First, each fact in the current knowledge base that shares any literals with *q* is postulated as a possible explanation for *q*. For example if *q* = *g(a)* and if *f(a)* is known, then one explanation postulated will be *f(a)⇒g(a)*. Confidence in this shallow explanation will initially be set to a very low value. Second, existing explanations (including those inside explanations) are expanded and justified by analogy to other explanations in knowledge.

Any potentially useful analogical mapping, e.g., $(a \Rightarrow b) \leftrightarrows (c \Rightarrow d)$, is computed by mapping the elements of *a,b* onto those of *c,d* using Holyoak and Thagard's (1989) ACME mapping algorithm. ACME's mapping strengths range between 0 and 1, and so translate conveniently into confidences. ACME combines structural isomorphism and semantic relationships. In ERIC, these semantic relationships are computed directly from the knowledge base (see Projectable Literals, below).

The best match produced by ACME is used as the basis for an analogy. This approach has two effects. First, the explanatory relation is justified by the analogical statement, using (3). Second, statements appearing in the analog but not in the current explanation are imported.

These two processes are applied to each explanation in the current set a fixed number of times (three in the current simulations). Each explanation in the final set justifies the conclusion; the result is the confidence in the conclusion.

Projectable Literals Analogical similarity integrates structural overlap and semantic relationships (Taylor & Hummel, 2009). That is, structural relations being equal, ERIC prefers analogies about identical or similar terms to comparisons among distantly related items.

The semantic similarity—more accurately, *projectability* (Simmons & Estes, 2008; Sloutsky, Kaminski, & Heckler, 2005)—of *a* onto *b*, p_{ab}, can come from either of two sources. If two terms have been related by past explanatory analogies, then the projectability is stored in the form of a mapping statement. The projectability of two previously unrelated terms is calculated from ERIC's knowledge:

$$p_{ab} = e^{-d_{ab}} \qquad (4)$$

where

$$d_{ab} = \alpha s_a + \beta s_b - \gamma s_{ab} - \delta m_{ab}$$

α, β, γ and δ are free parameters (15/40, 2/40, 1/40, and 17/40, respectively). Here s_a is the summed confidence in sentences in which *a* appears; s_b and s_{ab} are defined analogously. Intuitively, *a* is projectable onto *b* to the extent that they appear in similar relational roles in LTM (γs_{ab}) or to the extent that *b* is a kind of *a* (δm_{ab}) and to the extent that *a* does not appear in roles in which *b* does not and vice-versa ($\alpha s_a + \beta s_b$). If a mapping connection exists between *a* and *b* then ERIC uses the mapping strength as p_{ab}: ERIC learns that facts about *a* generally apply to *b*.

The differential applicability of known explanations to novel situations constructs a kind of soft domain separation. Although any knowledge can be applied to a new situation in principle, close knowledge will be applied with far more

confidence. As a result, cross domain analogies have most effect in the absence of other good explanations. This differential applicability of old explanation replaces the construction of explicit domains of explanations (Kemp & Tenenbaum, 2009) used in other approaches, and in general may implement generic symbolic rules (Gentner & Medina, 1998; Sun, 2006).

Knowledge Revision

In property induction, a certain number of premises (collectively π) are followed by a conclusion statement, *c*. In calculating confidence in a conclusion, ERIC first generates explanations of the premises. It uses these to update its knowledge base. If π consists of multiple premises, then each individual premise is explained; the full set of explanations is the set of all possible combinations of explanations for individual premises.

Learning a new premise means adding it to the knowledge base with confidence=1. Learning a new fact should inform the learner to the degree that the fact was surprising; it should increase confidence in things that would explain that fact. Both intuitions can be captured by Bayes law, if we are careful about where our terms come from.

$$p(e|\pi) = \frac{p(\pi|e)p(e)}{p(\pi)} \qquad (5)$$

The prior probability, $p(\pi)$, is the confidence in π resulting from the explanation process. Intuitively, $p(\pi|e)$ is the confidence we would have in π if some particular fact e were known with certainty. This value can be found by repeating the process of justifying π, setting the confidence of e to 1 for each fact that appears in explanations for π, including analogy sources, and assertions of analogical validity. It should be clear that the use of this law is not normative here, since the values are not strictly probabilities. However, the law forms one good way to incorporate evidence into belief systems. ERIC postulates that people use something like this kind of inference.

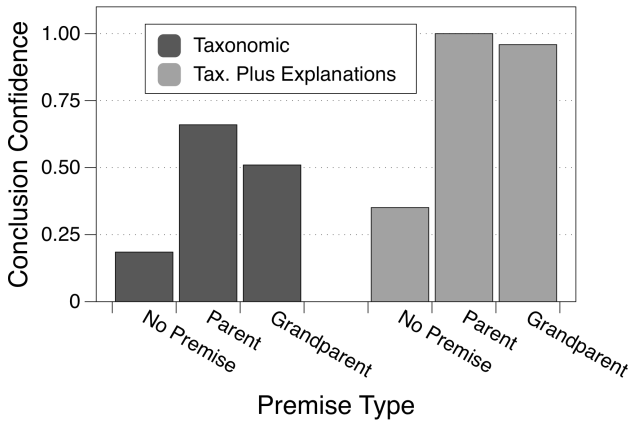


Figure 1: The strengths of induction of a property from one category to a related category. In general, ERIC makes stronger inductions from more closely related categories.

In property induction, ERIC uses the knowledge base that results from explaining the premise to explain the conclusion. Since each explanation justifies the conclusion, confidence in the conclusion results from the application of (1) once for each explanation.

In principle, the resulting confidence values could be matched directly to human probability estimates. In practice, current limitations of the model (especially its extremely impoverished "knowledge") make such point-by-point comparison uninformative, so our evaluation of the model will focus on the relative rankings of sets of explanations.

Simulations and Results

ERIC predicts that inductions, and even patterns of inductions, will be strongly dependent on knowledge, and particularly on contextually relevant knowledge. For this reason, conclusions about the predictions of ERIC must be made relative to some particular set of knowledge.

Taxonomic Simulations

Taxonomic relationships have received much attention in the literature on category inductions; we decided to explore two knowledge bases built largely around taxonomic knowledge. In the first, a taxonomic structure of "animals" was constructed with *isa* statements, including two mammals, six birds, and two reptiles. Animals were, in turn, defined by membership to the superordinate "living things." One general taxonomic explanation was included, over elements that did not appear in any other statements. The pattern of this explanation was: $isa(x,y) \wedge f(y) \Rightarrow f(x)$.

The second knowledge base included all of these taxonomic facts, but also included a fairly arbitrary set of about 200 facts, including property statements and casual explanations, both taxonomic and not taxonomic. This knowledge base tests the generality of the conclusions across a noisier knowledge base.[1]

Since inductions from a category to its subset are explanations, like all explanations, they are not certain. Furthermore, close ancestors generally provide more support than more distant ancestors. Figure 1 compares ERIC's inductions from immediate superordinates of a category ("parents"), and from the superordinates' superordinates ("grandparents" see Figure 1). Thus, a premise "birds have x" provides more support to the conclusion "robins have x" than does "animals have x". This pattern matches the empirically discovered category inclusion fallacy (Heit, 2000; Sloman, 1998). Figure 1 shows that this same pattern appears with the richer knowledge base, as well.

Within taxonomic categories at the same level (e.g., the species level), taxonomic proximity again can vary. Figure 2 shows the results of simulations varying the taxonomic proximity, and also the number of premises in the induction

[1] The full contents of all knowledge bases described here can be found online at http://www.richmond.edu/~dlandy/cogsci10/.

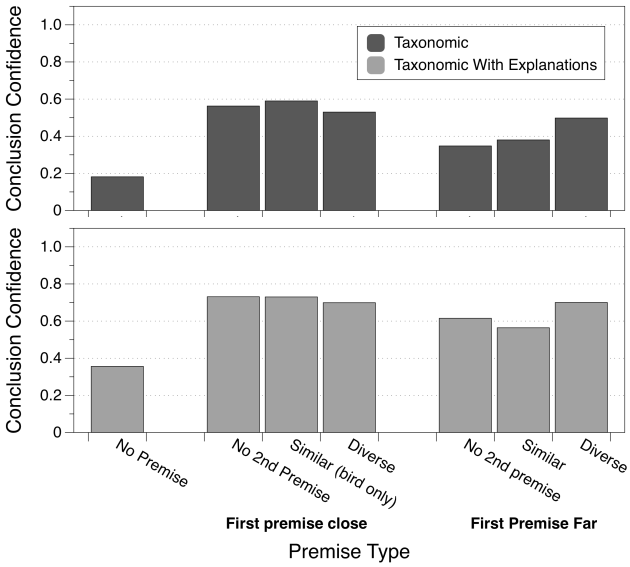


Figure 2: The strengths of induction of a property from zero, one or two categories to others at the same level.

(that is, the number of species of which the property was asserted). In the absence of knowledge, ERIC generally predicts that inductions tend to be stronger between categories that are closely related (see Figure 2). More premises tend to make inductions stronger; moreover, ERIC shows a general diversity effect: when multiple premises come from unrelated categories, that tends to increase inductions more than when they have a common superordinate. This is true in general because two close premises will tend to be best explained by explanations in terms of their common superordinate, while diverse premises are likely to be explained in terms of distant superordinates. This pattern is complicated, however, by an interaction between the diversity of the premises and their similarity to the conclusion. If one premise category is close to the conclusion category, a single premise category already generalizes fairly strongly, because most explanations for the premise are highly mappable into the conclusion; adding a second close premise improves the induction very slightly or not at all. However, if the second premise is from a very different category (making the premises more diverse), then ERIC's explanations are likely to be less finely tuned to the conclusion category, and confidence decreases slightly. This pattern again matches empirical literature (Osherson et al., 1990; Sloman, 1993).

Typicality

To explore how ERIC uses typicality information, we augmented the taxonomic knowledge base with two kinds of information. Both involved four members of a common animal family ("birds"), with four features. The *typical* member had the same four features. The *typical plus* member had the same four features plus an additional two not shared by other members. The *typical minus* had only two of the features, and no additional features. The final *atypical* member had two shared features, and two unique features. A second knowledge base had the same exemplars and features, plus explanations for each feature.

ERIC computed confidence in the induction of a blank property from each premise bird to the conclusion bird. Figure 3 displays the results. Generally, as with people (Heit, 2000), increased typicality led to higher inductive confidence. One interesting exception to this pattern was that in the features only case, inductions were slightly stronger from the premise category with relatively few features than from the premise category with many typical categories. This is because this "unknown" category was exceptionally projectable, due to having very few features. When more explanations were available, the relatively high number of good potential explanations for the typical category dominated, leading to strong inductions.

Causal Knowledge

Because ERIC extends its knowledge based on the overall analogical quality, the predicate attributed to a premise and conclusion category can also strongly impact induction, if facts involving that premise or a related one are part of prior knowledge. A predicate similar to those that appear as part of good, projectable explanations about similar categories sets will generally form strong inductions; projectable predicates known to apply to very different creatures, or those about which little is known, tend to project less well.

We illustrated this property by creating knowledge corresponding to the taxonomic and predatory structures explored by Shafto et al (2008). For a set of seven animals, predation and taxonomic facts were encoded in memory. Two generic explanations involved a "disease" spread by predation, and an "organ" shared by animals sharing a taxonomic category. Inductions were generated for each creature regarding a different "disease" and "bone."

As illustrated in Figure 4, inductions on the bone graded taxonomically. Premises involving species with the same parent (distance 0) generalized more strongly than more distantly related species. Diseases also showed a taxonomic structure, but less strongly than bones did. Furthermore, the disease was strongly affected by ecological relationships, generating an asymmetry such that predators were judged more likely to get diseases carried by their prey than were prey whose predator was known to catch the disease.

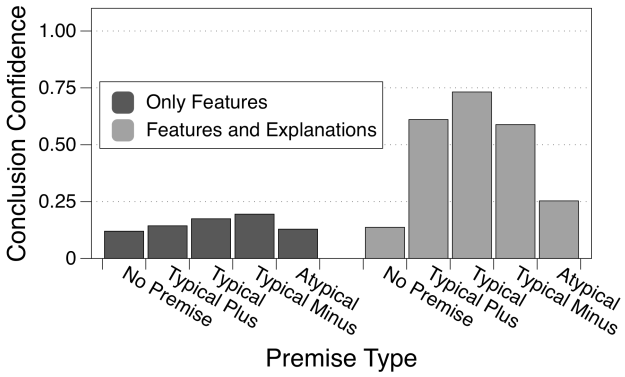


Figure 3: ERIC's predictions of induction strength, varying the typicality of the premise category.

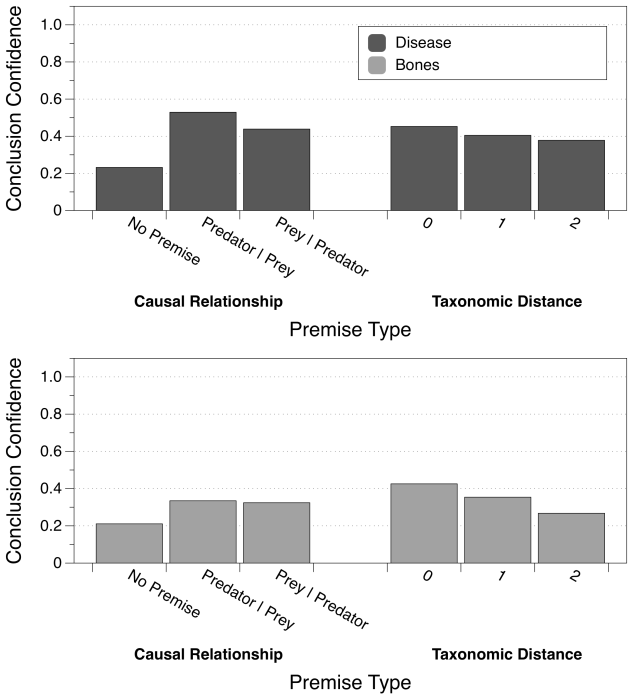


Figure 4: Dependency of inductive strength on both property and category relationships.

The latter still formed a strong induction in the disease case, because a prey carrying a disease made a good explanation for why a predator would have it; this explanation was thus well-supported during the premise explanation phase of ERIC's reasoning process. These patterns are quite similar to human judgments (Shafto et al., 2008), and demonstrate ERIC's ability to adjust the application of "rules" to different areas of knowledge.

Both properties showed taxonomic degradation. This is because both kinds of knowledge are in the system, and so both affect, to some degree, the same judgments. The model predicts that people will also blend different theories and domains of knowledge when making inductions.

Conclusions

ERIC combines deductive probabilistic inference with inductive analogical inference to generate and evaluate the likelihood of explanations, the propositions they comprise and the observations they explain. The resulting model, still in an early stage of development, successfully predicts and explains a wide range of phenomena in the property induction literature. Much work remains to be done (e.g., representing probabilities more realistically, allowing explanations to decrease as well as increase confidence, and making the generation of analogical explanations psychologically plausible rather than computationally exhaustive, among many others), but at this point ERIC seems a promising way to overcome the limitations of purely analogical, and purely Bayesian approaches to explanation generation and evaluation.

Acknowledgments

This research was funded by AFOSR grant # FA9550-07-1-0147. Thanks to Eric Taylor, Brian Ross, and Derek Devnich for thoughts and comments during the development of ERIC.

References

Falkenhainer, B., Forbus, K. D., & Gentner, D. (1989). The structure-mapping engine: algorithm and examples, *Artificial Intelligence, 41*, 1-63.

Gentner, D., & Medina, J. (1998). Similarity and the development of rules. *Cognition*, *65*(2-3), 263–297.

Heit, E. (2000). Properties of inductive reasoning. *Psychonomic Bulletin and Review*, *7*(4), 569–592.

Holyoak, K. J., & Thagard, P. (1989). Analogical Mapping by Constraint Satisfaction. *Cognitive Science*, *13*, 295-355.

Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review*, *104*(3), 427–466.

Hummel, J. E., & Landy, D. (2009). From analogy to explanation: Relaxing the 1:1 mapping constraint...Very carefully. In *New Frontiers in Analogy Research: Proceedings of the Second International Conference on Analogy*. Sofia, Bulgaria.

Kemp, C., & Tenenbaum, J. B. (2009). Structured Statistical Models of Inductive Reasoning. *Psychological Review*, *116*(1), 20-58.

Landy, D., & Hummel, J. E. (2009). Explanatory reasoning for inductive confidence. In *New Frontiers in Analogy Research: Proceedings of the Second International Conference on Analogy*. Sofia, Bulgaria.

Medin, D. L., Coley, J. D., Storms, G., & Hayes, B. K. (2003). A relevance theory of induction. *Psychonomic Bulletin & Review*, *10*(3), 517.

Osherson, D. N., Smith, E. E., Wilkie, O., López, A., & Shafir, E. (1990). Category-based induction. *Psychological Review*, *97*, 185-200.

Rips, L. J. (1975). Inductive judgments about natural categories. *Journal of verbal learning and verbal behavior*, *14*, 665-681.

Shafto, P., Kemp, C., Bonawitz, E. B., Coley, J. D., & Tenenbaum, J. B. (2008). Inductive reasoning about causally transmitted properties. *Cognition*, *109*(2), 175–192.

Sloman, S. A. (1993). Feature-based induction. *Cognitive Psychology*, *25*, 231-280.

Sloman, S. A. (1998). Categorical inference is not a tree: The myth of inheritance hierarchies. *Cognitive Psychology*, *35*(1), 1–33.

Sun, R. (2006). Accounting for a variety of reasoning data within a cognitive architecture. *Journal of Experimental and Theoretical Artificial Intelligence*, *18*(2), 169-191.

Wang, P. (2009). Formalization of Evidence: A Comparative Study. *Journal of Artificial General Intelligence*, *1*, 25–53.

Explanations make inconsistencies harder to detect

Sangeet Khemlani and P.N. Johnson-Laird
{khemlani, phil}@princeton.edu
Department of Psychology
Princeton University
Princeton, NJ 08540 USA

Abstract

What role do explanations play in reasoning about inconsistencies? We postulate that when people create explanations, they use them to resolve conflicting information. This hypothesis predicts that inconsistencies should be harder to detect once individuals have in mind an explanation of the inconsistency. We report four experiments that tested this prediction. Experiments 1a and 1b corroborated the effect when participants made inferences from inconsistent assertions. Experiment 2 compared the effect of explanations of inconsistencies with those of a similarly demanding task. Experiment 3 ruled out a potential confound.

Keywords: inconsistency, explanations, belief revision, reasoning, principle of resolution

Introduction

The word 'why' is used to elicit explanations for the mysteries of daily life. Why is my car making that noise? Why didn't the Redskins win last Sunday? Why isn't my experiment working? Indeed, a central feature of human rationality is the ability to construct explanations of observed behaviors and phenomena (Harman, 1965). Recent research has explored the function and developmental trajectory of explanatory reasoning (Keil, 2006; Wellman, Hickling, & Schult, 1997). There is consensus among researchers that explanations are related to causal inference (Johnson-Laird, Girotto, & Legrenzi, 2004; Sloman, 2005; Walsh & Johnson-Laird, 2009), and that explanations impact reasoning, categorization, and learning (Lombrozo, 2006). Less is known about the contexts under which explanations are generated, i.e., it is unclear when and how individuals decide to produce explanations.

How do you reveal what a person understands about some subject matter? One way is to ask the person to explain it, because explanations require individuals to communicate their knowledge and beliefs about the phenomenon in question. Explanations can also occur in other tasks that draw upon general knowledge. For instance, explanations are useful when you are learning new information (Amsterlaw & Wellman, 2006; Chi, De Leeuw, Chiu, & Lavancher, 1994; Crowley & Siegler, 1999; Rittle-Johnson, 2006), and they help to predict future behaviors (Anderson & Ross, 1980; Einhorn & Hogarth, 1986; Lombrozo & Carey, 2006; Ross, Lepper, Strack, & Steinmetz, 1977). Individuals spontaneously produce explanations when they try to form categories (Shafto & Coley, 2003) and when they judge how well concepts cohere with one another (Murphy & Medin, 1985; Palatano, Chin-Parker, & Ross, 2006). We propose that an additional function of explanatory reasoning is to resolve inconsistencies.

Explanations resolve inconsistencies

Consider the following:

If people are tired then they go to sleep.
A person was tired, but he did not go to sleep.

The two assertions are inconsistent, i.e., they cannot both be true. Given such an inconsistency, it is felicitous to ask: "why not?" But the same question is infelicitous when the assertions are obviously consistent with one another:

If people are tired then they go to sleep.
A person was not tired, and he did not go to sleep.

It seems strange to elicit an explanation for consistent assertions, and reasoners are likely to balk at such a request. Thus, an inconsistency calls for people to search for explanations, while an explanation is less appropriate when expectations are met.

We hypothesize that individuals resolve a set of inconsistent causal assertions by using an explanation to interpret each assertion, a view we call the *principle of resolution*. The principle assumes that when an inconsistency is detected among a set of assertions, reasoners construct explanations to restore consistency to the set (Johnson-Laird et al., 2004). They then interpret the assertions based on the consequences of the explanations. Consider the inconsistency above. One explanation for the person not going to sleep is that he was under some deadline, and so pursued his work despite his fatigue. The explanation provides an exception to the generalization that if people are tired they go to sleep. However, instead of abandoning it, reasoners are likely to construe it as an idealization that holds by default: it is true in many cases, but tolerates exceptions. The assertion may be interpreted as something akin to the generic assertion, i.e., 'people who are tired go to sleep' (Khemlani, Leslie, Glucksberg, & Rubio-Fernandez, 2007; Leslie, 2008). The principle of resolution thus allows individuals to use explanations to resolve inconsistencies by weakening the initial interpretation to that of an idealization rather than a universal truth.

One potential side effect of the principle is that when reasoners have an explanation of an inconsistency in mind, they may overlook the inconsistency on subsequent assessments of the assertions. If they interpret the conditional as an idealization, their new interpretation may

prevent them from detecting the conflict between the two assertions. Indeed, they may even forget that the reason for constructing the explanation in the first place was to resolve an inconsistency. To test this prediction, participants in four experiments were asked to detect an inconsistency after they had carried out various tasks.

Experiments 1a and 1b

Experiments 1a and 1b examined whether reasoners spontaneously construct explanations when faced with inconsistent scenarios, and whether those explanations made it more difficult to detect inconsistencies. They were presented with problems such as:

If a person is bitten by a viper then the person dies.

Someone was bitten by a viper, but did not die.

The participants in Experiment 1a evaluated the consistency of two assertions, either before or after they stated what follows from the assertions. The participants in Experiment 1b evaluated the consistency of the assertions before or after they responded to the question, "why not?" When individuals make an inference from inconsistent assertions, they should tend to infer explanations. The principle of resolution posits that when people create explanations, they interpret the assertions in the light of their explanation. It predicts an interaction: when individuals create an explanation first, they should be less accurate subsequently at detecting inconsistencies in comparison with those who have not created an explanation.

Method

Participants. 36 participants were recruited for Experiment 1a, and 40 participants were recruited for Experiment 1b. They volunteered through an online platform hosted through Amazon.com, and they completed the study for monetary compensation. None of the participants had received any training in logic.

Design and Procedure. On each trial, participants were given a set of consistent or inconsistent assertions (see Appendix A). Half of the problems presented a generalization (1) that was inconsistent with a categorical assertion (2), e.g.,

1. *If someone is very kind then he or she is liked by others.*
2. *Someone was very kind but was not liked by others.*

For the remaining problems, the inconsistency was eliminated by dropping the first clause in the categorical assertion (4), e.g.,

3. *If someone is very kind then he or she is liked by others.*
4. *Someone was not liked by others.*

Participants received an equal number of consistent and inconsistent problems, and carried out two tasks in succession for each problem, a consistency task and a task designed to elicit explanations. For the consistency task, participants had to answer the question, "Can both of these statements be true at the same time?" They responded by pressing one of two buttons marked "Yes" or "No". In Experiment 1a, participants also performed an inferential task, i.e., they answered the question, "What, if anything, follows from the statements above?" In Experiment 1b, they performed a more orthodox explanation task, i.e., they answered the question, "Why not?" They typed their responses into a text box provided on the screen. They were unable to see their response to the first task when they carried out the second task. In Experiment 1a, 20 participants performed the inferential task before the consistency task, and 16 participants performed the two tasks in the opposite order. In Experiment 1b, 20 participants performed the explanation task before the consistency task, and 20 performed the two tasks in the opposite order. All of the problems were similar to the two examples above, and participants received each set of contents only once. Each participant received the problems in a different random order.

Results and Discussion

Table 1 reports the proportions of trials on which participants correctly evaluated the assertions as consistent or inconsistent in Experiment 1a. Overall, participants were more accurate on consistent problems than inconsistent problems (77% vs. 50%, Wilcoxon test, $z = 3.27$, $p < .005$, Cliff's $d = .42$), and the group that carried out the consistency task first was marginally more accurate than the group that initially made an inference about the assertions (70% vs. 58%, Mann-Whitney test, $z = 1.66$, $p = .10$, Cliff's $d = .32$). These main effects were a consequence of the low rate of accuracy on inconsistent problems observed for the group that carried out the inferential task first. Their responses corroborated the principle of resolution, and the predicted interaction was significant: the group that initially carried out the inferential task was less accurate at detecting inconsistencies than consistencies, while the group that initially carried out the consistency task was just as accurate at detecting either type of problem (Mann-Whitney test, $z = 3.03$, $p < .005$, Cliff's $d = .59$). Accuracy in the evaluation of consistency in Experiment 1a therefore depended on whether or not participants initially made an inference about the assertions. The effect is likely to reflect the use of inferences that explain the inconsistency.

Table 1: The percentages of correct evaluations of consistency in Experiment 1a depending on whether participants carried out the evaluation or the inferential task first.

	Inconsistent problems	Consistent problems
Group that carried out the consistency task first	73	68
Group that carried out the inferential task first	33	84

Table 2 reports the proportions of correct responses in Experiment 1b. Participants were far more accurate at detecting consistencies than inconsistencies (89% vs. 45%, Wilcoxon test, $z = 4.00$, $p < .0001$, Cliff's $d = .69$). The group that initially evaluated the consistency of the assertions was more accurate than the group that initially provided an explanation (79% vs. 56%, Mann-Whitney test, $z = 3.07$, $p < .005$, Cliff's $d = .66$). And the predicted interaction was significant: the difference between accuracies on inconsistent vs. consistent problems was greater for the group that carried out the explanatory task first (Mann-Whitney test, $z = 2.02$, $p < .025$, Cliff's $d = .48$). As in Experiment 1a, participants in Experiment 1b were less accurate at detecting inconsistencies when they initially provided an explanation.

These results support the principle of resolution, which predicted that explanations would make it more difficult to detect inconsistencies. However, it is possible that the difficulty to detect inconsistencies could have occurred because the explanation and inferential tasks were inherently more difficult. In other words, there may not have been anything unique about the explanation task, and the same effects could have been observed had reasoners performed any task that increased processing load. The evidence for such an account is mixed: in Experiment 1a, participants who initially made an inference were more accurate at detecting consistencies than participants who initially carried out the consistency task (84% vs. 68%).

Table 2: The percentages of correct evaluations of consistency in Experiment 1b depending on whether participants carried out the evaluation or the explanation task first.

	Inconsistent problems	Consistent problems
Group that carried out the consistency task first	64	93
Group that carried out the explanation task first	27	86

Hence, a difference in processing load cannot readily explain this pattern of results. It should have decreased performance on both sorts of problem, but in fact the participants did better on the consistent problems. In contrast, a difference in processing load could explain the results of Experiment 1b, because in this case the participants who answered the question 'why not?' first, went on to evaluate the consistency of both sorts of problem worse than those participants who began with this evaluation task. Experiment 2 therefore sought to determine whether any demanding task could dull reasoners' sensitivity to inconsistencies, or whether explanations are unique in decreasing accuracy.

Experiment 2

To test whether explanations uniquely contribute to low rates of accuracy when individuals have to detect inconsistencies, the participants in this experiment evaluated the consistency of a set of assertions after carrying out one of two tasks: one group provided an explanation of the assertions and the other group decided whether some clauses of the assertions were more surprising than others. The surprisingness task was chosen because it required reasoners to take into account all the assertions, but it did not require them to construct explanations of inconsistencies. Those participants who performed the surprisingness task received trials such as the following one:

If the aperture on a camera is narrowed, then less light falls on the film

The aperture on this camera was narrowed but less light did not fall on the film

In light of these statements, which of the following is more surprising?

1. It's more surprising that the aperture on this camera was narrowed.

2. It's more surprising that less light did not fall on the film.

They received the same instructions for consistent trials, and responded by choosing between one of two alternative responses. Once their responses were registered, they carried out the consistency task. The other group of participants typed out their response to the question "Why not?" before completing the consistency task.

Method

Participants. 40 participants from the same online platform as in the previous studies and completed the experiment for monetary compensation.

Design and Procedure. Participants received an equal number of consistent and inconsistent problems, and received the same set of problems used in the previous study. Half the participants carried out the explanation task before the consistency task and the other half carried out the surprisingness task before the consistency task. They were unable to see their responses to the initial task when they carried out the consistency task. Participants received each set of contents only once, and each participant received the problems in a different randomized order.

Results and Discussion

Table 3 reports the proportions of correct responses in Experiment 2. The results again corroborated the principle of resolution. Participants were less accurate for inconsistent than consistent problems when they carried out the explanation task than when they carried out the surprisingness task (Mann-Whitney test, $z = 1.64$, $p = .05$, Cliff's $d = .30$). No decrease in accuracy was observed for consistent problems between the two groups (86% vs. 84%,

Mann-Whitney test, $z = .63$, $p = .53$). The results rule out the possibility that the effects reflected differences in processing load.

Table 3: The percentages of correct evaluations of consistency in Experiment 2 depending on whether participants carried out the surprisingness task first or the explanation task first.

	Inconsistent problems	Consistent problems
Group that carried out the surprisingness task first	75	86
Group that carried out the explanation task first	47	84

The experiment replicated the previous effect: participants who created explanations often went on to evaluate an inconsistent set of assertions as consistent, but the surprisingness task had no such effect. The study ruled out the possibility that any demanding mental task would yield the same results, because participants who rated how surprising the assertions were did not go on to err in their evaluation of the inconsistent problems. And both groups went on to evaluate consistent problems with no reliable difference in accuracy between them.

In Experiment 2, reasoners either carried out the surprisingness task or else the explanation task before judging the consistency of the assertions. That is, no participant was exposed to the two different task orders. Experiment 3 sought to extend the results to a context in which each participant carried out both tasks.

Experiment 3

Experiment 3 tested whether explanations impair evaluations of consistency more than judgments of surprisingness. On each trial, participants either provided an explanation, a judgment of surprisingness, or neither, before they evaluated the consistency of the assertions.

Method

Participants. 25 participants from the same online platform as in the previous studies completed the experiment for monetary compensation. None had received any training in logic.

Design and Procedure. Participants served as their own controls, and received an equal number of consistent and inconsistent problems. The materials consisted of those used in the previous studies. For a third of the trials, participants carried out only the consistency task; on another third, they carried out the surprisingness task before the consistency task; and on the remaining trials they carried out the explanation task before the consistency task. The three conditions were intermingled, and each participant received the problems in a different randomized order. Participants received each set of contents only once, and the contents were rotated over the three conditions so that each content occurred equally often in each condition in the experiment as a whole.

Results and Discussion

Table 4 provides the proportions of correct responses in Experiment 3. Participants were more accurate on consistent problems than inconsistent problems (71% vs. 52%, Wilcoxon test, $z = 2.38$, $p < .01$, Cliff's $d = .26$), and accuracy varied by the three types of trials (Friedman analysis of variance, $\chi^2 = 6.20$, $p < .05$). These main effects can be attributed to the drop in accuracy on inconsistent problems when participants had provided explanations.

The study yielded the predicted interaction between the type of trial and the consistency of the problem, i.e., participants were less accurate on inconsistent problems when they had carried out the explanation task than when they had carried out the surprisingness task or no prior task, whereas their accuracies for consistent problems were comparable to one another across the different tasks (Page's $L = 304.5$, $z = 2.55$, $p < .001$).

Table 4: The percentages of correct evaluations of consistency in Experiment 3 depending on whether participants carried out only the consistency task, the surprisingness task first, or the explanation task first.

	Inconsistent problems	Consistent problems
Consistency task only	60	70
Surprisingness task, then consistency task	56	76
Explanation task, then consistency task	40	68

As in the previous studies, Experiment 3 showed that explanations increased the likelihood that participants evaluated inconsistent assertions as consistent. The effect cannot be explained as a function of task demand, because participants did no better after they carried out the surprisingness task than after they had carried out no prior task. The study also extended the findings to a study in which the participants carried out all the different sorts of task. We conclude that the effect of explanations on consistency ratings is robust.

General Discussion

Across four experiments, participants erroneously evaluated inconsistent assertions as consistent after they had created an explanation for the inconsistency. Experiment 1a found that people produced the effect when they were asked to make inferences from the assertions, and Experiment 1b

extended the effect by directly eliciting explanations. Experiment 2 reproduced the effect by comparing those who formulated explanations with those who performed an unrelated task. Experiment 3 extended the effect to a context in which participants carried out the tasks in different orders. If participants had focused only on the assertions they were asked to read, the creation of an explanation should have had no effect on the evaluation of consistency in any of our experiments. Instead, the participants failed to detect inconsistencies as a result of creating explanations. When individuals resolve an inconsistency by explaining it, they are likely to establish a consistent interpretation of the facts of the matter and the original assertions. They have reasoned from inconsistency to consistency (see Johnson-Laird et al., 2004), and this newfound consistency makes it harder to detect the original inconsistency of the assertions.

Two gaps in the present account remain. First, the quality of the explanations that the participants created appeared to vary, but further research is need to interrelate this quality, say, to the latency of a correct evaluation of the inconsistent assertions. Second, the precise mechanism underlying the phenomenon has yet to be pinned down. When individuals explain an apparent inconsistency among a set of assertions, their explanation may sometimes rule out one of the assertions as false, and it may sometimes yield an idealized interpretation of a conditional generalization. For example, is the conditional assertion:

If a person is bitten by a viper then the person dies.

true or false? Given the further premise, say, that Viv was bitten by a viper, many people are likely to make the inference that Viv died. Yet, in answer to the preceding question, they might respond, "there are exceptions". In other words, the conditional expresses a truth that holds by default, i.e., a counterexample does not overturn it. In contrast, individuals are likely to judge that the conditional assertion:

If a person's brain is deprived of oxygen for 1 hour then the person dies.

is true unequivocally. And they might not be prepared to believe a description of an apparent counterexample.

The results of our experiments corroborate the principle of resolution, which states that when individuals detect an inconsistency, they formulate explanations to restore consistency. They subsequently can interpret the inconsistent assertions according to the consequences of their explanations. As a result they may treat conditional assertions as tolerating exceptions, which they can explain by invoking disabling conditions (Cummins, 1995). For example, consider the following problem:

If a person pulls the trigger then the pistol fires.

Someone pulled the trigger but the pistol did not fire.

If, like many of our participants, you explain the inconsistency by believing that there were no bullets in the pistol's chamber, then you have qualified the first assertion. It is true only when bullets are in the pistol's chamber, i.e., an enabling condition is satisfied. When bullets are not in the pistol's chamber, the conditional no longer hold (Johnson-Laird et al., 2004).

The present studies demonstrate the power and purpose of explanatory reasoning. Reasoners can draw inferences or answer the question 'why not?' without realizing that the set of assertions they reason about is inconsistent. The explanations they construct make it less likely that they will subsequently detect the inconsistency, because a plausible explanation serves to resolve the inconsistency. In some situations, this behavior is sensible and practical, because it allows individuals to revise their beliefs. In other situations, however, the behavior may account for striking lapses in reasoning. When a plausible explanation is available, regardless of whether it is true, reasoners may overlook glaring inconsistencies and behave in accordance with the explanation.

The present studies demonstrate the power and purpose of explanatory reasoning. Reasoners can draw inferences or answer the question 'why not?' without realizing that the set of assertions they reason about is inconsistent. The explanations they construct make it less likely that they will subsequently detect the inconsistency, because a plausible explanation serves to resolve the inconsistency. In some situations, this behavior is sensible and practical, because it allows individuals to revise their beliefs. In other situations, however, the behavior may account for striking lapses in reasoning. When a plausible explanation is available, regardless of whether it is true, reasoners may overlook glaring inconsistencies and behave in accordance with the explanation.

In sum, individuals who construct explanations of inconsistent assertions have difficulty evaluating those assertions as inconsistent. They do so erroneously, as the assertions remain in conflict with one another regardless of whether an explanation is available.

Acknowledgments

This research was supported by a National Science Foundation Graduate Research Fellowship to the first author, and by National Science Foundation Grant No. SES 0844851 to the second author to study deductive and probabilistic reasoning. We are grateful for helpful criticisms from Jeremy Boyd, John Darley, Sam Glucksberg, Adele Goldberg, Geoffrey Goodwin, Matt Johnson, Olivia Kang, Niklas Kunze, Max Lotstein, and Laura Suttle.

References

Amsterlaw, J., & Wellman, H.M. (2006). Theories of mind in transition: a microgenetic study of the development of false belief understanding. *Journal of Cognitive Development, 7,* 139-172.

Anderson, C.A., & Ross, L. (1980). Perseverence of social theories: The role of explanation in the persistence of discredited information. *Journal of Personality and Social Psychology, 39,* 1037-1049.

Byrne, R.M.J. (2005). *The rational imagination.* Cambridge, MA: MIT Press.

Chi, M., De Leeuw, N., Chiu, M.H., & Lavancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science, 18,* 439-477.

Crowley, K., & Siegler, R.S. (1999). Explanation and generalization in young children's strategy learning. *Child Development, 70,* 304-316.

Cummins, D.D. (1995). Naive theories and causal deduction. *Memory & Cognition, 23,* 646-658.

Einhorn, H. J., & Hogarth, R. M. (1986). Judging probable cause. *Psychological Bulletin, 99,* 3-19.

Harman, G. H. (1965). The inference to the best explanation. *Philosophical Review, 74,* 88-95.

Johnson-Laird, P.N. (2006). *How we reason.* Oxford: Oxford University Press.

Johnson-Laird, P. N., Girotto, V., & Legrenzi, P. (2004). Reasoning from inconsistency to consistency. *Psychological Review, 111,* 640-661.

Keil, F.C. (2006). Explanation and understanding. *Annual Review of Psychology, 57,* 225-254.

Khemlani, S., Leslie, S.J., Glucksberg, S., & Rubio-Fernandez, P. (2007). Do ducks lay eggs? How people interpret generic assertions. *In Proceedings of the 29th Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Leslie, S.J. (2008). Generics: Cognition and acquisition. *Philosophical Review, 117,* 1-47.

Lombrozo, T. (2006). The structure and function of explanations. *Trends in Cognitive Sciences, 10,* 464-470.

Lombrozo, T., & Carey, S. (2006). Functional explanation and the function of explanation. *Cognition, 99,* 167-204.

Murphy, G.L., & Medin, D.L. (1985). The role of theories in conceptual coherence. *Psychological Review, 92,* 289-316.

Patalano, A. L., Chin-Parker, S., & Ross, B. H. (2006). The importance of being coherent: Category coherence, cross-classification, and reasoning. *Journal of Memory and Language, 54,* 407-424.

Rittle-Johnson, B. (2006). Promoting transfer: the effects of direct instruction and self-explanation. *Child Development, 77,* 1-15.

Ross, L., Lepper, M.R., Strack, F., Steinmetz, J. (1977). Social explanation and social expectation: Effects of real and hypothetical explanations on subjective likelihood. *Journal of Personality and Social Psychology, 35,* 817-829.

Shafto, P., & Coley, J.D. (2003). Development of categorization and reasoning in the natural world: novices to experts, naïve similarity to ecological knowledge. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 29,* 641-649.

Sloman, S.A. (2005). *Causal models: how we think about the world and its alternatives.* New York: Oxford UP.

Walsh, C., & Johnson-Laird, P.N. (2009). Changing your mind. *Memory & Cognition, 37,* 624-631.

Wellman, H.M., Hickling, A.K., & Schult, C.A. (1997). Young children's psychological, physical, and biological explanations. *New Directions for Child Development, 75,* 7-25.

Appendix A

The assertions used in the experiments (generalizations were paired with consistent or inconsistent categorical assertions).

Domain	Generalization	Consistent Categorical	Inconsistent Categorical
Biology/physiology	If a person is bitten by a viper then they die	Someone did not die	Someone was bitten by a viper but did not die
Biology/physiology	If a person does regular aerobic exercises then that person strengthens his or her heart	Someone did not strengthen his heart	Someone did regular aerobic exercises but did not strengthen his or her heart
Mechanical	If a car's engine is tuned in the special way then its fuel consumption goes down	This car's fuel consumption did not go down	This car's engine was tuned in the special way but its fuel consumption did not go down
Mechanical	If graphite rods are inserted into a nuclear reactor, then its activity slows down	The nuclear reactor's activity did not slow down	Graphite rods were inserted into this nuclear reactor but its activity did not slow down
Mechanical	If the aperture on a camera is narrowed, then less light falls on the film	Less light did not fall on the film	The aperture on this camera was narrowed but less light did not fall on the film
Mechanical	If a person pulls the trigger then the pistol fires	The pistol did not fire	Someone pulled the trigger but the pistol did not fire
Natural	If a substance such as butter is heated then it melts	This piece of butter did not melt	This piece of butter was heated but it did not melt
Natural	If these two substances come into contact with one another then there is an explosion	There was no explosion	These two substances came into contact with one another but there was no explosion
Psychological	If someone is very kind then he or she is liked by others	Someone was not liked by others	Someone was very kind but was not liked by others
Psychological	If a person receives a heavy blow to the head then that person forgets some preceding events	Pat did not forget any preceding events	Pat received a heavy blow to the head but did not forget any preceding events
Social/economical	If people make too much noise at a party then the neighbors complain	The neighbors did not complain	People made too much noise at a party but the neighbors did not complain
Social/economical	If the banks cut interest rates then the economy increases	The economy did not increase	The banks cut interest rates but the economy did not increase

Why does explaining help learning? Insight from an explanation impairment effect

Joseph Jay Williams (joseph_williams@berkeley.edu)
Tania Lombrozo (lombrozo@berkeley.edu)
Department of Psychology, University of California, Berkeley

Bob Rehder (bob.rehder@nyu.edu)
Department of Psychology, New York University

Abstract

Engaging in explanation, even to oneself, can enhance learning. What underlies this effect? Williams & Lombrozo (in press) propose that explanation exerts *subsumptive constraints* on processing, driving learners to discover underlying patterns. A category-learning experiment demonstrates that explanation can enhance or impair learning depending on whether these constraints match the structure of the material being learned. Explaining can help learning when reliable patterns are present, but actually *impairs* learning when patterns are misleading. This *explanation impairment effect* is predicted by the subsumptive constraints account, but challenges alternative hypotheses according to which explaining helps learning by increasing task engagement through motivation, attention, or processing time. The findings have both theoretical and practical implications for learning and education.

Keywords: explanation; self-explanation; learning; constraints; impairment; category learning

Most teachers and tutors have had the experience of explaining a concept to another person and achieving greater understanding as a result. How does engaging in explanation generate this beneficial effect? This question's importance is underscored by the ubiquity of the phenomenon, and by converging evidence from cognitive science, education, and cognitive development confirming that explanation plays a significant role in learning.

Explanations have been implicated in theories of how conceptual knowledge is represented and how categories are learned (Carey, 1985; Gopnik & Meltzoff, 1997; Murphy & Medin, 1985). Education researchers have demonstrated that explaining has a potent effect on students' learning and fosters deep understanding that allows generalization to novel contexts (Chi et al, 1989; Chi et al, 1994). Research in cognitive development reveals even more profound effects (Wellman & Liu, 2006; Wellman, in press). Prompting children to explain can accelerate conceptual change, such as developing an understanding of number conservation (Siegler, 2002) and false belief (Amsterlaw & Wellman, 2006).

Many extant accounts of explanation's effects have emphasized the metacognitive benefits of explanation, such as prompting learners to identify and fill gaps in their knowledge (Chi et al, 1994; Chi, 2000). Explanations may also focus learners on uncovering the causes that underlie observed outcomes (Wellman & Liu, 2006), or may enhance learning by increasing task engagement in the form of additional motivation, attention, or processing time (for discussion see Siegler, 2002).

Williams and Lombrozo (in press) propose and find empirical support for the *subsumptive constraints* account, according to which explaining exerts constraints on processing that drive people to interpret what they are learning in terms of underlying patterns and regularities. The account is motivated in part by "subsumption" and "unification" theories of explanation from philosophy (Friedman, 1974; Kitcher, 1981), which propose that good explanations show how what is being explained is an instance of a unifying pattern: explanations cite generalizations that *subsume* what is being explained. If the explanations learners generate must satisfy this constraint, explaining will drive learners to reason and construct beliefs in the service of identifying patterns. When useful regularities exist, the subsumptive constraints account predicts positive effects of explanation through the discovery of generalizations. However, this account also predicts that seeking explanations can *impair* learning if there is a mismatch between the subsumptive constraints and the material being learned—for example, in situations in which patterns are nonexistent or misleading.

This paper tests the prediction that such an *explanation impairment effect* exists. Investigating the conditions under which explanation *hurts* learning can inform theories which aim to specify the mechanisms by which explanation *helps* learning, analogous to the study of visual illusions in perception. The conditions under which human perception or cognition succeeds can be less informative than those under which it breaks down and produces errors because the latter serve as a window onto the cognitive machinery underlying perception, in the case of visual illusions, or cognition, in the case of explanation and learning.

In fact, examining explanation's detrimental effects can discriminate the subsumptive constraints account from current theories, which to date have not predicted explanation impairment effects. In particular, a *task engagement* account advocates that engaging in explanation leads learners to be more engaged with the learning task, through increased motivation, attention, or time, which should benefit learning in virtually all contexts. The task engagement account provides an intuitive explanation for the beneficial effects of explaining, positing mechanisms that extend to contexts beyond explanation.

Some studies argue that explanation has effects that go beyond task engagement, showing that its effects surpass

control conditions that promote motivation, attention and processing time (e.g. Amsterlaw & Wellman, 2006; Chi et al, 1994; Williams & Lombrozo, in press). However, these studies cannot rule out the possibility that explaining simply engages these mechanisms to a greater degree than the control tasks, highlighting the difficulty of discriminating between competing accounts solely on the basis of explanation's beneficial effects.

Identifying explanation impairment effects is also of clear practical importance, as educators must know when prompted or spontaneous explanation will be detrimental (see also Kuhn & Katz, 2009). Moreover, a deeper understanding of the process by which explaining helps learning can inform educational interventions. If explaining simply boosts students' engagement with the task of learning or increases metacognitive awareness, then it can be expected to produce an 'all-purpose' benefit for learning. But if it helps through more specific mechanisms, such as constraining learners to find underlying principles, then it will be more helpful in some contexts than in others. Its effect may depend on the content being learned, learners' prior knowledge, and other factors.

As in previous work (Williams & Lombrozo, in press) , to investigate explanation our study utilizes category learning, which has been studied extensively and lends itself to carefully controlled artificial materials, permitting rigorous tests of competing accounts. Moreover, previous research supports the idea that explanation can and does play a role in category learning. When learners possess prior knowledge that explains why category features co-occur, they discover patterns underlying category membership and learn to classify items more quickly (Bott & Heit, 2000; Kaplan & Murphy, 2000; Murphy & Allopenna, 1994; Rehder & Ross, 2001; Wattenmaker, Dewey, Murphy, & Medin, 1986). There is also evidence that explanations influence the relative importance of features in learning novel categories (Lombrozo, 2009), and that explaining category membership can influence which features are used in categorization (Chin-Parker et al, 2006). Understanding how explaining influences category learning can thus shed light on the acquisition and representation of conceptual knowledge.

Experiment

Our category learning experiment tested the prediction that explanation can help or hinder learning, depending on the relationship between the material being explained and the subsumptive constraints imposed by explanation. Participants learned about two artificial categories of vehicles by classifying unlabeled items and then receiving feedback on their classification. After feedback and while studying the labeled item, participants in the *explain* condition were prompted to provide an explanation (out loud) for the item's category membership. In contrast, participants in the *classify* condition were free to use any study strategy and simply prompted to share what they were thinking out loud.

The category structures supported at least two bases for categorization, which are illustrated in Table 1 (materials adapted from Kaplan & Murphy, 2000). First, each of the 5 items in each category had a unique color feature. Remembering the 10 idiosyncratic color features always permitted accurate classification of all 10 items. Second, each item contained a feature that was associated with the unifying thematic pattern of jungle vehicles (e.g., drives in jungles, lightly insulated) or arctic vehicles (e.g., drives on glaciers, heavily insulated). In the *reliable pattern* condition, the theme could also be used to perfectly classify 10 out of 10 items based on the presence of an arctic or jungle vehicle feature. However, in the *misleading pattern* condition, the theme led to accurate classification for only 8 out of 10 items, and incorrect classification for the remaining 2 items. The experiment therefore used a 2 (study condition: *explain* vs. *classify*) x 2 (pattern type: *reliable* vs. *misleading*) design.

Dax	Kez
Theme Feature (1)	
Made in Norway	Made in Africa
Has Treads	Has Wheels
Heavily Insulated	Lightly Insulated
Used in Mountain Climbing	Used on Safaris
Drives on Glaciers	Drives in Jungles
Idiosyncratic Color Feature (1)	
Blue	Cyan
Silver	Magenta
Purple	Olive
Red	Maroon
Yellow	Lime
Irrelevant Features (3)	
Two doors/four doors	
Manual transmission/Automatic transmission	
Vinyl seats/Cloth seats	

Table 1. Features associated with each category. Each category item contained one theme feature, one idiosyncratic color feature, and three irrelevant features that were not diagnostic of category membership.

The subsumptive constraints account predicts that engaging in explanation should drive participants to discover and utilize the theme whether it is reliable or misleading, as the theme is more subsuming than the idiosyncratic color features. However, use of the theme should help learning when it is reliable but perpetuate classification errors when it is misleading, thereby *impairing* learning. In contrast, if explanation helps learning by boosting task engagement through increased motivation, attention, or processing time, it should produce a benefit regardless of pattern type.

Method

Participants There were 240 participants (60 in each of four conditions) from the UC Berkeley community who participated for monetary reimbursement or course credit.

Materials Each category was represented by five items, for a total of ten items. Each item was described by a list of five features (see Table 1): one *idiosyncratic color* feature (e.g. blue), one *theme*-related feature from either the arctic vehicle theme (e.g. heavily insulated) or the jungle vehicle theme (e.g. lightly insulated), and three *irrelevant* features that (a) occurred equally often in each category and so were not diagnostic and (b) were unrelated to the arctic/jungle themes (e.g. two doors). The pairing of theme and idiosyncratic color features was randomly chosen in each block of 10 items. The idiosyncratic color features were perfectly predictive of category membership (10 out of 10 items). The theme-related features were perfectly predictive (10 out of 10) in the *reliable* pattern condition, but predictive for only 8 out of 10 items in the *misleading* pattern condition. In each block, a different pair of theme features was randomly chosen to be misleading.

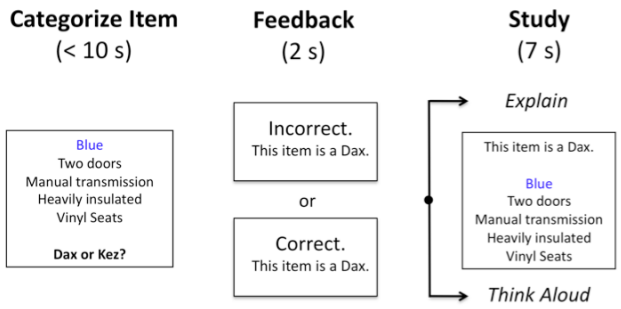


Figure 1: Structure of a single learning trial: item presentation and classification, feedback, and study.

Procedure The *reliable* and *misleading* conditions were run sequentially: data were first collected for 120 participants in the *misleading* pattern condition, then for 120 in the *reliable* condition. The *explain* and *classify* conditions were randomly interleaved, with participants randomly assigned to one or the other study condition. The experiment consisted of learning, test, and explicit report phases.

Learning phase. The structure of a learning trial is shown in Figure 1. On each learning trial an item description was presented as a list of five features, and participants had up to 10 seconds to categorize it as a "Dax" or a "Kez." The idiosyncratic color feature was always displayed on the first line in the color it named (e.g. the feature "red" was shown in red). All other features were presented below it in a random order, and shown in black. Feedback was provided after categorization, and the item was shown with the correct category label for 7 seconds. During this study period, participants in the *explain* condition were prompted (for example) to "Explain why this might be a Dax," and those in the *classify* condition were prompted with: "This item is a Dax. (Remember to say out loud whatever you are thinking.)" In both conditions participants spoke out loud to a voice recorder.

A random ordering of all 10 items constituted a block. Participants completed the experiment when they reached the learning criterion of correctly categorizing all 10 items in a single block, or the maximum of 15 blocks.

Classification test. Each of the 10 *idiosyncratic* and 10 *theme* features was individually presented onscreen and participants categorized it as belonging to a Dax or Kez, rated confidence in their decision (from 1 to 10), and how typical the feature was of its chosen category (1 to 7).[1] Idiosyncratic and theme features were presented in separate, randomly ordered blocks.

Ten *conflict* items were then presented in which an idiosyncratic feature was pitted against a theme feature. Features were paired so that using the idiosyncratic color features to categorize would generate an opposite response to using the theme features.[2]

Explicit report. At the end of the experiment participants were asked what differences might exist between categories and about their strategy for categorization; responses were typed onscreen.

Results

Learning measures, discovered differences between categories, and accuracy in the classification test are shown in Table 2. Significant differences between the *explain* and *classify* conditions are bolded.

	Reliable Pattern		Misleading Pattern	
Measures	Explain	Classify	Explain	Classify
Learning				
Perc. Reaching Criterion	93%	88%	**48%**	**75%**
Mean No. Blocks	6.9	7.9	**11.5**	**10.2**
Discovered differences between categories (from explicit reports)				
Theme Features	**62%**	**43%**	**28%**	**10%**
Color Features	**37%**	**57%**	**45%**	**70%**
Classification test accuracy				
Theme Features	**0.83**	**0.74**	**0.70**	**0.60**
Color Features	0.78	0.83	**0.81**	**0.89**
Conflict Items	**0.40**	**0.55**	**0.63**	**0.83**

Table 2. Measures of learning, discovered differences between categories, and classification test accuracy, as a function of *study condition* and *pattern type*. Significant differences between study conditions are bolded.

[1] These measures mirrored the results on classification accuracy and are not discussed further.

[2] After the classification test in the *reliable* condition, eight *transfer theme* features that were related to the arctic/jungle themes but had not been studied in the learning phase were presented for individual categorization and in *transfer conflict* items. Performance on these items was similar to those with studied theme features and are not discussed further.

Measures of learning. The mean number of blocks to reach the learning criterion is shown in Table 2, and frequency histograms in Figure 2, as a function of *study condition* and *pattern type*. A 2 (study condition*: explain* vs. *classify*) x 2 (pattern type: *reliable* vs. *misleading*) ANOVA on the number of blocks to learn revealed a significant interaction: the effects of explanation differed depending on whether the pattern was reliable or misleading, $F(1, 236) = 6.33$, $p < 0.05$.[3]

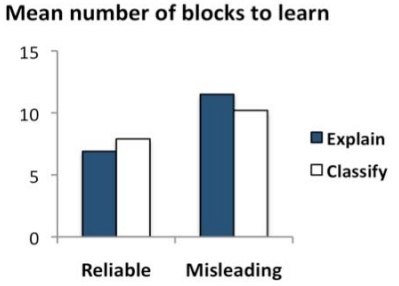


Figure 1: Mean number of blocks to reach the learning criterion of correctly categorizing all 10 items in a block, as a function of *study condition* and *pattern type*. Maximum number of blocks is 15.

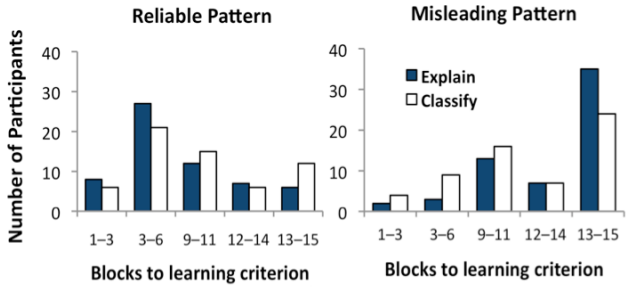


Figure 2: Frequency histogram of the number of blocks to reach learning criterion, as a function of *study condition* and *pattern type*. Bin size is three blocks.

The main effect of *pattern type* was also significant, $F(1, 236) = 44.49$, $p < 0.05$, suggesting that the *misleading* pattern slowed learning, although this interpretation should be qualified because participants were not randomly assigned to these two conditions. When the thematic pattern was reliable, there was a non-significant trend for the *explain* group to learn faster than the *classify* group, $t(118) = 1.43$, $p = 0.16$.[4] When it was misleading, the *explain* group took *longer* to learn, $t(118) = 2.11$, $p < 0.05$. In fact, the number of participants who learned how to classify (reached the learning criterion of correctly categorizing one block of 10 items) was lower in the *explain* condition than the *classify* condition, $\chi2\ (1) = 5.4$, $p < 0.05$. As predicted by the subsumptive constraints account, explanation's effects interacted with the structure of what was being learned, and actually *impaired* learning when a misleading pattern was present.

Discovered differences between categories. To test whether explaining exerted its effects through discovery of the theme, participants' explicit reports about the differences between categories and their categorization strategy were coded for mention of the theme-related and color features (see Fig. 3).[5] Participants in the *explain* condition more often reported theme features as a difference between categories than those in the *classify* condition, whether the pattern was reliable, $\chi2\ (1) = 4.04$, $p < 0.05$, or misleading, $\chi2\ (1) = 9.79$, $p < 0.05$. Participants in the *classify* condition more often reported color features (*reliable* pattern: $\chi2\ (1) = 4.82$, $p < 0.05$; *misleading* pattern: $\chi2\ (1) = 4.48$, $p < 0.05$). Explaining increased learning of theme-related category differences and decreased learning of theme-unrelated (color) features.

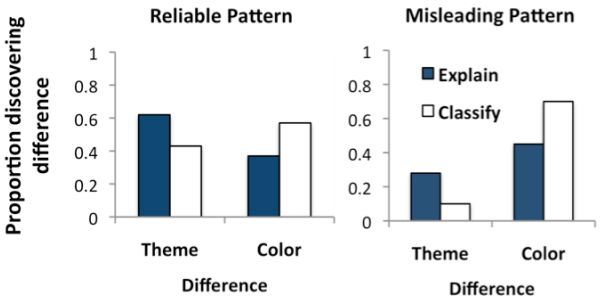


Figure 3: Proportion of participants whose explicit reports revealed discovery that theme and color features differed across categories, as a function of *study condition* and *pattern type*.

Classification test results. Accuracy in classifying theme and color features presented in Figure 4 shows that the *explain* and *classify* groups' different knowledge of theme versus color features also manifested itself in categorization performance. A 2 (study condition: *explain* vs. *classify*) x 2 (feature type: *theme* vs. *color*) repeated measures ANOVA on accuracy revealed a significant interaction for both the *reliable*, $F(1, 118) = 3.96$, $p < 0.05$, and *misleading*, $F(1, 118) = 9.85$, $p < 0.05$, conditions. Participants who explained learned which category the theme features were associated with better than those who classified, with the reverse pattern for color features.

The *conflict* items pitted an idiosyncratic color feature against a theme feature in a categorization decision, and the proportion of items categorized in accordance with the color features was defined as the *conflict score*. This measure was

[3] To address concerns about non-normality, we sorted the number of blocks to learning into five bins of three blocks (as in the histogram in Figure 1) and performed an ordinal regression with *study condition* and *pattern type* as factors. This analysis also found a significant interaction.

[4] To address concerns about non-normality, all *t*-tests reported in this paper were checked with non-parametric Mann-Whitney U tests, which generated the same conclusions.

[5] Agreement between two independent coders was 84% and reported results are for the first coder.

larger for the *classify* condition than the *explain* condition, whether the pattern was reliable, $t(118) = 2.00$, $p < 0.05$, or misleading, $t(118) = 3.42$, $p < 0.05$.

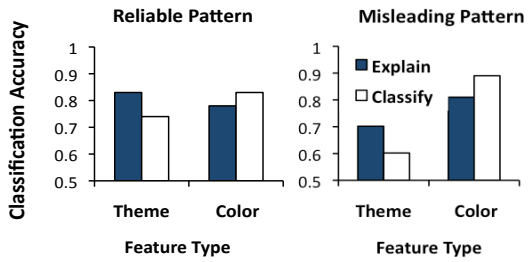


Figure 4: Accuracy in classifying theme and color features, as a function of *study condition* and *pattern type*.

Discussion

While most past research has documented explanation's positive effects, we found an explanation impairment effect: when a misleading pattern was present, explaining category membership impaired learning to categorize. Although counterintuitive, this impairment confirms a prediction of the subsumptive constraints account (Williams & Lombrozo, in press), according to which explaining exerts constraints that drive learners to interpret what they are learning in terms of underlying patterns.

The experiment provides evidence for an interaction between the subsumptive constraints exerted by explanation and the structure of the category. When compared to merely thinking aloud during study, explaining category membership further drove participants to rely on a unifying thematic pattern in categorization rather than use idiosyncratic features, even though both conditions engaged in the demanding task of classification learning with feedback. This produced (nonsignificant) positive consequences for learning when the thematic pattern was reliable, and (significant) negative consequences when it was misleading. Our explanation impairment effect provides evidence against a task engagement account of why explaining helps learning: if explaining merely increases motivation, attention, or processing time, it should not have impaired learning when the pattern was misleading.

A critical reader might have the intuition that the results of this experiment are unsurprising: prompting participants to explain tells them to find a pattern, which helps or harms learning depending on its existence. However, no previous account of explanation and learning has explicitly proposed that explaining constrains people to find patterns or predicted an impairment, instead focusing on metacognitive monitoring, identifying gaps in knowledge, or motivation and attention. A prompt to explain could have made participants attend more to their errors, justify individual categorizations by appeal to the salient and objective color features, or increased motivation to find a reliable basis for categorization. The subsumptive constraints account motivates our specific design and accounts for *why* people feel compelled to seek underlying patterns in response to explanation prompts.

Another criticism could be that this impairment effect is an artifact of an artificial lab task involving a "misleading" theme. However, our goal was precisely to characterize the conditions under which explanation's subsumptive constraints are detrimental. The finding that eliciting explanations impairs learning in any context is novel and consequential for current theories. Moreover, real-world cases involving misleading regularities and suspicious coincidences abound (Griffiths & Tenenbaum, 2007) and provide a promising direction for examining this effect outside the lab. It should be noted that *deeper processing* was not considered under the umbrella of task engagement. We do not see deeper processing as a specific competing account (like motivation) because we interpret the subsumptive constraints account as a specific proposal about the *nature* of the deeper processing explanation evokes.

Evidence for the subsumptive constraints account over the task engagement account has potential implications for education. If explaining does not merely produce an 'all-purpose' enhancement but exerts particular constraints on learning, more research is needed to understand the contexts in which self-explanation interventions are most effective and when they may be detrimental. First, one important question is how the explanation impairment effect varies with the quality of the explanatory pattern, that is, how misleading it is. In our misleading condition, the themes were misleading but only *partially* so: Classifying on the basis of theme features alone could result in moderately good accuracy (80%). The size of the learning impairment may have increased if the themes were even more misleading, but it is equally plausible that it would have *decreased* because subjects might choose to discard use of an explanatory pattern that is yielding poor performance (Murphy & Kaplan, 2000). The extent to which explaining may encourage learners to perseverate on a very low quality explanatory pattern remains to be determined.

Second, it is also important to assess the benefits of explanation *relative* to alternative learning activities, such as elaboration, direct instruction, or analogical comparison, and to examine how their complementary strengths and limitations can be combined. Williams and Lombrozo (in press) found that explaining drove discovery of underlying patterns but resulted in *worse* memory for details than describing. This is problematic because elaborating information in memory and receiving direct instruction may be more valuable at an early stage of learning. The subsumptive constraints account suggests that explaining will not necessarily be useful throughout a study episode (as would be predicted if it promoted task engagement), but will have its strongest effects when learners have already acquired factual background knowledge and need to discover and understand principles that underlie these facts. Successful demonstrations of the self-explanation effect may involve precisely such cases.

Other interesting directions for future research include the role of explanation in generating beliefs about both correct and misconceived underlying principles, in the effects of anomalies in belief revision (Chi, 2000), and in the deployment of prior knowledge (Chi et al, 1994; Williams & Lombrozo, in press). Such research will also be practically important for avoiding classroom manifestations of explanation impairment effects. For example, Kuhn and Katz (2009) suggested that requests for explanations on one task led children to later justify their knowledge of causal relationships by explaining how the relationships could exist, rather than citing observed evidence.

This is the first experiment to examine category learning through classification and feedback with (and without) additional prompts to explain. The learning differences generated by explaining suggest that category learning may involve processes beyond those that reduce immediate classification error. Bott et al. (2007) report that people learned about a thematic pattern underlying category membership (the same used in this experiment) in the *absence* of classification errors – a surprising violation of the classic *blocking effect* – while in the current experiment explaining drove learning about this pattern *despite* classification errors. A deeper understanding of these and other learning phenomena may be gained by considering the contribution of both classification error *and* the construction of knowledge that satisfies the constraints of explanation, whether it is prompted or spontaneous. For example, participants' spontaneous explanations may shed light on how prior knowledge is deployed, and when category learning is driven by explicit rule use versus bottom-up exemplar-based processing that reduces classification error.

The current research emphasizes the importance of subsumptive constraints in explanation's effects on learning, and demonstrates the value of explanation impairment effects for identifying the mechanisms by which explaining enhances learning. We are beginning to explore the relationship between prior knowledge and explanation (Williams & Lombrozo, in press (b)) and expect further investigation, in category learning and other learning contexts, to reveal a complex interaction between the constraints imposed by explanation, prior knowledge, and the structure of what is being explained.

Acknowledgments

We thank the Concepts and Cognition Lab, and in particular Preeti Talwai, for providing feedback on this project, assistance with analyzing data, and preparation of this paper. This work was partially supported by the McDonnell Foundation Collaborative Initiative on Causal Learning, and JJW was supported by an NSERC post-graduate scholarship.

References

Amsterlaw, J., & Wellman, H. (2006). Theories of mind in transition: A microgenetic study of the development of false belief understanding. *Journal of Cognition and Development, 7*, 139-172.

Bott, L., Hoffman, A., & Murphy, G. L. (2007). Blocking in category learning. *Journal of Experimental Psychology: General, 136*, 685-699.

Chi, M. T. H., Bassok, M., Lewis, M., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science, 13*, 145-182.

Chi, M.T.H., de Leeuw, N., Chiu, M.H., LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science, 18*, 439-477.

Chi, M.T.H. (2000). Self-explaining expository texts: The dual processes of generating inferences and repairing mental models. In R. Glaser (Ed.), Advances in Instructional Psychology, Hillsdale, NJ: Lawrence Erlbaum Associates. 161-238.

Chin-Parker, S., Hernandez, O., & Matens, M. (2006). Explanation in category learning. In R. Sun & N. Miyake (Eds.), *Proceedings of the 28th annual conference of the cognitive science society* (pp. 1098–1103*)*. Mahwah, NJ: Erlbaum.

Heit, E. & Bott, L. (2000). Knowledge selection in category learning. In D. Medin (Ed.), *Psychology of Learning and Motivation, 39*, 163-199. Academic Press.

Kaplan, A. S., & Murphy, G. L. (2000). Category learning with minimal prior knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 829-846.

Kitcher, P. (1981). Explanatory Unification. *Philosophy of Science, 48*, 507-31.

Kuhn, D., & Katz, J. (2009). Are self-explanations always beneficial? *Journal of Experimental Child Psychology, 103*, 386–394.

Lombrozo, T. (2009). Explanation and categorization: How "why?" informs "what?". *Cognition, 110*, 248-253.

Murphy, G. L., & Allopenna, P. D. (1994). The locus of knowledge effects in concept learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 904-919.

Murphy, G. L., & Kaplan, A. S. (2000). Feature distribution and background knowledge in category learning. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 53A*, 962-982.

Rehder, B. & Ross, B.H. (2001). Abstract coherent concepts. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27, 1261-1275.

Siegler, R. S. (2002). Microgenetic studies of self-explanations. In N. Granott & J. Parziale (Eds.), *Microdevelopment: Transition processes in development and learning* (pp. 31-58). New York: Cambridge University.

Wattenmaker, W. D., Dewey, G. I., Murphy, T. D., & Medin, D. L. (1986). Linear separability and concept learning: Context, relational properties, and concept naturalness. *Cognitive Psychology, 18*, 158-194.

Wellman, H. M. (in press). Reinvigorating explanations for the study of early cognitive development. *Child Development Perspectives*.

Wellman, H. M., & Liu, D. (2007). Causal reasoning as informed by the early development of explanations. *Causal learning: psychology, philosophy, and computation*, 261–279.

Williams, J. J., & Lombrozo, T. (in press). The role of explanation in discovery and generalization: evidence from category learning. *Cognitive Science*.

Williams, J. J., & Lombrozo, T. (in press (b)). Explanation constrains learning, and prior knowledge constrains explanation. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Explanation Constrains Learning, and Prior Knowledge Constrains Explanation

Joseph Jay Williams (joseph_williams@berkeley.edu)
Tania Lombrozo (lombrozo@berkeley.edu)
Department of Psychology, University of California, Berkeley

Abstract

A great deal of research has demonstrated that learning is influenced by the learner's prior background knowledge (e.g. Murphy, 2002; Keil, 1990), but little is known about the processes by which prior knowledge is deployed. We explore the role of explanation in deploying prior knowledge by examining the joint effects of eliciting explanations and providing prior knowledge in a task where each should aid learning. Three hypotheses are considered: that explanation and prior knowledge have independent and additive effects on learning, that their joint effects on learning are subadditive, and that their effects are superadditive. A category learning experiment finds evidence for a superadditive effect: explaining drives the discovery of regularities, while prior knowledge constrains which regularities learners discover. This is consistent with an account of explanation's effects on learning proposed in Williams & Lombrozo (in press).

Keywords: explanation; self-explanation; learning; prior knowledge; constraints; generalization; category learning

What processes underlie the critical capacity to acquire information and generalize to future situations? The topic of learning is one with a long history in cognitive science and development, and with important practical applications to education. While much research in cognitive science has focused on mechanisms that are independent of the specific knowledge people possess about a domain, studies have repeatedly and reliably demonstrated that prior background knowledge has profound effects on learning. This work suggests that characterizing how prior knowledge influences learning is a key issue for theories of learning.

Effects of prior knowledge have been particularly well characterized in the context of category learning. Prior knowledge that relates the features of a category allows learners to discover an underlying thematic pattern and learn the category more quickly (e.g., Murphy & Allopenna, 1994), and prior knowledge can also influence the construction of features in a way that supports classification (Wisniewski & Medin, 1994). Most broadly, prior knowledge has been seen as helpful because it exerts constraints on the process of knowledge acquisition (Keil, 1990), such as reducing the set of hypotheses learners entertain (Tenenbaum, Griffiths & Kemp, 2006). Most proposed mechanisms for category learning – such as encoding of exemplars, prototype formation, and other associative learning mechanisms – do not capture effects of prior knowledge (see Murphy, 2002), although more recent computational models attempt to incorporate such effects (e.g., Rehder & Murphy, 2003; Tenenbaum et al, 2006).

One possibility is that generating *explanations* plays a role in the effects of prior knowledge on learning. In this paper we consider the relationship between eliciting explanations and effects of prior knowledge. Engaging in explanation during study has been shown to promote learning and generalization in a range of knowledge-rich domains, for both adults (e.g. Chi, et al, 1994) and young children (for a review see Wellman & Liu, 2006). The process of "self-explaining" may be effective in part because explaining integrates new information with prior knowledge (Chi et al, 1994).

Previous work on eliciting explanations has considered the role of prior knowledge in mediating learning gains, but with mixed results. Some studies find that eliciting explanations has the greatest benefit for learners with low levels of prior domain-knowledge (e.g., Renkl et al., 1998), and that self-explanation training may be more useful for learners with low domain knowledge (McNamara, 2004). Other studies have not found a relationship between pre-test performance and the magnitude of post-test gains (e.g. Chi & VanLehn, 1991; Chi et al., 1994; Rittle-Johnson, 2006), although there is suggestive evidence that learners with more background produce higher-quality self-explanations (Renkl, 1997; Best, Ozuru, & McNamara, 2004).

Williams and Lombrozo (in press) propose a *subsumptive constraints* account of the role of explanation in learning that suggests how explanation and prior knowledge might interact to guide learning. The subsumptive constraints account is inspired by theories of explanation in philosophy which propose that explanations show how what is being explained is an instance of (subsumed by) a general pattern. If the explanations learners generate must satisfy this constraint, then attempting to explain should drive learners to discover regularities and underlying principles that are present in the material being explained. In support of this proposal, Williams and Lombrozo (in press) found that participants who explained items' category membership were more likely to discover a subtle regularity underlying category membership than participants who described category items, thought aloud, or engaged in free study.

The subsumptive constraints account suggests two ways in which explanation and prior knowledge could interact. First, explanations could determine *which prior knowledge* is deployed. According to the subsumptive constraints account, learners should invoke beliefs that demonstrate how what is being explained can be subsumed under general patterns. Second, the account suggests that prior knowledge could provide a source of constraint on *which subsuming generalizations* are considered explanatory. Consider the task of learning about the categories "psychology lecturer" and "psychology student" from the limited observation of a single lecture. The underlying bases for the categories could be that a psychology student is seated while a psychology

lecturer is standing, but this generalization seems like an implausible basis – and a poor explanation – for category membership. Distinguishing law-like generalizations from accidental generalizations is notoriously difficult (for discussion in philosophy see Caroll, 2008; and in psychology, Kalish, 2002), but prior knowledge may provide one source of constraint on which patterns are seen as explanatory, therefore determining which patterns participants are more likely to discover and employ in seeking explanations.

To investigate the relationship between explanation and prior knowledge, we restrict our focus to cases where explanation and prior knowledge would be expected to help learning, and consider whether their joint effects on learning are independent and additive, subadditive (less than the sum of their independent effects), or superadditive (greater than the sum of their independent effects).[1]

The proposed experiment uses a category-learning task in which there are patterns underlying category membership, and an explanation manipulation (explain vs. free study) is crossed with a prior knowledge manipulation (knowledge relevant to an underlying pattern is provided vs. no additional knowledge). The experiment aims to discriminate three alternative hypotheses about the joint effects of explanation and prior knowledge on learning.

One possibility is that explanation and prior knowledge have independent and additive effects. This hypothesis is a sensible default in the absence of evidence that eliciting explanations and prior knowledge interact, and no specific accounts have been proposed as to how prior knowledge might be deployed through explaining. Independent effects of explanations and prior knowledge would be likely if explaining helps learning through mechanisms that do not interact with those by which prior knowledge plays a role. For example, explaining might increase attention and motivation, while prior knowledge might independently constrain the hypotheses under consideration.

A second possibility is that prior knowledge and explanation have subadditive benefits. This could occur if the effects of explanation and prior knowledge are achieved through common mechanisms. For example, prompts to explain and the provision of prior knowledge may both guide learners to seek meaningful regularities in category structure. Explaining when prior knowledge is already available may therefore have little benefit above simply possessing prior knowledge.

[1] Whether explanation and prior knowledge help or hurt learning depends on the nature of what is being learned. Prior beliefs about a domain may be incorrect, or explaining may drive learners to unreliable patterns (Williams & Lombrozo, in press; Williams, Lombrozo, & Rehder, in press). In this paper we do not aim to investigate interactions of explanation and prior knowledge in settings where either will individually impair learning. In many real-world cases and educational contexts, both explaining and prior knowledge would be expected to benefit learning – for example, if there are regularities to discover and prior knowledge is correct – and this is the kind of setting we explore.

A final possibility is a superadditive effect of explanation and prior knowledge, such that explanation and prior knowledge interact in a way that produces a learning benefit that exceeds either of their independent effects. This could occur if explanations deploy prior knowledge that might otherwise be inert, or if prior knowledge influences the generation of explanations in a way that fosters more effective learning. The subsumptive constraints account suggests one way this might work: attempting to generate explanations (e.g. for category membership) could invoke prior beliefs in order to supply candidate subsuming patterns, and prior beliefs could simultaneously constrain which candidate subsuming regularities are deemed explanatory.

Experiment

There are many ways that prior knowledge could impact learning, and accordingly a multitude of ways in which prior knowledge could be manipulated. In this experiment, we provide category labels intended to activate prior knowledge relevant to which features might underlie membership.

We used eight category items, shown in Figure 1. There were two rules that could be used to categorize: an *antenna rule* (shorter left vs shorter right antenna) and a *foot rule* (pointy vs flat feet). The *prior knowledge* variable was operationalized by providing uninformative category labels that were neutral with respect to the two rules (*low* prior knowledge condition: items labeled as Glorp and Drent robots) versus labels that could be related to the foot rule (*high* prior knowledge condition: labeled as Outdoor and Indoor robots). The motivation for these rules was that participants' knowledge might account for Outdoor robots having pointy fleet and Indoor robots having flat feet, but not for why Outdoor or Indoor robots would have shorter left or right antennae.[2]

While all participants were informed that they would later be tested on their ability to categorize robots, those in the *explain* condition were prompted to explain the category membership of the Glorp and Drent (or Indoor & Outdoor) robots, while those in the *free study* condition were allowed to study the robots without specific prompts, yielding a *task* variable with two levels (explain vs. free study).

The two (*Task*: Explanation vs. Free Study) x two (*Prior knowledge*: Low vs. High) design therefore allowed for a test of whether the joint effect of explanation and prior knowledge on learning a basis for categorization is independent and additive, subadditive, or superadditive.

[2] Participants could have drawn on prior knowledge to explain why antenna length was related to being Outdoor/Indoor, or have had beliefs that conflicted with, for example, Outdoor robots having pointy feet, but the significant difference between conditions suggests this was not true for the majority of participants.

Participants

Two hundred and forty (60 in each condition) UC Berkeley students participated for course credit or monetary reimbursement (161 in the lab, 79 online).

Materials

The task involved *study items*, *test items*, and *transfer items*.

Study items. There were two categories of alien robots; the image participants saw in the *high prior knowledge* condition is displayed in Figure 1. The category labels were chosen based on whether the condition was *low* or *high* prior knowledge: the robots were labeled as *Glorps* and *Drents i*n the low prior knowledge condition, and as *Indoor* and *Outdoor* robots in the high prior knowledge condition.

Each robot was composed of six elements: left color (blue, green, red, yellow), right color (brown, cyan, grey, pink), body shape (square, circular), left antenna length (short, long), right antenna length (short, long), and foot shape (eight different geometric shapes). Color and body shape were uncorrelated with category membership: every right and left color occurred exactly once per category, and each category had two robots with square bodies and two with circular bodies. All four Outdoor (Glorp) robots had a shorter left antenna and all four Indoor (Drent) robots had a shorter right antenna. Although each robot had a unique geometric shape for feet, there was a subtle regularity across categories: all four Outdoor (Glorp) robots had pointy feet while all four Indoor (Drent) robots had flat feet. For simplicity, from this point on we refer to the robots in each category by their high prior knowledge label (*Outdoor/Indoor* robots).

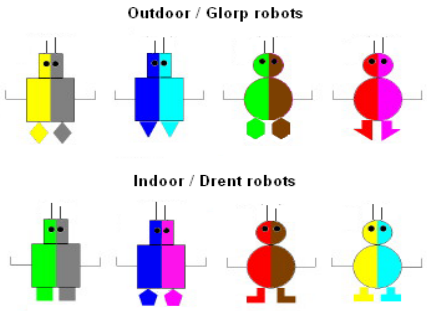


Figure 1: Study items.

This category structure supported at least three distinct bases for categorization. First, participants might not draw any generalizations about category membership, and instead categorize new items on the basis of their similarity to individual study items, where similarity is measured by tallying the number of shared features across items. We call this *item similarity*. Alternatively, participants could notice the antenna feature (Outdoor robots had shorter left antennas, Indoor robots shorter right antennas) and use it as a categorization rule: this is termed the *antenna rule*. Finally, participants could discover that although each robot had a unique geometric shape for feet, there was a subtle regularity termed the *foot rule*: Outdoor robots had pointy feet and Indoor robots had flat feet.

Test probe items. Three types of test item were constructed by taking novel combinations of the features used for the study items. Each type yielded a categorization judgment (of Outdoor/Indoor) that was diagnostic of one basis for categorization (item similarity, antenna rule, foot rule), by pitting that basis for categorization against the other two. For example, categorizing a yellow/gray robot with a shorter right antenna and pointy feet as an Indoor robot would suggest a participant relied on the antenna rule. We call these item similarity probes (three items), antenna rule probes (three items), and foot rule probes (four items). There was one extra item for which all three bases gave the same response.

Transfer Items. These four items used completely novel foot shapes to distinguish participants who genuinely drew an abstract generalization concerning "pointy" versus "flat" feet from those who simply recognized the importance of particular foot shapes. For each item, the foot rule was pitted against item similarity and the antenna rule.

Procedure

The task involved a study phase, a categorization phase, and additional measures designed to probe what participants had learned about the categories.

Study phase. Participants were instructed that they would be looking at two types of robots on the planet Zarn: Outdoor (Glorp) and Indoor (Drent) robots, with labels chosen based on being in the *high* or *low* prior knowledge condition. They were also informed that they would later be tested on their ability to remember the robots they had seen, and their ability to decide whether robots were Outdoor (Glorp) or Indoor (Drent) robots.

After advancing the instruction screen they saw a color image displaying the eight study items in a scrambled order, with each robot numbered 1 through 8 and category membership clearly indicated for each robot (the actual image for the high prior knowledge condition is shown in Figure 1). In both conditions participants were informed that they were seeing eight robots on ZARN and that the picture would be onscreen for two minutes. Participants in the *explain* condition were told "Explain why robots 1, 2, 3 & 4 might be Outdoor (Glorp) robots, and explain why robots 5, 6, 7 & 8 might be Indoor (Drent) robots."[3] Participants typed their explanations into a box onscreen. Those in the *free study* condition were told "Robots 1, 2, 3 & 4 are Outdoor robots, and robots 5, 6, 7 & 8 are Indoor robots." The image was onscreen for exactly two minutes and then the screen automatically advanced.

Categorization phase. The eleven test items were presented in random order, followed by the four transfer items in random order, with participants categorizing each robot as Outdoor (Glorp) or Indoor (Drent).

[3] In all quoted prompts, the alternative labels (Glorp/Drent instead of Outdoor/Indoor) are displayed in parentheses, but only one set of labels was actually displayed.

Probability of pattern. To assess participants' belief about the presence of a defining feature or rule, they were asked: "What do you think the chances are that there is one single feature that underlies whether a robot is Outdoor (Glorp) or Indoor (Drent) - a single feature that could be used to classify ALL robots?"

Category differences. Participants were explicitly asked "Were there any noticeable differences between Outdoor (Glorp) and Indoor (Drent) robots? If you think there were, please be SPECIFIC about what you thought the differences were."

Ranking of question informativeness.[4]

Features used for categorization. Participants were asked which features they used in categorizing robots. There was a separate line to enter features of Outdoor (Glorp) robots and features of Indoor (Drent) robots.[5]

Antenna Informativeness. Participants were asked if they could tell whether a robot was Outdoor (Glorp) or Indoor (Drent) by looking at its antenna, and if they could, to state what the difference was.

Antenna classification.[4]

Explanation self-report. All participants were asked if they were trying to explain the category membership of robots while the image of all 8 robots was onscreen.

Previous exposure. Participants were asked if they had seen the robots before, or already done an experiment using the materials.[6]

Foot informativeness. Participants were asked if they could tell what category a robot belonged to by looking at its feet, and if they could, to state what the difference was.

Results

In the interests of space, we do not report all dependent measures, especially as many support the same conclusions.

Each of the three kinds of test probe items pitted one basis for categorization against the other two, so participants' patterns of categorization over the full set was used to determine whether their basis for categorization was most consistent with 'item similarity', the 'antenna rule', or the 'foot rule', with ties coded as 'other'. The proportion of participants using each basis is shown in Table 1, as a function of condition. In addition to examining the basis participants' *used*, direct measures of *antenna rule discovery* and *foot rule discovery* were also coded from participants' responses to questions about whether they could classify robots based only on antenna or feet. These generally mirrored the findings on rule use. Figure 2 shows the proportion of participants who discovered the foot and antenna rules and Figure 3 shows the proportion that discovered *a* rule (antenna or foot), as a function of condition.

A log-linear analysis on *task* (explain vs. free study), *prior knowledge* (low vs. high), and *foot rule use* (used vs. did not use foot rule, as computed from inferred basis) revealed a significant three-way interaction, $\chi 2$ (1) = 7.27, $p < 0.01$, while that for *foot rule discovery* was marginal, $\chi 2$ (1) = 3.16, $p = 0.08$. Explanation and prior knowledge had a joint, superadditive effect on use of the foot rule. This interaction was driven by privileged use of the foot rule by participants who explained *and* had high prior knowledge (the explain-high PK condition): the combination of explaining and relevant prior knowledge exceeded the effects of each factor on its own. In fact, in the absence of explaining (i.e., the free study conditions) prior knowledge did not have an effect on foot rule use, $\chi 2$ (1) = 0.06, $p = 0.81$.

	Foot Rule	Antenna Rule	Item Similarity	Other
Explain- Low PK	0.32	0.60	0.05	0.03
Explain- High PK	0.67	0.25	0.06	0.02
Free Study- Low PK	0.35	0.22	0.38	0.05
Free Study- High PK	0.35	0.20	0.40	0.05

Table 1: Proportion of participants using each basis for categorization, by condition.

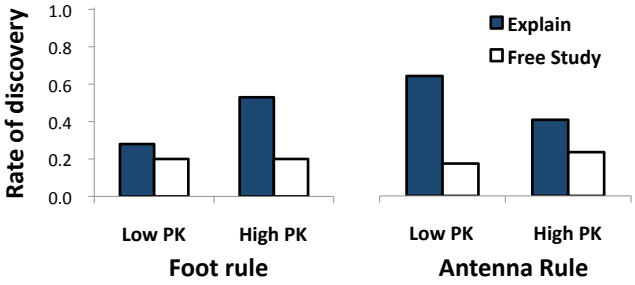


Figure 2: Proportion of participants who discovered the foot and antenna rules, by condition.

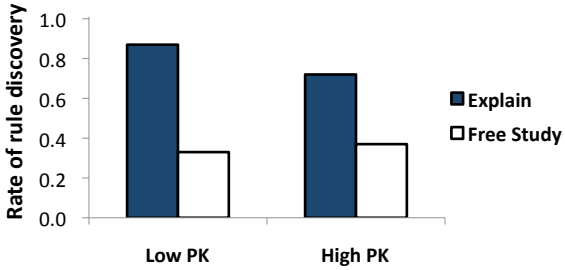


Figure 3: Proportion of participants who discovered a rule (antenna or foot), by condition.

There was also a three-way interaction between task, prior knowledge and both *antenna rule use*, $\chi 2$ (1) = 5.48, $p < 0.05$, and *antenna rule discovery*, $\chi 2$ (1) = 5.40, $p < 0.05$, driven by the explain-low PK condition. Overall, use of *a*

[4] This question asked participants to rank how informative different questions would be about membership, but is redundant with other reported measures and so omitted to save space.

[5] Some participants' categorization responses were reverse coded, if their explicit reports about the differences between categories or features used to categorize revealed they had reversed category labels, such as stating that outdoor robots had flat feet when in fact the opposite was true.

[6] Those who indicated previous participation were excluded.

rule (either antenna or foot) was higher for explainers (interaction between task and whether a rule was used, $\chi 2$ (1) = 42.76, $p < 0.001$, while reliance on *item similarity* was higher in the free study condition (interaction of task and item similarity use, $\chi 2$ (1) = 41.90, $p < 0.001$). Interestingly, overall rule discovery was actually higher in the explain-low PK than explain-high PK condition, $\chi 2$ (1) = 4.09, $p < 0.05$.

Discussion

In the context of category learning, we found that explanation and prior knowledge interacted, producing an effect on the discovery of a regularity related to prior knowledge that surpassed the independent effects of explanation or prior knowledge alone. This finding challenges the possibility that explaining and prior knowledge influence learning independently. Since a subadditive effect was not found, it also provides evidence against the hypothesis that explanation and prior knowledge draw on the same mechanisms or resources in promoting learning. The best explanation for the current findings is that explanation and prior knowledge influence learning by neither independent nor identical means, but have an interactive relationship.

This relationship can be understood in terms of the subsumptive constraints account of explanation and learning (Williams & Lombrozo, in press). If explaining exerts the constraint that learners generate explanations that show how what is being explained is subsumed by a general pattern, prior knowledge can provide constraints on which patterns support reasonable explanations. In the current experiment, explaining why items were Outdoor and Indoor robots drew on prior knowledge that constrained learners to explain membership in terms of the foot rule rather than a rule concerning antenna length. Not all subsuming patterns are equally explanatory; patterns must also make sense in light of prior knowledge.

An alternative account could instead implicate attentional mechanisms: Explaining promotes attention to items while prior knowledge exerts constraints on which item features are the focus of this attention, leading to an interactive effect on discovery of the foot rule. However, prior knowledge did not focus attention on the foot rule in the free study conditions. Moreover, Williams et al (in press) provide evidence that explaining can actually *impair* learning, suggesting that its effects go beyond increasing attention to exerting subsumptive constraints. If explaining influences attention, the evidence suggests it is not a generalized attentional boost to encode item details or monitor more information, but through constraints to attend to underlying patterns, which we would endorse as consistent with the subsumptive constraints account.

While we report a superadditive effect of explanation and prior knowledge, there are likely contexts in which different kinds of interactions would obtain. For example, it is known that the learning benefits of explanations (Williams et al, in press) and of prior knowledge (Wattenmaker et al, 1986) depend on the relationship between the constraints imposed by explanation or prior knowledge and the structure of the material being learned. If explanation exerts inappropriate constraints or prior knowledge is incorrect, their joint effects will be markedly different. Also, in cases where explanation automatically recruits prior knowledge or prior knowledge produces spontaneous explanation, their joint effect may appear to be independent or subadditive. The goal in the current work was to take a first and necessarily circumscribed step towards the ambitious goal of understanding the interactions between explanation and prior knowledge in learning.

Despite these limitations, the findings have implications for education and suggest interesting directions for applied research. Providing evidence that explaining invokes and is influenced by prior knowledge helps to explain why it has such powerful effects on learning. Explaining drives the discovery of regularities *and* guides learners to interpret what they are learning in terms of what they already know: an activity students may not engage in spontaneously even if they possess relevant prior knowledge.

If explaining promotes consistency with prior knowledge, its benefits may depend on having acquired correct and useful prior knowledge. Learning strategies that focus on acquiring background knowledge may be a necessary precursor to activities that involve explanation, and failures of explanation may suggest the need to develop background knowledge. The dangers inherent in incorrect prior knowledge are also brought into clear relief: effects of explaining may be reduced by incorrect or inappropriate prior knowledge, and may even be harmful. Examining the relationship between explanation and prior knowledge might therefore be one way to understand robust misconceptions and difficulties with conceptual change.

The current findings speak to the possibility that explanation is a mechanism by which prior knowledge is brought to bear in learning. In this experimental context, simply providing prior knowledge was insufficient to support learning: the high and low prior knowledge free study conditions did not differ in rule discovery. It may be that when learners explain and must satisfy subsumptive constraints, prior knowledge is accessed and deployed to inform which patterns are subsuming, so that explaining is a mechanism by which prior knowledge influences learning. Further research could explore what kinds of prior knowledge explaining might deploy, such as logical or causal inferences versus information stored in memory. Another issue concerns the amount of prior knowledge necessary for these interactive effects. The current experiments compared just two levels of prior knowledge, although prior knowledge spans a much broader continuum.

If explaining deploys prior knowledge in learning, it may be that spontaneously explaining category membership plays a role in knowledge effects on category learning. This possibility is bolstered by demonstrations that explaining increases use of features that are unified by prior knowledge into thematic patterns (Chin-Parker et al, 2006; Williams et

al, in press). Moreover, Wisniewski & Medin (1994) reported that activating prior knowledge through meaningful category labels drove the construction of novel and abstract features. The effects they report may in fact be best understood in terms of an interaction between prior knowledge and explanations for category membership, which the subsumptive constraints account can help explain.

Explanation's effects on category learning warrant an examination of the relationship between explanation-based learning and existing models of category learning. While the subsumptive constraints account aligns naturally with rule-based models (e.g. Nosofsky et al, 1994), the reported interaction shows how both our account and rule-based models need to be extended to account for effects of prior knowledge on *which* rules count as good bases for category membership. More broadly, while representations such as exemplars play one role in learning about a category, the effect of explanation may be to construct more abstract representations that are consistent with general prior knowledge about a category, such as its origin or function.

The current work suggests a number of future directions. Do different types of prior knowledge differentially support learning, such as prior knowledge about causal mechanisms vs. functions? When does prior knowledge help because it supplies candidate patterns that can subsume observations, versus help because it informs which patterns are subsuming? Given that subsumption and consistency with prior knowledge both constrain learning, how do they trade off? These and further questions await future research.

Acknowledgments

We thank Ania Jarosewicz and Preeti Talwai for collecting data and providing other assistance and feedback on this project. This work was partially supported by the McDonnell Foundation Collaborative Initiative on Causal Learning, and JJW was supported by an NSERC post-graduate scholarship.

References

Best, R., Ozuru, Y., & McNamara, D.S. (2004). Self-explaining science texts: strategies, knowledge, and reading skill. *Proceedings of the 6th international conference on learning sciences,* 89-96.

Carroll, J. W. (2008). Laws of nature, in Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy (Fall 2008 Edition)*, URL = <http://plato.stanford.edu/archives/fall2008/entries/laws-of-nature/>.

Chi, M.T.H., de Leeuw, N., Chiu, M.H., LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science, 18*, 439-477.

Chi, M., & VanLehn, K. (1991). The content of physics self-explanations. *Journal of the Learning Sciences, 1,* 69-105.

Chin-Parker, S., Hernandez, O., & Matens, M. (2006). Explanation in category learning. In R. Sun & N. Miyake (Eds.), *Proceedings of the 28th annual conference of the cognitive science society* (pp. 1098–1103*)*. Mahwah, NJ: Erlbaum.

Kalish, C. W. (2002). Gold, Jade, and Emeruby: The value of naturalness for theories of concepts and categories. *Journal of Theoretical and Philosophical Psychology, 22*, 45-56.

Keil, F. C. (1990). Constraints on constraints: Surveying the epigenetic landscape. *Cognitive Science*, *14*(1), 135–168.

McNamara, D.S. (2004). SERT: Self-explanation reading training. *Discourse Processes, 38*, 1-30.

Murphy, G. L. (2002). *The big book of concepts*. The MIT Press.

Murphy, G. L., & Allopenna, P. D. (1994). The locus of knowledge effects in concept learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 904-919.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, *101*, 53–53.

Rehder, B., & Murphy, G. L. (2003). A Knowledge-Resonance (KRES) model of category learning. Psychonomic Bulletin & Review, 10, 759-784.

Renkl, A. (1997). Learning from worked-out examples: a study of individual differences. *Cognitive Science, 21,* 1-29.

Renkl, A., Stark, R., Gruber, H., & Mandl, H. (1998). Learning from worked-out examples: the effects of example variability and elicited self-explanations. *Contemporary Educational Psychology, 23,* 90-108.

Rittle-Johnson, B. (2006). Promoting transfer: effects of self-explanation and direct instruction. *Child Development, 77*, 1-15.

Siegler, R. S. (2002). Microgenetic studies of self-explanations. In N. Granott & J. Parziale (Eds.), *Microdevelopment: Transition processes in development and learning* (pp. 31-58). New York: Cambridge University.

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, *10*, 309-318.

Wellman, H. M., & Liu, D. (2007). Causal reasoning as informed by the early development of explanations. *Causal learning: psychology, philosophy, and computation*, 261–279.

Williams, J. J., & Lombrozo, T. (in press). The role of explanation in discovery and generalization: evidence from category learning. *Cognitive Science*.

Williams, J. J., Lombrozo, T., & Rehder, B. (in press). Why does explaining help learning? Insight from an explanation impairment effect. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society

Wisniewski, E. J. & Medin, D. L. (1994). On the interaction of theory and data in concept learning. Cognitive Science, 18, 221-281.

Flux: Fundamental or Frivolous?

Lera Boroditsky (lera@stanford.edu)
Department of Psychology, 450 Serra Mall, Bldg 420
Stanford, CA, 94305-2130, USA

Helen Neville and Christina Karns (neville@uoregon.edu, ckarns@uoregon.edu)
Brain Development Lab, University of Oregon

Arthur B. Markman (markman@psy.utexas.edu)
Psychology, University of Texas at Austin

Michael J. Spivey (spivey@ucmerced.edu)
Cognitive Science, University of California at Merced

Symposium summary

A broad range of findings across the cognitive sciences has emerged revealing surprising flexibility and dynamic flux in a large range of cognitive domains. These include exciting new discoveries of neuroplasticity well into adulthood, discoveries of great cognitive variability as a function of the statistical properties of one's environment (from patterns in natural languages, to those in embodied experience), and discoveries of the surprisingly dynamic microstructure of cognition. Do such findings demonstrate that many fundamental aspects of cognition are indeed quite flexible? Or does finding that some aspect of cognition is flexible mean that it is therefore not fundamental? Or is flux the only truly fundamental thing about cognition in the first place? The talks in this symposium will speak to these questions from a variety of perspectives (incorporating ideas from development, neuroscience, computational insights, and cross-cultural approaches), and help us clarify our thinking about what such findings mean.

Variability and Specificity in Human Neuroplasticity: Flux is Fundamental!

Helen Neville and Christina Karns
Brain Development Lab, University of Oregon

The brain is in a state of constant change. In fact, one might argue that the reason some systems have such short critical periods is the constant pressure from competing systems in a rapidly changing brain. Different brain systems and related functions display markedly different degrees or 'profiles' of neuroplasticity in human development. Some systems are strongly determined and are not altered even when experience has been very different. Others are highly modifiable by experience and dependent on experience but only during particular time periods. There are several different such sensitive periods, even within a domain of processing. A third 'plasticity profile' is demonstrated by those neural systems that remain capable of change by experience throughout life. Neuroplasticity is a double-edged sword that permits both enhanceability and vulnerability. These findings contribute to a basic understanding of the nature, mechanisms and constraints of human brain plasticity, a fundamental player in all aspect of cognition. In addition, they can contribute information of practical significance in the design and implementation of educational programs.

Flexibility does not imply flux

Arthur B. Markman
University of Texas

Cognitive Science often focuses on the core aspects of cognitive processing that are common across individuals. Indeed, we often treat adaptability to context and variability across individuals as statistical error. Periodically, however, this variability comes into focus. This focus on variability is typically accompanied by calls for a fundamentally different way of characterizing cognitive processing such as dynamical systems, situated cognition, or embodied cognition. That is, there is an implicit assumption that the fluidity of cognitive processing is somehow incompatible with many of the core explanatory constructs in the field. I argue that variability and flexibility in cognitive processing is crucial for us to understand, but that they are explicable without having to give up most of the traditional representational and processing assumptions of cognitive science. I illustrate this point with examples from analogical reasoning, decision making, and motivation.

Surfing the Standing Wave of Cognition

Michael J. Spivey
University of California at Merced

Many of the most noticeable properties of cognition appear to be stable structures, concepts and categories in the mind that seem to function like static representations of things out in the world. This appearance of stability stands out in sharp relief at the time scale of several seconds, during momentary introspection or in a paper-and-pencil experiment. At finer and coarser time scales, there are dramatic patterns of change in those same cognitive structures, during neural processing and real-time responses and during long-term task performance and learning. This endemic property of flux that both underlies and overlays our subjectively stable mental entities has become the poster child for a theoretical framework in cognitive science calling itself the dynamical systems account of cognition (e.g., Chemero, 2009; Elman et al., 1996; Kelso, 1996; Port & Van Gelder, 1996; Spivey, 2007; Thelen & Smith, 1994; Ward, 2002). In this framework, the mental entities that appear as stable structures in cognition are much like standing waves which, if not examined at multiple time scales, cannot be properly understood. Essentially, even those things that appear stable in cognition are actually seen to be in flux when carefully analyzed. Therefore, if being in flux somehow prevents a property from being fundamental, then nothing in cognition is fundamental.

Solid as a rock, smart as a rock?

Lera Boroditsky
Stanford University

The human ability to flexibly adapt to a wide and unpredictable range of circumstances is the very trademark of human intelligence. When we study flexibility and diversity in human thought, we are approaching what may in fact be the human essence, those qualities that distinguish us from all other creatures. In this talk I will highlight a number of discoveries of radical diversity in human cognition, as a function of cultural and linguistic context. I will highlight four categories of differences that constitute different aspects of being "fundamental": differences that are deep, differences that are pervasive, differences that are big, and differences that are important. These findings demonstrate that many aspects of cognition that were previously thought to be static or pre-determined, are indeed quite flexible, the product of cultural invention and transmission. Studies of cross-cultural variation demonstrate that people can construct a variety of radically different perspectives on the same physical reality. I will argue that it is this flexibility that allows us to construct ever more complex and sophisticated conceptual tools, and adapt so successfully in cultural as opposed to in evolutionary time. When it comes to higher-level cognition, being solid as a rock may only be desirable if one wants to be as smart as one.

Author Index

Azizi Ab Aziz 435
Brandon Abbs 588
Sherief Abdallah 453
Rasha Abdel Rahman 576
Adele Abrahamsen 31
Richard Abrams 569
Kat Agres 600
Fatemeh Ahmadi-Fakhr 2615
Nayef Ahmar 2644
C. Athena Aktipis 1517
Ozge Alacam 2388
John Alderete 635
Vincent Aleven 2888
Robert Alexander 1222
Jamie Alexandre 1381
Martha Alibali 584, 628
Afra Alishahi 2452, 2615
Suvarna Alladi 670
Richard Alterman 1661
Maartje Ament 91
Craig Anderson 604
James Anderson 1276
Michael Anderson 1511, 1517
Nicola Anderson 604
Richard Anderson 644
Sarah Anderson 1192, 1198
Elena Andonova 2506
Jan Andrews 609
Giulia Andrighetto 1282
Florencia Anggoro 2426
Jana Appel 802
Mamiko Arata 1374
Richard Arias Hernández 526
Blair Armstrong 585
Inbal Arnon 760
Shin-ichi Asakawa 680
Richard Ashley 625
Richard Aslin 364, 2063, 2476
Stephen Asma 507
Joseph Austerweil 73
Ruqiyabi Naz Awan 453
Zohreh Azimifar 1529
Veerle Baaijen 1774
William Badecker 706
Hao Bai 593
Renee Baillargeon 180
Chris Baker 1184
Joseph Baker 552, 2799
Dare Baldwin 2710
Alan Bale 1178
Jerry Ball 1577, 1583
Linden Ball 1008
Thomas Balslev 1703
Martijn Balsters 2542
Raju Bapi 670
Laura Barca 266, 831
Neil Bardhan 364
Dave Barker-Plummer 808
David Barner 1178, 1246
Sam Baron 2548
Lawrence Barsalou 514
Daniel Bartels 483
James Bartolotti 532
Miriam Bassok 1571
Stuart Battersby 1998
Martin R. K. Baumann 1974
Peter Baumann 1204, 2218, 2438
Philip Beaman 554, 1014
Irina Beatu 814
William Bechtel 31
Nicole Beckage 2769
Benedikt Becker 2117
Stephen Begg 1792
Sam Behseta 1816
Michael Beigl 641
Sian Beilock 511
Daniel Belenky 459
Mikhail Belkin 694, 2414
Sieghard Beller 2767
Paul Bello 2022

Author Index

Tony Belpaeme 1362
Andrea Bender 2767
Viridiana Benitez 636, 2627
Giovanni Bennardo 2767
Brianna Bennett 633
Shlomo Bentin 1112
Benjamin Bergen 901
Leon Bergen 853
François Bernier 613
Vincent Berthiaume 682
Alain Berthoz 648
Catherine Best 1846
Ryan Best 2421
Klinton Bicknell 1142
Nik Nailah Binti Abdullah 2401
Jeffrey Bisanz 578
MARIA LUISA BIZZOTTO 1952
Elizabeth Ligon Bjork 1535
John Black 653, 2242, 2326
Mark Blair 236, 1354
Ann Blandford 91
Paulo Blikstein 2435
Svetoslav Bliznashki 2743
Mark Blokpoel 1643
Eliza Bobek 2650
Franziska Bocklisch 1974
Steffen Bocklisch 1974
Peter Bogunovich 1732
Luca Bonatti 2374
Elizabeth Baraff Bonawitz 2260, 2272
Jean-Francois Bonnefon 453
Dixon Bonnie 603
Walter Boot 2536
Julie Booth 1649
Anna Borghi 266
Lera Boroditsky 895, 1336, 2105, 2918
Arielle Borovsky 597, 1307
Jelmer Borst 513
Solveig Bosse 1052
James Boster 2767
Lewis Bott 509
Roberto Bottini 1342, 1348
Matthew Botvinick 525
Jason Braasch 1505
Chloe Bracis 1276
Timothy Brady 411
Nick Braisby 393
David Brang 632
David Braze 1429
Brooke Breaux 1601
Michelle Brewer 675
Katherine Brill 590
Chandra Brojde 1356
Neon Brooks 1178
Richard Brooks 586
Geoffrey Brookshire 1940
Randy Brou 582, 593
Kyler Brown 972
Scott Brown 1946, 2224, 2804
Michel Brudzinski 848
Duncan Brumby 91, 2389
Lionel Brunel 2392
Belinda Bruza 1792
David Buchanan 919, 1786
Leandra Bucher 1002
Daphna Buchsbaum 2858
Simon Buechner 553
Marc Buehner 1709
Magdalena Bugajska 672
Daniel Bullard 2425
Joseph Burling 2431
K.C. Burns 591
Nicholas Burns 2524
James Burton 2418
Jerome Busemeyer 1166, 2416
Kirsten Butcher 665, 2888
Lucas Butler 1417
Donna Byron 620
Heinrich Bülthoff 2500
Eva Cadez 2433

Author Index

Vladimir Cadez 2433
Jonathan Cagan 629
Kursat Cagiltay 2388
Aimee Callender 2123
Dustin Calvillo 557, 586, 608
Ellen Campana 515
Gwendolyn Campbell 1738
Ruth Campbell 1040
Angelo Cangelosi 1362
Lisa Cantrell 537
Jeremy Caplan 562, 669
Kathleen Carbary 109
Laura Carlson 569, 579
Richard Carlson 2397
Christopher Carroll 913, 1076
Mary Carroll 547
Daniel Casasanto 127, 512, 631, 1342, 1348, 1940
Adam Cassar 1709
Daniel Cassenti 2397
Deonne Castaneda 808
Enrico Castelli 831
Kefyn Catley 2656
Richard Catrambone 667
James Caverlee 2436
Fabian Cañas 1258
Fabián Cañas 1264
Thierry Chaminade 2716
Joel Chan 629
Margaret Chan 598
Michelle Chan 669
Raymond Chan 2430
Wen-Hsuan Chan 1483
Alicia Chang 662
Chun-Hao Chang 653, 2242
Julia Chariker 2016
NIck Chater 1720
Chi-hsin Chen 2236
Dawn Chen 871, 1875
Fang Chen 2404
Hsin-Chin Chen 676
Hung-Yi Chen 538
Jenn-Yeu Chen 534, 538, 1435
Lang Chen 2194
Sau-chin Chen 574
Yeechi Chen 639
Yung-Jung Chen 538
Joey Cheng 661
Karen Cheng 639
P. C-H. Cheng 1922
Patricia Cheng 913, 1076
Rong-Ju Cherng 538
Nadia Chernyak 2601
Kloser Chee Fung Cheung 1441
Min Chi 2870
Simone Chin 150
Seth Chin-Parker 2381
Joseph Chisholm 634
Eric Chiu 643
Pyeong Whan Cho 2200
Dongkyu Choi 2099
JaeHyuk Choi 572
MinGyung Choi 616
Tow Chong Chong 1106
Feng-Xuan Choo 2188
Xuan Choo 1100
Jessica Choplin 501, 1471
Morten Christiansen 618, 2686
Anna Chuderska 2749
Adam Chuderski 931
Meghan Clayards 663
James Clinton 679
James Close 1619
Caroline Cochrane-Braswell 145
Moreno Coco 1070, 1934
Andrew Cohen 2416
Jonathan Cohen 1270
Uriel Cohen Priva 43
Edward Cokely 405, 531, 2440
Gregory Colflesh 563, 647
Savéria Colonna 2218

Author Index

Eliana Colunga 399, 1356
Leisha Colyn 644
Luke Conlin 19
Delayne Connor 637
Rosaria Conte 1282
Christopher Conway 1393
Anne Cook 665
Rachel Cooper 1008
Richard Cooper 937
Peter Coppin 2418
Albert Corbett 2882
Adam Corner 1625
James Corter 1762, 2650, 2662
David Cosejo 649
Seana Coulson 632
Kenny Coventry 580, 614
Anna Cox 91
Greg Cox 429
Richard Cox 808
Edward Cranford 1768
Sarah Creel 1810
Ulrike Cress 784
Aimée Kay Crisp-Bright 67
Matthew Crocker 1637, 2458
Ashley Crockett Mazal 665
Scott Crossley 984
Caroline Crouch 664
Marco Cruciani 1028
William Cunningham 1234
Glennon Curran 507
Patrick Cushen 627
J. Cooper Cutting 679
Laura D'Andrea 2314
Sidney D'Mello 308, 2721
Rick Dale 121, 206, 594, 2419
Robert Dale 808
Olaf Dammann 682
Frederic Dandurand 688
Andreea Danielescu 515, 2566
Kevin Darby 2431
Adam Darlow 1088
Hia Datta 600, 602
Eddy Davelaar 85, 630, 937
Jodi Davenport 564
Nicolas Davidenko 595
Jim Davies 1553
Sarah Davies 665
Colin Dawson 79
Samuel Day 465
Kees de Glopper 1774
Wim De Neys 1020
Gilberto de Paiva 519
Paul De Palma 2410
Virginia de Sa 718
Philip DeCamp 2710
Daniel DeCaro 587
Marci DeCaro 536
Nina del Rosario 2389
Krista DeLeeuw 2390
Gary Dell 1447
Sophia Deng 230
Stephanie Denison 2272
Simon Dennis 97, 694, 842, 2414
Heeyeon Dennison 901
Stephen Denton 642, 2302
Jean-Louis Dessalles 1928
Barry Devereux 49
Aaron Dewald 665
Andrew Dewald 796
Gedeon Deák 278, 284, 2482
Judy Diamond 664
Connor Diemand-Yauman 2739
Zoltán Dienes 2206
Steve DiPaola 533
Kyung Soo Do 567, 572, 573
Stephanie Doane 558, 582, 593
Pennie Dodds 2804
Leonidas Domas 796
Peter Ford Dominey 542
Stéphane Donikian 648

Author Index

Christopher Donkin 1946
Chris Donkin 2804
Leonidas Doumas 2338
Mike Dowman 352
Lisa Dragoni 550
Amanda Drescher 961
Matthew Dry 1840
Vasanta Duggirala 670
David Dunbar 1613
John Dunn 97
Kelley Durkin 638
Justin Durtschi 679
Melody Dye 990
Myroslava Dzikovska 1738
Natasha Eapen 2548
Catherine Eberbach 133
Kathleen Eberhard 2051
Jörg Edelmann 2039
Berit Eika 1703
Alexander Eitel 555, 2822
Brianna Eiter 621
Chris Eliasmith 61, 1100, 2188
Jeffrey Elman 597, 1058, 1307
Lauren Emberson 2518
Paula Engelbrecht 2399
Randi Engle 477
Susan Epstein 996
Christopher Erb 495
Selda Eren 754
Carson Eric 2242
Kirk Erickson 2536
K. Anders Ericsson 2421
Eyal Ert 320
Zachary Estes 652
Marc Ettlinger 381
Owain Evans 853
Cameron Fadjo 653
Mandy Faretta 590, 596
Thomas Farmer 618
Lisa Fast 578
Caitlin Fausey 1330, 1336
Afsaneh Fazly 1529, 2452, 2615
Jessica Federman 550
Kara Federmeier 585
Aidan Feeney 67
Jessica Feigleson 2799
Michele Feist 1064, 1601
Jacob Feldman 2554
Laurie Beth Feldman 645
Veselina Feldman 1453
Adam Feltz 2440, 2560
Philip Fernbach 1088
Lisa Ferrara 665
Nicol Ferrier 614
Sara Finley 706
Anna Fisher 1493, 1499, 2488
Brian Fisher 526
Kristie Fisher 1571
Hartmut Fitz 2692
Robin Flanagan 724
Monique Flecken 547
Stephen Flusberg 595, 2105
Christopher Flynn 1505
Kenneth Forbus 1726, 2761
Tom Foulsham 543, 661
Trevor Fountain 1916
Adam Fouse 718
Neil Fox 601
Michael Frank 760, 1822
Julia Frankenstein 648
Lindsey Frederixon Byom 1957
Emily Freeman 97
Mary Freiman 1583
Robert French 681
Scott Freundschuh 579
Seth Frey 2093
Damian Fricker 2236, 2609
Ori Friedman 180
Scott Friedman 1726
Ellen Frohning 109

Author Index

Michael Fry 236
Katherine Fu 629
Wai-Tat Fu 2314, 2424, 2536
Karl Fua 25
Esther Fujiwara 562
Danilo Fum 338
Kotaro Funakoshi 447
Daniel Funk 2656
Robert G.M Hausmann 2401
Liane Gabora 955, 2350, 2432
Rami Gabriel 507
Soniya Gadgil 2583
Jean-François Gagnon 613
David Galbraith 1774

C. Randy Gallistel 517
Randy Gallistel 2793
Brian Gane 667
Tarun Gangwani 1595
Alexis Garland 591
Merideth Gattis 1046
Dedre Gentner 712
Severina Georgieva 1986
Lisa Geraci 2408
Jeffrey Gerard 218
Yannick Gerard 681
Emilie GERBIER 622

Joy Geren 2775
Alexander Gerganov 1986
Peter Gerjets 1703, 2039, 2278, 2828
LouAnn Gerken 79
Lisa Gershkoff-Stowe 2236
Samuel Gershman 1270
Tobias Gerstenberg 1697
Ben Gerstle 2182
Francesca Giardini 1282
Bryan Gibson 2320
Barry Giesbrecht 2390
Jill Gilkerson 121
Joan Gilkey 1547
Kuba Glazek 417
Vladimir Glebkin 2400
Jeremy Glick 668
Claudius Gläser 1744
Karrie Godwin 1493
Kelly Goedert 621
Ashok Goel 133, 1691, 2128
Winston Goh 423, 544
Alex Goldberg 609
Jill Goldstein 588
Robert Goldstone 465, 1216, 1405, 2093
Micah Goldwater 551
Sharon Goldwater 760
Dayna Gomes 557, 586, 608
Noah Goodman 180, 859, 1184, 2182, 2296, 2840
Andrew Goodwin 1276
Alison Gopnik 1228, 2272, 2852, 2858
Joshua Gordon 996
Art Graesser 2721
Jonathan Grainger 688
Steven Gray 133, 2128
Wayne Gray 848
Tera Marie Green 533
Michelle Greenwood 611
William Greville 1709
Thomas Griffin 583
Thomas Griffiths . 73, 242, 352, 2069, 2260, 2272, 2852
Tom Griffiths 2858
Maurice Grinberg 332
Sarah Gripshover 607
Marcello Guarini 2022
Matthew Guerdat 575
Diego Guerrero 2212
Markus Guhe 1992
Glenn Gunzelmann 2134
Ayush Gupta 19
Prahlad Gupta 650
Swati Gupta 521
Todd Gureckis 162, 248, 1216, 2607
Helmar Gust 1992
Ricardo Gutierrez-Osuna 2436

Author Index

Hyowon Gweon 2846
Nicholas Gwynne 1301
Jaime Gómez 626
Leah Hackman 549
Gerald Haeffel 2578
York Hagmayer 925, 2087
Udo Hahn 2146
Ulrike Hahn 1619, 1625
Harry Haladjian 2793
John Hale 2152
Tim Halverson 2134
Kiley Hamlin 1184
David Hammer 19
Rima Hanania 606
Sharon Hanlon 393
Joy Hanna 620
Thomas Hannagan 688
Bridgette Hard 2710
Mary Hare 1058
Mary Harmon-Vukic 2417
Adam Harris 1625
Maegan Harris 2560
Anthony Harrison 200
Joseph Harrison 2320
Katherine Harrison 1738
William Hartmann 620
Joshua Hartshorne 1186
Beth Hartzler 644
Shanta Hattikudur 584
David Havas 1804
Guy Hawkins 2224
Patrick Hays 308
Patrick G T Healey 2004
Patrick Healey 1998
Andrew Heathcote .. 1946, 2224, 2248, 2804
Andrew Heckler 139, 541
Neil Heffernan 2876
Jennifer Heil 2578
Evan Heit 1858, 2433
Christoph Held 784
Barbara Hemforth 2218, 2438
Pernille Hemmer 1130, 2402
Cynthia Henderson 2395
MIchelle Hendricks 1393
Andrew Hendrickson 1216
Petra Hendriks 617, 1325
Tania Henetz 512
Joseph Henrich 661
Marc Hernandez 527
Adolfo Hernando 626
Carlos Hernández 626
Stephanie Herppich 314
Julianne Herts 609
Ralph Hertwig 2391
Daniel Heussen 2033
Shohei Hidaka 1589, 2437
Barbara Hidalgo-Sotelo 820
Thomas Hills 168, 465, 2391, 2769
Cindy Hmelo-Silver 133, 2128
Valesia Ho 675
John Hoeks 1325
Aaron Hoffman 1354, 2755
Bob Hoffman 2882
Philip Hofmeister 224
Annette Hohenberger 754, 2388
Elena Hoicka 1040, 1046
Geoff Hollis 655
Kevin Holmes 2704
Kenneth Holmqvist 1703, 1968
Jana Holsanova 1968
Keith Holyoak 871, 1094, 1875
Hidehito Honda 772
Robert Honey 1619
JeeHye Hong 653
Shinichi Honiden 2401
Sameer Honwad 133, 2128
Jared Hotaling 2416, 2607
Joseph Houpt 1148
Christine Howes 2004
Evgenia Hristova 332, 1986

Author Index

Penka Hristova 1453
Andy Hsia 656
Janet Hui-wen Hsiao 1441
Anne Hsu 242, 1720
Jon-Fan Hu 574
Judith Hudson 2422
Stephanie Huette 1154, 1192
John Hummel 2894
Alycia Hund 523
Rafael Hurtado 2212
Ferenc Huszar 2810
Christoph Hölscher 553
Mutsumi Imai 1160, 1374
Birgit Imhof 2039
Bipin Indurkhya 2834
Keisuke Inohara 615
Thea Ionescu 606
Hiroshi Ishiguro 2716
Yoshihiro Itaguchi 680
Richard Ivry 1940
Daniel Jackson 889
Robert Jacobs 156, 2633
John Jacobson 658
Allison Jaeger 561
Georg Jahn 260
Sireesha Jala 670
Karin James 651
Randall Jamieson 1541
Jooyoung Jang 560
Susan Jang 2326
Christian Janssen 2389
R. Joanne Jao 284
Halszka Jarodzka 1703
Andrew Jarosz 563, 647
Kyle Jasmin 127, 631
Hector Jasso 278
Benjamin Jee 664, 2426
Gavin Jenkins 599
Patrick Jennet 2200
Kyle Jennings 978
Julia Jenvey 1804
Yoonkyung Jeong 612
Patrick Jeuniaux 613
Sandra Jhean-Larose 2412
Mark Johansen 1709
Roger Johansson 1968
Brendan Johns 55
Clint Johns 1429
Angie Johnson 580
Joseph Johnson 587
Phil Johnson-Laird 2028, 2900
Lamia Johnston 2010
Bevan Jones 2140
Dylan Jones 554
Jason Jones 2434
Matt Jones 1258, 1264
Michael Jones 55, 865, 877
Susan Jones 1655
Winston Jones 558
Jerome Scott Jordan 15, 344, 679
Kerry Jordan 552, 1459, 2799
Rebecca Jordan 133, 2128
Jürgen Jost 571
Frank Joublin 1744
David Joyner 133, 2128
Robert Jyung 2326
Linda Kaastra 526
George Kachergis ... 1216, 1595, 2362, 2464
Hilary Kalagher 1655
Aaron Kalb 808
Christopher Kalff 168
Charles Kalish 628, 2320
Michael Kalish 188, 1064
Deepthi Kamawar 578
Jennifer Kaminski 1828
Yvonne Kammerer 2278
Frank Kanayet 1234
Pentti Kanerva 865
Sean Kang 610
Justine Kao 990

Author Index

Vsevolod Kapatsinski 2010
Manu Kapur 2727
Themis Karaminis 730
Saraschandra Karanam 2834
Yaakov Kareev 1697
Fred Karlsson 2045
Linnea Karlsson 405

Samir Karmakar 826
christina karns 2918
Rajesh Kasturirangan 826
Margarita Kaushanskaya 1957
Alan Kawamoto 673, 2284
Artem Kaznatcheev 967, 972
Andrew Kehler 2057
Frank Keil 907
Frank Keller 218, 1070, 1559, 1934
Josh Keller 1289
Troy Kelley 2397
Philip Kellman 2409
Christopher Kello 655
Ronald Kellogg 1393
Charles Kemp 346
McRae Ken 1026
William Kennedy 672
Ingo Kennerknecht 571
Andrew Kenning 679
Trina Kershaw 1505
Alan Kersten 150
Bryan Kerster 655
Pranav Khaitan 623
Muhammad Ali Khalidi 194
Saera Khan 675
Aaron Kheifets 517
Sangeet Khemlani 2028, 2403, 2900
Naveen Khetarpal 358
Celeste Kidd 2476
Stephen Killingsworth 658
Dahee Kim 619
Eun Young Kim 1881
Eunsook Kim 2405
Hanna Kim 567
Jihie Kim 2344
Kyuhee Kim 573
Ryung Kim 2876
Taehwan Kim 2344
Yoon Kim 1881
Alan Kingstone 516, 543, 589, 604, 634, 661, 1489
Natasha Kirkham 646, 1228, 1863
David Kirsh 2864
Muneo Kitajima 529, 530
Audrey Kittredge 1447
Sachiko Kiyokawa 2206
Michael Kleiber 1679
Michel Klein 435
Heidi Kloos 495, 2266, 2425
Markus Knauff 1002
Pia Knoeferle 2446
Kazuki Kobayashi 447
Kenneth Koedinger 1649
Ken Koedinger 471
Judith Koehne 2458
Bryan Koenig 2396
Olivier KOENIG 622
Takatsugu Kojima 657
Boicho Kokinov 1453, 2081, 2743
Thorsten Kolling 2482
Takanori Komatsu 447
Lars Konieczny 1204, 2218, 2438
Anna Koop 549
Anna Korhonen 49
Danny Kostons 296
Kenneth Kotovsky 629
Emiel Krahmer 115, 736, 2542
Arthur Kramer 2536
Josef Krems 1974
Roger Kreuz 206, 2419
Trent Kriete 674
Sarah Kriz 639
Antje Krumnack 1002
Ulf Krumnack 1992

Author Index

John Kruschke 642
Nicole Krämer 802
Marta Kuas 2446
Sarah Kucker 2621
Gustav Kuhn 543, 589
Anuenue Kukona 1429
Sarah Kulkofsky 2420
Maithilee Kunda 1691
Niklas Kunze 2028
Kenneth Kurtz 605
Tamar Kushnir 656, 1184, 2601
Takashi Kusumi 570, 615
Marta Kutas 1058, 1307
Megumi Kuwabara 2627
Johan Kwisthout 1643
Mee-Kyoung Kwon 612
Kai-Uwe Kühnberger 1992
Elodie Labeye 2392
Paul Ladny 582, 593
Francy Ladusaw 2182
Daniel Lafond 613
David Lagnado 1697
Jun Lai 1387
Kaitlin Laidlaw 543
Tei Laine 521
Brenden Lake 778
Stéphane Lallée 542
Alessandro Lamberti Bocconi 266
David Landy 2164, 2894
Pat Langley 2368, 2566
Mirella Lapata 1916
Johann Larusson 1661
Lyuben Laskin 1892
Sarah Laszlo 585
Melissa Latham 675
Christine Lau 562
Marie Lauer-Schmaltz 624
Hee Seung Lee 1094
Hyo-hee Lee 567
Michael Lee 103, 387, 1118, 1124, 1565, 1840, 2607
Soo-Young Lee 581
Woo-yeol Lee 1881
Yoonha Lee 1881
Jo-Anne LeFevre 578
Richard Leibbrandt 2680
Stefan Leijnen 955
Alessandro Lenci 1886
Mate Lengyel 2810
Itamar Lerner 1112
Mathieu Lesourd 2392
Kimery Levering 605
Florent Levillain 2374
Daniel Levin 592
Susan Levine 612, 2816
Roger Levy 1142, 1313, 1483
Joshua Lewis 278, 718
Mark Lewis 1780
Michael Lewis 509
Owen Lewis 2332
Cheng-Yi Li 534, 1435
Jia Li 2344
Nan Li 2566
Ping Li 574, 2787
Tiziana Ligorio 996
Stephen Wee Hun Lim 423, 700
David Lin 961
Robert Lindsey 2332
Sally Linkenauger 2164
Diane Litman 2870
Daniel Little 520
Jeri Little 1535
Daniel Hsi-wen Liu 2413
Qiang Liu 2284
Ran Liu 602
Yang Liu 669
Yanping Liu 1136
Ken Livingston 609
Alejandro Lleras 489
Jeffrey Loewenstein 1289
Steffen Lohmann 2146

Author Index

Megan Lombardi 1471
Tania Lombrozo 528, 1301, 2906, 2912
Max Lotstein 2028
Max Louwerse 961
Bradley Love 2607, 2755
Andrew Lovett 2761
Jason Low 591
Robert Lowe 1607

Hongjing Lu 871, 1076, 2409
Christopher Lucas 2852
Chris Lucas 1184
George Luger 2410
Jií Lukavský 1380
Gary Lupyan 883, 1210
Simon Lynn 592
Don Lyon 510
Ian Lyons 511
Joseph MacInnes 585
Ben MacLaren 2882
Mohammed Iqbal Madakkatel 453
Christopher Madan 562
Carol Madden 542
W. Todd Maddox 654
Akihiro Maehigashi 943
Alfons Maes 115
James Magnuson 1429, 2200
Phil Maguire 748
Rebecca Maguire 748
Asifa Majid 358, 2767
Barbara Malt 358
Emmanuel Manalo 1852
Stefan Mangold 2111
Viorica Marian 532, 790
Doug Markant 248
Arther Markman 1762
Arthur Markman 1715, 2755, 2918
Ellen Markman 607, 1417
Jessecae Marsh 2420
John Marsh 554
Fischer Martin 1026
Fabiola Martinez 2412
Tony Martinez 2356
Hegarty Mary 603
David Mason 2662
Michael Matessa 1963
Robert Mathews 1750
Bryan Matlen 1493, 1499
Teenie Matlock . 524, 611, 1154, 1192, 1198, 1330
Tetsuya Matsuda 1374
Toshihiko Matsuka 772, 1762
Miki Matsumuro 766
Justin Matthews 524, 611
Percival Matthews 666
Richard Mayer 2390
Julien Mayor 836
Eva Mayr 1631
Ralf Mayrhofer 925, 1082
Brendan McCarthy 596
James McClelland 623, 668, 2395
Rachel McCloy 1014
Lauren McDonough 514
Katherine McEldoon 145
Charlotte McGinnis 2816
Keith McGreggor 1691
Jordan McGuire 582
Elizabeth McLaughlin 471
Matthew McLure 1726
Tyler McMillen 1816
Bob McMurray 2230
Danielle McNamara 984, 1319
Nicole McNeil 2578
Ken McRae 1058
Björn Meder 2087
Douglas Medin 2767
Ben Meijering 1423
Tobias Meilinger 2500
Zulfiqar Ali Memon 441
Einar Mencl 1429
Adam Mendelson 477
Stefania Mereu 489

Author Index

Angela Merritt 405
Everett Mettler 2409
Ross Metusalem 1058
Alex Metz 1685
D. J. K. Mewhort 1541
Meredith Meyer 2710
Kevin Mickey 579
Stieff Mike 603
Brent Miller 1130
Koji MINESHIMA 2668
Daniel Mirman 1447
Jennifer Misyak 618, 2686
Kazuhisa Miwa 766, 943
Kelly Mix 565
Lisette Mol 115, 736
Padraic Monaghan 618
Brian Monroe 521
Kristine Monteith 2356
J. Michelle Moon 2424
Michelle Moon 2536
Johanna Moore 1738
L. Richard Moore Jr. 2134
Emily Morgan 1559
Phillip Morgan 949
Ph.D.,Kara Morgan-Short 596
Kara Morgan-Short 590
Ryan Morris 550
Robert Morrison 2338
Anthony Morse 1362
Fermín Moscoso del Prado Martín 645
Jarrod Moss 558, 1319, 1768
Jerad Moxley 2421
Michael Mozer 610, 2332
Kathryn Mueller 2230
Neil Muggleton 2439
Edward Munnich 675
Christopher Myers 1583
Jay Myung 556, 2572
Daniel Müller 1204
Stephanie Müller 531
Farah Naaz 2016
Peter Nachbaur 609
Jonas Nagel 1082, 2111, 2595
Viswanath Naidu 670
Kuninori Nakamura 2393
Tetsuaki Nakamura 1898
Mikio Nakano 447
Laura Namy 518
Antonio Napoli 338
Mitchell Nathan 1804
Daniel Navarro 1411, 1792, 1834
James Negen 1252
Nicole Neiman 1477
Jelica Nejasmic 1002
Angela Nelson 254
helen neville 2918
Elissa Newport 2063
Phi Nguyen 477
Hannele Nicholson 2051
Jeffrey Nickerson 2662
Prolet Nikolova 1904
Marie Nilsenova 212
Jonna Nilsson 614
Yael Niv 1270
David Noelle 674, 678
Timothy Nokes 459, 662, 2583
Uta Noppeney 2810
Laura Novick 2656
Nikolay Novitskiy 1020
Lynne Nygaard 518
Marcus Nyström 1703
Matthias Nückles 314
Brian O'Connor 2432
Tim O'Donnell 1186
Tim Oates 1511
Stellan Ohlsson 649, 2099
Kayoko Ohtsu 2494
Hiroyuki Okada 1160, 1374
Yas Okan 2111
Jiro Okuda 1374

Author Index

Aude Oliva 820
D. Kimbrough Oller 121
Andrew Olney 37, 308
Luca Onnis 889
Herre van Oostendorp 2834
John Opfer 1234, 2572
Daniel Oppenheimer 2403, 2739
Mariela Orozco-Hormaza 2212
Judy Orton 2426
Andrew Ortony 25, 2396
Daniel Osherson 2530
Lee Osterhout 1571
Adam Osth 842
A. Ross Otto 1715
Bernd-Christian Otto 531
Yoshihiro Ouchi 2494
Patrick O'Connor 645
Fred Paas 296
Amanda Padgitt 523
Bozena Pajak 369
Themis Palpanas 218
John Pani 2016
Jaak Panksepp 507
Peter Pantelis 2554
Anna Papafragou 1052
Christopher Parisien 2674
Paula Parpart 548
Harold Pashler 610, 2332
Rebecca Passonneau 996
Melissa Patchan 660
Jigar Patel 670
John Patrick 509, 949
Tanya Patrick 949
David Pautler 2396
Alison Pease 1992
Alan Penaloza 586
Marcie Penner-Wilger 578
Alfredo Pereira 651
Ashley Perez 2560
Amy Perfors 1411, 1613, 1834, 2524
Lori Petersen 2578
Georgi Petkov 1904
Adam Petrashek 180
Tamella Pettitt 630
Giovanni Pezzulo 266, 831
David Pfeiffer 2425
Brenda Phillips 2656
Steven Phillips 1523
Steven Piantadosi 859, 2476
Nicholas Pilkington 49
Mark Pitt 556, 619, 2572
David Plaut 585
Timothy Pleskac 326
Kim Plunkett 836
Thierry Poibeau 49
Fenna Poletiek 1387
James Pooley 103
Hanna Popick 1863
Christopher Potts 808
David Powers 2680
Janani Prabhakar 2422
Richard Prather 565, 2394
Melissa Prince 2224, 2248
Wolfgang Prinz 2698
Kenneth Pugh 1429
Matthew Purver 2004
Zenon Pylyshyn 2793
Joël Pynte 2218
Boon-Kiat Quek 2396
Edys Quellmalz 564
Milena Rabovsky 576
Anna Rafferty 2069
Marco Ragni 2117
Iyad Rahwan 453
Maartje E. J. Raijmakers 545
Vilayanur Ramachandran 632
Kiruthika Ramanathan 1106
Jennifer Ramautar 1020
Sheela Ramesh 1499
Michael Ramscar 990, 1863

Author Index

Daniel Rasmussen 61
Louise Rasmussen 1756
Kristin Ratliff 2816
Wolfgang Rauch 624
Stephen Read 1465
Gabriel Recchia 865, 877
Gisela Redeker 1325
Stephen Reed 2882
Patricia Reeder 2063
Terry Regier 358
Bob Rehder 1354, 2906
Erik Reichle 1136
Frank Renkewitz 260
Alexander Renkl 314
William Revelle 25
Spyridon Revithis 568
Yun Jin Rho 2662
Rebecca Rhodes 676
Theo Rhodes 655
Jeffrey Richards 121
Daniel. Richardson 659
Daniel Richardson 290, 1228, 2548
J. Elizabeth Richey 662
Lindsey Richland 2338
Travis Ricks 2176
Joerg Rieskamp 174
Jörg Rieskamp 1910
Michael Riley 344
Monica Riordan 2419
Benoit Riou 2392
Evan Risko 516, 589, 604, 634
Bethany Rittle-Johnson 145, 536, 638
Laurie Robinette 1064
Christopher Robinson 1846
Chris Robinson 2639, 2644
Marybel Robledo 284, 2482
Jennifer Roche 206, 594
Stuart Rodgers 1583
K.S. Rodzon 552
Katrina Rodzon 1459
Timothy Rogers 525, 2194, 2320
Hannah Rohde 381
Marco Rolandi 639
Michael Romano 2429
Beverly Roskos-Ewoldsen 575
Benjamin Rottman 907
Michael Rouillé 648
Brandon Roy 1822
Deb Roy 1822, 2710
Dani Rubinstein 2518
Fernando Rubio 559
Spencer Rugaber 133, 2128
Natalie Ruiz 2404
Robert Ryan 990
Henrik Saalbach 1160
Ivan Sag 224
Eyal Sagi 640
Magnus Sahlgren 865
Sarah Sahni 2781
Motoyuki Saito 546
Maki Sakamoto 1034, 1898
Yasuaki Sakamoto 1869, 2158
Joanna Salapska-Gelleri 540
Carlos Salas 583
Nancy Salay 182
Ron Salden 2876
Dario Salvucci 1732
Alexei Samsonovich 2308
Larissa Samuelson 599, 2621
Catherine Sandhofer 677, 2470
Ricardo Sanz 626
Barbara Sarnecka 1252
Yuri SATO 2668
Rebecca Saxe 180
Ayse Pinar Saygin 2439, 2716
Megan Saylor 592, 658
Eleanor Sayre 139
Thomas Scaife 139, 541
Juliette Schaafsma 2542
Lennart Schalk 1160

Author Index

Anna Schapiro 525
Christoph Scheepers 2218
Benjamin Scheibehenne 1910
Katharina Scheiter .. 555, 1703, 2039, 2822, 2828
Kathrin Schielke 802
Sarah Schimke 2218
Tobias Schlicht 272
Christopher Marc Schlick 1679
Martin Schmidt 1992
Hedda Rahel Schmidtke 641
Lauren Schneider 656
Walter Schneider 1319
Michael Schoelles 848
Agnes Scholz 1974

Anne Schueler 2828
Holger Schultheis 569
Laura Schulz 2846
Christian Schunn .. 560, 629, 660, 662, 1319

Gerhard Schurz 2387
Sarah Schwind 2266
Anne Schüler 2822
Rose Scott 180
Gun Semin 961
Carissa Shafto 2775
Patrick Shafto 671, 2182
William Shankle 103
Helen Sharp 2401
Erin Shaw 2344
Hongyuan Shi 2158
Jenny Shi 1124, 2402
Luping Shi 1106
Richard Shiffrin 254, 520, 1946, 2302, 2362, 2416, 2464
Hayley Shilling 1804
Hideaki Shimada 566
Tsuneo Shimazaki 546

Hyunjung Shin 616, 2405
Anthony Shook 790
Oren Shriki 1112
Andrew Shtulman 302, 1295
Thomas Shultz 682, 814, 972, 1685
Pooja Sidney 584
Winston Sieck 1756
Patrick Simen 1816
Dan Simon 1465

Dylan Simon 2607
Chris Sims 848
Clare Sims 399
Manish Singh 2793
Suparna Sinha 133, 2128
Scott Sinnett 796, 1489
Susan SJones 651
Sheri-Lynn Skwarchuk 578
Steven Sloman 358
Vladimir Sloutsky 230, 842, 1828, 1846, 1980, 2639, 2644
Alan Smaill 1992
Jodi Smith 599
Kenny Smith 577
Linda Smith . 537, 565, 606, 636, 651, 1362, 1589, 2609, 2769
Linsey Smith 712
Nathaniel Smith 1313, 1483
Philip Smith 1014
Thomas Smith 2236
Brenda Smith-Chant 578
Tomasz Smole" 931
Filip Smolík 1380, 1667
Michael Smuc 1631
Jesse Snedeker 1186, 2775
David Sobel 919, 1786
Ahmad Sohrabi 2075
Werner Sommer 576
Morgan Sonderegger 375
Elizabeth Spelke 1184

John Spencer 599
Amy Spiegel 664
Michael Spivey .. 611, 643, 889, 1154, 1192, 1198, 2429, 2918
Simone Sprenger 522
Anne Springer 2698
Vishnu Sreekumar 694, 2414
William B. St. Clair 678
Jon Star 638

Author Index

Maria Staudte 1637
Laura Staum Casasanto 127, 224
Mark Steedman 1559
Nancy Stein 527
Natalie Steinhauser 1738
Douglas Stenstrom 1465
Rachel Stephens 1411
Suzanne Stevenson 2615, 2674
Terrence Stewart 1100
Mark Steyvers 1130, 1840, 2402
Cody Stitzel 565
Andrea Stocco 513
Todd Stoess 671
Rainer Stollhoff 571
Gert Storms 2033, 2290, 2407

Laurie Stowe 1325
David Stracuzzi 2566
Chris N.H. Street 659
Chris Street 290, 2548
Sandra Street 565
Andreas Stuhlmüller 2296
Andrew Stull 603
Jennifer Sturm 577
Joshua Sturm 609
Yanjie Su 2430
Yasutada Sudo 1186
Jess Sullivan 1246
Ron Sun 1750
Yanlong Sun 1399, 2512
Abigail Sussman 2403
Rich Sutton 549
Lidia Suárez 544
Marc Swerts 115, 2542
Daniel Swingley 1210
Christine Szostak 619
Guadalupe Sánchez 626
Edward T. Cokely 548
Niels Taatgen 513
Whiteny Tabor 1429
Ronnie Taib 2404
Tomohiro Taira 570
Ryo TAKEMURA 2668
Franklin Tamborello 2512
Joanne Tan 290
Seok Hui Tan 544
Daisuke Tanaka 2206
Michael Tanenhaus 109, 364
Yun Tang 2572
Roman Taraban 633
Peggy Tausche 2698
Mark Tawney 501
Eric Taylor 1673, 1715
Julia Taylor 2170
Leanne Taylor 1738
Joshua Tenenbaum 411, 778, 853, 859, 919, 1184, 2296, 2840
Treysi Terziyan 1547
Robert Teszka 589
Ursina Teuscher 632
Anand Theertha 2834
Jean-Pierre Thibaut 681
Erik Thiessen 1368
Serge Thill 1607
Emily Thom 677
Michael Thomas 730
Emine Mine Thompson 580
Patrick Thompson 652
Robin Thompson 601
Sharon Thompson-Schill 883
Harry Tily 760
Mike Timms 564
Stephen Tobin 2200
Ekaterina Todorova 1986
Katrin Tomanek 2146
Bruce Tomblin 2230
Fatemeh Torabi Asr 1529
Joseph Toscano 663, 2230
Alexia Toskos Dils 895, 2105
Sumeyra Tosun 676, 2408
Tyler Towne 2421
Corinne Townsend 1858

Author Index

James Townsend 1148
Makoto Toyota 529, 530
Jessica Tracy 661
J. Gregory Trafton 200, 593
Duc Tran 2627
Sebastien Tremblay 613
Jan Treur 435, 441
Jochen Triesch 278
Jonathan Trigg 188
Jennifer Trueblood 1166
Michael Tull 2399
Nicholas Turk-Browne 1240
Barbara Tversky 2662
Christina Tzeng 518
Mieko Ueno 2057
Yuri Uesaka 1852
Tomer Ullman 1184, 2840
Matthias Unterhuber 2387
M. Afzal Upal . 742, 2406, 2411, 2417, 2733
Thomas Urbach 2446
Oleg Urminsky 483
Miki Uruwashi 1186
Jeffrey Usher 2761
Akira Utsumi 1034, 1898
David Uttal 664
Kevin Uttich 528
Jyotsna Vaid 676, 2408
Frederic Vallee-Tourangeau 1172
Guillaume Vallet 2392
Marije van Amelsvoort 212

Theo P. van der Weide 1643
Hans P.A. Van Dongen 2134
Julie Van Dyke 1429
Erlijn van Genuchten 1922
Tamara van Gog 296
Bianca van Kemenade 2439
Leendert Van Maanen 1423, 2398
Jacolien van Rij 617
Hedderik van Rijn 513, 522, 617, 1423
Iris van Rooij 1643
Tessa J. P. van Schijndel 545
Whitney Vandiver 2427
Ivan Vankov 2081
Kurt VanLehn 1319, 2870
Wolf Vanpaemel 2290
Sashank Varma 1780
Swaroop Vattam 133, 2128
Erikka Vaughan 2739
Dan Ventura 2356
Rineke Verbrugge 1423
Steven Verheyen 2407
Rémy Versace 2392
Manuel Vidal 648
Gabriella Vigliocco 601
Gaelle Villejoubert 1172
Ad Vingerhoets 2542
David Vinson 601
Janhavi Viswanathan 2834
Jonathan Vitale 2242
Haley Vlach 2470
Bettina von Helversen 174
Christiane von Stutterheim 547
Momme von Sydow 2087
Wouter Voorspoels 2033, 2290
Soroush Vosoughi 1822
Chris Vredenburgh 656
Johan Wagemans 1020
Eric-Jan Wagenmakers 2398
Angela Wagner 2882
Michael Waldmann . 925, 1082, 2087, 2589, 2595
Esther Walker 516
Vincent Walsh 2439
Sierra Walton 675
Chung-Yu Wang 2423
Hongbin Wang 1399, 2512
Yi Wang 2536
Thomas Ward 575
Anne Warlaumont 121
Russell Warner 671
Christina Wasylyshyn 2254

Author Index

Marcus Watson 1354
Rebecca Weast 1477
Robert Weisberg 417
Rob Weitz 2876
Helmut Weldle 1204
Henry Wellman 2601
Matthew Welsh 1792, 1798
Avishai Wershbale 326
Robert West 2075
Ruud Wetzels 387
John Whitman 2152
Colin Widmer 619
Alex Wiegmann 2111, 2589
Jan Malte Wiener 168
Geraint Wiggins 539
Jennifer Wiley 561, 563, 627, 647, 2176
Meredith Wilkinson 1008
Claire Williams 308
Emma Williams 509
Joseph Jay Williams 2906, 2912
Jon Willits 551
Nicholas Wilson 1750
William Wilson 1523
Christine Wilson-Mendenhall 514
Florian Windhager 1631
Carsten Winkelholz 1679

Stephan Winter 802
Thomas Wisdom 1405
Jessica Witt 344
Jörg Wittwer 314
Sascha Wolfer 1204
Phillip Wolff 2704
Dan Woltz 559
Jeong-ae Won 1881
Ph.D.,Patrick Wong 596
Francis Wong 590, 596
Patrick Wong 590
Kristin Wood 629
Michael Wood 236
Darrell Worthy 654
Rachel Wu 646, 1228
Saul Wyner 2344
Dongxin Xu 121
Jing Xu 352
Yang Xu 346
Reiko Yakushijin 156
Ayumi Yamada 2206
Seiji Yamada 447
Takashi Yamauchi 2436
Jin Yan 1289
Patrick Yaner 1553
Lee-Xieng Yang 2423
Yoshio Yano 1852
Xin Yao 1980
Melvin Yap 544, 700
Eldad Yechiam 320
Sheng Kung Yi 1840
Ilker Yildirim 2633
Michael C. W. Yip 535
SangSuk Yoon 616
Hanako Yoshida 2431, 2627
Robert Youmans 2428
Andrew Young 628
Christopher Young 2572
Alan Yu 375
Chen Yu 646, 1589, 1595, 2236, 2362, 2464, 2609
Lixiu Yu 2662
Yue Yu 2430
Jinhui Yuan 2415
Jiwon Yun 2152
Daniel Yurovsky 646, 1589, 2609
Mohamed Zaoui 648
Alessandra Zarcone 1886
Lisa Zaval 162
Mathew Zeigenfuse 1565
Gregory Zelinsky 1222
Xiaofang Zeng 633
Jason Zevin 600, 602
Bo Zhang 2415
Shunan Zhang 1118

Author Index

Wei Zhang 1222
Jiaying Zhao 1240, 2530
Libo Zhao 650
Robert Zheng 559, 665
Xiaojin Zhu 2320
Yuwen Zhuang 694, 2414
Jürgen Ziegler 2146
Benjamin Zinszer 2787

Reviewers

Azizi Ab Aziz
Abdul Rehman Abbasi
Brandon Abbs
Kat Agres
Woo-Young Ahn
F.-Xavier ALARIO
Kyle Albarado
John Alderete
Jamie Alexandre
Nadia Ali
Afra Alishahi
Eshaa Alkhalifa
Paul Allopenna
Richard Alterman
Erik Altmann
Maartje Ament
Michael Anderson
Sarah Anderson
Elena Andonova
Janet Andrews
Mark Andrews
Blair Armstrong
Richard Ashley
Ash Asudeh
Joseph Austerweil
Marios Avraamides
Roger Azevedo
Jonathan Back
Chris Baker
Jerry Ball
Martijn Balsters
David Barner
Daniel Bartels
Miriam Bassok
Peter Baumann
Jacob Beal
Anthony Beavers
William Bechtel
Brian Beitzel
Daniel Belenky
Paul Bello
Andrea Bender
Nick Benesh
giovanni bennardo
Benjamin Bergen
Vincent G. Berthiaume
Robert Berwick
Brad Best
Catherine Best
Maryse Bianco
Klinton Bicknell
Breton Bienvenue
Dorrit Billman
Kim Binsted
Nik Nailah Binti Abdullah
Tamas Biro
Mark Blair
Stephen Blessing
Rens Bod
Catherine Bohn-Gettler
Elizabeth Baraff Bonawitz
Jean-Francois Bonnefon
Julie Booth
Anna M. Borghi
Aaron Bornstein
Arielle Borovsky
Jelmer Borst
Tibor Bosse
Fiemke Both
Ty Boyer
Gary Bradshaw
Timothy Brady
Nick Braisby
Erica Briscoe
James Brockmole
Chandra Brojde
Andrew Brook
Tobias Brosch
Meredith Brown
Duncan Brumby
Lionel Brunel
Tad Brunye
Monica Bucciarelli
David Buchanan
Norbou Buchler
Daphna Buchsbaum
Simon J. Buechner
Michele Burigo
Bruce Burns
Richard Burns
Mark Burstein
Jerome Busemeyer
Kirsten Butcher
RUTH BYRNE
Aimee Callender
Gwendolyn Campbell
Guillermo Campitelli
Lisa Cantrell
Alessandro Capone
Fernando Cardenas
Richard Carlson
Christopher Carroll
Sharon Carver
Daniel Cassenti
Nicholas Cassimatis
Cristiano Castelfranchi
Richard Catrambone
Gina Caucci
Daniel Cavagnaro
Margaret Chan
Myriam Chanceaux
Alicia Chang
Julia Chariker
Davida Charney
Sylvain Chartier
Nick Chater
Jessie Chin
Seth Chin-Parker
Susan Chipman
Jessica Choplin
Adam Chuderski
Bill Clancey
John Clapper
Timothy Clausner
Meghan Clayards
Catherine Clement
James Close
Jonathan Clucas
Moreno I. Coco

Reviewers

Michael Coen
Jennifer Collins
Louise Connell
Susan Cook
Rick Cooper
Javier Alejandro Corredor
James Corter
David Cosejo
Fintan Costello
Nicholas Costen
Gary Cottrell
Michelle Cowley
Anna Cox
Richard Cox
George Cree
Christopher Crick
Matthew Crocker
Scott Crossley
Fred Cummins
Gregory Currie
Mario Córdoba
Francesca D'Errico
Nils Dahlbäck
Gregory Dam
Coral Dando
Frederic Dandurand
David Danks
Adam Danz
Drew Dara-Abrams
Adam Darlow
Eddy Davelaar
Jodi Davenport
Jim Davies
Colin Davis
Samuel Day
Joachim De Beule
Bart de Boer
Oscar De bruijn
Wim De Neys
Marci DeCaro
Morteza Dehghani
Peter Delaney
Charles Delbé
Krista DeLeeuw
Gary Dell
Vera Demberg
Simon Dennis
Rutvik Desai
John Dewey
Kristina Charlotte Dietz
Denise Dillon
Eric Dimperio
Abidgani Diriye
Tali Ditman
Stephanie Doane
Wei Dong
Christopher Donkin
Tim Donovan
Leonidas Doumas
JUSTINE DRAKEFORD
Felix Dreyer
Ben du Boulay
Geoffrey Duggan
Susan Dunlap
John Dunn
Nicholas Duran
Gilles Dutilh
Kathleen Eberhard
Ullrich Ecker
SUZANNE EGAN
Inge-Marie Eigsti
Elsa Eiriksdottir
Michelle Ellefson
Lauren Emberson
Paula Engelbrecht
Paul Engelhardt
Eileen Entin
Susan Epstein
Orlando Espino
Owain Evans
Igor Farkas
Thomas Farmer
Caitlin Fausey
Aidan Feeney
Michele Feist
jacob Feldman
Naomi Feldman
Kimberly Fenn
Philip Fernbach
Emma Ferneyhough
Mark Finlayson
Sara Finley
Anna Fisher
Brian Fisher
Stanka Fitneva
Hartmut Fitz
Robin Flanagan
Jessica Fleck
Monique Flecken
Stephen Flusberg
Ken Forbus
Deborah Forster
Tom Foulsham
Michael Frank
Robert Frank
Stefan Frank
Stella Frank
Stan Franklin
Bob French
Scott Friedman
Jim Friedrich
Wai-Tat Fu
Karl Fua
Andy Fugard
Danilo Fum
Dominic Furniss
Soniya Gadgil
Francesco Gagliardi
Brian Gane
RAQUEL GARCIA JURADO
Ricardo Garcia
Simon Garrod
Hector Geffner
William Gehring
Dedre Gentner
Nathan Gerhart
LouAnn Gerken
Charlotte Gerritsen
Rachel Giora

Reviewers

Benoit Girard
Helene GIRAUDO
Jeremy Glick
Fernand Gobet
Kelly Goedert
Ashok Goel
Winston Goh
Joshua Goldberg
Rob Goldstone
David M Gomez
Pablo Gomez
Noah Goodman
Simon Goodson
Nicholas Gorski
Lawrence Gottlob
Amelie Gourdon
Justin Grace
Eileen Graf
Alberto Greco
Collin Green
Tera Marie Green
Carsten Griesel
Gina Griffiths
Tom Griffiths
Lisa Grimm
Qun Guan
Amal Guha
Markus Guhe
Glenn Gunzelmann
Gianchand Gupta
Swati Gupta
Todd Gureckis
Hyowon Gweon
Nicholas Gwynne
Göran Hagert
York Hagmayer
Ulrike Hahn
John Hale
Tim Halverson
James Hampton
Adam Harris
Kevin Harris
randy harris
Joshua Hartshorne
Peter Hastings
Robert Hausmann
Brett Hayes
Pat(rick) Hayes
Andrew Heckler
Neil Heffernan
Mary Hegarty
Sebastien Helie
Michael Helms
Barbara Hemforth
Pernille Hemmer
Cynthia Henderson
Andrew Hendrickson
Tania Henetz
Seth Herd
Mitchell Herschbach
Larry Hettinger
Daniel Heussen
Masako Hirotani
John Hoeks
Marieke Hoetjes
Aaron Hoffman
Kevin Holmes
Keith Holyoak
Alexandra Horowitz
William Horton
Autumn Hostetter
Jared Hotaling
Joseph Houpt
Andrew Howes
Penka Hristova
Roland Hubscher
Stephanie Huette
John Hummel
Alycia Hund
Julie Hupp
Edwin Hutchins
Christoph Hölscher
Mutsumi Iijima
Birgit Imhof
Terresa Jackson
Robert Jacobs
Joy Jacobs-Lawson
Georg Jahn
Jooyoung Jang
Christian Janssen
armina janyan
Kyle Jasmin
Benjamin Jee
Charlene Jennett
Kyle Jennings
Alan Jern
Ryan Jessup
Patrick Jeuniaux
Sandra Jhean-Larose
Yang Jiang
William Jimenez-Leal
Angie Johnson
Cheryl Johnson
James Johnston
Gary Jones
Lara Jones
Michael Jones
Jerome Scott Jordan
Marie Juanchich
Regina Jucks
Charles Kalish
Mike Kalish
Deepthi Kamawar
Jennifer Kaminski
Frank Kanayet
Ruogu Kang
Thomas Kannam
Vsevolod Kapatsinski
Themelis Karaminis
Minoru Karasawa
Michael Kaschak
Sophia Katrenko
Irvin Katz
Artem Kaznatcheev
Mark Keane
john kearns
Madeleine Keehner
Frank Keller
McRae Ken

Reviewers

William Kennedy
Trina Kershaw
Sangeet Khemlani
Peter Khooshabeh
Celeste Kidd
David Kieras
Stephen Killingsworth
Say Young Kim
ShinWoo Kim
Walter Kintsch
David Kirsh
Krystal Klein
Matthew Klenk
Heidi Kloos
Pia Knoeferle
Ken Koedinger
Judith Koehne
Boicho Kokinov
Talia Konkle
Ruud Koolen
Gert Kootstra
Vanja Kovic
Emiel Krahmer
Sarah Kriz
David Kronenfeld
Sven Kuehne
Maithilee Kunda
Aylin Kuntay
Christopher Kurby
Tamar Kushnir
Kai-Uwe Kühnberger
Tei Laine
John Laird
Brenden Lake
Kiran Lakkaraju
David Landy
Peter Lane
Pat Langley
Jill Lany
Jorge Larreamendy-Joerns
Brooke Lea
David Leake
Hee Seung Lee
Michael Lee
Benoit Lemaire
Bernard Lete
Sandro Leuchter
Dan Levin
William Levine
Roger Levy
Stephan Lewandowsky
Clayton Lewis
Joshua Lewis
Simon Li
Stephen Lim
Margarita Limon
Robb Lindgren
craig lindley
Shane Lindsay
Daniel Little
Daniel Hsi-wen Liu
Stefano Livi
Ken Livingston
Marcio Lobo Netto
Kate Lockwood
Jeffrey Loewenstein
Tania Lombrozo
Max Louwerse
Jessica Love
Will Lowe
Hongjing Lu
Christopher Lucas
claudio lucchiari
George Luger
Gary Lupyan
Dermot Lynott
Don Lyon
Andrew Maas
Jim MacGregor
Michael Mack
Carol J. Madden
Paul Maglio
Lorenzo Magnani
Frédéric Mailhot
Larry Maloney
Barbara Malt
Wenji Mao
Raymond A. Mar
Doug Markant
Art Markman
Jessecae Marsh
Sandra Marshall
Michael Matessa
Santosh Mathan
Moffat Mathews
Robert Mathews
Noboru Matsuda
Danielle Matthews
Justin Matthews
Camillia Matuk
Laura Matzen
Rich Mayer
Andre Mayers
Eva Mayr
Ralf Mayrhofer
Devin McAuley
Rachel McCloy
Drew McDermott
Bruce McLaren
Nicole McNeil
Katja Mehlhorn
Ben Meijering
Zulfiqar Ali Memon
David Mendonca
Stefania Mereu
David Miele
Gareth Miles
Craig Miller
John Paul Minda
Jelena Mirkovic
Daniel Mirman
Melanie Mitchell
Robert Mitchell
Antonija Tanja Mitrovic
Holger Mitterer
Naomi Miyake
Krishna Miyapuram
Lisette Mol
Padraic Monaghan

Reviewers

Daniel Montello
Bradley Morris
Anthony Morse
Johannes Moskaliuk
Jarrod Moss
Hanna Muenke
Edward Munnich
Paul Munro
Gregory Murphy
Kuninori Nakamura
Daniel Navarro
James Negen
Jonathan D. Nelson
Melissa Nelson
Hansjoerg Neth
Ben Newell
Penney Nichols-Whitehead
Marie Nilsenova
Farley Nobre
David Noelle
Timothy Nokes
Michael Noll-Hussong
Joseph Novak
Timothy O'Donnell
Padraig O'Seaghdha
Mike Oaksford
Klaus Oberauer
Amitash Ojha
Andrew Olney
Kristine Onishi
John Opfer
Tom Ormerod
Magda Osman
Anthony Otto
Ozge Ozturk
Martin Packer
Ulrike Pado
Martha Palmer
John Pani
Christopher Parisien
Gary Parker
Fey Parrill
Philippe Pasquier
Katerina Pastra
Andrea Patalano
James L. Pate
David Pautler
Lisa Pearl
Neal Pearlmutter
David Peebles
Marcie Penner-Wilger
Amy Perfors
Georgi Petkov
Alexander Petrov
Steven Phillips
Steven Piantadosi
Kathleen Pirog Revill
Mark Pitt
David Plaut
Timothy Pleskac
Isabella Poggi
Matthijs Pontier
James Pooley
Emmanuel Pothos
Sandeep Prasada
Richard Prather
Christoph Prof. Dr. Schommer
Athanassios Protopapas
Aryn Pyke
Gabriel Radvansky
Iyad Rahwan
Roxanne Raine
Ashwin Ram
Kiruthika Ramanathan
Michael Ranney
William Rapaport
David Rapp
Martina Rau
Wolfgang Rauch
William Raymond
Stephen Read
Stephen Reed
Patricia Reeder
Terry Regier
Bob Rehder
Jason Reiss
Roger Remington
Russell Revlin
Marjorie Rhodes
Daniel Richardson
Lindsey Richland
Tracy Riggins
Frank Ritter
Bethany Rittle-Johnson
Debi Roberson
Michael Roberts
Chris Robinson
Jennifer Roche
Florian Roehrbein
Ido Roll
Elena Cecilia Rosca
Paul Rosenbloom
Fred Rothganger
Brandon Roy
Timothy Rubin
J Rudine
Anna-Mari Rusanen
Manish Saggar
Eyal Sagi
Sarah Sahni
Shigeru Sakahara
Joanna Salapska-Gelleri
Ron Salden
Adam Sanborn
Catherine Sandhofer
Lelyn Saner
Ricardo Sanz
Brian Scassellati
Benjamin Scheibehenne
Katharina Scheiter
Paul Schermerhorn
Matthias Scheutz
Friederike Schlaghecken
Franz Schmalhofer
Ute Schmid
Hedda Rahel Schmidtke
Mike Schoelles
Lael Schooler
Anne Schueler

Reviewers

Holger Schultheis
Laura Schulz
Christian Schunn
Kristine Schuster Turko
Daniel Schwartz
Angela Schwering
J. Ignacio Serrano
Annalisa Setti
Carissa Shafto
Michael Shafto
Patrick Shafto
Alexei Sharpanskykh
Richard Shillcock
Hajime Shirouzu
Andrew Shtulman
Thomas Shultz
Martha Shumway
Anna Shusterman
Ekaterina Shutova
Winston Sieck
Noah Silbert
Vanessa Simmering
Nina Simms
Dylan Simon
Ut Na Sio
Jurgis Skilters
Steve Sloman
Vladimir Sloutsky
Filip Smolík
Melanie Soderstrom
Myeong-Ho Sohn
Ahmad Sohrabi
Firat Soylu
Ann Speed
Vishnu Sreekumar
Michelle St. Clair
Jon Star
Jim Staszewski
Maria Staudte
Laura Staum Casasanto
Courtney Stein
Rachel Stephens
Daniel Sternberg
Andrew Stewart
Terry Stewart
Mark Steyvers
gert storms
David Stracuzzi
Andrew Stull
Jennifer A. Sturm
Sakol Suethanapornkul
Jessica Sullivan
Ron Sun
gretchen sunderman
Bill Swartout
Christine Szostak
Niels Taatgen
federico tajariol
Monica Tamariz
Diana Tamir
alessandra tasso
Eric Taylor
Julia Taylor
Virginia Teller
David Temperley
Thora Tenbrink
Josh Tenenbaum
Robert Teszka
Paul Thagard
Jean-Pierre Thibaut
Paul Thibodeau
EriK Thiessen
John Thomas
Kevin Thomas
Michael Thomas
Clarissa Thompson
Ian Thornton
Barbara Tillman
Harry Tily
Maurizio Tirassa
Marc Tomlinson
Noriko Tomuro
Alexia Toskos Dils
Jozsef Toth
Greg Trafton
Jan Treur
Susan Trickett
Jennifer Trueblood
Jesse Tseng
Elio Tuci
Ryan Tweney
Christina Tzeng
Sebo Uithol
Tomer Ullman
David Uttal
Kevin Uttich
Jyotsna Vaid
Frederic Vallee-Tourangeau
Saskia van Dantzig
Ielka van der Sluis
Frank Van der Velde
Natalie van der Wal
Ludger van Eist
Erlijn van Genuchten
tamara van gog
Rianne van Lambalgen
Leendert van Maanen
Hedderik van Rijn
Arlette van Wissen
Bram Vandekerckhove
Ivan Vankov
wolf vanpaemel
Argiro Vatakis
Marga Vazquez
Richard Veale
Michelle Verges
Gaelle Villejoubert
Alessandro Vinciarelli
David Vinson
Ingmar Visser
Renan Vitral
Haley Vlach
John Voiklis
Regina Vollmeyer
Bettina von Helversen
Wouter Voorspoels
Edward Vul
Michael Waldmann
Erin Walker

Reviewers

David Waltz
Anne Warlaumont
Christopher Was
Jonathan Waskan
Christina Wasylyshyn
Kristin Weingartner
Matthew Welsh
Robert West
Gert Westermann
Rebecca White
Stefan Wierda
Harold Willaby
Joseph Jay Williams
Paul Williams
Andy Wills
William Wilson
Samuel Wintermute
Sascha Wolfer
John Wong
Shu-Chieh Wu
Ruth Wylie
Xu Xu
Takashi Yamauchi
Jie Yan
ChinLung Yang
Mark Yates
Sheng Kung Yi
Michael Yip
Robert Youmans
Richard Young
Erica Yu
Daniel Yurovsky
Carlos Zednik
Mathew Zeigenfuse
Hang Zhang
Shunan Zhang
Jiaying Zhao
Robert Zheng
Jing Zhu
Iraide Zipitria
Willem Zuidema
Ça˘grı Çöltekin